Data analysis result

Brent Schoenmakers

Inhoud

Introduction

The reason I am writing this report is because I was tasked by my client, a stakeholder investor in the video gaming market, to figure out what the most profitable market would be in the future. My client made a mistake in their calculation in the past, and they didn't want to repeat it. This time, they are trying out a data scientist to figure out what the most profitable market will be in the next few years.

The questions I thought were valuable in the prediction of profitable market were:

- What genres are providing the most amount of revenue?
- What platforms have released a lot of games?
- Are the global sales interesting to use for predicting other markets?

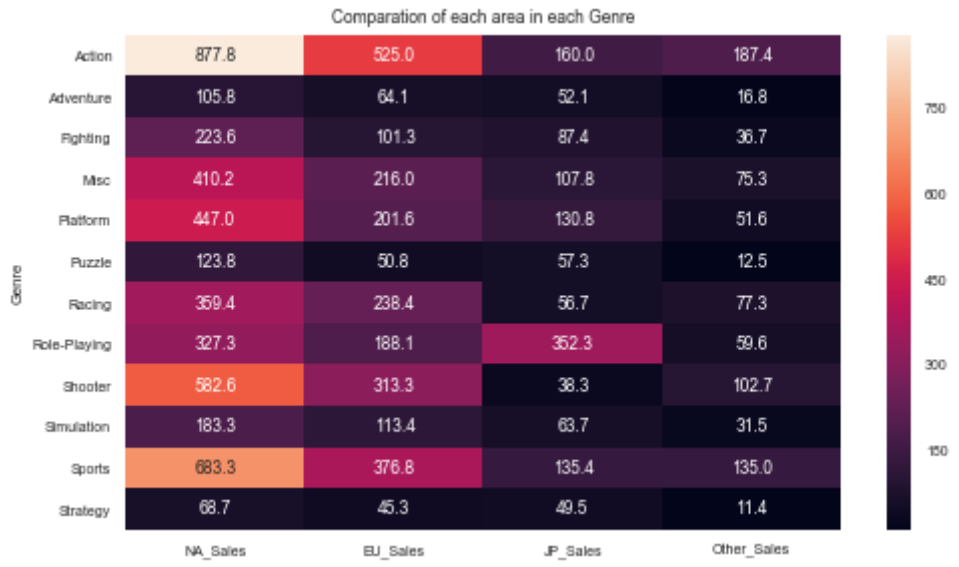
Data

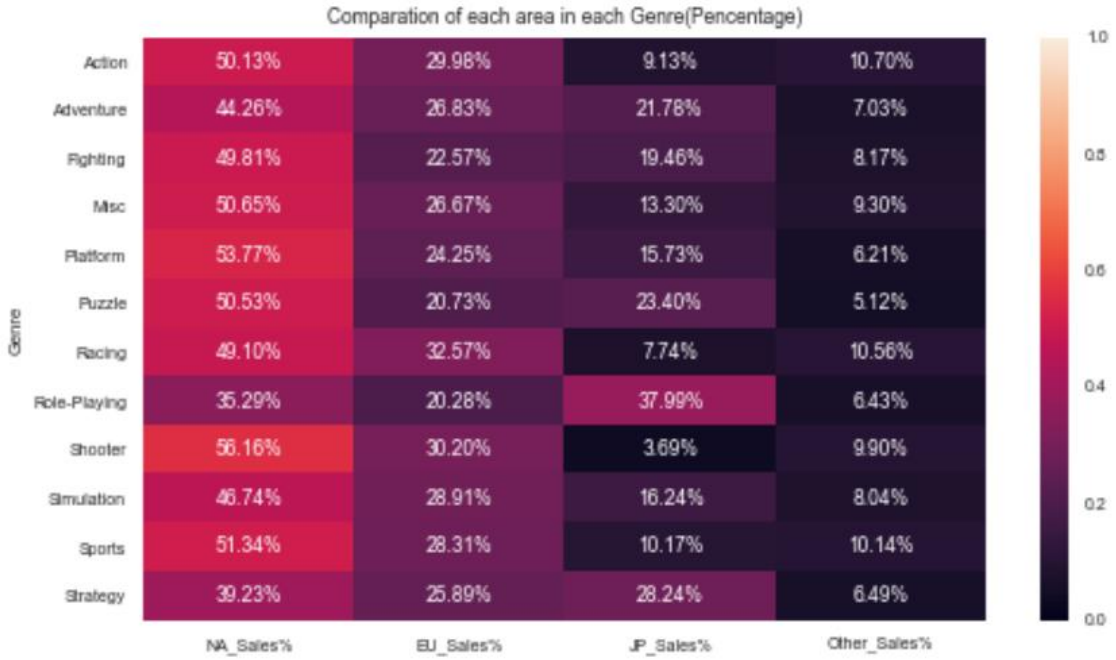
The client had provided me with a dataset, containing the sales from around 14600 games that have been released in between 1986 and 2016. This dataset contained information on several different markets on the topic, ranging from North American sales to Japanese sales to European sales.

This dataset was gathered by the client over the course of the 10 years that they've been active in their business. This data can be trusted as it has helped them gain many profits in the past, and has been praised by other partner companies to be highly successful.

Methods

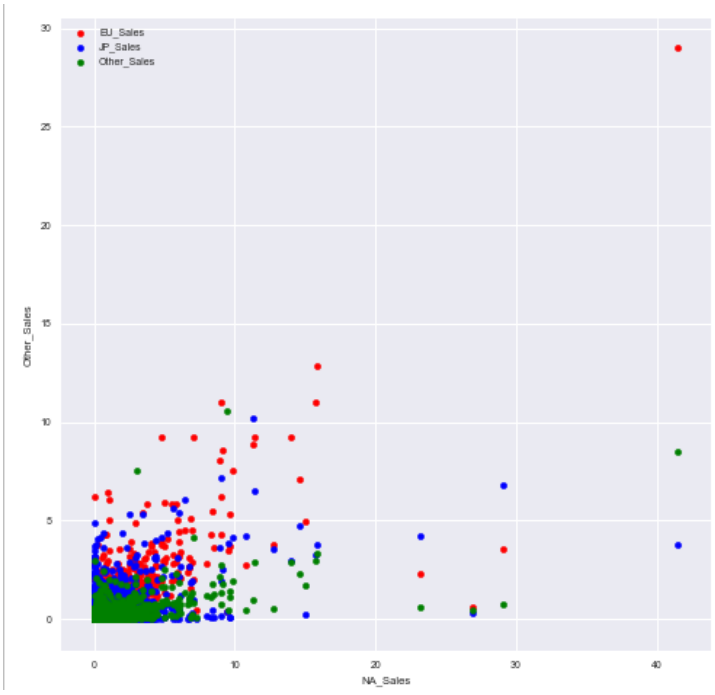
The first thing I needed to figure out, was what market is doing the best. I've visualized that data already in the business case, but here is it again:



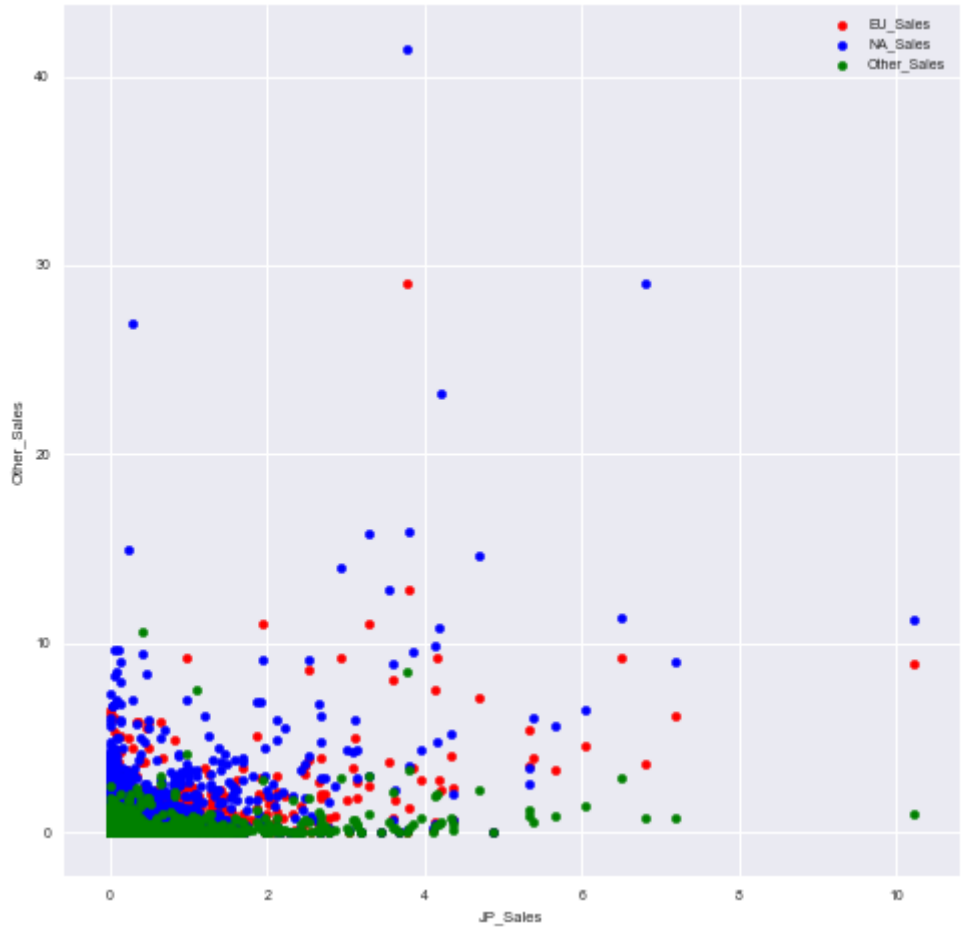


As you can see, the American sales are dominating the sales in every market, except from Role-Playing. I've thus concluded that this trend will stay the same for the next few years, and I'd decided to focus my efforts on the North American sales.

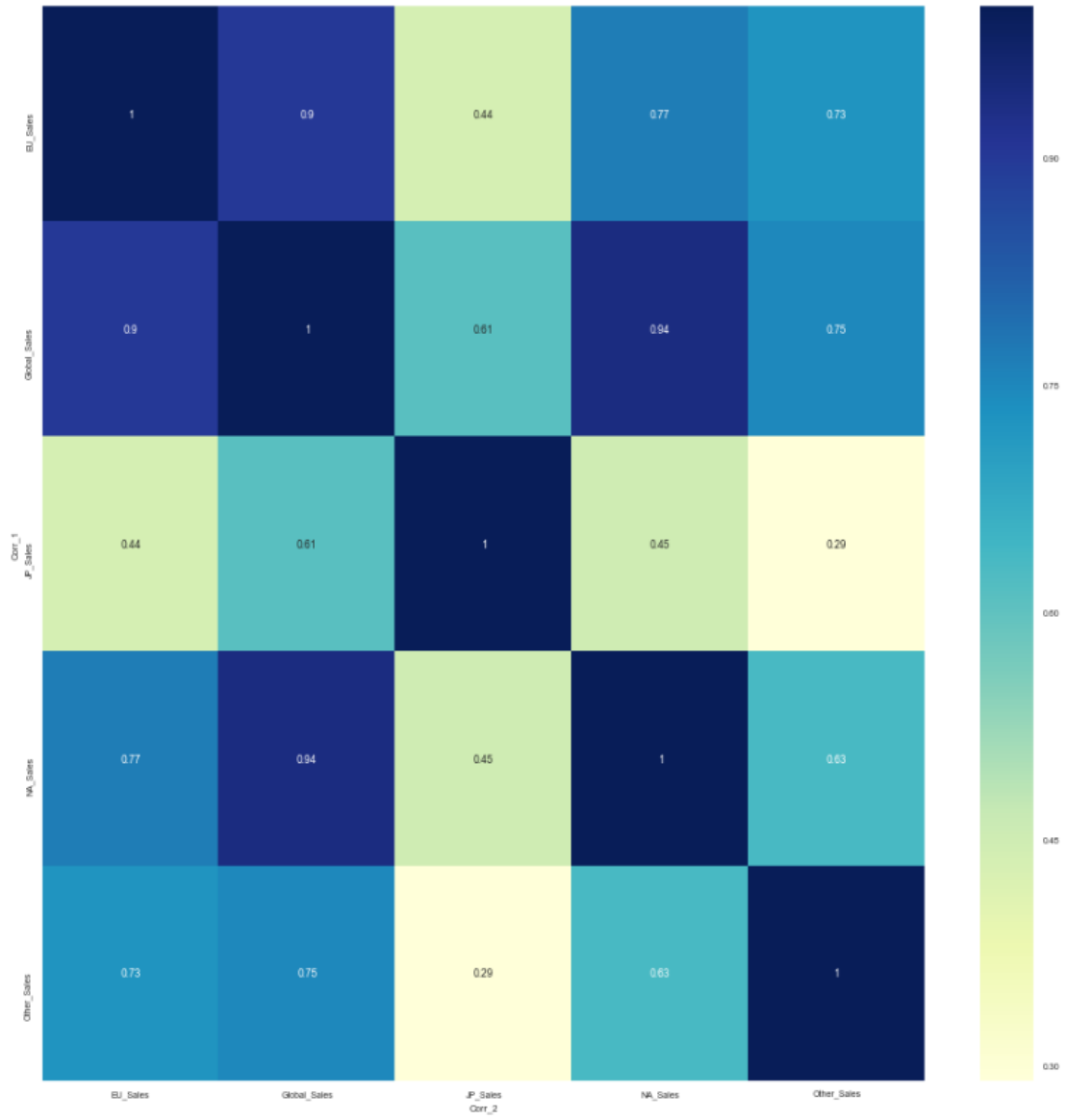
Next up, I was trying to find a correlation in the market sales. First I plotted all of the sales against each other as follows:



In the picture you can see that I plotted the North American sales against all of the other sales. I've done this for all of the sales:



These were all scatterplots, and didn't really give me a clear answer. Something I could work with. That's why I tried a different approach after that, using a heatmap:



In this picture it is clear to see what labels have a strong correlation and which don't. The highest one is the correlation between Global sales and North American sales. Based on this data, I'd decided to use the global sales to further investigate the north American sales.

Next up was the training and predicting part. First up, I had to reshape the dataframe a bit, so that it could be used for the regression methods I'd planned to use in the business case; Linear regression, decision tree and SVR.

The first method I'd explored was the Linear Regression:

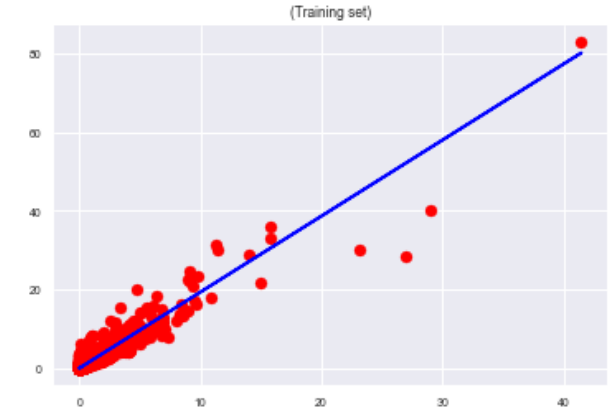


Figure 1: X-axis = North American Sales, Y-axis = Global Sales

The training set came with a score of 0.88. overall pretty good one could say.

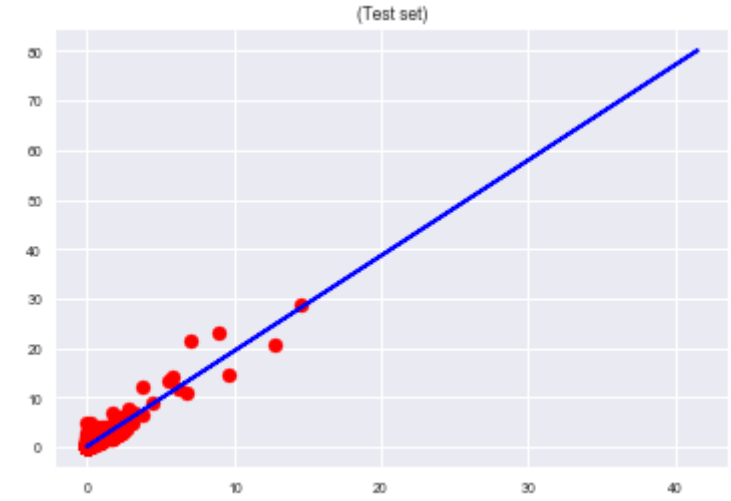
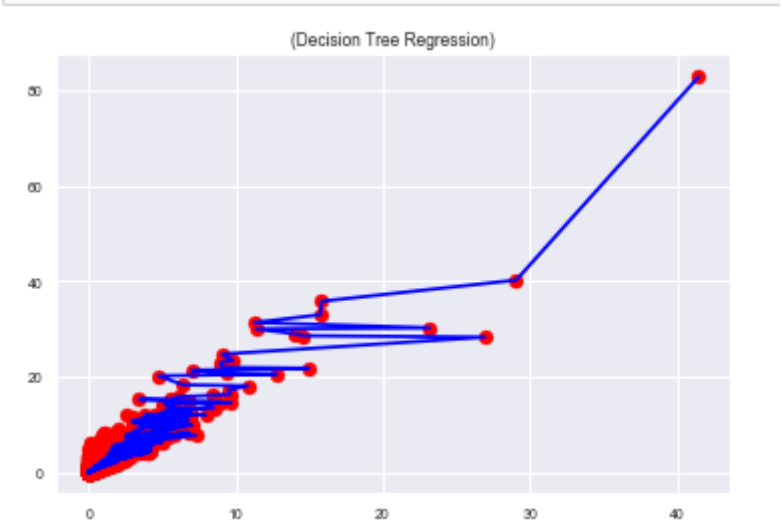


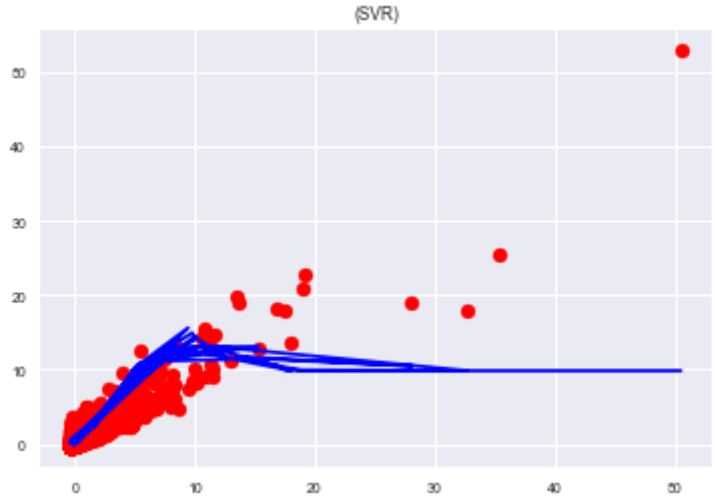
Figure 2: X-axis = North American Sales, Y-axis = Global Sales

The Test set came with a score of 0.90. Again, pretty good, but I was hoping in a bit higher than that.

Next up was the decision tree:



The decision tree managed to get a higher score than the linear regression, with a score of 0.95. This is probably the way to go for future attempts of predicting north American sales, but as I've promised in the business case, I wanted to test out the SVR too:



This method came in last with a score of 0.31. Concluding from these three methods, is that the decision tree and linear regression are the way to go for predicting north American sales, and that SVR is probably best used for something else.

Analysis

Based on my prediction results, it's safe to say that the North American market will continue to rise in profitability over the course of the next few years. The company should focus their efforts on the North American sales for two reasons:

- Based on my predictions, it will continue to be profitable in the next few years
- As it is the market which is by far the biggest, it also has the least chance of dropping huge amounts of profitability. Other smaller markets such as Japanese sales might face huge drops in profitability on a year by year basis.

Another interesting fact is that the American market dominates almost every genre, except for the Role playing genre. This is interesting for the client as it gives them a lot of choice on what particular share to invest in. The most popular genre being Action, this is a safe bet to invest in, as over 50% of the action games sold are being sold in North America.

Conclusion

In the introduction I stated a few questions that I thought that could help me with analyzing the problem, and I came up with answers for a few of them.

First of all, I've determined what genres are having the most impact on the sales of different markets. Action, Shooter and Sports seem to be the top 3 genres in terms of total revenue across the board, and will most likely be the safest bet when you want to buy shares in them.

Second of all, I determined that the global sales has good correlation with a lot of different markets. The strongest being the correlation between North American and global sales.

The important result here is that the North American market, the market that is by far the biggest of them all, is continuing in growing during the next few years, making a safe bet in buying shares there. It also has the least amount of risk of a potential drop in profitability, as again, it is the largest market.

Appendix A: Business case

2. Introduction

In this chapter I will shortly introduce this document and talk a little about the client in question and who I am. The rest of the document will contain information about how I want to try and find ways to help the organization to grow.

2.1 Description of the client

The client in question is a company that is interested in buying stocks from game producers. When saying game producers, think of the likes of Nintendo, Activision, Sony etc. The client has been very successful in the past, investing in the producers that are going to do well in the future years. But the last time the client invested in stocks, there was a miscalculation and they ended up with a huge amount of loss. There is no concrete plan given by the company, but the task given was to predict what company will end up on top in terms of global revenue in the next few years.

2.2 Our team

I am a student that recently graduated from the Fontys hogescholen university in Eindhoven. I currently work at Kaggle as a data scientist.

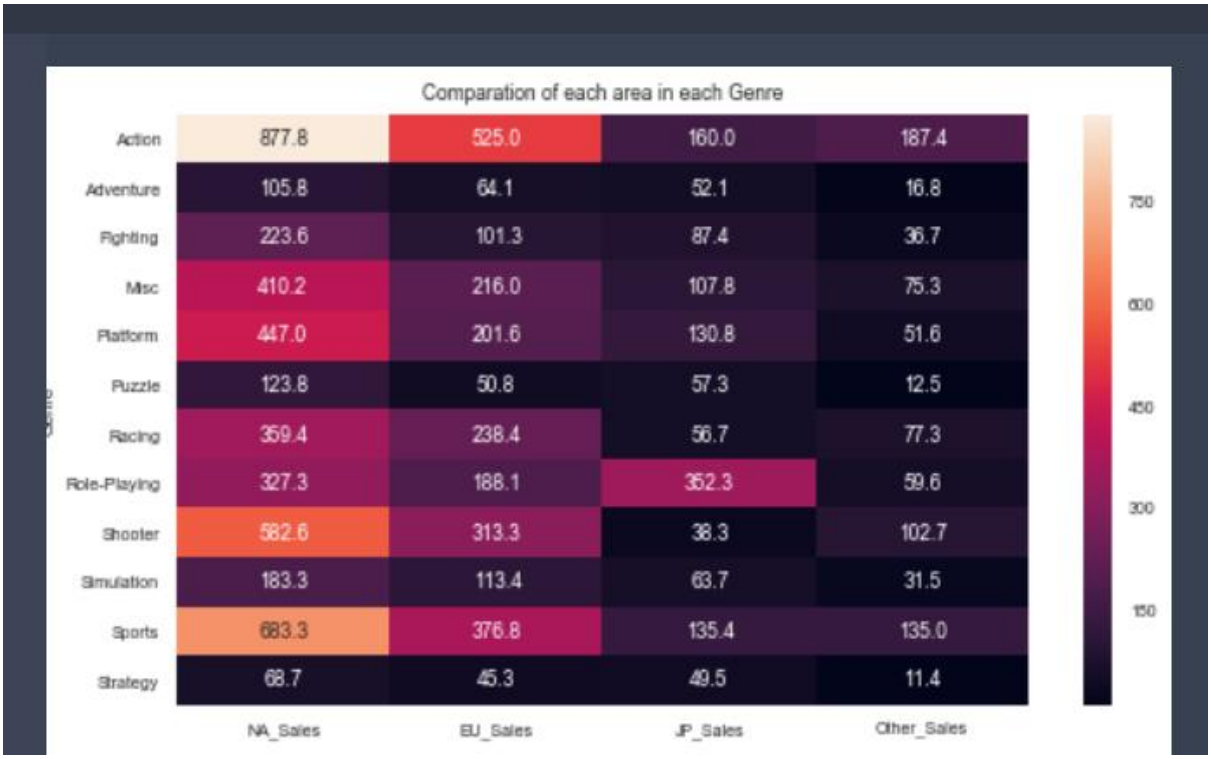
2.3 Planning

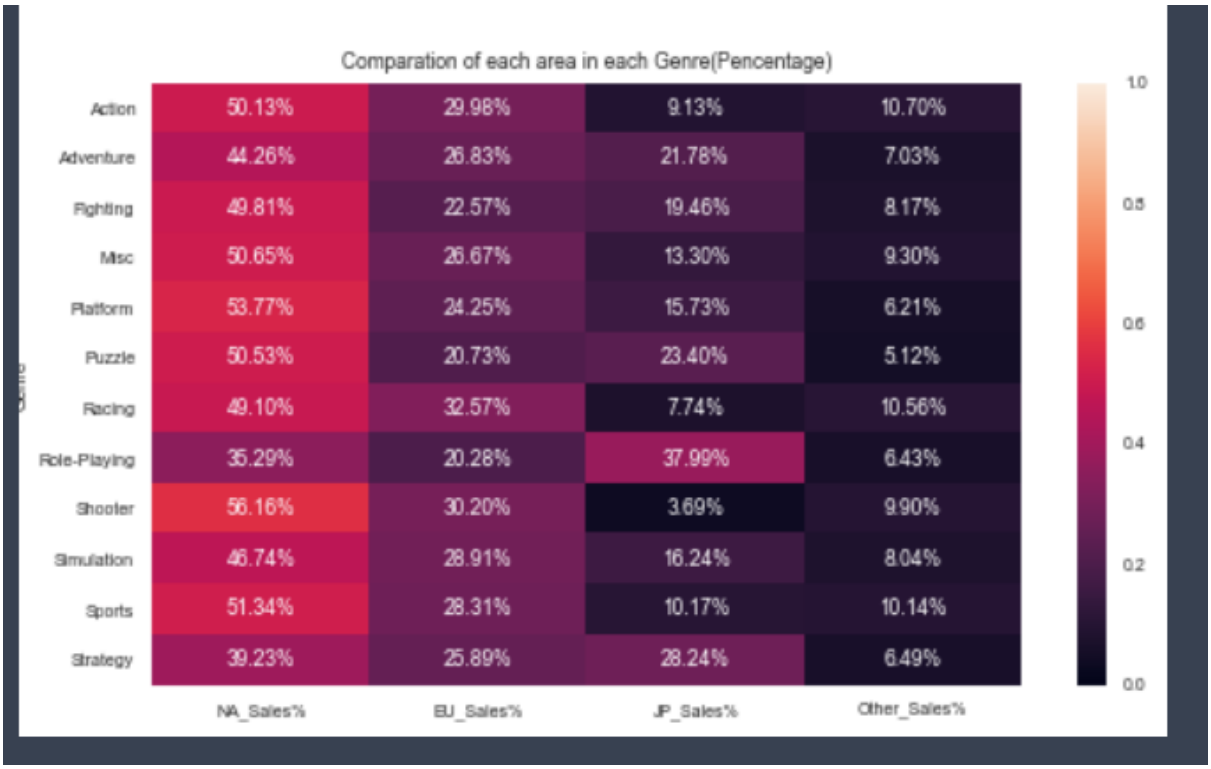
I will work according to the data science methodology written by Rollins (Rollins, 2015). It means I have three phases: Business case (this document), Data quality and Machine Learning & Reporting.

In the first phase I try to understand the client and define an analytic approach to solve the problem. In the second phase I specify the data content, formats and representations, I collect, understand and prepare the data for the chosen model. (I already received a dataset with information about the sales and customers, so I don't have to collect it by ourselves).

3.0 Business goal and problem

Popularity of genres change each year. This popularity has huge impact on the different markets. One year the action games are being very popular, while next year they could be outshined by another genre. The client in question made the miscalculation on what genre would be popular in the future, and lost a lot of money to it.





In the pictures above you can see the sales from the past 20 years (1986 to 2016). It is visible where, in the past, certain genres have already excelled. You might want to think that this trend will probably continue, but it might now be the case for the last few year, and for the next few years. I intend to find out what the best earning genre will be for the next 2 years.

While the client is mainly an European based company, the were also very interested in the popular genres of other regions for when they want to expand their company into other regions. It is then my client's question if they should continue investing in the European market, or switch to another market.

4.0 KPI's and metrics

Regarding the business and problem of the client, I decided to predict what market will be to most popular and profitable in the next two years. . Therefore, I need the concrete objectives and milestones for our project, so they can lead us to the right way. In this chapter, I'll specify the KPI's and Metrics of our project and explain their meaning and how I can measure the results.

- I want to predict what market will be popular and profitable In the next two years.
    - This can be achieved by looking at the data from the last years, and based on that data predicting the next few years.

- I want to increase the annual income of the client company by 5% compared to last years income.
    - This can be achieved by completing the goal of accurately predicting the popular and profitable genres for the next two years. After one year, there should already be an increase of 5% in the annual profits.

5.0 Analysis

In this chapter I will examine each 7 analytics topics in more details;

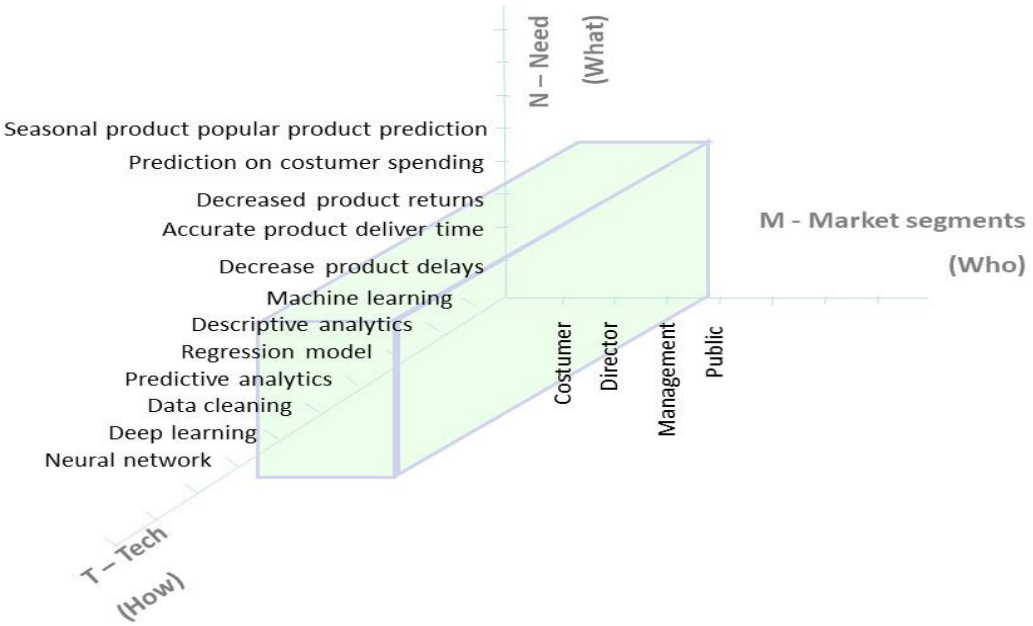
5.1 Assumptions

Assumptions are certain situations that I can expect to be true before starting the project. These assumptions must be true, otherwise I are unable to work further on the project:

- I assume that the company has a budget to invest in our idea.
    - This way I know that, should our idea get approved, I can work on the project I had in mind without finances hindering us.
- I assume that the data is accurate and real.
    - This way I can confidently work with the data without being insecure about the dataset.

5.2 Scope

For this case my business scope is centered in focusing on market profitability prediction. The Technical tools and knowledge I will use are centered on machine learning regression. This will all be done for the benefit for costumers and high management.



The timeframe of the dataset I received is all sales from 14600 video games from the years 1986 to 2016. This is the data I will use to base my predictions upon and all other data will be in the same timeframe.

The technologies I am going to be using to support and finish our project contain the following products and software:

- Python 3.x programming language; I use this to make the predictions and do the Machine Learning itself. I use this because it's the language I learned to use in the classes and this is the easiest and most streamlined language for Machine Learning.

5.4 Data sources and methods

The data sources that I will be using in the project is mainly the data I received from the client. This is the game sales data from 1986 to and including 2016 September 21st This will be the main data I will be using. I might use other datasets if I can find any reliable, but they will only be used to improve predictions.

Prediction wise, as my prediction will be a continuous value, I'm planning to use regression instead of classification. I will start off but using the following three algorithms :

- Linear Regression
- Decision tree regression
- SVR

The label that will be valuable will of course the different sales. Further correlation research needs to be done in order to be sure which ones will be useful for the prediction.

5.5 Cost/ Benefit Analysis

Implementing the market profitability prediction for the client is both beneficial and costly at the same time. I expect the prediction algorithm to be as accurate as possible so that the net income of the company can eventually increase. However, it should be reminded that the prediction algorithm itself costs a lot to be maintained after introduction, not even mentioning that a big financial investment is necessary in the first place to initially build the system. In this chapter, I will introduce which costs and benefits would be caused by implementing the prediction algorithm in detail.

Direct Costs

- Consulting fees and the actual payments for the data scientist
- System maintenance cost for collecting and cleaning the data

Indirect Costs

- Initial and ongoing training for the proper use of the machine learning algorithm
- Labor hours for implementation the algorithm

Benefit description

- Increased customer satisfaction
- Increased total annual income

5.6 Risk

Every business always has some sort of risk. The business case is focused on predicting the most profitable market. However, it is not guaranteed that I can accurately predict the most profitable market. The points mentioned above are risks that I can expect from our analysis because I are uncertain if I will truly obtain the right prediction.

5.7 Conclusion

With all the methods, benefits, costs and risks taken in consideration, I can be certain of the actions I plan to take. I will ensure that the benefits will outweigh the ethical and cost-effective risks. Given our selected scope for the solution, I have a clear vision on the what and how factors concerning our case for you. I can't say for sure if I will get accurate results, but even a slightly accurate case will most definitely be very helpful for the company in the future.

Appendix B: Code

See attached file: Code_Challenge.pdf

A interactive version of the same file will also be attached: BrentChallenge.ipynb