Analysis

Matt Johnson

10/5/2020

Introduction

The NFL Draft is one of the of the biggest sport drafts in the world. Teams will spend millions of dollars on players that they think will help their team win and reach the Superbowl. With all this money changing hands, comes a lot of risk and a lot of questions about what stats are important for deciding what players to draft. The NFL combine allows teams to see how players match up against each other in physical challenges such as the benchpress and the 40yd dash. Teams will use the results of these tests, along with college stats and performance, to decide if and when they should draft a player.

This project will shed light on what combine stats are most important for deciding when a player should be drafted dependent on the position of the player. Using predictive modeling techniques and machine learning, we will attempt to predict when a player will get drafted based on their performance in the combine and their position. Also, clustering methods will be used to asses the following hypothesis:

Hypothesis

Hypothesis: There are three types of players that enter the NFL Draft. Skill, Strength, and Mixed.

Explanation: A variable called "Type" has been added to denote the type of player. "Type" could be one of the following three options: Skill, Strength, Mixed. These columns were added as a hypothesis that these are the three types of players in the combine. Skill players are those that entered the combine as a WR, RB, S, CB. Strength are C, OG, OL, OT, DE, DT, DL. Mixed are LB, OLB, TE. These seemed fitting initially because of their relative positions on the field when the ball is snapped.

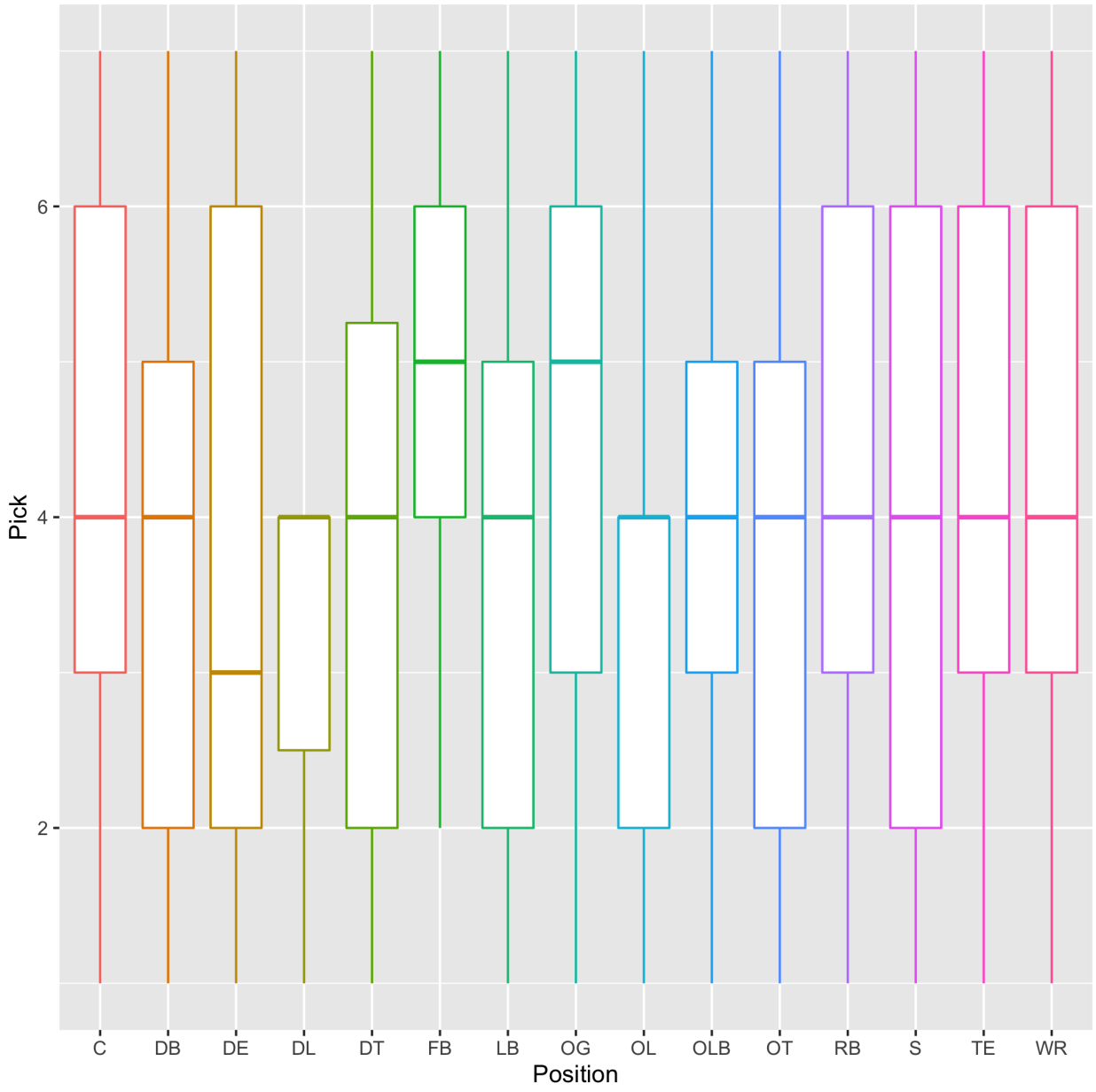
Data

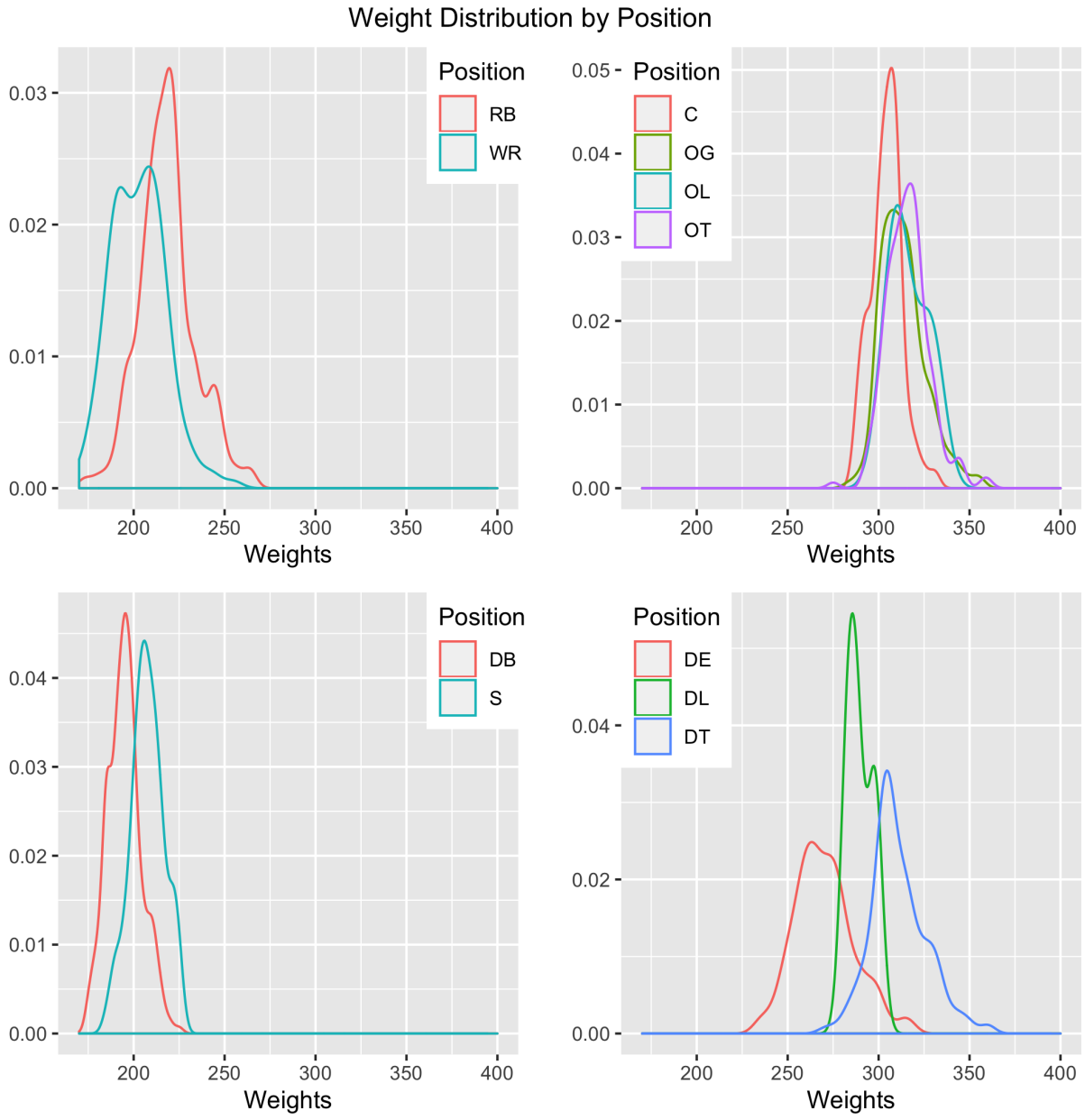
The two data sets used are combine.csv and draft.csv from: https://www.kaggle.com/toddsteussie/nfl-play-statistics-dataset-2004-to-present

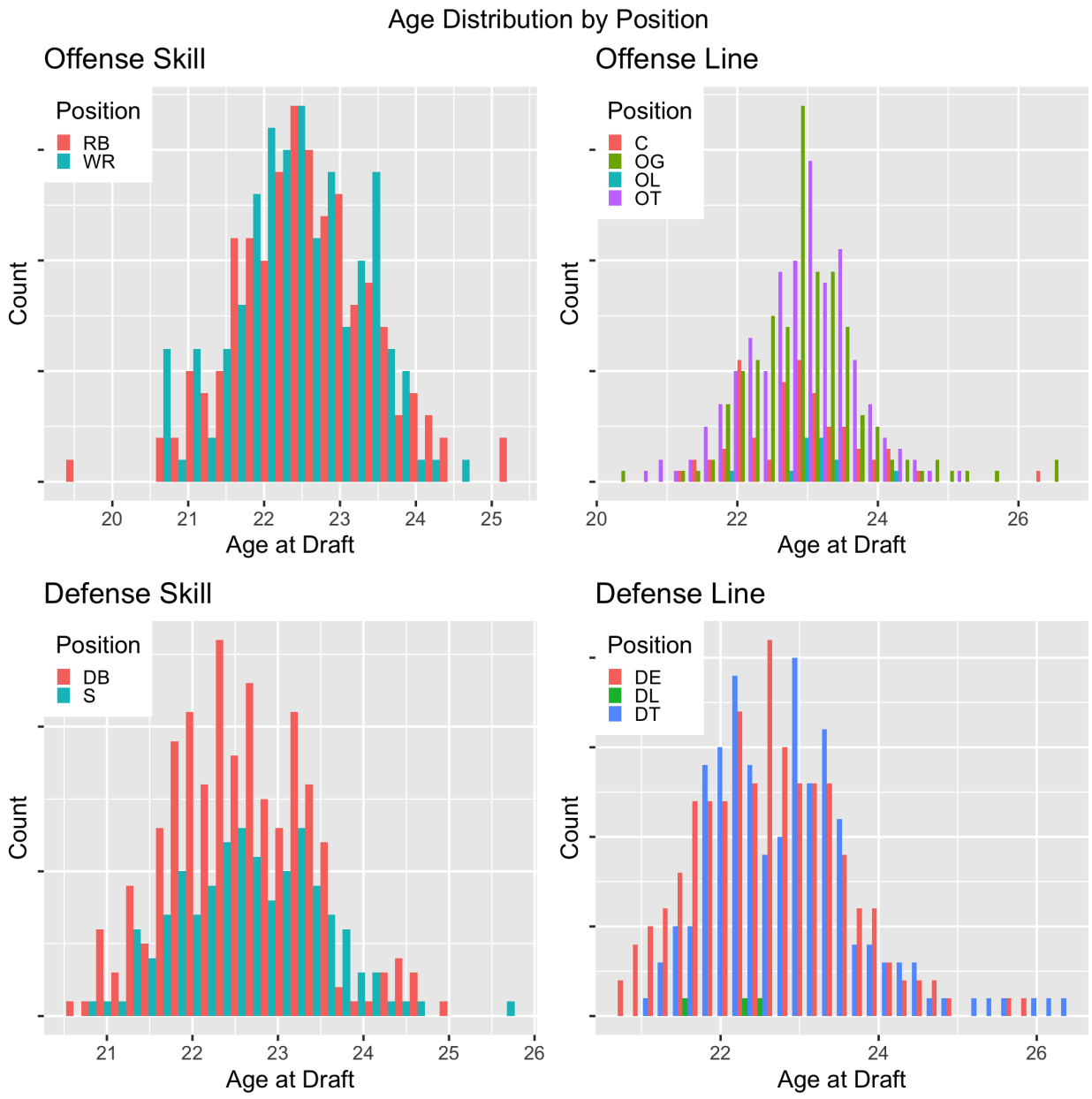
Data Cleaning and Joining

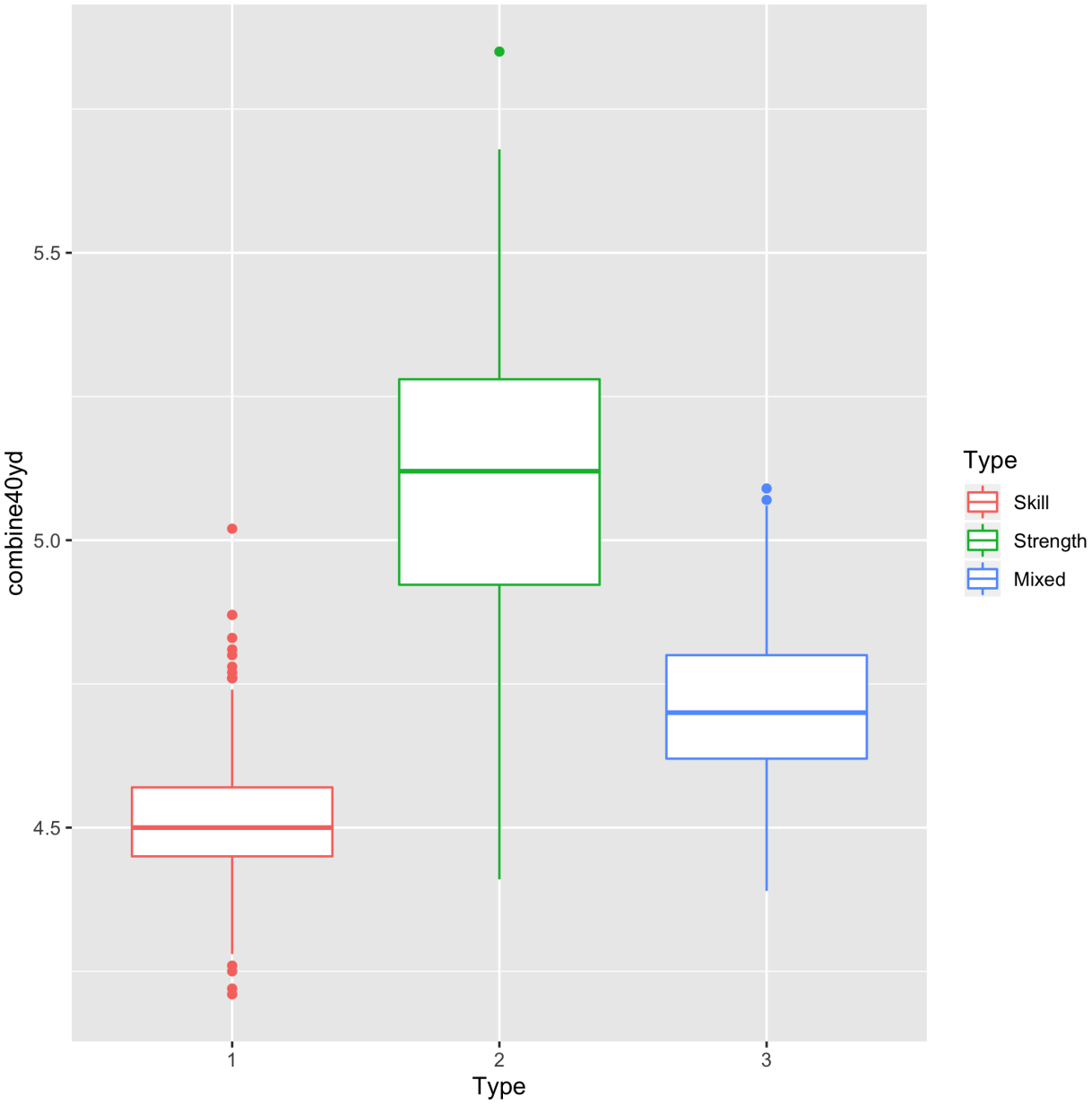
Only data from the years 2000-present are used for analysis. The NFL has changed significantly from the 1960s, this suggests only recent data is relevant for prediction. Data was split, tagged with "Type" column to test hypothesis previously stated, and rejoined to form prediction data set.

The following plots contain summary information for the data set.


Pick
6
4
2
C
DB
DE
DL
DT
FB
LB
OG
OL
OLB
OT
RB
S
TE
WR
Position


Weight Distribution by Position
Position
RB
WR
0.03
0.02
0.01
0.00
200
250
300
350
400
Weights
Position
C
OG
OL
OT
0.05
0.04
0.03
0.02
0.01
0.00
200
250
300
350
400
Weights
Position
DB
S
0.04
0.03
0.02
0.01
0.00
200
250
300
350
400
Weights
Position
DE
DL
DT
0.04
0.02
0.00
200
250
300
350
400
Weights


Age Distribution by Position
Offense Skill
Position
RB
WR
Count
20
21
22
23
24
25
Age at Draft
Offense Line
Position
C
OG
OL
OT
Count
20
22
24
26
Age at Draft
Defense Skill
Position
DB
S
Count
21
22
23
24
25
26
Age at Draft
Defense Line
Position
DE
DL
DT
Count
22
24
26
Age at Draft

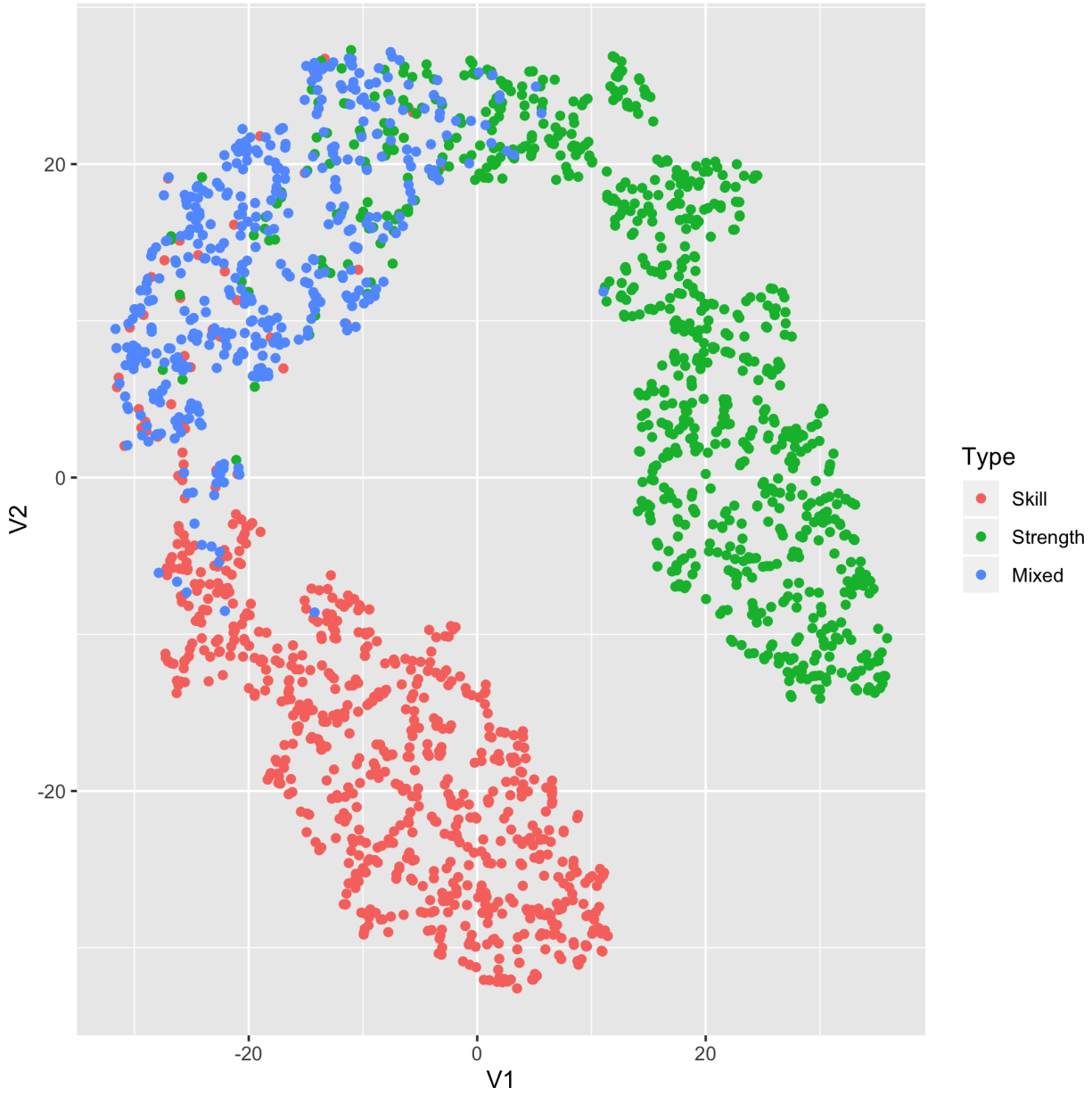


Methodology

Clustering

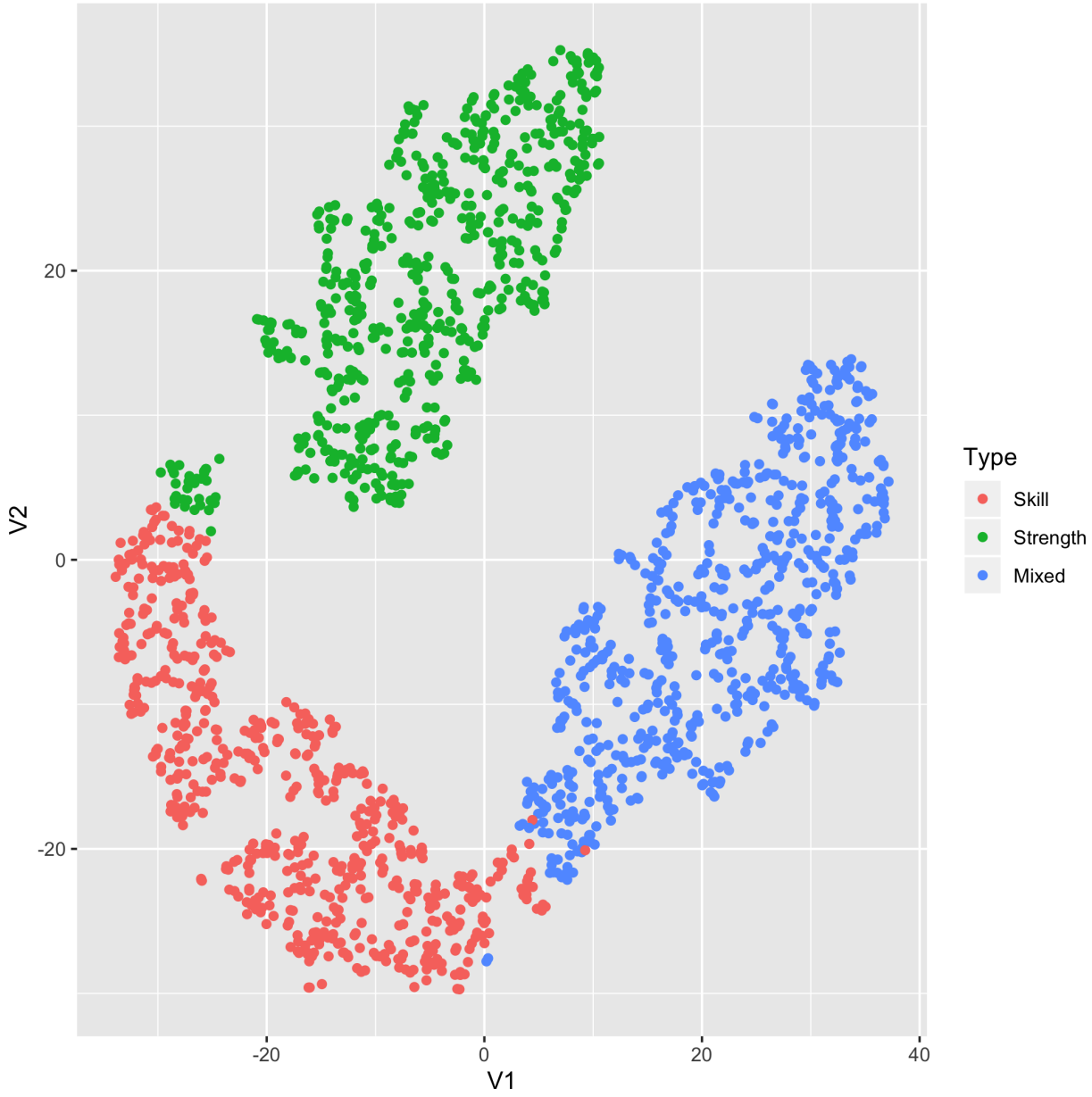
After suspecting there were three distinct groups, We tested this hypothesis with three clustering methods: TSNE, K-Means, and Principle Component Analysis (PCA).

TSNE



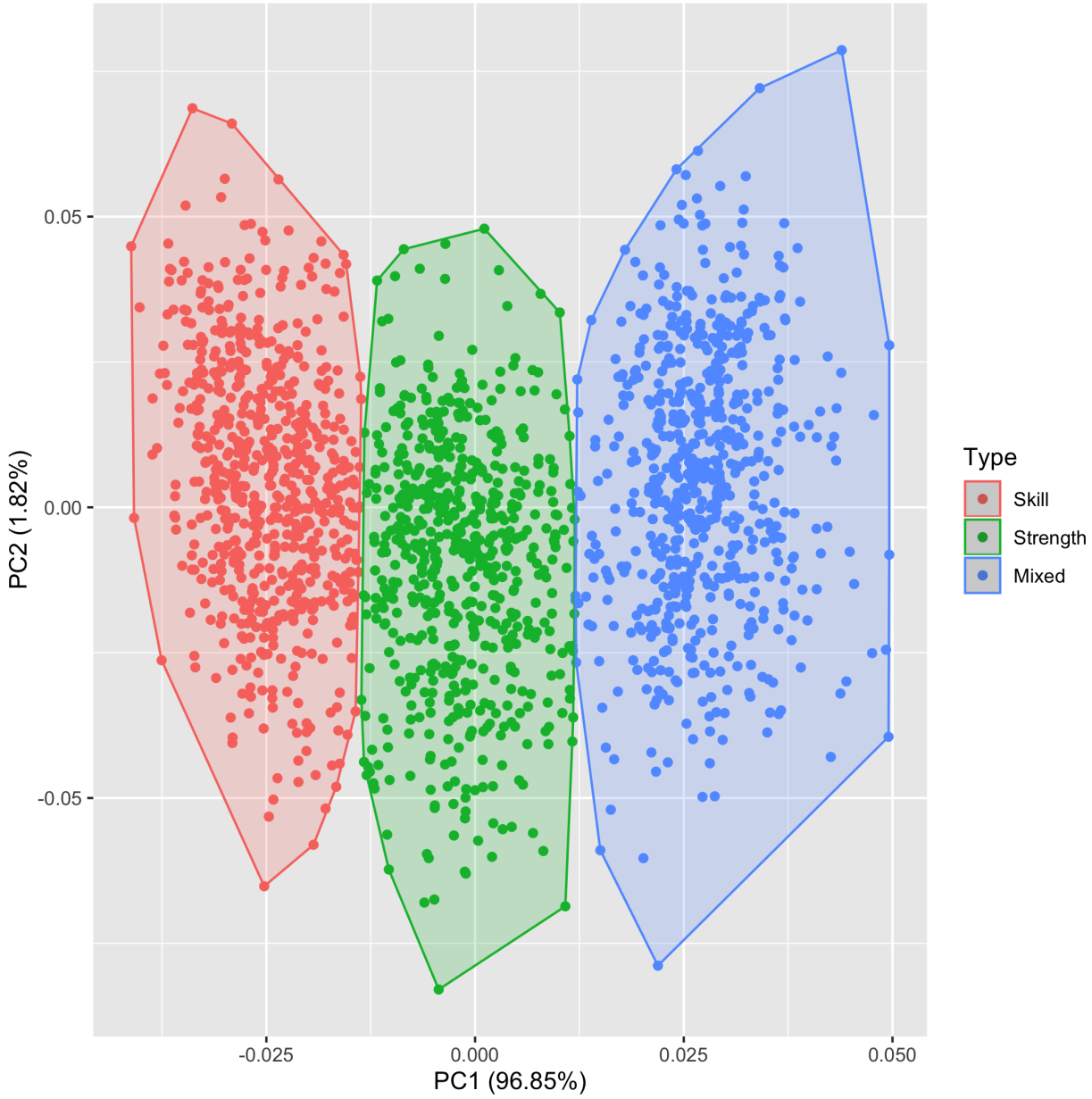
TSNE provided intial insight into the suspected clusters in the data. Since TSNE is non-linear, it could not be said for sure that these three groups truly existed.

K-Means



K-Means provides evidence that these clusters truly exist. There is little overlap between the three groups and there are clear centers in each of the clusters.

PCA



The PCA further justifies accpeting the hypothesis that there are indeed three groups of players in the NFL Draft. This plot shows the primary principle component plotted against the secondary principle component.

Modeling

We attempt to predict pick from the set of combine results as well as the height, weight, and age of the player. This basic full regression formula is as follows:

Pick ~ Weight + Height + Age + combine40yd + combineBench + combineVert + combine3cone + combineShuttle + combineBroad + ϵ.

This is done using multiple different regression techniques. Below is a table displaying the results of the different models. Models are assessed using RMSE and Rsquared of predicitons on randomly selected

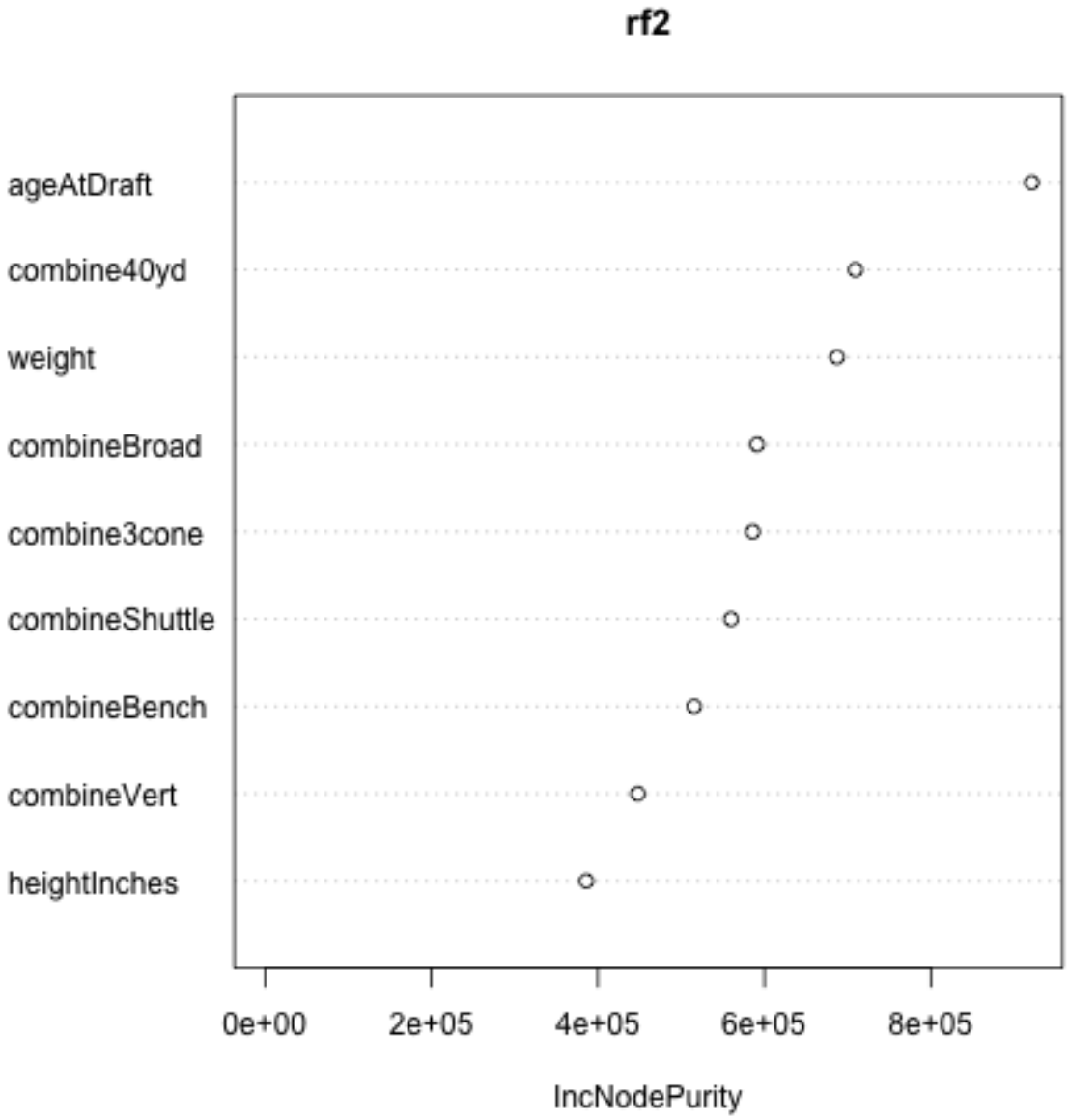
validation set.

Model	Rsquare	RMSE
Random Forest	0.5878970	45.02167
GBM	0.3528746	56.41738
Ridge	0.1476620	64.74774
Lasso	0.1464074	64.79537
PCR	0.1193670	65.81368
GLM	0.1100707	66.16014
LM	-2.0619531	122.72073

According to the table, the random forest model seemed to perfrom the best on the out of sample validation set. The importance of each variable in the random forest model is detailed below.

	IncNodePurity
heightInches	386189.7
weight	687344.9
ageAtDraft	921406.6
combineShuttle	560046.5
combineBroad	591326.8
combine3cone	586102.7
combineBench	515635.2
combine40yd	709496.8
combineVert	448071.6

Variance Importance Plot



According to the plot above, the age of the player seems to be quite significant in predicting where they will get drafted. This is an unexpected results and should be analysed in future experiments. Height having the least importance is a reasonable result considering the wide range of players that get drafted. Running-backs are often quite short, but still are valuable players that get drafted early. On the other hand, lineman are often tall and can just as likely get drafted early.

Conclusion

Results

Clustering

According to the clustering analysis, we have established three groups of players that enter the NFL draft and participate in the combine: Skill, Strength, and Mixed. Intuitvely, this division makes perfect sense considering where players lineup on the field. The stronger players that perform better in strength events all lineup right on the line of scrimmage. Skill position players that are agile and quicker lineup as recievers or in the backfield. Mixed players do both, tight ends can lineup as a blocker or a receiver and must be both strong and skilled. These three clustering analyses justify the conclusion that there are three groups of players.

Modeling

From the data used in this experiment, it does not seem reasonable to use this data to predict when a player will be selected in the nfl draft. The models built do not seem to have strong enough predicting power to make reasonable predictions.

Future Work

Future work should include data about how players performed in college. Statistics reasonable for each position should be included as aggregates over college careers. Picks in the draft can be unpredicitable because different teams have different needs. Future work should include a "blueprint" statistics to explain what type of player each NFL team needs. This could be as simple an an explanatory eigenvector describing the team in a vectorized way. Quarterbacks were removed from this analysis becasue of lack of data, future research could include QBs and how their performance in the combine affects draft results. Many QBs do not attend the combine, future work could assess if this helps or hurts draft stock.