Data Wrangling

Data wrangling is an essential part of data science project. Without a clean data, much of analysis that comes after will not return accurate information. For my capstone project 1, I used Jupyter Notebook to work on the dataset.

I imported all essential module into the python notebook. All I need for now is pandas. After that, I imported the data files that I will be working on using the command “pd.read_csv”. Before I started doing any modifications on the data, I summarized both data files to get an idea of what I need to clean up using the command “.info()”. After viewing the summaries, I realized that:

1. some of the columns in one of the data frames can be combined to form a new column

2. there are NA values in many columns

In the data frames “hotel”, to combined different columns to form a new column, I took the values from the dataset and appended the data into new lists separately so I can synthesize the list together later. I used for loop to get the data and convert string value into float value. Considering that I need to join the two files together, I purposely modify the formation of the data into “month/day/year”. After I obtain the list of string of date I have, I append the column to the data set “hotel” using “hotel['Date_time']=date”. Next, I merged the two data frames (“hotel” and “weather”) together based on the column “Date_time” using the “.merge()” method.

Looking at the newly formed data that I called “merged”, I now need to fill in the NA values. For “children”, “agent”, and “country” columns, I filled the blank space with their mode respectively. For “Heat_Index” column, I found formula online to calculate its value using “Average_Temperature” and “Relative_Humidity” from the data frame and fill the values in. I used the method “.fillna()” for this step. Since the column “company” has almost over 80% data missing, I decided to drop the entire column because it is useless in analysis with such amount of data missing. I used the method “.drop” to complete this step. Finally, I called “.info()” again just to check if there is any missing values.

In conclusion, in this part of the capstone project, I merged the two data files, filled in the missing values, and modified any irrelevant columns.