Homework 4

Due: April 28, 2025

This homework covers use of Pandas, Geopandas, and dictionaries for analysis and understanding of data relating to two applications: (1) Casualties in the sinking of the Titanic (1912), and (2) Types of vehicles involved in traffic accidents in Manhattan. Both applications make use of real-world data, which you should find is quite messy.

Question 1: 20 points.

On April 12, 1912, the RMS Titanic (a mail and passenger vessel) commenced her maiden voyage across the Atlantic, departing from Ireland and headed to New York. Three days later (April 15) the Titanic struck an iceberg and sank. Of the 2,224 passengers and crew aboard, approximately 1,500 died, making it the deadliest sinking of a ship at that time. Wikipedia has a nice writeup on the Titanic and the numerous deficiencies that lead to this disaster.

Problem Statement. In more recent times, the sinking of the Titanic has inspired numerous artistic works, including the 1997 romantic disaster film Titanic (directed by James Cameron). Like many movies, Titanic places an emphasis on romance and drama – a viewer might wonder, how much of this story is true? Who survived, and why? What does the data say?

Data Source. The data file `python-code.d/data/disaster/titanic.csv` contains information on 887 of the passengers and their attributes, including:

```
--- Survived: 1 means passenger survived; 0 for victims.
--- Pclass:   1, 2 and 3 for first, second and third class.
--- Name:     Master/miss first name, family name.
--- Sex:      male or female.
--- Age:      covers the range 0 to 80.
--- Siblings/Spouses Aboard
--- Parents/Children Aboard
--- Fare:     First class (1) tickets are the most expensive.
```

Things to do:

1. Write a Python program that will read `titanic.csv` into a Pandas dataframe.

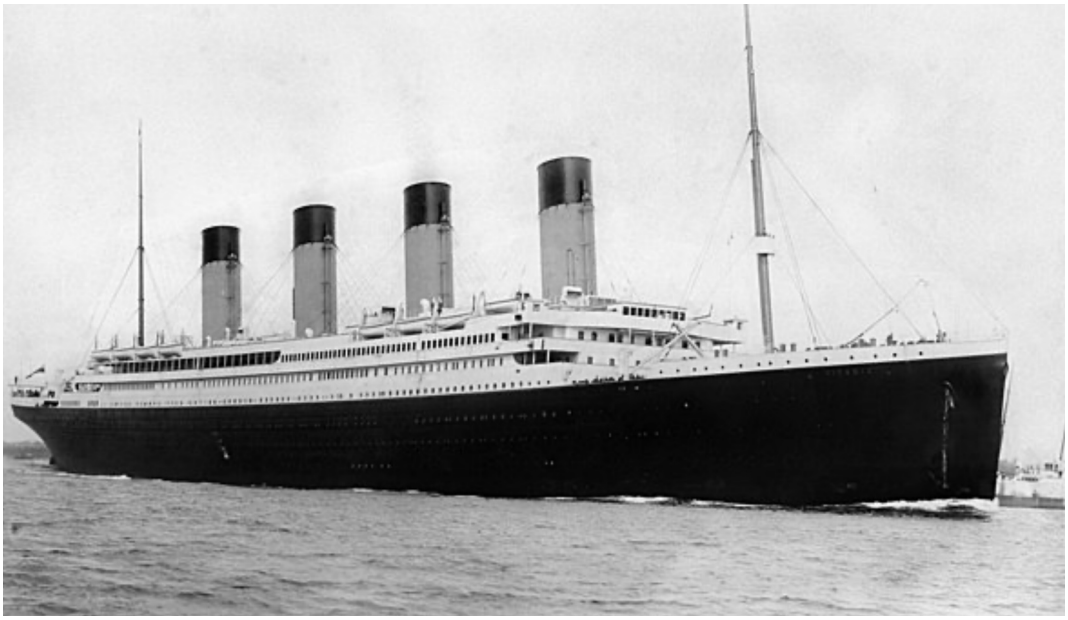

Figure 1: Titanic departing Southampton on April 10, 1912.

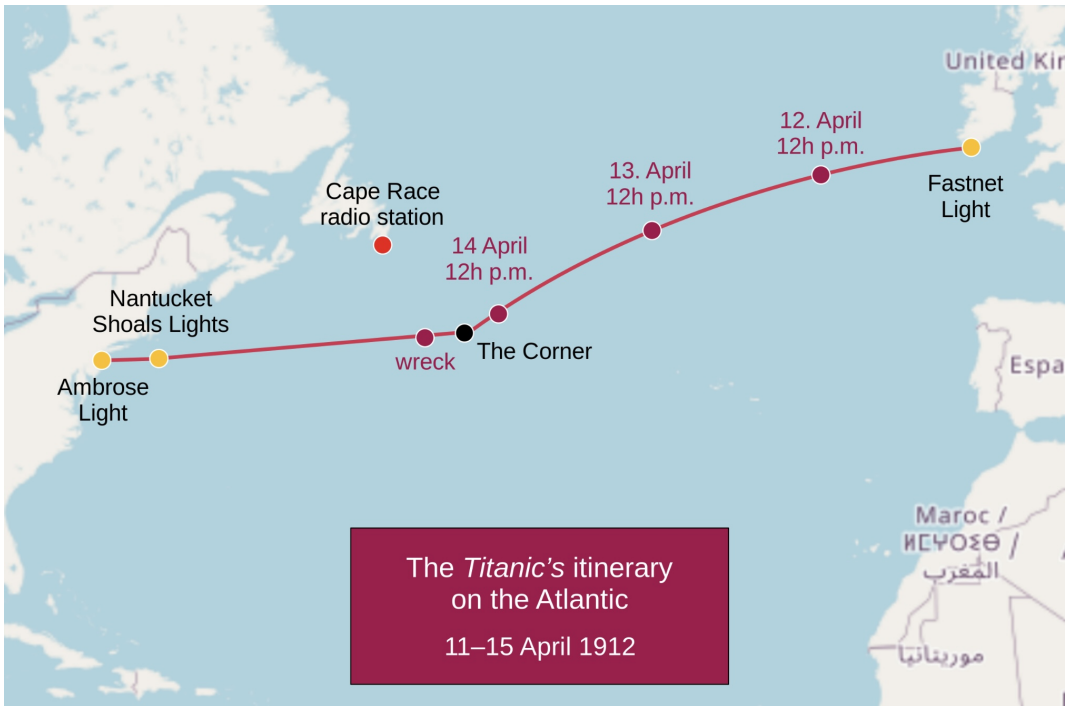

Figure 2: Trans-Atlantic route for the Titanic, from Ireland to New York.

2. Separate the data into two categories: passengers that survived, passengers that drowned. For each category compute the relevant statistics (e.g., how many people, ratio of males and females, number of passengers in each passenger class).

3. Generate histograms for the distribution of age among the survivors and victims.

In Cameron's movie, women and children were given priority to board a lifeboat, and hence survived.

4. Is this part of the story supported by the `titanic.csv` data, or not?

5. Is there any evidence in the data that first class passengers (class 1) were given priority in boarding a lifeboat?

Note: In the early 1900s a child would be someone younger than say 10 or 12 (not 18). And you'd expect children and their mothers would be given access to the lifeboats, regarless of their gender. So, a reasonable strategy is: isolate lists for each of these categories and compute appropriate percentages.

Question 2: 20 points.

Motor vehicle accidents in New York City are the leading cause of death for the city residents. Statistics indicate that close to one in four accidents results in someone being injured or losing their life. The problem is particularly acute for accidents involving high-speed, and/or when accidents involve pedestrians, bicyclists, or motor cycles. Root causes for this situation can be traced back to the city being flat and very walkable, as well as a significant biking culture.

Preliminary Work: During the Spring Semester, 2024, we took a first step toward formally analyzing data on motor vehicle accidents and, specifically, understand **where** and **when** vehicle accidents occur? Figure 3 shows the spatial distribution of 312,000 motor vehicle accidents in Manhattan. The graphic suggests that with the exception of Central Park, accidents occur everywhere. Figure 4 is heatmap of the temporal distribution of accidents. As expected, during the work week, accidents peak during the morning and afternoon rush hours. During the weekend, accidents peak early in the morning, presumably after people have been out socializing and are headed home.

Problem Statement: This question seeks to understand the number and types of vehicle (e.g., taxis, bicycle/bike, e-bike, e-scooter, sport utility vehicle, station wagon, sedan, van, fire truck, pick-up truck, tractor truck, ambulance, bus, tanker), involved in accidents, where and when such accidents occur, and whether or not accidents result in injuries and/or fatalities.

From a medical standpoint it would be interesting to understand whether or not the data contains information on the types of accident (i.e., vehicles involved; time of the day) that lead to fatalities.

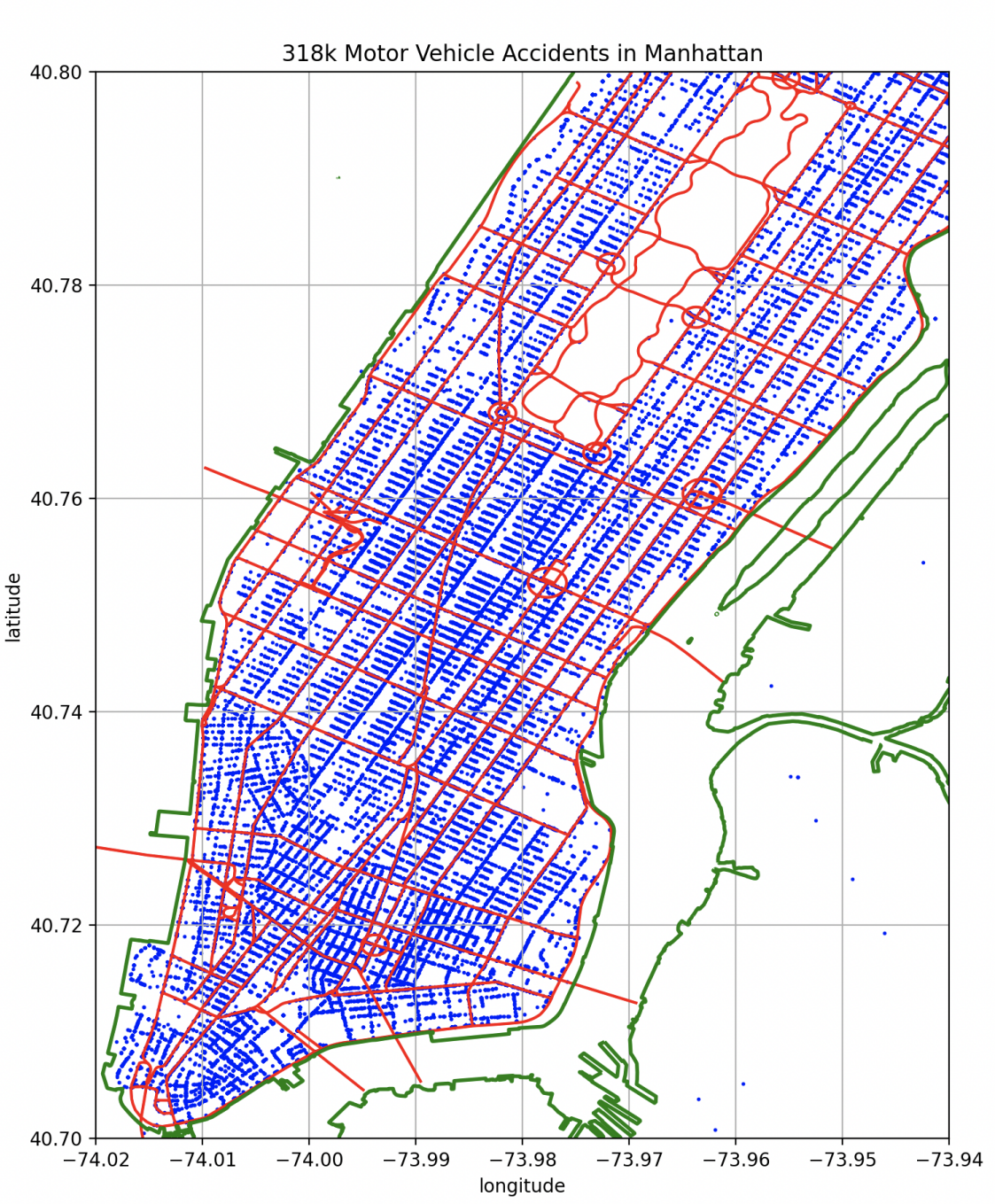

Figure 3: Spatial distribution of motor vehicle accidents in Manhattan.

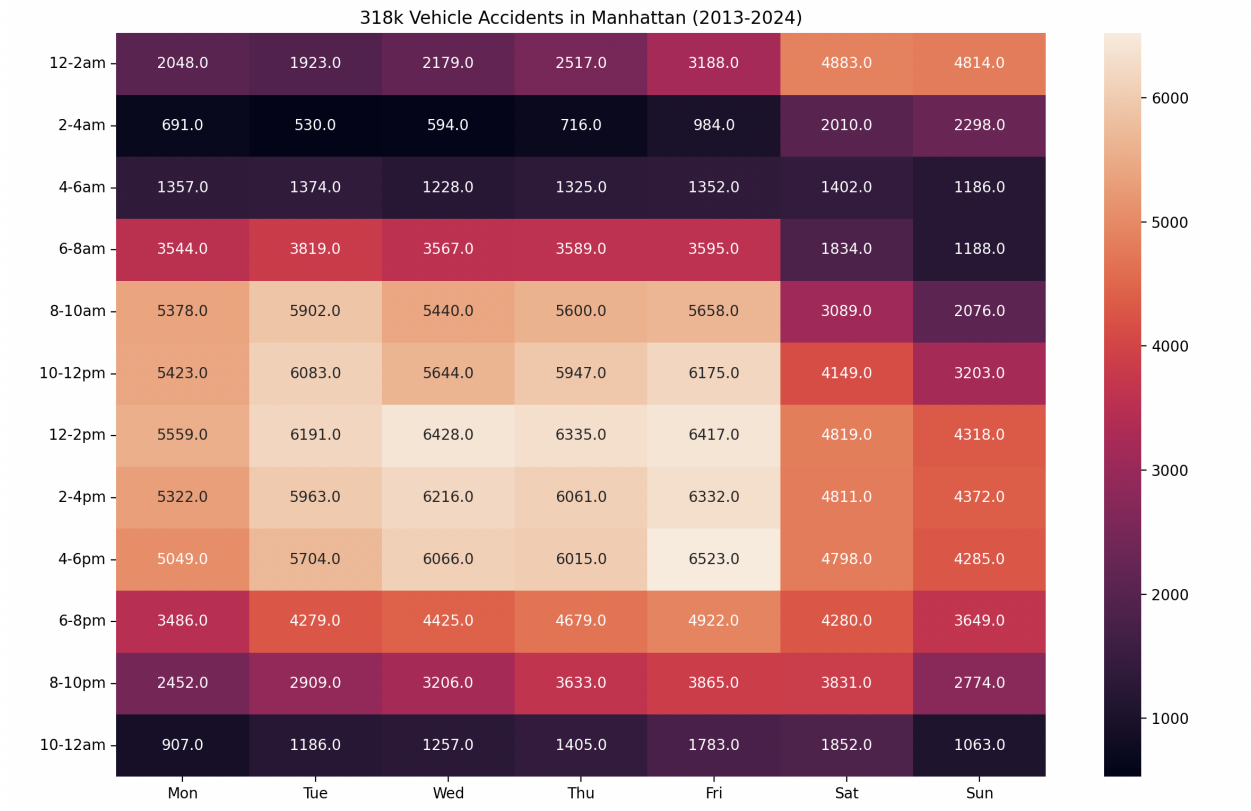

Figure 4: Temporal view of 318k motor vehicle accidents in Manhattan (2013-2024).

Our solution will gather data from multiple sources: (1) geospatial (to draw the city map, major streets, and coastline), (2) accident data.

Data Source 1 – Geospatial: Within the folder `data/cities/nyc` the data files `dcm-nyc-major-street.csv` and `nyc-shoreline.csv` contain geospatial data on the main streets and shoreline in the NYC area.

To run, see: python-code.d/applications/cities/nyc/TestMajorStreetsNYC.py.

Data Source 2 – Accident: The folder `python-code.d/data/cities/nyc/` contains data files that can be used in the analysis of motor vehicle accidents in NYC. The main file, Motor-Vehicle-Collisions.csv, comprises 2.06 million motor vehicle accidents recorded across the five boroughs of NYC and for about a decade.

The data is organized into 29 columns:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2062758 entries, 0 to 2062757
Data columns (total 29 columns):
 #   Column                          Dtype
---  ------                          -----
 0   CRASH DATE                      object
 1   CRASH TIME                      object
 2   BOROUGH                         object
 3   ZIP CODE                        object
 4   LATITUDE                        float64
 5   LONGITUDE                       float64
 6   LOCATION                        object
 7   ON STREET NAME                  object
 8   CROSS STREET NAME               object
 9   OFF STREET NAME                 object
 10  NUMBER OF PERSONS INJURED       float64
 11  NUMBER OF PERSONS KILLED        float64
 12  NUMBER OF PEDESTRIANS INJURED   int64
 13  NUMBER OF PEDESTRIANS KILLED    int64
 14  NUMBER OF CYCLIST INJURED       int64
 15  NUMBER OF CYCLIST KILLED        int64
 16  NUMBER OF MOTORIST INJURED      int64
 17  NUMBER OF MOTORIST KILLED       int64
 18  CONTRIBUTING FACTOR VEHICLE 1   object
 19  CONTRIBUTING FACTOR VEHICLE 2   object
 20  CONTRIBUTING FACTOR VEHICLE 3   object
 21  CONTRIBUTING FACTOR VEHICLE 4   object
 22  CONTRIBUTING FACTOR VEHICLE 5   object
 23  COLLISION_ID                    int64
 24  VEHICLE TYPE CODE 1             object
 25  VEHICLE TYPE CODE 2             object
 26  VEHICLE TYPE CODE 3             object
 27  VEHICLE TYPE CODE 4             object
 28  VEHICLE TYPE CODE 5             object
dtypes: float64(4), int64(7), object(18)
memory usage: 456.4+ MB
```

```
None
(2062758, 29)
```

Columns 10 through 17 store data on injuries/deaths to pedestrians, cyclists and motorists. Factors contributing to an accident are located in columns 18 through 22 (it's NYC, road rage is common). Columns 24 through 28 store the associated vehicle types (could be more than two) involved in the accident.

Things to do:

1. Download the latest version of python-code.zip, unpack, and run the python code that generates Figures 3 and 4. To generate the spatial view, run:

see: python-code.d/applications/cities/nyc/TestMotorVehicleAccidentsNYC01.py.

And to generate the temporal view (heatmap):

see: python-code.d/applications/cities/nyc/TestMotorVehicleAccidentsNYC02.py.

Both programs filter the accident data to only keep accidents occuring in Manhattan – this operation will reduce the number of accidents from 2 million to approximately 318,000. Then, they remove from further consideration accidents that do not have (lat,long) coordinates (c.f., there are about 10,000 of them).

2. The data indicates that approximately 17% of accidents result in injuries; slightly less than 0.1% of accidents result in fatalities. Write a Python program that will gather and print the total number of accidents and injuries/fatalities involving cyclists and pedestrians in Manhattan.

A snippet of the accident statistics might look like:

```
--- Total No Accidents     = xxx ...

--- No Persons Injured     = xxx ...
--- No Cyclists Injured    = xxx ...
--- No Pedestrians Injured = xxx ...
--- No Motorists Injured   = xxx ...

--- No Persons Killed      = xxx ...
--- No Cyclists Killed     = xxx ...
--- No Pedestrians Killed  = xxx ...
--- No Motorists Killed    = xxx ...
```

3. Extend your program to systematally assemble a dictionary (i.e., key-value pairs) of vehicle types and counts of accidents in which they are involved.

A snippet of dictionary content might look like:

```
Vehicle Type                 Accident Count
=========================================
e-scooter                          1,131
e-bike                             1,261
dump truck                         1,287
ambulance                          2,114
motorcycle                         4,054
bicycle                            6,069

  ... many lines of output removed ...

bike                              11,403

  ... more lines of output removed ...

bus                               17,535
van                               21,346
sport utility/station wagon       52,151
station wagon/sport utility vehicle 79,262
taxi                              79,867
sedan                            105,686
passenger vehicle                110,392
=========================================
```

4. From the previous section we see that over the 2013-2024 period, bikes (i.e., bike, bicycle, e-scooter, e-bike) have been involved in more than 17000 accidents in Manhattan alone. And more than 50% of these accidents result in injury and/or death.

Create heatmap views of pedestrian and cyclist injuries/fatalities in Manhattan, 2013-2024.

5. Briefly summarize and discuss your findings.