423 Project

Mohit Mohanraj

2025-03-06

Removing Outliers:

```
fifa <- read.csv("fifa_players.csv", stringsAsFactors = FALSE)

ggplot(data = fifa, aes(x=overall_rating))+
  geom_histogram(binwidth = 1)
```

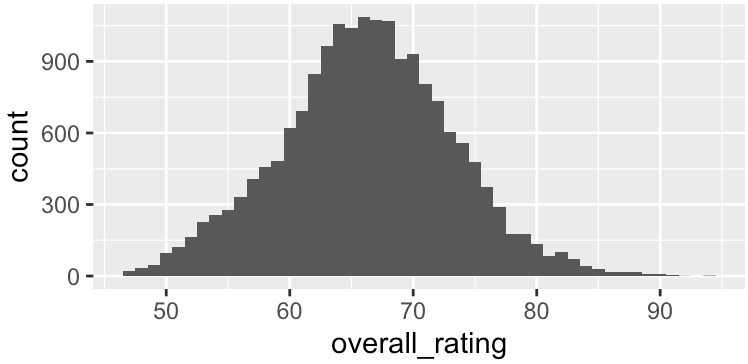


Fig. 1

As we can see, our main response variable, overall_rating, is approximately normally distributed and has no outliers.

```
ggplot(data = fifa, aes(x=height_cm, y=weight_kgs))+
  geom_point()
```

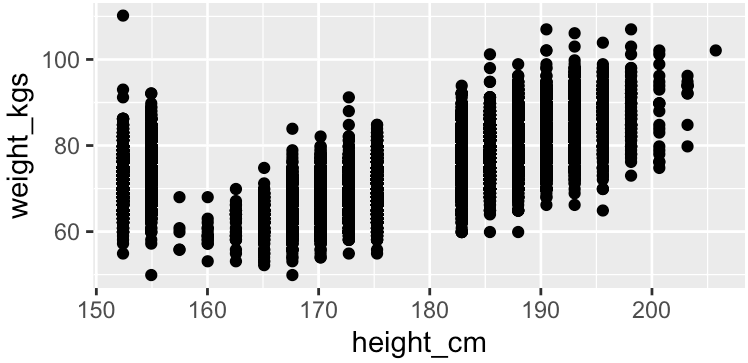


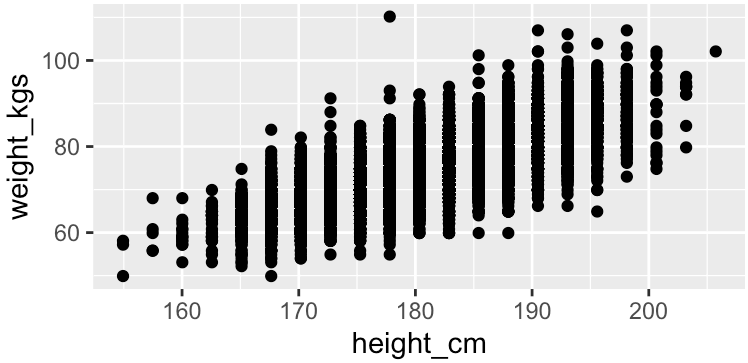
Fig. 2

We know from background information that height and weight should be positively correlated, and we expect to see that in this visualization. We mostly see this, but something is clearly off. We googled the heights to convert centimeters to inches and found that the outliers occur at heights 5'0" and 5'1". We also noticed that the data has an empty patch in the middle of the x axis, indicating that some heights are not included. We hypothesized that there was an error by the creators of the dataset in their data scraping process, in which 5'10" players were recorded as 5'0" and 5'11" was recorded as 5'1". The next code chunk shows us testing this hypothesis and

seeing that 5’10” and 5’11” were indeed the missing heights, so we decided to rewrite the data set so that all 5’0” entries were corrected to 5’10” and all 5’1” entries were corrected to 5’11”. However, the issue with this is that there could be a few players who really are 5’0” and 5’1” in the dataset and they need to be recorded with their accurate height. To solve this problem, we looked through the original data source (https://sofifa.com/players?col=hi&sort=asc&r=190043&set=true (https://sofifa.com/players?col=hi&sort=asc&r=190043&set=true)) and searched for 5’0” and 5’1” players. There were only 3, so we manually filtered by name to get their true height reflected in the dataset. After the following chunk the height column will have accurate data.

```
#wondering if 5'11 was scraped at 5'1 and 5'10 was scraped as 5'0
fifa_test <- fifa %>%
  mutate(height_cm = ifelse(height_cm == 152.4, 177.8, height_cm) ) %>%
  mutate(height_cm = ifelse(height_cm == 154.94, 180.34, height_cm) )  %>%
  mutate(height_cm = ifelse(full_name == "Kazuki Yamaguchi" |name == "H. Nakagawa" | ful
l_name == "Cristian Nahuel Barrios", 154.94, height_cm) )

ggplot(data = fifa_test, aes(x=height_cm, y=weight_kgs))+
  geom_point()
```

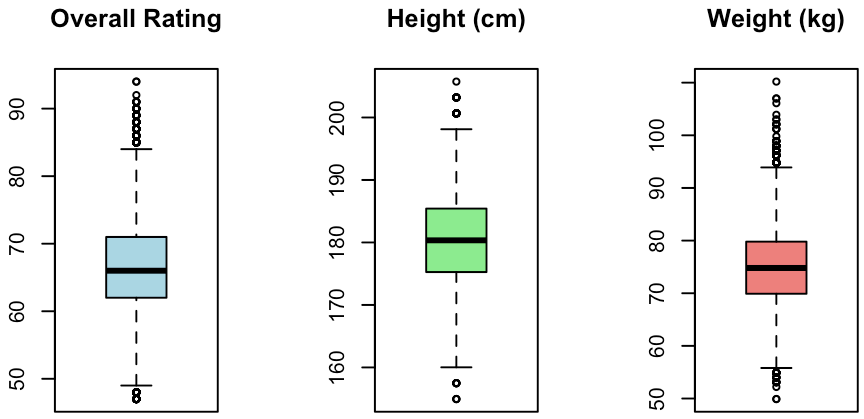



```
fifa <- fifa_test
```

Now our goal is to detect, analyze, and assess whether outliers in our dataset should be removed or mitigated in another way. To do this, we examined overall rating, height, and weight for potential extreme values. We began by generating boxplots for key numerical variables to visualize any extreme values.

Fig. 3

```
par(mfrow = c(1, 3))  # Arrange plots in one row
boxplot(fifa$overall_rating, main = "Overall Rating", col = "lightblue")
boxplot(fifa$height_cm, main = "Height (cm)", col = "lightgreen")
boxplot(fifa$weight_kgs, main = "Weight (kg)", col = "lightcoral")
```



From these boxplots we can see that height seems to be approximately distributed. However, there are outliers at both extremes for both Rating and Weight, with players rated above 85 and below 50 while in weight, some players are below 55kg or above 95kg. To anaylze these outliers further and see if they have an effect, we can identify the most extreme values and look at them:

Table 1 and 2

```
rating_outliers <- fifa[fifa$overall_rating < 50 | fifa$overall_rating > 85, c("name",
"overall_rating")]

head(rating_outliers[order(-rating_outliers$overall_rating), ], 10)  # Highest-rated pla
yers
```

```
##                   name overall_rating
## 1              L. Messi             94
## 17945 Cristiano Ronaldo             94
## 17944         Neymar Jr             92
## 17939         L. Suárez             91
## 17940         L. Modrić             91
## 17941         E. Hazard             91
## 17942      K. De Bruyne             91
## 17943            De Gea             91
## 17931      G. Chiellini             90
## 17932      Sergio Ramos             90
```

```
head(rating_outliers[order(rating_outliers$overall_rating), ], 10)   # Lowest-rated play
ers
```

```
##                      name overall_rating
## 4855           N. Fuentes             47
## 4856            S. Squire             47
## 4857 J. Norville-Williams             47
## 4858           L. Watkins             47
## 4859           A. Kaltner             47
## 4860           L. Collins             47
## 4861            C. Ehlich             47
## 4862         Zhang Yufeng             47
## 4863            Gao Yuqin             47
## 4871            G. Nugent             47
```

The highest-rated outliers include Lionel Messi (94), Cristiano Ronaldo (94), and Neymar Jr. (92). Since these values accurately reflect the skill level of these players, they should not be removed. On the other end, the lowest-rated players, such as N. Fuentes and S. Squire (47 overall rating), are likely reserve or youth players. While they appear as outliers, they are realistic and do not indicate data errors.

Table 3 and 4

```
weight_outliers <- fifa[fifa$weight_kgs < 55 | fifa$weight_kgs > 95, c("name", "weight_k
gs")]
head(weight_outliers[order(-weight_outliers$weight_kgs), ], 10)  # Heaviest players
```

```
##                name weight_kgs
## 11106  A. Akinfenwa      110.2
## 8267   L. Watkowiak      107.0
## 11686      C. Seitz      107.0
## 7073       M. Rhead      106.1
## 13091   F. Farnolle      103.9
## 3186  D. Telgenkamp      103.0
## 16937 L. Unnerstall      103.0
## 2840        T. Holý      102.1
## 7290       E. Louro      102.1
## 9732    E. Johansen      102.1
```

```
head(weight_outliers[order(weight_outliers$weight_kgs), ], 10)   # Lightest players
```

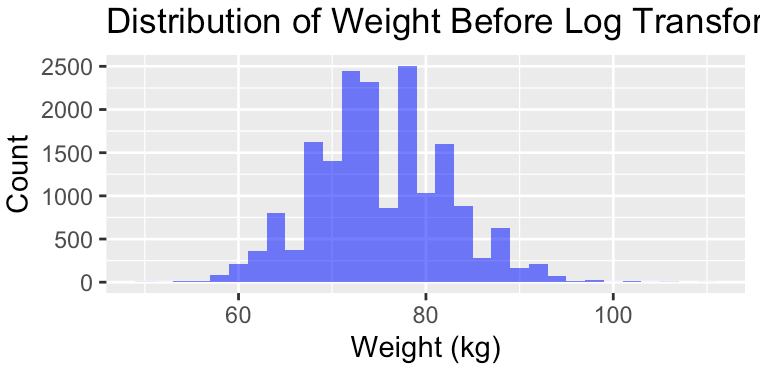
```
##                 name weight_kgs
## 4567    K. Yamaguchi       49.9
## 9061   B. Al Mutairi       49.9
## 14791       D. Rojas       52.2
## 4152     I. Al Talhi       53.1
## 4547       J. García       53.1
## 4661     M. Saavedra       53.1
## 5933    D. Takahashi       53.1
## 6191       Y. Khormi       53.1
## 1734    J. Carrascal       54.0
## 5228  M. Al Shudukhi       54.0
```

The heaviest players include Adebayo Akinfenwa (110.2 kg) and multiple goalkeepers who exceed 100 kg. Given that goalkeepers and defenders tend to be heavier, these values are valid and should not be removed. For the lightest players, we found that Kazuki Yamaguchi and B. Al Mutairi weigh around 49.9 kg, which aligns with expectations for smaller midfielders and forwards. These values are not errors and should also be kept in the dataset.

Fig. 4 and 5

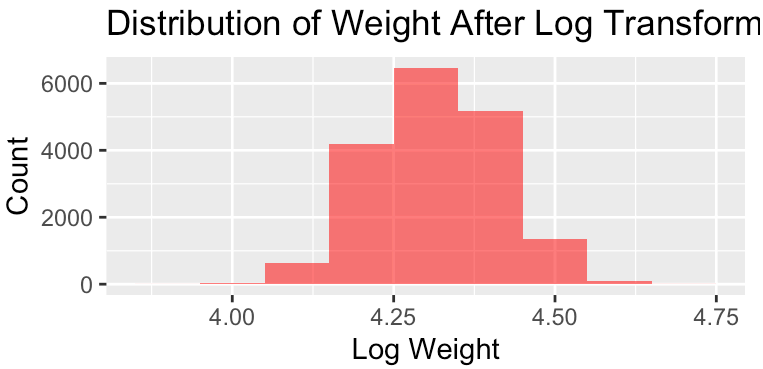
While our analysis confirmed that the outliers in overall rating and weight are valid, extreme values can still influence the regression model. Instead of removing them, we can apply methods to reduce their impact while keeping the dataset intact. The first step was to log-transform weight to make the distribution more normal.

```
ggplot(fifa, aes(x=weight_kgs)) +
  geom_histogram(binwidth = 2, fill = "blue", alpha = 0.6) +
  labs(title = "Distribution of Weight Before Log Transformation",
       x = "Weight (kg)", y = "Count")
```




```
fifa$log_weight <- log(fifa$weight_kgs)

ggplot(fifa, aes(x=log_weight)) +
  geom_histogram(binwidth = 0.1, fill = "red", alpha = 0.6) +
  labs(title = "Distribution of Weight After Log Transformation",
       x = "Log Weight", y = "Count")
```



Before applying the log transformation, the histogram of weight showed a clear right-skewed distribution, where a small number of very heavy players disproportionately influenced the overall weight distribution. After the transformation, the distribution became much more symmetrical, resembling a normal distribution. This adjustment ensures that extreme weight values do not overly influence the regression model while preserving all player observations.

Traditional ordinary least squares (OLS) regression is highly sensitive to extreme values. This means that a few very heavy or light players could disproportionately influence the regression coefficients. To prevent this, we implemented robust regression, which assigns less weight to extreme values while still considering them.

Figure 6

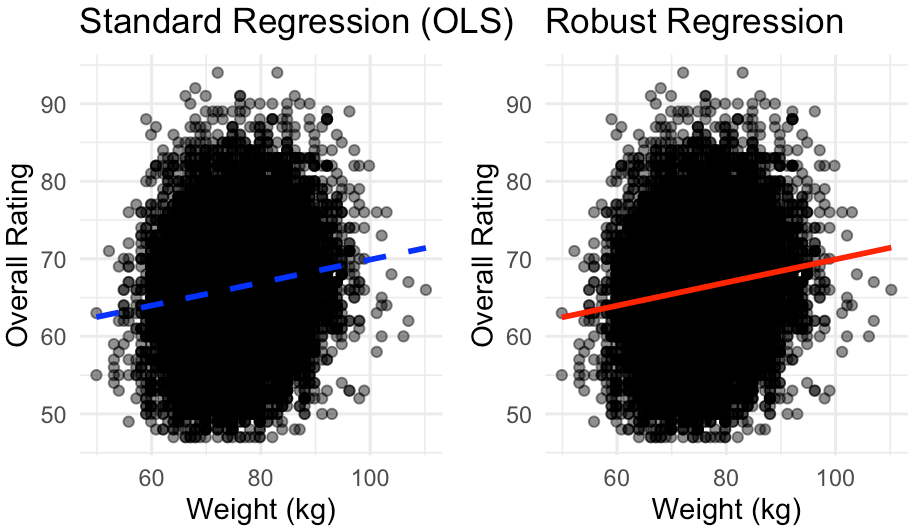
```
standard_model <- lm(overall_rating ~ weight_kgs, data = fifa)
robust_model <- rlm(overall_rating ~ weight_kgs, data = fifa)

library(patchwork)

plot_lm <- ggplot(fifa_test, aes(x = weight_kgs, y = overall_rating)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue", se = FALSE, linetype = "dashed") +
  labs(title = "Standard Regression (OLS)",
       x = "Weight (kg)", y = "Overall Rating") +
  theme_minimal()

plot_rlm <- ggplot(fifa_test, aes(x = weight_kgs, y = overall_rating)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "rlm", color = "red", se = FALSE) +
  labs(title = "Robust Regression",
       x = "Weight (kg)", y = "Overall Rating") +
  theme_minimal()

plot_lm + plot_rlm
```



The results of this comparison show a clear difference between standard regression and robust regression. In the scatter plot, the blue dashed line represents standard OLS regression, which is strongly influenced by extreme values. The red solid line represents robust regression, which follows the general trend of the data but is less affected by extreme values. By implementing robust regression, we ensure that our model is less sensitive to extreme weight values, making it more stable and interpretable.

To further validate whether keeping outliers affects model accuracy, we compared models with and without extreme weight values by examining the adjusted R^2 values and AIC information.

```
full_model <- lm(overall_rating ~ sprint_speed + dribbling + strength + stamina + log(we
ight_kgs), data = fifa)
clean_fifa <- fifa[fifa$weight_kgs >= 55 & fifa$weight_kgs <= 95, ]
clean_model <- lm(overall_rating ~ sprint_speed + dribbling + strength + stamina + log(w
eight_kgs), data = clean_fifa)

summary(full_model)$adj.r.squared
```

```
## [1] 0.2967695
```

```
summary(clean_model)$adj.r.squared
```

```
## [1] 0.2973633
```

```
AIC(full_model)
```

```
## [1] 114325.4
```

```
AIC(clean_model)
```

```
## [1] 113744.3
```

The Adjusted R^2 values for the full model (0.2968) and the clean model (0.2973) are nearly identical, suggesting that removing extreme weight values does not improve the model's explanatory power. The AIC values show a slight decrease when outliers are removed, with the clean model at 113,744.3 compared to 114,325.4 for the full model. While a lower AIC generally indicates a better model, the difference of about 580 points is relatively small considering the dataset size. This suggests that the improvement from removing outliers is minor and does not justify eliminating valid data points.

Conclusion:

Since removing outliers does not significantly improve predictive accuracy, we retain all players while applying transformations and robust modeling to ensure a more stable regression analysis. These adjustments allow us to preserve real-world player data while preventing extreme values from skewing the results, ultimately leading to a more reliable model for predicting FIFA player ratings.