Urban Mobility in SF Bay Area

Presented by:
Catherine Bui, Scott Xue, Steven Chen, & Nour El-Difrawy

General Motivation & Significance

- To study the effects of traffic-related delays on the extra fare passengers have to pay
- To use linear regression, machine learning, and K-Nearest Neighbors to understand the data

2.) Transparent pricing for trips.

Want to know about how much a taxi ride will cost? Here's a quick breakdown of taxi rates in San Francisco and for out-of-town trips.

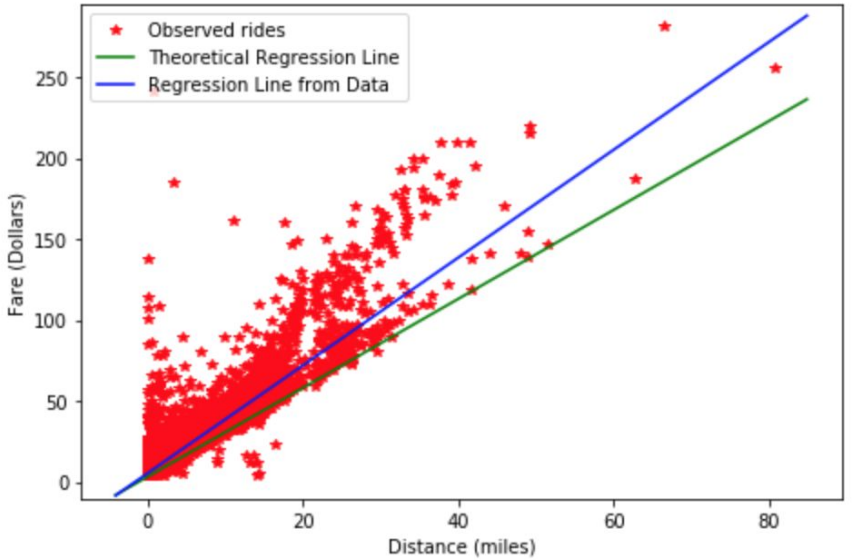
Taxi Service	Fare Amount
First one-fifth mile of flag rate	$3.50
Each additional one-fifth mile or fraction thereof	$0.55
Each minute of waiting or traffic time delay	$0.55
SFO Exit surcharge	$4.00

Initial Analysis

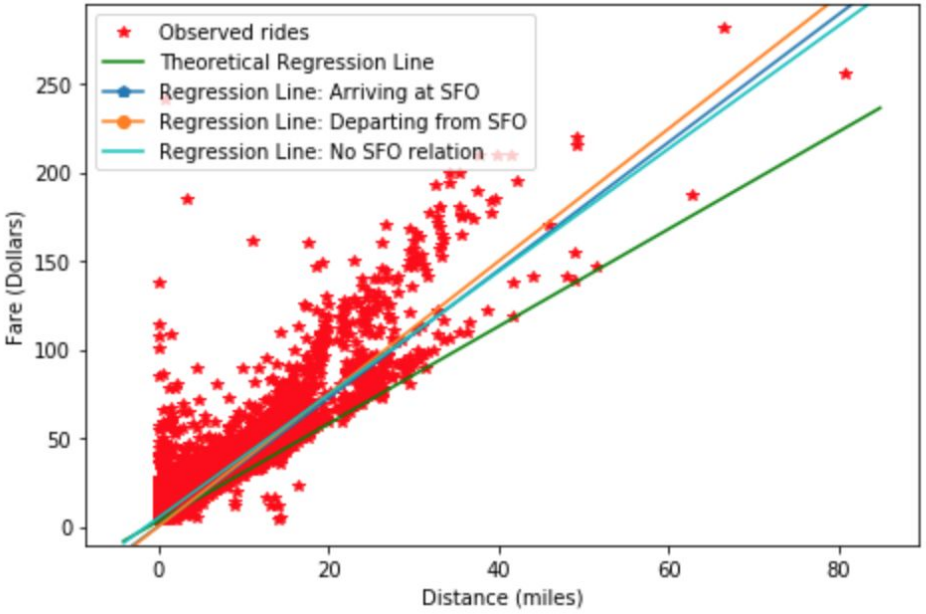
Part 1A: Data Exploration

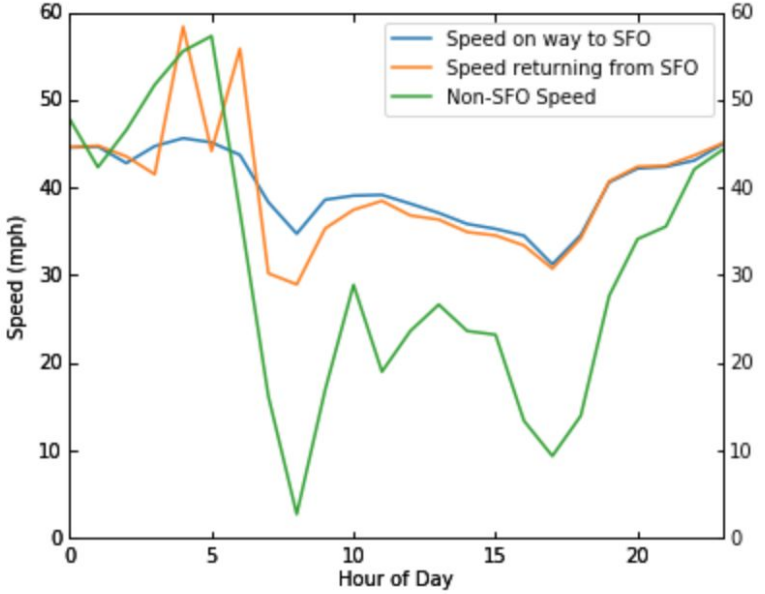
	Theoretical	Actual
Mean Squared Error	50.89	25.47

TAZ: 239



Part 1B: Data Exploration





Part 1B: Data Exploration

SFO trips:

	Mean	Median	Std. Dev
Dist (miles)	13.594	13.956	4.169
Duration (m)	21.933	21.000	8.809
Extra Fare	$8.87	$6.89	$10.39

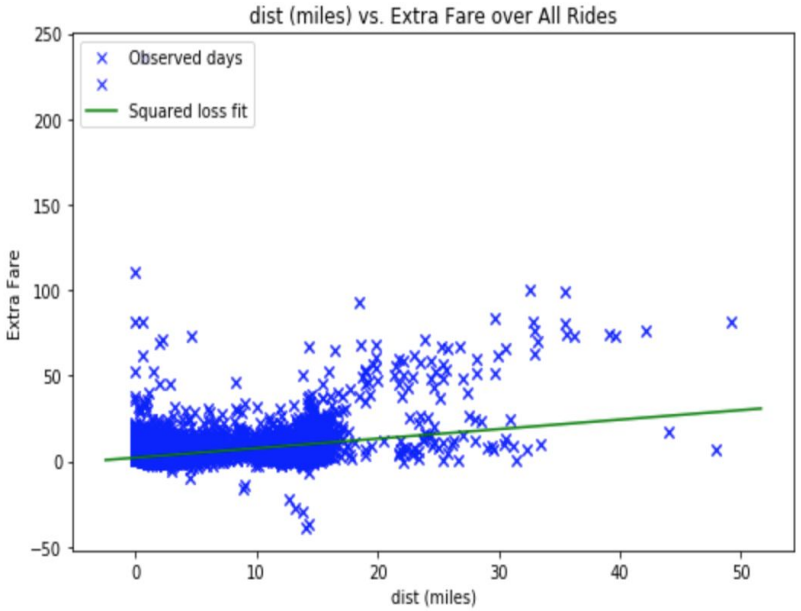
Non SFO trips:

	Mean	Median	Std. Dev
Dist (miles)	3.044	1.770	4.021
Duration (m)	10.661	9.000	7.446
Extra Fare	$4.11	$2.94	$5.62

Part 2: Machine Learning

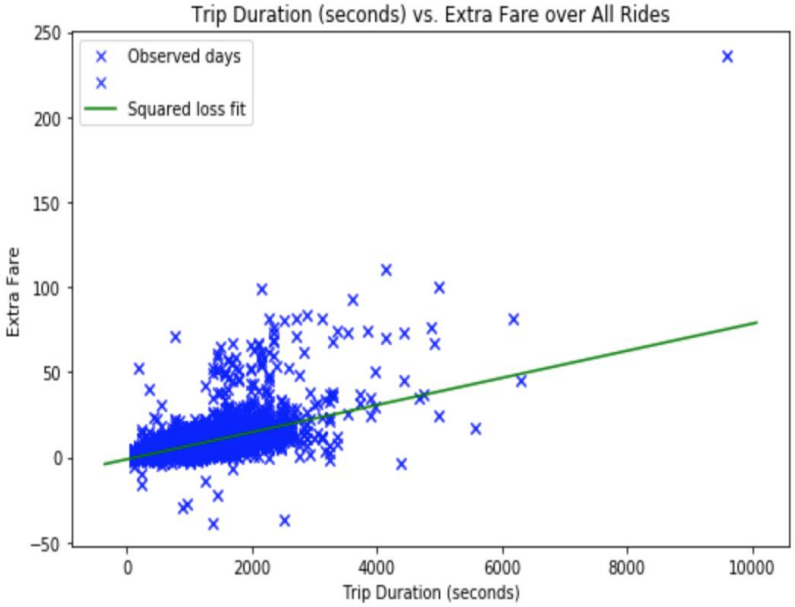
Distance

MSE: 29.843

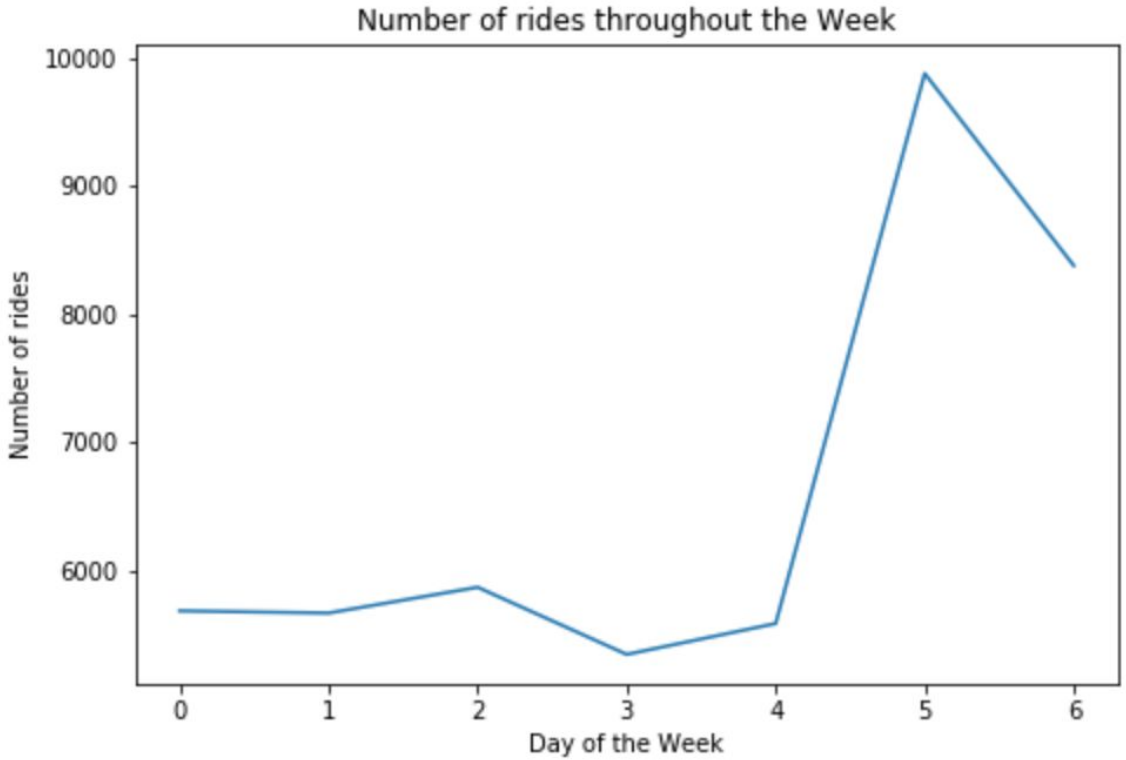


Duration

MSE: 21.573



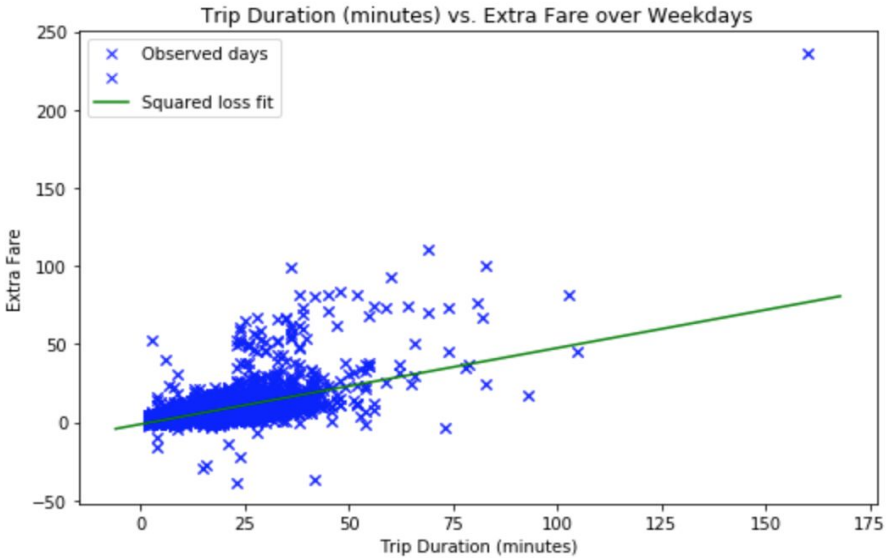
Separate Weekday/Weekend Regressions



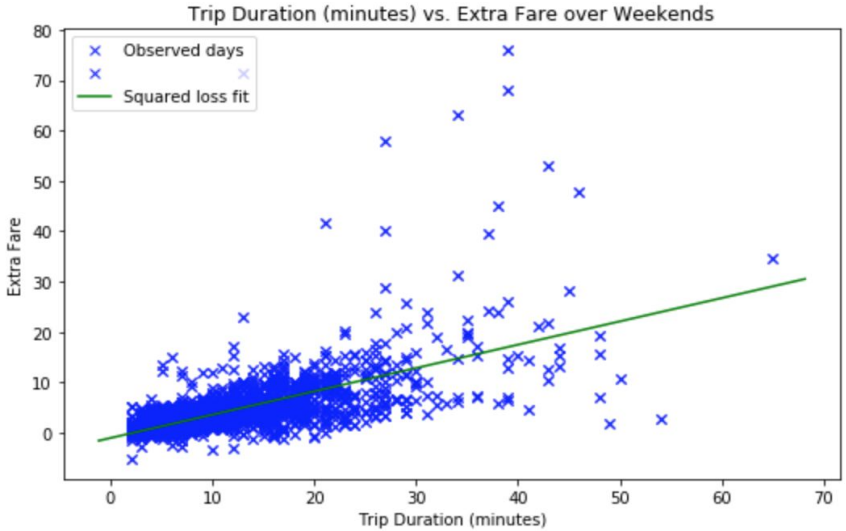
Separate Weekday/Weekend Regressions

Using duration, (average) LSE decreases from 21.573 to **18.659**.

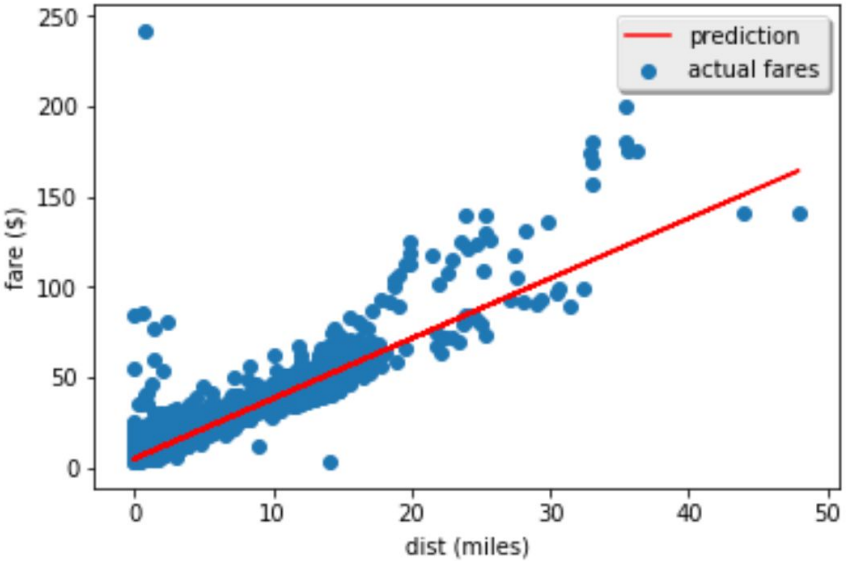
MSE = 23.188

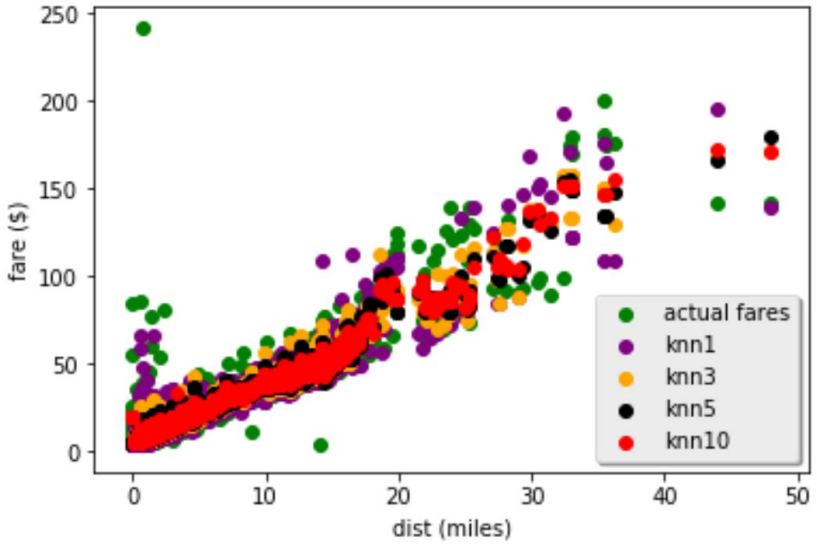


MSE = 14.130



Part 3: Linear Regression vs. K-nearest Neighbor





MSE KNN vs. Linear Regression

```
#Linear Regression
train_slope = sum(taxi_train.column(0)*(taxi_train.column(1)-np.mean(taxi_train.column(1))))/sum(taxi_train.column(0)*(taxi_train
train_intercept = (-1*train_slope*np.mean(taxi_train.column(0)))+np.mean(taxi_train.column(1))
predicted_fares = train_slope*taxi_test.column(0)+train_intercept
sum_of_errors_squared = sum((predicted_fares - taxi_test.column(1))**2)
sum_of_errors_squared
mean_squared_error(predicted_fares,taxi_test.column(1))
```

```
32.184562711291306
```

```
In [23]: sum((knn_pred(1) - taxi_test.column(1))**2)
         mean_squared_error(knn_pred(1), taxi_test.column(1))
Out[23]: 45.252737444339495
```

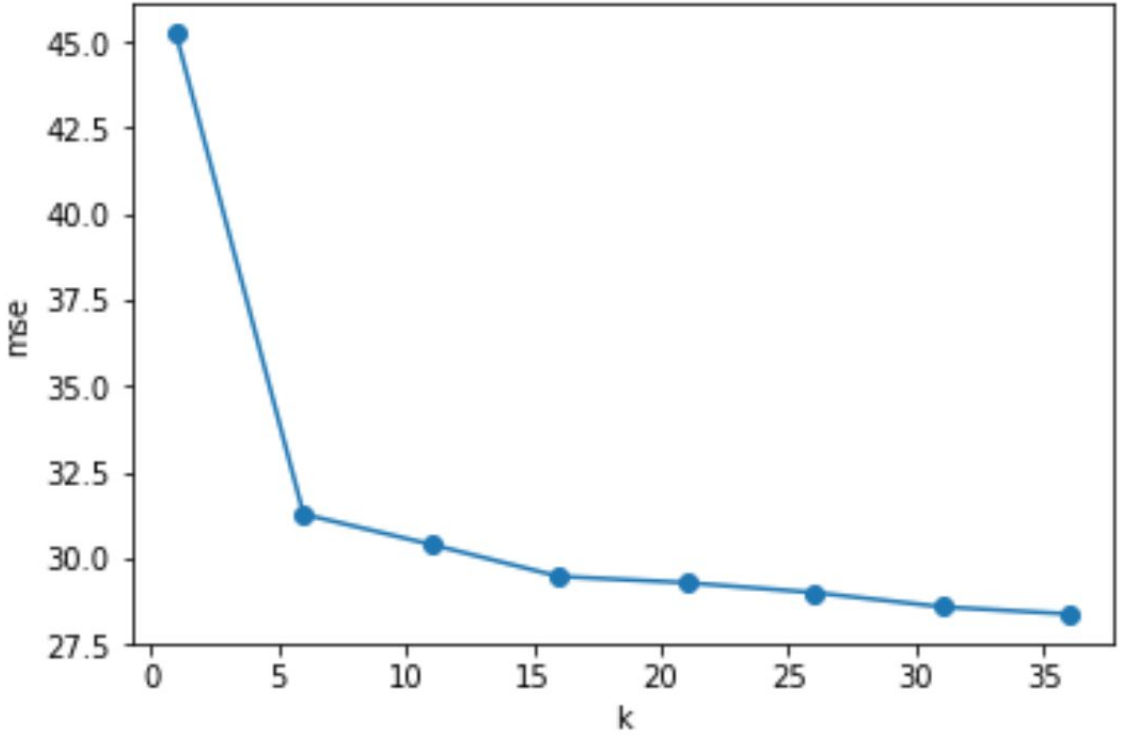
```
In [25]: sum((knn_pred(3) - taxi_test.column(1))**2)
         mean_squared_error(knn_pred(3), taxi_test.column(1))
Out[25]: 33.115535646711344
```

```
In [26]: sum((predictionknn - taxi_test.column(1))**2)
         mean_squared_error(knn_pred(5), taxi_test.column(1))
Out[26]: 31.847334267440292
```

```
In [24]: sum((knn_pred(10) - taxi_test.column(1))**2)
         mean_squared_error(knn_pred(10), taxi_test.column(1))
Out[24]: 30.608248306166509
```

For k=1,3, Linear Regression gives a better prediction. For k=5,10, KNN gives a better prediction.

K-Values vs. MSE



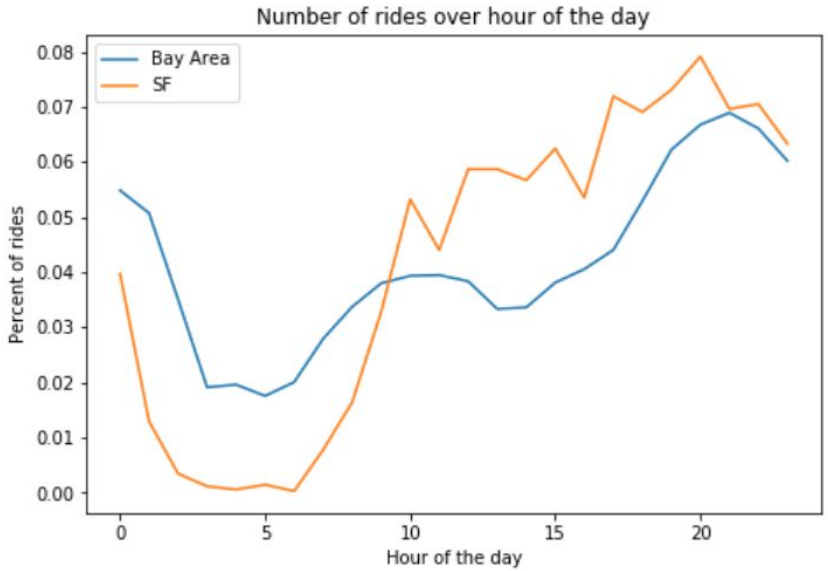
k	mse
1	45.2527
6	31.2797
11	30.4
16	29.4716
21	29.2885
26	29.0025
31	28.5875
36	28.3785

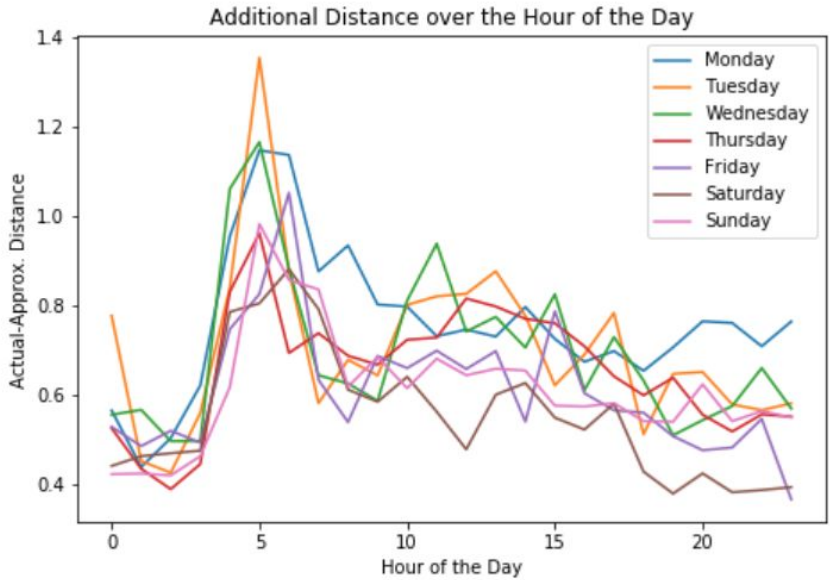
Additional Analysis

Using new metrics to improve analysis

- Time of day when trip was taken
- Day of the week when trip was taken
- Average speed of trip
- Approx. “extra distance” of trip
  - Difference between straight line distance and actual distance of trip
  - Imperfect but decent approximation

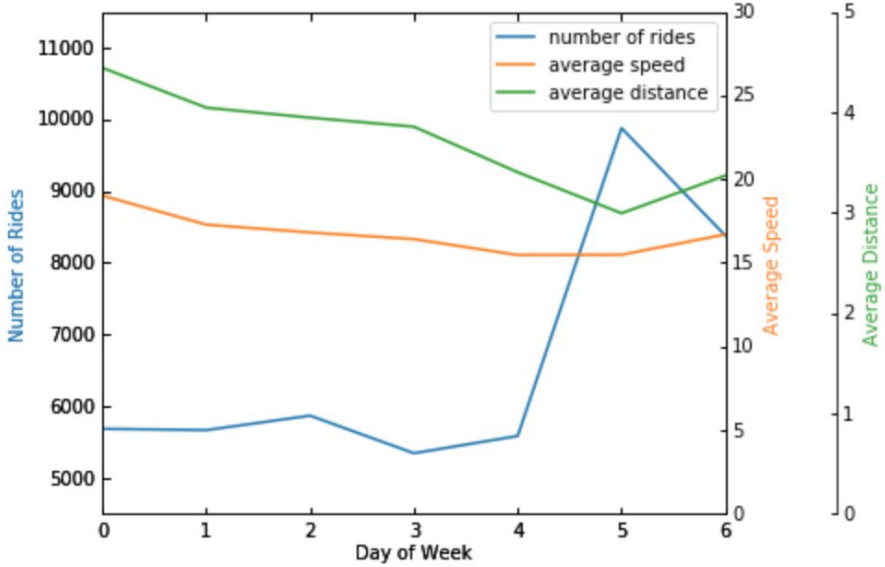
Does traffic activity affect additional distance?

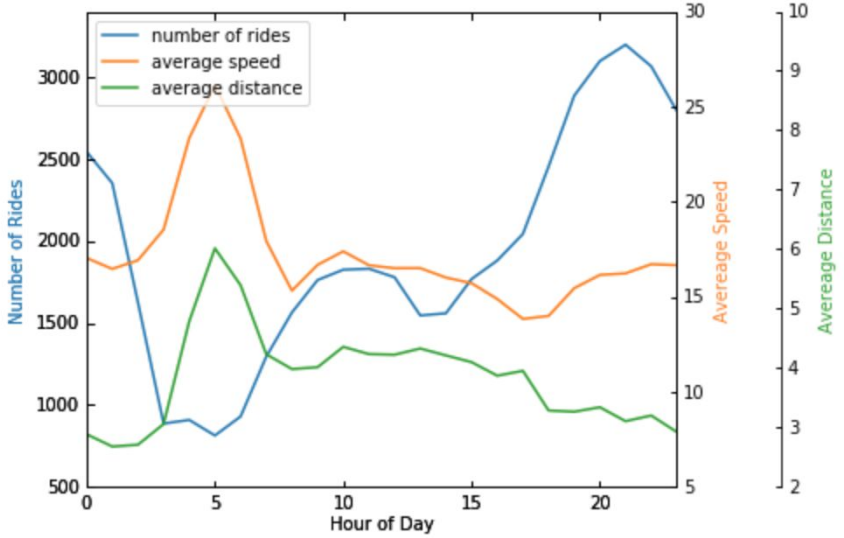




- Additional distance/Hour is the inverse of the Number of rides/Hour
- Traffic activity does influence drivers' behaviors

Does traffic activity affect average speed?



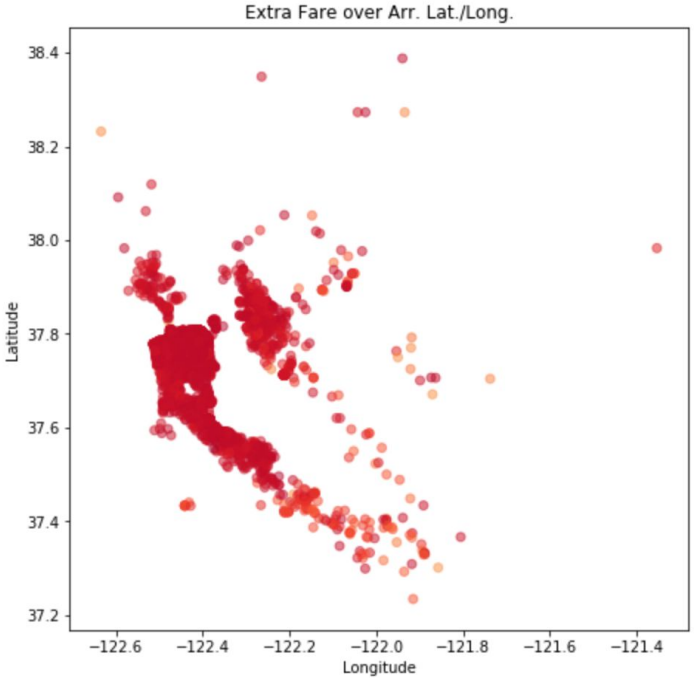


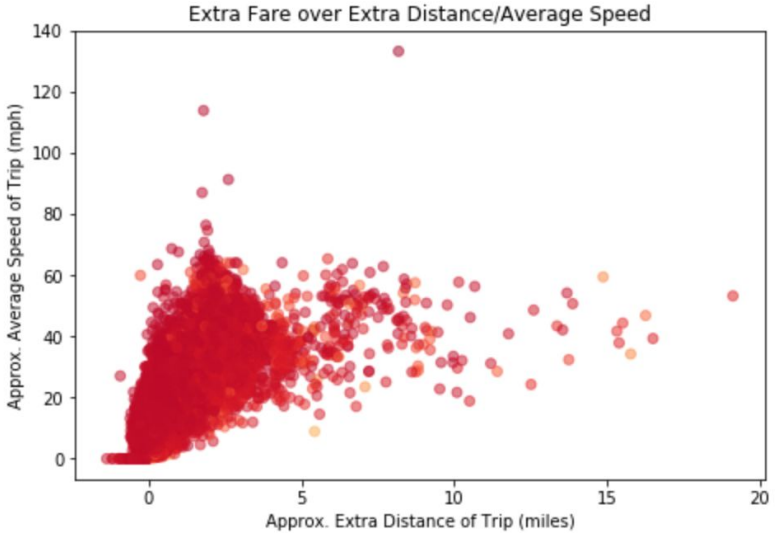
Multivariate Regression

- Using 12 features, can reduce MSE to **17.333**
- Using a subset of just 7 of those, can reduce MSE to **17.371**

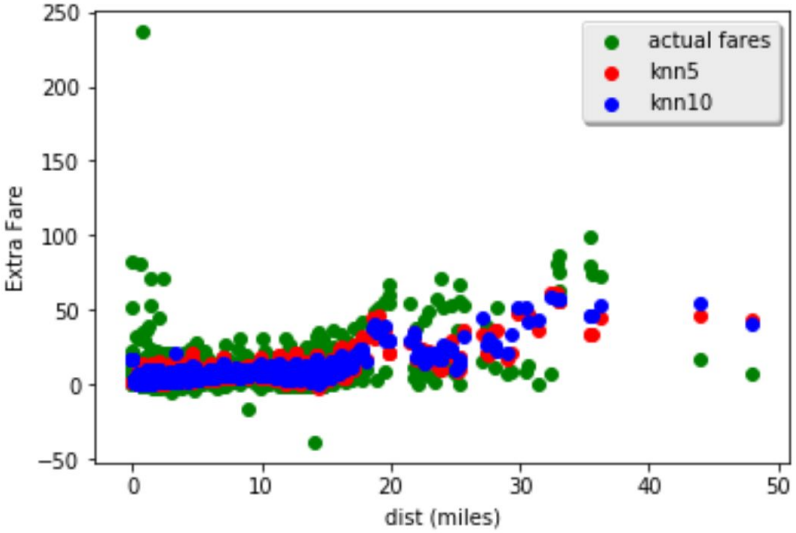
Features

1. Trip Duration
2. Trip distance
3. Hour of day of trip
4. **Weekend trip [0, 1]**
5. Average Speed of trip
6. Extra Distance of trip
7-8.Dep. Longitude, Latitude
9-10. Arr. Longitude, Latitude
11-12.Arr. SFO, Dep. SFO [0, 1]





KNN further analyses

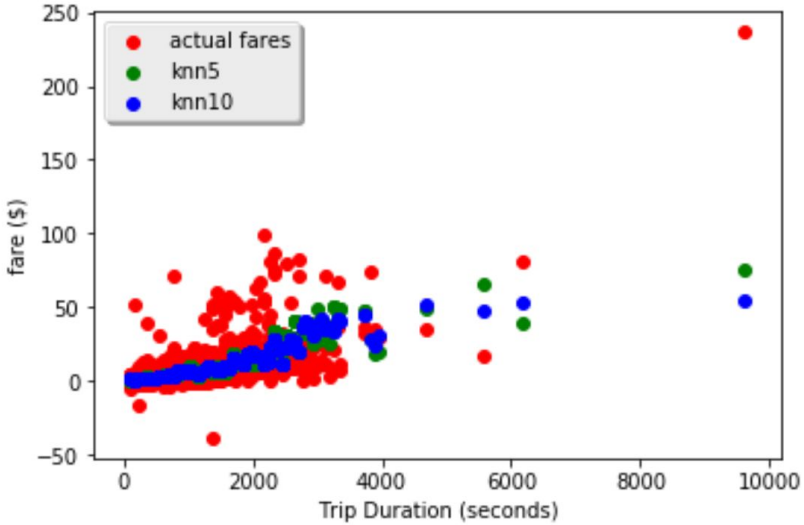



```
In [13]: mean_squared_error(extra_fares_5,taxi_test.column(1))
Out[13]: 31.883081117963339

In [14]: mean_squared_error(knn_pred(10),taxi_test.column(1))
Out[14]: 30.670621070762422

In [18]: mean_squared_error(knn_pred(20),taxi_test.column(1))
Out[18]: 28.970544624820374
```

KNN Predicting Extra Fares with Duration




```
mean_squared_error(k_5,taxi_test.column('Extra Fare'))
```

```
27.292066373505875
```

```
mean_squared_error(k_10,taxi_test.column('Extra Fare'))
```

```
26.259536458936864
```

```
mean_squared_error(k_20,taxi_test.column('Extra Fare'))
```

```
25.259152330832066
```

Conclusion

- Taxi fares aren’t as transparent as they say they are
- Multiple factors other than just those posted (distance and duration) can increase the actual fare that the customer pays