Analysis of Ford GoBike System

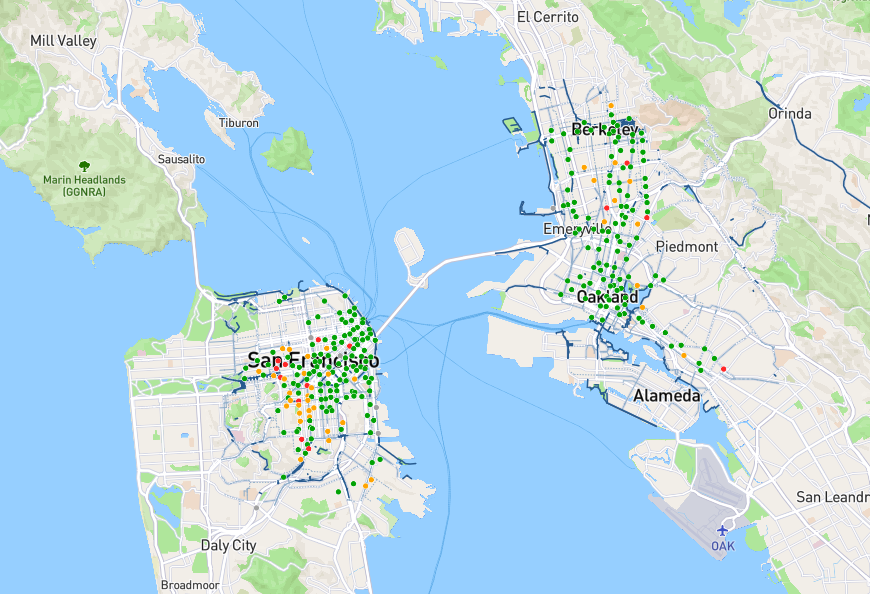
By : Chad Madding, Alec Lepe and Shantanu Godbole (MSDS 6372 405)

horizontal line

**Introduction:**

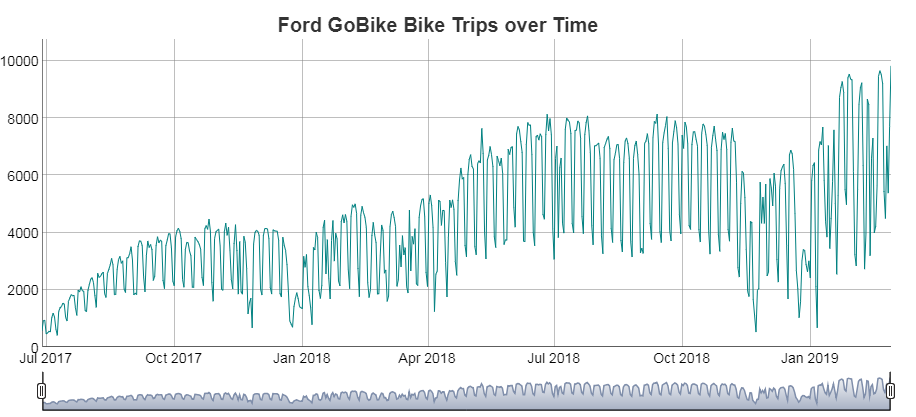
Ford GoBike (FGB) is the California Bay Area bike share program with over 300 stations and 4,000 bicycles. It spans 3 counties and 5 cities including: San Francisco, Emeryville, Oakland, San Jose, and Berkeley. We summarized, visualized, and analyzed their publicly available anonymized trip level data. Bike data begins in June 2017 and continues in monthly batches. To date, there are approximately 2.7 million trips taken, 2M taken solely within San Francisco. In addition to FGB data, we incorporated local weather statistics from the region’s most used airports, Oakland International Airport (OAK) and San Francisco International Airports (SFO), to potentially predict ride count. We focused on three languages to clean, visualize, and analyze various datasets.

**Ford GoBike Operations Map:**

****

**Ford GoBike Operations:**

While FGB operates in 5 cities, FGB trip data was split up into three distinct regions: East Bay, San Francisco, and San Jose. The East Bay region describes the three cities FGB serves: Emeryville, Oakland, and Berkeley while the other two region describes the cities of the same name. This is an important distinction because there are essentially no trips between these regions, which can also be described as intra-regional trips. Intra-regional trips are virtually non-existent because of the relative distance and relative infeasibility of an intra regional bike trip. Intra-regional trips, as a result, were ignored.

****

**Figure 1**

**Weather Data:**

We were interested to see how and if weather is impacting the Bike Rental. Frontier Weather [[1]](#footnote-1) provides hourly weather from various stations. The weather parameters they provide in which were most important for the analysis were: Temperature, Dewpoint, RH, Wind Speed and Precipitation. The weather data was carefully merged with the bike data to the hour for analyzing the impact of weather.

**Initial Analysis:**

The initial data was obtained [[2]](#footnote-2) from the main website of Ford-Go Bike. The main analysis is focused on the 2018 rental data. There were 1,638,328 transactions in the dataset across 2018 that contained following information:

|  |  |
| --- | --- |
| * Duration\_sec * Start\_time * End\_time * Start\_station\_id * Start\_station\_name * Start\_station\_latitude * Start\_station\_longitude * End\_station\_id * End\_station\_name * End\_station\_latitude * End\_station\_longitude * Bike\_id * User\_type (Subscriber/Customer) * Member\_birth\_year * Member\_gender * Bike\_share\_for\_all\_trip | **Some Observations**  Average Rental Duration ~ 15 mins  Average Subscriber Age ~ 35  Station Most used for renting a bike - ‘15’ [[3]](#footnote-3)  Station Most used for returning a bike - ‘67’[[4]](#footnote-4)  Subscribers rented bike 1,583,554 times  Males rented bikes 1,288,085 times  Females rented bikes 438,188 times |

**Analyzing Trends and Clustering:**

|  |  |
| --- | --- |
| Figure 2 | Figure 3 |

As we can see in Figure 2, there is a heavy bike usage at the initial station IDs as compared to the later stations. Ford Go Bike started in San Francisco, therefore the station numbering starts from the stations in San Francisco. The usage is less on Oakland, which are the later station IDs.

Several combinations were done for Clustering to see if the clustering can give us some more insight about trends in data, and the few patterns that stood out were:

1. Clustering on the basis of the ride Duration, Temperature and Precipitation. We ran k-means clustering to check how the clusters were formed and following output was observed.



Duration (in mins) - Temperature - Precipitation

Chart 1

There is not much variance observed in the temperature, but we can see that when the is no precipitation, the duration of bike rental is highest and then the duration significantly drops with the increase in amount of precipitation.

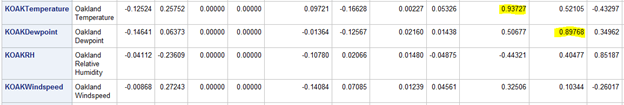
1. Age vs Duration in Mins: This was another observation that was interesting. As seen in figures [[5]](#footnote-5) , we can see that the most duration bike was rented was at a lower age for the customers.
2. Start Station ID vs Duration Clustering suggests that in the area around station 138, bike rental duration is very high. [[6]](#footnote-6) (Jersey St at Church St area in San Francisco)

**Principal Components Analysis Introduction:**

We will be using Principal Components Analysis to reduce the number of variables, while retaining as much of the information as possible. We would also like to run a PCA to identify patterns in the weather variables in regards to our bike share data

**Correlations:**





Before adjusting any of the data we wanted to look at all the original 24 variables to see if there were any correlation problems. The chart above shows an example of what we found. Four of the weather variables showed to have high correlation; .8 and above. Basically, meaning the cities are so close the weather data were measuring the same thing. To handle this, we took the average of Oakland and San Francisco’s temperature, dewpoint, relative humidity and perception, and ended up with four new variables for the weather data. (Avg Temp, Avg Dewpoint, AvgRH and Avg Precip).

Correlations between San Francisco and Oakland weather:

* Temperature = .94
* Dew Point = .90
* Relative Humidity = .85
* Precipitation = .81

There were also two date variables that had high correlation as well. For the date variable we dropped the Start Year because the same information was in the Start Date variable.

StartYear and StartDate = .83 - We kept Start Date

Since PCA is designed for continuous variables. There were six categorical variables that had almost no correlation at all, below .1. Because this could cause one or more of them to make its own principal component and with such low correlation this would not be useful they were removed from the PCA.

* StartSideN
* End City
* StartDayTypeN
* StartMonthLabelN
* KFSOWeatherN
* KOAKWeatherN

After adjusting for the major correlation problems, we ended up with 15 variables out of the original 24 to use in the PCA.

**Factors To Keep:**

We first decided to use the standard MINEIGEN option in SAS which follows Kaiser’s proposed dropping factors whose eigenvalues are less than one. Starting with the original 24 variables and running the PCA resulted in 10 factors that accounted for only 74% of the variance in the data.



After averaging the weather data and removing the data variables that did not fit the correlation we ended up with 15 variables that resulted in 6 factors that only accounted for 67% of the variance.



A cutoff of one was not capturing enough of the variance. [[7]](#footnote-7)

We decided to account for a range around 80% of the variance so we ended up selecting the first 8 factors.[[8]](#footnote-8)

Plotting the eigenvalues on a Scree Plot shows the ‘elbow’ falling in the 6 and 7 range and ending with 8.

In addressing how many factors should be kept, the scree plot shows that the first eight factors are indeed the largest. The cumulative percentages show that the first eight factors account for 79% of the variation. We felt that eight factors would adequately approximate these data.

**Interpretation of Factors: [[9]](#footnote-9)**

We can now step through a basic interpretation of each of the eight factors. Factor one appears to be a grouping of wind and temperature variables. Factor one contains significant weights in eight variables but the two on wind speed are the largest. Factor two appears to focus mostly on sky cover with relative humidity and the dew point also having some weight as well. Factor three shows a contrast of Starting Date and Average Dewpoint. Factor four is pointing to a contrast between the average temperature and humidity and wind. Factor five appears to shows an association with the barometric pressure variables. Factor six is waited with the count and shows a contrast with the rain data. Factor seven is a combination of Oakland’s barometric pressure and rainfall data. Factor eight shows a contrast of the barometric pressure variables of both Oakland and San Francisco.

**Varimax Rotation: [[10]](#footnote-10)**

To help clean up some of the ‘noise’ and improve our ability to interpret the meaning of each component we performed a rotation of the data.

Comparing the rotated factor patterns to the original factor patterns you can see the information has cleaned up a bit. This helps us to see the factors affecting different parts of the data.

**Multi-Variable Linear Modeling with R:**

Using a random variable selection method, R was able to generate linear models to predict total ride count, by region, per hour. Out of over 250,000 formulas, below are the best models, by adjusted R^2, per region.

|  |  |  |
| --- | --- | --- |
| Region | Formula | Adjusted R^2 |
| San Francisco | KFSOCldFrac+KFSODewpoint+KFSOMSLP+KFSORH+KFSOTemperature+KFSOWeather+KFSOWindDir+KFSOWindspeed+KOAKCldFrac+KOAKDewpoint+KOAKMSLP+KOAKPrecip+KOAKRH+KOAKTemperature+KOAKWeather+KOAKWindDir+KOAKWindspeed+StartDayType+StartHr+StartMonthLabel+StartYear | 0.29 |
| East Bay | EndCity+KFSOCldFrac+KFSODewpoint+KFSOMSLP+KFSORH+KFSOTemperature+KFSOWeather+KFSOWindDir+KFSOWindspeed+KOAKDewpoint+KOAKMSLP+KOAKPrecip+KOAKRH+KOAKTemperature+KOAKWeather+KOAKWindDir+KOAKWindspeed+StartCity+StartDayType+StartHr+StartMonthLabel+StartYear | 0.10 |
| San Jose | KFSOCldFrac+KFSODewpoint+KFSOMSLP+KFSOPrecip+KFSORH+KFSOTemperature+KFSOWeather+KFSOWindDir+KFSOWindspeed+KOAKDewpoint+KOAKMSLP+KOAKRH+KOAKTemperature+KOAKWeather+KOAKWindDir+KOAKWindspeed+StartDayType+StartHr+StartMonthLabel+StartYear | 0.44 |

**R - About the Generated Models:**

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. Adjusted R^2 values, as a result, were used rather than adjusted R^2 values to penalize multiple variables in models.

In regression analysis, we would have liked our regression model to have significant variables, low or extremely p-values, and to produce a high R-squared value. A combination of low P value, high statistical significance and, high R^2 indicates that changes in the predictors are related to changes in the response variable and that the model explains a lot of the response variability. However, our models do not have such a combination. We have a high statistical significance but relatively low adjusted R^2 values. Looking at a residuals graph of the East Bay, with the lowest adjusted R^2 of 0.10, we see high extremely variability amongst the residuals per fitted values. Along with an extremely high statistical significance (p<2.2e-16), there is a trend in the data but precision of this model, and other models with relatively low adjusted R^2 values, may be extremely imprecise.

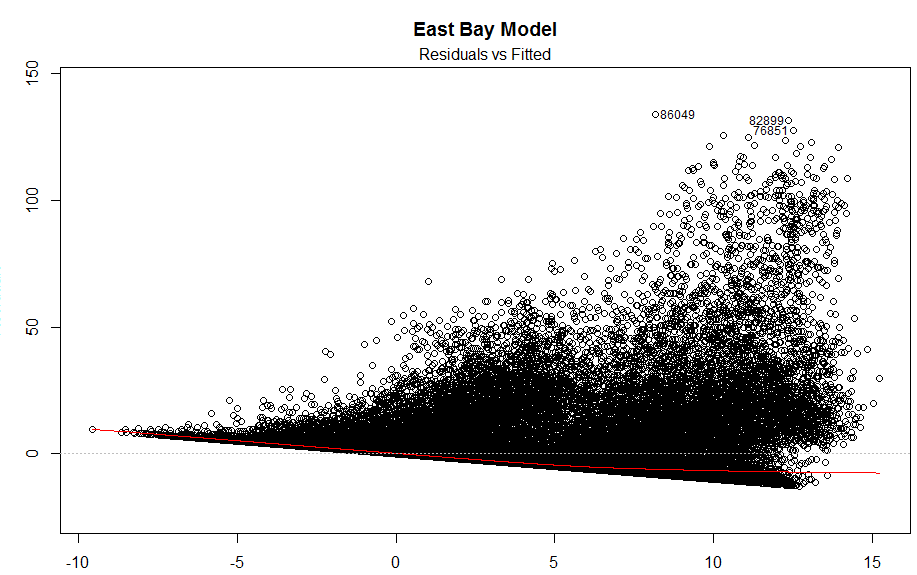


Figure 6

**Conclusion:**

Initial trends in the data were observed using K-means clustering, that gave us insight on how different variables could be correlated. Using this Principal Components Analysis, we could see which variables are related to each factor. PCA allowed us to sort through our original 24 variables and focus on eight factors without losing any data. Finally, it has allowed us to identify patterns in the weather variables in regards to our bike share data. Additionally, our randomly generated models did show overall statistical significance but lacked high adjusted R^2 showing imprecise, yet statistically significant, estimates of hourly intra-region bike usage. Data did not vary with temperature as initially expected, possibly because California does not have a lot of variance in the temperature. States that have more varied weather could be analyzed in the future to see if temperature makes a bigger impact on the bike rental.

Appendix A:

|  |  |
| --- | --- |
| Figure 1 | Duration (Mins) - Age  Chart 1 |

|  |  |
| --- | --- |
| Figure 2 | Duration - Station ID  Chart 2 |

|  |  |
| --- | --- |
| Figure 3 | Figure 4 |

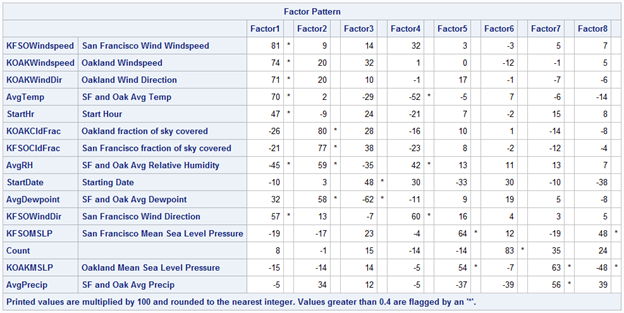


Figure 5

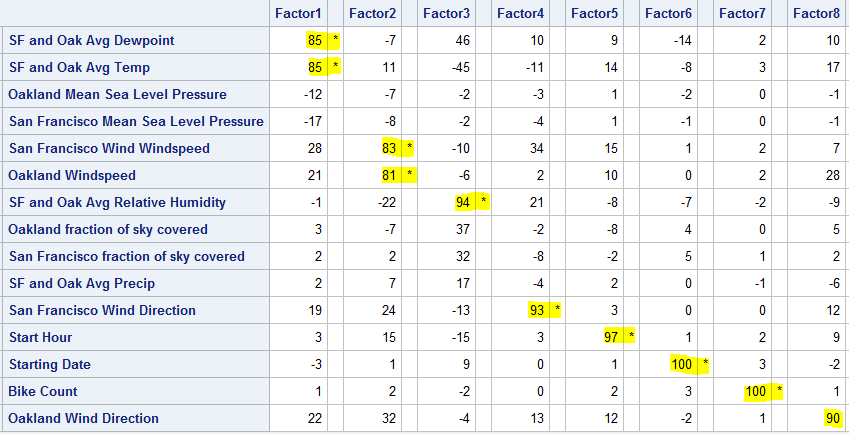


Figure 6

Appendix B:

Links

[Project GitHub](https://github.com/lepealec/Ford-GoBike-ML): <https://github.com/lepealec/Ford-GoBike-ML>

Ford GoBike Data: <https://s3.amazonaws.com/fordgobike-data/index.html>

Ford GoBike System Data: <https://www.fordgobike.com/system-data>

Hourly Weather Data: [www.frontierweather.com/historicaldataonly](http://www.frontierweather.com/historicaldataonly)

FGB Station Map: <https://member.fordgobike.com/map/>

Misc. Ford GoBike Data Tools:

Ford GoBike Personal Data Cleaner: <https://dieteto.shinyapps.io/FordGoBikeDataCleaner/>

Ford GoBike Daily Trip Visualizer: <https://dieteto.shinyapps.io/FGBTripVisualizer/>

Ford GoBike Live Data: <https://dieteto.shinyapps.io/FGBLiveData/>

Ford GoBike Maps: <https://dieteto.shinyapps.io/FGBMaps/>

1. [www.frontierweather.com](http://www.frontierweather.com) [↑](#footnote-ref-1)
2. <https://www.fordgobike.com/system-data> [↑](#footnote-ref-2)
3. San Francisco Ferry Building (Harry Bridges Plaza) [↑](#footnote-ref-3)
4. San Francisco Caltrain (Townsend St at 4th St) [↑](#footnote-ref-4)
5. Appendix A, Figure 1 and Appendix A Chart 1 [↑](#footnote-ref-5)
6. Appendix A, Figure 2 and Appendix A Chart 2 [↑](#footnote-ref-6)
7. Appendix A Figure 3 [↑](#footnote-ref-7)
8. Appendix A Figure 4 [↑](#footnote-ref-8)
9. Appendix A Figure 5 [↑](#footnote-ref-9)
10. Appendix A Figure 6 [↑](#footnote-ref-10)