Exploring Washington D.C. Bikeshare Ridership Behavior

A Generalized Estimating Equation (GEE) Approach

Alex McCreight
Yiyang Shi
Zixuan Zheng

May 6th, 2022

ABSTRACT

The increasing popularity of bike sharing solves the last-mile problem for the commuters in cities by deploying bike sharing stations at bus stops, metro stations, and libraries. Increasing gas prices, maintenance costs, and traffic jams are obstacles for commuters that travel by car. Bike sharing services are a cost-effective and eco-friendly alternative. This paper will discuss the influence and efficiency of the bike sharing system in DC by evaluating Capital Bikeshare stations' popularity with a longitudinal approach in 2021. We utilized a generalized estimating equation (GEE) to model the average daily rides per station by incorporating weather information, demographic characteristics, and spatial elements under the assumption of an exchangeable working correlation structure. We conclude that the temperature and distance to the Washington Monument are negatively correlated. At the same time, low humidity, low wind speeds, and the proportion of White residents per census tract are positively correlated with the number of daily riders.

INTRODUCTION

Bike sharing is a cost-effective and eco-friendly alternative to transportation. Capital Bikeshare is a company that provides bike sharing services in areas including Washington D.C., Arlington, Alexandria, Montgomery County, Prince George's County, Fairfax County, and Falls Church. Riders must pay $1 to unlock a bike and $0.05 per minute for a standard bike, and $0.15 per minute for an ebike. This service provides riders with a relatively cheap substitute for a short travel distance trip compared to other transportation such as taxis and ride-sharing services. Moreover, bike sharing systems also reduce greenhouse gas emissions. A 2020 study found that bike sharing systems across eight different cities reduced annual greenhouse gas emissions from 41 to 5417 tons of CO2-eq (Kou et al. 2020).

As commuters begin to resume their pre-pandemic routines, Washington D.C. transportation and industry officials have begun to expand the amount of Capital Bikeshare stations across the city (Lazo 2021). Officials believe that micro-mobility systems, such as bike sharing services, are critical to fulfilling Washington D.C.'s transportation needs. On October 25, 2021, D.C. Mayor Muriel Bowser and Lyft announced a 30-day free Capital Bikeshare membership for all D.C. Residents. Bowser hopes to ease travel disruptions caused by reduced Metrorail service. Throughout Bowser's time as mayor, she has created 80 new Capital Bikeshare stations, added 2,500 new e-bikes to the Capital Bikeshare inventory, and plans to create 30 miles of protected bike lanes over the next three years ("Mayor Bowser and Lyft Announce Free Capital Bikeshare Memberships for All DC Residents" 2021).

From a fiscal and policy-driven perspective, we are interested in investigating: 1) what factors account for the station-to-station difference in popularity, 2) how weather affects a rider's consumer behavior, and 3) how do spatial and demographic factors influence ridership?

DATA SET

In order to examine the bike sharing station-to-station differences in Washington D.C., we utilized data from the Capital Bikeshare company website (Motivate International, n.d.). Their database contains information about each bike trip, such as the trip date, the starting station name and location, and the membership type of the rider. While the Capital Bikeshare website has data dating back to 2010, we wanted to use 2021 to see how the Bikeshare industry recovered from the 2020 COVID-19 pandemic.

In addition to the bike sharing data, we also included daily 2021 weather data from the Weather Underground website ("DC Weather History," n.d.). We wanted to examine how meteorological predictors change consumer behavior patterns. These predictors include temperature (Fº), dew point (Fº), wind speed (mph), humidity (%), and precipitation (in). Furthermore, we included

federal holiday information for 2021 as we were curious to see if holidays affect consumer habits.

Finally, we merged demographic information collected by the US Census Bureau to our data set from the 'tidycensus' R package. This data set includes information on Washington D.C. about census tracts' median income, population, median age, race information, and location data. Unfortunately, census data is only available every ten years, so we use census data from 2020 to match our bike sharing data set.

DATA CLEANING

To begin cleaning our data, we wanted to add a couple of variables that we felt would be pertinent for our analysis. We first started out by adding a distance variable using the 'geosphere' R package for the bike sharing data. This variable calculated the distance between the start station and the end station of each ride using the Haversine formula. This formula determines the great-circle distance between two points on a sphere. However, one concern about this method is that it might not give us the accurate distance of each trip. Each bike rider can have multiple routes for this trip based on the traffic and viable road constructed. These estimates give us the shortest possible distance for each trip. We then added a ride duration predictor using the 'lubridate' R package. This allowed us to extract the time portion of the start and end date variables and then subtract the difference between the two to get the total duration in minutes. According to Capital Bikeshare company, they filtered out bike trips lasting less than 60 seconds due to the potential of false starts or users trying to re-dock a bike to ensure it is secure.

We sorted our data by start date and the ID of start stations. The Capital bikeshare stations are scattered over the DMV (DC, Maryland, Virginia) area, but we limited our focus to stations located in Washington D.C. Moreover, we did not have constant stations for our daily data since some stations were not included for some of the dates. This is because we only included stations with at least one ride per day, which caused the disappearance of some unpopular stations for days when they had no riders. After we wrangled our data set to be daily data, we took the average duration and distance for each station. Even though we lost specific information for each trip, the average duration and distance gave us a more general idea of bike trips by station.

To account for spatial correlation, we first extracted information about how close each census tract in Washington D.C. is from the closest interstate. We then calculated the distance from each station to its closest highway by using the spherical distance between two geometries in the 'sf' R package. We are also interested in investigating whether having tourist attractions nearby will affect the popularity of the bike stations. The National Mall is one of the busiest areas in Washington D.C. as it contains multiple tourist attractions such as the Washington Monument, the Lincoln Memorial, the World War II Memorial, the United States Holocaust Memorial

Museum, the Smithsonian National Museum of Natural History, and many other popular tourist destinations. To simplify our analysis, we calculated the distance between each station and the Washington Monument as it was the most central location in the National Mall.

We did all data cleaning, wrangling, and modeling in R Studio. Detailed steps can be retrieved from our GitHub.

VARIABLES OF INTEREST

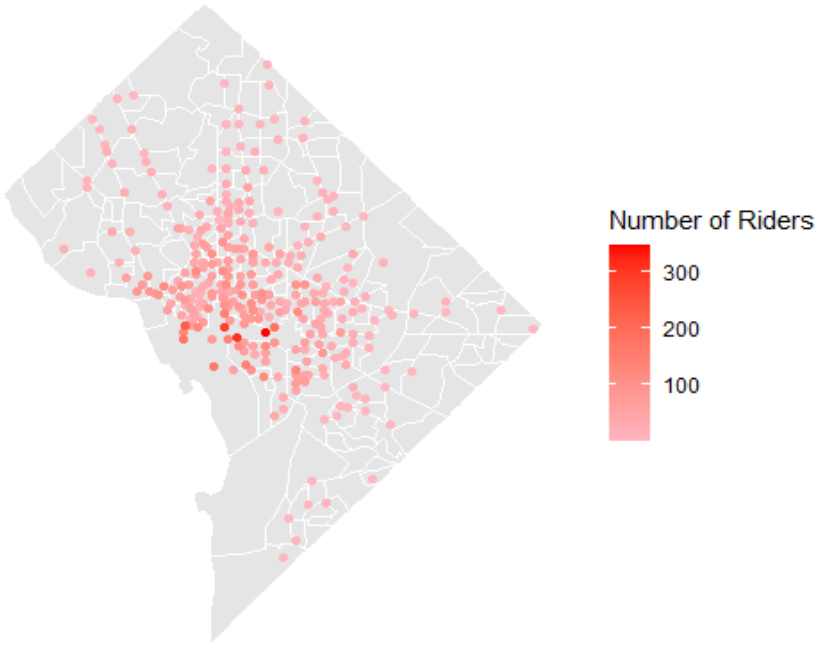


Figure 1: Total Rides from each station on July 4th, 2021

For our research, we focused on modeling each station's daily number of rides. The plot above shows each station's total number of rides on July 4th, 2021. Intuitively, we believe there is a higher probability of more tourists traveling to Washington D.C. on Independence day. This is why we chose to present July 4th because it is a national holiday and a Sunday, which makes this day have higher total ridership than other dates. According to figure 1, a few stations near the National Mall had over 300 riders on Independence day, whereas other stations further away do not have as many daily riders. So, we know that consumers are more likely to use the bike sharing systems around landmarks. One possible explanation for this is that parking lots might be difficult to locate, and public transportation might be tricky for travelers to use.

After our initial exploratory data analysis, we chose three predictors from the weather data to include in our model: temperature, wind speed, and humidity. We speculate that as temperature increases, people are more likely to participate in outdoor activities until it gets too hot; then, the total number of riders will probably diminish. As humidity increases, we speculate that people are more likely to stay inside, especially during the summer, as Washington D.C. has its highest

humidity during August. Finally, we postulate that higher wind speeds will decrease the number of rides as higher wind speeds might make it difficult for riders to control the direction of their bike.

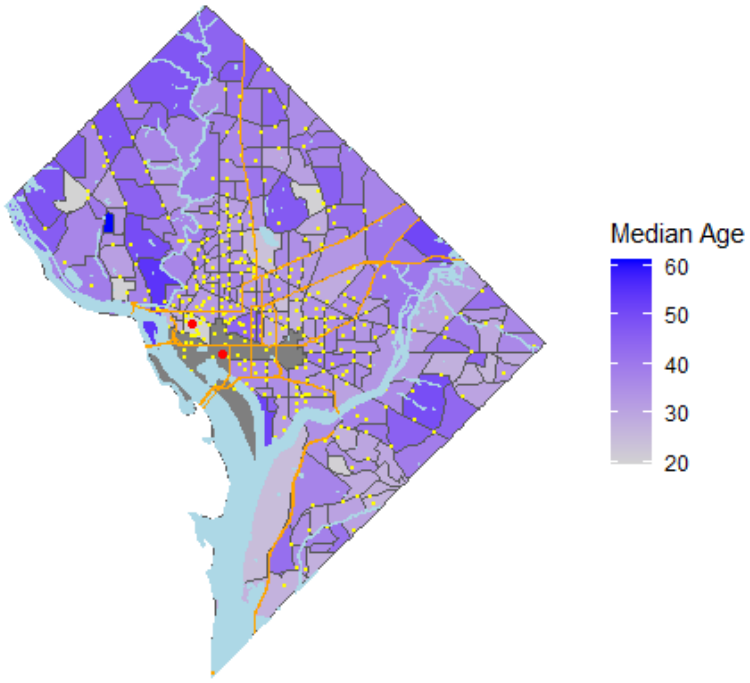


Figure 2: Median Age across Washington D.C. Census Tract

Besides weather information, we also consider the spatial information extracted from the 2020 census tract of DC. The graph above demonstrates the median age distribution, major highways (marked with orange), rivers (marked with light blue), bike stations (marked with yellow), and points of interest (marked with red). The red point on the left is George Washington University. We assume that there will be many riders in this area because most college students do not own cars or prefer not to drive in DC due to high gas prices, traffic, and parking fees. When we examine figure 2, we find that George Washington University has the largest number of stations. The surrounding area is one of the most popular areas in DC as it has many restaurants, bars, and hotels. The red point on the right is the Washington Monument. Even though the bike stations are not as condensed compared to George Washington University, there are still plenty of daily riders in that area since it is a popular tourist destination. Finally, the census tract from the northwest, especially the area close to Maryland, has the highest median household income, and a relatively high median age, but contains relatively few bike stations. We speculate this is due to residents of this census tract having more disposable income to afford cars.

STATISTICAL METHODS AND MODEL SELECTION PROCESS

Our data set is longitudinal as we have daily repeated measurements for each station in 2021. Since we have repeated measurements for each station and the stations are geographically close, we cannot assume that our observations are independent. One of the critical assumptions of an ordinary least squares (OLS) model is that the cases are independent of each other. Since the cases for our dataset are not independent, an OLS model would produce inaccurate standard errors; however, an OLS model would still have accurate slope coefficients (Heggeseth 2022). So, we will utilize a generalized estimating equations (GEE) model to account for this within unit variation.

Unlike a mixed effect model, a GEE model uses a sandwich robust estimator to attain robust standard errors. These robust standard errors are valid even if we incorrectly assume our correlation structure. So this will allow us to tune the correlation structure of our model until our standard errors and robust standard errors match, allowing for a more accurate final model. Predictions and conclusions of mixed effect models are only valid given the correct assumptions on fixed effect, random effects, and error. After we tried different working correlation structures for our model, such as AR1, independence, and exchangeable, we found that the model gives us the smallest difference in magnitude between the robust standard error and standard error under an exchangeable correlation structure.

RESULTS

After evaluating many different variable combinations and correlation structures, we created our final model:

Number of rides ~ f(Temperature, Humidity, Wind, Race, DistanceToWM)

Our outcome, *number of rides*, measures the total number of daily rides per station. *Temperature* is a binary categorical predictor showing how the daily number of riders changes for stations with daily temperatures above/at or below 65 degrees. *Humidity* is another binary categorical predictor showing how the daily number of riders changes for stations with daily humidities above/at or below 75%. *Wind* is our final binary categorical predictor that shows how the daily number of riders changes for stations with daily wind speeds that are either above/at or below 11 mph. *Age* is the median age of a station's census tract. *DistanceToWM* is the shortest distance measured in meters between the starting station and the Washington Monument.

Table 1: Model Coefficient Summary

	Estimates	Model SE	Robust SE	Wald	p-value
(Intercept)	22.140	2.808	2.494	8.877	0
Temperature.Low	-8.450	0.069	0.380	-22.230	0
Humidity.Low	4.362	0.082	0.196	22.300	0
Wind.Low	2.209	0.094	0.116	19.070	0
Race.White	13.530	2.945	2.437	5.550	0
DistanceToWM	-0.003	0.000	0.000	-9.339	0

By interpreting the slope coefficients of our model, we found some interesting trends in our data set. For our temperature variable, we found that holding all other variables constant, the average number of daily riders per station will decrease by 8.450 when the daily average temperature is below 65 degrees Fahrenheit compared to when the daily average temperature is at or above 65 degrees Fahrenheit. For our humidity variable, we found that holding all other variables constant, the average number of daily riders per station will increase by 4.362 when the daily average humidity is below 75% compared to when the daily average humidity is at/above 75%. For our wind speed variable, we found that holding all other variables constant, the average number of daily riders per station will increase by 2.209 when the daily average wind speed is below 11 mph compared to when the daily average wind speed is at/above 11 mph. For a 1 percent point change in proportion of white people in a station's census tract, the average daily number of riders per station increases by 13.530. Holding all other variables constant, for 1 meter increase in the distance from station to the Washington monument, the average daily number of riders per station decreased by 0.003.

Now, we will examine the validity of these slope coefficients and determine if we properly assumed our correlation structure. All of our slope coefficients have statistically significant p-values when utilizing a significance threshold of $\alpha = 0.05$. Our standard errors and robust standard errors are quite similar. Our categorical temperature and proportion of white residents variables have the largest absolute difference in standard error and robust standard errors at 0.311 and 0.508 respectively. The other variables all have less than a 0.2 difference in magnitude between the standard errors and robust standard errors. Since the standard errors and robust standard errors are so similar in magnitude, it is reasonable to assume that using an exchangeable correlation structure is close to modeling the correlations correctly.

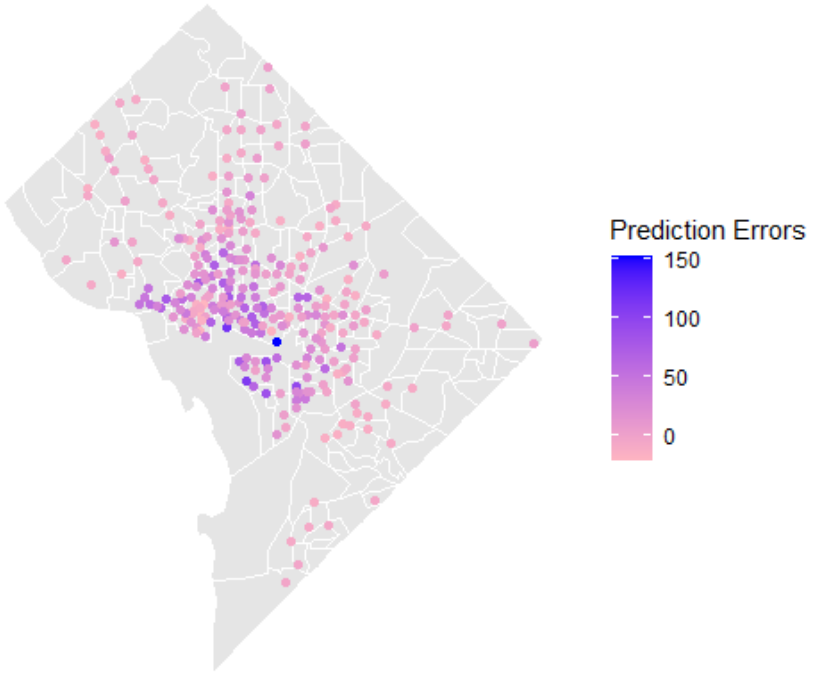


Figure 3: Model Residual Map on July 4th, 2021

We chose to examine the residual plot of our model predictions on Independence day, which is a relatively popular day of using the bike sharing system in Washington D.C.. In our plot, we found out that our model has a fairly good prediction accuracy on most of the stations for this specific day. However, the stations that are at the center of our graph tend to have a greater prediction error compared with other stations. This suggests that our model has some limitations, which we will discuss in the next section. However, we only show one day's residual plot, the performance of our model might vary due to different days.

LIMITATIONS

After comparing many models with different combinations of predictors and correlation structures, we arrived at our final model. Initially, we thought that distance to George Washington University would significantly affect the number of rides from each station. For travelers, using the bike sharing system will be an efficient way to explore the city, and for college students, it might be a cost-effective way to commute to school. However, we found that the distance to the University is, in fact, not statistically significant in our model.

The majority of stations with the highest residuals are located close to each other, suggesting that our model fails to account for the spatial correlation between stations. However, our residual plot only looks at independence day, the day with the highest number of riders in 2021. We focused on a single day as our model estimates the change in daily ridership instead of total ridership. So, more residual plots might be needed for us to justify the accuracy of our model.

Moreover, our spatial predictors such as median income, median age, and proportion of white residents in one census tract area tend to be highly correlated. For example, census tracts with a higher proportion of white residents or older residents tend to have more income than other census tracts. Additionally, older people and people with more money are less likely to use bike sharing services as they are more likely to have a car. So, in order to reduce the collinearity among race, income, and age, we decided only to include race in our model.

Another limitation of our model is that there is a chance for our predicted outcomes for a station to be negative. We can avoid this problem by changing our outcome to a log scale, but if we did this, we could not include stations in our model that had zero rides for a specific day. We also did not include any interaction effects in our model. For instance, the interaction between the proportion of white residents in the area and the median age might affect the number of daily riders.

CONCLUSION

After selecting predictors that might influence our outcome variable, *number of rides*, we used a GEE model with an exchangeable correlation structure to investigate the riders' behavior in Washington D.C.. We also utilized a residual plot to show the accuracy of our model's predictions. We found some interesting observations through our results. Our model predicted the number of rides would decrease as the stations were farther away from the Washington Monument. This suggests that tourists who like to sightsee landmarks in Washington D.C. might be a leading consumer group for the Capital Bikeshare Company. The Capital Bikeshare Company can also consider adding more bikes at these locations and creating promotions for travelers to maximize their profit.

However, we also found that some of our results do not match our intuition. For example, when we first examined the median age across D.C. census tracts, we thought that the number of daily riders would decrease as the median age of a census tract increased. Younger people are less likely to have a car and more willing to try new transportation methods than older people. Nevertheless, when we tried to implement this predictor into our model, we found it statistically insignificant. Also, we initially believed that a higher median income of a census tract would decrease the daily number of riders, but this was also found to be statistically insignificant.

Our model's primary limitation is that it does not explain the spatial correlation between geographically close. In the future, we could consider using spatial models such as a simultaneous autoregressive (SAR) or a conditional autoregressive (CAR) model to investigate consumer ridership behavior in the D.C. area. We could also extend the amount of time to our

model to a 5 year or even a 10 year period to incorporate more data into our model and see how ridership has changed over a longer period of time.

References

"DC Weather History." n.d. *Weather Underground*. https://www.wunderground.com/history/monthly/us/va/arlington/KDCA.

Kou, Zhaoyu, Xi Wang, Shun Fung (Anthony) Chiu, and Hua Cai. 2020. "Quantifying Greenhouse Gas Emissions Reduction from Bike Share Systems: A Model Considering Real-World Trips and Transportation Mode Choice Patterns." *Resources, Conservation and Recycling* 153: 104534. https://doi.org/https://doi.org/10.1016/j.resconrec.2019.104534.

Lazo, Luz. 2021. "Capital Bikeshare Gears up for Expansion as Commuters Resume Pre-Pandemic Routines." *The Washington Post*. WP Company. https://www.washingtonpost.com/transportation/2021/07/01/electric-bike-dc-capital-bikeshare/.

"Mayor Bowser and Lyft Announce Free Capital Bikeshare Memberships for All DC Residents." 2021. *Mayormb*. https://mayor.dc.gov/release/mayor-bowser-and-lyft-announce-free-capital-bikeshare-memberships-all-dc-residents.

Motivate International, Inc. n.d. "System Data." *Capital Bikeshare*. https://ride.capitalbikeshare.com/system-data.

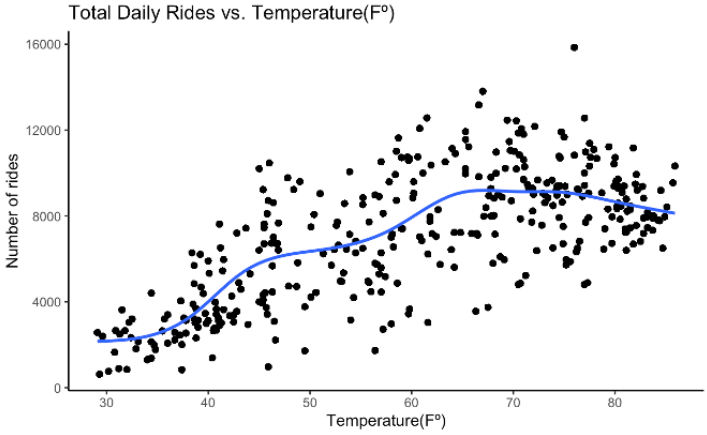
Walker, Kyle, and Matt Herman. 2022. *Tidycensus: Load US Census Boundary and Attribute Data as 'Tidyverse' and 'Sf'-Ready Data Frames*. https://CRAN.R-project.org/package=tidycensus.

Heggeseth, Brianna. 2022. "Correlated Data."

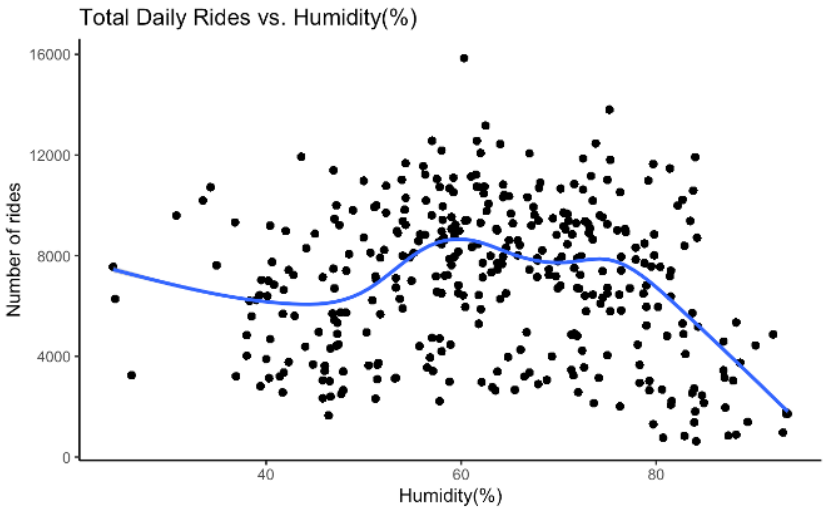
APPENDIX

Change quantitative into categorical variables

When reviewing our weather predictors such as temperature, humidity, and wind speed, we saw that the relationships between them and the daily number of riders was not globally linear. So, we decided to change these variables from quantitative predictors to categorical predictors. Looking at the temperature versus total daily number of riders graph, we see that there is a steady increase in the number of daily riders as temperature increases until it reaches about 65 degrees Fahrenheit. Once the temperature exceeds 65 degrees Fahrenheit, we see a slight decline in the total number of daily riders. Thus, we decided to create a binary temperature variable with a low level (> 65 degrees Fahrenheit) and a high level (65+ degrees Fahrenheit).



For our humidity graph, we plotted the daily average humidity percentage on the x-axis and the total number of daily riders on the y-axis. There is a slight increase in the total number of daily riders when humidity goes from 0% to 55%. This slight increase is followed by a relatively flat line when humidity goes from 55% to 75%, meaning that when humidity is within 55 to 75%, there is no change in the total number of daily riders. After the humidity exceeds 75%, we see a sharp decline in the total number of daily riders showing us that riders prefer lower-mid humidity levels. To simplify this relationship, we chose to incorporate two levels into the new categorical humidity variable: low humidity (0-74.99%) and high humidity (75-100%).



For our wind speed graph, we plotted the daily average wind speed on the x-axis and the total daily number of riders on the y-axis. There is relatively no change in the total number of daily riders for average daily wind speeds between 0 mph and 11 mph. However, after the daily average wind speeds exceed 11 mph, there is a negative relationship between wind speed and total number of daily riders. Meaning that as wind speed increases past 11 mph the total number of daily riders will decrease. So, we decided to cut our variable into two levels: low wind speed (< 11mph) and high wind speed (11+ mph).