# Week 2 - Visualization - Python

September 5, 2018

# 1 Data Warehousing and Data Mining

## 1.1 Labs

### 1.1.1 Prepared by Gilroy Gordon

**Contact Information** SCIT ext. 3643
ggordonutech@gmail.com
gilroy.gordon@utech.edu.jm

### 1.1.2 Week 2 - Visualization in Python

Additional Reference Resources:
https://matplotlib.org/gallery/index.html
https://matplotlib.org/users/pyplot_tutorial.html
https://seaborn.pydata.org/examples/index.html

## 1.2 Objectives

---

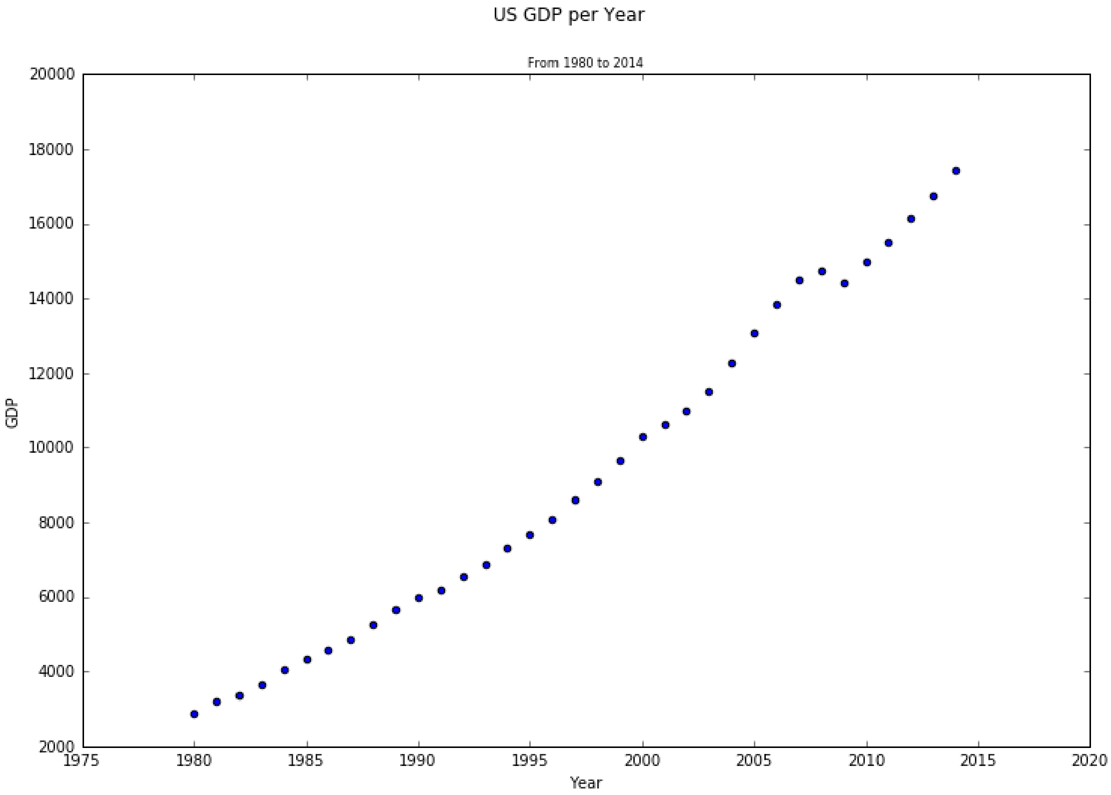
```
> Importing Data
    > csv
> 2D Visualization
     > Bar Plots
     > Scatter Plots
     > Box Plot
     > Histograms
     > Line Charts
```

```
In [10]: import pandas as pd # assists with managing data frames and data series, useful data st
         import numpy as np
         import os
         import matplotlib.pyplot as plt
         # indicates that we want our plots to be shown in our notebook and not in a sesparate w
         %matplotlib inline
```

```
In [11]: # What files are available in the current directory?
         os.listdir('.')

Out[11]: ['Week 2 - Visualization - Python.ipynb',
          'Week 2 - Visualization - R.ipynb',
          'data',
          '.ipynb_checkpoints']

In [12]: # What files are available in the "./data" directory?
         os.listdir('./data')

Out[12]: ['crime_incidents_2013_data.csv',
          'US GDP.csv',
          'crime_incidents_2013_location.csv',
          'NBA.csv']

In [13]: #read the contents of the 'crime_incidents_2013_data.csv' as a csv file and return the
         # store data in cr2013
         us_gdp = pd.read_csv('data/US GDP.csv')

         #preview the first 8 records of the dataset
         us_gdp.head(8)

Out[13]:    Year  US_GDP_BN  GDP_Growth_PC
         0  1980       2863            0.0
         1  1981       3211           12.2
         2  1982       3345            4.2
         3  1983       3638            8.8
         4  1984       4041           11.1
         5  1985       4347            7.6
         6  1986       4590            5.6
         7  1987       4870            6.1
```

### 1.2.1 Scatter Plot

```
In [37]: us_gdp.plot(kind="scatter", # or `us_gdp.plot.scatter(`
             x='Year',
             y='US_GDP_BN',
             title="US GDP per year",
             figsize=(12,8)
         )
         plt.title("From %d to %d" % (
             us_gdp['Year'].min(),
             us_gdp['Year'].max()
         ),size=8)
         plt.suptitle("US GDP per Year",size=12)
         plt.ylabel("GDP")

Out[37]: <matplotlib.text.Text at 0x7f2a60a0d1d0>
```

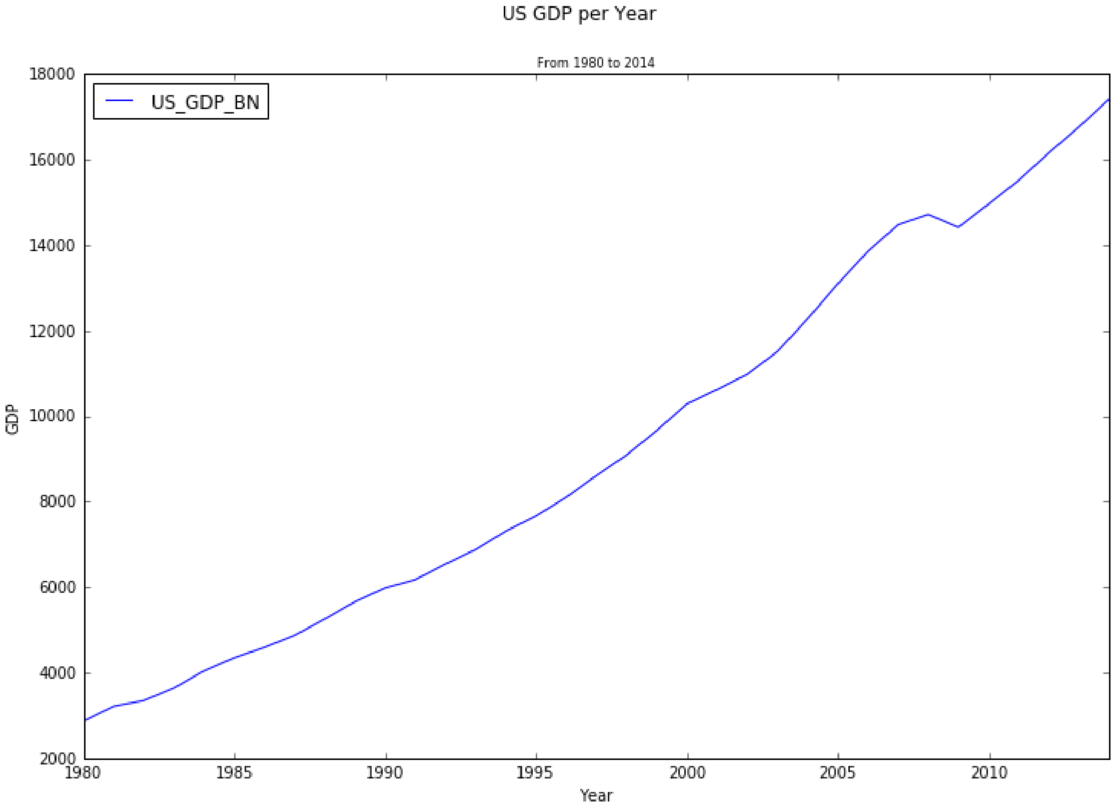


## 2 Line Graph

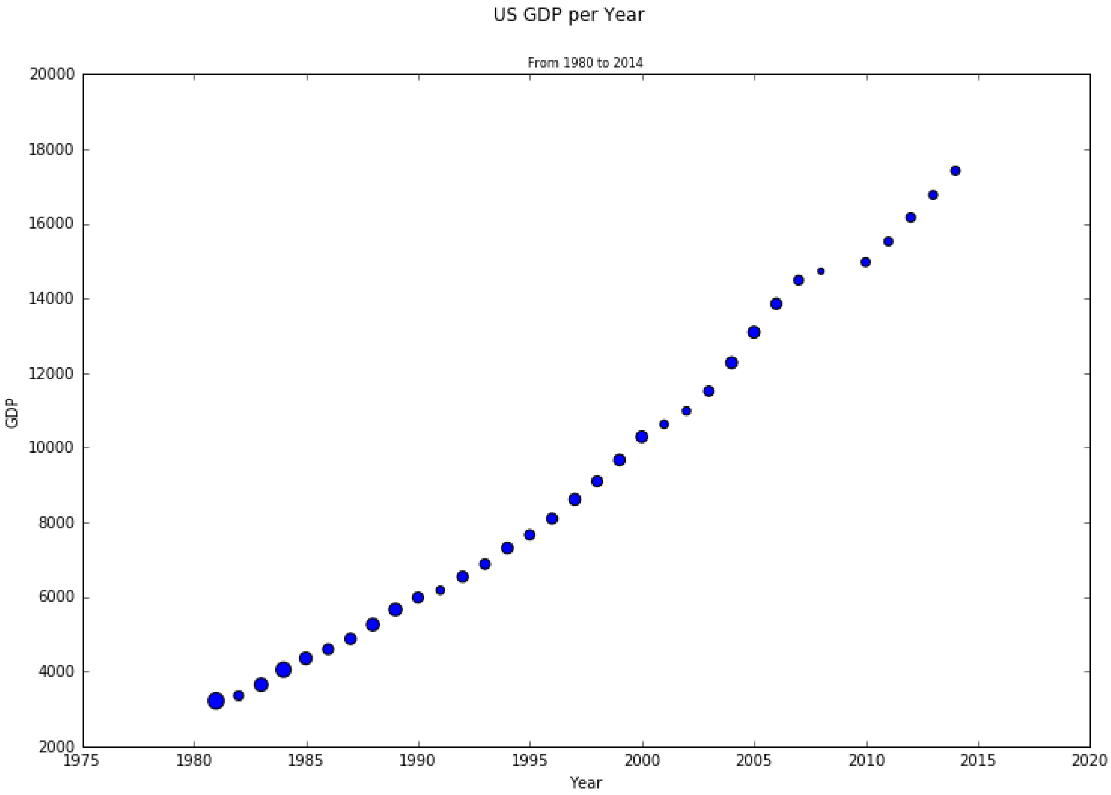
"Exploring the trend"

```
In [38]: us_gdp.plot(kind="line", # or `us_gdp.plot.line(`
             x='Year',
             y='US_GDP_BN',
             title="US GDP per year",
             figsize=(12,8)
         )
         plt.title("From %d to %d" % (
             us_gdp['Year'].min(),
             us_gdp['Year'].max()
         ),size=8)
         plt.suptitle("US GDP per Year",size=12)
         plt.ylabel("GDP")
```

```
Out[38]: <matplotlib.text.Text at 0x7f2a60d28588>
```

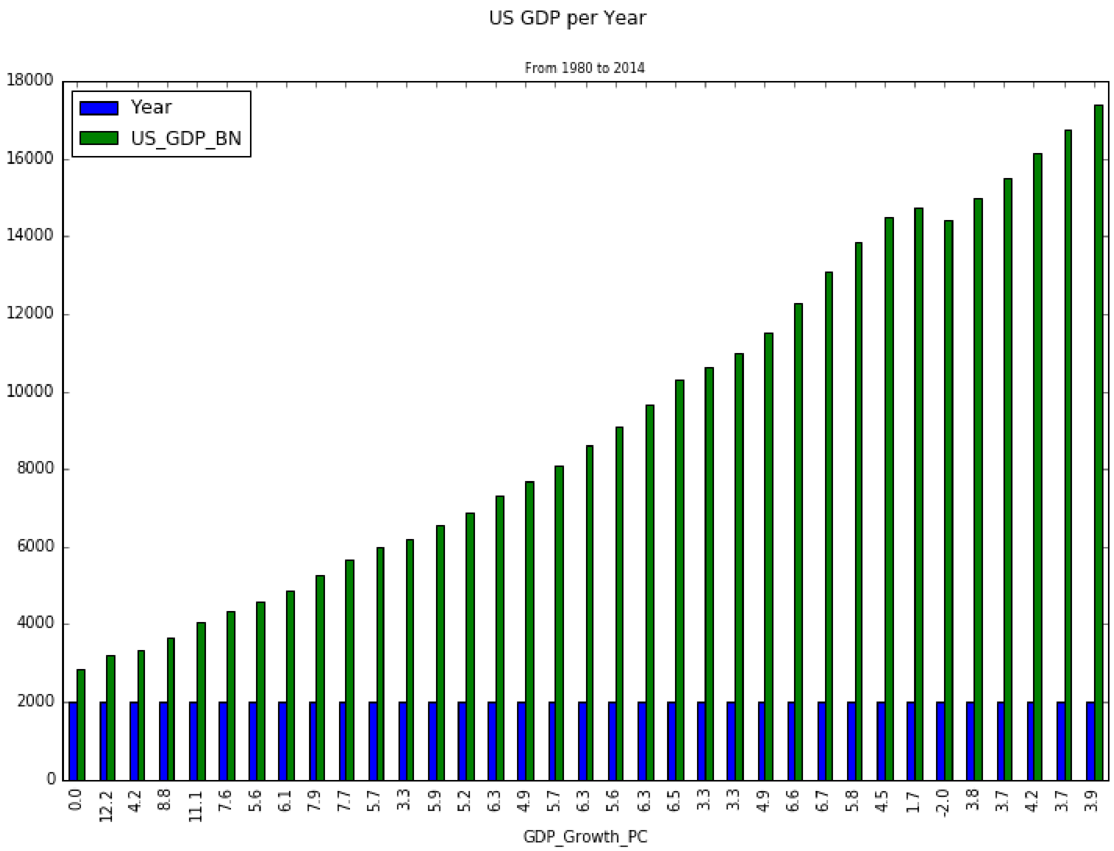


```
In [53]: us_gdp.plot(kind="scatter", # or `us_gdp.plot.line(`
             x='Year',
             y='US_GDP_BN',
             s=us_gdp['GDP_Growth_PC'].apply(lambda growth: 0 if growth < 0 else growth) * 9,
             title="US GDP per year",
             figsize=(12,8)
         )
         plt.title("From %d to %d" % (
             us_gdp['Year'].min(),
             us_gdp['Year'].max()
         ),size=8)
         plt.suptitle("US GDP per Year",size=12)
         plt.ylabel("GDP")
         "Size of each point on the Scatter plot is GDP Growth %"

Out[53]: 'Size of each point on the Scatter plot is GDP Growth %'
```



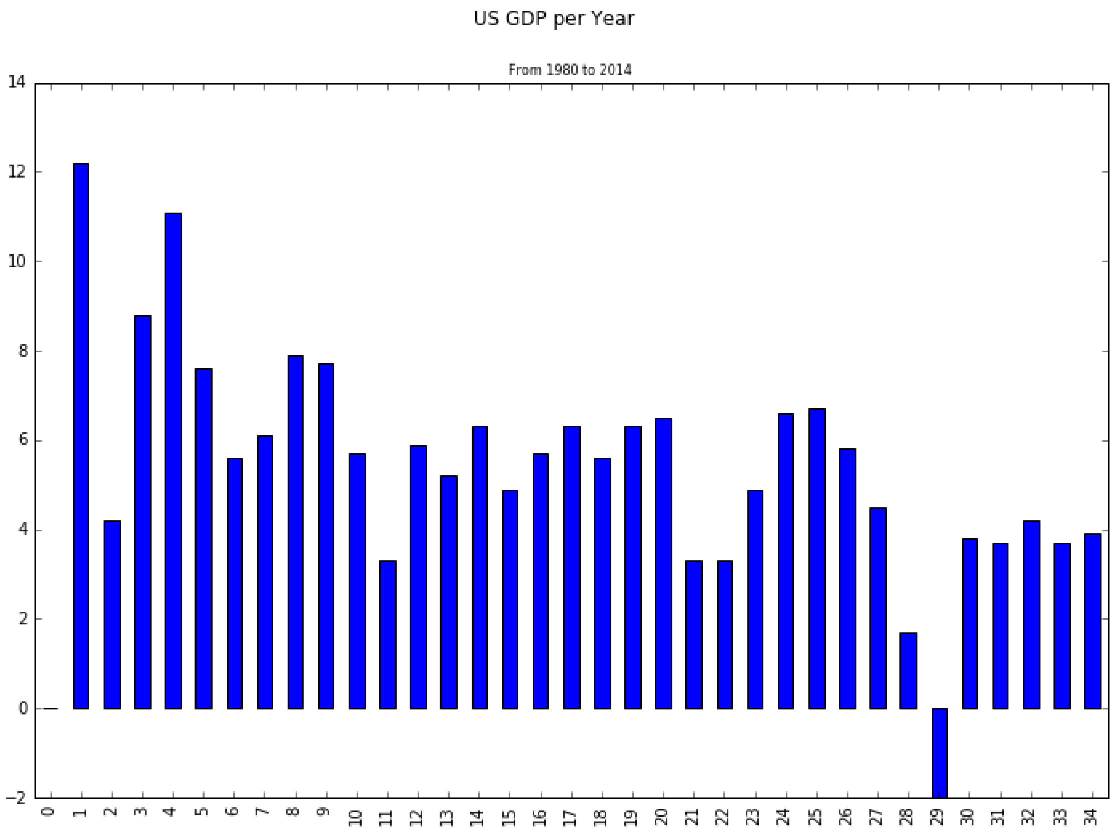
### 2.0.1 Bar Plot - Histogram

```
In [56]: us_gdp.plot(kind="bar", # or `us_gdp.plot.line(`
             x='GDP_Growth_PC',
             figsize=(12,8)
         )
         plt.title("From %d to %d" % (
             us_gdp['Year'].min(),
             us_gdp['Year'].max()
         ),size=8)
         plt.suptitle("US GDP per Year",size=12)

Out[56]: <matplotlib.text.Text at 0x7f2a5fff5ac8>
```



```
In [57]: us_gdp['GDP_Growth_PC'].plot(kind="bar", # or `us_gdp.plot.line(`

             figsize=(12,8)
         )
         plt.title("From %d to %d" % (
             us_gdp['Year'].min(),
             us_gdp['Year'].max()
         ),size=8)
         plt.suptitle("US GDP per Year",size=12)
```

```
Out[57]: <matplotlib.text.Text at 0x7f2a5fe80320>
```



```
In [60]: # What did we really just visualize?
         us_gdp['GDP_Growth_PC'].value_counts() #shows counts of value
```

```
Out[60]:  6.3     3
          3.3     3
          3.7     2
          5.7     2
          4.2     2
          5.6     2
          4.9     2
         -2.0     1
          12.2    1
          1.7     1
          6.1     1
          7.6     1
          8.8     1
          6.7     1
          11.1    1
          5.8     1
          3.9     1
          7.9     1
          5.9     1
          6.6     1
          3.8     1
          4.5     1
          6.5     1
          5.2     1
          7.7     1
          0.0     1
         Name: GDP_Growth_PC, dtype: int64
```

```
In [68]: GDP_Growth_PC_binned = pd.cut(us_gdp['GDP_Growth_PC'],5)
         print(GDP_Growth_PC_binned)
         print("*"*32)
         print(GDP_Growth_PC_binned.unique())
```

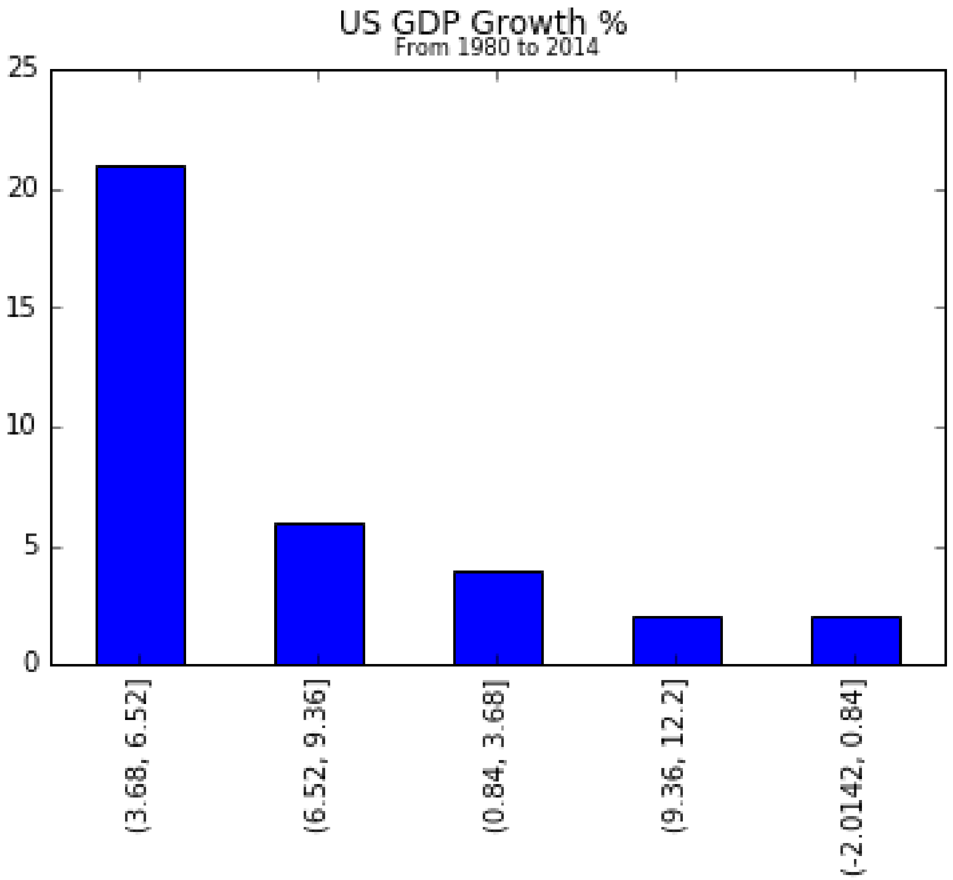
```
0     (-2.0142, 0.84]
1        (9.36, 12.2]
2        (3.68, 6.52]
3        (6.52, 9.36]
4        (9.36, 12.2]
5        (6.52, 9.36]
6        (3.68, 6.52]
7        (3.68, 6.52]
8        (6.52, 9.36]
9        (6.52, 9.36]
10       (3.68, 6.52]
11       (0.84, 3.68]
12       (3.68, 6.52]
13       (3.68, 6.52]
14       (3.68, 6.52]
15       (3.68, 6.52]
16       (3.68, 6.52]
17       (3.68, 6.52]
18       (3.68, 6.52]
19       (3.68, 6.52]
20       (3.68, 6.52]
21       (0.84, 3.68]
22       (0.84, 3.68]
23       (3.68, 6.52]
24       (6.52, 9.36]
25       (6.52, 9.36]
26       (3.68, 6.52]
27       (3.68, 6.52]
28       (0.84, 3.68]
29    (-2.0142, 0.84]
30       (3.68, 6.52]
31       (3.68, 6.52]
32       (3.68, 6.52]
33       (3.68, 6.52]
34       (3.68, 6.52]
Name: GDP_Growth_PC, dtype: category
Categories (5, object): [(-2.0142, 0.84] < (0.84, 3.68] < (3.68, 6.52] < (6.52, 9.36] < (9.36, 1
********************************
[(-2.0142, 0.84], (9.36, 12.2], (3.68, 6.52], (6.52, 9.36], (0.84, 3.68]]
Categories (5, object): [(-2.0142, 0.84] < (0.84, 3.68] < (3.68, 6.52] < (6.52, 9.36] < (9.36, 1
```

```
In [69]: GDP_Growth_PC_binned.value_counts()
```

```
Out[69]: (3.68, 6.52]       21
         (6.52, 9.36]        6
         (0.84, 3.68]        4
         (9.36, 12.2]        2
         (-2.0142, 0.84]     2
         dtype: int64
```

```
In [72]: GDP_Growth_PC_binned.value_counts().plot(kind='bar')
         plt.title("From %d to %d" % (
             us_gdp['Year'].min(),
             us_gdp['Year'].max()
         ),size=8)
         plt.suptitle("US GDP Growth %",size=12)
```

```
Out[72]: <matplotlib.text.Text at 0x7f2a609bed30>
```



### 2.0.2 Box plot

```
In [90]: us_gdp_groups = pd.cut(us_gdp.Year,3,labels=['Early 1990s','Late 1990s to Early 2000s',
         us_gdp_groups.value_counts()
```

```
Out[90]: After Early 2000s           12
         Early 1990s                 12
         Late 1990s to Early 2000s   11
         dtype: int64
```

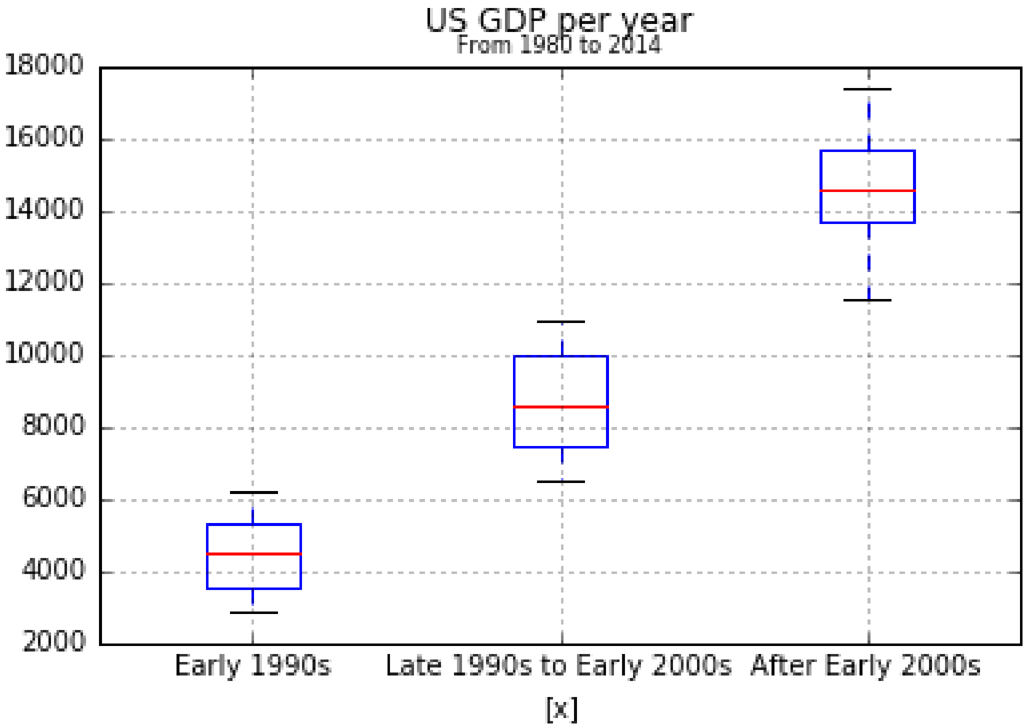
```
In [96]: us_gdp_bn_with_groups = pd.DataFrame({'x':us_gdp_groups,'y':us_gdp['US_GDP_BN']})
         us_gdp_bn_with_groups
```

```
Out[96]:                            x      y
         0                Early 1990s   2863
         1                Early 1990s   3211
         2                Early 1990s   3345
         3                Early 1990s   3638
         4                Early 1990s   4041
         5                Early 1990s   4347
```

```
6                  Early 1990s   4590
7                  Early 1990s   4870
8                  Early 1990s   5253
9                  Early 1990s   5658
10                 Early 1990s   5980
11                 Early 1990s   6174
12  Late 1990s to Early 2000s   6539
13  Late 1990s to Early 2000s   6879
14  Late 1990s to Early 2000s   7309
15  Late 1990s to Early 2000s   7664
16  Late 1990s to Early 2000s   8100
17  Late 1990s to Early 2000s   8609
18  Late 1990s to Early 2000s   9089
19  Late 1990s to Early 2000s   9661
20  Late 1990s to Early 2000s  10285
21  Late 1990s to Early 2000s  10622
22  Late 1990s to Early 2000s  10978
23          After Early 2000s  11511
24          After Early 2000s  12275
25          After Early 2000s  13094
26          After Early 2000s  13856
27          After Early 2000s  14478
28          After Early 2000s  14719
29          After Early 2000s  14419
30          After Early 2000s  14964
31          After Early 2000s  15518
32          After Early 2000s  16163
33          After Early 2000s  16768
34          After Early 2000s  17419
```

```
In [108]: us_gdp_bn_with_groups.boxplot(by='x')
          plt.title("From %d to %d" % (
              us_gdp['Year'].min(),
              us_gdp['Year'].max()
          ),size=8)
          plt.suptitle("US GDP per year",size=12)
```

```
Out[108]: <matplotlib.text.Text at 0x7f2a5eb8ccc0>
```



**Many other options available . Try seaborn!**