

**PROJECT TITLE**

**Data Analysis of**

**McDonalds Nutritional facts**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Topic** | **Page Number** |
| **1.** | **Problem Statement** | **3** |
| **2.** | **Problem Analysis** | **3** |
| **3.** | **Problem Design** | **4-6** |
| **4.** | **Programming Requirements** | **7-8** |
| **5.** | **Data/Input Output Description** | **9** |
| **6.** | **Algorithmic approach/Algorithm/DFD/ER diagram/Program Steps** | **10-11** |
| **7.** | **Implementation and Testing** | **12-13** |
| **8.** | **Output (Screenshots)** | **14-15** |

**Project Description**

**Problem Statement:** To analyze the McDonalds nutritional facts using statistics and probability concepts via python.

**Problem Analysis:** The dataset has the following components for analysis:

* Applying the concepts of Exploratory Data Analysis.
* Trying to find out combos of the various categories with balanced calorie content.
* Plotting PMF AND CDF graphs, modelling distribution plots for estimating probabilities.
* Analysing the sugar intake and guessing the chances of getting Diabetes via Bayes Theorem.
* Analysing the Sodium intake and guessing the chances of getting Thyroid via Bayes Theorem.
* Analysing the protein requirements that can be provide by the food in the menu.
* Analysing other essential nutrients present in the menu items.
* Plotting Scatter plot to find correlation between Carbohydrates and Cholesterol.
* Using Hypothesis testing to try to get food items with safer carbohydrate intake.

**Problem Design:** The dataset mainly focus on the nutritional value present in the menu items.

* Applying EDA concepts:-
* The dataset is first read using the pandas data frame and then validated and transformed.
* The data is checked for null values.
* The number of columns are calculated for ensuring no errors are present while importing the dataset.
* Using different plots to analyse different aspects of the dataset:-
* The barplots and scatter plots are used for determining components like Calories, Iron,Vitamins,etc.
* The plots help in establishing correlation between nutrients and cholesterol level to different negative effects from positive ones.
* The categories are analysed based on these nutrients for a better understanding of the nutritional value.
* Plotting PMF and CDF graphs and modeling distributions:-
* The PMF(Probability Mass Function) graphs are plotted in order to make the different nutrient values quantifiable and estimating their probabilities.
* The CDF(Cumulative Distribution Function) graph helps in differentiating the regions of high probabilities from that of lower probabilities.
* Modeling distributions helps in identifying the coreect form of distribution required for the data(Normal, Exponential, Lognormal, etc.).
* Applying Bayes Theorem: -
* In the imported data set Bayes Theorem is used for determining the chances of catching any severe disease like diabetes and thyroid.
* The Sodium and sugar intake of the different menu items are calculated and then compared with the facts gathered which finally gives us a probability of getting diagnosed with them due to the fast food consumption.
* Using Hypothesis Testing: -
* Hypothesis testing is used for trying to determine some safer consumption options in the McDonalds menu.
* A random sample selected from the population is analysed based on the facts gathered and using Z score as a medium to prove the hypothesis formed.

**Programming Requirements**

* **Functional Requirements:-**
* The program is constructed in **Python 3** and uses various libraries in order to execute data visualization and helps in building the data frames. The libraries used for this purpose are:
* **Matplotlib**
* **Numpy**
* **Pandas**
* **Seaborn**
* The **Thinkplot and Thinkstats2 libraries**  provided the desired functions for plotting the PMF and CDF graphs.
* The **sklearn.utils** library contains the **resample** to ensure that the selection bias is avoided and the values selected for the sample are random.
* The **scipy** library provided the modules for calculating the Z score.

**Data Input/output Description**

The program concepts of statistics and probability via Python to analyse different aspects of nutritional information provided by the data set.

* After the dataset is imported in the python environment, its goes through the validation stage to check if there are any errors during importing.
* Then it goes through the transformation and its checked for the null values and some values changed if required.
* The plots using Matplotlib and Seaborn helps in analyzing the different categories of food provided by the menu like trying to decide a calorie balanced meal, understanding the nutrients that harm and the ones that benefits.
* The thinkplot and thinkstats libraries helps in applying the probability concepts and thus making the data quantifiable for efficient analysis.
* The distribution modelling helps in understanding the properties of the data set.
* The hypothesis testing done through the z scores and p scores helps in verifying certain analysed facts.

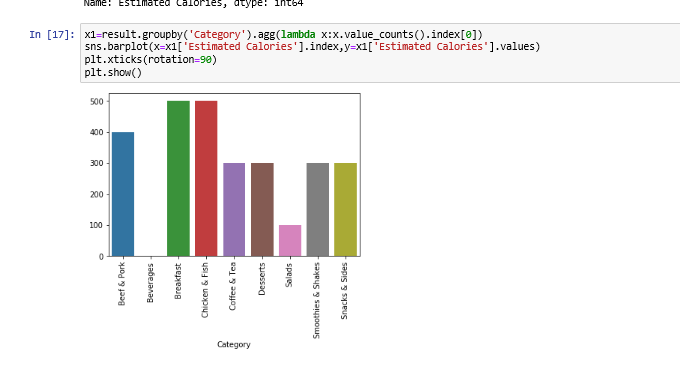
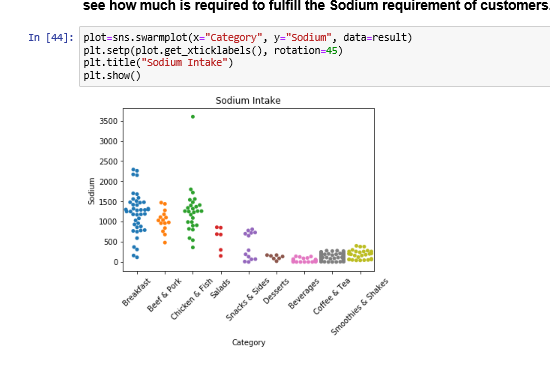
**Program Steps**

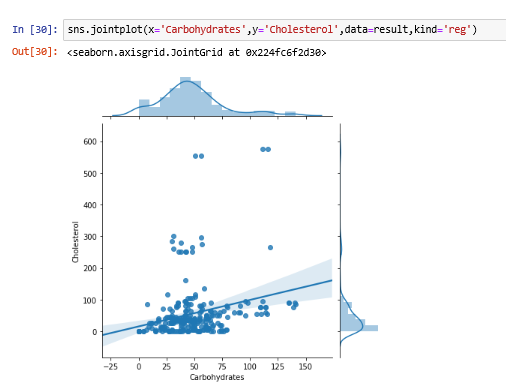
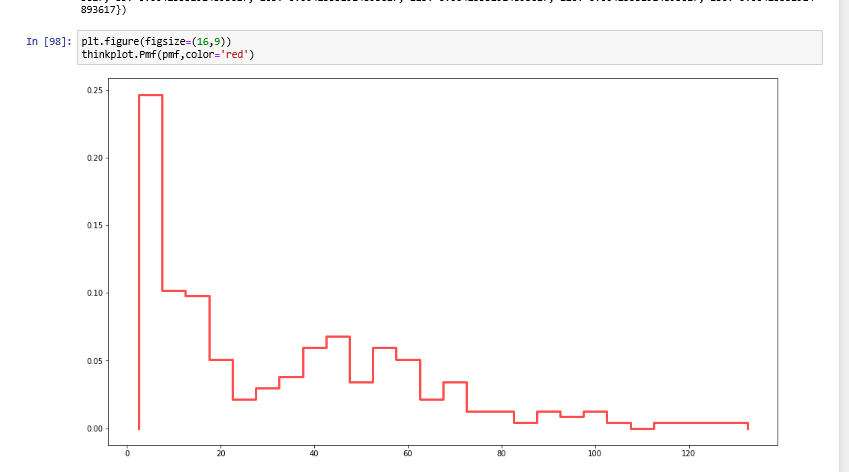
* The first step is importing the data using data frame via pandas library.
* After importing the dataset, check if there is any error while importing by checking the head and tail of the data set.
* Then check for null values and if any transformation is required in the dataset.
* Finally begin with the analysis on different components of the data set.
* Analyse the calorie content of different categories via barplot and try to make different combinations.
* Then analyse the salt and sugar intake followed by the application of bayes theorem to guess the chances of catching Thyroid or diabetes.
* The next analysis to be done on protein intake and how much requirement is the menu capable of fulfilling and analyse its distribution modeling.
* Then comes analyzing the essential nutrients like the iron, vitamins, dietary fiber, etc.
* Finally plot a scatter to analyse the correlation between the carbohydrates and cholesterol levels and verify your finding using hypothesis testing.

**Implementation and Testing**

|  |  |  |
| --- | --- | --- |
| S.No | Cases | Testing(Screenshots) |
| 1 | Ensuring the data set imported and validated. |  |
| 2 | Ensuring that the distribution modeling is accurate and correct for the dataset. |  |
| 3 | Checking for null values |  |
| 4 | Easy handling of numeric data for efficient plotting and interpretation |  |

**OUTPUT**



****

