Predicting Quarterback Success in the NFL-Data Wrangling

The data from this project was collected from multiple sources. The first dataset was NFL passing data from kaggle that was originally collected from pro-football-reference.com. This data set contained the years 2009-2018. Since the 2019 NFL season just finished, I went to the original source to collect the 2019 data and then concatenated all the data. Next, I did the same for college passing data from sports-reference.com/cfb for the same years. The data on this website was readily available in a csv format, therefore minimal effort was needed to obtain the data that was needed. This data was then inspected and cleaned.

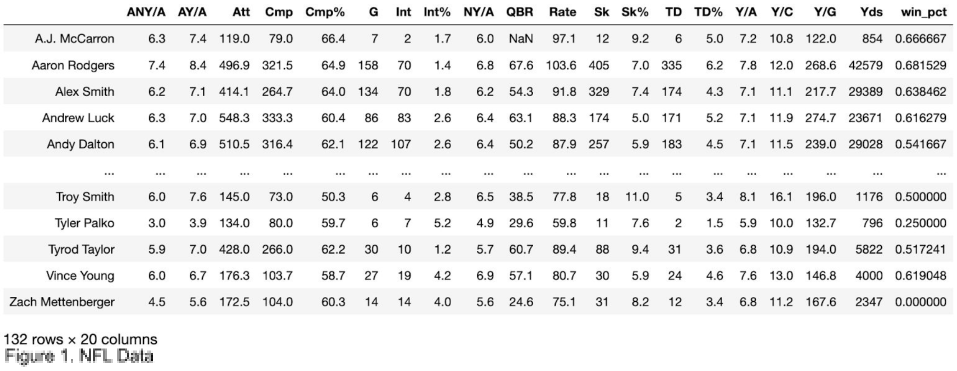
Data Cleaning

As mentioned above, the first thing done to the NFL data was concatenating it all into a single DataFrame. Once concatenated, the DataFrame was subsetted to only contain quarterbacks, as the dataset included statistics for players that are not listed as a quarterback that had thrown a pass that season, and also subsetted for players that had more than 100 passing attempts. The data was formatted by season therefore there were duplicate names for players that played in multiple seasons. The “Player” columns of these datasets contained different entries that included the following: “Name*\PlayerCode”, “Name/PlayerCode” and “Name*+\PlayerCode”. Since these player codes did not correspond to the player codes used in the college data, the Player column was split to only contain the player’s name.

In this DataFrame there is a “QBrec” column that has the starting record by season for each player. The entries in this column contained record data that was incorrectly put into a date format. This included “mm/dd/yyyy” and “mm-dd-yyyy.” The wins and losses were determined from these values and then a “win_pct” column was created.

The next step for this data was to use .groupby() aggregated by average to determine the career averages for each player. For columns such as ‘Int’, ‘TD’, ‘Yds’, ‘G’ and ‘Sk’, the career totals were used to replace the career seasonal average that was output by .groupby().

Once this was done, several columns that are not needed for the scope of this project were dropped. Only one value contained ‘NaN’ and that was the ‘QBR’ for A.J.McCarron. The NFL average was used to fill this value. This was done for simplicity and because it was only one value in a 132 rows X 20 columns DataFrame. The resulting DataFrame can be seen in Fig. 1.

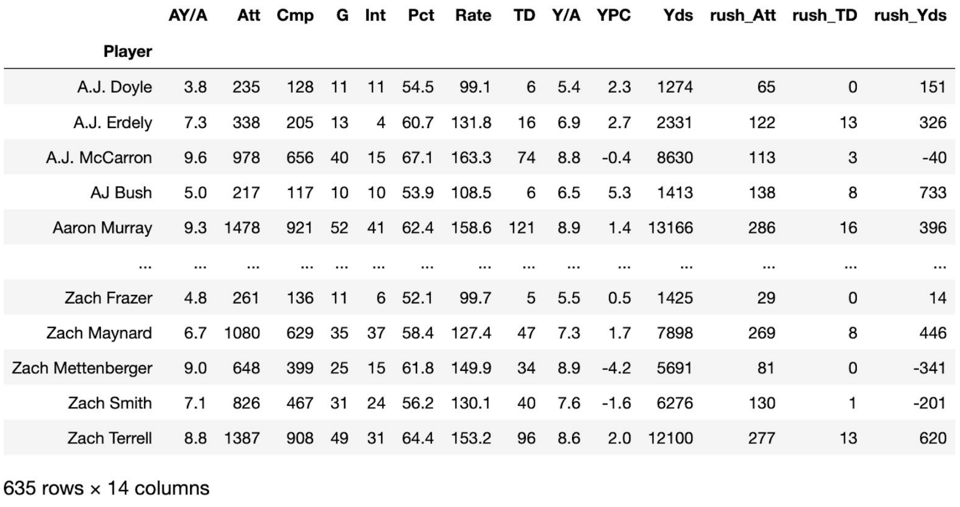


	ANY/A	AY/A	Att	Cmp	Cmp%	G	Int	Int%	NY/A	QBR	Rate	Sk	Sk%	TD	TD%	Y/A	Y/C	Y/G	Yds	win_pct
A.J. McCarron	6.3	7.4	119.0	79.0	66.4	7	2	1.7	6.0	NaN	97.1	12	9.2	6	5.0	7.2	10.8	122.0	854	0.666667
Aaron Rodgers	7.4	8.4	496.9	321.5	64.9	158	70	1.4	6.8	67.6	103.6	405	7.0	335	6.2	7.8	12.0	268.6	42579	0.681529
Alex Smith	6.2	7.1	414.1	264.7	64.0	134	70	1.8	6.2	54.3	91.8	329	7.4	174	4.3	7.1	11.1	217.7	29389	0.638462
Andrew Luck	6.3	7.0	548.3	333.3	60.4	86	83	2.6	6.4	63.1	88.3	174	5.0	171	5.2	7.1	11.9	274.7	23671	0.616279
Andy Dalton	6.1	6.9	510.5	316.4	62.1	122	107	2.6	6.4	50.2	87.9	257	5.9	183	4.5	7.1	11.5	239.0	29028	0.541667
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Troy Smith	6.0	7.6	145.0	73.0	50.3	6	4	2.8	6.5	38.5	77.8	18	11.0	5	3.4	8.1	16.1	196.0	1176	0.500000
Tyler Palko	3.0	3.9	134.0	80.0	59.7	6	7	5.2	4.9	29.6	59.8	11	7.6	2	1.5	5.9	10.0	132.7	796	0.250000
Tyrod Taylor	5.9	7.0	428.0	266.0	62.2	30	10	1.2	5.7	60.7	89.4	88	9.4	31	3.6	6.8	10.9	194.0	5822	0.517241
Vince Young	6.0	6.7	176.3	103.7	58.7	27	19	4.2	6.9	57.1	80.7	30	5.9	24	4.6	7.6	13.0	146.8	4000	0.619048
Zach Mettenberger	4.5	5.6	172.5	104.0	60.3	14	14	4.0	5.6	24.6	75.1	31	8.2	12	3.4	6.8	11.2	167.6	2347	0.000000

132 rows × 20 columns

Figure 1. NFL Data

For the college passing data, the data was concatenated into a single DataFrame and as with the NFL players, the college players had to be split from their player code. Like the NFL data, the necessary career totals were determined and this replaced the career averages calculated from using .groupby() aggregated by average. The resulting DataFrame can be seein in Fig. 2.



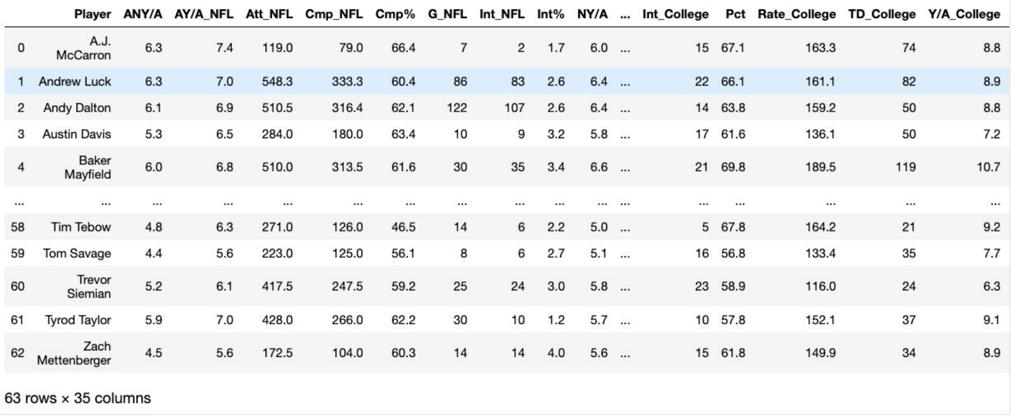
Player	AY/A	Att	Cmp	G	Int	Pct	Rate	TD	Y/A	YPC	Yds	rush_Att	rush_TD	rush_Yds
A.J. Doyle	3.8	235	128	11	11	54.5	99.1	6	5.4	2.3	1274	65	0	151
A.J. Erdely	7.3	338	205	13	4	60.7	131.8	16	6.9	2.7	2331	122	13	326
A.J. McCarron	9.6	978	656	40	15	67.1	163.3	74	8.8	-0.4	8630	113	3	-40
AJ Bush	5.0	217	117	10	10	53.9	108.5	6	6.5	5.3	1413	138	8	733
Aaron Murray	9.3	1478	921	52	41	62.4	158.6	121	8.9	1.4	13166	286	16	396
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Zach Frazer	4.8	261	136	11	6	52.1	99.7	5	5.5	0.5	1425	29	0	14
Zach Maynard	6.7	1080	629	35	37	58.4	127.4	47	7.3	1.7	7898	269	8	446
Zach Mettenberger	9.0	648	399	25	15	61.8	149.9	34	8.9	-4.2	5691	81	0	-341
Zach Smith	7.1	826	467	31	24	56.2	130.1	40	7.6	-1.6	6276	130	1	-201
Zach Terrell	8.8	1387	908	49	31	64.4	153.2	96	8.6	2.0	12100	277	13	620

635 rows × 14 columns

Figure 2. College Data

After creating these two DataFrames, they were exported into a csv file. In a separate Jupyter Notebook, an inner merge was performed on the “Player” column to obtain a DataFrame that consists of all the quarterbacks that were drafter between 2009 and 2018.

This DataFrame can be seen in Figure 3 and includes the NFL and college statistics of those 63 players. The columns with the same names were duplicated to depict which ones were for their NFL and College careers, respectively.



	Player	ANY/A	AY/A_NFL	Att_NFL	Cmp_NFL	Cmp%	G_NFL	Int_NFL	Int%	NY/A	...	Int_College	Pct	Rate_College	TD_College	Y/A_College
0	A.J. McCarron	6.3	7.4	119.0	79.0	66.4	7	2	1.7	6.0	...	15	67.1	163.3	74	8.8
1	Andrew Luck	6.3	7.0	548.3	333.3	60.4	86	83	2.6	6.4	...	22	66.1	161.1	82	8.9
2	Andy Dalton	6.1	6.9	510.5	316.4	62.1	122	107	2.6	6.4	...	14	63.8	159.2	50	8.8
3	Austin Davis	5.3	6.5	284.0	180.0	63.4	10	9	3.2	5.8	...	17	61.6	136.1	50	7.2
4	Baker Mayfield	6.0	6.8	510.0	313.5	61.6	30	35	3.4	6.6	...	21	69.8	189.5	119	10.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
58	Tim Tebow	4.8	6.3	271.0	126.0	46.5	14	6	2.2	5.0	...	5	67.8	164.2	21	9.2
59	Tom Savage	4.4	5.6	223.0	125.0	56.1	8	6	2.7	5.1	...	16	56.8	133.4	35	7.7
60	Trevor Siemian	5.2	6.1	417.5	247.5	59.2	25	24	3.0	5.8	...	23	58.9	116.0	24	6.3
61	Tyrod Taylor	5.9	7.0	428.0	266.0	62.2	30	10	1.2	5.7	...	10	57.8	152.1	37	9.1
62	Zach Mettenberger	4.5	5.6	172.5	104.0	60.3	14	14	4.0	5.6	...	15	61.8	149.9	34	8.9

63 rows × 35 columns

Figure 3. NFL and College Data of Quarterbacks drafted between 2009 and 2018

These three datasets will be useful for the scope of this project and should contain all the necessary data to determine the significant data in creating a model that predicts quarterback success in the NFL. If additional data is needed, it will most likely be data calculated from what is already in these datasets.