

DBC_Final_Gender

December 20, 2018

This project was completed by insert full name here in partial fulfilment of ECON-UB.0232, Data Bootcamp, Spring 2018. I certify that the NYU Stern Honor Code applies to this project. In particular, I have: Clearly acknowledged the work and efforts of others when submitting written work as our own. The incorporation of the work of others—including but not limited to their ideas, data, creative expression, and direct quotations (which should be designated with quotation marks), or paraphrasing thereof—has been fully and appropriately referenced using notations both in the text and the bibliography. And I understand that: Submitting the same or substantially similar work in multiple courses, either in the same semester or in a different semester, without the express approval of all instructors is strictly forbidden. I acknowledge that a failure to abide by NYU Stern Honor Code will result in a failing grade for the project and course. With this project we have executed a study of citibank through the study of demographics, mainly age and gender. We have come to the conclusion that citibike should mainly focus on marketing their program to males between the ages of 18-30 and 60-80 given that these users experience the highest average cost per minute from riding. If Citi Bike attracts customers who use the program least, they will be able to spread their assets amongst more customers. Unfortunately, we were unable to provide significant correlations between variables in our data and a further investigation with a larger dataset should be done to prove our conclusions.

```
In [1]: import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.formula.api as smf
%matplotlib inline
```

```
In [2]: citi = pd.read_csv('/Users/MartinSmit/Documents/NYUAD/Junior/Data Bootcamp/Final/CitiF
```

```
In [3]: del citi['Unnamed: 0']
citi=citi[citi['Age']<91]
```

```
In [4]: citi
```

```
Out[4]:
```

	tripduration	starttime	stoptime \
0	1346	1/1/2015 0:01	1/1/2015 0:24
1	363	1/1/2015 0:02	1/1/2015 0:08
2	346	1/1/2015 0:04	1/1/2015 0:10
3	182	1/1/2015 0:04	1/1/2015 0:07
4	969	1/1/2015 0:05	1/1/2015 0:21
5	496	1/1/2015 0:07	1/1/2015 0:15

6	152	1/1/2015 0:07	1/1/2015 0:09
7	1183	1/1/2015 0:08	1/1/2015 0:28
8	846	1/1/2015 0:09	1/1/2015 0:23
9	576	1/1/2015 0:10	1/1/2015 0:20
10	540	1/1/2015 0:10	1/1/2015 0:19
11	419	1/1/2015 0:11	1/1/2015 0:18
12	751	1/1/2015 0:13	1/1/2015 0:25
13	332	1/1/2015 0:13	1/1/2015 0:18
14	1099	1/1/2015 0:14	1/1/2015 0:32
15	649	1/1/2015 0:14	1/1/2015 0:25
16	614	1/1/2015 0:14	1/1/2015 0:24
17	1196	1/1/2015 0:16	1/1/2015 0:36
18	1426	1/1/2015 0:17	1/1/2015 0:40
19	1262	1/1/2015 0:18	1/1/2015 0:39
20	707	1/1/2015 0:18	1/1/2015 0:30
21	307	1/1/2015 0:18	1/1/2015 0:23
22	1053	1/1/2015 0:19	1/1/2015 0:37
23	446	1/1/2015 0:20	1/1/2015 0:27
25	797	1/1/2015 0:21	1/1/2015 0:35
27	1639	1/1/2015 0:22	1/1/2015 0:49
29	470	1/1/2015 0:22	1/1/2015 0:30
30	321	1/1/2015 0:23	1/1/2015 0:29
31	259	1/1/2015 0:23	1/1/2015 0:27
32	81	1/1/2015 0:23	1/1/2015 0:24
...
3942431	384	10/31/2015 23:57:28	11/1/2015 00:03:53
3942432	204	10/31/2015 23:57:30	11/1/2015 00:00:55
3942433	424	10/31/2015 23:57:31	11/1/2015 00:04:36
3942434	573	10/31/2015 23:57:34	11/1/2015 00:07:07
3942436	495	10/31/2015 23:57:42	11/1/2015 00:05:57
3942437	371	10/31/2015 23:57:47	11/1/2015 00:03:58
3942438	529	10/31/2015 23:57:46	11/1/2015 00:06:36
3942439	680	10/31/2015 23:57:52	11/1/2015 00:09:13
3942440	302	10/31/2015 23:57:51	11/1/2015 00:02:54
3942442	267	10/31/2015 23:57:57	11/1/2015 00:02:24
3942443	704	10/31/2015 23:58:00	11/1/2015 00:09:45
3942444	1127	10/31/2015 23:58:01	11/1/2015 00:16:49
3942445	288	10/31/2015 23:58:01	11/1/2015 00:02:50
3942446	474	10/31/2015 23:58:06	11/1/2015 00:06:00
3942447	292	10/31/2015 23:58:10	11/1/2015 00:03:03
3942448	470	10/31/2015 23:58:17	11/1/2015 00:06:07
3942449	320	10/31/2015 23:58:19	11/1/2015 00:03:40
3942450	710	10/31/2015 23:58:23	11/1/2015 00:10:14
3942451	513	10/31/2015 23:58:39	11/1/2015 00:07:12
3942452	883	10/31/2015 23:58:56	11/1/2015 00:13:40
3942454	194	10/31/2015 23:58:59	11/1/2015 00:02:13
3942459	1041	10/31/2015 23:59:12	11/1/2015 00:16:33
3942461	548	10/31/2015 23:59:26	11/1/2015 00:08:35

3942463	1621	10/31/2015	23:59:36	11/1/2015	00:26:38
3942464	341	10/31/2015	23:59:44	11/1/2015	00:05:25
3942465	1924	10/31/2015	23:59:46	11/1/2015	00:31:50
3942466	711	10/31/2015	23:59:50	11/1/2015	00:11:42
3942467	621	10/31/2015	23:59:54	11/1/2015	00:10:15
3942468	632	10/31/2015	23:59:56	11/1/2015	00:10:29
3942469	807	10/31/2015	23:59:57	11/1/2015	00:13:24

	start station latitude	start station longitude \
0	40.750020	-73.969053
1	40.743174	-74.003664
2	40.740964	-73.986022
3	40.683178	-73.965964
4	40.745168	-73.986831
5	40.750073	-73.998393
6	40.748549	-73.988084
7	40.739323	-74.008119
8	40.762272	-73.987882
9	40.748238	-73.978311
10	40.713126	-73.984844
11	40.751581	-73.977910
12	40.760203	-73.964785
13	40.716059	-73.991908
14	40.724537	-73.981854
15	40.756458	-73.993722
16	40.734546	-73.990741
17	40.743954	-73.991449
18	40.707179	-74.008873
19	40.750380	-73.983390
20	40.764618	-73.987895
21	40.745168	-73.986831
22	40.753231	-73.970325
23	40.757148	-73.972078
25	40.768254	-73.988639
27	40.734927	-73.992005
29	40.688226	-73.979382
30	40.734546	-73.990741
31	40.734546	-73.990741
32	40.715816	-73.994224
...
3942431	40.730477	-73.999061
3942432	40.745712	-73.981948
3942433	40.737262	-73.992390
3942434	40.730477	-73.999061
3942436	40.741444	-73.975361
3942437	40.730473	-73.986724
3942438	40.723684	-73.975748
3942439	40.719105	-73.999733

3942440	40.726281	-73.989780
3942442	40.715143	-73.944507
3942443	40.750967	-73.994442
3942444	40.742065	-74.004432
3942445	40.726281	-73.989780
3942446	40.750967	-73.994442
3942447	40.726281	-73.989780
3942448	40.750967	-73.994442
3942449	40.742388	-73.997262
3942450	40.728419	-73.987140
3942451	40.707678	-73.940162
3942452	40.727434	-73.993790
3942454	40.732219	-73.981656
3942459	40.750200	-73.990931
3942461	40.749156	-73.991600
3942463	40.734546	-73.990741
3942464	40.751551	-73.993934
3942465	40.748549	-73.988084
3942466	40.724910	-74.001547
3942467	40.730473	-73.986724
3942468	40.730473	-73.986724
3942469	40.765265	-73.981923

	end station latitude	end station longitude	bikeid	birth year	\
0	40.722293	-73.991475	18660	1960.0	
1	40.739355	-73.999318	16085	1963.0	
2	40.749013	-73.988484	20845	1974.0	
3	40.688515	-73.964763	19610	1969.0	
4	40.726218	-73.983799	20197	1977.0	
5	40.735238	-74.000271	20788	1969.0	
6	40.745168	-73.986831	19006	1972.0	
7	40.738177	-73.977387	17640	1985.0	
8	40.756458	-73.993722	15691	1991.0	
9	40.738177	-73.977387	17837	1991.0	
10	40.721816	-73.997203	16947	1979.0	
11	40.741473	-73.983209	14807	1980.0	
12	40.739126	-73.979738	16702	1987.0	
13	40.727791	-73.985649	17342	1988.0	
14	40.739445	-73.976806	19909	1983.0	
15	40.743174	-74.003664	19584	1979.0	
16	40.722055	-73.989111	19202	1959.0	
17	40.754557	-73.965930	20683	1971.0	
18	40.697601	-73.993446	21554	1977.0	
19	40.723180	-73.994800	16894	1980.0	
20	40.745497	-74.001971	14598	1972.0	
21	40.743943	-73.979661	18834	1962.0	
22	40.754666	-73.991382	15617	1987.0	
23	40.747804	-73.973442	15812	1988.0	

25	40.756458	-73.993722	16761	1969.0
27	40.763406	-73.977225	18320	1996.0
29	40.693083	-73.971789	20804	1986.0
30	40.739017	-74.002638	15777	1986.0
31	40.739017	-74.002638	21338	1986.0
32	40.714067	-73.992939	20781	1982.0
...
3942431	40.733320	-73.995101	18778	1978.0
3942432	40.748238	-73.978311	15163	1988.0
3942433	40.730477	-73.999061	19062	1981.0
3942434	40.722055	-73.989111	22597	1987.0
3942436	40.745168	-73.986831	21287	1960.0
3942437	40.732219	-73.981656	22923	1994.0
3942438	40.738177	-73.977387	23153	1987.0
3942439	40.720196	-73.989978	18911	1990.0
3942440	40.729538	-73.984267	23705	1982.0
3942442	40.723250	-73.943080	24078	1982.0
3942443	40.739355	-73.999318	23197	1967.0
3942444	40.725029	-73.990697	17391	1983.0
3942445	40.729538	-73.984267	22171	1982.0
3942446	40.762272	-73.987882	23482	1978.0
3942447	40.729538	-73.984267	22120	1983.0
3942448	40.762272	-73.987882	18211	1980.0
3942449	40.735354	-74.004831	22527	1979.0
3942450	40.720828	-73.977932	24230	1963.0
3942451	40.714133	-73.952344	22115	1980.0
3942452	40.713079	-73.998512	17554	1995.0
3942454	40.733143	-73.975739	23553	1982.0
3942459	40.734546	-73.990741	18253	1959.0
3942461	40.744449	-73.983035	16859	1962.0
3942463	40.766638	-73.953483	21038	1976.0
3942464	40.759291	-73.988597	15238	1971.0
3942465	40.718939	-73.992663	17292	1960.0
3942466	40.716059	-73.991908	23374	1982.0
3942467	40.724910	-74.001547	23503	1990.0
3942468	40.724910	-74.001547	22104	1980.0
3942469	40.744876	-73.995299	14597	1967.0

	gender	Age	Gender
0	2	58.0	Female
1	1	55.0	Male
2	1	44.0	Male
3	1	49.0	Male
4	1	41.0	Male
5	2	49.0	Female
6	1	46.0	Male
7	2	33.0	Female
8	1	27.0	Male

9	1	27.0	Male
10	1	39.0	Male
11	1	38.0	Male
12	1	31.0	Male
13	1	30.0	Male
14	1	35.0	Male
15	1	39.0	Male
16	1	59.0	Male
17	1	47.0	Male
18	1	41.0	Male
19	1	38.0	Male
20	1	46.0	Male
21	1	56.0	Male
22	1	31.0	Male
23	1	30.0	Male
25	1	49.0	Male
27	2	22.0	Female
29	1	32.0	Male
30	1	32.0	Male
31	1	32.0	Male
32	1	36.0	Male
...
3942431	1	40.0	Male
3942432	1	30.0	Male
3942433	1	37.0	Male
3942434	1	31.0	Male
3942436	1	58.0	Male
3942437	2	24.0	Female
3942438	1	31.0	Male
3942439	2	28.0	Female
3942440	1	36.0	Male
3942442	1	36.0	Male
3942443	1	51.0	Male
3942444	2	35.0	Female
3942445	2	36.0	Female
3942446	1	40.0	Male
3942447	2	35.0	Female
3942448	2	38.0	Female
3942449	1	39.0	Male
3942450	1	55.0	Male
3942451	1	38.0	Male
3942452	1	23.0	Male
3942454	2	36.0	Female
3942459	1	59.0	Male
3942461	1	56.0	Male
3942463	1	42.0	Male
3942464	1	47.0	Male
3942465	1	58.0	Male

```

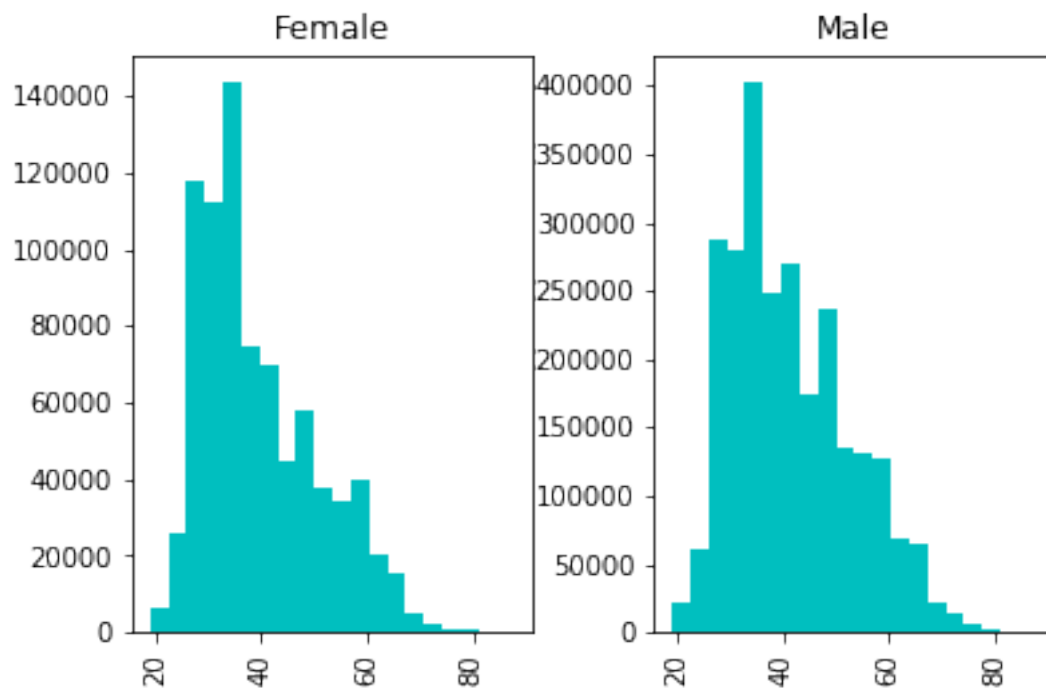
3942466      1  36.0    Male
3942467      1  28.0    Male
3942468      2  38.0  Female
3942469      1  51.0    Male

```

```
[3372554 rows x 12 columns]
```

```
In [5]: citi.hist('Age', by='Gender', bins=20, color='c')
```

```
Out[5]: array([<matplotlib.axes._subplots.AxesSubplot object at 0x1a486a66a0>,
               <matplotlib.axes._subplots.AxesSubplot object at 0x10ea44710>],
            dtype=object)
```

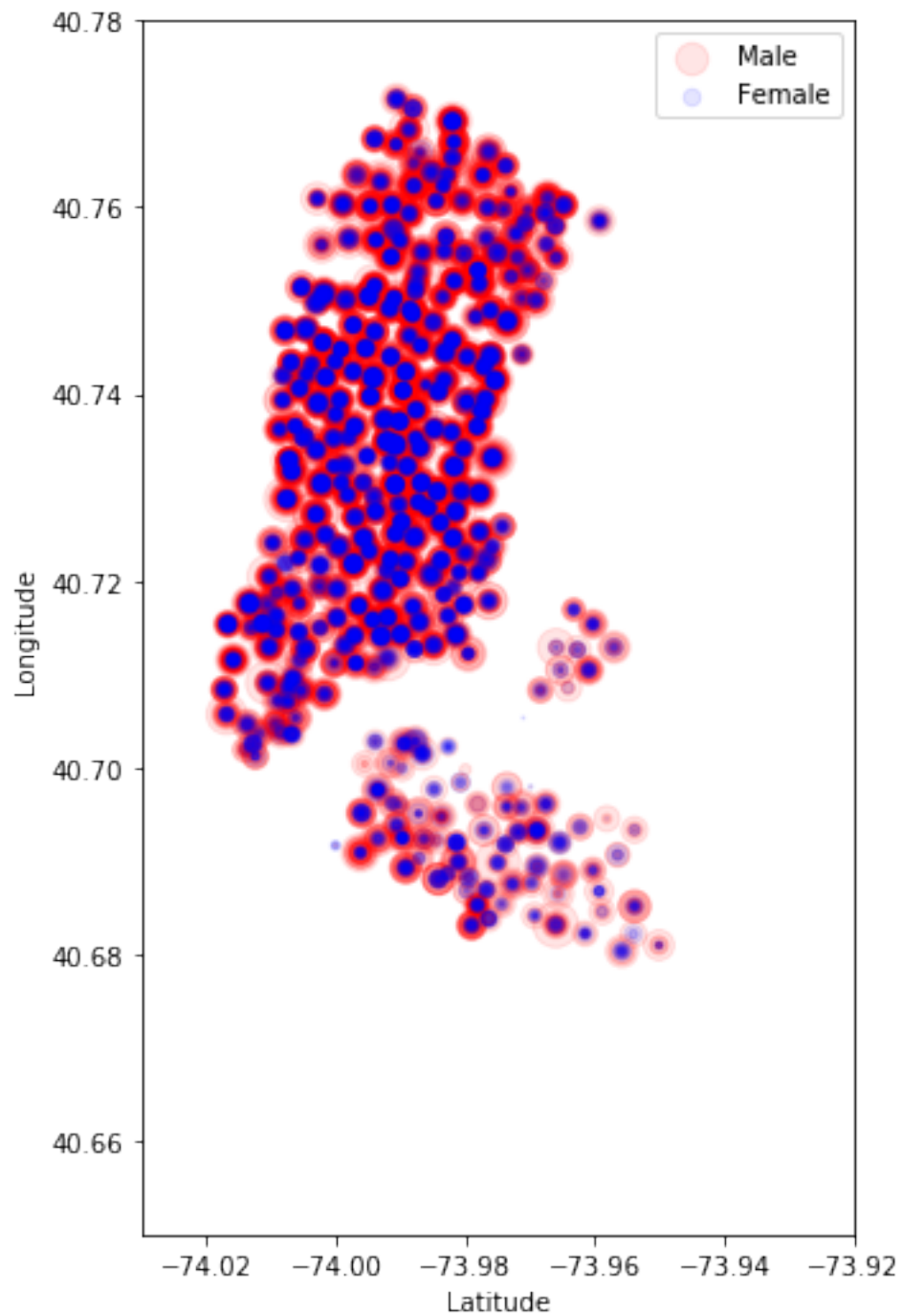


```

In [6]: citi_male = citi[citi['Gender']=='Male']
        citi_female = citi[citi['Gender']=='Female']
        plt.figure(figsize=(5,5*1.75))
        plt.scatter(citi_male['start station longitude'][:10000],citi_male['start station latitude'][:10000])
        plt.scatter(citi_female['start station longitude'][:10000],citi_female['start station latitude'][:10000])
        plt.axis([-74.03,-73.92,40.65,40.78])
        plt.xlabel('Latitude')
        plt.legend()
        plt.ylabel('Longitude')
        plt.suptitle('Heat Map Citibike Pick-ups by Gender', size=16, y=0.94)
        plt.style.use('bmh')
        plt.show()

```

Heat Map Citibike Pick-ups by Gender

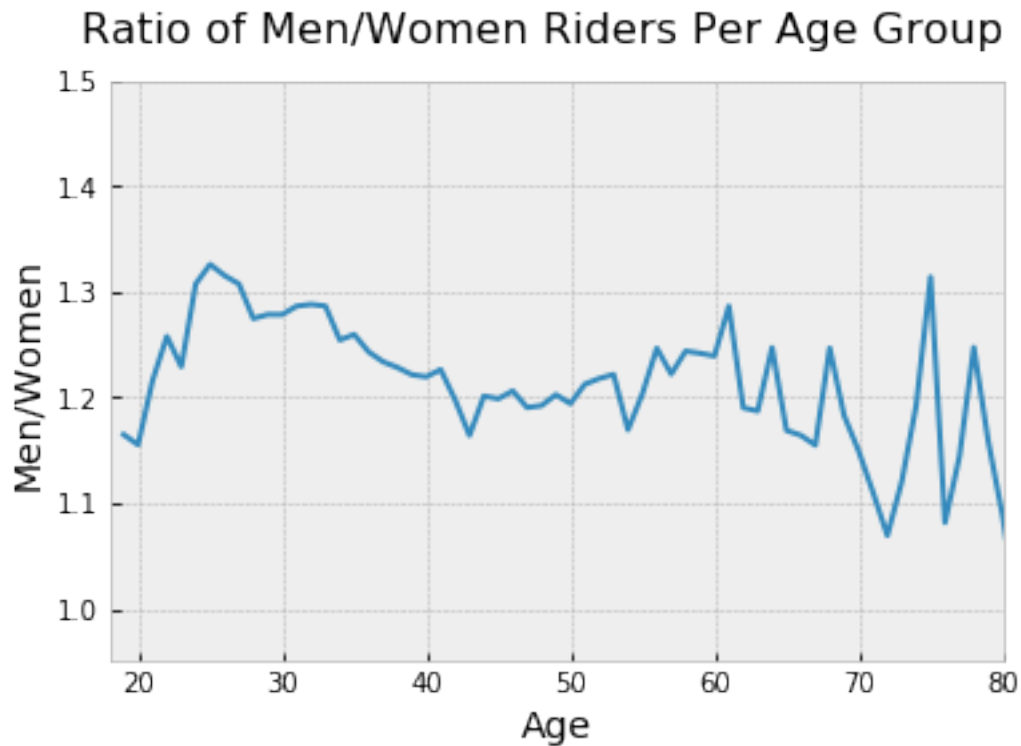


```
In [7]: plt.plot(citi.groupby('Age')['gender'].mean())  
plt.axis([18, 80, 0.95, 1.5])  
plt.style.use('fivethirtyeight')
```



```
plt.xlabel('Age', size=14)
plt.ylabel('Men/Women', size=14)
plt.suptitle('Ratio of Men/Women Riders Per Age Group', size=16, y=0.97)
```

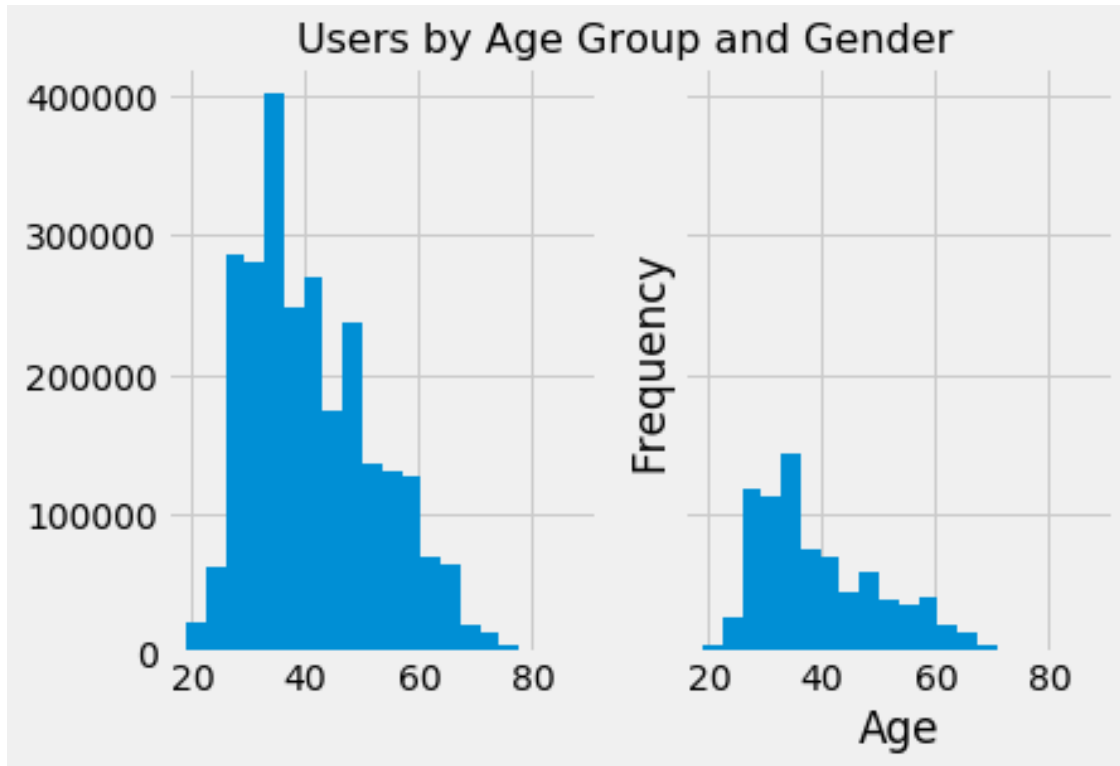
Out[7]: Text(0.5,0.97,'Ratio of Men/Women Riders Per Age Group')



```
In [8]: x=citi_female['Age']
        y=citi_male['Age']

fig, ax = plt.subplots(1, 2, sharey=True)
n_bins=20
plt.xlabel('Age')
plt.ylabel('Frequency')

ax[0].hist(y, bins=n_bins)
ax[1].hist(x, bins=n_bins)
plt.suptitle('Users by Age Group and Gender', size=16, y=0.94)
plt.show()
```



```
In [9]: citi_gender=pd.DataFrame()
citi_gender['sum_duration']=citi.groupby(['Gender']).tripduration.sum()
citi_gender['count']=citi.groupby(['Gender']).size()
citi_gender['duration_capita']=citi_gender['sum_duration']/citi_gender['count']*(len(c
citi_gender['cost_capita']=37.25/citi_gender['duration_capita']*60
citi_gender.reset_index(level=0,inplace=True)
citi_gender['Gender2']=citi_gender['Gender'].map({'Male':1,'Female':0})
# citi_gender['count']=citi.groupby(['Gender'])['tripduration'].count()
citi_gender
```

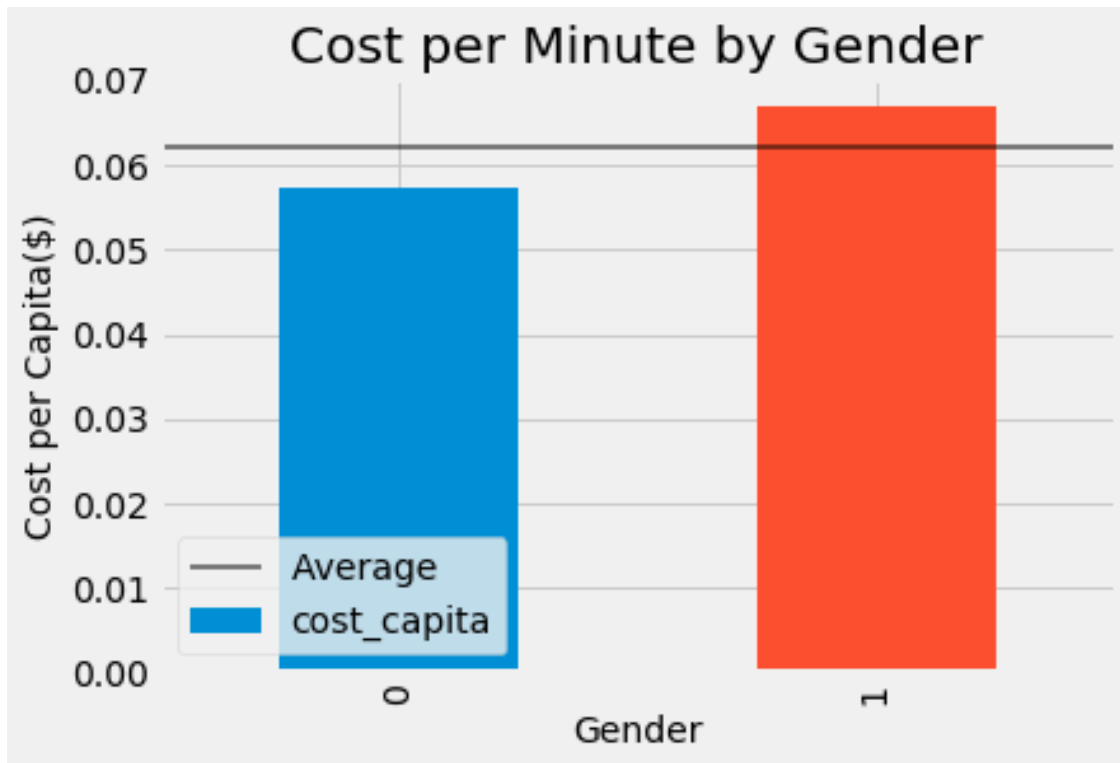
```
Out[9]:   Gender  sum_duration    count  duration_capita  cost_capita  Gender2
0  Female    670605848    811170    39033.707008    0.057258         0
1   Male    1809229479   2556766    33410.810800    0.066895         1
```

```
In [10]: citi_gender.shape
```

```
Out[10]: (2, 6)
```

```
In [11]: fig1, ax1=plt.subplots()
citi_gender['cost_capita'].plot(ax=ax1,kind='bar')
plt.axhline(citi_gender['cost_capita'].mean(),linewidth=2.0,color='black',alpha=0.5,
plt.legend()
plt.title('Cost per Minute by Gender')
plt.xlabel('Gender', size=14)
plt.ylabel('Cost per Capita($)', size=14)
```

```
Out[11]: Text(0,0.5,'Cost per Capita($)')
```



```
In [12]: print(smf.ols('cost_capita ~ Gender2',data=citi_gender).fit().summary())
```

```

                        OLS Regression Results
=====
Dep. Variable:          cost_capita    R-squared:                1.000
Model:                  OLS           Adj. R-squared:             nan
Method:                 Least Squares  F-statistic:               0.000
Date:                  Thu, 20 Dec 2018  Prob (F-statistic):       nan
Time:                  18:25:08         Log-Likelihood:            75.265
No. Observations:      2              AIC:                     -146.5
Df Residuals:          0              BIC:                     -149.1
Df Model:              1
Covariance Type:       nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.0573	inf	0	nan	nan	nan
Gender2	0.0096	inf	0	nan	nan	nan

```

=====
Omnibus:                nan    Durbin-Watson:                0.200
Prob(Omnibus):          nan    Jarque-Bera (JB):          0.333

```

Skew:	0.000	Prob(JB):	0.846
Kurtosis:	1.000	Cond. No.	2.62

=====

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```

/Users/MartinSmit/anaconda3/lib/python3.6/site-packages/statsmodels/stats/stattools.py:72: Val
"samples were given." % int(n), ValueWarning)
/Users/MartinSmit/anaconda3/lib/python3.6/site-packages/statsmodels/regression/linear_model.py
return 1 - np.divide(self.nobs - self.k_constant, self.df_resid) * (1 - self.rsquared)
/Users/MartinSmit/anaconda3/lib/python3.6/site-packages/statsmodels/regression/linear_model.py
return 1 - np.divide(self.nobs - self.k_constant, self.df_resid) * (1 - self.rsquared)
/Users/MartinSmit/anaconda3/lib/python3.6/site-packages/statsmodels/regression/linear_model.py
return self.ssr/self.df_resid
/Users/MartinSmit/anaconda3/lib/python3.6/site-packages/statsmodels/regression/linear_model.py
return np.dot(wresid, wresid) / self.df_resid

```

```
In [13]: print(smf.ols('tripduration ~ Age + gender',data=citi).fit().summary())
```

OLS Regression Results

```

=====
Dep. Variable:      tripduration    R-squared:                0.011
Model:              OLS            Adj. R-squared:            0.011
Method:             Least Squares   F-statistic:             1.927e+04
Date:               Thu, 20 Dec 2018 Prob (F-statistic):       0.00
Time:               18:25:10        Log-Likelihood:          -2.6006e+07
No. Observations:   3372554        AIC:                     5.201e+07
Df Residuals:       3372551        BIC:                     5.201e+07
Df Model:           2
Covariance Type:    nonrobust

```

```

=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept    483.6462      1.434     337.383      0.000     480.837     486.456
Age           2.4898      0.026     96.298      0.000         2.439         2.540
gender       121.5974      0.686    177.197      0.000     120.252     122.942
=====
Omnibus:            1883764.503    Durbin-Watson:           1.916
Prob(Omnibus):      0.000        Jarque-Bera (JB):        24671729.627
Skew:               2.431        Prob(JB):                0.00
Kurtosis:           15.326        Cond. No.:               218.
=====

```

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.