# Preface

1. Code environment
   1. Python 3.7 via Jupyter Notebooks was used to create this solution.

**Haversine formula inspiration:** <https://gist.github.com/rochacbruno/2883505>

# Divvy Bike Data

1. Data exploration
2. The data was found on the Divvy Bikes website: <https://www.divvybikes.com/system-data>
3. Once the quarterly trips data frames are concatenated, station data and coordinates are merged to each row, and distance and speed are calculated and added in as columns, there are 3,829,003 rows and 18 columns. The boxplot below shows the summary statistics for trip duration by user type after rides less than 2 minutes or greater than 6 hours are removed. The next graph shows the distribution of rides by trip duration in seconds. Almost all rides are under 5,000 seconds or 83.33 minutes. The bar chart shows that most of our riders are subscribers, followed by customers, with only a handful of dependent riders. The scatterplot shows the geographical lay out of the stations most frequently used- the stations on “the loop” or near Navy Pier are in downtown Chicago and near touristy areas, so they are most popular. The border on the right is Lake Chicago, so there are no stations there, but there is heavy usage from the stations in the east, riders seem to enjoy riding near the lake.

**Trip Duration:**

A picture containing screenshot, clock

Description automatically generated

A picture containing building, bridge

Description automatically generated

**User Types:**

A screenshot of a cell phone

Description automatically generated

**Station Locations:**

A screenshot of a cell phone

Description automatically generated

1. Data Visualization.
   1. Top 5 stations with the most starts (showing # of starts)

|  |  |  |
| --- | --- | --- |
| Station ID | Station Name | # of Trips |
| 35 | Streeter Dr. & Grand Ave. | 97,070 |
| 192 | Lake Shore Dr. & Monroe St. | 53,132 |
| 91 | Canal St. & Adams St. | 50,574 |
| 76 | Clinton St. & Washington Blvd. | 49,542 |
| 77 | Theater on the Lake | 47,627 |

The bar chart below shows that Streeter Dr. & Grand Ave. has the most starts by far.

**Stations with Most Starts:**

A close up of a logo

Description automatically generated

* 1. Trip duration by user type

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trip Duration by type | | | | |
|  |  |  | Minutes | |
| User type | Trips | Total Hours | Average | Median |
| Subscriber | 2,959,920 | 572,791.16 | 11.62 | 9.62 |
| Customer | 844 | 210.37 | 29.73 | 22.85 |
| Dependent | 7 | 1.46 | 12.53 | 13.40 |

The box plot below shows the IQR of trip duration for each user type. The bar chart shows average trip duration by user type. Customers have the longest trips on average, and their distribution is skewed right.

**Trip Duration by User Type:** A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

* 1. Most popular trips based on start station and stop station

|  |  |  |
| --- | --- | --- |
| Start Station | End Station | # of Trips |
| Lake Shore Dr. & Monroe St. | Streeter Dr. & Grand Ave. | 12,170 |
| Streeter Dr. & Grand Ave. | Streeter Dr. & Grand Ave. | 9,611 |
| Streeter Dr. & Grand Ave. | Theater on the Lake | 8,180 |
| Streeter Dr. & Grand Ave. | Lake Shore Dr. & North Blvd. | 7,992 |
| Lake Shore Dr. & North Blvd. | Streeter Dr. & Grand Ave. | 7,225 |

The bar chart below shows that there is a large difference in how many customer took the top 10 routes with nearly 8,000 rides separating the most popular and 10th most popular routes.

**Most popular trips:**

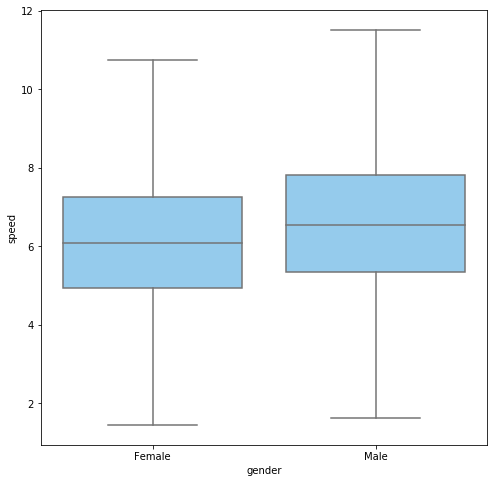
A screenshot of a cell phone

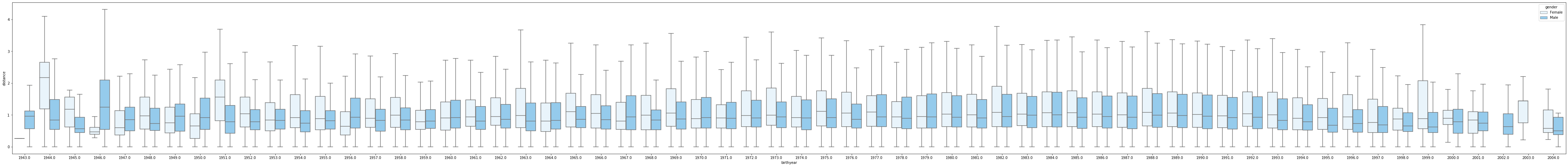
Description automatically generated

* 1. Rider performance by Gender and Age based on avg trip distance (station to station), median speed (distance traveled / trip duration)

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Age Group** | **Median Speed (miles/hr)** | **Avg. trip distance (miles)** |
| Female | 17-25 | 6.06 | 1.21 |
| 25-35 | 6.24 | 1.29 |
| 35-45 | 6.10 | 1.29 |
| 45-55 | 5.74 | 1.17 |
| 55-65 | 5.40 | 1.11 |
| 65-70 | 5.09 | 1.11 |
| Male | 17-25 | 6.44 | 1.06 |
| 25-35 | 6.74 | 1.23 |
| 35-45 | 6.59 | 1.22 |
| 45-55 | 6.31 | 1.12 |
| 55-65 | 5.84 | 1.04 |
| 65-70 | 5.55 | .97 |

**Average Speed by Gender:**



**Trip distance by Age and Gender:** ****

**Average Trip duration by Gender:** **A screenshot of a cell phone

Description automatically generated**

* 1. What is the busiest bike in Chicago in 2017? How many times was it used? How many minutes was it in use?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Time in Use | | |  | Trip Count | | |
| Bike ID | Trip Duration (Minutes) | Times Used |  | Bike ID | Times Used | Trip Duration (Minutes) |
| 2565 | 14,845.3 | 1,231 |  | 2565 | 14,845.3 | 1,231 |
| 3489 | 12,178.17 | 1,043 |  | 3489 | 12,178.17 | 1,043 |
| 2438 | 11,733.68 | 996 |  | 2438 | 11,733.68 | 996 |
| 3308 | 11,555.35 | 993 |  | 3308 | 11,555.35 | 993 |
| 5719 | 11,305.73 | 972 |  | 5719 | 11,305.73 | 972 |

The two bar charts below show that the busiest bike is the same when comparing most starts and most minutes used.

**Busiest Bike (# of starts):**

A screenshot of a cell phone

Description automatically generated

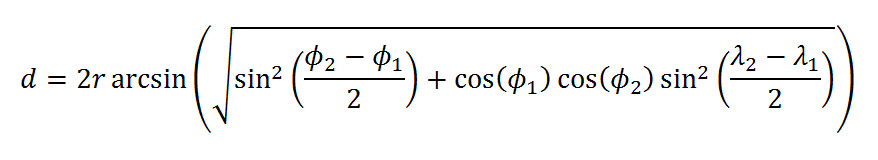
**Busiest Bike (minutes used):**

A screenshot of a cell phone

Description automatically generated

1. Data cleaning  
   Before building your model, you will need to clean your data. Describe the steps you take here.
   1. Unused columns were dropped (city, dpcapacity, name, online\_date). The different csv files from the Divvy Bikes website were combined by vertically concatenating the trip data for each quarter and merging the station data for the start and end stations.
   2. Assumptions: Treat all birth years as if birthday is 1/1 for calculating age groups. Rides less than 2 minutes where checked out and checked back in without leaving the station, they are dropped. Rides greater than 6 hours were abandoned bikes, they are dropped. Speeds over 40 MPH are dropped, it is unlikely any Olympians are using Divvy bikes. Ages under 10 are dropped, ages over 75 are dropped.
2. Modelling  
   Build a model that can predict how long a trip will take given a starting point and destination
   1. Feature engineering: Distance was created using the Haversine formula. The Haversine formula can be seen below. Speed was created using distance and duration. The categorical variables gender and rider type were made into dummy variables. Distance vs duration scatter plot can be seen below
   2. External data sources: All data was gathered from the Divvy Bikes website.
   3. Model selection: Models were selected and tuned by evaluating the R2 and MSE values and trying to create a model with the highest R2 and lowest MSE. The best results for each model (simple linear regression, multiple linear regression, ridge regression, and lasso regression) can be seen below.
   4. Feature selection: Features were selected by evaluating the R2 values before and after they were added into the multiple linear regression.
   5. Model validation: Models were evaluated using an 80/20 train/test split. Training RMSE for the final Lasso model is 207.12. Testing RMSE for the final Lasso model is 209.85. R2 for the final Lasso model is .7993, meaning that 79.93% of the variation in trip duration can be explained by the model.
   6. Improve baseline model: Model was improved by adding different variables and trying different modeling techniques to increase R2 and decrease MSE.

**Haversine Formula:**



**Distance vs Duration:**

A close up of a map

Description automatically generated

**Simple Linear Regression (R2 = .457):**

A screenshot of a cell phone

Description automatically generated

**Multiple Linear Regression (R2 = .537):**

**A picture containing pencil, sitting, filled, table

Description automatically generated**

**Ridge Regression (R2 = .7993, alpha = .05):**

**A close up of a mans face

Description automatically generated**

**Lasso Regression (R2 = .57):**

**A screenshot of a cell phone

Description automatically generated**

**Best Lasso Regression (R2 = .7993, alpha = .5):**

A screenshot of a cell phone

Description automatically generated

1. Conclusion and next steps:
   1. Model is more accurate for shorter trips because start and end stations are used to calculate distance, rather than using GPS tracking, so meandering is not taken into account. Future recommendations include using other data sets like weather, traffic patterns, and holidays, using GPS distances rather than calculating from start and end stations, and exploring other modeling techniques.