Video Game Analysis

Dinesh

January 3, 2018

My project contains a list of video games and sales data are analysed for more than 16,500 games. It also predicts in which year the video game was released, the publisher name and the global sales of the video game.

Setting working directory

```
setwd("C:/Users/Dinesh/MyFolder")
getwd()
```

```
## [1] "C:/Users/Dinesh/MyFolder"
```

Reading the dataset and visualizing the length and breadth of the dataset

```
video.df <- read.csv(paste("vgsales.csv", sep=""))
View(video.df)
dim(video.df)
```

```
## [1] 16598    11
```

Descriptive statistics of each variable

```
library(psych)
describe(video.df)
```

```
##              vars     n    mean      sd  median trimmed     mad  min
## Rank            1 16598 8300.61 4791.85 8300.50 8300.56 6152.05 1.00
## Name*           2 16598 5795.86 3324.01 5864.50 5810.22 4270.63 1.00
## Platform*       3 16598   16.71    8.29   17.00   16.67   10.38 1.00
## Year*           4 16598   27.61    6.00   28.00   28.00    5.93 1.00
## Genre*          5 16598    5.93    3.76    6.00    5.86    5.93 1.00
## Publisher*      6 16598  299.40  181.98  329.00  303.97  272.80 1.00
## NA_Sales        7 16598    0.26    0.82    0.08    0.13    0.12 0.00
## EU_Sales        8 16598    0.15    0.51    0.02    0.06    0.03 0.00
## JP_Sales        9 16598    0.08    0.31    0.00    0.02    0.00 0.00
## Other_Sales    10 16598    0.05    0.19    0.01    0.02    0.01 0.00
## Global_Sales   11 16598    0.54    1.56    0.17    0.27    0.21 0.01
##                  max    range  skew kurtosis    se
## Rank         16600.00 16599.00  0.00    -1.20 37.19
## Name*        11493.00 11492.00 -0.03    -1.21 25.80
## Platform*       31.00    30.00 -0.05    -1.00  0.06
```

```
## Year*           40.00    39.00 -0.86     1.68  0.05
## Genre*          12.00    11.00  0.07    -1.43  0.03
## Publisher*     579.00   578.00 -0.15    -1.40  1.41
## NA_Sales        41.49    41.49 18.80   648.86  0.01
## EU_Sales        29.02    29.02 18.87   755.71  0.00
## JP_Sales        10.22    10.22 11.20   194.15  0.00
## Other_Sales     10.57    10.57 24.23  1024.92  0.00
## Global_Sales    82.74    82.73 17.40   603.68  0.01
```

To find factors in the dataset

```
str(video.df)
```

```
## 'data.frame':    16598 obs. of  11 variables:
##  $ Rank        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name        : Factor w/ 11493 levels "'98 Koshien",..: 10991 9343 5532 10993 7370 9707 6648 10989
##  $ Platform    : Factor w/ 31 levels "2600","3DO","3DS",..: 26 12 26 26 6 6 5 26 26 12 ...
##  $ Year        : Factor w/ 40 levels "1980","1981",..: 27 6 29 30 17 10 27 27 30 5 ...
##  $ Genre       : Factor w/ 12 levels "Action","Adventure",..: 11 5 7 11 8 6 5 4 5 9 ...
##  $ Publisher   : Factor w/ 579 levels "10TACLE Studios",..: 369 369 369 369 369 369 369 369 369 369
##  $ 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 ...
```

From the above output, the categorical variables are Name, Platform, Year, Genre and Publisher.

Total Sales in a year for video game

```
tot <- aggregate(video.df$Global_Sales, by=list(Year=video.df$Year), sum)
tot
```

```
##    Year      x
## 1  1980  11.38
## 2  1981  35.77
## 3  1982  28.86
## 4  1983  16.79
## 5  1984  50.36
## 6  1985  53.94
## 7  1986  37.07
## 8  1987  21.74
## 9  1988  47.22
## 10 1989  73.45
## 11 1990  49.39
## 12 1991  32.23
## 13 1992  76.16
## 14 1993  45.98
## 15 1994  79.17
## 16 1995  88.11
## 17 1996 199.15
```

```
## 18 1997 200.98
## 19 1998 256.47
## 20 1999 251.27
## 21 2000 201.56
## 22 2001 331.47
## 23 2002 395.52
## 24 2003 357.85
## 25 2004 419.31
## 26 2005 459.94
## 27 2006 521.04
## 28 2007 611.13
## 29 2008 678.90
## 30 2009 667.30
## 31 2010 600.45
## 32 2011 515.99
## 33 2012 363.54
## 34 2013 368.11
## 35 2014 337.05
## 36 2015 264.44
## 37 2016  70.93
## 38 2017   0.05
## 39 2020   0.29
## 40  N/A 100.08
```

Finding in which year, maximum and minimum global sale occured

Maximum global sales

```
a <- max(tot$x)
maxsales <- tot$Year[tot$x==a]
maxsales
```

```
## [1] 2008
## 40 Levels: 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 ... N/A
```

From the above output, we can conclude that the maximum global sales occured in the year 2008.

Minimum global sales

```
a <- min(tot$x)
maxsales1 <- tot$Year[tot$x==a]
maxsales1
```

```
## [1] 2017
## 40 Levels: 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 ... N/A
```

From the above output, we can conclude that the minimum global sales occured in the year 2017.

Global Sales for specific genre

```
tots <- aggregate(video.df$Global_Sales, by=list(Genre=video.df$Genre), sum)
tots
```

```
##           Genre       x
## 1        Action 1751.18
## 2     Adventure  239.04
## 3      Fighting  448.91
## 4          Misc  809.96
## 5      Platform  831.37
## 6        Puzzle  244.95
## 7        Racing  732.04
## 8  Role-Playing  927.37
## 9       Shooter 1037.37
## 10   Simulation  392.20
## 11       Sports 1330.93
## 12     Strategy  175.12
```

One way contingency table - for Genre

```
mytable <- with(video.df, table(Genre))
mytable
```

```
## Genre
##       Action    Adventure     Fighting         Misc     Platform
##         3316         1286          848         1739          886
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##          582         1249         1488         1310          867
##       Sports     Strategy
##         2346          681
```

Percentages for the above one way contingency table

```
mytable <- with(video.df, table(Genre))
prop.table(mytable)*100
```

```
## Genre
##       Action    Adventure     Fighting         Misc     Platform
##    19.978311     7.747921     5.109049    10.477166     5.337993
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##     3.506447     7.525003     8.964936     7.892517     5.223521
##       Sports     Strategy
##    14.134233     4.102904
```

One way contingency table - for Publisher

```
mytable <- with(video.df, table(Platform))
margin.table(mytable,1)
```

```
## Platform
## 2600  3DO  3DS   DC   DS   GB  GBA   GC  GEN   GG  N64  NES   NG   PC PCFX
##  133    3  509   52 2163   98  822  556   27    1  319   98   12  960    1
##   PS  PS2  PS3  PS4  PSP  PSV  SAT  SCD SNES TG16  Wii WiiU   WS X360   XB
## 1196 2161 1329  336 1213  413  173    6  239    2 1325  143    6 1265  824
## XOne
##  213
```

Two way contingency table for year vs genre

```
mytable <- xtabs(~Year+Genre, data=video.df)
margin.table(mytable,1)
```

```
## Year
## 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
##    9   46   36   17   14   14   21   16   15   17   16   41   43   60  121
## 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
##  219  263  289  379  338  349  482  829  775  763  941 1008 1202 1428 1431
## 2010 2011 2012 2013 2014 2015 2016 2017 2020  N/A
## 1259 1139  657  546  582  614  344    3    1  271
```

Chi-squared test for sales in northamerica vs.europe

```
mytable <- xtabs(~NA_Sales + EU_Sales, data=video.df)
mytable1 <- margin.table(mytable,1)
chisq.test(mytable1)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  mytable1
## X-squared = 566500, df = 408, p-value < 2.2e-16
```

Here p-value is less than 0.05. Therefore we reject the null hypothesis and conclude that they are independent

Fisher's exact test for sales in Japan vs. others

Fisher's exact test - An alternate test to chi-squared test when the sample size is small

```
mytable <- xtabs(~JP_Sales + Other_Sales, data=video.df)
mytable1 <- margin.table(mytable,1)
chisq.test(mytable1)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  mytable1
## X-squared = 1618700, df = 243, p-value < 2.2e-16
```

Here p-value is less than 0.05. Therefore we reject the null hypothesis and conclude that they are independent

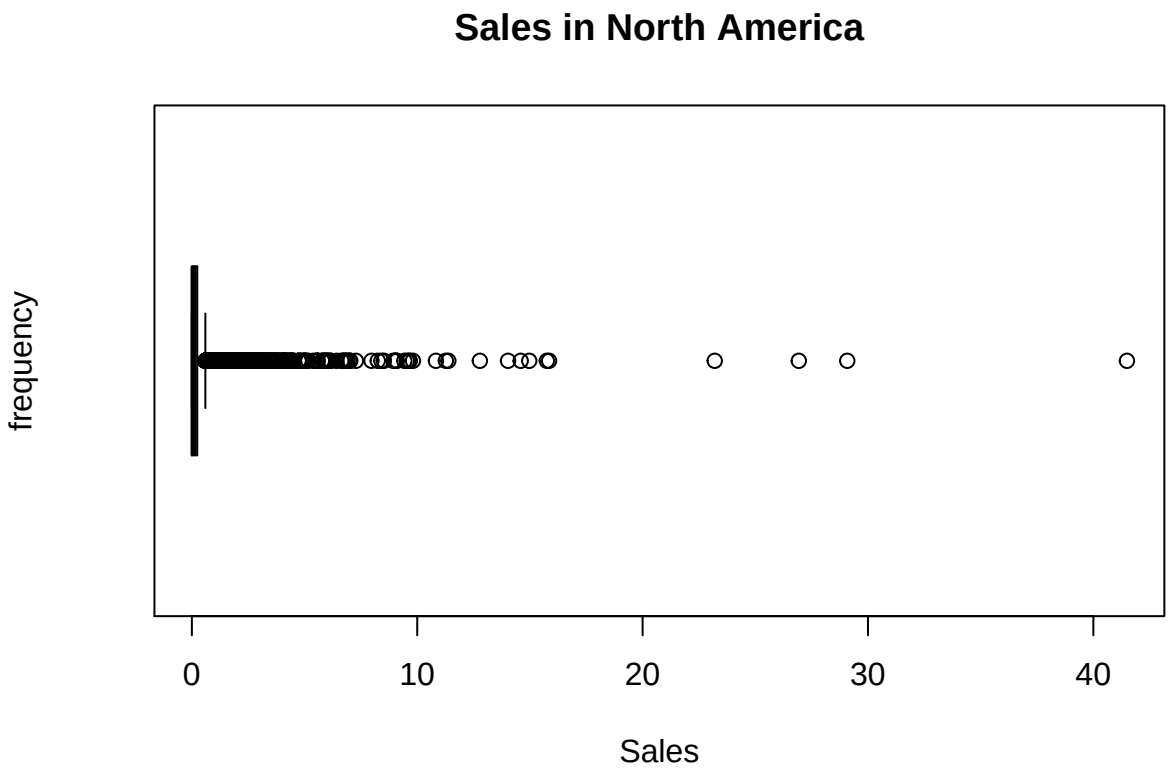
Two way contingency table for publisher vs genre

```
mytable <- xtabs(~Publisher+Genre, data=video.df)
head(mytable)
```

```
##                               Genre
## Publisher                      Action Adventure Fighting Misc Platform
##   10TACLE Studios                   0         1        0    0        0
##   1C Company                        0         0        0    0        0
##   20th Century Fox Video Games      4         0        0    0        0
##   2D Boy                            0         0        0    0        0
##   3DO                              17         3        1    0        1
##   49Games                           0         0        0    0        0
##                               Genre
## Publisher                      Puzzle Racing Role-Playing Shooter
##   10TACLE Studios                   1      0            0       0
##   1C Company                        0      1            1       0
##   20th Century Fox Video Games      0      0            0       1
##   2D Boy                            1      0            0       0
##   3DO                               1      0            1       5
##   49Games                           0      0            0       0
##                               Genre
## Publisher                      Simulation Sports Strategy
##   10TACLE Studios                       0      0        1
##   1C Company                            0      0        1
##   20th Century Fox Video Games          0      0        0
##   2D Boy                                0      0        0
##   3DO                                   0      6        1
##   49Games                               0      1        0
```

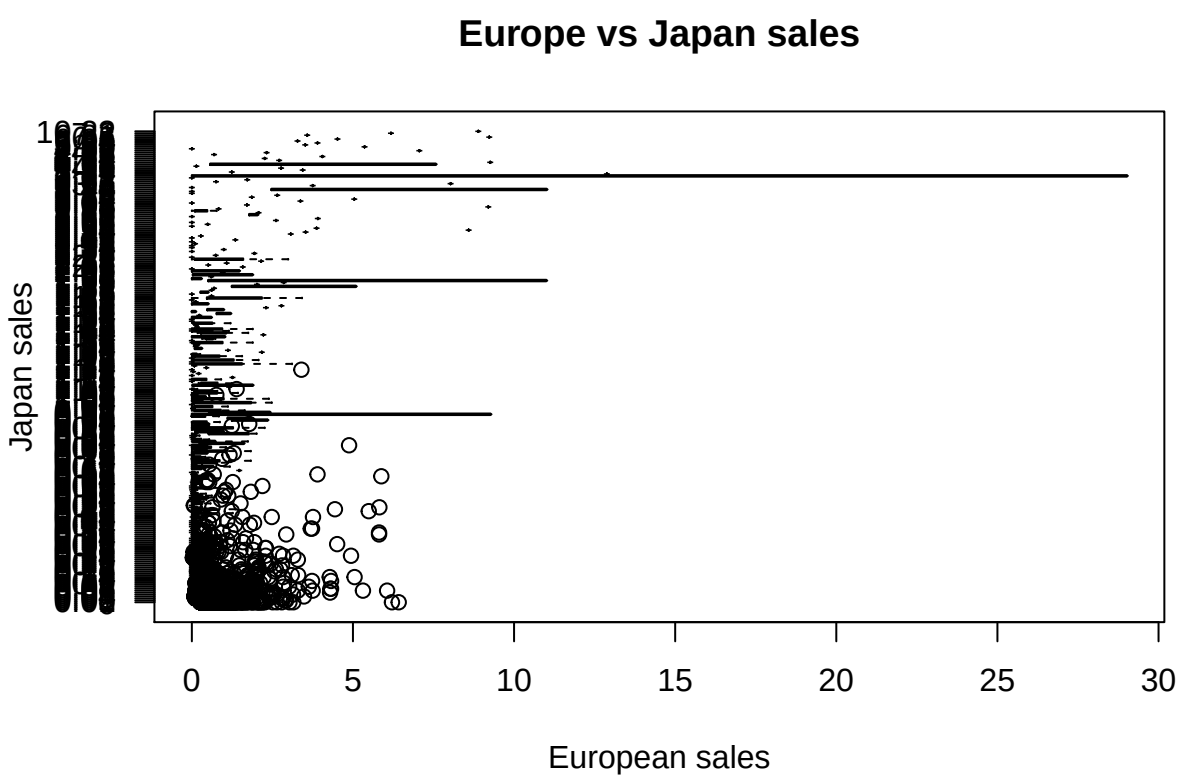
Boxplot for sales in North America (NA_Sales)

```
boxplot(video.df$NA_Sales, main="Sales in North America", xlab="Sales", ylab="frequency", horizontal = T
```



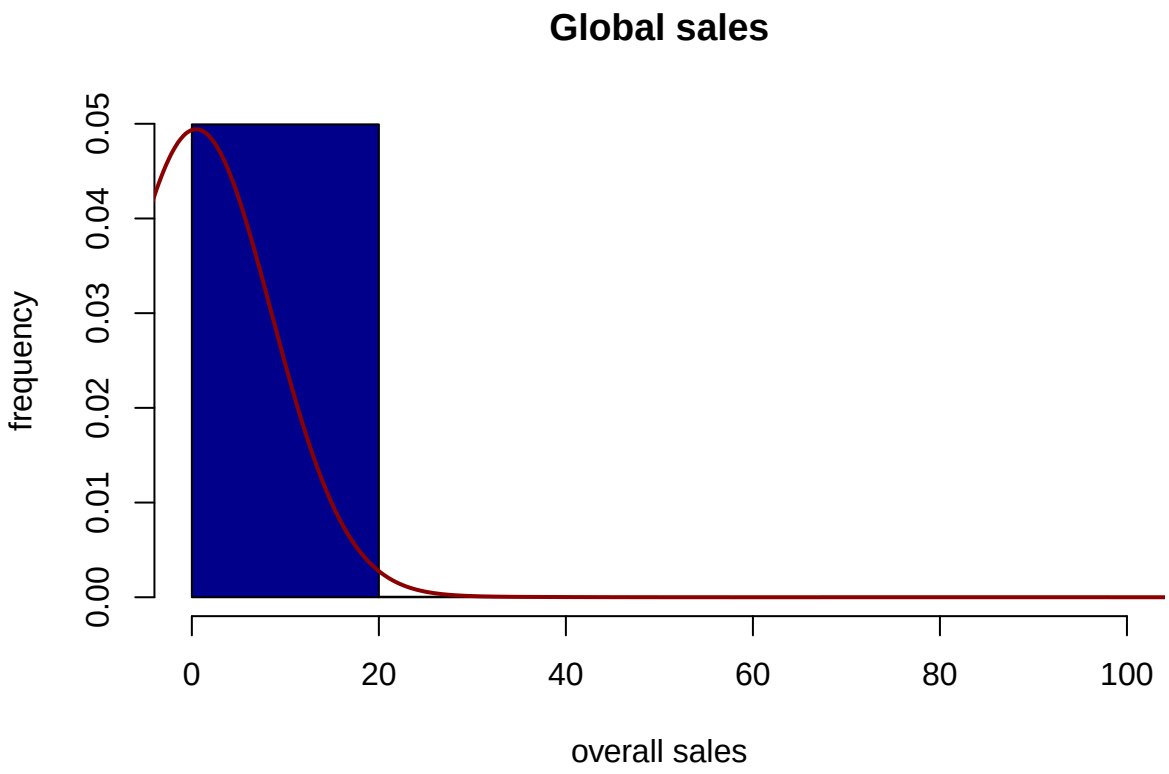
Boxplot for sales in europe vs japan

```
boxplot(video.df$EU_Sales ~ video.df$JP_Sales, horizontal=TRUE, xlab="European sales", ylab="Japan sales
```



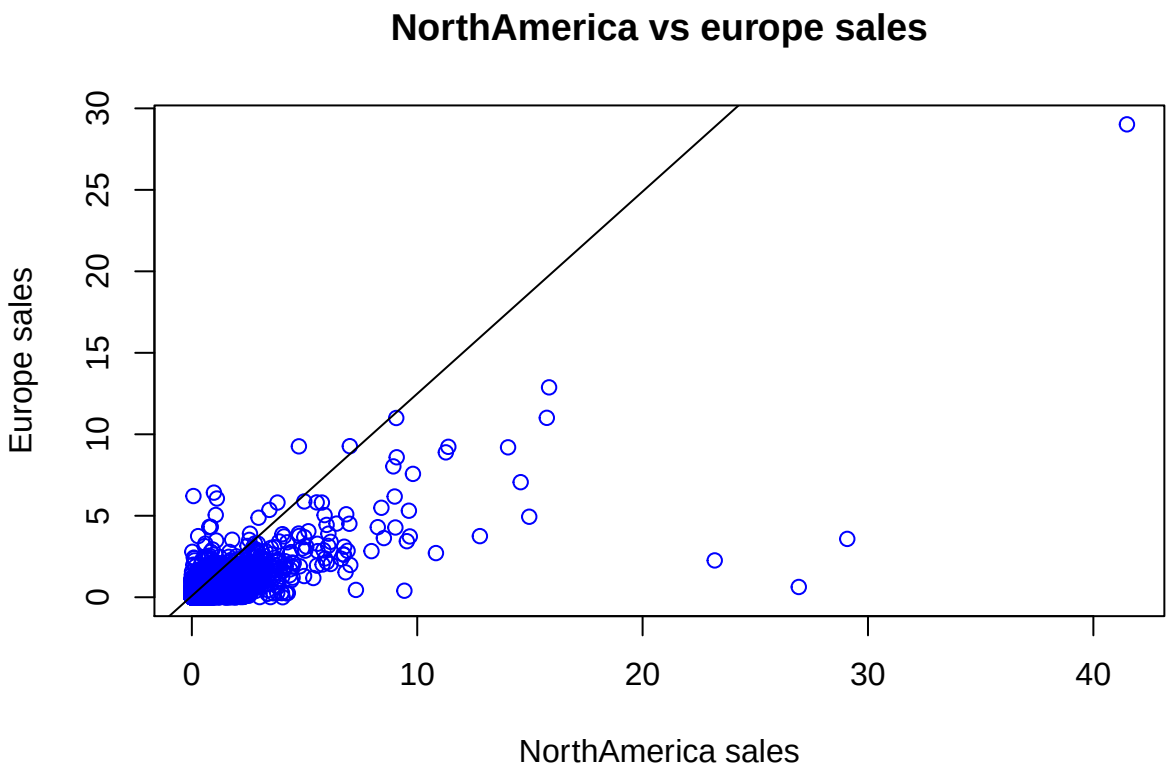
Histogram for Global sales

```
hist(video.df$Global_Sales, main="Global sales", xlab="overall sales", ylab="frequency", breaks=3, col=
lines(density(video.df$Global_Sales, bw=8), type="l", col="darkred", lwd=2)
```



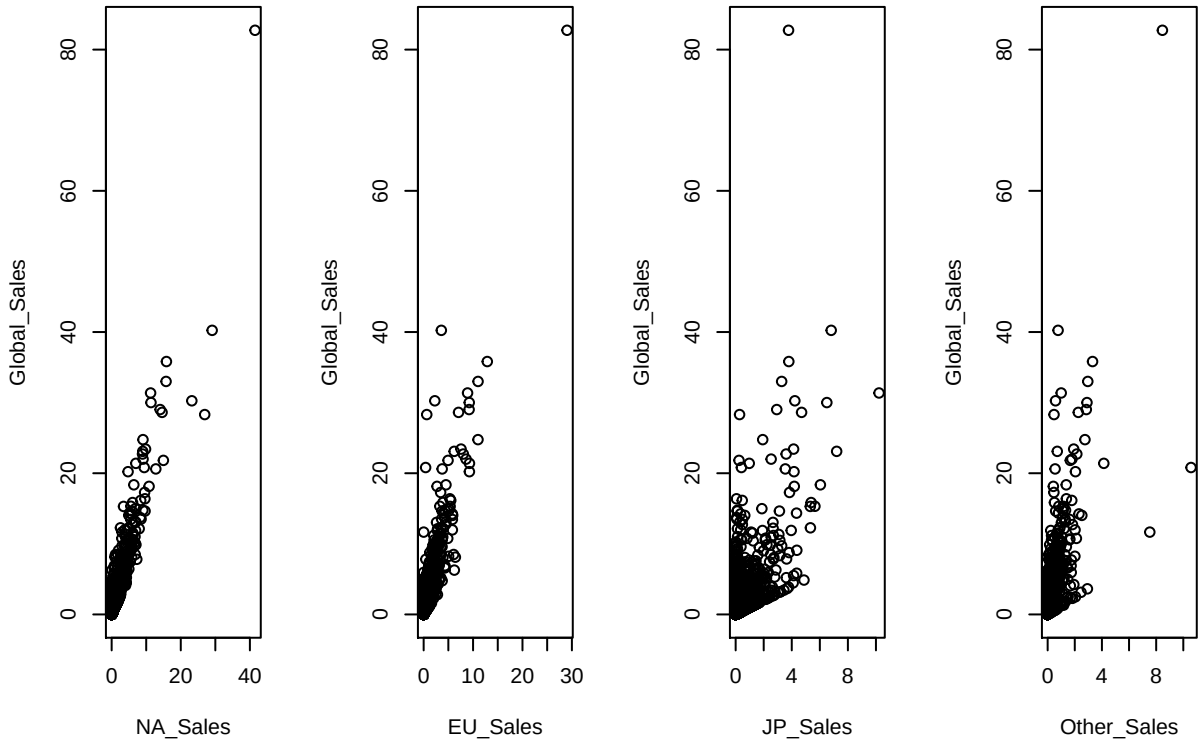
plotting north america sales vs europe sales

```
plot(video.df$NA_Sales, video.df$EU_Sales, col="blue", main="NorthAmerica vs europe sales", xlab="North
abline(lm(video.df$NA_Sales~video.df$EU_Sales))
```



plotting northamerica, europe, japan, other sales vs global sales

```
par(mfrow=c(1,4))
with(video.df, plot(NA_Sales, Global_Sales))
with(video.df, plot(EU_Sales, Global_Sales))
with(video.df, plot(JP_Sales, Global_Sales))
with(video.df, plot(Other_Sales, Global_Sales))
```




```
par(mfrow=c(1,1))
```

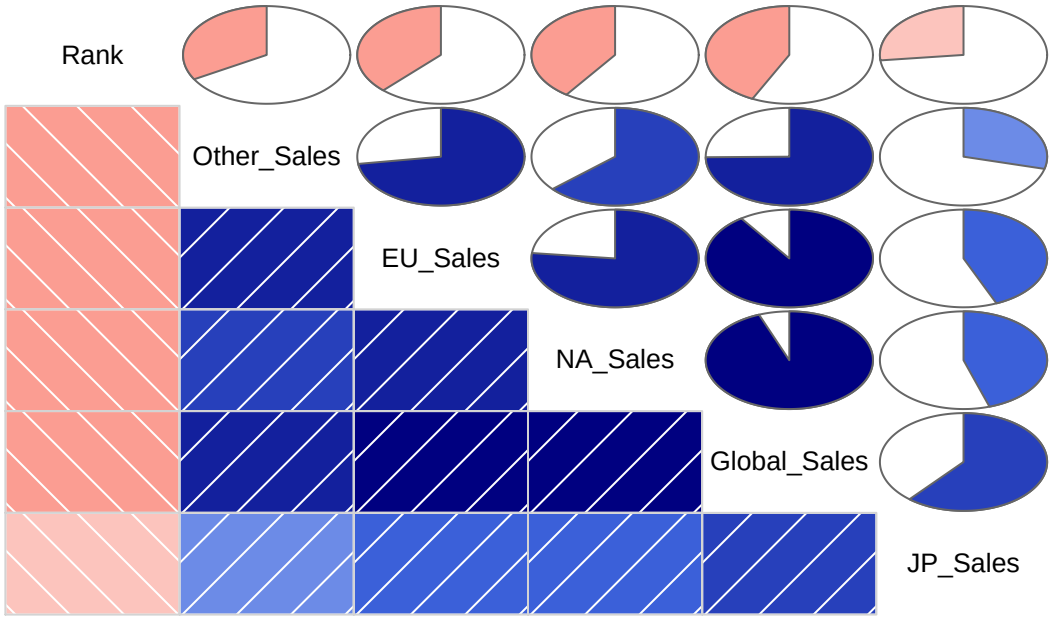
Creating correlation matrix for sales across different countries and overall sales

```
cor(video.df[, c(7:11)])
```

```
##                NA_Sales  EU_Sales  JP_Sales Other_Sales Global_Sales
## NA_Sales      1.0000000 0.7677267 0.4497874   0.6347373    0.9410474
## EU_Sales      0.7677267 1.0000000 0.4355845   0.7263849    0.9028358
## JP_Sales      0.4497874 0.4355845 1.0000000   0.2901862    0.6118155
## Other_Sales   0.6347373 0.7263849 0.2901862   1.0000000    0.7483308
## Global_Sales  0.9410474 0.9028358 0.6118155   0.7483308    1.0000000
```

Visualizing correlation matrix using corrgram

```
library(corrgram)
corrgram(video.df, order=TRUE, lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.t
```



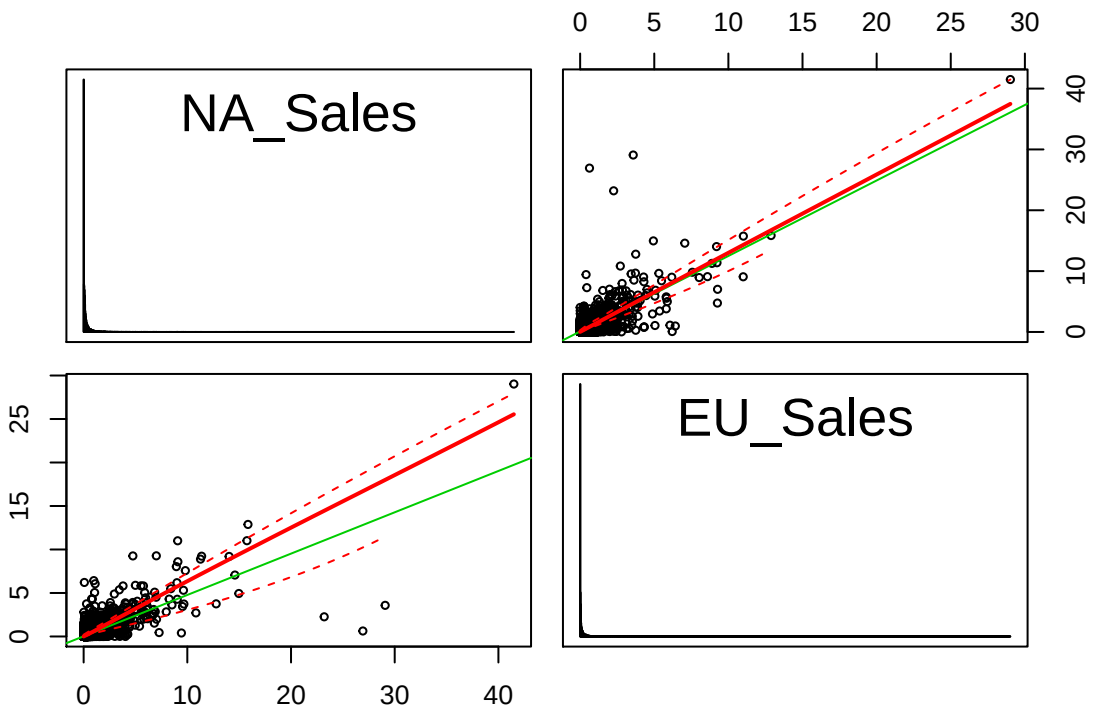
Scatterplot matrix for sales in northamerica and europe

```
library(car)
```

```
##
## Attaching package: 'car'
```

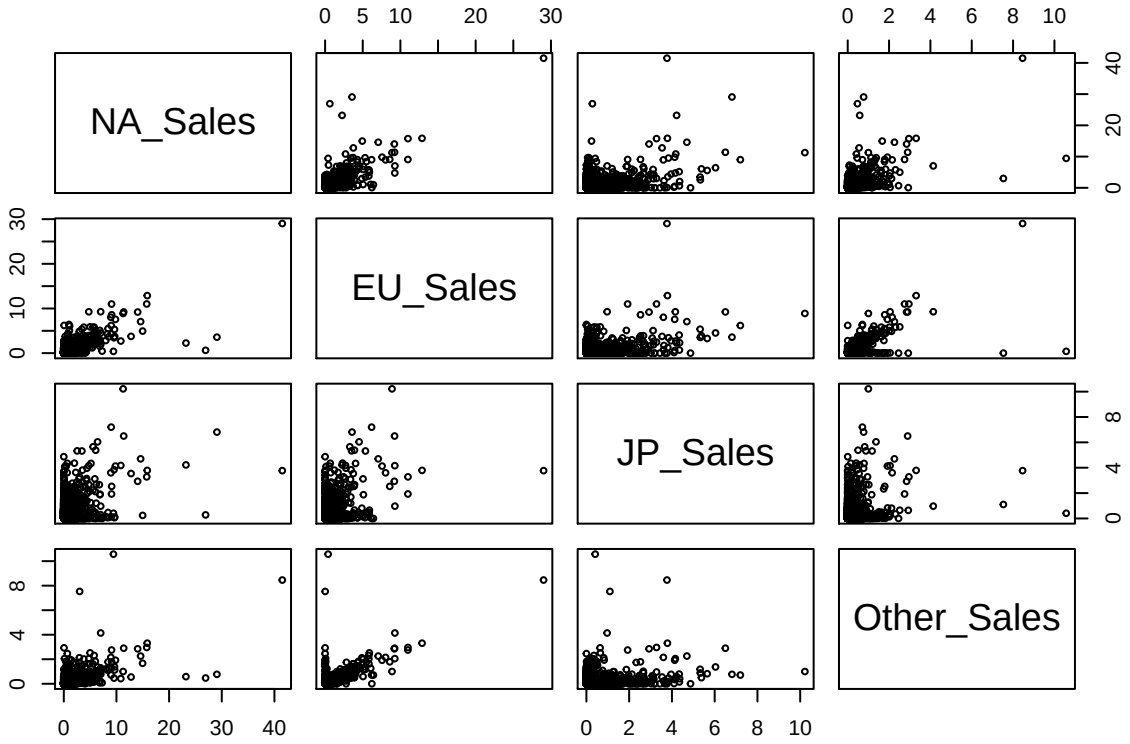
```
## The following object is masked from 'package:psych':
##
##     logit
```

```
scatterplotMatrix(formula = ~ NA_Sales + EU_Sales, cex=0.6, data=video.df, diagonal="histogram")
```



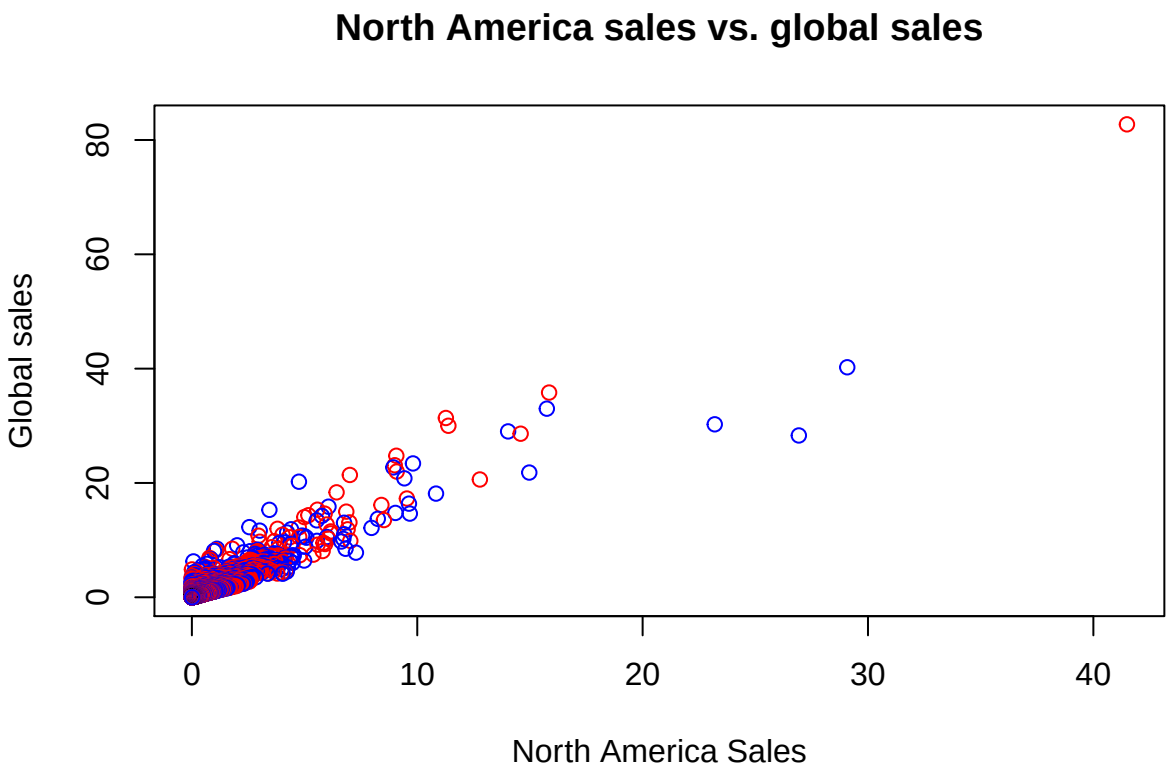
Scatterplot matrix for sales in northamerica, europe, japan and others

```
pairs(formula = ~ NA_Sales + EU_Sales + JP_Sales + Other_Sales, cex=0.6, data=video.df)
```



Jitter to plot year wise global sales

```
plot(jitter(video.df$NA_Sales), jitter(video.df$Global_Sales), xlab="North America Sales", ylab="Global
```



Test to check our hypothesis for suitable assumptions

Correlation test for northamerica sales vs european sales

```
cor.test(video.df$NA_Sales, video.df$EU_Sales)
```

```
##
##  Pearson's product-moment correlation
##
## data:  video.df$NA_Sales and video.df$EU_Sales
## t = 154.35, df = 16596, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7614064 0.7739012
## sample estimates:
##       cor
## 0.7677267
```

Since p-value is less than 0.05, they are independent variables

Correlation test for japan sales and other sales

```
cor.test(video.df$JP_Sales, video.df$Other_Sales)
```

```
##
##  Pearson's product-moment correlation
##
## data:  video.df$JP_Sales and video.df$Other_Sales
## t = 39.064, df = 16596, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2761922 0.3040573
## sample estimates:
##       cor
## 0.2901862
```

Since p-value is less than 0.05, they are independent variables

T-test to analyse the hypothesis

Hypothesis : The sales in northamerica is greater than that of europe

```
t.test(video.df$NA_Sales, video.df$EU_Sales)
```

```
##
##  Welch Two Sample t-test
##
## data:  video.df$NA_Sales and video.df$EU_Sales
## t = 15.831, df = 27682, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1034042 0.1326267
## sample estimates:
## mean of x mean of y
## 0.2646674 0.1466520
```

Null hypothesis : Two means must be equal. So we reject the null hypothesis and since p-value is less than 0.05, the variables are statistically significant.

Hypothesis : The sales in japan is less than that of europe

```
t.test(video.df$JP_Sales, video.df$EU_Sales)
```

```
##
##  Welch Two Sample t-test
##
## data:  video.df$JP_Sales and video.df$EU_Sales
## t = -14.976, df = 27501, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07788435 -0.05985634
## sample estimates:
##  mean of x  mean of y
## 0.07778166 0.14665201
```

Null hypothesis : Two means must be equal. So we reject the null hypothesis and since p-value is less than 0.05, the variables are statistically significant.

Formulating regression model

Model 1: Global sales vs. North America Sales

```
fit <- lm(Global_Sales ~ NA_Sales, data=video.df)
summary(fit)
```

```
##
## Call:
## lm(formula = Global_Sales ~ NA_Sales, data = video.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0071  -0.1086  -0.0528   0.0172  11.6456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.063202   0.004292   14.72   <2e-16 ***
## NA_Sales    1.791827   0.005000  358.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.526 on 16596 degrees of freedom
## Multiple R-squared:  0.8856, Adjusted R-squared:  0.8856
## F-statistic: 1.284e+05 on 1 and 16596 DF,  p-value: < 2.2e-16
```

From the above model, p-value = less than 0.001 Multiple r-squared and adjusted r-squared = 0.8856

Model 2: Global sales vs. European sales

```
fit <- lm(Global_Sales ~ EU_Sales, data=video.df)
summary(fit)
```

```
##
## Call:
## lm(formula = Global_Sales ~ EU_Sales, data = video.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1023  -0.1378  -0.0856   0.0274  30.1642
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.130021   0.005404   24.06   <2e-16 ***
## EU_Sales    2.778137   0.010271  270.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6687 on 16596 degrees of freedom
## Multiple R-squared:  0.8151, Adjusted R-squared:  0.8151
## F-statistic: 7.317e+04 on 1 and 16596 DF,  p-value: < 2.2e-16
```

From the above model, p-value = less than 0.001 Multiple r-squared and adjusted r-squared = 0.8151

Model 3: Global sales vs. Japan sales

```
fit <- lm(Global_Sales ~ JP_Sales, data=video.df)
summary(fit)
```

```
##
## Call:
## lm(formula = Global_Sales ~ JP_Sales, data = video.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.408  -0.319  -0.198   0.062  70.845
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.298181   0.009845   30.29   <2e-16 ***
## JP_Sales    3.076039   0.030871   99.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.23 on 16596 degrees of freedom
## Multiple R-squared:  0.3743, Adjusted R-squared:  0.3743
## F-statistic:  9929 on 1 and 16596 DF,  p-value: < 2.2e-16
```

From the above model, p-value = less than 0.01 Multiple r-squared and adjusted r-squared = 0.3743

The best model is chosen based upon two conditions: 1. The model must have lesser p-value than all. 2. It must also have higher multiple r-squared or adjusted r-squared value.

The model which satisfies the above two conditions is Model 3. Therefore, this is the best model of all.

To find beta-coefficients in a fitted model

Model 1: Global sales vs. North America Sales

```
fit <- lm(Global_Sales ~ NA_Sales, data=video.df)
fit$coefficients
```

```
## (Intercept)    NA_Sales
##  0.06320234  1.79182728
```

Global Sales(Y) = b0 + NorthAmerica Sales(b1) b0=-1, b1=1.7918 Global Sales = -1 + NorthAmerica Sales*1.7918

Model 2: Global sales vs. European sales

```
fit <- lm(Global_Sales ~ EU_Sales, data=video.df)
fit$coefficients
```

```
## (Intercept)    EU_Sales
##   0.1300213   2.7781369
```

Global Sales(Y) = b0 + European Sales(b1) b0=-1, b1=2.7781 Global Sales = -1 + European Sales*2.7781

Model 3: Global sales vs. Japan sales

```
fit <- lm(Global_Sales ~ JP_Sales, data=video.df)
fit$coefficients
```

```
## (Intercept)    JP_Sales
##   0.2981812   3.0760394
```

Global Sales(Y) = b0 + Japan Sales(b1) b0=-1, b1=3.0760 Global Sales = -1 + Japan Sales*3.0760

Model 4: Global Sales vs. other sales

```
fit <- lm(Global_Sales ~ Other_Sales + NA_Sales + EU_Sales + JP_Sales, data=video.df)
fit$coefficients
```

```
##  (Intercept)  Other_Sales     NA_Sales     EU_Sales     JP_Sales
## 0.0003229497 0.9995874903 0.9999405824 0.9999875831 0.9998838156
```

Global Sales(Y) = b0 + Other Sales(b1) + NorthAmerica Sales(b2) + European Sales(b3) + Japan sales(b4) b0=-1, b1=0.9995, b2=0.99994, b3=0.99998, b4=0.99988 Global Sales = -1 + Other Sales(0.9995) + NorthAmerica Sales(0.99994) + European Sales(0.99998) + Japan Sales(0.99988).

Fitted residuals and values are checked and the deviation was around 1000. because of large data points it's not suitable to show those in the output file

Conclusion:

From the above outputs, we have found out 1. In which year the global sales was high,low. 2. Comparisons between sales in northamerica, europe, japan and others 3. Genre which had the highest and lowest global sales.