The dataset used was pulled from the University of California Irvine Machine Learning Repository website, where it was donated by Moro et al (2017) for students to use to explore machine learning techniques. The dataset is composed of 504 total reviews from 21 different hotels on the Las Vegas Strip, extracted from TripAdvisor between January and August of 2015. There are 20 different attributes (Figure 1) that can be analyzed for importance in creating an accurate prediction of TripAdvisor scores based on the attributes of the reviewer and the hotel. The data came within a comma-separated file, which was looked through and one or two obvious errors in the data, such as a negative number in the 'member year' column in row 77, were manually changed to zero to represent less than one year. The comma separated file also had commas instead of decimals when indicating a fraction of a number in the 'hotel stars' column, which was manually changed within the comma separated file in order to import it as a .csv type file. After the data was found to be very imbalanced with a large skew towards hotel scores of four and five(Figure 2), a methodology was built to test if the imbalance was significant and how to account for it in the modeling.

The model was built using Python-based programming with many additional modules added to it. These modules include NumPy, Pandas, Scikit-Learn, and Imbalanced Learn modules. The data were split and trained to prepare it, and values of 'yes' and 'no' were changed to zero and one to make it easier for the machine to comprehend. All code was built and run on a cloud-based platform called Google Colab, which allowed the code to be run from any device with internet connection.

After initial data cleaning, results were obtained that made it clear the data was not properly cleaned for the purposes intended. A dictionary had been made and had edited the data, which did not provide results for categorical variables. It was decided to use the label binarize

function from Scikit-Learn on the original dataset, before it was cleaned and adjusted for negative values and zeroes. The original dataset was decided to be used for this to avoid using the dictionary made for the original methodology, which put weight on certain values based on their representative numerical value. Each column with categorical values were binarized and added to the original dataset, and the non-binarized column was dropped. For example, the 'User continent' feature had six different values throughout all rows, and once the values were binarized the original 'User continent' column was dropped and was replaced with the six categorical values as the feature name of the new columns. Because this was done with each categorical feature, the resulting dataset had 116 columns. The revised dataset had to undergo cleaning again, since the version of the data that had been binarized still had commas instead of decimals in the 'Hotel stars' column and a negative value in the 'Member years' column. Instead of manually changing all of the data, R was used to make these changes and the desired data was exported back into Colab.

In R, a linear model was created with each of the 20 columns and a backwards elimination was done to determine the significance of each variable to the model. A linear model was chosen due to restraints in the toolset of the researchers. Each run of the backwards elimination, the variable with the highest p-value was removed and the model was run again. All variables were deemed to be significant when p-values were at 0.05 or below, when only five variables remained. This method was used to find the best features for modelling, which resulted in the feature matrix including 'Dummy Pool', 'Dummy Free Internet', 'Saturday', 'China', 'Business', and 'Excalibur Hotel and Casino' being the most significant features, which was an increase from five to six features to use in training and testing the data. The binarized features

‘China’, ‘Business’, and ‘Excalibur Hotel and Casino’ come from the original columns named ‘User country’, ‘Traveler type’, and ‘Hotel name’.

An alternative feature matrix was made with more resulting features, which was also ran and compared to the first decision tree model. The features resulting from this matrix were ‘Dummy Pool’, Dummy Free Internet’, ‘Business’, ‘Families’, ‘China’, ‘Scotland’, ‘Spain’, ‘Dec-Feb’, ‘Excalibur Hotel and Casino’, ‘Hilton Grand Vacations at the Flamingo’, ‘Paris Las Vegas’, ‘The Westin las Vegas Hotel Casino and Spa’, ‘Treasure Island TI Hotel and Casino’, ‘December’, ‘February’, ‘January’, ‘Friday’, and ‘Saturday’. Many of these features are included in the original features selected, and classification metrics were run to compare the two sets of features. It was determined that the original six features yielded better accuracy and overall fit the data better than the alternative features, so the original features were used in conjunction with the cleaned dataset two for the remaining modelling.

The cleaned dataset two was then split, trained, tested, and validated. The data was also oversampled and weighted for use in the models. The values of the weight for each hotel score were a weight of 8.06 set for the hotel score value one, a weight of 4.2421 for a hotel score of two, a weight of 1.4392 for the hotel score of three, a weight of 0.6014 for the hotel score of four, and a weight of 0.4380 for a hotel score of five. The models run from the resampled cleaned dataset two were the decision tree classification, linear regression, logistic regression, Ada boost classification, random forest classification, and a K-means clustering method. For each model, the metrics precision, recall, accuracy, F1 score, Mean Absolute Error, Root Mean Squared Error, Mean Squared Error, and R^2 were calculated and compared. Confusion matrices were also used for each model to determine the success of the model in accurately predicting all five types of scores. Finally, each of the aforementioned models were cross-validated with the

SVM module from Scikit-Learn for precision values to draw conclusive results to compare each model.

	User country	Nr. reviews	Nr. hotel reviews	Helpful votes	Score	Period of stay	Traveler type	Pool	Gym	Tennis court	Spa	Casino	Free internet	Hotel name	Hotel stars	Nr. rooms	User continent	Member years	Review month	Review weekday
0	USA	11	4	13	5	Dec-Feb	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	9	January	Thursday
1	USA	119	21	75	3	Dec-Feb	Business	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	3	January	Friday
2	USA	36	9	25	5	Mar-May	Families	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	2	February	Saturday
3	UK	14	7	14	4	Mar-May	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	Europe	6	February	Friday
4	Canada	5	5	2	4	Mar-May	Solo	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	7	March	Tuesday

Figure 1. This table is the original dataset as it was downloaded from the UCI Machine Learning Repository, with 20 columns. There are numerical, categorical, and yes/no values in each column, instead of a consistent data reporting type. In the ‘Review Weekday’ column on the far right, it can be seen that not all of the categorical variables are even the same, which shows that there are many nuances to this dataset that will need to be worked on before any significant analysis or machine learning can begin.

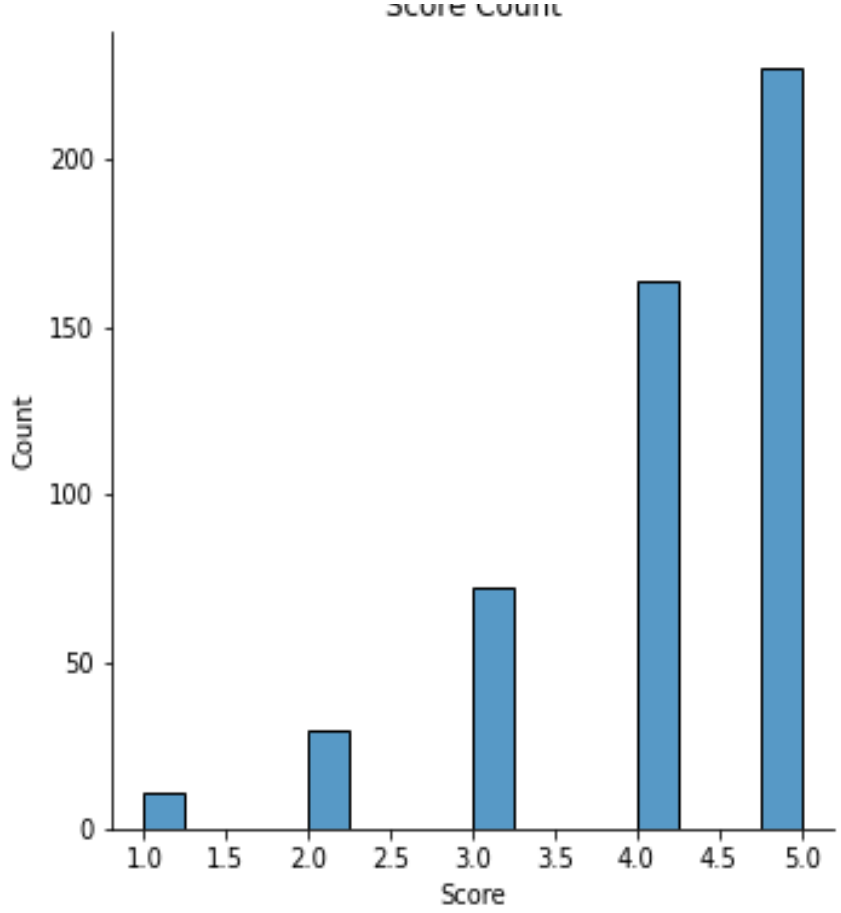

Figure 2. A histogram bar chart shows the overall distribution and count of hotel scores within the dataset. Counts are recorded in increments of fifty on the left side of the chart, with scores reported on the x-axis in 0.5 increments from 1.0 to 5.0. As scores were only reported in the dataset as whole numbers, counts from the n whole number to the n+0.5 tick represent the count of scores within the bin from whole number to whole number, for example the count from 1.0-1.5 represents the count for the bin from 1.0-2.0.