McDonalds_Menu - Linear Regression

March 3, 2018

1 McDonald's Menu Dataset

The data consists of nutrients value of all the items that are served @McDonald's Outlets. The dataset is taken from kaggle.

2 Load data

```
In [1]: import numpy as np
        import pandas as pd
        data = pd.read_csv("C:/Users/Aravind/Documents/menu.csv")
        data.head()
```

```
Out[1]:    Category                                Item    Serving Size  Calories  \
        0  Breakfast                        Egg McMuffin  4.8 oz (136 g)       300
        1  Breakfast                   Egg White Delight  4.8 oz (135 g)       250
        2  Breakfast                    Sausage McMuffin  3.9 oz (111 g)       370
        3  Breakfast           Sausage McMuffin with Egg  5.7 oz (161 g)       450
        4  Breakfast  Sausage McMuffin with Egg Whites  5.7 oz (161 g)       400

           Calories from Fat  Total Fat  Total Fat (% Daily Value)  Saturated Fat  \
        0                120       13.0                         20            5.0
        1                 70        8.0                         12            3.0
        2                200       23.0                         35            8.0
        3                250       28.0                         43           10.0
        4                210       23.0                         35            8.0

           Saturated Fat (% Daily Value)  Trans Fat          ...             \
        0                             25        0.0          ...
        1                             15        0.0          ...
        2                             42        0.0          ...
        3                             52        0.0          ...
        4                             42        0.0          ...

           Carbohydrates  Carbohydrates (% Daily Value)  Dietary Fiber  \
        0             31                             10              4
        1             30                             10              4
```

```
2            29                              10               4
3            30                              10               4
4            30                              10               4

   Dietary Fiber (% Daily Value)  Sugars  Protein  Vitamin A (% Daily Value)  \
0                             17       3       17                         10
1                             17       3       18                          6
2                             17       2       14                          8
3                             17       2       21                         15
4                             17       2       21                          6

   Vitamin C (% Daily Value)  Calcium (% Daily Value)  Iron (% Daily Value)
0                          0                       25                    15
1                          0                       25                     8
2                          0                       25                    10
3                          0                       30                    15
4                          0                       25                    10

[5 rows x 24 columns]
```

```
In [2]: data.shape
```

```
Out[2]: (260, 24)
```

There are 260 rows and 24 columns # Check for missing values

```
In [3]: data.isnull().sum()
```

```
Out[3]: Category                         0
        Item                             0
        Serving Size                     0
        Calories                         0
        Calories from Fat                0
        Total Fat                        0
        Total Fat (% Daily Value)        0
        Saturated Fat                    0
        Saturated Fat (% Daily Value)    0
        Trans Fat                        0
        Cholesterol                      0
        Cholesterol (% Daily Value)      0
        Sodium                           0
        Sodium (% Daily Value)           0
        Carbohydrates                    0
        Carbohydrates (% Daily Value)    0
        Dietary Fiber                    0
        Dietary Fiber (% Daily Value)    0
        Sugars                           0
        Protein                          0
        Vitamin A (% Daily Value)        0
        Vitamin C (% Daily Value)        0
        Calcium (% Daily Value)          0
        Iron (% Daily Value)             0
        dtype: int64
```

There are no missing values # Describe data

```
In [4]: data.describe()
```

```
Out[4]:           Calories  Calories from Fat   Total Fat  Total Fat (% Daily Value)  \
        count   260.000000         260.000000  260.000000                 260.000000
        mean    368.269231         127.096154   14.165385                  21.815385
        std     240.269886         127.875914   14.205998                  21.885199
        min       0.000000           0.000000    0.000000                   0.000000
        25%     210.000000          20.000000    2.375000                   3.750000
        50%     340.000000         100.000000   11.000000                  17.000000
        75%     500.000000         200.000000   22.250000                  35.000000
        max    1880.000000        1060.000000  118.000000                 182.000000

               Saturated Fat  Saturated Fat (% Daily Value)   Trans Fat  Cholesterol  \
        count     260.000000                     260.000000  260.000000   260.000000
        mean        6.007692                      29.965385    0.203846    54.942308
        std         5.321873                      26.639209    0.429133    87.269257
        min         0.000000                       0.000000    0.000000     0.000000
        25%         1.000000                       4.750000    0.000000     5.000000
        50%         5.000000                      24.000000    0.000000    35.000000
        75%        10.000000                      48.000000    0.000000    65.000000
        max        20.000000                     102.000000    2.500000   575.000000

               Cholesterol (% Daily Value)       Sodium          ...            \
        count                   260.000000   260.000000          ...
        mean                     18.392308   495.750000          ...
        std                      29.091653   577.026323          ...
        min                       0.000000     0.000000          ...
        25%                       2.000000   107.500000          ...
        50%                      11.000000   190.000000          ...
        75%                      21.250000   865.000000          ...
        max                     192.000000  3600.000000          ...

               Carbohydrates  Carbohydrates (% Daily Value)  Dietary Fiber  \
        count     260.000000                     260.000000     260.000000
        mean       47.346154                      15.780769       1.630769
        std        28.252232                       9.419544       1.567717
        min         0.000000                       0.000000       0.000000
        25%        30.000000                      10.000000       0.000000
        50%        44.000000                      15.000000       1.000000
        75%        60.000000                      20.000000       3.000000
        max       141.000000                      47.000000       7.000000
```

```
       Dietary Fiber (% Daily Value)      Sugars     Protein  \
count                     260.000000  260.000000  260.000000
mean                        6.530769   29.423077   13.338462
std                         6.307057   28.679797   11.426146
min                         0.000000    0.000000    0.000000
25%                         0.000000    5.750000    4.000000
50%                         5.000000   17.500000   12.000000
75%                        10.000000   48.000000   19.000000
max                        28.000000  128.000000   87.000000

       Vitamin A (% Daily Value)  Vitamin C (% Daily Value)  \
count                 260.000000                 260.000000
mean                   13.426923                   8.534615
std                    24.366381                  26.345542
min                     0.000000                   0.000000
25%                     2.000000                   0.000000
50%                     8.000000                   0.000000
75%                    15.000000                   4.000000
max                   170.000000                 240.000000

       Calcium (% Daily Value)  Iron (% Daily Value)
count               260.000000            260.000000
mean                 20.973077              7.734615
std                  17.019953              8.723263
min                   0.000000              0.000000
25%                   6.000000              0.000000
50%                  20.000000              4.000000
75%                  30.000000             15.000000
max                  70.000000             40.000000

[8 rows x 21 columns]
```

Some variables with daily values and item size, category, etc are not necessary for us. lets drop them.

```
In [5]: dataset = data.loc[:,['Calories','Total Fat','Cholesterol','Sodium','Carbohydrates','D
        dataset.head()
```

```
Out[5]:    Calories  Total Fat  Cholesterol  Sodium  Carbohydrates  Dietary Fiber  \
        0       300       13.0          260     750             31              4
        1       250        8.0           25     770             30              4
        2       370       23.0           45     780             29              4
        3       450       28.0          285     860             30              4
        4       400       23.0           50     880             30              4

           Sugars  Protein
        0       3       17
```

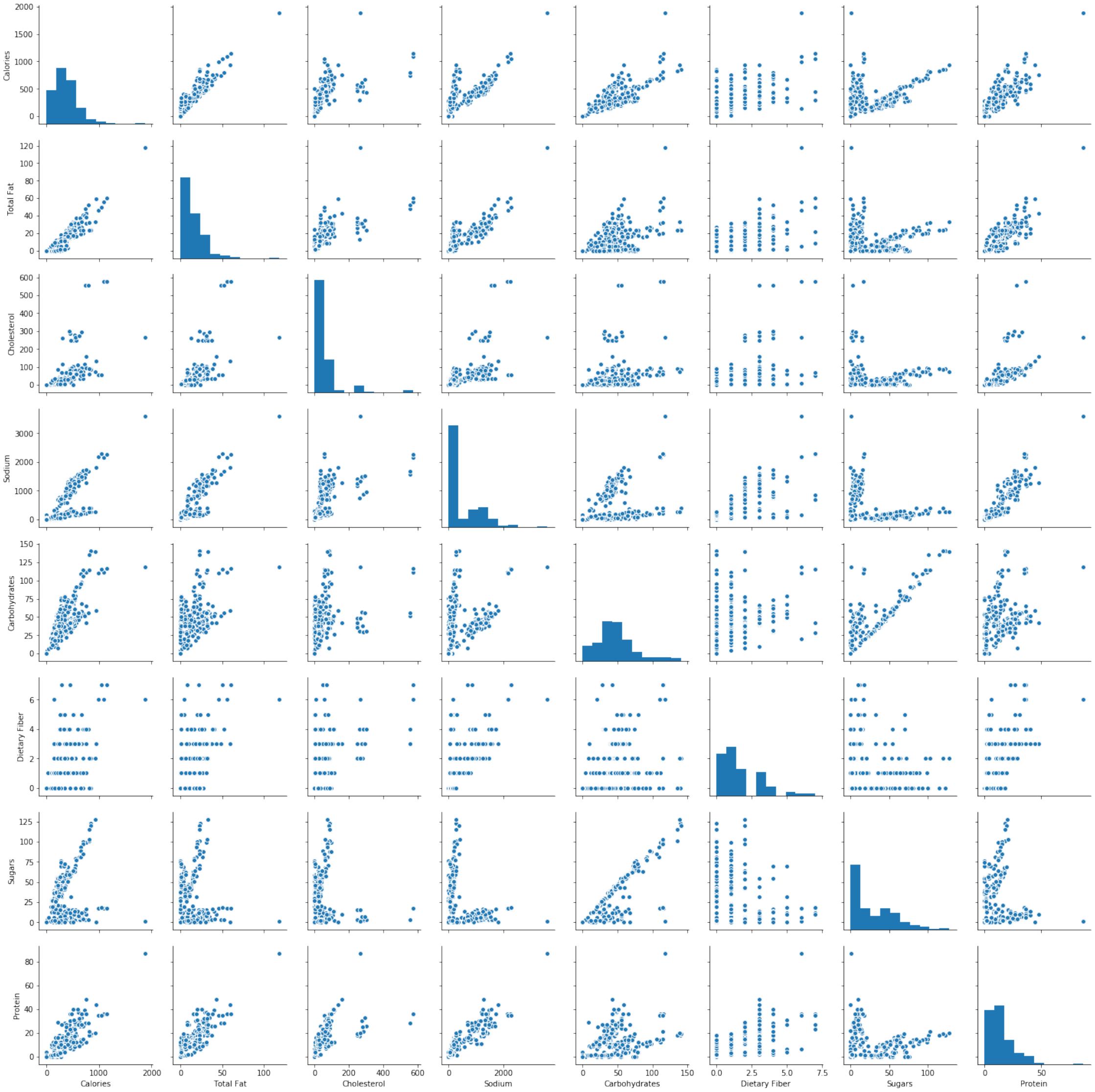
```
1         3         18
2         2         14
3         2         21
4         2         21
```

3 Visualization

Scatter plot

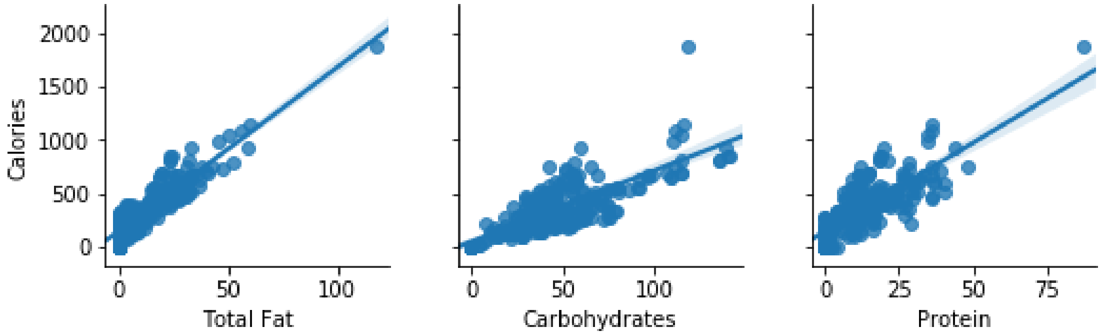
```
In [6]: import matplotlib as mpl
        import matplotlib.pyplot as plt
        import seaborn as sns
        %matplotlib inline
        sns.pairplot(dataset)
```

```
Out[6]: <seaborn.axisgrid.PairGrid at 0x1ff60bbf98>
```



```
In [43]: sns.pairplot(dataset, x_vars=['Total Fat','Carbohydrates','Protein'],y_vars='Calories
```

```
Out[43]: <seaborn.axisgrid.PairGrid at 0x4af29ab780>
```



From scatter plots we can see linear relationships between calories and other variable Lets see corrplots

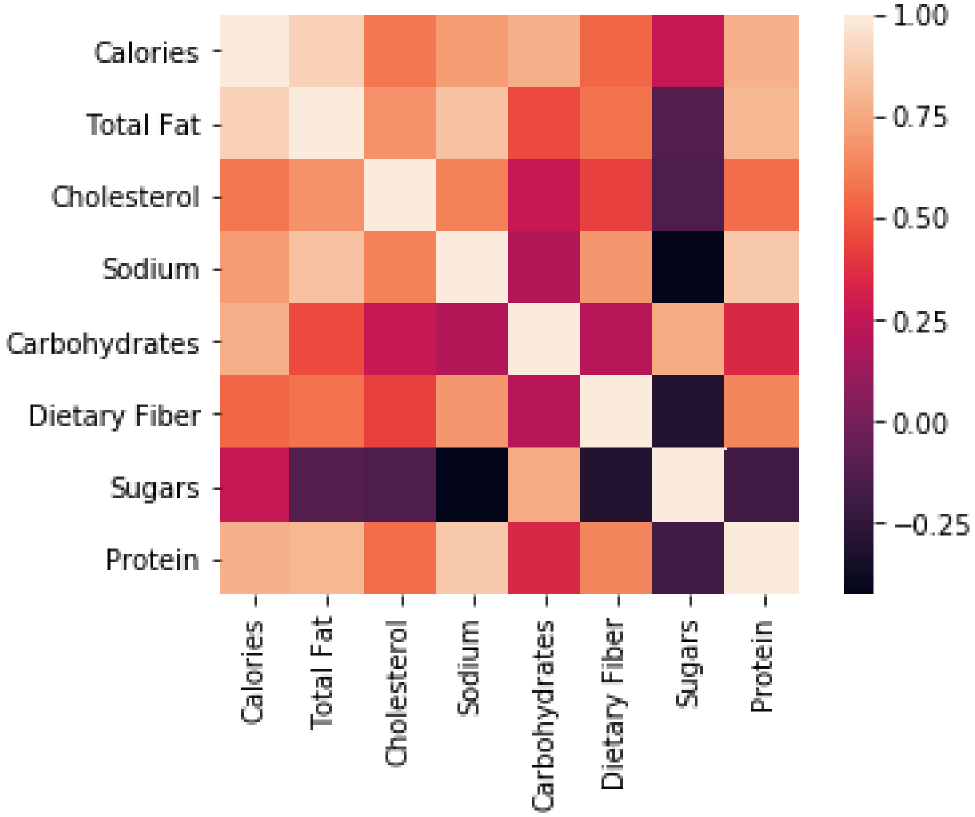
```
In [7]: cor = dataset.corr()
        print(cor)
```

```
               Calories  Total Fat  Cholesterol    Sodium  Carbohydrates  \
Calories       1.000000   0.904409     0.596399  0.712309       0.781539
Total Fat      0.904409   1.000000     0.680547  0.846158       0.461213
Cholesterol    0.596399   0.680547     1.000000  0.624362       0.270977
Sodium         0.712309   0.846158     0.624362  1.000000       0.200796
Carbohydrates  0.781539   0.461213     0.270977  0.200796       1.000000
Dietary Fiber  0.538894   0.580837     0.435575  0.694389       0.224577
Sugars         0.259598  -0.115446    -0.135518 -0.426536       0.762362
Protein        0.787847   0.807773     0.561561  0.869802       0.352122

               Dietary Fiber    Sugars   Protein
Calories            0.538894  0.259598  0.787847
Total Fat           0.580837 -0.115446  0.807773
Cholesterol         0.435575 -0.135518  0.561561
Sodium              0.694389 -0.426536  0.869802
Carbohydrates       0.224577  0.762362  0.352122
Dietary Fiber       1.000000 -0.295178  0.641345
Sugars             -0.295178  1.000000 -0.179940
Protein             0.641345 -0.179940  1.000000
```

```
In [50]: sns.heatmap(cor,square=True)
```

```
Out[50]: <matplotlib.axes._subplots.AxesSubplot at 0x4af5738ac8>
```



From corrplot we can see total fat,carbohydrates and protein has high positive correlation

4 Split data into train and test

```
In [13]: import sklearn
         from sklearn.model_selection import train_test_split
         X = dataset.loc[:,['Total Fat', 'Carbohydrates','Protein']]
         y = dataset.loc[:,'Calories']
         y.head()
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state

In [14]: X_train.shape , X_test.shape, y_train.shape, y_test.shape

Out[14]: ((182, 3), (78, 3), (182,), (78,))
```

5 Linear Regression Model

```
In [15]: ##Sklearn
         # import model
         from sklearn.linear_model import LinearRegression

         # instantiate
         linreg = LinearRegression()

         # fit the model to the training data (learn the coefficients)
         linreg.fit(X_train, y_train)
```

```
Out[15]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Intercept and Coefficients

```
In [16]: # print the intercept and coefficients
         print(linreg.intercept_)
         print(linreg.coef_)
```

```
-2.227172524597677
[9.00479505 3.99364571 4.05300657]
```

6 Predictions

```
In [17]: # make predictions on the testing set
         y_pred = linreg.predict(X_test)
```

```
In [18]: from sklearn import metrics

         print(np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
5.784963176081385
```

7 Accuracy

R^2 Value

```
In [20]: linreg.score(X,y)
```

```
Out[20]: 0.9994716750523599
```

~100% accuracy.! R^2 of 0.999(1) means that the independent variables(Total.Fat, Protein, Carbohydrates) are able to explain almost 100% of variance in Calories.