Capstone Project Proposal

**Mentor**: Andrew Maguire

**Mentee**: Sriprabha Kasturirangan

Dataset - This data comes from the [Open Food Facts database](http://world.openfoodfacts.org/data), a free, open, and collaborative database of food products around the world.

Exploratory analysis –

1. Plot French and UK Nutritional Score Values / Coverage of Nutritional Score
2. Identify how many Unique records / Missing records for important columns
3. Identify and Plot Top 10 Countries represented in dataset
4. How many products have allergens
5. How many countries publish allergens data
6. Top 10 Countries in dataset consuming Alcohol
7. Identify General Categories of food items
8. How many products have additives / Average no of Additives per Country
9. Identify top Enumbers / Additive used in the Dataset
10. Enumbers which are dangerous , cancerous
11. Top allergens in the food items
12. For top 10 Countries ,plot or identify the top 10 categories of food.

Target Columns to be used:

* [categories\_en] or [main\_category] - predict category from ingredients.
* [ingredients\_text] - model or market basket analysis to see what ingredients tend to co occur.
* [additives\_n] - predict number of additives.
* [additives] - predict presence of specific additives.
* [nutrition\_grade\_fr] - predict nutrition grade of foods.
* [energy\_100g] - predict energy per 100g.
* [fat\_100g] - predict fat per 100g.
* [saturated-fat\_100g]
* [carbohydrates\_100g]
* [sugars\_100g]
* [proteins\_100g]
* [salt\_100g]

Preditive Model –

Approach

* Collect all the data - this will require significant wrangling.
* Scope the problem down to make it both more tractable, as well as relevant.
* Model it as a regression problem to predict the additive number . Train the model on an earlier subset of a data set and test it on later subsets.
* Evaluate effectiveness of the model.