Introduction/Motivation: The primary objective of this study was to build a predictive model for body fat percentage based on anthropometric measurements.
Background Information: The dataset comprises observations of multiple body metrics for 253 men. Our analysis utilized linear regression models to establish a reliable predictor of body fat based on measurements such as age, weight, height, neck, and waist circumference.
Data Cleaning: Removing Outliers: To clean the data, we then removed for height and weight two particular values that were outliers. For the remaining predictors, we used the IQR ranges to filter out outliers, and the size of the dataset went from 252 observations down to 245 observations, which serves the purpose of cleaning the data; while not losing many data points out of an already small dataset.

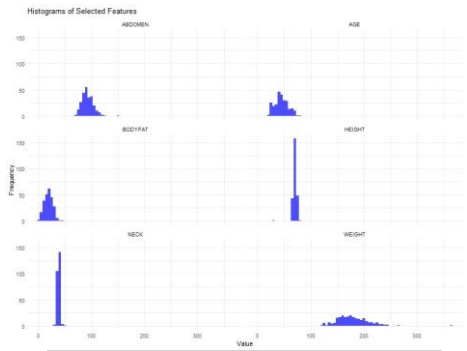

Figure 1. Histogram of original data

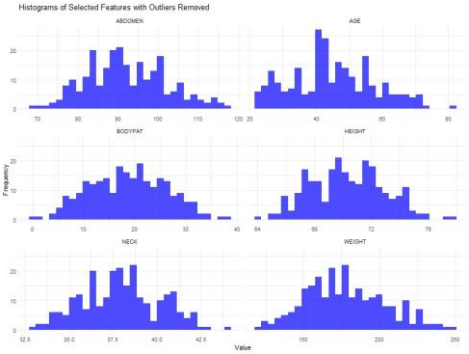

Figure 2. Histogram after removing outliers

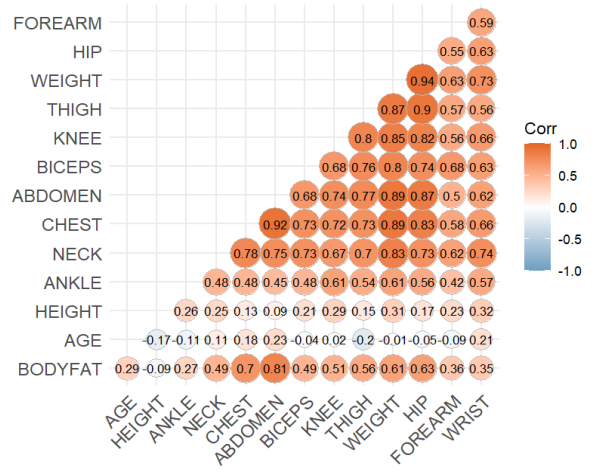

Figure 3. Visualize the Correlation Matrix using Heatmap

Feature Selection: We initially eliminated one of the predictors (density) because we believe that defeats the purpose of creating a "simple" body fat calculator. We then did some research on available literature and trusted online body fat estimators. We found the 5 features (Age, Height, Weight, Neck and Abdomen) to be agreed on in most of the references we found. [1-3]

Correlation Check: Before proceeding, we wanted to examine the correlations among the predictors. Based on our findings, we chose to test adding (Chest and Biceps).
Mode Parameters Hyper-tuning: We employed linear regression models for prediction. Two approaches were used: a traditional 80-20 train-test split and a 10-fold cross-validation. The latter was chosen for its robustness, based on the following linear regression model assumptions: **Linearity;** Checked using the "Residuals vs Fitted" plot**, Normality** using Q-Q plot, and **Homoscedasticity** using "Residuals vs Fitted" plot

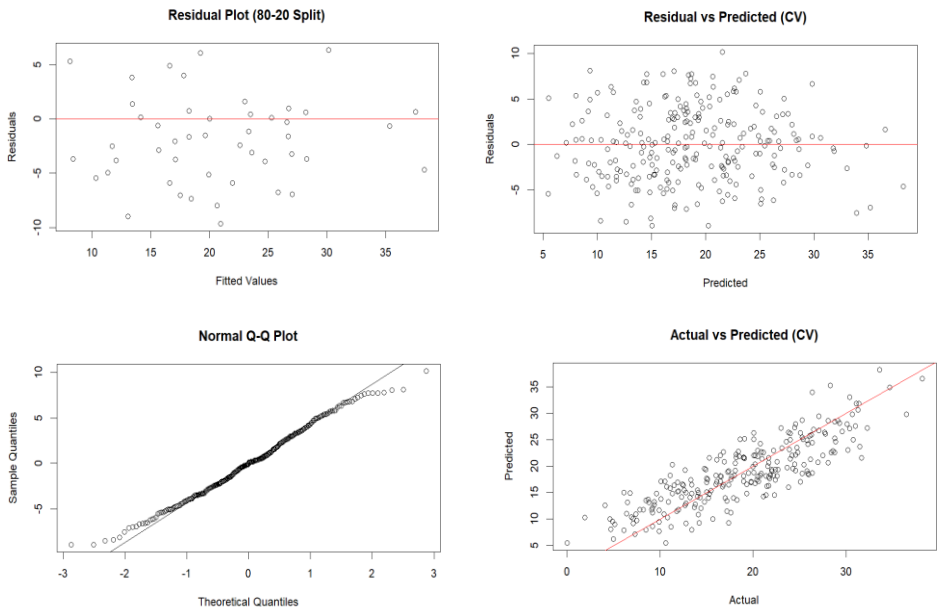

Figure 4. Model Results Plots for 80-20Split vs Cross-Validation

Finally, we build the 4 different models as follows: (A) Predictors: Age, weight, height, neck, and waist (B) Adding Chest (C) Adding Biceps (D) Adding Chest and Biceps

Results: Considering all the metrics in conjunction yields a robust evaluation. Based on all three, no significant improvement was noticed in adding more features, I.e., model D, resulted in minor improvement from model A, which is much simpler. As a result, we choose to go with model A.

Model	R2	RMSE	Adj_R2
fit_cv_A	0.71	3.98	0.71
fit_cv_B	0.71	3.97	0.71
fit_cv_C	0.71	3.97	0.71
fit_cv_D	0.72	3.95	0.71

Model and Coefficients: We established the following linear regression model: BODYFAT = -21.29 - 0.0073 * AGE - 0.0902 * WEIGHT - 0.1847 * HEIGHT - 0.3180 * NECK + 0.8876 * ABDOMEN. Each coefficient for the independent variables represents their impact on BODYFAT. ABDOMEN has the most significant impact, with an average increase of 0.8876 percentage points for every 1 unit increase in ABDOMEN.**Goodness of Fit:** With a multiple R-squared of 0.713, the model can explain 71.3% of variation in BODYFAT, indicating a good fit.**Significance:** The F-statistic is 95.9, and the corresponding p-value is much less than 0.05, indicating that the overall fit of the model is statistically significant.

Example Usage: For a 40-year-old male with a weight of 80 kg, height of 170 cm, neck circumference of 35 cm, and ABDOMEN measurement of 90 cm, the estimated BODYFAT percentage is approximately 15.25%. The 95% interval for this estimate is 14.5%-16%.

Strengths: (1) Robust cross-validation methodology (2) Model simplicity

Weaknesses: (1) May perform poorly on populationsoutside the dataset. (2) The model could not account for other potential confounders such as exercise. (3) Sensitive to outliers

Conclusion: The model provides a reasonably accurate, robust, and interpretable method for predicting body fat percentage based on easily obtainable measurements. While the model is useful, it is important to consider its limitations and the influence of unmeasured factors.

Reference

- Precision Nutrition. (n.d.). Body fat calculator: How to measure body fat percentage. Retrieved from https://www.precisionnutrition.com/body-fat-calculator
- Calculator.net. (n.d.). Body fat calculator. Retrieved from https://www.calculator.net/body-fat-calculator.html
- National Academy of Sports Medicine (NASM). (n.d.). Body fat calculator. Retrieved from https://www.nasm.org/resources/body-fat-calculator

Contributions:

	Contribution Table		
Contributions	Osama Kheshaifaty	CHUYI LIN	XIAOYANG DONG
Presentation	Made the slides for model evaluation and data cleaning	Made the presentation slides and added all the plots from summary	Created and summarized the model evaluation slides
Summary	Wrote the model's hypertuning part and the correlation check	Documented and wrote the model evaluation, feature selection, and data cleaning summary	Updated the introduction and the background/diagnosis proces, and made the contribution table
Code	Coded the model evaluation part for cross-validation/features	Conducted the statistical model evaluation	Coded the feature correlation analysis and the data cleaning
Shiny App	Developed the server components of the shiny app	Reviewed the team's input on the shiny app code	Developed the UI components of the shiny app