Student Name: Johnson Chang

Data Preparation

Error 1: Check if the data is loaded correctly

Verified the data has 119,392 rows, 32 columns, which is same as shown on csv file with 119,393 rows and 32 columns. (csv file include the column name row).

Rows	119393
Columns	32

```
hotel_bookings.shape
```

```
(119392, 32)
```

Error 2: Outliers in 'lead_time', 'adr'

For outliers, I check the summary statistics using describe(). Looking at the data 'lead_time', I think 'lead_time' > 465 are outliers, and I replace them with median value. When I check the average income per occupied room('adr'), I think the outliers are min (-6) and max(5400) values. Then I replace outliers in 'adr' with median values.

```
hotel_bookings['lead_time'].describe()
```

```
count    119392.000000
mean        104.170857
std         113.959838
min           0.000000
25%          18.000000
50%          69.000000
75%         160.000000
max       10000.000000
Name: lead_time, dtype: float64
```

```
hotel_bookings['adr'].describe()
```

```
count    119392.000000
mean        101.830768
std          50.535443
min          -6.380000
25%          69.290000
50%          94.560000
75%         126.000000
max        5400.000000
Name: adr, dtype: float64
```

Error 3: Duplicate values

For the duplicate Removal, I find duplicates using df.duplicated().sum() and drop them using df.drop_duplicates()

```
hotel_bookings.duplicated().sum()
```

```
31992
```

Error 4: Impossible or illogical value in ' adults ', ' previous_bookings_not_canceled and 'reservation_status_date''

For the sanity check, each reservation should have at least 1 adult. Hence, I change (adults < 0) into (adults = 1). Monving next, a repeated guest they must have at least 1 non-cancelled booking. Hence, I change (previous_bookings_not_canceled< 0) into (1). Lastly, the 'reservation_status_date' also shouldn't have '31/1/1900', so it was removed.

```
hotel_bookings[hotel_bookings['adults'] < 1]['adults'].value_counts()
```

```
adults
0    385
Name: count, dtype: int64
```

```
mask_sanatyCheck = (hotel_bookings['is_repeated_guest'] > 0) & (hotel_bookings['previous_bookings_not_canceled'] < 1)
hotel_bookings[mask_sanatyCheck]['previous_bookings_not_canceled'].value_counts()
```

```
previous_bookings_not_canceled
0    615
Name: count, dtype: int64
```

```
hotel_bookings['reservation_status_date'].value_counts()
```

```
18/3/2015        1
31/1/1900        1
Name: count, Length: 928, dtype: int64
```

Error 5: Missing value in 'children', 'arrival_date_month', ' country', ' is_canceled'. 'company' and 'agent'

For missing value, I use hotel_bookings.count() to check the missing values. I filled null value in 'children' with 0. Drop the rows with missing values from 'arrival_date_month', ' country' and ' is_canceled'. The columns 'company', 'agent' has too many missing value making it hard to analyse them, as a result I dropped the columns.

```
hotel_bookings['children'].isnull().value_counts()
```

```
children
False    87396
True         4
Name: count, dtype: int64
```

```
hotel_bookings['arrival_date_month'].isnull().value_counts()
```

```
arrival_date_month
False    87399
True         1
Name: count, dtype: int64
```

```
hotel_bookings['is_canceled'].isnull().value_counts()
```

```
is_canceled
False    87399
True         1
Name: count, dtype: int64
```

```
hotel_bookings['country'].isnull().value_counts()
```

```
country
False    86946
True       454
Name: count, dtype: int64
```

```
hotel_bookings['company'].isnull().value_counts()
```

```
company
True     81852
False     5092
Name: count, dtype: int64
```

```
hotel_bookings['agent'].isnull().value_counts()
```

```
agent
False    75075
True     11869
Name: count, dtype: int64
```

Error 6: Format standardization in 'reservation_status_date'

To ensure date format standardization I use regex to find and dropped the 'reservation_status_date' not following (DD/MM/YYYY) format.

```
matched_mask = hotel_bookings['reservation_status_date'].str.match(r'^(\d{1,3}\d|\d)/(1\d|\d)/\d{4}$')
matched_mask.value_counts()
```

```
reservation_status_date
True     86943
False        1
Name: count, dtype: int64
```

Data Exploration

Task 2.1

How I explore the data:

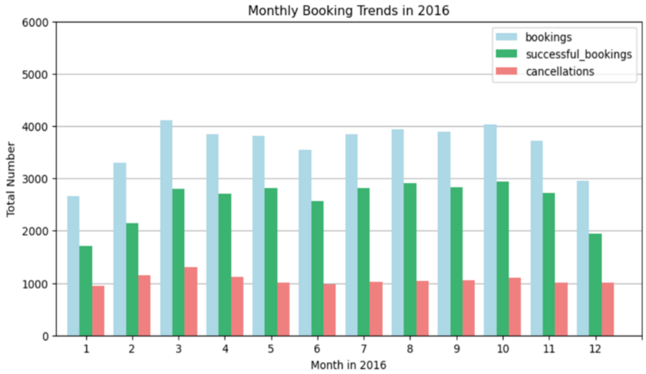
I created a copy of the booking dataframe for 2016 analysis. First, I change the reservation_status_date from string format (DD/MM/YYYY) to date type to filter records from 2016. I then added a month column for grouping data by month. Next, I created an empty dataframe (plt_df) specifically for plotting 2016 data. For each month, I calculated and added several columns: total bookings, cancellations, successful bookings, and average lead time.

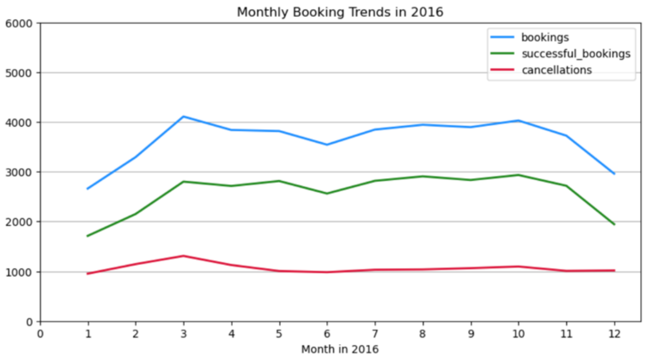
For the plot showing ' Monthly Booking Trends in 2016', I initially created a bar plot showing total bookings, cancellations, and successful bookings counts. Later, I added a line plot as it can show the trend over time better. For the "Days from booking to arrival in 2016" plot, I used a simple bar plot which communicate the lead time data effectively.

Key insights I gain:

From line and bar plot showing 'Monthly Booking Trends in 2016', the lowest booking volume occurred in December and January, so business should provide more incentive or discounts during those months to reduce room vacancy rates. Additionally, marketing and pricing strategies should target in May and November, which show lower cancellation rates with tailored promotions to increase revenue.

The plot "Days from booking to arrival" shows that travellers tend to plan in advance for trips from July to October, booking approximately 3 months in advance for these times.





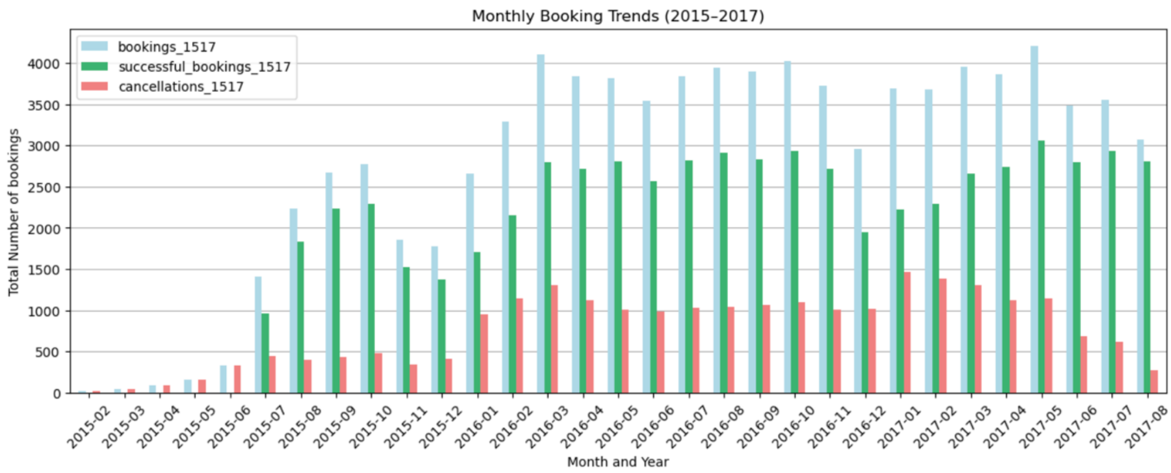
Task 2.2

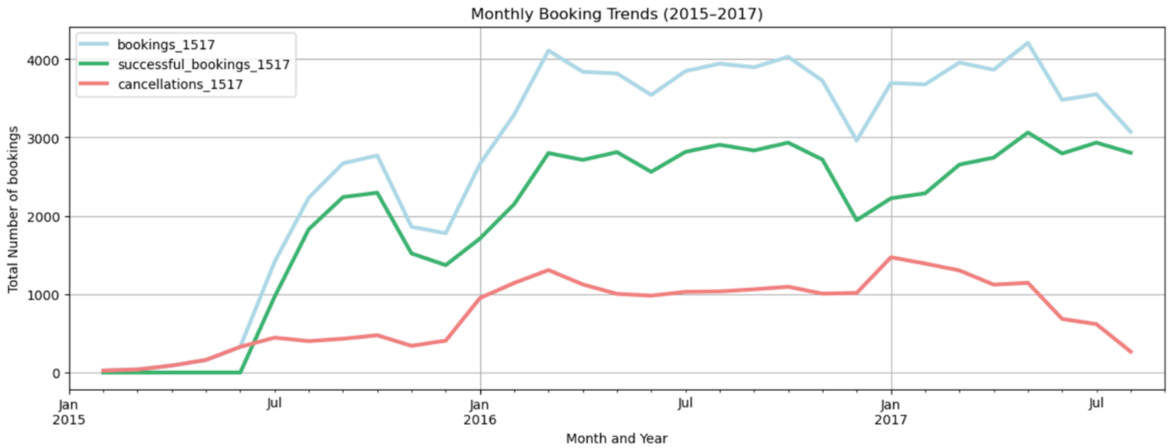
How I explore the data:

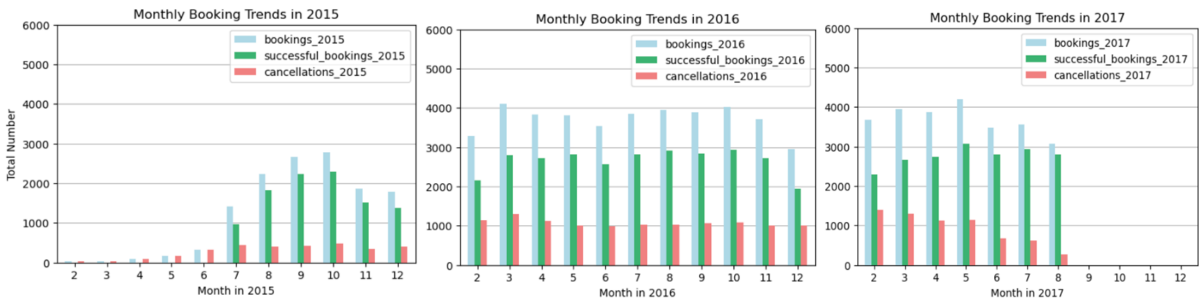
I created a copy of the booking dataframe for 2015-2017 analysis, changing the date format from string (DD/MM/YYYY) to date type. I added month, year, and month_with_years columns to streamline the calculations. Using the year column, I filtered for data from 2015-2017. When testing the plot, I found two stranging booking data 2015-01 and 2017-02(the booking numbers are equal to the cancellation). I cleaned them up but retained one sample row for plot. I then created an empty dataframe (plt_df_1517) for plotting and enriched it with calculated data (total bookings, cancellations, successful bookings, and average lead time) for each month. Finally, I plotted grouped bar and line charts to examine trends across 2015-2017 and created faceted bar plots to get a better comparison by year.

Key insights I gain:

The line plot and the bar plot show Monthly Booking Trends (2015–2017) reveals that the business began in January 2015 and the data collection came to the end in September 2017. The analysis of the 2016 data highlights the more stable business operation in that year indicating seasonal patterns in the travel market. The line chart has two dips in November and December, suggesting a decline in bookings during the late-year period.







Task 2.3

How I explore the data:

I created a copy of the booking dataframe for country-based analysis. Through exploration, I found that the top 5 countries account for 68.3 percent of total bookings, so I decided to focus on these countries. First, I created masks for these 5 countries for calculations, then computed values for each column (total bookings, average night stay, weekend versus weekday night stay, average lead time, cancellation rate, and repeat guest rate). Then I created a new dataframe called Top5_df enriched with all the computed data.

I created two bar plots showing 'Booking count of top 5 countries' and 'Total Nights Stayed', and found no big variance between them, both displaying same trends across countries. After that, I went further plotting bar plots of ' Avg_night_stay by Country', 'Cancellation_rate_by_country', 'Repeat_Guest_Rate', 'Average days between bookings and arrivals' to get insight into regional travel habits.

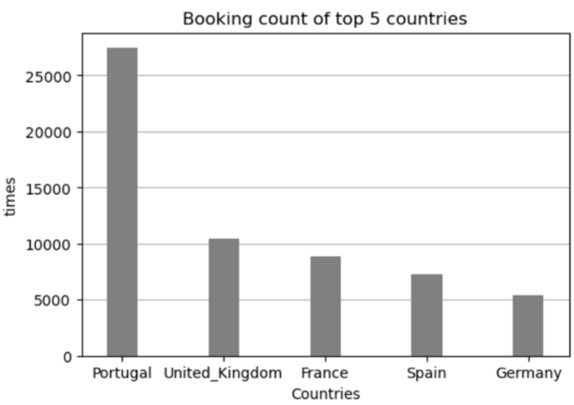
Key insights I gain:

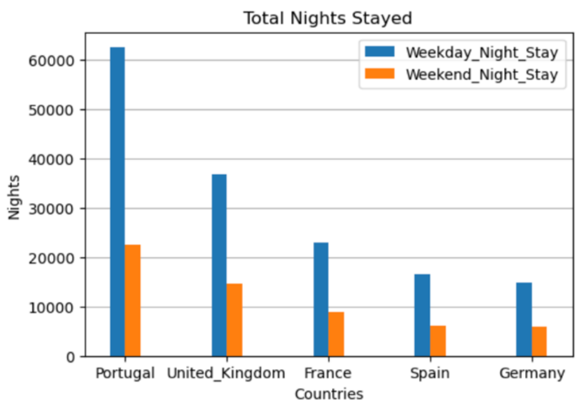
Top 5 countries account for 68.3 percent of the total bookings. From the plot ”Booking count of top 5 countries”, Portugal has the highest number with almost 50K bookings, followed by the United Kingdom and France with approximately 11,000 and 10,000 bookings respectively. That also explains why Portugal data leads the weekend and weekday night stay in the ”Total nights stayed” plot.

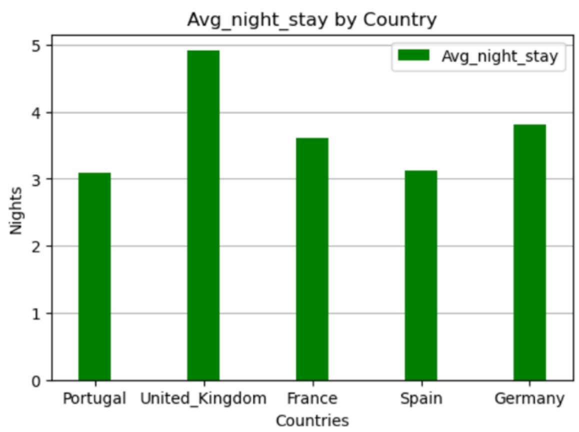
The plot showing the average length of stay per booking shows that Portugal and Spain has the short average stay for 3 nights. This pattern suggests that Portuguese and Spanish travelers tend to have frequent short trips over the long ones.

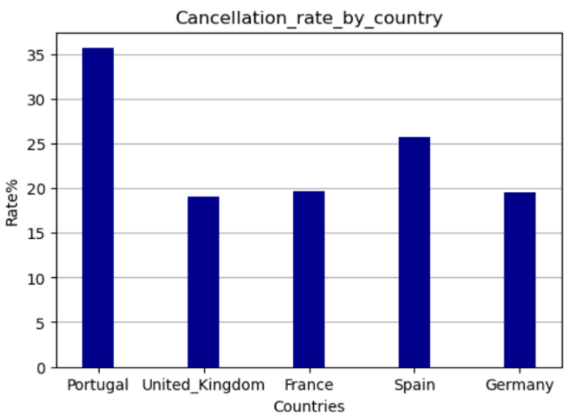
When we look at the plot for cancellation rate and repeat guest rate, they revel that Portugal is the highest in both. Comparing them to Portugal's high booking numbers and shortest average stay duration, we can infer that the hotel might locate in or near Portugal, and the people make and cancel reservations at a higher rate than visitors from other countries.

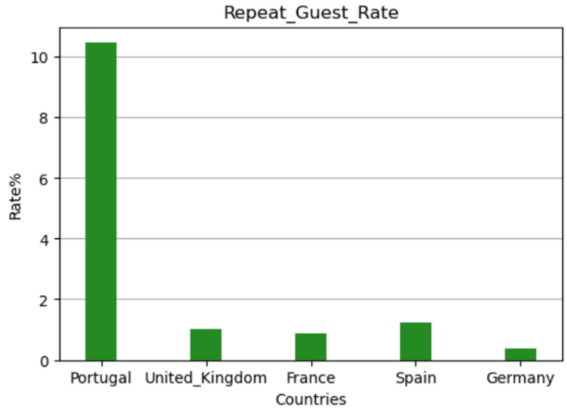
From the plot for cancellation rate, and Average time between bookings and arrivals, we can tell that people from the UK and Germany like to book further in advance than people from other four countries and german travelers have the lowest tendency to change their idea and cancel reservations once made.

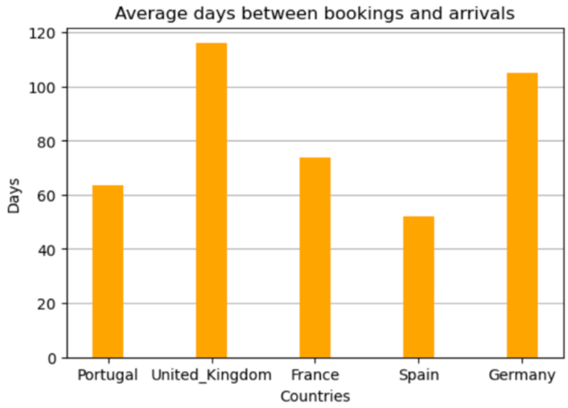









Repeat_Guest_Rate
Rate%
0
2
4
6
8
10
Portugal
United_Kingdom
France
Spain
Germany
Countries


Average days between bookings and arrivals
Days
0
20
40
60
80
100
120
Portugal
United_Kingdom
France
Spain
Germany
Countries