The Data

Initial summary statistics for the raw data revealed there are 11,914 observations (cars) across 16 variables with some anomalies (Summary in appendix 1.1) such as a maximum highway miles per gallon of 354, missing data and engines with no cylinders:

* 16 Total variables
* 1 Response variable
* 4 Explanatory variables
* 11 Categorical variables

On the initial pass we ran three individual tests for missing values so that we could utilize some built in features in the software that look for NA as well as known string variants such as, “NA,” “na ,” “n/a,” and number variants such as -9, -99 and 88 (see full list in appendix 2.variants).

The final combination showed potential missing values:

* Engine.Fuel.Type = 3
* Market.Category = 3,742
* Engine.HP = 69
* Engine.Cylinders = 30

A subsequent pass also showed there were nineteen missing values for Transmission.Type that were initially missing because they were labeled as, “Unknown.”

The first pass data cleanse resulted in the following variables:

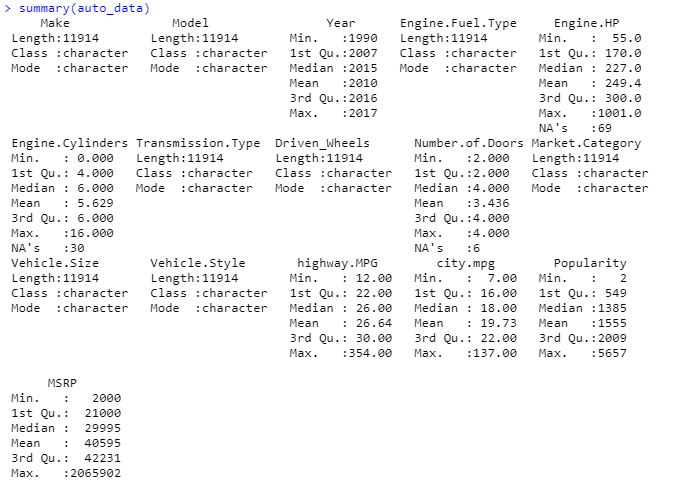


Popularity comment: Each Make has a unique Popularity Score regardless of model or any other attributes, so in this light we can ignore the variable and rely on the Make for easier model interpretation.

Appendix

Data

1.1 – Summary of Raw Data



2.Variants

