Inventado, Charles Fredric G.
Rodelas, John Vincent B.
Valles, James Vincent V.
4CSE - Tres Sigmas

Reflection Report

- **Which cleaning step had the biggest effect? What technique did you use for removing null values?**

  The cleaning step for replacing string Hyphen to 0. (From hyphen, to null, to 0) having the correct data type is necessary to do calculations and aggregates.

  For removing null values : Constant Value Imputation: replacing the null values with 0.

- **How did outlier removal affect the dataset? What data handling technique did you use?**

  Our member initially planned to remove the outliers but unfortunately this affected the dataset in a bad way since those outliers were actually necessary in our study since the ones that got removed were brands of cars with actual high values in some of the numerical categories it was under and in our case we wanted values such as speed, gas/electrical use, horsepower, etc. So in our case removing the outliers becomes detrimental for our study because we will lose necessary outliers for our study. If you are still interested in the technique we used to handle removing outliers in our initial plan, the member decided to remove values going far below 5th percentile and far above 95th percentile.

- **Why are encoding and scaling necessary?**

  Encoding is necessary, primarily when categorical variables exist in the chosen dataset, because most statistical models and algorithms can only process numerical values or data. Some of the columns in our dataset, namely the "Manufacturer", "Regulatory Class", and "Vehicle Type", are categorical variables. These variables are essential, but they're in the form of text instead of numbers. So, it is necessary to translate them into a quantitative format so the models can process them. By encoding, we can maintain the integrity of the data and enhance the depth and accuracy of our potential insights.

  On the other hand, scaling addresses the issue of magnitudes that vary among different variables. Specifically speaking, in our dataset, for instance, the Weight (lbs) is measured in thousands, while the Horsepower (HP) is measured in the hundreds. Currently, there are several algorithms, especially those that rely on measuring distances between points, that can be unfairly affected by variables with larger scales. A model might not correctly assume the exact value of variables due to the difference in their numerical values, specifically in decimal places. Scaling is needed to be performed so that all numerical values have equal footing during processing.

- **How can your flow be reused?**

Naturally, we can repeatedly use and run the Tableau Prep flow that we've built for data with similar structure such as the dataset we've been using. Our flow is designed to handle data from datasets containing similar variables involving real world fuel economy CO_2 emissions, vehicle attributes from different manufacturers and their general technological advancements. Given the setup of the flow that we created, we just need to refresh the input connection and run to flow to automate the necessary cleaning process needed. We can use our flow again as a template when it comes to data cleaning, so that we will be able to maintain consistency in the quality of data in our dataset.

If we opt to use different columns that weren't focused on the activity, we can reuse the flow for the normalization process again, and would be quicker, honestly the joining process took us time to understand so that would be a great sign of relief to not do these again. Join keys 😭, but that is a great lesson for use.

Deliverables:

1.).tflx flow file

2.) clean dataset (CSV)

3.) reflection report