**LANGUAGE USED:**

R

**DATA:**

*vgsales.csv*

****

**INTRODUCTION**

The following has data analysis and visualization of Video Game’s sales across various parts of the world based on various criteria such as year, platform etc. Exploratory data analysis performed using various data wrangling packages such a dplyr, tidyr. Also data visualization performed using ggplot, plotly.

**READ DATA**

vgsales <- read.csv("C:/Username /Video game sales/vgsales.csv")

View(vgsales)

library(dplyr)

library(tidyr)

library(ggplot2)

library(gridExtra)

library(xlsx)

**DATA WRANGLING**

The following has the percentage of global sales:

NA\_net\_sales- Percentage of North America sales

EU\_net\_sales- Percentage of sales in Europe

JP\_net\_sales- Percentage of sales in Japan

other\_net\_sales- Percentage of sales in rest of the world

vgs <- mutate(vgs, NA\_net\_sales= (NA\_Sales\*100)/ Global\_Sales)

vgs <- mutate(vgs, EU\_net\_sales= (EU\_Sales\*100)/ Global\_Sales)

vgs <- mutate(vgs, JP\_net\_sales= (JP\_Sales\*100)/ Global\_Sales)

vgs <- mutate(vgs, other\_net\_sales= (Other\_Sales\*100)/ Global\_Sales)



**EXPLORATORY DATA ANALYSIS**

***Summary statistics***

Overall summary statistics of video game sales

summary\_video\_game\_sales<-summary(vgs)

X Rank Name Platform

Min. : 1 Min. : 1 Need for Speed: Most Wanted: 12 DS :2163

1st Qu.: 4150 1st Qu.: 4151 FIFA 14 : 9 PS2 :2161

Median : 8300 Median : 8300 LEGO Marvel Super Heroes : 9 PS3 :1329

Mean : 8300 Mean : 8301 Madden NFL 07 : 9 Wii :1325

3rd Qu.:12449 3rd Qu.:12450 Ratatouille : 9 X360 :1265

Max. :16598 Max. :16600 Angry Birds Star Wars : 8 PSP :1213

(Other) :16542 (Other):7142

Year Genre Publisher

2009 :1431 Action :3316 Electronic Arts : 1351

2008 :1428 Sports :2346 Activision : 975

2010 :1259 Misc :1739 Namco Bandai Games : 932

2007 :1202 Role-Playing:1488 Ubisoft : 921

2011 :1139 Shooter :1310 Konami Digital Entertainment: 832

2006 :1008 Adventure :1286 THQ : 715

(Other):9131 (Other) :5113 (Other) :10872

NA\_Sales EU\_Sales JP\_Sales Other\_Sales

Min. : 0.0000 Min. : 0.0000 Min. : 0.00000 Min. : 0.00000

1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.00000 1st Qu.: 0.00000

Median : 0.0800 Median : 0.0200 Median : 0.00000 Median : 0.01000

Mean : 0.2647 Mean : 0.1467 Mean : 0.07778 Mean : 0.04806

3rd Qu.: 0.2400 3rd Qu.: 0.1100 3rd Qu.: 0.04000 3rd Qu.: 0.04000

Max. :41.4900 Max. :29.0200 Max. :10.22000 Max. :10.57000

Global\_Sales NA\_net\_sales EU\_net\_sales JP\_net\_sales

Min. : 0.0100 Min. : 0.00 Min. : 0.00 Min. : 0.00

1st Qu.: 0.0600 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00

Median : 0.1700 Median : 50.00 Median : 20.00 Median : 0.00

Mean : 0.5374 Mean : 45.55 Mean : 22.92 Mean : 24.23

3rd Qu.: 0.4700 3rd Qu.: 75.00 3rd Qu.: 37.50 3rd Qu.: 35.10

Max. :82.7400 Max. :100.00 Max. :100.00 Max. :100.00

other\_net\_sales

Min. : 0.000

1st Qu.: 0.000

Median : 5.556

Mean : 6.459

3rd Qu.: 10.253

Max. :100.000

***Summary statistics based on Platform, Publisher, Genre, Year***

Mean global sales based on platform, publisher, genre and year.

*Global Sales based on Platform*

summary\_GSales\_Platform<- summarise(group\_by(vgs, Platform), mean(Global\_Sales))



Mean global sales based on platform = 0.5762

*Global Sales based on Publisher*

summary\_GSales\_Publisher<- summarise(group\_by(vgs, Publisher), mean(Global\_Sales))



*Global Sales based on Genre*

summary\_GSales\_Genre<- summarise(group\_by(vgs, Genre), mean(Global\_Sales))



*Global Sales based on Year*

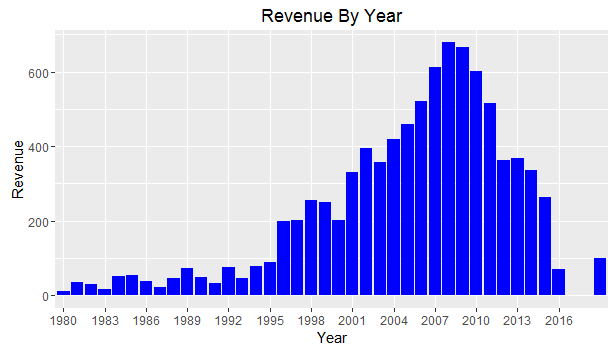
summary\_GSales\_Year<- summarise(group\_by(vgs, Year), mean(Global\_Sales))



**Revenue based on Year**

Revenue\_By\_Year<- vgs %>% group\_by(Year) %>% summarise(Revenue= sum(Global\_Sales))

ggplot(Revenue\_By\_Year, aes(Year, Revenue) + geom\_bar(fill=”Blue”, stat=”identity”) + ggtitle(“Revenue By Year”) + scale\_x\_discrete(breaks(1980,2016,3))

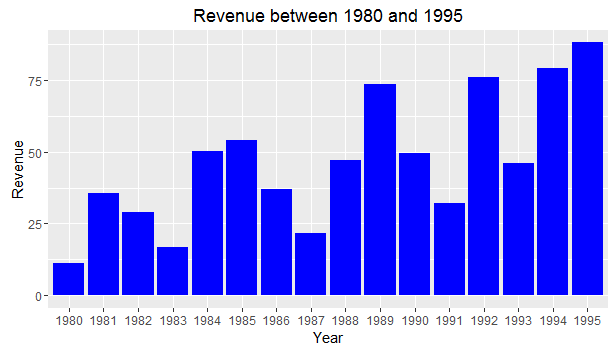
****

So, 1997/98 is the year, where the revenue became 4 times as big as previous years. Also, late 2000s marks the further increase in revenue. Following that, we are analyzing the reason for the sudden increase in revenue. The time period (1980 to 2016) is subset into four different time periods. This will help us analyze the real growth in sales of video games over the years.

*Revenue of video games between 1980 and 1995*

Revenue\_1980\_1995<- subset(Revenue\_by\_YNG, Year %in% c("1980": "1995"))

p1<- ggplot(Revenue\_1980\_1995, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 1980 and 1995") + ylim(0,800)

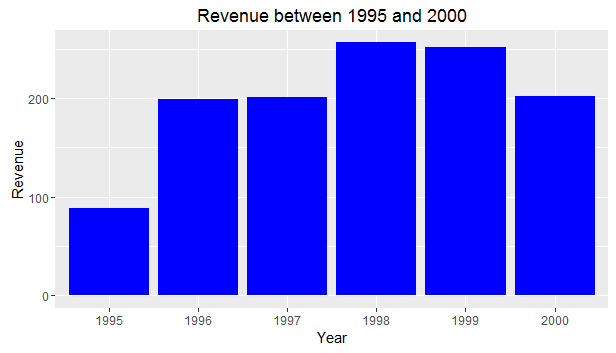


There is a steady and significant improvement between 1980 and 1995 with 1995 being the peak of sales.

*Revenue of video games between 1995 and 2000*

Revenue\_1995\_2000<- subset(Revenue\_by\_YNG, Year %in% c("1995": "2000"))

p2<- ggplot(Revenue\_1995\_2000, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 1995 and 2000") + ylim(0,800)

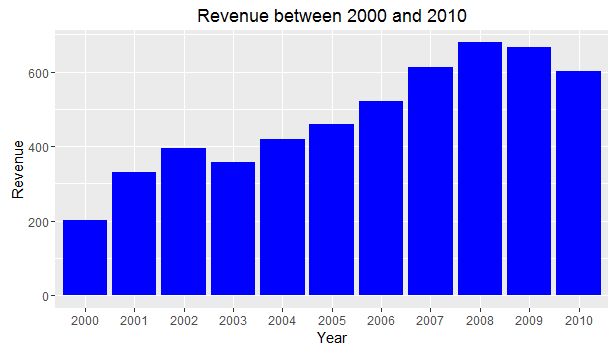


There is a significant level of improvement in sales across various genres of games between 1995 and 2000. Between 1995 and 2000, the discovery of computer systems and internet service has definitely played a huge role in the sales.

*Revenue of video games between 2000 and 2010:*

Revenue\_2000\_2010<- subset(Revenue\_by\_YNG, Year %in% c("2000": "2010"))

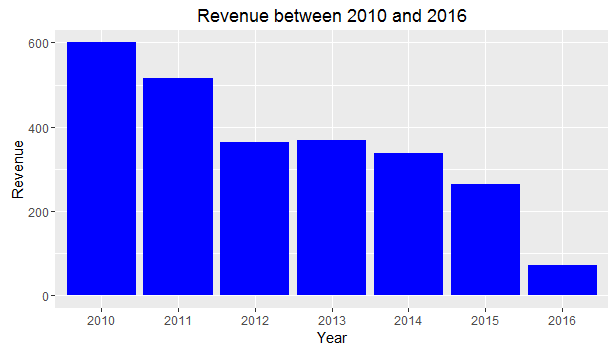
p3<- ggplot(Revenue\_2000\_2010, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 2000 and 2010") + ylim(0,800)



*Revenue of video games between 2010 and 2016*

Revenue\_2010\_2016<- subset(Revenue\_by\_YNG, Year %in% c("2010": "2016"))

p4<- ggplot(Revenue\_2010\_2016, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 2010 and 2016") + ylim(0,800)



*Overall Growth in Sales between 1980 and 2016*

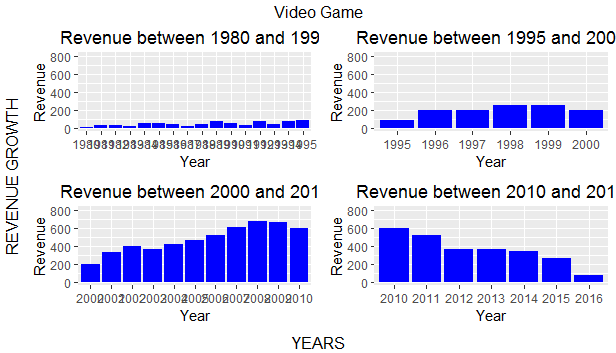
p1<- ggplot(Revenue\_1980\_1995, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 1980 and 1995") + ylim(0,800)

p2<- ggplot(Revenue\_1995\_2000, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 1995 and 2000") + ylim(0,800)

p4<- ggplot(Revenue\_2010\_2016, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 2010 and 2016") + ylim(0,800)

p3<- ggplot(Revenue\_2000\_2010, aes(Year,Revenue)) + geom\_bar(fill= "Blue", stat= "identity") + ggtitle("Revenue between 2000 and 2010") + ylim(0,800)

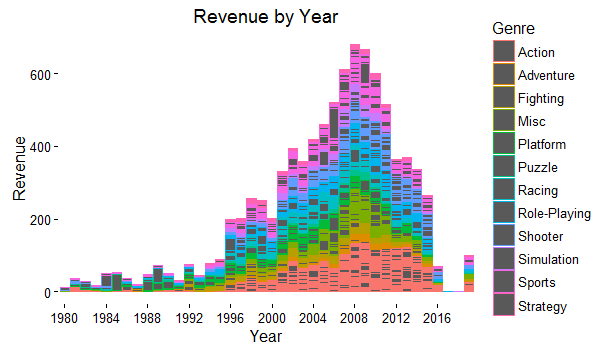
grid.arrange(p1,p2,p3,p4,top= "Video Game", left= "REVENUE GROWTH", bottom= "YEARS")



*Revenue by Year based on Genre*

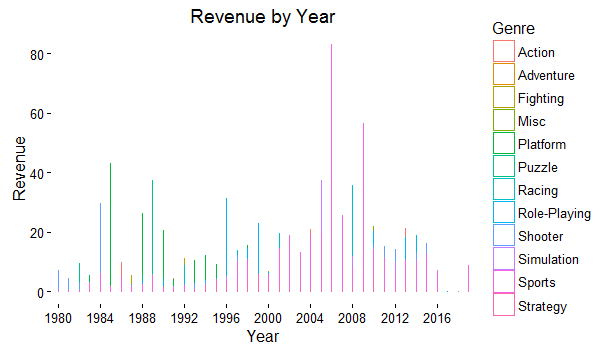
*Bar graph:*

ggplot(Revenue\_By\_Year, aes(Year, Revenue, na.omit(TRUE))) + geom\_bar(stat = "identity", aes(color= Genre)) + ggtitle("Revenue by Year") + scale\_x\_discrete(breaks=seq(1980,2016,4)) + theme\_classic()



*Density:*

ggplot(Revenue\_By\_Year, aes(Year, Revenue, na.omit("TRUE"))) + geom\_density(stat = "identity", aes(color= Genre)) + ggtitle("Revenue by Year") + scale\_x\_discrete(breaks=seq(1980,2016,4)) + theme\_classic()



*Jitter:*

ggplot(Revenue\_By\_Year, aes(Year, Revenue, na.omit("TRUE"))) + geom\_jitter(stat = "identity", aes(color= Genre)) + ggtitle("Revenue by Year") + scale\_x\_discrete(breaks=seq(1980,2016,4)) + theme\_classic()

