Final Project

INFO 523

Meghan Edgerton

2023-04-21

```r
library(dplyr)
library(readxl)
library(rgl)
library(adamethods)
library(ggplot2)
library(outliers)
library(DMwR2)
library(Rlof)
library(knitr)

p_path <- "data/PassingStats.xlsx"
re_path <- "data/ReceivingStats.xlsx"
ru_path <- "data/RushingStats.xlsx"

# Datasets
passing <- read_excel(p_path)
receiving <- read_excel(re_path)
rushing <- read_excel(ru_path)
```

This example demonstrates the use of the Local Outlier Factor (LOF) Algorithm for Outlier Detection on college football player data from the 2022 NCAA season. Passing, receiving, and rushing data was downloaded from www.sports-reference.com. The purpose of this application is to find outliers within a big population of potential football recruits and mimics a scenario of a football coach wanting to know which quarterbacks, running backs, and wide receivers stand out from the rest of the draft class.

```r
# Local Outlier Factor (Passing)

# Removing columns that aren't applicable to passing
p <- passing %>% select(`Passing Yards`, `Pass Completions`)

outlier.scores <- lofactor(p, k=10)

outliers <- order(outlier.scores, decreasing=T)[1:10]

p.outliers <- passing[outliers,]

best.qb <- arrange(p.outliers, desc(`Pass Completions`), desc(`Passing Yards`))

top5.qb <- head(best.qb, 5)
top5.qb.f <- top5.qb %>% select(Rank, Player, School, Conference,
                                `Pass Completions`,
                               `Pass Completion Percentage`, `Passing Yards`,
                               `Passing TDs`)

kable(top5.qb.f, caption = "Outliers Detected in 2022 Passing Data")
```

Table 1: Outliers Detected in 2022 Passing Data

Rank	Player	School	Conference	Pass Completions	Pass Completion Percentage	Passing Yards	Passing TDs
41	Austin Reed*	Western Kentucky	CUSA	389	64.5	4746	40
70	Kyle Vantrease*	Georgia Southern	Sun Belt	370	61.4	4247	27
24	Michael Penix Jr.*	Washington	Pac-12	362	65.3	4641	31
25	Drake Maye*	North Carolina	ACC	342	66.2	4321	38
5	Caleb Williams*	USC	Pac-12	333	66.6	4537	42

```r
ggplot() +
  geom_point(data = passing, aes(x = `Pass Completions`,
                                 y = `Passing Yards`)) +
  geom_point(data = top5.qb, aes(x = `Pass Completions`,
                                 y = `Passing Yards`), colour = "blue",
                                 size = 4, shape = 18) +
  scale_y_continuous(n.breaks = 5, limits = c(1000,5000)) +
  theme_bw()
```

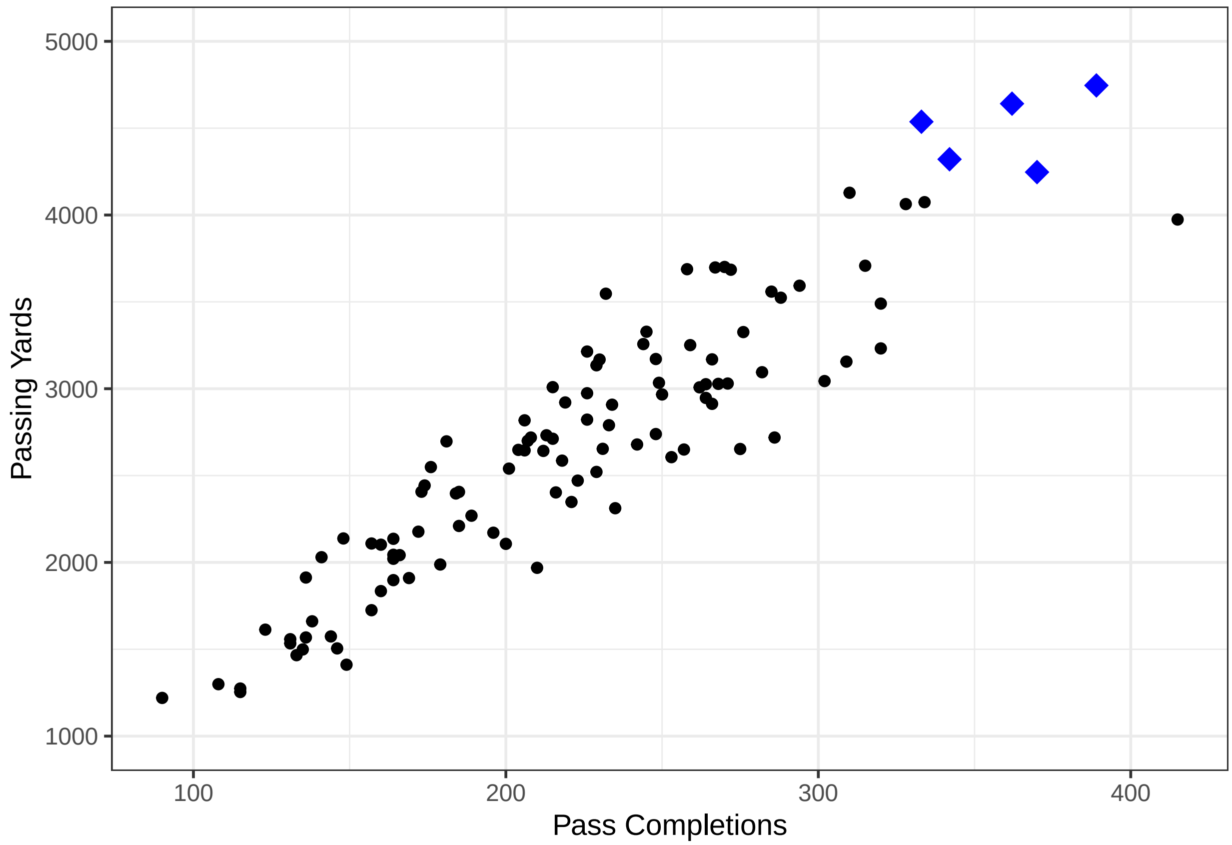

Figure 1 A scatterplot displaying the number of pass completions and the total passing yards for NCAA college quarterbacks from the 2022 season. The blue points represent the outliers detected using a local outlier factor (LOF) algorithm on pass completions and passing yards.

```r
# Local Outlier Factor (Receiving)

# Removing columns that aren't applicable to receiving
re <- receiving %>% select(Receptions, `Receiving Yards`)

outlier.scores2 <- lofactor(re, k=10)

outliers2 <- order(outlier.scores2, decreasing=T)[1:10]

re.outliers <- receiving[outliers2,]

best.rec <- arrange(re.outliers, desc(`Receptions`), desc(`Receiving Yards`))

top5.rec <- head(best.rec, 5)
top5.rec.f <- top5.rec %>% select(Rank, Player,School, Conference, Receptions,
                                  `Receiving Yards`, `Receiving TDs`,
                                  `Avg Receiving Yards Per Reception`)

kable(top5.rec.f, caption = "Outliers Detected in 2022 Receiving Data")
```

Table 2: Outliers Detected in 2022 Receiving Data

Rank	Player	School	Conference	Receptions	Receiving Yards	Receiving TDs	Avg Receiving Yards Per Reception
234	Charlie Jones*	Purdue	Big Ten	110	1361	12	12.4
200	Nathaniel Dell*	Houston	American	109	1398	17	12.8
201	Malachi Corley*	Western Kentucky	CUSA	101	1295	11	12.8
142	Rashee Rice*	SMU	American	96	1355	10	14.1
51	Marvin Harrison Jr.*	Ohio State	Big Ten	77	1263	14	16.4

```
ggplot() +
  geom_point(data = receiving, aes(x = `Receptions`,
                                   y = `Receiving Yards`)) +
  geom_point(data = top5.rec, aes(x = `Receptions`,
                                   y = `Receiving Yards`), colour = "red",
             size = 4, shape = 18) +
  scale_y_continuous(n.breaks = 5, limits = c(0,1500)) +
  theme_bw()
```

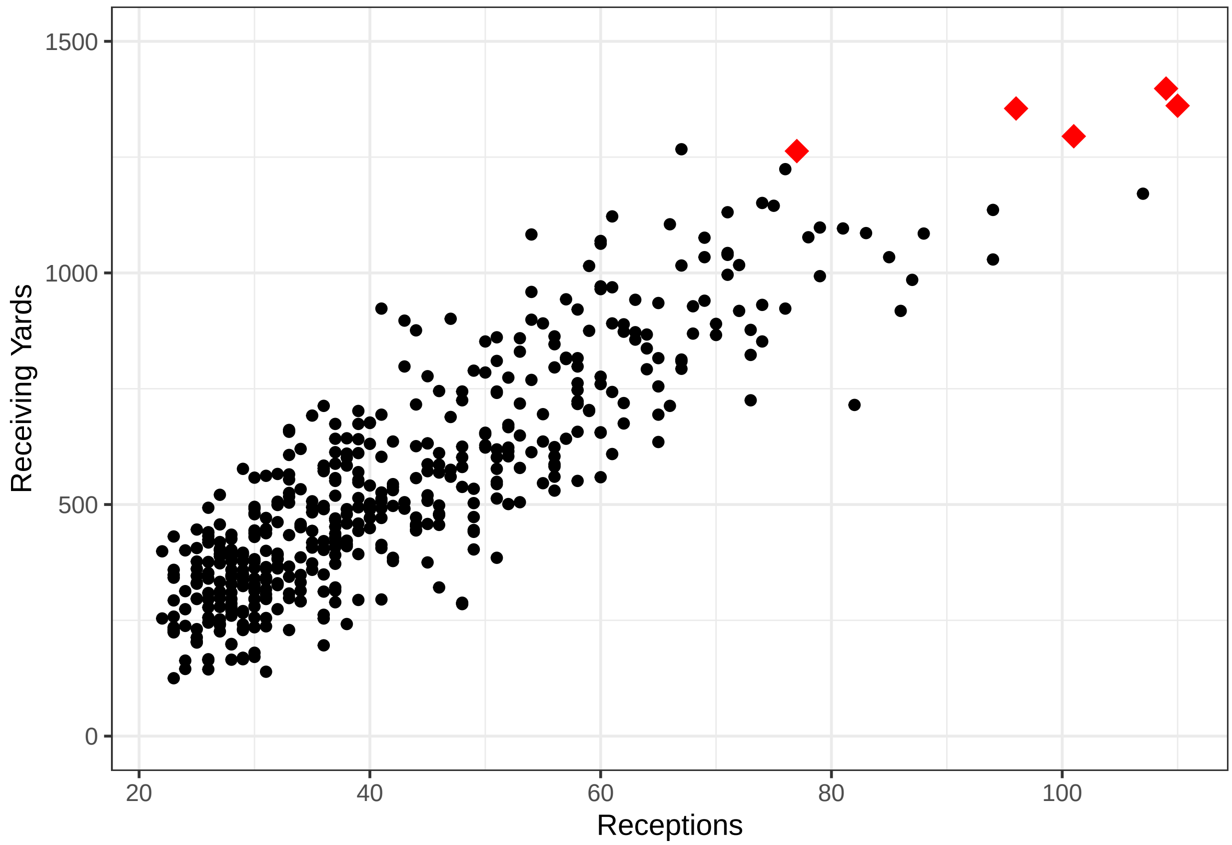

Figure 2 A scatterplot displaying the number of receptions and the total receiving yards for NCAA college wide receivers from the 2022 season. The red points represent the outliers detected using a local outlier factor (LOF) algorithm on total receptions and receiving yards.

```r
# Local Outlier Factor (Rushing)

# Removing columns that aren't applicable to rushing
ru <- rushing %>% select(`Rush Attempts`, `Avg Rushing Yards Per Attempt`) %>%
  filter(`Avg Rushing Yards Per Attempt` > 1.0)

outlier.scores3 <- lofactor(ru, k=10)

outliers3 <- order(outlier.scores3, decreasing=T)[1:10]

ru.outliers <- rushing[outliers3,]

best.rush <- arrange(ru.outliers, desc(`Rushing Yards`), desc(`Rush Attempts`))

top5.rush <- head(best.rush, 5)
top5.rush.f <- top5.rush %>% select(Rank, Player, School, Conference, `Rush Attempts`,
                                    `Rushing Yards`, `Rushing TDs`,
                                    `Avg Rushing Yards Per Attempt`)

kable(top5.rush.f, caption = "Outliers Detected in 2022 Rushing Data")
```

Table 3: Outliers Detected in 2022 Rushing Data

Rank	Player	School	Conference	Rush Attempts	Rushing Yards	Rushing TDs	Avg Rushing Yards Per Attempt
110	Brad Roberts*	Air Force	MWC	345	1728	17	5.0
95	Mohamed Ibrahim*	Minnesota	Big Ten	320	1665	20	5.2
111	Chase Brown*	Illinois	Big Ten	328	1643	10	5.0
31	Bijan Robinson*	Texas	Big 12	258	1580	18	6.1
44	Blake Corum*	Michigan	Big Ten	247	1463	18	5.9

```
ggplot() +
  geom_point(data = rushing, aes(x = `Rush Attempts`,
                                  y = `Rushing Yards`)) +
  geom_point(data = top5.rush, aes(x = `Rush Attempts`,
                                 y = `Rushing Yards`), colour = "green",
             size = 4, shape = 18) +
  scale_y_continuous(n.breaks = 6, limits = c(0,1800)) +
  scale_x_continuous(n.breaks = 5, limits = c(50,350)) +
  theme_bw()
```

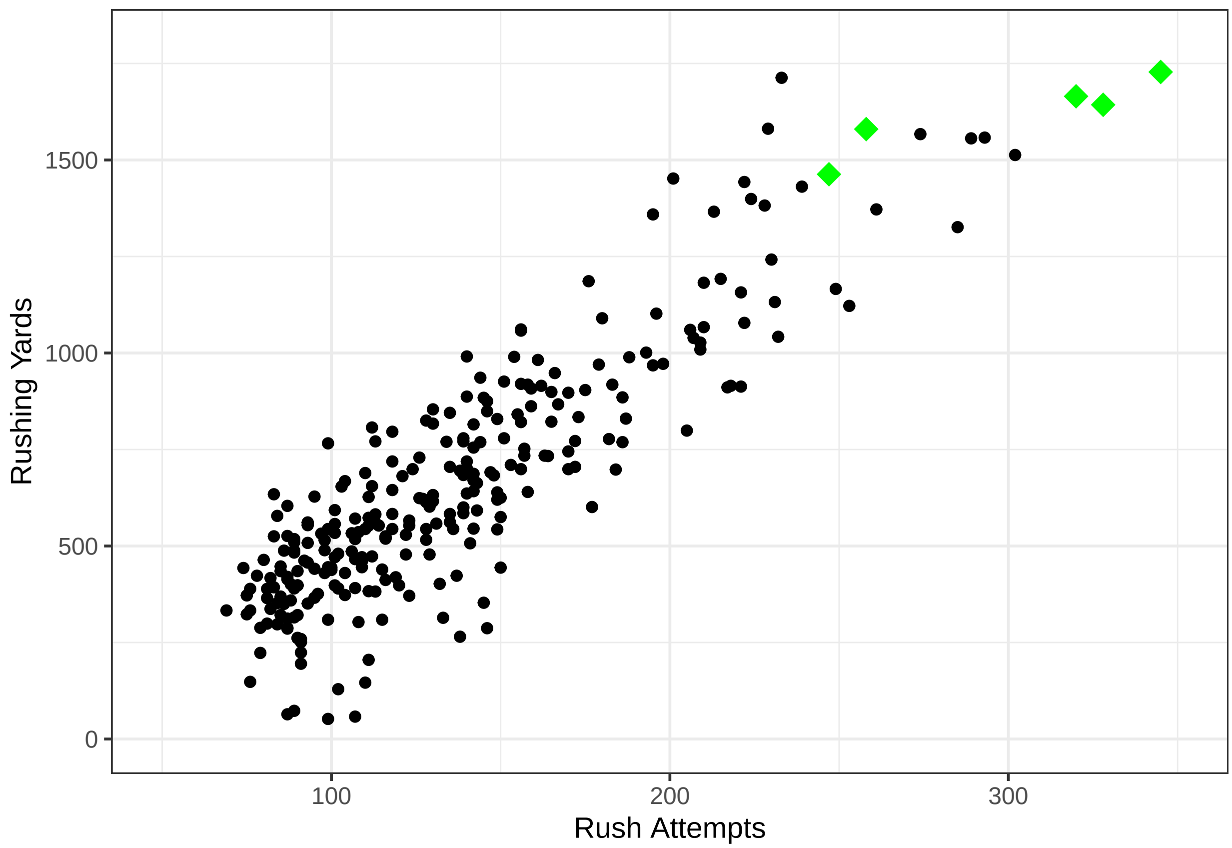

Figure 3 A scatterplot displaying the number of rushing attempts and the total rushing yards for NCAA college running backs from the 2022 season. The green points represent the outliers detected using a local outlier factor (LOF) algorithm on rushing attempts and average rushing yards per attempt. Average rushing yards per attempt was filtered to be greater than 1.0 to remove any additional players from the dataset that are not running backs.

Conclusions & Future Work

Using a local outlier factor algorithm to detect outliers in a dataset of college football player statistics worked relatively well - the detected outliers seemed to have impressive metrics compared to the rest of the dataset. I found it interesting that the outliers ended up having a wide range of different "ranks" that were already in the dataset. For future work I think it would be more interesting to dive deeper into this and use a more complex model, like a neural network.