**Project Report: Hotel Management System**

A green logo with a black background

Description automatically generated

***TEAM*: DEEPAK RAMBARKI (11703304)**

**MANEESH GEMBALI (11714953)**

**RENU SREE NAGIRIMADUGU (11667577)**

**CHANDANA ADLA (11664486)**

**VENKATLAXMAN PAVAN ATHI (11663816)**

**Content:**

1. Business Understanding
2. EDA/ Data Understanding & Data Preparation
3. Cancellation Trends by Month:
4. Customer Types and Their Meal Preferences:
5. Distribution of Total Length of Stay (up to 2 weeks):
6. Number of Bookings by Market Segment:

6.1 ANOVA Model – The effect of the meal type on Lead time:

6.2 ANOVA Model – The effect of market segment on the total stay duration.

7. Linear Regression model

7.1 Based on hotel type.

* 1. Based on meal type.

1. Linear Regression model
2. Regression model

**Business Understanding / Analytics Question(s):**

In this Project we are aimed to explore and understand hotel booking data through statistical model like ANOVA and regression analysis.

**ANOVA models -** The effect of meal type on lead time, The effect of Market segment on total stay duration.

**Regression Models -**

Predicting the number of special requests based on hotel type, meal type, customer type.

Predicting the likelihood of booking being Canceled using Lead time, total stay etc.

Predicting Days in waiting List using Hotel type, customer type and the total.

**EDA/ Data Understanding & Data Preparation:**

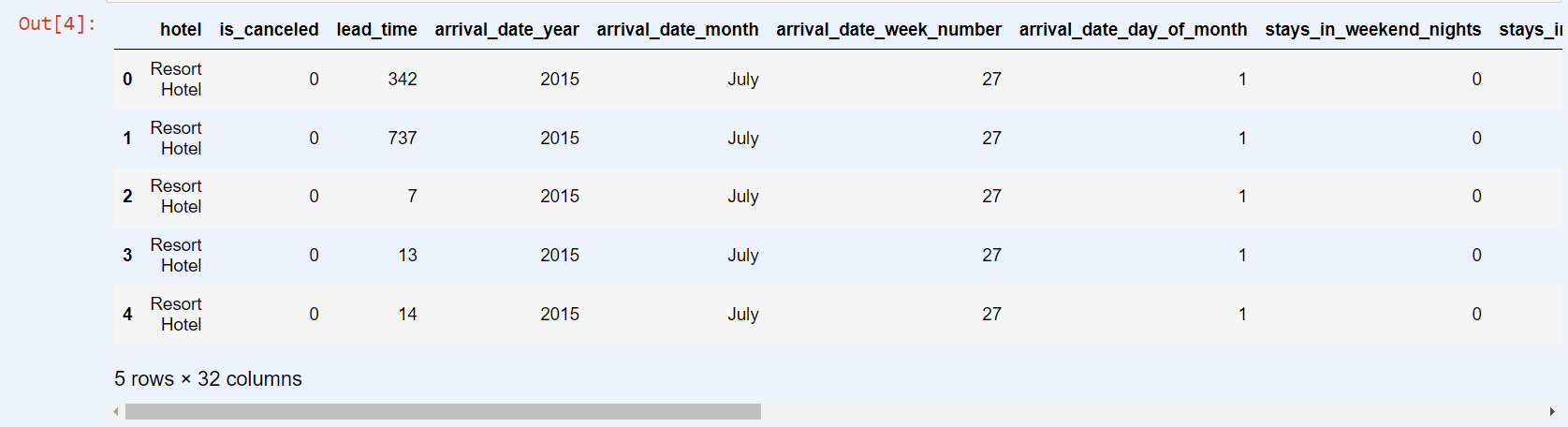
EDA is very import step in Data Analysis, To start the EDA we have imported the data into pandas dataframe, and performed the basic Statistical summary steps.

Code:

data = pd.read\_csv("hotel\_bookings.csv")

data.head()

**output:**



**Code:**

eda\_summary = {

'head': data.head(), # First 5 rows of the dataset

'describe': data.describe(include='all'), # Descriptive statistics for numeric columns

'info': data.info(), # Info about the dataset, including the data types and number of non-null values

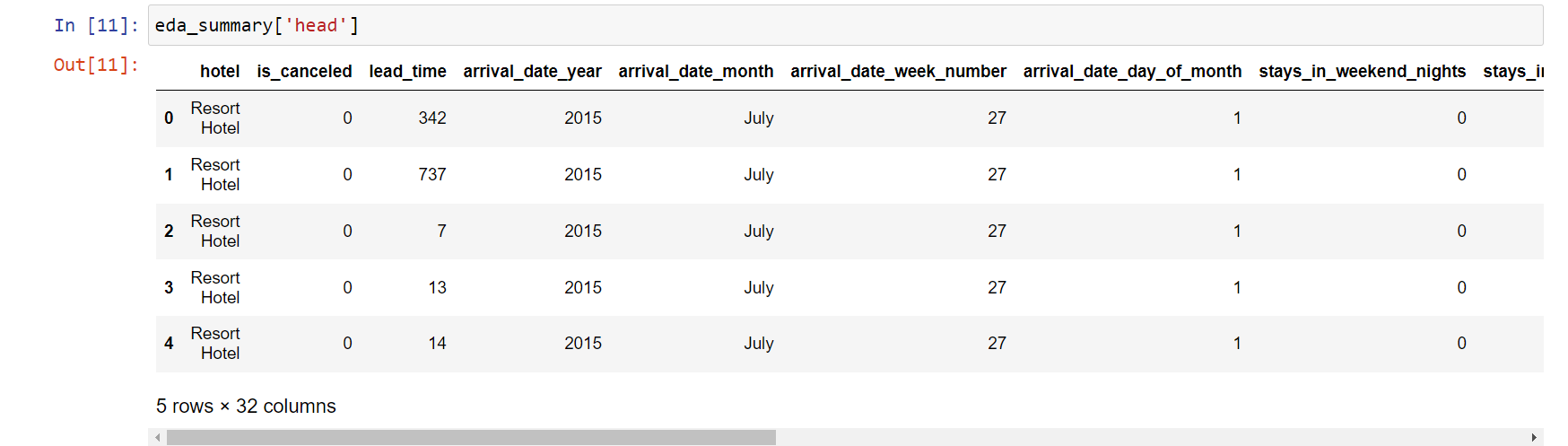
'null\_values': data.isnull().sum(), # Count of null values in each column

'unique\_values': data.nunique() # Number of unique values in each column

}

Above code is Dictionary which have crucial information like the top five rows, the statistical summary of the data set , the data types of the columns , null values and unique values. We access any of the attribute from the dictionary.

**Top 5 Rows in the dataset:**



**Null values in the Dataset:**

hotel 0

is\_canceled 0

lead\_time 0

arrival\_date\_year 0

arrival\_date\_month 0

arrival\_date\_week\_number 0

arrival\_date\_day\_of\_month 0

stays\_in\_weekend\_nights 0

stays\_in\_week\_nights 0

adults 0

children 4

babies 0

meal 0

country 488

market\_segment 0

distribution\_channel 0

is\_repeated\_guest 0

previous\_cancellations 0

previous\_bookings\_not\_canceled 0

reserved\_room\_type 0

assigned\_room\_type 0

booking\_changes 0

deposit\_type 0

agent 16340

company 112593

days\_in\_waiting\_list 0

customer\_type 0

adr 0

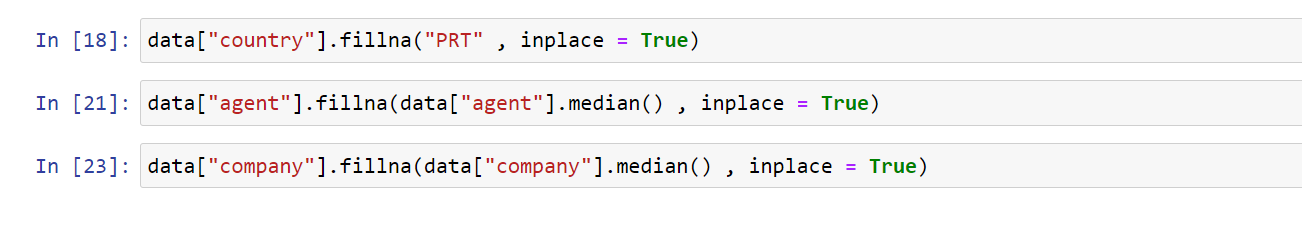
required\_car\_parking\_spaces 0

total\_of\_special\_requests 0

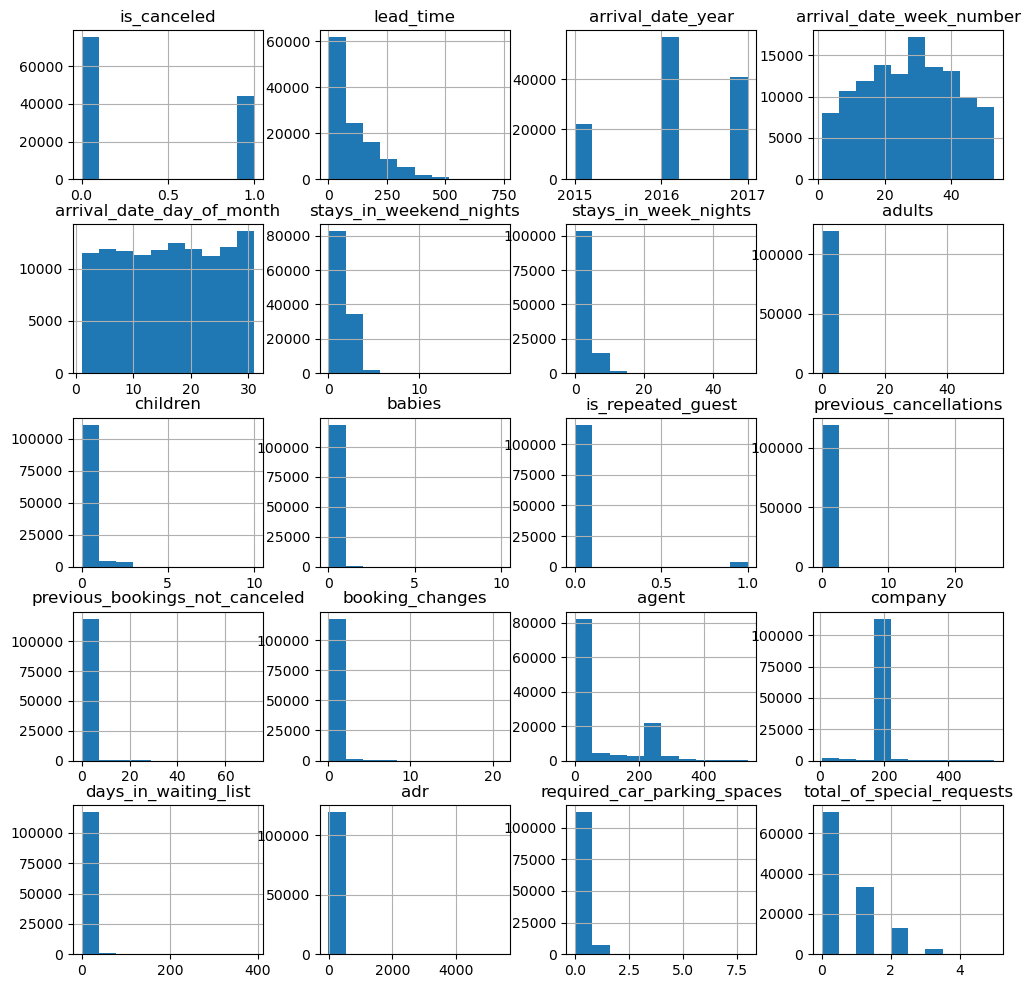
reservation\_status 0

reservation\_status\_date 0

There are a lot of missing values in the country column and we have replace missing values with the mode of the country column that is “PRT”, this method is imputation of missing values. We have performed the similar steps for the columns “agent” and “company” as well , as the “agent” and “company” are numerical values , we have replaced them with medium value.

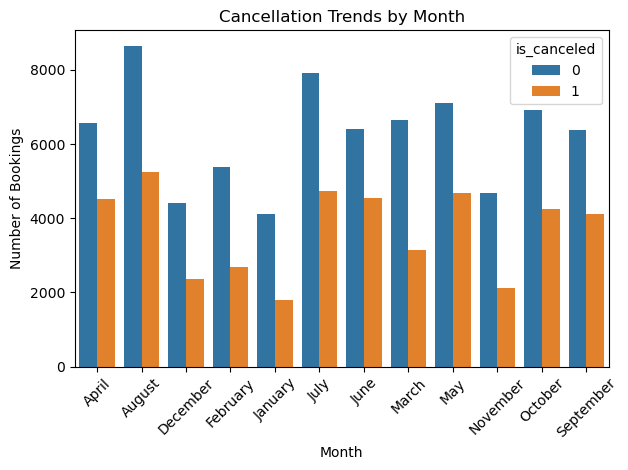


Below the histogram of the numerical columns in the dataset, by this we got know distribution of all the numerical columns. We can see that arrival\_date\_week\_number is not normal distribution similar case with the arrival\_date\_day\_of\_month. On the other hand, is canceled is binary column that mean it has only 2 value 0 and 1.



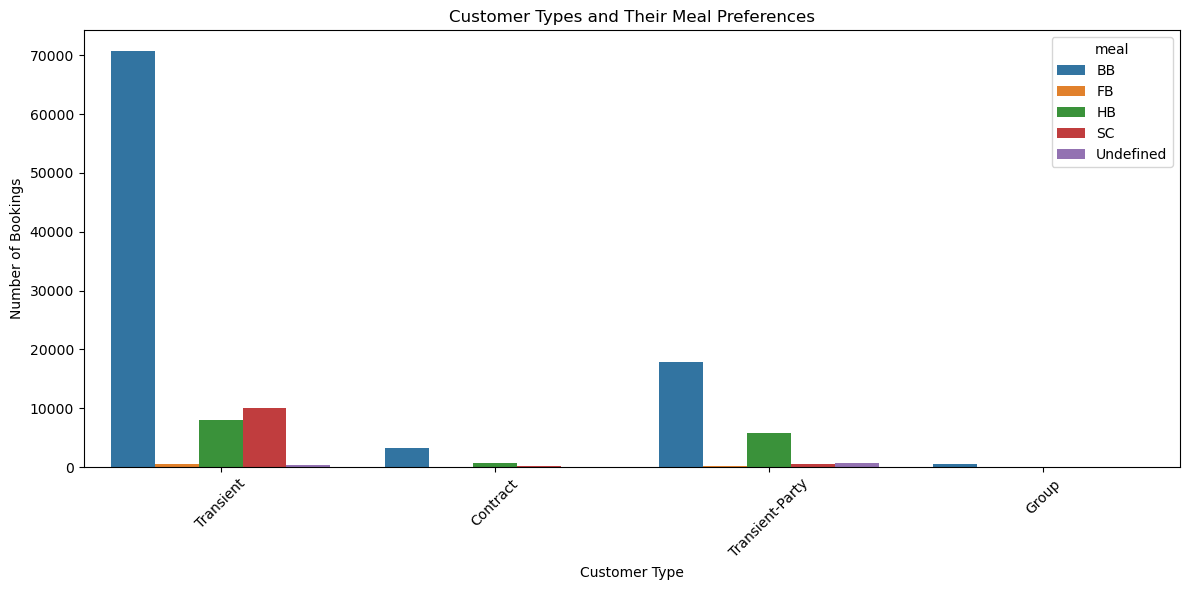
**Cancellation Trends by Month:**

Below is the bar plot for number of bookings for different months, as we can need the cancelation is max in the August then come the July and so on. At the same time the highest booking is also done in the month and August and least number of bookings are done in the month of January.



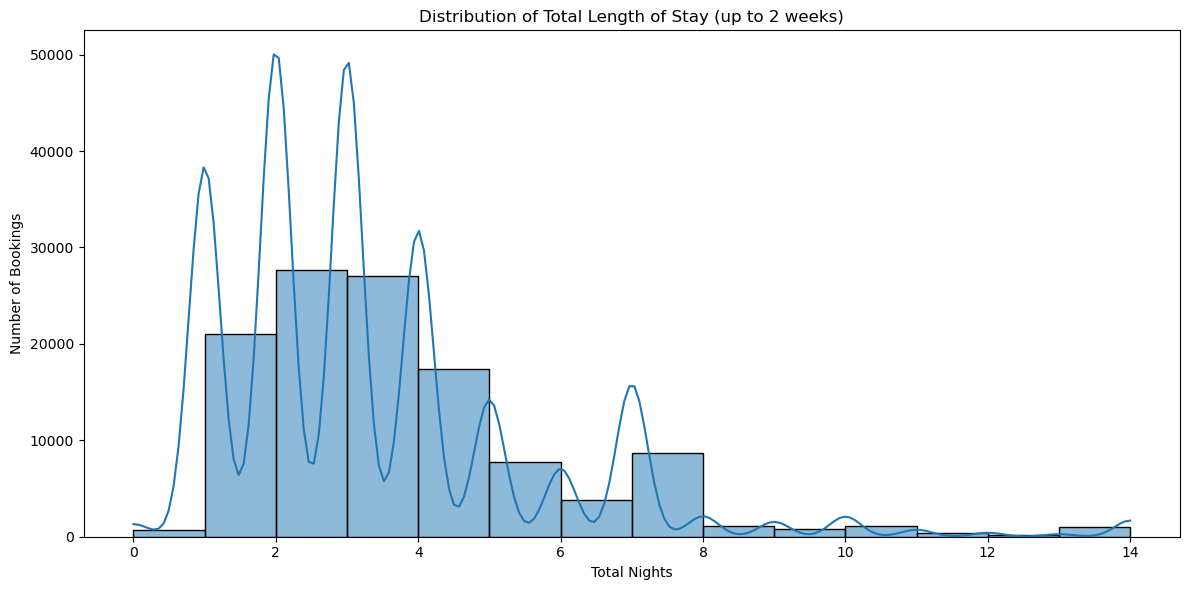
**Customer Types and Their Meal Preferences:**

Most of the customers prefer the BB type of meal and we can see that Transient type prefer the BB type meal the most and FB type the least, same goes with the customer type Contract and transient party, and the unexpected is that group type customer are not in any of the mean type.



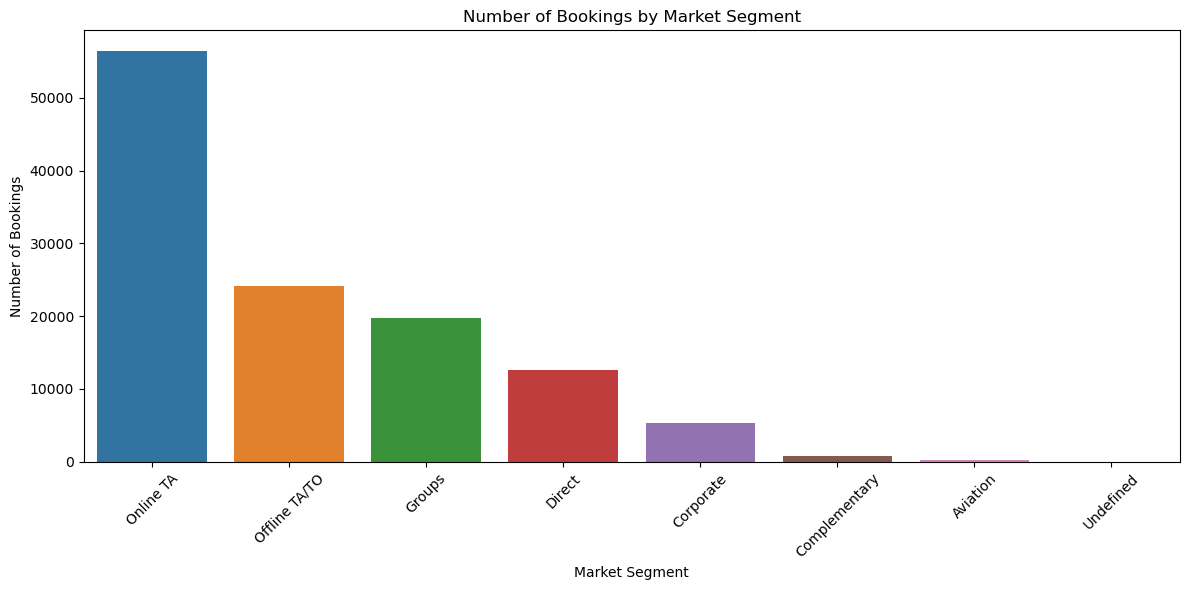
**Distribution of Total Length of Stay (up to 2 weeks):**

Below is the distribution of the Total Length stay by customer, as we can see most of the customer stay b/w 1 to 6 nights and again at 8 nights. The distribution is also not normal distribution, they are so many peaks and low in the distribution,



**aNumber of Bookings by Market Segment:**

This bar graphs display the count plots of different market segments, and we can see that Online TA is the most popular one among the customers are Aviation is the least popular one.



**ANOVA Model – The effect of the meal type on Lead time: (Methods / Modeling)**

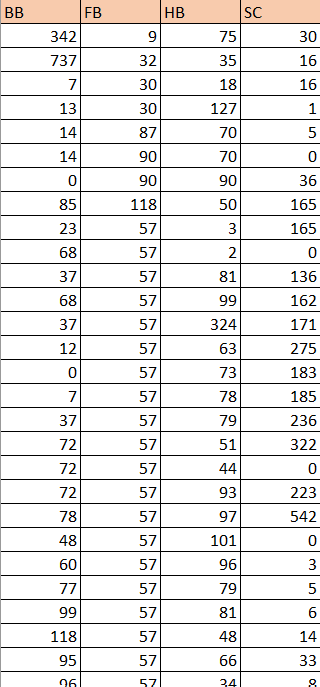
In this ANOVA test we are typing to see if there is any different in lead time by the customer based on the meal type,

If there is an effect on lead time based on preferred a meal type, we can improve the type of meal of change the meal.

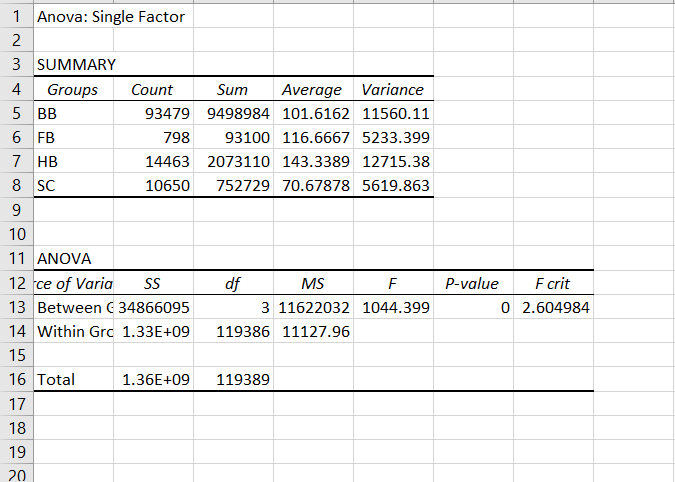
**Null Hypothesis**: – There is no difference in mean of lead time for different meal types like (BB, FB, HB , SC , undefined)

**Alternative hypothesis**: - There’s a difference in meal of lead time for different meal types (BB, FB , HB , SC , undefined)

To Perforce the ANOVA we have used excel in build ANOVA test, for use that that test the data should be in the below format.

****

**After the test we will be getting the below table: (Evaluation / Discussion)**



We can see that p value is 0 , that is less than Significance level (p < 0.05) , So we reject the , Null Hypothesis and the Alternative hypothesis. So there is difference in meal of lead time for different meal types (BB , FB , HB , SC , undefined).

After this we go to know that meal type BB customer are lead more time in hotel, So we can improve the BB meal type.

**ANOVA Model – The effect of market segment on the total stay duration.**

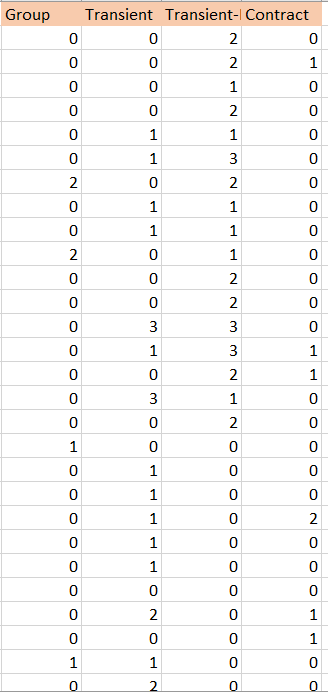
In this ANOVA test we are typing to see if there is any different in total stay duration by different market segment. (Complementary, Corporate, Direct, Groups, Offline TA/TO, Online TA, Undefined)

If there is different of the total stay duration for different market segments, then we can improve the customer of the group which have less duration.

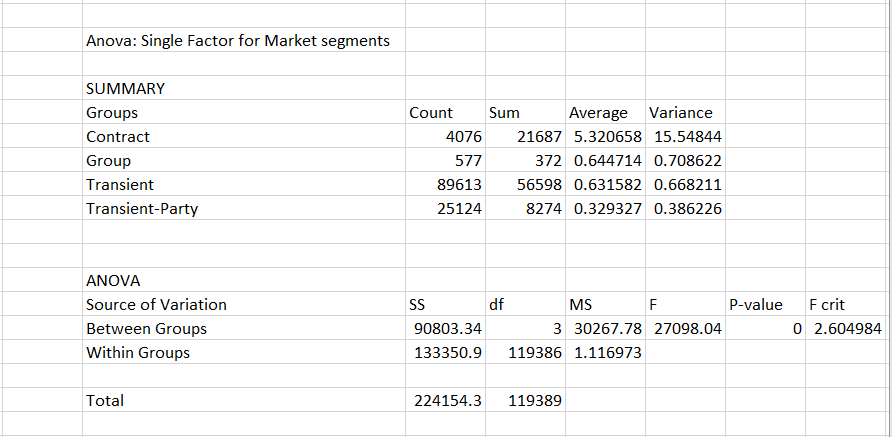
**Null Hypothesis: –** There is no difference in mean of total stay duration for different market segmentlike (Complementary, Corporate, Direct, Groups, Offline TA/TO, Online TA, Undefined)

**Alternative hypothesis: -** There is difference in mean of total stay duration for different market segmentlike (Complementary, Corporate, Direct, Groups, Offline TA/TO, Online TA, Undefined)

To Perforce the ANOVA we have used excel in build ANOVA test, for use that that test the data should be in the below format.

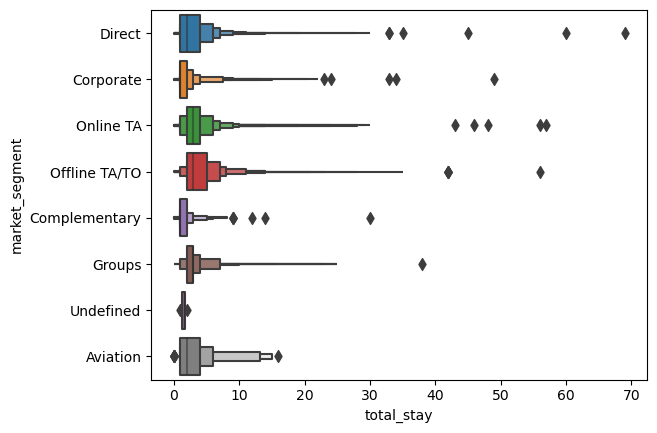


After the ANOVA test, we will be getting the below table which have the p values



Again the p value is 0, that is less than Significance level (p < 0.05) , So we reject the , Null Hypothesis and the Alternative hypothesis. So, there is difference in mean of total stay duration for different market segmentlike (Complementary, Corporate, Direct, Groups, Offline TA/TO, Online TA, Undefined)

Below is the box of market segment vs the total duration: -

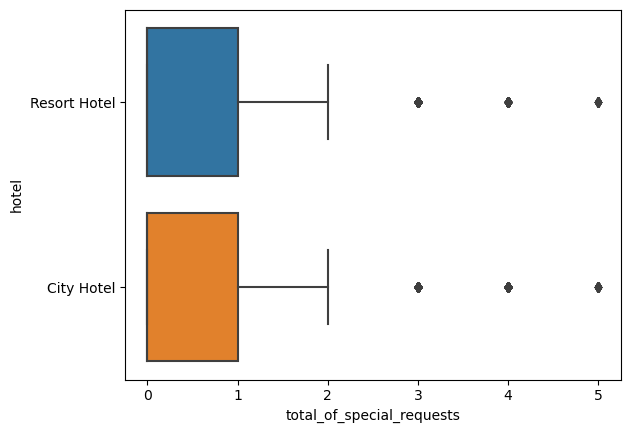


We can see that offline TA/TO has the largest total stay duration, and least is the Complementary market segment, So the hotel management need to focus on that group of customers.

**Linear Regression model:**

This model Predicts the number of special requests based on hotel type, meal type, and customer type. The request for number of bookings occurs based on the categories describes the demand in that respective category.

**Based on hotel type:**

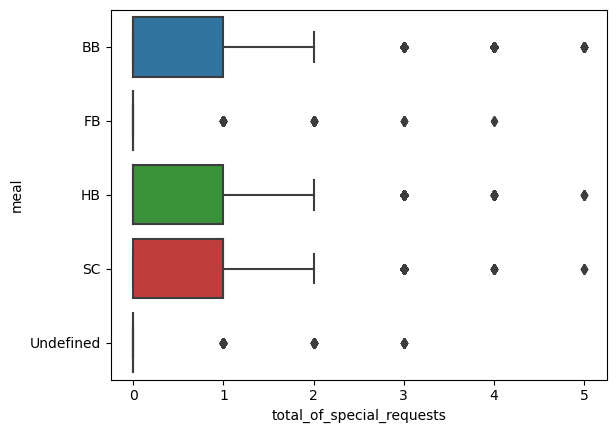


There is similar relation for getting special requests between city hotel and resort hotel which having same stats.

It also has some equal number of outliers for both city hotel and resort hotel start from number of requests three to five.

Both hotel types have 2 total highest number of requests and zero lowest number of requests. Havin upper quartile at 1 and lower quartile at 0.

**Based on meal type:**



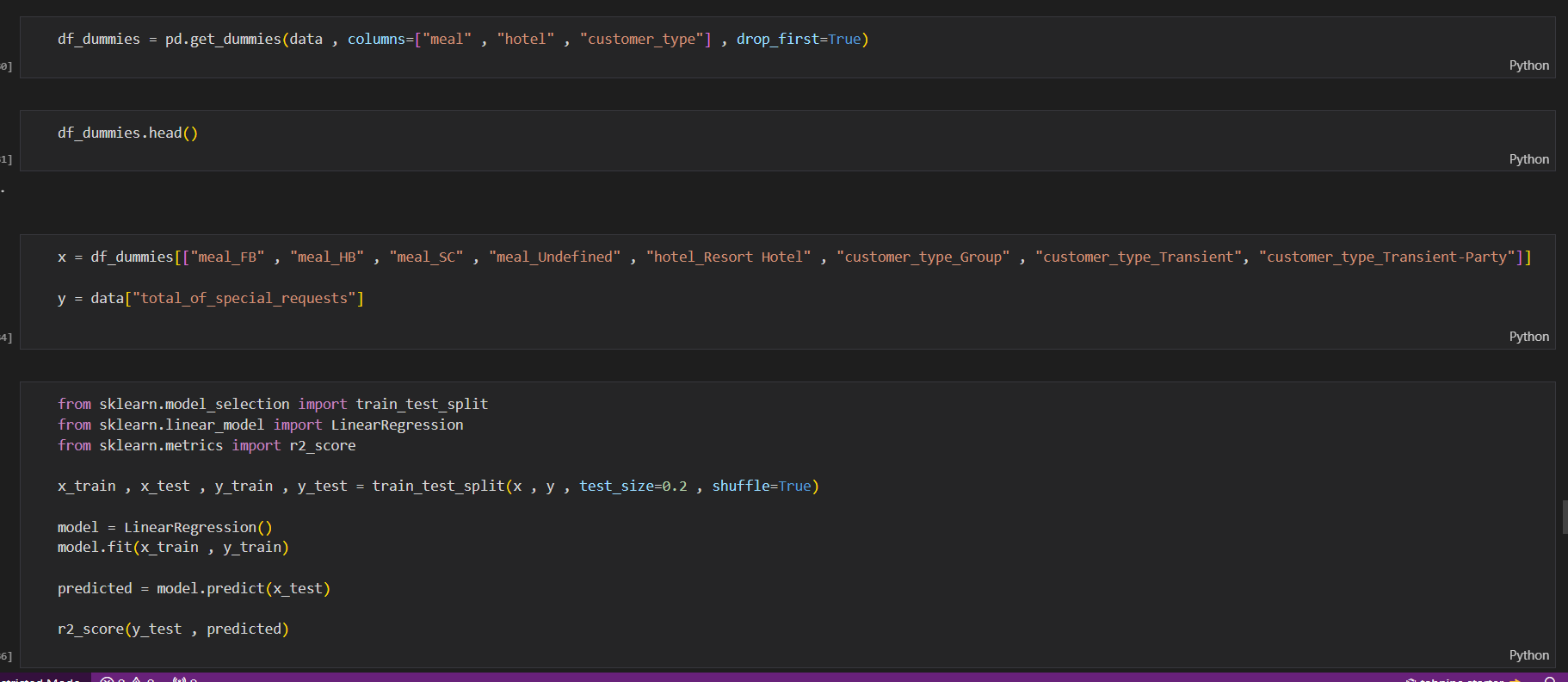
Three types of meals BB, HB, SC got same number of requests and outliers with highest of two requests and lowest of zero requests. Remaining two types of meals FB, Undefined has zero requests. Outliers for these three types range from three to five.

There are some outliers in each case of meal. Undefined and FB has highest number of outliers and zero requests.

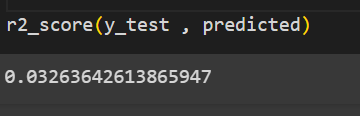
dummies data frame is created to encode categorical variables to numerical variables for regression model.

The dataset is splits in to x\_train , x\_test , y\_train , y\_test and model is fitted.

Predicted the model with x\_test data and got r2\_score value as 0.03434853987953823.

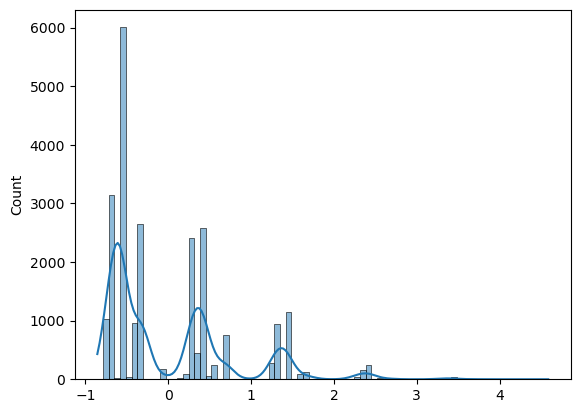


**RESULT:**



**Residual plot**:

It describes the normality of the distribution.



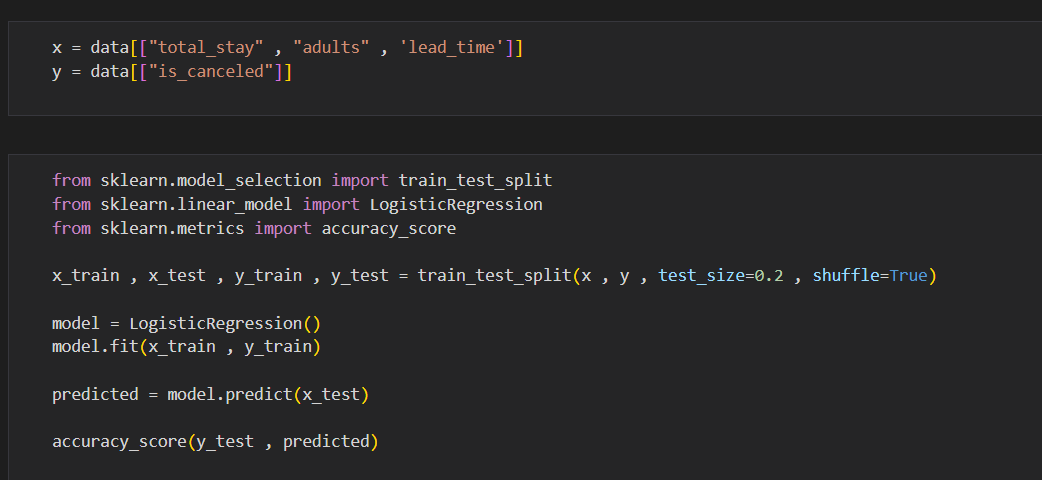
The graph clearly says that the normality is absent in the residual.

**Logistic regression model:**

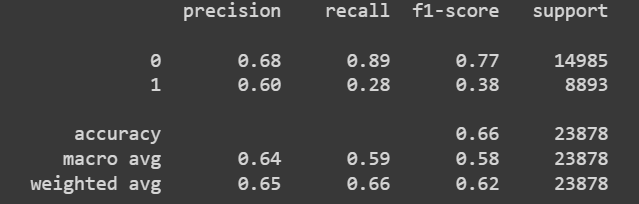
**Model:** Predicting the likelihood of a booking being canceled.

**Independent variables:** Lead time, total stay (sum of weekend and weekday stays), and the number of adults.

Model is trained with x\_train and y\_train and predicted using x\_test with accuracy of 0.6608.



**RESULTS:**



Here it’s canceled is a binary variable having 0 and 1 values. The classification report clearly says that:

For class 0 precision is 0.68, recall is 0.89, f1 sore is 0.77.

For class 1 precision is 0.60, recall is 0.28, f1 sore is 0.38.

The overall accuracy is 0.66.

Macro average precision is 0.64, recall is 0.59, and f1-score is 0.58 and for weighted average precision is 0.65, recall is 0.66, and f1-score is 0.62.

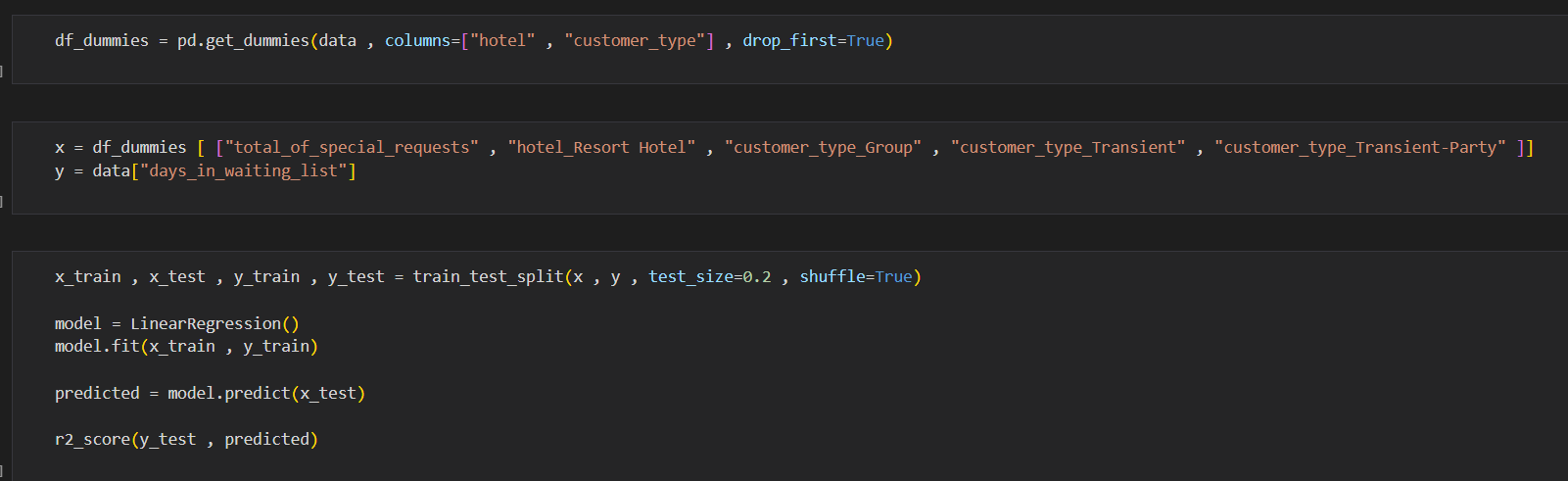
**Regression model:**

Predicting the number of days a booking stays on the waiting list.

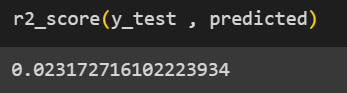
**Independent Variables:** Hotel type, customer type, and the total number of special requests.

Dummies data frame is created dropping the hotel and customer\_type.

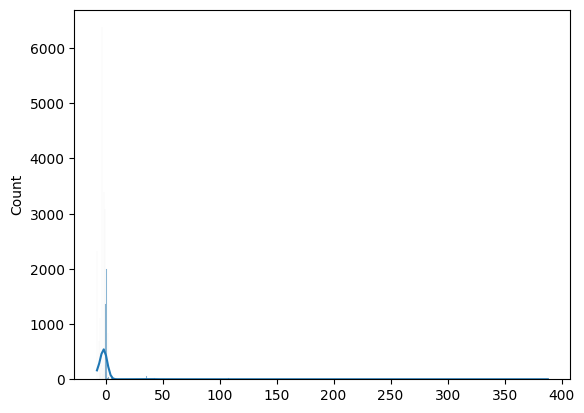
Model is trained and the x\_test data is predicted with r2\_score of 0.02351614929371748.



**RESULTS:**



**Residual plot:**



The picture shown is described as Residual plot. This picture clearly shows that there is some normality in the data used for predicting days in waiting list.