Jordan Tompkins

WGU  D214

Executive Summary

**Creating a multiple linear regression model to predict toal cinema ticket sales:**

**Problem statement and hypothesis**

One of the biggest industries in the world is the cinema industry. It provides millions of consumers worldwide the opportunity to unwind for a few hours and see the newest blockbuster hits. This industry brings in millions, if not billions of dollars per year and a large part of this industry is the theaters. Movie theaters are the first to be able to show these new movies to customes. While they are in the process of showing movies and providing a wonderful service, they are also businesses. And being a business means that the company needs to make money in order to stay afloat. Because of this, the question that was examined in this analysis was whether or not a statistically significant multiple linear regression model can be created to predict total sales for a theater. In this analysis, there were two hypotheses:

H0: The selected features will not be able to create a statistically significant model to predict total sales in the dataset and will have a p-value greater than 0.05

&

HA: The selected features will be able to create a statistically significant model to predict total sales in the data and will have a p-value less than 0.05.

**Data-analysis Process**

There were a few steps in order to conduct this analysis. The first step was to collect the data. The data used was from Kaggle.com and it contained “about eight months sales history of different cinemas with detailed data of screening, during 2018…” (Mobius, 2020). The dataset contained 145,525 rows with 14 variables, or columns, generated by an Integreated Cinema Ticket System. These variables included different metrics such as total sales, ticket price, how many tickets were sold, etc.

After the data was collected and loaded into Python, the next step was to clean and prepare the data. The open-source programing language, Python was used to conduct all of the analysis. Once loaded, Exploratory Data Analysis (EDA) was conducted. This consisted of checking the data for null or missing data, checking for any extreme outliers and also any duplicated values. Certain columns such as the ‘film\_code’, ‘cinema\_code’, and ‘date’ columns were dropped due to them being deemed unnecessary for the analysis. The film and cinema codes were anonymized to integer values so it was decided that they were to not be used. There were a few nulls and duplicates, so the next step was to fill those null values with data and drop the duplicated values. The final step was to explore the univariate and bivariate distribution of the variables. These were simple charts that showed the distribution of the data to give a better idea of what was being explored.

Once all the data was cleaned, the anlaysis could be conducted. Before creating the model, the variance inflation factors (VIF) were explored, as well as the correlation of the variables using a heatmap. This helped determine which variables were to be used within the model. Any VIF’s that had a score over 10 were not used, due to multicollinearity issues. A few variables, such as ‘month’ and ‘quarter’ were dropped because of their VIF scores. Once the variables were determined, an initial model was created. After exploring the initial model, a process known as Backwards Stepwise Elimination was used to reduce the model. This process involved removing variables one by one based on the p-value of that variable. Luckily, there was only one variable that had a p-value greater than the determined threshold of 0.05 and was subsequently dropped. This led to the determination of the final reduced model.

**Findings**

The final reduced model that was created contained six of the original variables, all with p-values less than 0.05. As shown below, the final model had a Prob (F-statistic), the p-value, less than 0.05. Because of this, it was concluded that a statisically significant model was created. Besides the p-value, other factors that were explored include the R-squared and adjusted R-squared values. Both these values were 0.873. What this showed was that 87.3% of the variation in the dependent variable, total\_sales, can be explained by the independent variables that were used.

A screenshot of a computer

Description automatically generated

**Limitations**

There were a few limitations that arose when conducting this analysis. One of those limitations has to do with the fact that the data that was used only covered one year. All of the data was from the year 2018. There was a lot of observations, over 145 thousand, however, there might be outside factors that could impact the model and data. Maybe 2018 was a good or bad year for movies and the number of ticket sold to increased or decreased accordingly. This could cause the total sales to be impacted in a more favorable or unfavorable manner. The theaters might be in areas with a high population of movie goers, or it might be far away from civilization. Using more years could give the theater using the model a better idea of how things are over time. Another limitation is that multiple linear regression is sensitive to outliers (Flom, 2019). Without the proper cleaning and preparation of the data, the regression model could be impacted due to extreme outliers.

**Propsed actions**

There is a lot of different actions that could be taken by a theater using this model. The first idea would be to get more data. As mentioned, the data only covers one year. If there was more data being studied, it is possible for a better model to be created. More than one year will give the theaters a lot more data and a much better look into total sales and how to use the multiple linear regression model to predict total sales. One could explore different variables along side the ones already explored. It might be beneficial to look into whether or not the movie was a sequel to a box office hit, or if it was in a large franchise. Having this information would allow the user to know how certain types of movies might impact the regression model. A final action that should be taken is to compare the prediction results to the actual results. This will allow the company to truly get a better understanding of whether or not the theater is making what it theoretically should be making. Is the company’s total sales more than what the model predicts? Is it less? Using the model to compare total sales between a predicted model and what the actual total sales were can give the theater a better understanding of how well, or how poorly, the theater is actually doing.

**Expected benefits**

One of the biggest benefits with the analysis is that a model was created for predicting total sales for the theater. Because of this, the theater can use this model to compare how well they actually performed throughout the year versus how well they should have performed based on the prediction model. This will allow them to determine whether or not they need to improve their overall business, and even determine the bare minimum they need to hit for certain variables such as tickets sold, or tickets used. Another benefit is they know what variables impact sales. They can see how much certain variables impact total sales and it will allow them to better focus their efforts to what has the greatest impact on their total sales for the year. Being able to focus on what variables might largely influence the total sales will be very beneficial for the theaters.

Flom, P. (2019, March 2). *The disadvantages of linear regression*. Sciencing. Retrieved March 4, 2023, from <https://sciencing.com/disadvantages-linear-regression-8562780.html>

Mobius. (2020, October 29). *Cinema tickets*. Kaggle. <https://www.kaggle.com/datasets/arashnic/cinema-ticket>