Part_II_slide_deck_template

August 12, 2022

1 Part II - FordGoBikes Analysis.

1.1 by (Tamuno-omi Jaja)

1.2 Investigation Overview

Ford GoBikes is a Bike sharing system that was renamed Bay Wheels in June 2019, after its acquisation by Lyft, It is based in the San Francisco Bay Area, California. In this investigation, I've looked at the bike ride trends and biker type of the bay Ford GoBike Share system. The main interest was to understand the relationship between biking duration & other trip information with focus on distance, the time (weekday, hour), and the user types.

1.3 Dataset Overview

The Ford GoBike System data includes information about individual rides covering the greater San Francisco Bay area, The dataset contains 183,412 individual rides and 16 features that outline customer information, the ride duration and start-end destination.

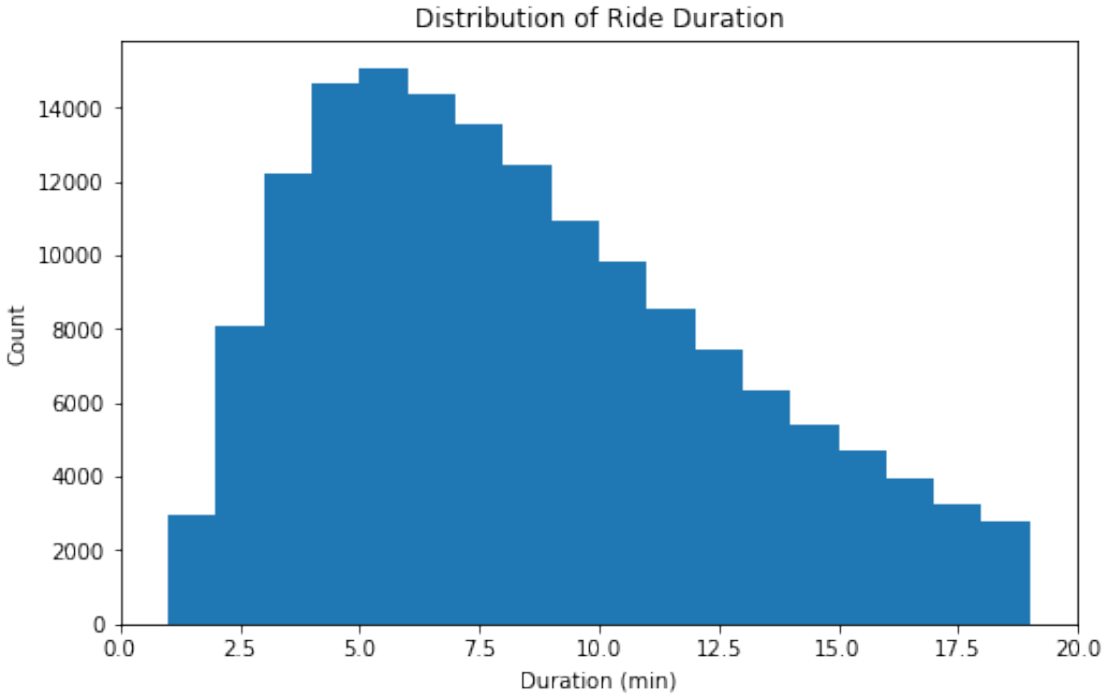
In [1]:
```
# import all packages and set plots to be embedded inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb

%matplotlib inline

# suppress warnings from final output
import warnings
warnings.simplefilter("ignore")
```

In [2]:
```
# load in the dataset into a pandas dataframe
bike_df = pd.read_csv('bike_trips_clean')
```

1.3.1 Distribution of Ride Duration (mins)

From the Visualization we can see that most rides are between three to twelve minutes and that Duration(min) is Right Skewed, pointing to a user preference for shorter trips.

```
In [3]: # investigating  Ride duration with  smaller bin size
        binsize = 1
        bins = np.arange(0, 20, 1)

        plt.figure(figsize=[8, 5])
        plt.hist(data = bike_df, x = 'duration_min', bins = bins)
        plt.xlim([0,20])
        plt.xlabel('Duration (min)')
        plt.ylabel('Count')
        plt.title('Distribution of Ride Duration')
        plt.show()
```

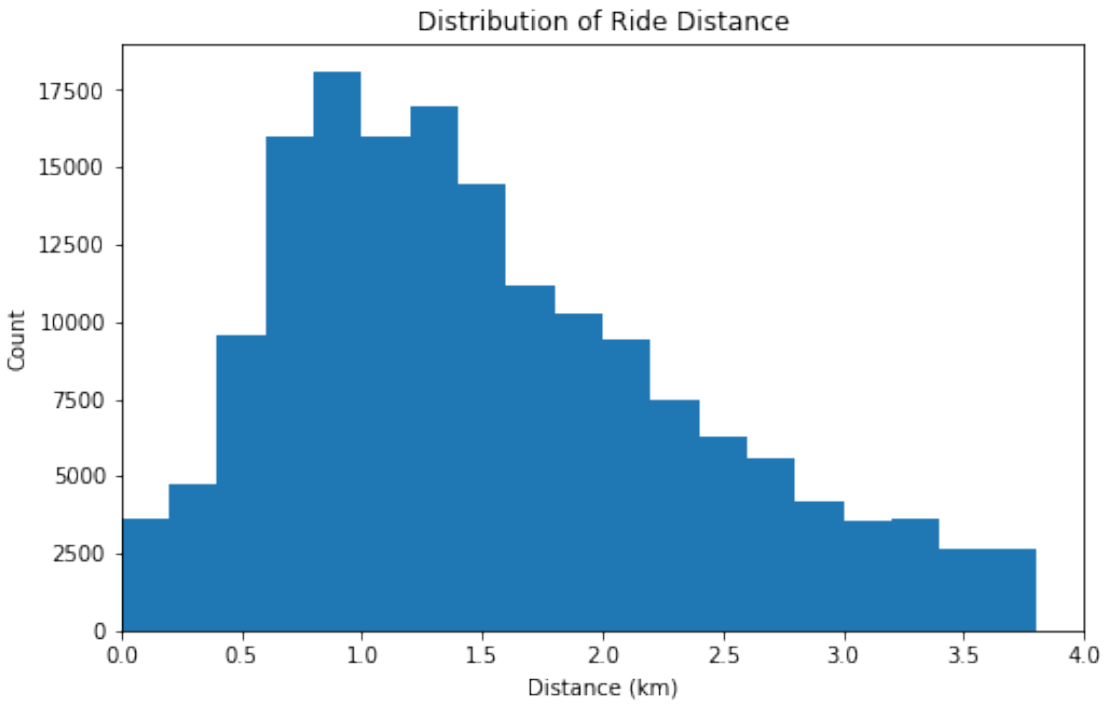


1.3.2 Distribution of Ride Distance (km)

The distance is the length of space between the Start and End Station, the limits of the visual is reduced to exclude outliers, the distribution looks roughly bimodal, with one peak between 0.8 and 1 km, and a second peak a little below 1.5 km. Futher evidence of users taking mainly short trips.

```
In [4]: # investigating distance
        bins = np.arange(0, 4, 0.2)
        plt.figure(figsize=[8, 5])
        plt.hist(data = bike_df, x = 'distance_km', bins = bins)
        plt.xlim([0,4])
        plt.xlabel('Distance (km)')
```

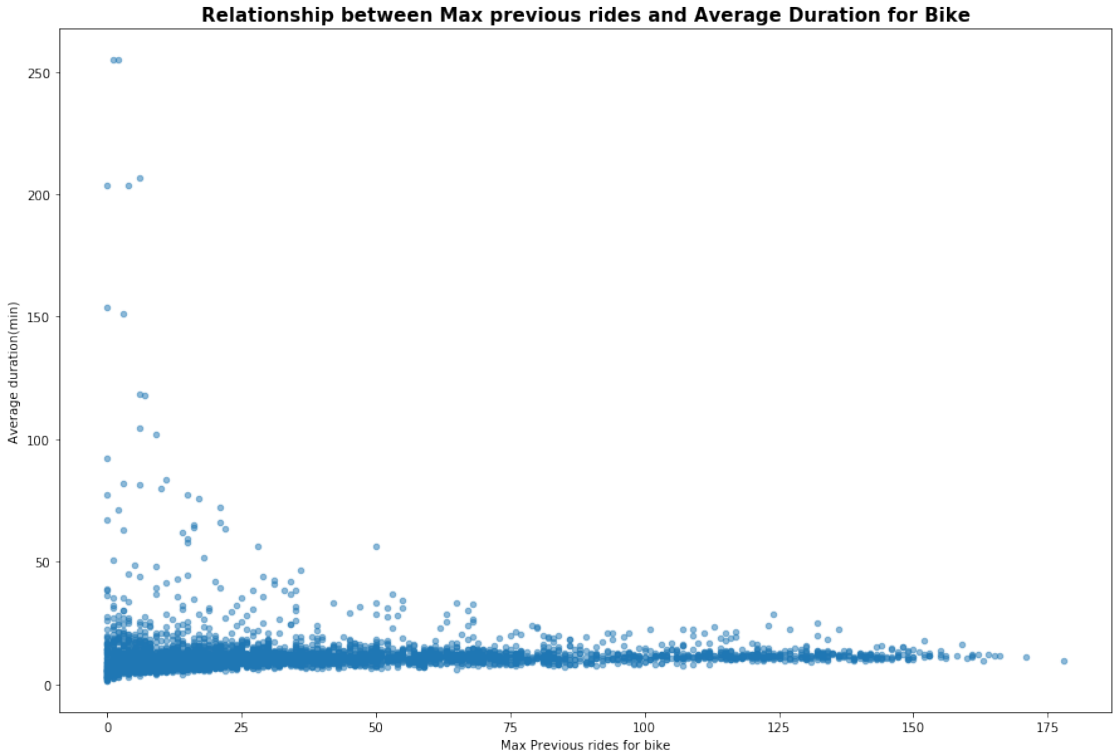
```
plt.ylabel('Count')
plt.title('Distribution of Ride Distance')
plt.show()
```



1.3.3 Max previous rides and Average Duration for Bike

It was observed that only bikes with less that 25 previous rides were able to accomplish ride duration of above 50 minutes, this could point to a maintainance issue on bikes as they get older.

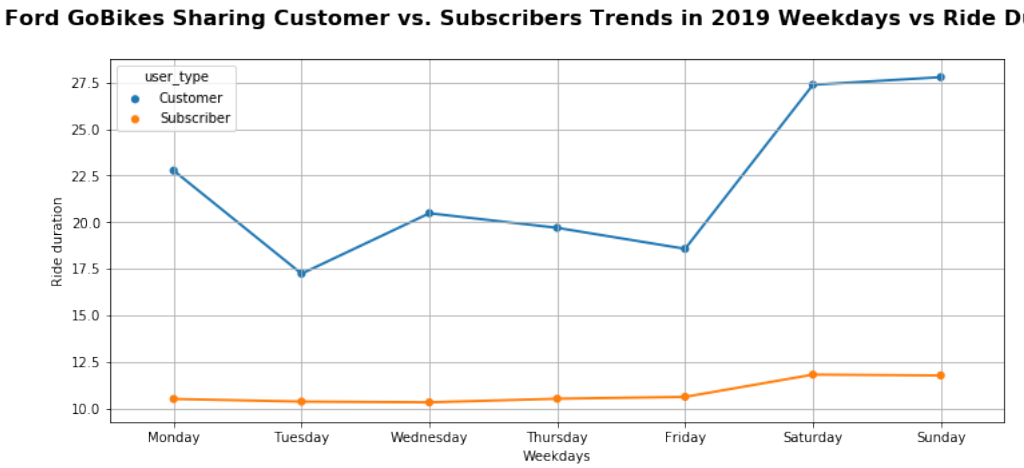
In [5]:
```
# Get max_number of previous ride taken by each bike
bike_rides = bike_df.groupby(['bike_id'], sort=False)['previous_bike_rides'].max().reset

# Relationship between Max previous rides and Average bike Duration
bike_duration = bike_df.groupby('bike_id')['duration_min'].mean().reset_index()
bike_info = bike_rides.merge(bike_duration, left_on = 'bike_id', right_on = 'bike_id', h
bike_info.head()

#scatterplot
x = bike_info.plot(kind = 'scatter', x = 'max_previous_rides', y = 'bike_avg_duration_mi
x.grid(False)
plt.xlabel('Max Previous rides for bike')
plt.ylabel('Average duration(min)')
plt.title('Relationship between Max previous rides and Average Duration for Bike', weigh
```



1.3.4 Ford GoBikes Sharing Customer vs. Subscribers Trends in 2019 Weekdays vs Ride Duration

Ride duration for both Member groups(Customer & Subscriber) peaks during weekends, probably because members aren't in a rush to get to their destination, unlike normal working days.

```
In [6]: plt.figure(figsize=(12, 5))

        df_cleaned_user_week = bike_df.groupby(['day', 'user_type'])['duration_min'].mean().rese
        weekday = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']

        ax = sb.pointplot(data=df_cleaned_user_week, x='day', y='duration_min', hue = 'user_type


        plt.title('Ford GoBikes Sharing Customer vs. Subscribers Trends in 2019 Weekdays vs Ride
        plt.xlabel('Weekdays')
        plt.ylabel('Ride duration');
        plt.grid()
        plt.show()
```

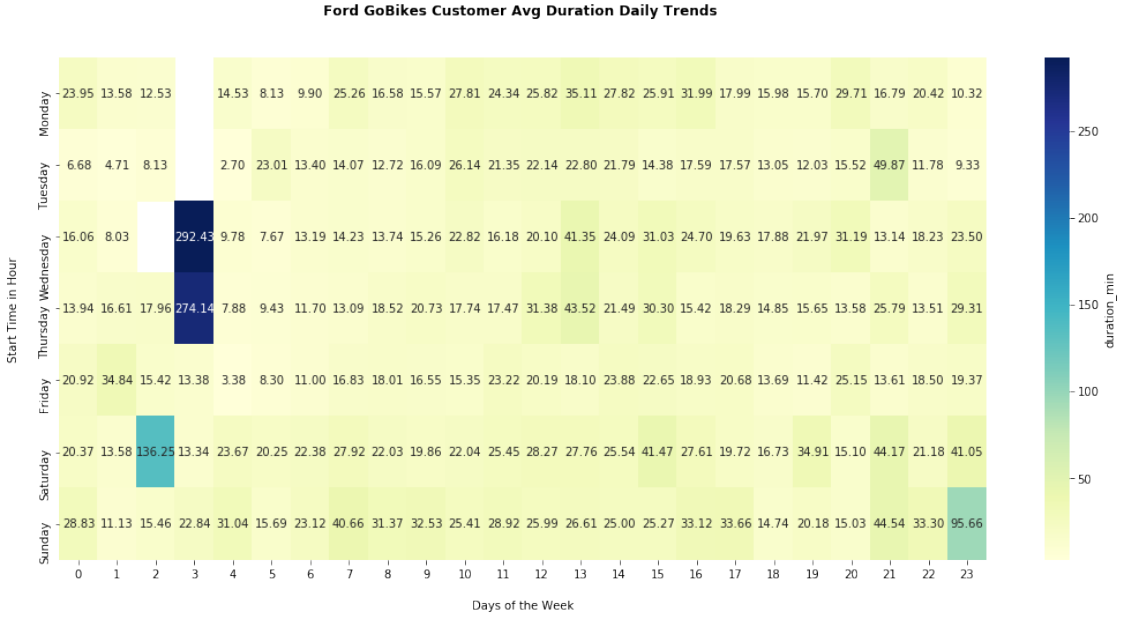


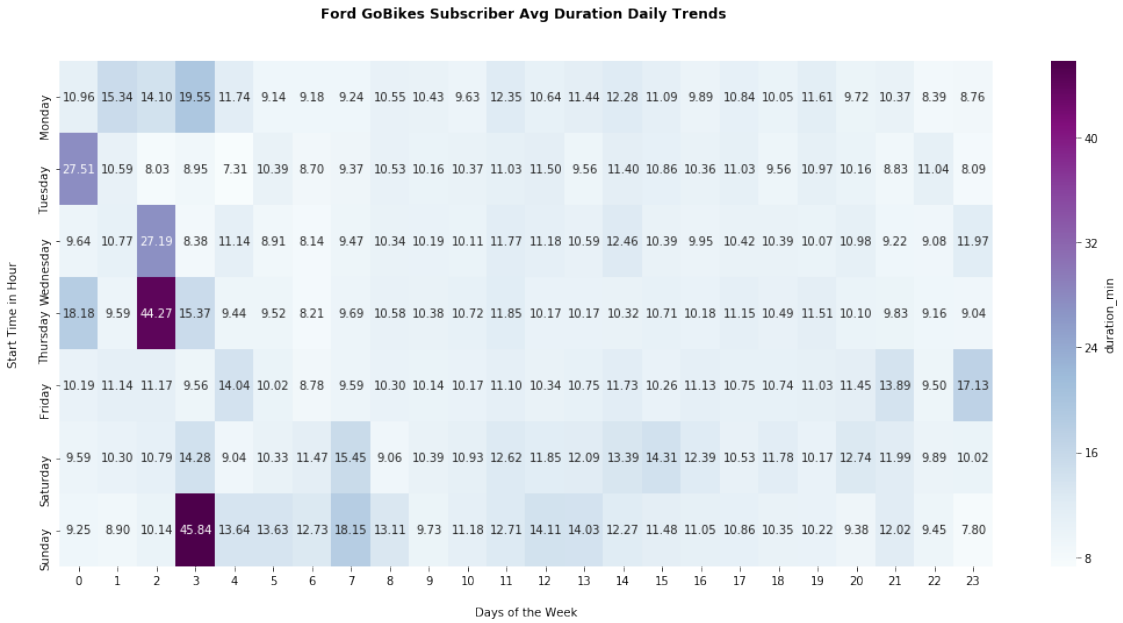
1.3.5 Duration Trends

In regards to the ride duration both groups saw a peak on Wednesday and Thursday for between 3 to 4 am for Customers and 2 to 3 am for Subscriber

```
In [11]: def plot_heat_map(df,group,variable,color="YlGnBu",precision = 2):
             '''df - dataframe
                group - user_type(Customer, Subscriber)
                variable - Feature of interest(duration_min, distance_km)
                color - The mapping from data values to color space

                returns - heat map plot segmented by group, of days and hours by variable
             '''
             # Set plot dimensions
             plt.figure(figsize = [18,8])
             # Select group
             bike_df = df.query('user_type == @group')
             # filter for columns of interest
             var_filter = bike_df.filter(items=['hour', 'day', variable])
             # Get average distance for day and hour
             var_filter = var_filter.groupby(['day','hour'])[variable].mean().reset_index()
             # Create pivot table
             var_trend = var_filter.pivot("day","hour",variable)
             # Sort index
             var_trend.index = pd.CategoricalIndex(var_trend.index, categories= ["Monday", "Tues
             var_trend.sort_index(level=0, inplace=True)
             # Plot heat_map
             ax = sb.heatmap(var_trend,cmap = color, annot = True, fmt = f".{precision}f",cbar_k
             return ax
```

```
# Customer Avg Duration Daily Trends

plot_heat_map(bike_df,'Customer','duration_min')
plt.title('Ford GoBikes Customer Avg Duration Daily Trends ', y=1.07, fontweight='bold'
plt.xlabel('Days of the Week', labelpad = 17)
plt.ylabel('Start Time in Hour',  labelpad = 17);

# Subscriber Avg Duration Daily Trends
plot_heat_map(bike_df,'Subscriber','duration_min','BuPu')

plt.title('Ford GoBikes Subscriber Avg Duration Daily Trends ', y=1.07, fontweight='bol
plt.xlabel('Days of the Week', labelpad = 17)
plt.ylabel('Start Time in Hour',  labelpad = 17);
```





1.4 Summary of Findings

There are two Client groups using the Ford GoBike service, namely Customers and Subscribers. Here are some of the key insights I ganered:

1. As envisaged Ride Duration is related to Ride distance, but another features which determines the duration of a trip is the previous usage of the bike, It was observed that bikes with less that 25 previous rides were able to accomplish ride duration of above 50 minutes, as a result were the only category of bikes to cover above 3km distance for trips

2. Both member groups see peak ride duration on Wednesday and Thursday for trips that started between 2 to 4 am.

3. FordGo members take more trips on weekdays which peak on thursdays and falls on weekends, in addition they cover a higher distances on weekdays for trips that started between 5 to 9 am.

1.4.1 Generate Slideshow

Once you're ready to generate your slideshow, use the `jupyter nbconvert` command to generate the HTML slide show.

```
In [16]: # Use this command if you are running this file in local
         #!jupyter nbconvert Part_II_slide_deck_template.ipynb.ipynb --to slides --post serve --
```

```
In [ ]:
```