Predictive Analysis of Success of Movies from Various Features

Karan Singh, Varun Venkatesh, Waiz Khan

Research Questions for Movies:

1. Which part of the world creates the highest-grossing movies?
    - Through this question, we aim to find out how movies' revenue in the box office is impacted by the location they are in. We initially were going to use a ratio of the gross revenue to the budget to account for inflation throughout the years, but instead decided it was more useful to measure the earnings themselves. This was because the gross is easier to see a difference between various countries, thereby making it easier to draw results from. This is a valuable piece of information because knowing where movies are most likely to make a profit can help target the movie to an audience based in that location. Using the information we just learned regarding Geopandas and plotting, we can visualize this data very easily, making it an interesting research question for us to look into.
    - We found that North America, Australia, Western Europe, and China had the highest-grossing movies.
2. Which part of the world produces movies with the highest rating-to-budget ratio?
    - For this question, we decided to use a ratio of rating-to-budget because it was easier to see differences between various countries with this data. For this question, we are trying to compute where the most cost-effective movies are produced. This analysis can be used to determine how production companies in parts of the world that are lacking in resources can still create high-quality content, and how bigger production companies can reduce the amount of money they spend on making movies.
    - We found that North America had the highest rating-to-budget ratio.
3. How do genres tend to correlate to ratings of their movies? Which genres tend to have the highest number of viewer votes?
    - Based on movie genres, we plan to parse through our data and figure out which genres tend to produce higher IMDb scores and viewer ratings. This is useful as it shows trends of how critics react to specific genres of movies. Knowing what genres end up having the most acclaim among the public audience and not just the critic's reviews are useful because it can help find what kinds of movies are perceived as the best. If we can find

    another data set that goes more in-depth regarding the intricacies of genres, we will utilize those features to further assist with answering these questions.
    - The model generally predicts higher scores for parameters fitting action and drama movies. We found that action, comedy, and drama had the highest user votes.
4. How do aspects of a movie’s production impact the gross revenue?
    - Using machine learning models, we plan to use features from our dataset to predict how well the movie will perform in the box office. We initially wanted to use a ratio of gross revenue to budget as our parameter, but instead decided to use the gross revenue so we could better discern differences between movies. This is an important feature because companies want to continually gain profit from their movies and knowing what features impact the movie’s revenue can help them achieve this. If possible, we will continue to search for data sets that contain information that could extend what we are currently looking for by providing even more features that could help with our predictions.
    - We found that budget, critic’s rating, viewer votes, year, and runtime impacted the gross revenue most.
5. Over the years, what are the trends of genres based on viewer ratings?
    - Based on critics and general public ratings, we will find out what types of movies performed the best over the years. We can either do this by year or by every few years and will decide this after looking into our data set more closely. This is useful as companies can make more informed decisions on what type of movies they want to produce. Similarly to the previous question, finding data sets that elaborate more on the specifics of genres can help us find more detailed answers to this question, so that will be one thing we will look further into. We initially were planning on only graphing the highest and lowest grossing genres by year, but felt that it was more useful to keep every genre for a more holistic viewpoint.
    - We found that over the years, action took over as number one around 2000 and comedy and drama have been consistently high over the years. Most other genres have been relatively low in terms of viewer votes.

Motivation and background:

All of our questions are targeted toward movie producers and marketing companies. We want to make it so that they create movies that result in the biggest benefit for both themselves and their target audiences. Therefore, we try to maximize those features that impact their profit and reviews. We believe that being able to find the answers to these questions will help these companies become more efficient with their time by focusing on those movies that will be most relevant, instead of wasting their time with movies that will not be successful.

Dataset:

https://www.kaggle.com/danielgrijalvas/movies
This dataset contains various features including movie budget, production company, country of origin, genre, revenue, release date, run time, reviews (from IMDb and viewer votes), starring actor/actress, and main writer. It covers 6820 movies from 1986 to 2016 (220 movies per year) and the data is scraped from IMDb by the person who posted the dataset.

https://www.kaggle.com/alenavorushilova/world-national-and-real-gdp-annualyquaterly
This dataset contains every country with its GDP from 2005 to 2019. This is a complementary data set that we will use in our first question to show alongside the highest-grossing to budget ratios based on countries.

GeoPandas World Map Dataset
There is no link to this dataset because it is loaded directly from the geopandas library. It contains various features that are used to plot a map of the world by country. The most important column is the shape column which is used for plotting.

Methodology/Analysis:

For our first research question, we will focus on data visualization. By finding how the various countries or regions fare in terms of how much money they make with each movie, we can use GeoPandas to draw this out. By coloring each country based on how much money they make through all of the movies that are produced there throughout the years, we can provide a detailed visualization that effectively represents this. One secondary visualization that would be useful to pair with this could be a graph of our regions and how much GDP they have had each year. Seeing the resulting data can help us understand what the common factors between the countries with highest-grossing movies are, helping us understand more about what results in revenue for these areas.

For our second research question, we will focus on data visualization again. Similarly to the first research question, we can use GeoPandas to represent what countries or regions have the highest rating-to-budget ratio. By coloring each country based on this value, we can provide a detailed visualization that effectively represents this. We initially thought an interesting component to analyze alongside this could be creating a second visualization that plots how this number has changed for each country or region every year. We decided to change this and compare it to each country's GDP because this would be more related to the budget of each country. Seeing the resulting data can help us understand what proportion of low-budget movies end up receiving positive reviews, and it can help justify if spending lots of money on a movie is worth it or not. The data visualization will also highlight which countries/regions have the highest rating to budget ratio.

For our third research question, we will create a linear regression model to predict what the critics' ratings will be for certain genres. We will pair this with data visualization of how genres are related to viewer votes through a bar graph of genre types vs average viewer ratings. The first step we will take is to cut out the features that are not highly correlated to the variable we want to predict (the ratings by critics) followed by cutting out features that are too highly correlated with each other. Using these features, we can predict the rating for a certain genre. The data visualization would use two features, the average viewer rating per genre on the y-axis and genre on the x-axis. If our model predicts high ratings and low ratings for certain genres, we can conclude that these genres are generally rated the highest by critics. Then we can compare this with our other data visualization to see how critic's ratings and viewer ratings differ or agree.

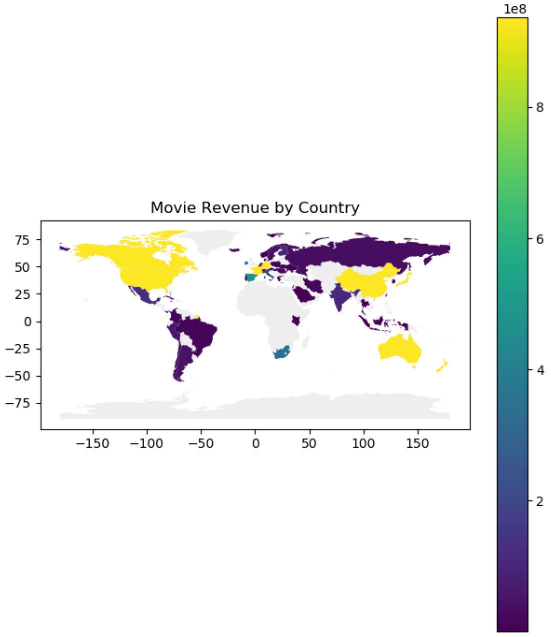
For our fourth research question, we will use a linear regression algorithm by training a decision tree regressor to predict future movies' box office based on features in our dataset. The label would be the 'gross revenue' and the features would be columns in our data that are not too distinct to avoid overfitting. We would use the same technique as the previous model when it comes to removing un-correlated and highly correlated features to select which features we will use in prediction. For instance, we would first cut out the movie name, as it is too distinct, and then remove features that have too high of a correlation with other predicting features. This is an interesting model as it allows companies to understand which factors of a movie will allow them to get maximum profit. If we see that certain movies' of a certain genre or by a certain company will not be grossing much money, then we can see that there are deciding factors for how a movie will do in the box office. If we cannot see consistent predictions, then we will not know if there are common features among successful box office movies.

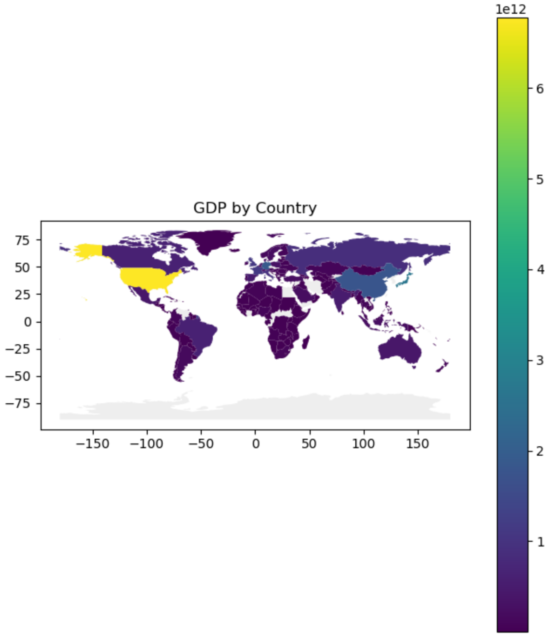
For our fifth research question, we will create a data visualization to help show the trends of many different genres on a year-by-year basis from 1986-2016. First, we will filter our data set to group user votes based on genre each year. We will then plot this on a line chart with the years on the x-axis, viewer ratings on the y-axis, and the hue of each line based on genre. If we see genres with large increases/decreases in user vote throughout the years, it will lead us to believe that the said genre is one of the most popular/unpopular genres among viewers and therefore does/does not bring in a lot of money.

Results:

1. Which part of the world creates the highest-grossing movies relative to budget?

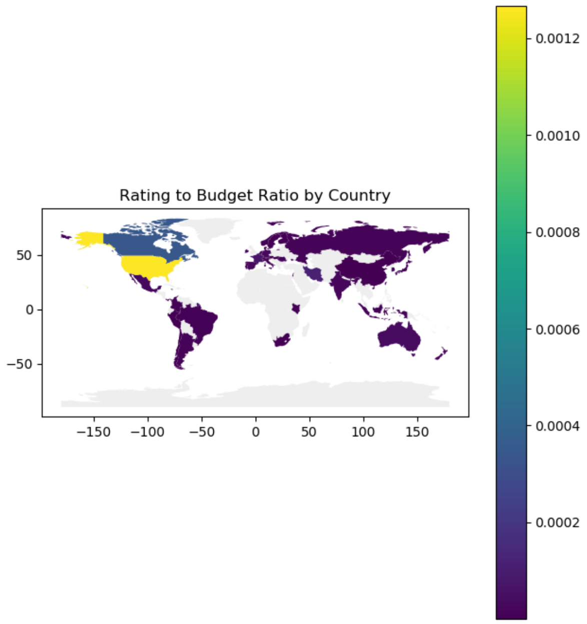
   From our data visualizations, we can see that GDP correlates to high grossing movies a large portion of the time. The three biggest exceptions are Canada, Australia, and South Africa. Canada and Australia both have GDPs on the lower end of the scale, but have total movie grossings at the very top of the list of countries. South Africa has a very low GDP, but has above average movie grossings. We believe this is due to the fact that many people film in these locations for movies because it is cheaper than other countries, and they end up releasing movies in these areas because of this. This overall shows us that GDP and movie grossings are heavily linked together and filming and releasing movies in high-GDP countries will result in a bigger profit. The most logical explanation for this is that the population in high GDP countries has a higher level of affluence and can afford to spend more money on movies as opposed to low-GDP countries.

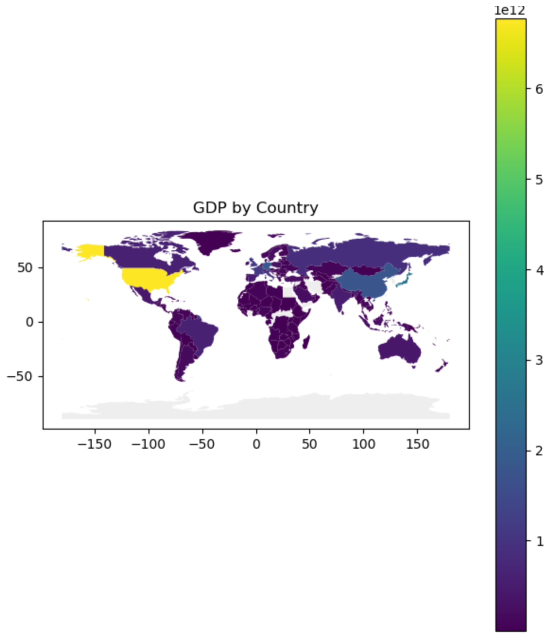




2. Which part of the world produces movies with the highest rating-to-budget ratio?

   From our data visualizations, we can see that the region that produces movies with the highest rating to budget ratio is North America. Some countries are not colored in as their budgets are 0 in the data set and we needed to filter them out. By comparing both of these graphs we see a correlation between high GDP and high budget to rating ratios, as the United States is the top of both. This shows that countries with higher GDPs tend to produce higher quality movies, perhaps because these nations can allocate more overall spending towards movies instead of essentials, like food or water. Canada is surprisingly higher in rating to budget ratio, but further analysis shows that this is because the budgets are much lower, perhaps because filming movies in Canada is cheaper.


Rating to Budget Ratio by Country


GDP by Country

3. How do genres tend to correlate to ratings of their movies? Which genres tend to have the highest number of viewer votes?

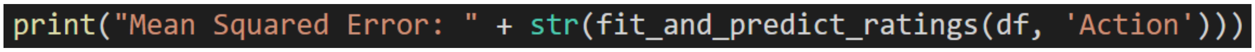
Our model is able to predict score with the mean square error usually being about 1.3. With score being on a scale of 0-10, this gives us an average mean square error of about .13. Not only is this a reasonable score, but the R^2 value also indicates that this model accurately predicts a large portion of the data.The visualization shows us that Action, Comedy, and Drama are the most popular genres when taking the average of the user votes per genre. The model generally predicts higher scores for parameters fitting action and drama movies, so we can see that a high number of viewer votes is associated with a higher critic rating for the most popular genres. When it comes to less popular genres, the viewer votes drop significantly, but the critic ratings, while being lower, do not drop as far as the viewer votes do. We think these are the most popular genres because past history of movie genres takes precedent when it has high gross revenue. Therefore, companies want to keep making movies that have a large audience, so these most popular genres continue to be the most popular.

Test cases for our model:
Input Features(in order):

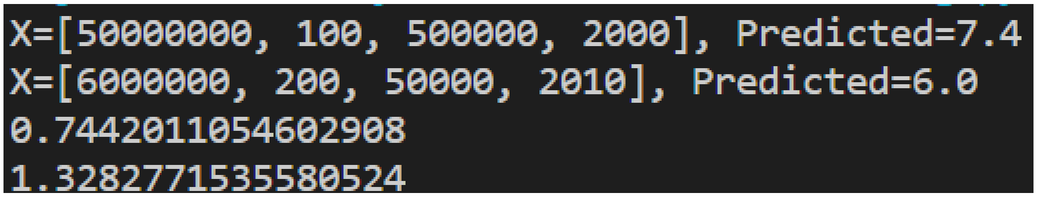
- Budget in dollars, runtime in minutes, number of votes, year of release

Output:

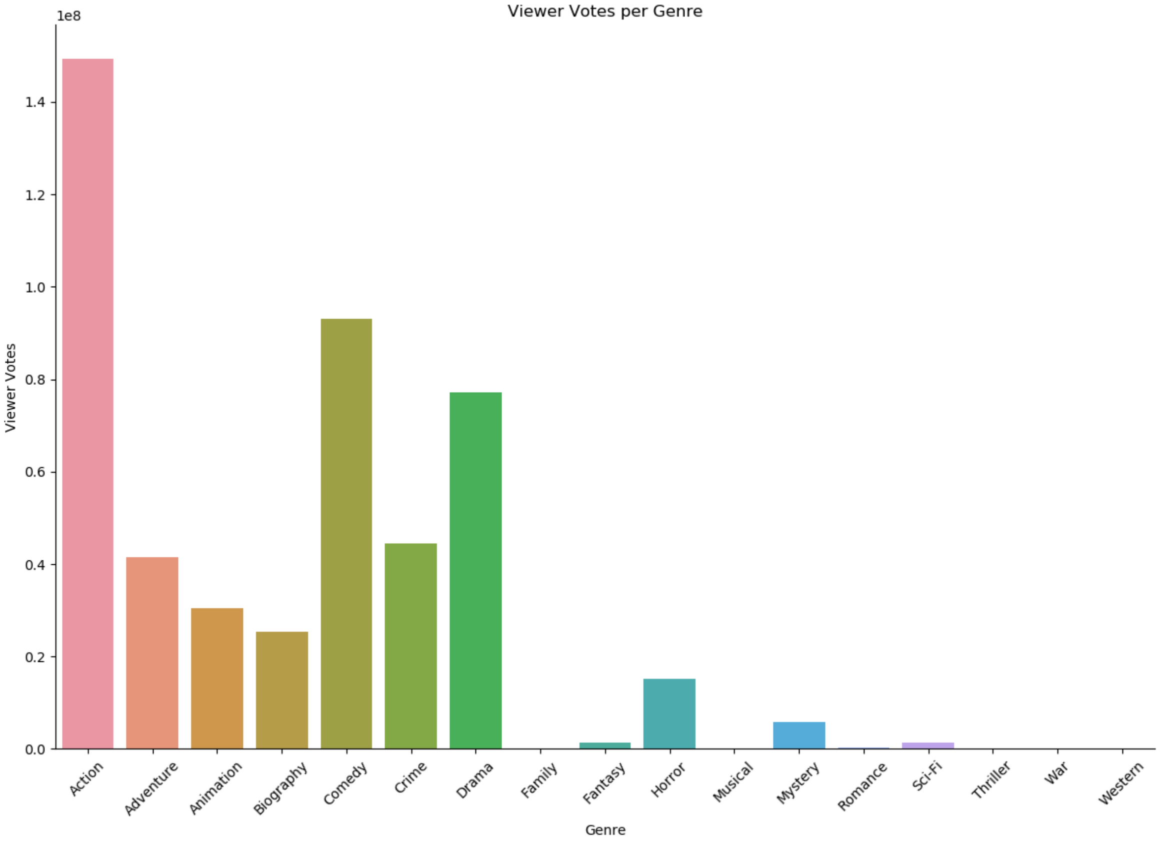
- Predicted IMDb rating of movie



```
print("Mean Squared Error: " + str(fit_and_predict_ratings(df, 'Action')))
```



```
X=[50000000, 100, 500000, 2000], Predicted=7.4
X=[6000000, 200, 50000, 2010], Predicted=6.0
0.7442011054602908
1.3282771535580524
```



4. How do aspects of a movie's production impact the gross revenue?

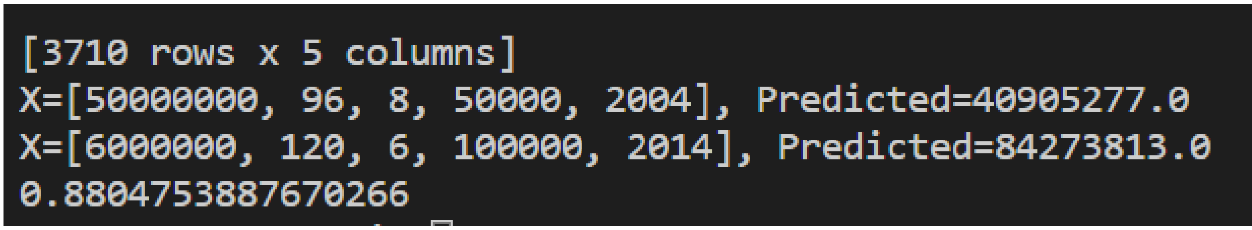
From our machine learning model's feature selection, five main features stood out as being highly correlated with gross revenue. These were the movie's budget, runtime, critics' rating, viewer votes, and year. Movie budgets seem highly correlated to gross revenue because a movie with a higher budget will be able to spend more time on production quality and large-scale filming, thereby helping it reach a wider audience and make more money. The runtime makes sense because people don't want to spend money on a movie too short or too long because it is not worth paying for it. It also makes sense that the critics' ratings and viewer votes are on this list because highly rated movies are going to be most popular. Finally, it seems the year is very highly correlated because over time more people could afford to watch movies in theaters and inflation has caused the price of tickets to go up over time. Therefore, it seems all of these features are equally important when it comes to the gross revenue of a movie.The R^2 value, which represents how much of the predicted data is from the features and not random chance, indicates that this is a model that accounts for almost all of the data, making it an accurate model. Additionally, the values predicted are exactly what we expected based on the parameters passed in during testing.

Test cases for our model:
Input Features(in order):

- Budget in dollars, runtime in minutes, critic rating, number of votes, year of release

Output:

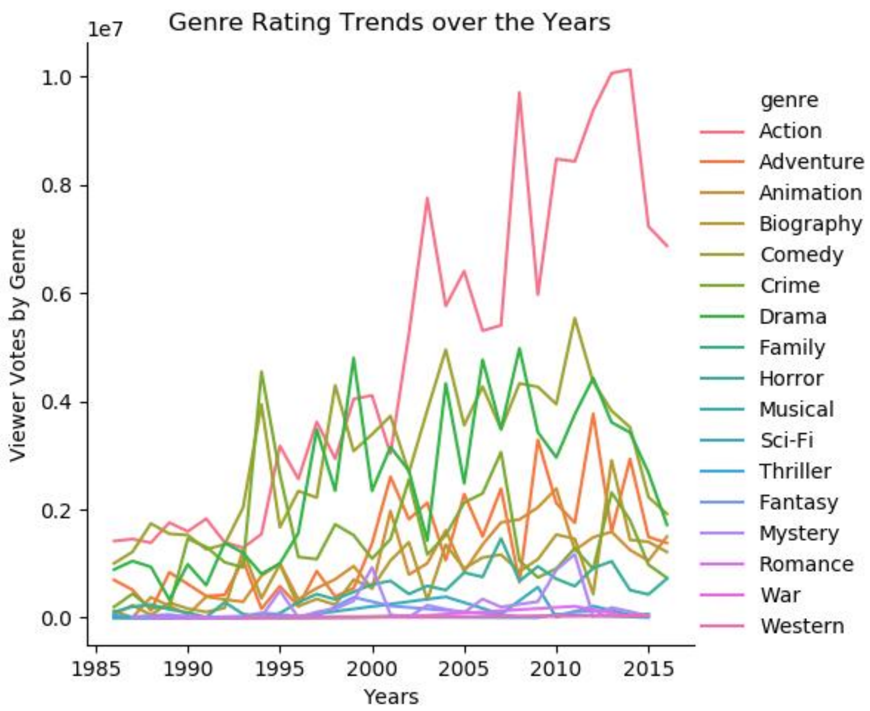
- Predicted gross revenue of movie



```
[3710 rows x 5 columns]
X=[50000000, 96, 8, 50000, 2004], Predicted=40905277.0
X=[6000000, 120, 6, 100000, 2014], Predicted=84273813.0
0.8804753887670266
```

5. Over the years, what are the trends of the highest-grossing genres based on viewer ratings?

   We can see that after 2000, action movies lap all other movies by a significant margin when it comes to viewer ratings. Starting off stronger, comedy and drama movies fall behind action but are consistently high up from 1985 till 2019. Adventure and animation are behind these two and all other movie genres have a relatively low amount of viewer votes. This visualization helps show us that the most common genres have rarely changed over the years with the only noticeable difference being that action went from 3rd in 1985-2000 to having a massive jump to number one overall. The most successful movies among viewers seem to be action movies overall. One interesting point to note is that almost all of the most popular genres have a downward trend at the very end of the graph (from around 2015 onwards). This is made up for with other less popular genres having a slight incline in their votes, perhaps showing that more rare genres are gaining popularity in recent times. Additionally, the years after 2015 have less total information in the dataset so it makes sense that the total number of votes is lower than previous years.



Reproducing Results:

In order to reproduce these results, the dataset can be downloaded from the links in the Dataset section of the report. The first dataset, containing the movie data should be saved as “movies.csv”, and the second dataset containing the GDP data should be saved as “gdp_csv.csv”. Both of these data sets should be saved in the same directory as the python files. The last dataset containing the shapes of the countries is loaded directly from the Geopandas library and does not require any external downloading.

1. Research Question 1 (Data Visualization): For this question, run the file in the terminal. It will produce an image output that can be saved.
    - The output contains two plots: Movie revenue by country and GDP by country. The legend can be used to deduce which countries have the highest movie revenues and how those metrics compare to that country’s GDP. Some countries appear in light gray on the maps as there is insufficient data for those countries.
2. Research Question 2 (Data Visualization): For this question, run the file in the terminal. It will produce an image output that can be saved.
    - The output contains two plots: Budget to Rating ratio by country, and the same GDP by country plot as the last question. The legend can be used to deduce which countries produce the most cost effective movies (eg., highest rating for the lowest budget), and how that matric relates to the

country’s GDP. Some countries appear in light gray on the maps as there is insufficient data for those countries.

3. Research Question 3 (Machine Learning): For this question, run the file in the terminal. It will produce output in the terminal and produce an image that can be saved.
    - The output produced will first print the data frame of the selected features to be analyzed by the model, and then the test cases along with the prediction that the algorithm produced. Additionally, it will print the R^2 value and the Mean Squared Error
    - The image output contains one plot of all genres with the average number of viewer votes per each genre as a bar. The units on the columns are scaled to 1 x 10^{8}.
4. Research Question 4 (Machine Learning): For this question, run the file in the terminal. It will produce an output in the terminal.
    - The output produced will print the result of a machine learning model with sample test cases of it running, along with the R^2 value.
5. Research Question 5 (Data Visualization): For this question, run the file in the terminal. It will produce an image output that can be saved.
    - The output produced contains one line plot with all genres’ total viewer votes over time (1986-2019). The legend can be used to deduce which line represents which genre. The units on the columns are scaled to 1 x 10^{7}.