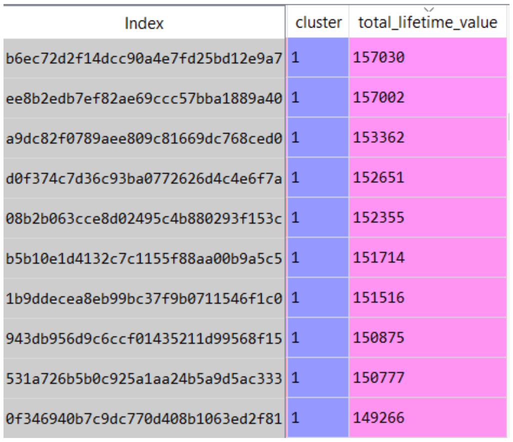
Lyft Data Challenge

The recommended Lifetime Values (in $) of the top 10 drivers (along with the driver id's) is shown in the table.



Index	cluster	total_lifetime_value
b6ec72d2f14dcc90a4e7fd25bd12e9a7	1	157030
ee8b2edb7ef82ae69ccc57bba1889a40	1	157002
a9dc82f0789aee809c81669dc768ced0	1	153362
d0f374c7d36c93ba0772626d4c4e6f7a	1	152651
08b2b063cce8d02495c4b880293f153c	1	152355
b5b10e1d4132c7c1155f88aa00b9a5c5	1	151714
1b9ddecea8eb99bc37f9b0711546f1c0	1	151516
943db956d9c6ccf01435211d99568f15	1	150875
531a726b5b0c925a1aa24b5a9d5ac333	1	150777
0f346940b7c9dc770d408b1063ed2f81	1	149266

Index represents driver_id

The steps we followed to arrive at our conclusion are as follows,
First, we calculated the cost per ride using the assumptions for the Lyft rate card given. There are 4 components which decide the value of the cost per ride: base fare, cost per mile, cost per minute & service fee. The Prime-time fee updates the cost per ride, and considers the base fare, cost per mile and cost per minute.
The formula thus obtained for calculating the cost per ride was
- Cost per ride = (1 + ride_prime_time/100) * {base_fare + [cost_per_mile*(ride_distance*0.000621)] + [cost_per_minute*(ride_duration/60)]} + service_fee
- Cost per ride = Min (400$, Cost per ride)
- Cost per ride = Max (5$, Cost per ride)

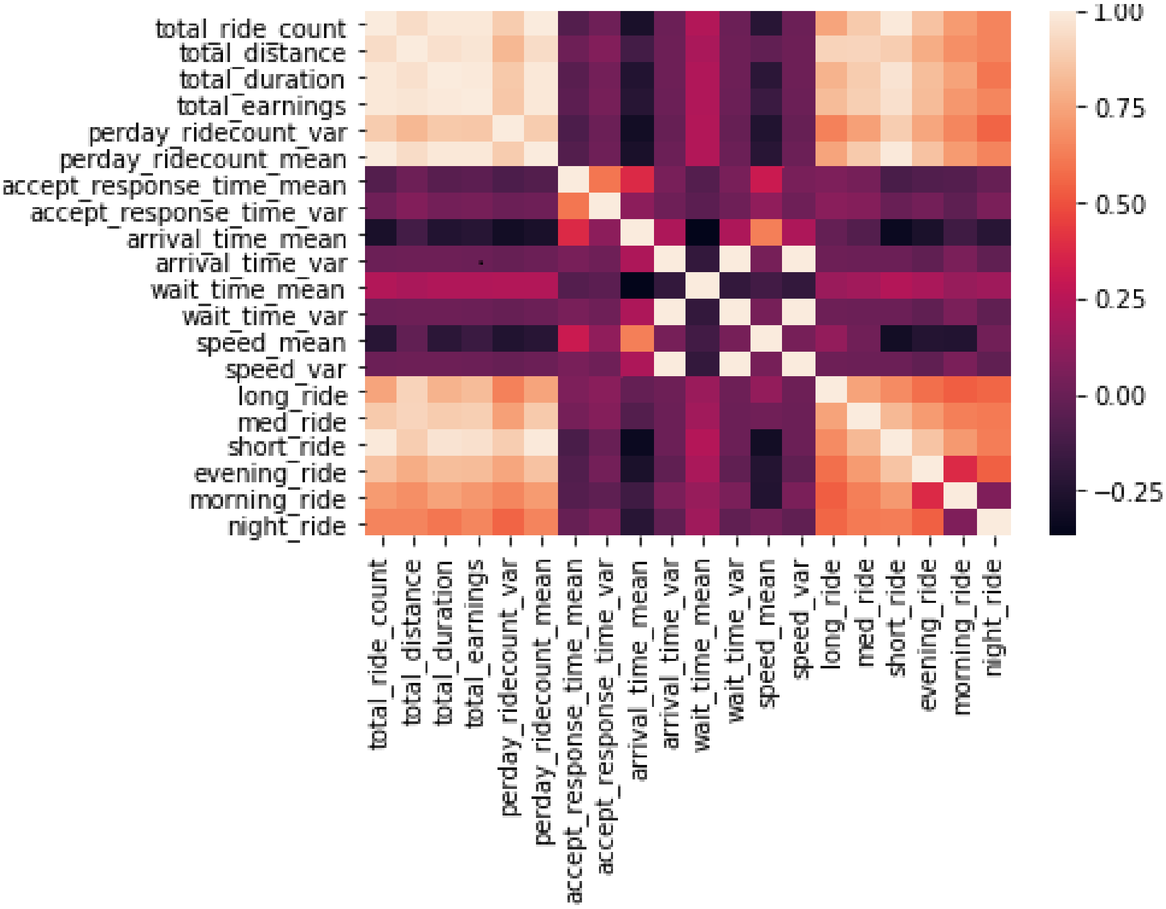
We next did feature extraction to extract the key and deciding factors that affect a driver's lifetime value. For each of the features, the mean and variance of the features were considered. It is so because for any feature, there will be instances where variance may give false results (consistently performing bad, for example), and there will be instances where mean may give false results (consider outliers for example). These features, or the factors that affect a driver's lifetime value are summarized below.

- **RIDE ACCEPT RESPONSE TIME:** It is the time a driver takes to accept a ride once it is requested. If the accept response time is low or negligible, that means the driver is loyal to his/her job. If the driver takes more time consistently in accepting a ride once it is requested, that concludes that the driver is not interested in his/her job and is an early sign of resignation.
- **RIDER ARRIVAL TIME:** It is the time a driver takes to reach the pickup location once a ride is accepted. Even though it is not a strong measure in determining the driver's lifetime value, but may result in the driver being tempted to leave the firm if that driver consistently gets rides with high arrival time, i.e., the pickup location is far away from the location where the driver accepted the ride. Or we can assume that the driver is willingly arriving late to pick up the rider. Both situations are negatively impacting the driver's lifetime value towards the firm.
- **DRIVER WAIT TIME:** It is the time a driver waits for the rider once the driver reaches the pickup location. If a driver consistently gets rides with higher wait times, that may be a factor for him/her to leave the firm.
- **RIDE AVERAGE SPEED:** It defines the average speed with which the driver has completed all the rides, depending on the total distance covered and total time taken to cover that distance. If the driver has an

average speed over a certain threshold speed, it is an indication that either the driver will get highly penalized and might lose his/her job or the driver may (unfortunately) die, thus, highly reducing his/her projected lifetime.

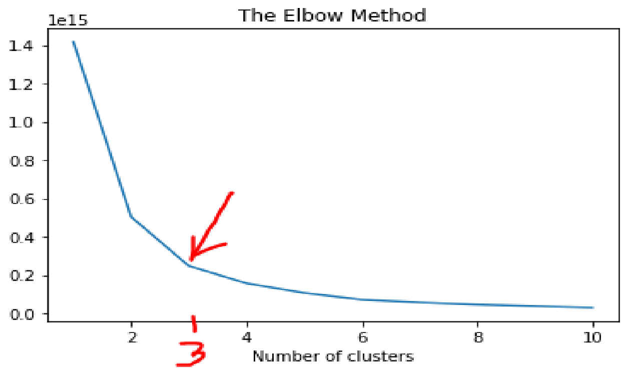
- **RIDE LENGTH:** The rides are divided into three broad categories: short rides (less than 8 km), medium rides (8-20 km), and long rides (more than 20 km). This indicates which type of rides, based on length a driver generally gets.
- **RIDE TIME:** The rides are divided into three broad categories: morning (6 am – 3 pm), evening (3 pm – 9 pm) and night (9 pm – 6 am). It determines at what time of the day a driver generally does his job.

After getting all the features, we tried to find the correlation between the extracted features, by plotting the correlation heat map. The heat map is shown below.



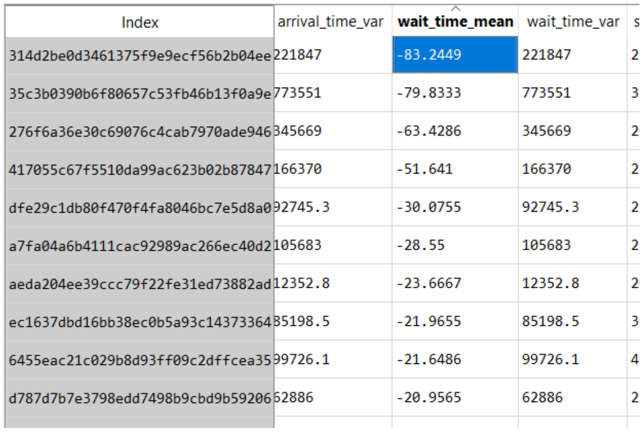
It was observed that all the extracted features were important as there was a low correlation among all the features (as shown by the darker squares).

Next, we tried to cluster the drivers based on the extracted features. For doing so, the first step was to calculate the number of clusters. We used Elbow method to get the number of clusters.

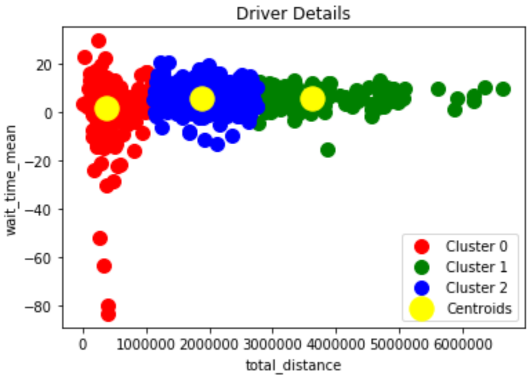


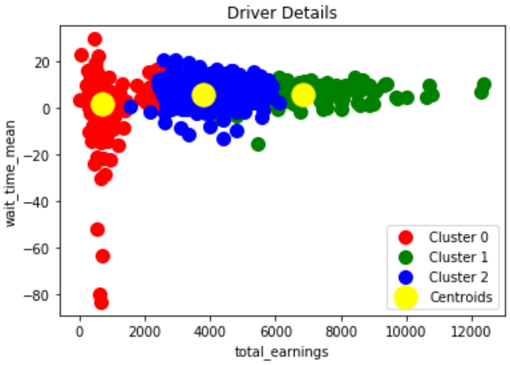
From the Elbow method, it was concluded that we require 3 clusters to classify our drivers. We would classify those 3 clusters as good, mediocre and bad, indicating good drivers (the ones who will stay with the firm for a longer time and are contributing heavily to the profit of the firm), the mediocre drivers (not good as well as not bad) and the bad drivers (the ones who are bound to leave the firm sooner and aren't comparatively contributing that much to the firm).

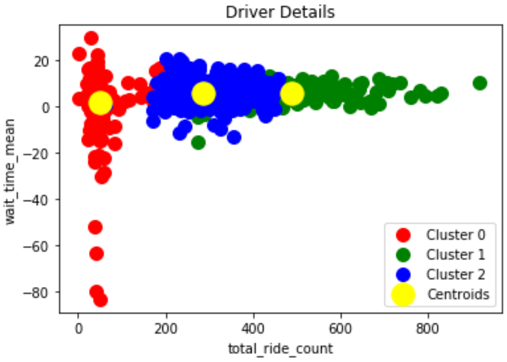
Next, we applied K-Means clustering, to classify the drivers based on the features and plotted various clustering scatter between two of the features, to arrive at significant conclusions. Some key observations are listed below.



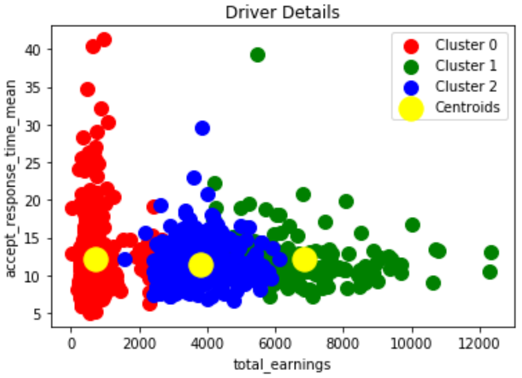
Index	arrival_time_var	wait_time_mean	wait_time_var	s
314d2be0d3461375f9e9ecf56b2b04ee	221847	-83.2449	221847	2
35c3b0390b6f80657c53fb46b13f0a9e	773551	-79.8333	773551	3
276f6a36e30c69076c4cab7970ade946	345669	-63.4286	345669	2
417055c67f5510da99ac623b02b87847	166370	-51.641	166370	2
dfe29c1db80f470f4fa8046bc7e5d8a0	92745.3	-30.0755	92745.3	2
a7fa04a6b4111cac92989ac266ec40d2	105683	-28.55	105683	2
aeda204ee39ccc79f22fe31ed73882ad	12352.8	-23.6667	12352.8	2
ec1637dbd16bb38ec0b5a93c14373364	85198.5	-21.9655	85198.5	3
6455eac21c029b8d93ff09c2dffcea35	99726.1	-21.6486	99726.1	4
d787d7b7e3798edd7498b9cbd9b59206	62886	-20.9565	62886	2

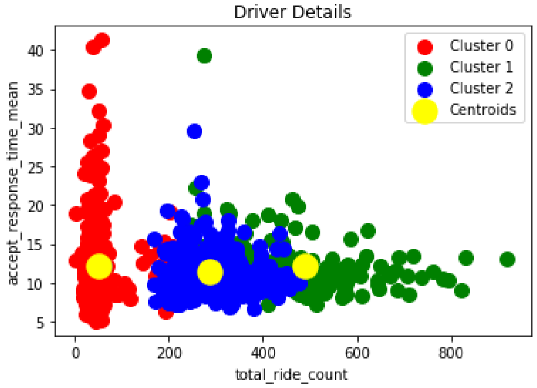




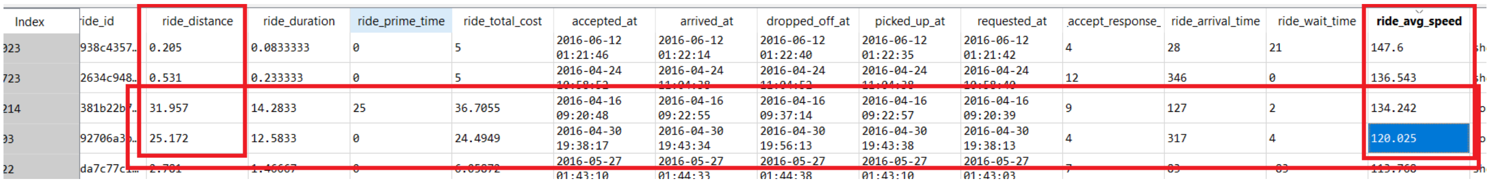


The above observations compare the mean of the wait time of a driver (the time a driver waits for the rider after arriving at the pickup location) with different features, such as total distance covered by a driver, total earnings by a driver and total rides done by a driver. It was observed that the drivers who have a wait time closer to 0 (the good drivers) are clustered in cluster 1 (shown in green), while the ones with abnormal wait times (the bad drivers) are clustered in cluster 0 (shown in red). We can also observe from the table that we have negative values of the wait time for the cases where the driver started a ride even before arriving at the pickup location. This is not desirable and should be totally avoided and such drivers must be penalized. Our clustering correctly identifies those bad drivers.



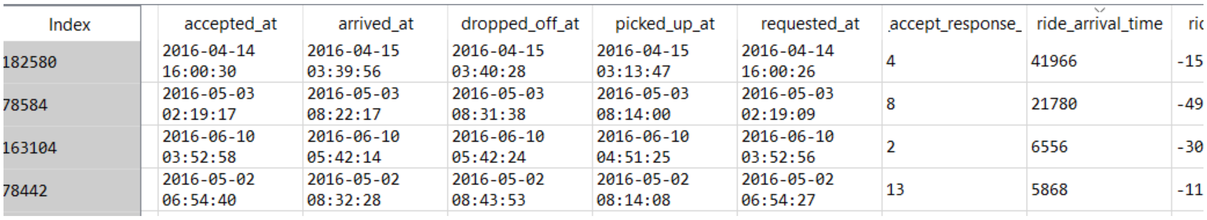


Another observation was the accept response time of a driver (the time a driver takes to accept a ride once requested by a rider), based on total earnings and total rides done by a driver. It was observed that bad drivers generally take more time to accept a ride while good drivers generally accept a ride instantly.



Index	ride_id	ride_distance	ride_duration	ride_prime_time	ride_total_cost	accepted_at	arrived_at	dropped_off_at	picked_up_at	requested_at	accept_response_	ride_arrival_time	ride_wait_time	**ride_avg_speed**	
023	938c4357…	0.205	0.0833333	0	5	2016-06-12 01:21:46	2016-06-12 01:22:14	2016-06-12 01:22:40	2016-06-12 01:22:35	2016-06-12 01:21:42	4	28	21	147.6	sh
723	2634c948…	0.531	0.233333	0	5	2016-04-24 10:58:52	2016-04-24 11:04:28	2016-04-24 11:04:52	2016-04-24 11:04:28	2016-04-24 10:58:40	12	346	0	136.543	sh
214	381b22b7…	31.957	14.2833	25	36.7055	2016-04-16 09:20:48	2016-04-16 09:22:55	2016-04-16 09:37:14	2016-04-16 09:22:57	2016-04-16 09:20:39	9	127	2	134.242	b
03	92706a3b…	25.172	12.5833	0	24.4949	2016-04-30 19:38:17	2016-04-30 19:43:34	2016-04-30 19:56:13	2016-04-30 19:43:38	2016-04-30 19:38:13	4	317	4	120.025	b
22	da7c77c1…	2.781	1.46667	0	6.05872	2016-05-27 01:43:10	2016-05-27 01:44:33	2016-05-27 01:44:38	2016-05-27 01:43:10	2016-05-27 01:43:03	7	83	-83	113.768	sh

It was also observed that there are some bad drivers who have done rides, far exceeding the threshold limit, as shown in the above figure. Consider for example, the drivers who drove 31.957 km at an average speed of 134.242 km/h. These speeds are not desirable and such drivers should be penalized.



Index	accepted_at	arrived_at	dropped_off_at	picked_up_at	requested_at	accept_response_	ride_arrival_time	ric
182580	2016-04-14 16:00:30	2016-04-15 03:39:56	2016-04-15 03:40:28	2016-04-15 03:13:47	2016-04-14 16:00:26	4	41966	-15
78584	2016-05-03 02:19:17	2016-05-03 08:22:17	2016-05-03 08:31:38	2016-05-03 08:14:00	2016-05-03 02:19:09	8	21780	-49
163104	2016-06-10 03:52:58	2016-06-10 05:42:14	2016-06-10 05:42:24	2016-06-10 04:51:25	2016-06-10 03:52:56	2	6556	-30
78442	2016-05-02 06:54:40	2016-05-02 08:32:28	2016-05-02 08:43:53	2016-05-02 08:14:08	2016-05-02 06:54:27	13	5868	-11

Another observation was the outlier data in the dataset. All the above-mentioned rides have a high ride arrival time, i.e., the driver took a long time to arrive at the pickup location. There could be a couple of reasons, either the driver selection algorithm is not that great (which is highly unlikely), or the driver deliberately took long time to pick the rider up (and thus decreasing the value of the firm in the eyes of the customers). Such drivers are also bad drivers.

We also observed that there were **some drivers without any rides but are associated with the firm** with an on-board date **(93 drivers in total)**. These drivers are not considered for the model and are classified as bad drivers.

In order to arrive at a recommended Driver's Lifetime Value (in numerals, given the limited dataset), we have assumed that **the bad drivers will likely stay with the firm for 1 year**, **the mediocre drivers will stay for 3 years** while **the good ones will stay for 5 years**. Based on these assumptions, we used the driver's total earnings to compute the 'Driver's Lifetime Value' for the good classified clusters.

Actionable recommendations for business are:

- Enhance the driver allotment algorithm such that a driver must travel the least distance to pickup a rider from the location where he/she accepted the ride. This will reduce the extra distance (arrival time) covered by the driver without an associated fare as well as save the time in covering more rides.
- Aim to reduce the driver 'wait time'. This can be accomplished by charging the riders a fee if they take more time after the driver has arrived at the pickup location. This will not only help the firm with maintaining the trust of the driver with the firm but will also help the firm cover more rides in the same time, and hence, increasing the efficiency.
- Increase the number of drivers during Prime time. This will help getting driver earn extra money and will also help the firm cover the fulfilments of all the riders in an area where there is a high demand for rides.
- Implement strict penalizations for drivers exceeding the speed limit. Give one warning and if the driver defaults again, take appropriate actions against the driver. It would not only set a good image of the firm in the eyes of its customers and inculcate a good traffic sense among all but would also make the roads safer for all.

Author

Ashwani Kumar Kashyap, axk190033@utdallas.edu
Anshul Pardhi, anshul.pardhi@utdallas.edu
Team Name: cizos