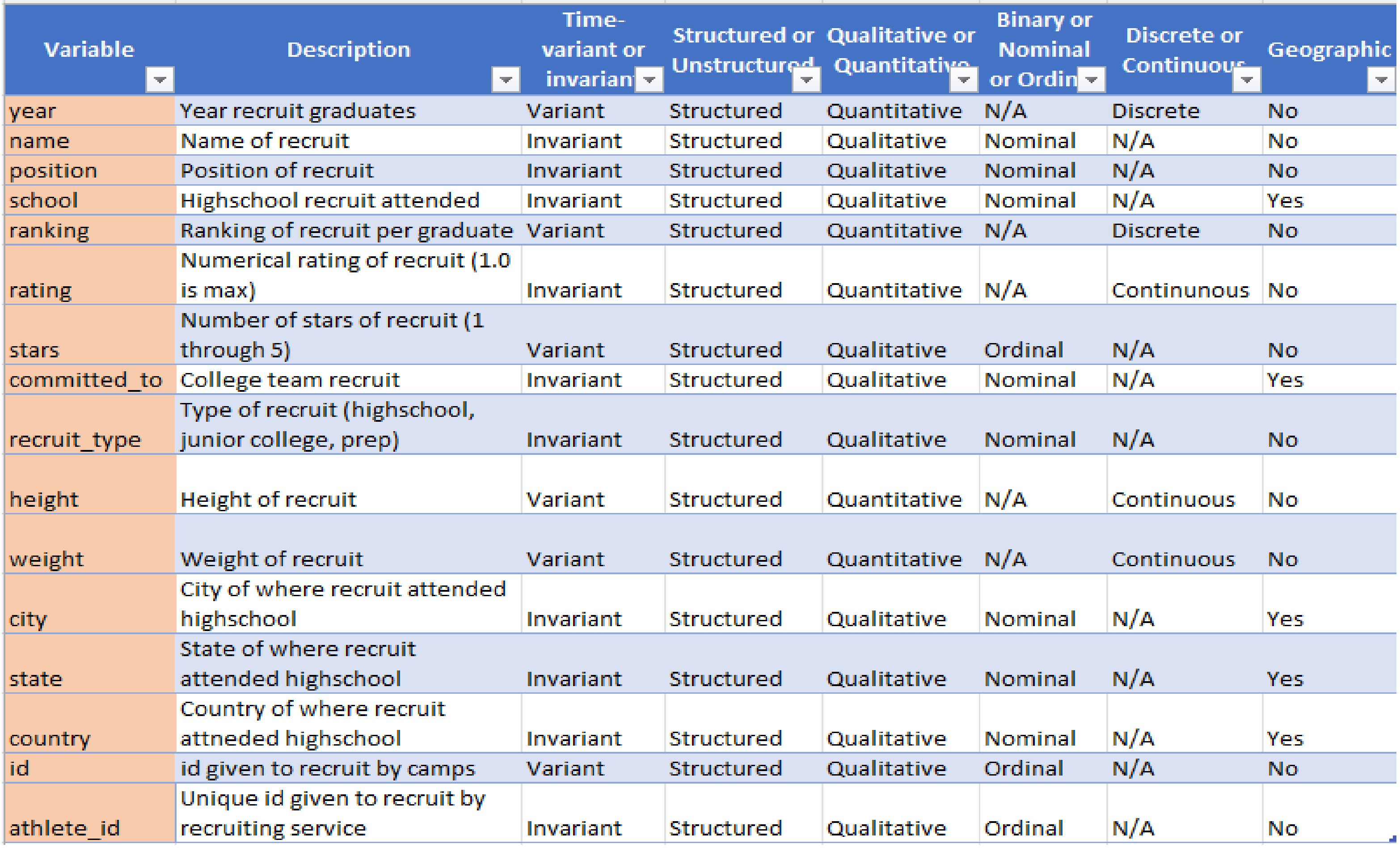
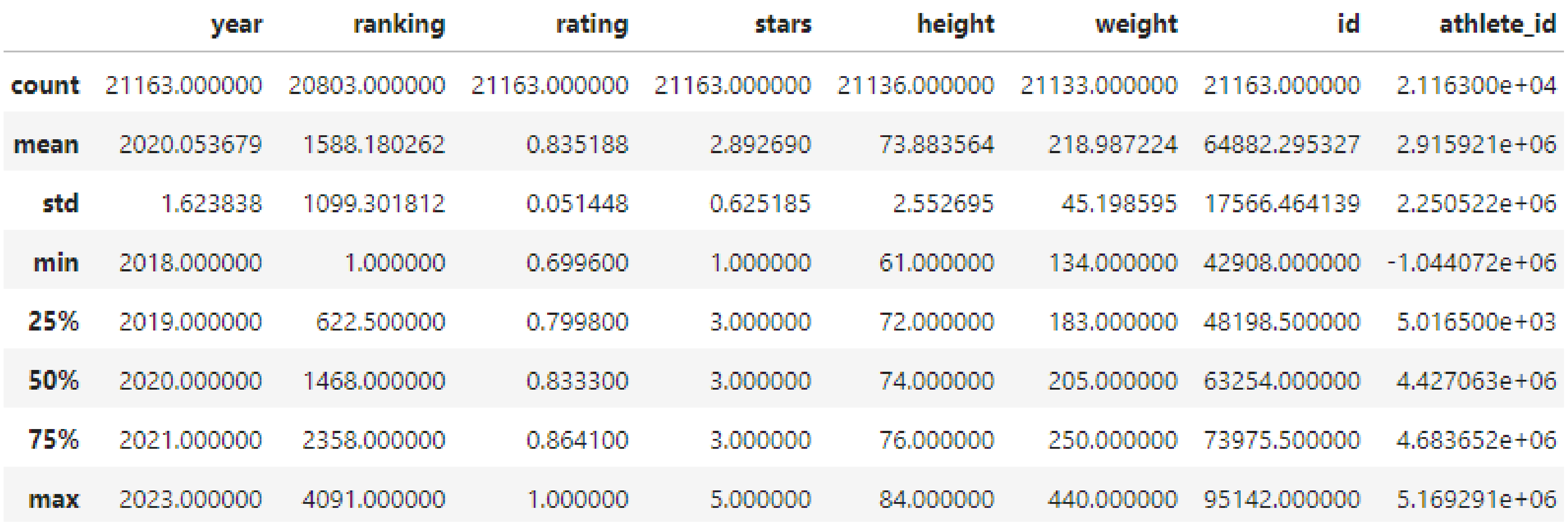
College Football Recruits Data

- This is external data collected by college football recruiting companies that are considered the industry standard. This data uses the 247 composite which is average player ranking of the three main recruiting companies (247, rivals and espn).
- This is administrative data collected by employees and represents the years 2018 through 2023. This data is updated quarterly.
- This data is grouped by year the player graduates, name, position, high school, rating ranking and stars, team committed to, recruit type, height and weight, city state and country, player id and athlete id.
- The main component of this data is manually collected including height, weight, year, name, position, school, committed to, recruit type, city, state, country, id, and athlete id. The rating is then created by industry experts by compiling game film, in game statistics, measurables (height, weight, arm length, etc.), and in person camps. Although this data can be seen as subjective it is considered the industry standard and can be seen as reliable and unbiased.
- Height and weight can be prone to large differences throughout the recruiting cycle. In most cases measurements are taken at camps hosted by recruiting sites or on college campuses and seen as reliable. However, not every player attends camps. In this case the high school the player attends or even the player himself can verify their own measurements. This can lead to bias as a player may exaggerate their measurements to achieve a higher rating.
- This data is very relevant to the project as the main questions center around recruiting rankings.

Recruits Data Profile



Variable	Description	Time-variant or invarian	Structured or Unstructured	Qualitative or Quantitative	Binary or Nominal or Ordin	Discrete or Continuous	Geographic
year	Year recruit graduates	Variant	Structured	Quantitative	N/A	Discrete	No
name	Name of recruit	Invariant	Structured	Qualitative	Nominal	N/A	No
position	Position of recruit	Invariant	Structured	Qualitative	Nominal	N/A	No
school	Highschool recruit attended	Invariant	Structured	Qualitative	Nominal	N/A	Yes
ranking	Ranking of recruit per graduate	Variant	Structured	Quantitative	N/A	Discrete	No
rating	Numerical rating of recruit (1.0 is max)	Invariant	Structured	Quantitative	N/A	Continunous	No
stars	Number of stars of recruit (1 through 5)	Variant	Structured	Qualitative	Ordinal	N/A	No
committed_to	College team recruit	Invariant	Structured	Qualitative	Nominal	N/A	Yes
recruit_type	Type of recruit (highschool, junior college, prep)	Invariant	Structured	Qualitative	Nominal	N/A	No
height	Height of recruit	Variant	Structured	Quantitative	N/A	Continuous	No
weight	Weight of recruit	Variant	Structured	Quantitative	N/A	Continuous	No
city	City of where recruit attended highschool	Invariant	Structured	Qualitative	Nominal	N/A	Yes
state	State of where recruit attended highschool	Invariant	Structured	Qualitative	Nominal	N/A	Yes
country	Country of where recruit attneded highschool	Invariant	Structured	Qualitative	Nominal	N/A	Yes
id	id given to recruit by camps	Variant	Structured	Qualitative	Ordinal	N/A	No
athlete_id	Unique id given to recruit by recruiting service	Invariant	Structured	Qualitative	Ordinal	N/A	No

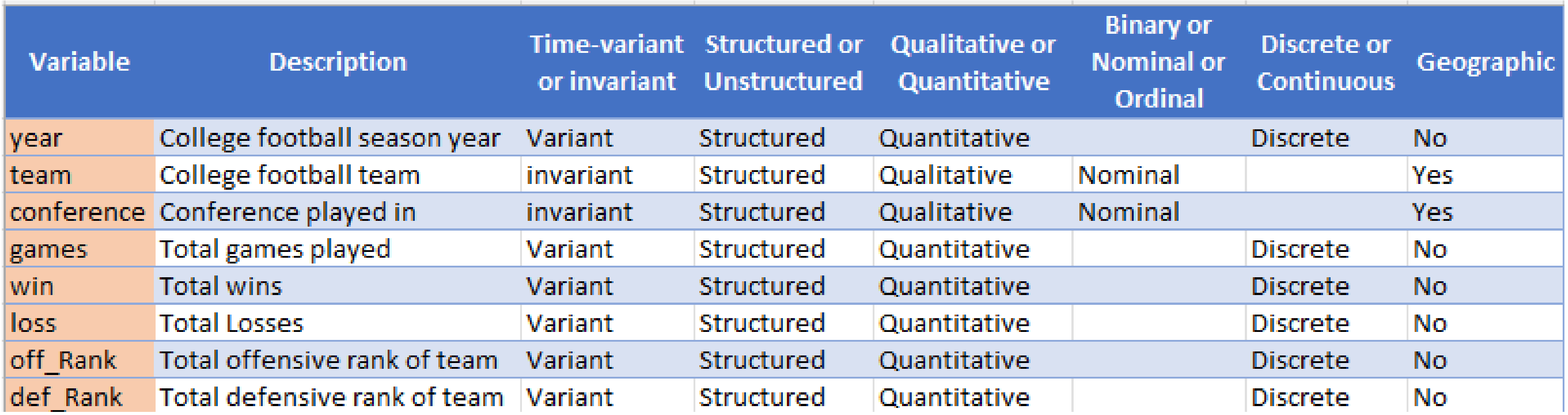


	year	ranking	rating	stars	height	weight	id	athlete_id
count	21163.000000	20803.000000	21163.000000	21163.000000	21136.000000	21133.000000	21163.000000	2.116300e+04
mean	2020.053679	1588.180262	0.835188	2.892690	73.883564	218.987224	64882.295327	2.915921e+06
std	1.623838	1099.301812	0.051448	0.625185	2.552695	45.198595	17566.464139	2.250522e+06
min	2018.000000	1.000000	0.699600	1.000000	61.000000	134.000000	42908.000000	-1.044072e+06
25%	2019.000000	622.500000	0.799800	3.000000	72.000000	183.000000	48198.500000	5.016500e+03
50%	2020.000000	1468.000000	0.833300	3.000000	74.000000	205.000000	63254.000000	4.427063e+06
75%	2021.000000	2358.000000	0.864100	3.000000	76.000000	250.000000	73975.500000	4.683652e+06
max	2023.000000	4091.000000	1.000000	5.000000	84.000000	440.000000	95142.000000	5.169291e+06

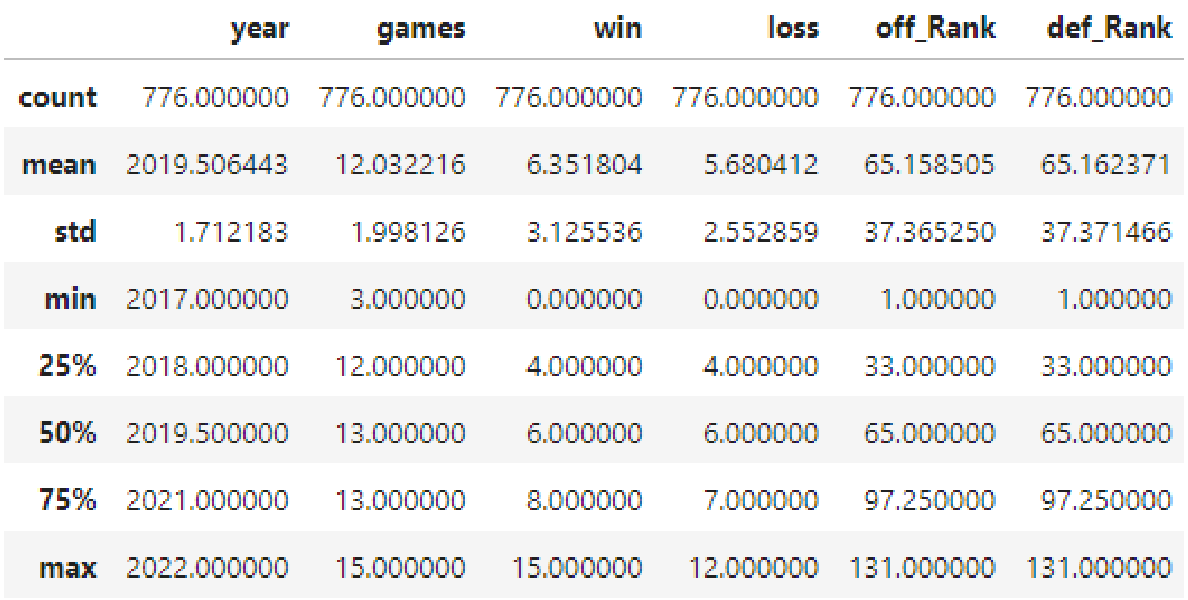
College Football Team Statistics

- This is internal data collected by the NCAA and listed on their website. It is manually collected for every game. It would be considered a trustworthy source.
- This data represents the years from 2017 through 2022 as that is the last full year of college football with available data.
- This data is grouped by team, conference, year, wins and losses, along with offence and defense ranking.
- This data is not biased, although it is possible there are errors in the offense and defense rankings. This could be due to errors in collecting data that is then compiled to create the offense and defense rankings.
- This data is very relevant to the project as win loss ratio and offense and defense ranking are a tangible way to attribute recruiting success in college football.

Team Statistics Data Profile

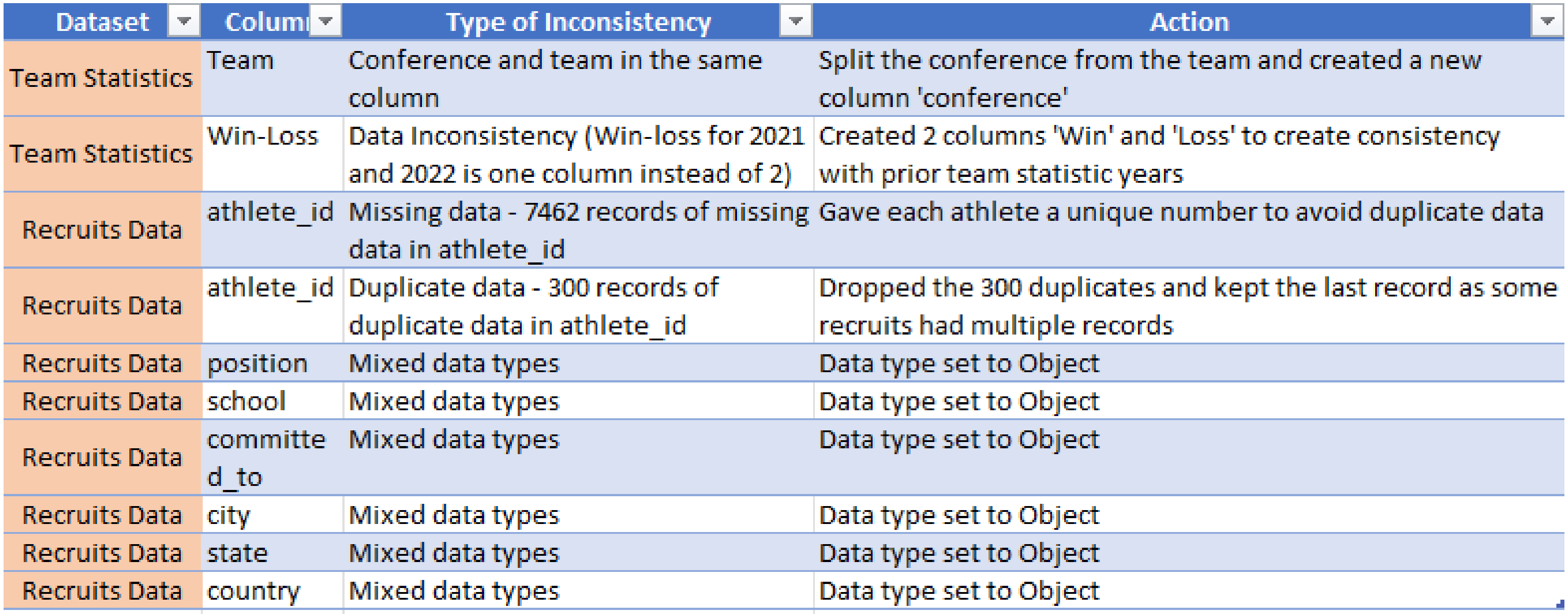


Variable	Description	Time-variant or invariant	Structured or Unstructured	Qualitative or Quantitative	Binary or Nominal or Ordinal	Discrete or Continuous	Geographic
year	College football season year	Variant	Structured	Quantitative		Discrete	No
team	College football team	invariant	Structured	Qualitative	Nominal		Yes
conference	Conference played in	invariant	Structured	Qualitative	Nominal		Yes
games	Total games played	Variant	Structured	Quantitative		Discrete	No
win	Total wins	Variant	Structured	Quantitative		Discrete	No
loss	Total Losses	Variant	Structured	Quantitative		Discrete	No
off_Rank	Total offensive rank of team	Variant	Structured	Quantitative		Discrete	No
def_Rank	Total defensive rank of team	Variant	Structured	Quantitative		Discrete	No



	year	games	win	loss	off_Rank	def_Rank
count	776.000000	776.000000	776.000000	776.000000	776.000000	776.000000
mean	2019.506443	12.032216	6.351804	5.680412	65.158505	65.162371
std	1.712183	1.998126	3.125536	2.552859	37.365250	37.371466
min	2017.000000	3.000000	0.000000	0.000000	1.000000	1.000000
25%	2018.000000	12.000000	4.000000	4.000000	33.000000	33.000000
50%	2019.500000	13.000000	6.000000	6.000000	65.000000	65.000000
75%	2021.000000	13.000000	8.000000	7.000000	97.250000	97.250000
max	2022.000000	15.000000	15.000000	12.000000	131.000000	131.000000

Cleaning Data



Dataset	Colum	Type of Inconsistency	Action
Team Statistics	Team	Conference and team in the same column	Split the conference from the team and created a new column 'conference'
Team Statistics	Win-Loss	Data Inconsistency (Win-loss for 2021 and 2022 is one column instead of 2)	Created 2 columns 'Win' and 'Loss' to create consistency with prior team statistic years
Recruits Data	athlete_id	Missing data - 7462 records of missing data in athlete_id	Gave each athlete a unique number to avoid duplicate data
Recruits Data	athlete_id	Duplicate data - 300 records of duplicate data in athlete_id	Dropped the 300 duplicates and kept the last record as some recruits had multiple records
Recruits Data	position	Mixed data types	Data type set to Object
Recruits Data	school	Mixed data types	Data type set to Object
Recruits Data	committed_to	Mixed data types	Data type set to Object
Recruits Data	city	Mixed data types	Data type set to Object
Recruits Data	state	Mixed data types	Data type set to Object
Recruits Data	country	Mixed data types	Data type set to Object