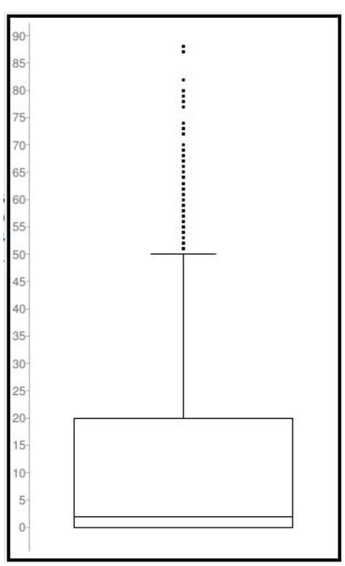
Kasha
Karina Ionkina and Daria (Dasha) Shifrina

Abstract

We ran through a number of different models and techniques to assess the contributing facets of Lyft drivers. We concluded that a Driver's Lifetime Value was around $798 based on average retention rate and sample mean monthly revenue. Our regression model successfully identified four independent factors that we hypothesized to most affect a driver's lifetime value: employment duration, average monthly ride distance, average monthly prime time rate per ride, and average monthly ride acceptance duration. In order to run the T test, we calculated the average projected lifetime of a driver to be 55 days or 1.8 months, as one of our models showed that once a driver had been inactive for over 20 days, they were most likely no longer working with Lyft. In our different tests, we found that drivers' behavior varies a lot, but tends to aggregate around a certain value and have a symmetrical distribution (our sample means and medians were always fairly close), and found a correlation between a driver's wait time and the amount of revenue they will bring. On average, drivers that worked more months since onboarding and had a higher monthly power rate showed a strong correlation for having a higher monthly revenue. Because of this, we recommend that Lyft creates more incentives that encourage drivers to work beyond a 55 day average by creating lotteries that do not cost Lyft more than a single driver's monthly revenue but intend to raise the retention rate.

Our Process

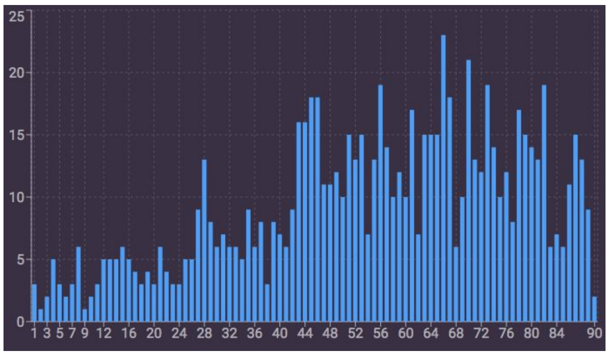
We used a relational database and Python for data comparison and manipulation. In order to assess the projected lifetime of a driver, we first wanted to distinguish employment from unemployment by comparing the drivers' most recent pickup to the most recent overall timestamp in the dataset. First, we extracted the most recent timestamp in the rider_ids dataset: 2016-06-27 00:50:50. Then we created driver_recent_ride(driver_id) to extract the most recent pickup timestamp of every driver. Afterwards, we found the difference between the two to calculate the number of consecutive idle work days per driver. Once we obtained the array, we plotted the data in a box-and-whisker plot. Click here for the full-sized image.

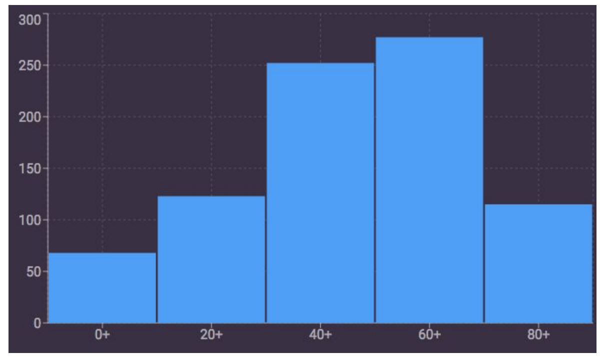


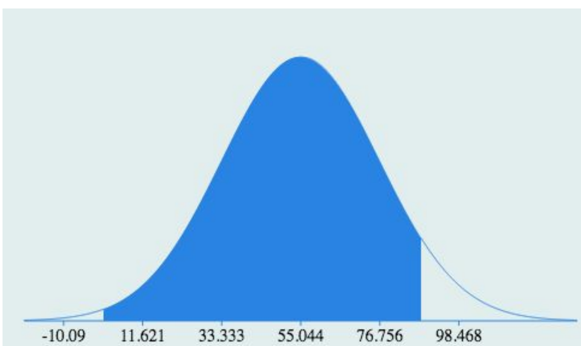
Sample size: 837
Median: 2
Minimum: 0
First quartile: 0
Third quartile: 20
Interquartile range: 20

We decided to assume that 20 days of inactivity indicate that a Lyft driver has quit the app. **Using this assumption, the data sample shows a ~24% attrition rate and ~76% retention rate.**

After this, we created another array to store the number of days that have elapsed between the drivers' onboarding dates and most recent pickups. Then we graphed the data to view the number of working days per employee. [Click here]() and [here]() and [here]() for the three diagrams.



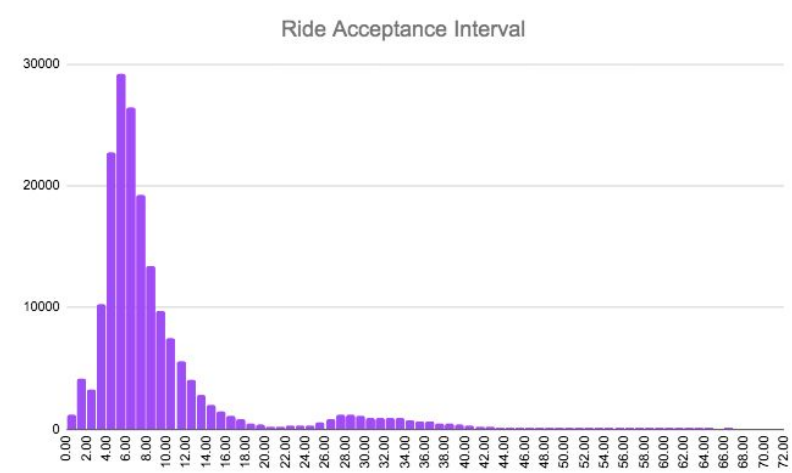




Avg: 55.0443
Standard dev: 21.7116
Median: 57
Middle 80% range: 23.4-82

These histograms and bell curve indicate that the average driver has an average lifetime of 55 days, or 1.8 months. **Therefore, an employee that stays with the company for 56+ days is above average.**

We created a histogram using a dataset of 184,209 different ride acceptance intervals to compare the drivers' response times. The x-axis shows the amount of time, in seconds, that it took each driver to accept a ride. The y-axis accounts for the frequency of trips with the given response time.



The data shows that for most trips, the driver takes up to 20 seconds to respond. Therefore **drivers that take less than 20s are above average.**

[Click here]() for a full-sized image.

Multivariable Linear Regression T-Test

We aimed to come up with a model to analyze which factors affect a driver's value. Before we could estimate a driver's lifetime value, we had to find a reasonable average monthly revenue.

Firstly, as previously stated, we identified a driver's amount of days worked by subtracting the onboarding date from the date of the most recent ride in the data set. Since we were not provided with the average revenue per ride, we used a Lyft-provided online equation to estimate their profits. We also used an online source that stated Lyft takes around 25% of a driver's revenue, alongside an additional service fee. The final equation we used was:

*0.25 * ride_prime_time * (2 + (0.22 * ride_duration) + (1.15 * ride_distance)) + 1.75*

Where 0.25 refers to Lyft's share of each ride, $1.75 is the added on service fee, and all other numerical values have been taken from the Rate Card provided in our challenge instructions.

We first used this equation to estimate Lyft's profit from each ride provided in the CSV files. Then we aggregated for each driver and each driver's monthly revenue, which we predicted by dividing total earnings by estimated active period. We were then able to extract factors like the arithmetic mean and standard deviation. Our calculations show a sample mean monthly revenue of $428.48. The median was around $400 with a first quartile of $134 and a third quartile of $621. Since the median and mean were close, we ended up approximating that a driver's standard average monthly value to Lyft had to come close to $428 to maintain profit.
Firstly, we used this value to approximate a Driver's Lifetime Value to Lyft. From this mean of average monthly revenue ($428), plus the mean employment duration (55 days), over the three months of data we were provided, we recommend a **driver's lifetime value to be around $798**. Secondly, we used this value to run a standard and multivariable regression T to analyze which independent variables could be correlated to increasing that value.

Then, we proceeded to run varied regression tests with different independent variables. In the end, our best model successfully identified four independent variables that we hypothesized to be associated with the revenue generated for Lyft monthly: amount of days worked, average distance of each ride monthly, average prime time rate of each ride monthly, and the monthly average time that a driver took to accept a ride after receiving a ride request. We concluded that these four were the main factors that affected a driver's lifetime value to Lyft. The model showed 14.5% variation in the data from the line of best fit regression. Out of the four variables, two were statistically significant, as well as the model as a whole.

Based on the following logic, we came up with these hypotheses for our variables:

- Amount of days worked (positive): the coefficient was expected to be positive because as an employee continues working with a company, they are normally expected to become more experienced in their tasks and better at identifying ways to maximize their own profit (understanding prime time, navigating passengers, etc.) and therefore maximizing Lyft's earnings.
- Average distance of each ride monthly (negative): expected to be negative because although distance increases profit per ride, it also increases the time it takes for a driver to complete a ride to get another customer and can especially slow down drivers in a busy city like San Francisco; the expected duration of the trip becomes less reliable as distance of trip increases.
- Average prime time rate of each ride monthly (positive): we rationalized that the higher the average monthly prime time rate was, the better the driver was at maximizing profit by catching more customers during times that created a multiplier on earnings.
- Average duration that a driver took to accept a ride since getting a request of each ride monthly (negative): a driver who has a higher acceptance duration is predicted to have less value as we anticipate that the reasons a driver takes longer to accept can be because they are motivated to accept riders that follow a route that works better for them rather than ones that maximize profit.

Below is a correlation matrix of the independent variables. None of them were greater than 0.7, so none of the variables are significantly related.

	Amount of days worked	**Average distance of each ride monthly**	**Average monthly percentage per**	**Average acceptance duration**
Amount of days worked	1	0.0112467	0.0444533	0.0128254
Average distance of each ride monthly	0.0112467	1	-0.224423	0.294417
Average prime time rate of each ride monthly	0.0444533	-0.224423	1	-0.110264
Average acceptance duration	0.0128254	0.294417	-0.110264	1

The Adjusted R-Square is 0.141237, meaning that around 14.124% of the variation from the expected linear regression is explained by the different variables in the model. In addition, since the significance of the F-value is less than 1e10, we can reject the null hypothesis at an alpha-level of both 0.05 and 0.01. Therefore, there is strong statistical evidence that our model as a whole is significant.

Below are the final test results. Good predictors were identified when they had a p value of less than 0.05 and highlighted in green:

	Coeff	**t-stat**	**lower t0.025(832)**	**upper t0.975(832)**	**p-value**
Intercept	110.163502	1.836667	-7.566606	227.89361	0.0666156
Amount of days worked	4.267382	8.887022	3.324873	5.209891	8.20E-14
Average distance of each ride monthly	-11.652444	-1.333293	-28.8067	5.501813	0.182801
Average prime time rate of each ride monthly	1040.288933	6.654704	733.453465	1347.124401	5.15E-11
Average acceptance duration	-2.935362	-1.131993	-8.025134	2.154409	0.257963

From this model and data, there is strong statistical evidence that a driver with a higher amount of total days working for Lyft who has a higher average prime time rate of each ride monthly will have a higher monthly revenue, which implies that Lyft gets a higher monthly revenue, and therefore has a higher lifetime value.

The regression of the line is as follows:

Average Monthly Revenue For Lyft = 110.163502 + 4.267382(**Amount of days worked**) - 11.652444(**Average distance of each ride monthly**) + 1040.288933(**Average prime time rate of each ride monthly**) - 2.935362 (**Average acceptance duration**)

As previously mentioned, we ran several regression tests before deciding this one was best. Other independent variables we tested and discarded were: quantity of monthly rides, average monthly ride duration, and average duration between accepting and picking up a passenger. We rejected these variables due to their generated T stat value, low p value, and their negative effect on the R Square Value. In the end, of the factors we tested, the four that seemed to have the greatest effect are outlined in the equation above and can be used to calculate an estimate for a driver's average monthly revenue. We can then compare the predicted value from these factors to the arithmetic mean of $428 and predict whether the value of their work will be below average.

If we are to follow regression test model, then it is easy to see that both correlation and regression equation indicate that the number of days an employee works are extremely important for the amount of income they can bring in monthly to Lyft. From this, we recommend that Lyft creates incentives for drivers to continue working until and beyond the average 55 days as predicted by another one of our earlier models. As we estimated, an average Lyft employee should bring an average of $428 of profit to Lyft a month. That number can also be used to create a ballpark for how much money Lyft is willing to invest in incentives to keep active drivers. For every $428 Lyft spends, it is essentially taking one driver's monthly earnings and redistributing them to multiply those earnings.

An example solution would be to create a tiered program with rewards based on employment duration. For instance, an existing program guarantees Lyft drivers $1000 if they are able to give 125 rides in their first 30 days. We propose some sort of reward for drivers who hit 100 days, 200 days, a year, etc. Possible rewards could be entering a monthly lottery where selected randomly chosen drivers win tickets to a resort. The price of the vacation would take into account the total amount of drivers that currently belong in that bracket as well as the average monthly revenue that drivers make. As drivers continue to work longer, they can be given the opportunity to earn higher costing prizes.