Week 2 - Visualization - R

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 R

Additional Reference Resources:
http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization
https://www.statmethods.net/graphs/dot.html

1.2 Objectives

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

1.2.1 Importing Data

```
In [7]: # Import the ggplot2 library to assist with data visualization
        library(ggplot2) # if you receive and error that the library is not available run `insta
```

```
In [9]: # Import the dplyr library to assist with data transformation
        library(dplyr) # if you receive and error that the library is not available run `install
```

```
Attaching package: dplyr

The following objects are masked from package:stats:

    filter, lag

The following objects are masked from package:base:

    intersect, setdiff, setequal, union
```

```
In [72]: # What files are available in the current directory?
         dir()
```

1. ’data’ 2. ’Week 2 - Visualization - Python.ipynb’ 3. ’Week 2 - Visualization - R.ipynb’

```
In [71]: # What files are available in the "./data" directory?
         dir('data')
```

1. ’crime_incidents_2013_data.csv’ 2. ’crime_incidents_2013_location.csv’ 3. ’NBA.csv’ 4. ’US GDP.csv’

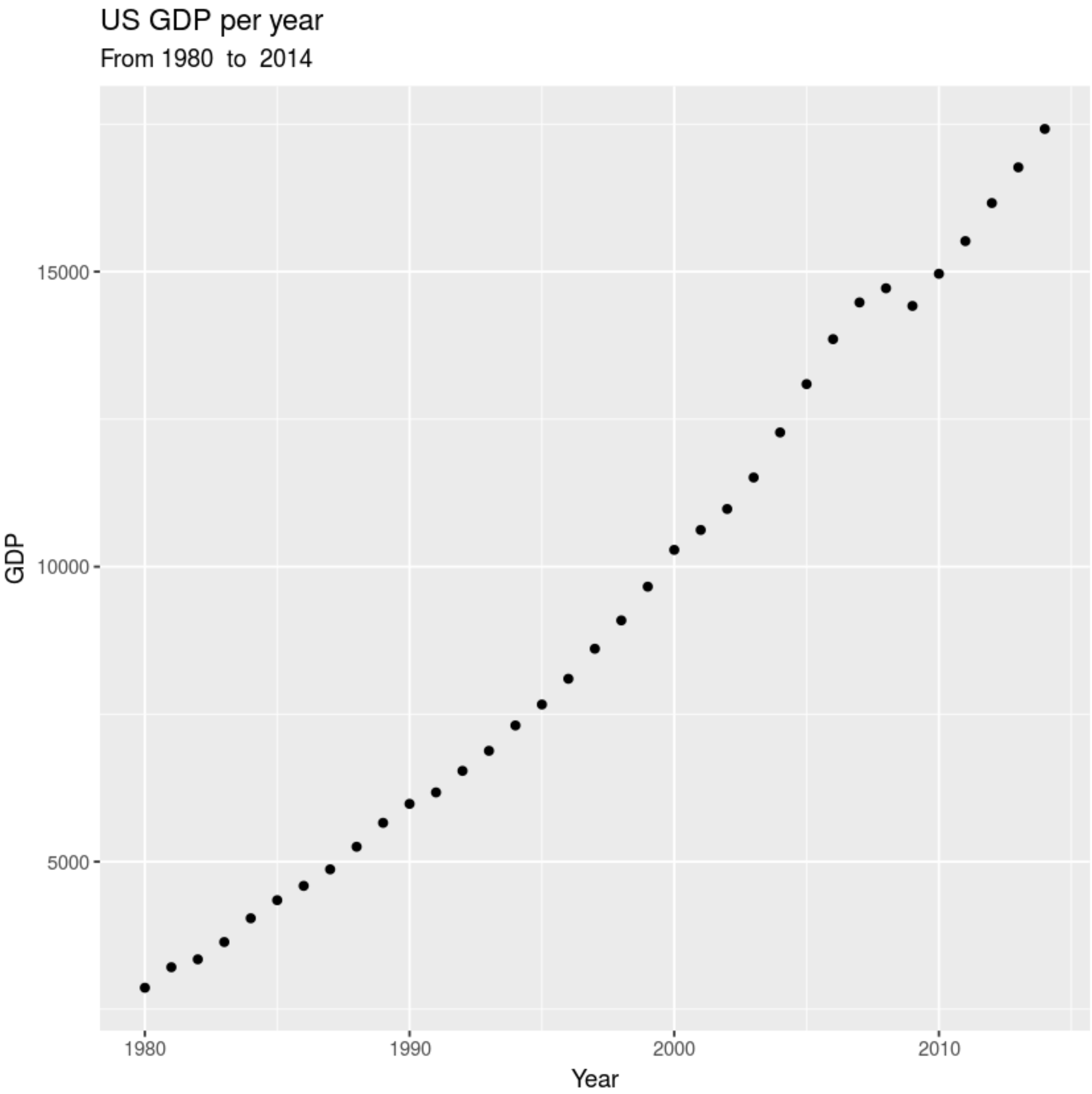
```
In [10]: #read the contents of the 'crime_incidents_2013_data.csv' as a csv file and return the
         # store data in cr2013
         us_gdp = read.csv('data/US GDP.csv')

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

Year	US_GDP_BN	GDP_Growth_PC
1980	2863	0.0
1981	3211	12.2
1982	3345	4.2
1983	3638	8.8
1984	4041	11.1
1985	4347	7.6
1986	4590	5.6
1987	4870	6.1

1.2.2 Scatter Plot

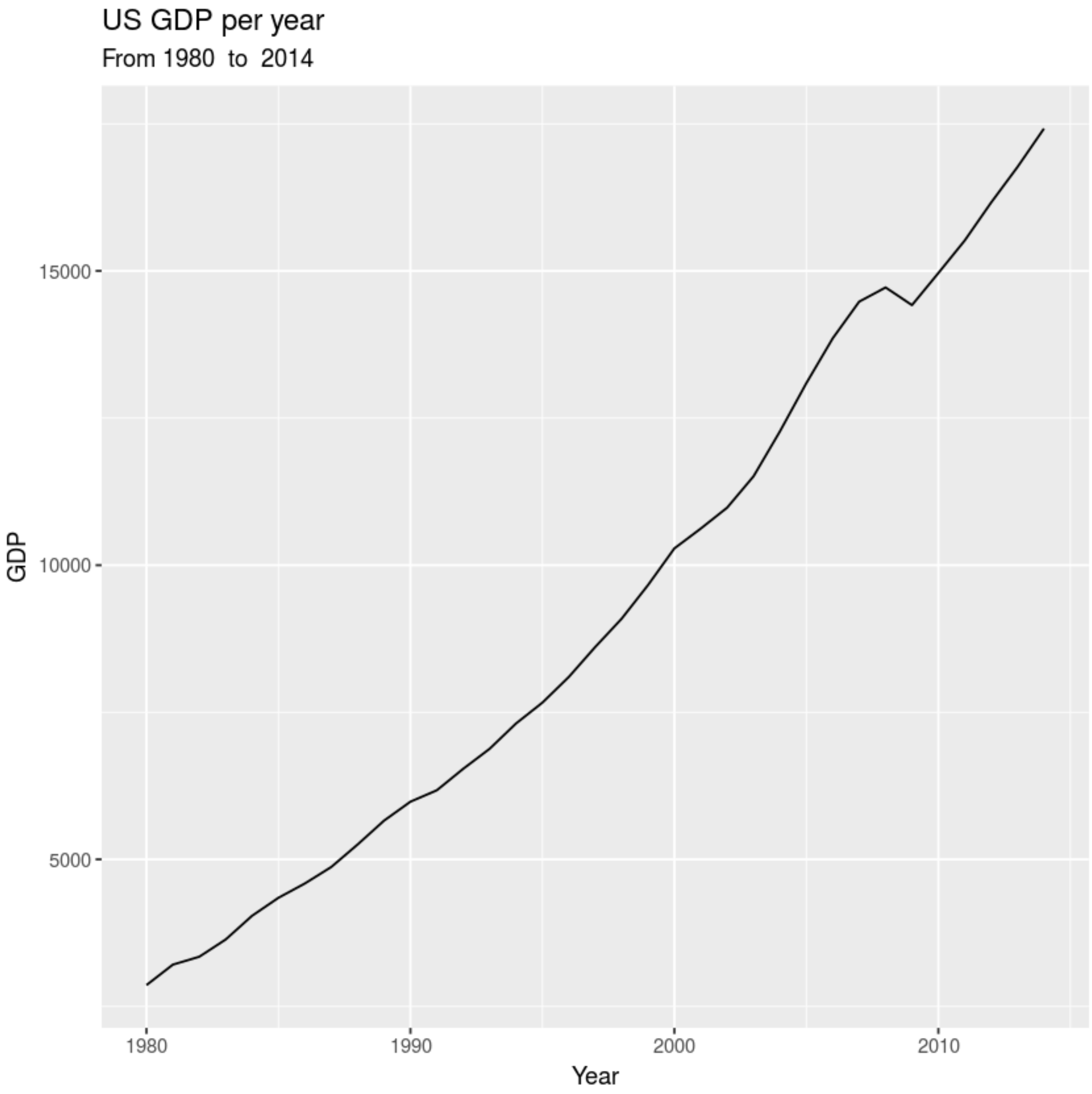
```
In [25]: ggplot(us_gdp, # start with the data
                aes(    # indicate which columns should be used where in the graph
                    x=Year,
                    y=US_GDP_BN
                )
               ) +
geom_point() + # use points i.e. a scatter plot
labs( # specify labels
    title="US GDP per year",
    subtitle=paste("From",min(us_gdp$Year)," to ",max(us_gdp$Year),collapse="")
) +
ylab("GDP")
```



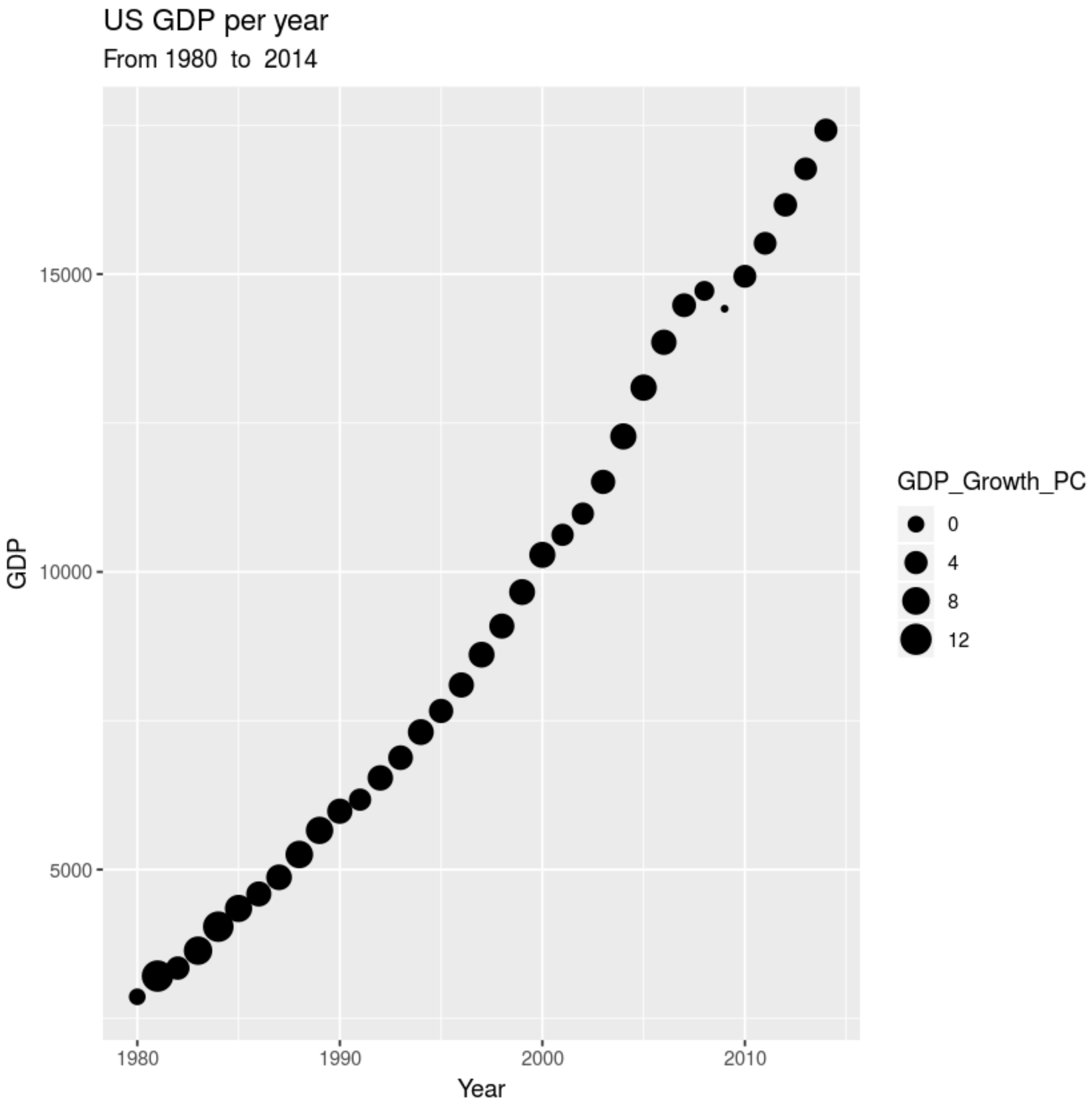
2 Line Graph

"Exploring the trend"

```
In [28]: ggplot(us_gdp, # start with the data
                aes(     # indicate which columns should be used where in the graph
                    x=Year,
                    y=US_GDP_BN
                )
               ) +
         geom_line() + # use a line
         labs( # specify labels
             title="US GDP per year",
             subtitle=paste("From",min(us_gdp$Year)," to ",max(us_gdp$Year),collapse="")
         ) +
         ylab("GDP")
```

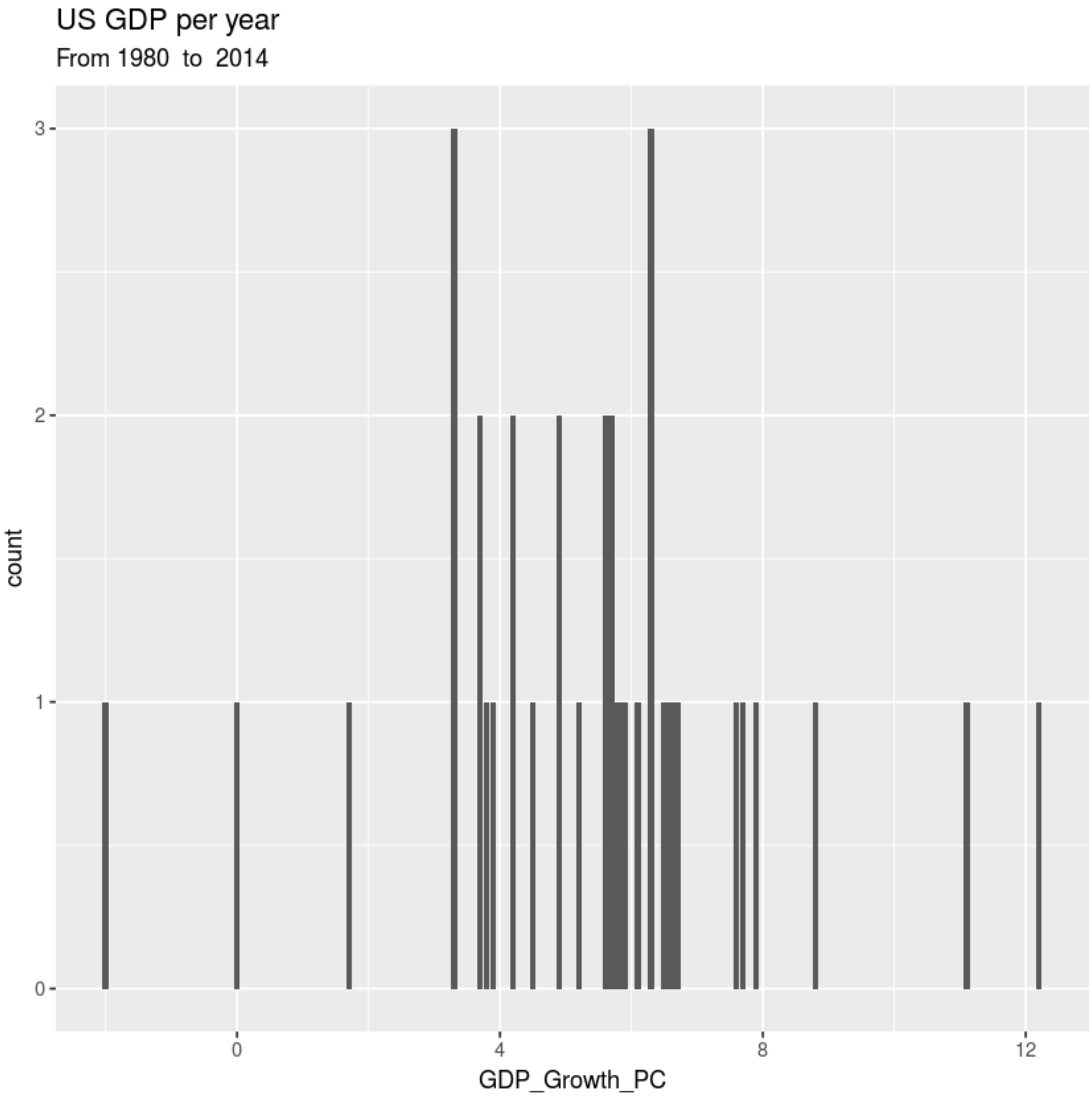


```
In [61]: ggplot(us_gdp, # start with the data
                aes(    # indicate which columns should be used where in the graph
           x=Year,
           y=US_GDP_BN,
           size=GDP_Growth_PC
       )
      ) +
  geom_point()+ # (add `+ geom_line()`)use a line + points
labs( # specify labels
    title="US GDP per year",
    subtitle=paste("From",min(us_gdp$Year)," to ",max(us_gdp$Year),collapse="")
) +
ylab("GDP")
```



2.0.1 Bar Plot - Histogram

```
In [37]: ggplot(us_gdp, # start with the data
                aes(    # indicate which columns should be used where in the graph
                    x=GDP_Growth_PC
                )
              ) +
         geom_bar() + # use bars i.e. a barplot / histogram
         labs( # specify labels
             title="US GDP per year",
             subtitle=paste("From",min(us_gdp$Year)," to ",max(us_gdp$Year),collapse="")
         )
```



```
In [36]: # What did we really just visualize?
```

```
         table(us_gdp$GDP_Growth_PC) #shows counts of value
```

```
  -2    0  1.7  3.3  3.7  3.8  3.9  4.2  4.5  4.9  5.2  5.6  5.7  5.8  5.9  6.1
   1    1    1    3    2    1    1    2    1    2    1    2    2    1    1    1
 6.3  6.5  6.6  6.7  7.6  7.7  7.9  8.8 11.1 12.2
   3    1    1    1    1    1    1    1    1    1
```

```
In [41]: # Import the library to assist with binning
         library(OneR) # if you receive and error that the library is not available run `install
```

```
In [45]: GDP_Growth_PC_binned = bin(us_gdp$GDP_Growth_PC,nbins=5)#labels parameter assigns group
         GDP_Growth_PC_binned
```

```
   1. (-2.01,0.84] 2. (9.36,12.2] 3. (3.68,6.52] 4. (6.52,9.36] 5. (9.36,12.2] 6. (6.52,9.36] 7. (3.68,6.52]
8. (3.68,6.52] 9. (6.52,9.36] 10. (6.52,9.36] 11. (3.68,6.52] 12. (0.84,3.68] 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. (3.68,6.52]
22. (0.84,3.68] 23. (0.84,3.68] 24. (3.68,6.52] 25. (6.52,9.36] 26. (6.52,9.36] 27. (3.68,6.52] 28. (3.68,6.52]
29. (0.84,3.68] 30. (-2.01,0.84] 31. (3.68,6.52] 32. (3.68,6.52] 33. (3.68,6.52] 34. (3.68,6.52] 35. (3.68,6.52]
   Levels: 1. '(-2.01,0.84]' 2. '(0.84,3.68]' 3. '(3.68,6.52]' 4. '(6.52,9.36]' 5. '(9.36,12.2]'
```

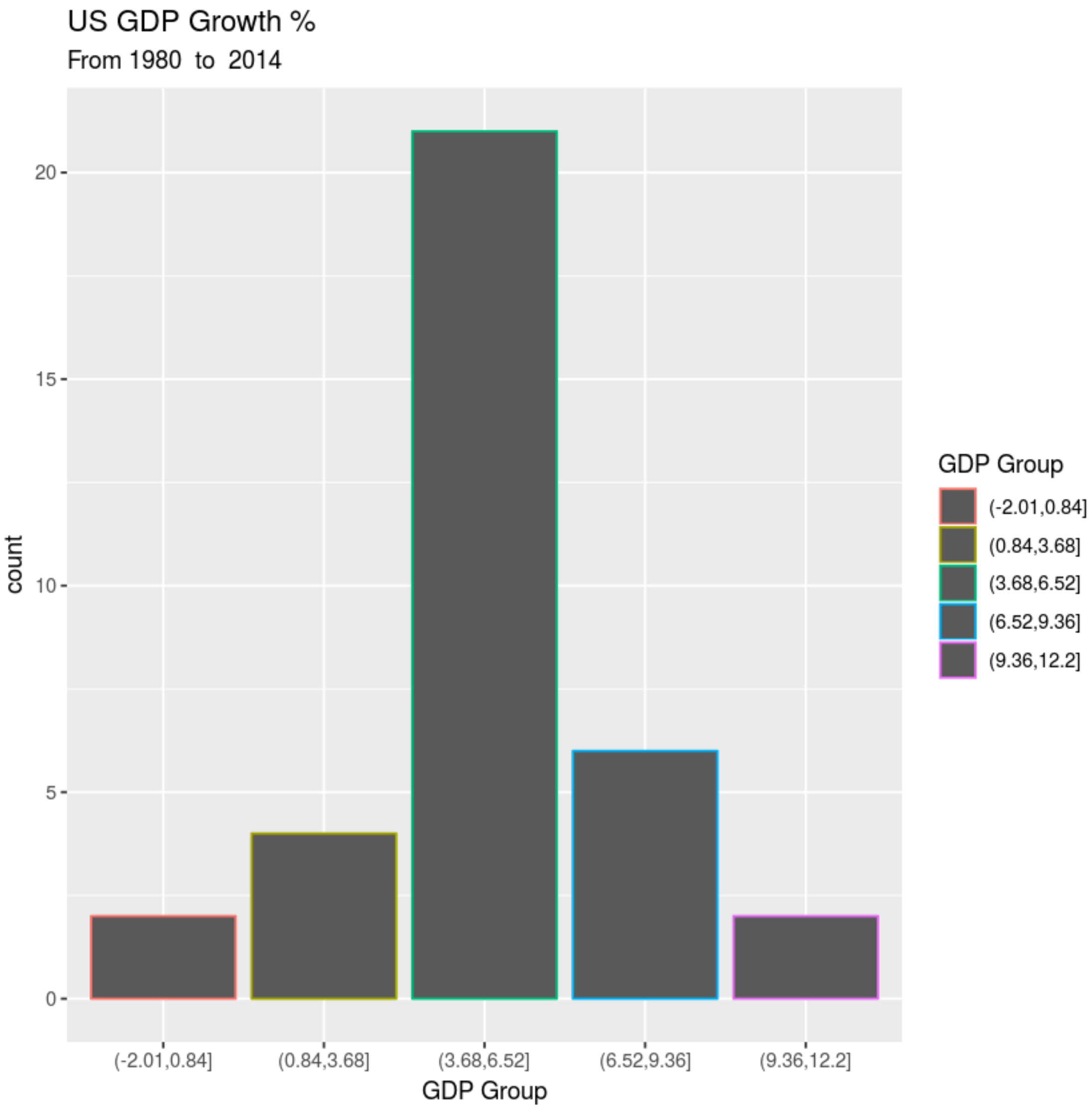
```
In [46]: table(GDP_Growth_PC_binned)
```

```
GDP_Growth_PC_binned
(-2.01,0.84]  (0.84,3.68]  (3.68,6.52]  (6.52,9.36]  (9.36,12.2]
           2            4           21            6            2
```

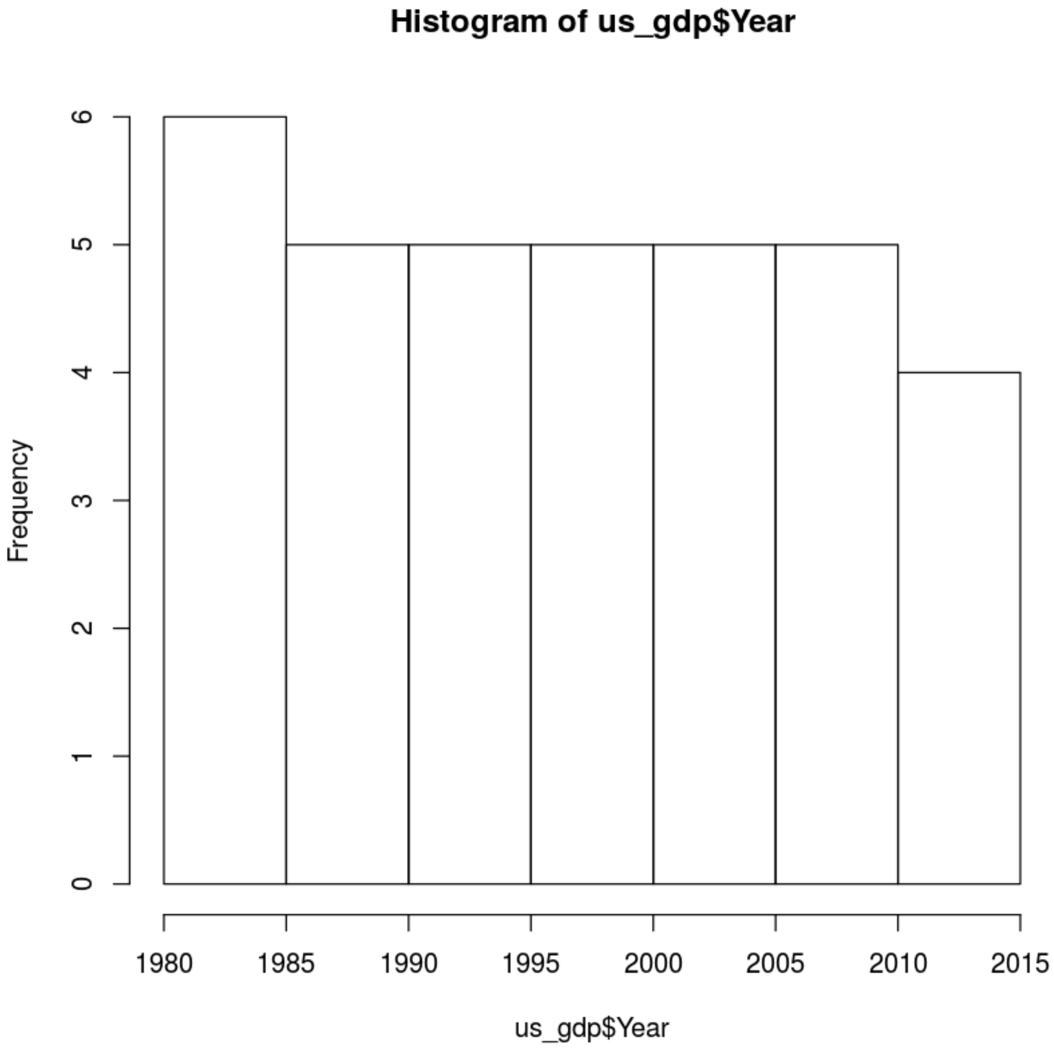
```
In [53]: GDP_Growth_PC_binned_df = data.frame('GDP Group'=GDP_Growth_PC_binned,check.names=FALSE
         GDP_Growth_PC_binned_df
```

GDP Group
(-2.01,0.84]
(9.36,12.2]
(3.68,6.52]
(6.52,9.36]
(9.36,12.2]
(6.52,9.36]
(3.68,6.52]
(3.68,6.52]
(6.52,9.36]
(6.52,9.36]
(3.68,6.52]
(0.84,3.68]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(0.84,3.68]
(0.84,3.68]
(3.68,6.52]
(6.52,9.36]
(6.52,9.36]
(3.68,6.52]
(3.68,6.52]
(0.84,3.68]
(-2.01,0.84]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]
(3.68,6.52]

```
In [74]: ggplot(GDP_Growth_PC_binned_df, # start with the data
                aes(    # indicate which columns should be used where in the graph
                    x=`GDP Group`,
                    color=`GDP Group`
                )
               ) +
         geom_bar() + # use bars i.e. a barplot / histogram
         labs( # specify labels
             title="US GDP Growth %",
             subtitle=paste("From",min(us_gdp$Year)," to ",max(us_gdp$Year),collapse="")
         )
```

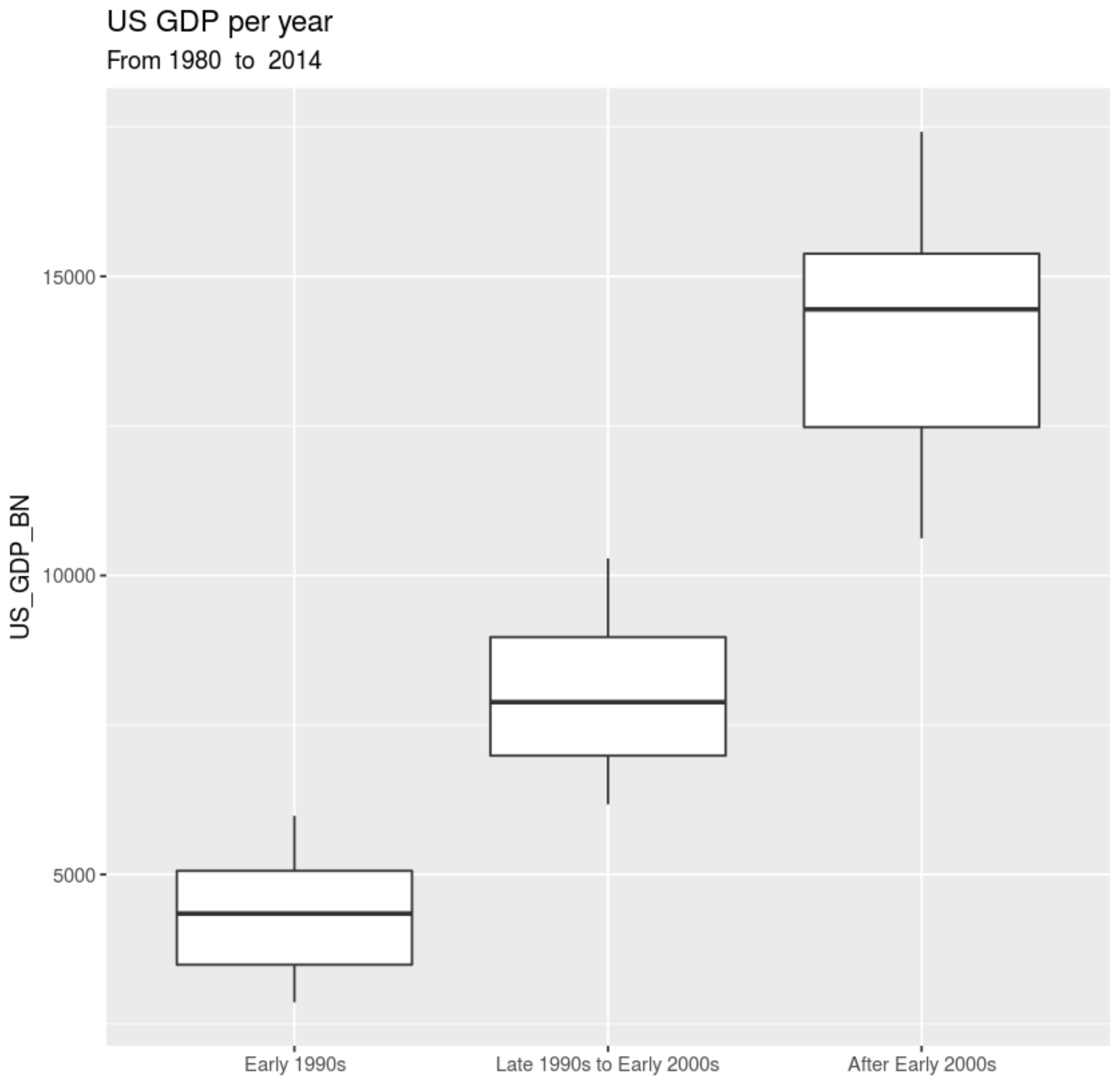


```
In [67]: hist(us_gdp$Year)
```

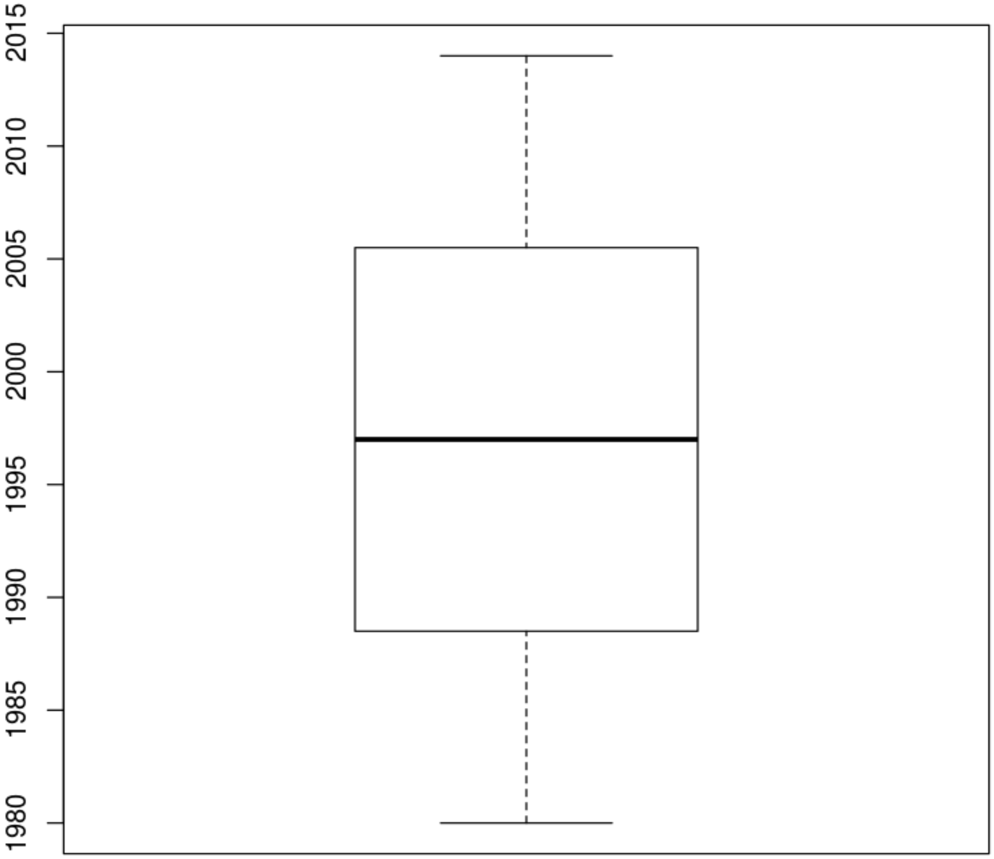


2.0.2 Box plot

```
In [79]: ggplot(us_gdp, # start with the data
                aes(    # indicate which columns should be used where in the graph
                    x=bin(Year,nbins=3,labels=c('Early 1990s','Late 1990s to Early 2000s','After
                    y=US_GDP_BN
                )
               ) +
         geom_boxplot() + # use bars i.e. a barplot / histogram
         labs( # specify labels
             title="US GDP per year",
             subtitle=paste("From",min(us_gdp$Year)," to ",max(us_gdp$Year),collapse="")
         )+xlab("")
```



```
In [68]: boxplot(us_gdp$Year)
```



2.0.3 Other Available Options and Types of Plot

```
ggplot2::geom_abline    Reference lines: horizontal, vertical, and
diagonal   Aliases: geom_abline, geom_hline, geom_vline ggplot2::geom_bar       Bar
charts   Aliases: geom_bar, geom_col ggplot2::geom_bin2d     Heatmap of 2d bin
counts   Aliases: geom_bin2d ggplot2::geom_blank     Draw nothing   Aliases:
geom_blank ggplot2::geom_boxplot   A box and whiskers plot (in the style of
Tukey)   Aliases: geom_boxplot ggplot2::geom_contour   2d contours of a 3d surface
Aliases: geom_contour ggplot2::geom_count     Count overlapping points   Aliases:
geom_count ggplot2::geom_crossbar                        Vertical intervals:
lines, crossbars &                         errorbars   Aliases: geom_crossbar,
geom_errorbar, geom_linerange,     geom_pointrange ggplot2::geom_density
Smoothed density estimates   Aliases: geom_density ggplot2::geom_density_2d
```

```
Contours of a 2d density estimate   Aliases: geom_density_2d, geom_density2d
ggplot2::geom_dotplot   Dot plot   Aliases: geom_dotplot ggplot2::geom_errorbarh
Horizontal error bars   Aliases: geom_errorbarh ggplot2::geom_freqpoly
Histograms and frequency polygons   Aliases: geom_freqpoly, geom_histogram
ggplot2::geom_hex       Hexagonal heatmap of 2d bin counts   Aliases: geom_hex
ggplot2::geom_jitter    Jittered points   Aliases: geom_jitter ggplot2::geom_label
Text   Aliases: geom_label, geom_text ggplot2::geom_map       Polygons from a
reference map   Aliases: geom_map ggplot2::geom_path      Connect observations
Aliases: geom_path, geom_line, geom_step ggplot2::geom_point     Points
Aliases: geom_point ggplot2::geom_polygon   Polygons   Aliases: geom_polygon
ggplot2::geom_qq_line   A quantile-quantile plot   Aliases: geom_qq_line, geom_qq
ggplot2::geom_quantile                         Quantile regression   Aliases:
geom_quantile ggplot2::geom_raster    Rectangles   Aliases: geom_raster, geom_rect,
geom_tile ggplot2::geom_ribbon    Ribbons and area plots   Aliases: geom_ribbon,
geom_area ggplot2::geom_rug       Rug plots in the margins   Aliases: geom_rug
ggplot2::geom_segment   Line segments and curves   Aliases: geom_segment,
geom_curve ggplot2::geom_smooth    Smoothed conditional means   Aliases:
geom_smooth ggplot2::geom_spoke     Line segments parameterised by location,
direction and distance   Aliases: geom_spoke ggplot2::geom_violin    Violin plot
Aliases: geom_violin
```