





# Analysis of Eating Habits Dataset through Linear Regression & Clustering

Yu Xuan Yong, PID: A16078479<sup>1</sup>, Marlon Cortez, PID: A14604525<sup>2</sup>, Cesar Grijalva, PID: A12682658<sup>3</sup>, and Servin Wayne Vartanian, PID: A14802532<sup>4</sup>

<sup>1</sup>University of California, San Diego, COGS 109

**Abstract** - This report presents an analysis on the Eating Habits dataset[1] from the UCI Machine Learning Repository. We utilized methods taught in the lecture such as Linear Regression, Principal Component Analysis (PCA) and Clustering. We decided on analyzing the attributes within the dataset to determine which were the optimal indicators that would actually predict the class label of the dataset: weight.

## 1. Introduction

Eating habits are a very essential part of our lives, especially seeing how those aspects could potentially lead to health issues and other complications in the future. Therefore, we saw fit to analyze the Eating Habits dataset [1] which contained attributes on factors that could influence weight. Our research was centered on discovering the optimal variables that influenced weight, which could help us understand what should be controlled within those eating habits to affect an individual's weight. Using the data analysis methods **Linear Regression** and **Clustering**, we were able to conclude that certain variables do have a more significant impact on weight than others.

## 2. Breakdown of Project

### 2.1 Research Question

Using exploratory linear regression and clustering, we aim to examine several attributes from the dataset to find which are the optimal indicators to predict the weight of an individual.

### 2.2 Description of Dataset

The dataset consists of data collected from individuals from Mexico, Peru, and Colombia. This data is useful for the estimation of the obesity levels based on eating habits and physical conditions. There are 2111 instances and 17 different attributes. Additionally, the data is classified using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

### 2.3 Attributes of Dataset

16 attributes were collected through a survey and one output attribute for obesity levels was determined for the individual based on their height and weight.

Eating habit attributes:

- **FAVC**: Frequent consumption of high caloric foods.
- **FCVC**: Frequency of consumption of vegetables.
- **NCP**: Number of main meals.
- **CAEC**: Consumption of food between meals.
- **CH2O**: Consumption of water daily.
- **CALC**: Consumption of alcohol.

Physical condition attributes:

- **SCC**: Calories consumption monitoring.
- **FAF**: Physical activity frequency.
- **TUE**: Time using technology devices (hours).
- **MTRANS**: Transportation method used.

Additional variables:

- **Gender**: The person's gender.
- **Age**: The person's age.
- **Height**: The person's height.
- **Weight**: The person's weight.
- **Family History With Overweight**: yes/no if their family has a history with obesity.
- **SMOKE**: Whether they smoke or not.

## 3. Methodology

### 3.1 Correlation Matrix

First, we created a Correlation Matrix to examine the correlation between all the attributes and the output variable: weight. This allowed for a more layered and exploratory route as we were able to chose an array of variables with positive, negative, and neutral correlations with weight. By doing this we aim to diversify our model creation process and allow us to create some truly unique models that can take in a variety of inputs and give out one continuous output, weight.

## 3.2 Linear Regression

We applied multivariate linear regression with cross validation using two different model types: Model 1 (M1) and Model 2 (M2).

For Model 1, we analyzed the relation each variable had with weight to see if any relevant results would emerge.

Model 2 utilized two or more variables to see if their joint presence had any different impact on weight. We then drew conclusions per model type about which combination of variables had a more drastic impact on weight.

This allowed us to make 15 different models, 5 univariate and 10 multivariate. Additionally, we split our data into training and testing partitions. We then solved our weights and got our prediction models which we used to plot a linear regression scatterplot to see how profound of an impact they have on the weight. Then we calculated their SSE and RMSE to see which 1st degree model best fit our data.

## 3.3 Clustering

Our first step for clustering was performing a PCA to reduce dimensionality of our data. After this, we took a portion of variables that were effective for our analysis and performed kMeans on them. This allowed us to determine the optimal number of clusters using the elbow method. Finally, we performed Parallel Coordinates to see what is the data distribution per cluster and see how weight plays into that.

# 4. Results

## 4.1 Correlation Matrix

**Fig 1.** shows the correlation matrix that we derived.

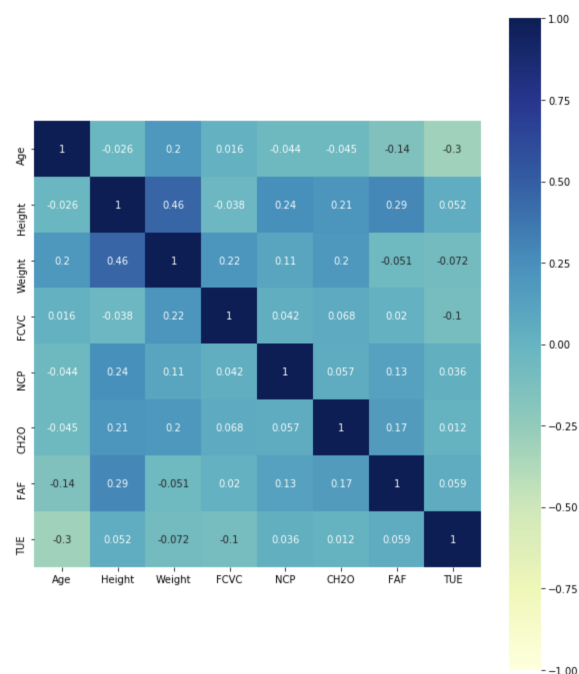
From the correlation matrix, we chose Physical activity frequency (FAF) and Time using technology (TUE) as our least correlated variables, Height and Age as the most positively correlated, and Gender as a non-listed categorical variable.

## 4.2 Linear Regression

**Fig. 2** shows the different linear regression graphs that were derived by combining different variables to form a regression line. We referenced the correlation matrix that we outputted to test those different attributes.

Using the attributes selected from the correlation matrix, 15 models were run, 5 univariate and 10 multivariate (**Fig. 2**). The univariate models are as follows:

- The weight vector for M1 (gender) is: [82.12869425 8.40258935].
- Model 1 (gender) is:  $\text{weight} = 82.1286942522292 + 8.402589350527581 * (\text{gender})$



**Fig. 1.** Correlation matrix showing relationship of different attributes in dataset.

- The weight vector for M1 (age) is: [64.33057853 0.89678302]
- Model 1 (age) is:  $\text{weight} = 64.33057852939457 + 0.8967830209346258 * (\text{age})$
- The weight vector for M1 (height) is: [-135.11204552 130.1521572 ]
- Model 1 (height) is:  $\text{weight} = -135.11204551897455 + 130.15215720030213 * (\text{height})$
- The weight vector for M1 (techtime) is: [88.14979663 -2.80772621]
- Model 1 (techtime) is:  $\text{weight} = 88.14979663382302 + -2.807726212623631 * (\text{techtime})$
- The weight vector for M1 (activetime) is: [87.79906189 -1.36027073]
- Model 1 (activetime) is:  $\text{weight} = 87.79906188914119 + -1.3602707335931097 * (\text{activetime})$

After this, we ran 10 multivariate models which are as follows:

- The Multivariate Model for Age and Gender is:  $\text{Weight} = [61.79726184] + [0.83814657] * (\text{age}) + [8.82741779] * (\text{gender})$
- The Multivariate Model for Age and Height is:  $\text{Weight} = [-155.96080573] + [0.88735306] * (\text{age}) + [130.00501184] * (\text{height})$
- The Multivariate Model for Age and Techtime is:  $\text{Weight} = [67.4664496] + [0.79617629] * (\text{age}) + [-0.62315962] * (\text{techtime})$

- The Multivariate Model for Age and Activetime is:  
Weight = [66.63620088] + [0.84058978] \*(age) + [-0.40168437] \*(activetime)
- The Multivariate Model for Gender and Height is:  
Weight = [-183.22404155] + [-10.11228695] \*(gender) + [161.64382122] \*(height)
- The Multivariate Model for Gender and Techtime is:  
Weight = [84.09573982] + [9.56581409] \*(gender) + [-3.3366158] \*(techtime)
- The Multivariate Model for Gender and Activetime is:  
Weight = [83.81280412] + [10.29049239] \*(gender) + [-2.80101073] \*(activetime)
- The Multivariate Model for Height and Techtime is:  
Weight = [-135.63267999] + [131.80294641] \*(height) + [-3.40192632] \*(techtime)
- The Multivariate Model for Height and Activetime is:  
Weight = [-157.70262952] + [147.22259209] \*(height) + [-6.00678105] \*(activetime)
- The Multivariate Model for Techtime and Activetime is:  
Weight = [89.66356657] + [-2.36668828] \*(techtime) + [-1.21972413] \*(activetime)

From these models, we created regression graphs for all univariate models and found that Height and Age were the most related while TUE, Gender, and FAF were the least correlated.

When looking at our regression graph for gender, we saw very little discrepancy in weight difference between being a male or female. Males do seem to weigh slightly heavier than females, but overall there's nothing noteworthy here to report.

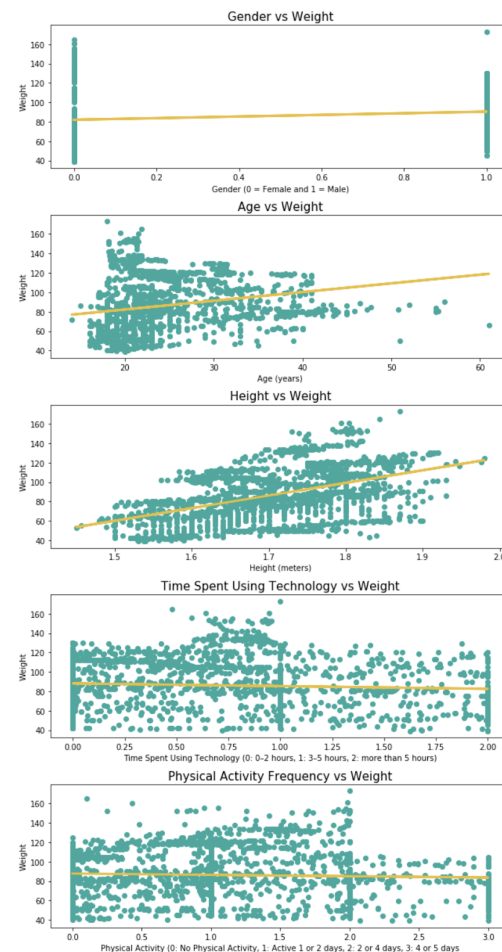
As for age, we found a strong correlation between age and weight with weight increasing as age does for both genders.

Another positive correlation was found between height and weight. While we can not determine whether that weight is muscle mass or fat, as the dataset does not provide a BMI or muscle mass/fat percentage data, we can reliably say taller people weigh more.

The graph for Time Using Technology showed a slight decrease in weight the more individuals interacted with technology.

When looking at the regression result for Physical Activity Frequency, a slight decrease in weight was shown the more individuals exercised.

It is important to note that just because this univariate model found no correlation between its variable and weight, it does not mean one does not exist once we moved into the multivariate realm. This is where our multivariate, M2,



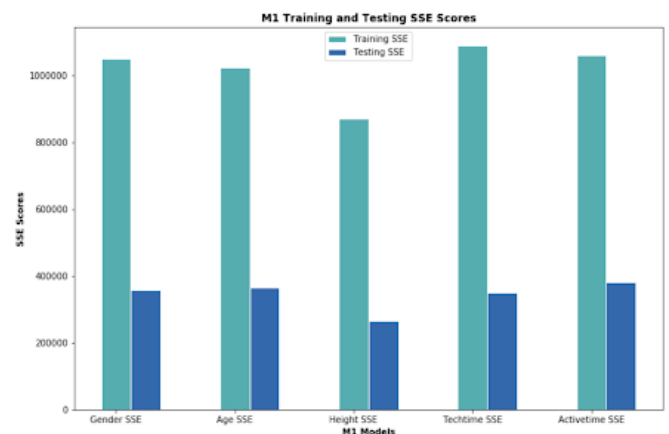
**Fig. 2.** Linear Regression graphs describing the relationship between weight and different variables.

models came into play.

### 4.3 RMSE and SSE

To examine even further, we computed SSE and RMSE scores for all models to determine which best fit our data.

The training and testing SSE for model 1 is shown in **Fig. 3** below.



**Fig. 3.** M1 Training and Testing Set Sum Squared Error.

The training and testing SSE for model 1 are as follows:

#### 1. Gender

- Training SSE for Model 1 (Gender) is: 1051795.1169674527
- Test SSE for Model 1 (Gender) is: 357879.2549451049

#### 2. Age

- Training SSE for Model 1 (Age) is: 1023198.6953659726
- Test SSE for Model 1 (Age) is: 365571.3896900322

#### 3. Height

- Training SSE for Model 1 (Height) is: 871340.1697636023
- Test SSE for Model 1 (Height) is: 265713.09665808873

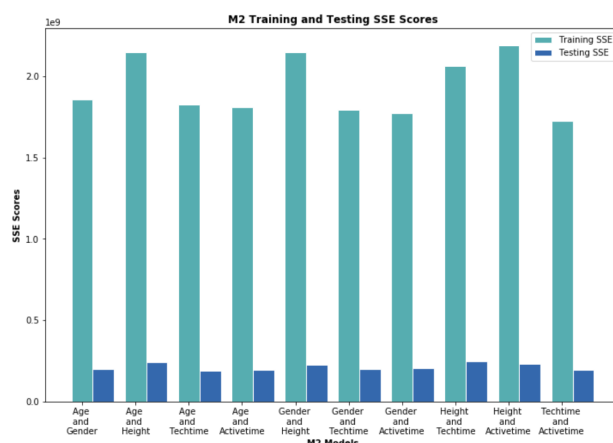
#### 4. Techtime

- Training SSE for Model 1 (Techtime) is: 1091033.5090157776
- Test SSE for Model 1 (Techtime) is: 349193.38630899345

#### 5. Activetime

- Training SSE for Model 1 (Activetime) is: 1061808.7560180882
- Test SSE for Model 1 (Activetime) is: 381905.50631207466

The training and testing SSE for model 2 is shown in **Fig. 4** below.



**Fig. 4.** M2 Training and Testing Set Sum Squared Error.

The training and testing models for Model 2 are as follows:

#### 1. Age and Gender

- Training SSE for Model 2 (Age and Gender) is: 1857642599.755895
- Test SSE for Model 2 (Age and Gender) is: 200356821.31118292

#### 2. Age and Height

- Training SSE for Model 2 (Age and Height) is: 2149608707.8878417
- Test SSE for Model 2 (Age and Height) is: 242565211.0513554

#### 3. Age and Techtime

- Training SSE for Model 2 (Age and Techtime) is: 1823810575.7459917
- Test SSE for Model 2 (Age and Techtime) is: 185750454.34667024

#### 4. Age and Activetime

- Training SSE for Model 2 (Age and Activetime) is: 1811704532.6721375
- Test SSE for Model 2 (Age and Activetime) is: 192389339.0987105

#### 5. Gender and Height

- Training SSE for Model 2 (Gender and Height) is: 2146467754.7578928
- Test SSE for Model 2 (Gender and Height) is: 224109319.62875098

#### 6. Gender and Techtime

- Training SSE for Model 2 (Gender and Techtime) is: 1791784032.1723177
- Test SSE for Model 2 (Gender and Techtime) is: 196365563.73336336

#### 7. Gender and Activetime

- Training SSE for Model 2 (Gender and Activetime) is: 1771367623.0359082
- Test SSE for Model 2 (Gender and Activetime) is: 204181524.42061839

#### 8. Height and Techtime

- Training SSE for Model 2 (Height and Techtime) is: 2063250345.961672
- Test SSE for Model 2 (Height and Techtime) is: 246454478.92518175

#### 9. Height and Activetime

- Training SSE for Model 2 (Height and Activetime) is: 2187229615.933541
- Test SSE for Model 2 (Height and Activetime) is: 227957664.44260058

## 10. Techtime and Activetime

- Training SSE for Model 2 (Techtime and Activetime) is: 1725251733.353066
- Test SSE for Model 2 (Techtime and Activetime) is: 192576685.2823746

The training and testing RMSE for model 1 is shown in **Fig. 5** below.



**Fig. 5.** M1 Training and Testing Set Root Mean Squared Error.

Training and testing for RMSE for Model 1 are as follows:

### 1. Gender

- Training RMSE for Model 1 (Gender) is: 22.321401457687838
- Test RMSE for Model 1 (Gender) is: 13.020394579640199

### 2. Age

- Training RMSE for Model 1 (Age) is: 22.015871017913216
- Test RMSE for Model 1 (Age) is: 13.159578624112662

### 3. Height

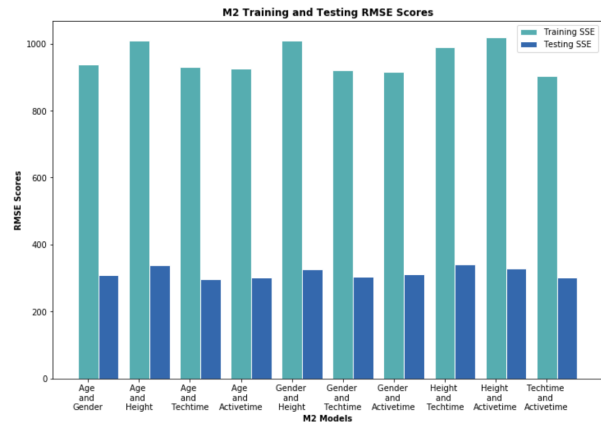
- Training RMSE for Model 1 (Height) is: 20.316540175921922
- Test RMSE for Model 1 (Height) is: 11.219212235798654

### 4. Techtime

- Training RMSE for Model 1 (Techtime) is: 22.733951440278744
- Test RMSE for Model 1 (Techtime) is: 12.861419036221973

### 5. Activetime

- Training RMSE for Model 1 (Activetime) is: 22.427405454133915
- Test RMSE for Model 1 (Activetime) is: 13.45035798572431



**Fig. 6.** M2 Training and Testing Set Root Mean Squared Error.

The training and testing RMSE for model 2 is shown in **Fig. 6.**

Training and testing for RMSE for Model 2 are as follows:

### 1. Age and Gender

- Training RMSE for Model 2 (Age and Gender) is: 938.0737088110044
- Test RMSE for Model 2 (Age and Gender) is: 308.07605884694675

### 2. Age and Height

- Training RMSE for Model 2 (Age and Height) is: 1009.1032146960586
- Test RMSE for Model 2 (Age and Height) is: 338.97692863739167

### 3. Age and Techtime

- Training RMSE for Model 2 (Age and Techtime) is: 929.4921968885682
- Test RMSE for Model 2 (Age and Techtime) is: 296.6339300612879

### 4. Age and Activetime

- Training RMSE for Model 2 (Age and Activetime) is: 926.4021809085441
- Test RMSE for Model 2 (Age and Activetime) is: 301.88837153362556

### 5. Gender and Height

- Training RMSE for Model 2 (Gender and Height) is: 1008.3657073184453
- Test RMSE for Model 2 (Gender and Height) is: 325.8260788318958

### 6. Gender and Techtime

- Training RMSE for Model 2 (Gender and Techtime) is: 921.2950003014479



- Test RMSE for Model 2 (Gender and Techtime) is: 304.99207010943223

#### 7. Gender and Activetime

- Training RMSE for Model 2 (Gender and Activetime) is: 916.0311337032932
- Test RMSE for Model 2 (Gender and Activetime) is: 311.0026605437416

#### 8. Height and Techtime

- Training RMSE for Model 2 (Height and Techtime) is: 988.6255875338773
- Test RMSE for Model 2 (Height and Techtime) is: 341.68368398013774

#### 9. Height and Activetime

- Training RMSE for Model 2 (Height and Activetime) is: 1017.8952136657148
- Test RMSE for Model 2 (Height and Activetime) is: 328.6116701657359

#### 10. Techtime and Activetime

- Training RMSE for Model 2 (Techtime and Activetime) is: 904.0284947679864
- Test RMSE for Model 2 (Techtime and Activetime) is: 302.0353232111348

Originally, Time Spent With Technology (TUE) and Physical Activity Frequency (FAF) showed little to no effect on increasing or decreasing weight. However, after having run our data through secondary multivariate (M2) models, we saw that models with Techtime and Activetime were the best fitting models for our dataset because they consistently scored lower than other models that did not include them.

Thus we can state that the top five models that fit our data the best are:

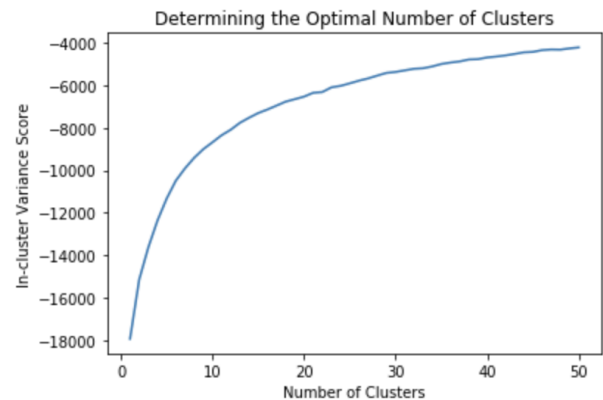
- **Techtime and Activetime:** With a test RMSE score of 305.27
- **Age and Activetime:** With a test RMSE score of 307.96
- **Age and Techtime:** With a test RMSE score of 309.96
- **Age and Gender:** With a test RMSE score of 310.81
- **Gender and Activetime:** With a test RMSE score of 314.92

### 4.4 Clustering

Before performing our clustering analysis, we refined the original dataset by dropping anything that might not be relevant to the clustering (**Fig 7**).

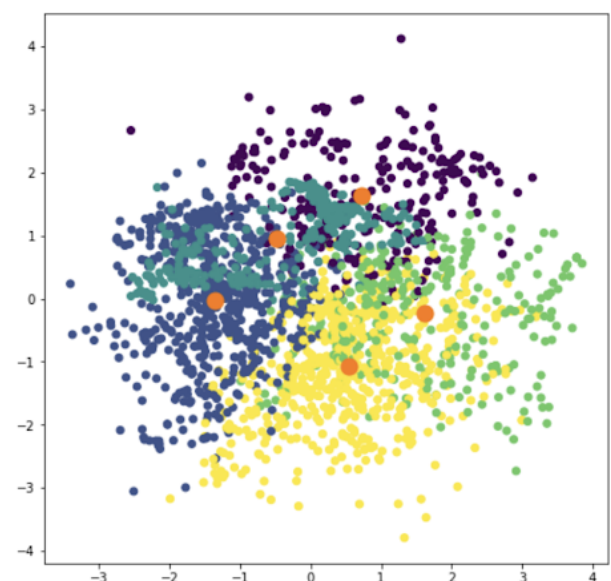
	Gender	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
0	Female	21.0	1.62	64.0	2.0	3.0	2.0	0.0	1.0
1	Female	21.0	1.52	56.0	3.0	3.0	3.0	3.0	0.0
2	Male	23.0	1.80	77.0	2.0	3.0	2.0	2.0	1.0
3	Male	27.0	1.80	87.0	3.0	3.0	2.0	2.0	0.0
4	Male	22.0	1.78	89.8	2.0	1.0	2.0	0.0	0.0

**Fig. 7.** Dataset relevant for Clustering.



**Fig. 8.** Graph of Number of Clusters vs Error.

For Clustering, we set our optimal number of clusters to  $k=5$  using the elbow method (**Fig. 8**) as the variance started flattening out as the number of clusters went past 10. Since we had 10 features in our dataset, we needed to conduct PCA on the data and project it onto two components so that we could plot our clusters (**Fig. 9**).



**Fig. 9.** Eating Habits dataset clustered with labels.

We then partitioned our data of eating habits into 5 clusters. To determine how these clusters were formed, we analyzed the parallel coordinate plot from the pandas library (**Fig. 10**).

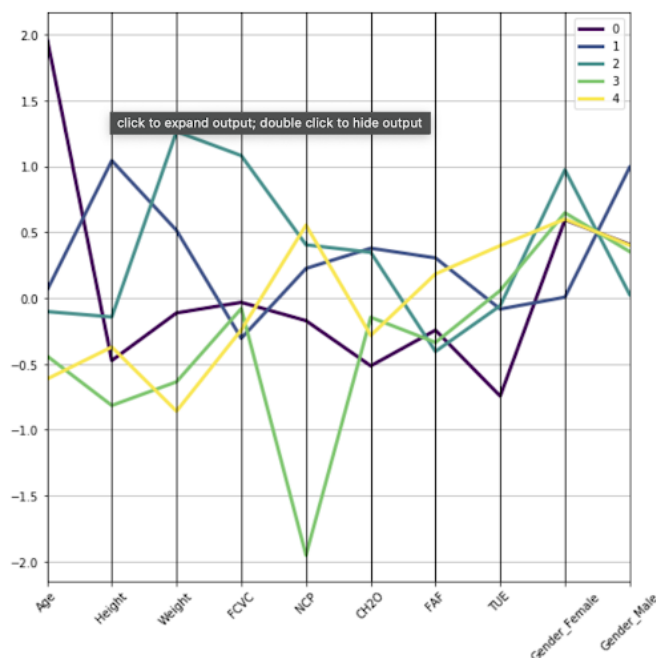


Fig. 10. Parallel coordinate plot to determine meaning of clusters.

#### • Cluster 0

- Centroid 1: [ 1.97054176 -0.47426073 -0.11276945 -0.032497 -0.17172933 -0.51486148 -0.24409226 -0.74389327 0.59375 0.40625 ]
- Number of observations: 256
- Maximum distance from centroid: 5.168246108982701
- Average distance from centroid: 2.394331622570507

#### • Cluster 1

- Centroid 2: [ 0.05912637 1.04344179 0.51361939 -0.3054928 0.2247369 0.37825849 0.30487472 -0.08341364 0.00655738 0.99344262 ]
- Number of observations: 610
- Maximum distance from centroid: 4.160046797828287
- Average distance from centroid: 2.2245743833865776

#### • Cluster 2

- Centroid 3: [-0.10249783 -0.14342319 1.26400701 1.08144661 0.40415272 0.34732841 -0.40541371 -0.05836476 0.97345133 0.02654867 ]
- Number of observations: 339
- Maximum distance from centroid: 3.6150385458284453

- Average distance from centroid: 1.7110953887568607

#### • Cluster 3

- Centroid 4: [-0.4398354 -0.8154928 -0.63580562 -0.07993277 -1.95165967 -0.1461136 -0.3368373 0.05729826 0.64726027 0.35273973 ]
- Number of observations: 292
- Maximum distance from centroid: 4.284178385015859
- Average distance from centroid: 2.4193983265509407

#### • Cluster 4

- Centroid 5: [-0.61457178 -0.3718964 -0.85876544 -0.24201985 0.55333885 -0.28340766 0.18290843 0.39800254 0.59934853 0.40065147 ]
- Number of observations: 614
- Maximum distance from centroid: 4.30122984505565
- Average distance from centroid: 2.3937750250514784

After formulating 5 clusters and using the parallel coordinates graph from above, we ranked the aforementioned clusters by the highest number of weight attributes in them, giving a better indicator of what affected weight.

Cluster 2, the teal cluster, came in first with being the cluster with most weight grouped into it. It was formed from a high proportion of females and weight.

Cluster 1, the blue cluster, came in second place for most weight data points in its cluster, and it was formed from a high proportion of males and height.

Cluster 0, the purple cluster, was ranked in the middle. It's data was mainly formed from a high proportion of age and females.

Cluster 3, the green cluster, was second to last. It's data clusters were slightly composed of weight inputs, but overall it was formed from a high proportion of FCVC and females.

Cluster 4, the yellow cluster, ranked at the bottom of our list with little to none of its data being clustered with weight. It was primarily formed from high proportions of females and NCP.

## 5. Discussion

### 5.1 Main Points

To sum up our results, through linear regression analysis we were able to determine which models gave us the best fit for our data. The attributes which helped predict weight the most were age, activetime, techtime, and gender according to our linear regression analysis. However, our clustering analysis showed that gender, age, and height were some of the data points most commonly grouped with weight.

Thus we determined through our analysis that Gender, and Age were the two leading predictors in weight, with height, techtime and activetime being secondary best-fits to our models.

### 5.2 Conclusion

*Was the research question answered? If so, what is the answer and how do you know? If not, why not? Was the analysis technique appropriate for the dataset?*

According to our findings in our univariate linear regression implementation, height is the most optimal indicator of the perceived weight of a person. However, according to our multivariate linear regression implementation, Techtime and Activetime are our most optimal variables when predicting the weight of an individual. Aside from these two variables there are several other linear regression models in our results that would also serve well in predicting the weight of a person.

In our clustering analysis, by running the k-means clustering algorithm using 5 centroids and then reducing the dimensions of our data from 17 variables to 10 (9 plus 1 for one-hot encoding gender). We concluded in our clustering analysis that cluster gender, age and height played a significant role in determining the weight of a person, indicating that they may be the most optimal indicators for predicting weight based on the variables of this dataset.

*This section may include suggestions for future work or a description of data that could be analyzed to answer a particular question.*

Data that could have furthered our analysis would have been the inclusion of the person's body mass index (BMI), as this would allow us to determine whether someone's weight is attributed to muscle mass or body fat. Instead of having data that tell us the number of main meals (NCP), a variable that tells us the caloric intake would be the most ideal. In the case of gender, while all genders are vastly different, grouping them together is not ideal, as what affects the weight of one gender may not necessarily be the case for the others. For example, in our clustering analysis we learned that some of our clusters had a higher proportion of females than males that represented that specific cluster.

While we can accurately deduce the variable from the set of given variables which may be the best predictor, we could always benefit from a larger dataset with more samples, and from a more diverse pool of individuals, as this dataset is entirely a sample representation of Mexico, Peru, and , Colombia, but if attempting to compare weight level on a global scale then that is where this data set would fall short. Even if a question regarding comparing weight levels across the globe might be an particular question that could potential be analyzed with a similar dataset the consists of more countries from around the world.

## 6. Links for Final Project

Code on Github: <https://github.com/yongyx/COGS109-Modeling-and-Data-Analysis>

Youtube link for Final Project: [https://www.youtube.com/watch?v=FKl\\_86BmQNE](https://www.youtube.com/watch?v=FKl_86BmQNE)

## References

- [1] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: <http://archive.ics.uci.edu/ml>.