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.

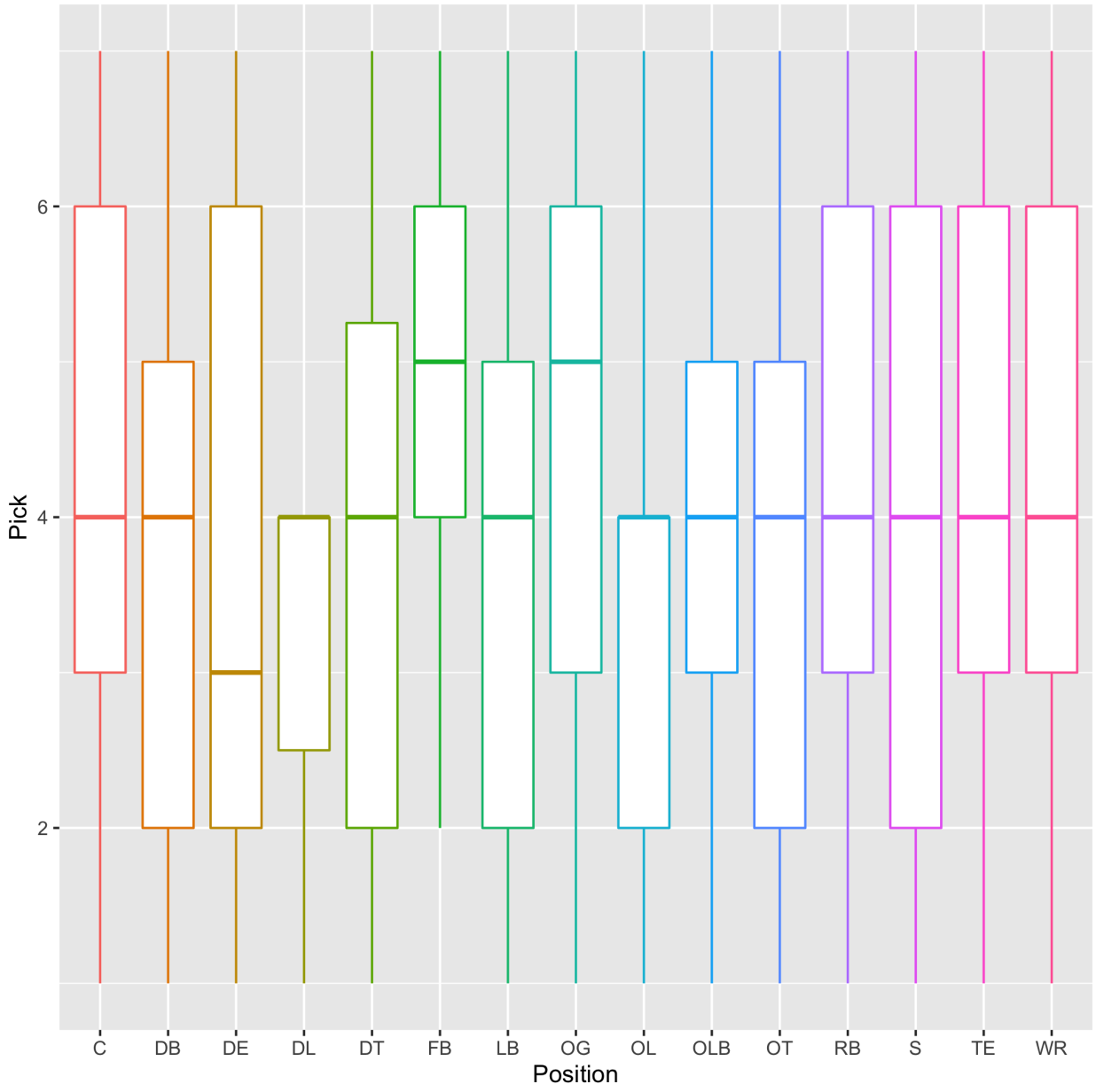
Data

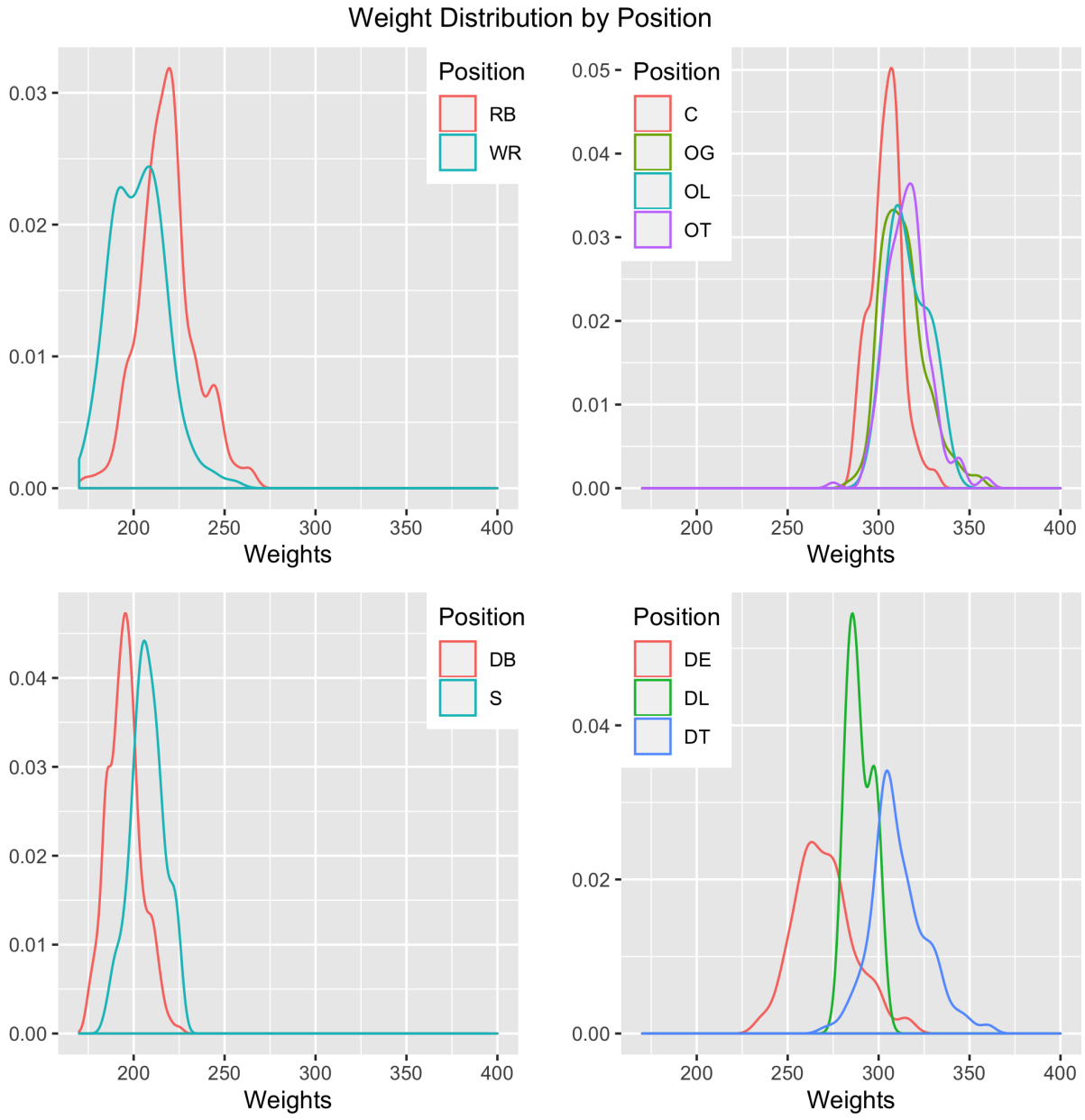
The two data sets used are combine.csv and draft.csv from: https://www.kaggle.com/toddsteussie/nfl-play-statistics-dataset-2004-to-present

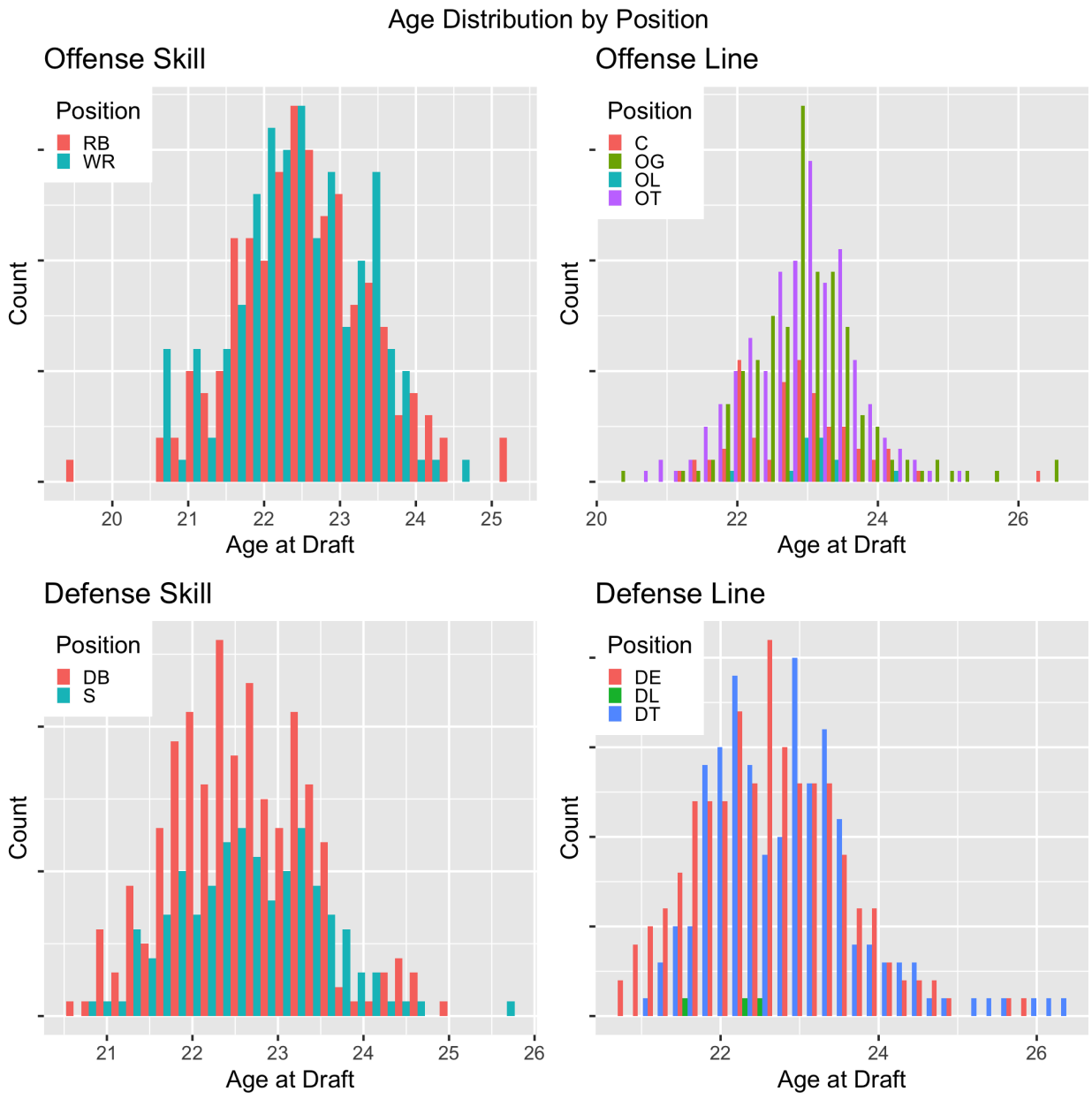
Data Cleaning and Joining

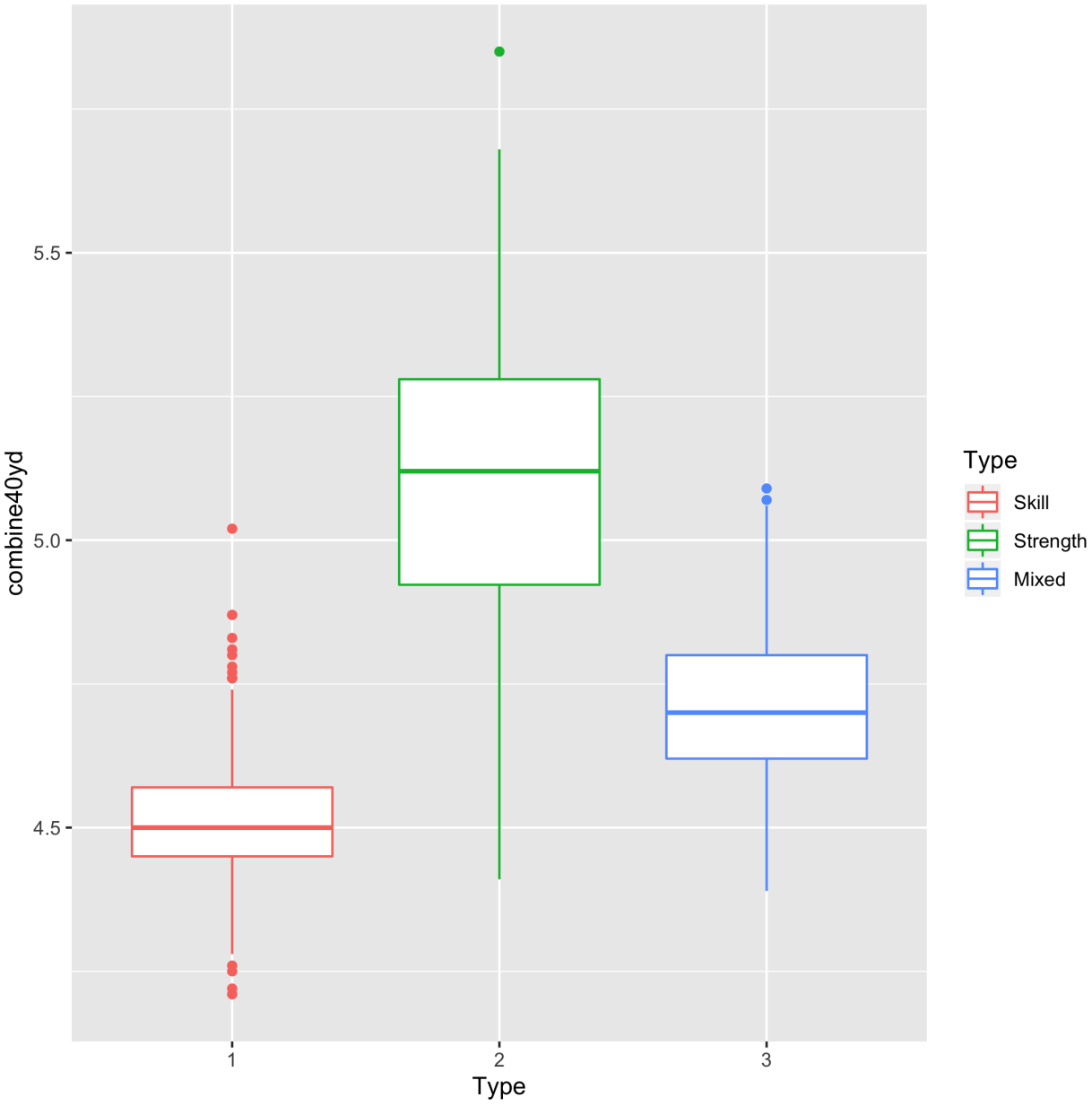
Only data from the years 2000-present are used for analysis. The NFL has changed significantly from the 1960s, this suggests only recent data is relevant for prediction. Data was split, tagged with "Type" column to test hypothesis previously stated, and rejoined to form prediction data set.

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

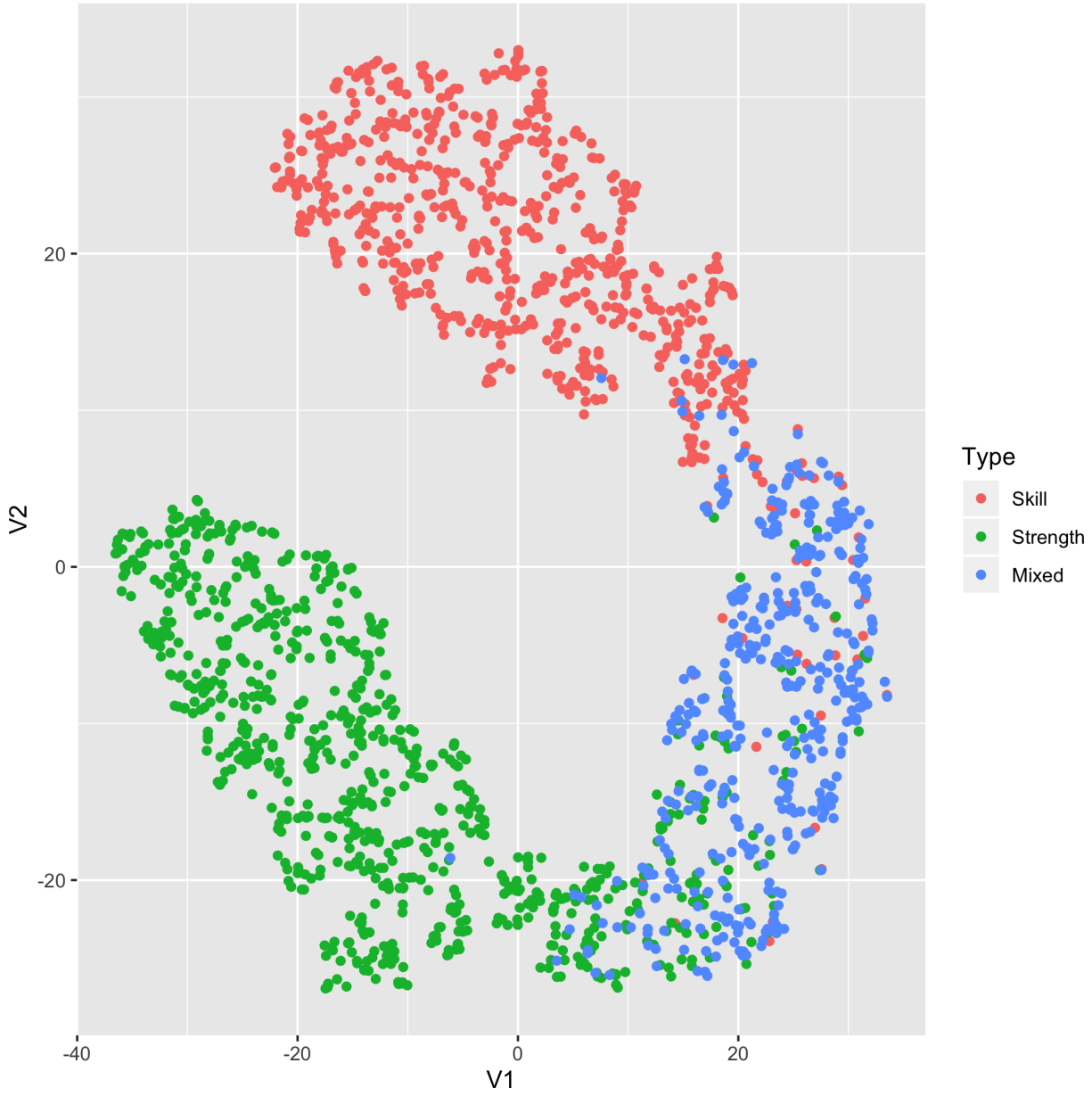


Methodology

Clustering

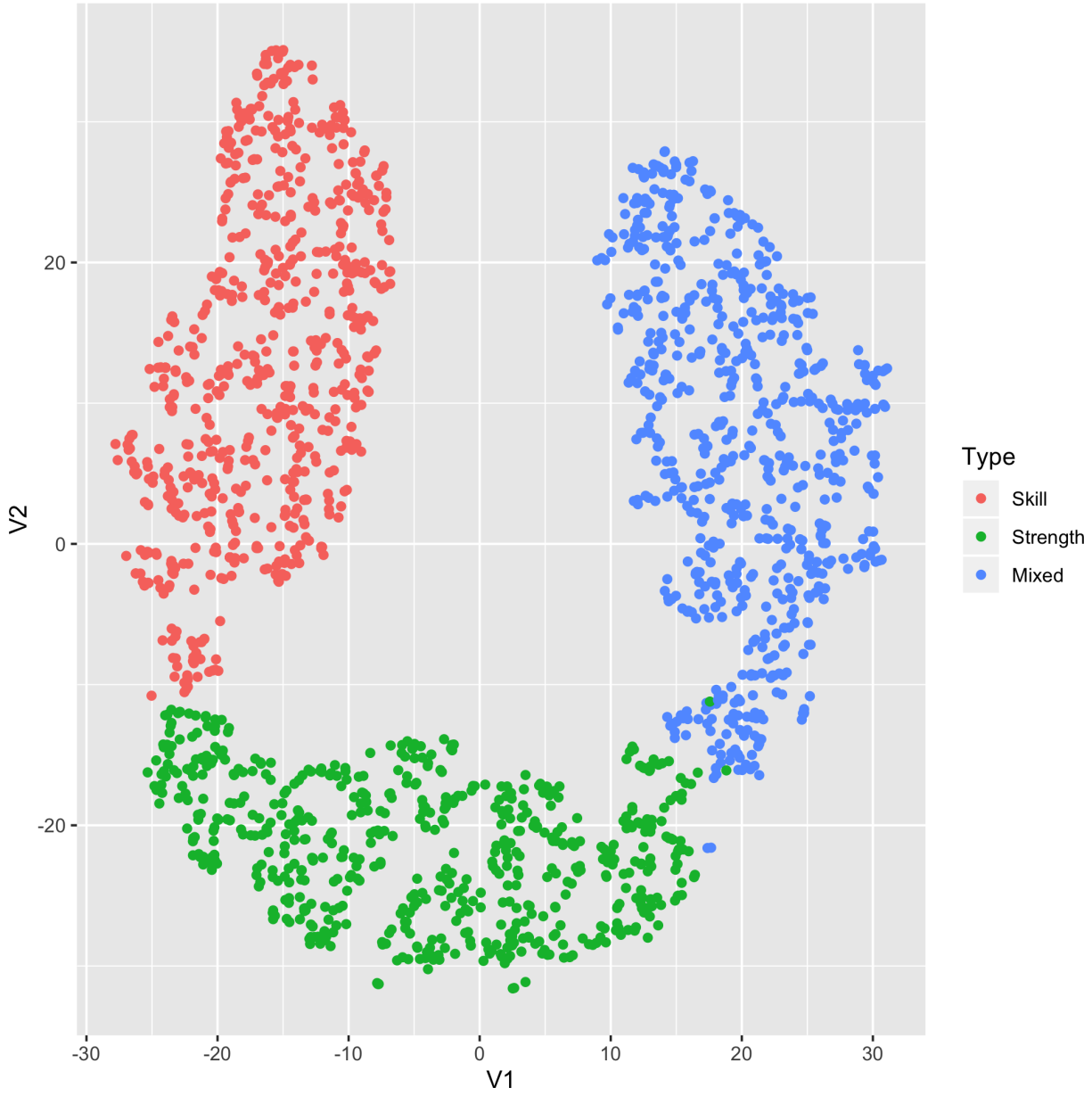
After suspecting there were three distinct groups, I tested this hypothesis with three clustering methods: TSNE, K-Means, and Principle Component Analysis (PCA).

TSNE



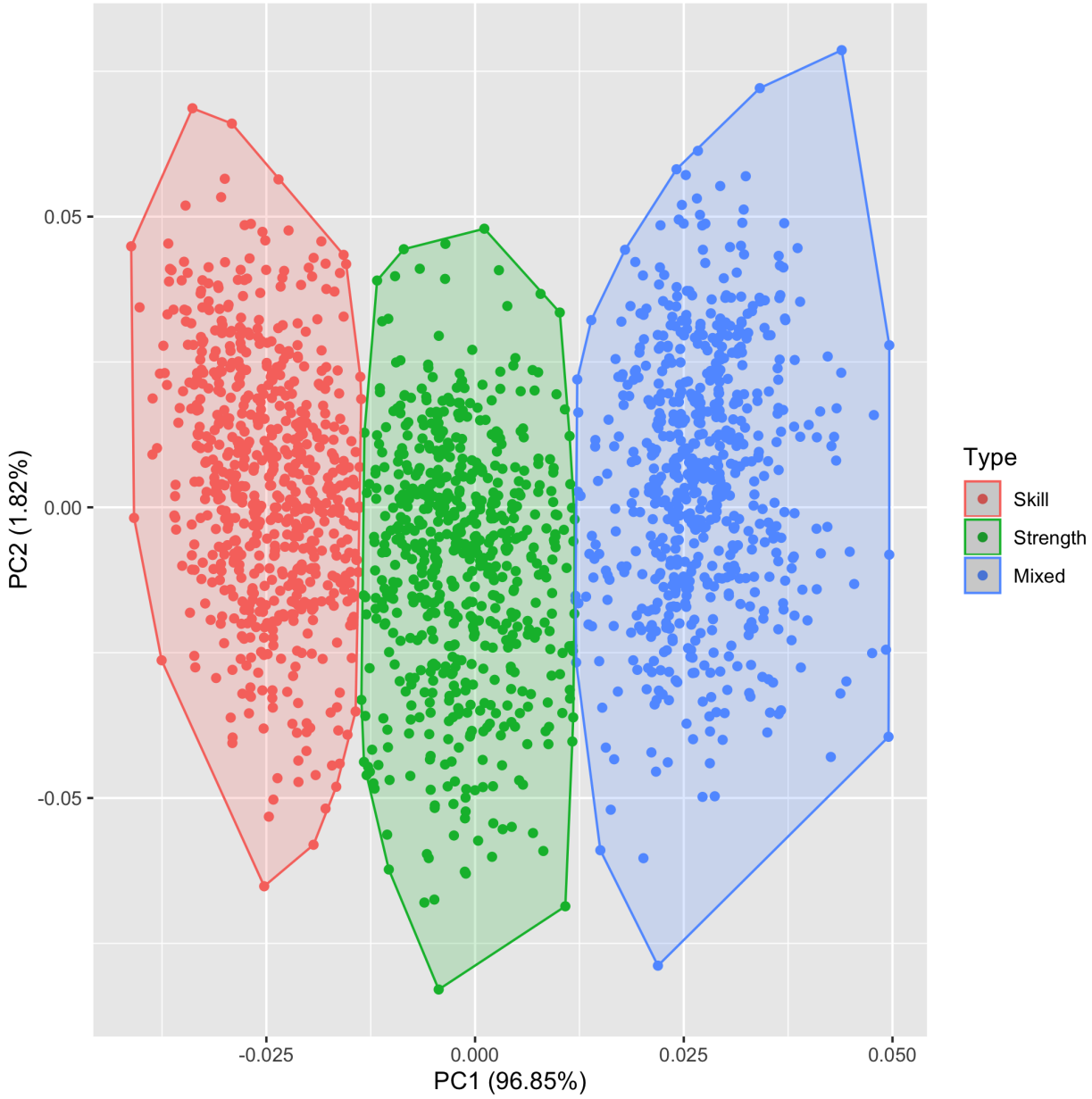
TSNE provided intial insight into the suspected clusters in the data. Since TSNE is non-linear, it could not be said for sure that these three groups truly existed.

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.

Clustering Analysis Conclusion

These three clustering analyses justify the conclusion that there are three groups of players. Intutitively it makes perfect sense as well. Running-back, Receiver, Safety, and cornerback are all positions known for speed and agility. These players would perfrom especially well in the 40yd dash, 3cone, and shuttle run drills during the combine. On the other hand, lineman on offense and defense need to be exceedingly strong and thus perform well at the benchpress. Also, since weight and height are assesed at the combine they are used to build these clusters.

Modeling

We attempt to predict pick from the set of combine times 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 model. Models are assessed using RMSE and Rsquared of predicitons on validation set.

Model	Rsquare	RMSE
Random Forest	0.5309678	46.49559
GBM	0.2876712	57.29948
Lasso	0.0867025	64.88085
Ridge	0.0856626	64.91778
PCR	0.0571834	65.92104
GLM	0.0562307	65.95434
LM	-1.6816115	111.17531

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	361955.1
weight	677694.2
ageAtDraft	910287.8
combineShuttle	564250.3
combineBroad	533554.2
combine3cone	621103.4
combineBench	525469.4
combine40yd	672784.0
combineVert	471443.3