Smarter Transit

Apr, 4, 2022

Pin-Yi Chiu, Nini Lin, An Lee, Hong-Ren Mao

Smarter Transit

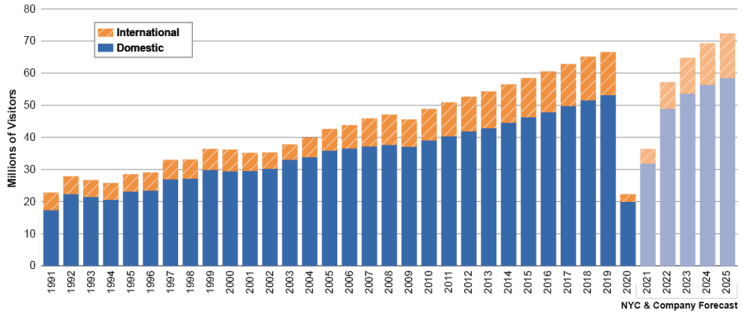
This report revealed the correlation of weather, transit, and crime in New York City, built on a panel of daily crime, taxi trip records, subway arrival time, and temperature, using the NYU peel cluster as the platform.

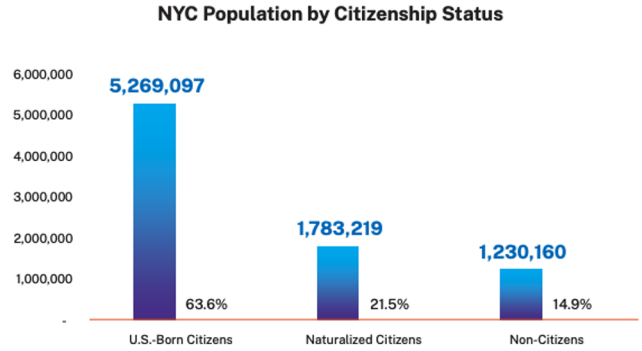
We identify the effect of weather on daily, seasonally crime by using the temperature and the number of crimes from 2006 to 2021.Looking into the relationship between transit and crime in New York City.

Smarter Transit

New York City is a big city but has a reputation for crime. Further, the recent subway attack in Brooklyn has raised more concerns about public transportation. Therefore, we decide to analyze the correlation between crime and transit.

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





Smarter Transit

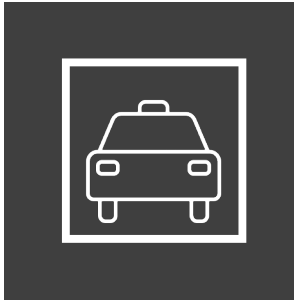
NYPD Complaint Data Historic

- 2.4 G
- All valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2020.

NYPD Complaint Data Current

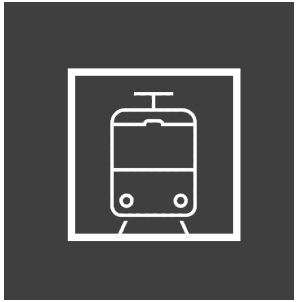
- 175MB
- All valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) for all complete quarters so far 2021.

Smarter Transit



TLC Trip Record Data

- 11.65 GB
- Monthly issued report containing four different taxi categories, including the time and location for pickup and drop off of each trip.
- The yellow taxi trip records (2019/01 - 2021/07)



Real-time subway feed from MTA

- 41 GB
- Collected the API response every 30 seconds during 4/14/2022 (Thu) 21:40 - 4/19/2022 (Tue) 15:51 for these subway lines: 123456BDFMNQRWJZ

NOAA public daily weather (ghcnd_all)

- 2.2 GB
- The dataset storage the weather of USW data from 1900s to now by region separately and contains weather conditions, including five core temperature(min & max), snow condition(snowfall & depth) and precipitation.

Smarter Transit

NYPD Complaint Data

- **Before data ingestion:**

```
[hl4674@hlog-1 used_dataset]$ head -n 3 NYPD_Complaint_Data_Historic.csv
CMPLNT_NUM,CMPLNT_FR_DT,CMPLNT_FR_TM,CMPLNT_TO_DT,CMPLNT_TO_TM,ADDR_PCT_CD,RPT_DT,KY_CD,OFNS_DESC,PD_CD,PD_DESC,CRM_ATPT_CPTD_CD,LAW
_CAT_CD,BORO_NM,LOC_OF_OCCUR_DESC,PREM_TYP_DESC,JURIS_DESC,JURISDICTION_CODE,PARKS_NM,HADEVELOPT,HOUSING_PSA,X_COORD_CD,Y_COORD_CD,S
USP_AGE_GROUP,SUSP_RACE,SUSP_SEX,TRANSIT_DISTRICT,Latitude,Longitude,Lat_Lon,PATROL_BORO,STATION_NAME,VIC_AGE_GROUP,VIC_RACE,VIC_SEX
394506329,12/31/2019,17:30:00,,,32,12/31/2019,118,DANGEROUS WEAPONS,793,WEAPONS POSSESSION 3,COMPLETED,FELONY,MANHATTAN,,STREET,N.Y.
 POLICE DEPT,0,,,,999937,238365,,,,,40.82092679700002,-73.94332421899996,"(40.82092679700002, -73.94332421899996)",PATROL BORO MAN N
ORTH,,UNKNOWN,UNKNOWN,E
968873685,12/29/2019,16:31:00,12/29/2019,16:54:00,47,12/29/2019,113,FORGERY,729,"FORGERY,ETC.,UNCLASSIFIED-FELO",COMPLETED,FELONY,BR
ONX,,STREET,N.Y. POLICE DEPT,0,,,,1022508,261990,,,,,40.885701406000074,-73.86164032499995,"(40.885701406000074, -73.86164032499995)
",PATROL BORO BRONX,,UNKNOWN,UNKNOWN,E
```

```
[hl4674@hlog-1 used_dataset]$ head -n 3 NYPD_Complaint_Data_Current__Year_To_Date_.csv
CMPLNT_NUM,ADDR_PCT_CD,BORO_NM,CMPLNT_FR_DT,CMPLNT_FR_TM,CMPLNT_TO_DT,CMPLNT_TO_TM,CRM_ATPT_CPTD_CD,HADEVELOPT,HOUSING_PSA,JURISDICT
ION_CODE,JURIS_DESC,KY_CD,LAW_CAT_CD,LOC_OF_OCCUR_DESC,OFNS_DESC,PARKS_NM,PATROL_BORO,PD_CD,PD_DESC,PREM_TYP_DESC,RPT_DT,STATION_NAM
E,SUSP_AGE_GROUP,SUSP_RACE,SUSP_SEX,TRANSIT_DISTRICT,VIC_AGE_GROUP,VIC_RACE,VIC_SEX,X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lat_Lon
,New Georeferenced Column
903695881,69,,12/17/2021,22:13:00,,,COMPLETED,,,,N.Y. POLICE DEPT,101,FELONY,OUTSIDE,MURDER & NON-NEGL. MANSLAUGHTER,,,,,,12/17/2021
,,25-44,BLACK,M,,25-44,BLACK,M,1011203,174515,40.64564719600002,-73.90287588699994,"(40.64564719600002, -73.90287588699994)",POINT (
-73.90287588699994 40.64564719600002)
400462399,113,,12/17/2021,06:21:00,,,COMPLETED,,,,N.Y. POLICE DEPT,101,FELONY,OUTSIDE,MURDER & NON-NEGL. MANSLAUGHTER,,,,,,12/17/202
1,,,,,,25-44,BLACK,F,1043252,187998,40.682499421000045,-73.78726915499993,"(40.682499421000045, -73.78726915499993)",POINT (-73.7872
6915499993 40.682499421000045)
```

- **After data ingestion:**

cmplnt_time	law_cat	boro_nm	prem_type	parks_nm	transit_district	station_nm
2007-03-15 15:00:00.0	FELONY	BROOKLYN	TRANSIT - NYC SUBWAY	NA	30	NOSTRAND AVENUE
2010-09-28 16:45:00.0	MISDEMEANOR	MANHATTAN	TRANSIT - NYC SUBWAY	NA	2	CANAL STREET
2006-03-28 20:00:00.0	MISDEMEANOR	QUEENS	TRANSIT - NYC SUBWAY	NA	20	BROADWAY
2006-12-28 15:25:00.0	FELONY	MANHATTAN	TRANSIT - NYC SUBWAY	NA	4	125 STREET
2009-10-01 14:10:00.0	MISDEMEANOR	MANHATTAN	TRANSIT - NYC SUBWAY	NA	4	125 STREET
2010-11-04 19:09:00.0	MISDEMEANOR	MANHATTAN	TRANSIT - NYC SUBWAY	NA	3	110 ST.-CATHEDRAL PKWY.
2009-12-02 17:30:00.0	MISDEMEANOR	MANHATTAN	TRANSIT - NYC SUBWAY	NA	2	34 ST.-HERALD SQ.
2006-01-05 07:56:00.0	FELONY	BROOKLYN	TRANSIT - NYC SUBWAY	NA	33	ROCKAWAY AVENUE
2008-04-26 02:55:00.0	MISDEMEANOR	MANHATTAN	TRANSIT - NYC SUBWAY	NA	1	59 ST.-COLUMBUS CIRCLE
2009-08-18 20:45:00.0	FELONY	MANHATTAN	TRANSIT - NYC SUBWAY	NA	2	CANAL STREET

Smarter Transit

TLC Trip Record Data

- **Before data ingestion:**
  Each monthly published data is in CSV format. The first line of each file is the schema/ field name, and the following lines are line-based records.

```
rayichiu@rayis-mbp yellow_taxi_data % cat yellow_tripdata_2021-01.csv | head -n 5
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,f
,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount,congestion_surcharge
1,2021-01-01 00:30:10,2021-01-01 00:36:12,1,2.10,1,N,142,43,2,8,3,0.5,0,0,0.3,11.8,2.5
1,2021-01-01 00:51:20,2021-01-01 00:52:19,1,.20,1,N,238,151,2,3,0.5,0.5,0,0,0.3,4.3,0
1,2021-01-01 00:43:30,2021-01-01 01:11:06,1,14.70,1,N,132,165,1,42,0.5,0.5,8.65,0,0.3,51.95,0
1,2021-01-01 00:15:48,2021-01-01 00:31:01,0,10.60,1,N,138,132,1,29,0.5,0.5,6.05,0,0.3,36.35,0
```

- **After data ingestion:**
  Pick-up date and time, the distance of the trip and the pick- up location ID.

```
[pc3095@hlog-2 pc3095]$ hdfs dfs -cat /user/pc3095/projectMR/output/part-m-00000 |
2020-03-07 18:39:07,9.70,132
2020-03-07 18:27:02,1.01,170
2020-03-07 18:40:08,1.31,163
2020-03-07 14:57:54,15.60,138
2020-03-07 18:19:51,1.12,48
```

Smarter Transit

Subway arrival time

- **Before data ingestion:**


```
{u'entity': [{u'id': u'000001N',
              u'tripUpdate': {u'stopTimeUpdate': [{u'[transit_realtime.nyct_stop_time_update]': {u'actualTrack': u'0B',
                                                                                                 u'scheduledTrack': u'0B'},
                                                  u'arrival': {u'time': u'1650314822'},
                                                  u'departure': {u'time': u'1650314822'},
                                                  u'stopId': u'D43S'}],
                              u'trip': {u'[transit_realtime.nyct_trip_descriptor]': {u'direction': u'SOUTH',
                                                                                     u'isAssigned': True,
                                                                                     u'trainId': u'1N 1535 DIT/STL'},
                                        u'routeId': u'N',
                                        u'startDate': u'20220418',
                                        u'startTime': u'15:35:00',
                                        u'tripId': u'093500_N..S'}}},
             {u'id': u'000002N',
              u'vehicle': {u'currentStatus': u'STOPPED_AT',
                           u'currentStopSequence': 27,
                           u'stopId': u'N10S',
                           u'timestamp': u'1650314697',
                           u'trip': {u'[transit_realtime.nyct_trip_descriptor]': {u'direction': u'SOUTH',
```


Subway arrival time

- **After data ingestion:**
  time_of_prediction, predicted_arrival_time, real_arrival_time, stopId, subway_line


```
1650154034, 1650157152, 1650157236, B08N, Q
1650154124, 1650157126, 1650157236, B08N, Q
1650154222, 1650157212, 1650157236, B08N, Q
1650155817, 1650157092, 1650157236, B08N, Q
1650155686, 1650157146, 1650157236, B08N, Q
1650156347, 1650157192, 1650157236, B08N, Q
```

NYC_Daily_Weather

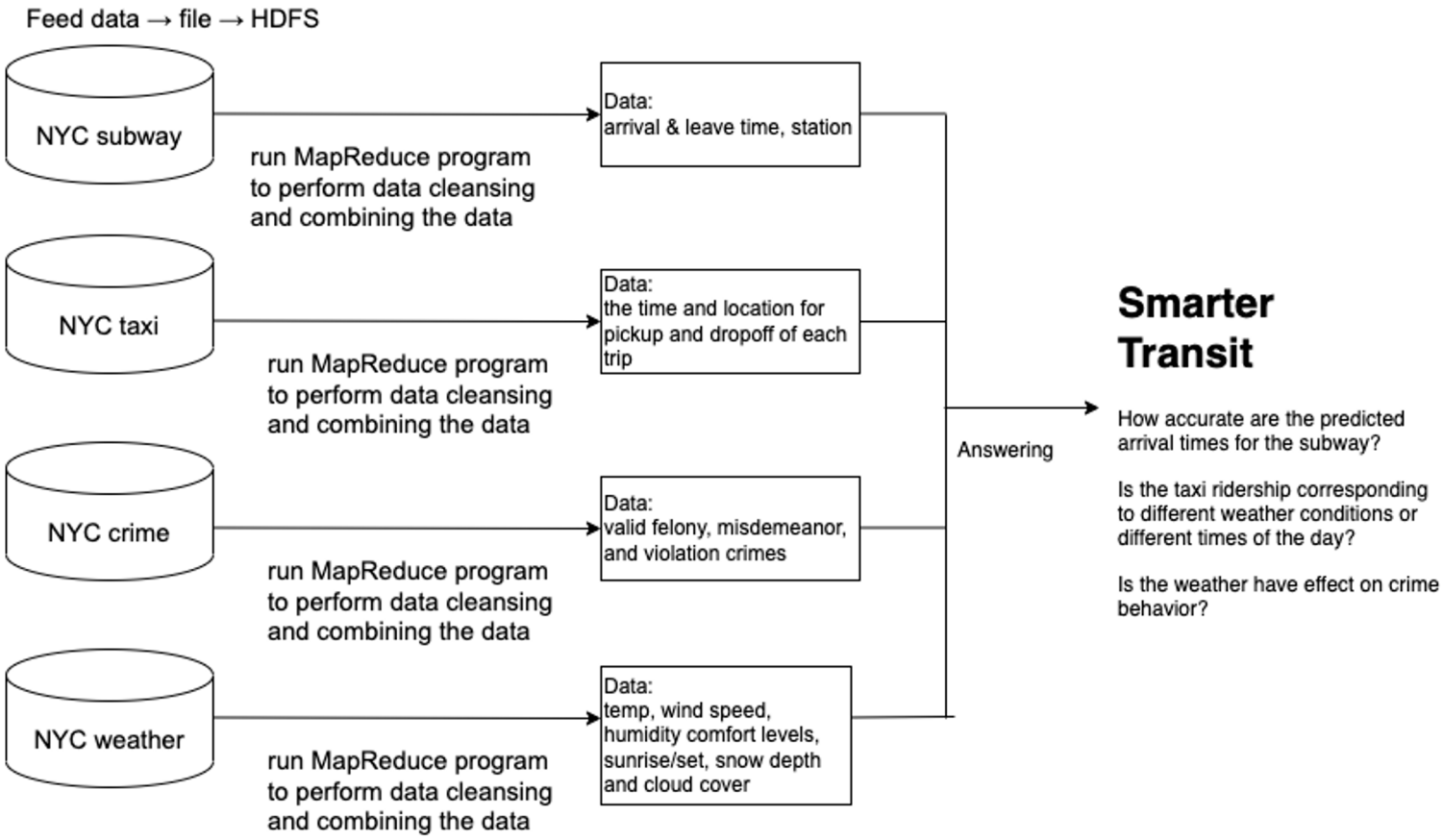
- Before data ingestion: each record in a file contains one month of daily data and quality flag

```
[al7527@hlog-1 project]$ hadoop fs -cat /user/al7527/weather_data/ghcnd_all/USW00094789.dly
USW00094789194807TMAX-9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -99
99   -9999   -9999     233  X  272  X  294  X  289  X  294  X  272  X  261  X  278  X  283  X  272  X  278  X  322  X  317  X  2
89  X  317  X
USW00094789194807TMIN-9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -99
99   -9999   -9999     178  X  211  X  211  X  206  X  217  X  217  X  217  X  200  X  183  X  194  X  217  X  228  X  217  X  2
17  X  233  X
USW00094789194807PRCP-9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -9999   -99
99   -9999   -9999      15  X    0T X    0T X    0  X  622  X    0T X  124  X    0  X    0  X    0  X  160  X    0T X    0  X
 0T X    0T X
```

- After data ingestion: set each daily element separately

```
+---------------------+------------------+-------------------+------------------+
| weather_new.region  | weather_new.ele  | weather_new.time  | weather_new.val  |
+---------------------+------------------+-------------------+------------------+
| USW00094789         | WT08             | 2022-03-18        |     1            |
| USW00094789         | WT04             | 2022-03-12        |     1            |
| USW00094789         | WT04             | 2022-03-09        |     1            |
| USW00094789         | WT03             | 2022-03-19        |     1            |
| USW00094789         | WT02             | 2022-03-19        |     1            |
+---------------------+------------------+-------------------+------------------+
```

Smarter Transit



Smarter Transit

NYPD Crime Data

- Designing MapReduce with multiple inputs to read files with special delimiter
- Merging data sources with inconsistent columns order.
- Check the validity of complaint datetime

NYC_Daily_Weather

- Using single mapper to transfer its format and filter out the related weather elements that we needed
- the challenge is to design a more readable format for others to use and recognize related region and weather elements.

Subway Arrival Time data

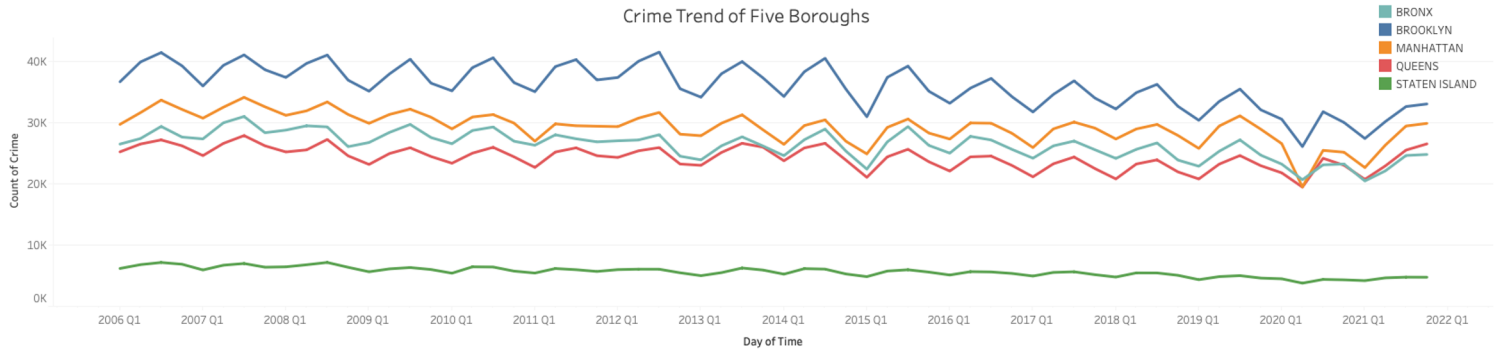
- Saving feed data
  * Protobuf deserialization (codegen, custom Python version)
  * Long-running program → learn how to use tmux

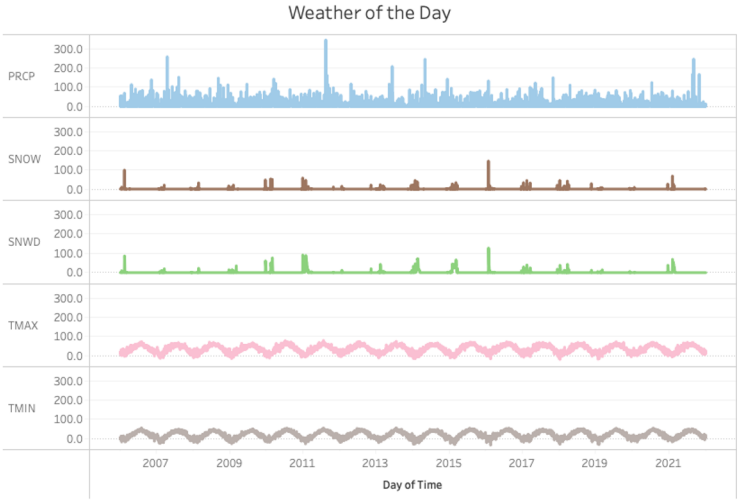
TLC Trip Record Data

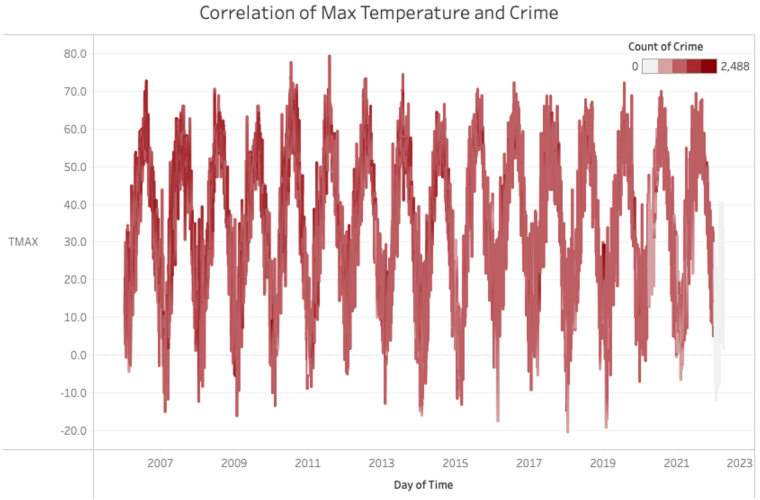
- Since the taxi data is relatively complete and clean, we can simply use a single mapper to filter out the inaccurate data,
- Furthermore, we use Hive to count the number of trips under different conditions.

Smarter Transit

Temperature Crime Behavior

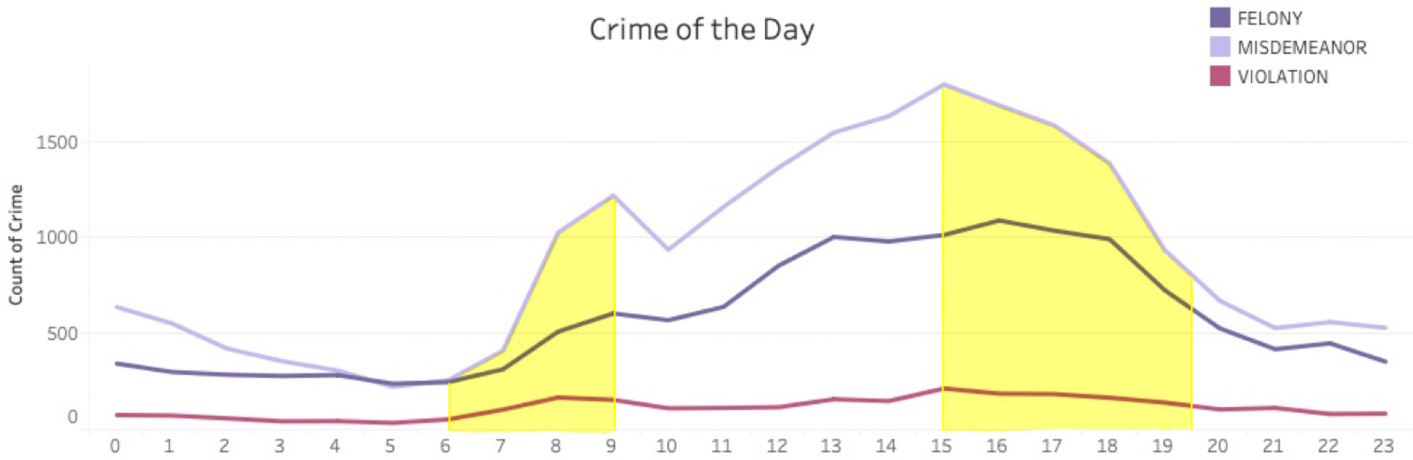


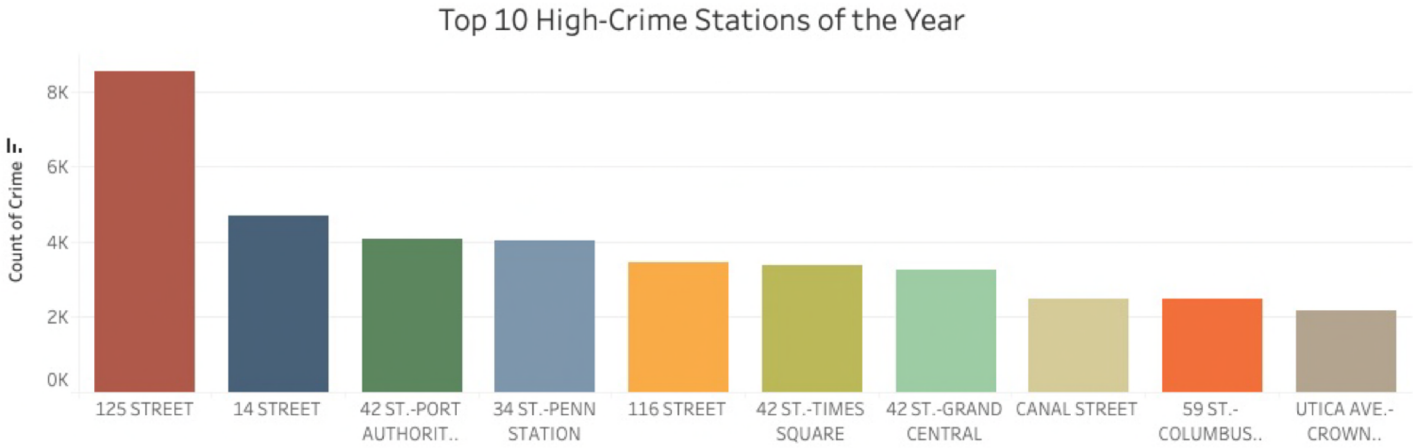




Smarter Transit

Transit Crime Behavior

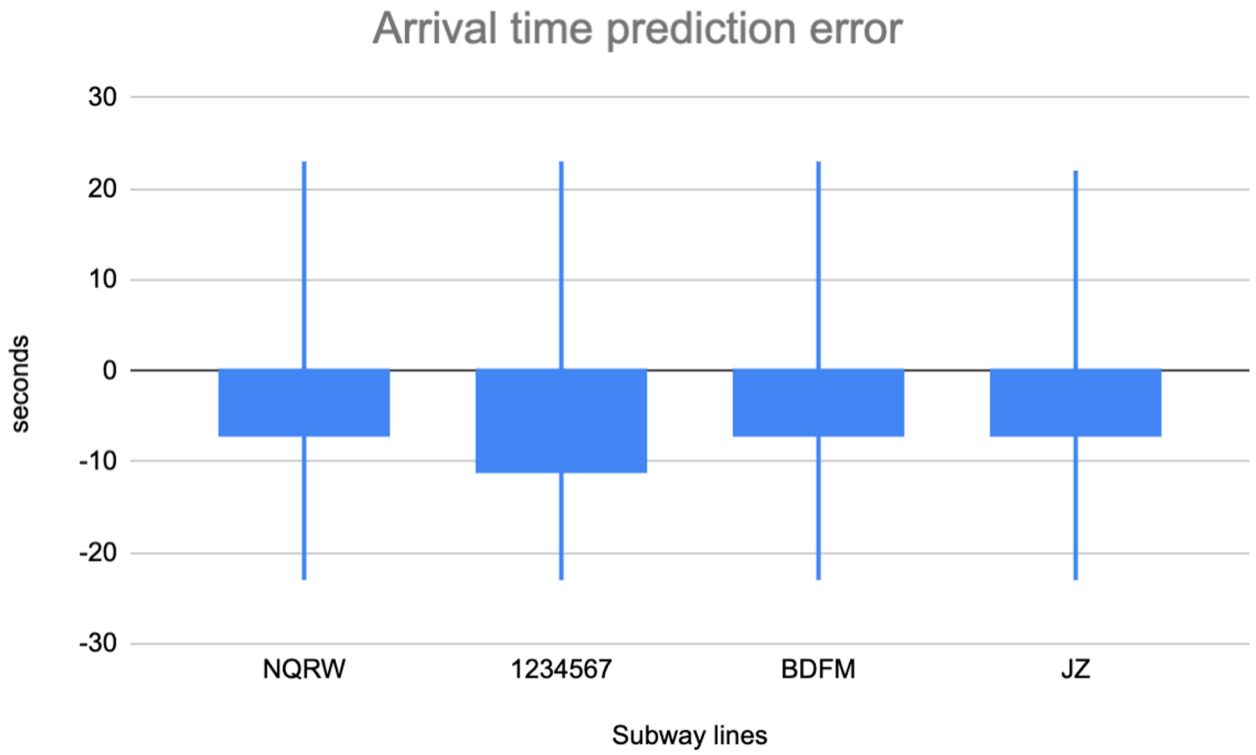




Smarter Transit

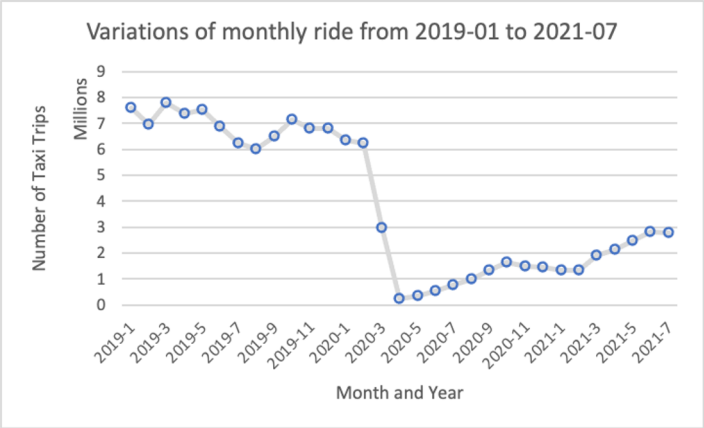
Subway arrival times:

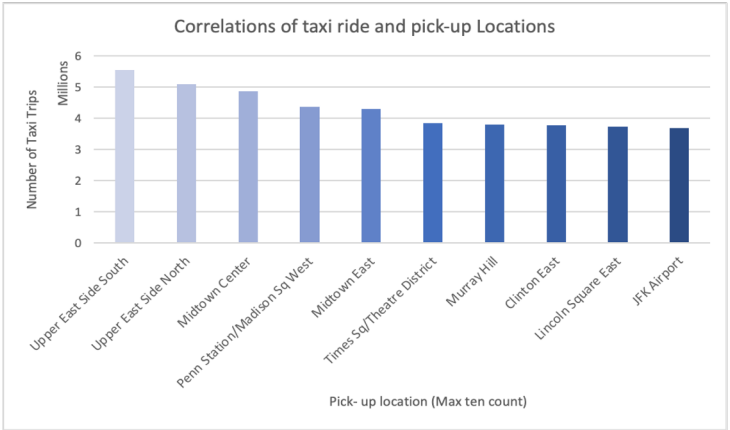
MTA's real-time predictions were pretty accurate within our data collection time frame

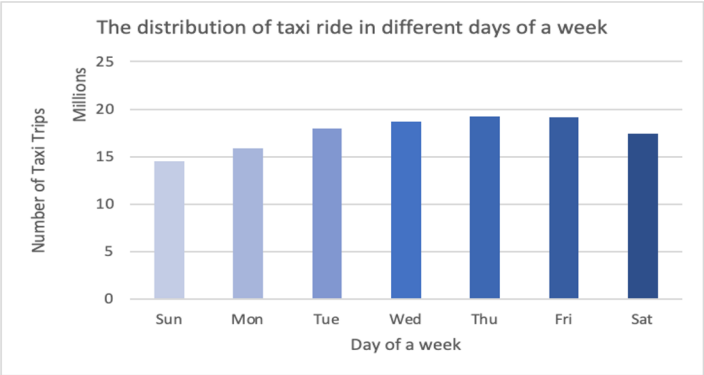


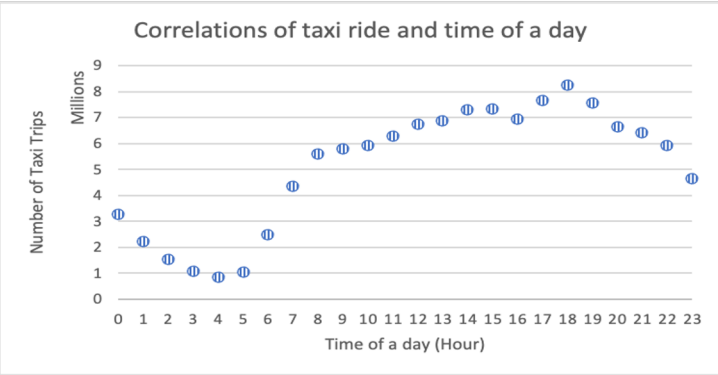
Smarter Transit

Taxi data observations



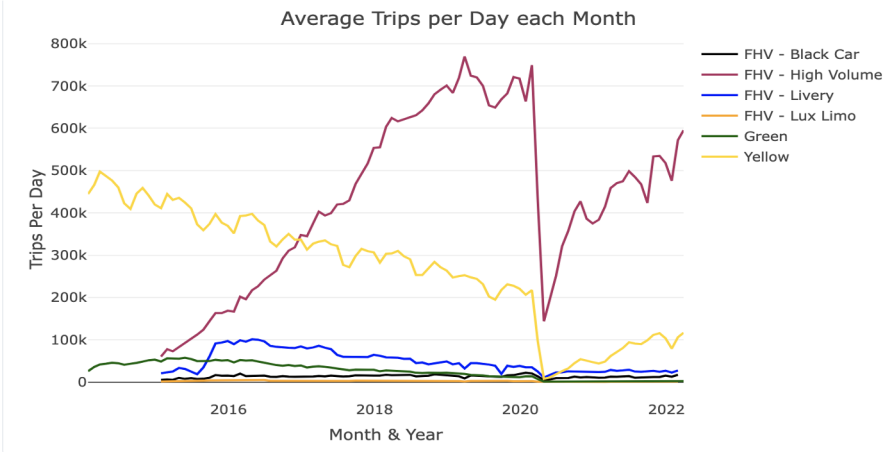




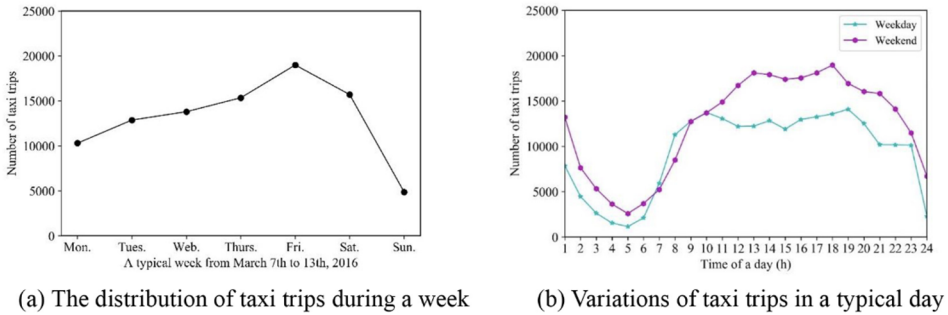


Smarter Transit

Taxi analysis



Average daily trips distribution according to analysis conducted by TLC organization: TLC Fast Dash



(a) The distribution of taxi trips during a week (b) Variations of taxi trips in a typical day

Fig. 2. The temporal distribution of taxi trips in Shanghai city.

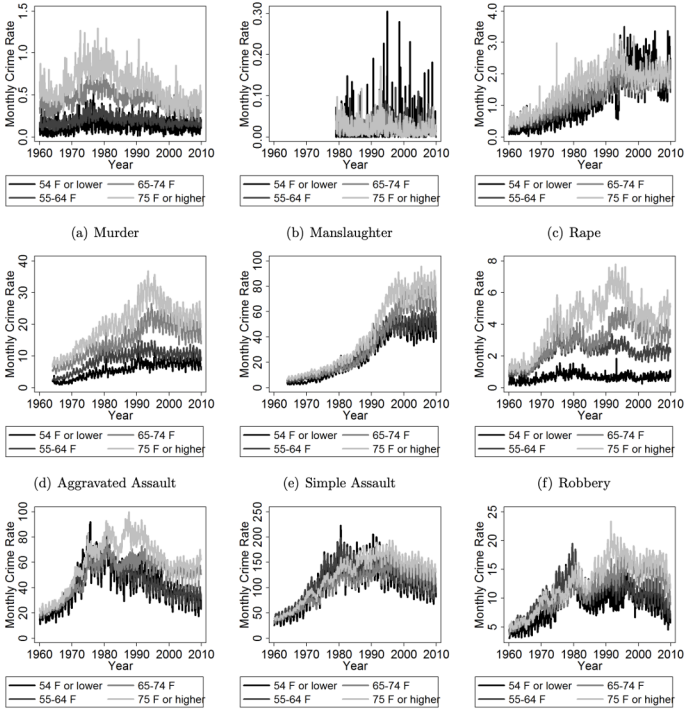
C. Chen. et al. Examining the spatial-temporal relationship between urban built environment and taxi ridership: Results of a semi-parametric GWPR model. https://doi.org/10.1016/j.jtrangeo.2021.103172

Smarter Transit

Subway arrival time analytics:

The subway trains have schedules, and a train operator says in a forum that these are strictly enforced

Crime analysis:



Crime Rate Trends, by Climate Zone

- **Processing saved feed data into a usable dataset was non-trivial**
  - Sift through many API response fields and read the documentation VERY carefully
  - Design custom MapReduce logic to group data by trip and deduce actual arrival time

Summary:

- The higher the temperature, the higher the occurrence probability of crime.
- Crimes happen the most during rush hours, instead of midnight
- Subway is on time in most cases, and people are more willing to take taxis close to the weekend

Acknowledgements

Thank:

- **HPC team** at New York University for making Peel cluster available.
- **NYPD**, **TLC**, **MTA** and **NOAA** for making all the data used for this project available.
- **Tableau** for offering us free license.
- **Prof. Tang** for the insightful comments and invaluable guidance in this semester.

References

- Divya Sharma Nag, Shubhangi Pandey and Pratima Singh
  The Safe Cities Index 2021 Report
  The Economist Intelligence Unit, 2021
- NYPD CompStat Unit
  City Wide Crime Statistics Weekly
  NYPD CompStat Unit, 2022
- Matthew Ranson
  Crime, Weather, and Climate Change
  Harvard Kennedy School, 2012
- Daniel Gayne
  Trouble underground as crime stats reveal varying success of policing on London's public transport
  SW Londoner, 2020
- Mayor's Office of Immigrant Affairs
  State of Our Immigrant City Annual Report
  Mayor's Office of Immigrant Affairs, 2018
- Office of New York State Comptroller
  State of Our Immigrant City Annual Report
  Office of New York State Comptroller, 2021

Thank you!