Министерство образования и науки

федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет инфокоммуникационных технологий

Отчет по дисциплине: **«Современные инструменты анализа данных»**

Лабораторная работа 2

Выполнил:

–

Проверила:

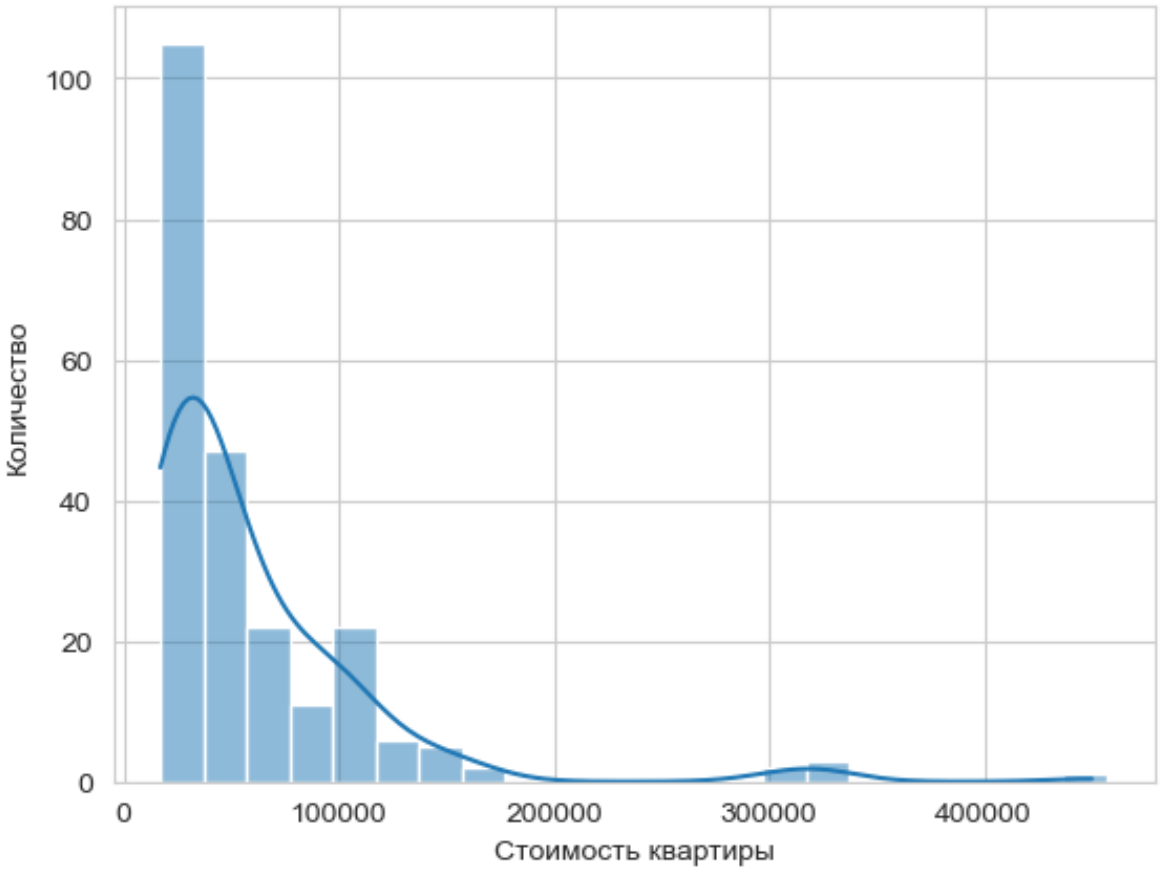
–

Санкт-Петербург

2023

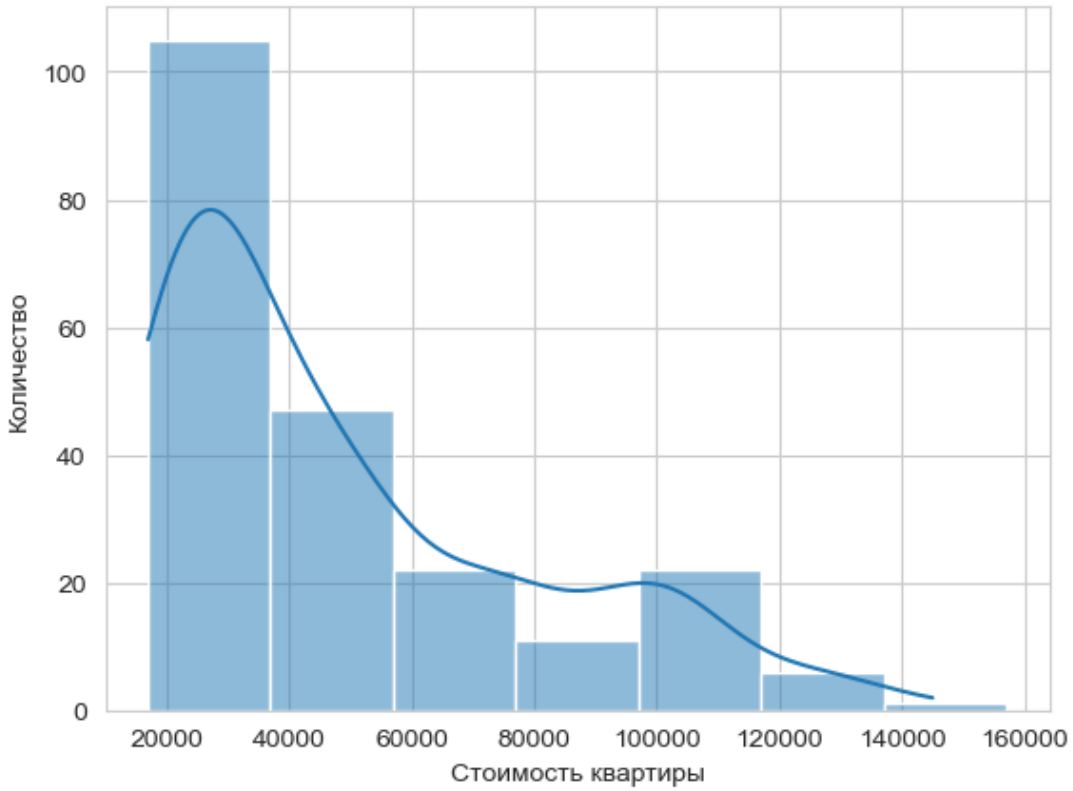
Первые две части делались с использованием **python** и последующих модулей для работы с данными и их визуализации: **pandas**, **seaborn**, **matplotlib**, **scipy**.

Часть 1. Анализ одномерных данных



1. Гистограмма распределения стоимости квартир до отбрасывания данных

Гистограмма была построена с шагом в 20.000. Можно заметить, что правая часть всего распределения слишком вытянута из-за выбросов в зарплатах, т.е. аномально высоких значений относительно всей выборки. Следующим шагом оставим только значения меньше 5-ого квантиля с правой стороны, иными словами, меньше 95-ого квантиля.



2. Гистограмма распределения стоимости квартир до 95-ого квантиля

Часть 2. Анализ одномерных данных

Необходимо проверить гипотезу о статистической значимости различия между доходами и рабочими часами двух групп работающих и получающих доход граждан Петербурга разными способами.

Группы:

- 1 группа - имеющие образование среднее и ниже,
- 2 группа - имеющие среднее специальное или высшее образование

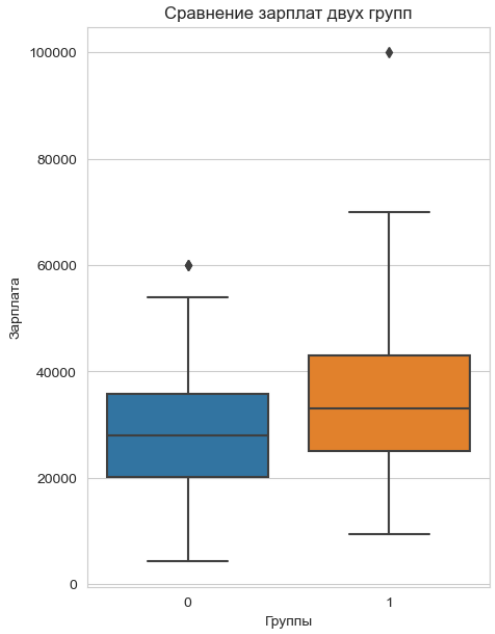
Способы:

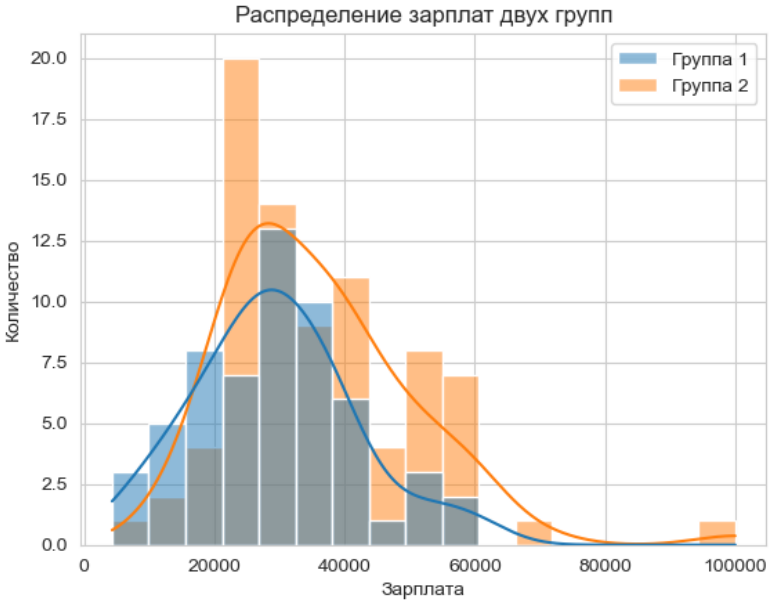
- Однофакторный дисперсионный анализ
- Т-тесты

Поставим две гипотезы для p-value:

- **Нулевая гипотеза (H0)**: Средние значения во всех группах равны, т.е., различий между группами нет.
- **Альтернативная гипотеза (H1)**: Средние значения хотя бы в одной из групп различаются от средних значений в других группах.

Исследование разницы в зарплатах:



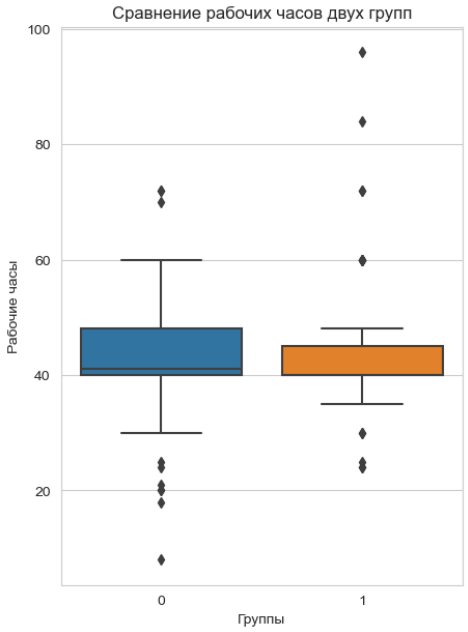


	p-value	statistic
f-value	0.005187	0.005187
T-test	8.068225	-2.840462

Результаты статистических проверок

В обоих подходах подсчет p-value совпал. Так как его значение меньше 0.05, то существуют статистически значимые различия между группами.

Исследование разницы в рабочих часах:



	p-value	statistic
f-value	0.680109	0.680109
T-test	0.170725	-0.413189

Результаты статистических проверок

В обоих подходах подсчет p-value совпал. Так как его значение больше 0.05, то нет статистически значимых различий между группами.

Часть 3. Проведение A/B тестов

Для первого задания результаты подсчетов представлены ниже. Итоговый размер выборки количества человек равняется 5547, а в каждом варианте по 2774 человек (вариант отправки по sms и по почте).

Для второго задания, исходя из расчетов, вышло, что лучше всего в качестве обратной связи использовать канал WhatsApp.