**DASC 5300/CSE 5300**

**Foundations of Computing**

**Instructor: Sharma Chakravarthy**

**Project I: Python Programming and Data Analysis**

**Project Team -13**

**Aditya Vardhan Marri, Ashik Maraliga Venkatesh**

Python Programming and Data Analysis for NYPD Dataset

In this study, we intend to share our effort, in which we processed, cleaned, and analyzed NYPD data for the years 2006 and 2018. Data analysis is the process of gathering, converting, cleaning, and modelling data to discover the underlying information needed. Data visualization is often used to represent data to find relevant patterns in the data.

Pandas come with support for various file formats and data sources (CSV, Excel, SQL, JSON, parquet, and so on), all of which begin with the prefix read\_\*. The read CSV () method in pandas allows you to read data from a CSV file into a panda's Data Frame. After reading the CSV file, the first step is to clean up the data by converting Nan values to zero. Next, we must use the is null() method to double-check the number of rows that have been replaced with zero.

You can make an unintentional modification, save it, and then discover that all your data is gone or has been modified in ways you didn't intend. The only way to get out of this pain is to save under a different name. You can use the name as it is, but add an a, b, or a 1 or 2 to the end. Then, once you've got the file exactly how you want it, you may delete any earlier efforts to change it. The main point is not to make significant changes to a file unless you have a backup copy.

You may utilize the date column (ARREST\_DATE) in the data frame as an index of type Datetime Index. We filtered the relevant year data using the panda's loc function. Using the seaborn, we plot the graphs that lead to analyzing some of the data with respect to different boroughs in New York. We grouped the year in a new column to better study it and displayed a histogram for the same, which analyses the number of people in each year. We added an additional column to generalize the work-class for Bar chart analysis.

File Descriptions

1) NYPD\_Arrests\_Data\_\_Historic\_.csv is an input file.

2)data – is the raw data frame converted from the input csv file using pandas.

3) data1 is a copy of the initial data frame converted from the csv file. Work with the copy to avoid unnecessary changes to the original file. We will have an original data frame(data) to compare the altered version with this.

4) dataFiltered is the new data frame created after filtering the 2006 and 2018 data using pandas.loc function.

5) dataFilteredSample is the sample we take first to develop the algorithm and test whether the code is working as expected. Once the functionality is satisfied, apply it in the sample space considering 2006 and 2018 data.

6) dataFilteredLocation is a data frame with the records of whose location of arrests is within 3 miles from the given coordinates.

7) m is the map containing all the locations that are less than or equal to three miles from the given coordinates.

Division of Labor

Because it was a new programming language for both of us, it took us time to learn Python, a new language, so for each part of the pin-making process, it takes time to master. If each worker focused on a single task and mastered it, productivity would be more efficient. We decided to get together and work on Python. The entire project was performed in a group setting. We had gone through some important libraries like Pandas, NumPy and maths to work on this project. After studying logic in each phase of the project, we utilized Google, YouTube, and Kaggle to figure out the syntax and analysis the data. This project took us about 70 hours to complete.

PROBLEMS ENCOUNTERED:

1) The first problem we ran into was figuring out how to replace the Different Age\_group more than 65 age with 65+. We have a different age group in the data set that we need to replace with 65+, and we could use the data frame lambda function to do so.

2) The second problem we encountered was plotting the histogram before grouping them into age groups. We tried a few different ways before categorizing the data by adding a new column to the dataset and after that, plotting the histogram was a lot easier.

3) The next issue was that, after obtaining median annual income by borough, we analyzed the total number of offences by borough. However, while the average annual income in Manhattan is higher when compared to other boroughs, there are many crimes. We ran several scenarios and discovered that the average cost of living in Manhattan is quite expensive.

4) The next problem we ran across was converting the object to a datetime index, so we tried a few different approaches. Finally, we have pd.to datetime, which allows us to quickly select dates based on the year we are interested in.

5) The next problem was we attempted a few different approaches to plot a map with markers, but we couldn't get as many points as we wanted within a three-mile radius, so we utilized folium to plot the map, which is a handy tool when working with maps. The marks may be readily plotted on the map.

**ANALYSIS:**

**a)** Analysis of 5 Boroughs of NY city.

Icon

Description automatically generatedChart, icon

Description automatically generated

The data is organized into five groups based on boroughs: B (Bronx), S (Staten Island), K (Brooklyn), M (Manhattan), and Q (Queens) (Queens). Using seaborn to plot the histogram, we can see that K(Brooklyn) has the highest total number of crimes and s (Staten Island) has the lowest total number of crimes in the overall data in 2006 and 2018. And during both the years the ratio of the crimes between the different boroughs remained the same. As a result, we go into detail and study each year, and LAW CAT CD is no exception.

Logo

Description automatically generated with medium confidence

When we look at the above graph by year, we can observe that the number of crimes decreases dramatically with time. In 2006, there were 371934 crimes reported; in 2018, there were 246773. By 2018, the crime rate fell by 33.655 per cent compared to 2006.

Chart, bar chart

Description automatically generated

Manhattan has the highest number of arrests for misdemeanor(M) in both 2006 and 2018, contributing 27.48 and 36.94 per cent of the total misdemeanor cases in 2006 and 2018 respectively, followed by Brooklyn, Bronx, Queens and Staten Island. We can observe that the number of cases reduced significantly between 2006 and 2018. Still, the ratio of cases between the boroughs almost remained the same.

When we look at the boroughs, we can also observe that from 2006 to 2018, all of them had a decline in crime, particularly in Manhattan, where the number of crimes with violations fell by 91.47 per cent. We may deduce from the preceding study that the number of Violations has fallen dramatically over time.

The above plots show that the number of arrests due to misdemeanor is always greater than violations in every borough in both years and the violations do not even makeup 1/10th of the total cases reported in each borough.

By using the information from the below website, we get Median annual income in each borough, New York City (NYC) Median Household Income 2017 Estimates

<https://www.baruch.cuny.edu/nycdata/income-taxes/med_hhold_income.htm>

Median Household income of Bronx(B) is $37,397, Brooklyn (K) is $56,942, Manhattan(M) is $85,071, Queens(Q) is $64,509, Staten Island (S) is $79,201.

From the above data, we can compare the number of crimes under M and V category, including both the years 2006 and 2018.

Chart, bar chart

Description automatically generated

Except for Manhattan, we may deduce from the preceding data that the boroughs with lower annual income have committed more crimes. We looked over all the facts on Manhattan, and we discovered that, despite having a higher yearly income, the cost of living in Manhattan is quite expensive. As a result, the number of crimes in Manhattan is extremely high. Hence The number of crimes is adversely correlated with average income.

Similarly, we investigated other features like which month number of crimes were more and plotted some graphs with respect to the month.

Icon

Description automatically generated Icon

Description automatically generated

From the given graph, the number of crimes was high in March, May and January, whereas October, November, and December appear safer. Based on this data, it's plausible to infer a positive relationship between temperature and crime.

We also looked at the Perp sex attribute and found that male members outnumbered female members in both 2006 and 2018 year. Ratio of arrest remained the same between male and female during year both the years.

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

We can see the ratio of crimes by the different races has seen no change in any of the boroughs when we compare the 2006 and 2018 with black people having major contribution in every boroughs except in the Staten Land.

**b) Age Group**

Chart, histogram

Description automatically generated

After cleaning the data, we divided the age into six groupings, namely 18-24,25-44,45-64,65+,0 and unknown. We plotted the graphs using this grouping and the following observations.

In both 2006 and 2018 the people in the age group of 25-44 had the highest number of arrests due to misdemeanor or violations, followed by people in 18-24 age,45-64 and 65+.

We can also observe that there has been a significant decrease in the total number of crimes in 2018 when compared to 2006. Among the various age groups, the arrests related to misdemeanor(M) alone on people ages 18-24 and 65+ had reduced by more than 45 per cent and 65 per cent by in 2018 compared to 2006. This implies that fewer people aged 65+ and youth were involved in criminal activity (M) in 2018.

While arrests on people in the age group of 45-64 has shown the lowest decrease in crimes by 2018 when compared to 2006 which is either in misdemeanor or violations with just 19.2 percent.

When it comes to violations, there were 19950 instances in 2006 but only 3017 in 2018, implying that there were far fewer cases arrested for violations in 2018.

We can see from the graph above that the age group 25-44 has a lot of crimes in both misdemeanor(M) and violations(V). Let us take a deeper view by considering the boroughs and their contribution to people arrested from age group 24-44.

Chart, bar chart

Description automatically generated

The first observation in the number of violations is drastically less compared to the misdemeanors and they do not even amount to 1/8th of the misdemeanors in every borough in 2006 and 2018.

Although the Misdemeanors has reduced a lot by 2018 when compared to 2006. Although the boroughs of Manhattan and Brooklyn have shown a significant decrease, Staten Island had remained stagnant in its numbers compared to both the years.

Manhattan contributed more violations in 2006 followed by K(Brooklyn), whereas K(Brooklyn) contributed more in 2018 followed by Manhattan. Whereas the remaining order remained the same in both the years with Queens in third place and followed by the Bronx and Staten Island.

From 2006 to 2018, we may deduce a significant reduction in the number of violations and misdemeanors committed in Manhattan while Staten Island almost remained stagnant.

**C) Randomly pick a location (i.e., latitude and longitude) for each Borough:**

We used a python library called folium, which is best for mapping markers and so many.

First, we applied the distance formula to calculate the distance between two given points. We calculate the distance between given coordinates and all the coordinates in the data frame using for loop and convert the value into miles.

We created a new column called “distance in miles” and we added the distance, we filtered data points which are less than or equal to 3 miles radius. Using folium, which is an interactive leaflet map, folium makes it simple to see data that has been edited in Python. It allows data to be bound to a map as well as the binding of data to a map. We have chosen a random place and plotted all the locations within 3 miles of it.

Map

Description automatically generated

Above points are the various locations where crime happened within miles of given location, given location latitude is 40.84 and longitude is -73.911

**References**

We looked at several websites and watched a lot of YouTube videos to better comprehend the data and analysis, and we have included the sources below.

1)<https://stackoverflow.com/questions/9744775/how-to-convert-integer-timestamp-to-python-datetime>

2)<https://www.analyticsvidhya.com/blog/2020/03/what-are-lambda-functions-inpython/#:~:text=We%20can%20do%20this%20with%20the%20apply()%20function%20in%20Pandas.&text=We%20can%20use%20the%20apply,gets%20applied%20to%20each%20row>.

3)<https://seaborn.pydata.org/generated/seaborn.catplot.html#:~:text=catplot,-seaborn.&text=Figure%2Dlevel%20interface%20for%20drawing,one%20of%20several%20visual%20representations>.

4) <https://www.kaggle.com/code>

5) <https://www.youtube.com/watch?v=t9Ed5QyO7qY>

6) [User Guide — pandas 1.4.1 documentation (pydata.org)](https://pandas.pydata.org/docs/user_guide/index.html#user-guide)