Project #1

STAT 494/873

Fall 2025

The National Football League (NFL) holds a scouting combine every year for college football players who would like to play football in the NFL. These players go through a number of evaluations during the combine so that NFL teams can assess their ability. For more information, please see <http://en.wikipedia.org/wiki/NFL_Scouting_Combine> and <http://www.nfl.com/combine>.

The NFLcombine.csv data file is available from my course website, and it contains information on some of the players who participated in a combine. The columns in the data file represent the following information:

* Player: Name of player being evaluated
* College: College that the player attended
* Position: The position of the player where DB = defensive back, LB = linebacker, OL = offensive linemen, RB = running back, S = safety, TE = tight end, WO = wide receiver; players who played other positions were excluded from the data file
* OverallGrade: The overall grade of the player based on the evaluations
* Height: Height in inches
* ArmLength: Arm length in inches
* Weight: Weight in pounds
* Dash40: 40-yard dash time in seconds
* BenchPress: Number of bench press repetitions of 225 pounds
* VerticalJump: Vertical jump in inches
* BroadJump: Broad jump in inches
* Cone3Drill: 3-cone drill time in seconds
* Shuttle20: 20-yard shuttle run in seconds

Below is how I read the data into R:

> fb <- read.csv(file = "NFLcombine.csv", stringsAsFactors = TRUE))

I used the stringsAsFactors argument with a value of TRUE to make sure categorical variables are treated as being in this format rather than simply as character strings. This is helpful when we want to examine differences among players in particular categories.

Complete the following problems below using the data. While you are welcome to use your football knowledge to help with interpretations, this is not needed to perform well on this project. Include your R program output with code inside of it for each part and any additional information needed to explain your answer. Your R code and output should be formatted in the same manner as in the lecture notes.

1. This part focuses on plotting the data and interpreting the corresponding plots.
   1. (4 points) Construct side-by-side scatter plots for the numerical variables. Compute the estimated correlation matrix for these same variables. Interpret the plots and the matrix in the context of the data.
   2. (6 points) Create a parallel coordinates plot for the numerical variables. The color of the lines for each player should correspond to their position in the following way:

*# Suppose the fb data frame contains the data*

library(plyr)

color.position <- palette()[1:length(levels(fb$Position))]

Position.color <- revalue(x = fb$Position, replace = c(DB = color.position[1],

LB = color.position[2], OL = color.position[3], RB = color.position[4],

S = color.position[5], TE = color.position[6], WO = color.position[7]))

Are there any trends in the plot corresponding to the position? If there are trends, explain what characteristics of the plot lead you to this conclusion. Are there any outliers? If there are outliers, identify them by observation number and player name along with discussing the characteristics of the plot that lead you to this conclusion.

* 1. (6 points) Create a stars plot for the numerical variables. Sort the data by position in the following manner before plotting:

fb[order(fb$Position),]

and use 10 columns of stars per row. Are there any trends in the plot corresponding to player position? If there are trends, explain what characteristics of the plot lead you to this conclusion. Are there any outliers? If there are outliers, identify them by observation number and player name along with discussing the characteristics of the plot that lead you to this conclusion.

* 1. (3 points) Would it be difficult to represent ALL of the numerical variables on one Trellis plot? If your answer is yes, fully explain why. If your answer is no, construct the plot and interpret it.

1. This part focuses on using PCA with the correlation matrix. **Exclude the OverallGrade variable in the analysis.**
   1. (3 points) Why is the correlation matrix more appropriate to use here than the covariance matrix with this data?
   2. (5 points) Determine the number of PCs needed. Fully justify your answer.
   3. (6 points) Interpret the PCs chosen from b).
   4. (4 points) Show how the first PC score for the first observation is found using matrix algebra in R and through using by-hand calculations. Use predict() or the scores component from princomp() to check your answer.
   5. (4 points) Construct a bubble plot of the first three PC scores. Include the observation number (black color) in the middle of each bubble. Colorize each bubble by position using

fg = fb$Position

in the symbols() function (the colors will be the same as in 1)b)).

* 1. (4 points) Identify outliers on the plot from e) with their observation number and player name. Explain why they are outliers relative to the characteristics of the PCs.
  2. (4 points) Again using the plot from e), discuss the differences between the PC scores with respect to the position of the player. Include characteristics of the PCs in your discussion.
  3. (4 points) Construct a 3D plot of the PC scores using plot3d() and common x-, y-, and z-axis limits. Color the plotting points using the position of players (same color scheme as in 1)b)). Are there any new items that were not apparent with the bubble plot? Explain.