Project Report – Video Games Analytics

1. Business objective
2. Exploratory Data Analysis
    a. Describe the data / Descriptive Statistics
    b. EDA
        i. Univariate analysis
        ii. Bivariate Analysis
        iii. Missing Value Treatment
        iv. Outlier Treatment
        v. Finding useful patterns and trends
        vi. Feature creation (if required)
        vii. Data Transformation (if required)
    c. Verify data quality
3. Data Preparation & Feature Engineering
    a. Data Imputation
    b. Features Creation
4. Modelling
    a. Select modelling technique
    b. Build and Assess the model
5. Evaluation / Executive Summary
6. Further Analysis

1. Business Objective

- Analyse and present the data analysis
- Predict
    - Probable sales for a *role-playing* (genre) game developed by EA (publishar)
    - Most profitable platform for developing a *Shooter* genre game (optional)s

Clarifications:

1. Period (years range) for which I have to forecast sales? The year when the game is published, Assume 2018
2. Do I need to do prediction / forecast for only TotalsSales? Yes
3. Publisher also is not "EA", it's "Square EA". Should I consider Square EA instead of "EA"?
   Square EA and EA Sports can be assumed to be the same company.

File	**platforms.csv**
Column	*Description*
Index	Row index
Rank	Platform rank
Platform	Platform name
HardwareSales	Hardware sales cumulative
SoftwareSales	Software sales cumulative
Games	Number of games available
File	**games.csv**
Column	*Description*
index	Row index
Name	Name the game
Platform_score	Game Platform
Year	Development year (First release)
Genre	Game genre
Publisher	Game developer
NorthAmericaSales	Sales in USA (million)
EuropeSales	Sales in Europe (million)
JapanSales	Sales in Japan (million)
RowSales	Sales in rest of the world (million)
TotalSales	Total sales (million)
VGScore	Rating given to the game by a popular website
CriticScore	Critics Rating
UserScore	Users feedback

2. EDA (Exploratory Data Analysis)

1. Descriptive Statistics

- The dataset used in this project has 78 platforms and 83,545 records in games for almost 40 years i.e. from 1980 till 2020

platforms.csv –

- 77 records with 7 attributes

games_data.csv –

- 83545 records with 15 attributes
- Out of 15 only 4 features has NA / missing data.
- Year has around 15% of missing data
- VGScore , CriticScore and userScore has more than 90% missing data. So imputing and use it is not possible. We tried to use them later analysis. Ignoring them for time being.

```
index          index.1                       Name           Year             Genre
 Min.   :    0   Min.   :    0   Plants vs. Zombies:  273   Min.   :1980    Misc     :16816
 1st Qu.: 9899   1st Qu.: 9899   Monopoly          :  210   1st Qu.:1998    Action   :13103
 Median :22284   Median :22284   Double Dragon     :  182   Median :2007    Sports   : 9770
 Mean   :22160   Mean   :22160   Space Invaders    :  144   Mean   :2004    Shooter  : 7019
 3rd Qu.:32946   3rd Qu.:32946   Angry Birds       :  143   3rd Qu.:2011    Platform : 6674
 Max.   :44646   Max.   :44646   Elite             :  132   Max.   :2020    Adventure: 6230
                                 (Other)           :82461   NA's   :12007   (Other)  :23933
          Publisher     NorthAmericaSales  EuropeSales       JapanSales         RowSales
               :11401   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.00000   Min.   : 0.00000
Sega           : 3800   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.: 0.00000
Activision     : 3779   Median : 0.0000   Median : 0.0000   Median :0.00000   Median : 0.00000
Electronic Arts: 3237   Mean   : 0.1313   Mean   : 0.0769   Mean   :0.01924   Mean   : 0.02693
Ubisoft        : 2950   3rd Qu.: 0.0800   3rd Qu.: 0.0200   3rd Qu.:0.00000   3rd Qu.: 0.01000
EA Sports      : 2443   Max.   :41.3600   Max.   :29.0100   Max.   :5.66000   Max.   :10.57000
(Other)        :55935
  TotalSales       Platform_score     VGScore       CriticScore      UserScore
Min.   : 0.0000   PC     :10574   N/A    :82437   N/A    :71311   N/A    :83115
1st Qu.: 0.0000   PS2    : 4661   8.4    :   84   8.0    :  606   9.0    :   48
Median : 0.0000   PS3    : 4626   8.0    :   74   7.0    :  563   8.0    :   46
Mean   : 0.2545   X360   : 4352   8.8    :   64   7.5    :  471   9.1    :   26
3rd Qu.: 0.1500   DS     : 3523   7.0    :   55   9.0    :  436   9.3    :   25
Max.   :82.6500   PS     : 3523   8.6    :   54   8.5    :  424   9.5    :   25
       (Other):52286   (Other):  777   (Other): 9734   (Other):  260
```

```
> sapply(df_main_games, class)
       index        index.1           Name           Year          Genre
   "integer"      "integer"       "factor"      "numeric"       "factor"
   Publisher NorthAmericaSales     EuropeSales     JapanSales        RowSales
    "factor"      "numeric"       "numeric"      "numeric"       "numeric"
  TotalSales   Platform_score        VGScore     CriticScore       UserScore
   "numeric"       "factor"       "factor"       "factor"        "factor"
```

Let us check the data format / attribute format -

Structure of the platform data:

```
> str(df_main_pf)
'data.frame':       77 obs. of  7 variables:
 $ index         : int  0 1 2 3 4 5 6 7 8 9 ...
 $ index.1       : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Rank          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Platform      : Factor w/ 77 levels "2600","2DS","3DO",..: 59 74 60 70 58 22 61 45 28 67 ...
 $ HardwareSales: num  157.7 85.8 86.9 101.6 104.2 ...
 $ SoftwareSales: num  1662 1008 975 966 962 ...
 $ Games         : num  3549 3678 3316 2809 2680 ...
```

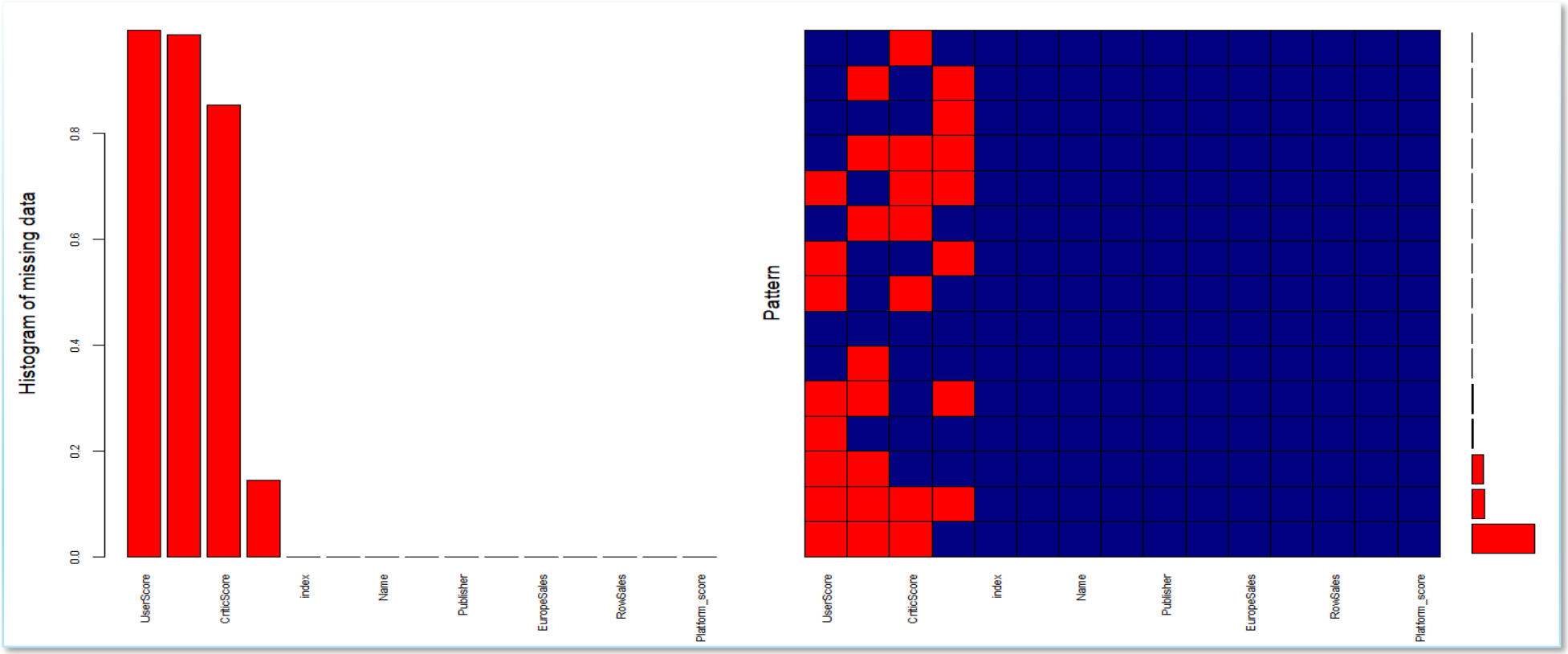
Structure of the games data :

```
> str(df_main_games)
'data.frame':       83545 obs. of  15 variables:
 $ index            : int  0 0 0 0 0 0 0 0 3920 3920 ...
 $ index.1          : int  0 0 0 0 0 0 0 0 3920 3920 ...
 $ Name             : Factor w/ 23623 levels "'70s Robot Anime: Geppy-X",..: 9003 9003 9003 9003 9003 9003 9003
9003 9003 9003 ...
 $ Year             : num  2004 2004 2004 2004 2004 ...
 $ Genre            : Factor w/ 17 levels "Action","Action-Adventure",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Publisher        : Factor w/ 1474 levels "","10TACLE Studios",..: 1080 1080 1080 1080 1080 1080 1080 1080 108
0 1080 ...
 $ NorthAmericaSales: num  9.43 9.43 9.43 9.43 9.43 9.43 9.43 9.43 0.09 0.09 ...
 $ EuropeSales      : num  0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.02 0.02 ...
 $ JapanSales       : num  0.41 0.41 0.41 0.41 0.41 0.41 0.41 0.41 0 0 ...
```

```
$ RowSales         : num  10.6 10.6 10.6 10.6 10.6 ...
$ TotalSales       : num  20.8 20.8 20.8 20.8 20.8 ...
$ Platform_score   : Factor w/ 77 levels "2600","3DO","3DS",..: 55 74 51 75 56 49 70 10 55 74 ...
$ VGScore          : Factor w/ 60 levels "2.6","3.0","3.1",..: 60 60 60 60 60 60 60 60 60 60 ...
$ CriticScore      : Factor w/ 87 levels "1.0","1.3","1.4",..: 82 79 81 87 87 87 87 87 82 79 ...
$ UserScore        : Factor w/ 41 levels "10.0","2.0","3.8",..: 41 41 41 41 41 41 41 41 41 41 ...
```

2. Missing Data Analysis (R Package Mice)

Usually thumb rule is that if the data less than 5% of missing values, one can impute those values by either mean or median values of the column or the most frequent value in the column.Let us look at the missing values graph.



Highlight:

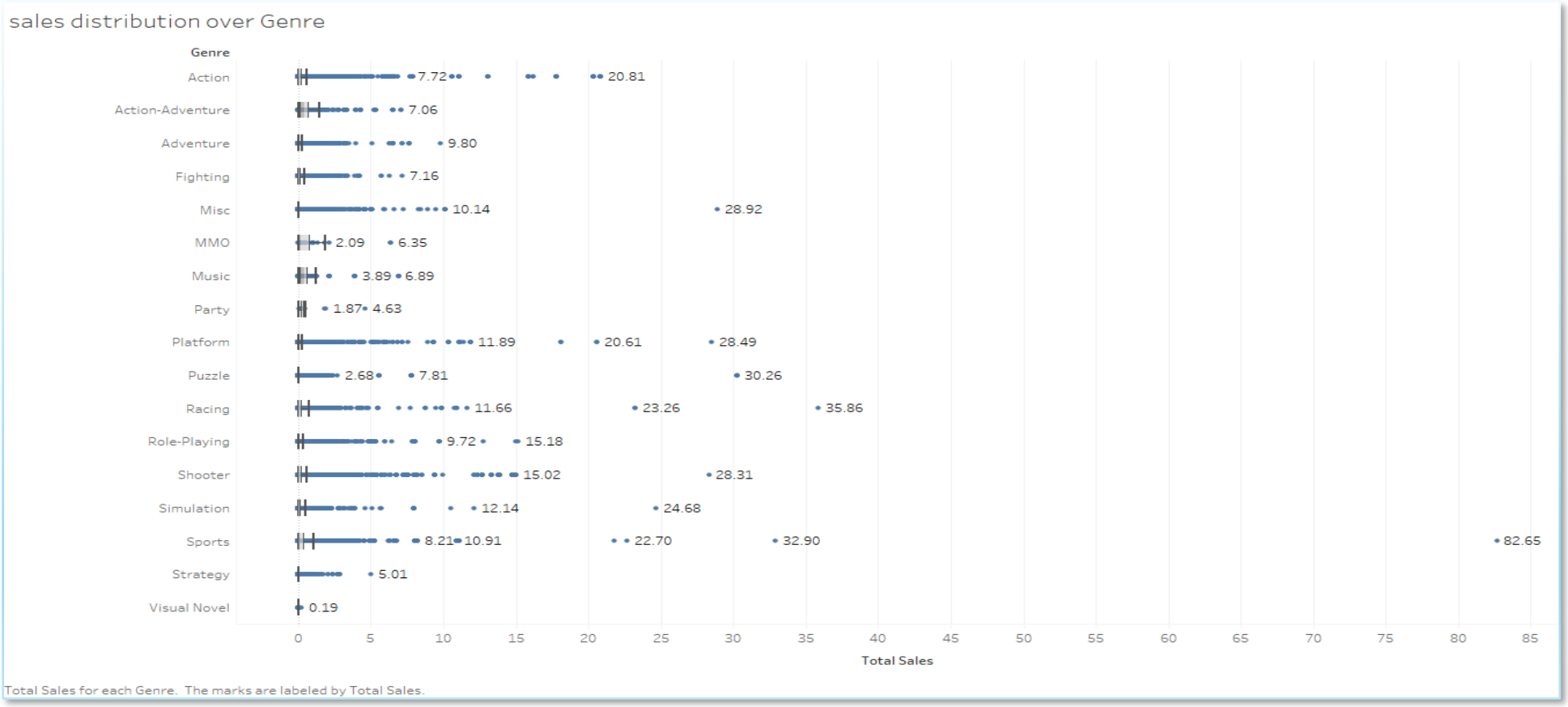
- As we can see from following bar chart **that more than 15% of data is missing** for year and same is true for VG, critic and user score. So, for time being lets drop idea of data imputation for these columns as huge amount of records are missing for these columns.

3. Outliers:

In platform data, iOS with more than 200k games can be considered as outlier.
As finally we have deal with Genre for prediction, I just wanted to see the spread of sales over Genre.

Highlight:

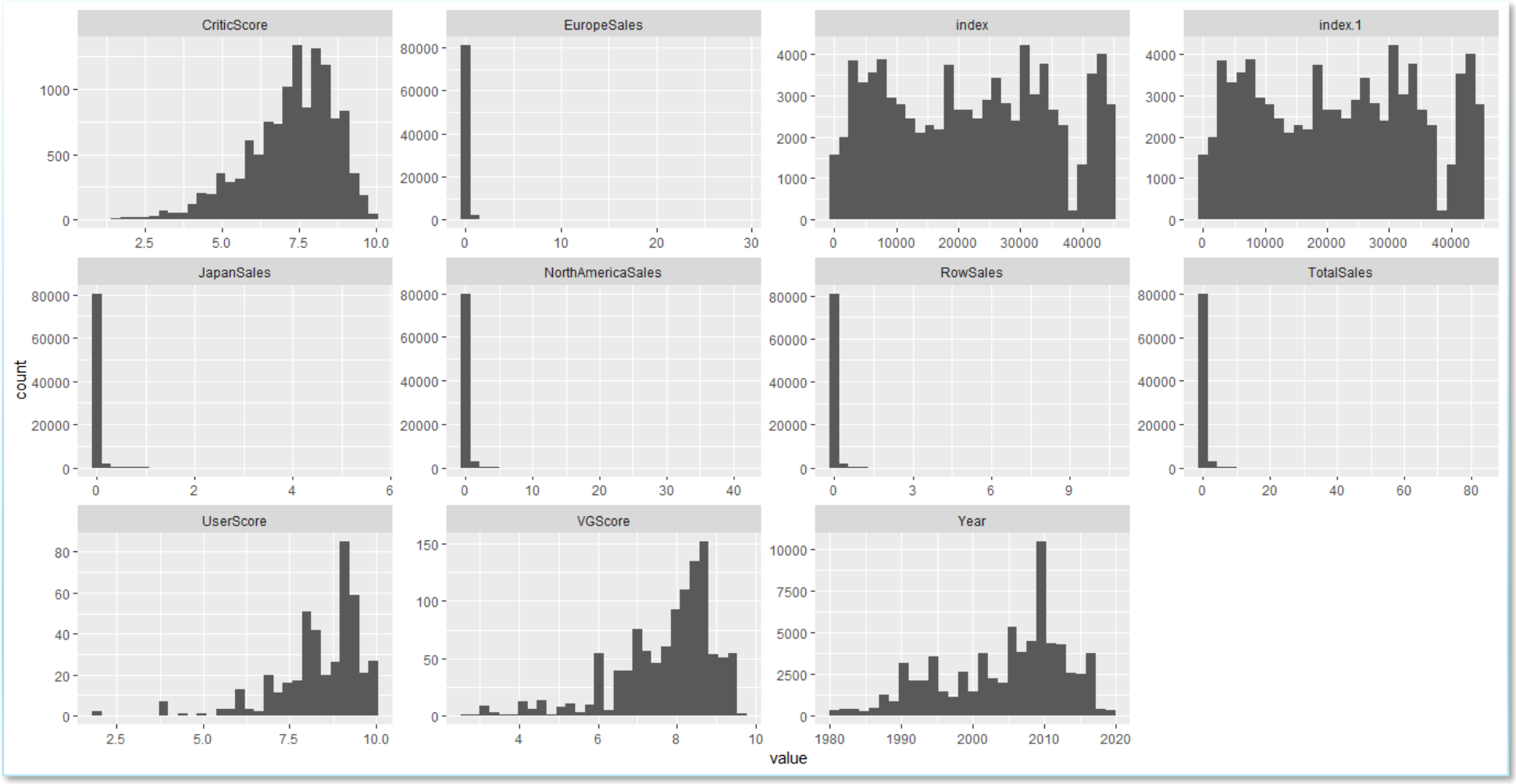
- We can see the clearly outlier over here for “Sports” Genre. Other than that, spread is normal.



Let us check the density of numeric data. Is data normally distributed or skedded.

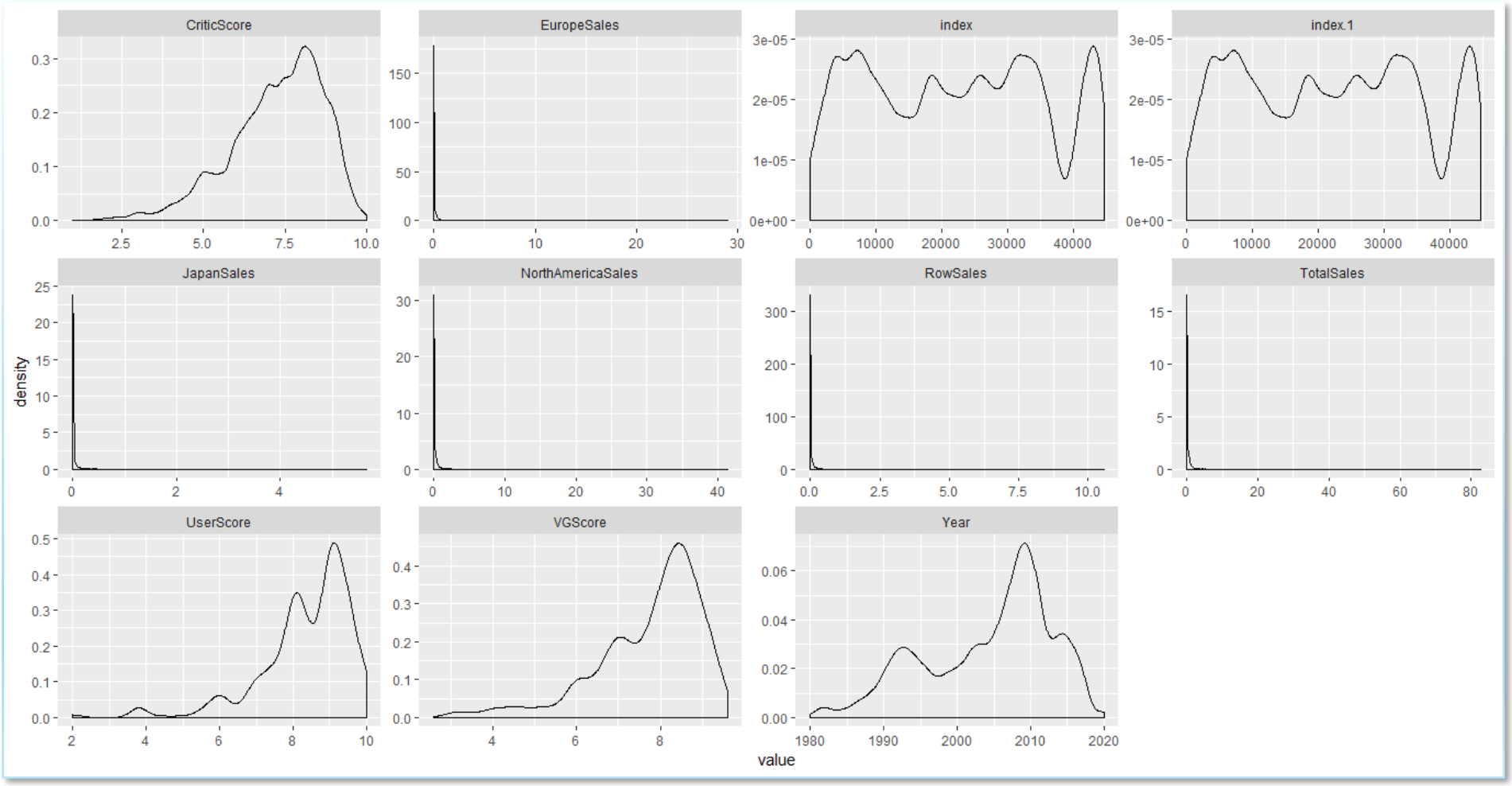
Highlight:

- There are some outliers with score and can be removed while model creation stage.



Highlight:

- All the scores are Right Skewed and can be used for prediction by normalizing them.



C. Verification of data quality

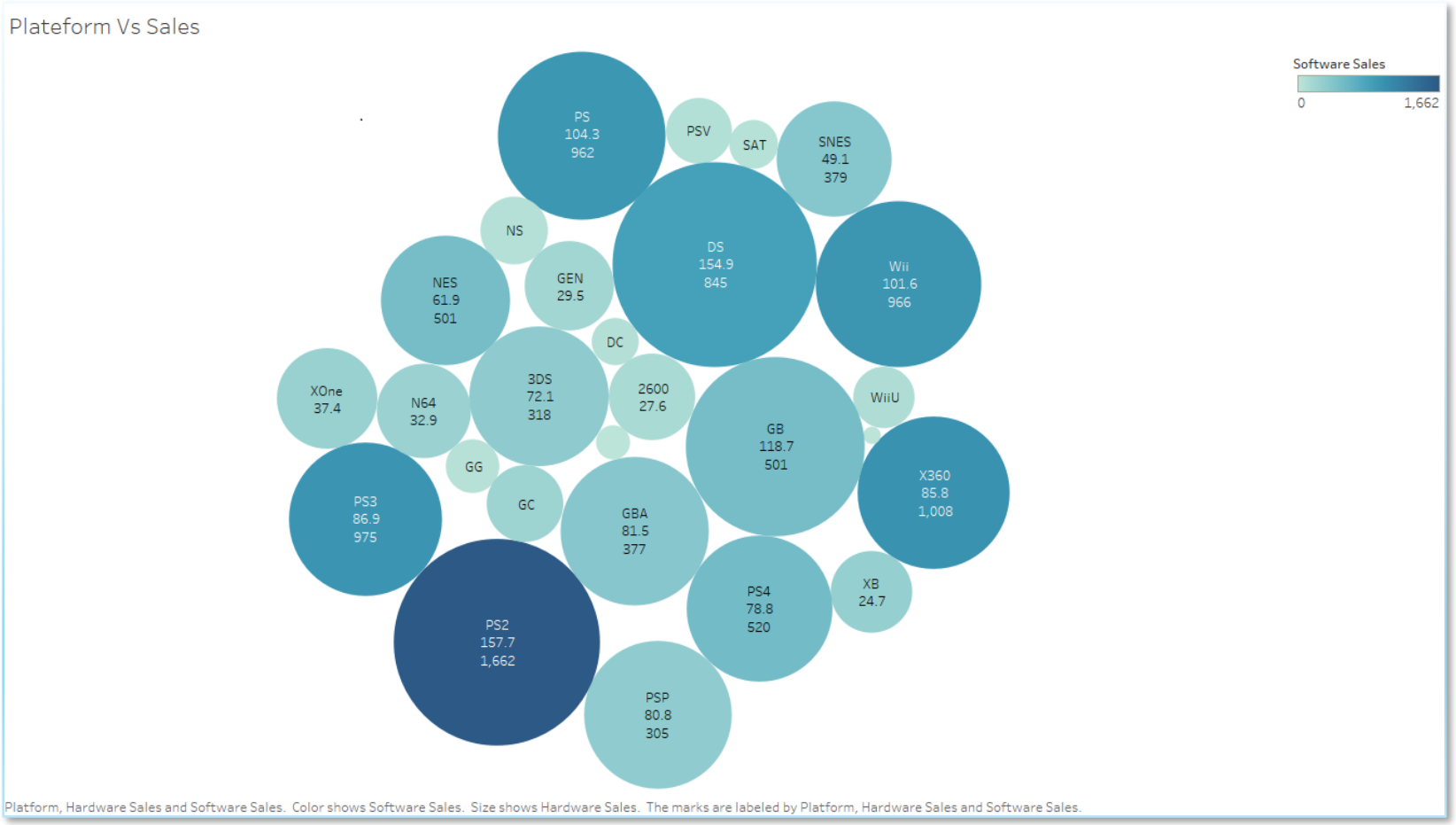
In general, it was found that the for few important feature, data is missing a lot. Imputation give biased results if we do it for more than 5% data.

Finding useful pattern and trends (Univariate/Bivariate Analysis):

1. Let us check platform related data. Hardware and Software sales against Platform.

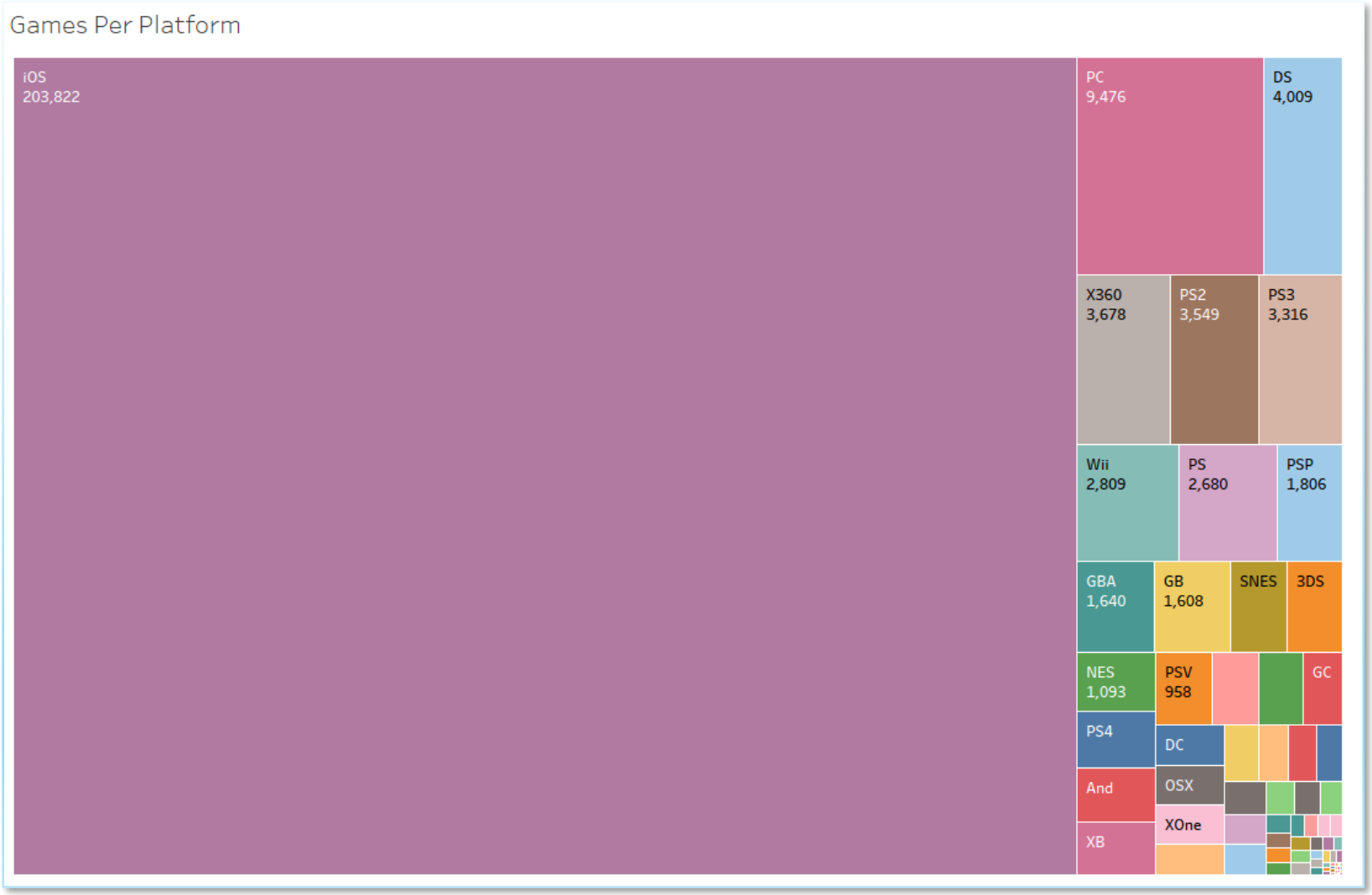
Highlights:

- PS platform is ahead and leader in market



Highlights:

- iOS platform has huge no of games on their platform



2. Though PS platform have only approximately. 5% share in total games but it has more than 40 % share in overall sales.

Platform	HardwareSales	SoftwareSales	Games	total_sales	games_per	sales_per
PS2	157.68	1661.95	3549	1819.63	1.393803485	14.2789097
PS	104.25	962.01	2680	1066.26	1.052519961	8.36710224
PS3	86.9	974.58	3316	1061.48	1.302297086	8.32959286
PS4	78.82	520.19	1049	599.01	0.411975164	4.70052136
PSP	80.82	304.61	1806	385.43	0.709272779	3.02452705
PSV	16	68.5	958	84.5	0.376236613	0.66308418

Highlights:

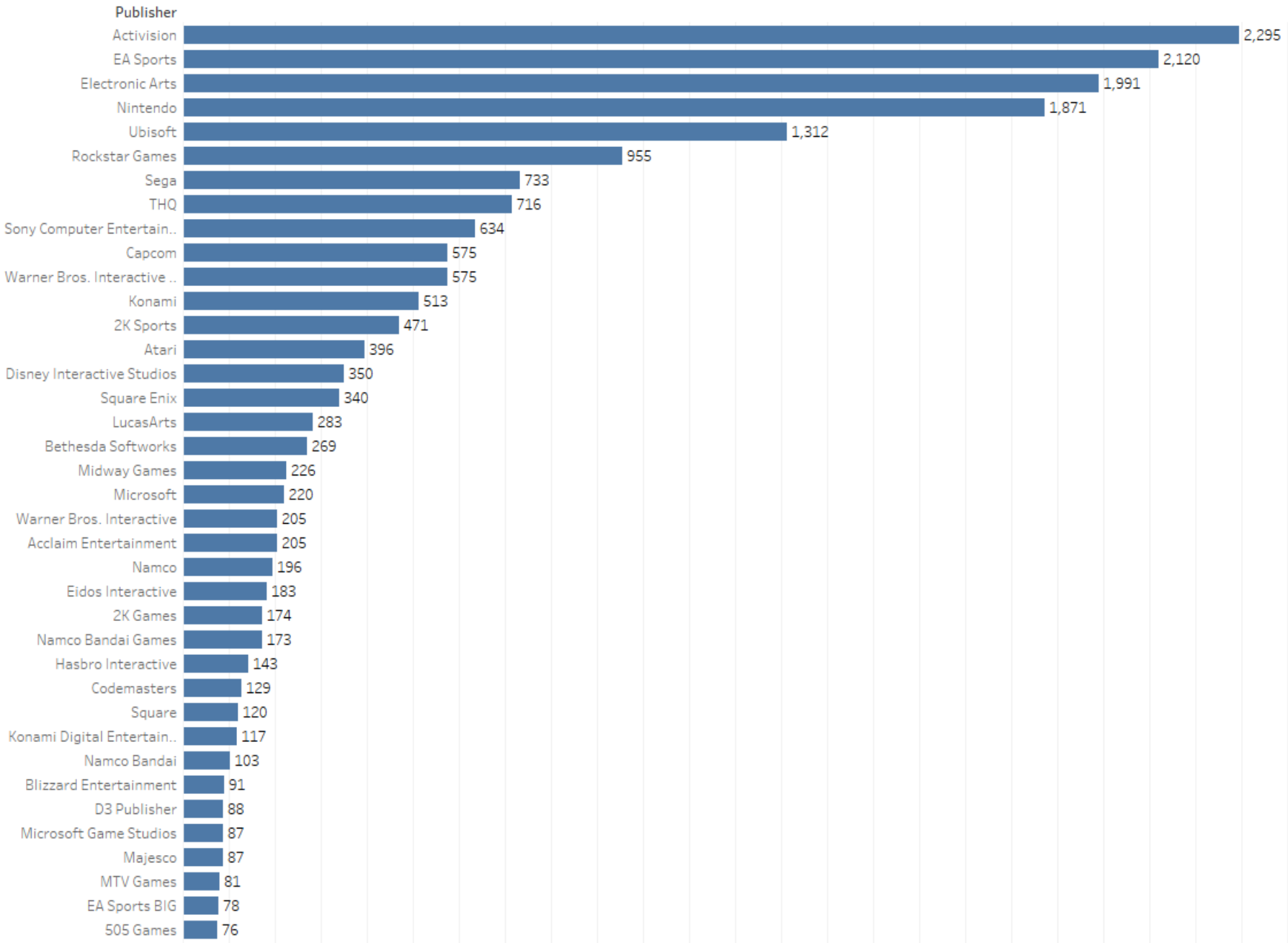
- PS platform is ahead and leader in sales which confirm our first conclusion.

3. Let us check for the big publishers

Highlights:

- Activision, EA Sports are among the top sales

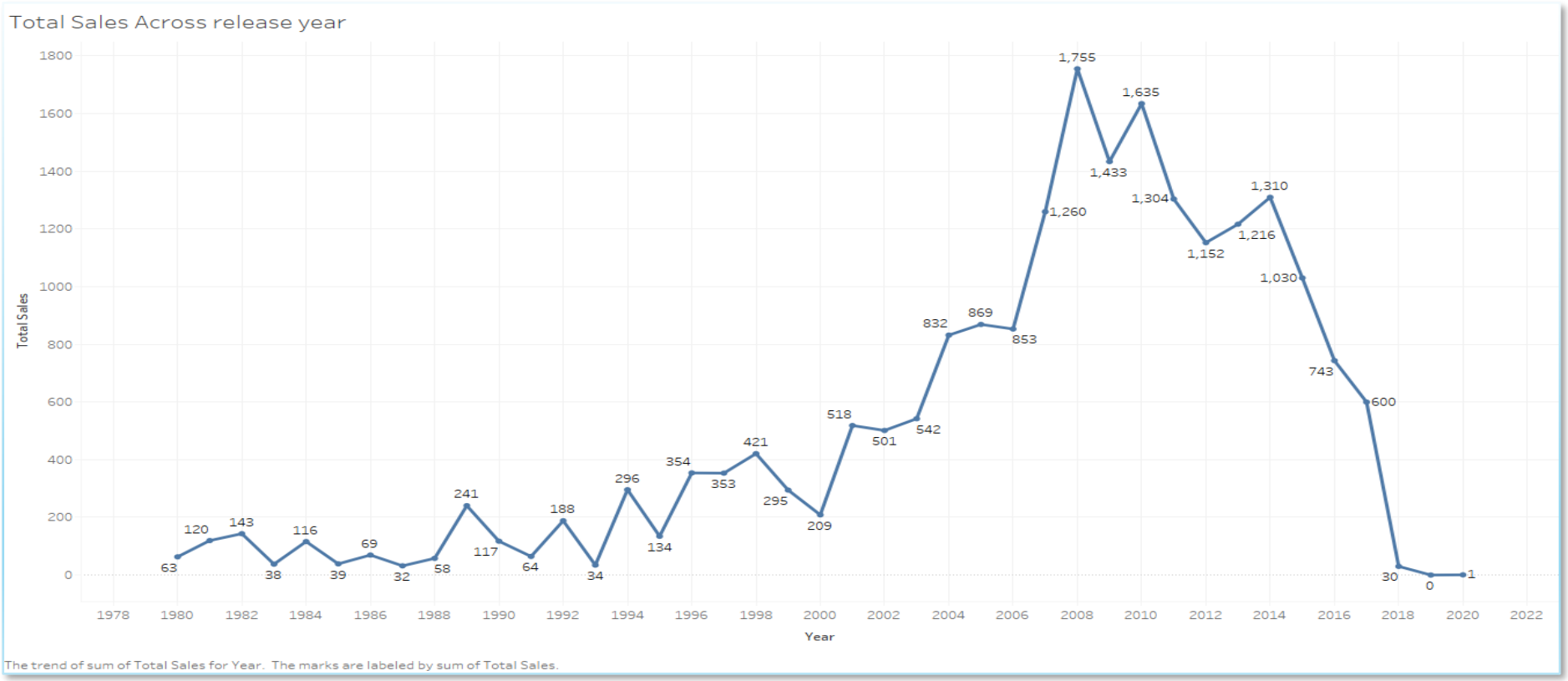
Publisher Vs Sales



4. Let us try to get idea about evolution of the video games.

Highlights:

- ⇨ During the period of 2007,2008,2009,2010, the sale is maximum.
- ⇨ We can during the period 2007 - 2010, user was interested in purchasing game and platform but after that trend changed to online gaming. Tough we have not that data but we can prove this fact.



3. Data Preparation & Feature Engineering

A. Missing Values Treatment / Data Imputation:

Highlights:

- ⇨ Missing values are present for Publisher, VGScore, CriticsScore.

Missing Data Statistics:

```
> missing_data_per
            Name             Year            Genre        Publisher NorthAmericaSales       EuropeSales
JapanSales
            0.00             0.00             0.00             0.33             0.00             0.00
      0.00
         RowSales       TotalSales          VGScore      CriticScore        UserScore         Platform
            0.00             0.00             1.00             0.58             0.95             0.00
```

There are 1474 unique publishers, 17 unique genres and 23623 unique games.

Feature Engineering -

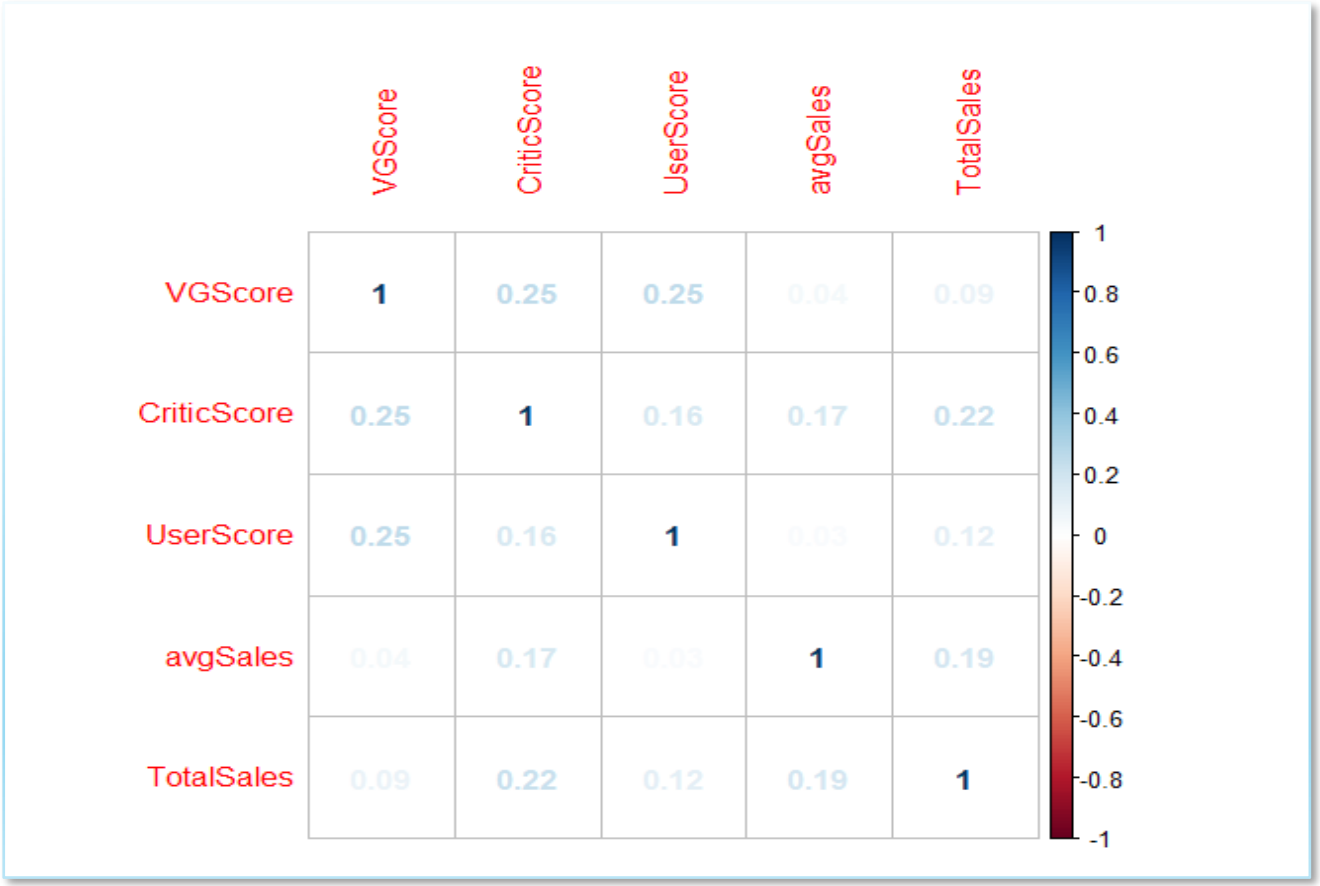
Publisher seems to be important while doing prediction, but it has been seen in our previous data exploration we might get new publisher after 3-4 years. So, model can fail if it gets new data which it hasn't seen. So, I will try to create features depends on Publisher.

To select a feature for model, we just need to check the importance of the features in prediction. We will do that before and after creation of the model. Right now, I am just thinking about the EA platform as we must build a model.

Correlation among features:

Highlights:

- ⇨ There is not very impressive correlation among features. But still we can do better with good feature engineering .



4. Modelling

A. Selecting Modelling Techniques

Data Splitting

Train: 80 % Test 20%

Modelling

we can try following models first.

- ⇨ Linear Regression
- ⇨ GLM
- ⇨ Random Forest
- ⇨ Support Vector Machine
- ⇨ Neural network

As mentioned earlier, the average units sold per game has evolved greatly since the 80s when the data set begins. This poses another difficulty to predictive modelling since models trained on the older data might generalize poorly to the newer data used for testing. Besides, the validation result might not be a good indication of the test result under these conditions.

To tackle this problem include attaching more weights to recent observations during training and choosing models that are robust to outliers such as support vector machine and random forest.

The three models evaluated here are **linear regression, support vector machine and random forest.**

Linear Regression Model Summary

By looking at p values we can get the significant predictors.

```
> summary(lm_model)
Call:
lm(formula = .outcome ~ ., data = dat, weights = wts)
Weighted Residuals:
       Min         1Q     Median         3Q        Max
-586998992   -2391856    -830473      -4790 2034990814
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             127.894214   2.850441  44.868  < 2e-16 ***
Year                     -0.063401   0.001412 -44.888  < 2e-16 ***
`GenreAction-Adventure`   0.077985   0.022762   3.426 0.000613 ***
GenreAdventure           -0.048343   0.010097  -4.788 1.69e-06 ***
GenreFighting             0.072761   0.012565   5.791 7.05e-09 ***
GenreMisc                -0.073281   0.007223 -10.146  < 2e-16 ***
GenreMMO                 -0.138025   0.070318  -1.963 0.049665 *
GenreMusic               -0.099110   0.024098  -4.113 3.91e-05 ***
GenreParty                0.388841   0.127789   3.043 0.002344 **
GenrePlatform            -0.016625   0.011244  -1.479 0.139268
GenrePuzzle               0.024185   0.014692   1.646 0.099740 .
GenreRacing              -0.015933   0.011570  -1.377 0.168484
GenreRolePlaying          0.017628   0.008141   2.165 0.030363 *
GenreShooter              0.077060   0.008839   8.718  < 2e-16 ***
GenreSimulation          -0.027253   0.011350  -2.401 0.016343 *
GenreSports               0.201941   0.009961  20.274  < 2e-16 ***
GenreStrategy            -0.125935   0.021950  -5.737 9.66e-09 ***
`GenreVisual Novel`      -0.091408   0.033886  -2.698 0.006987 **
avgSales                  0.288162   0.003872  74.431  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36980000 on 67356 degrees of freedom
Multiple R-squared:  0.1553, Adjusted R-squared:  0.1551
F-statistic: 688.1 on 18 and 67356 DF,  p-value: < 2.2e-16
```

Model selection

We can see the useful model from the following graph by looking at RMSE and RSquared values.

Modelling Issue:

To run the models like random forest it is requires high configuration, powerful machines. With only 4 GB Ram, models are taking so much time.

B. Build and Assess the model

Model	RMSE	Rsquared
Linear Regression	1.625143	0.15
Random Forest	Not available	
SVM		

1. Traditional Model Predictions:: TotalSales ~ Year+Genre+avgSales

Model	Year	Genre	avgSales	Publisher	Sales Prediction (in Million)
Linear Regression	2018	Role-Playing	0	EA	0.07605199
Random Forest	Models Are taking so much time				
SVM					

2. Custom Index Query Approach

Approach is like we take columns in query, here are genere and publisher and crete index and on top of that we can fire query like "Sales prediction for Role Playing' genre developed by 'EA' (refer Jupyter Notebook for details)

5. Executive Summary

1. **A. With Traditional ML approach**

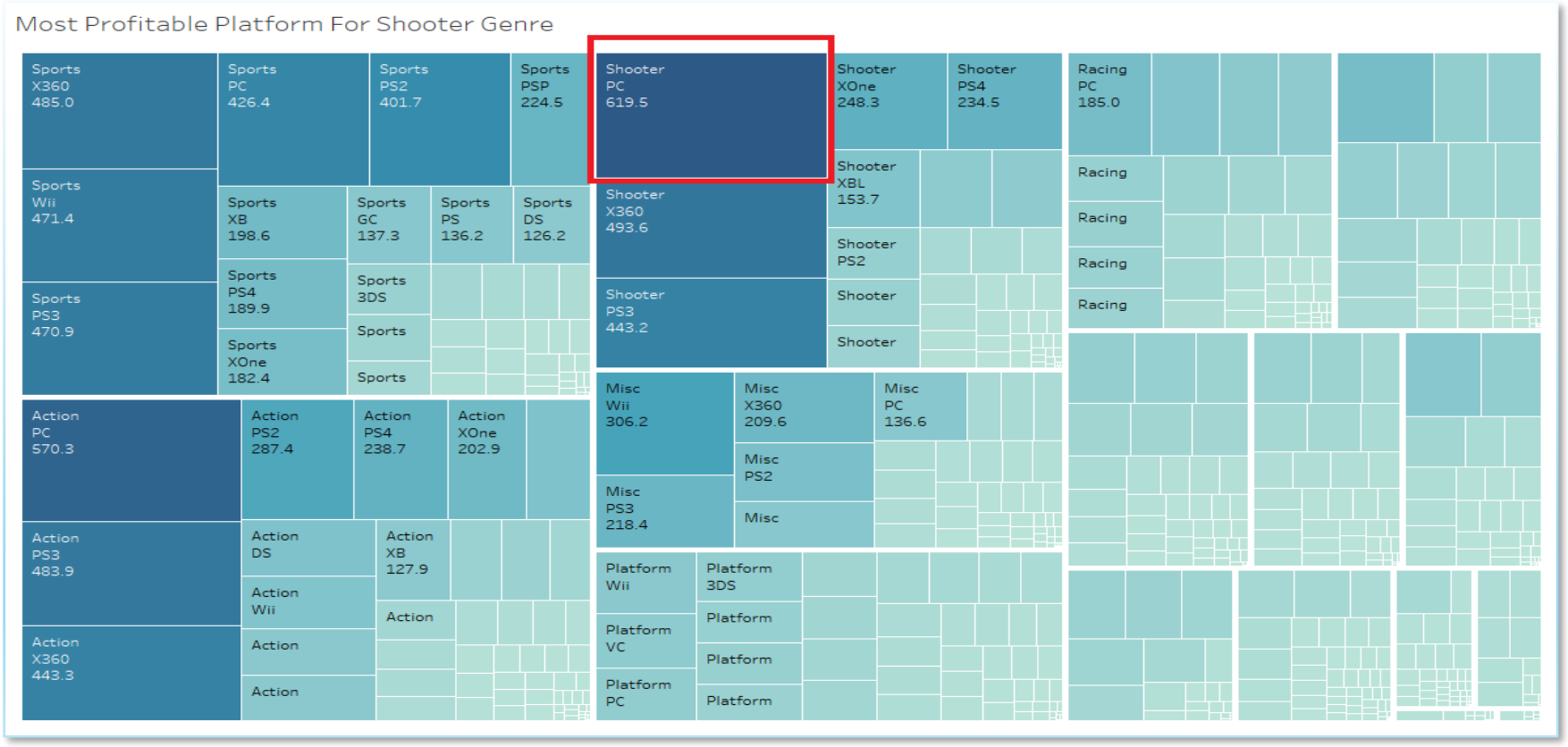
   For first query,Probable Sales for a “Role Playing’ genre developed by ‘EA’ in 2018 will be as follows –

Model	Year	Genre	avgSales	Publisher	Sales Prediction (in Million)
Linear Regression	2018	Role-Playing	0	EA	0.07605199
Random Forest	Models Are taking so much time				
SVM					

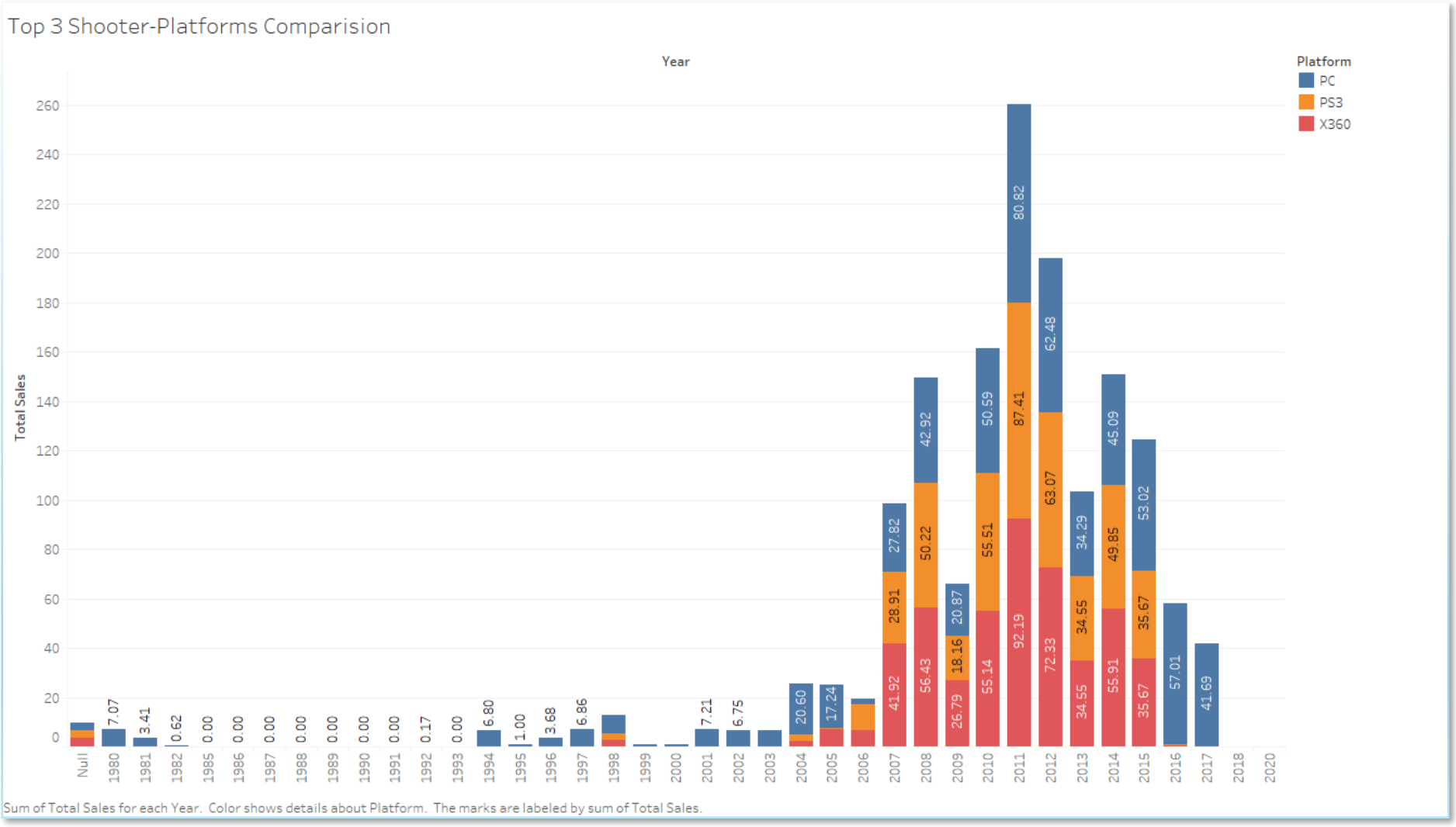
   B. Custom Index Query Approach

   Approach is like we take columns in query, here are genere and publisher and crete index and on top of that we can fire query like “Sales prediction for Role Playing’ genre developed by ‘EA’ (refer Jupyter Notebook for details)

2. To make the decision for profit or loss, it is required to have cost and revenue/sales with us. But as we are dealing with only sales values by following graph where we have genre plotted against platform in a heat map / mosaic chart, it is clear that PC is doing sale of more than 619 million dollars. So we can say probably that platform is making more profit compare to others.



Here I am doing **comparison between top 3 platforms for shooter genre**.
Most probably (not exactly because don't have exact data for cost and revenue) PC is most profitable platform for "Shooter" genre.


Top 3 Shooter-Platforms Comparision
Year
Platform
PC
PS3
X360
Total Sales
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Null
1980
1981
1982
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2020
7.07
3.41
0.62
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.17
0.00
6.80
1.00
3.68
6.86
7.21
6.75
20.60
17.24
27.82
28.91
41.92
42.92
50.22
56.43
20.87
18.16
26.79
50.59
55.51
55.14
80.82
87.41
92.19
62.48
63.07
72.33
34.29
34.55
34.55
45.09
49.85
55.91
53.02
35.67
35.67
57.01
41.69
Sum of Total Sales for each Year. Color shows details about Platform. The marks are labeled by sum of Total Sales.

6. Further Analysis

- ⇨ XGBOOST - I have selected only 3 models for regression. The model accuracy can be increased by using more advanced models like XGBoost.
- ⇨ Ensemble/Stacking of models can give better accuracy
- ⇨ DNN - We can also try deep neural networks
- ⇨ CV - I haven't used cross validation. By doing cross validation we can generalize model.
- ⇨ Feature Engineering - We can find more features which will help us to get more accurate results s