Group 10 Draft Report - TMDB Movies Analysis

By: Muqsit Momin, Prakhar Jain and Siddha Deshpande

Executive Summary

In our analysis, we aimed to identify the key factors that contribute to a movie's success in terms of revenue and ratings. We conducted a thorough examination of various aspects of a movie, such as genre, cast, runtime, and budget, to determine their importance in predicting revenue and ratings. Additionally, we investigated the correlation between revenue and ratings to understand how they relate to each other.

What makes our analysis unique is its comprehensiveness. We explored a wide range of movie characteristics and even identified the optimal runtime for a movie based on ratings. Our findings and visualizations provide a clear roadmap of our discovery process, highlighting the steps we took and the conclusions we reached.

Overall, our analysis sheds light on the critical factors that contribute to a movie's success and provides valuable insights for filmmakers, producers, and distributors seeking to make informed, data-driven decisions.

Visualization:

Major Steps:

Cleaning

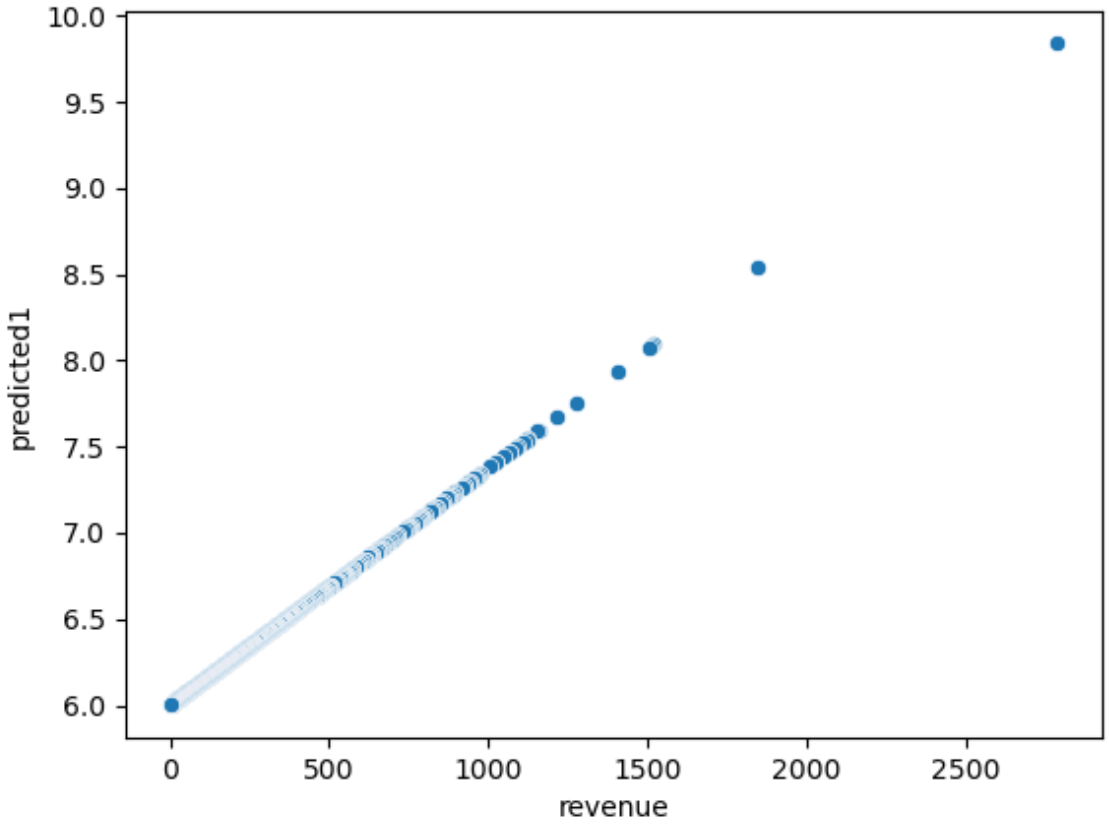
The provided code performs data cleaning and organizing tasks to prepare the data for regression analysis. It reads in two CSV files and creates data frames, cleans the data by removing unnecessary columns, converting data types, and extracting the first genre from the genres column. It then saves the cleaned data frame as a new CSV file. Finally, it loads the cleaned data frame and further cleans it by dropping rows with missing or infinite values to ensure that the data is ready for regression analysis.

Multiple visualizations and description of its analysis

OLS Linear Regression Analysis

OLS regression analysis where the dependent variable is vote_average(predicted1), and the independent variable is revenue. The results indicate that revenue has a significant positive effect on the vote_average. In other words as revenue increases so does rating.

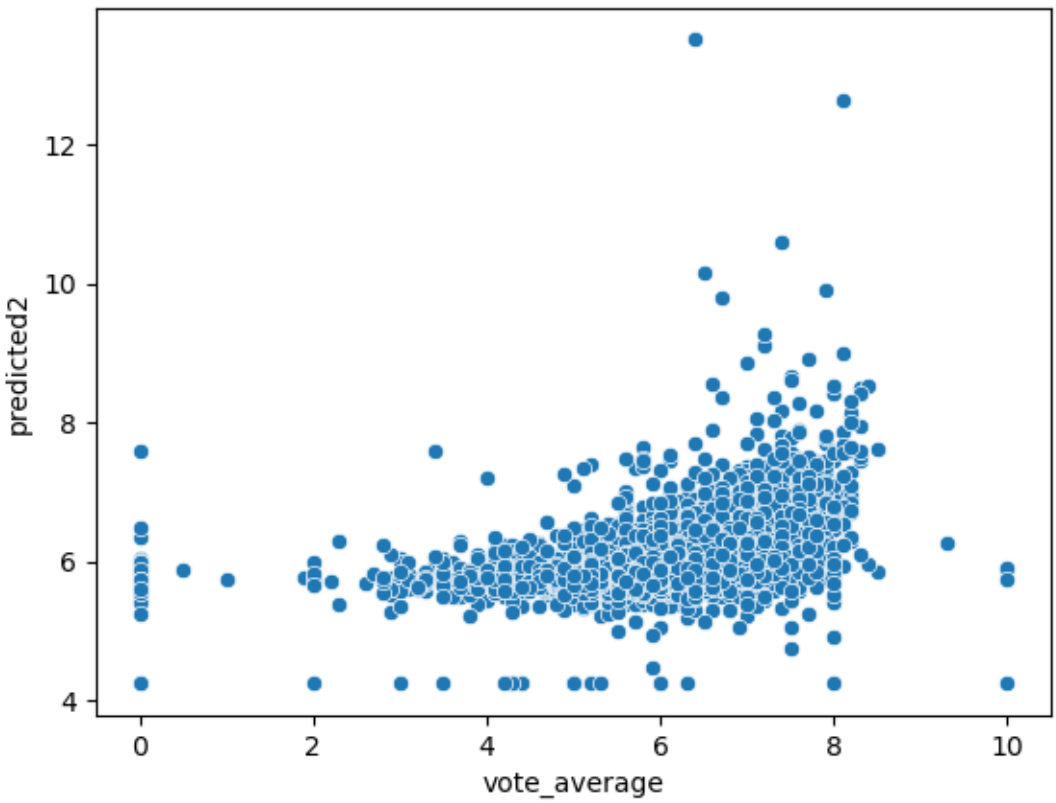
Shows all ranges of correlation between the characteristics that make up a movie. We can see extreme levels of correlation between vote_avg(Predicted1) and revenue.



OLS Multiple Linear Regression

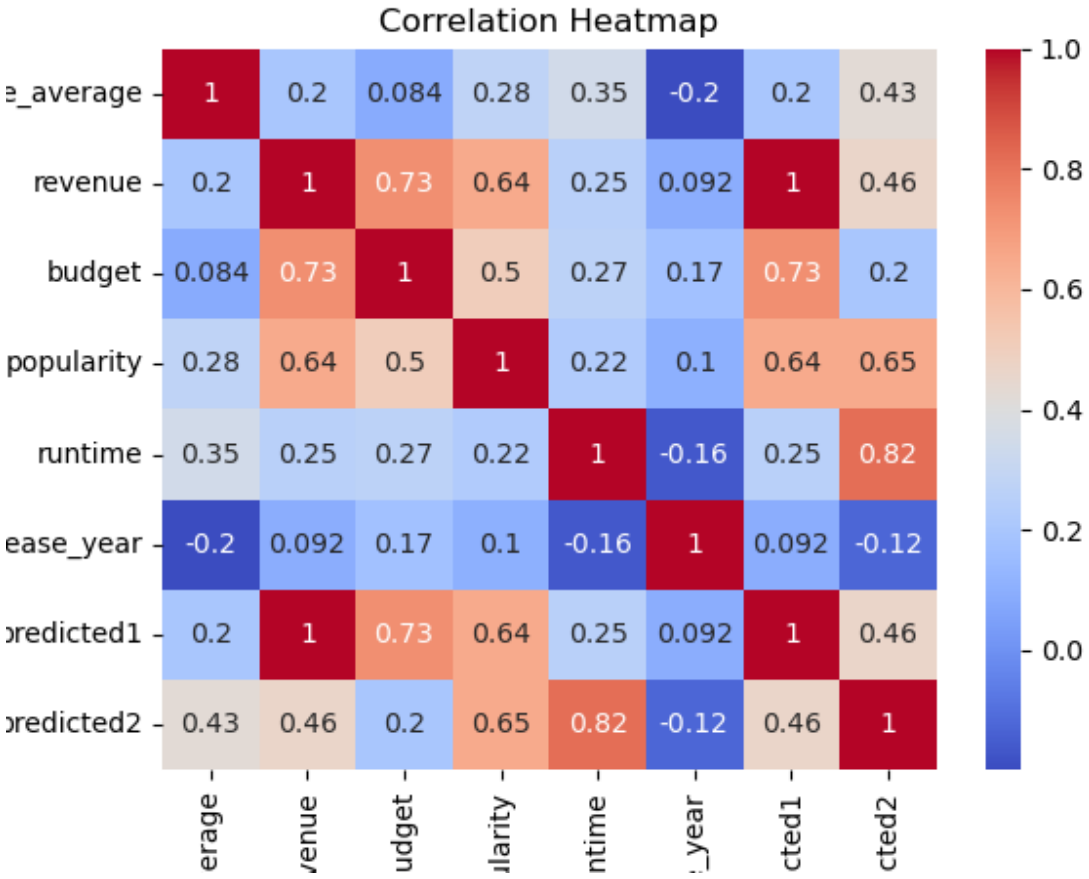
OLS regression analysis where the dependent variable is vote_average, and the independent variables are revenue, budget, popularity, and runtime. The results indicate that budget has a significant negative effect, while revenue, popularity, and runtime have a significant positive effect on the vote_average.

Overall, the second OLS regression model with multiple independent variables has a higher R-squared value, indicating a better fit.



Heat map of Correlations

Shows all ranges of correlation between the characteristics that make up a movie. We can see extreme levels of correlation between vote_avg(Predicted1) and revenue.


Correlation Heatmap
e_average
revenue
budget
popularity
runtime
ease_year
predicted1
predicted2
1 0.2 0.084 0.28 0.35 -0.2 0.2 0.43
0.2 1 0.73 0.64 0.25 0.092 1 0.46
0.084 0.73 1 0.5 0.27 0.17 0.73 0.2
0.28 0.64 0.5 1 0.22 0.1 0.64 0.65
0.35 0.25 0.27 0.22 1 -0.16 0.25 0.82
-0.2 0.092 0.17 0.1 -0.16 1 0.092 -0.12
0.2 1 0.73 0.64 0.25 0.092 1 0.46
0.43 0.46 0.2 0.65 0.82 -0.12 0.46 1
erage
venue
udget
ularity
ntime
_year
cted1
cted2
1.0
0.8
0.6
0.4
0.2
0.0