# PERSPECTIVES OF DATA SCIENCE

# FINAL PROJECT

# FALL 2022

# PROJECT TITLE

# McDonald’s India: Menu Nutrition

**By,**

**Indu Priya Chavva (K593E853)**

**Pavani Kondapalli (W388D345)**

**Prince David Sanam (M339J323)**

**Naga Lakshmi Gade (M555Z979)**

**Sarath Triveni Bavanam (S524A344)**

Table of Contents

Abstract 1

Introduction 1

problem statement: 3

hypothesis: 3

data wrangling. 4

exploratory data analysis (eda): 8

analysis and results: 17

conclusion: 26

leassons that we have learned: 26

**Problem Statement:**

Determination of finding whether the Order placed by the Customer in McDonald Store is giving sufficient amount of High Energy (Excess of Calories) or Low Energy (Shortage of Calories).

# Hypothesis Determination:

**Null Hypothesis**: H0: μ=μ0 where at least one order made by the customer should be giving enough Calories from the order placed.

**Alternative Hypothesis**: H1: μ > μ0 if not more than one order by the customer should be giving enough Calories.

# Importing Libraries:

# Here in this step, we import all the needed libraries.



# Taking The Data Set:

We took this data set from Kaggle. Here is the link.

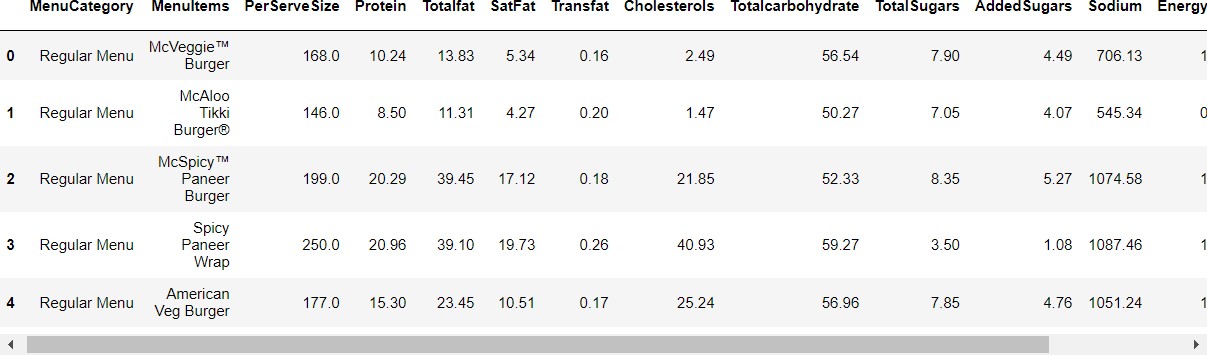
<https://www.kaggle.com/datasets/deepcontractor/mcdonalds-india-menu-nutrition-facts>

From the above-mentioned source link, we derived the data and did Pre-processing process and then we performed the different methods of Cleaning Techniques as per the required model development.

**Data Cleaning Process:**

# Reading The Dataset:

In this step we import the data set into a variable called “df”. And we also displayed the first 5 observations of the data set for clear understanding.

Here is the output of the above code

# Text Description automatically generated with medium confidenceGetting the Information of Each column of the Data Set:

# Table Description automatically generatedOutput of the info() is:

# Getting the shape of the Dataset:

The shape describes the number of rows and columns of the data set.

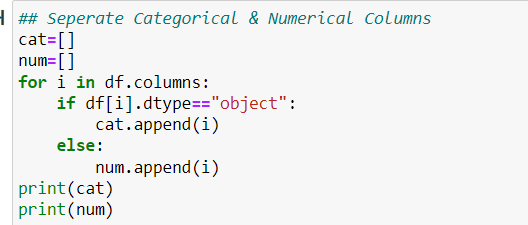


The output of the above code is:

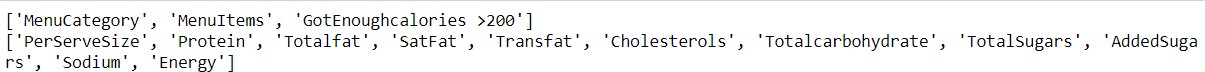


# Separating the Categorial & Numeric Columns:

Separating all the numerical and Categorial columns of the data set.



The output of the above code is:



# Finding the Missing Values:

We Find the missing values of the numerical and the categorial values of the data.

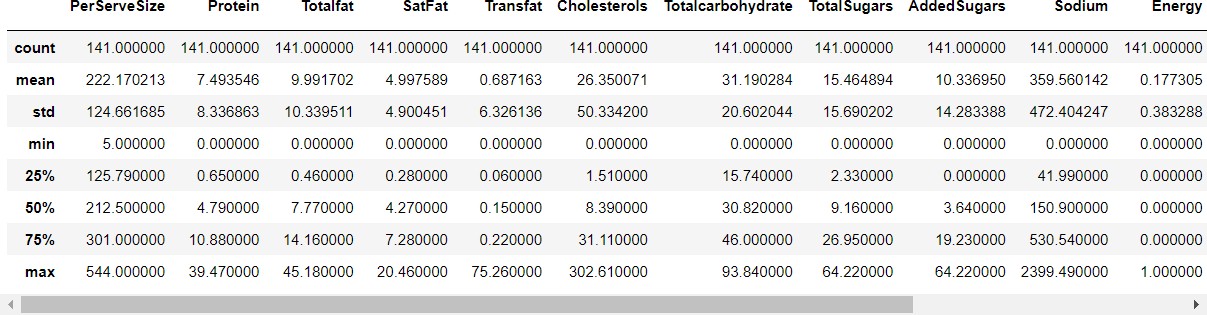


# Finding the Description of Dataset:

Here we find the description of the data from the Data set:

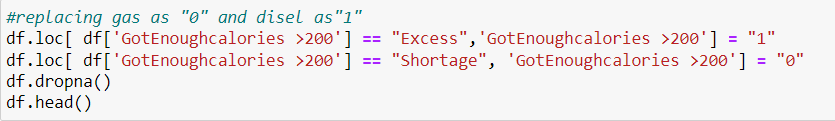


The output of the above code is:

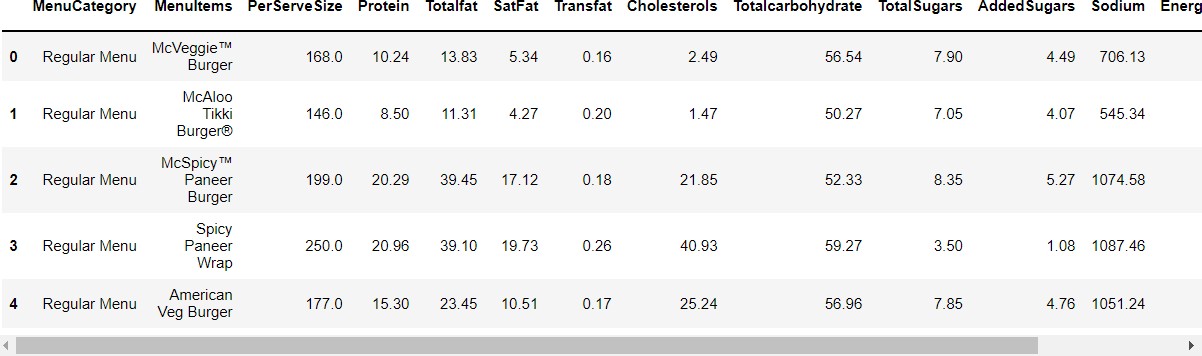


# Replacing the Categorical Data into Numeric Data:

Here we replace all the categorical data into the numeric data.



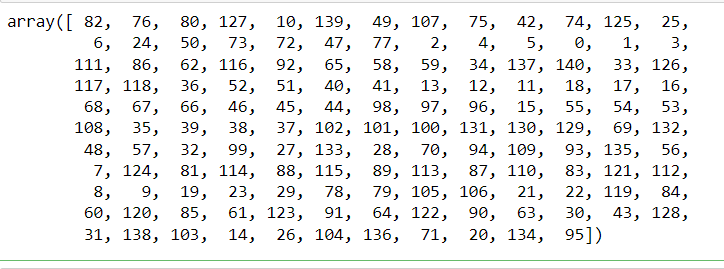
The output of the above code:



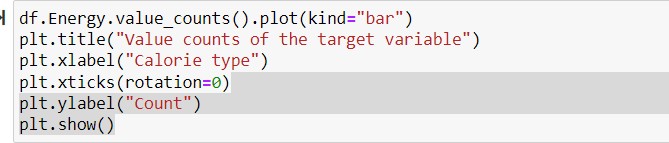
Replacing the menu Items column into numeric data:



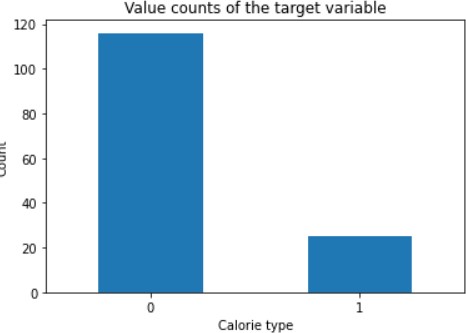
The output of the code is:



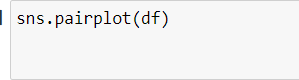
# EDA on Dataset:



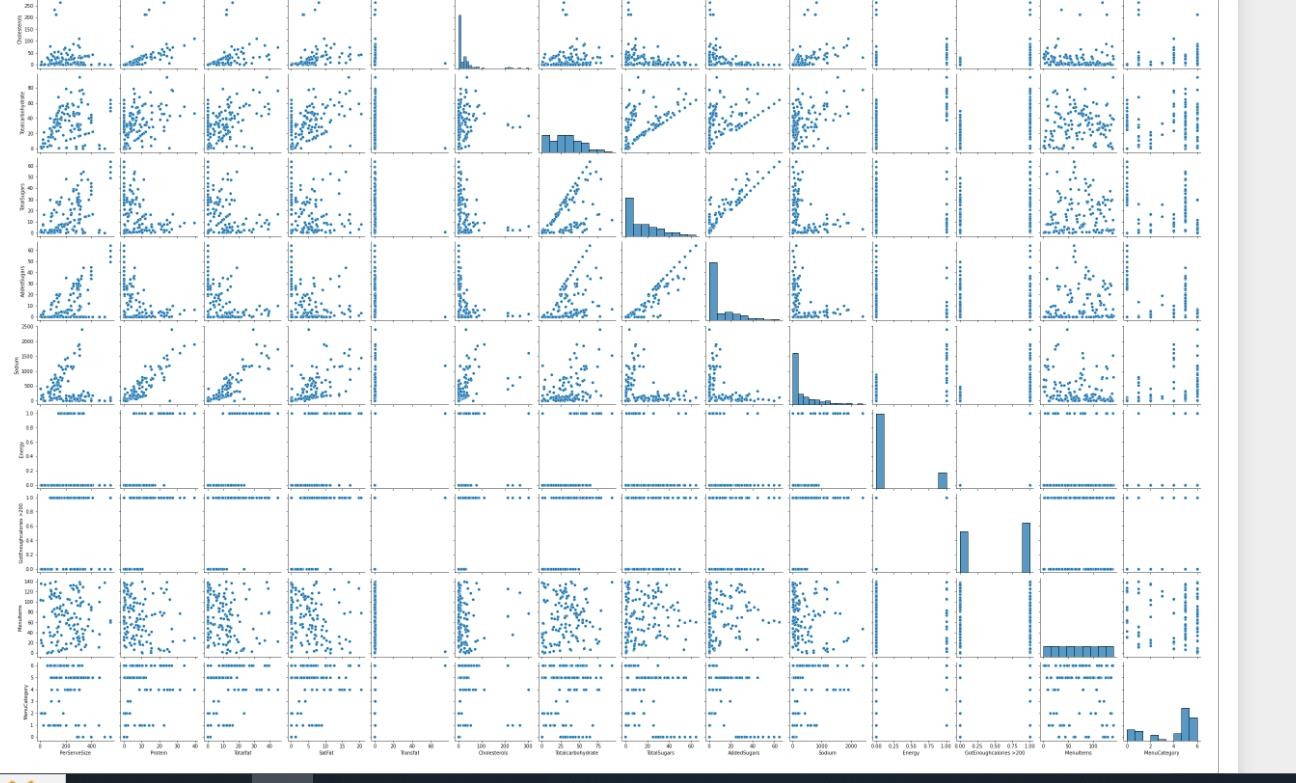
The output of the code is:



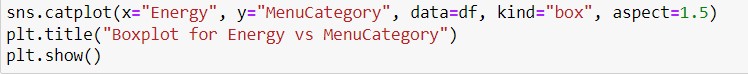
The pairwise plot of the data is:



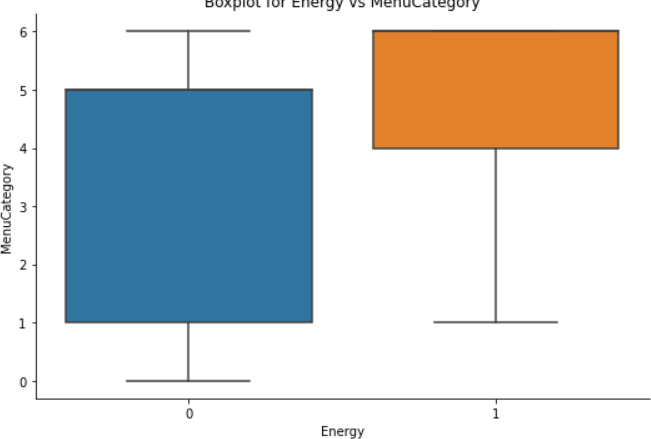
The output of the data is:



The boxplot of the data is:

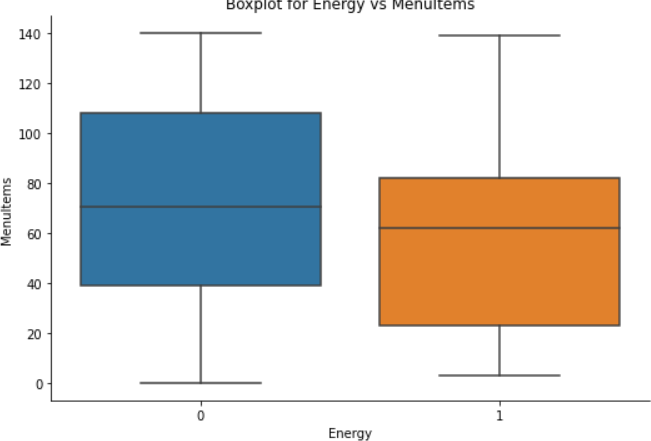


The output of the data is:

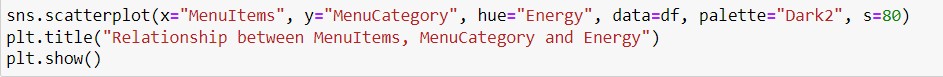


The boxplot between the Energy and Menu Items:

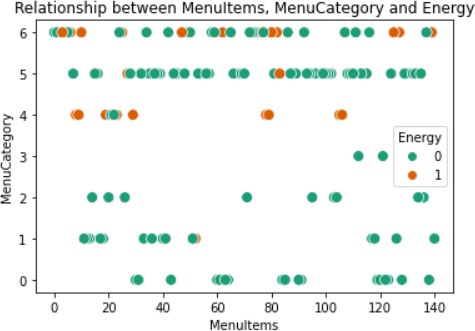


The output of the code is:

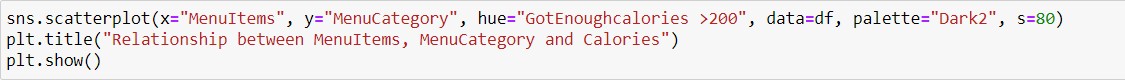
The scatterplot of the data is:



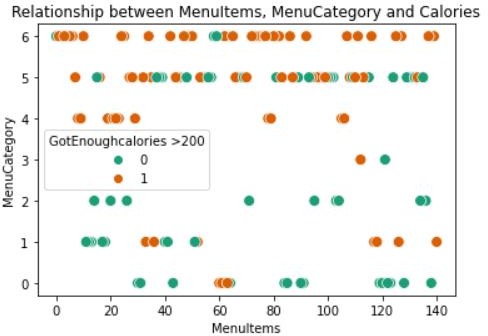
The output of the above code is:



The scatterplot between the different variables:



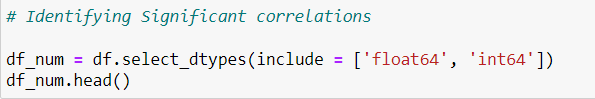
The output of the code is:



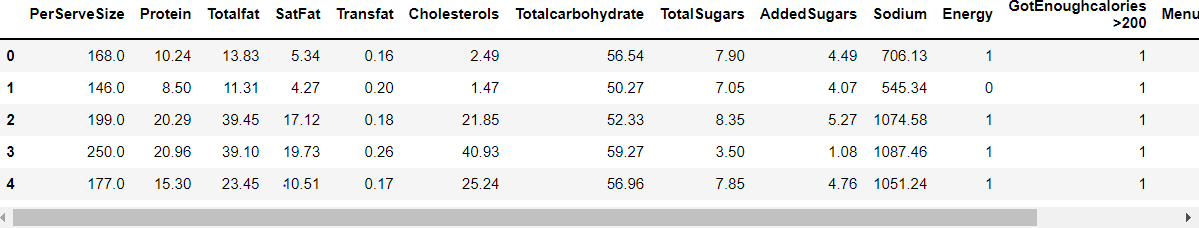
# Feature Selection:

# Filter Method:

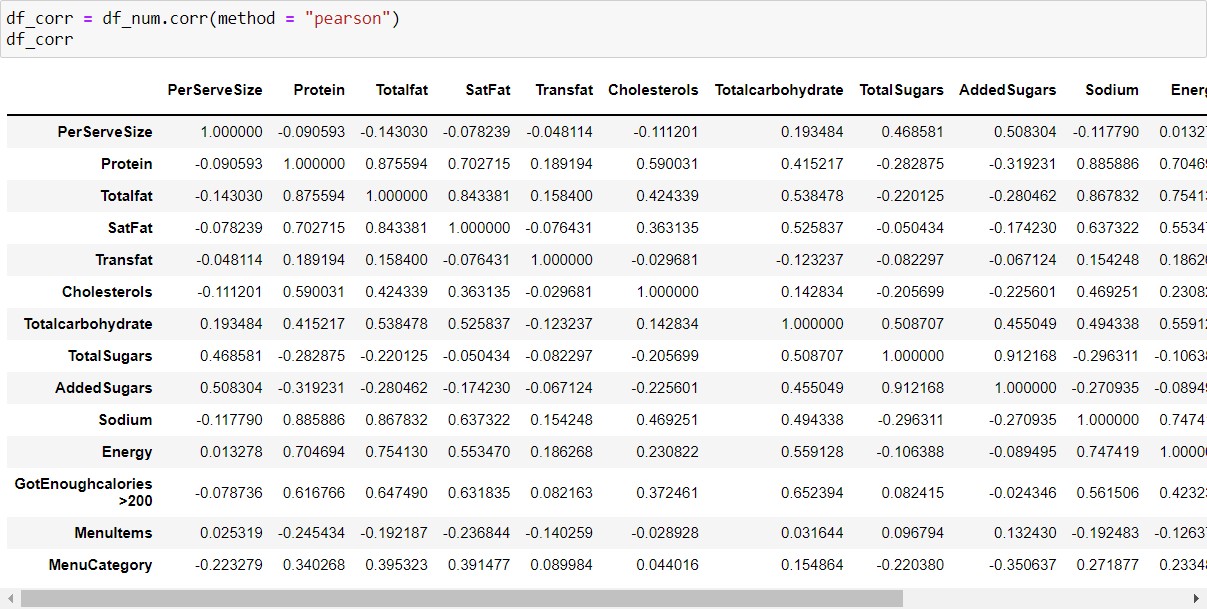
Identifying the significant correlation between the variables of the data set:



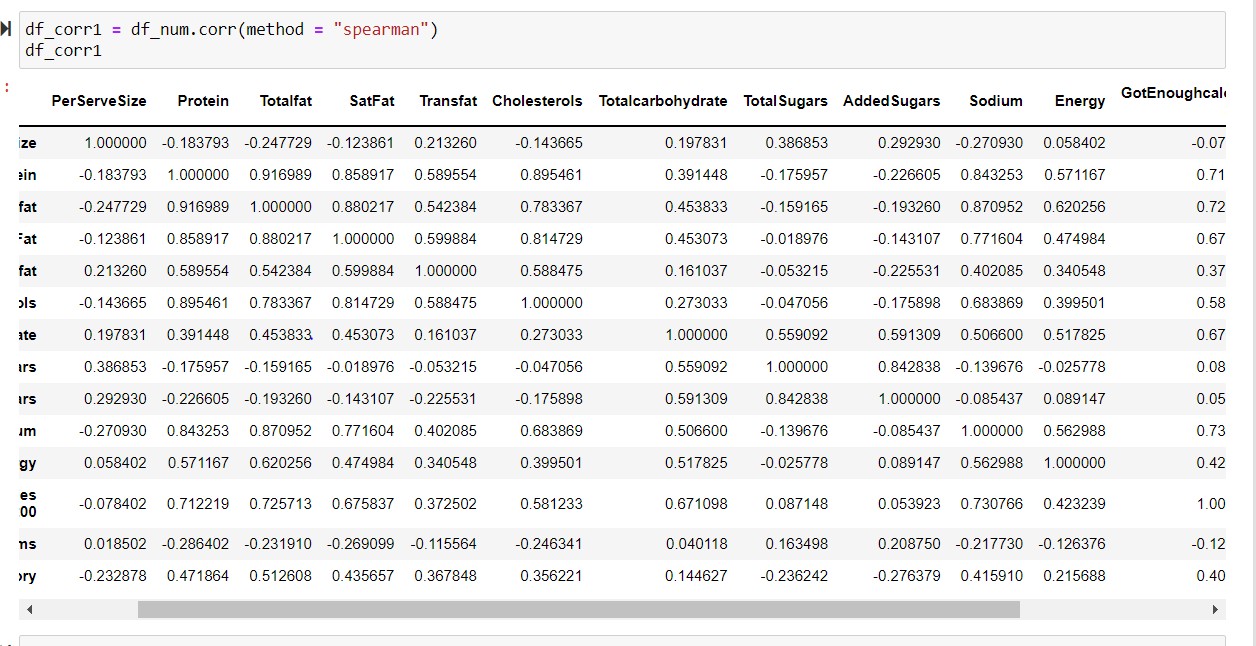
The output of the code is:



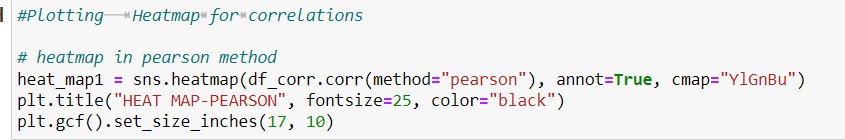
The correlation matrix of the Pearson in the data set is:



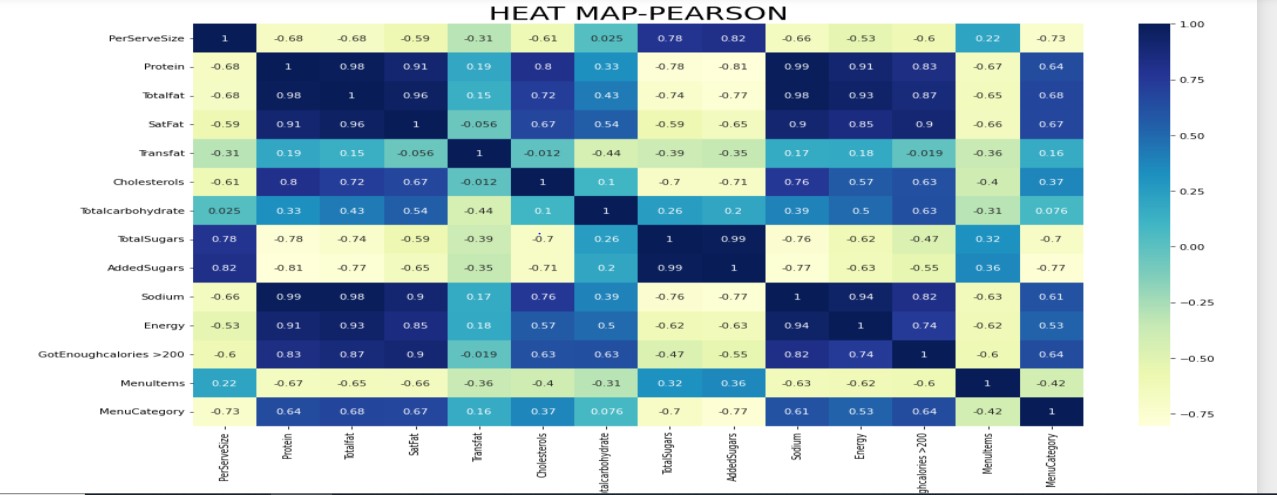
The correlation matrix of the spearmen is:



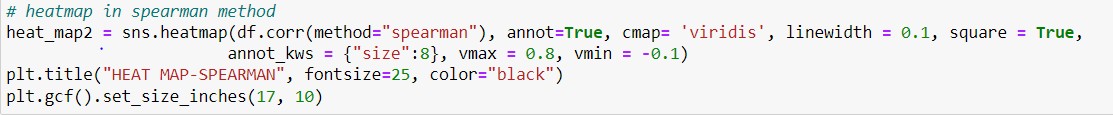
The heat map correlation of the variables in the data set is:



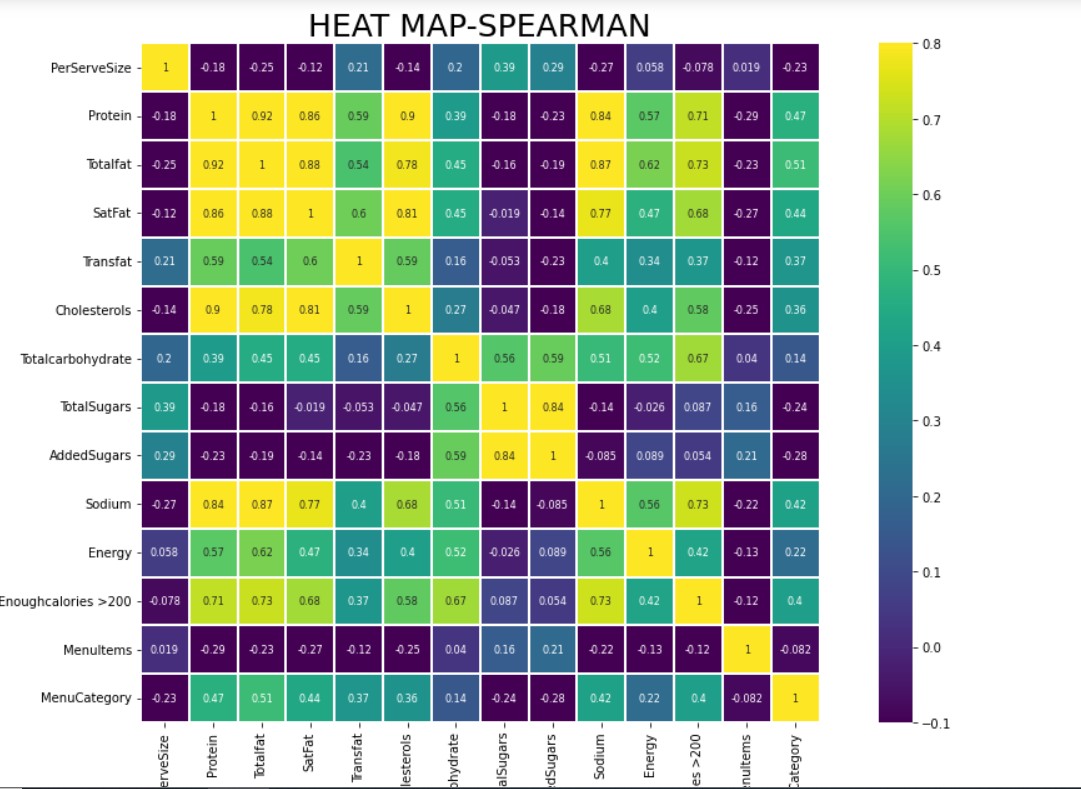
The output of the heat map is:



The heat map correlation of the spearmen method is:



The output of the spearmen method is:

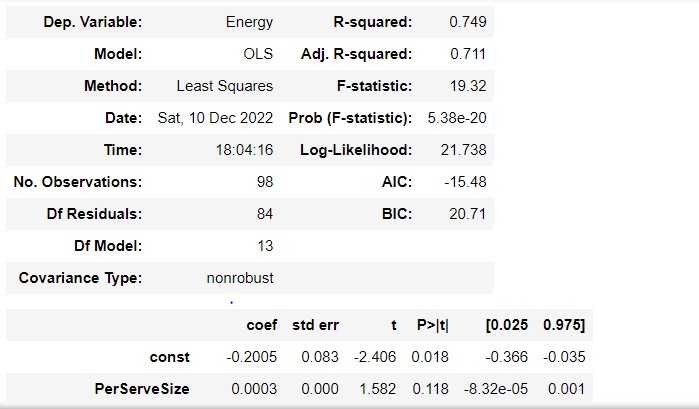


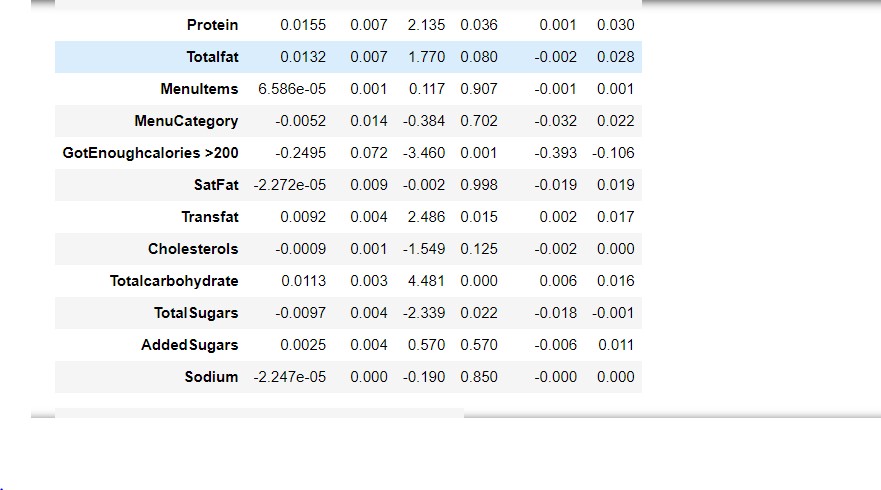
The variable that we got in the Filter Method is:

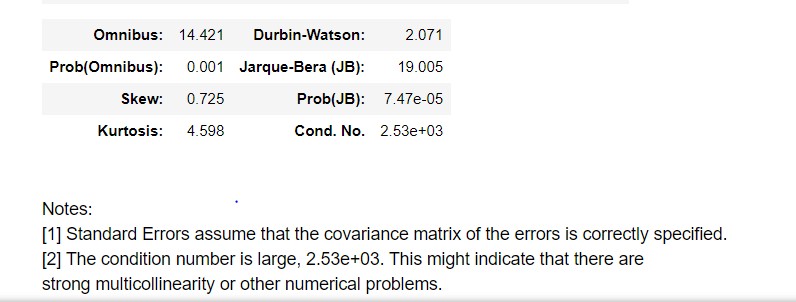
['AddedSugars', 'Sodium']

# Wrapper Method:

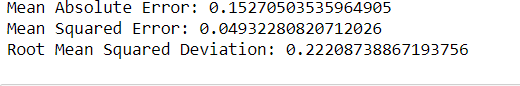
We used regression in this method for the data set. The OLS regression result is:



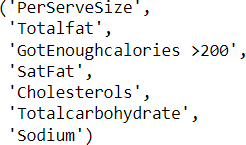




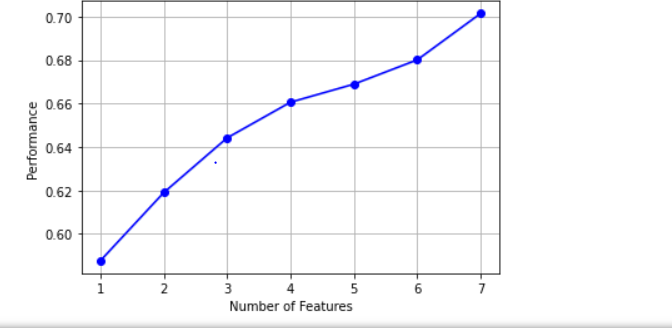
The mean absolute error values are:



The variables names that are obtained in the method is:



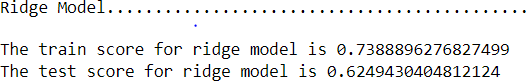
The curve obtained in this method is:



# Embedded Methods:

# Ridge Model:

The train and the test result of the ridge model is:



1. **Lasso Model:**

**Graphical user interface, text, application, email

Description automatically generated**

**Table

Description automatically generated**

**Conclusion of the Feature Selection Method:**

1. Therefore, Embedded methods i.e., Lasso and Ridge regression did not provide any information instead the scores produced are also less.
2. Wrapper Method implemented to determine the variables using sequential feature selection, but the scores determined are very less as per the regression model executed using OLS model.
3. Hence, the Filter Method is considered with the Correlation determined using Heat Map, and clearly two variables are shown i.e., "Added Sugars" & "Sodium".

**Classification Model:**

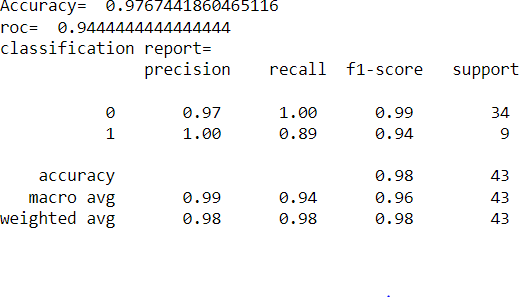
Define your X and Y: My Dependent Variable is Energy; the rest of the variables are your Independent Variables.

1. **Logistic Regression:**

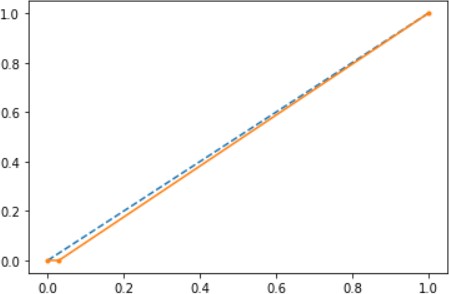
The confusion matrix of the Logistic regression model is:



The accuracy of the logistic regression model is:



The AUC curve of the logistic regression model is:

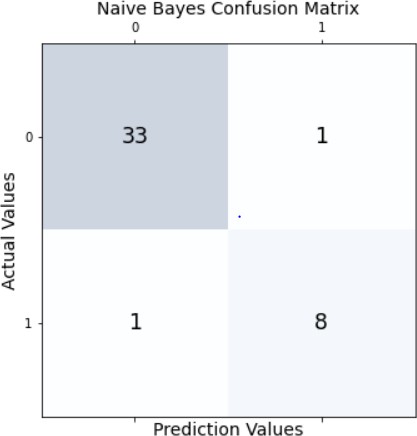


Classification Model Conclusion:

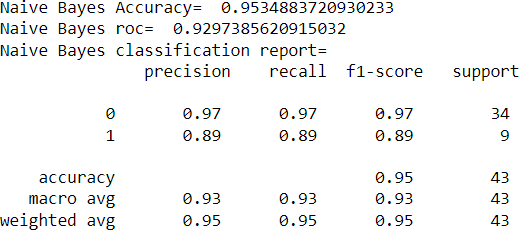
1. Accuracy: 97.6
2. Roc Value: 94.4
3. AUC Curve (Area Under Curve): 48.5

# Naïve bayes Classification model:

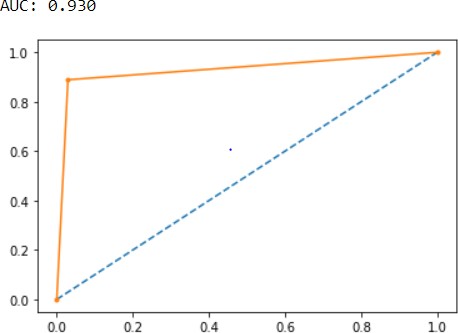
The confusion matrix of the naïve bayes algorithm is:



The accuracy of the naïve bayes algorithm is:



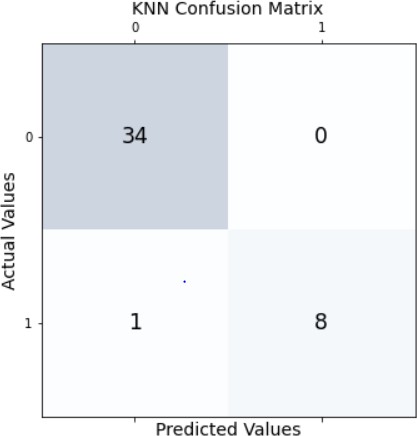
The AUC curve of the naïve bayes algorithm is:



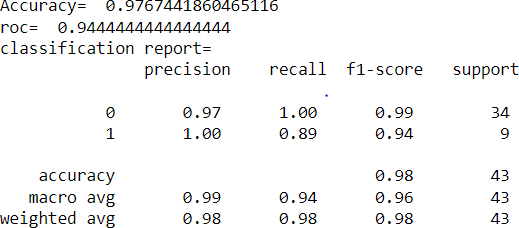
Naïve Bayes Classification Model Conclusion:

1. Accuracy: 95.34
2. Roc Value: 92.97
3. AUC Curve (Area Under Curve): 93

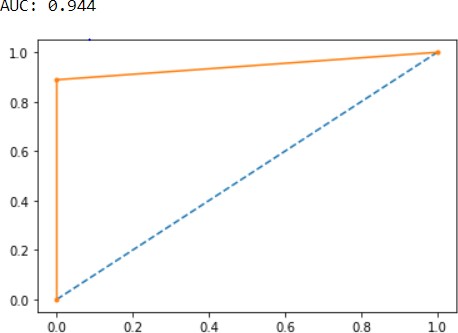
# KNN classifications model:

The confusion matrix of the KNN algorithm is:

The accuracy of the KNN algorithm is:



The AUC curve of this model is:

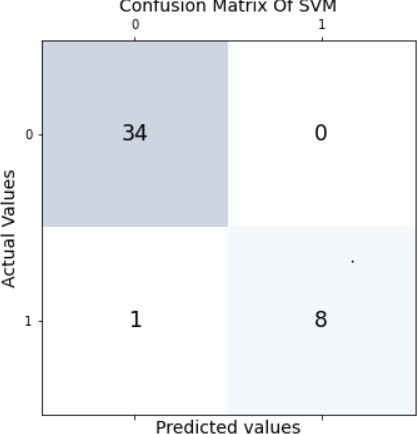


K-Nearest Neighbor Classification Model Conclusion:

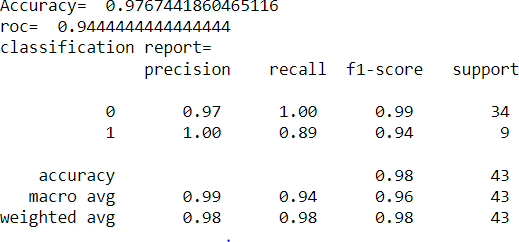
1. Accuracy: 97.67
2. Roc Value: 94.44
3. AUC Curve (Area Under Curve): 94.4

# SVM Classification Model:

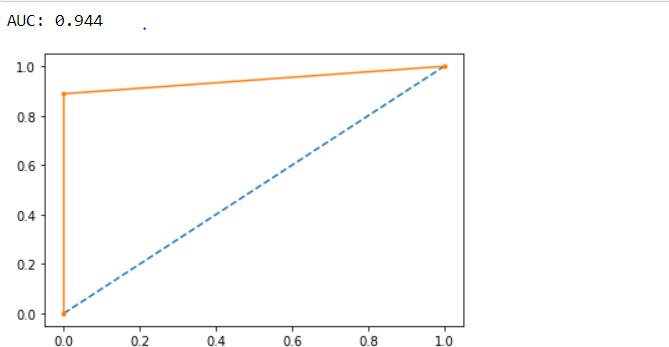
The confusion matrix of the SVM model is:



The accuracy of the SVM algorithm is:



The AUC curve of the algorithm is:



SVM Classification Model Conclusion:

1. Accuracy: 97.67
2. Roc Value: 94.44
3. AUC Curve (Area Under Curve): 94.4

# Support Vector Classifier:

The confusion matrix of the SVC algorithm is:

Graphical user interface, application, Word

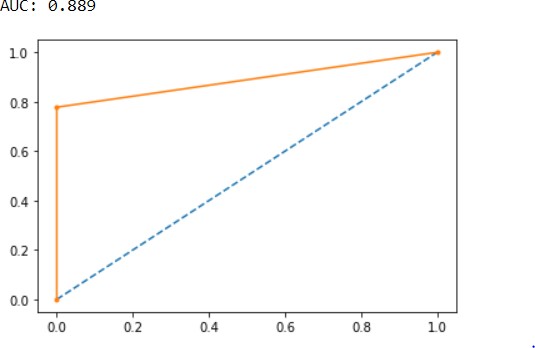
Description automatically generated

The accuracy of the algorithm is:

Table

Description automatically generated

The AUC curve of the algorithm is:



SVC Classification Model Conclusion:

1. Accuracy: 95.34
2. Roc Value: 88.88
3. AUC Curve (Area Under Curve): 88.9

# Conclusion Of The Model:

|  |  |  |  |
| --- | --- | --- | --- |
| **Models Implemented** | **Accuracy** | **ROC Value** | **AUC Curve** |
| Classification using Logistic Regression | 97.67 | 94.44 | 48.5 |
| Support Vector Machines | 97.67 | 94.44 | 94.4 |
| Naïve Bayes | 95.34 | 92.97 | 93 |
| KNN Classification | 97.67 | 94.44 | 94.4 |
| Support Vector Classifier | 95.34 | 88.88 | 88.9 |

1. **Therefore, from all the models i.e., Classification using Logistic Regression, Naive Bayes, Support Vector Classifier, K-Nearest Neighbor and Support Vector Machines the algorithm with maximum ROC value and Accuracy is given by 2 Algorithms**
2. **The best models are KNN and SVM as per the consideration of AUC curves too.**
3. **The Null Hypothesis mentioned that at least one order made by the customer that should be giving enough energy from the order placed is rejected. The Model produced is Statistically Significant as per the p values produced by the OLS Regression model.**
4. **Hence there are Multiple Attributes contributed for producing Excess number of Calories made by the customer towards Order Purchase.**