**Uber Vs Lyft, A Comparison.**

Adam Flory, Rory Berry, and Spencer Hurley.

**Dataset:**

We used Kaggle.com to find a large set of data collected on Uber and Lyft rides provided in Boston Massachusetts. The Data set we used is titled “Data Mining Project – Boston” uploaded by use SophiaLiu. This data set consisted of over 600,000 rides provided by the rideshare services over the course of 18 days in late November and early December 2018. The data set include the drop off and pick up points for each ride as well as the date, time, price, company and service used (ex. Uber Black), and various weather conditions (ex. Temperature, precipitation). Luckily the data set was already very clean so we only really had to narrow things down to the specific variables we wanted to use. The largest problem with the data set was that it was collected over such a short period of time. This made trying to track usage considering things like weather somewhat unfruitful, as it would be challenging to establish a verifiable pattern with such little data. This led us to look into other items most specifically the overall usage of each company and there services when taking things like time of day, type of service and pick-up/drop-off point.

**Does the market share between Lyft and Uber fluctuate as the total ride volume changes?**  
Null hypothesis – Market share will fluctuate as volume changes?I went into the project thinking there would likely be some relation between Uber and Lyft that would be a tug of war over market share with each company increasing ride volume as the other would be losing ride volume at the same time. I initially had to separate out the Lyft and Uber data and put it into a stacked bar graph to get an initial visual on how the data laid out. This showed me that there didn’t look to be any large fluctuating differences in market share over the referenced time period.From there, I created a line graph to show the rate of change in Uber’s lead in the market and the total volumes of both other services. These further showed no relation as there was no consistent pattern between volume and market share. Then following with the confirmation of my null hypothesis to show that the ranging volumes do not relate to market share.

**A close up of a map

Description automatically generatedA picture containing writing implement

Description automatically generated**

**Are rideshare services used more often to some areas of the city than others?**

Short answer yes, while the graph below does not appear to show a much difference between different destinations, running a Chi-Squared test rejects that each destination receives the same amount of traffic. To run the chi squared test I found how many rides ended in each destination, think the burrows of New York, to find the observed values. For the expected values I divided the total number of rides by the number of possible destinations, so every destination had the same number of rides. After running the Chi-Squared test which showed a p-value of .00129. This means that we can reasonably say some areas of Boston see more traffic from ride share services than other, which is not to surprising.

A screenshot of a cell phone

Description automatically generated

**Is the service used affected by the destination?**

Since the answer to the last question seemed pretty obvious we also decided to look at whether any areas of Boston see more than expected traffic from one ride share service or the other. in order to took into this we decided to again run a Chi-Squared test, however this time the observed values were calculated using the difference between the number of uber rides dropping off in a particular area and the Lyft rides dropping off in that area. The expected values where the difference between all Uber and Lyft rides in the data set divided by number of drop off points. This time the Chi-Squared test produced a value of .4004. which means that we can reasonably say that there is no difference in service preferences between different destinations.

A screenshot of a cell phone

Description automatically generated

**Does Uber X / XL / Black / SUV (i.e. 4-6 passenger car types) see more usage than comparable Lyft vehicles?**

Null hypothesis – Are there any differences/does any Uber car type matter?

1. NO - If we take a look at distance, we can quickly see that car type does not matter. They are all showing the same value.
2. YES - But, if we look at price, then 2 models show differences with 1 car type showing the largest difference. (based on the data)

Hurdles:

* Finding good and relevant data- We found data on Kaggle.
* Dataset – 2 months of Uber and Lyft data – Boston, MA (NYC favors Uber heavily)
* Collecting meaningful data… eliminating distance data
* Finding the right comparison points. We used 4 different car types and eliminated single rider and “shared rides” or “pools” that can skew the data.
* Finding the right order to process the data. Earlier, smaller datasets worked by simply reading the data, filtering the data and then displaying the chart. But, as we collected 2 months of data for Boston, we also needed to capture the “mean” and then “Group By” before charting.

What worked:

1. Kudos to the Team, Spencer and Adam, for making Github process work for us and for finding relevant data and then combining the Uber and Lyft data so that we could quickly filter and use it including for Statistical Analysis purposes.
2. We liked the speed and ease of working with Pandas/Jupyter Notebook. Errors quickly presented themselves. We always could view the latest data and the speed was excellent for processing 400,000 rows of data.
3. We could look at the data and column headings or charts fairly easily and then decide if changes or enhancements were needed.
4. Visibly, we can see that there is a “Mean” difference between Uber X and Lyft Lux (8) and then with Uber Black and Lyft Lux Black (4). With the amount of data that we found, it was difficult to find any differences between groups and these 2 differences are based on Price.
5. To conclude, I would say that we should “**Reject the Null Hypothesis**” (after using the ANOVA Procedure (i.e. for more than 2 values). We can visibly see differences based on Price.

**Slides:**



