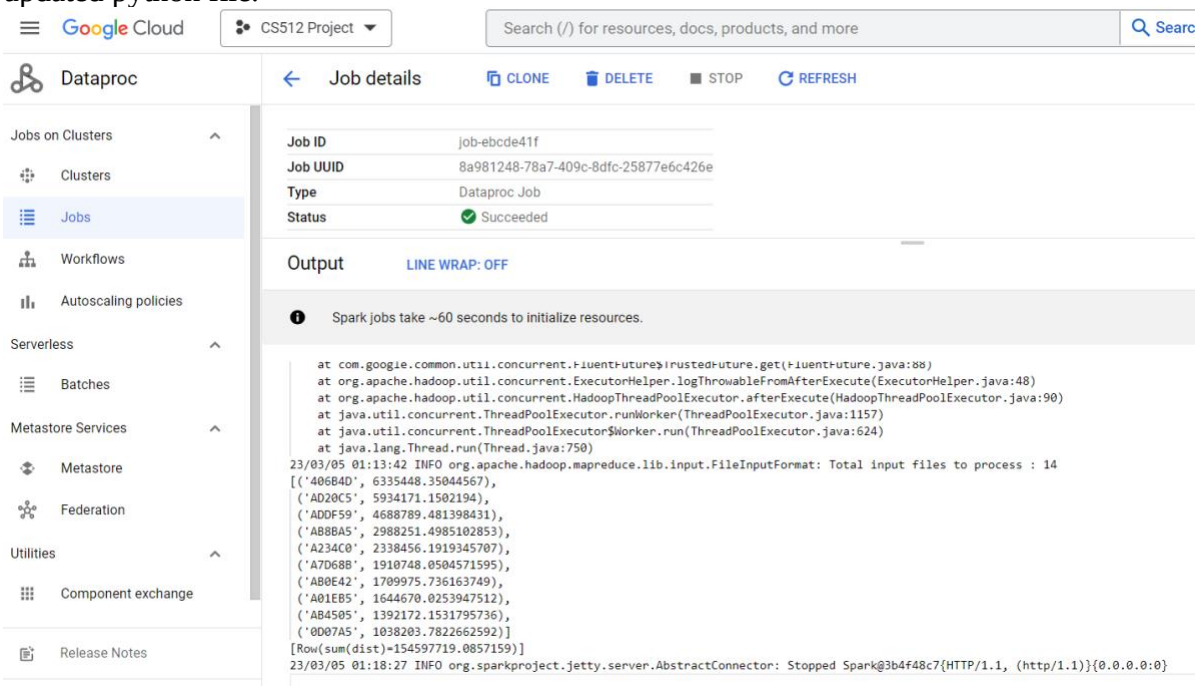
Spark Plane Distances Part 2 Report

Rahul Kumar Nalubandhu and Sandra Estrada

For this assignment Initially we downloaded the updated python file from the assignment page and updated with the respective project_id and storage details and later uploaded that into to the bucket and successfully ran the job on DataProc and below is the total distance found from using the given updated python file.

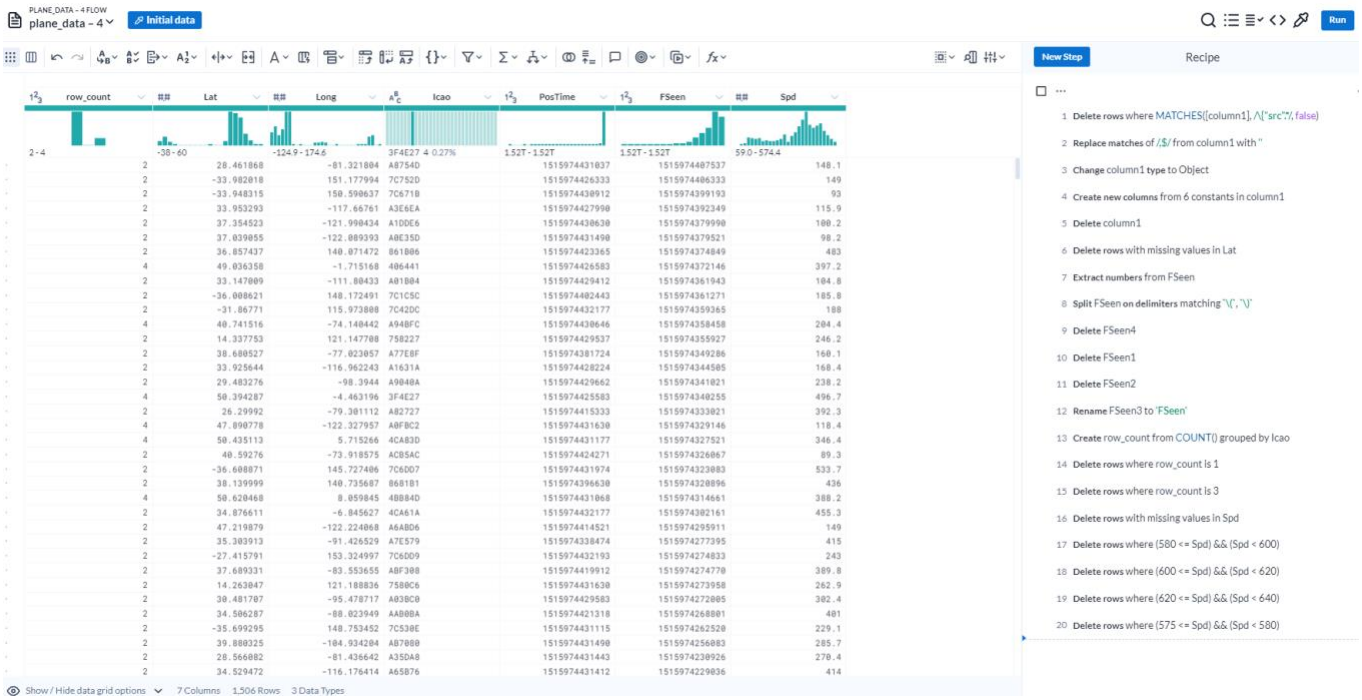


Description of Removing Bad data:

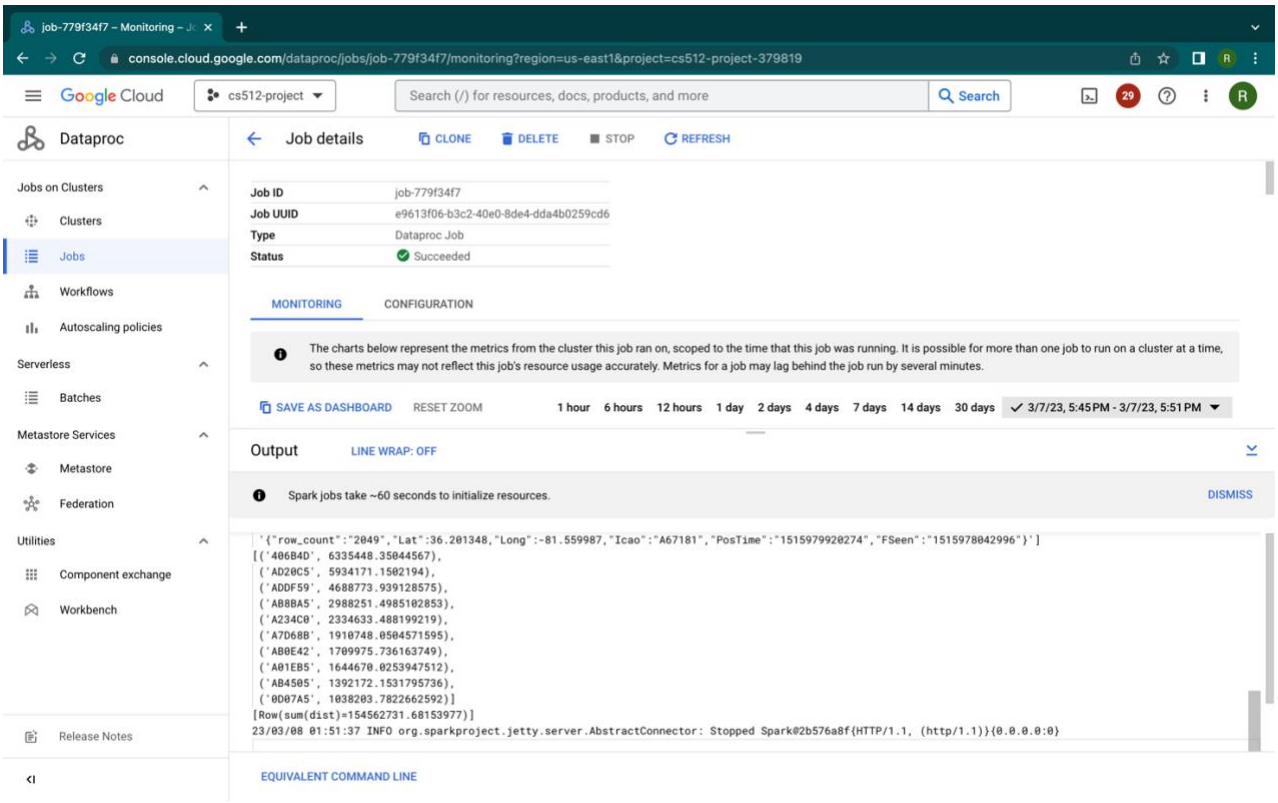
Upon running the job with cleaning from the previous assignment, 3,992 rows were filtered. For the current assignment, a filter was applied to the "Icao" column. To accomplish this, a new column was created using "groupby" and "count" functions to generate a count of values for each unique "Icao" value. The count for "Icao" was observed to range from 1 to 4. However, to ensure valid distance calculations, an even count for each "Icao" value was required, as two sets of latitude-longitude coordinates are needed for distance calculation. As a result, any values in the new column corresponding with odd numbers (1 and 3) were deleted, leaving only even values. This reduced the row count to 1,560 rows.

Later, the "Spd" column was examined, and rows were filtered by deleting null values. As part of the speed value constraints, the maximum speed range of a commercial airplane was obtained from the internet where the maximum speed range was found to be between 547-575 mph, and any values exceeding this range were deleted. After applying these constraints, the resulting dataset consisted of 1,506 rows of valid data.

Data cleaning:



DataProc:



Data Cleaning Attempts:

A few additional steps were taken that were well intentioned but did not end up making a difference in the data set. One of those failed attempts consisted of removing null values for the longitude column, but since all null values had already been removed for the latitude column in a previous week the data set was being analyzed, there were no null values to be removed under the longitude column. The next thing that came to mind was to check if any of the longitude and latitude values were outside of the range and remove those. New rules were added to the recipe to delete rows where the latitude values were outside of –90 to 90, and for longitude values outside of the –180 to 180 range. However, all values were within the required range, so these steps were not essential in cleaning up the data. Lastly, we tried to modify the latitude and longitude values to limit the decimal points to two values in an attempt to remove any duplicated rows in each ”Icao”. Unfortunately, during this process the value was rounded to the second decimal point. When the rule was created to remove duplicated rows, we were unable to see any impact to the dataset. These were all attempts and ideas aiming to reduce the number of outliers that made sense at the time, or that seemed like possible opportunities to clean the data but essentially had no impact on the data set.