Predicting Bike Availability in Washington DC

Group name: VanMoof

Prepared by: Clare Cruz

Tristal Li

Zhuoheng Han

Wei-Yu Tseng

For client: Capital Bikeshare

Presented in class: Perspectives in Data Science (36-601) & Professional Skills for Statisticians (36-611)
DATE: 11/30/2021

Agenda

Executive Summary

- Purpose: Capital Bikeshare has asked Vanmoof to build a sustainable model to predict availability of bikes by station for any given time period
- Methods: Statistical Modeling – Random Forest Regression Model using the average squared error as the accuracy measurement
- Results: Overall Accuracy – Mean Squared Error (MSE): 0.03
- Recommendations:
  - Bike reshuffling should adjust to the average temperature and hour
  - Our model should be put into production because it is unbiased, valid, and accurate

  Keeping the Model Working:
    - Make sure the data is clean and consistent
    - Run the data cleaning and modeling module
    - Input the station number and values for prediction

Introduction

- Why are we here?
  - Build a sustainable model to predict availability of bikes by station
  - The model is intended to improve bike availability by scheduling contractors who reshuffle bikes from/to bike stations
- Scope
  - Exploratory data analysis
  - Production level predictive model
  - Project plan and 3 status reports
  - Presentation deck

Background

- Capital Bikeshare Input
  - There is a relocation van that redistributes bikes
  - For large events, the capacity of bike stations increases(1)
  - The model should focus on bike availability (to be defined in the project)
  - The model needs to be dynamic, so that the analysts at Capital Bikeshare can continue to utilize it
- Assumptions
  - Because of the relocation van, a bike is not guaranteed to stay at the station where it was left
  - There may be new or relocated stations
  - Bikes can occasionally be stolen or broken
  - Aggregation information by time is representative of the true bike availability

(1). A fact we discovered in the data

Data Description

- Datasets:
  - Capital Bikeshare trip data in the year 2019[1 & 2] (To eliminate influence of Covid)
  - Capital Bikeshare station locations and capacity
  - Daily weather data from NCEI[3]
- Variables:

Category	Factor		Why include?
Predictive Variable	Bike Availability (%)		The factor we want to predict
Station	Station IDs		Unique identifier for each station
Time	Hour of day	Day of Week	Time can influence bike availability
Special events	Holidays		Special events may result in abnormal patterns
Weather	Precipitation (In)		People are less likely to ride a bike on rainy days
	Snow (In)		People are less likely to ride a bike on snowy days
	Temperatures (F)		People tend to not riding a bike in extreme temperatures

(1). Data accessed from: https://www.capitalbikeshare.com/system-data
(2). We used closing data from Dec. 2018 to help calculate bike availability, but it is not in the model
(3). National Centers for Environmental Information (NCEI), Data accessed from: https://www.ncei.noaa.gov/

Model Selection

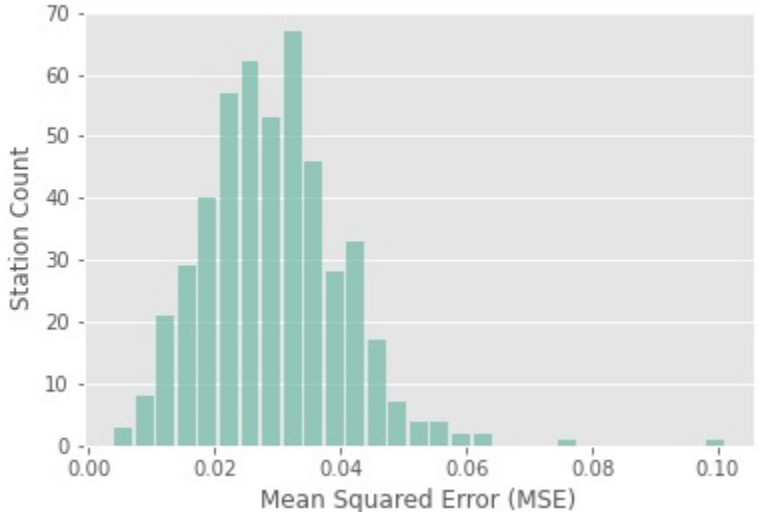
- Classification vs. Regression
  - We did not use classification for this project since the response is numerical, and dividing the groups would take more experimentation to see what would work best
  - We choose to use a **random forest regression** model since it has the power to handle a large data set with higher dimensionality [(1)]
- Accuracy Measurement
  - Mean Squared Error (MSE) = Avg[(Real value – Predicted value)^2]
  - Why MSE?[(2)]
    - Our model is regression
    - MSE is interpretable
  - The model is trained on 80% of the data and tested on the remaining 20%

(1). More detail on the MSE can be found in the technical appendix Page 16
(2). More detail on the models that were tested can be found in the technical appendix Page 14

Results

- Overall Accuracy – Measured by Mean Squared Error: 0.03
  - This means: For each predicted value from our model, its expected deviation from the true bike availability is $\sqrt{0.03} = 0.17$
  - In context, for a station that hold ten bikes, the model's predictions will be off by 2 bikes on average. For a station of 30 bikes, the model will only be off by 5 bikes or so

The distribution of MSE's are centered around 0.03



Recommendations

Our model should be put into production – why?

- The model is fairly accurate, errors are minimized
- The model is valid for every station and hour in the dataset
- The model is unbiased(1)

Potential Optimizations:

- Average Temperature – Van reshuffling schedules should adjust for the daily temperature
- Hour of the Day – Van reshuffling schedules should adjust to the hour of the day, especially *commuting hours*

(1). See Technical Appendix Page 17

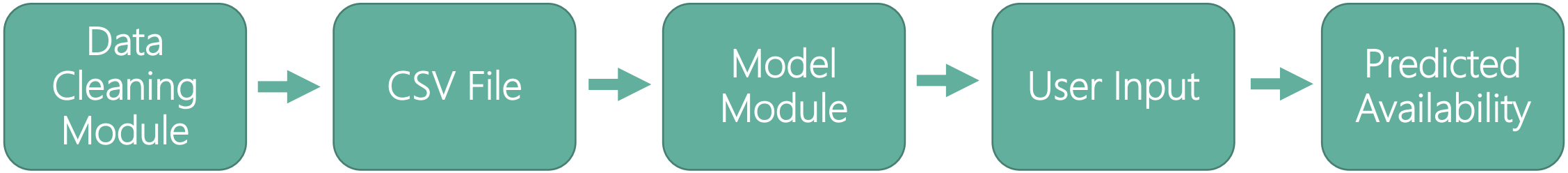
Keeping the Model Working

Data

- Update the bike trips, weather, and capacity database monthly
- Make sure the data is in consistent format (column names & data types)

User Input

- Apply the data cleaning module to clean the new data
- In the modeling module, the user inputs the station number and the values of our predictors for a prediction



Next Steps & Wish List

- Next Steps
  - Universal data processing algorithm
  - Add more variables that could be useful for prediction such as:
    - Metro and Bus Locations
    - Local Events Calendar
  - Restructure current datasets to include bike ID
- Wish List
  - Implement an online learning algorithm that can be automatically updated when more data comes in
  - Compile the Python functions into an app, which can be performed in a webpage to increase the user friendliness of the model
  - Structure the model to output potential van schedules

Q & A

Technical Appendix

Github Link to the works:
https://github.com/Tristal25/Capital_bikeshare_36601.git

Outline for Technical Appendix

- Key terminologies
  - Random Forest Regression Algorithm
  - Mean Squared Error (MSE)
- Model selection – Tested Model Algorithms
- Model evaluation – Accuracy by Locations
- EDA
  - Summary
  - Contents

Tested Model Algorithms

- Logistic Regression – It predicts values between 0 and 1 that follows a similar structure to simple linear regression
- K Nearest Neighbors Classification – It takes k nearest distances to predict the class. It predicts low (20%), medium (20%-80%), and high (80%) group with average accuracy 60% for one station
- Random Forest Classification - The output is the class selected from most trees. It predicts low (20%), medium (20%-80%), and high (80%) group with average accuracy 68% for one station

Random Forest Regression Algorithm

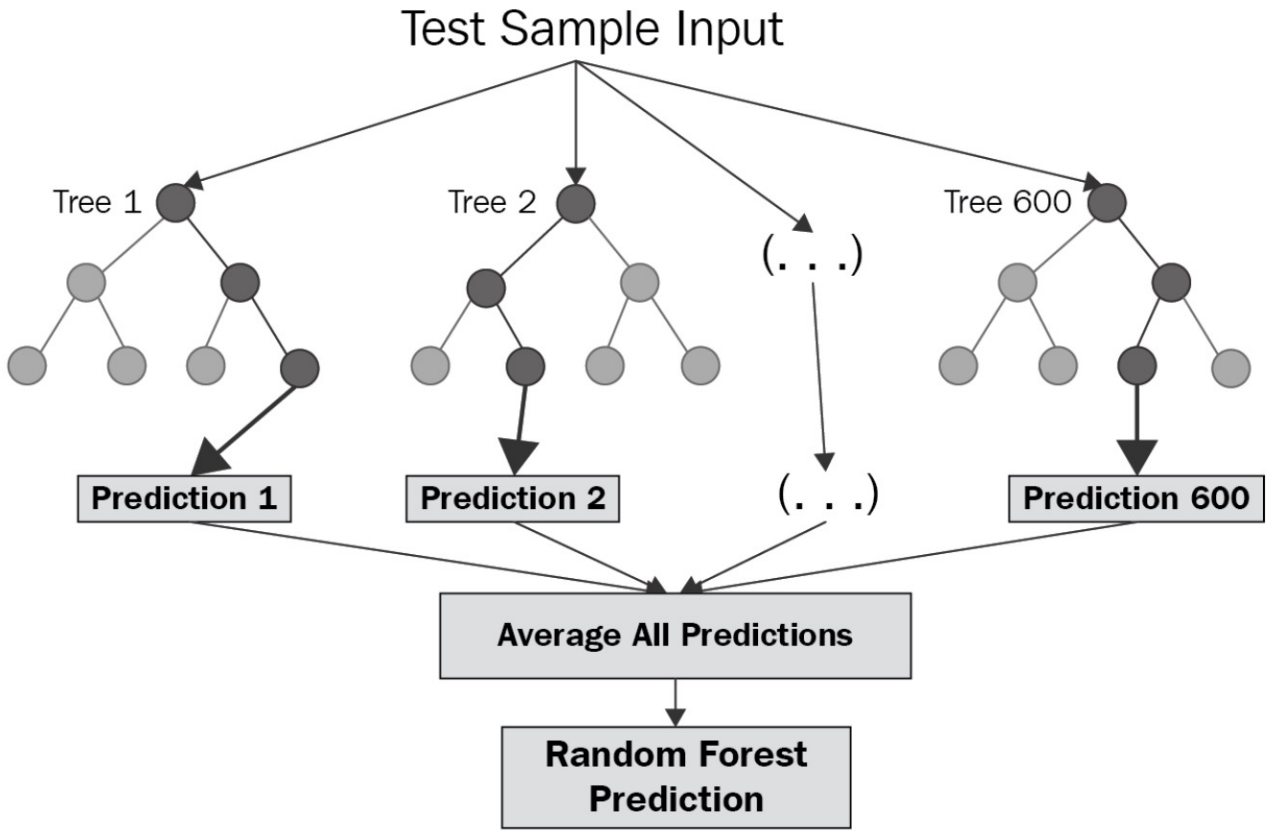
Step 1 – Sample the data with replacement X times. Where X is determined by the analyst

Step 2 – Fit a decision tree to each of the X data samples.

A decision tree is a model that utilizes several layers of binary decisions to predict an outcome

Step 3 – For each of the X samples, predict the value for an observation

Step 4 – Average the predictions across all the samples



Random Forest Structure

Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The average squared error between the actual bike availability (y_i) and the predicted availability from the model ($\hat{y}_i$) across all the hour in the year (N). The square root of the MSE describes the average error in the bike availability where lower values indicate better performance

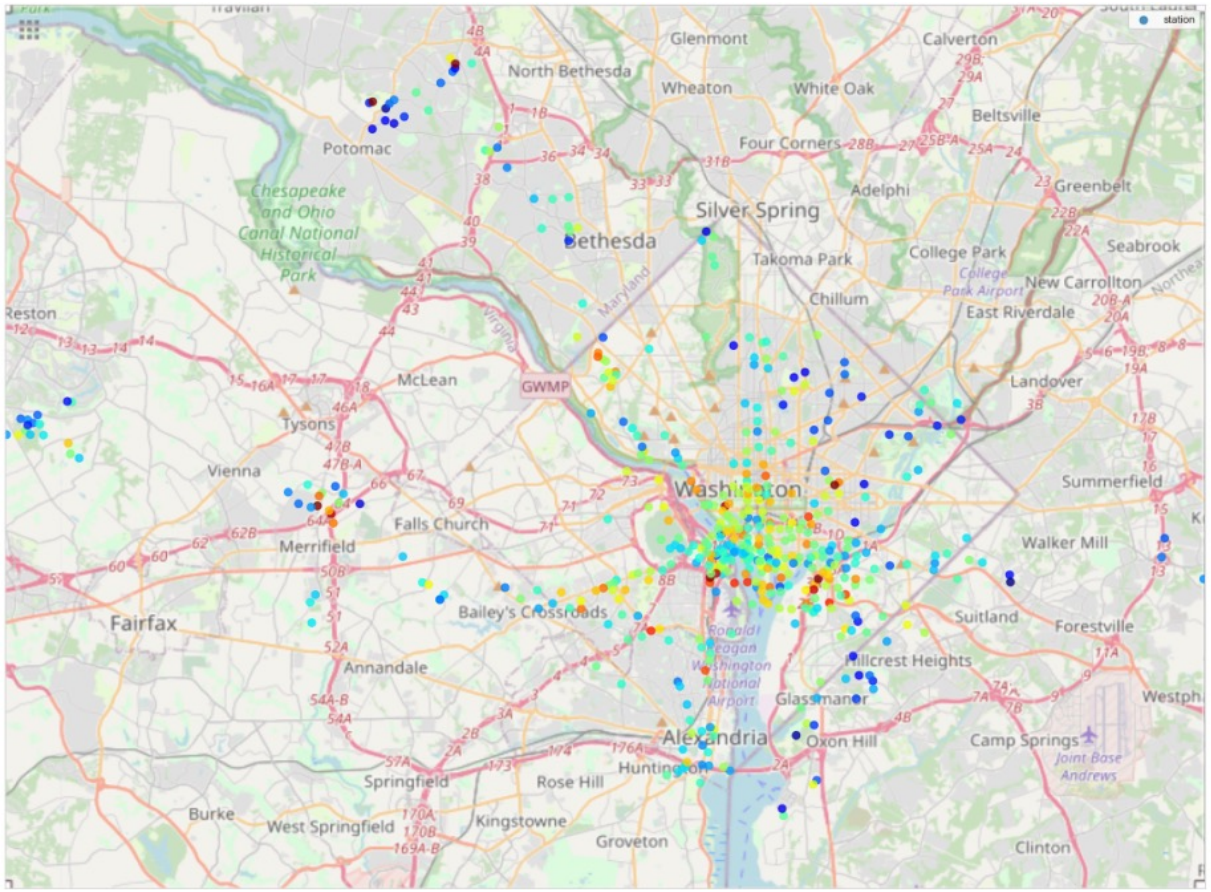
Pros

- Ensures that the model has no predictions with huge errors
- Easy to interpret and relate to the context of the problem

Cons

- Error is inflated if the model makes a single bad prediction

Model Accuracy by Location



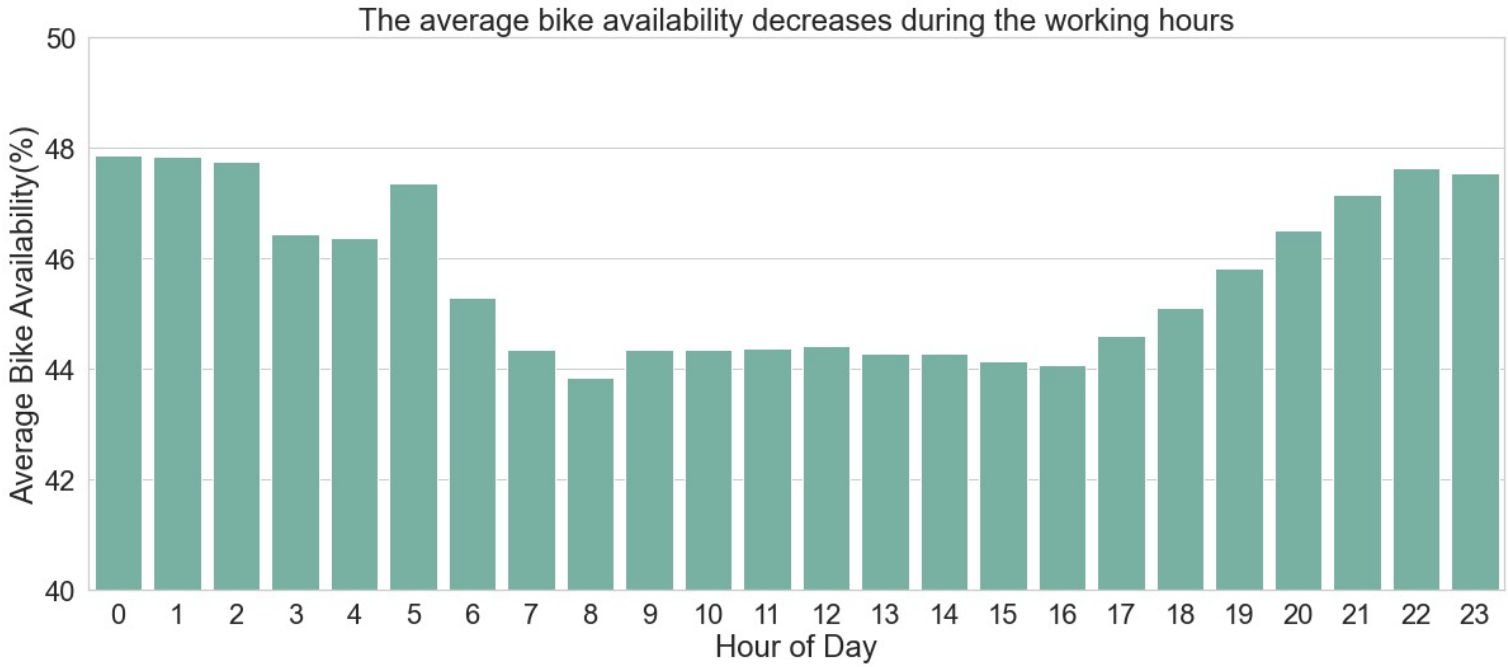
- Observations: Stations in the outskirts tend to have better accuracy. But the stations with the worst accuracy are not clustered into any specific area
- Conclusion: The model is not bias towards stations by location.

Exploratory Data Analysis

EDA Summary

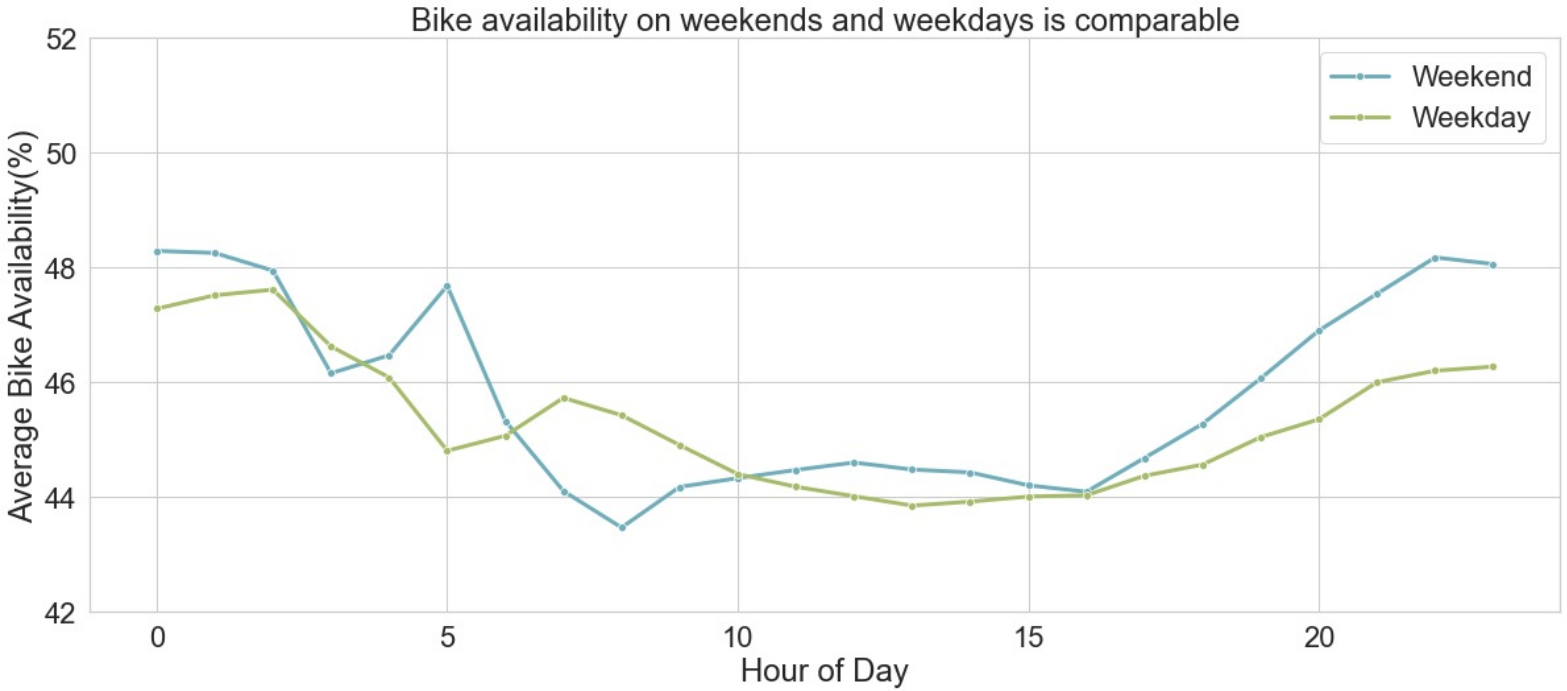
- Overall Bike Availability Over a Day:
  - Usually there is higher availability overnight than daytime
- Weekday vs. Weekend:
  - Bike availability on weekends and weekdays is comparable
- Bikeshare Station's Locations & Capacity:
  - More bikeshare stations and higher capacity in downtown than suburbs
- Commuting Patterns:
  - Higher bike availability in commercial areas after morning commute, and in residential areas after evening commute

Overall Bike Availability in a Day



- Higher average bike availability overnight than daytime
- Bike availability decrease sharply at 6 a.m., and increase gradually between 5 and 10 p.m.
- Conclusion: Average bike availability is lower in working hours
- Possible reason: More bikes are in use during daytime; Working and living schedules

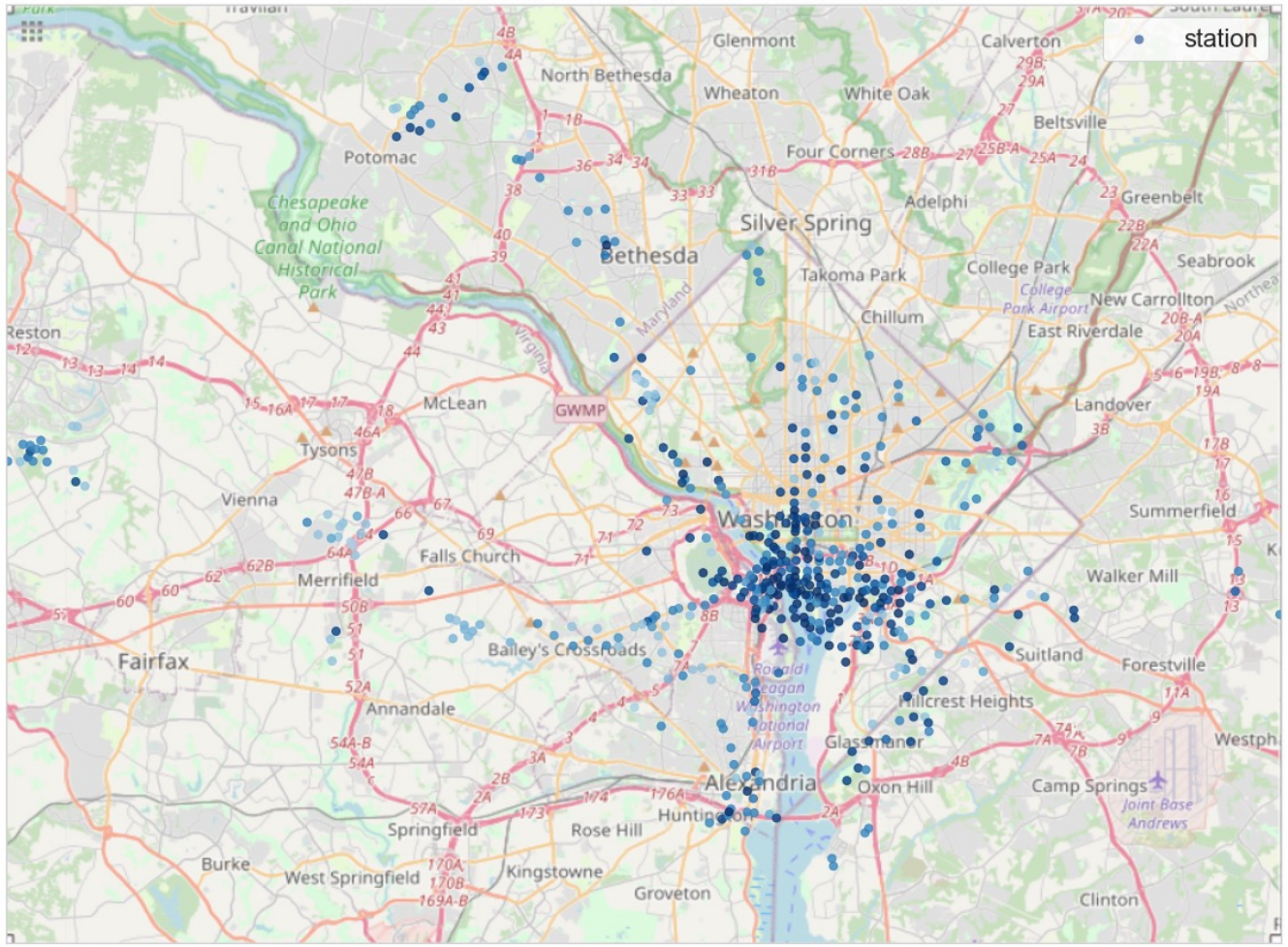
Weekday vs. Weekend

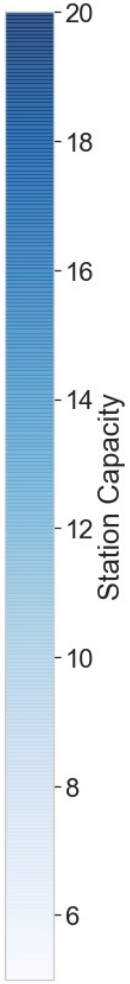


- The trend of two lines are similar except for some small fluctuations
- Conclusion: Although we originally thought that weekend vs. weekday can influence bike availability over a day, our analysis shows that they are similar

Bikeshare Station's Locations & Capacity

There are less stations in the non-downtown areas





Observations:

- Higher density of bikeshare stations in downtown than suburbs; Some small cluster of bike stations in outskirts
- Higher bikeshare station capacity in downtown areas

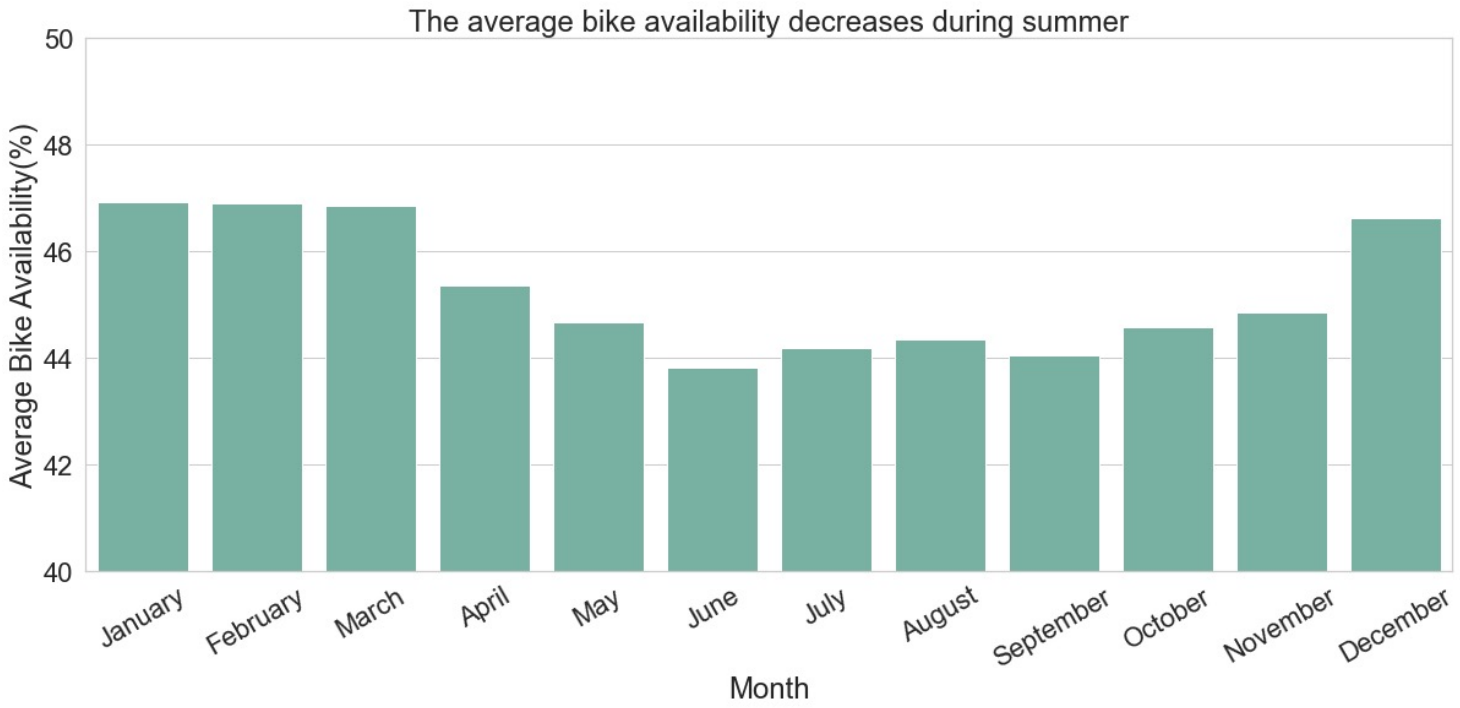
Possible reasons:

- Larger population & more companies in downtown
- Clusters of populations and companies in outskirts

Conclusion:

- Future analysis should separate downtown and suburban areas

Bike Availability by Month



- Conclusion: Average bike availability is lower in summer than other seasons
- Possible reason: More bikes are in use during summer because of the good weather