Practical Data Science with Python

S4098607

Assignment 1: Data Cleaning and Summarising

Rakshana N, rakshananagalingam21@gmail.

DEAKIN
BUSINESS
SCHOOL

CONTENTS

1.Introduction 2
2. Data Preparation 3
2.1. Missing Values 3
2.2. Duplicate Records 3
2.3 Standardization and Formatting 3
2.4 Outlier Detection and Handling 3
2.5 Final Cleaned Dataset 3
3. Data Exploration 4
3.1 Monthly Trends in 2016 4
3.2 Seasonal and Yearly Relationships (2015–2017) 5
3.3 Bookings by Country 6
4.Conclusion 7

Introduction

The first phase of any data science project is crucial in laying the groundwork for meaningful analysis. This assignment focuses on the essential initial steps in the data science workflow—data cleaning and exploration using Python and popular libraries such as pandas and matplotlib. The dataset provided contains hotel booking information for a city hotel and a resort hotel, including various features like reservation dates, guest demographics, booking durations, and cancellation records.
The primary goal of this project is to ensure the dataset is in a suitable format for analysis by identifying and resolving inconsistencies, missing values, and outliers. Once the data is cleaned and standardized, the second stage involves generating exploratory insights through visualizations and statistical summaries. These insights help uncover trends in bookings over time, seasonal patterns, and geographical customer behaviour. By following a structured and methodical approach, this assignment demonstrates the importance of preparing and understanding data before applying more advanced analytics or machine learning techniques. All coding tasks were performed in Jupyter Notebook using the specified Anaconda environment, ensuring reproducibility and consistency.

2. Data Preparation
This section outlines the steps followed to clean and prepare the dataset before analysis.

2.1 Missing Values
Missing values were detected using df.isnull().sum(). The columns with missing values included:

- children
- country
- agent
- company

Approach to Handling Missing Values:

- children: Filled with **0**, assuming bookings had no children.
- country: Filled with **'Unknown'** for undefined origins.
- agent and company: Filled with **0**, implying bookings were made directly.

2.2 Duplicate Records
Duplicates were identified using df.duplicated().sum() and revealed **32,011** duplicate rows.
Action: Removed using df.drop_duplicates().

2.3 Standardization and Formatting

- **Date Fields:** Converted reservation_status_date and arrival_date to datetime using pd.to_datetime().
- **Text Fields:** Standardized fields like hotel, customer_type, and market_segment using .str.strip().str.lower() to ensure consistency.

2.4 Outlier Detection and Handling

- Identified unusually high lead times (lead_time > 500) using boxplots and z-scores.
- Applied capping at the **99th percentile** to handle extreme lead time values.
- Removed records where both adults and children were zero, as they were invalid.

2.5 Final Cleaned Dataset

- **Total Rows:** 87,214
- **Total Columns:** 32
- The cleaned dataset was saved as cleaned_version.csv.

---

3. Data Exploration

This section focuses on discovering patterns in the cleaned dataset.

3.1 Monthly Trends in 2016

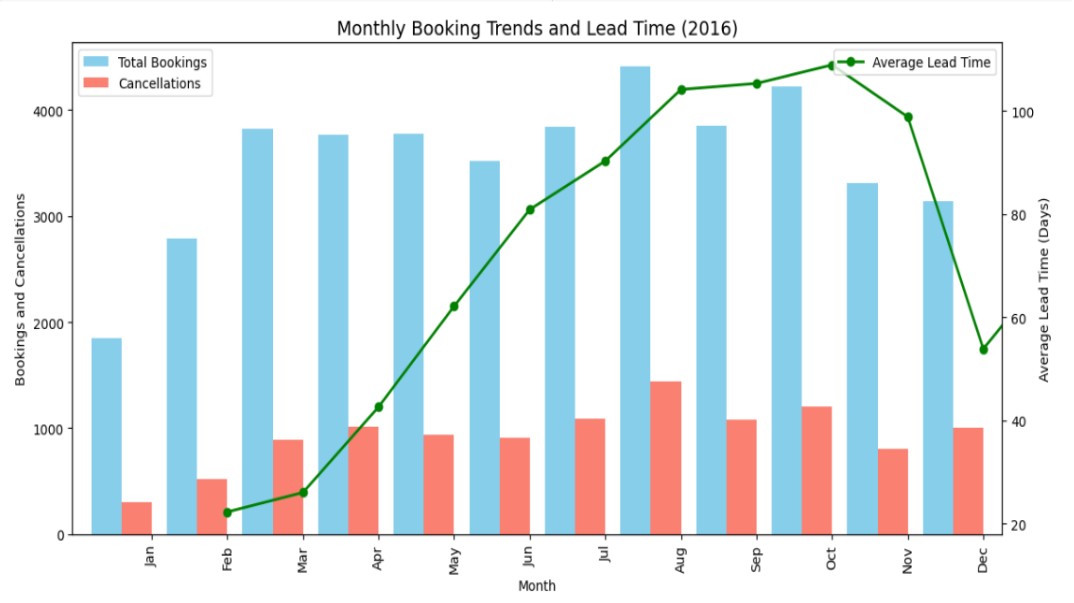
Metrics Analysed:

- Total Bookings
- Cancellations
- Average Lead Time

Key Findings:

- Bookings peaked in **July and August 2016**.
- **Cancellations** were highest in **August**.
- Lead times were longest in early 2016, suggesting early planning.

Visualizations:



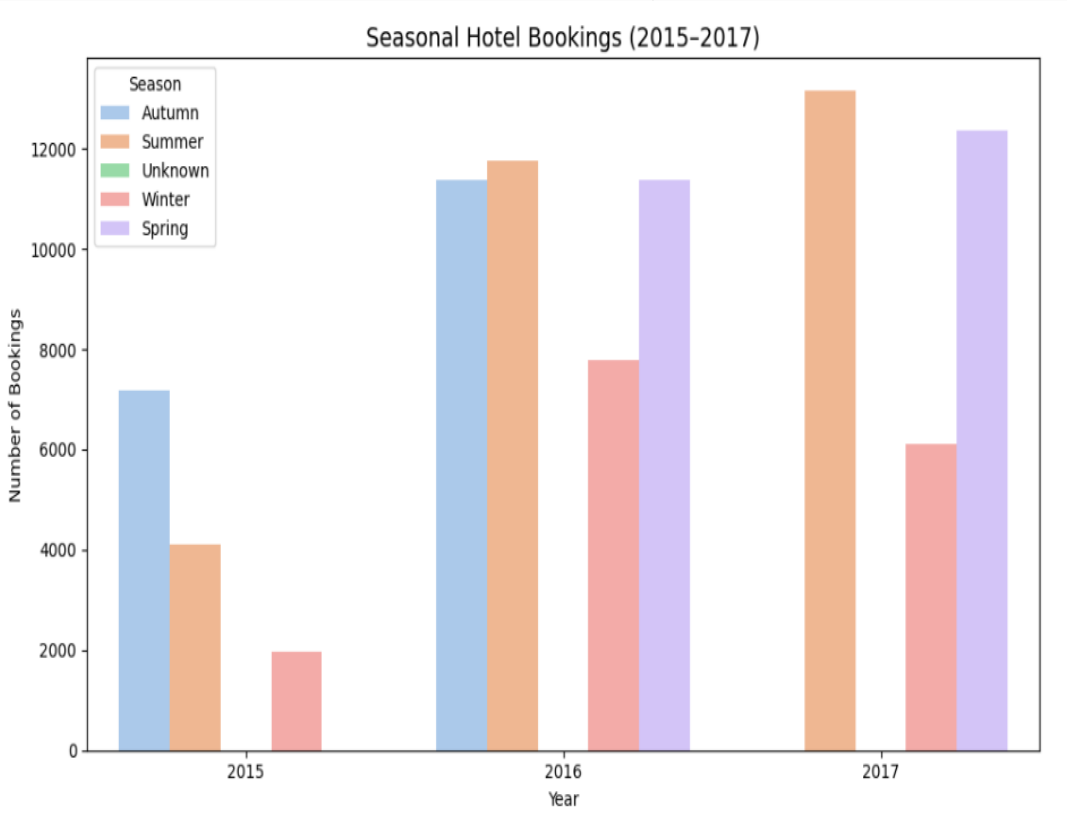
3.2 Seasonal and Yearly Relationships (2015–2017)

Booking patterns across different seasons and years were evaluated.

Observations:

- **Summer (high season)** saw increased bookings.
- **City hotels** experienced peak occupancy in Q3 annually.
- Seasonal changes had noticeable effects on customer behaviour.

Visualizations:



3.3 Bookings by Country

Analysed booking volume and customer behaviour by country.

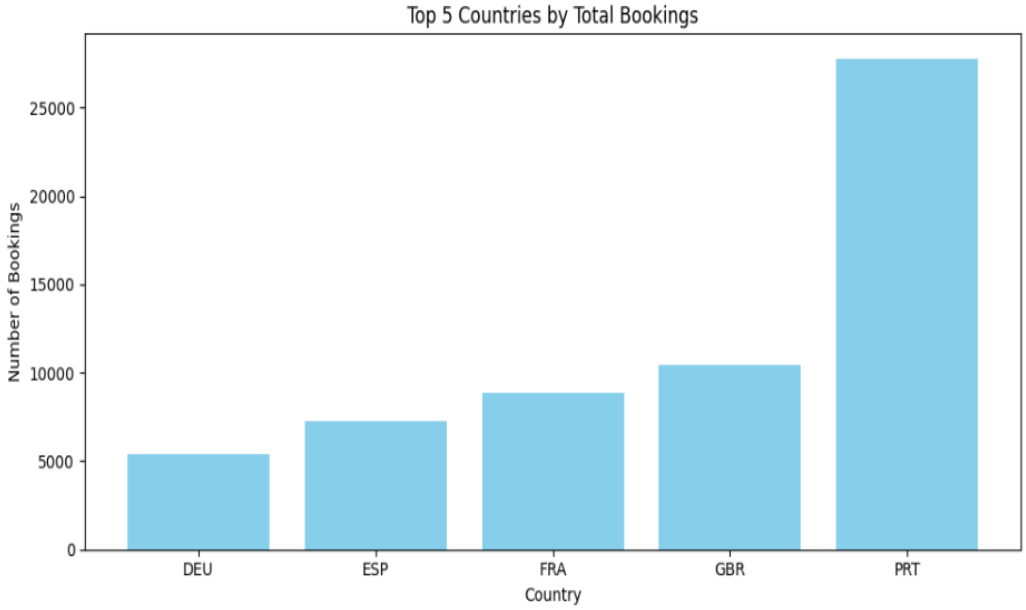
Top 5 Countries by Volume:

1. Portugal
2. United Kingdom
3. France
4. Spain
5. Germany

Customer Insights:

- **UK** visitors had longer stayed on average.
- **Portugal** had the **highest cancellation rate**.
- **France** and **Spain** were mostly associated with shorter stays.

Visualizations:



	country	cancellation_rate	avg_week_stays	avg_weekend_stays
0	DEU	0.195543	2.736305	1.080780
1	ESP	0.257040	2.282855	0.839177
2	FRA	0.196305	2.597529	1.011674
3	GBR	0.190348	3.510889	1.396431
4	PRT	0.353089	2.268092	0.812048

4. Conclusion

The analysis covered comprehensive data preparation and exploratory insights using the hotel bookings dataset. Through cleaning processes like handling missing values, duplicates, outliers, and standardization, the dataset was optimized for further analysis. Key insights revealed patterns in seasonal bookings, cancellation trends, and country-wise customer behaviour. These insights can guide hotel marketing strategies, improve service offerings, and ultimately increase customer satisfaction and retention.