Project 2: Investigating movie dataset

The dataset I analyzed is tmdb-movie dataset

Cleaning data:

I used df.info() checked the null value. Because I only need column genres, original title, release time, revenue, and popularity. Because popularity is float and revenue is integer. I don't need to worry about the data type. Also all the variables listed above, they don't have null value, except for the variable genre. To answer the questions I pose, genre plays an important role. So I decide to drop out all rows that has no genres. I used the df[]dropna(axis=0, how='any') function. Because genre is the only variable that has some null value, so the function will drop all null value located row.

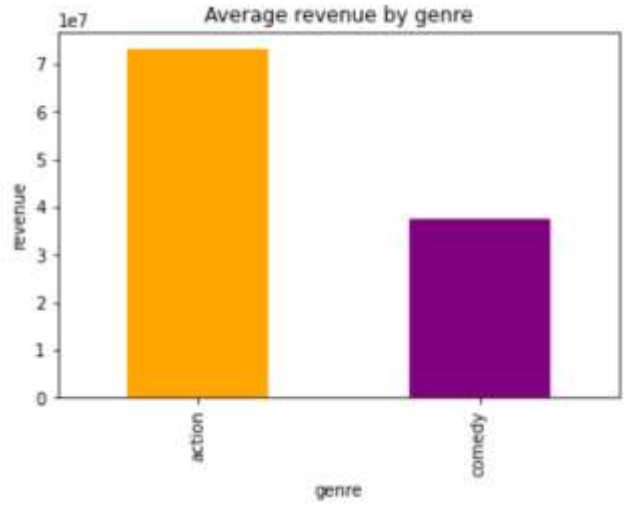
Two questions:

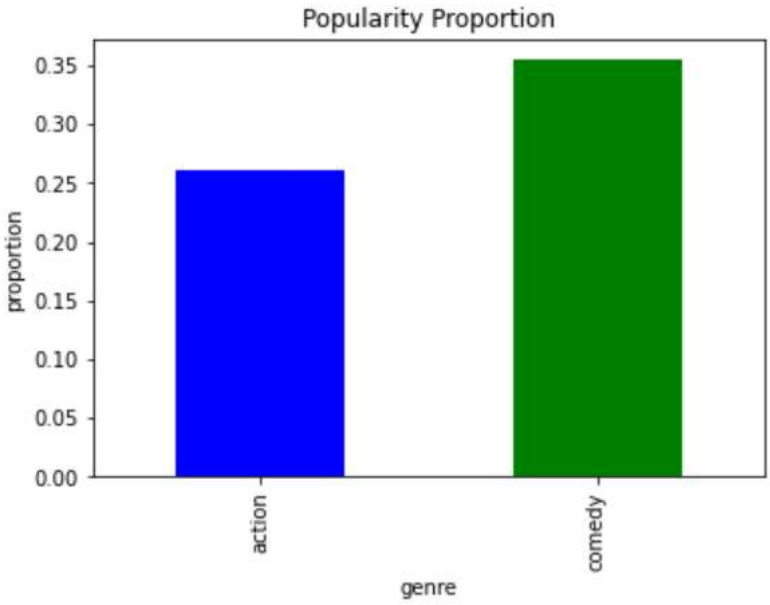
1. Which one receives higher revenue, comedy or action?
2. Which one are more likely to have higher popularity, comedy or action?

To answer my questions, I cleaned the data first, drop any null value related to genre and save it as a new dataset. Because my question is related to comedy and action, I only need collect all the data related to comedy and action. I use code "df_comedy=new_data[new_data['genres'].str.contains("Comedy")]" and " df_action=new_data[new_data['genres'].str.contains("Action")]" to generate 2 new data sets. In data set containing all action type of movie, I created a new column named genre and values are generated by using array function that using repeat function. "genre_action=np.repeat("action", 2365)" But these two functions above gives me all the row that contain genre "comedy" and "action", for later visualization, it is hard to compare. So I need to combine this two datasets and saved as "movie_df".
After combination of the dataset, I can analyze and answer the questions I poses. For the first question, I need to compare the revenue of comedy and action and compare it.

The second question is about which genre is more likely to have higher popularity, I need see the proportion of comedy action getting higher rating, here I use above equal to median popularity as higher rating. Code: popularity_high=movie_df.query('popularity >= 0.383921') later, I use groupby function to see how many action and comedy movie above median, for data accuracy, i use proportion instead count. I divide the comedy that having popularity >=median by the sum of the number of genre above median popularity and action that having popularity >=median by all which is greater than 0.384958.

visualization:





Conclusion:

Action movie receive more revenue than comedy movie

Comedy is more likely to receive higher popularity compared to action movie.