# Datamining Assessment 1 Report

In this project we made use of the large data set (and split it manually) after testing the code on the smaller sets. Initially, we explored the structure of the data and noted several attributes with large number of missing values (80% +) which would need to be dropped. Applying imputation would be inappropriate. Moreover, several attributes are irrelevant, such as the URL links which were removed. We also noted the outliers lying outside of mainland America, but did not drop these. This gave us an idea what to expect when cleaning the data, especially the heavily skewed features. The next step was to tidy and format the data appropriately. We therefore only included a handful of attributes (24 in total out of the 66) to use to predict the attribute price, using the correlations as a guide. It is worth noting that not all attributes could be displayed due to the fact they were categorical and needed to be formatted.

The process of cleaning using functions was straightforward, however, prior, we attempted to use pipelines which worked, but the code was incredibly messy. After dropping another attribute (engine_cylinders) we were left with a clean and scaled (not applied to the price attribute) data frame which we could now investigate using three models: linear regression, decision trees and random forests, and look at their Root Mean Square Errors (RMSEs) and their Mean Square Error (MSE).

Fitting a linear regression model and then testing it on the training data, we observed one negative price prediction (unrealistic) and then when looking at the associated RMSE error noted that the predictions were off by around 15772$. A decision tree model produced a slightly worse measure of error of roughly 15918$, whereas the random forest model yielded a RMSE of 4429$. It appears the last model might be the best. However, point estimates should be supported with confidence intervals. Thus, cross validation was performed on each of the three models with the following results and using folds =10.

| **Model used** | **RMSE** | **Mean** | **Standard deviation** | **MSE** |
|---|---|---|---|---|
| **Linear Regression** | 15772$ | 15232$ | 4145$ | 6726$ |
| **Decision Trees** | 15918$ | 15416$ | 4055$ | 7211$ |
| **Random Forests** | 4429$ | 10631$ | 5193$ | 1047$ |

Therefore, we choose a random forest model to test on our testing data. Applying the same data wrangling as with the training data, the final prediction errors are: (9737$, 16840$) with a RMSE of roughly 13755.

Overall, we have fitted three types of models: a linear regression model, decision trees and random forests. We, moreover, performed cross-validation on each of the models. In the end we can conclude that a random forest model fitted to our data, minimizes the RMSE more than the previous two. However, the error remains quite large with a 95% confidence interval which, when it comes to cars prices, is a considerable amount.