МЕТОД ОПИСАТЕЛЬНОЙ СТАТИСТИКИ НА ПРИМЕРЕ ДАТАСЕТА С ИНФОРМАЦИЕЙ О НАЕЗДЕ НА ПЕШЕХОДОВ

Цыганкова А.В.
НИ ТПУ, ИШИТР, 8К13, e-mail: avc99@tpu.ru

Введение

В практических задачах в большинстве случаев встречается совокупность наблюдений, по этой причине появляется задача компактного описания имеющихся данных. Описательная статистика или дескриптивная статистика занимается систематизаций и наглядным представлением в форме графиков или таблиц, а также их количественным описанием. Целью такой работы является подготовка больших данных к прогнозному анализу [1]. В данной статье представлены методы описательной статистики на примере датасета, содержащего информацию о дорожно-транспортных происшествиях с участием пешеходов в США, штат Мичиган, округ Уэйн во временном интервале с 2010 по 2018 года. Для создания описательной статистики использовались средства программы Microsoft Excel.

Основная часть

Исследуемый датасет содержит 6809 записей, в которых каждая строка содержит пятнадцать параметров. Необходимо составить общую сводную таблицу для наглядного представления о имеющихся данных.

Таблица 1

Общая сводная таблица исследуемого датасета

Атрибут	Имя	Формат	Размерность
Crash Year	Год ДТП	int	–
Crash Month	Месяц ДТП	char	–
Crash Day	День ДТП	int	–
Time of Day	Время ДТП	char	–
Day of Week	День недели	char	–
City or Township	Город	char	–
Crash: Intersection	Наличие перекрестка	char	–
Crash: Hit-and-run	Наезд и бегство	char	–
Lighting Conditions	Условия освещения	char	–
Weather Conditions (2016+)	Погодные условия (с 2016)	char	–
Speed Limit in Crash Site	Ограничение скорости	int	mph
Worst Injury in Crash	Степень тяжести травмы	char	–
Party Type	Виновник ДТП	char	–
Person Age	Возраст пострадавшего	int	–
Person Gender	Пол пострадавшего	char	–

Далее следует произвести удаление повторяющихся строк. В исследуемом датасете повторяющиеся строки не были найдены. На следующем шаге выполнено дезагрегирование данных по атрибутам и построено их табличное представление и графическое изображение. Для каждого параметра необходимо найти все значения, которые он принимает, если такое представляется возможным.

Для параметров год, месяц, день ДТП найдены максимальное, минимальное, среднее значения. По полученным результатам можно сделать предварительные выводы о том, что в зимние месяцы происходит увеличение числа аварий, что вероятнее всего связано с ухудшением погодных условий. В атрибуте день ДТП минимальное значение приходится на 31ое число, однако это число встречается только 7 раз в году, в отличие от других чисел месяца, что следует учесть при дальнейшем анализе. Параметр время ДТП содержит информацию в виде часового интервала, проанализировать данный атрибут посредством Microsoft Excel не представляется возможным. При составлении статистики столбца, содержащего информацию о дне недели ДТП, были получены дни с минимальным – воскресенье и максимальным – пятница количеством аварий, а также среднее значение. Можно предположить, что наименьший показатель связан с уменьшением автомобильного трафика в выходные дни, а

наибольший с усталостью и снижением внимания после трудовой недели, как со стороны водителей, так и пешеходов.

Следующий параметр предоставляет информацию о городе, в котором случилось ДТП. Выбрав уникальные значения, получаем сорок три населенных пункта округа Уэйн штата Мичиган. По количеству аварий лидирует Детройт, однако численность населения Детройта во много раз превосходит численность остальных городов. Более информативен показатель количества аварий на тысячу жителей населенного пункта, поэтому были найдены дополнительные данные о численности населения этих городов. Таким образом наибольший показатель количества аварий в городе Хайленд Парк –8,88, а наименьший в Гросс Иль – 0,29. Также получено среднее значение – 2,03 аварии на тысячу жителей.

Атрибуты «Наличие перекрестка» и «Наезд и бегство» при обработке посредством программирования могут быть представлены в формате boolean так как содержат только по два значения. Получив количественные показатели по этим столбцам и посчитав процентное соотношение, имеем результат – аварий на перекрестках меньше на 13,1%, сбежавших водителей меньше на 26,3% от общего числа.

Параметр «Условия освещенности» принимает восемь возможных значений, среди которых встречаются: другое, ошибка, неизвестные данные. Так как таких строк менее 5% при дальнейшем анализе эти строки можно не учитывать. Если бы количество таких строк находилось в интервале от 5% до 50%, то следовало бы использовать наиболее встречающиеся значения или вариант «ближайших соседей». В случае если количество отсутствующих данных превышало бы 50%, то имеет смысл исключить такой атрибут из дальнейшего анализа. По количеству аварий лидирует «дневное освещение», на последнем месте «рассвет». Очевидно, что в дневное время автомобильный трафик наиболее высокий, что следует учесть при дальнейшем анализе.

Вслед за этим рассмотрим «Погодные условия». Этот параметр тоже содержит два значения, при которых данных о погоде нет: ошибка и неизвестные данные. Информацию о подоге начали учитывать с 2016 года, поэтому при использовании этого параметра имеет смысл разбить датасет на две части: до и после 2016 года. Наибольшее количество аварий в ясную погоду, наименьшее во время смога. Результат по наибольшему количеству также вполне очевиден ввиду большего автомобильного трафика.

Следующий атрибут принимает значения от 5 до 70 миль в час с шагом в 5 и несет информацию о скоростном режиме на участке дороги, на котором произошло ДТП. Больше всего аварий наблюдается при скоростном ограничении в 25 миль в час. Это может быть связано с психологическим фактором: на относительно небольшой скорости водители менее сосредоточены на дороге, что чаще приводит к авариям. Также это может быть связано с численностью населения, так как в крупных городах скоростной режим, как правило, ниже, чем малонаселённых городах.

Далее следует параметр, который стоит выбрать в качестве целевой функции – степень тяжести полученных травм. Чаще всего фиксируется факт возможных травм, а реже всего смертельный исход.

Параметр о виновнике ДТП принимает единственное значение: Motor vehicle driver (водитель автомобиля), поэтому не представляет интереса для рассмотрения.

На примере следующего атрибута рассмотрим применение графического изображения данных. На рисунке 1 видно, что с возраста 14-15 лет график начинает быстро расти, рост сохраняется до отметки в 19-20 лет, затем идет на убыль. Так как в данных утеряно 2470 значений, что составляет 36,3% от общего количества, для дальнейшего анализа необходимо выбрать вариант замены отсутствующих значений. Это может быть среднее значение, медиана или «ближайшие соседи».

Рис. 1. График зависимости количества ДТП от возраста пострадавшего

Последний параметр содержит информацию о половой принадлежности пострадавшего и имеет 1911 строк утерянных данных, которые следует заменить, используя вариант «ближайших соседей». Исходя из имеющихся данных мужчины на 18,6% чаще попадают в ДТП. Это можно объяснить психологическим фактором: мужчины чаще идут на риск, чем женщины.

Заключение

В ходе описательной статистики мы получили общее представление о имеющихся параметрах, выделили целевую функцию, нашли «пробелы» в данных, определили интересующие нас параметры, сделали предварительный выводы и предположения.

Список использованных источников

1. Описательный анализ данных. [Электронный ресурс]. – URL: https://www.statmethods.ru/statistics-metody/opisatelnyj-analiz-dannykh (дата обращения 12.02.2023).
2. Auto Pedestrian Crashes. [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/syedasimalishah/auto-pedestrians-crashes (дата обращения 2.10.2022)