**Data Exploration and Cleaning**

**11)**

We had a very large dataset with over 1 million rows and 30+ columns. Many of the columns either had redundant or information that would not be useful to compare to CO2. Therefore, we were able to reduce the columns down to 12. We further cleaned the data by removing the null values, renaming columns to be more understandable. Our data had a mix of categorical and numerical data types so we spilt these columns into 2 separate tables to be able to better see what might fit together. We noticed several of the categorical columns would be interesting to compare to CO2 emissions, so we used LabelEncoder from the Sklearn python library to assign numeric values to these objects. We were able to see most of the variables had linear relationships to emissions which led us to choose a linear regression model for our machine learning which Hakeem will dive into shortly.

**12)**

Several screen shots of our data cleaning process. In the middle are the initial column names we had to decipher with the column descriptions that came with our dataset.

**ERD**

**14)**

This is our entity relationship diagram for our SQL tables showing the original cleaned data set and the 2 separate tables it was broken into with our numeric and categorical data.

**15)**

Our cleaned database was converted to a CSV which we then hosted on Amazon Web Services so the group could access the data. We decided to use Postgres and PGAdmin to manage the database and make our join using SQL.

**16)**

This shows a snip of our PGAdmin with our server and a few queries creating the tables.

I will now turn it over to Hakeem to discuss our machine learning process in more detail.