ANA 515 Assignment 2

Chester Kayonga

2024-06-12

# Load the dataset using the function read.csv  
url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/nba-draft-2015/historical\_projections.csv"  
Nba2015draft\_data <- read.csv(url)

# library for data manipulation  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Filter the dataset to include only rows where Projected.SPM is greater than or equal to 0.2  
filtered\_data <- Nba2015draft\_data %>%  
 filter(Projected.SPM >= 0.2)  
# Rename columns for better readability  
cleaned\_data <- filtered\_data %>%  
 rename(  
 Projected\_SPM = Projected.SPM,  
 Superstar\_Prob = Superstar,  
 Starter\_Prob = Starter,  
 Role\_Player\_Prob = Role.Player,  
 Bust\_Prob = Bust  
 )  
  
# Display the first few rows of the cleaned dataset  
head(cleaned\_data)

## Player Position ID Draft.Year Projected\_SPM  
## 1 Karl-Anthony Towns C karl-anthony-towns 2015 1.0306057  
## 2 Justise Winslow SF justise-winslow 2015 0.8753290  
## 3 Stanley Johnson SF stanley-johnson 2015 0.6794933  
## 4 Jahlil Okafor C jahlil-okafor 2015 0.5216613  
## 5 D`Angelo Russell PG d-angelo-russell 2015 0.5119667  
## 6 Dakari Johnson C dakari-johnson 2015 0.4917874  
## Superstar\_Prob Starter\_Prob Role\_Player\_Prob Bust\_Prob  
## 1 0.13476667 0.4271833 0.16308333 0.2749667  
## 2 0.08352857 0.5109048 0.17676667 0.2288000  
## 3 0.06780000 0.4237333 0.27850000 0.2299667  
## 4 0.05871667 0.4099000 0.23553333 0.2958500  
## 5 0.15203333 0.3422833 0.09658333 0.4091000  
## 6 0.02134078 0.3675436 0.41757200 0.1935437

# Characteristics of the Data

This dataframe has 99 rows and 9 columns. The names of the columns and a brief description of each are in the table below:

# Load necessary library for table creation  
library(knitr)   
  
# Create a dataframe for column descriptions  
column\_descriptions <- data.frame(  
 Column\_Name = c("Player", "College", "Year", "Position", "Height", "Wingspan",   
 "Projected\_SPM", "Superstar\_Prob", "Starter\_Prob",   
 "Role\_Player\_Prob", "Bust\_Prob"),  
 Description = c("Player's Name", "College attended", "Year of draft", "Position played",   
 "Height in inches", "Wingspan in inches",   
 "Projected Statistical Plus/Minus", "Probability of becoming a Superstar",   
 "Probability of becoming a Starter", "Probability of becoming a Role Player",   
 "Probability of becoming a Bust")  
)  
  
# Display the table  
kable(column\_descriptions, caption = "Column Names and Descriptions")

Column Names and Descriptions

| Column\_Name | Description |
| --- | --- |
| Player | Player’s Name |
| College | College attended |
| Year | Year of draft |
| Position | Position played |
| Height | Height in inches |
| Wingspan | Wingspan in inches |
| Projected\_SPM | Projected Statistical Plus/Minus |
| Superstar\_Prob | Probability of becoming a Superstar |
| Starter\_Prob | Probability of becoming a Starter |
| Role\_Player\_Prob | Probability of becoming a Role Player |
| Bust\_Prob | Probability of becoming a Bust |

# Summary Statistics

We will generate summary statistics for three columns: Projected\_SPM, Superstar\_Prob, and Starter\_Prob.

# Subsetting the Data

# Subset the dataset to include only the columns of interest  
data\_pick3 <- select(cleaned\_data, Projected\_SPM, Superstar\_Prob, Starter\_Prob)

# Summary of the Subset

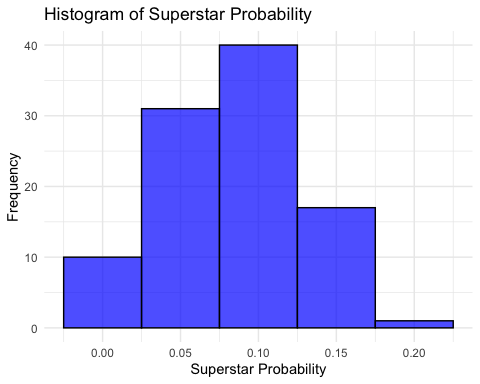
# Produce a summary of the subset  
Summarytable <- summary(data\_pick3) # Creates the summary  
Summarytable # Prints the summary in your output

## Projected\_SPM Superstar\_Prob Starter\_Prob   
## Min. :0.2042 Min. :0.00830 Min. :0.2515   
## 1st Qu.:0.3336 1st Qu.:0.05911 1st Qu.:0.3679   
## Median :0.4924 Median :0.08152 Median :0.4001   
## Mean :0.5922 Mean :0.08685 Mean :0.4007   
## 3rd Qu.:0.7923 3rd Qu.:0.11713 3rd Qu.:0.4247   
## Max. :1.7768 Max. :0.18198 Max. :0.6557

## Data Visualizations

## Histogram of superstar probability

# Load necessary library for data visualization  
library(ggplot2)  
  
# Create a histogram of Superstar\_Prob  
ggplot(cleaned\_data, aes(x = Superstar\_Prob)) +  
 geom\_histogram(binwidth = 0.05, fill = "blue", color = "black", alpha = 0.7) +  
 labs(title = "Histogram of Superstar Probability",  
 x = "Superstar Probability",  
 y = "Frequency") +  
 theme\_minimal()



##Scatterplot of Projected SPM vs. Scatterplot

# Create a scatter plot of Projected\_SPM vs. Starter\_Prob  
ggplot(cleaned\_data, aes(x = Projected\_SPM, y = Starter\_Prob)) +  
 geom\_point() +  
 labs(title = "Scatter Plot of Projected SPM vs. Starter Probability",  
 x = "Projected SPM",  
 y = "Starter Probability") +  
 theme\_minimal()

