2017 Operating Report:
Mayor De Blasio

Vinit Shah

Table of Contents

Background & Context

Data Exploration and Visuals

Data Preparation

Data Modelling

Background & Context

1

PURPOSE

This operating report aims to help the mayor obtain a better understanding of Citi Bike operations. The report thoroughly examines Citi Bike trips in the year of 2017. The data was obtained from Citi Bike's AWS portal available online. The main objective is to help Mayor de Blasio understand how tourists and the residents of New York City use Citi Bike. Lastly, the report gives a brief overview of the new feature the mayor desires: trip duration based on start and end station.

2

MAYOR DE BLASIO'S REQUESTS:

1. Top 5 stations with the most starts
2. Trip duration by user type
3. Most popular trips based on start station and stop station
4. Rider performance by Gender and Age based on average trip distance median speed
5. What is the busiest bike in NYC in 2017? How many times was it used? How many minutes was it in use?

Background & Context

Data Exploration and Visuals

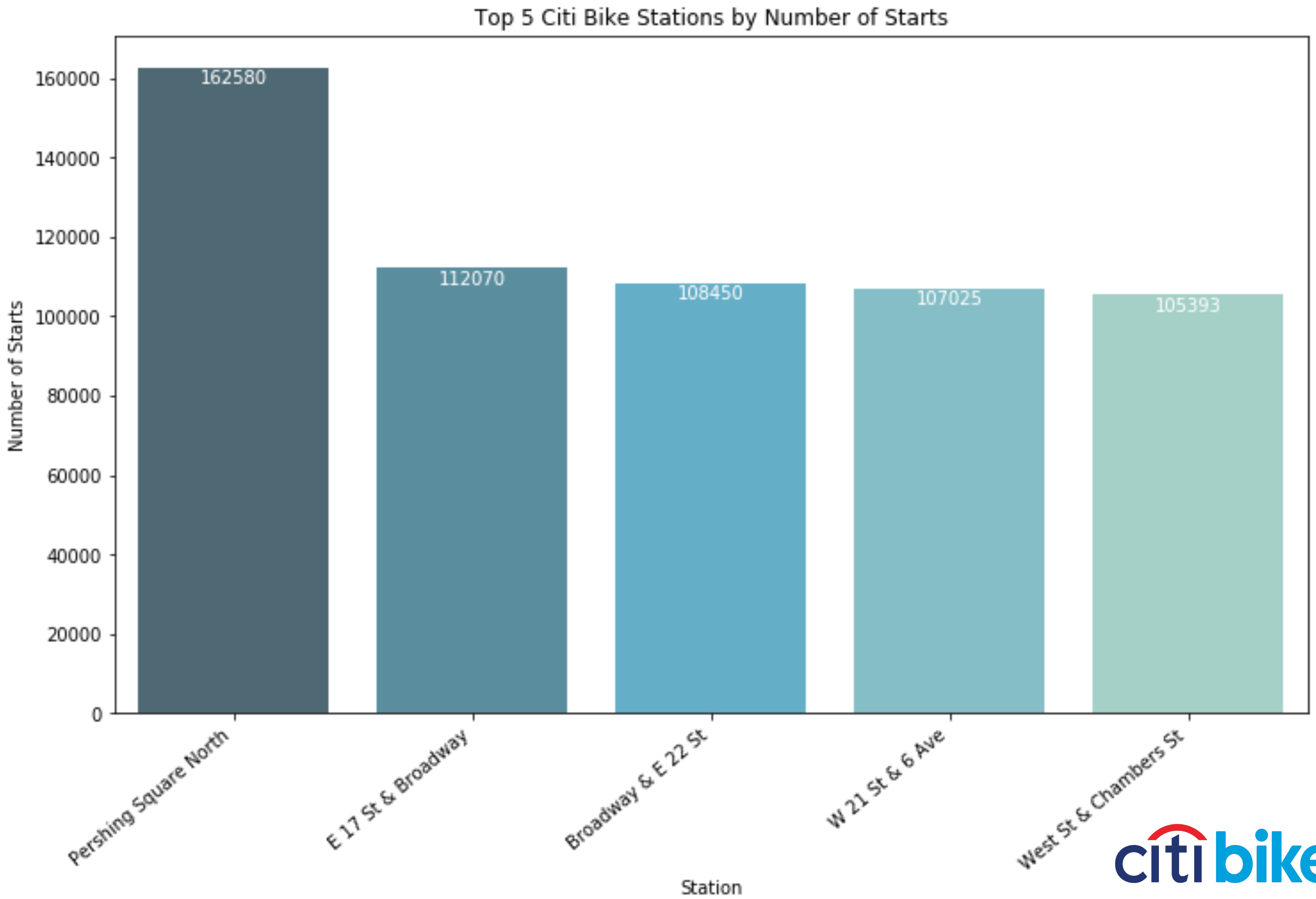
Data Preparation

Data Modelling

Top 5 Stations by Number of Starts

TOP 5 STATIONS

1. Pershing Square North
2. E 17 St. & Broadway
3. Broadway & E 22 St.
4. W 21 St. & 6th Ave.
5. West St. & Chambers St.



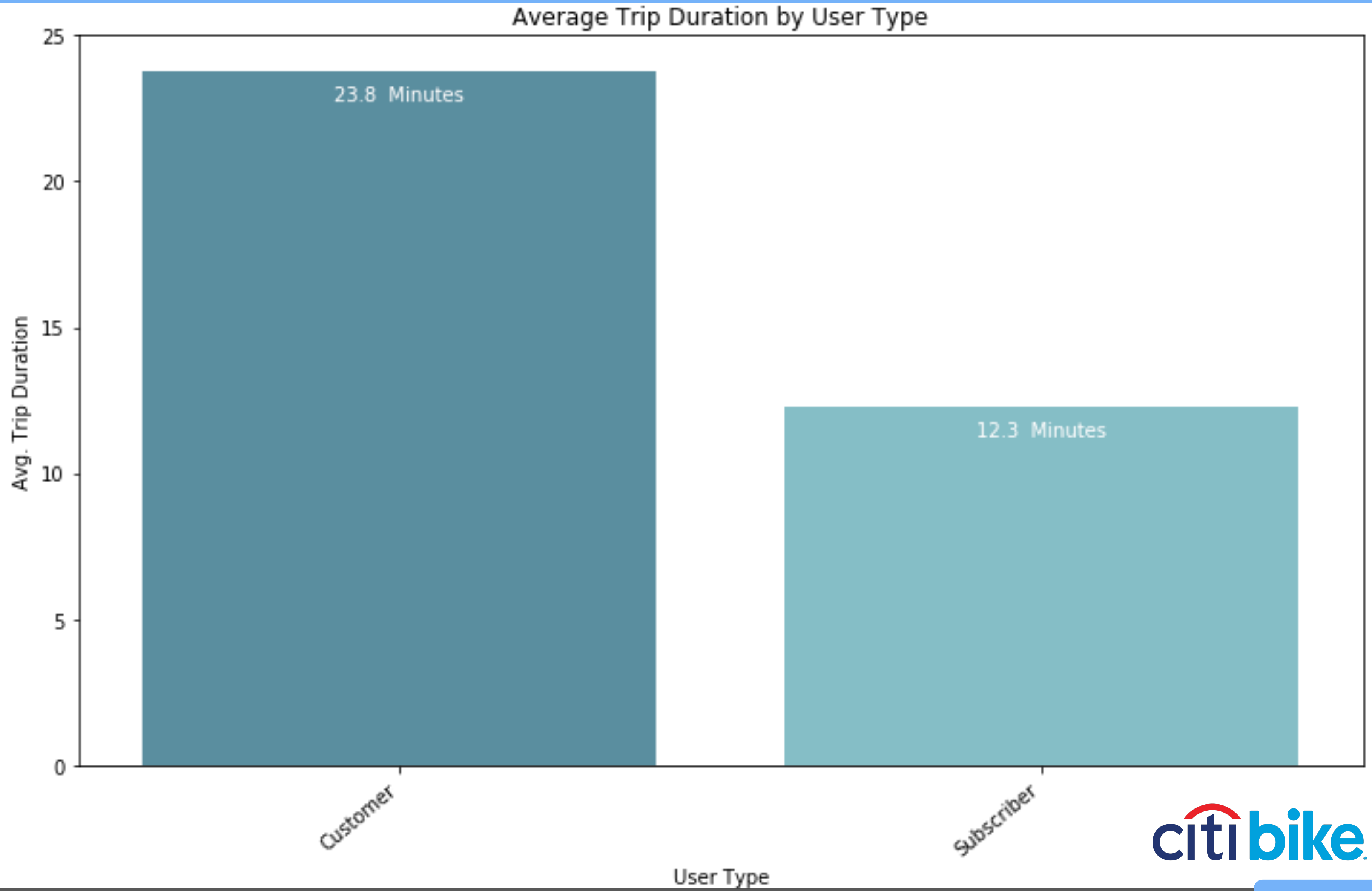
Trip Duration by User Type

Average Trip Duration for "Customer"

- 23.8 Minutes

Average Trip Duration for "Subscriber"

- 12.3 Minutes



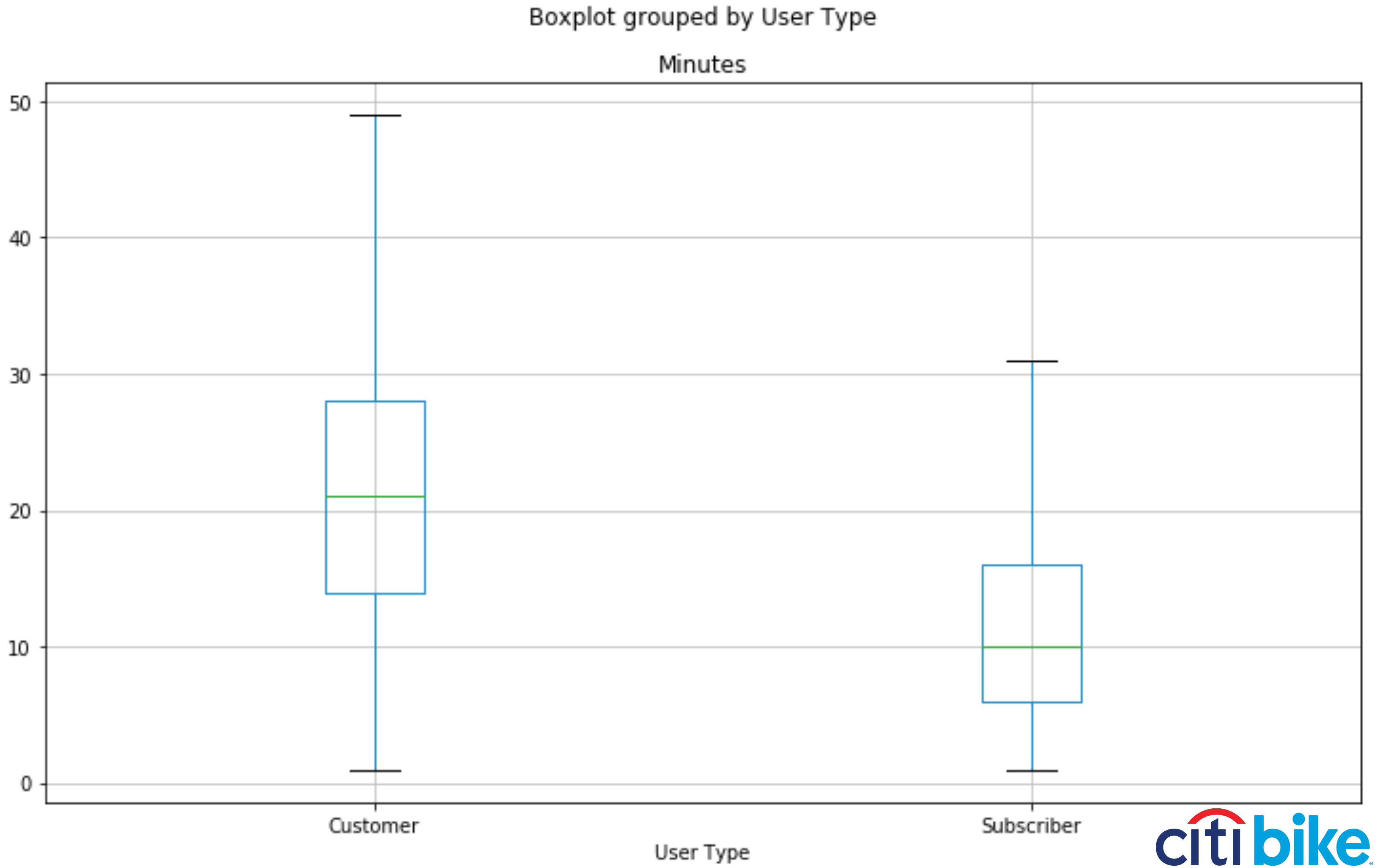
Trip Duration by User Type

Trip Duration for "Customer"

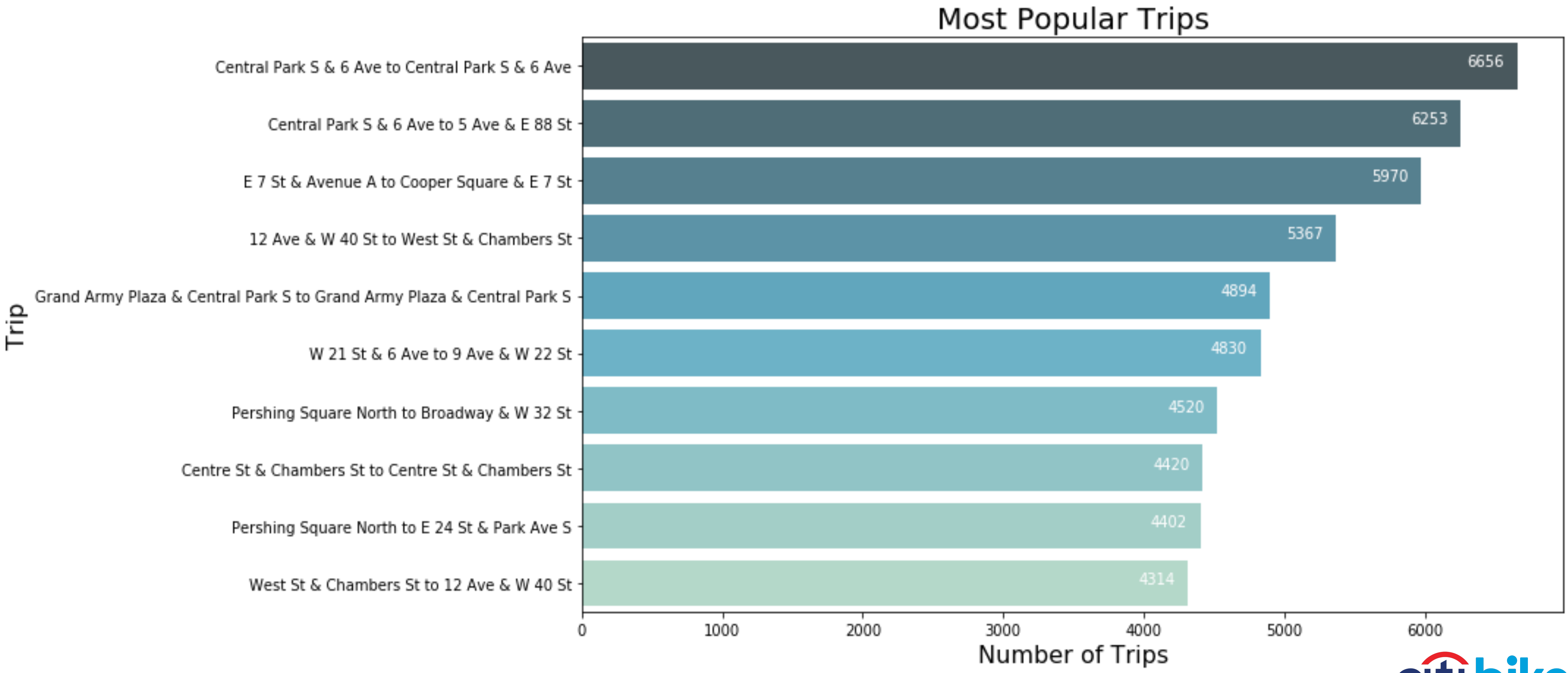
- Customers, who may normally be tourists , tend to use Citi Bike longer

Trip Duration for "Subscriber"

- Subscribers, who are most likely NYC residents, tend to ride for less time
- They most likely have standard routes and have identified the fastest route to work.



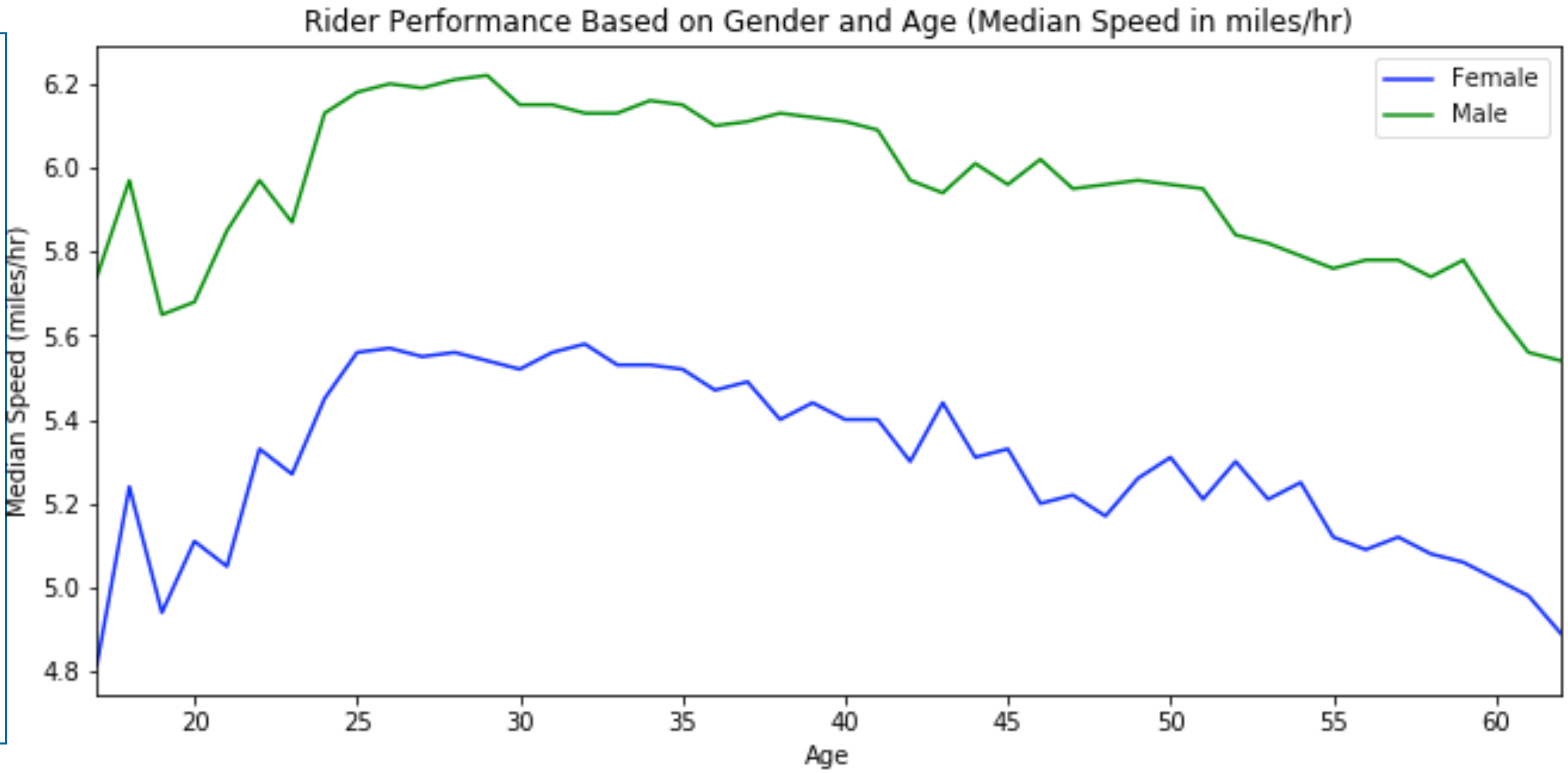
Most Popular Citi Bike Trips in New York City



Rider Speed Performance Based on Gender & Age

Rider Performance Based on Gender & Age:

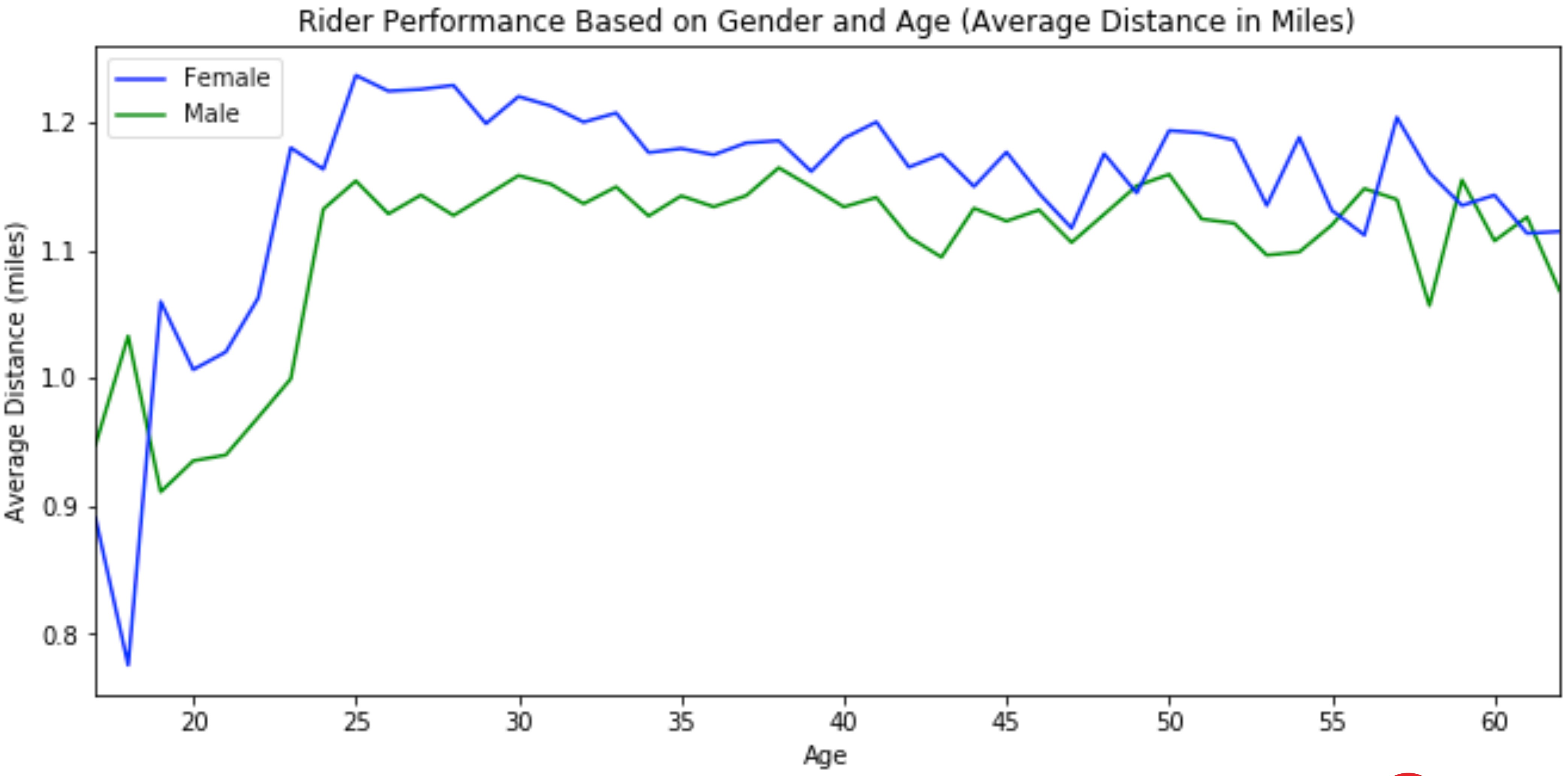
- Males tend to ride faster than females
- Could be explained by the fact that females ride cautiously and tend to stick to bike lanes
- There isn't a drastic difference between speed and age
- There's a slight increase, but it's negligible.



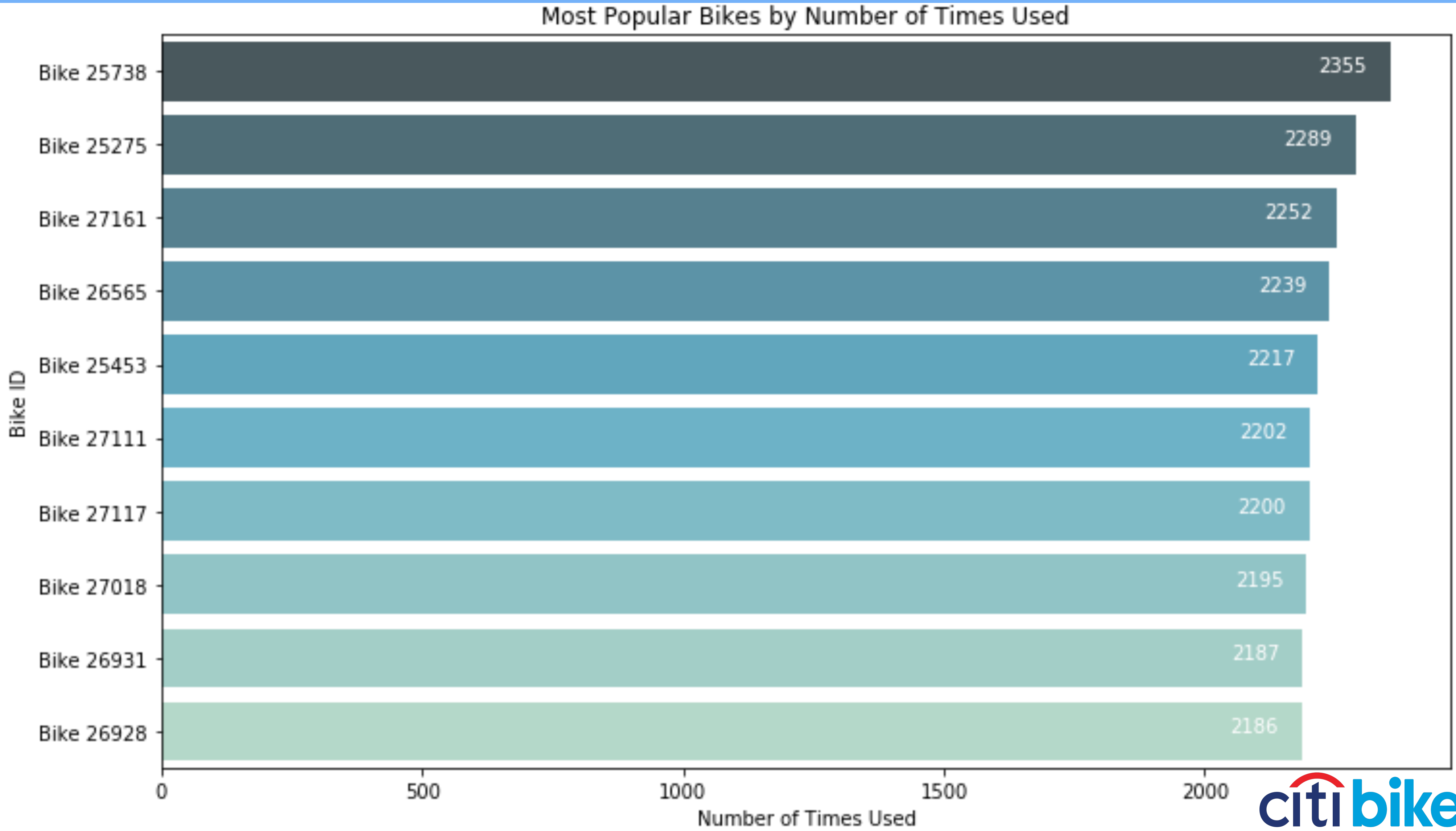
Rider Distance Based on Gender & Age

Rider Distance Based on Gender & Age:

- Females tend to travel further distances than males
- The difference in distance is negligible and could vary year by year
- On average older riders travel further distances
- Age correlation with distance could be due to the fact that younger riders bike long as well as a lot of short distances



Busiest Citi Bikes



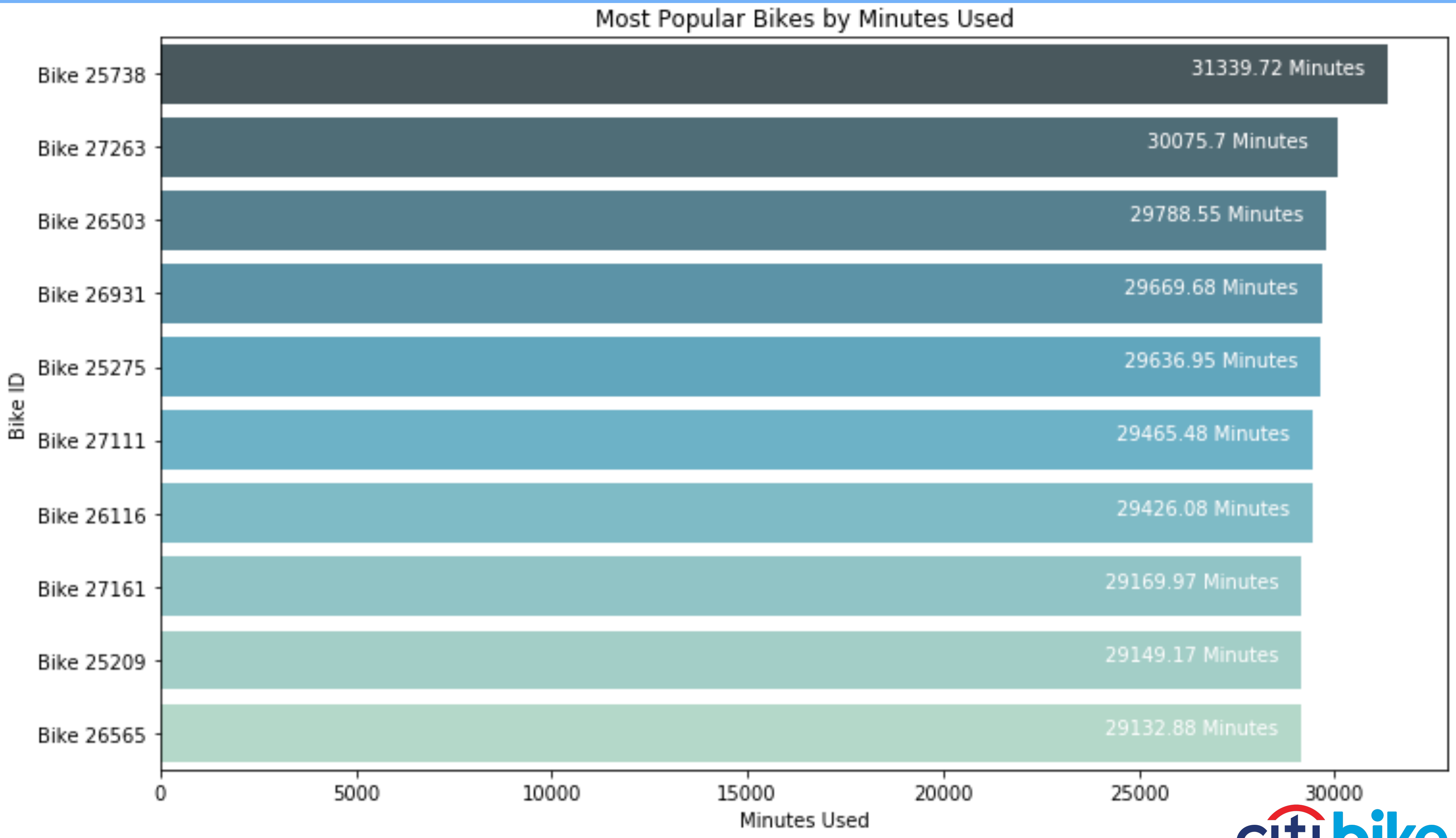
Busiest Bike Based on Minutes and Use:

- Bike 25738

Number of Times Used:

- 2,355 times

Busiest Citi Bikes



Busiest Bike Based on Minutes and Use:

- Bike 25738

Number of Minutes Used:

- 31,340 Minutes

Background & Context

Data Exploration and Visuals

Data Preparation

Data Modelling

Data Preparation

1

UNDERSTANDING THE KIOSK OF THE FUTURE

- Users will come up to the kiosk, swipe their Citi Bike fob, enter their start and end stations
- Based on the information from their Citi Bike fob and trip information, the kiosk will inform the rider how long they should expect the trip to take

2

WHY PREP THE DATA

- To develop the new feature for our Citi Bike kiosks, we will be using machine learning techniques. To be able to create a model which can accurately predict how long a trip will last
- To use these techniques, we need to feed the model the most appropriate data for it to learn from, hence data preparation

3

DATA CLEANED

- Any trip where the start and end station is the same
- Any trip which lasts longer than 45 minutes because no rider would purposefully intend to incur the penalty for going over the time
- Any rider who’s age is above 62
- Any trip where the bikers speed is above 20 miles per hour

Background & Context

Data Exploration and Visuals

Data Preparation

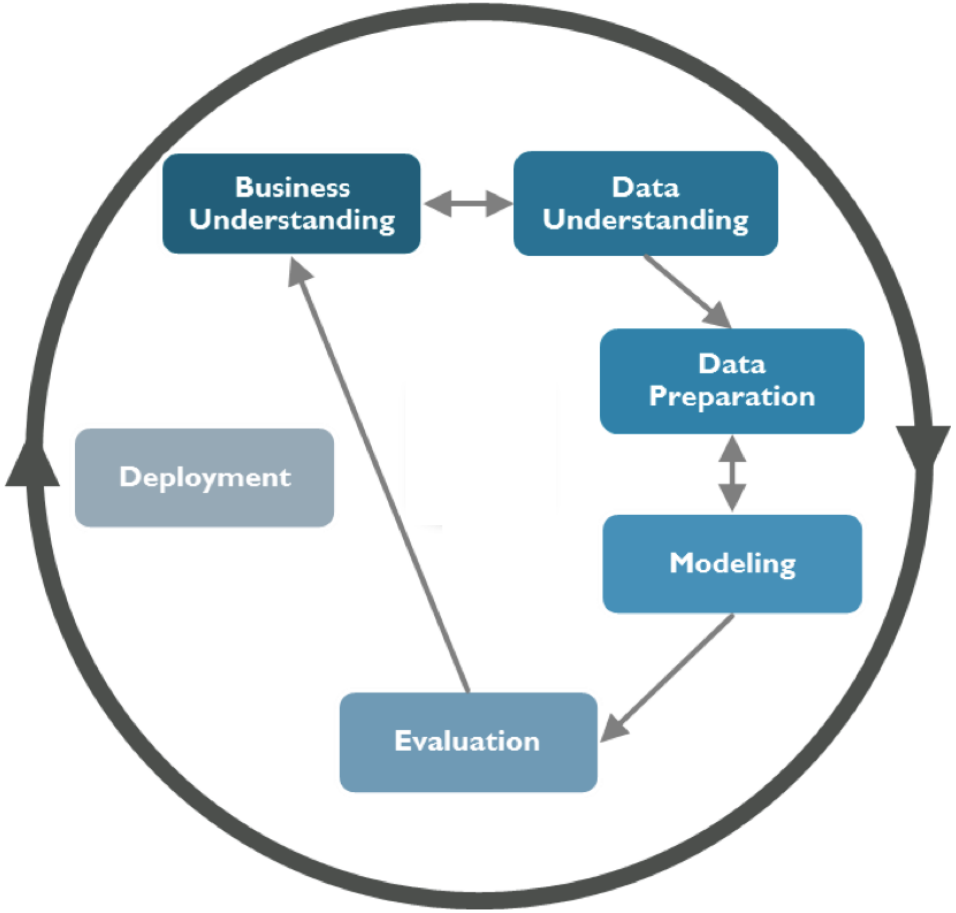
Data Modelling

Data Modelling Methodology

METHODOLOGY

CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project.

1. Understand the business and set objectives
2. Explore, analyze, and understand the data
3. Prepare the data for modelling, also known as data munging or data wrangling
4. Data Modelling
5. Evaluate the model based on objectives
6. Deploy final model after iterative improvements



Data Modelling Baseline Model

BASELINE MODEL

- This data set is extremely large making it difficult to iteratively improve models. Thus, we have decided to take a random yet representative sample of the data
- The baseline model will be based on **Gender, Distance,** and **User Type**

EVALUATING THE MODEL

- The model seems to perform decently well, however, we can't be so far off on our prediction for Citi Bike users
- R-Squared: **0.665**
- Next steps will be to include **Date** and **Time** information

Dep. Variable:	Minutes	**R-squared:**	0.665
Model:	OLS	**Adj. R-squared:**	0.665
Method:	Least Squares	**F-statistic:**	6.043e+05
Date:	Mon, 30 Apr 2018	**Prob (F-statistic):**	0.00
Time:	20:31:24	**Log-Likelihood:**	-3.7093e+06
No. Observations:	1218649	**AIC:**	7.419e+06
Df Residuals:	1218644	**BIC:**	7.419e+06
Df Model:	4		
Covariance Type:	nonrobust		

Data Modelling with Date

DATE BREAKDWON

We decided to add time based on the following format:

1. Is the ride on a **Weekend** or **Weekday**?
    - Weekday commutes are most likely for work
    - Weekend commutes for leisure
2. Is the ride in the **Morning, Afternoon, Evening,** or **Night** based on average trip duration based on hour of the day
    - Morning = **5am– 9am**
    - Afternoon = **9am - 2pm**
    - Evening = **2pm – 8pm**
    - Night = **8pm – 5am**
3. Is the ride in the **Winter, Summer, Fall,** or **Spring**

EVALUATING THE MODEL

- The model performs marginally better, with an R-Squared of **0.670.** Next steps will be to do some feature engineering

Dep. Variable:	Minutes	**R-squared:**	0.670
Model:	OLS	**Adj. R-squared:**	0.670
Method:	Least Squares	**F-statistic:**	2.246e+05
Date:	Mon, 30 Apr 2018	**Prob (F-statistic):**	0.00
Time:	20:37:38	**Log-Likelihood:**	-3.7005e+06
No. Observations:	1218649	**AIC:**	7.401e+06
Df Residuals:	1218637	**BIC:**	7.401e+06
Df Model:	11		
Covariance Type:	nonrobust		

Data Modelling with Feature Engineering

FEATURES ENGINEERED

Speed is extremely useful information. However, we cannot use that information in real-time because the kioskwill not know the speed the rider will bike at. Thus we will impute speed based on the following:

1. Include **Average Speed** based on: **Trip and User Type**
2. Include **Average Duration** for each trip based on: **Trip and User Type**
    - Some trips are up hill, others are down hill. Some routes, such as one through times square involve heavy traffic, based on intuition.
    - Tourists (usually customers), will usually ride more slowly with frequent stops than a Subscriber, according to the data.

EVALUATING THE MODEL

- The model performs marginally better, with an R-Squared of **0.7790.** Next steps will be to include weather data

Dep. Variable:	Minutes	**R-squared:**	0.779
Model:	OLS	**Adj. R-squared:**	0.779
Method:	Least Squares	**F-statistic:**	3.300e+05
Date:	Mon, 30 Apr 2018	**Prob (F-statistic):**	0.00
Time:	21:08:17	**Log-Likelihood:**	-3.4561e+06
No. Observations:	1218649	**AIC:**	6.912e+06
Df Residuals:	1218635	**BIC:**	6.912e+06
Df Model:	13		
Covariance Type:	nonrobust		

Data Modelling with Weather Data

WEATHER DATA

Weather could be an important predictor in how long someone will bike. There are three primary factors in the weather riders tend to consider:

1. How cold or hot is it?
   1. Captured by **TEMP. MIN**
   2. Captures by **TEMP. MAX**
2. How much rain or snow is there?
   1. Captured by **PRECIPITATION**

EVALUATING THE MODEL

- The model performs marginally better, with an R-Squared of **0.779.** The model does not perform any better and the weather just acts as noise. We also looked at p-values of each variable to come to this conclusion
- Final model will not include weather since it is ineffective

Dep. Variable:	Minutes	**R-squared:**	0.779
Model:	OLS	**Adj. R-squared:**	0.779
Method:	Least Squares	**F-statistic:**	2.684e+05
Date:	Mon, 30 Apr 2018	**Prob (F-statistic):**	0.00
Time:	21:22:21	**Log-Likelihood:**	-3.4556e+06
No. Observations:	1218649	**AIC:**	6.911e+06
Df Residuals:	1218632	**BIC:**	6.911e+06
Df Model:	16		
Covariance Type:	nonrobust		

Final Model

PREDICTORS

1. Distance
2. Gender
3. User Type
4. Time of Day
5. Season
6. Average Duration for Trip Based on Gender
7. Average Speed for Trip Based on Gender

NEXT STEPS

- We highly encourage further investment in developing the new feature for the kiosk
- Although the model is pretty good, it would be even better if integrated with big data tools and the Google Maps API
- Let's try to include more data beyond 2017 to get a better idea of seasonal effect on the data

Dep. Variable:	Minutes	**R-squared:**	0.746
Model:	OLS	**Adj. R-squared:**	0.746
Method:	Least Squares	**F-statistic:**	2.759e+06
Date:	Mon, 30 Apr 2018	**Prob (F-statistic):**	0.00
Time:	21:51:11	**Log-Likelihood:**	-3.5390e+07
No. Observations:	12186496	**AIC:**	7.078e+07
Df Residuals:	12186482	**BIC:**	7.078e+07
Df Model:	13		
Covariance Type:	nonrobust		

THANK YOU

Appendix

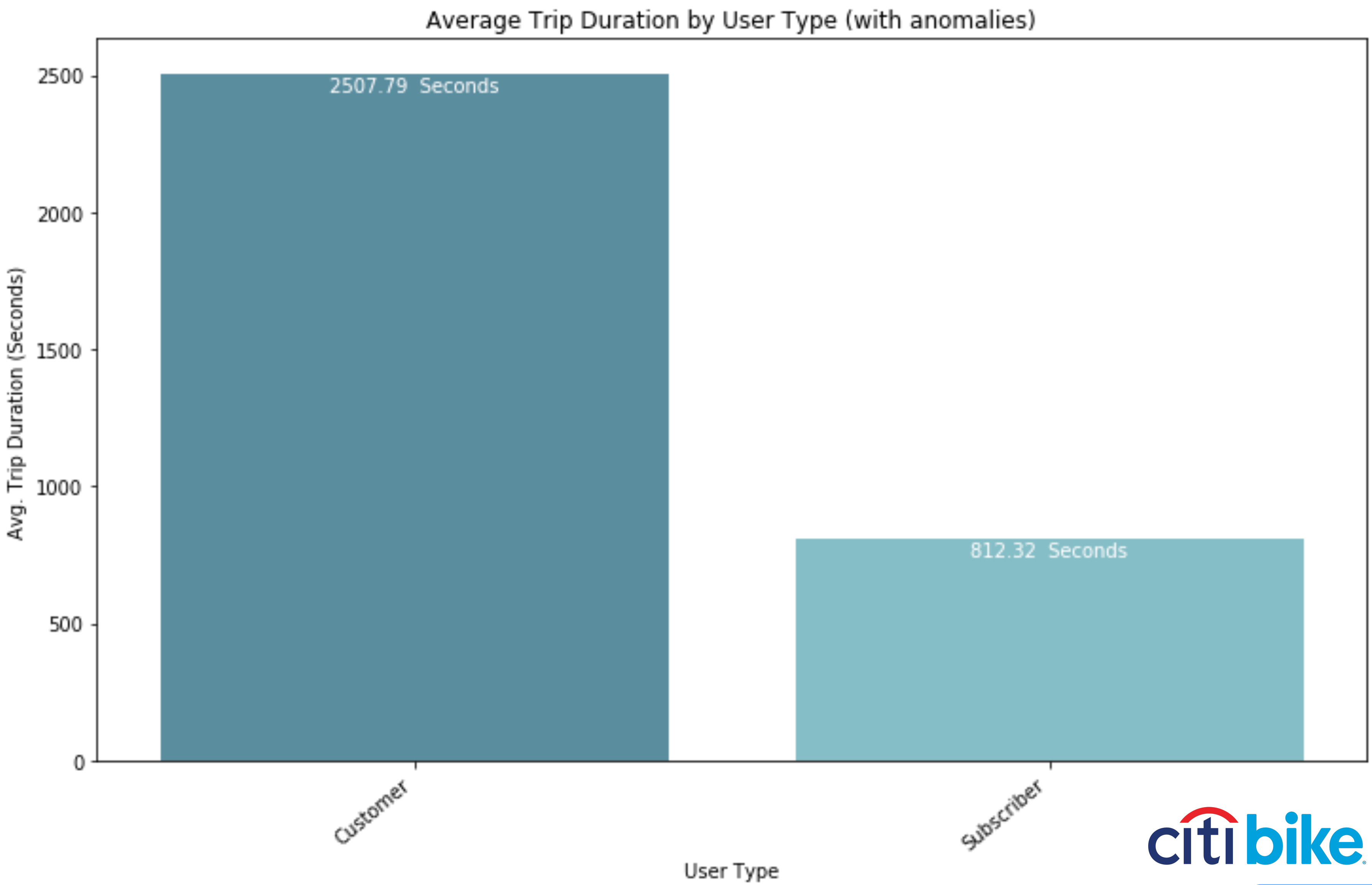
Trip Duration by User Type

Average Trip Duration for "Customer"

- 2507.79 Seconds

Average Trip Duration for "Subscriber"

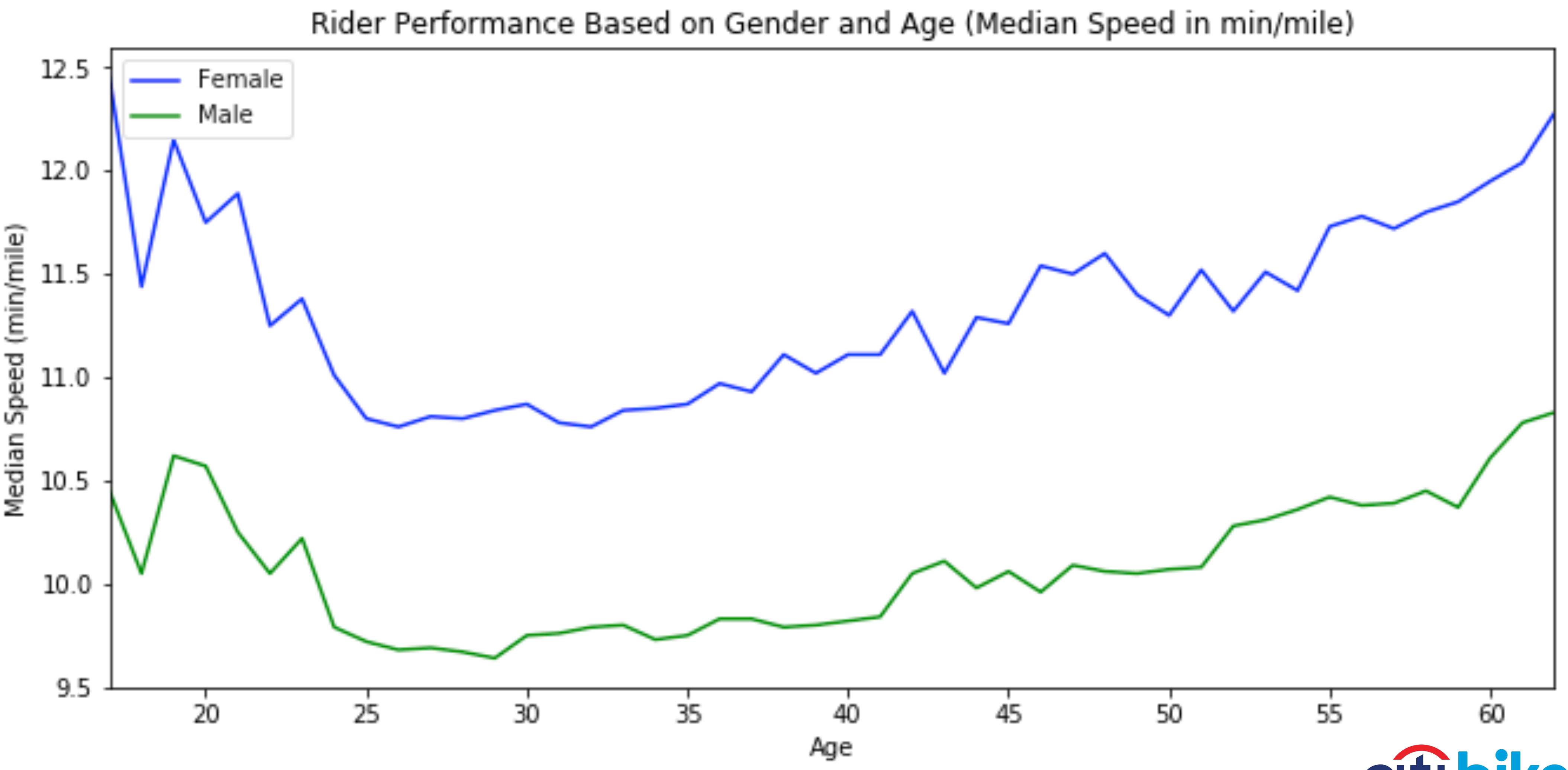
- 812.32 Seconds



Rider Speed Performance Based on Gender & Age

Rider Performance Based on Gender & Age:

- Females tend to take more minutes per mile than males
- There isn't a drastic difference between speed and age. There's a slight increase, but it's negligible.



 IBM Confidential