

This notebook will go through the processed csv and perform exploratory data analysis to find any issues that need to be fixed before model creation

```
In [2]: #import Libraries
import numpy as np
import pandas as pd
import plotly.express as px
import plotly.graph_objects as go
from IPython.display import display
import sweetviz as sv
from autoviz.AutoViz_Class import AutoViz_Class
AV = AutoViz_Class()

Imported AutoViz_Class version: 0.0.81. Call using:
    from autoviz.AutoViz_Class import AutoViz_Class
    AV = AutoViz_Class()
    AV.AutoViz(filename, sep=',', depVar='', dfte=None, header=0, verbose=0,
               lowess=False, chart_format='svg', max_rows_analyzed
               =150000, max_cols_analyzed=30)
Note: verbose=0 or 1 generates charts and displays them in your local Jupyter
notebook.
    verbose=2 saves plots in your local machine under AutoViz_Plots directo
ry and does not display charts.
```

Data Dictionary

- FoodID : Unique Identifier for the food (numerical)
- FoodDescription : Name and contents of the food (text)
- FoodGroup : 23 Groups of food (categorical)
- PROTValue : Value of Protein Nutreint in the food in g/100g
- FATValue : Value of Total Fat in the food in g/100g
- CARBValue : Value of Total Carbohydrate in the food in g/100g
- STARValue : Value of Starch in Carbohydrates in g/100g
- TSUGValue : Value of Sugar in Carbohydrates in g/100g
- TDFValue : Value of Dietary Fibre in Carbohydrates in g/100g
- TSATValue : Value of Saturated Fat in Fats in g/100g
- MUFAValue : Value of Monounsaturated Fat in Fats in g/100g
- PUFAValue : Value of Polyunsaturated Fat in Fats in g/100g

Important Information

- Get the user's age, height in cm, weight in kg and gender for user profile building
- Using Harris-Benedict Equation (Needs.pdf) for Basal Energy Expenditure (BEE), find the calorie intake required per day
- Requirement of Protein in grams = weight in kg
- Calorie of Protein = 4 * Protein in grams
- Find out the percentage of protein by dividing protein calorie by total calorie
- 60% of Calories should be Carbohydrates with more TDF,STAR and less TSUG
- Calculate carbohydrates in grams using CalorieCount/4
- Leftover calories need to be Fat with more MUFA,PUFA and less TSAT
- Calculate fat in grams using CalorieCount/9
- Input will be the requirement of Carbohydrates,Proteins,Fats calculated above
- Output should be Foods that meet the needs with carbohydrates favoring TDF and STAR and Fats favoring MUFA and PUFA
- TSAT should not be more than 10% of Fat intake
- TSUG should not be more than 10% of Carbohydrate intake

```
In [4]: food_data_df = pd.read_csv('FoodNutritionData.csv')
display(food_data_df.head(5))
```

	FoodID	FoodDescription	FoodGroup	PROTValue	FATValue	CARBValue	STARValue	TSUGValue
0	2	Cheese souffle	Mixed Dishes	9.54	15.70	5.91	0.00	;
1	4	Chop suey, with meat, canned	Mixed Dishes	4.07	2.80	5.29	0.00	;
2	5	Chinese dish, chow mein, chicken	Mixed Dishes	6.76	2.80	8.29	3.99	;
3	6	Corn fritter	Baked Products	8.55	21.24	38.62	0.00	;
4	7	Beef pot roast, with browned potatoes, peas and carrots	Mixed Dishes	21.29	5.25	10.72	0.00	;

```
In [5]: food_data_df.dtypes
```

```
Out[5]: FoodID           int64
FoodDescription    object
FoodGroup          object
PROTValue          float64
FATValue           float64
CARBValue          float64
STARValue          float64
TSUGValue          float64
TDFValue           float64
TSATValue          float64
MUFAValue          float64
PUFAValue          float64
dtype: object
```

```
In [6]: sv_analyzer = sv.analyze(food_data_df)
sv_analyzer.show_html()
```

Report SWEETVIZ_REPORT.html was generated! NOTEBOOK/COLAB USERS: the web browser MAY not pop up, regardless, the report IS saved in your notebook/colab files.

```
In [36]: food_data_df.describe(include='all')
```

```
Out[36]:
```

	FoodID	FoodDescription	FoodGroup	PROTValue	FATValue	CARBValue	S
count	5690.000000		5690	5690	5690.000000	5690.000000	5690.000000
unique		NaN	5689	23	NaN	NaN	NaN
top		NaN	Beans, hyacinth, raw	Vegetables and Vegetable Products	NaN	NaN	NaN
freq		NaN	2	785	NaN	NaN	NaN
mean	100722.967311		NaN	NaN	11.057056	9.954095	21.980529
std	197883.301111		NaN	NaN	10.843366	16.718251	26.537885
min	2.000000		NaN	NaN	0.000000	0.000000	0.000000
25%	2136.250000		NaN	NaN	2.110000	0.590000	0.500000
50%	3855.500000		NaN	NaN	7.625000	3.755000	10.265000
75%	5869.750000		NaN	NaN	18.757500	12.200000	32.075000
max	503380.000000		NaN	NaN	85.600000	100.000000	100.000000

Number of Rows and Columns in Database

```
In [37]: print('rows',food_data_df.shape, 'columns')
```

```
rows (5690, 12) columns
```

Total unique elements in each category

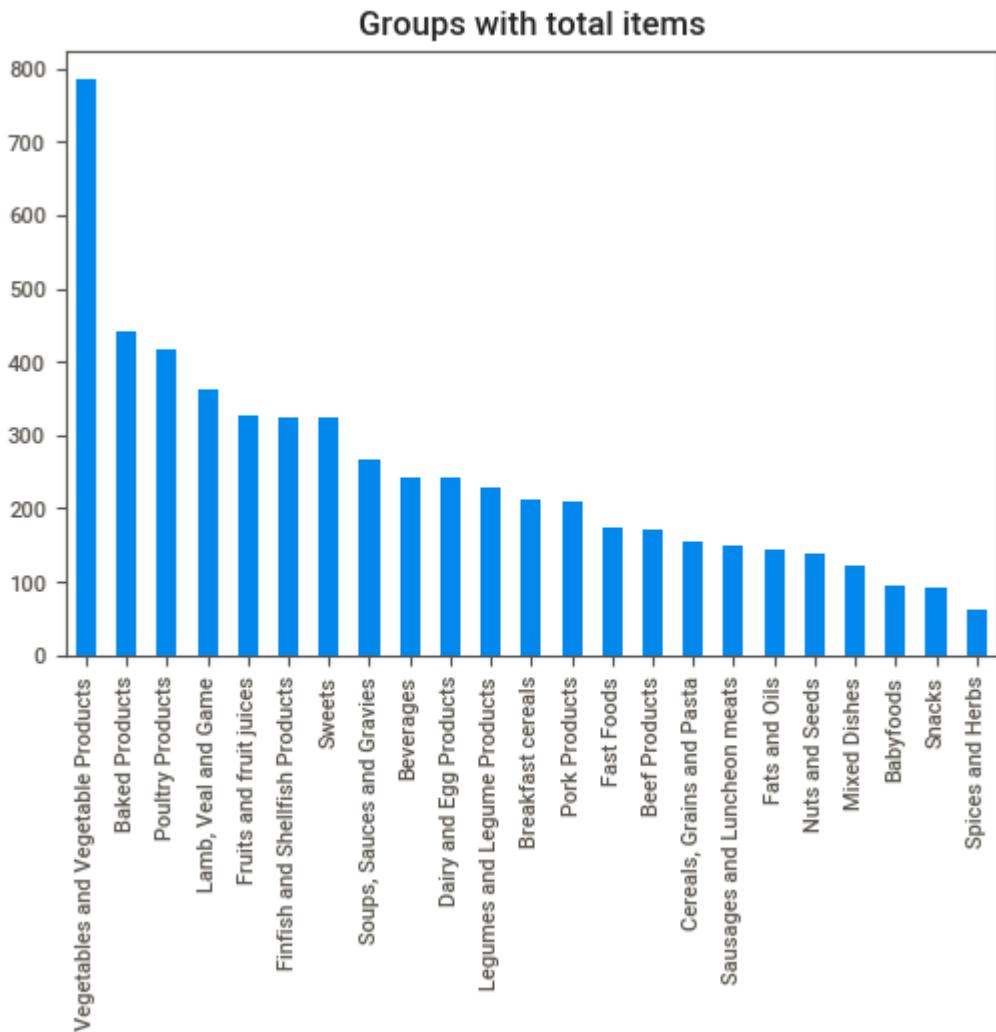
```
In [38]: print(food_data_df.nunique())
```

```
FoodID          5690  
FoodDescription 5689  
FoodGroup        23  
PROTValue       2261  
FATValue        1913  
CARBValue       2756  
STARValue        360  
TSUGValue       1505  
TDFValue         237  
TSATValue       2812  
MUFAValue       2880  
PUFAValue       2381  
dtype: int64
```

Total number of groups for different food along with number of contents in each category of group

```
In [39]: food_data_df['FoodGroup'].value_counts().plot.bar(title='Groups with total items')
```

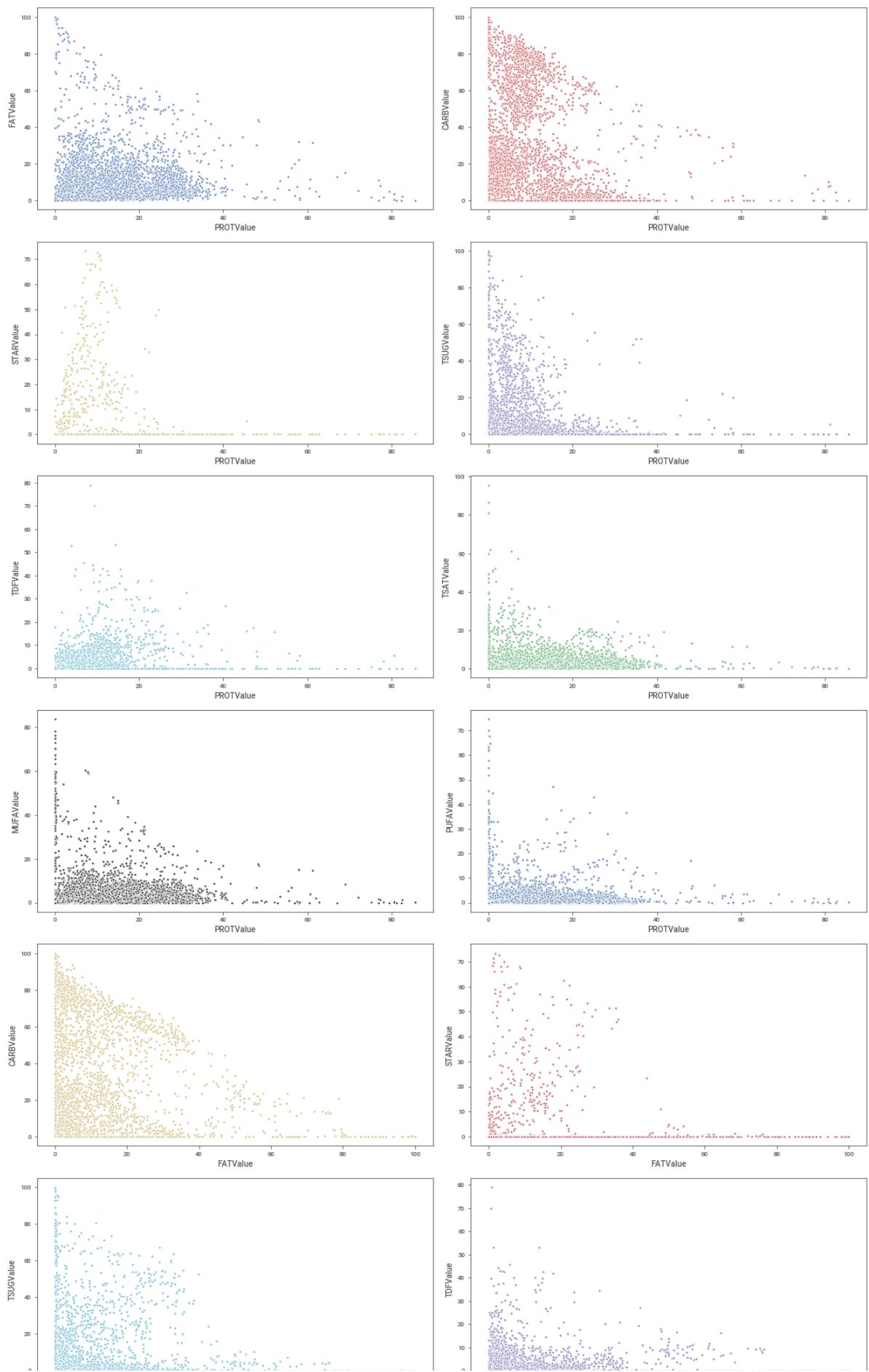
```
Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0x212c29384c0>
```

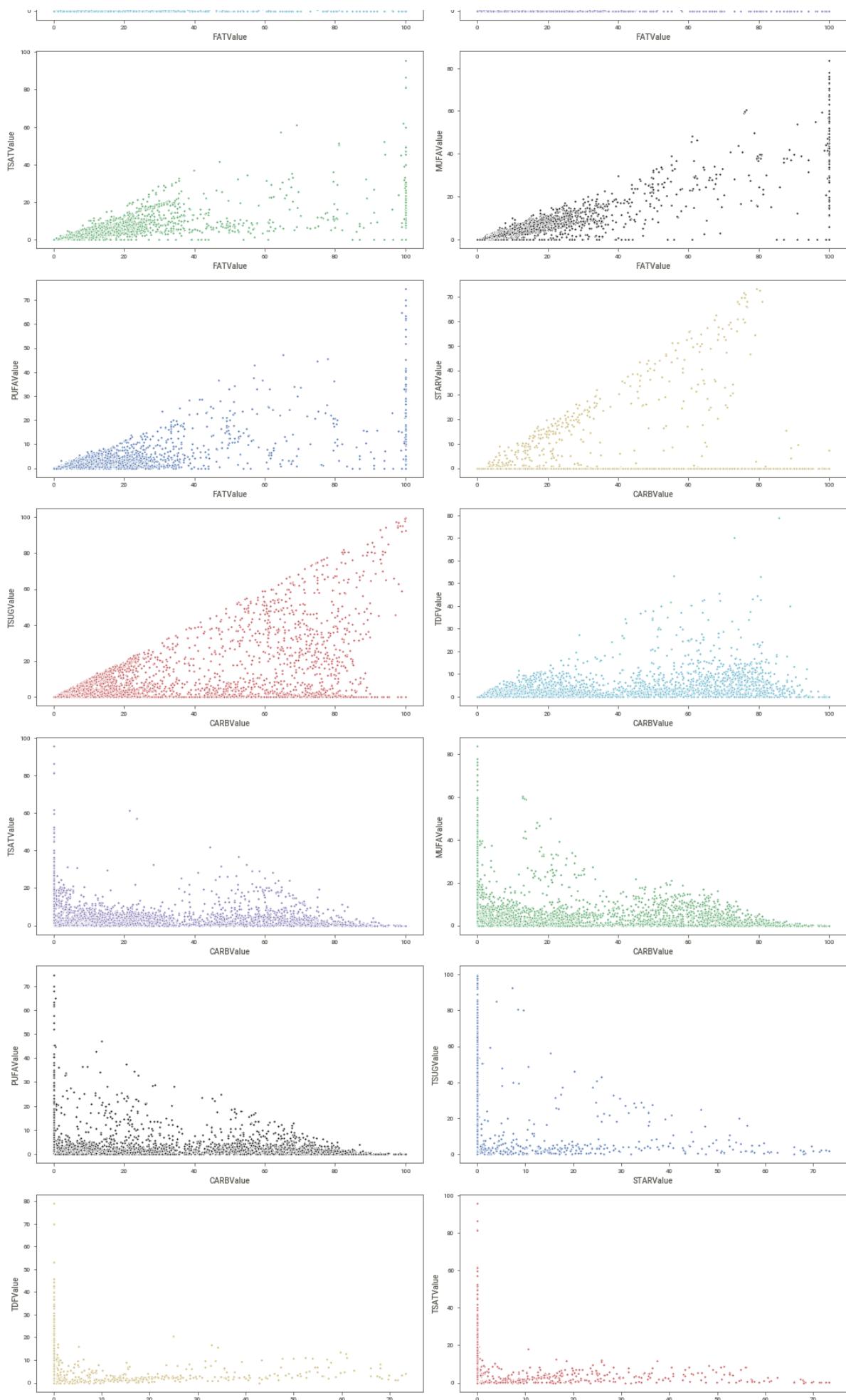


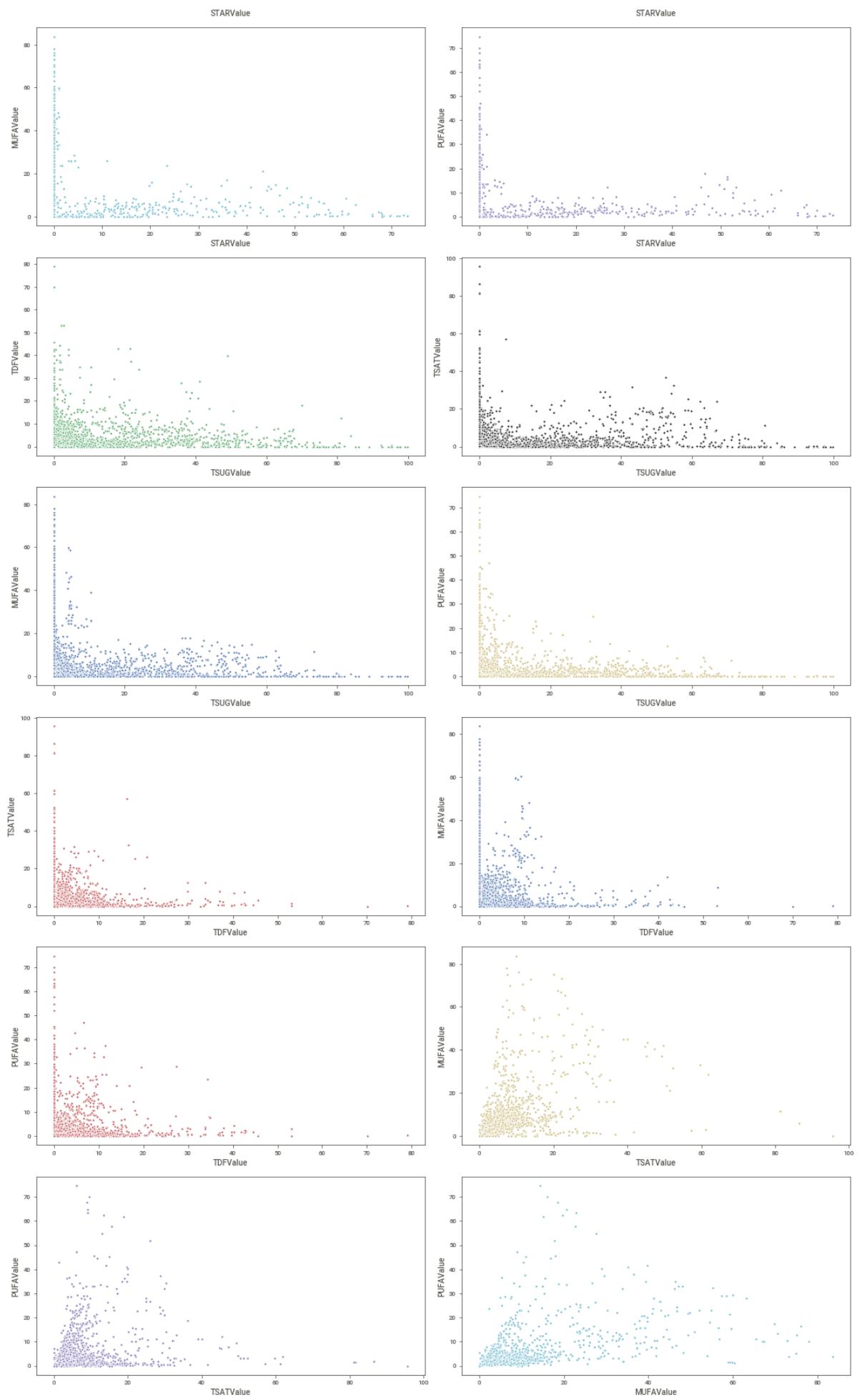
```
In [40]: df = AV.AutoViz('FoodNutritionData.csv')
```

Shape of your Data Set: (5690, 12)
CLASSIFIING VARIABLES #####
Classifying variables in data set...
Number of Numeric Columns = 9
Number of Integer-Categorical Columns = 0
Number of String-Categorical Columns = 1
Number of Factor-Categorical Columns = 0
Number of String-Boolean Columns = 0
Number of Numeric-Boolean Columns = 0
Number of Discrete String Columns = 1
Number of NLP String Columns = 0
Number of Date Time Columns = 0
Number of ID Columns = 1
Number of Columns to Delete = 0
12 Predictors classified...
This does not include the Target column(s)
2 variables removed since they were ID or low-information variables
Number of All Scatter Plots = 45

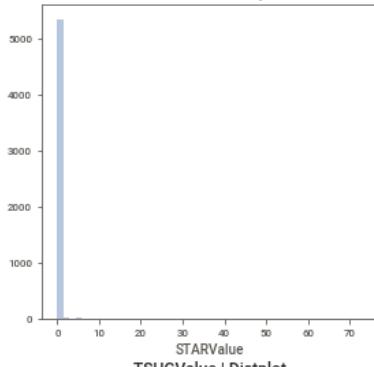
Pair-wise Scatter Plot of all Continuous Variables



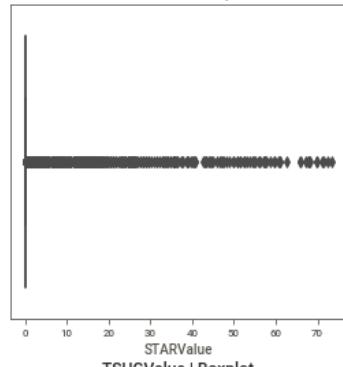




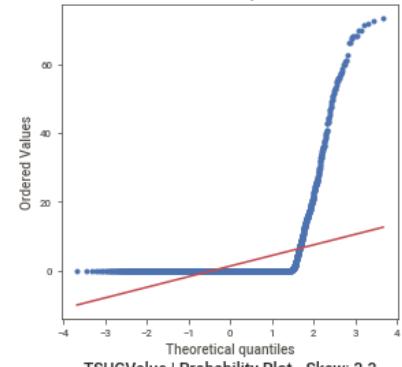
STARValue | Distplot



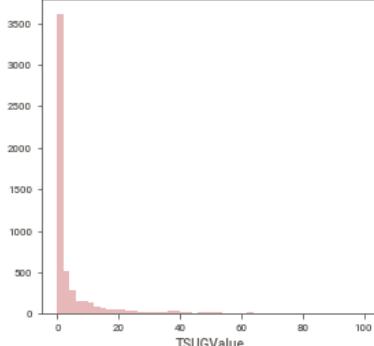
STARValue | Boxplot



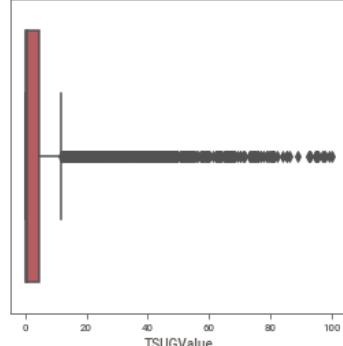
STARValue | Probability Plot - Skew: 6.6



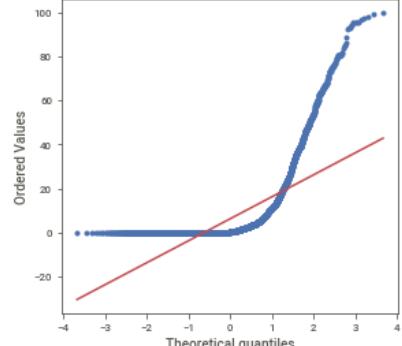
TSUGValue | Distplot



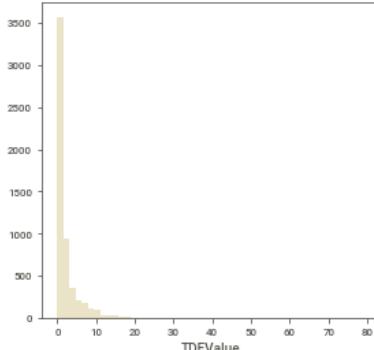
TSUGValue | Boxplot



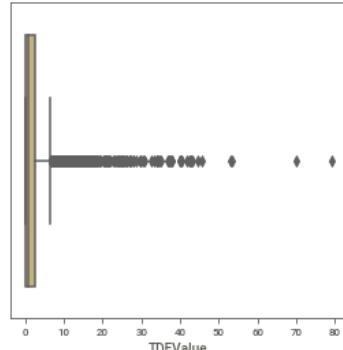
TSUGValue | Probability Plot - Skew: 3.3



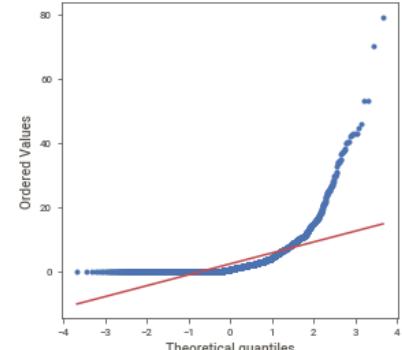
TDFValue | Distplot



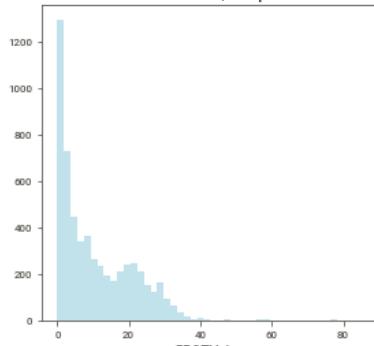
TDFValue | Boxplot



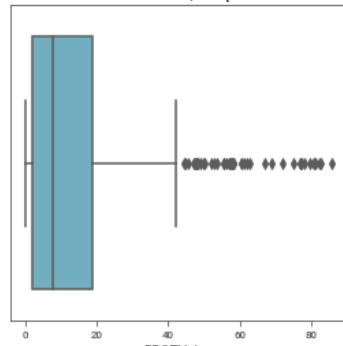
TDFValue | Probability Plot - Skew: 5.2



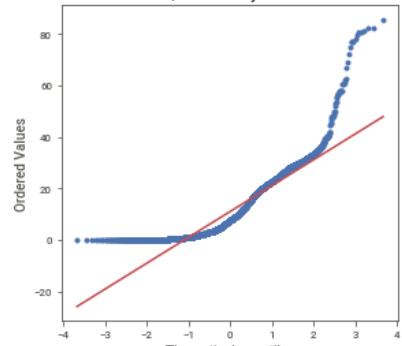
PROTValue | Distplot



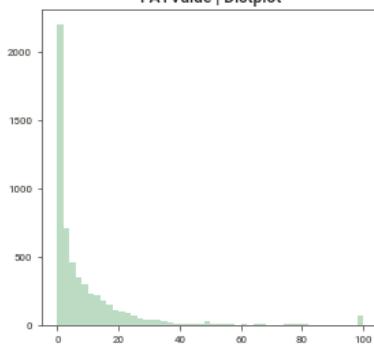
PROTValue | Boxplot



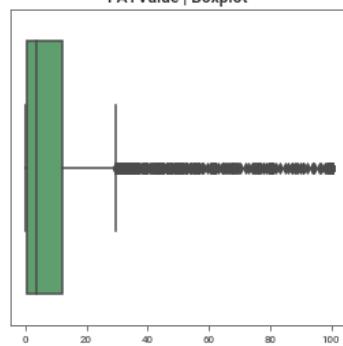
PROTValue | Probability Plot - Skew: 1.5



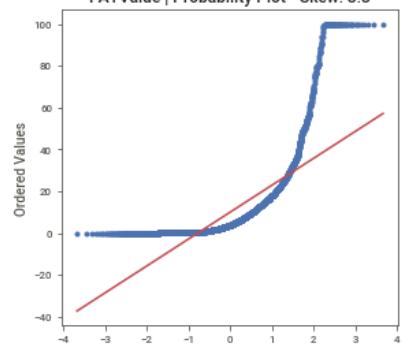
FATValue | Distplot



FATValue | Boxplot



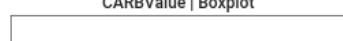
FATValue | Probability Plot - Skew: 3.3



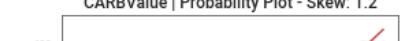
CARBValue | Distplot

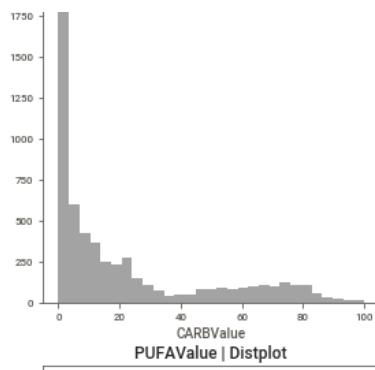


CARBValue | Boxplot

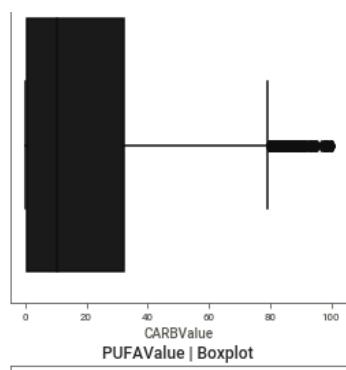


CARBValue | Probability Plot - Skew: 1.2

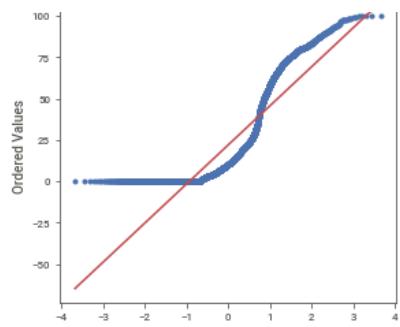




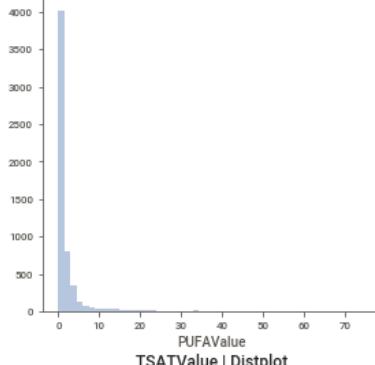
PUFAValue | Distplot



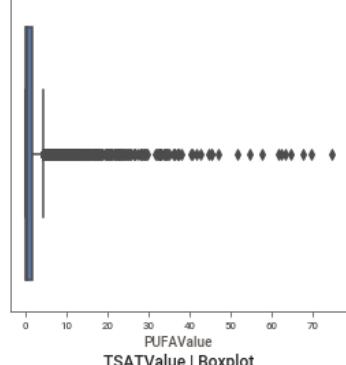
PUFAValue | Boxplot



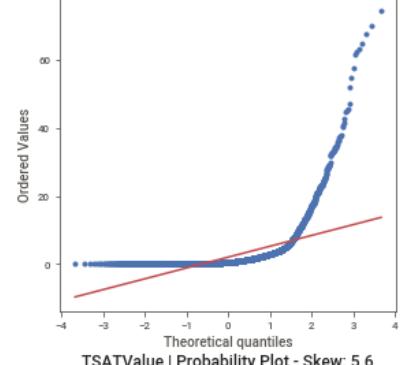
PUFAValue | Probability Plot - Skew: 6.3



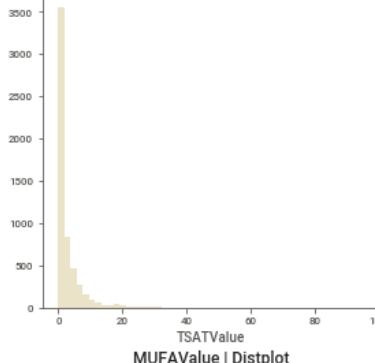
PUFAValue | Distplot



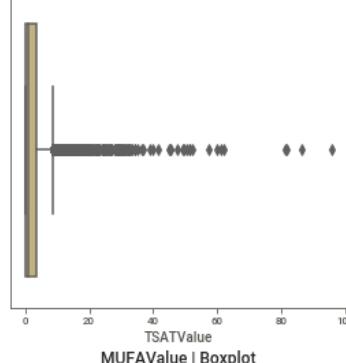
PUFAValue | Boxplot



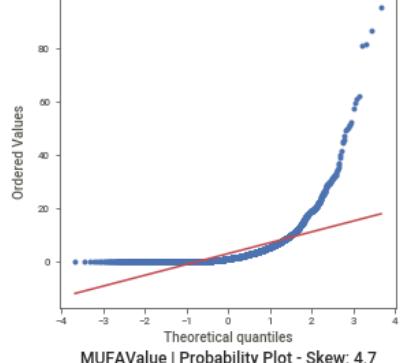
PUFAValue | Probability Plot - Skew: 6.3



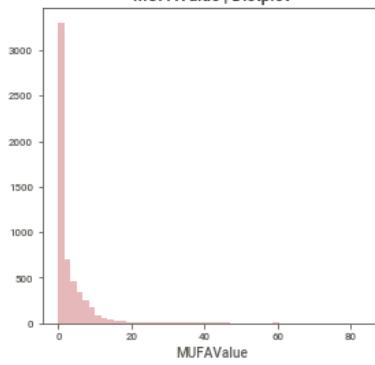
TSATValue | Distplot



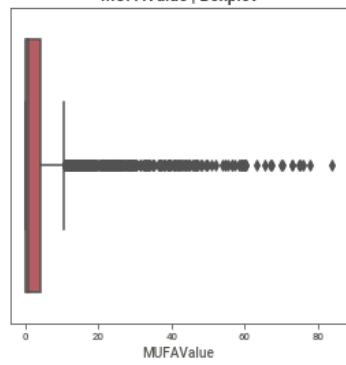
TSATValue | Boxplot



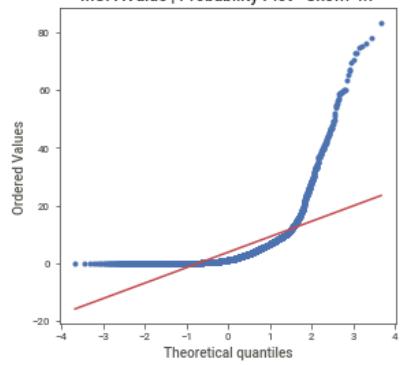
TSATValue | Probability Plot - Skew: 5.6



MUFAValue | Distplot

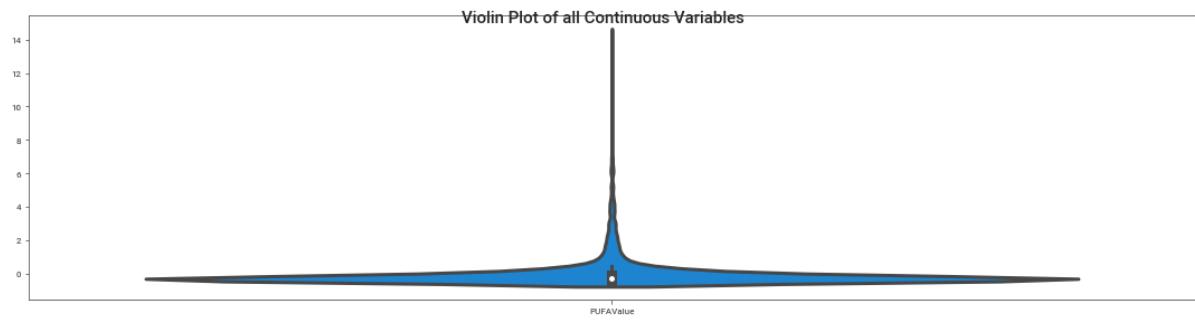
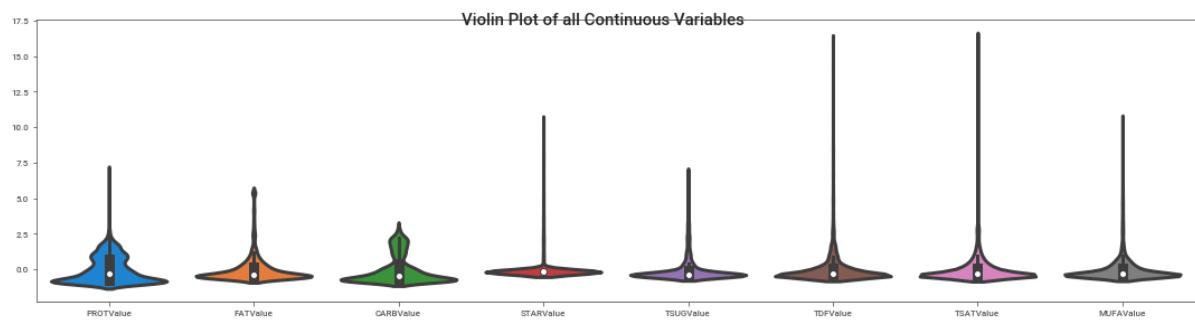
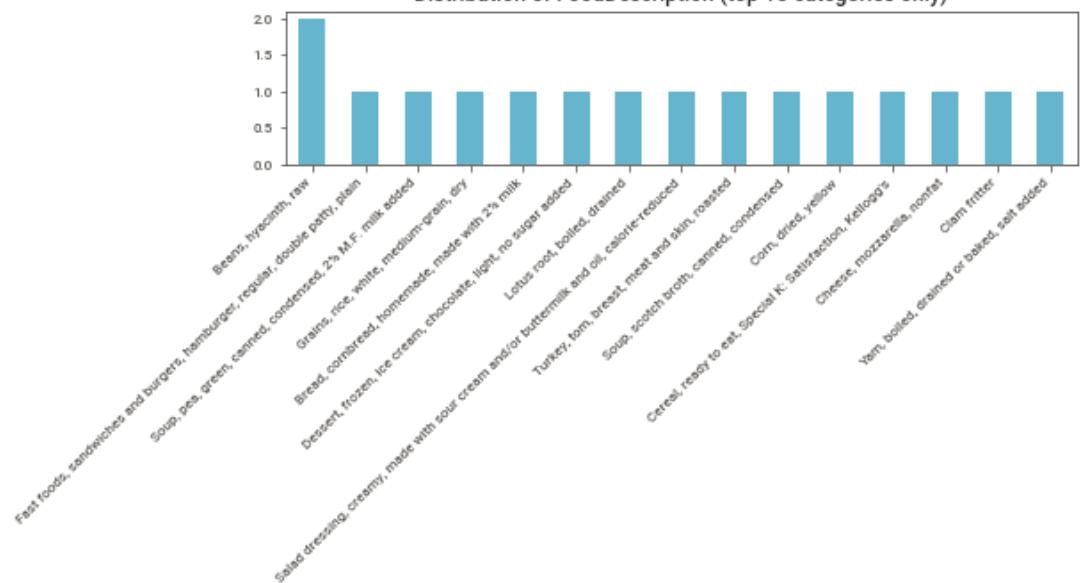


MUFAValue | Boxplot

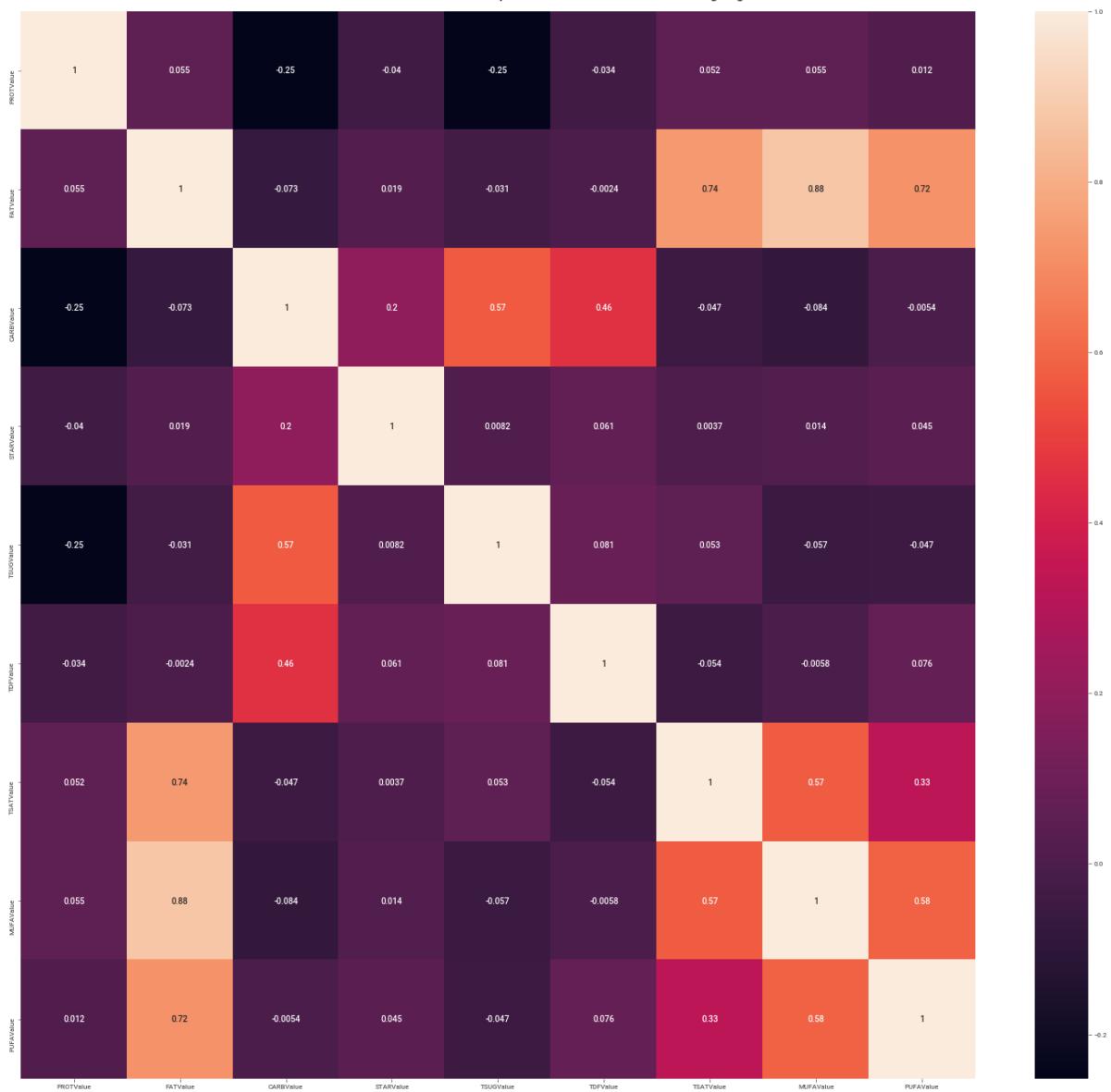


MUFAValue | Probability Plot - Skew: 4.7

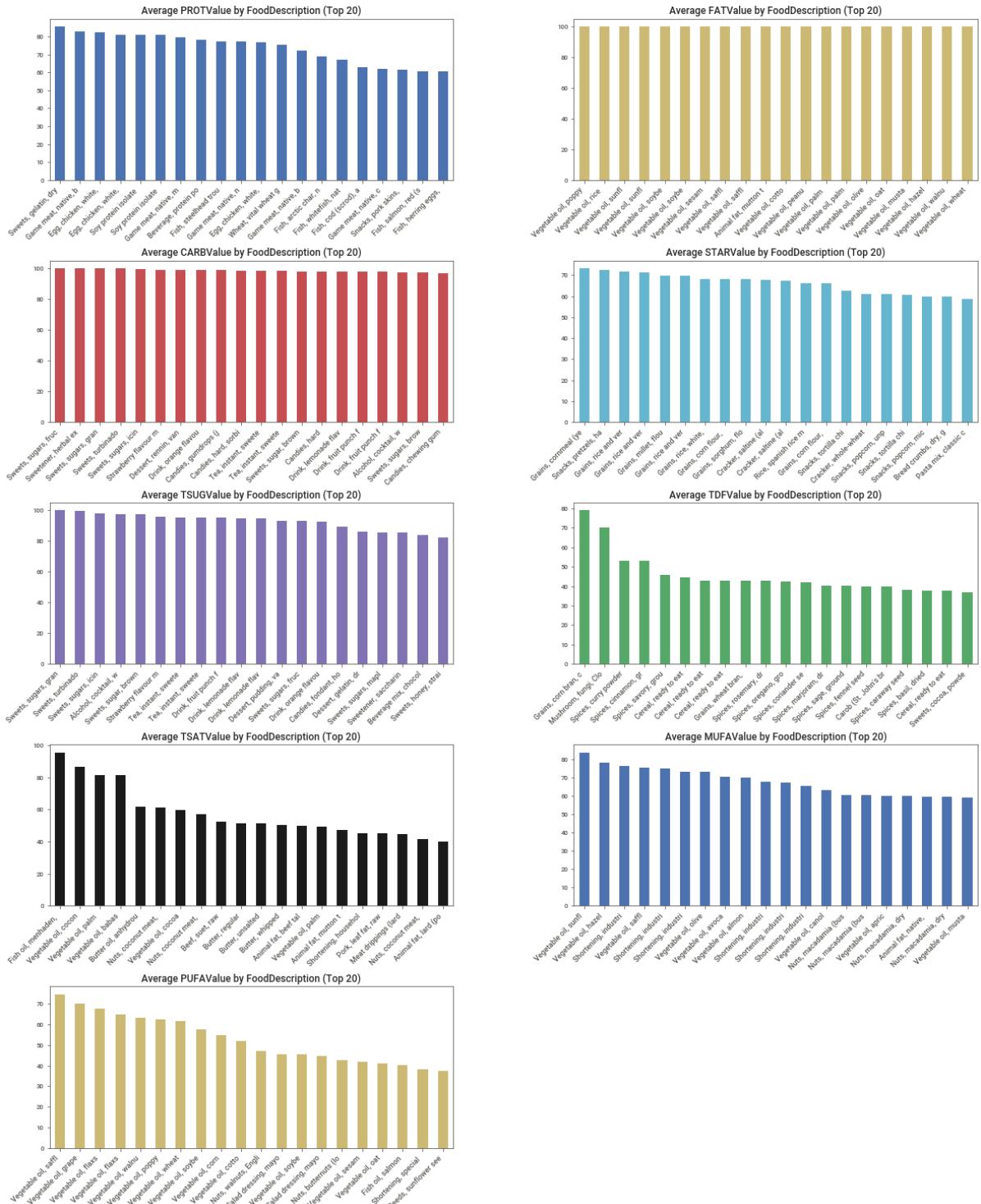
Histograms (KDE plots) of all Continuous Variables
Distribution of FoodDescription (top 15 categories only)



Heatmap of all Continuous Variables including target =



Bar plots for each Continuous by each Categorical variable



Time to run AutoViz (in seconds) = 10.098

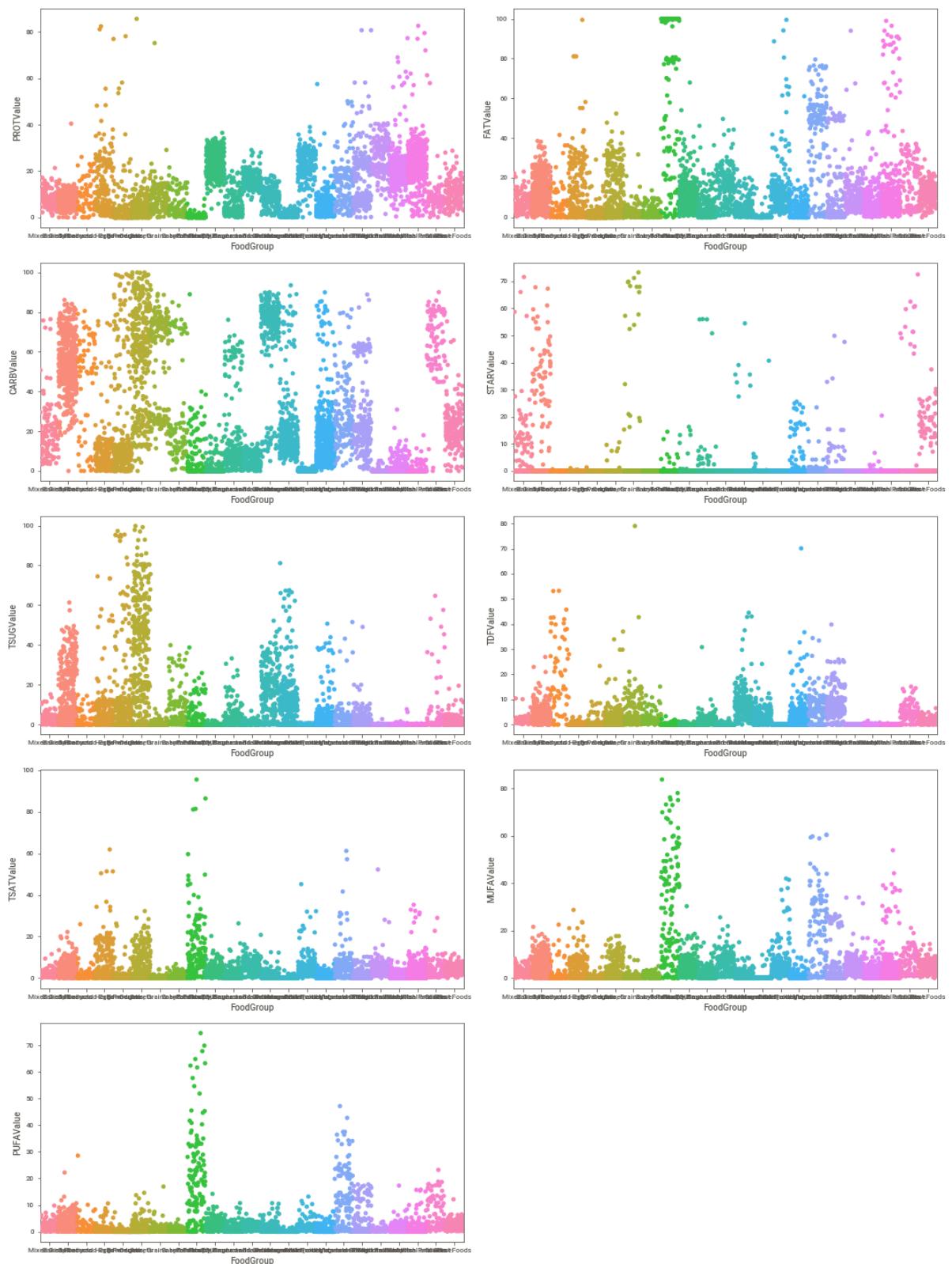
VISUALIZATION Completed

```
In [41]: f = AV.AutoViz('FoodNutritionData.csv', depVar='FoodGroup')
```

```
Shape of your Data Set: (5690, 12)
##### CLASSIFIING VARIABLES #####
Classifying variables in data set...
    Number of Numeric Columns = 9
    Number of Integer-Categorical Columns = 0
    Number of String-Categorical Columns = 1
    Number of Factor-Categorical Columns = 0
    Number of String-Boolean Columns = 0
    Number of Numeric-Boolean Columns = 0
    Number of Discrete String Columns = 0
    Number of NLP String Columns = 0
    Number of Date Time Columns = 0
    Number of ID Columns = 1
    Number of Columns to Delete = 0
11 Predictors classified...
    This does not include the Target column(s)
    1 variables removed since they were ID or low-information variables

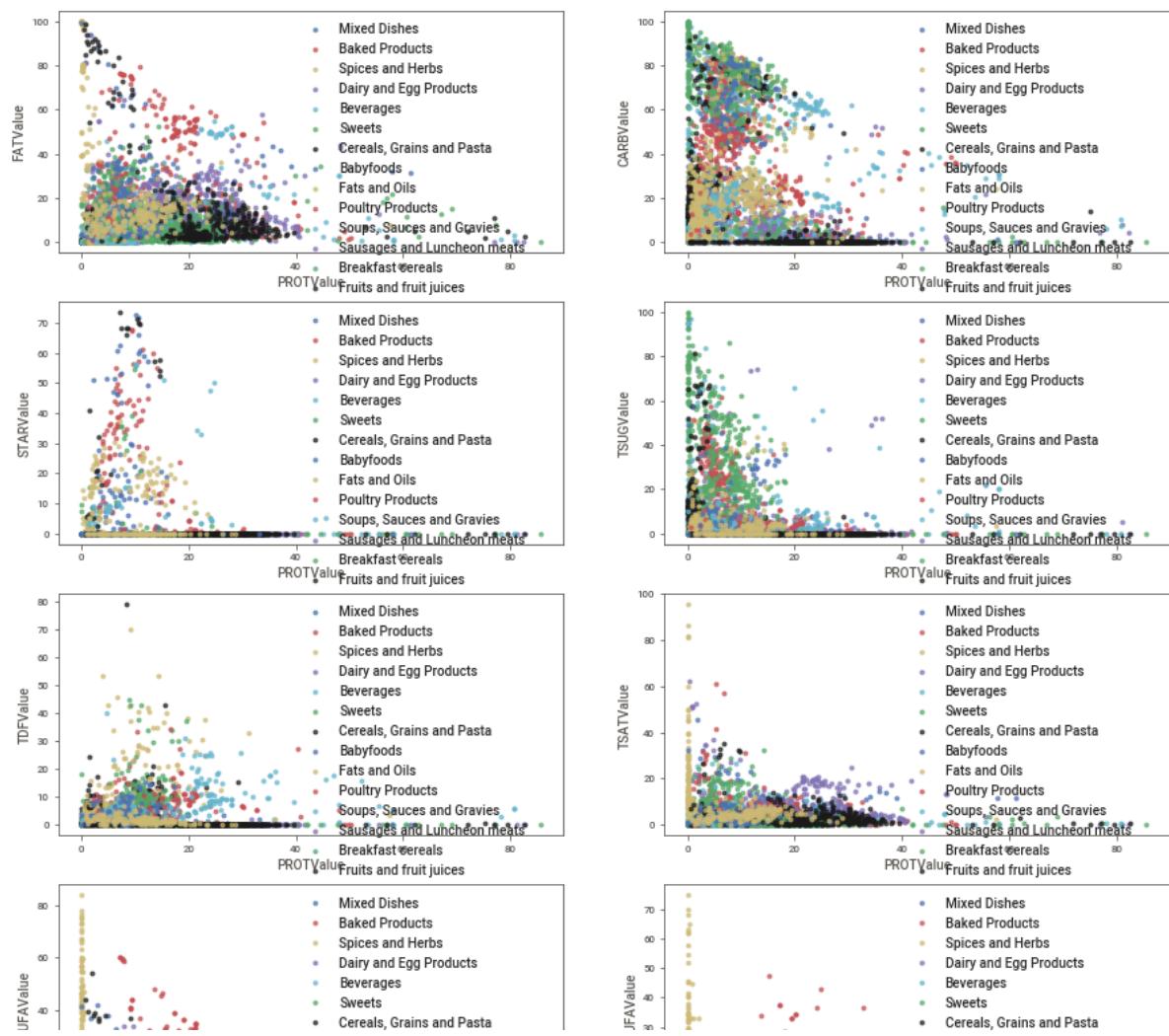
##### Multi_Classification VISUALIZATION Started #####
####
```

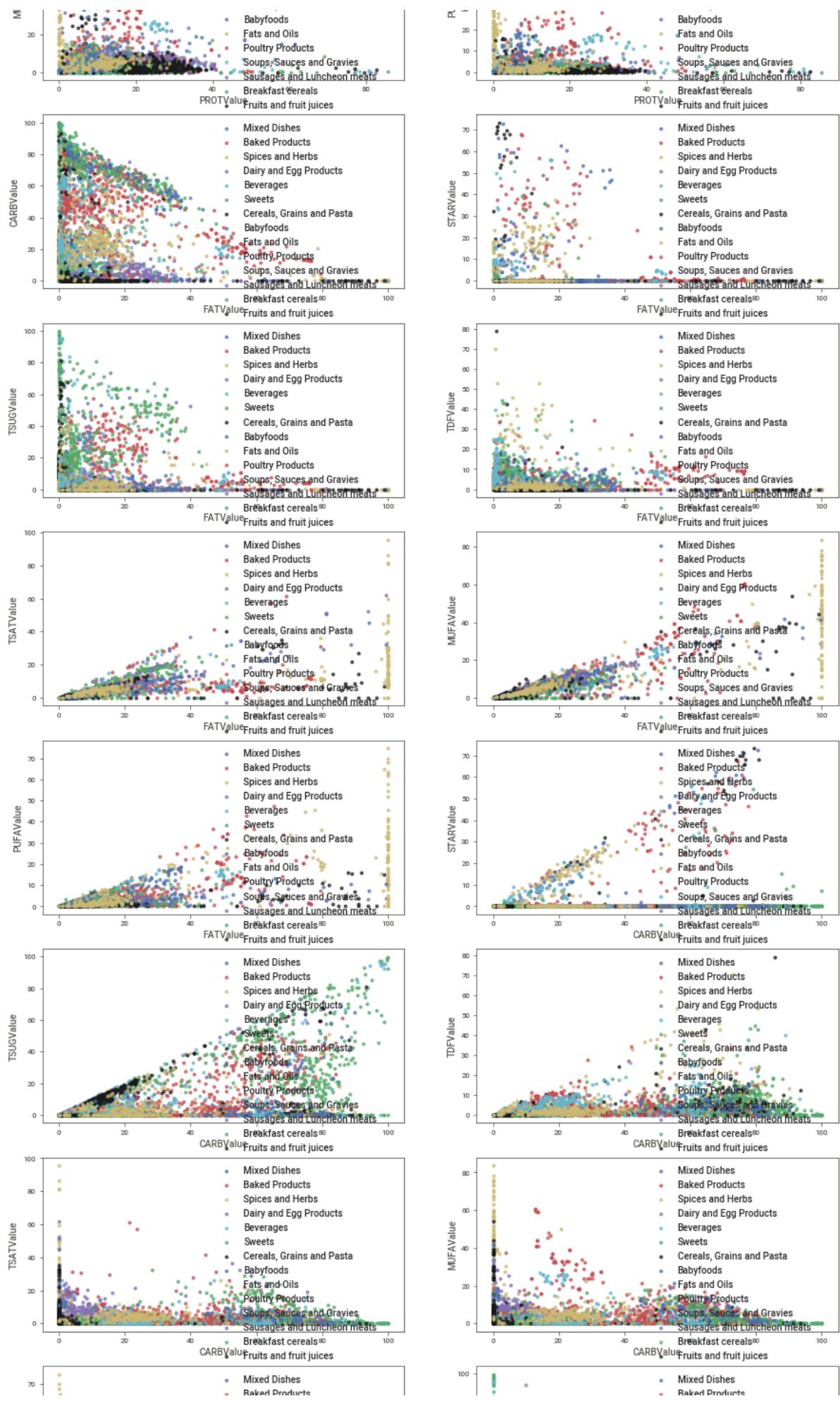
Scatter Plot of Continuous Variable vs Target (jitter=0.50)

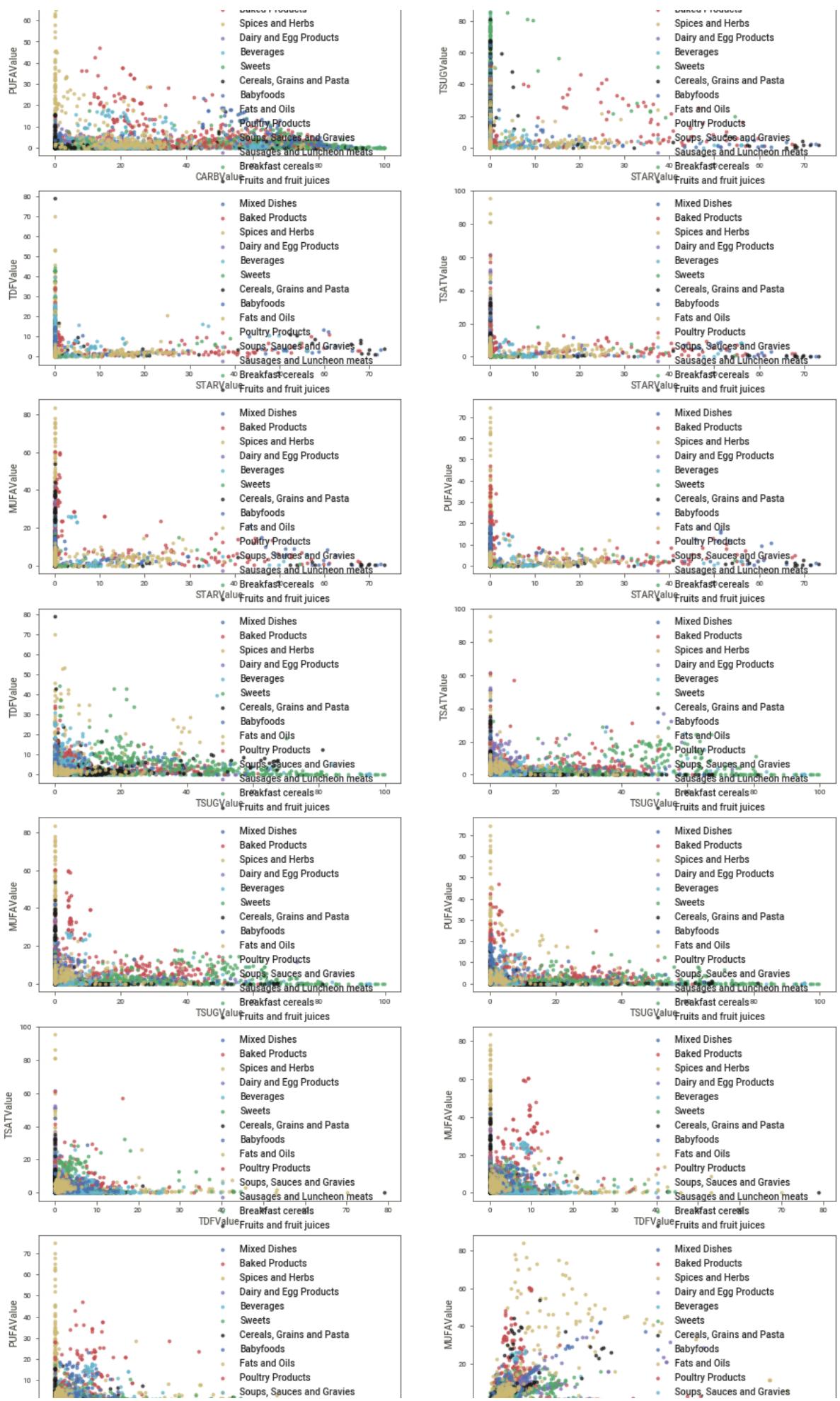


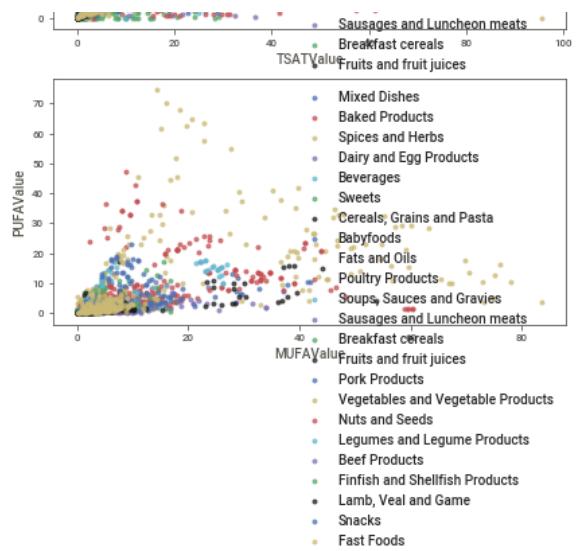
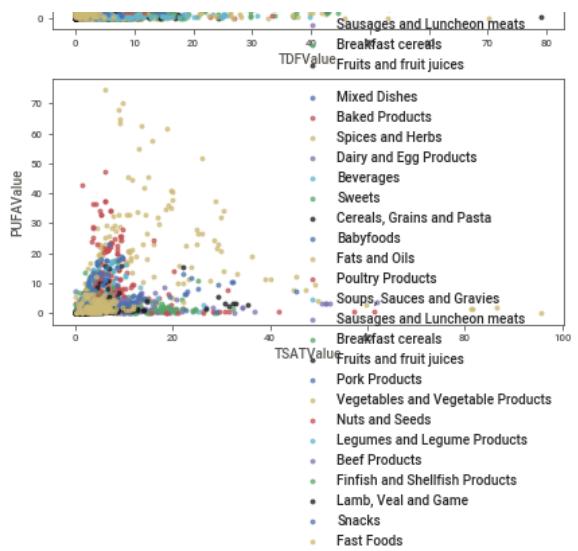
Total Number of Scatter Plots = 45

Pair-wise Scatter Plot of all Continuous Variables

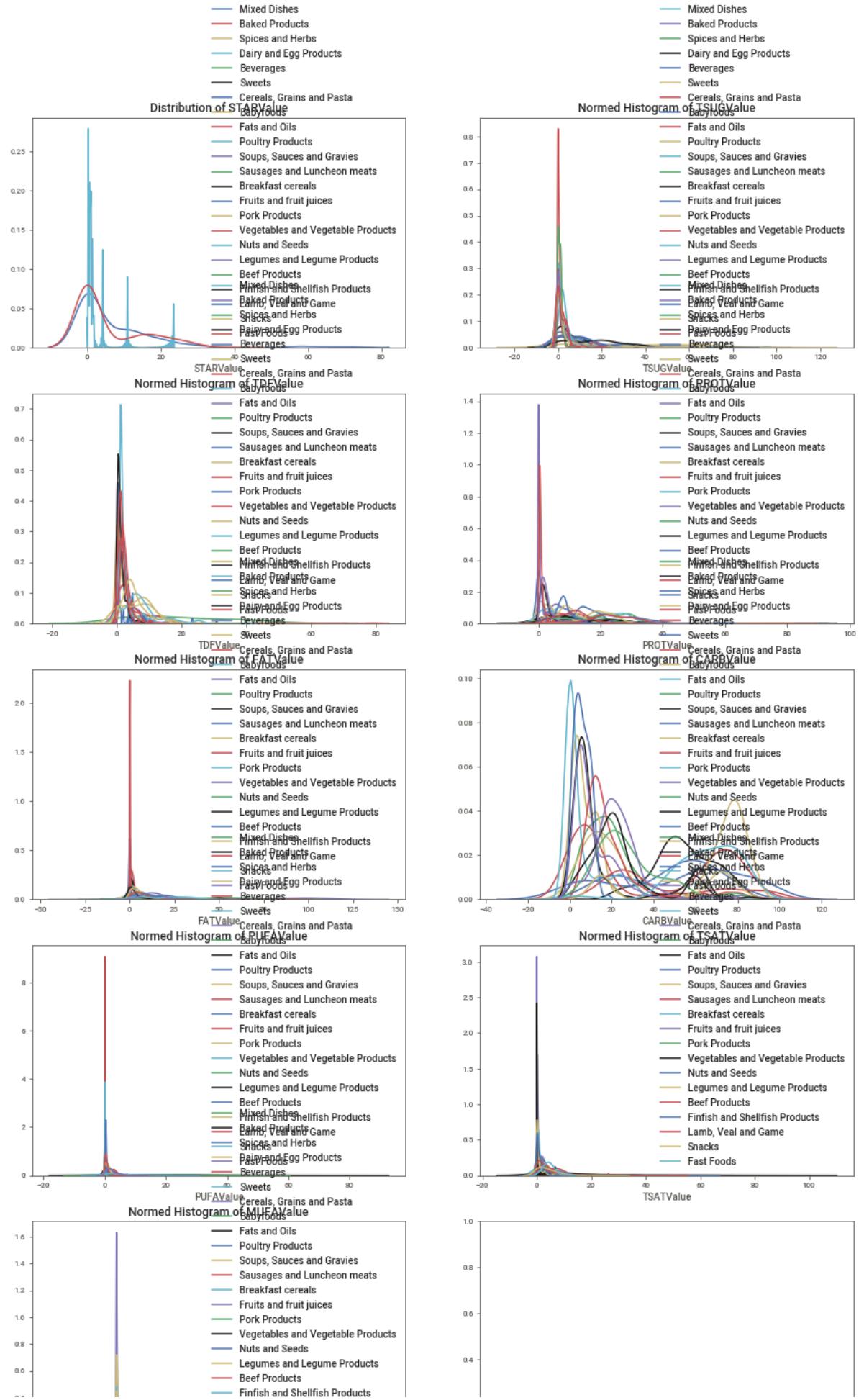


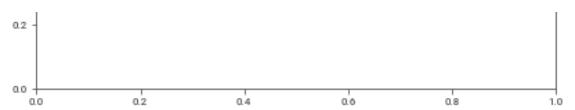
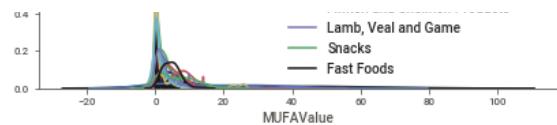


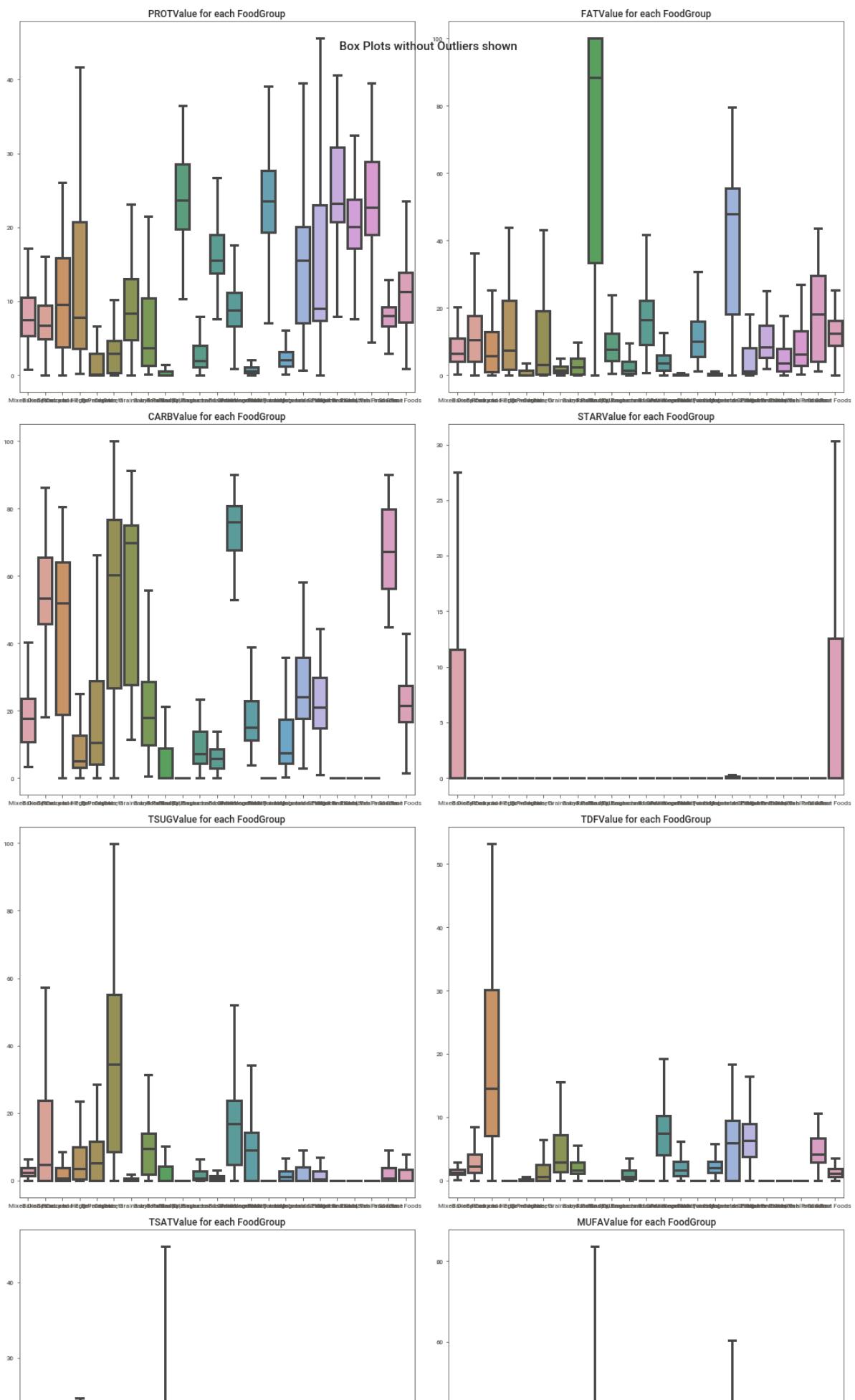


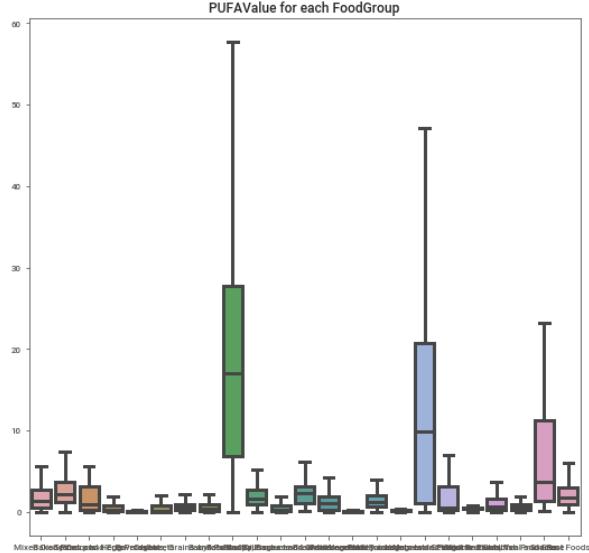
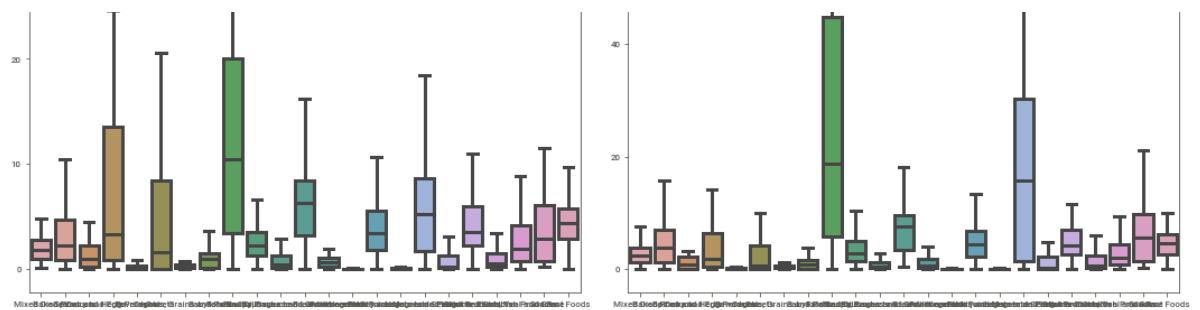


Could not draw Distribution Plots

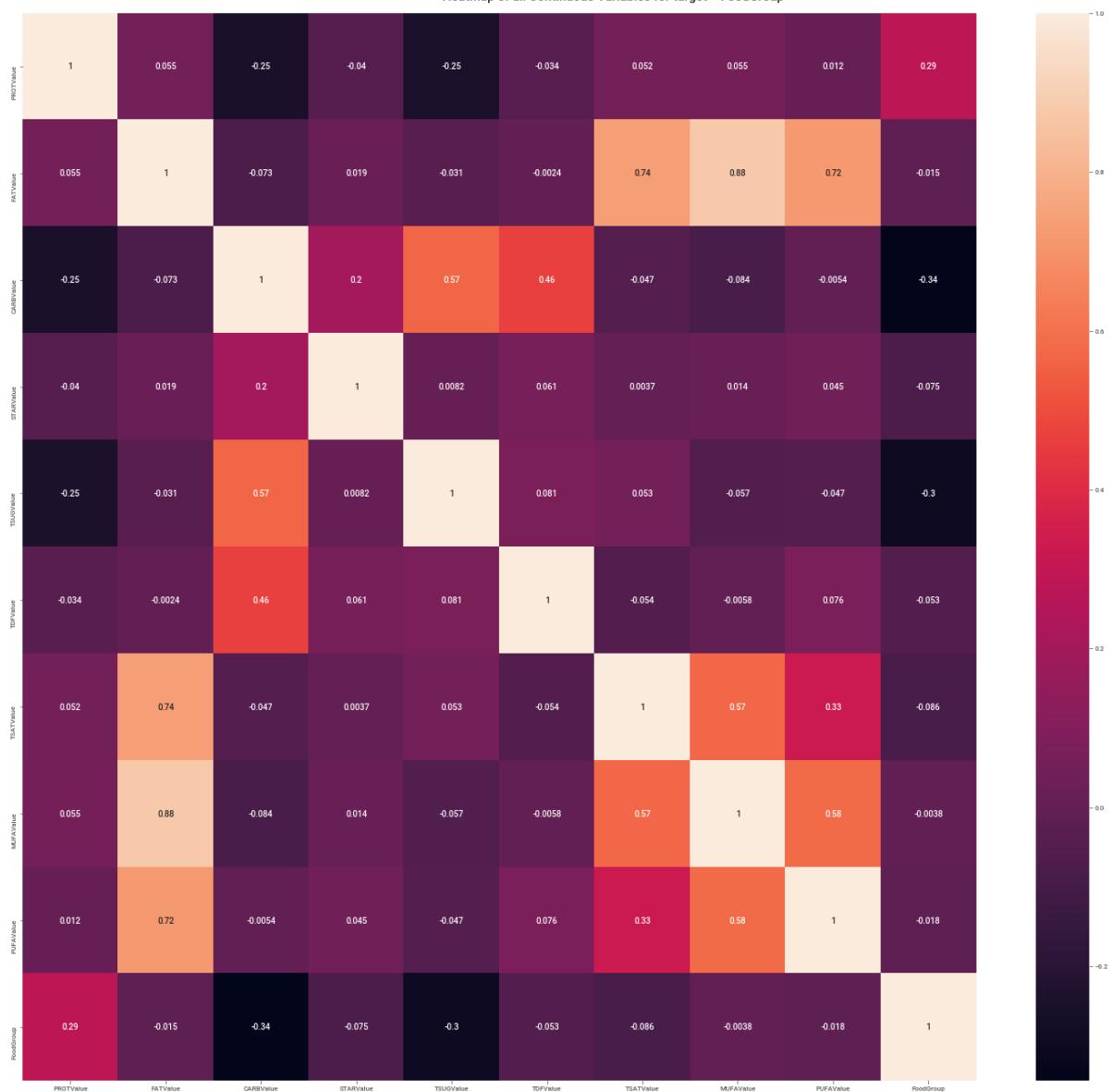




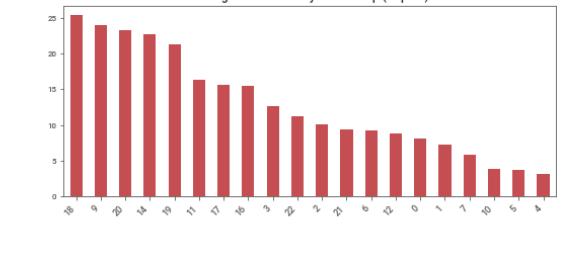
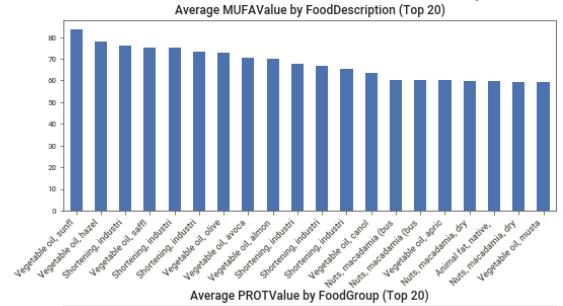
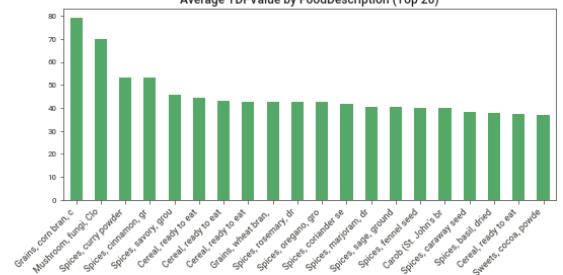
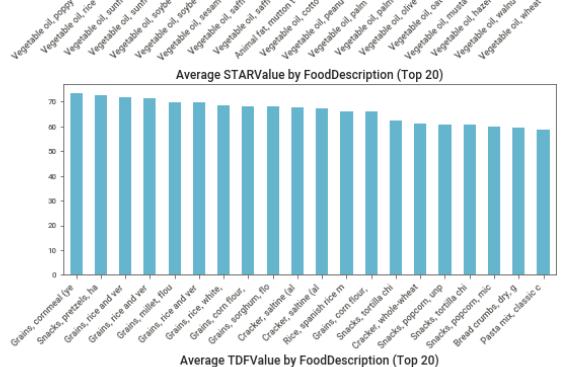
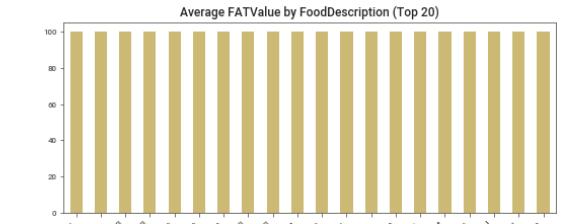
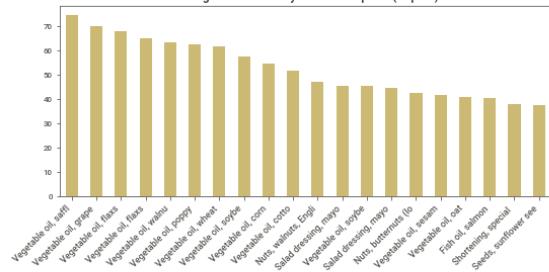
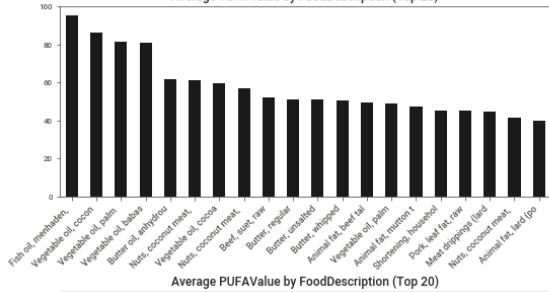
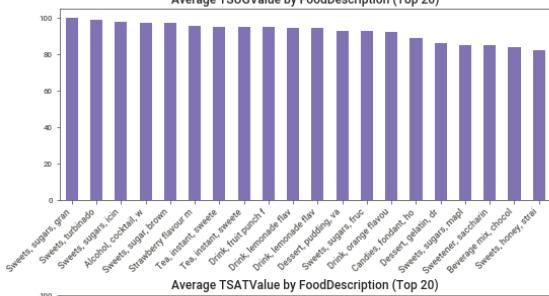
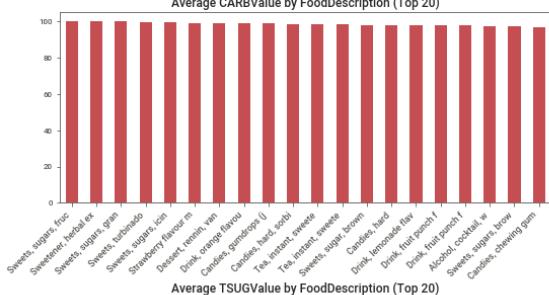
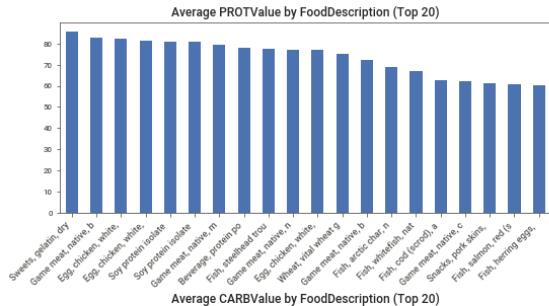


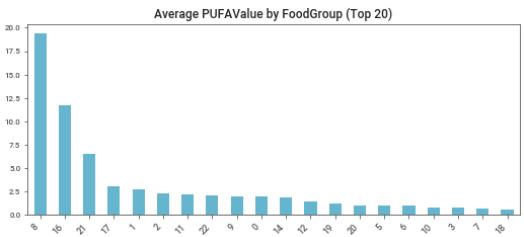
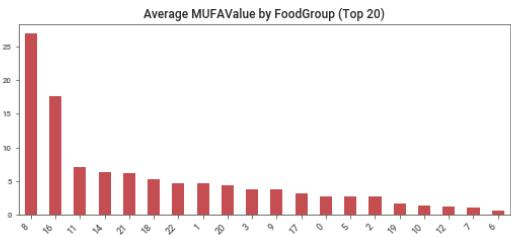
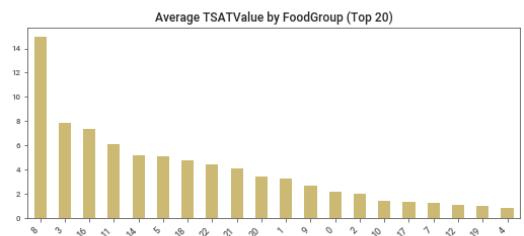
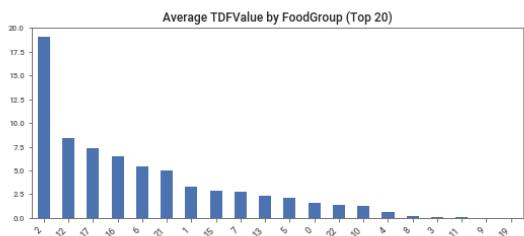
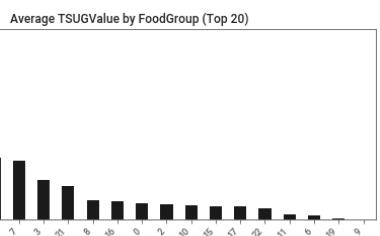
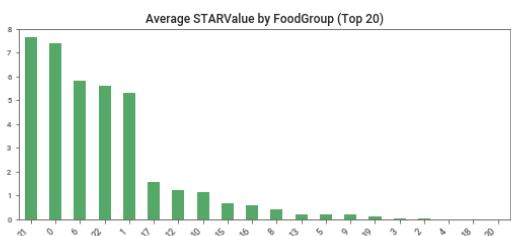
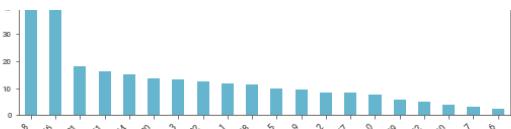


Heatmap of all Continuous Variables for target = FoodGroup



Bar plots for each Continuous by each Categorical variable





Time to run AutoViz (in seconds) = 29.156

VISUALIZATION Completed