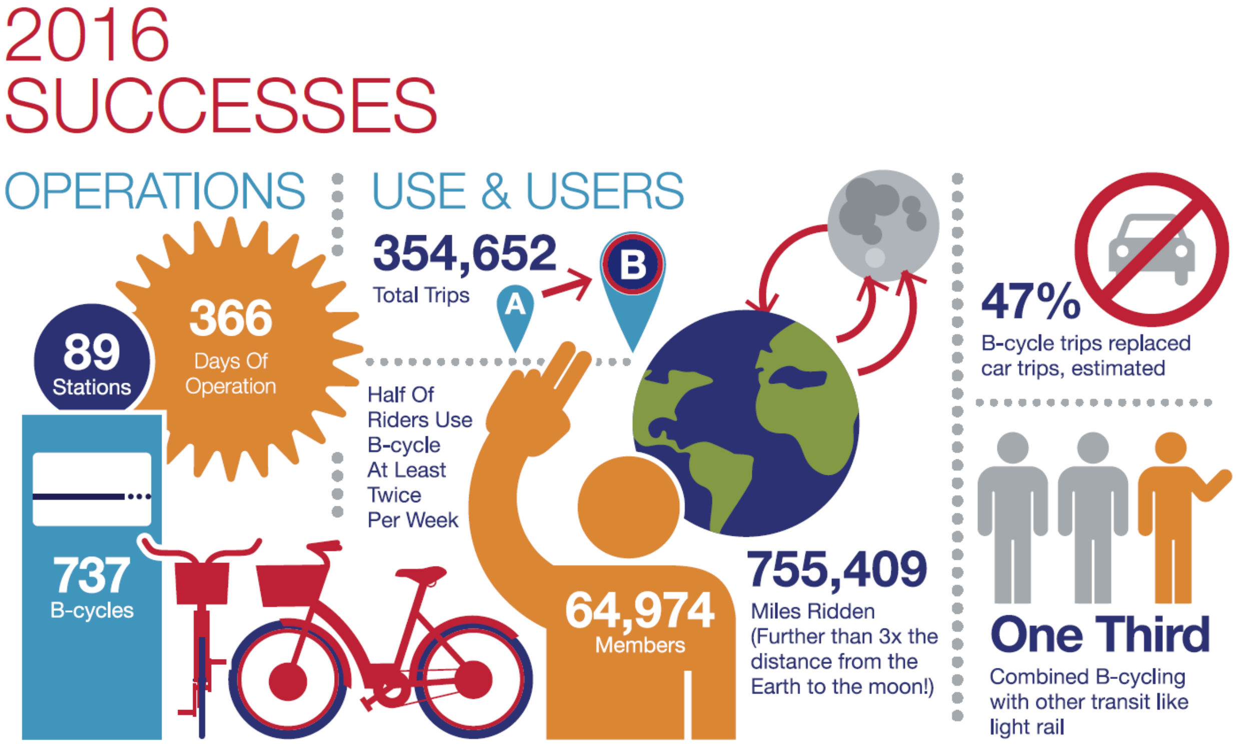
# Denver 2016 B-cycle Ridership Data Exploration


Source: http://denver.bcycle.com/docs/librariesprovider34/default-document-library/dbs_annualreport_2016_05.pdf

[Denver B-cycle](https://denver.bcycle.com) is a non-profit public bike sharing organization operating an automated bike sharing system called Denver B-cycle. Its mission is to "serve as a catalyst to fundamentally transform public thinking and behavior by operating a bike sharing system in Denver to enhance mobility while promoting all aspects of sustainability: quality of life, equity, the environment, economic development, and public health" its purpose, its organization and discuss its relevance to this exploration.

Denver B-cycle posts its trips data set on its website as soon as its annual report is released. Trips data have been available since 2010. The 2016 annual report and its associated dataset for this report were obtained from Denver B-Cycle website. The original plan was to use the 2015 dataset to continue the effort by Tyler Byler who published a report, Exploring 2014 Denver B-cycle Ridership. In his study Tyler indicated that "most calendar and clock variables were highly significant when predicting ridership, and weather variables such as temperature and amount of cloud cover appear to be as well". The original plan for this report was to use 2015 data to continue Tyler's work. However, the 2016 data became available at the end of February 2017, so gears had to be rapidly shifted to use this data instead. To this end, the reporting style will follow Tyler's study to provide seamless continuity and good reference on trends and analyses.

This study has three parts:

1. Explore the Trips datasets and visualize the data to provide useful and interesting information.
2. Deploy a variety of regression models to train and test the data.
3. Deploy a variety of classification models to train and test the data.

# Part 1: Data Exploration

## Data Acquisition

Data for this study was downloaded from several sources and combined using the following steps:

1. Downloaded B-cycle 2016 Trips and Kiosk data from Denver B-Cycle website. The columns names were changed to comply with Python code best practices.
2. Created a list of the 7921 combinations of the 89 checkout/return kiosks. Used Google Distance Matrix API to provide the bicycling distance and time between each checkout and return kiosk. Adopted Tyler's method of finding the average distance by taking the distance from each checkout-return pair's distance separately then averaging it. As he pointed out in his study, this approach was taken "because of the large number of one-way streets in the Denver downtown area where the kiosks are highly clustered". Google only supports a maximum of 2500 requests a day, it took four days to obtain this data.
3. Obtained daily and hourly weather data via Dark Sky API for all of 2016. Dark Sky supports up to 1000 requests per day.

### Basic Ridership Statistics

#### Number of Rides

The B-cycle data, as downloaded, contained 419,611 rows of trips data. Under normal circumstances this would mean that 419,611 B-cycle trips were taken in 2016. However, the 2016 Denver B-cycle annual report acknowledged 354,652 total trips for the year. The breakdown was as follows:

| Membership Type | Number of Trips |
|---|---|
| Annual (And Annual Plus) | 193,113 |
| Flex Pass | 3,565 |
| 30 Day | 54,004 |
| 24 hour online | 117 |
| 24-hour Kiosk | 103,853 |
| **Total Trips** | **354,652** |

The Trips dataset reported the following breakdown:

| Membership Type | Number of Trips |
|---|---|
| Annual (Denver B-cycle) | 82,199 |
| Annual Plus (Denver B-cycle) | 84,271 |
| Flex Pass | 3,565 |
| Monthly (Denver B-cycle) | 54,004 |
| 24 hour online (Denver B-cycle) | 117 |
| 24-hour Kiosk Only (Denver B-cycle) | 87,315 |
| **Total Trips** | **311,471** |

There were several other Membership Types that were also listed under "Denver B-cycle" in the User's Program:

| Membership Type | Number of Trips |
|---|---|
| Denver B-cycle Founder (Denver B-cycle) | 18,003 |
| Not Applicable | 64,959 |
| Single Ride (Denver B-cycle) | 16,526 |

In particular, the "Not Applicable" membership type accounted for more than 15% of the 419,611 trips. Perhaps some of these trips were used in the Denver B-cycle annual report.

Also over 2.3% of the Denver B-cycle rides (9,954 rides) had the same checkout station as return station with a trip duration of only 1 minute (Figure 1). Again, Tyler's explanation of why these trips should be removed from the dataset makes sense - "I believe these should be filtered out because I believe the majority of these "rides" are likely people checking out a bike, and then deciding after a very short time that this particular bike doesn't work for them. I believe that most of the same-kiosk rides under 5 minutes or so likely shouldn't count, but only culled the ones that were one minute long".

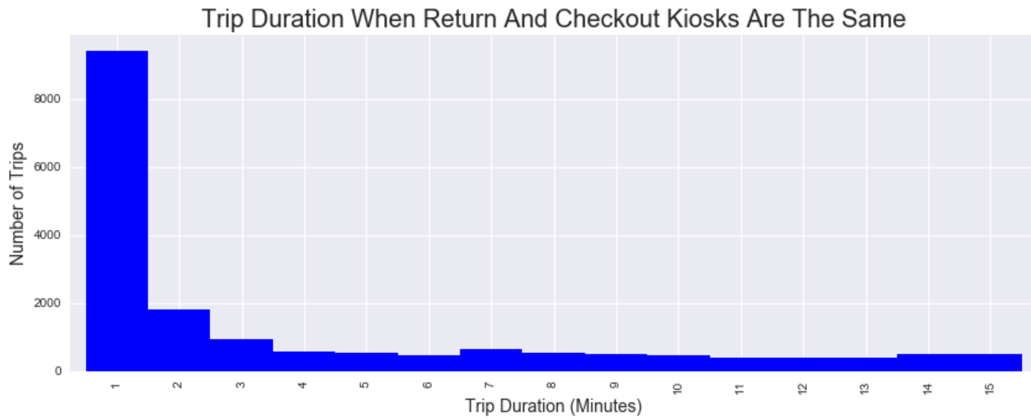



FIGURE 1: TRIP DURATION WHEN CHECKOUT AND RETURN KIOSKS ARE THE SAME

There were 6574 rows in the Trips dataset that had kiosk names not in the Kiosk Master List. These 6574 rows were removed accordingly.

Removing the 9,954 rows with a trip duration of 1 minute and 6574 rows with invalid kiosk names resulted in **394,431 Denver B-cycle rides in 2016**.

## Distance Traveled

To estimate the distance between checkout and return kiosks when they are the same, Tyler's method of using the "average speed of all the other rides (nominal distance ridden divided by the duration), and then applying this average speed to the same-kiosk trip durations" was adopted. This resulted in **670,802 miles ridden in 2016**.

## Most Popular and Least Popular Checkout and Return Kiosks

### Most Popular

The following ten kiosks were the most popular checkout kiosks by number of total bike checkouts in 2016.

| Checkout Kiosk | Number of Checkouts |
|---|---|
| 16th & Wynkoop | 11,174 |
| 16th & Broadway | 3,565 |
| 1350 Larimer | 10,837 |
| 18th & California | 9,865 |
| 1550 Glenarm | 9,441 |
| 18th & Arapahoe | **8,531** |
| 20th & Chestnut | 8,240 |
| 13th & Speer | 8,228 |
| REI | 8,218 |
| 16th & Little Raven | 8,198 |

The following ten kiosks were the most popular return kiosks by number of total bike checkouts in 2016.

| Return Kiosk | Number of Checkouts |
|---|---|
| 16th & Wynkoop | 11,289 |
| 1350 Larimer | 10,920 |
| 16th & Broadway | 10,870 |
| 18th & California | 9,863 |
| 1550 Glenarm | 9,501 |
| 18th & Arapahoe | 8,549 |
| 20th & Chestnut | 8,356 |
| REI | 8,284 |
| 13th & Speer | 8,272 |
| 16th & Little Raven | 8,267 |

## Least Popular

The following ten kiosks were the least popular checkout kiosks by number of total bike checkouts in 2016.

| Checkout Kiosk | Number of Checkouts |
|---|---|
| Pepsi Center | 1,795 |
| 32nd & Julian | 1,755 |
| 25th & Lawrence | 1,736 |
| Colfax & Garfield | 1,725 |
| 4th & Walnut | 1.663 |
| Decatur Federal Light Rail | 1,508 |
| Denver Zoo | 1,490 |
| Colfax & Gaylord | 1,421 |
| 17th & Curtis | 615 |
| 39th & Fox | 332 |

The following ten kiosks were the least popular return kiosks by number of total bike checkouts in 2016.

| Return Kiosk | Number of Checkouts |
|---|---|
| 21st & Market | 1,795 |
| 32nd & Julian | 1,767 |
| 25th & Lawrence | 1,758 |
| Colfax & Garfield | 1,743 |
| 4th & Walnut | 1,686 |
| Decatur Federal Light Rail | 1,537 |
| Denver Zoo | 1,468 |
| Colfax & Gaylord | 1,433 |
| 17th & Curtis | 632 |
| 39th & Fox | 345 |

## Map of Station Popularity

### Checkout Kiosks

The use of Tableau aided in the creation of the following map showing the popularity of the various Checkout Kiosks (Figure 2). The size of the circle is proportional to the number of checkouts from that kiosk in 2016.

Number of Checkouts by Checkout Kiosk

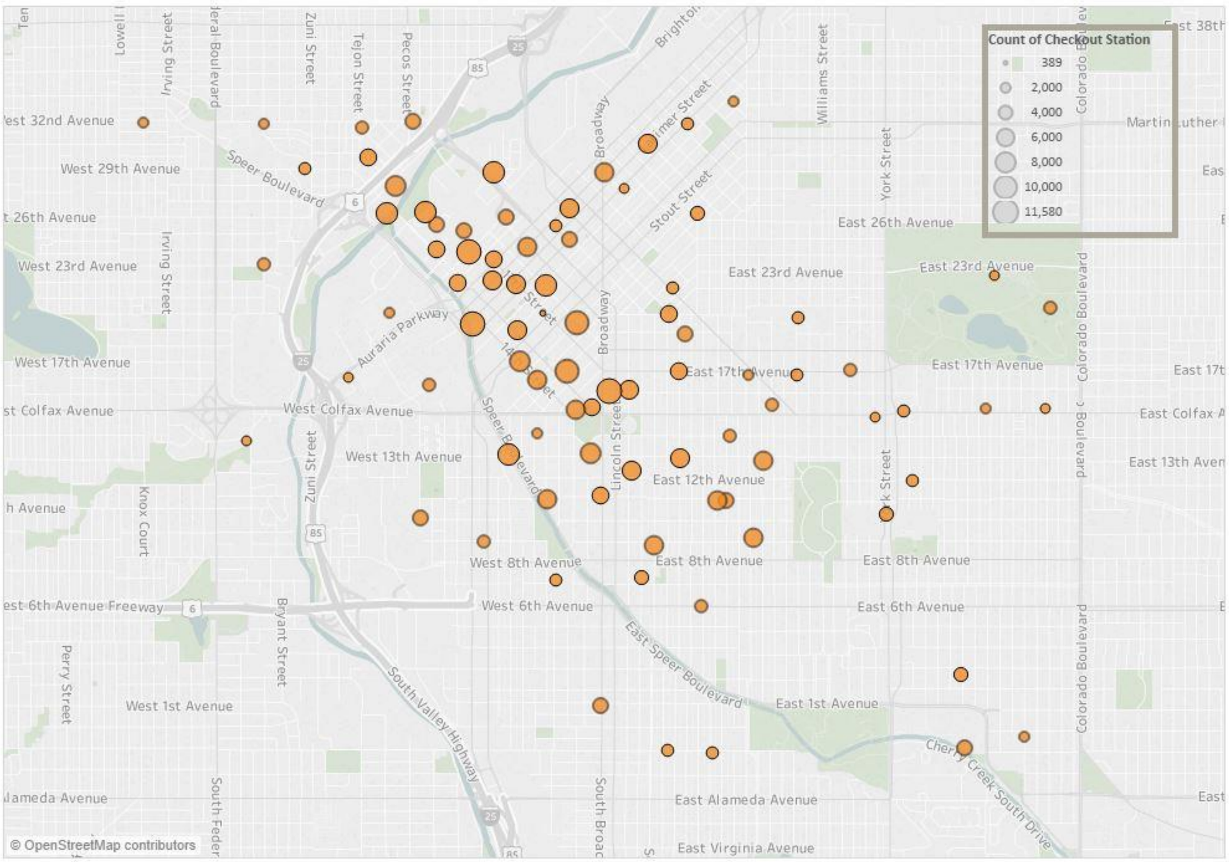



FIGURE 2: CHECKOUT KIOSK LOCATIONS AND NUMBER OF CHECKOUTS IN 2016

## Return Kiosks

Similarly, the use of Tableau aided in the creation of the following map showing the popularity of the various Return Kiosks (Figure 3). The size of the circle corresponds to the number of checkouts returned to that kiosk in 2016.

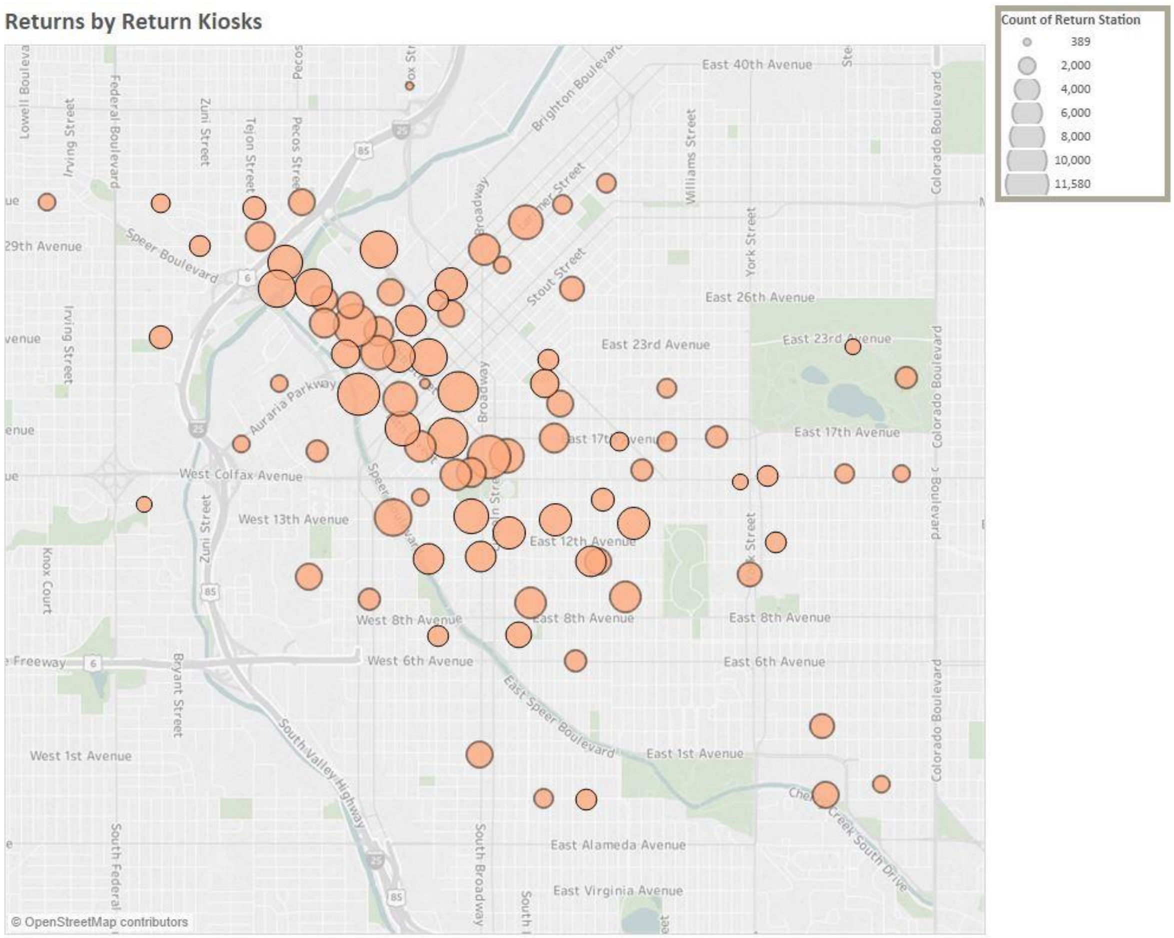



FIGURE 3: RETURN KIOSK LOCATIONS AND NUMBER OF RETURNS IN 2016

## Checkouts per Membership Type

Denver B-cycle has a number of different membership passes. The following were the top ten by number of checkouts in 2016 (Figure 4).

| Membership Type | Number of Checkouts |
|---|---|
| 24-hour Kiosk Only (Denver B-cycle | 85,680 |
| Annual Plus (Denver B-cycle) | 82,202 |
| Annual (Denver B-cycle) | 80,093 |
| Not Applicable | 56,250 |

| | |
|---|---|
| Monthly (Denver B-cycle) | 52,811 |
| Denver B-cycle Founder (Denver B-cycle) | 17,675 |
| Single Rider (Denver B-cycle) | 16,291 |
| Republic Rider (Annual) (Boulder B-cycle) | 5,930 |
| Flex Pass (Denver B-cycle) | 3,507 |
| Republic Rider (Boulder B-cycle) | 1,229 |

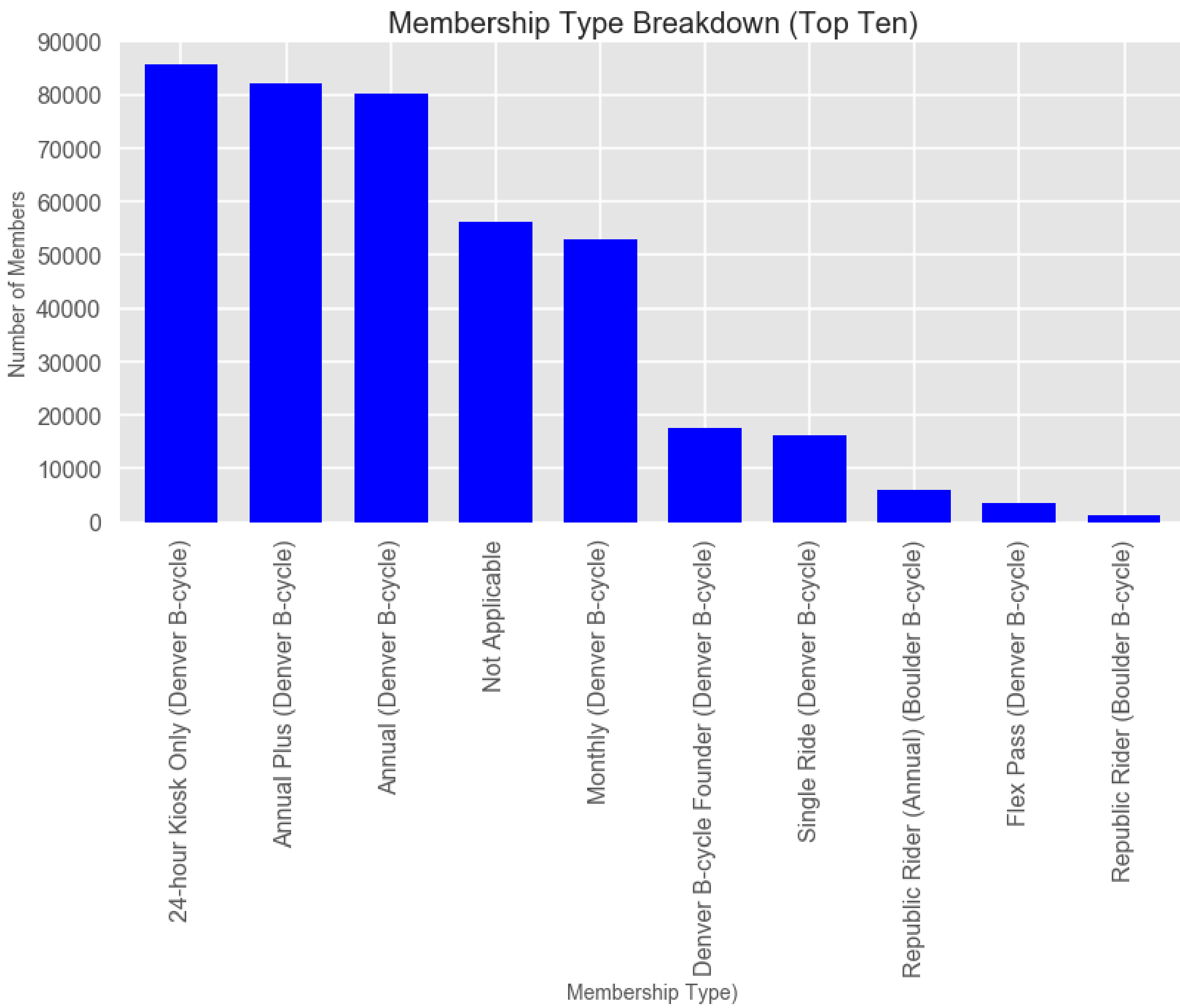



FIGURE 4: NUMBER OF CHECKOUTS BY MEMBERSHIP TYPE IN 2016

# Ridership by Calendar and Clock Variables

## Ridership by Hour

Bike checkout time is probably the most important attribute in the Trips dataset. Each checkout time was converted into its integer hour. For example, 7:02 AM or 7:59 AM would be converted to an integer of 7. In this way, total number of checkouts could be aggregated for the year and plotted against their hours of the day, as shown in Figure 5.

It appears that the highest number of checkouts occur between 4 PM and 5 PM with ridership increasing steadily from 10 AM onwards.

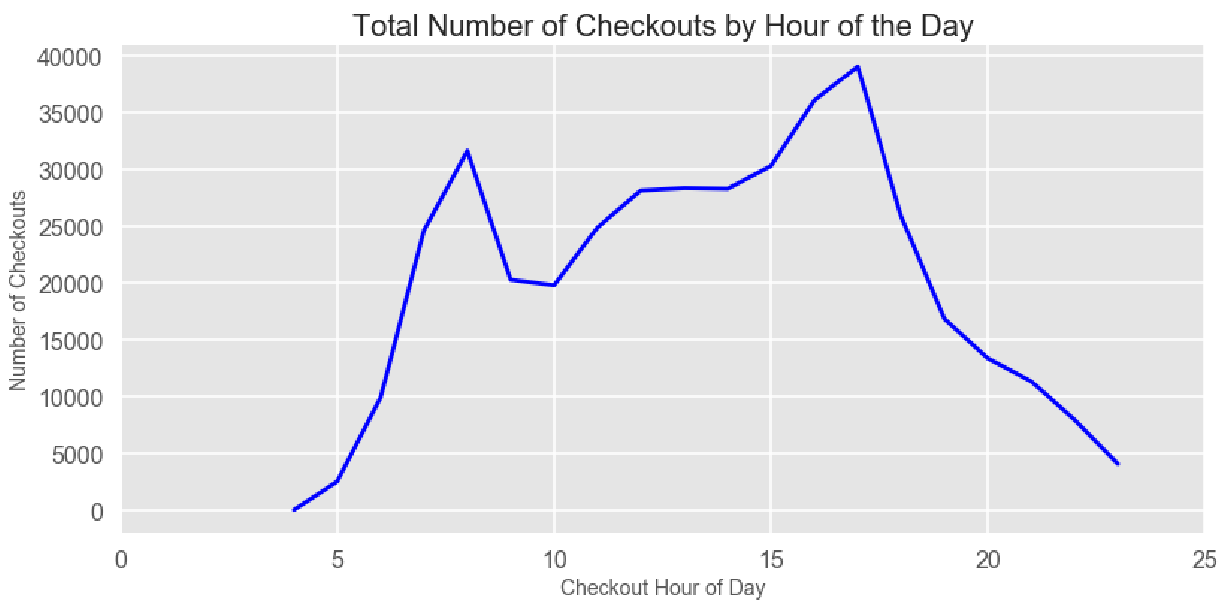



FIGURE 5: NUMBER OF CHECKOUTS BY HOUR IN 2016

Figure 6 shows the average distance ridden by the hour of the day in 2016. More distance is covered during the 10 AM period and declining steadily after 3 PM.

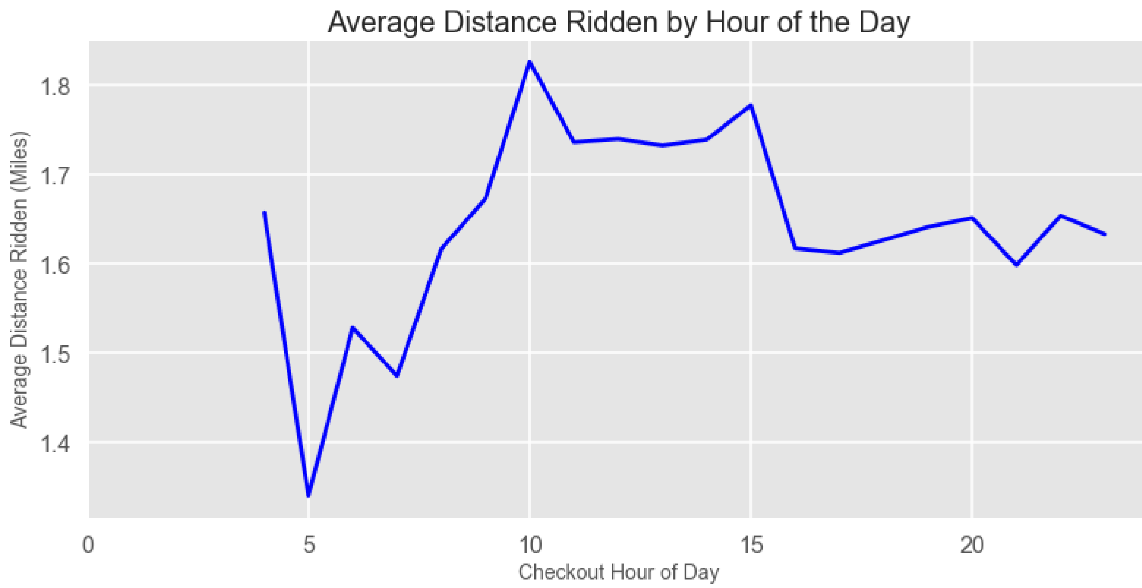



FIGURE 6: ESTIMATED AVERAGE MILES RIDDEN BY HOUR OF CHECKOUT IN 2016

## Ridership by Hour and Weekday

Figure 7 shows that weekday ridership patterns are similar. On the other hand weekend ridership demonstrate a busy afternoon (between 12 PM and 3 PM)

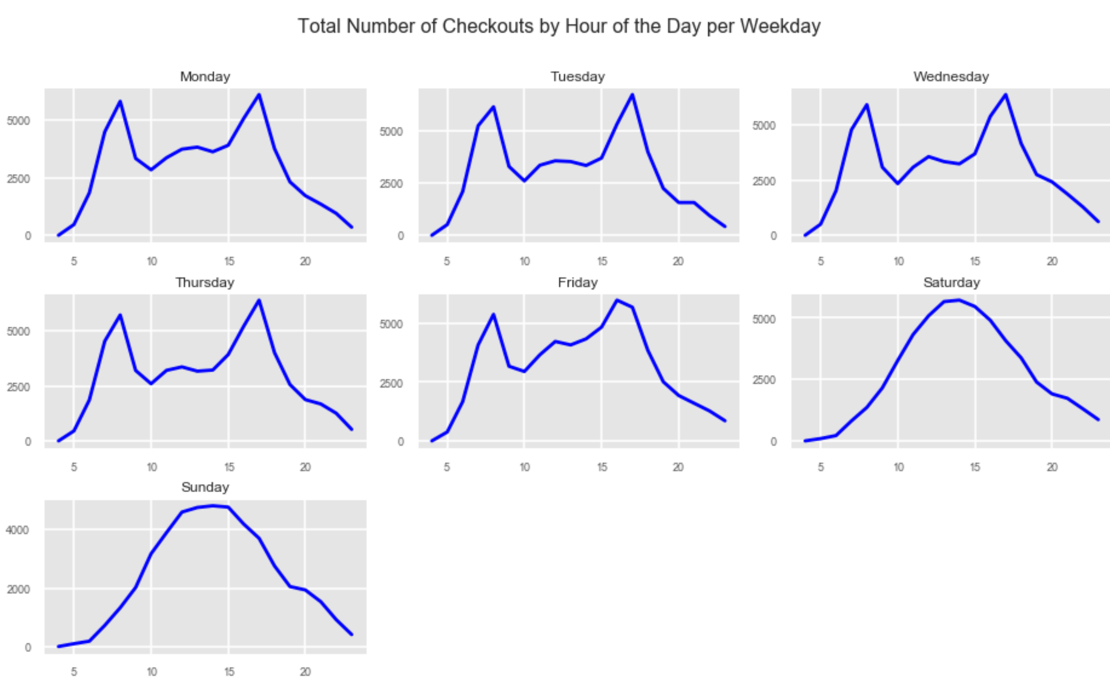



FIGURE 7: CHECKOUTS BY HOUR OF DAY PER WEEKDAY IN 2016

## Ridership by Month

Monthly checkouts, as shown in Figure 8, suggest high ridership during the summer months and low ridership during the winter months.

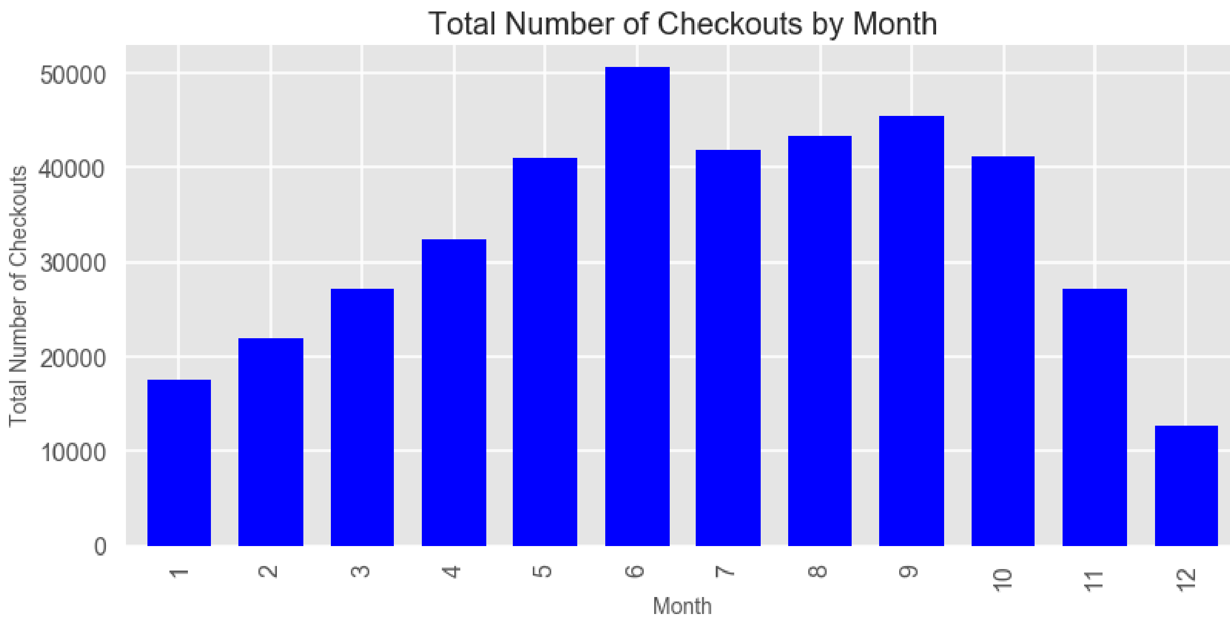



FIGURE 8: TOTAL CHECKOUTS BY MONTH IN 2016

# Merging with Weather

It is highly likely that weather plays a very important role in bike ridership and bike checkout times. This was shown in the previous plots on total checkouts per hour of the day, by weekday, and by month. To verify this, weather data obtained from [Dark Sky API] was merged with the Trips dataset and several graphs plotted to visualize the relationships.

## Checkouts vs. Daily Temperature

Figure 9 shows the total number of checkouts against maximum and minimum daily temperature. It clearly suggests that ridership increases as the temperature increases and vice-versa.

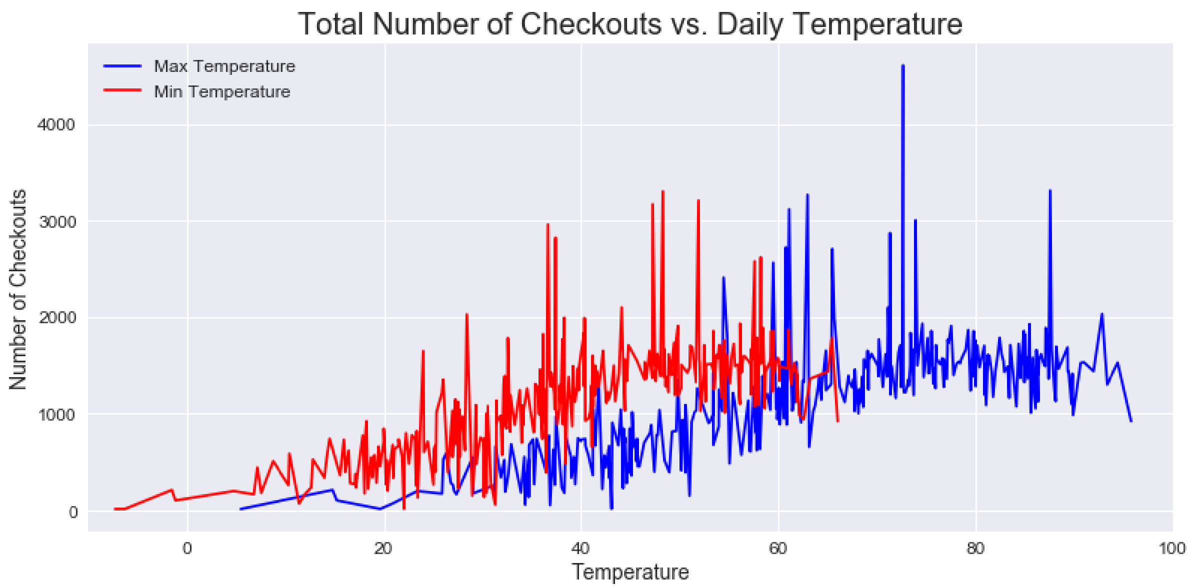



FIGURE 9: TOTAL CHECKOUTS BY DAILY TEMPERATURE IN 2016

Apparent temperature, as defined by Dark Sky, is “apparent (or “feels like”) temperature in degrees Fahrenheit”. It appears to have a subtle effect on bike ridership as shown in Figure 10.

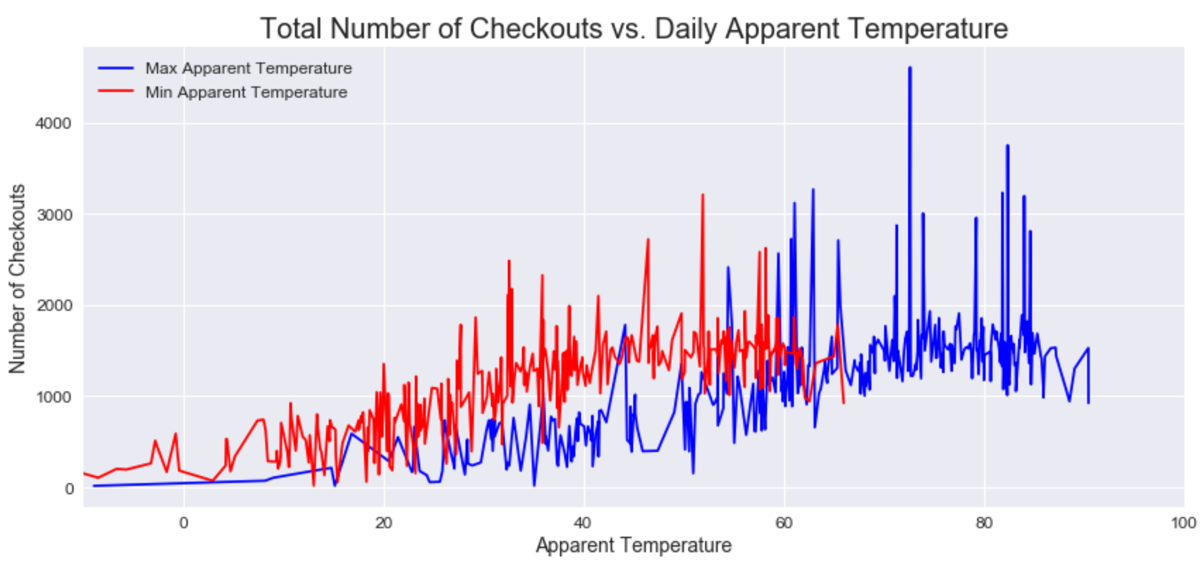



FIGURE 10: TOTAL CHECKOUTS BY DAILY APPARENT TEMPERATURE IN 2016

## Checkouts vs. Daily Cloud Cover

Dark Sky defines Cloud Cover as "the percentage of sky occluded by clouds, between 0 and 1, inclusive". Figures 11 shows the total number of checkouts against daily cloud cover. They clearly suggest that ridership is highest as the cloud cover stays at around 0.15.

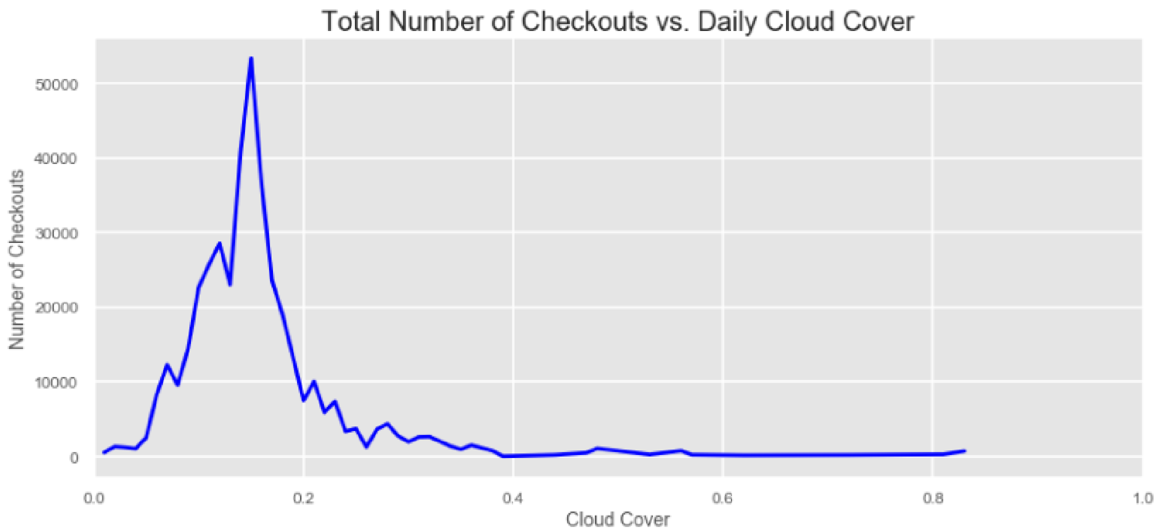



FIGURE 11: TOTAL CHECKOUTS BY DAILY CLOUD COVER IN 2016

## Checkouts vs. Daily Wind Speed

Wind speed is reported in miles per hour. As shown in Figure 12, ridership does not seem to be somewhat impacted by higher wind speeds.

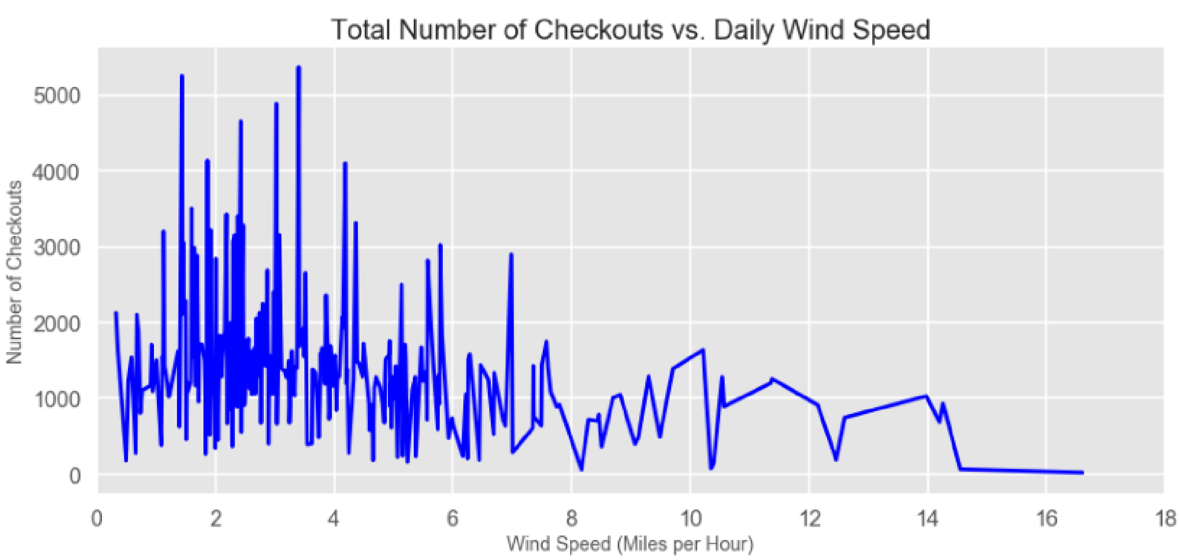



FIGURE 12: TOTAL CHECKOUTS BY DAILY WIND SPEED IN 2016

## Checkouts vs. Daily Humidity

Humidity is defined by Dark Sky as “relative humidity, between 0 and 1. Figure 13 shows decreased ridership at higher humidity levels.

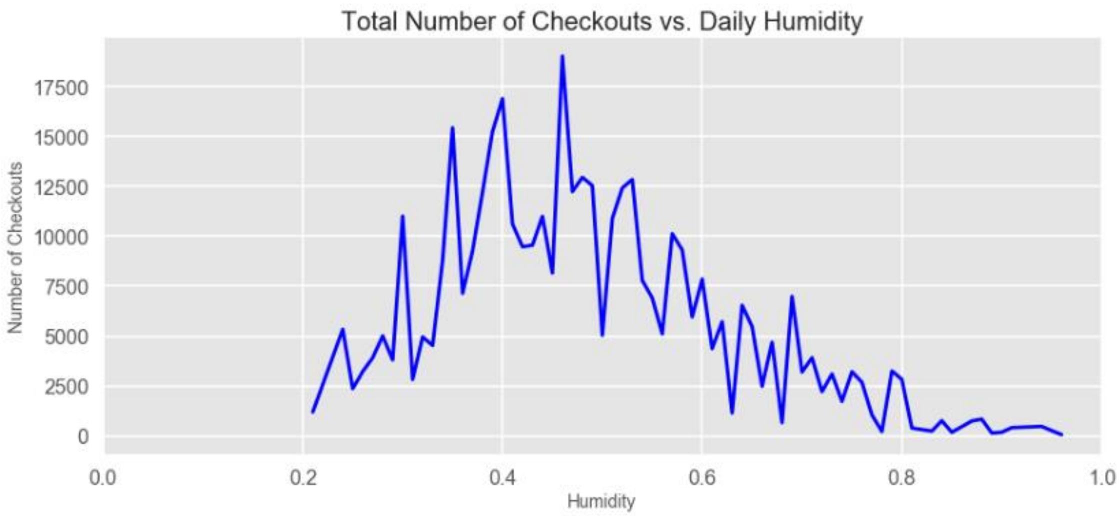



FIGURE 13: TOTAL CHECKOUTS BY DAILY HUMIDITY IN 2016

## Checkouts vs. Daily Visibility

Visibility is measured in miles and capped at 10 miles, according to Dark Sky. As Figure 14 shows, ridership peaks when visibility is at 10 miles.

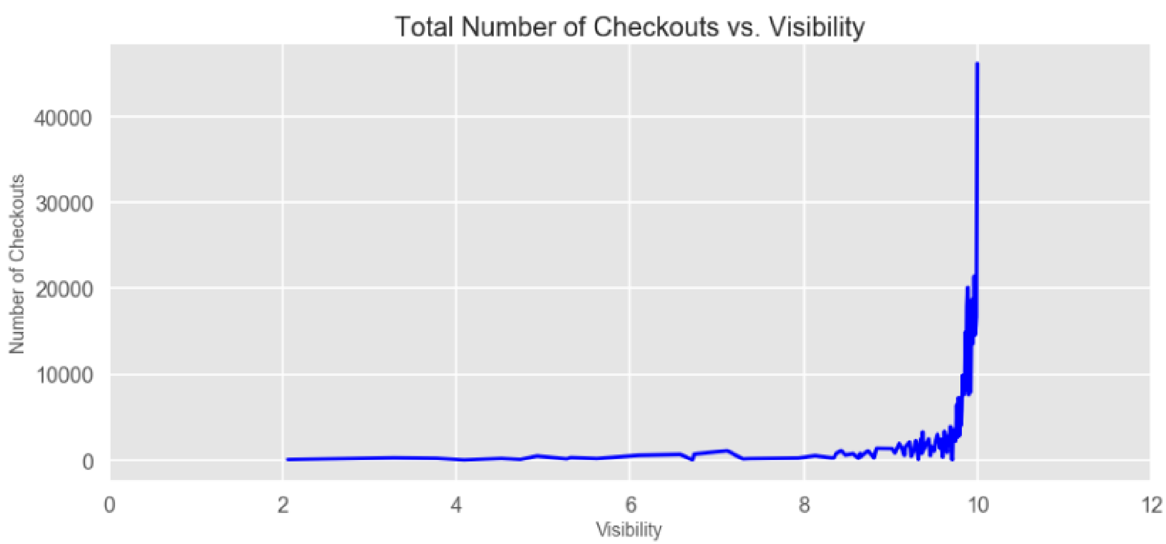



FIGURE 14: TOTAL CHECKOUTS BY DAILY VISIBILITY IN 2016

## Days with Highest/Lowest Ridership

Another interesting data discovery was the fact that Saturdays and Sundays had the highest and lowest ridership depending upon the weather. In his study, Tyler suggests that this may be due to "'weekend warriors' who rent B-cycles for pleasure and are highly affected by the weather in their decision to ride". This may well be the case.

### Highest Ridership

| Checkout Week Day | Checkout Date | Max Temperature | Min Temperature | Number of Checkouts |
|---|---|---|---|---|
| Sunday | 2016-05-29 | 71.090 | 44.100 | 2,100 |
| Saturday | 2016-05-28 | 65.650 | 40.330 | 1,990 |
| Friday | 2016-06-03 | 74.600 | 56.120 | 1,933 |
| Wednesday | 2016-06-15 | 85.430 | 51.980 | 1,927 |
| Saturday | 2016-06-21 | 77.510 | 49.790 | 1,909 |
| Monday | 2016-06-27 | 87.060 | 58.440 | 1,868 |
| Saturday | 2016-06-25 | 79.230 | 61.040 | 1,868 |
| Saturday | 2016-06-04 | 75.500 | 53.410 | 1,857 |
| Thursday | 2016-03-23 | 84.860 | 59.280 | 1,857 |
| Friday | 2016-09-02 | 79.770 | 59.500 | 1,855 |

### Lowest Ridership

| Checkout Week Day | Checkout Date | Max Temperature | Min Temperature | Number of Checkouts |
|---|---|---|---|---|
| Saturday | 2016-12-24 | 50.960 | 28.940 | 154 |
| Sunday | 2016-04-17 | 34.710 | 30.140 | 140 |
| Sunday | 2016-01-31 | 31.260 | 23.430 | 133 |
| Wednesday | 2016-12-07 | 15.250 | -1.110 | 105 |
| Tuesday | 2016-02-02 | 20.870 | 11.430 | 72 |
| Saturday | 2016-04-16 | 34.430 | 31.310 | 61 |
| Sunday | 2016-12-25 | 36.860 | 25.290 | 56 |
| Wednesday | 2016-03-23 | 43.070 | 22.040 | 18 |
| Sunday | 2016-12-18 | 19.640 | -6.220 | 17 |
| Saturday | 2016-12-17 | 5.490 | -7.220 | 16 |

## Checkouts vs. Hourly Weather Variables

Hourly weather conditions provide better resolution than daily weather conditions. To investigate this, number of checkouts against hourly weather variables were also plotted and compared with the plots using daily weather variables.

## Checkouts vs. Hourly Temperature

The scatter plots in Figure 15 and 16 show that the relationship between the number of checkouts and the hourly temperatures are not linear.

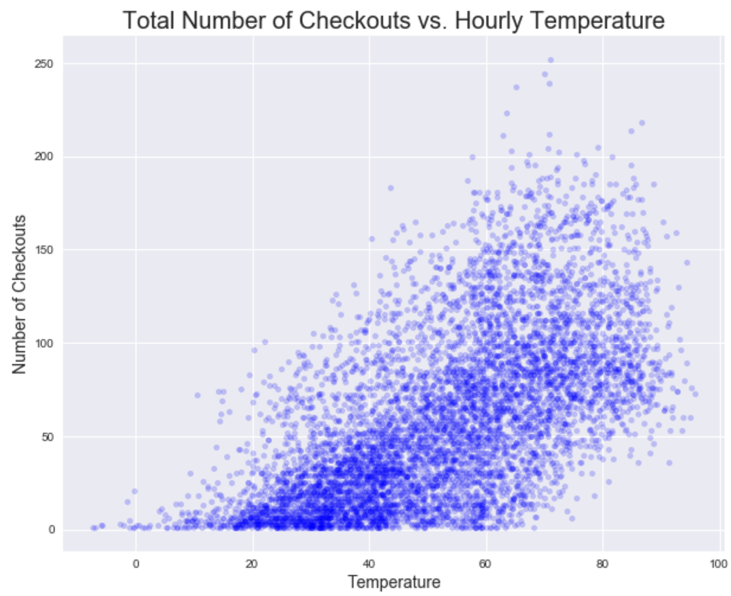



FIGURE 15: TOTAL CHECKOUTS BY HOURLY TEMPERATURE IN 2016

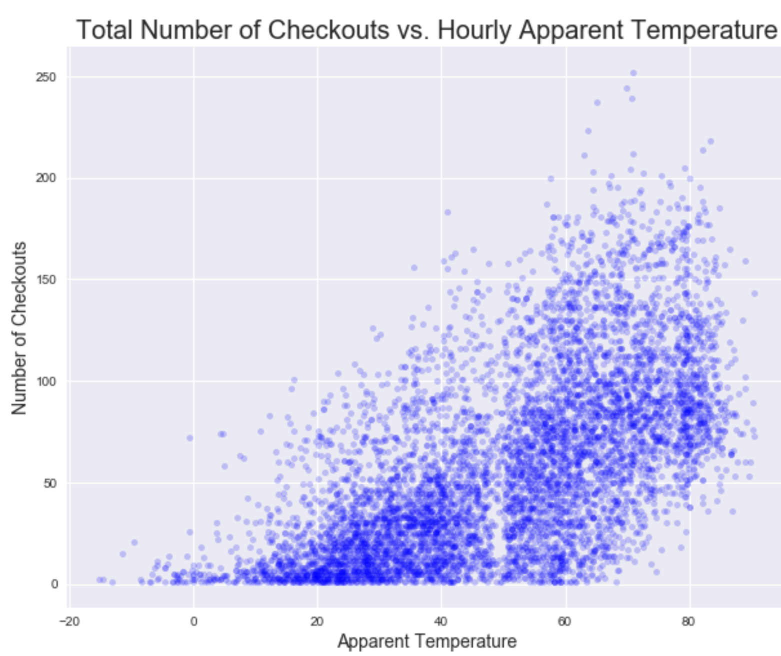



FIGURE 16: TOTAL CHECKOUTS BY HOURLY APPARENT TEMPERATURE IN 2016

## Checkouts vs. Hourly Humidity

Figure 17 shows that humidity affects ridership significantly.

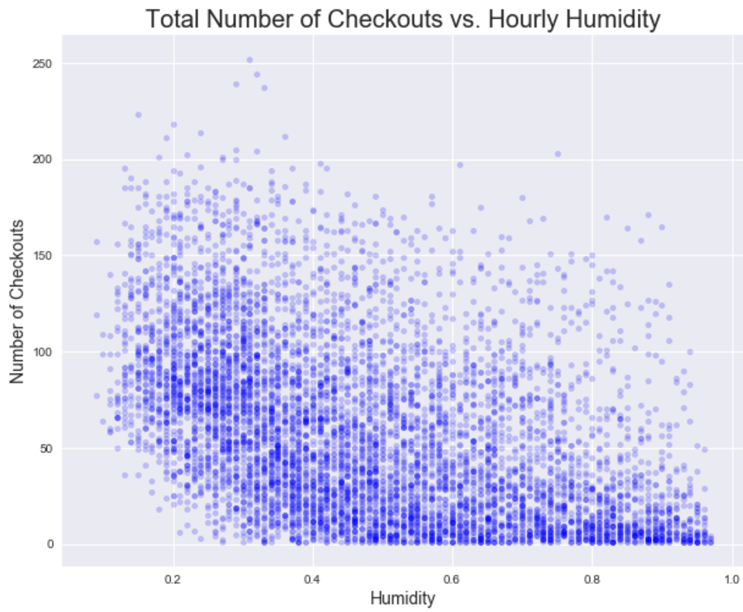



FIGURE 17: TOTAL CHECKOUTS BY HOURLY HUMIDITY IN 2016

## Checkouts vs. Hourly Cloud Cover

As shown in Figure 18 Cloud Cover certainly impacts ridership.

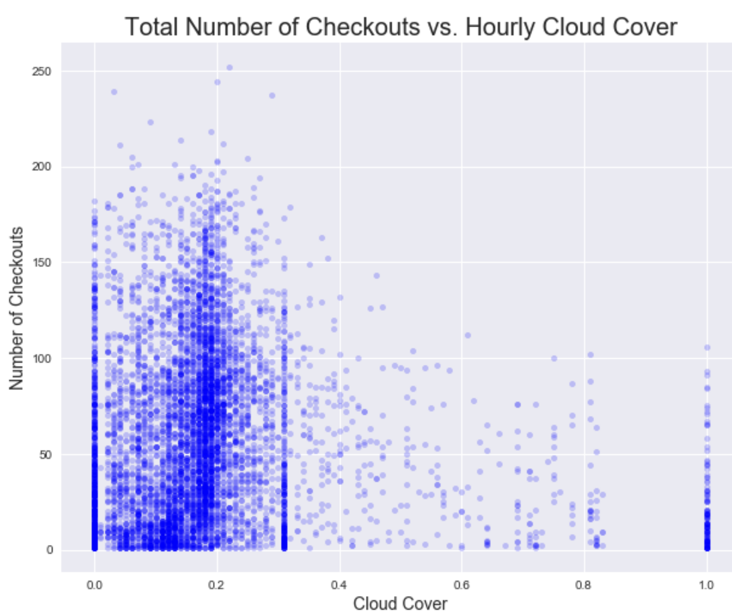



FIGURE 18: TOTAL CHECKOUTS BY HOURLY CLOUD COVER IN 2016

## Checkouts vs. Hourly Wind Speed

Data on wind speed indicates it is clustered heavily in 0 to 8 miles per hour range, as shown in Figure 19.

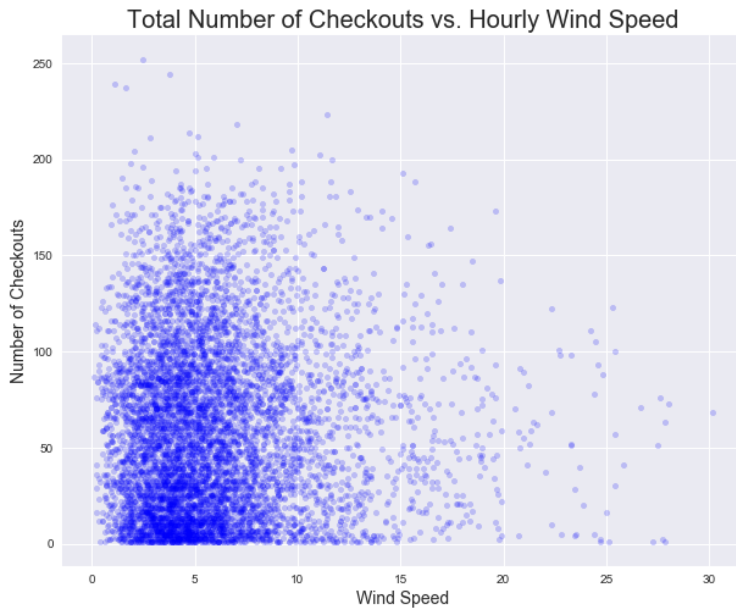



FIGURE 19: TOTAL CHECKOUTS BY HOURLY WIND SPEED IN 2016

## Checkouts vs. Hourly Visibility

As shown in Figure 20 visibility at 10 miles has the greatest impact on ridership.

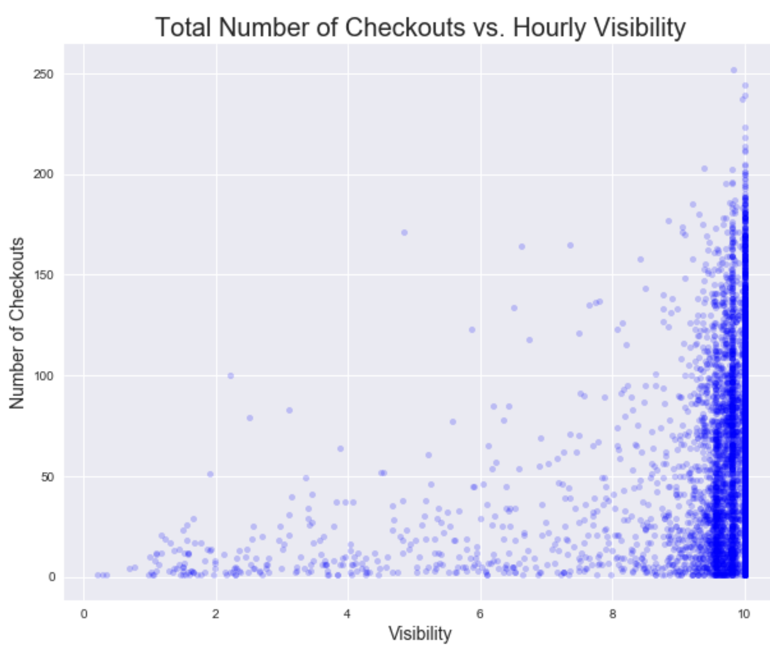



FIGURE 20: TOTAL CHECKOUTS BY HOURLY VISIBILITY IN 2016

# Part 2: Regression Modeling

In his study, Tyler attempted to create a linear regression model using a number of calendar and weather variables. Using temperature, temperature squared, humidity, month, weekday, hour of day, holiday and cloud cover as input variables he arrived at an R squared value of 0.7382 which meant that approximately 73.8% of the variation in the hourly ridership could be explained by the selected variables and the linear model he used to fit the data.

In this section various linear and non-linear regression models were used to test and train the Trips data that was merged with the weather data to try to predict the number of checkouts based on weather conditions.

The following regression models were used in this study:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Bayesian Ridge Regression
- Decision Tree Regression
- Random Forest Regression
- Extra Trees Regression
- Nearest Neighbors Regression

## Regression Modeling with Categorical Feature Set

The Checkout Month, Week Day and Hour numeric variables were converted to categorical features resulting in 45 total features for regression modeling.

Prior to applying the models a feature correlation was performed on all the features to see if any of the features were highly correlated to one another. As shown in Figure 21, Temperature and Apparent Temperature were highly correlated suggesting that one of them could be removed from the features in the model application.

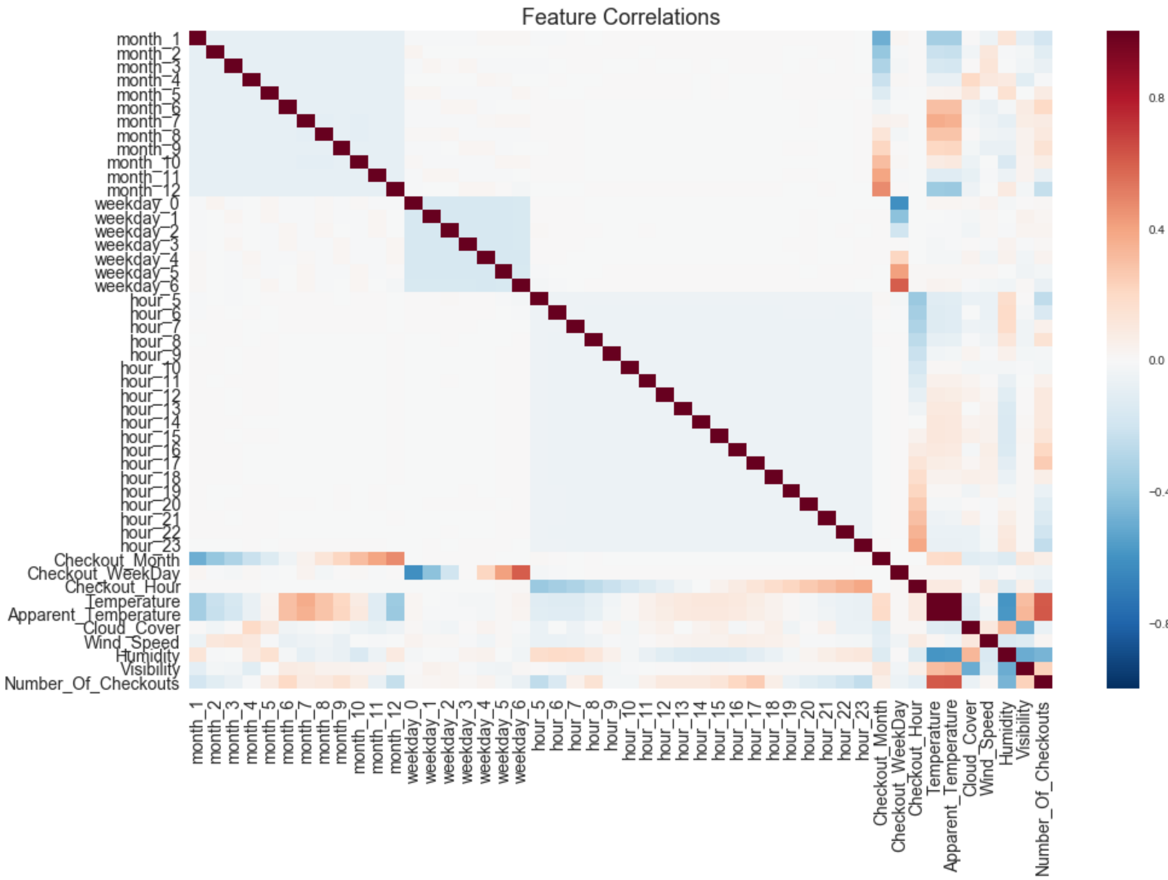



FIGURE 21: FEATURE CORRELATIONS

The models used for regression supported the use of several parameters that could be used to adjust or tune them for better performance. In most cases in this study, the parameters were set to default.

For each model the training and test scores, R Squared and RMSE results were collected and summarized. In addition, the Decision Tree, Random Forest and Extra Trees models also had their Feature Importance bar charts plotted. The chart for Extra Tree model is shown in Figure 22.

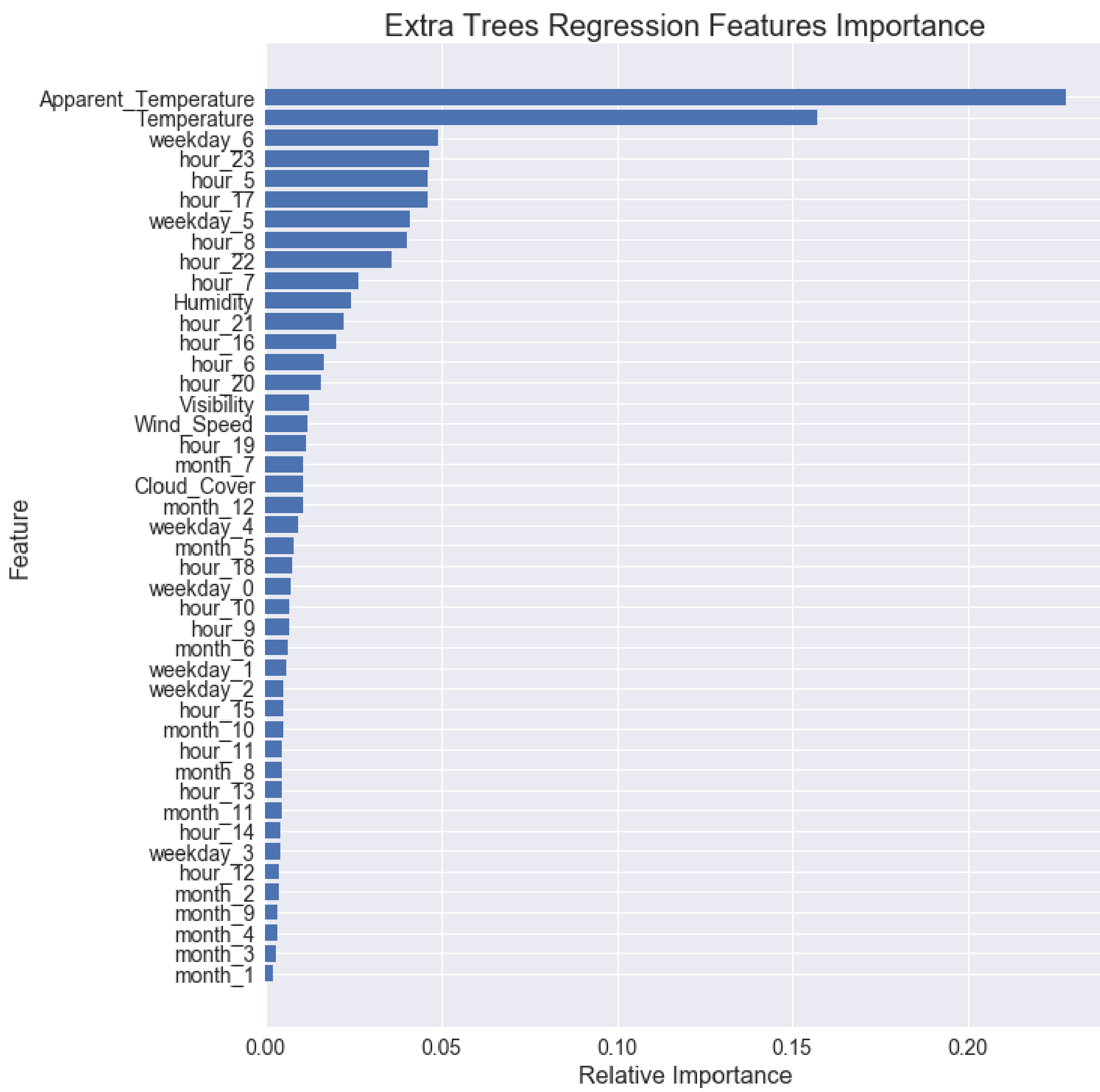



FIGURE 22: EXTRA TREES REGRESSION MODEL FEATURE IMPORTANCE CHART

## Regression Modeling Summary – Categorical Feature Set

| | Linear | Lasso | Ridge | Bayesian Ridge | Decision Tree | Random Forest | Extra Trees | Nearest Neighbors |
|---|---|---|---|---|---|---|---|---|
| Training Test Score | 0.676 | 0.676 | 0.676 | 0.676 | 1.000 | 0.969 | 1.000 | 0.575 |
| Test Set Score | 0.696 | 0.696 | 0.696 | 0.696 | 0.718 | 0.825 | 0.840 | 0.476 |
| R Squared | 0.834519 | 0.834457 | 0.834457 | 0.834448 | 0.847276 | 0.908443 | 0.916278 | 0.690249 |
| RMSE | 627.95439 | 628.16826 | 628.16826 | 628.19832 | 583.57445 | 361.43485 | 331.86035 | 1082.98114 |

The Extra Trees regression model achieved the highest accuracy and the lowest RMSE. All the linear models (Linear, Lasso, Ridge and Bayesian Ridge) had twice the RMSE value of the Extra Trees model.

## Regression Modeling with Numerical Feature Set

Using Checkout Month, Week Day and Hour numeric variables resulted in just 9 total features for regression modeling.

Prior to applying the models a feature correlation was performed on all the features to see if any of the features were highly correlated to one another. As shown in Figure 23, Temperature and Apparent Temperature were highly correlated suggesting that one of them could be removed from the features in the model application.

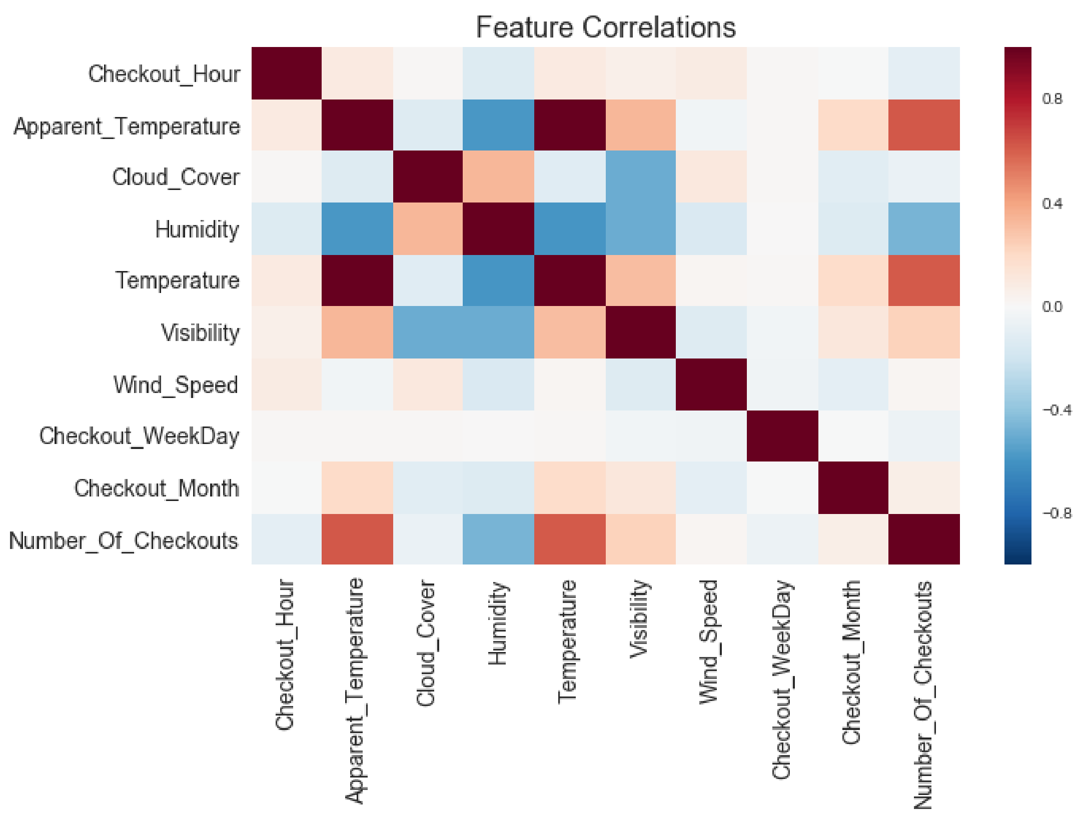



FIGURE 23: FEATURE CORRELATION

For each model the training and test scores, R Squared and RMSE results were collected and summarized. In addition, the Decision Tree, Random Forest and Extra Trees models also had their Feature Importance bar charts plotted. The chart for Extra Tree model is shown in Figure 24.

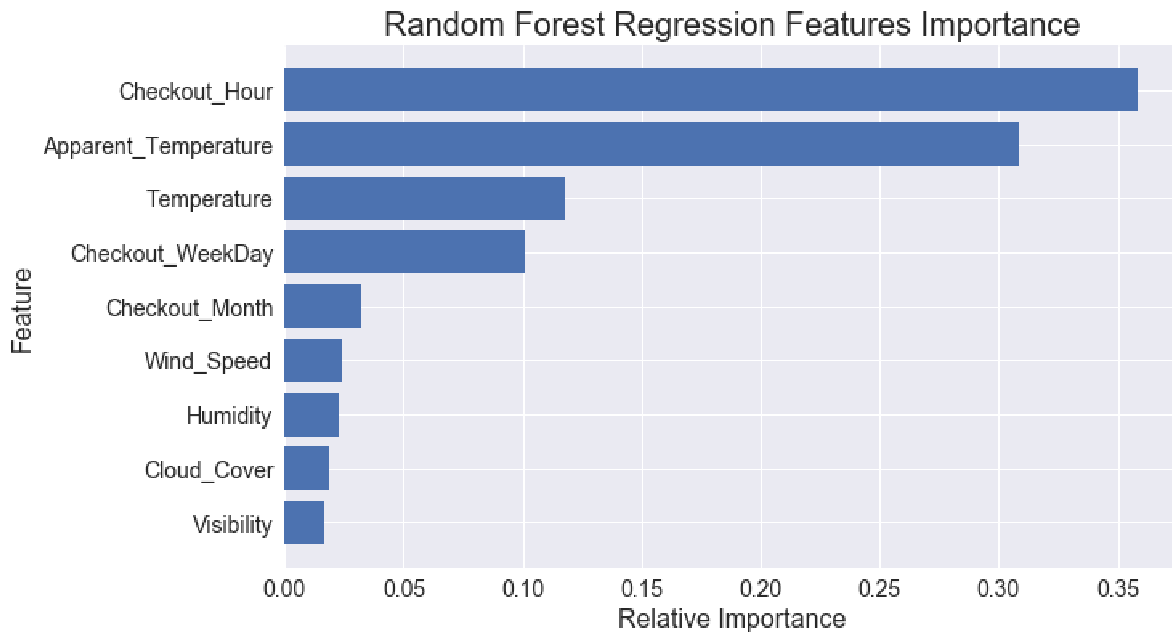



FIGURE 24: RANDOM FOREST REGRESSION MODEL FEATURE IMPORTANCE CHART

## Regression Modeling Summary – Numerical Feature Set

| | Linear | Lasso | Ridge | Bayesian Ridge | Decision Tree | Random Forest | Extra Trees | Nearest Neighbors |
|---|---|---|---|---|---|---|---|---|
| Training Test Score | 0.433 | 0.433 | 0.433 | 0.433 | 1.000 | 0.975 | 1.000 | 0.880 |
| Test Set Score | 0.448 | 0.447 | 0.447 | 0.447 | 0.741 | 0.854 | 0.838 | 0.646 |
| R Squared | 0.669090 | 0.668243 | 0.668243 | 0.668785 | 0.861079 | 0.924077 | 0.915609 | 0.803447 |
| RMSE | 1142.475 | 1144.818 | 1144.818 | 1143.319 | 534.800 | 302.172 | 334.397 | 733.229 |

## Regression Modeling Summary

- The data exploration phase of this study revealed the significance of weather variables on the ridership. The regression modeling phase confirmed this to be accurate. Looking at the feature importance graphs generated by the Extra Trees and Random Forest models, the weather attributes rank the highest.
- The non-linear regression models performed better than the linear models. In particular, even with a reduced feature set, the non-linear models such as the Random Forest and the Extra Trees were the best performers with R Squared values well above 0.9.

# Part 3: Classification Modeling

In this section various classification models were used to test and train the Trips data that was merged with the weather data to try to predict the checkout hour based on weather conditions.

The following classification models were used in this study:

- Linear (Logistic) Classification
- Decision Tree Classification
- Random Forest Classification
- Extra Trees Classification
- Naïve Bayes Classification
- Gradient Boosting Classification
- Nearest Neighbors Classification
- Multi-layer Perceptron Classification

The dataset was randomly spilt into 60% for training and 40% for testing. The class labels were defined as follows:

Class 0: Number of Checkouts >= 1 and <= 50

Class 1: Number of Checkouts >= 51 and <=75

Class 2: Number of Checkouts >=76 and <= 100

Class 3: Number of Checkouts >= 101 and <= 150

Class 4: Number of Checkouts >=151

A cross validation using the Shuffle Split method was performed on the dataset for each model using a training sample size of 50% and a testing sample size of 50% with 10 splits.

## Classification Modeling – Categorical Feature Set

As in the case of Regression modeling, feature correlation was carried out to determine if any features had a high correlation with one another. As shown in Figure 21, Temperature and Apparent Temperature were highly correlated suggesting that one of them could be removed from the features in the model application.

For each model the training and test scores, Accuracy, F1 (micro), F1 (macro), Precision (macro), Precision (micro), Recall (macro) and Recall (micro) results were collected and summarized. In addition, the Decision Tree, Random Forest and Extra Trees models also had their Feature Importance bar charts plotted.

## Classification Modeling Summary – Categorical Feature Set

| | Logistic | Decision Tree | Random Forest | Extra Trees | Naïve Bayes | Nearest Neighbors | Gradient Boosting | Multi-Layer Perceptron |
|---|---|---|---|---|---|---|---|---|
| Accuracy | 0.671898 | 0.639051 | 0.660949 | 0.69080 | 0.360584 | 0.549635 | 0.697080 | 0.713869 |
| F1 (macro) | 0.496528 | 0.524387 | 0.486226 | 0.576138 | 0.290211 | 0.322821 | 0.564894 | 0.556326 |
| F1 (micro) | 0.671898 | 0.639051 | 0.660949 | 0.697080 | 0.360584 | 0.549635 | 0.697080 | 0.713869 |
| Precision (macro) | 0.565525 | 0.524363 | 0.597933 | 0.603859 | 0.355717 | 0.393611 | 0.595219 | 0.620760 |
| Precision (micro) | 0.671898 | 0.639051 | 0.660949 | 0.697080 | 0.360584 | 0.549635 | 0.697080 | 0.713869 |
| Recall (macro) | 0.487240 | 0.524940 | 0.464222 | 0.557927 | 0.408047 | 0.319211 | 0.547159 | 0.550908 |
| Recall (micro) | 0.671898 | 0.639051 | 0.660949 | 0.697080 | 0.360584 | 0.549635 | 0.697080 | 0.713869 |

# Classification Modeling – Numerical Feature Set

Using Checkout Month, Week Day and Hour numeric variables resulted in just 9 total features for regression modeling.

As in the case of Regression modeling, feature correlation was carried out to determine if any features had a high correlation with one another. As shown in Figure 22, Temperature and Apparent Temperature were highly correlated suggesting that one of them could be removed from the features in the model application.

For each model the training and test scores, Accuracy, F1 (micro), F1 (macro), Precision (macro), Precision (micro), Recall (macro) and Recall (micro) results were collected and summarized. In addition, the Decision Tree, Random Forest, Extra Trees and Gradient Boosting models also had their Feature Importance bar charts plotted. The chart for the Gradient Boosting model is shown in Figure 25.

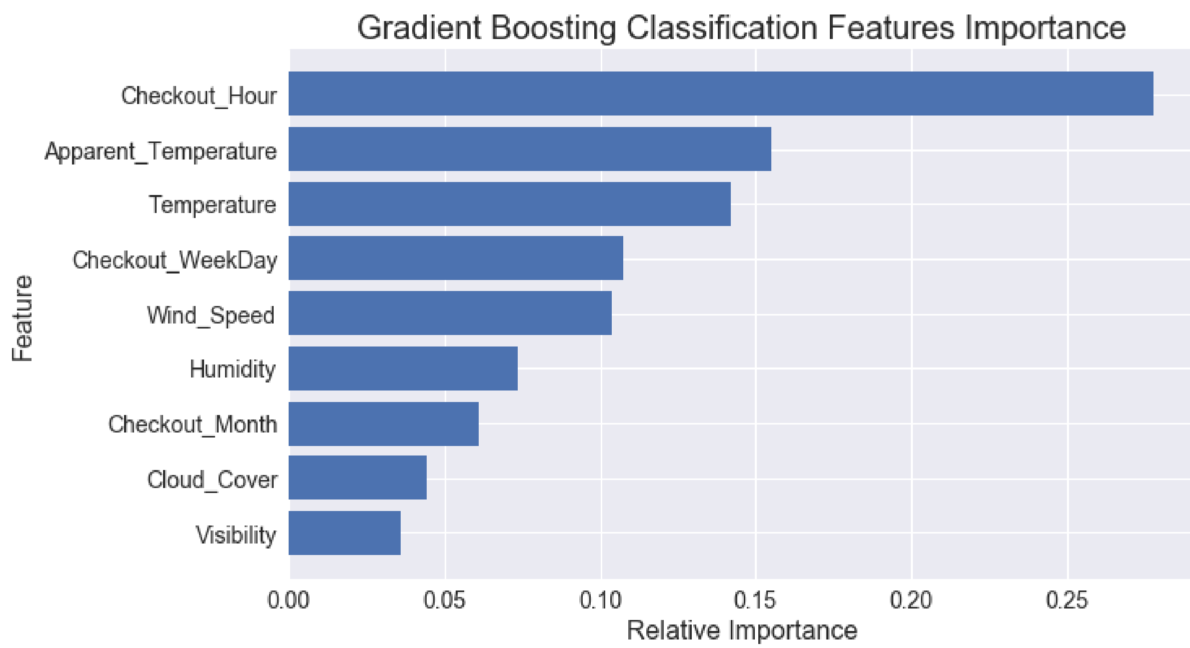



FIGURE 25: GRADIENT BOOSTING CLASSIFICATION MODEL FEATURE IMPORTANCE CHART

## Classification Modeling Summary – Numerical Feature Set

| | Logistic | Decision Tree | Random Forest | Extra Trees | Naïve Bayes | Nearest Neighbors | Gradient Boosting | Multi-Layer Perceptron |
|---|---|---|---|---|---|---|---|---|
| Accuracy | 0.577007 | 0.656934 | 0.670073 | 0.665328 | 0.500365 | 0.589781 | 0.70146 | 0.606569 |
| F1 (macro) | 0.325555 | 0.552091 | 0.50567 | 0.508443 | 0.299163 | 0.447329 | 0.571128 | 0.328199 |
| F1 (micro) | 0.577007 | 0.656934 | 0.670073 | 0.665328 | 0.500365 | 0.589781 | 0.70146 | 0.606569 |
| Precision (macro) | 0.337289 | 0.559901 | 0.5497 | 0.554362 | 0.388027 | 0.449879 | 0.617866 | 0.372159 |
| Precision (micro) | 0.577007 | 0.656934 | 0.670073 | 0.665328 | 0.500365 | 0.589781 | 0.70146 | 0.606569 |
| Recall (macro) | 0.334286 | 0.545906 | 0.489289 | 0.488079 | 0.348306 | 0.445305 | 0.550013 | 0.374967 |
| Recall (micro) | 0.577007 | 0.656934 | 0.670073 | 0.665328 | 0.500365 | 0.589781 | 0.70146 | 0.606569 |

The Gradient Boosting Classifier achieved the highest accuracy and the Naïve Bayes the lowest. While the Multi-Layer Perceptron model had better accuracy than the Gradient Boosting with the categorical feature set it did not fare as well in the numerical feature set.

### Classification Modeling Summary

- The multi-layer perceptron model attained the highest accuracy in classifying the four classes using the categorical feature set. The Naïve Bayes model performed the poorest.
- The Gradient Boosting Classifier achieved the highest accuracy and the Naïve Bayes the lowest with the numerical feature set. While the Multi-Layer Perceptron model had better accuracy than the Gradient Boosting with the categorical feature set it did not fare as well in the numerical feature set.
- None of the models used in this study were not able to achieve an accuracy greater than 71% either with the categorical or the numerical feature set.
- The non-linear regression models performed better than the linear models. In particular, even with a reduced feature set, the non-linear models such as the Random Forest and the Extra Trees were the best performers with R Squared values well above 0.9.

## Summary

This in-depth study on Denver 2016 Bike Share Trips data was undertaken to continue the work that Tyler started on the 2014 data. It agrees with his findings that by merging calendar, clock and weather attributes into the Trips dataset can reveal ridership patterns and allow regression and classification techniques to be applied for prediction purposes.

This study covered three areas:

1. Explored the Trips datasets and visualized the data and provided useful and interesting information.
2. Deployed a variety of supervised machine learning regression models to predict the number of checkouts using calendar, clock and weather attributes.
3. Deployed a variety of supervised machine learning classification models to predict the number of checkouts using calendar, clock and weather attributes.