Первые графики и выводы

Чтение данных из файла с использованием разделителей

```
# sep – разделитель столбцов
# decimal – разделитель десятичных знаков
data = read_csv('file.csv',  sep=';', decimal=',')
```

Уникальные значения в столбце

```
data['id'].unique() # список уникальных значений

len(data['id'].unique()) # количество уникальных значений
data['id'].nunique() # альтернативный способ подсчитать уникальные значения
```

Импорт библиотеки matpotlib

```
import matplotlib.pyplot as plt
```

Числовое описание данных для колонки

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

```
# минимум и максимум для оси X
plt.xlim(x_min, x_max)

# минимум и максимум для оси Y
plt.ylim(y_min, y_max)
```

Построение гистограммы

```
# строим гистограмму на 30 корзин для площади квартир
# отображаем только значения от 10 до 200 м²
real_estate['total_area'].hist(bins=30, range=(10, 200))
plt.show()
```

Первые графики и выводы

Диаграмма размаха («ящик с усами»)

```
real_estate.boxplot('total_area') # для одного столбца
plt.ylim(10, 200) # отображаем только квартиры с площадью от 10 до 200 м²
```

```
real_estate.boxplot(['total_area', 'living_area']) # для списка нужных столбцов
```

Глоссарий

Гистограмма — график, который показывает, как часто в наборе данных встречается то или иное значение.

Квартили (от лат. quartus — «четвёртый») — числа, которые разбивают упорядоченный набор данных на четыре части.

- **Первый квартиль** (Q1) отделяет первую четверть выборки: 25% элементов меньше, а 75% — больше него.
- **Медиана — второй квартиль** (Q2) — половина элементов больше и половина меньше неё.
- **Третий квартиль** (Q3) **— отсечка трёх четвертей, 75% элементов меньше и 25% элементов больше него.
- **Межквартильный размах** — расстояние между первым квартилем (Q1) и третьим квартилем (Q3).

Первые графики и выводы

Распределение — частота появления всех возможных значений переменной.

- При **нормальном распределении** чаще всего встречается среднее значение и близкие к нему, а крайние значения встречаются редко.
- **Распределение Пуассона** показывает число событий в единицу времени, если их средняя частота известна.

Стандартное отклонение показывает, насколько значения в выборке отличаются от среднего арифметического значения.

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

Числовое описание данных — среднее арифметическое значение, медиана, стандартное отклонение, количество наблюдений в выборке и разброс их значений.