HOTEL BOOKINGS ANALYSIS

PROJECT REPORT

ABSTRACT

Analysis of hotel bookings for marketing campaign recommendations based on hotel type, target audience and campaign timeline.

Hassanat, Tamsir and Tabatha

INTRODUCTION

We analysed the hotel booking data for a given company. Our aim was to enable the company to make informed decisions about marketing, leading to increased revenue. The company has two kinds of hotel: city and resort hotels. We were provided with bookings data collected from 2015 to 2017.

To make suggestions about targeted marketing, we came up with 5 questions that were answered with the data. The questions are:

1. What is the distribution of the average daily rate for each hotel type?
2. Which hotel type is mostly booked and then occupied?
3. Which months are the hotels mostly occupied?
4. Where do most of the guests come from?
5. Are the guests usually families or individuals?

The first three questions helped with determining the hotel type to be emphasised in the marketing campaign as well as the best time to carry out the campaign while the last two questions helped determine the company's target audience.

IMPLEMENTATION AND EXECUTION

DATASET: The dataset contained 100,000+ rows and 32 columns. Each row provided details of an individual booking. As part of the preprocessing, we removed redundant rows and columns. We also had to engineer new features such as the arrival_date and full_length. The following important columns were used in the analysis:

- **hotel**, categorical variable, described whether the row entries were for a city or resort hotel
- **arrival_date** indicated the date, month, and year of the guest's arrival at the hotel
- **adults**, **children** and **babies** columns, quantitative variables, showed their respective numbers for a given booking
- **country,** categorical variable, showed the origin of the booking

- **stays_in_weekend_nights,** quantitative variable, number of nights spent at the hotel on a weekend
- **stays_in_week_nights,** quantitative variable, number of nights spent at the hotel on a week day
- **adr** (Average Daily Rate), quantitative variable, showed the sum of all lodging transactions divided by the total number of staying nights for a given day
- **reservation_status**, categorical variable, indicated the last reservation status of the guest. There were three possibilities:
    - *Canceled* – booking was canceled by the customer;
    - *Check-Out* – customer has checked in but already departed;
    - *No-Show* – customer did not check-in and did not inform the hotel of their reason
- **is_family**, binary variable, indicated the bookings that were made by families i.e. adults and/or children and/or babies
- **full_length,** quantitative variable, the total number of nights spent at the hotel

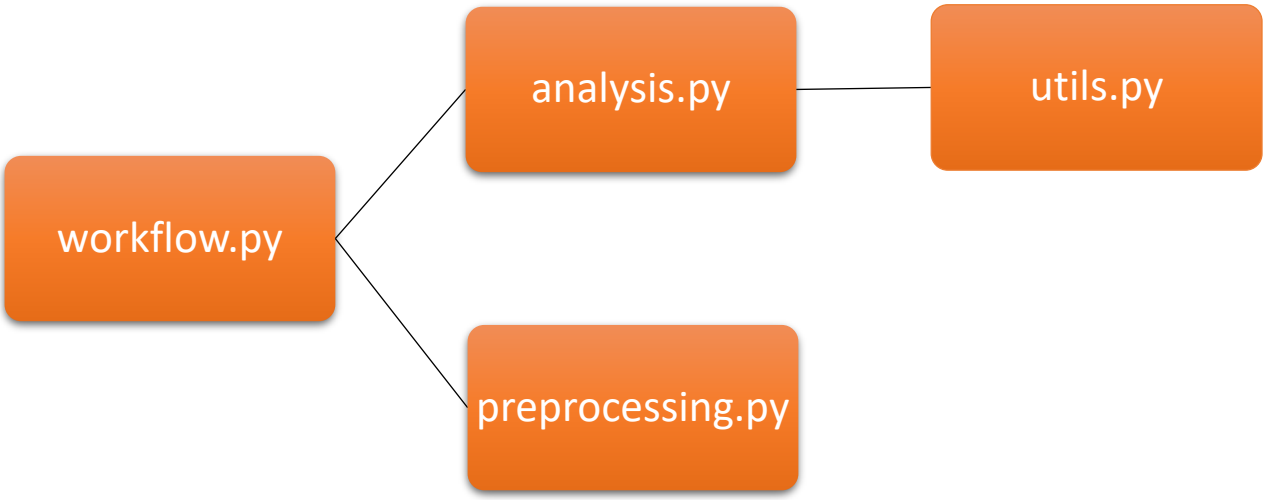


Figure 1: dependencies of the python scripts

PREPROCESSING: We created a script which contained functions to address each cleaning process for the dataset. The script was then imported into the workflow.py for the changes to be applied to the data. The functions contained in the preprocessing.py are:

1. **fillnull():** this function was used to fill null values with relevant values. Columns such as ‘children’, ‘agent’ and ‘company’ contained null entries. Some columns with null entries were ignored as the nulls were useful for the analysis such as the ‘country’ column. For the children column, we used the fillna function to fill the null values with 0.
2. **change_date():** this function was defined to unify the arrival details for the booking. The arrival date for each booking was split into day, month, week number and year. Firstly, the month was converted from the month name (e.g. July) to the month number (e.g. 7) using the calendar module. Then the day, month, and year were concatenated to replace the year column. Finally, the day, month and week number columns were dropped from the data.
3. **change_datatype():** this function was defined to change the data type for columns with the wrong data type. Arrival_date and reservation_status_date columns were changed to datetime using the pd.to_datetime() while children column was changed to integer using the astype()
4. **drop_column():** since a lot of the columns were not used in the analysis, we defined a function to remove all redundant columns.
5. **drop_duplicate():** the data contained a lot of duplicate entries. We defined this function to remove the duplicated rows and it was applied after the other cleaning functions had been applied.
6. **clean_data():** this function called on all the functions defined above and produced the final cleaned dataset.

ANALYSIS: Unlike the preprocessing, two scripts were created for the analysis. An **utils.py** which contained methods defined under a given class, each class was specific to an analysis question listed above. Then an **analysis.py** in which the classes from utils were imported and the methods applied. We also had a **workflow.py** which combined the processes from preprocessing.py and analysis.py.

A. utils.py: a module which contained six classes: A typical class answered an analysis question by containing methods that i) compute statistics, ii) plot visualisation and save to PDF iii) save summary statistics to CSV

```
class AdrStats:
    def __init__(self, hotel_data):

    def compute_stats(self):
            return adr_stats

    def plot_fig(self, filename):

    def summary_csv(self, filename):

class …
```

B. analysis.py: here, separate functions were defined to invoke all the methods from each class in the utils.py. A typical function takes in the dataset as a parameter, the first method which computes statistics is applied to the parameter and the result is saved as a variable. Then the other methods for visualising and transforming to a csv are applied to the variable. The file path to which the visuals and CSV files are saved is specified as the parameter for the methods. A last function called 'all_analysis' then combined all the functions defined earlier.

```
from utils import AdrStats, …

def analyse_adr(bookings):
    adr_stats = AdrStats(bookings)
    adr_stats.plot_fig('filepath')
    adr_stats.summary_csv(filepath')

def analyse_reservation_status (bookings) …

def all_analysis(bookings):
    analyse_adr(bookings)
    analyse_reservation_status(bookings)
```

C. workflow.py: this is the final script which combines the data cleaning and analysis processes. It imports the clean_data function from preprocessing.py and the all_analysis function from analysis.py. It reads the data using pandas read method and assigns it to a variable. Then the clean_data and all_analysis functions are applied to the variable to produce all the analysis results (statistics and graphs) at once. In this case, the cleaned data is not produced independently. The cleaned data is only produced when the preprocessing.py is directly executed.

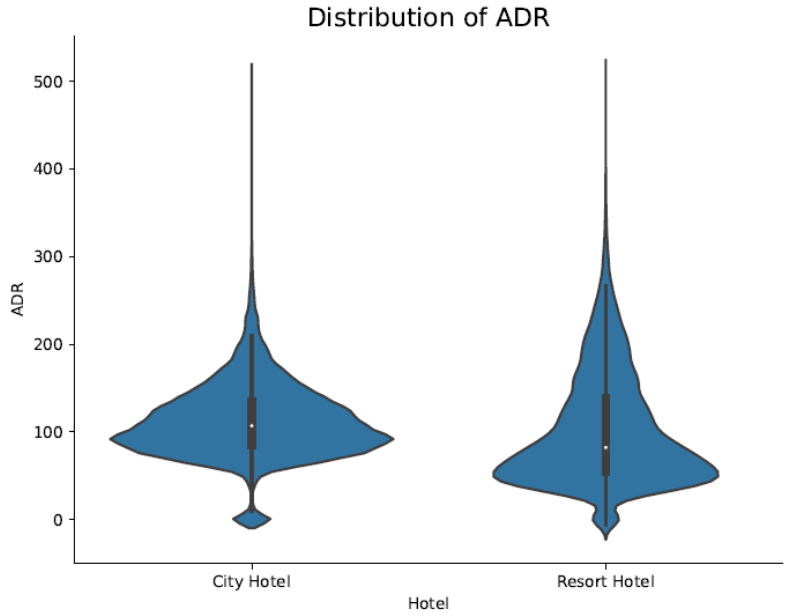
```
from preprocessing import clean_data
from analysis import all_analysis


if __name__ == "__main__":
    hotel_data = pd.read_csv(r'hotel_bookings.csv', encoding='utf-8')
    clean_data(hotel_data)
    all_analysis(hotel_data)
```

RESULTS

After the analysis, we were able to answer the proposed questions.

1. What is the distribution of the average daily rate for each hotel type?

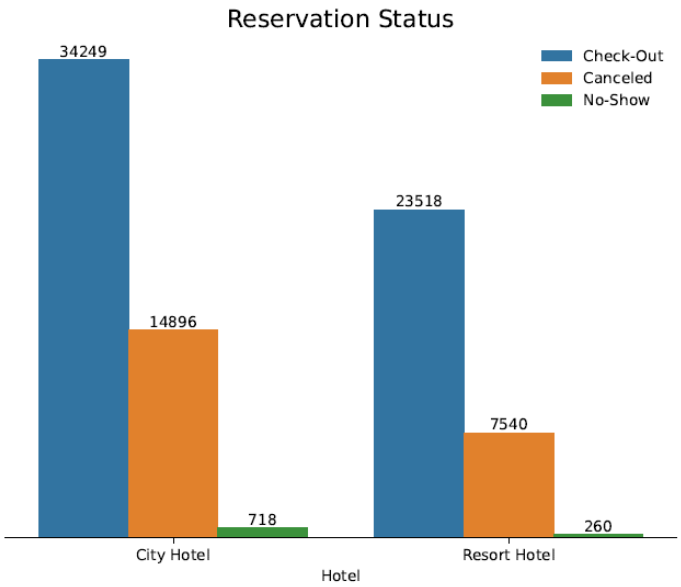
   Both hotels have a unimodal distribution for ADR. A lot of points for City hotel lie around 90-100 while the common ADR value for Resort hotel is around 50. For City hotel, there was a maximum outlier of 5400. After it was removed, the max ADR became 510 while that of Resort hotel was 508. Even then, these values lie in the outlier region as indicated in the violin plots used to show the distribution. Also, the outliers for the ADR in Resort hotel are slightly higher than that of City hotel.



2. Which hotel type is mostly booked and then occupied?

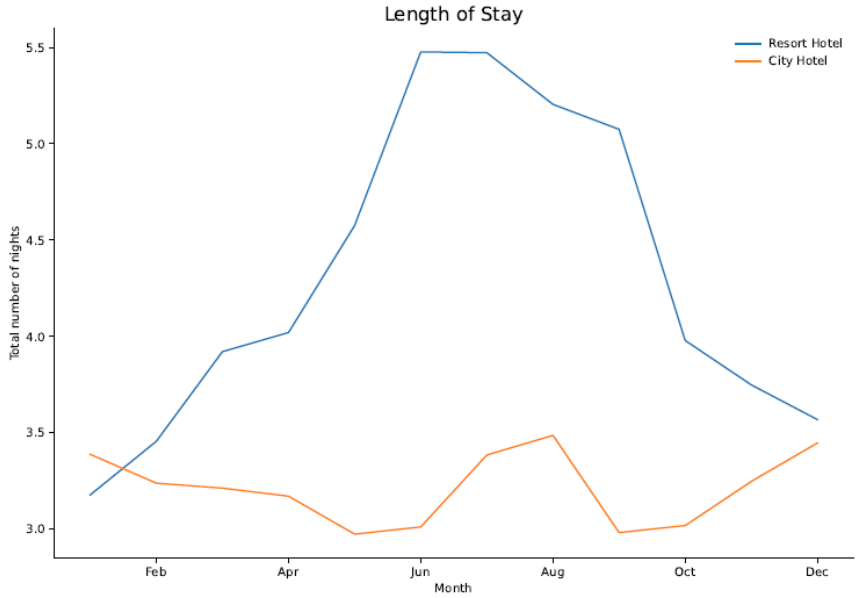
   City hotels had 61.42% of the total bookings while Resort had 38.58%. Comparing the ratio for Checkout, Canceled and No-show reservation status for each, we saw that city

bookings had (0.68 : 0.30 : 0.01) while resort had (0.75 : 0.24 : 0.008). This shows that there is a higher probability for a booking to be cancelled in a City hotel than in a Resort hotel. However, during the period that the data was collected, **City hotel was mostly booked,** and **Resort hotel had a higher proportion of check-out** (guests checked in and fulfilled their bookings).



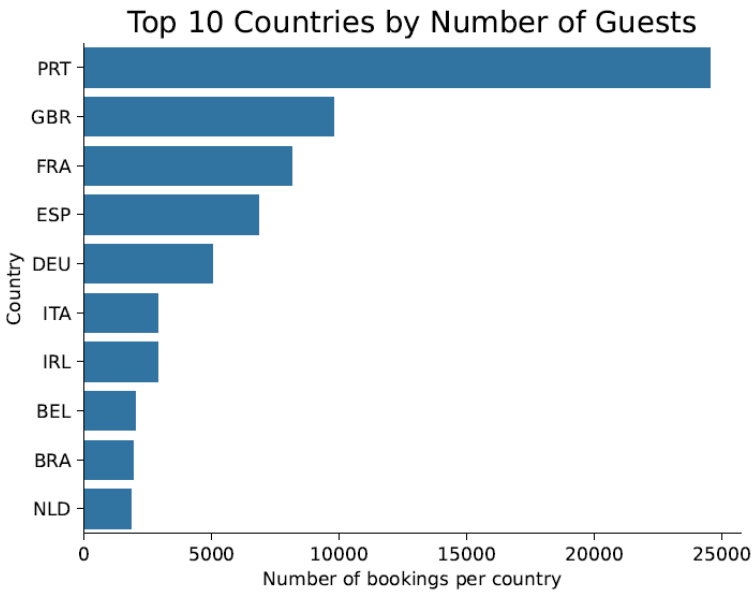
3. Which months are the hotels mostly occupied?

   Using data from the last two years, we observed that **guests tend to stay longer in the hotels during the summer**, especially for the Resort hotels. July and August had the mean highest number of nights which is 5 nights for Resort and 3.4 for City hotel. While January had the mean lowest number of nights for Resort, May had the mean lowest for City which seems contrasting with our earlier observation because May is a summer month.



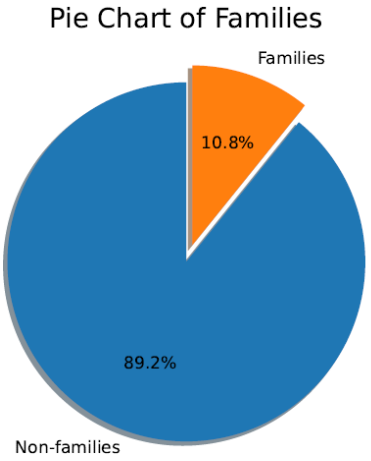
4. Where do most of the guests comes from?

   By filtering the results to the top 10 countries with the highest number of bookings, **Portugal had the highest count with 24,520 bookings**. The other countries on the list are all in Europe except Brazil and this suggests that the majority of the guests come from Europe. Some bookings did not indicate countries, they had a non-significant count.



5. Are the guests usually families or individuals?

   We sorted families as bookings that included adults and/or children and/or babies, and non-families as otherwise. We did not consider two or more adults (without children or babies) that could be couples or siblings because we cannot differentiate them from two or more adults that are friends or co-workers. Also, we did not consider bookings that were for only children or babies as they could be school kids on an excursion. Hence, the **percentage of families was 10.8%** which is really low compared to 89.2% for non-families.



CONCLUSION

The aim of the analysis was to make campaign recommendations for the marketing team. We streamlined our analysis to answer questions relating to the hotel type to emphasize in the campaign, the target audience, and the period during which the campaign should be carried out.

The average daily rate (ADR) was generally higher in City hotel than Resort allowing it to generate more revenue for the company however, there was a higher chance that a booking in a City hotel will be cancelled than in a Resort, leading to a loss of revenue. Hence a balance may be achieved by campaigning equally for both hotels.

Most of the guests were from Europe making it a great place to focus the campaign. Since very few numbers of guests were identified as families, the company might consider making the adverts more family-friendly, especially in the months approaching summer. In City hotels, special campaign may also be created for the winter holidays.