

hotel booking analysis

Final Project



Jack LinDen dass V 112721214006

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

hotel\_data = pd.read\_csv("hotel\_bookings.csv")

# Display the first few rows of the dataset

print(hotel\_data.head())

# Check for missing values

print(hotel\_data.isnull().sum())

# Summary statistics for numerical columns

print(hotel\_data.describe())

# Visualize the distribution of numerical columns

sns.histplot(data=hotel\_data, x='lead\_time', bins=30, kde=True)

plt.title('Lead Time Distribution')

plt.xlabel('Lead Time (days)')

plt.ylabel('Frequency')

plt.show()

sns.boxplot(data=hotel\_data, x='hotel', y='lead\_time')

plt.title('Lead Time by Hotel Type')

plt.xlabel('Hotel')

plt.ylabel('Lead Time (days)')

plt.show()

# Feature Engineering: Calculate total number of guests

hotel\_data['total\_guests'] = hotel\_data['adults'] + hotel\_data['children'] + hotel\_data['babies']

# Analysis: Relationship between total guests and special requests

sns.barplot(data=hotel\_data, x='total\_guests', y='total\_of\_special\_requests')

plt.title('Average Special Requests by Total Guests')

plt.xlabel('Total Guests')

plt.ylabel('Average Special Requests')

plt.show()

# Predictive Modeling: Example using logistic regression

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

# Select relevant features and target variable

X = hotel\_data[['lead\_time', 'total\_guests']]

y = hotel\_data['total\_of\_special\_requests']

# Split data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Train a logistic regression model

model = LogisticRegression()

model.fit(X\_train, y\_train)

# Make predictions on the testing set

y\_pred = model.predict(X\_test)

# Evaluate model performance

accuracy = accuracy\_score(y\_test, y\_pred)

print("Accuracy:", accuracy)

# Interpretation: Coefficients of logistic regression model

coefficients = pd.DataFrame({'Feature': X.columns, 'Coefficient': model.coef\_[0]})

print(coefficients)

hotel is\_canceled lead\_time arrival\_date\_year arrival\_date\_month \

0 Resort Hotel 0 342 2015.0 July

1 Resort Hotel 0 737 2015.0 July

2 Resort Hotel 0 7 2015.0 July

3 Resort Hotel 0 13 2015.0 July

4 Resort Hotel 0 14 2015.0 July

arrival\_date\_week\_number arrival\_date\_day\_of\_month \

0 27.0 1.0

1 27.0 1.0

2 27.0 1.0

3 27.0 1.0

4 27.0 1.0

stays\_in\_weekend\_nights stays\_in\_week\_nights adults ... deposit\_type \

0 0.0 0.0 2.0 ... No Deposit

1 0.0 0.0 2.0 ... No Deposit

2 0.0 1.0 1.0 ... No Deposit

3 0.0 1.0 1.0 ... No Deposit

4 0.0 2.0 2.0 ... No Deposit

agent company days\_in\_waiting\_list customer\_type adr \

0 NaN NaN 0.0 Transient 0.0

1 NaN NaN 0.0 Transient 0.0

2 NaN NaN 0.0 Transient 75.0

3 304.0 NaN 0.0 Transient 75.0

4 240.0 NaN 0.0 Transient 98.0

required\_car\_parking\_spaces total\_of\_special\_requests reservation\_status \

0 0.0 0.0 Check-Out

1 0.0 0.0 Check-Out

2 0.0 0.0 Check-Out

3 0.0 0.0 Check-Out

4 0.0 1.0 Check-Out

reservation\_status\_date

0 7/1/2015

1 7/1/2015

2 7/2/2015

3 7/2/2015

4 7/3/2015

[5 rows x 32 columns]

hotel 0

is\_canceled 0

lead\_time 0

arrival\_date\_year 1

arrival\_date\_month 1

arrival\_date\_week\_number 1

arrival\_date\_day\_of\_month 1

stays\_in\_weekend\_nights 1

stays\_in\_week\_nights 1

adults 1

children 5

babies 1

meal 1

country 488

market\_segment 1

distribution\_channel 1

is\_repeated\_guest 1

previous\_cancellations 1

previous\_bookings\_not\_canceled 1

reserved\_room\_type 1

assigned\_room\_type 1

booking\_changes 1

deposit\_type 1

agent 14134

company 90736

days\_in\_waiting\_list 1

customer\_type 1

adr 1

required\_car\_parking\_spaces 1

total\_of\_special\_requests 1

reservation\_status 1

reservation\_status\_date 1

dtype: int64

is\_canceled lead\_time arrival\_date\_year \

count 96522.000000 96522.000000 96521.000000

mean 0.457958 106.469116 2016.033184

std 0.498232 108.253308 0.698458

min 0.000000 0.000000 2015.000000

25% 0.000000 18.000000 2016.000000

50% 0.000000 72.000000 2016.000000

75% 1.000000 164.000000 2017.000000

max 1.000000 737.000000 2017.000000

arrival\_date\_week\_number arrival\_date\_day\_of\_month \

count 96521.000000 96521.000000

mean 27.206515 15.793610

std 13.436611 8.781417

min 1.000000 1.000000

25% 17.000000 8.000000

50% 28.000000 16.000000

75% 38.000000 23.000000

max 53.000000 31.000000

stays\_in\_weekend\_nights stays\_in\_week\_nights adults \

count 96521.000000 96521.000000 96521.000000

mean 0.948312 2.571648 1.856353

std 1.026630 1.996440 0.589514

min 0.000000 0.000000 0.000000

25% 0.000000 1.000000 2.000000

50% 1.000000 2.000000 2.000000

75% 2.000000 3.000000 2.000000

max 19.000000 50.000000 55.000000

children babies is\_repeated\_guest previous\_cancellations \

count 96517.000000 96521.000000 96521.000000 96521.000000

mean 0.103640 0.008257 0.035982 0.107479

std 0.399478 0.100579 0.186246 0.937749

min 0.000000 0.000000 0.000000 0.000000

25% 0.000000 0.000000 0.000000 0.000000

50% 0.000000 0.000000 0.000000 0.000000

75% 0.000000 0.000000 0.000000 0.000000

max 10.000000 10.000000 1.000000 26.000000

previous\_bookings\_not\_canceled booking\_changes agent \

count 96521.000000 96521.000000 82388.000000

mean 0.164513 0.205375 99.649476

std 1.660200 0.629318 115.139713

min 0.000000 0.000000 1.000000

25% 0.000000 0.000000 9.000000

50% 0.000000 0.000000 21.000000

75% 0.000000 0.000000 240.000000

max 72.000000 21.000000 535.000000

company days\_in\_waiting\_list adr \

count 5786.000000 96521.000000 96521.000000

mean 179.843761 2.737674 98.768716

std 129.705606 19.117495 51.946093

min 6.000000 0.000000 -6.380000

25% 51.000000 0.000000 65.000000

50% 174.000000 0.000000 90.000000

75% 246.000000 0.000000 121.500000

max 543.000000 391.000000 5400.000000

required\_car\_parking\_spaces total\_of\_special\_requests

count 96521.000000 96521.000000

mean 0.069923 0.503642

std 0.258448 0.753641

min 0.000000 0.000000

25% 0.000000 0.000000

50% 0.000000 0.000000

75% 0.000000 1.000000

max 8.000000 5.000000

