final_project

December 19, 2019

[40]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import statsmodels.formula.api as smf
df = pd.read_csv('suicide_rates.csv')
```

1 The project uses a data set of suicides number and many other factors to explore what factors has the most significant impact on the suicides rate.

[41]:
```
df.columns = ['country', 'year', 'sex', 'age', 'suicides_no', 'population',
       'suicides_rate', 'country-year', 'HDI',
       'gdp_for_year', 'gdp_per_capita', 'generation']
```

[42]:
```
# the data in year 2016 is not complete, so it should be dropped
df = df.loc[df['year']!=2016,:]
df['gdp_for_year'] = df['gdp_for_year'].str.replace(',','').astype(int)
df_year = df.groupby('year')['suicides_no'].sum()
```

[43]:
```
df
```

[43]:
```
          country  year     sex          age  suicides_no  population  \
0         Albania  1987    male  15-24 years           21      312900
1         Albania  1987    male  35-54 years           16      308000
2         Albania  1987  female  15-24 years           14      289700
3         Albania  1987    male    75+ years            1       21800
4         Albania  1987    male  25-34 years            9      274300
5         Albania  1987  female    75+ years            1       35600
6         Albania  1987  female  35-54 years            6      278800
7         Albania  1987  female  25-34 years            4      257200
8         Albania  1987    male  55-74 years            1      137500
9         Albania  1987  female   5-14 years            0      311000
10        Albania  1987  female  55-74 years            0      144600
11        Albania  1987    male   5-14 years            0      338200
12        Albania  1988  female    75+ years            2       36400
```

```
13        Albania  1988    male  15-24 years           17      319200
14        Albania  1988    male    75+ years            1       22300
15        Albania  1988    male  35-54 years           14      314100
16        Albania  1988    male  55-74 years            4      140200
17        Albania  1988  female  15-24 years            8      295600
18        Albania  1988  female  55-74 years            3      147500
19        Albania  1988  female  25-34 years            5      262400
20        Albania  1988    male  25-34 years            5      279900
21        Albania  1988  female  35-54 years            4      284500
22        Albania  1988  female   5-14 years            0      317200
23        Albania  1988    male   5-14 years            0      345000
24        Albania  1989    male    75+ years            2       22500
25        Albania  1989    male  25-34 years           18      283600
26        Albania  1989    male  35-54 years           15      318400
27        Albania  1989    male  55-74 years            6      142100
28        Albania  1989    male  15-24 years           12      323500
29        Albania  1989  female  35-54 years            7      288600
...           ...   ...     ...          ...          ...         ...
27790  Uzbekistan  2012  female  25-34 years          148     2556673
27791  Uzbekistan  2012  female  35-54 years           89     3474788
27792  Uzbekistan  2012    male   5-14 years           67     2701361
27793  Uzbekistan  2012  female  55-74 years           25     1283060
27794  Uzbekistan  2012  female    75+ years            4      338557
27795  Uzbekistan  2012  female   5-14 years           16     2578408
27796  Uzbekistan  2013    male  35-54 years          481     3346411
27797  Uzbekistan  2013    male  25-34 years          328     2644648
27798  Uzbekistan  2013  female  15-24 years          323     3039740
27799  Uzbekistan  2013    male  15-24 years          320     3171202
27800  Uzbekistan  2013    male  55-74 years          119     1202790
27801  Uzbekistan  2013    male    75+ years           13      221002
27802  Uzbekistan  2013  female  25-34 years          146     2647820
27803  Uzbekistan  2013  female  35-54 years           99     3547895
27804  Uzbekistan  2013  female    75+ years            8      345180
27805  Uzbekistan  2013    male   5-14 years           61     2720938
27806  Uzbekistan  2013  female  55-74 years           21     1356298
27807  Uzbekistan  2013  female   5-14 years           31     2595000
27808  Uzbekistan  2014    male  35-54 years          519     3421300
27809  Uzbekistan  2014    male  25-34 years          318     2739150
27810  Uzbekistan  2014  female  15-24 years          347     2992817
27811  Uzbekistan  2014    male  55-74 years          144     1271111
27812  Uzbekistan  2014    male  15-24 years          347     3126905
27813  Uzbekistan  2014    male    75+ years           17      224995
27814  Uzbekistan  2014  female  25-34 years          162     2735238
27815  Uzbekistan  2014  female  35-54 years          107     3620833
27816  Uzbekistan  2014  female    75+ years            9      348465
27817  Uzbekistan  2014    male   5-14 years           60     2762158
27818  Uzbekistan  2014  female   5-14 years           44     2631600
```

```
27819  Uzbekistan  2014  female  55-74 years           21     1438935

       suicides_rate    country-year    HDI  gdp_for_year  gdp_per_capita  \
0               6.71     Albania1987    NaN    2156624900             796
1               5.19     Albania1987    NaN    2156624900             796
2               4.83     Albania1987    NaN    2156624900             796
3               4.59     Albania1987    NaN    2156624900             796
4               3.28     Albania1987    NaN    2156624900             796
5               2.81     Albania1987    NaN    2156624900             796
6               2.15     Albania1987    NaN    2156624900             796
7               1.56     Albania1987    NaN    2156624900             796
8               0.73     Albania1987    NaN    2156624900             796
9               0.00     Albania1987    NaN    2156624900             796
10              0.00     Albania1987    NaN    2156624900             796
11              0.00     Albania1987    NaN    2156624900             796
12              5.49     Albania1988    NaN    2126000000             769
13              5.33     Albania1988    NaN    2126000000             769
14              4.48     Albania1988    NaN    2126000000             769
15              4.46     Albania1988    NaN    2126000000             769
16              2.85     Albania1988    NaN    2126000000             769
17              2.71     Albania1988    NaN    2126000000             769
18              2.03     Albania1988    NaN    2126000000             769
19              1.91     Albania1988    NaN    2126000000             769
20              1.79     Albania1988    NaN    2126000000             769
21              1.41     Albania1988    NaN    2126000000             769
22              0.00     Albania1988    NaN    2126000000             769
23              0.00     Albania1988    NaN    2126000000             769
24              8.89     Albania1989    NaN    2335124988             833
25              6.35     Albania1989    NaN    2335124988             833
26              4.71     Albania1989    NaN    2335124988             833
27              4.22     Albania1989    NaN    2335124988             833
28              3.71     Albania1989    NaN    2335124988             833
29              2.43     Albania1989    NaN    2335124988             833
...              ...             ...    ...           ...             ...
27790           5.79  Uzbekistan2012  0.668   51821573338            1964
27791           2.56  Uzbekistan2012  0.668   51821573338            1964
27792           2.48  Uzbekistan2012  0.668   51821573338            1964
27793           1.95  Uzbekistan2012  0.668   51821573338            1964
27794           1.18  Uzbekistan2012  0.668   51821573338            1964
27795           0.62  Uzbekistan2012  0.668   51821573338            1964
27796          14.37  Uzbekistan2013  0.672   57690453461            2150
27797          12.40  Uzbekistan2013  0.672   57690453461            2150
27798          10.63  Uzbekistan2013  0.672   57690453461            2150
27799          10.09  Uzbekistan2013  0.672   57690453461            2150
27800           9.89  Uzbekistan2013  0.672   57690453461            2150
27801           5.88  Uzbekistan2013  0.672   57690453461            2150
27802           5.51  Uzbekistan2013  0.672   57690453461            2150
```

```
27803           2.79  Uzbekistan2013  0.672   57690453461            2150
27804           2.32  Uzbekistan2013  0.672   57690453461            2150
27805           2.24  Uzbekistan2013  0.672   57690453461            2150
27806           1.55  Uzbekistan2013  0.672   57690453461            2150
27807           1.19  Uzbekistan2013  0.672   57690453461            2150
27808          15.17  Uzbekistan2014  0.675   63067077179            2309
27809          11.61  Uzbekistan2014  0.675   63067077179            2309
27810          11.59  Uzbekistan2014  0.675   63067077179            2309
27811          11.33  Uzbekistan2014  0.675   63067077179            2309
27812          11.10  Uzbekistan2014  0.675   63067077179            2309
27813           7.56  Uzbekistan2014  0.675   63067077179            2309
27814           5.92  Uzbekistan2014  0.675   63067077179            2309
27815           2.96  Uzbekistan2014  0.675   63067077179            2309
27816           2.58  Uzbekistan2014  0.675   63067077179            2309
27817           2.17  Uzbekistan2014  0.675   63067077179            2309
27818           1.67  Uzbekistan2014  0.675   63067077179            2309
27819           1.46  Uzbekistan2014  0.675   63067077179            2309

            generation
0         Generation X
1               Silent
2         Generation X
3      G.I. Generation
4              Boomers
5      G.I. Generation
6               Silent
7              Boomers
8      G.I. Generation
9         Generation X
10     G.I. Generation
11        Generation X
12     G.I. Generation
13        Generation X
14     G.I. Generation
15              Silent
16     G.I. Generation
17        Generation X
18     G.I. Generation
19             Boomers
20             Boomers
21              Silent
22        Generation X
23        Generation X
24     G.I. Generation
25             Boomers
26              Silent
27     G.I. Generation
```

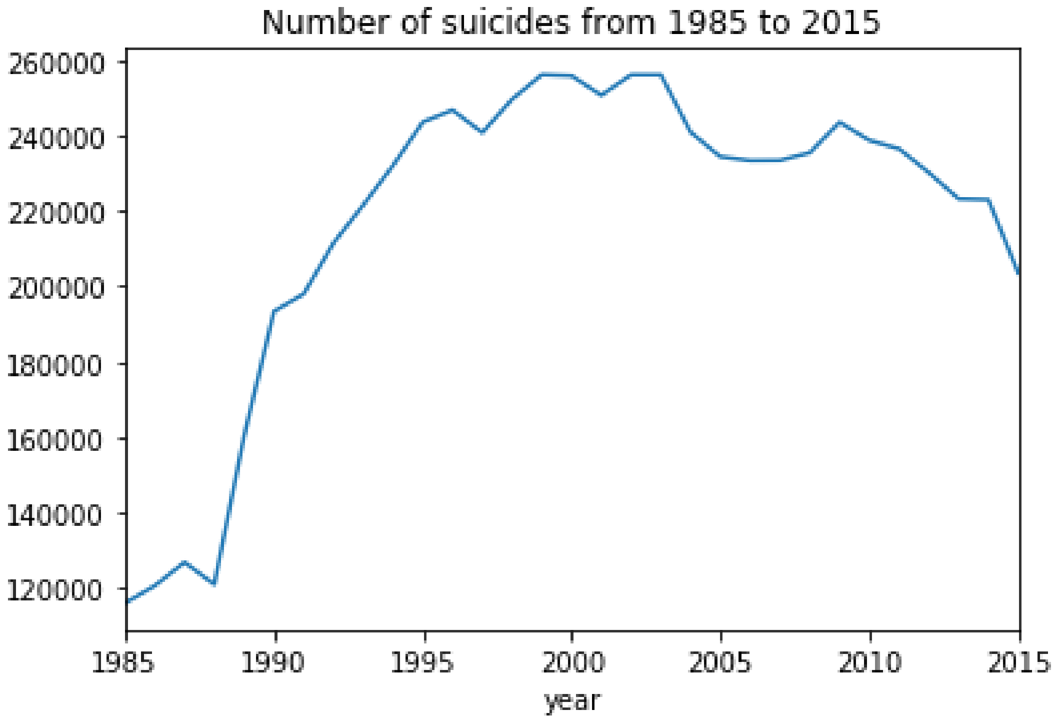
```
28       Generation X
29             Silent
...               ...
27790      Millenials
27791    Generation X
27792    Generation Z
27793         Boomers
27794          Silent
27795    Generation Z
27796    Generation X
27797      Millenials
27798      Millenials
27799      Millenials
27800         Boomers
27801          Silent
27802      Millenials
27803    Generation X
27804          Silent
27805    Generation Z
27806         Boomers
27807    Generation Z
27808    Generation X
27809      Millenials
27810      Millenials
27811         Boomers
27812      Millenials
27813          Silent
27814      Millenials
27815    Generation X
27816          Silent
27817    Generation Z
27818    Generation Z
27819         Boomers

[27660 rows x 12 columns]
```

[44]:

```
# show the change in suicides number per year
fig, ax = plt.subplots()
df_year.plot(x='year', y='suicides_no', ax = ax)
ax.set_title('Number of suicides from 1985 to 2015')
```

[44]:

```
Text(0.5, 1.0, 'Number of suicides from 1985 to 2015')
```



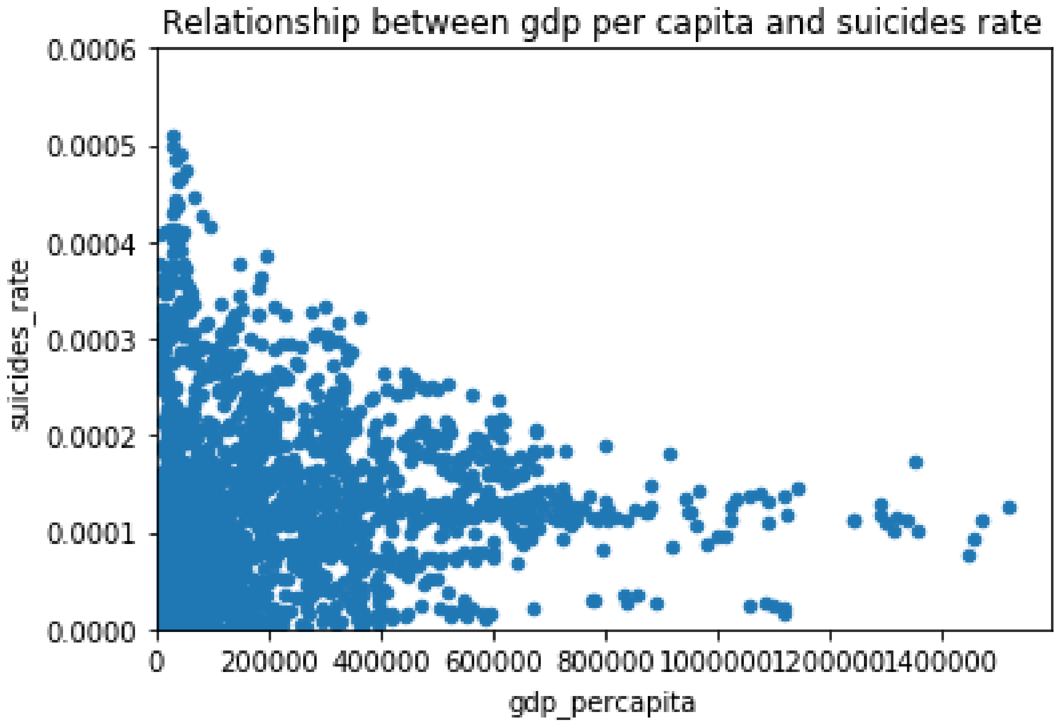
[45]:
```
df_country_year = df.groupby('country-year')['suicides_no'].sum()
df_population = df.groupby('country-year')['population'].sum()
df_country_pgdp = df.groupby('country-year')['gdp_per_capita'].sum()
df_country_gdp = df.groupby('country-year')['gdp_for_year'].sum()
```

[46]:
```
# explore the relationship between gdp per capita and suicides rate
suicide_pgdp = pd.DataFrame()
```

[47]:
```
suicide_pgdp['gdp_percapita'] = df_country_pgdp
suicide_pgdp['suicides_rate'] = df_country_year / df_population
```

[48]:
```
fig, ax = plt.subplots()
suicide_pgdp.plot.scatter(x='gdp_percapita', y='suicides_rate', ax=ax)
ax.set_ybound(lower=0, upper=0.0006)
ax.set_xbound(lower=0)
ax.set_title('Relationship between gdp per capita and suicides rate')
```

[48]:
```
Text(0.5, 1.0, 'Relationship between gdp per capita and suicides rate')
```



[49]:
```
reg = smf.ols('suicides_rate ~ gdp_percapita', data=suicide_pgdp).fit()
print(reg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          suicides_rate   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     8.542
Date:                Thu, 19 Dec 2019   Prob (F-statistic):            0.00351
Time:                        10:21:21   Log-Likelihood:                  18226.
No. Observations:                2305   AIC:                        -3.645e+04
Df Residuals:                    2303   BIC:                        -3.644e+04
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
=
                    coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-
Intercept         0.0001   2.49e-06     45.227      0.000       0.000
0.000
gdp_percapita  2.397e-11    8.2e-12      2.923      0.004    7.89e-12
4.01e-11
```

```
==============================================================================
Omnibus:                      440.100   Durbin-Watson:                   0.125
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              783.348
Skew:                           1.199   Prob(JB):                    7.91e-171
Kurtosis:                       4.550   Cond. No.                     4.06e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 4.06e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The t-value and p-value told us that there is a relationship between suicides rate and gdp_percapita, but we can find that the R_squared is very lower, so we may conclude that the gdp_percapita does not influence the suicides rate a lot.

[50]:

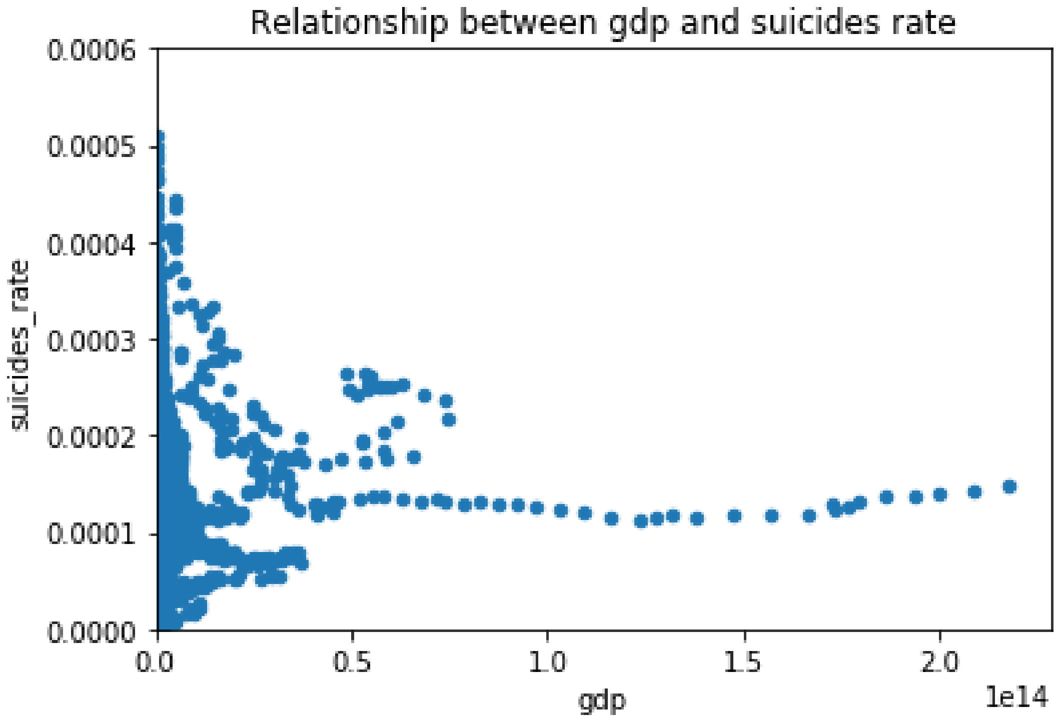
```
# explore the relationship between gdp and suicides rate
suicide_gdp = pd.DataFrame()
suicide_gdp['gdp'] = df_country_gdp
suicide_gdp['suicides_rate'] = df_country_year / df_population
suicide_gdp
```

[50]:

```
                                  gdp  suicides_rate
country-year
Albania1987               25879498800       0.000027
Albania1988               25512000000       0.000023
Albania1989               28021499856       0.000024
Albania1992                8513431008       0.000017
Albania1993               14736852456       0.000026
Albania1994               23828085576       0.000018
Albania1995               29093988108       0.000030
Albania1996               39778779504       0.000030
Albania1997               28318837296       0.000057
Albania1998               32485485264       0.000051
Albania1999               40977130980       0.000046
Albania2000               43584526896       0.000019
Albania2001               48729105648       0.000043
Albania2002               53220943776       0.000047
Albania2003               68963350956       0.000044
Albania2004               87778382112       0.000051
Albania2005               97902584604       0.000000
Albania2006              107911708188       0.000000
Albania2007              128412142764       0.000045
Albania2008              154576232256       0.000058
Albania2009              144530554848       0.000000
Albania2010              143123439108       0.000035
Antigua and Barbuda1985    2891087112       0.000000
Antigua and Barbuda1986    3485281776       0.000000
Antigua and Barbuda1987    4046098224       0.000000
Antigua and Barbuda1988    4783652892       0.000000
Antigua and Barbuda1989    5265537336       0.000000
Antigua and Barbuda1990    5513628888       0.000017
Antigua and Barbuda1991    5780475996       0.000000
Antigua and Barbuda1992    5991373776       0.000000
...                               ...            ...
Uruguay2007              280926871608       0.000187
Uruguay2008              364394557428       0.000169
Uruguay2009              379930935324       0.000164
Uruguay2010              483413779824       0.000175
Uruguay2012              615172681392       0.000190
Uruguay2013              690374800212       0.000173
Uruguay2014              686832157032       0.000186
Uruguay2015              639291650664       0.000197
Uzbekistan1990           160327295016       0.000085
Uzbekistan1991           164131466664       0.000080
Uzbekistan1992           155295568512       0.000074
Uzbekistan1993           157188166032       0.000073
Uzbekistan1994           154789883892       0.000074
Uzbekistan1995           160205627004       0.000076
Uzbekistan1996           167386706592       0.000086
Uzbekistan1997           176935245288       0.000076
Uzbekistan1998           179867654532       0.000078
Uzbekistan1999           204941591784       0.000084
Uzbekistan2000           165124493856       0.000088
Uzbekistan2001           136816217040       0.000086
Uzbekistan2002           116255412660       0.000070
Uzbekistan2003           121537348812       0.000062
Uzbekistan2004           144360282576       0.000054
Uzbekistan2005           171690118068       0.000052
Uzbekistan2009           404270684076       0.000055
Uzbekistan2010           471993251148       0.000057
Uzbekistan2011           550982294268       0.000063
Uzbekistan2012           621858880056       0.000070
Uzbekistan2013           692285441532       0.000073
Uzbekistan2014           756804926148       0.000077

[2305 rows x 2 columns]
```

[51]:
```
fig, ax = plt.subplots()
suicide_gdp.plot.scatter(x='gdp', y='suicides_rate', ax=ax)
ax.set_ybound(lower=0, upper=0.0006)
ax.set_xbound(lower=0)
ax.set_title('Relationship between gdp and suicides rate')
```

[51]:
```
Text(0.5, 1.0, 'Relationship between gdp and suicides rate')
```



[52]:

```
reg = smf.ols('suicides_rate ~ gdp', data=suicide_gdp).fit()
print(reg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          suicides_rate   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.005
Method:                 Least Squares   F-statistic:                     12.08
Date:                Thu, 19 Dec 2019   Prob (F-statistic):           0.000519
Time:                        10:21:24   Log-Likelihood:                  18228.
No. Observations:                2305   AIC:                        -3.645e+04
Df Residuals:                    2303   BIC:                        -3.644e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0001   1.94e-06     59.445      0.000       0.000       0.000
gdp         3.686e-19   1.06e-19      3.476      0.001    1.61e-19    5.77e-19
==============================================================================
Omnibus:                      422.152   Durbin-Watson:                   0.125
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              730.017
Skew:                           1.172   Prob(JB):                    3.01e-159
Kurtosis:                       4.453   Cond. No.                     1.91e+13
```

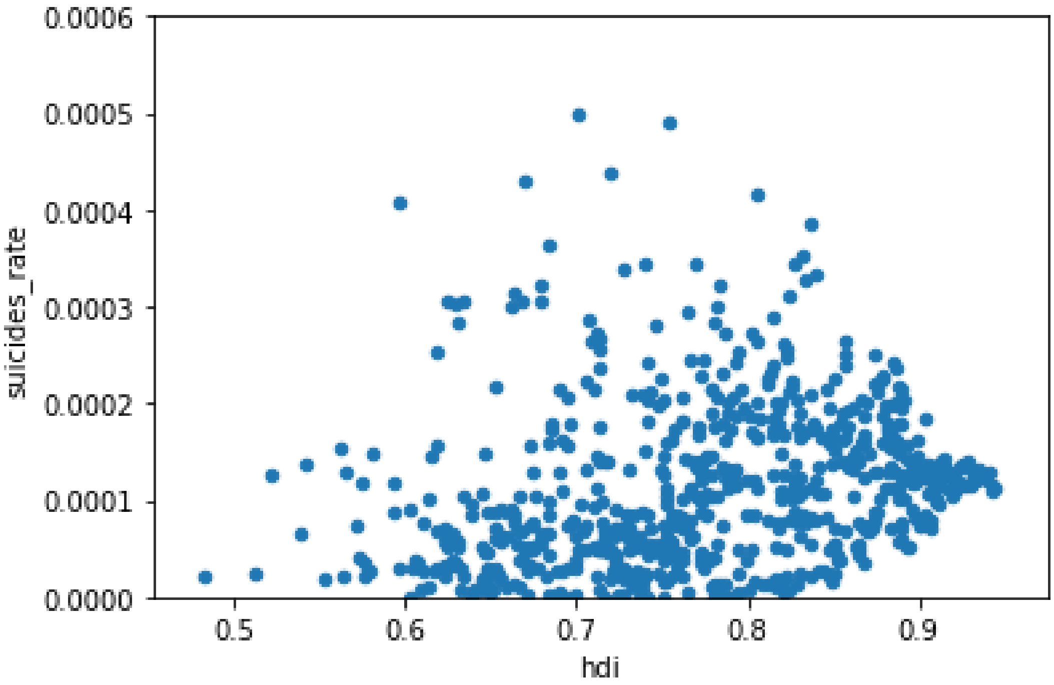
```
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.91e+13. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The t-value and p-value told us that there is a relationship between suicides rate and gdp, but we can find that the R_squared is very lower, so we may conclude that the gdp does not influence the suicides rate a lot.

[54]:
```python
df2 = df.dropna()
df2_suicide_rate = df2.groupby('country-year')['suicides_no'].sum() / df2.
 ↪groupby('country-year')['population'].sum()
df2_hdi = df2.groupby('country-year')['HDI'].mean()
```

[55]:
```python
# explore the relationship between hdi and suicides rate
suicide_hdi = pd.DataFrame()
suicide_hdi['hdi'] = df2_hdi
suicide_hdi['suicides_rate'] = df2_suicide_rate
```

[56]:
```python
fig, ax = plt.subplots()
suicide_hdi.plot.scatter(x='hdi', y='suicides_rate', ax=ax)
ax.set_ybound(lower=0, upper=0.0006)
```



[57]:
```
reg = smf.ols('suicides_rate ~ hdi', data=suicide_hdi).fit()
print(reg.summary())
```

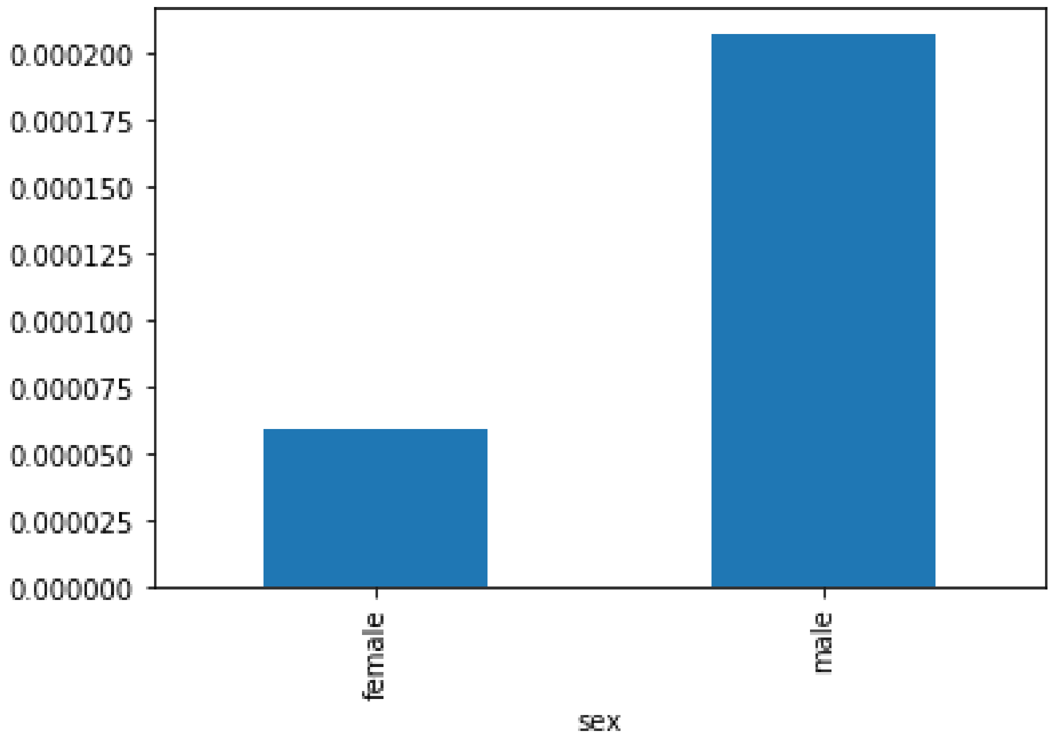
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          suicides_rate   R-squared:                       0.047
Model:                            OLS   Adj. R-squared:                  0.046
Method:                 Least Squares   F-statistic:                     34.23
Date:                Thu, 19 Dec 2019   Prob (F-statistic):           7.54e-09
Time:                        10:21:49   Log-Likelihood:                 5577.5
No. Observations:                 697   AIC:                        -1.115e+04
Df Residuals:                     695   BIC:                        -1.114e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3.629e-05    2.57e-05     -1.410      0.159   -8.68e-05    1.43e-05
hdi            0.0002    3.29e-05      5.851      0.000       0.000       0.000
==============================================================================
Omnibus:                      191.474   Durbin-Watson:                   0.411
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              470.773
Skew:                           1.424   Prob(JB):                    5.93e-103
Kurtosis:                       5.846   Cond. No.                         17.2
==============================================================================

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

The t-value and p-value told us that there is a relationship between suicides rate and hdi, and the R-square is higher, which is 0.047, but we still can not find a strong relationship between hdi and suicides rate.

[58]:
```
# explore the relationship between hdi and sex
male_female = df.groupby('sex')['suicides_no'].sum() / df.
 ↪groupby('sex')['population'].sum()
```

[59]:
```
male_female.plot.bar()
```

[59]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1a1c8e10>
```

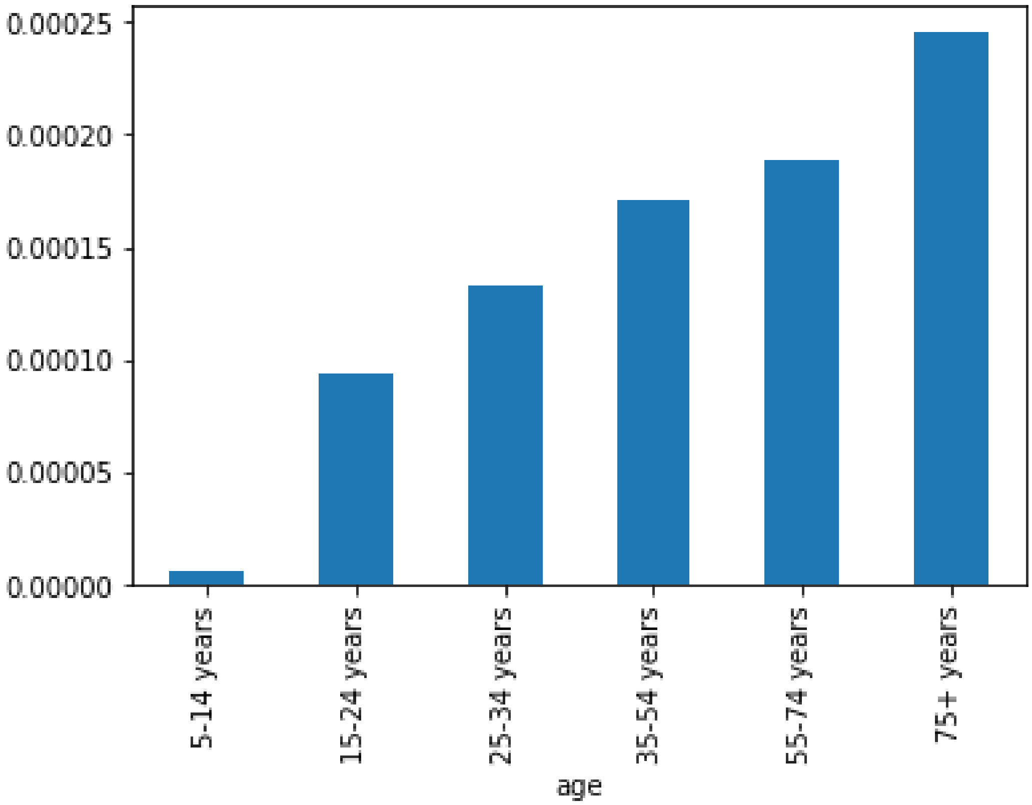


We can find a clear distinction between the suicides rate of male and female, this tells us that male has four times the possibility of suicide than female.

[60]:
```
ages = df.groupby('age')['suicides_no'].sum() / df.groupby('age')['population'].
 ↪sum()
```

[61]:
```
ages.sort_values().plot.bar()
```

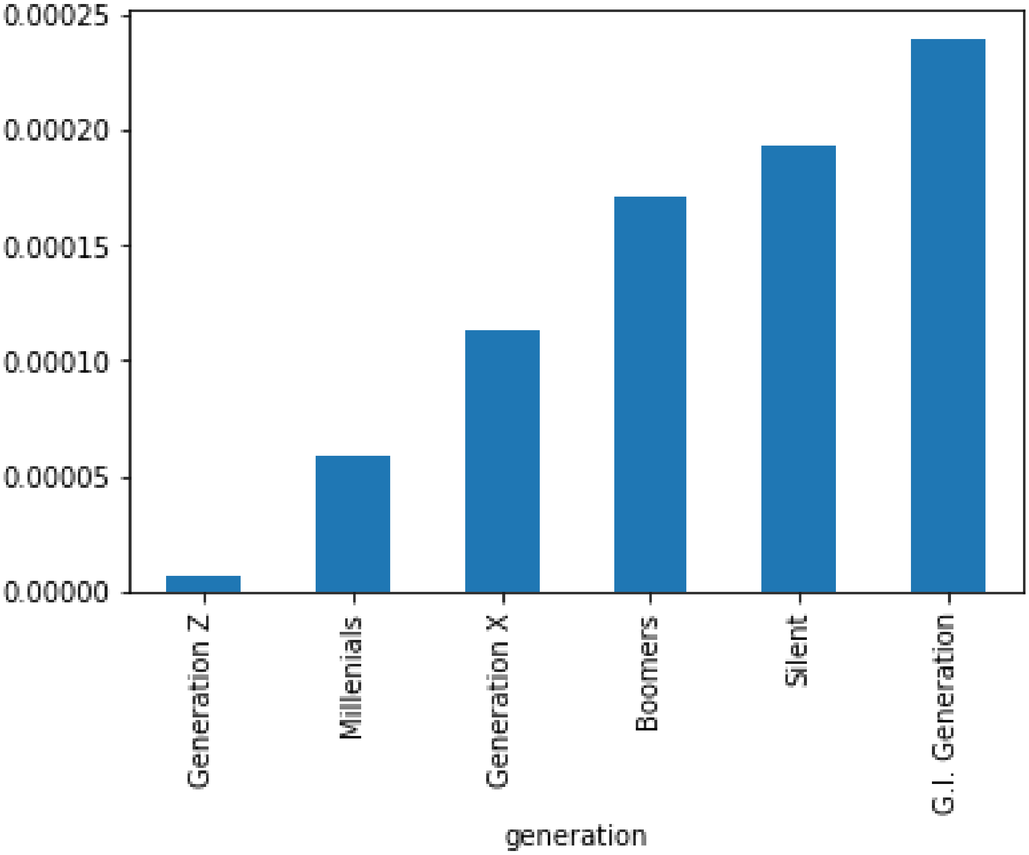
[61]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1ac8a6d8>
```



We can also find a relationship between suicides rate and age, generally speaking, as age increases, the probability of suicide increases.

[62]:
```
ages = df.groupby('generation')['suicides_no'].sum() / df.
 ↪groupby('generation')['population'].sum()
ages.sort_values().plot.bar()
```

[62]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1accf3c8>
```



The relationship between suicides rate and generation just corresponds to that between suicides rate and age.

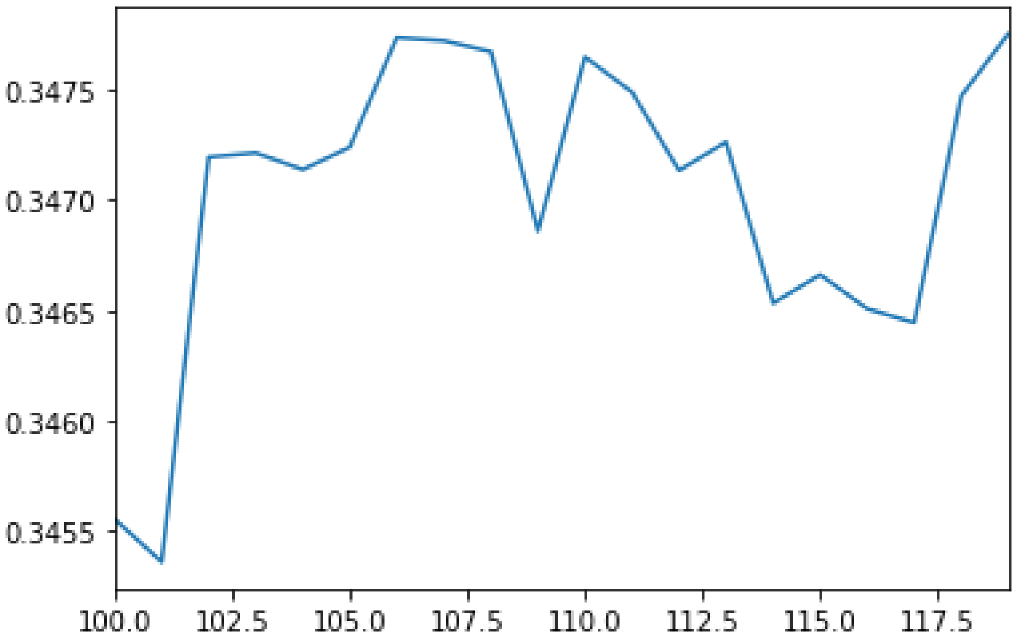
We have found out that sex and age influence suicides the most, so we are going to use these two factors to build KNeighborsRegreessor model to predict the suicides rate.

[66]:
```python
from patsy import dmatrices
from sklearn.neighbors import KNeighborsRegressor as knn
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import RandomForestRegressor as rf
```

[67]:
```python
# y, X = dmatrices('suicides_rate ~ HDI + gdp_for_year + gdp_per_capita', df2)
y, X = dmatrices('suicides_rate ~ sex + age', df2)
```

[68]:
```python
knn_scores = pd.Series()
for i in range(100, 120, 1):
    knn_scores.loc[i] = cross_val_score(knn(n_neighbors=i),X,np.ravel(y),cv=2).
 ↪mean()
knn_scores.plot()
```

[68]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a19fc9780>
```



[69]:
```
# We can find that the model is most useful when n_neighbors is set to 106
cross_val_score(knn(n_neighbors=106),X,np.ravel(y),cv=2).mean()
```

[69]:
```
0.3477382747360774
```

[70]:
```
knn(n_neighbors=106).fit(X,y).predict(X)
```

[70]:
```
array([[16.28245283],
       [26.47028302],
       [ 8.17698113],
       ...,
       [ 0.88415094],
       [ 0.52386792],
       [ 4.41745283]])
```