Конспект по теме "Первые графики и выводы"

Знакомство с задачей

Данные в формате *csv* в качестве разделителя могут иметь не только запятые, но и точки с запятой, знаки табуляции или другие символы. Десятичные дроби, записанные с запятой, тоже могут внести путаницу.

В параметрах функции `read_csv()` можно указать, *к*акими символами разделять колонки и дроби. Разделитель колонок задают параметром **sep** , а дробей — параметром **decimal**:

```
file = pd.read_csv('file.csv',  sep=';' , decimal=',')
```

Сводные таблицы для расчёта среднего

Занимаясь предобработкой данных, вы применяли `pivot_table()` — метод для построения сводных таблиц. Прежде значением *aggfunc* вы указывали *sum*, то есть складывали элементы столбца. Если параметр *aggfunc* не указывать, то по умолчанию метод `pivot_table()` рассчитает среднее арифметическое значений, указанных в параметре *values*.

Базовая проверка данных

В работе с данными почти всегда вас ждут сюрпризы:

- Из-за непонимания задачи или случайно выгрузили не те или неполные данные.
- Ошибки в алгоритмах, считающих нужное значение
- Не тот формат предоставляемых данных.
- Упущен какой-нибудь существенный факт

Словом, в данных может быть всё, что угодно. Именно вы как аналитик ручаетесь за их реалистичность. Попробуйте оценить, насколько они достоверны. Начните с базовых проверок. Например, несложно ответить на

вопросы по данным и самостоятельно либо с помощью коллег оценить, похожи ли результаты ваших расчётов на правду.

Базовая проверка может обнаружить проблему в данных. Или наоборот — свидетельствовать, что с ними всё в порядке. По крайней мере, пока.

Гистограмма

Гистограмма — это график, который показывает, как часто в наборе данных встречается то или иное значение. Гистограмма объединяет числовые значения по диапазонам, то есть считает частоту значений в пределах каждого интервала. Её построение подобно работе знакомого вам метода `value_counts()`, подсчитывающего количество уникальных значений в списке, однако `value_counts()` группирует строго одинаковые величины и хорош для подсчёта частоты в списках с категориальными переменными.

В *Pandas* гистограмму строит специальный метод `hist()`, применяемый к списку или к столбцу датафрейма. Метод `hist()` находит в наборе чисел минимальное и максимальное значения, а полученный диапазон делит на области, или корзины. Затем `hist()` считает, сколько значений попало в каждую корзину, и отображает это на графике. Параметр **bins** определяет, на сколько областей делить диапазон данных, по умолчанию таких `bins=10`. По умолчанию, гистограмма выводится для всех значений от минимального до максимального. Масштаб можно изменить вручную, указав параметр **range:** `range=(min_value, max_value)`.

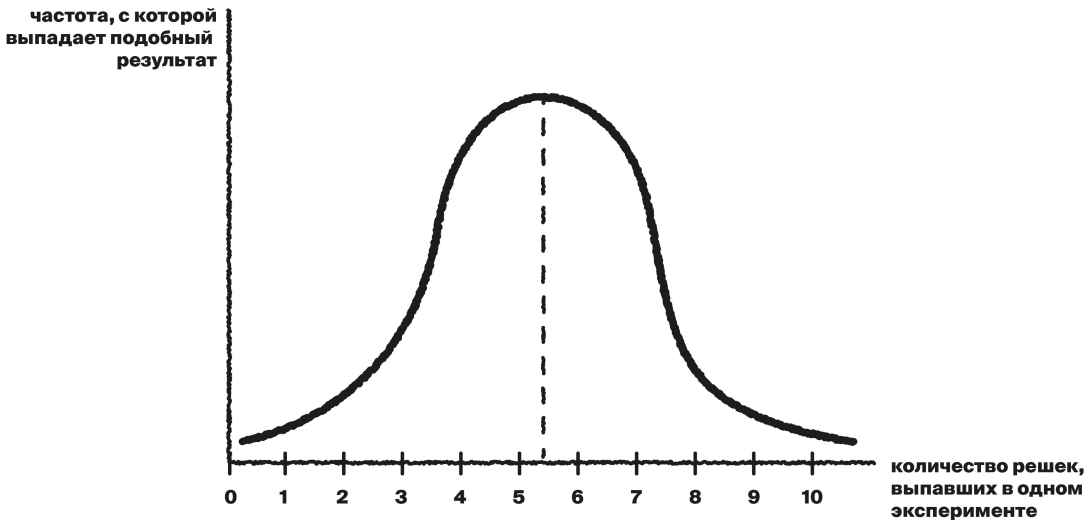
Для создания графиков (в том числе гистограмм), импортируют библиотеку **matplotlib**. Метод `show()` отображает графики.

```
import pandas as pd
import matplotlib.pyplot as plt # импортируем библиотеку, стандартно используется имя plt

pd.Series(...).hist(bins=n_bins, range=(min_value, max_value)

plt.show() # даём команду отобразить гистограмму
```

Распределения

Распределение - это все возможные значения переменной с частотой их появления. Различные распределения:

- **Нормальное**: чаще всего встречается среднее значение и близкие к нему, а крайние значения встречаются довольно редко



- **Распределение Пуассона:** число событий в единицу времени, если они в среднем происходят с измеренной частотой

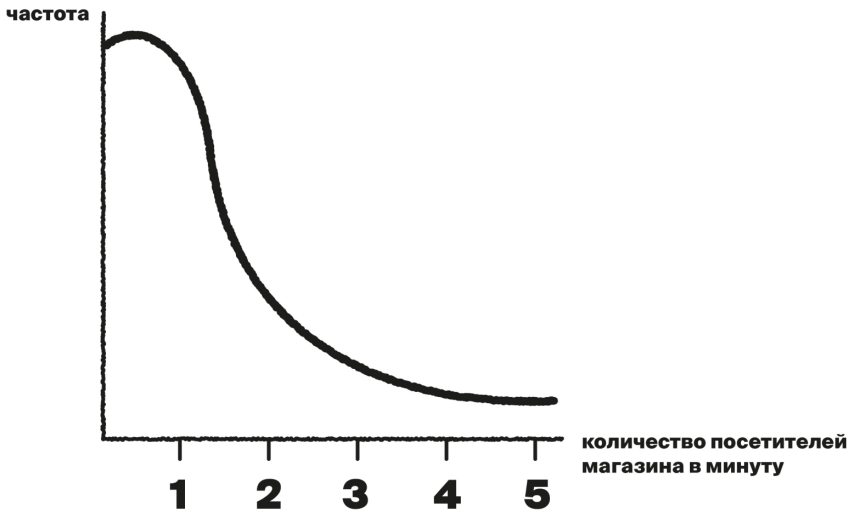


Диаграмма размаха

Характерный разброс — то, какие значения оказались вдали от среднего, и насколько их много.

Если считать характерным разбросом расстояние между минимальным и максимальным значением, то мы не всегда получим точное описание данных, на него могут повлиять выбросы. Поэтому, в качестве характерного разброса применяют межквартильный размах

Квартили разбивают упорядоченный набор данных на четыре части: **первый квартиль** *Q1* — число, отделяющее первую четверть выборки (25% элементов меньше, а 75% — больше него); **медиана — второй квартиль** Q2 (половина элементов больше и половина меньше неё); **третий квартиль** Q3 — это отсечка трёх четвертей (75% элементов меньше и 25% элементов больше него).

Межквартильный размах — это расстояние между первым квартилем *Q1* и третьим квартилем *Q3*.

Диаграмма размаха, или **ящик с усами**, позволяет отобразить все квартили для заданных данных.

«Ящик» ограничен первым и третьим квартилями. Внутри ящика обозначают медиану.

«Усы» простираются влево и вправо от границ ящика на расстояние, равное 1,5 **межквартильным размахам** (*IQR*). В размах «усов» попадают нормальные значения, а за пределами находятся выбросы, изображённые точками. Если правый «ус» длиннее максимума, то он заканчивается максимумом. То же — для минимума и левого уса.

В Python диаграмму размаха строят методом `boxplot()`, он позволяет визуально оценить характеристики распределения, не прибегая к гистограмме.

```
import matplotlib.pyplot as plt
data.boxplot()
plt.show()
```

Оси любого графика в pandas можно изменять, для этого нужно применить метод `xlim(x_min, x_max)` для оси X и `ylim(y_min, y_max)` для оси Y. Параметры в обоих случаях - минимальное и максимальное значение на графике

```
import matplotlib.pyplot as plt
plt.xlim(x_min, x_max)
plt.ylim(y_min, y_max)
```

Описание данных

С помощью гистограмм и диаграмм размаха можно получить графическое описание любого набора данных. Однако, не всегда по графикам можно

определить такие характеристики, как среднее, медиану, количество наблюдений в выборке и разброс их значений — **числовое описание данных**. В *Pandas* для этого применяется метод `describe()`.

Стандартное отклонение — числовая характеристика данных, входящая в числовое описание данных и характеризующая разброс величин, показывает, насколько значения в выборке отличаются от среднего арифметического. Оно позволяет понять природу распределения и определить, насколько значения однородны.

```
data['column'].describe()
```