



Make traffic safer in New York City

12.10.2021

Big Data - Group 26

Yiming Yang yy3797

Sichao Xu sx2101

Meng Zhou mz3043

Introduction

New York City is an international metropolis with great influence on the global economy, commerce, finance, media, politics, education, and entertainment. With over 20 million people in its metropolitan statistical area and 23,582,649 in its combined statistical area as of 2020, New York is one of the world's most populous megacities[1]. According to U.S. News and World Report, NYC ranks 4th for the worst traffic in the nation[2]. Since the heavily congested roads in a densely populated area result in a lot of traffic accidents. Our project is to use big data techniques to analyze the collisions in the five boroughs of NYC and then figure out any applicable method that could reduce the incidence of car accidents.

Problem formulation

Complain from commuters in NYC

The investigation came from New Yorkers' complaints about commuting. According to a survey by realtors.com, New York, with an average one-way commute of almost 36 minutes and an even more dreadful average of 89.4 hours spent in congestion every year by the typical New Yorker. The high accident rate has aggravated urban congestion. We need to reduce accidents from the source.

The safety of New Yorkers are at risk

Traffic accidents threaten the lives of citizens. Especially in Rush hours, a traffic accident will cause a chain reaction, causing many vehicles to collide, affecting the safety of commuters. We will analyze the injury situation distribution and produce recommendations to reduce the number of casualties.

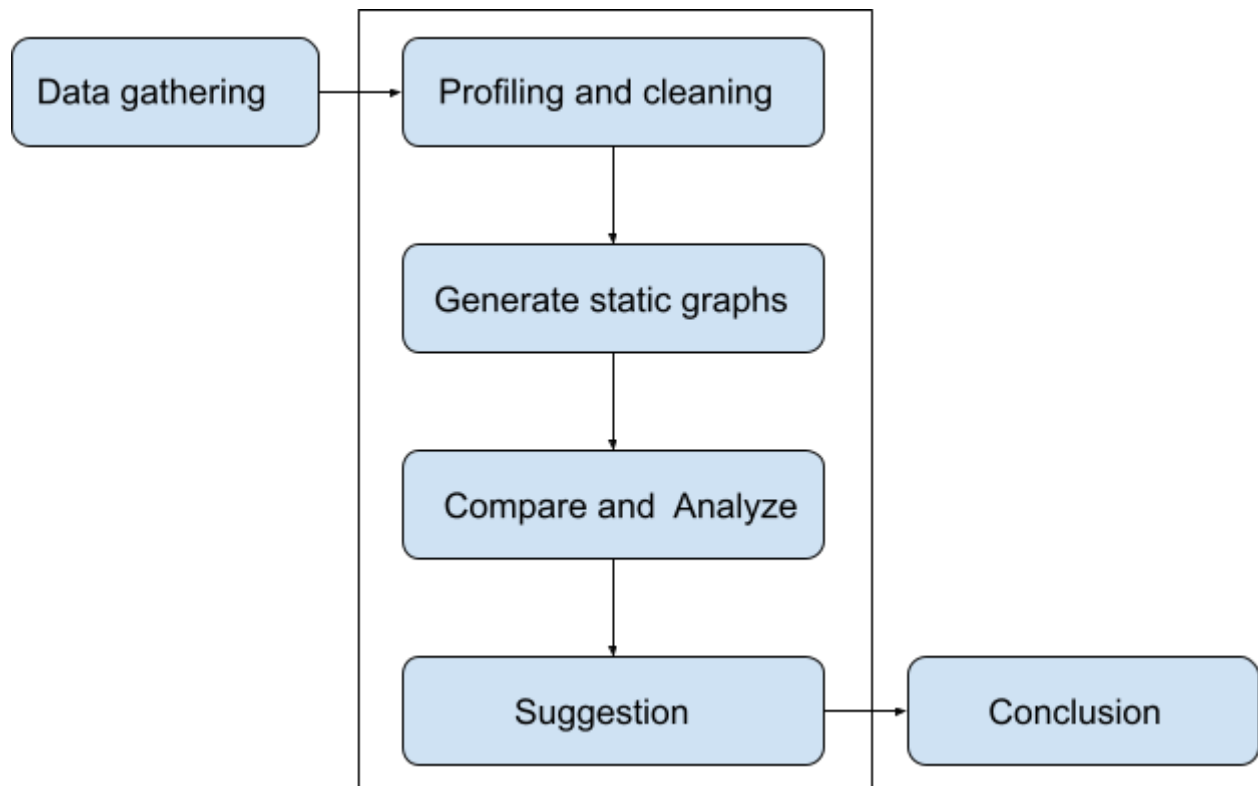
Subjective causes of traffic accidents

Besides the objective causes described above, we also need to investigate the subjective factor which directly causes the car collision. For example, while the driver is playing with the cell phone while driving, his/her attention is distracted and causes a car collision. We need to find out the most important factors and find ways to avoid them.

Worst and Best Commutes



Related work procedure



Methods, architecture, and design

I. Data Profiling and Cleaning

Data Profiling refers to the activity of creating small but informative summaries of a database. Data cleaning is the process of detecting and repairing corrupt or inaccurate records from a data set in order to improve the quality of data. In our case, the dataset is about motor vehicle collisions in NYC. Our goal is to have a cleaned dataset with validity, accuracy, completeness, and consistency.

Cleaning Strategy

There are 29 columns in the dataset. COLLISION_ID: This is the primary key of the dataset, we don't need to make any changes.

There is a lot of information in the accident report that may be manually filled in by the police or the person involved in the accident, and there will be a lot of information that is inconsistent or not filled in. For the rest 28 columns in the dataset, we justify the date to date data type and time to time data type. In the street name column, we limited the type to number, character, comma, pound, and period. The last step is to remove all invalid data including unaccepted data types and blank cells.

There are 1838945 rows in the original dataset and 1263421 rows after cleaning.

II. Improvement

To improve our cleaning strategy, instead of deleting all the empty rows, we re-analyzed each column of the dataset. Some column data can be directly retained. Although some columns have missing data, we can use other columns to calculate an approximate value of this blank. In this way, we can greatly reduce the deleted data and provide more effective samples for subsequent data analysis.

In the original dataset, we convert the address under the naming convention and fill empty cells with 'Unspecified' as another type of address.

In all date and time fields, column name: CRASH DATE, CRASH TIME

For Example,

Address: Street: 181 East street City: New York State: NY Zipcode: 10010

Date: 2021-12-09 (YYYY-MM-DD) Time: 16:44:00 (hh:mm:ss)


With columns that should be filled with non-negative numbers, we fill all the empty space with zero and set the data type to Int. For example, NUMBER OF PEDESTRIANS INJURED, NUMBER OF CYCLIST INJURED, and NUMBER OF MOTORIST KILLED.

We found 10 datasets, some of the data in this dataset are overlapping with the original dataset. We use these data to verify our data cleaning method effectiveness.

In datasets Automated_Traffic_Volume_Counts and Traffic-Signal-and-All-Way-Stop-Study, the column BOROUGH has some missing values. We fill 'Unspecified' as a new BOROUGH category.

In the dataset of Mobile-Telecommunications-Franchise-Pole-Reservation, it has BOROUGH and ZIPCODE. If the dataset is missing, we match the zip code with the BOROUGH name and fill in the empty blanks.

In the dataset of Motor-Vehicle-Collisions-Person and Motor-Vehicle-Collisions-Vehicles, the overlapping data are the CONTRIBUTING_FACTORS, we use KNN to find the reason that causes the collision. KNN is a trained model of machine learning, it will find the most similar reason which matches the column.



In the Vehicle-Classification-Counts dataset, we use a different attribute vehicle class type to find the most similar vehicle type.

In the dataset of Parking-Violations-Issued-Fiscal-Year-2022, the column to be cleaned is Vehicle Expiration Date. We convert the invalid data type to DateTime in the same way on the original dataset.

Comparison

First of all, after the improvement, our data integrity has been greatly improved. We used machine learning algorithms KNN to fill in many blank areas. We usually delete an entire column of missing data directly in the first part.

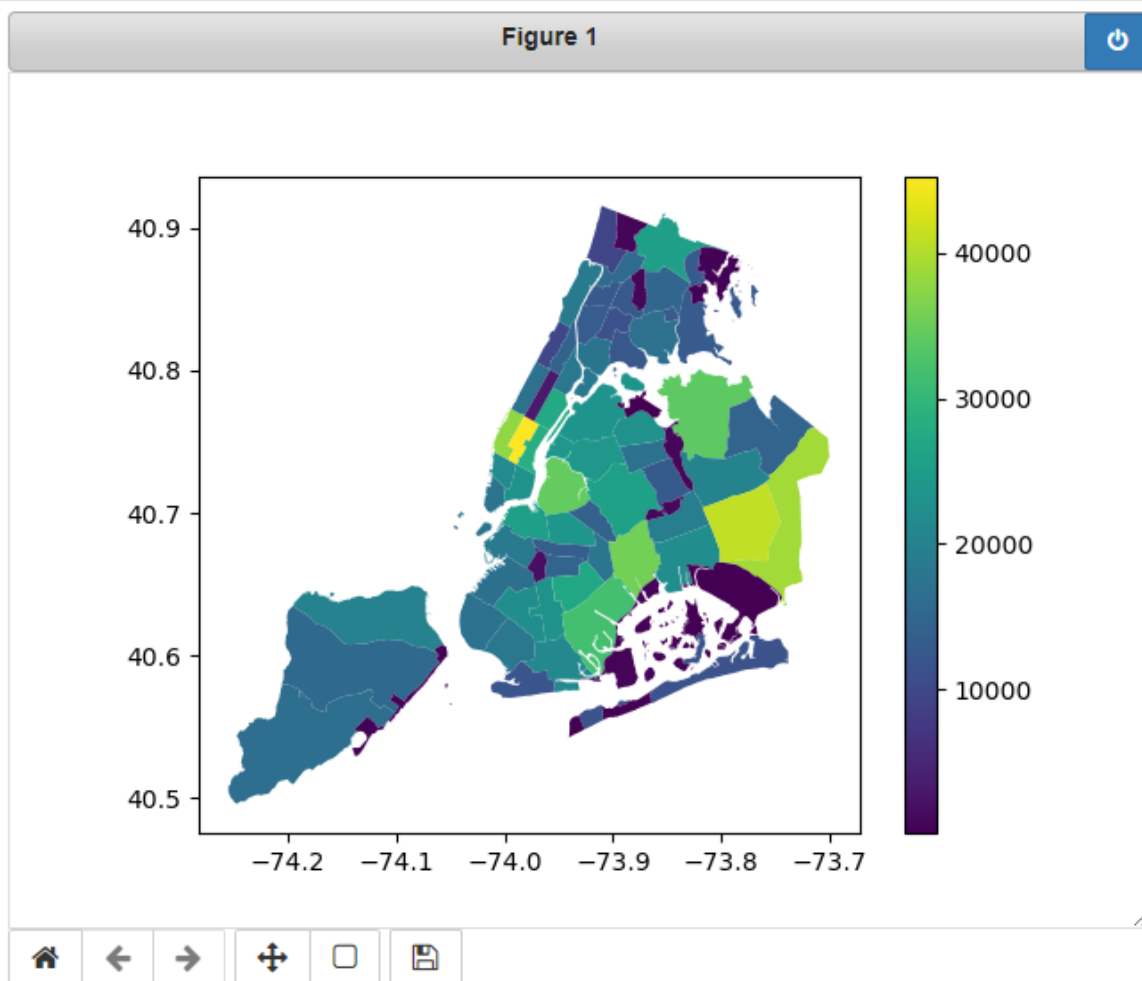
Second, because the zip codes and regions in New York City are very densely arranged, we can get more accurate latitude and longitude locations by matching the zip codes in Borough. The accuracy of the data has also been greatly improved.

Results

Visualization and Analysis

Figure 1: Case number by district

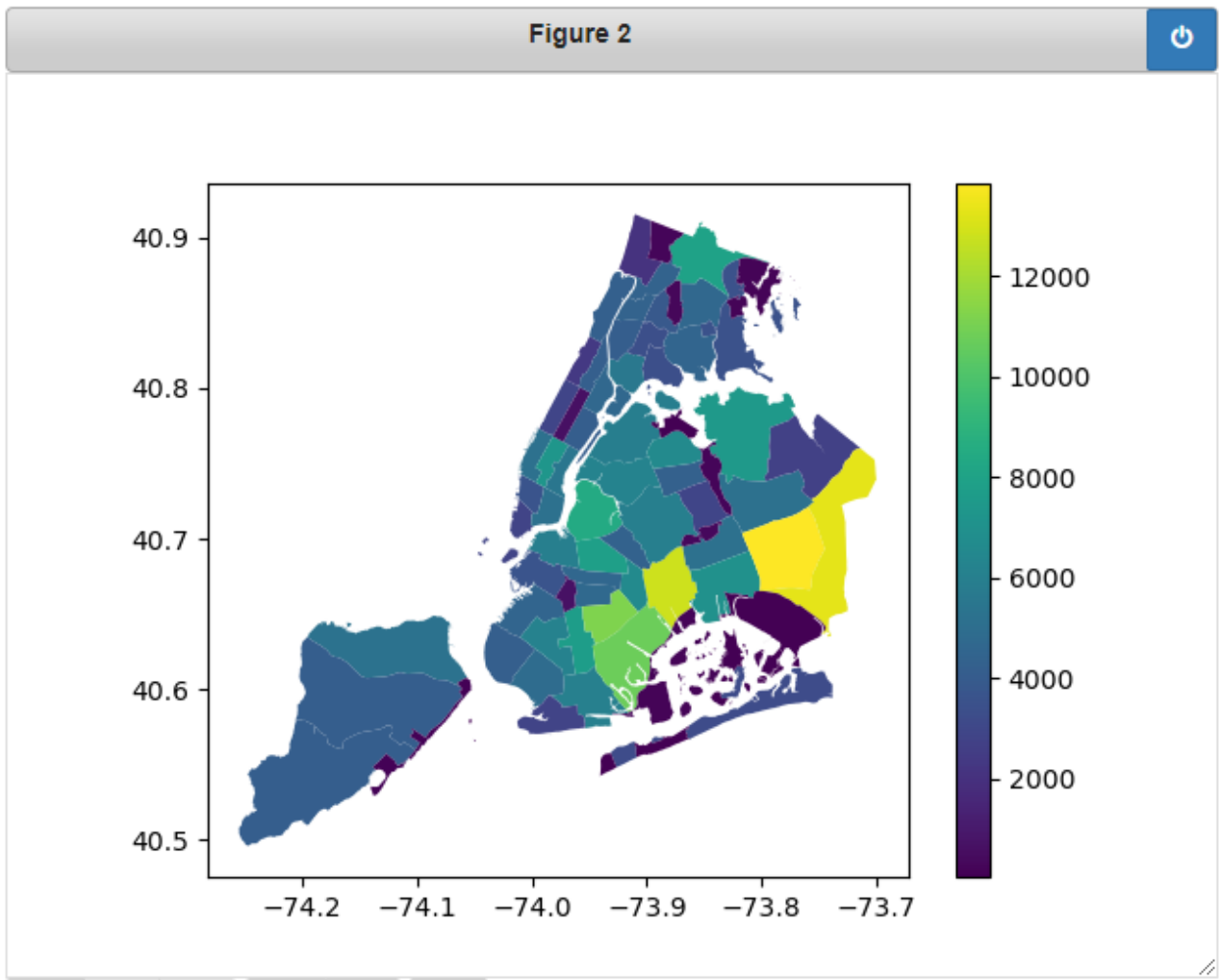
```
In [32]: mv_collision_geo.plot(column='case_num', legend=True)
```



```
Out[32]: <AxesSubplot:>
```

Figure 2: Injured number by district

```
mv_collision_injured_geo.plot(column='injured_num', legend=True)
```



Comparing Figures 1 and 2, it can be concluded that the case number is lower in the urban and densely populated areas, whereas in the suburbs with lower population density, the injured num is higher than the case number. The possible related factors: low-speed limits in areas with high urban population density and not so low-speed limits in suburban areas, and better transportation facilities in urban areas.

Figure 3: Killed number by district

```
mv_collision_injured_geo.plot(column='killed_num', legend=True)
```

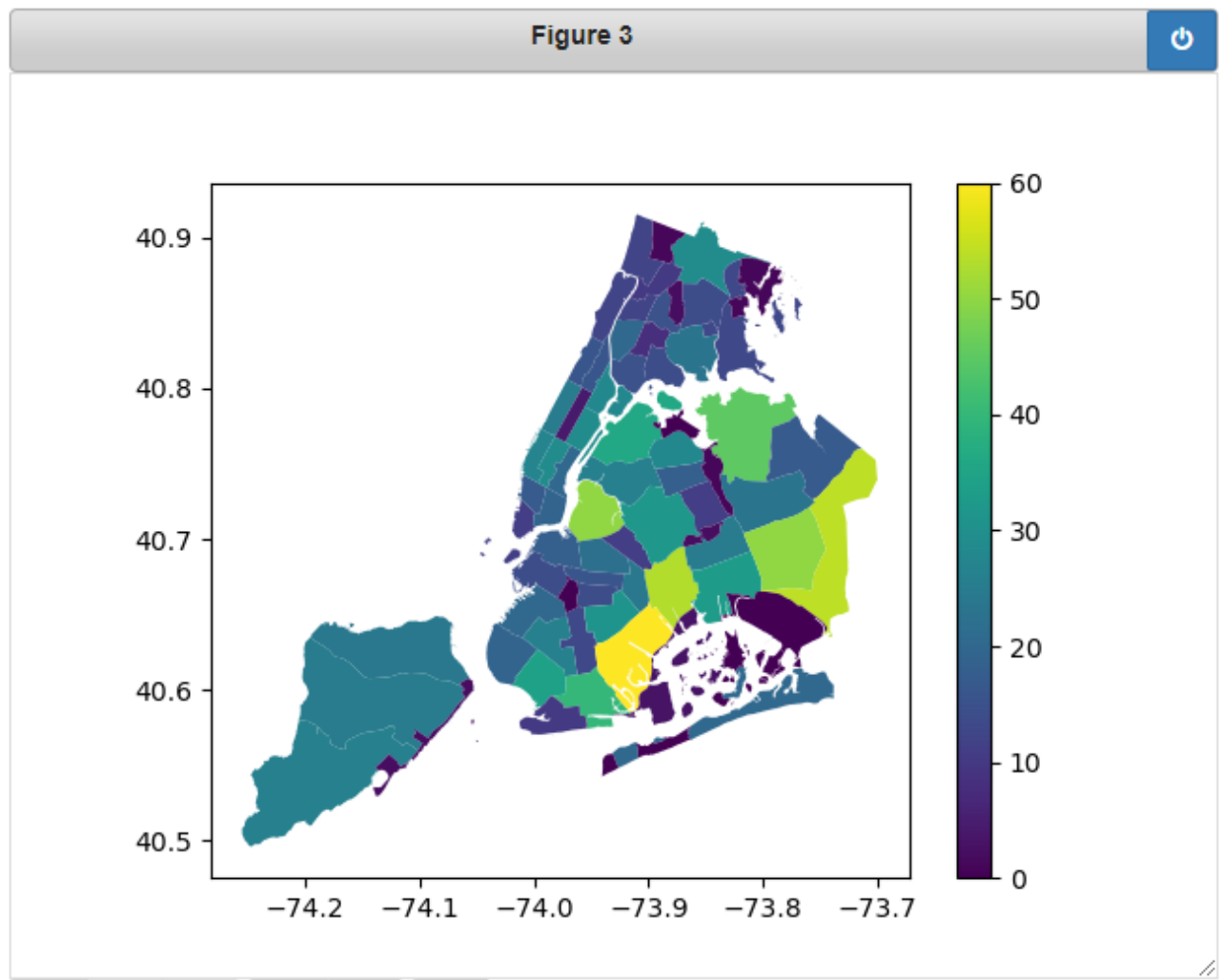
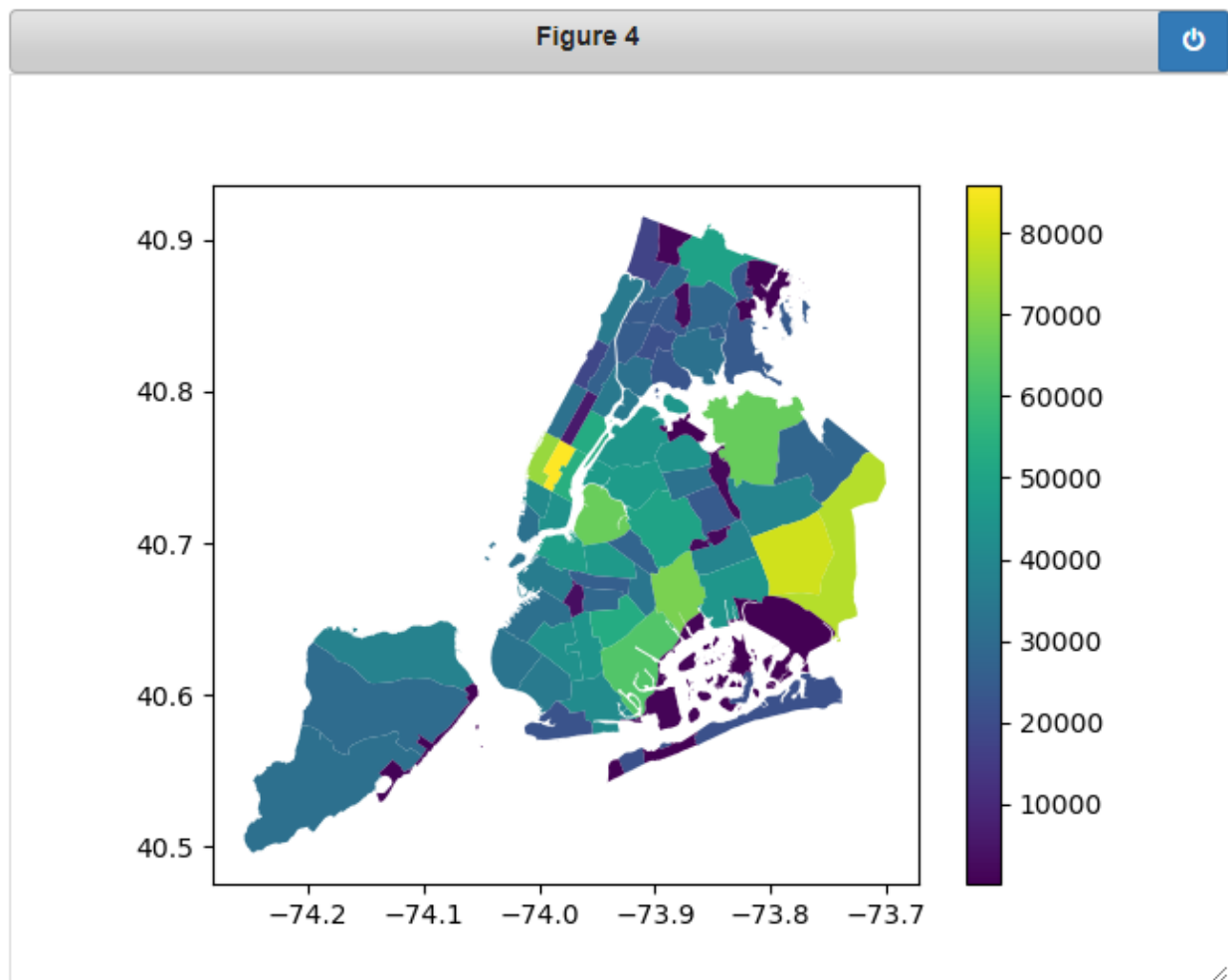


Figure 3 and the first two figures and their conclusions can be drawn that the proportion of deaths in urban areas is lower than that in suburban areas. On the one hand, suburbs are larger in area, with larger case numbers, and more injured numbers and killed numbers. On the other hand, it is closely related to vehicle speed and transportation facilities.

Figure 4: Evolved vehicle number by district

```
mv_collision_injured_geo.plot(column='vehicle_num', legend=True)
```



From Figure 4, even if the land area in urban areas is small, the proportion of land area occupied by vehicles is comparable to that in suburban areas. Therefore, it can be concluded that there are more vehicles in urban areas, which are more likely to have traffic, so the speed is lower. Even if the case number is high, the severity of the case is very likely to be less than the severity of the traffic accident in the suburbs.

Figure 5: Collision case number per Month

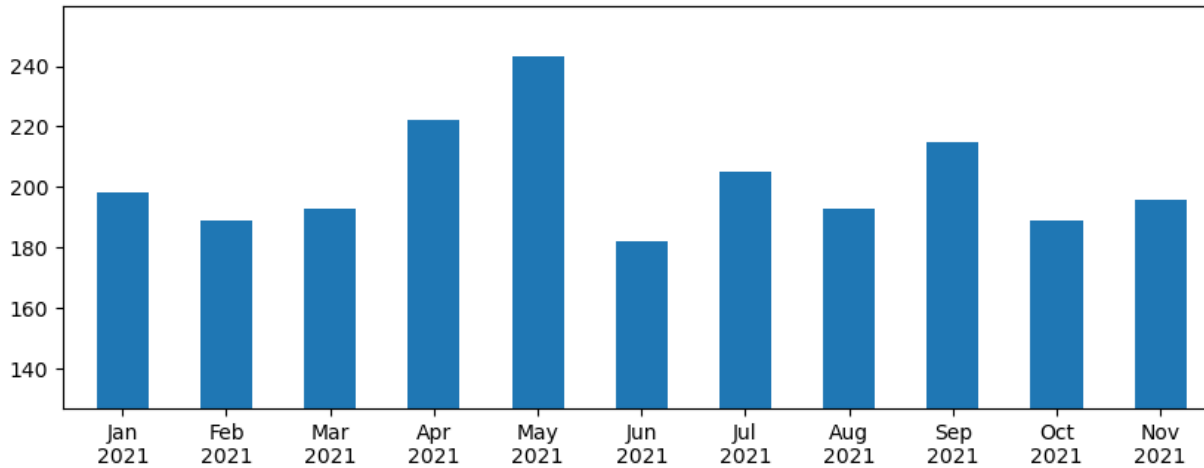
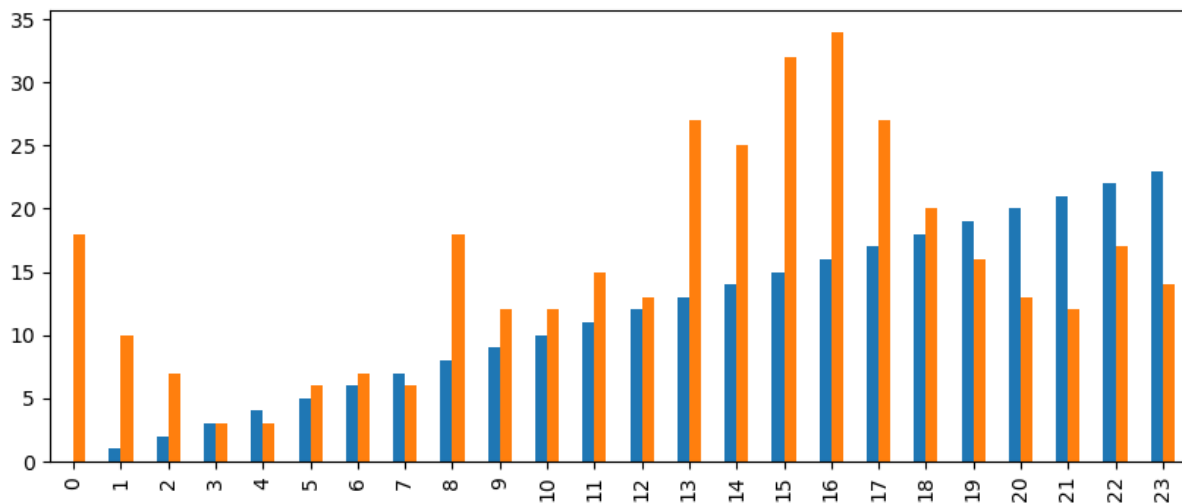


Figure 6: Collision case number per Hour



At around 12 o'clock in the evening, the drivers on the road drove fast, and there was a small peak of accidents. The highest accident rate is in the afternoon when children are out of school and office workers are off work.

Suggestion

1. Concentrate when driving, don't do distracting things like playing with mobile phones, eating, etc.
2. If you can take the bus and subway, try not to drive as much as possible to reduce congestion and avoid traffic accidents
3. Choose car models with a higher safety rate
4. Avoid driving during rush hours when accidents rates are high
5. Maintain safety distance between cars on congested roads to avoid chain accidents, especially in urban areas.
6. Do not drive too fast in the suburbs, which can effectively reduce the severity of the accident.

Conclusion

In the process of optimizing data cleaning, we realized that data cleaning is a very time-consuming process. We have to think about our cleaning goals and what purpose the data is used for. Analyze the format of each column and various constraints. Only then can we work out a more matching way to efficiently complete data cleaning.

In the process of data analysis and visualization, we have a deeper understanding of the importance of data profiling and cleaning. Only valid data can make a meaningful visualization graph, and the data analysis carried out could be meaningful.

References

[1]. Wikipedia_New York City https://en.wikipedia.org/wiki/New_York_City

[2] <https://jknylaw.com/new-york-car-accident-lawyer/statistics/>

<https://www.nytimes.com/2018/02/22/realestate/commuting-best-worst-cities.html>

<https://opendata.cityofnewyork.us/>