Jimmy Chen
Data Analytics Assignment 6
CSCI 4600 (4000 Level)
Dr. Ahmed Eleish
chenj62@rpi.edu

Analyzing Obesity Levels Based on Habits and Physical Condition

https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Exploratory Data Analysis

For this assignment, I chose the Estimation of Obseity Levels based on eating habits and physical condition dataset. After running a skimr, I was able to see that there was no data quality issues. All the data we needed was present, and there was no NA values so it made things much easier.

```
> skim_without_charts(data)
── Data Summary ────────────────────────
                           Values
Name                       data
Number of rows             2111
Number of columns          18
_____________________
Column type frequency:
  factor                   9
  numeric                  9
_____________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable                  n_missing complete_rate ordered n_unique top_counts
1 gender                                 0             1 FALSE          2 Mal: 1068, Fem: 1043
2 family_history_with_overweight         0             1 FALSE          2 yes: 1726, no: 385
3 favc                                   0             1 FALSE          2 yes: 1866, no: 245
4 caec                                   0             1 FALSE          4 Som: 1765, Fre: 242, Alw: 53, no: 51
5 smoke                                  0             1 FALSE          2 no: 2067, yes: 44
6 scc                                    0             1 FALSE          2 no: 2015, yes: 96
7 calc                                   0             1 FALSE          4 Som: 1401, no: 639, Fre: 70, Alw: 1
8 mtrans                                 0             1 FALSE          5 Pub: 1580, Aut: 457, Wal: 56, Mot: 11
9 n_obeyesdad                            0             1 FALSE          7 Obe: 351, Obe: 324, Obe: 297, Ove: 290

── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd    p0    p25    p50    p75   p100
1 age                   0             1 24.3    6.35   14    19.9   22.8    26     61
2 height                0             1  1.70   0.0933  1.45  1.63   1.70    1.77   1.98
3 weight                0             1 86.6   26.2    39    65.5   83     107.   173
4 fcvc                  0             1  2.42   0.534   1     2      2.39    3      3
5 ncp                   0             1  2.69   0.778   1     2.66   3       3      4
6 ch2o                  0             1  2.01   0.613   1     1.58   2       2.48   3
7 faf                   0             1  1.01   0.851   0     0.125  1       1.67   3
8 tue                   0             1  0.658  0.609   0     0      0.625   1      2
9 bmi                   0             1 29.7    8.01   13.0  24.3   28.7    36.0   50.8
```

Figure 1: Skim results, showcasing each variable's statistics

Upon closer analysis by graphing some numerical distributions, we can see that the data is based on a young, university-aged crowd, which could be a potential cause of bias in our model results, as we don't account much more older individuals, especially over 40, where the count drops significantly. I calculated the BMI myself, which is weight / height^2, and used this as what I wanted to predict. The most frequent BMi was around 26, and the most frequent height was around 1.77, and weight around 80 kgs.

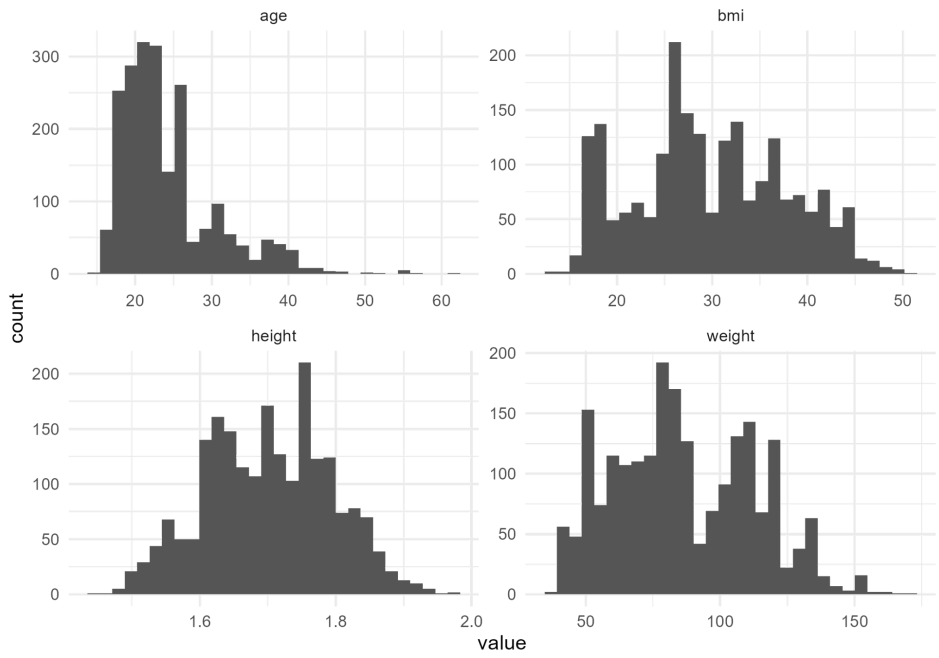

Figure 2: Numeric Distributions Graphs

The categorial balance bar chart tell us that there's pretty even distribution on each weight type, but more people fall into the type 1 obesity factor. Overall, the class distribution is reasonably broad across all levels, which is helpful for modeling each category.

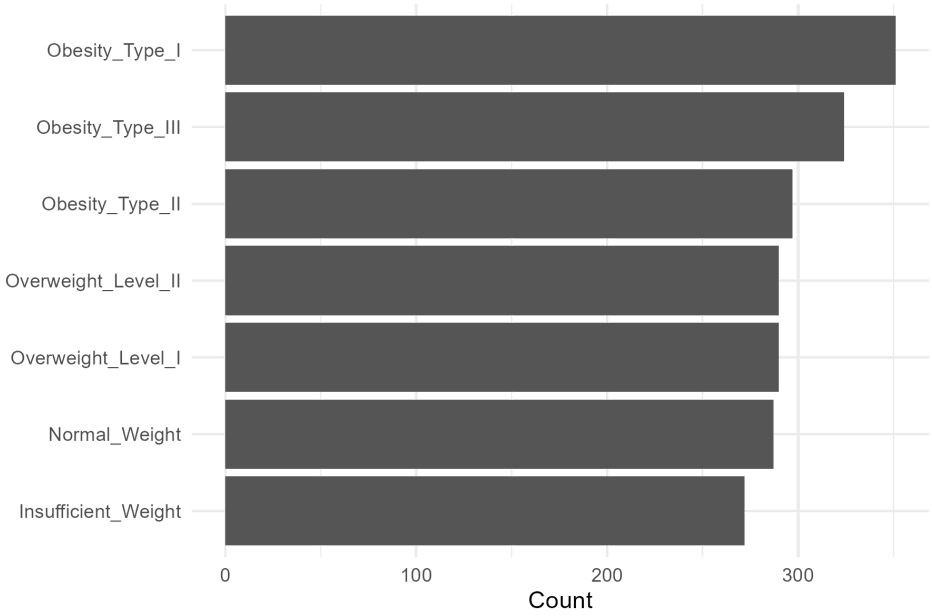

Figure 3: Categorical Balance bar chart

I also generated a correlation heatmap for the values that were nuermical. As expected, weight and BMI are strongly correlated, as BMI is based on weight. Heigh has a moderate positive correlation with weight at r=0.46, showing that taller individuals tend to weigh more. Age shows only a mild correlation with weight and BMI, which shows that older individuals in this dataset tend to have slightly higher BMI on average. Based on everything we analyzed, we can potentially have a model to predict the BMI, and also a model to find out how likely someone is to fall into each one of the weight categories.

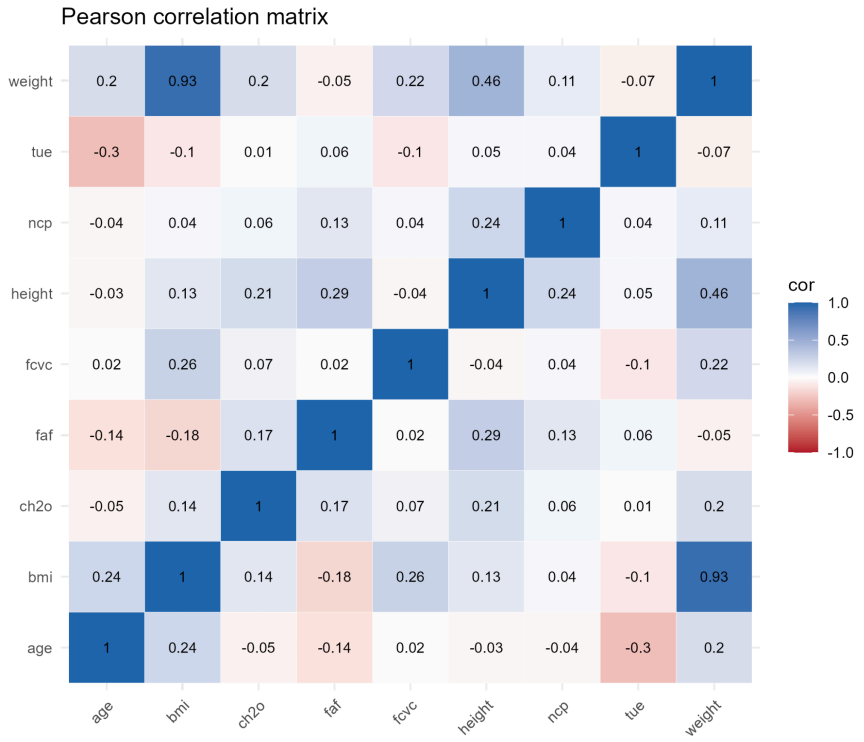

Figure 4: Correlation heatmap for the numeric features

Model Development, Validation, and Optimization

I developed two models using the dataset. I used a linear regression model to predict BMI, and a Multinomial Logistic Regression to predict the categorical obesity level.

To predit BMI, I used a linear regression model and fitted all the relevant predictors like age, weight, height, gender, and various dietary and exercise habits. I expect weight and heigh to be the primary determinants of BMI, but the addition of lifestyle factors might provide more insights about the BMI beyond weight and height. I chose linear regression because it's simple, and interpretable, and is easy to use to predict numerical values. Before I ran the training, I first normalized and encoded the data in the recipe, and then I trained the model on it's training set.

The model ended up actually achieving a really good, fit with R^2 being ~.98, but I'm a little bit sketpical as the data we had only included young individuals.

Taking a closer look at the residuals vs fitted for the BMI model, the residuals seemed randomly scattered around zero, across the range of fitted BMI values, with no obvious curvature. This suggests that the relationship between predictors and BMI is well captured. The residual spread is roughly constant for low, medium, and high predicted BMI. Only at the very extreme high end of predicted BMI (around 50) do we see slightly larger residual deviations, but there are very few.

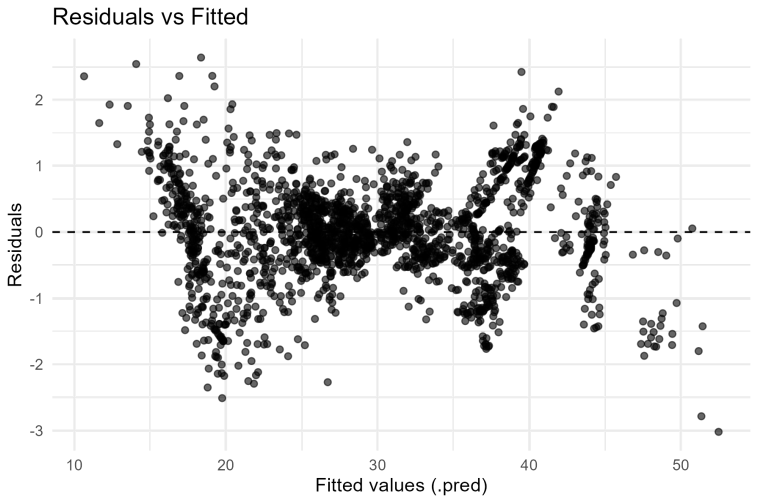

Figure 5: Residuals vs. Fitted Linear Regression BMI

In the Q-Q plot of residuals we can see that the residuals follow the line pretty closely, but more so in the middle range. There are some deviations at the tails, with the most extreme residuals straying away only at the very end of the tails, which means that the prediction for very high and very low BMIs were harder. Overall there were good results, with some hints at bias.

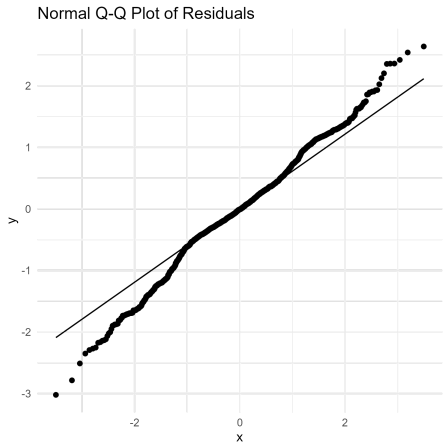

Figure 6: Normal Q-Q Plot of Residuals for Linear Regression BMI

In terms of the multinomial logistic regression, I wanted to predict the obesity classification for each invidual, which meant that there could be seven outcomes. I applied a pre-processing recipe again, where all categorial predictor variables like gender, family history obesity, and habits, were scaled to their normalized ranges. The logistic model was trained on the training set and validated with 5-fold cross-validation during tuning, then evaluated on the test set. It achieved high accuracy with about 95% of individuals' obesity levels were correctly classified on the test data. This strong performance indicates the model captures the distinctions between classes very well. Given the ordinal nature of obesity levels, most of the errors that did occur were understandable misclassifications between adjacent categories, like confusing Overweight with the nearest Obesity level in some borderline cases.

The model's predictions were highly accurate across all classes. Each obesity category has the majority of its instances on the diagonal of the matrix. For example, of those actually in the Normal_Weight category, almost all were correctly predicted as Normal (52 out of 56), with only a couple mislabeled as the next level up (Overweight Level I). Similarly, Obesity_Type_I and Obesity_Type_II individuals were overwhelmingly classified into their correct groups. The overall test set accuracy was ~95%, and metrics like weighted F1-score were similarly high, reflecting the excellent performance. The high accuracy implies that the logistic model could be a useful automated classifier of obesity risk levels in practice.

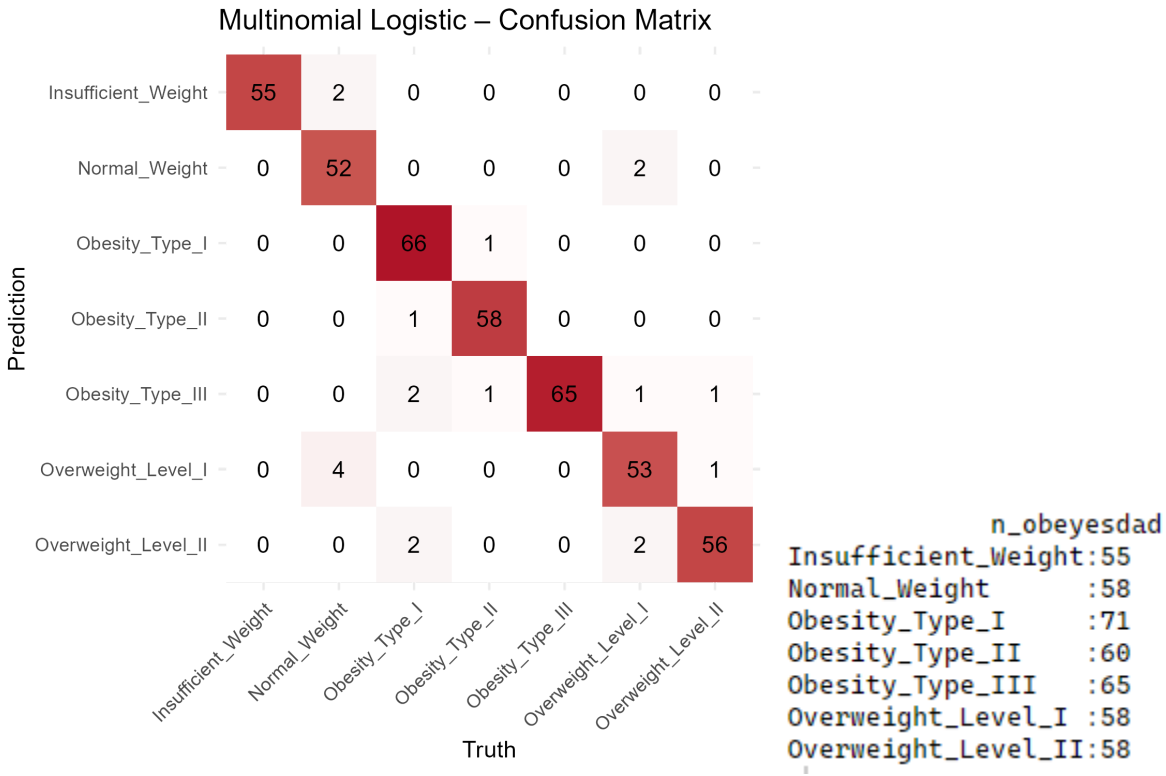

Figure 7: Confusion Matrix of the Multinomial Logstic Model

Looking deeper, I noticed that BMI, weight and height were the top predictors, which shows that body measurements are the most dominant factors to classifying obesity. The dietary and activity behaviors provide addition help, but these factors are correspondent with an individual's BMI.

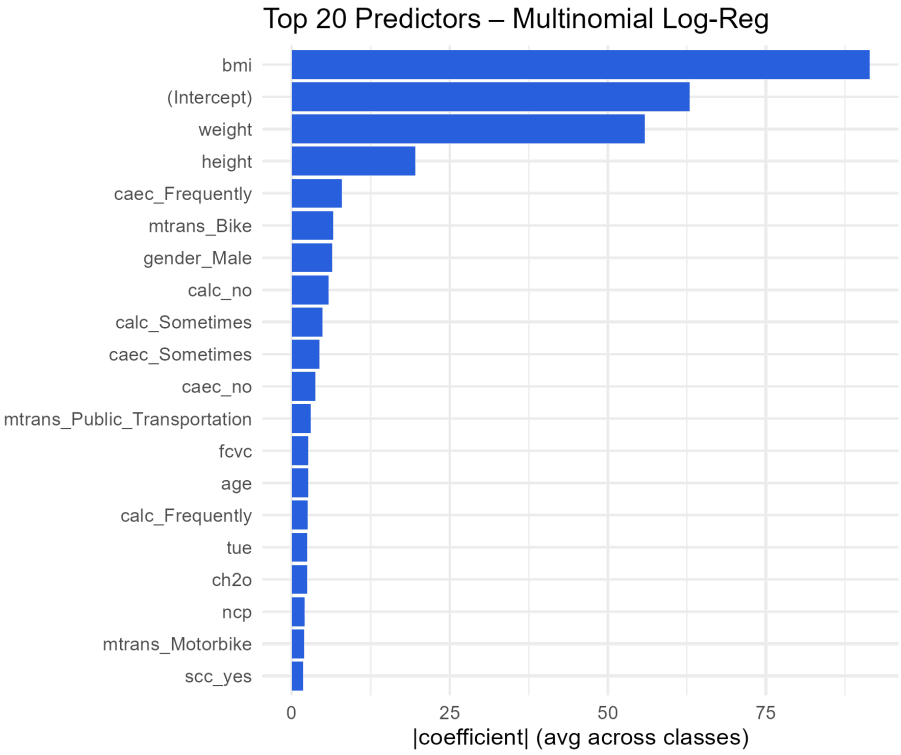

Figure 8: Top 20 Predictors

Decisions

The linear regression accurately predicted BMI, closely replicating the actual BMI formula using height and weight. Diagnostics confirmed stable residuals and normal distribution. Practically, while BMI can be directly calculated, this model confirms that height and weight alone greatly influences BMI determination, indicating that weight management is essential for BMI control.

The multinomial logistic regression demonstrated great accuracy of ~95%, reliably categorizing obesity risks. The confusion matrix confirmed any minimal significant errors, making the model very useful in cases like a clinical or an app-based predictions. The model primarily used BMI (weight and height) to categorize obesity risk accurately. Additionally, lifestyle factors such as reduced snacking (CAEC) and active transportation moderately influenced outcomes, showing

Both models offer strong and accurate predictions suitable for practical decision-making. The linear regression efficiently predicts BMI from basic metrics, while the classification model reliably identifies obesity risk categories. These results emphasize that public health strategies

should prioritize weight reduction, supported by behavioral interventions targeting diet and physical activity.