The three datasets used for my final project are a CSV file of vehicle engine and emissions data, a website of 2019 vehicle sales, and an API of vehicle dimensions. I had to perform three overall steps: 1) Convert each dataset into a data frame, 2) Store the frames in an SQL database and merge them into one SQL table, and 3) Generate at least 5 visualizations using data from the merged table.

For the CSV file, I imported the file using Pandas and removed all categorical variables besides make and model, as they are irrelevant for me. I then performed the following data cleaning steps:

1. Capitalizing make/model names.
2. Removing hybrid cars since they are outliers.
3. Aggregating duplicate cars by grouping make and model then calculating the means. I had to learn how to use the *groupby()* function for this.
4. Dropping the trim names from the model names to make searching easier.
5. Removing the make name from Mazda car models to avoid confusion.

For the website, I learned how to use the BeautifulSoup library to import a website as HTML and parse through it to get my data. I then converted the data to a data frame and applied the following cleaning steps:

1. Capitalizing make/model names.
2. Renaming the columns for ease of comprehension.
3. Splitting the car name column into two columns for make and model.
4. Dropping the trim names from the model names, as in the CSV file.
5. Removing hybrid-electric cars since they are outliers.

For the API, I wrote a function that will take a car’s make/model as input and output its dimension data from the NHTSA API. I knew how to use the requests library and try-except statements, but I had to learn how to integrate the function into a loop that would iterate over make/model names taken from the CSV file. Once the data was obtained and converted to a data frame, I could apply the data cleaning:

1. Removing all null data instances.
2. Dropping the trim names from the model names as before.
3. Removing the make name from Mazda car models to avoid confusion as before.
4. Dropping duplicate rows from the data frame.
5. Removing numbers from a Lexus model.

With all three frames cleaned, I imported all of them into one file. I learned how to create a new SQL database, insert the frames into the database and merge all three into one table within that database. I did some extra cleaning and fixed typos before the merge to prevent data loss in the join. I then generated 7 visualizations:

1. Scatter plot of fuel economy (CSV file) vs wheelbase (API)
2. Line chart of emissions (CSV file) vs vehicle curb weight (API)
3. Bar chart of car sales (website) for each make (CSV file)
4. Same bar chart as a pie chart
5. Boxplots of emissions for each make
6. Histogram of emissions
7. Density plot of emissions