MICROSOFT MOVIE PROJECT

1. BUSINESS UNDERSTANDING

1.1 :OVERVIEW

Microsoft , an American multinational technology company, wishes to create original video content , which has become quite popular . The company wishes to achieve this by creating a new movie studio tasked with movie production.However, they do not know anything about creating movies and need insight on the kind of genres they will need to produce based on certain factors. This project will entail the analysis of data provided from various sources such as IMDB and rotten tomatoes to help generate insight on this project so that the genres recommended at the end of the analysis will be the best fit in terms of Box office performance .

1.2: PROBLEM STATEMENT

This analysis will revolve around finding the best genre of movies that the new Microsoft Movie studio can produce. The project will centre its analysis based on movie ratings , the gross money returns as well as do an in depth analysis on the relationship between Box office performance and the production costs to get an approximated project budget.

2. DATA UNDERSTANDING

2.1: DATA COLLECTION

The data that will be key in the analysis that is about to begin has been gathered from various movie review and summary websites.These websites entail Box Office Mojo (https://www.boxofficemojo.com/) IMDB (https://www.imdb.com/) Rotten Tomatoes (https://www.rottentomatoes.com/) TheMovieDB (https://www.themoviedb.org/) The Numbers(https://www.the-numbers.com/) However, i will only include the dataset from IMDB and The Numbers.

2.2: DATA DESCRIPTION.

The datasets that will be in use are the IMDB and The Numbers datasets. The IMDB dataset is in the form of an sqlite database while The Numbers dataset is in the form of a CSV file.The IMDB database contains two tables each containing a common column which is the movie id column. The Numbers dataset, assigned the variable name movie_budget_df has 5782 rows and 6 columns, with the column names as id ,release_date , movie ,production_budget ,domestic_gross and worldwide_gross. The IMDB dataset assigned the variable name Im.db_df has two relevant tables to the analysis, namely ; movie_ratings table and movie_basics table. The two tables have been converted to dataframes and the movie_ratings table, assigned the variable name ratings_df has a total of 73856 rows and 3 columns. The column names are movie_id, averagerating and numvotes.The average rating is affected by the numvotes column. The movie_basics table has been converted to a dataframe with the variable name movie_basics_df. All these datasets have been merged , beginning with the dataframe movie_budget_df and the the movie_basics_df and the merged dataframe assigned the variable name merge_df. The final merge has been conducted between the merge_df data frame and the ratings_df and the final data frame which will be in use for the remaining part of the analysis assigned the variable name final_merge_df.
The final_merge_df data frame contains some missing values.

2.3 : DATA PREPARATION AND CLEANING

The final_merge_df data frame has a total of 118 missing values in the runtime_minutes and 8 missing values in the genres column. The approach taken to clean this dataset is to remove the missing values. This has been achieved by dropping the entire runtime_minutes column since it will not be essential in my analysis as well as dropping the rows which have missing values in the genres column. The next step is to check for any duplicate values and in the final_merge_df dataframe .In this case, there are quite a number of duplicated values , shown by the primary_title and id columns.The duplicated values have therefore been dropped and the number of rows decreased.The dataset now has 2,126 rows and 13 columns There are no outliers in the data frame as well as extraneous values

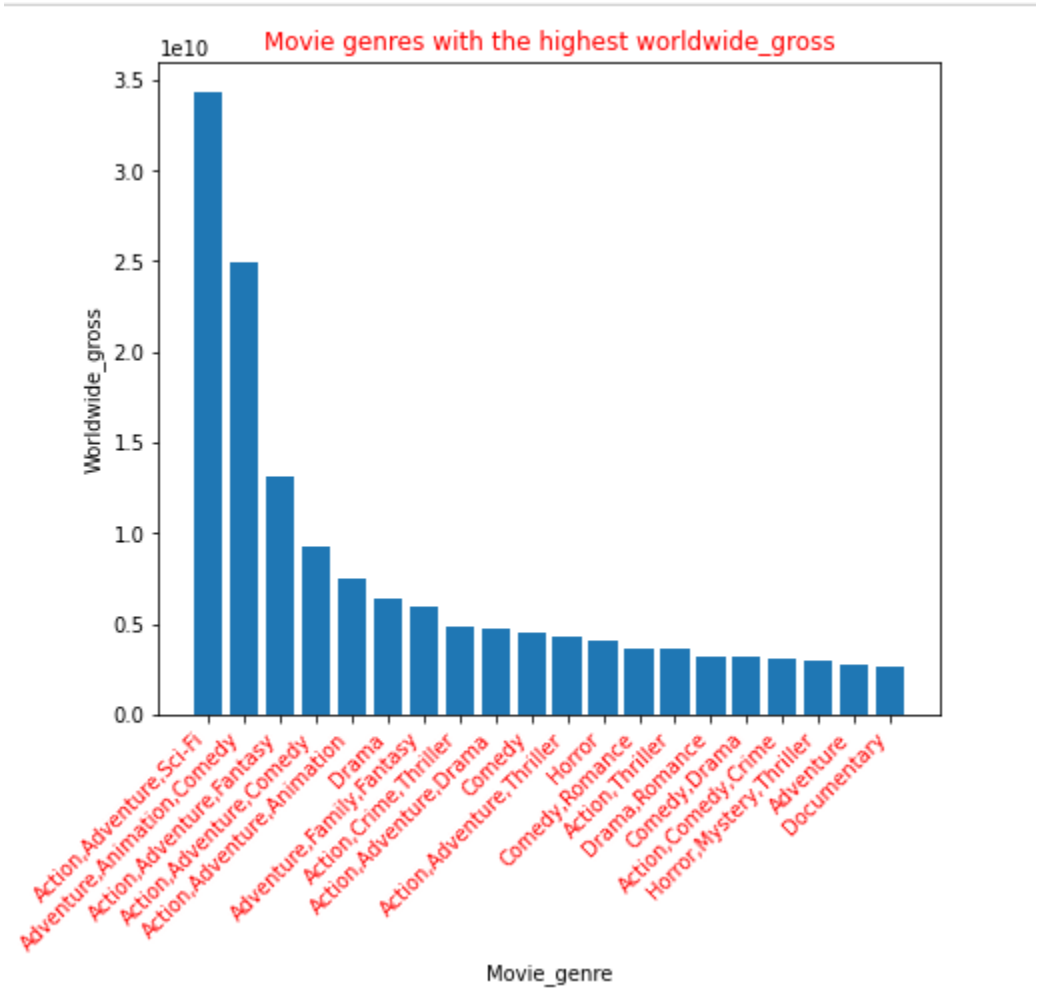
3.DATA ANALYSIS

The method used in this analysis is the Exploratory Data Analysis with pandas , EDA with pandas,where each analysis has been conducted based on theperformance of a single variable.

3.1: Univariate Analysis

3.1.1: Analysis based on the worldwide_gross income generation

The final_merge_df data frame used in the analysis was clean and only required particular columns for this analysis.The columns required were the genres column and the worldwide_gross column which was already inclusive of the domestic_gross column. The new dataframe created was assigned the variable name income_generation_df and will be in use henceforth. This new dataframe was only inclusive of the genres and worldwide_gross and the primary_titles were irrelevant in this case so i had to group these datasets based on genres, and the resulting worldwide_gross column was now an average of the worldwide_gross values having the same genre. The variable name was updated with these changes and now the analysis could be done using these to columns to determine which movie genres had the highest worldwide_gross income.

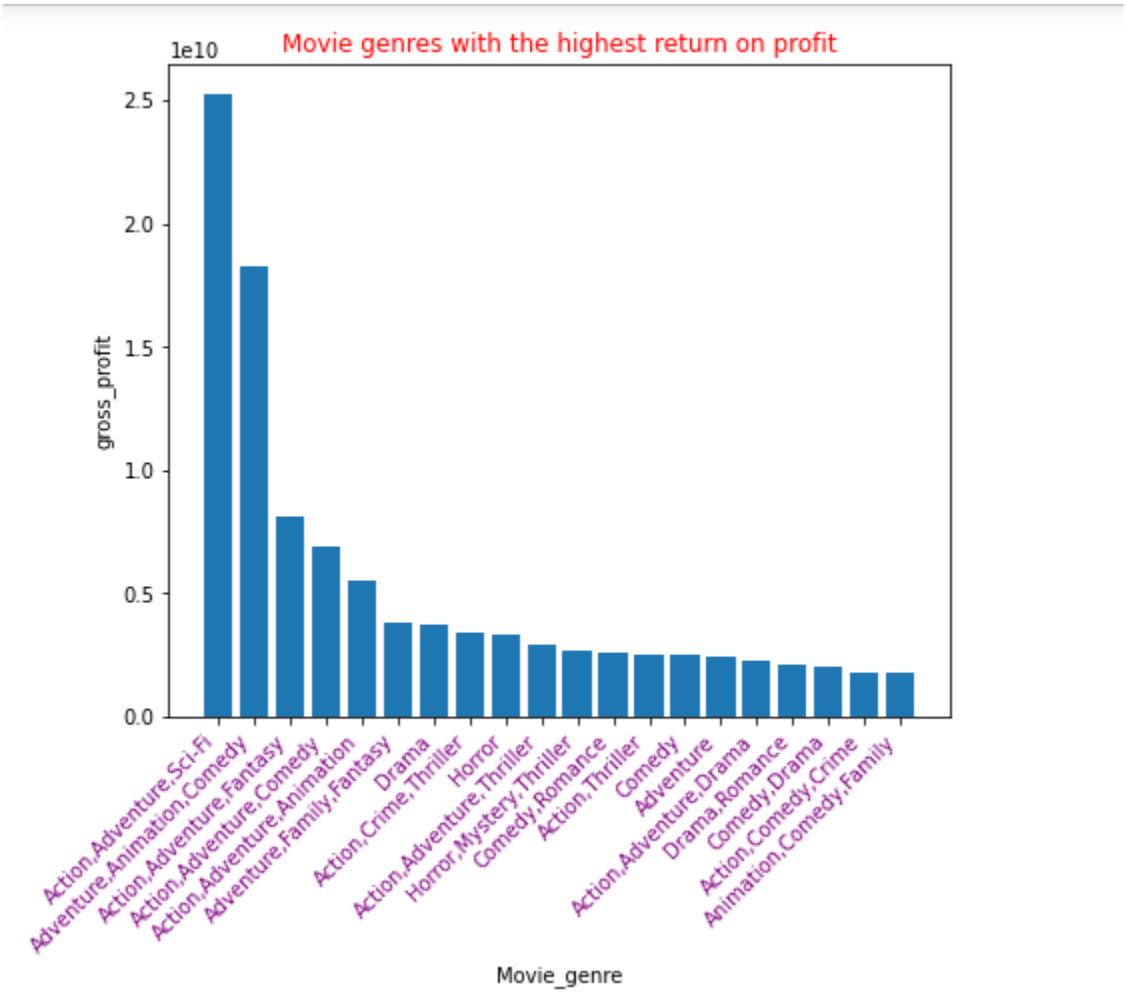


Movie genres with the highest average worldwide_gross.

Based on the bar chart presentation above , the genre that seems to generate the highest gross income world_wide would be the combined Action,Adventure,Sci-Fi genre.

3.1.2 :Analysis of genre based on the gross_profit

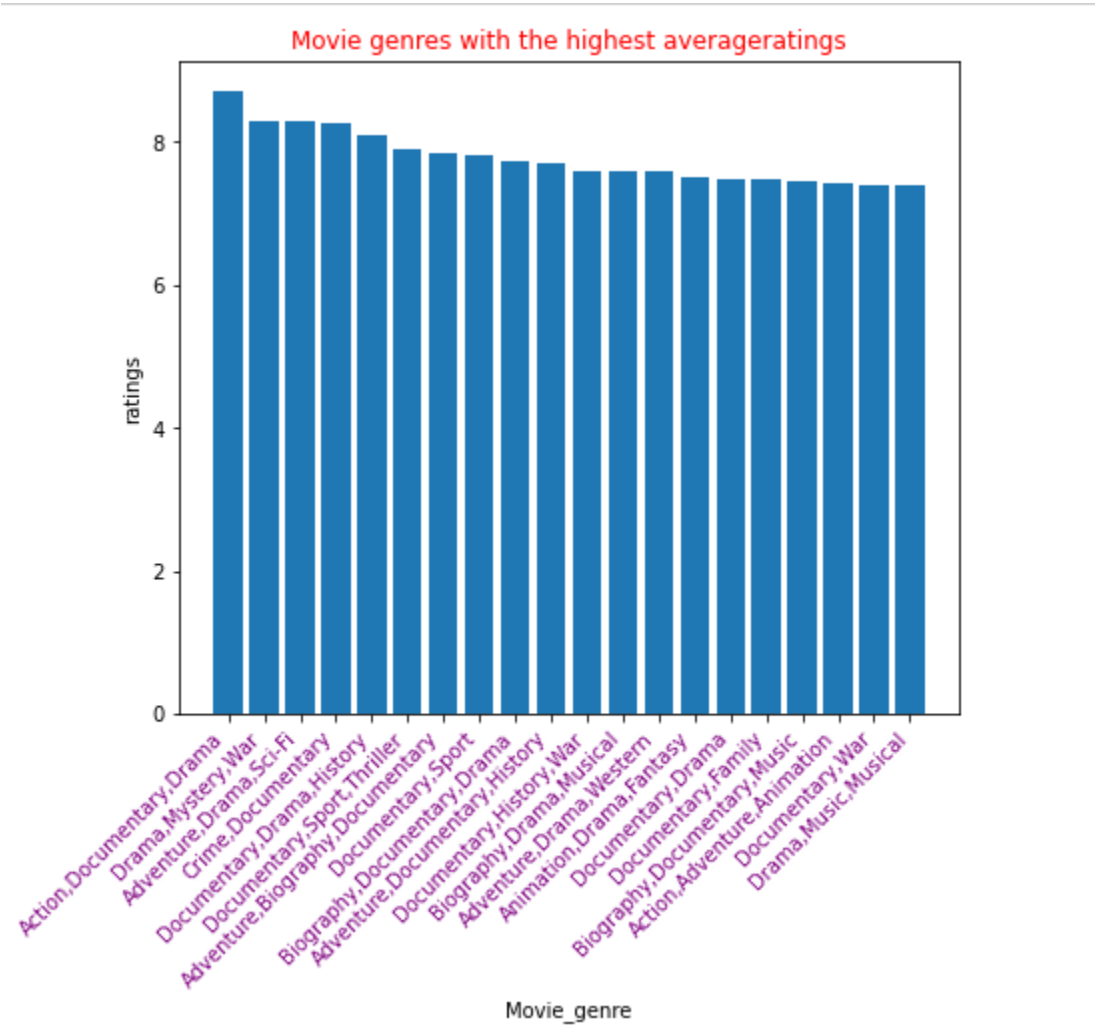
This analysis was conducted to obtain the movie genre that generated the highest profit having gotten the returns on the production budget. For this analysis, the columns that will be in use will be the genre column and the gross_ profit column . The gross profit column was obtained by getting the difference between the worldwide_gross column and the production_budget columns and assigned its variable name. The next step was to create a dataframe with the columns of interest and assign it the variable name gross_profit_df. Once the dataframe was created , the next step entailed grouping the data frame by genres and reassigning the variable name to update on the changes. The resulting data frame contained the genres column and average gross_profit which were plotted in a bar graph as shown below.



Similar to the worldwide_gross income table, the genre that seems to generate the greatest income , having compensated for the production budget is the combined Action,Adventure,Sci-Fi genre which is the best choice putting return on investment into account and probably the Adventure,Animation,Comedy genre as a second choice

3.1.3 : Analysis of genre based on average ratings

For this part of the analysis, the columns in use were the genres column and averagerating columns.I created a dataframe called rate_df which only contained the columns of interest and grouped them by genre to obtain the resulting data frame having the unique genre in a single columns and the averages of the averagerating values as a single column. I then used these values to plot a bar chart as illustrated below to generate some findings.



Based on the barchart illustrated above , the genre that had the highest average ratings was the combined Action,Documentary,Drama genre.
This however may seem odd since a good number of the movie genres that demonstrated the highest worldwide gross and gross profits are not included in the chart. To determine if the average rating was affected by the number of votes , i conducted a
t- test to investigate this claim.I used the null hypothesis stating that the number of votes does not affect the average rating and obtained a t-value of 0.0, showing that there is a significant relationship between the number of votes and average ratings. In an attempt to obtain a more accurate chart on genres and average rating , i conducted an analysis to determine if the gross_profit and production budget were related in any way.

3.2 : Bivariate Analysis

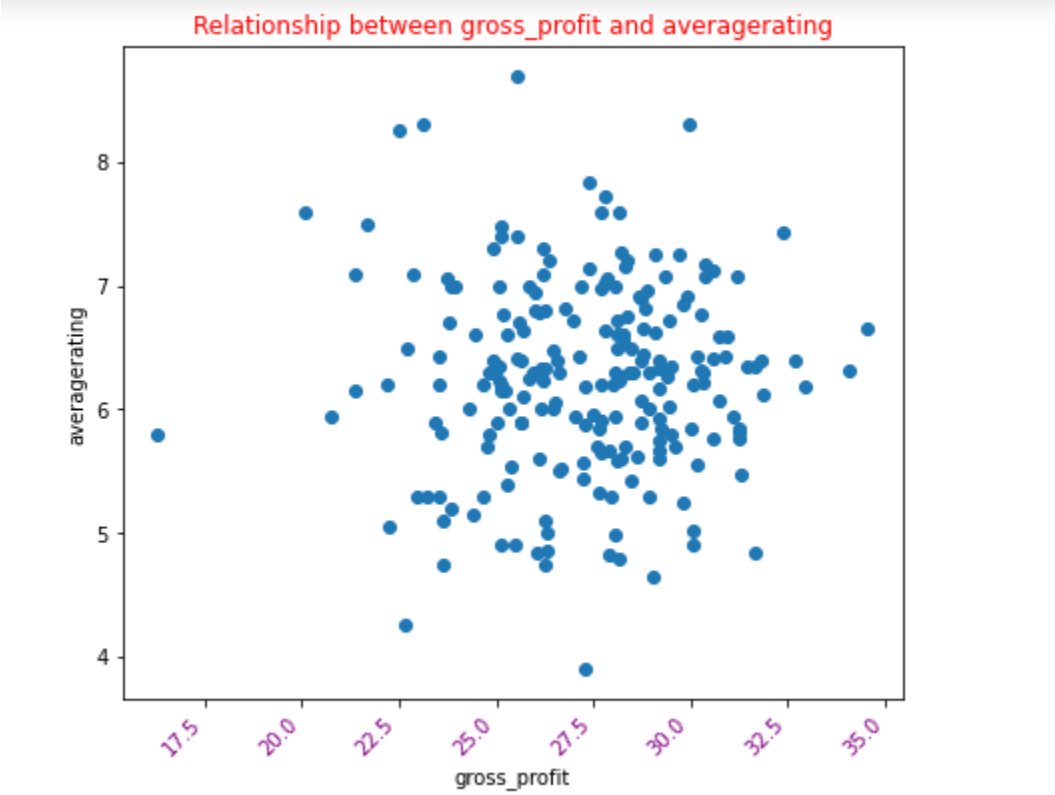
This analysis entailed testing whether the averageratings could be determined based on their relationship with some given factors.

3.2.1: Is there any relationship between the average ratings and gross profit?

To determine if there was any relationship between the averageratings column and the gross_profit column, i had to create a data frame, profits_and_ratings_df and assign it three columns , genre,gross_profit and averagerating,obtained from the main data frame , ie.the final_merge_df,like all the previous columns. Once grouped, the resulting data frame was used to plot a scatter plot to determine if there was a relationship. However , the gross_profit column values had to be normalised so as to get a more distinct look on their relationship.

The scatterplot will be illustrated in the following page

As stated earlier, this test was to show if there was any relationship between the gross profit and the ratings and if a regression line is included in the scatter plot , there will be a linear relationship between them. In this case, there is no relationship between the gross profit and averagerating.This can be further backed up by the correlation coefficient value of 0.009 showing no relationship.The accurate averaterating based on genre cannot be depicted by this factor alone.Let us then investigate on whether there is a linear relationship between the production budget and ratings.



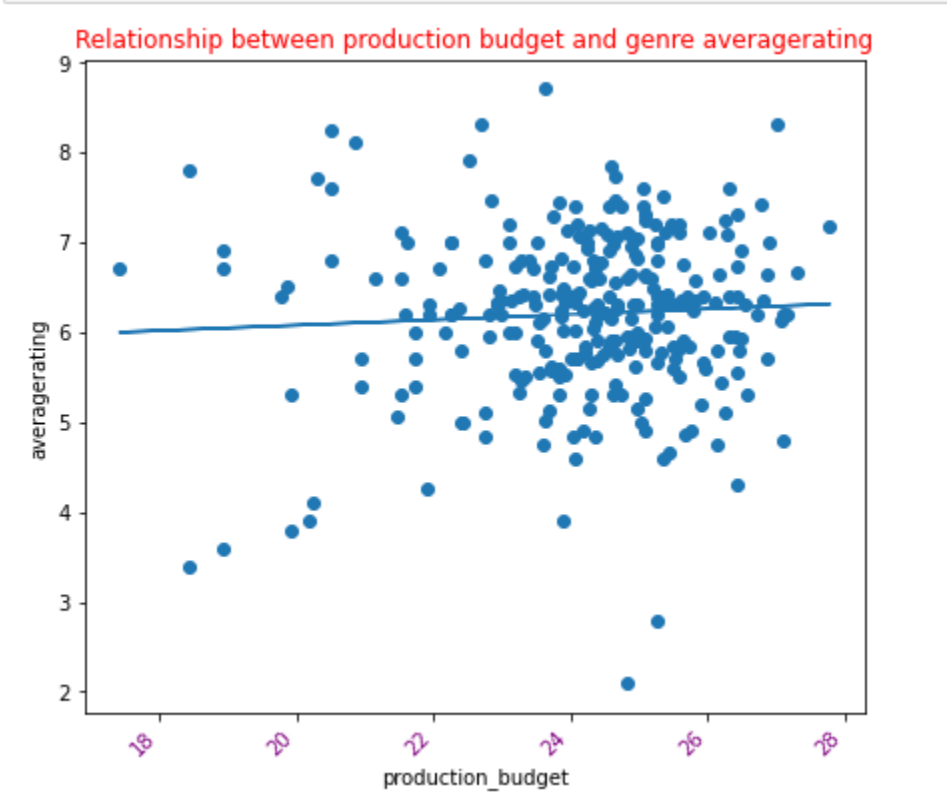
3.2.2 : **Is there any relationship between the production_budget and ratings?**

Since there was no linear relationship between the gross_profits and average ratings,the next test to be conducted was on whether or not there was a relationship between production_budget and average ratings.To conduct this analysis , the procedure is the same and involved obtaining our columns of interest from the final_merge_df and assigning the new dataframe a variable name. In this case, the columns of interest are the genres, production_budget and averagerating. The new data frame created was assigned the variable name budget_and_rating_df. The data frame was then grouped using genre ,normalised and plotted as shown in the following scatter plot

The scatter plot illustration will be illustrated in the following page

As clearly indicated in the scatter plot below, there is no relationship between averageratings and the production budget. This can be proven further by the correlation coefficient of 0.0589 Based on the analysis conducted above as well as its prior , there is no accurate way of determining the success of a movie based on its averageratings obtained from a relationship

with a single different factor. This is due to the fact that there are so many factors that can affect the results of the average rating,including a direct cause on the number of votes. Regardless of the uncertainty of the analysis done earlier on the movie_genres and average_ratings, it is what i will use in the analysis



3.2.3 : Relationship between production_budget and gross_profits

The next step was to test if there was a relationship between the production_budget and the gross_profits. To test this relationship , I obtained the production_budget and gross_profits columns and create the dataframe profits_and_budget_df. I then grouped these rows based on genres to obtain unique rows based on genres. I then plotted the scatter plot without having to normalize the data values and obtained the scatter plot illustrated below;

Based on the scatter plot above... There is a positive linear relationship between the production budget and gross_profits as well as a strong correlation coefficient to further prove this . This shows that movie_genres that had the highest production cost also had the highest gross_profits. Microsoft should therefore put this factor in consideration in the case that they intend to produce the most successful movie genres in terms of gross profits.This can further be proven by the correlation coefficient value of 0.638 showing a strong positive correlation between the production_budget and gross_profits.

4. CONCLUSION

In order to produce a certain movie genre, some factors have to be put into consideration based on where the interests of the production company are. If they wish to maximise on profits, then they would put emphasis on the movie genres that generated the highest profits. If their interests were on minimising on the production budgets, then the focus would be centred around the movie genres that had the lowest production budgets. If their interests were focused on the movie genres that had the lowest production budgets but still generated the highest gross profits, then the analysis would be based on that. This analysis has been centred around the movie genres that generated the highest gross profits, that had the highest worldwide gross generated and averagerating. We also went a step further by analysing the relationship that exists between the production_budgets and

grossprofits and averagerating and it was safe to conclude that the averageratings could not be predicted based on these factors alone.

5: RECOMMENDATION

Based on the analysis conducted:

a)If Microsoft wishes to produce a movie genre that has the highest worldwide_gross income, then the recommended genre would be the combined Action,Adventure and Sci-Fi movie genre

b)If Microsoft wishes to produce a movie genre that has the highest gross profits based on the production budget, then the recommended genre would be the combined Action,Adventure and Sci-Fi movie genre.

c)If Microsoft wishes to produce a movie genre that has the highest movie ratings, then the recommended genre would be the combined Action,Documentary,Drama movie genre.

d)In order to produce the best genre based on either gross_profit and worldwide_gross, Microsoft would need to consider the probability of a high production budget