

**TASK**

**Exploratory Data Analysis on the Movies Data Set**

[](https://www.hyperiondev.com/)

**Introduction**

The Dataset is a collection of data about movies, this includes movie titles, producing companies, genres, popularity, budgets and profits etc. This data can be used to find meaningful information for companies planning on investing or making new movies.

**DATA CLEANING**

* Removing columns that are not needed for processing by dropping them.
* Removing duplicate rows

**MISSING & INCORRECT FORMATED DATA**

* Removing rows with missing data, this included rows where the value = 0
* Some of the columns (budget& revenue) needed to be converted to integers as well which I converted to int64
* The ‘movie release date’ needed to be formatted to just the year as this is easier to use in calculations
* The data for genres, production countries & spoken languages are in JSON, I converted them so that the data is easier to read and calculate.

**DATA STORIES AND VISUALISATIONS**

#### **Which are the 5 most expensive movies?**

* + Pirates of the Caribbean: On Stranger Tides
  + Pirates of the Caribbean: At World's End
  + Avengers: Age of Ultron
  + Superman Returns
  + John Carter

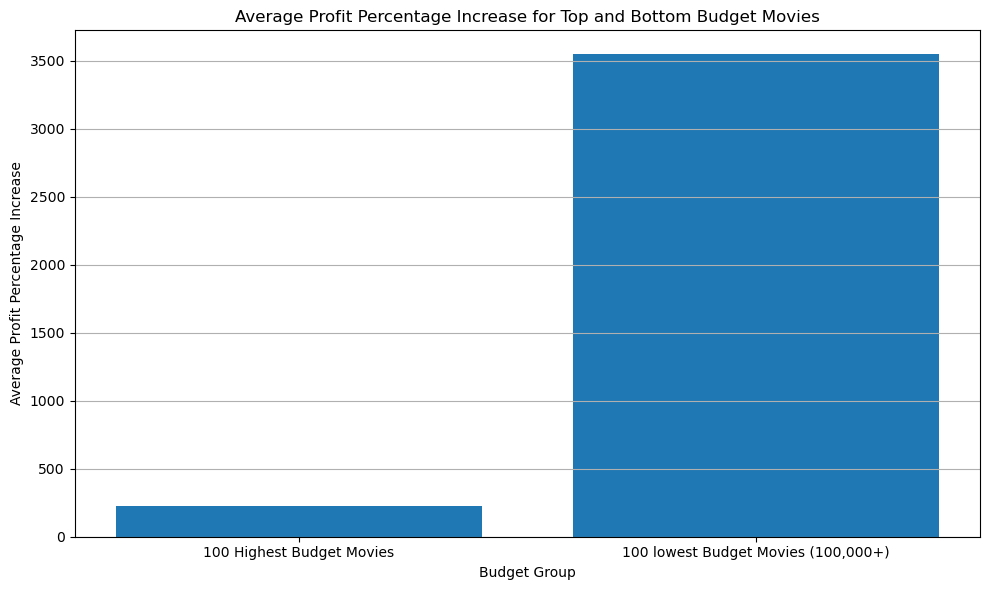
#### **What are the top 5 most profitable movies?**

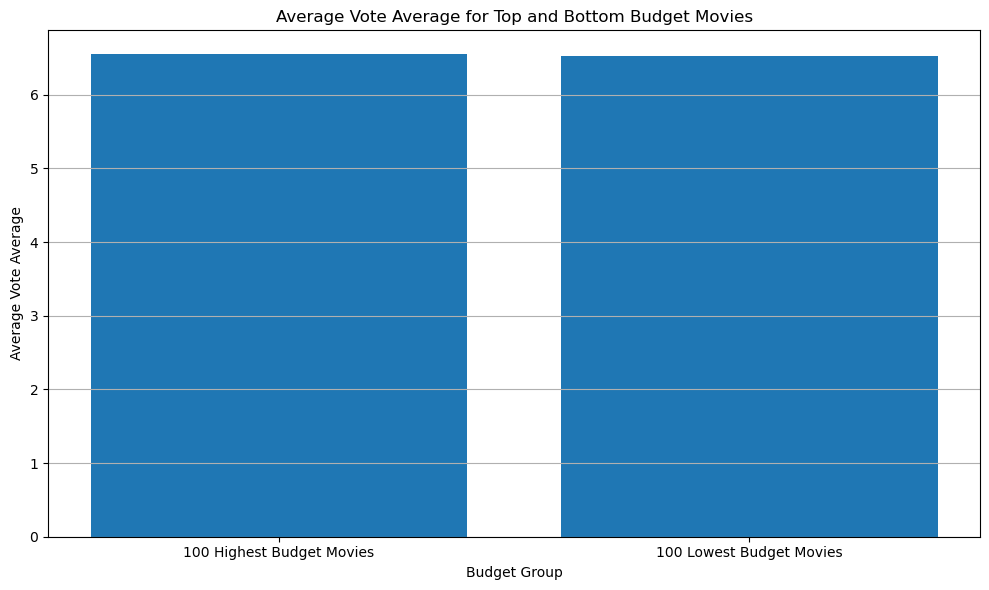
* + Avatar
  + Titanic
  + Jurassic World
  + Furious 7
  + The Avengers

#### **What are the top 5 least profitable movies?**

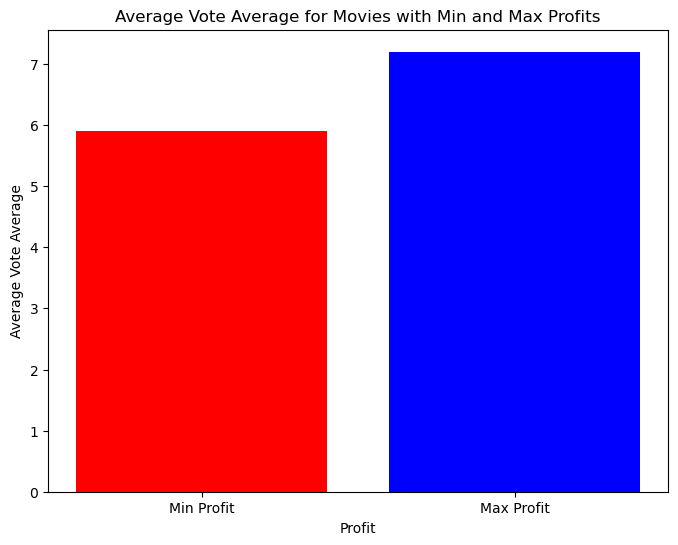
* + The Lone Ranger
  + The Alamo
  + Mars Needs Moms
  + The 13th Warrior
  + The Adventures of Pluto Nash

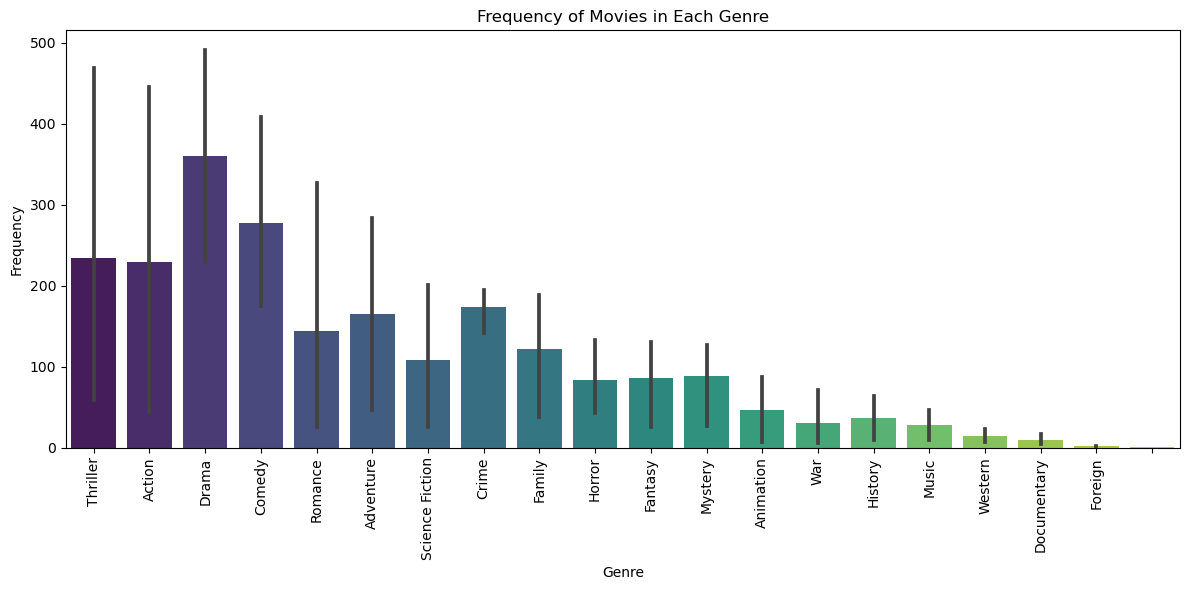
**Visualisations:**

  
**The graph shows that although highest budget movies generate more Revenue the profit percentage increase is actually higher for the lower budget movies**

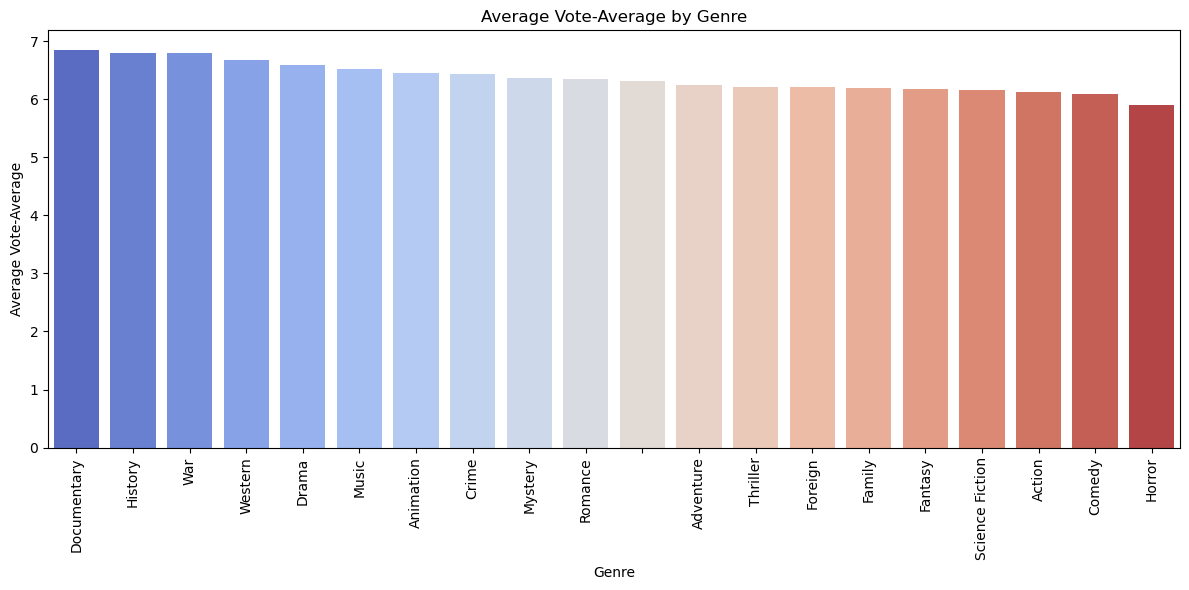
****

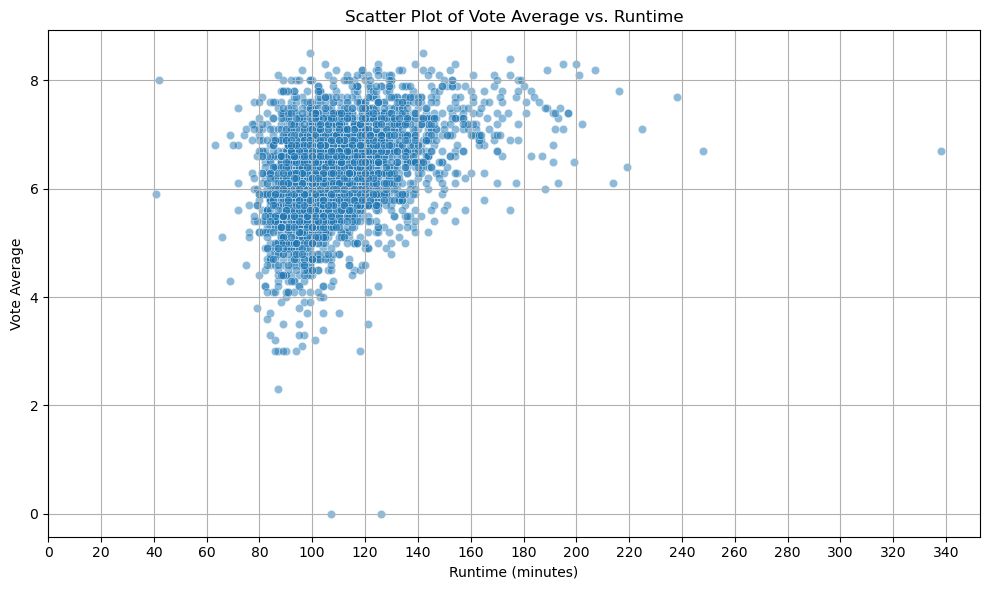
#### **The graph shows that the average vote is about the same value for lower and higher budget movies**

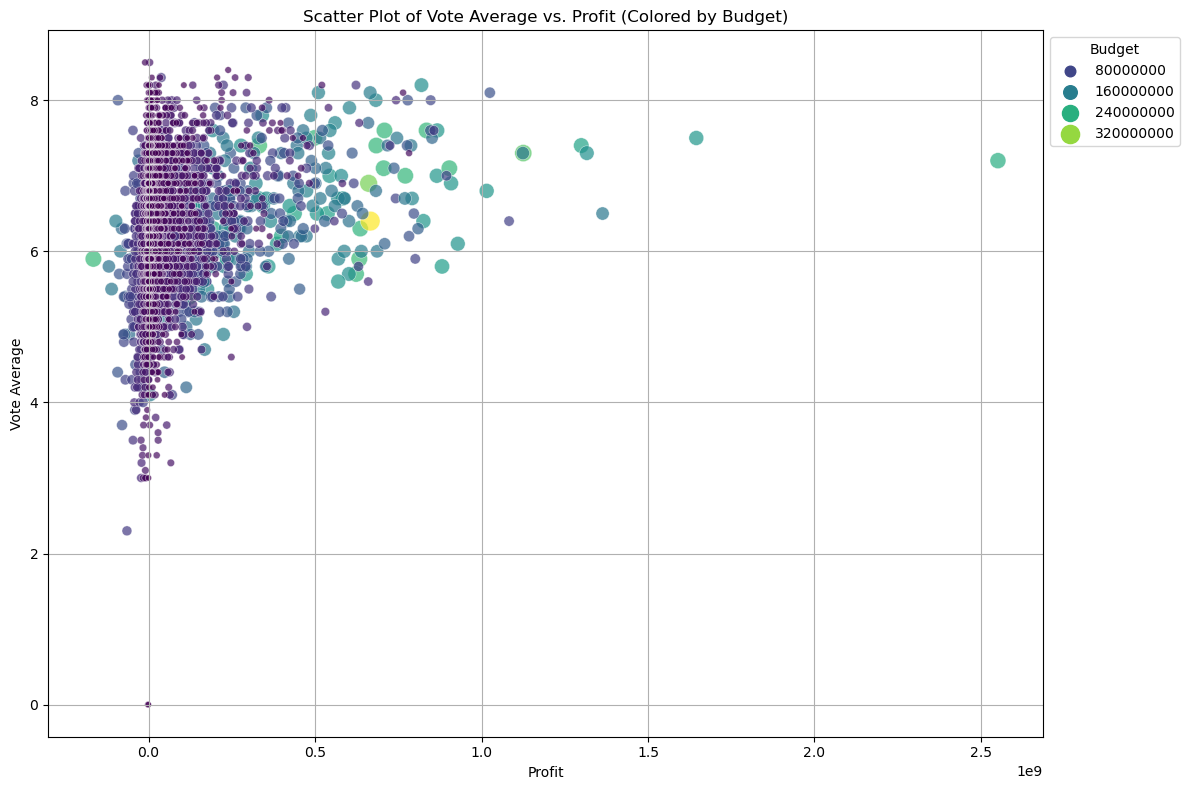
**The graph shows that movies that made less profit had a lower vote average**

****

**The graph shows that drama has the most frequency of movies and foreign has the least, this could be due to the fact that either drama movies are easier to make or are more in demand where as due to language barriers foreign movies may not be as in demand**

**The Graph shows that Documentary, history and war genres have the highest ratings whereas horror, comedy and action have the least. A point to consider would be that comedy and action have a lot more movies created from any of the producers in the world where as documentaries, history and war are aimed and produced with a specific group of people that have interests in them. There may be less reviews but the reviews are by interested parties.**

**The graph shows that as movies increase in runtime, less vote averages are low**[**¶**](http://localhost:8888/notebooks/movies.ipynb#The-graph-shows-that-as-movies-increase-in-runtime,-less-vote-averages-are-low)**most movies are between 80 - 140 minutes, the movies thereafter are much less but maintain higher voting averages**

**The graph shows the higher the budget, the more stable the vote average the more profit is made but less movies are created with higher budgets**

**THIS REPORT WAS WRITTEN BY : RYAN COX**

