**Requirements (delete prior to final submission)**

* No more than 7 written pages (10 pages for appendix) (Max total = 17)
* single-spaced to doubled spaced
* 11-point font
* 1-inch margins
* page limitation includes graphics and tables
* Each graphic or table should be clearly labeled and discussed in the text.
* You may include more tables and figures in an appendix (no more than 10 pages)
* R code should also be included in the appendix
* Any table or figure that is included in the appendix must be referenced in the text.
* *build regression models*
* *identify key relationships and interpreting those relationships*
* *Address the importance of the “Popularity” variable*
* *How much general popularity can play a role in the retail price of a vehicle*
* *Provide detailed information on summary statistics, EDA, and your model building process*
* *Provide interpretation of the regression coefficients of your final model including hypothesis testing, interpretation of regression coefficients, and confidence intervals.*
* *Mention the Practical vs Statistical significance of the predictors.*
* *Answer any additional questions using your model that you deem are relevant.*
* *Use training data set for EDA and model fitting*
* *Use test set to compare models to make a final call*
* *Compare multiple models with the goal of developing a model that can predict the best and do well on future data*
* *Use training and test set to build at least one additional multiple linear regression model with more complexity than the interpretable model*
* *Use the ISLR text book (and internet) to identify one nonparametric technique to build a regression model*
* *Select from k-nearest neighbors’ regression or regression trees.*
* *Total you should have at least 3 models, 2 linear regression models and 1 nonparametric.*
* *For each of the three models, provide measures of fit for comparisons: test ASE and validation ASE are mandatory.*
* *You may also include additional metrics for completeness like R squared/Adjusted R squared, AIC, and BIC where applicable (only point where the validation set is being used).*
* *Providing additional insight as to why one model is better than the other, or why they are all the same is encourage.*
* After 2000 the MRSP has massive shift – explain it and use data, or split into 2 datasets, or drop/delete one of the two sides of data
* For each model interpret the MEANING of at least 1 numerical and 1 categorical variable. Show table of ALL variable results.
* Make sure to discuss the PROCESS and mindset of the EDA, not just the graphs
* 3-7 pages of WRITTEN text is sweet spot
* Start with “shotgun” graphs to look for correlation but then focus down into particular variables and display larger (more comprehensive) graphics

**Analysis of the Automotive Industry**

**Authors:** Nicole Assenza, Andrew Taylor and Ranjan Karki

In 2021, the worldwide automotive industry had an estimated value of almost 3 trillion dollars. More than 60 international manufacturers compete for their share of the market which sells roughly 76 million cars every year. As the amount of relevant data surrounding the industry continues to grow at an exponential rate, these companies are continually looking for reliable trends that yield even the slightest competitive advantage. This comprehensive report is intended to aid one such company in analysis and interpretation of the available data through multiple regression and nonparametric modeling.

The data set of interest includes 11,914 records of discrete automobiles in production between 1990 and 2017. Each record hosts data points for 16 variables (9 numerical and 7 categorical) such as sale price (MSRP), Engine HP, Manufacturer, and “popularity score” to name a few. The full list of variables and their descriptions can be seen below.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| MSRP | Numeric | Sale Price (Response variable) |
| Car Make | Factor | The company that made the car. Ex: Honda, Toyota, etc. |
| Car Model | Factor | The model of the car. Ex: 4Runner, Accord, etc. |
| Year | Numeric | Year the car was produced |
| Engine Fuel Type | Factor | Type of fuel the car accepts. Ex: Regular unleaded, Diesel, etc. |
| Engine HP | Numeric | Horsepower of the car’s engine. |
| Engine Cylinders | Factor | Number of cylinders in the car’s engine. |
| Transmission Type | Factor | Type of transmission in the car. Ex: manual, automatic, etc. |
| Driven Wheels | Numeric | Wheels powered by engine. Ex: Front Wheel, Rear Wheel, etc. |
| Number of Doors | Numeric | The number of doors that the car has. Ex: 2, 4, etc. |
| Market Category | Factor | Various special factors for each car. Ex: Exotic, Luxury, etc. |
| Vehicle Size | Factor | The size of the vehicle. Ex: Midsize, Large, Compact, etc. |
| Vehicle Style | Factor | Body type of the vehicle. Ex: Coupe, Convertible, etc. |
| Highway MPG | Numeric | Fuel efficiency on the highway in MPG |
| City MPG | Numeric | Fuel efficiency in the city in MPG |
| Popularity | Numeric | A numerical popularity score for each car. |
| Crossover | Boolean |  |
| Diesel | Boolean |  |
| Exotic | Boolean |  |
| Factory Tuner | Boolean |  |
| Flex Fuel | Boolean |  |
| Hatchback | Boolean |  |
| Hybrid | Boolean |  |
| Luxury | Boolean |  |
| Performance | Boolean |  |

**Appendix**