D.C. Parking Tickets

By: Reed, Boris, Jason, Karan

Tickets are the worst!

You woke up today with a smile on your face, ready to take on the day. You have all these things to get done, and are thankful that you own a car in the city so you don't have to rely on the schedule of public transportation. You arrive at your car to find a little pink slip attached to your windshield; it's a parking ticket. You pull it out and take a look; your heart sinks. We've all been there. Getting a parking ticket is a sure-fire way to bring down your mood, and potentially your bank account balance.

Problem Statement

- We hope to be able to find trends that might decrease your chances of getting a parking ticket or having your fine dismissed if you get a ticket
- Using data from Open Data DC on parking violations issued from July 2017 - July 2019 in DC, we want to develop a classification model that will predict the likelihood of having a ticket being liable or dismissed.
- We will gridsearch over machine learning algorithms such as KNN, Logistic Regression, and k-means clustering to develop the most robust model that optimizes predictive accuracy.
- We will also perform EDA on the categorical features in the dataset to try and unearth trends that might reduce the chances of getting a parking ticket.

Data Cleaning/ Pre-Processing

- Concatenated monthly datasets on parking violations over a 2 year period from july 2017 - july 2019 from Open Data DC.
- Dropped all columns with over 80% nulls, and columns that were not relevant to our analysis.
- Extracted quadrant data from location column to create individual one-hot encoded quadrant categories
- Then reverse one-hot encoded using .idxmax() method to get categorical column of all the quadrants.
  - Dropped rows with missing quadrant data or more than one quadrant represented.

Disclaimer - data cleaning

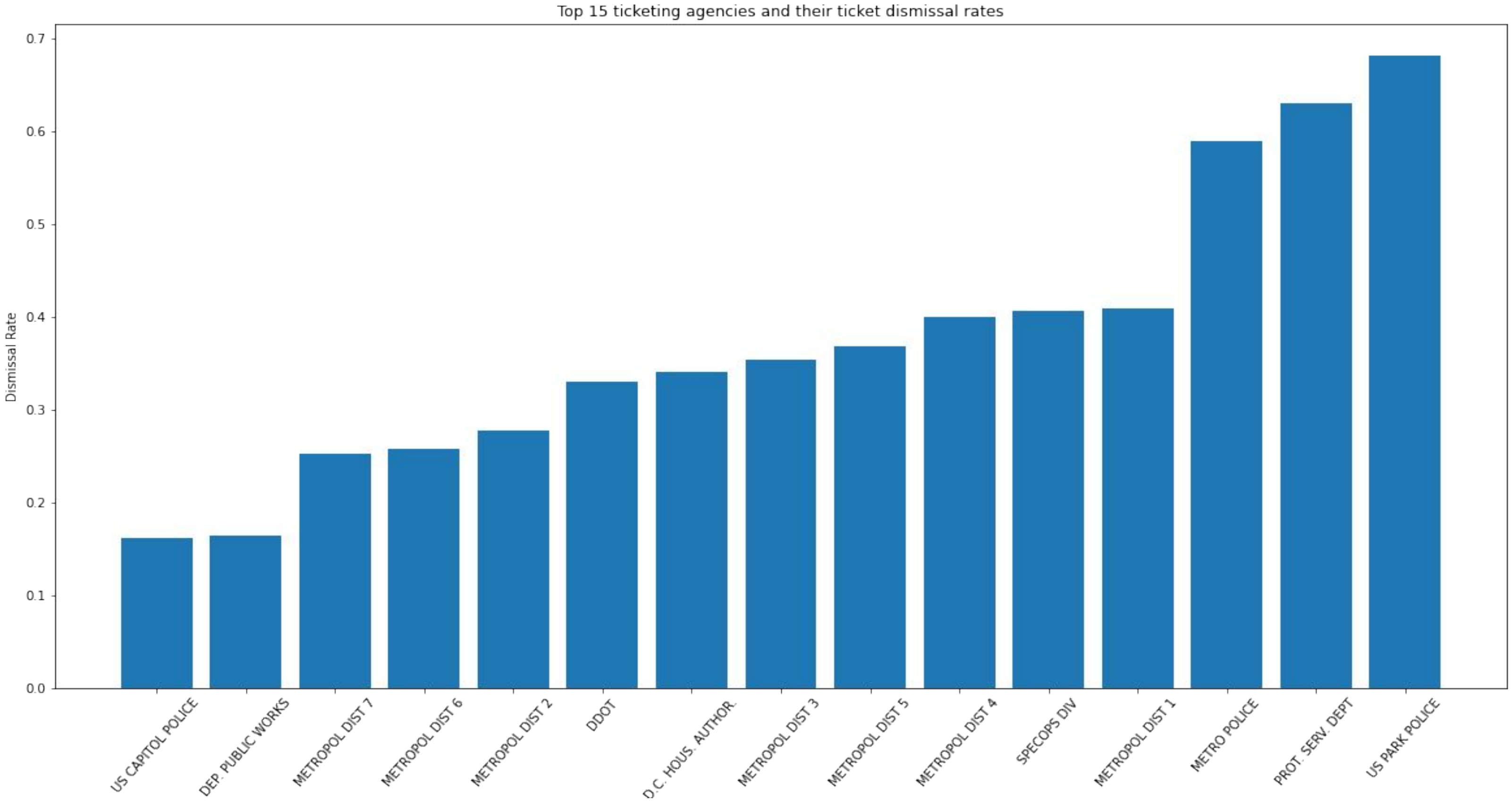
Without a proper data dictionary, it is difficult to decipher what some of the classes in the disposition_type column mean. A request was made out to the open data dc offices for a data dictionary on this dataset; however, based on responses it doesn't seem they will be able to get one to us in time to complete this project by the due date.

Therefore, we will attempt to perform some empirical analysis on the dispostion_type category to see if we can spot some trends that will help us in identifying the different classes such as 'other', 'void', 'continued', and 'administrative' in this dataset.

For further research on disposition-type, we will be dropping all rows containing 'other', 'void', 'continued', and 'administrative' as their disposition_type, and grouping together the 'liable', 'liable_system', and 'liable - traffic' into one 'liable' class. We should then be left with only two classes in the dispostion_type column: 'liable' and 'dismissed'.¶

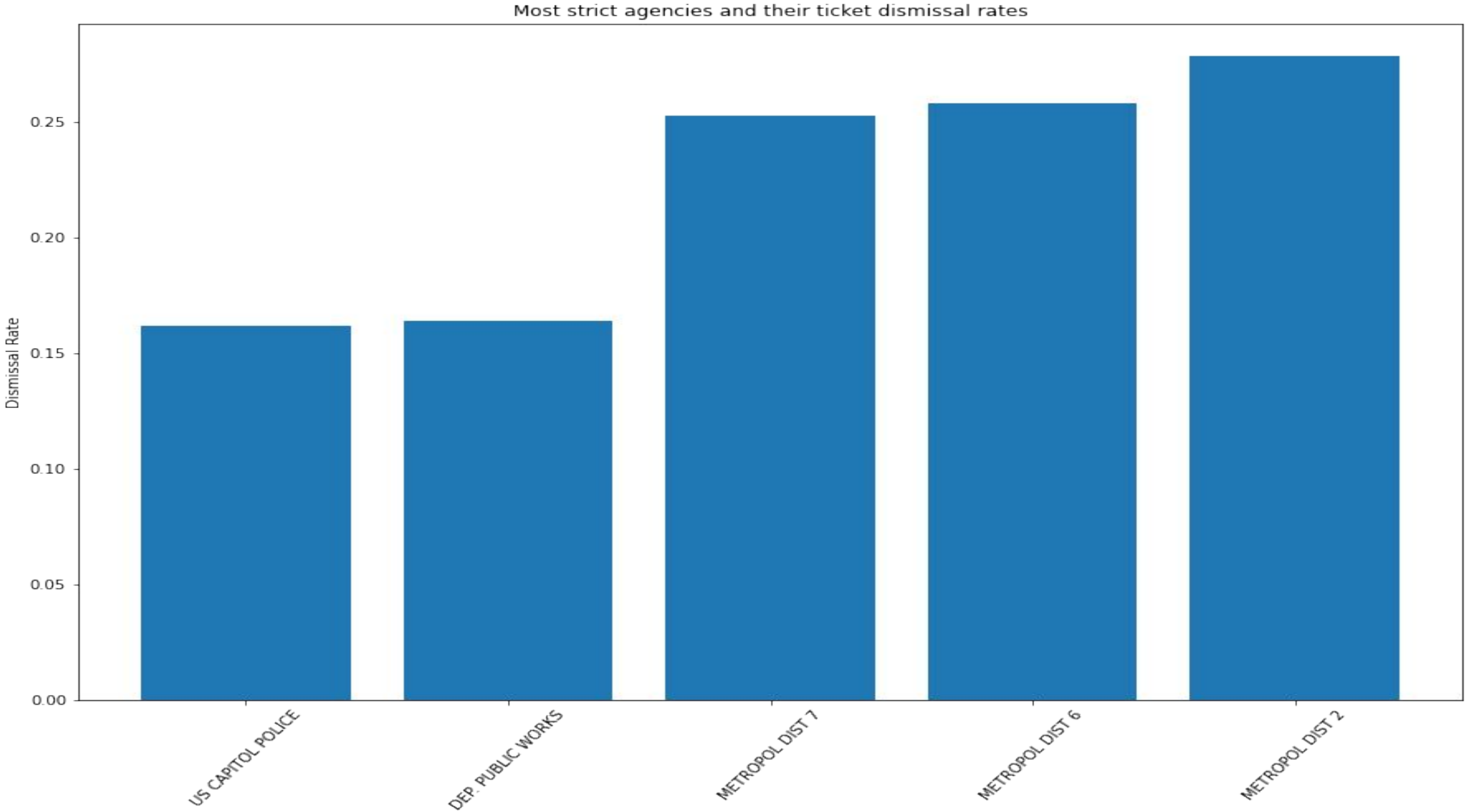
We understand that dropping a large amount of data from our dataset like this is likely to introduce some bias into our model that may affect the accuracy of any predictions or inferences made from our model. However, given the scope and time constraints of this project, we feel it is necessary.

Agencies



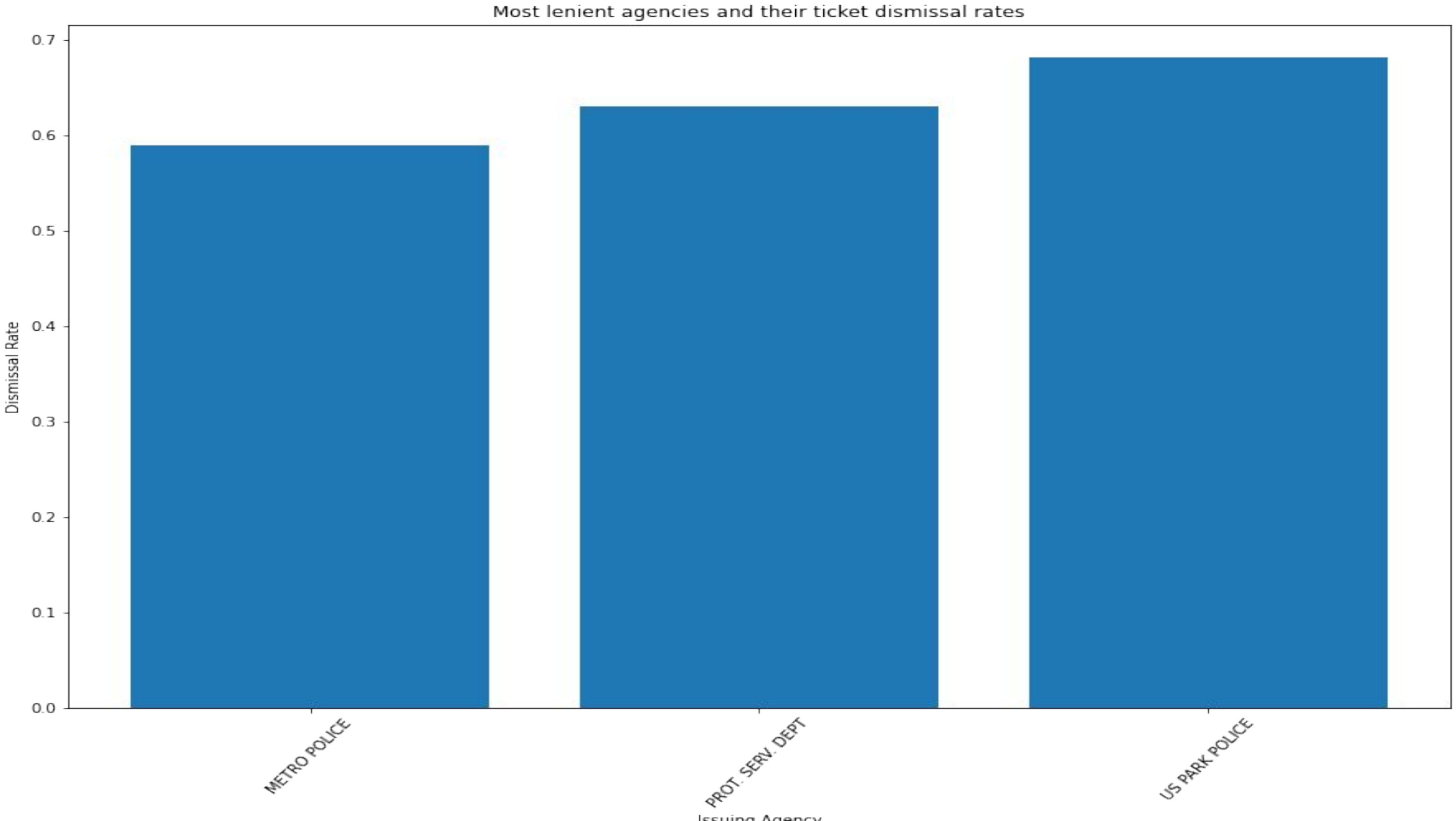
Top 15 ticketing agencies and their ticket dismissal rates

Most Strict



Most strict agencies and their ticket dismissal rates
Dismissal Rate
0.25
0.20
0.15
0.10
0.05
0.00
US CAPITOL POLICE
DEP. PUBLIC WORKS
METROPOL DIST 7
METROPOL DIST 6
METROPOL DIST 2

Most Lenient



Models

- Logistic Regression, KNN Classifier, Clustering
- Baseline Rate: Liable - 81.99%   Dismissed - 18.01%
- Logistic Regression
  - Score on training data: 0.8193
  - Score on testing data: 0.8196
  - Low variance
- KNN Classifier
  - Data was too large so used a sample of the data
  - Score on training data: 0.8264
  - Score on testing data: 0.8102
  - Low variance

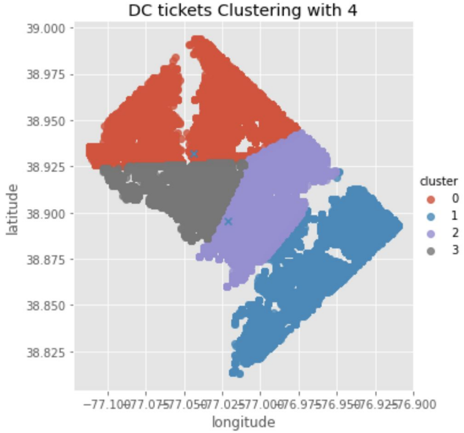
Data EDA

- The original dataset contains more than 2 million data and the cleaned dataset contains close to 800,000 data have very similar distributions.

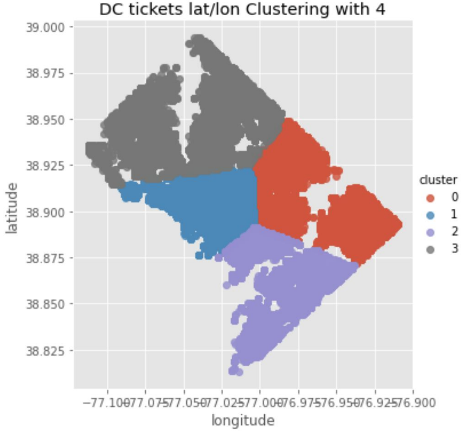
- 4 clusters clustering for two datasets
  - Cluster 0 in the original dataset and cluster 3 in cleaned dataset are at similar locations, they are both the third place for sum and average fine-amount and total-paid amount.
  - Cluster 3 in the original dataset and cluster 1 in the cleaned dataset are also having similar location and similar sum and average fine-amount. (Highest sum but lowest average) -- Which implies that this location has very large number of tickets.
- 5 clusters clustering for datasets
  - 5 clusters do not have the identical cluster for both datasets. But there are also interesting similarities between the locations.
  - Cluster 4 of the original dataset is at a similar location as cluster 3 in the 4 clustering, so it has the similar sum and average fine-amount and total-paid amount(highest sum but lowest average).
  - Cluster 4 and 3 of the trained dataset are in the similar location as cluster 2 and 3 in the 4 clustering . (low sum but high average)

Data Clustering

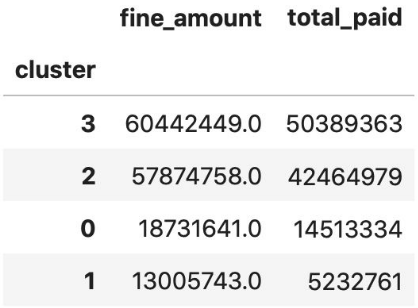
Clustering with all 2 Million data



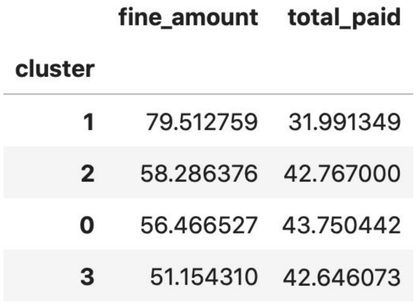
Clustering with Selected 896755 data



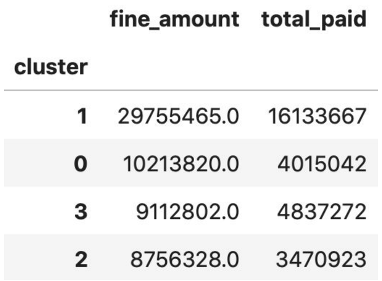
Sum and average of fine-amount and total-paid for each cluster



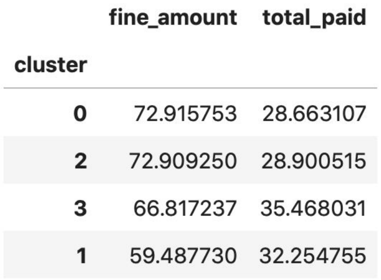
cluster	fine_amount	total_paid
3	60442449.0	50389363
2	57874758.0	42464979
0	18731641.0	14513334
1	13005743.0	5232761



cluster	fine_amount	total_paid
1	79.512759	31.991349
2	58.286376	42.767000
0	56.466527	43.750442
3	51.154310	42.646073



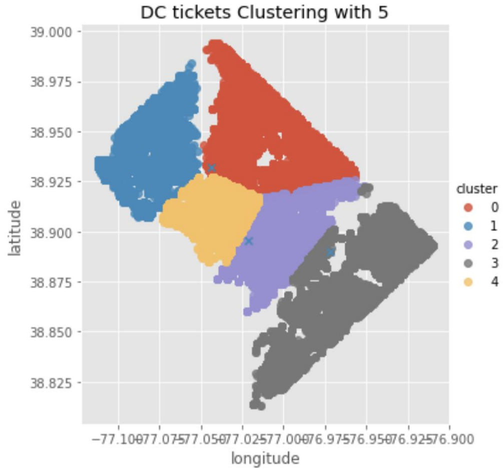
cluster	fine_amount	total_paid
1	29755465.0	16133667
0	10213820.0	4015042
3	9112802.0	4837272
2	8756328.0	3470923

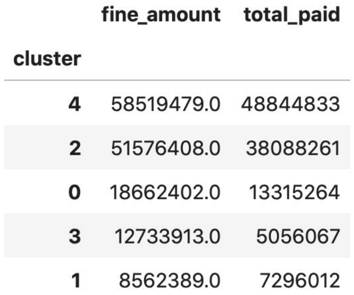


cluster	fine_amount	total_paid
0	72.915753	28.663107
2	72.909250	28.900515
3	66.817237	35.468031
1	59.487730	32.254755

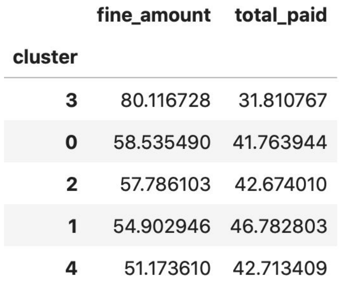
Data Clustering

Clustering with all 2 Million data





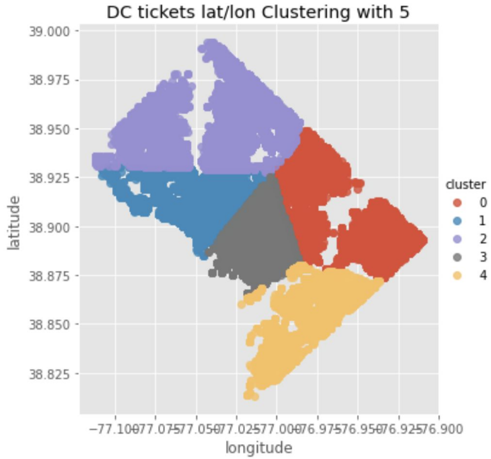
	fine_amount	total_paid
cluster		
4	58519479.0	48844833
2	51576408.0	38088261
0	18662402.0	13315264
3	12733913.0	5056067
1	8562389.0	7296012

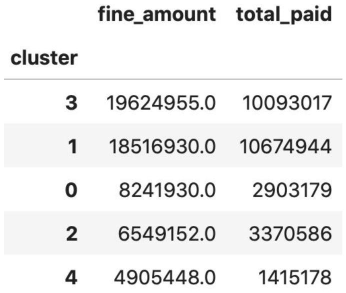


	fine_amount	total_paid
cluster		
3	80.116728	31.810767
0	58.535490	41.763944
2	57.786103	42.674010
1	54.902946	46.782803
4	51.173610	42.713409

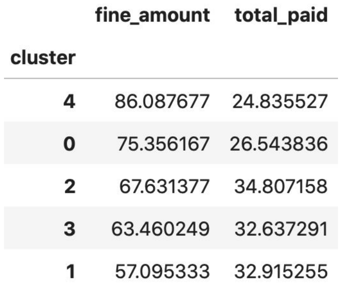
Sum and average of fine-amount and total-paid for each cluster

Clustering with Selected 896755 data



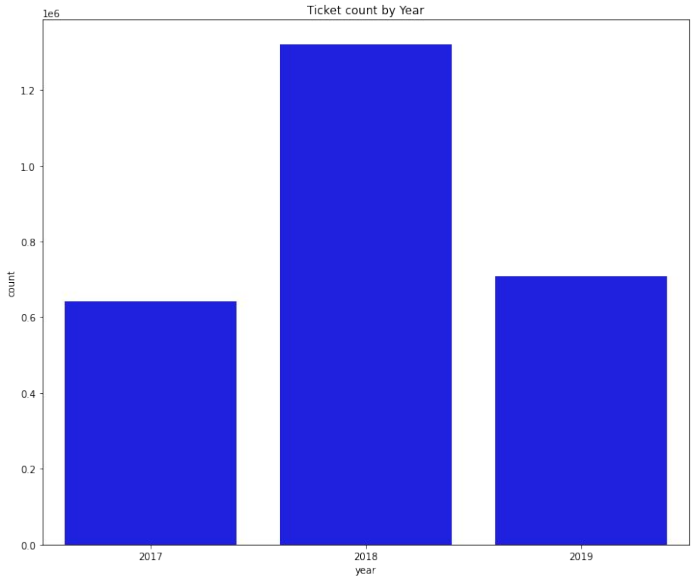


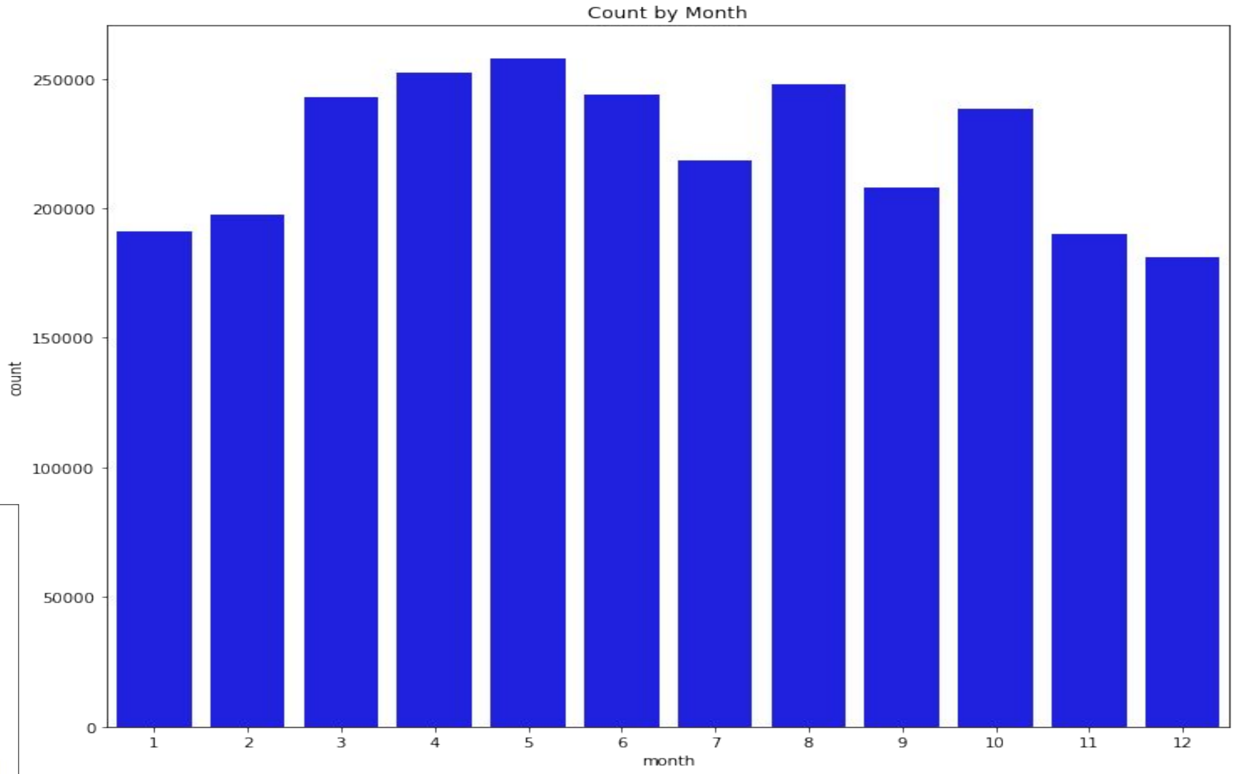
	fine_amount	total_paid
cluster		
3	19624955.0	10093017
1	18516930.0	10674944
0	8241930.0	2903179
2	6549152.0	3370586
4	4905448.0	1415178



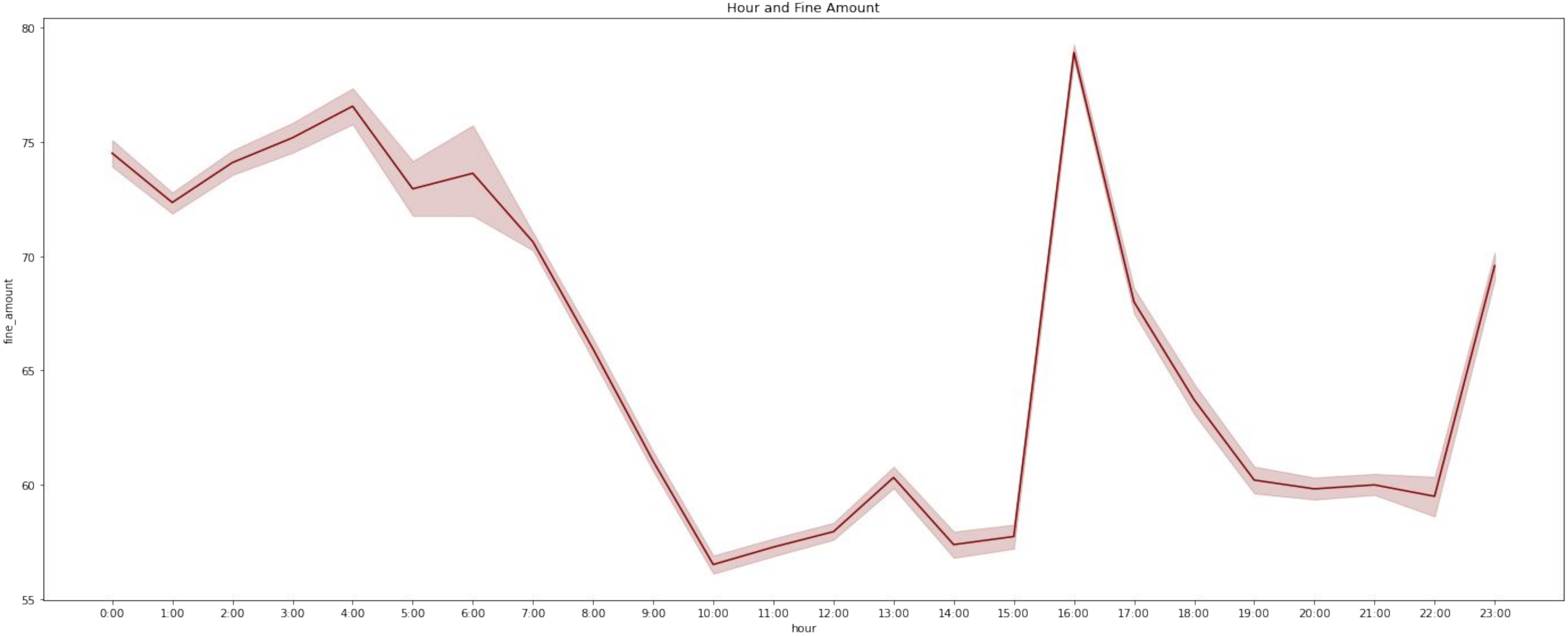
	fine_amount	total_paid
cluster		
4	86.087677	24.835527
0	75.356167	26.543836
2	67.631377	34.807158
3	63.460249	32.637291
1	57.095333	32.915255

Time

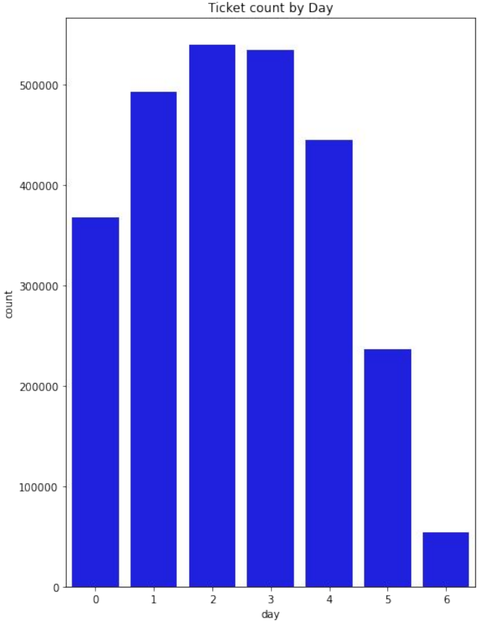

Ticket count by Year
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
count
2017
2018
2019
year

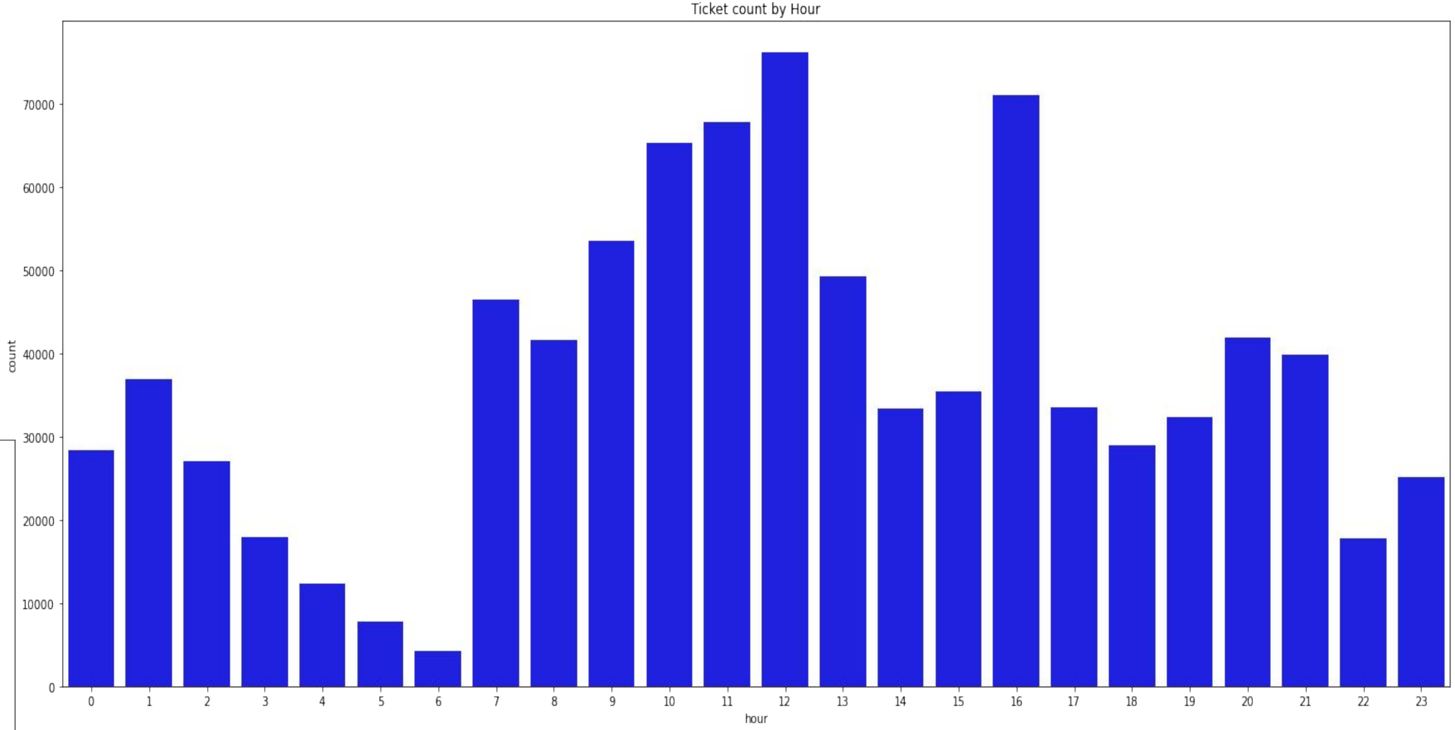

Count by Month
250000
200000
150000
100000
50000
0
count
1
2
3
4
5
6
7
8
9
10
11
12
month

Time and Fine Amounts

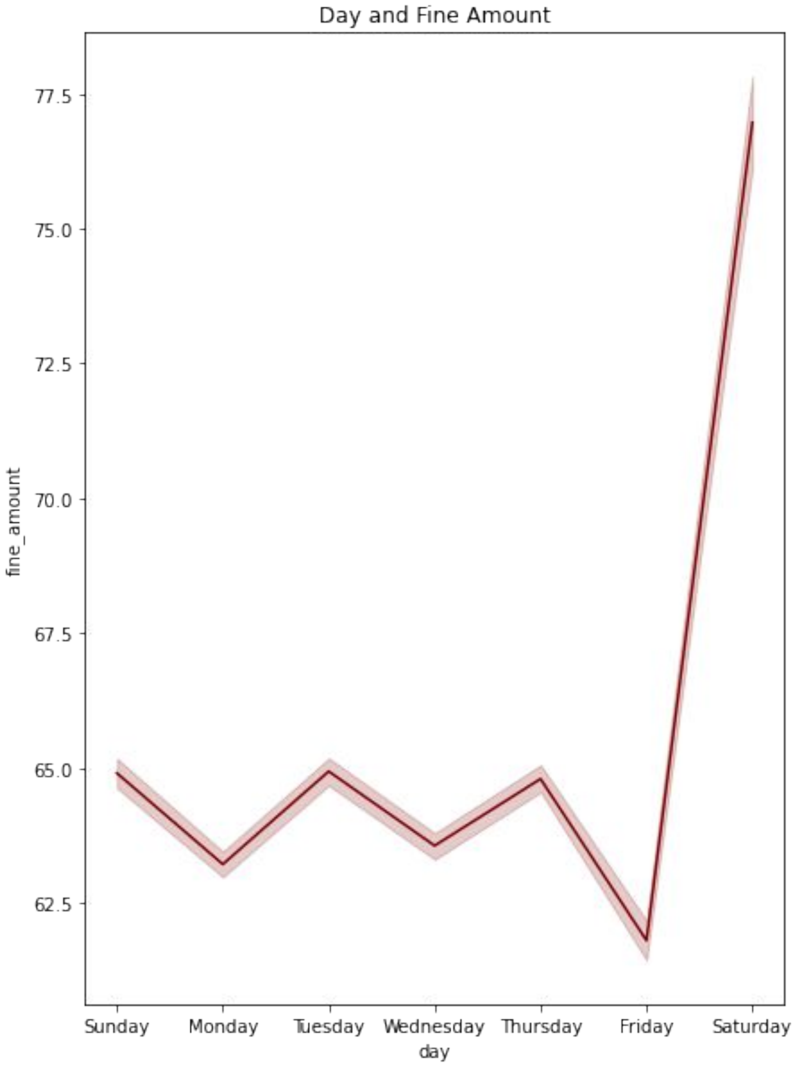


Time

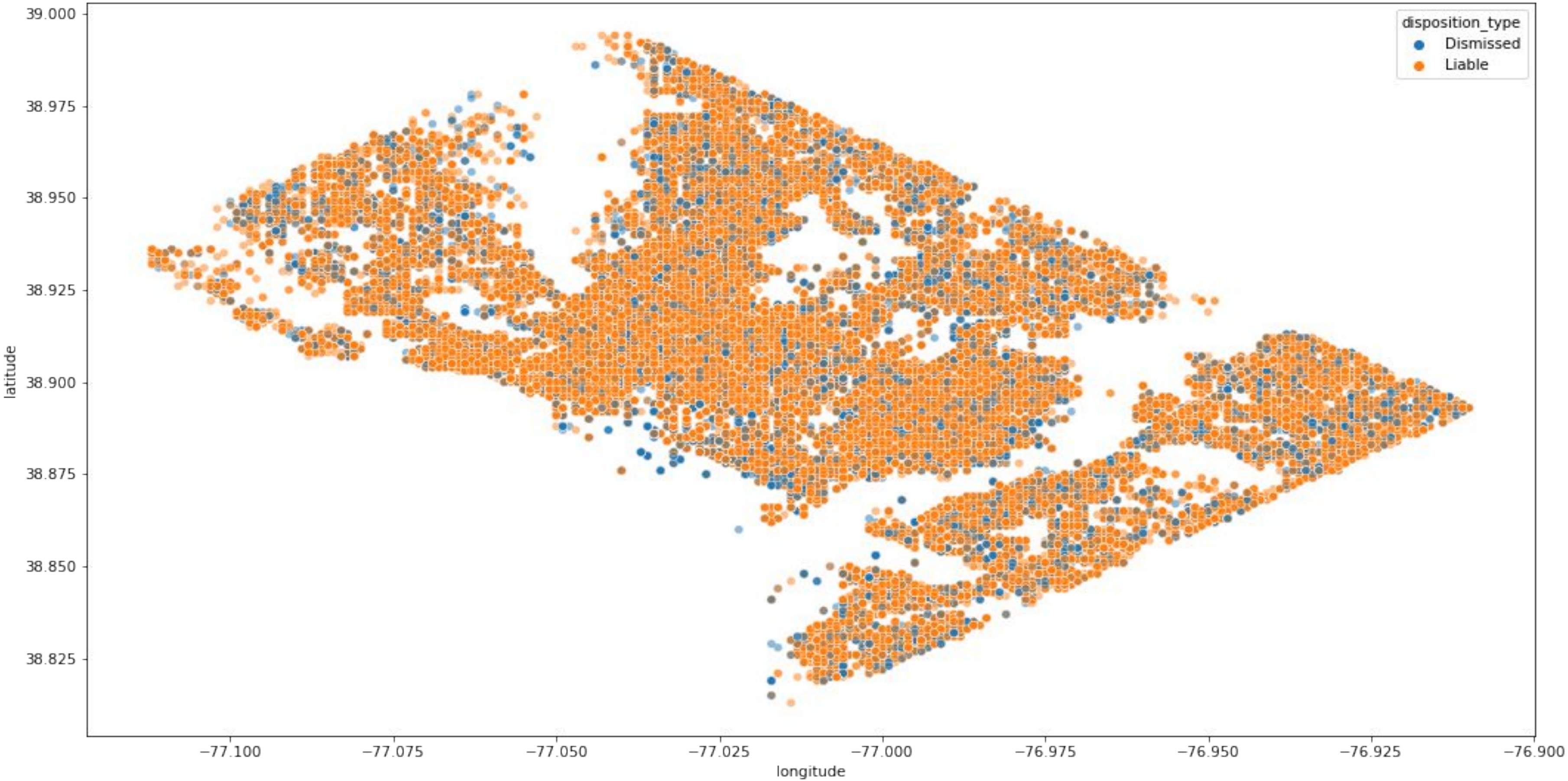

Ticket count by Day
count
0
100000
200000
300000
400000
500000
day
0
1
2
3
4
5
6

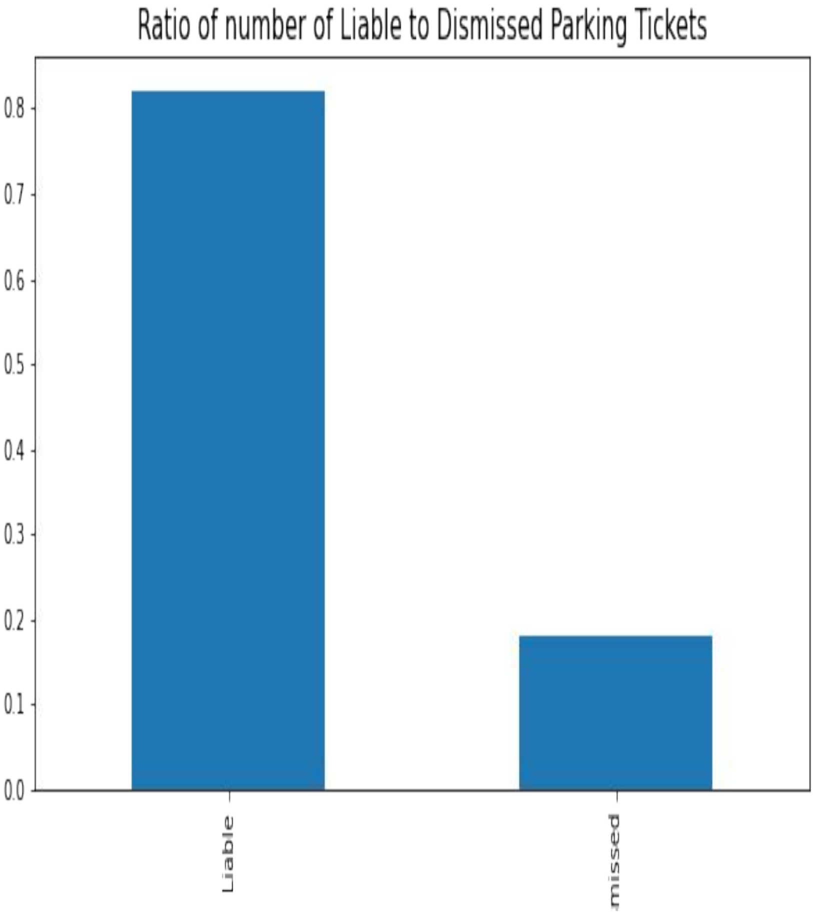

Ticket count by Hour
count
0
10000
20000
30000
40000
50000
60000
70000
hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Time and Fine Amounts

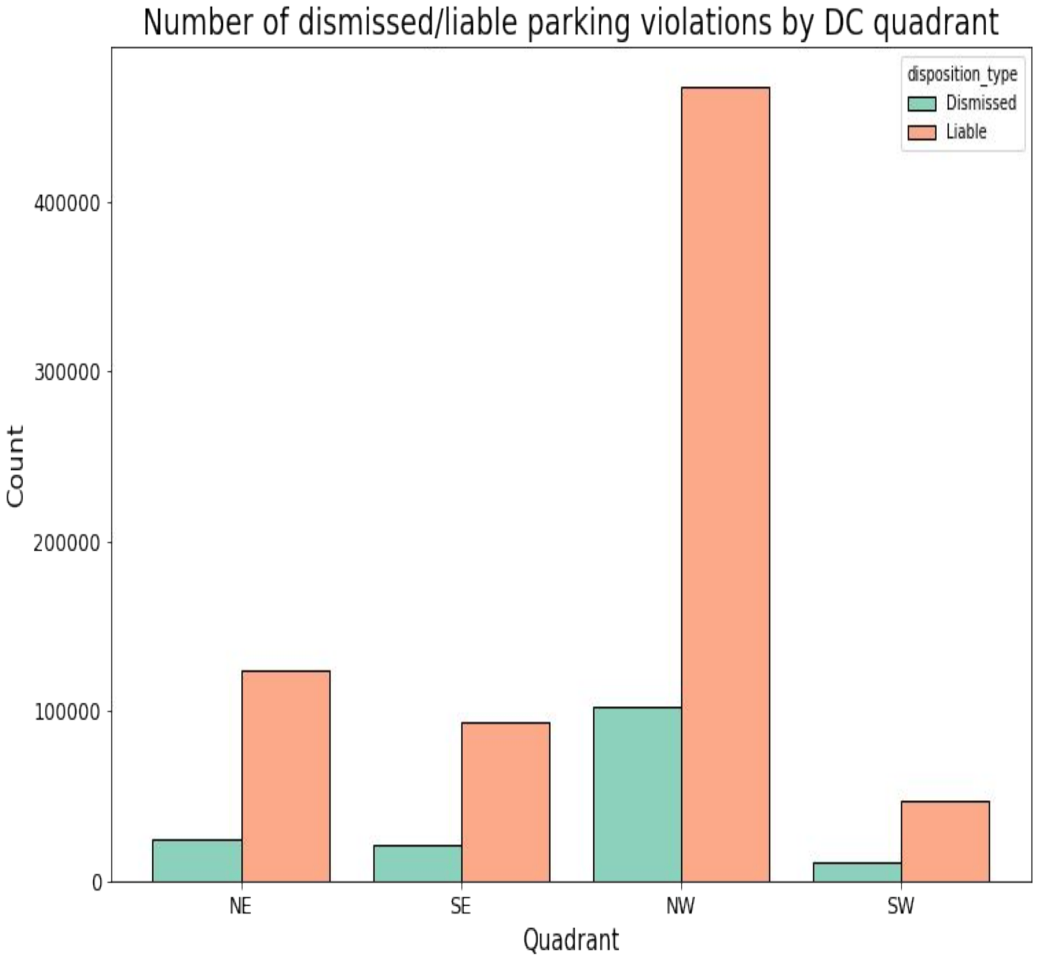


Location


disposition_type
● Dismissed
● Liable
39.000
38.975
38.950
38.925
38.900
38.875
38.850
38.825
latitude
−77.100
−77.075
−77.050
−77.025
−77.000
−76.975
−76.950
−76.925
−76.900
longitude



Ratio of number of Liable to Dismissed Parking Tickets



Number of dismissed/liable parking violations by DC quadrant

Location

10 Interesting things you can get ticketed for:

CODE	DESCRIPTION	FINE AMT	NO. OF TIX
P017	Excessive idling	$500	96
P029	Park with left wheel to the curb	$20	1110
P011	Park more than 12 in from curb	$20	403
P077	Motor Running Unattended	$50	279
P040	Fail to park parallel	$20	163
P300	Fail to turn wheel tom curb	$20	52
P072	Fail to lock and remove key from ignition when pkd	$50	27
P305	Park with 25 ft of a mailbox	$20	18
P107	Excessive smoke	$15	1
P100	Glass left in street	$25	1

Conclusions

Agencies

- Top 5 most strict ticketing agencies: United States Capitol Police, Department of Public Works, Metropolitan Police Department Districts 7, 6, and 2
- Top 5 most lenient ticketing agencies: United States Park Police, Protective Services Department, Metro Police, Metropolitan Police Department District 1, and Special Operation Div & Traffic Div
  - Should you receive a ticket from any one of these agencies, it would be worth contesting, as they are more likely to get dismissed.

Location

- Although dismissal rates were pretty equal among the quadrants (around 18%), the SW quadrant accounted for only 6.5% of total tickets in the dataset, whereas the NW quadrant had the most with 63.9% of total tickets issued.

If we had more time/resources

- Investigate the different variations of liable tickets from the dataset
- Use hyperparameters to gridsearch and fine tune our models
- Classify based on neighborhoods/wards to get more detailed location-based analysis.
- Get more data about features we had to drop due to an abundance of null values, like car make/model.

The End