

Abstract geometric lines in black on a white background, forming various overlapping polygons and shapes, primarily concentrated in the upper left quadrant.

# **EXPLORATORY DATA ANALYSIS FOR MACHINE LEARNING**

*IBM MACHINE LEARNING PROJECT  
SYED ALI ASGHAR  
AUGUST 2024*

# CONTENTS:

- Goal.
- Brief description of the dataset.
- Attributes summary.
- Data overview.
- Initial plan for data exploration.
- Data cleaning.
- Data visualization.
  - i. No. of guests with respect to countries.
  - ii. Total no. of guests per month.
  - iii. Frequency of Total Stay Duration.
  - iv. Percentage of booking cancellation.
  - v. Cancellation per month.
  - vi. Cancellation rate over time.
  - vii. Lead time effect on cancellation rate.
- Hypothesis testing.
- Suggestions for next steps in analyzing this data



# GOAL:

## **Explore The Attributes Effecting Cancellations Of The Bookings.**

The goal of this exploratory data analysis (EDA) report is to thoroughly investigate the attributes that influence booking cancellations. By delving into various features, such as lead time, market segments, and seasonal patterns, we aim to identify significant relationships and trends that affect cancellation rates. This analysis seeks to uncover how different attributes interact with booking cancellations, providing insights that can be used to optimize hotel management strategies and improve predictive models. By visualizing and analyzing these attributes, we aim to develop a comprehensive understanding of the factors driving cancellations, ultimately guiding more informed decision-making and enhancing overall operational efficiency.

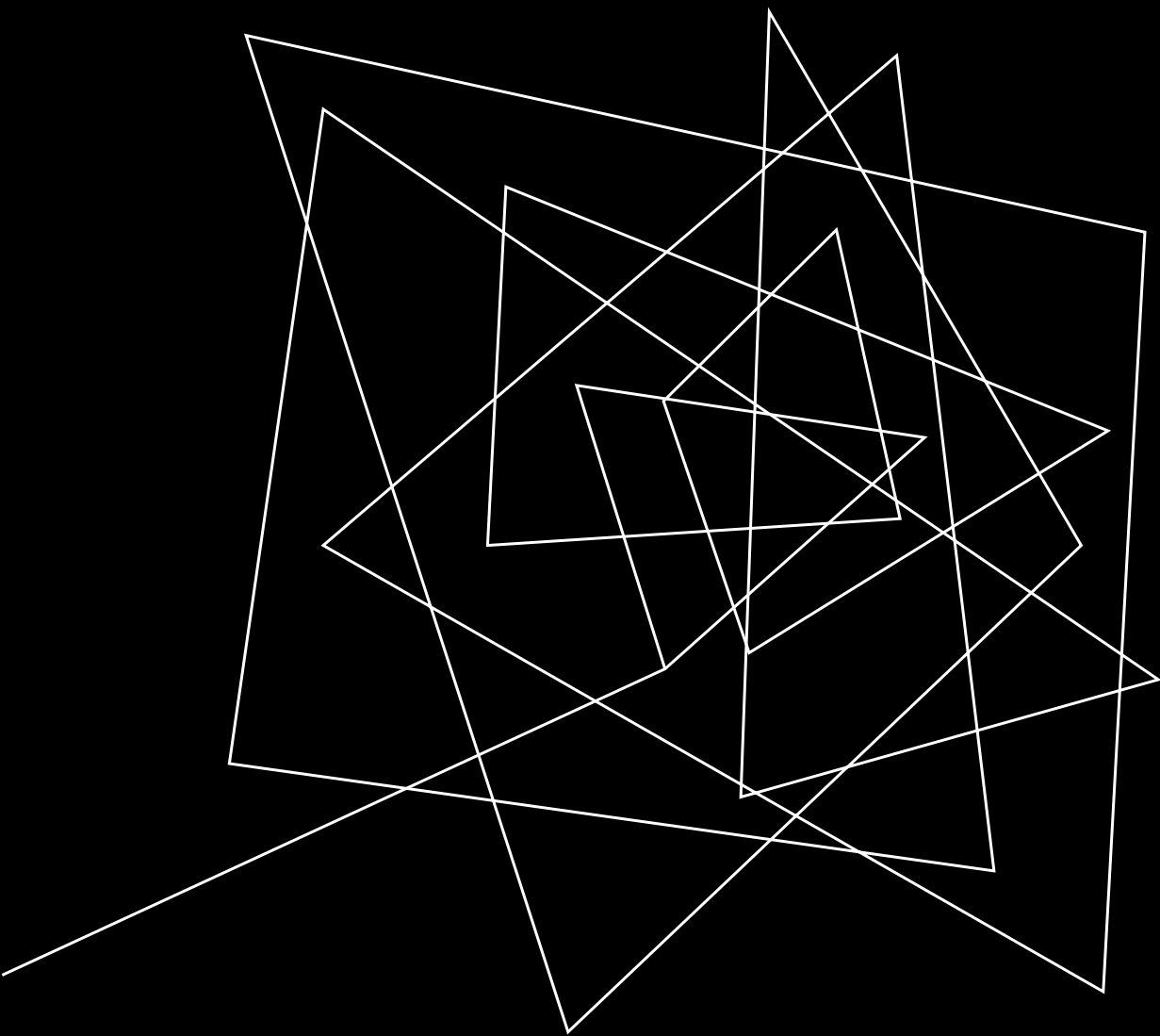
# BRIEF DESCRIPTION OF THE DATASET:

## Content

- This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.
- All personally identifying information has been removed from the data.
- The data have 32 columns(features) and 119390 rows.
- The data contains "bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017".

## Acknowledgment:

- The data is originally from the article [Hotel Booking Demand Datasets](#), written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.




# ATTRIBUTES SUMMARY




## Describing the ambiguous columns, as documented on Kaggle:

### CATEGORICAL FEATURES:

- hotel : There are only two hotel types: Resort Hotel or City Hotel.
- meal: Type of meal booked. Categories are presented in standard hospitality meal packages:
  - Undefined/SC – no meal package
  - BB – Bed & Breakfast
  - HB – Half board (breakfast and one other meal – usually dinner)
  - FB – Full board (breakfast, lunch and dinner)
- country: Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- market\_segment: Market segment designation.
  - “TA” means “Travel Agents”
  - “TO” means “Tour Operators”
- distribution\_channel: Booking distribution channel.
  - “TA” means “Travel Agents”
  - “TO” means “Tour Operators”
- reserved\_room\_type: Code of room type reserved.


- 
- assigned\_room\_type: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request.
  - deposit\_type: Indication on if the customer made a deposit to guarantee the booking.
    - This variable can assume three categories:
    - No Deposit – no deposit was made
    - Non Refund – a deposit was made in the value of the total stay cost
    - Refundable – a deposit was made with a value under the total cost of stay
  - customer\_type: Type of booking, assuming one of four categories:
    - Contract – when the booking has an allotment or other type of contract associated to it
    - Group – when the booking is associated to a group
    - Transient – when the booking is not part of a group or contract, and is not associated to other Transient parties
    - Transient Party – when the booking is transient, but is associated to at least other transient booking.


- 
- reservation\_status: Reservation last status, assuming one of three categories:
    - 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 inform the hotel of the reason why.
  - reservation\_status\_date: Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

#### **NUMERIC FEATURES:**

- agent: ID of the travel agency that made the booking
- company: ID of the company/entity that made the booking or responsible for paying the booking.
- days\_in\_waiting\_list: Number of days the booking was in the waiting list before it was confirmed to the customer.
- is\_canceled: "Value indicating if the booking was canceled (1) or not (0)
- lead\_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date



- 
- stays\_in\_weekend\_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
  - stays\_in\_week\_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
  - is\_repeated\_guest: Value indicating if the booking name was from a repeated guest (1) or not (0)
  - previous\_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking
  - previous\_bookings\_not\_canceled: Number of previous bookings not cancelled by the customer prior to the current booking
  - booking\_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.
  - adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
  - required\_car\_parking\_spaces: Number of car parking spaces required by the customer

- 
- total\_of\_special\_requests: Number of special requests made by the customer (e.g. twin bed or high floor)
  - reservation\_status: Reservation last status, assuming one of three categories:
    - 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 inform the hotel of the reason why
  - reservation\_status\_date: Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

# DATA OVERVIEW:

Data columns (total 32 columns):

#	Column	Non-Null Count	Dtype
0	hotel	119390 non-null	object
1	is_canceled	119390 non-null	int64
2	lead_time	119390 non-null	int64
3	arrival_date_year	119390 non-null	int64
4	arrival_date_month	119390 non-null	object
5	arrival_date_week_number	119390 non-null	int64
6	arrival_date_day_of_month	119390 non-null	int64
7	stays_in_weekend_nights	119390 non-null	int64
8	stays_in_week_nights	119390 non-null	int64
9	adults	119390 non-null	int64
10	children	119386 non-null	float64
11	babies	119390 non-null	int64
12	meal	119390 non-null	object
13	country	118902 non-null	object
14	market_segment	119390 non-null	object
15	distribution_channel	119390 non-null	object
16	is_repeated_guest	119390 non-null	int64
17	previous_cancellations	119390 non-null	int64
18	previous_bookings_not_canceled	119390 non-null	int64
19	reserved_room_type	119390 non-null	object
20	assigned_room_type	119390 non-null	object
21	booking_changes	119390 non-null	int64
22	deposit_type	119390 non-null	object
23	agent	103050 non-null	float64
24	company	6797 non-null	float64
25	days_in_waiting_list	119390 non-null	int64
26	customer_type	119390 non-null	object
27	adr	119390 non-null	float64
28	required_car_parking_spaces	119390 non-null	int64
29	total_of_special_requests	119390 non-null	int64
30	reservation_status	119390 non-null	object
31	reservation_status_date	119390 non-null	object

dtypes: float64(4), int64(16), object(12)



# INITIAL PLAN FOR DATA EXPLORATION:

## **Data cleaning:**

- We'll take care of the missing values in the features accordingly.
- The irrelevant features and inconsistent data will be handled appropriately.
- Some categorical attributes are also feature engineered and turned into numeric one.

## **Data visualization:**

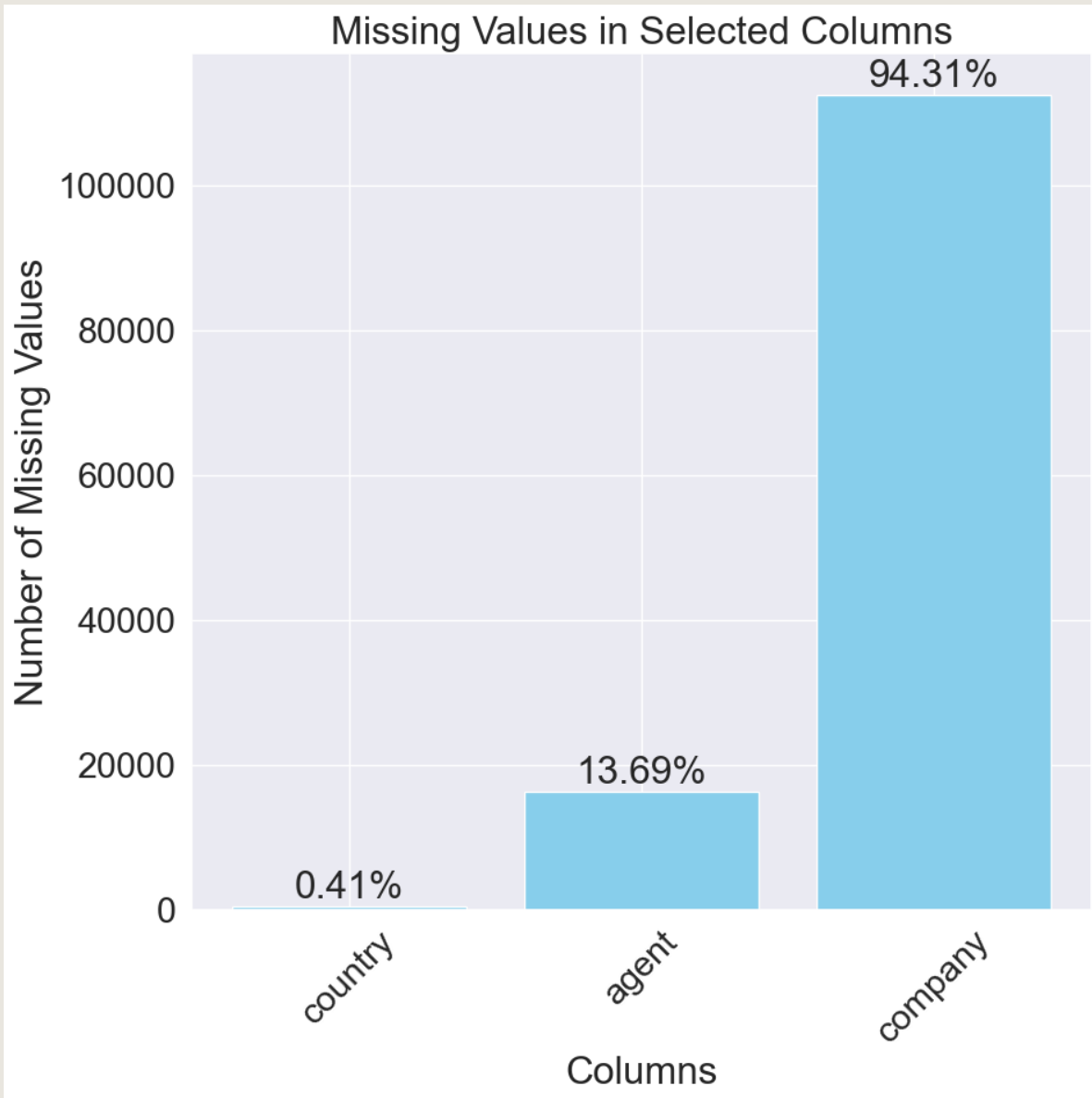
- Matplotlib, folium, plotly and seaborn will be used to plot different features.
- Line graph, histograms, bar graphs and other type of graphs are used to visualize the relation between different attributes of the data.



# DATA CLEANING

*In the upcoming slides, we will address missing, irrelevant, and inconsistent data. Additionally, a statistical summary of both categorical and numerical attributes will be provided.*

# MISSING VALUES:



- The missing values in “Company” attribute were replaced with 0.
- The Nan values in “Country” feature were filled with “unknown” .
- The “Agent” is a ID category hence we substitute the nan values with 0.

# Handling inconsistencies and irrelevant features:

*This section also involves the feature engineering that we have done.*

- In some instances the “stays\_in\_weekend\_nights” and “stays\_in\_week\_nights” are both 0 that cant be by logic so we errase all record where this happens.
- The column “assigned\_room\_type” dosent make sense in this problem because a room only is assigned on the check in day that means the booking was not canceled, so this is not a good predictor.
- There is one negative “adr” value which does not make sense. Replace negative adr with median of adr column.
- From the descriptive statistics we see that the max of children and babies is 10. since we only have 1 instance for tose outliers children and babies, we will remove them from the data.
- The “is\_canceled” column was feature engineered and transformed into two numerical attributes using binary encoding.
- Arrival date, week, month and year was combined into a single feature for easy handling.

# NUMERICAL ATTRIBUTES

## STATISTICAL SUMMARY:

	count	mean	std	min	25%	50%	75%	max
is_canceled	119210.0	0.370766	0.483012	0.0	0.0	0.00	1.0	1.0
lead_time	119210.0	104.109227	106.875450	0.0	18.0	69.00	161.0	737.0
arrival_date_year	119210.0	2016.156472	0.707485	2015.0	2016.0	2016.00	2017.0	2017.0
arrival_date_week_number	119210.0	27.163376	13.601107	1.0	16.0	28.00	38.0	53.0
arrival_date_day_of_month	119210.0	15.798717	8.781070	1.0	8.0	16.00	23.0	31.0
stays_in_weekend_nights	119210.0	0.927053	0.995117	0.0	0.0	1.00	2.0	19.0
stays_in_week_nights	119210.0	2.499195	1.897106	0.0	1.0	2.00	3.0	50.0
adults	119210.0	1.859206	0.575186	0.0	2.0	2.00	2.0	55.0
children	119210.0	0.104043	0.398836	0.0	0.0	0.00	0.0	10.0
babies	119210.0	0.007961	0.097509	0.0	0.0	0.00	0.0	10.0
is_repeated_guest	119210.0	0.031499	0.174663	0.0	0.0	0.00	0.0	1.0
previous_cancellations	119210.0	0.087191	0.844918	0.0	0.0	0.00	0.0	26.0
previous_bookings_not_canceled	119210.0	0.137094	1.498137	0.0	0.0	0.00	0.0	72.0
booking_changes	119210.0	0.218799	0.638504	0.0	0.0	0.00	0.0	18.0
agent	119210.0	74.889078	107.168884	0.0	7.0	9.00	152.0	535.0
company	119210.0	10.735400	53.830143	0.0	0.0	0.00	0.0	543.0
days_in_waiting_list	119210.0	2.321215	17.598002	0.0	0.0	0.00	0.0	391.0
adr	119210.0	101.969942	50.433035	0.0	69.5	94.95	126.0	5400.0
required_car_parking_spaces	119210.0	0.062553	0.245360	0.0	0.0	0.00	0.0	8.0
total_of_special_requests	119210.0	0.571504	0.792876	0.0	0.0	0.00	1.0	5.0



# CATEGORICAL ATTRIBUTES

## SUMMARY:

	count	unique	top	freq
hotel	119210	2	City Hotel	79163
arrival_date_month	119210	12	August	13861
meal	119210	4	BB	92236
country	119210	178	PRT	48483
market_segment	119210	8	Online TA	56408
distribution_channel	119210	5	TA/TO	97750
reserved_room_type	119210	9	A	85873
assigned_room_type	119210	11	A	74020
deposit_type	119210	3	No Deposit	104461
customer_type	119210	4	Transient	89476
reservation_status	119210	3	Check-Out	75011
reservation_status_date	119210	926	2015-10-21	1460



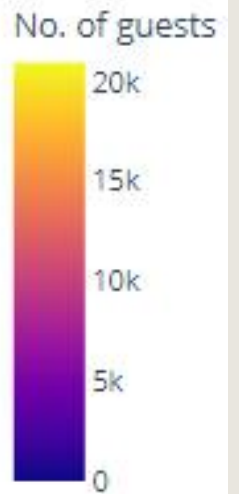
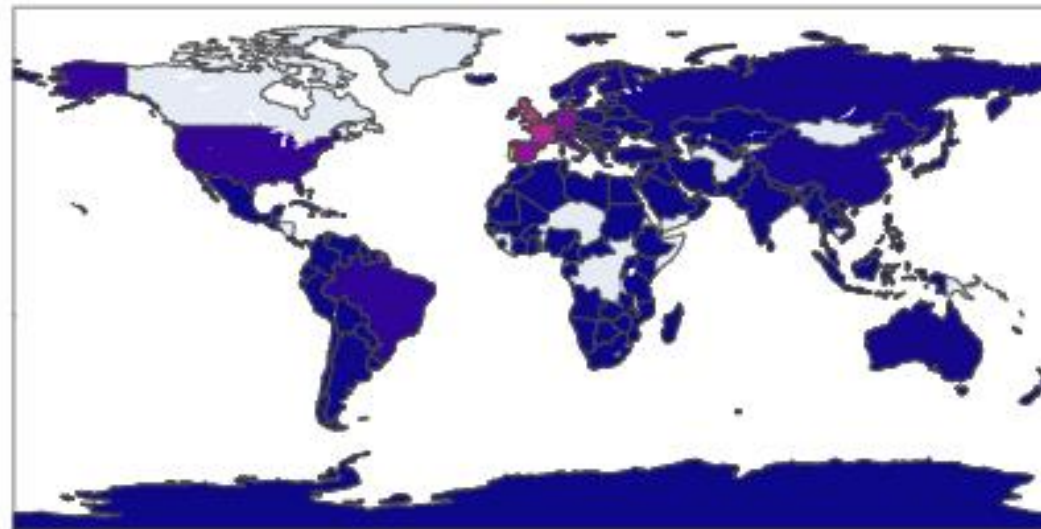
# DATA VISUALIZATION

*The following slides will explore various data attributes through visualizations, with a primary focus on the cancellation rate and the factors that influence it.*

## NO. OF GUESTS WITH RESPECT TO THE COUNTRIES:

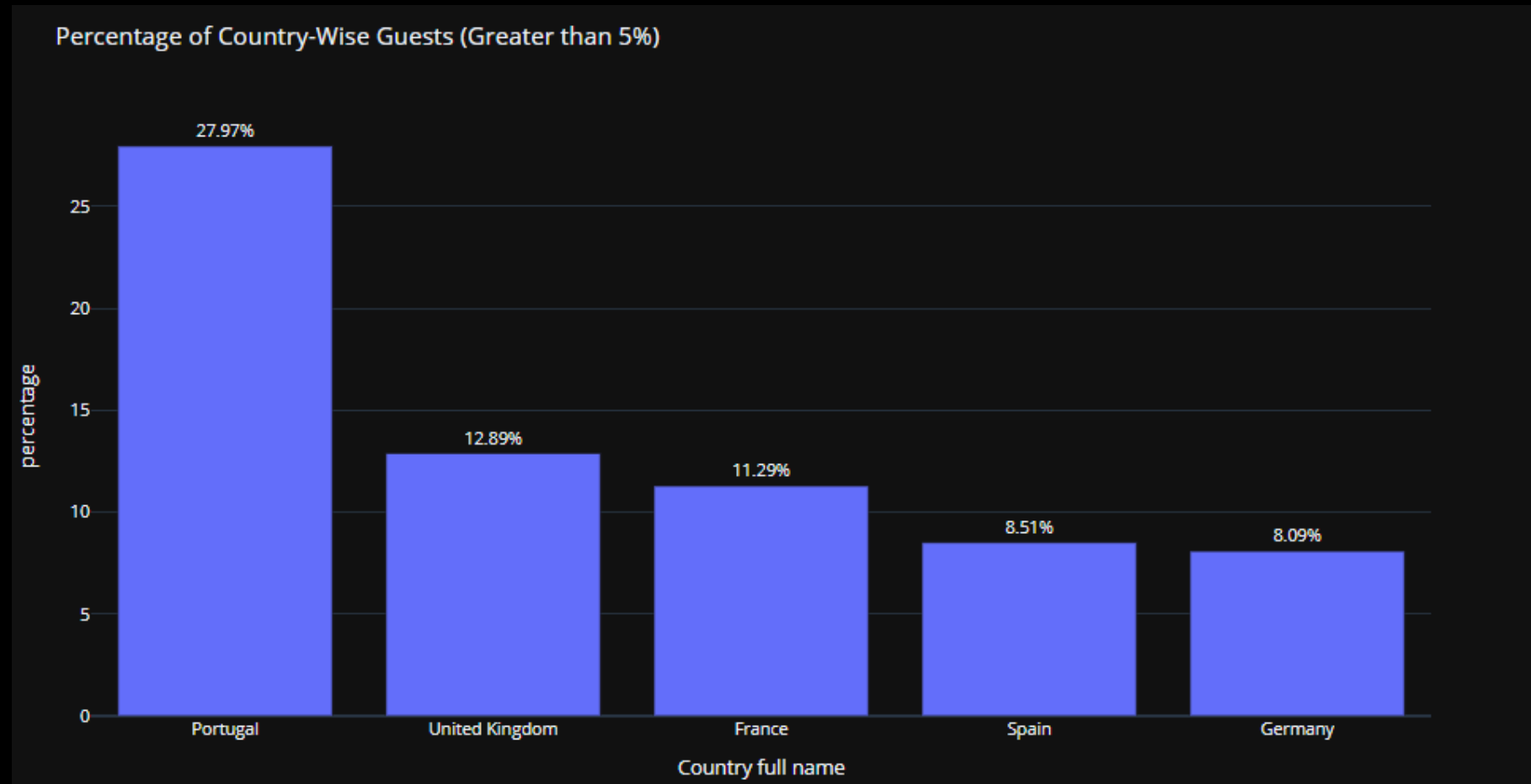
Folium, Plotly, and Matplotlib were used to visualize the origins of hotel guests. The majority of visitors are from Portugal, with over 20,000 guests, followed by the United Kingdom with over 9,000 visitors. Most guests come from Europe.

Below is a world map that displays the countries, color-coded according to the number of visitors.



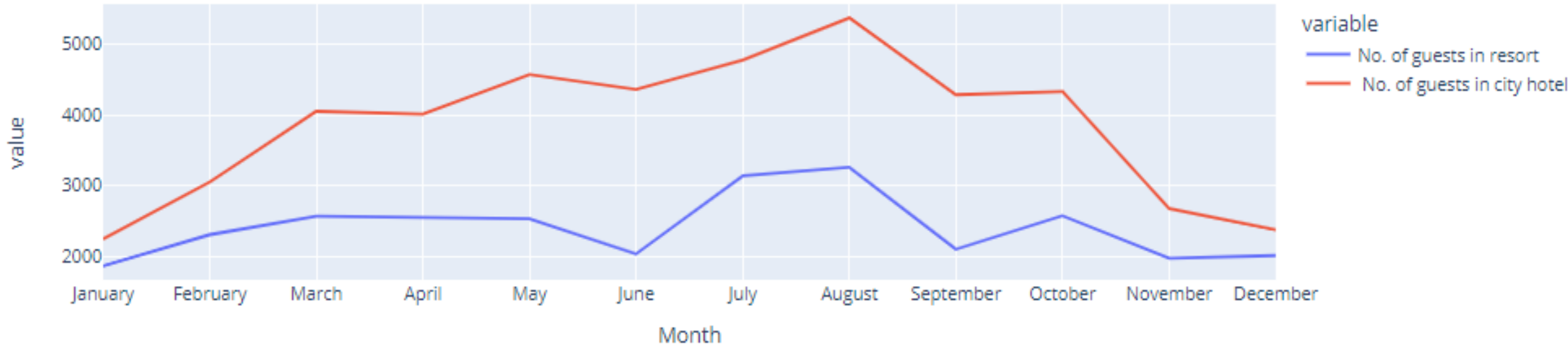
The graph below displays the exact percentage of visitors from countries where the percentage of visitors exceeds 5%.

The country names were initially in acronyms, so we utilized the 'pycountry' library to convert these labels to full names for better readability.



# TOTAL NO. OF GUESTS PER MONTH:

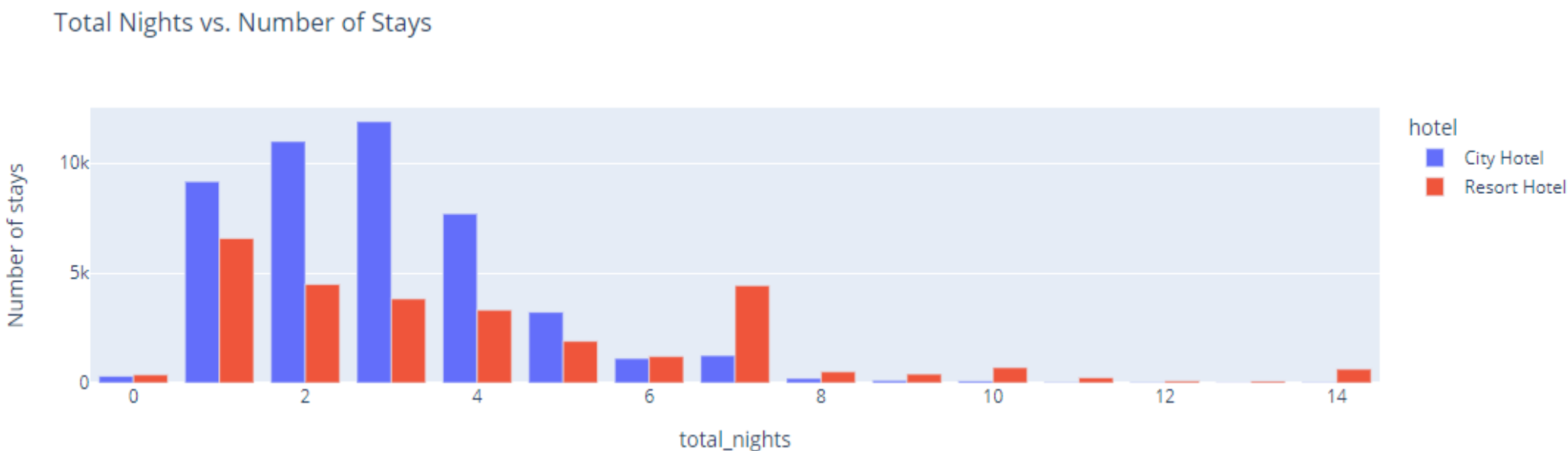
Total guests per month



The above graph shows that the most guests visits during spring and autumn. And least number of guests were recorded in winters around January.

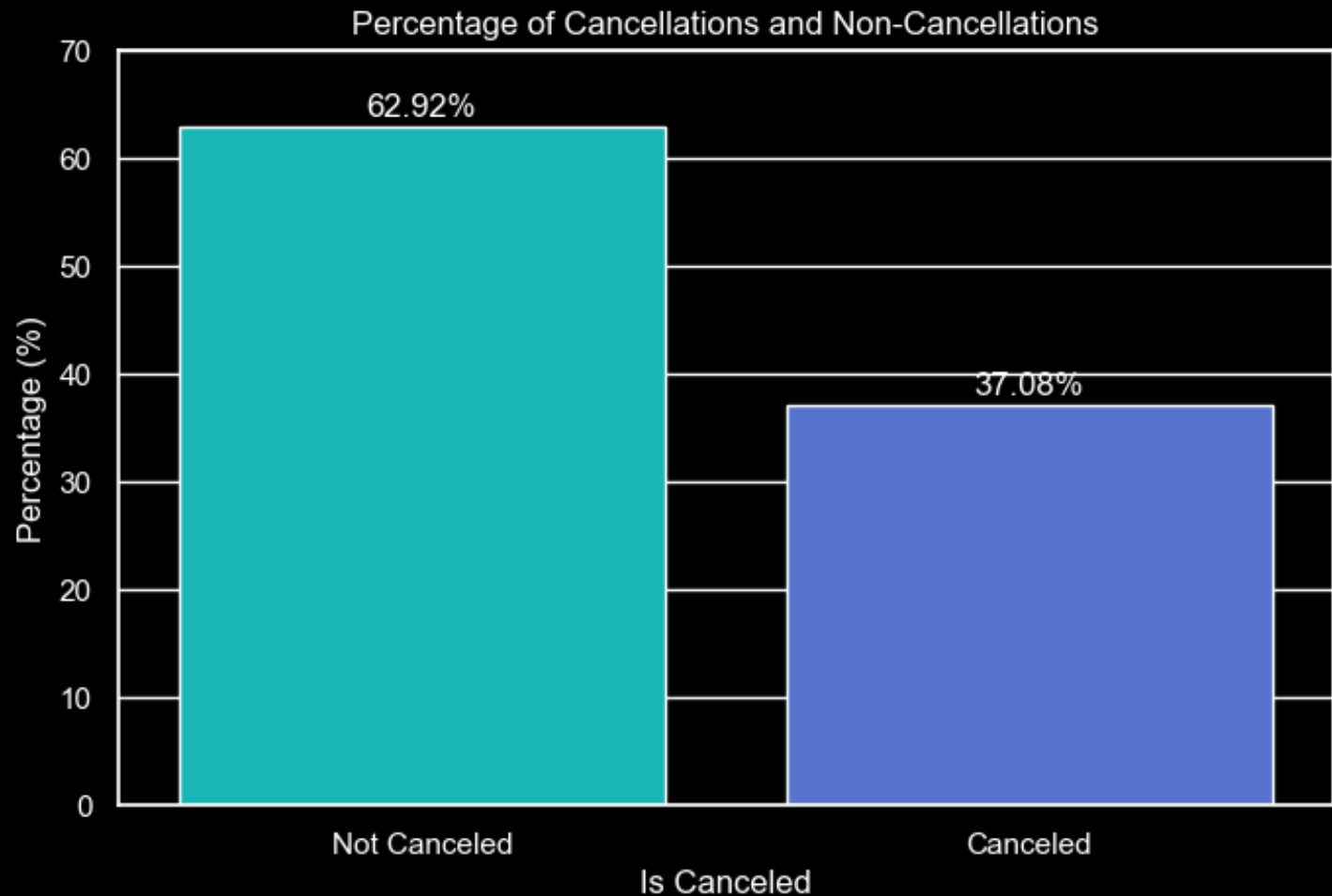
# FREQUENCY OF TOTAL STAY DURATION:

The graph illustrates the frequency of different stay durations (total nights) for both City Hotel and Resort Hotel. The data shows that most guests stayed between 1 to 4 nights, with the highest number of stays occurring at 3 nights for the City Hotel and 2 nights for the Resort Hotel. Stays longer than 7 nights are less frequent, with a notable decrease in the number of stays as the duration increases. The data suggests that shorter stays are more common, particularly for City Hotel guests.



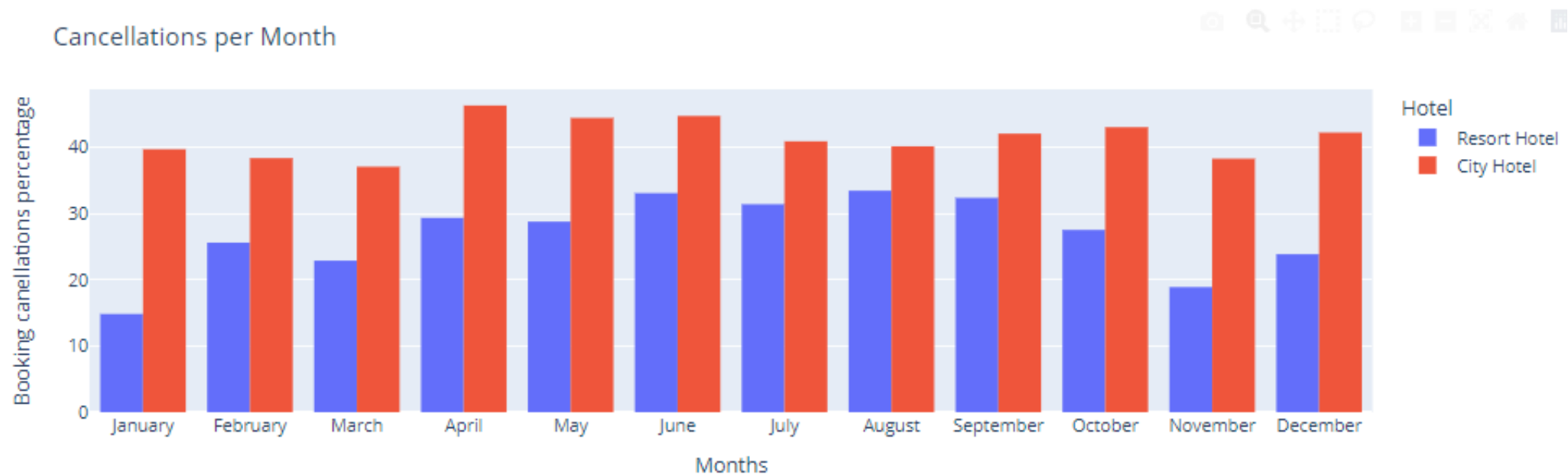
# PERCENTAGE OF BOOKING CANCELLATION:

The data indicates that approximately 37.08% of bookings were canceled, while 62.92% of bookings were completed. This significant cancellation rate highlights the importance for hotels to develop strategies for predicting and managing cancellations to optimize occupancy rates and resource planning.



# CANCELLATIONS PER MONTH:

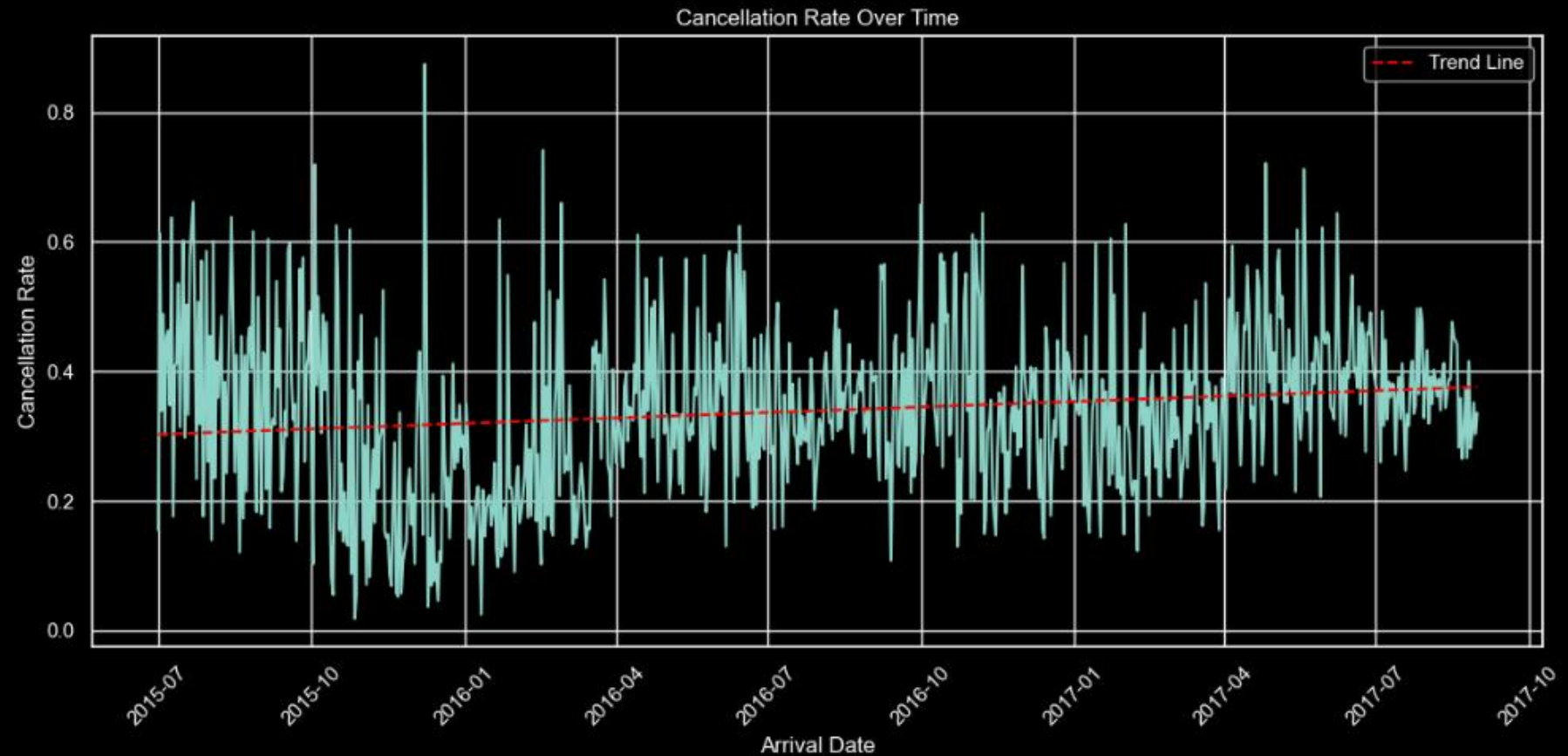
The graph below of this data will represent the monthly trend of hotel bookings and cancellations for both Resort Hotels and City Hotels. It will show the total number of bookings each month, alongside the percentage of those bookings that were canceled. By comparing the data across months and between the two types of hotels, the graph can reveal patterns such as peak cancellation periods, differences in cancellation rates between Resort Hotels and City Hotels, and how seasonality impacts booking behaviors.





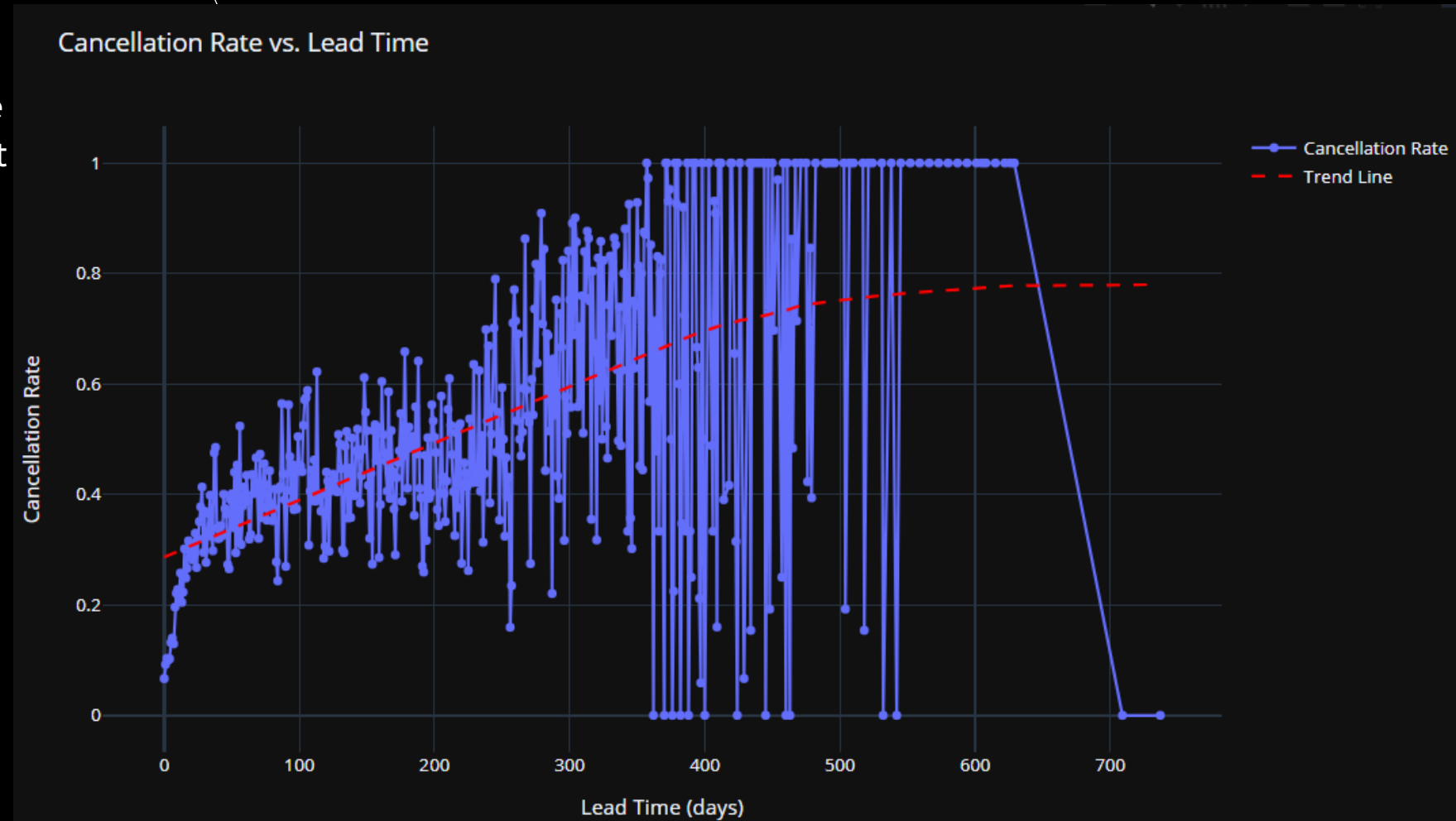
# CANCELLATION RATE OVER TIME:

The graph shows hotel booking cancellation rates from mid-2015 to late 2017, with fluctuations over time. The trend line indicates a stable cancellation rate with a slight upward trend.



# LEAD TIME EFFECT ON CANCELLATION RATE:

The graph indicates that the cancellation rate increases as the lead time (in days) extends, but it stabilizes once the lead time surpasses 500 days. As the duration between booking and guest arrival grows longer, the likelihood of cancellations also increases.



The background of the slide features several thin, dark gray lines that intersect at various points, creating a complex geometric pattern. These lines are positioned primarily on the left side of the slide, with some extending towards the center.

# HYPOTHESIS TESTING

*In the upcoming slides, we will formulate three hypothesis regarding the dependence of different attributes on cancellation rate. Then we'll test those hypothesis and give conclusion.*

# HYPOTHESIS NO.1:

The likelihood of booking cancellations varies by the months.

**Null Hypothesis (H0):** The cancellation rate does not significantly differ between months.

**Alternative Hypothesis (Ha):** There is a significant difference in the cancellation rate between at least some months.

## RESULTS:

• **Months with Highly Significant Results:** The months with very low p-values indicate strong evidence that the cancellation rates for these months are significantly different from the average cancellation rate of the other months. Specifically, January, March, November, June, April, February, May, September, and December all show strong statistical significance.

• **Non-Significant Result:** July has a p-value greater than 0.05, indicating that the cancellation rate for July is not significantly different from the average cancellation rate of the other months.

**Therefore, we do not reject the hypothesis that the cancellation rate varies by month,** but we note that July does not show a significant difference compared to other months.

Cancellation Rate by Month Test Results:		
	t-value	p-value
Month		
January	-11.242033	4.641163e-29
March	-10.669682	1.866189e-26
November	-10.508271	1.180510e-25
June	9.836049	9.437055e-23
April	8.350548	7.459167e-17
February	-7.139893	1.003932e-12
May	6.145385	8.187140e-10
September	4.651028	3.336525e-06
December	-3.619827	2.967134e-04
October	2.317563	2.048796e-02
August	1.823674	6.821832e-02
July	0.953109	3.405497e-01

# HYPOTHESIS NO.2:

Certain market segments have higher cancellation rates.

**Null Hypothesis (H0):** There is no significant difference in the cancellation rates among different market segments.

**Alternative Hypothesis (Ha):** There is a significant difference in the cancellation rates among at least some market segments.

## RESULTS:

Since all p-values are extremely small (below the conventional alpha level of 0.05), **we reject the null hypothesis** for all market segments. This indicates that there are significant differences in the cancellation rates across different market segments.

Market Segments Test Results:		
	t-value	p-value
Market Segment		
Direct	-68.407863	0.000000e+00
Groups	76.447182	0.000000e+00
Undefined	449.797421	0.000000e+00
Corporate	-34.459408	8.402745e-238
Complementary	-20.444680	4.069585e-74
Offline TA/TO	-10.026756	1.244111e-23
Aviation	-5.512365	9.284044e-08
Online TA	-2.152140	3.138832e-02

## HYPOTHESIS NO.3:

The likelihood of booking cancellations varies by the months.

**Null Hypothesis (H<sub>0</sub>):** There is no correlation between lead time and cancellation rate.

**Alternative Hypothesis (H<sub>a</sub>):** There is a significant correlation between lead time and cancellation rate.

Pearson correlation coefficient: 0.2929

## RESULTS:

With a Pearson correlation coefficient of 0.2929, there is a **positive correlation** between lead time and cancellation rate, suggesting that as the lead time increases, the cancellation rate also tends to increase.

- **Null Hypothesis is rejected** because the correlation is significantly different from zero.
- **Alternative Hypothesis is accepted:** There is a significant positive correlation between lead time and cancellation rate.

Thus, the analysis supports the hypothesis that guests with higher lead times are more likely to cancel their bookings.

Thus, the analysis supports the hypothesis that guests with higher lead times are more likely to cancel their bookings.

A decorative graphic consisting of several thin, white, intersecting lines on a black background, located in the top-left corner of the slide.

# SUGGESTIONS FOR NEXT STEPS IN ANALYZING THIS DATA

You can enhance your analysis by visualizing additional features against each other, particularly focusing on their relationship with the cancellation rate. This could reveal more significant patterns and interactions between attributes that influence booking cancellations. To make the data more useful for machine learning models, consider incorporating further feature engineering based on the specific models you plan to use.

We can perform cluster analysis to segment customers based on their booking and cancellation behavior. This can help in understanding different customer types and their cancellation patterns.