Capstone 1 Milestone Report - Lyft Baywheels Analysis

Context

This summer Lyft rebranded its Ford GoBikes to create a fully Lyft branded suite of share bikes, black with the quintessentially Lyft-like pink rims. I started to see them all over the place in SF and the East Bay and began to wonder about them.

Doing some digging, I realized that this comes out of a longstanding lawsuit with the city of San Francisco and is an effort by Lyft to re-establish control over the bikeshare market in SF. The issue was that Uber's JUMP bikes had encroached on a supposed 10 year agreement that Lyft had, stating they had sole operational privilege of share bikes in the city.

And this just in, it appears the bikes are having troubles of their own! These data can be used as a historical account of Lyft Baywheels (or previously Ford Gobikes) rides in the Bay Area, but may not be available going forward since they're pulling the fleet.

In what can be presumed to be an effort of good will towards the data-savvy SF public, Lyft opened up their anonymized Baywheels trip data, showing 2.5 years of rides as well as API access to real-time trip data - the locations of every bike. I discovered the historical trip data and wanted to dive in to explore a number of questions. The live data may be incorporated as well in the future.

Problem statement

In looking at these data, I hope to answer the following questions:

1. What is the distribution of categorical and discrete variables like user type, age, gender, etc.?
2. What is the optimal number of bikes in the system to optimize revenue?
    a. How does system revenue scale with each new bike?
    b. Can the cost of a new bike be modeled?
3. Are there any obvious patterns such as: men contribute more to revenue, people in x age group result in more rides, more rides happen after 10:00 PM, etc.?

In answering these questions and in performing EDA on these data, a lot of interesting implications for how people move about SF can be gleaned as well that are ancillary to this main thesis.

Stakeholders

The key stakeholders would be Lyft, people involved in the share economy more generally, and perhaps data-savvy bike-share users as well. Perhaps this analysis could show to the city of

San Francisco that with enough investment and enablement, Baywheels could provide the city's bike sharing needs. The main reason the city of SF is fighting back on the lawsuit was that they wanted more bikesharing offerings. Maybe it's about volume, or maybe it's about a fair and evenly distributed market. The reason it matters is that there may be more optimal places to position stations/bikes to improve SF mobility and boost revenue.

The Data

Lyft has made available a directory of data hosted on AWS, illustrating bike-share rides over the past two and a half years. Details on these data and the link to the real-time data feed is available on their Systems Data page. I wrote a Python script and scraped all historical trip data, concatenating them together into 3.4M rows of rides.

Acquiring the Data

Once I acquired a list of all the filenames from the directory, I downloaded all the zip folders containing them and unzipped each locally.

The Master Dataframe

The final dataframe I ended up working with is a dictionary of dataframes I unzipped and finally concatenated. I also grabbed the station and region IDs from the real-time system data API, joined those on station IDs in the master dataframe, and then filtered by the SF region ID because the dataframe was unnecessarily large for modeling. It originally contained SF, the East Bay, and San Jose bike rides.

Note: I did not include all data from the original repository because 2019 is only up to May. If I did any analysis on the year or month, the 2019 data would be underrepresented in the sample and thus would make it difficult to draw inferences from.

Cleaning the Data

For cleaning the data, I decided to remove the names of the stations and the bike_share_for_all_trip columns because I already have station latitudes and longitudes so the names are distracting/unhelpful. The bike_share_for_all_trip column is a marketing campaign that was a very small subset of the data and I didn't care to analyze it.

Glancing at the data I saw that rides were in seconds, so I added a column duration_min, which converted it to minutes - this seemed to make more sense and be more useful. I also converted the start and end times to pandas datetime objects since they were strings initially. This was so I could do Time Series analysis.

I encoded the gender columns as floats because these process faster than strings. I was also able to interpolate the age of the riders from their member_birth_year column.

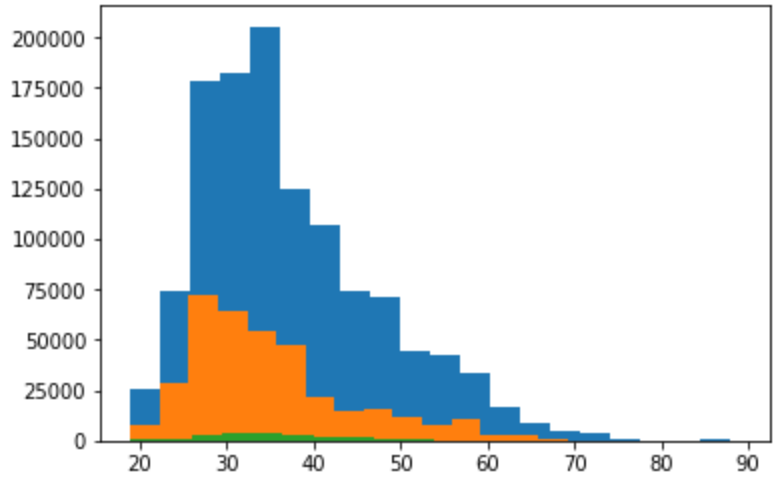
I created columns for day, hour, etc. from the datetime objects previously created. I wanted to look at a heatmap of rides by day and hour. However, because not all months of the year have 31 days, I removed the 31st day. This was hard to add into the main dataframe since it didn't make sense that each row corresponds to a ride but would also have a frequency of rides by day column. As such this sits in a different dataframe: days.

Exploratory Data Analysis (EDA)

Distributions

In order to understand what the data looks like, I looked at the distributions.

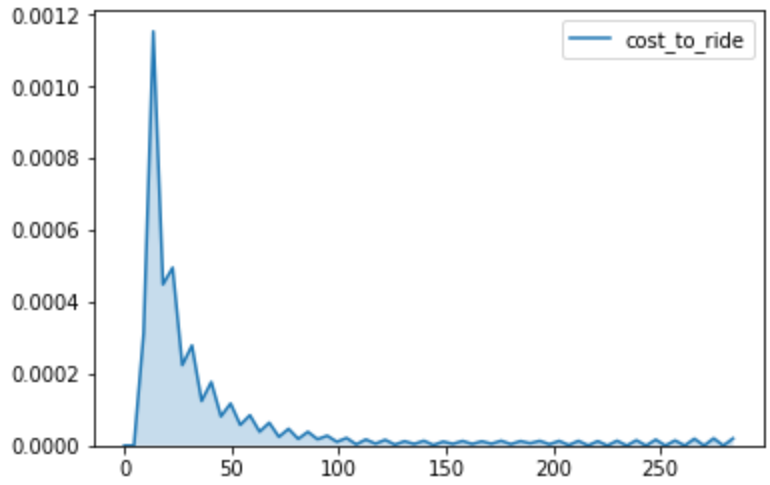
Age by gender histogram (blue=male, orange=female, green=other):



From the above chart we can glean several things:

1. The age distributions for all genders appears to be right skewed
2. The median age seems to be somewhere between 30 and 35
3. There are more men than any other gender in these data
4. There are about half as many women than men in these data
5. Other is the least common gender type

The revenue gains from rides appears to be exponentially distributed (and is highly correlated with the duration of the ride since that determines the cost of the ride):

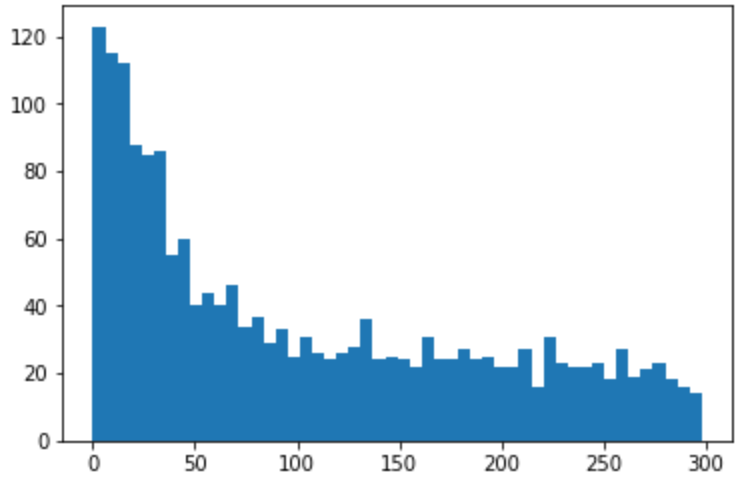


The majority of costs are below $25 it appears for a given ride.

Interesting findings

If we had assumed the revenue generated by bikes was normally distributed, then the number of people who took bikes and the duration of rides would also be normally distributed. Each bike would also have a normally distributed likelihood of being selected. This was in fact not the case (and of course none of the above variables would be normally distributed because bikes could be anywhere in the city, and people who need bikes are likely concentrated in the financial district and similar areas):

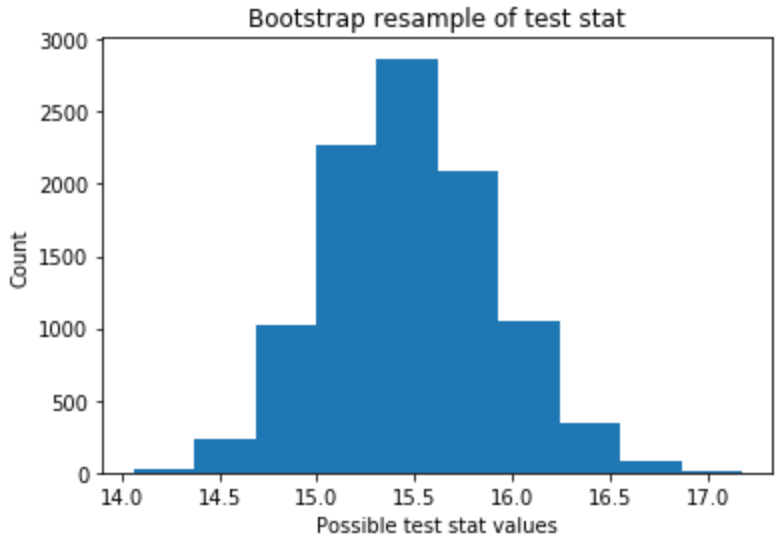
Revenue by Bike ID (Y-axis = total money made):



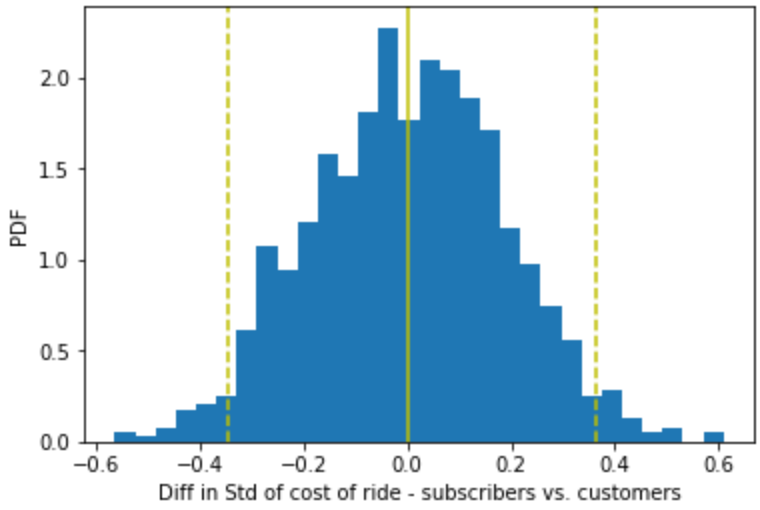
There are bikes that make well over $25 total from rides in these data.

Statistics

I performed bootstrap resamples of test statistics to see that I could be 95% confident they came from the population, and that was indeed the case as shown below for the average duration in minutes of rides:



I was also curious about whether or not there was a difference between customers and subscribers and how much they spent on a per-ride basis. The average difference in the standard deviations of their costs of rides did not differ significantly either. Yellow dashed lines indicate the cut-off points of the 95% confidence interval.



Machine Learning

This segment is coming soon! Given the state of the data and the variables and nuances discovered in the cleaning and EDA steps, I can perform regression modeling on the following: How much new revenue is injected into the bike system with each new bike added?

I can also perform a classification algorithm to answer questions of the type:
Given a trip of x duration in minutes and y ending location, is the rider male or female?

I also hope to answer the question: *can the cost of adding a bike to the system be modeled?* This will help me understand the opportunity-cost of adding bikes to the system as it probably doesn't scale linearly towards positive infinity.