Ford GoBike System Data Exploration¶

Dataset

This data set includes information about individual rides made in a bike-sharing system for the year 2019. There are 183412 rows × 16 columns in total.In this notebook we'll be exploring and visualising this dataset. The data set was gotten from https://www.fordgobike.com/system-data

There are 183412 rows × 16 columns in total.In this notebook we'll be exploring and visualising this dataset which contains members, users types, users age and gender, stations of starting and ending trips, duration of trips..

After cleaning,wrangling and analysing the data I was able to extract the age,dates,days of the week, e.t.c for easier analysis.

Summary of Findings

- There are more male riders in the dataset
- Most 5am bike riders were subscribers.
- There were more rides between 4pm and 12am this could be due to numerous reasons,one being 'office closing hours' e.t.c
- Thursdays were busy days and there were more subscribers sharing rides as opposed to customers.
- Weekends had the least number of rides.This could be attributed to different factors one being, weekends are resting days.
- Ages 32-35 show a maximum indulgence in the sport.
- Customers had the highest mean rides.
- 'Market St at 10th St' was the most frequented end station with a count rate of 3709 rides for afternoon riders.. This station was also used as a start station at some point too.
- 16:00pm-6:00pm tend to be the busiest hours.
- Customers took longer trips than subscribers.
- There are no kids present in the dataset as most of the riders age start from 20years old.

Key Insights for Presentation

In this presentation, the influence of the data variables and how they interact i.e user_types, gender, member_age, trip duration can be seen from different plots.I started off first by plotting a standard scaled plot of the member_ages to see it’s distribution before extending my analysis to other variables.

Using a countplot to visualise my categorical variables, i was able to see the effect of each categorical column on the data set.Histograms were implemented to see the the distribution of member’s _age per user_type to know the distribution of ages of each user type in the dataset. Scatterplots were also used to show the relationship between numerical columns, then I introduce each of the categorical variables one by one.