GENERAL ASSEMBLY

Boston Bike hire

Blue bike Dataset
Report 2019 & 2020

Presented By
Mashitah Susanto

Hypothesis

How well can we predict a user's trip duration?

Goal

To establish which feature has the most effect on the duration

---

Mashitah Susanto

Blue Bike Dataset
Report 2025

Outline

Mashitah Susanto

Blue Bike Dataset
Report 2019 & 2020

Cycling in Boston

Massachusetts, is one of the most bike-friendly states in the country, with Boston and Cambridge considered the best places for cyclists. With a population of 4.87 million, Boston is ranked No. 6 most bikeable city in the US.

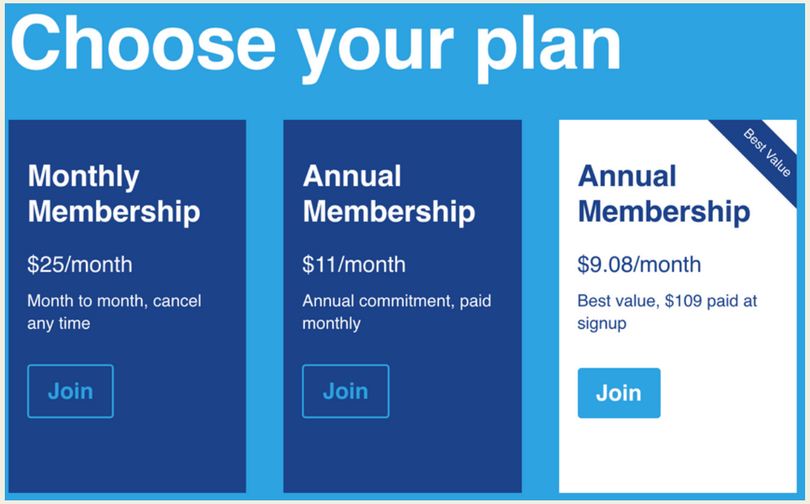
The state worked hard for by investing in updated infrastructure, building roads with bicycle tracts, founding one of the nation’s first public bike-sharing systems

---

Mashitah Susanto

Blue bike Dataset
Report 2019 & 2020

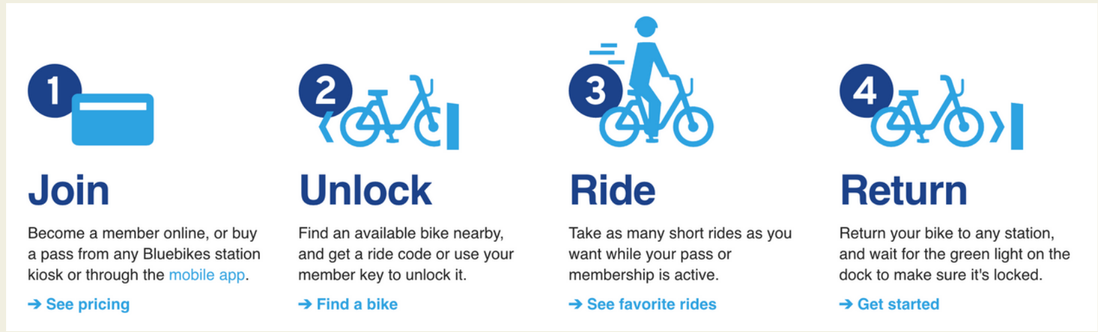
BLUEbikes SM



400+ docking stations & 4000+ bikes

BlueBikes company is a government-owned bike-share company, all across Arlington, Boston, Brookline, Cambridge, Chelsea, Everett, Newton, Revere, Salem, Somerville, or Watertown.

A user can pick up a bike at any station dock, ride it for a specific amount of time, and then return it to any station for re-docking.



Mashitah Susanto

Blue bike Dataset
Report 2019 & 2020

Project Framework

1 ACQUIRE THE DATA
Data Collection

2 DATA CLEANING
EDA

3 EXPLORATION
Data Visualisation

4 MODELING
Linear Regression & Logistic Classification, Gridsearch

5 EVALUATION
Limitations, Next Step, Conclusion

Will the hypothesis be solved?

Data Source(s)

KAGGLE

where I obtain the dataset

BLUEBIKE WEBSITE

BOSTON NEWS

The Data

```
In [2]: df= pd.read_csv('/Users/mashitahsusanto/Desktop/GA/CAPSTONE/bluebikes_tripdata_2019.csv')
df
```

	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype
0	790	2019-12-01 00:01:25.3240	2019-12-01 00:14:35.3350	370	Dartmouth St at Newbury St	42.350961	-71.077828	33	Kenmore Square	42.348706	-71.097009	5133	Subscrit
1	166	2019-12-01 00:05:42.8610	2019-12-01 00:08:29.3830	80	MIT Stata Center at Vassar St / Main St	42.362131	-71.091156	67	MIT at Mass Ave / Amherst St	42.358100	-71.093198	2653	Subscrit
2	323	2019-12-01 00:08:28.5560	2019-12-01 00:13:52.4340	381	Inman Square at Springfield St.	42.374384	-71.100157	221	Verizon Innovation Hub 10 Ware Street	42.372509	-71.113054	4875	Subscrit
3	709	2019-12-01	2019-12-01	185	Third at	42.365445	-71.082771	184	Sidney Research Campus/	42.357753	-71.103934	2116	Subscrit

2,5 m

Rows

18

Columns

```
In [3]: df2= pd.read_csv('/Users/mashitahsusanto/Desktop/GA/CAPSTONE/bluebikes_tripdata_2020.csv')
df2
```

	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	
0	1793	2020-11-01 00:00:18.3990	2020-11-01 00:30:12.2630	186	Congress St at Northern Ave	42.348100	-71.037640	186	Congress St at Northern Ave	42.348100	-71.037640	4896	C
1	1832	2020-11-01 00:00:34.3330	2020-11-01 00:31:07.2920	186	Congress St at Northern Ave	42.348100	-71.037640	186	Congress St at Northern Ave	42.348100	-71.037640	5630	C
2	262	2020-11-01 00:01:54.8450	2020-11-01 00:06:17.4090	186	Congress St at Northern Ave	42.348100	-71.037640	7	Fan Pier	42.353391	-71.044571	5634	St
3	419	2020-11-01 00:04:00.1950	2020-11-01 00:10:59.5590	74	Harvard Square at Mass Ave/ Dunster	42.373268	-71.118579	76	Central Sq Post Office / Cambridge City Hall a...	42.366426	-71.105495	6071	C

Mashitah Susanto

Blue bike Dataset

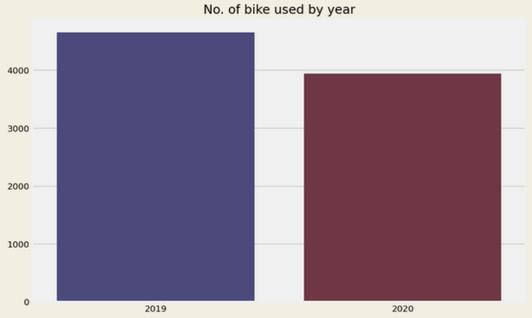
Report 2019 & 2020

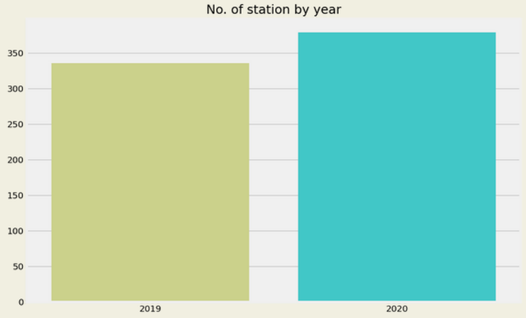
The Data

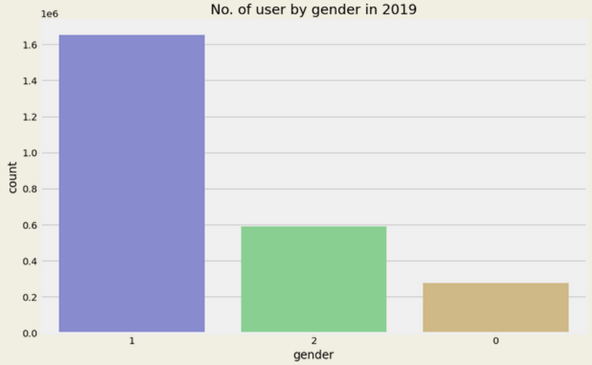
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1999446 entries, 0 to 1999445
Data columns (total 18 columns):
 #   Column                    Dtype
---  ------                    -----
 0   tripduration              int64
 1   starttime                 object
 2   stoptime                  object
 3   start station id          int64
 4   start station name        object
 5   start station latitude    float64
 6   start station longitude   float64
 7   end station id            int64
 8   end station name          object
 9   end station latitude      float64
 10  end station longitude     float64
 11  bikeid                    int64
 12  usertype                  object
 13  postal code               object
 14  year                      int64
 15  month                     int64
 16  birth year                float64
 17  gender                    float64
dtypes: float64(6), int64(6), object(6)
memory usage: 274.6+ MB
```

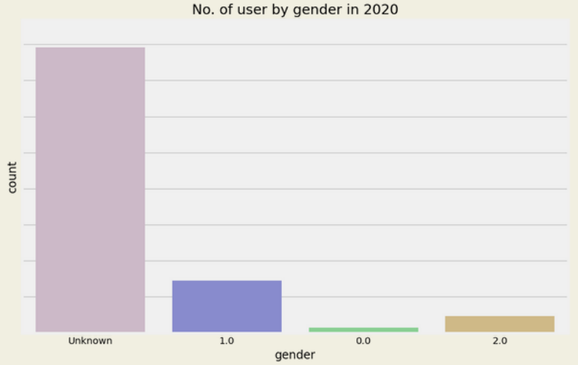
TARGET ARRAY

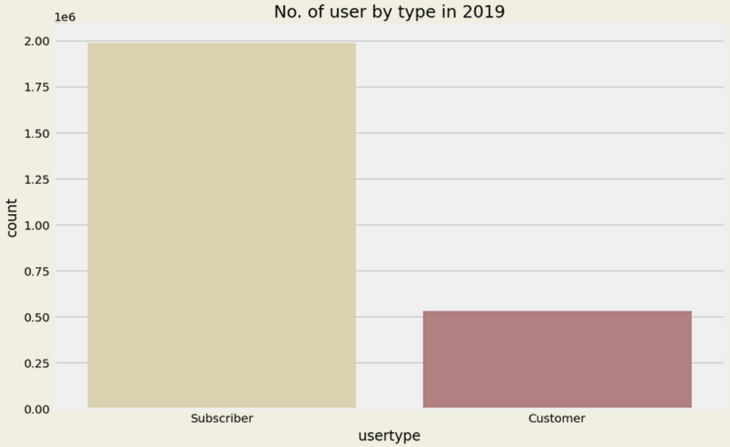
Data Visualisation

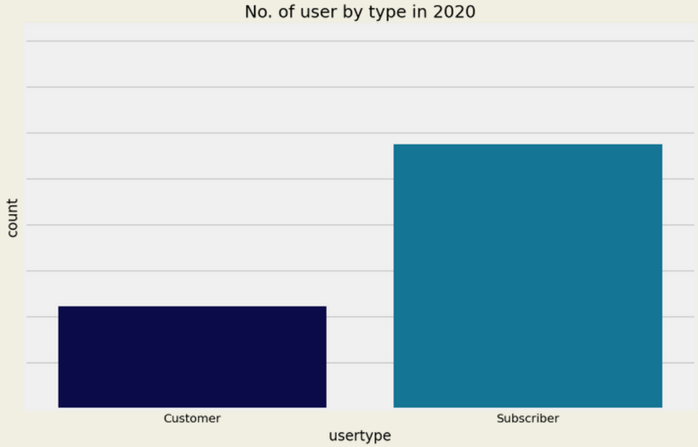










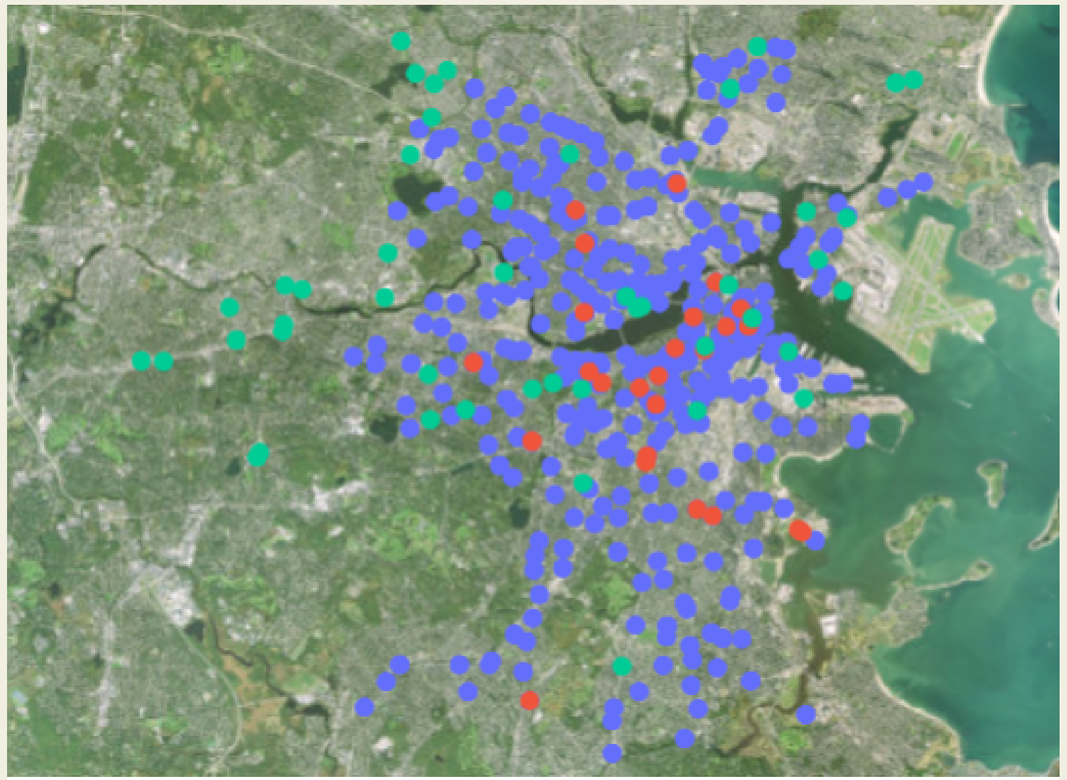


Mashitah Susanto

Blue bike Dataset

Report 2019 & 2020

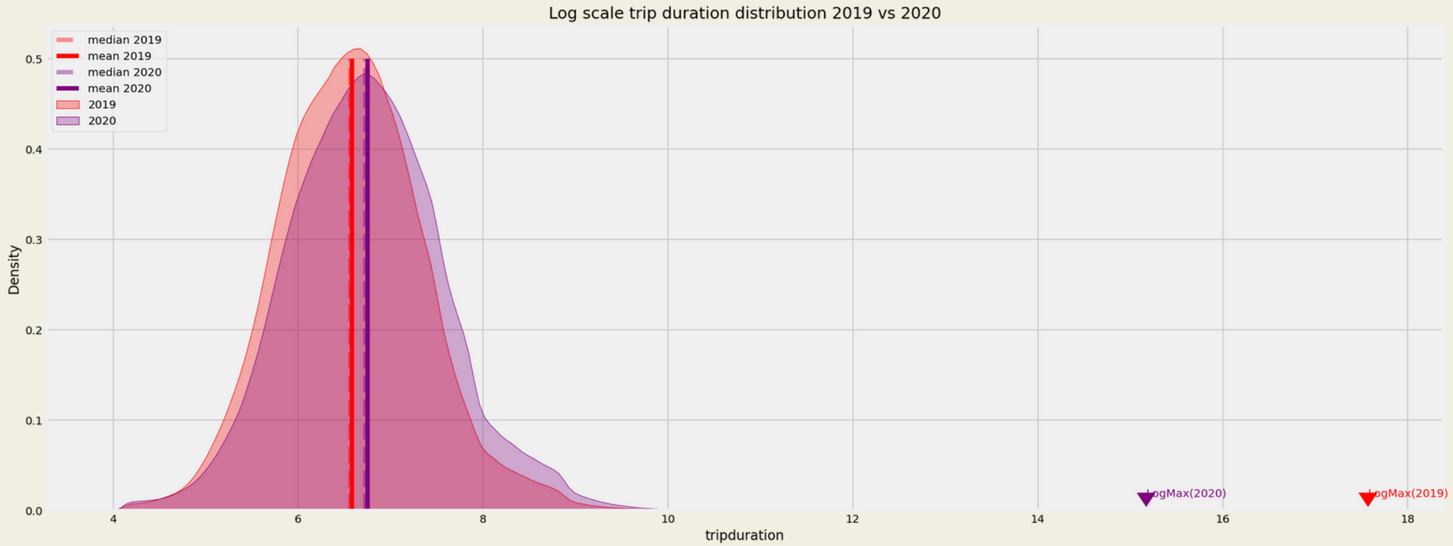
Data Visualisation

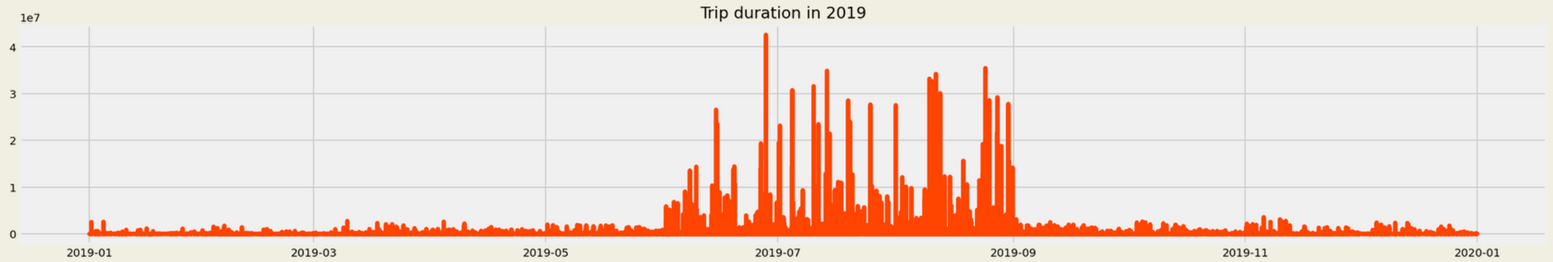


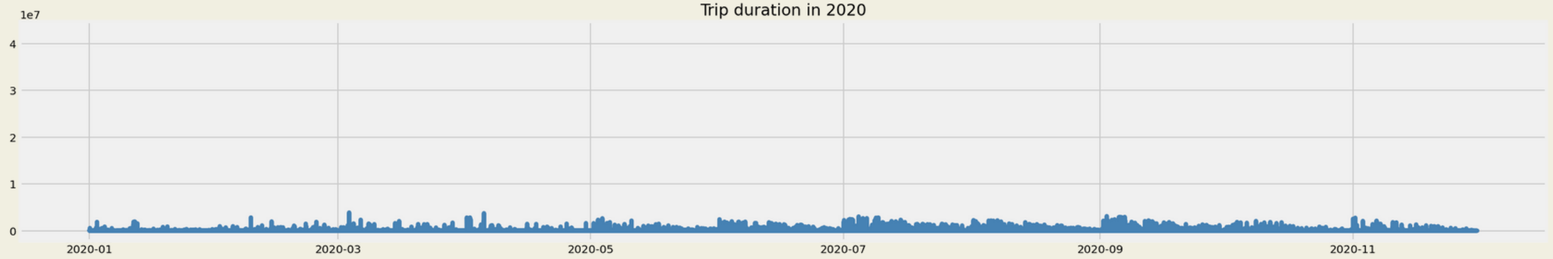
Mashitah Susanto

Blue bike Dataset

Report 2019 & 2020

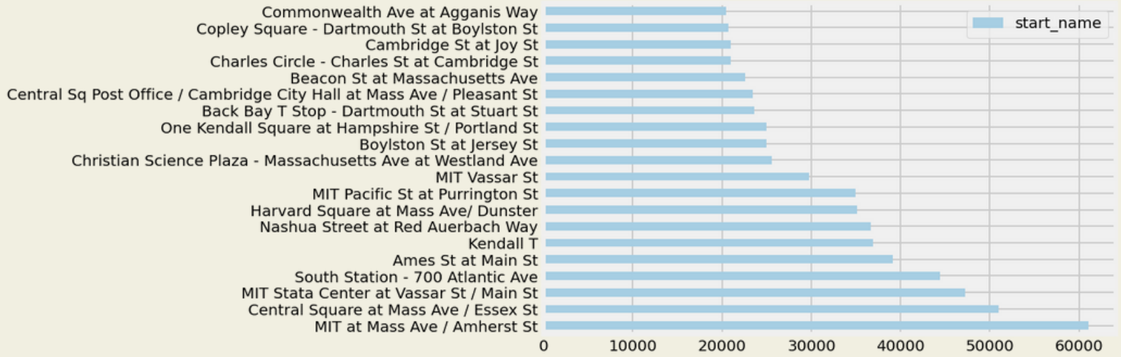


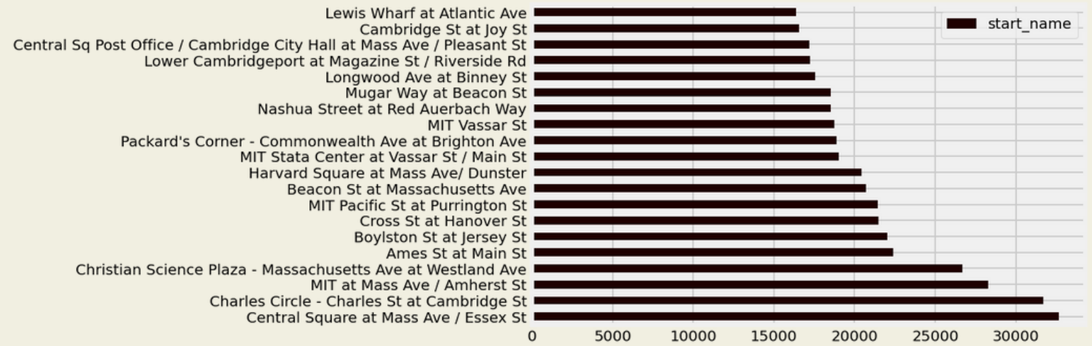


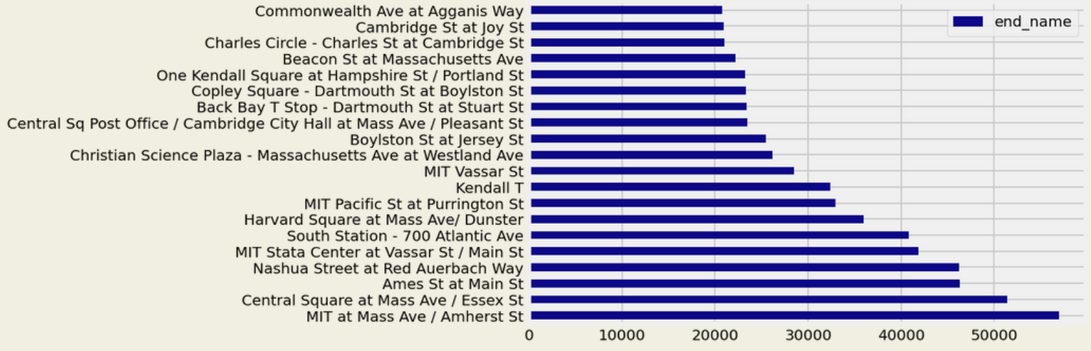


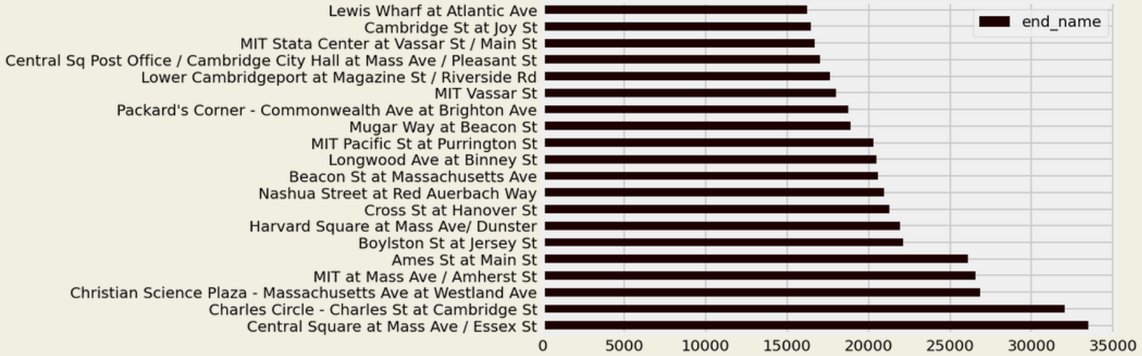
Mashitah Susanto Blue bike Dataset

Report 2019 & 2020





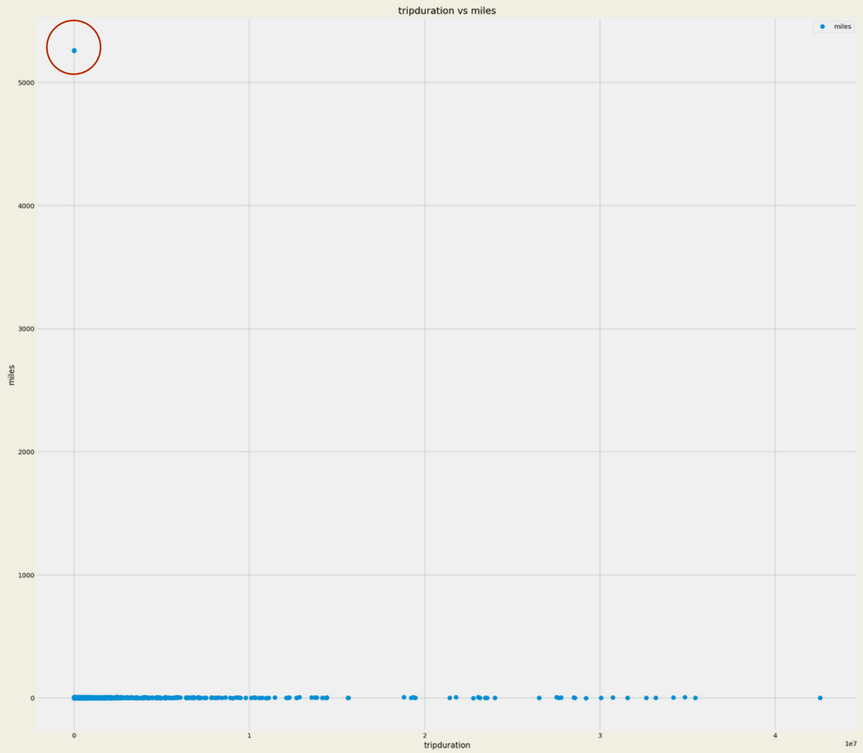




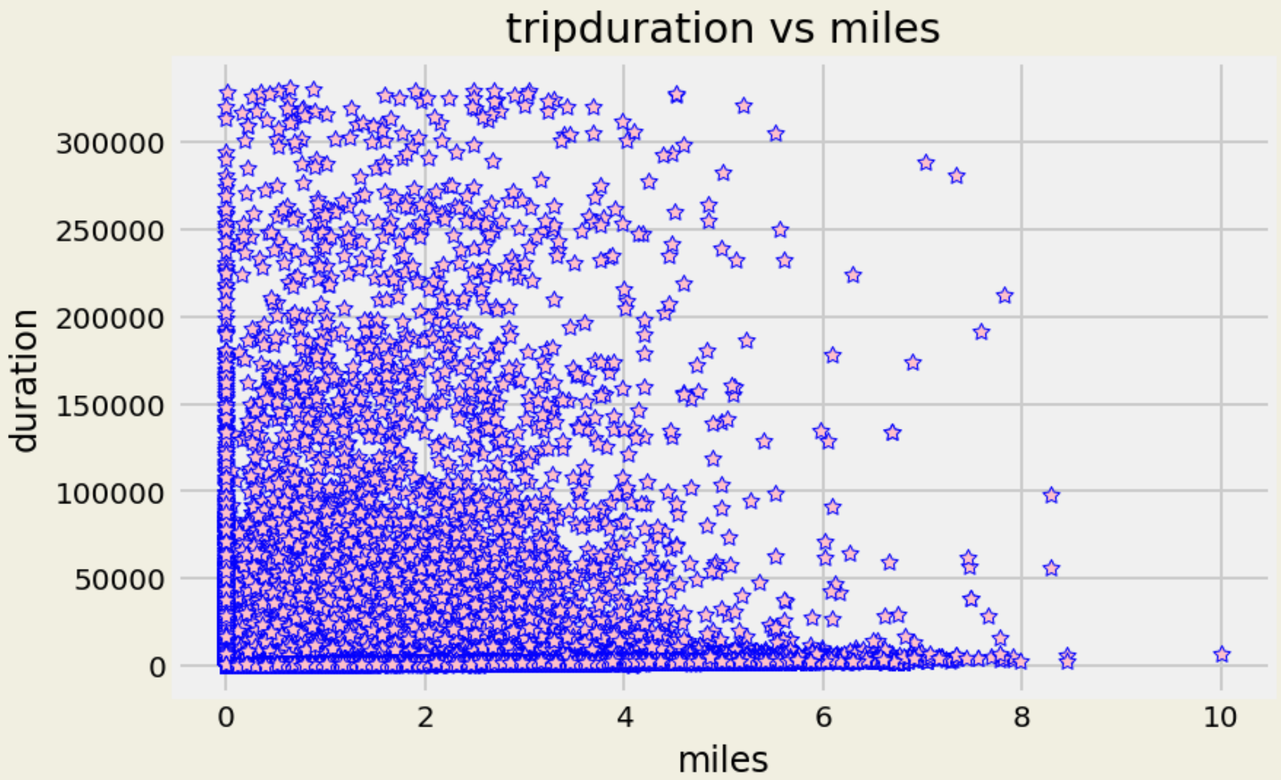
Mashitah Susanto Blue bike Dataset

Report 2019 & 2020

outlier



After removing outlier



Linear Regression

test size: 90/10

```
Cross-validated training scores: [-0.0302192   0.00058617  0.00021457 -0.00059153 -0.0157843
4]
Mean cross-validated training score: -0.009158865453166487
Training Score: 0.0006259519983532469
Test Score: -0.0027113732702048754
```

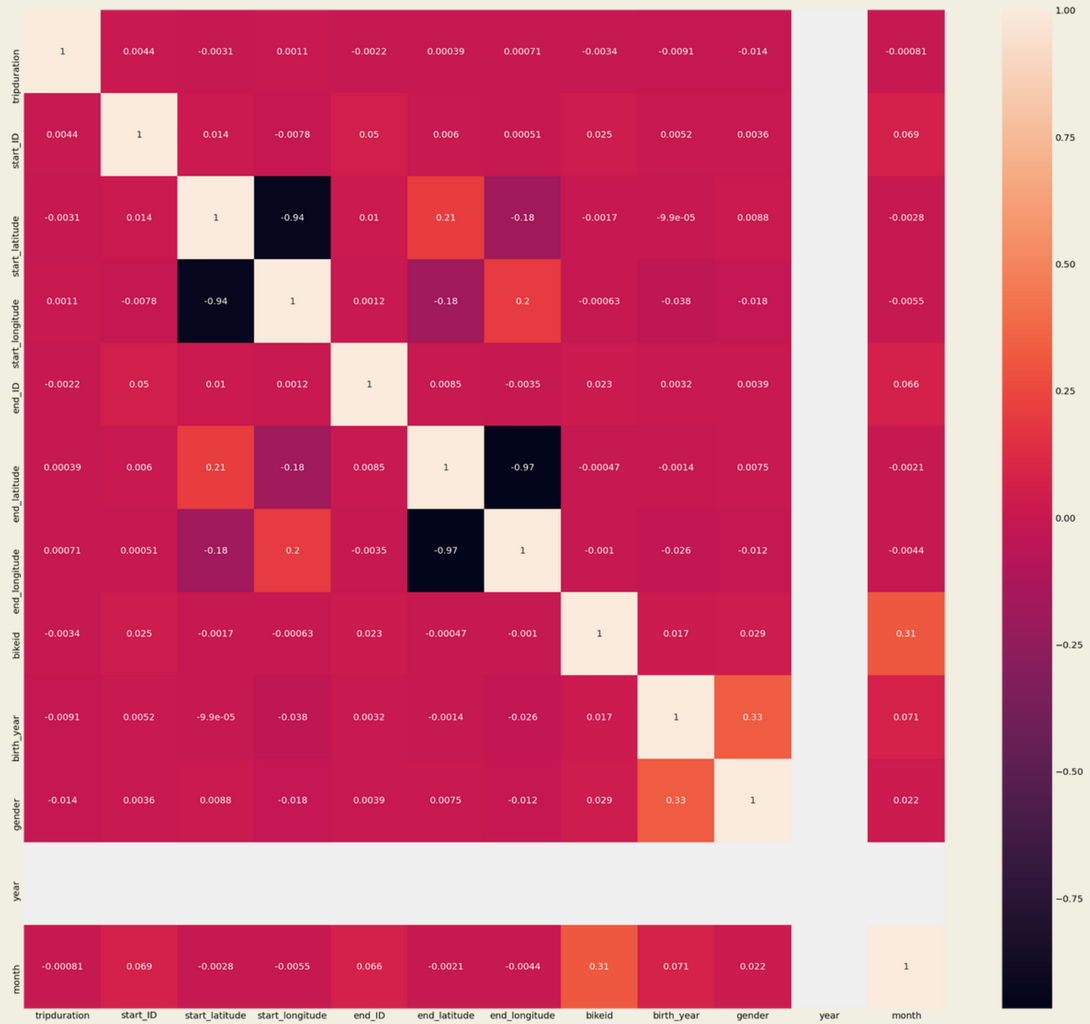
very low score!

try to increase the accuracy level using Polynomial Features?

```
Cross-validated training scores - poly: [-3.28347740e-01  8.59566872e-03 -3.97975549e-05 -5.35837170e-02
 -3.20861473e+04]
Mean cross-validated training score - poly: -6417.304134445026
Training Score - poly: 0.006448120658014922
Test Score - poly: -48784.53188217088
```

still very low....

```
R2 score: 0.0027950981300092215
Intercept: -1720.6791594389338
```



Mashitah Susanto Blue bike Dataset

Report 2019 & 2020

Logistic Classification

```
# baseline for DF value
baseline = y.value_counts(normalize=True).max()
baseline
```

```
0.5004925139856135
```

test size: 30/70

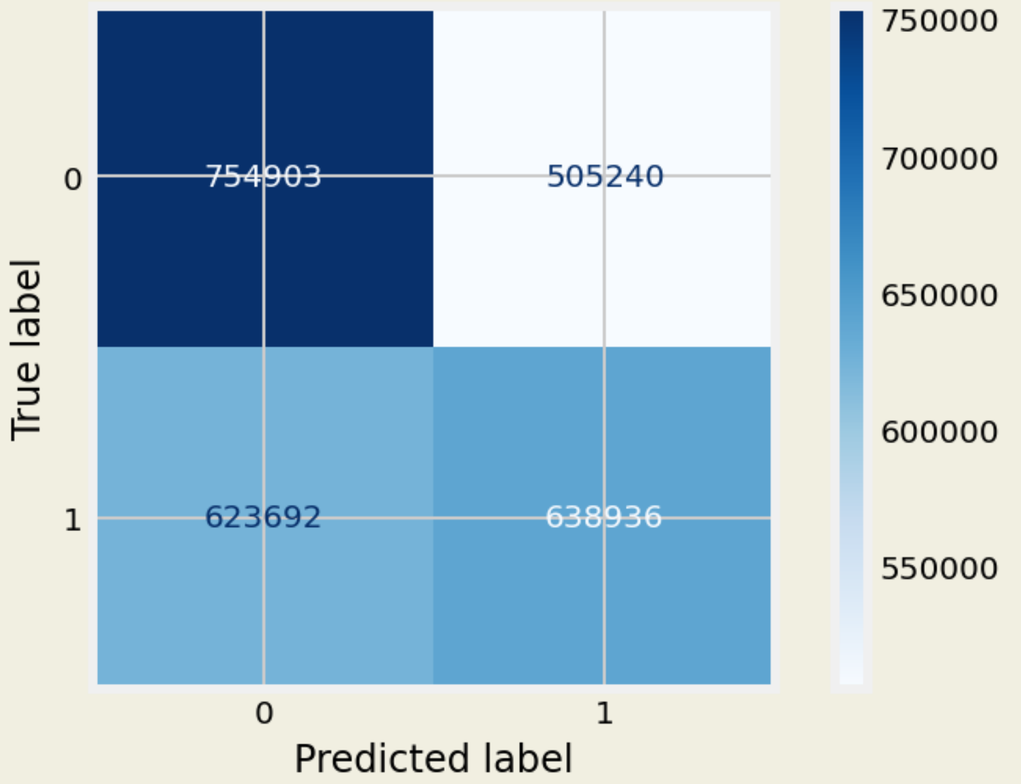
```
Cross-validated training scores: [0.55096184 0.55307856 0.5532173  0.55615851 0.55418845]
Mean cross-validated training score: 0.5535209313571986
Training Score: 0.5529096984663683
Test Score: 0.5520966619258498
```

5% higher than baseline!

what if we scale data?

```
lr = LogisticRegression(multi_class='ovr')
cross_val_score(lr, X_test_std, y_test, cv=5).mean()
```

```
0.5656882908521996
```



GridSearch & Feature importance

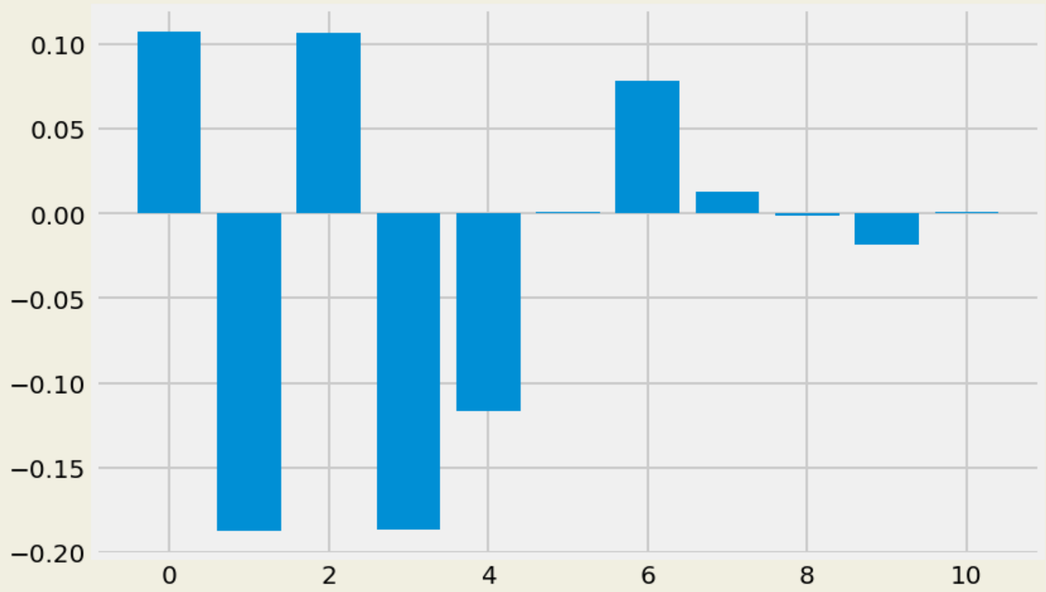
```
Cross-validated training scores: [0.57275646 0.56928809 0.57201467 0.57147954 0.56979487]
Mean cross-validated training score: 0.5710667246276735
 KNN Training Score: 0.6492744086856908
KNN Test Score: 0.5740874012439584
```

```
Cross-validated training scores: [0.65379737 0.65266767 0.65246259 0.65113467 0.6551977 ]
Mean cross-validated training score: 0.6530520009313692
 KNN Training Score: 0.7089152003551652
```

```
0.6576802228942248
```

```
knn_gridsearch.best_score_
```

```
0.686356659680789
```



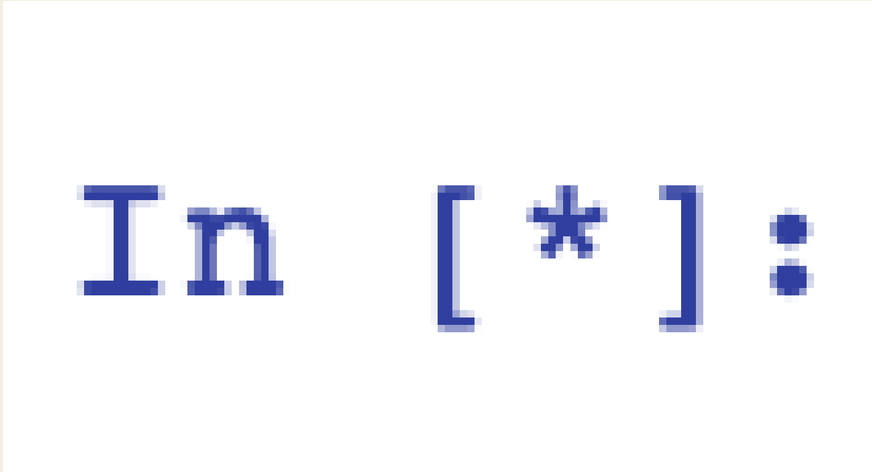
Start latitude, End latitude & Daystart have the highest impact on feature importance

Evaluation

What are the feature that has the most effect on trip duration?

Day of the week

Limitations



- Limited Columns in the Dataset
- Time Restrictions
- Is the Data Accurate?
- Too much time on EDA & data cleaning

Conclusion

Did we solve the hypothesis?

yes... to a certain extent

Next Steps

- Improving my scores in Gridsearch & Logistic model
- explore more using Decision Tree
- Explore more plots (ROC curve, Gradientboosting classifier, & Tableau)
- Boosting & Random Forest
- Further explore the Dataset with a new Hypothesis: Does the birth year affect the station that they are heading?

Resources

- https://www.bluebikes.com/
- https://www.kaggle.com/jackdaoud/bluebikes-in-boston
- https://www.bikeattorney.com/cycling-in-boston-an-overview.html
- https://www.boston.com/news/local-news/2021/07/30/if-boston-builds-more-protected-bike-lanes-cyclists-will-come-advocates-say/