Meena Mall
Data Analytics
Professor Ahmed
12/2/24

Assignment 7

1) Exploratory Data Analysis (EDA) of Dataset

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

This analysis uses the Obesity Dataset to explore and identify key statistical patterns. The dataset includes various variables such as demographics, lifestyle factors, and obesity status. The goal is to understand the distribution and relationships among the data, spot any issues like outliers, and identify potential features for modeling. The process involved loading the dataset, checking for missing values, summarizing statistics, and conducting visual analysis to uncover patterns, distributions, outliers, and relationships between the variables.

Data Summary and Cleaning

The dataset consists of 2,111 rows and 17 columns, containing both categorical variables (e.g., gender, family history of overweight) and numerical variables (e.g., age, height, weight, BMI). The first step in the cleaning process was checking for missing values. Since none were found, no imputation or removal of data was necessary.

Next, outlier detection was conducted using visual tools like boxplots for numerical variables, particularly focusing on height, weight, and BMI. Based on these plots, no significant outliers were identified that would necessitate removal or transformation.

To ensure consistency and accuracy in the analysis, we calculated BMI (Body Mass Index) using the weight (kg) and height (m) columns, as the dataset contained BMI as an existing variable. We cross-checked the new BMI calculations with the provided values to ensure no discrepancies.

Additionally, summary statistics (e.g., mean, median, standard deviation) were calculated for all numerical columns to assess data distribution and identify any unusual patterns.

In terms of smoothing techniques, no significant smoothing was applied in this case, as the dataset did not appear to have high noise or irregular patterns that would require such methods.

```
> print(outliers)
numeric(0)
```

Figure 1.1: Outliers

Figure 1.1 shows the output of 0 - proving there were no outliers in this dataset.

```
> # Check for missing values
> sum(is.na(obesity_data))
[1] 0
```

Figure 1.2: Missing Values

Figure 1.2 shows the output of 0 - proving there were no missing values in this dataset.

```
> # Check the summary of BMI and other statistics
> summary(obesity_data_clean$BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   24.33   28.72   29.70   36.02   50.81
> # Calculate the standard deviation for key numeric variables
> sd(obesity_data_clean$BMI)
[1] 8.011337
> sd(obesity_data_clean$Age)
[1] 6.345968
> sd(obesity_data_clean$Height)
[1] 0.09330482
> sd(obesity_data_clean$Weight)
[1] 26.19117
```

Figure 1.3: Summary Statistics

Figure 1.3 shows the summary statics within this dataset, including date on BMI in which I calculated to perform these statistical aspects.

Summary Statistics:

- **Age**: The ages of individuals range from 14 to 61 years, with an average age of 24.31 years.
- **Height**: Heights range from 1.45 meters to 1.98 meters, with an average height of 1.70 meters.
- **Weight**: Weights range from 39 kg to 173 kg, with an average weight of 86.59 kg.
- **BMI**: BMI values range from 13.00 to 50.81, with an average BMI of 29.70

Below are the graphics for the Exploratory Data Analysis (EDA), showing data distributions and summaries, along with insights for model approaches.

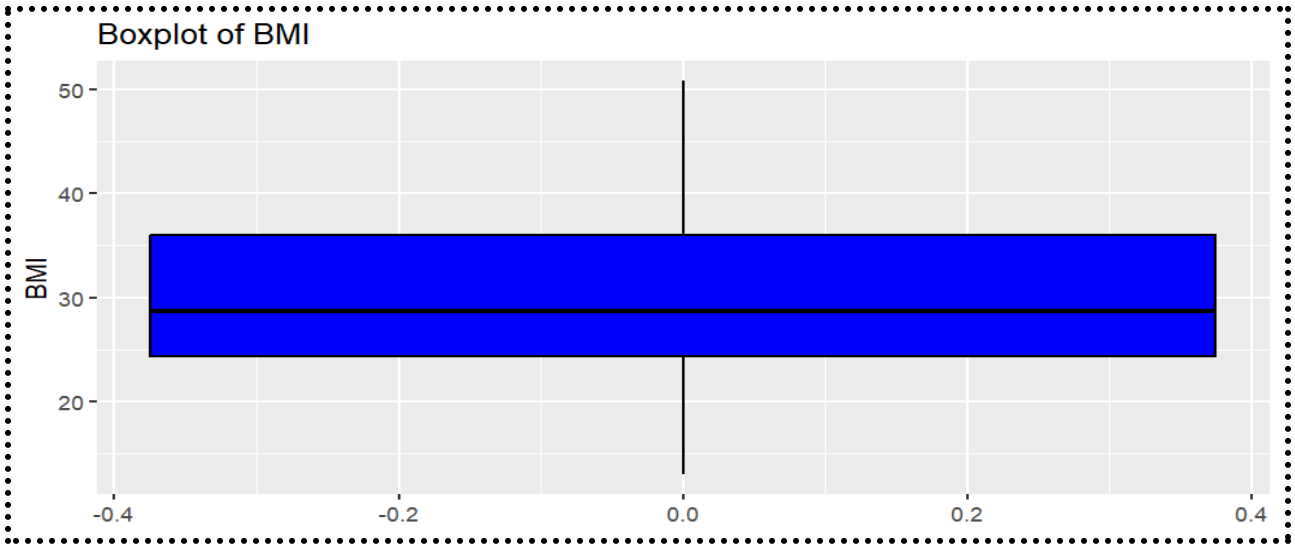

Figure 2: Boxplot of BMI

Figure 2 was created to check for outliers. It showed most values are around the median, with no clear extreme outliers. The data is mostly balanced, with most people in the normal or overweight categories.

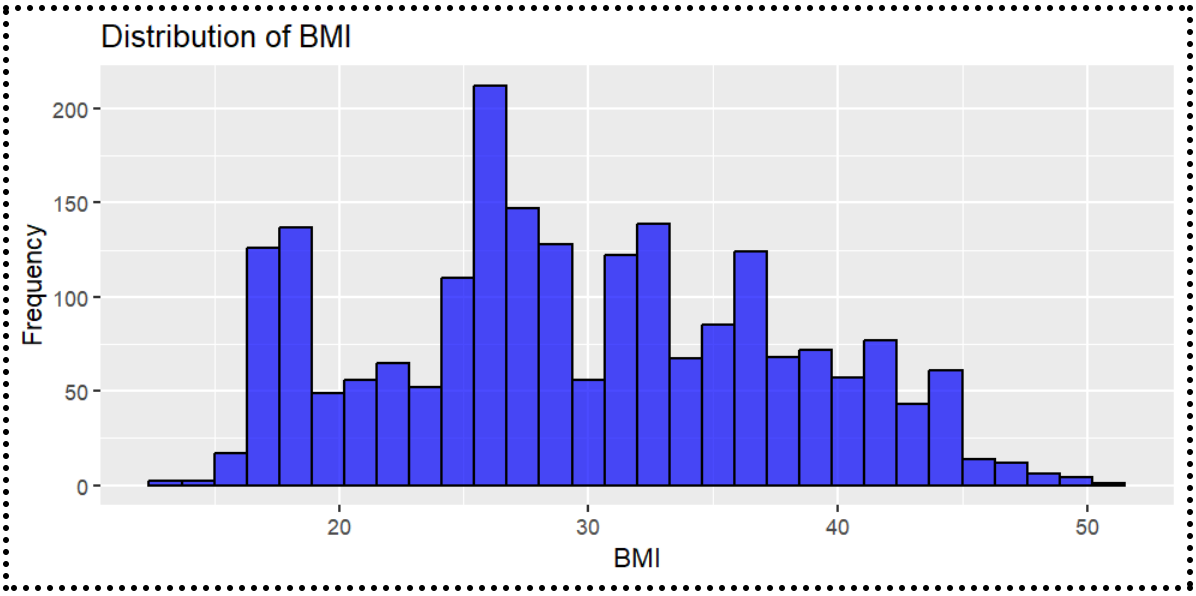

Figure 3: Histogram of BMI

Figure 3 shows a balanced distribution, with most values clustering around the 25-30 BMI range. This suggests that a significant portion of individuals fall into the overweight or borderline obese category, with fewer individuals in the very low or high BMI ranges.

	Age	Height	Weight	BMI
Age	1.00000000	-0.02595813	0.2025601	0.2441631
Height	-0.02595813	1.00000000	0.4631361	0.1317845
Weight	0.20256010	0.46313612	1.0000000	0.9348057
BMI	0.24416312	0.13178454	0.9348057	1.0000000

Figure 4: Correlation Matrix

Figure 4 shows the different factors that relate to BMI. The analysis found a very strong correlation (0.93) between BMI and weight, meaning weight directly affects BMI. There was a moderate correlation (0.24) between BMI and age, showing that older people tend to have slightly higher BMIs. Lastly, BMI and height had a weak negative correlation (-0.13), suggesting taller people might have slightly lower BMIs.

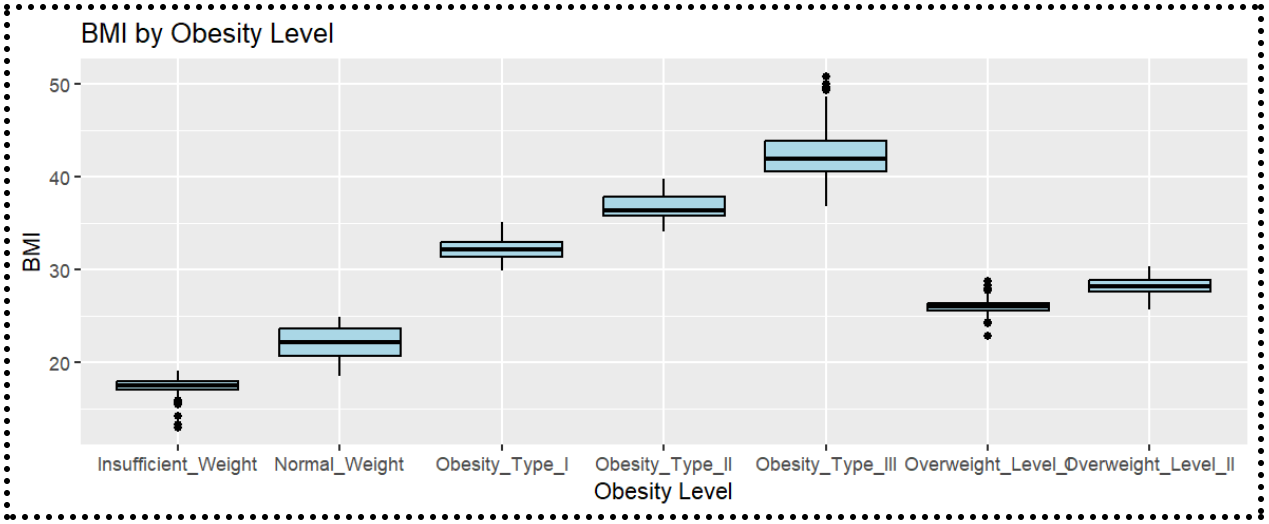

Figure 5: BMI by Obesity Level

Figure 5 shows that people who are classified as "Overweight" or "Obese" generally have higher BMI values compared to those who are classified as "Normal Weight."

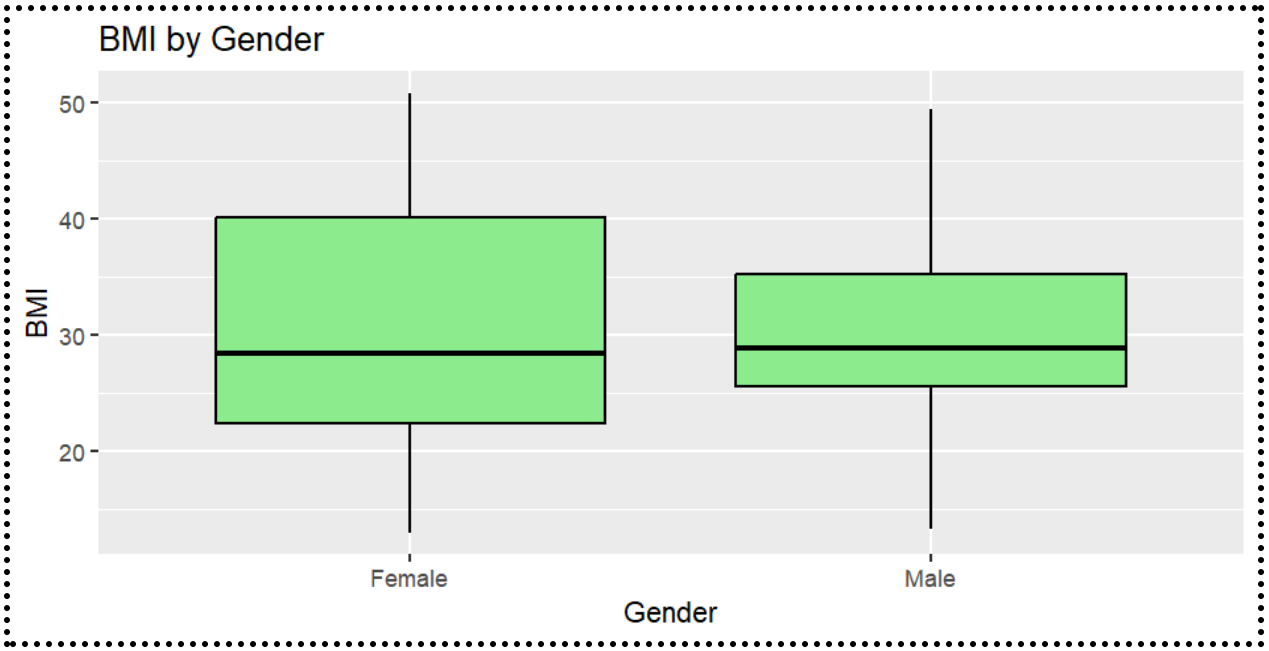

Figure 6: BMI by Gender

Figure 6 shows a small difference in BMI between males and females, with females having a slightly lower average BMI than males.

2) Model Development, Validation and Optimization

For this analysis, I chose three models for different reasons: Linear Regression, Logistic Regression, and K-means Clustering. Linear Regression is used to predict BMI, which is a number that can change. Logistic Regression helps determine whether someone is obese or not. K-means Clustering groups similar data together to find patterns without any pre-set categories.

I chose these models to address different goals in the analysis. Linear Regression was used to predict continuous values like BMI, as it helps estimate a number that can vary. Logistic Regression was selected to classify obesity status, determining whether someone is obese or not. Lastly, K-means Clustering was chosen to find natural groups within the data, helping to identify patterns without using pre-set categories.

Validation Approach:

- **Linear Regression:** I used cross-validation with the caret package to check how reliable the model is.
- **Logistic Regression:** I evaluated the model's performance using a confusion matrix, which measures accuracy, precision, recall, and F1-score.
- **K-means Clustering:** I looked at the clustering results visually to see how well the data was grouped into meaningful sections.

Model Results

The **Linear Regression** model, Figure 7, worked really well with an R-squared value of 0.9211, which means it's good at predicting BMI. The Mean Squared Error (MSE) was 5.00, and the main factors that affected BMI were age, gender, and weight. The results showed that being male lowered BMI, while age and weight increased it. After testing the model with cross-validation, the RMSE was 2.28, the R-squared improved slightly to 0.9215, and the Mean Absolute Error (MAE) was 1.83, showing that the model's predictions were more accurate.

```
Linear Regression

1477 samples
   4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1330, 1330, 1329, 1329, 1329, 1330, ...
Resampling results:

  RMSE      Rsquared   MAE
  2.276648  0.9214519  1.827438

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 7: Linear Regression

For **Logistic Regression**, Figure 8, the goal was to predict obesity. The key factors in predicting obesity were weight and gender (male). A family history of being overweight didn't significantly affect the prediction. The model's accuracy, measured with a confusion matrix, showed 323 correct predictions of non-obese (true negatives), 261 correct predictions of obese (true positives), 28 false positives (non-obese predicted as obese), and 22 false negatives (obese predicted as non-obese).

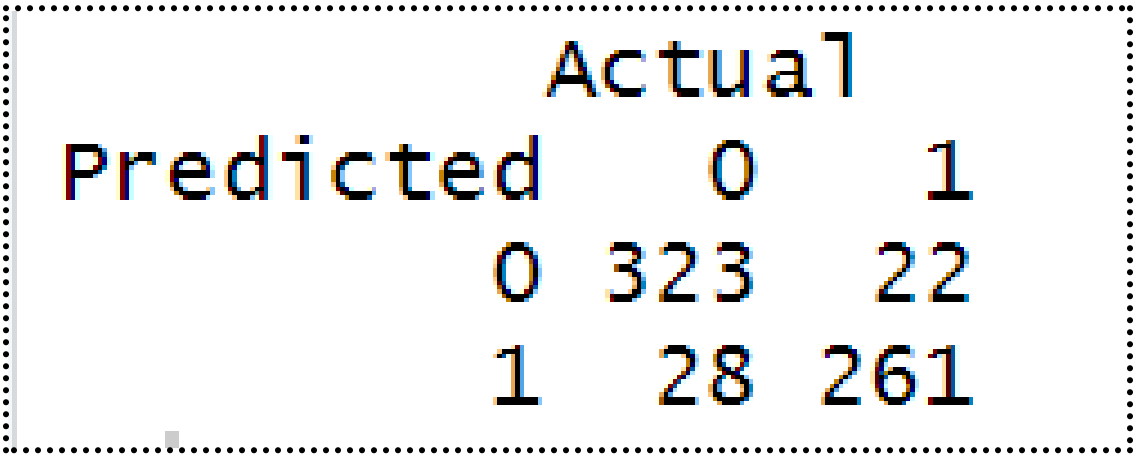

Figure 8: Logistic Regression

In the **K-means Clustering** model, Figure 9, three clusters were found with 811, 918, and 382 individuals in each. These clusters were based on factors like age, gender, weight, and family history of overweight. Looking at the clusters visually showed meaningful groupings based on these factors.

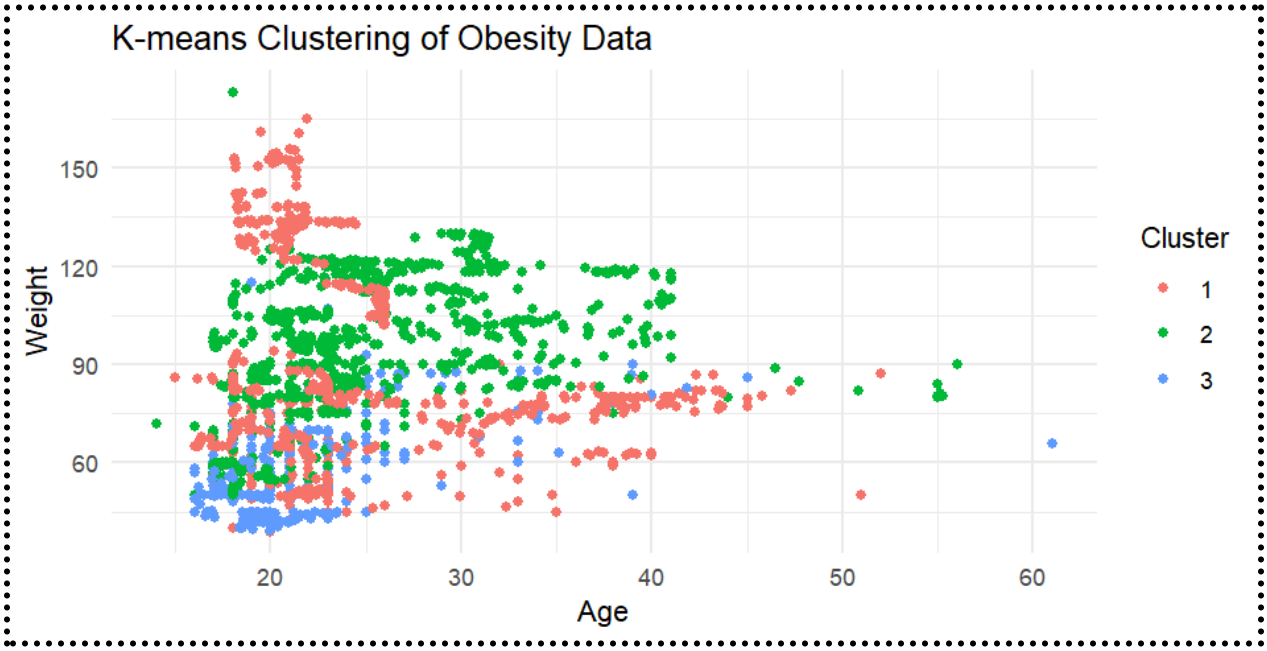

Figure 9: K Clustering

Result Summary:

The Linear Regression model had an R-squared of 0.9211, meaning it explains most of the changes in BMI, with weight and age being the main factors. The Logistic Regression model used a confusion matrix to check how accurate it was at predicting obesity, with weight and gender being important predictors. In K-means Clustering, three separate groups were found based on age, weight, and gender, helping to find patterns in the data.

3) Decisions

After building and testing the models, I looked at the results to understand how well each model worked and how it could help with decision-making.

Linear Regression Model (BMI Prediction):
The linear regression model for predicting BMI showed strong results, with an R-squared value of 0.9211, meaning it explains over 92% of the variation in BMI using factors like age, gender, and weight. This high value shows the model is good at predicting BMI. Key factors like age and weight had positive links to BMI, while being male had a negative impact. The model's prediction errors were low, with a Mean Squared Error (MSE) of 5.00 and a Root Mean Squared Error (RMSE) of 2.28, showing it was accurate. This model can be useful for making general predictions and understanding trends, but it has some limitations. It might not catch all the patterns, especially in more complex or unusual cases. While it can help in many situations, it should be used carefully and not relied on too much when dealing with diverse or complicated data.

Logistic Regression Model (Obesity Classification):
The logistic regression model did a good job of predicting whether someone was obese. It correctly identified 323 non-obese people and 261 obese people, showing it was accurate in classifying obesity risk. However, it also made some mistakes, misclassifying 28 non-obese people as obese and 22 obese people as non-obese. This shows there's room for improvement. This model could help healthcare professionals spot people at risk of obesity and offer early help. The key factors—weight and gender (male)—are known to be linked to obesity. To make the model more accurate, it could include other factors like diet or exercise.

K-means Clustering:
The K-means clustering model found three separate groups based on factors like age, gender, weight, and family history. The sizes of these groups were 811, 918, and 382. These groups can help us understand different types of people with similar traits, which can be useful for things like health plans, marketing, or creating personalized wellness programs. By looking at a visual of the clusters, we can see patterns and trends in the data.

Overall Conclusion:
In summary, the three models give a well-rounded analysis of the dataset:

- The Linear Regression model provides reliable BMI predictions, helping healthcare professionals track health changes.
- The Logistic Regression model identifies individuals at risk of obesity, giving healthcare providers valuable information for targeted interventions.

- K-means Clustering finds natural groups in the data, which could be useful for personalized health plans or marketing strategies, but it needs further analysis to be truly useful.

These models have practical uses in fields like healthcare and wellness, but to make better decisions in the real world, they need continuous updates and improvements, such as adding more data or features to improve accuracy.