Programming for big data

Project

Magda Leszczynska – Student number

Table of Contents

Introduction

Project is made for 'Programming for big data' module and it consist 2 separate analyses: Video Games Sales Analysis and Bitcoin price over years. Both are made based on this same methodology described in Methodology section. Environment: Spyder (Python 2.7) and R on Anaconda.

Packages used:
-tidyverse - ggplot
-pandas
-numpy

Methodology

Above project is based on well-established approach to data analysis - Cross Industry Standard Process for Data Mining (CRISP-DM)[1]. We focus mainly on:

- Business understanding
- Data understanding
- Data Preparation
- Evaluation
- Deployment

1. Video Games Sales Analysis

1.1 Business Understanding and Objectives of Analysis

Video Games business is very competitive business. Recent years shows that entry to that market become much easier than before. But is that make games profitable? What decide about success in different markets like in USA, Europe, Japan and Others? Is there a genre of a game which makes it more profitable than others? Platform, distributor is important? Is the analysis of past data will answer for those questions?

Domain for this analysis is Finance as we will be working on Sales in different markets mostly.

First Objective of Video Games Sales Analysis is to give answers for questions how good data stored in datasets is, how well entering data activates has been made and how accurate data is.

Second Objective is to analyse data to find patterns for example what kind of games and on what platform makes the most of Sales in history. What games are most profitable?

Third objective is test hypothesis that Sales of games depends from Distribution Platform by using relevant statistical test ANOVA. Also to prove that Japanese has different taste in games.

1.2 Data Understanding

Data has been downloaded from Kaggle.com. It was generated by a scrape of vgchartz.com at 26/10/2016. It contains Video games with sales over 100 000 copies.

Dataset contains 16588 entries, there is one integer data column, 6 float and 4 string columns. 11 Columns in Total:

- Rank – the column with ranking of most profitable game to the least profitable game in global - Integer

[1] Azevedo, A. and Santos, M. F. (2008); KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 184.

- Name – Object – String
- Platform – Object – String
- Year - Float
- Genre – Object – String
- Publisher – Object – String
- NA_ Sales – Northern America Sales in millions - Float
- EU_Sales – Europe Sales in millions – Float
- JP_Sales – Japan Sales in millions - Float
- Other_Sales – Other countries sales in millions - Float
- Global_Sales – Global Sales in millions – Float

```
Rank            int64
Name           object
Platform       object
Year          float64
Genre          object
Publisher      object
NA_Sales      float64
EU_Sales      float64
JP_Sales      float64
Other_Sales   float64
Global_Sales  float64
dtype: object
```

Picture 1. Data type

```
Index([u'Rank', u'Name', u'Platform', u'Year', u'Genre', u'Publisher',
       u'NA_Sales', u'EU_Sales', u'JP_Sales', u'Other_Sales', u'Global_Sales'],
      dtype='object')
```

Picture 2. Data headers

From the Example of 5 first rows in Picture above we can see that the most successful games were Wii Sports on Wii Platform realised in 2006. Publisher was Nintendo and globally it made sales of 82.74 million copies. Also we can see that this game was not so popular in Japan. In future chapter of that analysis we will look closer into Japan market as it is little different than the others.

	Rank	Name	Platform	Year	Genre	Publisher	\
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	

	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	41.49	29.02	3.77	8.46	82.74
1	29.08	3.58	6.81	0.77	40.24
2	15.85	12.88	3.79	3.31	35.82
3	15.75	11.01	3.28	2.96	33.00
4	11.27	8.89	10.22	1.00	31.37

Picture 3. Five first rows in Datasets

Dataset information collected from Python you can see on Picture 4:

```
In [34]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1056 entries, 0 to 1055
Data columns (total 8 columns):
Date      1056 non-null object
High      1056 non-null float64
Low       1056 non-null float64
Mid       1056 non-null float64
Last      1056 non-null float64
Bid       1056 non-null float64
Ask       1056 non-null float64
Volume    1056 non-null float64
dtypes: float64(7), object(1)
memory usage: 66.1+ KB
```

Picture 4. Dataset info

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	\
count	16598.000000	16327.000000	16598.000000	16598.000000	16598.000000	
mean	8300.605254	2006.406443	0.264667	0.146652	0.077782	
std	4791.853933	5.828981	0.816683	0.505351	0.309291	
min	1.000000	1980.000000	0.000000	0.000000	0.000000	
25%	4151.250000	2003.000000	0.000000	0.000000	0.000000	
50%	8300.500000	2007.000000	0.080000	0.020000	0.000000	
75%	12449.750000	2010.000000	0.240000	0.110000	0.040000	
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	

	Other_Sales	Global_Sales
count	16598.000000	16598.000000
mean	0.048063	0.537441
std	0.188588	1.555028
min	0.000000	0.010000
25%	0.000000	0.060000
50%	0.010000	0.170000
75%	0.040000	0.470000
max	10.570000	82.740000

Picture 5. Describe the dataset

1.3 Data Preparation

Data preparation started from making a copy of existing file to make sure that we have back up. After examination of data for empty boxes called Nan values we discovered that there are empty values in Year and Publisher Column. We have 271 Nan Values in Year and 58 Nan Values in Publisher Column.

```
Rank            False
Name            False
Platform        False
Year             True
Genre           False
Publisher        True
NA_Sales        False
EU_Sales        False
JP_Sales        False
Other_Sales     False
Global_Sales    False
dtype: bool
Rank              0
Name              0
Platform          0
Year            271
Genre             0
Publisher        58
NA_Sales          0
EU_Sales          0
JP_Sales          0
Other_Sales       0
Global_Sales      0
dtype: int64
[]
```

Picture 6. Nan Values in Dataset

This Nan values were in columns that were not related to our hypothesis. That's why they just have been left as it is. There was no action taken to replace that with 0 or mean value.

All columns have been check for leading space. There was no leading space Also columns has been check for duplicate entries. All duplicates have been agreed that it is normal course of this kind of data.

The abnormal activities has been check in Years Column. We discovered that 3 games were with date realise for future.

After data preparation new file has been saved to make sure that we have back up file.

1.4 Exploratory Analysis

Firstly please see in Picture 1 summary of Video Games Sales Dataset. Rank shows the Overall Revenue from game and we see that max is 16600 it mean that we have less entries than rank number. I believe that it because data contains only games sold more than 100 000 copies. In Name Column we are able to see that game Need for Speed: Most Wanted is in our table 12 times. It is because this game was made on 12 different platforms.

Most common platform is DS – Nintendo with 2163 entries. Second is PlayStation 2 with 2161 games published within period.

The Range of Years within we are dealing it is from 1980 – 2020. The median is year 2007 and mean 2006.

Genre of games tells us what kind of game we have available. The most common is Action game. Second is Sports games.

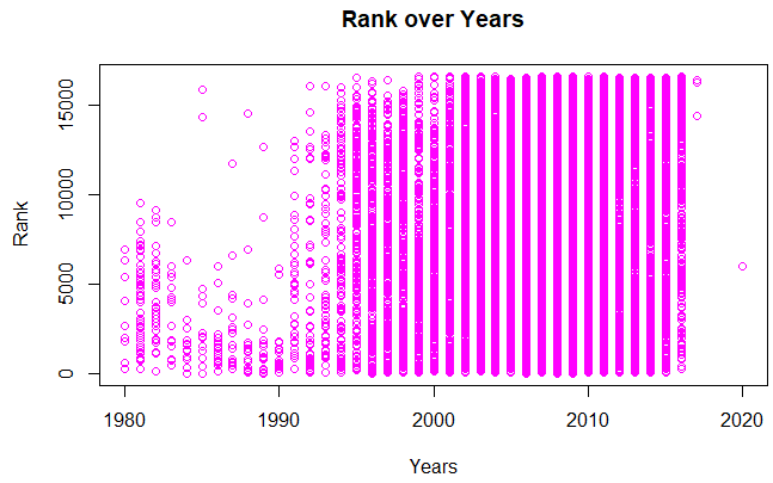
Publisher who made the largest number of profitable games is Electronic Arts, second is Activision.

About Sales we have them divided between region and Global sales.
From that summary we can see that the biggest game sale we had in North America it was 41.49 million. In Europe the biggest was 29.02 million, in Japan it was 10.22 million and other markets maximum was 10.57 million. But compare to the maximum values the mean is very small, it is respectively: 0.2647, 0.1467, 0.07778, 0.04806 in millions.

```
> summary(dat)
       X                Rank                                  Name          Platform          Year     
 Min.   :    0   Min.   :    1   Need for Speed: Most Wanted:   12   DS      :2163   Min.   :1980  
 1st Qu.: 4149   1st Qu.: 4151   FIFA 14                    :    9   PS2     :2161   1st Qu.:2003  
 Median : 8298   Median : 8300   LEGO Marvel Super Heroes   :    9   PS3     :1329   Median :2007  
 Mean   : 8298   Mean   : 8301   Madden NFL 07              :    9   Wii     :1325   Mean   :2006  
 3rd Qu.:12448   3rd Qu.:12450   Ratatouille                :    9   X360    :1265   3rd Qu.:2010  
 Max.   :16597   Max.   :16600   Angry Birds Star Wars      :    8   PSP     :1213   Max.   :2020  
                                 (Other)                    :16542   (Other):7142   NA's   :271   
          Genre                                Publisher       NA_Sales          EU_Sales       
 Action       :3316   Electronic Arts                :1351   Min.   : 0.0000   Min.   : 0.0000  
 Sports       :2346   Activision                     : 975   1st Qu.: 0.0000   1st Qu.: 0.0000  
 Misc         :1739   Namco Bandai Games             : 932   Median : 0.0800   Median : 0.0200  
 Role-Playing:1488   Ubisoft                        : 921   Mean   : 0.2647   Mean   : 0.1467  
 Shooter      :1310   Konami Digital Entertainment   : 832   3rd Qu.: 0.2400   3rd Qu.: 0.1100  
 Adventure    :1286   THQ                            : 715   Max.   :41.4900   Max.   :29.0200  
 (Other)      :5113   (Other)                        :10872                                      
    JP_Sales          Other_Sales        Global_Sales     
 Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.0100  
 1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.: 0.0600  
 Median : 0.00000   Median : 0.01000   Median : 0.1700  
 Mean   : 0.07778   Mean   : 0.04806   Mean   : 0.5374  
 3rd Qu.: 0.04000   3rd Qu.: 0.04000   3rd Qu.: 0.4700  
 Max.   :10.22000   Max.   :10.57000   Max.   :82.7400  
```

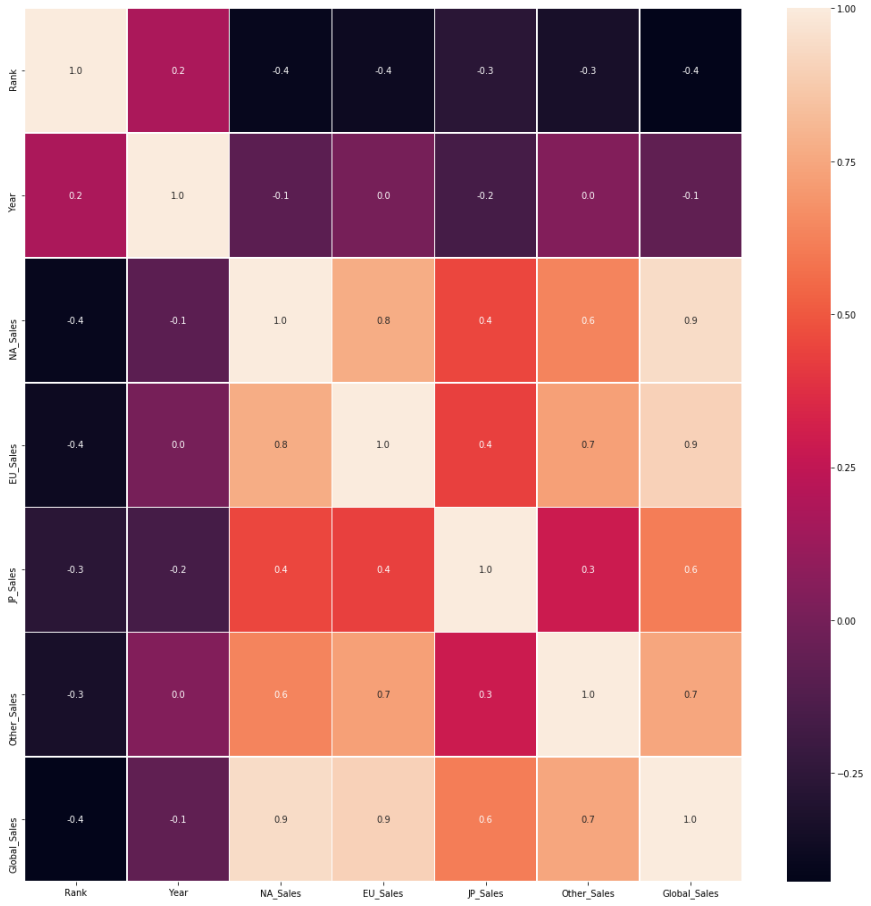
Picture7. Summary of Video Games Sales dataset

Graph 1 show us global rank of most profitable Games over Years. It clears that between 2000 and 2012 there were golden years for sellers and producers of games. It was decade of most profitable games made. Recent years show that the market is slowing down. But we need to take account that data was collected in the middle of 2016.



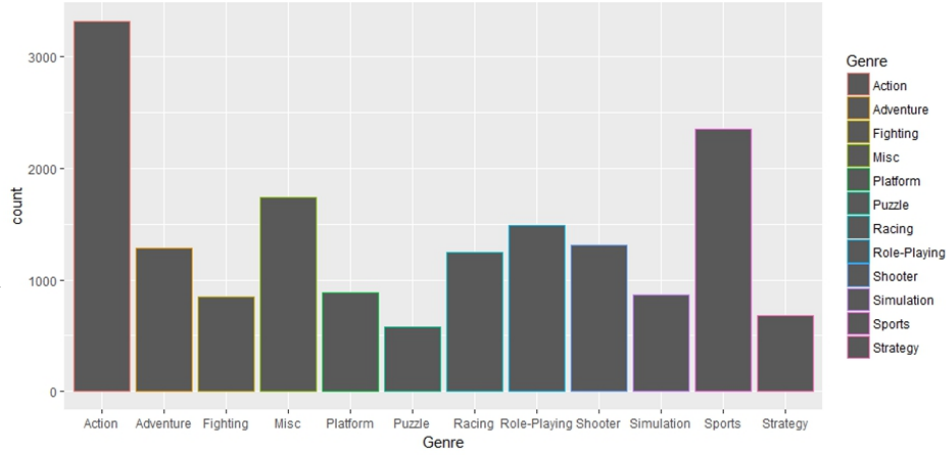
Graph 1. Global Rank of most profitable Games over Years.

This Heat map in Graph 2 show us correlation between separate markets and Rank and Global Market. Zero (purple) mean no correlation. As closer to '1' then the correlation is stronger. For example we see very strong correlation between Global markets and EU markets and NA markets - 0.9. It means that these same games are selling in both markets. These markets have similar taste to games. The smallest correlation is between this markets and Japan market 0.4. Japanese game lovers have different taste. Product for them has to be different.



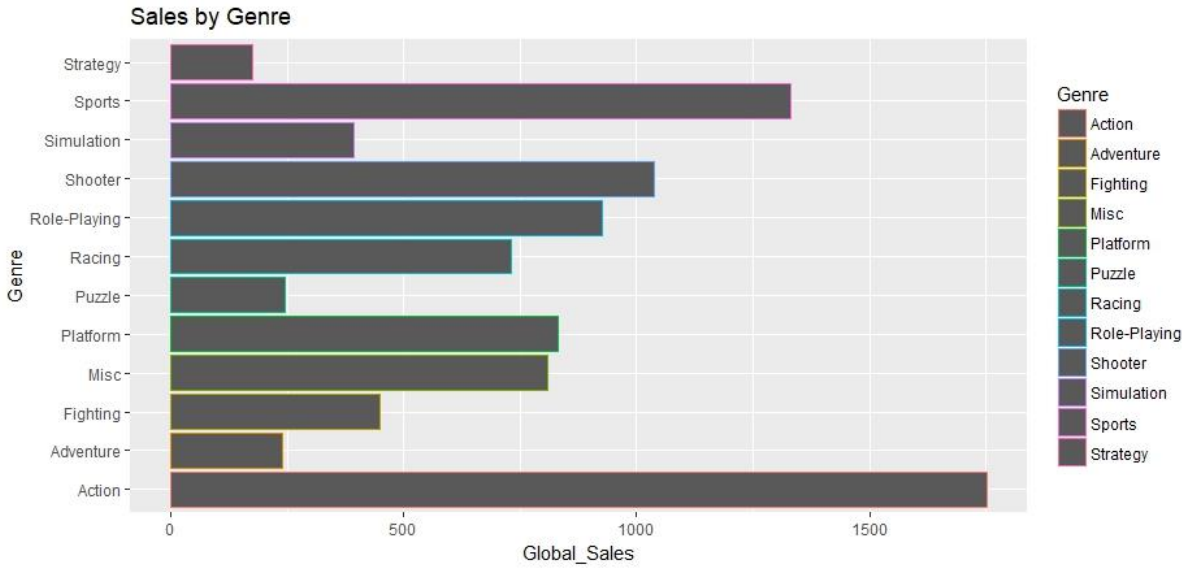
Graph 2. Heatmap of correlation

How many games on specific genre show in our database with sales over 100 000 copies? On Graph 3 we can see that Action games were the most common genre bought in last almost 40 years. Second place have sports games.



Graph 3. Count of games sold by Genre

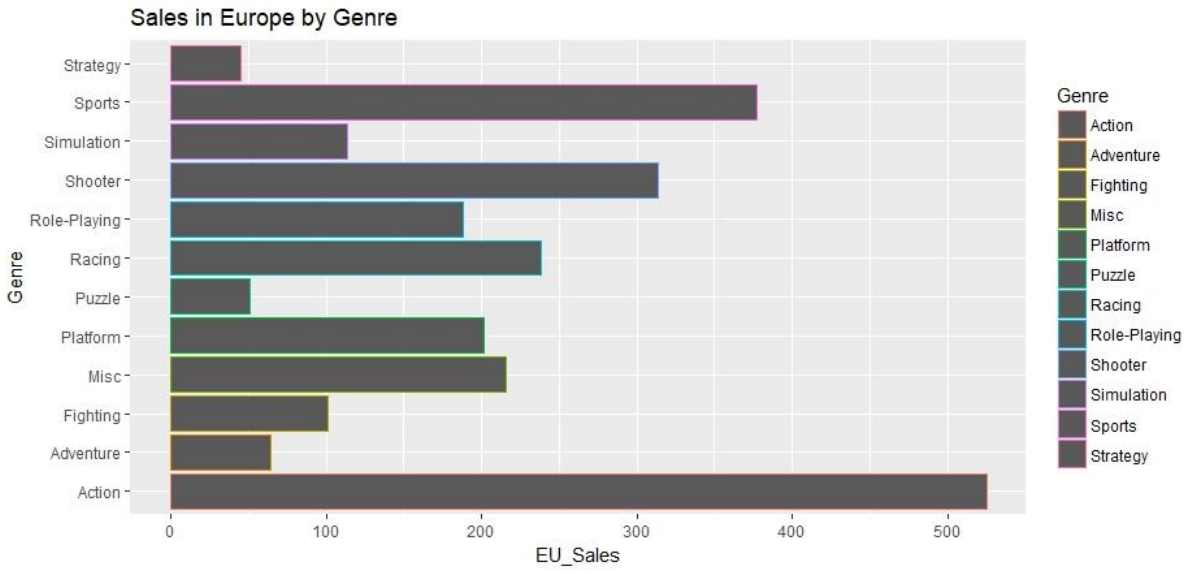
Now we can compare it with Sales per Genre. Is Sales is any different than number of games sold?
The future analysis shows that Sales are similar to number of games sold and still Action Games are the most profitable and most sold games in recent amount 4 decades (Graph 4). The least successful



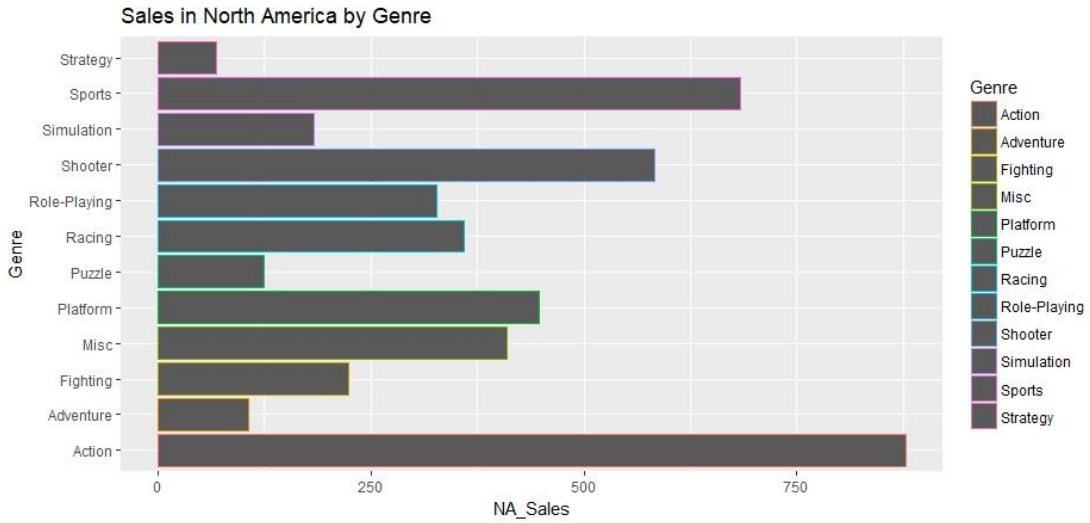
are Puzzle, Strategy and Adventure games.

Graph 4. Global Sales per Genre

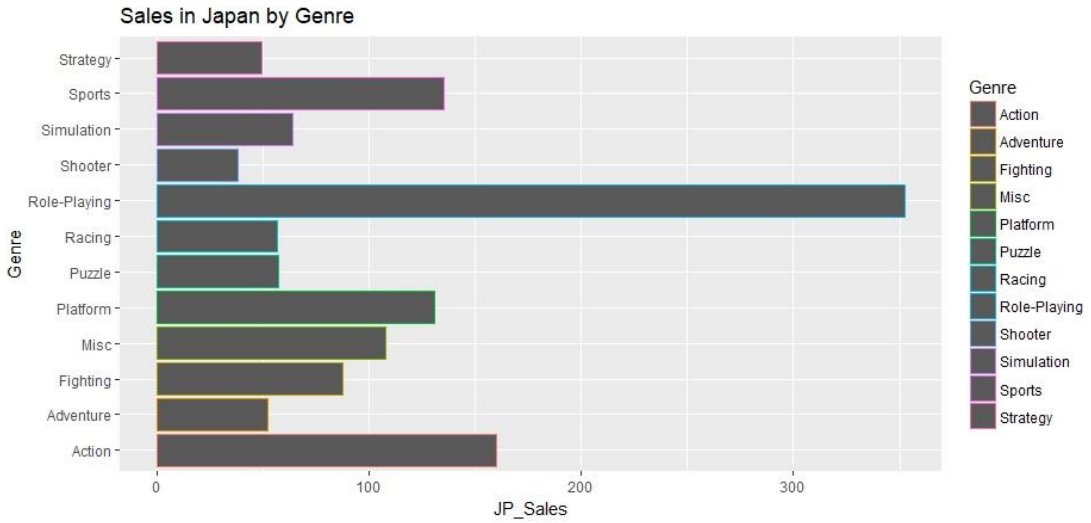
Let's now look closer on specific markets and what genre (kind of) games are most popular in specific regions:



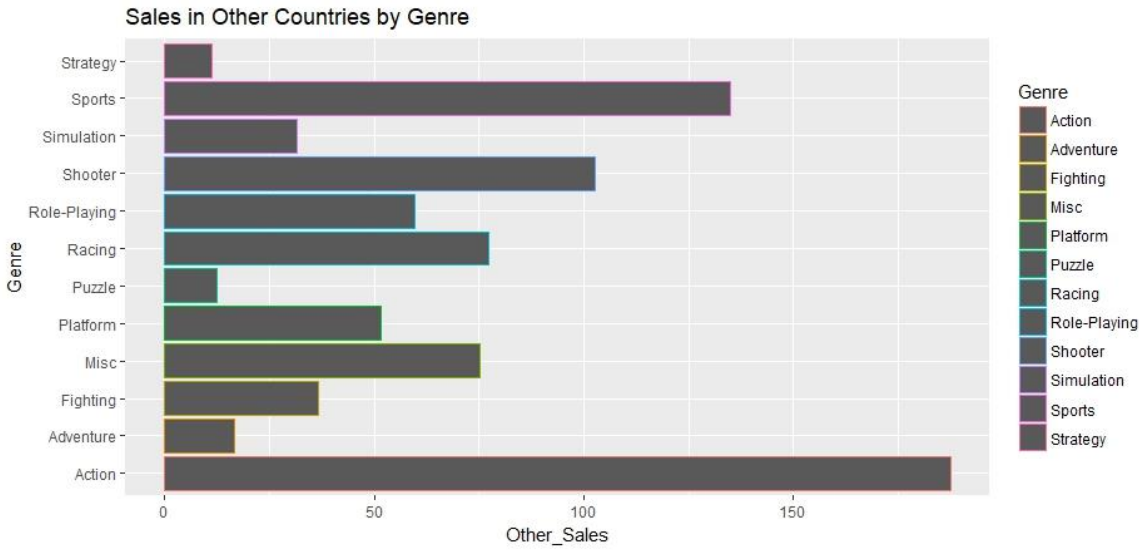
Graph 5. Sales in Europe by Genre



Graph 6. Sales in North America by Genre



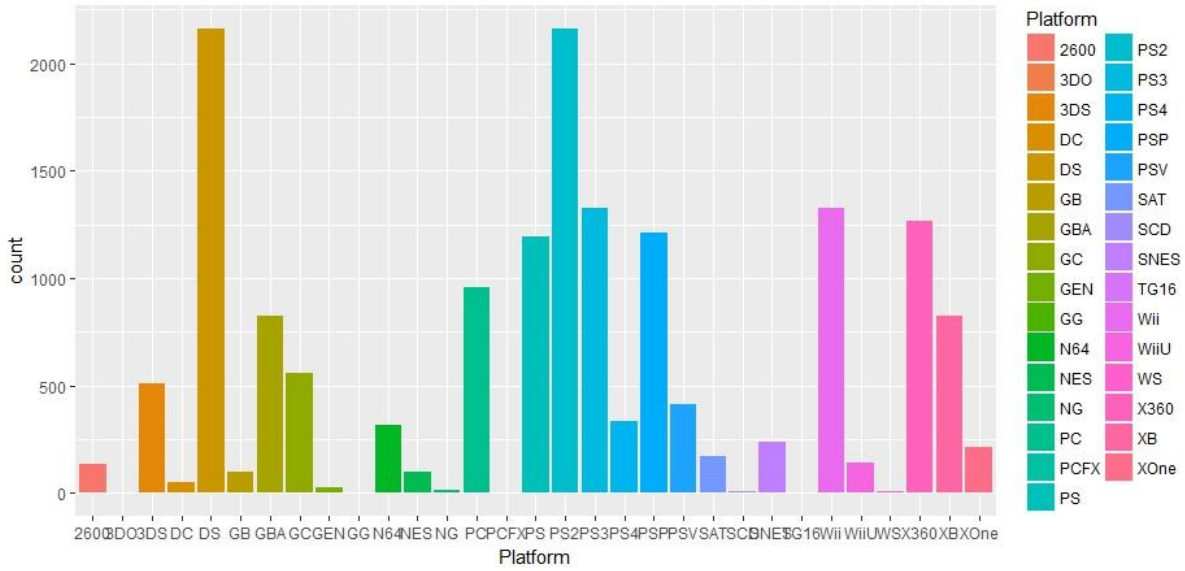
Graph 6. Sales in Japan by Genre



Graph 7. Sales in Other Countries by Genre

It is very interesting but distribution of Sales from games and divided by Genre in different markets show us that Global Sales are similar to markets in North America, Europe and Others but Japan Market it is different. The most successful Genre in that region is Role – Playing and it has more than double sales then Action games.

Addition to that analysis is the small graph about platform used over the last almost 40 years (Graph 8). The most popular platform in Global market is PS2. Second is DS – Nintendo.



Graph 8. Sales by Platform

1.5 Primary analysis – Evaluation and Deployment

Hypothesis that mean of Global Sales and mean of Platform used are equal has been check in ANOVA analysis in R. In Picture above you can see that critical F value is smaller than our F statistical based on p value of 0.05. It means we cannot reject hypothesis zero.
It means that there is a significant correlation between Sales and Platform of Game.

```
> #ANOVA
> anova <- aov(Global_Sales~Platform, data=vg)
> summary(anova)
              Df Sum Sq Mean Sq F value Pr(>F)
Platform      30   1450   48.34    20.7 <2e-16 ***
Residuals   16567  38683    2.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Picture 2. ANOVA

1.6 Challenges whilst handling datasets.

First challenge during analysis of Video Games Analysis was the Nan Values. Firstly after discovering them, calculate how big is this error and secondly what to do with them. I didn't want to skip any rows as other values like Sales and Name of games were present and they could be with significant meaning.

Second challenge was to make graph. Due to distribution of values many of graphs were not readable. On market we could observe that there was few games with very high Sales in millions and then thousands with little above 100 000 sales. All histograms were not bringing any value to the analysis except information about not symmetrical distribution. I decide not to include them. To overcome that challenge I decide to use different kind of graphs. Better choice in that situation was simple plot. After small research I decided to use ggplot from tidyverse package.

Third challenge was with games date. It happens from data cleaning processes that I discovered that 3 games had date in future time. I didn't know what to do with them. Future study is needed. I didn't want to replace it with mean – 2007, as I start to think that maybe there was a game already bought and paid but will be realised later. Like for example pre-sales. But Pre sales with date 2020 seem unreasonable. Anyway no action was taken.

1.7 Video Games Sales Analysis Conclusions

Games business is not easy, before entering to that industry I would recommend to make analysis of markets which the investor would like to enter. Predict the success before invest.

Our short analysis already shows that success of the game depends from genre. We have proven that in our primary analysis. There are more popular kinds like Sports and Action games and the least popular like Strategy and Puzzle also the platform is important as people need to have option to play your game. I would recommend availability on PS and Nintendo as a must in nowadays.

Another factor is region when the entrepreneur is planning to operate as we prove in our hypothesis that Japan market is different.

2. Bitcoin Price over years

2.2 Business Understanding and Objectives of Analysis

Bitcoin is the most famous cryptocurrency. Recent months show that is widely adopted into public and price raised since last year over 2101%[2]. We already can call that economic phenomenon.

Domain for this analysis is Finance as we will be working on Sales in different markets mostly.

First Objective of Bitcoin Price analysis is to give answers for questions how good data stored in datasets is, how well entering data activates has been made and how accurate data is.

Second Objective of this project is to visualize the data set and show the structure of Opening price and closing price for Bitcoin for every single day.

Third objective is test hypothesis that there is a correlation between Volume of transactions on market and price of bitcoin.

2.2 Data Understanding

Data has been downloaded from Kaggle.com
https://www.kaggle.com/fayomi/bitcoinetheriumlitecoin-exchange-price/data.

It is simple dataset of everyday price of bitcoin: High Price, Low Price, Mid Price, Last Price, Bid Price and Asking Price together with Volume. "Volume is the number of shares or contracts traded in a security or an entire market during a given period of time. For every buyer, there is a seller, and each transaction contributes to the count of total volume"[3]

Dataset contains 1056 rows, there is one integer data column, 6 float and 4 string columns. 8 Columns in Total:

-Date – object
-High – float
-Low – float
-Mid – float
-Last – float
-Bid – float
-Ask – float
-Volume – float

```
In [25]: print(data.dtypes)
Date       object
High      float64
Low       float64
Mid       float64
Last      float64
Bid       float64
Ask       float64
Volume    float64
dtype: object
```

```
In [24]: print(data.shape)
(1056, 8)
```

```
In [28]: print(df.columns)
Index([u'Date', u'High', u'Low', u'Mid', u'Last', u'Bid', u'Ask', u'Volume',
       u'mnth_yr', u'YEAR', u'MONTH'],
      dtype='object')
```

[2] http://markets.businessinsider.com/currencies/BTC-USD
[3] https://www.investopedia.com/terms/v/volume.asp

```
In [23]: print(data.head())
       Date    High     Low      Mid    Last     Bid     Ask        Volume
0  01-01-15  322.90  313.81  315.130  315.03  315.06  315.20   7523.671787
1  02-01-15  316.74  313.28  315.185  315.20  315.17  315.20   3784.155803
2  03-01-15  316.01  286.49  288.015  287.80  287.81  288.22  38745.725020
3  04-01-15  290.78  255.03  260.675  260.62  260.61  260.74  87644.268260
4  05-01-15  279.50  258.06  278.985  279.50  278.08  279.89  46559.163410
```

```
In [22]: print(df.tail())
           Date     High          Low      Mid         Last      Bid      Ask  \
1051 2017-12-07  13599.0  11590.00000  13508.0  13503.00000  13507.0  13509.0
1052 2017-12-08  16649.0  13139.00000  16610.0  16607.00000  16608.0  16612.0
1053 2017-12-09  17171.0  13722.00000  15833.0  15816.78052  15822.0  15844.0
1054 2017-12-10  16300.0  13010.00000  14660.5  14660.00000  14660.0  14661.0
1055 2017-12-11  15741.0  12730.63686  14932.0  14937.00000  14926.0  14938.0

            Volume        mnth_yr  YEAR  MONTH
1051   83642.06608  December-2017  2017     12
1052  141225.69290  December-2017  2017     12
1053  124036.20960  December-2017  2017     12
1054   77819.74691  December-2017  2017     12
1055  107145.64600  December-2017  2017     12
```

2.3 Data Preparation

Data preparation started from making a copy of existing file to make sure that we have back up. Column with Dates has been check for entries. We had to make sure that it contains only Dates, no empty spaces and no names or any other values. All columns have been check for leading space. There was no leading space

Data has been check for empty values – Nan Values. There were no Nan values in dataset.

```
In [29]: print(df.isnull().any())
Date       False
High       False
Low        False
Mid        False
Last       False
Bid        False
Ask        False
Volume     False
mnth_yr    False
YEAR       False
MONTH      False
dtype: bool
```

In addition we check duplication in dates by counting data. All data entered was unique.

```
2017-11-11    1
2017-11-12    1
2017-11-13    1
2017-11-14    1
2017-11-15    1
2017-11-16    1
2017-11-17    1
2017-11-18    1
2017-11-19    1
2017-11-20    1
2017-11-21    1
2017-11-22    1
2017-11-23    1
2017-11-24    1
2017-11-25    1
2017-11-27    1
2017-11-28    1
2017-11-29    1
2017-11-30    1
2017-12-01    1
2017-12-02    1
2017-12-03    1
2017-12-04    1
2017-12-05    1
2017-12-06    1
2017-12-07    1
2017-12-08    1
2017-12-09    1
2017-12-10    1
2017-12-11    1
Name: Date, Length: 1056, dt
```

Also in part of preparation First column contains Date has been reformatted for datetime in Pandas. After importing datetime (DatetimeIndex).

```
[31]: df['Date'] = pd.to_datetime(df['Date'], dayfirst=True)

[32]: df['mnth_yr'] = df['Date'].apply(lambda x: x.strftime('%B·
  ...: df['YEAR'] = pd.DatetimeIndex(df['Date']).year
  ...: df['MONTH'] = pd.DatetimeIndex(df['Date']).month
```

For better readable graphs I decided to create a new column with only months and year called ‘mnth_yr’.

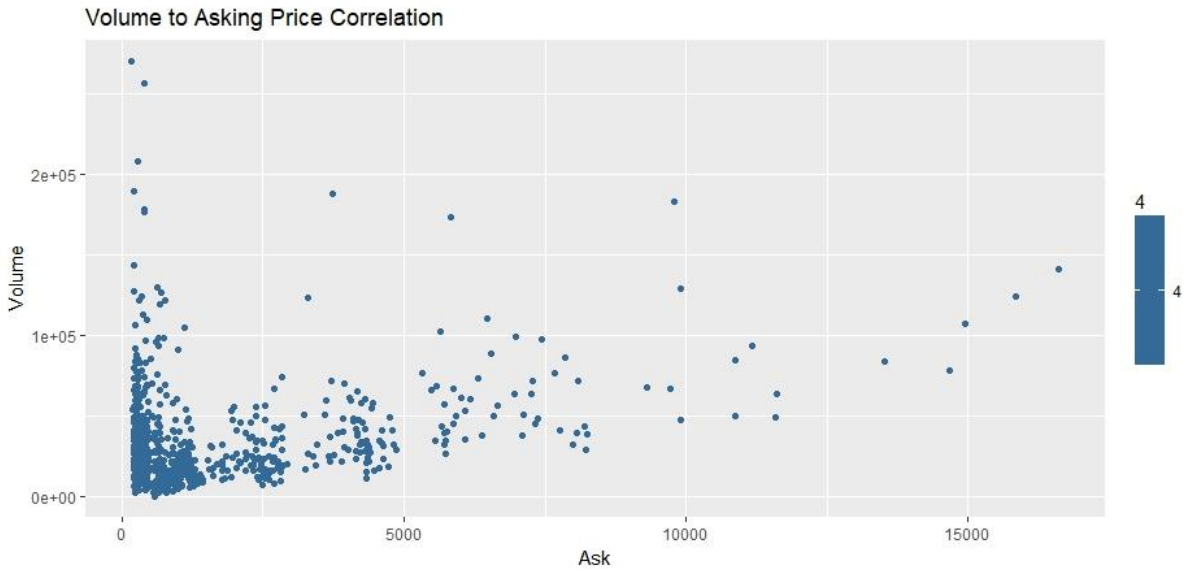
Also data has been check for extraordinary values. Values of bitcoin has been check for any entries above 200 000k. There were 3 entries.

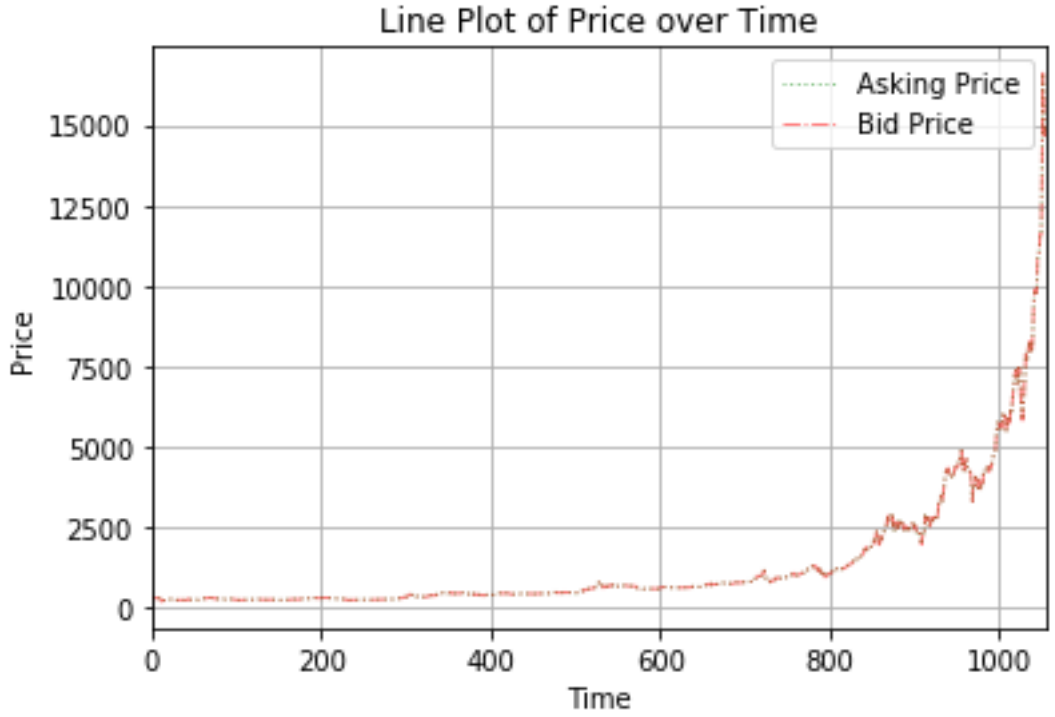
All seems to be valid. No data has been skipped.

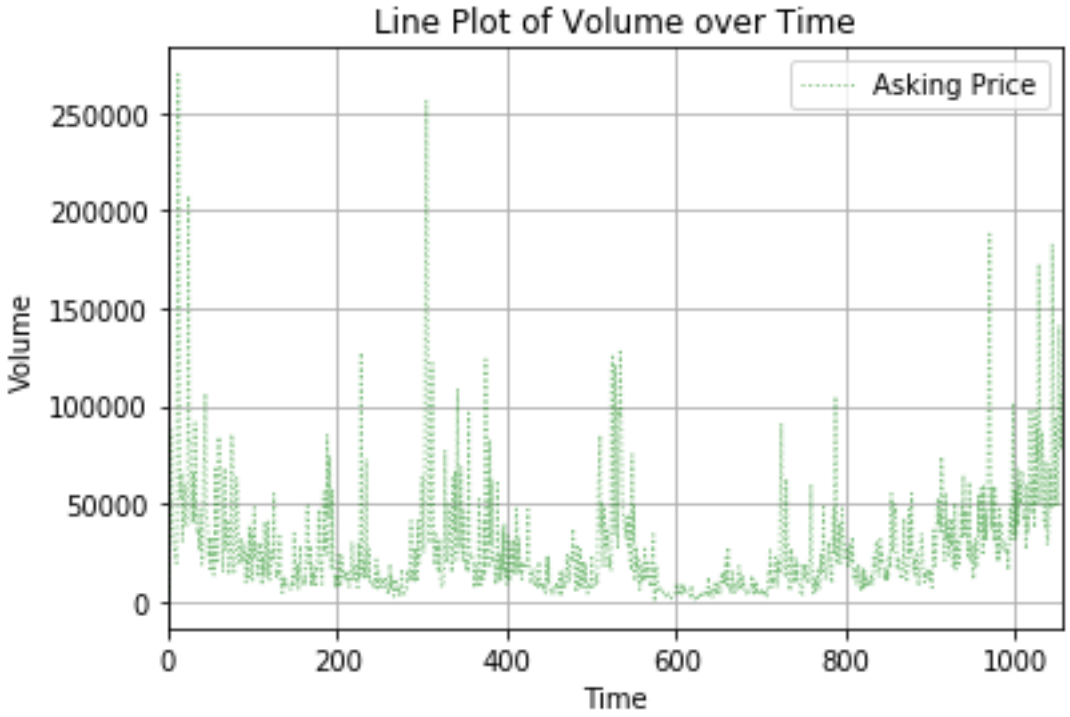
	Date	High	Low	Mid	Last	Bid	Ask	Volume
12	14-01-15	229.99	166.45	177.980	178.00	177.96	178.00	269676.7632
24	26-01-15	315.00	252.04	274.145	274.11	274.11	274.18	207546.8671
305	04-11-15	504.00	366.66	391.995	392.99	391.00	392.99	256445.1394

All new documents have been saved in separate file 'bitcoin_clean_data.csv'.

2.4 Exploratory Analysis


Volume to Asking Price Correlation
Volume
2e+05
1e+05
0e+00
0
5000
10000
15000
Ask
4
4


Line Plot of Price over Time
Asking Price
Bid Price
Price
15000
12500
10000
7500
5000
2500
0
0
200
400
600
800
1000
Time



2.6 Primary analysis – Evaluation and Deployment

2.5 Challenges whilst handling datasets.

First challenge while making analysis of this datasets was Column with Date. On beginning entries were strings and when making graphs it was not readable. I had to convert it into datetime in Pandas but after simple code of ‘df['Date'] = pd.to_datetime(df['Date'])’ it change first 12 days in month as 1 of January, 1 of February, 1 of March,… and then rest of month was correct. After some research about that Index I decide to use second factor ‘dayfirst=True’ in the conversion index.

2.7 Bitcoin Analysis Conclusions

References

1. Azevedo, A. and Santos, M. F. (2008); KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 184.
2. http://markets.businessinsider.com/currencies/BTC-USD

3. Investopedia - https://www.investopedia.com/terms/v/volume.asp
4. Kaggle.com

Appendices

1. Python script for Video games analysis
2. Python script for Bitcoin analysis
3. R script for Video games analysis
4. R script for Bitcoin analysis
5. Dataset of Bitcoin Price
6. Dataset of Video Games Sales