Project 2

Bjerke, Martin & Spremic, Mina

4/26/2021

Introduction

The aim of this project is to predict the sales of video games using random forest. The dataset is scraped video game sales until 2016, obtained from Kaggle. Additionally we wish apply some of the models from the XAI part of the course to explain which covariates were the most important in the prediction, since random forests is an ensemble type of method and not interpretable. We give some explanation of the dataset along with summary of how the dataset has been treated. We present some exploratory analysis of the dataset. First we divide the dataset into train and test set and apply the random forests, Finally, we discuss the intepretation and results and conclude the report.

Dataset and its preparation

Preparation of the dataset can be found in a separate .R file in the repository labelled as "data_cleaning.R". We give a summary of the way we proceeded to clean it and impute some of the missing data.

Firstly we removed the columns which are not entering our data analysis, which are the other types of sales that are not Global, and these are North American, Japenese, European and Other sales. Secondly we removed the rows for which both the critics score and the user score, as well as the counts are empty, tba or NA. Next step was to add the missing year of release, as there were not too many observations missing this covariate after the previous removals, and it is possible to obtain this information. There were between 150 and 200 rows missing the year of release, and these were added manually.

There have been some rows where either the critic score and count or the user score and count were either missing, to be determined (tbd) or simply NA. Here we have chosen to impute the values, based on a mean of other observations which have a similar score (+/- 5), either the user or critic score (depending on which we are imputing for).

For some of the other columns which were just missing a few entries (less than 10), we added these in to the existing Unknown category.

The covariate Name is also removed and not included in the statistical analysis.

Here we present the covariates as they are, before the dummy variable coding is conducted for the statistical (and exploratory data) analysis:

- Platform: Gaming platform, 17 different categories
- Year of Release: When the video game was releases, ranges from 1985 to 2016
- Genre: type of the video game, 12 different categories
- Publisher: which studio published the game, 310 different publishers

- Global Sales: how much game was sold, ranges from $0.01m to $82.53m

- Critic Score: evaluation of the critics, ranges from 13 to 98 (out of 100)
- Critic Count: number of critics that reviewed the game, ranges from 3 to 113
- User Score: evaluation of the game by user, ranges from 0 to 9.7 (out of 10)
- User Count: number of users that reviewed the game, ranges from 4 to 10665
- Developer: studio that developed the game, 1516 different developers
- Rating: ESBR rating of the game, 6 different categories

More details on the covariates are presented in the next section.

The covariates that are turned into dummy variables for the analysis later are Genre, Publisher, Delevoper and Platform. Rating is not made into a dummy variable as it has very few categories.

There are in total 8706 observations in this cleaned dataset. There might be some inconsistencies when it comes to some entries in the dataset. For example, we have observed that the publisher for one of the titles was incorrect. These have not been corrected and we have not actively looked for such inconsistencies.

Reference for why imputing and not surrogate split: http://arxiv.org/pdf/0811.1645.pdf

Exploratory analysis

First we present the summary of the dataset, before it is dummy coded.

```
##      Name              Platform         Year_of_Release    Genre
##  Length:8706        Length:8706        Min.   :1985    Length:8706
##  Class :character   Class :character   1st Qu.:2004    Class :character
##  Mode  :character   Mode  :character   Median :2007    Mode  :character
##                                        Mean   :2007
##                                        3rd Qu.:2010
##                                        Max.   :2016
##   Publisher          Global_Sales      Critic_Score    Critic_Count
##  Length:8706        Min.   : 0.0100   Min.   :13.00   Min.   :  3.00
##  Class :character   1st Qu.: 0.0900   1st Qu.:61.00   1st Qu.: 12.00
##  Mode  :character   Median : 0.2400   Median :70.43   Median : 23.00
##                     Mean   : 0.6703   Mean   :69.01   Mean   : 26.51
##                     3rd Qu.: 0.6300   3rd Qu.:79.00   3rd Qu.: 35.00
##                     Max.   :82.5300   Max.   :98.00   Max.   :113.00
##    User_Score      User_Count        Developer           Rating
##  Min.   :0.000   Min.   :    4.00   Length:8706        Length:8706
##  1st Qu.:6.600   1st Qu.:   11.00   Class :character   Class :character
##  Median :7.200   Median :   31.34   Mode  :character   Mode  :character
##  Mean   :7.118   Mean   :  150.02
##  3rd Qu.:8.100   3rd Qu.:   77.00
##  Max.   :9.700   Max.   :10665.00
```

We observe that mean and the median of the sales are quite far apart, meaning that there most probably are a few games that have sold considerably more than others, and that it is more common for games to not sell that well, which makes sense. Additionally we can see that there are more users reviewing games than critics, both looking at mean and median, which also makes sense.

Looking at the first correlation plot in Figure 1, we observe that user score and critic score have the highest correlation among the covariates. This is logical as normally, a good critic score and review indicates the quality of the game, and it is probable that the users are also going to review the game favourably. User count is also negatively correlated with year of release, which could mean that either quality of the games have decreased, or that the number of games made has increased, and that more of them are of poorer quality. Concerning Critic Count and User Count, it is no susprise that these covariates are positively correlated with

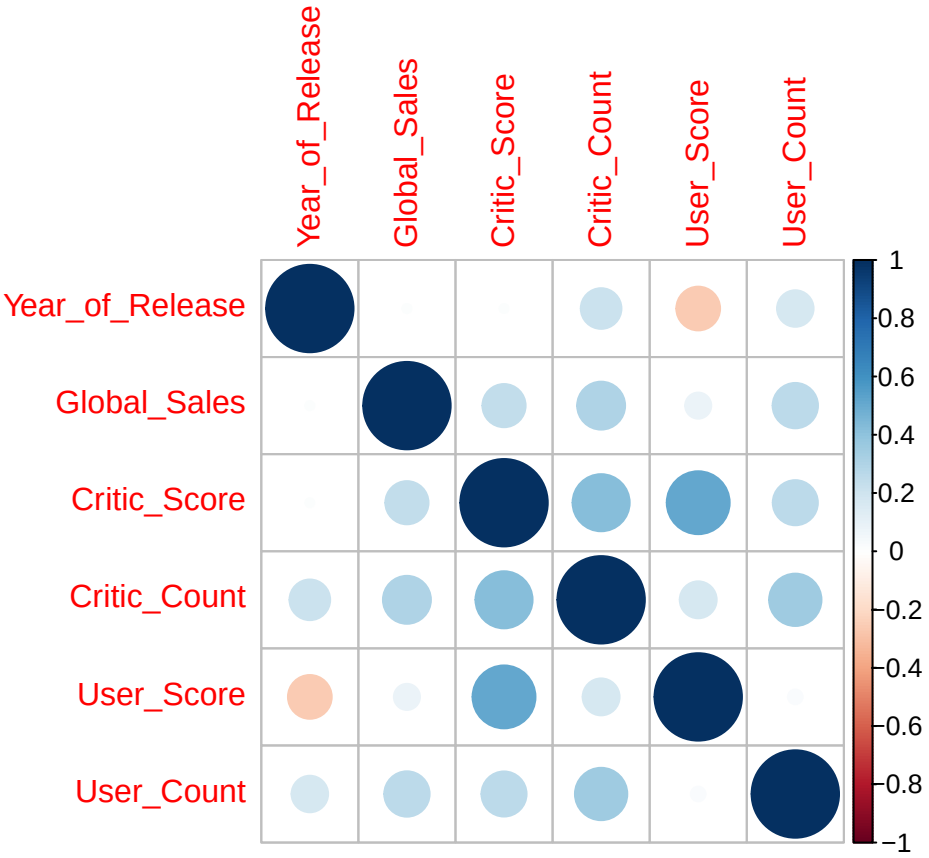

Figure 1: Correlation plot between the numerical variables.

Year of Release due to improvement in technology and the reviewing culture, meaning it has become much easier and more accessable to review games (and other things) in the recent years. Critic score and critic count are positively correlated as well, as expected, the games with higher score are usually very popular meaning that a lot of critics (and users) review them. Finally, again, unsurprisingly enough positively correlated with user count, critic count and critic score. The more people review the game, the more have bought it and the more positive reviews are, the more people are going to buy the game.

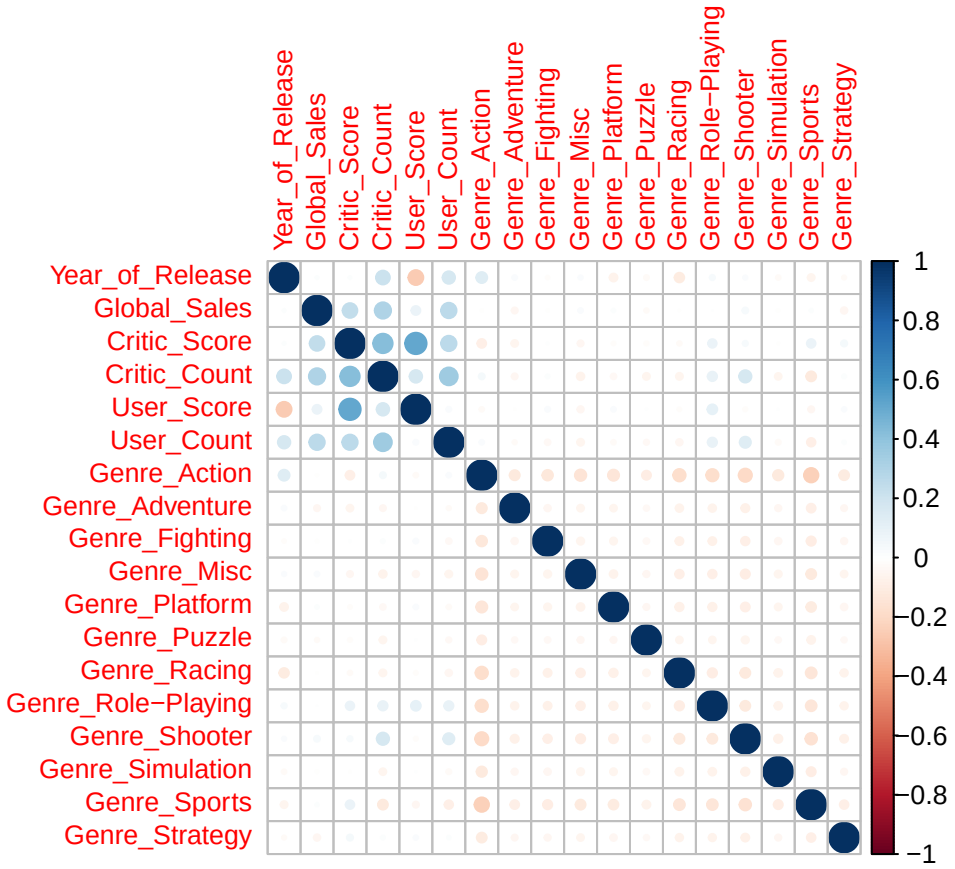

Figure 2: Correlation plot for genre that has been dummy coded and the other numerical variables.

Now we have a look at a correlation plot in Figure 2, with added dummy variables for game genres. Here we comment some of the relations that can be observed from the plot. We observe there is positive correlation between action genre and year of release and negative between platform, racing, sports and year of release. This most probably means that there have been released more games of the action genre and less of the other three. We also see that more critics and users review shooter and roleplaying games, as the correlation is slightly positive, and they review less of the other genre.

Looking at the final correlation plot in Figure 3, we will comment on some of the interesting relationships shown. We see various correlations between the platforms and year of release, and this of course is related to the period in time each of the platforms has been in use. We see that X360 has positive correlation with the critics count, meaning critics reviwed a lot of games made for this platform. We observe the same for the relationship between User Count and PC platform. We see also some positive correlations between certain genres and platforms, meaning there have been developed more games of those specific types for the platforms in questions.

Due to their rather large number, the different developers and publishers are not included in any of the correlation plots (it was also not possible to visualize it or see anything, graphically).

In Figure 4 we see the number of video games released each year. The number starts out low in the begining,

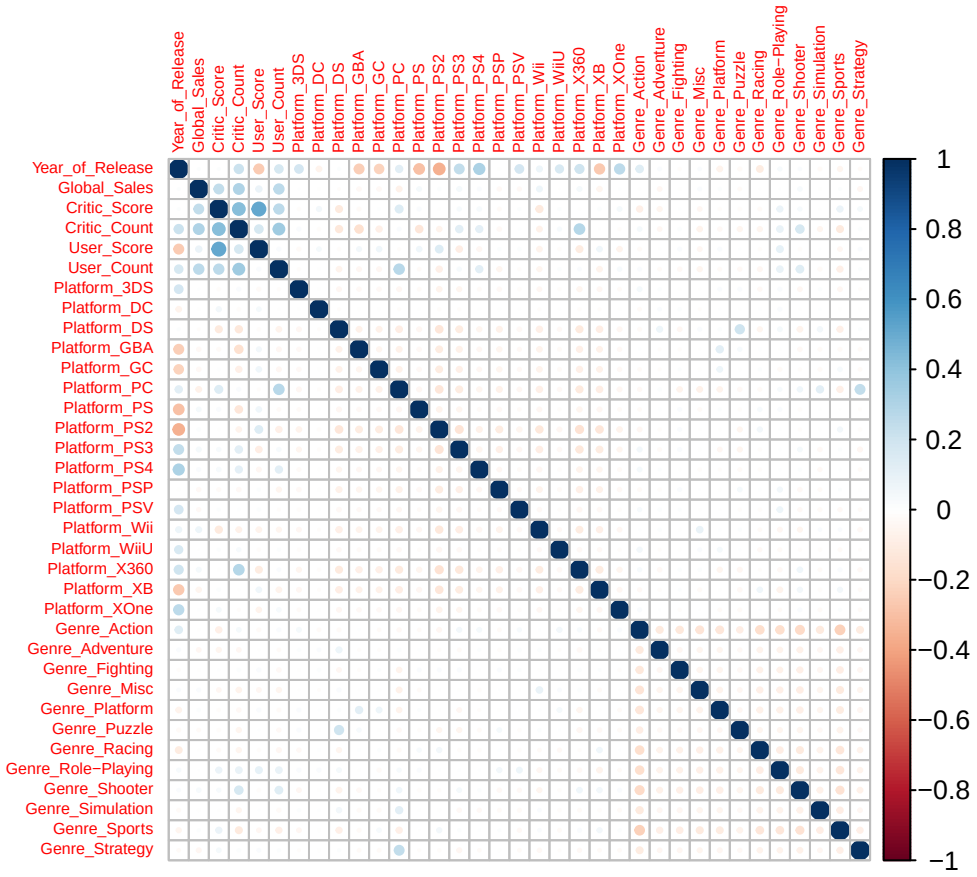

Figure 3: Correlation plot for genre and platform types that have been dummy coded and other numerical variables.

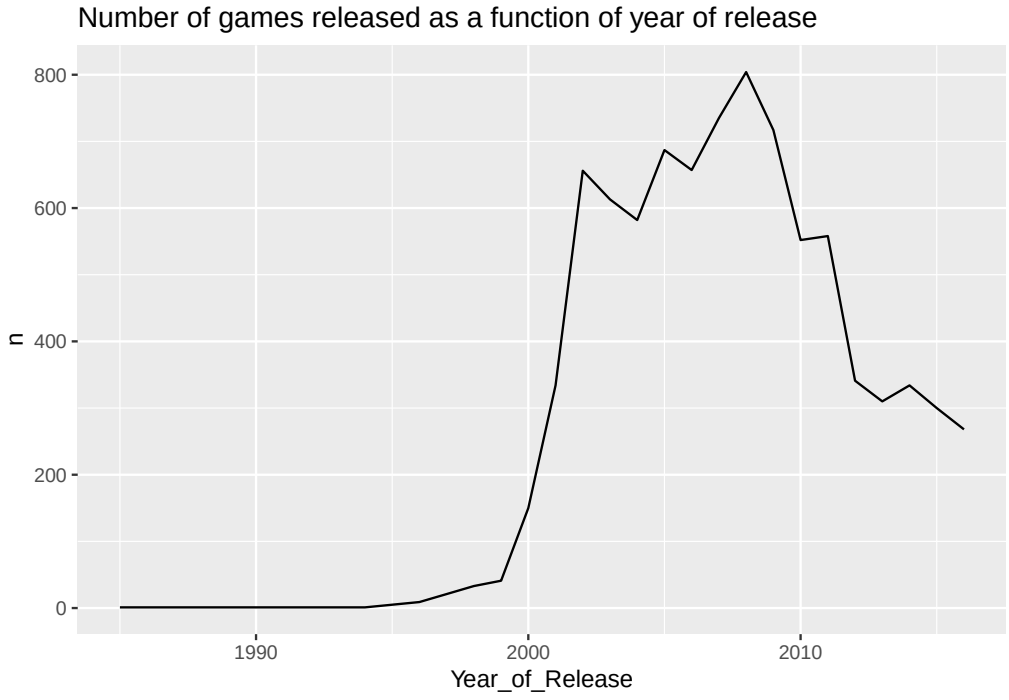

Figure 4: Number of games released each of the given years.

and we suspect this is due to the low number of video games registered in the dataset from this period. The number proceeds to increase until it peaks around 2008, and has been decreasing since. The decrease can perhaps be explained with there being fewer games that are larger, meaning more effort is put in developing them, with demanding graphics with large amount of possible play hourse.

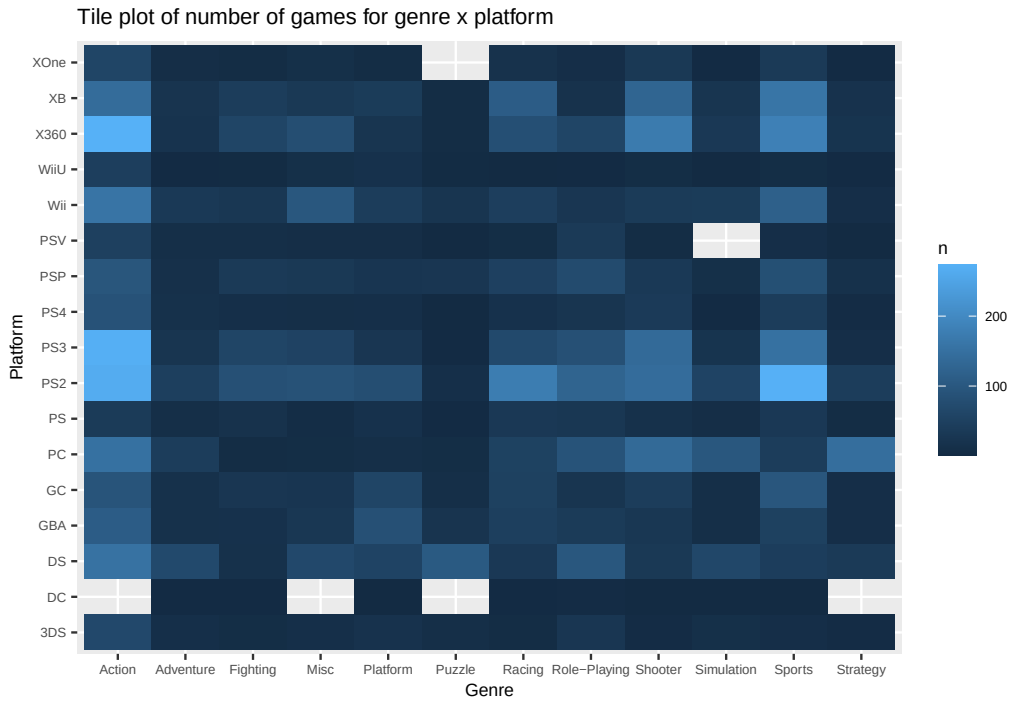

Figure 5: The tile plot where showing number of games for each combination of platform and genre.

Figure 5 shows the number of games released for each combination of genre and platform. We observe that the Action and Sports are genres with a lot of games for PS2 (Playstation 2) and Action genre has a lot of games released for X360 and PS3 as well. Usually the same games are released on multiple platform so this is not surprising. We also observe that some genres are not represented on certain platforms. We see that most games released for PC platform are from genre like Strategy, Shooter and Action.

In Figure 6 we see the total global sales shown per each genre. Wee see the most selling genres are Action, Sports and Shooter, which is not very surprising as there are the most games released in these genres.

```
## Selecting by sum_gs
```

Figure 8 shows top ten developers with highest number of total global sales. We see here that Nintendo tops this list, followed by EA Sports and Rockstar North.

```
## Selecting by sum_gs
```

Figure 8 shows top ten publishers with highest total global sales. Here we see again EA on the top of the list, along with Nintendo. These two companies both develop and publish games. Following up are publishers like Activision (most famous for publishing Call of Duty) and Sony Computer Entertainment (published a lot of succesful titles for Playstation).

Method

Random forest

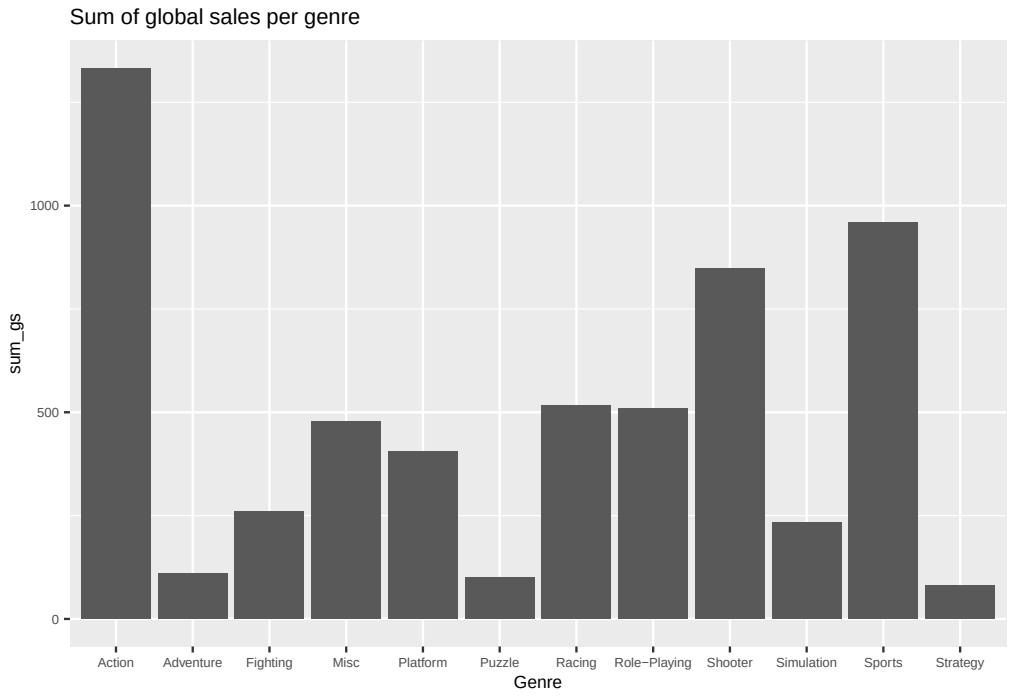

Figure 6: Sum of global sales per video game genre.

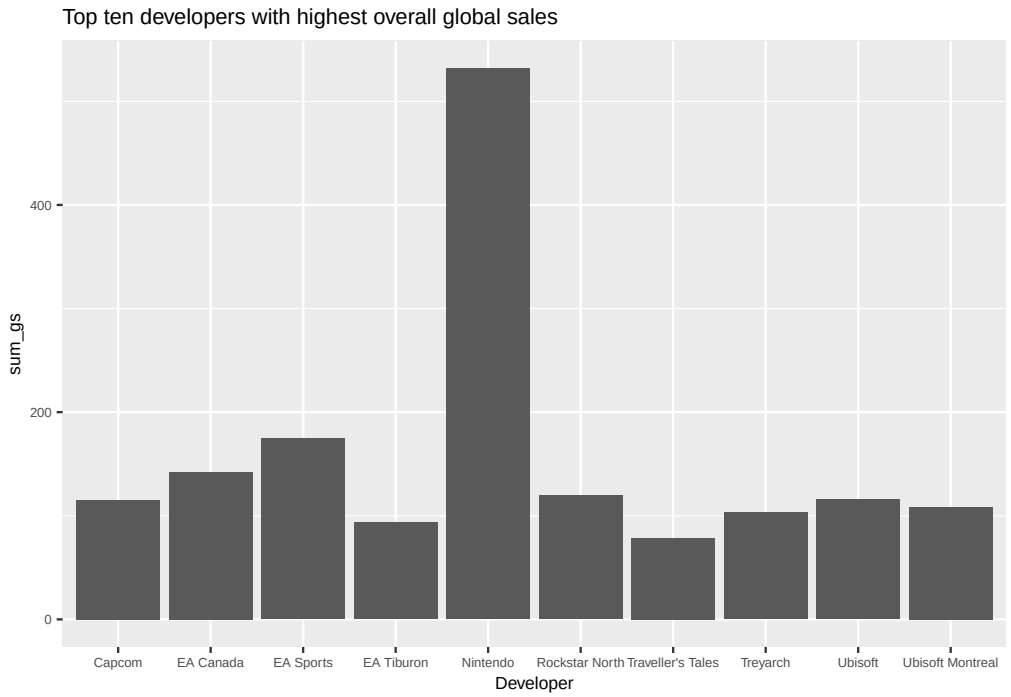

Figure 7: Sum of global sales for each developer, here top ten developers with highest global sales shown.

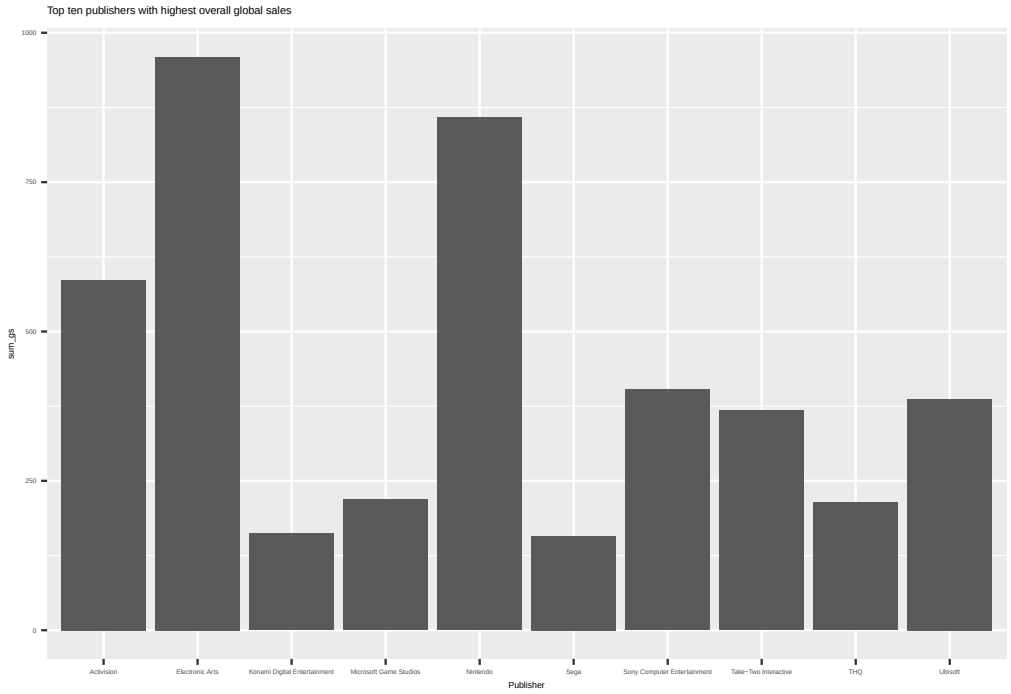

Figure 8: Global sales per publisher, here top ten publisher with highest global sales shown.

```
ds_vgs <- as.data.frame(read.csv("vgs_22_2016_cleaned_dummied_final_2.csv"))

ds_vgs <- subset(ds_vgs, select = -c(X))
#set.seed(1337, kind = "Mersenne-Twister", normal.kind = "Inversion")
set.seed(1337)
data_set_size <- floor(nrow(ds_vgs))

indexes <- sample(1:data_set_size, size=data_set_size*0.8)

training_set <- ds_vgs[indexes,]
test_set <- ds_vgs[-indexes,]

print(test_set$Name[1:5])
```

```
## [1] "Wii Sports"
## [2] "Wii Sports Resort"
## [3] "Kinect Adventures!"
## [4] "Brain Age: Train Your Brain in Minutes a Day"
## [5] "Call of Duty: Black Ops II"
```

Results

Prediction

Explainability

Conclusion

References

[1]: “Video Games Sales Dataset” - https://www.kaggle.com/sidtwr/videogames-sales-dataset
[2]: corrplot package - https://cran.r-project.org/web/packages/corrplot/corrplot.pdf
[3]: tidyverse package - https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf
[4]: fastDummies package - https://cran.r-project.org/web/packages/fastDummies/fastDummies.pdf
[5]: randomForest package - https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
[6]: iml - https://cran.r-project.org/web/packages/iml/iml.pdf
[7]: ranger - https://cran.r-project.org/web/packages/ranger/ranger.pdf
[8]: lime - https://cran.r-project.org/web/packages/lime/lime.pdf
[9]: shapr - https://cran.r-project.org/web/packages/shapr/shapr.pdf
In addition, all the dependencies were loaded as well for the packages mentioned above to work. These are not listed in detail.