

```

library(tidyverse)
library(ggplot2)
library(janitor)

# Import

mlb_batters_original <- read.csv("mlb_2024_batters.csv")
str(mlb_batters_original)
summary(mlb_batters_original)

# Data Cleaning
mlb_batters_all <- mlb_batters_original
mlb_batters_all <- clean_names(mlb_batters_all)

mlb_batters_all <- mlb_batters_all %>%
  select(-x, -x_1, -x_2, -player_additional)

mlb_batters_all <- mlb_batters_all %>%
  mutate(player = str_remove_all(player, "[*#]"))

mlb_batters_all <- mlb_batters_all %>%
  filter(!(team %in% c("2TM", "3TM")))

mlb_batters_all <- mlb_batters_all %>%
  filter(!is.na(ba))

colnames(mlb_batters_all)[colnames(mlb_batters_all) == "x2b"] <- "2b"
colnames(mlb_batters_all)[colnames(mlb_batters_all) == "x3b"] <- "3b"

# Data Structure
mlb_batters_all <- mlb_batters_all %>%
  mutate(pos2 = if_else(str_detect(pos, ","),
    str_trim(str_extract(pos, "(?<=,).*")),
    NA_character_))

mlb_batters_all <- mlb_batters_all %>%
  mutate(pos = str_trim(str_extract(pos, "^[^,]+")))

mlb_batters_all <- mlb_batters_all %>%
  mutate(pos_type = case_when(
    pos %in% c("1", "2", "3", "4", "5", "6") ~ "infielder",
    pos %in% c("7", "8", "9") ~ "outfielder",
    pos == "DH" ~ "DH",
    TRUE ~ NA_character_
  ))

```

```

))

mlb_batters_all <- mlb_batters_all %>%
mutate(extra_bases = as.numeric(`2b`) + as.numeric(`3b`))

# Frequency Tables

ba_range_ft <- mlb_batters_all %>%
mutate(ba_range = case_when(
  ba < 0.200 ~ "< .200",
  ba >= 0.200 & ba < 0.250 ~ ".200 - .250",
  ba >= 0.250 & ba < 0.300 ~ ".250 - .300",
  ba >= 0.300 & ba < 0.350 ~ ".300 - .350",
  ba >= 0.350 & ba < 0.400 ~ ".350 - .400",
  ba >= 0.400 ~ ">= .400"
)) %>%
count(ba_range) %>%
mutate(Percent = round(n / sum(n) * 100, 1)) %>%
rename('# of Players' = n)

team_batting_avg_ft <- mlb_batters_all %>%
group_by(team) %>%
summarise(
  avg_ba = round(mean(ba, na.rm = TRUE), 3),
  avg_hr = round(mean(hr, na.rm = TRUE), 3)
)

# Cross-tabulations

hr_vs_pos <- mlb_batters_all %>%
group_by(pos_type) %>%
summarise(avg_hr = round(mean(hr, na.rm = TRUE), 2))

### Distribution of Home Runs - Frequency

ggplot(mlb_batters_all, aes(x = hr)) +
geom_histogram(fill = "red", color = "black", bins = 20) +
ggtitle("Distribution of Home Runs") +
xlab("Home Runs") +
ylab("Frequency") +
theme_minimal() +
theme(

```

```

panel.grid = element_blank(),
panel.background = element_rect(fill = "white", color = NA)
)
num_players_more_than_30_hr <- sum(mlb_batters_all$hr > 30, na.rm = TRUE)
num_players_more_than_30_hr

### Calculate BA by Position for plots below

pos_ba <- mlb_batters_all %>%
group_by(pos) %>%
summarise(avg_ba = mean(ba, na.rm = TRUE))
pos_ba

### Batting Average by Position - Bar

ggplot(pos_ba, aes(x = pos, y = avg_ba)) +
geom_bar(stat = "identity", fill = "green", color = "black") +
ggtitle("Batting Averages by Position") +
xlab("Position") +
ylab("Batting Average") +
theme_minimal() +
theme(
panel.grid = element_blank(),
panel.background = element_rect(fill = "white", color = NA)
)

### Position vs Batting Avg - Boxplot

ggplot(mlb_batters_all, aes(x = pos, y = ba)) +
geom_boxplot(fill = "lightblue", color = "darkblue") +
labs(title = "Boxplot of Batting Averages by Position",
x = "Position",
y = "Batting Average") +
theme(
panel.grid = element_blank(),
panel.background = element_rect(fill = "white", color = NA)
)

### Batting Average vs Home Runs - Scatter

ggplot(mlb_batters_all, aes(x = ba, y = hr)) +
geom_point(color = "darkgreen", alpha = 0.6) +
geom_smooth(method = lm, se = TRUE, color = "red") +

```

```

labs(title = "Batting Average vs. Home Runs",
x = "Batting Average",
y = "Home Runs") +
theme(
panel.grid = element_blank(),
panel.background = element_rect(fill = "white", color = NA)
)

```

At Bats vs Batting Average - Scatter

```

ggplot(mlb_batters_all, aes(x = ab, y = ba)) +
geom_point(color = "steelblue", alpha = 0.6) +
geom_smooth(method = lm, se = TRUE, color = "green") +
labs(title = "At-Bats vs. Batting Average",
x = "At-Bats",
y = "Batting Average") +
theme(
panel.grid = element_blank(),
panel.background = element_rect(fill = "white", color = NA)
)

```

Extra Bases vs RBI - Scatter

```

ggplot(mlb_batters_all, aes(x = extra_bases, y = rbi)) +
geom_point(color = "darkorange", alpha = 0.7) +
geom_smooth(method = "lm", se = TRUE, color = "green") +
labs(title = "RBI vs. Extra-Base Hits",
x = "Extra-Base Hits (2B + 3B)",
y = "Runs Batted In (RBI)") +
theme(
panel.grid = element_blank(),
panel.background = element_rect(fill = "white", color = NA)
)

```