NBA_Statistical_Analysis

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 Provenance

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.

3 Setup & Data Cleaning

```r
library(dplyr)
library(janitor)
library(ggplot2)
library(tidyr)
library(readr)
library(scales)
library(patchwork)
library(stringr)

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

# Clean Player Info
player_clean <- player_info %>%
  mutate(
    height = as.numeric(str_extract(height, "[0-9]+")),
    weight = as.numeric(str_extract(weight, "[0-9]+")),
    across(c(height, 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))

# Join Datasets
nba_data <- inner_join(player_clean, draft_clean, by = "person_id")
```

4 Exploratory Data Analysis

4.1 Glimpse of Data

```
glimpse(nba_data)
```

```
Rows: 2,985
Columns: 46
$ 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                           <dbl> 6, 7, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6~
$ 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~
$ 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

```r
num_summary <- nba_data %>%
  select(height, weight, season_exp, round_number, round_pick, draft_type, player_profile_fla
  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)
```

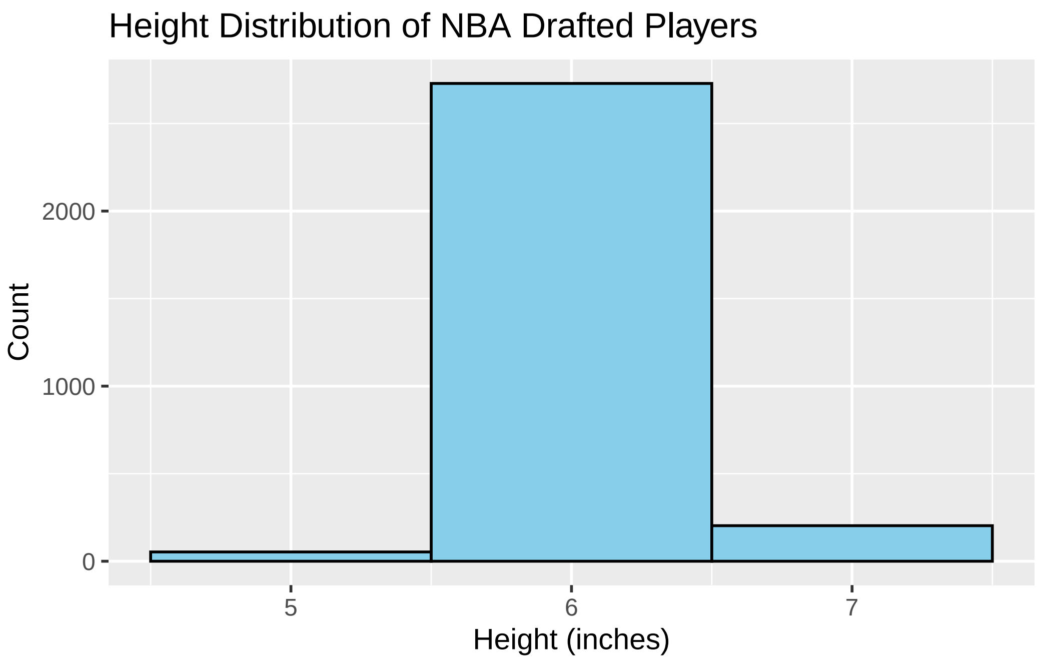
```
# A tibble: 7 x 7
  variable             mean median      sd   min   max n_missing
  <chr>               <dbl>  <dbl>   <dbl> <dbl> <dbl>     <int>
1 height               6.05      6  0.289      5     7         0
2 overall_pick        30.6      24 29.8        0   221         0
3 player_profile_flag  1.00      1  0.0183     0     1         0
4 round_number         2.06      2  1.92       0    20         0
```

```
5 round_pick           11.0        9  8.09       0    30          0
6 season_exp            5.98       4  4.70       0    22          0
7 weight               212.      210 26.4      133   325          0
```

5 Graphs

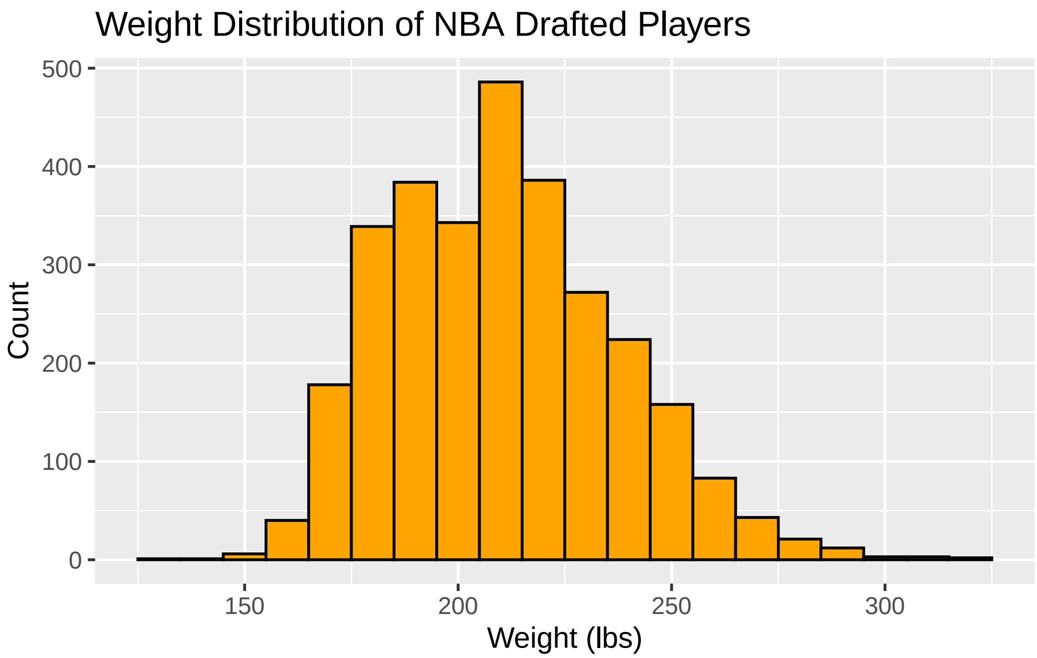
5.1 Height Distribution

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



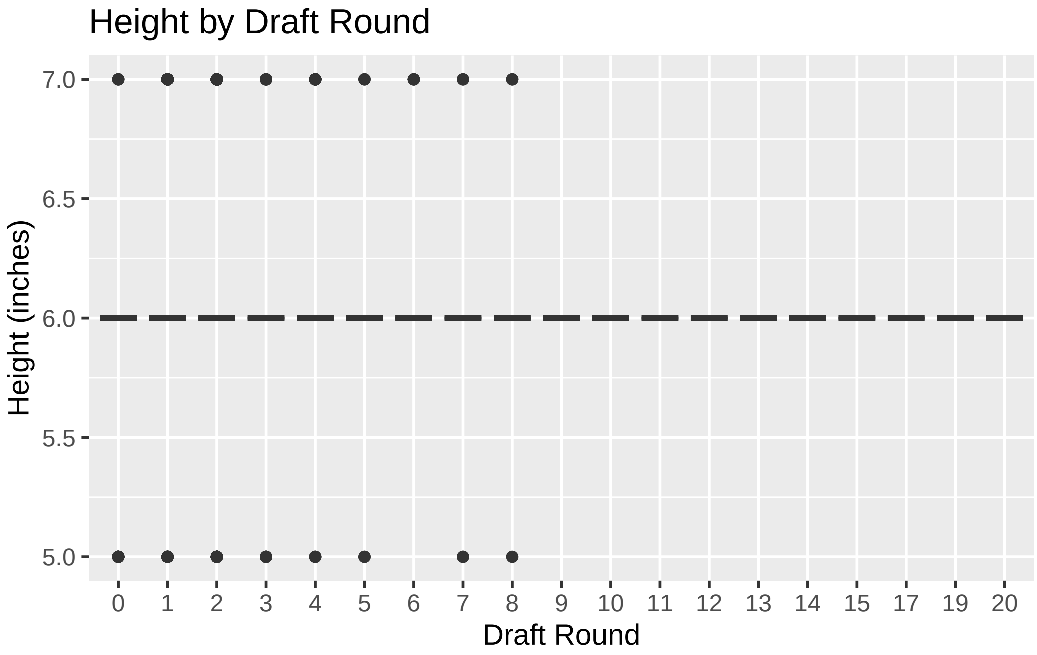
5.2 Weight Distribution

```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"
  )
```



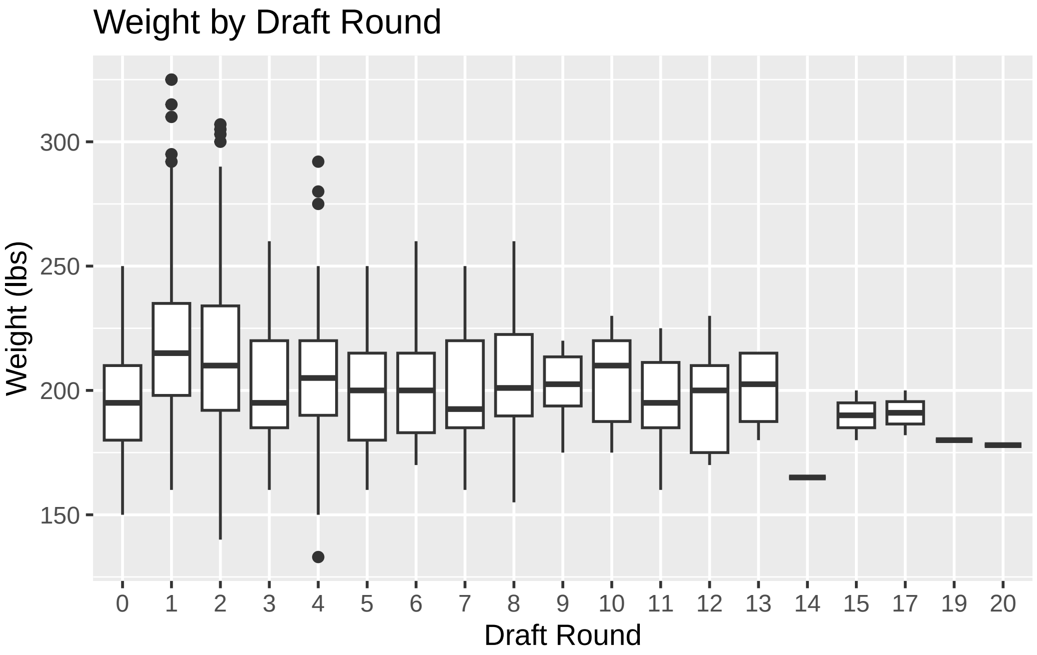
5.3 Draft Round vs Height

```r
nba_data %>%
  ggplot(aes(x = factor(round_number), y = height)) +
  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)"
  )
```



6 Narrative Summary

The majority of NBA drafted players cluster around standard height and weight ranges. 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 overall variation still exists among players across all rounds.

8 Code Appendix

```r
library(dplyr)
library(janitor)
library(ggplot2)
library(tidyr)
library(readr)
library(scales)
library(patchwork)
library(stringr)

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

# Clean Player Info
player_clean <- player_info %>%
  mutate(
    height = as.numeric(str_extract(height, "[0-9]+")),
    weight = as.numeric(str_extract(weight, "[0-9]+")),
    across(c(height, 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))

# Join Datasets
nba_data <- inner_join(player_clean, draft_clean, by = "person_id")
glimpse(nba_data)
num_summary <- nba_data %>%
  select(height, weight, season_exp, round_number, round_pick, draft_type, player_profile_fla
  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)
ggplot(nba_data, aes(x = height)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(
```

```r
    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)) +
  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)"
  )
```

9 References

- O'Walsh, W. (2025). *Basketball Data* [Data set]. Kaggle.

- NBA. (n.d.). *Official Player Stats.* NBA.com.