**Project Proposal**

For our project, we decided to compare new and used cars data from <http://www.cars.com>. In particular, we looked at all Toyota vehicles. We chose to scrape the site for data on the price, mileage, make, model, and year of the vehicles. In addition, we also looked at the dealership and the distance at which the shops were located. For our second dataset, we collected consumer ratings data from Kaggle (<https://www.kaggle.com/ankkur13/edmundsconsumer-car-ratings-and-reviews?select=Scrapped_Car_Reviews_Toyota.csv>.) This dataset consisted of a list of all costumer reviews and ratings for Toyota vehicles. We found that this data set had a wide collection of customer reviews for almost every Toyota vehicle from 1997 to 2018. After extracting and cleaning the data, we loaded the data into Postgres, a relational database.

**Project Report**

**Extract**

Within Jupyter notebook, we imported and setup Splinter in order to access our site. From there, we used BeautifulSoup to extract and save the data in a data frame. For our ratings data, we read the csv file into a data frame in Jupyter. We separated the Vehicle\_Title column into its various components and created a new column titled Vehicle\_Name that consisted of just the year, make, and model of the Toyota vehicle. We then used the Vehicle\_Name to group the ratings and calculate the average ratings for each vehicle.

**Transform**

To start, we removed all duplicate and null rows from both of the datasets. We renamed the column headers to remove all spaces and replace them with an underscore, and we also removed the dollar symbol ($) from the Price column. In the cars dataset, we adjusted the values in the Price and Distance from 30332 columns in order to set these as numeric or integer values. We often had to remove extraneous symbols and letters, including the dollar symbol ($), in order to obtain our results. In addition, we also created separate columns for the year, make, and model of the vehicle, and we concatenated the three into a column title Vehicle\_Name. Lastly, we determined that the Vehicle\_Name would be the primary key for both datasets and the column at which we would merge the data.

**Load**

Finally, we loaded our data into Postgres, a relational database. We joined the datasets on the column Vehicle\_Name and chose to add all columns into our joined dataset. This dataset was very helpful for us to see the price differences between new and used vehicles. We were also able to compare the differences to the vehicle ratings to see how much customer opinions influenced price. The techniques by which vehicle are priced is truly interesting, and we are curious to explore more with this dataset.