

ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÁO CÁO MÔN HỌC

LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN

***Đề tài: Xây dựng data pipeline mô phỏng
và phân tích dữ liệu NFL***

Giảng viên hướng dẫn: PGS.TS Nguyễn Bình Minh

Mã lớp học: 136051

Danh sách sinh viên thực hiện:

STT	Họ và tên	MSSV	Lớp
1	Hữu Tường Tú	20194395	CTTN-KHMT-K64
2	Nguyễn Phúc Tân	20194163	CTTN-KHMT-K64
3	Nguyễn Thành Phong	20192016	CTTN-KHMT-K64
4	Trần Tiến Bằng	20193988	CTTN-KHMT-K64

MỤC LỤC

I. Giới thiệu đề tài	4
1. Đặt vấn đề	4
2. Mục tiêu và phạm vi của đề tài.....	4
II. Thực hiện đề tài	6
1. Tổng quan hệ thống.....	6
2. Xây dựng HDFS.....	6
3. Xây dựng Spark	8
4. Mô phỏng và phân tích dữ liệu.....	9
5. Xây dựng interface trực quan hóa dữ liệu.....	23
III. Kết luận và định hướng.....	25
IV. Tài liệu tham khảo	25

I. Giới thiệu đề tài

1. Đặt vấn đề

Bóng bầu dục Mỹ hay còn gọi là bóng đá kiểu Mỹ (American football hoặc Gridiron football) là một môn thể thao phổ biến tại nhiều nơi trên thế giới.

Mỗi trận đấu bóng bầu dục sẽ có các thống kê của từng đội hay từng cầu thủ. Qua các mùa giải một số lượng lớn trận đấu sẽ được diễn ra và việc xử lý thông tin của các trận đấu có thể mang lại những tri thức về chỉ số thống kê của các cầu thủ, chiến thuật thường hay sử dụng, những nơi xuất ra những cầu thủ giỏi, những nơi xem nhiều giải đấu...

Chính vì vậy, chúng em sẽ thu thập, lưu trữ, xử lý và phân tích dữ liệu lớn cần thiết để có thể đưa ra những mô phỏng và phân tích cho dữ liệu này.

2. Mục tiêu và phạm vi của đề tài

Thông qua khảo sát và phân tích, đề tài của chúng em sẽ lấy dữ liệu trận đấu bóng bầu dục NFL năm 2017 đến năm 2019. Sau đó chúng em sẽ xử lý và thống kê những chỉ số như độ tuổi, chiều cao, cân nặng các cầu thủ, các nơi xuất ra được nhiều cầu thủ nhất, các chiến thuật hay sử dụng, các vị trí cầu thủ hay xuất phát khi ở trên sân... từ đó để đánh giá những phương án và xu hướng, chiến thuật các đội.

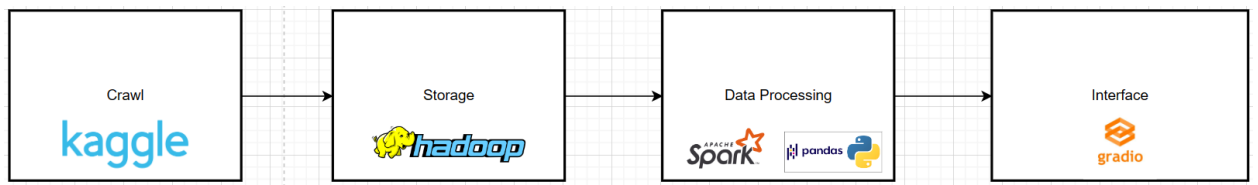
Trong đề tài này, cả nhóm chúng em đều cùng thực hiện nghiên cứu và phân tích các công nghệ và tài liệu liên quan. Cụ thể, mỗi người đã thực hiện những công việc như sau:

Nhiệm vụ	Người thực hiện
Crawl dữ liệu và lưu trữ dữ liệu lên HDFS	Nguyễn Phúc Tân
Truy vấn, xử lý, mô phỏng dữ liệu, xây dựng interface	Hữu Tường Tú
Phân tích dữ liệu, làm slide và báo cáo	Nguyễn Thành Phong
Phân tích dữ liệu, làm slide và báo cáo	Trần Tiến Bằng

II. Thực hiện đề tài

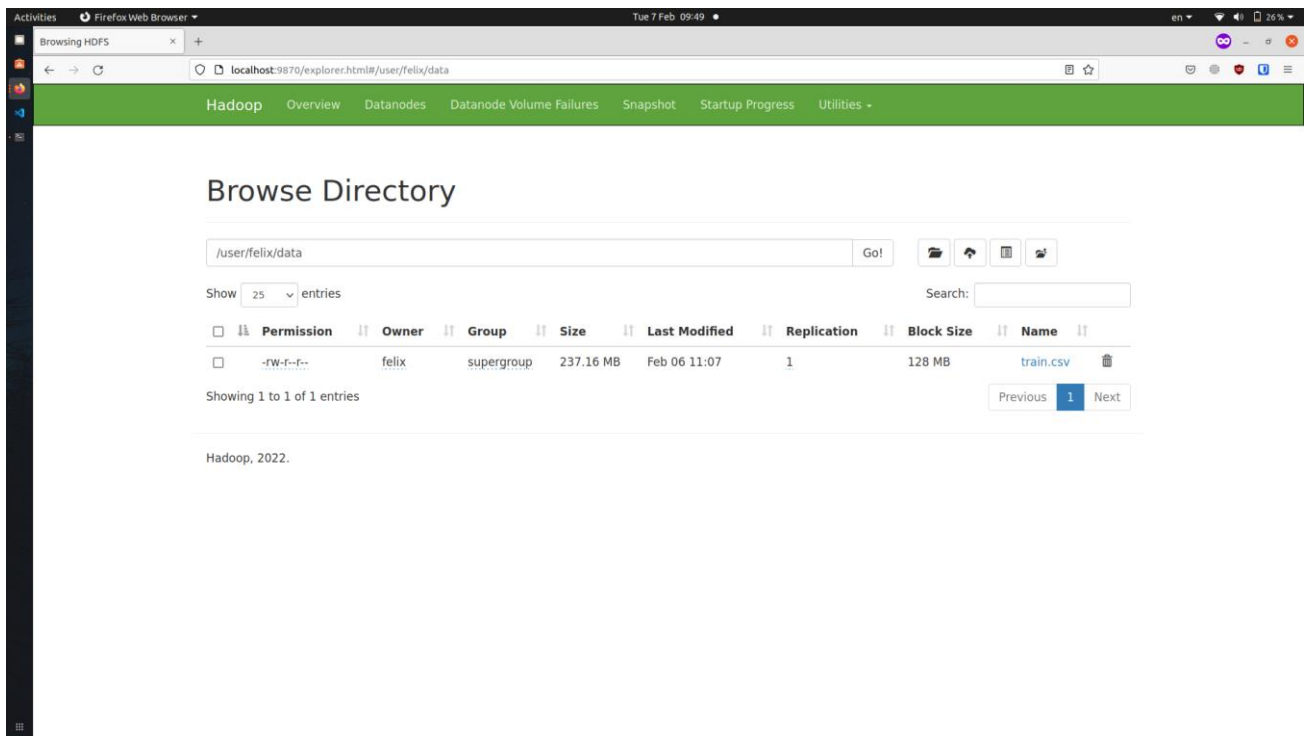
1. Tổng quan hệ thống

Đầu tiên, ta crawl dữ liệu NFL trên Kaggle xuống. Sau đó dữ liệu lấy về sẽ được lưu lại dưới dạng file csv ở HDFS. Từ dữ liệu csv, ta sử dụng PySpark để truy vấn, phân tích và xử lý dữ liệu kết hợp với các thư viện của python để mô phỏng dữ liệu. Cuối cùng dữ liệu trực quan hóa trên một nền tảng xây dựng bởi Gradio.

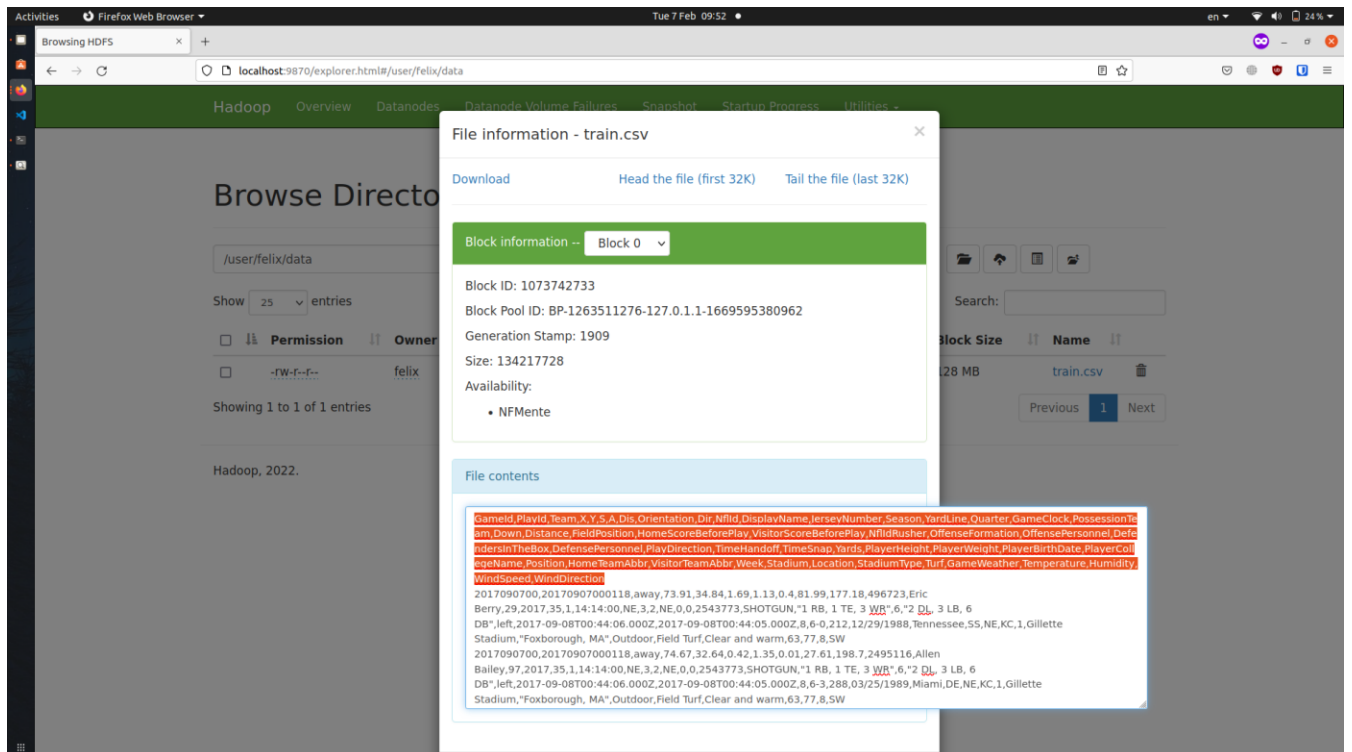


2. Xây dựng HDFS

Các thư mục trên Hdfs:



Dữ liệu được lưu trên Hdfs như sau:



3. Xây dựng Spark

Cấu hình của Spark khi xử lý dữ liệu

```
spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

df = spark.read.csv("train.csv",header=True,sep=",")
```

Tổng lượng dữ liệu ta có:

```
print("Total Row Number: {0} \nTotal Col Number: {1}".format(df.count(), len(df.dtypes)))

Total Row Number: 682154
Total Col Number: 49
```

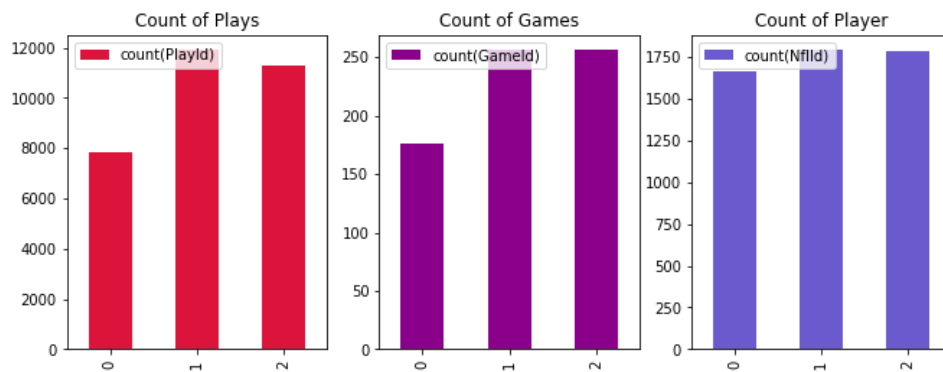
Dữ liệu ta thu được sau khi đọc csv bởi Pyspark được lưu dưới dạng Dataframe. Dưới đây là 20 dữ liệu đầu tiên:

GameId	PlayId	Team	X	Y	S	A	Dis	Orientation	Dir	NflId	DisplayName	JerseyNumber	Season	YardLine	Quarter
2017090700	20170907000118	away	73.91	34.84	1.69	1.13	0.4	81.99	177.18	496723	Eric Berry	29	2017	35	1
2017090700	20170907000118	away	74.67	32.64	0.42	1.35	0.01	27.61	198.7	2495116	Allen Bailey	97	2017	35	1
2017090700	20170907000118	away	74	33.2	1.22	0.59	0.31	3.01	202.73	2495493	Justin Houston	50	2017	35	1
2017090700	20170907000118	away	71.46	27.7	0.42	0.54	0.02	359.77	105.64	2506353	Derrick Johnson	56	2017	35	1
2017090700	20170907000118	away	69.32	35.42	1.82	2.43	0.16	12.63	164.31	2530794	Ron Parker	38	2017	35	1
2017090700	20170907000118	away	75.06	24	1.01	0.32	0.18	308.34	95.01	2543494	Dee Ford	55	2017	35	1
2017090700	20170907000118	away	74.11	16.64	1.11	0.83	0.02	357.23	322.59	2543637	Terrance Mitchell	39	2017	35	1
2017090700	20170907000118	away	73.37	18.73	1.24	0.74	0.13	328.52	270.04	2543851	Phillip Gaines	23	2017	35	1
2017090700	20170907000118	away	56.63	26.9	0.26	1.86	0.28	344.7	55.31	2550257	Daniel Sorensen	49	2017	35	1
2017090700	20170907000118	away	73.35	38.83	4.55	0.76	0.51	75.47	190.84	2552488	Marcus Peters	22	2017	35	1
2017090700	20170907000118	away	74.15	28.9	0.72	0.73	0.01	342.58	274.14	2556369	Chris Jones	95	2017	35	1
2017090700	20170907000118	home	75.82	17.56	2.3	1.39	0.55	178.97	284.15	2649	Danny Amendola	80	2017	35	1
2017090700	20170907000118	home	74.78	33.21	1.71	0.82	0.19	178.82	215.9	497240	Rob Gronkowski	87	2017	35	1
2017090700	20170907000118	home	75.43	32.41	1.5	1.36	0.32	207.08	222.76	2495131	Marcus Cannon	61	2017	35	1
2017090700	20170907000118	home	75.9	25.12	1.38	0.8	0.19	133.01	198.55	2495232	Nate Solder	77	2017	35	1
2017090700	20170907000118	home	79.76	29.49	0.84	1.22	0	192.18	110.86	2504211	Tom Brady	12	2017	35	1
2017090700	20170907000118	home	76.47	36.91	5.15	0.77	0.59	112.02	195.09	2530515	Chris Hogan	15	2017	35	1
2017090700	20170907000118	home	74.7	19.19	2.1	1.48	0.51	152.14	278.52	2543498	Brandin Cooks	14	2017	35	1
2017090700	20170907000118	home	78.75	30.53	3.63	3.35	0.38	161.98	245.74	2543773	James White	28	2017	35	1
2017090700	20170907000118	home	74.6	31.88	1.86	1.51	0.17	218.49	267.32	2552563	Shaq Mason	69	2017	35	1

Từ dữ liệu này, chúng ta sẽ sử dụng các hàm trong Pyspark như groupBy, agg, sum, count, ... để tính toán ra các trường theo tần số, bỏ những trường null và tiến hành truy vấn ra những dữ liệu cần thiết của các bảng.

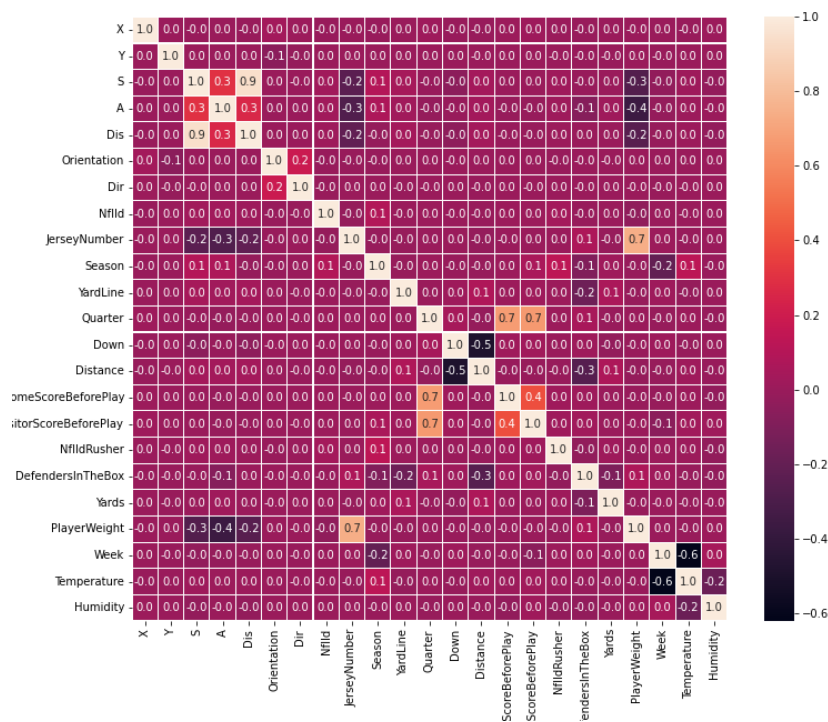
4. Mô phỏng và phân tích dữ liệu

1. Tổng lượng game/playid/players



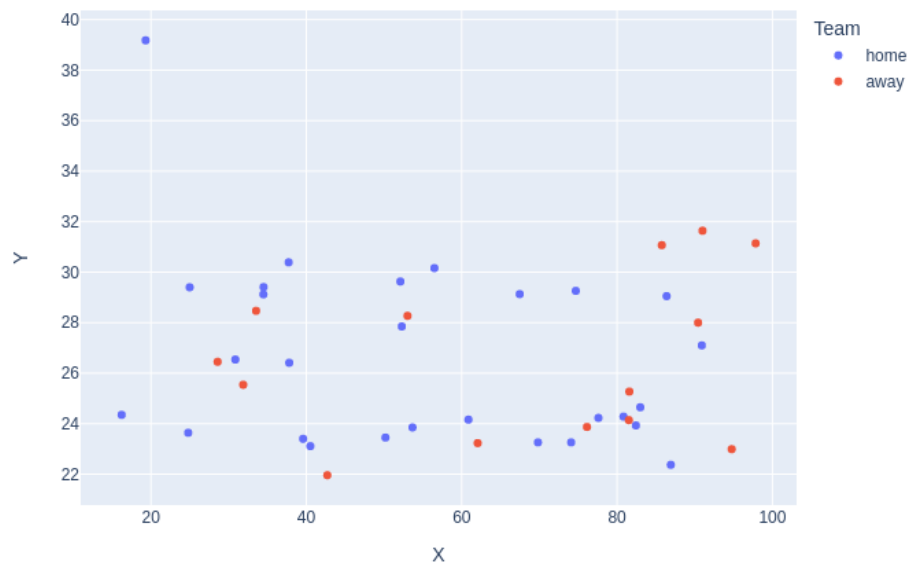
Đây là biểu đồ số lượt plays, games, players. Từ bên trái qua là 2019, 2017 và 2019. 1 Games sẽ thường có nhiều lượt plays (Ta có thể coi là quarter). Do 1 team có 11 thành viên, 1 trận thường có 4 quarters do đó khi giống từ cột plays sang cột games sẽ gấp nhau khoảng 44 – 45 lần.

2. Biểu đồ tương quan giữa các trường dữ liệu



Từ biểu đồ tương quan, ta dễ thấy S,A,Dis có tỉ lệ tương quan với nhau cao vì khoảng cách sẽ phụ thuộc vào vận tốc và gia tốc. Tỉ lệ tương quan giữa quarter với scorebeforeplay và scoreafterplay dễ thấy do nếu hiệp đấu hòa thì cần có quarter 5 (Hiệp phụ). Điều thú vị ở đây là Số áo đấu với PlayWeight lại có mối tương quan khá cao với nhau.

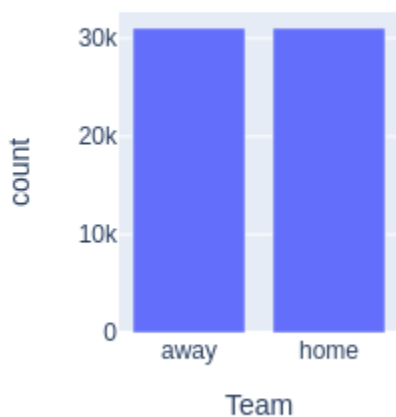
3. Biểu đồ vị trí 1 cầu thủ role CB trong game



Đây là biểu đồ vị trí trong 1 game của 2 cầu thủ 2 team cùng role là “CB”. Ta có thể thấy đa phần 2 cầu thủ sẽ chơi từ Yards 30 hất về 0. Có 1 điểm ngoại lai ở Yards 40.

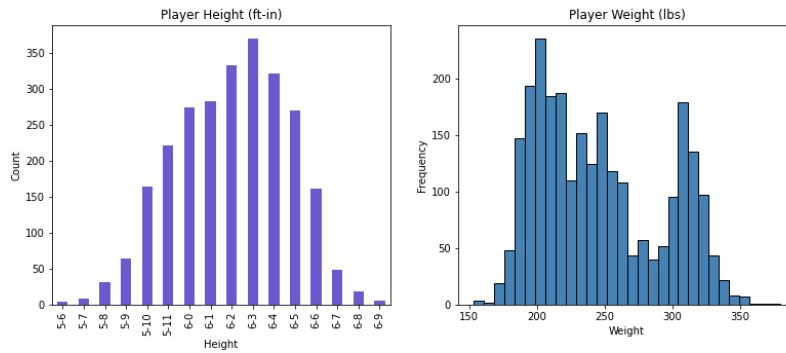
4. Biểu đồ Away-Home

Team (Away or Home)



Đây là biểu đồ đá ở sân nhà hay sân khách. Hiển nhiên số lượt sân nhà và lượt sân khách là bằng nhau.

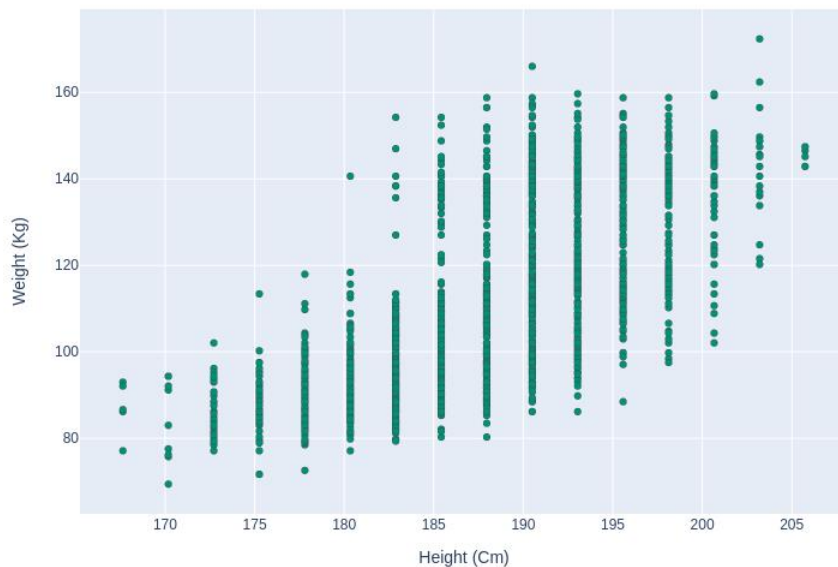
5. Biểu đồ số lượng cân nặng, chiều cao



Ta thấy biểu đồ chiều cao có tuân theo phân phối chuẩn với chiều cao có tần số cao nhất là 6.3 ft tương đương với khoảng 1.9m. Ở biểu đồ cân nặng ta thấy gần như phân bố thành tổng 3 phân phối chuẩn. Kết hợp với biểu đồ tương quan tổng quan ta có thể suy ra rằng giữa vị trí chơi, số áo và cân nặng sẽ có một mối liên hệ nào đó với nhau.

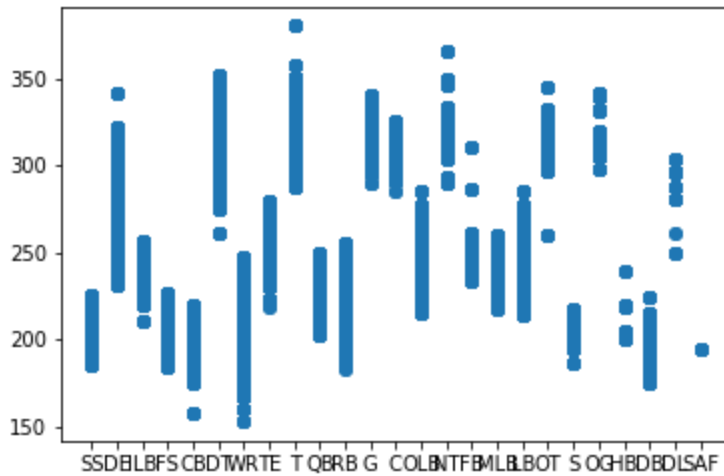
6. Biểu đồ tương quan cân nặng - chiều cao

Weight Height Correlation



Biểu đồ tương quan chiều cao cân nặng chỉ ra rằng một cầu thủ càng cao thì cân nặng sẽ càng lớn. Việc này dễ thấy đúng trong thực tế. Việc ở khoảng 1m85 đến 1m95 thấy phủ đều cân nặng cũng chứng tỏ rằng chiều cao đó chính là vùng có tần số cao nhất.

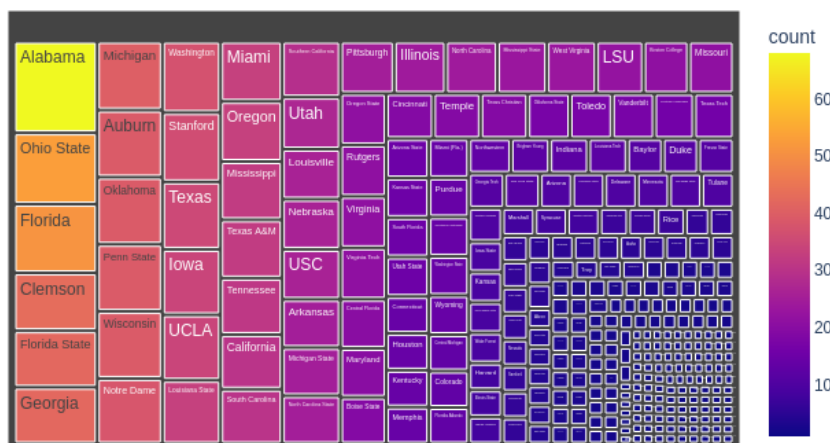
7. Biểu đồ tương quan vị trí - cân nặng



Ta thấy được sự phân phối tương quan rõ ràng ở vị trí các cầu thủ với cân nặng của họ. Mỗi role sẽ thường có một khoảng cân nặng cố định. Kết hợp với thông tin số áo vị trí cân nặng ở trên ta có thể đoán rằng với mỗi vị trí trên sân sẽ được khoảng số áo đầu nhất định. Điều này khá thú vị vì trong thực tế khi đọc luật chơi thì việc mỗi vị trí cũng có 1 khoảng áo đầu phân bổ.

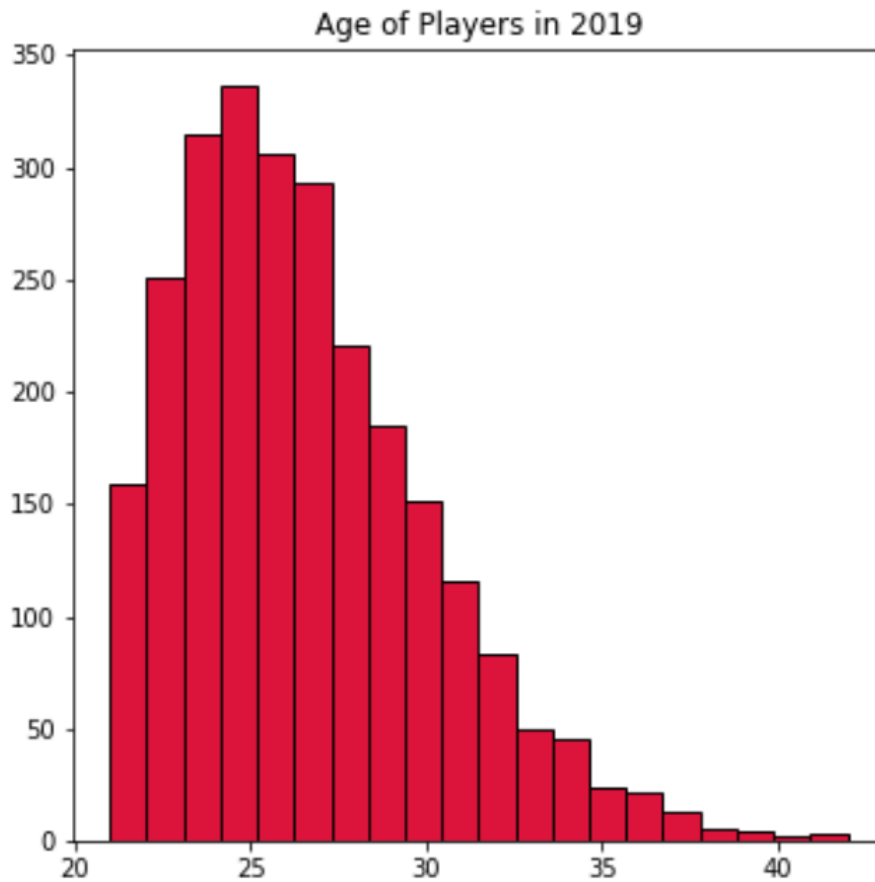
8. Biểu đồ số lượng cầu thủ từng học đại học

Arrangement Of Colleges From Highest To Lowest Number Of Players



Biểu đồ cho thấy ở trường Alabama có số lượng tuyển thủ cao nhất. Từ biểu đồ này, các huấn luyện viên có thể biết được nên đi tìm thành viên clb mình ở trường nào.

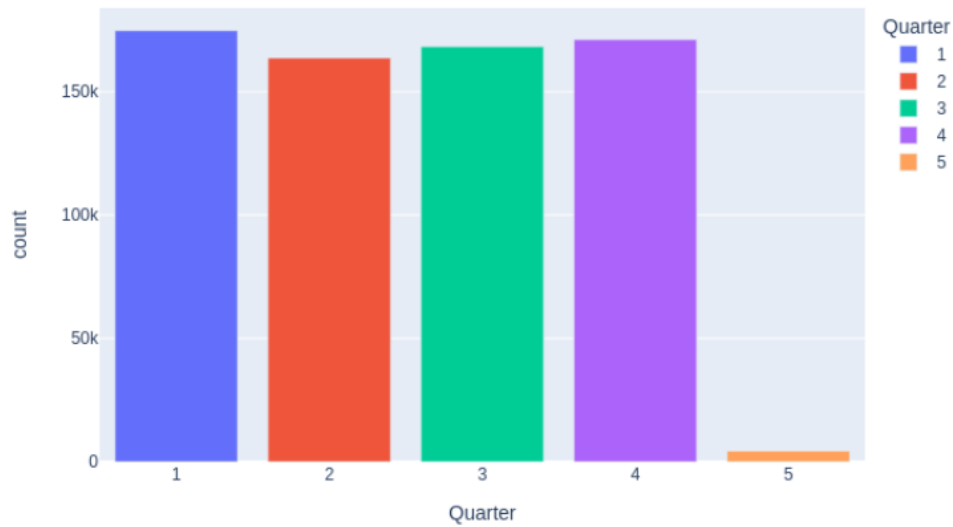
9. Biểu đồ tần số tuổi



Lứa tuổi chơi bóng bầu dục chủ yếu là từ 21 đến 31 tuổi. Phổ biến nhất là 25 tuổi. Tuổi thấp nhất là 21, cao nhất là 41. Cũng như bóng đá ta có thể thấy tầm trên 30 tuổi là tuổi ở bên kia sườn dốc do đó biểu đồ thấy được sự giảm liên tục.

10. Biểu đồ các hiệp trong trận đấu

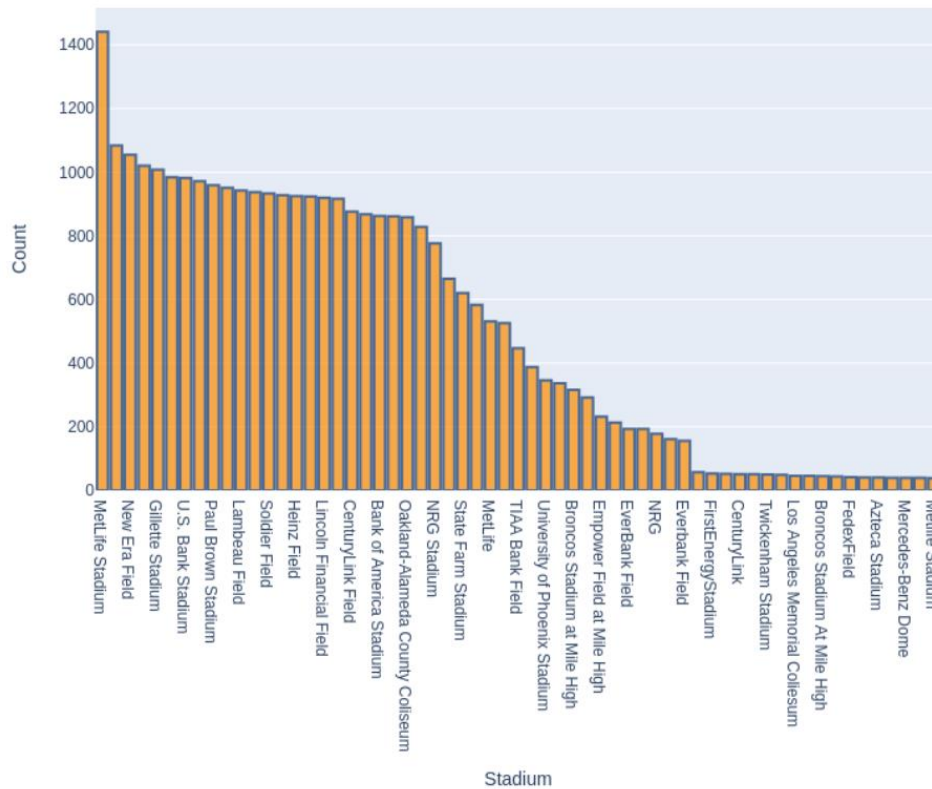
Distinct Count Of NFL Game Quarters



Do tập dữ liệu còn nhiều chỗ thiếu sót nên số hiệp 1,2,3,4 có tần số gần bằng nhau. Qua biểu đồ, ta cũng thấy được tần số của các hiệp chính gấp cỡ 40 lần so với hiệp phụ. Suy ra, các trận đấu trong 4 hiệp chính có tỉ lệ hòa là khoảng 2.5%.

11. Biểu đồ số lượng game trên sân vận động

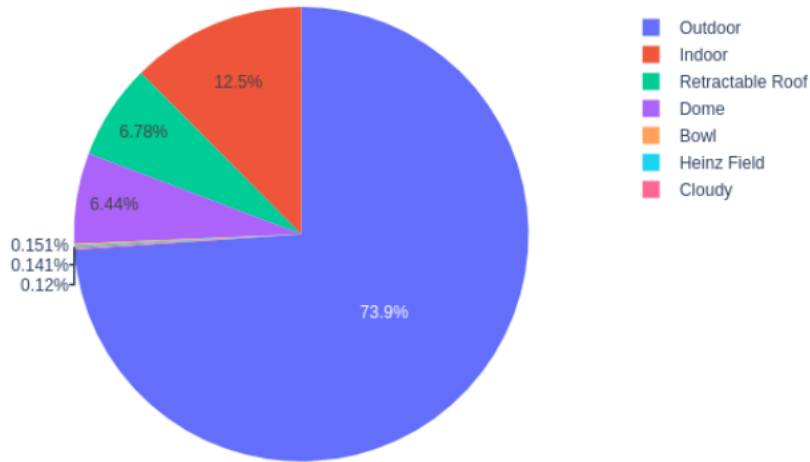
Games by Stadium



Sân vận động được sử dụng nhiều nhất cho bóng bầu dục là sân vận động MetLife. Qua biểu đồ, ta thấy được các sân vận động được chia thành 3 nhóm, nhóm được sử dụng nhiều ($\text{count} > 800$), vừa ($800 > \text{count} > 200$) và ít ($\text{count} < 200$). Sân metLife qua tìm hiểu thì đây là sân vận động duy nhất được chia sẻ bởi 2 câu lạc bộ do đó có lẽ đây là lý do cột cao đến vậy.

12. Biểu đồ số lượng game bởi loại sân vận động

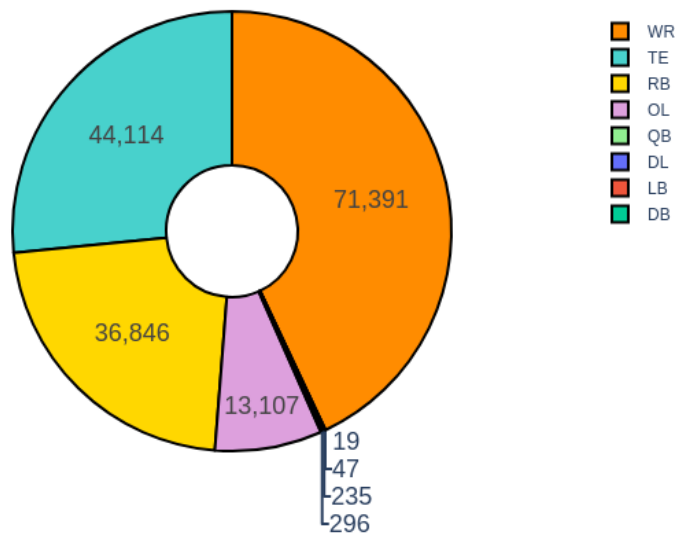
Games by Stadium Type



Đa phần các game sẽ được tổ chức trên các sân vận động ngoài trời, số còn lại sẽ được tổ chức trên các sân vận động trong nhà, mái vòm hoặc có mái nhà có thể thu đóng

13. Biểu đồ tỉ lệ role tấn công tuyến thủ

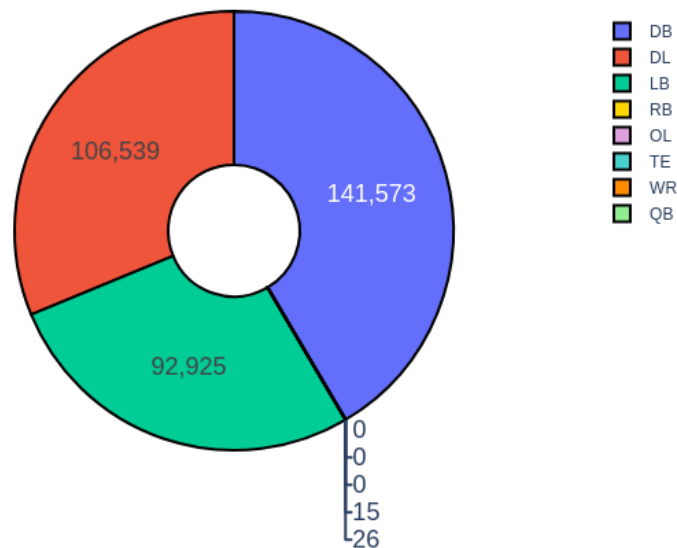
Offensive Players Ratio in All Plays



Tỉ lệ cầu thủ trong vai trò tấn công trong tất cả các pha bóng. Bốn vai trò phổ biến nhất là Wide Receiver, Tight End, Running Back và Outside Linebacker.

14. Biểu đồ tỉ lệ role phòng thủ tuyến thủ

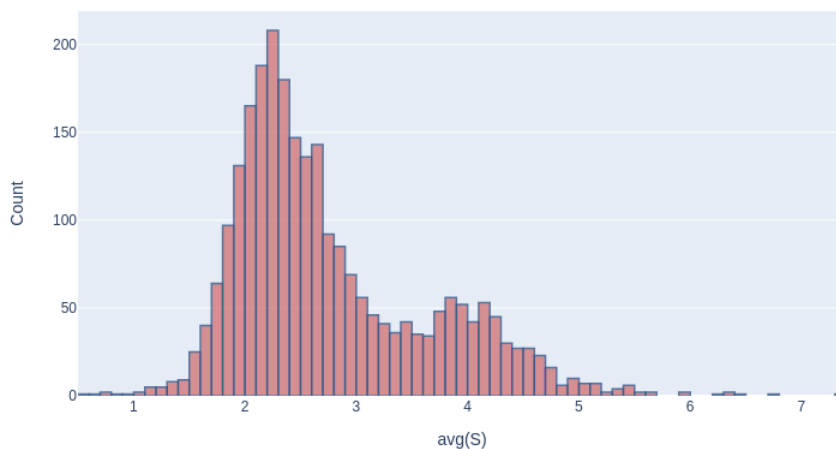
Defense Players Ratio in All Plays



Tỉ lệ cầu thủ trong vai trò phòng thủ trong tất cả các pha bóng. Ba vai trò phổ biến nhất là Defensive Back, Defensive Line, Linebacker.

15. Phân phối tốc độ

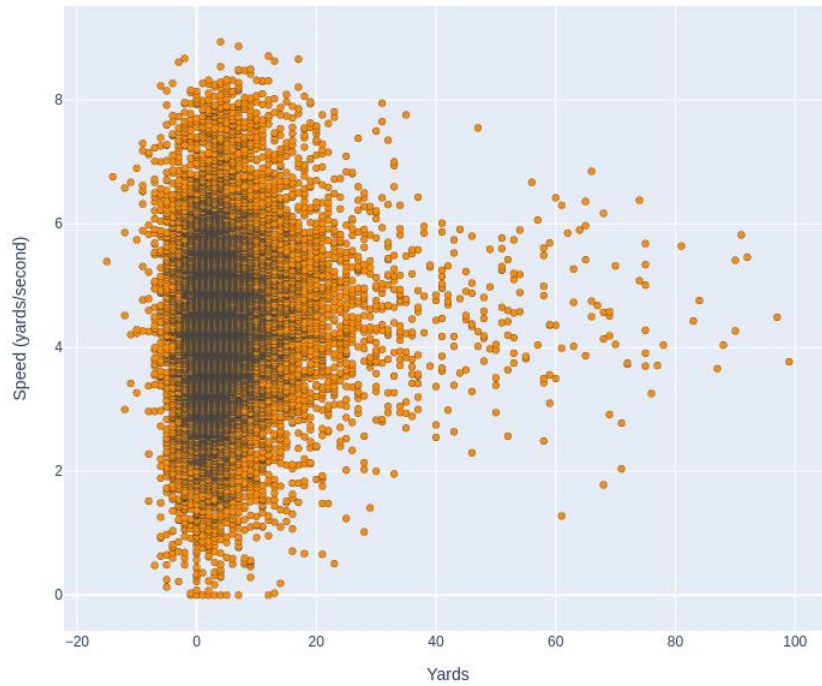
Average Speed Distribution of Players



Phân phối tốc độ có hình dạng tổng của 2 phân phối chuẩn với mode=2,3 và mode=4. Điều này có nghĩa là một số cầu thủ có tốc độ cao thuộc một số role thiên về tấn công được sử dụng nhiều.

16. Biểu đồ tốc độ trên yards

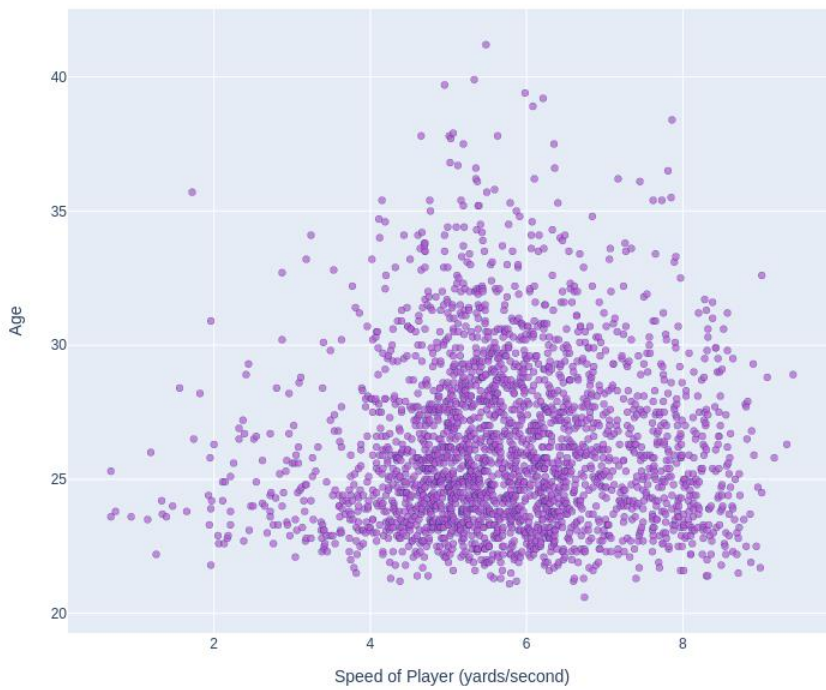
Speed to Yards



Từ biểu đồ ta thấy ở Yards -5 đến 5 là có tần số cao nhất vì đây là vị trí gần goal line tức là ở vị trí này thường là điểm nóng của trận đấu. Tốc độ trung bình sẽ thuộc khoảng 4-5 yards/s.

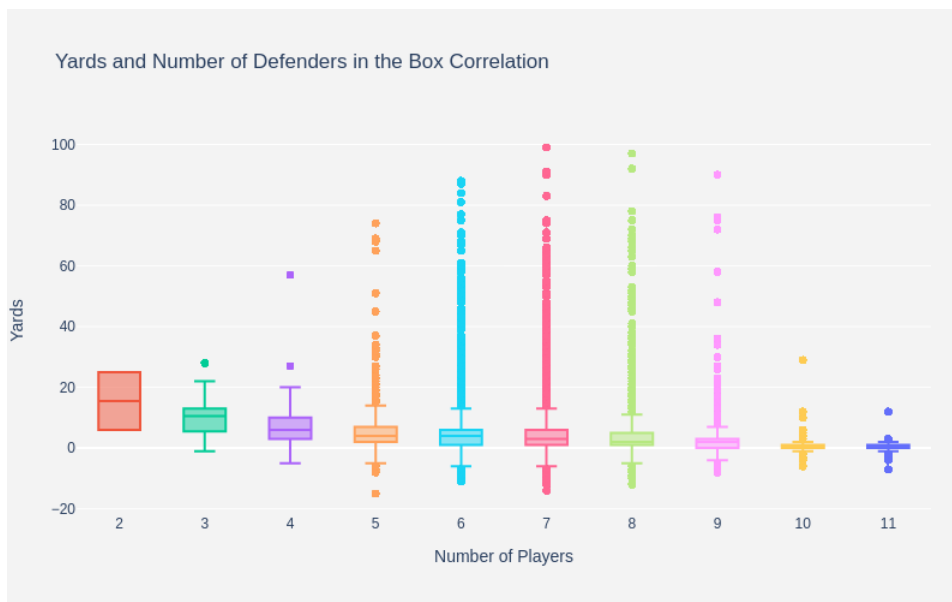
17. Biểu đồ độ tuổi đỉnh cao của các cầu thủ

Age of The Players When They Make Their Maximum Speed



Đa phần các cầu thủ đạt đỉnh về tốc độ ở độ tuổi 23-27. Đa phần trong khoảng 4-8 yards/second

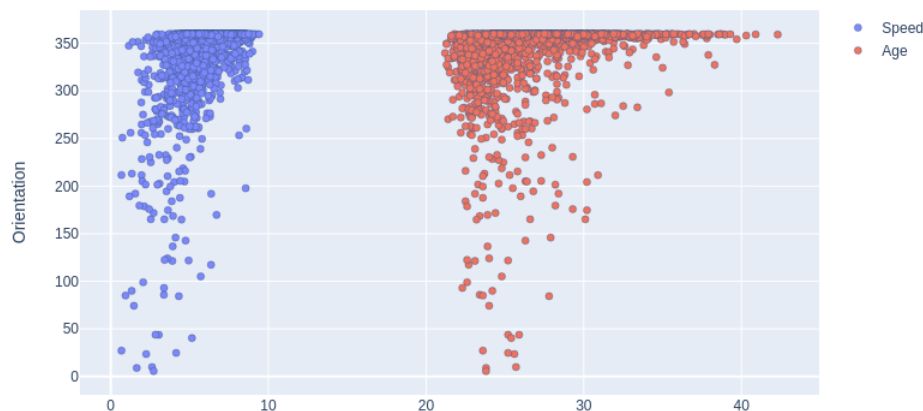
18. Biểu đồ defender in the box



Biểu đồ nghĩa là số người thủ goalie xếp hàng gần goalie, kéo dài theo chiều rộng của đường tấn công. Ta thấy càng về gần yards 0 số người hàng thủ càng tăng do đây là vị trí gần goalline nên cần bảo vệ khỏi việc đối phương ghi bàn

19. Biểu đồ orientation – age – speed

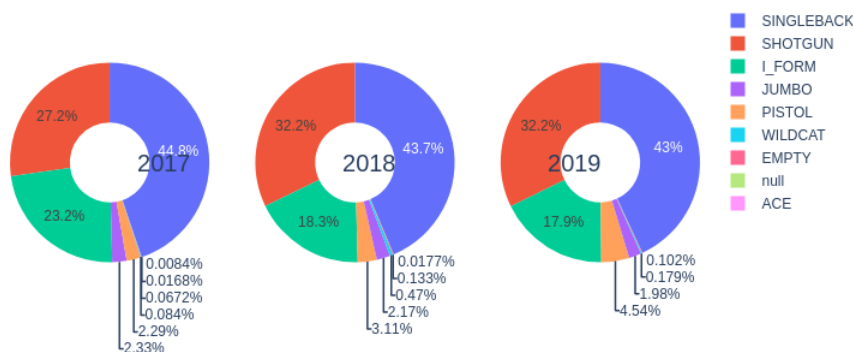
Orientation & Speed + Orientation & Age Correlation



Biểu đồ định hướng chạy tương quan với tốc độ và tuổi tác, ta thấy các cầu thủ trẻ thường có xu hướng chạy nhiều hướng khác nhau do có sức khỏe nhiều hơn. Các cầu thủ già hơn thì thường chạy thẳng về goalline đối phương tuy nhiên nhìn chung thì các cầu thủ đều có xu hướng chạy thẳng lên goal line

20. Biểu đồ đội hình team

Favorite Offense Formation in 2017 and 2018 and 2019

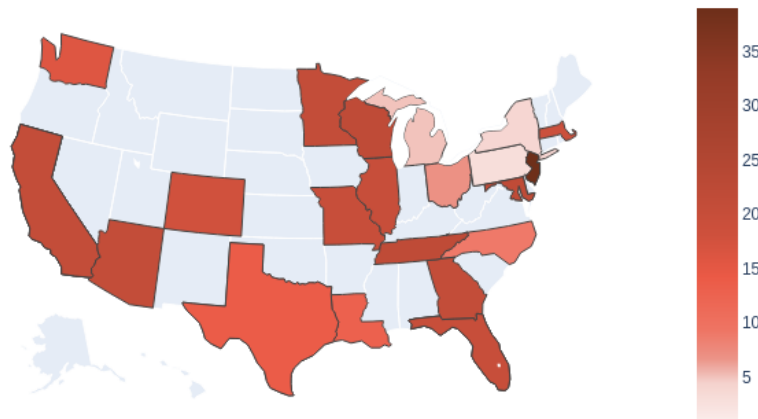


Từ biểu đồ trên, ta thấy có một sự thay đổi nhẹ giữa các năm, song thứ tự tần suất của các đội hình gần như không thay đổi. Ba đội hình tấn công được sử dụng phổ biến nhất ở mỗi năm đều là: SINGLEBACK, SHOTGUN, I_FORM.

Các đội hình tấn công như SHOTGUN, PISTOL đang có xu hướng được sử dụng tăng nhẹ qua các năm. Số đội sử dụng đội hình tấn công I_FORM và các đội hình có tần suất thấp đang giảm mạnh. Đội hình SINGLEBACK có xu hướng giảm nhẹ, song vẫn là đội hình được sử dụng nhiều nhất qua các năm

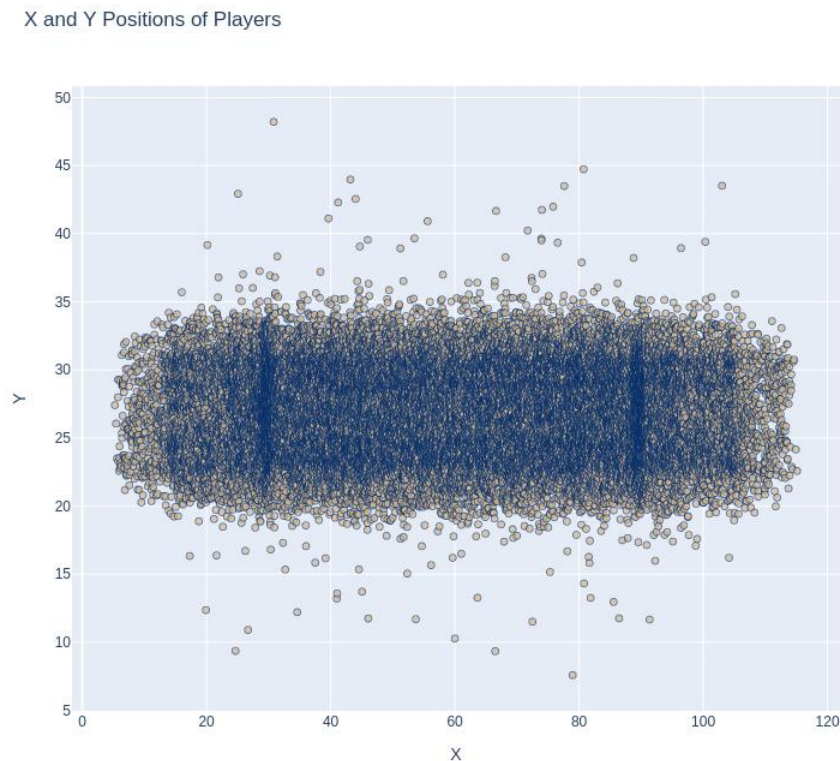
21. Biểu đồ tổng game trên 1 bang tại Mỹ

Total Number of Games by States



Dựa vào biểu đồ, các trận đấu thường được tổ chức tại các bang có kinh tế lớn hơn, thường là gần biển, có điều kiện thuận lợi, du lịch phát triển, được đầu tư nhiều sân đấu chuyên nghiệp.

22. Biểu đồ phân phối vị trí cầu thủ



Dựa vào biểu đồ, đa số các cầu thủ thường tập trung ở giữa sân. Ta dễ dàng quan sát thấy có 4 đường kẻ tập trung lượng lớn cầu thủ, là do đây là 4 đường quan trọng gọi là Hash mark. Dựa vào 4 đường này thì một tuyển thủ có thể biết được vị trí của mình trên sân và cũng từ đó có thể dễ dàng liên kết với đồng đội mình để chuyền bóng.

5. Xây dựng interface trực quan hóa dữ liệu

Trong đề tài này, chúng ta lựa chọn gradio để xây dựng làm interface. Ở interface, chúng ta cho 1 dropdownlist để chọn những thuộc tính cần phân tích. Sau đó khi ta submit thì ở bên phần output của chúng ta sẽ có 3 output. Output thứ nhất là mô phỏng của dữ liệu chúng ta đã chọn thuộc tính. Output thứ hai và output thứ ba dùng để so sánh tốc độ (hiệu suất) giữa việc truy vấn bằng pandas và truy vấn bằng Spark. Ở cuối trang là các trường của dữ liệu và giải thích ý nghĩa của từng trường.

Giao diện màn hình chính:

Big Data - NFL Analysis

Attribute_choice

Clear Submit

output 0

output 1



output 2

Flag

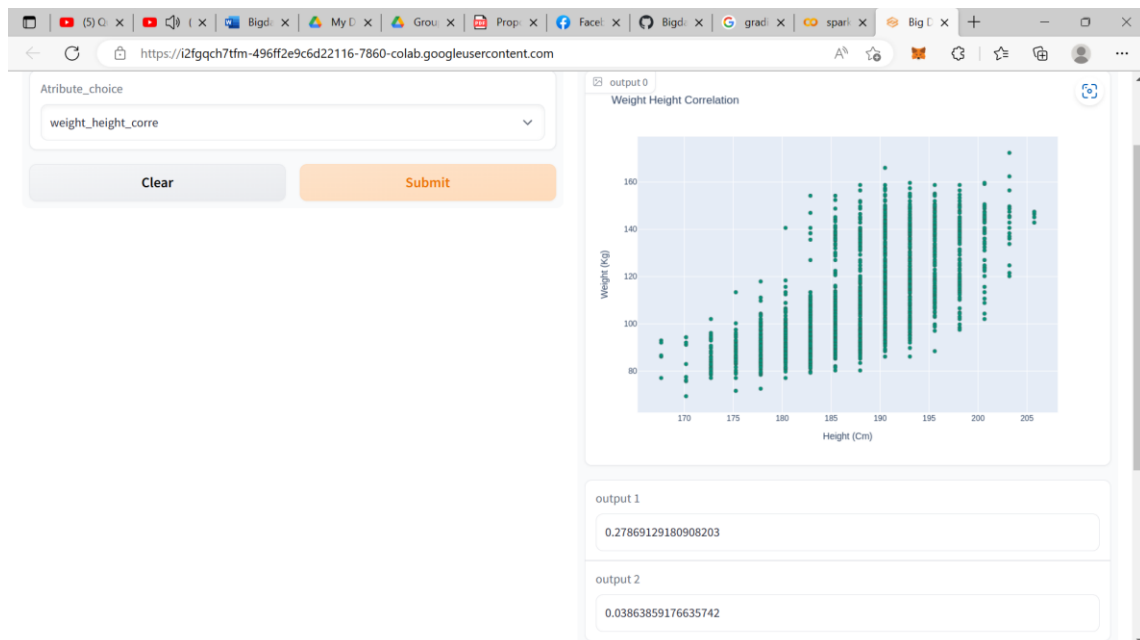
Vị trí ý nghĩa của các trường dữ liệu:

Gamelid - a unique game identifier PlayId - a unique play identifier Team - home or away X - player position along the long axis of the field. See figure below.
 Y - player position along the short axis of the field. See figure below. S - speed in yards/second A - acceleration in yards/second^2 Dis - distance traveled from prior time point, in yards
 Orientation - orientation of player (deg) Dir - angle of player motion (deg)

Pages: 1 2 3 4 5

Use via API  · Built with Gradio 

Giao diện khi ta chọn 1 thuộc tính:



III. Kết luận và định hướng

Qua đề tài lần này, nhóm chúng em đã tìm hiểu, nghiên cứu và xây dựng được một data pipeline cơ bản từ các công nghệ như Hdfs, Spark, Gradio. Trong tương lai, chúng em mong là có thể phát triển data pipeline của mình ngày càng lớn hơn hỗ trợ được thêm nhiều tính năng hơn trên nền tảng của mình như có thể để người dùng sử dụng dữ liệu streaming dữ liệu, hỗ trợ phân tích nhiều cách khác nhau hơn...

IV. Tài liệu tham khảo

Tài liệu tham khảo:

<https://www.kaggle.com/>

<https://www.kaggle.com/code/sanjayv007/nfl-big-data-bowl-beginner-s-complete-eda/notebook>

[Apache Spark™ - Unified Engine for large-scale data analytics](#)

[Apache Hadoop](#)

Link source code: [huutuongtu/Bigdata \(github.com\)](https://github.com/huutuongtu/Bigdata)