#### Министерство науки и высшего образования Российской Федерации

#### Федеральное государственное бюджетное образовательное учреждение

#### высшего образования

#### «Владимирский государственный университет

#### имени Александра Григорьевича и Николая Григорьевича Столетовых»

**(ВлГУ)**

**Кафедра информационных систем и программной инженерии**

Рейтинг-контроль №2

по дисциплине

"Технологии искусственного интеллекта"

Выполнил:

ст. гр. ПРИм-124

Парахин К.В.

Приняла:

доц. Кафедры ИСПИ

Озерова М.И.

Владимир, 2025 г.

**Цель работы**

Познакомиться с библиотеками Python для реализации задачи анализа данных датасэтов.

Задание из презентации с титаником  
  
Cсылка на Google Colab: <https://colab.research.google.com/drive/1yXYWFPb9DzADs0m1YCbwk0SDQRT_d6ks?usp=sharing>   
  
Все входные данные были загружены в отдельную папку на моем Google Drive – и к ним был оставлен доступ по ссылке

Чтобы смочь грамотно импортировать csv файл в Google Colab – я открыл свой файл, находящийся на диске через Google Sheets – а затем оттуда уже взял ссылку на инструмент по его преобразованию из .gsheet формата в .csv формат.

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

sheet\_id = "1Lho5kk-l3w6nBBFhMzSZ-0a58fLJK2c5XJl8Jq85YH8"

url = f"https://docs.google.com/spreadsheets/d/{sheet\_id}/export?format=csv"

df = pd.read\_csv(url)

#print(df.to\_markdown())

males\_total = len(df.loc[(df['Sex'] == 'male')])

females\_total = len(df.loc[(df['Sex'] == 'female')])

males\_survived = len(df.loc[((df['Sex'] == 'male') & (df['Survived'] == 1))])

females\_survived = len(df.loc[((df['Sex'] == 'female') & (df['Survived'] == 1))])

# Визуализируем

survivors\_counts = {

    'Выжили': [males\_survived, females\_survived],

    'Погибли': [males\_total - males\_survived, females\_total - females\_survived]

}

fig, ax = plt.subplots()

bottom = np.zeros(2)

for key, count in survivors\_counts.items():

    p = ax.bar(('Мужчины', 'Женщины'), count, width=0.6, label=key, bottom=bottom)

    bottom += count

    ax.bar\_label(p, label\_type='center')

ax.set\_title('Распределение выживших в зависимости от пола')

ax.set\_ylabel('Количество человек')

ax.legend()

plt.show()

from collections import Counter

# Выберем списки пассажиров в отдельные группы

survivors = df.loc[df['Survived'] == 1]

male\_survivors = df.loc[(df['Sex'] == 'male') & (df['Survived'] == 1)]

female\_survivors = df.loc[(df['Sex'] == 'female') & (df['Survived'] == 1)]

# Число возрастов

num\_of\_ages = len(dict(Counter(survivors['Age'])).keys())

# Визуализируем

fig, axs = plt.subplots(2, 2)

fig.set\_figwidth(20)

fig.set\_figheight(10)

names = [['Все пассажиры', 'Выжившие'], ['Выжившие мужчины', 'Выжившие женщины']]

for i, surv in enumerate([[df, survivors], [male\_survivors, female\_survivors]]):

    for j, subsurv in enumerate(surv):

        axs[i][j].hist(subsurv['Age'], bins=num\_of\_ages)

        axs[i][j].set\_title(names[i][j])

        axs[i][j].set\_xlabel('Возраст, лет')

        axs[i][j].set\_ylabel('Количество человек')

        axs[i][j].set\_ylim(0, 30)

        axs[i][j].set\_xlim(0, 70)

plt.show()

df.loc[((df['Survived'] == 1) & (df['Age'] > 0) & (df['Age'] > 18))].Age.hist()

В данном случае были загружены данные в Dataframe – а далее по разным измерениям входных данных (параметрам, связанным с распределением людей по возрасту, полу, классификатору выжившего человека – Survived – и остальным) были выполнены простые математические аналитические операции

Например, было посчитано число человек выживших и погибших в зависимости от пола – и это распределение отображено на гистограмме.

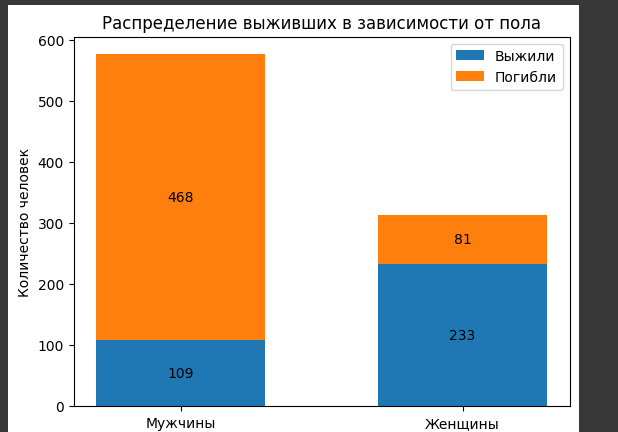
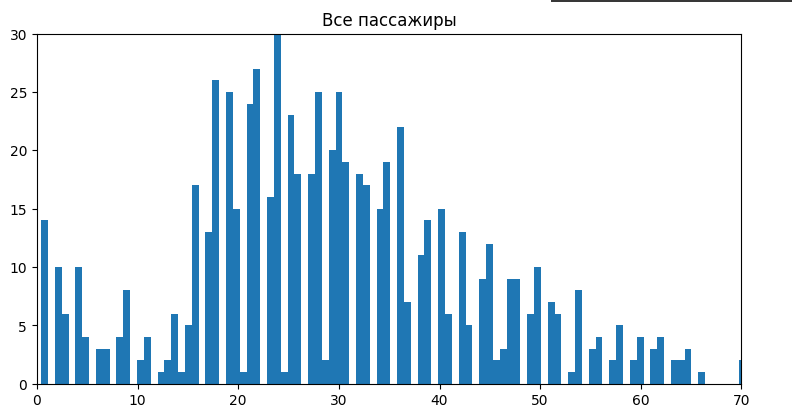


Рисунок 1.1. Распределение выживших в зависимости от пола

Также была в отдельные коллекции отфильтрована выборка выживших людей в зависимости от возраста – в этой выборке по абсциссам идет дискретный интервал возрастов – от 0 лет до 70 лет – а по ординатам на графике – кол-во пассажиров, которые относятся к каждому из значений внутри возрастного интервала



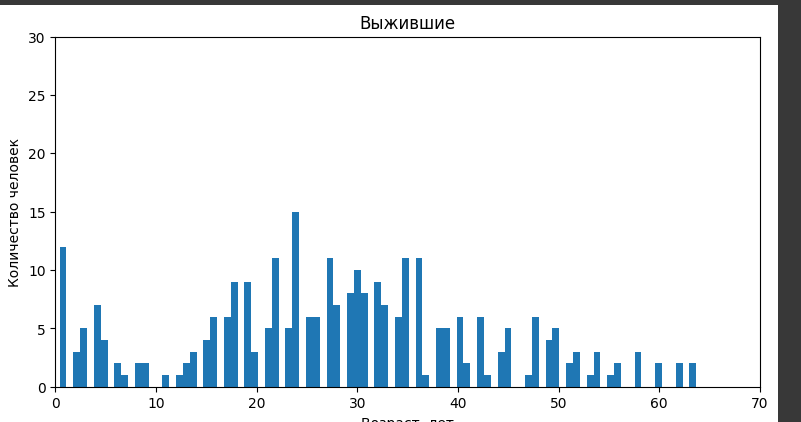


Рисунок 1.2-1.3. Столбчатые диаграммы вне зависимости от пола выживших

Эти же манипуляции были выполнены и в разрезе пола выжившего человека, на одном графике показана стобчатая диаграмма для выживших мужчин – а на другом графике аналогичная диаграмма для выживших женщин

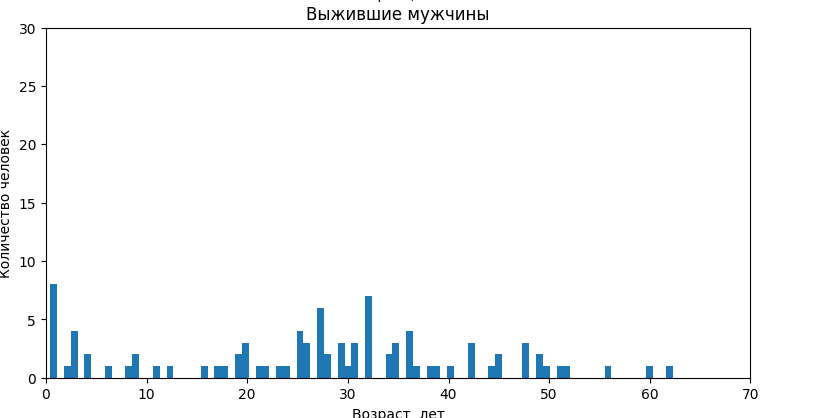




Рисунок 1.4-1.5. Столбчатые диаграммы в зависимости от пола выживших

Кроме того, можно посмотреть распределение процента выживших среди определенной группы людей (до 18 лет – детей), сделать это на гистограмме:

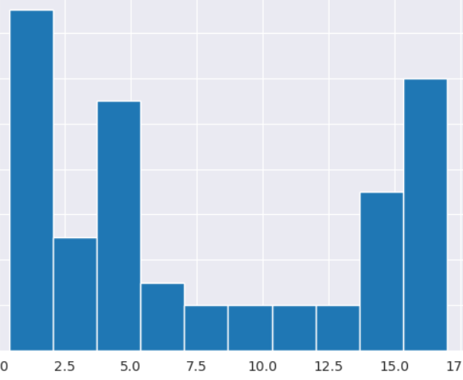


Рисунок 1.6. Распределение по выжившим детям

Далее выполняем подобный анализ для данных по населению городов России в 1959 году

Ссылка на Google Colab <https://colab.research.google.com/drive/13YE9LUzT9wqL-unOAwJv0_iXZ6tr0Ugk>

Для начала выполняем импорт данных из Google Sheets таблиц .csv файла с гугл диска:

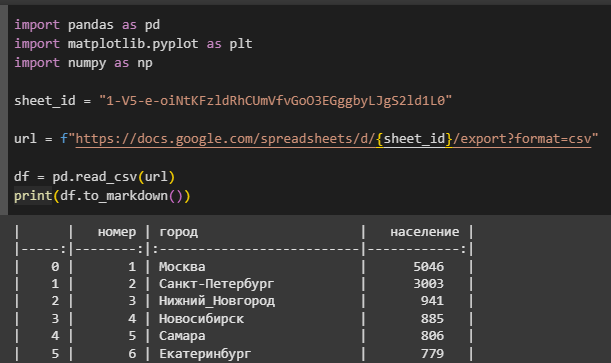
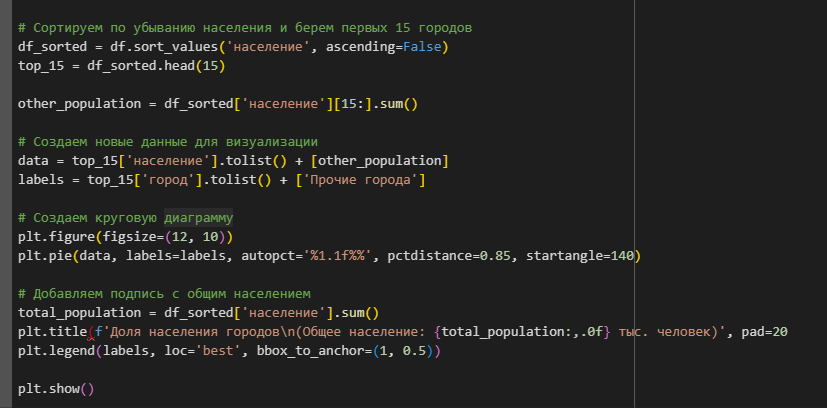


Рисунок 2.1. Импорт данных

Сразу же хочется выполнить анализ распределения численности между крупными городами того времени – для этого необходимо отсортировать данные по населению

Далее я понял, что визуализировать на графике всю тысячу городов будет ненаглядно – поэтому я решил взять только первые 15 городов – а остальные численности населения просуммировать в группу «Прочие города» - и затем в процентном соотношении к общей численности изобразить распределение на круговой диаграмме



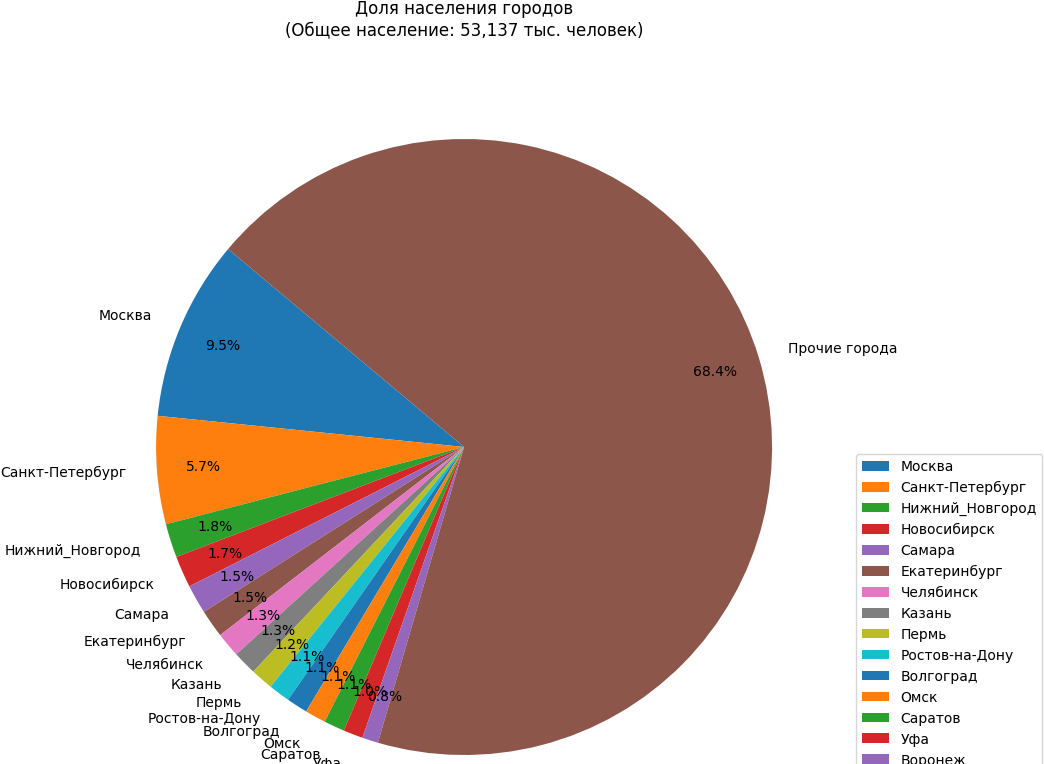


Рисунок 2.2 – 2.3. Распределение численности конкретных городов на круговой диаграмме

Далее я решил сделать анализ населения городов по группам – группы формируются по принадлежности к категории города по его численности населения:

    'Не города (<10 тыс.)',

    'Микро города (10-30 тыс.)',

    'Маленькие города (30-100 тыс.)',

    'Обычные города (100-250 тыс.)',

    'Средние города (250-500 тыс.)',

    'Крупные города (500 тыс.-1 млн)',

    'Огромные города (>1 млн)'

Тем самым возможно собрать статистику по имеющимся данным и понять, как непосредственно распространены те или иные категории городов по стране – удобнее всего это сделать с помощью гистограммы – отсортированной по убыванию числа городов в каждой категории:

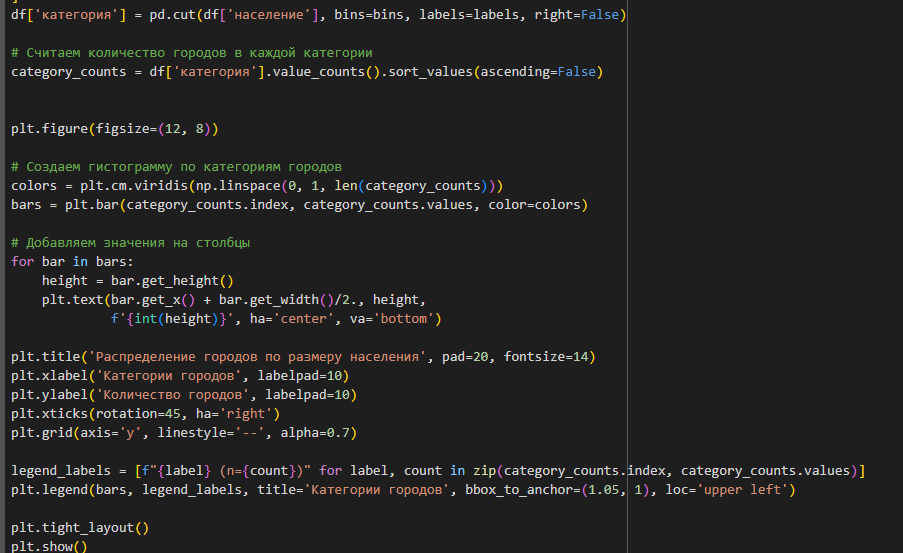


Рисунок 3.1. Просчет принадлежности городов к категориям, создание и визуализация гистограммы данных, оформление легенды

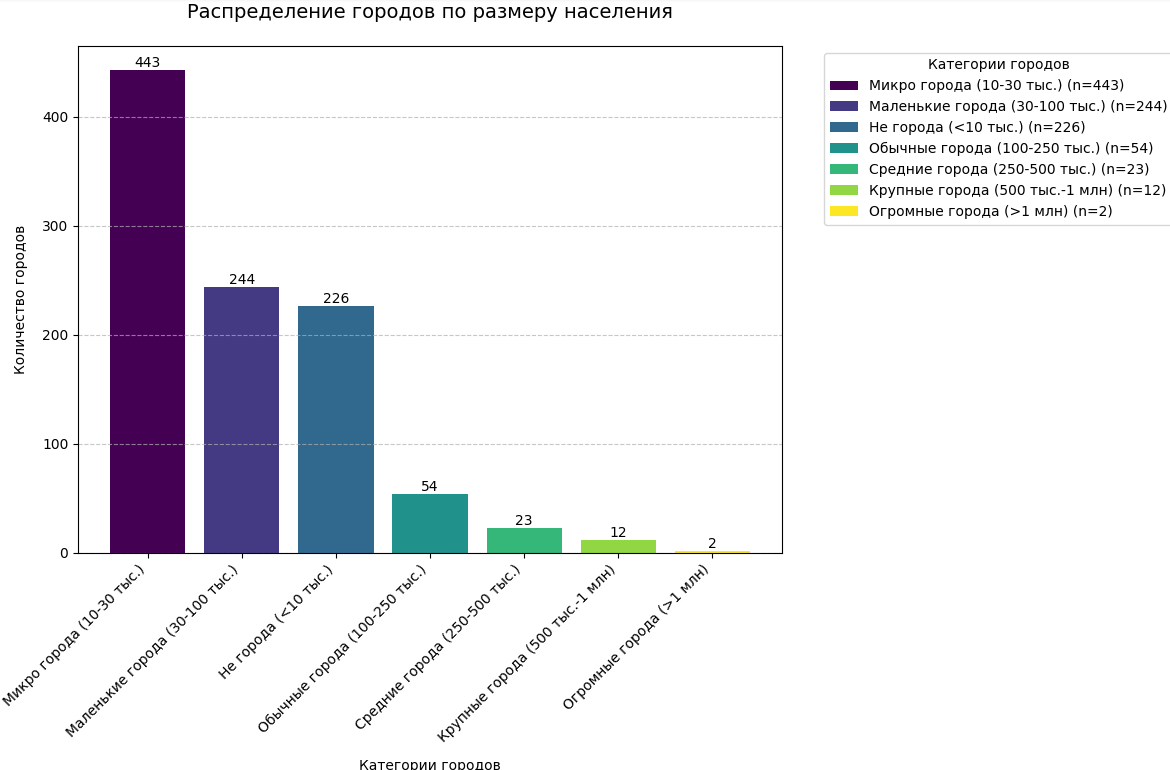


Рисунок 3.2. Гистограмма распределения категорий городов по стране

В рамках последнего сценария анализа я решил сравнить суммарное население по полученным категориям городов – при этом 2 города с населением больше 1 миллиона: Москва и Санкт-Петербург взять отдельными категориями

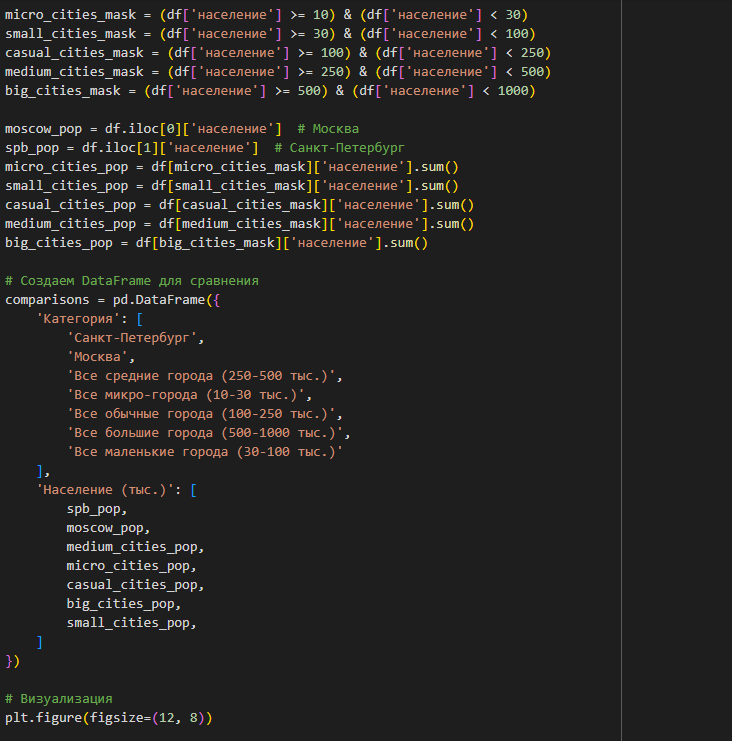


Рисунок 4.1. Подсчет суммарной численности по категориям

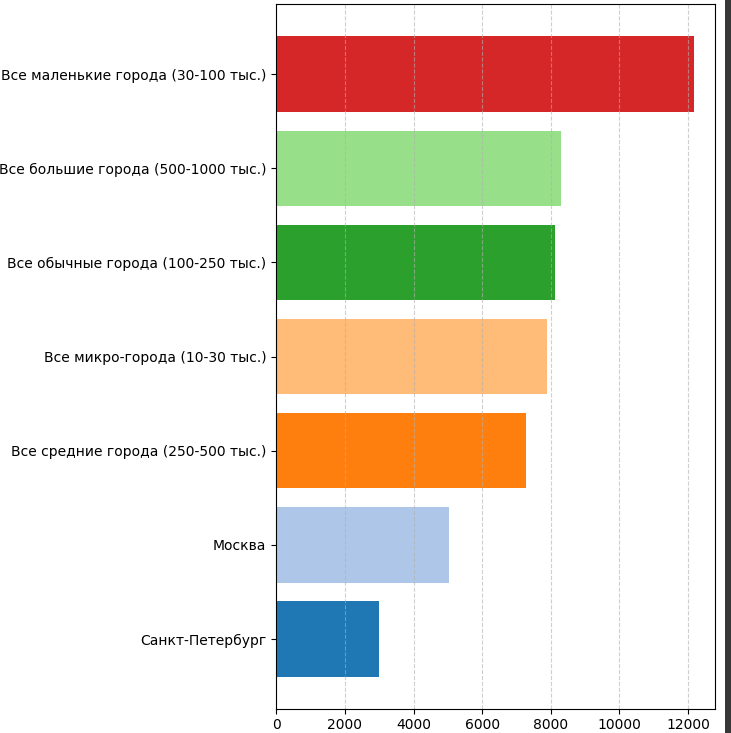


Рисунок 4.2. Визуализация в виде вертикальной отсортированной гистограммы

Вывод

В результате выполнения работы я познакомился с основными средствами анализа данных по датасэту в Google Colab средствами языка Python.