Untitled

November 22, 2019

1 question one

(a) Find the Polity IV dataset and import it into your Jupyter notebook using whatever method you prefer. Show the first five observations of the imported dataset.

[1]:

```
import pandas as pd #import pands, call it pd
import numpy as np #import numpy,call it np
import statsmodels.api as sm #import statsmodels.api, call it sm
import statsmodels.formula.api as smf #import statsmodels.formula.api, call it␣
 ↪smf
import matplotlib.pyplot as plt #import matplotlib.pyplot, call it plt
```

[2]:

```
url='https://raw.githubusercontent.com/ajr348/polity4/master/p4v2017.csv'
#import from github, name it url

data=pd.read_csv(url)
#read the data, call it data

pd.set_option('display.max_columns',36)
#display all columns of the data

pd.set_option('display.float_format','{:2.2f}'.format)
#don't show scientific notation

pd.set_option('precision',2)
#set the decimal precision output to 2 decimal places

data[:5]
#show the first five rows of the dataset
```

[2]:

```
   cyear  ccode scode        country  year  flag  fragment  democ  autoc  \
0  21800      2   USA  United States  1800     0       nan      7      3
1  21801      2   USA  United States  1801     0       nan      7      3
2  21802      2   USA  United States  1802     0       nan      7      3
3  21803      2   USA  United States  1803     0       nan      7      3
4  21804      2   USA  United States  1804     0       nan      7      3

   polity  polity2  durable  xrreg  xrcomp  xropen  xconst  parreg  parcomp  \
0       4     4.00      nan      3       3       4       7       4        2
```

```
1       4    4.00      nan       3       3       4       7       4        2
2       4    4.00      nan       3       3       4       7       4        2
3       4    4.00      nan       3       3       4       7       4        2
4       4    4.00      nan       3       3       4       7       4        2

   exrec  exconst  polcomp  prior  emonth  eday  eyear  eprec  interim  \
0   8.00        7     2.00    nan     nan   nan    nan    nan      nan
1   8.00        7     2.00    nan     nan   nan    nan    nan      nan
2   8.00        7     2.00    nan     nan   nan    nan    nan      nan
3   8.00        7     2.00    nan     nan   nan    nan    nan      nan
4   8.00        7     2.00    nan     nan   nan    nan    nan      nan

   bmonth  bday   byear  bprec  post  change   d4  sf  regtrans
0    1.00  1.00 1800.00   1.00  4.00   88.00 1.00 nan       nan
1     nan   nan     nan    nan   nan     nan  nan nan       nan
2     nan   nan     nan    nan   nan     nan  nan nan       nan
3     nan   nan     nan    nan   nan     nan  nan nan       nan
4     nan   nan     nan    nan   nan     nan  nan nan       nan
```

(b) What countries had the lowest polity score in 2017? Show the code you used to find this out (just scrolling and looking doesn't count!).

[3]:

```
data2017=data[data['year']==2017]  #show only scores in 2017
data2017.sort_values('polity2', ascending = True).head(1)
#sort data by scores and only show the lowest one
```

[3]:

```
         cyear  ccode scode  country  year  flag  fragment  democ  autoc  \
14091  6922017    692   BAH  Bahrain  2017     0      0.00      0     10

       polity  polity2  durable  xrreg  xrcomp  xropen  xconst  parreg  \
14091     -10   -10.00     5.00      3       1       1       1       4

       parcomp  exrec  exconst  polcomp  prior  emonth  eday  eyear  eprec  \
14091        1   1.00        1     1.00    nan     nan   nan    nan    nan

       interim  bmonth  bday  byear  bprec  post  change  d4  sf  regtrans
14091      nan     nan   nan    nan    nan   nan     nan nan nan       nan
```

Bahrain has the lowest polity score in 2017.

(c) The researcher is curious about how Polity IV scores over all countries may (or may not) have shifted between 2007 and 2017. Compare the mean, median, and modal polity scores for all countries in each of these two years. How have the scores changed (if at all) according to each statistic?

[4]:

```
data2017['polity2'].describe() #describe the data of 2017
```

[4]:

```
count   166.00
mean      4.13
```

```
std        6.16
min      -10.00
25%       -1.00
50%        6.00
75%        9.00
max       10.00
Name: polity2, dtype: float64
```

[5]:
```
data2007=data[data['year']==2007] #show only scores in 2007
data2007['polity2'].describe() #describe the data of 2007
```

[5]:
```
count   162.00
mean      3.69
std       6.44
min     -10.00
25%      -2.00
50%       6.00
75%       9.00
max      10.00
Name: polity2, dtype: float64
```

From 2007 to 2017, mean increases from 3.69 to 4.13, median does not change.The modes are the same.

(d) What is the interquartile range (IQR) of polity scores in 2017 and the IQR in 2007? What can be concluded about any changes between the two years according to this statistic?

[6]:
```
data2017['polity2'].quantile(0.75)-data2017['polity2'].quantile(0.25)
#calculate IQR of polity scores in 2017 by subtracting 25 percentile from 75␣
 ↪percentile
```

[6]:
```
10.0
```

[7]:
```
data2007['polity2'].quantile(0.75)-data2007['polity2'].quantile(0.25)
#calculate IQR of polity scores in 2007 by subtracting 25 percentile from 75␣
 ↪percentile
```

[7]:
```
11.0
```

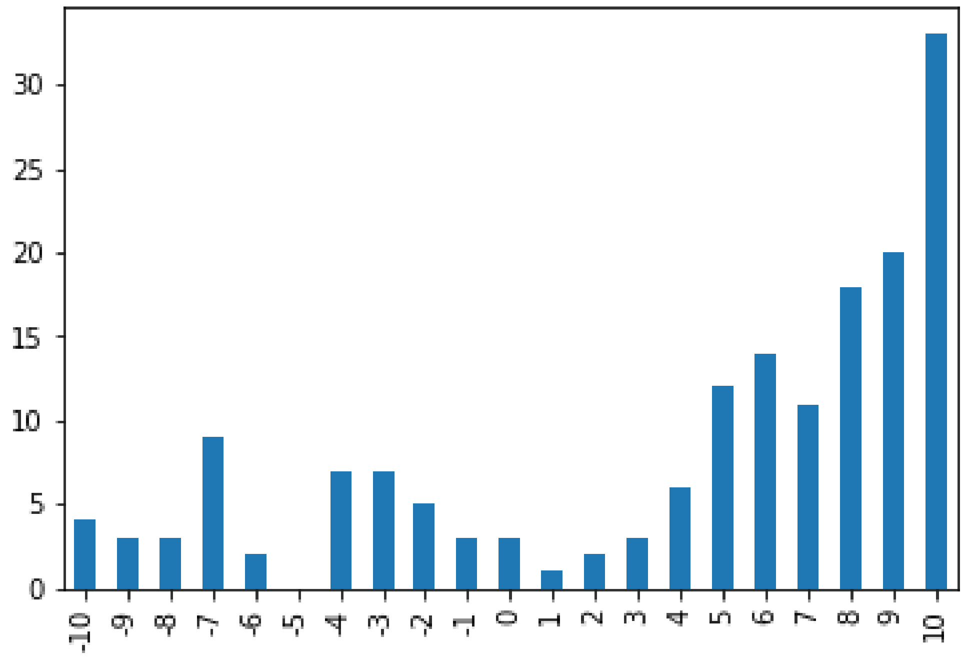
IQR decreases about 1.00 from 2007 to 2017. It shows that the scores more concentrated and have lessvariablity.

(e) Generate two bar charts of the polity scores, one for 2017 and one for 2007, where the x-axis for both goes from -10 to 10 (in that order from left to right). Does this visualization shed any further insight on how scores have changed over the last ten years? If so, what? If not, why not?

[8]:
```
order = [-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7,8,9,10]
#define x-axis to go from -10 to 10

data2017['polity2'].value_counts().loc[order].plot.bar(x =␣
 ↪[-10,-9,-8,-7,-6,-5,-4,
                                                                                        ␣
 ↪-3,-2,-1,0,1,2,3,4,5,
                                                                                         6,7,8,9,10])
#make up the bar chart for 2017
```

```
/share/apps/jupyterhub/2019-FA-DS-UA-111/lib/python3.7/site-
packages/ipykernel_launcher.py:4: FutureWarning:
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.

See the documentation here:
https://pandas.pydata.org/pandas-docs/stable/indexing.html#deprecate-loc-
reindex-listlike
  after removing the cwd from sys.path.
```

[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2af8c64339e8>
```

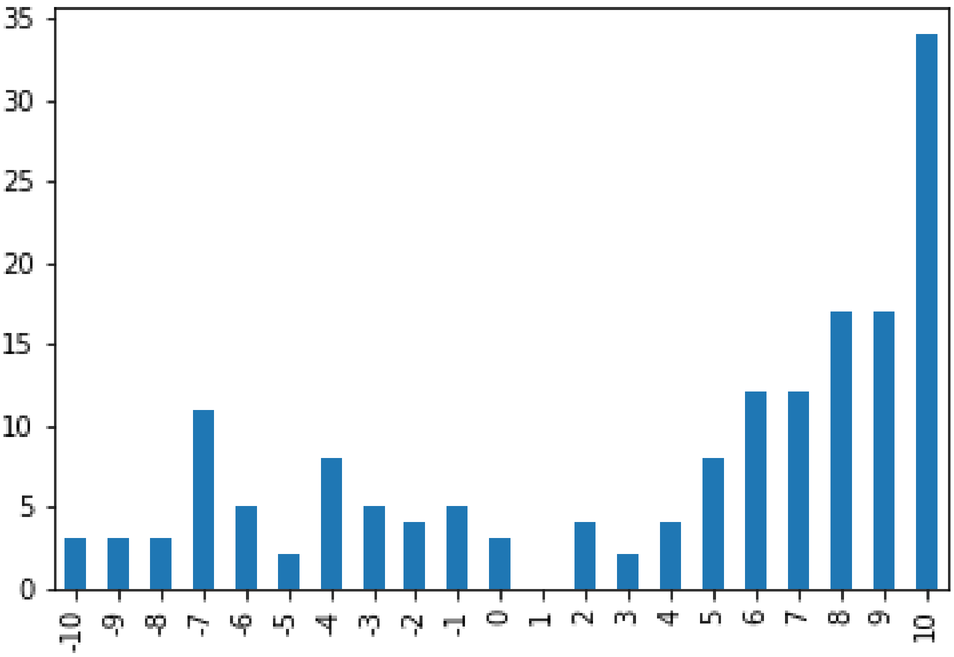


[9]:
```
order = [-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7,8,9,10]
#define x-axis to go from -10 to 10

data2007['polity2'].value_counts().loc[order].plot.bar(x =␣
 ↪[-10,-9,-8,-7,-6,-5,-4,
                                                                                        ␣
 ↪-3,-2,-1,0,1,2,3,4,5,
                                                                                         6,7,8,9,10])
```

```
#make up the bar chart for 2007
```

```
/share/apps/jupyterhub/2019-FA-DS-UA-111/lib/python3.7/site-
packages/ipykernel_launcher.py:4: FutureWarning:
Passing list-likes to .loc or [] with any missing label will raise
KeyError in the future, you can use .reindex() as an alternative.

See the documentation here:
https://pandas.pydata.org/pandas-docs/stable/indexing.html#deprecate-loc-
reindex-listlike
  after removing the cwd from sys.path.
```

```
[9]: <matplotlib.axes._subplots.AxesSubplot at 0x2af8c8e69550>
```

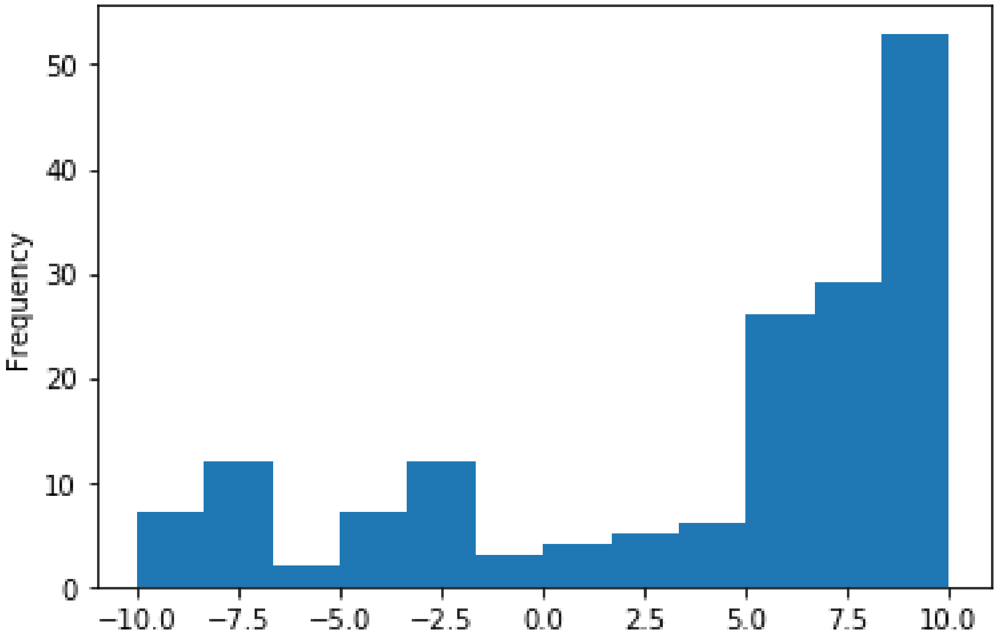


This visualization does not show the score change over ten years very clearly. It's because x-axias are both -10 to 10. In this way, the distributionsof the two seem to be similar.

(f) Do the same as the previous question, but using histograms instead of bar charts with the bins set to 12 for both. What additional insight, if any, can be gained from this?

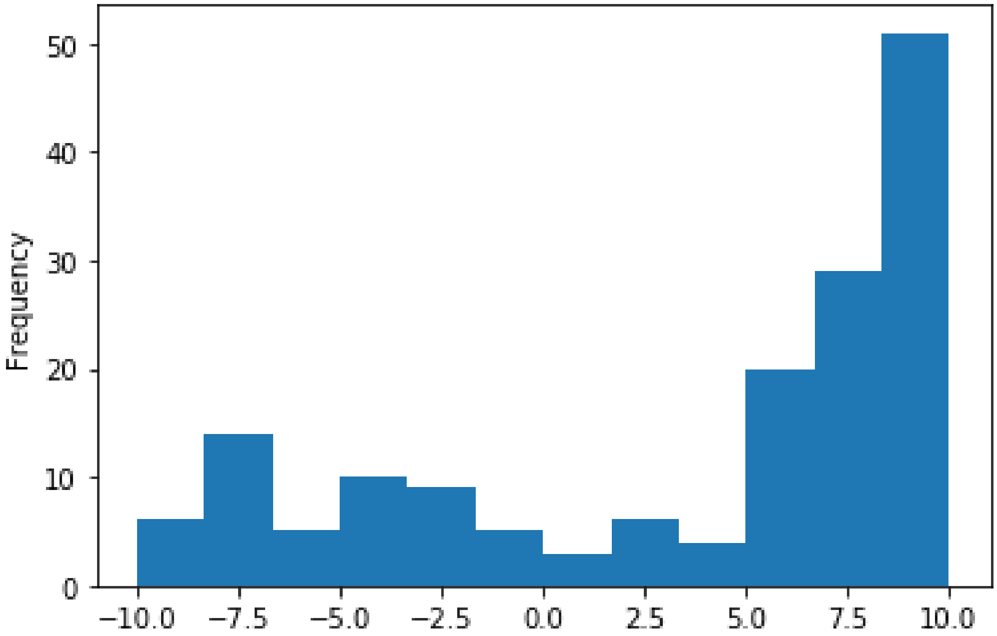
```
[10]: data2017['polity2'].plot.hist(bins = 12)
      #creat a histogram for data in 2017
```

```
[10]: <matplotlib.axes._subplots.AxesSubplot at 0x2af8c8e691d0>
```



[11]:
```
data2007['polity2'].plot.hist(bins = 12)
#creat a histogram for data in 2007
```

[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2af8c94099b0>
```

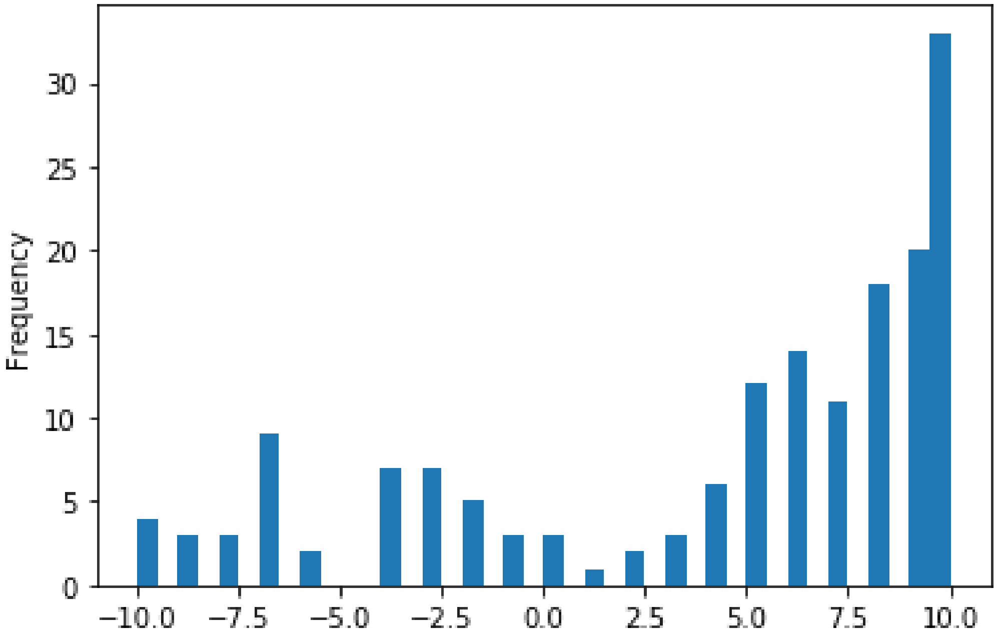


The distributions of two period have a similar pattern. The propotion of small values maintain a small proportion, and values larger than 5 increases and takes a large proportion.

(g) Now set the bins to 40 for each of the two years and show the new visualizations. Do you think this improves the visualizations? Why or why not?

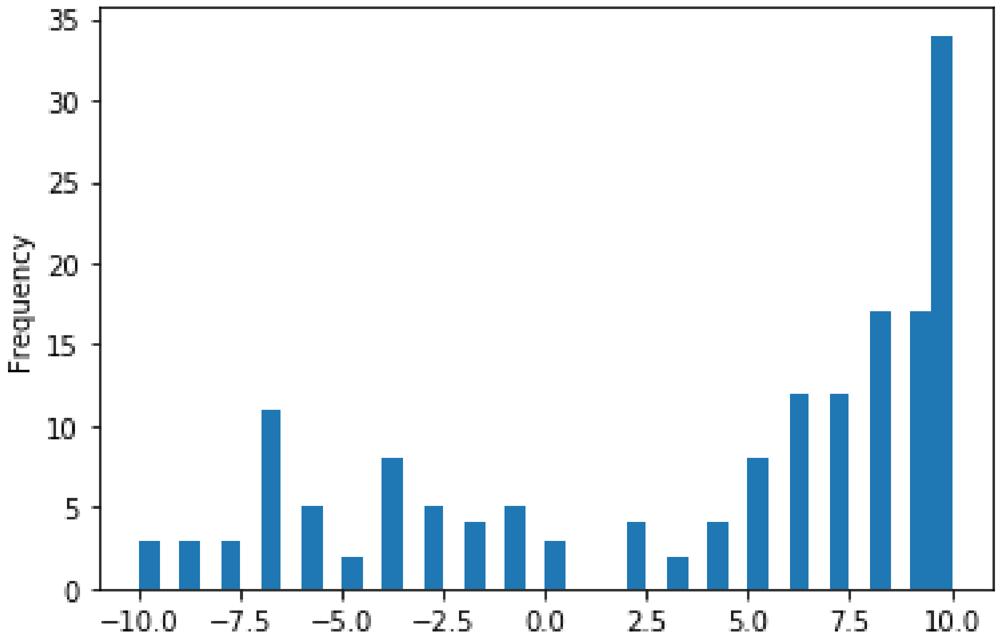
[12]:
```
data2017['polity2'].plot.hist(bins = 40)
#creat a histogram for data in 2017
```

[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2af8c9480eb8>
```



[13]:
```
data2007['polity2'].plot.hist(bins = 40)
#creat a histogram for data in 2007
```

[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2af8c94fa208>
```



Yes. Because using a larger bin shows the detials better. We can see clearyly which part of the distribution changes.

(h) Having explored patterns in 2007 and 2017 from a number of perspectives, what are two overall conclusions the researcher can draw about the distribution of regime types between these ten years?

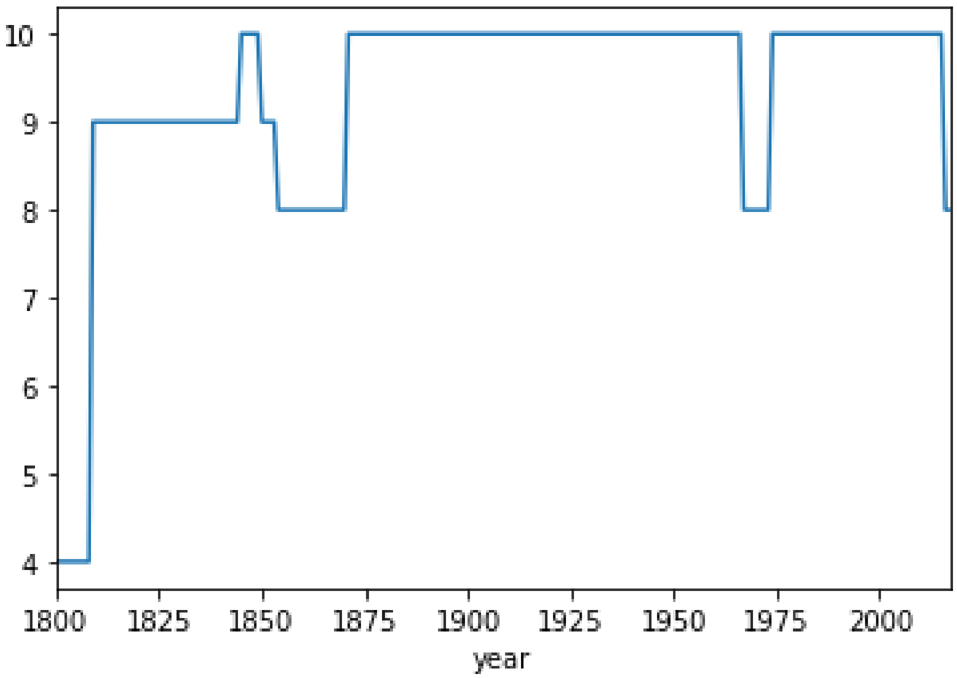
1. The distribution shape does not vary too much from 2007 to 2017. It just change slightly and becomes a little more concentrated to those with larger polity scores. Basically, the regime scores are tend to concentrate at higher levels.

2. The mean of polity scores increased from 2007 to 2017. The proportion of high policy scores has increased, so they pull the mean larger.

(i) The researcher is now curious about how the regime type of the US compares with that of India over time. Generate two line graphs, one for the US and one for India, of their regime types over all years available for each, with the years clearly labelled on the x-axes.

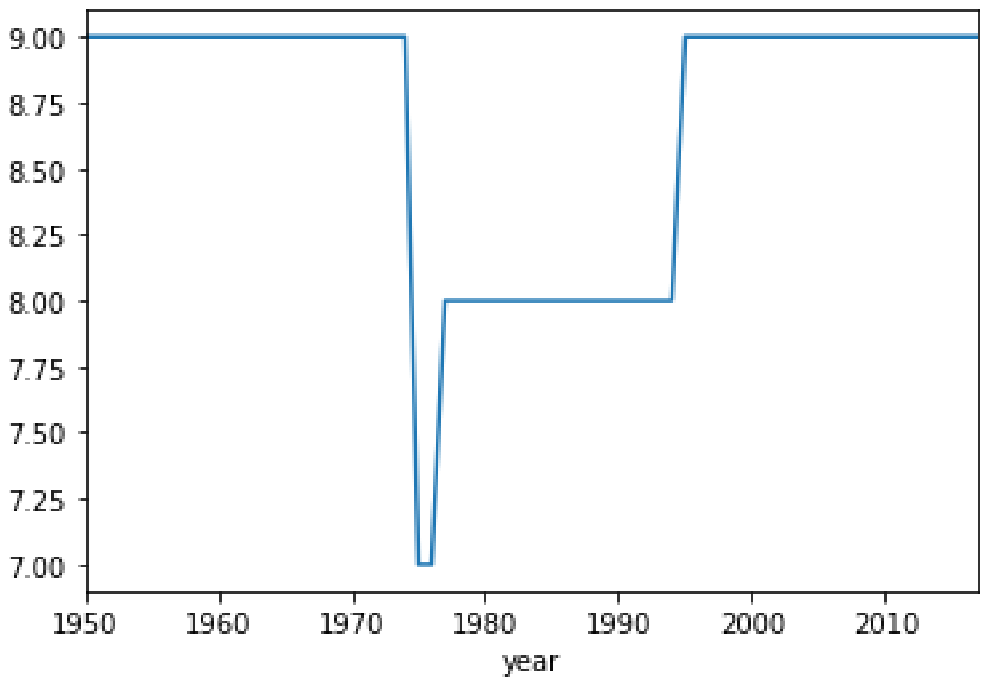
[14]:
```
dataus=data[data['country']=='United States']
#site the datas of United States only

dataus.set_index('year')['polity2'].plot.line()
#make a plot between year and polity2
```

[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2af8c9632a90>
```



[15]:
```
dataus=data[data['country']=='India']
#site the datas of India only

dataus.set_index('year')['polity2'].plot.line()
#make a plot between year and polity2
```

[15]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2af8c96a8ac8>
```



(j) What are at least two findings that can be inferred from these visualizations?

1. We have more data about United States than India before 1950. The United States experienced a blust in the number of regime types in 1800, and a flutuation in the next 75 years. It shows that United States has become a democratic country since then.
2. Both India and United States experienced a decrease in regime types in around 1975.It takes India more time to get back to normal level of regime types than United States. World economic depression influenced both country's policy in around 1975.

(k) Suppose the researcher wanted to visualize the association between regime type in the US and India. What kind of graph would be the most effective choice to visualize this? Scatter plot would be the most effective choice to visualize this.

(l) What is the correlation between regime type and durability for all countries and years in the dataset?

[16]:
```
data.corr() #show all the correlation between each two variables
```

[16]:
```
          cyear  ccode  year  flag  fragment  democ  autoc  polity  polity2  \
cyear      1.00   1.00  0.21  0.03      0.01  -0.05   0.03   -0.09    -0.18
ccode      1.00   1.00  0.21  0.03      0.01  -0.05   0.03   -0.09    -0.18
year       0.21   0.21  1.00  0.11     -0.03   0.04  -0.09    0.11     0.33
flag       0.03   0.03  0.11  1.00      0.03   0.01  -0.01    0.02     0.05
fragment   0.01   0.01 -0.03  0.03      1.00  -0.35  -0.35   -0.34    -0.04
democ     -0.05  -0.05  0.04  0.01     -0.35   1.00   0.92    0.98     0.24
```

```
autoc       0.03   0.03 -0.09 -0.01      -0.35   0.92   1.00    0.82    -0.20
polity     -0.09  -0.09  0.11  0.02      -0.34   0.98   0.82    1.00     0.44
polity2    -0.18  -0.18  0.33  0.05      -0.04   0.24  -0.20    0.44     1.00
durable    -0.08  -0.08  0.08 -0.04      -0.09   0.23   0.16    0.24     0.19
xrreg      -0.01  -0.01 -0.02  0.00      -0.36   0.98   0.98    0.92     0.02
xrcomp     -0.02  -0.02 -0.00  0.00      -0.36   0.99   0.97    0.94     0.07
xropen     -0.03  -0.03  0.01  0.00      -0.36   0.98   0.97    0.93     0.06
xconst     -0.03  -0.03  0.02  0.01      -0.36   0.99   0.95    0.96     0.15
parreg     -0.02  -0.02 -0.02 -0.00      -0.36   0.98   0.98    0.92     0.01
parcomp    -0.03  -0.03 -0.00  0.00      -0.36   0.99   0.96    0.94     0.09
exrec      -0.04  -0.04  0.05  0.01      -0.36   0.99   0.94    0.96     0.16
exconst    -0.03  -0.03  0.02  0.01      -0.36   0.99   0.95    0.96     0.15
polcomp    -0.05  -0.05  0.02  0.01      -0.36   0.99   0.93    0.97     0.19
prior       0.01   0.01  0.16  0.08      -0.17   0.06   0.00    0.08     0.30
emonth      0.02   0.02 -0.00 -0.03       0.09  -0.04  -0.04   -0.03     0.01
eday       -0.05  -0.05 -0.13  0.01      -0.09  -0.04  -0.04   -0.03     0.02
eyear       0.37   0.37  1.00  0.20      -0.12   0.10   0.06    0.12     0.25
eprec      -0.02  -0.02 -0.05 -0.02       0.05   0.14   0.13    0.14     0.01
interim     0.10   0.10  0.21  0.03      -0.24   0.57   0.58    0.58     0.04
bmonth      0.05   0.05  0.13  0.02       0.04  -0.01  -0.02   -0.00     0.01
bday        0.04   0.04  0.20  0.07      -0.05   0.01  -0.01    0.02     0.03
byear       0.21   0.21  0.69  0.06      -0.17   0.01  -0.09    0.07     0.17
bprec      -0.02  -0.02  0.02 -0.07      -0.03   0.03   0.01    0.03     0.03
post       -0.09  -0.09  0.13  0.06      -0.15   0.88   0.42    0.94     0.65
change     -0.03  -0.03 -0.15 -0.01      -0.47   0.52   0.52    0.51    -0.01
d4           nan    nan   nan   nan        nan    nan    nan     nan      nan
sf           nan    nan   nan   nan        nan    nan    nan     nan      nan
regtrans   -0.02  -0.02 -0.10 -0.01      -0.46   0.51   0.51    0.50    -0.03

           durable  xrreg  xrcomp  xropen  xconst  parreg  parcomp  exrec  \
cyear        -0.08  -0.01   -0.02   -0.03   -0.03   -0.02    -0.03  -0.04
ccode        -0.08  -0.01   -0.02   -0.03   -0.03   -0.02    -0.03  -0.04
year          0.08  -0.02   -0.00    0.01    0.02   -0.02    -0.00   0.05
flag         -0.04   0.00    0.00    0.00    0.01   -0.00     0.00   0.01
fragment     -0.09  -0.36   -0.36   -0.36   -0.36   -0.36    -0.36  -0.36
democ         0.23   0.98    0.99    0.98    0.99    0.98     0.99   0.99
autoc         0.16   0.98    0.97    0.97    0.95    0.98     0.96   0.94
polity        0.24   0.92    0.94    0.93    0.96    0.92     0.94   0.96
polity2       0.19   0.02    0.07    0.06    0.15    0.01     0.09   0.16
durable       1.00   0.19    0.19    0.19    0.21    0.20     0.20   0.20
xrreg         0.19   1.00    1.00    1.00    0.99    1.00     1.00   0.99
xrcomp        0.19   1.00    1.00    1.00    1.00    1.00     1.00   0.99
xropen        0.19   1.00    1.00    1.00    0.99    0.99     1.00   0.99
xconst        0.21   0.99    1.00    0.99    1.00    0.99     1.00   0.99
parreg        0.20   1.00    1.00    0.99    0.99    1.00     1.00   0.99
parcomp       0.20   1.00    1.00    1.00    1.00    1.00     1.00   0.99
exrec         0.20   0.99    0.99    0.99    0.99    0.99     0.99   1.00
```

```
exconst     0.21   0.99    1.00    0.99    1.00    0.99     1.00   0.99
polcomp     0.22   0.98    0.99    0.98    0.99    0.98     0.99   0.99
prior       0.03   0.03    0.03    0.03    0.04    0.03     0.03   0.05
emonth     -0.00  -0.04   -0.04   -0.04   -0.04   -0.04    -0.04  -0.03
eday        0.00  -0.04   -0.04   -0.04   -0.04   -0.04    -0.04  -0.04
eyear       0.04   0.08    0.09    0.09    0.09    0.08     0.09   0.11
eprec       0.03   0.14    0.14    0.14    0.14    0.14     0.14   0.14
interim     0.26   0.58    0.58    0.58    0.58    0.58     0.58   0.58
bmonth      0.04  -0.02   -0.01   -0.01   -0.01   -0.01    -0.02  -0.00
bday       -0.03  -0.00    0.00    0.00    0.00   -0.00     0.00   0.02
byear       0.05  -0.03   -0.02   -0.01   -0.00   -0.04    -0.02   0.03
bprec       0.04   0.02    0.02    0.02    0.03    0.02     0.03   0.02
post        0.05   0.72    0.76    0.74    0.82    0.67     0.77   0.83
change      0.08   0.52    0.52    0.52    0.52    0.52     0.52   0.52
d4           nan    nan     nan     nan     nan     nan      nan    nan
sf           nan    nan     nan     nan     nan     nan      nan    nan
regtrans    0.08   0.51    0.51    0.51    0.51    0.51     0.51   0.51

          exconst  polcomp  prior  emonth  eday  eyear  eprec  interim  \
cyear       -0.03    -0.05   0.01    0.02 -0.05   0.37  -0.02     0.10
ccode       -0.03    -0.05   0.01    0.02 -0.05   0.37  -0.02     0.10
year         0.02     0.02   0.16   -0.00 -0.13   1.00  -0.05     0.21
flag         0.01     0.01   0.08   -0.03  0.01   0.20  -0.02     0.03
fragment    -0.36    -0.36  -0.17    0.09 -0.09  -0.12   0.05    -0.24
democ        0.99     0.99   0.06   -0.04 -0.04   0.10   0.14     0.57
autoc        0.95     0.93   0.00   -0.04 -0.04   0.06   0.13     0.58
polity       0.96     0.97   0.08   -0.03 -0.03   0.12   0.14     0.58
polity2      0.15     0.19   0.30    0.01  0.02   0.25   0.01     0.04
durable      0.21     0.22   0.03   -0.00  0.00   0.04   0.03     0.26
xrreg        0.99     0.98   0.03   -0.04 -0.04   0.08   0.14     0.58
xrcomp       1.00     0.99   0.03   -0.04 -0.04   0.09   0.14     0.58
xropen       0.99     0.98   0.03   -0.04 -0.04   0.09   0.14     0.58
xconst       1.00     0.99   0.04   -0.04 -0.04   0.09   0.14     0.58
parreg       0.99     0.98   0.03   -0.04 -0.04   0.08   0.14     0.58
parcomp      1.00     0.99   0.03   -0.04 -0.04   0.09   0.14     0.58
exrec        0.99     0.99   0.05   -0.03 -0.04   0.11   0.14     0.58
exconst      1.00     0.99   0.04   -0.04 -0.04   0.09   0.14     0.58
polcomp      0.99     1.00   0.05   -0.04 -0.04   0.10   0.14     0.58
prior        0.04     0.05   1.00    0.04 -0.02   0.17   0.02     0.02
emonth      -0.04    -0.04   0.04    1.00  0.01  -0.00  -0.04    -0.02
eday        -0.04    -0.04  -0.02    0.01  1.00  -0.13   0.08    -0.11
eyear        0.09     0.10   0.17   -0.00 -0.13   1.00   0.00     0.24
eprec        0.14     0.14   0.02   -0.04  0.08   0.00   1.00     0.18
interim      0.58     0.58   0.02   -0.02 -0.11   0.24   0.18     1.00
bmonth      -0.01    -0.02  -0.00    0.65  0.00   0.02   0.01     0.03
bday         0.00     0.01  -0.01   -0.00  0.29   0.18  -0.17     0.02
byear       -0.00    -0.00   0.10   -0.03 -0.07   0.62   0.01     0.14
```

```
bprec        0.03     0.03  -0.01    0.01  0.12  -0.02   0.88     0.13
post         0.82     0.85   0.21    0.03  0.01   0.10   0.05    -0.05
change       0.52     0.52   0.08   -0.02  0.03  -0.00   0.16     0.11
d4            nan      nan    nan     nan   nan    nan    nan      nan
sf            nan      nan    nan     nan   nan    nan    nan      nan
regtrans     0.51     0.51   0.16   -0.01  0.03  -0.01   0.16     0.13

          bmonth  bday  byear  bprec  post  change  d4  sf  regtrans
cyear       0.05  0.04   0.21  -0.02 -0.09   -0.03 nan nan     -0.02
ccode       0.05  0.04   0.21  -0.02 -0.09   -0.03 nan nan     -0.02
year        0.13  0.20   0.69   0.02  0.13   -0.15 nan nan     -0.10
flag        0.02  0.07   0.06  -0.07  0.06   -0.01 nan nan     -0.01
fragment    0.04 -0.05  -0.17  -0.03 -0.15   -0.47 nan nan     -0.46
democ      -0.01  0.01   0.01   0.03  0.88    0.52 nan nan      0.51
autoc      -0.02 -0.01  -0.09   0.01  0.42    0.52 nan nan      0.51
polity     -0.00  0.02   0.07   0.03  0.94    0.51 nan nan      0.50
polity2     0.01  0.03   0.17   0.03  0.65   -0.01 nan nan     -0.03
durable     0.04 -0.03   0.05   0.04  0.05    0.08 nan nan      0.08
xrreg      -0.02 -0.00  -0.03   0.02  0.72    0.52 nan nan      0.51
xrcomp     -0.01  0.00  -0.02   0.02  0.76    0.52 nan nan      0.51
xropen     -0.01  0.00  -0.01   0.02  0.74    0.52 nan nan      0.51
xconst     -0.01  0.00  -0.00   0.03  0.82    0.52 nan nan      0.51
parreg     -0.01 -0.00  -0.04   0.02  0.67    0.52 nan nan      0.51
parcomp    -0.02  0.00  -0.02   0.03  0.77    0.52 nan nan      0.51
exrec      -0.00  0.02   0.03   0.02  0.83    0.52 nan nan      0.51
exconst    -0.01  0.00  -0.00   0.03  0.82    0.52 nan nan      0.51
polcomp    -0.02  0.01  -0.00   0.03  0.85    0.52 nan nan      0.51
prior      -0.00 -0.01   0.10  -0.01  0.21    0.08 nan nan      0.16
emonth      0.65 -0.00  -0.03   0.01  0.03   -0.02 nan nan     -0.01
eday        0.00  0.29  -0.07   0.12  0.01    0.03 nan nan      0.03
eyear       0.02  0.18   0.62  -0.02  0.10   -0.00 nan nan     -0.01
eprec       0.01 -0.17   0.01   0.88  0.05    0.16 nan nan      0.16
interim     0.03  0.02   0.14   0.13 -0.05    0.11 nan nan      0.13
bmonth      1.00  0.06   0.13   0.05 -0.00   -0.08 nan nan     -0.02
bday        0.06  1.00   0.16  -0.16  0.01   -0.06 nan nan     -0.02
byear       0.13  0.16   1.00   0.02  0.08   -0.17 nan nan     -0.11
bprec       0.05 -0.16   0.02   1.00  0.02   -0.15 nan nan     -0.14
post       -0.00  0.01   0.08   0.02  1.00    0.06 nan nan      0.07
change     -0.08 -0.06  -0.17  -0.15  0.06    1.00 nan nan      0.99
d4           nan   nan    nan    nan   nan     nan nan nan       nan
sf           nan   nan    nan    nan   nan     nan nan nan       nan
regtrans   -0.02 -0.02  -0.11  -0.14  0.07    0.99 nan nan      1.00
```

The correlation between regime type and durability for all countries and years in the dataset is basically 0.19.

(m) What is the correlation between regime type and durability in 2017 and in 2007? Briefly interpret each. What do you make of the difference, if any, between the two?

[17]:
```
data2017.corr() #show all the correlation between each two variables in 2017
```

[17]:
```
             cyear  ccode  year  flag  fragment  democ  autoc  polity  polity2  \
cyear         1.00   1.00   nan  0.01      0.01  -0.17   0.02   -0.22    -0.39
ccode         1.00   1.00   nan  0.01      0.01  -0.17   0.02   -0.22    -0.39
year           nan    nan   nan   nan       nan    nan    nan     nan      nan
flag          0.01   0.01   nan  1.00      0.09   0.05   0.05    0.07    -0.00
fragment      0.01   0.01   nan  0.09      1.00  -0.12  -0.10   -0.12    -0.05
democ        -0.17  -0.17   nan  0.05     -0.12   1.00   0.88    0.98     0.39
autoc         0.02   0.02   nan  0.05     -0.10   0.88   1.00    0.77    -0.15
polity       -0.22  -0.22   nan  0.07     -0.12   0.98   0.77    1.00     0.57
polity2      -0.39  -0.39   nan -0.00     -0.05   0.39  -0.15    0.57     1.00
durable      -0.20  -0.20   nan -0.30     -0.07   0.21   0.15    0.20     0.15
xrreg        -0.07  -0.07   nan  0.06     -0.12   0.97   0.97    0.90     0.12
xrcomp       -0.09  -0.09   nan  0.07     -0.12   0.98   0.96    0.92     0.17
xropen       -0.08  -0.08   nan  0.08     -0.11   0.97   0.96    0.91     0.15
xconst       -0.11  -0.11   nan  0.08     -0.11   0.99   0.93    0.95     0.26
parreg       -0.06  -0.06   nan  0.06     -0.12   0.96   0.97    0.89     0.09
parcomp      -0.10  -0.10   nan  0.07     -0.13   0.98   0.95    0.93     0.19
exrec        -0.12  -0.12   nan  0.08     -0.12   0.99   0.92    0.95     0.27
exconst      -0.11  -0.11   nan  0.08     -0.11   0.99   0.93    0.95     0.26
polcomp      -0.15  -0.15   nan  0.06     -0.13   0.99   0.91    0.96     0.32
prior        -0.20  -0.20   nan  0.18       nan   0.41   0.06    0.24     0.24
emonth        0.72   0.72   nan -0.44       nan  -0.32   0.49   -0.40    -0.40
eday          0.13   0.13   nan -0.68       nan   0.38  -0.54    0.45     0.45
eyear         0.47   0.47   nan -1.00       nan  -0.29   0.18   -0.25    -0.25
eprec        -0.20  -0.20   nan  0.76       nan   1.00   1.00    0.99     0.21
interim       1.00   1.00   nan -1.00       nan  -1.00  -1.00   -1.00    -1.00
bmonth        0.70   0.70   nan -0.37       nan  -0.31   0.49   -0.39    -0.39
bday          0.25   0.25   nan -0.84       nan   0.19  -0.35    0.26     0.26
byear          nan    nan   nan   nan       nan    nan    nan     nan      nan
bprec          nan    nan   nan   nan       nan    nan    nan     nan      nan
post         -0.12  -0.12   nan  0.25       nan   0.98  -0.95    1.00     1.00
change       -0.41  -0.41   nan  0.87       nan   0.99   0.99    1.00     0.38
d4             nan    nan   nan   nan       nan    nan    nan     nan      nan
sf             nan    nan   nan   nan       nan    nan    nan     nan      nan
regtrans     -0.42  -0.42   nan  0.88       nan   1.00   1.00    1.00     0.33

             durable  xrreg  xrcomp  xropen  xconst  parreg  parcomp  exrec  \
cyear          -0.20  -0.07   -0.09   -0.08   -0.11   -0.06    -0.10  -0.12
ccode          -0.20  -0.07   -0.09   -0.08   -0.11   -0.06    -0.10  -0.12
year             nan    nan     nan     nan     nan     nan      nan    nan
flag           -0.30   0.06    0.07    0.08    0.08    0.06     0.07   0.08
fragment       -0.07  -0.12   -0.12   -0.11   -0.11   -0.12    -0.13  -0.12
democ           0.21   0.97    0.98    0.97    0.99    0.96     0.98   0.99
autoc           0.15   0.97    0.96    0.96    0.93    0.97     0.95   0.92
polity          0.20   0.90    0.92    0.91    0.95    0.89     0.93   0.95
```

```
polity2      0.15   0.12    0.17    0.15    0.26    0.09     0.19   0.27
durable      1.00   0.17    0.17    0.16    0.18    0.19     0.17   0.16
xrreg        0.17   1.00    1.00    1.00    0.99    1.00     1.00   0.99
xrcomp       0.17   1.00    1.00    1.00    1.00    0.99     1.00   0.99
xropen       0.16   1.00    1.00    1.00    0.99    0.99     0.99   0.99
xconst       0.18   0.99    1.00    0.99    1.00    0.98     0.99   1.00
parreg       0.19   1.00    0.99    0.99    0.98    1.00     0.99   0.98
parcomp      0.17   1.00    1.00    0.99    0.99    0.99     1.00   0.99
exrec        0.16   0.99    0.99    0.99    1.00    0.98     0.99   1.00
exconst      0.18   0.99    1.00    0.99    1.00    0.98     0.99   1.00
polcomp      0.17   0.98    0.99    0.98    0.99    0.97     0.99   0.99
prior        0.48   0.35    0.24    0.15    0.25    0.40     0.93   0.11
emonth      -0.35   0.00   -0.31   -0.53   -0.24   -0.08    -0.03  -0.42
eday         0.59   0.39    0.47    0.33    0.47   -0.57    -0.31   0.46
eyear        0.24   0.00   -0.16   -0.20   -0.14   -0.32    -0.43  -0.24
eprec        0.20   1.00    1.00    1.00    1.00    1.00     1.00   1.00
interim       nan  -1.00   -1.00   -1.00   -1.00   -1.00    -1.00  -1.00
bmonth      -0.38   0.00   -0.30   -0.53   -0.23   -0.05     0.01  -0.41
bday         0.52   0.29    0.30    0.18    0.31   -0.53    -0.38   0.27
byear         nan    nan     nan     nan     nan     nan      nan    nan
bprec         nan    nan     nan     nan     nan     nan      nan    nan
post         0.35   0.78    0.96    0.66    0.97   -0.56     0.19   0.97
change       0.28   0.99    0.99    0.99    0.99    0.99     0.99   1.00
d4            nan    nan     nan     nan     nan     nan      nan    nan
sf            nan    nan     nan     nan     nan     nan      nan    nan
regtrans     0.30   1.00    1.00    1.00    1.00    1.00     1.00   1.00

          exconst  polcomp  prior  emonth   eday  eyear  eprec  interim  \
cyear       -0.11    -0.15  -0.20    0.72   0.13   0.47  -0.20     1.00
ccode       -0.11    -0.15  -0.20    0.72   0.13   0.47  -0.20     1.00
year          nan      nan    nan     nan    nan    nan    nan      nan
flag         0.08     0.06   0.18   -0.44  -0.68  -1.00   0.76    -1.00
fragment    -0.11    -0.13    nan     nan    nan    nan    nan      nan
democ        0.99     0.99   0.41   -0.32   0.38  -0.29   1.00    -1.00
autoc        0.93     0.91   0.06    0.49  -0.54   0.18   1.00    -1.00
polity       0.95     0.96   0.24   -0.40   0.45  -0.25   0.99    -1.00
polity2      0.26     0.32   0.24   -0.40   0.45  -0.25   0.21    -1.00
durable      0.18     0.17   0.48   -0.35   0.59   0.24   0.20      nan
xrreg        0.99     0.98   0.35    0.00   0.39   0.00   1.00    -1.00
xrcomp       1.00     0.99   0.24   -0.31   0.47  -0.16   1.00    -1.00
xropen       0.99     0.98   0.15   -0.53   0.33  -0.20   1.00    -1.00
xconst       1.00     0.99   0.25   -0.24   0.47  -0.14   1.00    -1.00
parreg       0.98     0.97   0.40   -0.08  -0.57  -0.32   1.00    -1.00
parcomp      0.99     0.99   0.93   -0.03  -0.31  -0.43   1.00    -1.00
exrec        1.00     0.99   0.11   -0.42   0.46  -0.24   1.00    -1.00
exconst      1.00     0.99   0.25   -0.24   0.47  -0.14   1.00    -1.00
polcomp      0.99     1.00   0.61   -0.50   0.11  -0.46   1.00    -1.00
```

```
prior        0.25     0.61   1.00    0.25 -0.15  -0.18    nan      nan
emonth      -0.24    -0.50   0.25    1.00 -0.19   0.44    nan      nan
eday         0.47     0.11  -0.15   -0.19  1.00   0.68    nan      nan
eyear       -0.14    -0.46  -0.18    0.44  0.68   1.00    nan      nan
eprec        1.00     1.00    nan     nan   nan    nan   1.00      nan
interim     -1.00    -1.00    nan     nan   nan    nan    nan     1.00
bmonth      -0.23    -0.48   0.27    1.00 -0.25   0.37    nan      nan
bday         0.31    -0.07  -0.17    0.00  0.97   0.84    nan      nan
byear         nan      nan    nan     nan   nan    nan    nan      nan
bprec         nan      nan    nan     nan   nan    nan    nan      nan
post         0.97     0.71   0.24   -0.40  0.45  -0.25    nan      nan
change       0.99     0.99  -0.58   -0.52  0.49  -0.07   0.98    -1.00
d4            nan      nan    nan     nan   nan    nan    nan      nan
sf            nan      nan    nan     nan   nan    nan    nan      nan
regtrans     1.00     1.00  -0.57   -0.64  0.56  -0.04   1.00    -1.00

          bmonth  bday  byear  bprec  post  change  d4  sf  regtrans
cyear       0.70  0.25    nan    nan -0.12   -0.41 nan nan     -0.42
ccode       0.70  0.25    nan    nan -0.12   -0.41 nan nan     -0.42
year         nan   nan    nan    nan   nan     nan nan nan       nan
flag       -0.37 -0.84    nan    nan  0.25    0.87 nan nan      0.88
fragment     nan   nan    nan    nan   nan     nan nan nan       nan
democ      -0.31  0.19    nan    nan  0.98    0.99 nan nan      1.00
autoc       0.49 -0.35    nan    nan -0.95    0.99 nan nan      1.00
polity     -0.39  0.26    nan    nan  1.00    1.00 nan nan      1.00
polity2    -0.39  0.26    nan    nan  1.00    0.38 nan nan      0.33
durable    -0.38  0.52    nan    nan  0.35    0.28 nan nan      0.30
xrreg       0.00  0.29    nan    nan  0.78    0.99 nan nan      1.00
xrcomp     -0.30  0.30    nan    nan  0.96    0.99 nan nan      1.00
xropen     -0.53  0.18    nan    nan  0.66    0.99 nan nan      1.00
xconst     -0.23  0.31    nan    nan  0.97    0.99 nan nan      1.00
parreg     -0.05 -0.53    nan    nan -0.56    0.99 nan nan      1.00
parcomp     0.01 -0.38    nan    nan  0.19    0.99 nan nan      1.00
exrec      -0.41  0.27    nan    nan  0.97    1.00 nan nan      1.00
exconst    -0.23  0.31    nan    nan  0.97    0.99 nan nan      1.00
polcomp    -0.48 -0.07    nan    nan  0.71    0.99 nan nan      1.00
prior       0.27 -0.17    nan    nan  0.24   -0.58 nan nan     -0.57
emonth      1.00  0.00    nan    nan -0.40   -0.52 nan nan     -0.64
eday       -0.25  0.97    nan    nan  0.45    0.49 nan nan      0.56
eyear       0.37  0.84    nan    nan -0.25   -0.07 nan nan     -0.04
eprec        nan   nan    nan    nan   nan    0.98 nan nan      1.00
interim      nan   nan    nan    nan   nan   -1.00 nan nan     -1.00
bmonth      1.00 -0.07    nan    nan -0.39   -0.54 nan nan     -0.65
bday       -0.07  1.00    nan    nan  0.26    0.35 nan nan      0.40
byear        nan   nan    nan    nan   nan     nan nan nan       nan
bprec        nan   nan    nan    nan   nan     nan nan nan       nan
post       -0.39  0.26    nan    nan  1.00    0.65 nan nan      0.64
```

```
change     -0.54  0.35    nan    nan  0.65    1.00 nan nan       1.00
d4           nan   nan    nan    nan   nan     nan nan nan        nan
sf           nan   nan    nan    nan   nan     nan nan nan        nan
regtrans   -0.65  0.40    nan    nan  0.64    1.00 nan nan       1.00
```

[18]:
```
data2007.corr() #show all the correlation between each two variables in 2007
```

[18]:
```
          cyear  ccode  year  flag  fragment  democ  autoc  polity  polity2  \
cyear      1.00   1.00   nan   nan      0.03  -0.17   0.05   -0.24    -0.46
ccode      1.00   1.00   nan   nan      0.03  -0.17   0.05   -0.24    -0.46
year        nan    nan   nan   nan       nan    nan    nan     nan      nan
flag        nan    nan   nan   nan       nan    nan    nan     nan      nan
fragment   0.03   0.03   nan   nan      1.00  -0.44  -0.47   -0.41     0.03
democ     -0.17  -0.17   nan   nan     -0.44   1.00   0.89    0.98     0.42
autoc      0.05   0.05   nan   nan     -0.47   0.89   1.00    0.79    -0.21
polity    -0.24  -0.24   nan   nan     -0.41   0.98   0.79    1.00     0.61
polity2   -0.46  -0.46   nan   nan      0.03   0.42  -0.21    0.61     1.00
durable   -0.15  -0.15   nan   nan     -0.09   0.21   0.15    0.20     0.15
xrreg     -0.06  -0.06   nan   nan     -0.47   0.97   0.97    0.91     0.10
xrcomp    -0.08  -0.08   nan   nan     -0.47   0.98   0.96    0.93     0.16
xropen    -0.07  -0.07   nan   nan     -0.46   0.97   0.96    0.92     0.14
xconst    -0.10  -0.10   nan   nan     -0.46   0.99   0.94    0.95     0.26
parreg    -0.05  -0.05   nan   nan     -0.47   0.96   0.98    0.90     0.07
parcomp   -0.09  -0.09   nan   nan     -0.47   0.98   0.96    0.93     0.18
exrec     -0.12  -0.12   nan   nan     -0.46   0.99   0.93    0.96     0.27
exconst   -0.10  -0.10   nan   nan     -0.46   0.99   0.94    0.95     0.26
polcomp   -0.13  -0.13   nan   nan     -0.46   0.99   0.92    0.97     0.33
prior     -0.48  -0.48   nan   nan     -0.06   0.52  -0.17    0.41     0.41
emonth     0.03   0.03   nan   nan      0.09   0.24   0.22    0.26     0.14
eday      -0.11  -0.11   nan   nan     -0.16   0.15   0.15    0.14    -0.01
eyear     -0.15  -0.15   nan   nan      0.13  -0.06  -0.09   -0.06     0.13
eprec     -0.19  -0.19   nan   nan     -0.24  -0.36  -0.38   -0.36     0.03
interim    0.23   0.23   nan   nan     -0.70   1.00   0.99    1.00     0.13
bmonth     0.07   0.07   nan   nan      0.40  -0.10  -0.03   -0.05    -0.05
bday       0.16   0.16   nan   nan     -0.06  -0.27   0.02   -0.18    -0.18
byear       nan    nan   nan   nan       nan    nan    nan     nan      nan
bprec     -0.10  -0.10   nan   nan     -0.27   0.13   0.04    0.07     0.07
post      -0.82  -0.82   nan   nan      0.09   0.95  -0.90    1.00     1.00
change    -0.08  -0.08   nan   nan     -0.72   0.99   0.98    0.99     0.33
d4          nan    nan   nan   nan       nan    nan    nan     nan      nan
sf          nan    nan   nan   nan       nan    nan    nan     nan      nan
regtrans  -0.07  -0.07   nan   nan     -0.73   0.99   0.99    0.99     0.25

          durable  xrreg  xrcomp  xropen  xconst  parreg  parcomp  exrec  \
cyear       -0.15  -0.06   -0.08   -0.07   -0.10   -0.05    -0.09  -0.12
ccode       -0.15  -0.06   -0.08   -0.07   -0.10   -0.05    -0.09  -0.12
year          nan    nan     nan     nan     nan     nan      nan    nan
flag          nan    nan     nan     nan     nan     nan      nan    nan
```

```
fragment    -0.09  -0.47   -0.47   -0.46   -0.46   -0.47    -0.47  -0.46
democ        0.21   0.97    0.98    0.97    0.99    0.96     0.98   0.99
autoc        0.15   0.97    0.96    0.96    0.94    0.98     0.96   0.93
polity       0.20   0.91    0.93    0.92    0.95    0.90     0.93   0.96
polity2      0.15   0.10    0.16    0.14    0.26    0.07     0.18   0.27
durable      1.00   0.16    0.17    0.16    0.17    0.19     0.17   0.16
xrreg        0.16   1.00    1.00    1.00    0.99    1.00     1.00   0.99
xrcomp       0.17   1.00    1.00    1.00    1.00    0.99     1.00   0.99
xropen       0.16   1.00    1.00    1.00    0.99    0.99     0.99   0.99
xconst       0.17   0.99    1.00    0.99    1.00    0.99     1.00   1.00
parreg       0.19   1.00    0.99    0.99    0.99    1.00     0.99   0.98
parcomp      0.17   1.00    1.00    0.99    1.00    0.99     1.00   0.99
exrec        0.16   0.99    0.99    0.99    1.00    0.98     0.99   1.00
exconst      0.17   0.99    1.00    0.99    1.00    0.99     1.00   1.00
polcomp      0.17   0.98    0.99    0.98    0.99    0.98     0.99   0.99
prior        0.41   0.37    0.60    0.35    0.30   -0.08     0.34   0.55
emonth      -0.09   0.24    0.25    0.24    0.23    0.24     0.23   0.26
eday         0.09   0.16    0.15    0.14    0.14    0.14     0.15   0.14
eyear        0.21  -0.08   -0.08   -0.08   -0.08   -0.09    -0.06  -0.09
eprec        0.42  -0.38   -0.38   -0.38   -0.37   -0.38    -0.37  -0.38
interim      0.30   1.00    1.00    1.00    1.00    1.00     1.00   1.00
bmonth      -0.19   0.10    0.00    0.06   -0.22    0.25     0.27   0.04
bday        -0.16   0.13   -0.26   -0.37   -0.14   -0.23    -0.17  -0.29
byear         nan    nan     nan     nan     nan     nan      nan    nan
bprec        0.50   0.22    0.10    0.12    0.16    0.05     0.25  -0.07
post         0.37   0.63    0.83    0.55    0.92   -0.55     0.36   0.94
change       0.29   0.99    0.99    0.99    0.99    0.99     0.99   0.99
d4            nan    nan     nan     nan     nan     nan      nan    nan
sf            nan    nan     nan     nan     nan     nan      nan    nan
regtrans     0.31   1.00    1.00    1.00    1.00    1.00     1.00   1.00

          exconst  polcomp  prior  emonth  eday  eyear  eprec  interim  \
cyear       -0.10    -0.13  -0.48    0.03 -0.11  -0.15  -0.19     0.23
ccode       -0.10    -0.13  -0.48    0.03 -0.11  -0.15  -0.19     0.23
year          nan      nan    nan     nan   nan    nan    nan      nan
flag          nan      nan    nan     nan   nan    nan    nan      nan
fragment    -0.46    -0.46  -0.06    0.09 -0.16   0.13  -0.24    -0.70
democ        0.99     0.99   0.52    0.24  0.15  -0.06  -0.36     1.00
autoc         0.94     0.92  -0.17    0.22  0.15  -0.09  -0.38     0.99
polity        0.95     0.97   0.41    0.26  0.14  -0.06  -0.36     1.00
polity2       0.26     0.33   0.41    0.14 -0.01   0.13   0.03     0.13
durable       0.17     0.17   0.41   -0.09  0.09   0.21   0.42     0.30
xrreg         0.99     0.98   0.37    0.24  0.16  -0.08  -0.38     1.00
xrcomp        1.00     0.99   0.60    0.25  0.15  -0.08  -0.38     1.00
xropen        0.99     0.98   0.35    0.24  0.14  -0.08  -0.38     1.00
xconst        1.00     0.99   0.30    0.23  0.14  -0.08  -0.37     1.00
parreg        0.99     0.98  -0.08    0.24  0.14  -0.09  -0.38     1.00
```

```
parcomp      1.00      0.99   0.34     0.23  0.15  -0.06  -0.37      1.00
exrec        1.00      0.99   0.55     0.26  0.14  -0.09  -0.38      1.00
exconst      1.00      0.99   0.30     0.23  0.14  -0.08  -0.37      1.00
polcomp      0.99      1.00   0.20     0.23  0.15  -0.06  -0.36      1.00
prior        0.30      0.20   1.00     0.18  0.04   0.41   0.22     -0.20
emonth       0.23      0.23   0.18     1.00  0.15  -0.18   0.07      0.33
eday         0.14      0.15   0.04     0.15  1.00   0.31   0.21      0.16
eyear       -0.08     -0.06   0.41    -0.18  0.31   1.00   0.24     -0.23
eprec       -0.37     -0.36   0.22     0.07  0.21   0.24   1.00     -1.00
interim      1.00      1.00  -0.20     0.33  0.16  -0.23  -1.00      1.00
bmonth      -0.22      0.11   0.03     0.36  0.24   0.46   0.20     -0.21
bday        -0.14     -0.03  -0.57     0.01  0.79  -0.05   0.08      0.74
byear         nan       nan    nan      nan   nan    nan    nan       nan
bprec        0.16      0.23   0.16     0.23  0.29   0.27   1.00       nan
post         0.92      0.73   0.35     0.11 -0.11   0.19   0.09      0.72
change       0.99      0.99  -0.45     0.25  0.14  -0.18  -0.38      0.99
d4            nan       nan    nan      nan   nan    nan    nan       nan
sf            nan       nan    nan      nan   nan    nan    nan       nan
regtrans     1.00      1.00  -0.76     0.22  0.13  -0.11  -0.39      0.99

          bmonth  bday  byear  bprec  post  change  d4  sf  regtrans
cyear       0.07  0.16    nan  -0.10 -0.82   -0.08 nan nan     -0.07
ccode       0.07  0.16    nan  -0.10 -0.82   -0.08 nan nan     -0.07
year         nan   nan    nan    nan   nan     nan nan nan       nan
flag         nan   nan    nan    nan   nan     nan nan nan       nan
fragment    0.40 -0.06    nan  -0.27  0.09   -0.72 nan nan     -0.73
democ      -0.10 -0.27    nan   0.13  0.95    0.99 nan nan      0.99
autoc      -0.03  0.02    nan   0.04 -0.90    0.98 nan nan      0.99
polity     -0.05 -0.18    nan   0.07  1.00    0.99 nan nan      0.99
polity2    -0.05 -0.18    nan   0.07  1.00    0.33 nan nan      0.25
durable    -0.19 -0.16    nan   0.50  0.37    0.29 nan nan      0.31
xrreg       0.10  0.13    nan   0.22  0.63    0.99 nan nan      1.00
xrcomp      0.00 -0.26    nan   0.10  0.83    0.99 nan nan      1.00
xropen      0.06 -0.37    nan   0.12  0.55    0.99 nan nan      1.00
xconst     -0.22 -0.14    nan   0.16  0.92    0.99 nan nan      1.00
parreg      0.25 -0.23    nan   0.05 -0.55    0.99 nan nan      1.00
parcomp     0.27 -0.17    nan   0.25  0.36    0.99 nan nan      1.00
exrec       0.04 -0.29    nan  -0.07  0.94    0.99 nan nan      1.00
exconst    -0.22 -0.14    nan   0.16  0.92    0.99 nan nan      1.00
polcomp     0.11 -0.03    nan   0.23  0.73    0.99 nan nan      1.00
prior       0.03 -0.57    nan   0.16  0.35   -0.45 nan nan     -0.76
emonth      0.36  0.01    nan   0.23  0.11    0.25 nan nan      0.22
eday        0.24  0.79    nan   0.29 -0.11    0.14 nan nan      0.13
eyear       0.46 -0.05    nan   0.27  0.19   -0.18 nan nan     -0.11
eprec       0.20  0.08    nan   1.00  0.09   -0.38 nan nan     -0.39
interim    -0.21  0.74    nan    nan  0.72    0.99 nan nan      0.99
bmonth      1.00  0.17    nan   0.20 -0.03   -0.15 nan nan      0.06
```

```
bday       0.17  1.00   nan   0.08 -0.20   0.24 nan nan    0.39
byear       nan   nan   nan    nan   nan    nan nan nan     nan
bprec      0.20  0.08   nan   1.00  0.09  -0.01 nan nan   -0.07
post      -0.03 -0.20   nan   0.09  1.00   0.61 nan nan    0.36
change    -0.15  0.24   nan  -0.01  0.61   1.00 nan nan    0.99
d4          nan   nan   nan    nan   nan    nan nan nan     nan
sf          nan   nan   nan    nan   nan    nan nan nan     nan
regtrans   0.06  0.39   nan  -0.07  0.36   0.99 nan nan    1.00
```

The correlation between regime type and durability for all countries in 2017 the dataset is basically 0.15. The correlation between regime type and durability for all countries in 2007 the dataset is also basically 0.15. The correlation between regime type and durability maintain almost the same from 2007 to 2017.

(n) Conduct a linear regression analysis between polity score and durability for all years, where the polity score is the independent variable. Show the output.

[19]:

```
results=smf.ols('durable~polity2',data=data).fit()
#make a regression

results.summary()
#show the result
```

[19]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                durable   R-squared:                       0.037
Model:                            OLS   Adj. R-squared:                  0.037
Method:                 Least Squares   F-statistic:                     610.3
Date:                Fri, 22 Nov 2019   Prob (F-statistic):          2.98e-132
Time:                        17:40:52   Log-Likelihood:                -74163.
No. Observations:               15895   AIC:                         1.483e+05
Df Residuals:                   15893   BIC:                         1.483e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     21.9371      0.204    107.570      0.000      21.537      22.337
polity2        0.7148      0.029     24.704      0.000       0.658       0.771
==============================================================================
Omnibus:                     7398.657   Durbin-Watson:                   0.067
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            44403.334
Skew:                           2.191   Prob(JB):                         0.00
Kurtosis:                       9.917   Cond. No.                         7.05
==============================================================================

Warnings:
```

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

(o) Interpret the coefficient, confidence interval, and p-value for the coefficient on x With each unit increase in polity score, durable would increase by 0.7148 units. The 95% confidence interval shows that 95% the simulation [0.658,0.771] catches the true value or not. P value is basically zero that it show the data is statistically significant to reject the null hypothesis.

(p) Interpret the R2 for this regression with respect to the variance of y explained. 3.7% of the variance in the polity can be explained by the model.

(q) Based on these results, what can the researcher conclude about the relationship between durability and polity scores? Please discuss this conclusion first statistically in terms of the null and alternative hypotheses and then substantively in terms of the magnitude of the estimated effect. The null hypothesis is: there's no association between durability and regime types.

The Alternative hypothesis is: there's association between durability and regime types.

Since P-value is much smaller than the significance level, the result is statically significant to reject the null hypothesis. There's a linear relationship between durability and polity scores. When polity score increases by 1 unit, durability increases by 0.7148 unit. When Polity score is 0, durability would be 21.9371.

(r) Find the codebook for this dataset and examine how durable is measured. Briefly describe how it is measured, and then comment on at least one strength and one weakness of how it is measured. In calculating the DURABLE value, the first year during which a new (post-change) polity is established is coded as the baseline "year zero" (value = 0) and each subsequent year adds one to the value of the DURABLE variable consecutively until a new regime change or transition period occurs.

Strength: The measure calculates the year of durability in a simple and effective way. Every time a new regime change occurs, the value would be stagnant. We can search for the number directly to see the length of stable regimes.

Weakness: This measure cannot show the extent of regime change. A tiny change would affect the durability to be stagnant, as well as a huge decrease or increase in regime types. Durability is calculated with high sensitivity that it might consider unnecessary tiny changes and be totally influenced.

(s) Given your new understanding of the durability measure, how comfortable are you with the findings from your regression analysis? Why? Since durability basically depends on regime changes, the regression analysis makes sense. When there's a change in regime types, it would directly cause durability to change.

(t) What is a different way to measure durability that might address the weakness you identified? How might this new measure change the conclusions drawn from the regression analysis? Measure:In calculating the DURABLE value, the first year during which a new (post-change) polity is established is coded as the baseline "year zero" (value = 0) and each subsequent year adds one times polity score to the value of the DURABLE variable consecutively until a new regime change or transition period occurs.

This new measure takes the extent of regime change into account. This would make the regression analysis with higher power. The association would be greater between the two.

1.1 question two

2.(a) Pre-step: Describe briefly a broad topic you're interested in and would like to explore. Why are you interested in this and what do you hope to discover? As an international student, I have to fly faraway every semester. I am interested in researching airline safety problems, in order to discover if I should avoid those that have had crashes in the past. I want to find out which airline has the most incidents in the past.

(b) Dataset: Find a dataset that may help you explore at least some of these questions. Briefly describe the dataset (main variables of interest, how it's structured) and how you went about finding it. The main variables of ineterst are available seat kilometers flown every week, incidents number, total fatal accidents, total fatalities. They are divided into mainly to groups by year, 1985-1999 and 2000-2014. I find it directly from the lecture.

(c) Why did you choose this dataset? Comment on the process of looking for the dataset –Was it difficult? Easy? Were you deciding between multiple possible datasets? If you chose one of the datasets recommended in lecture, what drew you to that one? I think the variables of interest is suitalbe to research the topics. Counting both incidents and deaths makes the dataset persuasive. I also like the way it is structured. By dividing the data into two time periods, we can see if the safety problem is persistent. It grabs my eyes because it' is easy to follow and analyze.

(d) What are some of the strengths and limitations of the dataset that you can see so far? Please provide at least one of each. Strength: The most crucial elements to measure safety are all included in this dataset.For example, if only fatality is considered, it might be a simple huge plane crash that caused all the deaths. The airline might have only one accident in the past decades but be considered more dangerous than those that have several accidents. Although, old datas might weight less now when considering how safe the airline is.

Limitation: The dataset does not include other counfounders like eathquakes or breakup of the Soviet Union during that period. But these do cause some airlines to undergo high rate of incidents and fatalities.

(e) Import the dataset into Jupyter using any method you like and show the first ten observations. If you had to do any pre-work to get the data into an uploadable format please describe it briefly. (If you didn't, please say so as well.)

[20]:
```
url1='https://raw.githubusercontent.com/fivethirtyeight/data/master/
 ↪airline-safety/airline-safety.csv'
#import from github, name it url1

data_airline=pd.read_csv(url1)
#read the data, call it data_airline
```

```
pd.set_option('display.max_columns',8)
#display all columns of the data

data_airline[:10]
#show the first ten rows of the dataset
```

[20]:

```
                 airline  avail_seat_km_per_week  incidents_85_99  \
0             Aer Lingus               320906734                2
1              Aeroflot*              1197672318               76
2  Aerolineas Argentinas               385803648                6
3            Aeromexico*               596871813                3
4             Air Canada              1865253802                2
5             Air France              3004002661               14
6             Air India*               869253552                2
7       Air New Zealand*               710174817                3
8       Alaska Airlines*               965346773                5
9               Alitalia               698012498                7

   fatal_accidents_85_99  fatalities_85_99  incidents_00_14  \
0                      0                 0                0
1                     14               128                6
2                      0                 0                1
3                      1                64                5
4                      0                 0                2
5                      4                79                6
6                      1               329                4
7                      0                 0                5
8                      0                 0                5
9                      2                50                4

   fatal_accidents_00_14  fatalities_00_14
0                      0                 0
1                      1                88
2                      0                 0
3                      0                 0
4                      0                 0
5                      2               337
6                      1               158
7                      1                 7
8                      1                88
9                      0                 0
```

(f) Initial analysis: Conduct at least two manipulations of your now-ready table that help you understand something of interest about the dataset (e.g., you might explore options like sort, shape, value counts, groupby, etc.). Why did you choose these two, and what have you learned? (Hint: You may need to do a bit work to get the data into a format that is usable for you – e.g., renaming columns, changing data types, etc. If any of this was necessary, show your code and

briefly explain why you made these changes)

[21]:
```python
incident=data_airline[['airline','incidents_85_99','incidents_00_14']]
#show only data of incidents

incident['total']=incident['incidents_85_99']+incident['incidents_00_14']
#calcute the total incidents influence

incident.sort_values('total', ascending = False).head(5)
#show the five airlines that have largest value of total
```

```
/share/apps/jupyterhub/2019-FA-DS-UA-111/lib/python3.7/site-
packages/ipykernel_launcher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/indexing.html#indexing-view-versus-copy
  after removing the cwd from sys.path.
```

[21]:
```
                  airline  incidents_85_99  incidents_00_14  total
1                Aeroflot*               76                6     82
19      Delta / Northwest*               24               24     48
11               American*               21               17     38
51    United / Continental*              19               14     33
22       Ethiopian Airlines              25                5     30
```

The first method I choose is to add a new column and calculate total incidents. I find it direct and clear to see which airline has the most incidents in the past. By sorting the data,I discover the five airlines having most incidents in the past.

[22]:
```python
incident['total'].describe()
#analyze the created variable
```

[22]:
```
count    56.00
mean     11.30
std      13.52
min       0.00
25%       4.00
50%       8.00
75%      12.25
max      82.00
Name: total, dtype: float64
```

By discribing the data, I could see the mean and median. I could know the basic distribution of total incidents of all airlines.

(g) Generate two graphs of any kind that are useful to you to better understand what you're interested in. They don't need to be formatted particularly beautifully, but you do need to use two different types of graphs (e.g., a bar chart and a scatterplot) and explain what you hoped

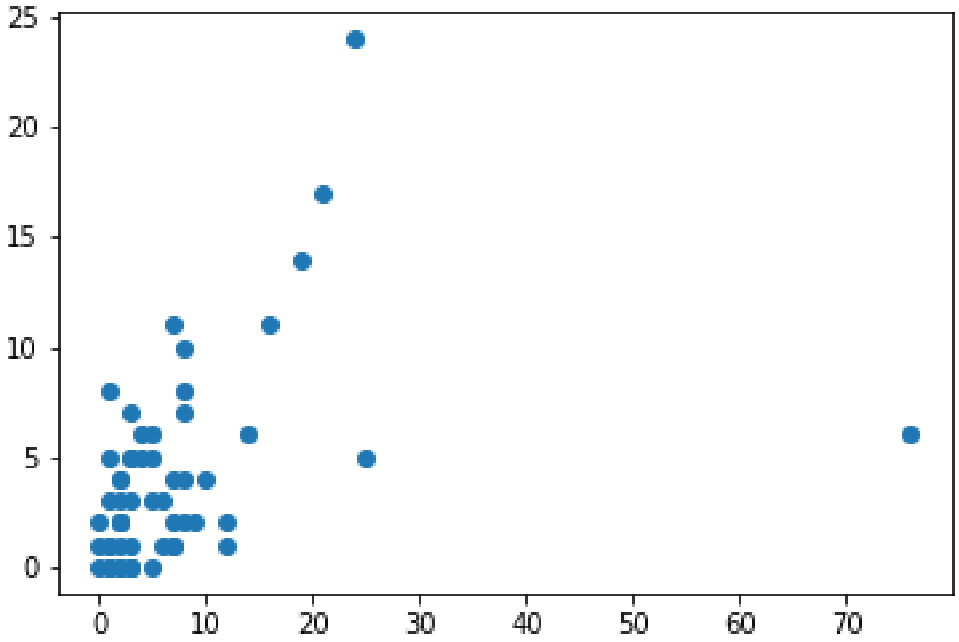
to understand, why you chose these graphs, and whether they're useful in improving your understanding.

[23]:

```
plt.scatter(data_airline['incidents_85_99'],data_airline['incidents_00_14'])
#this graph helps me to see the incidents of both periods clearly.
#The points that are faraway from orgin are those more dangerous airlines.
```

[23]:

```
<matplotlib.collections.PathCollection at 0x2af8c98095f8>
```

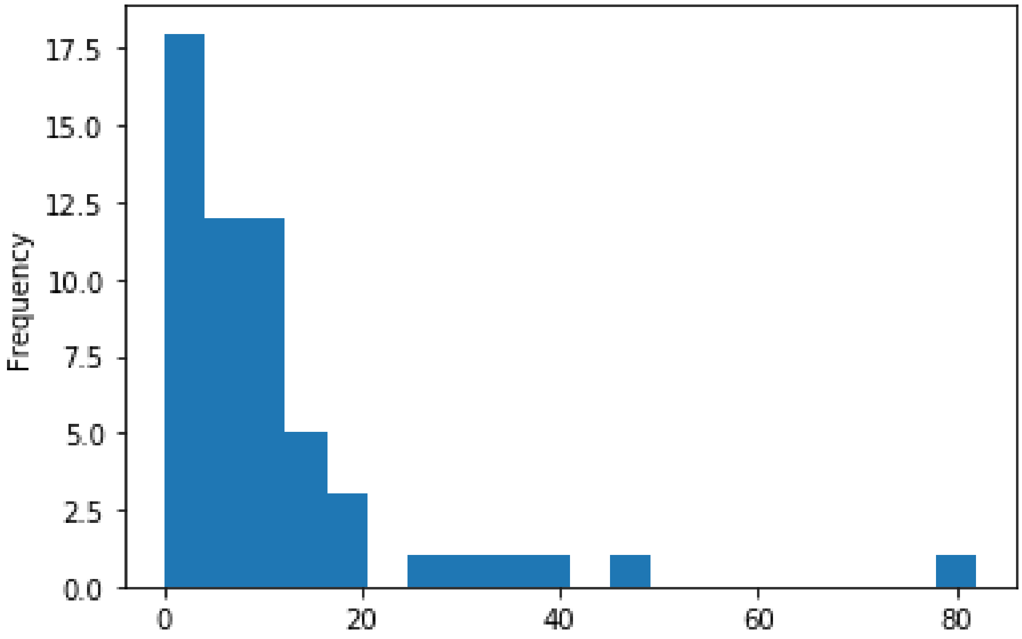


[24]:

```
incident['total'].plot.hist(bins=20)
#This graph helps me to know the overall level of safety of all airlines.
#The distribution is skewed to the right.
#This shows me that median is more valuable than mean when measuring the center.
```

[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2af8c9accd30>
```



(h) Hypothesis formation: What is your dependent variable and independent variable? Briefly describe how they are measured in this dataset. (Remember, they'll both need to be continuous variables.) My dependent variable: "incidents_85_99", which is the inceidents in 1985-1999. My independent variable:"incidents_00_14", which is the inceidents in 2000-2014.

(i) Calculate the correlation coefficient between your two variables and interpret the result.

```
[25]: incident.corr() #show all the correlations
```

```
[25]:                  incidents_85_99  incidents_00_14  total
 incidents_85_99              1.00             0.40   0.95
 incidents_00_14              0.40             1.00   0.66
 total                        0.95             0.66   1.00
```

The correlation coefcient between incidents_00_14 and incidents_85_99 is basically 0.40. The two variables have a moderately weak positive association.

(j) Write out your regression model as an equation. incidents_85_99=+incidents_00_14

(k) Write out your null and alternative hypotheses. Null hypothesis: There's no association between incidents_85_99 and incidents_00_14. Alternative hypothesis: There's association between incidents_85_99 and incidents_00_14.

(l) Regression: Estimate the regression equation you specified above and show the regression output.

[26]:

```
results=smf.ols('incidents_85_99~incidents_00_14',data=incident).fit()
#make the regression

results.summary()
#show the result
```

[26]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:        incidents_85_99   R-squared:                       0.162
Model:                            OLS   Adj. R-squared:                  0.147
Method:                 Least Squares   F-statistic:                     10.47
Date:                Fri, 22 Nov 2019   Prob (F-statistic):            0.00207
Time:                        17:40:53   Log-Likelihood:                -208.46
No. Observations:                  56   AIC:                             420.9
Df Residuals:                      54   BIC:                             425.0
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
===
                      coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---
Intercept           3.1421      1.847      1.701      0.095      -0.561
6.845
incidents_00_14     0.9785      0.302      3.236      0.002       0.372
1.585
==============================================================================
Omnibus:                      102.067   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2816.837
Skew:                           5.395   Prob(JB):                         0.00
Kurtosis:                      36.027   Cond. No.                         8.38
==============================================================================

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

(m) Interpretation & diagnostics: What do the results in the regression output tell you? Interpret the coefficient, p-value, and confidence interval for your independent variable (you don't have to do the intercept) and the R2. When there's 1 increase in incidents_85_99, there's 0.9785 increase in incidents_00_14. The 95% confidence interval means that 95% of the interval [0.372, 1.586] catches the true value or not. 16.2% of the variance in the dependent variable could be explained by the model. P-value is 0.002, which is smaller than the significance level. We can reject the null hypothesis. There's an association between incidents_85_99 and incidents_00_14.

(n) Which hypothesis do you reject and fail to reject, and why? We can reject the null hypothesis that there's no association between incidents_85_99 and incidents_00_14 because P-value equals 0.002, is much smaller than the 0.05 significance level.There's an association between incidents_85_99 and incidents_00_14.

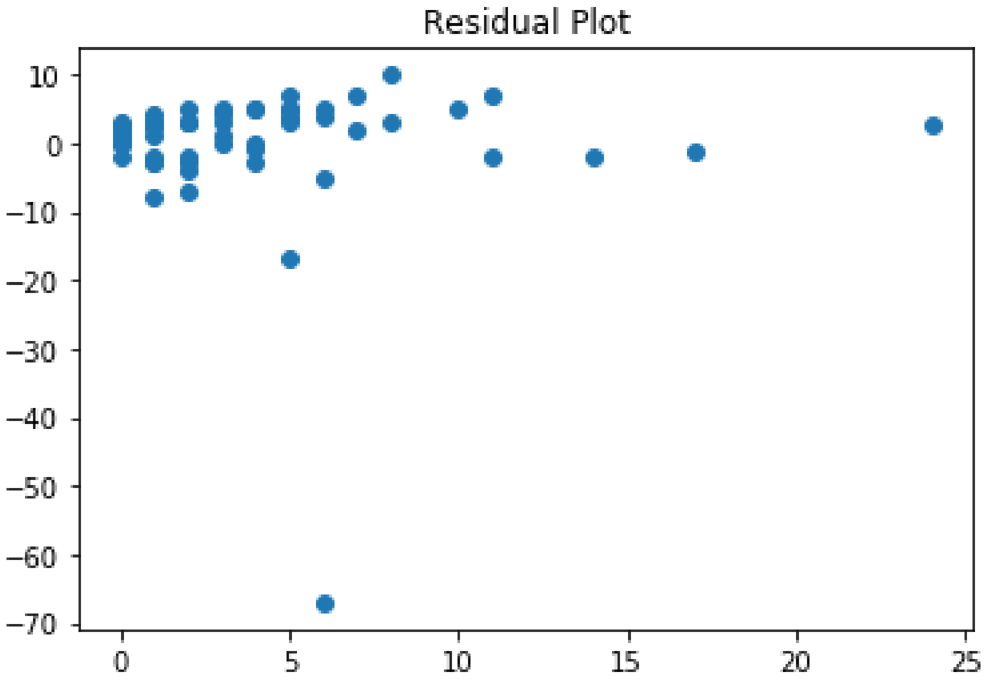
(o) Generate the residual plot and comment on any heteroskedasticity. What does this imply for your inference?

[27]:

```
intercept=results.params[0]
#define intercept

slope=results.params[1]
#define slope

predictions=intercept+slope*incident['incidents_00_14']
#define predictions

resid=(predictions-incident['incidents_85_99'])
#define residuls

plt.scatter(incident['incidents_00_14'], resid)
#make up the scatter plot

plt.title('Residual Plot')
#name the plot as Residual Plot
```

[27]:

```
Text(0.5, 1.0, 'Residual Plot')
```



There's basically no heteroskedasticity in the graph. It shows that a linear regression is suitable for modeling the data. The two outliers might influence the coefficients, but does not mean that linear model is bad for the problem.

(p) Conclusions: What biases might be present in the sample itself that could be affecting the outcome? Discuss at least two possible types of bias.

1. There might be data recorded incorrectly that might influence the result. Missing some crashes might affect the outcome.

2. Information bias: Some airlines might report fewer crashes than the actual amount. This might affect the outcome.

(q) Considering all the work you've done, including the regression output, the results of your hypothesis tests, and any biases present in the data, what conclusions, however tentative, can you draw from your analysis about the relationship between your two variables of interest? The trend of incidents does not vary too much as the time passes. We can predict future trend by analyzing these two data. There's a positive linear association between incidents_85_99 and incidents_00_14. Dispite the fact that there might be special economic crisis or natural disasters causing more incidents, basically Delta/Northwest, American are some unsafe airlines. They continue to have high rate of incidents.

(r) What are at least two changes or improvements that could be made that you think would strengthen this analysis, and why would they help? (These can be about anything, e.g., data, biases, analysis, results, and/or diagnostics.)

1. Investigate the time period when incidents blust. If there's certain natural disasters or wars, we might take that period out from our research. This could take outliers out and make our model more precise.

2. we can analyze death per incident next time to see the capability of an airline company to take emergenct actions when facing incidents. Although not experiencing incidents is always the best, but if one could be saved in most incidents, it might sometimes be a better choice.