Documentation of any cleaning or manipulation of data

The tool that I used to process and clean the data was *Google Sheets*. I had prepared the dataset by importing the respective csv files into *Google Sheets*, which was a very helpful tool to immediately capture anomalies within the data. This was the driving factor behind using *Google Sheets* to complete the process step of the data analysis process.

I had quickly noticed that there were blanks present for fields concerning station ids and station names. The first action that I had taken to clean the dataset was to remove any instances of an empty field for *start\_station\_name*, *start\_station\_id*, *end\_station\_name*, and *end\_station\_id*.

Once the blank station ids/names were removed, I had created an additional field to determine whether the start and end times for each trip were identical. This was accomplished by using a conditional if statement. For the records, where the fields *started\_at* and *ended\_at* were identical, I had deleted these records. It doesn’t make sense for the times to be identical even if the user had returned to the same station because time elapses. This was the reasoning behind the deletion of these entries.

For the next step of the data cleaning process, I had used a pivot table to evaluate if a station name contained multiple entries for its station id. I believe that this is an anomaly because each station name should contain one distinct station id. If a station name had multiple station ids,I would tally up the number of instances based upon station id. The station id with the lowest count would be removed from the dataset.

I had found records where a station id can have multiple station names. In order to resolve this issue, I had determined the distances between the multiple station names and the total number of entries of each station within the dataset.

If the distances between the stations were small, I had changed the station name with the fewest entries to the station name with the most entries. If the distances were large, then the station names with the fewest entries were deleted. Lastly, if the number of entries amongst the station names were relatively equal, I had removed the records containing that station\_id from the dataset. The details about this particular analysis can be found on the *Google Sheet* entitled “4-Analysis STN ID AND STN NAME-Rev 2”.

For the final step of the cleaning process, I had removed the fields *start\_lat*, *start\_lng*, *end\_lat*, *end\_lng*, and *Does started\_at = ended\_at* because they will not be required during the analysis step of the data analysis process. However, I had added the fields *ride\_length*, *weekday\_started\_at*, and *weekday\_ended\_at* which will be required for the analysis step. These are calculated fields, where its operands are the fields *started\_at* and *ended\_at*.