

Министерство науки и высшего образования Российской Федерации
 федеральное государственное автономное
 образовательное учреждение высшего образования
 «Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
 Направление подготовки 09.04.04 Программная инженерия
 Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Analysis of Olympic Games data (1896-2016) using clustering elements (Анализ данных Олимпийских игр (1896-2016) с использованием элементов кластеризации)

УДК 004.65:004.451:519.23:796.032.2

Студент

Группа	ФИО	Подпись	Дата
8ПМ0И	Сунь Хуншуай		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Меньшикова Е. В.	к.ф.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А.О.	к.т.н.		

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП
по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований

ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базами данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

Министерство науки и высшего образования Российской Федерации
 федеральное государственное автономное
 образовательное учреждение высшего образования
 «Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
 Направление подготовки (специальность) 09.04.04 Программная инженерия
 Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
 Руководитель ООП
 _____ Савельев А.О.
 (подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ПМОИ	Сунь Хуншуай

Тема работы:

Analysis of Olympic Games data (1896-2016) using clustering elements (Анализ данных Олимпийских игр (1896-2016) с использованием элементов кластеризации)	
Утверждена приказом директора (дата, номер)	№ 145-46/с от 25.05.2022

Срок сдачи студентом выполненной работы:	15.06.2022
--	------------

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

<p>Исходные данные к работе</p> <p><i>(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).</i></p>	<p>На сегодняшний день проведено 32 летние Олимпийские игры и 23 зимние Олимпийские игры. Мы проанализируем эти данные и разберемся в истории. Анализ позволит выявить любые тенденции в олимпийских играх, например, страны, которые доминируют в определенных видах спорта на протяжении 120 лет, и качественный анализ, чтобы ответить на вопрос, почему это явление происходит.</p>
---	---

<p>Перечень подлежащих исследованию, проектированию и разработке вопросов</p> <p><i>(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).</i></p>	<ol style="list-style-type: none"> 1. Подготовьте набор данных и поймите основную информацию о данных. 2. Считайте набор данных и выполните предварительную обработку данных. 3. Проанализируйте данные и сделайте выводы. 4. Построить модель машинного обучения для прогнозирования пола спортсмена на основе роста и веса. 5. Работа над разделом по финансовому менеджменту. 6. Работа над разделом по социальной ответственности.
<p>Перечень графического материала</p> <p><i>(с точным указанием обязательных чертежей)</i></p>	<ol style="list-style-type: none"> 1. Скриншот программы. 2. Диаграмма Ганта.

<p>Консультанты по разделам выпускной квалификационной работы</p> <p><i>(с указанием разделов)</i></p>	
<p>Раздел</p>	<p>Консультант</p>
<p>Основная часть</p>	<p>Доцент ОИТ ИШИТР, к.ф.-м.н., доцент Губин Е. И.</p>
<p>Финансовый менеджмент, ресурсоэффективность и ресурсосбережение</p>	<p>Доцент ОСГН ШБИП, к.ф.н., доцент Меньшикова Е. В.</p>
<p>Социальная ответственность</p>	<p>Доцент ООД ШБИП, к.б.н., доцент Антоневиц О. А.</p>

<p>Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику</p>	<p>1.03.2022</p>
--	------------------

Задание выдал руководитель:

<p>Должность</p>	<p>ФИО</p>	<p>Ученая степень, звание</p>	<p>Подпись</p>	<p>Дата</p>
<p>доцент ОИТ ИШИТР</p>	<p>Губин Е. И.</p>	<p>к.ф.-м.н., доцент</p>		<p>1.03.2022</p>

Задание принял к исполнению студент:

<p>Группа</p>	<p>ФИО</p>	<p>Подпись</p>	<p>Дата</p>
<p>8ПМОИ</p>	<p>Сунь Хуншуай</p>		<p>1.03.2022</p>

Министерство науки и высшего образования Российской Федерации
 федеральное государственное автономное
 образовательное учреждение высшего образования
 «Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
 Направление подготовки (специальность) 09.04.04 Программная инженерия
 Уровень образования магистратура
 Отделение школы (НОЦ) Информационных технологий
 Период выполнения весенний семестр 2021 /2022 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	15.06.2022
--	------------

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2022	Основная часть	70
10.06.2022	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2022	Социальная ответственность	10
10.06.2022	Английский язык	10

СОСТАВИЛ:

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

СОГЛАСОВАНО:

Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А. О.	к.т.н.		

**TASK FOR SECTION
«FINANCIAL MANAGEMENT, RESOURCE EFFICIENCY AND RESOURCE
SAVING»**

To the student:

Group	Full name
8PMOI	Hongshuai Sun

School	Information Tech & Robotics	Division	Big Data Solution
Degree	Master	Educational Program	09.04.04.Software Engineering

Title of graduation thesis:

Research on Credit Risk of Bank Lending in the Context of Big Data	
Input data to the section «Financial management, resource efficiency and resource saving»:	
1. <i>Resource cost of scientific and technical research (STR): material and technical, energetic, financial and human</i>	– Salary costs – 196363.4 – STR budget – 140287.5
2. <i>Expenditure rates and expenditure standards for resources</i>	– Electricity costs – 5,8 rub per 1 kW
3. <i>Current tax system, tax rates, charges rates, discounting rates and interest rates</i>	– Labor tax – 27,1 %; – Overhead costs – 30%;
The list of subjects to study, design and develop:	
1. <i>Assessment of commercial and innovative potential of STR</i>	– comparative analysis with other researches in this field;
2. <i>Development of charter for scientific-research project</i>	– SWOT-analysis;
3. <i>Scheduling of STR management process: structure and timeline, budget, risk management</i>	– calculation of working hours for project; – creation of the time schedule of the project; – calculation of scientific and technical research budget;
4. <i>Resource efficiency</i>	– integral indicator of resource efficiency for the developed project.
A list of graphic material (with list of mandatory blueprints):	
1. <i>Competitiveness analysis</i> 2. <i>SWOT- analysis</i> 3. <i>Gantt chart and budget of scientific research</i> 4. <i>Assessment of resource, financial and economic efficiency of STR</i> 5. <i>Potential risks</i>	

Date of issue of the task for the section according to the schedule	
--	--

Task issued by adviser:

Position	Full name	Scientific degree, rank	Signature	Date
Associate professor	E.V. Menshikova	PhD		

The task was accepted by the student:

Group	Full name	Signature	Date
8PMOI	Hongshuai Sun		

**TASK FOR CHAPTER
«SOCIAL RESPONSIBILITY»**

Student:

Group 8PMOI		Name Hongshuai Sun	
School	School of Engineering of Information Technology and Robotics	Division	Big Data Solution
Educational level	Master degree	Course/Specialty	09.04.04 Software Engineering

Topic of FQW:

Analysis of Olympic Games data (1896-2016) using clustering elements	
Initial data for the chapter «social responsibility»:	
1. Characteristics of the researched object (substance, material, device, algorithm, technique, working area)	<ul style="list-style-type: none"> – Data preparation and data analysis for the 120-year Olympic Games from the Athens Olympics in 1896 to the Rio de Janeiro Olympics in 2016. – Analysis and prediction using multiple machine learning algorithms.
List of questions to be researched, designed and developed:	
1. Legal and organizational issues of occupational safety <ul style="list-style-type: none"> – consider special (specific to the projected work area) law norms of labor legislation. – indicate the features of the labor legislation in relation to the specific conditions of the project. 	<ul style="list-style-type: none"> – GOST 12.2.032-78 SSBT. Workplace when performing work while sitting General ergonomic requirements. – GOST 12.1.019-2017 Electrical safety. General requirements and nomenclature of types of protection.
2. Occupational safety: 2.1. Analysis of the identified harmful and dangerous factors: the source of factor, the impact on human's body 2.2 Suggest measures to reduce the impact of identified harmful and dangerous factors	<ul style="list-style-type: none"> – Increased voltage in an electrical circuit, the closure of which can pass through the human body – Lack or lack of natural light, insufficient illumination – Physical overload (static - long-term preservation of a certain posture) – Monotony of work – Increased noise level
3. Environmental Safety: Influence on the atmosphere, hydrosphere, lithosphere	<ul style="list-style-type: none"> – Atmosphere: Computers and peripherals contain materials that can be harmful to the environment. These materials can contain high concentrations of heavy metals such as cadmium, lead, or mercury. – Hydrosphere, lithosphere: Some electronic scrap components, such as CRTs, may contain contaminants such as Pb, Cd, Be or brominated flame retardants.
4. Emergency Safety: describe the most likely emergency situation	<ul style="list-style-type: none"> – Fire-
Date issue of the task for the chapter	

Consultant:

Post	Name	Academic degree	Date	Signature
Associate professor	Antonevich O.A	PhD		

Student:

Group	Name	Date	Signature
8PMOI	Hongshuai Sun		

Abstract

Final qualifying work 72 pages, 46 figures, 20 tables, 22sources.

Keywords: data analysis, data preparation, data cleaning, linear regression, decision tree, random forest, Gender Prediction Model.

Today, 32 Summer Olympic Games and 23 Winter Olympic Games have been held. We will analyze this data and understand the history. Before that, we will data preparation, data cleaning. The analysis will find out any trends in Olympics games such as the country that dominating in certain sports for 120 years and qualitative analysis to answer the question why this phenomenon happens. Further, our journey steps to the side of statistics with regression problem to estimate the missing value in certain variables. It's interesting to find out specific, rightful, and useful regression method to handle this problem.

Publications:

XIX Международной научно-практической конференции студентов, аспирантов и молодых ученых.

МОЛОДЕЖЬ И СОВРЕМЕННЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
МСИТ-2022

Томск, 21 -25 марта 2022г.

Сунь Хуншуай (Томский политехнический университет) “Analysis of Olympic Games data (1896-2016) using clustering elements”

Content

Terms, abbreviations and convention	13
INTRODUCTION.....	14
1. Dataset introduction:	15
2.Prepare dataset.....	16
2.1 Questions.....	17
2.2 Assumptions.....	17
2.3 Approaches.....	17
2.4 Handle outliers in dataset	18
2.5 Handle NA value.....	19
2.5.1 Linear Regression model to Pre-processing.....	22
2.5.2 Decision Tree Regression model to Pre-processing	24
2.6 Duplicate value.....	26
3. Data analysis	26
3.1 Bar plot of Countries.....	27
3.2 Bar plot of top ten sport	28
3.3 Scatterplot between Number of Sport and Total of Medals	29
3.4 The number of participants is increasing year by year	29
3.5 The number of female athletes is gradually increasing	30
3.6 The variety of competitions is gradually enriched.....	31
3.7 Sports with the most participation	31
3.8 The best of country / region	32
3.8.1 The country with the most participation in the Olympic Games.....	32
3.8.2 Which country sends the most athletes	33

3.8.3 Which city has hosted the most Olympic Games?.....	33
3.8.4 Which country won the most awards?	34
3.9 The best of athlete	34
3.9.1 The youngest player.....	35
3.9.2 The oldest player	35
3.9.3 The shortest player	36
3.9.4 The tallest player.....	36
3.9.5 The lightest player	36
3.9.6 The heaviest player.....	37
3.9.7 Athlete with the most participation in the Olympic Games.	37
4.Model building	37
4.1 Decision tree model.....	39
4.2 Random forest model	40
5.Financial management, resource efficiency and resource saving.....	42
5.1 Competitiveness analysis of technical solutions.....	42
5.2 SWOT analysis.....	44
5.3 Project Initiation.....	45
5.4 Scientific and technical research budget.....	48
5.5 Evaluation of the comparative effectiveness of the project.....	54
5.6 Conclusion of financial management.....	58
6. Social responsibility	59
6.1 Introduction	59
6.2 Legal and organizational issues of occupational safety	59
6.3 Basic ergonomic requirements for the correct location and arrangement of researcher's workplace	61
6.4 Occupational safety	61
6.5 Ecological safety	66

6.6 Safety in emergency	67
6.7 Conclusion of social responsibility	69
6.8 Reference of social responsibility	70
Conclusion.....	71
Reference.....	72

Terms, abbreviations and convention

EDA - Exploratory Data Analysis

RMSE - root-mean-square error

MAE - mean absolute error

LR - Linear Regression

DTs - Decision Tree

RF – Random Forests

INTRODUCTION

The Olympic Games originated in ancient Greece more than two thousand years ago and got its name because it was held in Olympia. The first Olympic Games was held in 1896 and the first Winter Olympic Games was held in 1924. It is the most influential sports event in the world.

In 1896, the Olympic Games, which had been suspended for 1,500 years, were finally re-hosted. This was also the first modern Olympic Games. Today, 32 Summer Olympic Games and 23 Winter Olympic Games have been held.

We can analyze this data and understand the history with the following questions.

1. Geographically, which countries host the most Olympic Games? The most athletes participating? Most awarded?
2. Personally, how have male and female athletes performed over the years?
3. In terms of projects, are there any projects that are the strengths of certain countries/regions?
4. By using machine learning models. Is it possible to use height and weight to predict gender?

We need to do following steps.

1. The analysis will find out any trends in Olympics games such as the country that dominating in certain sports for 120 years and qualitative analysis to answer the question why this phenomenon happens.
2. Further, our journey steps to the side of statistics with regression problem to estimate the missing value in certain variables.
3. It's interesting to find out specific, rightful, and useful regression method to handle this problem.

Analysis tools: Power BI + Excel + Python

1. Dataset introduction:

Data Sources:

<https://www.heywhale.com/mw/dataset/5b62ca77a711e60010ab1154>

There are two pieces of data, one athlete_events.csv, which contains the basic biological data and medal results of participating athletes.

A noc_regions.csv is the 3-letter code of the National Olympic Committee and the corresponding country information.

The athlete data includes the data of each athlete participating in the previous Olympic Games from 1896 to 2016, with a total of 271,116 rows and 15 fields, each row corresponds to the information of each athlete participating in the Olympic Games.

The general data situation is shown in the following table.

	Variable Name	Description	Type
0	ID	The unique number of each athlete, a total of 135571 numbers	integer
1	Name	Athlete name	object
2	Sex	Athlete gender, F is female, M is male	object
3	Age	Athlete's age	float
4	Height	Athlete's height, in cm	float
5	Weight	Athlete weight in kg	float

	Variable Name	Description	Type
6	Team	Athlete teams such as China	object
7	NOC	National Olympic Committee three-letter code	object
8	Games	which olympics the athlete participated in	object
9	Year	years	integer
10	Season	season	object
11	City	host city such as Beijing	object
12	Sport	sports such as basketball	object
13	Event	specific programs, such as men's basketball	object
14	Medal	Medals such as gold, silver, bronze or none	object

Table 1. athlete_events's dataset basic info

2.Prepare dataset

Before data analysis, the data set needs to be processed. The data set is generally repeated rows, noise values, noise labels, etc., which need to be corrected step by step for the problems of the data set.

2.1 Questions

To ensure the effectiveness of the research, the following question will be answered systematically:

- 1.How does data pre-processing will be developed?
- 2.What kind of methods is more useful to handle and fill missing value in the certain columns?
- 3.Are there unique trends visually in the data?

2.2 Assumptions

The scopes of research are listed:

- 1.Each rows is the unique person in difference time of period
- 2.The chosen columns must have high correlation

2.3 Approaches

To answer the questions , the following methods and approaches will be considered:

Several columns will be dropped because it has no enough correlation with the main analysis.

To handle missing value, the linear regression and decision tree regressor will be compared.

Evaluation metrics use root mean square error, mean absolute error, and Pearson correlation.

Explanatory Data Analysis (EDA) is rightful method and mostly used to find out pattern in the whole data.

2.4 Handle outliers in dataset

In order to find outliers, we need to analyze age, weight, height. we use histogram to display the characteristics of the data.

Core code:

```
Age = df_used.Age
Weight = df_used.Weight
Height = df_used.Height
bin_width = 5
bins = int((max(Age)-min(Age))/bin_width)
res1 = plt.hist(Age, bins = bins)
plt.show()

bin_width = 7
bins = int((max(Weight)-min(Age))/bin_width)
res2 = plt.hist(Weight, bins = bins)
plt.show()

bin_width = 7
bins = int((max(Height)-min(Age))/bin_width)
res3 = plt.hist(Height, bins = bins)
plt.show()
```

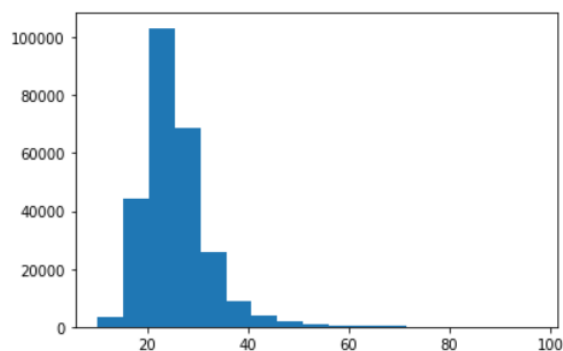


Figure 1. Histogram of Athlete's Age

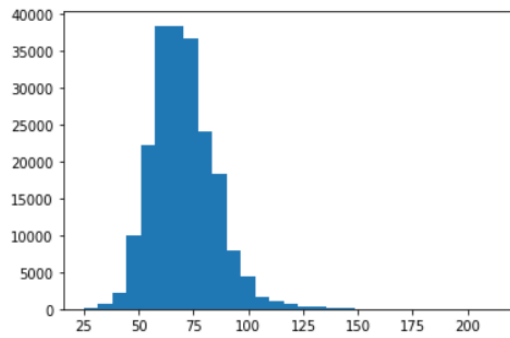


Figure 2. Histogram of Athlete's Weight

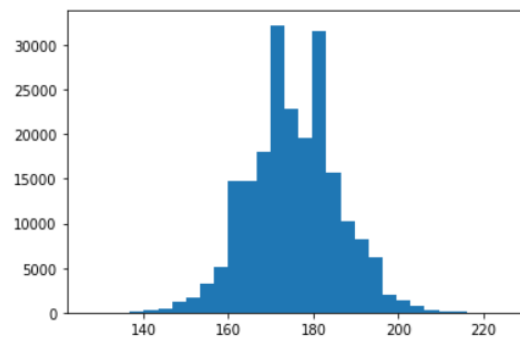


Figure 3. Histogram of Athlete's Height

According to those histogram, athlete's age and weight are right-skewed while the height is bell-shape, Normal distribution.

- The oldest of athlete's age is 97. It's unnatural. So, we need to do pre-processing
- The maximum of athlete's weight is about 214 kg. This is why the histogram would be rightskewed

2.5 Handle NA value

This data has missing values in the Age, Height, Weight, and Medal columns:

In order to fill missing value, We need to know the basic information of our data.

Core code :

```
data =
```

```

pd.read_csv('D:/Dataset/athlete_events.csv')

df = pd.DataFrame(data)

df_copy = df.copy()

df.duplicated()

df.drop_duplicates()

df.info()

```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aafink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

Figure 4. athlete_events's dataset value info

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   ID           271116 non-null  int64
1   Name        271116 non-null  object
2   Sex         271116 non-null  object
3   Age         261642 non-null  float64
4   Height      210945 non-null  float64
5   Weight      298241 non-null  float64
6   Team        271116 non-null  object
7   NOC         271116 non-null  object
8   Games       271116 non-null  object
9   Year        271116 non-null  int64
10  Season      271116 non-null  object
11  City        271116 non-null  object
12  Sport       271116 non-null  object
13  Event       271116 non-null  object
14  Medal       39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB

```

Figure 5. Description of column type

```

Dimension of training data:
271116 rows and 15 columns

ID           0
Name         0
Sex          0
Age         9474
Height      60171
Weight      62875
Team        0
NOC         0
Games       0
Year        0
Season      0
City        0
Sport       0
Event       0
Medal      231333
dtype: int64

```

Figure 6. Missing value each columns

```
ID          135571
Name        134732
Sex         2
Age         74
Height      95
Weight      220
Team        1184
NOC         230
Games       51
Year        35
Season      2
City        42
Sport       66
Event       765
Medal       3
dtype: int64
```

Figure 7. Unique value each columns

From the above information, we can find that:

Three important variables for deep analysis need manipulation. These are athlete's age, weight, and height. So, it needs to find out the best method to fill those missing value properly.

We use scatter plots to analyze the correlation between variables.

Core code:

```
scatter1 = plt.scatter(Age, Height)
plt.show()
scatter2 = plt.scatter(Age, Weight)
plt.show()
scatter3 = plt.scatter(Weight, Height)
plt.show()
```

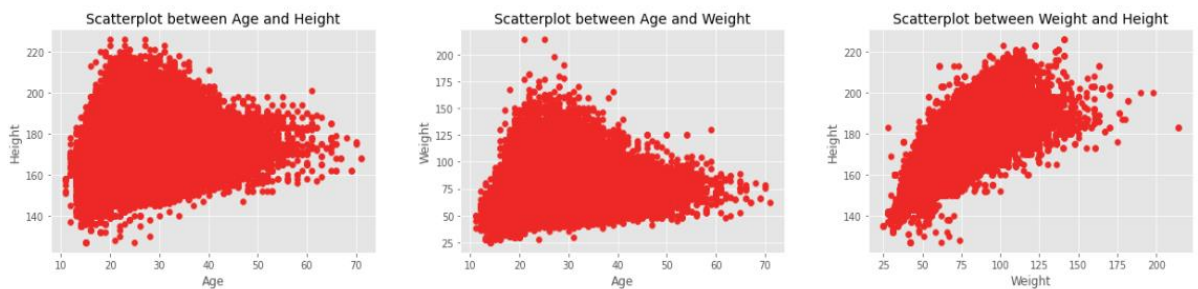


Figure 8. Scatterplot between Age, Height and Weight

We can find that, the correlation is the statistic indicating the

relationship between two variables in the data. After exploring the numerical variables, the correlation between athlete's weight and height is high.

Then we need to choose a suitable model to deal with missing values.

2.5.1 Linear Regression model to Pre-processing

Linear regression(LR) was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

Core code:

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
x_new = Height
y_new = Weight
x_train, x_test, y_train, y_test =
train_test_split(x_new, y_new, test_size=0.5)
lr.fit(x_train, y_train)
plt.scatter(x, y)
plt.show()
lr.fit(x_test, y_test)
plt.scatter(x, y)
plt.show()
w = lr.coef_[0][0]
b = lr.intercept_[0]
print("Intercept: ", b)
print("Coefficient: ", w)
```

First, we use linear Regression model to train our data.

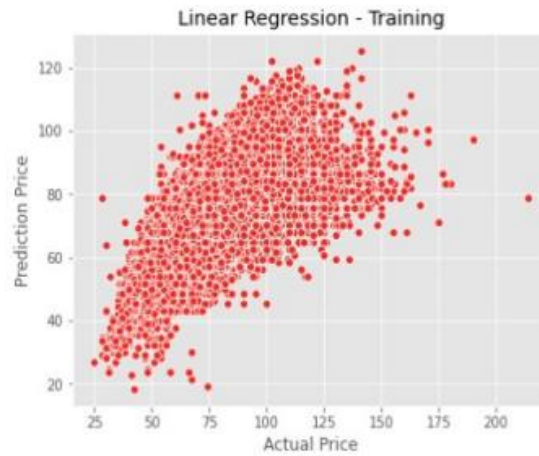


Figure 9. Linear Regression - Training

Then, we need to test our data.

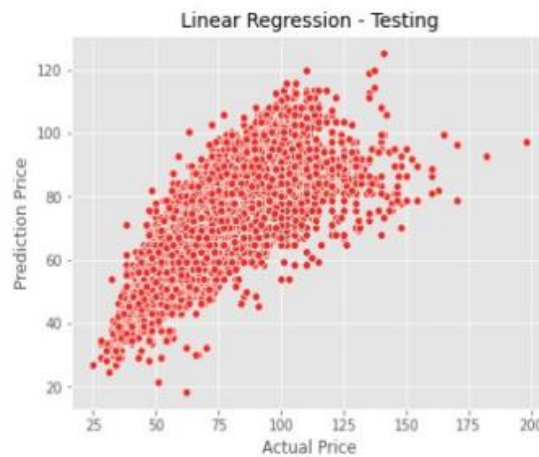


Figure 10. Linear Regression - Testing

And, we use Cross validation method to evaluate our model.

```

RMSE in CV - 1: 8.774727 and MAE: 6.258586
RMSE in CV - 2: 8.55525 and MAE: 6.065735
RMSE in CV - 3: 8.793303 and MAE: 6.19477
RMSE in CV - 4: 8.461461 and MAE: 6.039466
RMSE in CV - 5: 8.589853 and MAE: 6.084789
RMSE in CV - 6: 8.575798 and MAE: 6.123329
RMSE in CV - 7: 8.577335 and MAE: 6.086527
RMSE in CV - 8: 8.65246 and MAE: 6.145842
RMSE in CV - 9: 8.91851 and MAE: 6.243718
RMSE in CV - 10: 8.766976 and MAE: 6.186461
Average of RMSE: 8.666567278592114
Average of MAE: 6.142922312460948

```

Figure 11. 10 times – Cross validation

From the above information, we can find that:

The RMSE of prediction is about 8.66 where it is comparable with

the standard deviation of response variable. So, the linear regression model is quite good. The model equation is:

Intercept: -116.85

Coefficient: 1.06

2.5.2 Decision Tree Regression model to Pre-processing

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

We will use Decision Tree Regression.

Core code:

```
from sklearn import tree
clf = tree.DecisionTreeRegressor()
x_new = Height
y_new = Weight
x_train, x_test, y_train, y_test =
train_test_split(x_new, y_new, test_size=0.5)
clf.fit(x_train, y_train)
plt.scatter(x, y)
plt.show()
clf.fit(x_test, y_test)
plt.scatter(x, y)
plt.show()
```

This time, we use Decision Tree Regression model to train our data.



Figure 12. Decision Tree Regression – Training

Then, we need to test our data.



Figure 13. Decision Tree Regression - Testing

We use Grid-search to get optimum hyper parameters.

```
Best hyperparameters :
{'max_depth': 10, 'min_samples_leaf': 100, 'min_samples_split': 2}

Best evaluation :
-8.634977365979429

Best model of Decision Tree:
DecisionTreeRegressor(max_depth=10, min_samples_leaf=100)
```

Figure 14. Optimum hyper parameters

And, we compare different models.

Regression Model	RMSE Training	RMSE Validation	MAE Training	MAE Validation	Pearson Training	Pearson Validation
Linear Regression	8.66746	8.68068	6.14282	6.1264	0.79574	0.79806
Decision Tree Baseline	8.62923	8.64788	6.11897	6.11031	0.79777	0.79977
Decision Tree Grid-Search	8.63171	8.65194	6.12056	6.11259	0.79764	0.79956

Figure 15. RMSE,MAE and Pearson of different models

Final, we use Linear Regression Model. because it is simplicity.

2.6 Duplicate value

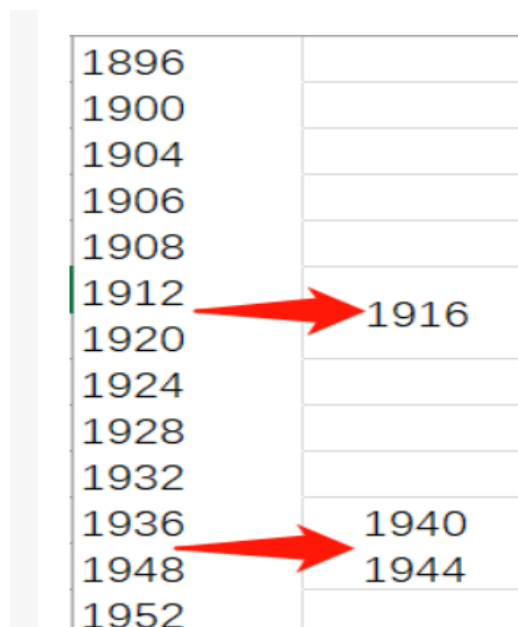
An ID represents an event that an athlete participates in, so it is normal for IDs to be duplicated because an athlete may participate in more than one competition.

3. Data analysis

The Summer Olympics have been held every 4 years since 1896. This data is up to 2016, with a total of 29 held and the Winter Olympics held 22 times.

How come the Summer Olympics here are only held 29 times? Which 3 sessions have not been held?

In fact, you can see the clues by looking at the years. Because of the two world wars, the three Olympic Games originally planned to be held in 1916, 1940 and 1944 became blank.



1896	
1900	
1904	
1906	
1908	
1912	
1916	1916
1920	
1924	
1928	
1932	
1936	1940
1948	1944
1952	

Figure 16. the years that did not hold successfully

3.1 Bar plot of Countries

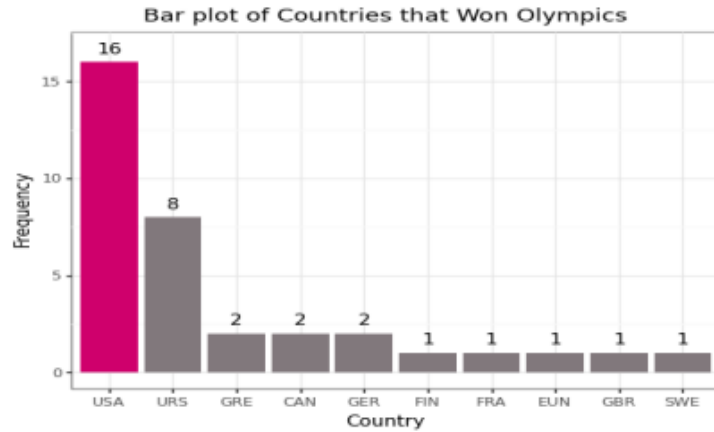


Figure 17. bar plot of countries that won Olympics

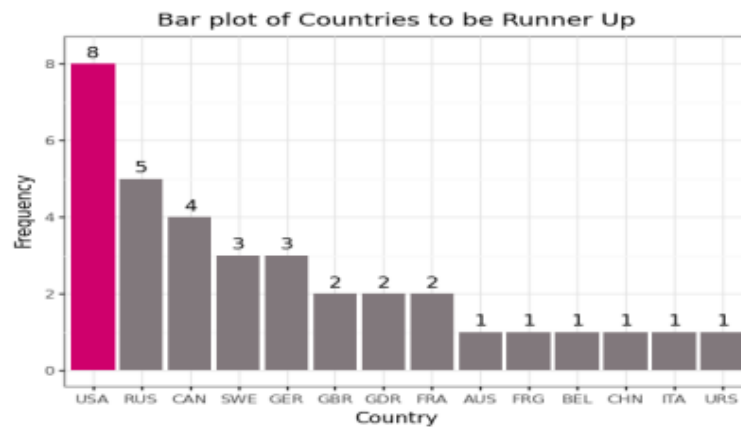


Figure 18. bar plot of countries to be Runner up

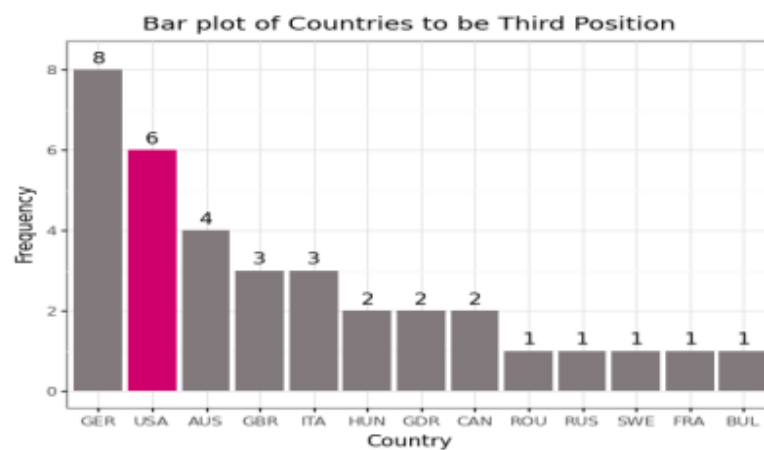


Figure 19. bar plot of countries to be Third Position

Findings:

- For all Olympics event, United State of America (USA) have won the competition 16 times as general champion. Further, Uni Soviet has 8 times as

general champion .

- Despite not being 1st position, USA also active as runner up and 3rd position .
- Uni Soviet is a rival of USA .
- German and Canada are the other rival of USA with good potency .

3.2 Bar plot of top ten sport

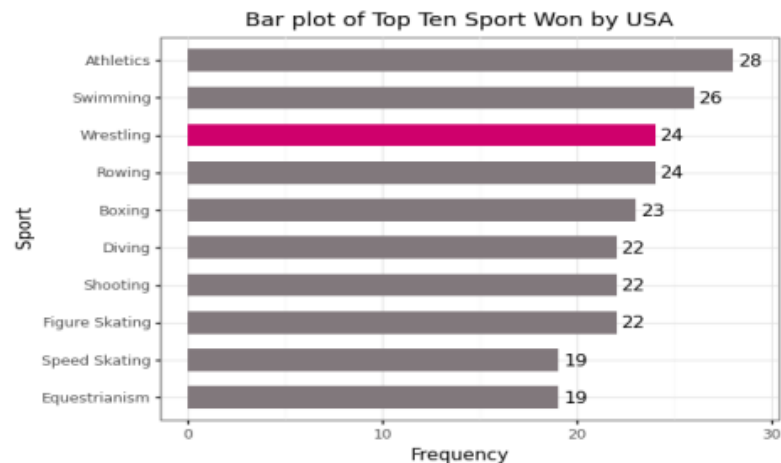


Figure 20. bar plot of top ten sport Won by USA

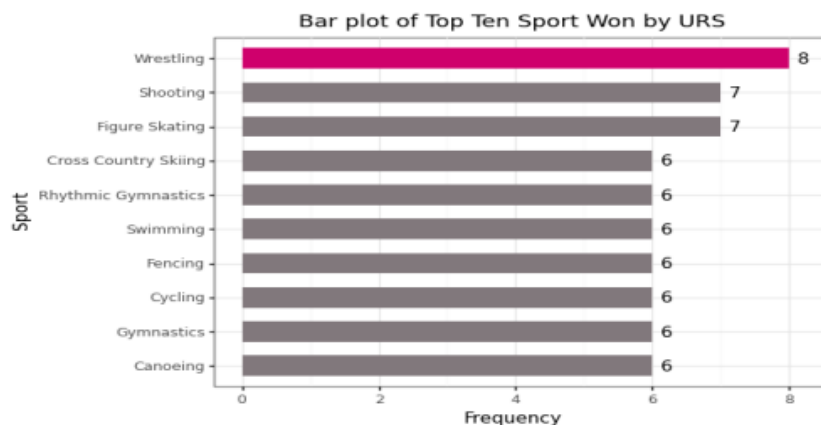


Figure 21. bar plot of top ten sport Won by URS

Findings:

- As the rival of USA, Uni Soviet has the strongest sport with highest number of medals, that is wrestling .
- The USA's sport with highest number of medals is athletics (28). It doesn't include in top ten sport won by the Uni Soviet.
- Rowing, boxing, and diving can be optimized by USA in order to beat the real

rival of Uni Soviet.

3.3 Scatterplot between Number of Sport and Total of Medals

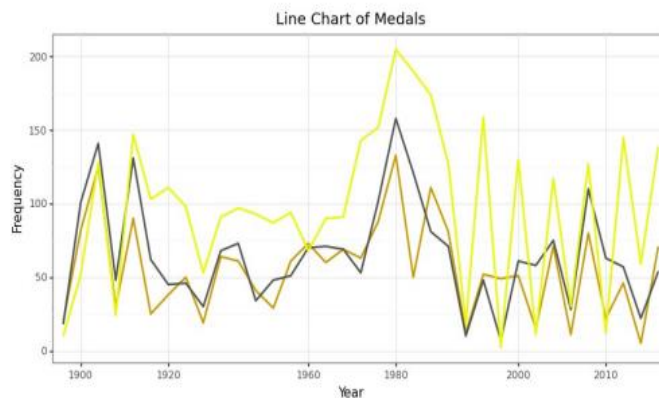


Figure 22. line Chart of Medals

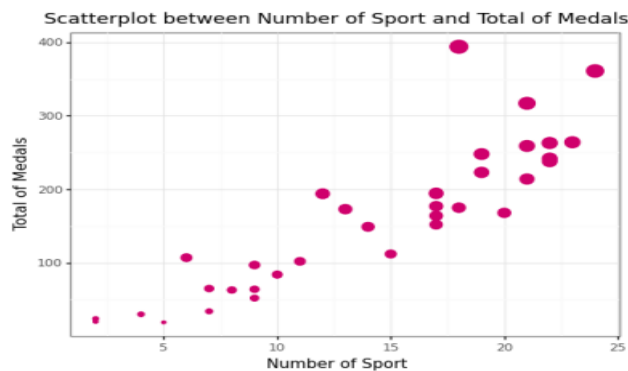


Figure 23. Scatterplot between number of sport and total of medals

Findings:

- In 1980, it was a year with the highest number of medals to be contested.
- Of course, there is high positive correlation between number of sports with the total medals won by country (0.883).

3.4 The number of participants is increasing year by year

From the first modern Olympic Games in 1896 with 176 athletes participating in 12 countries, to the 2016 London Olympics with 11,179 athletes participating in 206 countries, the number of athletes participating has gradually increased (11,669 athletes participating in the 2020 Tokyo Olympics, 204 countries) / region), the figure below is a graph of the number of athletes participating in the Summer

Olympics and the number of parameter countries.

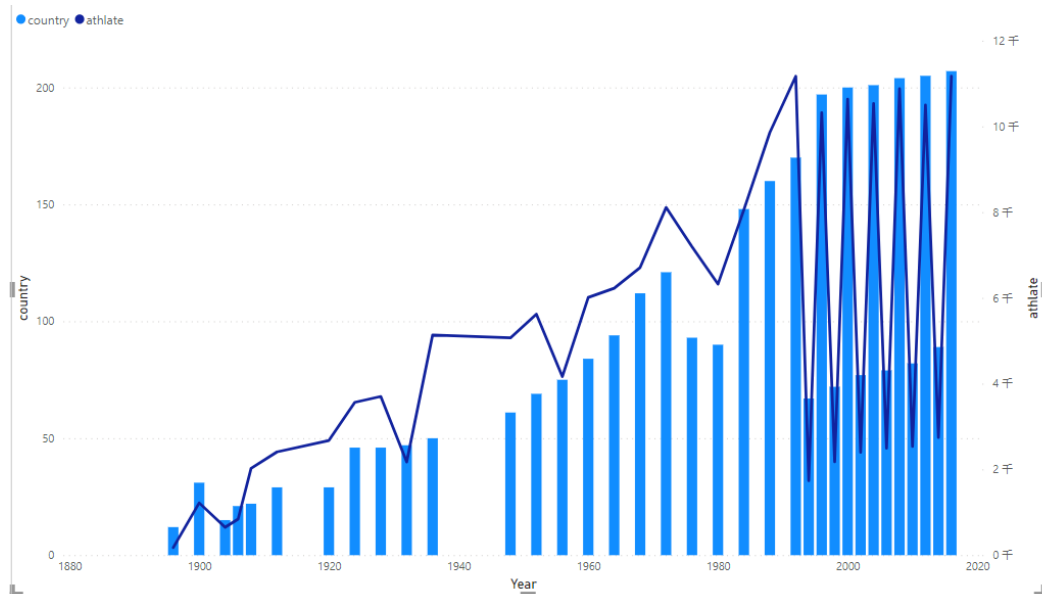


Figure 24. Number of participating countries and athletes over the years

3.5 The number of female athletes is gradually increasing

In 1900, 23 women participated in the Olympic Games for the first time, accounting for 1.87%. Since 1980, the number of women participating in the Olympic Games has increased significantly. By 2016, 5,034 female athletes participated, accounting for 45%.

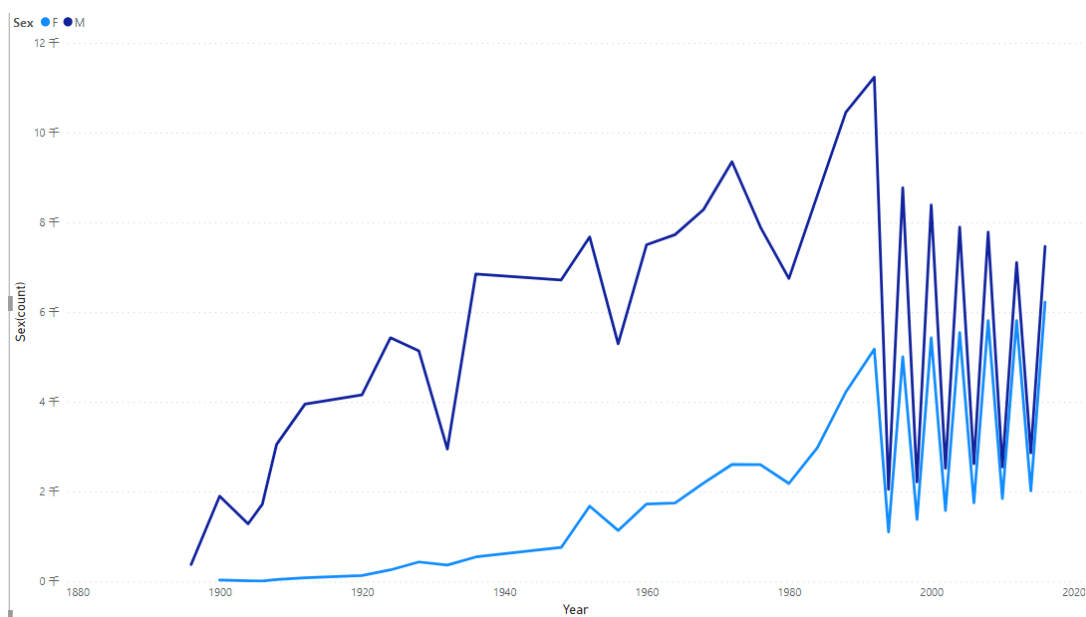


Figure 25. Number of male and female athletes

The ratio of male to female athletes in history.

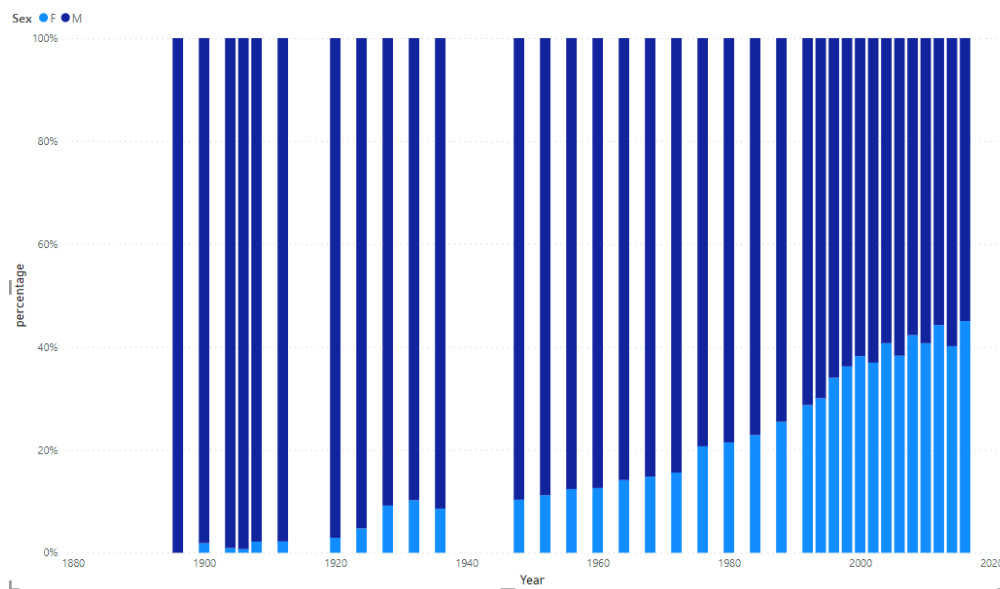


Figure 26. The ratio of men to women at the Olympic Games

3.6 The variety of competitions is gradually enriched

The types of events in the previous Olympic Games have also gradually increased. In the 1896 Summer Olympics, there were only 9 events, and by 2016, there were 36 events.

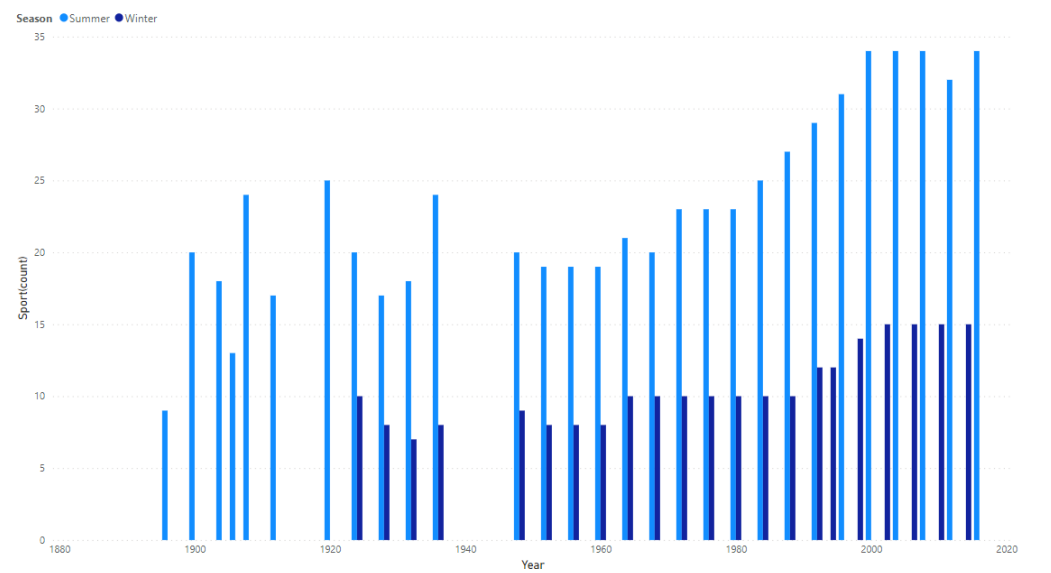


Figure 27. The number of projects

3.7 Sports with the most participation

The sport with the most participation in history is Athletics, followed by swimming, rowing, and football.



Figure 28. The most attended sport in history

The proportion of the number of male and female athletes participating in these events is shown in the figure below. In 13 events including baseball, Nordic biathlon (Winter Olympics), tug-of-war, rugby, polo, and lacrosse, no female athletes participated at all, but in rhythmic gymnastics, There are no male athletes in synchronized swimming or softball.

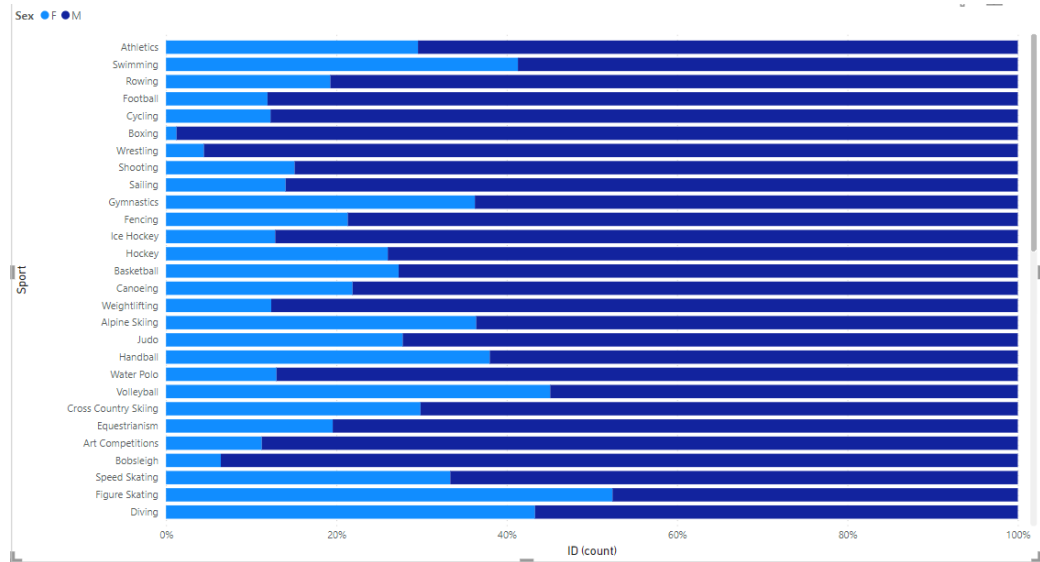


Figure 29. Percentage of men and women participating in projects

3.8 The best of country / region

3.8.1 The country with the most participation in the Olympic Games

A total of 208 countries/regions have participated in the Olympic Games in history, Australia, France, Greece, Italy, Sweden have participated in all 29 Summer

Olympics.

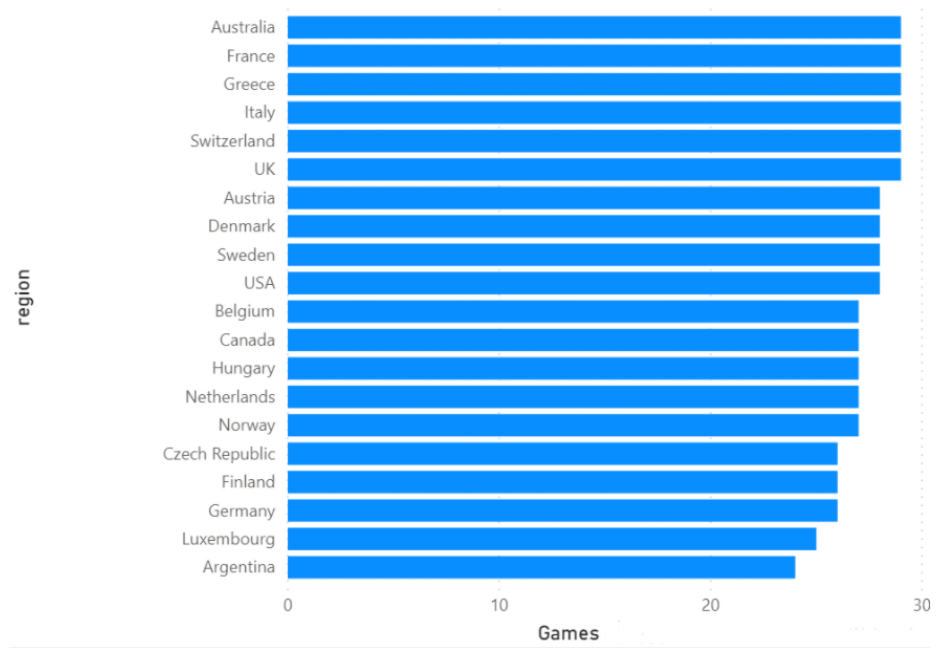


Figure 30. Number of countries participating in the Olympic Games

3.8.2 Which country sends the most athletes

It can be seen that the United States dispatched the largest number of people to participate in the Olympic Games in history, followed by Germany.

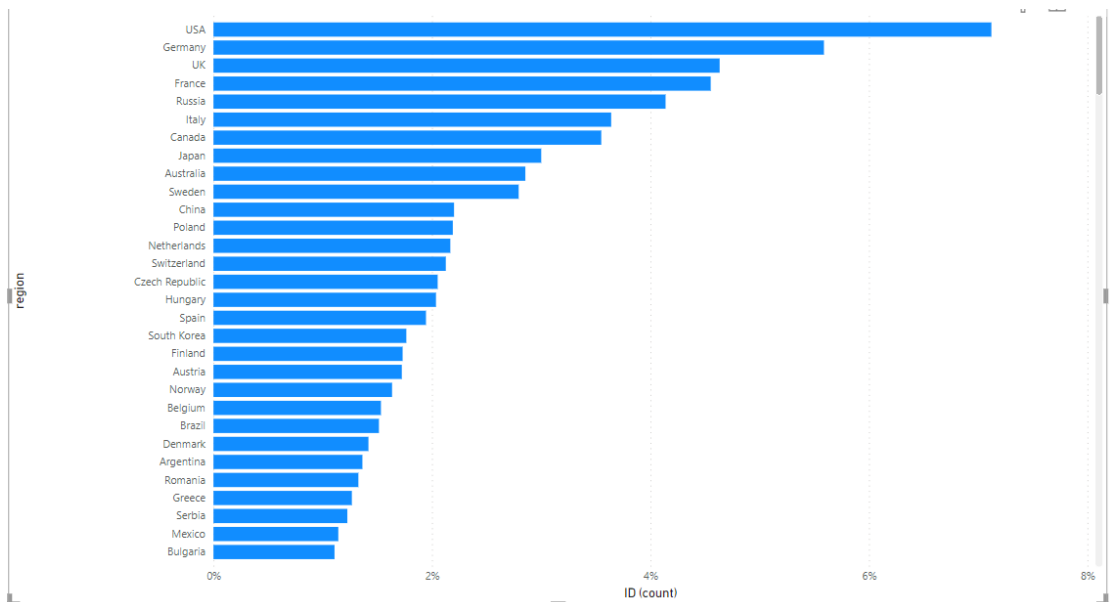


Figure 31. number of athletes

3.8.3 Which city has hosted the most Olympic Games?

A total of 42 cities have hosted the Olympic Games in history, of which Athens and London have held three, Innsbruck, Lake Placid, Los Angeles, Paris, St.

Moritz, and Stockholm have held two, and the remaining cities have only held once.

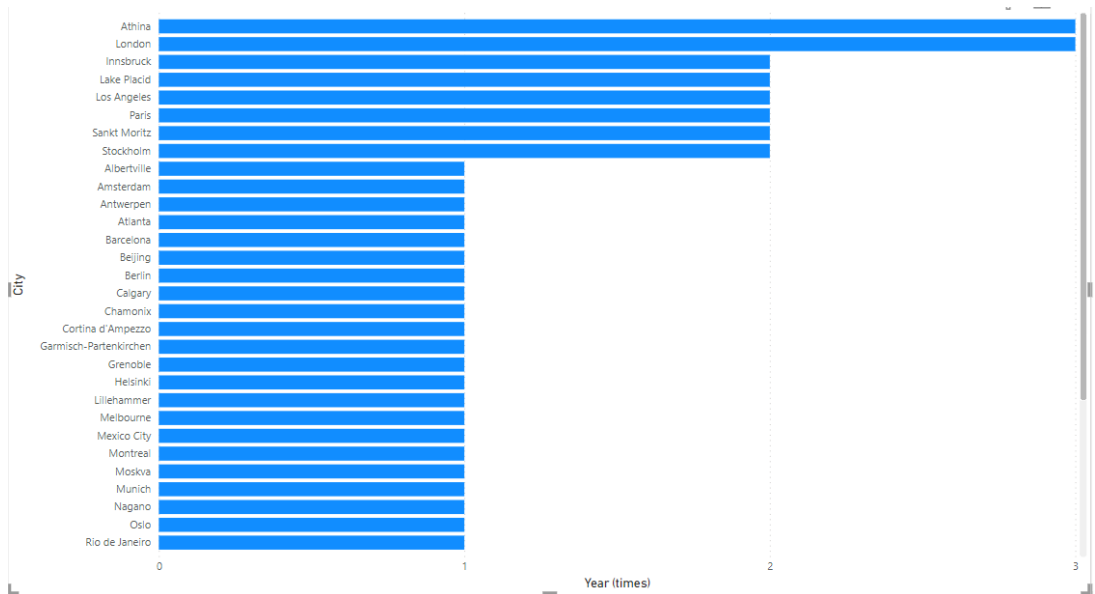


Figure 32. city hosting the olympics

3.8.4 Which country won the most awards?

The country with the most medals in history is the United States, followed by Russia, Germany, and the United Kingdom. At this year's Tokyo Olympics, we won 38 gold medals and 88 medals.

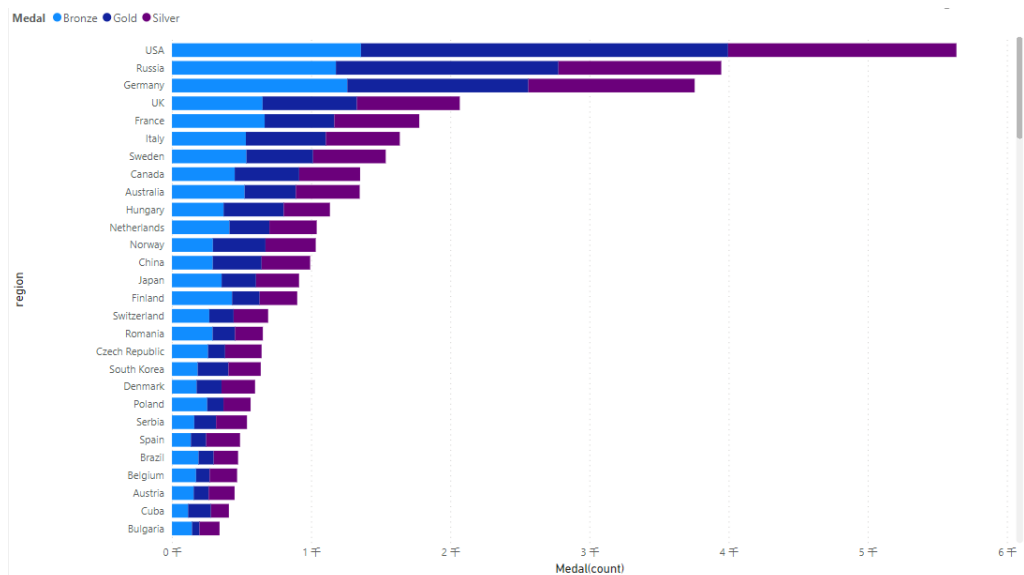


Figure 33. city hosting the olympics

3.9 The best of athlete

By looking at the age distribution of athletes, it can be known that the number of athletes aged 21 to 24 is the largest, and both male and female athletes are similar.

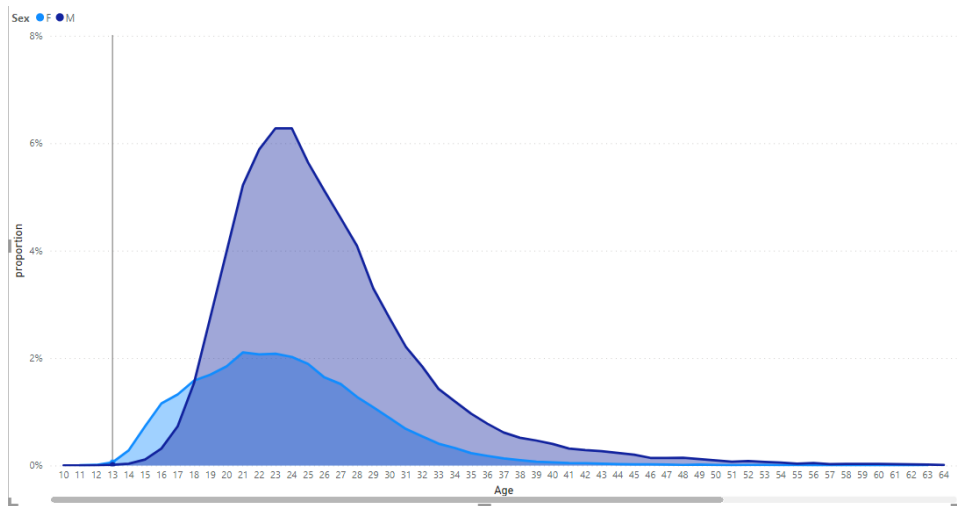


Figure 34. Age distribution of athletes

It can also be seen from the age distribution of the players who won the medals, the players aged 22-23 won the most prizes.

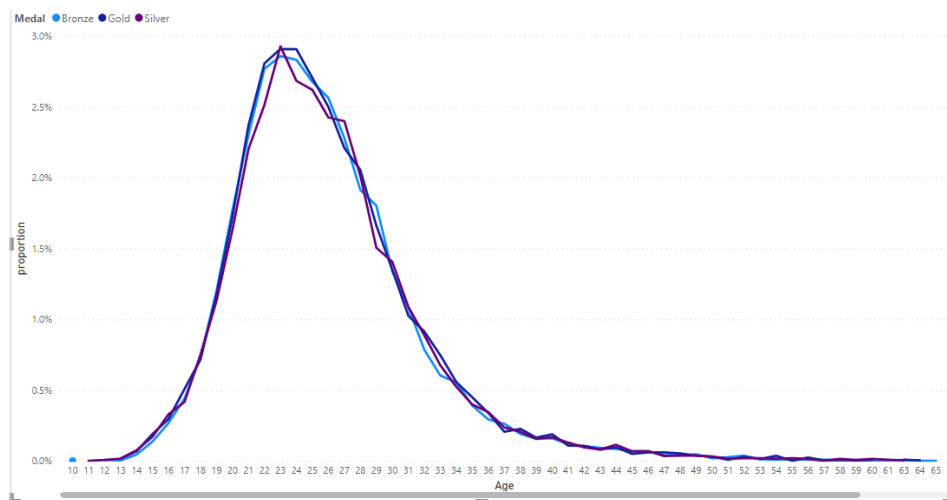


Figure 35. Age distribution of medalists

3.9.1 The youngest player

The youngest is 10 years old. I checked it and found it to be true. Dimitrios Loundras, a 10-year-old boy, won the bronze medal in the gymnastics men's team at the 1896 Athens Olympics, the youngest in the history of the Olympic Games athlete.

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
71691	Dimitrios Loundras	M	10	NA	NA	Ethnikos Gymnast	GRE	1896 Summer	1896	Summer	Athina	Gymnastics	Gymnastics Men's Parallel Bars, Teams	Bronze

Figure 36. The youngest player

3.9.2 The oldest player

Then the 97-year-old athlete, I don't think it is an outlier. This John Quincy Adams Ward participated in the 1928 Amsterdam Olympic Games. Although he did not win a medal in the art sculpture project, he became the oldest at the age of 97. Olympian.

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
128715	John Quincy Adams Ward	M	97	NA	NA	United States	USA	1928 Summer	1928	Summer	Amsterdam	Art Competitions	Art Competitions Mixed Sculpturing, Statues	NA

Figure 37. The oldest player

3.9.3 The shortest player

There are two players with the lowest height, both 127cm, one male and one female.

One is Rosario Briones, a female gymnast from Mexico who competed in the 1968 Mexico Olympics.

The other is Lyton Levison Mphande, a male boxer from Malawi who competed in the 1988 Seoul Olympics.

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
15150	Rosario Briones	F	15	127	42	Mexico	MEX	1968 Summer	1968	Summer	Mexico City	Gymnastics	Gymnastics Women's Uneven Bars	NA
15150	Rosario Briones	F	15	127	42	Mexico	MEX	1968 Summer	1968	Summer	Mexico City	Gymnastics	Gymnastics Women's Individual All-Around	NA
15150	Rosario Briones	F	15	127	42	Mexico	MEX	1968 Summer	1968	Summer	Mexico City	Gymnastics	Gymnastics Women's Balance Beam	NA
15150	Rosario Briones	F	15	127	42	Mexico	MEX	1968 Summer	1968	Summer	Mexico City	Gymnastics	Gymnastics Women's Team All-Around	NA
15150	Rosario Briones	F	15	127	42	Mexico	MEX	1968 Summer	1968	Summer	Mexico City	Gymnastics	Gymnastics Women's Floor Exercise	NA
82769	Lyton Levison Mphande	M	25	127	62	Malawi	MAW	1988 Summer	1988	Summer	Seoul	Boxing	Boxing Men's Light-Welterweight	NA
15150	Rosario Briones	F	15	127	42	Mexico	MEX	1968 Summer	1968	Summer	Mexico City	Gymnastics	Gymnastics Women's Horse Vault	NA

Figure 38. The shortest player

3.9.4 The tallest player

The tallest is Yao Ming, 226cm, who participated in the Olympic basketball events in 2000, 2004 and 2008.

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
132627	Yao Ming	M	23	226	141	China	CHN	2004 Summer	2004	Summer	Athina	Basketball	Basketball Men's Basketball	NA
132627	Yao Ming	M	27	226	141	China	CHN	2008 Summer	2008	Summer	Beijing	Basketball	Basketball Men's Basketball	NA
132627	Yao Ming	M	20	226	141	China	CHN	2000 Summer	2000	Summer	Sydney	Basketball	Basketball Men's Basketball	NA

Figure 39. The tallest player

3.9.5 The lightest player

The lightest athlete is this female gymnast all-around athlete from North Korea, only 25kg, really light as a swallow, and participated in the 1980 Moscow Olympic Games.

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
21049	Choi Myong-Hui	F	14	135	25	North Korea	PRK	1980 Summer	1980	Summer	Moskva	Gymnastics	Gymnastics Women's Balance Beam	NA
21049	Choi Myong-Hui	F	14	135	25	North Korea	PRK	1980 Summer	1980	Summer	Moskva	Gymnastics	Gymnastics Women's Individual All-Around	NA
21049	Choi Myong-Hui	F	14	135	25	North Korea	PRK	1980 Summer	1980	Summer	Moskva	Gymnastics	Gymnastics Women's Uneven Bars	NA
21049	Choi Myong-Hui	F	14	135	25	North Korea	PRK	1980 Summer	1980	Summer	Moskva	Gymnastics	Gymnastics Women's Floor Exercise	NA
21049	Choi Myong-Hui	F	14	135	25	North Korea	PRK	1980 Summer	1980	Summer	Moskva	Gymnastics	Gymnastics Women's Horse Vault	NA
21049	Choi Myong-Hui	F	14	135	25	North Korea	PRK	1980 Summer	1980	Summer	Moskva	Gymnastics	Gymnastics Women's Team All-Around	NA

Figure 40. The lightest player

3.9.6 The heaviest player

The heaviest player is the male judo player from Guam, 214kg, who participated in the 2008 and 2012 Olympic Games.

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
12177	Ricardo Blas, Jr.	M	25	183	214	Guam	GUM	2012 Summer	2012	Summer	London	Judo	Judo Men's Heavyweight	NA
12177	Ricardo Blas, Jr.	M	21	183	214	Guam	GUM	2008 Summer	2008	Summer	Beijing	Judo	Judo Men's Heavyweight	NA

Figure 41. The heaviest player

3.9.7 Athlete with the most participation in the Olympic Games.

An equestrian named Ian Millar has participated in 10 Olympic Games. Since 1972, he has represented Canada in the Olympic Games. Until 2012, it was his 10th Summer Olympics. The second time I won the team silver medal in the equestrian event is really a very inspirational story.

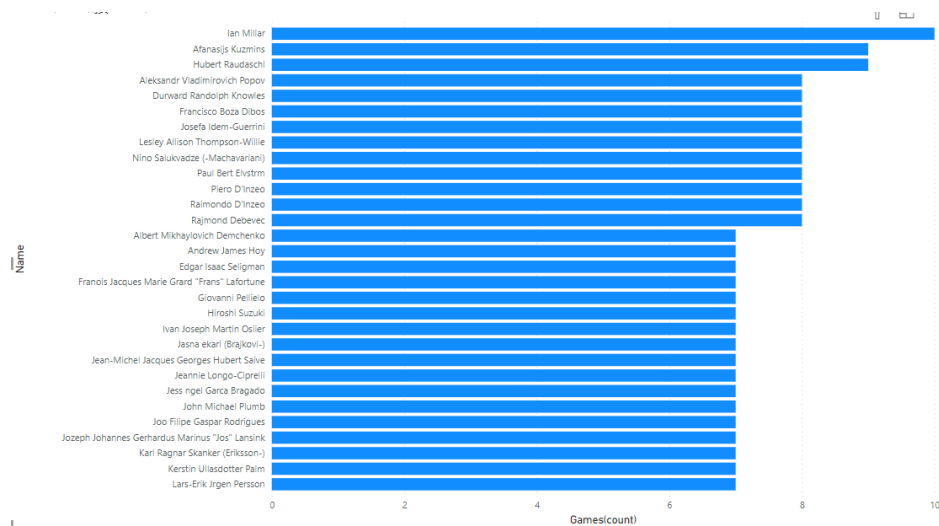


Figure 42. Athlete with the most participation in the Olympic Games

4. Model building

Olympic athletes have the best physical fitness. Their physical fitness is far superior to ordinary people. We can build machine learning models to analyze their height

and weight. Predict their gender based on their height and weight. We can build the following two classification models:

1).decision tree model

2).random forest model

First we need to preprocess the data. We need to deal with missing values. Prepare feature vector x and target y.

Code:

```
data = pd.read_csv('D:/Dataset/athlete_events.csv')
df = pd.DataFrame(data)
df_copy = df.copy()
df.duplicated()
df.drop_duplicates()
df_used = df.drop(['ID', 'Name', 'Age', 'Team', 'NOC',
'Games', 'Year', 'Season', 'City', 'Sport', 'Event',
'Medal'], axis=1)
df_used.dropna(axis=0, inplace=True)
df_used.Sex=df_used.Sex.astype(str).map({'M':0, 'F':1})
y=df_used.Sex
x=df_used.drop('Sex', axis=1)
```

```
df_used
```

	Sex	Height	Weight
0	0	180.0	80.0
1	0	170.0	60.0
4	1	185.0	82.0
5	1	185.0	82.0
6	1	185.0	82.0
...
271111	0	179.0	89.0
271112	0	176.0	59.0
271113	0	176.0	59.0
271114	0	185.0	96.0
271115	0	185.0	96.0

206853 rows × 3 columns

Figure 43. feature vector x and target y

4.1 Decision tree model

We have used a decision tree model before.

This time we will use Decision Tree Classifier.

Core code:

```
x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size=0.37)
dtc = DecisionTreeClassifier()
dtc.fit(x_train, y_train)
dt_predict = dtc.predict(x_test)
print("DecisionTreeClassifier report")
print(dtc.score(x_test, y_test))
print(sklearn.metrics.classification_report(y_test,
dt_predict))
```

```

DecisionTreeClassifier report
0.795377338768684

```

	precision	recall	f1-score	support
0	0.83	0.88	0.85	51750
1	0.71	0.63	0.67	24786
accuracy			0.80	76536
macro avg	0.77	0.75	0.76	76536
weighted avg	0.79	0.80	0.79	76536

Figure 44. Decision tree classifier report

From the above figure, we find that accuracy is 0.8, this result is very good, so we can use weight and height to predict sex.

4.2 Random forest model

Random forest (RF) or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

Core code:

```

x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size=0.37)
rfc = RandomForestClassifier()
rfc.fit(x_train, y_train)
rfc_y_predict = rfc.predict(x_test)
print("RandomForestClassifier report")
print(rfc.score(x_test, y_test))
print(sklearn.metrics.classification_report(y_test,
rfc_y_predict))

```



```

RandomForestClassifier report
0.7978990279084353
      precision    recall  f1-score   support

     0       0.83     0.88     0.86     51732
     1       0.72     0.62     0.67     24804

 accuracy                   0.80     76536
 macro avg                   0.77     76536
 weighted avg                 0.79     76536

```

Figure 45. Random forest classifier report

From the above figure, we find that accuracy is 0.8, this result is same as decision tree.

5. Financial management, resource efficiency and resource saving

The purpose of this section discusses the issues of competitiveness, resource efficiency and resource saving, as well as financial costs regarding the object of study of Master's thesis. Competitiveness analysis is carried out for this purpose. SWOT analysis helps to identify strengths, weaknesses, opportunities and threats associated with the project, and give an idea of working with them in each particular case. For the development of the project requires funds that go to the salaries of project participants and the necessary equipment, a complete list is given in the relevant section. The calculation of the resource efficiency indicator helps to make a final assessment of the technical decision on individual criteria and in general.

5.1 Competitiveness analysis of technical solutions

In order to find sources of financing for the project, it is necessary, first, to determine the commercial value of the work. Analysis of competitive technical solutions in terms of resource efficiency and resource saving allows to evaluate the comparative effectiveness of scientific development. This analysis is advisable to carry out using an evaluation card.

First of all, it is necessary to analyze possible technical solutions and choose the best one based on the considered technical and economic criteria.

Evaluation map analysis presented in Table 1. The position of your research and competitors is evaluated for each indicator by you on a five-point scale, where 1 is the weakest position and 5 is the strongest. The weights of indicators determined by you in the amount should be 1. Analysis of competitive technical solutions is determined by the formula:

$$C = \sum W_i \cdot P_i,$$

C - the competitiveness of research or a competitor;

W_i – criterion weight;

P_i – point of i-th criteria.

P₁₁ – Ability to handle missing values

P₁₂ – Accuracy of predictions on data

Since the first modern Olympic Games were held in 1896, the Olympic medal table has always been the focus of attention of countries all over the world, because it is not only a manifestation of sports performance, but also a symbol of a country's identity. Therefore, the analysis of the relevant factors affecting the Olympic medals naturally also has rich research significance. Therefore, more and more data analysts are analyzing this data set.

Compared to other data analysts, we have better handling of missing datasets. It's interesting to find out specific, rightful, and useful regression method to handle this problem. Therefore, more accurate predictions can be made on the data.

Table5.1. Evaluation card for comparison of competitive technical solutions

Evaluation criteria	Criterion weight	Points			Competitiveness Taking into account weight coefficients		
		P_f	P_{i1}	P_{i2}	C_f	C_{i1}	C_{i2}
1	2	3	4	5	6	7	8
Technical criteria for evaluating resource efficiency							
1. Reliability of dataset	0.14	5	3	5	0.7	0.42	0.7
2. Ease of operation	0.20	4	5	3	0.8	1.0	0.6
3. Ability to connect to PC	0.10	4	3	5	0.4	0.3	0.5
4. Smart interface quality	0.16	4	5	4	0.64	0.8	0.64
Economic criteria for performance evaluation							
1. Development cost	0.14	4	3	4	0.56	0.42	0.56
2. Development efficiency	0.08	3	4	4	0.24	0.32	0.32
3. Prediction accuracy	0.18	4	3	4	0.72	0.54	0.72
Total	1	28	26	29	4.06	3.8	4.04

Data analysis industry is widely used. At present, the commercial application of data analysis is mainly based on their own transaction data and customer data,

external data is supplemented by descriptive data analysis, predictive data modeling is supplemented, and business customers are the main business.

5.2 SWOT analysis

Complex analysis solution with the greatest competitiveness is carried out with the method of the SWOT analysis: Strengths, Weaknesses, Opportunities and Threats. The analysis has several stages. The first stage consists of describing the strengths and weaknesses of the project, identifying opportunities and threats to the project that have emerged or may appear in its external environment. The second stage consists of identifying the compatibility of the strengths and weaknesses of the project with the external environmental conditions. This compatibility or incompatibility should help to identify what strategic changes are needed.

Table 5.2 SWOT analysis

	<p>Strengths:</p> <p>S1. Some of the information with commercial value can be analyzed through data mining.</p> <p>S2. Explore the potential value of sports to athletes.</p> <p>S3. Big data analysis can improve business decisions and provide data support.</p>	<p>Weaknesses:</p> <p>W1. Noisy data in the dataset needs to be preprocessed.</p> <p>W2. Requires analysis using a large dataset to improve accuracy.</p>
<p>Opportunities:</p> <p>O1. Analyze the ratio of height and weight of athletes to predict the gender of athletes and help us better understand the body structure of athletes.</p> <p>O2. Athlete information visualization processing, can more intuitively understand</p>	<p><i>Strategy which based on strengths and opportunities: Analyze the height ratio of athletes through machine learning methods and visualize the information.</i></p>	<p><i>Strategy which based on weaknesses and opportunities: Use different data preprocessing methods whenever possible and analyze the results multiple times.</i></p>

the various indicators of the athlete.		
Threats: T1. The prediction accuracy is not ideal. T2. Data sets for individual sports are not large enough, leading to problems with prediction overfitting.	<i>Strategy which based on strengths and threats: The initial data needs to be feature-engineered before prediction.</i>	<i>Strategy which based on weaknesses and threats: Use cross-validation methods to prevent overfitting problems when performing data analysis.</i>

5.3 Project Initiation

The initiation process group consists of processes that are performed to define a new project or a new phase of an existing one. In the initiation processes, the initial purpose and content are determined and the initial financial resources are fixed. The internal and external stakeholders of the project who will interact and influence the overall result of the research project are determined.

Table 5.3 Stakeholders of the project

Project stakeholders	Stakeholder expectations
Data Analyst	Easy to use, high accuracy of scoring model.
User of smart interface	Understanding Model Prediction Systems

Table 5.4 Purpose and results of the project

Purpose of project:	The purpose of the project is to analyze previous Olympic data, draw corresponding conclusions, and predict result by analyzing their height and weight.
Expected results of the project:	Visually display various indicators of the athlete's body, and predict the gender of the athlete based on the athlete's height ratio.
Criteria for acceptance of the project result:	Predicting the gender of athletes by building a model with an accuracy rate of up to 80%.
Requirements for the	1. The project must be completed before May 31, 2022 of

project result:	the current year.
	2. The results obtained must meet the acceptance criteria of the project results.

It is necessary to solve the some questions: who will be part of the working group of this project, determine the role of each participant in this project, and prescribe the functions of the participants and their number of labor hours in the project.

Table 5. 5 Structure of the project

№	Participant	Role in the project	Functions	Labor time, hours (working days (from table 7) × 6 hours)
1	Supervisor	Head of project	Suggest project direction and review master's thesis.	492 hours
2	Student	Executor	1. Analyze the dataset to find a suitable model. 2. Writing master's dissertations.	450 hours

Project limitations are all factors that can be as a restriction on the degree of freedom of the project team members.

Table 5.6 Project limitations

Factors	Limitations / Assumptions
3.1. Project's budget	415000 RUB
3.1.1. Source of financing	TPU
3.2. Project timeline:	10/1/2022 to 30/05/2022

3.2.1. Date of approval of plan of project	20/03/2022
3.2.2. Completion date	30/05/2022

As part of planning a science project, you need to build a project timeline and a Gantt Chart.

Table 5.7 Project Schedule

Job title	Duration, working days	Start date	Date of completion	Participants
General Technical supervision	22 days	10/01/2022	08/02/2022	Supervisor
Research and analysis of literature	19 days	09/02/2022	10/03/2022	Supervisor/ Student
Clean data and build machine learning models	21 days	11/03/2022	09/04/2022	Supervisor/ Student
Building a graphical interface based on machine learning models	20 days	10/04/2022	09/05/2022	Supervisor/ Student
Preparing of dissertation	15 days	10/05/2022	30/05/2022	Student

A Gantt chart, or harmonogram, is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph shows the duration of each activity.

Table 5.8 A Gantt chart

№	Activities	Participants	T _c , days	Duration of the project																
				January			February			March			April			May				
				1	2	3	1	2	3	1	2	3	1	2	3	1	2	3		
1	General Technical supervision	Supervisor	22																	
2	Research and analysis of literature	Supervisor / Student	19																	
3	Clean data and build machine learning models	Supervisor / Student	21																	
4	Building a graphical interface based on machine learning models	Supervisor / Student	20																	
5	Preparing of dissertation	Student	15																	

5.4 Scientific and technical research budget

The amount of costs associated with the implementation of this work is the basis for the formation of the project budget. This budget will be presented as the lower limit of project costs when forming a contract with the customer.

To form the final cost value, all calculated costs for individual items related to the manager and the student are summed.

In the process of budgeting, the following grouping of costs by items is used:

- Material costs of scientific and technical research;
- costs of special equipment for scientific work (Depreciation of equipment used for design);
- basic salary;

- additional salary;
- labor tax;
- overhead.

Calculation of material costs

The calculation of material costs is carried out according to the formula:

$$C_m = (1 + k_T) \cdot \sum_{i=1}^m P_i \cdot N_{consi}$$

where m – the number of types of material resources consumed in the performance of scientific research;

N_{consi} – the amount of material resources of the i -th species planned to be used when performing scientific research (units, kg, m, m², etc.);

P_i – the acquisition price of a unit of the i -th type of material resources consumed (rub./units, rub./kg, rub./m, rub./m², etc.);

k_T – coefficient taking into account transportation costs.

Prices for material resources can be set according to data posted on relevant websites on the Internet by manufacturers (or supplier organizations).

Table 5.9 Material costs

Name	Unit	Amount	Price per unit, rub.	Material costs, rub.
Electricity of computer	kWh	130	5.8	754
Papers		120	1.0	120
Pen		3	150	450
Printing on A4 sheet		210	4	840
Total				2164

Costs of special equipment

This point includes the costs associated with the acquirement of special equipment (instruments, stands, devices and mechanisms) necessary to carry out work on a specific topic.

Table 5.10 a. Costs of special equipment (+software)

№	equipment identification	Quantity of equipment	Price per unit, rub.	Total cost of equipment, rub.
1.	Laser printer	1	12000	12000

OR

Calculation of the depreciation. Depreciation is not charged if an equipment cost is less than 40 thousand rubles, its cost is taken into account in full.

If you use available equipment, then you need to calculate depreciation:

$$A = \frac{C_{\text{перв}} * H_a}{100}$$

A - annual amount of depreciation;

$C_{\text{перв}}$ - initial cost of the equipment;

$H_a = \frac{100}{T_{\text{ср}}}$ - rate of depreciation;

$T_{\text{ср}}$ - life expectancy.

Table 5.10 b. Depreciation of special equipment (+software)

№	equipment identification	Quantity of equipment	Total cost of equipment, rub.	Life expectancy, year	Depreciation for the duration of the project, rub.
1.	Personal Computer	1	84000	6	14000

Basic salary

This point includes the basic salary of participants directly involved in the implementation of work on this research. The value of salary costs is determined based on the labor intensity of the work performed and the current salary system

The basic salary (S_b) is calculated according to the formula:

$$S_b = S_a \cdot T_w, \quad (3.3)$$

where S_b – basic salary per participant;

T_w – the duration of the work performed by the scientific and technical worker, working days;

S_d - the average daily salary of an participant, rub.

The average daily salary is calculated by the formula:

$$S_d = \frac{S_m \cdot M}{F_v}, \quad (3.4)$$

где S_m – monthly salary of an participant, rub .;

M – the number of months of work without leave during the year:
at holiday in 48 days, $M = 11.2$ months, 6 day per week;

F_v – valid annual fund of working time of scientific and technical personnel (251 days).

Table 5.11 The valid annual fund of working time

Working time indicators	
Calendar number of days	365
The number of non-working days	
- weekend	52
- holidays	14
Loss of working time	
- vacation	48
- isolation period	
- sick absence	
The valid annual fund of working time	251

Monthly salary is calculated by formula:

$$S_{month} = S_{base} \cdot (k_{premium} + k_{bonus}) \cdot k_{reg}, \quad (x)$$

where S_{base} – base salary, rubles;

$k_{premium}$ – premium rate;

k_{bonus} – bonus rate;

k_{reg} – regional rate.

Table 5.12 Calculation of the base salaries

Performers	S_{base} , rubles	$k_{premium}$	k_{bonus}	k_{reg}	S_{month} , rub.	W_d , rub.	T_p , work days	W_{base} , rub.
Supervisor	37700			1.3	49010	1633.7	82	133963.4
Student	19200				24960	832	75	62400

Additional salary

This point includes the amount of payments stipulated by the legislation on labor, for example, payment of regular and additional holidays; payment of time associated with state and public duties; payment for work experience, etc.

Additional salaries are calculated on the basis of 10-15% of the base salary of workers:

$$W_{add} = k_{extra} \cdot W_{base}, \quad (x)$$

where W_{add} – additional salary, rubles;

k_{extra} – additional salary coefficient (10%);

W_{base} – base salary, rubles.

Labor tax

Tax to extra-budgetary funds are compulsory according to the norms established by the legislation of the Russian Federation to the state social insurance (SIF), pension fund (PF) and medical insurance (FCMIF) from the costs of workers.

Payment to extra-budgetary funds is determined of the formula:

$$P_{social} = k_b \cdot (W_{base} + W_{add}) \quad (x)$$

where k_b – coefficient of deductions for labor tax.

In accordance with the Federal law of July 24, 2009 No. 212-FL, the amount of insurance contributions is set at 30%. Institutions conducting educational and scientific activities have rate - 27.1%.

Table 5.13 Labor tax

	Project leader	Engineer
Coefficient of deductions	27.1%	
Salary (basic and additional), rubles	133963.4	62400
Labor tax, rubles	36304.08	16910.4

Overhead costs

Overhead costs include other management and maintenance costs that can be allocated directly to the project. In addition, this includes expenses for the maintenance, operation and repair of equipment, production tools and equipment, buildings, structures, etc.

Overhead costs account from 30% to 90% of the amount of base and additional salary of employees.

Overhead is calculated according to the formula:

$$C_{ov} = k_{ov} \cdot (W_{base} + W_{add})$$

where k_{ov} – overhead rate.

Table 5.14 Overhead

	Project leader	Engineer
Overhead rate	30%	
Salary, rubles	133963.4	62400
Overhead, rubles	40189.02	18720

Other direct costs

Energy costs for equipment are calculated by the formula:

$$C = P_{el} \cdot P \cdot F_{eq},$$

where P_{el} – power rates (5.8 rubles per 1 kWh);

P – power of equipment, kW;

F_{eq} – equipment usage time, hours.

Formation of budget costs

The calculated cost of research is the basis for budgeting project costs.

Determining the budget for the scientific research is given in the table 14.

Table 5.15 Items expenses grouping

Name	Cost, rubles
1. Material costs	2164
2. Equipment costs	26000
3. Basic salary	196363.4
4. Additional salary	0
5. Labor tax	53214.48
6. Overhead	58909.02
7. Other direct costs	0
Total planned costs	336650.9

5.5 Evaluation of the comparative effectiveness of the project

Determination of efficiency is based on the calculation of the integral indicator of the effectiveness of scientific research. Its finding is associated with the definition of two weighted average values: financial efficiency and resource efficiency.

The integral indicator of the financial efficiency of a scientific study is obtained in the course of estimating the budget for the costs of three (or more) variants of the execution of a scientific study. For this, the largest integral indicator of the implementation of the technical problem is taken as the calculation base (as

the denominator), with which the financial values for all the options are correlated.

The integral financial measure of development is defined as:

$$I_f^d = \frac{C_i}{C_{max}} \quad (x)$$

where I_f^d – integral financial measure of development;

C_i – the cost of the i-th version;

C_{max} – the maximum cost of execution of a research project (including analogues).

The obtained value of the integral financial measure of development reflects the corresponding numerical increase in the budget of development costs in times (the value is greater than one), or the corresponding numerical reduction in the cost of development in times (the value is less than one, but greater than zero).

Since the development has one performance, then $I_f^d = 1$.

The integral indicator of the resource efficiency of the variants of the research object can be determined as follows:

$$I_m^a = \sum_{i=1}^n a_i b_i^a \quad I_m^p = \sum_{i=1}^n a_i b_i^p$$

where I_m – integral indicator of resource efficiency for the i-th version of the development;

a_i – the weighting factor of the i-th version of the development;

b_i^a, b_i^p – score rating of the i-th version of the development, is established by an expert on the selected rating scale;

n – number of comparison parameters.

The calculation of the integral indicator of resource efficiency is presented in the form of table 5.15.

Table 5.16 – Evaluation of the performance of the project

Criteria	Weight criterion	Points
1. Reliability of dataset	0.14	13
2. Ease of operation	0.20	12
3. Ability to connect to PC	0.10	12
4. Smart interface quality	0.16	13
Economic criteria for performance evaluation		
1. Development cost	0.14	11
2. Development efficiency	0.08	11
3. Prediction accuracy	0.18	11
Total	1	83

The integral indicator of the development efficiency (I_e^p) is determined on the basis of the integral indicator of resource efficiency and the integral financial indicator using the formula:

$$I_e^p = \frac{I_m^p}{I_f^p}, \quad I_e^a = \frac{I_m^a}{I_f^a} \quad (=)$$

$$I_{\text{исп.2}} = \frac{I_{\text{р-исп2}}}{I_{\text{финр}}} \text{ и т.д.}$$

Comparison of the integral indicator of the current project efficiency and analogues will determine the comparative efficiency. Comparative effectiveness of the project:

$$E_c = \frac{I_e^p}{I_e^a} \quad (=)$$

Thus, the effectiveness of the development is presented in table 16.

Table 5.17 – Efficiency of development

№	Indicators	Points
1	Integral financial measure of development	12
2	Integral indicator of resource efficiency of development	15
3	Integral indicator of the development efficiency	13

Comparison of the values of integral performance indicators allows us to understand and choose a more effective solution to the technical problem from the standpoint of financial and resource efficiency.

5.6 Conclusion of financial management

Thus, in this section was developed stages for design and create competitive development that meet the requirements in the field of resource efficiency and resource saving.

These stages includes:

- development of a common economic project idea, formation of a project concept;
- organization of work on a research project;
- identification of possible research alternatives;
- research planning;
- assessing the commercial potential and prospects of scientific research from the standpoint of resource efficiency and resource saving;
- determination of resource (resource saving), financial, budget, social and economic efficiency of the project.

6. Social responsibility

6.1 Introduction

The Olympic Games originated in ancient Greece more than two thousand years ago and got its name because it was held in Olympia . We're going to analyze the Olympic data . In order to better understand the secrets behind the data .The development of the program is only carried out with the help of computer.

In this section, harmful and dangerous factors affecting the work of personnel will be considered, the impact of the developed program on the environment, legal and organizational issues, measures in emergency situations will be considered.

The work was carried out in the hall of residence of TPU (2th floor). Room 204 was a research execution place. The layout of the apartment is shown in Figure:

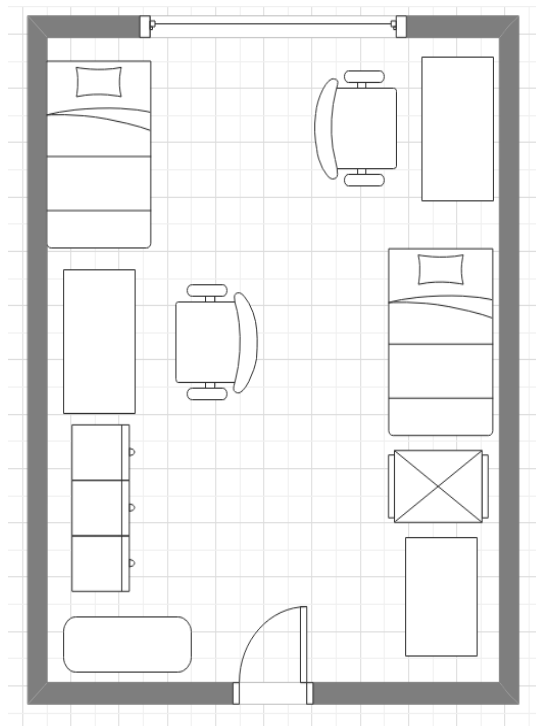


Figure 6.1. Apartment layout 204

6.2 Legal and organizational issues of occupational safety

Today, one of the main ways to fundamentally improve all prevention efforts is the widespread implementation of an integrated occupational safety and health management system to reduce overall accident rates and occupational morbidity.

This means combining isolated activities into a targeted system of action at all levels and stages of the production process.

Occupational safety is a system of legislative, socio-economic, organizational, technological, hygienic and therapeutic and prophylactic measures and tools that ensure the safety, preservation of health and human performance in the work process.

- According to the Labor Code of the Russian Federation, every employee has the right:
 - To have a workplace that meets Occupational safety requirements;
 - To have a compulsory social insurance against accidents at manufacturing and occupational diseases;
 - To receive reliable information from the employer, relevant government bodies and public organizations on conditions and Occupational safety at the workplace, about the existing risk of damage to health, as well as measures to protect against harmful and (or) hazardous factors;
 - To refuse carrying out work in case of danger to his life and health due to violation of Occupational safety requirements;
 - Be provided with personal and collective protective equipment in compliance with Occupational safety requirements at the expense of the employer;
 - For training in safe work methods and techniques at the expense of the employer;
 - For personal participation or participation through their representatives in consideration of issues related to ensuring safe working conditions in his workplace, and in the investigation of the accident with him at work or occupational disease;
 - For extraordinary medical examination in accordance with medical recommendations with preservation of his place of work (position) and secondary earnings during the passage of the specified medical examination;
 - For warranties and compensation established in accordance with this Code, collective agreement, agreement, local regulatory act, an employment

contract, if he is engaged in work with harmful and (or) hazardous working conditions.

The labor code of the Russian Federation states that normal working hours may not exceed 40 hours per week, The employer must keep track of the time worked by each employee.

Rules for labor protection and safety measures are introduced in order to prevent accidents, ensure safe working conditions for workers and are mandatory for workers, managers, engineers and technicians.

6.3 Basic ergonomic requirements for the correct location and arrangement of researcher's workplace

The workplace when working with a PC should be at least 6 square meters. The legroom should correspond to the following parameters: the legroom height is at least 600 mm, the seat distance to the lower edge of the working surface is at least 150 mm, and the seat height is 420 mm. It is worth noting that the height of the table should depend on the growth of the operator.

The following requirements are also provided for the organization of the workplace of the PC user: The design of the working chair should ensure the maintenance of a rational working posture while working on the PC and allow the posture to be changed in order to reduce the static tension of the neck and shoulder muscles and back to prevent the development of fatigue.

The type of working chair should be selected taking into account the growth of the user, the nature and duration of work with the PC. The working chair should be lifting and swivel, adjustable in height and angle of inclination of the seat and back, as well as the distance of the back from the front edge of the seat, while the adjustment of each parameter should be independent, easy to carry out and have a secure fit.-GOST 12.2.032-78 SSBT

6.4 Occupational safety

Workplace safety is the responsibility of everyone in the organization.

Occupational hygiene is a system of ensuring the health of workers in the process of labor activity, including legal, socio-economic, organizational and technical, sanitary and hygienic, treatment and prophylactic, rehabilitation and other measures.

Working conditions - a set of factors of the working environment and the labor process that affect human health and performance.

Harmful production factor is a factor of the environment and the work process that can cause occupational pathology, temporary or permanent decrease in working capacity, increase the frequency of somatic and infectious diseases, and lead to impaired health of the offspring.

Hazardous production factor is a factor of the environment and the labor process that can cause injury, acute illness or sudden sharp deterioration in health, death.

In this subsection it is necessary to analyze harmful and hazardous factors that can occur during research in the laboratory, when development or operation of the designed solution (on a workplace).

GOST 12.0.003-2015 "Hazardous and harmful production factors. Classification" must be used to identify potential factors, that can effect on a worker(employee).

Table 6.1 - Potential hazardous and harmful production factors

Factors (GOST 12.0.003-2015)	Stages of work			Legislation documents
	developing	manufacturing	operation	
1. Increased noise level	+	+		GOST 12.1.003-2014 Occupational safety standards system. Noise. General safety requirements

2. Lack or lack of natural light, insufficient illumination	+			SanPiN 2.2.1/2.1.1.1278-03 Hygienic requirements for natural, artificial and mixed lighting of residential and public buildings
3. Physical overload (static-long-term preservation of a certain posture).		+	+	GOST 12.2.032-78 SSBT. Workplace when performing work while sitting General ergonomic requirements.
4. Increased voltage in an electrical circuit, the closure of which can pass through the human body		+	+	Sanitary rules GOST 12.1.038-82 SSBT. Electrical safety. Maximum permissible levels of touch voltages and currents.

Increased noise level

Noise worsens working conditions; have a harmful effect on the human body, namely, the organs of hearing and the whole body through the central nervous system. It results in weakened attention, deteriorated memory, decreased response, and increased number of errors in work.

Noise can be generated by operating equipment, air conditioning units, daylight illuminating devices, as well as spread from the outside.

When working on a PC, the noise level in the workplace should not exceed 50 dB . In order to study in a quiet environment, irrelevant applications of the computer should be closed to reduce computer power consumption, thereby reducing computer noise, and windows should also be closed to reduce environmental noise.

Lack or lack of natural light, insufficient illumination

Light sources can be both natural and artificial. The natural source of the light in the room is the sun, artificial light are lamps. With long work in low illumination conditions and in violation of other parameters of the illumination, visual perception decreases, myopia, eye disease develops, and headaches appear.

According to the SanPiN 2.2.1/2.1.1.1278-03 standard., the illumination on the table surface in the area of the working document should be 300-500 lux. Lighting should not create glare on the surface of the monitor. Illumination of the monitor surface should not be more than 300 lux.

The brightness of the lamps of common light in the area with radiation angles from 50 to 90 ° should be no more than 200 cd/m, the protective angle of the lamps should be at least 40 °. The ripple coefficient should not exceed 5%.

Physical overload

As a computer worker, you often need to sit in front of the computer for a long time to work. This static working posture will cause physical overload for a long time. If you sit for a long time, your body will protest, and many people will have occupational diseases. If you don't pay attention to timely improvement, it will have a great impact on your health.

Our eyes are damaged due to static - prolonged computer viewing posture, sore eyes and deepened vision: When many people work in front of the computer for a long time, if they do not pay attention to rest in time, their eyes will feel tired. There is often a sour feeling, and the degree of myopia has deepened.

Wrist pain and decreased finger flexibility: Some friends who are engaged in writing work need to keep typing on the keyboard at the computer when they go to work. If the hand performs a single activity for a long time, the wrist joints, fingers and other parts may appear In the case of faint pain, the flexibility of the fingers will also decrease.

Lumbar and cervical vertebrae are often painful: when many people sit for a long time, they can maintain a straight posture at first, and the whole person can sit

very upright, but with the development of time, many people will begin to hunched over. The neck will also stretch forward, and the lumbar spine, cervical spine and other parts of the human body will experience faint pain, and some people may even have problems such as lumbar disc herniation.

According to the GOST 12.2.032-78 SSBT standard, The design of the workplace and the relative position of all its elements (seats, controls, ways of displaying information, etc.) must conform to anthropometric, physiological and psychological requirements and the nature of the work.

It is recommended to prepare a comfortable small pillow at ordinary times and place the pillow on the chair, which will help relieve the pressure on the lumbar spine of the human body, and turn the neck more in leisure time, which can move the muscles and bones well.

Therefore, after hitting the keyboard for about an hour, you should stop and move your wrist joints. You can stretch your fingers together, which can also relax our arms.

For office workers, after sitting for a long time, the body is prone to some occupational diseases. Everyone should pay attention to avoid it in peacetime, and pay attention to replenishing the water in the body every day. The daily water intake is 2000 ml, which can be very good. Promote water circulation in the body, and get up and walk in moderation after sitting for a long time, which is more conducive to the health of the body.

Abnormally high voltage value in the circuit , the closure which may occur through the human body

The mechanical action of current on the body is the cause of electrical injuries. Typical types of electric injuries are burns, electric signs, skin metallization, tissue tears, dislocations of joints and bone fractures.

The following protective equipment can be used as measures to ensure the safety of working with electrical equipment:

- disconnection of voltage from live parts, on which or near to which work will be carried out, and taking measures to ensure the impossibility of applying voltage to the workplace;
- posting of posters indicating the place of work;
- electrical grounding of the housings of all installations through a neutral wire;
- coating of metal surfaces of tools with reliable insulation;
- inaccessibility of current-carrying parts of equipment (the conclusion in the case of electroporation elements, the conclusion in the body of current carrying parts).

6.5 Ecological safety

Presently section discusses the environmental impacts of the project development activities, as well as the product itself as a result of its implementation in production. The software product itself, developed during the implementation of the master's thesis, does not harm the environment either at the stages of its development or at the stages of operation. However, the funds required to develop and operate it can harm the environment.

There is no production in the laboratory. The waste produced in the premises, first of all, can be attributed to waste paper, plastic waste, defective parts of personal computers and other types of computers. Waste paper is recommended accumulate and transfer them to waste paper collection points for further processing. Place plastic bottles in specially designed containers.

Modern PCs are produced practically without the use of harmful substances hazardous to humans and the environment. Exceptions are batteries for computers and mobile devices. Batteries contain heavy metals, acids and alkalis that can harm the environment by entering the hydrosphere and lithosphere if not properly disposed of. For battery disposal it is necessary to contact special organizations specialized in the reception, disposal and recycling of batteries .

Fluorescent lamps used for artificial illumination of workplaces also require special disposal, because they contain from 10 to 70 mg of mercury, which is an extremely dangerous chemical substance and can cause poisoning of living beings, and pollution of the atmosphere, hydrosphere and lithosphere. The service life of such lamps is about 5 years, after which they must be handed over for recycling at special reception points. Legal entities are required to hand over lamps for recycling and maintain a passport for this type of waste. An additional method to reduce waste is to increase the share of electronic document management.

6.6 Safety in emergency

In the working environment of the PC operator, the following manufactured emergencies may occur:

- Fires and explosions in buildings and communications;
- Collapse of buildings.

Possible natural disasters include meteorological (hurricanes, showers, frosts), hydrological (floods, floods, flooding), and natural fires.

Emergencies of a biological and social nature include epidemics, epizootics, and epiphytotic. Environmental emergencies can be caused by changes in the state, lithosphere, hydrosphere, atmosphere and biosphere as a result of human activities.

The most typical for the object where the working rooms are located, equipped with a personal computer, the emergency is a fire. Premises for work of PC operators according to the classification system of categories premises for explosion and fire hazard belongs to category D (out of 5 categories A, B, B1-B4, D, D), because applies to premises with noncombustible substances and materials in a cold state.

All employees of the organization must be familiar with the fire safety instructions, undergo safety instructions and strictly observe it. It is forbidden to use electrical appliances in conditions that do not meet the requirements of the manufacturer's instructions, or have various kinds of malfunctions that, in

accordance with the instructions for use, may lead to a fire, as well as use electrical wires and cables with damaged or lost protective properties of insulation.

Before leaving the office, it is required to inspect it, close the windows, and make sure that there are no sources of possible ignition in the room, all electrical appliances are turned off and the lighting is turned off.

With a frequency of at least once every three years, it is necessary to measure the insulation resistance of current-carrying parts of power and lighting equipment. The increase in sustainability is achieved through the implementation of appropriate organizational and technical measures, training of personnel to work in emergencies.

Upon detecting a fire or signs of combustion (smoke, burning smell, temperature increase, etc.), an employee must:

- It is required to stop work, call the fire department by phone "01";
 - If possible, take measures to evacuate people and material values;
 - Disconnect electrical equipment from the mains;
 - Start extinguishing the fire with the available fire extinguishing means;
 - Inform the immediate or superior supervisor and notify the surrounding employees;
- In case of a general signal of danger, leave the building in accordance with the "Plan for the evacuation of people in case of fire and other emergencies."

To extinguish a fire, use manual carbon dioxide fire extinguishers (type OU-2, OU-5) located in the office premises, and a fire hydrant internal fire-fighting water supply. They are designed to extinguish the initial fires of various substances and materials, with the exception of substances that burn without air access. Fire extinguishers must be kept in good working order at all times and ready for action. It is strictly forbidden to extinguish fires in office premises using chemical foam fire extinguishers (type OHP-10).

6.7 Conclusion of social responsibility

Each employee must carry out professional activities with taking into account social, legal, environmental and cultural aspects, issues health and safety, be socially responsible for the solutions, be aware of the need for sustainable development.

In presently section covered the main issues of observance of rights employee to work, compliance with the rules for labor safety, industrial safety, ecology and resource conservation.

It was found that the researcher's workplace satisfies safety and health requirements during project implementation, and the harmful impact of the research object on the environment is not exceeds the norm.

6.8 Reference of social responsibility

1. GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements.
2. SanPiN 2.2.2 / 2.4.1340-03. Sanitary-epidemiological rules and standards "Hygienic requirements for PC and work organization".
3. GOST 12.1.003-2014 SSBT. Noise. General safety requirements.
4. SanPiN 2.2.1 / 2.1.1.1278-03. Hygienic requirements for natural, artificial and combined lighting of residential and public buildings.
5. SanPiN 2.2.2 / 2.4.1340-03 "Hygienic requirements for personal computers and work organization "
6. GOST 12.1.038-82 Occupational safety standards system. Electrical safety
7. Federal Law "On the Fundamentals of Labor Protection in the Russian Federation" of 17.07.99 № 181 – FZ
8. GOST R ISO 1410-2010. Environmental management. Assessment of life Cycle. Principles and structure.
9. GOST R12.1.004-85 Occupational safety standards system. Fire safety
10. GOST 12.2.003-91 Occupational safety standards system. Industrial equipment. General safety requirements
11. GOST Industrial equipment. General safety requirements to working places
12. GOST 12.2.003-91 Occupational safety standards system. Industrial equipment. General safety requirements

Conclusion

1. The number of athletes in the Olympic parameters increased from 176 in the first session to 11,669 in the 32nd session, and the Olympic Games continued to cover more people.
2. Female athletes have made a great breakthrough from the initial proportion of less than 2% to the current 45%.
3. The types of competitions have also increased from 9 to 36 today, with more and more types.
4. The event with the most participation in history is Athletics, followed by swimming, rowing, soccer.
5. Australia, France, Greece, Italy, Sweden participated in all 29 Summer Olympics.
6. The U.S. sent the most to the Olympics, followed by Germany.
7. A total of 42 cities have hosted the Olympic Games in history, of which Athens and London have hosted 3 times.
8. The country with the most medals is the United States, followed by Russia, Germany, the United Kingdom.
9. By using machine learning models, we can use weight and height to predict gender.

Reference

- 1.Olympic Games – URL: https://en.wikipedia.org/wiki/Olympic_Games
- 2.Olympic history data: thorough analysis – URL: <https://www.kaggle.com/heesoo37/olympic-history-data-a-thorough-analysis>
- 3.Big data visualization analysis of the Olympic Games – URL: <https://zhuanlan.zhihu.com/p/397957667>
- 4.Linear Regression for Machine Learning – URL: <https://machinelearningmastery.com/linear-regression-for-machine-learning/>
- 5.Decision tree – URL: https://en.wikipedia.org/wiki/Decision_tree
- 6.Random forest – URL: https://en.wikipedia.org/wiki/Random_forest
- 7.Root-mean-square deviation – URL: https://en.wikipedia.org/wiki/Root-mean-square_deviation
- 8.Mean absolute error – URL: https://en.wikipedia.org/wiki/Mean_absolute_error
- 9.GridSearchCV – URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- 10.Huang Shan, Gubin E. Data cleaning for data analysis. Томск, 2018г. - С. 387-389.