DSC 540

Inman, Gracie

Final Project

11/18/23

Reflection

Throughout this project, I have had to learn numerous skills. One of the most important things I learned throughout this project is debugging code. While working on the project, I had numerous issues with code not working and having to figure out why. Thanks to this project I have gotten a lot better at understanding the error outputs in Jupyter Notebook which allows me to more efficiently fix my code. Also, due to having multiple different types of data in this project, I feel more confident in obtaining data from sources other than a pre-made dataset. Merging the datasets required a lot of cleaning and preparation. During this project, I had to continually look up different methods for cleaning and fixing data issues. If one wasn't working and I couldn't figure out why, I would try another. This really showed me how many ways can be used to complete a single task. Each dataset was individually cleaned before being put together. After being put together, I had to further fix type issues and other things that I had not thought about previously to get my dataset to merge seamlessly.

There are many ethical implications throughout my project. One implication is my decision to drop entries from the dataset that were not complete. This removes data points from the analysis and could skew results. A lot of unused columns were dropped from the dataset as well which could lead to an inaccurate representation of the dataset. There were many outliers within the cars data frame. These were assumed to be from luxury cars and were kept. However, I should have done an analysis with and without the outliers to show their effect on the data. This assumption could have skewed the results and lead to an inaccurate representation of the dataset. Also, 57.2% of the data belonged to the price category 20 to 40 thousand. A large part of the dataset being from one category can lease to bias. Different car brands also cost different prices and that should have been considered in analysis. After completing this milestone and performing visualizations, I realized that a large amount of the dataset was from one state. This causes bias when considering temperature as a deciding factor. Most of the cars sold in the data set were from Texas which a notoriously hot state. This causes bias and an inaccurate analysis when considering the relationship between temperature and other factors. This shows that the dataset is biased and the original goal of relating temperature to price and other factors would lead to inaccurate representations and would be unethical until the data set it transformed to eliminate bias or a new dataset was found with less of a bias.