GROUP 11 - MODULE 2 - EXECUTIVE SUMMARY

Introduction:

The core focus of this report is to explore the methodology behind developing a simplified, yet cost-effective technique for estimating body fat percentage. The ambition was to craft a model that blends simplicity, robustness, and accuracy, making body fat estimation both practical and economical. Recognizing that body fat estimation is multifaceted and lacks a universal solution, this project intricately tries to balance these elements. By delving into statistical analyses and leveraging foundational knowledge, we demonstrate our decision-making process. Such a model can greatly benefit the healthcare and fitness industries, enabling professionals to make more informed decisions about individual health and fitness regimes.

Data pre-processing

During preprocessing, the data undergoes transformations to improve its modeling readiness. Redundant columns are discarded, and anomalies are removed to ensure data consistency. The dataset's richness is augmented by computing the Body Mass Index (BMI) using weight and height, subsequently leading to the exclusion of 'HEIGHT' and 'WEIGHT' columns for simplicity. Outliers are managed using the Interquartile Range (IQR) method, and missing values are addressed using KNN imputation, with prior standardization for accuracy. Features showing significant skewness are log-transformed to suit linear modeling. Finally, a Robust Scaler ensures the data's resilience against outliers and maintains consistent scales between features and the 'BODYFAT' target. This strategy aims to harmonize simplicity, robustness, and accuracy: **simplicity** through concise feature selection, **robustness** via outlier handling, and **accuracy** through imputation, transformation, and other such carefully considered techniques.

Final Model Statement:

Regression Type	Feature combination	R^2	AIC	BIC
Linear	(3)	0.729	187.66	205.27
Robust	(4)	0.729	186.92	208.05
Ridge	(7)	0.714	181.09	205.63

The exact combination can be seen in the reference page, where “*” means variable being selected and “ ” means not. These combinations are selected based on the ‘regsubset’ function in the ‘leaps’ package. It selects 8 relatively best models based on BIC value.

The data illustrates the performance of three regression models: linear, robust, and ridge, each characterized by different feature combinations. The linear regression, with a feature combination of 3, achieves an R^2 value of 0.729, and the associated AIC and BIC values are 187.66 and 205.27, respectively. The robust regression, despite having an additional feature (total of 4), achieves a similar R^2 value of 0.729, but slightly improves the AIC to 186.92 while increasing the BIC to 208.05. Meanwhile, the ridge regression, with the most extensive feature combination of 7, yields a slightly lower R^2 of 0.714. However, it presents the best AIC value at 181.09, with a BIC similar to the linear model at 205.63. Out of these, we chose **linear regression** as our final model.

The expression for the final model can be written as:

$$BODYFAT = 0.1156 + (0.1113 * AGE) + (0.8928 * ABDOMEN) + (-\ 0.2328 * WRIST)$$

Conclusion/Discussion:

Throughout the course of this project, our aim was consistently clear: to predict body fat percentage by prioritizing three foundational pillars - simplicity, robustness, and accuracy. These guiding principles provided a structured framework, ensuring the resulting models and applications were both interpretable and effective.

Our journey began with rigorous data preprocessing. By cleaning, scaling, and transforming the data, we laid a solid foundation, ensuring that subsequent models could perform optimally. This step was crucial, as even the most sophisticated models can falter if the underlying data is not well-prepared.

Following this, our model exploration phase delved into various algorithms, evaluating their strengths, weaknesses, and fit for our dataset. The diverse exploration led to a strategic decision in the model selection phase: a linear regression model with just three pivotal features. This choice effectively highlighted the trade-offs we often had to make between our three guiding pillars, especially given constraints like a limited dataset and time.

The culmination of our efforts was realized in the form of a Shiny app. This interactive platform not only showcases the project's outcomes but also offers real-world applicability. Users can input their data and receive immediate insights into their body fat percentage, bridging the gap between complex data analysis and actionable information.

However, the project wasn't without challenges. The limited dataset posed constraints, and the balance between simplicity, robustness, and accuracy often necessitated tough decisions. Yet, these challenges served as valuable learning experiences, underscoring the intricacies and nuances of data science projects.

Looking to the future, there's a vast horizon of opportunities. The groundwork laid by this project can be expanded to more advanced applications. Envision a platform with real-time data input from wearables, instantaneous feedback on body fat percentage, and even personalized health recommendations. The integration of real-time data would elevate the project, providing users with dynamic insights into their health.

However, such advancements will undoubtedly introduce new challenges, from handling real-time data influx to ensuring the Shiny app's scalability. The ever-present balance between our guiding principles will remain pivotal, especially as the project's complexity increases.

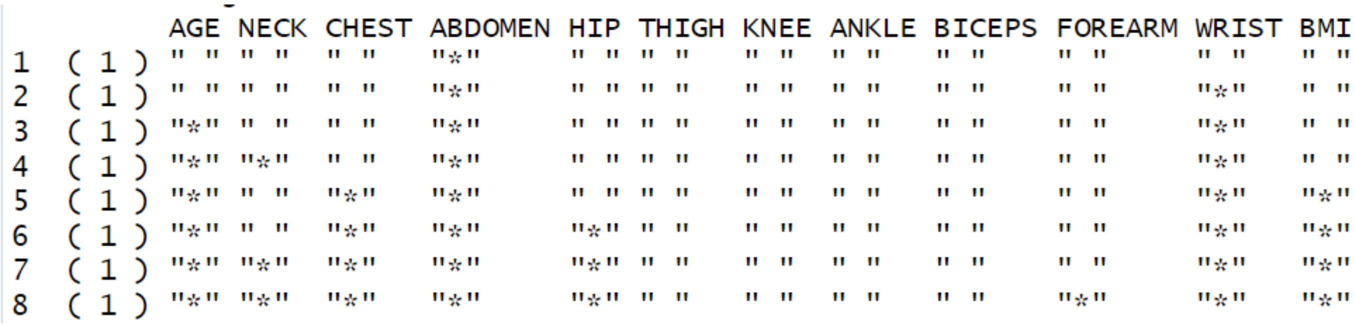
In conclusion, this project, with its four distinct yet interlinked components, underscores the importance of a holistic approach, the challenges inherent in data-driven projects, and the potential transformative impact such endeavors can have in real-world scenarios.

Contributions:

Contribution	Chixu Ni	Shourya Maheshwari	Sreeja Kodati
Presentation	Section 2, 3	Introduction and Section 1, Final edits	Section 4-5
Executive Summary (section-wise)	Final Model Statement	Data Preprocessing, Conclusion, and final edits	Introduction and Contribution
Model Scripting	Model Building, test and Selection	EDA, Data Cleaning and Preprocessing	Model Building
Application	Helped in building the App	Built a Streamlit App (a backup to Shiny)	Built the Shiny App

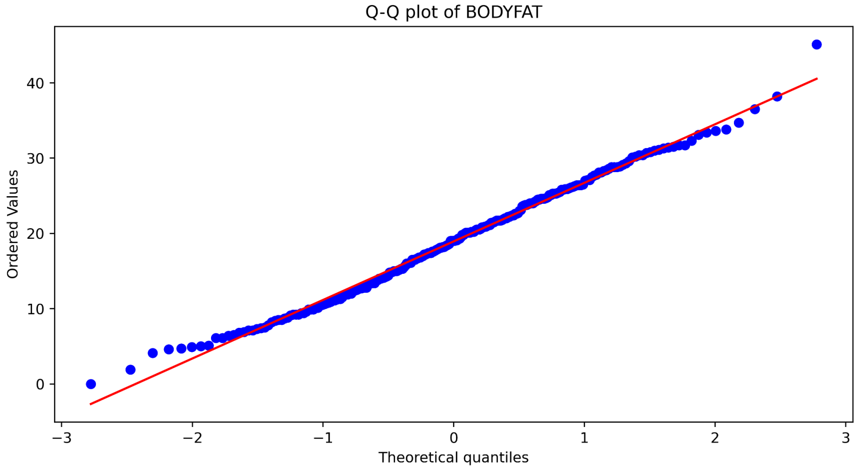
Visualizations:

- **Eight relatively best model combinations**

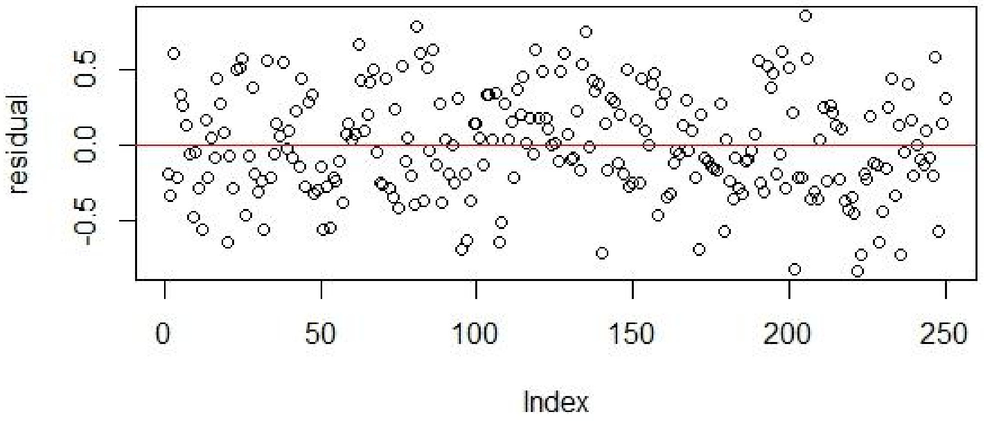


```
         AGE NECK CHEST ABDOMEN HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST BMI
1  ( 1 ) " " " "  " "   "*"     " " " "   " "  " "   " "    " "     " "   " "
2  ( 1 ) " " " "  " "   "*"     " " " "   " "  " "   " "    " "     "*"   " "
3  ( 1 ) "*" " "  " "   "*"     " " " "   " "  " "   " "    " "     "*"   " "
4  ( 1 ) "*" "*"  " "   "*"     " " " "   " "  " "   " "    " "     "*"   " "
5  ( 1 ) "*" " "  "*"   "*"     " " " "   " "  " "   " "    " "     "*"   "*"
6  ( 1 ) "*" " "  "*"   "*"     "*" " "   " "  " "   " "    " "     "*"   "*"
7  ( 1 ) "*" "*"  "*"   "*"     "*" " "   " "  " "   " "    " "     "*"   "*"
8  ( 1 ) "*" "*"  "*"   "*"     "*" " "   " "  " "   " "    "*"     "*"   "*"
```

- **Q-Q Plot of BODYFAT:**



- **Residual performance of best model:**



References:

1. R Documentation for Robust Regression: rlm function - RDocumentation
2. Robust Regression and Outlier Detection with R: (PDF) Outlier Detection for Compositional Data Using Robust Methods (researchgate.net)
3. "Shiny Tutorial" on DataCamp: datacamp/01_R/Reporting/Building Dashboards with shinydashboard/12_case_study.R at master · Georgits/datacamp (github.com)