Springboard Foundations of Data Science Capstone Project Final Report: Predicting Nice Ride MN Bikeshare Volume

Tony Tushar Jr

March 31, 2018

Contents

Executive Summary **2**

Background **3**

Data Wrangling **3**
Datasets and Descriptions . . . 3
Ride Activity Data Preview - Nice_Ride_trip_history_2017_season.csv . . . 3
Station Location Characteristics - Nice_Ride_2017_station_locations.csv . . . 3
2017 Local Climatological Data, Daily Averages . . . 4
Key Identifications . . . 4
Data Preparation Summary . . . 5

Exploratory/Statistical Data Analysis (EDA) **5**
Load Packages, Load Data . . . 5
1. General exploration of dataset: . . . 5
Distribution of trips by month, for 2016-2017 seasons . . . 5
Distribution of trips per day of the week, by account type . . . 6
Distibution of trips per year by account type . . . 8
Summary of trip distance . . . 8
Summary of trip duration in minutes . . . 8
2. Correlation testing for the effects of primary weather variables on bikeshare volume: . . . 9
Effect of average temperature . . . 9
Effect of average wind speed . . . 10
Effect of precipitation . . . 10
Effect of relative humidity . . . 11
3. Welch Two Sample t-tests for measuring variances in average general biking characteristics and primary weather variables for member and casual bike account types: . . . 12
Trip duration . . . 12
Trip distance . . . 13
Temperature . . . 13
Wind Speed . . . 13
Precipitation . . . 14
Relative humidity . . . 14
4. EDA Conclusion . . . 15
Brief summary of findings: . . . 15
Possible further exploration: . . . 15

Predictive Modeling: Machine Learning Application **15**
Modeling Strategies Considered: . . . 15
Main features (predictors) based on EDA results . . . 16
Properties of Model Evaluation . . . 16

Modeling Caveat . . . 16
Load Modeling Data . . . 16
Geospatial Exploration . . . 16
Calculate distance data frame from start station coordinates for clustering and plot on same map 17
Daily ride counts . . . 18
Check dataset structure prior modeling . . . 19
Random Forest Modeling . . . 19
Test model 1 predictive ablity on test dataset . . . 19
Divide by mean of daily ride count(n) for interpretation as percentage of the mean . . . 19
Run model 2 removing tuneGrid component . . . 19
Test model 2 predictive ablity on test dataset . . . 20
Divide by mean of daily ride count(n) for interpretation as percentage of the mean . . . 20
Run model 3 doubling ntree parameter . . . 20
Test model 3 predictive ablity on test dataset . . . 21
Divide by mean of daily ride count(n) for interpretation as percentage of the mean . . . 21
Plot variables in relation to RMSE contribution . . . 22
Conclusion . . . 22

Executive Summary

This project analyzes the Nice Ride MN bike share program 2016-2017 seasons alongside historical daily weather data in order to predict daily ridership. The primary solution this analysis aims to provide - a reliable estimation of daily bikeshare use based on historical behavior and correlations to weather - can apply to the frequent problem of when to redistribute bikes to various docking stations and/or determine maintenance strategies based on user behavior. Addressing these issues can lead to greater operational efficiency, lower costs, and higher revenue for Nice Ride MN.

Questions covered in the analysis:

- How does weather - temperature, precipitation, humidity, and wind speed - affect bike use?
- What are the busiest docking stations and how is this represented from a geospatial perspective?
- How does bike use vary between weekdays and weekends?
- How does use behavior differ between members and nonmembers? Behavioral use questions to include mean, median, mode, and range for ride length, ride distance, and day of the week.

Potential business applications for this research:

- Optimization model for maintaining bike stock at docking stations based on demand forecasting
- Predictive model for volume of bike use based on a combination of weather variables
- Exploration of shifting from flat rate pricing structure to dynamic pricing for revenue optimization
- Predictive model for best future locations of bike docking stations

The project analysis occurs in three stages: (1) data wrangling, (2) exploratory/statistical data analysis, and (3) predictive modeling. Findings from this analysis include:

- Riders with membership subscriptions comprise the majority of daily use, tend to utilize the bikeshare during work commute hours, and have a higher tolerance for unfavorable weather conditions.
- Bikeshare volume density is greatest in downtown Minneapolis and less utilized in St. Paul.
- Bike ride volume is comparable between the two seasons analyzed, however outside factors (road construction) might have contributed to minor difference between the seasons.
- Daily average temperature plays the largest influence on bike use, followed by humidity, precipitation, and wind speed.

Background

Nice Ride MN was formed in 2008 from a city of Minneapolis initiative - Twin Cities Bike Share Project. The initiative reviewed other municipal bike share programs and settled upon a non-profit structure that utilizes both public and private funding. Bike rides began in 2010 with over 100,000 rides on 700 bikes across 65 stations. Since 2015 the system has included over 1700 bikes and 190 stations with annual rides over 450,000. Source: Nice Ride MN - About

Dockless stations are a growing trend in bike share programs across the world, but they present a new set of challenges known as "bike pollution". Dockless stations address the operational inefficiency of bike rebalancing, however they can lead to disrupting other environmental aspects of city life when users do not adhere to dockless rules. The objectives of this research have been asked by others and more of this type of research might lead to increasing the efficiencies of bike stations or assist in determing solutions for the issues arising from dockless programs.

Data Wrangling

Datasets and Descriptions

Nice Ride MN provides annual datasets for all bike rental activity and dock station characteristics. Nice Ride MN

Ride Activity Data Preview - Nice_Ride_trip_history_2017_season.csv

Variable	Description
Start date	Date and time the rental began
Start station	Descriptive name for the station where the rental began, changes if station is moved
Start terminal	Logical name for the station/terminal where the rental began, does not change when station is moved
End date	Date and time the rental ended
End station	Descriptive name for the station where the rental ended
End terminal	Logical name for the station/terminal where the rental ended
Total duration	Total length of the rental, in seconds
Account type	Values are Member or Casual, Members are users who have an account with Nice Ride, Casuals are walk up users who purchased a pass at the station based on half hour increments

Station Location Characteristics - Nice_Ride_2017_station_locations.csv

Variable	Description
Terminal	Logical name of station - matches Start terminal / End terminal in trip history

Variable	Description
Station	Station name used on maps, xml feed and station poster frame - matches Start station / End station in trip history
Latitude	Station location decimal latitude
Longitude	Station location decimal longitude
Nb Docks	Total number of bike docking points at station - indicates station size

Local climatological data are available from the National Centers for Environmental Information's Integrated Surface Data (ISD) dataset. NOAA Weather Data Library

2017 Local Climatological Data, Daily Averages

Variable	Desecription
Station	Station identification number
Station name	Name of station
Elevation	Station elevation
Latitude	Latitude of station
Longitude	Longitude of station
Date	Date of recorded observations
Report type	Reporting method characteristics
Average daily dry bulb temp F	Dry bulb measured temperature in degrees Fahrenheit
Average daily relative humidity	Humidity level
Average daily wind speed	Wind speed in miles per hour
Average daily precipitation	Precipitation in inches

Key Identifications

Dataset Strengths By joining the three datasets described above, a rich dataset offers the opportunity to understand public bike share behavior for two differing price structures - memberships and casual rides. Exploration of the data shows differing use behavior for casual and member customers based on day of the week, time of day and responses to weather scenarios. Analyzing bike use behavior and correlations to weather scenarios allows for insights related to optimal maintenance scheduling and any potential price restructuring for strategies related to revenue and growth.

Dataset Limitations/Caveats Limitations of the dataset include the effects of historical city construction projects. The data include two bike seasons, 2016-2017, and span a duration from early April to early November. During exploratory data analysis a trend of greater biking volume was noticed for 2017 in comparison to 2016. Based on research and potentially backed by domain knowledge, there was a greater volume of city construction projects in 2016 compared to 2017, this might be an uncontrollable variable in the analysis.

There are over 800,000 bike ride observations in this dataset. Through this analysis outliers were uncovered for roughly 1500 observations for the trip duration variables. They represent less than one percent of the total observations and were removed prior conducting statistical tests.

Data Preparation Summary

The Nice Ride MN bikeshare datasets required little data wrangling other than renaming a few columns based on preference, formatting the date and time columns to match with the weather data, and a full join to match the dock station data with the trip history data. Joining bikeshare volume with daily weather averages provided to be a more challenging task. Initially the data was formatted based on hourly weather and riding observations, however, the hourly weather data provided multiple observations per hour causing excessive noise. It was decided that a more reliable and consistent dataset could be created using daily weather averages for joining with ride observations.

Exploratory/Statistical Data Analysis (EDA)

Load Packages, Load Data

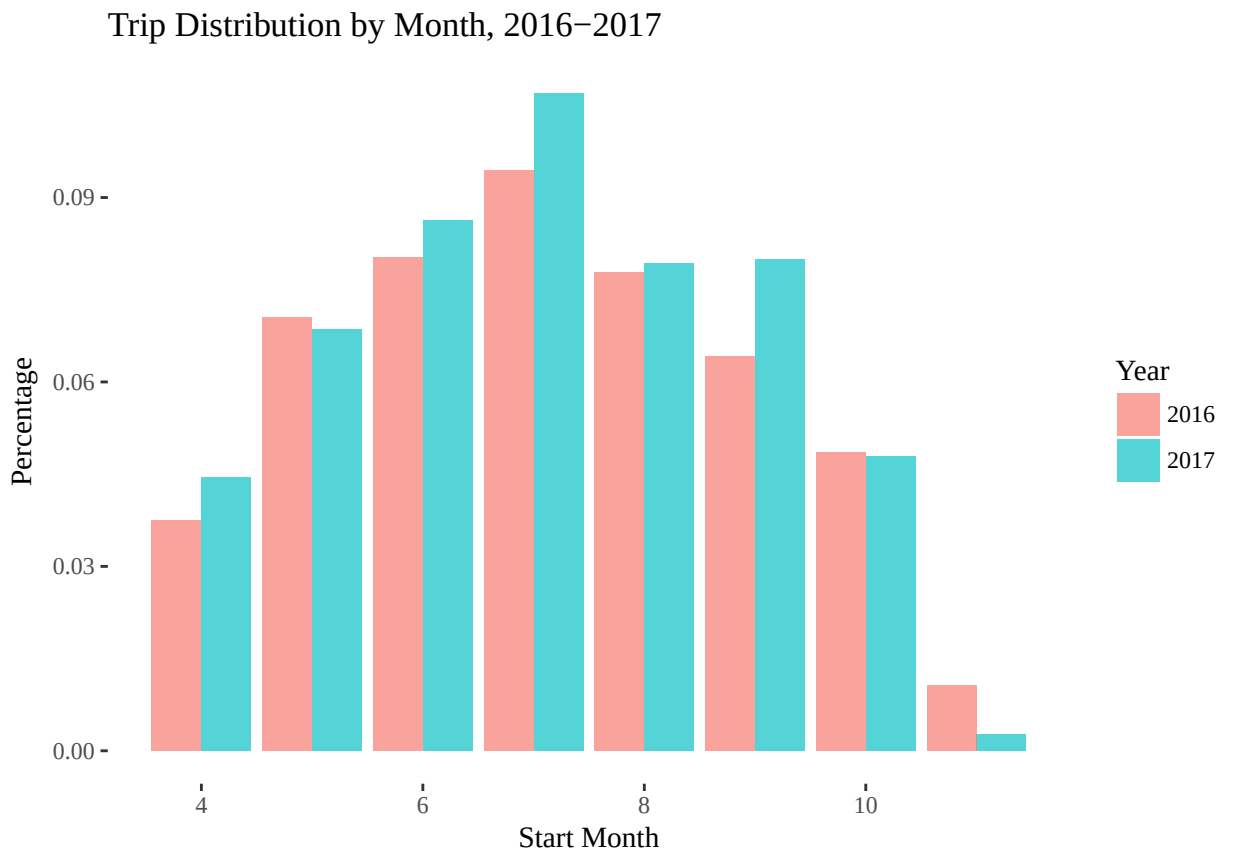
```
head(Rides1617)
```

```
## # A tibble: 6 x 30
##   Start_DoWeek Weekend Start_Year Start_Month Start_Day Start_Station
##   <chr>          <int>      <int>       <int>     <int> <chr>
## 1 Mon                0       2016           4         4 4th Street & 17th~
## 2 Mon                0       2016           4         4 Washington & Cedar
## 3 Mon                0       2016           4         4 Sanford Hall
## 4 Mon                0       2016           4         4 Social Sciences
## 5 Mon                0       2016           4         4 S 8th Street & 2n~
## 6 Mon                0       2016           4         4 Loring Park
## # ... with 24 more variables: Start_Latitude <dbl>, Start_Longitude <dbl>,
## #   Start_Docks <int>, End_Year <int>, End_Month <int>, End_Day <int>,
## #   End_Station <chr>, End_Latitude <dbl>, End_Longitude <dbl>,
## #   End_Docks <int>, Total_Duration_Seconds <int>,
## #   Total_DurationMin <dbl>, Trip_DistanceMiles <dbl>, Account_Type <chr>,
## #   Avg_Temp <int>, Avg_Wind <dbl>, Precip <dbl>, Snow <int>,
## #   Rel_Humidity <int>, Fog <int>, Heavy_Fog <int>, Thunder <int>,
## #   Hail <int>, Haze <int>
```

1. General exploration of dataset:

Distribution of trips by month, for 2016-2017 seasons

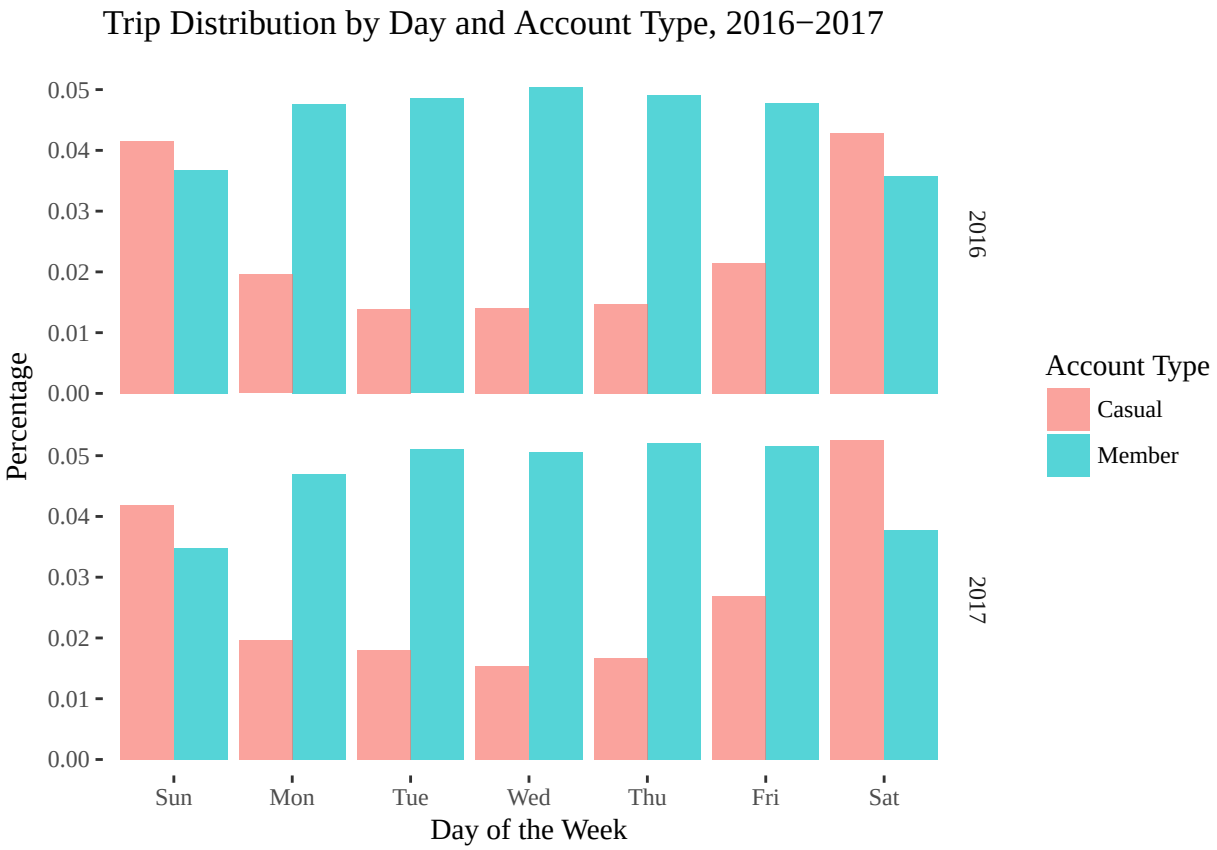
```
## # A tibble: 6 x 4
##   Start_Year Start_Month     n Percent_Year
##        <int>       <int> <int>        <dbl>
## 1       2016           4 33447       0.0375
## 2       2016           5 62985       0.0705
## 3       2016           6 71716       0.0803
## 4       2016           7 84324       0.0944
## 5       2016           8 69542       0.0779
## 6       2016           9 57301       0.0642
```



Trip distrubution shows greatest variance year-over-year during the summer months. It is possible road construction played a larger role in 2016 than 2017, impacting ride volume negatively. However, it is difficult to control for road construction from one year to the next, this is simply a hypothesis to note but not investigate in the scope of this project.

Distribution of trips per day of the week, by account type

```
## # A tibble: 6 x 5
##   Start_Year Start_DoWeek Account_Type     n Percent_Year
##        <int> <ord>        <chr>        <int>        <dbl>
## 1       2016 Sun          Casual       37048       0.0415
## 2       2016 Sun          Member       32820       0.0368
## 3       2016 Mon          Casual       17486       0.0196
## 4       2016 Mon          Member       42558       0.0477
## 5       2016 Tue          Casual       12422       0.0139
## 6       2016 Tue          Member       43336       0.0485
```



Daily distribution by year confirms consistency year-over-year, member riders are utilizing bikshare for commuting purposes while casual riders utilize bikeshare for leisure, inverse of one another. An hourly breakdown of bike volume would further detail this observation, however, this is not relevant to the greater scope of our project purpose.

Distibution of trips per year by account type

Member rides dominant casual year-over-year. Both bike seasons demonstrate a convex shape visually, ride volume peaks in the summer months. It is worth noting that 2017 demonstrates more frequent peak volume points for casual riders above members, perhaps these max points are related to holidays or city-wide events.

Summary of trip distance

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.047   1.944   2.523   3.324  29.046
```

Median trip distance is 1.944 miles while the max is 29 miles, is this logical in comparison to trip duration?

Summary of trip duration in minutes

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      1.00      6.80     12.72     34.81     24.50 280043.68
```

It is odd to see a max ride observation of over 280,000 minutes, this equates to 194 days! How many bike rides are greater than one day? This is the process by which we can determine outliers as previously mentioned.

```
# What percentage of bike rides are over one day in length?
# Trips equal to or less than one day
Rides_Day <- Rides1617 %>% mutate(Total_DurationDay = (Total_DurationMin <= 24*60))

Day_Length <- nrow(Rides_Day)
```

```r
sum(Rides_Day$Total_DurationDay/Day_Length)
```

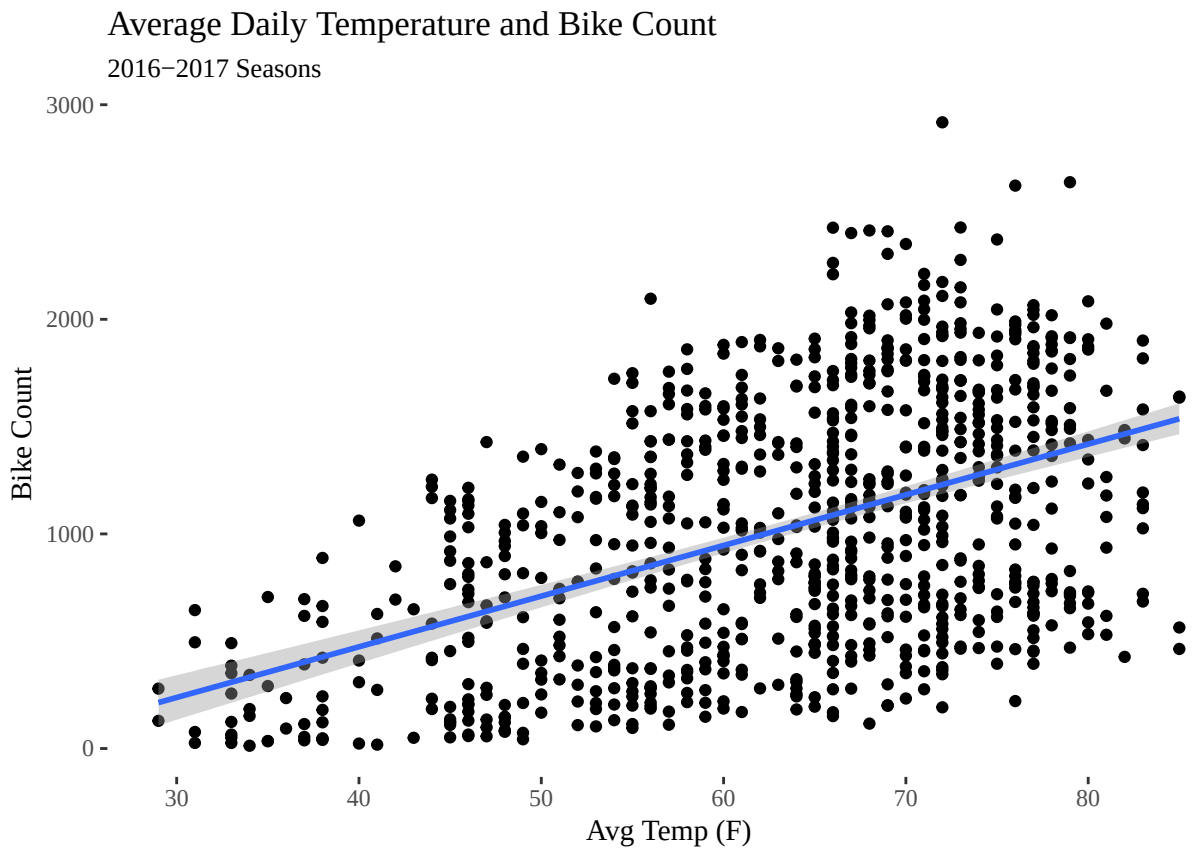
```
## [1] 0.9983225
```

There are 1,498 observations in which ride duration is greater than one day, this accounts for less than 0.02% of our dataset. As our intended outcome of this study is to create a predictive model for daily trips as related to daily weather, we consider these observations as outliers and remove them.

2. Correlation testing for the effects of primary weather variables on bikeshare volume:

Effect of average temperature

```
##
##  Pearson's product-moment correlation
##
## data:  Temp$Avg_Temp and Temp$n
## t = 15.861, df = 866, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4212174 0.5244458
## sample estimates:
##       cor
## 0.4744612
```

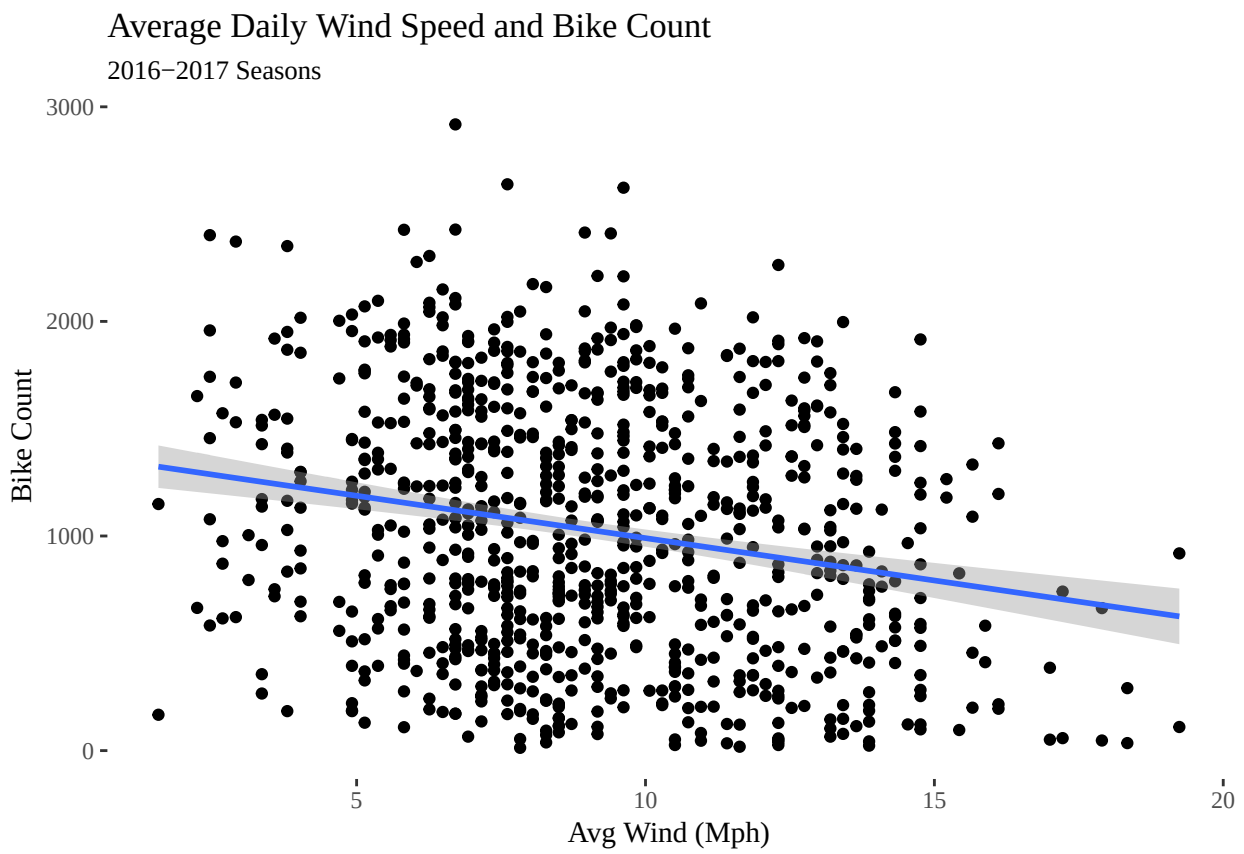


Null hypothesis: there is no correlation between average daily temperature and bike volume. We reject the

null hypothesis based on the statistical significance of a p value less than 0.05 and a correlation coefficient of 0.47. Heteroskedasticity is also visible from this plot, the variance in bike counts increases as temperature increases. This will have to be reconsidered when in the modeling phase of the analysis.

Effect of average wind speed

```
##
##  Pearson's product-moment correlation
##
## data:  Wind$Avg_Wind and Wind$n
## t = -6.3699, df = 866, p-value = 3.067e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2742415 -0.1470882
## sample estimates:
##        cor
## -0.2115599
```

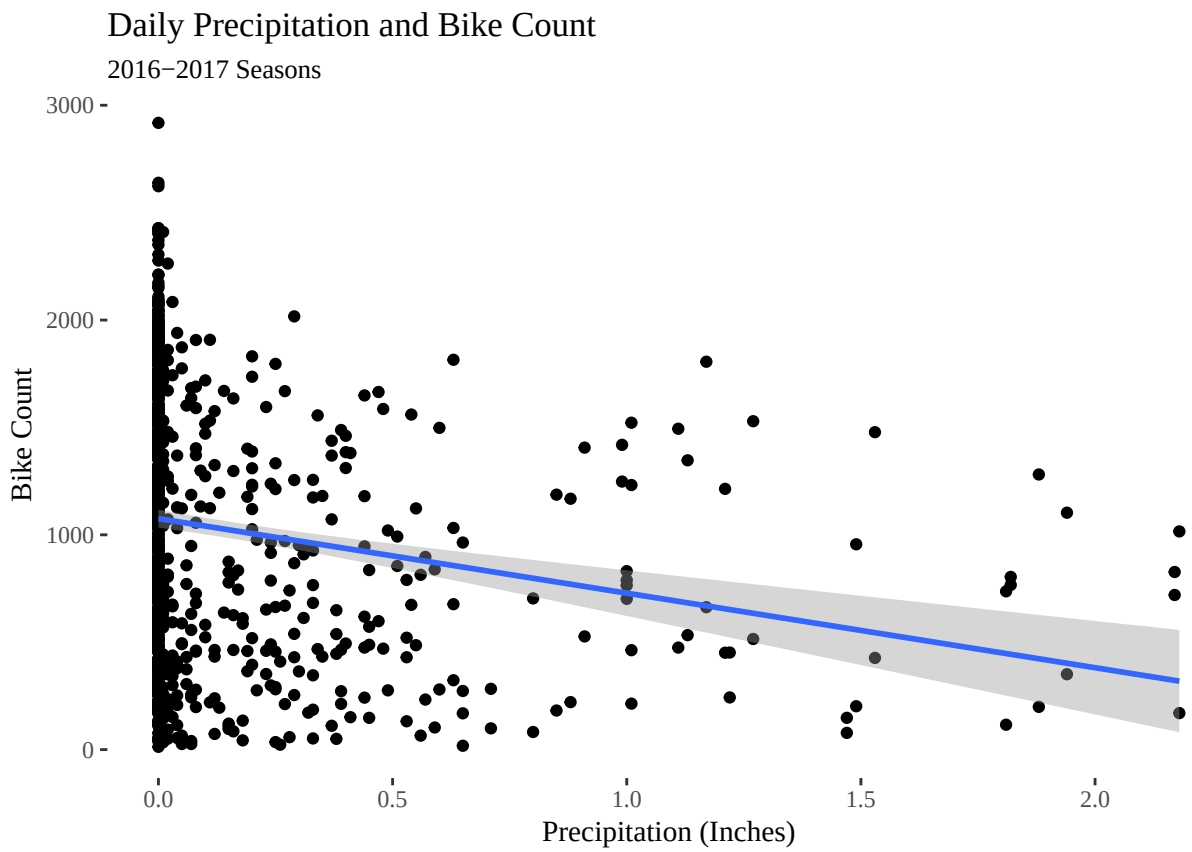


Null hypothesis: there is no correlation of average daily wind to daily bike count. We reject the null hypothesis based on the statistical significance of a p value less than 0.05 and a correlation coefficient of -0.21.

Effect of precipitation

```
##
##  Pearson's product-moment correlation
##
```

```
## data:  Precip$Precip and Precip$n
## t = -5.9179, df = 866, p-value = 4.695e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2602785 -0.1323447
## sample estimates:
##       cor
## -0.1971508
```



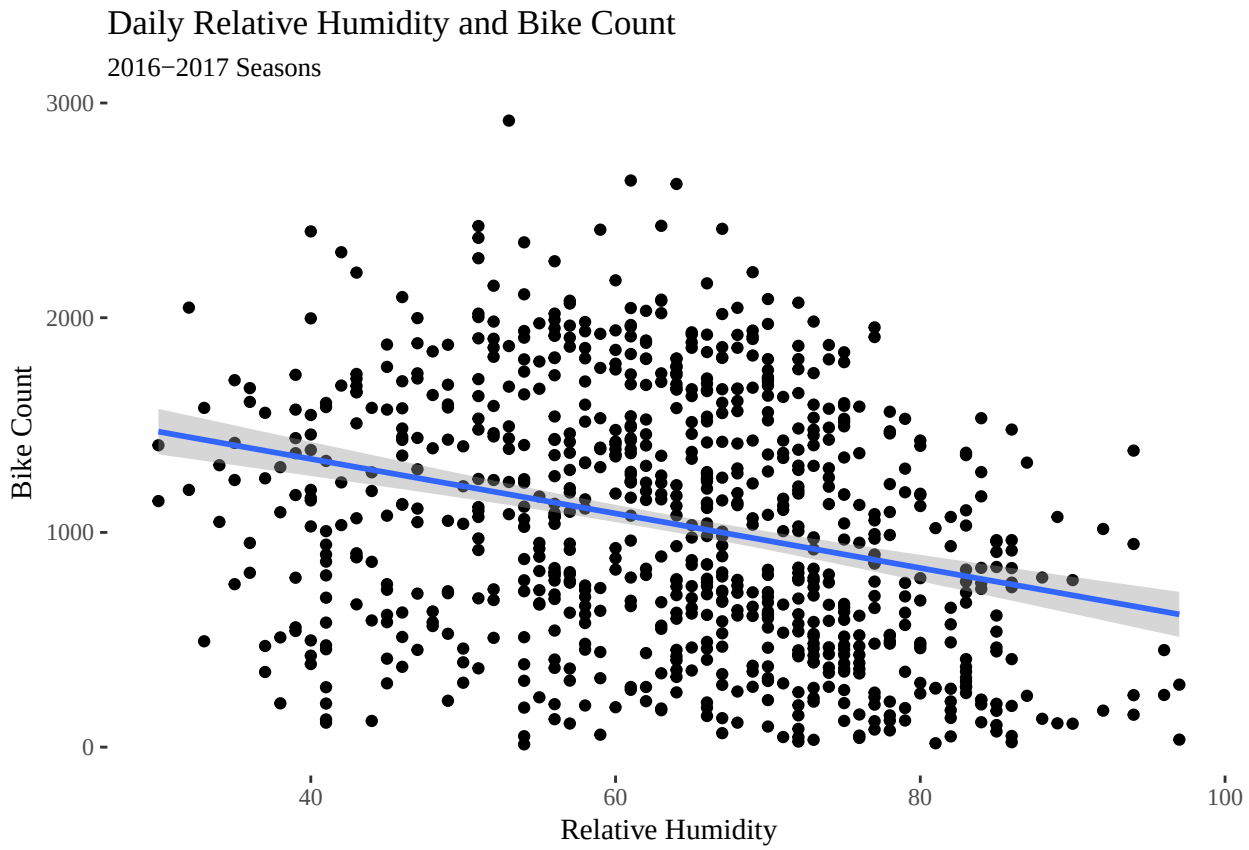
Null hypothesis: There is no correlation between precipitation and bike volume. We reject the null hypothesis based on the statistical significance of a p value less than 0.05 and an inverse correlation coefficient of -0.20.

Effect of relative humidity

```
##
##  Pearson's product-moment correlation
##
## data:  Humidity$Rel_Humidity and Humidity$n
## t = -8.5128, df = 816, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3473439 -0.2213840
## sample estimates:
##       cor
## -0.2855969
```

```
## Warning: Removed 50 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 50 rows containing missing values (geom_point).
```



Null hypothesis: There is no correlation between relative humidity and bike volume. We reject the null hypothesis based on the statistical significance of a p value less than 0.05 and an inverse correlation coefficient of -0.29.

3. Welch Two Sample t-tests for measuring variances in average general biking characteristics and primary weather variables for member and casual bike account types:

Trip duration

```
##
##  Welch Two Sample t-test
##
## data:  Total_DurationMin by Account_Type
## t = 155.89, df = 362350, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  22.78977 23.37012
## sample estimates:
## mean in group Casual mean in group Member
##             37.31497             14.23503
```

Null hypothesis (H0): There is no difference in the mean trip duration of member and casual account types. Alternative hypothesis (H1): There is a difference in the mean trip duration of member and casual account

types. In this case, we reject the null hypothesis as the probability of there being no difference in mean trip duration for account types is very, very small. *On average, casual bike riders favor bike trips longer than members.*

Trip distance

```
##
##  Welch Two Sample t-test
##
## data:  Trip_DistanceMiles by Account_Type
## t = 3.5525, df = 578700, p-value = 0.0003817
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.008584695 0.029715775
## sample estimates:
## mean in group Casual mean in group Member
##             2.532404             2.513254
```

Null hypothesis (H0): There is no difference in the mean trip distance of member and casual account types. Alternative hypothesis (H1): There is a difference in the mean trip distance of member and casual account types. In this case, we reject the null hypothesis as the probability of there being no difference in mean trip distance for account types is very, very small. *On average, casual bike riders favor a slightly longer riding distance.*

Temperature

```
##
##  Welch Two Sample t-test
##
## data:  Avg_Temp by Account_Type
## t = 87.543, df = 747400, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.777764 1.859190
## sample estimates:
## mean in group Casual mean in group Member
##             67.82874             66.01027
```

Null hypothesis (H0): There is no difference in the mean temperature of member and casual account types. Alternative hypothesis (H1): There is a difference in the mean temperatures of member and casual account types. In this case, we reject the null hypothesis as the probability of there being no difference in mean temperature for account types is very, very small. *On average, casual bike riders favor a slightly warmer average temperature than member riders.*

Wind Speed

```
##
##  Welch Two Sample t-test
##
## data:  Avg_Wind by Account_Type
## t = -29.747, df = 664330, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.2109829 -0.1848990
## sample estimates:
## mean in group Casual mean in group Member
##             8.553975             8.751916
```

Null hypothesis (H0): There is no difference in the mean wind speed of member and casual account types. Alternative hypothesis (H1): There is a difference in the mean wind speed of member and casual account types. In this case, we reject the null hypothesis as the probability of there being no difference in mean riding distance for account types is very, very small. *On average, casual bike riders favor a slightly lower wind speed for riding compared to member riders.*

Precipitation

```
##
##  Welch Two Sample t-test
##
## data:  Precip by Account_Type
## t = -49.317, df = 747490, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03113073 -0.02875090
## sample estimates:
## mean in group Casual mean in group Member
##           0.08238605           0.11232687
```

Null hypothesis (H0): There is no difference in the mean precipiation of member and casual account type rides. Alternative hypothesis (H1): There is a difference in the mean precipiation of member and casual account type rides. In this case, we reject the null hypothesis as the probability of there being no difference in mean precipitation for account types is very, very small. *On average, casual bike riders favor less precipitation while riding compared to member riders.*

Relative humidity

```
##
##  Welch Two Sample t-test
##
## data:  Rel_Humidity by Account_Type
## t = -86.829, df = 657360, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.433958 -2.326502
## sample estimates:
## mean in group Casual mean in group Member
##             60.05355             62.43378
```

Null hypothesis (H0): There is no difference in the mean relative humidity of member and casual account type rides. Alternative hypothesis (H1): There is a difference in the mean relative humidity of member and casual account type rides. In this case, we reject the null hypothesis as the probability of there being no difference in mean relative humidity for account types is very, very small. *On average, casual bike riders favor lower relative humidity while riding compared to member riders.*

4. EDA Conclusion

Brief summary of findings:

All t-tests resulted in rejecting the null hypothesis as there was little probabilty for a difference of zero between member and casual account type ride characteristics in relation to the average observations of trip distance, trip duration, and primary weather variables. Running t-tests to compare trip distance and duration between the two groups, we determined a high probabilty that casual rider trip duration is on average over twice as long as member trip duration. As for trip distance, we reject the null hypothesis that there is no difference in trip distance between account types. However this was our weakest p value for all t-tests conducted. While there is a difference in average trip distance by account type, the difference is minimal.

In comparison to casual riders, member riders appear more willing to handle average weather variables in the following manner:

- Lower temperatures
- Higher wind speed
- Higher precipitation
- Higher relative humidty

Possible further exploration:

We have not analyzed the secondary, binomial weather variables that detail various weather types as "yes" or "no" for a given ride observation. This study did include propability tests comparing the likelihood that weather types such as fog, heavy fog, thunder, hail, and haze; in which we found that probabilities for bike rides by members were higher for these weather conditions present than for casual riders. We will explore the impact of these categorical independent variables in greater detail when conducting regression analysis and weighting the variables for machine learning applications.

Predictive Modeling: Machine Learning Application

Modeling Strategies Considered:

The main question for this project is how does a given daily weather scenario affect the bike share use of casual and member riders? Framing this as a machine learning problem involves modeling the relationship between weather and bike riders via historical data in order to apply the modeling toward future weather scenarios in the next riding season.

We start the predictive modeling under the premise that this problem is supervised and in the form of a regression. Secondly, we consider hierarchial clustering as a preliminary step knowing that daily bike ride density varies throughout the system. It might be best to cluster stations based on the mean distance between all stations and run a regression for each cluster. Initial attempts of this approach are likely to yield poor results due to the previously observed heteroskedasticity of the weather variables in relation to daily counts.

A third approach is considered: random forests. Applying random forests to the total dataset while including the cluster group variable might provide the best outcome under the current scope of this project. Random forest models draw random samples from the data set with replacement and utilize a regression tree approach in place of the linear form. The model reaches an optimal regression tree based on te dominant outcome from the numerous random sampling. Utilizing the train function from the 'caret' package, the model chooses an optimal number of predictor variables in the dataset for each tree, resulting in the highest Rsquared predictive power while accounting for model overfit.

Main features (predictors) based on EDA results

For the first process of clustering, the important predictors are geospatial data points - latitude and longitude, clusters of bike share stations related to a centroid point, based on the distance properties gleaned from the geospatial coordinates.

For the secondary phase of random forests, in addition to the cluster variable, time variables along with the four main daily weather variables and ride account type are considered:

- Year, Month, and Day
- Day of the Week
- Account Type
- Temperature
- Wind Speed
- Precipitation
- Humidity

Properties of Model Evaluation

A random sampling of 80% of the data set is utilized for training the model and the remainder is used to test the model. Adjusted Rsquared and the Root Mean Squared Error (RMSE) will serve as measurements of the accuracy and efficiency of the random forest regression tree modeling. Maximizing Rsquared and minimizing the RMSE in proportion to the dependent variable mean will result in a reliable model for predicting daily bike volume.

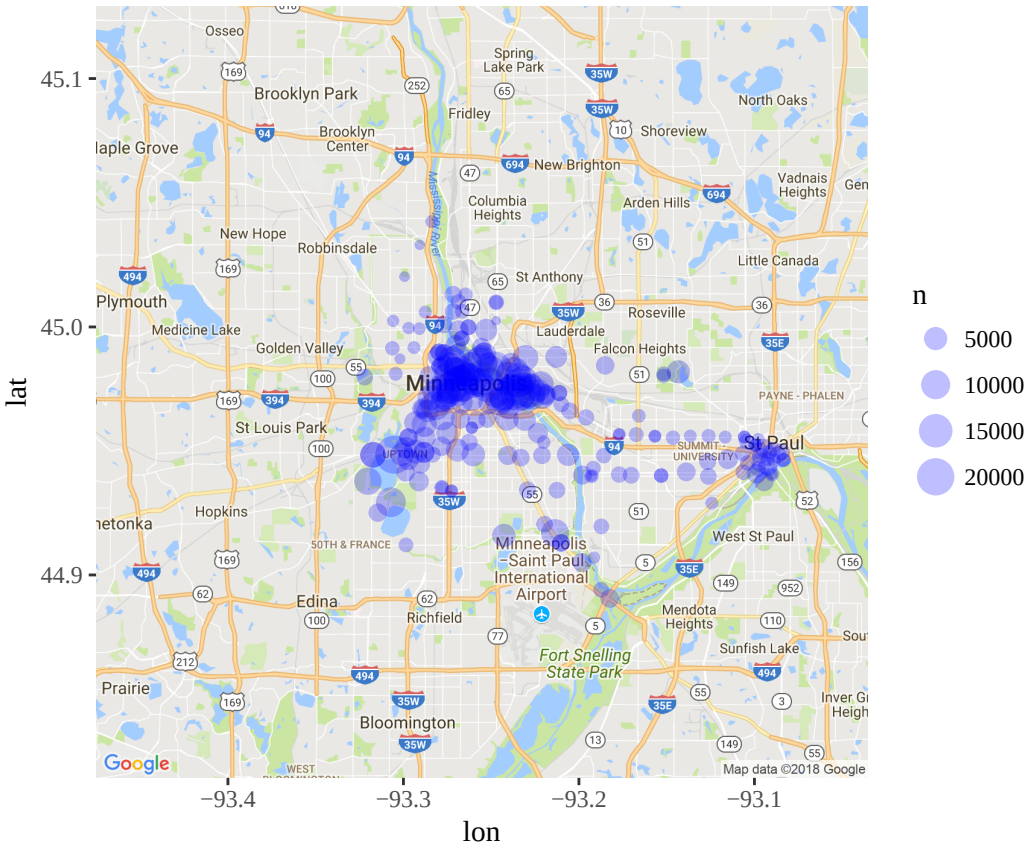
Modeling Caveat

This phase of the analysis began with the intention to produce a model for each bike station cluster, however, due to time constraints and the reality that a model for the total dataset serves as a satisfactory process in itself, the clustering serves as a feature rather than modeling divergent.

Load Modeling Data

Geospatial Exploration

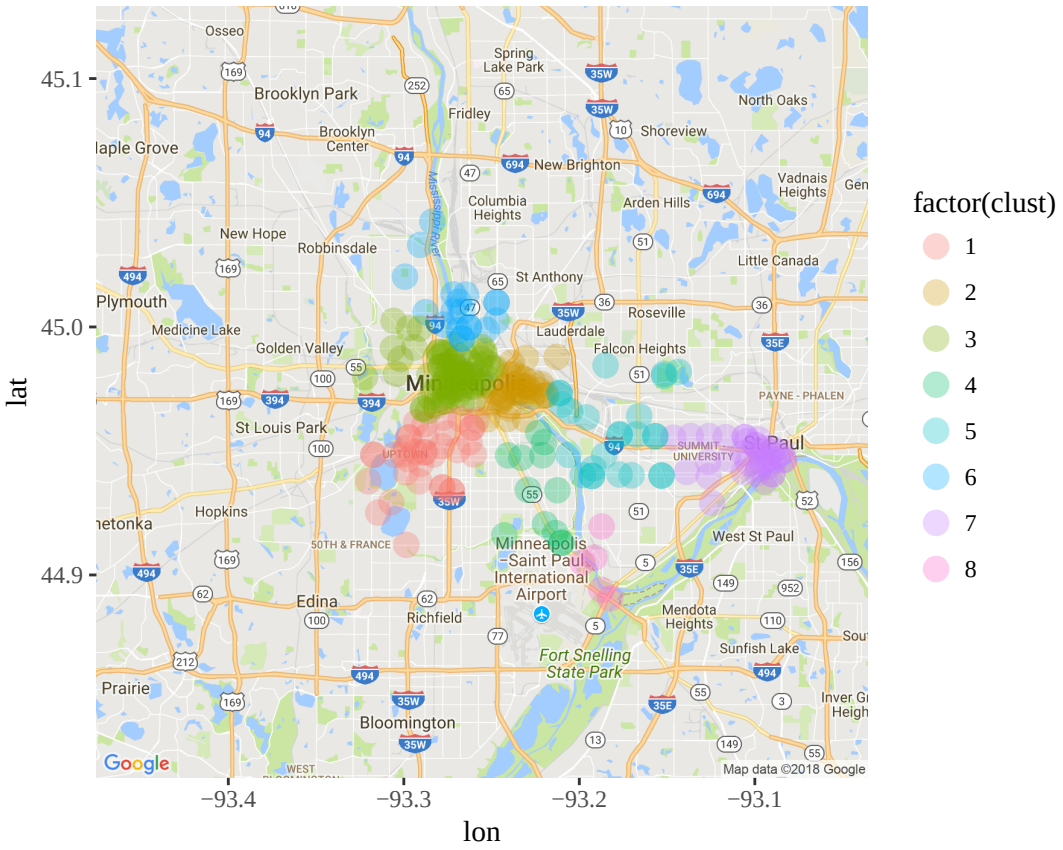
Observe the busiest starting stations for Minneapolis-St.Paul in order to estimate k number of clusters for additional variable.



Reviewing the density of bike station use on the city map it is clear that the majority of activity occurs in Minneapolis while most activity occurs in St. Paul along the most popular residential streets and downtown. We estimate the optimal clustering somewhere between 6-10 clusters.

Calculate distance data frame from start station coordinates for clustering and plot on same map

Utilizing the geocoordinates for each bike station we can calculate a piecewise vector for the distance between every possible combination of bike stations. We determine a possible number of clusters from this output by applying the mean distance between station combinations as our cutoff point.



Clustering the stations based on the mean distance between stations resulted in eight clusters.

We will use the cluster station ID's for joining the clustering to our original dataset. There is a possible argument that the cluster variable should not be included in the modeling. We argue for its inclusion based on the idea that clusters with greater daily bike count density carry higher weight in the modeling process. High density stations are more closely laid out in the city and are likely to be areas of greater concern in daily operations as they harbor the most use.

Daily ride counts

Add a column for count(n) per day by account type - member or casual. This is our dependent variable for modeling. We will clean up the data set prior modeling by dropping all variables except for clustering, data/time, account type, daily weather averages, and daily counts.

```
## # A tibble: 6 x 11
## # Groups:   clust [3]
##   clust Start_DoWeek Start_Year Start_Month Start_Day Account_Type
##   <int> <chr>             <int>       <int>     <int> <chr>
## 1     1 Mon                2016           4         4 Casual
## 2     1 Mon                2016           4         4 Member
## 3     2 Mon                2016           4         4 Casual
## 4     2 Mon                2016           4         4 Member
## 5     3 Mon                2016           4         4 Casual
## 6     3 Mon                2016           4         4 Member
## # ... with 5 more variables: Avg_Temp <int>, Avg_Wind <dbl>, Precip <dbl>,
## #   Rel_Humidity <dbl>, n <int>
```

Check dataset structure prior modeling

This step ensures factor variables are correct.

Random Forest Modeling

We will run a random forest model on 80% of the NiceRide.Counts data and test for prediction accuracy with the remaining 20%:

```
## [1] "This took 32.2 seconds"
```

```
## Random Forest
##
## 5437 samples
##   10 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 5437, 5437, 5437, 5437, 5437, 5437, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   108.5483  0.7864496  68.13897
##
## Tuning parameter 'mtry' was held constant at a value of 2
```

Model 1 outcome has an Rsquared of 0.79 and RMSE of 108.5. A decent first model, however, the RMSE will be more interpreable if made proportional to the mean daily ride count (n). We will make this conversion for the prediction outcome from our test dataset.

Test model 1 predictive ablity on test dataset

```
## [1] 100.495
```

The Root Mean Squared Error is ~100 bike rides per day, convert this to a percentage of the mean for (n) to better interpret model accuracy

Divide by mean of daily ride count(n) for interpretation as percentage of the mean

```
## [1] 0.8017891
```

This is a poor outcome if error rate accounts for 80% of the test set mean. Re-run model and allow 'caret' package free roaming for determining how many columns to include per tree (mtry):

Run model 2 removing tuneGrid component

```
## [1] "This took 358.2 seconds"
```

```
## Random Forest
##
## 5437 samples
##   10 predictor
##
## No pre-processing
```

```
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 5437, 5437, 5437, 5437, 5437, 5437, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    2    108.05501  0.7913703  67.79845
##   11     58.09365  0.9084784  30.24016
##   21     59.40336  0.9041902  30.52005
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 11.
```

Model 2 produced a much better outcome since caret was allowed to determine the mtry floor automatically. After a similar review of the model predictability, lets consider model 3 with double the ntree parameter.

Test model 2 predictive ablity on test dataset

```
## [1] 48.25269
```

```
## [1] -0.51985
```

Allowing 'caret' to find the optimal mtry value has reduced the RMSE by 52%, checking RMSE in proportion to mean for NR.test$n:

Divide by mean of daily ride count(n) for interpretation as percentage of the mean

```
## [1] 0.384979
```

The result for RMSE.2 is much smaller in proportion to the mean for NR.test$n, running one final model with 100 trees instead of 50:

Run model 3 doubling ntree parameter

```
## [1] "This took 709.9 seconds"
```

```
## Random Forest
##
## 5437 samples
##   10 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 5437, 5437, 5437, 5437, 5437, 5437, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    2    107.48863  0.7990741  67.49143
##   11     57.66042  0.9099156  29.97403
##   21     59.19493  0.9048651  30.42767
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 11.
```

Test model 3 predictive ablity on test dataset

```
## [1] 48.67531
```

```
## [1] 0.0087584
```

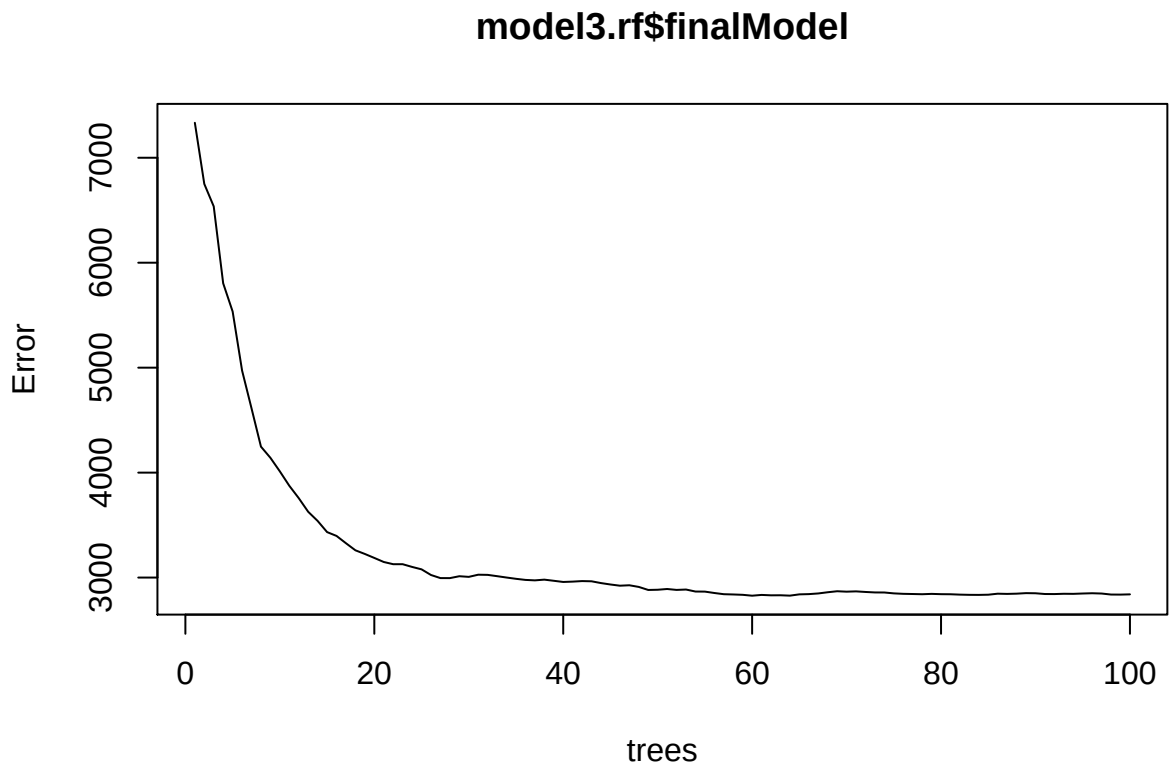
Model 3 provided only a slight improvement to Rsquared at mtry = 11 and the proportional difference of RMSE.3 to RMSE.2 is actually just under 1%. Not much improvement for more processing time.

Divide by mean of daily ride count(n) for interpretation as percentage of the mean

```
## [1] 0.3883508
```

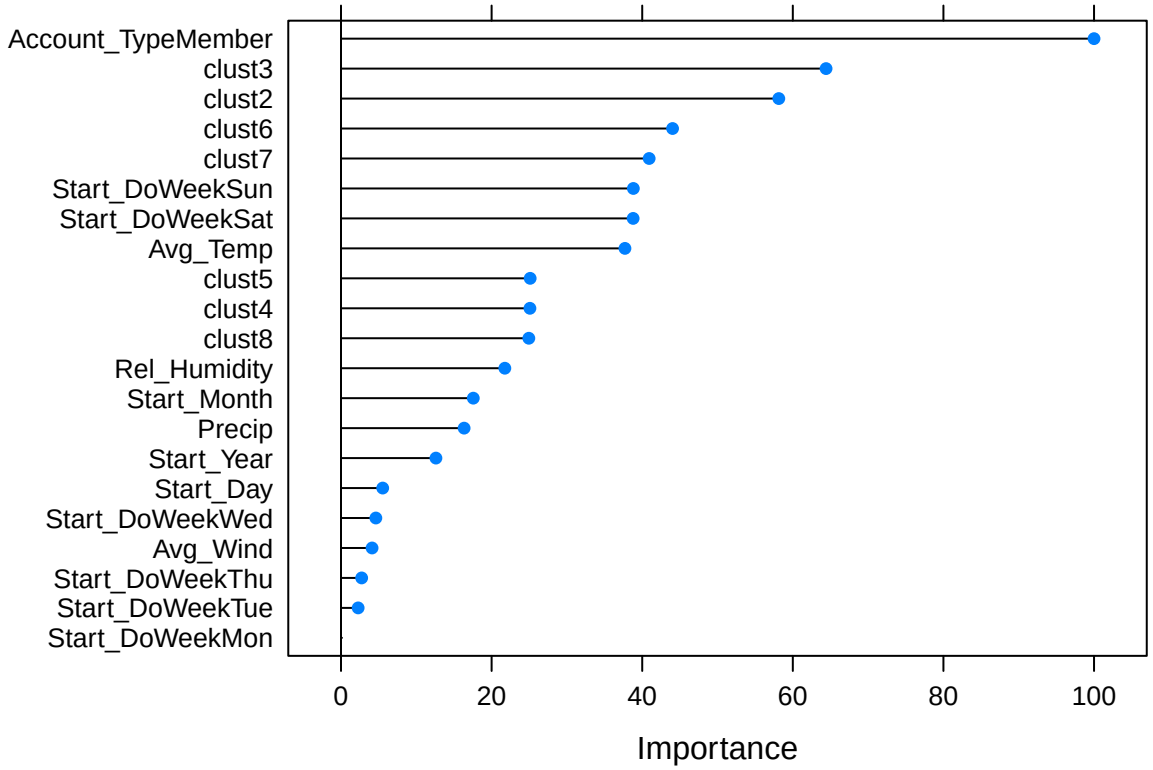
A sligthly less desireable RMSE proportion to the mean for model 3 compeared to model 2.

We will consider model 2 our final model and plot the error rate in relation to the number of trees:



This plot confirms that not much improvement occurs when ntree = 100 vs. ntree = 50.

Plot variables in relation to RMSE contribution



Variable importance plotting might allude to a limitation of the random forest approach when conducted on the entire data set rather than subsets based on clustering. The train function is giving importance to the cluster groups that outweighs the input for the weather variables.

Conclusion

The machine learning application for this business problem originally began with the intention of clustering the dataset based on the mean distance between bike stations and applying a linear regression to each cluster. The approach shifted to random forests due to observing heteroskedasticity from the relationship of the weather variables to the dependent daily ride count. The heteroskedasticity caused poor regression modeling. The random forest model with 50 trees and mtry = 11 provided a model with a much higher Rsquared (.90 in place of ~0.70). However, it might be ideal to apply eight separate models to data subsets based on the clustering to give higher consideration to the weather variables and their influence on daily ride counts within each cluster rather than as a whole. Modeling on this level is beyond the scope of this current application and can be considered in future analysis.

Regardless, this analysis has provided several important insights:

Rider behavior shows statistically significant differences which ought to be considered for any planned re-evaluation of bike ride subscription types and pricing, especially in lieu of RFP's for dockless systems. Temperature plays a large role in daily bike count volume. Nice Ride MN might consider promotional strategies around changes in temperature, humidity, precipitation, and wind speed. This model could be use to prepare a dashboard interface for predicting daily ridership based on various weather variables combinations and thresholds. From this data product Nice Ride MN promotional efforts could be planned, providing discounted riding during certain parts of the season. *More work would need to be done with this analysis to

provide insight and/or support for daily operational activities such as bike station rebalancing. This is the biggest limitation of this analysis, running models on each cluster might improve the outcome.