**3 Questions Data Case Analysis [~1,000 words in word (.docx) and python script (.ipynb)]**

With the proliferation of the Web 2.0 technologies product, ‘experiences’ can be conveniently shared on the internet which can reduce information uncertainty for decision making. Online comments reflect others’ product experience and therefore help in making a purchase decision. Information gathered via word-of-mouth (WOM) significantly influences product evaluations and purchase decisions. eWOM (online WOM) has shifted the power to the consumer. A staggering 92 percent of consumers around the world say they trust earned media, such as recommendations from friends and family, above all other forms of advertising. Therefore, sometimes it does not matter how effective your campaign is because a bad review through the internet can destroy it quickly. Hence, it is important to understand the experiences of the consumer, both negative as well as positive to be able to capture the complete picture, adopt the right marketing strategy, and improve their experience in the future.

A fundamental objective of the motion picture industry has been to understand the overall experience of moviegoers/ spectators/audience and consequently derive better financial remuneration from its theatrical exhibition. Classical film theorists conceived the spectator as a passive participant in receiving the film as a mediated message. However, there is a substantial transformation in understanding the moviegoers’ satisfaction from mere spectators watching films to “experiencing” the film. Considering that the ability of a film to provide a memorable experience colored with emotions, affects and fantasies is dependent on the pleasures the movie offers, the desires it elicits and most importantly the motivations behind the viewer’s watching of the movie, the key to understand this experience is to understand the nature of the film spectator’s response to the film. Consumers including moviegoers want experiences which provide a novel and creative escape from everyday life. The film provides such opportunity through the vicarious experiences it provides, thus making an indelible impression on their memories and intersecting with their lives in significant ways.

Suppose you are a movie producer and want to learn about what consumers have been sharing online and how it will influence box office revenues. A data set from IMDB has been collected for this purpose. You are only required to analyse and interpret the data provided. This will test your knowledge and ability in analysing quantitative data by using a variety of methods learned. You will be required to use various machine learning techniques to address the following questions by analysing the data provided.

**You may need to clean up the data before actual analysis.**

**Q1: Use topic modelling to figure out what users are talking about on IMDB. Are the topics for action the same as the ones for comedy?**

**Interpretation:**

**Action topics:**

Based on the results of the topic modelling analysis, the two most prominent topics in the action genre were identified. Topic 1 comprises words such as 'movie', 'film', 'good', 'like', and 'action', while Topic 2 includes terms like 'bond', 'film', 'movie', 'action', and 'like'. From these findings, it can be inferred that users on IMDB were discussing action movies and expressing positive sentiments in their reviews by employing terms like "good" and "like". Consequently, it can be concluded that the action movies were well-received by the audience based on the available information.

**Comedy topic:**

Based on the outcomes of the topic modeling analysis, the two dominant topics within the Comedy genre have been identified. Topic 1 consists of words such as 'movie', 'funny', 'film', 'good', and 'really', while Topic 2 encompasses terms like 'movie', 'parody', 'film', 'scary', and 'scream'. Based on this information, it can be inferred that users on IMDB were actively discussing Comedy movies and expressing positive sentiments in their reviews, indicated by the presence of terms such as "funny" and "parody". Consequently, it can be concluded that Comedy movies were favorably received by the audience, as suggested by the available information.

* **what test(s) was (were) used**

In order to enhance the quality and usability of our dataset, we conducted essential pre-processing tasks, including lowercasing, tokenization, and stop word removal. These steps were implemented to standardize the text and eliminate irrelevant words that may not contribute significantly to the analysis. Following the pre-processing stage, we proceeded with data vectorization, a technique used to represent textual data in a numerical format suitable for further analysis.

To extract meaningful insights from the action and comedy genres, we employed topic modelling techniques. Specifically, we employed topic modelling algorithms to identify and categorize the underlying themes or topics within these genres. In this case, our analysis revealed a total of five distinct topics within the action and comedy genres.

Subsequently, we explored the topic distribution within the action and comedy genres. This involved quantifying the prevalence and significance of each topic across the respective genres, providing valuable information on the relative importance and representation of different themes within these categories.

By conducting these rigorous data processing, analysis, and topic modelling procedures, we aimed to ensure a professional and robust approach to uncovering the key topics within the action and comedy genres.

* **what information from the survey e.g., variable(s) was used?** 
  + In our analysis, we leveraged the review column to extract pertinent information regarding the reviews, while utilizing the genre column to identify and categorize action and comedy movies of interest. This approach enabled us to obtain valuable insights and facilitate a more comprehensive examination of the dataset. By employing these specific columns, we aimed to ensure a professional and efficient methodology in our research process.
* **What are the results? Explain & Interpret**
  + We can see that the distributions are different. Each topic is represented by a probability value, and we can observe that the probabilities for each topic vary between action and comedy genres.
  + This suggests that the topics being discussed in reviews for action movies are not the same as the topics being discussed in reviews for comedy movies. The difference in topic distributions indicates that users are likely discussing different aspects, themes, or elements when reviewing action and comedy movies.

**Q2: Use sentiment analysis to estimate ratings of the top 2 topics for action and comedy movies and interpret the results. (You can apply sentiment analysis on the tweets highly relevant to each of the topics separately. For example, the top 30% of the tweets which are highly relevant to a topic.)**

* **what test(s) was (were) used**
* **what information from the survey e.g., variable(s) was used?**
* **What are the results? Explain & Interpret**

**Interpretation:**

**Action movie sentiment score 2 topic:**

Based on the sentiment analysis conducted on the top 2 topics for comedy genre movies, here are some observations:<br>

The majority of the sentiment scores fall in the positive range, with values above 0.5. This indicates that the sentiment towards most action movies was positive.

There are a few negative sentiment scores below 0, suggesting a negative sentiment associated with some action movies. These movies might have received negative feedback or were perceived unfavorably by the audience.

Some sentiment scores are close to 0, indicating a neutral sentiment or an ambiguous interpretation of the movies. These scores suggest that the sentiment towards these movies may not be strongly positive or negative, and the audience's perception might be more mixed or uncertain.

Overall, the sentiment scores suggest that the majority of the action movies analyzed received positive sentiment. This indicates that the audience generally had a favorable perception of these movies.

**Comedy genre sentiment score of 2 topic:**

To summarize the sentiment analysis conducted on the top 2 topics for comedy genre movies, here are some observations:

The majority of the sentiment scores are positive, with values above 0.5. This indicates a generally positive sentiment towards comedy-related items or instances.

There are a few negative sentiment scores below 0, suggesting a negative sentiment associated with some comedy movies. These movies might not have been perceived as humorous or might have received negative feedback from the audience.

Some sentiment scores are close to 0, indicating a neutral sentiment or an ambiguous interpretation. This suggests that the sentiment towards these movies may not be strongly positive or negative, and the audience's perception might be mixed or uncertain.

Overall, the sentiment scores suggest that the majority of the comedy movies were perceived positively. This indicates that the audience generally had a favorable perception of these comedic elements.

**Q3: Use regression analysis to figure out how the sentiment scores (you obtained from Q2) for the top 2 topics for action and comedy movies influence the box office revenue. What** **are the differences in results for action and comedy movies? What are the managerial implications from the results?**

* **what test(s) was (were) used**
* **what information from the survey e.g., variable(s) was used?**
* **What are the results? Explain & Interpret**

**Comedy regression analysis**

**Mean Squared Error: 829301651351364.4**

**R-squared: 0.02774524950135082**

In the given comedy regression analysis, the Mean Squared Error (MSE) value of 829,301,651,351,364.4 is quite high, indicating a large amount of error between the predicted comedy ratings and the actual ratings. This suggests that the model's predictions are not very accurate or precise.

The R-squared (R^2) value of 0.02774524950135082 suggests that only approximately 2.77% of the variance in comedy ratings can be explained by the independent variables used in the analysis. This means that the included independent variables, such as sentiment score, have limited explanatory power in predicting comedy ratings accurately.

Considering these results, it seems that the regression model for comedy is not performing well. The high MSE and low R-squared indicate that the model's predictions are not reliable and that the included independent variables do not strongly influence comedy ratings. It might be necessary to consider alternative or additional variables to improve the model's predictive ability for comedy ratings.

**Action regression analysis**

**Mean Squared Error: 1222277061325680.8**

**R-squared: 0.042426693486870226**

In the action regression analysis, the Mean Squared Error (MSE) value of 1,222,277,061,325,680.8 indicates a relatively high level of error between the predicted and actual box office revenue for action movies. This suggests that the model's predictions are not very accurate or precise.

The R-squared (R^2) value of 0.042426693486870226 suggests that approximately 4.24% of the variance in box office revenue for action movies can be explained by the independent variable of sentiment scores. This indicates that sentiment scores have a limited ability to explain the variation in box office revenue for action movies accurately.

Based on these results, it can be concluded that the regression model for action movies may not be performing well. The high MSE and relatively low R-squared value indicate that the model's predictions are not closely aligned with the actual box office revenue, and the included independent variable (sentiment scores) has limited explanatory power.

To improve the predictive ability of the model for action movie box office revenue, it might be necessary to consider additional independent variables or explore other factors that could better explain the variability in box office performance.

**Comparison**

Comparing the results of the regression analysis for action and comedy movies, we can observe the following:

Mean Squared Error (MSE):

Action: The MSE for action movies is 1,222,277,061,325,680.8. Comedy: The MSE for comedy movies is 829,301,651,351,364.4. The MSE values indicate the average squared difference between the predicted and actual box office revenue or ratings for each genre. In both cases, the MSE values are relatively high, suggesting a significant amount of error or deviation in the predictions. However, the MSE for action movies is slightly higher, indicating a potentially larger error compared to comedy movies.

R-squared (R^2) Coefficient:

Action: The R-squared value for action movies is 0.042426693486870226. Comedy: The R-squared value for comedy movies is 0.02774524950135082. The R-squared values represent the proportion of variance in the box office revenue or ratings that can be explained by the independent variables (sentiment scores). In both cases, the R-squared values are relatively low, indicating that the sentiment scores have limited explanatory power for both genres. However, the R-squared value for action movies is slightly higher, suggesting a marginally better ability to explain the variation in box office revenue compared to comedy movies.

Overall, based on these results, it seems that the regression model for action movies performs slightly worse than the one for comedy movies. Both models have relatively high MSE values and low R-squared values, indicating limitations in accurately predicting box office revenue or ratings based on sentiment scores alone. Additional factors or independent variables might be needed to improve the models' predictive abilities for both genres.

**Managerial implications:**

For comedy movies: Given the higher R-squared value, sentiment scores seem to have a more significant impact on the box office revenue of comedy movies. Filmmakers and studios can pay closer attention to audience sentiment and use it as a potential indicator of the movie's success. They may consider investing in marketing strategies that target the sentiments associated with comedy genres to attract more viewers.

For action movies: The lower R-squared value suggests that sentiment scores have a relatively weaker influence on the box office revenue of action movies. Filmmakers and studios should focus on other factors such as star power, action sequences, and storyline to attract audiences and drive box office success. While sentiment scores may still provide some insights, they may not be as reliable in predicting revenue for action movies.

Overall, understanding the differential impact of sentiment scores on different genres can help movie industry professionals make more informed decisions regarding production, marketing, and audience targeting strategies for comedy and action movies.

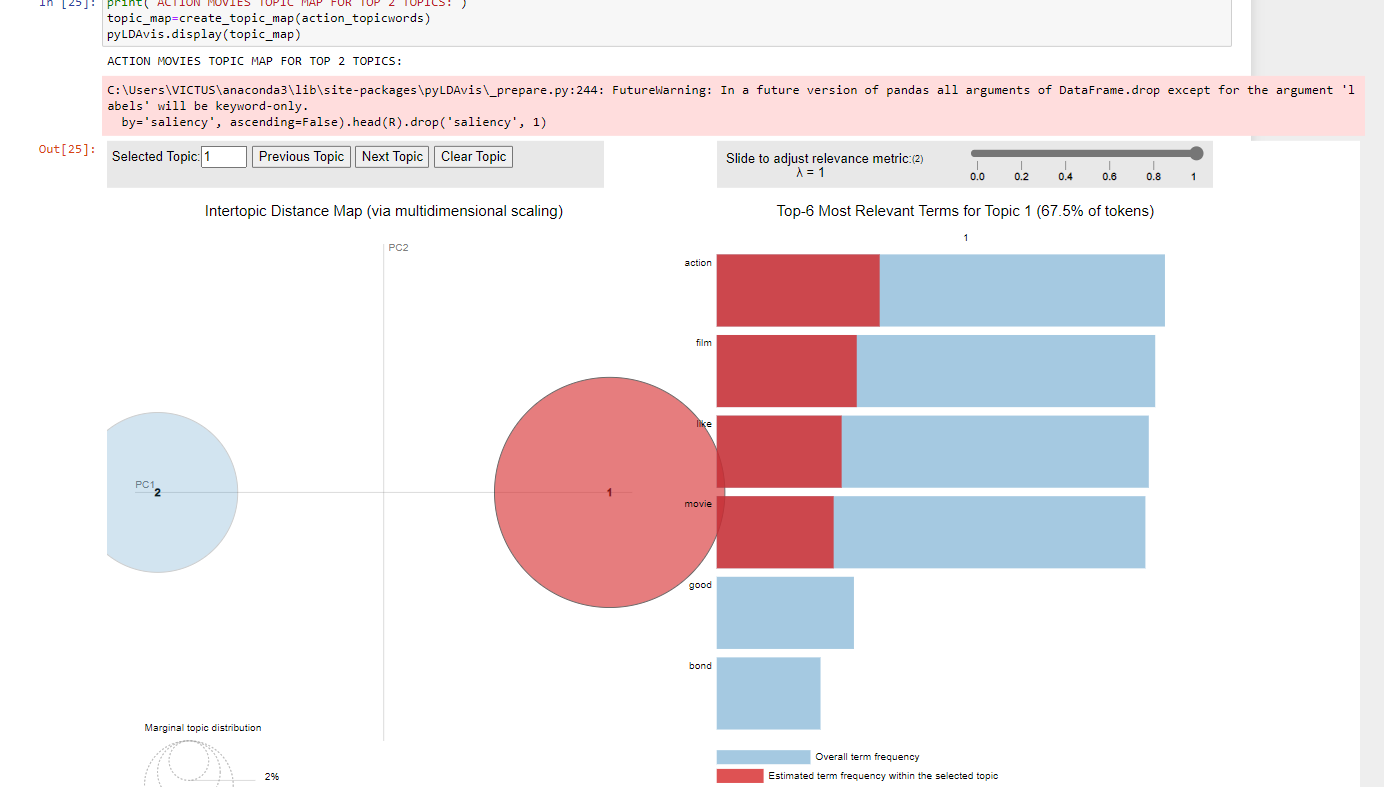
**Data Description**

The data file is *dataset\_9.csv*. It contains:

* movie: movie name
* imdbid: unique IMDB ID
* review\_post\_date: the data that a review was posted on IMDB
* review: review on IMDB
* rating: rating on IMDB (10 point scale-max: 10; min:1)
* user\_name: the name of the user who posted this review num\_helpful: the number of yes votes for helpfulness on IMDB
* num\_helpful; the total number of votes for helpfulness on IMDB (i.e., yes + no votes)
* box\_office\_revenue: the total sales ($) of the movie
* movie\_distributor: movie distributor
* budget: movie budget ($)
* release\_date: movie release date
* close\_date: movie close date
* mpaa: movie ratings by Motion Picture Association (i.e., G: General audiences – All ages admitted; PG: Parental guidance suggested – Some material may not be suitable for children; PG-13: Parents strongly cautioned – Some material may be inappropriate for children under 13; R: Restricted – Under 17 requires accompanying parent or adult guardian.)
* genre: the movie genre (i.e., Action, Comedy, Drama, Fantasy, and Horror)
* max\_screens: the maximum number of screens shown on for this movie

Graph ss:

Action one:



A screenshot of a computer

Description automatically generated

COMEDY ONE:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated