Теория вероятностей и статистика в Машинном Обучении

Описательная статистика разброса данных

Размах

Стандартное отклонение

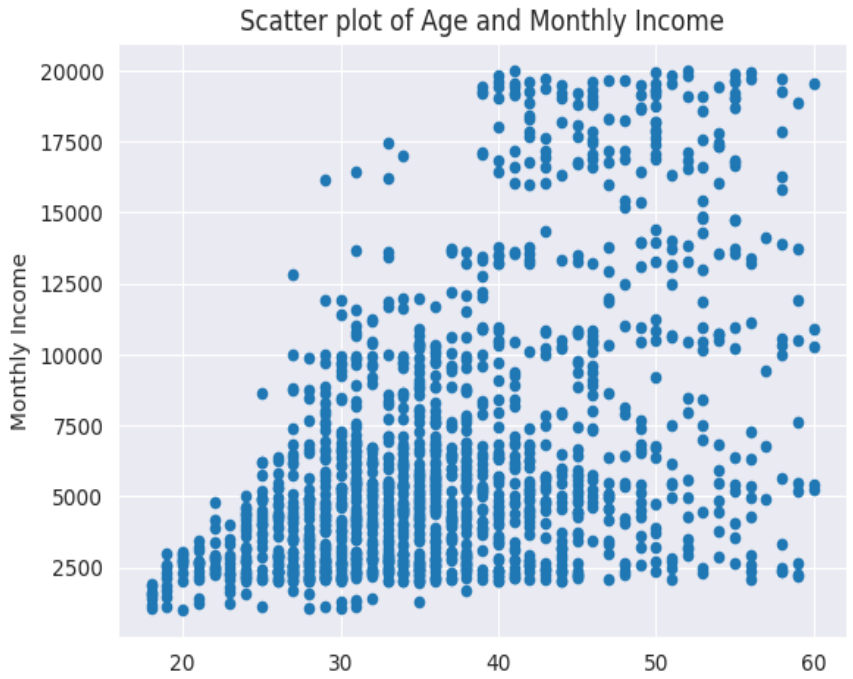
Квартили

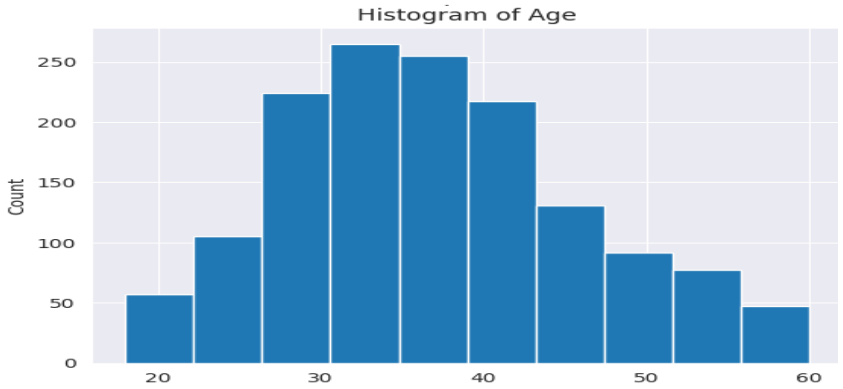
Межквартильный размах

Диаграммы размаха (выбросы)

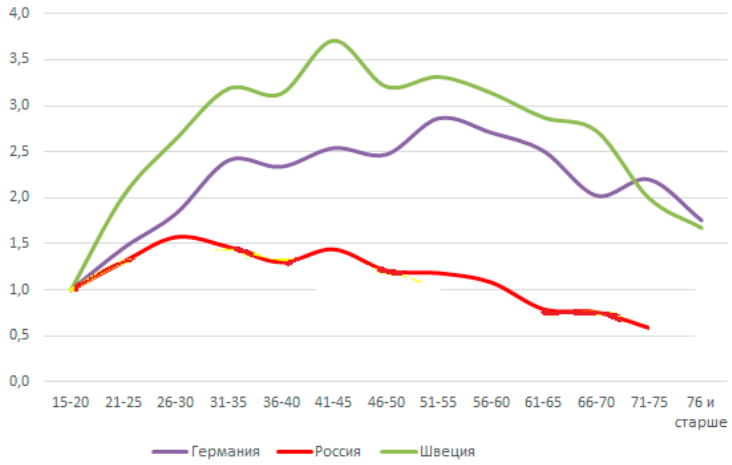
Разброс данных

Зависимость ежемесячного дохода от возраста

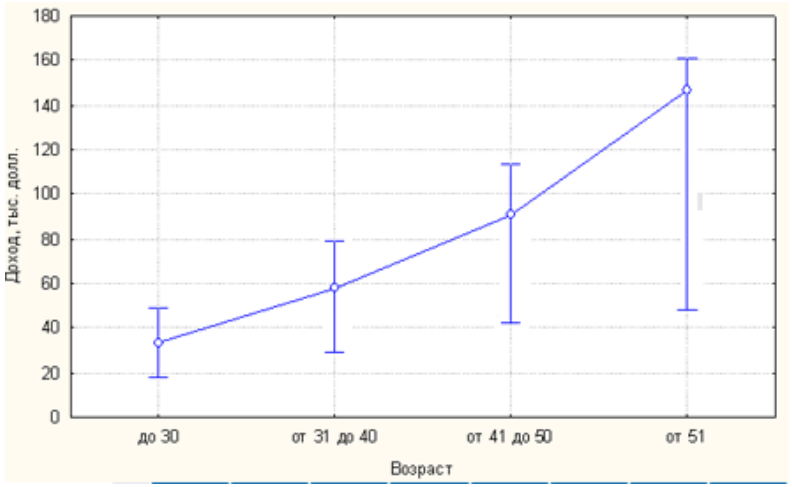


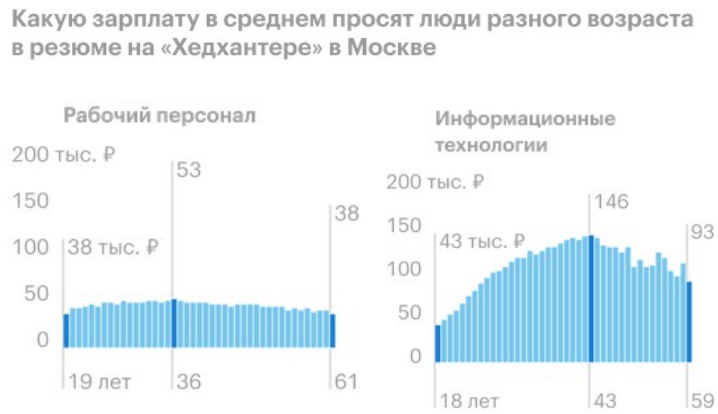


Различие по странам



Размах дохода от возраста





Измерение разброса данных

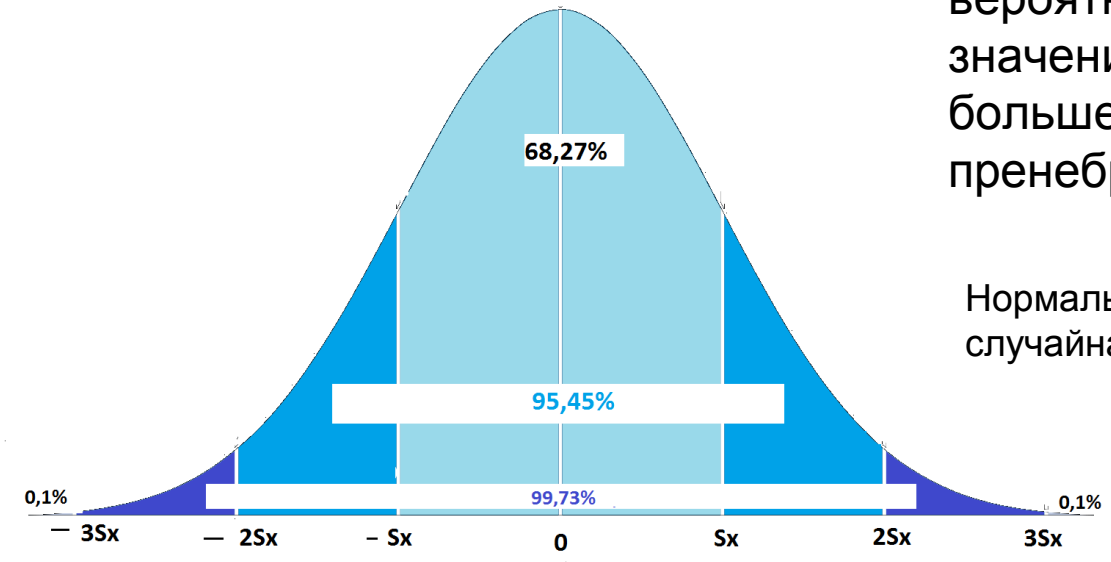
*Чтобы делать выводы на основании данных нужно знать среднюю тенденцию и понимать, насколько далеко от средней тенденции может **отклониться** исследуемый признак.*

1. Размах (интервал): $R_X = x_{(\max)} - x_{(\min)}$

2. Стандартное отклонение (σ) $s_X = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}.$

Стандартное отклонение

Для нормального закона:



Правило 3 сигм

Если случайная величина распределена по нормальному закону, вероятность того, что эта случайная величина примет значение, отклоняющееся от математического ожидания больше чем на три среднеквадратических отклонения пренебрежимо мала (не превышает 0,28%)

Нормальный закон распределения применим и тогда, когда изучаемая случайная величина является суммой большего числа случайных слагаемых

68,27 % всех значений такого признака отклоняются от среднего не больше чем **на одно** стандартное отклонение Sx

95,45 % всех значений такого признака отклоняются от среднего не больше чем **на два** стандартных отклонений Sx

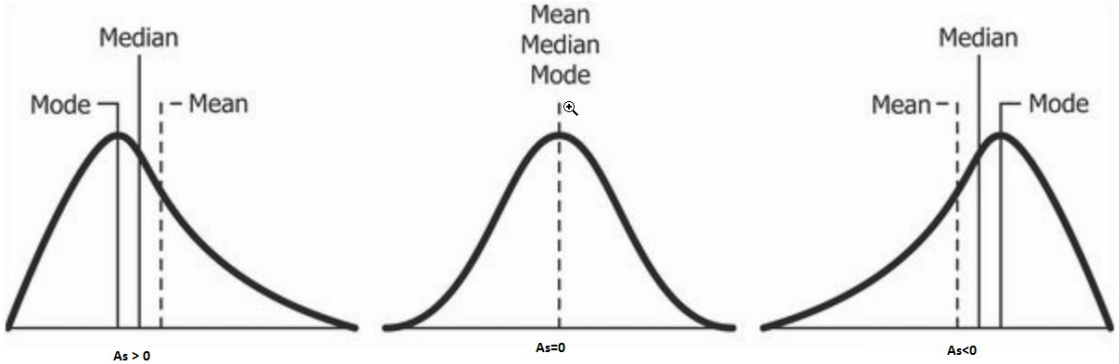
99,73 % всех значений такого признака отклоняются от среднего не больше чем **на три** стандартных отклонения Sx

Асимметрия распределения

$$As = \frac{\sum \left(x_i - \overline{x} \right)^3}{n s_x^3}$$

Асимметричность (коэффициент асимметрии) - показывает, насколько распределение набора данных является скошенным:

- у распределений, скошенных вправо, коэффициент асимметрии >0
- у идеально симметричных распределений коэффициент асимметрии = 0
- у распределений, скошенных влево, коэффициент асимметрии < 0



*Коэффициент асимметрии положителен, если **правый** хвост распределения **длиннее** левого, и отрицателен в противном случае.*

Квартили

Квартили — числа, делящие организованный в порядке неубывания ряд значений признака X на четыре равные по численности части:

25% всех значений признака X не больше **первой** (или нижней) квартили $\mathbf{X}_{0,25}$, а оставшиеся 75% значений — не меньше $\mathbf{X}_{0,25}$

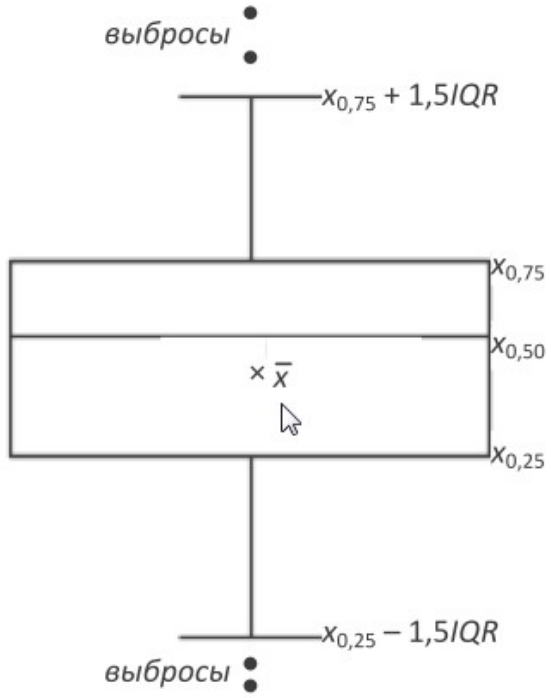
50% всех значений признака X не больше **второй** (или средней) квартили $\mathbf{X}_{0,50}$, а оставшиеся 50% значений — не меньше $\mathbf{X}_{0,50}$ ($\mathbf{X}_{0,5}$ = медиана)

75% всех значений признака X не больше **третьей** (или верхней) квартили $\mathbf{X}_{0,75}$, а оставшиеся 25% значений — не меньше $\mathbf{X}_{0,75}$

Межквартильный размах: $IQR = x_{0,75} - x_{0,25}$

Диаграмма размаха («ящик с усами»)

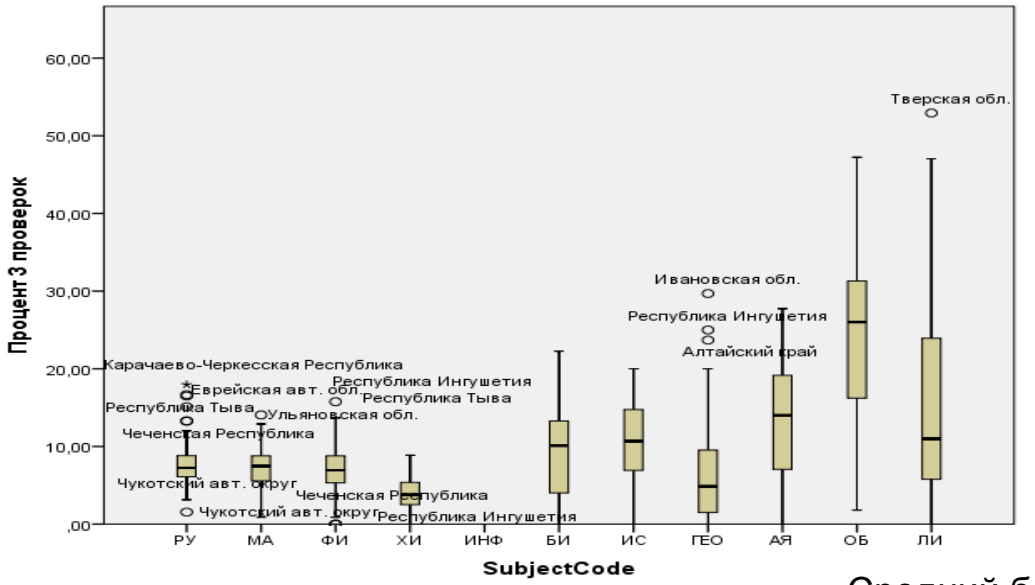
Выбросы: Значения не попадающие в интервал $[x_{0,25} - 1{,}5IQR;\ x_{0,75} + 1{,}5IQR]$



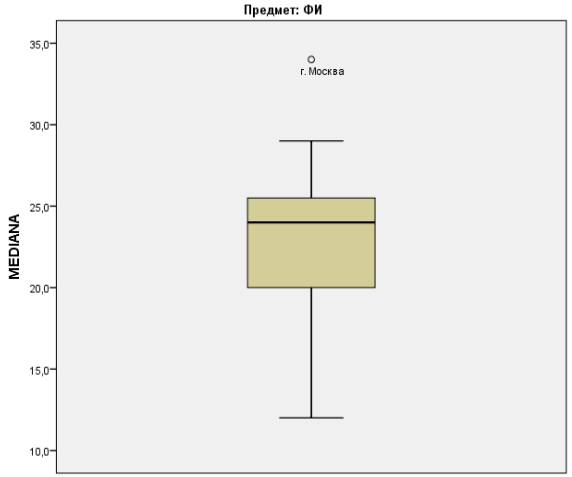
«Усы»: $\left[\max\left\{x_{(\min)};\ x_{0,25} - 1{,}5IQR\right\};\ \min\left\{x_{0,75} + 1{,}5IQR;\ x_{(\max)}\right\}\right].$

Примеры использования диаграммы размаха

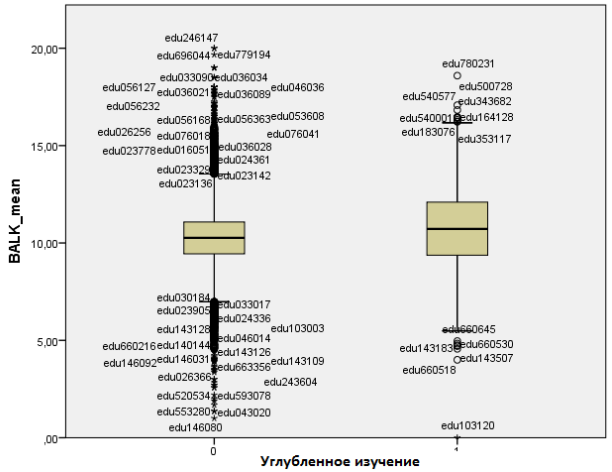
Оценка качества проверки экспертами



Медиана баллов для поиска необъективности



Средний балл ВПР математика по ОО



Предварительная обработка данных

1. Пропущенные значения, повторяющиеся строки

2. Нахождение выбросов

3. Описательная статистика:

4. Исследование распределения

Теория вероятностей и статистика в Машинном Обучении

Лабораторная работа №1.1

Описательная статистика, визуализация, предварительная обработка данных

Часть 1

1. Найти на сайте РосСтат https://rosstat.gov.ru/free_doc/new_site/zdor22/PublishSite_2022/index.html данные, распределение которых было бы близко к
   a) нормальному (визуально выглядит как кривая Гаусса - «колокол»),
   b) равномерному.

2. Найти распределение со смещенной медианой относительно среднего (~15% размаха) и несмещенной.

3. Посчитать описательные характеристики распределения выбранных данных (среднее, мода, медиана), дать визуальное представление данных (условное форматирование Excel, построить график рассеивания/ гистограмму/ круговую диаграмму).

Теория вероятностей и статистика в Машинном Обучении

Лабораторная работа №1.2

4. Найти датасет с аномальными значениями, которые можно выявить, построив диаграмму размаха «ящик с усами». Попробовать выдвинуть гипотезу о причинах аномальности, которую можно подтвердить дальнейшими исследованиями, используя синтетические признаки (или корреляционный анализ и т.п.).

https://statanaliz.info/excel/diagrammy/diagramma-yashhik-s-usami-boxplot-v-excel-2016/

https://excel2.ru/articles/blochnaya-diagramma-v-ms-excel