**ANALYSIS OF HOTEL BOOKINGS**

**By**

**Haritha Nallam -11670566**

**Priyanka Malkapet -1172775**

**Pravallika Obulapuram - 11716198**

**Dinesh Kumar Reddy Yeduguri - 11664333**

**University of North Texas**

**Final Project**

**ADTA 5130 - Data Analytics I**

**Dr. Sarah Quintanar**

**Contents**

Problem Statement and Hypothesis………………………………………………………...3

Dataset……………………………………………………………………………………...4

Data Preparation……………………………………………………………………………5

Null Values…………………………………………………………………………5

Dummies……………………………………………………………………………6

Basic Statistics and Visualizations………………………………………………………….7

Approach……………………………………………………………………………………15

Problem Statement 1………………………………………………………………..16

ANOVA…………………………………………………………………………….16

Problem Statement 2………………………………………………………………..18

Regression………………………………………………………………………….18

Discussion…………………………………………………………………………………..23

Limitations………………………………………………………………………………….24

**Introduction and Business Understanding**

**Introduction:**

This project aims to analyze hotel booking data to produce valuable insights to understand the booking criteria, maximize profits, and reduce booking cancellations. This helps the hotel industry to enhance customer satisfaction by using different statistical approaches and comparing different variables present in the data.

**Business Understanding:**

In this analysis, the first phase is to understand the business. Here we are dealing with the business of hotel bookings. We are studying how the booking patterns vary over time. In this data set, we are focused on the Resort Hotel and City Hotel. This dataset contains 119,390 booking entries and 32 variables, which can provide detailed information about each booking.

**Research Questions and Hypothesis:**

1. What is the relationship between booking cancellation and lead time?

Null Hypothesis (Ho): There is no interaction between the lead time of booking and the cancellation rate.

Alternate Hypothesis (Ha): There is an interaction between the lead time of booking and the cancellation rate.

1. Predict which bookings are more likely to be cancelled and forecast the demand.

Null Hypothesis (Ho): The predictive model is no better than random chance at predicting booking cancellations.

Alternate Hypothesis (Ha): The predictive model performs significantly better than random chance.

3. What impact do market segments and distribution channels have on cancellation rates?

Null Hypothesis (Ho): There is no interaction in cancellation rates across customer market segments.

Alternate Hypothesis (Ha): There is an interaction in cancellation rates depending on the market segment.

4. Which type of customers usually cancel their bookings? Perform Customer segmentation.

Null Hypothesis (Ho): There is no significant difference in cancellation rates between customer types.

Alternate Hypothesis (Ha): Cancellation rates differ significantly between customer types.

5. How do booking patterns vary throughout the year?

Null Hypothesis (Ho): There is no significant variation in booking patterns throughout the year.

Alternate Hypothesis (Ha): There is a significant variation in booking patterns throughout the year.

**Data Understanding**

The variables in the data set are as below:

|  |
| --- |
| Hotel – It gives the information on the type of hotel |
| is\_canceled – This variable tells whether the booking was cancelled or not |
| lead\_time – It is the number of days between the booking date and arrival date |
| arrival\_date\_year – It gives the year of the customer's arrival date to the hotel |
| arrival\_date\_month – It gives the month of the customer's arrival date to the hotel |
| arrival\_date\_week\_number – It gives the week number of the customer's arrival date to the hotel |
| arrival\_date\_day\_of\_month – It gives the day of the month of the customer's arrival date to the hotel |
| stays\_in\_weekend\_nights – It gives the number of weekend nights customers stayed. |
| stays\_in\_week\_nights – It gives the number of weekday nights customers stayed. |
| Adults/Children/Babies – Gives information on the number of adults, children and babies respectively stayed. |
| Meal – It shows the type of food booked by customers. |
| Country – Information of country they are from. |
| market\_segment – It gives how the bookings are made, market segment. |
| distribution\_channel – It gives the information on from which channel bookings are made |
| is\_repeated\_guest – It tells whether the guest is revisiting or a new guest. |
| previous\_cancellations – To find any previous cancellations done by that guests |
| previous\_bookings\_not\_canceled – To find any previous bookings not cancellations done by those guests |
| reserved\_room\_type – Type of room booked or reserved |
| assigned\_room\_type – Type of room assigned to the guest |
| booking\_changes – Number of changes made to that booking |
| deposit\_type – Type of deposit made by guests like non-refund, no deposit, refund. |
| Agent – This shows which travel agent has booked |
| Company – This gives which company has done this booking |
| days\_in\_waiting\_list – Number of days this booking is under the waiting list |
| customer\_type – This gives information on the type of customer |
| Adr – This is the Average Daily Rate of the booking. Average per day cost. |
| required\_car\_parking\_spaces – Total number of parking spaces need by a guest. |
| total\_of\_special\_requests – Total count of special requests made by the guest. |
| reservation\_status – Status of the reservation or booking |
| reservation\_status\_date – The date on which the above status was noted. |

**Data Preparation**

In this phase, we do data cleaning, data manipulation wherever needed if there are any abnormalities in the data. Data is prepared so that it will be suitable for designing models and analysis.

**To find Null Values**

We need to check the presence of Null values in the dataset. These Null values are the blank values that are not present in the dataset for many reasons and can be a problem for the Analysts. The Null values must be handled by removing them or replacing them in the dataset.

Let’s check for the Null values using Python by running the ‘**isnull().sum()**’ function.

A screenshot of a computer

Description automatically generated

We found that the variables, children, country, agent, and company have Null values. These must be handled before proceeding further. We follow different strategies for handling missing values. This depends on the percentage of missing data and the significance of the variable. Columns with high percentage of missing values may drop that variable if it is not vital.

Here we have a high missing percentage of 94% and 14% for the “company” and “agent” variables respectively. As these features are not critical, we can remove these variables.

**Dummies**

Since we have 2 categorical variables ‘Integrated Wireless?’ and ‘Bundled Applications?’ with ‘Yes’ or ‘No’ as their values, we need to change those values into dummies which will help in better analyzing. We need to change these categorical values into numerical dummies which will help in developing a proper regression model.

We will create dummy variables for ‘Integrated Wireless?’ and ‘Bundled applications?’ using Excel.

We have used the formula “**=IF(F2="Yes",1,0)”** for ‘Integrated Wireless’ and **“=IF(H2="Yes",1,0)”** for ‘Bundled Application?’ in new columns and got the required dummies.

Graphical user interface

Description automatically generated

We can see that the new columns ‘Integrated Wireless’ and ‘Bundled Applications’ are created with dummies.

We will properly arrange the variables and remove the categorical variables and keep only the new dummy variables created.

Below is the final prepared data.

Table

Description automatically generated

**Basic Statistics and Visualizations**

Now, we’ll be looking at different descriptive statistics of various variables present in the dataset. I am using Python for statistical values and visualizations.

**Configuration**

**Table

Description automatically generatedGraphical user interface, text, application

Description automatically generated**There are a total of 864 unique values for ‘Configuration’ in the dataset starting from 1. These 864 different ‘Configuration’ values are divided between 99999 values based on the laptop.

The mode of the ‘Configuration’ variable in 61, means that the ‘Configuration’ type 61 has occurred mostly in the dataset.

**Screen Size (Inches)**

**Graphical user interface, text, application

Description automatically generated**

We can see that the minimum value of ‘Screen Size’ is 15 inches and the maximum value of ‘Screen Size’ is 17 inches.

Chart, bar chart

Description automatically generated

We can see that around 70,000 laptops have 15 inches of display, and the remaining 30,000 laptops are with 17 inches of display.

**Battery Life**

**Graphical user interface, table

Description automatically generated**

There are 3 different values present in the ‘Battery Life’ variable. 6 hours is the most occurred value in the ‘Battery Life’.

The count plot for the ‘Battery Life (Hours)’ is shown below.

Chart, bar chart

Description automatically generated

**RAM (GB)**

**Graphical user interface, application

Description automatically generated**

There are 3 unique values for ‘RAM (GB)’ which are 4, 8, and 16. 8 GB of RAM has the maximum count in the dataset.

The count plot for ‘RAM’ is shown below.

Chart, bar chart

Description automatically generated

**Processor Speeds (GHz)**

**Graphical user interface

Description automatically generated with medium confidence**

There are 3 unique values for ‘Processor Speeds (GHz)’ which are 1.5, 2, and 2.4 GHz. Processor Speed of 2 GHz has the highest count in the dataset.

The count plot for the ‘Processor Speeds (GHz) is shown below.

Chart, bar chart

Description automatically generated

**HD Size (GB)**

**Table

Description automatically generated**

There are 4 unique values present in the ‘HD Size (GB)’ variable and they are 40, 80, 120, and 320. The 120 GB HD size has the highest count in the dataset.

The count plot for the ‘HD Size (GB)’ is shown below.

Chart, bar chart

Description automatically generated

**Integrated Wireless?**

**Chart, bar chart

Description automatically generated**

The newly created variable using dummies ‘Integrated Wireless?’ has 2 values 1 and 0, in which ‘1’ shows it is Integrated Wireless and ‘0’ shows not Integrated Wireless. The value count of ‘1’ is higher which tells that there are more Integrated Wireless laptops in the dataset.

**Bundled Applications?**

**Chart, bar chart

Description automatically generated**

The newly created variable using dummies ‘Bundled Applications?’ has 2 values 1 and 0, in which ‘1’ shows it has Bundled Applications and ‘0’ shows do not have Bundled Applications. The value count of ‘1’ is higher which means that the count of laptops with Bundled Applications is higher in the dataset.

**Price**

**Table

Description automatically generated**

The dependent value Price has a range of values. The minimum value of the laptop Price is 1000 and the maximum value is 1890.

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated  
The above box plot shows the median Price is around 1500. The box plot also shows the lower quartile and upper quartile.

The histogram plot shows the distribution of the Price values.

**Approach**

Now we will be working on different statistical techniques to work on our problem statements and find the best statistical answer.

**Problem Statement 1**

To find whether there is any significant difference between the mean prices of the laptops based on whether they have any Bundled Applications in them, we will collect a sample of data from the dataset which will tell the laptop price based on the Bundled Application variable.

We can add filters and collect sample data.

I have taken 60 samples of laptop prices for each for Bundled Applications -Yes and Bundled Applications - No.

Table

Description automatically generated with low confidenceBelow is the sample data.

Table

Description automatically generated

Table

Description automatically generated

Now we will use **ANOVA- One Factor Analysis** to find out whether there is any significant difference between mean laptop prices.

**ANOVA- Single Factor**

Analysis of Variance or ANOVA is a method that is used to compare the means of two or more group values. In a Single Factor ANOVA there is only one independent group present for comparison.

Different statistical comparisons can be done using an ANOVA table using Microsoft Excel or R Studio.

**Steps for ANOVA: Single Factor**

* Add Analysis Tool Pack in the Microsoft Excel.
* Click on ‘Data’.
* Select ‘Data Analysis’ present in the top right corner.
* Select ANOVA: Single Factor.
* Enter the input range.
* Check the box if the text labels are present.
* Enter the significance level.
* Select the cell for the output.
* Click OK

For the above sample data for Problem Statement 1, the ANOVA table is obtained using Microsoft Excel.

**Application, table, Excel

Description automatically generated**

We have obtained different statistics from our sample.

But for the problem statement, we will be focusing on the F-statistics value, P-value, and the F-Critical value.

The **F- Statistical value** is **~8.6** which is greater than the **F- Critical value** which is **3.92**.

Since our **significance leve**l is **95%** our **Alpha value** is **0.05**

The P-value obtained is **0.0040** which is less than the Alpha value which is **0.05**.

Since **F- Statistic > F-Critical and P-Value < Alpha**, we can reject our Null Hypothesis and accept the Alternate Hypothesis.

**We can conclude that there is a difference between the means of laptop prices based on the presence of Bundled Applications. This means that the mean laptop prices for in-built Bundled Applications are comparatively higher than the laptops that don’t have any Bundled Applications. This conclusion cannot have been drawn by just looking at the data, as the prices were almost similar. So, ANOVA helped in getting a proper conclusion on Mean Laptop prices for Bundled Applications.**

**Problem Statement 2**

To predict the laptop price which is a dependent variable based on different independent variables such as Screen Size, Battery Life, RAM, Processor Speed, Integrated Wireless, Hard Disk Size, and Bundled Applications.

We will be using the Multiple Regression technique to predict laptop prices. This can be done using Microsoft Excel.

**Regression**

Regression is a statistical method that gives the relationship between one dependent variable and one or more independent variables. A Regression model tells us that the change in the dependent variable is due to the change in independent variables.

The dependent variable is denoted by ‘Y’ and the independent variables are denoted by X1, X2, X3,…….

The prediction of the dependent variable ‘Y’ is given by the below equation

**Y = Intercept + (Coefficient of X1) \*X1 + (Coefficient of X2) \*X2 + (Coefficient of X2) \*X2……**

The values of Intercept, X1, X2 can be achieved by Regression Table using Analysis tool pack in Microsoft Excel.

For our problem statement,

**Steps for Regression**

* Add Analysis Tool Pack in the Microsoft Excel.
* Click on ‘Data’.
* Select ‘Data Analysis’ present in the top right corner.
* Select Regression
* Select the input range.
* Check the labels.
* Enter the significance level.
* Select the cell for Output.
* Click OK.

For our problem statement, we are considering Screen Size, Battery Life, RAM, Processor Speed, Integrated Wireless, Hard Disk Size, and Bundled Applications as our independent variables. We are supposed to enter the respective cell values of these independent variables in the Regression analysis.

Below is the regression table for our data.

Table

Description automatically generated

We will be focusing mainly on the R Square value, Intercept, and the Coefficients of independent variables for our prediction.

**R Square** value of **0.31** which is also known as the coefficient of determination, as it explains how good the regression model is. 0.31 R Square value is moderate and tells that there is a slight effect on the dependent variable.

31% of the R Square value means 31% of Laptop Prices can be predicted by our independent variables.

So, the general **Regression equation** for our data is,

**Price=259.032 + 46.730\*(Screen Size (Inches)) + 46.581\*(Battery Life (Hours)) + 11.223\*(RAM GB) + 46.557\*(Processor Speeds (GHz) +19.06\*(Integrated Wireless?) + 0.39\*(HD Size (GB)) + 47.544\*(Bundled Applications?)**

From the Regression Equation, we can interpret that,

* Price is determined by Screen Size (Inches) with an increase in the factor of 46.730
* Price is determined by Battery Life (Hours) with an increase in the factor of 46.581
* Price is determined by RAM (GB) with an increase in the factor of 11.223
* Price is determined by Processor Speeds (GHz) with an increase in the factor of 46.557
* Price is determined by Integrated Wireless with an increase in the factor of 19.06
* Price is determined by HD Size (GB) with an increase in the factor of 0.39
* Price is determined by Bundled Applications with an increase in the factor of 47.544

Since we have Categorical Variables in our dataset, this regression equation will change according to the values of these categorical variables. Since, Yes =1, No =0 based on the dummy values present.

If the value of Integrated Wireless is Yes and Bundled applications is Yes. Then the Regression equation will change to

**Price=259.032 + 46.730\*(Screen Size (Inches)) + 46.581\*(Battery Life (Hours)) + 11.223\*(RAM GB) + 46.557\*(Processor Speeds (GHz) +19.06 + 0.39\*(HD Size (GB)) + 47.544**

If the value of Integrated Wireless is Yes and Bundled applications is No. Then the Regression equation will change to

**Price=259.032 + 46.730\*(Screen Size (Inches)) + 46.581\*(Battery Life (Hours)) + 11.223\*(RAM GB) + 46.557\*(Processor Speeds (GHz) +19.06 + 0.39\*(HD Size (GB))**

If Integrated Wireless is No and Bundled applications is Yes. Then the Regression equation is given by

**Price=259.032 + 46.730\*(Screen Size (Inches)) + 46.581\*(Battery Life (Hours)) + 11.223\*(RAM GB) + 46.557\*(Processor Speeds (GHz) + 0.39\*(HD Size (GB)) + 47.544**

If Integrated Wireless is No and Bundled applications is No. Then the Regression equation is given by

**Price=259.032 + 46.730\*(Screen Size (Inches)) + 46.581\*(Battery Life (Hours)) + 11.223\*(RAM GB) + 46.557\*(Processor Speeds (GHz) + 0.39\*(HD Size (GB))**

Now we will check how close is the predicted price from the regression equation to the value from the dataset price.

For a laptop, Screen Size (Inches) = 15, Battery Life (Hours) = 5, RAM (GB) = 8, Processor Speeds (GHz) = 2.4, Integrated Wireless? = Yes, HD Size (GB) = 120, Bundled Applications? = No

The Price of the dataset is **1490.**



The Price predicted using the Regression Equation is **1459.543**

**Price=259.032 + 46.730\*(15) + 46.581\*(5) + 11.223\*(8) + 46.557\*(2.4) +19.06(1) + 0.39\*(120) = 1459.543**

We can see that the price from the dataset and the predicted price are slightly different but do not have a big difference. Because of our low R Square value, we are getting the difference in actual and predicted prices.

Now, let’s predict the laptop price for Screen Size (Inches) = 18, Battery Life (Hours) = 8, RAM (GB) = 16, Processor Speeds (GHz) = 2.4, Integrated Wireless? = Yes, HD Size (GB) = 360, Bundled Applications? = Yes

**Price=259.032 + 46.730\*(18) + 46.581\*(8) + 11.223\*(16) + 46.557\*(2.4) +19.06\*(1) + 0.39\*(360) + 47.544\*(1) = 1968.936**

We have predicted that the laptop price for the above specifications would be around **1968.936**

**Discussion**

Based on the ANOVA results, we can clearly see that there was a difference between the laptop prices based on the presence of Bundle Applications present in it. It clearly tells us that the Bundled Applications play an important role in determining the laptop price.

Based on the Regression results, we are easily able to predict the laptop prices from the model we built based on the configuration and specifications we need in our laptop. The Linear Regression model was moderate in predicting the laptop prices. There was a certain relationship between the independent variables and dependent variables by which we were able to predict the laptop prices. More data preprocessing could help in making a better model which would help in predicting laptop prices more accurately.

**Limitations**

There are certain limitations to this dataset. The variables used are very less to determine the laptop price. There are other variables like weight, graphics, brand, operating system used, and touchscreen options that might also help in considering the laptop price. Therefore, only some basic variables might not alone help in predicting the laptop price.