3. Nilsson T, Wahlström J, Burström L Hand-arm vibration and the risk of vascular and neurological diseases. A systematic review and meta-analysis. [Электронный ресурс] PLoS ONE 2017; 12(7): e0180795. URL: https://doi.org/10.1371/journal.pone. 0180795 (дата доступа 30.08.2019)
4. Денисов Э.И., Кравченко О.К. Локальная вибрация и риск вибрационной болезни. – В кн. Профессиональный риск для здоровья работников (Руководство) / Под ред. Н.Ф.Измерова и Э.И.Денисова. – М.:Тровант, 2003. – С.115-124.
5. ISO 5349:1986 Mechanical vibration – Guidelines for the measurement and the assessment of human exposure to hand-transmitted vibration [Электронный ресурс]. URL: https://www.iso.org/standard/11369.html (дата доступа 30.09.2019)
6. Thompson RP. Causality, mathematical models and statistical association: dismantling evidence-based medicine//J Eval Clin Pract.- 2010.- №16(2).- P. 267-75.
7. Денисов Э.И., Чесалин П.В., Профессионально обусловленная заболеваемость, основы методологии//Медицина труда и пром.экология. -2006.- №8.- С. 5-9.
8. Максимов С.А., Цыганкова Д.П., Артамонова Г.В. Применение регрессионного анализа и деревьев классификации для расчета дополнительного популяционного риска ишемической болезни сердца //Анализ риска здоровью. – 2017. – № 3. – С. 31–39.
9. Тырсин А. Н., Калев О. Ф., Яшин Д. А., Лебедева О. В. Оценка состояния здоровья популяции на основе энтропийного моделирования//Матем. биология и биоинформ.- 2015.- том 10.- выпуск 1.-С. 206–219.
10. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, Kwak TY. Artificial Intelligence in Pathology// J Pathol Transl Med. -2019.- №53(1).-P.1-12.
11. Волчек Ю.А., Шишко О.Н., Спиридонова О.С., Мохорт Т.В. Положение модели искусственной нейронной сети в медицинских экспертных системах //Juvenis scientia. -2017.- №9. –С.4-9.
12. Wang L.X., Mendel J.M. Generating Fuzzy Rules by Learning from Examples//IEEE Transactions on Systems, Man, and Cybernetics.- November/December 1992.- vol.22.- №6.- P 1414-1427.

ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ ДЛЯ УГЛУБЛЕННОГО АНАЛИЗА НЕФТЕГАЗОВОГО МЕСТОРОЖДЕНИЯ

Н.И. Журбич
(г. Томск, Томский политехнический университет)
e-mail: niz1@tpu.ru

PREPARATION OF INITIAL DATA FOR IN-DEPTH ANALYSIS OF THE OIL AND GAS FIELD

N.I. Zhurbich
(Tomsk, Tomsk Polytechnic University)

Abstract: The article is devoted to the analysis of a file containing information about the oil and gas field, the main parameters and indicators of the oil and gas field under consideration.
Keywords: python, k-nearest algorithm imputation, missings, outliers, data analysis.

Введение. Нефтегазовые компании в процессе своей деятельности получают петабайты данных каждый день, которые необходимо обрабатывать и анализировать для повышения эффективности их работы. Для этой цели лучше использовать современные технологии обработки больших данных, которые предоставляют различные инструменты для анализа и

предсказания будущих трендов в области геологии, инженерии и нефтегазового производства.

По мнению экспертов консалтинговой компании Molten, многие нефтегазовые предприятия «безответственно» распоряжаются своими данными. По их подсчетам, крупные нефтегазовые компании тратят от $ 1 до $ 3 млрд в год на сбор данных, однако расходы на поддержание и обработку накопленной информации зачастую составляют менее 1 % от этой суммы. В то же время высокая конкуренция на рынке обуславливает принятие обоснованных решений для поддержки высокого уровня производительности нефтегазовых компаний [1].

Задание. Провести анализ файла, содержащего информацию о нефтегазовом месторождении, его основных параметрах и показателях. Файл содержит 3368 записей и более 120 атрибутов.

В ходе выполнения задания, необходимо было решить следующие задачи:

- Устранение ошибок и восстановление пропущенных значений данных в файле с помощью алгоритма поиска ближайшего соседа (*k*-nearest algorithm imputation)
- Построить графические представления нескольких атрибутов для демонстрации работы алгоритма.

Для решения первой задачи необходимо было выбрать технологию для восстановления пропущенных значений (missing values), а также применить алгоритм для восстановления этих значений. Для решения второй задачи было решено использовать библиотеку для визуализации рассматриваемых данных MatPlotLib.

Выбор инструментов разработки. Подготовка и чистка исследуемого датасета для дальнейшей визуализации производилась на языках программирования Python и R.

Также были использованы следующие инструменты и библиотеки:

1. NumPy – это библиотека языка Python, добавляющая поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых (и очень быстрых) математических функций для операций с этими массивами [2].

2. Pandas – программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня. [3].

3. Seaborn – это библиотека визуализации данных Python, основанная на Matplotlib. Она предоставляет высокоуровневый интерфейс для рисования информативной статистической графики.

4. R Studio – среда разработки программного обеспечения с открытым исходным кодом для языка программирования R, который предназначен для статистической обработки данных и работы с графикой.

Подготовка и чистка данных. Прежде всего, нужно подготовить данные для анализа. Для этого нужно понять, есть ли в наборе данных отсутствующие или нулевые значения. Наша цель – улучшить значения в наборе данных, если это возможно. Данный этап анализа является очень важным, потому что если датасет содержит некорректные данные на входе, будут получены некорректные результаты. Пример такой проверки представлен на рисунке 1.

Скважина	False
Дата	False
ГТМ	True
Метод	True
Характер работы	True
Состояние	True
Время работы, ч	True
Время накопления	True
Нефть, т	True
Попутный газ, м3	True
Закачка, м3	True
Природный газ, м3	True
Газ из газовой шапки, м3	True
Конденсат, т	True
Простой, ч	True
Причина простоя	True
Приемистость, м3/сут	True
Обводненность (вес), %	True
Агент закачки	True
Нефть, м3	True
Жидкость, м3	True
Дебит конденсата	True
Добыча растворенного газа, м3	True
Дебит попутного газа, м3/сут	True
Пласт МЭР	True
Куст	True

Рисунок 1. Проверка на наличие пропущенных значений

Для восстановления пропущенных значений использовался пакет R Studio с языком программирования R. На первом этапе было необходимо считать файл и заменить пустые или пропущенные значения на пустые (NA). На рисунке 2 продемонстрирован данный процесс.

	Скважина	Дата	ГТМ	Метод	Характер.работы	Состояние	Время.работы..ч
1	002ff5b8a6dc271f58581e1b4fa2c5fc	01.12.2016	1	ФОН	НЕФ	ОСВ ТГ	0
2	008d0347e572a5d938a9c40c29e539fc	01.10.2013	NA	<NA>	<NA>	<NA>	NA
3	00b40cb7bb8c9fd1ac26b4cc86f2b291	01.02.2018	NA	<NA>	<NA>	<NA>	NA
4	01ba18d8b6d29875a18d4bca4eb201d7	01.05.2014	0	ЭЦН/ФОН	НЕФ	РАБ.	120
5	024ec6f6e3f9c5150ecf525bf8b7a6a3	01.06.2017	1	ФОН	НЕФ	ОСВ ТГ	0
6	0254a227c6c2c31a419126700cfcddc2	01.05.2017	1	ЭЦН/ФОН	НЕФ	ОСТ.	193

Рисунок 2. Чтение файла в среде разработки R Studio

После этого этапа можно увидеть количество пустых значений (NA). На рисунках 3 и 4 приведено подробное описание определённых атрибутов.

Variables sorted by number of missings:

Variable	Count
Причина.простоя	215
ГТМ	45
Метод	45
Характер.работы	45
Состояние	45
Время.работы..ч	45
Попутный.газ..м3	45
Простой..ч	45
Обводненность..вес....	45
Добыча.растворенного.газа..м3	45
Дебит.попутного.газа..м3.сут	45
Скважина	0
Дата	0

Рисунок 3. Количество пропущенных значений в каждом атрибуте

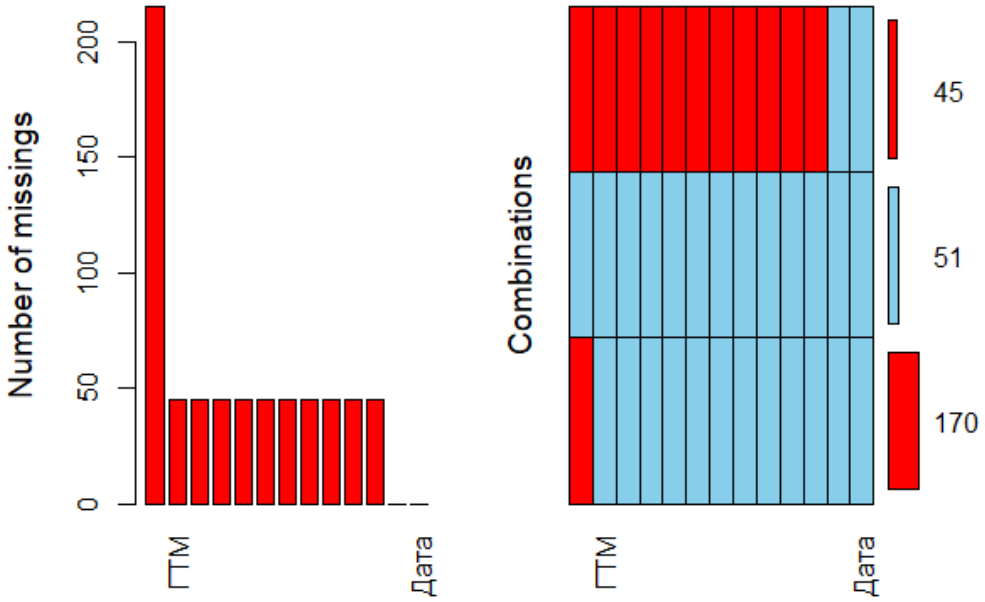

Рисунок 4. Количество пропущенных значений в каждом атрибуте

Для восстановления пропущенных значений в этом наборе данных был выбран алгоритм поиска ближайшего соседа (*k*-nearest algorithm imputation). KNN – это алгоритм, который применяется для сопоставления точки с её ближайшими *k*-соседями в многомерном пространстве. Он может использоваться для данных, которые являются непрерывными, дискретными, порядковыми и категориальными, что делает его особенно полезным для работы со всеми видами недостающих данных.

Аргументом в пользу применения алгоритма KNN для пропущенных значений является то, что значение точки может быть аппроксимировано значениями ближайших к нему точек на основе других переменных [5].

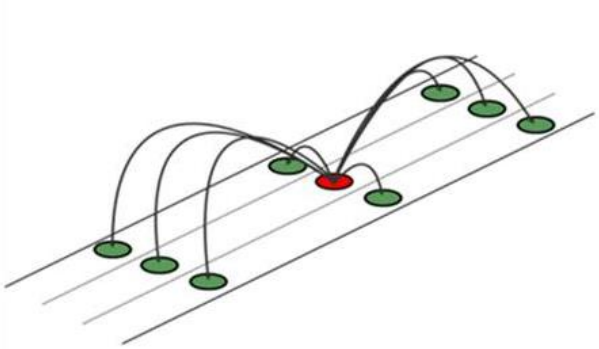

Рисунок 5. Принцип работы алгоритма поиска ближайшего соседа

После применения данного алгоритма для вставки пропущенных значений, можно проверить переменные. Проверка корректности работы алгоритма поиска ближайшего соседа (*k*-nearest algorithm imputation) представлена на рисунке 6.

```
Variables sorted by number of missings:
                          Variable Count
                          Скважина     0
                              Дата     0
                               ГТМ     0
                            Метод      0
                  Характер.работы      0
                         Состояние     0
                  Время.работы..ч      0
                 Попутный.газ..м3      0
                       Простой..ч      0
                 Причина.простоя       0
          Обводненность..вес....      0
  Добыча.растворенного.газа..м3      0
   Дебит.попутного.газа..м3.сут      0
                  Скважина_imp       0
                      Дата_imp       0
                       ГТМ_imp       0
                     Метод_imp       0
           Характер.работы_imp       0
                  Состояние_imp      0
           Время.работы..ч_imp       0
          Попутный.газ..м3_imp       0
                Простой..ч_imp       0
          Причина.простоя_imp        0
    Обводненность..вес...._imp       0
Добыча.растворенного.газа..м3_imp    0
 Дебит.попутного.газа..м3.сут_imp    0
```

Рисунок 6. Результат работы алгоритма поиска ближайшего соседа

Исходя из результатов проверки можно сделать вывод о том, что в ходе работы данного алгоритма все значения были успешно восстановлены.

Визуализация. После подготовки данных исследуемого датасета можно приступить к визуализации данных. Например, на рисунке 7 можно увидеть самые распространенные методы, используемые на буровых скважинах. Первое место по использованию занимает электроприводной центробежный насос (ЭЦН) – это наиболее широко распространённый в России аппарат механизированной добычи нефти [6]. Второе позицию занимает комбинация ЭЦН и ФОН (фонтанный способ добычи нефти). Третье место – ФОН (фонтанный способ добычи нефти).

```
import seaborn as sns
sns.countplot(neft1['Метод'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xe5cfc90>
```

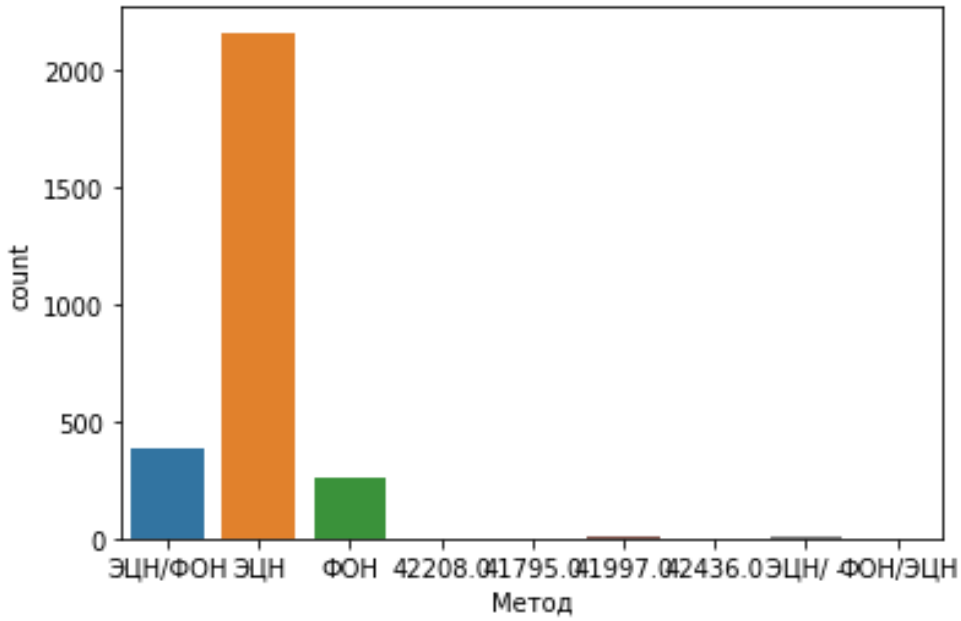

Рисунок 7. Методы, используемые на буровых скважинах

На рисунке 8 показано наличие проведения геолого-технических мероприятий (ГТМ) – это работы, проводимые на скважинах с целью регулирования разработки месторождений и поддержания целевых уровней добычи нефти.

В данном случае используется словарь данных:

- Значение «1» – геолого-технические мероприятия проводились на данной скважине.
- Значение «0» – геолого-технические мероприятия не проводились на данной скважине.

```
sns.countplot(neft1['ГТМ'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xe6e8ab0>
```

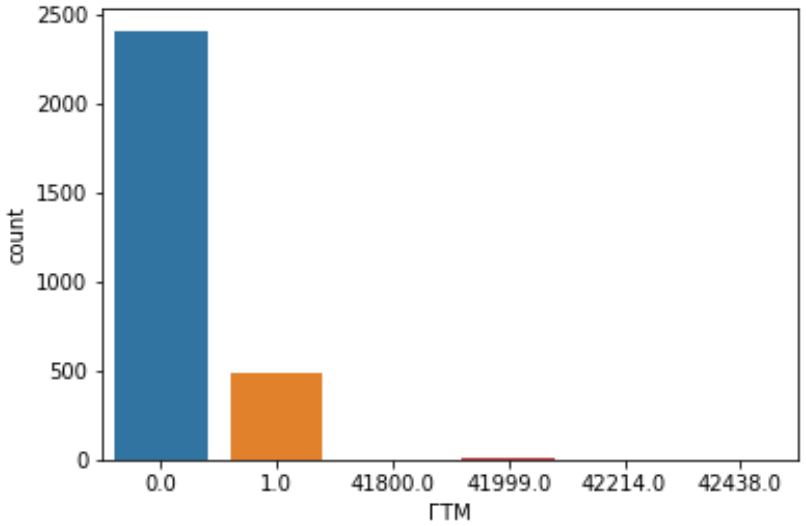

Рисунок 8. Наличие проведения ГТМ

Одно из главных преимуществ визуализации перед другими методами является возможность графически представить большое количество информации и понять, какие ошибки могли появиться после этапа подготовки данных для дальнейшего анализа. В данном случае можно заметить, что в этом наборе данных присутствуют значения, которые выбиваются из допустимого диапазона. Следовательно, рекомендуется удалить такие данные.

Заключение. В результате исследования были выполнены следующие задачи:

1. Применение алгоритма восстановления пропущенных значений для рассматриваемого набора данных.
2. Устранение ошибок и восстановление пропущенных значений с помощью алгоритма поиска ближайшего соседа (k-nearest algorithm imputation).
3. Визуализация атрибутов набора данных для демонстрации корректной работы используемого алгоритма.

ЛИТЕРАТУРА

1. Лидерами нефтегаза станут компании, использующие Big Data. [Электронный ресурс]. CNEWS. URL: http://www.cnews.ru/news/top/liderami_neftegaza_stanut_kompanii (дата обращения: 03.09.2019).

2. Геолого-технические мероприятия (ГТМ). [Электронный ресурс]. CNEWS. URL: https://www.petroleumengineers.ru/forum/39 (дата обращения: 27.08.2019).

3. Breazley, D. Python Cookbook, Third Edition / D. Breasley, B. K. Jones. – USA: O'Reilly Media, 2013. – 688 p.

4. McKinney, W. Python for Data Analysis. – USA: O'Reilly Media, 2013. – 453 p.

5. The use of KNN for missing values. [Электронный ресурс]. Towards Data Science. URL: https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637 (дата обращения: 03.09.2019).