Project Name: BMW Sales Analysis

- Franklin Doane
- David Slate Lee
- Shaye Shankar
- Will Ward
- AJ Wood

1. Problem Statement and Background (15%)

We will develop a model that will determine the price of a vehicle by the following features: model, year, transmission, mileage, tax, fuelType, mpg, engineSize. The data comes from European BMW sales and contains 10,781 observational units. We hope to develop a model that can explain 85% or more of the variation in price and accurately predict a price value within 10% of the true value. This data is useful for a seller who wants to determine what price to set on a vehicle, or a buyer who wants to compare a listing to historical values. This information could limit future overpricing. For our current data set, we are unable to determine exact location or recent or historical biases that could affect the data. If we had more background knowledge we could take steps to mitigate the effects of any unknown issues present in the data.

2. Data and Exploratory Analysis (15%)

Our features are car model, mpg, mileage, year, engine size, transmission, tax, price, fuel type. We noted that 47 cars have an engine size of 0, 45 of which are not electric. We would likely remove these instances. On the flip side, there is 1 electric vehicle with an engine size of 1, which we will set to 0. For mpg, there are many nonsensical outliers in the electrical, hybrid, and other fuel type categories, leading us to think we should remove that feature for all non-diesel and non-gasoline vehicles by imputing it with an NA or 0 value. We didn’t find any other messiness in our data, but we created many charts and notes on the dataset in our attached markdown file. The dataset can be found here.

3. Methods (10%)

[Describe the methods you are planning on exploring (usually algorithms, or data cleaning or data wrangling approaches). Justify your methods in terms of the problem statement. What did you consider but *not* use? In particular, be sure to include every method you tried, even if it eventually does not "work". When describing methods that didn't work, make clear how they failed and any evaluation metrics you used to decide so.]

4. Tools (10%)

[Describe the tools that you used and the reasons for their choice. Justify them in terms of the problem itself and the methods you want to use. Tools will probably include machine learning, and possibly data wrangling and visualization. Please discuss all of them. How did you employ them? What features worked well and what didn't? What could be improved? Describe any tools that you tried and ended up not using. What was the problem?]

5. Results (35%)

[Give a detailed summary of the results of your work. Here is where you specify the exact performance measures you used. Usually there will be some kind of accuracy or quality measure. There may also be a performance (runtime or throughput) measure. Please use visualizations whenever possible. Include links to interactive visualizations if you built them. You should attempt to evaluate a primary model and in addition a "baseline" model. The baseline is typically the simplest model that's applicable to that data problem, e.g. Naive Bayes for classification, or K-means on raw feature data for clustering. If there isn't a plausible automatic baseline model, you can e.g. compare with human performance by having someone hand-solve your problem on a small subset of data. You won't expect to achieve this level of performance, but it establishes a scale by which to measure your project's performance. Compare the performance of your baseline model and primary model and explain the differences. Note: everyone on your Team should code/test/document results from at least one model.]

6. Summary and Conclusions (10%)

[In this section give a high-level summary of your results. If the reader only reads one section of the report, this one should be it, and it should be self-contained. You can refer back to the "Results" section for elaborations. This section should be less than a page. In particular, emphasize any results that were surprising. Include lessons learned and any potential future work.]

7. Appendix (5%)

GitHub repository (link)

Kaggle set (link)