Project Final Report

|  |
| --- |
| **Title of your data science project**  The Factors that Affect Total Game Sales on Steam |
| **Members of your project**  PONGSAKORN TANUPATRASAKUL1702060008  KAMIN PROMMAS 1706120001  THANAKORN KAEWPLOY 1801050001 |
| **(update) Role of each member**  PONGSAKORN TANUPTRASAKUL- Coding, Documentation |

**1. BACKGROUND AND OBJECTIVES**

1. **Description of your project**

We are analysing factor that affect game sale on Steam.The main goal of this project is to find multiple factors that affect the total amount of each game that is sold. These days there are so many games in the market, and with Steam being one of the most popular gaming retailers, there are many people trying to sell their games on Steam. This project will hopefully find the many ways that can affect a game being sold, such as reviews, whether they are positive or negative, and even the price.  
The information found from this project, can possible help a game developer to understand their customers, and what they can do to maybe help their sales go up. The logical prediction from our data set so far is that, positive reviews will mean more total sales, negative reviews will mean less total views and the larger the price, the lower the total sales. However, this project will give us the insight on whether our predictions are right or wrong.

*State the motivation or problem that you are interested. Argue why it is a important problem*

It is important because in the future if we created or work in gaming industry we need to consider what type of game will make it into the market and what aspect of the game would lead to higher sales.

*State three specific data science questions you want to answer*

*The 3 specific data science question is that does positive review affect Sale, does game price affect Sale and discount event vs Sale. This is the factor that we think it going to affect the sale the most.*

*State your general idea how you gonna tackle this problem. Argue why it is a convincing solution*

*We going to tackle the problem by comparing all the factor with the total sale by using*

*Scatter graph.*

*State your general expected results:*

It depends on the mostly on the price and positive review of the game and also discount event.

1. **Goals and objectives**

*State your goals and objectives. Elaborate what exactly you wanna do. Follow SMART goals.*

**Goal : Factors that affect the sales of the games**

**Specific-** We want to find out does reviews affect the sale of games.

**Measurable -** When Reviews start to fluctuate based on the ratings.

**Attainable-**It can attain by having a wide range of data available so it can be more accurate.

**Realistic-** By the result we can predict the outcome of the research**.**

**Time Bound-** The dataset is collect from2012 to 2018.

1. **How do you measure that you have achieved your goals and objectives**

We measure our goal by if the data we get in the top 10 game that game with most sale most likely to be a popular game because most game that make lot sale is alway an outlier from our data that we have because most common game is alway in the same range so we can measure by if the popular game is the same as the information that other post online.

**2. PROCEDURES, METHODS, AND SCOPE**

1. **Describe your procedure**

*State where and how you gonna gather the data:*

The data so far has been collected from two websites, SteamSpy and SteamDB. The data has been entered manually into an excel file, then converted to csv. At the moment we have only 86 entries, but in the future we will expand this list, as it is a time-consuming process. We plan to use the linear regression technique, with a line of best fit, that can also help us to predict our data.   
We also expect to clean the outliner out to make the graph more clear and use “log” to make the graph look more tidy. As we collect more and more data, we may use a prediction model to fill in any missing data.

*State what are the expected cleaning process you have to do*

We will replace the missing values with a value such as NaN or use the mean as a placeholder. As our data is fairly limited since Steam don’t really make all this data readily available to the public.

*State what are the main data analysis algorithm you plan to use*

*We did data mining on steamDB and Steamspy. These both website is a sub website from steam which takes record of the sales and all game database which is sold on steam.*

*Then use linear regression for a scatter graph to plot the two different variables against each other.*

*State what is your plan regarding splitting training, testing, and validation set*

We plan to find more x to solve y by introducing discount sale frequency because some people will buy or try the game when it is discount so this might be one of the important factor because some game are way to expensive and the discount on steam is very effective to many gamer. Hence, this could be an important factor in the rise of the sales in the game.

*State what is your baseline model that you plan to compare with your proposed model*

We plan to compare by using the median and mean to see if the average match or nearly the same.

*State how you gonna know your model works well*

If the model work well it will most likely show the same results as in the steam website but it can be more accurate and easy to read ways. This will also logically read the same result as reality like game A is very popular so the data we got also need to show Game A is popular.

1. **Limitations**

The limitation is that it is very hard to find the genre other than that mostly our game is action game. There some game that have way high sale than other which mean outlier mostly it gather in the same range but for some sequel and popular company it sale seem higher than other not including the price and the review. All data was entered manually which is time consuming, and took up a lot of our time, because without the data we cannot use the model. Manually entering in data can also cause problems because humans are easy to make mistakes, so it’s possible that our data may not be totally accurate. We are also not expert in Python so it’s very difficult for us to fully comprehend the project we are facing, we took all the knowledge from research to attempt to use in this project. It’s also difficult for us to find the data, as Steam do not publish the date so publicly. We relied on 2 sites to get the data for us, which also contributed to why we had to enter in all the data manually. Also our data for total sales was an estimation, meaning that our data cannot be totally accurate.

**3. PROGRESS**

1. **Report your progress**

We used logarithm to expand our graph, as before it was very close together and cannot interpret the data at all to see the result. We also added in more x variables, such as the genre and also the frequency of times that the game had been put on discount to compare with the total sales.

**b. Report obstacles if any**

We tried to include a line of best fit for prediction purposes which was achieved, however, due to lack of knowledge in python, it was agreed upon by the group that the line did not look correct.

There were several attempts to fit a line of best fit, that were not successful.

The data entry aspect really held us back, as it was time consuming and since our project really relied on the data, we spent a long time on collection for not a lot of data.

**4. RESULTS**

**Report your results**

*As of the results, we tend to have compared results of the steam game in various aspects to know which criteria affects the sales of the game. However, after collecting data and running the test, we found out that Total Sales and Positive Reviews have a positive correlation as the more positive reviews the sales of that game rises. Moving further we tested the sales of the game with negative reviews and we discovered that it had positive correlation but with variances as not all games were affected with the negative review. In addition, we further testing the aspect of the sales of games with the prices of the games and we studied that it didn’t really have a correlation and certain point of view it didn’t really affect if the games were expensive possibly some hardcore gamers still prefered going for it. Moving further ahead, We tested the aspect of sales with the frequency of discounts offered and we studied that discounts on games had a positive correlation with plateau, Basically discounted games increased the sales of the games but just to a certain level.*

**5. DISCUSSION**

**Discuss your results**

There seems to be a strong positive correlation between the positive reviews and the total sales, confirming our theory that having more positive reviews will mean that the game would get more total sales. However, there is also a positive correlation between the negative review and the total sales, but compared to positive reviews, it is not as strong a correlation and has more variance in the data. This is a little unexpected as the logical conclusion would be that more negative sales would mean that they would get less total game sales, however, this doesn’t seem to be the case with our data. There is a positive correlation between the total sales and total reviews, as it could be implied that more reviews negative or positive, would boost popularity for the game and therefore increase sales as well. The game price factor, against total sales is a little difficult to interpret from our graph. As there doesn’t seem to be much of a correlation, where we expected maybe that the cheaper the game, the more sales that it has, but the graph indicates that the price of the game really does not matter too much in how many sales it gets. There is also the case of the frequency of discount. The obvious answer would be that the more times the game has been discounted, the more sales it would get. This can be seen from our results that there is indeed a positive correlation between the two, however at a certain point, it plateaus. This means that it appears that once a game has been discounted so many times, that maybe, all the people who wanted to buy the game, have already bought the game the first few times it was on sale, therefore, there is no more people willing to buy the game. So once it has reached a certain point, there just simply aren’t many people interested anymore, and a discount would not have a great effect on their influence to buy the game.

**Limitations and future work**

**Self reflection**

This project has helped us to self-learn Python, as we went into this project not really knowing anything about it. A lot of it was trial and error, however, it proved useful and the research that we conducted in order to learn this, hopefully help us succeed with the outcome of the project.

We could have expanded more on other models, rather than focus a lot on linear regression technique, we could have used a prediction model and others for data analysis, however time constraints and low skill level made this expansion difficult within our project.