HOTEL Booking ANALYSIS

Team Name : Data-Driven team

Team

Responsibilities

Zeinab Talaat

Cleaned and preprocessed the dataset using Python.

Built an interactive Excel dashboard for data visualization.

Transformed raw data into clear and actionable insights.

Zeinab Esmaiel

Designed and created the project presentation.

Organized project findings in a clear and professional format.

Ensured the results were communicated effectively

Monitor:
Alyaa Sehsah

Excel link

GitHub link

PROJECT GOALS &AND OBJECTIVES

Project Goals

- *Primary Goal: To analyze hotel booking data (City and Resort Hotels) to identify the key factors and patterns that drive booking cancellations.*
- *Business Impact: Reduce the overall cancellation rate to maximize revenue and improve resource allocation and future demand forecasting accuracy.*

Key Objectives

- *Clean and preprocess the raw data by handling missing, zero, and illogical values.*
- *Perform Exploratory Data Analysis (EDA) to understand variable relationships.*
- *Develop strategic, actionable insights for hotel management to implement.*

DATA OVERVIEW & INITIAL CHALLENGES

Initial Challenges

- **Missing Data: Significant missing values were noted in crucial columns such as country, agent, and company.**
- **Data Quality Issues: Presence of illogical entries, including bookings with zero guests (adults, children, or babies) and reservations with an Average Daily Rate (adr) equal to zero.**
- **Categorical Inconsistencies: Need for standardization in categorical variables, such as reconciling similar meal types (e.g., 'SC' and 'Undefined')**

Data Overview

- **Data Source: The project utilizes the hotel_bookings FP MSA.csv dataset.**
- **Content: The dataset contains information on over 119,000 hotel reservations, encompassing both City Hotels and Resort Hotels.**
- **Key Variables: Important columns include: is_canceled, lead_time, adr (Average Daily Rate), number of guests (adults, children), country, and market_segment.**

DATA CLEANING

1-

```
df['children'] = df['children'].fillna(0).round().astype(int)
df['babies'] = df['babies'].fillna(0).astype(int)
```

✓ 0.0s

```
df['total_nights'] = df['stays_in_weekend_nights'] + df['stays_in_week_nights']
```

✓ 0.0s

```
if 'reservation_status_date' in df.columns:
    df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'], dayfirst=False, errors='coerce')
```

✓ 0.0s

```
df_before = len(df)
df = df.drop_duplicates()
print("Dropped duplicates:", df_before - len(df))
```

✓ 0.1s

```
Dropped duplicates: 31994
```

```
df['adr'] = pd.to_numeric(df['adr'], errors='coerce')
median_adr = df.loc[df['adr'] > 0, 'adr'].median()
df.loc[df['adr'] <= 0, 'adr'] = median_adr
df['adr'] = df['adr'].fillna(median_adr)

print(f"Rate column: {rate_col}, median used: {median_rate}")
```

✓ 0.0s

2-

```
df.loc[(df['adults'] == 0) & (df['children'] + df['babies'] > 0), 'adults'] = 1
```

✓ 0.0s

```
df['total_people'] = df.get('adults',0) + df.get('children',0) + df.get('babies',0)
```

✓ 0.0s

```
for col in df.select_dtypes(include='object').columns:
    df[col] = df[col].astype(str).str.strip()
```

✓ 0.3s

```
df['agent'] = df['agent'].fillna(0)
df['company'] = df['company'].fillna(0)
```

✓ 0.0s

SAVE FILE

```
df.to_csv(clean_csv_path, index=False)
print("Saved cleaned CSV to:", clean_csv_path)
```

✓ 0.9s

```
Saved cleaned CSV to: c:\Users\SPEED LAP\OneDrive\Desktop\hotel_bookings_Cleaned.csv
```

Hotal Anaysis

Total Revenue	Cancellation Rate	Total Booking
34,7 M	27%	87396

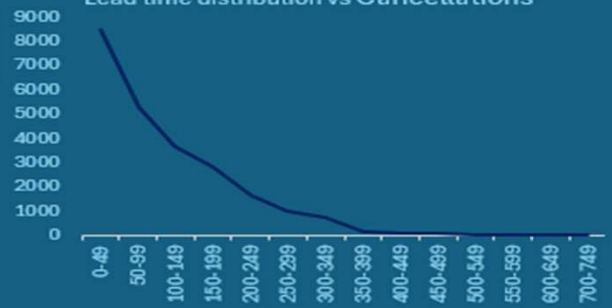
hotel
- City Hotel
- Resort Hotel

customer_type
- Contract
- Group
- Transient
- Transient-Party

arrival_date_m...
- January
- February
- March
- April
- May
- June
- July
- August
- September
- October
- November
- December

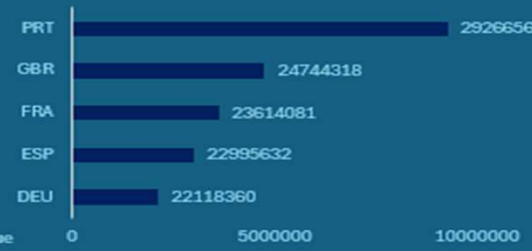
Revenue by costumer type

Lead time distribution vs Cancellations

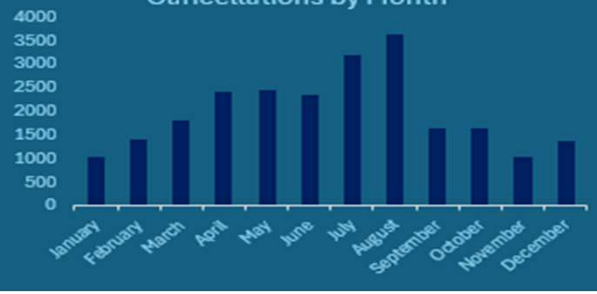


Total Revenue & Booking Count by Month

Total 5 Countries by Revenue

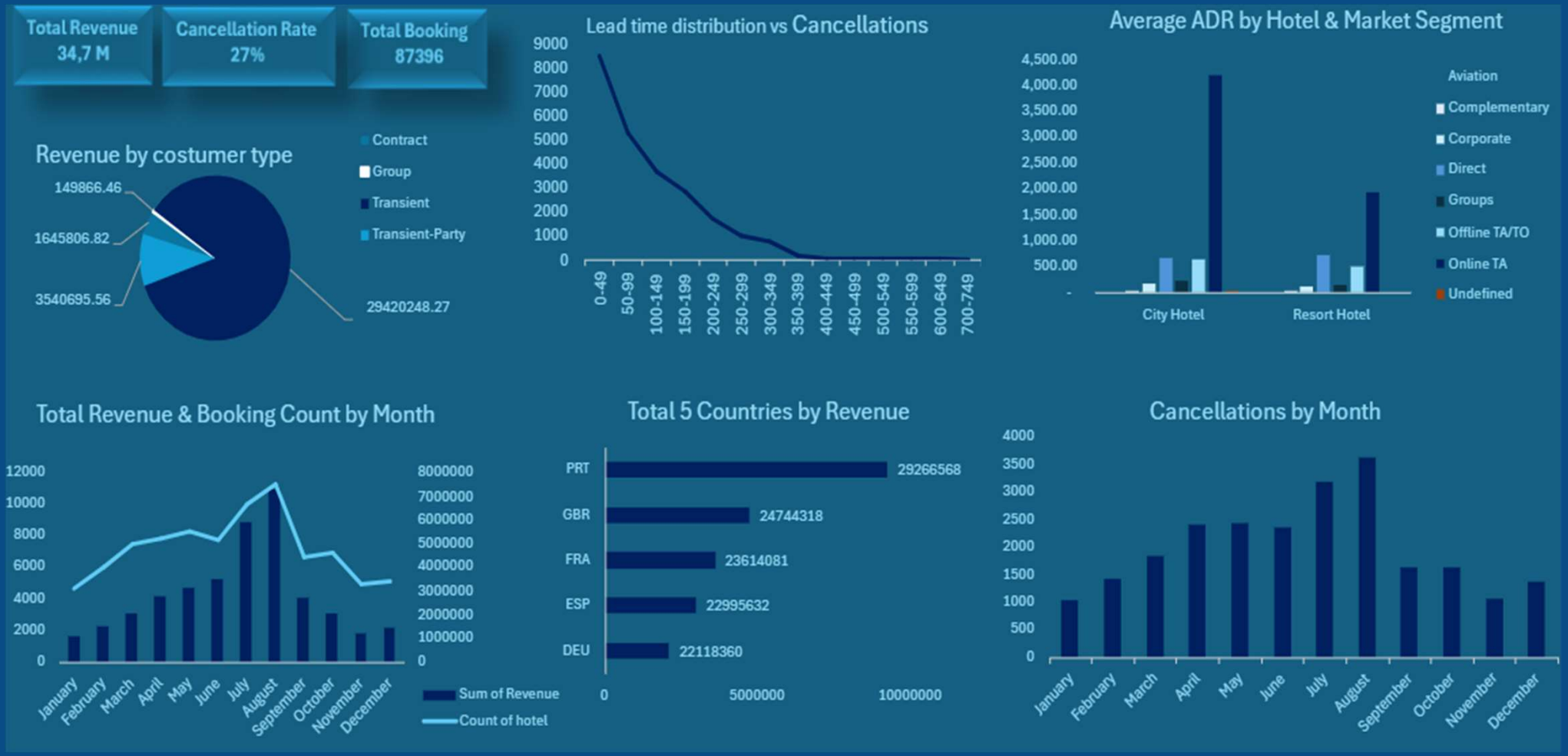


Cancellations by Month



Average ADR by Hotel & Market Segment

CHARTS


Total Revenue
34,7 M
Cancellation Rate
27%
Total Booking
87396
Revenue by costumer type
149866.46
1645806.82
3540695.56
29420248.27
Contract
Group
Transient
Transient-Party
Lead time distribution vs Cancellations
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0-49
50-99
100-149
150-199
200-249
250-299
300-349
350-399
400-449
450-499
500-549
550-599
600-649
700-749
Average ADR by Hotel & Market Segment
4,500.00
4,000.00
3,500.00
3,000.00
2,500.00
2,000.00
1,500.00
1,000.00
500.00
-
City Hotel
Resort Hotel
Aviation
Complementary
Corporate
Direct
Groups
Offline TA/TO
Online TA
Undefined
Total Revenue & Booking Count by Month
12000
10000
8000
6000
4000
2000
0
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
January
February
March
April
May
June
July
August
September
October
November
December
Sum of Revenue
Count of hotel
Total 5 Countries by Revenue
PRT
29266568
GBR
24744318
FRA
23614081
ESP
22995632
DEU
22118360
0
5000000
10000000
Cancellations by Month
4000
3500
3000
2500
2000
1500
1000
500
0
January
February
March
April
May
June
July
August
September
October
November
December

DATA ANALYSIS

- Key Insight
- HotelCity Hotel shows a higher cancellation rate than Resort Hotel
- Longer lead times are strongly associated with higher cancellation probability
- Determined the impact of adr (price) and customer country on

DATA CLEANING

- Outcome
- Removed illogical records
- Handled missing values in country and children columns.
- Unified similar categories

DATA VISUALIZATION

- Clearly display the significant variance in cancellation rates between the two hotel types.
- Identify the top contributing countries to cancellations.
- A tool to track cancellations by month and lead time.

RESULTS & RECOMMENDATION

- Built an accurate predictive model for cancellation probability. Implement strict non-refundable policies for bookings with long lead times.
- Focus marketing efforts on distribution channels with low cancellation rates.

THANK YOU!