Datahacks 2020 Report
Andrea Sudharta and Rui Zheng

Table of Contents

Introduction

We are doing the Business Track for this year's Datahacks competition, and have used data given to us to gather information about traffic trends in San Francisco. The dataset that we are mainly exploring is barts_hotspots.csv. The dataset consists mainly of categorical data and numerical data. We removed all the rows in which a categorical value is missing, and imputed all the missing values in the numerical columns with mean. The second thing we have done is that we created a different version of this table. We took all the time_framed_columns, such as AM Mean Travel Time (Seconds), and created categorical columns. This will help us to visualize what happened to average travel time during a day. We also created another column based on the date of the ride: we converted the date into weekdays. This nre column will help us to visualize what happens to travel time during a week.

Link to Code

Our data was analyzed using Python and can be found here in the form of a Jupyter Notebook: https://github.com/annsudhart/DataHacks2020/tree/master/Workshop_Notebooks

Data Exploration and Observations

All the visualizations we have made can be found in our Jupyter Notebook, but an interactive visualization can be found in https://annsudhart.github.io/DataHacks2020/index.html.

We explored data from many aspects. First we looked at which bart station is the best to take off to minimize the travel time to a Hotspot. We find out for hotspots_3396, the best bart station is 3603. For hotspot 3792, the best bart station is 3692, for hotspot 3394, the best bart station is 3603. We then wanted to explore further on how travel time between different bart stations and hotspots change during a week. Then we find out that the travel time peaks in the midweek and drops significantly during the weekend. We also looked at the direction of each bart-hotspot travel. We searched online if the travel direction is to the center of san francisco or not, and analyzed time traveled according to this, We find out that it takes much more time to travel to

the center than away from the center. We also looked how travel time changes during different time of the day, and find out that the travel time is significantly higher in PM than AM. In addition to this, we noticed a significant drop in time traveled on saturday and Sundays for both AM and PM. We also find out that early morning travel time is very stable throughout the week. That suggests early morning travel time can be the time required to travel between barts and hotspots without all the traffic in san francisco.

Hypothesis Testing
We have made 2 hypothesis testings:

Hypothesis Testing 1
Null Hypothesis：early morning distribution throughout the week is the same as that of other time periods.
Alternative Hypothesis:early morning distribution throughout the week is lower than that of other time periods

This null hypothesis is rejected, suggesting that the travel time is less in early morning so people can go out early to minimize their travel time.

Hypothesis Testing 2
Null Hypothesis：The time length distribution to center is the same as that of from center.
Alternative Hypothesis:The time length distribution to center is not the same as that of from center

This null hypothesis is rejected, suggesting that the travel time is less if you are travelling to the center, so people should take that into account and go out early to avoid being late if they are going out.

Creating a Machine Learning Model
We think three features are significant for this forecast: tocenter, dow(day of a week), Time (time of a day). We hot encoded these features and used linear regression to predict the future travelling time.

Conclusions
- Travel time drop significantly on weekends
- Travel time to the center is usually higher than that of away from center. Most hotspots are away from the center, most bart stations in the station.
- Travel time peaks during midweek, usually Thursday.
- Travel time peaks in PM and midday
- Travel time minimizes in early morning
- If you are going to center, make sure to encourage carpooling options

- Encourage drivers to operate (with carpooling recommended) near the city and during PM and midday, as they are the busiest and have high demand