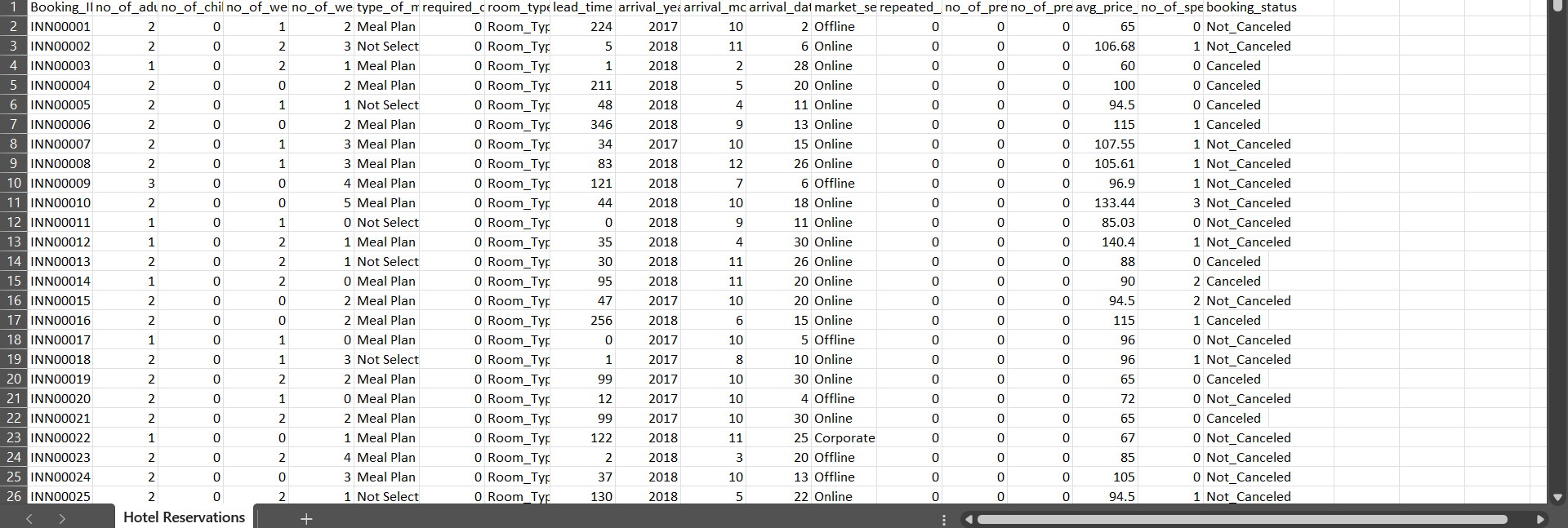
**Machine Learning and Data Science Capstone Project**

**1.Data Collection:**

I have take **Hotel Reservation** dataset

**Upon analyzing the dataset, over 30% of the guests cancelled their reservations. what drives these cancellations? Is it the cost of the booking, the time between booking and check-in (lead time), or perhaps the booking method (online vs. offline)?**

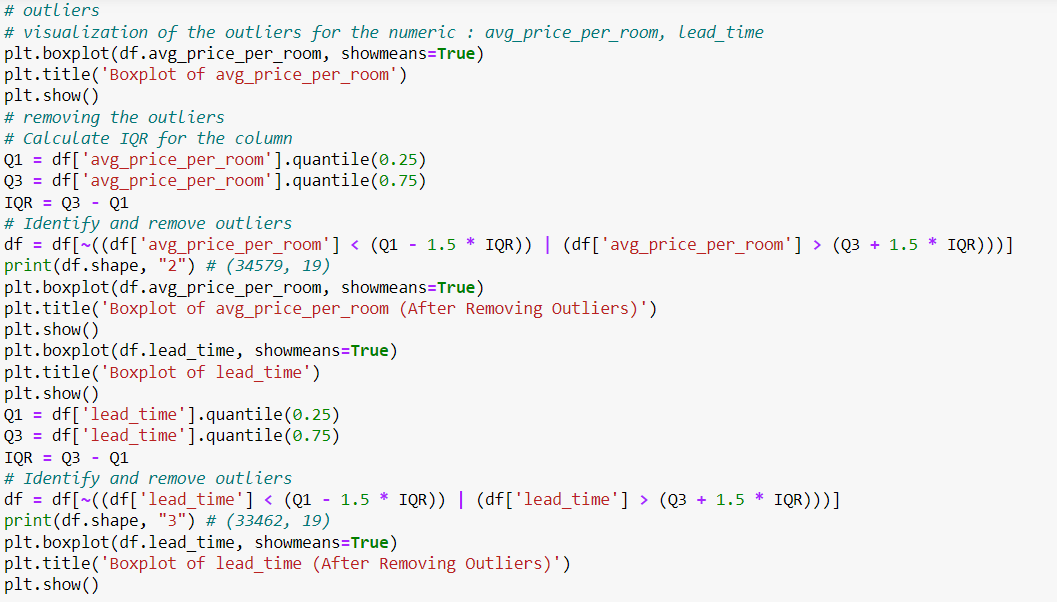


Above image is the screenshot of the **Hotel Reservation** dataset.

**2.Data PreProcessing:**

Check the dataset using **dataset. isnull().sum()** function

Handling Outliers through boxplots, notable outliers were identified in both **'avg\_price\_per\_room' and 'lead\_time'.**Using the **Interquartile Range (IQR)** method, these outliers were removed. The data was trimmed from 34,579 entries down to 33,462, highlighting the significant impact of outlier removal.

****

3. Univariate nd Bivariate :

For Univariate and Bivariate I use **Tableau ,** below are the statements which are done using tableau.

1.How many num of weekend nights were reserved at which type of room.

**No\_ of\_ weekend\_ nights = room\_ type\_ reserved**

2. How many num of week nights were reserved at which type of room.

**No\_ of\_ week\_ nights = room\_ type\_ reserved**

3.what is the price of room for pervious cancellations.

**No\_ of\_ previous\_ cancellations = avg \_ price\_ per\_ room**

4.what is the price of room for perivous bookings not cancelled.

**No\_ of\_ previous\_ bookings\_ not\_ cancelled = avg \_ price\_ per\_ room**

5.Booking status for average price per room

**Avg \_ price\_ per\_ room = booking\_ sts**

6.Booking status for market segment type.

**Market\_ segment\_ type = booking\_ sts**

7.Type of meal type preferred based on the room type reserved.

**Type\_ of\_ meal\_ plan = room\_ type\_ reserved**

8.what is the price of room for repeated guest

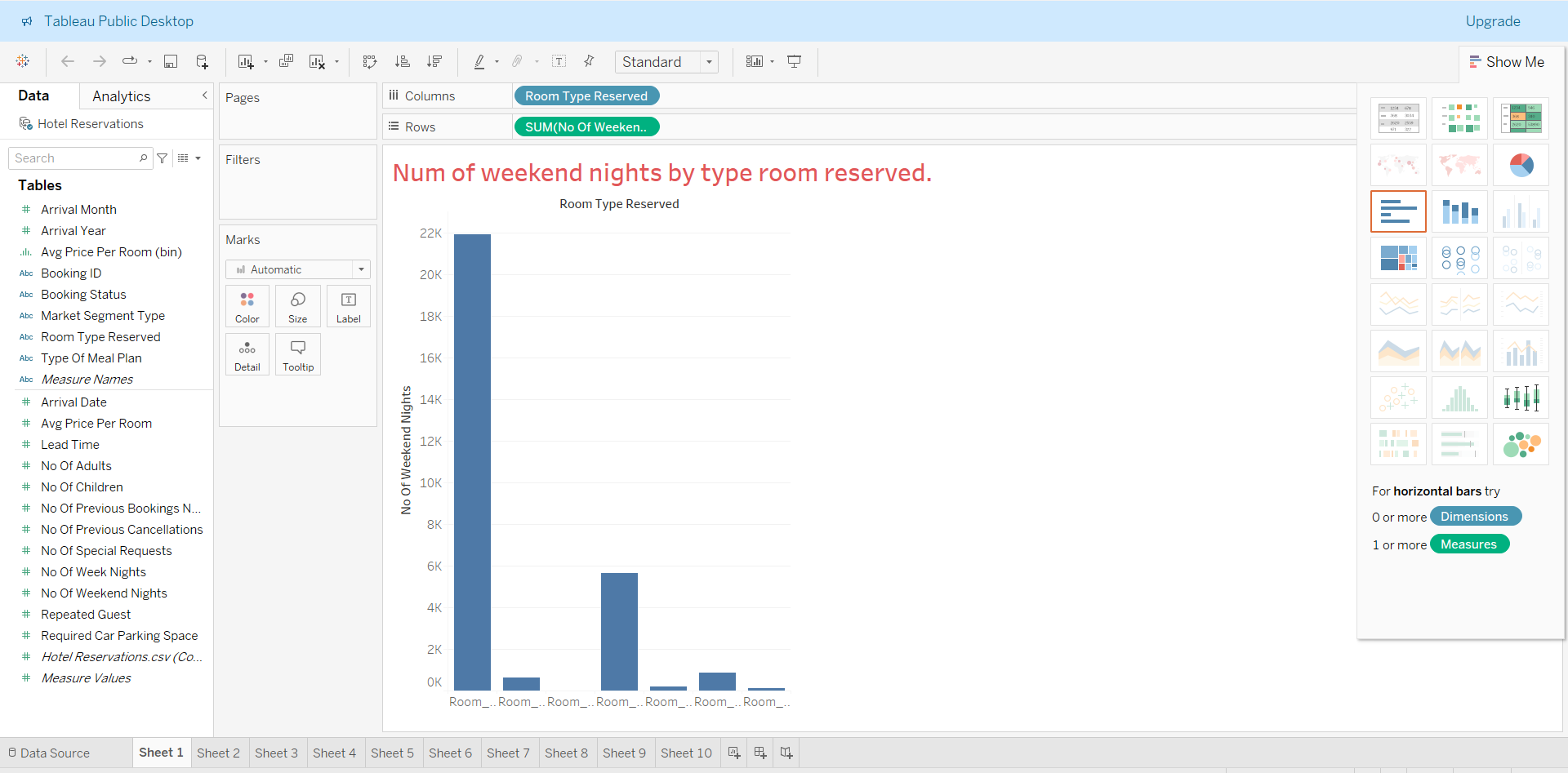
**Repeated\_ guest = price\_ per\_ room**

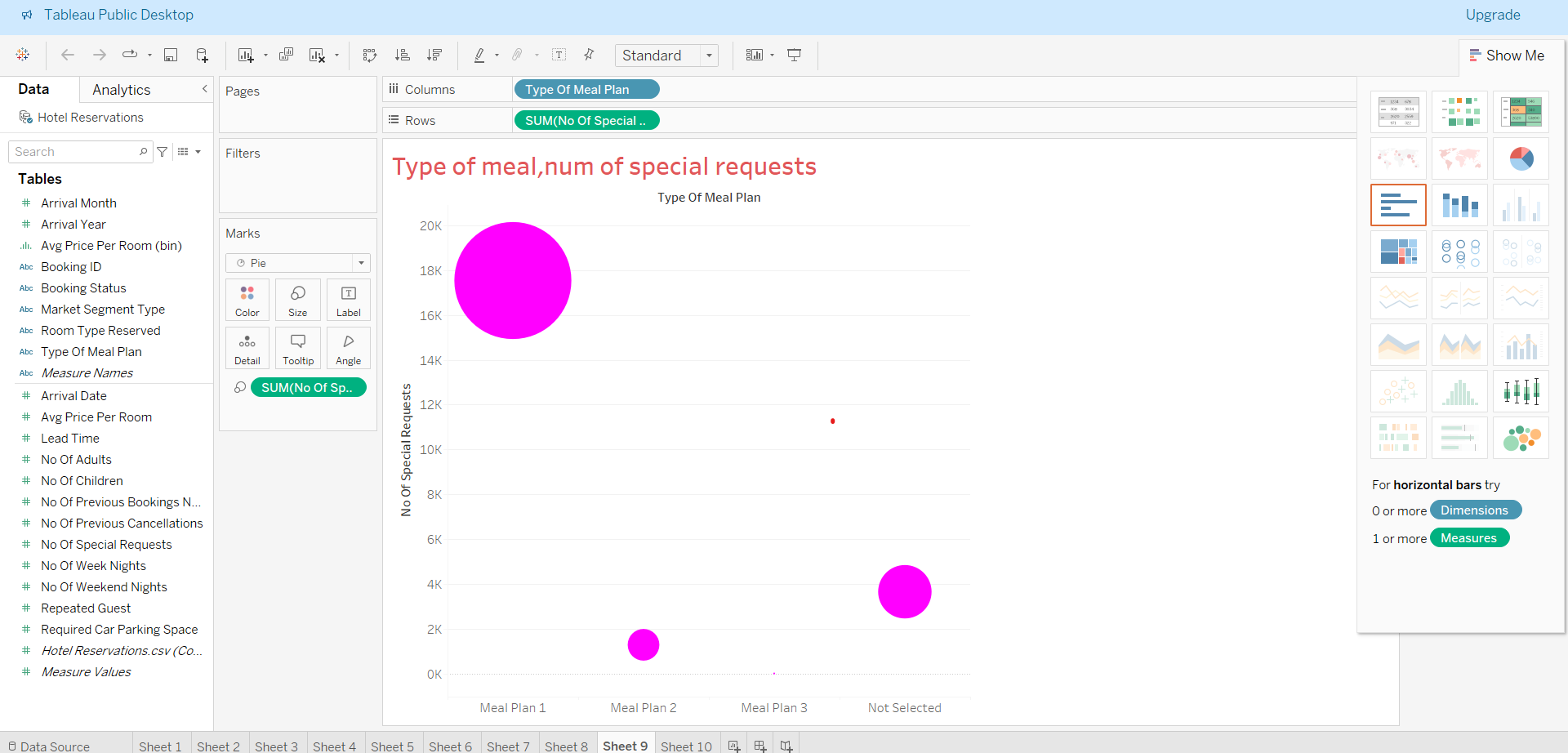
9.Type Mean plan for special request

**Special\_ guest = meal\_ plan**

10.Is room price is increased based on num of adults

**No\_ of\_ adults = price\_ per\_ room**





**4.Model Creation:**

* The primary objective is to **predict hotel booking cancellations** to mitigate revenue loss due to unoccupied rooms.
* Dataset contains detailed information about hotel bookings, including customer specifics and whether the booking was cancelled or not.
* Aim to develop a **Classification model** that can predict future booking cancellations accurately.
* This model can enable hotel management to identify potential cancellations early.

**Success Metrics**

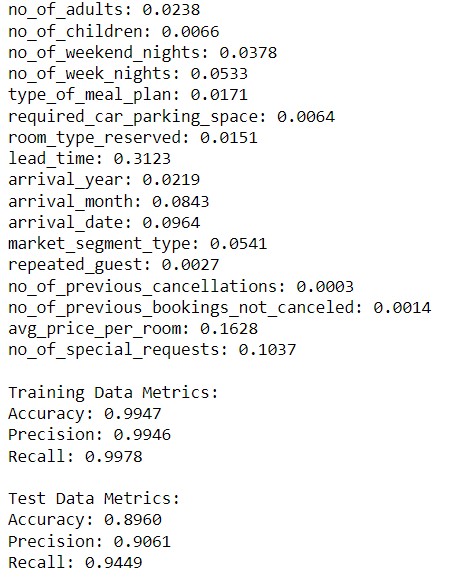
The success of model will be evaluated based on its predictive performance.

* **Accuracy:** This is the most intuitive performance measure. It is simply the ratio of correctly predicted observations to the total observations.
* **Precision:** Precision looks at the ratio of correct positive observations to the total predicted positives. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
* **Recall :** Recall is the ratio of correct positive observations to the all observations in actual class. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
* **F1 Score:** The F1 Score is the weighted average of Precision and Recall. This score tries to balance both precision and recall. It is suitable for uneven class distribution problems.

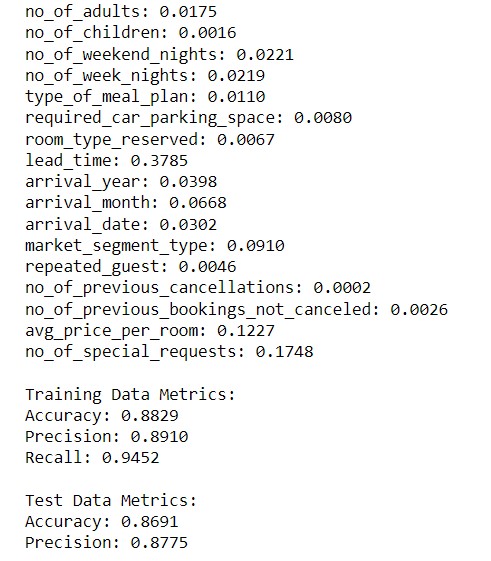
**5.Final Model:**

1. **Encoding Categorical Features**: The categorical columns ('type\_ of\_ meal\_ plan', 'room\_ type\_ reserved', 'market\_ segment\_ type', 'booking\_ status') in the dataset are encoded using One\_ Hot\_ Encoder to convert them into numeric values suitable for machine learning models.
2. **Data Splitting:** The dataset is split into predictors (features) and target (booking status). This data is then divided into a training set (70% of data) and a test set (30% of data).
3. **Model Building and Evaluation:** Initially, a **Random Forest classifier** is built and trained, followed by its evaluation on both training and test sets. Given signs of overfitting in this model, a second Random Forest classifier with additional hyperparameters (like max\_depth, min\_samples\_split, etc.) is constructed to prevent overfitting. This revised model is trained, evaluated, and its feature importance is also displayed.

Final Model:



After overfitting:



Conclusion:

The refined model offers a reliable prediction on whether a guest will cancel. A key takeaway from this analysis is the direct correlation between order lead time and cancellation .