R Markdown :

|  |
| --- |
| --- |
|  | title: "Video Game Sales Project" |
|  | author: "Sahil Garg" |
|  | date: "29 December 2017" |
|  | output: |
|  | word\_document: default |
|  | pdf\_document: default |
|  | html\_document: default |
|  | --- |
|  | ```{r setup, include=FALSE} |
|  | knitr::opts\_chunk$set(echo = TRUE) |
|  | ``` |
|  | **### Project Title: Understanding and Analyzing Video Game Sales** |
|  | **### NAME: Sahil Garg** |
|  | **### EMAIL: sahilgarg3098@gmail.com** |
|  | **### COLLEGE / COMPANY: Manipal Institute of Technology, Manipal** |
|  |  |
|  | **## 1. Introduction** |
|  | A video game is an electronic game that can be played on a computing device, such as a personal computer, gaming console or mobile phone. Depending on the platform, video games can be subcategorized into computer games and console games. In recent years however, the emergence of social networks, smartphones and tablets introduced new categories such as mobile and social games. Video games have come a long way since the first games emerged in the 1970s. Today's video games offer photorealistic graphics and simulate reality to a degree which is astonishing in many cases. |
|  |  |
|  | **## 2. Overview of the Study** |
|  | The video game industry is the economic sector involved in the development, marketing, and monetization of video games. It encompasses dozens of job disciplines and its component parts employ thousands of people worldwide. |
|  |  |
|  | The computer and video game industry has grown from focused markets to mainstream. They took in about US$9.5 billion in the US in 2007, 11.7 billion in 2008, and 25.1 billion in 2010 (ESA annual report). |
|  |  |
|  | Modern personal computers owe many advancements and innovations to the game industry: sound cards, graphics cards and 3D graphic accelerators, faster CPUs, and dedicated co-processors like PhysX are a few of the more notable improvements. |
|  |  |
|  | **## 3. An analytical study of Video game industry and sales across the world** |
|  |  |
|  | **## 3.1 Overview** |
|  |  |
|  | We study the given data set of sales of video games to find relations between different factors that affect video game sales. Our main objectives are - |
|  | - Finding which platform is more popular in which region and affects sale by how much |
|  | - Finding which year and period was more popular and how has it affected sales |
|  | - Finding which publisher is more popular in which region and affects sale by how much |
|  | - Finding which genre is more popular in which region and affects sale by how much |
|  | - How are the sales of America, Europe, Japan and other regions of the world correlated and is there a common pattern amongst all, or do they follow different trends |
|  | - Realising and understanding the needs and trends in the video game industry to target the maximum customers in a new release of a game etc. |
|  |  |
|  | **## 3.2 Data** |
|  | This dataset contains a list of video games with sales greater than 100,000 copies. It was generated by a scrape of vgchartz.com. |
|  |  |
|  | Fields include |
|  |  |
|  | Rank - Ranking of overall sales |
|  |  |
|  | Name - The games name |
|  |  |
|  | Platform - Platform of the games release (i.e. PC,PS4, etc.) |
|  |  |
|  | Year - Year of the game's release |
|  |  |
|  | Genre - Genre of the game |
|  |  |
|  | Publisher - Publisher of the game |
|  |  |
|  | NA\_Sales - Sales in North America (in millions) |
|  |  |
|  | EU\_Sales - Sales in Europe (in millions) |
|  |  |
|  | JP\_Sales - Sales in Japan (in millions) |
|  |  |
|  | Other\_Sales - Sales in the rest of the world (in millions) |
|  |  |
|  | Global\_Sales - Total worldwide sales. |
|  |  |
|  | The script to scrape the data is available at https://github.com/GregorUT/vgchartzScrape. It is based on BeautifulSoup using Python. There are 16,598 records. 2 records were dropped due to incomplete information. |
|  |  |
|  | **## Reading the data** |
|  |  |
|  | Read your dataset in R and visualize the length and breadth of your dataset. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | setwd("D:/manipal-year2/internship/IIML\_dataAnalytics/project/videogamesales") |
|  | game.df <- read.csv(paste("vgsales.csv",sep="")) |
|  | View(game.df) |
|  | head(game.df) |
|  | dim(game.df) |
|  | ``` |
|  |  |
|  | **## Describe each variable** |
|  |  |
|  | Create a descriptive statistics (min, max, median etc) of each variable. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | attach(game.df) |
|  | summary(game.df) |
|  | library(psych) |
|  | describe(game.df) |
|  |  |
|  | str(game.df) |
|  |  |
|  | min(NA\_Sales) |
|  | max(NA\_Sales) |
|  | mean(NA\_Sales) |
|  | median(NA\_Sales) |
|  |  |
|  | min(EU\_Sales) |
|  | max(EU\_Sales) |
|  | mean(EU\_Sales) |
|  | median(EU\_Sales) |
|  |  |
|  | min(JP\_Sales) |
|  | max(JP\_Sales) |
|  | mean(JP\_Sales) |
|  | median(JP\_Sales) |
|  |  |
|  | min(Other\_Sales) |
|  | max(Other\_Sales) |
|  | mean(Other\_Sales) |
|  | median(Other\_Sales) |
|  |  |
|  | min(Global\_Sales) |
|  | max(Global\_Sales) |
|  | mean(Global\_Sales) |
|  | median(Global\_Sales) |
|  | ``` |
|  |  |
|  | We can see that the mean sales in North America are the highest, followed by Europe, then Japan and then the rest of the world. The maximum sales also show a similar trend, however the sales of Japan and the rest of the world are quite comparable. |
|  | Electronic Arts is the most popular publisher for video games. Action games are higher in number and are more preffered according to the sales statistics. |
|  |  |
|  | **## 3.3 Model** |
|  | We now run regression models to understand the fields that affect sales in NA, EU, JP, Other regions and globally. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | fit1 <- lm(Rank~Platform,data = game.df) |
|  | summary(fit1) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | fit1 <- lm(Rank~Year,data = game.df) |
|  | summary(fit1) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | fit1 <- lm(Rank~Genre,data = game.df) |
|  | summary(fit1) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | fit1 <- lm(Rank~Publisher,data = game.df) |
|  | summary(fit1) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | fit1 <- lm(Rank~NA\_Sales+EU\_Sales+JP\_Sales+Other\_Sales+Global\_Sales,data = game.df) |
|  | summary(fit1) |
|  | fit1$coefficients |
|  | confint(fit1) |
|  | ``` |
|  |  |
|  | We established the effect of platform, year, publisher, genre and sales on the rank of a video game with the simplest model. We regressed Rank on platform, year, publisher, genre and sales in NA, EU, JP, other regions and globally. We estimated model, using linear least squares. If there was more sales in the different regions, we expected the rank of the game to be lesser. |
|  |  |
|  | **## 3.4 Results** |
|  |  |
|  | The above regression models show that the rank of a game depends on - |
|  | (since the p-values are less than 0.05) |
|  | 1. The platform game is developed on, for all platforms. |
|  | 2. The year the game was developed in, after 2005. |
|  | 3. The genre of the game, except for racing and role-playing. |
|  | 4. The publisher for a few specific publishers |
|  | 5. Sales of all the regions (NA, EU, JP, other regions and global). |
|  |  |
|  | **## 4. Conclusion and Inference** |
|  |  |
|  | - Sales in North america are the most, accounting to almost to half of the global sales. |
|  | - The years when the video game sales were at their peak were from 2007-2011 globally. |
|  | - DS, PS2, PS3 are the most popular consoles to develop games on. |
|  | - Action games are an all-time favourite, followed by sports, shooter and role-playing. |
|  | - Big names in the market such as EA Sports, Ubisoft, Activision have high sales of video games in NA and the rest of the world except Japan. Japan prefers more of role-playing games, by producers such as capcom, namcoi bandai and nintendo. |
|  | - Trends in sales are almost similar in all regions across the world. |
|  | - The rank of a game is directly correlated to the sales of all regions, whether it be NA, EU, JP, other regions or global sales. |
|  | - The sales of a region are also correlated to the sales of other regions, thus, sales of an area are in sync with other regions. It is a high possibility that a game successful in one region will be successful in other regions also. |
|  |  |
|  | **## 5. References** |
|  |  |
|  | \bibitem{1} Kaggle link of video games sales data and overview on it, Available from: <https://www.kaggle.com/gregorut/videogamesales>. [29 Dec 2017]. |
|  |  |
|  | \bibitem{2}Wikipedia link for video games, Available from: <https://en.wikipedia.org/wiki/Video\_game>. [30 Dec 2017]. |
|  |  |
|  | \bibitem{2}Wikipedia link for video games industry, Available from: <https://en.wikipedia.org/wiki/Video\_game\_industry>. [30 Dec 2017]. |
|  |  |
|  | **## Table 1: Aggregate NA sales based on Genre** |
|  |  |
|  | Genre NA\_Sales |
|  | ------------------------ |
|  | Action 877.83 |
|  | Adventure 105.80 |
|  | Fighting 223.59 |
|  | Misc 410.24 |
|  | Platform 447.05 |
|  | Puzzle 123.78 |
|  | Racing 359.42 |
|  | Role-Playing 327.28 |
|  | Shooter 582.60 |
|  | Simulation 183.31 |
|  | Sports 683.35 |
|  | Strategy 68.70 |
|  |  |
|  | **## Table 2: Regression Analysis in the video games Study based on sales** |
|  |  |
|  | Estimate Std. Error t value Pr(>|t|) |
|  | ----------------------------------------------------- |
|  | (Intercept) 8996.7 35.6 252.681 <2e-16 \*\*\* |
|  | NA\_Sales -55850.7 6422.9 -8.696 <2e-16 \*\*\* |
|  | EU\_Sales -55472.6 6424.0 -8.635 <2e-16 \*\*\* |
|  | JP\_Sales -56046.2 6423.4 -8.725 <2e-16 \*\*\* |
|  | Other\_Sales -56823.1 6425.7 -8.843 <2e-16 \*\*\* |
|  | Global\_Sales 54538.9 6422.9 8.491 <2e-16 \*\*\* |
|  | ----------------------------------------------------- |
|  |  |
|  | Figure 1: Popular video games in the market, topping the sales charts. (Source: http://hiconsumption.com/2015/07/best-video-games-of-all-time/) |
|  |  |
|  | ![Best Video Games of 2015](image1.jpeg) |
|  |  |
|  | Figure 2: Popular video game consoles in the market, as of 2016. (Source: http://www.finehomesandliving.com/Best-Gaming-Consoles-for-2015/) |
|  |  |
|  | ![Best Video Game Consoles of 2016](image2.jpg) |
|  |  |
|  | **## Appendix 1 - Plots** |
|  |  |
|  | **### BoxPlots** |
|  |  |
|  | Draw a boxplot of the variables that belong to your study. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | boxplot(NA\_Sales,horizontal = TRUE, |
|  | xlab = "NA Sales in millions", |
|  | main = "North America Sales") |
|  |  |
|  | boxplot(EU\_Sales,horizontal = TRUE, |
|  | xlab = "EU Sales in millions", |
|  | main = "Europe Sales") |
|  |  |
|  | boxplot(JP\_Sales,horizontal = TRUE, |
|  | xlab = "JP Sales in millions", |
|  | main = "Japan Sales") |
|  |  |
|  | boxplot(Other\_Sales,horizontal = TRUE, |
|  | xlab = "Other Sales in millions", |
|  | main = "Other Sales") |
|  |  |
|  | boxplot(Global\_Sales,horizontal = TRUE, |
|  | xlab = "Global Sales in millions", |
|  | main = "Global Sales") |
|  |  |
|  | boxplot(NA\_Sales~Platform,horizontal = TRUE, |
|  | xlab = "NA Sales in millions", |
|  | ylab = "Platform", |
|  | main = "North America Sales based on platform") |
|  |  |
|  | boxplot(NA\_Sales~Year,horizontal = TRUE, |
|  | xlab = "NA Sales in millions", |
|  | ylab = "Year", |
|  | main = "North America Sales based on year") |
|  |  |
|  | boxplot(NA\_Sales~Genre,horizontal = TRUE, |
|  | xlab = "NA Sales in millions", |
|  | ylab = "Genre", |
|  | main = "North America Sales based on genre") |
|  |  |
|  | boxplot(NA\_Sales~Publisher,horizontal = TRUE, |
|  | xlab = "NA Sales in millions", |
|  | ylab = "Publisher", |
|  | main = "North America Sales based on publisher") |
|  |  |
|  | boxplot(EU\_Sales~Platform,horizontal = TRUE, |
|  | xlab = "EU Sales in millions", |
|  | ylab = "Platform", |
|  | main = "Europe Sales based on platform") |
|  |  |
|  | boxplot(EU\_Sales~Year,horizontal = TRUE, |
|  | xlab = "EU Sales in millions", |
|  | ylab = "Year", |
|  | main = "Europe Sales based on year") |
|  |  |
|  | boxplot(EU\_Sales~Genre,horizontal = TRUE, |
|  | xlab = "EU Sales in millions", |
|  | ylab = "Genre", |
|  | main = "Europe Sales based on genre") |
|  |  |
|  | boxplot(EU\_Sales~Publisher,horizontal = TRUE, |
|  | xlab = "EU Sales in millions", |
|  | ylab = "Publisher", |
|  | main = "Europe Sales based on publisher") |
|  |  |
|  | boxplot(JP\_Sales~Platform,horizontal = TRUE, |
|  | xlab = "JP Sales in millions", |
|  | ylab = "Platform", |
|  | main = "Japan Sales based on platform") |
|  |  |
|  | boxplot(JP\_Sales~Year,horizontal = TRUE, |
|  | xlab = "JP Sales in millions", |
|  | ylab = "Year", |
|  | main = "Japan Sales based on year") |
|  |  |
|  | boxplot(JP\_Sales~Genre,horizontal = TRUE, |
|  | xlab = "JP Sales in millions", |
|  | ylab = "Genre", |
|  | main = "Japan Sales based on genre") |
|  |  |
|  | boxplot(JP\_Sales~Publisher,horizontal = TRUE, |
|  | xlab = "JP Sales in millions", |
|  | ylab = "Publisher", |
|  | main = "Japan Sales based on publisher") |
|  |  |
|  | boxplot(Other\_Sales~Platform,horizontal = TRUE, |
|  | xlab = "Other Sales in millions", |
|  | ylab = "Platform", |
|  | main = "Other Sales based on platform") |
|  |  |
|  | boxplot(Other\_Sales~Year,horizontal = TRUE, |
|  | xlab = "Other Sales in millions", |
|  | ylab = "Year", |
|  | main = "Other Sales based on year") |
|  |  |
|  | boxplot(Other\_Sales~Genre,horizontal = TRUE, |
|  | xlab = "Other Sales in millions", |
|  | ylab = "Genre", |
|  | main = "Other Sales based on genre") |
|  |  |
|  | boxplot(Other\_Sales~Publisher,horizontal = TRUE, |
|  | xlab = "Global Sales in millions", |
|  | ylab = "Publisher", |
|  | main = "Global Sales based on publisher") |
|  |  |
|  | boxplot(Global\_Sales~Platform,horizontal = TRUE, |
|  | xlab = "Global Sales in millions", |
|  | ylab = "Platform", |
|  | main = "Global Sales based on platform") |
|  |  |
|  | boxplot(Global\_Sales~Year,horizontal = TRUE, |
|  | xlab = "Global Sales in millions", |
|  | ylab = "Year", |
|  | main = "Global Sales based on year") |
|  |  |
|  | boxplot(Global\_Sales~Genre,horizontal = TRUE, |
|  | xlab = "Global Sales in millions", |
|  | ylab = "Genre", |
|  | main = "Global Sales based on genre") |
|  |  |
|  | boxplot(Global\_Sales~Publisher,horizontal = TRUE, |
|  | xlab = "Global Sales in millions", |
|  | ylab = "Publisher", |
|  | main = "Global Sales based on publisher") |
|  | ``` |
|  |  |
|  | North america and europe have more sales for the PS, whereas japan has good sales for the nintendo and DS too. |
|  | Sales have increased in all regions over the years, and have reached a high around 2012-2014. |
|  | Action, misc, racing and sports are the popular genres. |
|  |  |
|  | **### Histograms** |
|  |  |
|  | Draw Histograms for your suitable data fields. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | hist(NA\_Sales, |
|  | xlab = "NA Sales in millions", |
|  | ylab = "Count", |
|  | xlim = c(0,10), |
|  | col = "lightblue", |
|  | main = "North America Sales") |
|  | hist(NA\_Sales, |
|  | xlab = "NA Sales in millions", |
|  | ylab = "Count", |
|  | col = "lightblue", |
|  | main = "North America Sales") |
|  |  |
|  | hist(EU\_Sales, |
|  | xlab = "EU Sales in millions", |
|  | ylab = "Count", |
|  | xlim = c(0,10), |
|  | col = "lightblue", |
|  | main = "Europe Sales") |
|  | hist(EU\_Sales, |
|  | xlab = "EU Sales in millions", |
|  | ylab = "Count", |
|  | col = "lightblue", |
|  | main = "Europe Sales") |
|  |  |
|  | hist(JP\_Sales, |
|  | xlab = "JP Sales in millions", |
|  | ylab = "Count", |
|  | xlim = c(0,10), |
|  | col = "lightblue", |
|  | main = "Japan Sales") |
|  | hist(JP\_Sales, |
|  | xlab = "JP Sales in millions", |
|  | ylab = "Count", |
|  | col = "lightblue", |
|  | main = "Japan Sales") |
|  |  |
|  | hist(Other\_Sales, |
|  | xlab = "Other Sales in millions", |
|  | ylab = "Count", |
|  | xlim = c(0,10), |
|  | col = "lightblue", |
|  | main = "Other Sales") |
|  | hist(Other\_Sales, |
|  | xlab = "Other Sales in millions", |
|  | ylab = "Count", |
|  | col = "lightblue", |
|  | main = "Other Sales") |
|  |  |
|  | hist(Global\_Sales, |
|  | xlab = "GLobal Sales in millions", |
|  | ylab = "Count", |
|  | xlim = c(0,20), |
|  | col = "lightblue", |
|  | main = "Global Sales") |
|  | hist(Global\_Sales, |
|  | xlab = "Global Sales in millions", |
|  | ylab = "Count", |
|  | col = "lightblue", |
|  | main = "Global Sales") |
|  | ``` |
|  |  |
|  | Sales in North America and Europe range between 0-2 million, and those in Japan range from 0-1 million. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | library(lattice) |
|  | histogram(~NA\_Sales|Platform,type = "count") |
|  | histogram(~NA\_Sales|Year,type = "count") |
|  | histogram(~NA\_Sales|Genre,type = "count") |
|  | histogram(~EU\_Sales|Platform,type = "count") |
|  | histogram(~EU\_Sales|Year,type = "count") |
|  | histogram(~EU\_Sales|Genre,type = "count") |
|  | histogram(~JP\_Sales|Platform,type = "count") |
|  | histogram(~JP\_Sales|Year,type = "count") |
|  | histogram(~JP\_Sales|Genre,type = "count") |
|  | histogram(~Other\_Sales|Platform,type = "count") |
|  | histogram(~Other\_Sales|Year,type = "count") |
|  | histogram(~Other\_Sales|Genre,type = "count") |
|  | histogram(~Global\_Sales|Platform,type = "count") |
|  | histogram(~Global\_Sales|Year,type = "count") |
|  | histogram(~Global\_Sales|Genre,type = "count") |
|  | ``` |
|  |  |
|  | All regions show a similar gaming trend, with PS2 and DS being the popular consoles, and years 2007-11 being heavy on the industry. |
|  |  |
|  | **### Plots** |
|  |  |
|  | Draw suitable plot for your data fields. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | plot(NA\_Sales,EU\_Sales, |
|  | main = "North America sales vs Europe sales", |
|  | xlab = "NA Sales", |
|  | ylab = "EU Sales") |
|  |  |
|  | plot(NA\_Sales,JP\_Sales, |
|  | main = "North America sales vs Japan sales", |
|  | xlab = "NA Sales", |
|  | ylab = "JP Sales") |
|  |  |
|  | plot(NA\_Sales,Other\_Sales, |
|  | main = "North America sales vs Other sales", |
|  | xlab = "NA Sales", |
|  | ylab = "Other Sales") |
|  |  |
|  | plot(EU\_Sales,JP\_Sales, |
|  | main = "Europe sales vs Japan sales", |
|  | xlab = "EU Sales", |
|  | ylab = "JP Sales") |
|  |  |
|  | plot(EU\_Sales,Other\_Sales, |
|  | main = "Europe sales vs Other sales", |
|  | xlab = "EU Sales", |
|  | ylab = "Other Sales") |
|  |  |
|  | plot(JP\_Sales,Other\_Sales, |
|  | main = "Japan sales vs Other sales", |
|  | xlab = "JP Sales", |
|  | ylab = "Other Sales") |
|  |  |
|  | plot(Global\_Sales,NA\_Sales, |
|  | main = "Global sales vs NA sales", |
|  | xlab = "Global Sales", |
|  | ylab = "NA Sales") |
|  | plot(Global\_Sales,JP\_Sales, |
|  | main = "Global sales vs Japan sales", |
|  | xlab = "Global Sales", |
|  | ylab = "Japan Sales") |
|  | plot(Global\_Sales,EU\_Sales, |
|  | main = "Global sales vs Europe sales", |
|  | xlab = "Global Sales", |
|  | ylab = "Europe Sales") |
|  | plot(Global\_Sales,Other\_Sales, |
|  | main = "Global sales vs Other sales", |
|  | xlab = "Global Sales", |
|  | ylab = "Other Sales") |
|  | ``` |
|  |  |
|  | Comparing sales of different regions, America is the most intense consumer of video games, followed by Europe and then Japan. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | barchart(NA\_Sales~Platform) |
|  | barchart(NA\_Sales~Genre) |
|  | barchart(NA\_Sales~Year) |
|  |  |
|  | barchart(EU\_Sales~Platform) |
|  | barchart(EU\_Sales~Genre) |
|  | barchart(EU\_Sales~Year) |
|  |  |
|  | barchart(JP\_Sales~Platform) |
|  | barchart(JP\_Sales~Genre) |
|  | barchart(JP\_Sales~Year) |
|  |  |
|  | barchart(Other\_Sales~Platform) |
|  | barchart(Other\_Sales~Genre) |
|  | barchart(Other\_Sales~Year) |
|  |  |
|  | barchart(Global\_Sales~Platform) |
|  | barchart(Global\_Sales~Genre) |
|  | barchart(Global\_Sales~Year) |
|  |  |
|  | bwplot(Platform~NA\_Sales,horizontal = TRUE) |
|  | bwplot(Genre~NA\_Sales,horizontal = TRUE) |
|  | bwplot(Year~NA\_Sales,horizontal = TRUE) |
|  |  |
|  | bwplot(Platform~EU\_Sales,horizontal = TRUE) |
|  | bwplot(Genre~EU\_Sales,horizontal = TRUE) |
|  | bwplot(Year~EU\_Sales,horizontal = TRUE) |
|  |  |
|  | bwplot(Platform~JP\_Sales,horizontal = TRUE) |
|  | bwplot(Genre~JP\_Sales,horizontal = TRUE) |
|  | bwplot(Year~JP\_Sales,horizontal = TRUE) |
|  |  |
|  | bwplot(Platform~Other\_Sales,horizontal = TRUE) |
|  | bwplot(Genre~Other\_Sales,horizontal = TRUE) |
|  | bwplot(Year~Other\_Sales,horizontal = TRUE) |
|  |  |
|  | bwplot(Platform~Global\_Sales,horizontal = TRUE) |
|  | bwplot(Genre~Global\_Sales,horizontal = TRUE) |
|  | bwplot(Year~Global\_Sales,horizontal = TRUE) |
|  | ``` |
|  |  |
|  | Sports is a popular category in North America and Europe, and role-playing is preffered in Japan. Other regions show more interest in action games. The trend in years is more or less similar in all regions. Japan uses more of DS consoles, whereas America and Europe use PS, Wii and DS(not to as much extend as Japan). |
|  |  |
|  | **## Appendix 2 - Correlation Matrix and Scatterplot matrix** |
|  |  |
|  | **### Correlation Matrix** |
|  |  |
|  | Create a correlation matrix. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | round(cor(game.df[,7:11]),2) |
|  | library(corrplot) |
|  | corrplot(cor(game.df[,7:11]),method = "circle") |
|  | ``` |
|  |  |
|  | Sales of all regions are inter-dependent and strongly correlated. |
|  |  |
|  | **### Corrgram** |
|  |  |
|  | Visualize your correlation matrix using corrgram. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | library(corrgram) |
|  | corrgram(game.df,order=TRUE, |
|  | lower.panel = panel.shade, |
|  | upper.panel = panel.pie, |
|  | text.panel = panel.txt, |
|  | main = "Corrgram for video game sales" |
|  | ) |
|  | ``` |
|  |  |
|  | The sales of each region show relation with each other, but the rank is not necessarily related directly and is a seperate field. |
|  |  |
|  | **### ScatterPlot Matrix** |
|  |  |
|  | Create a scatter plot matrix for your data set. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | library(car) |
|  | scatterplotMatrix(game.df[,c(1,7:11)], spread=FALSE, smoother.args=list(lty=2), main="Scatter Plot Matrix") |
|  | ``` |
|  |  |
|  | The scatter plot matrix shows relations between all the sales, which is obvious since years where sales is higher shows increased sales across all regions of the world. |
|  |  |
|  | **## Appendix 3 - Cor.test, chisq.test and t.test** |
|  |  |
|  | **### Running some tests** |
|  |  |
|  | Run a suitable test to check your hypothesis for your suitable assumptions. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | cor.test(Rank,NA\_Sales) |
|  | cor.test(Rank,EU\_Sales) |
|  | cor.test(Rank,JP\_Sales) |
|  | cor.test(Rank,Other\_Sales) |
|  | cor.test(Rank,Global\_Sales) |
|  | ``` |
|  |  |
|  | We reject the null hypothesis and the p-value less than 0.05 shows that rank of games is directly related to each of the regional sales. |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | cor.test(Global\_Sales,NA\_Sales) |
|  | cor.test(Global\_Sales,EU\_Sales) |
|  | cor.test(Global\_Sales,JP\_Sales) |
|  | cor.test(Global\_Sales,Other\_Sales) |
|  | ``` |
|  |  |
|  | We reject the null hypothesis and the p-value less than 0.05 shows that global sales is directly related to each of the regional sales. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | cor.test(EU\_Sales,NA\_Sales) |
|  | cor.test(JP\_Sales,NA\_Sales) |
|  | cor.test(Other\_Sales,NA\_Sales) |
|  | cor.test(JP\_Sales,EU\_Sales) |
|  | cor.test(Other\_Sales,EU\_Sales) |
|  | cor.test(Other\_Sales,JP\_Sales) |
|  | ``` |
|  |  |
|  |  |
|  | We reject the null hypothesis and the p-value less than 0.05 shows that each of the regional sales are directly related to all of the other region's sales. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | mytable <- xtabs(~NA\_Sales,Platform) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~NA\_Sales,Year) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~NA\_Sales,Genre) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~NA\_Sales,Publisher) |
|  | chisq.test(mytable) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | mytable <- xtabs(~EU\_Sales,Platform) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~EU\_Sales,Year) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~EU\_Sales,Genre) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~EU\_Sales,Publisher) |
|  | chisq.test(mytable) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | mytable <- xtabs(~JP\_Sales,Platform) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~JP\_Sales,Year) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~JP\_Sales,Genre) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~JP\_Sales,Publisher) |
|  | chisq.test(mytable) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | mytable <- xtabs(~Other\_Sales,Platform) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~Other\_Sales,Year) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~Other\_Sales,Genre) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~Other\_Sales,Publisher) |
|  | chisq.test(mytable) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | mytable <- xtabs(~Global\_Sales,Platform) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~Global\_Sales,Year) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~Global\_Sales,Genre) |
|  | chisq.test(mytable) |
|  | mytable <- xtabs(~Global\_Sales,Publisher) |
|  | chisq.test(mytable) |
|  | ``` |
|  |  |
|  | The chi-square tests help us to further establish our hypothesis that sales vary and depend on the different platforms, genres, year of the game and publisher. |
|  | We reject the null hypothesis and the p-values less than 0.05 shows that all the fields are related. |
|  | All the sales show some or the other correlation with platform, year, genre and publisher of the video games. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | corr.test(game.df[,c(1,7:11)],use="complete") |
|  | ``` |
|  |  |
|  | Rank is negatively correlated to all the sales and each of the sales are highly positively correlated to global sales. America and Europe sales show high levels of correlation (0.77), and so do Europe and Other Sales, with a high positive correlation of 0.73. |
|  |  |
|  | **### Running T-tests** |
|  |  |
|  | Run a t-test to analyse your hypothesis. |
|  | To further prove the hypothesis that the fields are related, which is also indicated by the chi-square tests and cor.tests, we run the t-tests. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | t.test(NA\_Sales,EU\_Sales) |
|  | t.test(NA\_Sales,JP\_Sales) |
|  | t.test(NA\_Sales,Other\_Sales) |
|  | t.test(NA\_Sales,Global\_Sales,paired = TRUE) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | t.test(EU\_Sales,JP\_Sales) |
|  | t.test(EU\_Sales,Other\_Sales) |
|  | t.test(EU\_Sales,Global\_Sales,paired = TRUE) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | t.test(JP\_Sales,Other\_Sales) |
|  | t.test(JP\_Sales,Global\_Sales,paired = TRUE) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | t.test(Other\_Sales,Global\_Sales,paired = TRUE) |
|  | ``` |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | t.test(Rank,NA\_Sales) |
|  | t.test(Rank,EU\_Sales) |
|  | t.test(Rank,JP\_Sales) |
|  | t.test(Rank,Other\_Sales) |
|  | t.test(Rank,Global\_Sales) |
|  | ``` |
|  |  |
|  | Hence, from all the tests that we have carried out, we can find out that, since the p-value is less than 0.05 (all of them being < 2.2e-16), we can reject the null hypothesis. So, the sales of each region are correlated to each other, as well as to the global sales and to the rank of the game. Also, there lie differences between different regional sales based on year and platform mostly, as seen from the graphs and plots. |
|  |  |
|  | **## Appendix 4 - Aggregate, By and apply methods** |
|  |  |
|  | We aggregate various fields to find out relations and patterns between variables. |
|  | Here, our main aim is to compare the different sales, based on the region, platform and year. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | aggregate(NA\_Sales,by=list(Platform = Platform),sum) |
|  | aggregate(NA\_Sales,by=list(Platform = Platform),mean) |
|  |  |
|  | aggregate(NA\_Sales,by=list(Year = Year),sum) |
|  | aggregate(NA\_Sales,by=list(Year = Year),mean) |
|  |  |
|  | aggregate(NA\_Sales,by=list(Genre = Genre),sum) |
|  | aggregate(NA\_Sales,by=list(Genre = Genre),mean) |
|  |  |
|  | aggregate(NA\_Sales,by=list(Publisher = Publisher),sum) |
|  | aggregate(NA\_Sales,by=list(Publisher = Publisher),mean) |
|  | ``` |
|  |  |
|  | DS, PS2, PS3 and Wii are widely used in the US. 2007-2010 have the highest sales in america, taking it to an all time high of 351.44 million. The sales have declined further on. Action, shooter and sports are the widely played categories.505 games, activision, disney interactive studios, electronic arts, konami digital entertainment are the high selling producers. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | aggregate(EU\_Sales,by=list(Platform = Platform),sum) |
|  | aggregate(EU\_Sales,by=list(Platform = Platform),mean) |
|  |  |
|  | aggregate(EU\_Sales,by=list(Year = Year),sum) |
|  | aggregate(EU\_Sales,by=list(Year = Year),mean) |
|  |  |
|  | aggregate(EU\_Sales,by=list(Genre = Genre),sum) |
|  | aggregate(EU\_Sales,by=list(Genre = Genre),mean) |
|  |  |
|  | aggregate(EU\_Sales,by=list(Publisher = Publisher),sum) |
|  | aggregate(EU\_Sales,by=list(Publisher = Publisher),mean) |
|  | ``` |
|  | PS2, PS3 and XBOX360 are widely used in Europe. 2007-2010 have the highest sales in europe, taking it to an all time high of 191.59 million. The sales have declined gradually further on. Action, shooter and sports are the widely played categories.Nintendo, activision, electronic arts, konami digital entertainment, microsoft game studios, sony computer entertainment, ubisoft are the high selling producers. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | aggregate(JP\_Sales,by=list(Platform = Platform),sum) |
|  | aggregate(JP\_Sales,by=list(Platform = Platform),mean) |
|  |  |
|  | aggregate(JP\_Sales,by=list(Year = Year),sum) |
|  | aggregate(JP\_Sales,by=list(Year = Year),mean) |
|  |  |
|  | aggregate(JP\_Sales,by=list(Genre = Genre),sum) |
|  | aggregate(JP\_Sales,by=list(Genre = Genre),mean) |
|  |  |
|  | aggregate(JP\_Sales,by=list(Publisher = Publisher),sum) |
|  | aggregate(JP\_Sales,by=list(Publisher = Publisher),mean) |
|  | ``` |
|  |  |
|  | DS, PS2 and PS are widely used in Japan. 2006 had the highest sales in japan, with 73.73 million units and sales were high during 2007-10. The sales have declined gradually further on. Role-playing is the widely played category in Japan. A few play action and sports. Nintendo, capcom, namcom bandai games, konami digital entertainment, sega, sony computer entertainment are the high selling producers. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | aggregate(Other\_Sales,by=list(Platform = Platform),sum) |
|  | aggregate(Other\_Sales,by=list(Platform = Platform),mean) |
|  |  |
|  | aggregate(Other\_Sales,by=list(Year = Year),sum) |
|  | aggregate(Other\_Sales,by=list(Year = Year),mean) |
|  |  |
|  | aggregate(Other\_Sales,by=list(Genre = Genre),sum) |
|  | aggregate(Other\_Sales,by=list(Genre = Genre),mean) |
|  |  |
|  | aggregate(Other\_Sales,by=list(Publisher = Publisher),sum) |
|  | aggregate(Other\_Sales,by=list(Publisher = Publisher),mean) |
|  | ``` |
|  |  |
|  | DS, PS2, PS3 and Wii are widely used in the Other areas. 2007-2010 have the highest sales in other regions, taking it to an all time high of 82.39 million. The sales have declined further on. Action, shooter and sports are the widely played categories. Activision, Take-Two interactive, THQ, disney interactive studios, electronic arts, konami digital entertainment, Nintendo, Sony entertainment, Ubisoft are the high selling producers. |
|  |  |
|  | ```{r ,message=FALSE,error=FALSE} |
|  | aggregate(Global\_Sales,by=list(Platform = Platform),sum) |
|  | aggregate(Global\_Sales,by=list(Platform = Platform),mean) |
|  |  |
|  | aggregate(Global\_Sales,by=list(Year = Year),sum) |
|  | aggregate(Global\_Sales,by=list(Year = Year),mean) |
|  |  |
|  | aggregate(Global\_Sales,by=list(Genre = Genre),sum) |
|  | aggregate(Global\_Sales,by=list(Genre = Genre),mean) |
|  |  |
|  | aggregate(Global\_Sales,by=list(Publisher = Publisher),sum) |
|  | aggregate(Global\_Sales,by=list(Publisher = Publisher),mean) |
|  | ``` |
|  |  |
|  | DS, PS2, PS3, XBOX360 and Wii are widely used globally. 1996-2000 had high sales. 2007-2010 have the highest sales globally, taking it to an all time high of 678.90 million. The sales have declined gradually further on. Action, shooter and sports are the widely played categories. Activision, Atari, capcom, electronic arts, konami digital entertainment, nintendo, namcoi bandai, sony computer entertainment and ubisoft are the high selling producers. |
|  |  |
|  | **## Appendix 5 - Contingency Tables** |
|  |  |
|  | **### One-Way Contingency Tables** |
|  |  |
|  | Create one-way contingency tables for the categorical variables in your dataset. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <- table(Platform) |
|  | myTable |
|  | ``` |
|  |  |
|  | DS and PS2 are the most popular consoles on which games are released, followed by PS and XBOX360. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <- table(Year) |
|  | myTable |
|  | ``` |
|  |  |
|  | The video game industry has been seeing more video games every year, and the high was seen in the years 2007-2011. Maximum video games were released in 2008 and 2009. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <- table(Genre) |
|  | myTable |
|  | ``` |
|  |  |
|  | Most of the games released were action games, followed by sports, misc, role-playing, shooting and racing. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <- table(Publisher) |
|  | myTable |
|  | ``` |
|  |  |
|  | Activision and Nintendo are having the maximum number of video game releases, with almost 975 and 700 games in the market. |
|  |  |
|  | **### Two-Way Contingency Tables** |
|  |  |
|  | Create two-way contingency tables for the categorical variables in your dataset. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <-xtabs(~Platform+Year) |
|  | library(gmodels) |
|  | addmargins(myTable) |
|  | myTable |
|  | ``` |
|  |  |
|  | The PS was popular between 1998 to 2000, and then came the DS between 2003 and 2007, after which the PSP was more sought after. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <-xtabs(~Platform+Genre) |
|  | addmargins(myTable) |
|  | myTable |
|  | ``` |
|  |  |
|  | The action games are more popular on almost all consoles, but sports games are also played on the PS and the XBOX360. |
|  | Misc. games on DS and racing games on PS2 are also more in numbers. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <-xtabs(~Platform+Publisher) |
|  | addmargins(myTable) |
|  | myTable |
|  | ``` |
|  |  |
|  | Imagic is the publisher that makes more games on 2600 console. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <-xtabs(~Genre+Year) |
|  | addmargins(myTable) |
|  | myTable |
|  | ``` |
|  |  |
|  | Action games have been an all-time favourite, however, misc, role playing and adventure games are also gaining popularity in the recent years. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <-xtabs(~Publisher+Year) |
|  | addmargins(myTable) |
|  | myTable |
|  | ``` |
|  |  |
|  | Acclaim entertainment had a lot of releases earlier, before the 2005. Activision and 505 games have a good amount of game releases over the years. |
|  |  |
|  | ```{r , message=FALSE, warning=FALSE} |
|  | myTable <-xtabs(~Genre+Publisher) |
|  | addmargins(myTable) |
|  | myTable |
|  | ``` |
|  |  |
|  | Atari, Interactive studios and ubisoft are the major contributors to action games. |