b

MH3511 Data Analysis with Computer - Group Project

Han Jun (U1820665L)
Hu Zhuangyu (U1821776L)
Jiang Lingling (U1840036A)
Laura Lit Pei Lin (U1821546A)
Tao Weijing (U1820897C)
Zeng Shijia (U1840980G)

Project Title: Analysis of Video game sales and its factors

Abstract:

Our study provides information on the video game and gaming industry, from data on the factors affecting the global sales in the business to statistics on the output and consumption of games. A study has found that Singaporeans are the most frequent gamers in Asia, and the third most frequent in the world, spending a total of 7.44 hours each week playing video games (Sean, 2019). Forrest Li, the Singaporean founder of a gaming company is the world's second person to become a billionaire thanks to an online game (Chia, 2019). With the growth in the video games industry, our project aims to investigate how the various factors affect the global sales of video games. The analysis of data by performing various statistical methods such as variance test, t-test, Pearson Chi-squared test, construction of multiple linear regression model etcetera, allow us to scrutinize the correlations between the variables and the global sales of the video game. From the results, we are then able to identify the considerable factors which may be useful in increasing sales of video games for future ventures in the industry.

Content Page

1. Introduction

There have been many theories made showing correlation between global sales of video games and other associated conditions like the platform of the game. As the video games industry continues to evolve and the diversity of users continue to rise, there have been many new factors that affect global sales of video games. Hence, we would like to further examine the other potential contributing factors that may impact the sales of video games.

In our project, a dataset of video game sales globally with their related variables is used for analysis and testing to achieve these following objectives.

1. To investigate whether there is any association between global sales and year of release of video games
2. To investigate whether scores graded by two groups of people have the same mean
3. To determine whether different genres and different platforms will affect the mean of global sales
4. To determine whether there is a correlation between global sales of video games and the selected variables

2. Data Description

Our team obtained the dataset, from kaggle. The original dataset contains a dataframe of 16720 observations and 16 variables. This dataset is being used to determine any relationship between the variables and video games sales.

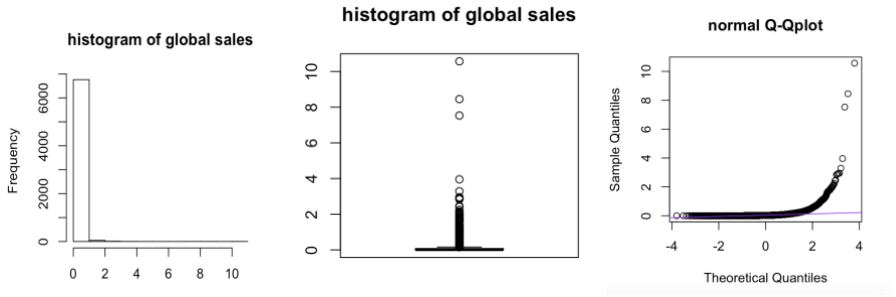
The dataset contains 16 variables: Name, Platform, Year_of_Release, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count, Developer and Rating. Out of all these, we will be using variables, Platform, Year_of_Release, Genre, Gobal_Sales, Critic_Score, Critic_Count, User_Score and User_Count. Details are in Appendix Table 1.

Some data are being cleaned and removed so that only numeric datasets are used in certain analysis. The summary of these numeric data is in Appendix Table 2.

3. Data Analysis

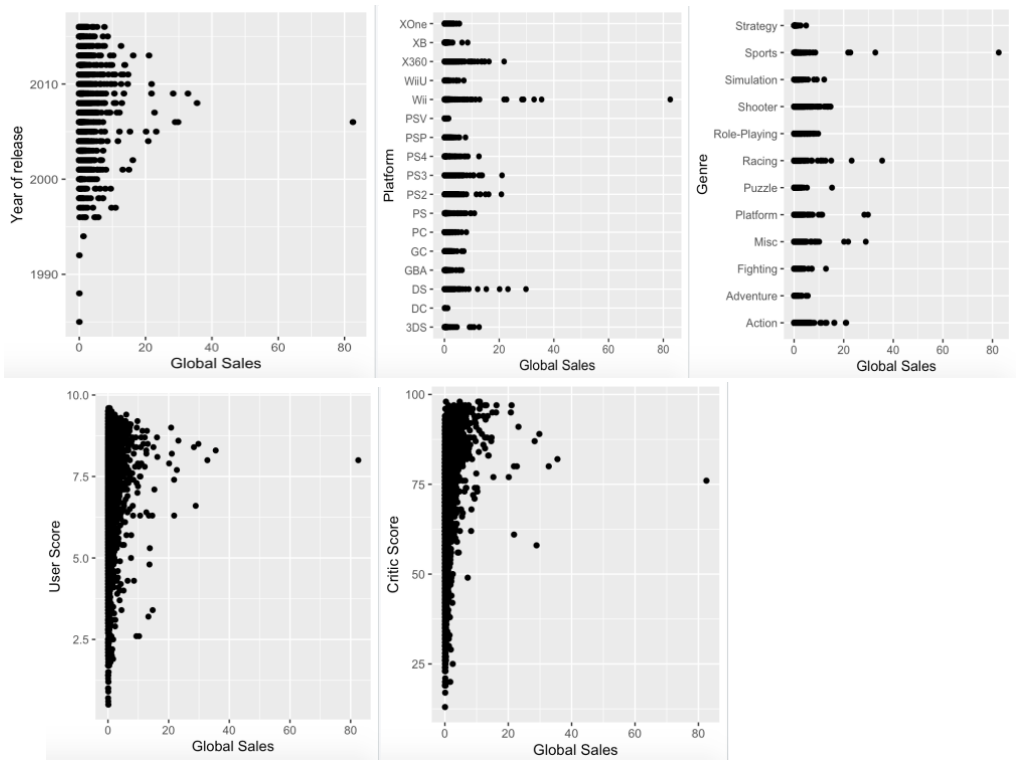
3.1 Description of Dataset

3.1.1 Summary Statistic & Normality check for global sales of video games



Based on the plots shown, the histogram of global sales, it reveals that it is a right skewed distribution as most of the data is distributed on the left hand side.It can also be seen that most of its global sales are between 0 and 2.
In addition, from the histogram of global sales, it can be seen that most of the outliers exist on the top of the boxplot as such a test should be done to determine if the outlier needs to be removed from the dataset.
From the qqplot, it can be seen that most of the data fits nicely to a qqline, hence we can deduce that the data of global sales follows a normal distribution to a certain extent. However, on the right tail of the qq plot, it can be seen that it does not follow the normal distribution thus implying that there are more extreme values on the right tail.

3.1.2 Scatterplot

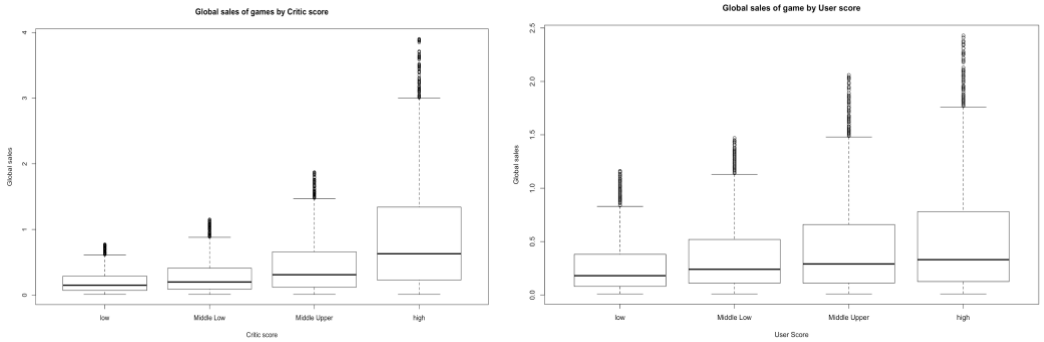


We have used scatterplot to identify any potential relationship between Global_Sales and other variables like Year_of_Release, Platform, Genre, User_Score and Critic_Score.

Based on the plot of the numerical variables, it seems that there is a positive correlation between user score and critic score with global sales. There is higher global sales for games that have higher user scores and critic scores.
For categorical variables like Year of release, platform and genre, there seems to be higher global sales for games released between 2005 and 2010s. In addition, games played on wii and are of sports genre seem to have gotten the highest global sales.

3.1.3 Boxplot

3.1.3.1. Overview & determine if Critic_Score and User_Score of games affect the Global_Sales of games



We categorised the critic score and user score into 4 groups, low, middle low, middle upper and high. The Low category is formed by taking values less than the first quartile.The middle low category is formed by taking values between the first quartile and median. The middle high category is formed by taking values between the median and third quartile. The high category is formed by taking values greater than the third quartile.

From the first plot shown,there is an increase in the median of global sales of games as critic score increases. This can be seen as the low critic score category has a lower median of global sales than the middle low score category. This means that the critic score does have an impact on the global sales of games, where a higher score would result in an increase in sales. A higher critic score could further motivate users to buy the games thus resulting in a higher global sales. There is a huge increase in median between high critic score category and middle upper category. This could be because high critic score categories contain games that are award winning, thus gaining higher global sales.

Looking at the second plot, we can also observe that there is an increase in the median of global sales as user score increases. To add on, there is also an increase in maximum global score as user score increases. This shows that user score has an impact on global sales of games, where the higher the user score, the greater the sales.

3.2 Outlier Analysis

Based on the above boxplot, it can be seen that there are many outliers present. As such, an outlier analysis should be done to decide if there is a need to remove the outlier. We will compare the 0.05 trimmed mean with the original mean to see if there is any significant difference present. The trimmed mean would help to reduce the impact of extreme values on the dataset.

Variables	Original Mean	Trimmed Mean
Global_Sales	0.7775897	0.5135683
Critic_Score	70.27209	70.95442
User_Score	7.185626	7.300244

After computing the trimmed mean and original mean, it can be seen that there is not any significant difference between the two. As a result, it is not necessary to remove the outlier present. Although from the above data plot there are many outliers present,there are about 6 thousand data present, as a result the number of outliers relative to the total data present may not be very significant.

3.3 Categorical Data Analysis

3.3.1 Hypothesis Testing between Global_Sales and Year_of_Release

We would like to perform categorical analysis to see if there is any association between global sales and year of release for video games. We classified global sales into four categories based on the quartiles and the median, and classified year of release by decades. After that, we conducted a hypothesis test using the Chi-squared test at the significance level of 0.05.

H0: There is no association between Global_Sales and Year_of_Release
H1: There is association between Global_Sales and Year_of_Release

	Global_Sales			
Year_of_Release	Low	Middle lower	Middle upper	High
1980s	3	7	51	144
1990s	216	449	520	586
2000s	2038	2488	2415	2251

2010s	1532	1387	1191	1169

Result:

Pearson's Chi-squared test
data: table
X-squared = 520.58, df = 9, p-value < 2.2e-16

Based on the Pearson's Chi-squared test, p-value obtained is less than 2.2*10^(-16) which is much less than the significance level of 0.05. Hence, we reject the null hypothesis and conclude that there is association between global sales and year of release for video games. According to the table above, it seems that the games that were published later were more likely to have lower global sales. This might be explained by the fact that, as time passes by, people have more alternative options of games such as mobile phone games, resulting in a decrease in the proportion of video games that have higher sales.

3.4 Sample Tests

3.4.1 Hypothesis Testing of Critic_Score and User_Score

To investigate whether critic score and user score have the same mean, Variance test and two-samples t-test will be used to determine whether homoscedasticity assumption is violated and if the variables have the same mean at the significance level of 0.05. If the p-value is less than 0.05, we will reject the null hypothesis. Noted that critic score and user score were graded on different scales, thus we need to adjust them to the same scale first.

Variance test:
H0: Variance of Critic_Score and variance of User_Score are equal
H1: Variance of Critic_Score and variance of User_Score are not equal

T-test:
H0: Mean of Critic_Score and mean of User_Score are the same
H1: Mean of Critic_Score is less than mean of User_Score

Variables	Variance Test	T-test	Is mean of Critic_Score less than mean of User_Score
Global_Sales by Critic_Score	Heteroscedastic (p-value = 0.001919)	p-value = 3.054e-11	Yes, the p-value is less than 0.05

Based on the result shown above, if the p-value is above the significance level of 0.05, we do not reject the null hypothesis and deduce that the mean of critic score is the same as the mean of user score. However, the p-value is less than 0.05, which reveals that the mean of critic score is less than the mean of user score.

3.4.2 Analysis of Variance (ANOVA)

From the analysis, we realised that the genre of games and platform where the game is running have a correlation with the global sales. Therefore, we would want to determine the mean global sales at different genres and at different platforms to see whether the means remain the same.

H0: μ1=μ2=μ3= μ4=……= μi

H1: not all μi are equal

To conduct ANOVA test, we would have to assume the following:

- Data are normally distributed (sample size >= 30)
- Homoscedasticity (variances are the same)
- Samples are independent

3.4.2.1 ANOVA on the mean Global_Sales at different Genre of games

Since the p-value = 3.02e-08 < 0.05, we reject the null hypothesis and conclude that the mean global sales at different genres of games differs. Hence, the different genres of games do affect the global sales.

3.4.2.2 ANOVA on the mean Global_Sales at different Platform of games

Since the p-value < 2e-16 < 0.05, we reject the null hypothesis and conclude that the mean global sales differs at different platforms of games. Hence, we can observe that the different platforms of games do affect the global sales.

We go on to check whether our assumptions are satisfied, that is to check whether the variances of samples at different level factors are the same. The var.test is not suitable here because there are multiple samples. Thus, by searching on the internet, bartlett.test will be used here. Conducting bartlett.test on the global sales among different types of games and different platforms, we realized that the p-value = 3.02e-08 < 0.05 and p-value < 2.2e-16 < 0.05 respectively. Therefore, we reject the null hypothesis that the variances are the same. Hence the assumption of homoscedasticity is not satisfied. That means ANOVA model is not suitable here to determine whether the means are the same and we cannot conclude the mean global sales differ at different factor levels. However, we are unable to find a better model within this course scope. Therefore, we need further knowledge to improve this.

3.5 Correlation and Regression

3.5.1 Correlation

To investigate the correlation of the factors, Pearson's Correlation Coefficients (ρ) was used to measure the strength of a linear relationship between two variables. This was followed by the use of a correlation test to ensure the association between the variables. The null hypothesis (H0) of correlation test is that there was no association between paired samples, while alternative hypothesis (H1) states otherwise.

From *Table 3.5.1* below, all factors have positive correlation with the global sales value, as seen from all positive ρ values from the first row of the table.
Critic count has the highest correlation with the global sales (ρ = 0.2903967) and is supported by the correlation test which showed the p-value (2.2 e-16) which is less than the significance level of 0.05. Hence, the null hypothesis of Pearson's Correlation Coefficient Test was therefore rejected, indicating that there is a correlation between global sales and critic count.

Similarly, critic score,user score and user count have affected positively towards the global sales with a correlation of 0.2375557, 0.08834853, 0.2641508 towards the global sales respectively. Their p-values are 2.2 e-16, 2.637e-13, 2.2e-16 respectively, all of which are lower than the significance level of 0.05.
However, the year of release of the games did not play a significant role in the global sales, with a correlation coefficient of 0.006660842 and a p value of 0.5822, which is much greater than the significance level of 0.05. Hence, in this case, the null hypothesis is not rejected, indicating there is no strong correlation between global sales and the year of release of the games.

	Global_Sales	Ctritic_Score	Critic_Count	User_Score	User_Count	Year_of_Release
Global_Sales	-	ρ=0.2375557	ρ=0.2903967	ρ=0.0883485 3	ρ=0.2641508	ρ=0.006660842
Critic_Score	p-value=2.2e-16	-	ρ=0.3964782	ρ=0.5803184	ρ=0.2656387	ρ=-0.007660526
Critic_Count	p-value=2.2e-16	p-value=2.2e-16	-	ρ=0.1950873	ρ=0.3656026	ρ=0.2033363
User_Score	p-value=2.637e-13	p-value=2.2e-16	p-value=2.2e-16	-	ρ=0.0175460 4	ρ=-0.2539137
User_Count	p-value=2.2e-16	p-value=2.2e-16	p-value=2.2e-16	p-value=0.1472	-	ρ=0.1993475

Year_of_Release	p-value=0.5822	p-value=0.5269	p-value=2.2e-16	p-value=2.2e-16	p-value=2.2e-16	-

Table 3.5.1

3.5.2 Multiple Linear Regression

In order to have a better understanding of the relationship between the variables and global sales, Multiple Linear Regression (MLR) was used to understand the prediction of the value and to find a suitable model that established the relationship between global sales and the various factors.

Suggested Model:

$$Y_i = \beta_0 + \beta_1 \times CriticCount + \beta_2 \times CriticScore + \beta_3 \times UserCount + \beta_4 \times UserScore$$

Y_i: Global_Sales
β_0: Intercept of the linear model,
β_1: Standardised Regression Coefficient for Critic_Count,
β_2: Standardised Regression Coefficient for Critic_Score,
β_3: Standardised Regression Coefficient for User_Count,
β_4: Standardised Regression Coefficient for User_Score,
With the use of *summary()* function in R, *Figure 3.5.2* is obtained.

```
Call:
lm(formula = games$Global_Sales ~ games$Critic_Score + games$Critic_Count +
    games$User_Score + games$User_Count)

Residuals:
   Min     1Q Median     3Q    Max
-6.111 -0.563 -0.230  0.181 81.184

Coefficients:
                   Estimate Std. Error t value
(Intercept)       -9.464e-01  1.287e-01  -7.356
games$Critic_Score  2.001e-02  2.153e-03   9.293
games$Critic_Count  1.857e-02  1.318e-03  14.096
games$User_Score   -4.357e-02  1.925e-02  -2.264
games$User_Count    5.371e-04  4.165e-05  12.894
                  Pr(>|t|)
(Intercept)       2.12e-13 ***
games$Critic_Score  < 2e-16 ***
games$Critic_Count  < 2e-16 ***
games$User_Score     0.0236 *
games$User_Count    < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.836 on 6820 degrees of freedom
Multiple R-squared:  0.126,     Adjusted R-squared:  0.1255
F-statistic: 245.8 on 4 and 6820 DF,  p-value: < 2.2e-16
```

Figure 3.5.2 Linear Model

The multiple of R^2 is 12.6% which shows a weak linear relation between the global sales and the respective variables. Likewise, F-statistics is using the ratio between the two mean squares (Sum of Square Regression and Sum of Square Error). Thus, this model sets the null hypothesis to be that of not being able to show the overall significance in explaining the model effectively. F-statistics rejected the null hypothesis (p-value < 2.2e-

16) and indicates that the variables are explaining the overall significance effectively on the global sales.

From *Figure 3.5.2*, estimation of the regression coefficients are,

β_0= *-0.9464054554,* β_1= *0.0185736669,* β_2 = *0.0200065164, ,* β_3= *0.0005370832,* β_4= *-0.0435734022*

Each of the $\beta_i (i = 0,1,\ldots,5)$indicates the regression coefficients that affects the Y_i when a unit of the variable increases. Hence if β_i is negative, it indicates that it will have a negative impact on Y and vice versa. From *Table 3.5.1*, a few things are notably important. First, critic count plays a significant role in the model as the p-value is below the level of significance which is 0.05 in this case. This is due to the large influence of the critics and users would like to follow their recommendation in majority. Thus it is not surprising that the critic count plays the most significant role here. Second, the user score has a negative impact on the global sales, it is implied that the global sales will decrease by 0.04 unit with an increase of user score by 1 unit. While users tend to check the feedback before buying a new game, therefore the feedback from the old users are quite persuasive and reliable for them. This could further affect if the new users would like to purchase for the game or not. However, using the model suggested in *Figure 3.5.2*, the residuals from the fitted model do vary a lot and almost form a straight line. This implies that a linear regression model is probably not appropriate for this data. Thus we would like to suggest other models for further studies.

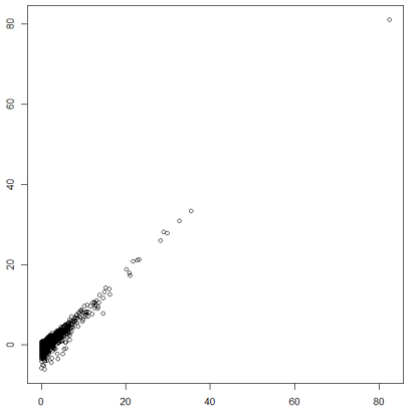


Figure 3.5.3 Residuals

4. Conclusion

Using dataframe Video_Games_Sales_as_at_22_Dec_2016 from kaggle, we are able to identify the response variable Global_Sales along with its predictor variables consisting of numerical and categorical variables.

Next, we used statistical methods such as plotting of histogram and qqplot to visualise the distribution and normality of our response variable. Thereafter, we plotted several scatterplots of numerical predictor variables to compare their relationships with the response variable.

We then broke down two of the numerical predictor variables, Critic_Score and User_Score, into four groups (low, middle-low, middle-upper and high) and used boxplot to analyse how global sales differ for these different groups using boxplot.

Additionally, we applied outlier analysis and then did not remove any outliers as they do not have much influence on the mean of the data. Further analysis was also done through the use of Pearson Chi Square test, two sample t-test and ANOVA.

Furthermore, we computed the correlation and p-values between the different variables to further confirm our visual analysis. Finally, MLR model was constructed to examine how the response variable (Global_Sales) changes with respect to its predictor variables (Critic_Count, Critic_Score, User_Count, User_Score).

However, there were some limitations in our analysis. Our analysis may not be comprehensive because the dataset we used did not contain some variables that would have significant effects on global sales. For example, marketing plays an essential role for customers to decide whether or not to purchase the video game. Also, old series of games have an advantage on acquiring users because it has an existing customer base.

Nevertheless, even with this limitation, we were able to analyse and come up with a comparatively reliable conclusion based on current variables.

Using the above results and applying it in the field of video game development, we could give some advice to video game companies based on our analysis. One possible suggestion to them is through the increase of scores graded by critics. However, Critic_Score cannot be directly affected by video game companies. One possible way to increase Critic_Score is to improve the quality of the video games. In addition, video game companies should pay more attention to critics' opinions and take their views into considerations for future development of games.

5. Appendix

References:

Rachel Chia. 2019. Singapore's newest billionaire, Sea founder Forrest Li, says he chose his English name after watching Forrest Gump.

Available from: https://www.businessinsider.sg/its-official-singaporeans-spend-the-most-time-playing-video-games-in-all-of-asia-skipping-out-on-sleep-and-meals-study-finds

Sean Lim. 2019. It's official: Singaporeans spend the most time playing video games in all of Asia – skipping out on sleep and meals, study finds.
Available from: https://www.businessinsider.sg/its-official-singaporeans-spend-the-most-time-playing-video-games-in-all-of-asia-skipping-out-on-sleep-and-meals-study-finds

Description of variables used

Variables	Description
Platform	Platform where the game can be played (eg PS4, X360)
Year_of_Release	The year when the game was first released
Genre	Classification of gaming type (eg sports,puzzle)
Gobal_Sales	Total sales value obtained globally
Critic_Score	Rating given by critics, total score being 100
Critic_Count	Number of critics giving rating for the game
User_Score	Rating given by critics, total score being 10
User_Count	Number of users giving rating for the game

Table 1

Summary statistics of the variables used

Variable	Mean	Standard deviation	Minimum	Maximum
Year_Of_Release	2007	4.21	1985	2016
Global_Sales	0.7776	1.96	0.01	82.53

Critic_Score	70.27	13.88	13	98
Critic_Count	28.93	19.22	3	113
User_Score	7.186	1.44	0.5	9.60
User_Count	174.7	587.43	4	10665

Table 2

Codes:

Section 3.1: Description of data

#plotting of graph

```
> globals = cleansales$Global_Sales
> years = cleansales$Year_of_Release
> plot(globals,years, xlab = "Global sales", ylab= "Year of release")
```

plotting of boxplot to be repeated with other variables:Platform, Genre, User_Score and Critic_Score.

```
> globalsale = cleansales$Global_Sales
> hist(globalsale, main="histogram of global sales")
> boxplot(globalsale, main="boxplot of global sales")
> qqnorm(globalsale)
> qqline(globalsale)
```

analysing the critic score to categorize them into the groups

```
> critic = cleansales$Critic_Score
> user = cleansales$User_Score
> summary(critic)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   62.00   72.00   70.27   80.00   98.00
```

grouping the various critic score

```
> cleansales$Categorycritic[cleansales$Critic_Score<=62]='low'
Warning message:
Unknown or uninitialised column: 'Categorycritic'.
> cleansales$Categorycritic[cleansales$Critic_Score>62 &cleansales$Critic_Score<=72
 ]='Mid low'
>
> cleansales$Categorycritic[cleansales$Critic_Score>72 &cleansales$Critic_Score<=80
 ]='Mid high'
>
> cleansales$Categorycritic[cleansales$Critic_Score>80]='High'
> boxplot(Global_Sales~Categorycritic, data = cleansales, xlab = "Critic score", yla
b = "Global sales")
>
```

Section 3.2: Outliers Analysis

```
        # finding the mean for outlier analysis
> mean(cleansales$Global_Sales)
[1] 0.7775897
> mean(cleansales$Global_Sales,trim=0.05)
[1] 0.5135683
> mean(cleansales$Critic_Score)
[1] 70.27209
> mean(cleansales$Critic_Score,trim=0.05)
[1] 70.95442
> mean(cleansales$User_Score)
[1] 7.185626
> mean(cleansales$User_Score,trim=0.05)
[1] 7.300244
>
```

Section 3.3: Categorical analysis

```
 # change data type of Year_of_Release from factor to numeric
> games$Year_of_Release = as.numeric(as.character(games$Year_of_Release))

 # remove NA values in Year_of_Release
> games = games[!is.na(games$Year_of_Release), ]

 # check the quartiles and the median of Global_Sales for classification
> summary(games$Global_Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0100  0.0600  0.1700  0.5363  0.4700 82.5300

 # find the number of games under specific categories with codes
 # an example:
> str(subset(games, Year_of_Release >= 1980 & Year_of_Release <1990 & Global_Sales >= 0.06
 & Global_Sales < 0.17))
'data.frame':   7 obs. of  16 variables:
 # repeat until all the required values in the 2-way contingency table are found

 # create the 2-way contingency table and conduct the Chi-squared test
> table = matrix(c(3,7,51,144,216,449,520,586,2038,2488,2415,2251,1532,1387,119
1,1169), ncol = 4, byrow = TRUE)
> colnames(table) = c("Low","Middle lower","Middle upper","High")
> rownames(table) = c("1980s","1990s","2000s","2010s")
> chisq.test(table)

        Pearson's Chi-squared test

data:  table
X-squared = 520.58, df = 9, p-value < 2.2e-16
```

Section 3.4: Sample Test

```
> games_score<-subset(games,select=c(Critic_Score,User_Score))
> str(games_score)
'data.frame':   6825 obs. of  2 variables:
 $ Critic_Score: num  76 82 80 89 58 87 91 80 61 80 ...
 $ User_Score  : num  8 8.3 8 8.5 6.6 8.4 8.6 7.7 6.3 7.4 ...
> #adjust critic_score to the same sale
> games_score$Critic_Score<-games_score$Critic_Score/10
> str(games_score)
'data.frame':   6825 obs. of  2 variables:
 $ Critic_Score: num  7.6 8.2 8 8.9 5.8 8.7 9.1 8 6.1 8 ...
 $ User_Score  : num  8 8.3 8 8.5 6.6 8.4 8.6 7.7 6.3 7.4 ...
> attach(games_score)
```

```
> #determine whether two variances are the same
> var.test(Critic_Score,User_Score)

        F test to compare two variances

data:  Critic_Score and User_Score
F = 0.92763, num df = 6824, denom df = 6824, p-value = 0.001919
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8846341 0.9727109
sample estimates:
ratio of variances
         0.9276277
```

```
> #conduct t-test
> t.test(Critic_Score, User_Score,var.equal=F,alternative="less")

        Welch Two Sample t-test

data:  Critic_Score and User_Score
t = -6.5463, df = 13629, p-value = 3.054e-11
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1186104
sample estimates:
mean of x mean of y
 7.027209  7.185626
```

```
> #conduct ANOVA model
> summary(aov(games$Global_Sales~factor(games$Genre)))
                     Df Sum Sq Mean Sq F value   Pr(>F)
factor(games$Genre)   11    220  19.991   5.221 3.02e-08 ***
Residuals           6813  26087   3.829
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(games$Global_Sales~factor(games$Platform)))
                       Df Sum Sq Mean Sq F value Pr(>F)
factor(games$Platform)   16    692   43.27    11.5 <2e-16 ***
Residuals              6808  25615    3.76
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #determine whether two variances are the same
> bartlett.test(games$Global_Sales,games$Genre)

        Bartlett test of homogeneity of variances

data:  games$Global_Sales and games$Genre
Bartlett's K-squared = 2341, df = 11, p-value < 2.2e-16

> bartlett.test(games$Global_Sales,games$Platform)

        Bartlett test of homogeneity of variances

data:  games$Global_Sales and games$Platform
Bartlett's K-squared = 5105.5, df = 16, p-value < 2.2e-16
```

Section 3.5.1: Coefficient values

"games" refer to cleaned data, which only includes the numeric data : year of release, global sales, critic score, critic count, user score and user count values.

```
> cor(games)
                     X Year_of_Release Global_Sales Critic_Score Critic_Count  User_Score  User_Count
X           1.00000000    -0.010506930 -0.464819311 -0.369455840   -0.4208513 -0.16446098 -0.21762339
Year_of_Release -0.01050693     1.000000000  0.006660842 -0.007660526    0.2033363 -0.25391372  0.19934754
Global_Sales    -0.46481931     0.006660842  1.000000000  0.237555722    0.2903967  0.08834853  0.26415078
Critic_Score    -0.36945584    -0.007660526  0.237555722  1.000000000    0.3964782  0.58031837  0.26563871
Critic_Count    -0.42085135     0.203336345  0.290396685  0.396478161    1.0000000  0.19508730  0.36560257
User_Score      -0.16446098    -0.253913721  0.088348526  0.580318371    0.1950873  1.00000000  0.01754604
User_Count      -0.21762339     0.199347540  0.264150778  0.265638706    0.3656026  0.01754604  1.00000000
```

Figure 3.5.4 Coefficient Values

P values are found by using cor.test() on individual pairs of variables.
Combining these values, we will get Figure 3.5.1 on page 9.

Section 3.5.2: Multiple Linear Regression

With the use of *summary()* function in R, *Figure 3.5.2* in page 10 of the report is obtained.