Math 408X Project

Tyler Mogensen

Introduction

- There tons of different nutrients inside of food
- I used data to see which factors are most important when looking at calories
- Calories are important for any diet

Problem Definition

- Accurate analysis of food
- Looking at what nutrients tend to lead to more calories
- We wanted to see if we could accurately predict calories for a new food to data set
- Categorizing food to make appropriate substitutes

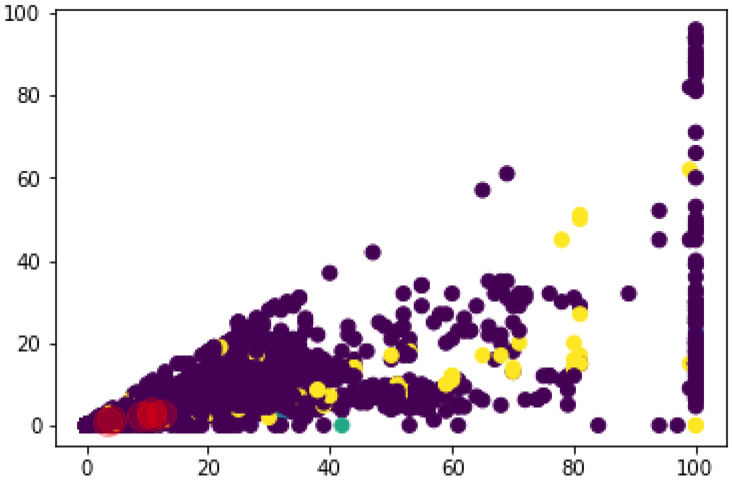
What We Did vs What Has Been Done

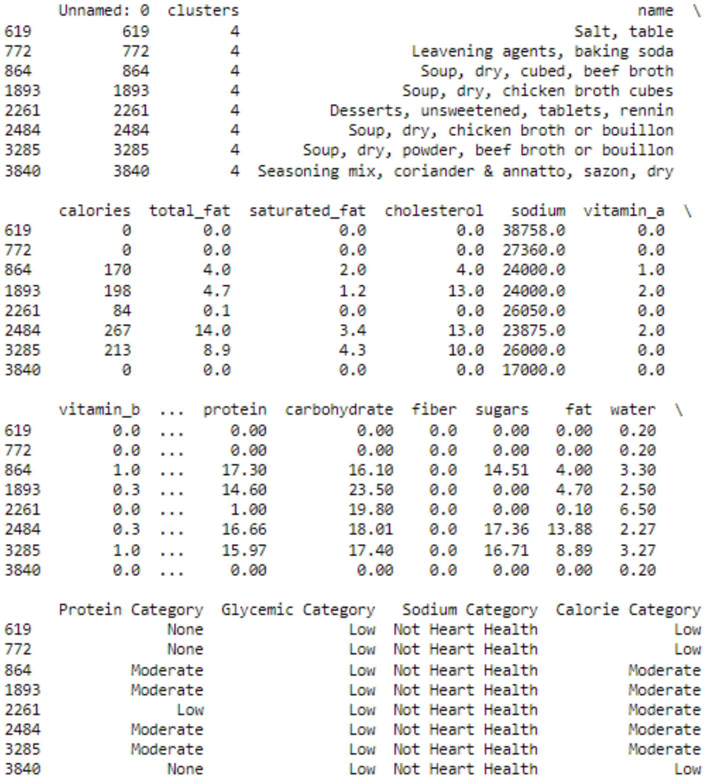
- Nutritionists and Dieticians
- InBody Scans
- Genetics

Proposed Method

- Data Collection
  - Data comes from CSV file called nutrition.csv from kaggle
  - Narrowed data down removing somewhat redundant rows
- Data Analysis
  - Multiple Linear Regression
  - Split data into test and training data
  - Clustering
  - Categorization
- Data Visualization
  - Correlation matrix heat map
  - OLS regression
  - Pairplot

Experiment 1





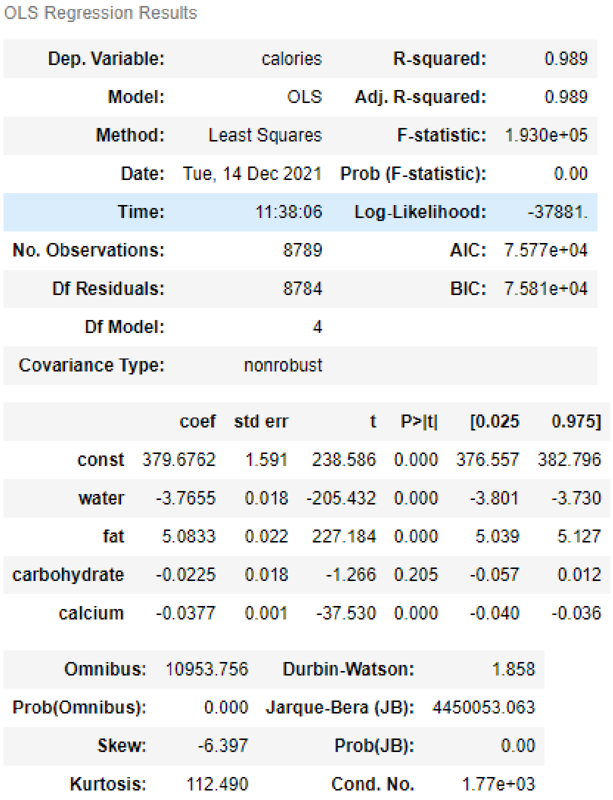
```
      Unnamed: 0  clusters                                            name  \
619          619         4                                     Salt, table
772          772         4                   Leavening agents, baking soda
864          864         4                    Soup, dry, cubed, beef broth
1893        1893         4                  Soup, dry, chicken broth cubes
2261        2261         4          Desserts, unsweetened, tablets, rennin
2484        2484         4            Soup, dry, chicken broth or bouillon
3285        3285         4       Soup, dry, powder, beef broth or bouillon
3840        3840         4  Seasoning mix, coriander & annatto, sazon, dry

      calories  total_fat  saturated_fat  cholesterol   sodium  vitamin_a  \
619          0        0.0            0.0          0.0  38758.0        0.0
772          0        0.0            0.0          0.0  27360.0        0.0
864        170        4.0            2.0          4.0  24000.0        1.0
1893       198        4.7            1.2         13.0  24000.0        2.0
2261        84        0.1            0.0          0.0  26050.0        0.0
2484       267       14.0            3.4         13.0  23875.0        2.0
3285       213        8.9            4.3         10.0  26000.0        0.0
3840         0        0.0            0.0          0.0  17000.0        0.0

      vitamin_b  ...  protein  carbohydrate  fiber  sugars    fat  water  \
619         0.0  ...     0.00          0.00    0.0    0.00   0.00   0.20
772         0.0  ...     0.00          0.00    0.0    0.00   0.00   0.20
864         1.0  ...    17.30         16.10    0.0   14.51   4.00   3.30
1893        0.3  ...    14.60         23.50    0.0    0.00   4.70   2.50
2261        0.0  ...     1.00         19.80    0.0    0.00   0.10   6.50
2484        0.3  ...    16.66         18.01    0.0   17.36  13.88   2.27
3285        1.0  ...    15.97         17.40    0.0   16.71   8.89   3.27
3840        0.0  ...     0.00          0.00    0.0    0.00   0.00   0.20

      Protein Category  Glycemic Category   Sodium Category  Calorie Category
619               None                Low  Not Heart Health               Low
772               None                Low  Not Heart Health               Low
864           Moderate                Low  Not Heart Health          Moderate
1893          Moderate                Low  Not Heart Health          Moderate
2261               Low                Low  Not Heart Health          Moderate
2484          Moderate                Low  Not Heart Health          Moderate
3285          Moderate                Low  Not Heart Health          Moderate
3840              None                Low  Not Heart Health               Low
```

Experiment 2



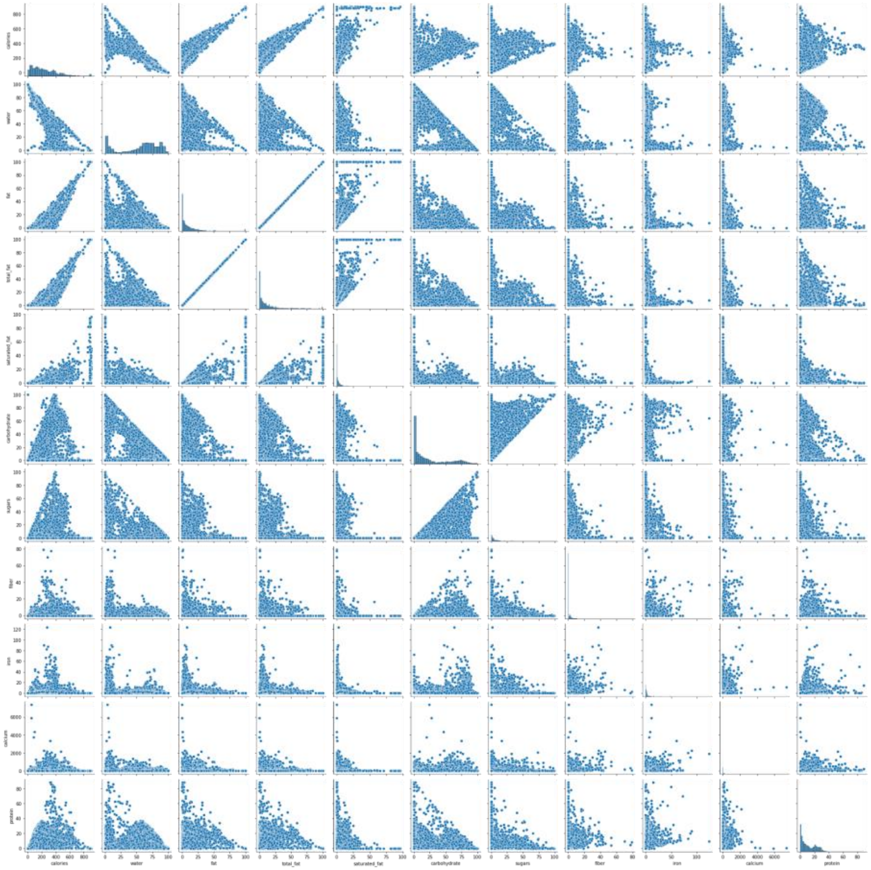
OLS Regression Results

Dep. Variable:	calories	**R-squared:**	0.989
Model:	OLS	**Adj. R-squared:**	0.989
Method:	Least Squares	**F-statistic:**	1.930e+05
Date:	Tue, 14 Dec 2021	**Prob (F-statistic):**	0.00
Time:	11:38:06	**Log-Likelihood:**	-37881.
No. Observations:	8789	**AIC:**	7.577e+04
Df Residuals:	8784	**BIC:**	7.581e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	379.6762	1.591	238.586	0.000	376.557	382.796
water	-3.7655	0.018	-205.432	0.000	-3.801	-3.730
fat	5.0833	0.022	227.184	0.000	5.039	5.127
carbohydrate	-0.0225	0.018	-1.266	0.205	-0.057	0.012
calcium	-0.0377	0.001	-37.530	0.000	-0.040	-0.036

Omnibus:	10953.756	**Durbin-Watson:**	1.858
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4450053.063
Skew:	-6.397	**Prob(JB):**	0.00
Kurtosis:	112.490	**Cond. No.**	1.77e+03

Experiment 3



Conclusion and Discussion

- I was able to group foods into clusters using the clustering algorithm
- I was able to see how the variables correlated with each other
- This analysis could go further