nyc_taxi_tips

My take on predicting tips on the NYC taxi dataset

Input data

Data Sampling

For the purposes of modeling we keep 20% of each dataset. While a more indepth look was conducted on each file separetely.

Data processing

There are some critical issues on the given dataset, which are adressed as follows:

Issue	Action
Missing values exist in most of attributes	Drop NA
Negative Values on fares,tips amounts,tolls and extra	Select only positive
Wrong dates in the dataset (wrong year and months)	Filter only the ones targeted
Vendor IDs are not concistent across different datasets	Select only

Feature Engineering

To explore more options while modeling there are a number of new features that are being introduced

Feature	Comment
Month, day, hour	Transform PU & DO time of the trip
Duration_s	Duration of the trip measured in seconds
PULocationID / DOLocation	Transform to categorical vector using OHE
Airport,congestion and mta tax Flags	Transform values to a binary flag

Data Insights

![](path/to/image.png)

Chapter 2. Data exploration

Data overview

Spatial Distribution

Spatial information for each trip, is registered using a zone system. Specifically there are 258 zones that cover the five boroughs of New York.

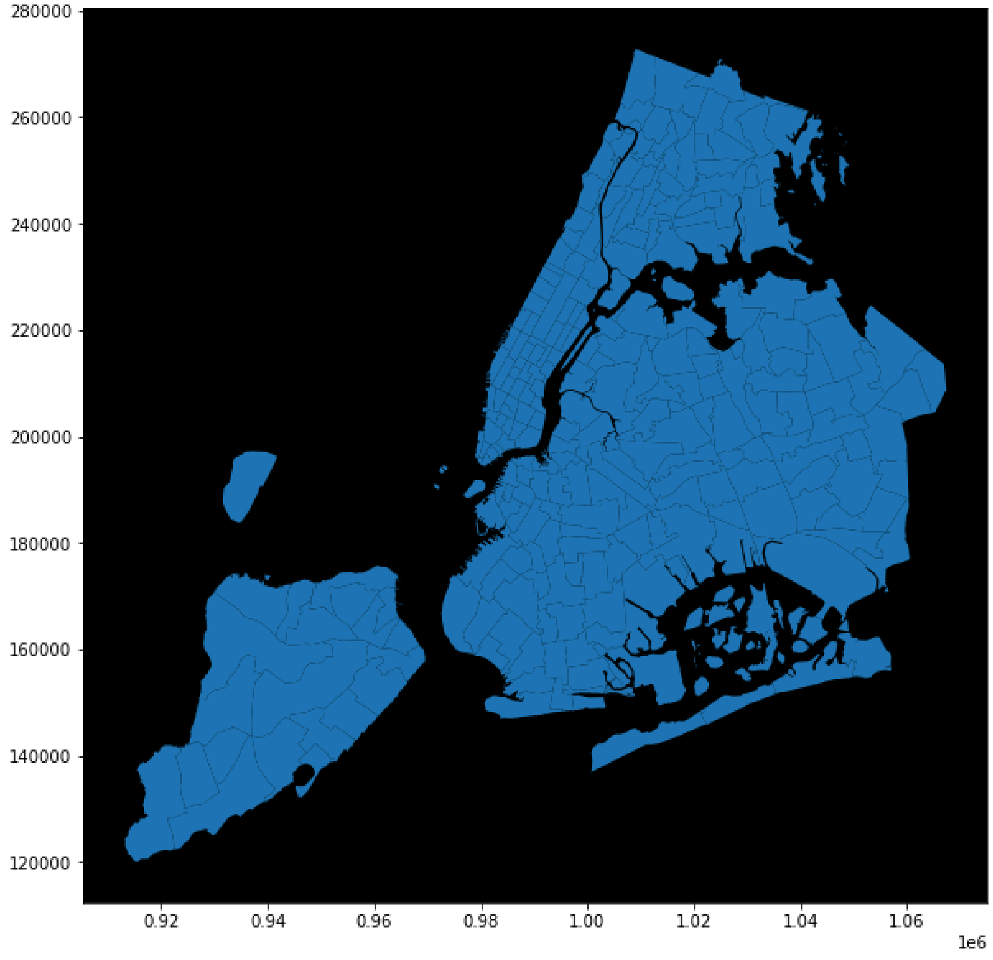


Figure X: Taxi zones.

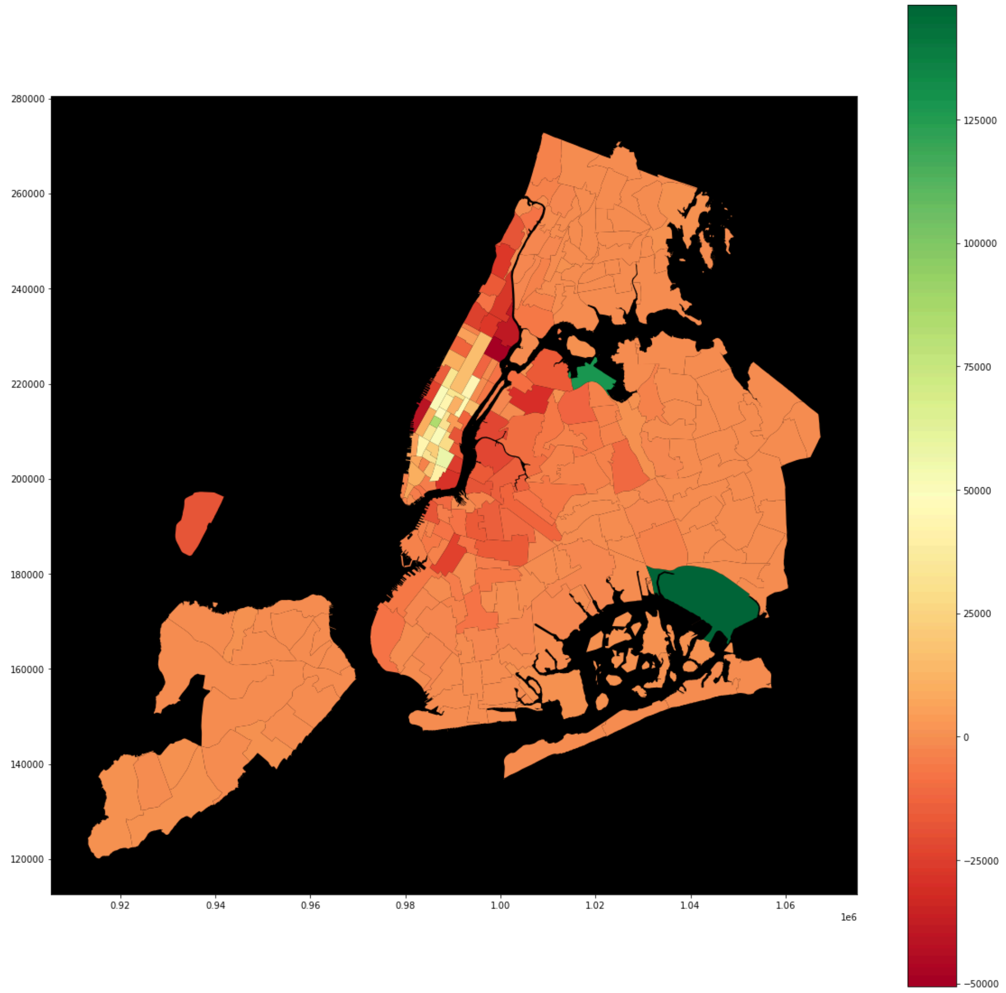


Figure X: Taxi trips. Pickups to Drop-offs difference.

most people take a taxi from the airport to the city, touristic zones as Madison Sq, Theatre District might have more pickups due to their late nature of their activities, so people might go there with a public transport but when they finish their activity it might be late so they choose to take a taxi on their way back...

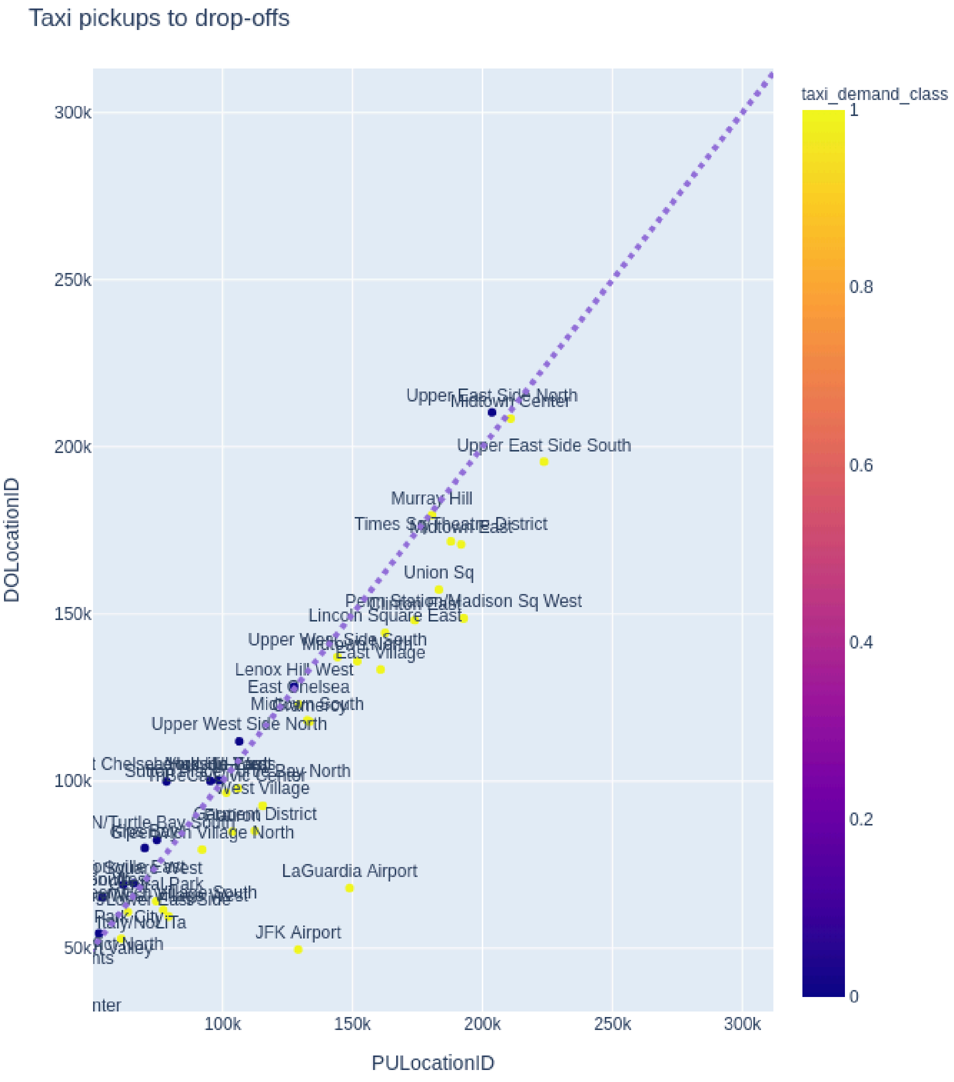


Figure X: Taxi trips. Pickups to Drop-offs difference.

Temporal Distribution

We look on the distribution of trips per month.There is a huge drop on count of trips during November 27 when it is the Thanksgiving Day. After searching for this specific period, during June,2024 there were some significant heatwave and record-breaking temperatures.

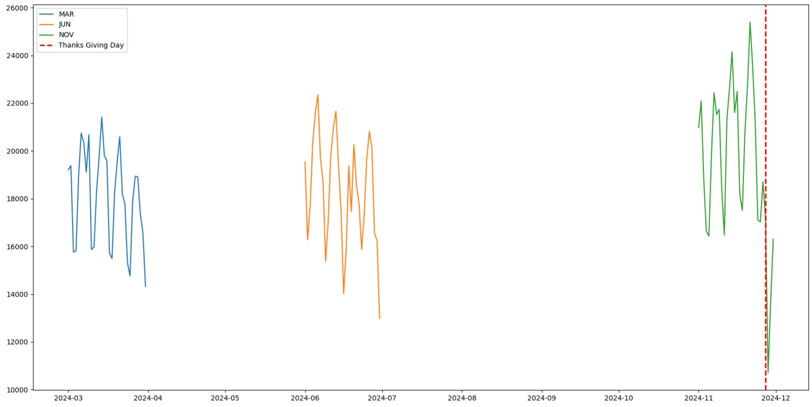
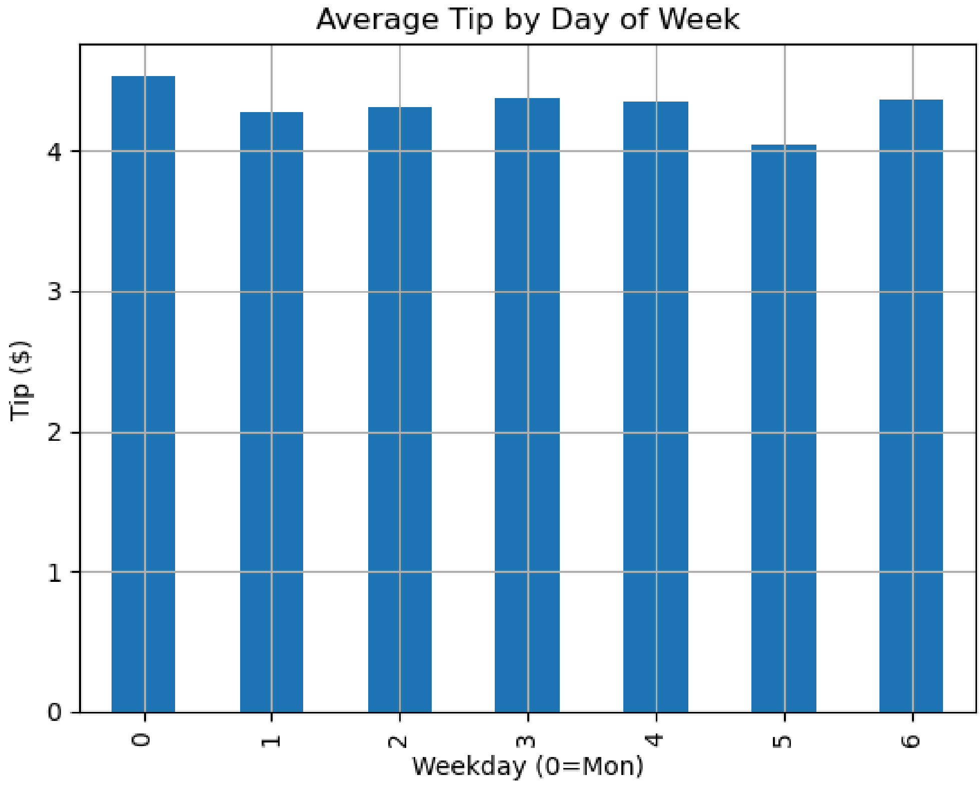
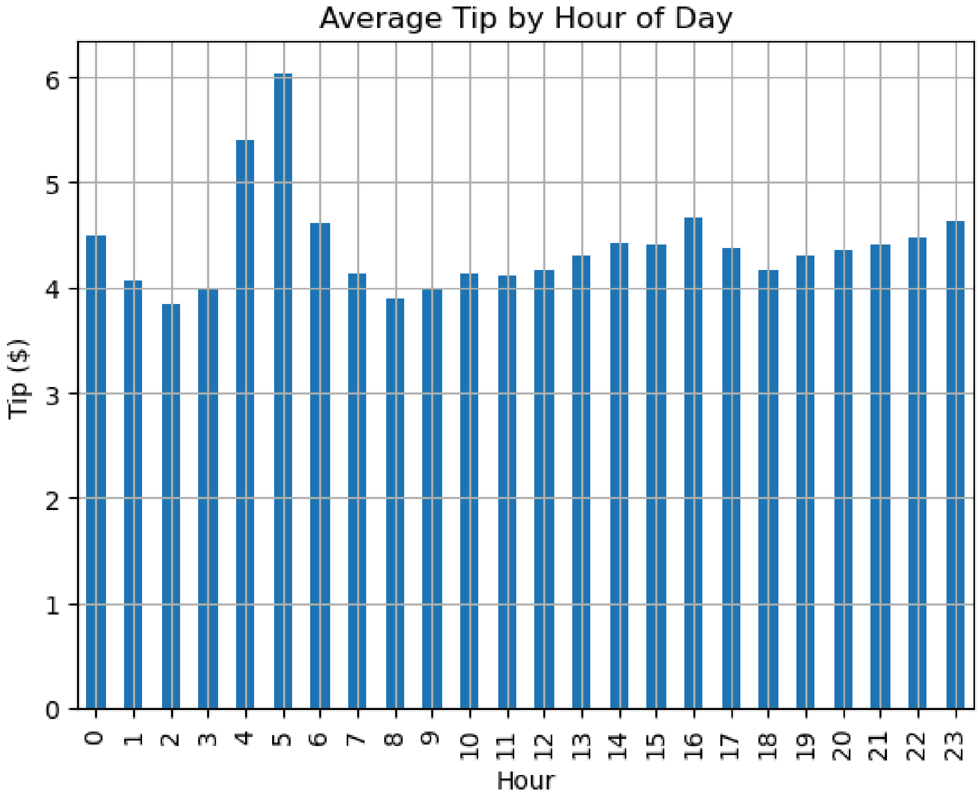


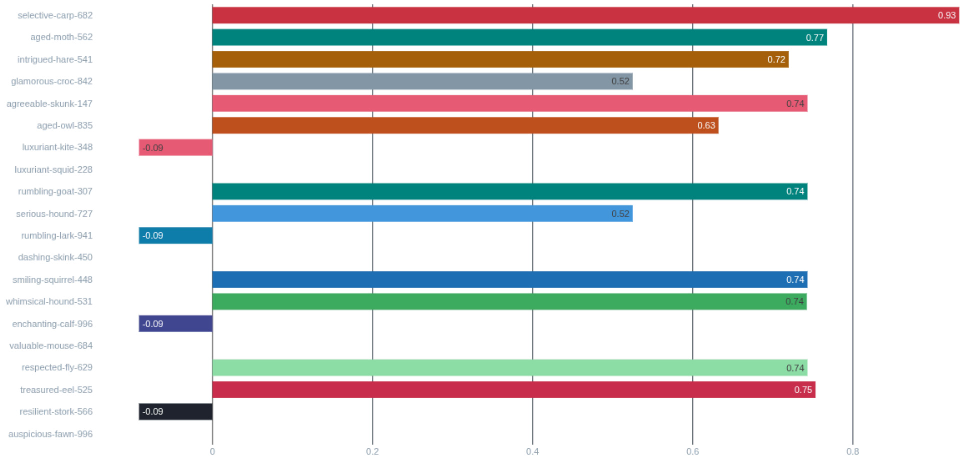
Figure X: Count of taxi trips during March.

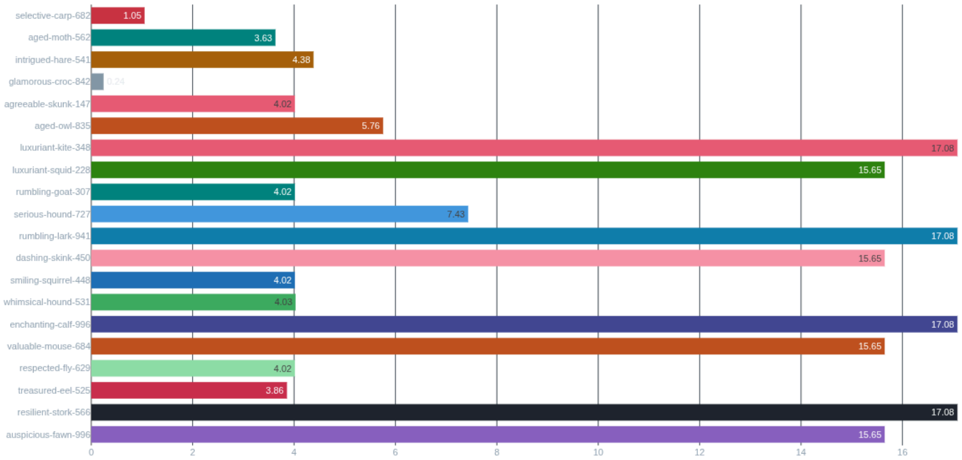
Reviewing distribution of the amount of tip per week and hour, there is a pick in early morning hours(04:00-06:00)



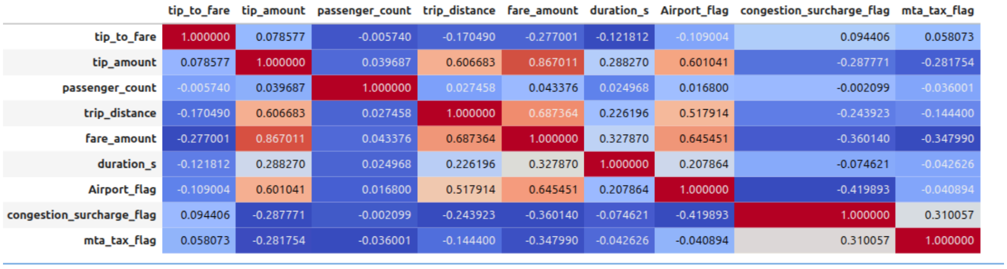


Modeling Experiments





First inuition on looking on some correlated values and try to fit a linear model.



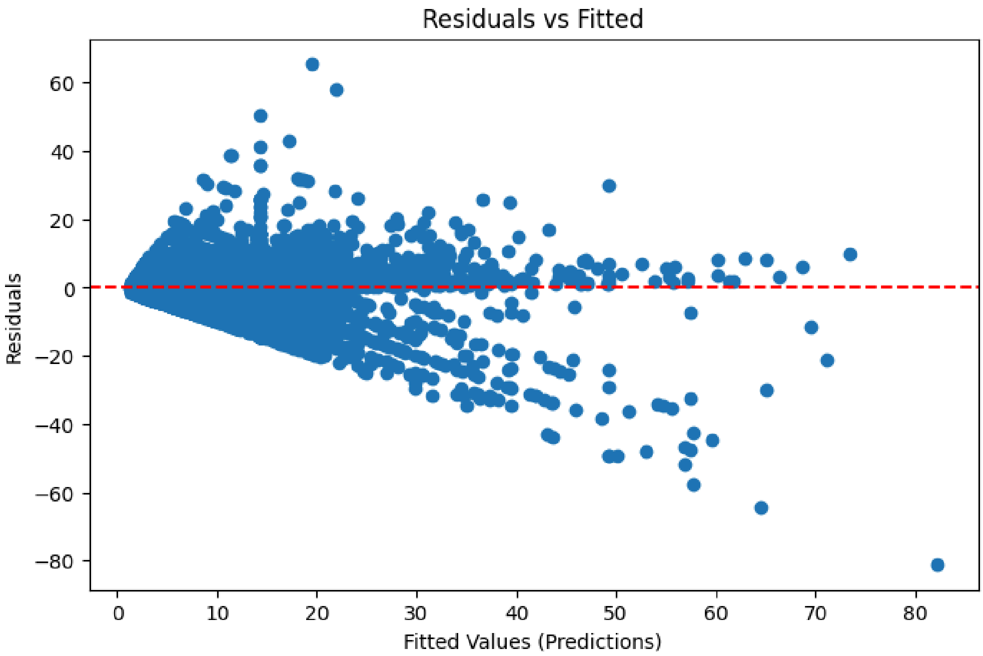
	tip_to_fare	tip_amount	passenger_count	trip_distance	fare_amount	duration_s	Airport_flag	congestion_surcharge_flag	mta_tax_flag
tip_to_fare	1.000000	0.078577	-0.005740	-0.170490	-0.277001	-0.121812	-0.109004	0.094406	0.058073
tip_amount	0.078577	1.000000	0.039687	0.606683	0.867011	0.288270	0.601041	-0.287771	-0.281754
passenger_count	-0.005740	0.039687	1.000000	0.027458	0.043376	0.024968	0.016800	-0.002099	-0.036001
trip_distance	-0.170490	0.606683	0.027458	1.000000	0.687364	0.226196	0.517914	-0.243923	-0.144400
fare_amount	-0.277001	0.867011	0.043376	0.687364	1.000000	0.327870	0.645451	-0.360140	-0.347990
duration_s	-0.121812	0.288270	0.024968	0.226196	0.327870	1.000000	0.207864	-0.074621	-0.042626
Airport_flag	-0.109004	0.601041	0.016800	0.517914	0.645451	0.207864	1.000000	-0.419893	-0.040894
congestion_surcharge_flag	0.094406	-0.287771	-0.002099	-0.243923	-0.360140	-0.074621	-0.419893	1.000000	0.310057
mta_tax_flag	0.058073	-0.281754	-0.036001	-0.144400	-0.347990	-0.042626	-0.040894	0.310057	1.000000

Starting simple we use only the fare_amount and the results are the following

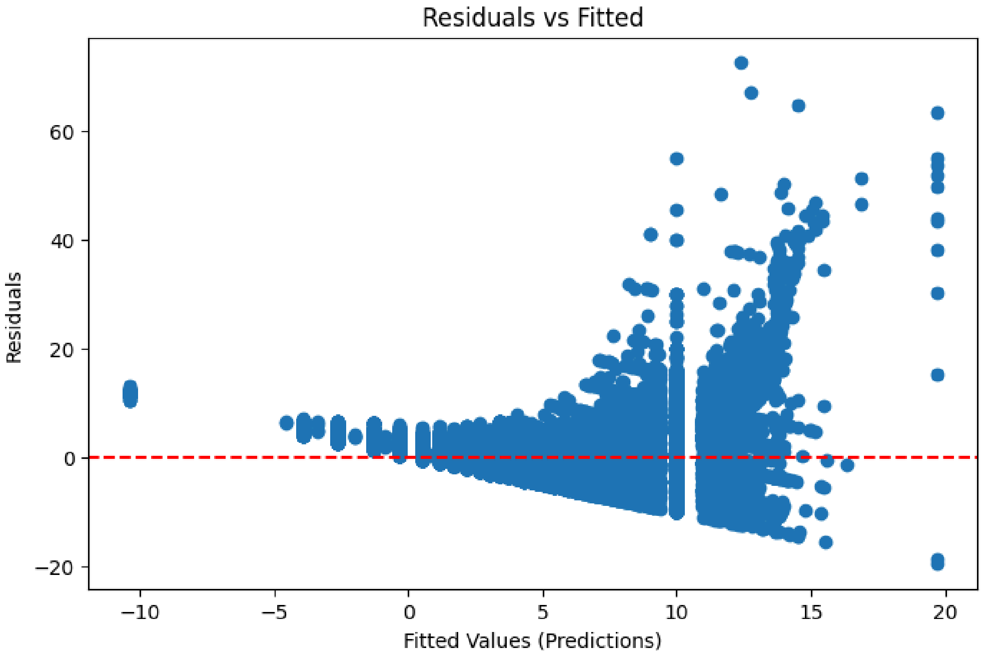
Model Evaluation:

Model	**MSE**	**R²**
Dummy Mean	MSE: 15.6497	R²: -0.0000
Dummy Median	MSE: 17.0848	R²: -0.0917
Linear Regression	MSE: 3.8599	R²: 0.7534
Decision Tree	MSE: 4.0165	R²: 0.7433

Although the statistical metrics are positive, the residuals do not distribute normally and the conical shape of the Residuals vs Fitted values suggests heteroscedacity in the model



Transforming the input data



Regularisation of the lm model

Trying to resolve heteroscedacity with Regualirsation and Transfoming the values

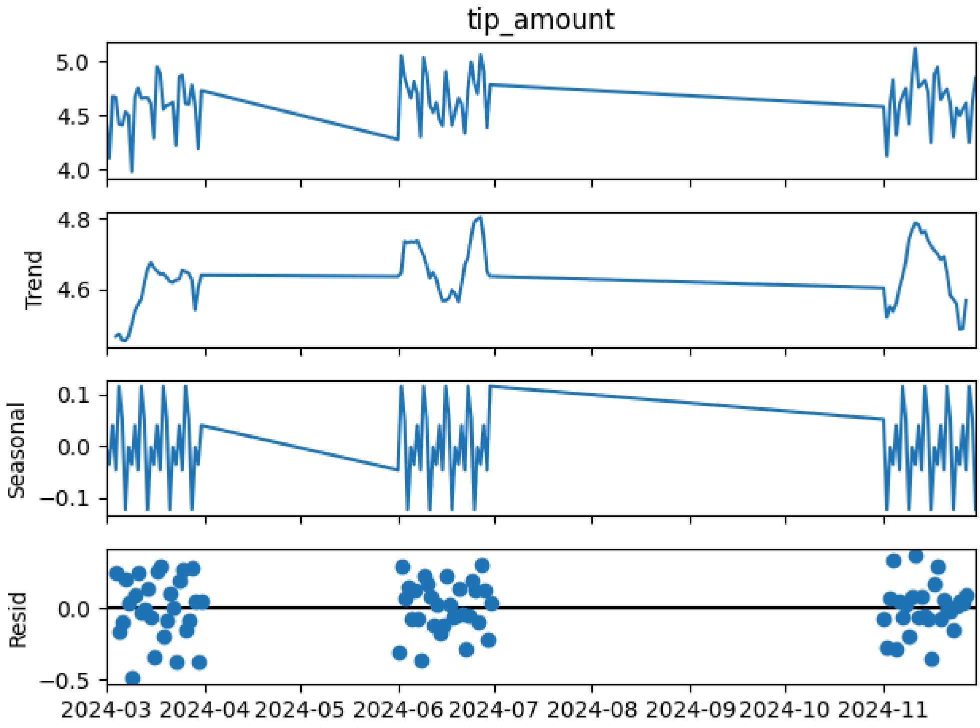
```
Elastic Net          | MSE: 4.3846 | R²: 0.7198

tip_amount = 0.9210 + 0.1745*fare_amount + 0.0000*Airport_flag + 0.0801*trip_distance +
0.0000*congestion_surcharge_flag + 0.0000*mta_tax_flag
```

CatBoost

The best result was performed when using CatBoost and introducing some extra features as the PULocationID and DOLocationID

Other experiments on Spatio-temporal analysis

Q: Is there a siggnificant temporal corelation to the tip amount?



Conduct time corelation between thet tip amount and the PU datetime

```
Spearman Correlation: -0.0029, p-value: 1.4924e-05
```

Q: Is there a significant spatial corelation between the tip amount given to a borough and its neightborhood?

Further Steps