**IS 497 Project Proposal – Talal Hashmi**

I have a huge passion for cars. Luckily, I also have a huge passion for data science. For my final project, I will be attempting to import and clean a large scale data set full of cars. I have been searching the web trying to find the perfect dataset with wholesome values that are relevant to my interests. This includes every specification possible for a road car including:

* Dimensions (height, width, length)
* Year/edition
* Brand/model
* Engine/transmission types
* Fuel economy (both city, highway)
* Horsepower + torque

I was able to find the dataset below on Kaggle thanks to “Jose Diniz” publishing in 2018 (<https://www.kaggle.com/datasets/dinizjr/cars2csv>). It has 5076 observations of different cars and 18 columns holding unique descriptions and statistics about each make and model. See below:

A screenshot of a computer

Description automatically generated with medium confidence

The reason I chose this specific dataset is because it is very large, I can get accurate depictions of the data I am looking for. I am planning on cleaning the file by confining and condensing the entire dataframe. There are too many columns and rows so first I want to remove any null/missing or duplicate rows because after looking at the dataset I noticed there are many cases where the same car, model, and year is repeated more than once. Then I want to remove columns that do not provide me with value engine.id or transmission.info because I cannot visualize them. I wil rename the column headers as the names are messy and lowercase. I wish to mainly use horsepower and torque data to explore which cars and brands produce the best performing engines in their categories and what cars are ultimately the fastest in scope of the dataset. I want to showcase some data transformation and manipulation skills learned from class so restructuring the dataset and perhaps adding a column of my own might just be it. I did realize my dataset does not include a category dedicated to the nation’s that the brand/make originates from so I might map out brands to certain manufacturing nations (Germany, Italy, Korea, for example). With this I can visualize which nations manufacture the quickest and slowest cars organizing them by engine type and drivetrain transmissions. Also highlighting certain distributions that share more knowledge of this large dataset will be added to my jupyter notebook scripts. As for project organization, I plan on including the following:

Diagram

Description automatically generated