# AIT-580 DL1 - Big Data To Information

Final Project On

# McDonald’s Nutrition Analysis

By

Sai Hemanth Nirujogi
G01065588

George Mason University

Fairfax, VA.

Sai Hemanth Nirujogi
G01065588.

## Introduction - McDonalds

McDonald's is a massive food service chain with over 34,000 restaurants in 118 countries with more than 75 million customers around the globe. With a daily consumer traffic of 62 million, they sell nearly 75 burgers per second. The company's annual revenue of $27 million and more than 750,000 employees working has made McDonald's world's leading restaurant chain. One billion pounds of beef is consumed in a year in United States alone [1].

Over the years, McDonald's has become a data centric organization that makes decisions which are data driven. McDonald's depends on its data to extract consumer and measure food quality and take decisions according to the data collected like location, supply chain and nutrition check. McDonald's shares a wide variety of data with the analysts and students to perform their own analysis on them. Users can share and publish their results with others on platforms like "Kaggle", a company which provides datasets for analysis. The dataset used in this report is obtained from Kaggle, which was shared by the company McDonalds through the website. Since Kaggle is an open source which high quality public datasets there are no high privacy policies with the dataset.

In this dataset we analyze various McDonalds categories of its menu items and the nutrition values of the food. Sharing the nutrition values like sodium, fat, cholesterol, iron, vitamins and etc of the food on their menu with the customers to balance their meal with other items has been a key highlight of how McDonald's leading the food industry game.

## Need

As Mcdonalds has become a data centric company, it has created a professional specialization teams to develop and discover new possibilities for the organization. This phase has led to ideas and development in data driven insights. McDonald's has created this dataset to provide the customers with the information that makes them take correct decisions about moderation and balance in their diet. This nutrition information has been derived from conducting tests in accredited labs, resources and from the McDonald's suppliers. All the information is derived from average values of the ingredients from the suppliers around the United States and by meeting the periodic changes in product formulations. [2] Variations in the nutrient content of products can be found in some restaurants because of the portions of ingredients used to make an item. These are questions to be answered in this dataset:

- What food items contain dangerous amounts of fat content?
- What are the nutrients contributing to the calorie intake?
- What is the lowest and highest calorie meal combination?
- Visualizations to observe the data.

## Data Description

The dataset consists of 24 attributes, with one categorical Variable, one String, one Multivalued discrete variable, ten continuous variables and eleven discrete variables. The dataset is a blend of different types of attributes:

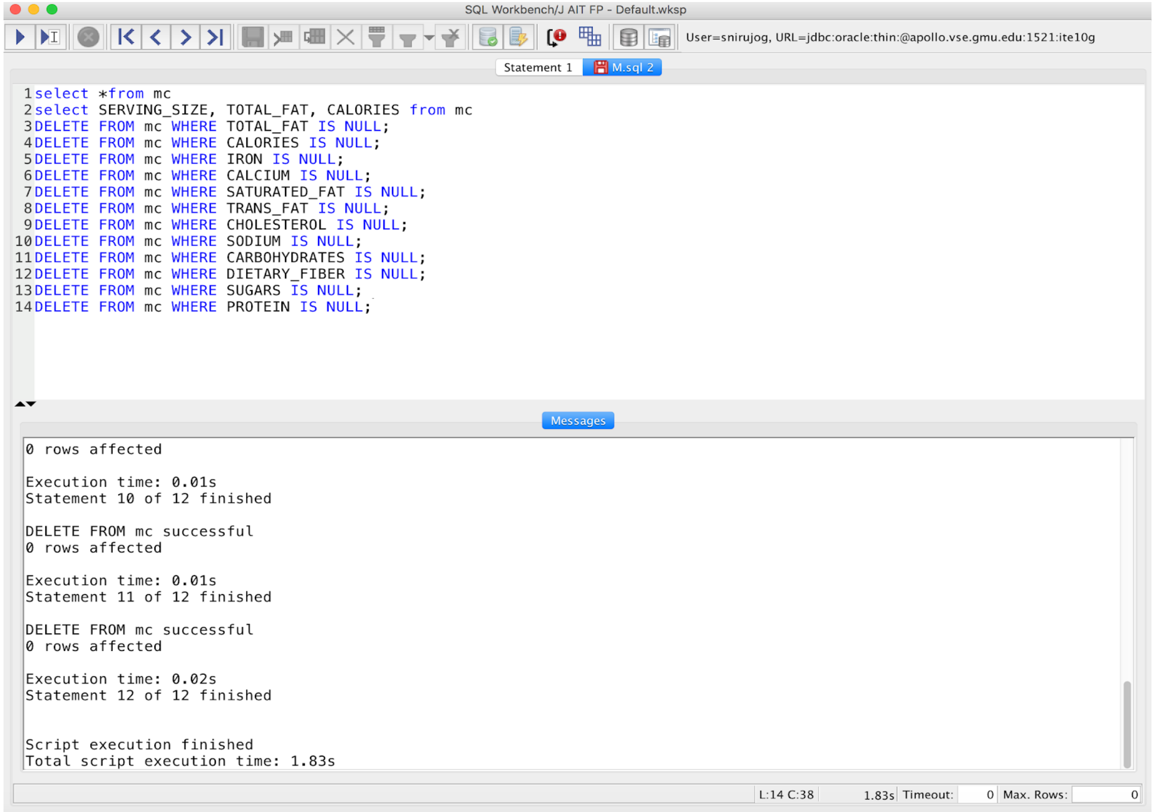
- Category - Categorical Variable
- Item - String (Unique for each instance)
- Serving Size - Multivalued Discrete
- Calories - Continuous
- Calories From Fat - Continuous
- Total Fat - Continuous
- Total Fat (% Daily Value) - Discrete
- Saturated Fat - Continuous
- Saturated Fat (% Daily Value) - Discrete
- Trans Fat - Continuous
- Cholesterol - Continuous
- Cholesterol (% Daily Value) - Discrete
- Sodium - Continuous
- Sodium (% Daily Value) - Discrete
- Carbohydrates Continuous - Discrete
- Carbohydrates (% Daily Value) - Discrete
- Dietary Fibre - Continuous
- Dietary Fibre (% Daily Value) - Discrete
- Sugars - Continuous
- Protein - Continuous
- Vitamin A (% Daily Value) - Discrete
- Vitamin C (% Daily Value) - Discrete
- Calcium (% Daily Value) - Discrete
- Iron (% Daily Value) - Discrete

## Resources and tools

To analyze this dataset open source tools and resources like R, Tableau and SQL are used. SQL is used to understand the variable types and to add missing data in the dataset. Tableau and R are used to design most of the visualizations in this project. For finding the lowest and highest calorie intake in a meal using aggregate and merge functions in R.

## Metadata using SQL

To understand the variable and the type, nutrition data is inserted into the SQL database by creating a table named “mc”. Almost all the data was clean and did not require much preprocessing. Although there are some null values and some missing information in the beverages section that needs to be added in the table. Null values from every column in the table is is removed using DELETE FROM function shown in Fig-1. Four rows of missing data has been inserted into the table shown in Fig-2 and Fig-3 shows the displayed rows in the table.

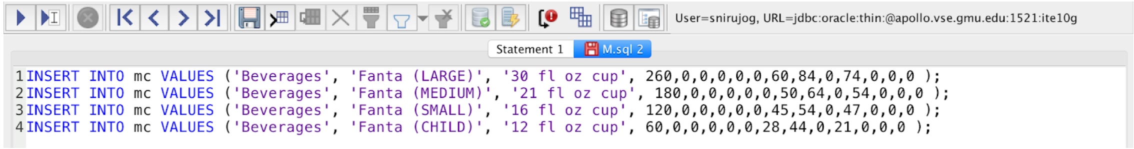


User=snirujog, URL=jdbc:oracle:thin:@apollo.vse.gmu.edu:1521:ite10g

```
select *from mc
select SERVING_SIZE, TOTAL_FAT, CALORIES from mc
DELETE FROM mc WHERE TOTAL_FAT IS NULL;
DELETE FROM mc WHERE CALORIES IS NULL;
DELETE FROM mc WHERE IRON IS NULL;
DELETE FROM mc WHERE CALCIUM IS NULL;
DELETE FROM mc WHERE SATURATED_FAT IS NULL;
DELETE FROM mc WHERE TRANS_FAT IS NULL;
DELETE FROM mc WHERE CHOLESTEROL IS NULL;
DELETE FROM mc WHERE SODIUM IS NULL;
DELETE FROM mc WHERE CARBOHYDRATES IS NULL;
DELETE FROM mc WHERE DIETARY_FIBER IS NULL;
DELETE FROM mc WHERE SUGARS IS NULL;
DELETE FROM mc WHERE PROTEIN IS NULL;
```

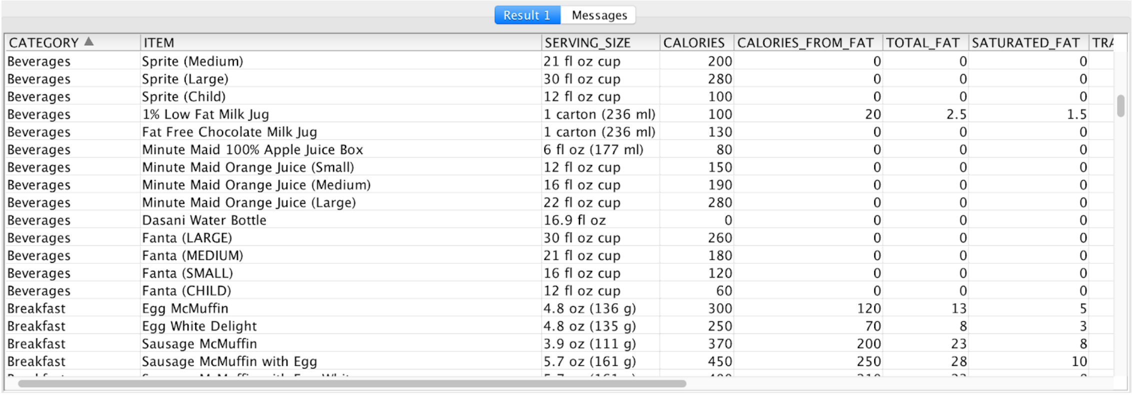
```
0 rows affected

Execution time: 0.01s
Statement 10 of 12 finished

DELETE FROM mc successful
0 rows affected

Execution time: 0.01s
Statement 11 of 12 finished

DELETE FROM mc successful
0 rows affected

Execution time: 0.02s
Statement 12 of 12 finished


Script execution finished
Total script execution time: 1.83s
```

*Fig-1 Removing null values from the table*


User=snirujog, URL=jdbc:oracle:thin:@apollo.vse.gmu.edu:1521:ite10g
Statement 1
M.sql 2

```sql
INSERT INTO mc VALUES ('Beverages', 'Fanta (LARGE)', '30 fl oz cup', 260,0,0,0,0,0,60,84,0,74,0,0,0 );
INSERT INTO mc VALUES ('Beverages', 'Fanta (MEDIUM)', '21 fl oz cup', 180,0,0,0,0,0,50,64,0,54,0,0,0 );
INSERT INTO mc VALUES ('Beverages', 'Fanta (SMALL)', '16 fl oz cup', 120,0,0,0,0,0,45,54,0,47,0,0,0 );
INSERT INTO mc VALUES ('Beverages', 'Fanta (CHILD)', '12 fl oz cup', 60,0,0,0,0,0,28,44,0,21,0,0,0 );
```

*Fig-2 Inserting values into the table*

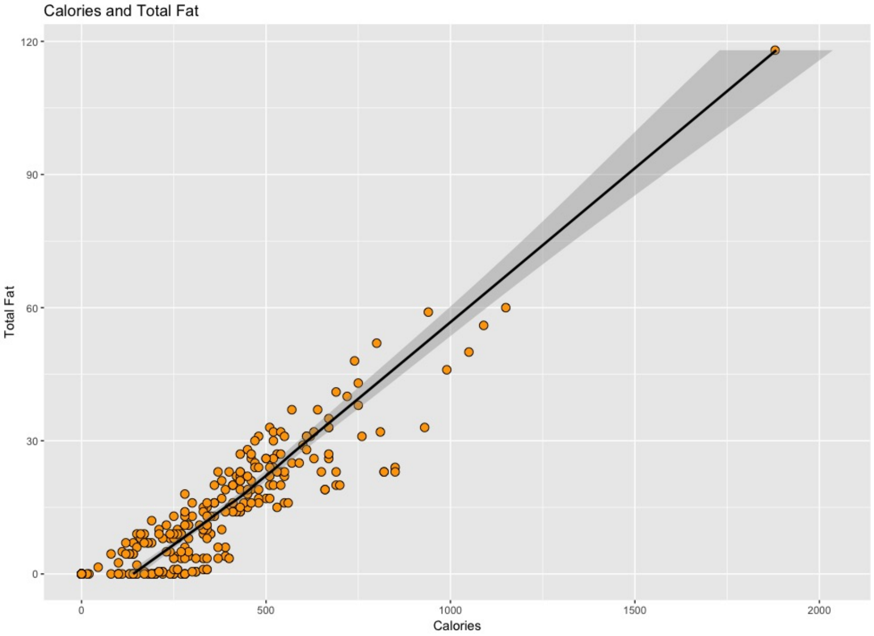


| CATEGORY ▲ | ITEM | SERVING_SIZE | CALORIES | CALORIES_FROM_FAT | TOTAL_FAT | SATURATED_FAT | TRA |
|---|---|---|---|---|---|---|---|
| Beverages | Sprite (Medium) | 21 fl oz cup | 200 | 0 | 0 | 0 | |
| Beverages | Sprite (Large) | 30 fl oz cup | 280 | 0 | 0 | 0 | |
| Beverages | Sprite (Child) | 12 fl oz cup | 100 | 0 | 0 | 0 | |
| Beverages | 1% Low Fat Milk Jug | 1 carton (236 ml) | 100 | 20 | 2.5 | 1.5 | |
| Beverages | Fat Free Chocolate Milk Jug | 1 carton (236 ml) | 130 | 0 | 0 | 0 | |
| Beverages | Minute Maid 100% Apple Juice Box | 6 fl oz (177 ml) | 80 | 0 | 0 | 0 | |
| Beverages | Minute Maid Orange Juice (Small) | 12 fl oz cup | 150 | 0 | 0 | 0 | |
| Beverages | Minute Maid Orange Juice (Medium) | 16 fl oz cup | 190 | 0 | 0 | 0 | |
| Beverages | Minute Maid Orange Juice (Large) | 22 fl oz cup | 280 | 0 | 0 | 0 | |
| Beverages | Dasani Water Bottle | 16.9 fl oz | 0 | 0 | 0 | 0 | |
| Beverages | Fanta (LARGE) | 30 fl oz cup | 260 | 0 | 0 | 0 | |
| Beverages | Fanta (MEDIUM) | 21 fl oz cup | 180 | 0 | 0 | 0 | |
| Beverages | Fanta (SMALL) | 16 fl oz cup | 120 | 0 | 0 | 0 | |
| Beverages | Fanta (CHILD) | 12 fl oz cup | 60 | 0 | 0 | 0 | |
| Breakfast | Egg McMuffin | 4.8 oz (136 g) | 300 | 120 | 13 | 5 | |
| Breakfast | Egg White Delight | 4.8 oz (135 g) | 250 | 70 | 8 | 3 | |
| Breakfast | Sausage McMuffin | 3.9 oz (111 g) | 370 | 200 | 23 | 8 | |
| Breakfast | Sausage McMuffin with Egg | 5.7 oz (161 g) | 450 | 250 | 28 | 10 | |

*Fig-3 Displaying the inserted rows*

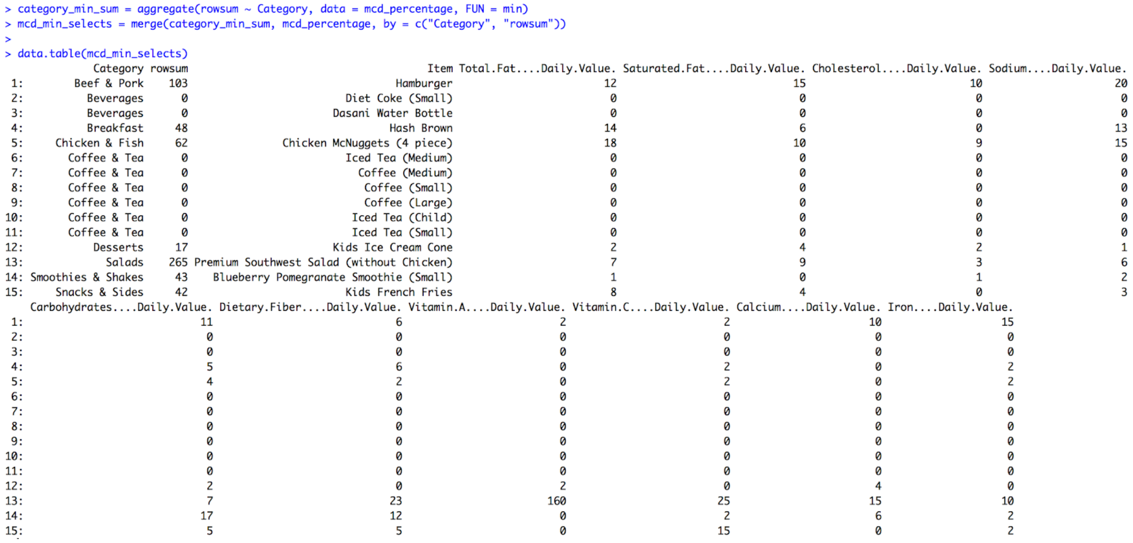
## Linear Regression and Analysis using R

To find out the predictors of nutrition in the menu, we are applying linear regression. The graph obtained from the model shows that calories and total fat are the predictors for the nutrition values.



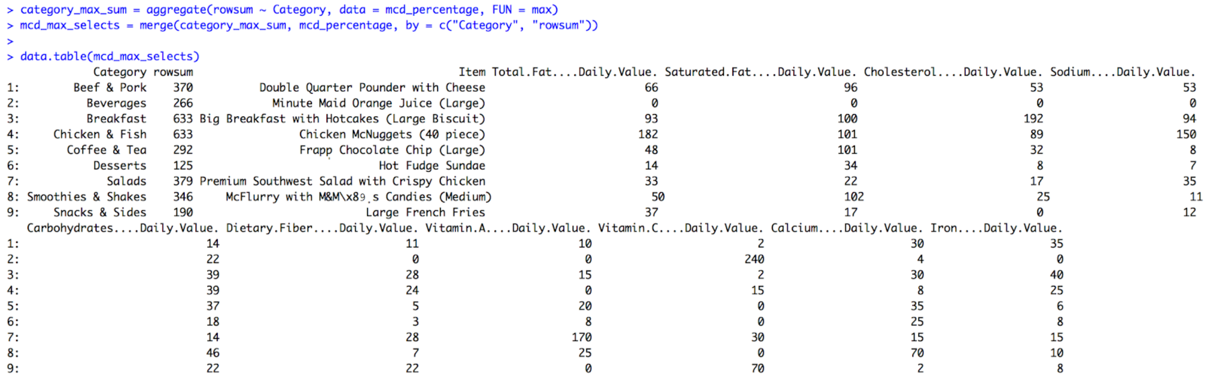
From the graph, we can see that the highest number of calories in an item is 1880 with total fat content of 118, the item is “Chicken McNuggets  (40 piece)”

To find out the lowest and highest calorie meal combination, we used the aggregate and merge function in R. Aggregate function in R parts the information into subsets, processes outline measurements for each, and restores the outcome in a user friendly shape. The rows in categories are added together to find out the total daily value of items and that can be used to find the minimum and maximum calorie meal combination. To order meal with minimum calorie count customer can choose from the items shown in Fig-4 and to order meal with maximum calorie count one can choose items from the output table in Fig-5.



```
> category_min_sum = aggregate(rowsum ~ Category, data = mcd_percentage, FUN = min)
> mcd_min_selects = merge(category_min_sum, mcd_percentage, by = c("Category", "rowsum"))
>
> data.table(mcd_min_selects)
              Category rowsum                                    Item Total.Fat....Daily.Value. Saturated.Fat....Daily.Value. Cholesterol....Daily.Value. Sodium....Daily.Value.
 1:        Beef & Pork  103                                Hamburger                        12                            15                          10                     20
 2:          Beverages    0                        Diet Coke (Small)                         0                             0                           0                      0
 3:          Beverages    0                      Dasani Water Bottle                         0                             0                           0                      0
 4:          Breakfast   48                               Hash Brown                        14                             6                           0                     13
 5:     Chicken & Fish   62              Chicken McNuggets (4 piece)                        18                            10                           9                     15
 6:       Coffee & Tea    0                        Iced Tea (Medium)                         0                             0                           0                      0
 7:       Coffee & Tea    0                          Coffee (Medium)                         0                             0                           0                      0
 8:       Coffee & Tea    0                           Coffee (Small)                         0                             0                           0                      0
 9:       Coffee & Tea    0                           Coffee (Large)                         0                             0                           0                      0
10:       Coffee & Tea    0                         Iced Tea (Child)                         0                             0                           0                      0
11:       Coffee & Tea    0                         Iced Tea (Small)                         0                             0                           0                      0
12:           Desserts   17                      Kids Ice Cream Cone                         2                             4                           2                      1
13:             Salads  265 Premium Southwest Salad (without Chicken)                        7                             9                           3                      6
14: Smoothies & Shakes   43   Blueberry Pomegranate Smoothie (Small)                         1                             0                           1                      2
15:     Snacks & Sides   42                        Kids French Fries                         8                             4                           0                      3
    Carbohydrates....Daily.Value. Dietary.Fiber....Daily.Value. Vitamin.A....Daily.Value. Vitamin.C....Daily.Value. Calcium....Daily.Value. Iron....Daily.Value.
 1:                            11                             6                         2                         2                      10                   15
 2:                             0                             0                         0                         0                       0                    0
 3:                             0                             0                         0                         0                       0                    0
 4:                             5                             6                         0                         2                       0                    2
 5:                             4                             2                         0                         2                       0                    2
 6:                             0                             0                         0                         0                       0                    0
 7:                             0                             0                         0                         0                       0                    0
 8:                             0                             0                         0                         0                       0                    0
 9:                             0                             0                         0                         0                       0                    0
10:                             0                             0                         0                         0                       0                    0
11:                             0                             0                         0                         0                       0                    0
12:                             2                             0                         2                         0                       4                    0
13:                             7                            23                       160                        25                      15                   10
14:                            17                            12                         0                         2                       6                    2
15:                             5                             5                         0                        15                       0                    2
```

*Fig-4 Lowest calorie meal combination*



```
> category_max_sum = aggregate(rowsum ~ Category, data = mcd_percentage, FUN = max)
> mcd_max_selects = merge(category_max_sum, mcd_percentage, by = c("Category", "rowsum"))
>
> data.table(mcd_max_selects)
             Category rowsum                                        Item Total.Fat....Daily.Value. Saturated.Fat....Daily.Value. Cholesterol....Daily.Value. Sodium....Daily.Value.
1:        Beef & Pork   370          Double Quarter Pounder with Cheese                        66                            96                          53                     53
2:          Beverages   266            Minute Maid Orange Juice (Large)                         0                             0                           0                      0
3:          Breakfast   633 Big Breakfast with Hotcakes (Large Biscuit)                        93                           100                         192                     94
4:     Chicken & Fish   633                Chicken McNuggets (40 piece)                       182                           101                          89                    150
5:       Coffee & Tea   292                Frapp Chocolate Chip (Large)                        48                           101                          32                      8
6:           Desserts   125                            Hot Fudge Sundae                        14                            34                           8                      7
7:             Salads   379 Premium Southwest Salad with Crispy Chicken                        33                            22                          17                     35
8: Smoothies & Shakes   346       McFlurry with M&M\x89‚s Candies (Medium)                         50                           102                          25                     11
9:     Snacks & Sides   190                          Large French Fries                        37                            17                           0                     12
   Carbohydrates....Daily.Value. Dietary.Fiber....Daily.Value. Vitamin.A....Daily.Value. Vitamin.C....Daily.Value. Calcium....Daily.Value. Iron....Daily.Value.
1:                            14                            11                        10                         2                      30                   35
2:                            22                             0                         0                       240                       4                    0
3:                            39                            28                        15                         2                      30                   40
4:                            39                            24                         0                        15                       8                   25
5:                            37                             5                        20                         0                      35                    6
6:                            18                             3                         8                         0                      25                    8
7:                            14                            28                       170                        30                      15                   15
8:                            46                             7                        25                         0                      70                   10
9:                            22                            22                         0                        70                       2                    8
```

*Fig-5 Highest calorie meal combination*

## Visualizations

### Correlation

This visualization on the dataset is conducted using Tableau and R. As observed from the relationship plots Fig-6 , one would already be able to see that clearly tie into each other (the more yellow segments of the plot). For instance serving size and calories. The size of the rectangles shows how strong the attributes are correlated and the color shows if they are positively or negatively correlated.

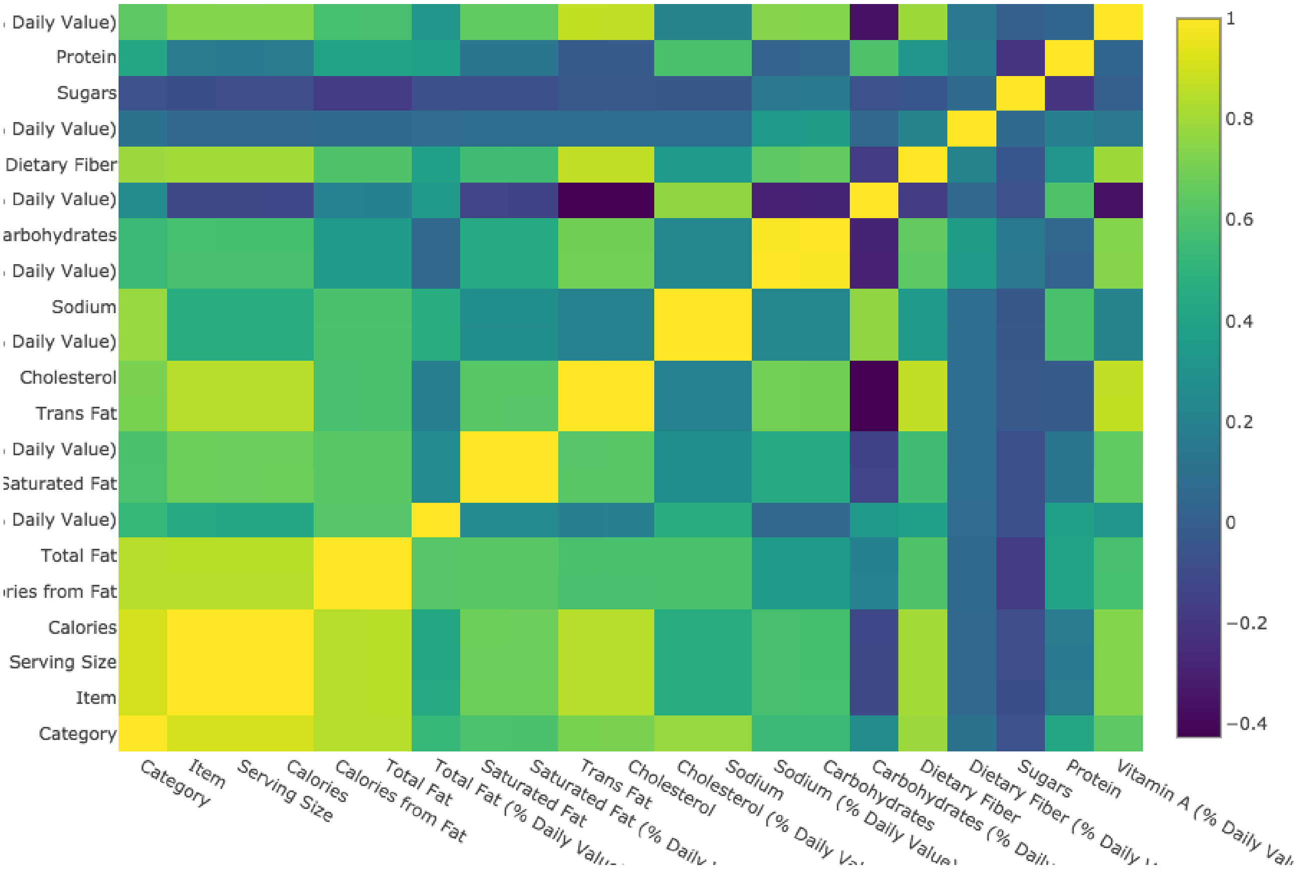


*Fig-6 Correlation of all nutrition metrics*

We can observe there are some pretty weak correlations like one between total fats and trans/Saturated fats. The main motive of creating a correlation graph is to check if there are any issues with the data quality. Carbohydrates section is negatively correlated to the other sections, which might rise a question regarding the data quality. But carbohydrates loaded foods does not have any much nutrients in them except from carbohydrates, which justifies the negative correlations.

## Bar Graphs

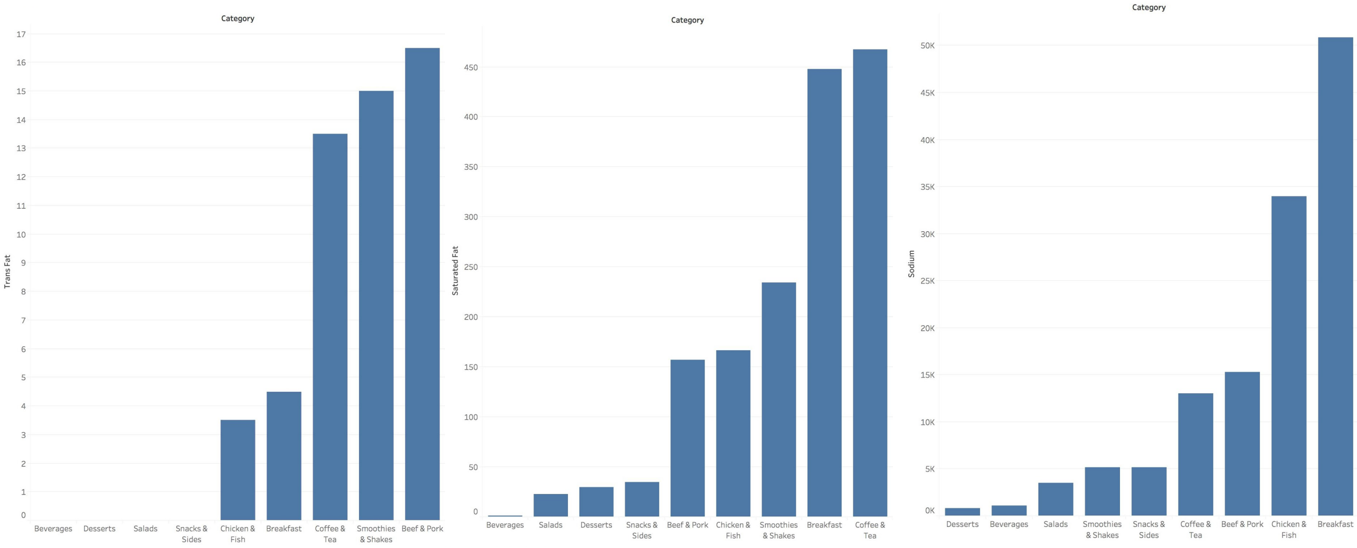


*Fig-7 Bar Graphs showing categories with Trans fat, Saturated fat and Sodium content in ascending order*

From the above visualizations we can see the categories which have high or low fat and sodium content. The first bar plot from the left shows the section containing highest trans fat, which is Beef and pork. Coffee and tea have the high saturated fats and with breakfast being almost equal to coffee. Beverages contains of the least amount of saturated fats among all categories. Breakfast leads the sodium content and desserts being the least.

From the plot, we can observe the sections with high and low calorie, carbohydrates and Total Fat. Coffee & Tea leads the charts on both calorie and carbohydrates content. The section with high total fat content is breakfast. Beverages have the least total fat content of all the other categories. Desserts have the least carbohydrate content.

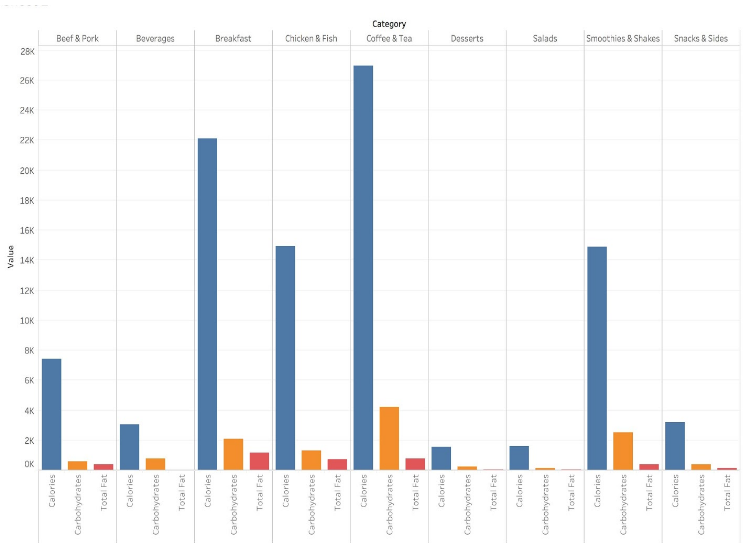


*Fig-8 Histogram showing calorie, total fat and carbs content in each category*

## Scatter Plots

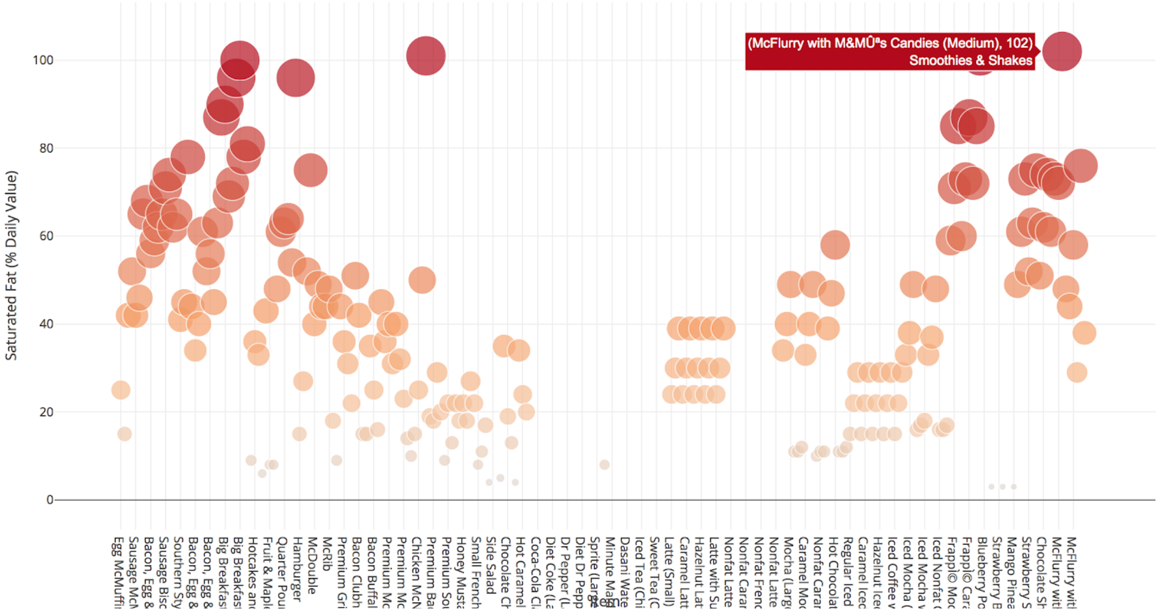


*Fig-9 Saturated Fat (% Daily Value) per item on menu*

The plot Fig-9 shows only a few of McDonald's food items that comprise dangerous amounts of saturated fat content. The graph depicts that a food item might contain saturated fat closer to a customer's daily recommended allowance.

The items which contain a higher amount of saturated fats are **McFlurry with M&M candies, Frappe Chocolate Chip, Chicken McNuggets (40 piece) and Big breakfast with Hotcakes etc.**

The plot Fig-10 shows the items with dangerous amount of cholesterol contents. The circles size and color varies with high or low cholesterol (%daily value). The item with most Cholesterol content is **Big Breakfast with Hotcakes (Large Biscuits**)

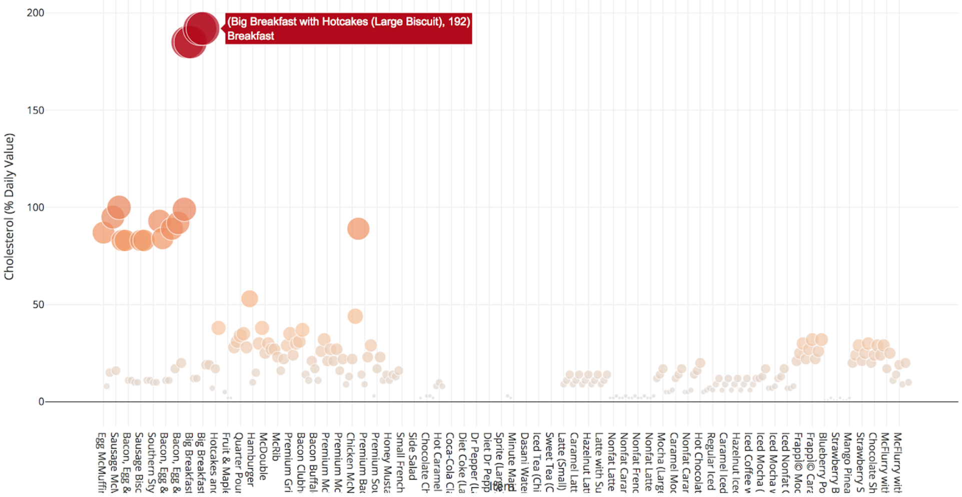


*Fig-10 Cholesterol (% daily value) per item on menu*

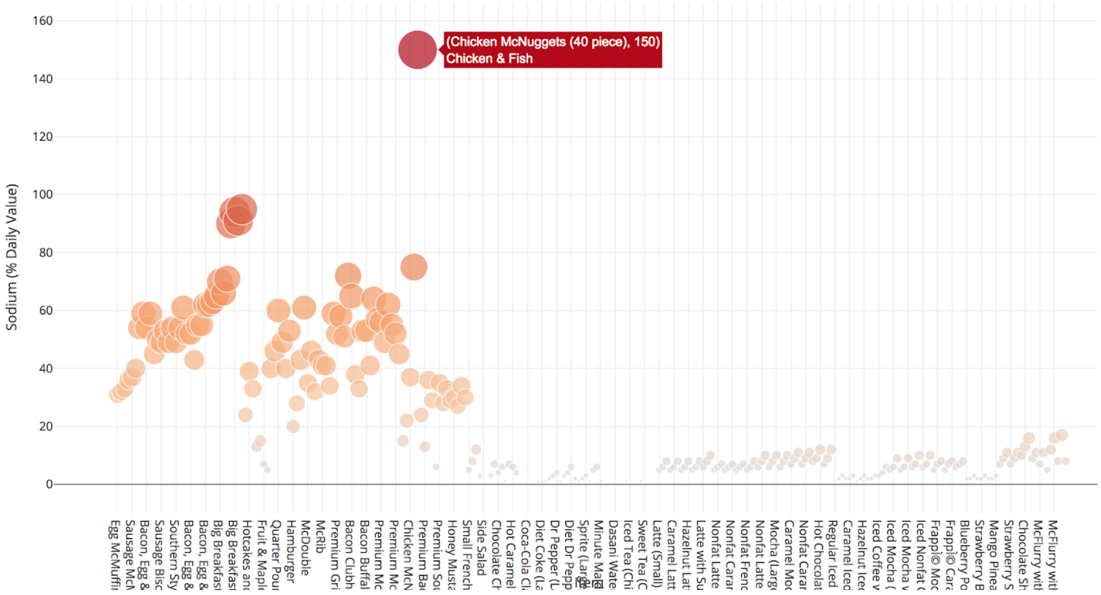


*Fig-11 Sodium (% Daily Value) per item on menu*

The large dark red circle depicts the item with highest sodium content. This plot with sodium (% daily value) has a similar point distribution as of the cholesterol.

The item with highest sodium intake is **Chicken McNuggets (40 piece).** With the Big Breakfast with Hotcakes as near second.

### Comparative Linear Regression

R is used to visualize the nutrients which contribute to the calories. Using the Linear regression and comparing with other nutrients gave us the plot.

From this plot we can see that cholesterol, carbohydrates, sugars, saturated fat and protein are the nutrients contributing to calories.

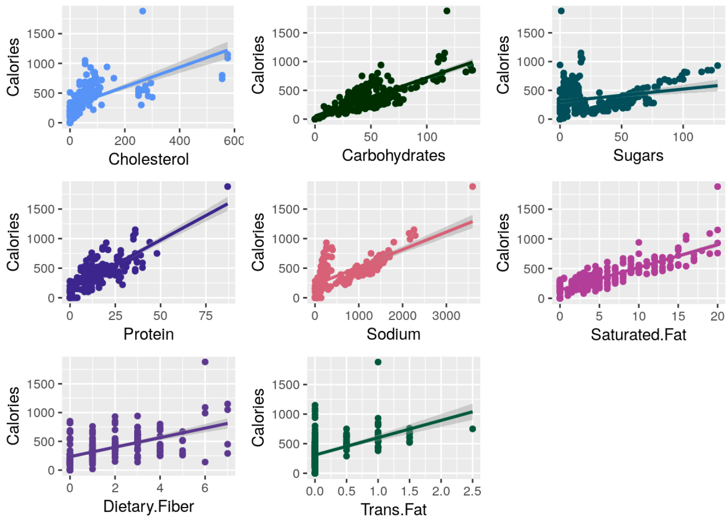


*Fig-12 Nutrients contributing to calories*

## Conclusion

By analyzing the dataset and with some data exploration using tableau and R, here are the some of the findings. With the help of R, linear regression models are created and the nutrients which contribute to the calories of the food items are cholesterol, carbohydrates, sugars, saturated fat and protein. The food items with dangerous amounts of Saturated fat, cholesterol and sodium content in McDonald's menu are found. The lowest calorie meal combination is found using R, which helps the customer to order a meal according to their daily intake. This analysis of McDonald's menu will help the customers to choose what food items are the best fit for their health and diet. This can help parents to choose the food items that's best for their children which can prevent common health problems like obesity.

## Explanation of terms

**Linear Regression:** This is Statistical method that allows to summarize and understand the relationship between two or more variables that are correlated with each other [3]

**Correlation:** It is calculated to find the relation between two variables.

## Citations

1. McDonald's. Nutrition Facts for McDonald's Menu | Kaggle, 3 Mar. 2017, www.kaggle.com/mcdonalds/nutrition-facts
2. "From Big Data to Big Mac; how McDonalds leverages Big Data." *Datafloq - Connecting Data and People*, datafloq.com/read/from-big-data-to-big-mac-how-mcdonalds-leverages-b/403
3. "Lesson 1: Simple Linear Regression." *Lesson 1: Simple Linear Regression | STAT 501*, https://onlinecourses.science.psu.edu/stat501/node/250