Project 2 = Big Game Census Analytics

Problem Statement:

This Big Game Census data visualization takes a fun look at where Super Bowl 52 players come from, the related population figures, and opens up pathways (via embedded links) to additional census data points.

The dataset came about when two hapless data nerds had their teams eliminated from the playoffs, thus turning to data to try to find more rooting interests for Super Bowl 52. The rosters for both, competing teams are included, with the corresponding roster information and birthplace and state population information. The developers utilized census data pulled from census.gov, and roster information from Yahoo Sports and designed the data visualization within the Tableau platform.

In [171…

```
# Import the necessary libararies
import pandas as pd
```

In [174…

```
# Load all the datasets
Data = pd.read_csv("/Users/nikhilreddyponnala/Desktop/Big Game Census Analyt
Data1 = pd.read_csv("/Users/nikhilreddyponnala/Desktop/Big Game Census Analy
Data2 = pd.read_csv("/Users/nikhilreddyponnala/Desktop/Big Game Census Analy
```

Display dataset structure

Data = All Places Census 2016 Population Estimates.csv Data1 = All states Census 2017 Population Estimates.csv Data2 = Big Game Census data.csv

In [177…

```
Data
```

Out[177]:

	Geographic ID	GEOID 2	Geography, full name (City, State)	April 1, 2010 - Census	April 1, 2010 - Estimates Base	Population Estimate (as of July 1) - 2010	Populati Estima (as of Ju 1) - 20
0	1620000US0100124	100124	Abbeville city, Alabama	2688	2688	2683	26
1	1620000US0100460	100460	Adamsville city, Alabama	4522	4522	4517	44
2	1620000US0100484	100484	Addison town, Alabama	758	756	754	7
3	1620000US0100676	100676	Akron town, Alabama	356	356	355	3
4	1620000US0100820	100820	Alabaster city, Alabama	30352	31066	31176	313
...	...	...	...	...	...	...	
19505	1620000US5681300	5681300	Wamsutter town, Wyoming	451	451	450	4
19506	1620000US5683040	5683040	Wheatland town, Wyoming	3627	3627	3629	36
19507	1620000US5684925	5684925	Worland city, Wyoming	5487	5487	5494	54
19508	1620000US5685015	5685015	Wright town, Wyoming	1807	1807	1807	18
19509	1620000US5686665	5686665	Yoder town, Wyoming	151	151	152	1

19510 rows × 12 columns

In [179…

```
Data1
```

Out[179]:

	GEOID	GEOID2	Geography Name	April 1, 2010 - Census	April 1, 2010 - Estimates Base	Population Estimate (as of July 1) - 2010	Population Estimate (as of July 1) - 2011
0	0400000US01	1	Alabama	4779736	4780135	4785579	4798649
1	0400000US02	2	Alaska	710231	710249	714015	722259
2	0400000US04	4	Arizona	6392017	6392309	6407002	6465488
3	0400000US05	5	Arkansas	2915918	2916031	2921737	2938640
4	0400000US06	6	California	37253956	37254518	37327690	37672654
5	0400000US08	8	Colorado	5029196	5029325	5048029	5116411
6	0400000US09	9	Connecticut	3574097	3574114	3580171	3591927
7	0400000US10	10	Delaware	897934	897936	899712	907884
8	0400000US11	11	District of Columbia	601723	601766	605040	620336
9	0400000US12	12	Florida	18801310	18804594	18846461	19097369
10	0400000US13	13	Georgia	9687653	9688690	9712696	9810595
11	0400000US15	15	Hawaii	1360301	1360301	1363817	1378323
12	0400000US16	16	Idaho	1567582	1567650	1570912	1583180
13	0400000US17	17	Illinois	12830632	12831565	12841196	12862298
14	0400000US18	18	Indiana	6483802	6484125	6490029	6515358
15	0400000US19	19	Iowa	3046355	3046869	3050223	3063690
16	0400000US20	20	Kansas	2853118	2853130	2858403	2868756
17	0400000US21	21	Kentucky	4339367	4339340	4347948	4368505
18	0400000US22	22	Louisiana	4533372	4533478	4544871	4574388
19	0400000US23	23	Maine	1328361	1328362	1327568	1327968
20	0400000US24	24	Maryland	5773552	5773807	5788099	5843115
21	0400000US25	25	Massachusetts	6547629	6547808	6564943	6612178
22	0400000US26	26	Michigan	9883640	9884129	9876731	9876199
23	0400000US27	27	Minnesota	5303925	5303924	5310711	5345967
24	0400000US28	28	Mississippi	2967297	2968103	2970437	2977452
25	0400000US29	29	Missouri	5988927	5988925	5995681	6010280
26	0400000US30	30	Montana	989415	989414	990507	996866
27	0400000US31	31	Nebraska	1826341	1826327	1829956	1841641
28	0400000US32	32	Nevada	2700551	2700691	2702797	2718170
29	0400000US33	33	New Hampshire	1316470	1316460	1316700	1318345
30	0400000US34	34	New Jersey	8791894	8791953	8803708	8844694
31	0400000US35	35	New Mexico	2059179	2059207	2064607	2077744
32	0400000US36	36	New York	19378102	19378110	19405185	19526372
33	0400000US37	37	North Carolina	9535483	9535721	9574247	9662940
34	0400000US38	38	North Dakota	672591	672585	674518	684830

	GEOID	GEOID2	Geography Name	April 1, 2010 - Census	April 1, 2010 - Estimates Base	Population Estimate (as of July 1) - 2010	Population Estimate (as of July 1) - 2011	P (
35	0400000US39	39	Ohio	11536504	11536730	11539282	11543332	
36	0400000US40	40	Oklahoma	3751351	3751598	3759529	3785232	
37	0400000US41	41	Oregon	3831074	3831072	3837073	3865845	
38	0400000US42	42	Pennsylvania	12702379	12702857	12711063	12742811	
39	0400000US44	44	Rhode Island	1052567	1052945	1053169	1052154	
40	0400000US45	45	South Carolina	4625364	4625381	4635834	4672744	
41	0400000US46	46	South Dakota	814180	814197	816227	823338	
42	0400000US47	47	Tennessee	6346105	6346295	6355882	6396281	
43	0400000US48	48	Texas	25145561	25146100	25241648	25644424	
44	0400000US49	49	Utah	2763885	2763889	2775260	2815430	
45	0400000US50	50	Vermont	625741	625741	625842	626210	
46	0400000US51	51	Virginia	8001024	8001043	8025206	8107548	
47	0400000US53	53	Washington	6724540	6724545	6741386	6819155	
48	0400000US54	54	West Virginia	1852994	1853006	1854315	1854891	
49	0400000US55	55	Wisconsin	5686986	5687288	5690403	5705812	
50	0400000US56	56	Wyoming	563626	563767	564376	567602	
51	0400000US72	72	Puerto Rico	3725789	3726157	3721525	3678732	

```
In [181…  Data2
```

Out[181]:

	Player Name	Player Jersey Number	Player Position	Player Age	Player Weight (lbs.)	Years Played	Player Birthplace (city, town, etc.)	Player Birth State	
0	Devin McCourty	#32	S/FS	30.0	195.0	8	Nanuet	New York	N
1	Danny Amendola	#80	PR/WR/KR	32.0	190.0	9	The Woodlands	Texas	
2	Johnson Bademosi	#29	CB/SPTM/RCB	27.0	206.0	6	Silver Spring	Maryland	
3	Chris Hogan	#15	WR	29.0	210.0	5	Wyckoff	New Jersey	
4	James Develin	#46	RB/FB	29.0	255.0	5	Gilbertsville	Pennsylvania	C P
...	...	...	...	...	...	...	...	...	
116	Kamu Grugier-Hill	#54	LB/SPTM/RLB	23.0	220.0	2	Honolulu	Hawaii	
117	Isaac Seumalo	#73	G/ROG	24.0	303.0	2	Honolulu	Hawaii	
118	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
119	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
120	Player data from Yahoo Sports	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

121 rows × 24 columns

In [183…

```
Data.isna().sum()
```

Out[183]:

```
Geographic ID                                0
GEOID 2                                      0
Geography, full name (City, State)           0
April 1, 2010 - Census                       0
April 1, 2010 - Estimates Base               0
Population Estimate (as of July 1) - 2010    0
Population Estimate (as of July 1) - 2011    0
Population Estimate (as of July 1) - 2012    0
Population Estimate (as of July 1) - 2013    0
Population Estimate (as of July 1) - 2014    0
Population Estimate (as of July 1) - 2015    0
Population Estimate (as of July 1) - 2016    0
dtype: int64
```

In [185…

```
Data1.isna().sum()
```

Out[185]:

```
GEOID                                        0
GEOID2                                       0
Geography Name                               0
April 1, 2010 - Census                       0
April 1, 2010 - Estimates Base               0
Population Estimate (as of July 1) - 2010    0
Population Estimate (as of July 1) - 2011    0
Population Estimate (as of July 1) - 2012    0
Population Estimate (as of July 1) - 2013    0
Population Estimate (as of July 1) - 2014    0
Population Estimate (as of July 1) - 2015    0
Population Estimate (as of July 1) - 2016    0
Population Estimate (as of July 1) - 2017    0
dtype: int64
```

In [187…

```
Data2.isna().sum()
```

Out[187]:

```
Player Name                                                 2
Player Jersey Number                                        3
Player Position                                             3
Player Age                                                  3
Player Weight (lbs.)                                        3
Years Played                                                3
Player Birthplace (city, town, etc.)                        3
Player Birth State                                          3
Player Birthplace (Combo)                                   3
Player College                                              3
Player Team                                                 3
Conference                                                  3
2016 Population Estimates (except where otherwise noted)    3
State GEO ID                                                3
Full GEOID                                                  3
Latitude (player birthplace)                                3
Longitude (player birthplace)                               3
Number from City                                            3
Number of Records                                           3
American FactFinder Link for more Census data points        3
Quickfacts Link                                             3
State Data Link                                             3
Source (Population States 2017)                             3
Birthplace, Population Data Source                          3
dtype: int64
```

Cleaning the dataset 'All Places Census 2016 Population Estimates.csv'

Check for missing values and fill/drop if necessary

In [191…

```
print(
    f"All Places Census:\n{Data.isna().sum()}\nAll States Census:\n{Data1.is
)
```

```
All Places Census:
Geographic ID                                0
GEOID 2                                      0
Geography, full name (City, State)           0
April 1, 2010 - Census                       0
April 1, 2010 - Estimates Base               0
Population Estimate (as of July 1) - 2010    0
Population Estimate (as of July 1) - 2011    0
Population Estimate (as of July 1) - 2012    0
Population Estimate (as of July 1) - 2013    0
Population Estimate (as of July 1) - 2014    0
Population Estimate (as of July 1) - 2015    0
Population Estimate (as of July 1) - 2016    0
dtype: int64
All States Census:
GEOID                                        0
GEOID2                                       0
Geography Name                               0
April 1, 2010 - Census                       0
April 1, 2010 - Estimates Base               0
Population Estimate (as of July 1) - 2010    0
Population Estimate (as of July 1) - 2011    0
Population Estimate (as of July 1) - 2012    0
Population Estimate (as of July 1) - 2013    0
Population Estimate (as of July 1) - 2014    0
Population Estimate (as of July 1) - 2015    0
Population Estimate (as of July 1) - 2016    0
Population Estimate (as of July 1) - 2017    0
dtype: int64
Big Game Census:
Player Name                                                 2
Player Jersey Number                                        3
Player Position                                             3
Player Age                                                  3
Player Weight (lbs.)                                        3
Years Played                                                3
Player Birthplace (city, town, etc.)                        3
Player Birth State                                          3
Player Birthplace (Combo)                                   3
Player College                                              3
Player Team                                                 3
Conference                                                  3
2016 Population Estimates (except where otherwise noted)    3
State GEO ID                                                3
Full GEOID                                                  3
Latitude (player birthplace)                                3
Longitude (player birthplace)                               3
Number from City                                            3
Number of Records                                           3
American FactFinder Link for more Census data points        3
Quickfacts Link                                             3
State Data Link                                             3
Source (Population States 2017)                             3
Birthplace, Population Data Source                          3
dtype: int64
```

Drop rows where essential columns have missing values

In [194…

```
Data.dropna(
    subset=["Geographic ID", "GEOID 2", "Geography, full name (City, State)"
    inplace=True,
)

Data1.dropna(
    subset=["Geography Name", "Population Estimate (as of July 1) - 2017"],
)

Data2.dropna(
    subset=[
        "Player Name",
        "Player Birthplace (city, town, etc.)",
        "Player Birth State",
        "Player College",
    ],
    inplace=True,
)
```

Check for duplicates and remove them

In [197…

```
Data.drop_duplicates(inplace=True)

Data1.drop_duplicates(inplace=True)

Data2.drop_duplicates(inplace=True)
```

Ensure proper data types

In [200…

```
Data["Geographic ID"] = Data["Geographic ID"].astype(str)
Data["GEOID 2"] = Data["GEOID 2"].astype(str)

Data1["Population Estimate (as of July 1) - 2017"] = Data1[
    "Population Estimate (as of July 1) - 2017"
].astype(int)

Data2["Number from City"] = Data2["Number from City"].astype(int)
Data2["Number of Records"] = Data2["Number of Records"].astype(int)
```

Display cleaned data

In [203…

```
Data.head()
```

Out[203]:

	Geographic ID	GEOID 2	Geography, full name (City, State)	April 1, 2010 - Census	April 1, 2010 - Estimates Base	Population Estimate (as of July 1) - 2010	Population Estimate (as of July 1) - 2011	Po E (a 1
0	1620000US0100124	100124	Abbeville city, Alabama	2688	2688	2683	2685	
1	1620000US0100460	100460	Adamsville city, Alabama	4522	4522	4517	4495	
2	1620000US0100484	100484	Addison town, Alabama	758	756	754	753	
3	1620000US0100676	100676	Akron town, Alabama	356	356	355	345	
4	1620000US0100820	100820	Alabaster city, Alabama	30352	31066	31176	31362	

In [205…

```
Data1.head()
```

Out[205]:

	GEOID	GEOID2	Geography Name	April 1, 2010 - Census	April 1, 2010 - Estimates Base	Population Estimate (as of July 1) - 2010	Population Estimate (as of July 1) - 2011	Popul Esti (as of 1) -
0	0400000US01	1	Alabama	4779736	4780135	4785579	4798649	481
1	0400000US02	2	Alaska	710231	710249	714015	722259	73
2	0400000US04	4	Arizona	6392017	6392309	6407002	6465488	654
3	0400000US05	5	Arkansas	2915918	2916031	2921737	2938640	294
4	0400000US06	6	California	37253956	37254518	37327690	37672654	3801

In [207…

```
Data2.head()
```

Out[207]:

	Player Name	Player Jersey Number	Player Position	Player Age	Player Weight (lbs.)	Years Played	Player Birthplace (city, town, etc.)	Player Birth State	Bi
0	Devin McCourty	#32	S/FS	30.0	195.0	8	Nanuet	New York	Nan
1	Danny Amendola	#80	PR/WR/KR	32.0	190.0	9	The Woodlands	Texas	Wo
2	Johnson Bademosi	#29	CB/SPTM/RCB	27.0	206.0	6	Silver Spring	Maryland	
3	Chris Hogan	#15	WR	29.0	210.0	5	Wyckoff	New Jersey	Ne
4	James Develin	#46	RB/FB	29.0	255.0	5	Gilbertsville	Pennsylvania	Gilb Pen

5 rows × 24 columns

Step 2: Dataset Cleaner

Get a cleand datasets

In [214…

```
Data.to_csv("/Users/nikhilreddyponnala/Desktop/Big Game Census Analytics/Dat
Data1.to_csv("/Users/nikhilreddyponnala/Desktop/Big Game Census Analytics/Da
Data2.to_csv("/Users/nikhilreddyponnala/Desktop/Big Game Census Analytics/Da
```

Step 3: Exploratry Data Analysis of Big Game Census

In [217…

```
# Import the necessary libs

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [221…

```
# Load the datasets
cleaned_dataset = pd.read_csv("/Users/nikhilreddyponnala/Desktop/Big Game Ce
places_cleaned = pd.read_csv("/Users/nikhilreddyponnala/Desktop/Big Game Cen
states_cleaned = pd.read_csv("/Users/nikhilreddyponnala/Desktop/Big Game Cen
```

In [223…

```
# Display the first few rows of each dataset to understand their structure

cleaned_dataset_head = cleaned_dataset.head()
places_cleaned_head = places_cleaned.head()
states_cleaned_head = states_cleaned.head()

cleaned_dataset_head, places_cleaned_head, states_cleaned_head
```

Out[223]:

```
(         Player Name Player Jersey Number Player Position  Player Age  \
 0    Devin McCourty                  #32             S/FS        30.0   
 1    Danny Amendola                  #80         PR/WR/KR        32.0   
 2  Johnson Bademosi                  #29      CB/SPTM/RCB        27.0   
 3       Chris Hogan                  #15               WR        29.0   
 4     James Develin                  #46            RB/FB        29.0   

    Player Weight (lbs.) Years Played Player Birthplace (city, town, etc.)  
\
 0                 195.0            8                               Nanuet   
 1                 190.0            9                        The Woodlands   
 2                 206.0            6                        Silver Spring   
 3                 210.0            5                              Wyckoff   
 4                 255.0            5                        Gilbertsville   

   Player Birth State    Player Birthplace (Combo) Player College  ...  \
 0           New York            Nanuet, New York        Rutgers  ...   
 1              Texas        The Woodlands, Texas     Texas Tech  ...   
 2           Maryland      Silver Spring, Maryland       Stanford  ...   
 3         New Jersey          Wyckoff, New Jersey       Monmouth  ...   
 4       Pennsylvania  Gilbertsville, Pennsylvania          Brown  ...   

     Full GEOID Latitude (player birthplace) Longitude (player birthplace)  
\
 0  0400000US36                     41.088707                    -74.013473   
 1  0400000US48                     30.173419                    -95.504686   
 2  0400000US24                     38.990666                    -77.026088   
 3  0400000US34                     41.009542                    -74.172922   
 4  0400000US42                     40.320097                    -75.610184   

    Number from City Number of Records  \
 0                 1                  1   
 1                 1                  1   
 2                 1                  1   
 3                 1                  1   
 4                 1                  1   

    American FactFinder Link for more Census data points  \
 0  https://factfinder.census.gov/bkmk/cf/1.0/en/p...      
 1  https://factfinder.census.gov/bkmk/cf/1.0/en/p...      
 2  https://factfinder.census.gov/bkmk/cf/1.0/en/p...      
 3  https://factfinder.census.gov/bkmk/cf/1.0/en/p...      
 4  https://factfinder.census.gov/bkmk/cf/1.0/en/p...      

                                    Quickfacts Link  \
 0  https://www.census.gov/quickfacts/fact/table/N...   
 1  https://www.census.gov/quickfacts/fact/table/T...   
 2  https://www.census.gov/quickfacts/fact/table/S...   
 3  https://www.census.gov/quickfacts/fact/table/W...   
 4  https://www.census.gov/quickfacts/fact/table/G...   

                                    State Data Link  \
 0  https://factfinder.census.gov/bkmk/cf/1.0/en/s...   
 1  https://factfinder.census.gov/bkmk/cf/1.0/en/s...   
 2  https://factfinder.census.gov/bkmk/cf/1.0/en/s...   
 3  https://factfinder.census.gov/bkmk/cf/1.0/en/s...   
 4  https://factfinder.census.gov/bkmk/cf/1.0/en/s...   

                         Source (Population States 2017)  \
 0  U.S. Census Bureau, 2017 Annual Estimates of t...   
 1  U.S. Census Bureau, 2017 Annual Estimates of t...   
 2  U.S. Census Bureau, 2017 Annual Estimates of t...   
 3  U.S. Census Bureau, 2017 Annual Estimates of t...   
 4  U.S. Census Bureau, 2017 Annual Estimates of t...   
```

```
                  Birthplace, Population Data Source
0  U.S. Census Bureau, 2012-2016 American Communi...
1  U.S. Census Bureau, 2012-2016 American Communi...
2  U.S. Census Bureau, 2012-2016 American Communi...
3  U.S. Census Bureau, 2012-2016 American Communi...
4  U.S. Census Bureau, 2012-2016 American Communi...

[5 rows x 24 columns],
      Geographic ID  GEOID 2 Geography, full name (City, State)  \
0  1620000US0100124   100124           Abbeville city, Alabama
1  1620000US0100460   100460          Adamsville city, Alabama
2  1620000US0100484   100484             Addison town, Alabama
3  1620000US0100676   100676               Akron town, Alabama
4  1620000US0100820   100820           Alabaster city, Alabama

  April 1, 2010 - Census  April 1, 2010 - Estimates Base  \
0                   2688                            2688
1                   4522                            4522
2                    758                             756
3                    356                             356
4                  30352                           31066

   Population Estimate (as of July 1) - 2010  \
0                                       2683
1                                       4517
2                                        754
3                                        355
4                                      31176

   Population Estimate (as of July 1) - 2011  \
0                                       2685
1                                       4495
2                                        753
3                                        345
4                                      31362

   Population Estimate (as of July 1) - 2012  \
0                                       2647
1                                       4472
2                                        748
3                                        345
4                                      31663

   Population Estimate (as of July 1) - 2013  \
0                                       2631
1                                       4447
2                                        748
3                                        342
4                                      31960

   Population Estimate (as of July 1) - 2014  \
0                                       2619
1                                       4428
2                                        747
3                                        337
4                                      32167

   Population Estimate (as of July 1) - 2015  \
0                                       2616
1                                       4395
2                                        740
3                                        337
4                                      32751
```

```
    Population Estimate (as of July 1) - 2016
 0                                       2603
 1                                       4360
 2                                        738
 3                                        334
 4                                      32948  ,
          GEOID  GEOID2 Geography Name  April 1, 2010 - Census  \
 0  0400000US01       1        Alabama                 4779736
 1  0400000US02       2         Alaska                  710231
 2  0400000US04       4        Arizona                 6392017
 3  0400000US05       5       Arkansas                 2915918
 4  0400000US06       6     California                37253956

    April 1, 2010 - Estimates Base  Population Estimate (as of July 1) - 2
010  \
 0                         4780135                                    4785
579
 1                          710249                                     714
015
 2                         6392309                                    6407
002
 3                         2916031                                    2921
737
 4                        37254518                                   37327
690

    Population Estimate (as of July 1) - 2011  \
 0                                    4798649
 1                                     722259
 2                                    6465488
 3                                    2938640
 4                                   37672654

    Population Estimate (as of July 1) - 2012  \
 0                                    4813946
 1                                     730825
 2                                    6544211
 3                                    2949208
 4                                   38019006

    Population Estimate (as of July 1) - 2013  \
 0                                    4827660
 1                                     736760
 2                                    6616124
 3                                    2956780
 4                                   38347383

    Population Estimate (as of July 1) - 2014  \
 0                                    4840037
 1                                     736759
 2                                    6706435
 3                                    2964800
 4                                   38701278

    Population Estimate (as of July 1) - 2015  \
 0                                    4850858
 1                                     737979
 2                                    6802262
 3                                    2975626
 4                                   39032444

    Population Estimate (as of July 1) - 2016  \
 0                                    4860545
```

```
1                                        741522
2                                       6908642
3                                       2988231
4                                      39296476

   Population Estimate (as of July 1) - 2017
0                                       4874747
1                                        739795
2                                       7016270
3                                       3004279
4                                      39536653  )
```

In [225…

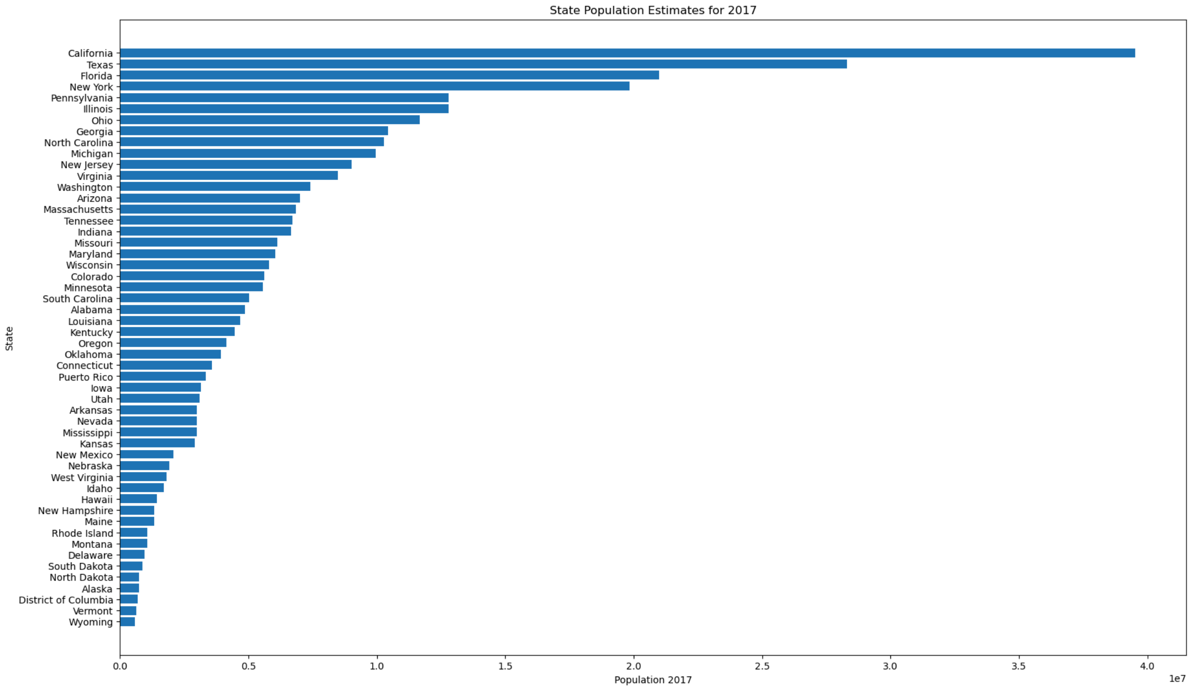
```
# Extract 2017 population data using the correct column name for states

states_population_2017 = states_cleaned[
    ["Geography Name", "Population Estimate (as of July 1) - 2017"]
].copy()
states_population_2017.columns = ["State", "Population 2017"]
```

In [227…

```
# Sort states by population for better visualization

states_population_2017 = states_population_2017.sort_values(
    by="Population 2017", ascending=False
)
```

In [229…

```
# Plot the data
plt.figure(figsize=(20, 12))
plt.barh(
    states_population_2017["State"],
    states_population_2017["Population 2017"],
)
plt.xlabel("Population 2017")
plt.ylabel("State")
plt.title("State Population Estimates for 2017")
plt.gca().invert_yaxis()
plt.show()
```



In [231…

```
# Checking the columns in places_cleaned to identify the correct column for

places_cleaned.columns.tolist()
```

Out[231]:

```
['Geographic ID',
 'GEOID 2',
 'Geography, full name (City, State)',
 'April 1, 2010 - Census',
 'April 1, 2010 - Estimates Base',
 'Population Estimate (as of July 1) - 2010',
 'Population Estimate (as of July 1) - 2011',
 'Population Estimate (as of July 1) - 2012',
 'Population Estimate (as of July 1) - 2013',
 'Population Estimate (as of July 1) - 2014',
 'Population Estimate (as of July 1) - 2015',
 'Population Estimate (as of July 1) - 2016']
```

In [233…

```python
# Extract state information from the 'Geography, full name (City, State)' co

places_cleaned["State"] = places_cleaned["Geography, full name (City, State)
    lambda x: x.split(", ")[-1]
)
```

In [235…

```python
# Check the new column

places_cleaned[["Geography, full name (City, State)", "State"]].head()
```

Out[235]:

	Geography, full name (City, State)	State
0	Abbeville city, Alabama	Alabama
1	Adamsville city, Alabama	Alabama
2	Addison town, Alabama	Alabama
3	Akron town, Alabama	Alabama
4	Alabaster city, Alabama	Alabama

In [237…

```python
# Define a mapping of states to regions

state_to_region = {
    "Northeast": [
        "Connecticut",
        "Maine",
        "Massachusetts",
        "New Hampshire",
        "Rhode Island",
        "Vermont",
        "New Jersey",
        "New York",
        "Pennsylvania",
    ],
    "Midwest": [
        "Illinois",
        "Indiana",
        "Michigan",
        "Ohio",
        "Wisconsin",
        "Iowa",
        "Kansas",
        "Minnesota",
        "Missouri",
        "Nebraska",
        "North Dakota",
        "South Dakota",
    ],
    "South": [
        "Delaware",
        "Florida",
        "Georgia",
        "Maryland",
        "North Carolina",
        "South Carolina",
        "Virginia",
        "District of Columbia",
        "West Virginia",
        "Alabama",
        "Kentucky",
        "Mississippi",
        "Tennessee",
        "Arkansas",
        "Louisiana",
        "Oklahoma",
        "Texas",
    ],
    "West": [
        "Arizona",
        "Colorado",
        "Idaho",
        "Montana",
        "Nevada",
        "New Mexico",
        "Utah",
        "Wyoming",
        "Alaska",
        "California",
        "Hawaii",
        "Oregon",
        "Washington",
    ],
}
```

In [239…

```
# Create a reverse mapping from state to region

state_to_region_rev = {
    state: region for region, states in state_to_region.items() for state i
}
```

In [241…

```
# Assign regions to players based on their birth state

cleaned_dataset["Region"] = cleaned_dataset["Player Birth State"].map(
    state_to_region_rev
)
```

In [243…

```
# Aggregate player data by region

region_wise_players = cleaned_dataset["Region"].value_counts().reset_index(
region_wise_players.columns = ["Region", "Player Count"]
```

In [245…

```
# Display the result

region_wise_players
```

Out[245]:

	Region	Player Count
0	South	53
1	West	27
2	Midwest	20
3	Northeast	15

In [247…

```python
# Extracting population estimates from states_cleaned

population_estimates = states_cleaned[
    [
        "Geography Name",
        "Population Estimate (as of July 1) - 2010",
        "Population Estimate (as of July 1) - 2011",
        "Population Estimate (as of July 1) - 2012",
        "Population Estimate (as of July 1) - 2013",
        "Population Estimate (as of July 1) - 2014",
        "Population Estimate (as of July 1) - 2015",
        "Population Estimate (as of July 1) - 2016",
        "Population Estimate (as of July 1) - 2017",
    ]
]
```

In [249…

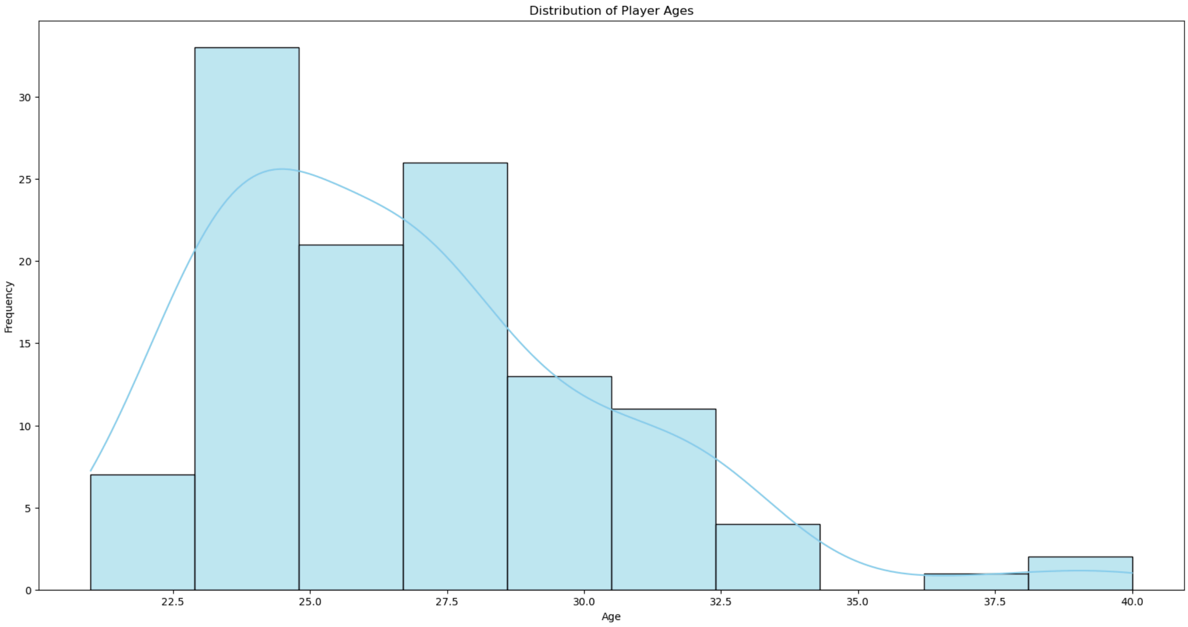
```python
# Rename columns for clarity

population_estimates.columns = [
    "State",
    "2010",
    "2011",
    "2012",
    "2013",
    "2014",
    "2015",
    "2016",
    "2017",
]

population_estimates.head()
```

Out[249]:

	State	2010	2011	2012	2013	2014	2015	2016
0	Alabama	4785579	4798649	4813946	4827660	4840037	4850858	4860545
1	Alaska	714015	722259	730825	736760	736759	737979	741522
2	Arizona	6407002	6465488	6544211	6616124	6706435	6802262	6908642
3	Arkansas	2921737	2938640	2949208	2956780	2964800	2975626	2988231
4	California	37327690	37672654	38019006	38347383	38701278	39032444	39296476

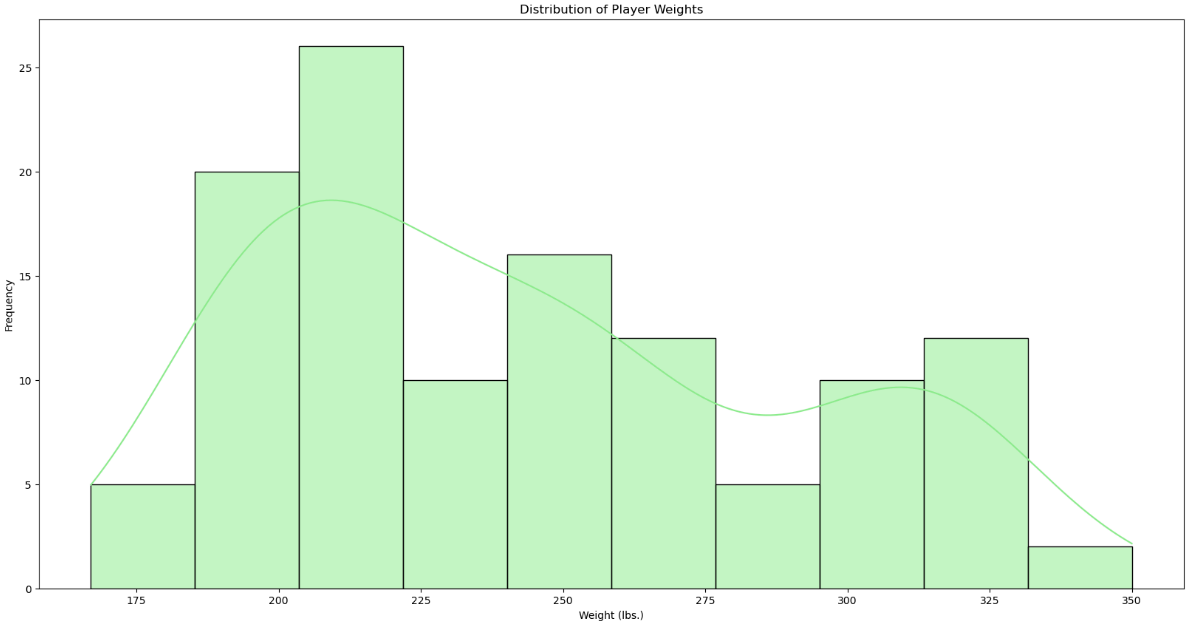
In [251…

```python
# Distribution of Player Ages

plt.figure(figsize=(20, 10))
sns.histplot(
    cleaned_dataset["Player Age"],
    bins=10,
    kde=True,
    color="skyblue",
)
plt.title("Distribution of Player Ages")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```



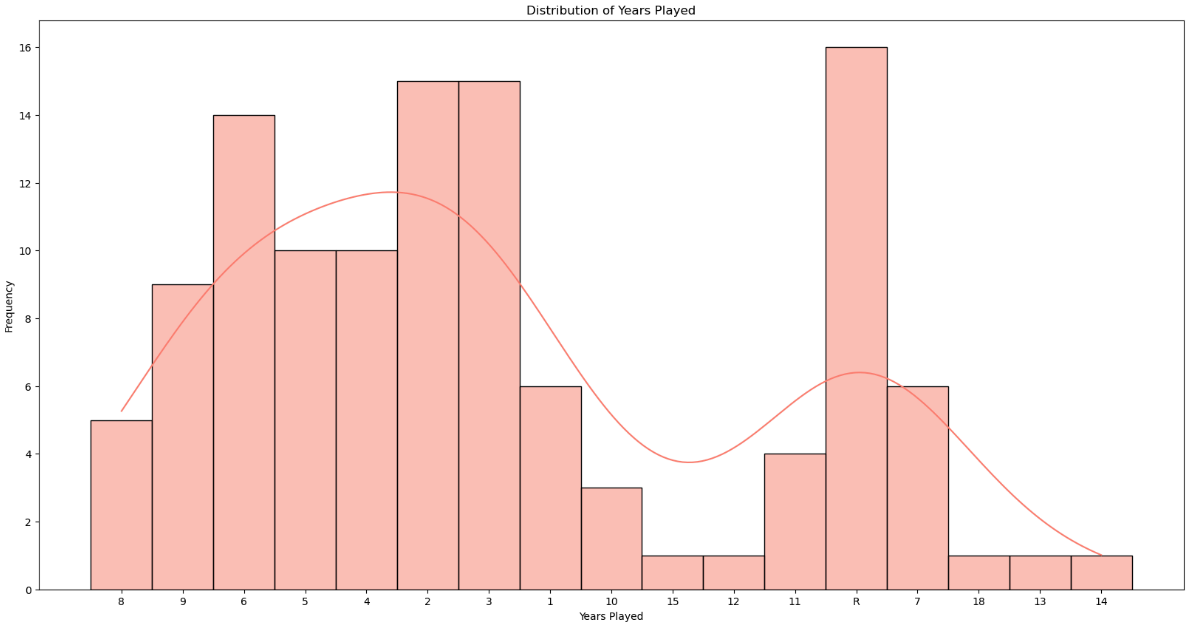
In [253…

```
# Distribution of Player Weights

plt.figure(figsize=(20, 10))
sns.histplot(
    cleaned_dataset["Player Weight (lbs.)"],
    bins=10,
    kde=True,
    color="lightgreen",
)
plt.title("Distribution of Player Weights")
plt.xlabel("Weight (lbs.)")
plt.ylabel("Frequency")
plt.show()
```



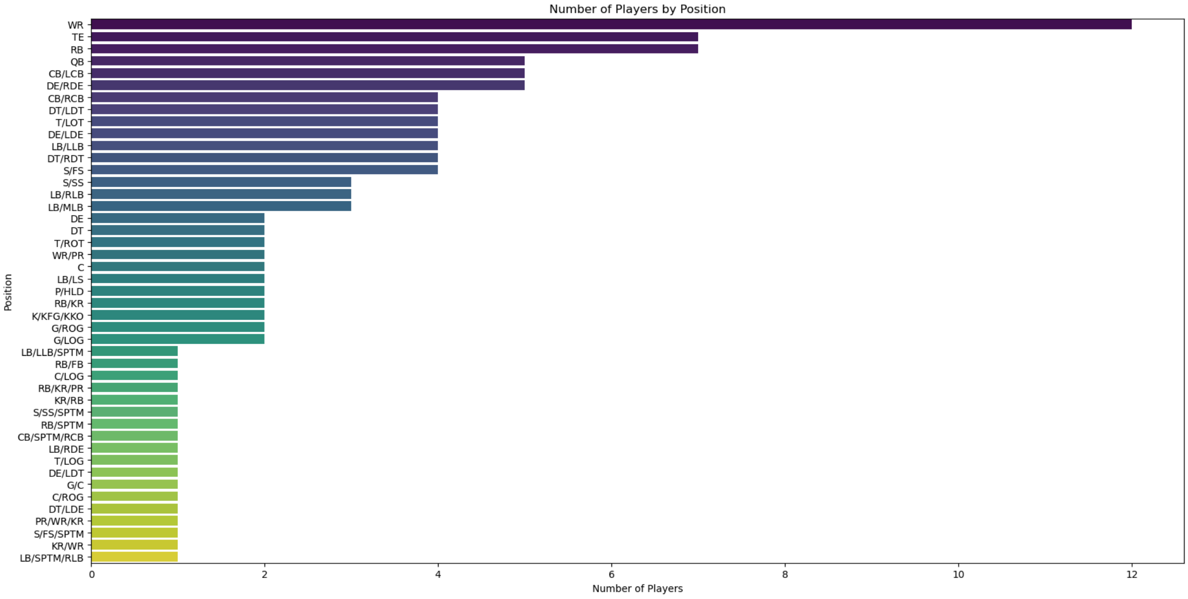
In [255…

```
# Distribution of Years Played

plt.figure(figsize=(20, 10))
sns.histplot(cleaned_dataset["Years Played"], bins=10, kde=True, color="salr
plt.title("Distribution of Years Played")
plt.xlabel("Years Played")
plt.ylabel("Frequency")
plt.show()
```



In [257…

```
# Position-wise Analysis

plt.figure(figsize=(20, 10))
sns.countplot(
    y=cleaned_dataset["Player Position"],
    order=cleaned_dataset["Player Position"].value_counts().index,
    palette="viridis",
)
plt.title("Number of Players by Position")
plt.xlabel("Number of Players")
plt.ylabel("Position")
plt.show()
```



In [259…

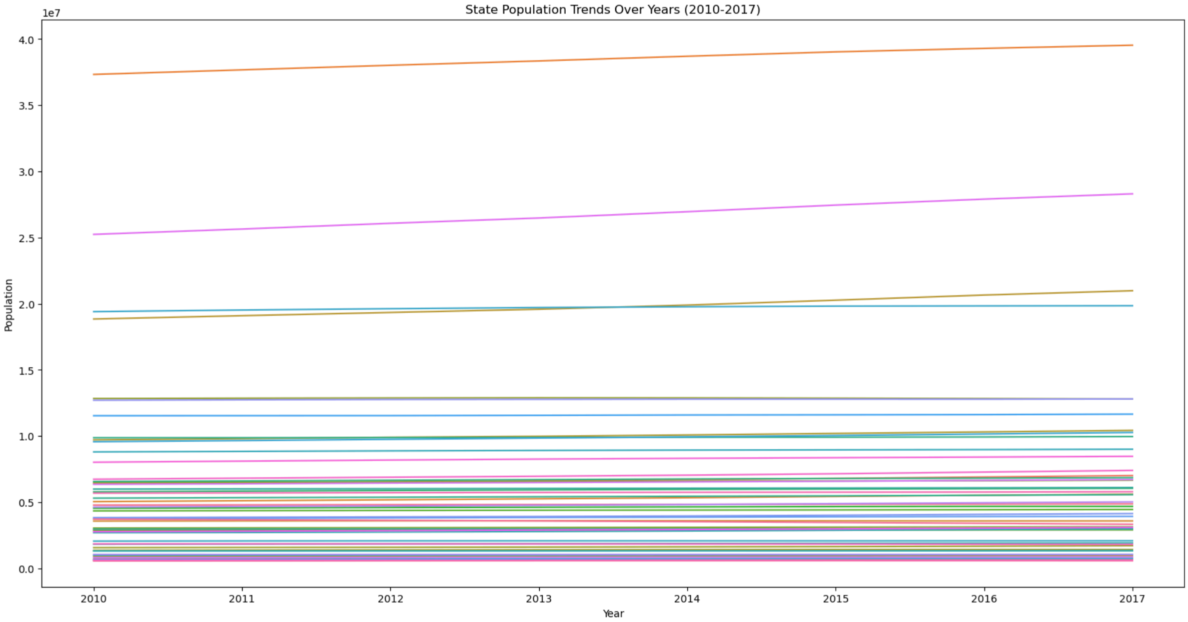
```
### Plotting state population trends over the years
## Reshaping the population estimates data for better visualization

population_trends = population_estimates.melt(
    id_vars=["State"], var_name="Year", value_name="Population"
)
```

In [261…

```
# Converting 'Year' to a numerical format for plotting

population_trends["Year"] = population_trends["Year"].astype(int)
```

In [263…

```
# Plot the state population trends

plt.figure(figsize=(20, 10))
sns.lineplot(data=population_trends, x="Year", y="Population", hue="State",
plt.title("State Population Trends Over Years (2010-2017)")
plt.xlabel("Year")
plt.ylabel("Population")
plt.show()
```

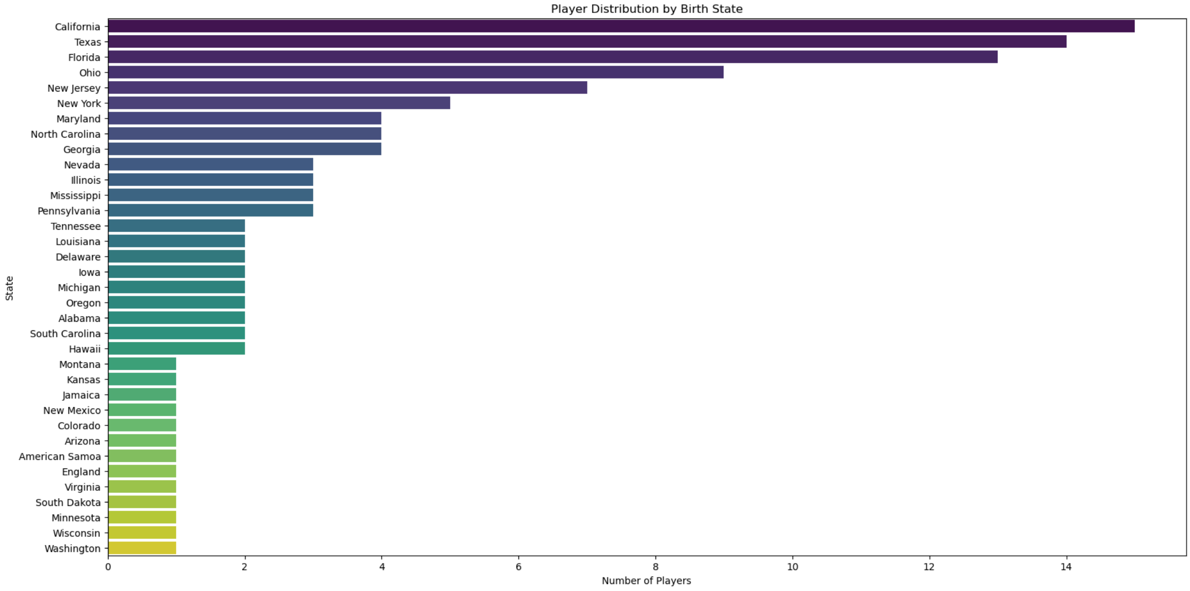


In [265…

```
### Analyzing player distribution by state
## Count the number of players from each state

player_distribution_by_state = (
    cleaned_dataset["Player Birth State"].value_counts().reset_index()
)
player_distribution_by_state.columns = ["State", "Player Count"]
```

In [267…

```
# Plot the player distribution by state

plt.figure(figsize=(20, 10))
sns.barplot(
    y=player_distribution_by_state["State"],
    x=player_distribution_by_state["Player Count"],
    palette="viridis",
)
plt.title("Player Distribution by Birth State")
plt.xlabel("Number of Players")
plt.ylabel("State")
plt.show()
```



In [269…

```
### Comparing player count with state population
## Merge player count data with population estimates for 2017

player_population_comparison = pd.merge(
    player_distribution_by_state,
    states_cleaned[["Geography Name", "Population Estimate (as of July 1) –
    how="left",
    left_on="State",
    right_on="Geography Name",
)
```

In [271…

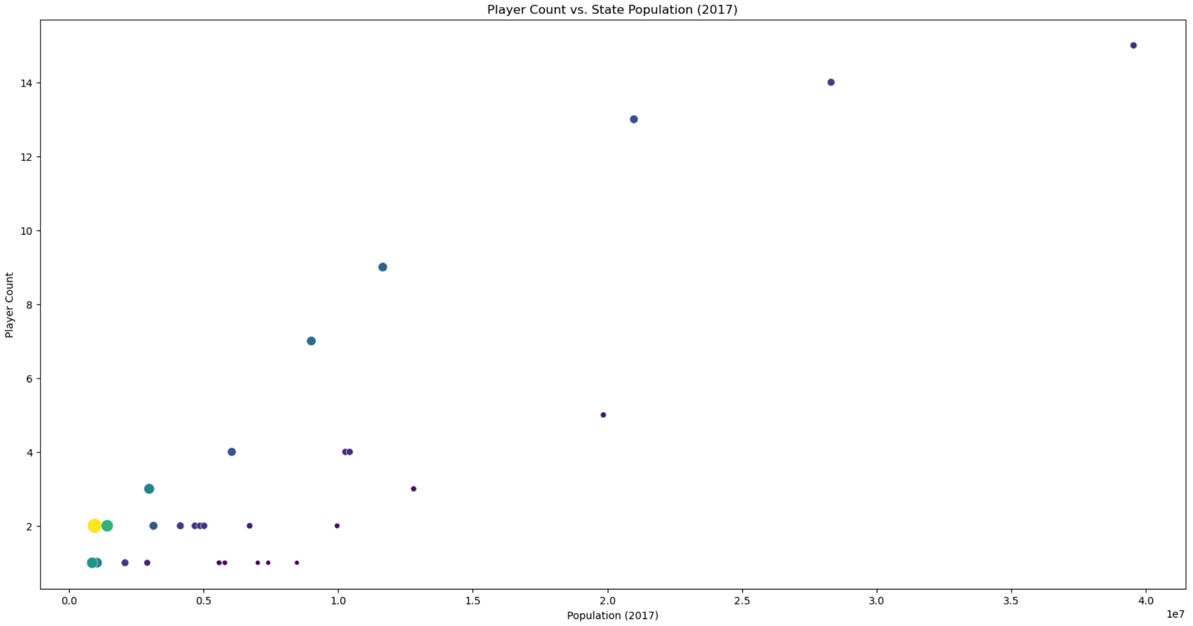
```
# Drop unnecessary columns and rename for clarity

player_population_comparison = player_population_comparison[
    ["State", "Player Count", "Population Estimate (as of July 1) – 2017"]
]
player_population_comparison.columns = ["State", "Player Count", "Populatior
```

In [273…

```
# Calculate players per capita (players per million residents)

player_population_comparison["Players per Million"] = player_population_comp
    "Player Count"
] / (player_population_comparison["Population 2017"] / 1_000_000)
```

In [275…

```
# Plot the comparison

plt.figure(figsize=(20, 10))
sns.scatterplot(
    data=player_population_comparison,
    x="Population 2017",
    y="Player Count",
    hue="Players per Million",
    size="Players per Million",
    sizes=(20, 200),
    palette="viridis",
    legend=None,
)
plt.title("Player Count vs. State Population (2017)")
plt.xlabel("Population (2017)")
plt.ylabel("Player Count")
plt.show()

player_population_comparison.head()
```



Out[275]:

	State	Player Count	Population 2017	Players per Million
0	California	15	39536653.0	0.379395
1	Texas	14	28304596.0	0.494619
2	Florida	13	20984400.0	0.619508
3	Ohio	9	11658609.0	0.771962
4	New Jersey	7	9005644.0	0.777290

In [277…

```
# Analyzing players by college

college_distribution = cleaned_dataset["Player College"].value_counts().rese
college_distribution.columns = ["College", "Player Count"]
```

In [279…

```
# Top 10 colleges by number of players

top_colleges = college_distribution.head(10)
```

In [281…

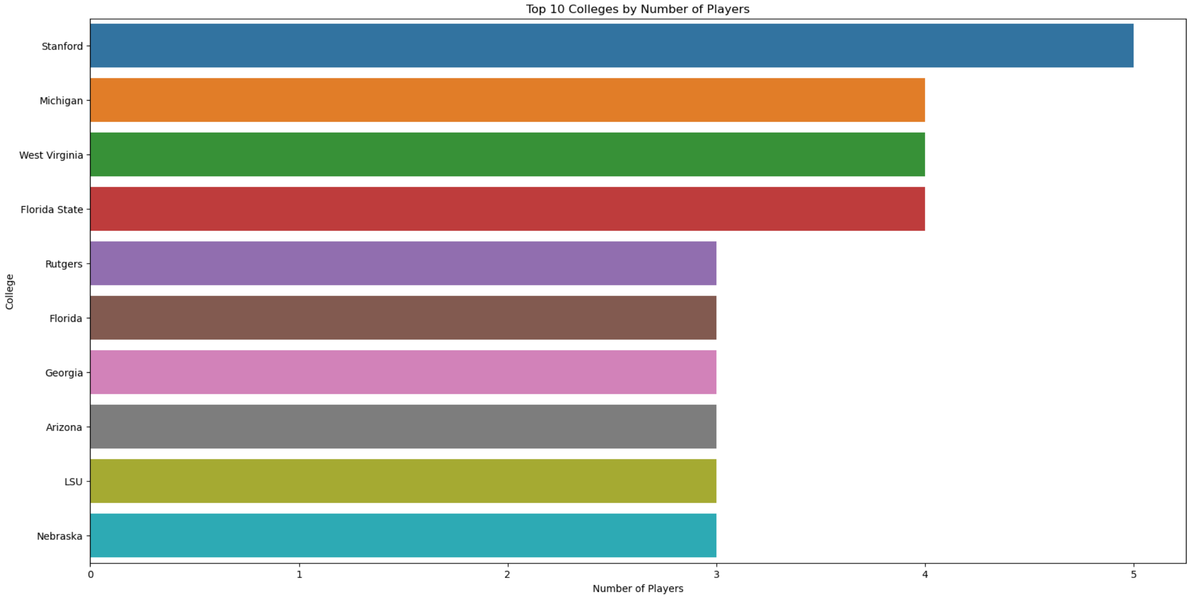
```
# Displaying detailed player information

player_information = cleaned_dataset[
    [
        "Player Name",
        "Player Jersey Number",
        "Player Position",
        "Player Age",
        "Player Weight (lbs.)",
        "Years Played",
        "Player Birth State",
        "Player College",
    ]
]

player_information.head()
```

Out[281]:

	Player Name	Player Jersey Number	Player Position	Player Age	Player Weight (lbs.)	Years Played	Player Birth State	Player College
0	Devin McCourty	#32	S/FS	30.0	195.0	8	New York	Rutgers
1	Danny Amendola	#80	PR/WR/KR	32.0	190.0	9	Texas	Texas Tech
2	Johnson Bademosi	#29	CB/SPTM/RCB	27.0	206.0	6	Maryland	Stanford
3	Chris Hogan	#15	WR	29.0	210.0	5	New Jersey	Monmouth
4	James Develin	#46	RB/FB	29.0	255.0	5	Pennsylvania	Brown

In [283…

```python
# Plot the distribution of players by college

plt.figure(figsize=(20, 10))
sns.barplot(
    y=top_colleges["College"],
    x=top_colleges["Player Count"],
)
plt.title("Top 10 Colleges by Number of Players")
plt.xlabel("Number of Players")
plt.ylabel("College")
plt.show()
```



In [285…

```python
# Cleaning the 'Years Played' column to ensure all values are numeric

cleaned_dataset["Years Played"] = pd.to_numeric(
    cleaned_dataset["Years Played"], errors="coerce"
)
```

In [287…

```python
# Drop rows with NaN values in 'Years Played' after conversion

cleaned_dataset_clean = cleaned_dataset.dropna(subset=["Years Played"])
```

In [289…

```python
# Re-calculate performance by college (average years played)

college_performance_clean = (
    cleaned_dataset_clean.groupby("Player College")["Years Played"].mean().
```

```
)
top_college_performance_clean = college_performance_clean[
    college_performance_clean["Player College"].isin(top_colleges["College"
]
```

In [291…

```
# Detailed list of all colleges and the number of players they have produce

all_college_affiliations = (
    cleaned_dataset_clean["Player College"].value_counts().reset_index()
)
all_college_affiliations.columns = ["College", "Player Count"]


all_college_affiliations.head(20)
```

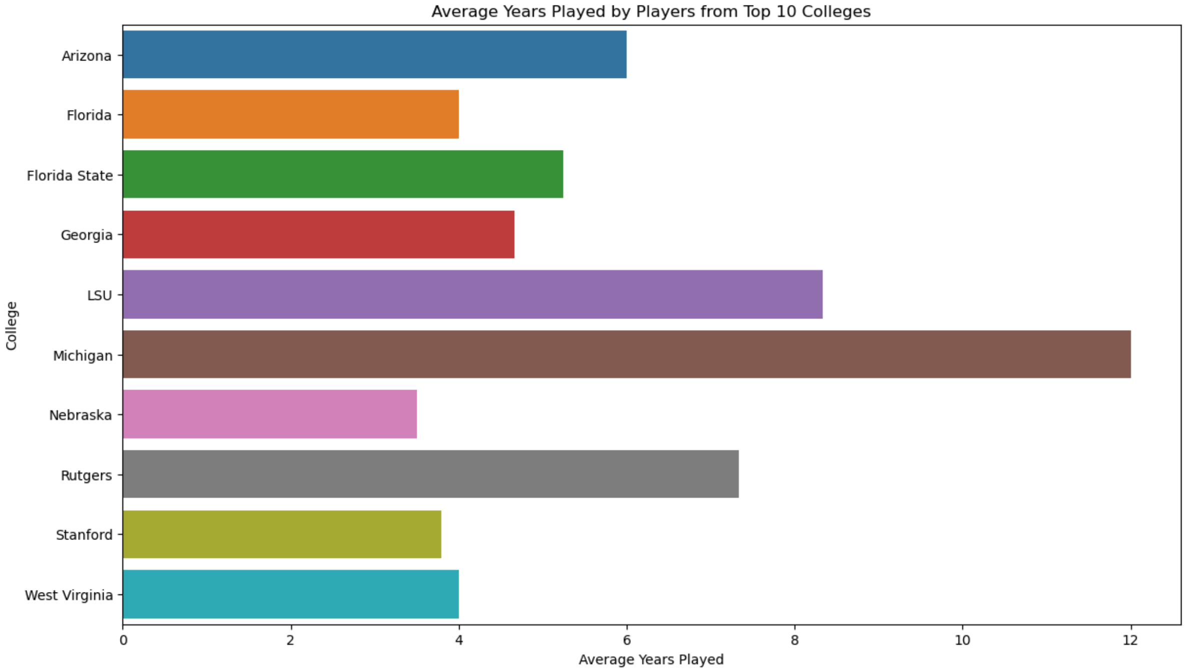
Out[291]:

	College	Player Count
0	Stanford	5
1	Florida State	4
2	Michigan	4
3	Rutgers	3
4	Florida	3
5	Arizona	3
6	Georgia	3
7	LSU	3
8	Oregon	3
9	West Virginia	2
10	Texas	2
11	Oklahoma	2
12	South Carolina	2
13	Virginia	2
14	Texas Tech	2
15	Auburn	2
16	Wisconsin	2
17	Pittsburgh	2
18	Oregon State	2
19	Washington State	2

In [293…

```
# Plot the average years played by college

plt.figure(figsize=(14, 8))
sns.barplot(
    y=top_college_performance_clean["Player College"],
    x=top_college_performance_clean["Years Played"],
)
plt.title("Average Years Played by Players from Top 10 Colleges")
plt.xlabel("Average Years Played")
plt.ylabel("College")
plt.show()
```

```
top_college_performance_clean
```



Out[293]:

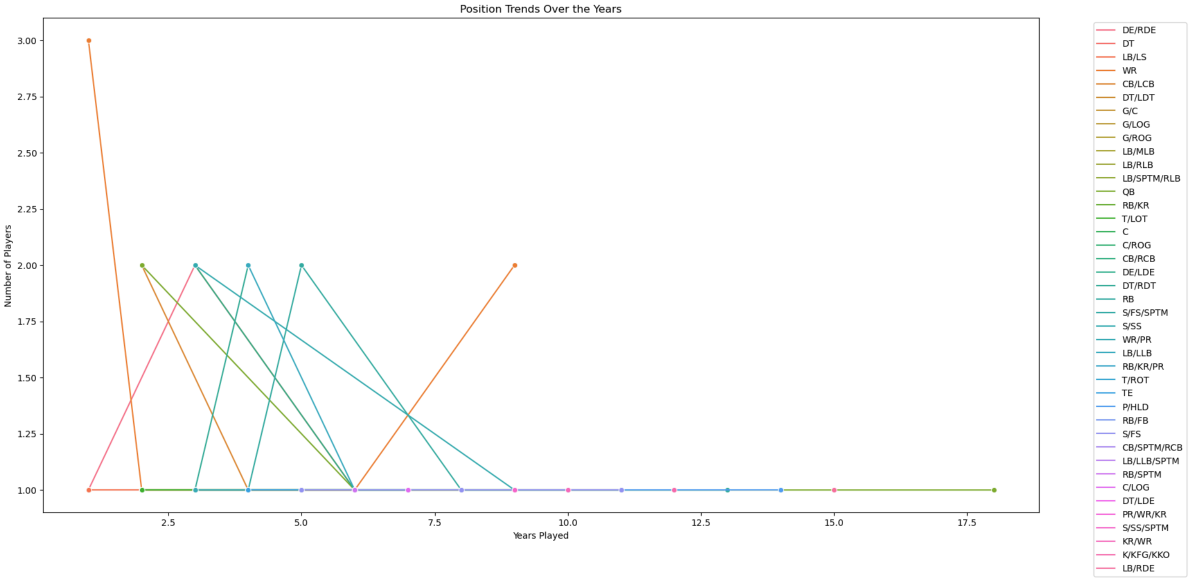
	Player College	Years Played
1	Arizona	6.000000
15	Florida	4.000000
16	Florida State	5.250000
17	Georgia	4.666667
24	LSU	8.333333
31	Michigan	12.000000
37	Nebraska	3.500000
48	Rutgers	7.333333
51	Stanford	3.800000
64	West Virginia	4.000000

In [295…

```
# Analyzing position trends over the years

position_trends = (
    cleaned_dataset_clean.groupby(["Years Played", "Player Position"])
    .size()
    .reset_index(name="Count")
)
```

In [297…

```
# Plot the position trends over the years

plt.figure(figsize=(20, 10))
sns.lineplot(
    data=position_trends, x="Years Played", y="Count", hue="Player Position'
)
plt.title("Position Trends Over the Years")
plt.xlabel("Years Played")
plt.ylabel("Number of Players")
plt.legend(bbox_to_anchor=(1.05, 1), loc="upper left")
plt.show()
```


Position Trends Over the Years
Number of Players
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Years Played
DE/RDE
DT
LB/LS
WR
CB/LCB
DT/LDT
G/C
G/LOG
G/ROG
LB/MLB
LB/RLB
LB/SPTM/RLB
QB
RB/KR
T/LOT
C
C/ROG
CB/RCB
DE/LDE
DT/RDT
RB
S/FS/SPTM
S/SS
WR/PR
LB/LLB
RB/KR/PR
T/ROT
TE
P/HLD
RB/FB
S/FS
CB/SPTM/RCB
LB/LLB/SPTM
RB/SPTM
C/LOG
DT/LDE
PR/WR/KR
S/SS/SPTM
KR/WR
K/KFG/KKO
LB/RDE

In [ ]: