Date: 22 October 2020

COS60008 – INTRODUCTION TO DATA SCIENCE ASSIGNEMNT – 2

CUSTOMER FEEDBACK ON LAS VEGAS HOTELS THROUGH DATA ANALYSIS

By,
VAIBHAV SANGAMKAR
Student Id: 101894149
Email: 101894149@student.swin.edu.au
Ph. No.: +61 470202488

ABSTRACT:

This report provides the analysis on the hotels in Las Vegas which is based on the TripAdvisor dataset. The main purpose of this report is Identification and Classification of problem based data-analysis, conceiving the problems, and finding out the data which is to be solved. The data is extracted from the UCI repository. There are 3 phases of data science process, Data preparation, Data Exploration and Data Modelling. In the first phase, the data is loaded, cleaned, and prepared for the next phase. The second phase is to explore the data with all attributes with its relationships between all dataset attributes. In addition, a research question is raised and tried to get the answer using statistical or graphical visualizations. The final phase is the Data Modelling, the dataset is split into Training and Testing dataset and two classification models are applied namely Random Forest and Decision Tree. Both the models are compared, and the results are visualized.

NOTE: To compare between two classifier models, the ROC and Precision Recall Curve takes only Binary data, it is not possible to show the difference between these two models. Therefore, Bar graph is used, and it shows clear difference.

INTRODUCTION:

In today`s world, Online Travel Agencies plays a crucial role, in terms on travel booking, accommodation and transportation. As almost every customer can be able to access internet and is easy to express their thought in terms of feedback on various hotels, travel agencies etc. People tend to go look for the feedback of the other people before making their decision of booking a particular hotel. The main aim of this report is to provide the ratings (Score) by using of Predictive Analysis and by various classification models. Some visualizations are also shown which helps in clear understanding. Visualizations can also be used to know all the columns and their relationships and Feature Importance is given which helps in understanding the important columns which can be affected by Score. Classification models are used to Train and Test the dataset to know the Accuracy and Score of the data and comparison of these models is also shown with a help of a graph.

DATASET INFORMATION:

The dataset is extracted from UCI repository, which contains the Quantitative and Categorical data from reviews which are collected from 21 hotels in Las Vegas. It has got 504 reviews in total which are gathered from all around the world. This dataset consists of 504 rows with 20 columns. The current dataset contains of all the data on review features, user features and hotel features information. All the reviews are featured from the Traveler type, Period of Stay, Review Month or Weekday. It also includes user information such as User Country, User Continent, Member Years and their reviews and votes on hotels. The hotel data consists of their hotel name, number of rooms and the amenities of hotels such as Gym, Casino, Spa, Free-Internet, Tennis Court etc.

Dataset is downloaded from UCI Repository and the link is provided below.

https://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip

TASK 1 - DATA ACQUISITION AND PREPARATION:

1.1 PROBLEM FORMULATION:

TripAdvisor must get the analysis of the reviews made by people on the hotels on Las Vegas. Score is the major factor which is considered by every person as everyone rates the Hotels and gives reviews on internet, based on these reviews other people tend to attract to that hotel. Different set of people go to different hotels and stay for a period and then provide reviews, these reviews can be non-identical to one another. All the features are analysed and calculated based on Score. The challenge the hotels face is to maintain the score, to solve this issue the features are analysed which can affect score and the score is predicted. Now the hotels can know which features attracts the customers the most, to keep their score flawless.

1.2 DATA ACQUISITION:

The Las Vegas Trip dataset is a extracted from UCI repository which contains all the descriptions. The file has all its Headers for each column and the Score is set as Class label in the dataset which satisfies all the criteria provided i.e., it has got more than 150 rows and more than 5 columns with categorical and numerical columns.

1.3 DATA PREPARATION:

At the beginning, the dataset was not properly in condition as it was not in a proper format. Therefore, the table was created and changed the column names for ease.

Importing the Libraries and Packages:

All the essential Libraries and Packages were imported such as Pandas, NumPy, Matplotlib, Seaborn, sklearn and others which are needed for modelling.

Loading the Dataset:

The Las Vegas Trip dataset is loaded and renamed as **LasVegasTrip.csv** into Jupyter Notebook with all 504 rows and 20 columns.

Cleaning the Dataset:

The data needs to be cleaned to further analyse. To clean the data, following methods were used.

- **Removing NULL Values:** The data is checked for NULL values and some NULL values can be found in few columns such as No_of_rooms, User_continent, Member_years, Review_month and Review_weekday. These NULL values were removed.
- **Checking Duplicates:** The dataset is checked thoroughly, and NO duplicate value can be found in either rows or columns.
- **Changing the Datatypes:** The dataset is checked for their datatypes. Few datatypes were changed for the columns Member_years, No_of_rooms and hotelstars for convenience in understanding the data and to visualize.
- **Checking for Unique Values:** The Unique Values are now checked to know whether there are any blank spaces, spelling errors or any other errors.

Descriptive Statistics: The Descriptive statistics is performed differently for Numerical columns and Categorical columns to analyse more about the dataset.

Feature Importance:

The Feature Importance is used to find the most contributing features columns for Score. As there are 20 columns, it will be easy to analyse first 10 columns which are related to Score. The steps which are used to find the Most Featured Column are:

- Select and Define all the columns in the dataset.
- Transform all the Categorical columns into Numbers which helps to calculate and compare all the other columns.
- Removed the "Score label" from the dataset as it is easy to compare and the Score label will not be included to find the features columns.
- Prepared the Training and Testing datasets.
- Applied the Random Forest classification and executed Recursive Feature elimination which helped us to find the Most Contributed Features for Score.

The Most Contributed Features for Score are:

```
**Most Contibuted Feature:**
['Period_of_stay' 'Traveler_type' 'Pool' 'Tennis_court' 'Spa' 'Casino'
 'Free_internet' 'hotelstar' 'User_continent' 'Review_weekday']
```

TASK 2 – DATA EXPLORATION:

Based on the Most Contributed Features columns, Data Exploration is performed.

2.1 EXPLORING ALL COLUMNS:

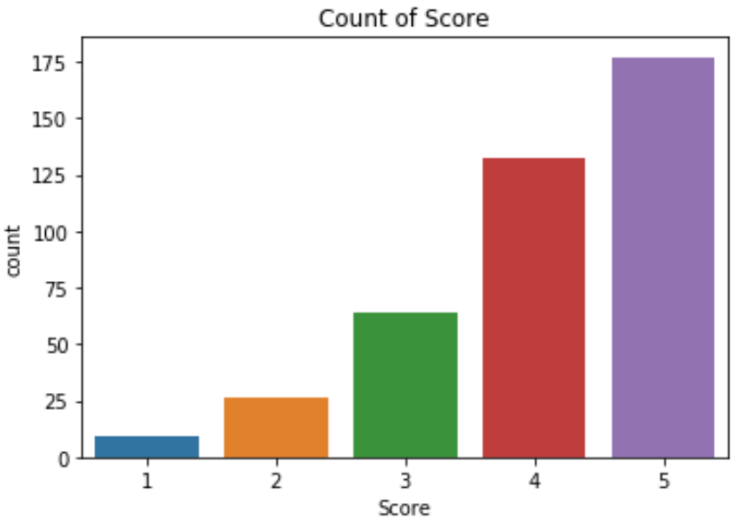
Descriptive Statistics:

	No_of_reviews	No_of_hotelreviews	Helpful_votes	Score	hotelstar	No_of_rooms	Member_years
count	408.000000	408.000000	408.000000	408.000000	408.000000	408.000000	408.000000
mean	43.848039	14.764706	29.767157	4.083333	4.235294	2540.529412	-0.068627
std	72.075080	23.086283	46.142629	1.019667	0.807538	1177.605288	89.674624
min	1.000000	0.000000	0.000000	1.000000	3.000000	315.000000	-1806.000000
25%	12.000000	5.000000	8.000000	4.000000	4.000000	1467.000000	2.000000
50%	22.000000	9.000000	16.000000	4.000000	4.000000	2916.000000	4.000000
75%	47.250000	17.000000	31.000000	5.000000	5.000000	3348.000000	7.000000
max	775.000000	263.000000	365.000000	5.000000	5.000000	4027.000000	13.000000

The above Statistics are for all Numerical Values. From the statistics we can say that,

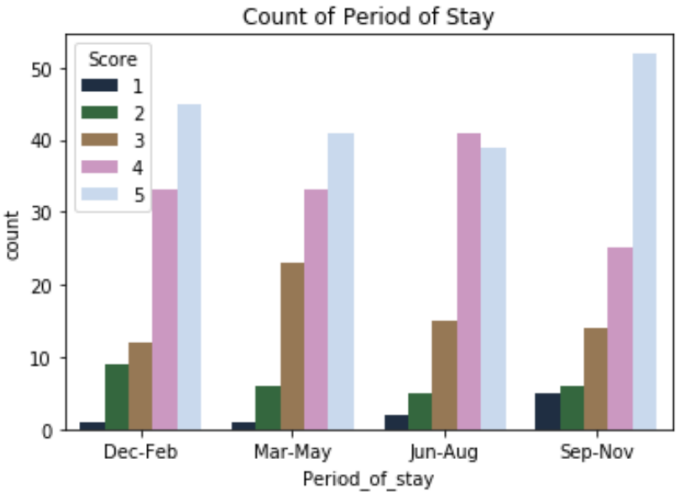
- Score, which is the class label has range from 1 to 5 and the average value is 4.
- Hotel Star rating has a range from 3 to 5 and the mean value is 4. Hence, all the hotels are high rated, and as the ratings are high the score will also be high.

- As most of Hotels are highly rated, we can say that the Quality and Infrastructure of the hotels are good and therefore average rooms in the hotels are 2540.



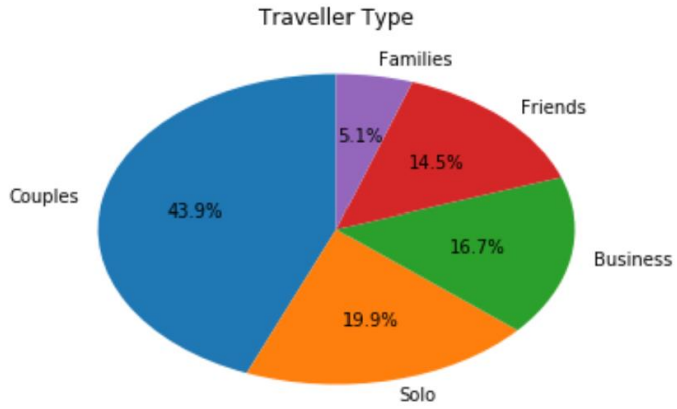
The above Bar graph represents the count of Score. As the maximum score is above 100, we can say that majority of scores are 4 and 5.

Period of Stay based on Score:



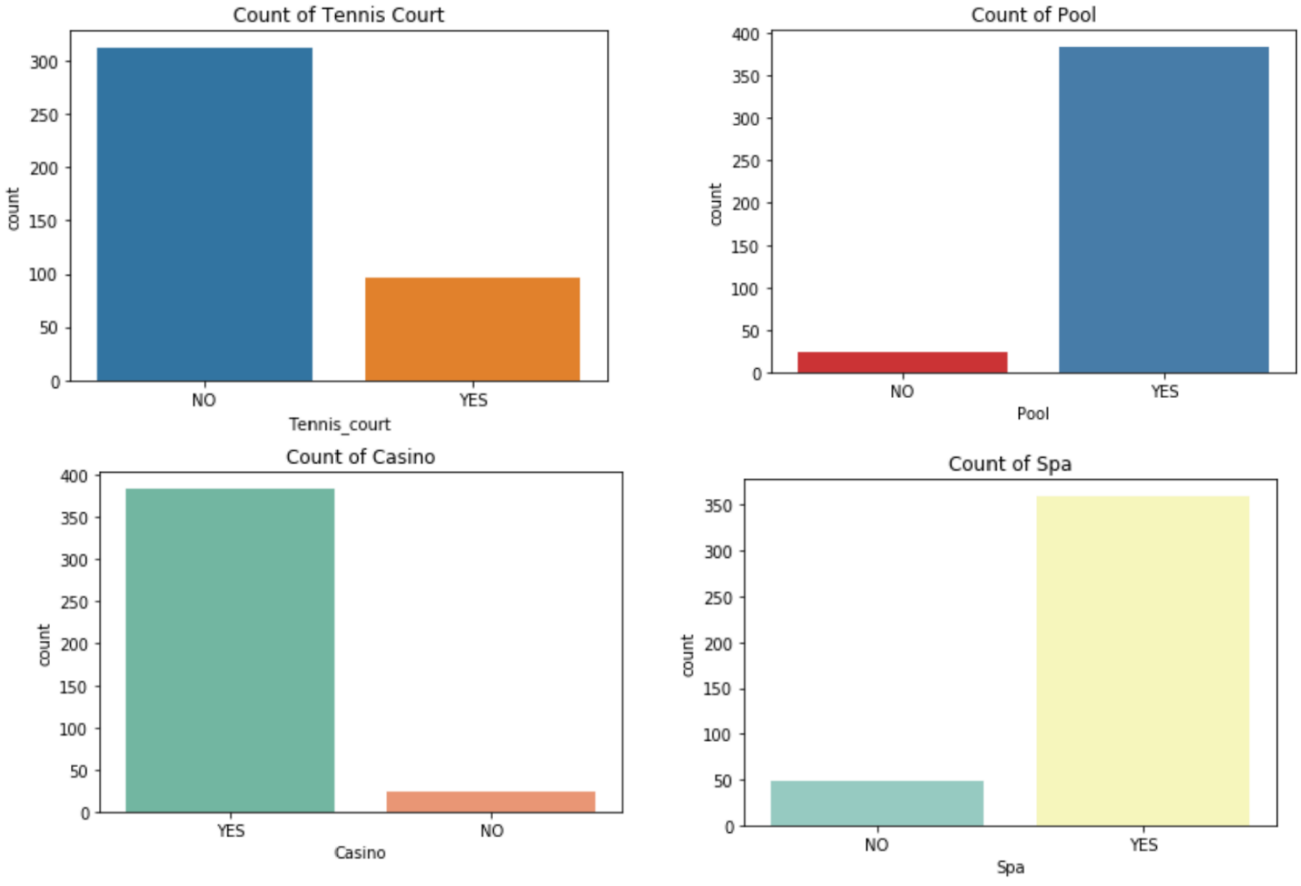
As we know that the maximum Score ranges between 4 to 5, we can see from above graph that maximum score on period of stay is also ranging from 4 to 5. Therefore, it helps in estimating the score ranges in each period.

Percentage of Traveller Type:



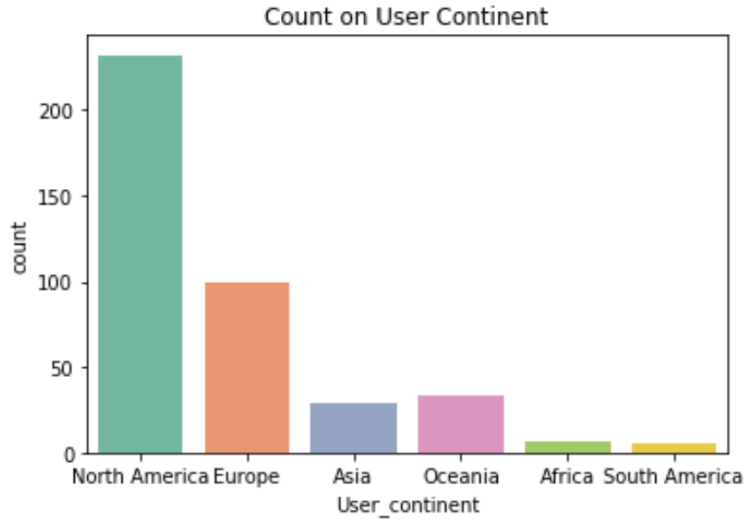
The above illustration shows the Percentage of Traveller Type such as Couples, Friends, Family, Business and Solo. As per the Pie graph, 43.9% Couples travel to Las Vegas and Families are the least with only 5.1% when compared to other traveller types. By knowing this analysis, we can suggest that it is good to keep some packages for Couples to attract more people. The hotels can advertise the services provided for families as they are the least traveller types.

Different Types of Services provided by the Hotels:



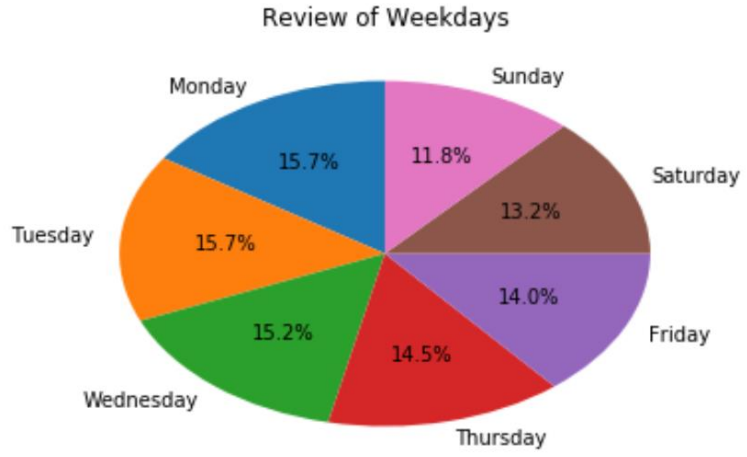
The above are few amenities or services which are provided by the Hotels. It is clearly understandable that most of Hotels provide all the services but only few provide Tennis Courts service as it takes a huge space and there are less chances of providing Tennis Courts for the Hotels.

Count on Users from different Continents:



Therefore, by above graph we can say that the maximum number of people come from NORTH AMERICA and the least number of people come from SOUTH AMERICA.

Percentage of Reviews on Weekdays:



By the above Pie graph, we can say that most people come on Mondays and Tuesday.

2.2 RELATIONSHIPS BETWEEN COLUMNS:

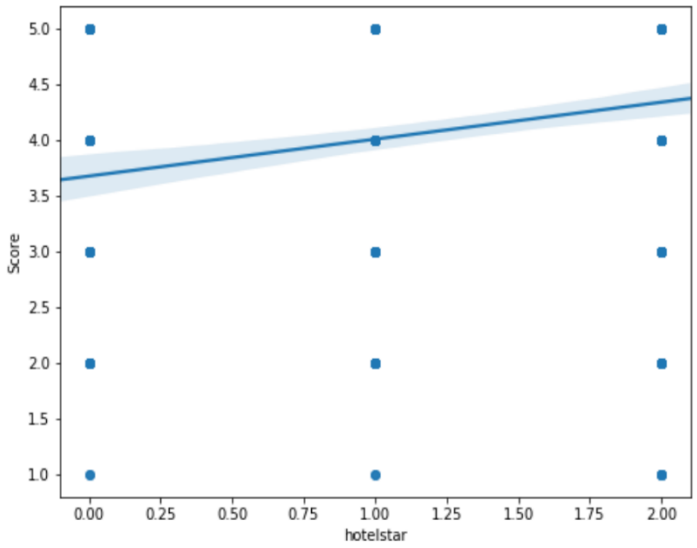
Correlation Matrices:

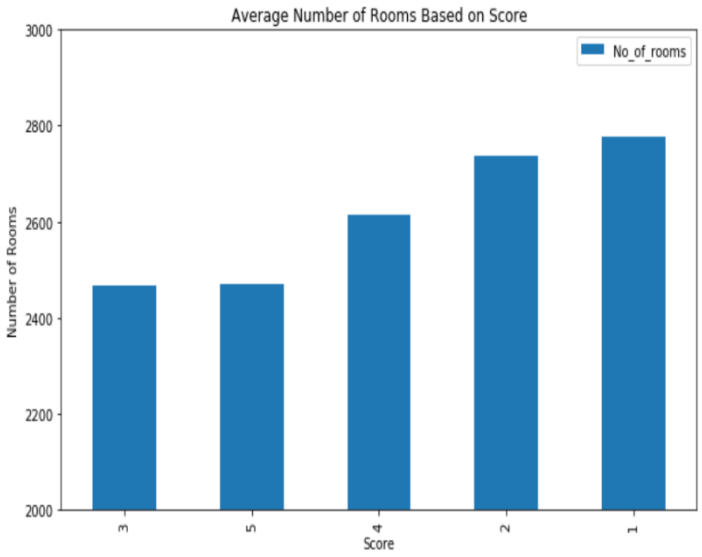
	Period_of_stay	Traveler_type	Pool	Tennis_court	Spa	Casino	Free_internet	hotelstar	User_continent	Review_weekday
Period_of_stay	1	-0.042	-0.035	-0.0049	-0.024	0.0022	0.0022	-0.013	-0.034	-0.084
Traveler_type	-0.042	1	-0.11	-0.027	-0.063	0.019	0.038	-0.056	-0.058	0.00079
Pool	-0.035	-0.11	1	0.14	0.68	-0.062	-0.063	0.38	-0.047	0.031
Tennis_court	-0.0049	-0.027	0.14	1	0.2	0.14	0.14	-0.16	0.022	-0.021
Spa	-0.024	-0.063	0.68	0.2	1	0.68	-0.091	0.56	-0.087	-0.0066
Casino	0.0022	0.019	-0.062	0.14	0.68	1	-0.062	0.38	-0.072	-0.04
Free_internet	0.0022	0.038	-0.063	0.14	-0.091	-0.062	1	0.073	0.088	0.0006
hotelstar	-0.013	-0.056	0.38	-0.16	0.56	0.38	0.073	1	-0.031	-0.019
User_continent	-0.034	-0.058	-0.047	0.022	-0.087	-0.072	0.088	-0.031	1	0.022
Review_weekday	-0.084	0.00079	0.031	-0.021	-0.0066	-0.04	0.0006	-0.019	0.022	1

The above is the Correlation Matrix for Categorical Columns (Most Contributed Features).

	No_of_reviews	No_of_hotelreviews	Helpful_votes	Score	hotelstar	No_of_rooms	Member_years
No_of_reviews	1	0.59	0.77	-0.025	-0.029	-0.088	0.022
No_of_hotelreviews	0.59	1	0.73	0.013	-0.078	-0.09	0.022
Helpful_votes	0.77	0.73	1	0.019	-0.0056	-0.061	0.023
Score	-0.025	0.013	0.019	1	0.26	-0.05	-0.042
hotelstar	-0.029	-0.078	-0.0056	0.26	1	0.27	0.016
No_of_rooms	-0.088	-0.09	-0.061	-0.05	0.27	1	-0.016
Member_years	0.022	0.022	0.023	-0.042	0.016	-0.016	1

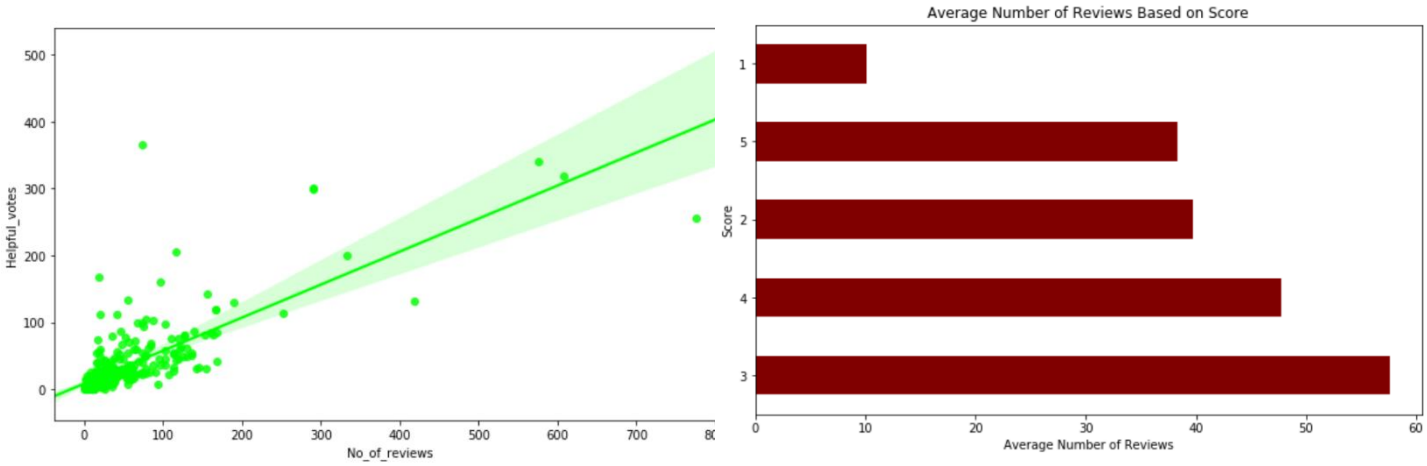
The above is the Correlation Matrix for Numerical Columns.





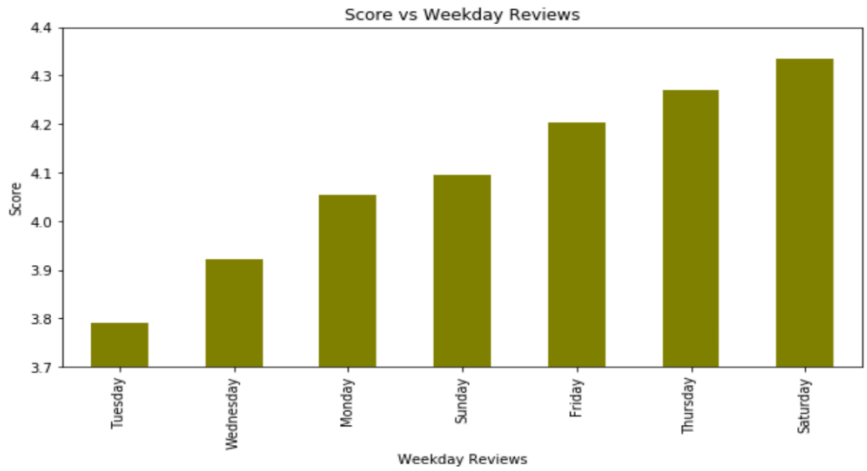
A Linear Regression graph is shown which represents the Hotel Stars based on the Score. By knowing the graph, we can say that it is positive yet weakly correlated. Hence, when the Score increases the Hotel Stars also increases. The bar graph at the right side illustrates the Number of rooms based on the score. By seeing the graph, we can say that when Score increases, the number of rooms decreases. Therefore, there are a greater number of rooms in low scored hotels.

The Linear Regression graph represents the correlation between Number of Reviews and



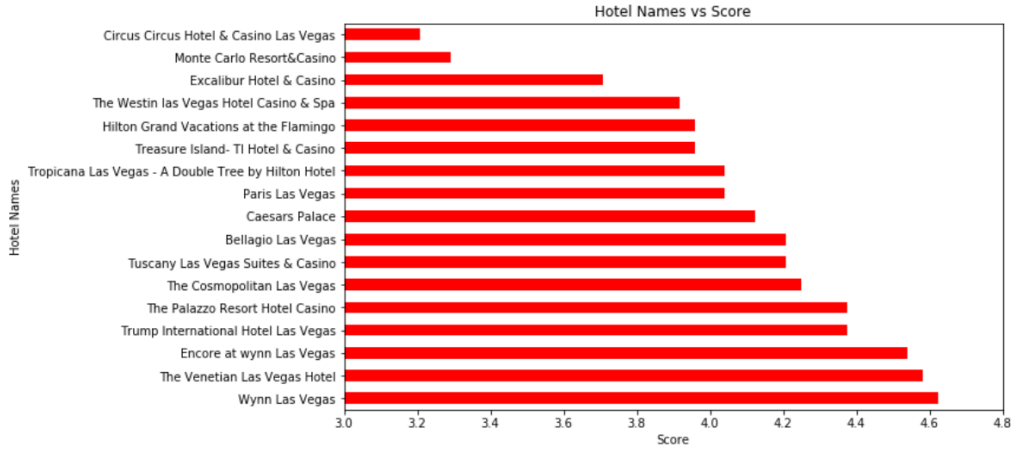
the Helpful votes. It clearly depicts that it is positively correlated. The maximum number of reviews are between 0 to 200. The bar graph at the right side illustrates the Average Number of Reviews based on Score. We can see that there are more number of reviews on 3 and 4 scored hotels, instead of a 5 scored hotel. Hence, the 5 cored hotels might need improvement to increase their reviews.

Reviews on Weekdays based on Scores:



The above Bar plot illustrates the Reviews on Weekdays which are based on Scores. By knowing the result, we can say that maximum number of Scores are given on SATURDAY. And the least scores are given on TUESDAY.

Average Review Scores for Hotel Names:

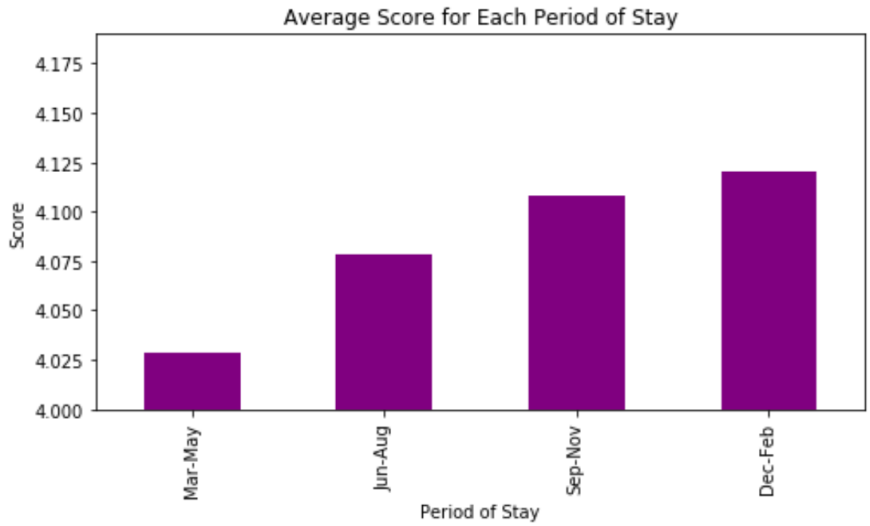


The above bar graph represents the Hotel Names with their highest Scoring. By the analysis, we can say that WYNN LAS VEGAS Hotel scores the maximum when compared to other hotels which makes it the most attractive and liveable hotel in Las Vegas. It can be known by calculating the Average scoring with all the hotels, and it gives the result of highest scoring hotel.

2.3 QUESTION:

What is the ideal time period to Stay in Las Vegas?

Below is the graph which depicts the Period of Stay in Las Vegas based on Scores. We can see that the maximum number of Scores refer to period from December to February with more than 4 scores. Hence, most of the people tend to stay during this period to enjoy the Christmas and New Year and as Las Vegas is located near hot dessert, the tourist can visit this place during cooler months (December to February). Therefore, December to February time period is ideal to visit and stay in Las Vegas.



TASK 3 – DATA MODELLING:

3.1 SPLITTING THE DATA:

The splitting of Data involves the following steps:

1. Importing all the necessary Libraries and Packages such as Train, test and Split to help in splitting the data into training and testing datasets.
2. Splitting is divided onto x and y for both Training and Testing.
3. In the X1 column, except Score all the columns are taken.
4. In Y1 column, only the Score column in taken.
5. Now, the X1 and Y1 are split into x_train, y_train, x_test, and y_test using train_test_split.
6. After passing the Train and test dataset, we must set the size of test based on splitting.
7. For **Suite1:** 50% for training and 50% for testing the test size is set to 0.5
8. For **Suite2**: 60% for training and 40% for testing the test size is set to 0.4
9. For **Suite3**: 80% for training and 20% for testing the test size is set to 0.2

3.2 CHOOSING THE MODELS:

Random Forest Classifier and Decision Tree Classifier are chosen to Train and Test the data.

3.2.1 Identification Method:

- The method used for Random Forest Classifier is "**ensemble**" i.e.,

  sklearn.ensemble import RandomForestClassifier.

- The method used for Decision Tree Classifier is "**tree**" i.e.,

  sklearn.tree import DecisionTreeClassifier.

3.2.2 Parameters Used to train Models:

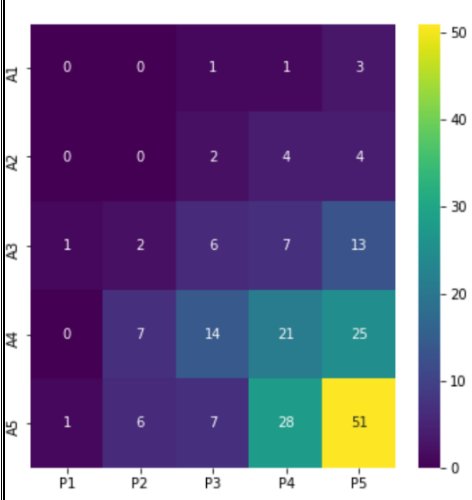
- Random Forest Classifier: The parameter used for Random Forest Classifier is "Random State" (random_state = 0) which helps in controlling the Randomness of the model. "Max_depth" parameter is kept as default.
- Decision Tree Classifier: The parameter used in Decision Tree Classifier is "Random State" (random_state = 0). All the other parameters have kept as default.

3.2.3 Evaluating the Performances of the Models:

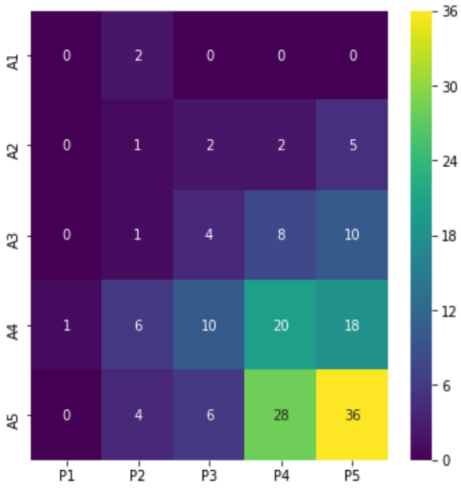
1. RANDOM FOREST CLASSIFIER MODEL:

The Confusion Matrix by Random Forest Classifier for the three Suites are given below.

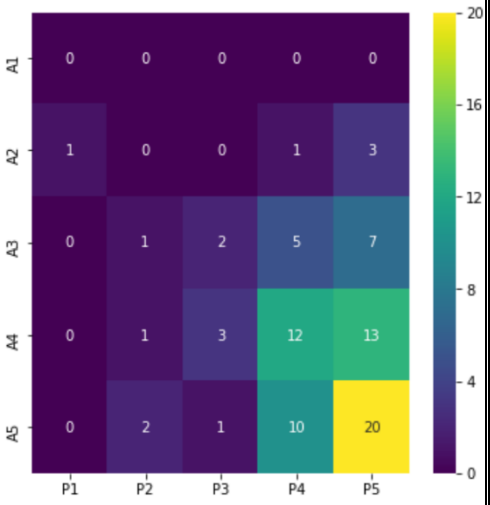
SUITE 1:



SUITE 2:



SUITE 3:



SUITE	TRAIN ACCURACY	TEST ACCURACY	SCORE	PRECISION	RECALL	F1-SCORE	SUPPORT
1	0.97	0.38					
			1	0	0	0	5
			2	0	0	0	10
			3	0.20	0.21	0.20	29
			4	0.34	0.31	0.33	67
			5	0.53	0.55	0.54	93
2	1.0	0.37					
			1	0	0	0	2
			2	0.07	0.10	0.08	10
			3	0.18	0.17	0.18	23
			4	0.34	0.36	0.35	55
			5	0.52	0.49	0.50	74
3	0.97	0.41					

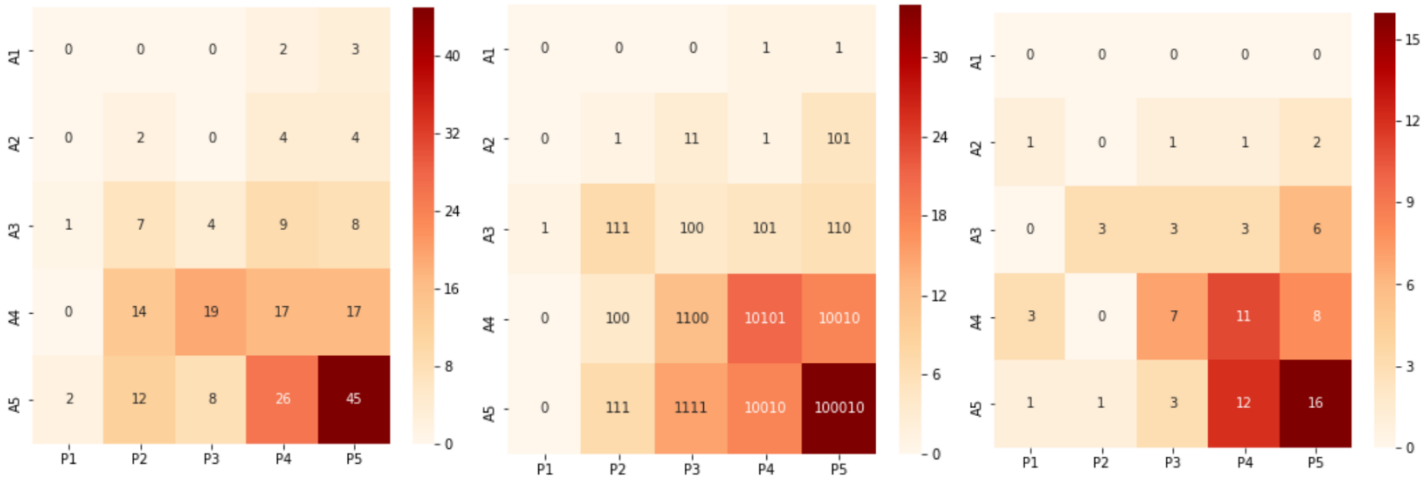
			1	0	0	0	0
			2	0	0	0	5
			3	0.33	0.13	0.19	15
			4	0.43	0.41	0.42	29
			5	0.47	0.61	0.53	33

The below table consists of the data acquired from the Confusion Matrix such as Training Accuracy, Testing Accuracy, Score, Precision, Recall, F1-Score and Support for all the three Suites.

2. **DECISION TREE CLASSIFIER MODEL:**

The Confusion Matrix by the Decision Tree Classifier for the three suites are given below.

SUITE 1: **SUITE 2:** **SUITE 3:**



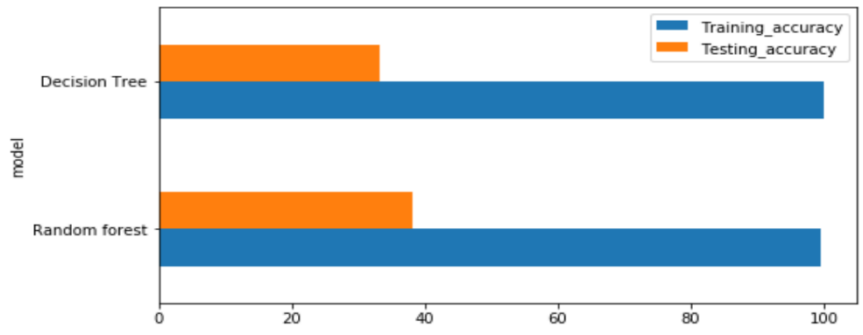
The below table consists of the data acquired from the Confusion Matrix such as Training Accuracy, Testing Accuracy, Score, Precision, Recall, F1-Score and Support for all the three Suites.

SUITE	TRAIN ACCURACY	TEST ACCURACY	SCORE	PRECISION	RECALL	F1-SCORE	SUPPORT
1	1.0	0.33					
			1	0	0	0	5
			2	0.06	0.20	0.09	10
			3	0.13	0.14	0.13	29
			4	0.29	0.25	0.27	67
			5	0.58	0.48	0.53	93
2	1.0	0.37					
			1	0	0	0	2
			2	0.05	0.10	0.07	10
			3	0.12	0.17	0.14	23
			4	0.46	0.38	0.42	55
			5	0.53	0.46	0.49	74
3	1.0	0.36					
			1	0	0	0	0
			2	0	0	0	5
			3	0.21	0.20	0.21	15
			4	0.41	0.38	0.39	29
			5	0.50	0.48	0.49	33

3.3 COMPARISION OF MODELS:

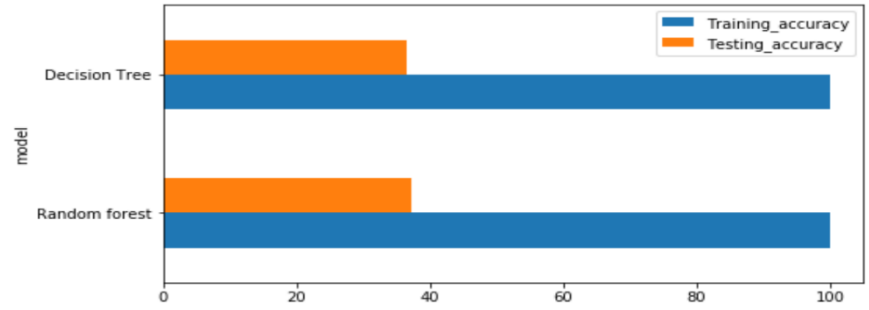
SUITE 1:

```
           model  Training_accuracy  Testing_accuracy
0  Random forest              99.51             38.24
1  Decision Tree             100.00             33.33

<matplotlib.axes._subplots.AxesSubplot at 0xbd395d0>
```



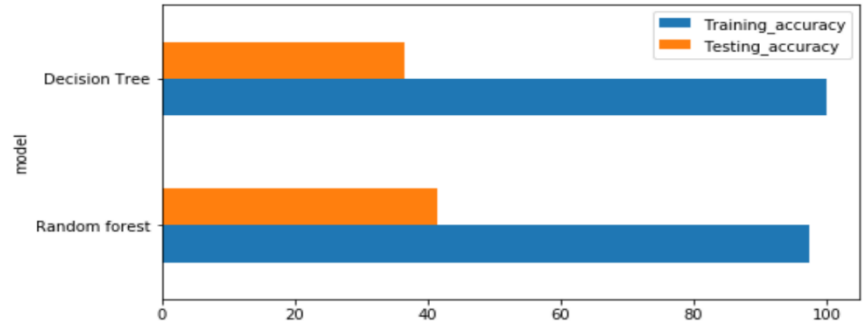
SUITE 2:

```
           model  Training_accuracy  Testing_accuracy
0  Random forest              100.0             37.20
1  Decision Tree              100.0             36.59

<matplotlib.axes._subplots.AxesSubplot at 0xdacb850>
```



SUITE 3:

```
           model  Training_accuracy  Testing_accuracy
0  Random forest              97.55             41.46
1  Decision Tree             100.00             36.59

<matplotlib.axes._subplots.AxesSubplot at 0xdc3ccb0>
```



The above Bar graphs are used to compare between two models. By comparing all the three suites, it clearly illustrates that Training Accuracy is higher than the Testing Accuracy. Both the classifier models show equal accuracy, but Decision Tree shows moderately more than the Random Forest.

DISCUSSIONS AND CONCLUSION:

Now a days, Online feedbacks are very important to rate something as it can be impactful. Here, the reviews mean a lot to people who tend to visit Las Vegas. All the hotels are high in quality and score as they have a lot of services to offer. From the analysis, we can say that Couples visit are maximum during weekends as per the reviews. Also, the ideal period of stay is between December and February. There might be good chance to improve the score (reviews) by maintaining the Hotel star (ratings). The modelling was conducted with Random Forest and Decision Tree classifiers by splitting the data into Training and Testing datasets. By the result of modelling, we can say that Decision Tree was rather good than Random Forest, even with low accuracy. Moreover, the hotels are in between 3 to 5 stars as per reviews and the services, it is adequate to amplify the journey before travelling.