# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 24 April 2024 |
| Team ID | 740663 |
| Project Title | RESERVATION CANCELLATION PREDICTION |
| Maximum Marks | 6 Marks |

**Data Exploration and Preprocessing Template**

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

| Section | Description |
|---|---|
| Data Overview | |

```
[ ] train_data.describe()
```

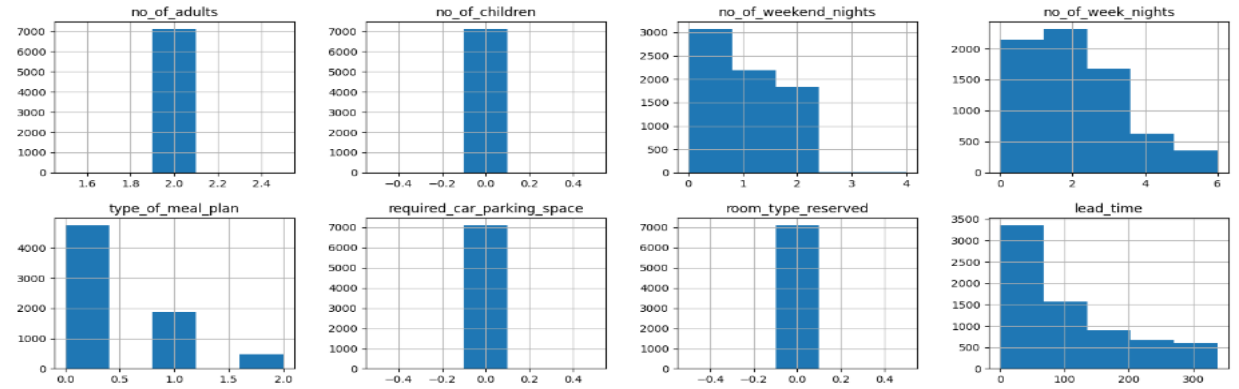
| | no_of_adults | no_of_children | no_of_weekend_nights | no_of_week_nights | type_of_meal_plan |
|---|---|---|---|---|---|
| count | 18137.000000 | 18137.000000 | 18137.000000 | 18137.000000 | 18137.000000 |
| mean | 1.846777 | 0.107515 | 0.811104 | 2.208965 | 0.318465 |
| std | 0.516020 | 0.408901 | 0.873470 | 1.426365 | 0.629140 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 2.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| 50% | 2.000000 | 0.000000 | 1.000000 | 2.000000 | 0.000000 |
| 75% | 2.000000 | 0.000000 | 2.000000 | 3.000000 | 0.000000 |
| max | 4.000000 | 9.000000 | 7.000000 | 17.000000 | 3.000000 |

```
[ ] test_data.describe()
```

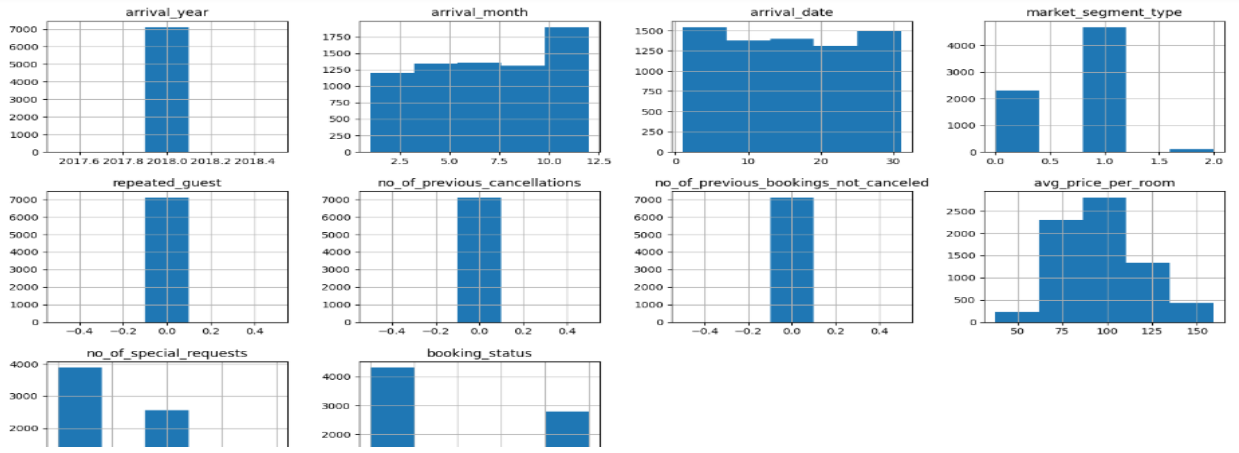
| | no_of_adults | no_of_children | no_of_weekend_nights | no_of_week_nights | type_of_meal_plan |
|---|---|---|---|---|---|
| count | 18138.000000 | 18138.000000 | 18138.000000 | 18138.000000 | 18138.000000 |
| mean | 1.843147 | 0.103043 | 0.810343 | 2.199636 | 0.329639 |
| std | 0.521403 | 0.396295 | 0.867833 | 1.395298 | 0.639016 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 2.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| 50% | 2.000000 | 0.000000 | 1.000000 | 2.000000 | 0.000000 |
| 75% | 2.000000 | 0.000000 | 2.000000 | 3.000000 | 0.000000 |
| max | 4.000000 | 10.000000 | 6.000000 | 16.000000 | 3.000000 |

Univariate Analysis

```
[ ] filtered_data.hist(bins=5, figsize=(18, 18))
    plt.show()
```

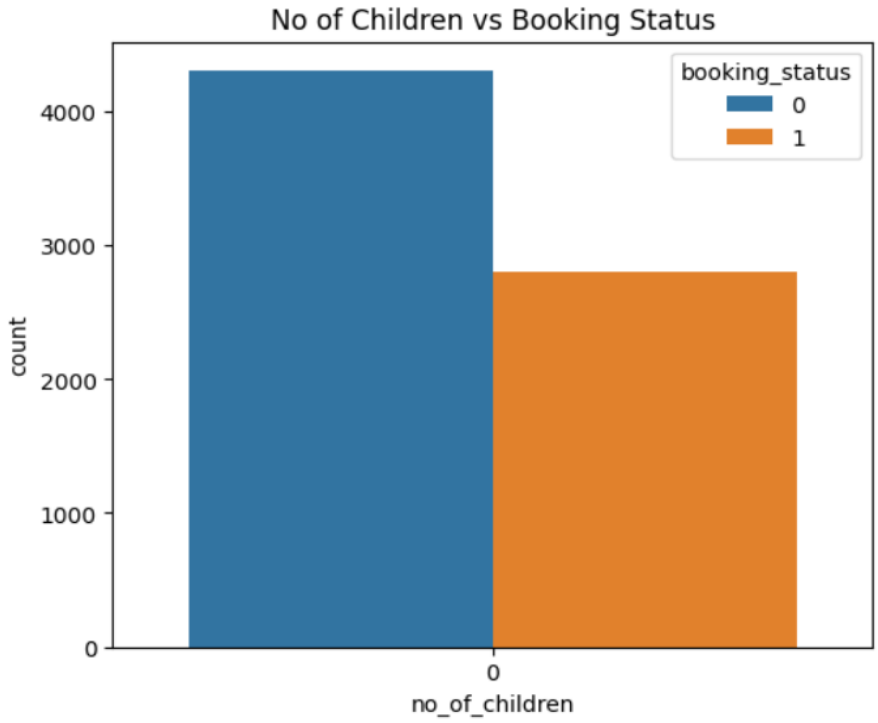



```
[ ]
```

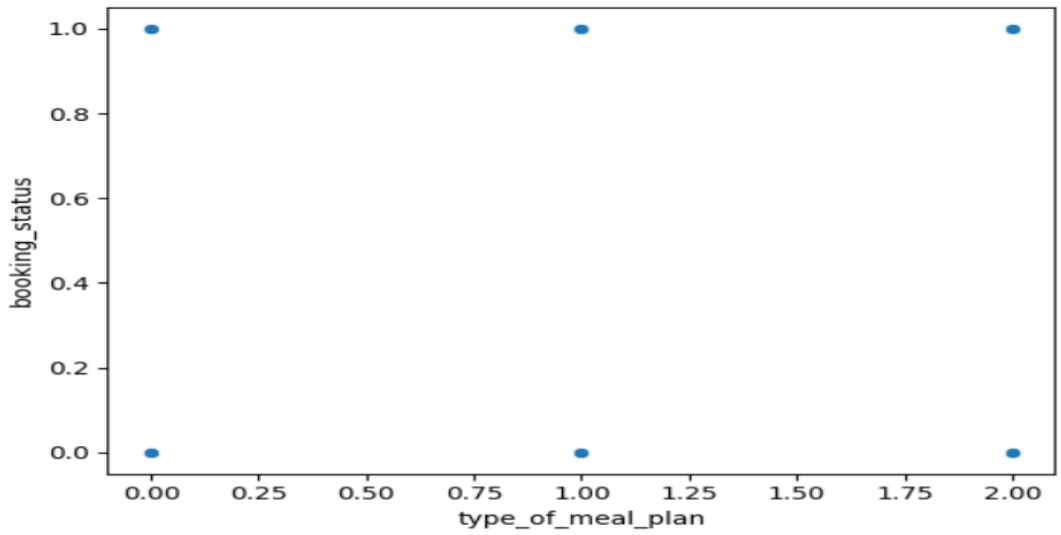


Bivariate Analysis

```
countplot_of_z(x='no_of_children', hue='booking_status', title='No of Children vs Booking Status')
```



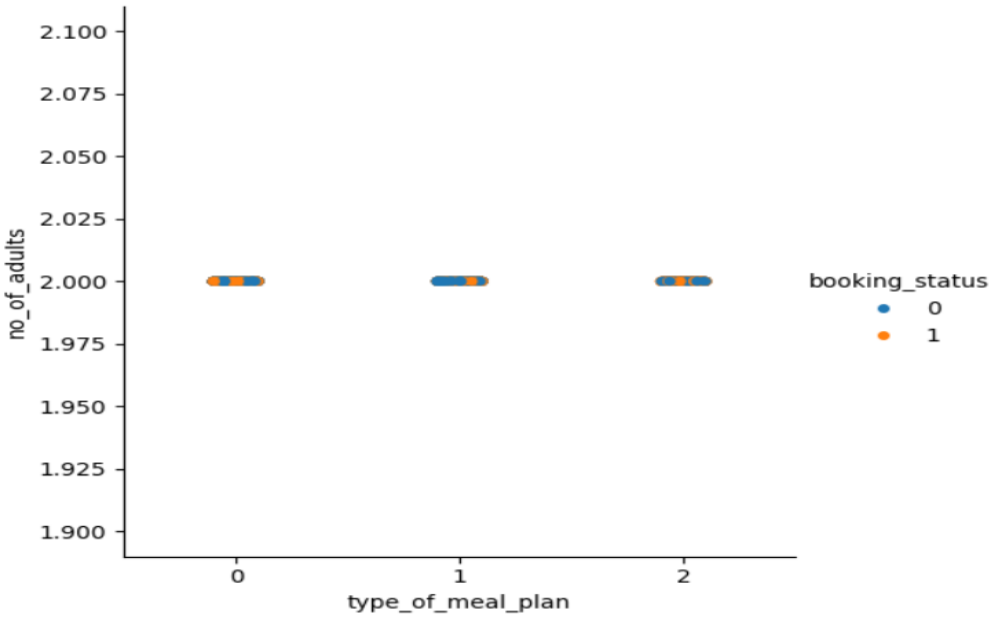

```
[ ] sns.scatterplot(data=filtered_data, x='type_of_meal_plan', y='booking_status',)
    plt.show()
```



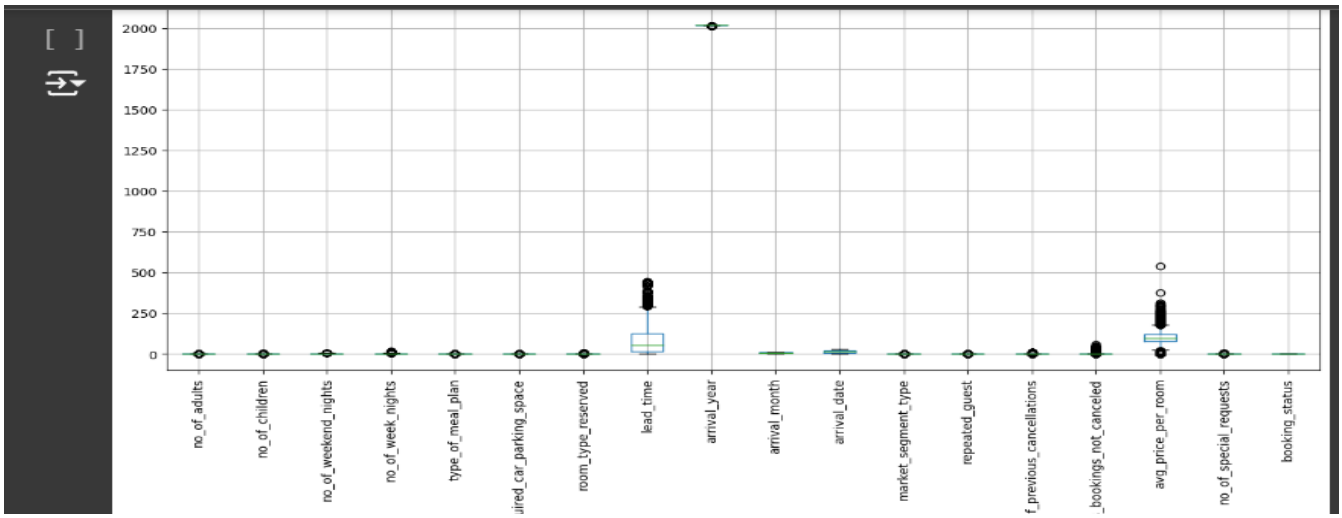
Multivariate Analysis

| | no_of_adults | no_of_children | no_of_weekend_nights | no_of_week_nights | type_of_meal_plan | required_car_parking_space | room_type_reserved | le |
|---|---|---|---|---|---|---|---|---|
| no_of_adults | nan | nan | nan | nan | nan | nan | nan | |
| no_of_children | nan | nan | nan | nan | nan | nan | nan | |
| no_of_weekend_nights | nan | nan | 1.000000 | 0.002152 | -0.048610 | nan | nan | ( |
| no_of_week_nights | nan | nan | 0.002152 | 1.000000 | -0.108686 | nan | nan | ( |
| type_of_meal_plan | nan | nan | -0.048610 | -0.108686 | 1.000000 | nan | nan | ( |
| required_car_parking_space | nan | nan | nan | nan | nan | nan | nan | |
| room_type_reserved | nan | nan | nan | nan | nan | nan | nan | |
| lead_time | nan | nan | 0.013533 | 0.189781 | 0.019528 | nan | nan | |
| arrival_year | nan | nan | nan | nan | nan | nan | nan | |
| arrival_month | nan | nan | 0.008513 | 0.082517 | -0.024910 | nan | nan | ( |
| arrival_date | nan | nan | 0.053746 | -0.006413 | 0.054452 | nan | nan | -( |
| market_segment_type | nan | nan | 0.037351 | -0.031379 | 0.055280 | nan | nan | -( |
| repeated_guest | nan | nan | nan | nan | nan | nan | nan | |
| no_of_previous_cancellations | nan | nan | nan | nan | nan | nan | nan | |
| no_of_previous_bookings_not_canceled | nan | nan | nan | nan | nan | nan | nan | |
| avg_price_per_room | nan | nan | -0.125499 | -0.112414 | 0.148880 | nan | nan | -( |
| no_of_special_requests | nan | nan | 0.051803 | 0.013128 | 0.063569 | nan | nan | -( |
| booking_status | nan | nan | -0.055431 | 0.026674 | 0.112614 | nan | nan | ( |

```
sns.catplot(x='type_of_meal_plan', y='no_of_adults', hue='booking_status', data=filtered_data)
plt.show()
```



Outliers




**Data Preprocessing Code Screenshots**

Loading Data

```
#reading the test data
test_data=pd.read_csv('/content/test___dataset.csv')
#reading the train data
train_data=pd.read_csv('/content/train__dataset.csv')
```

```
test_data.head()
```

| | no_of_adults | no_of_children | no_of_weekend_nights | no_of_week_nights | type_of_meal_plan | rec |
|---|---|---|---|---|---|---|
| 0 | 2 | 0 | 1 | 2 | 0 | |
| 1 | 2 | 0 | 0 | 2 | 0 | |
| 2 | 1 | 0 | 2 | 3 | 0 | |
| 3 | 2 | 0 | 2 | 0 | 2 | |
| 4 | 2 | 0 | 1 | 4 | 0 | |

```
train_data.head()
```

| | no_of_adults | no_of_children | no_of_weekend_nights | no_of_week_nights | type_of_meal_plan | rec |
|---|---|---|---|---|---|---|
| 0 | 2 | 0 | 1 | 4 | 0 | |
| 1 | 2 | 1 | 0 | 2 | 0 | |
| 2 | 1 | 0 | 1 | 5 | 0 | |
| 3 | 1 | 0 | 2 | 4 | 0 | |
| 4 | 2 | 0 | 0 | 4 | 1 | |

Handling Missing values

```
[ ] test_data.isna().sum()
```

```
no_of_adults                            0
no_of_children                          0
no_of_weekend_nights                    0
no_of_week_nights                       0
type_of_meal_plan                       0
required_car_parking_space              0
room_type_reserved                      0
lead_time                               0
arrival_year                            0
arrival_month                           0
arrival_date                            0
market_segment_type                     0
repeated_guest                          0
no_of_previous_cancellations            0
no_of_previous_bookings_not_canceled    0
avg_price_per_room                      0
no_of_special_requests                  0
dtype: int64
```

```
[ ] train_data.isna().sum()
```

```
no_of_adults                            0
no_of_children                          0
no_of_weekend_nights                    0
no_of_week_nights                       0
type_of_meal_plan                       0
required_car_parking_space              0
room_type_reserved                      0
lead_time                               0
arrival_year                            0
arrival_month                           0
arrival_date                            0
market_segment_type                     0
repeated_guest                          0
no_of_previous_cancellations            0
no_of_previous_bookings_not_canceled    0
avg_price_per_room                      0
no_of_special_requests                  0
booking_status                          0
dtype: int64
```

| | |
|---|---|
| Save Processed Data | [ ] `import joblib # Import the pickle m`<br>`joblib.dump(model, 'model.pkl')`<br><br>`['model.pkl']` |