NBA Player Performance

Tim Chen, Ying Jiang, Mohammed Alshamsi

2025-04-28

Table of contents

1 Introduction

We're exploring the relationship between physical characteristics (height, weight) and draft outcomes in the NBA.
Research questions:
- What's the distribution of height and weight among drafted players? - How do physical attributes relate to draft pick or round? - Any general patterns in the dataset?

2 Data Source

Data come from Wyatt O'Walsh's Kaggle repo (https://www.kaggle.com/datasets/wyattowalsh/basketball/data originally collected by the NBA. Cases = individual players; variables = physical stats and draft history.

2.1 FAIR

- Findable: Yes, indexed on Kaggle with metadata;
- Accessible: Direct download (requires Kaggle account);
- Interoperable: CSV format;
- Reusable: CC BY 4.0 license.

2.2 CARE

- Collective Benefit: Publicly shared for analytics;
- Authority to Control: Dataset creator controls upload; no community governance;
- Responsibility: Ethical sourcing, but provenance unclear;
- Ethics: No apparent privacy issues (public sports data).

3 Setup & Data Cleaning

```
# Load necessary library
library(dplyr)
library(janitor)
library(ggplot2)
library(tidyr)
library(readr)
library(stringr)
```

```r
# Read Data
player_info <- read_csv("https://raw.githubusercontent.com/jiangyeee0/STAT-184-/main/common_
draft_history <- read_csv("https://raw.githubusercontent.com/jiangyeee0/STAT-184-/main/draft_

# Clean Player Info
player_clean <- player_info %>%
  mutate(
    feet = as.numeric(str_extract(height, "^[0-9]+")),
    inches = as.numeric(str_extract(height, "(?<=-)[0-9]+")),
    height_in = feet * 12 + replace_na(inches, 0),
    weight = as.numeric(str_extract(weight, "[0-9]+")),
    bmi = (703 * weight) / (height_in^2),
    across(c(height_in, weight), ~replace_na(., median(., na.rm = TRUE))))

# Clean Draft History
draft_clean <- draft_history %>%
  mutate(across(c(overall_pick, round_number, round_pick), as.numeric))

# Merge two databases
nba_data <- inner_join(player_clean, draft_clean, by = "person_id")
```

4 Exploratory Data Analysis

4.1 Glimpse of Data

```r
glimpse(nba_data)
```

```
Rows: 2,985
Columns: 50
$ person_id                  <dbl> 76001, 76003, 1505, 949, 76005, 76006~
$ first_name                 <chr> "Alaa", "Kareem", "Tariq", "Shareef",~
$ last_name                  <chr> "Abdelnaby", "Abdul-Jabbar", "Abdul-W~
$ display_first_last         <chr> "Alaa Abdelnaby", "Kareem Abdul-Jabba~
$ display_last_comma_first   <chr> "Abdelnaby, Alaa", "Abdul-Jabbar, Kar~
$ display_fi_last            <chr> "A. Abdelnaby", "K. Abdul-Jabbar", "T~
$ player_slug                <chr> "alaa-abdelnaby", "kareem-abdul-jabba~
$ birthdate                  <dttm> 1968-06-24, 1947-04-16, 1974-11-03, ~
$ school                     <chr> "Duke", "UCLA", "San Jose State", "Ca~
```

```
$ country                          <chr> "USA", "USA", "France", "USA", "USA",~
$ last_affiliation                 <chr> "Duke/USA", "UCLA/USA", "San Jose Sta~
$ height                           <chr> "6-10", "7-2", "6-6", "6-9", "6-7", "~
$ weight                           <dbl> 240, 225, 235, 245, 220, 180, 200, 22~
$ season_exp                       <dbl> 5, 20, 7, 13, 5, 1, 3, 3, 3, 1, 6, 7,~
$ jersey                           <chr> "30", "33", "9", "3", "5", "6", NA, "~
$ position                         <chr> "Forward", "Center", "Forward-Guard",~
$ rosterstatus                     <chr> "Inactive", "Inactive", "Inactive", "~
$ games_played_current_season_flag <chr> "N", "N", "N", "N", "N", "N", "N", "N~
$ team_id.x                        <dbl> 1610612757, 1610612747, 1610612758, 1~
$ team_name.x                      <chr> "Trail Blazers", "Lakers", "Kings", "~
$ team_abbreviation.x              <chr> "POR", "LAL", "SAC", "VAN", "GOS", "P~
$ team_code                        <chr> "blazers", "lakers", "kings", "grizzl~
$ team_city.x                      <chr> "Portland", "Los Angeles", "Sacrament~
$ playercode                       <chr> "HISTADD_alaa_abdelnaby", "HISTADD_ka~
$ from_year                        <dbl> 1990, 1969, 1997, 1996, 1976, 1956, 2~
$ to_year                          <dbl> 1994, 1988, 2003, 2007, 1980, 1956, 2~
$ dleague_flag                     <chr> "N", "N", "N", "N", "N", "N", "N", "N~
$ nba_flag                         <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y~
$ games_played_flag                <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y~
$ draft_year                       <chr> "1990", "1969", "1997", "1996", "1976~
$ draft_round                      <chr> "1", "1", "1", "1", "3", NA, "2", "1"~
$ draft_number                     <chr> "25", "1", "11", "3", "43", NA, "32",~
$ greatest_75_flag                 <chr> "N", "Y", "N", "N", "N", "N", "N", "N~
$ feet                             <dbl> 6, 7, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6~
$ inches                           <dbl> 10, 2, 6, 9, 7, 3, 6, 8, 5, 0, 11, 7,~
$ height_in                        <dbl> 82, 86, 78, 81, 79, 75, 78, 80, 77, 7~
$ bmi                              <dbl> 25.09221, 21.38656, 27.15401, 26.2513~
$ player_name                      <chr> "Alaa Abdelnaby", "Kareem Abdul-Jabba~
$ season                           <dbl> 1990, 1969, 1997, 1996, 1976, 1956, 2~
$ round_number                     <dbl> 1, 1, 1, 1, 3, 0, 2, 1, 2, 2, 2, 2, 1~
$ round_pick                       <dbl> 25, 1, 11, 3, 9, 0, 2, 20, 30, 0, 16,~
$ overall_pick                     <dbl> 25, 1, 11, 3, 43, 0, 32, 20, 60, 0, 4~
$ draft_type                       <chr> "Draft", "Draft", "Draft", "Draft", "~
$ team_id.y                        <dbl> 1610612757, 1610612749, 1610612758, 1~
$ team_city.y                      <chr> "Portland", "Milwaukee", "Sacramento"~
$ team_name.y                      <chr> "Trail Blazers", "Bucks", "Kings", "G~
$ team_abbreviation.y              <chr> "POR", "MIL", "SAC", "VAN", "LAL", "S~
$ organization                     <chr> "Duke", "California-Los Angeles", "Sa~
$ organization_type                <chr> "College/University", "College/Univer~
$ player_profile_flag              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

4.2 Summary of Numeric Variables

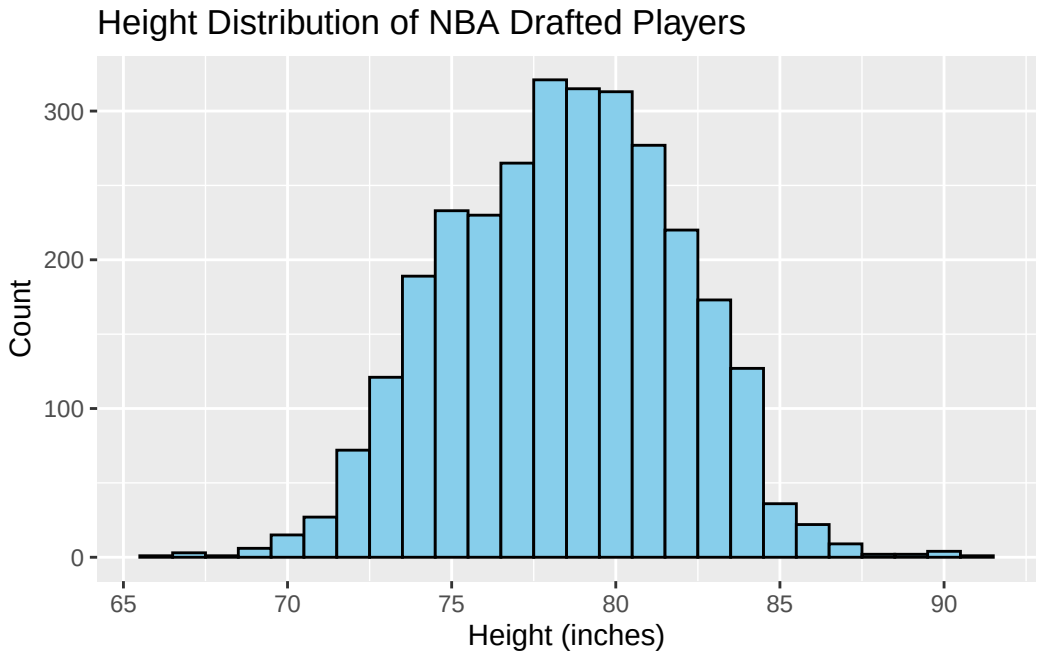
```
# Summarizing Player Stats
num_summary <- nba_data %>%
  select(bmi, height_in, weight, season_exp, round_number,
         round_pick, draft_type, player_profile_flag, overall_pick) %>%
  select(where(is.numeric)) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "value") %>%
  group_by(variable) %>%
  summarise(
    mean = mean(value, na.rm = TRUE),
    median = median(value, na.rm = TRUE),
    sd = sd(value, na.rm = TRUE),
    min = min(value, na.rm = TRUE),
    max = max(value, na.rm = TRUE),
    n_missing = sum(is.na(value)),
    .groups = 'drop'
  )

print(num_summary)
```

```
# A tibble: 8 x 7
  variable              mean median      sd   min   max n_missing
  <chr>                <dbl>  <dbl>   <dbl> <dbl> <dbl>     <int>
1 bmi                  24.2    24.1  1.77    17.4  33.8         50
2 height_in            78.4    79    3.49    66    91            0
3 overall_pick         30.6    24   29.8      0   221            0
4 player_profile_flag   1.00    1    0.0183   0     1            0
5 round_number          2.06    2    1.92     0    20            0
6 round_pick           11.0     9    8.09     0    30            0
7 season_exp            5.98    4    4.70     0    22            0
8 weight              212.    210   26.4    133   325            0
```

5 Graphs

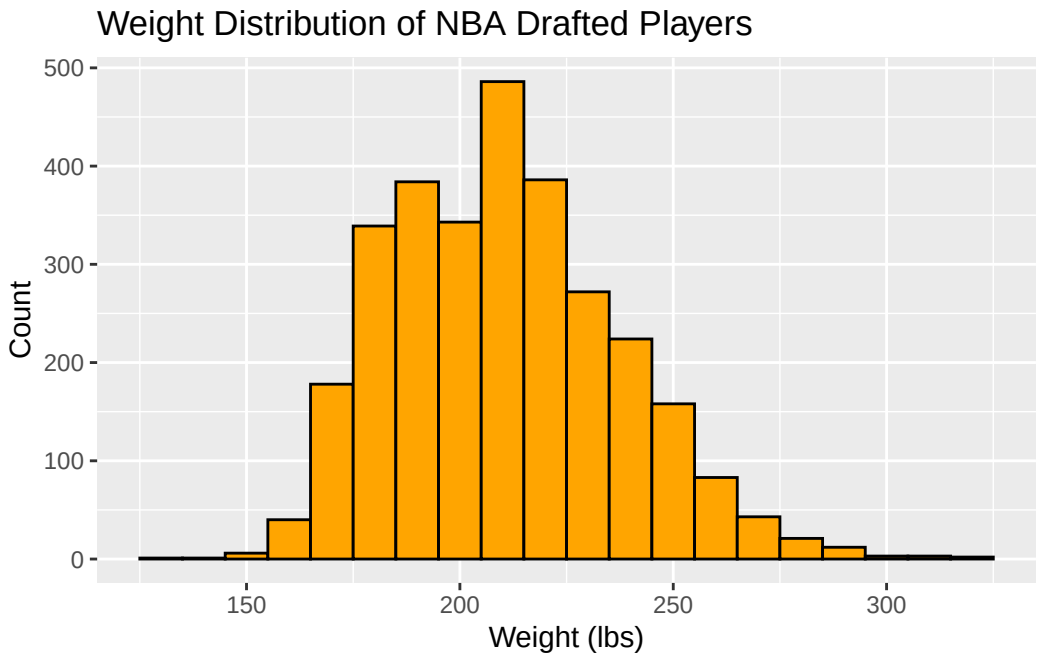
5.1 Height Distribution

```r
ggplot(nba_data, aes(x = height_in)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(
    title = "Height Distribution of NBA Drafted Players",
    x = "Height (inches)",
    y = "Count"
  )
```



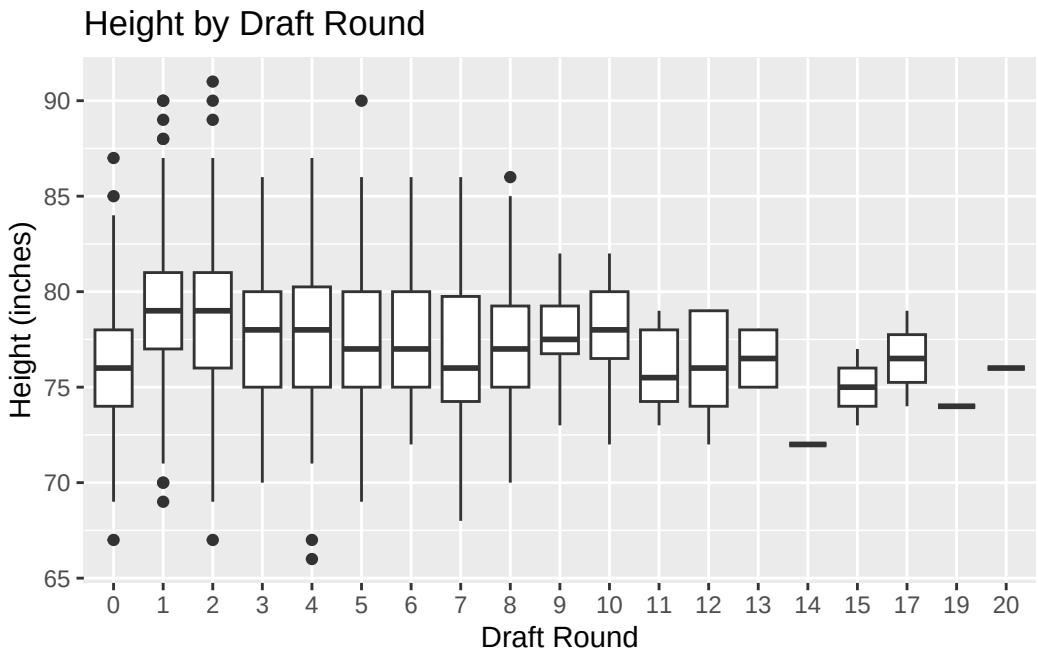
5.2 Weight Distribution

```r
ggplot(nba_data, aes(x = weight)) +
  geom_histogram(binwidth = 10, fill = "orange", color = "black") +
  labs(
    title = "Weight Distribution of NBA Drafted Players",
    x = "Weight (lbs)",
    y = "Count"
  )
```



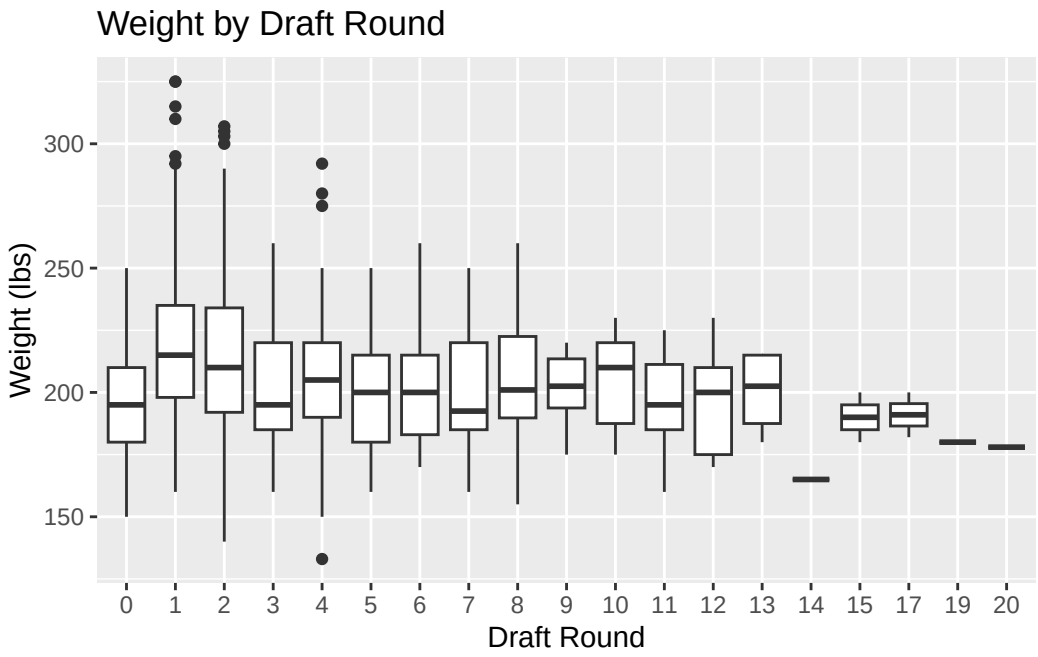
5.3 Draft Round vs Height

```
nba_data %>%
  ggplot(aes(x = factor(round_number), y = height_in)) +
  geom_boxplot() +
  labs(
    title = "Height by Draft Round",
    x = "Draft Round",
    y = "Height (inches)"
  )
```



5.4 Draft Round vs Weight

```
nba_data %>%
  ggplot(aes(x = factor(round_number), y = weight)) +
  geom_boxplot() +
  labs(
    title = "Weight by Draft Round",
    x = "Draft Round",
    y = "Weight (lbs)"
  )
```



5.5 Height and Weight by Draft Round

```
round_summary <- nba_data %>%
  group_by(round_number) %>%
  summarise(
    Avg_Height = mean(height_in, na.rm = TRUE),
    Median_Height = median(height_in, na.rm = TRUE),
    Avg_Weight = mean(weight, na.rm = TRUE),
    Median_Weight = median(weight, na.rm = TRUE),
    n_players = n(),
    .groups = 'drop'
  ) %>%
  mutate(round_number = paste("Round", round_number))

knitr::kable(round_summary, caption = "Height and weight by draft round")
```

Table 1: Height and weight by draft round

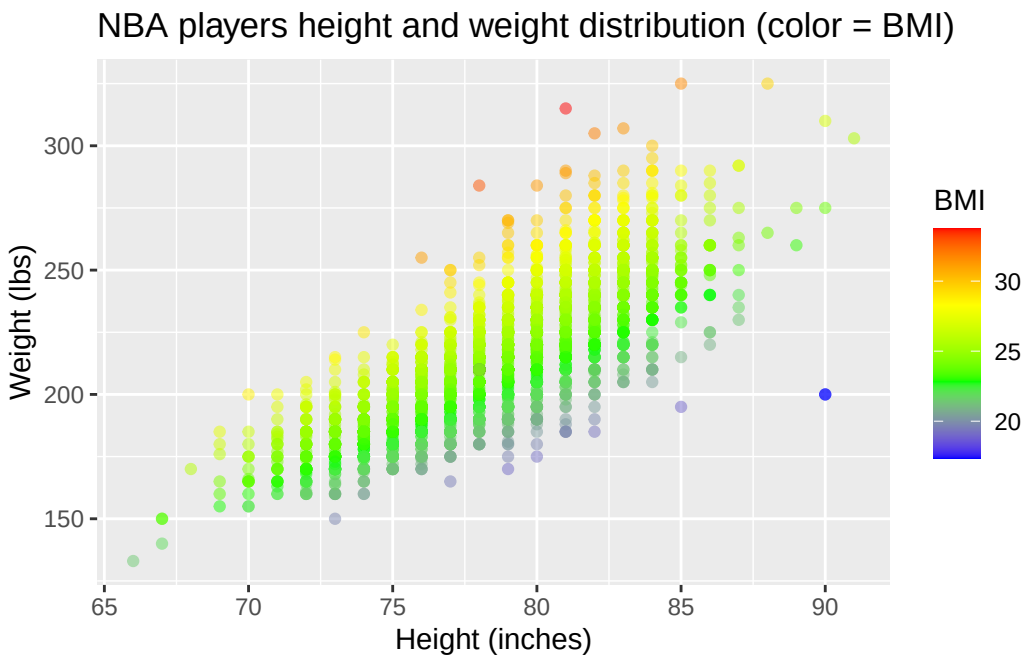
round_number	Avg_Height	Median_Height	Avg_Weight	Median_Weight	n_players
Round 0	75.89167	76.0	195.0750	195.0	120

round_number	Avg_Height	Median_Height	Avg_Weight	Median_Weight	n_players
Round 1	78.96402	79.0	217.2886	215.0	1334
Round 2	78.50734	79.0	214.2222	210.0	954
Round 3	77.71698	78.0	201.8632	195.0	212
Round 4	77.93966	78.0	205.1034	205.0	116
Round 5	77.31507	77.0	199.3425	200.0	73
Round 6	77.67347	77.0	201.1224	200.0	49
Round 7	76.84211	76.0	200.0789	192.5	38
Round 8	77.47222	77.0	205.6667	201.0	36
Round 9	77.75000	77.5	202.8125	202.5	16
Round 10	77.63636	78.0	205.4545	210.0	11
Round 11	76.00000	75.5	196.0000	195.0	10
Round 12	76.00000	76.0	197.0000	200.0	5
Round 13	76.50000	76.5	200.0000	202.5	4
Round 14	72.00000	72.0	165.0000	165.0	1
Round 15	75.00000	75.0	190.0000	190.0	2
Round 17	76.50000	76.5	191.0000	191.0	2
Round 19	74.00000	74.0	180.0000	180.0	1
Round 20	76.00000	76.0	178.0000	178.0	1

Regarding "Round 0" in NBA graphs/tables: The dataset used 0 to represent undrafted players.

5.6 Height and Weight Distribution (color = BMI)

```
ggplot(nba_data, aes(x = height_in, y = weight)) +
  geom_point(aes(color = bmi), alpha = 0.5) +
  scale_color_gradientn(
    name = "BMI",
    colors = c("blue", "green", "yellow", "red"),
    breaks = c(20, 25, 30)
  ) +
  labs(
    title = "NBA players height and weight distribution (color = BMI)",
    x = "Height (inches)",
    y = "Weight (lbs)"
  )
```



6 Narrative Summary

Most NBA draft picks fall within the standard height and weight range. Generally, earlier rounds feature slightly taller and lighter players. Most players have heights around the mid-to-high 70 inches and weights between 180-240 lbs.

7 Conclusion

Draft outcomes show slight tendencies towards specific physical profiles, though clear gaps remain between players selected in different rounds.

8 Code Appendix

```
# Load necessary library
library(dplyr)
library(janitor)
library(ggplot2)
library(tidyr)
```

```r
library(readr)
library(stringr)

# Read Data
player_info <- read_csv("https://raw.githubusercontent.com/jiangyeee0/STAT-184-/main/common_
draft_history <- read_csv("https://raw.githubusercontent.com/jiangyeee0/STAT-184-/main/draft_

# Clean Player Info
player_clean <- player_info %>%
  mutate(
    feet = as.numeric(str_extract(height, "^[0-9]+")),
    inches = as.numeric(str_extract(height, "(?<=-)[0-9]+")),
    height_in = feet * 12 + replace_na(inches, 0),
    weight = as.numeric(str_extract(weight, "[0-9]+")),
    bmi = (703 * weight) / (height_in^2),
    across(c(height_in, weight), ~replace_na(., median(., na.rm = TRUE))))

# Clean Draft History
draft_clean <- draft_history %>%
  mutate(across(c(overall_pick, round_number, round_pick), as.numeric))

# Merge two databases
nba_data <- inner_join(player_clean, draft_clean, by = "person_id")
glimpse(nba_data)
# Summarizing Player Stats
num_summary <- nba_data %>%
  select(bmi, height_in, weight, season_exp, round_number,
         round_pick, draft_type, player_profile_flag, overall_pick) %>%
  select(where(is.numeric)) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "value") %>%
  group_by(variable) %>%
  summarise(
    mean = mean(value, na.rm = TRUE),
    median = median(value, na.rm = TRUE),
    sd = sd(value, na.rm = TRUE),
    min = min(value, na.rm = TRUE),
    max = max(value, na.rm = TRUE),
    n_missing = sum(is.na(value)),
    .groups = 'drop'
  )

print(num_summary)
```

```r
ggplot(nba_data, aes(x = height_in)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(
    title = "Height Distribution of NBA Drafted Players",
    x = "Height (inches)",
    y = "Count"
  )
ggplot(nba_data, aes(x = weight)) +
  geom_histogram(binwidth = 10, fill = "orange", color = "black") +
  labs(
    title = "Weight Distribution of NBA Drafted Players",
    x = "Weight (lbs)",
    y = "Count"
  )
nba_data %>%
  ggplot(aes(x = factor(round_number), y = height_in)) +
  geom_boxplot() +
  labs(
    title = "Height by Draft Round",
    x = "Draft Round",
    y = "Height (inches)"
  )
nba_data %>%
  ggplot(aes(x = factor(round_number), y = weight)) +
  geom_boxplot() +
  labs(
    title = "Weight by Draft Round",
    x = "Draft Round",
    y = "Weight (lbs)"
  )
round_summary <- nba_data %>%
  group_by(round_number) %>%
  summarise(
    Avg_Height = mean(height_in, na.rm = TRUE),
    Median_Height = median(height_in, na.rm = TRUE),
    Avg_Weight = mean(weight, na.rm = TRUE),
    Median_Weight = median(weight, na.rm = TRUE),
    n_players = n(),
    .groups = 'drop'
  ) %>%
  mutate(round_number = paste("Round", round_number))
```

```r
knitr::kable(round_summary, caption = "Height and weight by draft round")
ggplot(nba_data, aes(x = height_in, y = weight)) +
  geom_point(aes(color = bmi), alpha = 0.5) +
  scale_color_gradientn(
    name = "BMI",
    colors = c("blue", "green", "yellow", "red"),
    breaks = c(20, 25, 30)
  ) +
  labs(
    title = "NBA players height and weight distribution (color = BMI)",
    x = "Height (inches)",
    y = "Weight (lbs)"
  )
```

9 References

- O'Walsh, W. (2025). *Basketball Data* [Data set]. Kaggle.
- NBA. (n.d.). *Official Player Stats*. NBA.com.