HW1: Data Preparation of Moneyball Training Data

{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE) # Homework 1: Data Preparation of Moneyball Training Data ## 1. Load and Inspect Data

```
# Load libraries
library(readxl)
library(dplyr)
library(ggplot2)
library(reshape2)
library(corrplot)
library(knitr)
library(naniar)
library(car)

# Load the dataset
moneyball_training_data <- read_excel("/Users/leslietavarez/Downloads/moneyball-training-data.xlsx")

# Basic info
dim(moneyball_training_data)      # number of rows and columns
colnames(moneyball_training_data) # variable names
summary(moneyball_training_data)  # quick summary stats
```

Summary Statistics and Boxplots

```
# Summary stats (mean, median, sd)
summary_stats <- data.frame(
  Variable = colnames(moneyball_training_data),
  Mean = sapply(moneyball_training_data, mean, na.rm = TRUE),
  Median = sapply(moneyball_training_data, median, na.rm = TRUE),
  SD = sapply(moneyball_training_data, sd, na.rm = TRUE)
)
kable(summary_stats, caption = "Summary Statistics (Mean, Median, SD)")

# Boxplots for all variables
melted_num <- melt(moneyball_training_data)
ggplot(melted_num, aes(x = variable, y = value)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Box Plots of Variables", x = "Variables", y = "Values")
```

3. Target Variable Distribution

```
# Histogram of target variable (TARGET_WINS)
```

```
ggplot(moneyball_training_data, aes(x = TARGET_WINS)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Distribution of TARGET_WINS", x = "Wins", y = "Count")
```

4. Correlation Analysis

```
# Correlation matrix
# Correlation matrix
cor_matrix <- cor(moneyball_training_data, use = "pairwise.complete.obs")

# Top 5 positive and negative correlations with TARGET_WINS
target_corr <- sort(cor_matrix[,"TARGET_WINS"], decreasing = TRUE)
top_pos <- head(target_corr, 6)   # includes self-correlation
top_neg <- tail(target_corr, 5)
kable(round(c(top_pos, top_neg), 3), caption = "Top Correlations with TARGET_WINS")

# Heatmap
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.6)
```

5. Missing Data Analysis

```
# Missing data summary
missing_summary <- sapply(moneyball_training_data, function(x) sum(is.na(x)))
missing_summary <- data.frame(Variable = names(missing_summary), MissingCount = missing_summary)
kable(missing_summary, caption = "Missing Data Summary")
# Visualize missing data
vis_miss(moneyball_training_data) + labs(title = "Missing Data Visualization")
```

7. Multicollinearity Check

```
# VIF calculation
vif_values <- vif(lm(TARGET_WINS ~ ., data = moneyball_training_data))
vif_df <- data.frame(Variable = names(vif_values), VIF = vif_values)
kable(vif_df, caption = "Variance Inflation Factor (VIF) for Each Variable")

# Identify variables with high VIF
high_vif <- vif_values[vif_values > 5]
if(length(high_vif) > 0) {
  kable(as.data.frame(high_vif), caption = "Variables with High VIF (>5)")
} else {
  print("No variables with VIF > 5")
}
```

8. Scatterplots with Target Variable

Scatterplots help us explore how individual predictors relate to `TARGET_WINS`.
We focus on three variables that represent different aspects of the game:
- **Home Runs (TEAM_BATTING_HR / log_HR):** Power hitting.

- **Walks (TEAM_BATTING_BB):** Plate discipline and baserunners.
- **Errors (TEAM_FIELDING_E):** Defensive mistakes.

These were chosen because they are interpretable, relatively complete (not heavily missing), and together balance offensive and defensive perspectives. Variables with extreme missingness (e.g., Hit by Pitch) were excluded.

```
# Home Runs vs Wins
ggplot(moneyball_training_data, aes(x = TEAM_BATTING_HR, y = TARGET_WINS)) +
  geom_point(alpha = 0.5, color = "darkred") +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Home Runs vs Wins", x = "Home Runs", y = "Wins")

# Walks vs Wins
ggplot(moneyball_training_data, aes(x = TEAM_BATTING_BB, y = TARGET_WINS)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Walks vs Wins", x = "Walks", y = "Wins")

# Errors vs Wins
ggplot(moneyball_training_data, aes(x = TEAM_FIELDING_E, y = TARGET_WINS)) +
  geom_point(alpha = 0.5, color = "darkgreen") +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Errors vs Wins", x = "Errors", y = "Wins")
```