Project A
- Regression Analysis -
Body Fat Percentage Estimation Based on Physical Measurements

Author: Driton Krasniqi

• Introduction

This project aims to estimate body fat percentage based on various physical measurements that can be easily collected using a scale and measuring tape. Body fat percentage is an important health indicator that can be difficult to measure directly without specialized equipment, making predictive models based on simple measurements a valuable tool.

Project analyzes the relationship between body fat percentage and various physical measurements, developing a regression model to predict body fat percentage using indicators that can be easily measured with a scale and measuring tape.

The analysis explores the relationship between body fat percentage and various indicators such as age, weight, height, and ten body circumference measurements, with the goal of developing a reliable regression model for predicting body fat percentage.

• Data Understanding

- Dataset Description

The dataset contains measurements from 252 individuals, including:
- Target variable: PercentBodyFat
- Features: Age, Weight, Height, and ten body circumference measurements (Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist)

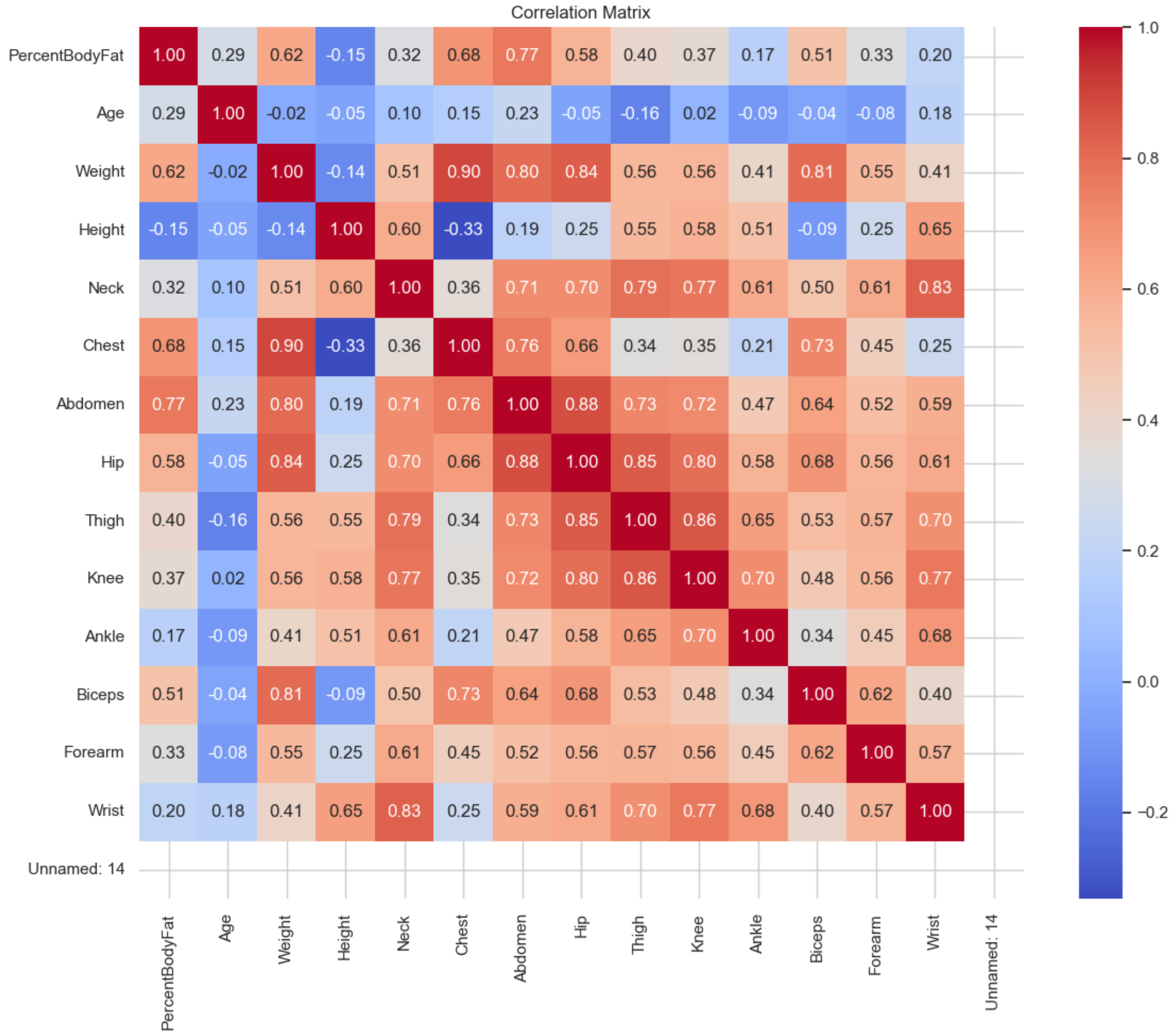
- Exploratory Data Analysis

Initial data exploration revealed:
- The dataset contained 252 observations with 15 variables
- Some extreme values in the body fat percentage (including values near 0%, which are physiologically implausible)
- After removing extreme values (< 3% body fat), 249 observations remained for analysis

The distribution of body fat percentage follows an approximately normal distribution, with some variation in the range.

- Correlation Analysis

The correlation matrix revealed strong relationships between many variables:



Correlation Matrix

	PercentBodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist	Unnamed: 14
PercentBodyFat	1.00	0.29	0.62	-0.15	0.32	0.68	0.77	0.58	0.40	0.37	0.17	0.51	0.33	0.20	
Age	0.29	1.00	-0.02	-0.05	0.10	0.15	0.23	-0.05	-0.16	0.02	-0.09	-0.04	-0.08	0.18	
Weight	0.62	-0.02	1.00	-0.14	0.51	0.90	0.80	0.84	0.56	0.56	0.41	0.81	0.55	0.41	
Height	-0.15	-0.05	-0.14	1.00	0.60	-0.33	0.19	0.25	0.55	0.58	0.51	-0.09	0.25	0.65	
Neck	0.32	0.10	0.51	0.60	1.00	0.36	0.71	0.70	0.79	0.77	0.61	0.50	0.61	0.83	
Chest	0.68	0.15	0.90	-0.33	0.36	1.00	0.76	0.66	0.34	0.35	0.21	0.73	0.45	0.25	
Abdomen	0.77	0.23	0.80	0.19	0.71	0.76	1.00	0.88	0.73	0.72	0.47	0.64	0.52	0.59	
Hip	0.58	-0.05	0.84	0.25	0.70	0.66	0.88	1.00	0.85	0.80	0.58	0.68	0.56	0.61	
Thigh	0.40	-0.16	0.56	0.55	0.79	0.34	0.73	0.85	1.00	0.86	0.65	0.53	0.57	0.70	
Knee	0.37	0.02	0.56	0.58	0.77	0.35	0.72	0.80	0.86	1.00	0.70	0.48	0.56	0.77	
Ankle	0.17	-0.09	0.41	0.51	0.61	0.21	0.47	0.58	0.65	0.70	1.00	0.34	0.45	0.68	
Biceps	0.51	-0.04	0.81	-0.09	0.50	0.73	0.64	0.68	0.53	0.48	0.34	1.00	0.62	0.40	
Forearm	0.33	-0.08	0.55	0.25	0.61	0.45	0.52	0.56	0.57	0.56	0.45	0.62	1.00	0.57	
Wrist	0.20	0.18	0.41	0.65	0.83	0.25	0.59	0.61	0.70	0.77	0.68	0.40	0.57	1.00	
Unnamed: 14															

Key findings from correlation analysis:
- Abdomen circumference had the strongest correlation with body fat percentage (r = 0.80)
- Other strong correlates include Chest (r = 0.69) and Hip (r = 0.61)
- Height showed a weak negative correlation with body fat percentage (r = -0.09)

Many measurements are highly intercorrelated, particularly the various circumference measurements, indicating potential multicollinearity issues in modeling.

- **Methodology**

- Data Preprocessing

The following steps were performed:
1. Removed an unused column from the dataset
2. Identified and handled outliers in the target variable
3. Removed physiologically implausible values (body fat < 3%)
4. Split the data into training (80%) and testing (20%) sets

- Feature Selection Methods

Three different feature selection approaches were used to identify the most important predictors:

1. **Random Forest Feature Importance**
   - Used the built-in feature importance measure from Random Forest
   - Provides a measure of each feature's contribution to prediction accuracy

2. **Recursive Feature Elimination (RFE)**
   - Sequentially removed the least important features
   - Used linear regression as the base estimator
   - Selected the top 5 features

3. **Sequential R^2 Analysis**
   - Calculated individual R^2 for each feature
   - Measured the incremental change in R^2 as features were added in order of importance
   - Provided insight into the unique contribution of each feature

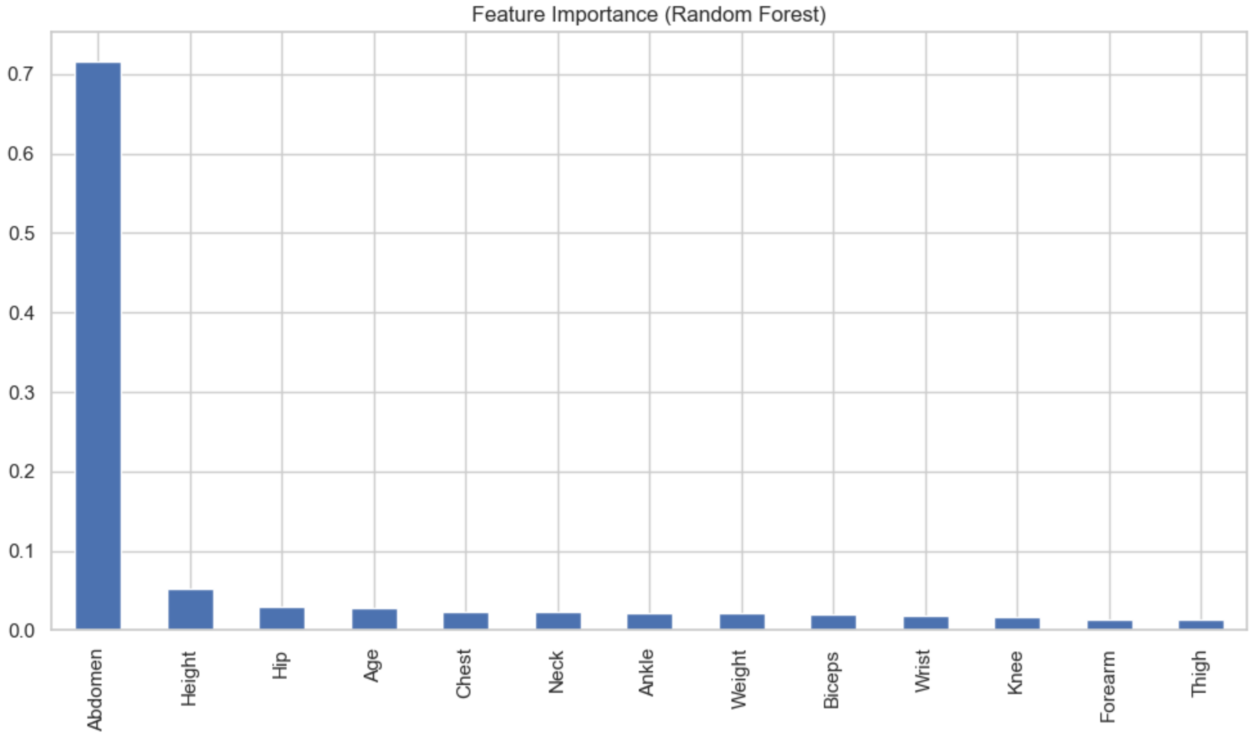
- Model Building

Several multiple linear regression models were developed:
1. Full model with all features
2. Reduced model with the top 5 features from R^2 analysis
3. Model with features selected by RFE

Cross-validation was used to assess model stability and generalization ability.

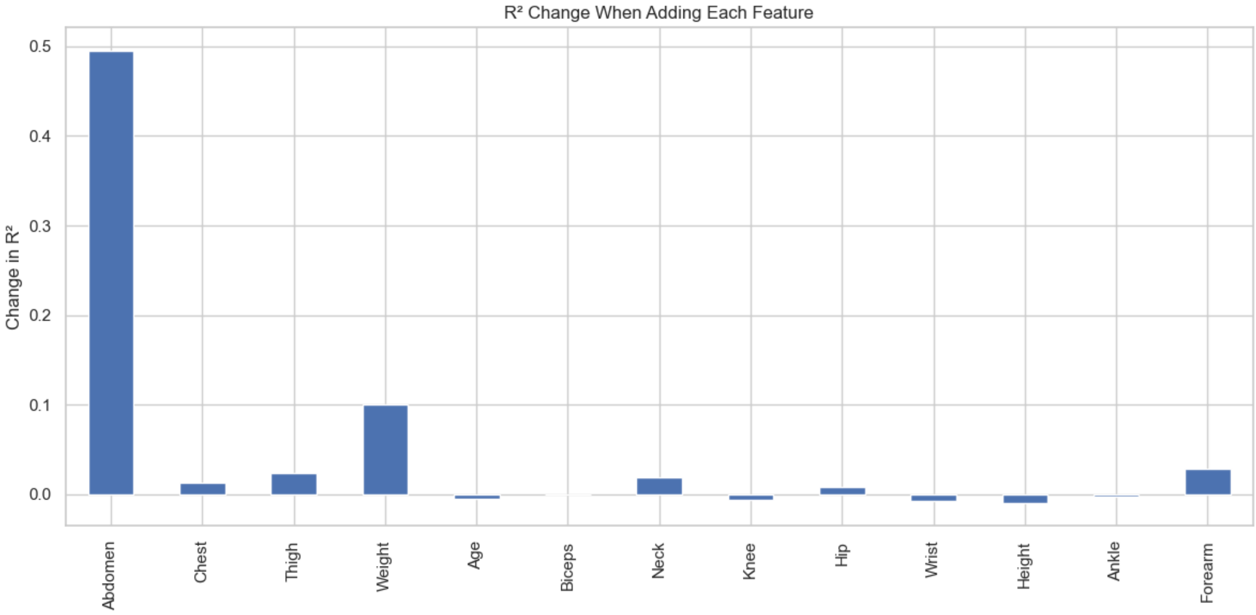
- **Results**

- Feature Importance
Random Forest Feature Importance



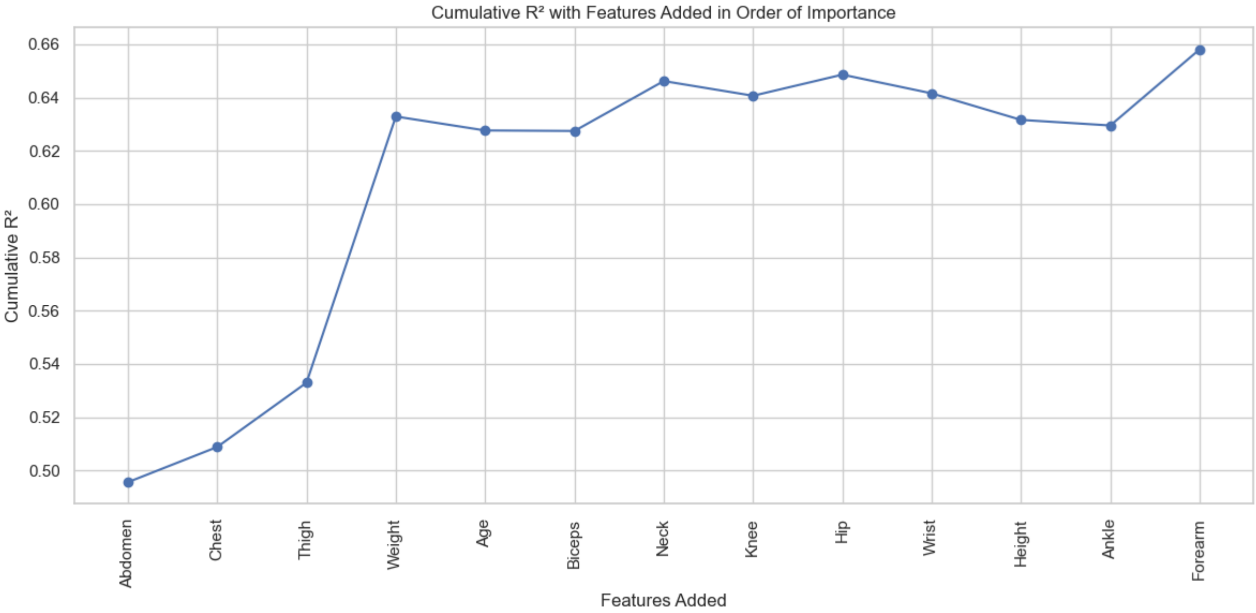
The Random Forest analysis identified Abdomen as the most important feature by a significant margin, followed by Height, Hip, Age, and Chest.

Incremental R² Analysis



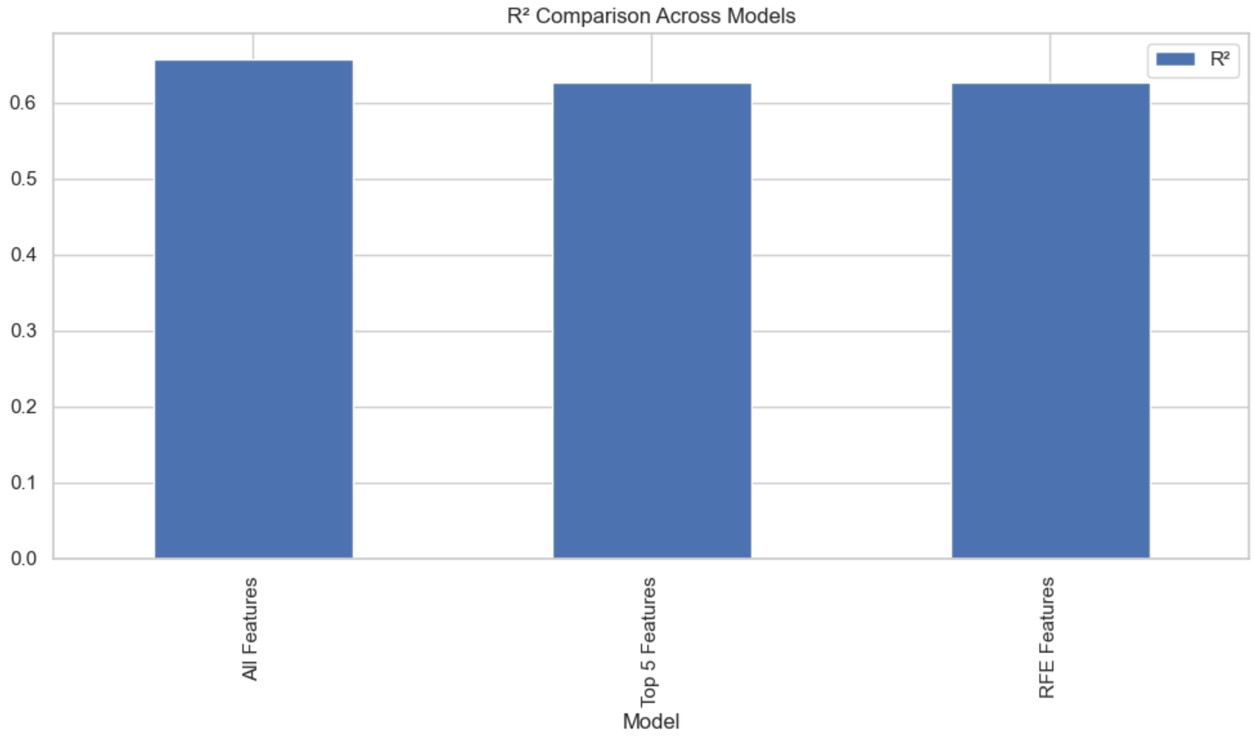
The R² analysis showed that:

- Abdomen alone explains approximately 50% of the variance
- Weight and Chest provide significant additional predictive value
- The incremental benefit diminishes with additional features



The cumulative R² chart shows that most of the predictive power is achieved with just a few features.

- Model Performance Comparison



Model performance comparison:
- Full model with all 13 features: $R^2 = 0.658$, RMSE = 4.49
- Top 5 features model: $R^2 = 0.628$, RMSE = 4.68
- RFE selected features model: $R^2 = 0.628$, RMSE = 4.68

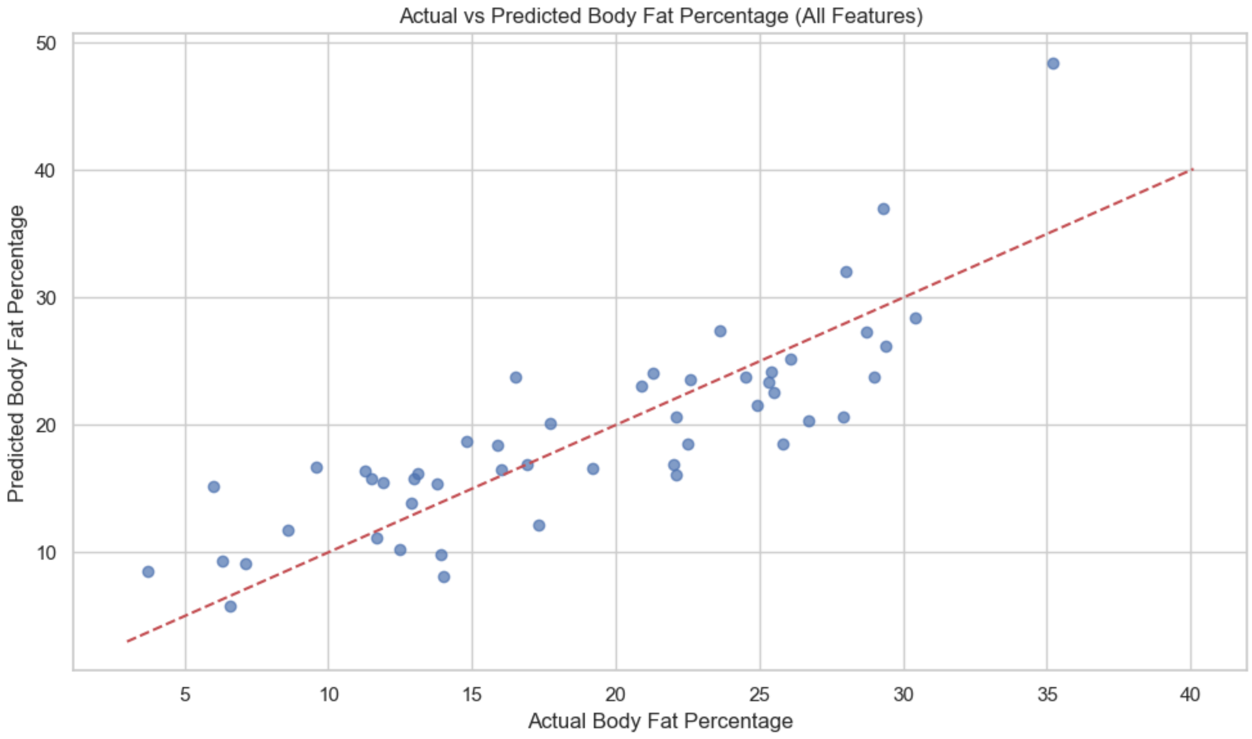
The full model provides the best performance, but the reduced models offer comparable performance with fewer features.

- Regression Equation

The best regression model is:

PercentBodyFat = -13.1545 + 0.9156 × Abdomen + 0.2633 × Forearm + 0.2194 × Ankle + 0.2180 × Thigh + 0.2062 × Biceps + 0.0708 × Age - 0.0075 × Knee - 0.0638 × Weight - 0.0663 × Chest - 0.1240 × Height - 0.1676 × Hip - 0.2883 × Neck - 1.8070 × Wrist

Key observations from the coefficients:
- Abdomen has the largest positive coefficient (0.916)
- Wrist has the largest negative coefficient (-1.807)
- Some coefficients have unexpected signs due to multicollinearity



• Interpretation of Results

The analysis provides answers to the three main questions:

1. **Which indicators provide reliable estimation of body fat percentage?**
 - Abdomen circumference is the most reliable single predictor
 - Other important indicators include Chest, Hip, and Weight
 - The combination of multiple measurements significantly improves prediction accuracy

2. **Best regression model for body fat percentage prediction:**
 - The full model with all features provides the best performance ($R^2 = 0.658$)
 - However, a simplified model with fewer features can achieve nearly as good performance

3. **Most important variables based on R-squared:**
 - Abdomen alone explains about 50% of variance in body fat percentage
 - Adding Weight increases R^2 by approximately 10%
 - Chest, Neck, and Forearm also provide meaningful contributions

- Limitations and Challenges

Several limitations and challenges were encountered:

1. **Multicollinearity:**
   - Many of the physical measurements are highly correlated
   - This affects coefficient interpretability and stability
   - The condition number in the statistical analysis indicates strong multicollinearity

2. **Sample representativeness:**
   - The dataset's representativeness of the general population is unknown
   - Model may not generalize well to all demographic groups

3. **Measurement precision:**
   - Physical measurements may contain errors
   - Standardization of measurement technique is important for reliable predictions

- Practical Applications

The developed model has several practical applications:

1. **Health assessment:**
   - Simple way to estimate body fat without specialized equipment
   - Can be used for initial screening in health assessments

2. **Fitness tracking:**
   - Tool for monitoring changes in body composition over time
   - Helps evaluate the effectiveness of fitness or weight management programs

3. **Health research:**
   - Method for estimating body fat in large-scale studies
   - Reduces the need for expensive measurement equipment

• Conclusion

This project successfully developed a regression model to estimate body fat percentage using easily accessible physical measurements. The analysis identified Abdomen circumference as the most important predictor, followed by other circumference measurements like Chest and Hip.

The best model achieved an R^2 of 0.658, indicating that approximately 66% of the variance in body fat percentage can be explained by the physical measurements included in the model. This level of accuracy is suitable for many practical applications where direct measurement of body fat is not feasible.

Future work could focus on addressing the multicollinearity issue, perhaps through dimension reduction techniques, and testing the model on more diverse populations to ensure generalizability.