The Journal of Open Source Software

Subscribers versus Customers – Bike Share Analysis using bikr R Package and Binary Logistic Regression

Daniel Van Veghel[1]

1 MSc Candidate, McMaster University

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Introduction

A public Bike Share system is one that rents out bicycles in different locations across a particular region, to users for membership fees – akin to a Transit Pass – or via a pay-per-distance-unit system (Hosford et al., 2019). The first Bike Share system was known as *Witte Fietsen* and was developed in Amsterdam in 1965 (Zee, 2016). Since this time, Bike Share systems have grown greatly in popularity, and since 1994, 31 successful Bike Share programs have been started in North American urban markets alone(Shaheen, Martin, Cohen, Finson, et al., 2012). One such Bike Share program is *Bike Share Pittsburgh, Inc.*, which was established in 2012 and launched its *Healthy Ride* bike system in 2015; it has expanded to more than 100 bike share hubs since then (Pittsburgh Bike Share, 2021). With Bike Share programs Pittsburgh Bike Share growing in popularity as a sustainable, affordable, and healthy means of intra-urban transportation, an analysis of rides and ridership is important for future planning within the organizations themselves, as well as at the municipal or regional government levels.

This paper leverages the `R` statistical and data-visualization programming language, to further analyze Pittsburgh Bike Share Ridership trends. This includes the development of a new `R` package: `bikr`, available for download publicly. The package includes built-in Bike Share trip data, as well as several functions capable of providing data distribution visualization and significance testing. Finally, a simple, Binary Logistic Regression Model is presented, using `R` functionality, to predict whether the Bike Share Trip is completed by a "Subscriber" or a "Customer," using a couple of variable inputs.

In keeping with the importance of openness and accessibility – as promoted by systems like bike share programs – all material, data, packages and source coding for this project are included within this paper itself, or this paper provides direct links to the public data repositories housing the additional materials.

Methods

Data

Data for this project was a Dataframe generated initially from a CSV file available here (Western Pennsylvania Regional Data Center, 2021). The Dataframe contains 10 variables (listed below) and 14,619 observations – individual Bike Share trips for the first Quarter of 2021 (January to March). The data was enhanced using a publicly-available Dataframe from the Western Pennsylvania Regional Data Centre, to include Latitude and Longitude coordinates for Start and End Hubs, as well as the number of bike racks at each hub. The dataframe was also enriched with several additional estimated or binary variables. The Maximum Distance capable in trip ("Max_Dist") was estimated using the Trip Duration (converted to hours) and an average cyclist biking speed of 22 km/h. Finally, a binary variable was coded for trips that started and ended at the same hubs.

The final aggregated and pre-processed dataset contained 18 variables. These variables, along with their datatypes, are described in Table 1 below:

Table 1: Variables and Variable Data Types

Name	Alias	Datatype
trip.id	Trip ID	Nominal
Starttime	Trip Start Time	Date/Time
Stoptime	Trip End Time	Date/Time
Bikeid	Bike Share Bike ID	Nominal
TripDuration	Length of Trip in Minutes	Continuous
From.station.id	Origin Hub ID	Nominal
To.station.id	Destination Hub ID	Nominal
From.station.name	Origin Hub Name	Nominal
To.station.name	Destination Hub Name	Nominal
Usertype	Bike Share User Designation	Nominal
Number of Racks	Number of Racks Available at Hub	Discrete
Start_Lat	Latitude of Start Hub	Continuous
Start_Log	Longitude of Start Hub	Continuous
End_Lat	Latitude of End Hub	Continuous
End_Log	Longitude of End Hub	Continuous
Max_Dist	Estimated Max Distance of Trip	Continuous
Same_Dest	Start and End Hubs are the same	Binary, Nominal
Is_Subscriber	Is a Subscriber	Binary, Nominal

Descriptive Statistics

Table 2: Continuous Variable Descriptive Statistics

Variable	Mean	Median	StDev	Min	Max
Trip Duration	4880.41	1136.00	14400.18	60.00	161392.00
# of Racks	12.29	13.00	6.72	5.00	31.00
Max Distance	29.82	6.94	88.00	0.37	986.28
Start Latitude	39.84	40.44	6.99	-40.47	40.48
Start Longitude	-79.97	-79.96	0.03	-80.01	-79.90
End Latitude	39.94	40.44	6.36	-40.47	40.48
End Longitude	-79.97	-79.97	0.03	-80.01	-79.90

'bikr' `R` Package

A simple, but complete, `R` Package was created as part of the analysis. It contained two functions and the dataset used in the analysis. The two functions are `visualize_tripdurations` and `compare_means`. The first function is a parameter-free function which provides several data visualization plots of the distribution of trip durations in the dataset. These visualizations include a histogram, a boxplot, and a barplot of average durations by user type (Subscriber, Customer or Not Indicated). The second function is a statistical tool which performs an independent-sample t-test calculation. It has several parameters: `mean1`, `mean2`, `var1`, `var2`, `n1` and `n2`, which are the first sample mean, second sample mean, first sample variance, second sample variance, first sample size and second sample size, respectively.

The package and statistical function within `bikr` will be used in the subsequent analysis, as an example of a means to determine whether the population of bike share riders are heterogeneous or homogenous, which will be important to know for the subsequent Binary Logistic Regression.

Methods

Sample homogeneity

Sample homogeneity is important in Multivariate Analysis, and so the data could be tested to determine if there were significant differences in Trip Durations between samples of Subscribers and Customers. To exemplify how to determine sample homo or heterogeneity, the `compare_means` tool from the `bikr` package was run on randomly-generated 30-case samples. This was done using process below. The first output value from the code is the t-value, and the second is the corresponding p-value.

```r
## Random sample (n=30) of the data
sam <- as.data.frame(join2[sample(nrow(join2), 30), ])

## Calculate means, variances and counts by User Type in the sample
avgs <- aggregate(sam$Tripduration, by = list(sam$Usertype), FUN = mean)
vars <- aggregate(sam$Tripduration, by = list(sam$Usertype), FUN = var)
ns <- aggregate(sam$Tripduration, by = list(sam$Usertype), FUN = length)

## Run Independent Sample t-test
bikr::compare_means(avgs[1, 2], avgs[2, 2], vars[1, 2], vars[2, 2], ns[1, 2], ns[2
    2])
```

```
## [[1]]
## [1] 1.572833
##
## [[2]]
## [1] 0.1269905
```

While this sample is very small, considering the overall size of the dataset, and a more complete and large-scale test of homogeneity is necessary, the `bikr` package provides a means of testing the data for Multivariate Assumptions and homogeneity.

Multivariate data assumptions

The multivariate data assumptions are Normality, Linearity, and Homoscedasticity (Hair, Black, & Babin, 2019). While more rigorous testing is generally needed than that provided in this paper (such as Shapiro-Wilk testing, e.g.), evaluations of these assumptions can be made visually through plots. While bikr offers data visualization capabilities, the complex plots offered by the `ggplot2` package are best suited for this visual analysis. The normality of three continuous variables have been tested: “Trip Duration,” “Start Hub Latitude” and “Start Hub Longitude.” The figures below showcase these variables’ distributions.

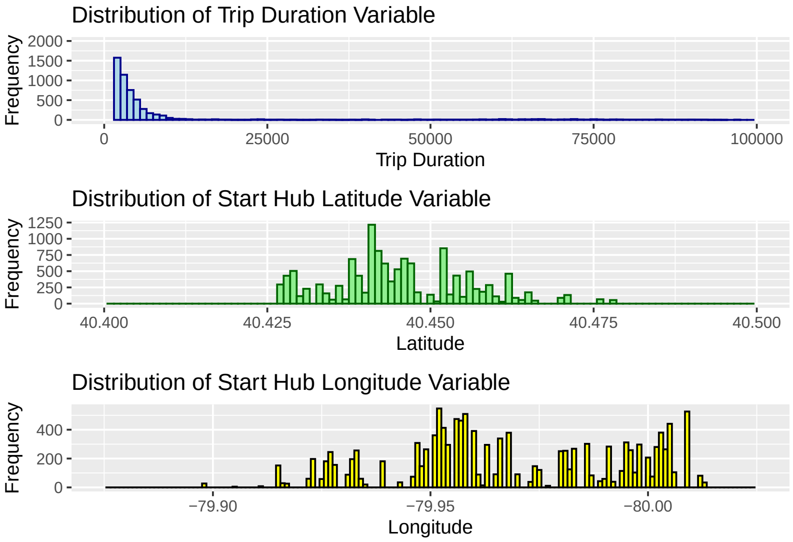

Figure 1: Continuous Variable Distributions

These variables are not normally distributed. They are each skewed; the Trip Duration variable is heavily skewed to the right. This data violates the normality assumption.

Next, the variables are tested for the Multivariate Assumption of Linearity. This can be seen in the following correlation matrix.

Table 3: Correlation Matrix for Continuous Variables

	TripDuration	X._Racks	Max_Distance	Start_Lat	Start_Long	End_Lat	End_Long
TripDuration	1.00	0.06	1.00	0.02	-0.11	0.01	-0.11
#_Racks	0.06	1.00	0.06	0.07	-0.21	0.09	-0.22
Max_Distance	1.00	0.06	1.00	0.02	-0.11	0.01	-0.11
Start_Lat	0.02	0.07	0.02	1.00	-0.14	0.51	-0.13
Start_Long	-0.11	-0.21	-0.11	-0.14	1.00	-0.12	0.82
End_Lat	0.01	0.09	0.01	0.51	-0.12	1.00	-0.12
End_Long	-0.11	-0.22	-0.11	-0.13	0.82	-0.12	1.00

Several variables exhibit high correlations, including Trip Duration and Max Distance (*1.00*), as well as Starting Longitude and Ending Longitude (*0.82*).

Binary Logistic Regression

A Binary Logistic Regression is a multivariate statistical data model, which uses a variety of categorical or numerical variables as independent variables, to predict the probability or odds of the occurrence of an event (Hair et al., 2019). The dependent variable in the model is a binary variable, where 1 indicates the occurrence of an event, or the presence of a quality, and 0 indicates the absence of the event or quality. While the data was well-suited for this model type, the Binary Logistic Regression method was also used because the continuous variables, as shown in the *Data* section of the paper, did not meet the multivariate data assumptions of Normality, Linearity, and Homoscedasticity. Binary Logistic Regression does not require treatment to meet these assumptions, and so it was the optimal analysis technique for the data.

The standard multiple linear regression follows the format:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Where y is a metric dependent variable, β_0 is the intercept of the multivariate plane-of-best-fit, β_n is the regression coefficient for x_n, an individual independent variable, and epsilon is representative of the model error.

The Binary Logistic Regression equation, however, is defined as:

$$P(Y) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

Where $P(Y)$ is the probability of the occurrence of event Y, α is the y-intercept, β are the variable coefficients and X is an independent variable.

The generalized function follows an S-Curve of the form:

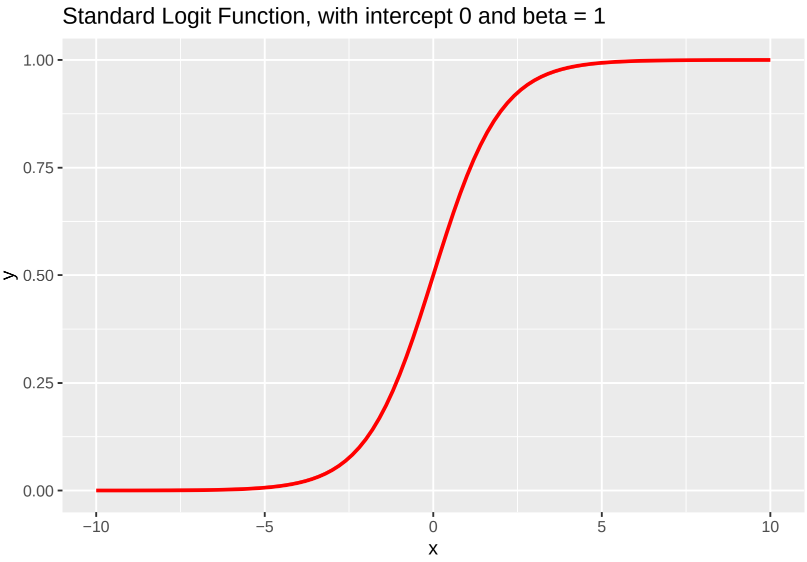

Figure 2: Basic Logit Function

Three potential variables are to be considered in the analysis: "Trip Duration," "Max Distance" and "Same Destination." First, correlations between variables must be checked. Any variable pairing with a correlation greater than 0.7 (particularly those approaching perfect correlation) must be addressed (Hair et al., 2019). The variable pair's correlation can be inspected visually:

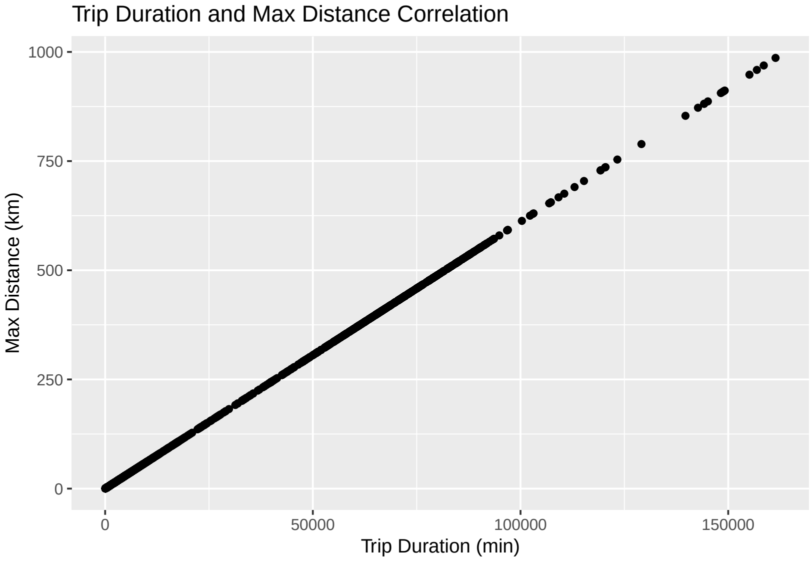

Figure 3: Trip Duration and Max Distance

These variables exhibit singularity – a perfect correlation. This intuitively makes sense; the "Max Distance" variable was generated by applying multiplication and division to the "Trip Duration" variable. Therefore, because the one is perfectly scaled with the other, it is perfectly positively correlated. The Max Distance variable was removed from the analysis, and the measured variable was used instead.

Results

The Model

The model was generated using the `glm` function in `R`.

Table 4: Logistic Regression Results

Predictor	B	Std. Error	t-Statistic	p-Value
Intercept (B0)	1.172	0.032	36.676	0
Trip Duration	-0.001	0.000	-35.066	0
Same Destination as Origin	-1.058	0.050	-21.085	0

Note:
p-values of '0' could be a result of large sample size, and therefore large degree of freedom

A simple 'goodness of fit' parameter can be calculated for the estimated model using the formula:

$$R^2 = 1 - (\frac{R_d}{N_d})$$

Where R_d is the Residual Deviance, and N_d is the Null Deviance. This produces an R^2 of 0.244. The final Logit Model is written as:

$$Logit_i = 1.172 - 0.00065t_t - 1.058s_d$$

Where t_t is the Travel Duration, and s_d is the binary “Same Destination” variable. As the continuous variable increases, there is a decrease in the log(odds) of the individual taking this trip being a subscriber. Subsequently, having the destination the same as the origin for the trip presents a significant decrease in the log(odds) of a Subscriber taking the trip, with respect to a trip with different start and end hubs.

Model Accuracy

The Figures below indicate the proportions of actual occurrences of all observations predicted as “Subscribers” by the model:

```r
## New Same_Dest variable that is not a 'factor'
join2$Same_Dest_2 <- ifelse(join2$To.station.id == join2$From.station.id, 1, 0)
join2$Prob <- (exp(1.172 - 0.00065*join2$Tripduration) -
                 1.058*(join2$Same_Dest_2))/
  (1 + (exp(1.172 - 0.00065*join2$Tripduration) - 1.058*(join2$Same_Dest_2)))

join2$Prediction <- ifelse(join2$Prob > 0.50, 1, 0)


ggplot(join2, aes(x="", y=Prediction, fill=Usertype))+
geom_bar(width = 1, stat = "identity") + xlab(" ")
```

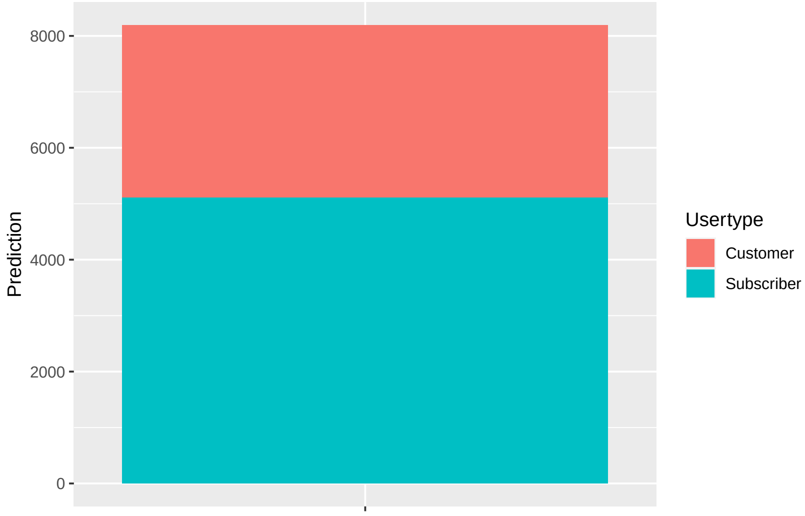

Figure 4: Actual Make-Up of Predicted Subscriber Trips

The Model was validated through the manual generation of a confusion matrix. This prevented the need for further package installation.

Table 5: Model Predictive Accuracy Confusion Matrix

		Actual Occurrence	
		1	0
Predicted Value			
Pred_Positives	1	5105	3089
Pred_Negatives	0	4112	719

This model produces a Sensitivity ($\frac{TP}{TP+FP}$) of **0.623** and a Recall ($\frac{TP}{TP+FN}$) of **0.553**. These values are quite low for model prediction accuracy.

Discussion

While this model would be unacceptable to present as a potential means of predicting User Types accurately, it is successful in applying statistical techniques to the dataset presented. However, there is merit to the model produced and its associated coefficients. To begin, the β_0 intercept value is 1.172. As the intercept, is no other values for other independent variables were present, this would mean the Logit value was 1.172. We convert Logits to Probabilities of Event Occurrence via:

$$P(Y) = \frac{e^{Logit_i}}{1 + e^{Logit_i}}$$

This means that if the Trip Duration was 0 seconds (i.e. no trip) and therefore the destination variable was removed as well, the probability of the user being a subscriber would be $\frac{e^{1.172}}{1+e^{1.172}}$ or 76.3%. Additionally, Trip Durations is seen as having a near-zero impact on the probability of event occurrence. This makes sense based on *Healthy Rides Pittsburgh* payment plans. The company's payments include 2 dollars for 30 minutes, or membership rates of 12 or 20 dollars/month for unlimited trips between 30 and 60 minutes (Pittsburgh Bike Share, 2021). This means that, because all trips are based on time, for the average person who is not regularly using bike share systems for longer trips, there is less benefit to being a member; if the trip length is beyond the 30 or 60 minute length, they would pay more – regardless of their subscriber status. This means, Trip Durations have very little impact on probability of subscriber status. The Same Destination variable, however, has a strong negative impact on the probability of Subscriber status. One could assume that subscribers are more regular users of the bike share system. Several studies have been conducted, via surveys and trip data, to determine the most prominent demographics of regular bike share users. Fishman et al., (2016), in a comprehensive review of bike share literature, found that "Commuting is the most common trip purpose for annual members" of bike share systems – citing work by Buck et al. (2013). As commuting is intuitively a trip from Point A to B (i.e. with a different start and end point), the model confirms the findings of other studies in the literature. Trips with the same origin and destination hubs are much less likely to be from a Subscriber than those with different start and end points.

Finally, the `bikr` package, though in its early development stages, presents an intriguing path toward more complex, publicly-available analysis toolkits for active transportation researchers. The ability for data analysis packages to be publicly accessed is important for promoting more widespread study of topics like bike share. While the package only has a single dataset, and two simple functions, it is a starting point for deeper and more meaningful analytical tools, as well as a wider range of datasets to explore.

This paper is subject to several distinct limitations. To begin, the dataset is limited in variables provided, and so several variables were derived via estimations. Analysis of a dataset with more environmental variables, such as weather or temperature, as well as more locational data would be beneficial for further exploration. More variables included in the Binary Logistic Regression could have also potentially produced a more accurate, well-fitting model than that which the study proposes. Additionally, more efficient, and stronger analyses could have been conducted with additional statistical packages; however, the study attempted to accomplish a general data analysis manually – using either `base R` functions or tools created in the `bikr` package. Finally, the study is limited in its scope – having focused only on the city of Pittsburgh and on one bike share system. Further work should be done on a larger geographic and organizational scope, to better understand bike share influences and trip patterns.

Conclusion

This study presents a rudimentary data analysis and basic Logistic Regression Model, on a sample bike share trip dataset for *Healthy Rides Pittsburgh* bike share system. It created a novel bike share trip analysis `R` package called `bikr`, and performed some basic modelling. It found that the length of the trip (in minutes) and whether the trip had the same start and end hubs influenced the probability of users being Subscribers to the *Healthy Rides* system. Finally, the paper kept with the principles of openness and accessibility – promoted by public-use systems like bike share – and contains all necessary code, data or links to data, in order for the work to be reproduced or enhanced by other researchers.

References

Buck, D., Buehler, R., Happ, P., Rawls, B., Chung, P., & Borecki, N. (2013). Are bike-share users different from regular cyclists? *Transportation Research Record: Journal of the Transportation Research Board*, *2387*(1), 112–119. doi:10.3141/2387-13

Fishman, E. (2016). Bikeshare: A review of recent literature. *Transport Reviews*, *36*(1), 92–113. doi:10.1080/01441647.2015.1033036

Hair, J. F., Black, W. C., & Babin, B. J. (2019). Multivariate data analysis.

Pittsburgh Bike Share. (2021). The organization: Healthy ride pittsburgh. Healthy Ride Pittsburgh. Retrieved from https://healthyridepgh.com/organization/

Shaheen, S. A., Martin, E. W., Cohen, A. P., Finson, R. S., et al. (2012). *Public bikesharing in north america: Early operator and user understanding.* Mineta Transportation Institute.

Western Pennsylvania Regional Data Center. (2021). Healthy ride trip data- 2021 Q1. *WPRDC*. University of Pittsburgh. Retrieved from https://data.wprdc.org/dataset/healthyride-trip-data/resource/7aaac45e-bc45-4219-a856-4d26c1706fdb

Zee, R. van der. (2016, April). *The Guardian.* Guardian News; Media. Retrieved from https://www.theguardian.com/cities/2016/apr/26/story-cities-amsterdam-bike-share-scheme