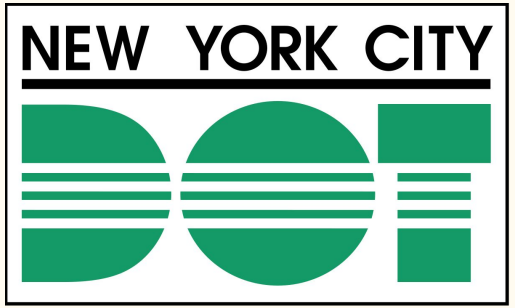
NYC Traffic Dynamics Assessment

Yashvardhan Singh - ys5608
Harsh Bansal - hb2709
Lakshay Tyagi - lt2504
Utkarsh Tyagi - ut2028

Abstract



- Department of Transportation (DOT)
  - Traffic Volume
  - Traffic Speed
  - Motor Vehicle Crashes
  - Parking Violations
- Individual datasets for interesting insights
- Analysing Correlations: Speed, Volume, Crash Data and Parking Violations
- Surprising Discoveries: Dispelling some misconceptions

  Does higher traffic speed lead to more accidents?

Motivation

End user :

- Government officials for data driven policy making
- NYPD for effective resource and personnel allocation

Beneficiaries :

- The people of NYC !! - Public will benefit from effective governance

Importance :

- Identifying streets with high incidence of accidents can save lives
- Streets that have high volume can be expanded or traffic can be re-routed
- Parking spaces can be built in precincts with high frequency of parking violations

Goodness

1. **Data Trustworthiness:** All of our datasets were obtained from the Department of Transportation(DoT) making them trusted and reliable.
2. **Bias Elimination:** To eliminate potential bias, we utilized a subset of the data that was shared across datasets when deriving insights.
3. **Hypothesis Confirmation:** The outcomes align with our initial hypothesis regarding the interplay of different traffic dynamics.
4. **Outcome validation:** Our outcomes matches with the outcomes from other studies[5].

Data source - Motor Vehicle Collisions

Updated: Daily (last updated in December 2023)
Number of rows: 2.04 M
Size - ~414 MB
Number of columns: 29
Relevant columns:

- Crash Date - Date at which the collision occurred
- Crash Time - Time at which the collision occurred
- On Street Name - The name of the street where the incident occured
- Persons Injured - Number of people injured in the incident
- Persons Killed - Number of people killed in the incident
- Borough - The borough where the incident occured

```
 crash_date | crash_time |  borough  | on_street_name | number_of_persons_injured | number_of_persons_killed
------------+------------+-----------+----------------+---------------------------+-------------------------
 11/18/2015 | 7:57       | QUEENS    | 46 ave         |                         1 |                        0
 11/18/2015 | 19:30      | MANHATTAN | east 36 st     |                         0 |                        0
(2 rows)
```

Data sources - Automated Traffic Volume Counts

Updated: Annually (last updated in December 2022)
Number of rows: 27.4 M
Number of columns: 14
Size: ~3.2GB
Relevant columns:

- RequestID - Id corresponding to the vehicle count entries
- Date[Yr,M,D,HH,MM] - date and time at which vehicle count was conducted
- Vol - the number of vehicles counted at given timestamp
- Street - the name of the street
- Boro - the Borough of the street

requestid	boro	yr	m	d	hh	mm	vol	street
16082	Manhattan	2013	11	21	19	0	226	holland tun ext
16082	Manhattan	2013	11	21	18	45	223	holland tun ext
16082	Manhattan	2013	11	21	18	30	230	holland tun ext
16082	Manhattan	2013	11	21	18	15	243	holland tun ext
16082	Manhattan	2013	11	21	18	0	235	holland tun ext

(5 rows)

Data sources - DOT Traffic Speed NBE

Updated: Daily (last updated in December 2023)
Number of rows: 76.9 M
Size of Dataset: 33.4 GB
Number of columns: 13
Relevant columns:

- Speed - Speed of vehicle through the link
- Travel_Time - Time taken to cross the link
- Date_Time - Time when the speed was recorded
- Borough - Borough (Brooklyn, Manhattan, etc)
- From_Street - Entry street of link
- To_Street - Exit street of link

```
 speed | travel_time | status |        date_time        | borough |   from_street    |    to_street
-------+-------------+--------+-------------------------+---------+------------------+-----------------
 54.68 |         137 |      0 | 2022-06-22 07:44:12.000 | Bronx   | CASTLE HILL AVE | GRISWOLD AVE
 54.68 |          85 |      0 | 2022-06-22 07:44:12.000 | Bronx   | STRATFORD AVE   | CASTLE HILL AVE
(2 rows)
```

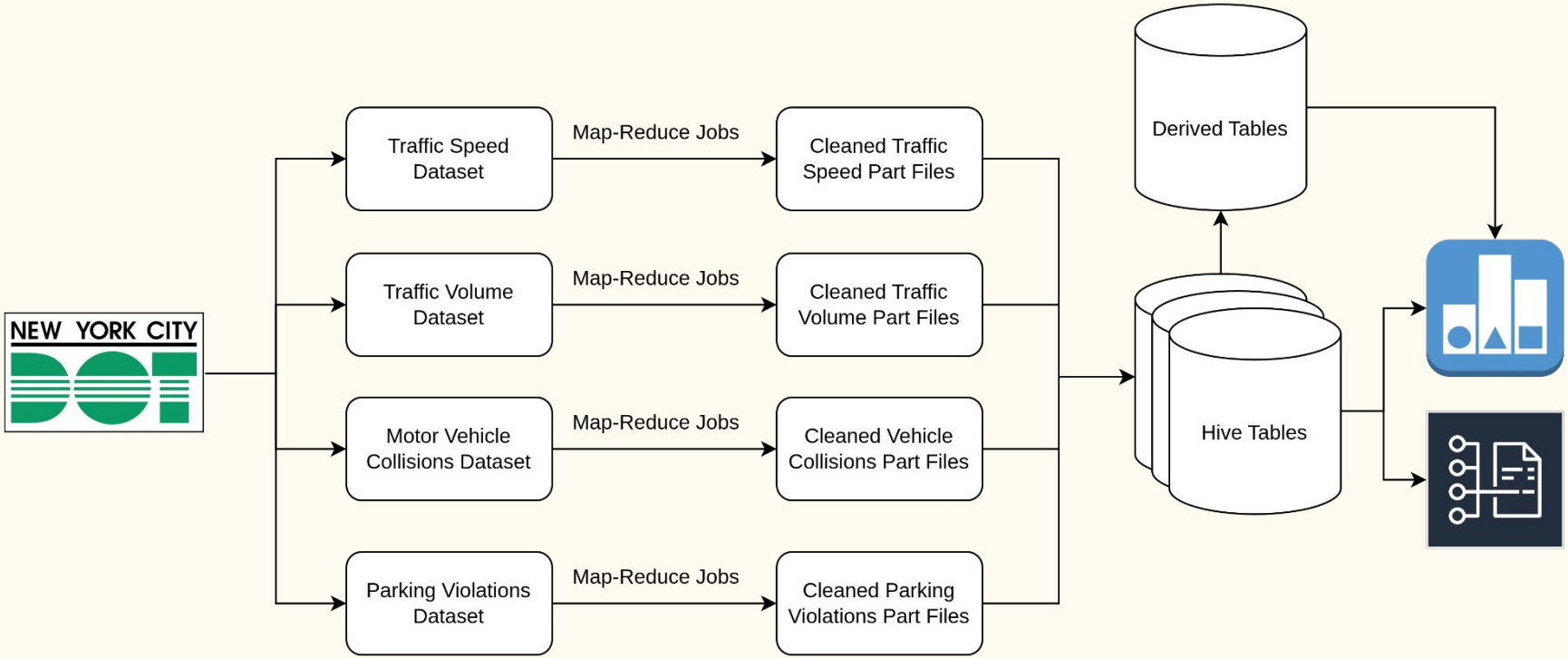
Data sources - Parking Violations Issued

- **Updated:** Annually (last updated in October 2023)
- **Number of rows:** 120.3 M
- **Number of columns:** 44
- Relevant columns:
  - Issue Date - Date the violation was issued
  - From Time - The time the parking violation started
  - End Time - The time the violation ended
  - Violation Precinct - The precinct the violation occurred in
  - Street Name - Name of the street the violation occurred in
  - State - The state violating vehicle belonged to

issue_date	violation_code	violation_precinct	intersecting_street	vehicle_yr	from_hours	to_hours	feet_from_curb
FL	38	103	91 ave	0700	0	0830	0
FL	38	106	liberty ave	0700	0	0900	0
FL	38	106	liberty ave	0700	0	0900	0
FL	38	106	liberty ave	0700	0	0900	0
FL	38	107	main st	0700	0	0730	0
FL	38	107	manton st	0700	0	0800	0
FL	38	108	40 st	0700	0	0900	0
FL	38	108	roosevelt ave	0700	0	0900	0
FL	38	109	college point blvd	0700	0	0800	0
FL	38	112	metropolitan ave	0700	0	0800	0

(10 rows)

Design Diagram



Code Challenge - Motor Vehicle Collisions

Challenge: Street name standardization. In order to find correlations in data from different datasets based on the street name, we need to standardize them so that they would match across different datasets. For example, 5th street is present in different datasets as 5 street, 5th st etc.

Solution:

1. Remove “(“ and “)” from the name
2. Remove ["ST", "ND", "RD", "TH"] after numerals
3. Map (“STREET” : “ST”, “AVENUE” : “AVE”, “AV” : “AVE”, “BOULEVARD”: “BLVD”, “ROAD”: “RD”, “PARKWAY”: “PKY”, “HIGHWAY”: “HWY”, “COURT”: “CT”, “PLACE”: “PL”, “SQUARE”: “SQ”, “TURNPIKE”: “TPKE”, “LANE”: “LN”, “POINT”: “PT”, “PLAZA”: “PZ”)
4. Remove these strings, if present, and the word preceding them: ["LEVEL"]
5. Remove these strings, if present, and the word after them: ["EXIT"]

Code Challenge - Automatic Traffic Volume

Challenge : In order to find the trends across different days of the week, we first had to find the day from the date. In this dataset the date was spread across 3 different columns. Merging these 3 together through SQL query and then finding the day was a tough task.

Solution:

```
SELECT day_of_week,AVG(vol) FROM (
SELECT date_format(date_parse(formatteddate, '%Y-%m-%d'), '%W') AS day_of_week,vol
FROM (SELECT FormattedDate,vol FROM (SELECT Yr,M,D,
CONCAT( CAST(yr AS VARCHAR), '-', LPAD(CAST(m AS VARCHAR), 2, '0'), '-', LPAD(CAST(d AS
VARCHAR), 2, '0')) AS FormattedDate, vol FROM atvc)))
GROUP BY day_of_week
ORDER BY day_of_week;
```

Code Challenge - Traffic Speed Dataset

Challenge: CSV Parsing - My original dataset was in the csv format, but even within a single column, there were values separated by commas(‘,’). Eg: The geocode had a list of lat-long values. Eg: “40.60,-74.14,40.61”. This proved to be a challenge in parsing the row in the Mapper.

Solution:

Initially I observed that the Geo-Codes were enclosed within quotes, i.e., “lat-longs” while no other column had quotes. My initial workaround was to extract the segment within quotes, and replace the commas in this segment with a blank space.

After initial profiling, I observed the column Status has a value of “-101” for invalid rows. This allowed me to further skip these rows in the final cleaned data.

Code Challenges - Parking Violations Dataset

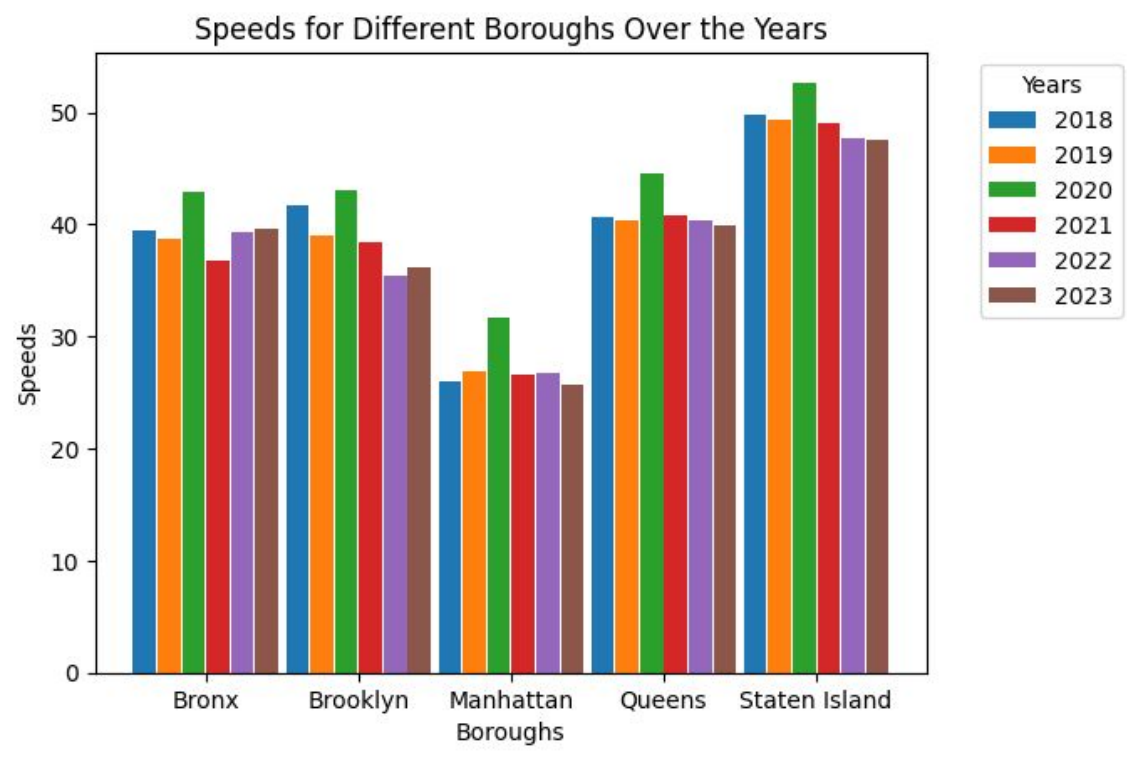
Challenge:

- Date inconsistency (ex - Issue date - 10/23/2030)
- Invalid timestamps (ex - 15:00 PM)
- Non existent state codes

Solution:

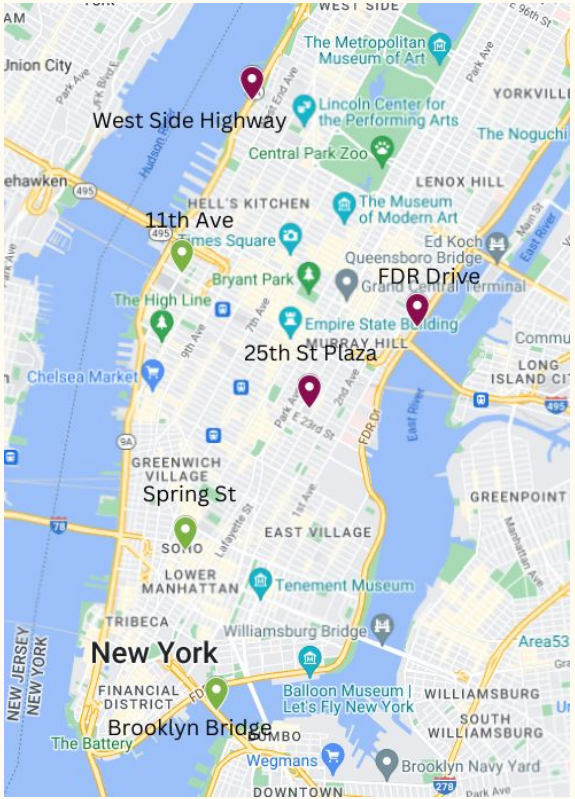
- Robust Map Reduce filtering code removes inconsistencies of years beyond 2023
- Incosistent timestamps are all converted to 24 hour format
- All state codes are checked against a list of street codes to ensure validity

Results

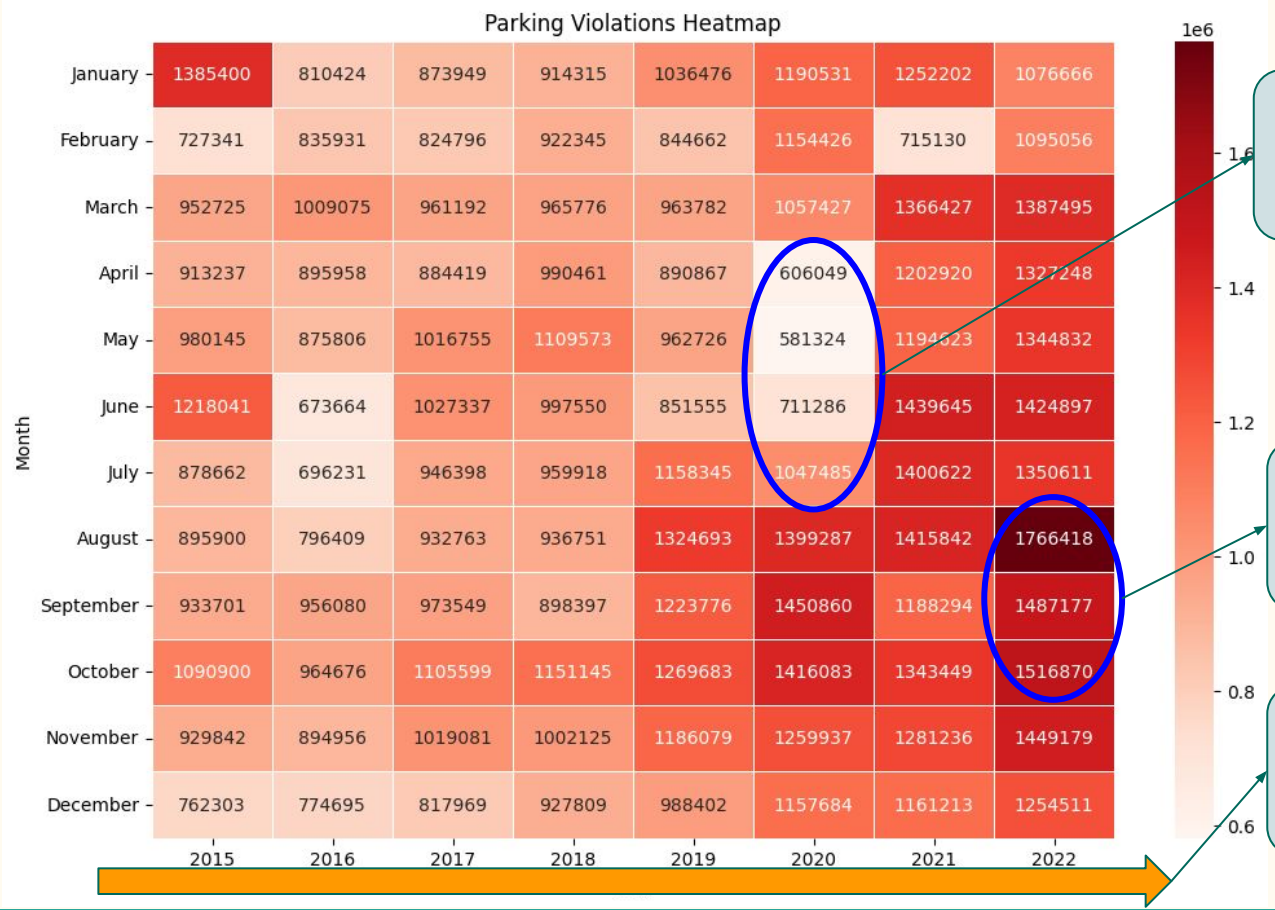


Fastest Streets

Slowest Streets



Parking Violation Results

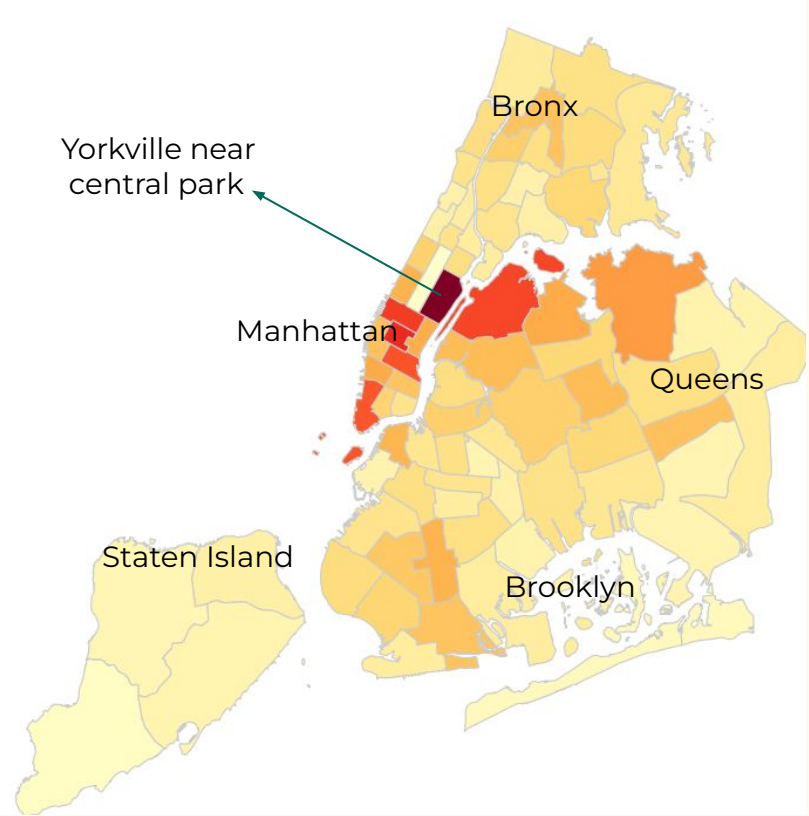


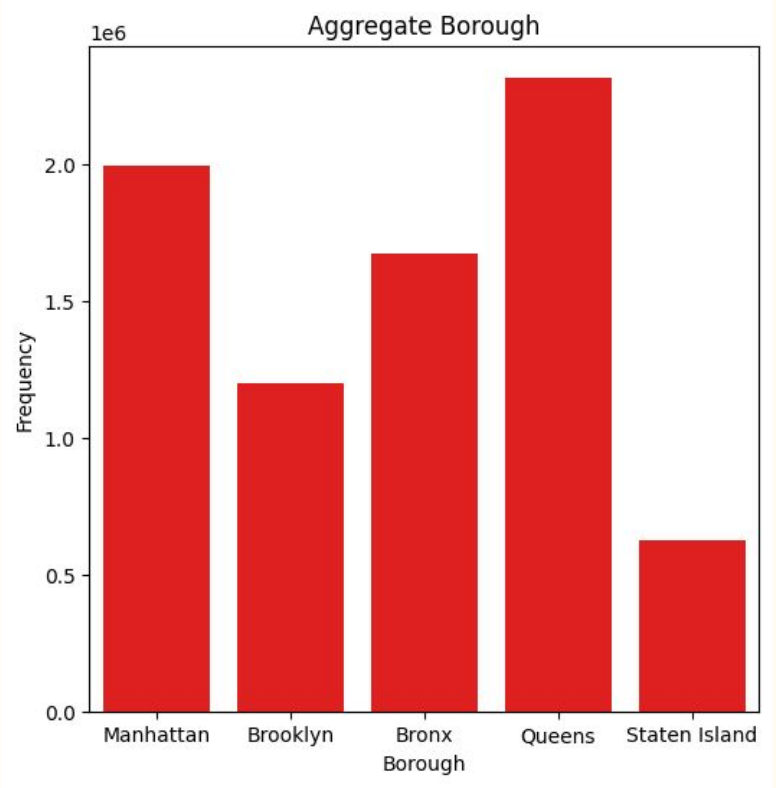
COVID Lockdown!

Peak during Aug-Oct

Gradual increase Over the years

Parking Violations Popular precincts





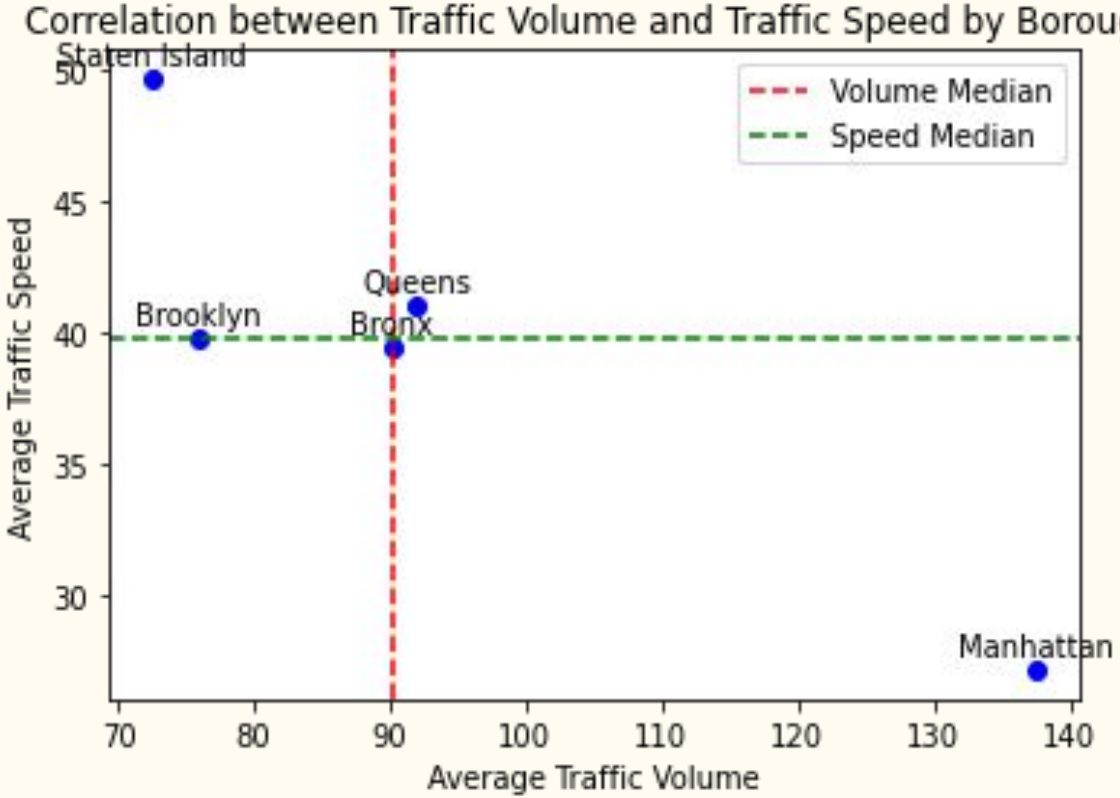
Streetwise Correlations between Traffic Dynamics

Correlation values	
Speed vs Volume	0.054
Volume vs Injuries	0.039
Injuries vs Volume	-0.013
(Speed x Volume) vs Injuries	-0.029

Karl Pearson's Coefficient

$$R = \frac{\sum_{i=1}^{N_s}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}^{N_s}(x_i - x)^2(y_i - y)^2}}$$

- Derived Table which stores the avg speed and volume, total injuries for each street
- Finding the R value using a HiveQL query on the Derived Table

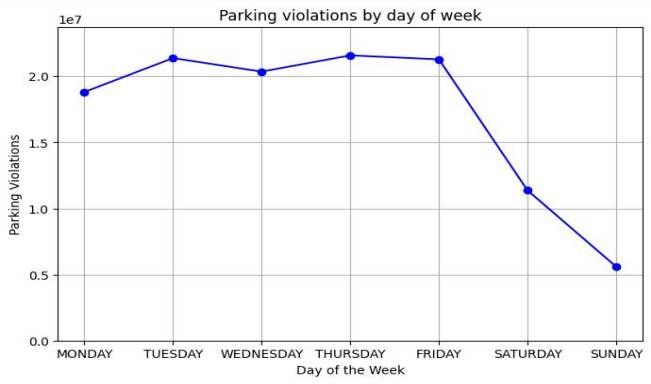


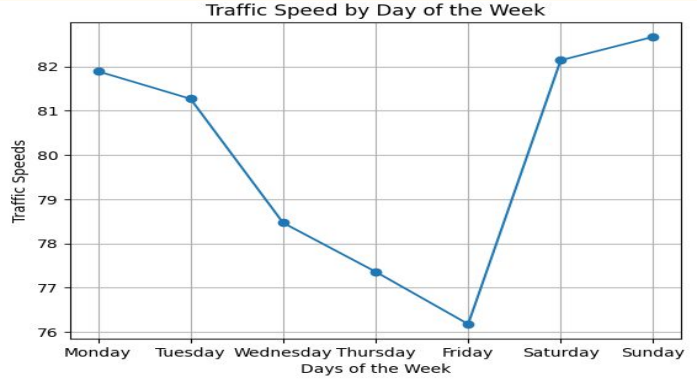
Correlation -
Speed vs Volume (for Boroughs):

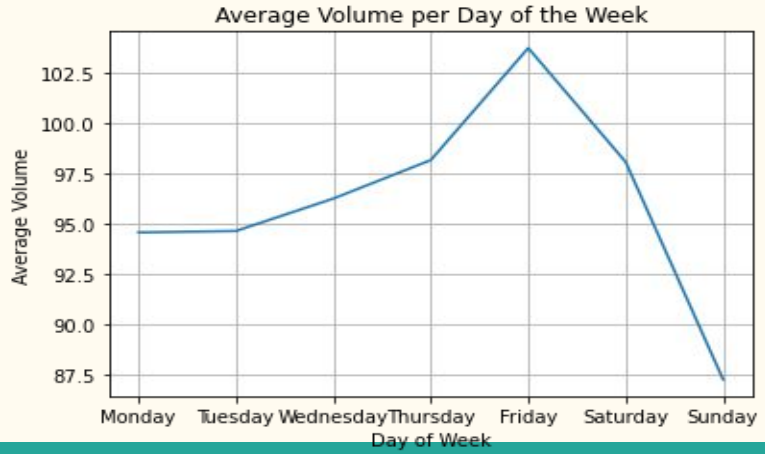
$R = -0.914$

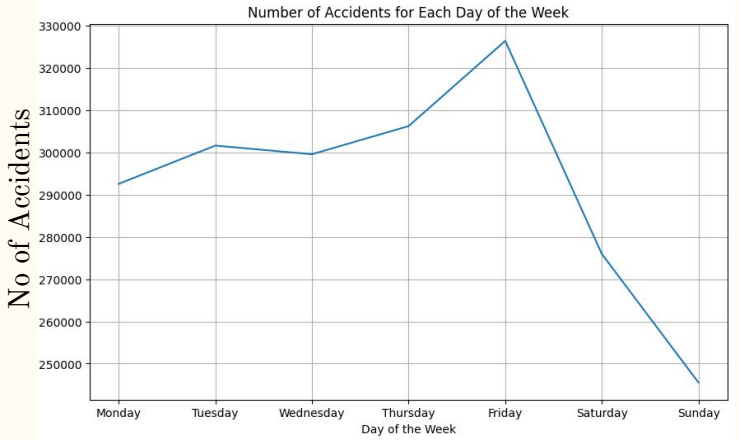
Day Wise Correlations between Traffic Dynamics

Volume - Speed: -0.756	**Volume - Injuries**: 0.856	**Volume - Parking Violations**: 0.658	**Speed - Injuries**: -0.811



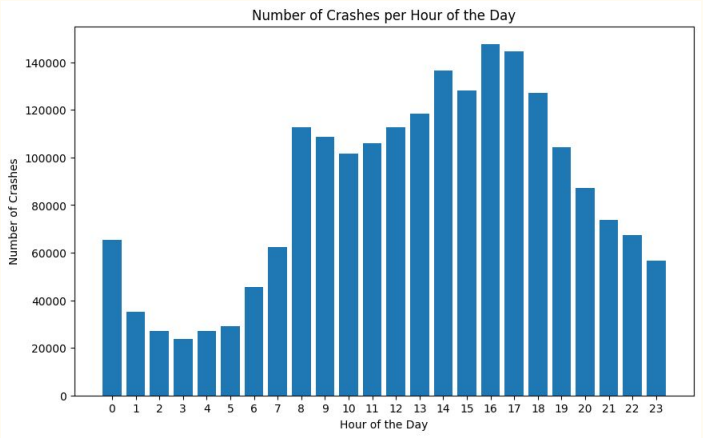


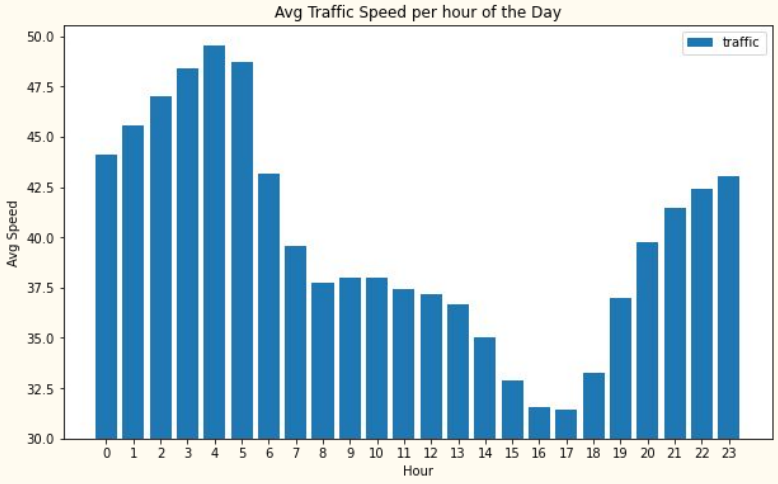


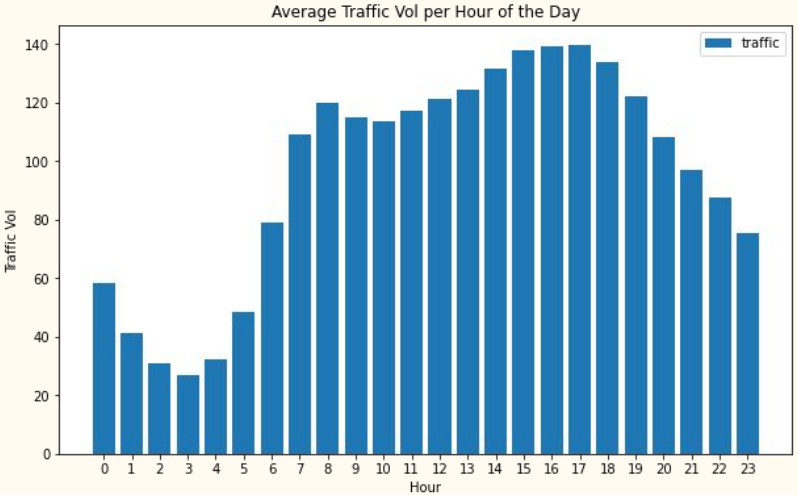


Hour Wise Correlations between Traffic Dynamics

Injuries - Volume	0.948
Speed - Volume	-0.966
Speed - Injuries	-0.972







Obstacles

1. Non-Uniform Street Sampling by Borough: Streets sampling per borough was non-uniform over different datasets which might have lead to some biases.
2. Inconsistent Number of Common Streets: Number of common streets between all the dataset were not equal. Some data was collected over a smaller number of streets.
3. Diverse Time Periods for Datasets: Different datasets have data collected over different time periods.

All can lead to data imbalance

Summary

- Correlation does not imply causation - Traffic speed has a negative correlation with Traffic incidents
- High traffic speeds are linked with low volumes and low collisions
- Increased traffic volume is associated with significantly more collisions
- Traffic dynamics varied unexpectedly during the weekends.
- Parking violations are increasing at much higher pace than traffic volume with certain precincts in Manhattan and Queens being important flashpoints

Acknowledgement

- New York Department of Transport for proving the datasets
- NYU HPC Team for providing the clusters
- Prof. Yang Tang for his guidance and support throughout the semester

References

[1] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

[2] https://data.cityofnewyork.us/Transportation/Automated-Traffic-Volume-Counts/7ym2-wayt

[3] https://data.cityofnewyork.us/City-Government/Parking-Fiscal-Year-2023/869v-vr48

[4] https://www.nyc.gov/site/nypd/bureaus/patrol/find-your-precinct.page

[5] https://www.sciencedirect.com/science/article/pii/S0022437522001098

Thank You!