Predicting Hotel Cancellations

Valerie Zixi Li

Sociology Ph.D. Student, Affiliated with Population Studies and Training Center

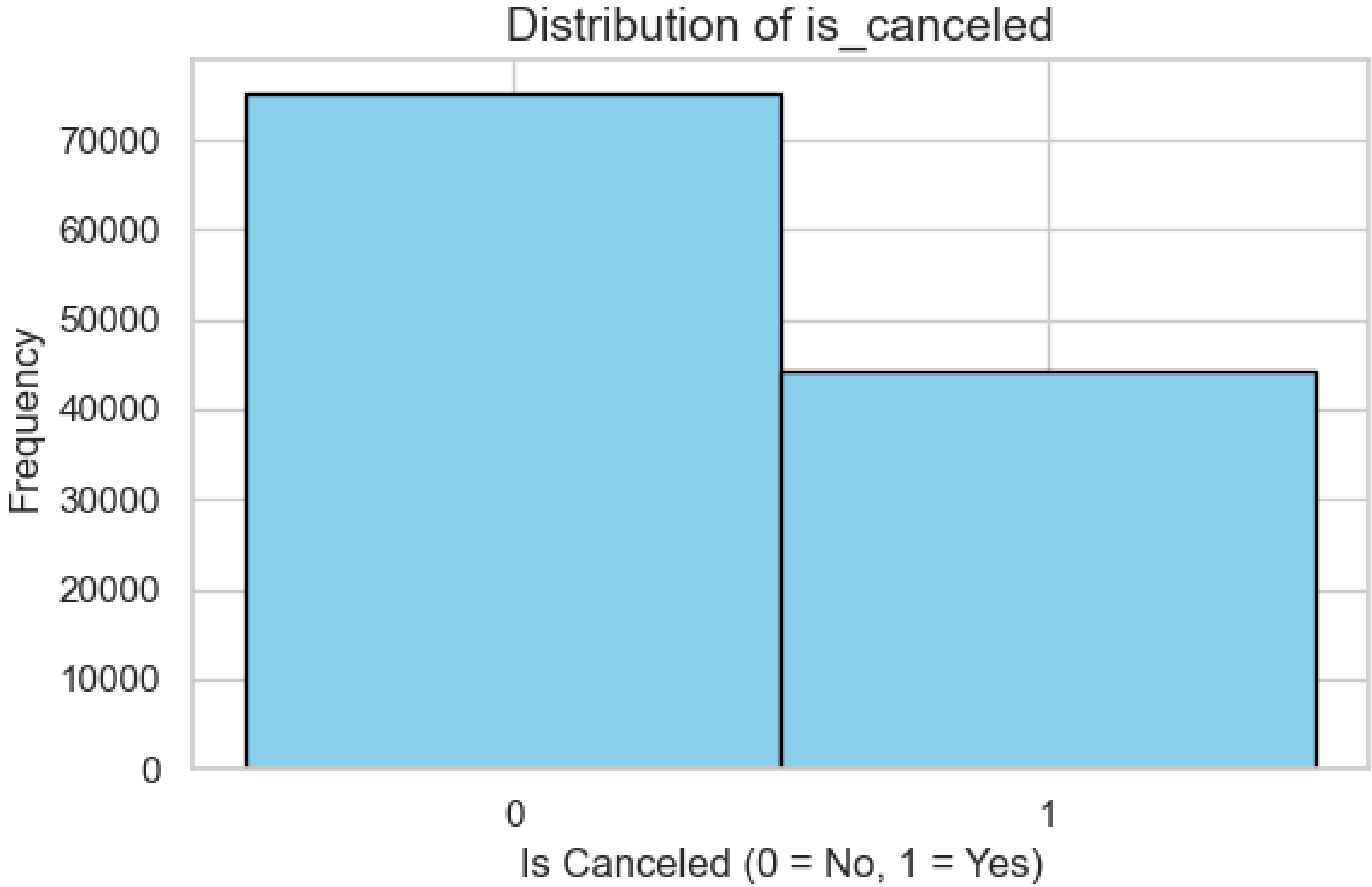
https://github.com/ZixiLi76/Hotel-Cancellation-Prediction

Oct 25 2024

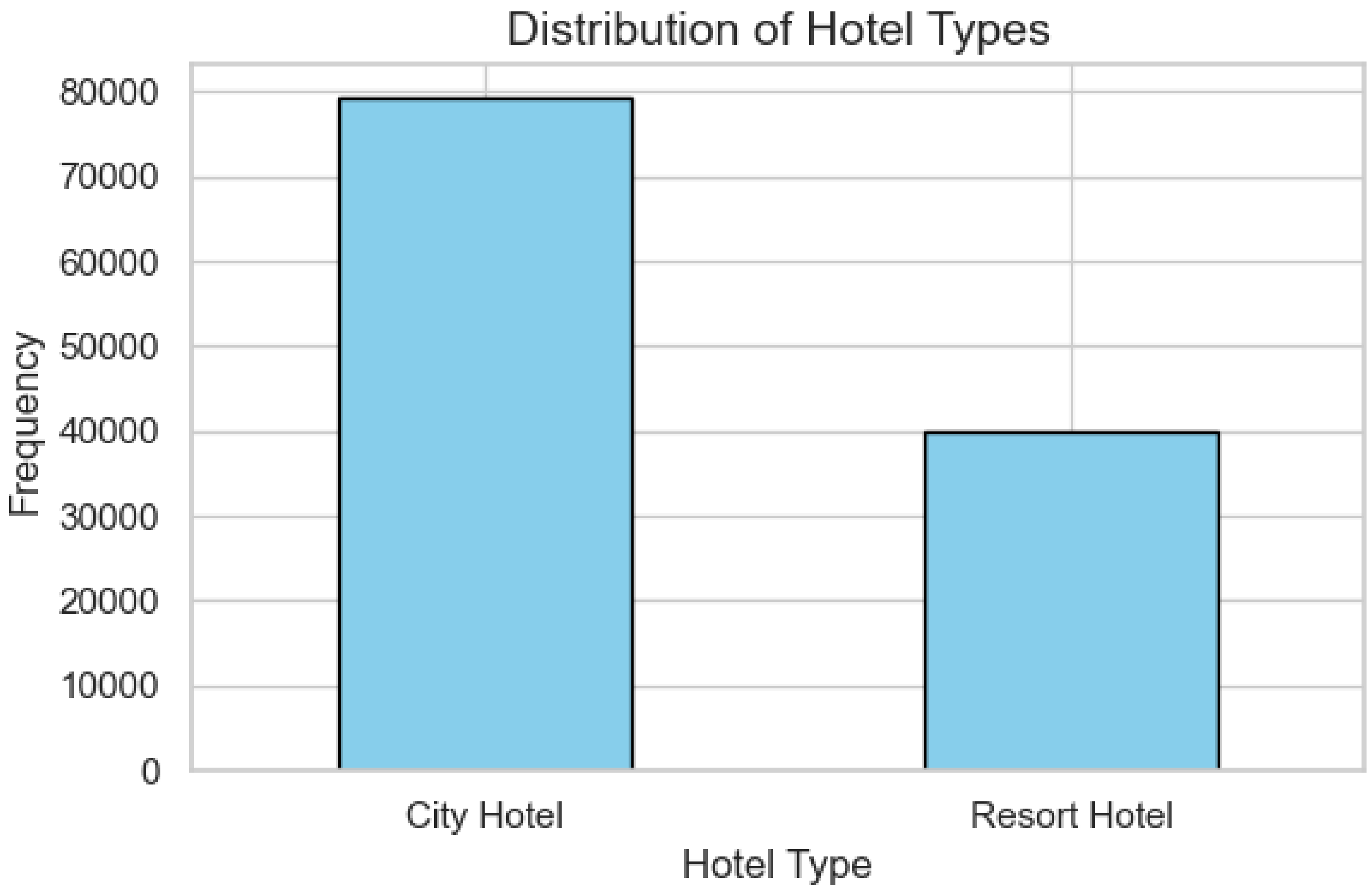
Question & Data

- Question: What factors affect hotel cancellations?
    - Optimize revenue management, enhance customer experience, improve operational efficiency
- Hotel booking demand dataset from Kaggle: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data
- The dataset: booking & cancellation details for city and resort hotels, featuring variables like booking dates, length of stay, guest count, # of special requests, …, with all personal information removed
- Collection methodology: data was sourced from ScienceDirect and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday, available on GitHub.
- Target variable (y): is_canceled (dichotomous) – classification problem

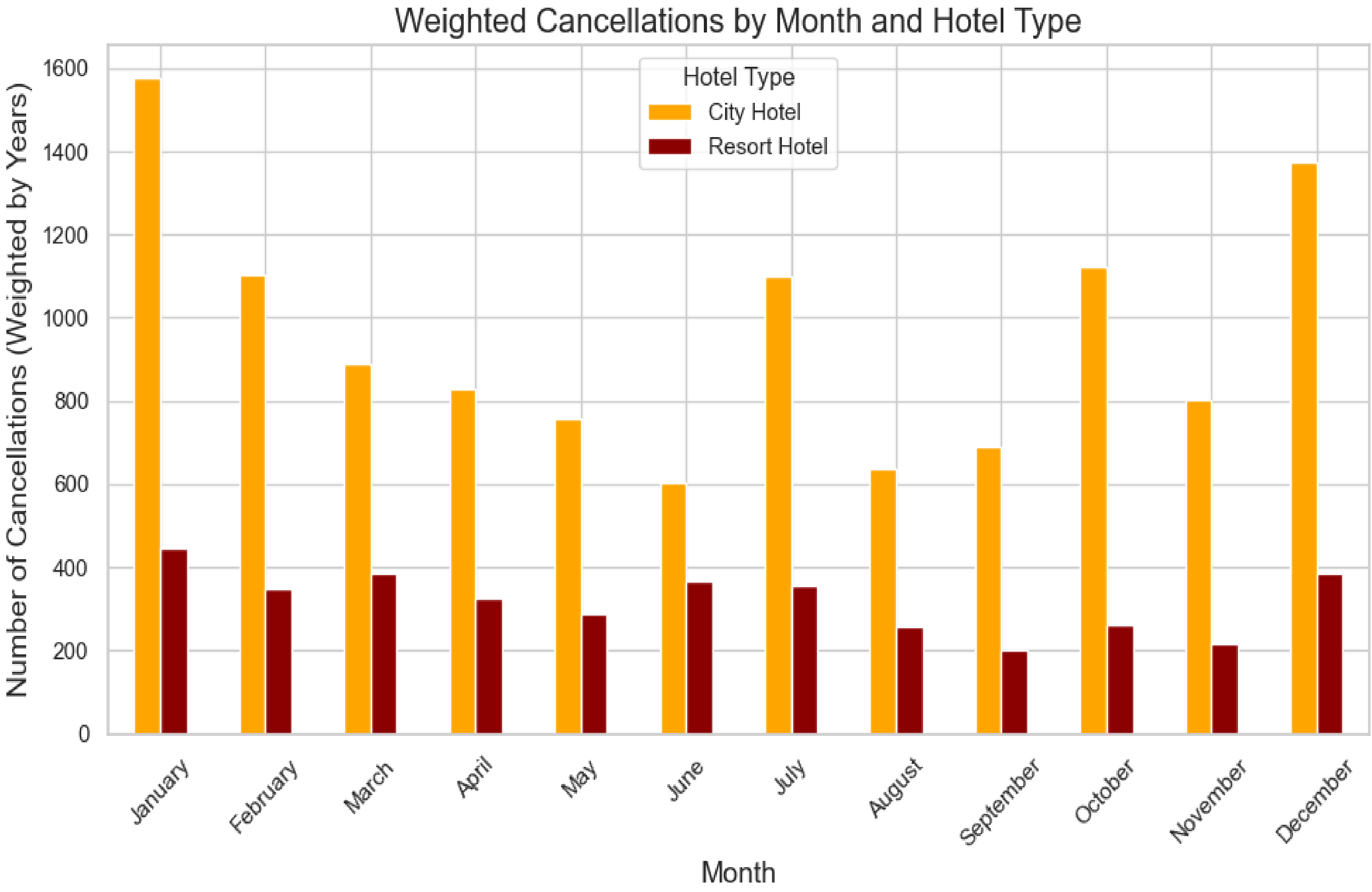
EDA – Cancellations & Hotel Type



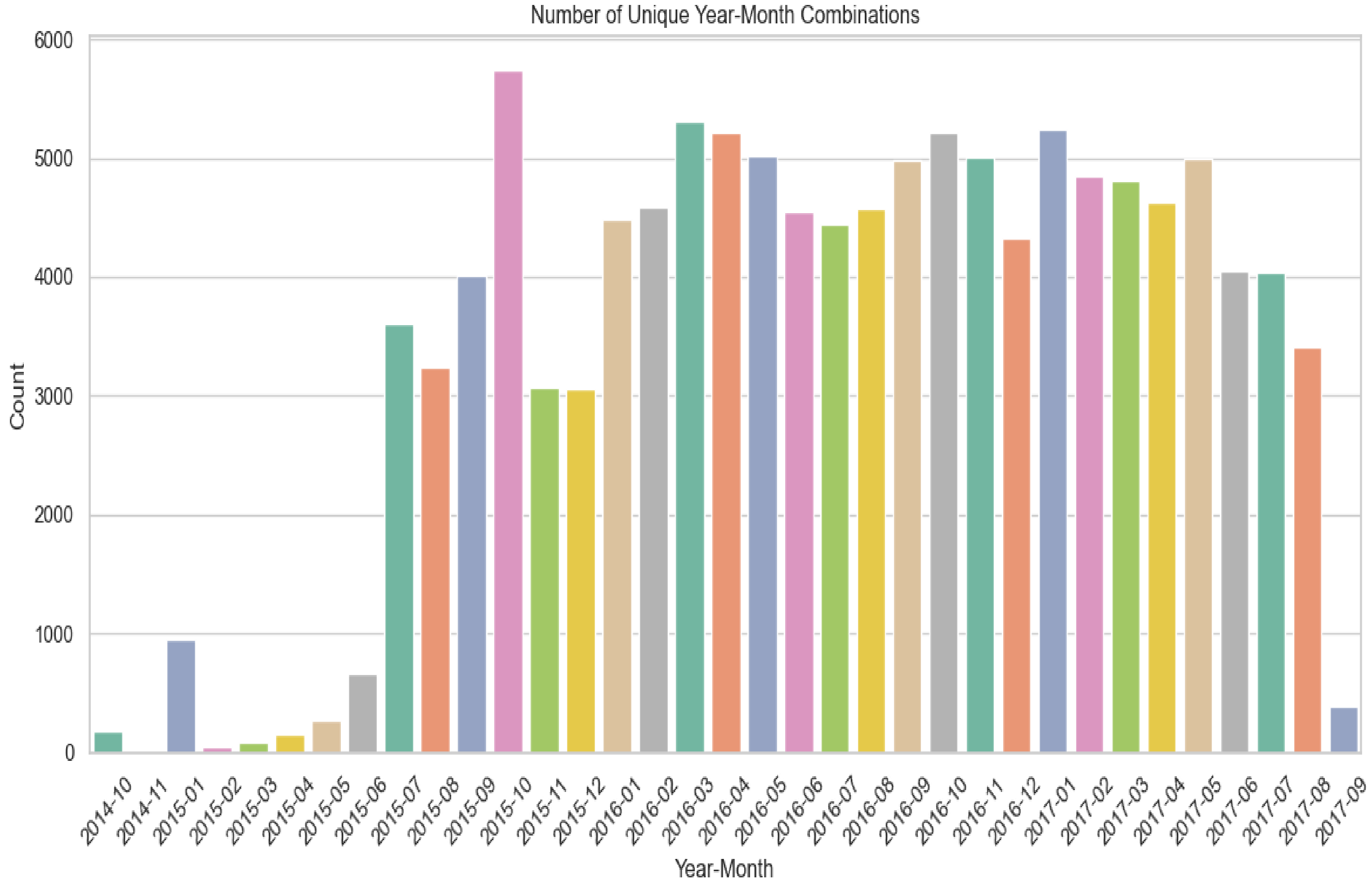
EDA – Cancellations & Hotel Type



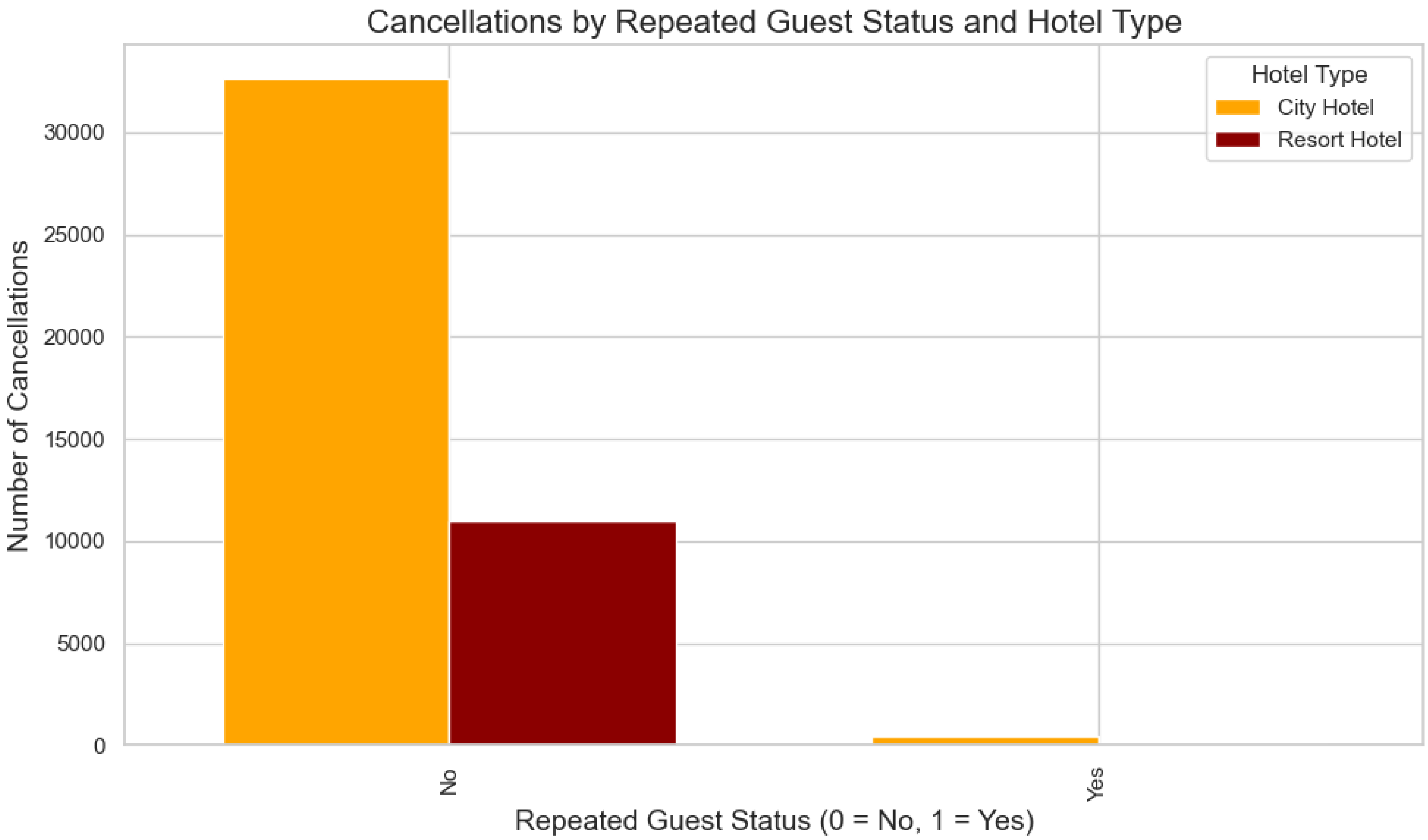
EDA – Cancellations & Hotel Type



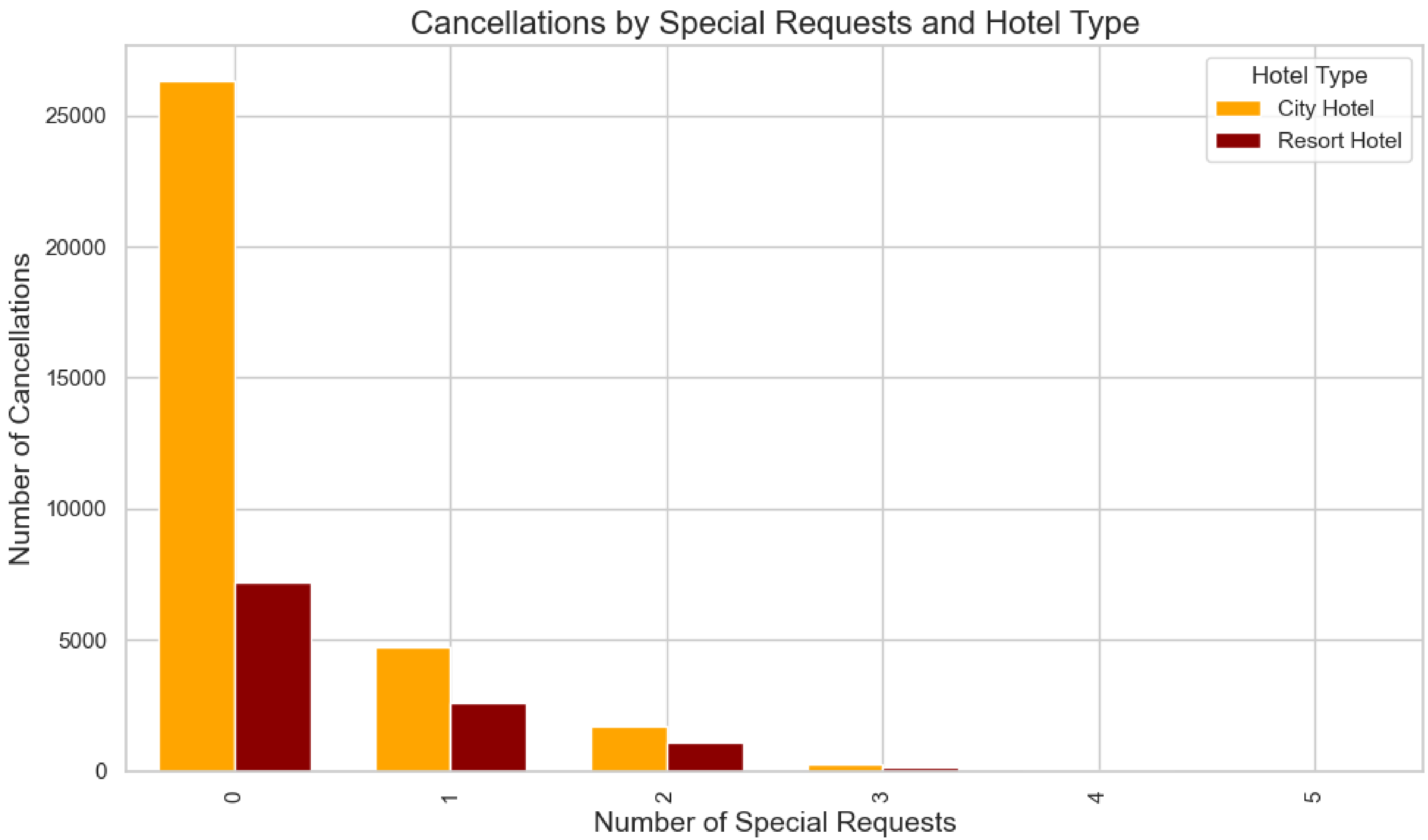
EDA – Unique Year-Month Combinations



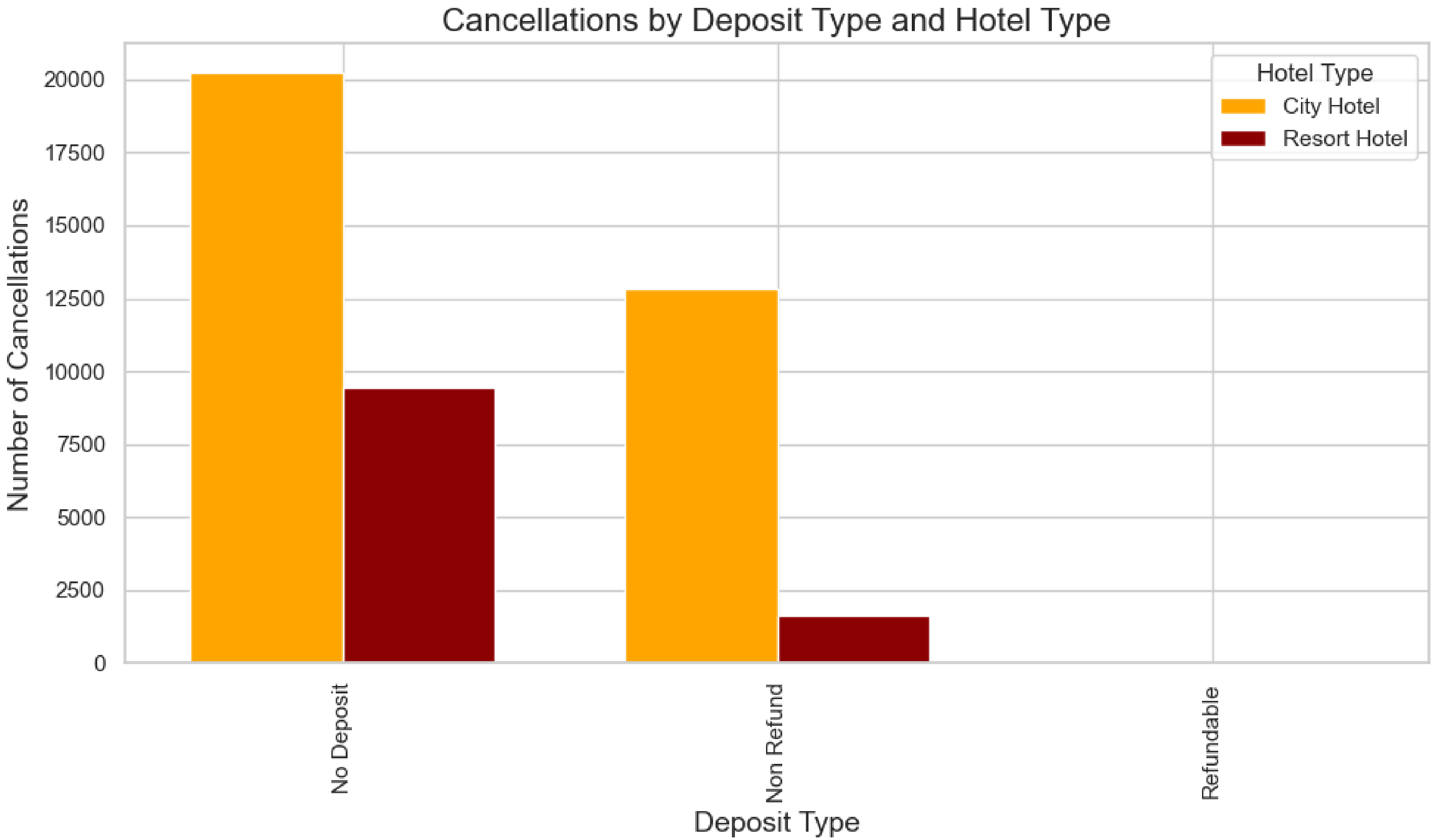
EDA – Cancellations by …



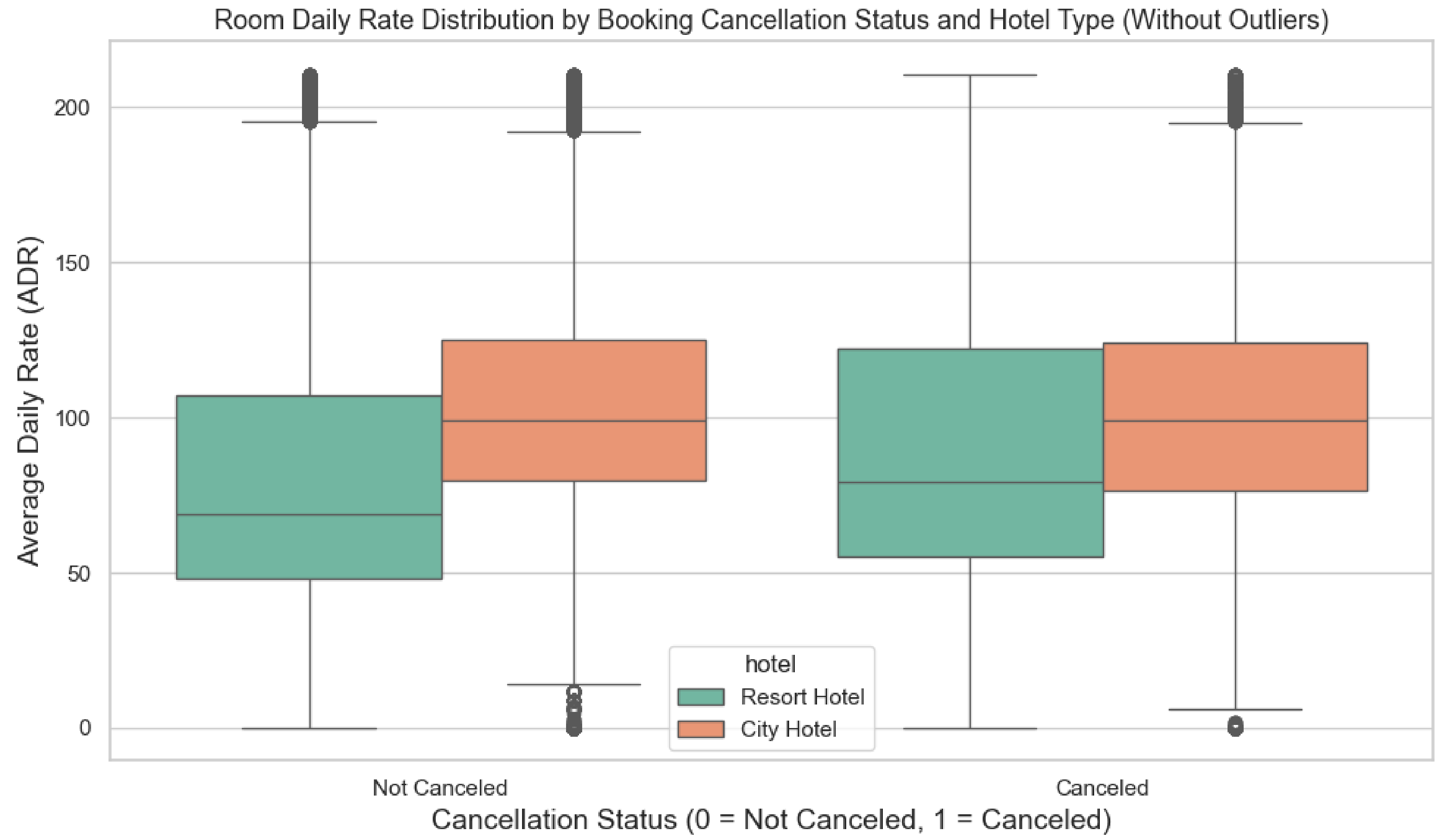
EDA – Cancellations by …



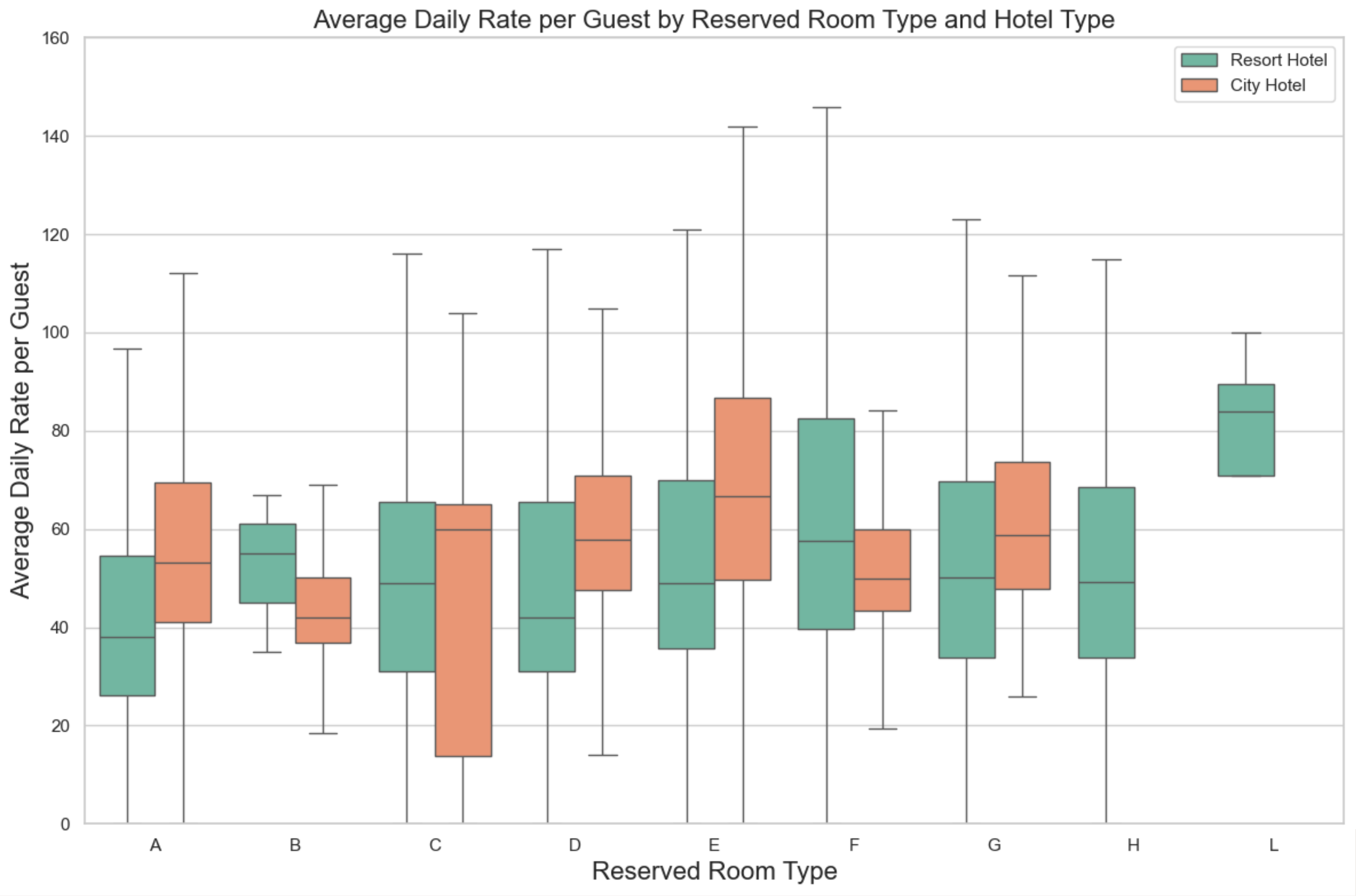
EDA – Cancellations by …



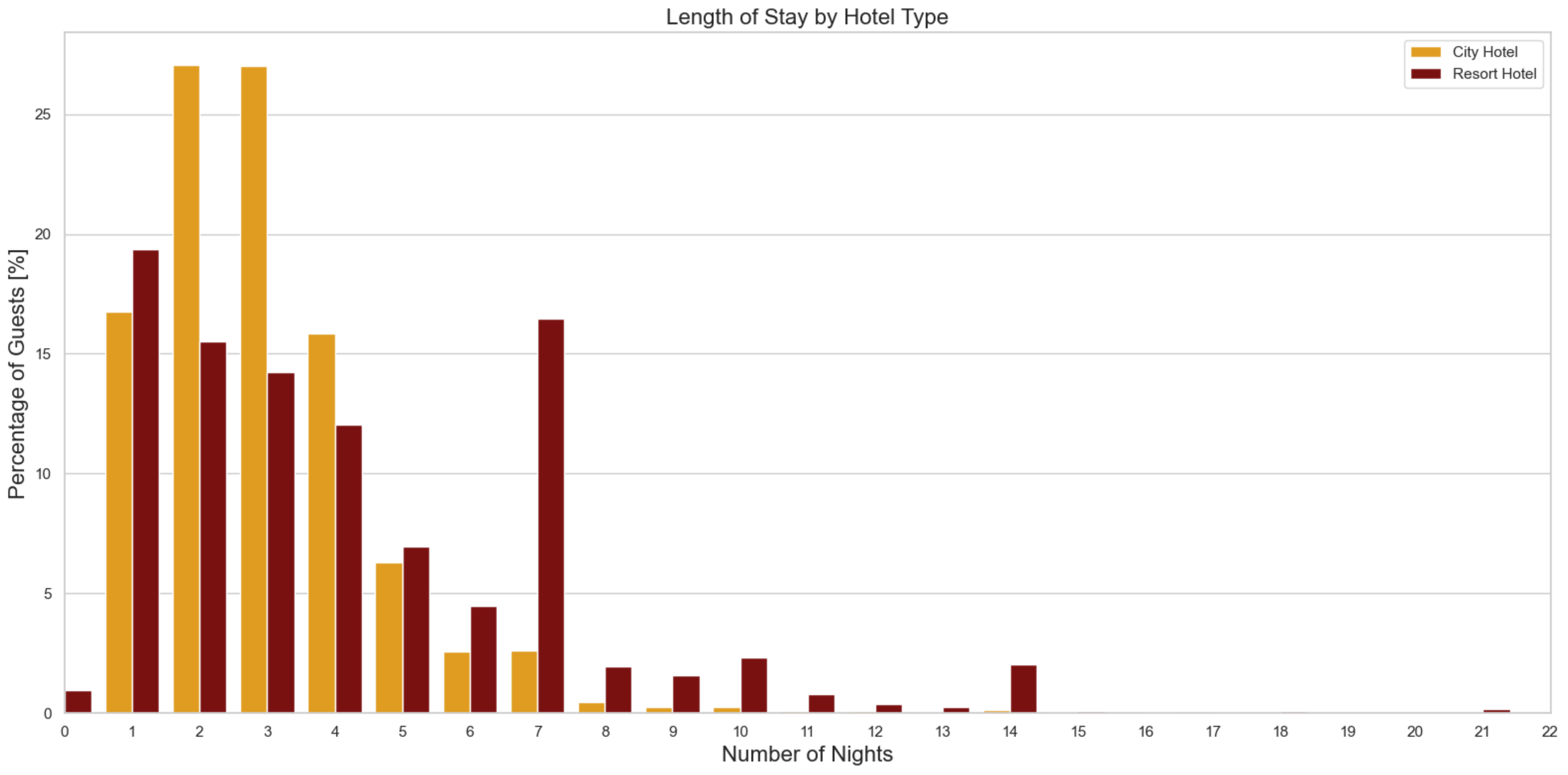
EDA – Cancellations by …



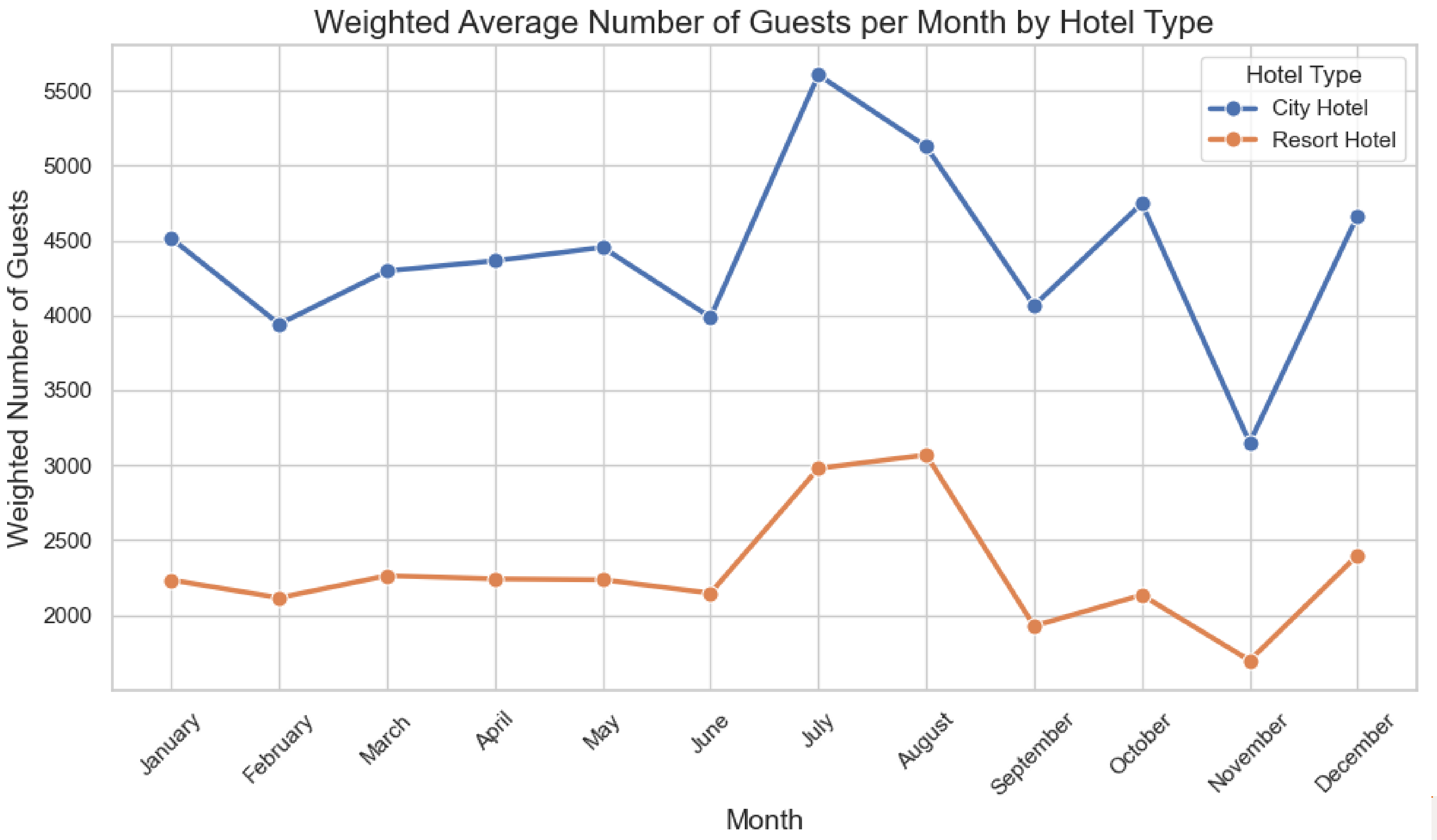
EDA – Descriptives



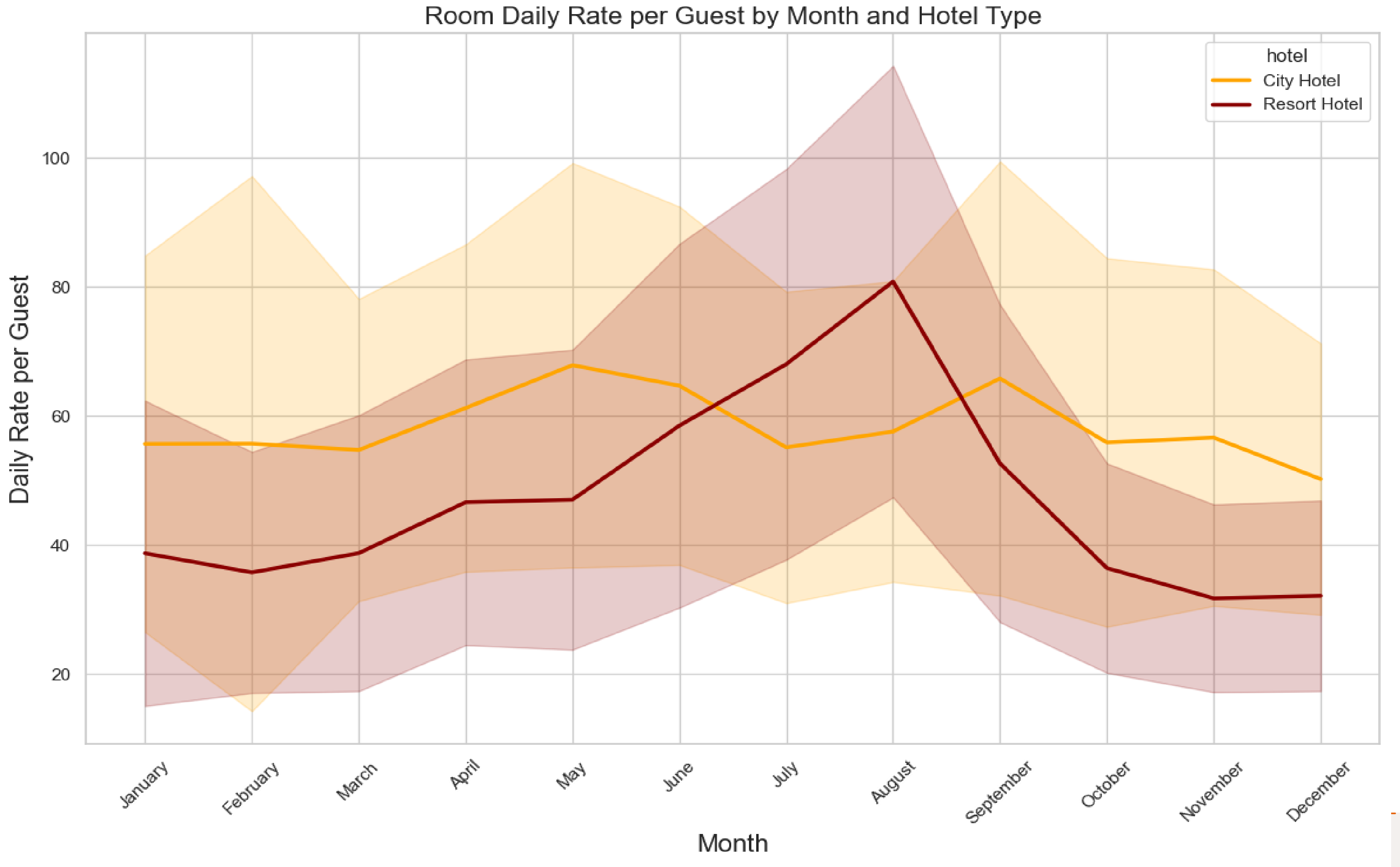
EDA – Descriptives



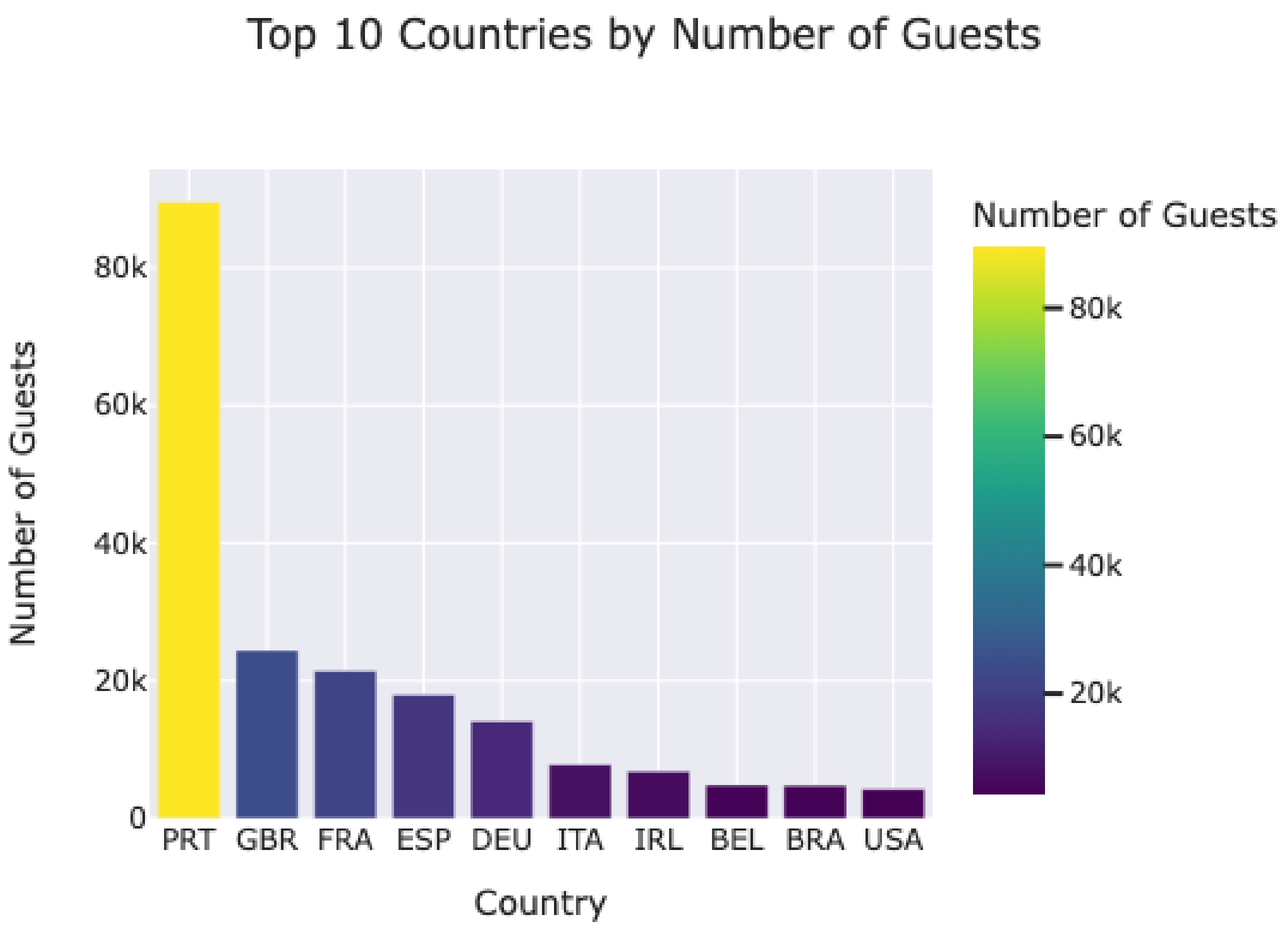
EDA – Descriptives



EDA – Descriptives

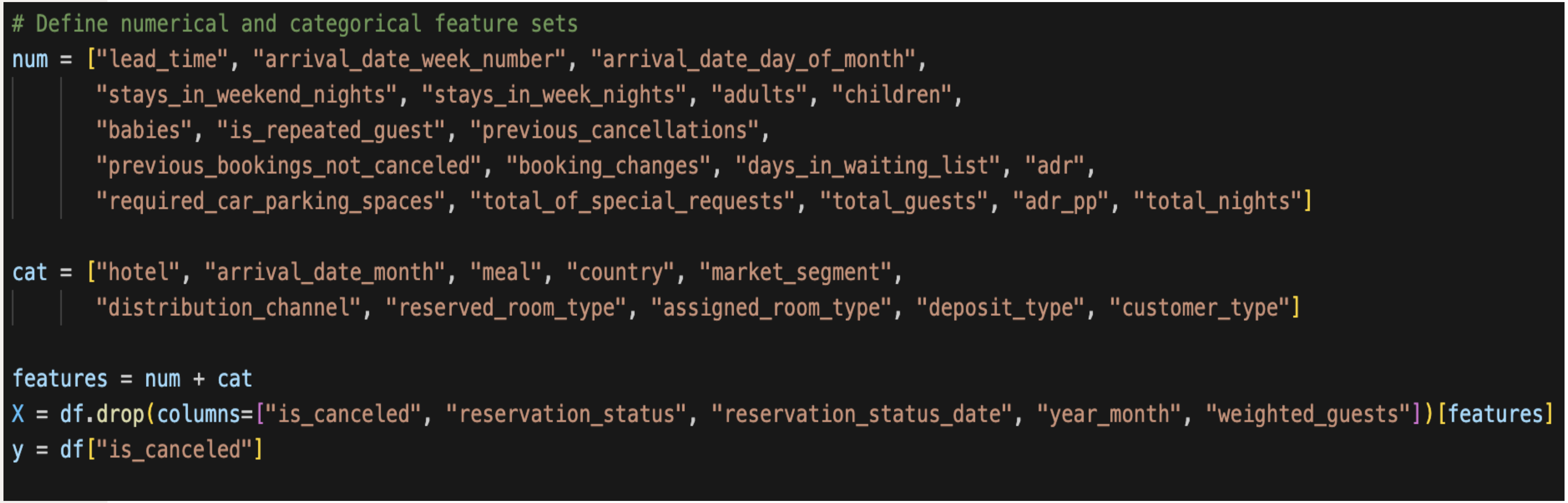


EDA – Descriptives



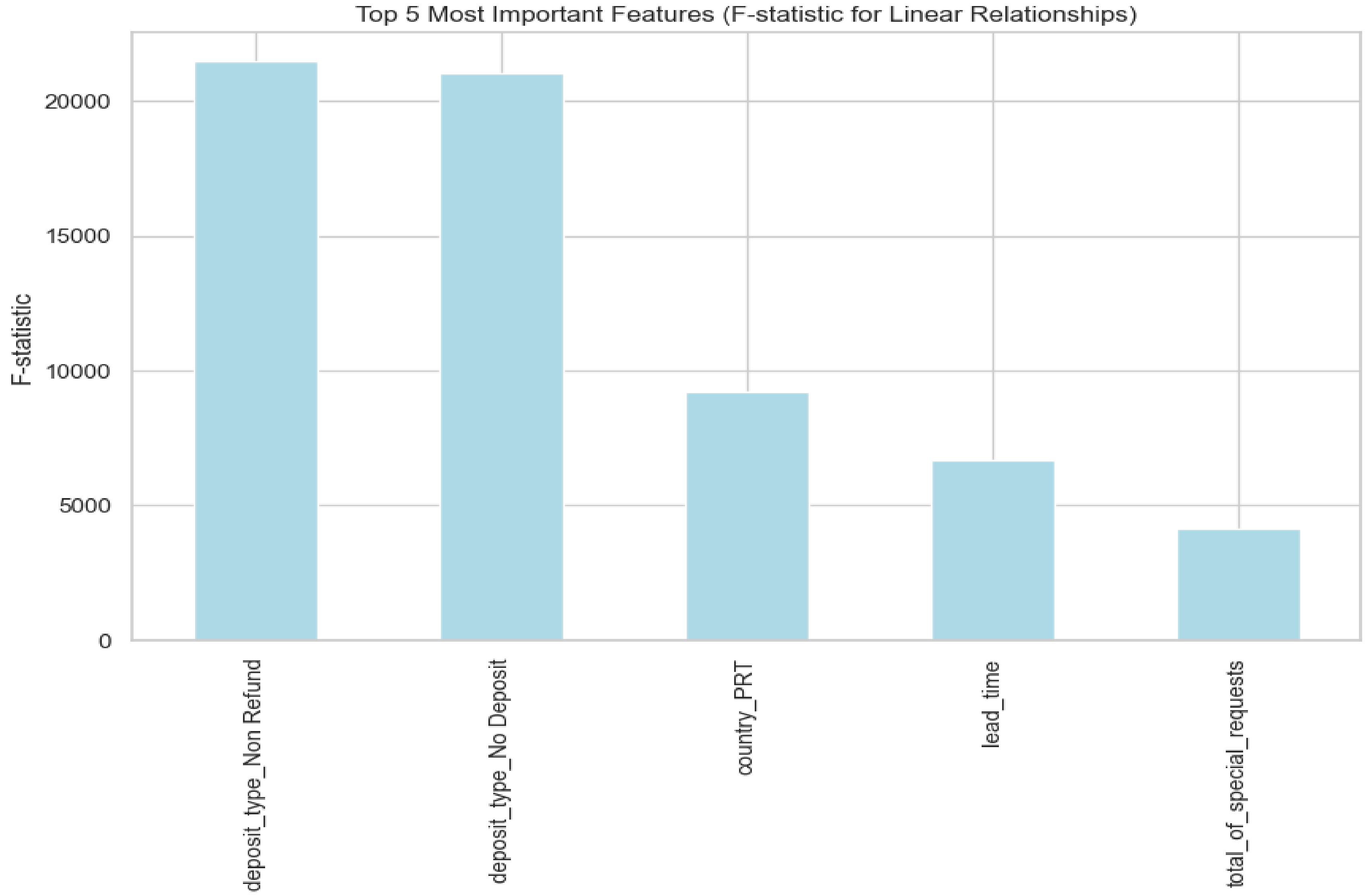
Preprocessing

- Splitting strategy: data split into **60% training, validation** (20%) and **test** (20%); **random state** set to ensure reproducibility
- Preprocessors:
  - **Categorical features**: missing values in *'country'* (41%) imputed with **'Unknown'** using SimpleImputer; one-hot encoding applied to categorical features
  - **Numerical features**: standardized using StandardScaler to ensure features are on the same scale
- Features & data points: original training set shape – (71523, 29) – 10 categorical 19 numerical; preprocessed training set shape – (71523, 240)
- Missing values
  - Country: 41%, imputed & one-hot encoded
  - Children: 0.3%, number of children, dropped (only 4 rows)
  - Company: 94.3%, ID of booking company, dropped this column because no need for ID
  - Agent: 13.7%, ID of travel agency, dropped this column because no need for ID

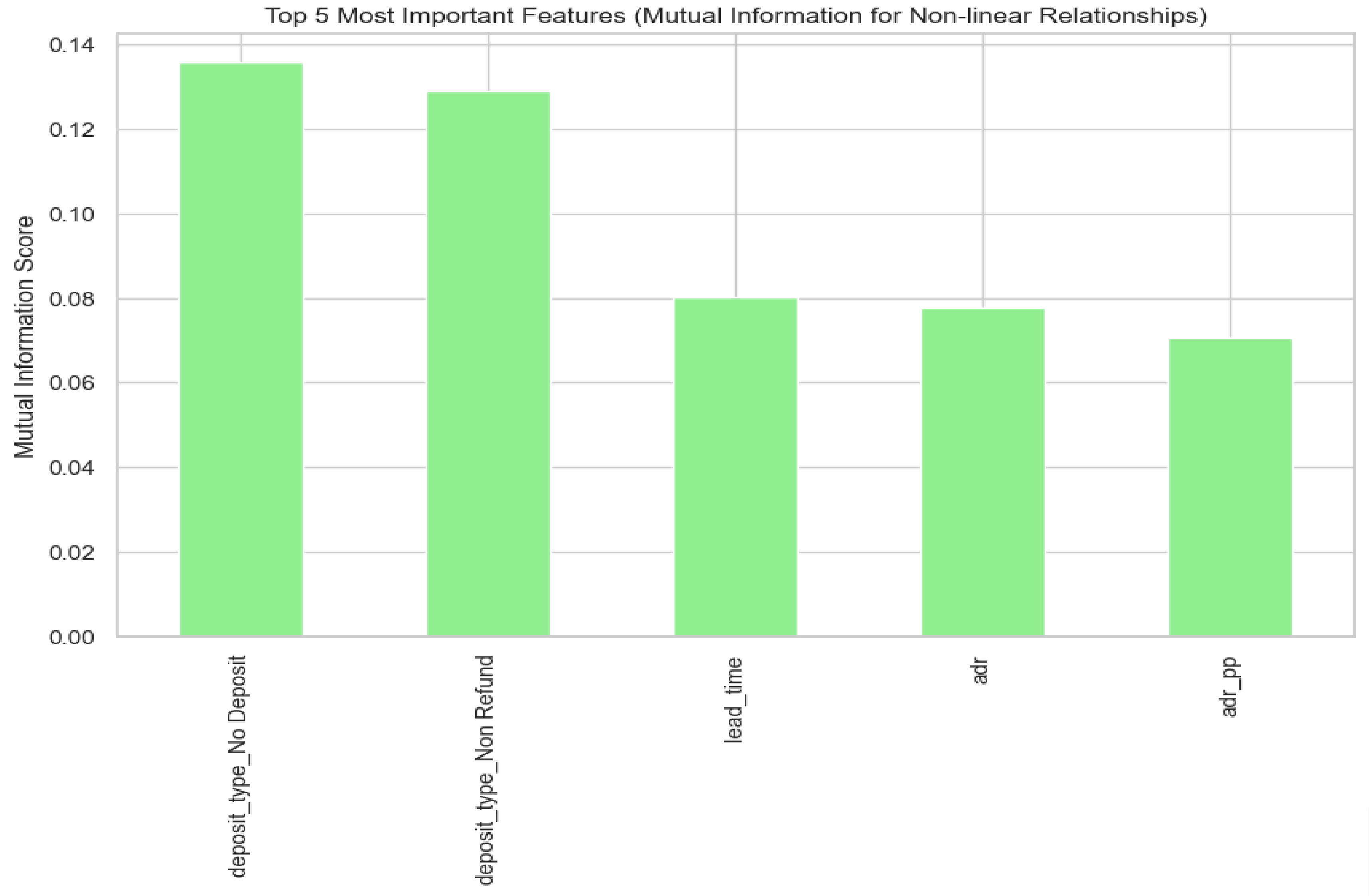


```python
# Define numerical and categorical feature sets
num = ["lead_time", "arrival_date_week_number", "arrival_date_day_of_month",
       "stays_in_weekend_nights", "stays_in_week_nights", "adults", "children",
       "babies", "is_repeated_guest", "previous_cancellations",
       "previous_bookings_not_canceled", "booking_changes", "days_in_waiting_list", "adr",
       "required_car_parking_spaces", "total_of_special_requests", "total_guests", "adr_pp", "total_nights"]

cat = ["hotel", "arrival_date_month", "meal", "country", "market_segment",
       "distribution_channel", "reserved_room_type", "assigned_room_type", "deposit_type", "customer_type"]

features = num + cat
X = df.drop(columns=["is_canceled", "reservation_status", "reservation_status_date", "year_month", "weighted_guests"])[features]
y = df["is_canceled"]
```

Top 5 Features



Top 5 Features


Top 5 Most Important Features (Mutual Information for Non-linear Relationships)
Mutual Information Score
deposit_type_No Deposit
deposit_type_Non Refund
lead_time
adr
adr_pp