Capstone Project 1 Data Wrangling

Timothy Baney

June 15th, 2017

As mentioned in my capstone intro, I downloaded all the necessary csv files publicly available on Kaggle. I created a new jupyter notebook, and moved all the downloaded csv files, as well as the notebook to its own directory. I imported all the required Python libraries I needed which were Pandas, math, and Numpy. I started by using Panda’s ‘read csv’ method to turn the AirBnB csv files into data frames. One of the csv files, sessions, was too large, and used up too much memory which kept causing the jupyter notebook kernel to crash. To solve this issue I had to load the csv files into a list by chunk.

When all the chunks had been processed, and appended to a list, I concatenated all the chunks into one data frame using Panda’s concat method. After having all the data in data frames ready to be manipulated, and cleaned I use the .head(), and .isnull().any() methods to get a rough idea of what kind of data I’m looking at for each data frame, without getting too much into exploratory data analysis, as well as to see if there are any null or missing values that are in the data. Sessions had null values in almost all of its columns, one of which being the user\_id column. I deleted all sessions that didn’t have an identifier since the data wouldn’t be any good to me otherwise.

After dropping all rows without IDs, all the other columns with null values also disappeared. The target variable, ‘country destination’, in the Users data frame had values of NDF or ‘No Destination Found’, rows with a destination of NDF also had a null value for its first\_booking\_date column. I deleted all the rows that had a value of NDF for its destination, which wouldn’t be useful to me in making predictions. As suspected all date\_first\_booking null values also disappeared. The age column not only had missing values, but also had some impossible outliers, ages all the way up in the 2000s. I used an inplace .fillna method on the users dataframe with the average age of all users as the filler. I than mapped through the dataframe, and replaced all values above 120 with the average age, since it is reasonable to assume that no one over theage of 120 years old is going to be going on AirBnB.

The ‘first affiliate tracked’ column in the Users table already has an 'untracked' category, so for any NaN values, I replaced it with 'untracked'. After filling in all the missing data, I started to combine the dataframe using Pandas merge method. I merged the Users and Countires data frames on the shared ‘country\_destination’ column. The age\_gender\_buckets dataframe was a tricky one to merge in since I needed to merge on multiple keys: gender, country destination, and age. However the age column in age\_gender\_buckets was a string representation of the range of ages for that row, e.g ’14-19’. I mapped through the rows, applying a method that would split the string range into an actual list of integers. I than had create a new data frame which had repeated rows for each age in its age\_list column.