Jordan Tompkins

[Company name]  [Company address]

Creating a multiple linear regression model to predict total cinema ticket sales

**Research Question**

The cinema industry is one of the biggest industries in the world. They do more than just show movies to audiences; they provide experiences that most people cannot find at home. When someone goes to the movies they get the opportunity to kick back and relax, enjoy some popcorn, candy and soda, and enjoy the experience of seeing their favorite actors/actresses and stories being portrayed on the big screen for a much more immersive viewing experience compared to their couch at home. While movie theaters are in the industry of providing this service to the public, they are still a business and as a business, they need to make money in order to operate. Because of this, many theaters could benefit from a predictive model to help predict how much total sales they are able to bring in. If theaters were to use a multiple linear regression model in an attempt to predict total sales, they could see if there are any factors that might influence total sales thus allowing them to get a good idea as to what ways the can improve their business in order to make money. This leads to the research question being explored: Can a statistically significant multiple linear regression model be created to predict total sales? In the exploration of this, the hypotheses are as follows:

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 Collection**

The data that was used for this analysis came from a Kaggle.com data set uploaded by a user who goes by the handle Mobius. It was “about eight months sales history of different cinemas with detailed data of screening, during 2018 with encoded anonymized locations” (Mobius, 2018). The data set simply titled “Cinema Ticket” consists of 145,525 rows with 14 variables generated by an Integrated Cinema Ticket System. These variables include a wide variety of different metrics such as total sales (which was the target variable), ticket price, the time the movie was showing, how many tickets were sold and how many were used, etc. It even included specific day and month data which can help explore how the timing of the release of the movie might impact total sales. Some advantages to using this data set was that it was readily available and contains such a large amount of data related to cinema ticket sales that it was perfect for analysis. One of the disadvantages was the dataset only contains information from one specific year, 2018. While the total amount of observations is extremely high, there is the possibility that outside factors, such as many box office hits being released in 2018, that could potentially allow for the creation of a model that might not be as useful in the long run.

**Data Extraction and Preparation**

The main tool that will be used for this analysis is Python and JupyterNotebook with the python kernel. Python is an free to use, open source programing language that has many useful packages and easy to use syntax that will be helpful in creating the multiple linear regression model (“R or Python.”, 2022). Within Python, the Pandas, Numpy, Matplotlib and statsmodels libraries will be used to perform exploratory data analysis, multiple linear regression and create any visualizations needed for the analysis.

There are many benefits when it comes to using all the tools and techniques listed above. To start, python is one of the most widely used programs for data analysis. As stated above, it has many useful packages and libraries that are being used, as well as an easy to use syntax. Python is also “…a robust tool to handle, process, and model data. It has an array of packages for linear regression modeling” (Li, 2019). Jupyter Notebook with the Python kernel provides a space where one can both work with the Python script as well as provide a narrative output.

Pandas allows for all of the data gathered to be placed into a single dataframe. The NumPy library provides mathematical functions needed to perform the analysis to other libraries and packages such as Pandas. MatPlotLib will allow for the creation of any necessary data visualizations needed for the analysis. Statsmodels provides the “ols” function needed for the creation of a multiple linear regression model. All of these libraries will used in together in order to create the multiple linear regression model.

Exploratory Data analysis was the first step in data preparation and cleaning. Exploratory data analysis (EDA) is used to summarize the data, visualize and identify any trends, outliers, missing data etc. It’s one of the first steps in data analysis because it will allows for the user to gain a better understanding of the data that is being studied and identify any issues that need to be corrected before the regression model is created.

The first step to begin EDA was to read the data into the python kernel within Jupyter Notebook. Once that was loaded in, the first 5 rows were called using df.head(5) to make sure everything was loaded in properly, before using df.info() to view some basic information surrounding the data set, such as how many non-nulls it contains, what type of variables were being explored, etc.

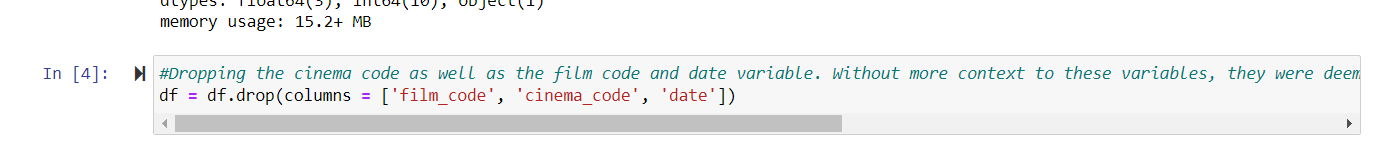
**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Upon further investigation of the variables, it was determined that 3 variables appeared to be unnecessary to the analysis. Those variables include ‘cinema\_code’, ‘film\_code’ and ‘date’. The reason being, the cinema code and film code were integer variables, and anonymized versions of the cinema’s name and the film’s name. While these could be helpful, especially if this model was used for one specific cinema, due to the anonymized responses, these were deemed unnecessary for creating the multiple linear model. The ‘date’ variable was also dropped due to it being a datetime variable, which does not work with creating a linear regression model.



A few of the other variables, specifically, ‘month’ and ‘day’ were already in the dataset and are sufficient enough to use in place of the actual date. Because these three variables were deemed unnecessary, they were dropped from the dataframe.

Once those variables were dropped, the next step was to check for null values and duplicates within the data set. There was 657 rows that were duplicates which were subsequently dropped. Upon inspection of the nulls, it was noted that there were two variables who each had 125 null values.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

To handle those, first the distributions were explored. Looking at the histograms of these two variables, both had skewed distributions.

A screenshot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

Because of the skewed distribution, the way to fill in those missing values was to use the .fillna() method, filing in the missing values with the median of the column.

A screenshot of a computer

Description automatically generated

After handling null values, using df.describe() the description of the data, such as the count, mean, minimum and maximum values, etc. was explored.

A screenshot of a computer

Description automatically generated

It was noted that within the ‘capacity’ column, the minimum was a negative number. Basic knowledge of cinemas allows for one to assume that these values might have been a mistake. Because of this, any rows where the capacity was a negative number was dropped.

A screenshot of a computer

Description automatically generated

Another item that was noticed upon further investigation of the description of the data, was that there could potentially have been some outliers. However, without knowing more information surrounding the cinemas (can they really fit thousands of people in the cinema thus leading to such high numbers elsewhere, etc.?), it was decided that the outliers were to stay in the data set.

After the data was treated, and missing values were handled, the next step was exploring the data by looking at univariate and bivariate visualizations of the data. See below for the univariate and bivariate visualizations:

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

**Analysis**

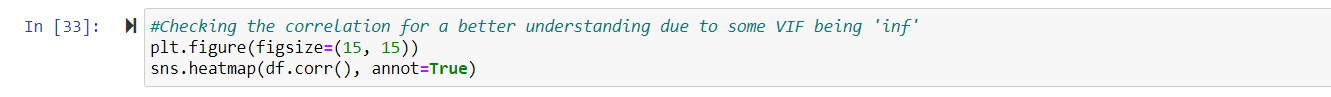
Once the data was prepared, cleaned, and EDA was performed, the analysis was able to be conducted and the model could be created. Linear regression was the mode that was used for this analysis. Linear regression is used to model relationships between a dependent variable and a number of independent variables. Multiple linear regression was used because the target, or response variable, total\_sales, was a continuous variable. The reason why multiple linear regression was used instead of simple linear regression is because the model that was being created had multiple explanatory variables. A multiple regression model allows for the use of several explanatory variables rather than a single one which is used in simple linear regression (Hayes, 2023). Because there are more than one predictor variable that is being explored in the analysis, multiple linear regression was needed. Using multiple linear regression allows researchers to understand and predict the relationship between the dependent and independent variables.

The first step was checking the variance inflation factor of the variables. This “…is a measure of the amount of multicollinearity in regression analysis” (“Variance inflation factor”, n.d) The VIF can estimate how much variance of a regression coefficient is inflated due to multicollinearity. See the VIF below:

A screen shot of a computer

Description automatically generated

Typically, one would want to not use a variable that has a VIF above a certain threshold. For this analysis, it was decided that 10 was just that. It is noted that there were a few variables that had an “inf” for their VIF. Because of this, a heatmap of the correlations was used below to get a much more general idea of the correlation between variables.



A screenshot of a screen

Description automatically generated

In this case, most variables had moderate correlation to total\_sales, the target variable. A couple did have high correlations; however, it was decided that they were to be included in the regression model because when taken out in testing, the model did not perform as well. Once correlation and VIF was explored, the initial model was to be created below:

A screenshot of a computer

Description automatically generated

To determine ways to reduce the model, there are many different processes that could be used. The one used in this analysis was Backwards Stepwise Elimination. Backwards Stepwise elimination is a process used when creating a multiple linear regression model. The first model created will not always be the best model due to the potential of having too many insignificant variables. Using backwards stepwise elimination, researchers are able to eliminate any variables based on significance thus allowing them to reduce the model. It is a reduction technique that allows for the use of all of the determined initial variables and then variables are eliminated from the model based on their significance by looking at the variables p-values (Choueiry, 2022). The threshold for determining what variables can be eliminated using Backwards Stepwise Elimination, was any variables with a p-value greater than 0.05. There was only one variable that had a p-value greater than 0.05, and that was ‘ticket\_use’. Because of this, this variable was eliminated and a final reduced model was created:

A screenshot of a computer

Description automatically generated

**Data Summary and Implications**

The model that was created did in fact have a p-value, or the Prob (F-statistic), less than 0.05, so a statistically valid model was in fact created. Both the initial model and the reduced model had Prob (F-statistics) of 0.00. Assuming it is not actually zero, it is pretty close to zero when rounded. This tells us that the regression is a meaningful regression. Another metric that was explored was the R-squared value. This metric shows the percent variation of the dependent variable that is explained by the independent variables. In the model created, the R-squared and adjusted R-squared values were 0.873, which means that a little over 87% of the variation in the dependent variable is explained by the independent variables. All off these metrics taken into consideration, it can be concluded that a statistically valid model can be created within the dataset using multiple linear regression.

While a statistically valid model was able to be created, there are some limitations to using multiple linear regression. One of those limitations is that multiple linear regression is sensitive to outliers (Flom, 2019). There did appear to be potential outliers within the dataset, however, it was determined that all data was to be used as to not eliminate relevant information. With more information related to the different cinemas in which that data was taken from, perhaps those outliers can be properly handled. However, for this analysis, they were kept in to create the model.

There is a few different directions that future studies of the dataset could take. One of those directions is to continue gathering data related to the total sales of cinema tickets, however, doing so over multiple years. From there any issues with outliers can be dealt with and multicollinearity can be explored and addressed. Once those issues are addressed, another multiple linear regression model can be created. Perhaps a better model can even be created. Another direction is to look at the dataset as a timeseries dataset and conduct forecasting and time series analysis. There is already a date variable in the “YYYY-MM-DD” format. A future study can look at the dataset and do a forecast of total sales over time.

Choueiry, G. (2022, November 25). *Quantifying health*. QUANTIFYING HEALTH. Retrieved March 4, 2023, from <https://quantifyinghealth.com/stepwise-selection/>

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

Hayes, A. (2023, January 11). *Multiple linear regression (MLR) definition, formula, and example*. Investopedia. Retrieved March 4, 2023, from <https://www.investopedia.com/terms/m/mlr.asp>

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

Li, L. (2019, February 5). *Introduction to linear regression in python*. Medium. Retrieved March 4, 2023, from <https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0#:~:text=Understanding%20how%20to%20implement%20linear,packages%20for%20linear%20regression%20modelling>.

“R Or Python.” *Western Governors University*, 16 Nov. 2022, <https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html>.

*Variance inflation factor (VIF)*. Investopedia. <https://www.investopedia.com/terms/v/variance-inflation-factor.asp>