# Predicting Bike Share Availability

**Team KaZAM**

Woochan Lee
Ziyan Xia
Yueni Wang

# Agenda



**Team KaZAM**

Ziyan Xia

Woochan Lee

Yueni Wang

# Objectives and Assumptions

## Objectives:

- Build a predictive model to improve bike availability and reshuffling strategies across Washington DC
- Leverage existing bike share usage data through years to lower operational costs
- Improve user experience and sustain existing market share for client’s core business

## Assumptions:

- Model does not take into account unpredictabilities during COVID pandemic
- Reshuffling happens exactly halfway in time between a bike’s start trip and end trip disparities
- Starting number of bikes in each station is total number of bikes that arrived at each station in previous month and did not leave
- Weather data for a subset of stations is used to represent the overall weather for Washington DC area

# Overview of the Raw Dataset

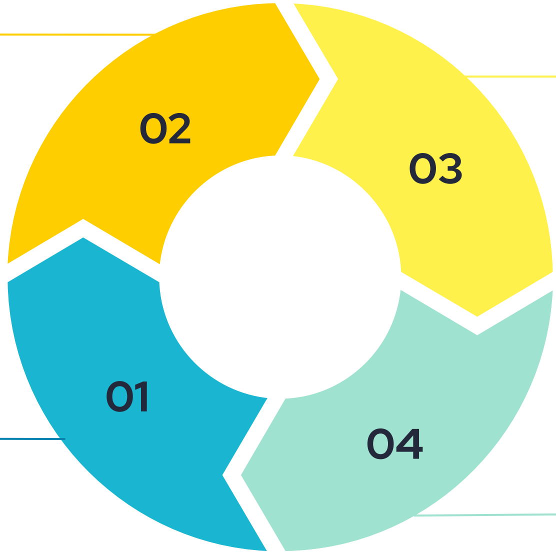
## Raw data:

- Dates: 2019 January 1st ~ 2019 December 31st
- Historical and real time
- Each observation is a trip history of Capital Bikeshare bikes - the data includes:
  - Duration
  - Start Date
  - End Date
  - Start Station
  - End Station
  - Bike Number
  - Member Type
  - Having electric bikes or not, etc.

## Additional effort:

- Retrieved weather data of Washington DC from NOAA website:
  - Precipitation
  - Temperature
  - Wind level
  - etc…
- Utilized Capital Bikeshare's API to retrieve
  - Longitude and latitude of each station
  - Total capacity of each station

# Transformation of the Raw Dataset




### 1. Splitting raw data

Each observation of trips in the raw data was split into 2 separate trips:

- Bike picked up
- Bike dropped off

### 2. Adding reshuffle trips

We assumed that reshuffling occurred when bike last arrived at station A, but was recently picked up from station B

### 3. Starting bike numbers

We assumed that the starting number of bikes in 2019 January were the total bikes that arrived in 2018 December

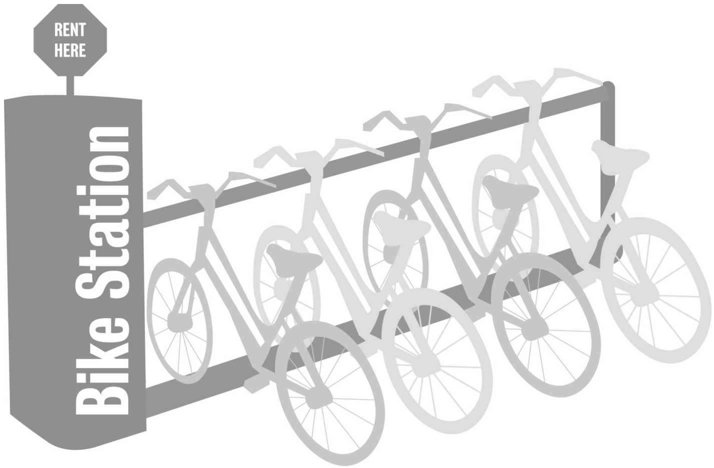
### 4. Defining availability

We defined availability to be the **proportion** of available bikes in a station at a certain time

# Overview of the Transformed Data

**Response Variable:**

- **availability**:
  - YES:  station has **at least** 20% bikes available
  - NO: station has **less than** 20% bikes available




**Explanatory Variables:**

- **is_weekday**: the day is a weekday
- **month**: month of the trip
- **PRCP**: precipitation rate during day of the trip
- **TAVG**: average temperature during day of the trip
- **hour_minute**: during which 30 minutes of the day the bike ride happened
- **station_id**: station number

# Clustering of Stations

- Clustering based on:
  - Longitude
  - Latitude
  - Variance of availability
- From the 5 resulting clusters, we focused on the central cluster with supposedly higher demand
- From this central cluster, we then chose the **top 20** stations with **highest variance** of bike availability
  - To ensure the 20 stations portrayed similar characteristics in availability trends
  - To ensure stations are not too far apart from each other (reshuffling efficiency)

## Selected 20 Stations

# Final Model and Results - Random Forest

- We sampled the data to ensure that it consist of 50% “Yes” and 50% “No” in availability so that  the model can learn both cases equally
- We split the data into **95% training set** and **5% test set**
- Using the transformed data for the 20 selected stations, we chose to fit a **Random Forest** model to predict availability (“Yes” or “No”) after comparing this model with Logistics Regression Model

| **Actual** | **Predicted** | |
|---|---|---|
| | Yes | No |
| Yes | 4,164 | 843 |
| No | 607 | 4,386 |

→

| | |
|---|---|
| **Overall Accuracy** | 85.5% |
| **Accuracy for “Yes”** | 83.2% |
| **Accuracy for “No”** | 87.8% |

# Model in Practice

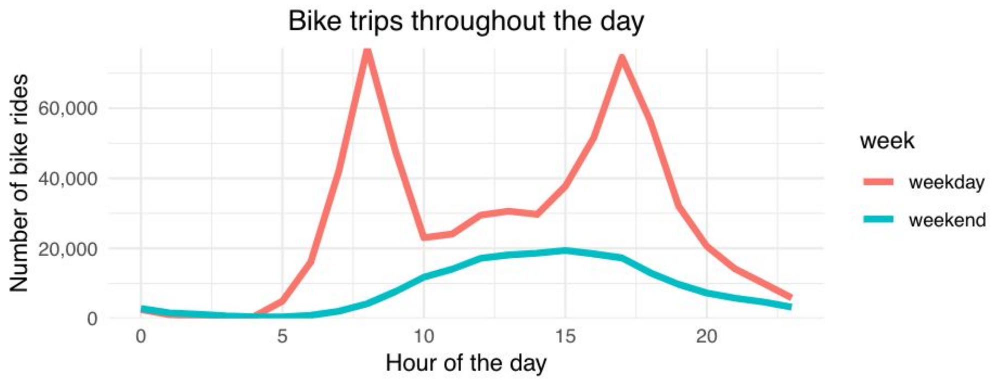
- To use the model for prediction, the inputs from a user-perspective should be:
  - **is_weekday**
  - **month** → Auto-assigned using the date of the user's request
  - **hour_minute**
  - **PRCP**: → These can be acquired and auto-updated using weather forecast APIs
  - **TAVG**:
  - **station_id**: → This is the station the user is interested in (the only "manual" input needed)
- The model can predict the availability status of any stations **within our chosen cluster** at a user defined time interval for customers
- The further the date the user wants to predict, the less accurate the predictions will be because trends can change

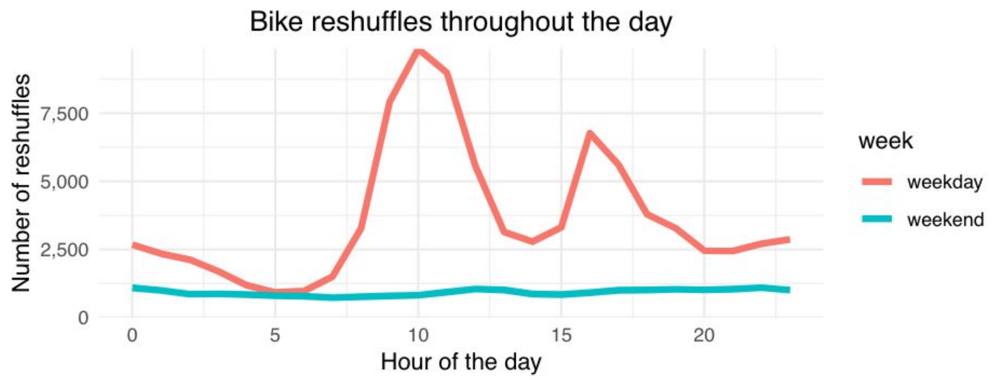
# Maintenance of the Model

To maintain the model, we should retrain the model every month to update the model based on recent trends

- Maintain a **“window” of one year**: delete the oldest month, and add recent corresponding month
  - e.g. **[**August 2020-August 2021**]** → **[**Sep 2020 **-** Sep 2021**]** → **[**Oct 2020 **-** Oct 2021**]** …
- Most importantly, we need to ensure that all relevant variables are tracked consistently (e.g. 2020 March stops tracking the bike_id per trip)

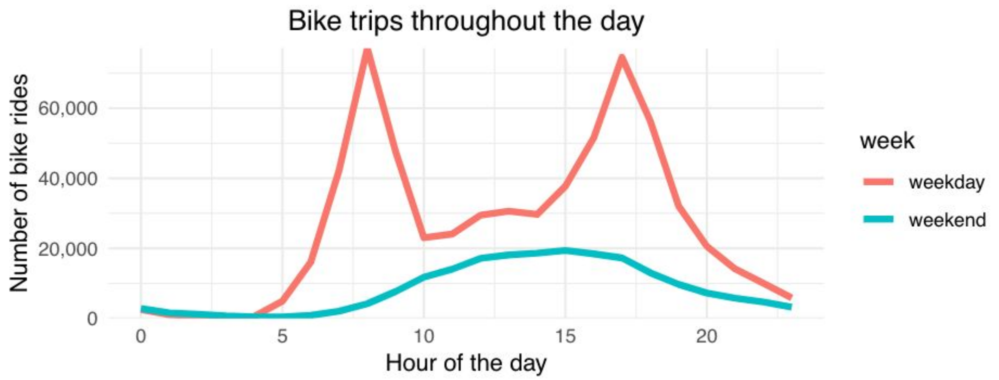
# Understanding the Reshuffling Patterns

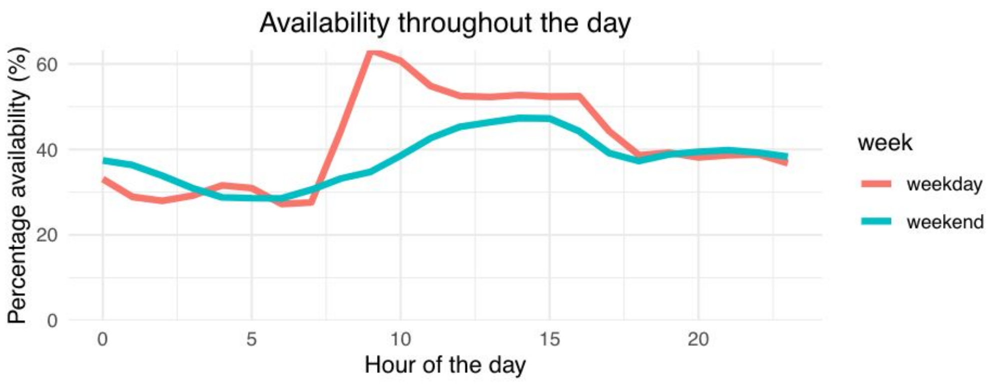



- Our cluster of stations are nearby busy districts and office buildings
- Bike rides reach peaks during two rush hours
  - 7AM ~ 9AM (morning)
  - 5PM ~ 7PM (evening)
- Bike reshuffles peaks:
  - After morning rush hours
  - Before evening rush hours

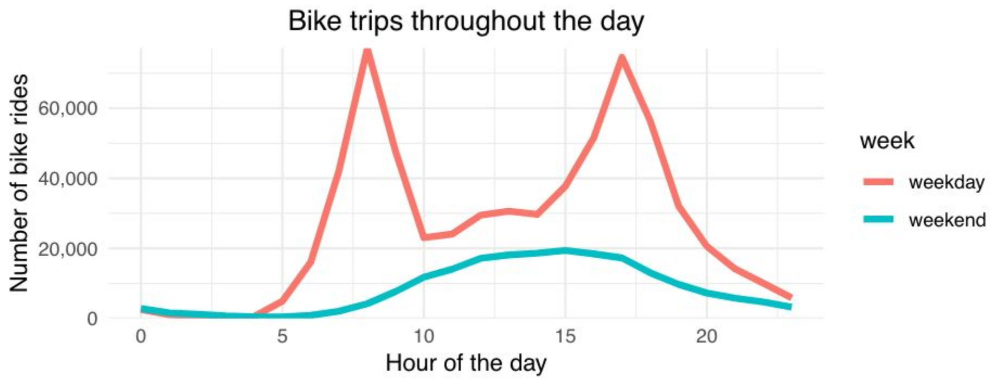
# Availability Trends and Reshuffling Strategy

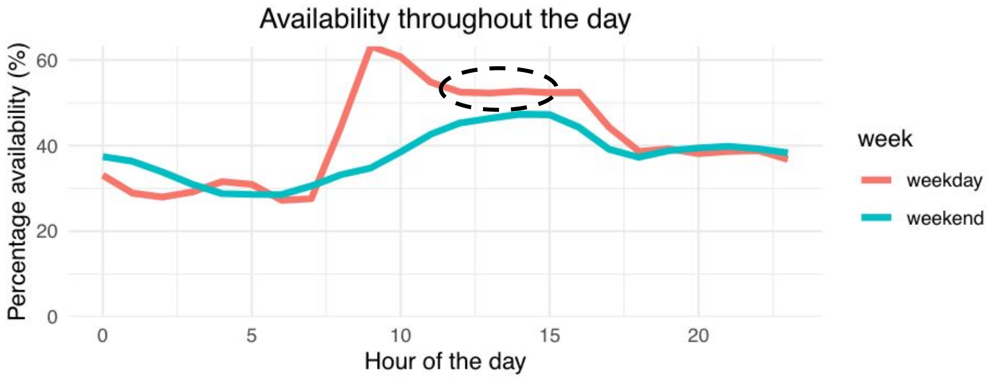



- Morning rush:
  - Huge influx of bikes that shoots up bike availability to 60%
  - After morning rush, reshuffling is done to prevent stations overloading with bikes
- Evening rush:
  - Large decrease in bike availability to 40%
  - Before evening rush, bike availability remains high
  - Reshuffling is done to maintain bikes that people can ride back home

# Availability Trends and Reshuffling Strategy




- Availability is high between the two rush hours, although not many bike riders

- This suggests idle bikes not being utilized efficiently
  - They should be re-allocated to other clusters / stations in need

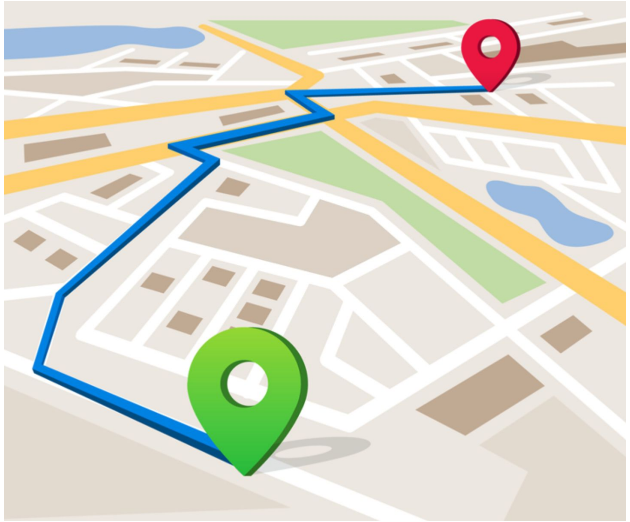
- Use our model to predict which specific stations or clusters suffer from low availability at the time interval

# Actionable Recommendation

- Much like our chosen cluster, different clusters would have differing needs at certain time of the day
  - Some clusters could be short of bikes at a certain time of the day
  - Some clusters could have more bikes than needed (underutilized)
- We could expand this concept to fit models for different clusters and capture their own characteristics by predicting availability patterns
  - Adjust our future reshuffling strategies

# Optimization of Operational Cost

- Capture additional data on reshuffling field-workers
  - Price of gasoline
  - Wage for the driver
  - Capacity of bikes a truck can take
  - Distance between stations
- Create “optimal” route for field-workers to reshuffle bikes
- Our aim would be to **minimize total operation costs** while achieving balanced bike availabilities that meet the needs of different station clusters
  - Reduce unnecessary reshuffles
  - Increase necessary reshuffles



# Next Steps for Phase 2

- Identify other distinct clusters of stations
  - Rural areas
  - Parks
  - Schools
- Expand our model concept to be able to predict for these additional stations and clusters
- Create separate models for 2020 / 2021 that can account for COVID pandemic
  - Adds complexity, but more relevant to today's time

# Thank You!

Any Questions?

# Technical Appendix

# Transformation of the Raw Dataset

- Each observation of trips in the raw data was split into 2 separate trips:
  - Bike picked up
  - Bike dropped off

- Reshuffling of bikes was not accounted for in the raw dataset, and needed to be treated as separate trips
  - We assumed that reshuffling occurred when bike last arrived at station A, but was recently picked up from station B

- We added starting number of bikes at each station in January 1st 2019, by using data from 2018 December
  - Total number of bikes that arrived in 2018 December

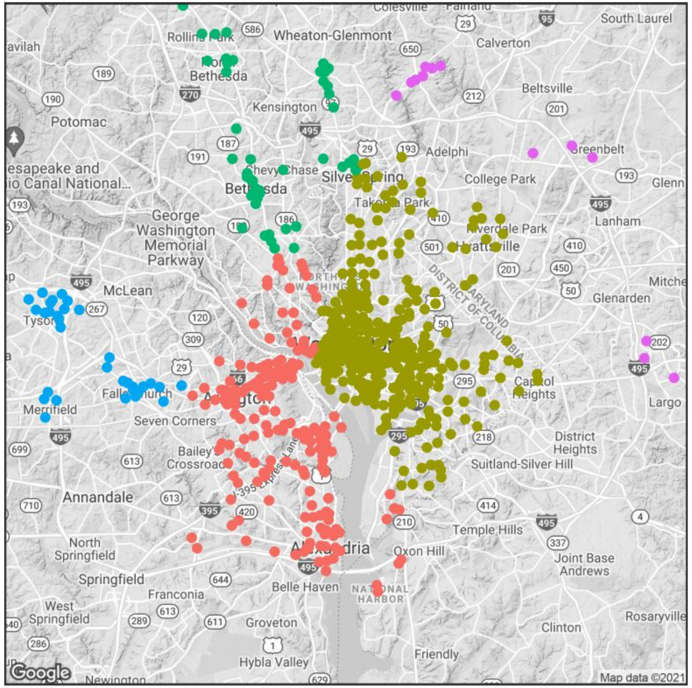
- We defined **availability** to be the **proportion of available bikes at a station in a given time**
  - Availability = Number of bikes at a station in a given time / Total capacity of that station

# Clustering of Stations

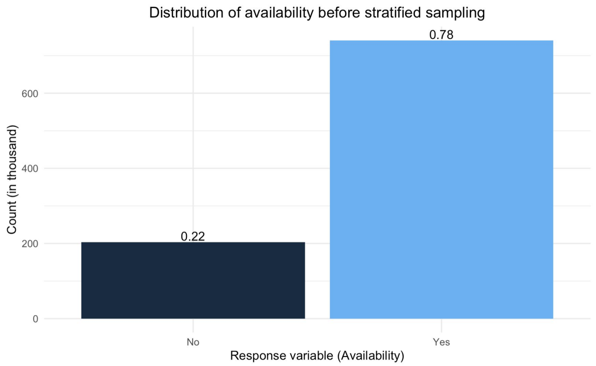
- We first performed hierarchical clustering based on:
  - Longitude
  - Latitude

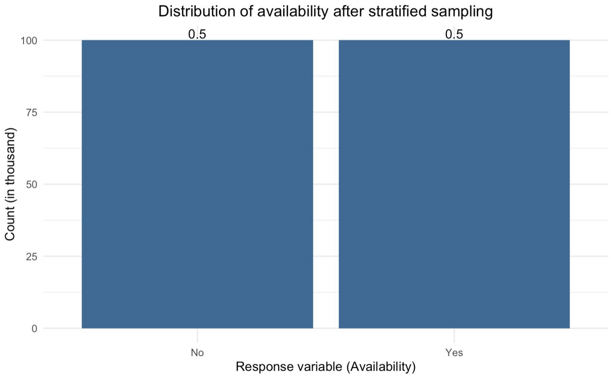
- We ended up producing **5** different cluster groups
- We wanted to focus specifically on the central clusters where we hypothesized there will be higher demands
  - Urban area
  - Busier streets & office buildings

Clusters on WashingtonDC Map



# Handling Data Imbalance

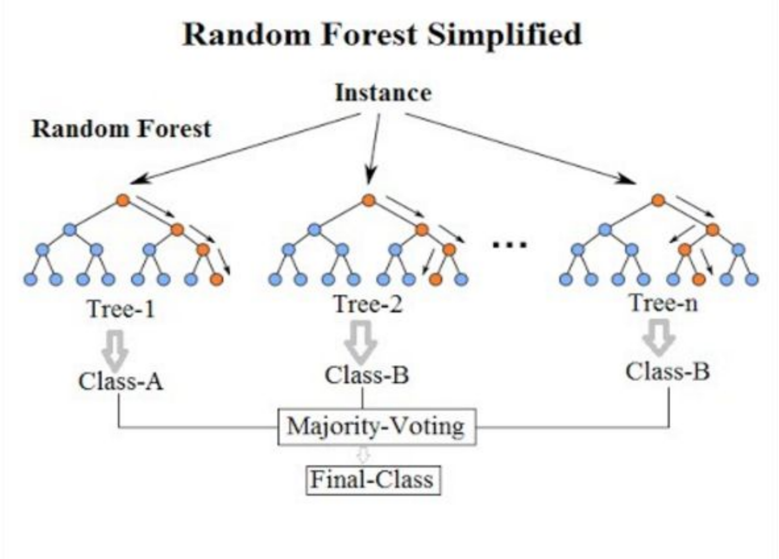



- Because of our 20% availability threshold, there was a high imbalance in our response variable
    - ~80% “Yes”, ~20% “No”
- The model failed to learn about the “No” cases
- We used stratified sampling to ensure that our training sample data had 50:50 ratio between y = “Yes” and y = “No” observations
- Then, with the correctly trained model, we created a test set sampled from the true population to validate our model accuracies
- This significantly improved our model’s performance on the negative y=”No” cases

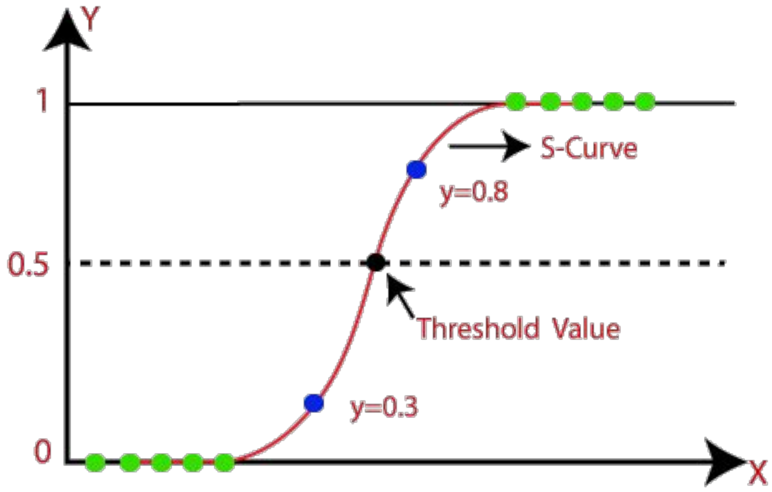
# Summary of Methodology

## Random Forest




Random Forest is a common learning method for classification using decision trees.

## Logistics Regression




Logistic Regression uses a logistic function to model the dependent variable. The dependent variable is dichotomous in nature, i.e. there could only be two possible classes

# Model Results

## Random Forest

Confusion Matrix of the Actual and the Predicted

| Actual | Predicted | |
|---|---|---|
| | Yes | No |
| Yes | 4,164 | 843 |
| No | 607 | 4,386 |

Overall Accuracy: 85.5%
Accuracy for “Yes”: 83.2%
Accuracy for “No”: 87.8%

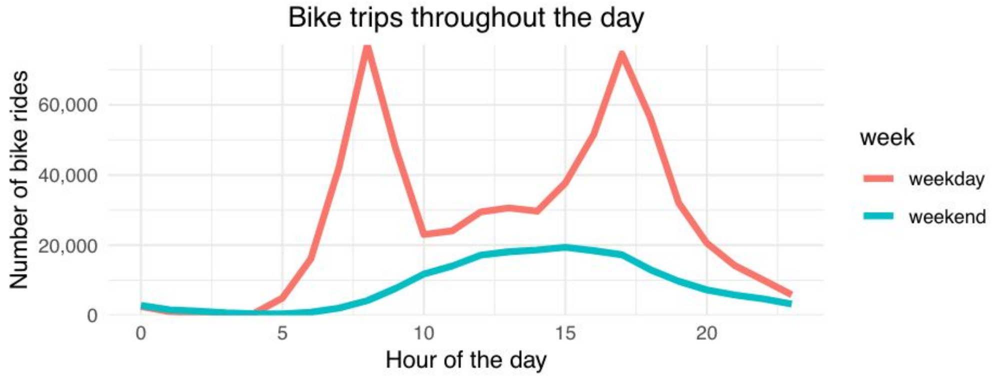
## Logistics Regression

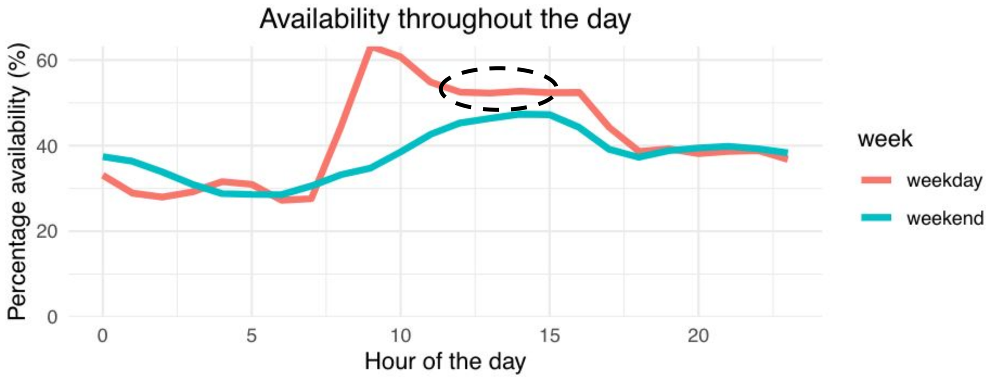
Confusion Matrix of the Actual and the Predicted

| Actual | Predicted | |
|---|---|---|
| | Yes | No |
| Yes | 3,436 | 1,571 |
| No | 1,925 | 3,068 |

Overall Accuracy: 58.2%
Accuracy for “Yes”: 54.8%
Accuracy for “No”: 61.7%

# Availability Trends and Reshuffling Strategy




- We fit a Logistic regression model:
    - 10% increase in availability(%) leads to 4% increase in odds of reshuffle

| Reshuffle type | Average availability |
|---|---|
| Bring in bikes (+1) | 37.8% |
| Take out bikes (-1) | 48.5% |

- We found that the trucks were more likely to take away the bikes as the availability reached higher levels.
- The majority of the reshuffling that happen at the peak availability, was the trucks taking away the bikes.