fordbike

February 26, 2021

0.1 Final Project

0.1.1 Introduction

We are going to analyze fordbike dataset and explore all the questions to study the best improvments needed to boost the profit

```
In [1]: # All imports needed for analysis
        import pandas as pd
        import numpy as np
        import random
        import matplotlib.pyplot as plt
```

Reading the csv file

```
In [2]: df = pd.read_csv('fordbike.csv')
        df['age'] = 2021-df.member_birth_year
        df.head(20)
```

```
Out[2]:     duration_sec                start_time                  end_time  \
        0          52185  2019-02-28 17:32:10.1450  2019-03-01 08:01:55.9750
        1          42521  2019-02-28 18:53:21.7890  2019-03-01 06:42:03.0560
        2          61854  2019-02-28 12:13:13.2180  2019-03-01 05:24:08.1460
        3          36490  2019-02-28 17:54:26.0100  2019-03-01 04:02:36.8420
        4           1585  2019-02-28 23:54:18.5490  2019-03-01 00:20:44.0740
        5           1793  2019-02-28 23:49:58.6320  2019-03-01 00:19:51.7600
        6           1147  2019-02-28 23:55:35.1040  2019-03-01 00:14:42.5880
        7           1615  2019-02-28 23:41:06.7660  2019-03-01 00:08:02.7560
        8           1570  2019-02-28 23:41:48.7900  2019-03-01 00:07:59.7150
        9           1049  2019-02-28 23:49:47.6990  2019-03-01 00:07:17.0250
        10           458  2019-02-28 23:57:57.2110  2019-03-01 00:05:35.4350
        11           506  2019-02-28 23:56:55.5400  2019-03-01 00:05:21.7330
        12          1176  2019-02-28 23:45:12.6510  2019-03-01 00:04:49.1840
        13           915  2019-02-28 23:49:06.0620  2019-03-01 00:04:21.8670
        14           395  2019-02-28 23:56:26.8480  2019-03-01 00:03:01.9470
        15           208  2019-02-28 23:59:18.5480  2019-03-01 00:02:47.2280
        16           548  2019-02-28 23:50:41.6070  2019-02-28 23:59:49.9530
        17           674  2019-02-28 23:48:25.0950  2019-02-28 23:59:40.0920
        18           557  2019-02-28 23:49:01.8510  2019-02-28 23:58:19.8090
```

```
19           874  2019-02-28 23:43:05.1830  2019-02-28 23:57:39.7960

    start_station_id                                start_station_name  \
0               21.0   Montgomery St BART Station (Market St at 2nd St)
1               23.0                      The Embarcadero at Steuart St
2               86.0                            Market St at Dolores St
3              375.0                            Grove St at Masonic Ave
4                7.0                                Frank H Ogawa Plaza
5               93.0                       4th St at Mission Bay Blvd S
6              300.0                               Palm St at Willow St
7               10.0                         Washington St at Kearny St
8               10.0                         Washington St at Kearny St
9               19.0                               Post St at Kearny St
10             370.0                                Jones St at Post St
11              44.0  Civic Center/UN Plaza BART Station (Market St ...
12             127.0                             Valencia St at 21st St
13             252.0                       Channing Way at Shattuck Ave
14             243.0                        Bancroft Way at College Ave
15             349.0                               Howard St at Mary St
16             131.0                              22nd St at Dolores St
17              74.0                              Laguna St at Hayes St
18             321.0                                   5th St at Folsom
19             180.0                           Telegraph Ave at 23rd St

    start_station_latitude  start_station_longitude  end_station_id  \
0                37.789625              -122.400811            13.0
1                37.791464              -122.391034            81.0
2                37.769305              -122.426826             3.0
3                37.774836              -122.446546            70.0
4                37.804562              -122.271738           222.0
5                37.770407              -122.391198           323.0
6                37.317298              -121.884995           312.0
7                37.795393              -122.404770           127.0
8                37.795393              -122.404770           127.0
9                37.788975              -122.403452           121.0
10               37.787327              -122.413278            43.0
11               37.781074              -122.411738           343.0
12               37.756708              -122.421025           323.0
13               37.865847              -122.267443           244.0
14               37.869360              -122.254337           252.0
15               37.781010              -122.405666            60.0
16               37.755000              -122.425728            71.0
17               37.776435              -122.426244           336.0
18               37.780146              -122.403071            75.0
19               37.812678              -122.268773           180.0

                                     end_station_name  end_station_latitude  \
0                      Commercial St at Montgomery St             37.794231
```

```
1                                 Berry St at 4th St             37.775880
2        Powell St BART Station (Market St at 4th St)             37.786375
3                             Central Ave at Fell St             37.773311
4                              10th Ave at E 15th St             37.792714
5                                 Broadway at Kearny             37.798014
6                           San Jose Diridon Station             37.329732
7                             Valencia St at 21st St             37.756708
8                             Valencia St at 21st St             37.756708
9                                 Mission Playground             37.759210
10  San Francisco Public Library (Grove St at Hyde...             37.778768
11                               Bryant St at 2nd St             37.783172
12                                Broadway at Kearny             37.798014
13                        Shattuck Ave at Hearst Ave             37.873676
14                      Channing Way at Shattuck Ave             37.865847
15                              8th St at Ringold St             37.774520
16                            Broderick St at Oak St             37.773063
17                       Potrero Ave and Mariposa St             37.763281
18                          Market St at Franklin St             37.773793
19                          Telegraph Ave at 23rd St             37.812678

    end_station_longitude  bike_id   user_type  member_birth_year  \
0             -122.402923     4902    Customer             1984.0
1             -122.393170     2535    Customer                NaN
2             -122.404904     5905    Customer             1972.0
3             -122.444293     6638  Subscriber             1989.0
4             -122.248780     4898  Subscriber             1974.0
5             -122.405950     5200  Subscriber             1959.0
6             -121.901782     3803  Subscriber             1983.0
7             -122.421025     6329  Subscriber             1989.0
8             -122.421025     6548  Subscriber             1988.0
9             -122.421339     6488  Subscriber             1992.0
10            -122.415929     5318  Subscriber             1996.0
11            -122.393572     5848  Subscriber             1993.0
12            -122.405950     5328    Customer             1990.0
13            -122.268487     5101  Subscriber                NaN
14            -122.267443     4786  Subscriber             1988.0
15            -122.409449     6361  Subscriber             1993.0
16            -122.439078     6572  Subscriber             1981.0
17            -122.407377     5343  Subscriber             1975.0
18            -122.421239     5854  Subscriber             1990.0
19            -122.268773     5629    Customer             1978.0

   member_gender bike_share_for_all_trip   age
0           Male                      No  37.0
1            NaN                      No   NaN
2           Male                      No  49.0
3          Other                      No  32.0
4           Male                     Yes  47.0
```

```
5          Male                     No  62.0
6        Female                     No  38.0
7          Male                     No  32.0
8         Other                     No  33.0
9          Male                     No  29.0
10       Female                    Yes  25.0
11         Male                     No  28.0
12         Male                     No  31.0
13          NaN                     No   NaN
14         Male                     No  33.0
15         Male                    Yes  28.0
16         Male                     No  40.0
17         Male                     No  46.0
18         Male                     No  31.0
19         Male                     No  43.0
```

Checking out the total number of entries, total number of columns, and total number of nan values per column

```
In [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 17 columns):
duration_sec               183412 non-null int64
start_time                 183412 non-null object
end_time                   183412 non-null object
start_station_id           183215 non-null float64
start_station_name         183215 non-null object
start_station_latitude     183412 non-null float64
start_station_longitude    183412 non-null float64
end_station_id             183215 non-null float64
end_station_name           183215 non-null object
end_station_latitude       183412 non-null float64
end_station_longitude      183412 non-null float64
bike_id                    183412 non-null int64
user_type                  183412 non-null object
member_birth_year          175147 non-null float64
member_gender              175147 non-null object
bike_share_for_all_trip    183412 non-null object
age                        175147 non-null float64
dtypes: float64(8), int64(2), object(7)
memory usage: 23.8+ MB
```

```
In [4]: # Changing from object to datetime
        df['start_time'] = pd.to_datetime(df['start_time'])
        df['end_time'] = pd.to_datetime(df['end_time'])
```

```
In [5]: # Lets check the distribution of the numerical columns
        df.describe()

Out[5]:          duration_sec  start_station_id  start_station_latitude  \
        count  183412.000000     183215.000000           183412.000000
        mean      726.078435        138.590427               37.771223
        std      1794.389780        111.778864                0.099581
        min        61.000000          3.000000               37.317298
        25%       325.000000         47.000000               37.770083
        50%       514.000000        104.000000               37.780760
        75%       796.000000        239.000000               37.797280
        max     85444.000000        398.000000               37.880222

               start_station_longitude  end_station_id  end_station_latitude  \
        count            183412.000000   183215.000000         183412.000000
        mean               -122.352664      136.249123             37.771427
        std                   0.117097      111.515131              0.099490
        min                -122.453704        3.000000             37.317298
        25%                -122.412408       44.000000             37.770407
        50%                -122.398285      100.000000             37.781010
        75%                -122.286533      235.000000             37.797320
        max                -121.874119      398.000000             37.880222

               end_station_longitude        bike_id  member_birth_year            age
        count          183412.000000  183412.000000      175147.000000  175147.000000
        mean             -122.352250    4472.906375        1984.806437      36.193563
        std                 0.116673    1664.383394          10.116689      10.116689
        min              -122.453704      11.000000        1878.000000      20.000000
        25%              -122.411726    3777.000000        1980.000000      29.000000
        50%              -122.398279    4958.000000        1987.000000      34.000000
        75%              -122.288045    5502.000000        1992.000000      41.000000
        max              -121.874119    6645.000000        2001.000000     143.000000
```

```
In [6]: #check for duplicated rows
        df.duplicated()

Out[6]: 0         False
        1         False
        2         False
        3         False
        4         False
        5         False
        6         False
        7         False
        8         False
        9         False
        10        False
        11        False
```

```
12        False
13        False
14        False
15        False
16        False
17        False
18        False
19        False
20        False
21        False
22        False
23        False
24        False
25        False
26        False
27        False
28        False
29        False
          ...
183382    False
183383    False
183384    False
183385    False
183386    False
183387    False
183388    False
183389    False
183390    False
183391    False
183392    False
183393    False
183394    False
183395    False
183396    False
183397    False
183398    False
183399    False
183400    False
183401    False
183402    False
183403    False
183404    False
183405    False
183406    False
183407    False
183408    False
183409    False
183410    False
```

```
        183411    False
        Length: 183412, dtype: bool
```

```
In [7]: # Drop duplicates if they occur and recheck the number of entries to see if any rows are
        df.drop_duplicates(inplace=True)
```

```
In [8]: # Droppin rows with na values in station names and ids and gender
        df = df[df['start_station_id'].notna()]
        df = df[df['member_gender'].notna()]
```

```
In [9]: #Fill in birth year with median value
        df['member_birth_year'].fillna((df['member_birth_year'].median()), inplace=True)
```

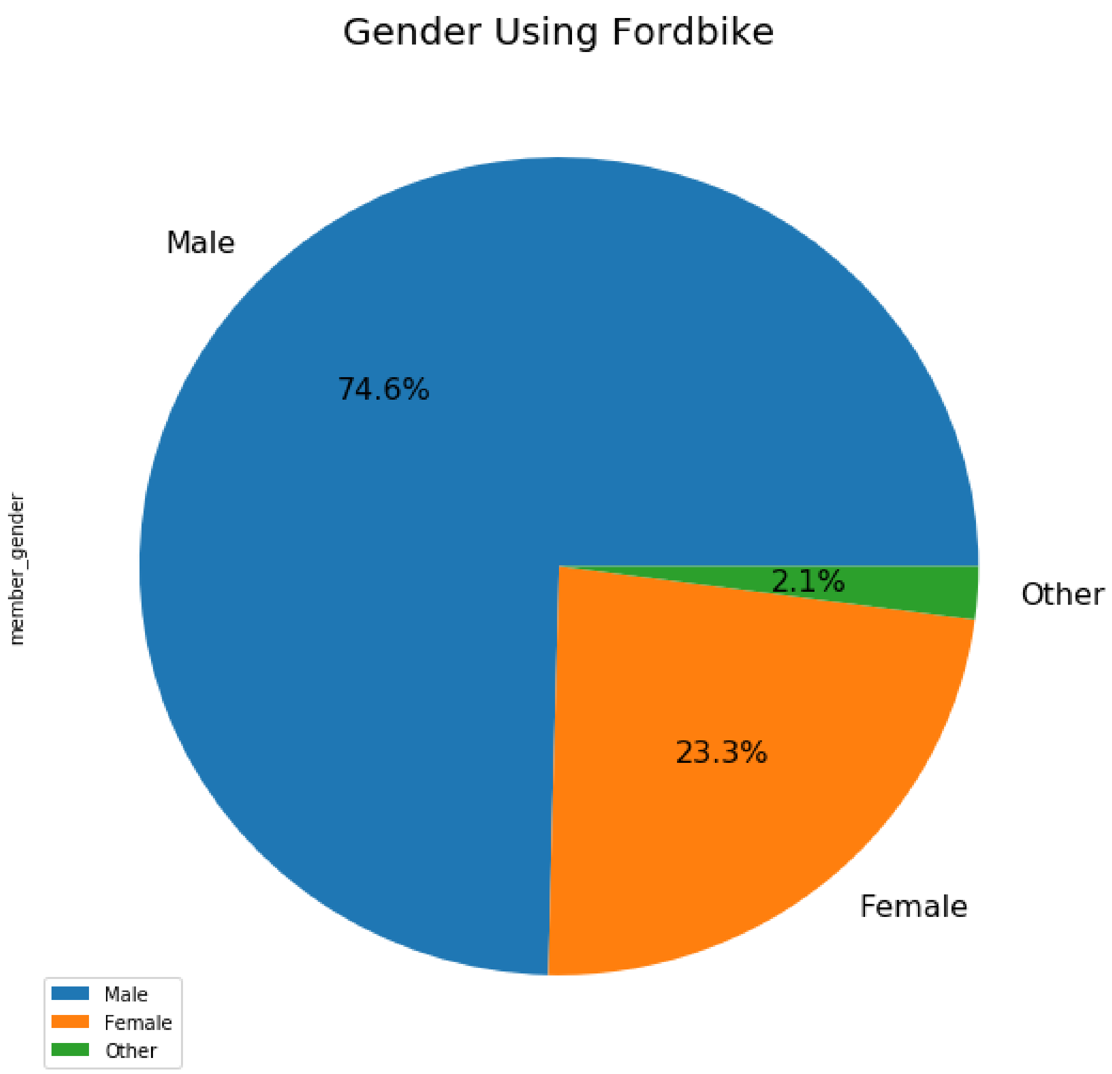
```
In [10]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 174952 entries, 0 to 183411
Data columns (total 17 columns):
duration_sec               174952 non-null int64
start_time                 174952 non-null datetime64[ns]
end_time                   174952 non-null datetime64[ns]
start_station_id           174952 non-null float64
start_station_name         174952 non-null object
start_station_latitude     174952 non-null float64
start_station_longitude    174952 non-null float64
end_station_id             174952 non-null float64
end_station_name           174952 non-null object
end_station_latitude       174952 non-null float64
end_station_longitude      174952 non-null float64
bike_id                    174952 non-null int64
user_type                  174952 non-null object
member_birth_year          174952 non-null float64
member_gender              174952 non-null object
bike_share_for_all_trip    174952 non-null object
age                        174952 non-null float64
dtypes: datetime64[ns](2), float64(8), int64(2), object(5)
memory usage: 24.0+ MB
```

0.1.2 Exploration

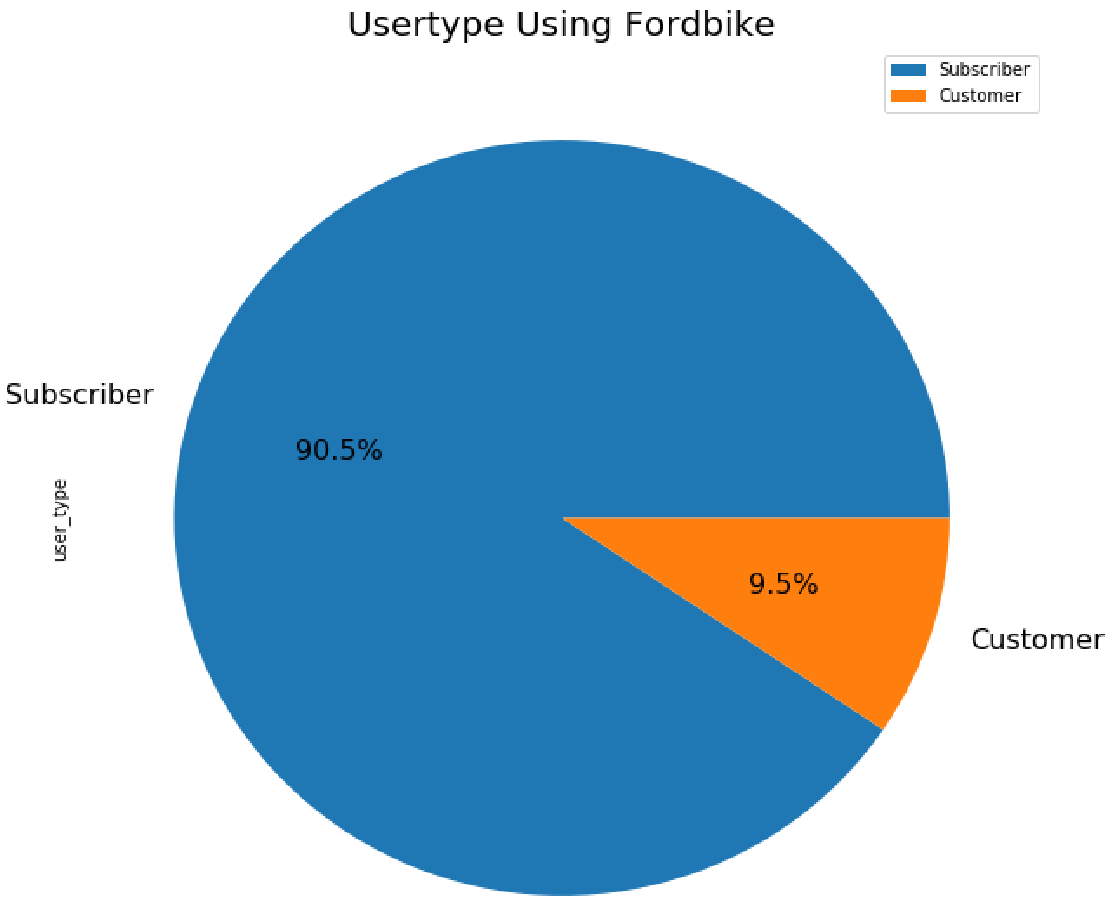
Research Question 1 (Which gender use the fordbike more?)

```
In [11]: df_gender = df[['member_gender']]
         genderPie = df_gender['member_gender'].value_counts()
         pieChart = genderPie.plot.pie(figsize=(10,10), autopct='%1.1f%%', fontsize = 16);
         pieChart.set_title("Gender Using Fordbike", fontsize = 20);
         plt.legend();
```



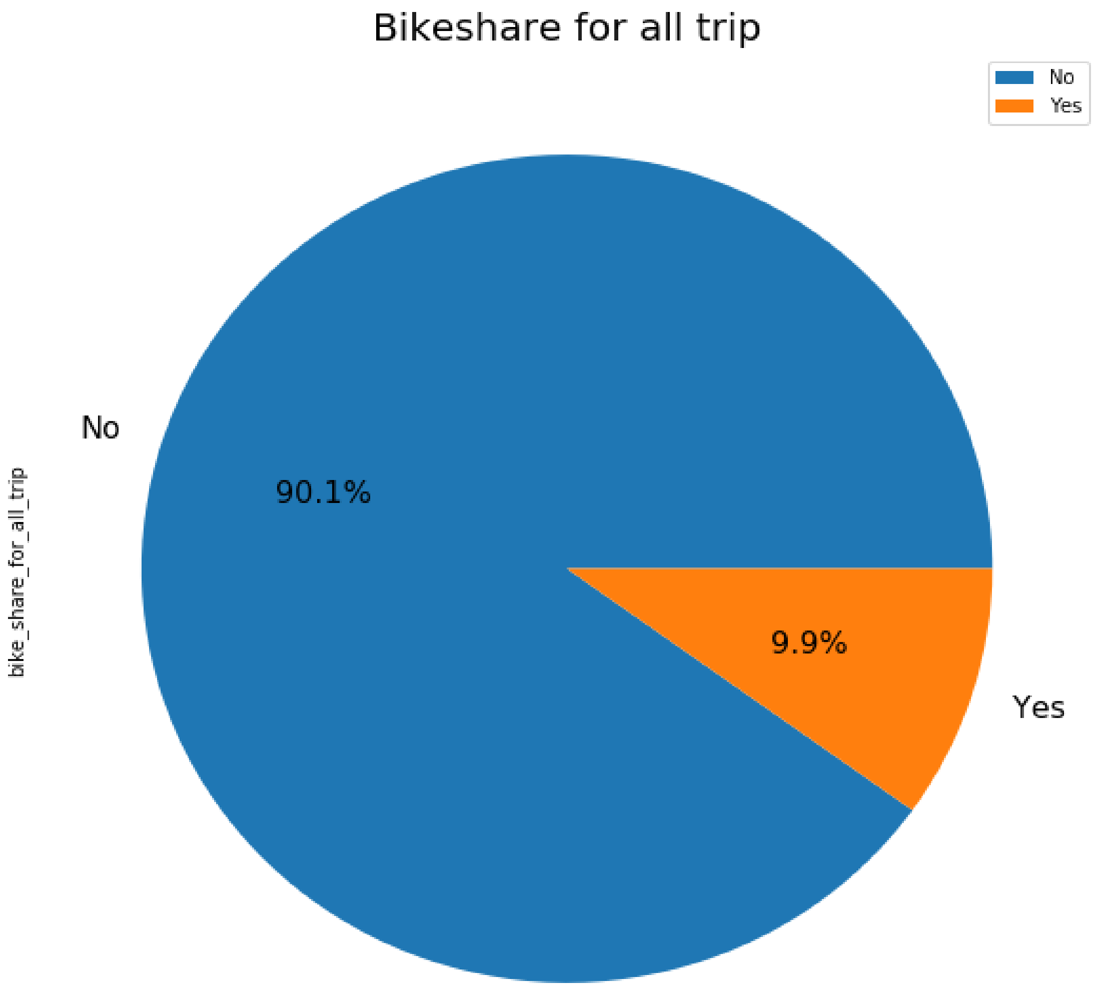
Research Question 2 (What is the ratio of subscribers to customers?)

```
In [12]: df_user_type = df[['user_type']]
         genderPie = df_user_type['user_type'].value_counts()
         pieChart = genderPie.plot.pie(figsize=(10,10), autopct='%1.1f%%', fontsize = 16);
         pieChart.set_title("Usertype Using Fordbike", fontsize = 20);
         plt.legend();
```



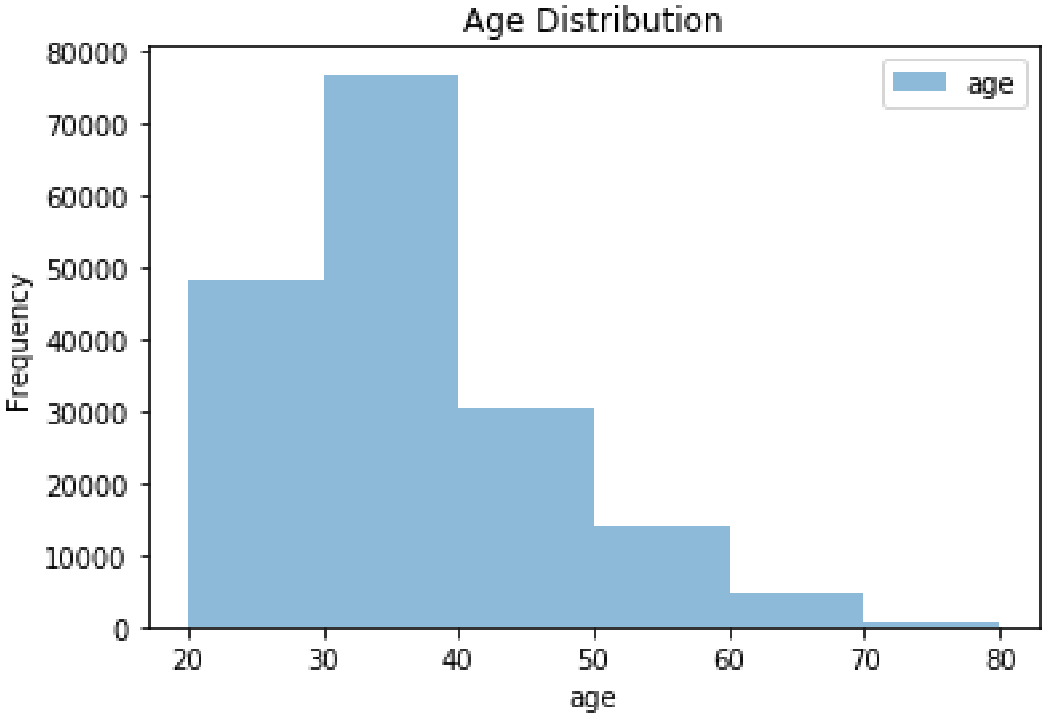
Research Question 3 (What is percentage of bikeshare for all the trips?)

```
In [13]: df_bike_share_for_all_trip = df[['bike_share_for_all_trip']]
         genderPie = df_bike_share_for_all_trip['bike_share_for_all_trip'].value_counts()
         pieChart = genderPie.plot.pie(figsize=(10,10), autopct='%1.1f%%', fontsize = 16);
         pieChart.set_title("Bikeshare for all trip", fontsize = 20);
         plt.legend();
```

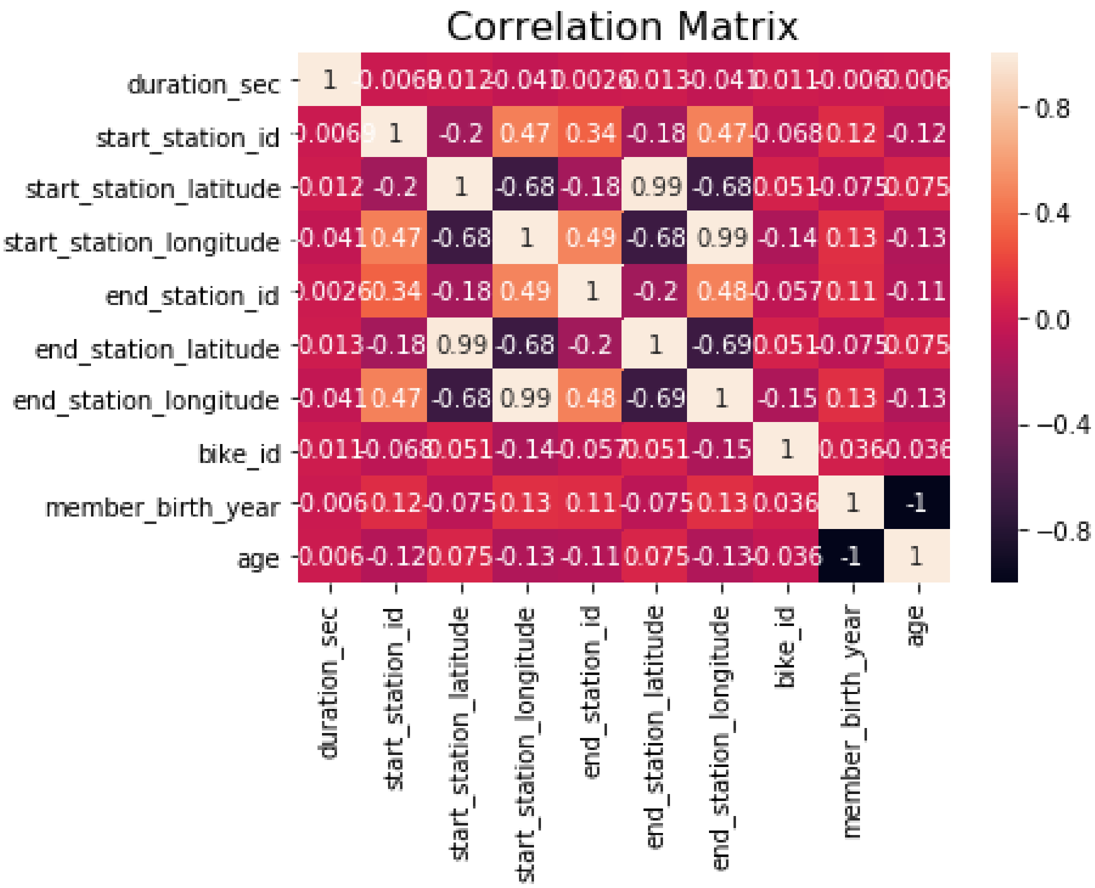


Research Question 4 (What is the range age of riders?)

```
In [20]: df_age = df[['age']]
         nowshow_age = df_age.plot.hist(bins=6,range=[20, 80], alpha=0.5, title="Age Distributio
         plt.xlabel('age')
         plt.show()
```



```
In [32]: import pandas as pd
         import seaborn as sn
         import matplotlib.pyplot as plt
         corrMatrix = df.corr()
         sn.heatmap(corrMatrix, annot=True)
         plt.title('Correlation Matrix', fontsize=16);
```



	duration_sec	start_station_id	start_station_latitude	start_station_longitude	end_station_id	end_station_latitude	end_station_longitude	bike_id	member_birth_year	age
duration_sec	1	0.0069	0.012	-0.041	0.0026	0.013	-0.041	0.011	-0.006	0.006
start_station_id	0.0069	1	-0.2	0.47	0.34	-0.18	0.47	-0.068	0.12	-0.12
start_station_latitude	0.012	-0.2	1	-0.68	-0.18	0.99	-0.68	0.051	-0.075	0.075
start_station_longitude	-0.041	0.47	-0.68	1	0.49	-0.68	0.99	-0.14	0.13	-0.13
end_station_id	0.0026	0.34	-0.18	0.49	1	-0.2	0.48	-0.057	0.11	-0.11
end_station_latitude	0.013	-0.18	0.99	-0.68	-0.2	1	-0.69	0.051	-0.075	0.075
end_station_longitude	-0.041	0.47	-0.68	0.99	0.48	-0.69	1	-0.15	0.13	-0.13
bike_id	0.011	-0.068	0.051	-0.14	-0.057	0.051	-0.15	1	0.036	-0.036
member_birth_year	-0.006	0.12	-0.075	0.13	0.11	-0.075	0.13	0.036	1	-1
age	0.006	-0.12	0.075	-0.13	-0.11	0.075	-0.13	-0.036	-1	1

```
In [33]: upyter nbconvert presentation.ipynb --to slides
```

```
      File "<ipython-input-33-1d1ca0382e64>", line 1
    upyter nbconvert presentation.ipynb --to slides
                   ^
SyntaxError: invalid syntax
```

```
In [ ]:
```