Success from NCAA to NFL Hall of Fame

2022-12-7

Introduction

College football excites hundreds of thousands of people across America year by year, and many collegiate players dream of playing professionally in the NFL. Only a small percentage (1.6%) of college-level players fulfill that dream and get drafted into the NFL. Within this extremely selective process, certain universities and their respective conferences produce more NFL-bound athletes than others. However, does draft success equate to professional success? We are looking to investigate which NCAA conference produces the most successful players in the NFL.

Our final dataset will be composed of 3 individual datasets. Our first dataset (https://www.pro-football-reference.com/schools/ (https://www.pro-football-reference.com/schools/)) lists 847 different colleges/universities in America based on how many of their players have gone on to play in the NFL. The dataset also lists the number of current active professional players, number of Hall of Famers, number of Pro Bowl selections, total NFL games played amongst alumni, total touchdowns scored by alumni, and the alumnus with the highest career Approximate Value. Each unique row represents an individual school, and the dataset contains both numeric and categorical variables. Our second dataset (https://collegefootballdata.com/exporter/teams/fbs (https://collegefootballdata.com/exporter/teams/fbs)) lists every school that has an intercollegiate football program, and also lists each school's conference. In this dataset, a unique row also represents an individual school, and only categorical variables are present. Our third dataset (https://www.pro-football-reference.com/leaders/career_av_career.htm (https://www.pro-football-reference.com/leaders/career_av_career.htm)) lists the top 255 NFL players of all time based on their weighted Approximate Value, which is a measure of the player's impact on the team based on performance and longevity. The dataset also lists the years of a player's career, as well as the number of different teams a player played for. In this dataset, a row represents an individual NFL player, and both numeric and categorical variables are present.

Based on these datasets, we may expect to see a positive correlation between a conference's average weighted Approximate Value and the number of professional players it has produced.

One research question we wish to analyze is which individual player metrics correlate most with induction into the NFL Hall of Fame. Another research question we will look to analyze is whether school conference predicts induction into the NFL Hall of Fame.

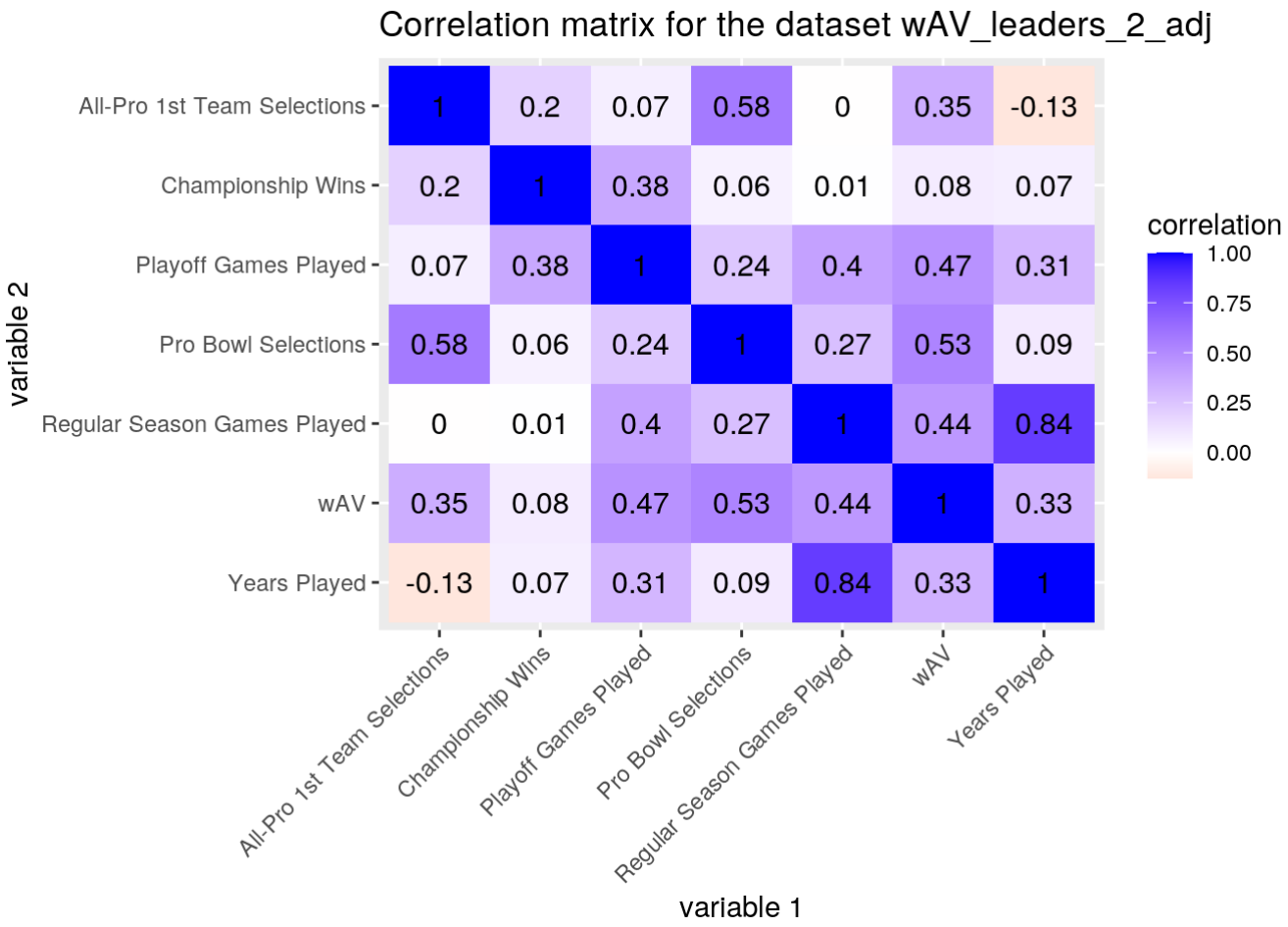
Exploratory Data Analysis

For the wAV_leaders_2 dataset, there were originally 255 observations (players) and 5 unique IDs (Rank, Player, wAV, Years, Tm). Since all 5 of the IDs were unique to the wAV_leaders dataset when compared to the other datasets, the School ID variable was added beforehand to provide a common variable for joining. Additionally, the Hall of Famer, Active, Pro Bowl Selections, All-Pro 1st Team Selections, and Playoff Games Played ID variables were manually added to the dataset to provide more statistics on individual players; datasets containing these specific variables together were difficult to find.

By creating a correlation matrix, we can see that the variables that correlate the most are Years Played and Regular Season Games Played. This is expected, as playing for a longer time results in more opportunities to play games. On the other hand, the variables that correlate the least are All-Pro 1st Team Selections and Years Played. This is also somewhat expected, as playing for longer does not necessarily mean that a player performs at an especially high level consistently.

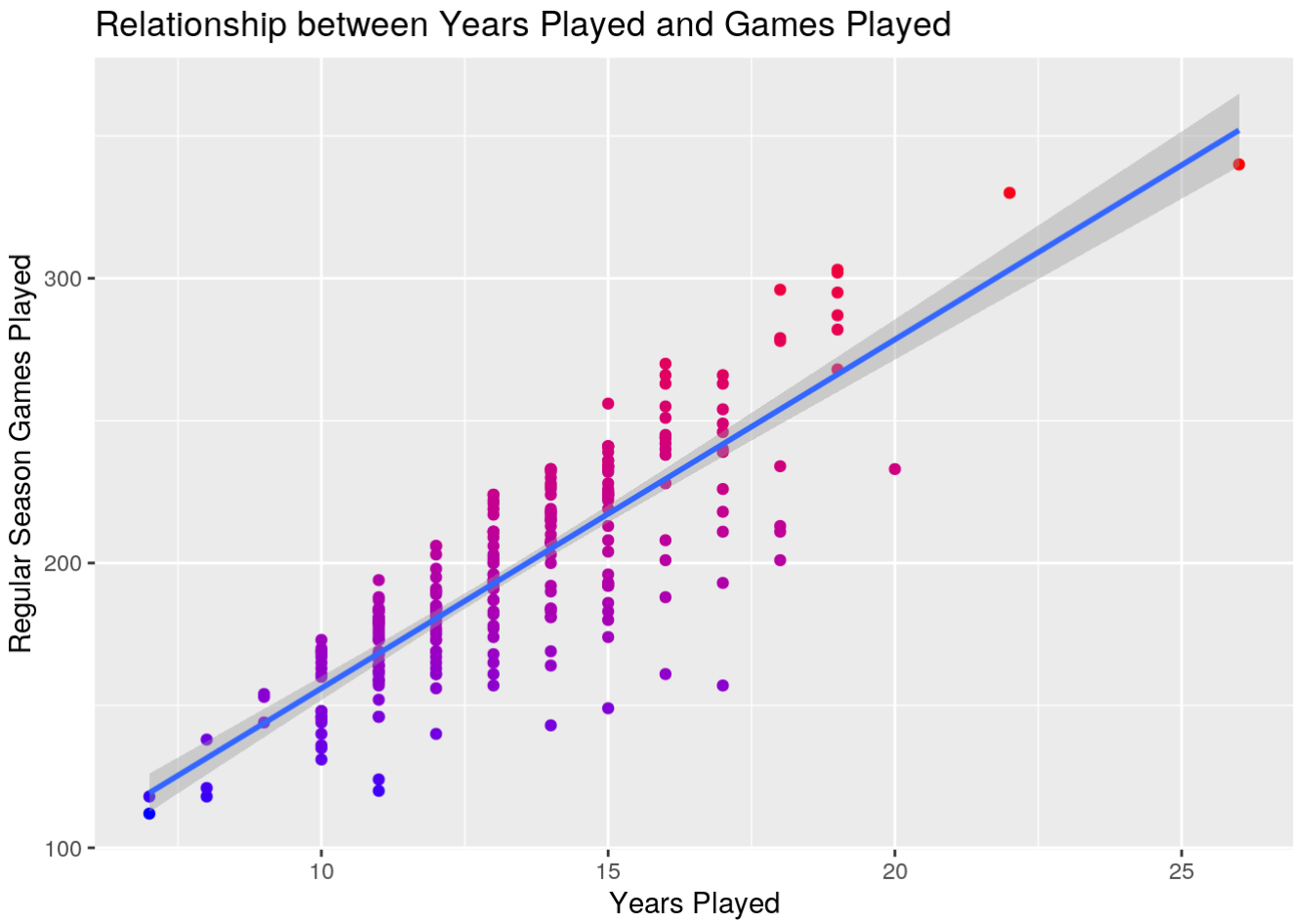
```
# Create new object in environment
wAV_leaders_2_adj <- wAV_leaders_2 %>%
  select(-Rank,-Tm) %>% # Remove specific variable columns
  separate(Years, into = c("Rookie Year", "Final Year")) %>% # Separate variable into tw
o columns
  mutate(`Rookie Year` = as.numeric(`Rookie Year`), # Change variable class
         `Final Year` = as.numeric(`Final Year`),
         `Years Played` = `Final Year` - `Rookie Year`) %>% # Create new Years Played va
riable
  select(-`Rookie Year`,-`Final Year`) %>% # Remove specific variable columns
  relocate(`Years Played`,.after = wAV) # Move variable column to specific place in data
frame

wAV_leaders_2_adj %>%
  select(-Player,-School,-`Hall of Famer`, -Active) %>% # Remove specific variable colum
ns
  cor(use = "pairwise.complete.obs") %>% # Create correlation matrix
  as.data.frame %>% # Save as a data frame
  rownames_to_column %>% # Convert row names to an explicit variable
  pivot_longer(-1,
               names_to = "other_var",
               values_to = "correlation") %>% # Pivot so that all correlations appear in
the same column
  ggplot(aes(x = rowname,
             y = ordered(other_var, levels = rev(sort(unique(other_var)))), # Define ggp
lot (reorder values on y-axis)
             fill = correlation)) +
  geom_tile() +   # Heat map with geom_tile
  scale_fill_gradient2(low = "red", mid = "white", high = "blue") +  # Change the scale
 to make the middle appear neutral
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) + # Overlay va
lues
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) + # Angle the x-axis label to
45 degrees
  labs(title = "Correlation matrix for the dataset wAV_leaders_2_adj", # Give title and
 labels
       x = "variable 1", y = "variable 2")
```

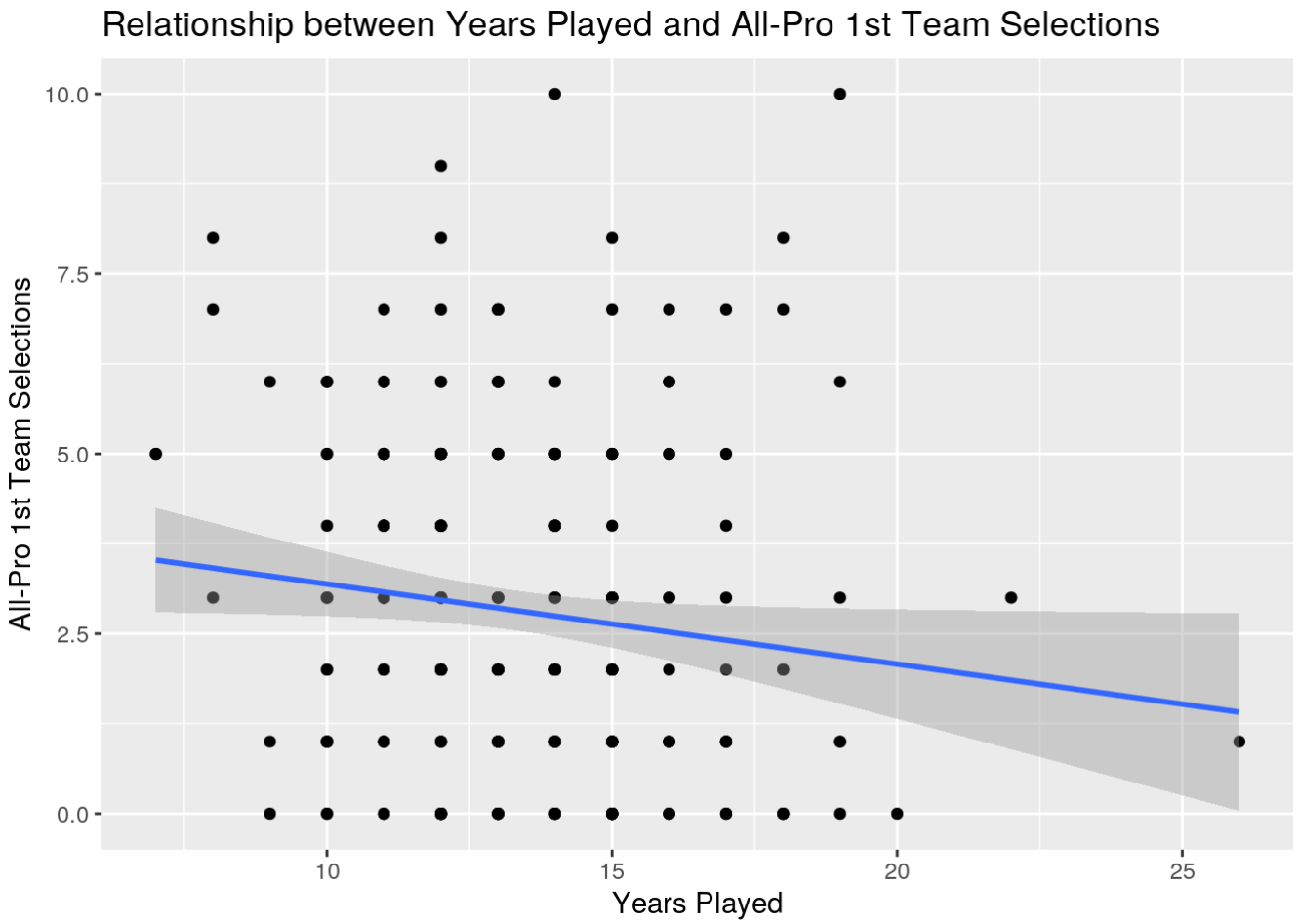


With visualizations, we can clearly see the strong, positive correlation between Years Played and Regular Season Games Played, as well as the absence of correlation between Years Played and All-Pro 1st Team Selections. These correlations are within reason, as playing for longer results in more chances to play games, while playing longer does not always lead to playing at a high level in a given year.

```
# Create ggplot based on Years Played and Games Played
ggplot(wAV_leaders_2_adj, aes(x = `Years Played`, y = `Regular Season Games Played`, col
or = `Regular Season Games Played`)) +
  scale_colour_gradient(low="blue", high="red") +
  theme(legend.position="none") +
  geom_point() + geom_smooth(method = "lm") +
  labs(title = "Relationship between Years Played and Games Played")
```




```
# Create ggplot based on Years Played and All-Pro 1st Team Selections
ggplot(wAV_leaders_2_adj, aes(x = `Years Played`, y = `All-Pro 1st Team Selections`)) +
  geom_point() + geom_smooth(method = "lm") +
  labs(title = "Relationship between Years Played and All-Pro 1st Team Selections")
```



To prepare the joining of the first three datasets (college, conference, wAV), the conference dataset's names for each school must match the names in the college dataset.

```
# Save object in environment
ncaa_schools_adj <- ncaa_schools %>%
  mutate(School = ifelse(School == "Miami",str_replace(School, "Miami", "Miami (FL)"),Sc
hool), # Replace specific string with a different string
         School = ifelse(School == "Grambling",str_replace(School, "Grambling","Gramblin
g State"),School),
         School = ifelse(School == "NC State",str_replace(School, "NC State","North Caro
lina State"),School),
         School = ifelse(School == "Ole Miss",str_replace(School, "Ole Miss","Mississipp
i"),School),
         School = ifelse(School == "Mississippi Valley State",str_replace(School, "Missi
ssippi Valley State","Miss. Valley State"),School),
         School = ifelse(School == "Southern Mississippi",str_replace(School, "Southern
 Mississippi","Southern Miss"),School),
         School = ifelse(School == "UTEP",str_replace(School, "UTEP","Texas-El Paso"),Sc
hool),
         School = ifelse(School == "Prairie View",str_replace(School, "Prairie View","Pr
airie View A&M"),School),
         School = ifelse(School == "Maryland-Eastern Shore",str_replace(School, "Marylan
d-Eastern Shore","Md-Eastern Shore"),School))
```

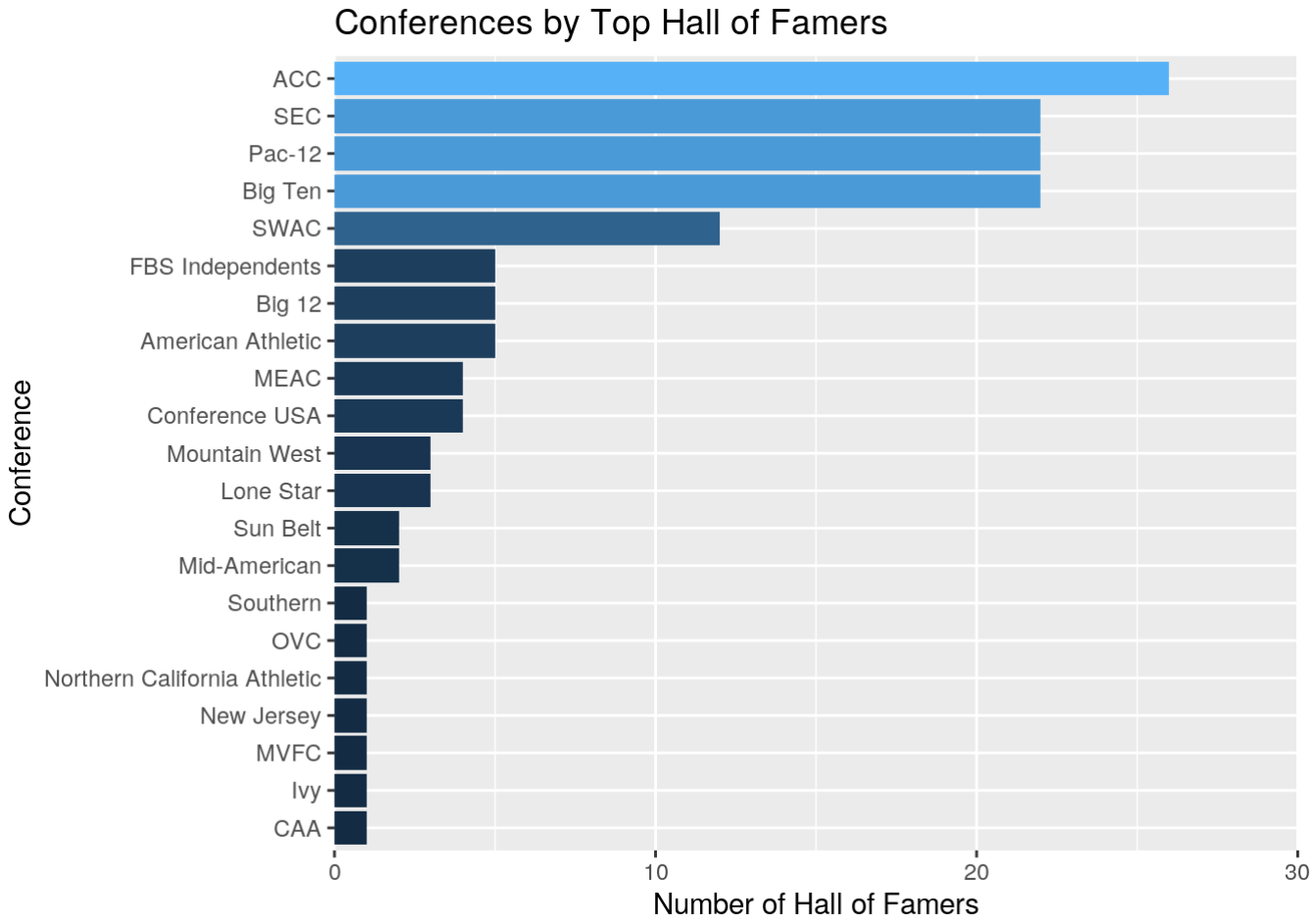
Before joining the datasets together, some column variables in the college dataset were removed. Since the focus is on metrics that correlate to induction into the Hall of Fame, statistics that do not help contribute to correlation were removed. Additionally, the names of certain schools were adjusted to match with the names in the conference dataset.

After joining the wAV dataset to the college dataset and filtering out any schools that did not produce any of the top 255 players, there were 255 observations left, matching the wAV dataset's original number of observations. We will also filter out any players who are active in the league, as active players are ineligible for the Hall of Fame. This leaves us with 238 observations, and we will now remove the Active variable because all remaining observations are retired players. After joining the conference dataset to the new college dataset, the number of observations remained constant because the joining simply added the Conference ID to the college dataset.

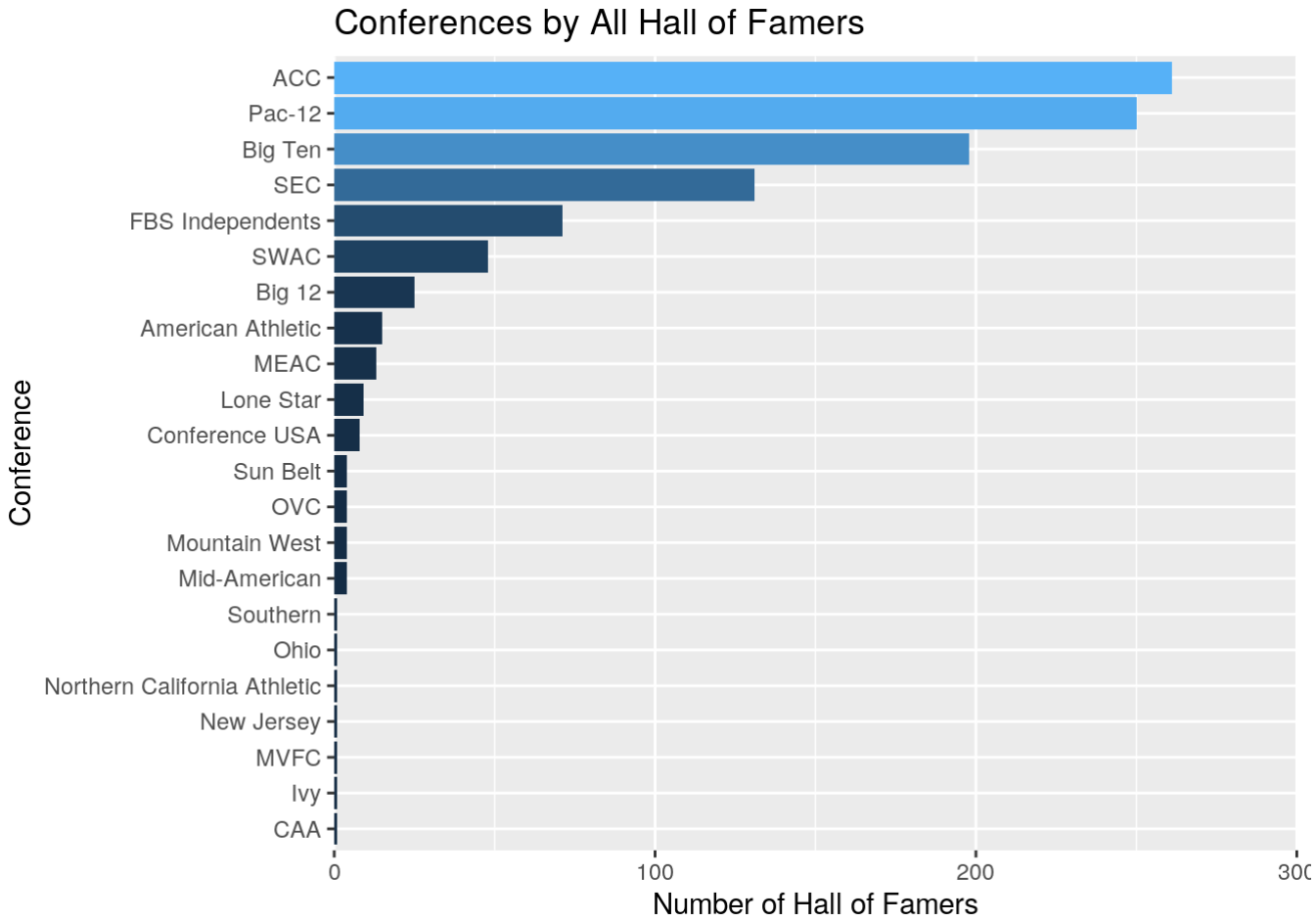
```
# Create new object in environment
hall_of_famers <- college_stats %>%
  select(-Rk,-State,-`Active Players`,-Touchdowns,-`Highest Career Approximate Value`,-`
Approximate Value`) %>% # Remove specific column variables
  rename(`Total Hall of Famers` = `Hall of Famers`,
         `Total Pro Bowl Selections` = `Pro Bowl Selections`,
         `Total Games Played` = `Games Played`) %>% # Rename variables
  left_join(wAV_leaders_2_adj,by="School") %>% # Apply left_join
  filter(!is.na(Player), # Remove observations that are empty in the School variable
         !Active == "TRUE") %>% # Remove players who are currently active in the NFL
  select(-Active) %>%
  mutate(School = str_replace(School, "St\\.$", "State"), # Replace specific string with
a different string
         School = str_replace(School, "Col\\.", "College"),
         School = str_replace(School, "East\\.", "Eastern")) %>%
  left_join(ncaa_schools_adj, by = "School") %>% # Apply left_join
  relocate(Conference, .after=School) # Move conference variable to display following sc
hool variable
```

When looking at the distribution of Hall of Famers by conference, we see a significant disparity between four of the Power Five conferences (ACC, Big Ten, Pac-12, SEC) and the rest of the conferences, which is to be expected. These conferences are the most prominent and highest-earning conferences, meaning they have the resources to consistently produce top-tier football talents who have the skills to become all-time greats in the sport. Interestingly, the remaining Power Five conference (Big 12) has produced significantly fewer Hall of Famers than the rest of the Power Five, but this may be due to only observing the top 255 players. However, when observing all players throughout the history of the NFL, this trend remains; the Big 12 conference has produced fewer Hall of Famers than two non-Power Five conferences (FBS Independents, SWAC). In both visualizations, the ACC has the most inductees into the Hall of Fame. Based on these visualizations, we can conclude that playing football at a Power Five conference has the highest chance for being inducted into the Hall of Fame. However, whether that is due to the conferences themselves producing high-level talent or the individual players having the talent and consistency to produce high-level performances over many years is yet to be determined.

```
# Create summary statistic for distribution of Hall of Famers/non-Hall of Famers by conf
erence
hall_of_famers %>%
  mutate(`Hall of Famer` = ifelse(`Hall of Famer` == TRUE, 1, 0)) %>% # Rewrite variable
s
  group_by(Conference, `Hall of Famer`) %>% # Create subsets by conference, by Hall of F
ame status
  summarize(n=n()) %>% # Count number of observations in each subset
  arrange(desc(n)) %>% # Arrange in descending order
  filter(!`Hall of Famer` == 0) %>% # Remove schools that do not have any Hall of Famers
  ggplot(aes(x=n,y=reorder(Conference,n),fill=n))+ # Create ggplot based on conference a
nd number of Hall of Famers
  geom_bar(stat="identity") + # Create bar graph
  scale_x_continuous(expand=c(0,0), # Remove space between data and axes
                     limits=c(0,30)) + # Set x-axis to start at 0, end at 50
  labs(x="Number of Hall of Famers",y = "Conference", title="Conferences by Top Hall of
 Famers") + # Create labels
  theme(legend.position = "none") # Remove legend
```



```
hall_of_famers %>%
  group_by(Conference) %>% # Create subsets by conference
  summarize(n=sum(`Total Hall of Famers`)) %>% # Count number of historical Hall of Fame
rs for each conference
  arrange(desc(n)) %>% # Arrange in descending order
  filter(!n == 0) %>% # Remove schools that do not have any Hall of Famers
  ggplot(aes(x=n,y=reorder(Conference,n),fill=n))+ # Create ggplot based on Conference a
nd number of Hall of Famers
  geom_bar(stat="identity") + # Create bar graph
  scale_x_continuous(expand=c(0,0), # Remove space between data and axes
                     limits=c(0,300)) + # Set x-axis to start at 0, end at 50
  labs(x="Number of Hall of Famers",y = "Conference", title="Conferences by All Hall of
 Famers") + # Create labels
  theme(legend.position = "none") # Remove legend
```



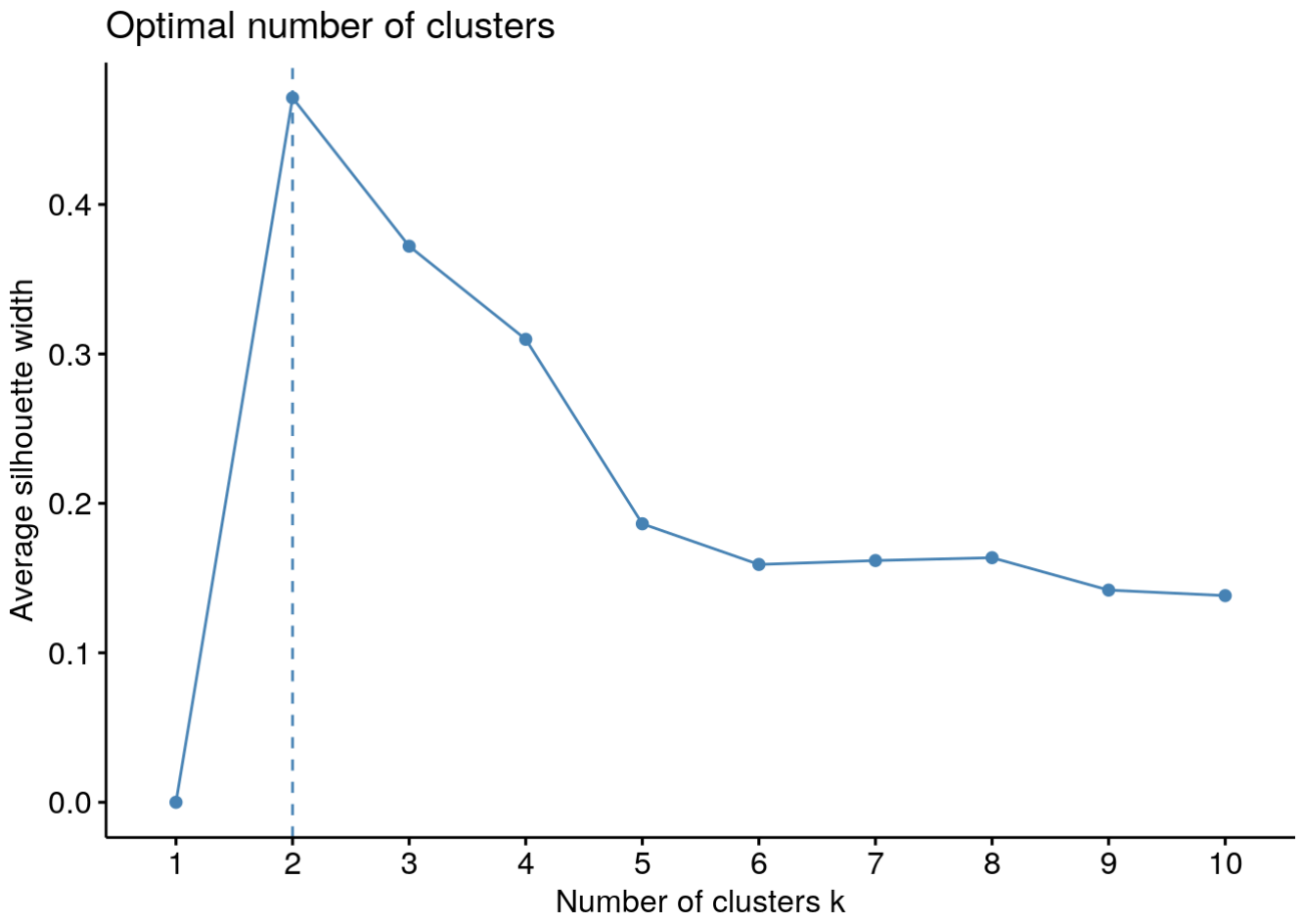
Clustering/Dimensionality Reduction

Before clustering, we first removed any variables that were not individual-specific because this clustering focused on individual players rather than colleges or conferences. After performing clustering based on Gower's dissimilarities, we decided that the optimal number of clusters for the dataset would be 2, as that would lead to the highest average silhouette width.

```
# Create new object in environment
hall_of_famers_reduced <- hall_of_famers %>%
  # Drop some categorical variables with too many categories, drop numerical variables t
hat are not individual-specific
  select(-School,-Conference,-Player,-`Total Hall of Famers`, -`Total Pro Bowl Selection
s`, -`Total Games Played`) %>%
  # Consider categorical variables as factors
  mutate_if(is.character, as.factor) %>%
  # Ignore missing values
  drop_na

# Calculate Gower distances between observations
hall_of_famers_reduced %>%
  # No need to scale when calculating the Gower's distance
  daisy(metric = "gower") %>%
  # Save as a matrix
  as.matrix -> hall_of_famers_gower

# Use the silhouette on the matrix of distances
fviz_nbclust(hall_of_famers_gower, pam, method = "silhouette")
```



In terms of the numeric variables, cluster 1 has greater values than cluster 2 in all variables other than Years Played. This is expected because playing for a longer period of time does not necessarily correlate with consistently amazing performances. Based on these individual-specific numeric variable summary statistics, cluster 1 appears to encompass the Hall of Famers, while cluster 2 apepars to contain the non-Hall of Famers.

```
# Apply PAM on the dissimilarity object (specify diss = TRUE)
pam_results <- pam(hall_of_famers_gower, k = 2, diss = TRUE)

# Save cluster assignment as a column in your dataset
hall_of_famers_pam <- hall_of_famers_reduced %>%
  mutate(cluster = as.factor(pam_results$clustering))

# Summary statistics of numeric variables
hall_of_famers_pam %>%
  group_by(cluster) %>%
  summarize_if(is.numeric, mean, na.rm = T)
```

```
## # A tibble: 2 × 9
##   cluster Players   wAV `Years Played` Regular…¹ Pro B…² All-P…³ Champ…⁴ Playo…⁵
##   <fct>     <dbl> <dbl>          <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1          284.  111.           13.3      198.    7.85    3.65   1.46     12.5
## 2 2          241.  101.           13.4      195.    4.96    1.59   0.585    10.7
## # … with abbreviated variable names ¹`Regular Season Games Played`,
## #   ²`Pro Bowl Selections`, ³`All-Pro 1st Team Selections`,
## #   ⁴`Championship Wins`, ⁵`Playoff Games Played`
```

Looking at the categorical variable summary statistics, what we assumed based on the numerical statistics was correct. Cluster 1 contains Hall of Famers, while cluster 2 is only comprised of non-Hall of Famers.

When observing the center of each cluster, the two players are not too far apart in statistics, which was interesting to see. The Hall of Famer (Gary Zimmerman) actually played fewer regular-season games compared to the non-Hall of Famer (Justin Smith), though the rest of the metric comparisons all came out as expected. This shows that there are most likely metrics that contribute to a player's induction into the Hall of Fame other than the ones we selected.

The silhouette width came out to be around 0.43, which is decent, but not amazing.

```
# Summary statistics of the categorical variable Hall of Famer
hall_of_famers_pam %>%
  group_by(cluster, `Hall of Famer`) %>%
  summarize(freq = n())
```

```
## # A tibble: 2 × 3
## # Groups:   cluster [2]
##   cluster `Hall of Famer`  freq
##   <fct>   <lgl>           <int>
## 1 1       TRUE              144
## 2 2       FALSE              94
```

```
# Look at the final medoids
hall_of_famers[pam_results$id.med,]
```

```
## # A tibble: 2 × 15
##   School   Confer…¹ Players Total…² Total…³ Total…⁴ Player   wAV Years…⁵ Hall …⁶
##   <chr>    <chr>      <dbl>   <dbl>   <dbl>   <dbl> <chr>  <dbl>   <dbl> <lgl>
## 1 Oregon   Pac-12       284       6      86   12976 Gary …   103      11 TRUE
## 2 Missouri SEC          239       2      45    9357 Justi…   101      13 FALSE
## # … with 5 more variables: `Regular Season Games Played` <dbl>,
## #   `Pro Bowl Selections` <dbl>, `All-Pro 1st Team Selections` <dbl>,
## #   `Championship Wins` <dbl>, `Playoff Games Played` <dbl>, and abbreviated
## #   variable names ¹Conference, ²`Total Hall of Famers`,
## #   ³`Total Pro Bowl Selections`, ⁴`Total Games Played`, ⁵`Years Played`,
## #   ⁶`Hall of Famer`
```

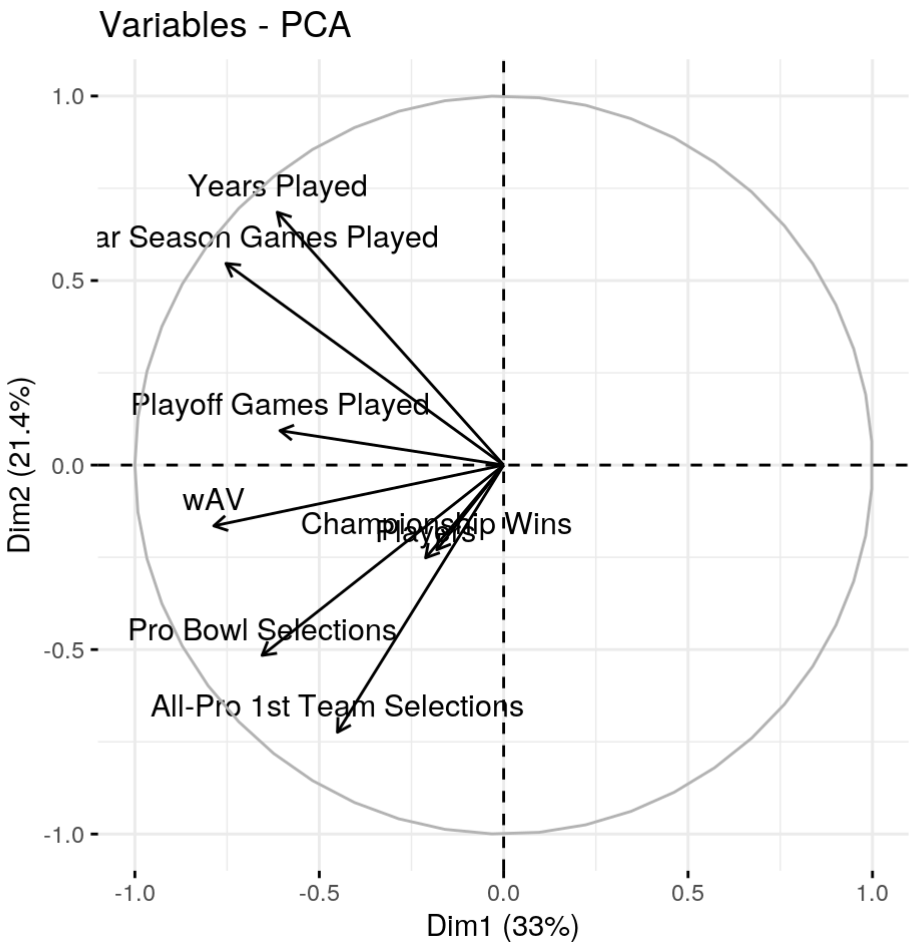
```
# Average silhouette width
pam_results$silinfo$avg.width
```

```
## [1] 0.3868916
```

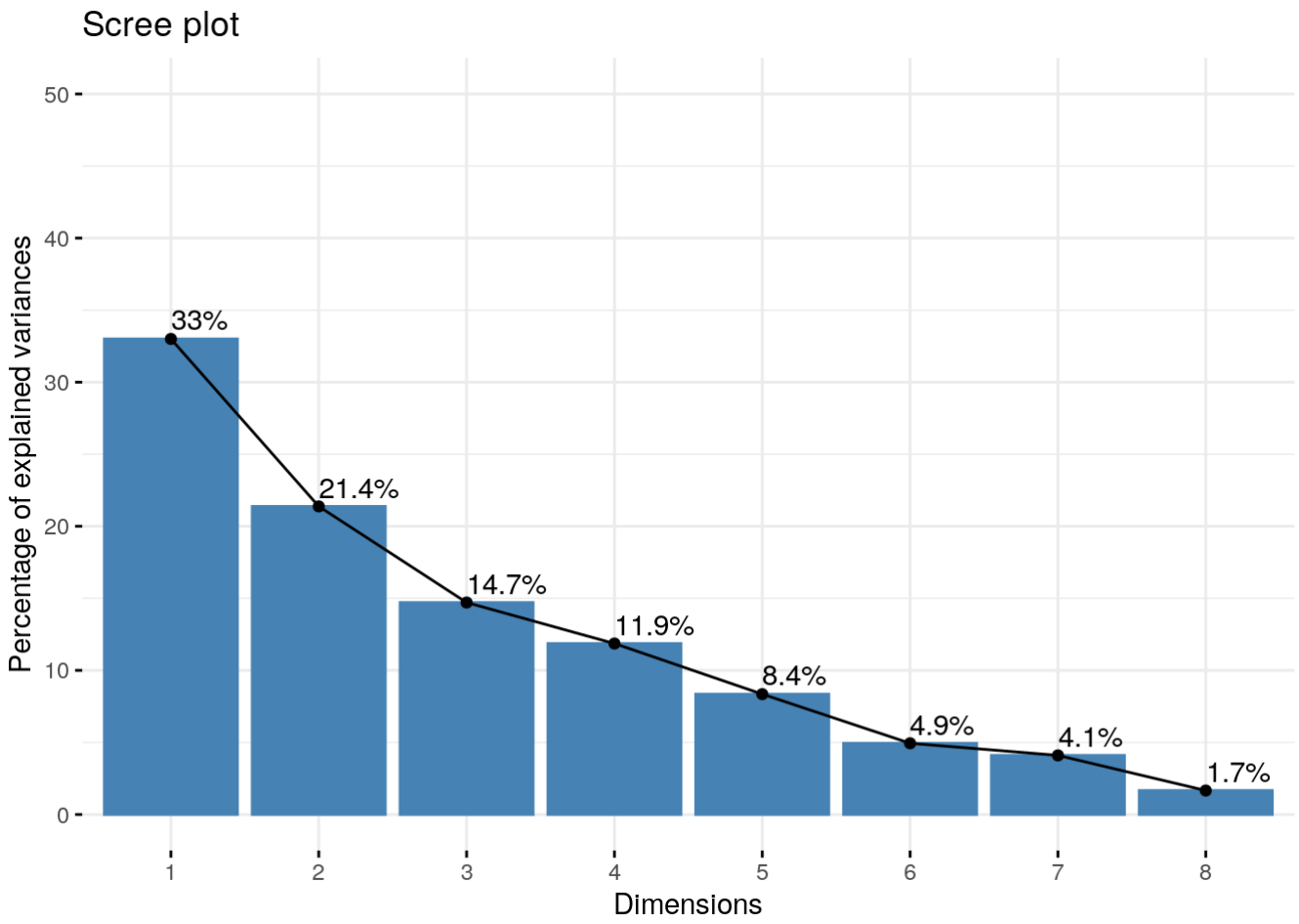
After applying PCA, we noticed that first two principal components cover a large percentage of the variation explained (61.3%), but the explained variation could be better with more variables/metrics.

Scoring higher on the first component indicates a high number of championship wins, while scoring lower indicates a high wAV. Scoring higher on the second component indicates a high number of years and games played, while scoring lower indicates a high number of All-Pro 1st Team and Pro Bowl selections.

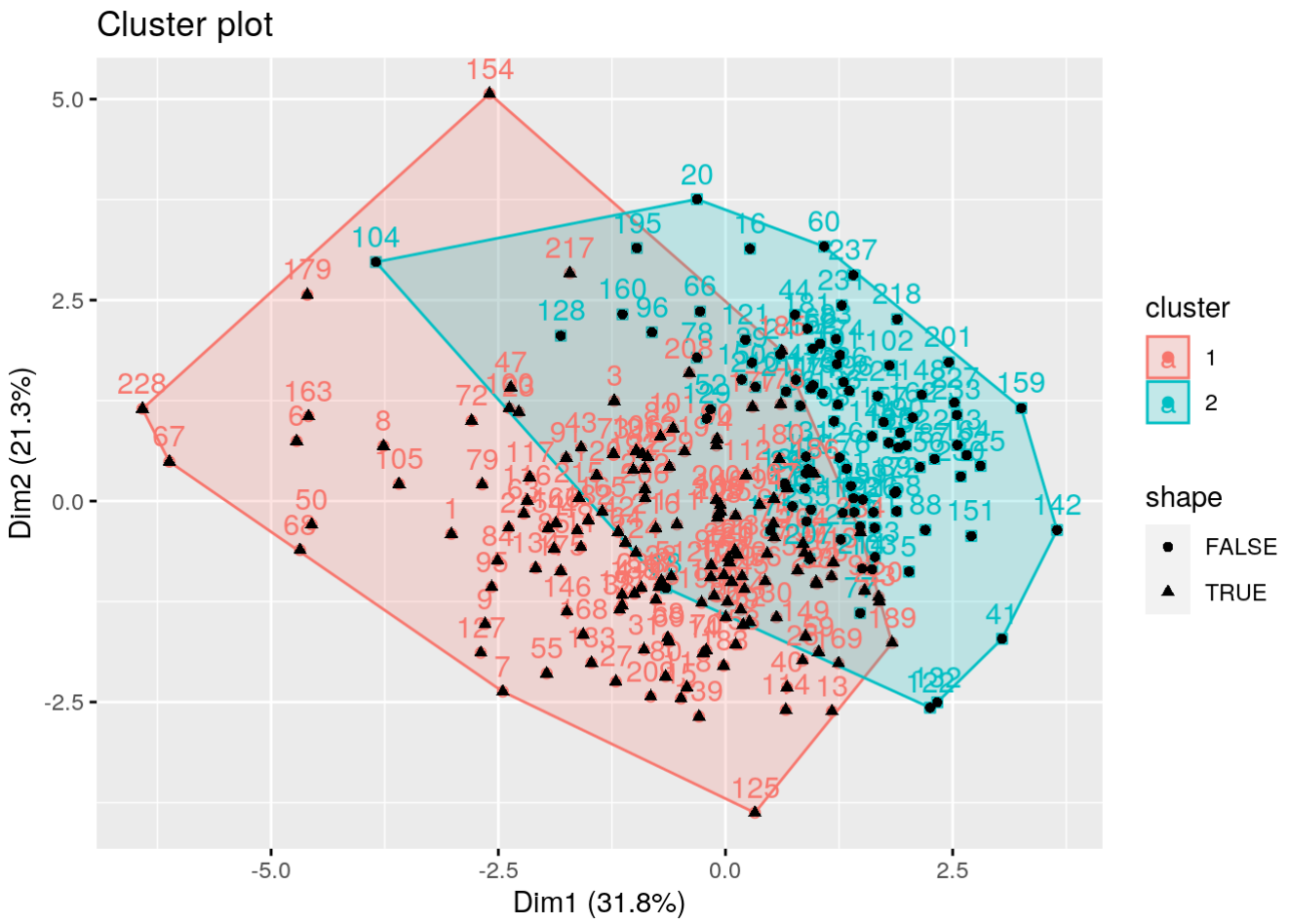
```
# Apply PCA to numeric variables only
pca <- hall_of_famers_reduced %>%
  select_if(is.numeric) %>%
  scale %>% # remember to scale
  prcomp
fviz_pca_var(pca) # Visualize contribution of elements to principal components
```




```
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 50)) # Percentage of variance explained for
 each PC in a scree plot
```




```
# Represent the clusters on PC1 and PC2
pam_results$data = hall_of_famers_reduced
fviz_cluster(pam_results, data = hall_of_famers_reduced,
             shape = hall_of_famers_reduced$`Hall of Famer`) +
  geom_point(aes(shape = hall_of_famers_reduced$`Hall of Famer`)) +
  guides(shape = guide_legend(title = "shape"))
```



Classification and Cross-Validation

After looking at how all of the variables and metrics correlate to Hall of Fame status, we can use classification to predict a binary variable based on other variables in our dataset. In this case, we will predict Hall of Fame status based on Championship Wins and Pro Bowl Selections.

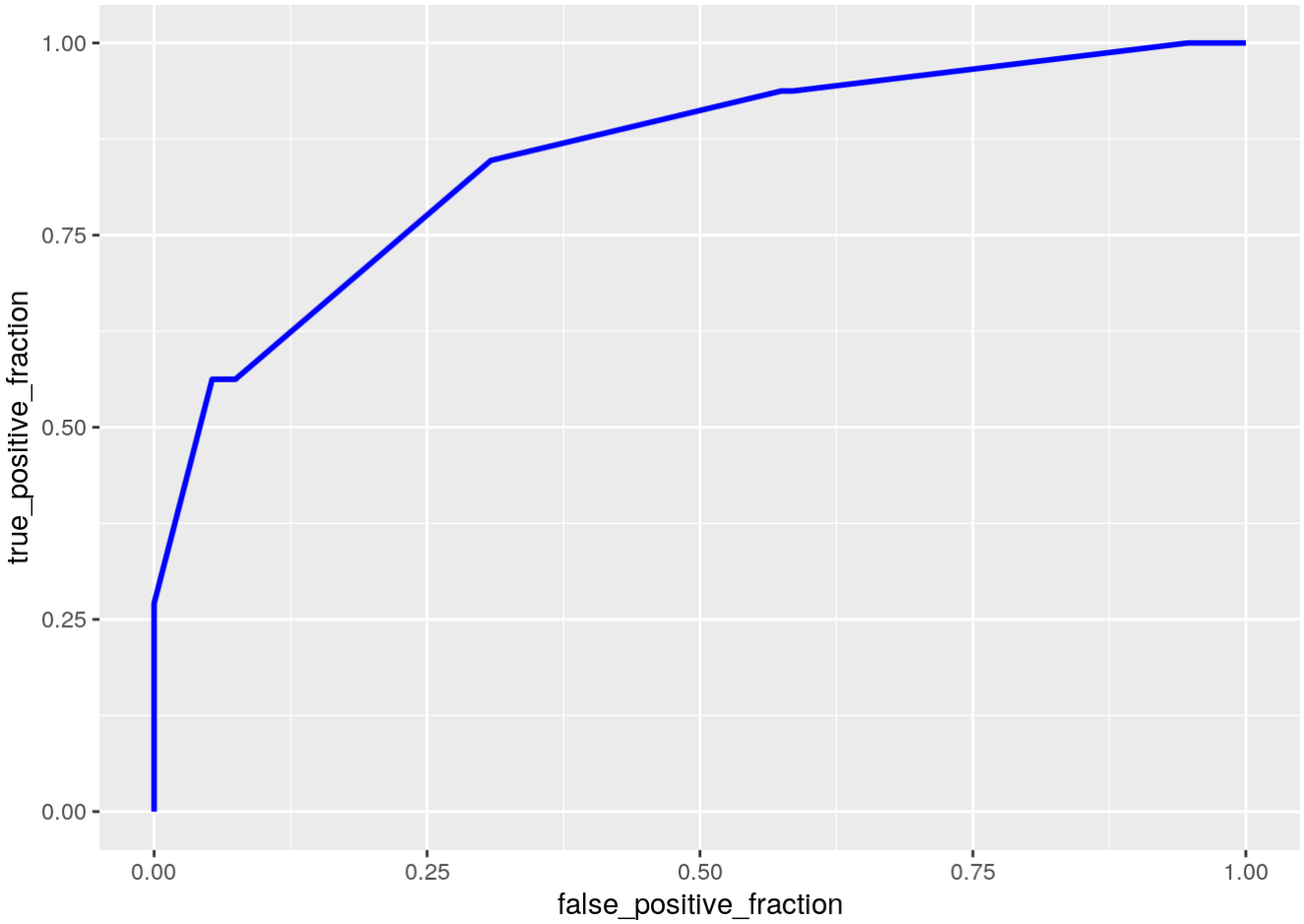
Based on the ROC curve, the variables we selected do not seem to be amazing predictors of Hall of Fame status, as there are likely other metrics that contribute to Hall of Fame induction. Interestingly, the AUC score came out to be roughly 0.85, which is a great classification score.

```
# Overwrite object in environment
hall_of_famers <- hall_of_famers %>%
  rename(hall_of_famer = `Hall of Famer`, # Rename variables
         pro_bowl_selections = `Pro Bowl Selections`,
         championship_wins = `Championship Wins`,
         years_played = `Years Played`,
         regular_season_games_played = `Regular Season Games Played`,
         all_pro_1st_team_selections = `All-Pro 1st Team Selections`,
         playoff_games_played = `Playoff Games Played`) %>%
  select(-Player, -School,-Conference,-Players,-`Total Hall of Famers`,-`Total Pro Bowl
 Selections`,-`Total Games Played`) %>% # Remove specific variables
  mutate(hall_of_famer = ifelse(hall_of_famer == TRUE, 1, 0)) # Rewrite variable values

# kNN with k = 5
hall_of_famers_kNN <- knn3(hall_of_famer ~ .,data = hall_of_famers, k = 5) # number of n
eighbors

# Save new object in environment
hall_of_famers_pred <- hall_of_famers %>%
  mutate(predictions_kNN = predict(hall_of_famers_kNN, hall_of_famers)[,2]) # Create new
predictions variable

# Create ggplot ROC curve
ROC <- ggplot(hall_of_famers_pred) +
  geom_roc(aes(d = hall_of_famer, m = predictions_kNN), color = "blue", n.cuts = 0)
ROC
```




```
calc_auc(ROC)$AUC # Calculate AUC performance
```

```
## [1] 0.8499926
```

Using k-fold cross-validation, we can validate our dataset's performance over multiple test sets. Fitting a kNN model and repeating for each k-fold, we obtained an average AUC performance of roughly 0.69, which is less than the AUC of the kNN model trained on the entire data (.85), indicating that the kNN model is a decent predictor a majority of the time. Thus, there are no significant signs of overfitting, which means the model will perform well in future scenarios with new datasets.

```
# Choose number of folds
k = 10

# Randomly order rows in the dataset
data <- hall_of_famers[sample(nrow(hall_of_famers)), ]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)

# Initialize a vector to keep track of the performance
perf_k <- NULL

# Use a for loop to get diagnostics for each test set
for(i in 1:k){
  # Create train and test sets
  train <- data[folds != i, ] # all observations except in fold i
  test <- data[folds == i, ]  # observations in fold i

  # Train model on train set (all but fold i)
  hall_of_famers_kNN <- knn3(hall_of_famer ~ .,
                data = train,
                k = 5) # number of neighbors

  # Test model on test set (fold i)
  df <- data.frame(
    predictions = predict(hall_of_famers_kNN, test)[,2],
    hall_of_famer = test$hall_of_famer)

  # Consider the ROC curve for the test dataset
  ROC <- ggplot(df) +
    geom_roc(aes(d = hall_of_famer, m = predictions))

  # Get diagnostics for fold i (AUC)
  perf_k[i] <- calc_auc(ROC)$AUC
}

# Average performance
mean(perf_k)
```

```
## [1] 0.6941405
```

Acknowledgements

Thank you to Phillip Kim for creating visualizations, summary statistics, clustering, and dimensionality reduction.

Thank you to Srikar Sriram for creating visualizations, classification, and cross-validation.