ETL Project:

The Process

The project that we decided to do is about medical financial, focusing on coronary heart disease. We found three datasets that are related to each other. The first data that we saw (IPPS-all.csv) contains that the DRG, financial information, and the provider id. The following data (OSHPD.csv) contains the OSHPD Id. Finally, the last data (cover ca.csv) have both provider id and OSHPD id, which are used to join the three tables together.

For the IPPS data, we decided to focus only on the state of California. We also filter the DRG that contains 233, 234, 235, and 236. The following steps that we did were renaming some of the columns that would be proper for our database.

Next, for the cover ca dataset, we chose for columns: entity name, Medicare provider number, and oshpd id. We dropped all the rows that contain null. We also converted the Medicare provider number and OSHPD to an integer type. Next, we rename the column to the appropriate columns name.

Finally, our last data set, OSHPD. To open the file, we needed to use an encoding, which in this case, we used Unicode\_escape. We filter out the table to the column that we want and look for a specific year, which is 2017. We dropped rows that contain no value, although it doesn’t say null. We also convert OSHPD ID to float and rename all the columns to fit better with other data sets.

The next step that we do after cleaning all the data, we decided to use PostgreSQL and export our data frame and create a table. Some of the challenges that we face were not being able to query any of the tables that we exported. We found out that PostgreSQL cannot read the upper-case letter table, so we had to convert all the tables into a lowercase table to work.