Smarter Transit: An analytic for building good life for New Yorker

Pin-Yi Chiu, Nini Lin, An Lee, Hong-Ren Mao
New York University Realtime and Big Data Analytics

ABSTRACT

This report revealed the correlation of weather, transit, and crime in New York City, built on a panel of daily crime, taxi trip records, subway arrival time, and temperature, using the NYU peel cluster, Hive and Tableau as the platform. We identify the effect of weather on daily, seasonally crime by using the temperature and the number of crimes from 2006 to 2021. Looking into the relationship between transit and crime in New York City.

Keywords

Analytics, big data, crime, weather, taxi, subway, New York, correlation

I. INTRODUCTION

New York is one of the biggest and most dynamic cities in the US. Lots of people living in New York City rely on public transit and do not own a car. Taxi services also play an essential role in transporting travelers. While enjoying the convenience and availability of the urban-transport systems, residents are also aware of the safety aspects of urban transportation.

A new Quinnipiac University poll shows that nearly 65% of residents worry about becoming the next victim of a crime. [1] Crime and traffic are long-term issues, discarding that the government has invested lots of effort to improve, that individuals cannot reduce their vigilance.

Therefore, our analysis combined the crime, transportation, and weather data to target dangerous moments and places, providing more information and warning for residents in New York City. In the paragraph on data sources, we list the dataset and describe each dataset's application and data size. The design and implementation section shows how we clean up and extract relevant elements to perform the study of the subject.

At the end of the paper, we separately conclude the multiple correlations among data sources and further introduce in-depth analyses within the scope.

II. MOTIVATION

New York City is a big city - a world city with a population from all over the world. However, NYC has a reputation for crime, from pickpocketing to absolute violence. Moreover, the recent subway attack in Brooklyn has raised more concerns about public transportation, despite efforts by authorities to reduce crime.

Therefore, we decided to take a deeper look at the correlation between crime and transit, expecting to provide information for people to understand whether NYC transportation is safe.

In addition, we consider weather conditions as unavoidable factors in this analysis, as studies have shown that climate has considerably impacted criminal behavior and transit ridership.

III. DATA SOURCES

1. NYPD Complaint Data Historic [2]

Historic data includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD). This dataset can be used to explore the nature of the criminal activity and even combine weather and NYC traffic data to observe their correlations with patterns of criminal activity.

The size of historical data is 2.4GB and updates annually, which covers reported crimes from 2006 to the end of 2020.

2. NYPD Complaint Data Current [3]

To complement the possibility of missing data in historic complaint data, we also collect the current dataset, which also includes all valid felony, misdemeanor, and violation reported crime, to enrich and complete the data sample even though its size is only 175 MB. The current dataset updates quarterly and covers from 2006 to the end of 2021.

3. TLC Trip Record Data [4]

Monthly issued report containing four different taxi categories, including the time and location for pickup and drop-off of each trip. Since the data quality of the yellow taxi is the best among these four, we choose yellow taxi data to conduct the analysis. In selected time periods from January 2019 to July 2021 the yellow taxi trip records' dataset size is approximately 11.65 GB.

4. Real-time subway feed from the MTA [5]

The MTA has some archived subway arrival time data, but we discovered that it is very incomplete and difficult to analyze. Therefore, we chose to build our own dataset by first accumulating data from the MTA's real-time subway feed API. After collecting API responses every 30 seconds during 4/14/2022 (Thu) 21:40 - 4/19/2022 (Tue) 15:51 for the 123456BDFMNQRWJZ subway lines, we accumulated approximately 41GBs of data.

5. NOAA public daily weather [6]

The essential data that connected both criminal rate and traffic condition. The USW weather dataset size is 2.2GB from NOAA (National Oceanic and Atmospheric Administration).

We filter out the daily information from the 1900s to now by region in NYC separately and contain weather conditions, including five core elements temperatures (minimum and maximum), snow condition (snowfall and depth), and precipitation.

Moreover, to find out the correlation among daylight, traffic and criminal rate, we will use the sunrise, sunset data from Sunrise and sunset times in New York City to provide more diverse perspectives.

V. DESIGN AND IMPLEMENTATION

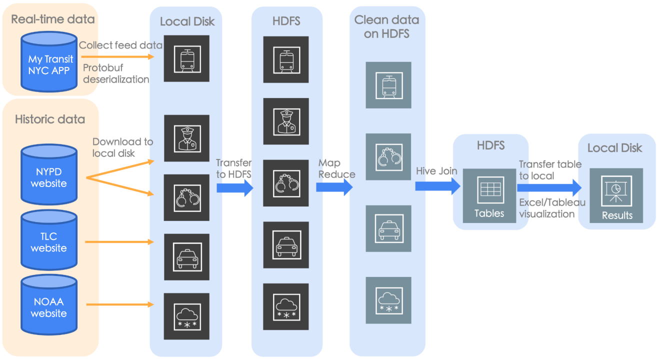

Fig 1: Design diagram of the analytics process.

First, we downloaded the historical data files to our local disk from TLC, NYPD, and NOAA websites. The next step is to use MapReduce to perform data cleansing and profiling.

For subway data, we wrote a Python script to accumulate API responses from the MTA's real time subway feed. Since the subway feed is serialized using Protobuf, we deserialized API responses before storing them to reduce data processing complexity further down the road.

As for crime data, we designed MapReduce jobs with multiple inputs to merge files with inconsistent column order. Also, we wrote customized logic in the mapper and reducer since the value of some columns has special delimiters, which would split columns incorrectly if directly using the built-in function - split. Then, using job chaining to merge the clean data into a single output.

The daily weather data from NOAA is complete and clean. We use a single mapper to transfer its format and filter out the related weather elements that we need. The challenge of weather data is to design a more readable format for others to use and recognize related regions and weather elements.

Since the taxi data is relatively clean, we can simply use a single mapper to filter out the inaccurate data, and retrieve the columns that are of interest. The detailed MapReduce description can refer to attachment A.

After obtaining the organized data, we used Hive to create table and join tables intra and inter datasets, the Hive coding section can refer to attachment C. Finally, we downloaded the table to local disks and used Excel/Tableau for data visualization.

VI. RESULTS

The three graphs below show the relation between weather conditions and crime counts. In figure 2-1, no matter which borough, the number of crimes is frequently higher in the thirst quarter, which begins in July and ends in September; on the other hand, the relatively lower number of crimes period is usually in the first or fourth quarter, that is the winter season.

To explore the correlation between temperature and crime in-depth, we observe that, in figure 2-3, the average maximum temperature from 2006 to 2021 is 17.3℃ (degrees Celsius). When the temperature is above the average maximum temperature, the number of crimes is generally higher, represented by the darker red. Otherwise, the red is lighter, meaning the number of crimes is relatively lower.

In the 2012 research, Crime, Weather, and Climate Change [7], also shows that, in most crime categories, high frequency varies in crime rates due to seasonality.

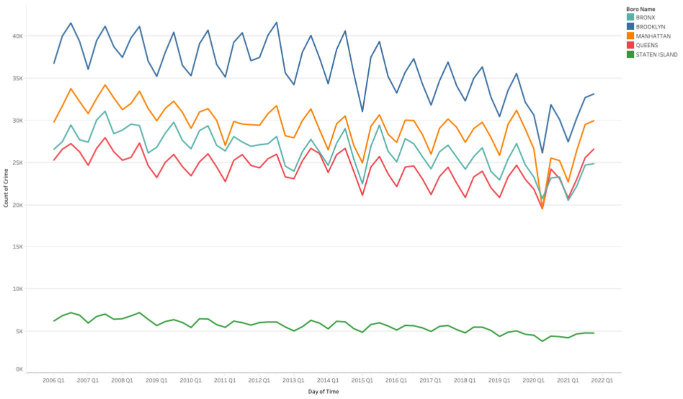

Fig. 2-1 Crime trend of five boroughs

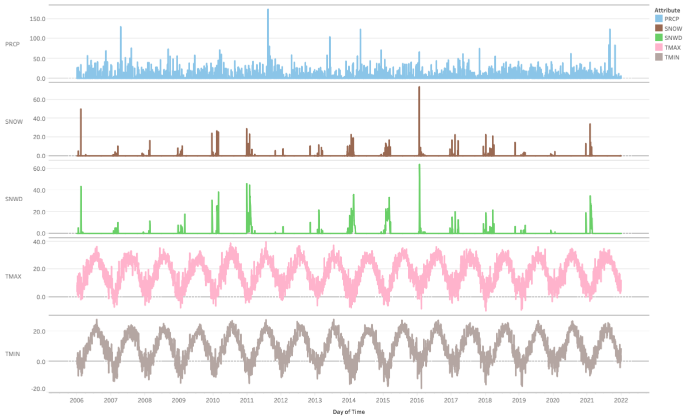

Fig. 2-2 Weather of the day

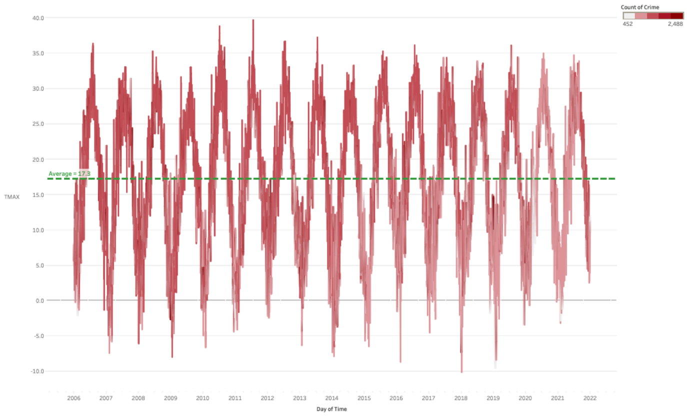

Fig. 2-3 Correlation of Max Temperature and Crime

In the mind of almost every New Yorker, people consider the subway more dangerous at night than in the daytime. According to figure 3-1, surprisingly, the peak of the transit crime is actually in the morning and evening rush hours, 6:30 a.m. to 9:30 a.m. and 3:30 p.m. to 8:00 p.m., as defined by the MTA. Additionally, most offenses type that occurred in the subway is misdemeanor, typically a crime punishable by less than 12 months in jail, such as petit larceny.

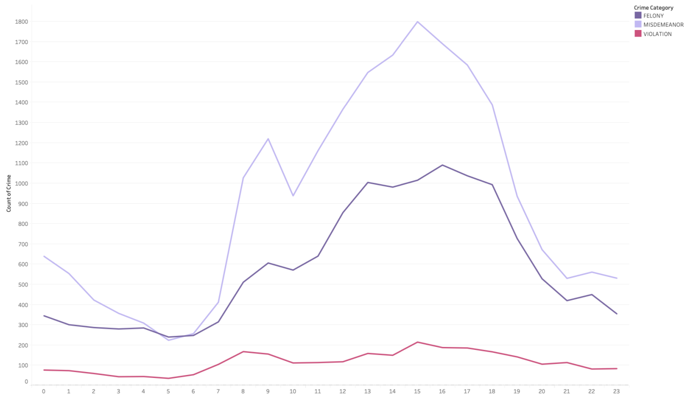

Fig. 3-1 Crime of the day occuring in NYC subway

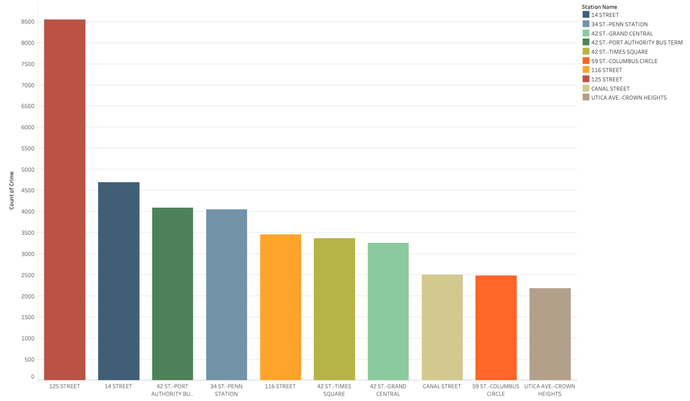

Fig. 3-2 Top 10 high-crime stations of the year in NYC

Below is the analysis done on our subway arrival time dataset. As we can see, arrival time prediction error is within 30 seconds. Since subway trains in NYC actually have set timetables [8], it makes sense that arrival time predictions would be accurate if trains adhere to their schedules. According to a subway train operator [9], the MTA does indeed push for the schedules to be accurate.

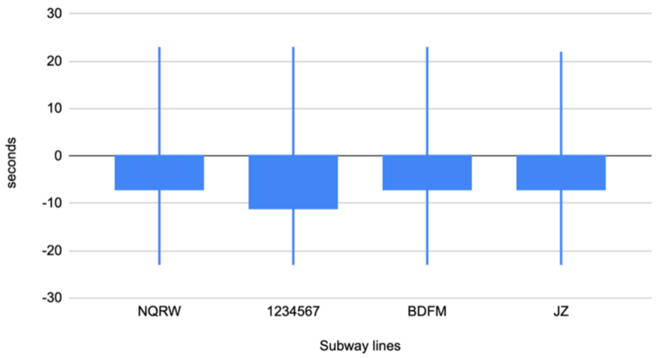

Fig. 4 Box plot of subway arrival prediction inaccuracy (predicted time - actual time) using data collected from 4/14/2022 (Thu) 21:40 - 4/19/2022 (Tue) 15:51

We aggregated the trip count from January 2019 to July 2021 based on the day of the week and obtained Figure 5-1. To further zoom in the data, we investigated the daily ride count in a typical month and the plot in figure 5-2 shows a weekly pattern. The orange circulated data points are Sunday and Monday. From these two figures we can observe that on Sunday and Monday the taxi ridership is lower than the other days in general. In figure 5-3, the plot shows the correlation between the time of the day in hour and the trip count. We can find that from 8am to 10pm is the taxi hailing peak period.

A previous research examining influential factors of the taxi ridership in another metropolitan (Shanghai city) shows similar trends as our study [10].

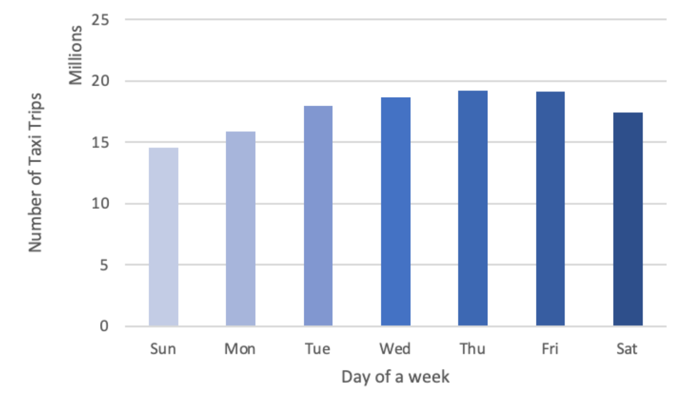

Fig. 5-1 The distribution of taxi ride in different days of a week

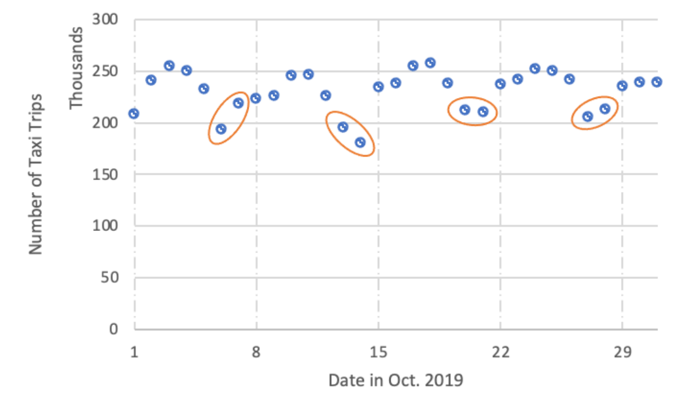

Fig. 5-2 The distribution of daily Ride Count in Oct. 2019

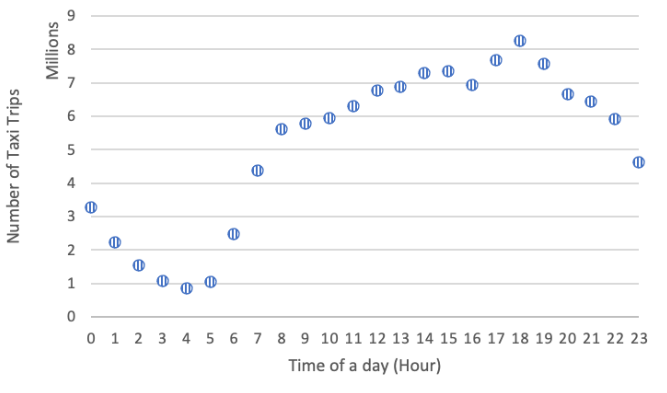

Fig. 5-3 Correlations of taxi ride and time of a day

From figure 5-4, we can observe that there is a sudden drop of trip count at the beginning of 2020 which is because most of the passengers stayed home at the covid time.

The goodness of the analytic can be tested by comparing the data with TLC aggregated reports [11] and TLC fast dash [12]. In the section of average daily trips per month, both of the reports show a pattern similar to our analysis.

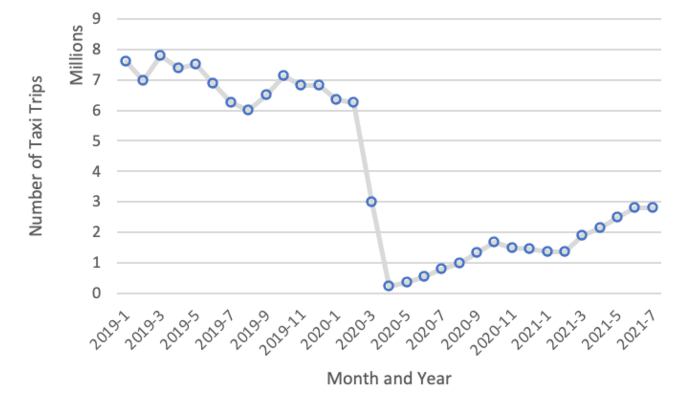

Fig. 5-4 Variations of monthly ride from 2019-01 to 2021-07

Further, we analyzed the spatial distribution of taxi trips according to pick up locations. As presented in figure 5-5, upper east side, midtown and Penn station are popular taxi hailing spots. The analysis done by S. Sawant also shows similar behavior [13].

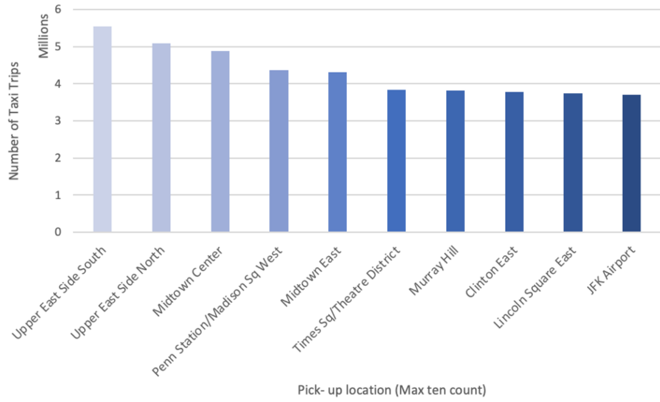

Fig. 5-5 Correlations of taxi ride and pick-up Locations

We also examined how the weather conditions impact the taxi operation and found out that rainfall doesn't influence the taxi ride significantly. By joining the weather and taxi data in the year of 2019, we obtained that the average trip per day in 2019 is 229,155 and average trip per day on a rainy day is 232,982 (1.67% higher than the yearly average). The research done by T. Schneider to analyze NYC taxi and uber trips also shows similar findings [14].

VII. FUTURE WORK

In this paper, we focus on the correlation between crime incidents and transportation based on the past data, providing insights into the transit and crime pattern with different seasons.

To observe the deeper relation, we can further observe how the crime incidents affect the ridership of each subway station and taxi; or how the traffic of public transit impacts the crime incidents. We will need to join the taxi and subway dataset with crime data by specific foreign keys, offering crime statistics, within a distance range, to different taxi pickup locations.

VIII. CONCLUSION

Our analysis focused on crime and transportation data in New York City to find insights that could provide more aspects of New York's daily life. We combined several data sources to find the pattern of crime incidents and the operation of subway and taxi services. In addition, the correlations between crime/ taxi and weather are studied.

One of the most remarkable findings is that transit crime behavior is more likely to occur at the rush hours instead of midnight or early morning. Furthermore, the main category of transit crime is the so-called pickpocketing. In other words, passengers should pay more attention to this kind of petit crime in their daily lives.

We also discovered that the MTA's predictions for subway arrival times were quite accurate within our data collection period of around 5 days, having error less than 30 seconds.

Finally, the spatial and temporal distribution of taxi ridership patterns is investigated. The coronavirus pandemic had significant impacts on the taxi services market and in our findings we can see a post-covid business recovery. Furthermore, by concatenating with weather data, our results show that the taxi ride change affected by rainfall is minor.

ACKNOWLEDGMENTS

We would like to show our gratitude to the HPC team at New York University for making Peel cluster available where we mainly developed our project. Also, we would like to thank the NYPD, TLC, MTA and NOAA for making all the available data. Again, to Tableau for offering us a free license. Finally, we are immensely grateful to Prof. Tang for the insightful comments and invaluable guidance in this semester.

REFERENCES

[1] Quinnipiac University New York City Poll: https://poll.qu.edu/poll-release?releaseid=3845

[2] NYPD Complaint Data Historic: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[3] NYPD Complaint Data Current (Year To Date): https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243

[4] TLC Trip Record Data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[5] MTA Realtime Data Feeds: https://api.mta.info/#/landing

[6] NOAA public daily weather: https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail

[7] M. Ranson, Crime, Weather, and Climate Change, Harvard Kennedy School (2012)

[8] MTA Schedules: https://new.mta.info/schedules

[9] Train Operator On-the-Job Experience: https://www.nyctransitforums.com/topic/51274-train-operator-on-the-job-experience/

[10] C. Chen. et al. Examining the spatial-temporal relationship between urban built environment and taxi ridership: Results of a semi-parametric GWPR model. Journal of Transport Geography 96 (2021), pp. 3.

[11] TLC Aggregated Reports: https://www1.nyc.gov/site/tlc/about/aggregated-reports.page

[12] TLC Fast Dash: https://tlcanalytics.shinyapps.io/tlc_fast_dash/

[13] Exploratory Data analysis of New York City Taxi Ride: https://medium.com/analytics-vidhya/data-science-of-new-york-city-taxi-ride-data-b55a4f9145de

[14] Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance: https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/