Островский Кирилл

Теория вероятностей и статистика в машинном обучении

Лабораторная работа №1

Описательная статистика, визуализация, предварительная обработка данных

Excel-таблицы со всеми вычислениями собраны на страницах одного документа “ТВИМС\_Lab1.xlsx”

Часть 1

1. Поиск данных на сайте Росстат
   1. Нормальное распределение

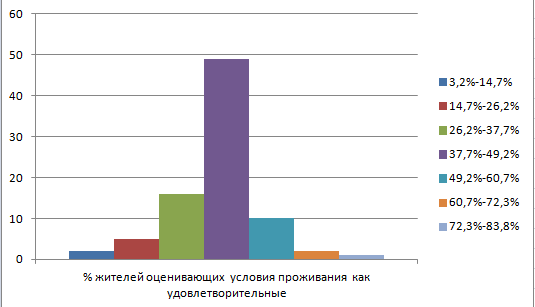


Рис.1 – Пример нормального распределения

Представленное гистограммой на рис.1 распределение близко к нормальному. Значения, находящиеся в середине выборки, являются самыми популярными, создавая так называемый “колокол” нормального распределения.

Источник: <https://rosstat.gov.ru/free_doc/new_site/GKS_KOUZH_2022/index.html>

* 1. Равномерное распределение



Рис.2 – Пример равномерного распределения

Гистограмма на рис.2 отображает количество поездок совершенных россиянами заграницу с 2014 по 2023 годы. Распределение равномерное, так как каждый год количество поездок изменялось не сильно. В 2020 году из-за эпидемии коронавируса произошел резкий спад поездок, но постепенно их количество возвращается к своему обычному значению. В данной выборке отсутствует информация за более ранний период, но можно предположить, что тогда ситуация была аналогична промежутку 2014-2019.

Источник: <https://rosstat.gov.ru/statistics/turizm>

1. Распределения со смещенной медианой относительно среднего и несмещенной

Для примера можно также взять данные оценки жилищных условий, использованные в первом пункте данной лабораторной работы. Так, на рисунке 1 показано распределение с несмещенной медианой (<15%).

Медиана = 42,8%

Среднее = 42,0%

Смещенная медиана (>15%) хорошо видна на другой гистограмме той же выборки, которая приведена на рис.3.



Рис.3 – Пример распределения со смещенной медианой

Медиана = 2,9%

Среднее = 3,4%

1. Описательные характеристики распределения

В качестве еще одной описательной характеристики, наряду с медианой и средним значением используется мода. Воспользуемся статистикой цен на продукты общего потребления, чтобы найти подходящий пример (рис.4).

Медиана = 17,87 руб/кг

Среднее = 19,05 руб/кг

Мода = 26,41 руб/кг

В данном случае медиана смещена относительно среднего значения всего на 4% и распределение является нормальным.

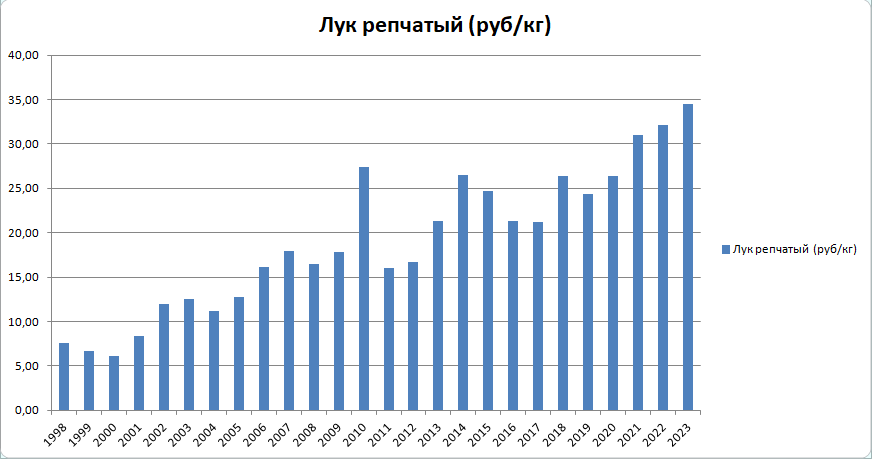


Рис. 4 – Цены на репчатый лук по годам

1. Выбросы и диаграмма “ящик с усами”

Для поиска выбросов воспользуемся тем же набором данных, что и в пунктах 1 и 2. Определим все необходимые характеристики наборов данных для построения графиков. Они представлены на рисунке 5.

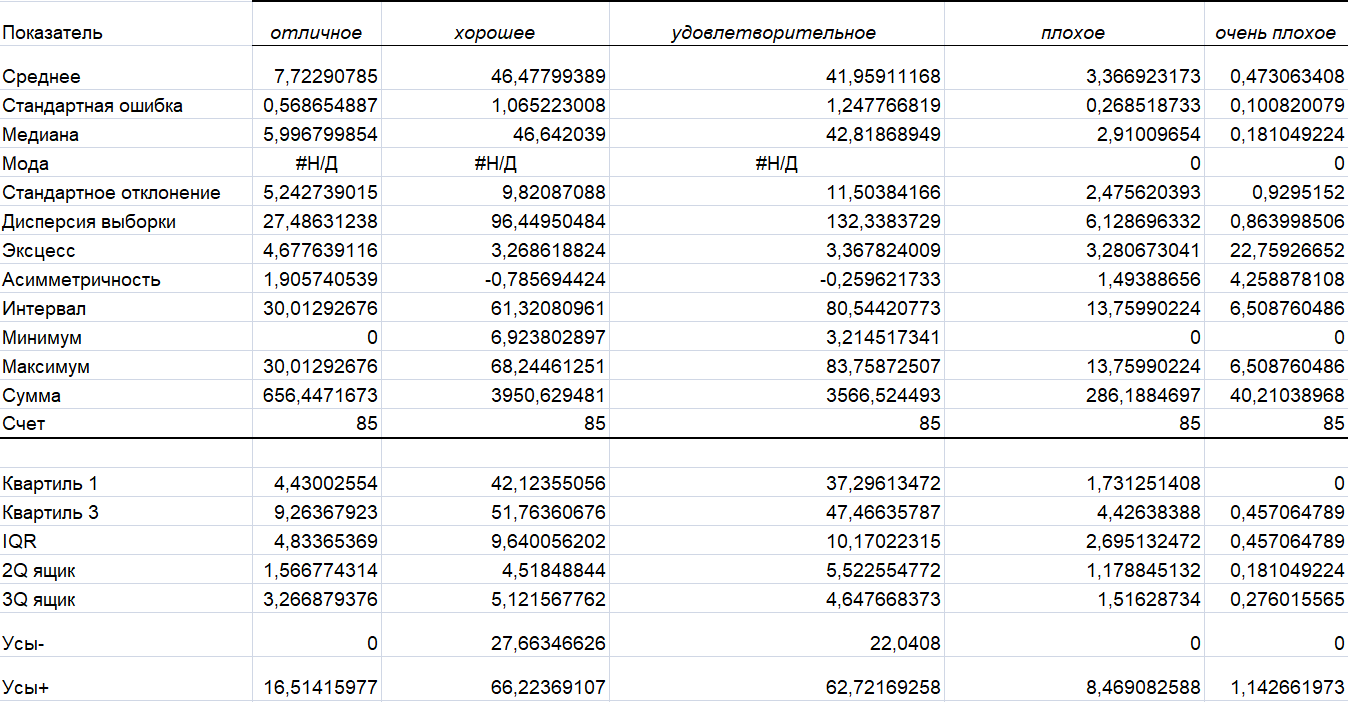


Рис. 5 – Характеристики набора данных

Используя эти характеристики, можно построить для каждой оценки диаграмму “ящик с усами” (рис.7), а также определить выбросы (рис.6), выходящие за пределы допустимых отклонений.

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

Рис. 6 – Выбросы значений (отмечены красным)

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

Рис. 7 – Диаграммы “ящик с усами” для каждой оценки

В данном наборе данных количество выбросов довольно велико, но это и неудивительно, учитывая большую разницу в условиях проживания между различными регионами нашей страны. Так, очевидно, в Москве будет выброс в большую сторону для оценки “отлично”. А в северных и дальневосточных регионах, где условия жизни намного тяжелее, а возможностей для обновления жилых помещений гораздо меньше, присутствуют аналогичные выбросы, но уже для оценки “очень плохо”. В некоторых регионах на полученных значения может сказываться культура и характер жителей региона. Так, в Чеченской республике абсолютное большинство жителей полностью удовлетворены условиями своего проживания.