Help User, Help Uber

A Proposal Connecting Uber & Users with Insights to Innovate for Better Efficiency

Section 1: Data Cleaning/Preprocessing

Conditional Fill:

For a specific (*origin*, *destination*, *weekday)* pair, we fill the null according to the group median. This will preserve the originality because we can maintain the difference between weekday and weekend. This improved **data completeness from ~50% to ~85%**.

Quantitle Fill:

Since some columns still contain NaN and we want to preserve them because everything else is alright, we also did a quantitle fill based on (origin, destination, weekday) pair and its mean daily travel time quantitle relative to early morning travel time. That is, for a row that contains NaN in early morning travel time, we calculate the quantitle of its mean daily travel time and fill in an early morning travel time that matches the same quantitle. This allow us to preserve the most originality.

Proprocessing:

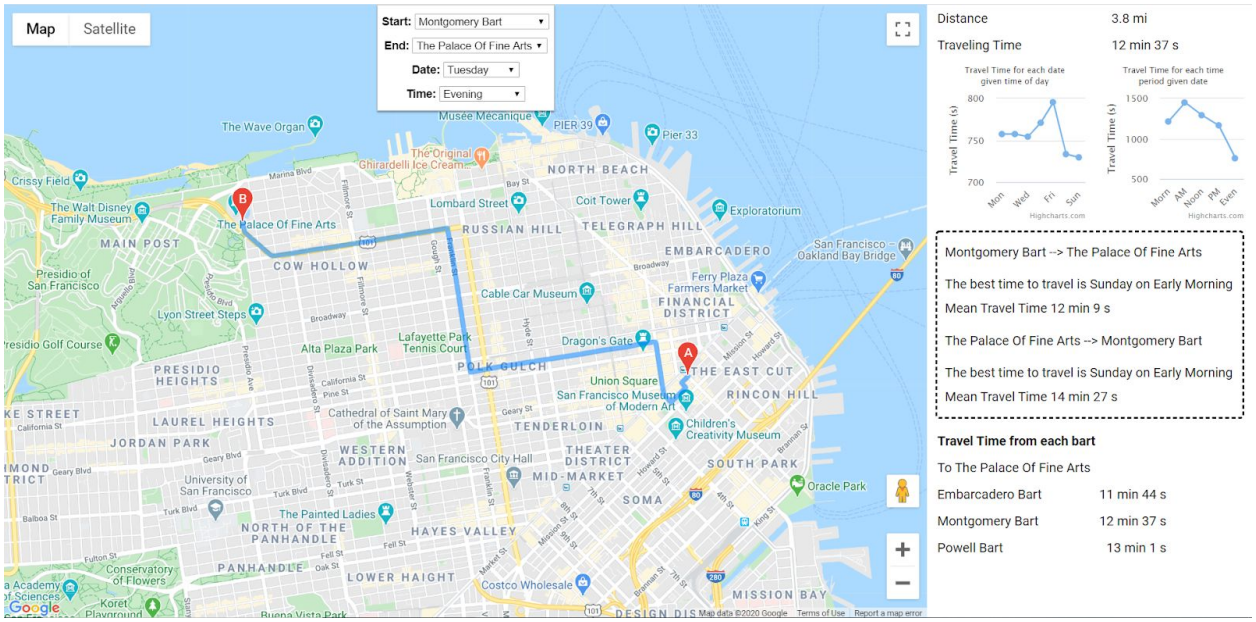
We added weekday(Mon-Sun) and month from date. We also extracted the number of rides from *hour_q1* and *hour_q2* as features in each of Barts and Hotspots locations for machine learning. Since some pairs are not available, including (Powell, Palace), (Palace, Powell), (Powell, Fisherman's), (Fisherman's, Powell), we later disgarded the the number of rides from *hour_q1* and *hour_q2* as features because it might introduce bias.

	filled
Origin ID	1.000000
Destination ID	1.000000
Daily Mean Travel Time	0.982198
Daily Range - Lower	0.982198
Daily Range - Upper	0.982198
AM Mean Travel Time	0.550031
AM Range - Lower	0.550031
AM Range - Upper	0.550031
PM Mean Travel Time	0.715163
PM Range - Lower Bound Travel Time	0.715163
PM Range - Upper Bound Travel Time	0.715163
Midday Mean Travel Time	0.776857
Midday Range - Lower	0.776857
Midday Range - Upper	0.776857
Evening Mean Travel Time	0.685083
Evening Range - Lower	0.685083
Evening Range - Upper	0.685083
Early Morning Mean Travel Time	0.259975
Early Morning Range - Lower	0.259975
Early Morning Range - Upper	0.259975
weekday	1.000000

	filled
Origin ID	1.000000
Destination ID	1.000000
Daily Mean Travel Time	1.000000
Daily Range - Lower	1.000000
Daily Range - Upper	1.000000
AM Mean Travel Time	0.857274
AM Range - Lower	0.857274
AM Range - Upper	0.857274
PM Mean Travel Time	0.960098
PM Range - Lower Bound Travel Time	0.960098
PM Range - Upper Bound Travel Time	0.960098
Midday Mean Travel Time	1.000000
Midday Range - Lower	1.000000
Midday Range - Upper	1.000000
Evening Mean Travel Time	0.921117
Evening Range - Lower	0.921117
Evening Range - Upper	0.921117
Early Morning Mean Travel Time	0.588091
Early Morning Range - Lower	0.588091
Early Morning Range - Upper	0.588091
weekday	1.000000

Section 2: Data Visualization

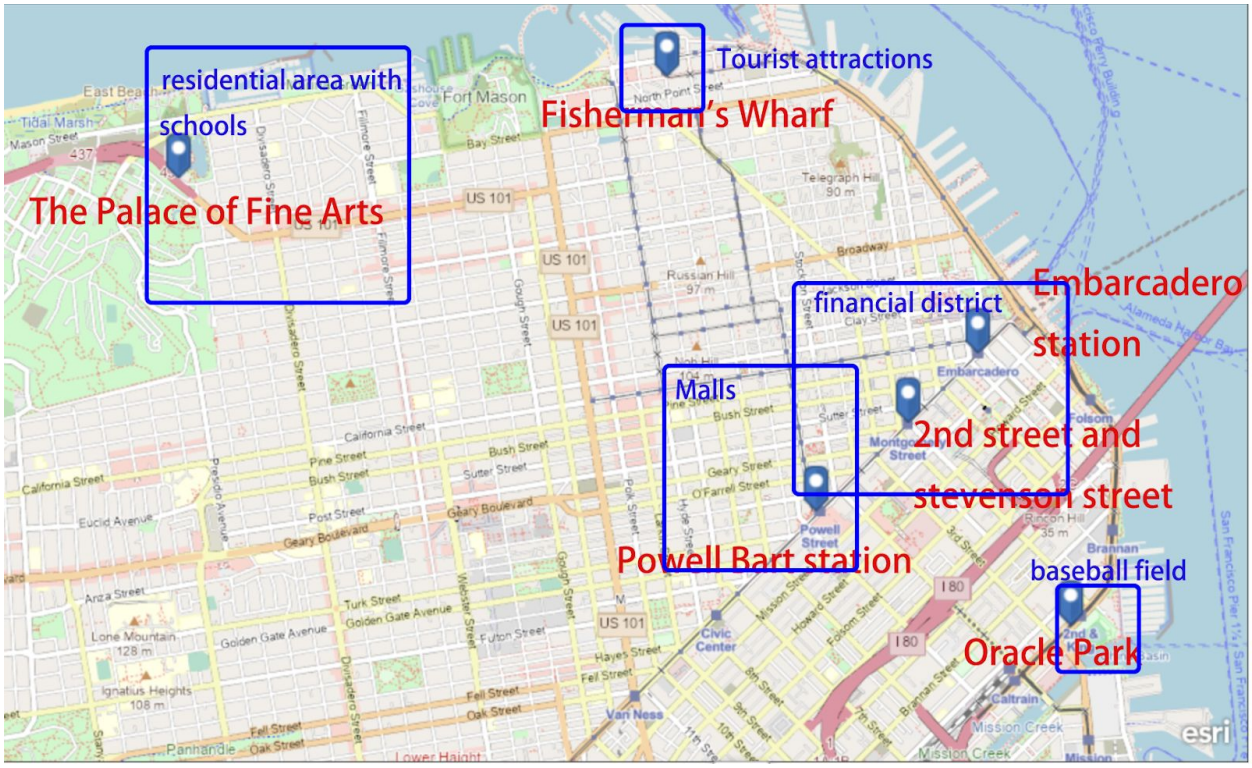
Tableau, ArcGIS, PhotoShop, JavaScript, HTML, CSS, Google Cloud, Google Map, Statsmodels and Matplotlib … delivering the best visualizations for your needs



[Figure 1]

User website (Plan your trip): shy218.github.io

This is our screenshot of our user website. From this website, after users choose the origin and destination, we will provide them some **actionable recommendations**, such as the best time to go there. We will also provide the suggestion of the best time to **go back from each hotspot** to bart to help them plan their trips. In addition, we will also give them a **google map route, predicted traveling time** based on Uber data to show them how to go to the destination to save their time. Moreover, we also provide some **data visualizations to show the trends of traveling time**. Finally, we will tell them the **best bart** to go to their destination hotspot.



In this map, we have 6 locations:

These 6 locations are well representing traffic in San Francisco because they cover a wide range of location type, such as financial district, tourist attraction, malls, sport field, and residencial region.

1: 2nd street and stevenson street:

The "2nd street and stevenson street" is located in a **financial district** where many offices and some restaurants are around. Many people take uber to commute there for work.

2: Embarcadero station

The "Embarcadero station" is located in a farther **financial district** which has a viewpoint "Embarcadero" around. Many people take uber to commute there for work or the Embarcadero viewpoint.

3: Fisherman's Wharf

The "Fisherman's Wharf" is a **famous tourist attraction** located in the Northern part of San Francisco. Many people takes BART (Bay Area Rapid Transportation) to the "Embarcadero station" first and then uber to the Fisherman's Wharf. There are also a few other famous in a walkable or uber distance around the Fisherman's Wharf.

4 : Oracle Park

"Oracle Park" is a **baseball field** that is located in the Southern part of San Francisco. Many famous baseball games and football games were hosted there. People might uber there for games.

5: Powell Bart station

"Powell Bart station" nexts to the **largest commericial area** that includes Macys and Nike, as well as office buildings. People might Uber there for working or shopping.

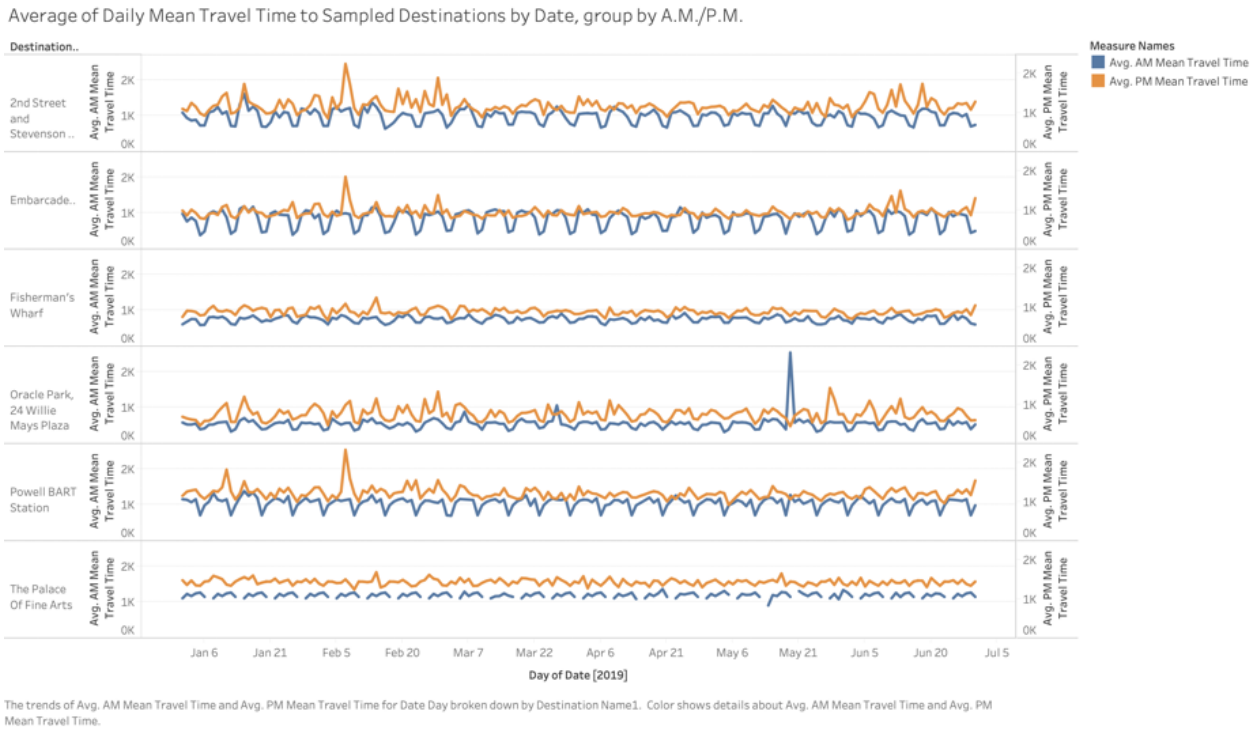
6: The Palace of Fine Arts

"The Palace of Fine Arts" is an art exhibition which is located in a **residential area** that nexts to schools and parks.

Data Visualization

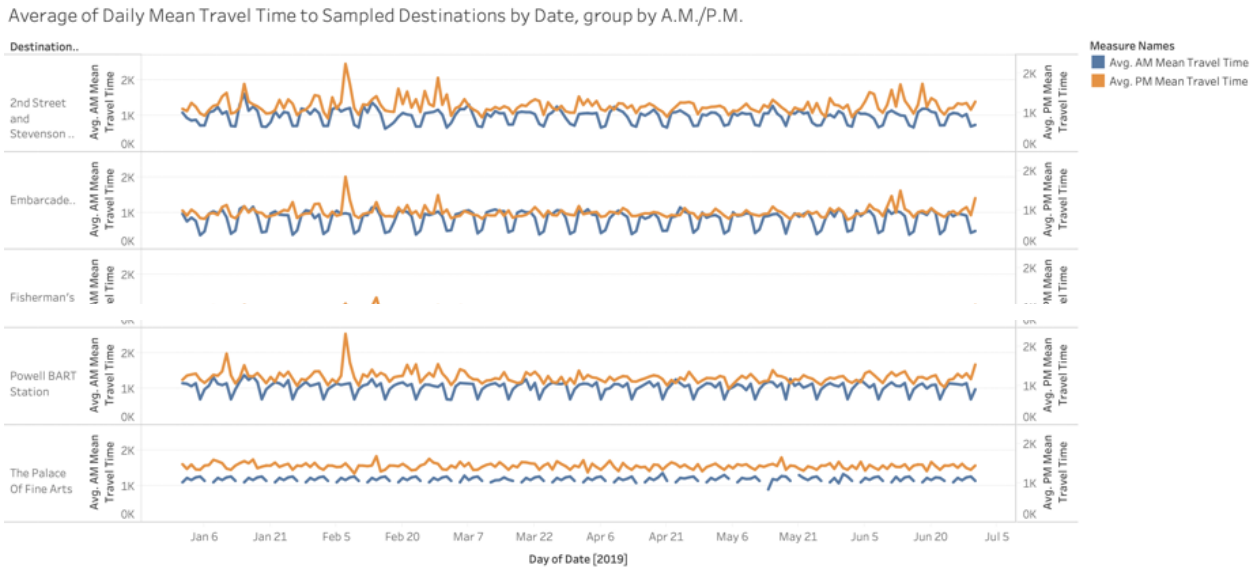
Assumptions:
1: We consider the travel time as the combination of the waiting time and the actual traveling time.
2: We joined the Q1 and Q2 data to see the overall trends for different locations. Although we see that the average traveling times for Q2 are larger than Q1, but the trends are very similar.

Finding 1



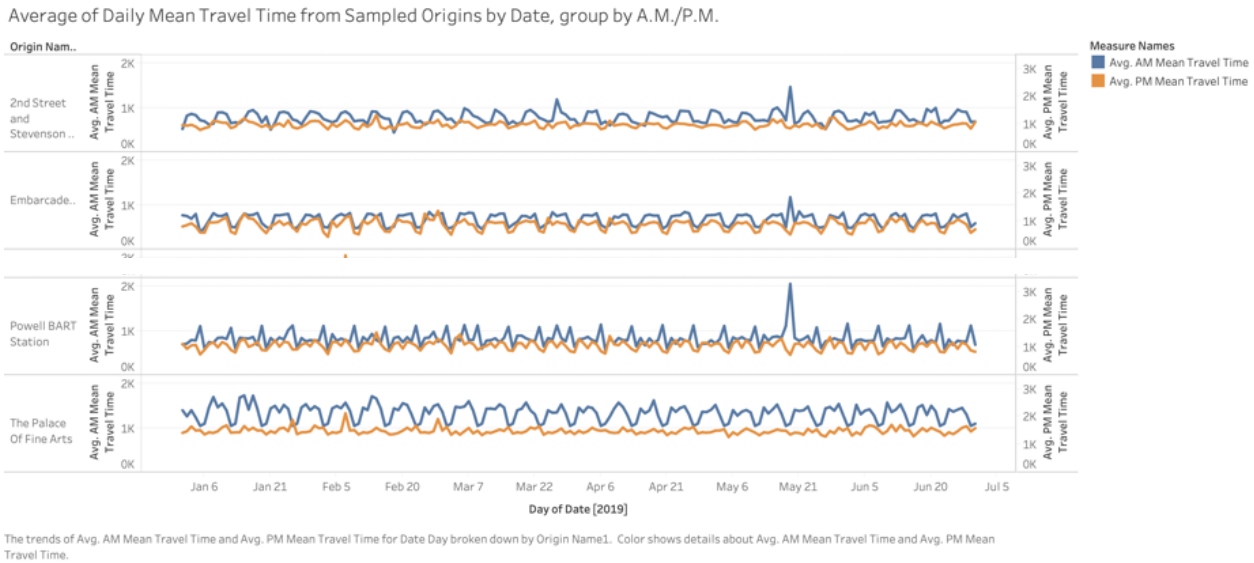
In this chart, we are showing the average daily mean travel time to sampled destinations by date, grouped by A.M. and P.M. Overally, the P.M. mean travel time is much larger than the A.M. travel time. The reason behind this might be that dispatch problem for Uber. During the evening peak, Ubers are more concentrated in the busier part of the city, like the 6 locations we showed in the chart. Then if the passenger is leaving from other parts of the city, they might need a longer waiting time for their Uber.

Finding 2



Another pattern we see is the arched pattern in the "2nd street and stevenson street", " Embarcadero station", "Oracle Park", "Powell Bart station", and "The Palace of Fine Arts" for the morning travelling time. It can be explained by the function of those places. The "2nd street and stevenson street", " Embarcadero station" and "Powell" are next to office buildings which means that people are Ubering there for work regularly for work during weekdays, and not going there so often during the weekend. This can explain why the weekday's time is much larger than the weekend time.

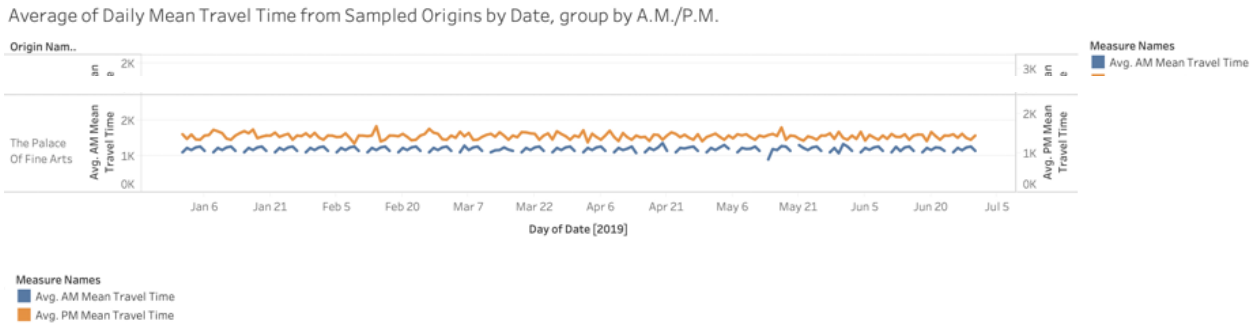
Finding 3



The trends of Avg. AM Mean Travel Time and Avg. PM Mean Travel Time for Date Day broken down by Origin Name1. Color shows details about Avg. AM Mean Travel Time and Avg. PM Mean Travel Time.

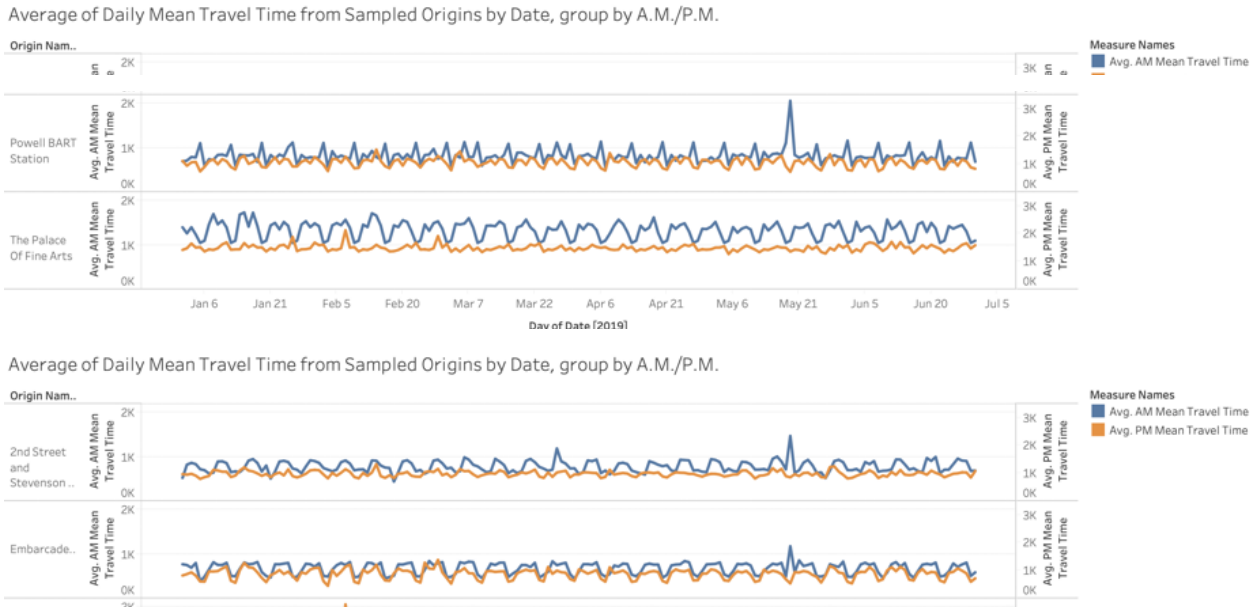
This AM peak for people who are Ubering to these locations also influence people who are leaving those locations, causing the similar arch pattern..

Finding 4



"The Palace of Fine Art" also shows this arch pattern but it's arch pattern has shorter traveling time generally since it is a combination of Residential area and school. During the weekday morning peaks, traffic is busy since people are leaving for work and for school, and less busy during the weekend.

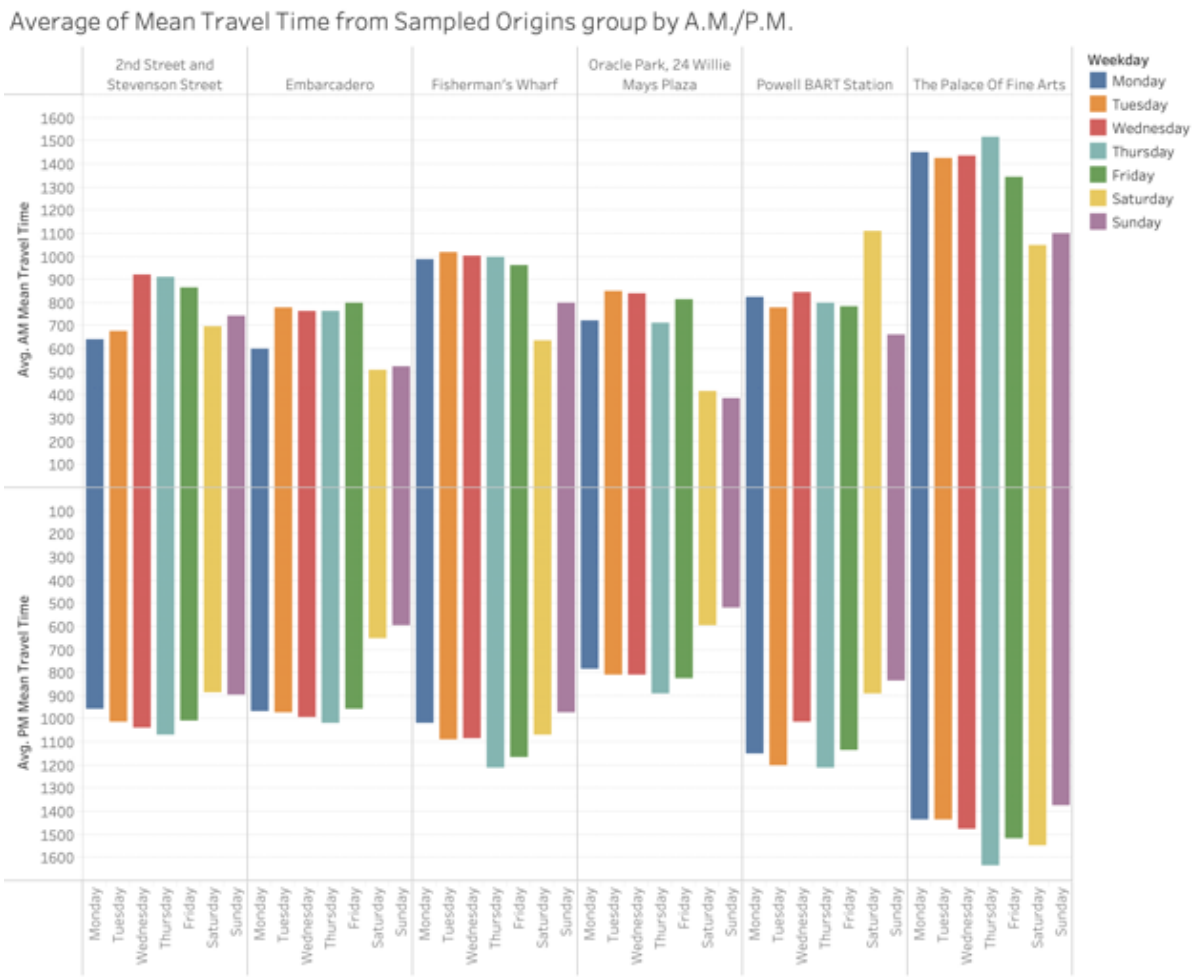
Finding 5



We also identified some weekend P.M. peak for the "Powel", since people go shopping more during the weekend rather than weekdays.

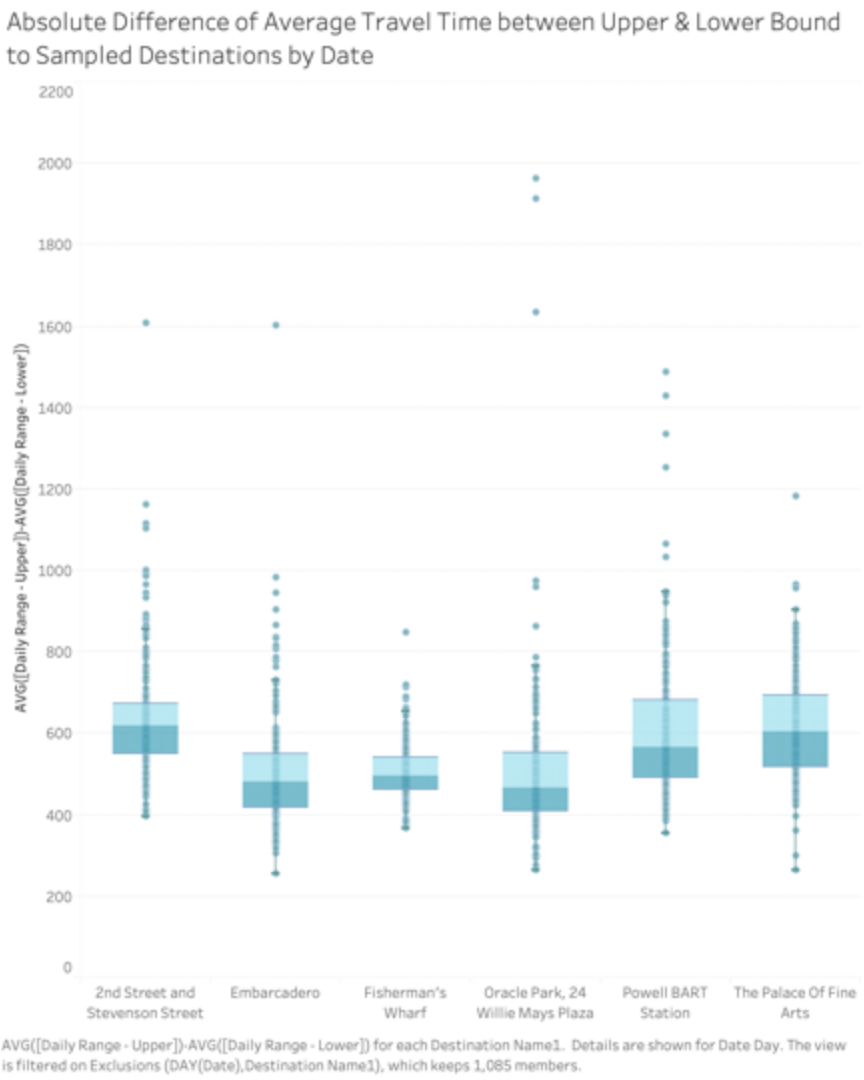
Another interesting point we see is the super peak in the Oracle park on May 19 during the morning hours. With more research, we see that there is a famous sports game happening on May 19 in the evening. The reason that for the morning peak is that people are Ubering to the game area in advance in the morning while the Uber system are not entirely prepared yet. As showed in the seconds chart, there's no P.M. peak since there's more Ubers in the P.M. to reduce the traffic pressure. Also, as showed in other locations, the morning peak origins in other places are also high since people are coming from all over the SF to watch this game.

Finding 6



- In this chat, we are showing the average of mean travel time from samples origins grouped by A.M. and P.M.. The top part indicates the A.M. and the bottom part indicates the P.M. value. We can see that the average time for people who are Ubering to leave the residential area is much larger than the other places. We identified two reasons for it: First, this residential area is less bustling than the financial district, so there's less Ubers in that region which caused a longer waiting time for cars. Second, since it is a residential region, people might Uber to and from somewhere farther than only SF for work. Another reason is that people might leave for somewhere farther during the weekend from the residential area which might also increase the travelling time.
- Suggestion: Preparing more Uber in the residential region and less for the financial regions.

Finding 7



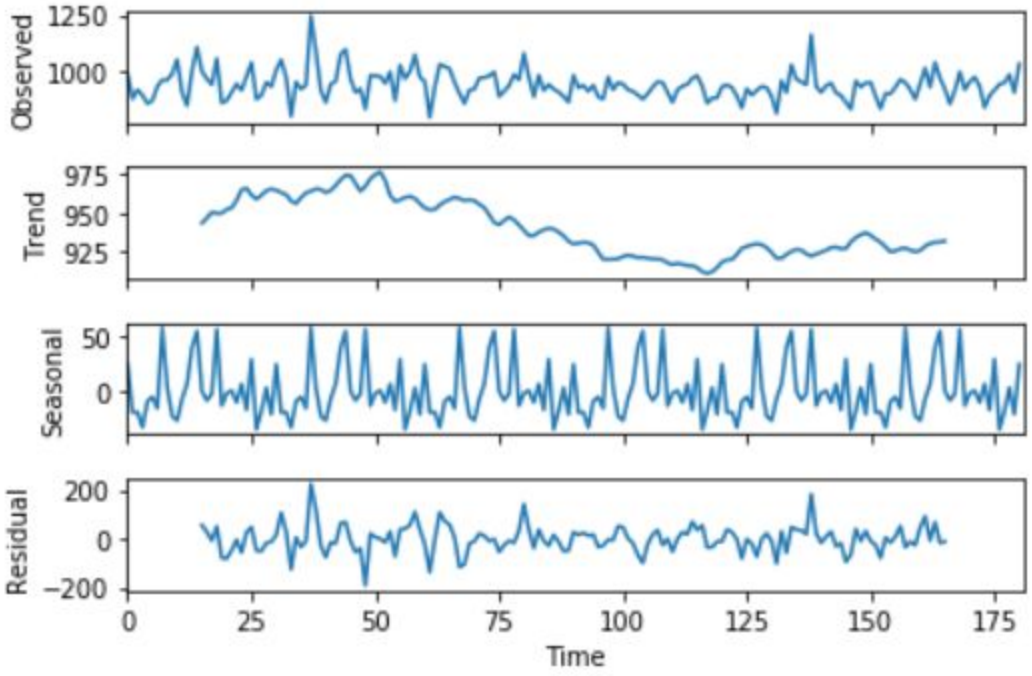
1. The Embarcadero and Fisherman's wharf have a shorter average traveling time since they are related (People usually take Bart to Embarcadero and then uber to the near by Fisherman's wharf). Also, this chart shows that the traffic peak has more influence to places like Powell and the palace of fine arts, since the average is higher.

Section 3: Forecasting/Machine Learning

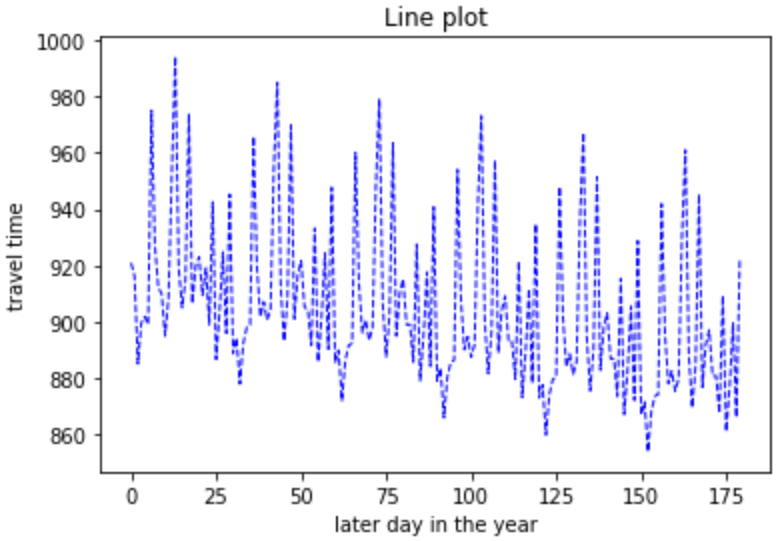
In this part, we provide two way of machine learning analysis.
The first method is applying ***time series specific statistical machine learning model***, it include three models to analyze and forecast the data, seasonal decomposition, SARIMAX, Regressive Analysis grouped by features. The first two models are to show the seasonal trend and long-term

trend in time-series data and forecast.

Model 1.



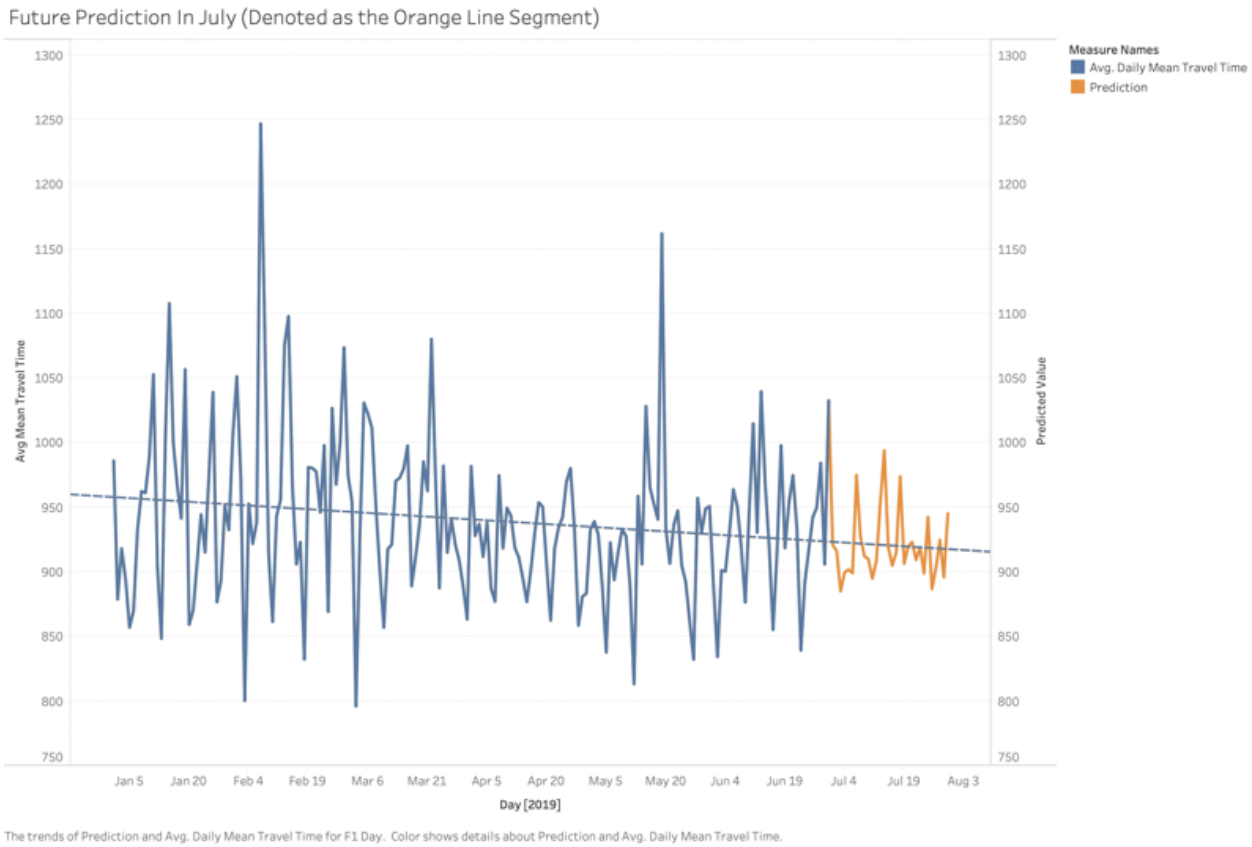
[figure 9]



[figure 9X]

The seasonal decomposition segregates the long-term trend from seasonal trend. From the figure 9 and 9X, the long-term trend indicates that the travel time in SF is decreasing overall, when the monthly trend is eliminated.

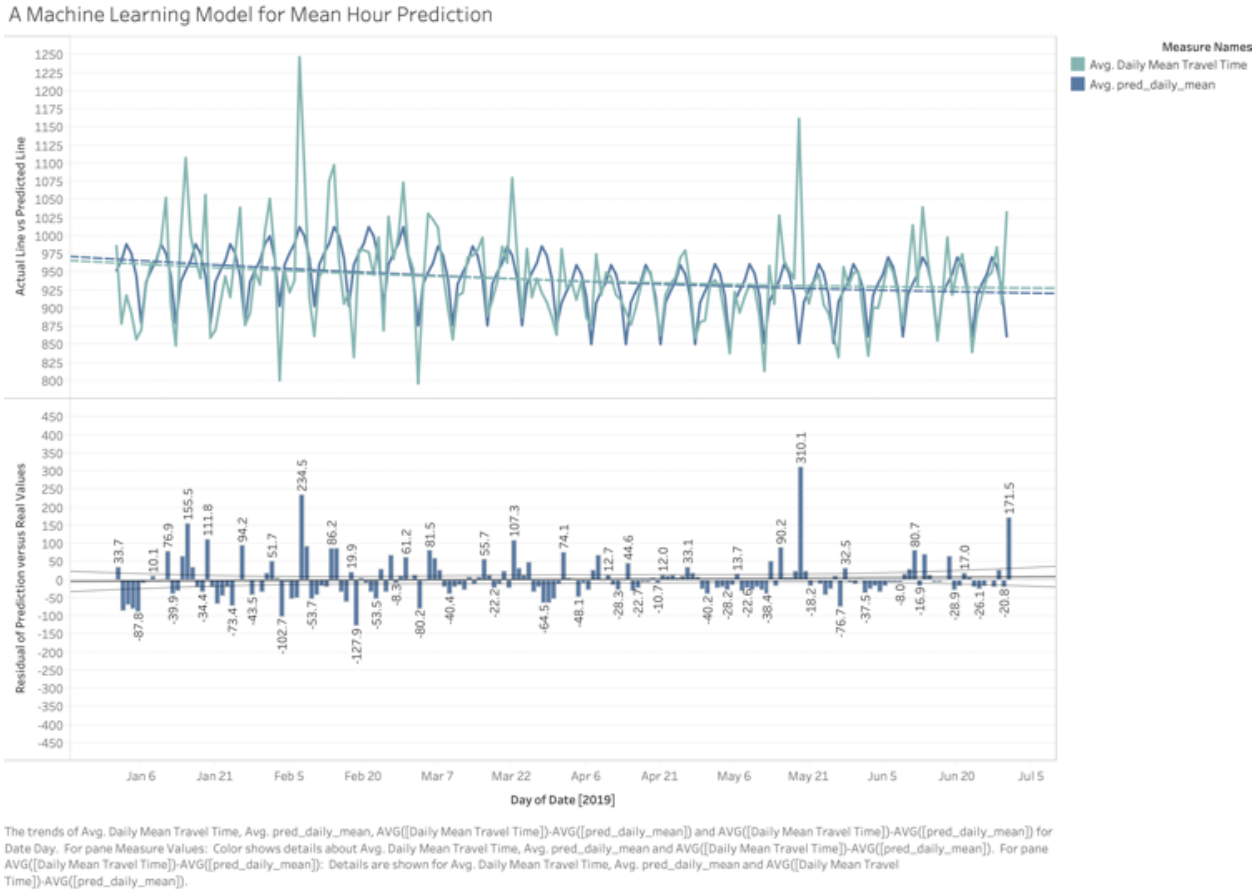
Model 2



[figure 10]

Based on monthly seasonal trend, we define the SARIMA (Seasonal Auto-Regressive Integrated Moving Average) to forecast the next month traveling time from past data. SARIMA is a machine learning model, which fits to explore the seasonal trend in data. In our case, we set the seasonal period as 30 days. In our result (figure 10), the yellow part is the forecasting data from training data in bart_to_all and hotspot_to_all. The line within the chart is the long-term trend in data. From the result, we can know the long-term trend of traveling time in Uber is decreasing, which means that the traffic in SF is getting better, at least in 2019.

Model 3

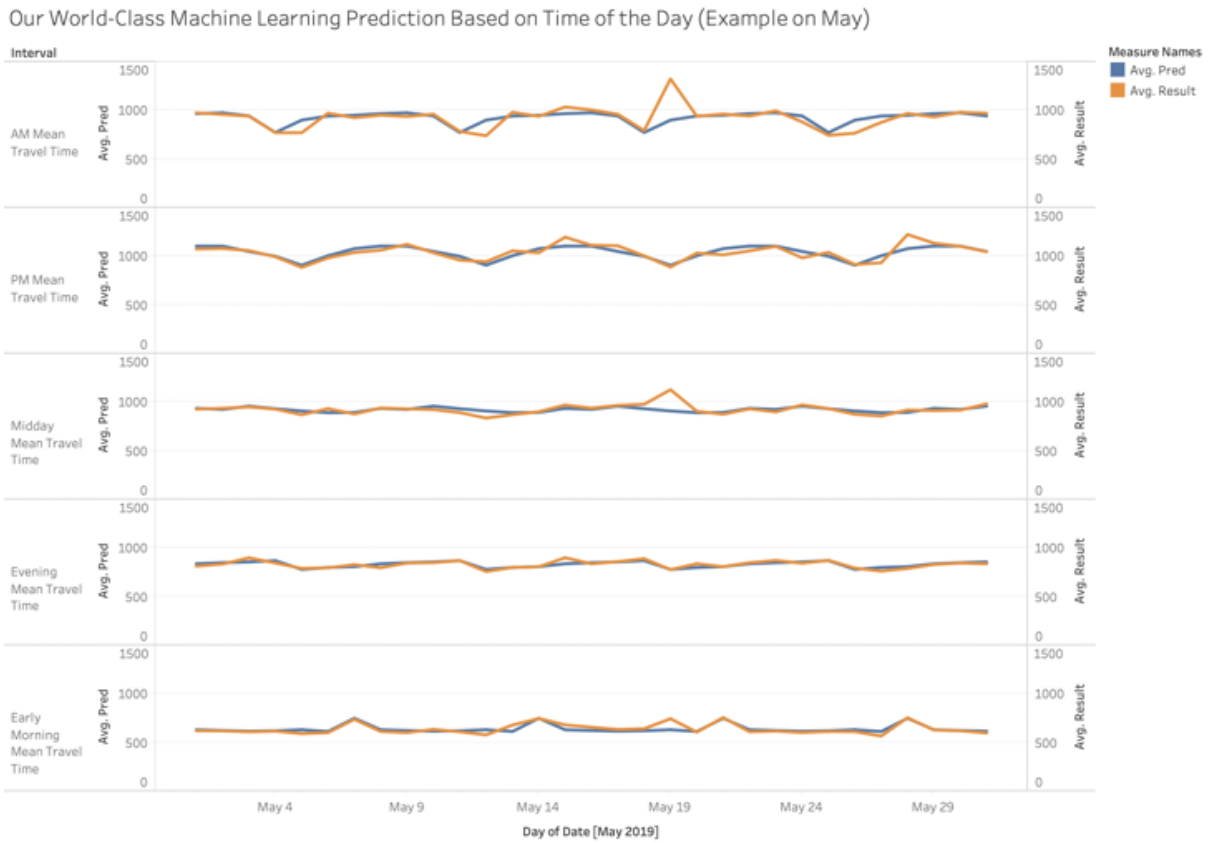


[figure 11]

The second way is to apply strong prediction model into the prediction of time travel.

This graph shows a time-series machine learning prediction for daily mean travel time. We one-hot encoded the categorical labels(origin, destination, month, weekday, day interval)and applied logistic regression with grid search validation with 25% and 75% train test split. RMSE is 130. To increase the performance of the model, we add more features and use bagging of predictors to predict.

Model 4



[figure 12]

Also, in order to make better predictions. We add another feature(day interval, like AM or PM...) to this section. We had applied prediction on the travel time of all 18 combinations of BART-hotspots routes. We first use one-hot encoding to split the original one row data of time interval for the day(AM,PM…) into a different row. And then, we try to predict the travel time for given origin and destination, given month(like January), giving weekday(like Monday) and day interval(like AM). We had applied 25-75 train test split and 5 fold cross validation in the training process. Our final model includes the ***ensemble regressor of decision tree, random forest and gradient boosting.*** We had achieve very high accuracy compare to the original data(RMSE error of 93s)

We had also tried stronger deep learning models like ***GRU and LSTM*** into the prediction of time. Since the number of data between BARTS and hotspots is not enough to well fit the model. We end up didn't improve our accuracy compared to our ensemble regressor.

Section 4: Business Analysis and proposal

Business Proposal:

Introduction:

These days, with the growing car-sharing service, more people commute with Uber. As an international city with a large population, San Francisco faces a serious traffic challenge during the traffic hours. Reducing the traffic pressure for Uber, such as long waiting time or traffic jams, is important for both the rider and the Uber company. We are analysing the Uber dataset to make the Uber riding more efficient.

Problem: Travel time for some Hotpots and BARTs are too long during the weekdays

Solution:

- Suggestions: Uber should send more drivers to the financial area during the weekdays and send more drivers to the mall area during the weekend.
- Reasoning: From section 2 and 3 we see that there are clear patterns for peaks during the weekdays and less traffic time during the weekends for most of the stops beside the the mall area and tourist area. That might be caused by high waiting time due to the shortage of Ubers in that area or busy traffic.
- Testing:
    - We assume that moving more uber drivers to the financial area during the weekdays and more uber drivers to the malls area during the weekend. Then we collect the new average travelling time.
    - If the travel time reduces, then the high travelling time during the weekdays is caused by the shortage of drivers. Otherwise, the high travelling time during the weekdays might be caused by too much traffic during the weekdays.

Problem: Travel time for some special events are too long

Solution:

- Suggestions: Plan more Uber drivers on the major games date for the Oracle park.
- Reasoning: From section 2 we see that there is a super big peak on May 19 for the SF Giants vs. Arizona Diamondbacks. We can see that some big games might influence the Uber travelling time a lot. That might be caused by high waiting time due to the shortage of Ubers in that area or busy traffic.
- Testing:
    - We assume that paying higher to the uber drivers during the game days can shorten the wait time for people who are going to the games. Then we collect the new average travelling time on those big game days to see whether the traveling time still increases a lot.

    - If the travel time doesn't increase a lot, then the high travelling time during the the game days are caused by the shortage of drivers. Otherwise, the high travelling time during the game days might be caused by too much traffic or traffic control.

Problem: During weekdays, the travel time for ubers from other part of the SF to the financial region is much higher than the traveling time from the financial region to other parts of the SF.

Solution:

- Suggestions: Plan less Uber drivers to the business centers and more to the surrounding area.
- Reasoning: From section 2& 3 we see that the travel time for ubers from other parts of the SF to the financial region is much higher than the traveling time from the financial region to other parts of the SF. That might be caused by the shortage of Ubers in the surrounding area.
- Testing:
    - We assume that reducing the riding price for the pooling option for people who are leaving the business area will reduce the need of uber in there, so that we can send more ubers to the soundring areas to pick up people. Then we collect the new average travelling time during the weekdays to see whether the traveling time. We compared the leave and arrive travel time, if both show a decrease, it means our strategy is working.
    - If the travel time doesn't increase a lot, then the high travelling time during the the game days are caused by the shortage of drivers. Otherwise, the high travelling time during the game days might be caused by too much traffic or traffic control.

Section 5: Conclusion

We analyzed the traveling time data among 6 different main locations in SF. In general, we can observe a clear trend that there's less traffic early morning and 100% more traffic on midday of weekday. On weekends, there's an 100-150% increase in the traffic of Powell because it's close to many malls in Union Square so we suggest Uber incentize drivers to work on weekends in the Union Square area.

We used several Machine Learning models, such as gradient boosting, randomForest regressors, seasonal decomposition, SARIMAX and bagging of predictors, to forecast the traveling situation in SF, and we get the result that in long-term trend, the traffic will get better because of a slight decrease of travel time. We suggest the Uber to incentize more drivers during the weekdays, and encourage people to do pooling more during the traffic peak, spread more drivers to regions that get less attention during the traffic peaks, and plan more drivers ahead for big events like big sport games.