Hotel Booking Analysis-Technical Documentation

ASHISH AVDHESH GUPTA, MOHAMMAD SUHAIL M K, SAHIL SANJAY CHOCHE

COHORT SHIVALIK, ALMABETTER PRO MEMBER

Contents -:

Acknowledgement

First of all, I would like to thank **Almabetter** for giving me opportunity to learn amazing things of python for data science. I would like to thank my all teachers and student success manager for helping me to solve my doubts.

Beside from my lecturer, I like to thank my teammates for helping to understand the project related questions more clearly. They gave their best. I thank them for their effort.

Hotel Booking Analysis-Technical Documentation

Introduction

Word "HOTEL" as we are familiar with this word and come across with this word very often. What does HOTEL mean? a hotel is a building where travelers can pay for lodging and meals and other service. The hospitality industry is one of the blooming industries all over the world. Hotel operations vary in size, function, complexity, and cost. Most hotels and major hospitality companies have set industry standards to classify hotel types. An upscale full-service hotel facility offers luxury amenities, full-service accommodations, an on-site restaurant, and the highest level of personalized service Small, lower-priced hotels may offer only the most basic guest services and facilities.

The dataset describes two types of hotels, city and resort hotel. This has total 1, 19,390 rows and 32 columns. We perform exploratory data analysis (EDA) with python to get insights from the data, to help hotel industry take key decisions which will help them improve their performances in market as well as profits too.

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? So in this project we are going to look into the given data and try to analyze it.

2.1 Why Hotel?

1. The hotel prices varies from consumer to consumer
2. The priority of hotel is customer service
3. Hotel provides best facility to the customer
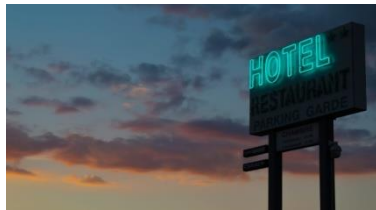
2.2 Past 10 years…

1. High demand from business travelers and Internationals tourists

2. Many facets of the hotel industry have shifted to adapt to new technology and changing consumer behavior, which have in turn changed how hoteliers conduct business.

3. The availability of internet marketing has led to increased competition and has created pressure on room rates. As costs continue to rise at a higher rate, the net effect on the [net operating income] has overall been negative. "Profit margins have also decreased because of higher commissions from [online travel agencies] and increases in operating costs such as workforce, and the increased amenities demanded by consumers."

Basic Insights

3.1 Marketing and Distribution

Marketing and distribution costs are the most significant changes the industry has seen over the past couple decades. "The delivery cost for getting customers into hotels was changed significantly by the business model, which was primarily based on brand and travel agents that were being paid at a 10-percent rate to the OTAs".

the traditional marketing whether it was billboards newspaper, television, radio—that hotel companies did has been replaced by digital formats in the form of ads through social networking, Google, etc."

3.2 Basic Terminologies

Market segment: Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators".

Distribution channel: Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators".

Lead Time: At a hotel, the time taken between when a customer makes a reservation and their actual arrival is called the Lead Time. Sometimes this time period has to be restricted.

Categorical variable: categorical variable takes on a limited, and usually fixed, number of possible values Examples are gender, social class, blood type, country affiliation, observation time.

Continuous variables: Continuous variables are numeric variables that have an infinite number of values between any two values. A continuous variable can be numeric or date/time. For example, the length of a part or the date and time a payment is received.

No-show: This consists in the non-appearance, without prior notice, of a confirmed reservation. The client did not release the room and we no longer have the opportunity to sell it, implying an opportunity cost.

Data Preparation
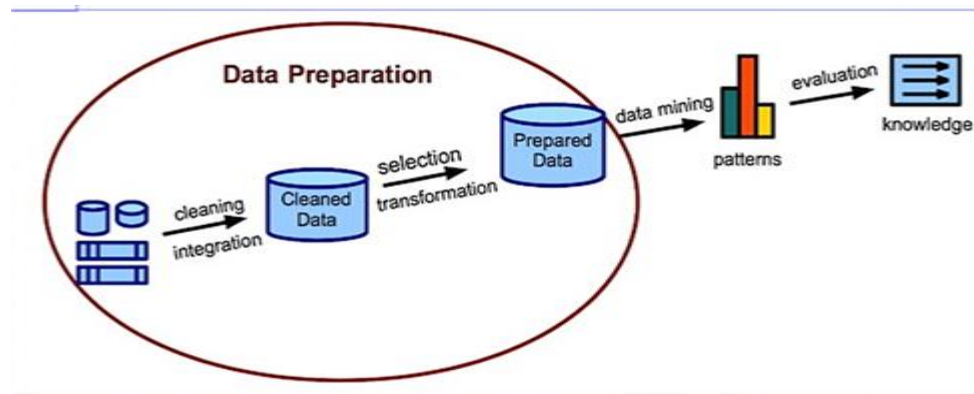
Data preparation is very important in exploratory data analysis. Because our output is depends on how our data is. First we check for the null and Na values. If very fewer amounts of data are Na or null then we just remove those data and if high amount of data is missing then we just replace it with mean and mode, depends on terminology. After this we look for duplicates. We check in data. If there are duplicates in our data then we drop those duplicate data. Now in our

data there are zero null, Na and duplicates data. Then we check for outliers in our data using box plot. It is very easy to check for outliers but replacing that outlier is very hard .because if we replace outlier with wrong values then it will lead to wrong direction. Therefore imputing outlier is very important. There are 3 types of outlier as following:

Type 1: Global Outliers (aka Point Anomalies)

Type 2: Contextual Outliers (aka Conditional Anomalies)

Type 3: Collective Outliers

There some ways which we can replace outlier. First is, we replace outlier with the mean and second is replace with IQR (inter quartile range)

4.1 What is data preparation?

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making correlations to data and the combining of data sets to enrich data.

4.2 Why data preparation?

Data preparation ensures accuracy in data, which leads to accurate insights. Without data preparation, it's Possible that insights will be off due to junk data. i.e. **No quality data, no quality mining results!**

4.3 What are the techniques provided in data preparation?

1. **Data cleaning:** Fill in missing values, identify and/or outliers and resolve inconsistencies.

2. **Data integration:** integration of multiple datasets, data cubes, or files

3. **Data transformation:** Normalization and aggregation.

4. Data reduction: obtains reduced representation in volumes but produces the same or similar analytical result.
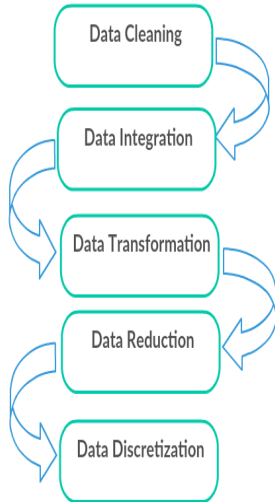
EDA

EDA is exploratory data analysis. In this we will analyze trends of our data more deeply. We will take each and every column and try to analyze it and get some insights from data. Same for continuous and categorical variables.

5.1 Univariate Analysis

Univariate analysis means explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable.

We did univariate analysis for column name as canceled, country, meal, segment, customer type, year, month and reservation. For this we define function to identify continuous and categorical variable. For numeric column we use histogram and count plot graphs to get some insights. Then for categorical variable we used pie chart.

From this we got some of the following conclusion:-

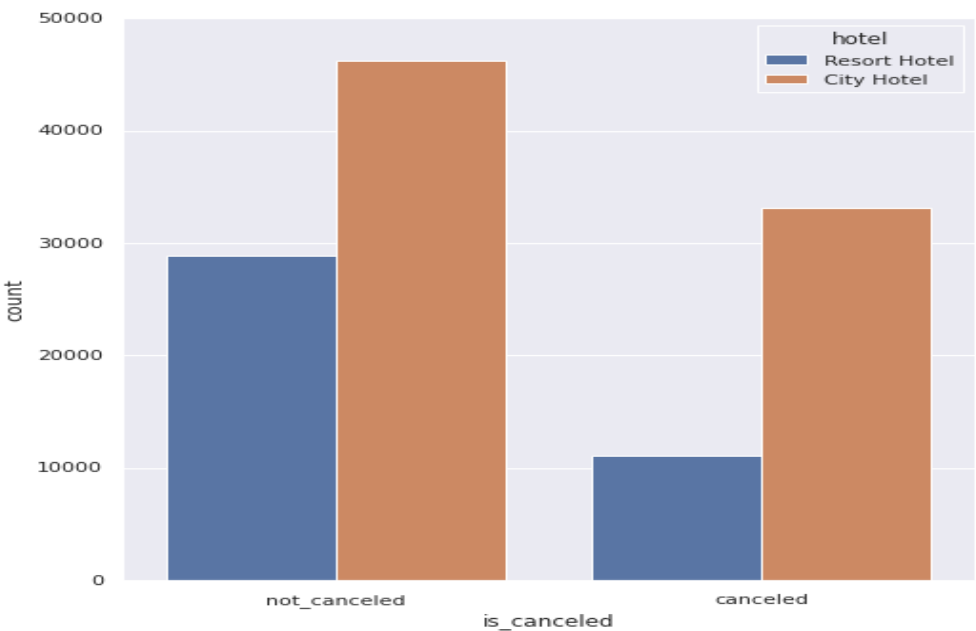1) In terms of hotel most number of reservations has city hotel compared to resort hotel.
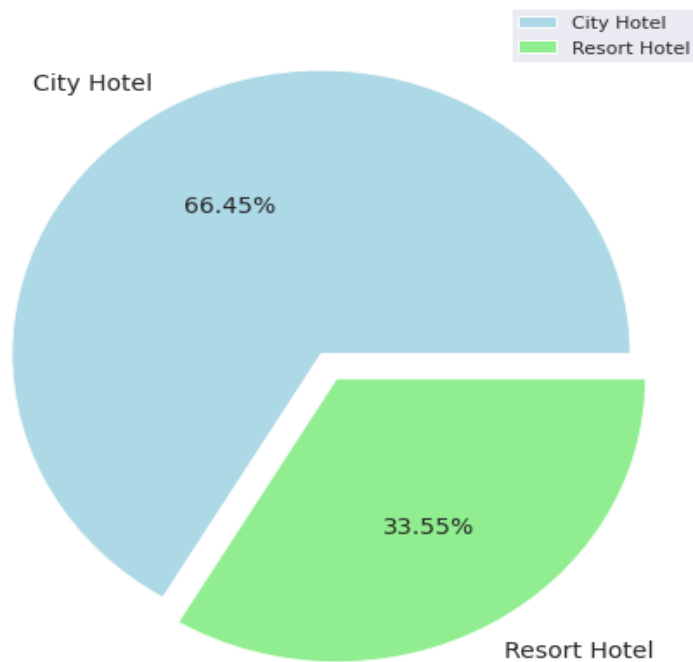
2) Overall bookings were canceled which is around 27%.
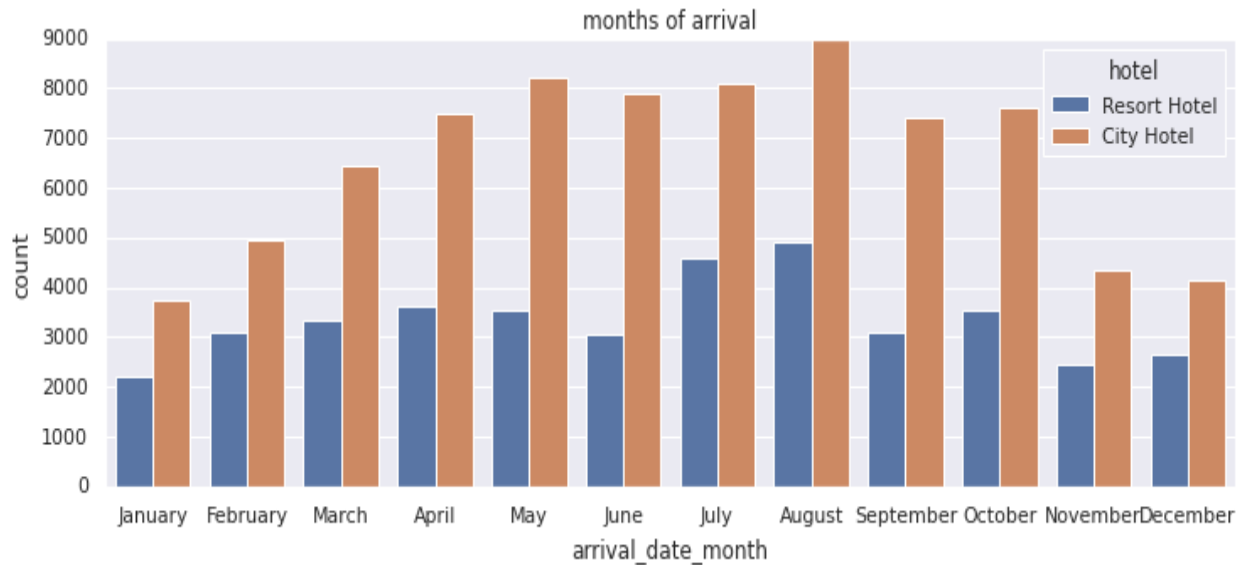
3) Around 40.69% of all bookings were booked from Portugal, Great Britain 10.15% & France 8%.

```
PRT      40.698551
GBR      10.159142
FRA       8.723511
ESP       7.176480
DEU       6.103526
           . . .
MMR       0.000838
PYF       0.000838
DJI       0.000838
BFA       0.000838
LCA       0.000838
```
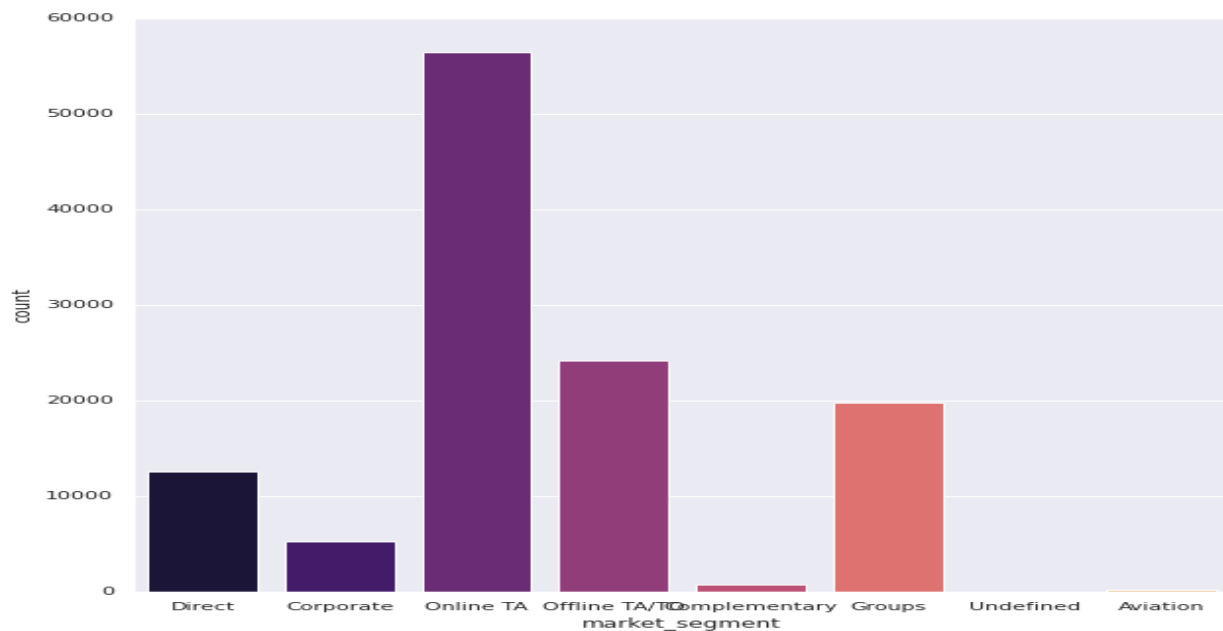
4) August is the most occupied or busiest month with 12.9% bookings and January is the most unoccupied month with 5.36% bookings.
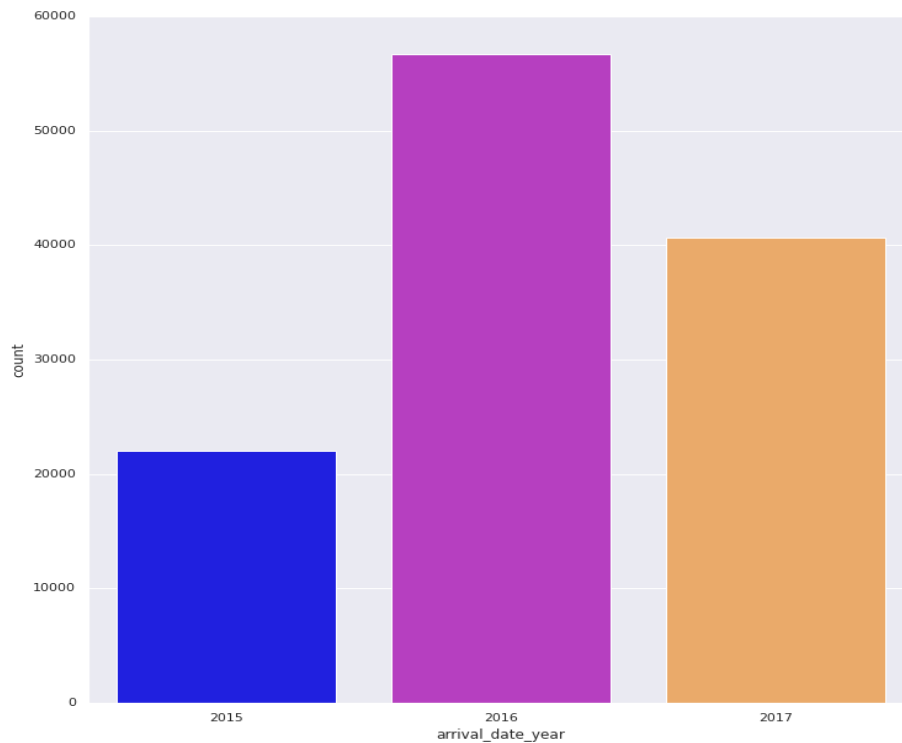

5) Highest number of booking made via online travel agent i.e. 59%, almost 15.8% of bookings are made via Offline Travel Agents, and less than 15% are direct bookings without any other agents.

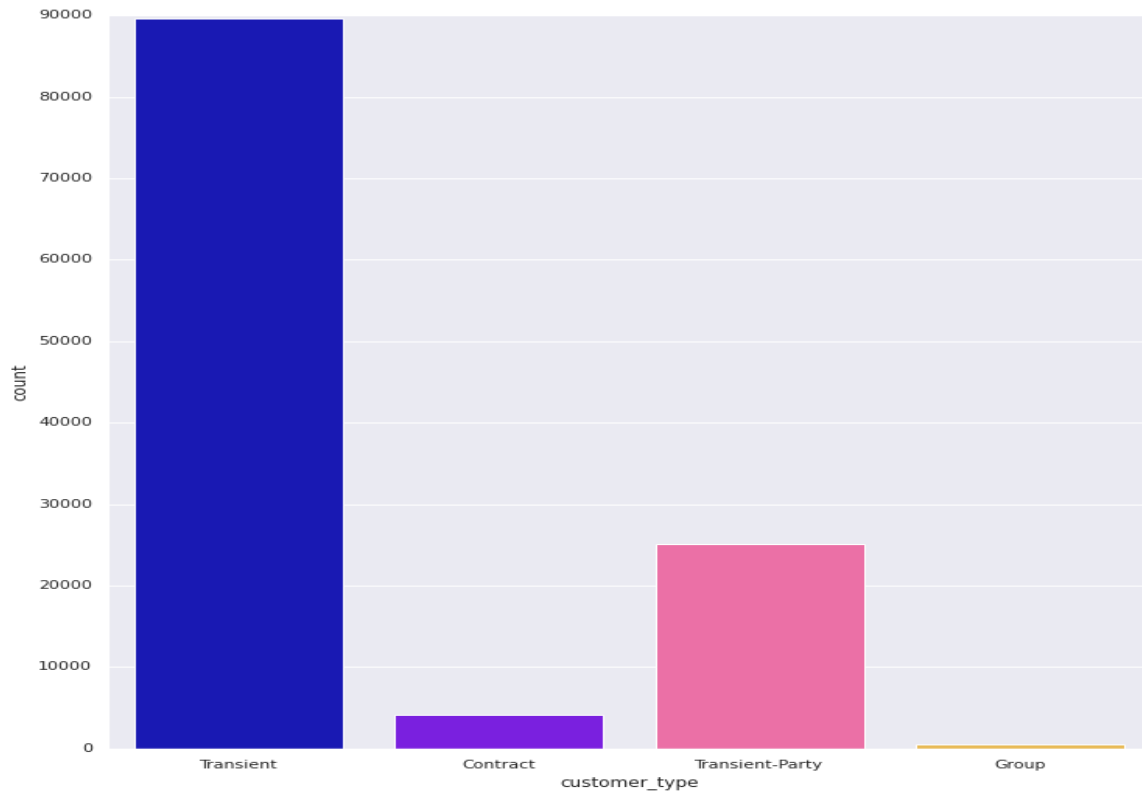6)48.5% bookings were done in 2016, 36.2% in 2017 and 15% in 2015.

7) BB (Bed & Breakfast) is the most ordered meal which is around 77.31%, followed by HB (Half Board), SC (no meal package), Undefined and FB (Full Board).

BB	**77.318033**
HB	**12.114080**
SC	**8.920345**
Undefined	**0.979144**
FB	**0.668398**

8) Transient (person) type of customers is the more around 82%.

5.2 Bivariate Analysis

Bivariate analysis is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them.

When we talk about bivariate analysis, it means analyzing 2 variables. Since we know there are numerical and categorical variables, there are 3 ways we can analyze these variable. i) Numerical vs. Numerical ii) Categorical vs. Numerical iii) Two Categorical Variables. In this we used joint plot and bar plot using seaborn library. Seaborn library has awesome functionality, we used some of it.

From this we got some of the following conclusion:-

1) Maximum number of booking changed by Transient customer and most of Change Booking done by Transient-party. Very less change in booking done by contract and group type.
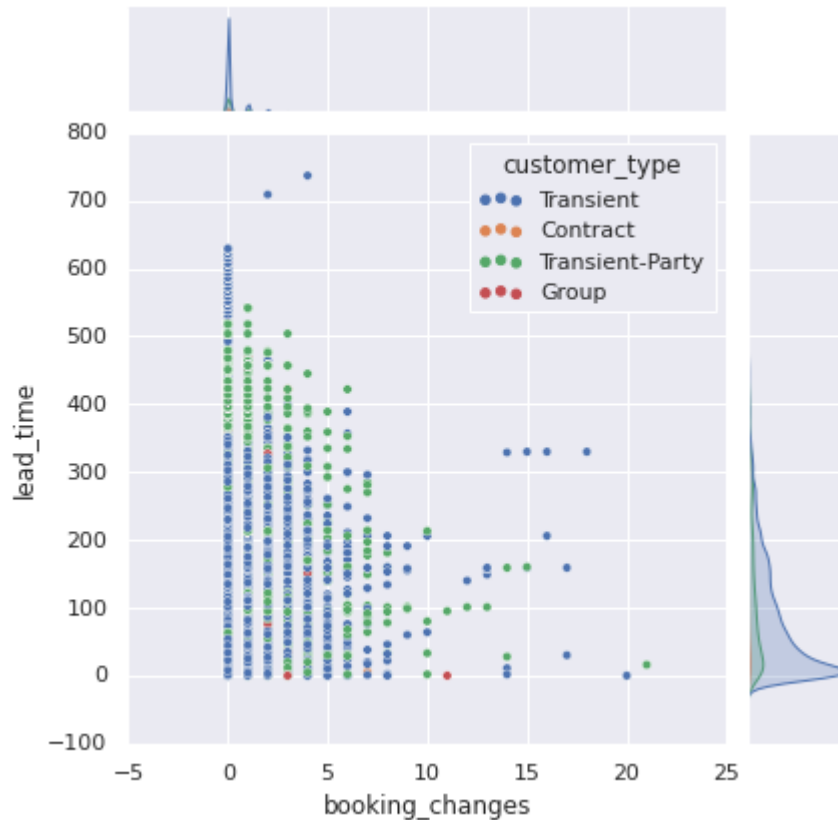
2) Both hotels have maximum Previous Booking done by group, followed by transient and transient-party and very less previous booking done by contract.
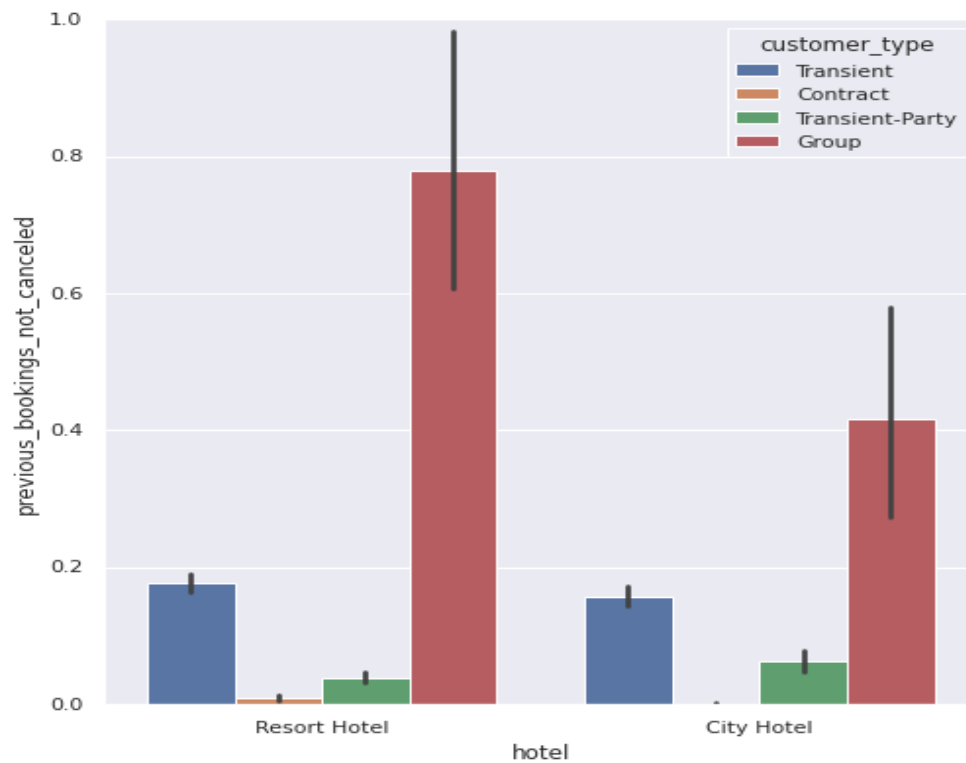
5.3 Correlation of data

Correlation is a **statistical measure that expresses the extent to which two variables are linearly related** (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.
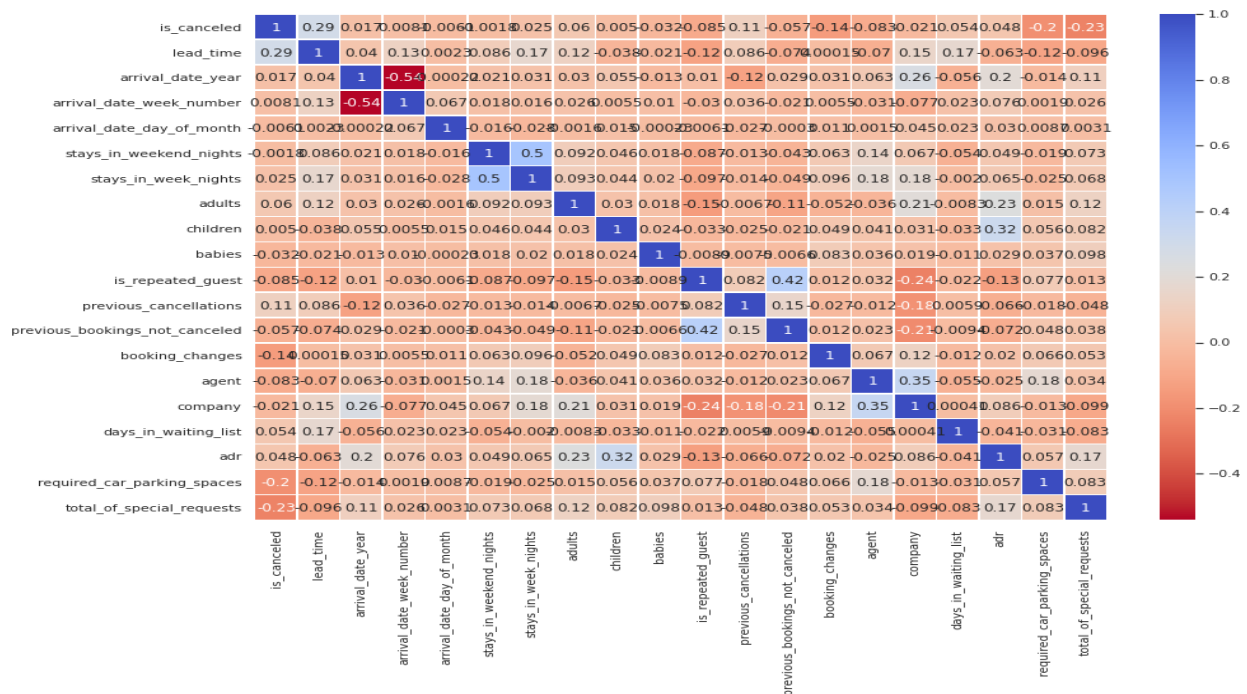
Hypothesis
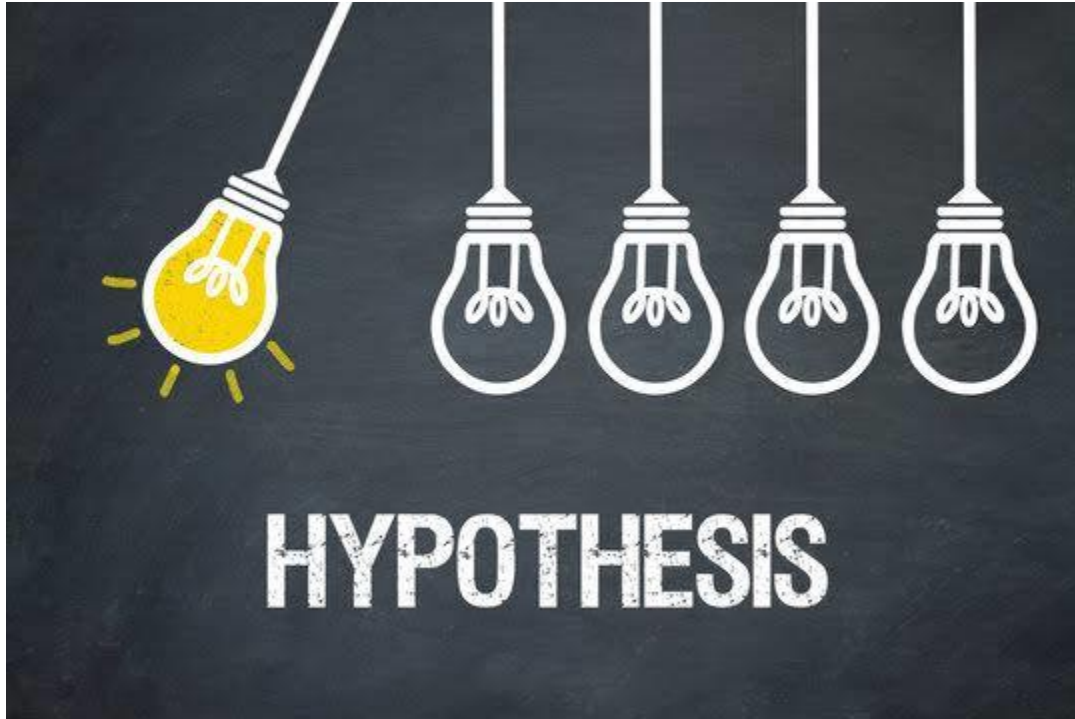
Hypothesis is a tentative prediction or explanation of the relationship between two variables."It implies that there is a systematic relationship between an independent and dependent variable".

Hypothesis means mere assumption or suppositions which are to be proved or disproved. Hypothesis is a formal question that is intended to resolve.

5.4 Does market segment have any impact on repeated guest?

We have done some hypothesis questions which help us to get more insights and useful outcome.
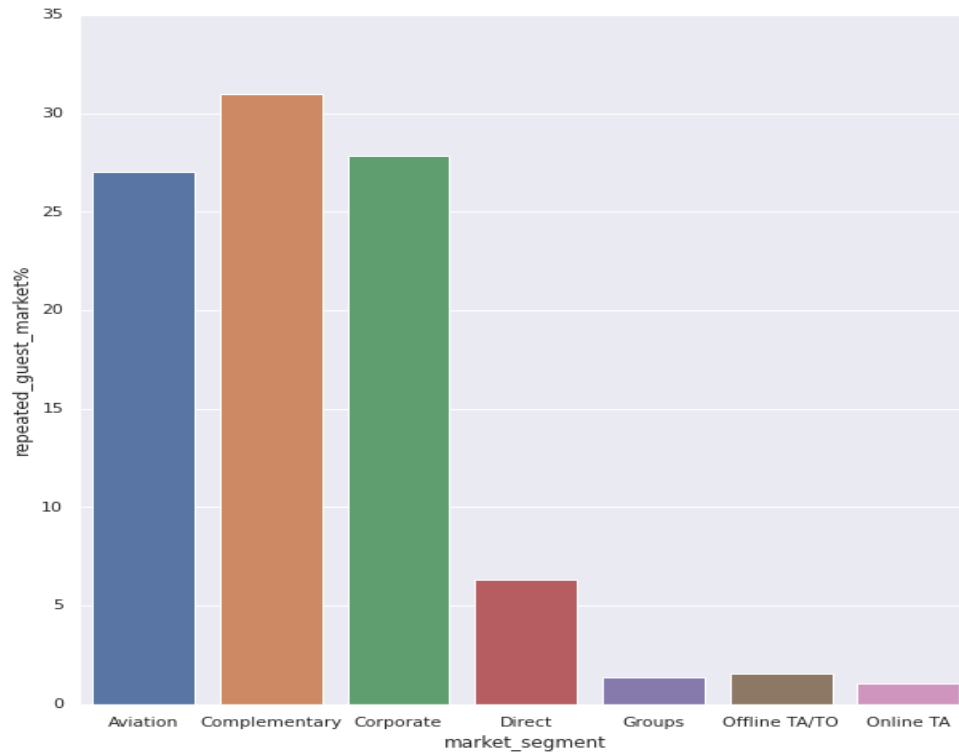
While working on first hypothesis question, dose market segment have any impact on repeated guest? We saw that there are very few people who repeat the same market segment and most of people was not repeating same market segment.

5.5 Is effect of customer types and reservation status?
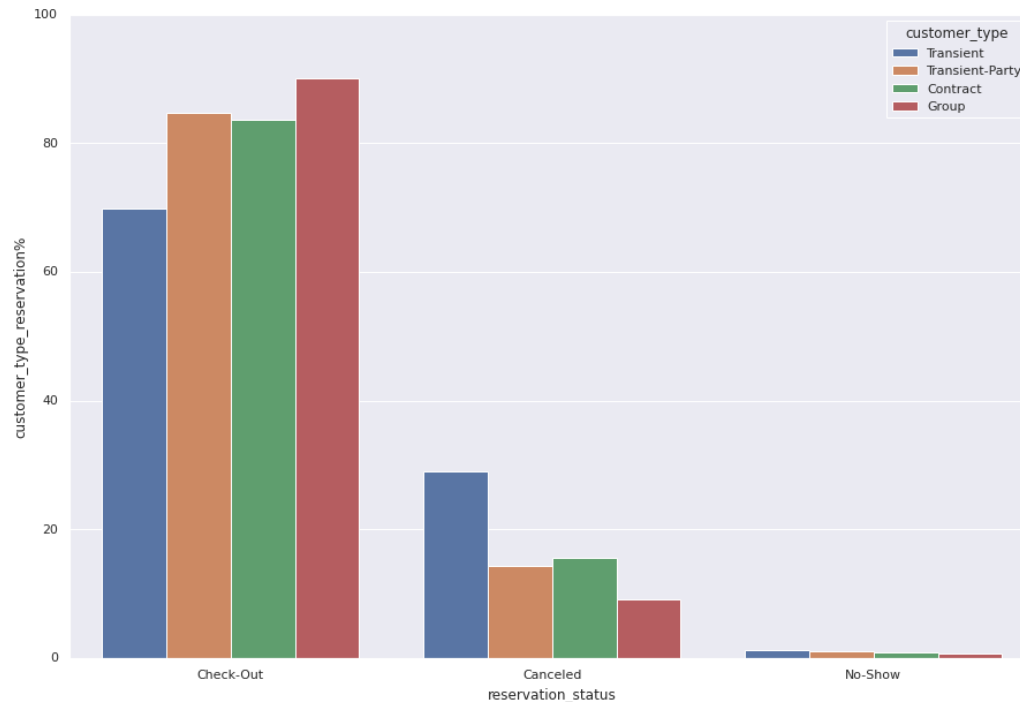
Second hypothesis question, is effect of customer types and reservation status?

- We can see, the group type in check-out status has more percentage

- And in cancelled status transient type has more reservation percentage

Price Analysis

In this we compared average daily rate on the basis of types of hotels, types of assigned room and month and years. In hotel vise comparison city hotel has high adr then that of resort hotel.

```
hotel
City Hotel     110.985944
Resort Hotel    99.025346
Name: adr, dtype: float64
```

If we checked on assign room type then H, G and F have high adr also L & P has negligible adr compared to others

```
assigned_room_type
A      95.849738
B      95.368533
C     116.900878
D     109.067211
E     119.199101
F     152.870074
G     167.424868
H     171.881530
I      40.801933
K      53.829674
L       8.000000
P       0.000000
Name: adr, dtype: float64
```

If we check month vise adr then August 2017 has highest adr and the lowest is November 2015.

	arrival_date_month_year	adr
0	April 2016	90.972653
1	April 2017	115.128432
2	August 2015	122.994391
3	August 2016	148.833968
4	August 2017	168.549327
5	December 2015	71.540293
6	December 2016	87.687599
7	February 2016	69.931764
8	February 2017	78.749541
9	January 2016	62.698459
10	January 2017	74.830759
11	July 2015	112.399379
12	July 2016	131.396209
13	July 2017	147.613122
14	June 2016	108.865351
15	June 2017	128.761827
16	March 2016	79.483631
17	March 2017	83.821445
18	May 2016	100.088238
19	May 2017	120.373019
20	November 2015	59.236336
21	November 2016	79.513523
22	October 2015	78.925463
23	October 2016	97.311953
24	September 2015	99.878190
25	September 2016	121.077505

```
arrival_date_month_year
April 2016        10471.0
April 2017        11529.0
August 2015        8040.0
August 2016       11600.0
August 2017       11131.0
December 2015      5470.0
December 2016      7726.0
February 2016      7215.0
February 2017      8084.0
January 2016       3835.0
January 2017       6689.0
July 2015          5715.0
July 2016         10101.0
July 2017         11791.0
June 2016         10101.0
June 2017         11380.0
March 2016         9261.0
March 2017         9171.0
May 2016          10420.0
May 2017          12036.0
November 2015      3807.0
November 2016      8005.0
October 2015       9084.0
October 2016      11960.0
September 2015     9726.0
September 2016    10640.0
Name: Total_guest, dtype: float64
```

Conclusion

1) Even percentage of total booking more in city hotel as compared to resort hotel but the number of cancellation in city hotels maximum compared to resort hotel. Overall cancelation rate is 27%. Approximately 20% of the total booking is getting cancelled in city hotel. Hence the percentage of confirmed booking is more in resort hotel.
2) The "online TA" is most used market segment and "TA/TO" is maximum in distribution channel.
3) Transient type of customer is most coming customer compared to others.
4) Most preferred meal by customer is BB (Bed & Breakfast) type.
5) Around 40.69% of all bookings were booked from Portugal, Great Britain 10.15% & France 8%.

Some future prediction we can make to help hotel industry, take key decisions to maximize profit.

- ✓ In cancellation of room, hotel should take the cancellation charges. But cancellation charges should be less, then it will not affect consumer also.
- ✓ If no-show of consumer seems then within some timeframe room should assign to someone else.(timeframe – take follow-up of customer)
- ✓ Hotel should consider maximum number of special guest then it will lead to minimum cancellation.

Reference

- ❖ TowardsDataScience
- ❖ Kaggle
- ❖ Google
- ❖ Medium