Final Project

Video Game Sales

Nam Jun Lee

Jun 16, 2021

Contents

1 Big Question

- This dataset provides a quick look at video games that we use as our hobbies in our daily lives. I have played video games from many companies and like video games, so I decided to analyze this dataset.
- The **Big Question** I analyze is using *Video game sales from 1980 to 2020* to figure out **what accounts for a large portion of total sales in each variables (Game, Publisher, Genre, Platform, Year, Each Country Sales)**.
- This file contains sales volume and total sales volume for each country, so it is a good choice to answer Big Question.
- Further analysis will be done to analyze the *Top 10 Platforms, Publishers and Games that affect sales.* And also analyze *which year has the highest and lowest sales*, and I will check the *trend of video games.* Finally, I will find the *Top 3 best-selling games in the year when sales were highest.*

2 Dataset Information

- This dataset was collected by vgchartz.com scraps and contains a list of video games with sales of more than 100,000 units. The detailed list includes *overall rankings, game names, platforms, years, genres, game publishers, sales in North America, sales in Europe, sales in Japan, sales in the rest of the world, and total sales worldwide.*
- There are **16,598** records and **2** records were dropped due to incomplete information.

3 Data Collection

3.0.1 Load the vgsales.csv data into R

```
# read Video game sales csv file
VGS <- read.csv("./vgsales.csv", encoding = 'UTF-8')

# head for VGS file
head(VGS)
```

```
##   Rank                     Name Platform Year         Genre Publisher NA_Sales
## 1    1               Wii Sports      Wii 2006        Sports  Nintendo    41.49
## 2    2        Super Mario Bros.      NES 1985      Platform  Nintendo    29.08
## 3    3           Mario Kart Wii      Wii 2008        Racing  Nintendo    15.85
## 4    4        Wii Sports Resort      Wii 2009        Sports  Nintendo    15.75
## 5    5 Pokemon Red/Pokemon Blue       GB 1996 Role-Playing  Nintendo    11.27
## 6    6                   Tetris       GB 1989        Puzzle  Nintendo    23.20
##   EU_Sales JP_Sales Other_Sales Global_Sales
## 1    29.02     3.77        8.46        82.74
## 2     3.58     6.81        0.77        40.24
## 3    12.88     3.79        3.31        35.82
## 4    11.01     3.28        2.96        33.00
## 5     8.89    10.22        1.00        31.37
## 6     2.26     4.22        0.58        30.26
```

Load the video game sales data needed for data analysis and obtain rows for the first six to see what is there.

3.0.2 Data information & Tukey's five point summary

```
# View the data information
str(VGS)
```

```
## 'data.frame':    16598 obs. of  11 variables:
##  $ Rank        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name        : chr  "Wii Sports" "Super Mario Bros." "Mario Kart Wii" "Wii Sports Resort" ...
##  $ Platform    : chr  "Wii" "NES" "Wii" "Wii" ...
##  $ Year        : chr  "2006" "1985" "2008" "2009" ...
##  $ Genre       : chr  "Sports" "Platform" "Racing" "Sports" ...
##  $ Publisher   : chr  "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...
##  $ NA_Sales    : num  41.5 29.1 15.8 15.8 11.3 ...
##  $ EU_Sales    : num  29.02 3.58 12.88 11.01 8.89 ...
##  $ JP_Sales    : num  3.77 6.81 3.79 3.28 10.22 ...
##  $ Other_Sales : num  8.46 0.77 3.31 2.96 1 0.58 2.9 2.85 2.26 0.47 ...
##  $ Global_Sales: num  82.7 40.2 35.8 33 31.4 ...
```

```
# Tukey's five point summary for each of the variables
summary(VGS)
```

```
##       Rank           Name              Platform            Year          
##  Min.   :    1   Length:16598       Length:16598       Length:16598      
##  1st Qu.: 4151   Class :character   Class :character   Class :character  
##  Median : 8300   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 8301                                                           
##  3rd Qu.:12450                                                           
##  Max.   :16600                                                           
##     Genre            Publisher            NA_Sales          EU_Sales       
##  Length:16598       Length:16598       Min.   : 0.0000   Min.   : 0.0000  
##  Class :character   Class :character   1st Qu.: 0.0000   1st Qu.: 0.0000  
##  Mode  :character   Mode  :character   Median : 0.0800   Median : 0.0200  
##                                        Mean   : 0.2647   Mean   : 0.1467  
##                                        3rd Qu.: 0.2400   3rd Qu.: 0.1100  
##                                        Max.   :41.4900   Max.   :29.0200  
##     JP_Sales         Other_Sales       Global_Sales     
##  Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.0100  
##  1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.: 0.0600  
##  Median : 0.00000   Median : 0.01000   Median : 0.1700  
##  Mean   : 0.07778   Mean   : 0.04806   Mean   : 0.5374  
##  3rd Qu.: 0.04000   3rd Qu.: 0.04000   3rd Qu.: 0.4700  
##  Max.   :10.22000   Max.   :10.57000   Max.   :82.7400
```

This dataset has a total of *16598* data, *11* variables, and contains data frame properties. Some variables are character variables, making it difficult to see the results in the summary command.

4 Data Preprocessing

4.0.1 Data Preprocessing [Cleaning data (missing value)]

```
# Find missing values in VGS file
VGS[VGS=='N/A']<-NA

# Show sum of all variables NA
colSums(is.na(VGS))
```

```
##         Rank         Name     Platform         Year        Genre    Publisher
##            0            0            0          271            0           58
##     NA_Sales     EU_Sales     JP_Sales  Other_Sales Global_Sales
##            0            0            0            0            0
```

```
# remove missing values in VGS
VGS <- na.omit(VGS)

# remove no need col (Rank)
VGS <- VGS[,-1]

#change data types
VGS$Year <- as.numeric(as.character(VGS$Year))
VGS$Genre <- as.factor(as.character(VGS$Genre))
VGS$Platform <- as.factor(as.character(VGS$Platform))
VGS$Name <- as.factor(as.character(VGS$Name))
VGS$Publisher <- as.factor(as.character(VGS$Publisher))
```

After checking the missing values in the data, found that there were *271* missing values in the **Year** variable and *58* missing values in the **Publisher** variable.

Before the analysis, the *missing values* were removed from the dataset and the *types of variables* were transformed. Also remove the **Rank** variable because it is a column that is not required for analysis.

5 Exploratory Data

5.0.1 Recall fixed data information & Tukey's five point summary

```
# View the fixed data information
str(VGS)
```

```
## 'data.frame':    16291 obs. of  10 variables:
##  $ Name        : Factor w/ 11325 levels "'98 Koshien",..: 10831 9216 5451 10833 7264 9572 6558 10829
##  $ Platform    : Factor w/ 31 levels "2600","3DO","3DS",..: 26 12 26 26 6 6 5 26 26 12 ...
##  $ Year        : num  2006 1985 2008 2009 1996 ...
##  $ Genre       : Factor w/ 12 levels "Action","Adventure",..: 11 5 7 11 8 6 5 4 5 9 ...
##  $ Publisher   : Factor w/ 576 levels "10TACLE Studios",..: 368 368 368 368 368 368 368 368 368 368
##  $ NA_Sales    : num  41.5 29.1 15.8 15.8 11.3 ...
##  $ EU_Sales    : num  29.02 3.58 12.88 11.01 8.89 ...
##  $ JP_Sales    : num  3.77 6.81 3.79 3.28 10.22 ...
##  $ Other_Sales : num  8.46 0.77 3.31 2.96 1 0.58 2.9 2.85 2.26 0.47 ...
##  $ Global_Sales: num  82.7 40.2 35.8 33 31.4 ...
```

```
# Tukey's five point summary for each of the fixed variables
summary(VGS)
```

```
##                          Name          Platform         Year
##  Need for Speed: Most Wanted:   12   DS     :2131   Min.   :1980
##  FIFA 14                    :    9   PS2    :2127   1st Qu.:2003
##  LEGO Marvel Super Heroes   :    9   PS3    :1304   Median :2007
##  Ratatouille                :    9   Wii    :1290   Mean   :2006
##  Angry Birds Star Wars      :    8   X360   :1234   3rd Qu.:2010
##  Cars                       :    8   PSP    :1197   Max.   :2020
##  (Other)                    :16236   (Other):7008
##           Genre                              Publisher        NA_Sales
##  Action      :3251   Electronic Arts             : 1339   Min.   : 0.0000
##  Sports      :2304   Activision                  :  966   1st Qu.: 0.0000
##  Misc        :1686   Namco Bandai Games          :  928   Median : 0.0800
##  Role-Playing:1470   Ubisoft                     :  918   Mean   : 0.2656
##  Shooter     :1282   Konami Digital Entertainment:  823   3rd Qu.: 0.2400
##  Adventure   :1274   THQ                         :  712   Max.   :41.4900
##  (Other)     :5024   (Other)                     :10605
##     EU_Sales          JP_Sales         Other_Sales        Global_Sales
##  Min.   : 0.0000   Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.0100
##  1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.: 0.0600
##  Median : 0.0200   Median : 0.00000   Median : 0.01000   Median : 0.1700
##  Mean   : 0.1477   Mean   : 0.07883   Mean   : 0.04843   Mean   : 0.5409
##  3rd Qu.: 0.1100   3rd Qu.: 0.04000   3rd Qu.: 0.04000   3rd Qu.: 0.4800
##  Max.   :29.0200   Max.   :10.22000   Max.   :10.57000   Max.   :82.7400
##
```

This dataset shows a data frame with a total of 16291 data and 10 variables. The summary command computes the values for each list in the variables Name, Platform, Year, Genre, and Publisher. It has a total of **11325** Game Name Levels, **31** Platform Levels, **12** Genre Levels, and **576** Publisher Levels. Additional,

summary command computes the NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales. Max is very high for the median. Through this, sales values are skewed. Also using outliers formula,

$$Observations < 25thPercentile - 1.5 * IQR$$

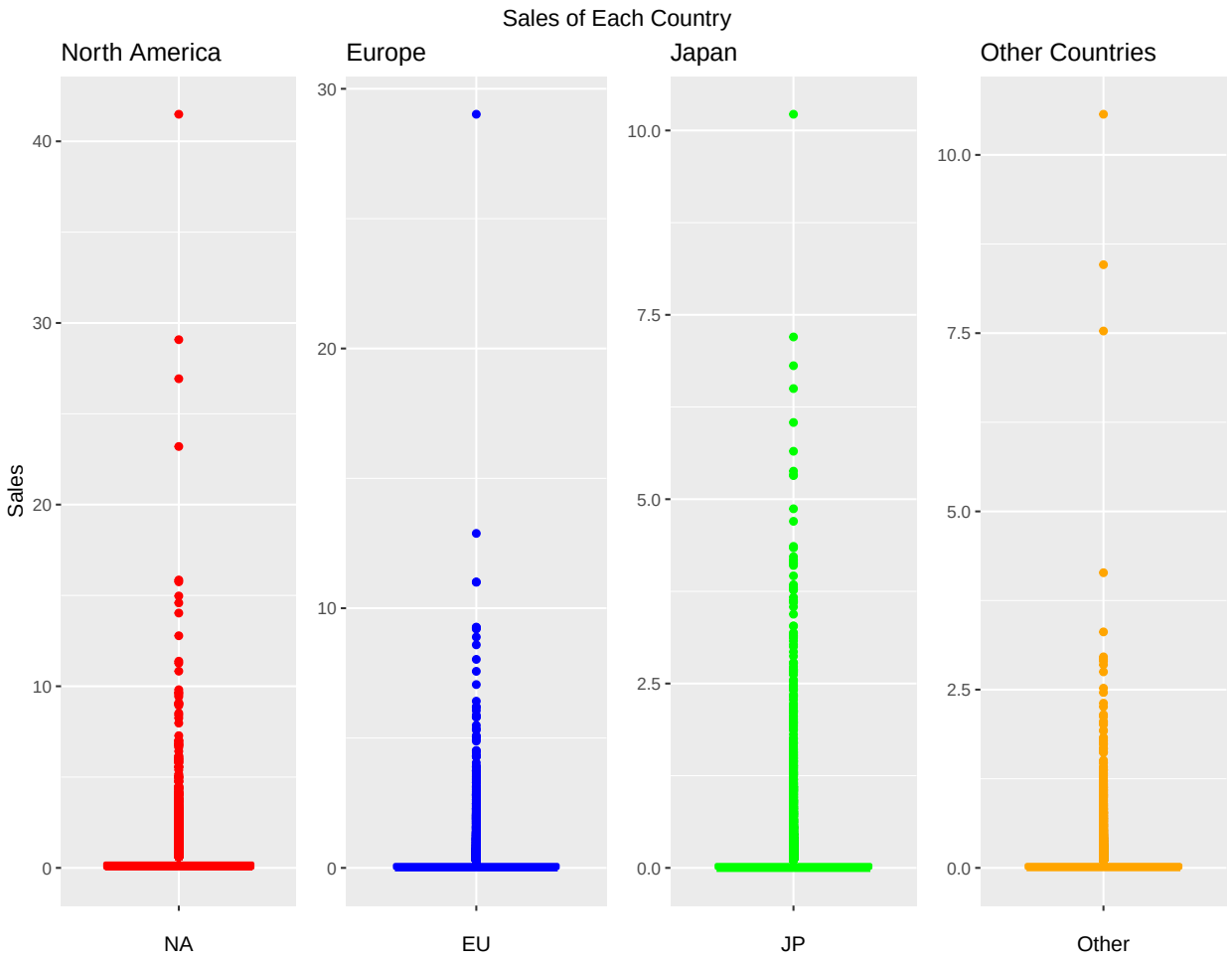
or

$$Observations > 75thPercentile + 1.5 * IQR$$

These Sales variables have outliers.

5.0.2 Each Country Sales Boxplot

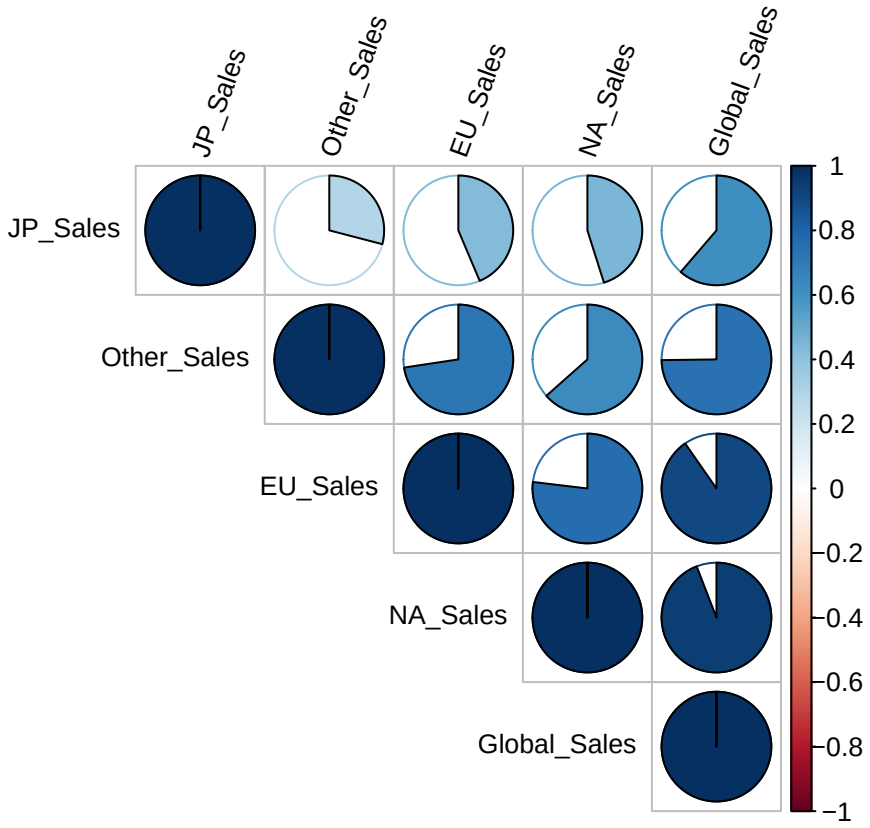
```
# North America Sales boxplot
nab <- ggplot(VGS, aes(x='',y=NA_Sales)) +
  geom_boxplot(color='red', lwd=1.0) +
  ggtitle('North America') +
  xlab('NA') +
  ylab('Sales')

# Europe Sales boxplot
eub <- ggplot(VGS, aes(x='',y=EU_Sales)) +
  geom_boxplot(color='blue', lwd=1.0) +
  ggtitle('Europe') +
  xlab('EU') +
  theme(axis.title.y = element_blank())

# Japan Sales boxplot
jpb <- ggplot(VGS, aes(x='',y=JP_Sales)) +
  geom_boxplot(color='green', lwd=1.0) +
  ggtitle('Japan') +
  xlab('JP') +
  theme(axis.title.y = element_blank())

# Other Sales boxplot
otb <- ggplot(VGS, aes(x='',y=Other_Sales)) +
  geom_boxplot(color='orange', lwd=1.0) +
  ggtitle('Other Countries') +
  xlab('Other') +
  theme(axis.title.y = element_blank())

# show four graphs on one page
grid.arrange(top='Sales of Each Country',nab,eub,jpb,otb, ncol=4)
```



Notice that the box plot of each country's sales, it can see that the data is *positively skewed*. It can also be confirmed that there is *outliers*. The outliers found here represent sales for each game. Sales can be small or large, so I would not remove it.

5.0.3 Correlation of All Sales Factor

```
# Show correlation
all_corr <- cor(VGS[,c(6:10)]) # Take the variables and tie them up

# visualize the strength of linear relationships between variables.(correlation matrix)
corrplot(all_corr, method='pie', type='upper',
         tl.col='black',
         tl.srt=70, tl.cex=0.8, order='hclust')
```



The correlation between NA_Sales value, EU_Sales value, JP_Sales value, and Other_Sales price was identified at a glance. Overall, the linear relationship to the *Global Sales* value is **all quantitative linear**. The *North America Sales* value shows a **very strong quantitative linear relationship** to the *Global Sales* value. Also, the *Japan Sales* value shows the **moderate quantitative linear relationship**. Also, that *American Sales* have a **lot of influence** on *Europe Sales* and *Other Sales* have **little influence** on *Japan Sales*.

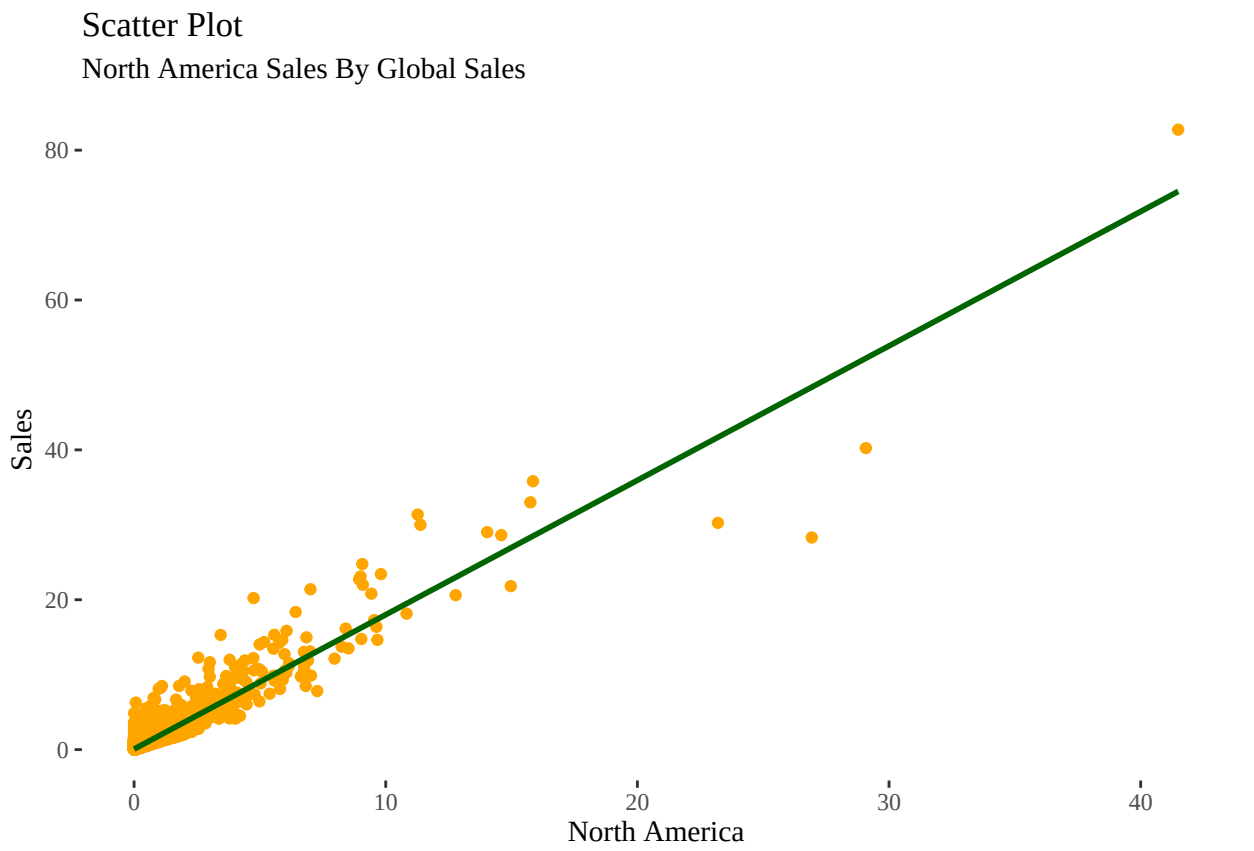
5.0.4 Linear Regression of North America Sales by Global Sales

```
# Regression
sales_na <- lm(data=VGS,Global_Sales~NA_Sales)
summary(sales_na)
```

```
##
## Call:
## lm(formula = Global_Sales ~ NA_Sales, data = VGS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0619  -0.1096  -0.0541   0.0177  11.6350
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.064389   0.004357   14.78   <2e-16 ***
## NA_Sales    1.793817   0.005042  355.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5292 on 16289 degrees of freedom
## Multiple R-squared:  0.886,  Adjusted R-squared:  0.886
## F-statistic: 1.266e+05 on 1 and 16289 DF,  p-value: < 2.2e-16
```

```
ggplot(sales_na, aes(NA_Sales,Global_Sales)) +
  geom_point(color='orange') +
  geom_smooth(method='lm', formula=y~x,color='darkgreen') +
  theme_tufte() +
  ggtitle('Scatter Plot', subtitle = 'North America Sales By Global Sales') +
  labs(x='North America',y='Sales')
```



Regression (strongest correlation among Sales) of *North America Sales values* to *Global Sales values*. Allows us to see that the estimate line has equation

$$Y = 0.064389 + 1.793817X$$

with coefficient of determination. This value, *lower than the p-value of 0.05*, is *statistically significant for the entire regression model*. *Coefficients* have y-intercept values and p-values for variables. The above * indicates that it is statistically significant at a glance, and the **North America Sales value is likely to**

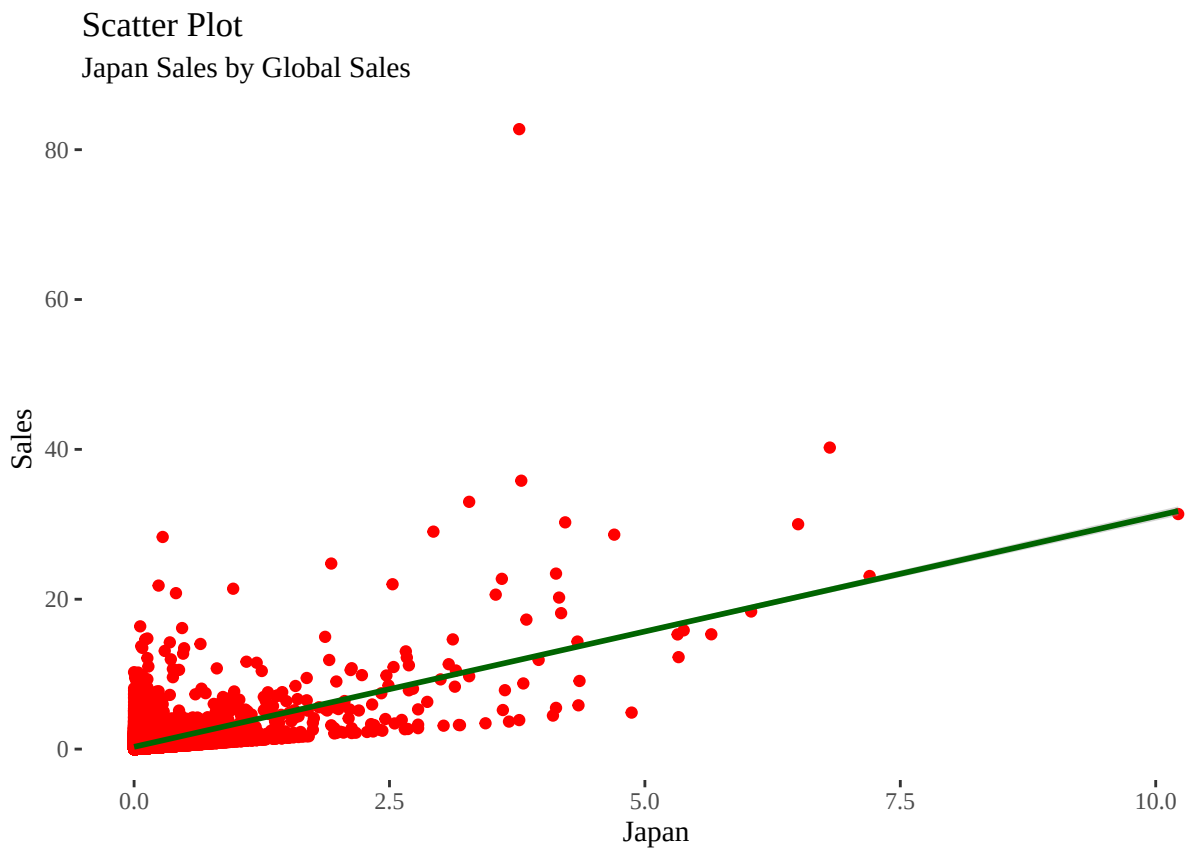
be statistically significant. It also indicates that the *adjusted R squared* value is *0.886* and the *explanatory power* is **88.6%**. This scatterplot shows that the data are positive skewed and very strongly linear (positive correlation). It also shows that there are outliers.

5.0.5 Linear Regression of Japan Sales by Global Sales

```
# Regression
sales_jp <- lm(data=VGS,Global_Sales~JP_Sales)
summary(sales_jp)
```

```
##
## Call:
## lm(formula = Global_Sales ~ JP_Sales, data = VGS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.425  -0.319  -0.198   0.062  70.832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.29814    0.01001   29.79   <2e-16 ***
## JP_Sales     3.07948    0.03112   98.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.239 on 16289 degrees of freedom
## Multiple R-squared:  0.3755, Adjusted R-squared:  0.3755
## F-statistic:  9794 on 1 and 16289 DF,  p-value: < 2.2e-16
```

```
ggplot(sales_jp, aes(JP_Sales,Global_Sales)) +
  geom_point(color='red') +
  geom_smooth(method='lm', formula=y~x,color='darkgreen') +
  theme_tufte() +
  ggtitle('Scatter Plot', subtitle = 'Japan Sales by Global Sales') +
  labs(x='Japan', y='Sales')
```



Regression (moderate correlation among Sales) of *Japan Sales values* to *Global Sales values*. Allows us to see that the estimate line has equation

$$Y = 0.29814 + 3.07948X$$

with coefficient of determination. This value, *lower than the p-value of 0.05*, is *statistically significant for the entire regression model*. *Coefficients* have y-intercept values and p-values for variables. The above * indicates that it is statistically significant at a glance, and the **Japan Sales value is likely to be statistically significant**. It also indicates that the *adjusted R squared* value is *0.3755* and the *explanatory power* is **37.55%**. This scatterplot shows that the data are positive skewed and strongly linear (positive correlation). It also shows that there are outliers.

5.0.6 Multiple linear Regression summary of All Sales by Global Sales

```
all_sales <- lm(data=VGS,Global_Sales~NA_Sales+EU_Sales+JP_Sales+Other_Sales)
summary(all_sales)
```

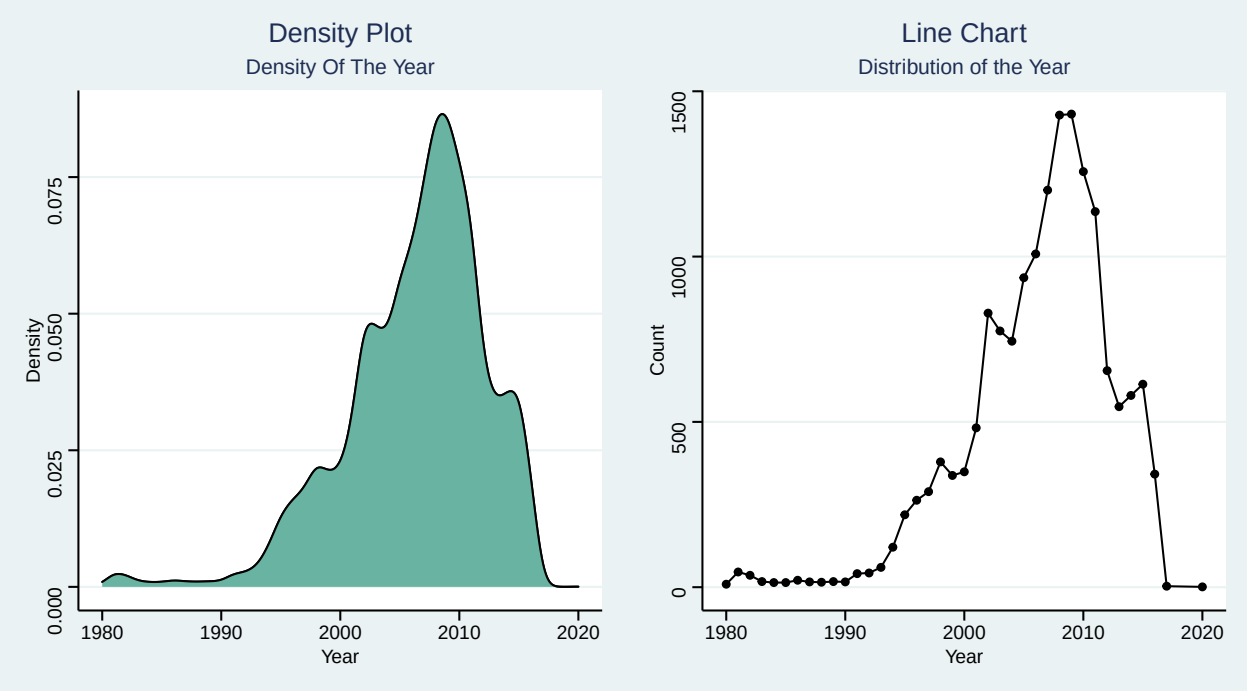
```
##
## Call:
## lm(formula = Global_Sales ~ NA_Sales + EU_Sales + JP_Sales +
##     Other_Sales, data = VGS)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0202835 -0.0003148 -0.0003044 -0.0002506  0.0200711
```

```
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 3.171e-04  4.337e-05     7.312 2.76e-13 ***
## NA_Sales    1.000e+00  8.087e-05 12365.241  < 2e-16 ***
## EU_Sales    1.000e+00  1.457e-04  6860.840  < 2e-16 ***
## JP_Sales    9.999e-01  1.494e-04  6693.685  < 2e-16 ***
## Other_Sales 9.996e-01  3.190e-04  3133.030  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005224 on 16286 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.666e+08 on 4 and 16286 DF,  p-value: < 2.2e-16
```

Multiple Regression of *All Sales values* to *Global Sales values*. This value, *lower than the p-value of 0.05*, is *statistically significant for the entire regression model*. *Coefficients* have y-intercept values and p-values for variables. The above * indicates that it is statistically significant at a glance, and the **All Sales value is likely to be statistically significant**. It also indicates that the *adjusted R squared* value is *1*. This means the model completely fit and explained all variance.

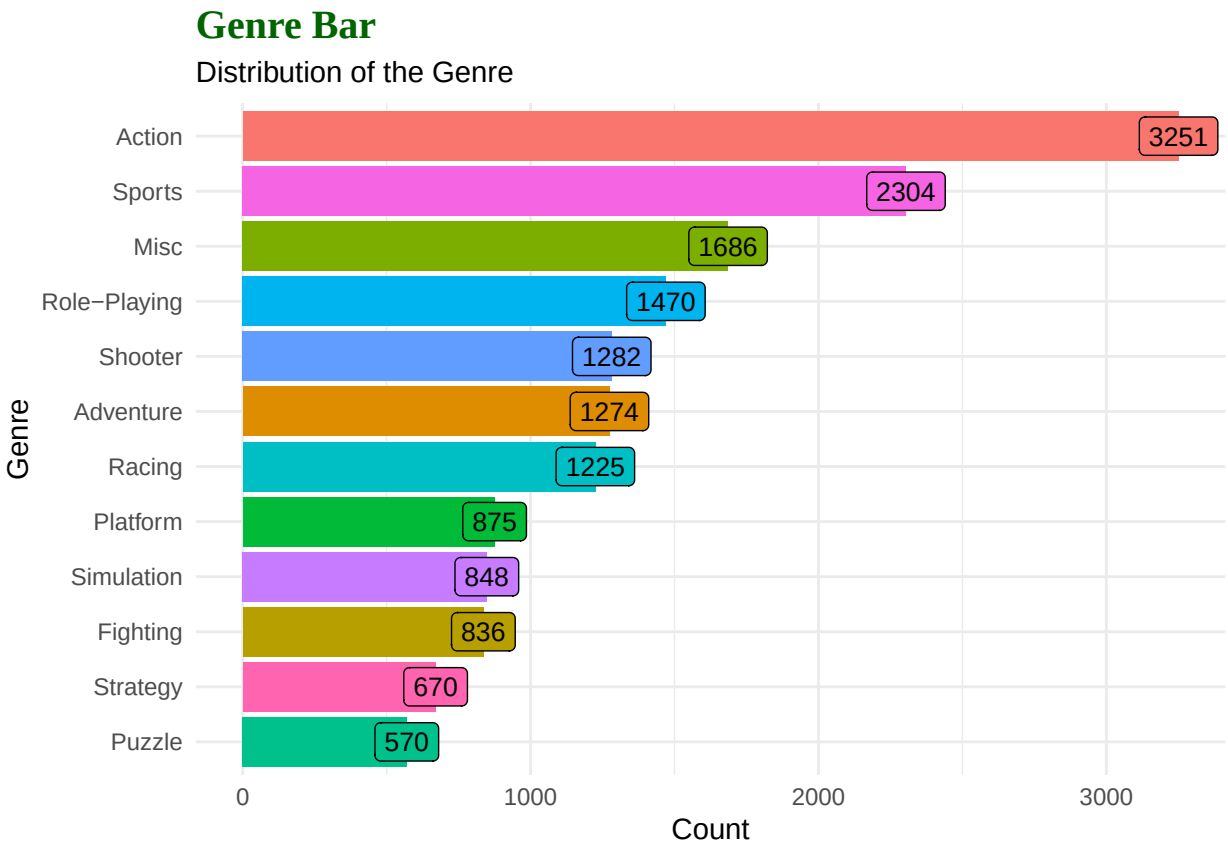
5.0.7 Distribution by Year & Genre

```
# density by year
dens_yr <- ggplot(VGS, aes(x = Year, y = ..density..)) +
  geom_density(color='black', size=0.5) +
  geom_density(fill='#69b3a2') +
  theme_stata() +
  scale_color_stata() +
  ggtitle('Density Plot', subtitle = 'Density Of The Year') +
  xlab('Year') +
  ylab('Density')

yr <- VGS %>%
  group_by(Year) %>%
  dplyr::summarise(count = n())

# distribution by year
dist_yr <- ggplot(yr, aes(x = Year, y = count)) +
  geom_line() +
  geom_point() +
  theme_stata() +
  scale_color_stata() +
  ggtitle('Line Chart', subtitle = 'Distribution of the Year') +
  xlab('Year') +
  ylab('Count')

grid.arrange(dens_yr,dist_yr, ncol=2) # show two graphs on one page
```



Notice that **density and line of the Year**. The data are *clearly skewed to the left* and *non-symmetric*. It can be seen that *game releases* were high between **2000** and **2010**.

```
group_genre <- VGS %>%
  group_by(Genre) %>%
  dplyr::summarise(count=n()) %>%
  arrange(desc(count))

ggplot(group_genre, aes(x = reorder(Genre,count),y=count, fill=Genre)) +
  geom_bar(stat='identity') +
  geom_label(aes(label=count),size=3.4) +
  coord_flip() + # flipped coordinates.
  ggtitle('Genre Bar', subtitle = 'Distribution of the Genre') +
  theme_minimal() +
  xlab('Genre') +
  ylab('Count') +
  theme(legend.position = 'none') +
  theme(plot.title = element_text(family = 'serif',
                                  face='bold',
                                  color = 'darkgreen',
                                  size=15))
```

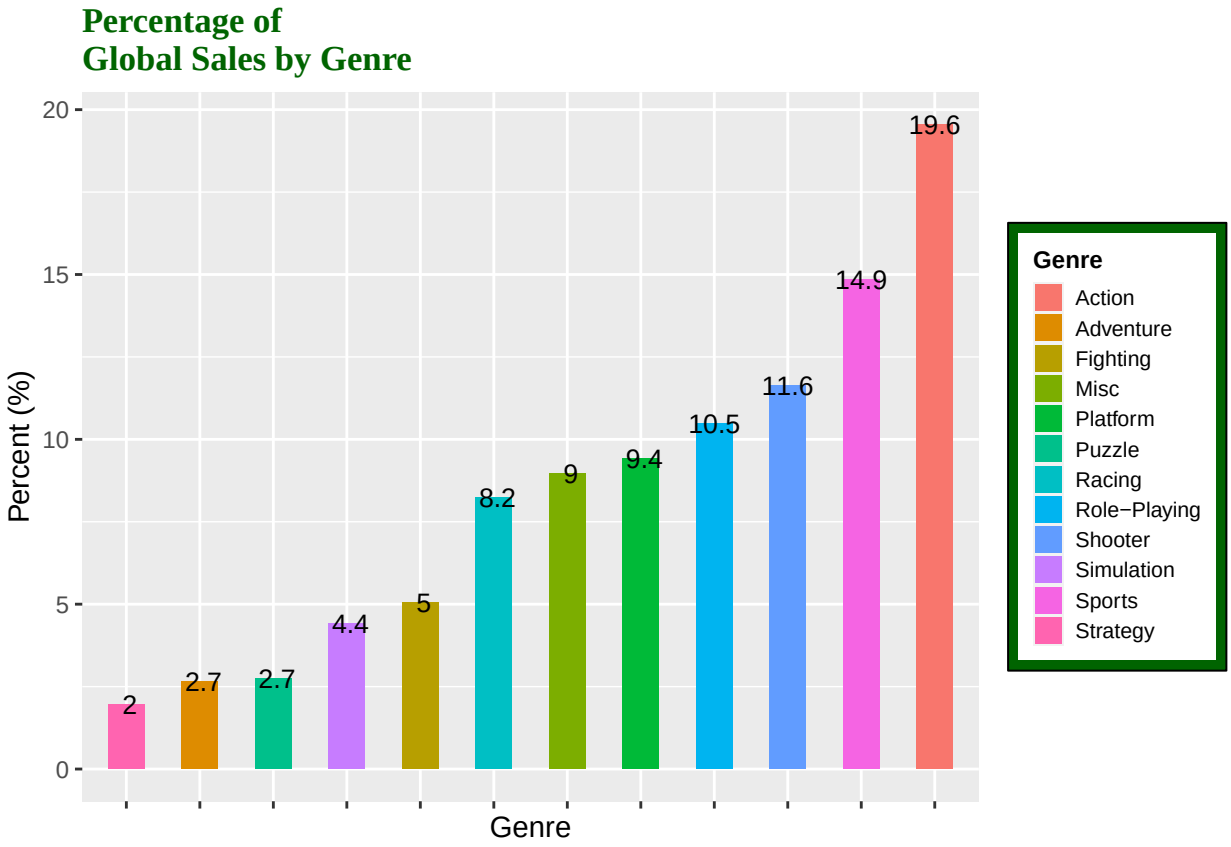


It can see that **Action genre game releases are the highest** at **3251** and **Puzzle game releases are the lowest** at **570**.

6 Data Analysis

6.0.1 Percentage of Global Sales by Genre

```
genre <- VGS %>%
  group_by(Genre) %>%
  dplyr::summarise(sales = sum(Global_Sales),
            .groups = 'drop') %>%
  mutate(percent = sales/sum(sales) * 100) %>%
  arrange(desc(sales))

ggplot(genre, aes(reorder(x=Genre,sales), y=percent, fill=Genre)) +
  geom_bar(stat='identity', width=.5, position='dodge') +
  labs(title='Percentage of\nGlobal Sales by Genre') +
  xlab('Genre') +
  ylab('Percent (%)') +
  theme(axis.text.x = element_blank()) +
  theme(legend.key.height = unit(0.4, 'cm'),
        legend.key.width =  unit(0.4, 'cm'))  +
  # show percentage of the each genre on the graph.
  geom_text(aes(label=format(percent,
                             digits=2,
                             drop0trailing = TRUE)),
            colour='black',
            size=3.4) +
  theme(legend.box.background = element_rect(fill='darkgreen'),
        legend.box.margin = margin(4,4,4,4)) +
  theme(plot.title = element_text(family='serif',
                                  color = 'darkgreen',
                                  face='bold')) +
   theme(legend.title = element_text(size=9, face='bold')) +
  theme(legend.text = element_text(size=7.5))
```

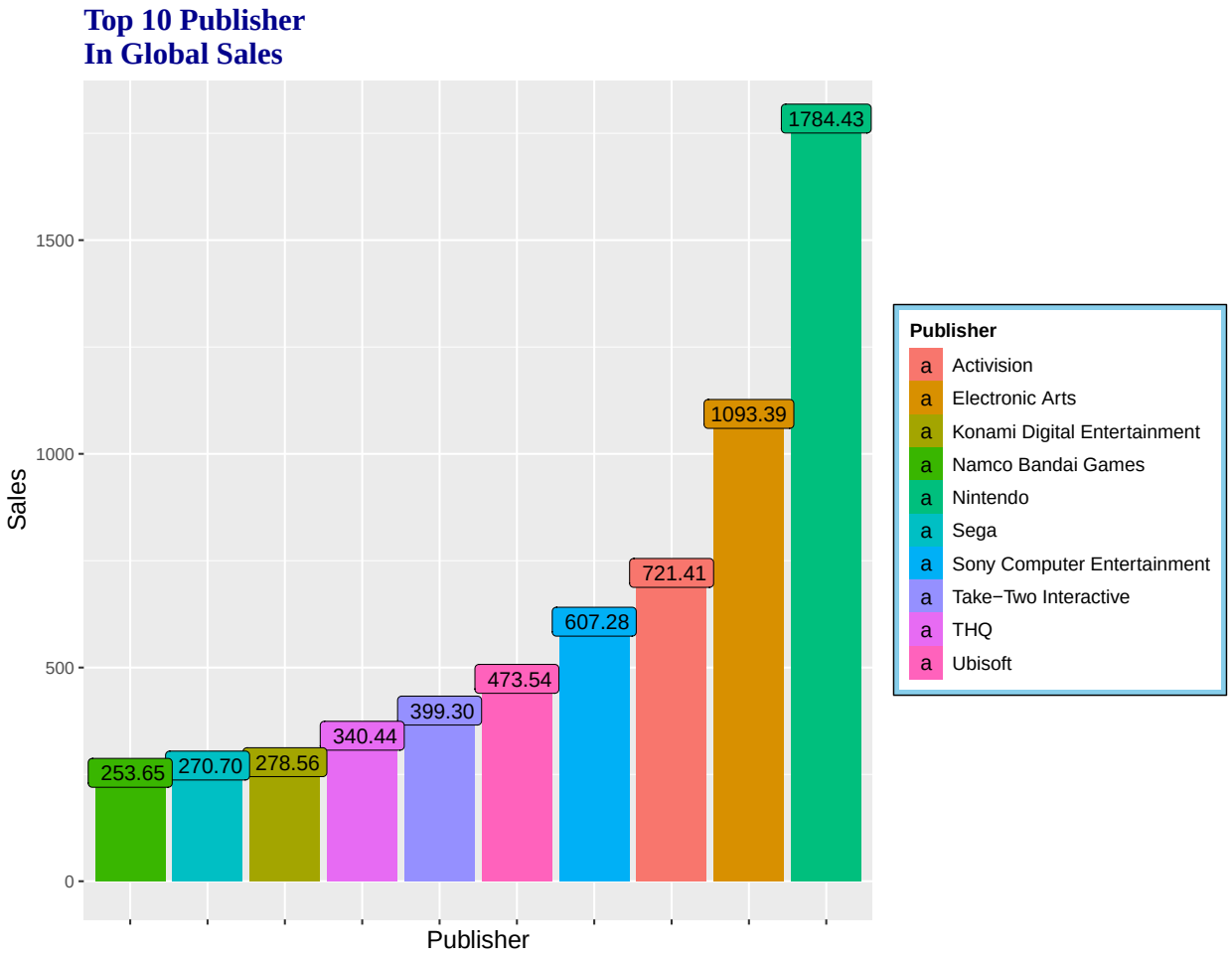


Notice that percentage of global sales by genre bar, it shows that **Action** is highest sales rate **19.6** (%). Conversely, **Strategy** is lowest sales rate **2** (%).

6.0.2 Top 10 Global Sales by Publisher, Platform, Game

```
# Top 10 Global Sales by Publisher
pub_top10 <- VGS %>%
  group_by(Publisher) %>%
  dplyr::summarise(sales = sum(Global_Sales),
            .groups = 'drop') %>%
  arrange(desc(sales)) %>%
  head(10)

ggplot(pub_top10, aes(reorder(x=Publisher,sales),y=sales,fill=Publisher)) +
  geom_bar(stat = 'identity') +
  geom_label(aes(label=format(sales)), size=4) +
  labs(title = 'Top 10 Publisher\nIn Global Sales') +
  theme(axis.text.x = element_blank()) +
  xlab('Publisher') +
  ylab('Sales') +
  theme(plot.title = element_text(family = 'serif',
                                  face='bold',
                                  size=16,
                                  color='darkblue')) +
```

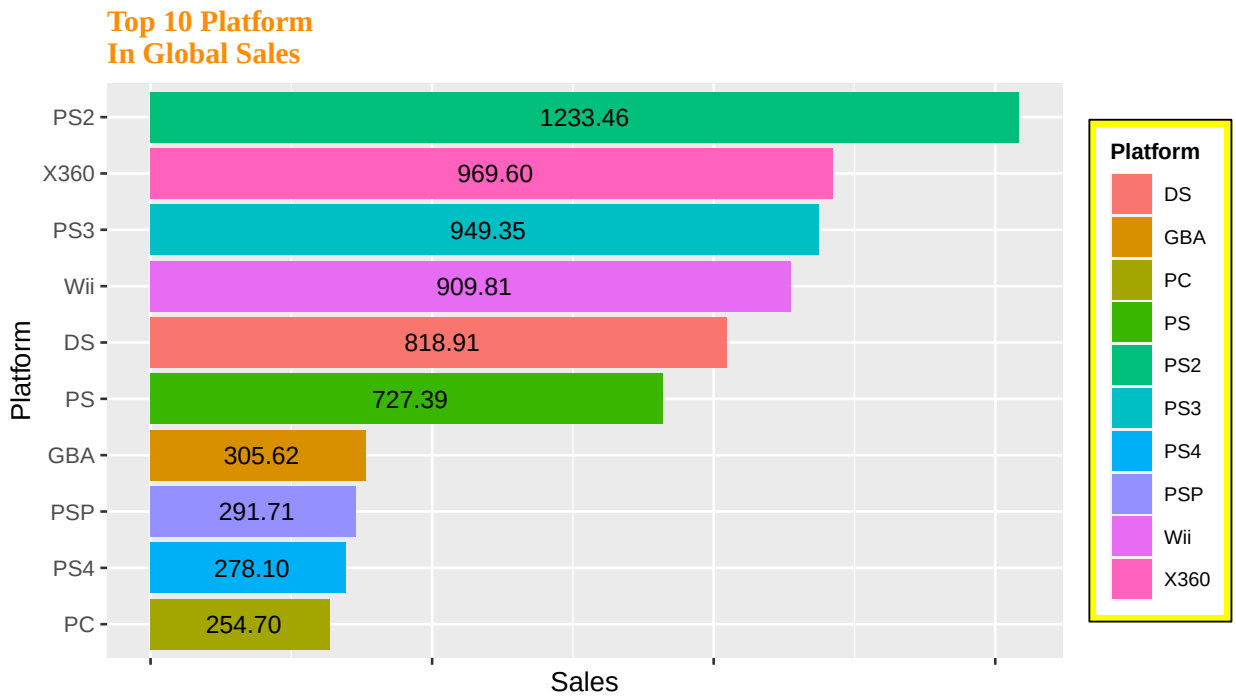
```
  theme(legend.box.background = element_rect(fill = 'skyblue'),
        legend.box.margin = margin(3,3,3,3)) +
  theme(legend.title = element_text(size=10, face='bold')) +
  theme(legend.text = element_text(size=9.5)) +
   theme(axis.title = element_text(size=13))
```



There are many Publisher in this dataset. So I expressed Publisher who had on impact on sales in a bar chart. **Nintendo**'s sales were overwhelming at **1784.43**(M).

```
#Top 10 Global Sales by Platform
plat_top10 <- VGS %>%
  group_by(Platform) %>%
  dplyr::summarise(sales = sum(Global_Sales),
            .groups = 'drop') %>%
  arrange(desc(sales)) %>%
  head(10)

ggplot(plat_top10, aes(reorder(x=Platform,sales),y=sales,fill=Platform)) +
  geom_bar(stat = 'identity') +
  geom_text(aes(label=format(sales)), size=3.6,position = position_stack(vjust = 0.5)) +
```

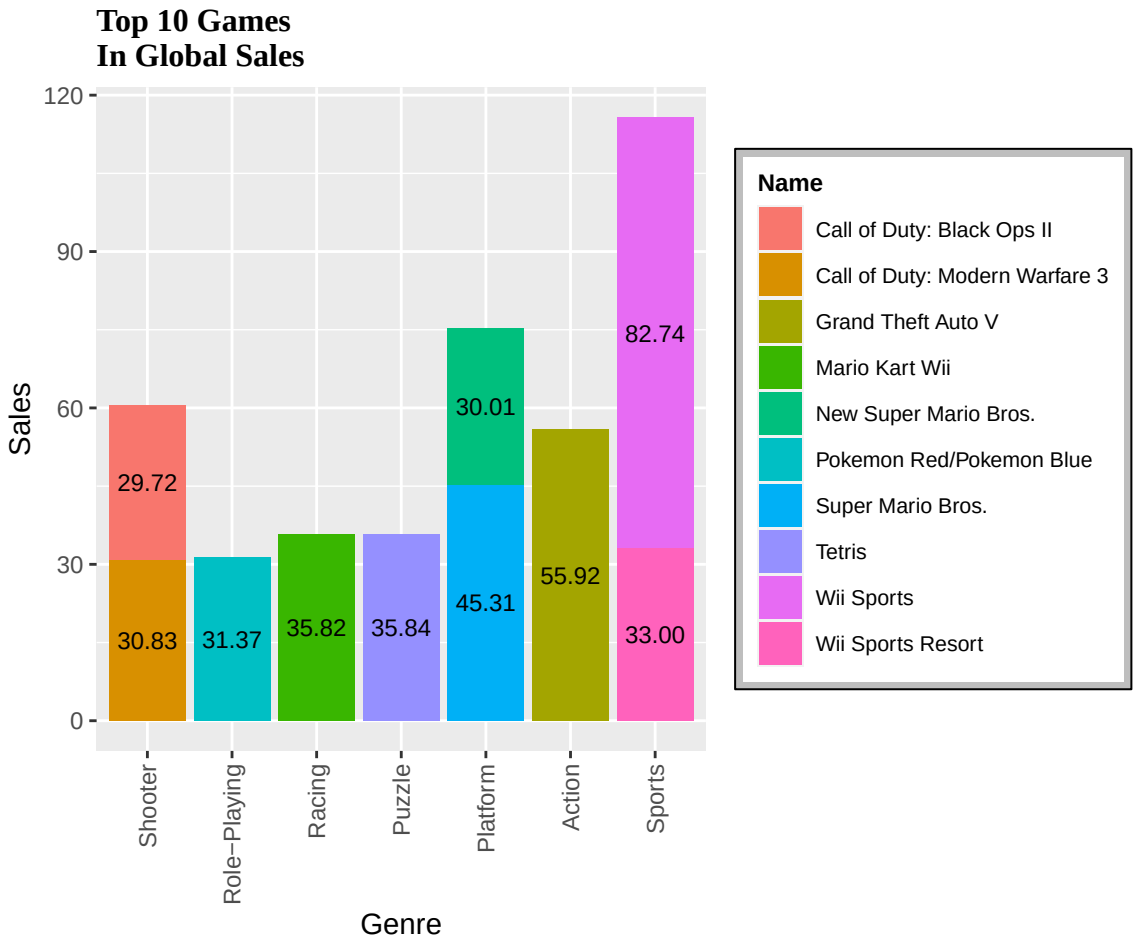
```
  labs(title = 'Top 10 Platform\nIn Global Sales') +
  xlab('Platform') +
  ylab('Sales') +
  coord_flip() +
  theme(axis.text.x = element_blank()) +
  theme(plot.title = element_text(family = 'serif',
                                  face='bold',
                                  size=12,
                                  color='darkorange')) +
  theme(legend.box.background = element_rect(fill = 'yellow'),
        legend.box.margin = margin(3,3,3,3)) +
  theme(legend.title = element_text(size=9, face='bold')) +
  theme(legend.text = element_text(size=7.5))
```



There are many Platform in this dataset. So I expressed Platform who had an impact on sales in a bar chart. Notice that *Top 10 Platform bar chart*, **PS2**'s sales were the highest at **1233.46**(M). It can also be seen that **X360** is **969.60**(M) and **PS3** is **949.35**(M), ranking second and third, respectively.

```
# Top 10 Global Sales by Games
game_top10 <- VGS %>%
  group_by(Name, Genre) %>%
  dplyr::summarise(sales = sum(Global_Sales),
            .groups = 'drop') %>%
  arrange(desc(sales)) %>%
  head(10)

ggplot(game_top10, aes(reorder(x=Genre,sales),y=sales,fill=Name)) +
  geom_bar(stat = 'identity') +
```

```
  geom_text(aes(label=format(sales)), size=3,position = position_stack(vjust = 0.5)) +
  labs(title = 'Top 10 Games\nIn Global Sales') +
  xlab('Genre') +
  ylab('Sales') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5 )) +
  theme(legend.box.background = element_rect(fill = 'gray'),
        legend.box.margin = margin(3,3,3,3)) +
  theme(legend.title = element_text(size=9, face='bold')) +
  theme(legend.text = element_text(size=7.5)) +
  theme(plot.title = element_text(family = 'serif', face='bold', size=12))
```

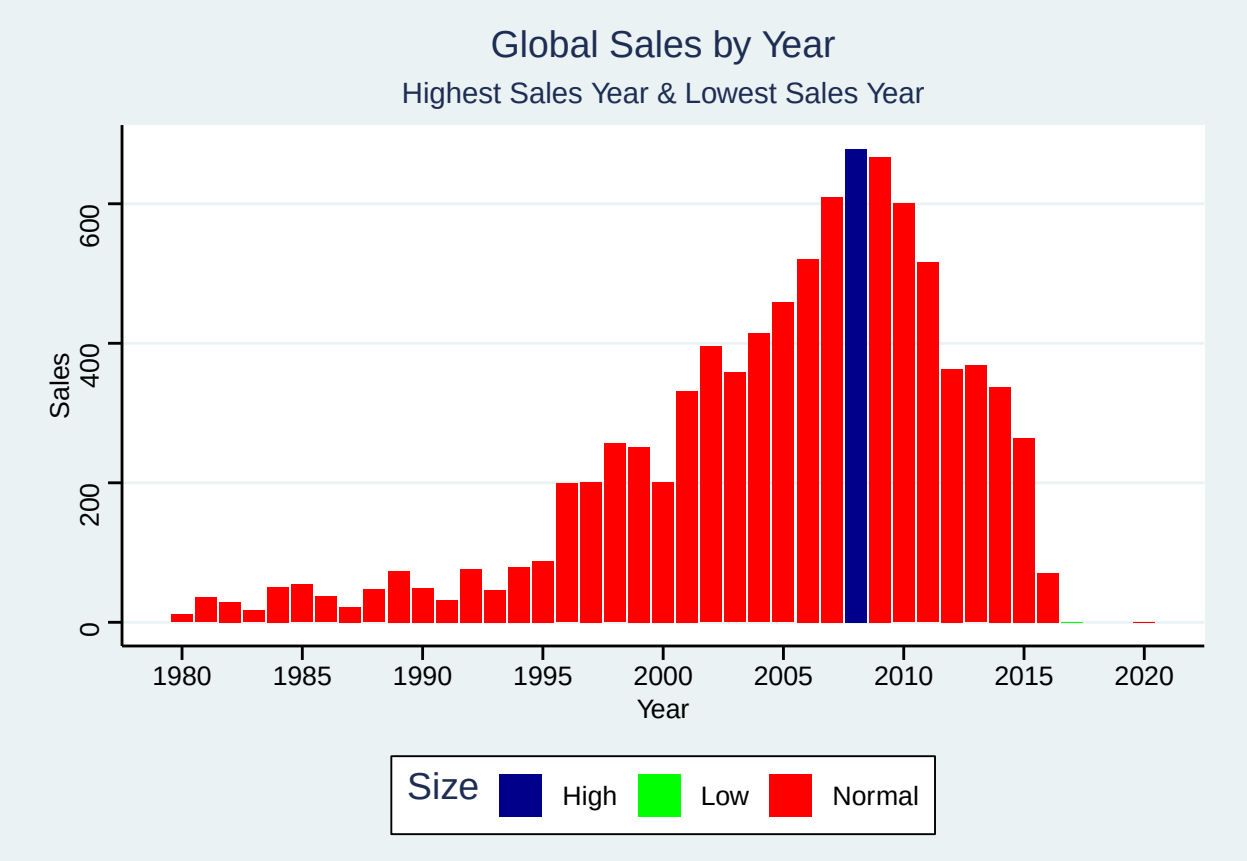


There are also many Games in this dataset. So I looked up the Top 10 Games in total sales. Notice that *Top 10 Games sales bar chart*, **Wii Sports**'s sales were overwhelming at **82.74(M)**. It can also be seen that **Grand Theft Auto V** is **55.92(M)** and **Super Mario Bros.** is **45.31(M)**, ranking second and third, respectively.

6.0.3 Highest Sales Year & Lowest Sales Year

```
Global_yr <- VGS %>%
  group_by(Year) %>%
```

```
  dplyr::summarise(sales = sum(Global_Sales),
            .groups = 'drop') %>%
  # find highest sales and lowest sales year
  mutate(Size = ifelse(sales == max(sales), 'High',
                       ifelse(sales==min(sales),
                              'Low','Normal')))

ggplot(Global_yr, aes(x = Year, y = sales)) +
  geom_bar(stat = 'identity', aes(fill=Size)) +
  scale_fill_manual(values = c('darkblue','green','red')) +
  theme_stata() +
  scale_color_stata() +
  ylab('Sales') +
  scale_x_continuous(breaks = c(1980,1985,1990,1995,2000,2005,2010,2015,2020)) +
  ggtitle('Global Sales by Year', subtitle = 'Highest Sales Year & Lowest Sales Year')
```




```
tail(Global_yr, n = 11)
```

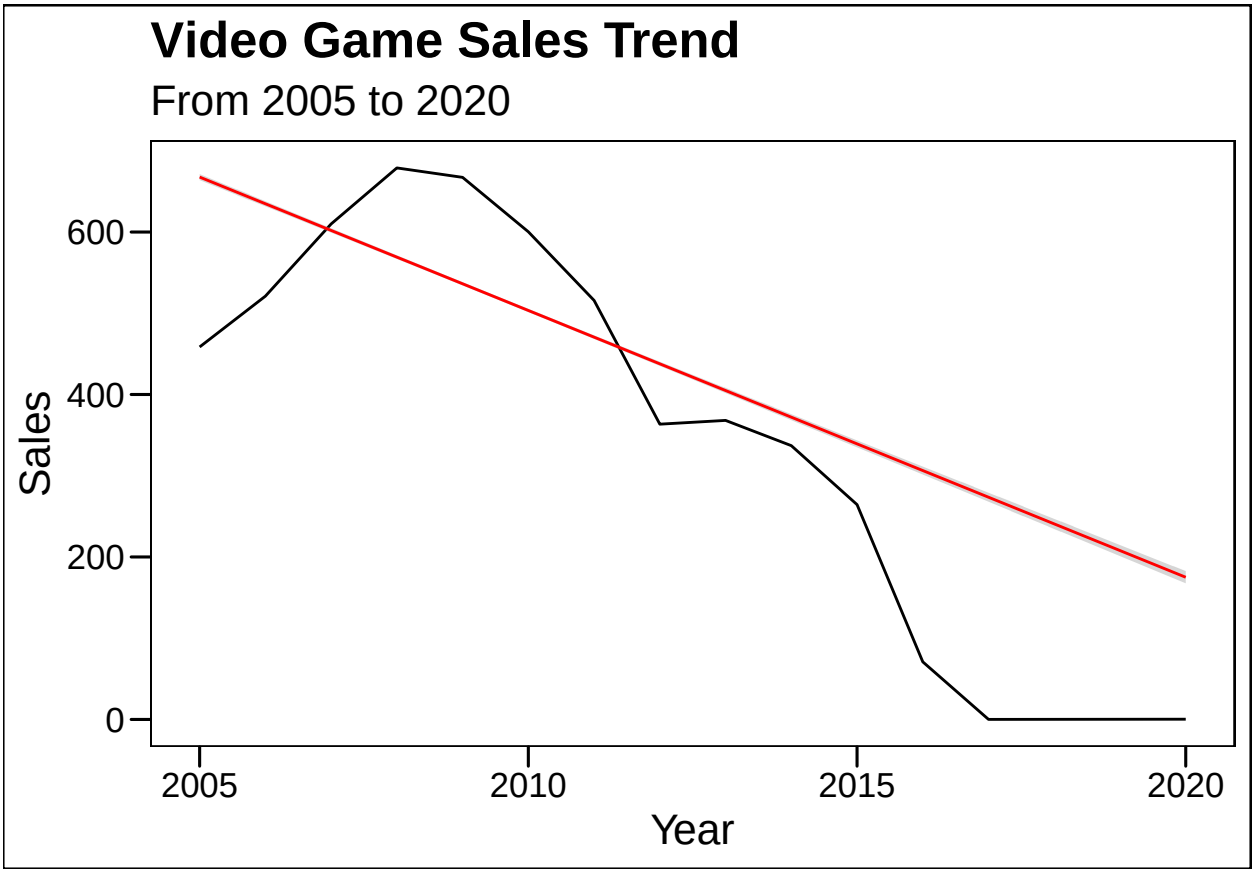
```
## # A tibble: 11 x 3
##     Year  sales Size
##    <dbl>  <dbl> <chr>
##  1  2008 679.   High
##  2  2009 667.   Normal
##  3  2010 600.   Normal
```

```
##  4  2011 516.   Normal
##  5  2012 363.   Normal
##  6  2013 368.   Normal
##  7  2014 337.   Normal
##  8  2015 264.   Normal
##  9  2016  70.9  Normal
## 10  2017   0.05 Low
## 11  2020   0.29 Normal
```

Let's look at the sales of games by year. Before the 2000s, sales were significantly lower. The *highest sales* year* was **2008**, when sales were **678.90**(M). The year when sales were the lowest is too low to be checked on the graph. The figures show the *lowest sales year* was **2017** at **0.05**(M).

6.0.4 Video Game Sales Trend

```
trend <- VGS %>%
  filter(Year %in% c(2005:2020)) %>%  # 2005 ~ 2020
  group_by(Year) %>%
  dplyr::mutate(sales = sum(Global_Sales))

ggplot(trend, aes(Year, sales)) +
  geom_line() +
  stat_smooth(method='lm', formula = y~x, colour='red', size=0.5) +
  labs(title='Video Game Sales Trend', subtitle = 'From 2005 to 2020', y='Sales') +
  theme_base()
```

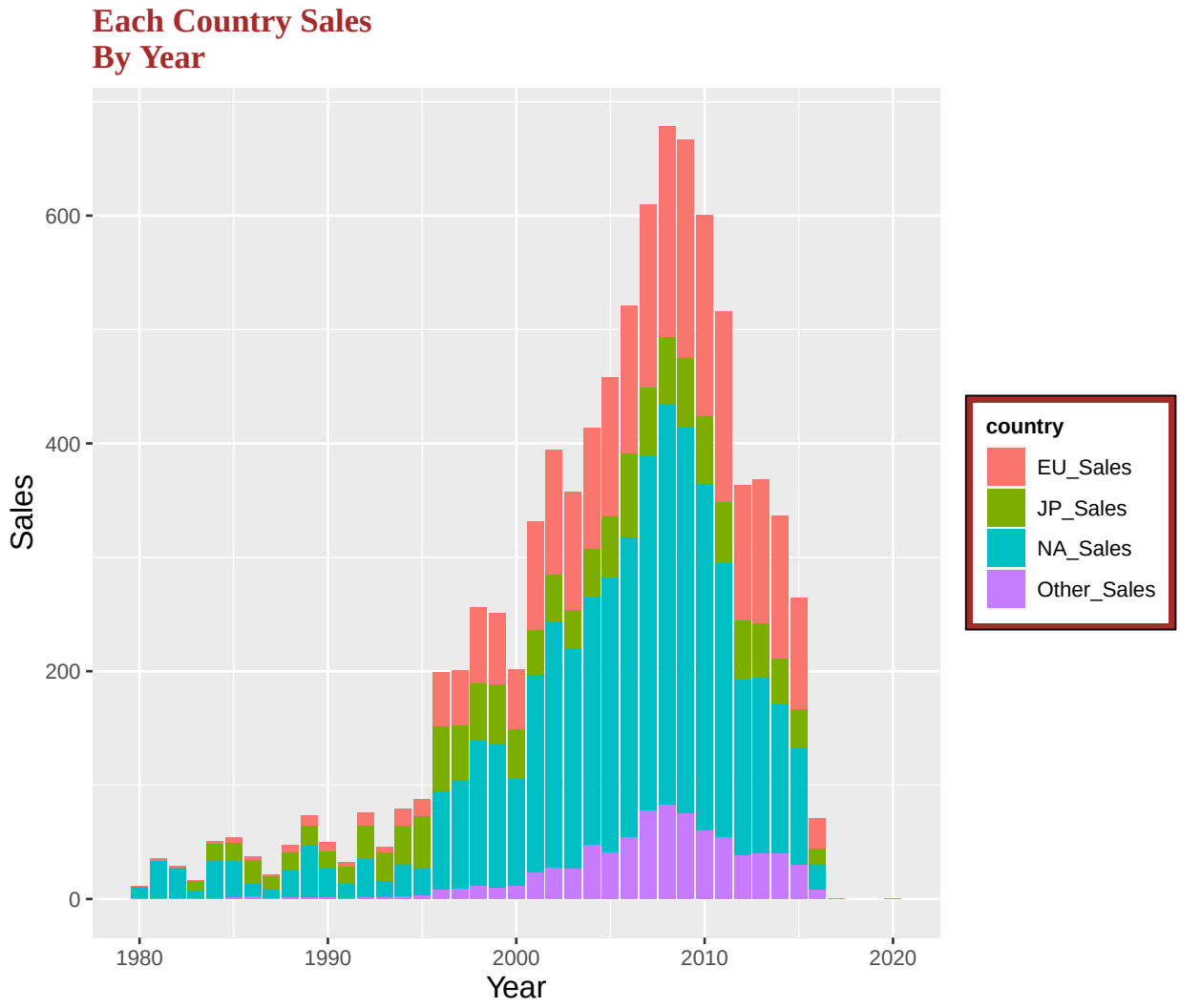


Notice that the graph of video game sales from 2005 to 2020. The slope of the regression line is gradually decreasing from 2005 to 2020. This shows that **Video game sales trend** are *decreasing*.

6.0.5 Each Country Sales by Year

```
yr_sales <- VGS %>%
  group_by(Year) %>%
  dplyr::summarise(Global_Sales = sum(Global_Sales),
            NA_Sales = sum(NA_Sales),
            JP_Sales = sum(JP_Sales),
            Other_Sales = sum(Other_Sales),
            EU_Sales = sum(EU_Sales))

yr_all_sales <- gather(yr_sales, country, sales, NA_Sales:EU_Sales)

ggplot(yr_all_sales, aes(x=Year, y=sales, fill= country)) +
  geom_bar(stat='identity') +
  labs(title='Each Country Sales\nBy Year',y='Sales') +
  theme(plot.title = element_text(family = 'serif',
                                  face='bold',
                                  size=14,
                                  color='brown')) +
  theme(legend.box.background = element_rect(fill = 'brown'),
        legend.box.margin = margin(3,3,3,3)) +
```

```
  theme(legend.title = element_text(size=9, face='bold')) +
  theme(legend.text = element_text(size=8.5)) +
  theme(axis.title = element_text(size=13))
```



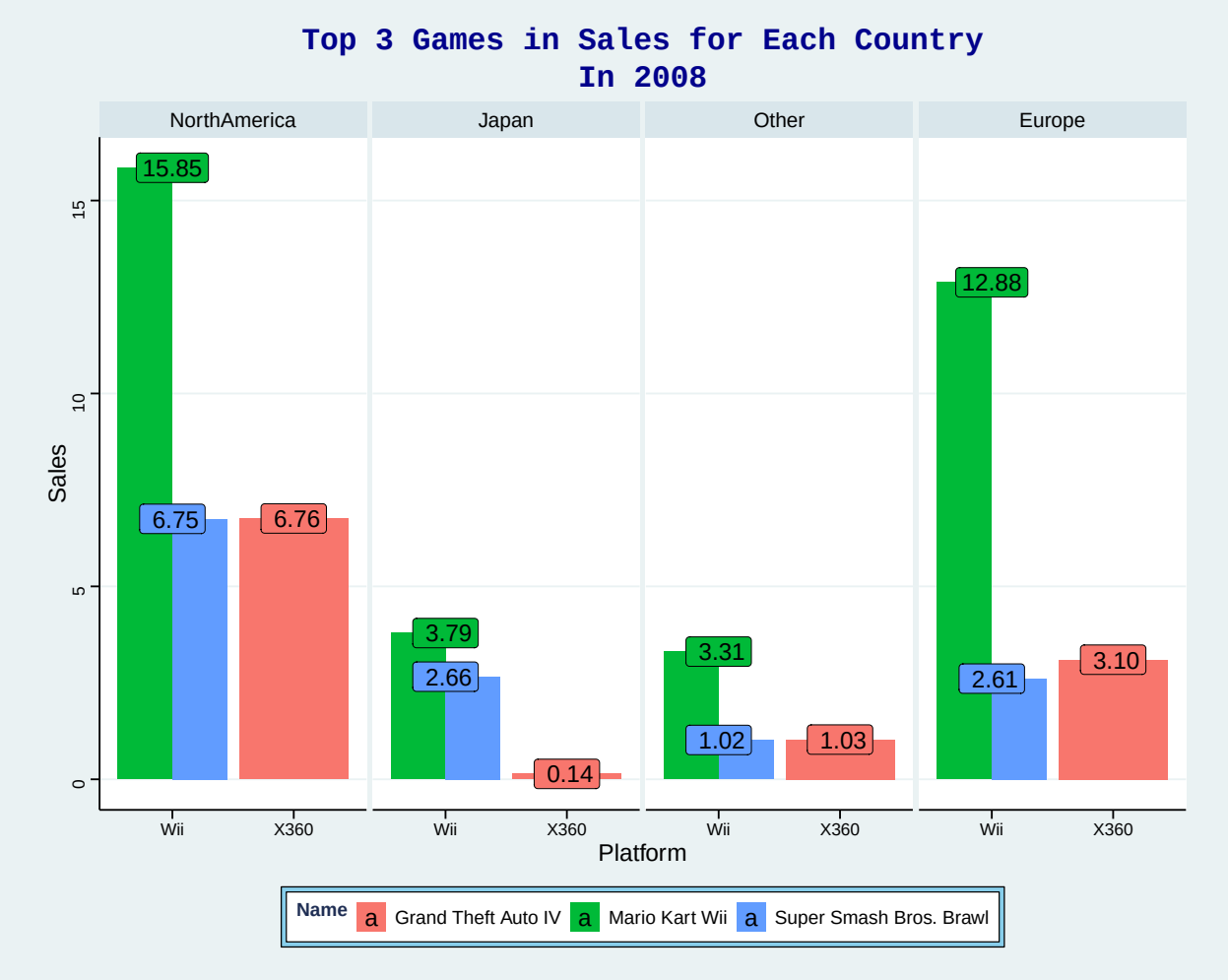
Let's check the sales volume of each country by year. It can be seen that game sales in the **North America** are the **highest**. On the contrary, we can see that sales of games in **Other Countries** are **significantly lower**. It can also be determined that there were almost no games sold in **Other Countries Sales** from *1980 to 1995*.

Sales Rank: North America > Europe > Japan > Other

6.0.6 Top 3 Games in Sales for Each Country in the Highest Sales Year

```
high_yr_top5_sales <- VGS %>%
  filter(Year %in% c(2008)) %>%
```

```
  group_by(Name, Platform) %>%
  dplyr::summarise(Global = sum(Global_Sales),
            NorthAmerica = sum(NA_Sales),
            Japan = sum(JP_Sales),
            Other = sum(Other_Sales),
            Europe = sum(EU_Sales),
            .groups = 'drop') %>%
  arrange(desc(Global)) %>%
  head(3)

top5 = melt(high_yr_top5_sales)
names(top5) = c('Name','Platform','Country','Sales')
top5 <- subset(top5, Country != 'Global')

ggplot(top5, aes(x=Platform, y=Sales, fill=Name)) +
  geom_bar(stat = 'identity', position='dodge') +
  facet_grid(~Country) +
  theme_stata() +
  scale_color_stata() +
  labs(title='Top 3 Games in Sales for Each Country\nIn 2008') +
  theme(plot.title = element_text(family = 'mono',
                                  face='bold',
                                  size=18,
                                  color='darkblue')) +
  theme(legend.box.background = element_rect(fill = 'skyblue'),
        legend.box.margin = margin(3,3,3,3)) +
  theme(legend.title = element_text(size=10.5, face='bold')) +
  theme(legend.text = element_text(size=10.5)) +
  geom_label(aes(label=format(Sales)), size=5) +
   theme(axis.title = element_text(size=14))
```



Game sales were the *highest in 2008*. So I checked three of the best-selling games in each country in 2008 and the platform it belongs to. The best-selling game of *2008* was **Mario Kart Wii, a part of the Wii platform**, which *topped the list in all countries. Second* and *third* are **Super Smash Bros. Brawl, which is part of the Wii platform**, and **Grand Theft Auto IV, which is part of the X360 platform**, with sales of the two games similar except for Japan.

7 Conclusion

- For the analysis of Big question, exploratory analysis showed that each country's sales were influential in total sales and that outliers existed. Also found that the data were generally skewed to the left (Negative Skewness). Also use visual packages to see how much sales has been affected by each values.
- In summary, the **higher the annual game release**, the **higher the annual game sales**. The country that had the *biggest impact on video game sales* shows that the **North America** is overwhelming. People were mostly interested in **Action-genre games**, and I can see that there were many action-genre games released. The **Wii**, the Publisher, who had a significant impact on total sales and **PS2**, the Platform, which had a significant impact on total sales. The best-selling games so far show that **Wii Sports** is overwhelmingly high, and compared to game sales by genre, the game has had a significant impact on sports genre sales. Finally, after checking the *trend of video games*, it is **decreasing**. Sales were the *highest in 2008* and sales were very *low in 2017*. Video games used to be very popular, but video game prospects are not good present.