|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ **Информатика, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ и системы  
 управления**

КАФЕДРА **Компьютерные системы и сети (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/07 Интеллектуальные системы анализа,**

**обработки и интерпретации больших данных**

**Отчет**

**по лабораторной работе №1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент | ИУ6-32М |  |  | А.М. Панфилкин |
|  | (Группа) |  | (Подпись, дата) | (И.О. Фамилия) |
|  |  |  |  |  |
| Преподаватель |  |  |  | М.А. Захаров |
|  |  |  | (Подпись, дата) | (И.О. Фамилия) |

**Задание**

1. Считать набор данных 'data/homicide-data.csv'
2. Вывести информацию про первых 5 пассажиров, всех, кроме последних 2х и с 3 по 15 включительно
3. Проанализировать на предмет наличия особенностей, таких как:
   1. Средний возраст, максимальный и минимальный
   2. Узнать количество женщин и мужчин
   3. Количество женщин, которым больше 18 лет
   4. Количество мужчин, которые "Closed by arrest" (то есть преступник был пойман)
4. Отсортировать всех людей по возрасту по возрастанию и по убыванию и вывести
5. Для каждого штата найти год, где было максимальное число пострадавших. Вывести таблицу

**Выполнение**

Листинг 1 – Выполнение задания

|  |
| --- |
| # Подключение библиотек  import pandas as pd  from IPython.display import display  # Считать набор данных 'data/homicide-data.csv'  homicide\_data = pd.read\_csv('data/homicide-data.csv', encoding='ISO-8859-1')  # Вывести информацию про первых 5 пассажиров, всех, кроме последних 2х и с 3 по 15 включительно  display(homicide\_data.head())  display(homicide\_data[:-2])  display(homicide\_data[3:16])  # Проанализировать на предмет наличия особенностей, таких как:  # 1. Средний возраст, максимальный и минимальный  # 2. Узнать количество женщин и мужчин  # 3. Количество женщин, которым больше 18 лет  # 4. Количество мужчин, которые "Closed by arrest" (то есть преступник был пойман)  # Отсортировать всех людей по возрасту по возрастанию и по убыванию и вывести  # 1. Средний возраст, максимальный и минимальный  # P.S. В колонке victim\_age есть значения "Unknown", которые не надо учитывать  clean\_df = homicide\_data[homicide\_data.victim\_age != 'Unknown']  # В одну строку можно было бы сделать так:  print(f'Средний возраст: {clean\_df.victim\_age.astype(int).mean()}')  print(f'Максимальный возраст: {clean\_df.victim\_age.astype(int).max()}')  print(f'Минимальный возраст: {clean\_df.victim\_age.astype(int).min()}')  # 2. Узнать количество женщин и мужчин (victim\_sex =  / Female)  print(f'Количество женщин: {homicide\_data[homicide\_data.victim\_sex == "Female"].shape[0]}')  print(f'Количество мужчин: {homicide\_data[homicide\_data.victim\_sex == "Male"].shape[0]}')  # 3. Количество женщин, которым больше 18 лет  print(f'Количество женщин, которым больше 18 лет: {clean\_df[(clean\_df.victim\_sex == "Female") & (clean\_df.victim\_age.astype(int) > 18)].shape[0]}')  # 4. Количество мужчин, которые "Closed by arrest" (то есть преступник был пойман)  print(f'Количество мужчин, которые "Closed by arrest": {homicide\_data[(homicide\_data.victim\_sex == "Male") & (homicide\_data.disposition == "Closed by arrest")].shape[0]}')  # Отсортировать всех людей по возрасту по возрастанию и по убыванию и вывести  print('Возраст по возрастанию:')  display(homicide\_data.sort\_values(by=['victim\_age']))  print('Возраст по убыванию:')  display(homicide\_data.sort\_values(by=['victim\_age'], ascending=False))  # Таблица - для каждого штата найти год, где было максимальное число пострадавших  homicide\_data['year'] = homicide\_data['reported\_date'].astype(str).str[:4]  grouped = homicide\_data.groupby(['state', 'year']).size().reset\_index(name='incidents')  max\_incidents = grouped.loc[grouped.groupby('state')['incidents'].idxmax()]  result\_df = max\_incidents.rename(columns={'incidents': 'max\_incidents'})[['state', 'year', 'max\_incidents']]  display(result\_df) |

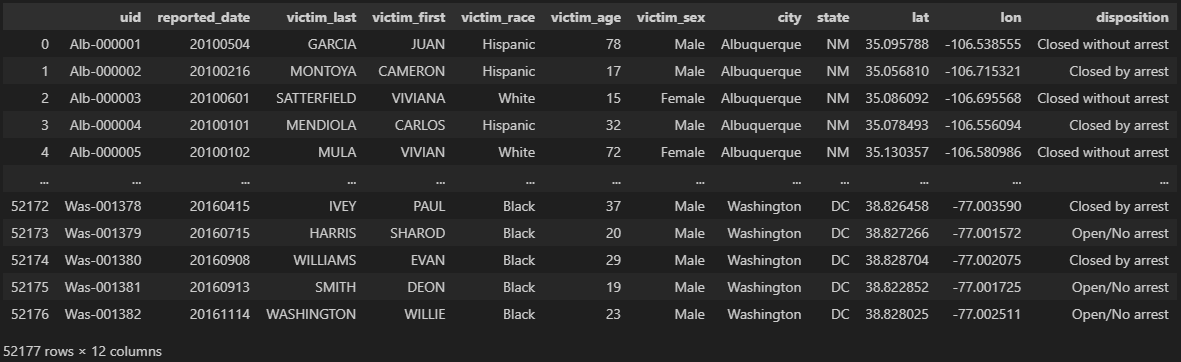
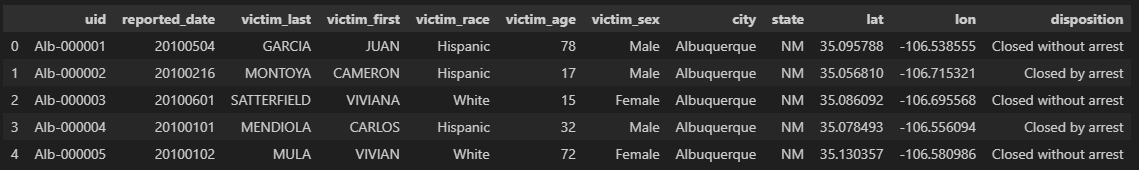


Рисунок 1 – Результат выполнение пункта 2 задания

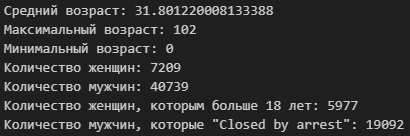


Рисунок 2 – Результат выполнения пункта 3 задания

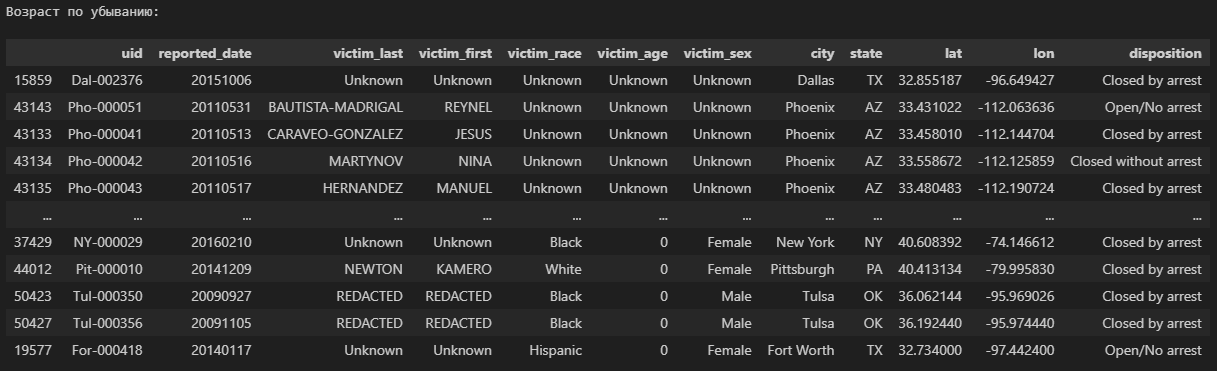
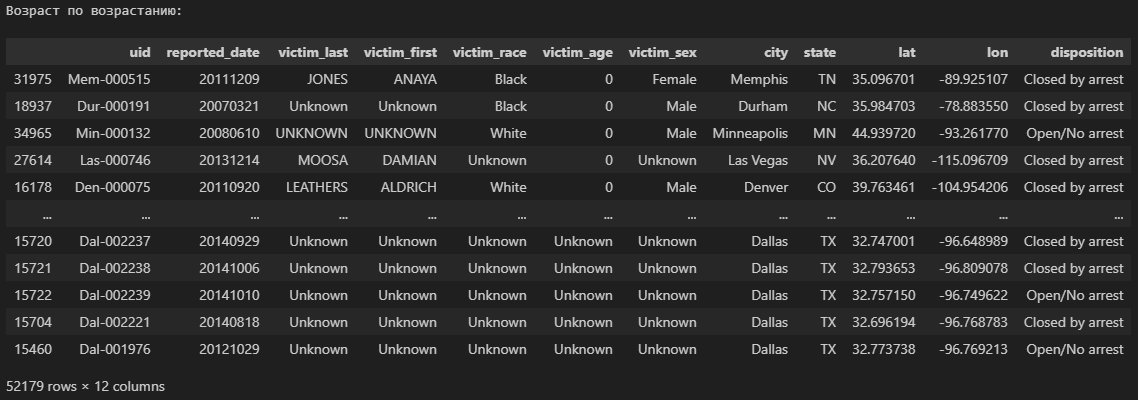


Рисунок 3 – Результат выполнения пункта 4 задания



Рисунок 4 – Результат выполнения пункта 5 задания

**Вывод**

В ходе лабораторной работы мы освоили навыки работы с анализом данных на примере набора 'data/homicide-data.csv'. Полученные навыки включают в себя умение извлекать и выводить информацию о подмножестве данных, проводить базовый статистический анализ (в данном случае, возраст и пол), а также умение сортировать данные и выделять максимальные значения в разрезе групп (например, максимальное число пострадавших в каждом штате).