**Final Project Report**

**Introduction to Data Analytics**

**Project Title:**

**Prediction/Analysis of booking cancellations using Online travel booking company data**

**Prepared by:**

**Justice Arthur (N01613631)**

**Luting Chiu (N01604196)**

**Sachindra (N……….)**

**AIGC 5000 IGA – Fall 2023**

**Humber College**

**1. Problem Statement**

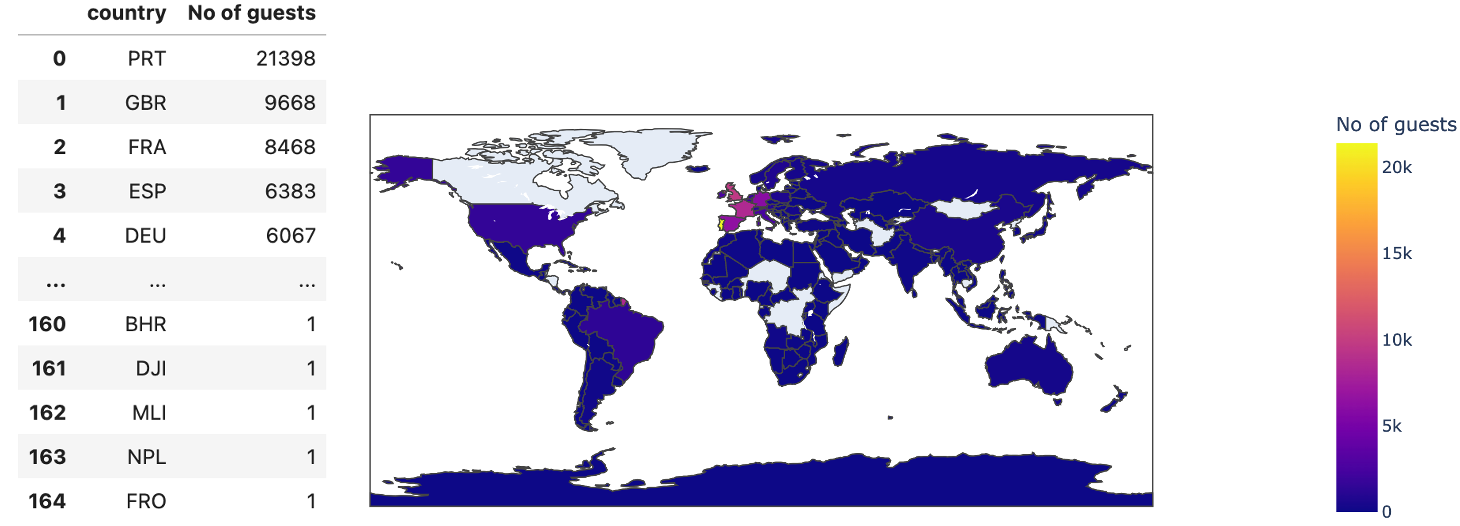
* Prediction/Analysis of booking cancellations using Online travel booking company data

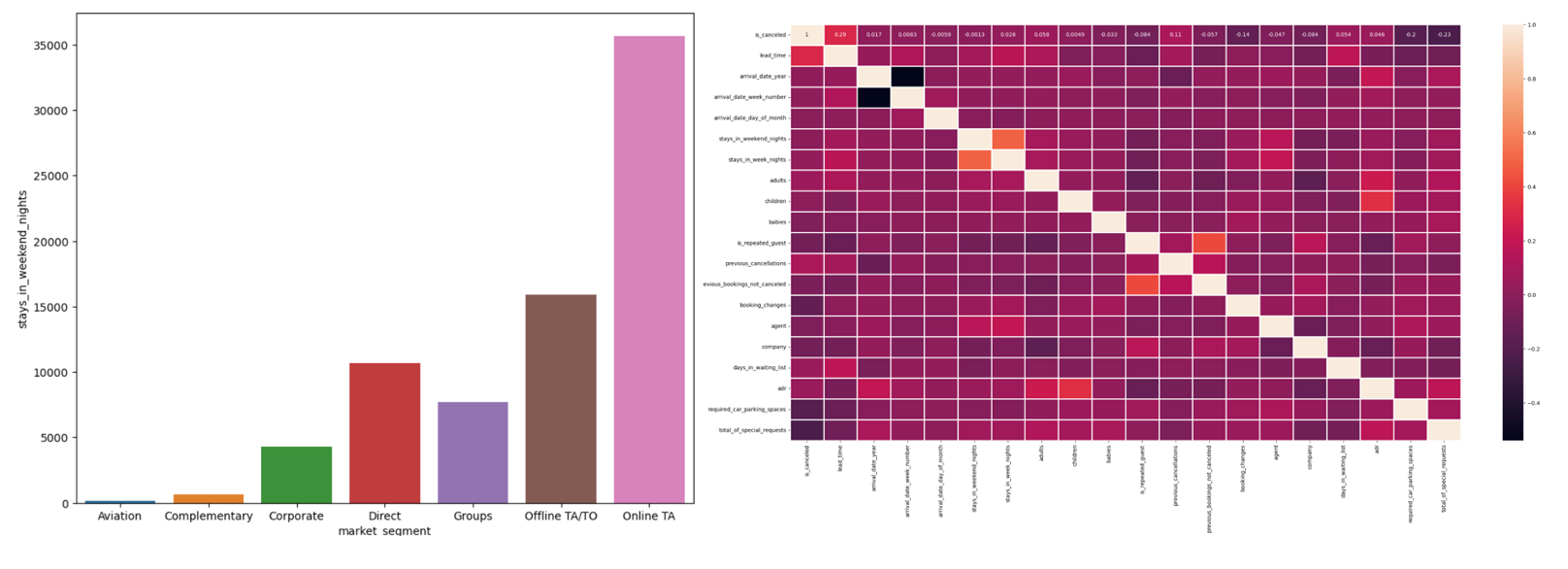
**2. Dataset Description**

* The cancellation status (is\_canceled: canceled:1; not canceled: 0) depends on various factors, including customer behavior, time, country, and individual background.
* Dependent variables such as hotel, lead\_time, arrival\_data\_year, arrival\_date\_month, arrival\_date\_week\_number, arrival\_date\_day\_of\_month, meal, contry, market\_segment, distribution\_channel, previous\_cancellations, previous\_booking\_not\_canceled, deposit\_type, days\_in\_waiting\_list, customer\_type, AveragedailyRate which I used to train the model for predicting cancellation.

**3. Dataset Analysis and Observations**

* For dataset analysis, we identify the country with the majority of not-canceled reservations for univariate analysis. Subsequently, we create a bar plot for bivariate analysis to explore the relationship between market\_segment and stay\_in\_weekend\_nights in not-canceled population. Next, we create a heatmap to identify variables correlated with the cancellation status.
* **Observation:**



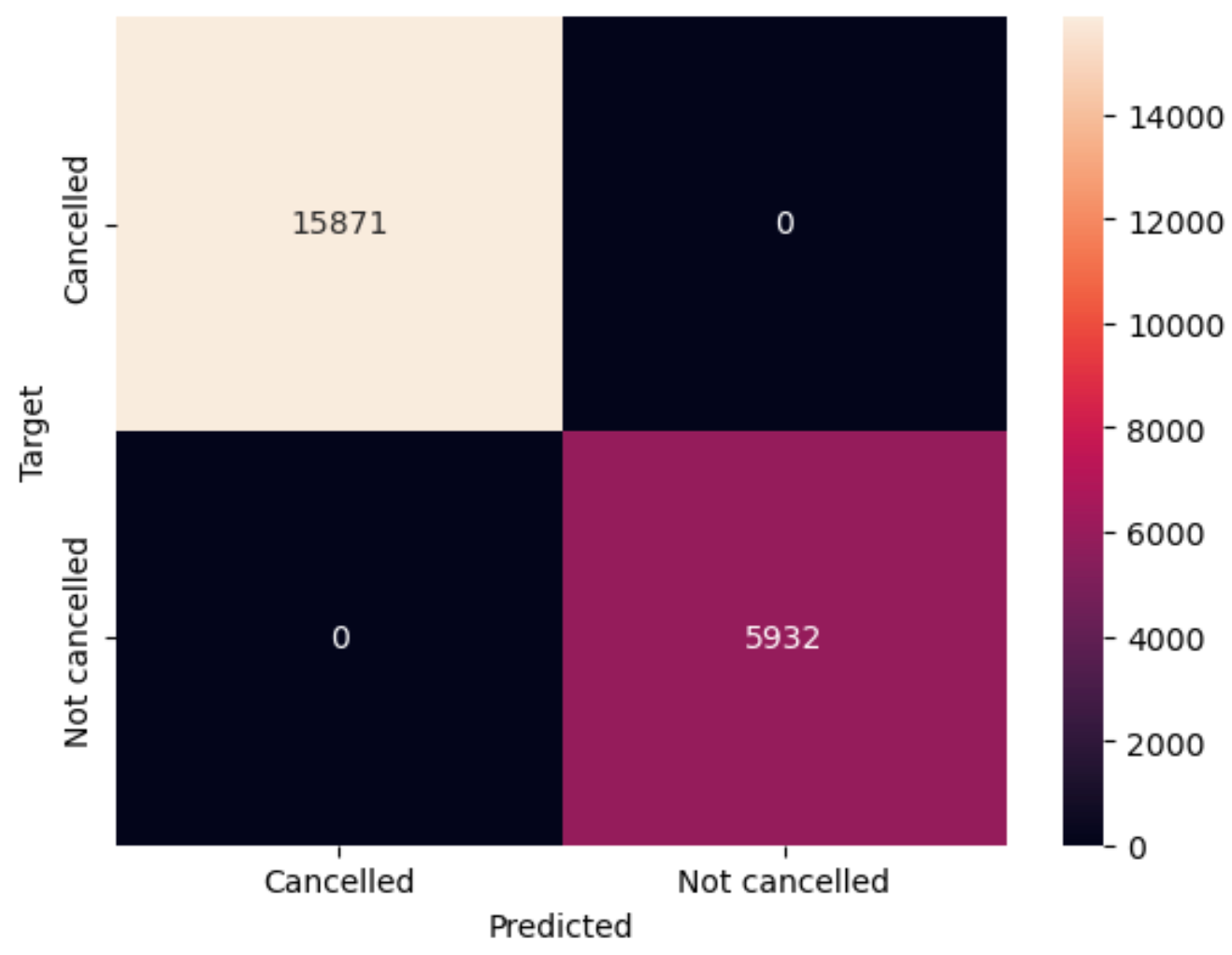
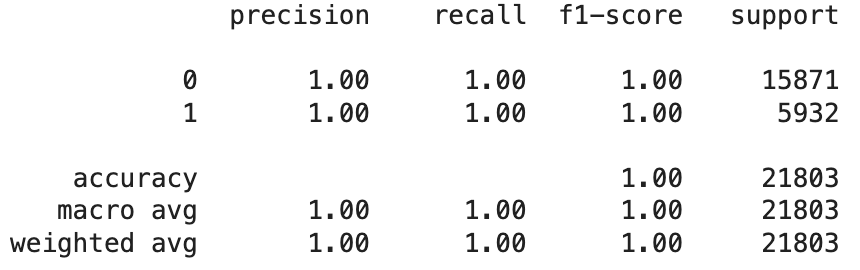


**4. Proposed Analytical/Prediction Model**

* Based on the above data exploration and engineering, we incorporate the following variables into our model: hotel, lead\_time, arrival\_date\_year, arrival\_date\_month, arrival\_date\_week\_number, arrival\_date\_day\_of\_month, meal, country, market\_segment, distribution\_channel, previous\_cancellations, previous\_booking\_not\_canceled, deposit\_type, days\_in\_waiting\_list, customer\_type, and AverageDailyRate.
* We split our dataset, allocating 75% to training data and the remaining 25% to testing data.
* Since our target variable is categorical (0 and 1), we apply a logistic regression model (binomial) to this dataset.

**5. Results and Discussions**

* We create a confusion matrix to observe our prediction results. Here, we observe that the number of false positives and false negatives is zero, indicating a 100% accuracy in predicting the cancellation status.
* Here we perform precision, recall and f1-score for this prediction:
  + Precision is the ratio of correctly predicted positive observations to the total predicted positives. It assesses the accuracy of the positive predictions made by the model.
  + Recall is the ratio of correctly predicted positive observations to the total actual positives. It measures the ability of the model to capture all the positive instances.
  + The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when there is an imbalance between the classes.
* As FP and FN are zero, both Precision and Recall are equal to 1. Therefore, F1, which combines Precision and Recall, is also equal to 1.

****