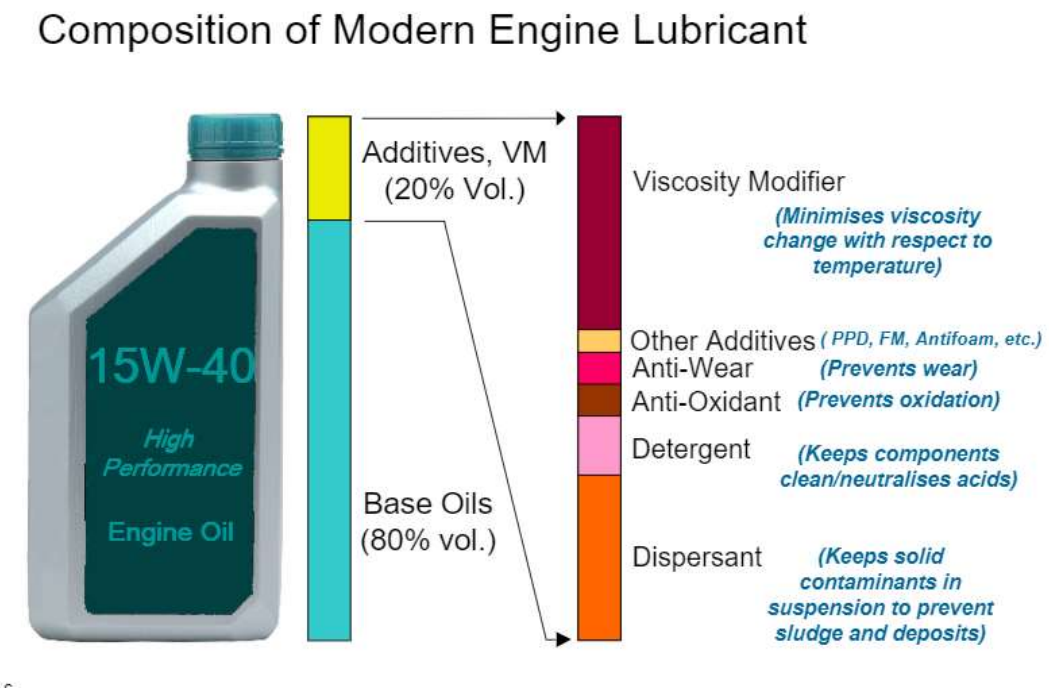
Notes/help with the data:

The data is from a bench test that model the propensity for an engine oil to protect against rust. 100 is a completely clean (no rust) – this is as good a response as you can get. A value greater than 50 is a pass.

Lubrizol make the additive package and viscosity modifiers (VMs) that go into an engine oil. The chemical properties of the engine oil are the predictor variables in the data.



Unfortunately, we had to relabel/annonomise our data. You will see variable name start with "ingredient", and variable name start with "Group". The group variables represent properties for a type of chemistry that goes into the additive package or represent the base oil. For example, the properties of the dispersants in the engine oil will be represented by a single group of variables.

There are many potential ingredients that can be used within a chemistry group. However, each ingredient will only relate to a single group of variables. For example, ingredient_43 will only influence (be correlated to) some of the Group10 variables. How the ingredients relate to the group variables is determined by chemical structure/properties of that ingredient. You might see that the group variables are correlated to each other.

The only group that will not have any associated ingredients is Group13.

We prefer to model with the group variables. If a new ingredient came along in the future assuming we could relate the ingredient to the group variables, we could still predict how that ingredient would perform in our model without having to test it. However, sometimes there is no good group variable that picks up the property of the ingredient, so we end up putting an ingredient variable in the model.