Weekly Update

Week2: Oct.5, 2020 - Oct.9, 2020
Project: Greenwich
Group: Gannett Peak
Members: Isaac Choi, Matt Ko, Binqi Shen, Congda Xu

We recall Diego's lecture informing us that it is common to spend most time cleaning data in any data-related projects. His teaching became a real-life scenario this week for us.

This week, our group mainly focused on downloading the data and preparing the Greenwich data. The main challenge we faced was uploading and combining 17GB worth of part files (.part_00000) to their corresponding tables onto the server. However, two conditions made our first task even more difficult. The first condition is that, according to our conversation with Greenwich, their data is typically not 'clean' and intentionally kept as such because many Greenwich's clients prefer to see the raw data. The second is that, with their proprietary software, Greenwich scrapes data from many different sources. These two conditions facilitated numerous text encoding issues in our part files, and we first had to fix various characters that cause uploading issues. We then configured tables so that the tables can accommodate the inconsistent datatypes present. The process took longer than what we expected, but we were able to successfully and collaboratively upload all the part files for all five tables in the end.

After successfully uploading the data, we performed some preliminary analyses on the tables in order to plan out our next phase of data cleansing. We learned that on the Master table, the earliest record according to modify_timestamp starts from "2019-03-03", which differs from Greenwich's data specification. Also, there are about 16.5 thousand companies, but among them, only about 3,900 companies have been normalized, and only about 3,300 companies have corresponding tickers. We may have to normalize the fields ourselves to widen the data visibility or else risk the missing data severely limiting our equity analysis later on. And even inside the ticker, we observed some other miscoded characters, which we will have to clean later on. The ticker with the most records was CRCM which is from care.com. They claim that they are "the world's largest online destination for care." This ranking may indicate some sampling biases in the data, and we will have to adjust for it later on. On Titles, Roles, and Tags, the number of unique job_id --primary key to the Master table, and foreign key to other tables-- records did not match, and the starting timestamp differed from table to table.

The rankings provided a similar story, as the top rankings were associated with sales, retails, and general managerial positions.

Starting this weekend, we will continue preliminary analyses of our dataset, and plan and execute the next phase of data cleansing processes. We also plan to tentatively decide on our input vectors that we are going to use for our model and schedule a meeting with our client by the end of next week. One of the potential challenges we see to proceed is the process of combining the financial data with the Greenwich data. For example, the financial data table might not have the same datetime interval as the Greenwich data's 'post_date' column, which can cause some errors when integrating the financial aspects to the original dataset. But we are excited to tackle any obstacles that come our way.