Likelihood and Correlation Estimation of NYC Parking Violations: A scalable approach using PySpark

Maria Mahbub[1] and Tasmia Rahman[2]

Bredesen Center, UTK & ORNL[1], EECS, UTK[2]

Motivation

- Where and when are tickets most likely to be issued?
- What type of codes are mostly violated?
- What are the most common color and types of cars to be ticketed?
- How much NYC is earning from parking violations?

Data Sets

- NYC Parking Violations Data- Fiscal year 2019*:
  - 3.95 million data points with 43 dimensions
- Supporting Dataset: 99 Parking Violation Codes* with fine amount and descriptions

*Provider: US Department of Finance

Data Analysis Workflow (Scalable)

Data Preprocessing

Tool: Pandas

NYC Open Data (Raw Data)

Fine Amount (Supporting Data)

- Filter missing values
- Fix inconsistencies
- Remove irrelevant attributes
- Split the time
- Incorporate fine data
- Sample randomly from the dataset

50000 Sample Data (Prepared Data)

Data Analysis

Platform , Framework & Tools: JetStream, Apache PySpark, Pandas

Parallelization (PySpark context to build rdds)

- Likelihood Analysis:
  - Observe frequency of factors using MapReduce that scales to larger dataset
- Pattern Analysis:
  - Build maps for correlation analysis of different features
  - Generate likelihood correlations

Data Visualization

Tool: Matplotlib

Visualization

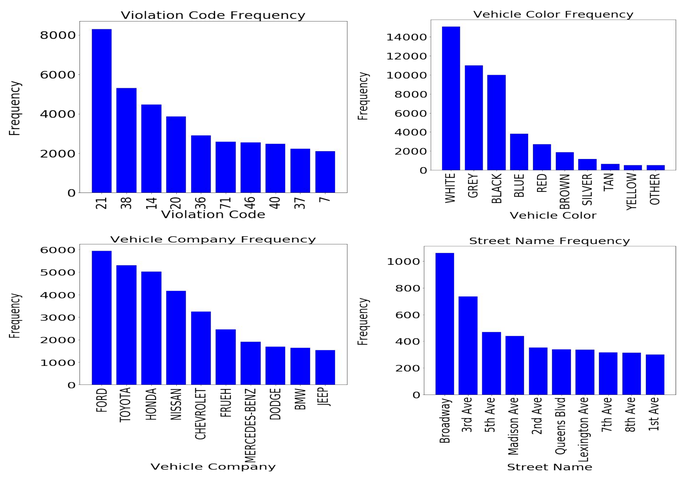
- Likelihood Analysis
  - Bar Plot
  - Pie Chart
- Pattern Analysis
  - Scatter Plot

Analysis and Findings

Likelihood Analysis

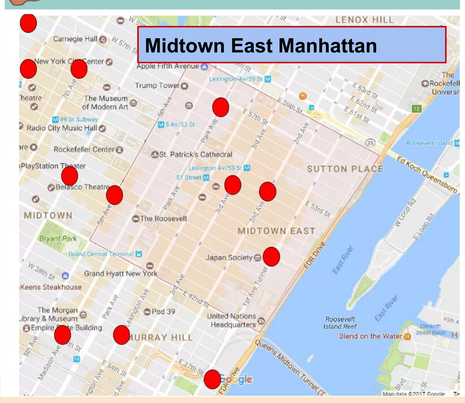
➢ **Which codes seem to get violated the most?**
➢ **Are there some common car companies that get tickets issued frequently, or color?**
➢ **Is there any risky street?**

- Most occurred violation codes, car companies and streets are analyzed with Map Reduce
- Mostly violated
  - **Code**: 21,38,14
    - **21: No parking sign violation**
    - **38: Failing to show a receipt or tag in the windshield.**
    - **14: Standing not allowed**
  - **Street:** Broadway, 3rd Ave, 5th Ave
  - **Vehicle Company :** Ford, Toyota, Honda
  - **Vehicle Color:** White, Grey and Black

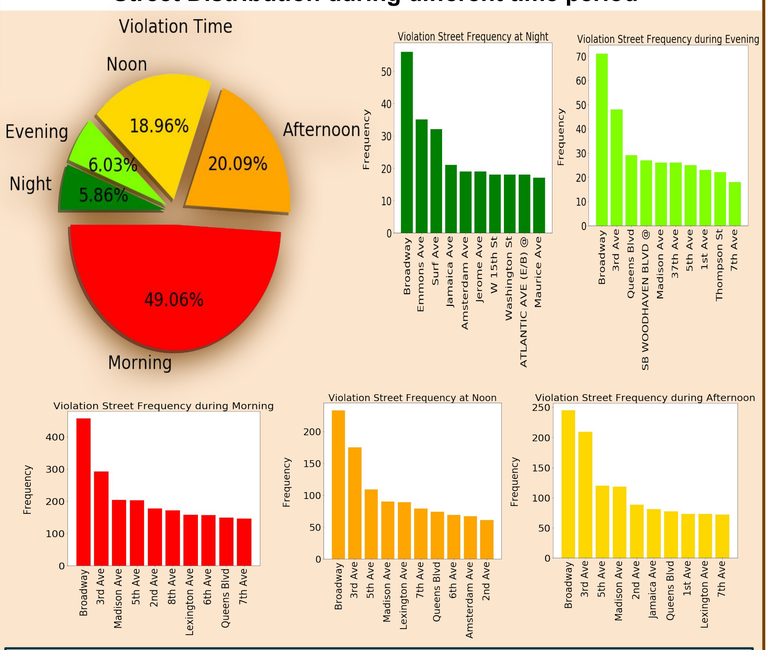


➢ **Can we estimate the time of the day when tickets are most likely to be issued?**
➢ **If we can what are their relationship with any specific street?**

Map of Boroughs of New York City



Street Distribution during different time period

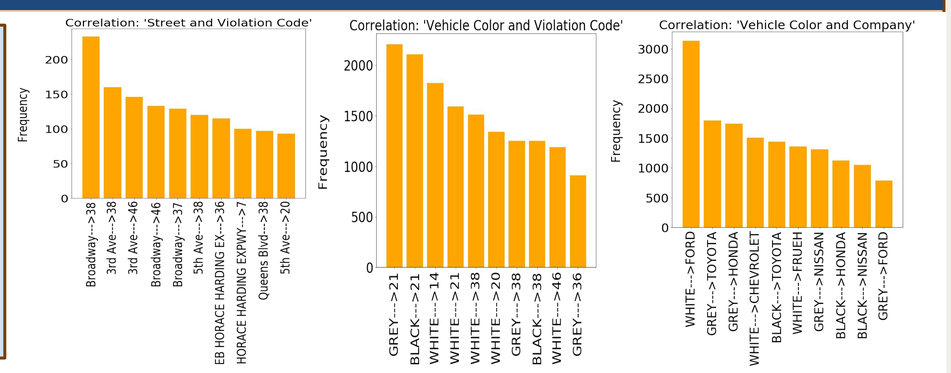


- **Morning is the most busy time, hence parking codes are most likely to get violated during this time**
- **Broadway is the riskiest zone regardless of time period**

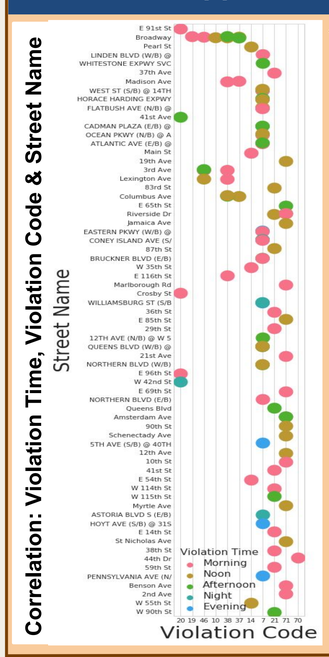
Correlation Analysis

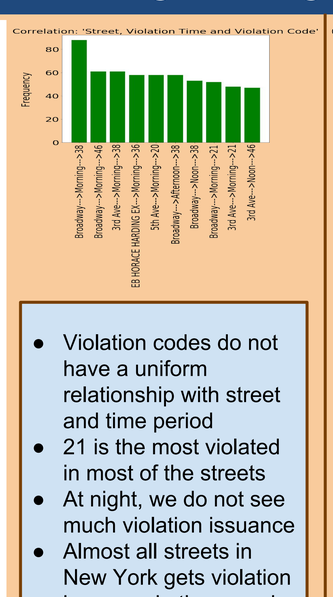
➢ **What happens when 2 categories are paired? Does it make any difference?**

- Code 21 is most violated when paired with a color
- When paired with street, 38 seems to be most violated
- White, grey colors prove to remain most likely even when paired with others

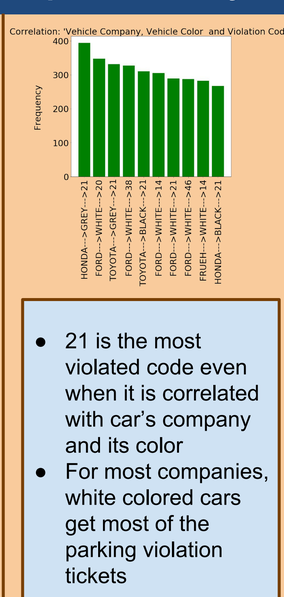


➢ **What happens when 3 categories are grouped? Are they somewhat correlated?**

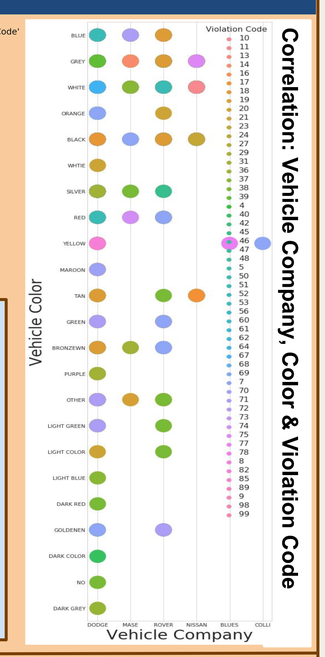


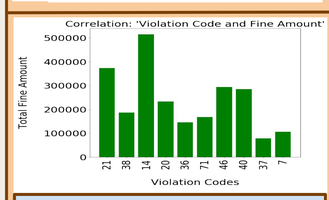


- Violation codes do not have a uniform relationship with street and time period
- 21 is the most violated in most of the streets
- At night, we do not see much violation issuance
- Almost all streets in New York gets violation issuance in the morning



- 21 is the most violated code even when it is correlated with car's company and its color
- For most companies, white colored cars get most of the parking violation tickets

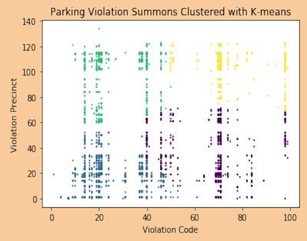




- **Fine amount for code 14 is higher than that of code 21 ---> hence more revenue from code 14**

Future Work:

- Add dimensions to the clustering model
- Build prediction algorithm to predict parking violations based on other features
- Group similar violation codes



Conclusion:

- Likelihood analysis of the features reveals the frequency of violation
- Pattern analysis can demonstrate a correlation among the features
- Incorporating fine amount have also provided insights on the revenue generation

THE UNIVERSITY OF TENNESSEE KNOXVILLE