NUTRITIONAL DIETARY DATA

GROUP 13

SIMPAO, CHARIZE R.

TANYAG, LORD EXZEL JHONNE L.

INTRODUCTION

In this report, the exploratory analysis and regression of a set of data Nutritional Dietary data that highlights nutritional intake, physical activity, and body composition are provided. The data set will offer clues about the variables that can possibly affect the body fat percentage among other health measurements. In this analysis, the aim is to get distributions of these variables, relationships and to model prediction of Body Fat Percentage. The column " Patient.ID " was omitted in the preprocessing phase, and lines with missing values were omitted to guarantee the quality of data.

METHODOLOGY

The analysis involved the following steps:

Data Cleaning
Summary Statistics
Visualization

- Scatter Plot
- Histogram
- Boxplot
- Bar Plot of Physical Activity by BMI Group

Multiple Linear Regression (MLR)

To investigate the dataset, different visualizations and analyses were carried out. The relationship between them was measured in a scatter plot of the Muscle Mass (kg) against the Body Fat Percentage with a linear trendline. All the numeric variables were used to prepare histograms and boxplots to consider sample distribution, tendencies, and outliers. Distribution of the physical activity according to BMI catagories was presented using bar plot, where the catagories were identified according to WHO standards. A multiple linear regression was created in order to seek beforehand the Body Fat Percentage based on a number of presages, which were muscle mass, nutrient consumption, and water intake. An assessment of the model was done with summary () function and diagnostic plots to ensure assumptions. Another measure of the performance of the model was the scatter plot of the predicted versus the actual values.

RESULTS AND FIGURES

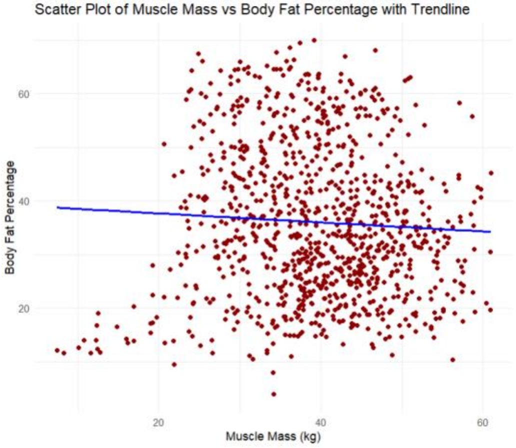
The data, for convenience, consists of 1000 observations, 11 numeric variables that are tied to body composition, physical activities, and diet intake after clean-up. The summary statistics of these variables are presented in table 1. Namely, the Body Fat Percentage varies by quite a lot and goes between 3.86 percent and 70 percent, with the average that amounts to 36.03 percent. The Muscle Mass:kg means 39.46 kg with about 2449 kcal a day of caloric.

Table 1. Summary Statistics for Numeric Variables in the Cleaned Nutritional and Dietary Dataset.

Statistic	Body_Fat_percent	Muscle_Mass_kg	BMI	Physical_Activity_Hours_Week	Daily_Caloric_Intake_kcal	Protein_intake_g	Fat_intake_g	Carbohydrate_intake_g	Vitamin_C_mg	Iron_mg	Water_intake_ml
Min	3.86	7.44	14.8	0	311	10	5	21	50	5	1000
Max	70	60.98	30.4	16	4509	240	200	499	2000	40	4000
Mean	36.03	39.46	23.01	7.8	2449.12	126.23	102.52	255.38	743.3	18.17	2586
Median	33.83	39.7	23.05	7	2482.5	126.5	103	255	500	18	3000
SD	14.08	9.21	2.8	4.82	797.77	67.2	56.55	138.37	689.14	13.99	1118.86
Count	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

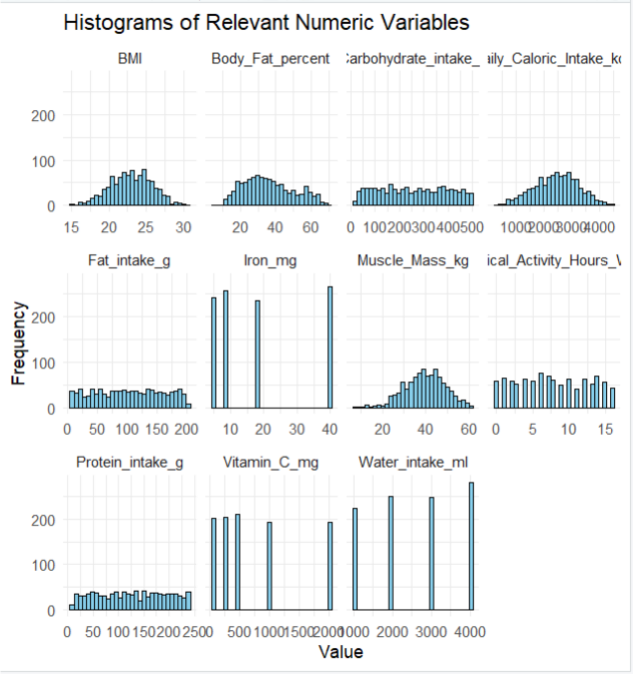
This graph displays barely negative or rather insignificant relationship between body fat percentage and muscle mass. Although a few other people in a higher muscle mass might be a tad lower in the body fat, the general correlation seems not so strong. The body fat percentage is plotted against Muscle Mass (kg) in the scatter plot you see here. A blue line shows the linear regression fit which qualifies the overall tendency of the two variables. Every point in red is the combination of Muscle Mass and Body Fat Percentage measurements of one person.

Figure 1. Scatter Plot of Muscle Mass vs Body Fat Percentage with Trendline.



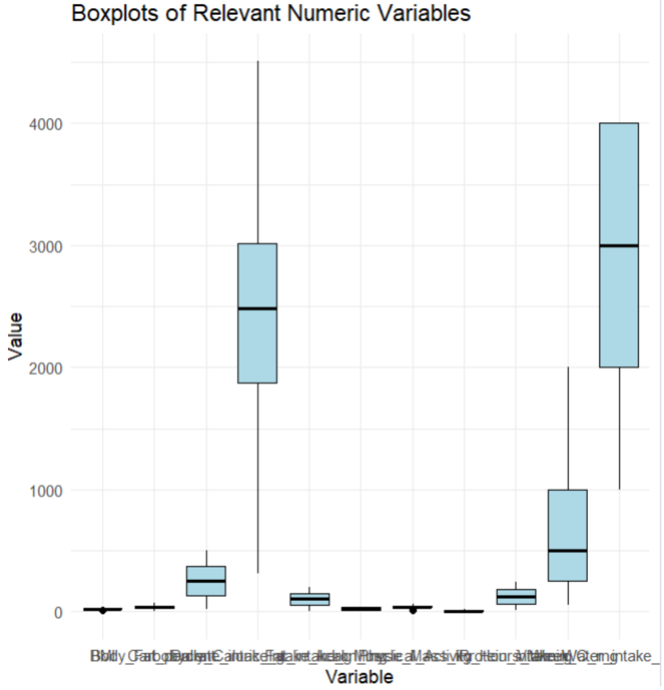
There are lumpy distributions as depicted by the histograms. BMI, Body_Fat_percent, Muscle_Mass_kg, Daily_Caloric_Intake_kcal, and Fat_intake_g are a little skewed or fairly normal distributed. On the other hand, such variables as Iron_mg, Vitamin_C_mg, and Water_intake_ml have very different peaks or even more discrete distributions, which may be explained by standardized intake recommendations or measurement. (Figure 2).

Figure 2. Histograms of Relevant Numeric Variables.



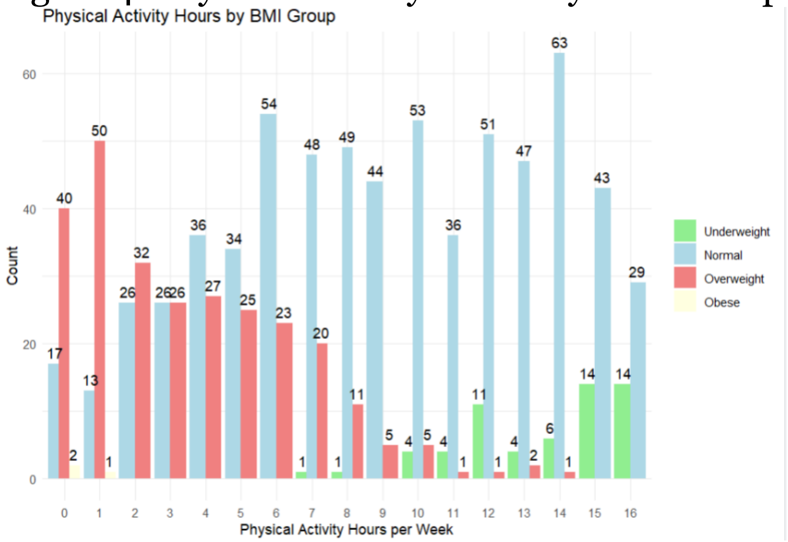
Boxplots allow visualization of the distributions and showing possible outliers. It is important to note that the range and spread of Winder_intake_ml was broad and with relatively high spread as well, whereas the spread of Daily_Caloric_Intake_kcal was also large. There are also variables like iron_mg and vitamin_c_mg that are more discrete in their clustering or number unique objects. A boxplot is a statistical summary of a single variable; the median (the horizonal line in the box), the interquartile range (the box) and possible outliers (the points out of the whiskers). The boxes are light blue and have black lines. (Figure 3).

Figure 3. Boxplots of Relevant Numeric Variables.



This bar plot informs about the representation of categories of BMI at various levels of the weekly physical activity. It recommends that people who fit in the category of Normal BMI are more likely to have a broader number of hours involving physical activity and a consistent proportion of diverse levels of activities. The category of this group includes also the group of Overweight and is notably represented, particularly in lower and moderate activity. The Count of individuals is lower in the groups of Underweight and Obese and demonstrate a different pattern by the volume of Physical Activity Hours per Week. his bar plot indicates the number of people in various classifications (Underweight, Normal, Overweight, Obese) according to the level of Physical Activity Hours per Week. Bars have been avoided to compare the BMI groups at each level of physical activity and the different colors are assigned to different BMI groups, (lightgreen to Underweight, lightblue to Normal, lightcoral to Overweight, lightyellow to Obese). (Figure 4).

Figure 4. Physical Activity Hours by BMI Group.



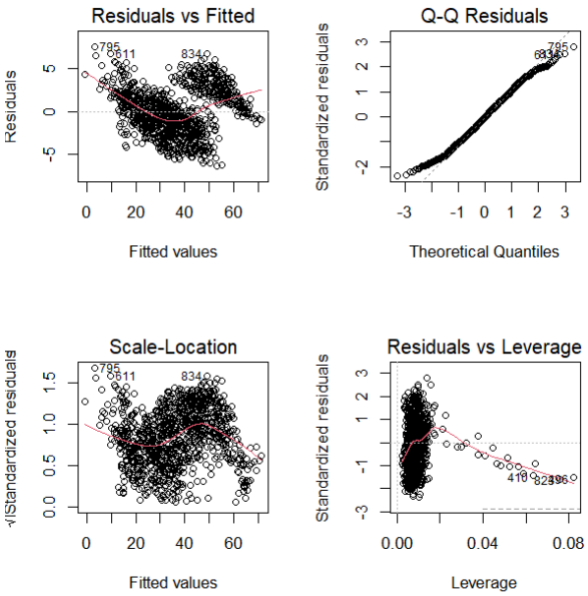
Multiple Linear Regression Model

- Body_Fat_percent was predicted using a fitted multiple linear regression model. In the output, the model summary shows that the adjusted R-squared is high at 0.9622 which implies that the model explains about 96.22 percent of the variance in Body Fat Percentage.

- Model Diagnostic Plots (Figure 5): This is a set of plots that evaluates the assumption of linearity, independence of errors, the normality of residuals and the homoscedasticity. The Residuals vs Fitted scatterplot indicates that the residuals are scattered about zero, but it can give a hint of the non-linearity or heteroscedasticity (such as U-shape pattern). Normal Q-Q plot shows that residuals are normally distributed which has few deviations on its tails. Scale-Location plot indicates the equality of residuals distribution across the range of predictors, and it looks fairly consistent. The residuals vs leverage plot is useful to detect influential obsera vations; some of the points show seemingly high leverage.

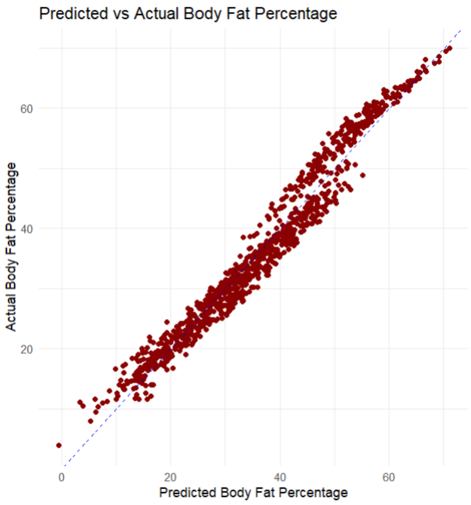
- This multi-paneled display shows four typical diagnostic plots of the multiple-linear regression model predicting Body fat Percentage. These are: (top-left) Residuals vs Fitted values, (top-right) Q-Q Plot of Residuals, (bottom-left) Scale-Location plot and (bottom-right) Residuals vs leverage plot. To check the assumption of the model (linearity along with the normality of the residuals and homoscedasticity), these plots are available in order to determine as well the influential data points.

Figure 5. Diagnostic Plots for the Multiple Linear Regression Model.



The scatter plot depicts effective prediction based on the model. The points are more close to the dashed blue line (which signify perfect prediction), which means that the model would be a highly strong fit to predict Body Fat Percentage. This scatter plot has the actual Body Fat Percentage values within the dataset and the predicted Body Fat Percentage values within the multiple linear regression model. Individual data points are plotted as red points and the dashed blue line (y=x) with the slope of 1 and intercept of 0 represents perfect prediction. The fact that points match this line closely indicate the high degree of predictive performance of the model. (Figure 6).

Figure 6. Predicted vs Actual Body Fat Percentage.



INTERPRETATION AND CONCLUSION

Through this nutritional and dietary data analysis, this study can draw important conclusions regarding the aspect of body fat percentage. Descriptive statistics demonstrate the variety of composition of the body and dietary habits among the cohort. The distribution patterns of different measurements of health-related variables also can be visualized, i.e. the histograms of the boxplots, some of which have unusual discrete distributions. The bar plot of physical activity by level of BMI shows that personnel with a BMI of Normal group mostly depict a consistent kind of engagement with the varying types of physical activity levels.

It was observed that the multiple linear regression model was highly capable of predicting Body Fat Percentage, as the adjusted R-squared value was very high. Some of the key results in a regression are that the association between Muscle Mass, Physical Activity, Protein Intake, and Fat Intake and the Body Fat Percentage is strong and negative. On the contrary, Daily Caloric Intake revealed significant positive correlation. The advantages of the diagnostic plots are that they basically confirm that the model assumptions are met, however, there exists a slight trend in the residuals which should have some minor attention. All in all, these findings highlight that muscle mass, PA, and balanced intake of macronutrients tip the scales toward fat sometime some way. Since the predictive effect of this model is very high, these five factors combine as a strong indicator of determining body fat percentage in this data set.