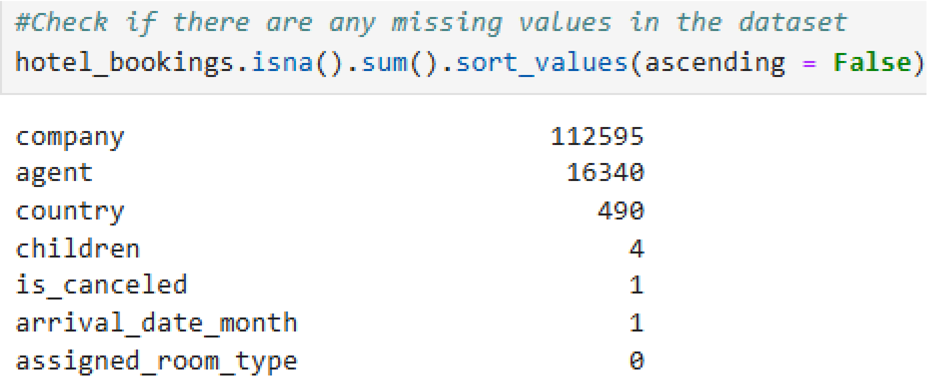
Data Preparation

Missing values:

To begin data preprocessing and preparation, I first need to understand what I am working with. Hence is why I code dtypes(). This tells us the column names, as well as their data type. Additionally, info() gives me further information, and head() demonstrates the information displayed in the dataframe.

After understanding the data, the first hurdle is to identify missing values. This is done through the function isna() and is also summed to demonstrate the total amount of missing values in the column.

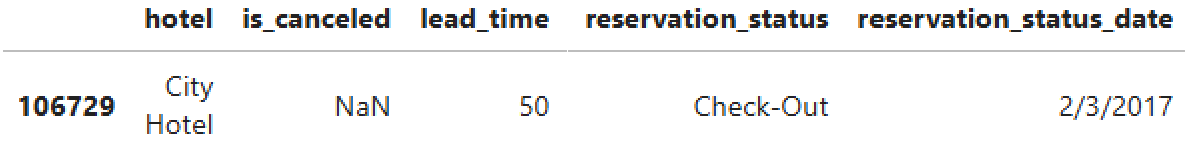


```
#Check if there are any missing values in the dataset
hotel_bookings.isna().sum().sort_values(ascending = False)
```

```
company                     112595
agent                        16340
country                        490
children                         4
is_canceled                      1
arrival_date_month               1
assigned_room_type               0
```

Figure 1. Analysing the number of missing values in the columns

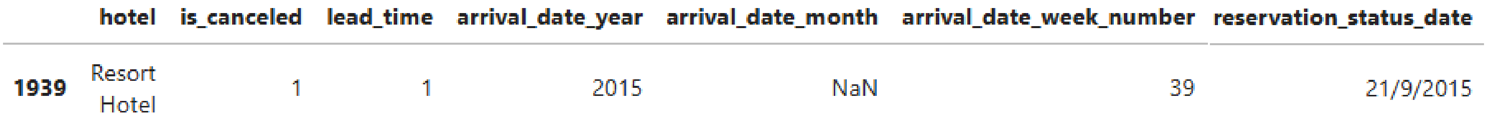
In figure 1, it could be observed that both company and agent have many missing values, which made sense to me because not everyone is affiliated with a company or agent, so these were subsequently filled with 0 to indicate no ID, because each number in the columns represented a unique ID. Also made sure that these were integers, not floats. Countries could not be inferred, as no other information could be related to the country of origin, so they were filled with 'Unknown'. For the children column, if the entry was missing, I thought it would be better to fill it with 0, because it is most likely they didn't have any with them, instead of dropping the rows.



	hotel	is_canceled	lead_time	reservation_status	reservation_status_date
106729	City Hotel	NaN	50	Check-Out	2/3/2017

Figure 2. This merged image shows the missing entry in is_canceled

In figure 2, the is_canceled column had 1 missing value, and when I filtered for the missing value, it showed that the reservation_status was in 'Check-Out' instead of 'Canceled' signifying that the customer had proceeded with the stay. So this was updated to '0'.



	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	reservation_status_date
1939	Resort Hotel	1	1	2015	NaN	39	21/9/2015

Figure 3. This merged image displays the missing value in arrival_date_month

In figure 3, the arrival date month is missing, however with the week number being 39, and the last update on the reservation, it can be understood that this entry is supposed to be 'September'.

When checking the missing values again with the code:

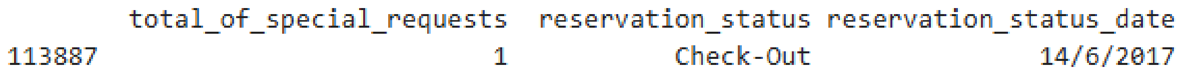
hotel_bookings.isna().sum().sort_values(ascending = False), The missing values are all 0, and I proceeded to the next step of data preprocessing.

Dropping duplicates:

By checking duplicates, I can see that there are 31992 duplicate rows. I drop these rows, reducing the size of the dataset. It is important to note that dropping duplicate rows before addressing missing values is critically wrong, because the whole dataset is not pictured. Some observations may not be dropped before entering missing values since they appear different, however after filling missing values they may be the same, thus dropping a ‘real’ duplicate. Best practice to drop after fixing missing values.

Checking unique values and validity:

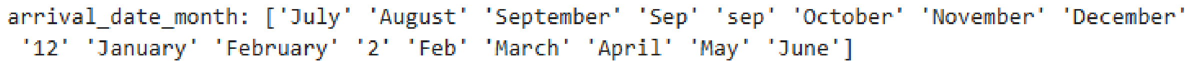
I then check the unique values of each column to address spelling mistakes, or illogical errors. The first outlier present was a lead time of 10,000 days. This equates to 27 years. This is clearly mistyped or entered wrong. I figured that no other information could really tell me what the lead time could be, so I hid it from the dataset. The next error was an arrival date of 1999. Same with the previous, the booking was cancelled, so I decided to hide it from the dataset as well. The next outlier was ‘17’.



```
        total_of_special_requests  reservation_status reservation_status_date
113887                          1           Check-Out               14/6/2017
```

Figure 4. Outlier in arrival_date_year

In figure 4, the reservation status date was set at 2017, and from the entry of 17, I inferred that this could be changed to 2017.



```
arrival_date_month: ['July' 'August' 'September' 'Sep' 'sep' 'October' 'November' 'December'
 '12' 'January' 'February' '2' 'Feb' 'March' 'April' 'May' 'June']
```

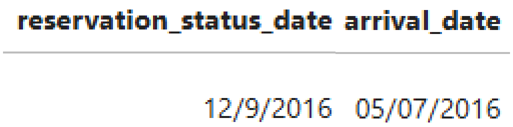
Figure 5. The unique values of the arrival date month

Figure 5, depicts the unique names of the arrival date month. It is messy, where there are different spellings or entries for similar months. These were all changed to be uniform.

The next strange entry was an arrival date week number entry of 53. At first, it doesn’t make sense because there are 52 weeks in a year. However, it is 52 weeks and a few days in a year, those days counting towards the 53rd week. Furthermore, all the entries appeared in late December, which is concurrent with the reason.

Variable creation:

I created a new date using arrival date year, month and day of month. This date was called arrival date, and was formatted into d/m/y (string). This was so that I was able to compare carefully with the reservation status date, as well as I believed that those 3 columns provided was somewhat redundant, and I would have preferred to see 1 column representing the arrival date. It was later converted back to datetime format when assessing the graphs in Task 2.



reservation_status_date	arrival_date
12/9/2016	05/07/2016

Figure 6. The difference between dates

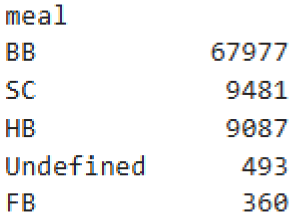
Sanity checks:

With this arrival date (Figure 6) when it was in datetime format, I checked if there were invalid dates such as 30/2. None were present.

Logic checks:

With the amount of days stayed at the hotel, I took the highest value to check if the stay period is correct. I thought 19 weekends seemed very long, stretching out for 2 months. A rough estimate of their stay period can be calculated from the arrival date and their reservation status date. Although they stayed 19 weekend, the entry has no errors. This is the same with the 50 week day entries.

Both entries of 10 children and another of 10 babies seemed farfetched. The entry with 10 children were a no-show, the entry with 10 babies with 2 adults didn't cancel. I don't have enough context or enough reason to doubt it, so I kept it.



```
meal
BB           67977
SC            9481
HB            9087
Undefined      493
FB             360
```

Figure 7. The sum of different meal types

Figure 7 represents the type of meal count ordered. I left 'Undefined' as is, because there were 493 entries. This signifies that they probably did not have a meal option chosen.

I also checked the observations with 8 required car parking spaces, because it seemed like a lot, especially for just 2 adults. But I lack context, so I left them.

The reservation status date had a few entries with the wrong format. I created a mask, this turns that entry into a Boolean array. When using the .loc method, it updates the mask in the column to my newly entered data.

I recheck the structure of the dataset with head(), and it turns out great. Saved the csv file as cleaned_hotel_bookings.csv

Data Exploration

Task 2.1: What insights can be gained from exploring booking trends over time, such as monthly trend of booking statistics (e.g., total bookings, cancellations, average lead time) in the year of 2016?

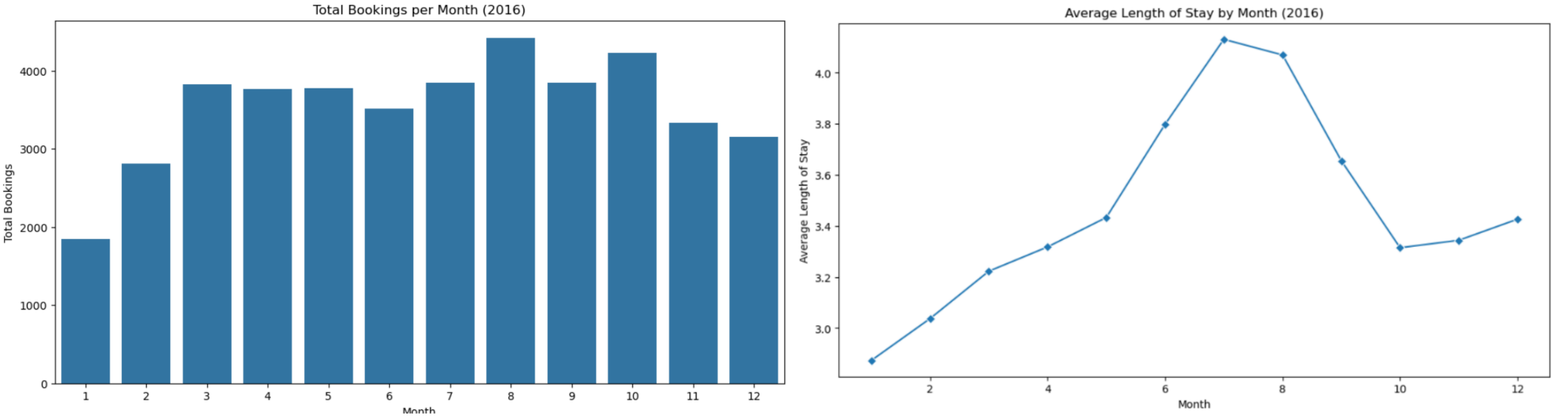


Figure 8. Monthly statistics of bookings and average length of stay in 2016

It is important to understand first that this data originates within Europe. This is understood through the most common visitors being Portugal, Great Britain, France, Spain and Netherland s. The highest amount bookings are in August, however it is evenly spread out apart from Jan, Feb, Nov and Dec. August is summertime in the Northern Hemisphere, so it makes sense that people are taking vacation all throughout Spring, Summer and Autumn. This relates to the length of stay in particularly July and August, which is Summer, where the longest stay period is significantly greater than other periods, passing an average of 4 days per hotel booking. What can be taken from these graphs is that I have identified the on and off peak seasons to travel, and it corresponds to the season. A shorter trip may represent a weekend getaway.

Thus, it is seen that the most popular time is near the middle of the year to the tail end besides Winter. This can help understand the businesses busiest and most profitable period, and when they need to account for large volumes of bookings, and staff employment to tackle the demand.

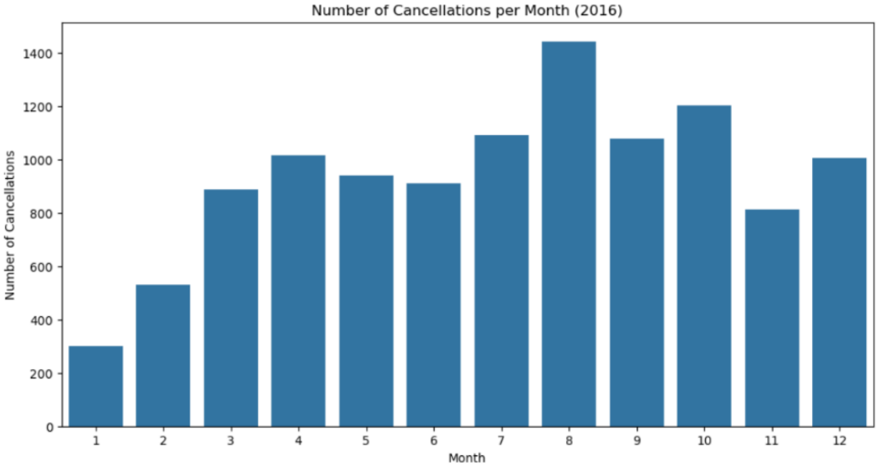


Figure 9. Amount of cancellations per month in 2016

Highest booking cancellations are also prominent in the busiest period like August. With a larger amount of bookings, there also comes a larger number of cancellations, so it makes sense. However, it needs to be understood why people are cancelling. Maybe the hotel was so overbooked that they needed to cancel, or maybe there was a shift in consumer preferences due to competition. With context, the reason should be understood to further the business.

Task 2.2: How do we describe and visualize the complex relationships between bookings, seasons and years for data between 2015 and 2017? What can be learned from this analysis?

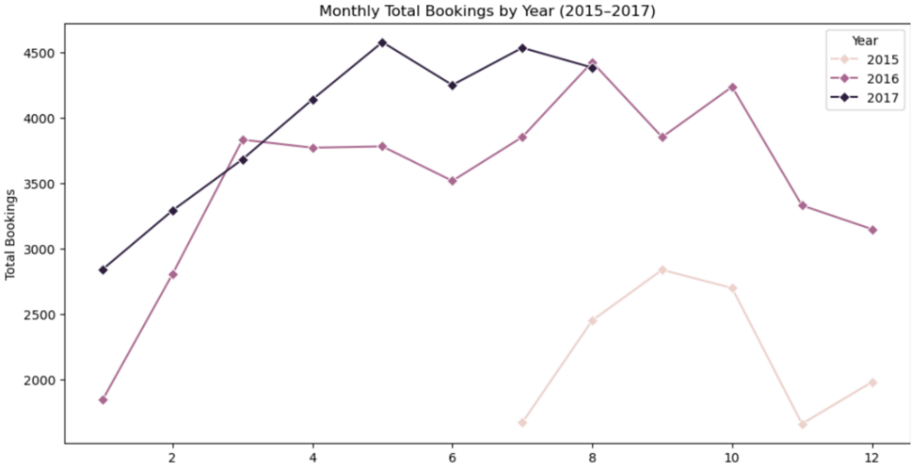


Figure 10. Monthly bookings within a 2-year span between 2015 and 2017

Figure 10 depicts the monthly bookings from July 2015 to July 2017. It can be observed that there is an increasing trend in the total bookings. This signifies that the hotel is gaining in popularity. The January booking total in 2017 is roughly 2800, which is greater than the best month of 2015 in September. Overall, the dark purple trend (2017) is above the pink trend (2016), demonstrating and increase in popularity. It can also be seen that the most popular months still appear to be from June to August, the summer months of the Northern Hemisphere. So most people are taking their vacations in Summer, and least in Winter.

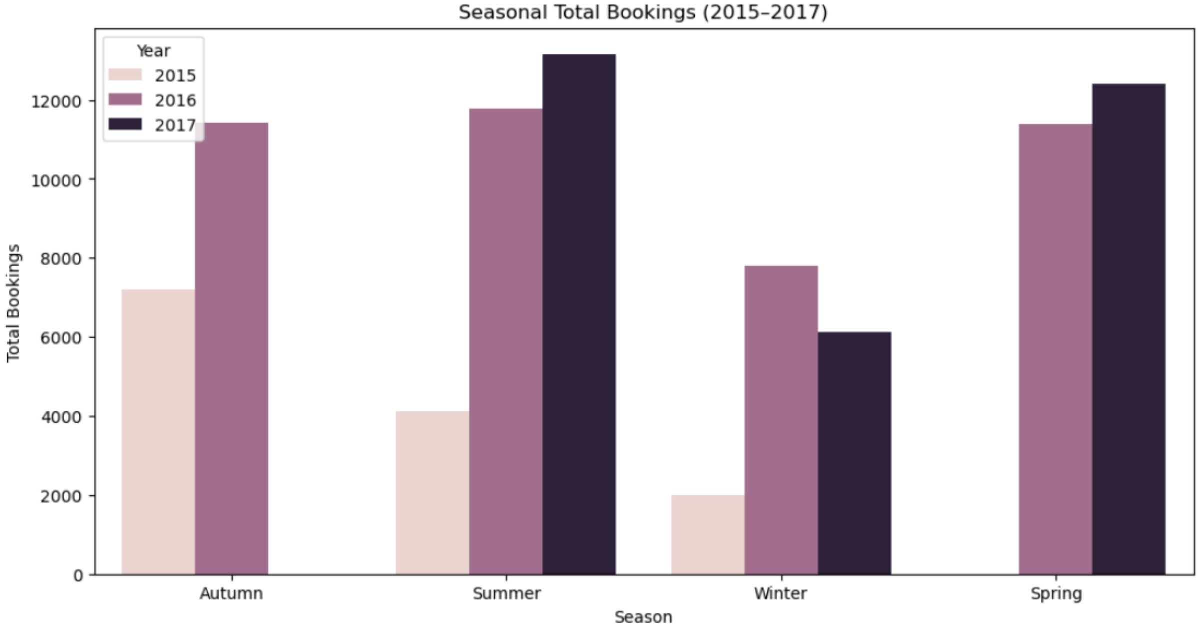


Figure 11. Seasonal total booking numbers from 2015 to 2017

This seasonal graph shows the total number of bookings in that year in each season. In the year 2016 and 2017, the trend remains, where summer is the most popular, but not by much. The number of bookings is quite similar. What is interesting is that 2015 summer has very little bookings compared to autumn. This may be because the hotel only just opened. The data records start at July 2015, so half of Summer was left, this was compared to Autumn with 3 full months of September to November. However, there may be other factors, and this anomaly may be due to external factors.

Task 2.3: **What stories emerge when we analyze the distribution of bookings by geographic region (like top 5 countries), and how do these insights inform our understanding of their customer behavior and preferences (e.g., average length of stay and cancellation rates)?**

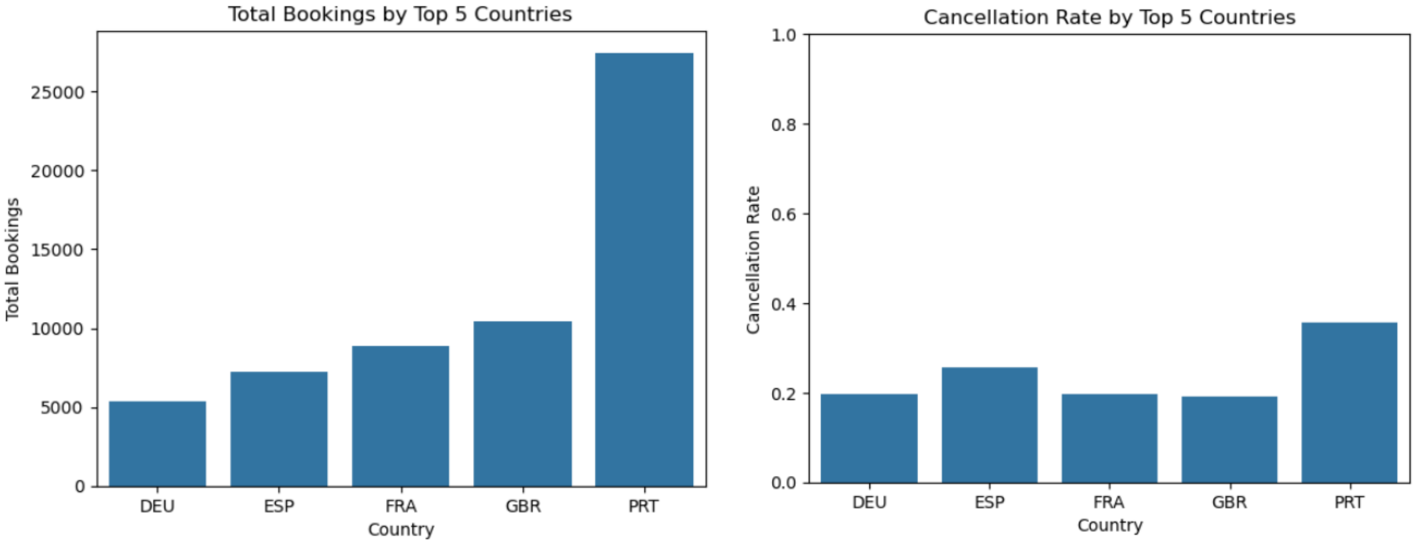


Figure 12. Booking and cancellations by visitors from the top 5 countries

What can be taken from this is that Portugal is likely the host country of this hotel. The reason being is that many Portuguese may go for a holiday like a weekend getaway and stay in their own country. The more bookings, the more likelihood for cancellations. Spain is quite a bit higher than the other 3 countries. This potentially could be because Spain is geographically close to Portugal, and they would rather stay in their own country.

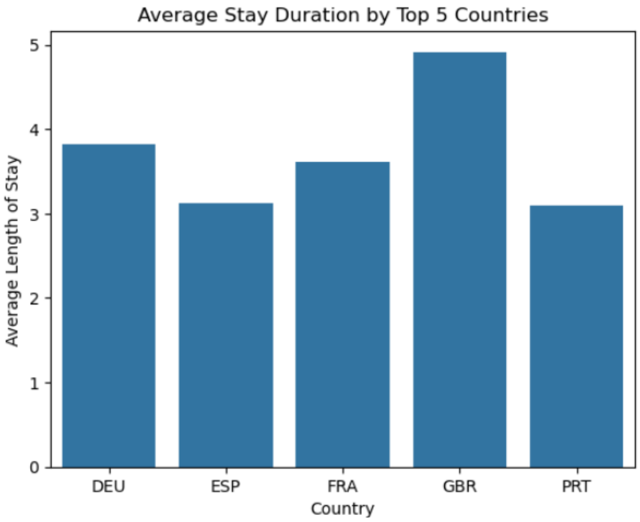


Figure 13. Average stay duration by the top 5 country visitors

Figure 13 also matches the hypothesis that since Spain is a geographical neighbour to Portugal, thus leading to a shorter stay. This could be because of shorter stays and weekend trips, similarly to the Portuguese. International guests besides Spain typically stay longer, especially the UK people, and international guests cancel less. UK stay longer because their weather is often characterised as rainy, gloomy and cloudy, whereas Portuguese weather is sunnier, which would be much more appealing to a UK citizen. To sum, this hotel is based in Portugal due to the large number of bookings. International people other than Spain are more likely to not cancel and stay for longer. This is because Spain is a geographical neighbour to Portugal.