Part_II_slide_deck

November 22, 2022

1 Part II - Ford Go Bike Trip Data

1.1 by Mustafe Abdulahi

1.2 Investigation Overview

In this project investigation, my aim is to create a meaningful key insights from the data we have and perform Exploratory Data Analysis in short (EDA). I am mainly focusing on the frequencies by hours of the day, days of the week and customer type. I want know when most trips occur/take place, what hours of the day, days of the week, and which user types made on these trips and how these variable relate to each other.

1.3 Dataset Overview

This data set contains a single csv file and consists of information about individual bike-sharing system covering the greater San Francisco Bay area. The data features include tripduration (secs), start_time, end_time, user information i.e (user_type, age), and some other variable.

[1]:
```
# import all packages and set plots to be embedded inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib import rcParams
import seaborn as sb
import datetime as dt
from datetime import datetime
plt.style.use('ggplot')
%matplotlib inline
```

[2]:
```
# load in the dataset into a pandas dataframe
df = pd.read_csv("fordgobiketrip_cleaned_data.csv")
df.head()
```

[2]:

	duration_sec	start_time	end_time	\
0	52185	2019-02-28 17:32:10.145	2019-03-01 08:01:55.975	
1	42521	2019-02-28 18:53:21.789	2019-03-01 06:42:03.056	
2	61854	2019-02-28 12:13:13.218	2019-03-01 05:24:08.146	
3	36490	2019-02-28 17:54:26.010	2019-03-01 04:02:36.842	

```
4          1585  2019-02-28 23:54:18.549  2019-03-01 00:20:44.074

                              start_station_name  \
0  Montgomery St BART Station (Market St at 2nd St)
1                      The Embarcadero at Steuart St
2                            Market St at Dolores St
3                            Grove St at Masonic Ave
4                                Frank H Ogawa Plaza

                             end_station_name  bike_id   user_type  \
0               Commercial St at Montgomery St     4902    Customer
1                           Berry St at 4th St     2535    Customer
2  Powell St BART Station (Market St at 4th St)     5905    Customer
3                        Central Ave at Fell St     6638  Subscriber
4                         10th Ave at E 15th St     4898  Subscriber

   member_birth_year member_gender bike_share_for_all_trip       day  hour  \
0             1984.0          Male                      No  Thursday    17
1                NaN           NaN                      No  Thursday    18
2             1972.0          Male                      No  Thursday    12
3             1989.0         Other                      No  Thursday    17
4             1974.0          Male                     Yes  Thursday    23

   dur_per_minute   age age_group
0             869  38.0     Adult
1             708   NaN       NaN
2            1030  50.0     Adult
3             608  33.0     Adult
4              26  48.0     Adult
```

[3]:
```
base_color = sb.color_palette()[1]
```

Note that the above cells have been set as "Skip"-type slides. That means that when the notebook is rendered as http slides, those cells won't show up.

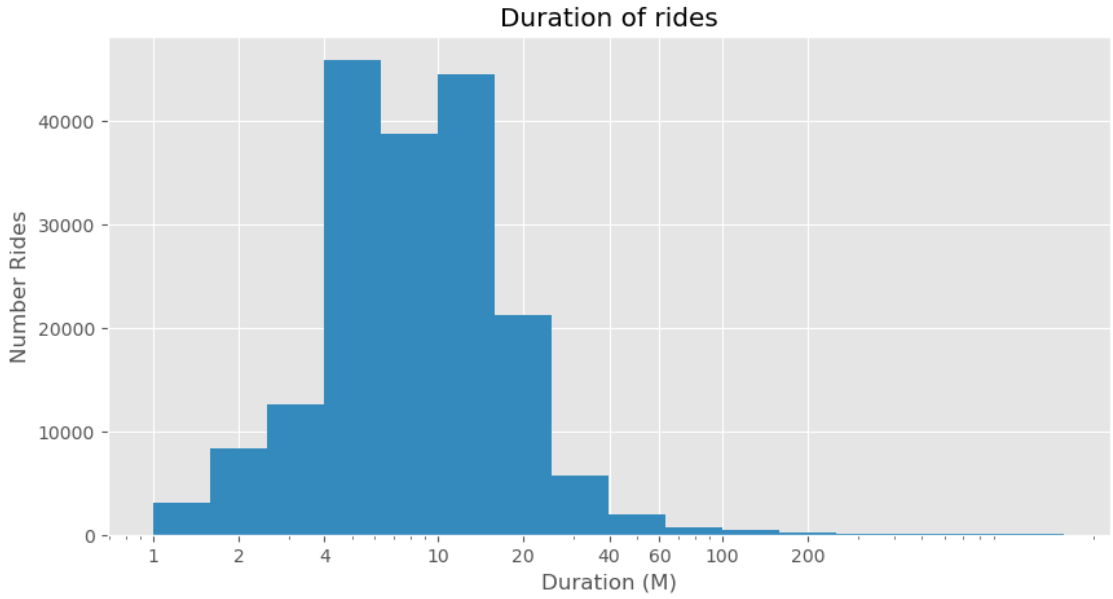
1.3.1 Distribution of Ride Duration

The original trip duration in the data was was measured in Seconds, so I converted into minutes and visualized in the Exploratory section in part I and found there was a long tail of duration distribution, so I have applied to logarithmic scale transformation and used smaller binsize to get a more detailed distribution. As we see in this histogram, most rides took between 8-12 minutes and very few rides lasted more than an one hour (60 minutes). We also confirmed the trip avearge duration is about 12 minutes.

I used proper labels, title and base color to depict the figure.

[4]:
```
# investigation Ride Duration
def histogram():
```

```
    plt.figure(figsize=[10,5])
    ticks = [1, 2, 4, 10, 20, 40, 60, 100, 200 ]
    bin_edges = 10 ** np.arange(0.0, np.log10(df.dur_per_minute.max())+0.2, 0.2)
    plt.hist(data = df, x = 'dur_per_minute', bins = bin_edges,color =␣
 ↪base_color)
    plt.xscale('log')
    plt.xticks(ticks, ticks)
    plt.xlabel('Duration (M)')
    plt.ylabel("Number Rides")
    plt.title("Duration of rides");
histogram()
```



[5]:
```
# check the average trip duratons
df['dur_per_minute'].mean()
```

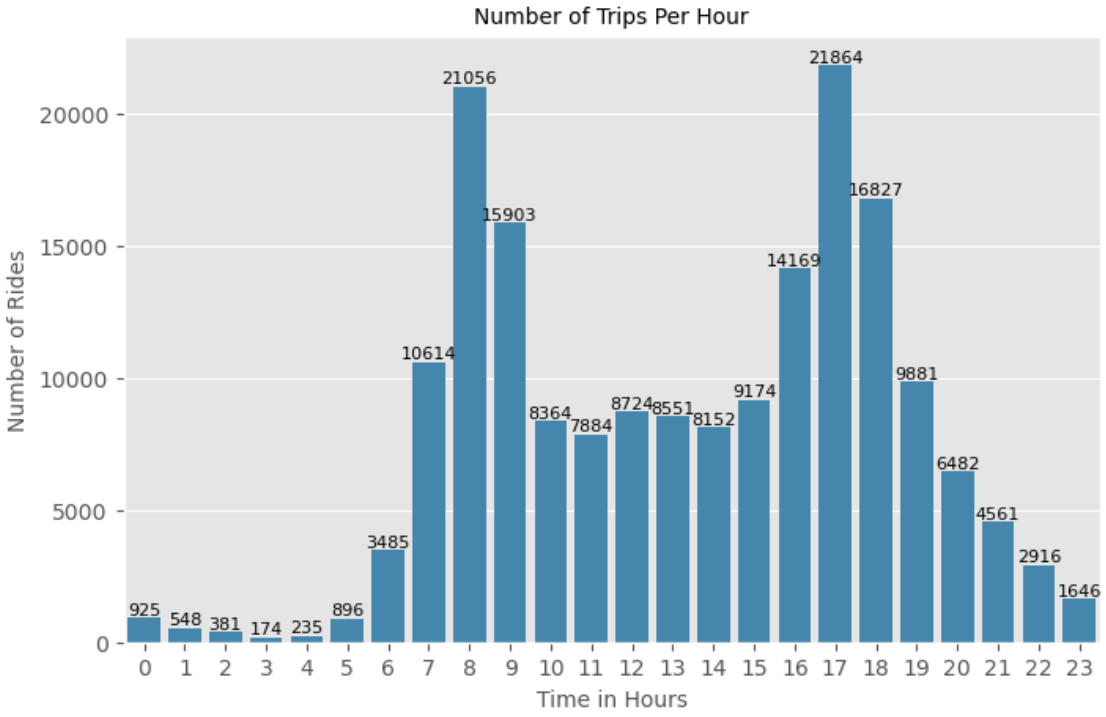
[5]: 11.60939306043225

1.3.2 Distribution of trips per hour

I have used countplot to visualize the total number of rides per hour and added proper labels, descriptive title and base_color.

In this figure, I found that Most trips were taken at commute hours: 8th,9th,17th,and 18th hour. From this insight we can infere that this is because people going work to between 8-9 hr (mornings) and is coming back from work to home at 17-18 hour which is closing work.

[6]:
```
# let take a look at the trip duration per hour frequence
plt.figure(figsize=(8, 5))
myplot = sb.countplot(data= df,
             x=df['hour'].sort_values(ascending=True),
color= base_color)

myplot.bar_label(myplot.containers[0], size = 8)
plt.title("Number of Trips Per Hour", size = 10)
plt.xlabel('Time in Hours', size = 10)
plt.ylabel("Number of Rides", size = 10);
```



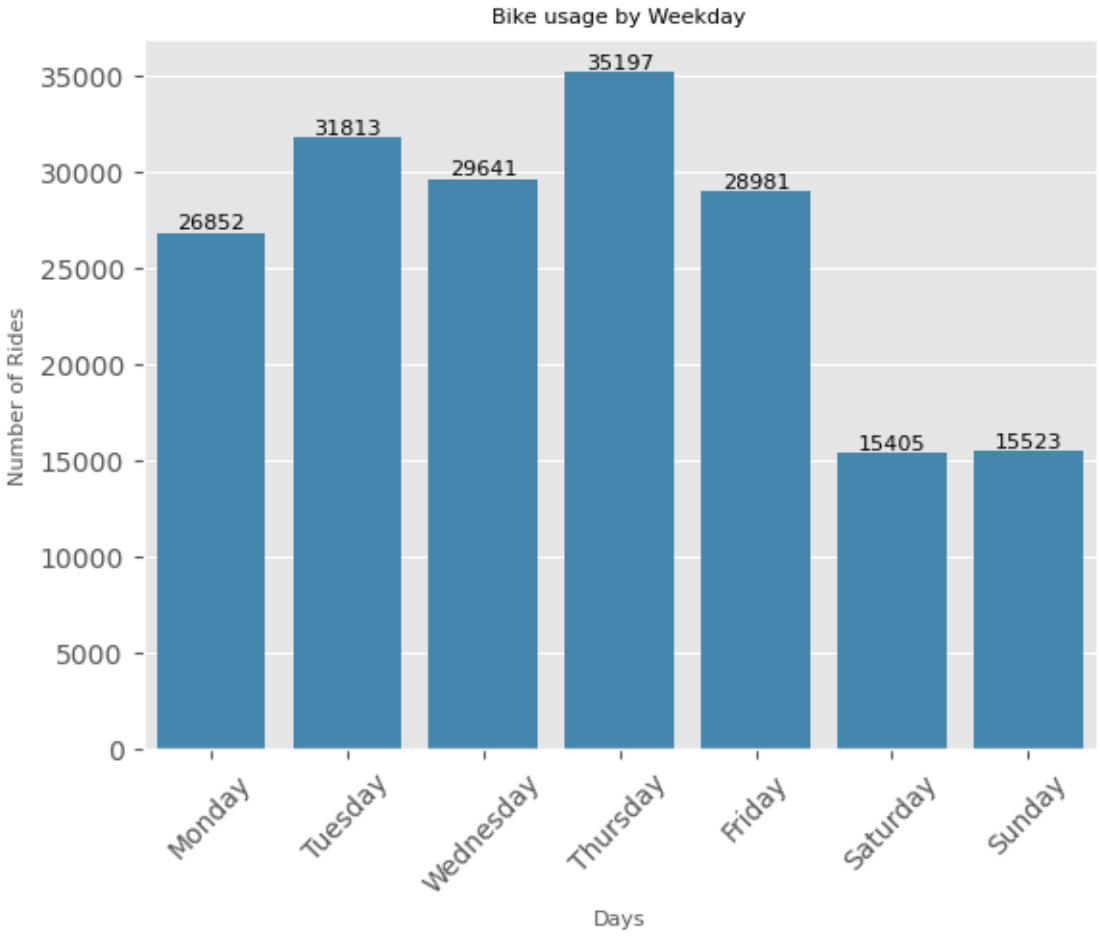
1.4 Bike Usage by Weekday

Similarly, I have choosed countplot to depict the bike use over weekdays. I added proper labels, added descriptive titles, and ordered weekdays manually and and included font size to make it legible.

From this graph, we drived that most trips were taken Thrusday, followed by Tuesday. We also see that weekends (sat and sun) trips are smaller compared to weekday.

[7]:
```
# let take a look at the average trip duration per day frequence
def horizontal_bar():
    order_days =␣
 ↪["Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"]
```

```
    myplot=sb.countplot(x='day',
                        data=df, color=base_color, order= order_days)
    myplot.bar_label(myplot.containers[0], size = 8)
    plt.xticks(rotation=46)
    plt.title("Bike usage by Weekday", size = 8)
    plt.xlabel("Days", size = 8)
    plt.ylabel("Number of Rides", size = 8)
horizontal_bar()
```



2 Age distribution

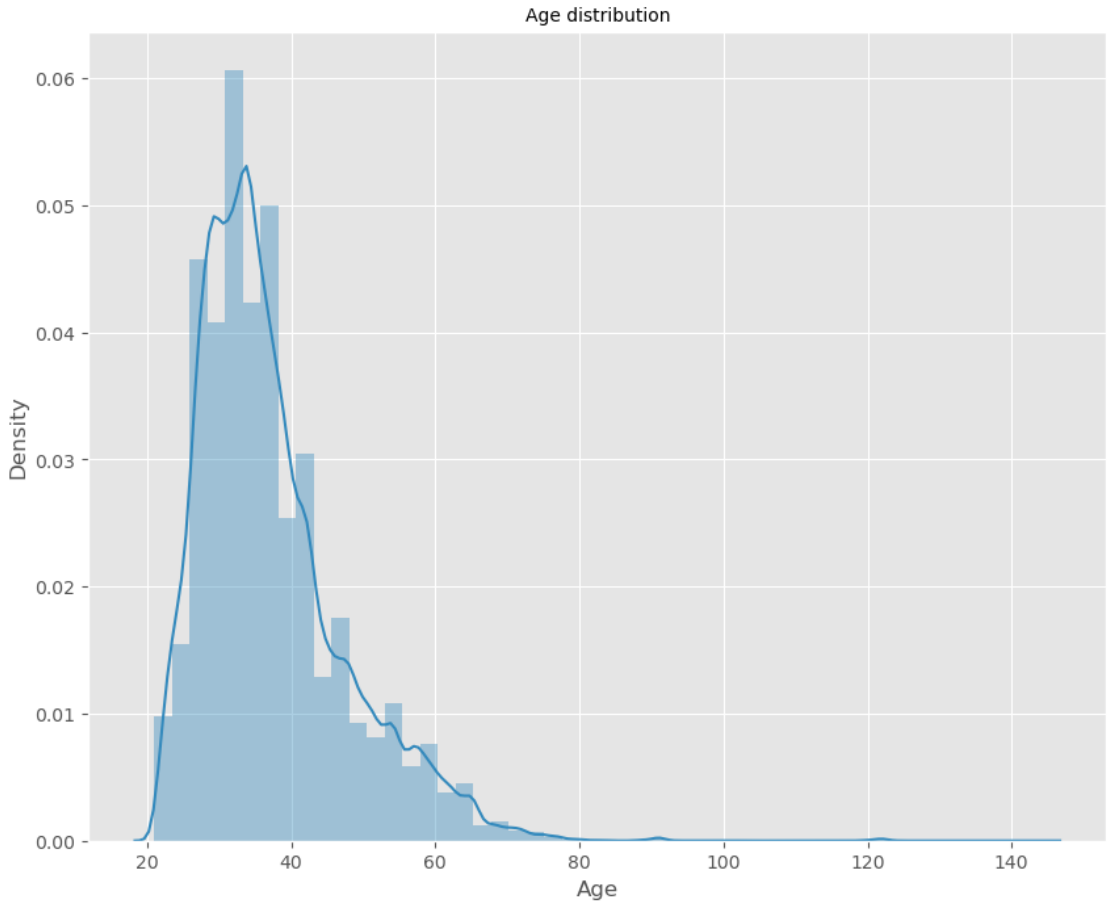
For the age distribution, I have chosed to depict displot graph, see figure (1). In this figure we see that the age distribution is right skewed a little bit, so I decided to look at it more and see if I can drive a detailed insight.

I added the mean by calculating the age distribution mean see figure(2). From this displot graphs we see that most of the users are between 30s - 40s with an average about 37.

[8]:
```
# Investigating the distribution of age
rcParams['figure.figsize'] = 10,8
x = df["age"].values
sb.distplot(x, color= base_color)
plt.title("Age distribution", size =10)
print("******* Fig (1)*********")
plt.xlabel("Age");
```

```
/Users/mabdulahi/opt/anaconda3/lib/python3.9/site-
packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a
deprecated function and will be removed in a future version. Please adapt your
code to use either `displot` (a figure-level function with similar flexibility)
or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
******* Fig (1)*********
```



[9]:
```
# Let us know check out the age distribution by adding the mean.
rcParams['figure.figsize'] = 10,8
```

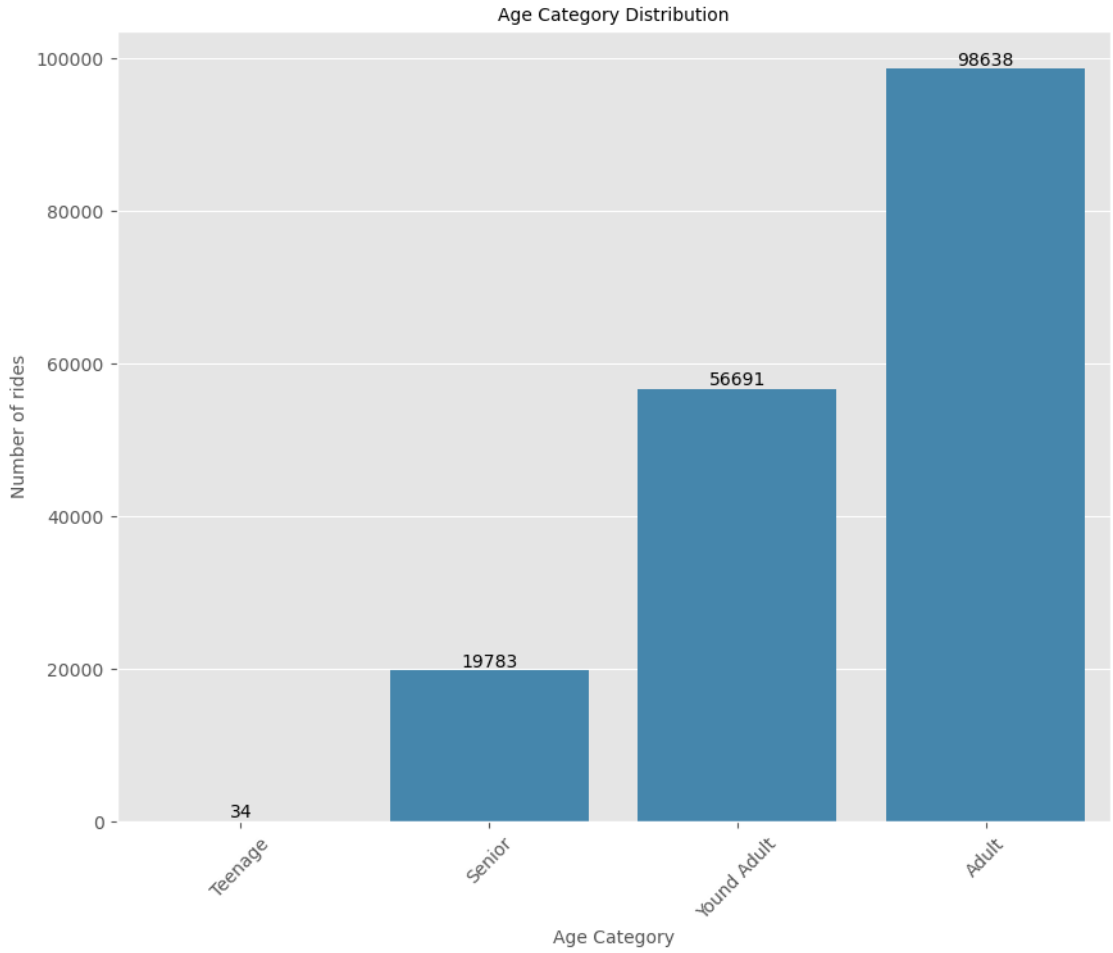
```
x = df['age'].values
sb.displot(df, x="age", kind="kde", fill= True, color= base_color)

# Calculating the mean
mean = df['age'].mean()

#ploting the mean
plt.axvline(mean, 0,2, color = 'red')
print("******* Fig (2)*********")
plt.title("Age distribution")
plt.xlabel("Age");
```

******* Fig (2)*********

3 Age category Distribution

Analazing the group age and user type distribution I found that the majority of the ride were made by adults which defined a as ages 31 - 49.

[10]:
```
# coun the number of rides by age category
df1 =df.groupby("age_group")["age"].count().sort_values().reset_index()
myplot=sb.barplot(x='age_group',y= 'age',data=df1,color=base_color)
myplot.bar_label(myplot.containers[0])
plt.title('Age Category Distribution', size=10)
plt.xlabel("Age Category",size = 10)
plt.ylabel("Number of rides", size= 10)
plt.xticks(rotation=46);
```



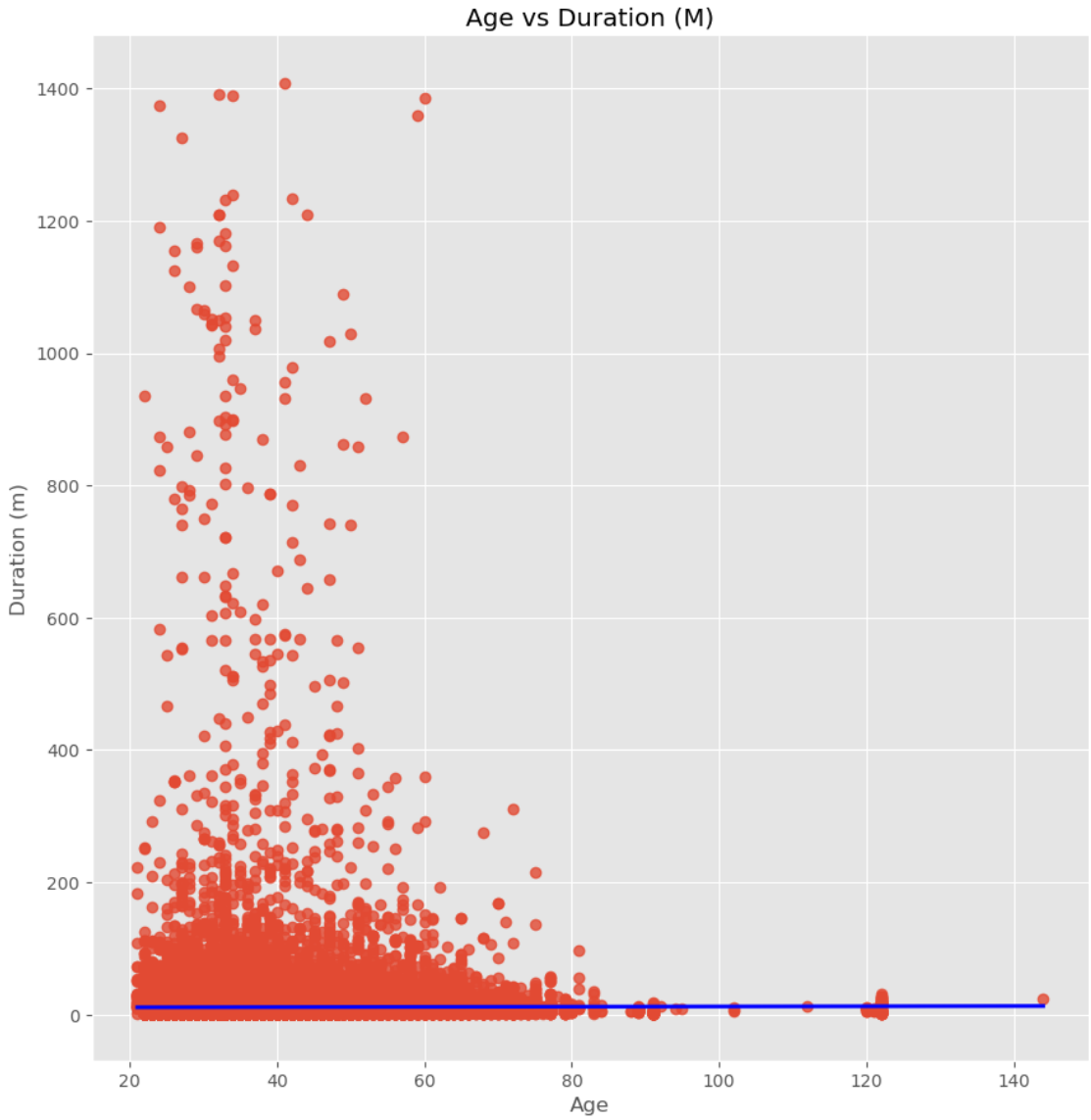
3.0.1 Relationship Between Age and Duration Per Minute Distribution

In order to Look more closely the relationship between age and duration I have plotted lmplot to test its relationship, but Unfortunately, I found that age doesn't seem to have a good relatioship

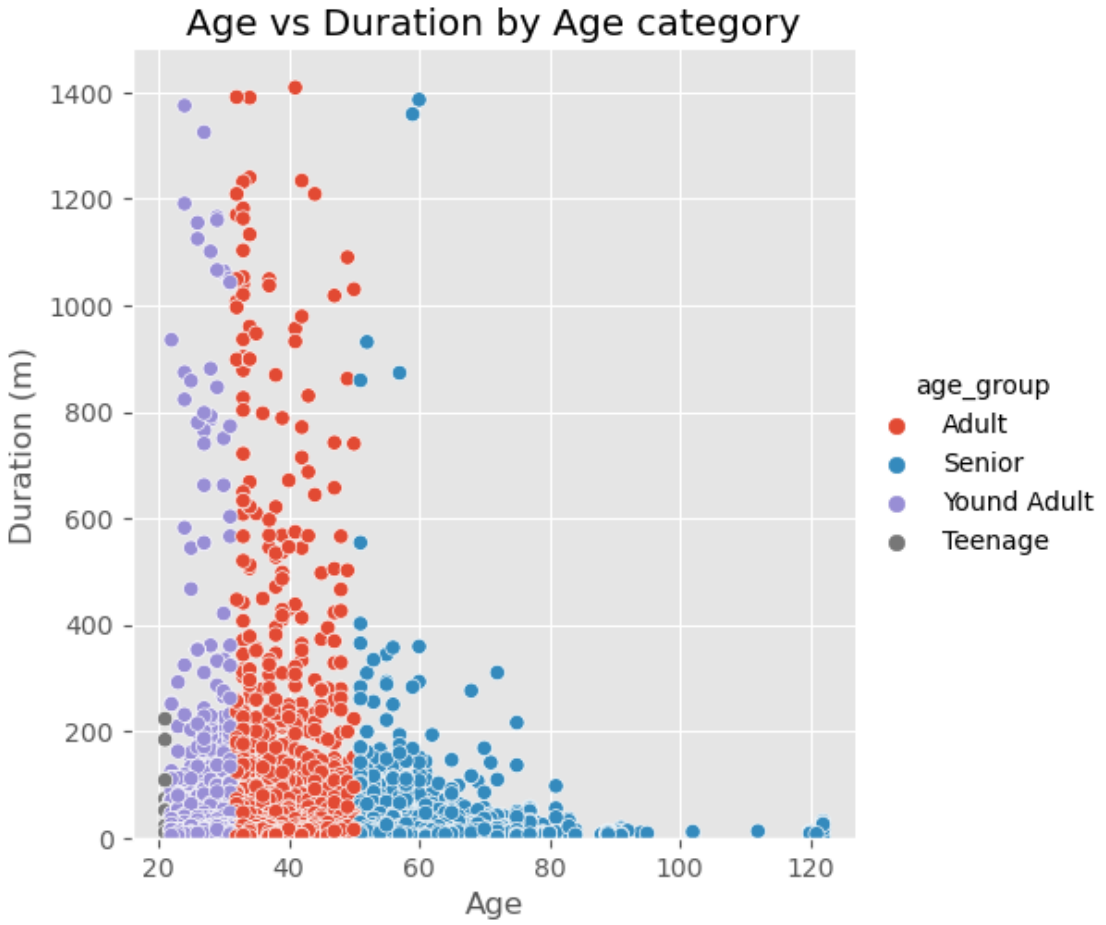
with duration because the regression is so close to the horizantal which indicates that there good relationship.

In Figure (2) I examined if age has any effection on the duration and noticed as the age increased the duraion is decreasing.

[11]:
```
ax=sb.lmplot(x="age", y="dur_per_minute", data=df, height=9,line_kws={'color':␣
↪'blue'})
ax.set_xlabels("Age")
ax.set_ylabels("Duration (m)")
plt.title("Age vs Duration (M)");
```



[12]:
```
# Relationship between age and duration by age category
sb.relplot(x="age", y="dur_per_minute", hue="age_group", data=df)
plt.ylim(0)
plt.xlabel("Age")
plt.ylabel("Duration (m)")
plt.title("Age vs Duration (M)");
plt.title("Age vs Duration by Age category");
```

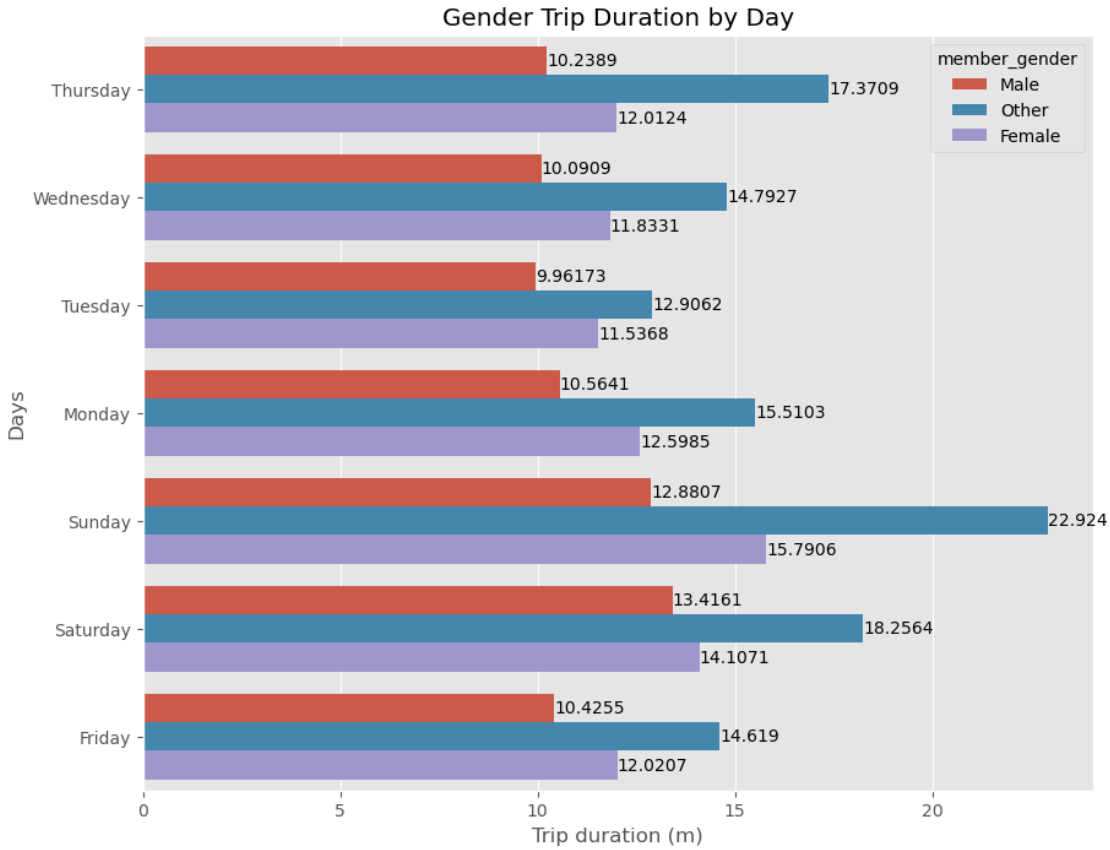


Distribution of Gender Trip Duration The distribution shows that Females and Other genders take longer trip durations compared Males which was surprising to me because because in the previous section we saw that most trips were made by men.

[13]:
```
# show which genre takes the longest trip
# Compare daily trip duration by gender
rcParams['figure.figsize'] = 10,8
ax = sb.barplot(y="day", x="dur_per_minute",
                hue="member_gender",
```

```
            data=df, ci=False)
for container in ax.containers:
    ax.bar_label(container)
plt.title('Gender Trip Duration by Day')
plt.xlabel('Trip duration (m)')
plt.ylabel('Days')
```

[13]: `Text(0, 0.5, 'Days')`

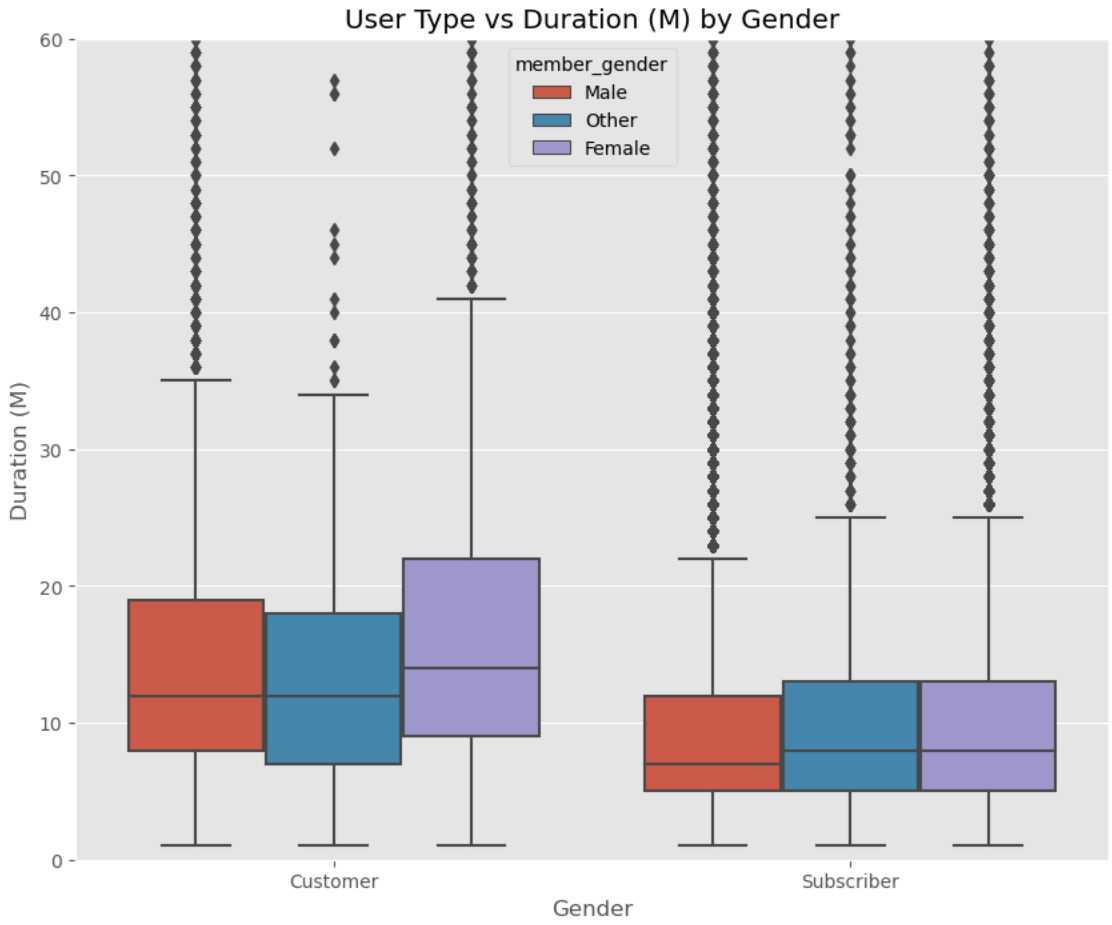


3.0.2 User Type Distribution Duration across all the genders

- In my observation for `Customer Type users`, I found that the females take longer trips, followed by male.
- On other hand, the `subscriber` boxplot depicts that female and other genders are leveled while the male duration is smalled compared to female and other genders. Therefore, we can say, from this figure that females take longer trips than any other gender.

[14]:
```
# Investigating the distribution of user type and duration by gender
sb.boxplot(x='user_type', y='dur_per_minute', data = df, hue="member_gender")
plt.ylim(0, 60)
```

```
plt.title('User Type vs Duration (M) by Gender')
plt.xlabel('Gender')
plt.ylabel('Duration (M)');
```



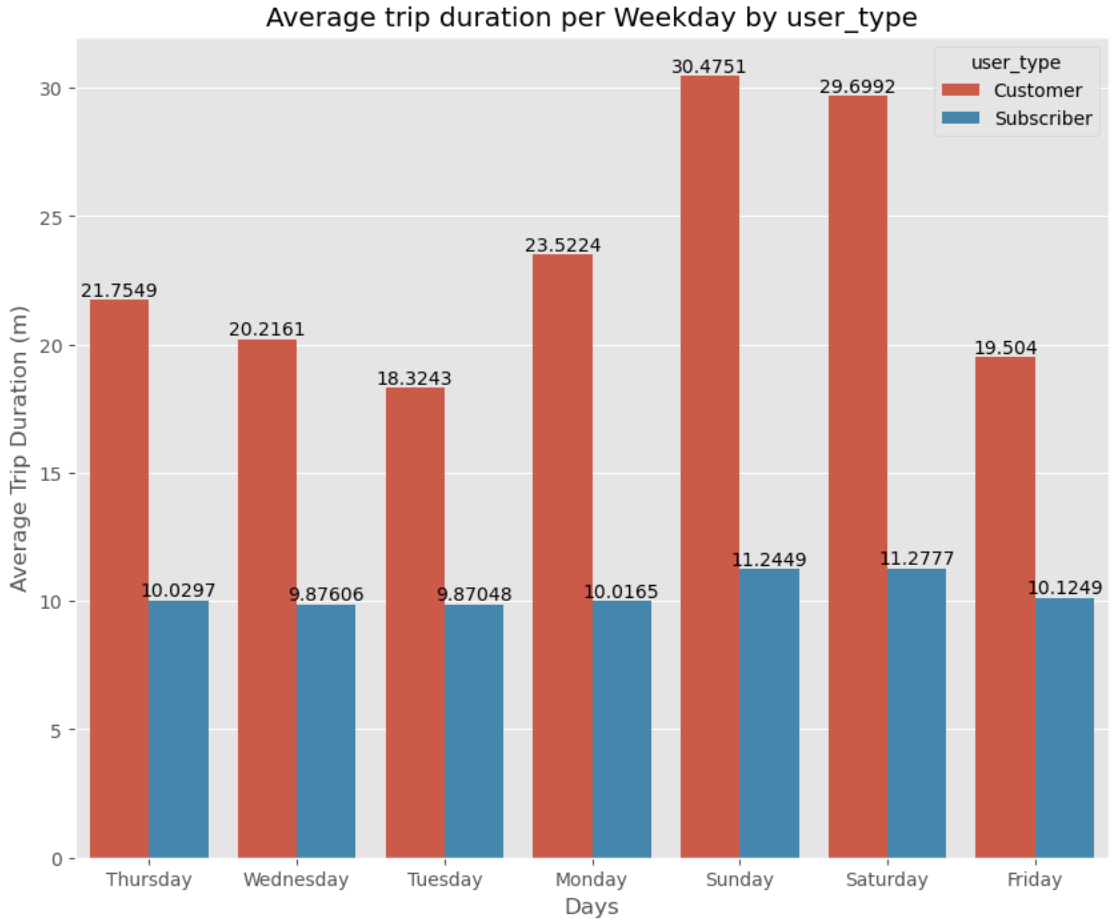
3.0.3 # Compare the average trip duration per Weekday by user type

I have a ploted a grouped barplat to compare the average durartion of rides per day by user type. the graph is properly plotted and polished. I added proper labels and descriptive title.

The graph Visual shows that The customer user types take longer trips than subscribe users on week days.

[15]:
```
# Compare Average trip duration per Weekday by user_type
ax = sb.barplot(data = df, x ='day', y = 'dur_per_minute', hue =␣
 ↪'user_type',ci= False)
for container in ax.containers:
    ax.bar_label(container)
plt.title('Average trip duration per Weekday by user_type')
plt.xlabel('Days')
```

```
plt.ylabel('Average Trip Duration (m)')
plt.xlabel('Days');
```



3.0.4 Display the total number of users_types and their age catagory.

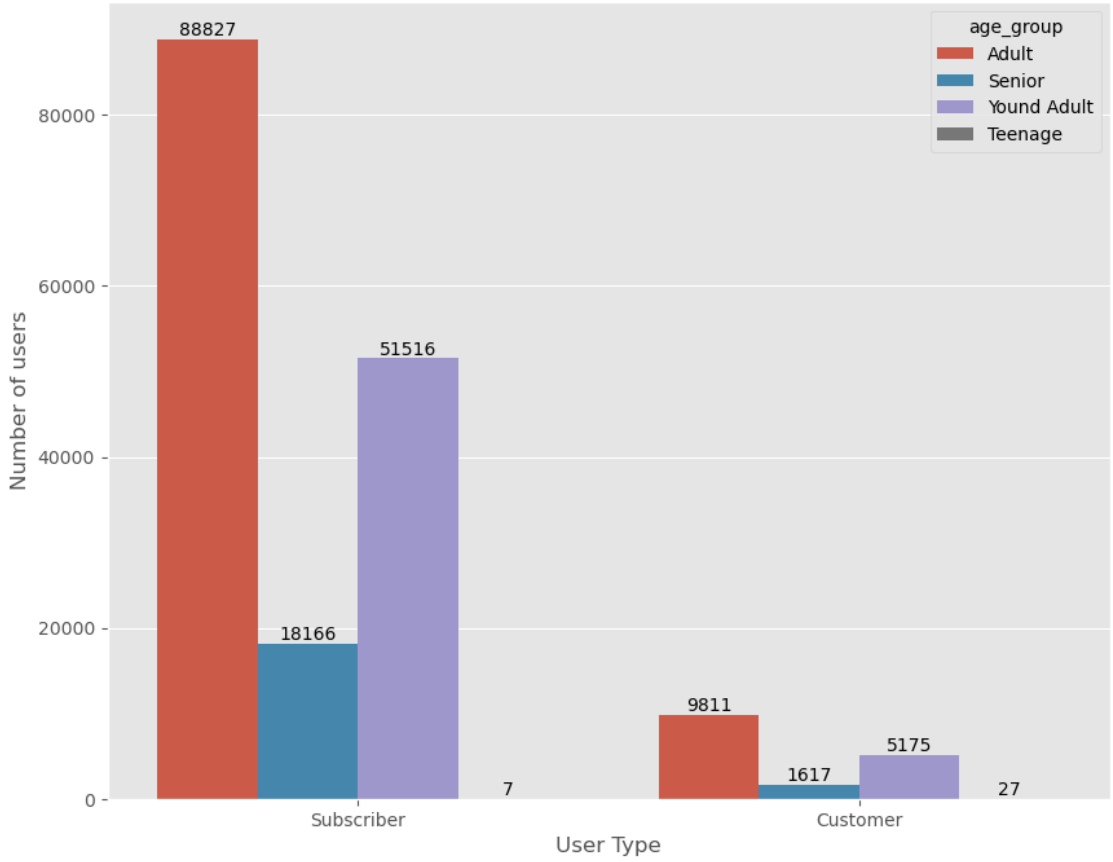
To display the total numbers of users, their type and and their age category, I selected countplot with proper labels, titles, legends to drive a meaningful in sights about our data and answer quick question about our users.

In our data we, have 7 teenagers (ages12-20), 51516 youth adults ages between 21 through 30. 88827 Adult subscribers between age 31 and 49, and 18166 seniors age 50+

For Customer user types, we have 5175 young adults 9811 youth, 1617 seniors, and 27 teenagershe numbers of users, their user_type and and their age category

[16]:
```
#display the numbers of users, their user_type and and their age category
ax = sb.countplot(data=df, x="user_type", hue="age_group",
                  order=df.user_type.value_counts().index)
for container in ax.containers:
```

```
    ax.bar_label(container)
plt.xlabel('User Type')
plt.ylabel('Number of users');
```



3.1 Summery and Conclusions

- The data contains information about individual bike-sharing system covering the greater San Francisco Bay area.
- The average trips is about 12 Minute long, the most trips are between 8-12 minute.
- people start their trips between 8th, 9th and end 17th and 18th o'clock. start and closing work hours.
- Most trips were taken Thrusday, followed by Tuesday. weekends (sat and sun) trips are smaller compared to weekday.
- As age increases trip duration decreases
- Customer user type trips take a longer duration compared to subscribers.
- Female gender take longer trips than other genders
- We also look at the different age categories we have in the data and foun that we have:
  - 7 teenagers ages betweeb(12-20),
  - 51516 youth adults ages between 21 through 30.
  - 88827 Adult subscribers between age 31 and 49, and

– 18166 seniors age 50+

3.2 Limitations

Chosing the right visualization was the hardest in this project for me, I was using python 3.6 in my Udacity work space, so I felt I have very limited graphs visualization in my seaborn library, I was not able to updates latest version of seaborn, so I have been missing most of the recent seaborn plots like (Catplot, scatterplot and many more), Also time was not my best friend for the past couble months I work 12-13 hrs shift,so I believed if I would have spent more time I would have done better, I am sure this analysis is not 100% guarenteed to be proof error solution.

3.2.1 Sources

https://seaborn.pydata.org/generated/seaborn.regplot.html https://stackoverflow.com/questions/55104819/displa
count-on-top-of-seaborn-barplot https://deepnote.com/@dain-russell/bike-exploration-328b5ba1-
25e4-4a35-aaad-e70146c9e182 https://seaborn.pydata.org/generated/seaborn.boxplot.html
https://seaborn.pydata.org/generated/seaborn.countplot.html https://stackoverflow.com/questions/26597116/sea
plots-not-showing-up https://stackoverflow.com/questions/67723105/how-
to-convert-time-from-24-hour-format-to-12-hour-format-am-pm-with-pandas-p
https://dataindependent.com/pandas/pandas-to-datetime-string-to-date-pd-to_datetime/
https://stackoverflow.com/questions/49153253/pandas-rounding-when-converting-float-to-integer

3.2.2 Generate Slideshow

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

[17]:
```
# Use this command if you are running this file in local
!jupyter nbconvert <Part_II_Filename>.ipynb --to slides --post serve --no-input␣
↪--no-prompt
```

```
zsh:1: no such file or directory: Part_II_Filename
```

> In the classroom workspace, the generated HTML slideshow will be placed in the home folder.
>
> In local machines, the command above should open a tab in your web browser where you can scroll through your presentation. Sub-slides can be accessed by pressing 'down' when viewing its parent slide. Make sure you remove all of the quote-formatted guide notes like this one before you finish your presentation! At last, you can stop the Kernel.

3.2.3 Submission

If you are using classroom workspace, you can choose from the following two ways of submission:

1. **Submit from the workspace**. Make sure you have removed the example project from the /home/workspace directory. You must submit the following files:
   - Part_I_notebook.ipynb
   - Part_I_notebook.html or pdf
   - Part_II_notebook.ipynb
   - Part_I_slides.html

- README.md
- dataset (optional)

2. **Submit a zip file on the last page of this project lesson**. In this case, open the Jupyter terminal and run the command below to generate a ZIP file.

```
zip -r my_project.zip .
```

The command abobve will ZIP every file present in your /home/workspace directory. Next, you can download the zip to your local, and follow the instructions on the last page of this project lesson.

[ ]: