Part 1

In this data there is an overall upward trend over time. There are clear weekly patterns. Weekends show much higher volume than weekdays, with Monday showing the lowest volume and each day increasing toward Saturday. Sunday shows slightly lower traffic than Saturday. During the weekdays, the peak usage is around 11:15 or 11:30. The lowest usage is around 6:30 am. Usage slows considerably after midday and then picks up again in the evening. For each day of the week from Monday through Friday, the evening peak gets higher and higher until Friday when the evening peak is higher than the Friday lunch peak. From then on during the weekend, there are consistently high usage rates in the evening/night/early morning with the peak times around 4:45 am. After the peak the usage drops off precipitously. There tends to be low usage until about 11 am and then it steadily picks up again through the evening hours.

When grouping by day of week, I noticed a few outliers with Thursday having a low outlier. That is explained by incomplete data for the first day in the series. It was also New Year's Day, a holiday, which may have had an impact on usage. Looking at usage after 10 pm for that day, it is still lower than the other Thursdays after 10 pm. That would make sense since the night before was New Year's Eve, when more people tend to go out, and the next days was a weekday, meaning people would likely stay in.

There were high outliers in March for Tuesday, Wednesday and Thursday. The Tuesday was St. Patrick's Day, which may have caused a spike in usage that carried over for a few days. It would be interesting to see if the other high days coincided with any other significant events such as a sporting event or a religious holiday.

Part 2

1. The best measure depends on what data is available for use. If there is a way of tagging a given trip as being in one city or the other, the most direct way to measure this would be to take the ratio of trips in Gotham to Metropolis. If there was an even split between the two cities, this ratio would be 1. The metric therefore would be the absolute value of the difference from 1. If not, and it is possible to count the number of trips over the bridge per driver, this would be a good indicator because more trips across would mean they were serving both cities. If that information is not available, and only time of login is available, like in the logins.json file, then taking a ratio between night and day drives may work since one city needs drivers at night and the other during the day. In this case, the absolute value of the difference from 1 would be the measure. This is based on the hunch that a driver would work nights in Gotham and midday in Metropolis since those are presumably the highest traffic times in the respective cities. If I took this approach, I would probably
2. To test the efficacy of this program, the company should conduct an AB Test, comparing the results of a control and an experimental group.

a. The company should select the drivers at random to participate in either the control group or the experimental group. The experimental group would be notified that they will receive reimbursement for all toll trips, and the control group will not receive any notification. The experiment will be conducted over a period of 4 weeks. The null hypothesis would be that there is no difference between the groups. The alternative hypothesis is that there is a difference.

b. I recommend a 2 tailed t-test to compare the difference between two means for the selected measure for the two different groups. I would look for a p-score less than 0.05.

c. If the p-score was less than alpha, I would reject the null hypothesis. I would make sure to plot the metrics for each driver to understand the data and identify any outliers. I would let the city operations team know if the null hypothesis was rejected. But if not, I would still provide the summary statistics for each group along with the p-score to show if there were any tendencies in the different groups.

Part 3

1. To explore the data, I converted the signup date and the last ride date to date time objects, then extracted the month of last ride. I calculated the delta of signup to last ride and used that as the y variable in a series of scatterplots and correlations with each numeric field as the x variable. taking the number of riders with last ride dates on or after June 1, 2014. I created a heat map to chart the correlations and created histograms to compare categorical variables, binning the frequencies based on last ride date. I calculated the percentage retained by taking the number of riders with last ride dates on or after June 1, 2014 and dividing it by all riders: 38%.
2. I decided to use a logistic regression model from Sci Kit Learn library since this problem is looking for an output of either Active or Not Active. Based on the correlations and histograms, I noticed a relationship between recent activity as the y variable and phone type, city location, number of trips in first 30 days, and ultimate black user as x variables. I observed that ultimate black user and number of trips in the first 30 days appeared to be confounding, so I included only the ultimate black user feature and not the number of trips in first 30 days. I split the data with 75% training and 25% test. I considered increasing the number of features but did not find significant improvement in scores and I did not want to risk overfitting the training set. My model showed an overall accuracy of 71% with a recall of 89% for predicting Not Active, and precision of 72% for predicting Not Active. The 89% recall is the most important score because I am seeking to minimize false negatives for the Not Actives – in other words I want to avoid the cost of predicting someone as Active when they are truly Not Active because it will distort revenue projections, and it will mean that I will miss the opportunity to entice them back into activity.
3. To approach this problem, I would try to entice Not Active people back into Activity by offering discounts on rides, and perhaps a special trial period as an ultimate black user. I would consider running a pilot discount initiative in Astapor to see if I could make improvements there, since they showed the lowest rate of long-term activity. I would consider the same kind of targeted outreach to Android users for the same reason.