The overall results collected from obtaining information on 426K cars, showed that the greater the year of the car, manufacturer such as mercedes benz, and new condition all displayed factors that make a car more expensive. The CRISP-DM framework guided the analysis and modeling of car price data. The project began with defining the goal: predicting car prices based on factors like odometer readings, drivetrain, and manufacturer. Data exploration revealed missing values, formatting issues, and categorical variables, which were cleaned and prepared using techniques like encoding and handling NaNs. Linear Regression and Random Forest models were developed, with hyperparameters optimized using GridSearchCV and RandomizedSearchCV, and validated through cross-validation. Evaluation focused on R^2 and error metrics to ensure model reliability. The final model can assist in pricing strategies and may be deployed as a tool for buyers or sellers to predict car values.

In depth, after importing the file, making gento be equal to reading the data for data cleaning. I saw a lot of NaN within the data and noticed that columns such as year and odometer had this weird format with .0 behind each value. I then dropna() and noticed that the code still didn't get rid of all of them so I chose to coerce the data for the year column. After some research I was able to remove .0 behind every value by converting all data within the column to an integer then back to a string. Doing this twice to fix the odometer column. After gento was cleaned I used a linear regression model and fit it between the year and price of a car. Found the prediction is correct that greater the year of the car correlated to a more expensive car. To give a clear recommendation to a used car dealership, I wanted to know which manufacturers resulted to a higher price point. Using the group by manufacturer and price, I see that the chart displays Mercedes Benz, Volvo, toyota, jeep, and chevrolet as the top performers for average price. These manufacturers show the best cars to have in your dealership as they are the most expensive. In this data, it also shows to avoid manufacturers such as mercury, saturn, land rover, and pontiac because the average price is very low. Another process for Linear regression model, training the data for a correlation between odometer and year with price. I am able to see the y=mx+b formula. Comparing the actual and predicted data we can see the MSE is very high which means the model isn't too good. Plotting the residual plot vs. predicted prices. Using this model, I compared both actual and predicted price and found the model needs more tweaking because it isn't as close as I want it to be. Moreover, using one hot coding for the condition of the car. Plotting this data using actual vs. predicted prices and even residual plot using the condition we see the data shows a bad correlation. Lastly, The form of Gridsearch using random forest regressor, we found the best model out of 50 fits. To conclude, out of all this data, research of manufacturers such as mercedes benz, a decrease in odometer, new condition and greater year model all lead to a more expensive vehicle.