Hantao Lin
Springboard Data Science Career Track
February 2020 Cohort
Capstone Project 1

Data Wrangling

Problem: Using the provided data, predict the possibility of a client cancel their hotel appointment.

Dataset: There are two datasets used in this project. The first dataset is directly downloaded from Kaggle (https://www.kaggle.com/jessemostipak/hotel-booking-demand) in csv file. The dataset 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. It has 32 columns and 119390 entries. The other dataset is collected from wunderground.com also in csv file. The dataset includes the average temperature, wind speed, precipitation and other things at Lisbon, Portugal from 7/1/2015 to 8/31/2017. It has 9 columns and 793 entries.

Cleaning the Data: The datasets were joined for the analysis. In order to do so, a new column(“Date_time”) is created in Kaggle dataset based on the “arrival_year”, “arrival_month”, and “arrival_day” in the dataset. After joining the two datasets with “Date_time” as the common column, missing value were filled different in the new dataset. Missing value in “Heat_Index” is calculated based on formula provided online. Other missing values in other columns are fill with the mode of the column. Column “company” is dropped from the dataset because it is missing 80% of its values. After the cleaning, the final dataset has 40 columns in total and 119390 entries.

Saving the Data: The final dataset is saved as a csv file called “data” using the command ‘dataframe.to_csv()’ locally for easy access.