videogamessaleanalysis

Ashish Das

10/11/2021

R Markdown

```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
df <- read.csv("E:\\Semester 5\\CSE3505 - FDA\\Project\\Gaming\\vgsales.csv")
head(df)
```

```
##   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
```

```
df_clean <- na.omit(df)

sum(is.na(df_clean))
```

```
## [1] 0
```

```
sum(is.na(df$Rank))
```

```
## [1] 0
```

```
sum(is.na(df$Name))
```

```
## [1] 0
```

```
sum(is.na(df$Platform))
```

```
## [1] 0
```

```
sum(is.na(df$Year))
```

```
## [1] 0
```

```
sum(is.na(df$Genre))
```

```
## [1] 0
```

```
sum(is.na(df$Publisher))
```

```
## [1] 0
```

```
sum(is.na(df$NA_Sales))
```

```
## [1] 0
```

```
sum(is.na(df$EU_Sales))
```

```
## [1] 0
```

```
sum(is.na(df$JP_Sales))
```

```
## [1] 0
```

```
sum(is.na(df$Other_Sales))
```

```
## [1] 0
```

```
sum(is.na(df$Global_Sales))
```

```
## [1] 0
```

```
unique(df$NA_Sales)
```

```
##   [1] 41.49 29.08 15.85 15.75 11.27 23.20 11.38 14.03 14.59 26.93  9.07  9.81
##  [13]  9.00  8.94  9.09 14.97  7.01  9.43 12.78  4.75  6.42 10.83  9.54  9.63
##  [25]  8.41  6.06  5.57  3.44  6.85  9.03  5.89  9.67  5.17  5.77  4.99  8.25
##  [37]  8.52  5.54  6.99  6.75  5.98  2.55  4.74  7.97  3.80  4.40  6.91  3.01
##  [49]  6.16  4.23  6.76  4.02  4.89  2.96  4.76  5.99  4.34  5.08  6.05  6.72
##  [61]  7.03  5.55  3.66  6.63  4.09  5.84  3.88  5.91  4.36  5.58  2.01  4.46
##  [73]  5.03  3.54  1.11  1.79  6.82  3.81  2.91  1.06  0.98  5.80  2.58  2.28
##  [85]  2.82  7.28  2.90  2.93  2.80  4.10  3.78  5.39  3.24  4.79  3.83  4.52
##  [97]  3.51  2.85  3.27  3.68  4.41  3.13  2.47  4.12  4.14  0.78  2.71  2.77
## [109]  3.23  3.50  4.15  3.10  0.84  1.67  2.79  0.79  3.25  3.74  2.64  4.98
## [121]  2.57  3.64  3.70  4.01  0.07  3.11  3.92  4.05  2.45  4.47  2.63  3.18
## [133]  2.41  1.88  0.66  2.26  2.49  2.97  2.54  2.95  3.28  2.70  2.99  0.47
## [145]  3.14  2.62  3.21  2.72  2.07  1.97  1.74  2.18  3.02  1.62  1.92  3.33
## [157]  1.22  2.30  4.26  0.65  2.43  2.32  1.08  1.90  2.10  0.96  1.64  1.98
## [169]  3.59  3.22  1.96  2.66  1.70  0.60  3.40  2.05  3.42  2.59  3.36  3.06
## [181]  3.49  3.39  1.85  2.31  3.98  2.89  0.00  2.74  2.56  1.91  0.57  0.28
## [193]  2.36  1.73  3.05  1.87  1.94  2.08  2.29  2.42  2.60  1.89  1.78  1.55
## [205]  3.19  4.18  4.21  3.63  0.20  1.54  2.67  0.10  2.19  2.03  3.03  2.20
## [217]  0.92  2.75  4.00  2.51  2.11  2.23  1.41  3.00  1.46  0.88  1.30  1.28
## [229]  2.25  2.02  3.38  2.04  3.79  1.40  4.03  1.65  0.71  2.14  1.42  2.13
## [241]  2.65  2.35  0.12  1.68  1.12  2.78  1.38  2.15  1.18  1.33  0.67  1.53
## [253]  1.15  0.93  2.12  2.48  0.16  0.87  2.21  1.44  1.49  1.14  2.40  1.82
## [265]  1.37  1.93  0.58  1.59  2.53  2.33  0.05  1.61  2.38  1.57  1.56  1.23
## [277]  1.66  1.17  2.84  0.59  2.09  2.39  1.34  1.13  0.86  1.75  0.46  1.43
## [289]  1.63  1.45  1.47  1.99  1.50  0.80  1.36  0.50  0.25  0.95  1.27  0.03
## [301]  1.72  0.73  1.76  1.35  1.48  1.52  1.86  2.06  0.68  0.91  1.69  0.08
## [313]  1.29  2.17  2.50  1.01  1.58  1.04  2.22  1.83  0.61  1.84  0.99  1.51
## [325]  0.09  0.40  2.52  1.32  0.02  1.05  0.29  1.19  0.89  0.30  1.20  1.24
## [337]  1.25  1.07  1.02  0.69  1.95  2.00  0.76  0.63  0.90  0.48  0.64  0.37
## [349]  1.31  0.15  1.21  0.49  0.13  1.81  1.26  0.81  0.77  1.00  1.16  1.39
## [361]  0.85  0.52  0.51  0.38  0.62  1.09  1.71  1.03  0.34  1.60  0.54  0.14
## [373]  0.01  0.82  0.83  0.11  0.94  1.77  0.70  0.97  0.75  0.35  0.72  0.74
## [385]  0.18  1.10  0.56  0.26  0.21  0.22  0.53  0.55  0.23  0.39  0.32  0.45
## [397]  0.41  0.31  0.24  0.06  0.43  0.44  0.19  0.04  0.17  0.36  0.33  0.27
## [409]  0.42
```

1. Which region has performed the best in terms of sales?

We will utilize the average sales made per region and compare the results. Before we do that, let's make sure we know how to calculate the average. We observe that our output is coming in decimals, to convert the values in millions, let's multiple the result with 10,00,000. The final code should look like this.

```
x <- mean(df$NA_Sales,na.rm = TRUE)*1000000
y <- mean(df$EU_Sales,na.rm = TRUE)*1000000
z <- mean(df$JP_Sales,na.rm = TRUE)*1000000
q <- mean(df$Other_Sales,na.rm = TRUE)*1000000
p <- mean(df$Global_Sales,na.rm = TRUE)*1000000
```

```
print(paste("The average sales in North America =", x))
```

```
## [1] "The average sales in North America = 264667.429810821"
```

```
print(paste("The average sales in Europe =", y))
```

```
## [1] "The average sales in Europe = 146652.006265815"
```

```
print(paste("The average sales in Japan =", z))
```

```
## [1] "The average sales in Japan = 77781.660441017"
```

```
print(paste("The average sales in other regions =", q))
```

```
## [1] "The average sales in other regions = 48063.0196409206"
```

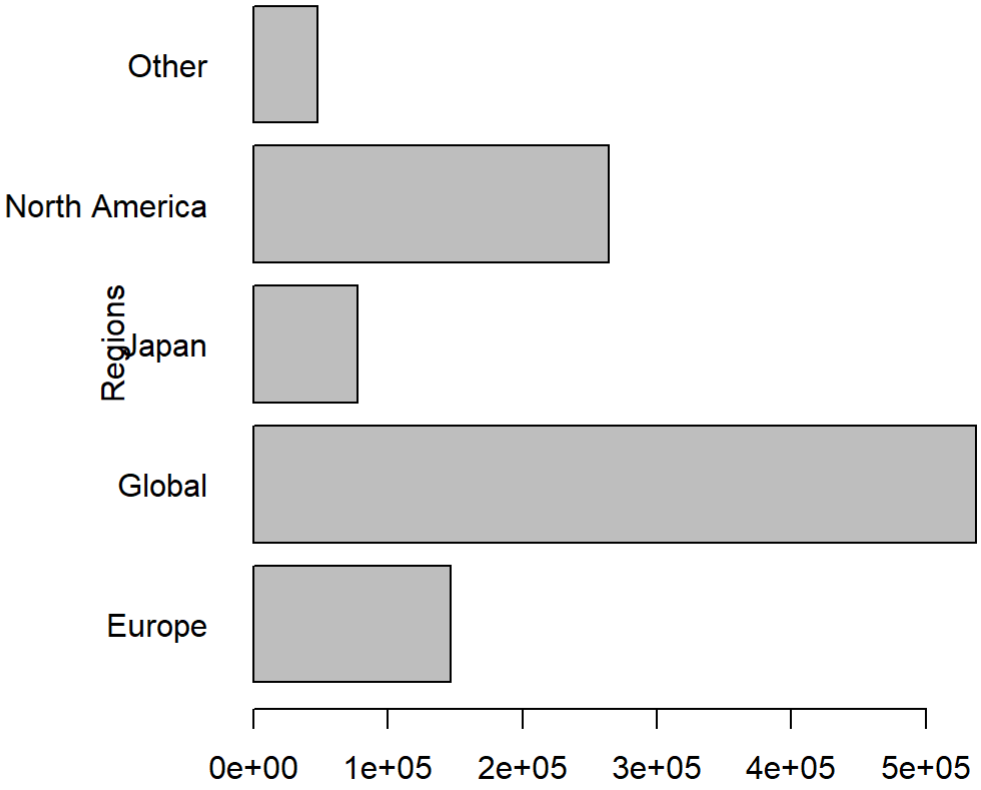
```
print(paste("The average sales globally =", p))
```

```
## [1] "The average sales globally = 537440.655500663"
```

```
X <- c(537440.656,264667.430, 146652.006, 77781.660, 48063.020)
Y <- c('Global','North America', 'Europe', 'Japan', 'Other')
```

Let us now plot our findings on a bar chart.

```
par(mar=c(3, 15, 3, 1))
barplot(X ~ Y,las=1, ylab = "Regions", horiz = TRUE)
```



2. The top gaming consoles are Microsoft (Xbox), Sony (Playstation) and Nintendo, with Google acting as a new competitor.

```
unique(df$Platform)
```

```
##  [1] "Wii"       "NES"       "GB"        "DS"        "X360"      "PS3"
##  [7] "PS2"       "SNES"      "GBA"       "3DS"       "PS4"       "N64"
## [13] "PS"        "XB"        "PC"        "Atari2600" "PSP"       "XOne"
## [19] "GC"        "WiiU"      "GEN"       "DC"        "PSV"       "SAT"
## [25] "SCD"       "WS"        "NG"        "TG16"      "3DO"       "GG"
## [31] "PCFX"
```

Grouping the Global sales based on each platform

```
Platform_Global = subset(df, select=c(Platform, Global_Sales))
head(Platform_Global)
```

```
##   Platform Global_Sales
## 1      Wii        82.74
## 2      NES        40.24
## 3      Wii        35.82
## 4      Wii        33.00
## 5       GB        31.37
## 6       GB        30.26
```

Grouping the North America sales based on each platform

```
Platform_NA = subset(df, select=c(Platform, NA_Sales))
head(Platform_NA)
```

```
##   Platform NA_Sales
## 1      Wii    41.49
## 2      NES    29.08
## 3      Wii    15.85
## 4      Wii    15.75
## 5       GB    11.27
## 6       GB    23.20
```

Grouping the Europe sales based on each platform

```
Platform_EU = subset(df, select=c(Platform, EU_Sales))
head(Platform_EU)
```

```
##   Platform EU_Sales
## 1      Wii    29.02
## 2      NES     3.58
## 3      Wii    12.88
## 4      Wii    11.01
## 5       GB     8.89
## 6       GB     2.26
```

Grouping the Japan sales based on each platform

```
Platform_JP = subset(df, select=c(Platform, JP_Sales))
head(Platform_JP)
```

```
##   Platform JP_Sales
## 1      Wii     3.77
## 2      NES     6.81
## 3      Wii     3.79
## 4      Wii     3.28
## 5       GB    10.22
## 6       GB     4.22
```

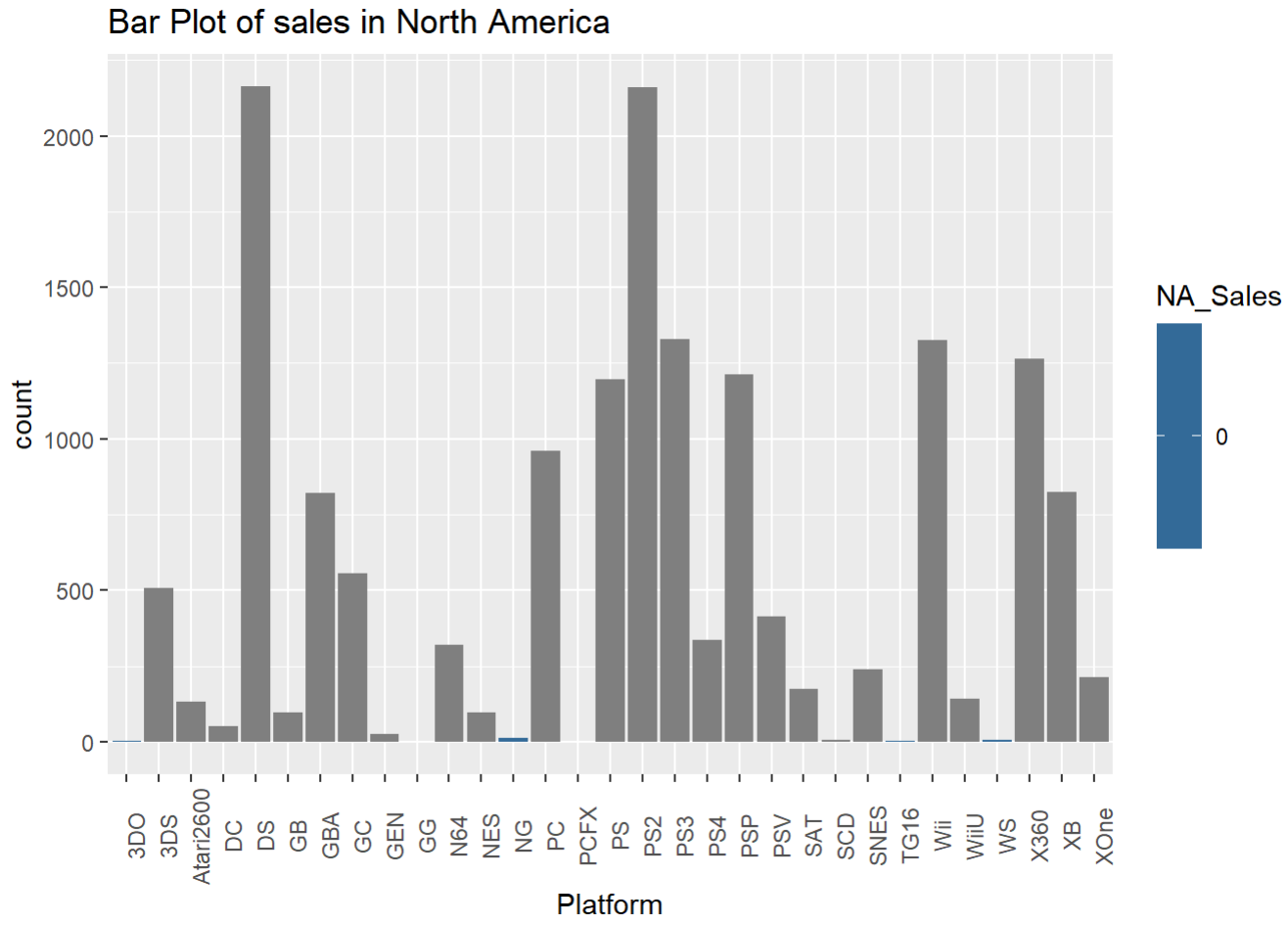
Grouping the other countries sales based on each platform

```
Platform_Other = subset(df, select=c(Platform, Other_Sales))
head(Platform_Other)
```

```
##   Platform Other_Sales
## 1      Wii        8.46
## 2      NES        0.77
## 3      Wii        3.31
## 4      Wii        2.96
## 5       GB        1.00
## 6       GB        0.58
```

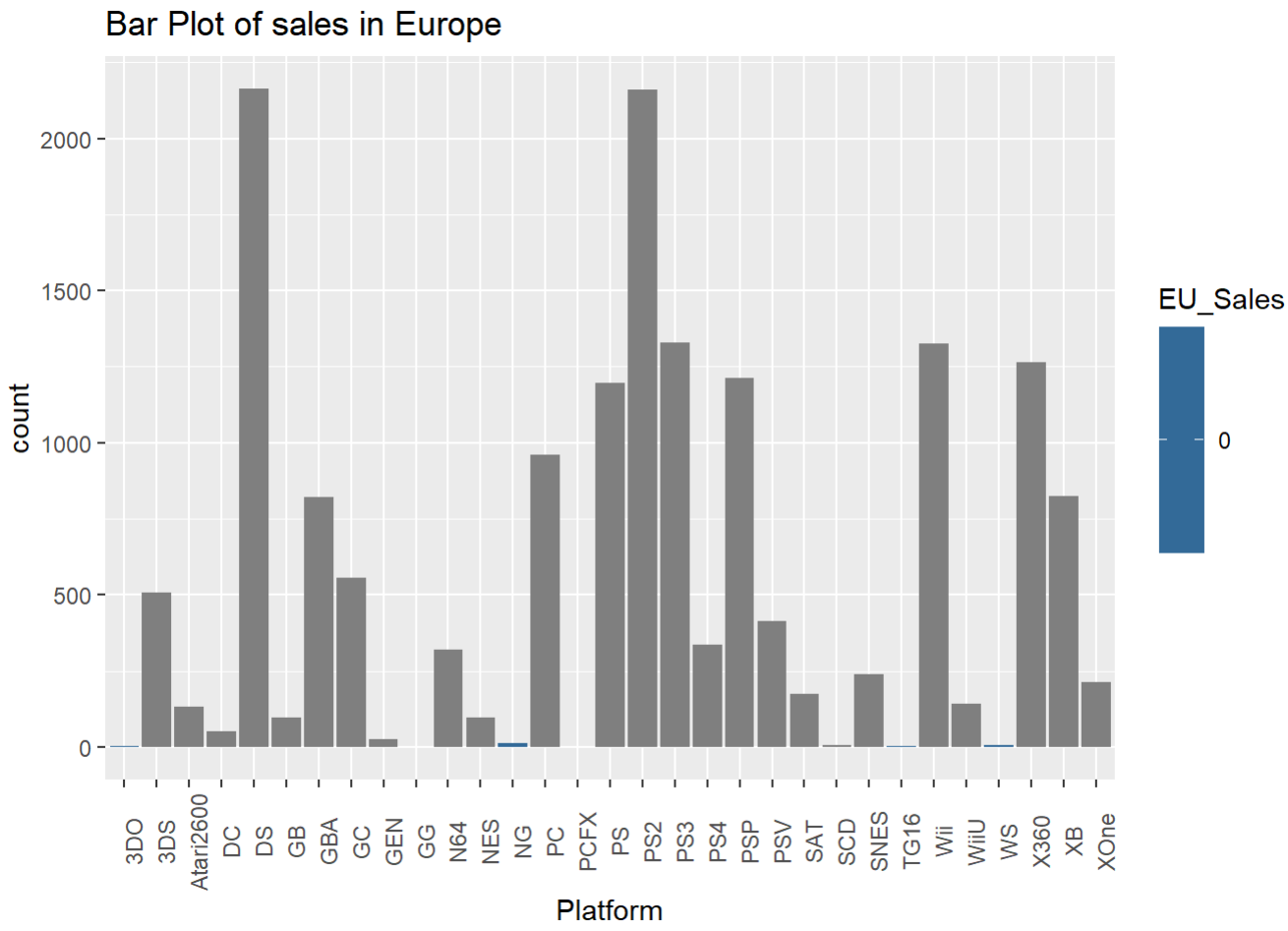
North America top Platforms

```
ggplot(data=Platform_NA,mapping=aes(x=Platform, fill=NA_Sales))+geom_bar() + ggtitle("Bar Plot o
f sales in North America") + theme(axis.text.x = element_text(angle = 90))
```



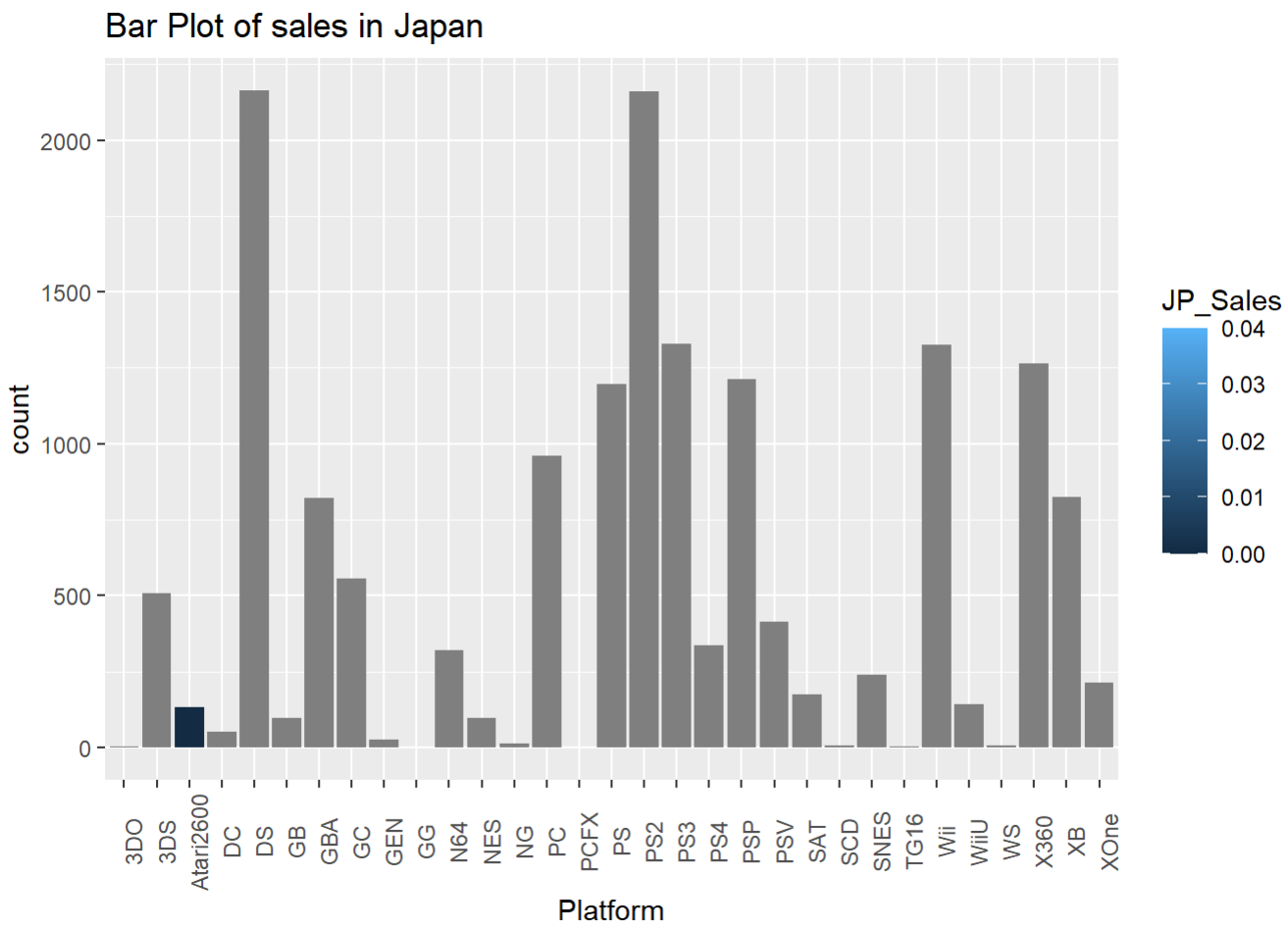
Europe top Platforms

```
ggplot(data=Platform_EU,mapping=aes(x=Platform, fill=EU_Sales))+geom_bar() + ggtitle("Bar Plot o
f sales in Europe") + theme(axis.text.x = element_text(angle = 90))
```



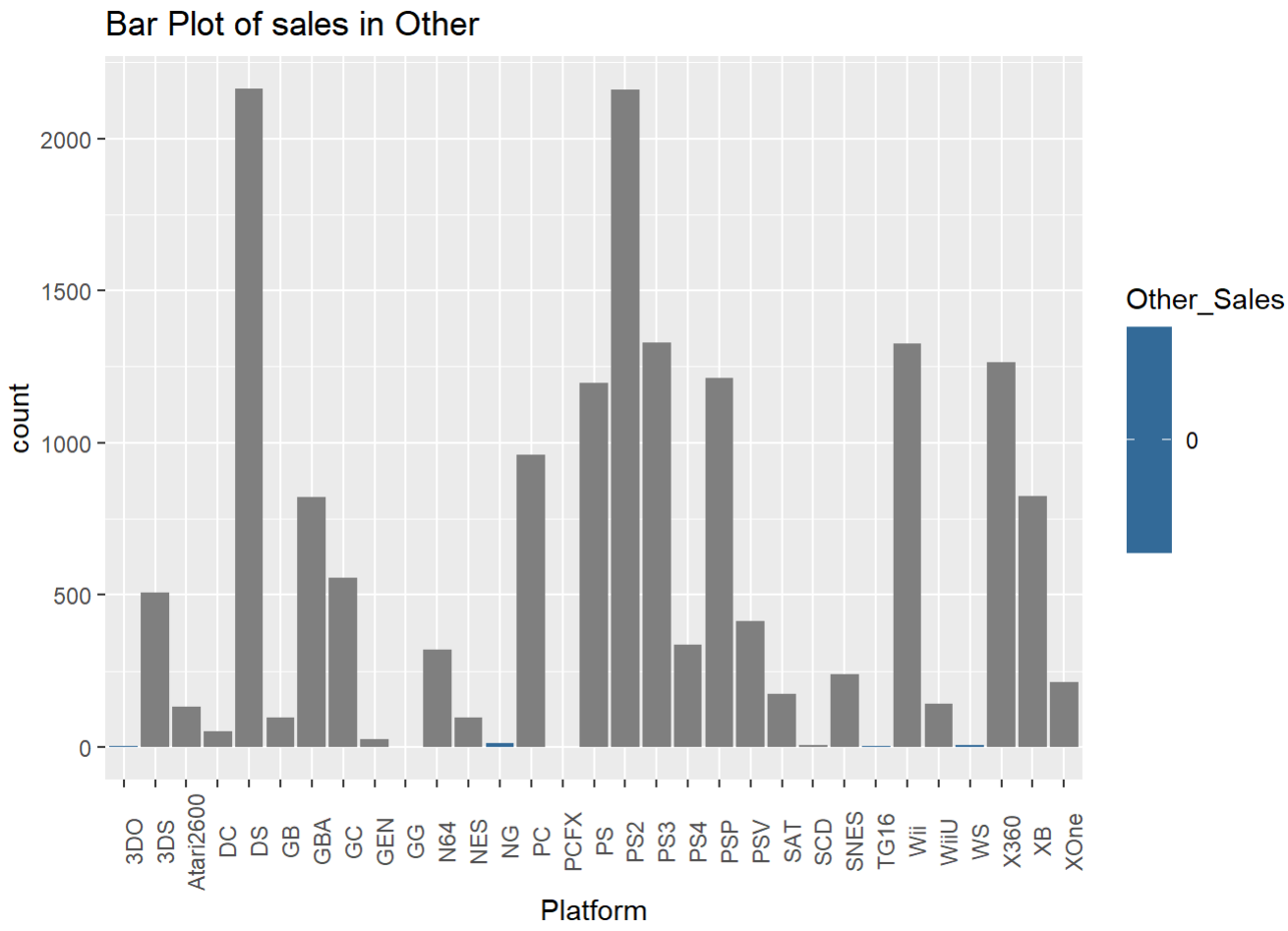
Japan top Platforms

```
ggplot(data=Platform_JP,mapping=aes(x=Platform, fill=JP_Sales))+geom_bar() + ggtitle("Bar Plot o
f sales in Japan") + theme(axis.text.x = element_text(angle = 90))
```



Other top Platforms

```
ggplot(data=Platform_Other,mapping=aes(x=Platform, fill=Other_Sales))+geom_bar() + ggtitle("Bar
 Plot of sales in Other") + theme(axis.text.x = element_text(angle = 90))
```



3. What are the top 10 games currently making the most sales globally?

We will use a similar approach by grouping the games with respect to the global sales and observe the top 10 games.

```
df2 = subset(df, select=c(Name,Global_Sales))
head(df2)
```

```
##                       Name Global_Sales
## 1               Wii Sports        82.74
## 2        Super Mario Bros.        40.24
## 3           Mario Kart Wii        35.82
## 4        Wii Sports Resort        33.00
## 5 Pokemon Red/Pokemon Blue        31.37
## 6                   Tetris        30.26
```

```
df3<-head(df2,10)
df3<-df3 %>%
  group_by(Name)%>%
  arrange(desc(Global_Sales))
```

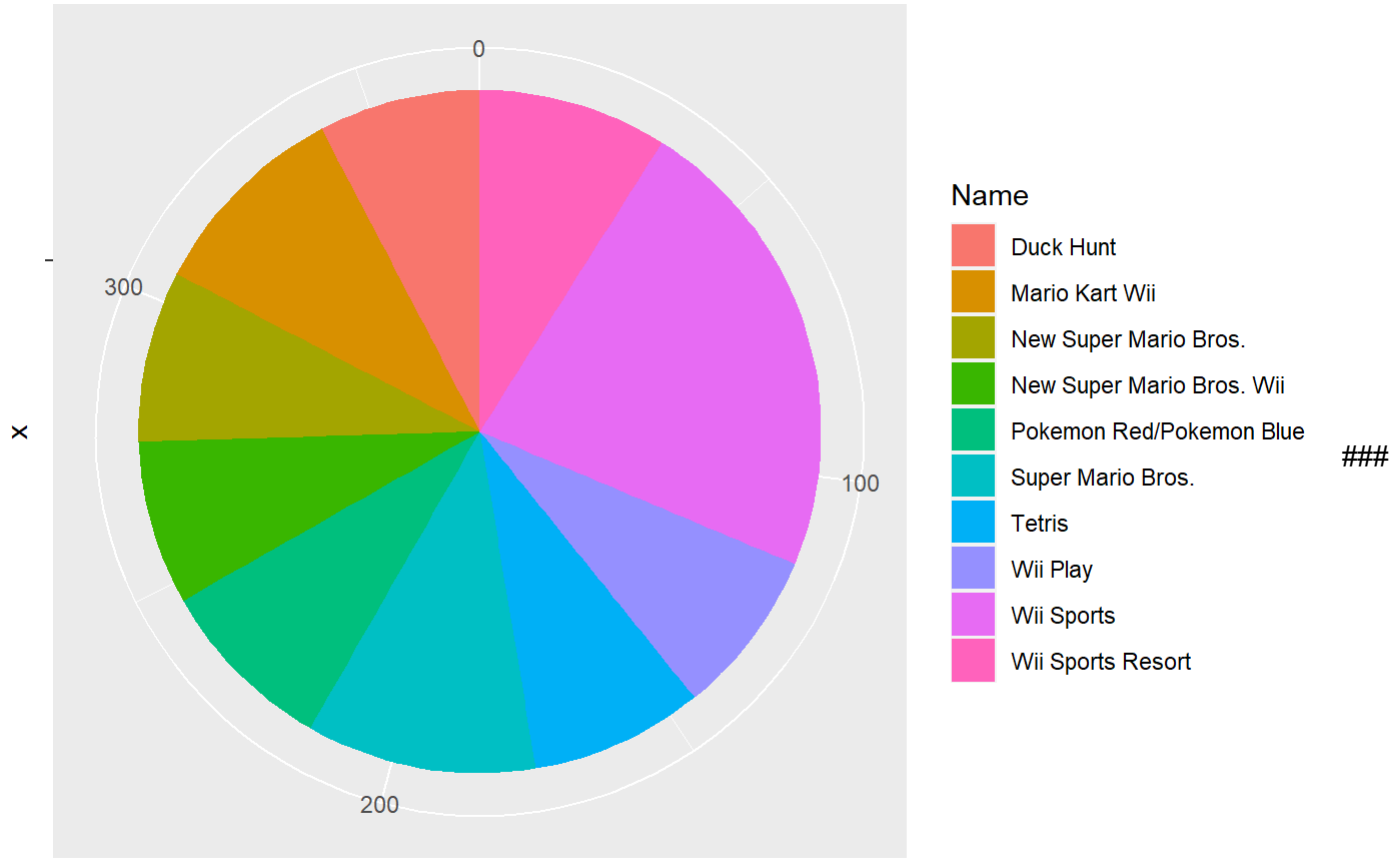
We see the most played game is Wii Sports making a total of $82.74 million globally.

We will plot the above using a pie chart.

```
df3
```

```
## # A tibble: 10 x 2
## # Groups:   Name [10]
##    Name                      Global_Sales
##    <chr>                            <dbl>
##  1 Wii Sports                        82.7
##  2 Super Mario Bros.                 40.2
##  3 Mario Kart Wii                    35.8
##  4 Wii Sports Resort                 33
##  5 Pokemon Red/Pokemon Blue          31.4
##  6 Tetris                            30.3
##  7 New Super Mario Bros.             30.0
##  8 Wii Play                          29.0
##  9 New Super Mario Bros. Wii         28.6
## 10 Duck Hunt                         28.3
```

```
ggplot(df3, aes(x="",y=Global_Sales,fill=Name))+geom_bar(width=1,stat='identity')+coord_polar(
"y",start = 0)
```



The pie chart also shows the proportion of sales each game holds, while also depicting the results.

4. What are the top games for different regions?

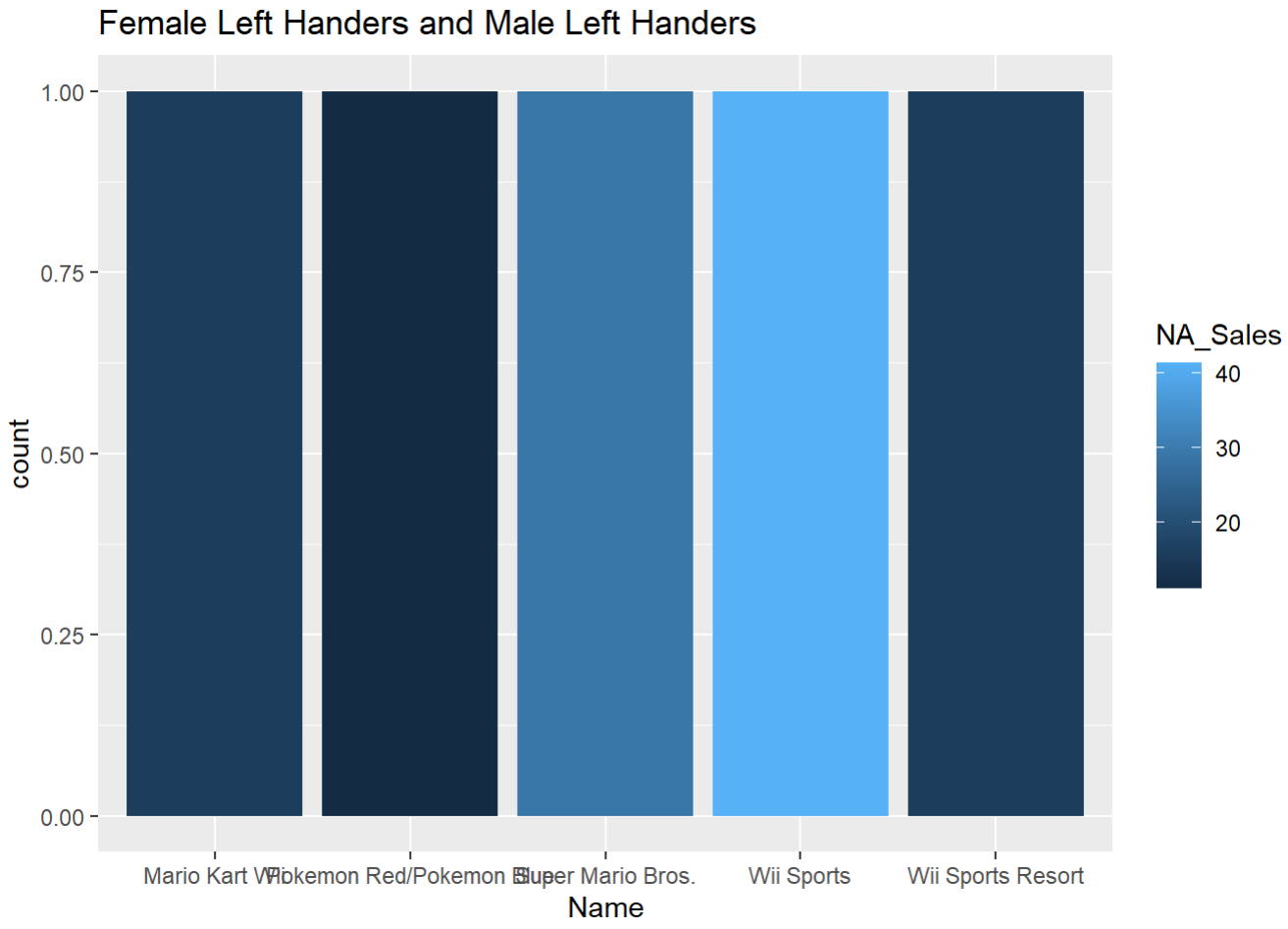
We will have to compare the sales made by different games regionally. We will use the same approach we did while analyzing the platform.

```
df4 = subset(df, select=c(Name,NA_Sales))
df4<-head(df4,5)
df4<-df4 %>%
  group_by(Name)%>%
  arrange(desc(NA_Sales))
df4
```

```
## # A tibble: 5 x 2
## # Groups:   Name [5]
##   Name                      NA_Sales
##   <chr>                        <dbl>
## 1 Wii Sports                    41.5
## 2 Super Mario Bros.             29.1
## 3 Mario Kart Wii                15.8
## 4 Wii Sports Resort             15.8
## 5 Pokemon Red/Pokemon Blue      11.3
```

Plotting it in a similar fashion to understand the results obtained.

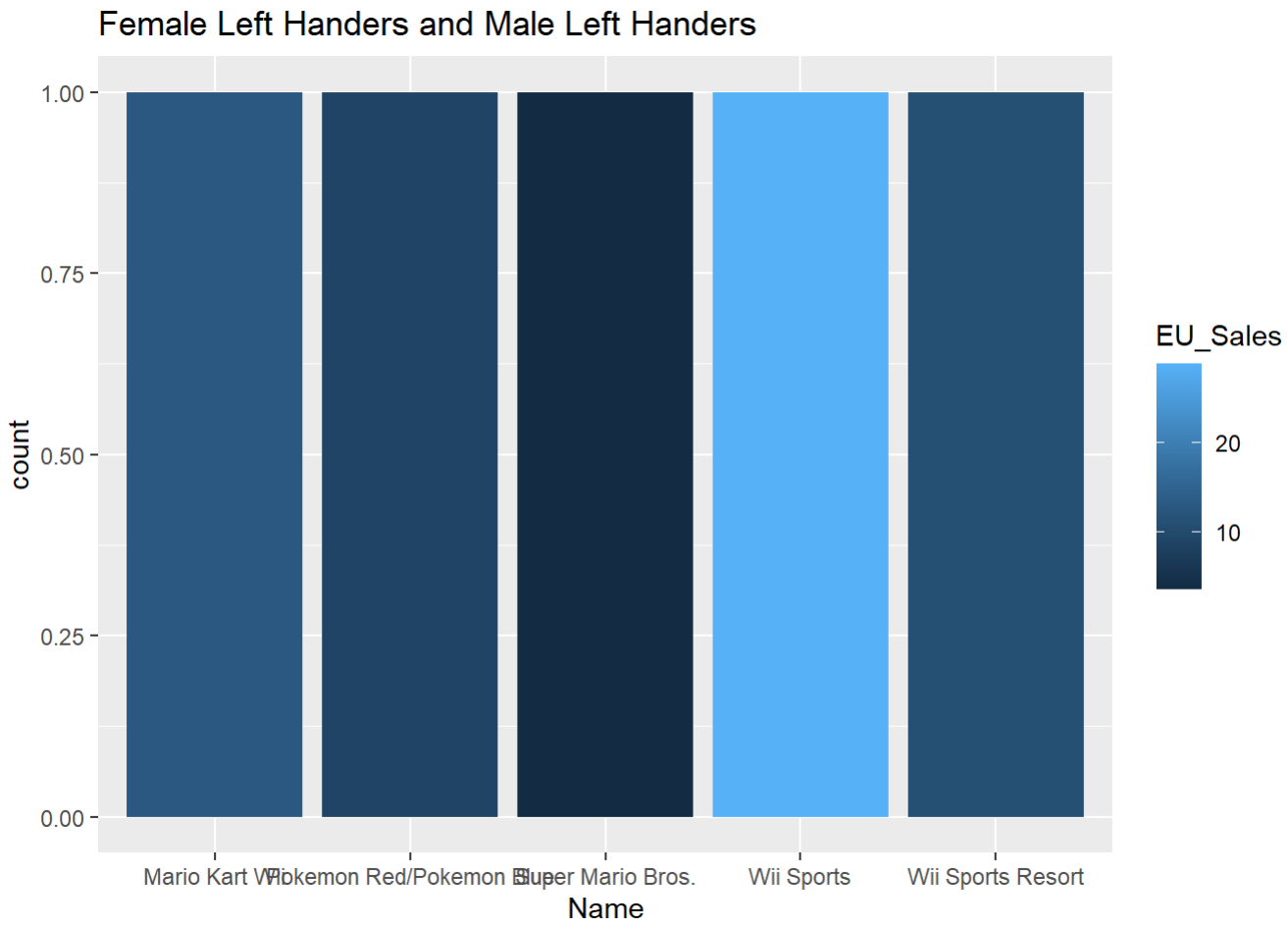
```
ggplot(data=df4,mapping=aes(x=Name, fill=NA_Sales))+geom_bar() + ggtitle("Female Left Handers and Male Left Handers")
```




```
df4 = subset(df, select=c(Name,EU_Sales))
df4<-head(df4,5)
df4<-df4 %>%
  group_by(Name)%>%
  arrange(desc(EU_Sales))
df4
```

```
## # A tibble: 5 x 2
## # Groups:   Name [5]
##   Name                      EU_Sales
##   <chr>                        <dbl>
## 1 Wii Sports                    29.0
## 2 Mario Kart Wii                12.9
## 3 Wii Sports Resort             11.0
## 4 Pokemon Red/Pokemon Blue       8.89
## 5 Super Mario Bros.              3.58
```

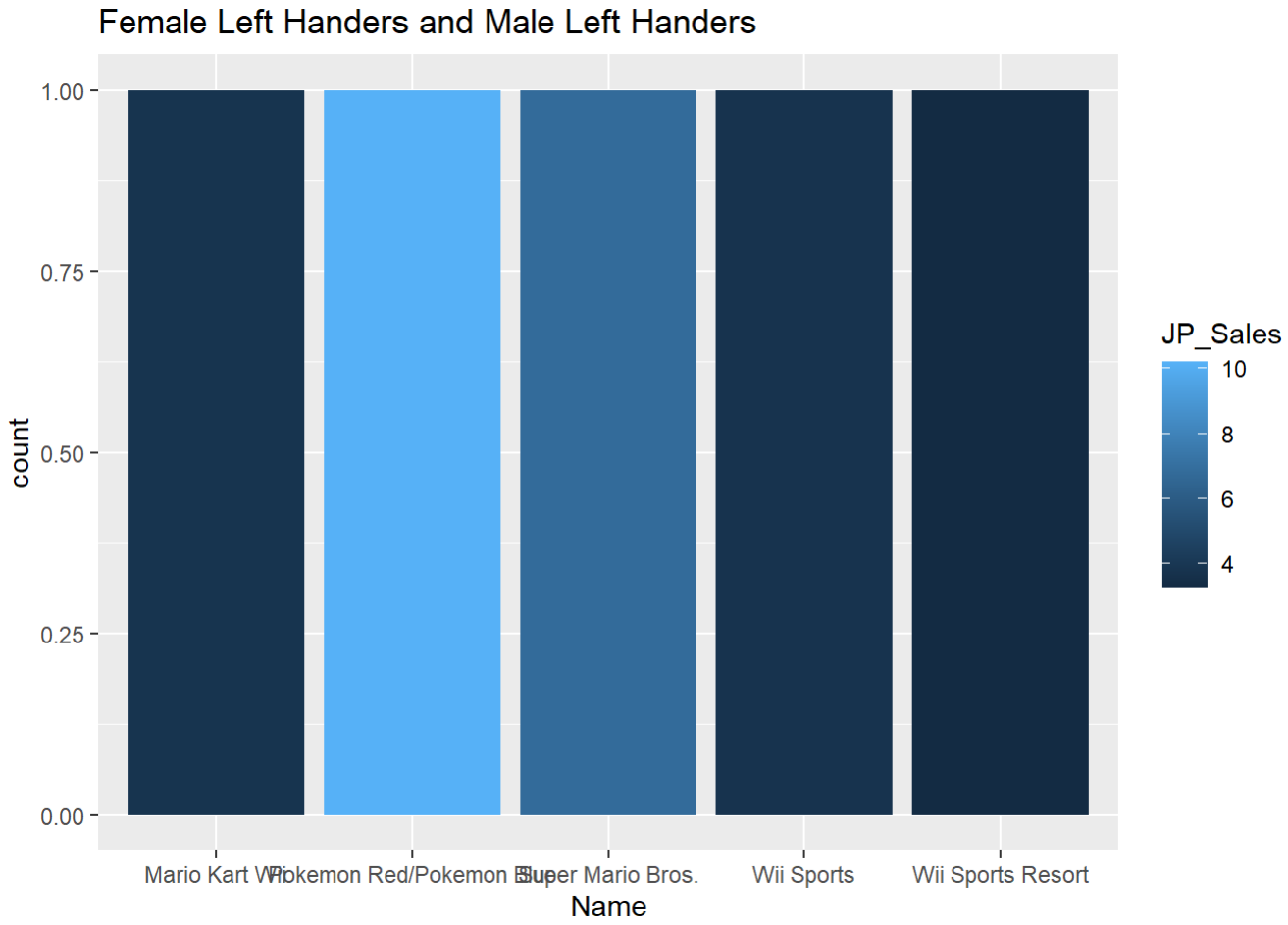
```
ggplot(data=df4,mapping=aes(x=Name, fill=EU_Sales))+geom_bar() + ggtitle("Female Left Handers an
d Male Left Handers")
```




```
df4 = subset(df, select=c(Name,JP_Sales))
df4<-head(df4,5)
df4<-df4 %>%
  group_by(Name)%>%
  arrange(desc(JP_Sales))
df4
```

```
## # A tibble: 5 x 2
## # Groups:   Name [5]
##   Name                     JP_Sales
##   <chr>                       <dbl>
## 1 Pokemon Red/Pokemon Blue    10.2
## 2 Super Mario Bros.            6.81
## 3 Mario Kart Wii               3.79
## 4 Wii Sports                   3.77
## 5 Wii Sports Resort            3.28
```

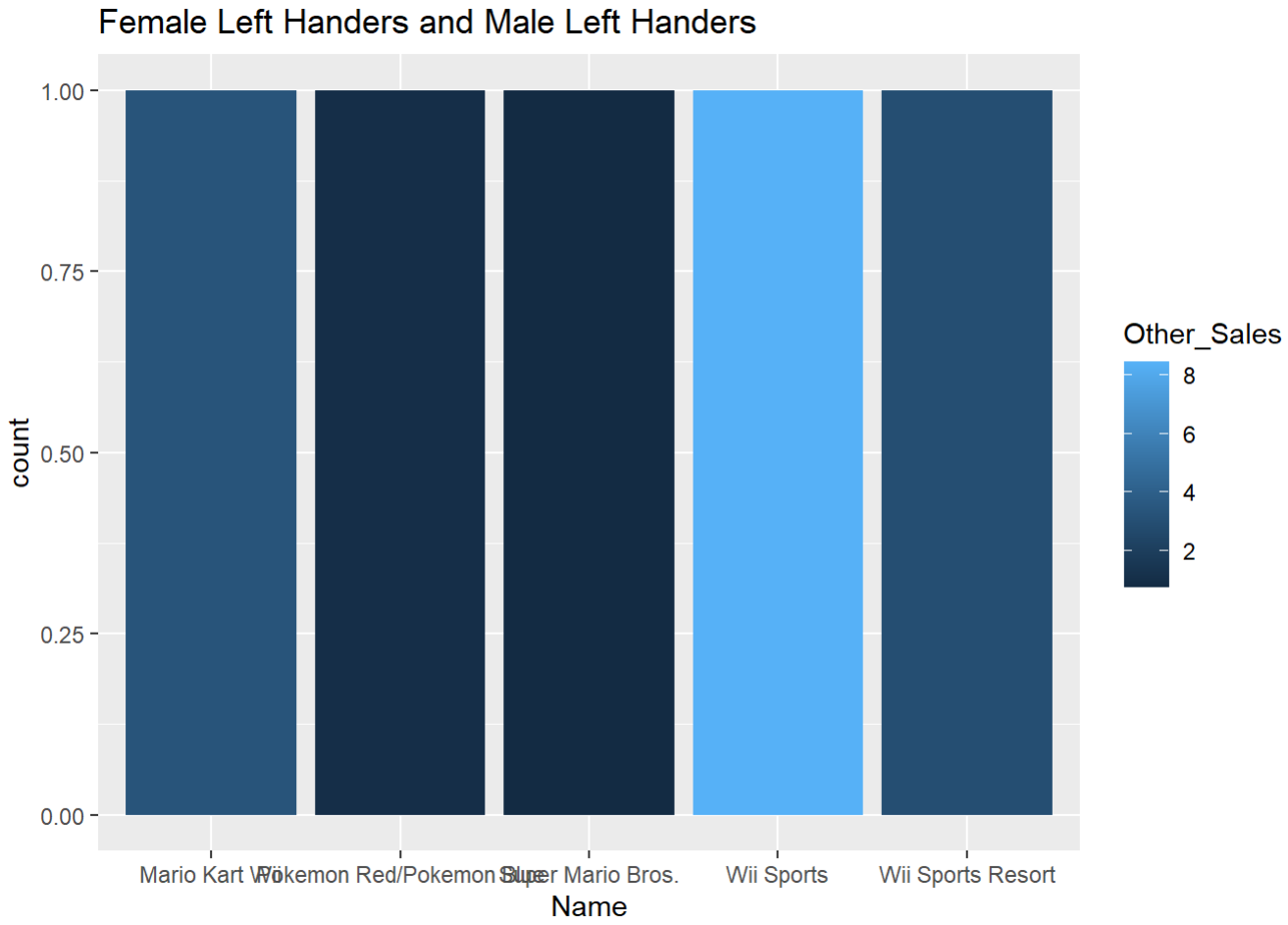
```
ggplot(data=df4,mapping=aes(x=Name, fill=JP_Sales))+geom_bar() + ggtitle("Female Left Handers an
d Male Left Handers")
```




```
df4 = subset(df, select=c(Name,Other_Sales))
df4<-head(df4,5)
df4<-df4 %>%
  group_by(Name)%>%
  arrange(desc(Other_Sales))
df4
```

```
## # A tibble: 5 x 2
## # Groups:   Name [5]
##   Name                      Other_Sales
##   <chr>                           <dbl>
## 1 Wii Sports                       8.46
## 2 Mario Kart Wii                   3.31
## 3 Wii Sports Resort                2.96
## 4 Pokemon Red/Pokemon Blue         1
## 5 Super Mario Bros.                0.77
```

```
ggplot(data=df4,mapping=aes(x=Name, fill=Other_Sales))+geom_bar() + ggtitle("Female Left Handers and Male Left Handers")
```



The graph shows us the top games preferred by users in different regions and also globally. We observe the following:

Wii Sports has been the top game in North America, Europe, other regions.

Pokemon Red/Pokemon Blue is the top game in Japan.

5. Are there any games with release year older than 2000 that are still making high sales? What are they?

```
old_games = filter(df, Year<2000)
head(old_games)
```

```
##   Rank                         Name Platform Year        Genre Publisher
## 1    2           Super Mario Bros.      NES 1985     Platform  Nintendo
## 2    5    Pokemon Red/Pokemon Blue       GB 1996 Role-Playing  Nintendo
## 3    6                      Tetris       GB 1989       Puzzle  Nintendo
## 4   10                   Duck Hunt      NES 1984      Shooter  Nintendo
## 5   13 Pokemon Gold/Pokemon Silver       GB 1999 Role-Playing  Nintendo
## 6   19           Super Mario World     SNES 1990     Platform  Nintendo
##   NA_Sales EU_Sales JP_Sales Other_Sales Global_Sales
## 1    29.08     3.58     6.81        0.77        40.24
## 2    11.27     8.89    10.22        1.00        31.37
## 3    23.20     2.26     4.22        0.58        30.26
## 4    26.93     0.63     0.28        0.47        28.31
## 5     9.00     6.18     7.20        0.71        23.10
## 6    12.78     3.75     3.54        0.55        20.61
```

```
a = old_games$Global_Sales
quantile(a, c(.99))
```

```
##    99%
## 7.8235
```

```
filter(old_games, Global_Sales>7.8235)
```

```
##    Rank                                    Name Platform Year        Genre
## 1     2                     Super Mario Bros.      NES 1985     Platform
## 2     5                Pokemon Red/Pokemon Blue       GB 1996 Role-Playing
## 3     6                                  Tetris       GB 1989       Puzzle
## 4    10                              Duck Hunt      NES 1984      Shooter
## 5    13             Pokemon Gold/Pokemon Silver       GB 1999 Role-Playing
## 6    19                       Super Mario World     SNES 1990     Platform
## 7    22                        Super Mario Land       GB 1989     Platform
## 8    23                    Super Mario Bros. 3      NES 1988     Platform
## 9    31 PokÃ©mon Yellow: Special Pikachu Edition       GB 1998 Role-Playing
## 10   47                          Super Mario 64      N64 1996     Platform
## 11   51       Super Mario Land 2: 6 Golden Coins       GB 1992    Adventure
## 12   53                            Gran Turismo       PS 1997       Racing
## 13   58                   Super Mario All-Stars     SNES 1993     Platform
## 14   64                           Mario Kart 64      N64 1996       Racing
## 15   67                      Final Fantasy VII       PS 1997 Role-Playing
## 16   70                         Gran Turismo 2       PS 1999       Racing
## 17   72                    Donkey Kong Country     SNES 1994     Platform
## 18   77                       Super Mario Kart     SNES 1992       Racing
## 19   85                          GoldenEye 007      N64 1997      Shooter
## 20   88                     Final Fantasy VIII       PS 1999 Role-Playing
##                     Publisher NA_Sales EU_Sales JP_Sales Other_Sales
## 1                    Nintendo    29.08     3.58     6.81        0.77
## 2                    Nintendo    11.27     8.89    10.22        1.00
## 3                    Nintendo    23.20     2.26     4.22        0.58
## 4                    Nintendo    26.93     0.63     0.28        0.47
## 5                    Nintendo     9.00     6.18     7.20        0.71
## 6                    Nintendo    12.78     3.75     3.54        0.55
## 7                    Nintendo    10.83     2.71     4.18        0.42
## 8                    Nintendo     9.54     3.44     3.84        0.46
## 9                    Nintendo     5.89     5.04     3.12        0.59
## 10                   Nintendo     6.91     2.85     1.91        0.23
## 11                   Nintendo     6.16     2.04     2.69        0.29
## 12 Sony Computer Entertainment     4.02     3.87     2.54        0.52
## 13                   Nintendo     5.99     2.15     2.12        0.29
## 14                   Nintendo     5.55     1.94     2.23        0.15
## 15 Sony Computer Entertainment     3.01     2.47     3.28        0.96
## 16 Sony Computer Entertainment     3.88     3.42     1.69        0.50
## 17                   Nintendo     4.36     1.71     3.00        0.23
## 18                   Nintendo     3.54     1.24     3.81        0.18
## 19                   Nintendo     5.80     2.01     0.13        0.15
## 20                 SquareSoft     2.28     1.72     3.63        0.23
##    Global_Sales
## 1         40.24
## 2         31.37
## 3         30.26
## 4         28.31
## 5         23.10
## 6         20.61
## 7         18.14
## 8         17.28
## 9         14.64
```

```
## 10        11.89
## 11        11.18
## 12        10.95
## 13        10.55
## 14         9.87
## 15         9.72
## 16         9.49
## 17         9.30
## 18         8.76
## 19         8.09
## 20         7.86
```

```
genre_df = subset(df, select=c(Genre, Global_Sales)) %>%
  arrange(desc(Global_Sales))
head(genre_df)
```

```
##          Genre Global_Sales
## 1       Sports        82.74
## 2     Platform        40.24
## 3       Racing        35.82
## 4       Sports        33.00
## 5 Role-Playing        31.37
## 6       Puzzle        30.26
```

```
df %>%
  group_by(Genre) %>%
    summarize(sum(Global_Sales))
```

```
## # A tibble: 12 x 2
##    Genre        `sum(Global_Sales)`
##    <chr>                      <dbl>
##  1 Action                     1751.
##  2 Adventure                   239.
##  3 Fighting                    449.
##  4 Misc                        810.
##  5 Platform                    831.
##  6 Puzzle                      245.
##  7 Racing                      732.
##  8 Role-Playing                927.
##  9 Shooter                    1037.
## 10 Simulation                  392.
## 11 Sports                    1331.
## 12 Strategy                   175.
```

```
genre_dff = df %>%
    dplyr::select(Genre, Global_Sales) %>%
      group_by(Genre) %>%
        summarize(sum(Global_Sales))
genre_dff
```

```
## # A tibble: 12 x 2
##    Genre        `sum(Global_Sales)`
##    <chr>                      <dbl>
##  1 Action                     1751.
##  2 Adventure                   239.
##  3 Fighting                    449.
##  4 Misc                        810.
##  5 Platform                    831.
##  6 Puzzle                      245.
##  7 Racing                      732.
##  8 Role-Playing                927.
##  9 Shooter                    1037.
## 10 Simulation                  392.
## 11 Sports                    1331.
## 12 Strategy                   175.
```

```
genre_dff
```

```
## # A tibble: 12 x 2
##    Genre        `sum(Global_Sales)`
##    <chr>                      <dbl>
##  1 Action                     1751.
##  2 Adventure                   239.
##  3 Fighting                    449.
##  4 Misc                        810.
##  5 Platform                    831.
##  6 Puzzle                      245.
##  7 Racing                      732.
##  8 Role-Playing                927.
##  9 Shooter                    1037.
## 10 Simulation                  392.
## 11 Sports                    1331.
## 12 Strategy                   175.
```

```
#ggplot(data=genre_dff,mapping=aes(x=Genre, fill=Global_Sales)) + geom_bar() + ggtitle("Bar Plot of Global Sales Genre wise") + theme(axis.text.x = element_text(angle = 90))
```