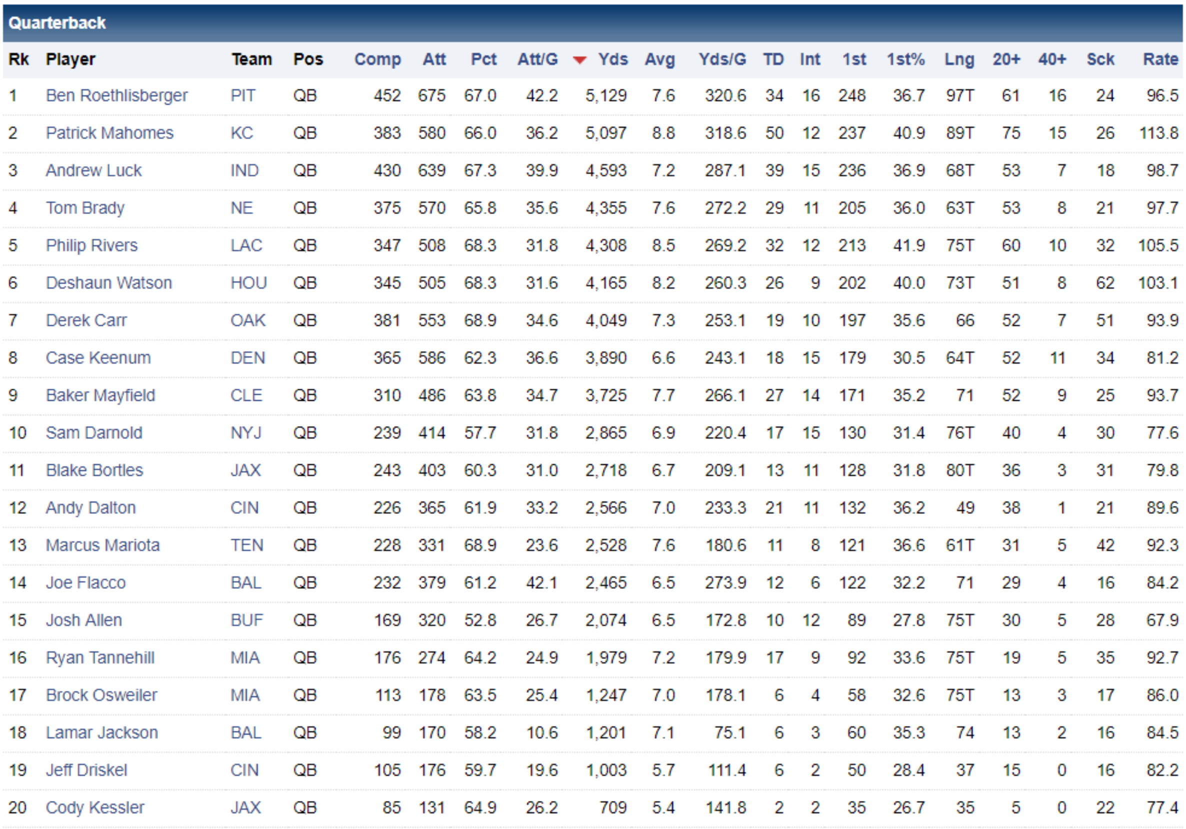
AIT-580 Data Analysis Individual Project
(Kausik Valeti, G01178711, Sec-004)

SPORTS

Deliverable 1 – Dataset Selection & description:

I have selected a data set in Sports which is from Pro Football Reference.com (Forman). The sport I selected is NFL(National Football League).

Data description:



Quarterback

Rk	Player	Team	Pos	Comp	Att	Pct	Att/G	▼ Yds	Avg	Yds/G	TD	Int	1st	1st%	Lng	20+	40+	Sck	Rate
1	Ben Roethlisberger	PIT	QB	452	675	67.0	42.2	5,129	7.6	320.6	34	16	248	36.7	97T	61	16	24	96.5
2	Patrick Mahomes	KC	QB	383	580	66.0	36.2	5,097	8.8	318.6	50	12	237	40.9	89T	75	15	26	113.8
3	Andrew Luck	IND	QB	430	639	67.3	39.9	4,593	7.2	287.1	39	15	236	36.9	68T	53	7	18	98.7
4	Tom Brady	NE	QB	375	570	65.8	35.6	4,355	7.6	272.2	29	11	205	36.0	63T	53	8	21	97.7
5	Philip Rivers	LAC	QB	347	508	68.3	31.8	4,308	8.5	269.2	32	12	213	41.9	75T	60	10	32	105.5
6	Deshaun Watson	HOU	QB	345	505	68.3	31.6	4,165	8.2	260.3	26	9	202	40.0	73T	51	8	62	103.1
7	Derek Carr	OAK	QB	381	553	68.9	34.6	4,049	7.3	253.1	19	10	197	35.6	66	52	7	51	93.9
8	Case Keenum	DEN	QB	365	586	62.3	36.6	3,890	6.6	243.1	18	15	179	30.5	64T	52	11	34	81.2
9	Baker Mayfield	CLE	QB	310	486	63.8	34.7	3,725	7.7	266.1	27	14	171	35.2	71	52	9	25	93.7
10	Sam Darnold	NYJ	QB	239	414	57.7	31.8	2,865	6.9	220.4	17	15	130	31.4	76T	40	4	30	77.6
11	Blake Bortles	JAX	QB	243	403	60.3	31.0	2,718	6.7	209.1	13	11	128	31.8	80T	36	3	31	79.8
12	Andy Dalton	CIN	QB	226	365	61.9	33.2	2,566	7.0	233.3	21	11	132	36.2	49	38	1	21	89.6
13	Marcus Mariota	TEN	QB	228	331	68.9	23.6	2,528	7.6	180.6	11	8	121	36.6	61T	31	5	42	92.3
14	Joe Flacco	BAL	QB	232	379	61.2	42.1	2,465	6.5	273.9	12	6	122	32.2	71	29	4	16	84.2
15	Josh Allen	BUF	QB	169	320	52.8	26.7	2,074	6.5	172.8	10	12	89	27.8	75T	30	5	28	67.9
16	Ryan Tannehill	MIA	QB	176	274	64.2	24.9	1,979	7.2	179.9	17	9	92	33.6	75T	19	5	35	92.7
17	Brock Osweiler	MIA	QB	113	178	63.5	25.4	1,247	7.0	178.1	6	4	58	32.6	75T	13	3	17	86.0
18	Lamar Jackson	BAL	QB	99	170	58.2	10.6	1,201	7.1	75.1	6	3	60	35.3	74	13	2	16	84.5
19	Jeff Driskel	CIN	QB	105	176	59.7	19.6	1,003	5.7	111.4	6	2	50	28.4	37	15	0	16	82.2
20	Cody Kessler	JAX	QB	85	131	64.9	26.2	709	5.4	141.8	2	2	35	26.7	35	5	0	22	77.4

Fig-1: Dataset Image (Gracenote). (NFL)

The size of the data that I have is 6kb which might be small because in sports we can analyse only a limited no.of players which we don't have like huge data sets in terms of GB's. Regarding the Data items I have player names and their statistics from AFC with all Quarterbacks in 2018 regular season from all 16 teams. This data is stored in form of Excel on my PC, which can also be stored on server or any where because of negligible size.

The data was collected by Pro Football Reference.com (Forman) maintaining stats of all the players who are playing in NFL. The purpose of collecting stats of players make available to fans who are interested in looking their favourite player stats and how they have been performing since they started their career in NFL.

AIT-580 Data Analysis Individual Project
(Kausik Valeti, G01178711, Sec-004)

As we are having data collected only for quarterback's of 2018 regular season of AFC, the main problem is to find out who is the best Quarterback that suits a team in season 2019 of all the players if they want to buy. There is no privacy, quality or ethical issues with the data because the data was present on official NFL website which is available for fans and public.

Deliverable 1 – Dataset Selection & description:

After studying the data that I have, the plan was to go for clustering technique, if possible some other models and can be improved on the data items. Categorising quarterbacks' depending up on their skills, which a team coach want to buy a Quarterback for their game plans in season 2019.

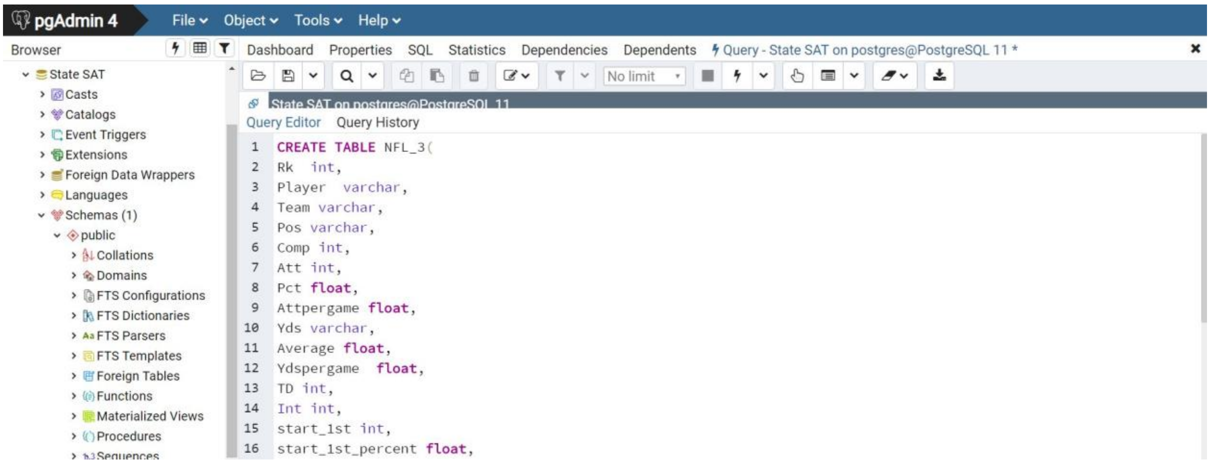
A best Quarterback can be selected on following stats like, attempts, yards per attempt, Longest pass play, Touchdown passes, interceptions finally Ratings. Considering all these factors can decide of selecting the best Quarterback for a team in upcoming season according to the plans made by head coach.

The software's that may require are R, Python, SQL and the hardware that might be required is win10-OS, Intel-I7 7th Gen, 8GB RAM, min Hard disk, Nvidia GeForce 940MX 2GB graphics.

AIT-580 Data Analysis Individual Project
(Kausik Valeti, G01178711, Sec-004)

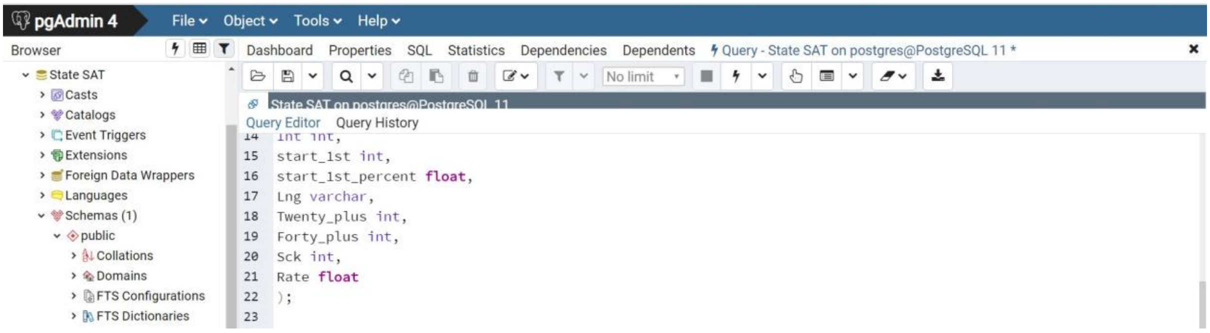
Deliverable 2 – Data Analysis & interpretation Report:

The first step I did was visualizing the data in SQL and writing some quires in sorting the data which was complex in Python and R when compared to SQL.



```
CREATE TABLE NFL_3(
Rk  int,
Player  varchar,
Team varchar,
Pos varchar,
Comp int,
Att int,
Pct float,
Attpergame float,
Yds varchar,
Average float,
Ydspergame  float,
TD int,
Int int,
start_1st int,
start_1st_percent float,
```

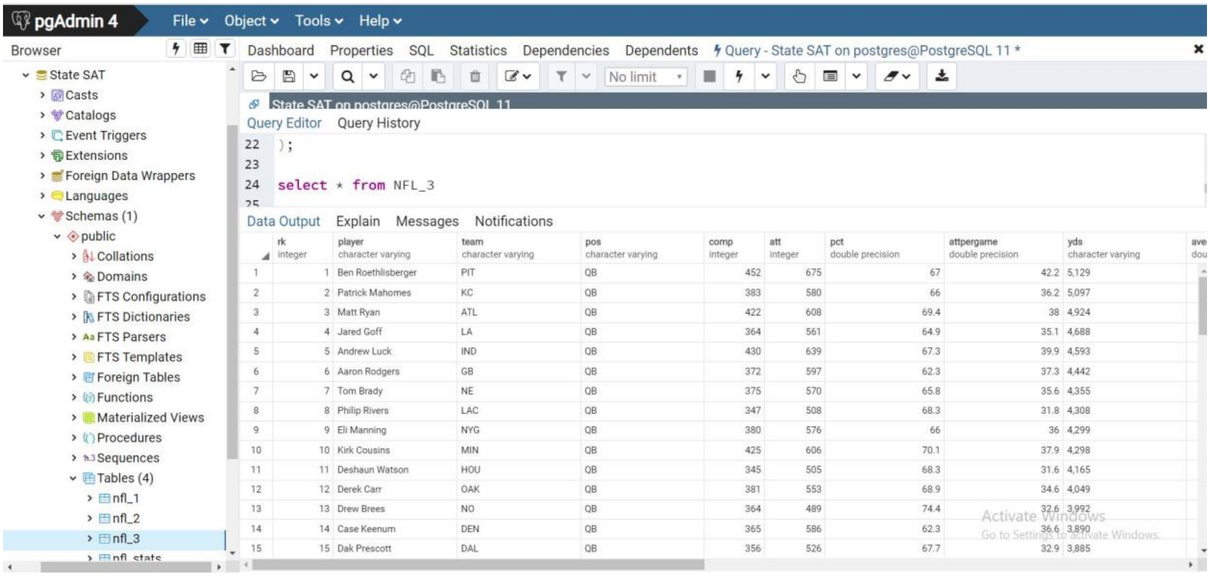
Fig-2(a): Creating a Table in SQL



```
Int int,
start_1st int,
start_1st_percent float,
Lng varchar,
Twenty_plus int,
Forty_plus int,
Sck int,
Rate float
);
```

Fig-2(b): Creating a Table in SQL

AIT-580 Data Analysis Individual Project
(Kausik Valeti, G01178711, Sec-004)

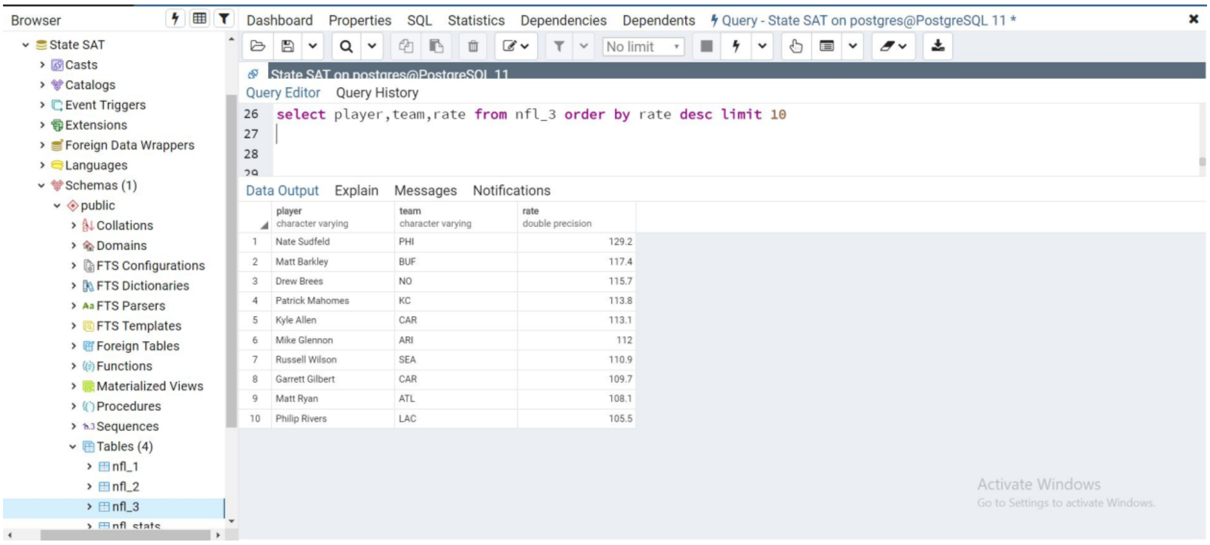


```
);

select * from NFL_3
```

	rk integer	player character varying	team character varying	pos character varying	comp integer	att integer	pct double precision	attpergame double precision	yds character varying	ave dou
1	1	Ben Roethlisberger	PIT	QB	452	675	67	42.2	5,129	
2	2	Patrick Mahomes	KC	QB	383	580	66	36.2	5,097	
3	3	Matt Ryan	ATL	QB	422	608	69.4	38	4,924	
4	4	Jared Goff	LA	QB	364	561	64.9	35.1	4,688	
5	5	Andrew Luck	IND	QB	430	639	67.3	39.9	4,593	
6	6	Aaron Rodgers	GB	QB	372	597	62.3	37.3	4,442	
7	7	Tom Brady	NE	QB	375	570	65.8	35.6	4,355	
8	8	Philip Rivers	LAC	QB	347	508	68.3	31.8	4,308	
9	9	Eli Manning	NYG	QB	380	576	66	36	4,299	
10	10	Kirk Cousins	MIN	QB	425	606	70.1	37.9	4,298	
11	11	Deshaun Watson	HOU	QB	345	505	68.3	31.6	4,165	
12	12	Derek Carr	OAK	QB	381	553	68.9	34.6	4,049	
13	13	Drew Brees	NO	QB	364	489	74.4	32.6	3,992	
14	14	Case Keenum	DEN	QB	365	586	62.3	36.6	3,890	
15	15	Dak Prescott	DAL	QB	356	526	67.7	32.9	3,885	

Fig-3: Viewing the table in SQL

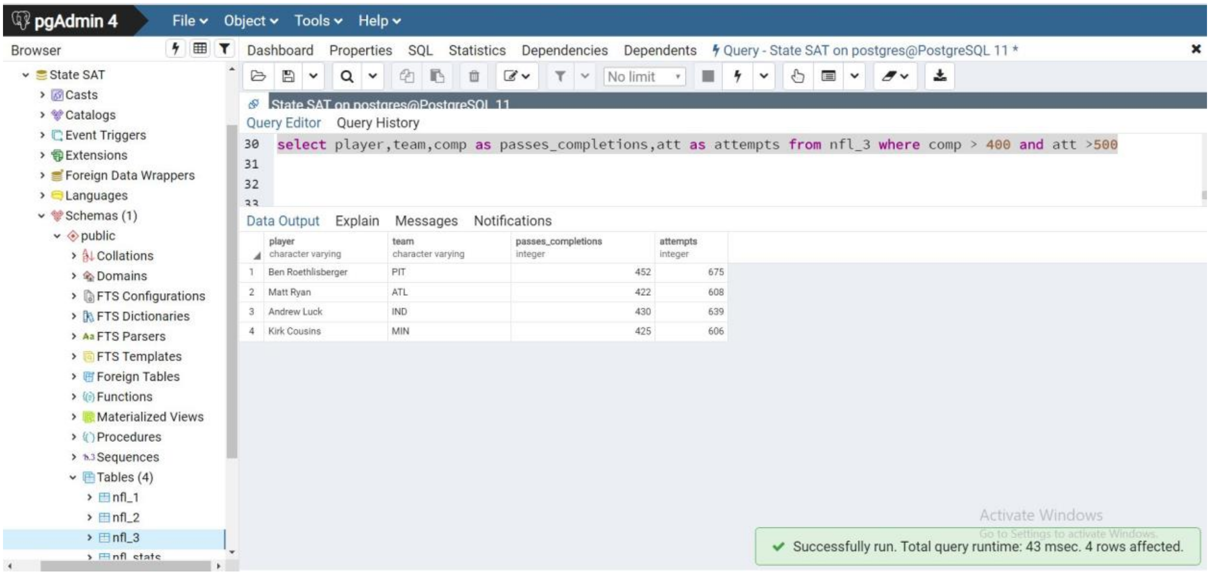


```
select player,team,rate from nfl_3 order by rate desc limit 10
```

	player character varying	team character varying	rate double precision
1	Nate Sudfeld	PHI	129.2
2	Matt Barkley	BUF	117.4
3	Drew Brees	NO	115.7
4	Patrick Mahomes	KC	113.8
5	Kyle Allen	CAR	113.1
6	Mike Glennon	ARI	112
7	Russell Wilson	SEA	110.9
8	Garrett Gilbert	CAR	109.7
9	Matt Ryan	ATL	108.1
10	Philip Rivers	LAC	105.5

Fig-4: Player’s with Top-10 rating

In the above graph we are viewing top 10 players with rating which are written in SQL query.



```
select player,team,comp as passes_completions,att as attempts from nfl_3 where comp > 400 and att >500
```

	player character varying	team character varying	passes_completions integer	attempts integer
1	Ben Roethlisberger	PIT	452	675
2	Matt Ryan	ATL	422	608
3	Andrew Luck	IND	430	639
4	Kirk Cousins	MIN	425	606

✓ Successfully run. Total query runtime: 43 msec. 4 rows affected.

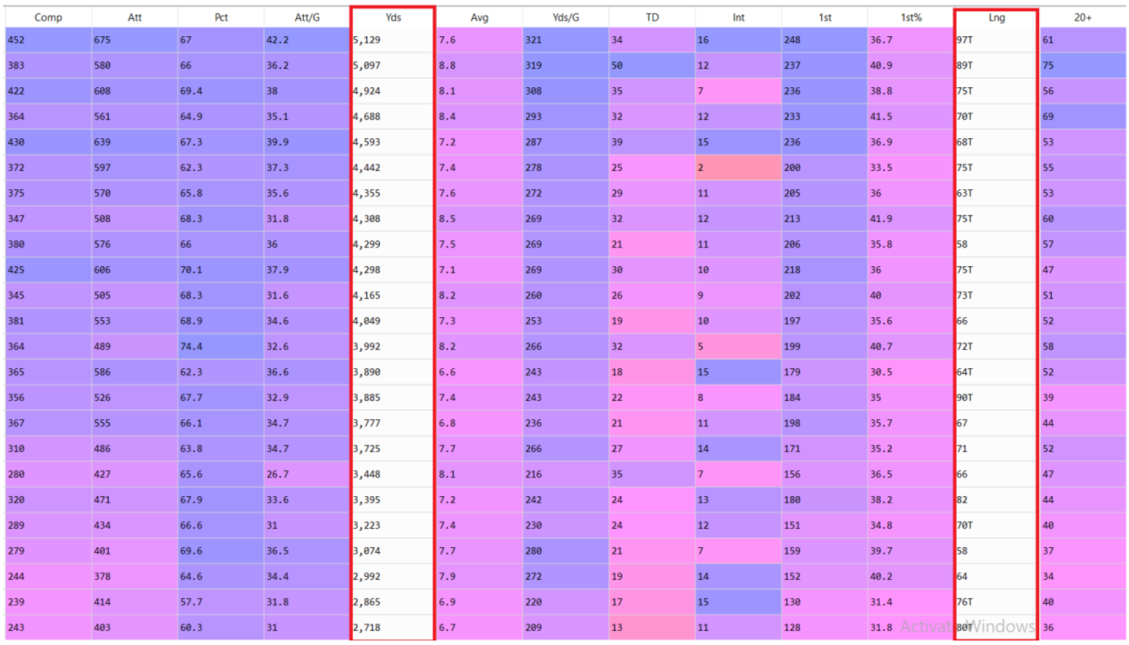
Fig-5: Player's with Pass Completion greater than 400 with more than 500 Attempts

From the SQL queries we have sorted such a way that sorting is easy in SQL when compared to Python and R. Because in other two languages we need to do if condition or loop of the data which is time taking but in SQL its just a single line of code makes simple and easier.

AIT-580 Data Analysis Individual Project
(Kausik Valeti, G01178711, Sec-004)

<u>Deliverable 2 – Data Analysis & interpretation Report:</u>

The second step that I did was cleaning the data, I have three options to go for whether SQL, Python or R. I would like to go for Python because data cleaning in Python is efficient and less time taking. Moreover, we have pandas library which can make our data looks better and easy to understand. In data cleaning part what I have done is removing the "commas" column called Yards. The reason for doing so is having comma in that column might make it as a character instead of numeric value. Same as in column called "Lng" which is Longest pass play replacing T after numeric value to make it number instead of character.

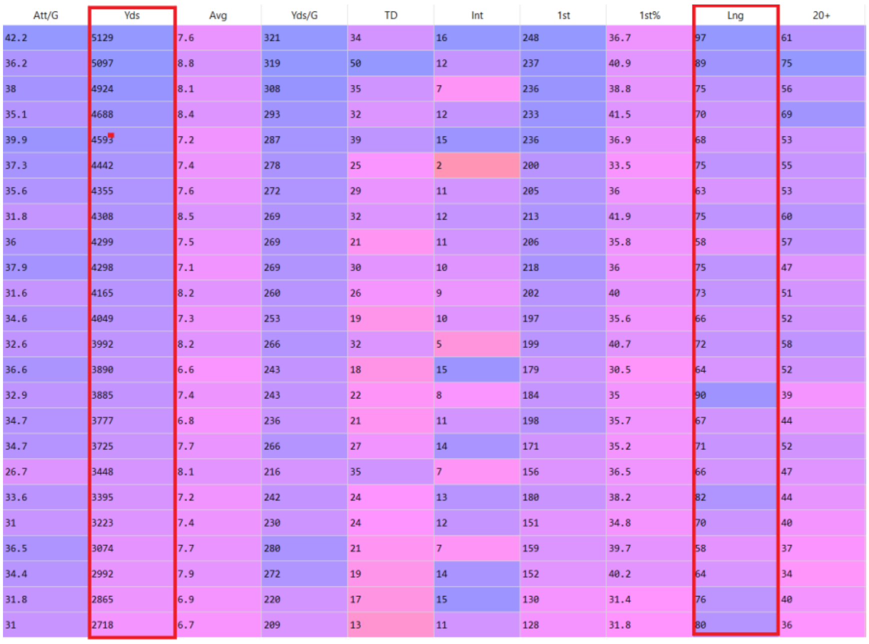


Comp	Att	Pct	Att/G	Yds	Avg	Yds/G	TD	Int	1st	1st%	Lng	20+
452	675	67	42.2	5,129	7.6	321	34	16	248	36.7	97T	61
383	580	66	36.2	5,097	8.8	319	50	12	237	40.9	89T	75
422	608	69.4	38	4,924	8.1	308	35	7	236	38.8	75T	56
364	561	64.9	35.1	4,688	8.4	293	32	12	233	41.5	70T	69
430	639	67.3	39.9	4,593	7.2	287	39	15	236	36.9	68T	53
372	597	62.3	37.3	4,442	7.4	278	25	2	200	33.5	75T	55
375	570	65.8	35.6	4,355	7.6	272	29	11	205	36	63T	53
347	508	68.3	31.8	4,308	8.5	269	32	12	213	41.9	75T	60
380	576	66	36	4,299	7.5	269	21	11	206	35.8	58	57
425	606	70.1	37.9	4,298	7.1	269	30	10	218	36	75T	47
345	505	68.3	31.6	4,165	8.2	260	26	9	202	40	73T	51
381	553	68.9	34.6	4,049	7.3	253	19	10	197	35.6	66	52
364	489	74.4	32.6	3,992	8.2	266	32	5	199	40.7	72T	58
365	586	62.3	36.6	3,890	6.6	243	18	15	179	30.5	64T	52
356	526	67.7	32.9	3,885	7.4	243	22	8	184	35	90T	39
367	555	66.1	34.7	3,777	6.8	236	21	11	198	35.7	67	44
310	486	63.8	34.7	3,725	7.7	266	27	14	171	35.2	71	52
280	427	65.6	26.7	3,448	8.1	216	35	7	156	36.5	66	47
320	471	67.9	33.6	3,395	7.2	242	24	13	180	38.2	82	44
289	434	66.6	31	3,223	7.4	230	24	12	151	34.8	70T	40
279	401	69.6	36.5	3,074	7.7	280	21	7	159	39.7	58	37
244	378	64.6	34.4	2,992	7.9	272	19	14	152	40.2	64	34
239	414	57.7	31.8	2,865	6.9	220	17	15	130	31.4	76T	40
243	403	60.3	31	2,718	6.7	209	13	11	128	31.8	80T	36

Activate Windows

Fig-6: Before modification of data in Python

After modification of data it looks as below

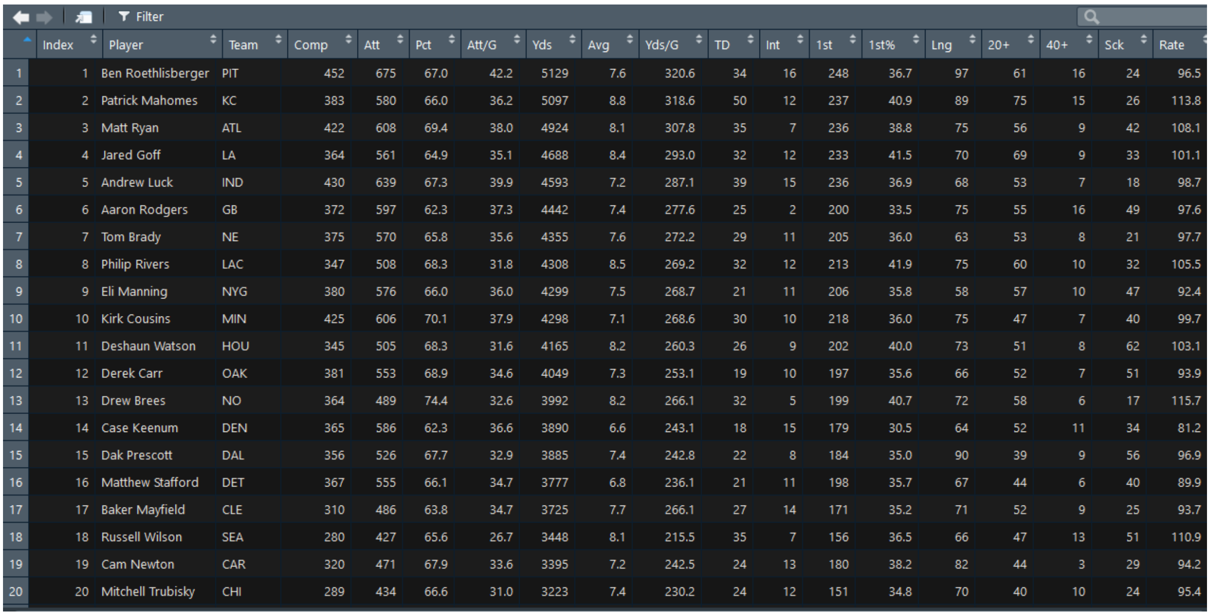


Att/G	Yds	Avg	Yds/G	TD	Int	1st	1st%	Lng	20+
42.2	5129	7.6	321	34	16	248	36.7	97	61
36.2	5097	8.8	319	50	12	237	40.9	89	75
38	4924	8.1	308	35	7	236	38.8	75	56
35.1	4688	8.4	293	32	12	233	41.5	70	69
39.9	4593	7.2	287	39	15	236	36.9	68	53
37.3	4442	7.4	278	25	2	200	33.5	75	55
35.6	4355	7.6	272	29	11	205	36	63	53
31.8	4308	8.5	269	32	12	213	41.9	75	60
36	4299	7.5	269	21	11	206	35.8	58	57
37.9	4298	7.1	269	30	10	218	36	75	47
31.6	4165	8.2	260	26	9	202	40	73	51
34.6	4049	7.3	253	19	10	197	35.6	66	52
32.6	3992	8.2	266	32	5	199	40.7	72	58
36.6	3890	6.6	243	18	15	179	30.5	64	52
32.9	3885	7.4	243	22	8	184	35	90	39
34.7	3777	6.8	236	21	11	198	35.7	67	44
34.7	3725	7.7	266	27	14	171	35.2	71	52
26.7	3448	8.1	216	35	7	156	36.5	66	47
33.6	3395	7.2	242	24	13	180	38.2	82	44
31	3223	7.4	230	24	12	151	34.8	70	40
36.5	3074	7.7	280	21	7	159	39.7	58	37
34.4	2992	7.9	272	19	14	152	40.2	64	34
31.8	2865	6.9	220	17	15	130	31.4	76	40
31	2718	6.7	209	13	11	128	31.8	80	36

Fig-7: After modification of data in Python

After modifing data I moved into R, the purpose of doing in R is I want to cluster the players according to the stats where they are and how each player differs from one for particular stat. In R I am using K-means model to identify the clusters of players which are similar and how they are classified. In addition to that we use Elbow method for clustering the players. Number of clusters can be identified from the Elbow method. We clustered the players using K-means model with scatter plots. The following graphs explains as follows:

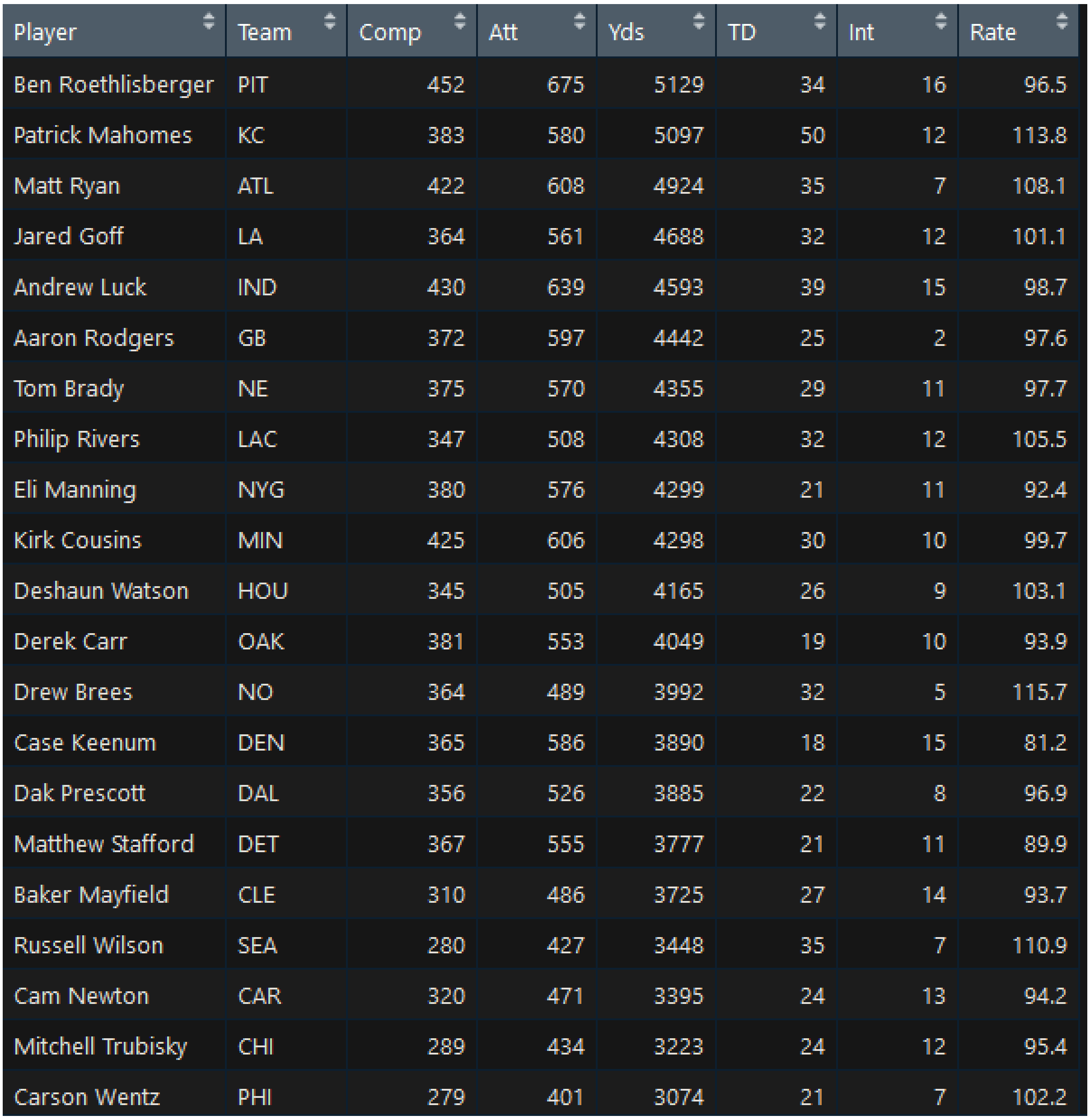
Dataset in R looks as follows:



	Index	Player	Team	Comp	Att	Pct	Att/G	Yds	Avg	Yds/G	TD	Int	1st	1st%	Lng	20+	40+	Sck	Rate
1	1	Ben Roethlisberger	PIT	452	675	67.0	42.2	5129	7.6	320.6	34	16	248	36.7	97	61	16	24	96.5
2	2	Patrick Mahomes	KC	383	580	66.0	36.2	5097	8.8	318.6	50	12	237	40.9	89	75	15	26	113.8
3	3	Matt Ryan	ATL	422	608	69.4	38.0	4924	8.1	307.8	35	7	236	38.8	75	56	9	42	108.1
4	4	Jared Goff	LA	364	561	64.9	35.1	4688	8.4	293.0	32	12	233	41.5	70	69	9	33	101.1
5	5	Andrew Luck	IND	430	639	67.3	39.9	4593	7.2	287.1	39	15	236	36.9	68	53	7	18	98.7
6	6	Aaron Rodgers	GB	372	597	62.3	37.3	4442	7.4	277.6	25	2	200	33.5	75	55	16	49	97.6
7	7	Tom Brady	NE	375	570	65.8	35.6	4355	7.6	272.2	29	11	205	36.0	63	53	8	21	97.7
8	8	Philip Rivers	LAC	347	508	68.3	31.8	4308	8.5	269.2	32	12	213	41.9	75	60	10	32	105.5
9	9	Eli Manning	NYG	380	576	66.0	36.0	4299	7.5	268.7	21	11	206	35.8	58	57	10	47	92.4
10	10	Kirk Cousins	MIN	425	606	70.1	37.9	4298	7.1	268.6	30	10	218	36.0	75	47	7	40	99.7
11	11	Deshaun Watson	HOU	345	505	68.3	31.6	4165	8.2	260.3	26	9	202	40.0	73	51	8	62	103.1
12	12	Derek Carr	OAK	381	553	68.9	34.6	4049	7.3	253.1	19	10	197	35.6	66	52	7	51	93.9
13	13	Drew Brees	NO	364	489	74.4	32.6	3992	8.2	266.1	32	5	199	40.7	72	58	6	17	115.7
14	14	Case Keenum	DEN	365	586	62.3	36.6	3890	6.6	243.1	18	15	179	30.5	64	52	11	34	81.2
15	15	Dak Prescott	DAL	356	526	67.7	32.9	3885	7.4	242.8	22	8	184	35.0	90	39	9	56	96.9
16	16	Matthew Stafford	DET	367	555	66.1	34.7	3777	6.8	236.1	21	11	198	35.7	67	44	6	40	89.9
17	17	Baker Mayfield	CLE	310	486	63.8	34.7	3725	7.7	266.1	27	14	171	35.2	71	52	9	25	93.7
18	18	Russell Wilson	SEA	280	427	65.6	26.7	3448	8.1	215.5	35	7	156	36.5	66	47	13	51	110.9
19	19	Cam Newton	CAR	320	471	67.9	33.6	3395	7.2	242.5	24	13	180	38.2	82	44	3	29	94.2
20	20	Mitchell Trubisky	CHI	289	434	66.6	31.0	3223	7.4	230.2	24	12	151	34.8	70	40	10	24	95.4

Fig-8: Dataset in R

But we are not intrested in all the variables so we subset the data only for specific varaibles which are as follows



Player	Team	Comp	Att	Yds	TD	Int	Rate
Ben Roethlisberger	PIT	452	675	5129	34	16	96.5
Patrick Mahomes	KC	383	580	5097	50	12	113.8
Matt Ryan	ATL	422	608	4924	35	7	108.1
Jared Goff	LA	364	561	4688	32	12	101.1
Andrew Luck	IND	430	639	4593	39	15	98.7
Aaron Rodgers	GB	372	597	4442	25	2	97.6
Tom Brady	NE	375	570	4355	29	11	97.7
Philip Rivers	LAC	347	508	4308	32	12	105.5
Eli Manning	NYG	380	576	4299	21	11	92.4
Kirk Cousins	MIN	425	606	4298	30	10	99.7
Deshaun Watson	HOU	345	505	4165	26	9	103.1
Derek Carr	OAK	381	553	4049	19	10	93.9
Drew Brees	NO	364	489	3992	32	5	115.7
Case Keenum	DEN	365	586	3890	18	15	81.2
Dak Prescott	DAL	356	526	3885	22	8	96.9
Matthew Stafford	DET	367	555	3777	21	11	89.9
Baker Mayfield	CLE	310	486	3725	27	14	93.7
Russell Wilson	SEA	280	427	3448	35	7	110.9
Cam Newton	CAR	320	471	3395	24	13	94.2
Mitchell Trubisky	CHI	289	434	3223	24	12	95.4
Carson Wentz	PHI	279	401	3074	21	7	102.2

Fig-9: After Subsetting of original dataset in R

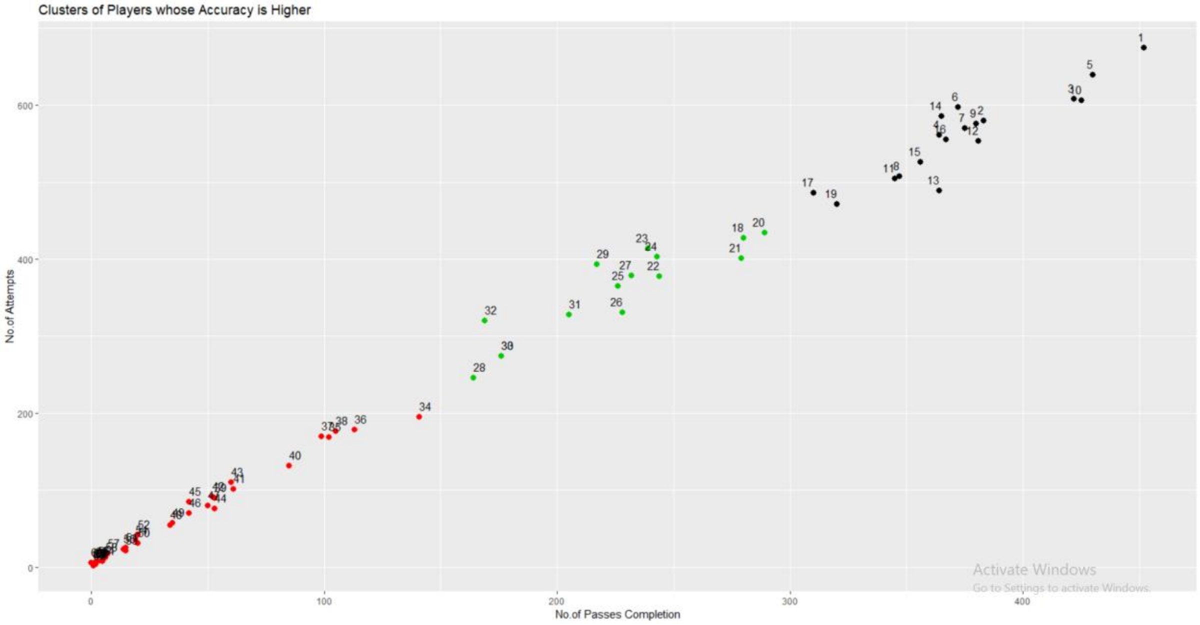


Fig-10: Cluster’s of Player’s whose Accuracy is Higher

From the above graph we can say that players who are in black colour has more accuracies when compared to the rest of them.

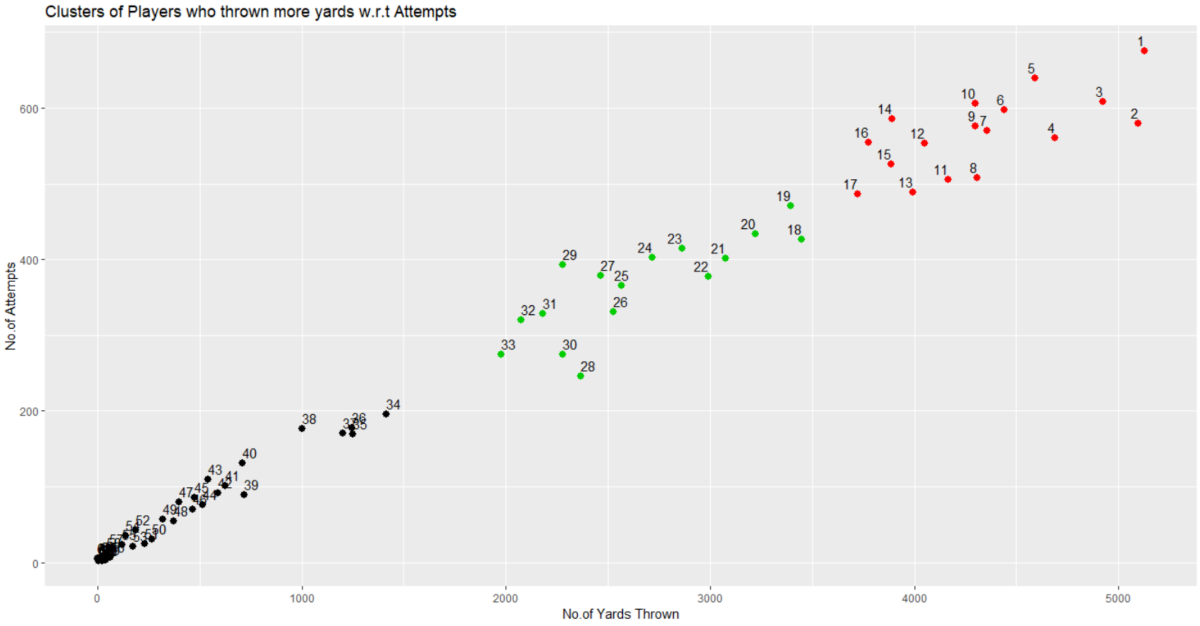


Fig-11: Cluster’s of Player’s who can throw more yard’s w.r.t Attempts

From the above graph we can say that players who are in Red colour can throw more yards for every attempt which makes them different from the others.

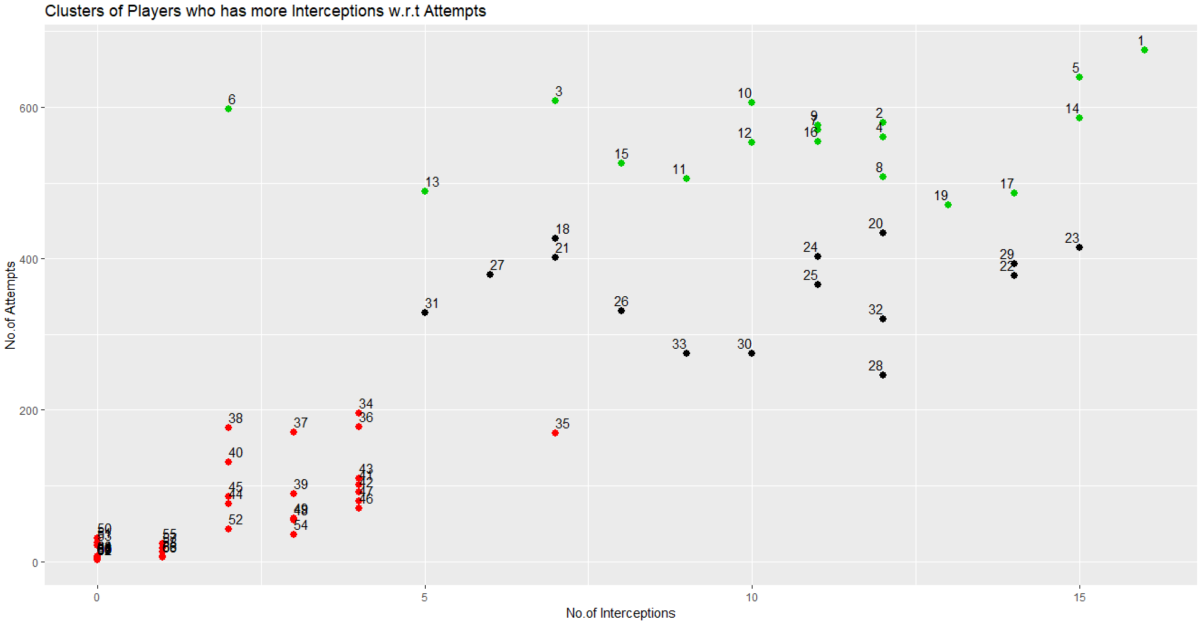


Fig-12: Cluster's of Player's who have ball more time with them

From the above graph player's who are in Green colour has the ball more time with them which can make them to score for their team and eventually a Win.

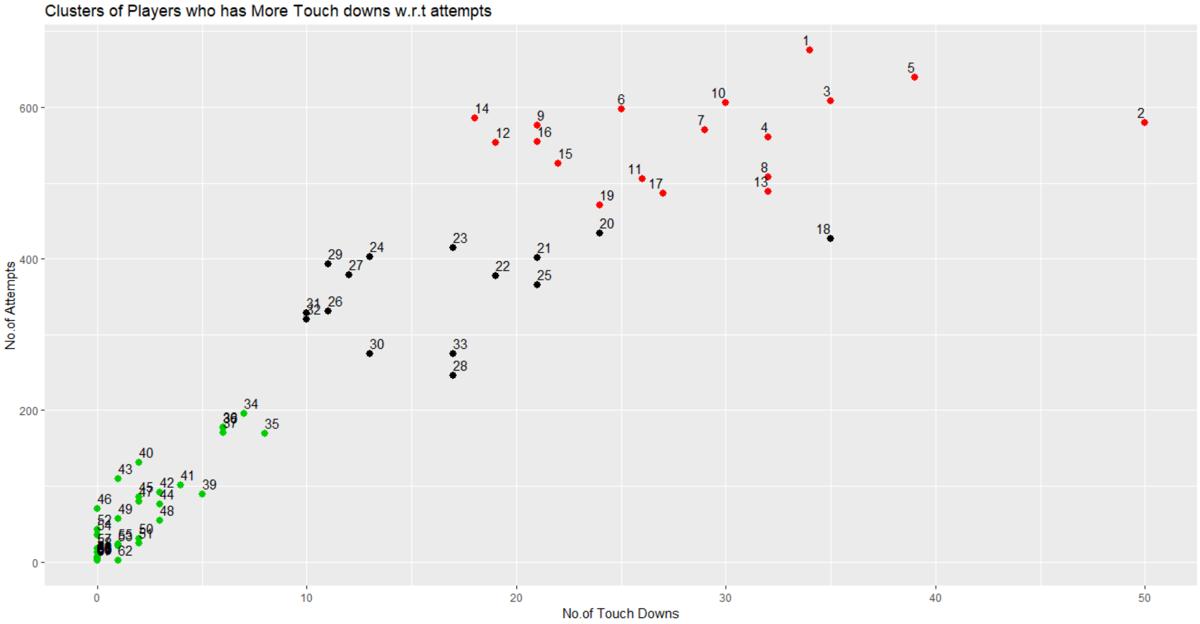


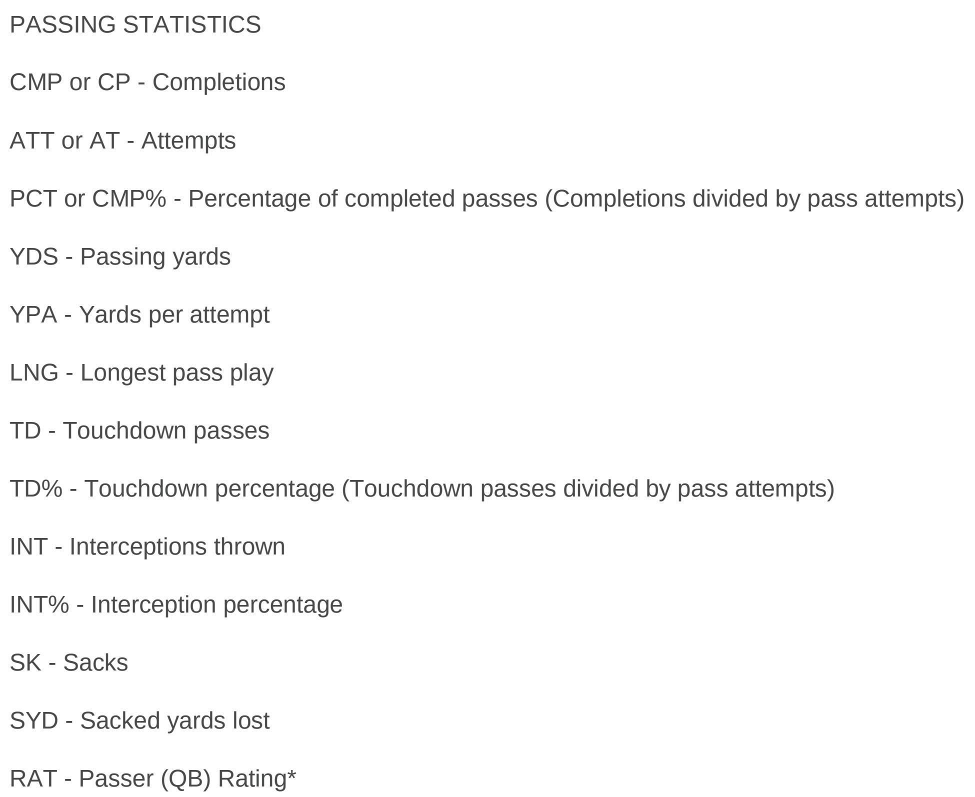
Fig-13: Cluster's of Player's who can score more points for their team

From the above graph, Player's with Red colour scores more points when compared with other quarter back's.

AIT-580 Data Analysis Individual Project
(Kausik Valeti, G01178711, Sec-004)

Interpreting the Results:

From the Results we can say that after clustering the players, there are certain criterias where the player's stand. The way the clustering was done makes the players distinct from each other and how they perform with the similar player's skills. Now its up to the head coach who want to pick a Quarter Back for the plans he make on how accurately he wants the player to throw ball long distance or in particular range.



PASSING STATISTICS

CMP or CP - Completions

ATT or AT - Attempts

PCT or CMP% - Percentage of completed passes (Completions divided by pass attempts)

YDS - Passing yards

YPA - Yards per attempt

LNG - Longest pass play

TD - Touchdown passes

TD% - Touchdown percentage (Touchdown passes divided by pass attempts)

INT - Interceptions thrown

INT% - Interception percentage

SK - Sacks

SYD - Sacked yards lost

RAT - Passer (QB) Rating*

Fig-14: NFL Passing Statistics glossary (NFL)

References

Forman, Sean. *Pro-Football-Reference.com*. 03 01 2003. 29 03 2019.

Gracenote. *Pro Football Reference*. 29 03 2019. 29 03 2019.

http://www.nfl.com/stats/categorystats?archive=false&conference=0011&statisticPositionCategory=QUARTERBACK&season=2018&seasonType=REG&experience=&tabSeq=1&qualified=true&Submit=Go. *NFL*. n.d. 04 05 2019.

NFL.

http://www.espn.com/nfl/news/story?id=2128923

NFL, ESPN. *NFL Statistics Glossary*. 28 08 2015. 04 05 2019.