**Iron Insights - Phase One Report**

Project Overview

Powerlifting is a sport that focuses on three lifts: squat, bench press, and deadlift, each lifter gets three attempts at performing one repetition for each lift. In powerlifting, it can be challenging as a coach or handler to tell your athlete how much weight they should attempt to go for in their latter attempts. We are trying to create a predictive model which takes the lifter’s first attempt for Squat, Bench Press, and Deadlift to predict their total after all three attempts for each lift.

Our data mining project aims to address this by constructing a predictive model which will be built and trained on historical meet data, which includes the weights of each attempt and the final totals for each lifter.

To ensure the model's predictions are precise and useful, we will evaluate its performance through both traditional statistical metrics like mean absolute error and root mean squared error and practical benchmarks, such as the accuracy of its predictions compared to the actual subsequent attempts made by lifters. The goal is to achieve a lower error margin, indicating that the model can reliably guide lifters and coaches in choosing weights that maximize total lift without exceeding the athlete's capability.

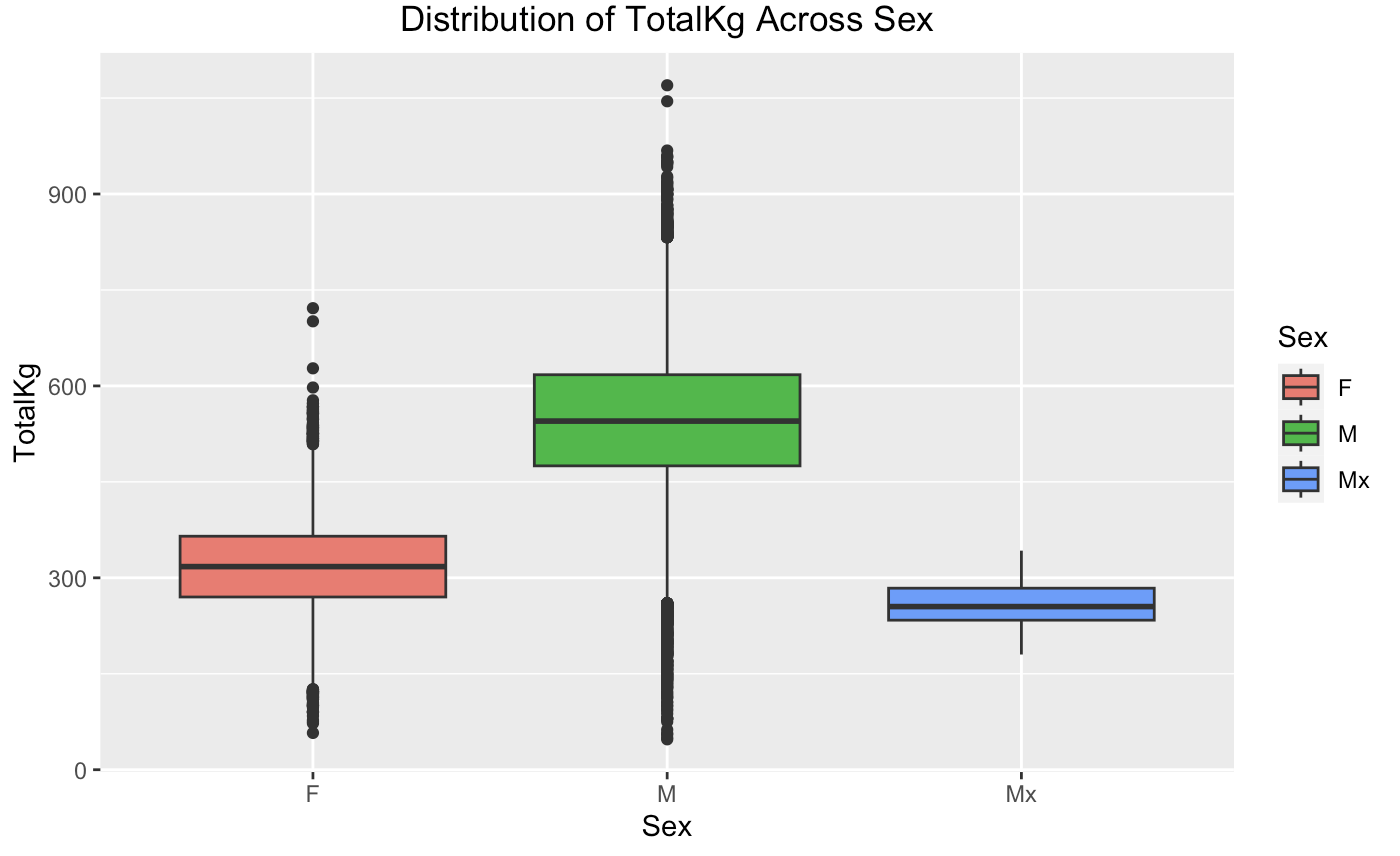
A successful model will be one that can be confidently used by coaches and athletes to inform strategy during competitions, potentially leading to better performance and reduced risk of injury due to overestimation of capabilities.

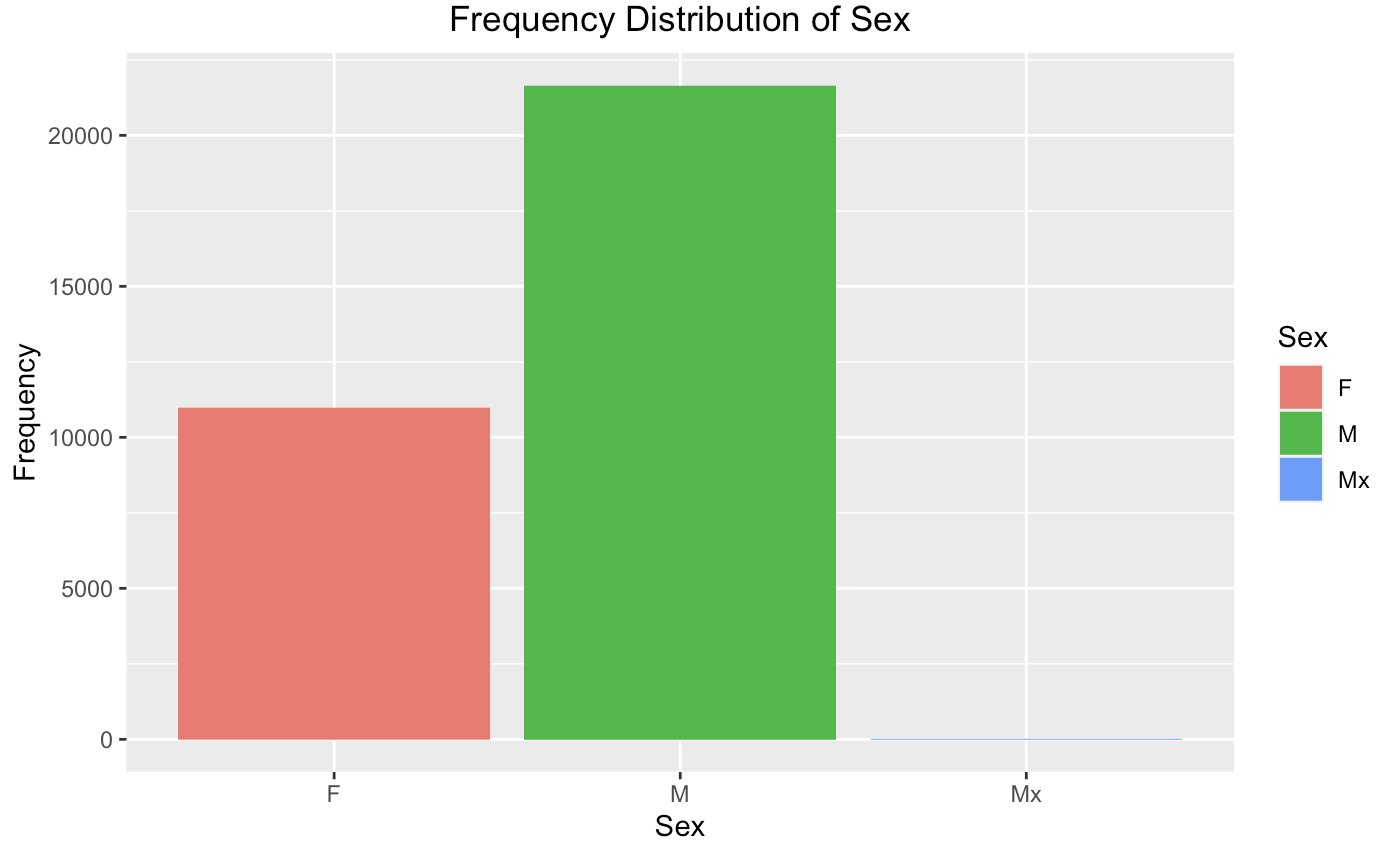
Data Audit

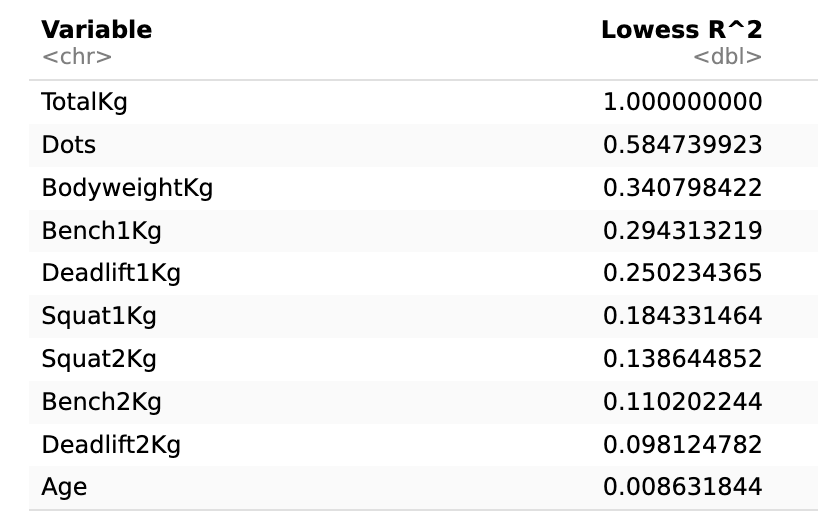
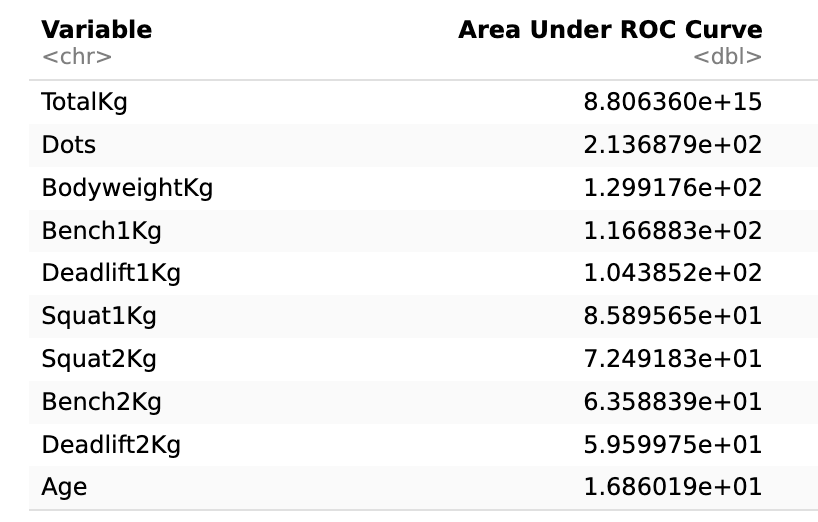
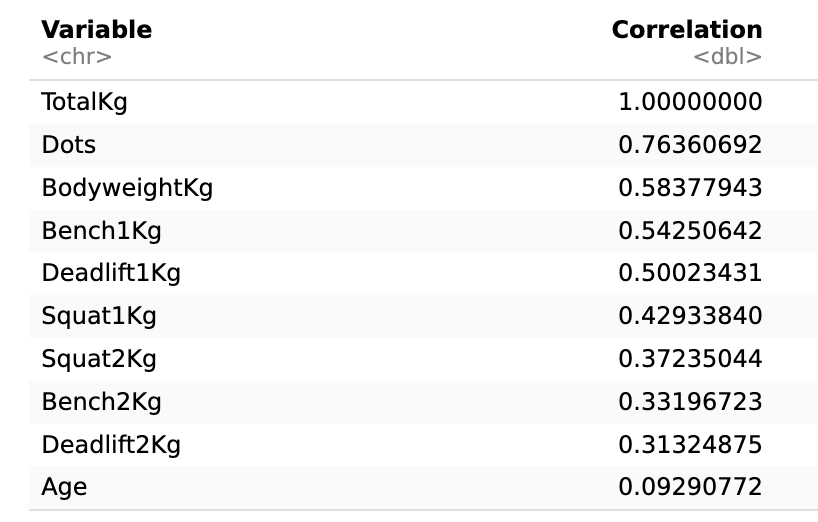
| Name | Model Role | Measurement Level | Description |
| --- | --- | --- | --- |
| Name | ID | Nominal | The name of the lifter |
| TotalKg | Target | Interval | The sum of all best lifts in kg |
| Bench1Kg | Input | Interval | First attempt bench press in kg |
| Bench2Kg | Input | Interval | Second attempt bench press in kg |
| Squat1Kg | Input | Interval | First attempt squat in kg |
| Squat2Kg | Input | Interval | Second attempt squat in kg |
| Deadlift1Kg | Input | Interval | First attempt deadlift in kg |
| Deadlift2Kg | Input | Interval | Second attempt deadlift in kg |
| BodyweightKg | Input | Interval | Bodyweight of lifter in kg |
| Age | Input | Interval | Age of liter |
| Sex | Input | Nominal | Sex of lifer |

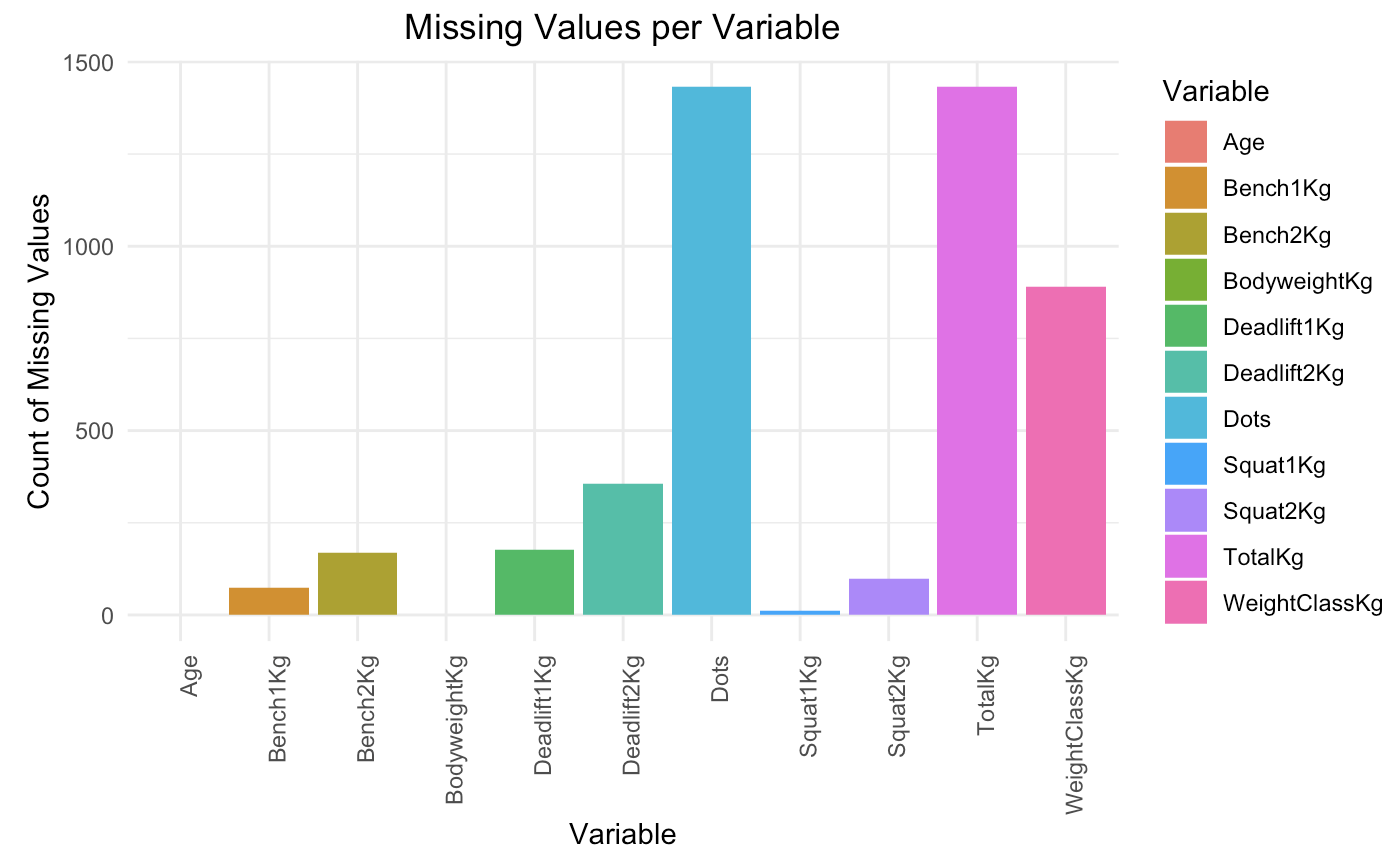
Intuitively, we would predict the above variables to have the greatest correlation with TotalKg.

Descriptive Analysis

To analyze our interval level input variables, we will look at both weighted and unweighted average values for each athlete's first attempts at Squat, Bench Press, and Deadlift, as well as the standard deviation. The mean first attempts for Squat was 124.42 kg, the median was 147.5 kg, with a standard deviation of 61.08 kg. For Bench Press, the mean first attempt was 84.41 kg, the median was 95 kg, with a standard deviation of 108.13 kg. For the Deadlift, the lifter’s first attempt had a mean of 161.78 kg, a median of 175 kg, with a standard deviation of 91.1 kg. It is important to look at all descriptive statistics because in powerlifting, the weight of a lifter immensely impacts the amount of weight the lifter can move. The bodyweight of the lifters had a mean of 82.02 kg, a median of 80.40 kg, and a standard deviation of 19.70 kg. The similarity in averages for bodyweight and low standard deviation relative to the previous statistics is important to note. 



The frequency distribution of the 'Sex' category within the dataset indicates a higher number of male (M) participants compared to female (F) and gender-neutral (Mx) participants, with males dominating the dataset. This is not unexpected in powerlifting datasets but should be considered when analyzing trends and drawing conclusions. The distribution of 'TotalKg' lifted across these groups shows a higher median and a wider interquartile range for males, suggesting greater variability in their performance. For females, the median total lifted is lower, with a narrower interquartile range, which might reflect less variability or a more homogenous group in terms of performance. The 'Mx' category has the smallest sample size, shown by the narrower box plot, which can affect the reliability of statistical inferences for this group. There are some extreme outliers in male and female that would significantly distort the overall analysis. In powerlifting, it's not uncommon for a small number of athletes to achieve totals that are significantly higher than the average due to factors like advanced training, greater experience, superior technique, or genetic advantages. For male, the broader range of physical characteristics and training backgrounds can lead to a wider spread in performance, which might explain the greater number of outliers. 

The R outputs showcase variable importance using three different measures: Area Under ROC Curve, Pearson Correlation, and Lowess \( R^2 \). 'TotalKg' unsurprisingly ranks highest across all measures as it is the target variable. 'Dots' score highly, suggesting strong predictive power, while 'BodyweightKg' and lift attempts like 'Bench1Kg' and 'Deadlift1Kg' also show notable importance. These metrics help prioritize variables for model inputs, indicating 'Dots' and 'BodyweightKg' could be key predictors for the target 'TotalKg'. 'Age' has the least influence, which may suggest it's a less critical predictor in this context. This analysis will guide the variable selection for predictive modeling in the next phase of the project.

The dataset shows a varying number of missing values across different variables. 'TotalKg' has the highest number of missing values, followed by 'WeightClassKg' and 'Squat2Kg'. Since missing 'TotalKg' values represent lifters who did not complete all three lifts, imputing these would not be meaningful—such entries are better excluded to ensure accurate analysis of performance. Similarly, missing 'WeightClassKg' might indicate missing weigh-in data, which is also critical to powerlifting analysis and thus supports the decision not to impute. For other variables like 'Squat1Kg' and 'Deadlift1Kg', the absence of values indicates incomplete attempts, which again, should not be imputed, as they represent actual events in the competition. The plan to drop rows with missing values aligns with the nature of the sport, where completed attempts are crucial for predicting final outcomes.