**Business Understanding**

Problem statement:

In this application, you will explore a dataset from Kaggle. The original dataset contained information on 3 million used cars. The provided dataset contains information on 426K cars to ensure speed of processing.  Your goal is to understand what factors make a car more or less expensive.  As a result of your analysis, you should provide clear recommendations to your client -- a used car dealership -- as to what consumers value in a used car.

**Data Understanding**

Reviewing the data we can find some issues with developing a model due to several factors I will summarize below:

* Some entries had a price of $0
* Data has units from 1905, but for most dealers that isn’t useful. I chose a cutoff of 1950
* Some units weren’t fully operational or were labeled missing
* The VIN number would not be helpful
* Region was removed and state was used to simplify the investigation
* Model was removed and used in favor of the manufacturer

As a first pass investigation I decided to keep the model to a simpler level but keep as many unique top-level categories as possible.

**Data Preparation**

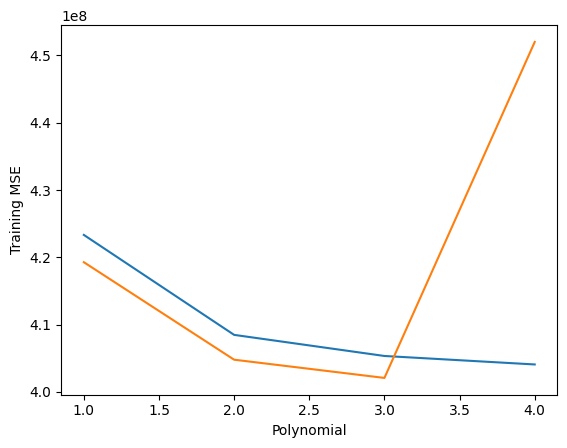
I removed data that was problematic given the issues above then mapped the categorical data into Boolean (0, 1) so that a model could be created for it.

The categorical and continuous data was separated to see the impact they have on the data independently and this resulted in two end models.

**Modeling**

The two models were developed, and both give some intuitive insight on the price of used vehicles and what matters most.

Treating year as a numerical value I found that the optimum number of polynomials to use was 3 to keep the validation data set MSE lower.



**Evaluation**

**A graph with text on it

Description automatically generatedA graph with different colored squares

Description automatically generatedA graph with different colored bars

Description automatically generatedA graph with different colored squares

Description automatically generated**Below are some highlights for the plot

**A graph with different colored bars

Description automatically generatedA graph with different colored squares

Description automatically generatedA graph of a number of different colored bars

Description automatically generated with medium confidence**

**Deployment**

Customers are most interested in the following types of cars:

* Excellent/Good/Like New Condition
* 4/6/8 cylinders
* Automatic/Manual Transmission
* SUV/Sedan/Truck Type
* Alternative/Electric Fuel type
* Rear Wheel Drive
* Luxury Car brands
* Newer models
* Lower milage

Next Steps

* Investigate region specific machine learning models
* Investigate car model specific machine learning models