## ***TASK 1***

![Text

Description automatically generated]()

Line 4 imports the csv file and stores the data in a variable called mydata. Line 7, 8 selects only the rows mentioned in the pdf document from the dataset, while line 16 gets rid of all missing variable values from the dataset. Finally, after all the cleanup, the dataset mydata is left with 3661 rows and 11 columns.

---------------------------------------------------------------------------------------------------------------------------

Part a.

![Chart, box and whisker chart

Description automatically generated]()

**Figure1:** Hidden.Gem.Score vs Runtime Boxplot



The above lines of were used to generate a boxplot of hidden gem scores against runtime. As can be seen from the boxplot (figure 1), the median is approximately equal – with > 2 hour movies having slightly higher hidden gem score than others - for movies of all 4 different runtimes. However, the hidden gem score variance for these movies are quite high except 30-60 mins runtime (mainly because there are very few movies of this runtime category). Thus, it can be safely stated that runtime and hidden gem score do not have a strong association.

![Text

Description automatically generated]()

Line 28 in the above code was firstly used to create a column named “FirstLanguage” which only lists the first language in the column “Languages” for each row. For example, if the column entry for “Languages” is “Spanish, German” – the FirstLanguage column entry is “Spanish.” This was done under the assumption that the language listed first was the main language the movie was available in – and to simplify the process of finding the association between Languages and Hidden Gem score.

Line 30-32 generates the boxplot of figure 2. To back this boxplot, lines 34-37 generate the leftmost table of figure 3 – which shows the mean hidden gem score based on the ‘FirstLanguage’ of each movie in a descending order. As can be seen, Tibetan movies had the highest mean at 8.9 while Chinese movies had the lowest mean at 1.7 – a similar result can be seen from the median in the boxplot of figure 2.

However, since the number of movies in each ‘FirstLanguage’ is not equal, the middle table of figure 3 is included to show what the frequency of each ‘FirstLanguage’ is in the dataset using lines 39-41 in the code.

Next, the mean Hidden Gem score is found out to be approximately 3.55 using line 45 for the entire dataset – and then the rightmost table of figure 3 is generated to show how many movies of each first language are above average and how many are below average by creating a ‘cutoff’ point at 3.55 (using lines 47-50 of the code).

As we can see - Arabic, French, Japanese, etc. movies are more likely to receive a higher than average Hidden Gem Score, while English and Thai movies are more likely to receive a lower than average Hidden Gem Score. This data is standardized to a proportion using lines 53-56 of the code and shown in figure 4. Thus, it would NOT be wrong to claim that movies of particular ‘FirstLanguage’ like Arabic receive higher Hidden Gem Score than movies of other ‘Firstlanguage’ like English. This claim is nevertheless partially weakened due to the fact that English has an incredibly high number of entries in the column of ‘FirstLanguage’ compared to other languages in the column.

![Text

Description automatically generated]()

Furthermore, to evaluate an association between the number of Languages a movie is offered in (for example, if a movie is offered in “English, Spanish” then Num\_Languages is 2) and the Hidden Gem Score of the movie, lines 61-63 were written in the above chunk of code.

Finally, lines 69-71 were used to illustrate the relationship between these two attributes – which is shown in figure 5 below. Line 73 was written to find the mean for each value of Num\_Languages, which is also included in figure 5.

As can be seen, both the mean and median for each of these are approximately equal, with a comparatively low mean for ‘10’ and comparatively high median for ‘9’. However, these are not sufficient to conclude that there is a strong association between the number of languages a movie is offered in and the hidden gem score.

Chart

Description automatically generated

**Figure 2:** FirstLanguage vs Hidden Gem Score Boxplot

![Timeline

Description automatically generated with low confidence]() ![A picture containing text

Description automatically generated]() ![A picture containing text, chain, metalware

Description automatically generated]()

**Figure 3 [left to right] :** mean by FirstLanguage -> frequency of FirstLanguage -> Score Range of FirstLanguage

Chart, bar chart

Description automatically generated

**Figure 4:** Barplot for numberOfMovies receiving ‘lower than average’ and ‘higher than average’ hidden gem score for each ‘FirstLanguage’

![Chart, box and whisker chart

Description automatically generated]() ![Text

Description automatically generated]()

**Figure 5 [left to right] :** Boxplot of Hidden Gem Score against Num\_Languages -> Mean for the number of languages movies are offered in

Part b.

![Text

Description automatically generated with medium confidence]()

The lines of code from 80-84 were written to generate 3 different scatterplots of Hidden Gem Score against each of IMDb, Rotten Tomatoes, and Metacritic respectively. The geom\_smooth function was used to show the overall trend and correlation between the scores.

Next, lines 86-89 were written to combine them together in a single diagram as shown below. We can see that the correlation between Hidden Gem Score and each of the 3 scores is quite strong – given we overlook some outliers. This can be seen from the nature of the trendline (blue, red, and green respectively). Thus, it is safe to say that Hidden Gem Score has a direct correlation with each of the 3 scores.

![Graphical user interface

Description automatically generated with medium confidence]()

Part c.

![Text

Description automatically generated]()

Lines 94-105 of the code were written to find the year when each movie was released from its release date, and then categorize the years into 4 segments: 1920-1945, 1945-1970, 1970-1995, 1995-2020

Then, lines 108-112 were used to generate the table shown below which shows the mean Hidden Gem Score for each Release year category based on the runtime of the movies. As we can see, if we consider movies longer than 2 hours to be ‘long’ movies, the mean Hidden Gem Score has fallen from 4.500 to 3.617. The story is the same for ‘moderately long’ movies between 1-2 hours whose mean Hidden Gem score has fallen from 4.743 to 3.540.

If the hidden gem score is a representation of how acceptable movies are to users, then the theory that longer movies have gained more acceptance over time is incorrect.

However, if the hidden gem score is a representation of how well the movies are ‘hidden’, it would mean more users are now watching these longer movies causing the movies to NOT remain ‘hidden’ anymore and resulting in them having a lower hidden gem score. In that case, the theory proves to be correct.

![A picture containing text, scoreboard, plaque, receipt

Description automatically generated]()