LYFT DATA ANALYSIS

Bernard Wong, Jonathan Gong

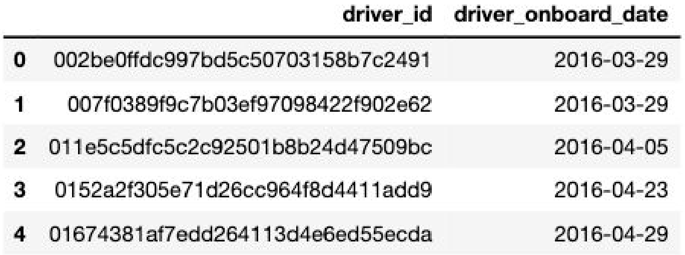
INTRODUCTION

The ride sharing industry has increased significantly in the past few years. With the Ride Sharing Market predicted to reach 220.5 billion USD by 2025 and company goals of expansion to areas such as Asia and Europe, it's undeniable that companies such as Lyft are playing a dominant role in the way that society thinks of transportation. With our analysis on Lyft drivers and the business as a whole, we hope that we'll be able to provide valuable insight on what the next best steps are for Lyft so that they can successfully increase business while still providing high quality ride sharing.

UNDERSTANDING AND CLEANING THE DATA

We were given 3 different datasets; one giving us the information about drivers (driver_ids), one giving us information about rides (ride_ids), and one giving us information about the timing with rides (ride_timestamps). Before any data manipulation and analysis, we made sure that we understood the data and that the data was pure and wasn't missing or erroneous.

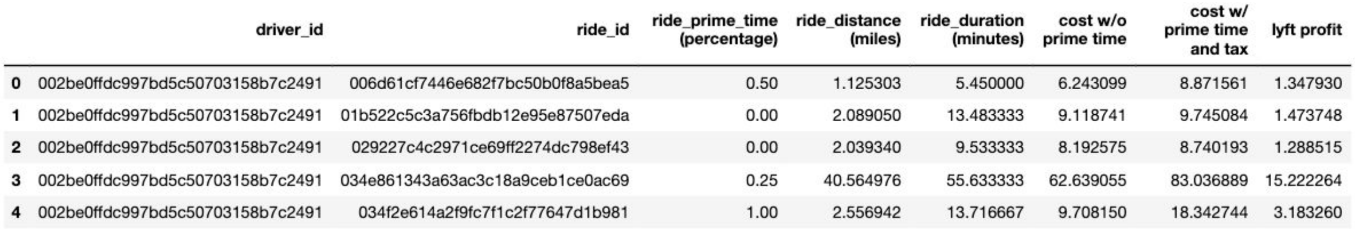
Drivers Dataset



	driver_id	driver_onboard_date
0	002be0ffdc997bd5c50703158b7c2491	2016-03-29
1	007f0389f9c7b03ef97098422f902e62	2016-03-29
2	011e5c5dfc5c2c92501b8b24d47509bc	2016-04-05
3	0152a2f305e71d26cc964f8d4411add9	2016-04-23
4	01674381af7edd264113d4e6ed55ecda	2016-04-29

The Drivers dataset is relatively simple and clean. The dataset contains driver IDs and when each driver joined Lyft. Overall, the onboard dates were realistic and there were no repeats of driver IDs, making this dataset ready for use.

Rides Dataset



	driver_id	ride_id	ride_prime_time (percentage)	ride_distance (miles)	ride_duration (minutes)	cost w/o prime time	cost w/ prime time and tax	lyft profit
0	002be0ffdc997bd5c50703158b7c2491	006d61cf7446e682f7bc50b0f8a5bea5	0.50	1.125303	5.450000	6.243099	8.871561	1.347930
1	002be0ffdc997bd5c50703158b7c2491	01b522c5c3a756fbdb12e95e87507eda	0.00	2.089050	13.483333	9.118741	9.745084	1.473748
2	002be0ffdc997bd5c50703158b7c2491	029227c4c2971ce69ff2274dc798ef43	0.00	2.039340	9.533333	8.192575	8.740193	1.288515
3	002be0ffdc997bd5c50703158b7c2491	034e861343a63ac3c18a9ceb1ce0ac69	0.25	40.564976	55.633333	62.639055	83.036889	15.222264
4	002be0ffdc997bd5c50703158b7c2491	034f2e614a2f9fc7f1c2f77647d1b981	1.00	2.556942	13.716667	9.708150	18.342744	3.183260

The Ride dataset contains driver IDs, ride IDs, ride distances in meters, ride duration in seconds, and the prime time percentage. Unlike the Drivers Dataset, the Rides Dataset has repeats of Driver ID. However, each ride ID is unique which means that we have distinct information for a variety of different Lyft rides.

Because we knew we wanted to do an analysis regarding different types of costs, we decided to create extra data about the cost to the rider with and without prime time as well as calculate the profit that Lyft would get with each ride. Because our mileage rates were in miles and our duration rates were in minutes, we first converted the distance and duration to the proper units. Afterwards, we created formulas for the cost of the ride for the rider including prime time and San Francisco tax and the profit Lyft made per ride assuming that they take 20% of the cost of the ride (according to Money Under 30). Here were the formulas that we created:

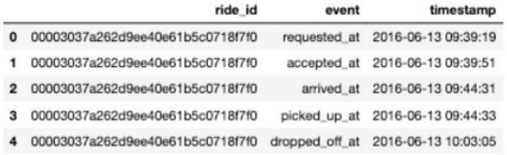
Base Fare = $2.00 + (ride distance in miles) * $1.15 + (ride duration in minutes) * $0.22
Cost per ride = $1.75 + Base Fare + Base Fare * (ride prime time percentage) + Base Fare * (.085)
Lyft Profit = (Base Fare + Base Fare * (ride prime time percentage)) * .20

LEGEND: 🟥 *= base fare,* 🟧 *= cost involving mileage,* 🟨 *= cost involving duration,* 🟩 *= service fee,* 🟦 *= prime time increase,* 🟪 *= cost involving San Francisco Tax*

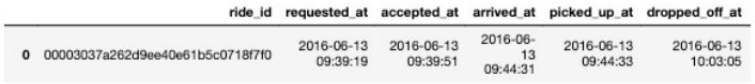
As a result, we now had a dataframe that also included information about the cost to the rider as well as the profits that Lyft makes for each ride.

Time Dataset

The time dataset contained information about ride IDs, types of events (which include when the ride was requested by the rider, when a driver accepted the request, when the driver arrived at the requested location, when the rider entered the driver's vehicle, and when the driver dropped off the rider), and the timestamp of each event occurring. We noticed that each ride ID repeated five times and instead had unique timestamps for the five different events that occurred in each ride. As a result, rather than have a ride repeat five times, we decided to transform and pivot the data frame to make it more organized.

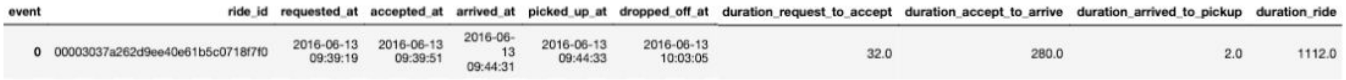


	ride_id	event	timestamp
0	00003037a262d9ee40e61b5c0718f7f0	requested_at	2016-06-13 09:39:19
1	00003037a262d9ee40e61b5c0718f7f0	accepted_at	2016-06-13 09:39:51
2	00003037a262d9ee40e61b5c0718f7f0	arrived_at	2016-06-13 09:44:31
3	00003037a262d9ee40e61b5c0718f7f0	picked_up_at	2016-06-13 09:44:33
4	00003037a262d9ee40e61b5c0718f7f0	dropped_off_at	2016-06-13 10:03:05



	ride_id	requested_at	accepted_at	arrived_at	picked_up_at	dropped_off_at
0	00003037a262d9ee40e61b5c0718f7f0	2016-06-13 09:39:19	2016-06-13 09:39:51	2016-06-13 09:44:31	2016-06-13 09:44:33	2016-06-13 10:03:05

Along with that, we created some additional columns that had the measurements between each of the events. This information tells us how long the rider has to wait to get a ride, how long the driver has to wait to pick up the rider, and how long the trip is overall.



event	ride_id	requested_at	accepted_at	arrived_at	picked_up_at	dropped_off_at	duration_request_to_accept	duration_accept_to_arrive	duration_arrived_to_pickup	duration_ride
0	00003037a262d9ee40e61b5c0718f7f0	2016-06-13 09:39:19	2016-06-13 09:39:51	2016-06-13 09:44:31	2016-06-13 09:44:33	2016-06-13 10:03:05	32.0	280.0	2.0	1112.0

While double checking the data for validity and missingness, we noticed that the amount of time the driver had to wait to pick up a rider was incorrect for some of the rides. Some of the measurements were negative, meaning that the driver had picked up the rider before arriving at the location. Because there was a reasonably sized portion of rides with negative pick up times, we decided to replace the times with a placeholder (np.NaN) rather than get rid of the data. We were cognisant of this erroneous data and as a result we didn't do too much data manipulation regarding that column. However, the other columns had relatively reasonable data points and no missing data or negative times, so as a result we feel that this data can still prove to be valuable. After rotating and cleaning, the dataset is now ready for use.

COMBINING THE DATAFRAMES

While cleaning the data, we noticed that the driver dataframe had driver_id's to help identify each driver, the timestamps data frame had ride_id's to help identify each ride, and the ride dataset had both driver_id's and ride_id's. As a result, we thought that it would be valuable to merge all three based off of each ID so that we could have a master dataset that included information about the drivers, the ride, and the times.

By using the ride data frame as a middle ground, we'd be able to merge the driver data frame and ride data frame based off of driver ID's and merge the time and ride data frame off of ride ID's.

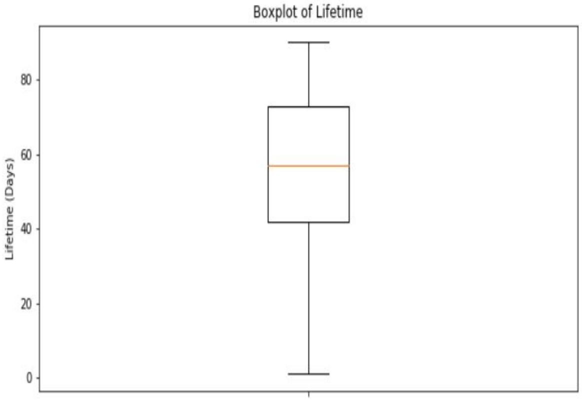
While merging the datasets together we noticed that not all driver IDs from the driver dataset matched the driver IDs from the ride dataset, as well as not all the ride IDs from the time dataset matching the ride IDs from the ride dataset. Because we wanted temporal data regarding the drivers, the merge between all three datasets would still be valuable, so we decided to drop any of the data that didn't have matching IDs to the ride dataset. There were a few factors that helped us come to this decision. One reason is that we still had plenty of data to work with; if we only used data points that had complete information on drivers, rides, and times, we had 184209 data points or 95% of the ride data which is a large amount of data. Along with that, data imputation for the driver or time dataset wouldn't have been very valuable. We weren't sure whether the drivers drove no rides or simply were not measured in the rides dataset and the missing time data only accounted for 1%, so imputation wouldn't have been a good usage of time.

In the end, we were able to merge everything and create a master dataset of 184209 points that contains information about the ride, the driver, and the times. This would prove to be extremely useful as we work on analyzing the data.

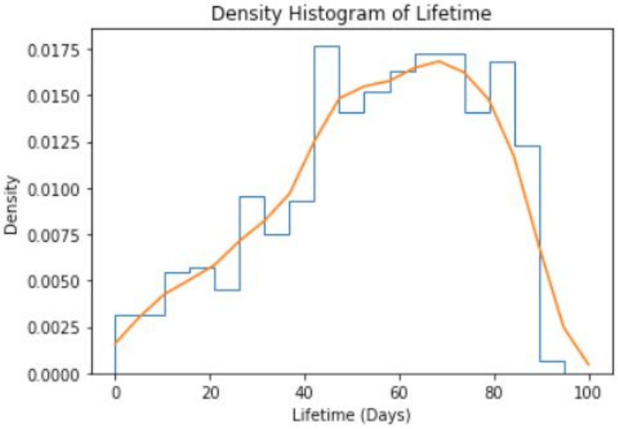
ANALYSIS AND RESULTS

What is the average projected lifetime of a driver?

Our final merged data frame combines both information about driver and times which means we can start to predict average projected lifetime of a driver. Before we could even begin to calculate averages though, we needed to determine what the lifetime of each Lyft driver in our dataset was. We determined that the number of days between the driver's onboard date and the driver's last drive would represent the lifetime of a driver. As a result, we were able to manipulate the data frame and come up with a lifetime expectancy for each of the drivers in our dataset.



Now that we have data on the lifetime of each driver we can start to do some statistical analysis to determine whether the mean or median would be a better representation of lifetime. Because the mean is strongly affected by outliers and skewness, we decided to take a look at the distribution of driver lifetimes.



The first item that we created was a boxplot of driver lifetimes. This boxplot helps us easily identify any outliers that lie outside the 1.5 IQR range. Because there are no data points outside the bounds we see that there are no outliers in driver lifetimes, which is a great sign for the data.

We also calculated the skewness and kurtosis of the lifetimes, which we found to be -0.48 and -0.56 respectively. The skewness tells us that the data is slightly skewed to the left, but because the absolute value of the skewness is still less than .5 we can conclude that the data is relatively normal. The kurtosis value tells us that there's a low chance for outliers, which reinforces the results we got from the boxplot.

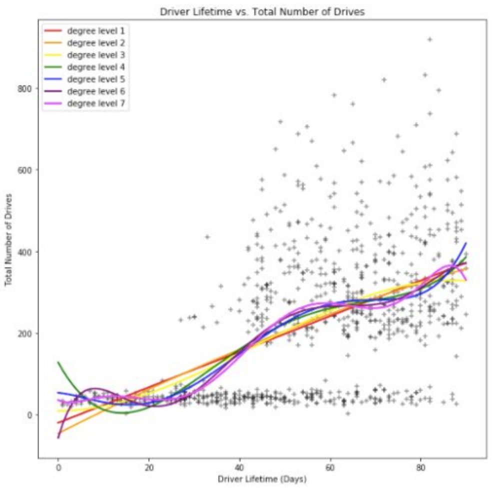
Plotting out the density histogram of lifetime confirms our previous results. We see that the lifetimes are relatively normal meaning that there shouldn't be too big of a difference between the mean and median being a representation of the average lifetime. Because there is still a very slight skewness though, our conclusion is that the median will most likely be a better representation of the average projected lifetime, which means that <u>once a driver is onboarded, we predict that they will typically continue driving for Lyft for 55 days.</u>

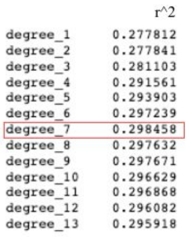
<u>What is a Driver's Lifetime Value, or the value a driver brings to Lyft over the average projected lifetime?</u>

There are a few factors that we believe make a driver valuable to Lyft which include profitability, quantity of business, and quality of business. Profitability is the most obvious and measurable amount and is the amount of profit Lyft makes (in USD). However, profit is not the only thing that makes a company successful; quantity of quality of their business also strongly determines its success. As a result, we decided to also measure the number of rides (which is a measure of the quantity of business that Lyft does) and the speed of service (which is a measure of the quality of business and can be correlated to customer satisfaction) that a rider provides in a projected lifetime.

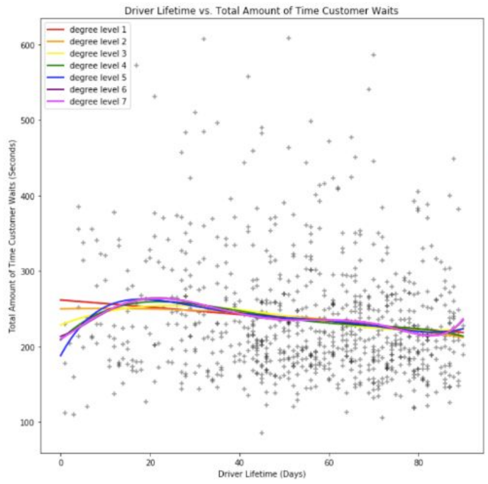
We went ahead and plotted Driver Lifetime vs. Lyft profit, Driver Lifetime vs. Number of Rides, and Driver Lifetime vs. Total Amount of Time a Customer Waits in order to find some trends. Because we wanted to predict these values as well, we plotted a few lines of best fit to help us with our predictions. For each one of these factors, we utilized machine learning and linear regression algorithms to find the line of best fit.

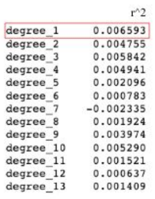
Our process was simple. For each of these factors, we created 200 70/30 testing/training datasets and then used them to create polynomial lines of best fit from the 1st degree to the 13th degree. We then took the average r^2 for each of these polynomial lines and selected the polynomial line that had to largest r^2 to use on our data. By dividing up the data into testing/training data, we avoid overfitting or underfitting the data.



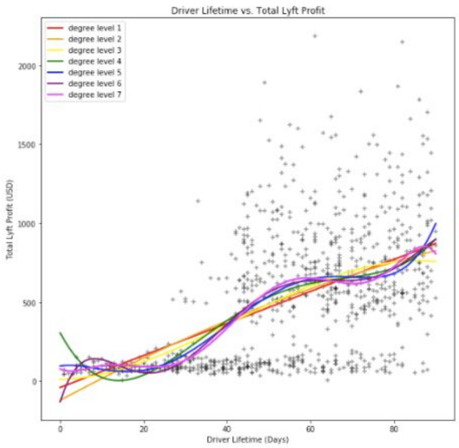


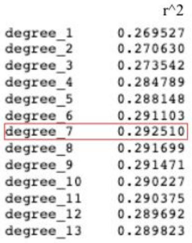
```
                r^2
degree_1   0.277812
degree_2   0.277841
degree_3   0.281103
degree_4   0.291561
degree_5   0.293903
degree_6   0.297239
degree_7   0.298458
degree_8   0.297632
degree_9   0.297671
degree_10  0.296629
degree_11  0.296868
degree_12  0.296082
degree_13  0.295918
```





```
                 r^2
degree_1    0.006593
degree_2    0.004755
degree_3    0.005842
degree_4    0.004941
degree_5    0.002096
degree_6    0.000783
degree_7   -0.002335
degree_8    0.001924
degree_9    0.003974
degree_10   0.005290
degree_11   0.001521
degree_12   0.000637
degree_13   0.001409
```





```
                r^2
degree_1   0.269527
degree_2   0.270630
degree_3   0.273542
degree_4   0.284789
degree_5   0.288148
degree_6   0.291103
degree_7   0.292510
degree_8   0.291699
degree_9   0.291471
degree_10  0.290227
degree_11  0.290375
degree_12  0.289692
degree_13  0.289823
```

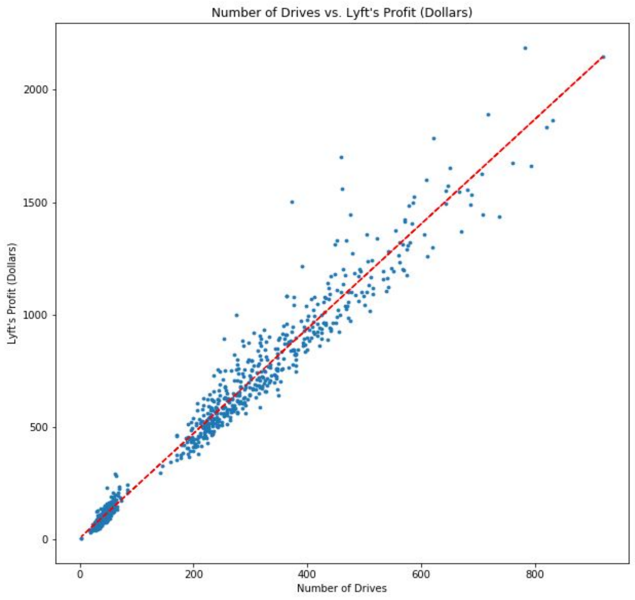
Using these methods, we were able to find the lines of best fit for each of these factors. After plugging in the expected driver lifetime of 55 days to each of the respective lines of best fit, we were able to find that the average driver brings in around $637.04 in profits, completes around 267 rides, and on average makes a potential rider wait around 235.45 seconds for a ride for their projected lifetime of 55 days.

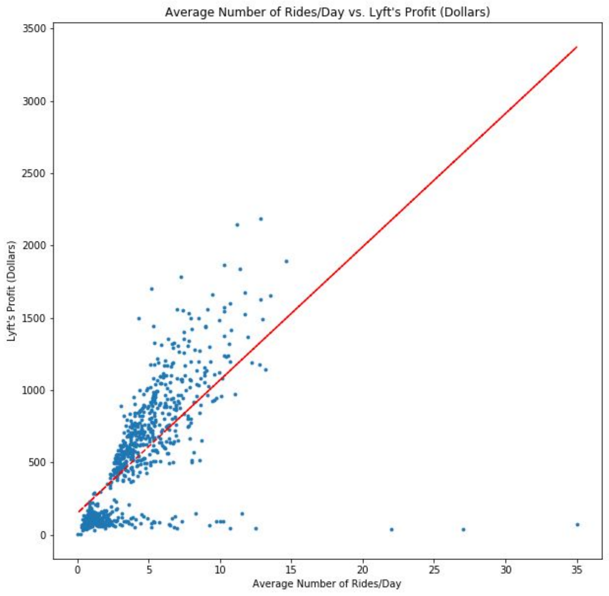
What are the main factors that affect a driver's lifetime value?

We defined our Driver's Lifetime Value based off of three factors: the pure capital they bring to Lyft, the amount of business that they bring (in number of rides), and customer satisfaction (inadvertently based off of the low wait time). We hypothesized that the primary factors that would impact these values would be the number of rides a driver completes in his/her lifetime, a driver's average ride duration, a driver's average ride distance, the average time it takes for the driver to accept a rider's request, the average time it takes for the driver to arrive at their rider's location after accepting the request, and the average time it takes for the driver to successfully pick up the rider.

Pure Capital

The factor we found most impactful on the sheer capital that a driver brings in was the number of rides that a driver completes. We plotted both drivers' lifetime number of rides and average number of rides per day versus their value to Lyft (*Number of Drives vs. Lyft's Profit (Dollars), Average Number of Rides/Day vs. Lyft's Profit (Dollars)*).





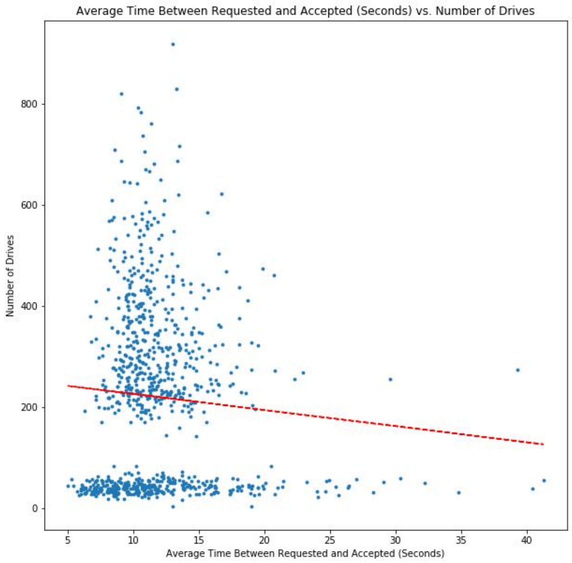
Both relations show clear positive linear correlations. This intuitively makes sense; the more rides a driver completes, the more money the driver makes and the more revenue is generated for Lyft. In addition to this, it seems there’s a slight negative correlation between the driver’s arrival time after accepting a request versus profit. There are many possible explanations for this. One possible one is that drivers who are quick to arrive after accepting a request tend to me more efficient and industrious, and therefore generate more profit.

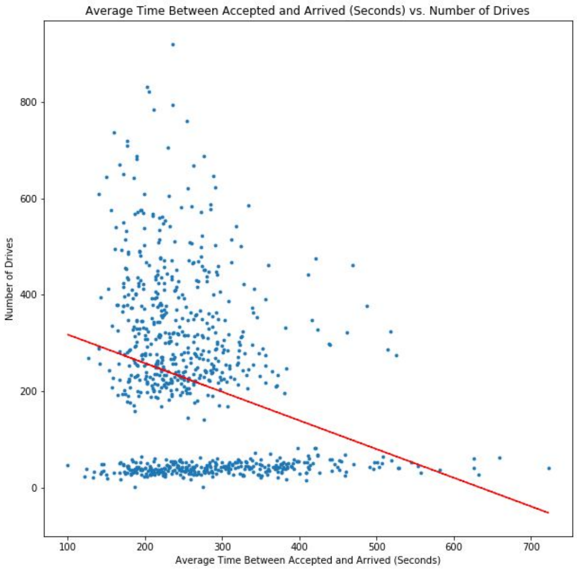
There is also a slight positive correlation between the time it takes a driver to pick up their rider after arriving versus profit. Again, there are many possible explanations for this. A possible reason is that drivers who wait for less time are perhaps happier with their jobs and are more likely to continue doing more rides and not quitting.

Business

Business, which is closely related to the company’s publicity, is another form of value that a driver can bring. The quantity of business that a driver presents can be naively represented by the number of rides that a driver performs in his/her lifetime; more rides mean more exposure for Lyft.

We compared the times that it took drivers to accept drives after they had been requested versus the lifetime number of drives those drivers performed as well as the times that it took drivers to arrive at their riders’ locations after accepting the requests versus the drivers’ lifetime drives. Both showed slightly negative correlations, seeming to imply that drivers who experience less downtime before their drives tend to be able to complete more drives over their lifetimes and bring more business to the company.





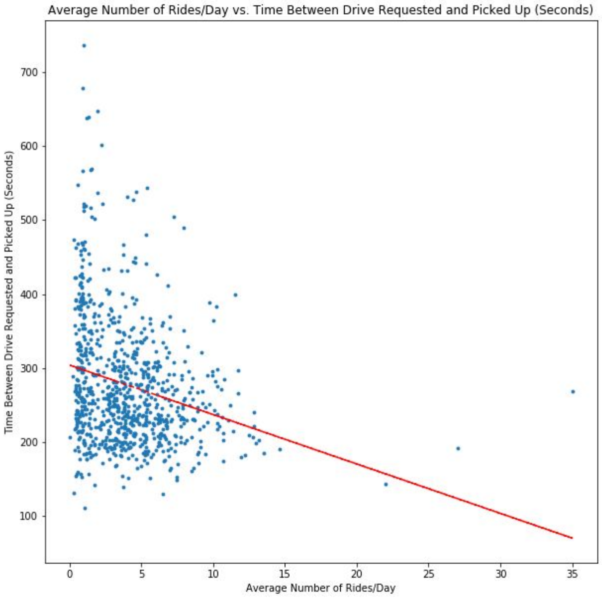
Customer Satisfaction

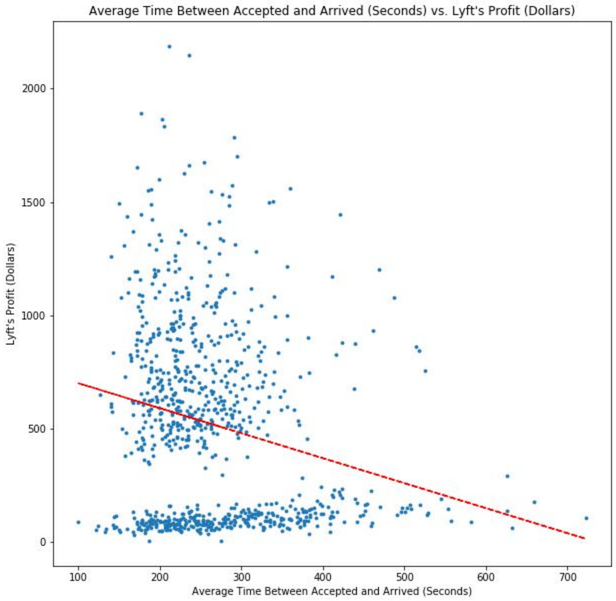
Customer satisfaction is the third and final factor that goes into the value that a driver can bring Lyft. If a driver gives a rider a good experience, the rider is more likely to do another ride with Lyft again in the future, bringing in more potential revenue.

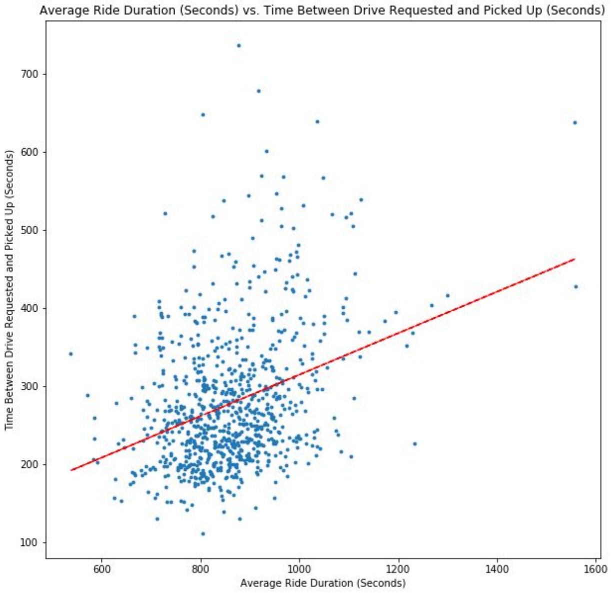
We naively represented customer satisfaction as the time difference between a rider requesting a ride and the driver picking the rider up. There's many other factors that can potentially impact customer satisfaction, such as the quality of the driver's car/driving, however we don't have access to data describing these factors. The aforementioned time difference tends to be a large factor that customers consider when using ride sharing services like Lyft; riders don't like having to wait significant amounts of time for their ride to arrive. It should be an accurate gauge of customer satisfaction, at least to some degree.

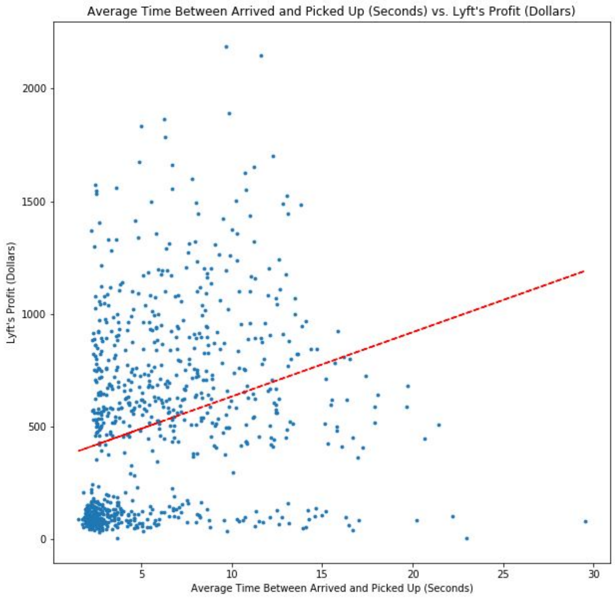
We found a negative correlation between a driver's average number of rides/day and customer satisfaction. There are numerous reasons why this might be the case. It's possible that as drivers do more rides throughout the day, they begin to treat successive rides with less thought and care than they might if they did fewer rides per day. We also found a positive correlation between a driver's average ride duration and customer satisfaction. Perhaps drivers who conduct longer rides tend to maintain better driver-rider relationships due to increased interaction time.

Both correlations show relatively small statistical significance; points are scattered in the plot like a cloud. Thus, only the correlations' positivity or negativity were considered and any deductions discussed should be taken with a grain of salt. Overall, however, the data suggests that drivers who give more and longer rides tend to garner greater customer satisfaction.


Average Number of Rides/Day vs. Time Between Drive Requested and Picked Up (Seconds)

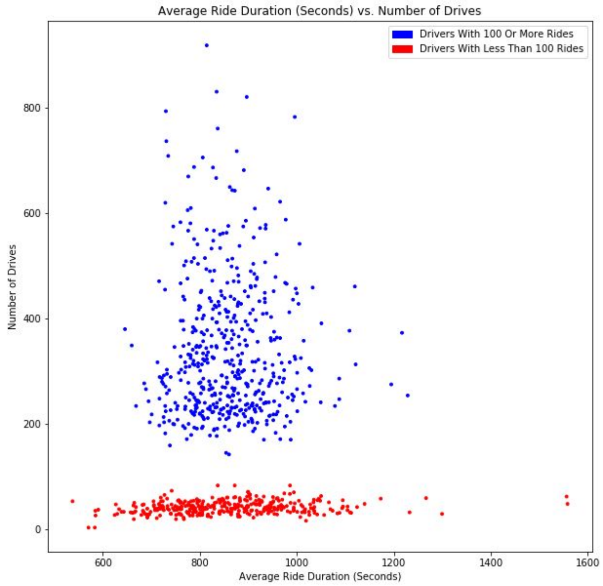

Average Time Between Accepted and Arrived (Seconds) vs. Lyft's Profit (Dollars)

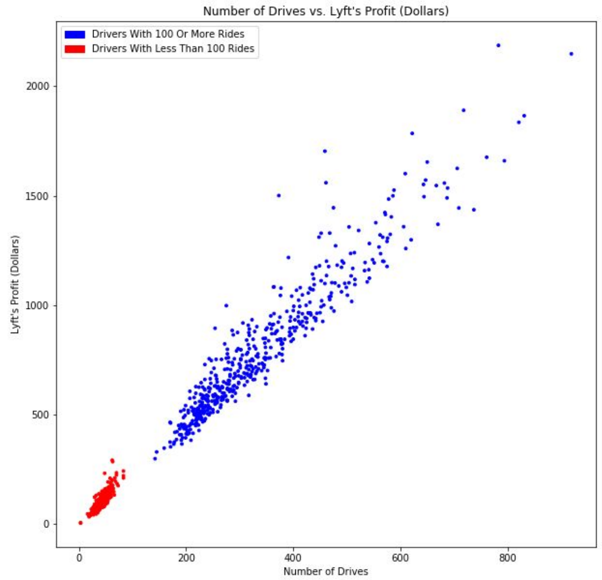

Average Ride Duration (Seconds) vs. Time Between Drive Requested and Picked Up (Seconds)


Average Time Between Arrived and Picked Up (Seconds) vs. Lyft's Profit (Dollars)

Do all drivers act alike?

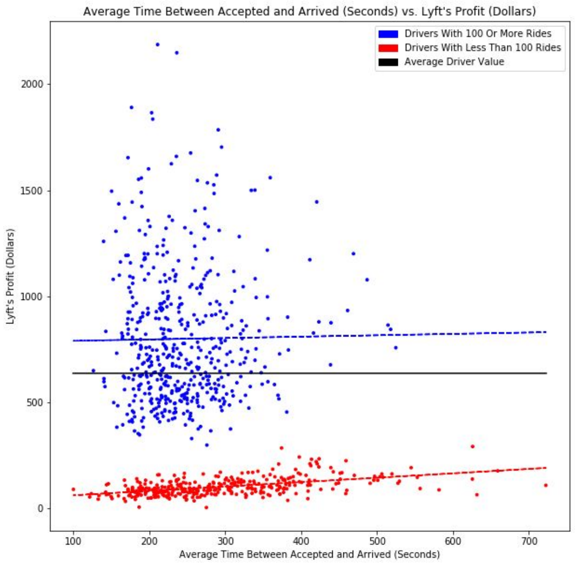
On many of the plots that we've obtained, there was a distinct separation of drivers into two groups. This difference could be observed most clearly when plotting the number of drives a driver undertook versus various other factors, such as average ride duration and profitability (*Average Ride Duration (Seconds) vs. Number of Drives*, *Number of Drives vs. Lyft's Profit (Dollars)*). Drivers that had either greater than or less than 100 lifetime rides tended to group together.

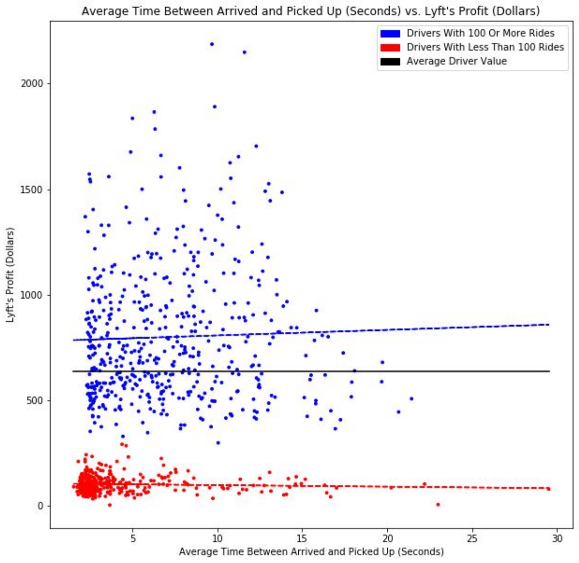


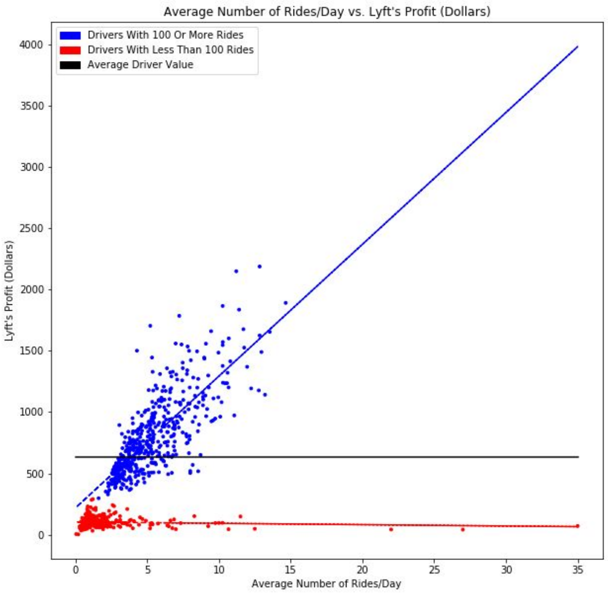


A likely explanation of this separation may involve the learning curve associated with becoming a more experienced driver. Less experienced beginner drivers tend to behave similarly to one another, as do more experienced drivers. In terms of performance and profitability, the two groups have very different tendencies. Across the board, beginner drivers underperform compared to the average profitability.

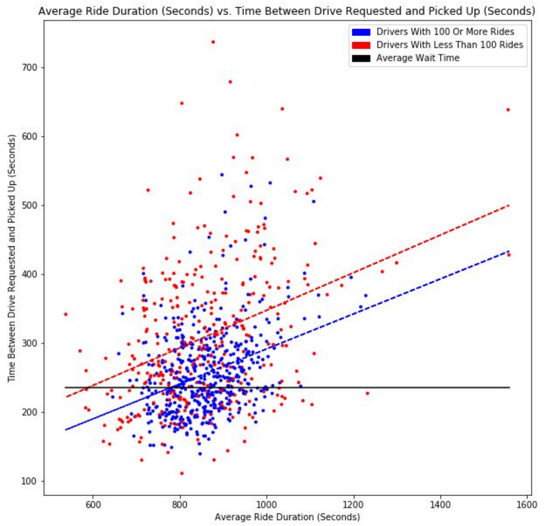
For sheer capital profitability, experienced drivers show a strong positive relationship between the number of rides that they perform per day versus their profitability. Less experienced drivers, however, show little correlation at all between the two factors; their below average profitability remains unaffected by the number of rides they perform per day (*Average Number of Rides/Day vs. Lyft's Profit (Dollars)*). Other factors, such as the time between a request accepts and driver arrival (*Average Time Between Accepted and Arrived (Seconds) vs. Lyft's Profit (Dollars)*) or time between driver arrival and rider pickup (*Average Time Between Arrived and Picked Up (Seconds) vs. Lyft's Profit (Dollars)*), don't show differences in the direction of the correlation however still illustrate the fact that beginner drivers consistently perform below average.

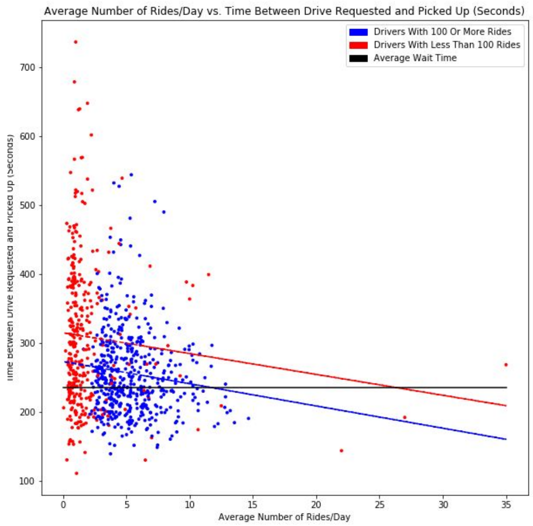






Quality of business is the final measure of profitability. Comparing the two groups in *Average Ride Duration (Seconds) vs. Time Between Drive Requested and Picked Up (Seconds)* and *Average Number of Rides/Day vs. Time Between Drive Requested and Picked Up (Seconds)*, it can be seen that the positivity of the correlations remain consistent between the two groups however, once again, beginner drivers consistently tend to take longer to pick up their riders than the more experienced drivers.





For all notable measures of profitability, experienced drivers generate more value for Lyft. This should make sense; drivers who take their roles seriously and commit to more rides tend to be the ones who make more money. Action should be taken to maximize the number of drivers that fall into this category.

What actionable recommendations are there for the business?

As we discussed in the previous question, there seems to be a large disparity between beginner drivers who have done less than 100 drives and more experienced drivers. There are a few potential explanations for this disparity. One reason could be due to the fact that beginners are not getting paid enough to have the desire to continue. Less experienced drivers are consistently making less profits for Lyft (and therefore making less money themselves) and consistently have a smaller lifetime (meaning that they're quitting early).

Our recommendation for the company is to focus on this issue and perhaps create incentives that promote being a long time driver for Lyft. By incentivizing drivers to drive more than 100 rides, we believe that Lyft will be able to maintain a long term relationship that will be beneficial both to Lyft and the driver. By either offering a bonus after hitting 100 drives or potentially adding a 'prime time rate' for beginners, this could potentially accelerate the activity beginner drivers have and generally increase the average value drivers bring to Lyft.

CONCLUSION

Overall, we were able to discover many valuable lessons from the data that was measured. However, there are a few points that are worth acknowledging.

The data primarily consisted of rides within a three month period which could greatly affect many things, the largest being the projected life time. We assumed that the ride dataset contained the full careers of drivers but this may not necessarily be true. In the future it would be useful to have information about a broader range of rides so that we can account items such as seasonal changes and general changes in driver behavior.

We also strongly valued business quality and customer satisfaction, and while we used rider waiting time as a measure of business quality, there are many other measurements that could be used as well. Rider ratings and rider tip amounts would be a great way to measure the quality of drivers and could help us quantify business quality accurately.

As the ride sharing industry grows, the impact Lyft has on the transportation industry and on drivers will grow with it. We believe that these points that we've discovered will not only help Lyft grow as a company but also ensure that the impact that Lyft has on its driver and the transportation industry is a positive one.