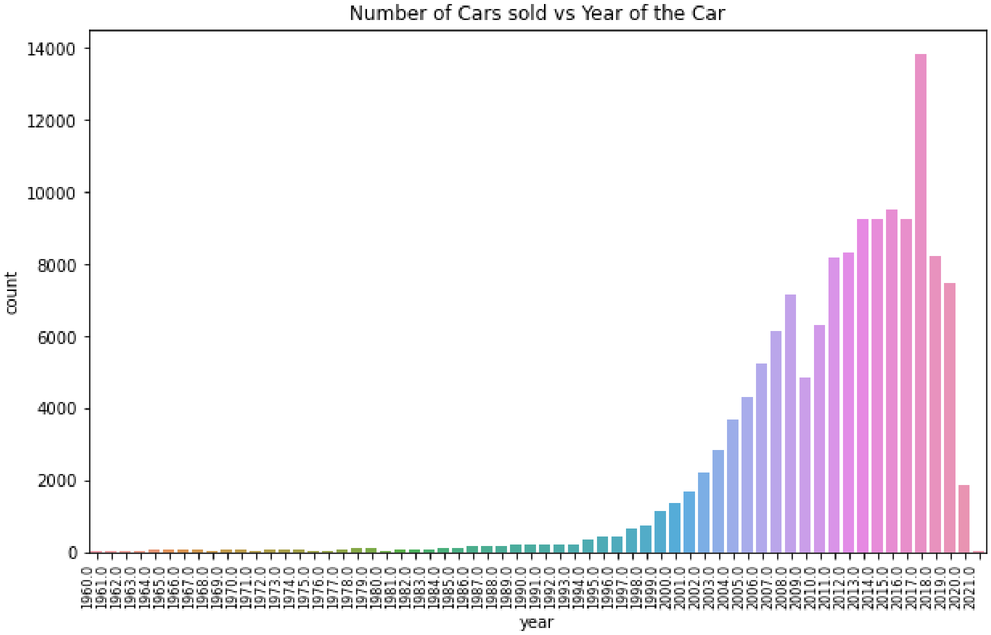
Eric Gleiter
Anyesha Ray
Final Project Part Two
11/10/2020

Our dataset is a premade dataset that is available from the website kaggle (https://www.kaggle.com/austinreese/craigslist-carstrucks-data?select=vehicles.csv). The data is gathered via a web scraper that pulls information from craiglist every few months. Then the updated data is published to kaggle.

In order to clean our data we first dropped unnecessary data, such as the url of the craigslist post (which is often outdated), the latitude and longitude of the car, (as we have access to location via county and state), and other columns relating to the craigslist post (image_url, region_url, etc…). Then we began to look deeper into the data. We noticed that the vast majority of car sales were from the year 1960 and onward (see plot below). Additionally we reasoned that as the years go by older cars are less and less likely to be for sale. Thus we decided to limit the tail end of the data set to the year 1960.


Number of Cars sold vs Year of the Car

Next we looked at other variables to identify similar issues. We noticed that a large chunk of data had a price of 0. Upon further investigation we discovered that this was because the seller was looking for the “best offer” and did not set a price. Due to this we removed any row which had a zero for price as we have no way of knowing the true sale price for that vehicle. Similarly we had issues with missing data many posts simply did not include some details such as county or state of the sale. In order to solve this problem we decided to remove any columns with more than 50% of the data missing. These columns were: price, year, manufacturer, model, fuel, odometer, title_status, transmission, drive, type, paint_color, and state. Our cleaned dataset has 12 variables and over 130,000 data points to work with. Outside of the stated anomalies of 0 prices, we have one potential problem. While we removed columns with an extremely large number of missing variables we still have to deal with missing values in the data set. When we perform feature selection we will have to be careful about how we deal with these variables. Please see the accompanying jupyter notebook to see the code for our data preparation and cleaning.