UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

DAKD Final Course Essay

CRISP-DM
Steam Store Games

Anass Benali

January 2020

Contents

1 Introduction

This project was performed for the Data Analysis and Knowledge Discovery subject of the Master in Innovation and Research in Informatics at UPC-FIB. The purpose of this project is to apply the CRISP-DM methodology to analyze the Steam Store Games. While it was not compulsory for the project, it analyses a topic using fairly recent data which has not yet been fully explored, analyzed and modeled.
When looking for other analysis for the Steam platform, many websites providing basic information for the Steam Store show up. However, the provided information is mainly compromised by descriptive data, like the number of concurrent players [1], the top played games at the moment and similar [2] [3]. Hence, it seems like there are almost no public data analysis projects from the point of view of data mining for the Steam Store [4]. In fact, no other comprehensive analysis was found apart from the one performed by the owner of the dataset that will be used in this project [5].

2 Problem understanding

2.1 Background

Steam is a video game digital distribution service by Valve [6]. The initial release was at September 2003 [7]. Currently Steam is the leader of digital video game distribution being the the largest and most used platform with lots of years of leadership [8]. Gaming as for today it already transitioned from the classic physical distribution to digital. As such the PC gaming world is a growing market for platforms and stores like Steam. Part of that success is given by the amount of complementary services that Steam offers for their platform (such as achievements or cloud saves, etc) and the fact that they host frequently many seasonal big sales [9].

2.2 Goals

If one were to decide to enter the gaming market as a developer or publisher, being in the Steam platform would be a decision to make. To make that decision it would be great to have information and a understanding of how the games perform in the platform. Also we would like to know how the overall store looks like and see around which price margin games are more successful. Moreover, we would like to model the price of the games such that given some features we are able to predict at which price margin corresponds.

2.3 Success criteria

We will consider we succeeded in the project if we are able to find or extract patterns or characteristics about the store and which kind of games perform better. The idea is to give insights for the marketing decisions of the game, such as which price it should have, which genres are more common, etc.

2.4 Tools

The project will be developed in Python language using a Jupyter Notebook. However, if some tasks are troublesome and difficult to perform in Python then R language will be used instead. All the source code will be provided attached with the report (also found on the GitHub repository 'DAKD Steam Store Project').

3 Data understanding

3.1 Data Selection

The dataset that will be used combines information pulled from the APIs of both Steam and SteamSpy amounting to 27075 games. The data was collected by Nik Davis around May 2019 and posted on Kaggle as 'Steam Store Games' [10].

The considered columns are:

- **appid**: Unique identifier for each title.
- **release_date**: Release date in format YYYY-MM-DD.
- **english**: Language support: 1 if English is supported, 0 elsewhere.
- **developer**: Name (or names) of developer(s). Semicolon delimited if multiple.
- **publisher**: Name (or names) of publisher(s). Semicolon delimited if multiple.
- **platforms**: Semicolon delimited list of supported platforms.
- **required_age**: Minimum required age according to PEGI UK standards. Many with 0 are unrated or unsupplied.
- **categories**: Semicolon delimited list of game categories.
- **genres**: Semicolon delimited list of game genres.
- **steamspy_tags**: Semicolon delimited list of top SteamSpy game tags, similar to genres but community voted.
- **achievements**: Number of in-games achievements.
- **positive_ratings**: Number of positive ratings, from SteamSpy.
- **negative_ratings**: Number of negative ratings, from SteamSpy.
- **average_playtime**: Average user playtime, from SteamSpy.
- **median_playtime**: Median user playtime, from SteamSpy.
- **owners**: Estimated number of owners. Contains lower and upper bound.
- **price**: Current full price of title in GBP (pounds sterling).

3.1.1 appid

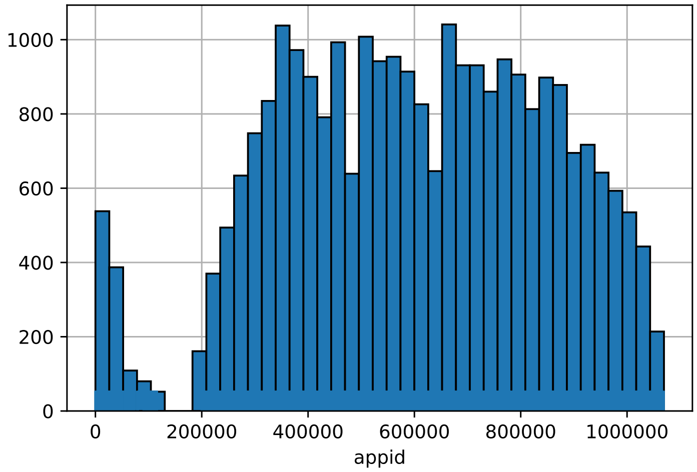

Figure 1: Distribution of appid

There is a gap around 200000 where there is no appid. Which either means the range was reserved for future use or some change of system was made in that time.

3.1.2 english

English	98.11%
Other	1.89%

Table 1: Distribution of language

Since we will focus in a english language market we will remove the non-english games.

3.1.3 PEGI rating

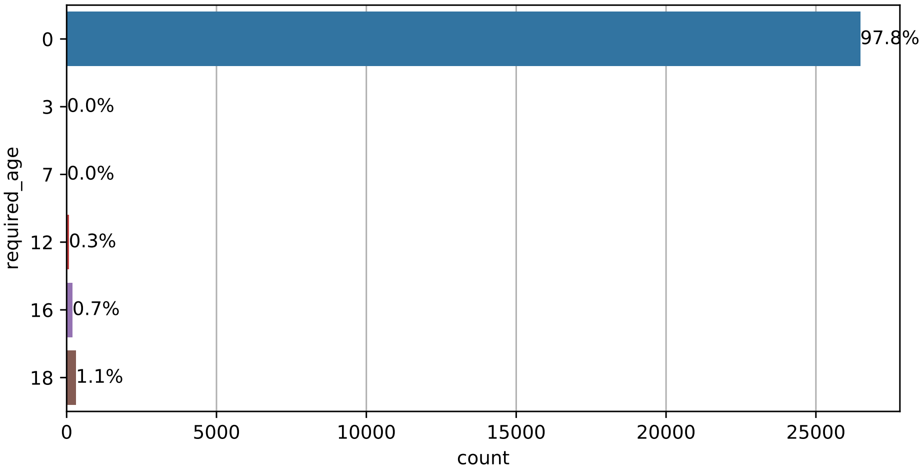

Figure 2: Distribution of PEGI ratings

It is unlikely that the real PEGI rating distribution is like in figure 2, therefore we conclude that most of the games are just unlabeled. Still, the few that are labeled may still have some use.

3.1.4 platforms

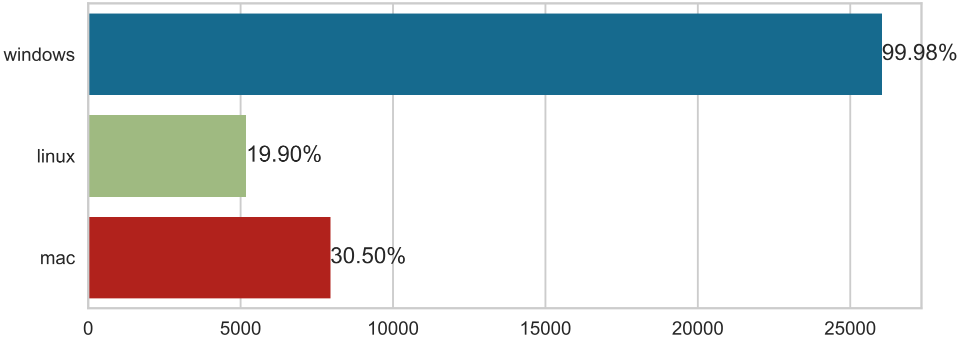

Figure 3: Distribution of platforms

The distribution is within expected. We can see that most almost all the games support windows and that a small portion of them support the other platforms.

3.1.5 price

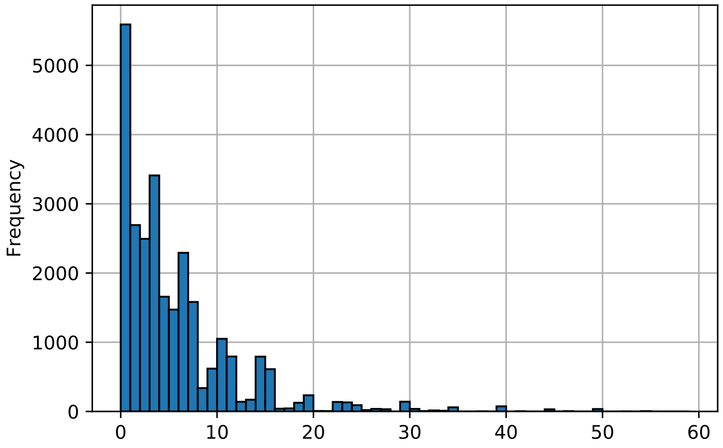

Figure 4: Distribution of price

We know that the videogames usually cost at most $60. There are a few titles (30) which cost more than that.

3.1.6 genres, categories and tags

The columns ’categories’ and ’steamspy_tags’ contain overlapping information with the genres. Hence, only the ’genres’ column will be kept since it is the most complete.

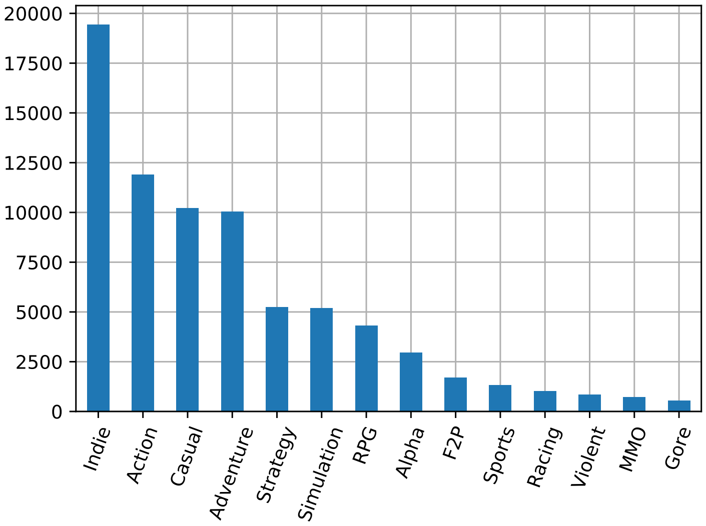

Figure 5: Distribution of genres

It was surprising to find that most of the games are Indie and Action. This pattern along with how the other attributes are distributed, suggests that in fact, the store contains a large amount games from small developers.

3.1.7 owners

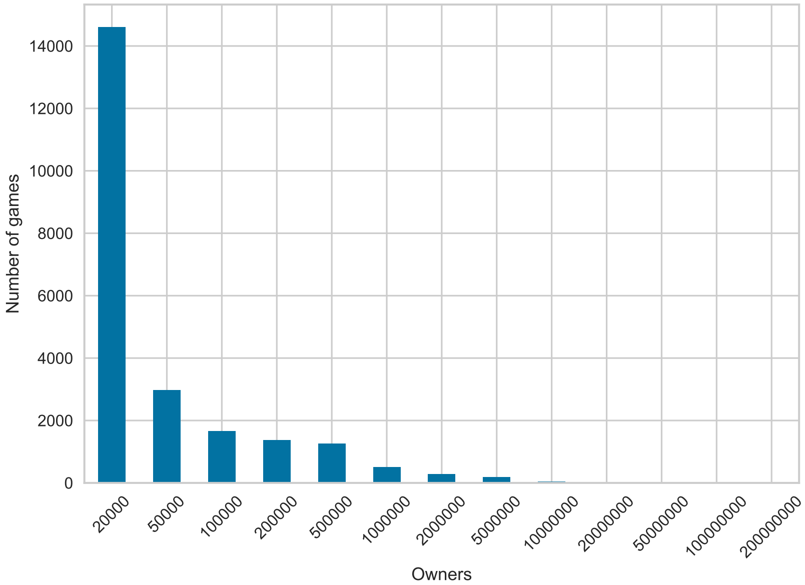

Figure 6: Distribution of owners

Most of the games of the store have between 0 and 20000 owners, which basically mean, that there is a lot of low quality games or niche indie games. The ownership follows a decreasing exponential distribution.

3.1.8 achievements

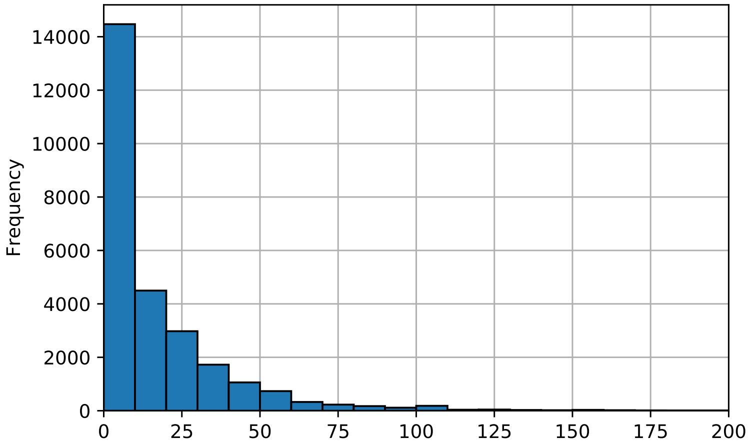

Figure 7: Distribution of achievements

Most of the games have little to none achievements. Again, we see a exponentially decreasing distribution.

3.1.9 playtime

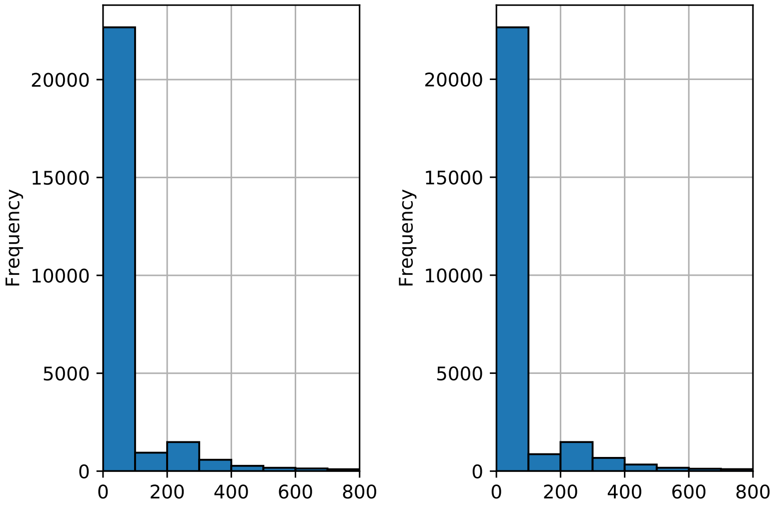

Figure 8: Distribution of average and median playtime

The average and median playtime look exactly the same, and as will be seen after, they have very high correlation, which essentially means that, for the most part, are interchangeable.

4 Data preparation

The data is mostly in good state. The missing values and outliers are taken care of by the owner of the dataset. However, there is still some preprocessing to be done. The objective of this part is to specify the transformations and modifications done to the data. The decisions taken here are influenced by both what has been seen in the descriptive analysis and what will be seen in the modelling. The decisions taken here at the preprocessing will have an impact on the conclusions and analysis.

The following preprocessing steps were performed:

- Create a year column from release date since the full date is not useful as it is.
- Transform to one-hot encoding the multi-valued columns: platforms and genres.
- Remove the duplicate columns 'steamspy_tags' and 'categories' because they contain duplicate information already found in the 'genres' column.
- Only keep the genres that appear at least in 1% of the games.
- Compute the score of the games using the positive and negative ratings.
- Keep the average playtimes and remove the median playtimes.
- Keep only the upper bound for the estimated owners.
- Create four categories for price: free, cheap, mid-range and expensive.
- Keep English games only (target market) and remove the attribute English.
- Filter out all the games with owners between 0 and 20000.
- Filter out all the games that do not have at least one positive and negative rating.
- Remove the windows platform column since all the games support it apart from some outliers.

After the preprocessing 8329 rows were left.

5 Modelling

5.1 Exploratory analysis

We will start with some exploratory analysis and visualization of the variables.

5.1.1 Correlation heatmap

To will begin with, we will be focusing in the correlation of the numerical variables. The correlation between the variables will give more insight on the useful variables for classification and prediction. Moreover, we may be able to either discover or confirm patterns or conclusions.

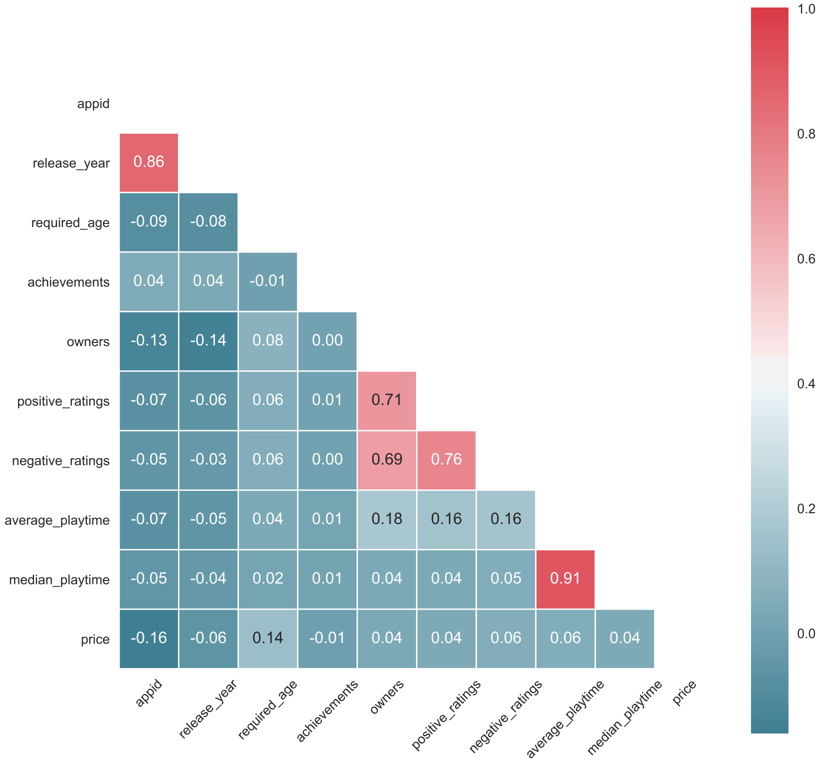

Figure 9: Heatmap of attributes

As expected, the positive and negative ratings are positively correlated, which essentially means that popular games have more reviews. Moreover, following this reasoning we can also see that the ratings are also positively correlated with the owners.

Also, average and median play times are highly correlated given that they basically are similar measures. Additionally we observe that the appid and year of release, are strongly correlated which makes sense since the appid is an incremental identifier.

The other correlations are not as strong, however, we can see interesting patterns. There are more affordable games as the years increases. Games with more owners and ratings have more average time played. Owners decrease as year increases which may be a result of older games having more time to be bought and discovered by users. The higher PEGI rated games are to some degree more expensive. We must remember that correlation does not imply causality, however, those are the patters found in the data. A similar analysis can be performed analyzing the PCA Biplots that can be found attached and on the R script.

5.1.2 Price analysis

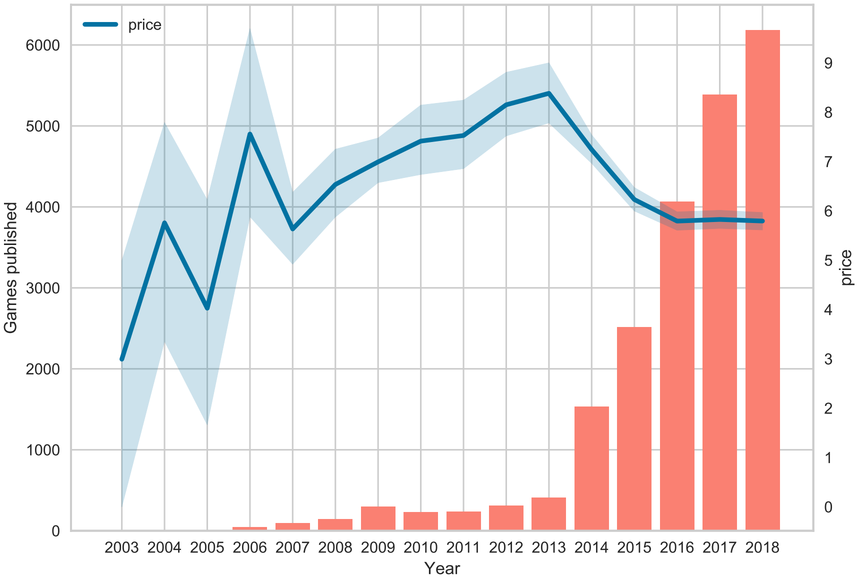

Figure 10: Price and published games by year

The price of the games was rather low and with high variance at the early days of steam platform. As the store gained more popularity and the market grew the games were increasingly more expensive, at least, until 2013. However, from that point, as the number of the number of published games increased, the average price per game per year dropped. Most likely there was an influx of small indie developers offering cheaper games.

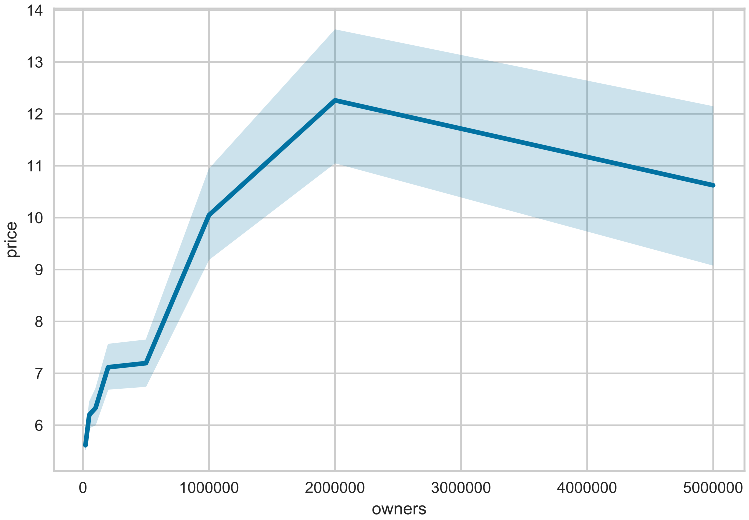

Figure 11: Price by ownership

We should remember that we took the upper bound of the ranges of owners. We can see that the most profitable margin is from 2 million owners and above. The variability in price is low for niche games in comparison to more owned games.

Taking into account only the price sale of the game, the most profitable price point to have is around £10. The plot suggests that, for small and mid-sized publishers, one should target the game to deserve the higher price possible up to around the £10 range. Obviously, the plot does not illustrate other income sources, as for example, the money earned thought micro-transactions of the marketing model of free-to-play games.

5.2 Clustering of genres

We will perform hierarchical clustering. Given that the we encoded the genres using one-hot encoding we used the cosine distance to obtain the following dendrogram.

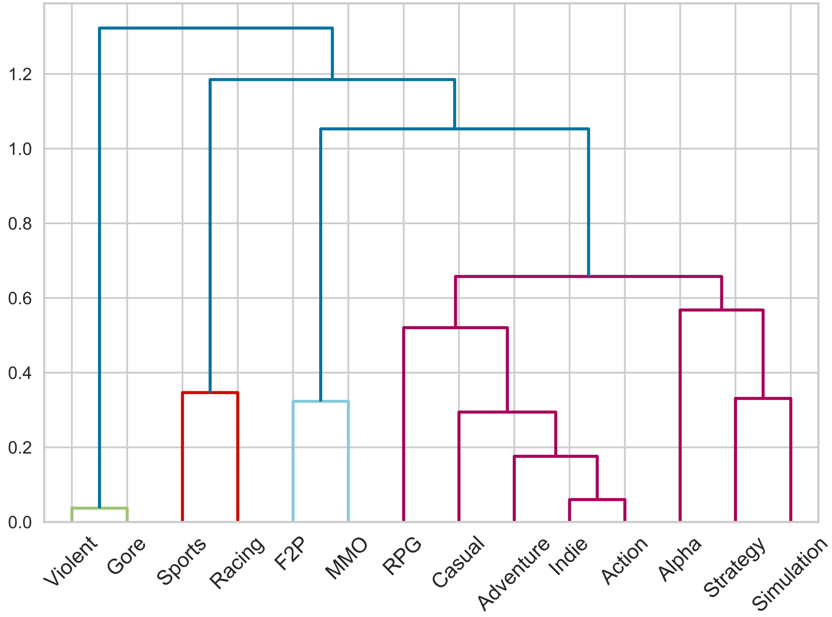

Figure 12: Dendrogram of genres

The dendrogram illustrates intuitively the closeness of the genres. We in fact can see different clusters: Violent and Gore, Sports and Racing, F2P and MMO, Strategy and Simulation and then we have Alpha (Early Access) and finally the others ones.

5.3 Classification of price range

For the prediction of the price range we will take into account only those games which have an ownership above 20000 users. The reason is because games with few owners or without reviews do not make many impact, either because they did not reach any users or they are not serious game. With the data available they will all look roughly the same adding only noise.

We will model the pricing in four categories. We will categorize prince in Free, Cheap, Mid-range and Expensive. Free-to-play games as the games with price equal to zero. The other categories will be the discretized in three equal sized slices. The percentiles will be used (cheap as the $\frac{1}{3}$ percentile, mid-range as between $\frac{1}{3}$ and $\frac{2}{3}$ percentile, the rest will be expensive).

The dataset will be split into 80% training and 20% testing, uniformly at random without replacement.

The following methods were tested: K-Means, SVM, AdaBoost (with decision trees) and Random Forest. Among the methods, the one which worked better was the Random Forest. An accuracy of about 69% was obtained. This is a considerable improvement compared to the baseline of a random predictor which would have an accuracy of about $\frac{1}{4}$.

	precision	recall	f1-score	support
Cheap	0.69	0.83	0.75	268
Expensive	0.63	0.70	0.66	183
Free	0.98	0.86	0.91	137
Midrange	0.48	0.28	0.35	144
weighted avg	0.69	0.70	0.68	732

Table 2: Classification report of Random Forest (validation test)

We can see that the Free category is the most easy to classify most likely with the genres of the games. For the other categories, it struggles more but the method was able to take advantage of the correlations a good extent.

5.4 Prediction of price

For the prediction of price as a numerical variable, many regression methods were performed. The tested methods are Logistic regression, SVM, AdaBoost with Decision Trees, Random Forest and a simple Neural Network. Like before, Random Forest outperformed the other methods obtaining a R^2 of 0.61 and $RMSE$ of 4.96.

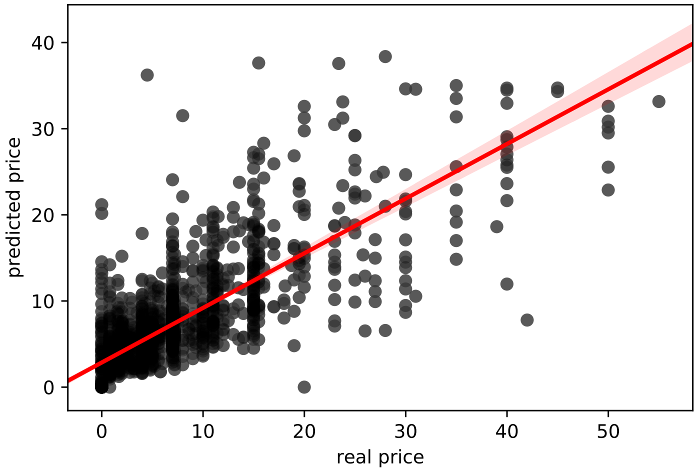

Figure 13: Scatter plot of predicted versus real price

We can observe from the plot the predicted prices versus the real ones. The plot contains a line where the predictions should be to be totally accurate. We can see that specially for the price range of 0 to 10 it works decently, however, it is far from perfect.

6 Evaluation

I believe that the goals set at the beginning of the project were meet with success. We were able to do an descriptive analysis and exploratory analysis for some characteristics of the games of the store. With that we were able to get insights on the platform as a whole.

Moreover, we were able to find clues about the pricing of a game through the visualizations, and we were able to model it using both classification and regression to a decent degree.

Also, we identified which combinations of genres are more popular and which combinations are most usual in the store with hierarchical clustering. That information could prove useful to decide which kind of game to develop so it fits the the store. It could even be used to try to position a game in an new niche of different combination of genres.

7 Conclusions

We got insights about how the titles on the steam platform are distributed thorough exploratory analysis. We found that Indie video games are the most common genre of game and that there is a lot of niche, cheap and with low ownership titles. We saw that from the game sales price revenue standpoint, around £10 is ideal. Moreover, we classified the genres using hierarchical clustering to visualize which combinations of genres are more common on the store.

We modeled a classifier and predictor for the price. With the information available we achieved around 70% of precision for the classification of the price range and a R^2 of 0.61 and $RMSE$ of 4.96 for the numerical predictor. The models could be used to have an insight of which price point a game should be in given his attributes and to predict how a title yet to be released could perform.

References

[1] *Steam stats*. URL: https://store.steampowered.com/stats.

[2] *Steam Charts*. URL: https://steamcharts.com.

[3] *SteamDB*. URL: https://steamdb.info/graph.

[4] *Steam Lack Data Scientists*. URL: https://www.gamasutra.com/blogs/BurakTezateser/20190930/350922/Steam_has_a_lack_of_data_scientists.php.

[5] *Steam data exploration*. URL: https://nik-davis.github.io/posts/2019/steam-data-exploration.

[6] *Steam Store*. URL: https://store.steampowered.com.

[7] *Wikipedia entry for Steam*. URL: https://en.wikipedia.org/wiki/Steam_(service).

[8] *Steam largest digital platform for PC gaming*. URL: https://www.pcgamer.com/market-data-firm-claims-valve-made-730-million-last-year.

[9] *Steam features and services*. URL: https://partner.steamgames.com/doc/features.

[10] *Steam store games*. URL: https://www.kaggle.com/nikdavis/steam-store-games.