Citi Bike Investigation

Michael Harder
Brown University - Data Science Initiative
October 25, 2019
https://github.com/Michael-Harder/NYC_Citi_Bike_-Analysis

Agenda

- Introduction
- Preprocessing
- EDA
- Questions

Introduction

Citi Bike	• **New York City public bike share program** launched in an effort to not only reduce traffic, carbon emissions, and roadwear, but also improve public health • Operational since 2013 • Via the **NYC Open Data** initiative the city has publicly published various data sets including **Citi Bike trips data from 2013 to present** ○ Data for this project was collected via www.citibikenyc.com/system-data
Problem	• Like any customer based business model, Citi Bike can benefit from understanding more about their **customers' behavior** • Citi Bike trips data can illuminate how **annual subscription riders differ from 24-hour or 3-day pass riders** • Applying **machine learning classification** we can predict if a **trip was conducted by a subscription rider or an everyday customer** - providing an interesting lense into how their behaviors differ

Agenda

- Introduction
- Preprocessing
- EDA
- Questions

Preprocessing – initial investigating

- Dataset:
  - Limited data to trips from August 2017 to August 2019
  - 759807 rows of trips data by 12 columns
    - Feature columns included - start time, end time, trip duration, start station name, end station name, start station longitude, start station latitude, end station longitude, end station latitude, user type, birth year, gender
- Initial Cleaning:
  - Dropped the following columns
    - Start Station ID - data set includes start station name. ID used for internal purposes
    - End Station ID - data set includes end station name. ID used for internal purposes
    - Bike ID - ID number used for internal purposes
  - Start time and end time:
    - Provided as strings in format “yyyy-mm-dd hh:mm:ss.ssss”
    - Trimmed this string to get the time and converted it to seconds from start of day so it could be preprocessed with standard scaler as a float64

Preprocessing – encoding

One-Hot Encode	• Applied to **categorical** variables:
	  ○ Start station name
	  ○ End station name
	  ○ gender

Standard Scaler	• Applied to **continuous** variables:
	  ○ Trip duration
	  ○ Start station longitude
	  ○ End station longitude
	  ○ Start station latitude
	  ○ End station latitude
	  ○ Start time
	  ○ End time
	  ○ Birth year

Label Encoder	• Applied to the **categorical target variable:** ○ User type

Preprocessing – missing values

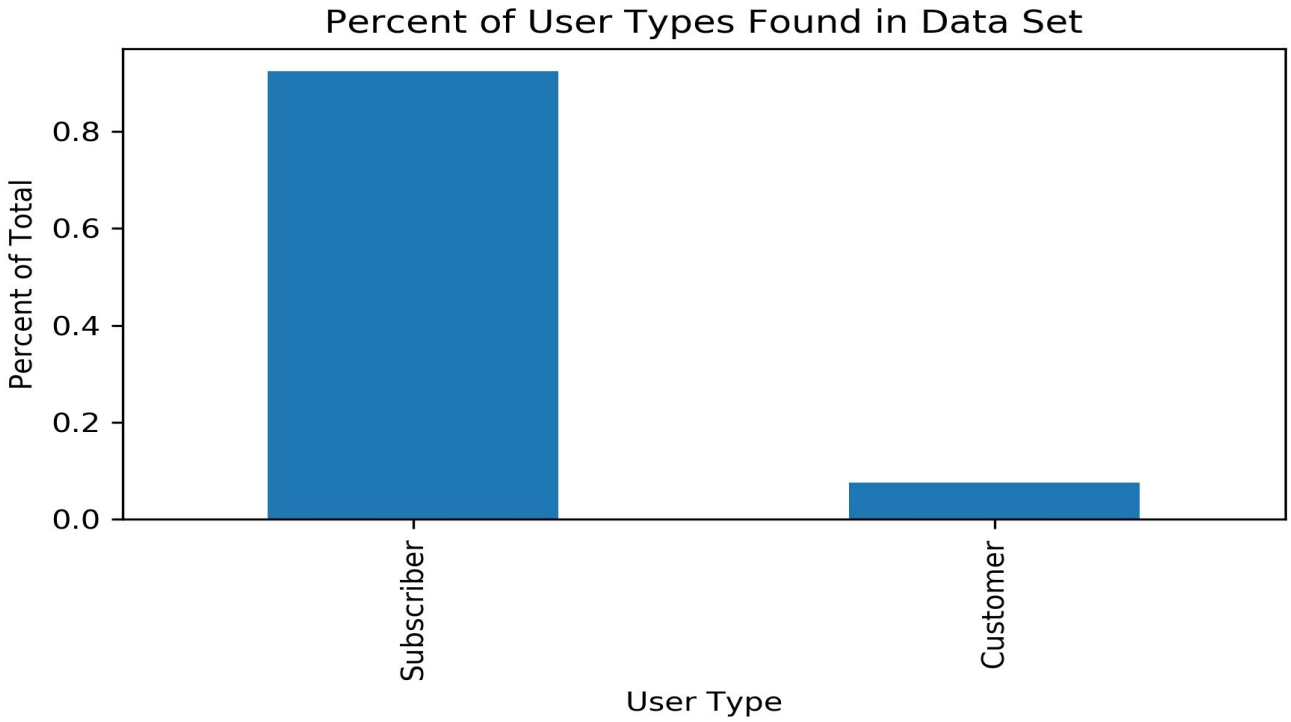
- Before standard scalar was applied there were missing values to consider:
  - 1.12% of rows contained missing data
  - The only feature containing missing data was Birth Year
- MCAR test was applied to investigate the MCAR p value
  - Received error Andras “has never seen before”
- Considering this small percentage of points with NaNs, the small fraction of Nans in each feature, and the difficulties with the MCAR test I dropped the rows with missing values
  - Note: this was sanctioned per Andras

Agenda

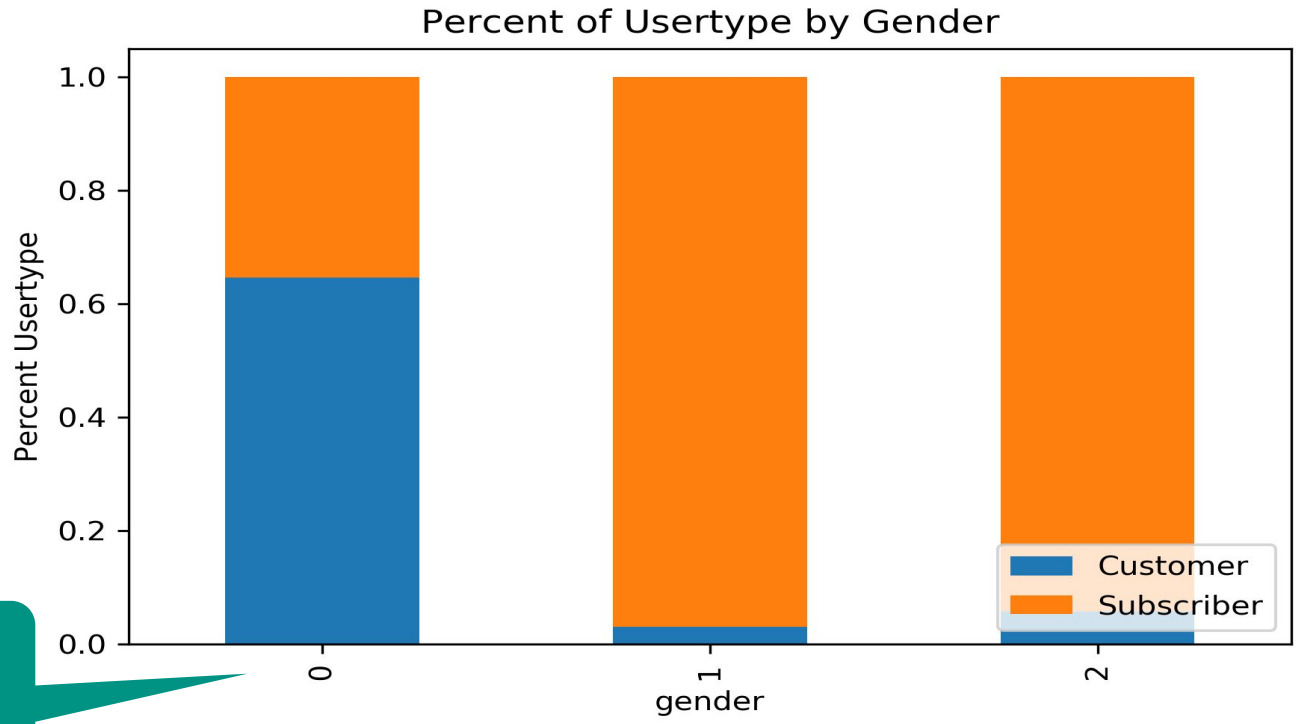
- Introduction
- Preprocessing
- EDA
- Questions

EDA

Dataset seems to be unbalanced



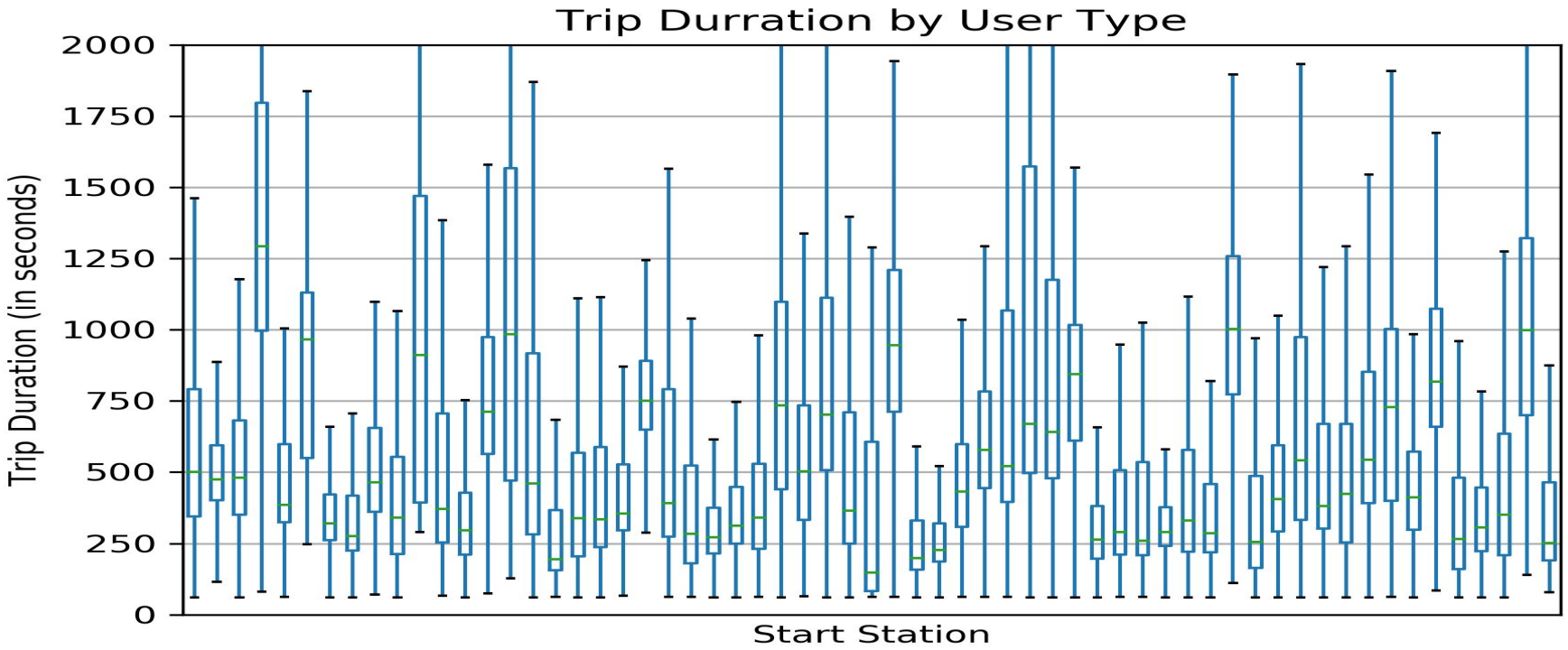
EDA
Less information is known for everyday customers



0 = unknown
1 = male
2 = female

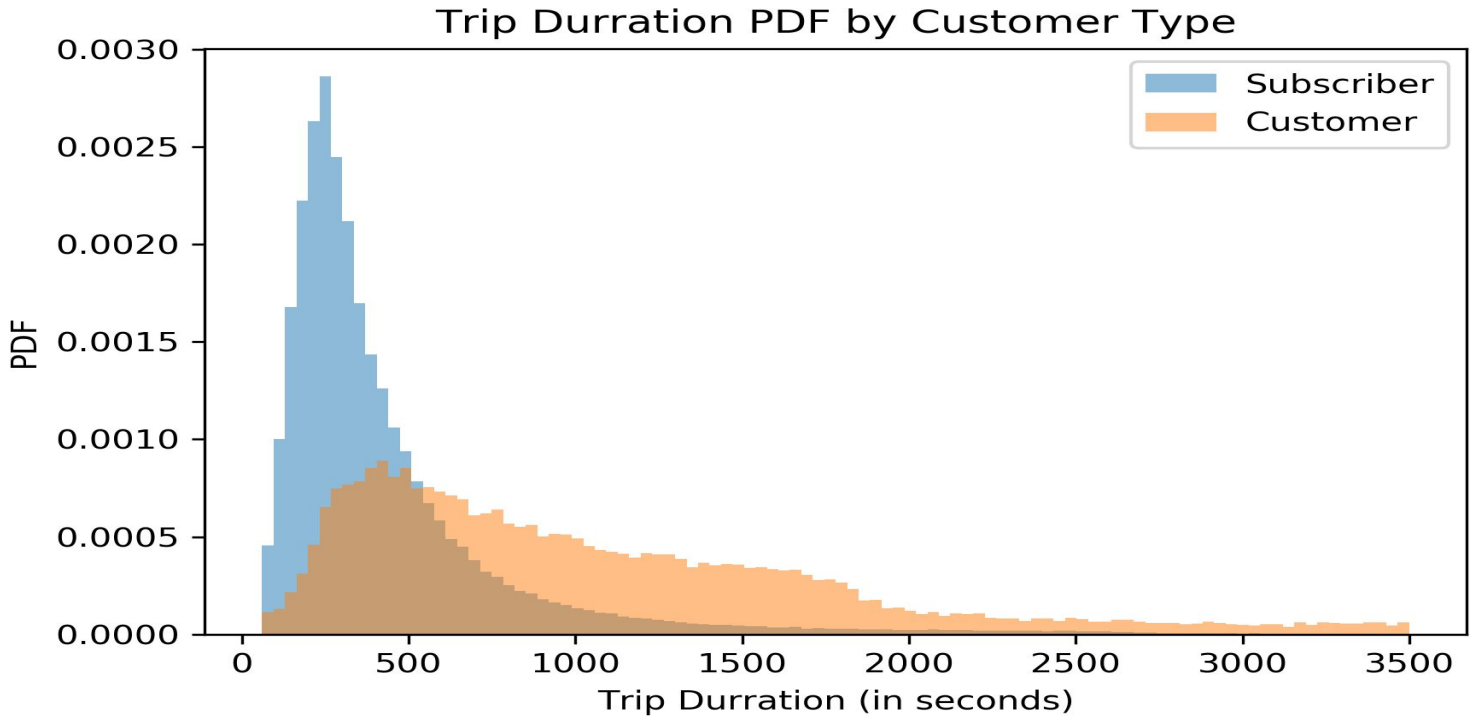
EDA

Start station may be correlated to trip duration



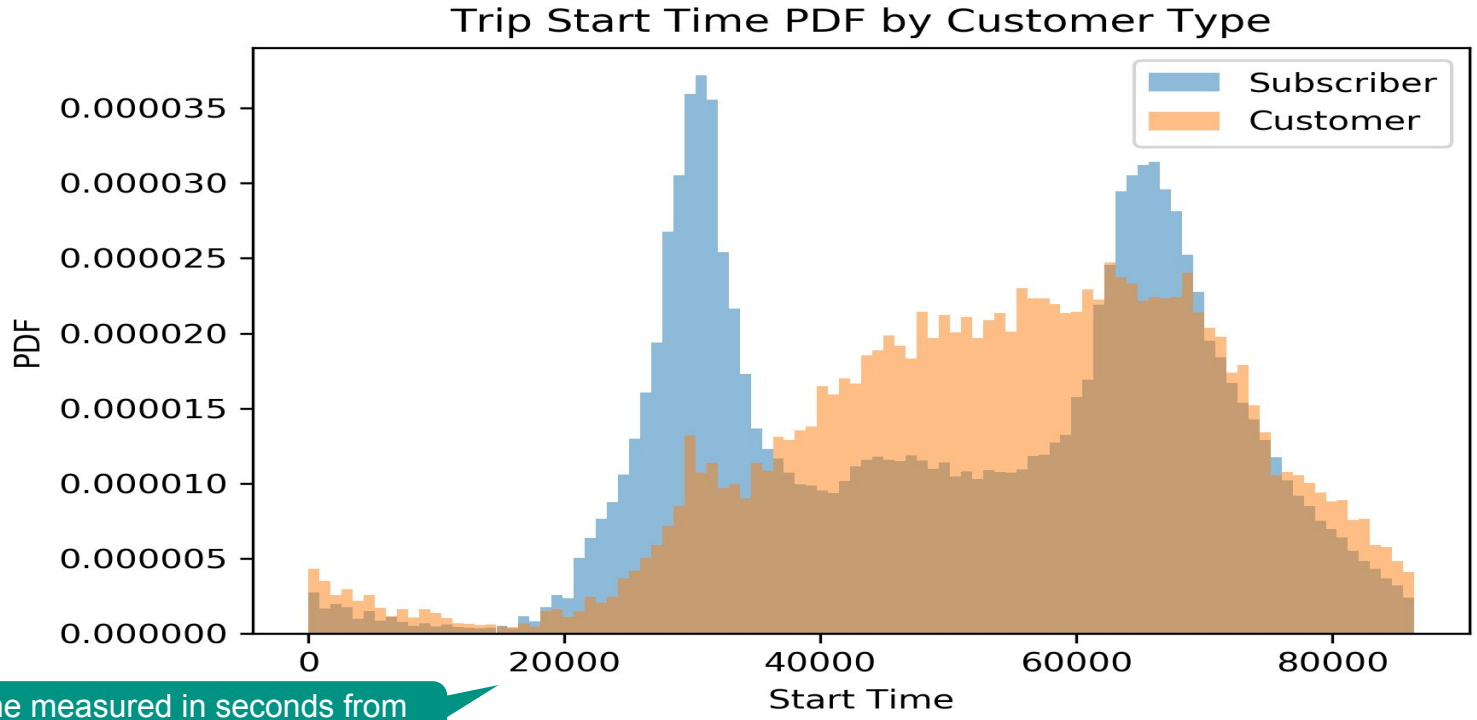
EDA

Subscribers tend to take shorter trips



EDA

Subscribers' start times have two peaks over the course of the day - this could represent commuter behavior



Start time measured in seconds from the start of each day (midnight)

Agenda

- Introduction
- Preprocessing
- EDA
- Questions