PandasSQL

PandasSQL is a Python library that provides a convenient way to query pandas DataFrames using SQL syntax. It allows you to leverage your SQL skills to manipulate and analyze data stored in pandas DataFrames, providing a familiar interface for those comfortable with SQL. Here's an explanation and an example to illustrate its usage:

Overview of pandasql:

PandasSQL bridges the gap between SQL and pandas by allowing you to write SQL queries directly against pandas DataFrames. It internally converts SQL queries into pandas operations, making it easier to perform complex data manipulations and aggregations using SQL syntax.

In this analysis, i shall be using a data obtained fron kaggle called: home_school_district and home_school_state obtained during the pandemic in USA

home_school_district

In [1]:
```
import pandas as pd
data_district = pd.read_csv('home_school_district.csv')
data_district=data_district.sort_values(by=['state'],ascending=True)
data_district
```

Out[1]:

	lea_name	lea_id	state	year	homeschool_students
0	Alma School District	502250	AR	2017-18	142.0
945	Mulberry/Pleasant View Bi-County Schools	510290	AR	2021-22	71.0
944	Mulberry School District	510290	AR	2020-21	73.0
943	Mulberry School District	510290	AR	2019-20	54.0
942	Mulberry School District	510290	AR	2018-19	37.0
...	...	...	...	...	...
22986	Platte County School District #2	5603180	WY	2018-19	9.0
22987	Platte County School District #2	5603180	WY	2019-20	0.0
22988	Platte County School District #2	5603180	WY	2020-21	21.0
23009	Sublette County School District #1	5604860	WY	2017-18	56.0
23063	Washakie County School District #2	5605820	WY	2017-18	0.0

37674 rows × 5 columns

In [2]:
```
data_district.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 37674 entries, 0 to 23063
Data columns (total 5 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   lea_name             37637 non-null  object
 1   lea_id               37639 non-null  object
 2   state                37674 non-null  object
 3   year                 37674 non-null  object
 4   homeschool_students  37192 non-null  float64
dtypes: float64(1), object(4)
memory usage: 1.7+ MB
```

In [3]:
```
data_district=data_district.dropna()
data_district
```

Out[3]:

	lea_name	lea_id	state	year	homeschool_students
0	Alma School District	502250	AR	2017-18	142.0
945	Mulberry/Pleasant View Bi-County Schools	510290	AR	2021-22	71.0
944	Mulberry School District	510290	AR	2020-21	73.0
943	Mulberry School District	510290	AR	2019-20	54.0
942	Mulberry School District	510290	AR	2018-19	37.0
...	...	...	...	...	...
22986	Platte County School District #2	5603180	WY	2018-19	9.0
22987	Platte County School District #2	5603180	WY	2019-20	0.0
22988	Platte County School District #2	5603180	WY	2020-21	21.0
23009	Sublette County School District #1	5604860	WY	2017-18	56.0
23063	Washakie County School District #2	5605820	WY	2017-18	0.0

37167 rows × 5 columns

Data Transformation

In [4]:
```
data_district['state'].unique()
```

Out[4]:
```
array(['AR', 'CA', 'CO', 'DC', 'DE', 'FL', 'GA', 'HI', 'KS', 'KY', 'LA',
       'MA', 'MD', 'ME', 'MN', 'MS', 'ND', 'NH', 'NM', 'NY', 'OH', 'PA',
       'RI', 'SC', 'SD', 'TN', 'VA', 'WA', 'WI', 'WY'], dtype=object)
```

In [5]:
```
district_state={'AR':'ARKANSAS',
                        'CA':'CALIFORNIA',
                        'CO':'COLORADO',
                        'DC':'DISTRICT OF COLUMBIA',
                        'DE':'DELAWARE',
                        'FL':'FLORIDA',
                        'GA':'GEORGIA',
                        'HI':'HAWAII',
                        'KS':'KANSAS',
                        'KY':'KENTUCKY',
                        'LA':'LOUISIANA',
                        'MA':'MAINE',
                        'MD':'MARYLAND',
                        'ME':'MASSACHUSETTS',
                        'MN':'MINNESOTA',
                        'MS':'MISSISSIPPI',
                        'ND':'NORTH DAKOTA',
                        'NH':'NEW HAMPSHIRE',
                        'NM':'NEW MEXICO',
                        'NY':'NEW YORK',
                        'OH':'OHIO',
                        'PA':'PENNSYLVANIA',
                        'RI':'RHODE ISLAND',
                        'SC':'SOUTH CAROLINA',
                        'SD':'SOUTH DAKOTA',
                        'TN':'TENNESSEE',
                        'VA':'VIRGINIA',
                        'WA':'WASHINGTON',
                        'WI':'WISCONSIN',
                        'WY':'WYOMING'
                }
data_district['state'] = data_district['state'].replace(district_state)
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_10848\1806540588.py:32: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data_district['state'] = data_district['state'].replace(district_state)
```

In [6]:
```
data_district
```

Out[6]:

	lea_name	lea_id	state	year	homeschool_students
0	Alma School District	502250	ARKANSAS	2017-18	142.0
945	Mulberry/Pleasant View Bi-County Schools	510290	ARKANSAS	2021-22	71.0
944	Mulberry School District	510290	ARKANSAS	2020-21	73.0
943	Mulberry School District	510290	ARKANSAS	2019-20	54.0
942	Mulberry School District	510290	ARKANSAS	2018-19	37.0
...	...	...	...	...	...
22986	Platte County School District #2	5603180	WYOMING	2018-19	9.0
22987	Platte County School District #2	5603180	WYOMING	2019-20	0.0
22988	Platte County School District #2	5603180	WYOMING	2020-21	21.0
23009	Sublette County School District #1	5604860	WYOMING	2017-18	56.0
23063	Washakie County School District #2	5605820	WYOMING	2017-18	0.0

37167 rows × 5 columns

In [7]:
```
data_district['state'].unique()
```

Out[7]:
```
array(['ARKANSAS', 'CALIFORNIA', 'COLORADO', 'DISTRICT OF COLUMBIA',
       'DELAWARE', 'FLORIDA', 'GEORGIA', 'HAWAII', 'KANSAS', 'KENTUCKY',
       'LOUISIANA', 'MAINE', 'MARYLAND', 'MASSACHUSETTS', 'MINNESOTA',
       'MISSISSIPPI', 'NORTH DAKOTA', 'NEW HAMPSHIRE', 'NEW MEXICO',
       'NEW YORK', 'OHIO', 'PENNSYLVANIA', 'RHODE ISLAND',
       'SOUTH CAROLINA', 'SOUTH DAKOTA', 'TENNESSEE', 'VIRGINIA',
       'WASHINGTON', 'WISCONSIN', 'WYOMING'], dtype=object)
```

Save the dataset in a csv

In [8]:
```
data_district.to_csv("data_district.csv",index=False)
```

import CSV file data into a PostgreSQL table

In [9]:
```
from sqlalchemy import create_engine
import pandas as pd
df = pd.read_csv('data_district.csv')

from sqlalchemy import create_engine
engine = create_engine('postgresql://postgres:kayode@localhost:5432/Ex1')

df.to_sql("home_school_district1", engine, if_exists='replace')
```

Out[9]:
```
167
```

In [10]:
```
sql_query = pd.read_sql_query("SELECT * FROM home_school_district1", engine)
sql_query
```

Out[10]:

	index	lea_name	lea_id	state	year	homeschool_students
0	0	Alma School District	502250	ARKANSAS	2017-18	142.0
1	1	Mulberry/Pleasant View Bi-County Schools	510290	ARKANSAS	2021-22	71.0
2	2	Mulberry School District	510290	ARKANSAS	2020-21	73.0
3	3	Mulberry School District	510290	ARKANSAS	2019-20	54.0
4	4	Mulberry School District	510290	ARKANSAS	2018-19	37.0
...	...	...	...	...	...	...
37162	37162	Platte County School District #2	5603180	WYOMING	2018-19	9.0
37163	37163	Platte County School District #2	5603180	WYOMING	2019-20	0.0
37164	37164	Platte County School District #2	5603180	WYOMING	2020-21	21.0
37165	37165	Sublette County School District #1	5604860	WYOMING	2017-18	56.0
37166	37166	Washakie County School District #2	5605820	WYOMING	2017-18	0.0

37167 rows × 6 columns

In [11]:
```
from pandasql import sqldf
import pandas as pd
district = sqldf("SELECT * FROM sql_query")
district
```

Out[11]:

	index	lea_name	lea_id	state	year	homeschool_students
0	0	Alma School District	502250	ARKANSAS	2017-18	142.0
1	1	Mulberry/Pleasant View Bi-County Schools	510290	ARKANSAS	2021-22	71.0
2	2	Mulberry School District	510290	ARKANSAS	2020-21	73.0
3	3	Mulberry School District	510290	ARKANSAS	2019-20	54.0
4	4	Mulberry School District	510290	ARKANSAS	2018-19	37.0
...	...	...	...	...	...	...
37162	37162	Platte County School District #2	5603180	WYOMING	2018-19	9.0
37163	37163	Platte County School District #2	5603180	WYOMING	2019-20	0.0
37164	37164	Platte County School District #2	5603180	WYOMING	2020-21	21.0
37165	37165	Sublette County School District #1	5604860	WYOMING	2017-18	56.0
37166	37166	Washakie County School District #2	5605820	WYOMING	2017-18	0.0

37167 rows × 6 columns

In [12]:
```
query= """SELECT 'TOTAL', COUNT(*) as total_count
        FROM district"""

Total_district = sqldf(query)
Total_district
```

Out[12]:

	'TOTAL'	total_count
0	TOTAL	37167

In [13]:
```
query = """SELECT *
FROM(SELECT state, lea_name, COUNT(*) OVER(PARTITION BY lea_name) lea_name_count
FROM district) a
GROUP BY state, lea_name, lea_name_count
ORDER BY lea_name_count"""

District_count = sqldf(query)
District_count
```

Out[13]:

	state	lea_name	lea_name_count
0	CALIFORNIA	Alpaugh Unified	1
1	CALIFORNIA	Aromas/San Juan Unified	1
2	CALIFORNIA	Belridge Elementary	1
3	CALIFORNIA	Dixie Elementary	1
4	CALIFORNIA	Ducor Union Elementary	1
...	...	...	...
6750	WASHINGTON	Highland School District	18
6751	WASHINGTON	Monroe School District	18
6752	WISCONSIN	Highland School District	18
6753	WISCONSIN	Monroe School District	18
6754	WISCONSIN	Salem School District	18

6755 rows × 3 columns

In [14]:
```
query = """SELECT lea_name,COUNT(*) Count_of_lea_name
FROM district
GROUP BY lea_name
ORDER BY COUNT(*) asc"""

Lea_name_count = sqldf(query)
Lea_name_count
```

Out[14]:

	lea_name	Count_of_lea_name
0	Alpaugh Unified	1
1	Apollo	1
2	Aromas/San Juan Unified	1
3	Ashland County-West Holmes	1
4	Bamberg 03	1
...	...	...
6626	Perry Local	18
6627	Pioneer Union Elementary	18
6628	Salem School District	18
6629	Southern Local	18
6630	Springfield Local	18

6631 rows × 2 columns

In [15]:

```
query="""SELECT lea_name, SUM(homeschool_students) homeschool_sum
FROM district
GROUP BY lea_name
ORDER BY lea_name asc;
"""

homeschool_sum = sqldf(query)
homeschool_sum
```

Out[15]:

	lea_name	homeschool_sum
0	A.c.g.c. Public School District	292.0
1	Abbeville 60	983.0
2	Abbotsford School District	39.0
3	Abc Unified	363.0
4	Aberdeen School Dist	286.0
...	...	...
6626	Zane Trace Local	202.0
6627	Zanesville City	400.0
6628	Zeeland 4	1.0
6629	Zillah School District	223.0
6630	Zumbrota-Mazeppa School District	117.0

6631 rows × 2 columns

In [16]:

```
query="""SELECT lea_name,COUNT(*) Count_of_lea_names
         FROM district
         GROUP BY lea_name
         UNION ALL
         SELECT 'TOTAL', COUNT(*)
         FROM district
         ORDER BY Count_of_lea_names
"""

Lea_name_total = sqldf(query)
Lea_name_total
```

Out[16]:

	lea_name	Count_of_lea_names
0	Alpaugh Unified	1
1	Apollo	1
2	Aromas/San Juan Unified	1
3	Ashland County-West Holmes	1
4	Bamberg 03	1
...	...	...
6627	Pioneer Union Elementary	18
6628	Salem School District	18
6629	Southern Local	18
6630	Springfield Local	18
6631	TOTAL	37167

6632 rows × 2 columns

In [17]:

```
query="""SELECT state, COUNT() State_count
    FROM(SELECT state, COUNT(*) OVER(PARTITION BY state) State_count
    FROM district) a
    GROUP BY state
    UNION ALL
    SELECT 'TOTAL', COUNT(*)
    FROM(SELECT state, COUNT(*) OVER(PARTITION BY state) State_count
    FROM district) a
    ORDER BY state_count"""

state_total = sqldf(query)
state_total
```

Out[17]:

	state	State_count
0	DISTRICT OF COLUMBIA	6
1	HAWAII	6
2	DELAWARE	96
3	MARYLAND	144
4	RHODE ISLAND	205
5	WYOMING	288
6	FLORIDA	402
7	NEW MEXICO	410
8	LOUISIANA	414
9	SOUTH CAROLINA	454
10	MASSACHUSETTS	524
11	TENNESSEE	577
12	MISSISSIPPI	747
13	NORTH DAKOTA	765
14	VIRGINIA	779
15	SOUTH DAKOTA	893
16	NEW HAMPSHIRE	919
17	KENTUCKY	956
18	COLORADO	1080
19	GEORGIA	1186
20	ARKANSAS	1403
21	MAINE	1449
22	WASHINGTON	1566
23	KANSAS	1716
24	MINNESOTA	1975
25	PENNSYLVANIA	2495
26	WISCONSIN	2525
27	OHIO	3523
28	NEW YORK	4315
29	CALIFORNIA	5349
30	TOTAL	37167

In [18]:
```
query="""SELECT state
FROM district
GROUP BY state"""

Unique_states = sqldf(query)
Unique_states
```

Out[18]:

	state
0	ARKANSAS
1	CALIFORNIA
2	COLORADO
3	DELAWARE
4	DISTRICT OF COLUMBIA
5	FLORIDA
6	GEORGIA
7	HAWAII
8	KANSAS
9	KENTUCKY
10	LOUISIANA
11	MAINE
12	MARYLAND
13	MASSACHUSETTS
14	MINNESOTA
15	MISSISSIPPI
16	NEW HAMPSHIRE
17	NEW MEXICO
18	NEW YORK
19	NORTH DAKOTA
20	OHIO
21	PENNSYLVANIA
22	RHODE ISLAND
23	SOUTH CAROLINA
24	SOUTH DAKOTA
25	TENNESSEE
26	VIRGINIA
27	WASHINGTON
28	WISCONSIN
29	WYOMING

In [19]:
```
query = """SELECT lea_name, state, homeschool_students,
RANK() OVER(PARTITION BY state ORDER BY homeschool_students desc) Ranking
FROM district"""

Rank_state = sqldf(query)
Rank_state
```

Out[19]:

	lea_name	state	homeschool_students	Ranking
0	Bentonville School District	ARKANSAS	1613.0	1
1	Bentonville School District	ARKANSAS	1532.0	2
2	Bentonville School District	ARKANSAS	1503.0	3
3	Pulaski Co. Spec. School Dist.	ARKANSAS	1233.0	4
4	Bentonville School District	ARKANSAS	1121.0	5
...	...	...	...	...
37162	Sheridan County School District #3	WYOMING	0.0	213
37163	Park County School District #16	WYOMING	0.0	213
37164	Park County School District #16	WYOMING	0.0	213
37165	Platte County School District #2	WYOMING	0.0	213
37166	Washakie County School District #2	WYOMING	0.0	213

37167 rows × 4 columns

In [20]:

```
query = """SELECT state, sum(homeschool_students) as Sum_district
         FROM district
         GROUP BY state
         ORDER BY Sum_district"""

Grouped_by_state = sqldf(query)
Grouped_by_state
```

Out[20]:

	state	Sum_district
0	DISTRICT OF COLUMBIA	4235.0
1	RHODE ISLAND	11860.0
2	WYOMING	18082.0
3	KANSAS	20260.0
4	NEW HAMPSHIRE	21518.0
5	NORTH DAKOTA	21725.0
6	DELAWARE	22026.0
7	HAWAII	25797.0
8	MASSACHUSETTS	29423.0
9	SOUTH DAKOTA	38314.0
10	COLORADO	59500.0
11	NEW MEXICO	63884.0
12	MAINE	64765.0
13	TENNESSEE	65883.0
14	MISSISSIPPI	134129.0
15	MINNESOTA	144096.0
16	WISCONSIN	149976.0
17	ARKANSAS	154733.0
18	SOUTH CAROLINA	155294.0
19	LOUISIANA	159993.0
20	PENNSYLVANIA	162070.0
21	WASHINGTON	163713.0
22	KENTUCKY	188024.0
23	MARYLAND	196616.0
24	NEW YORK	243683.0
25	OHIO	246460.0
26	CALIFORNIA	248183.0
27	VIRGINIA	315343.0
28	GEORGIA	479288.0
29	FLORIDA	743022.0

State with the highest enrollment is Florida

In [21]:

```
query = """SELECT *
         FROM district
```

```
        WHERE state == 'FLORIDA'
        """

Florida_state = sqldf(query)
Florida_state
```

Out[21]:

	index	lea_name	lea_id	state	year	homeschool_students
0	7934	Leon	1201110	FLORIDA	2020-21	2007.0
1	7935	Lee	1201080	FLORIDA	2021-22	3698.0
2	7936	Leon	1201110	FLORIDA	2019-20	1650.0
3	7937	Leon	1201110	FLORIDA	2018-19	1801.0
4	7938	Leon	1201110	FLORIDA	2017-18	2026.0
...	...	...	...	...	...	...
397	8331	St. Johns	1201740	FLORIDA	2019-20	2224.0
398	8332	St. Johns	1201740	FLORIDA	2018-19	1974.0
399	8333	St. Johns	1201740	FLORIDA	2017-18	1745.0
400	8334	Seminole	1201710	FLORIDA	2022-23	4202.0
401	8335	St. Lucie	1201770	FLORIDA	2017-18	1414.0

402 rows × 6 columns

In [22]:

```
query = """SELECT lea_name
        FROM(SELECT *
        FROM district
        WHERE state == 'FLORIDA') a
        GROUP BY lea_name
        """

lea_name = sqldf(query)
lea_name
```

Out[22]:

	lea_name
0	Alachua
1	Baker
2	Bay
3	Bradford
4	Brevard
...	...
63	Union
64	Volusia
65	Wakulla
66	Walton
67	Washington

68 rows × 1 columns

The highest enrollment in Florida district and the local education agency name (lea_name)

In [23]:

```
query = """SELECT lea_name, SUM(homeschool_students) as Sum_homeschool_students
        FROM(SELECT *
        FROM district
        WHERE state == 'FLORIDA') a
        GROUP BY lea_name
        ORDER BY Sum_homeschool_students asc;
        """

lea_name_sum = sqldf(query)
lea_name_sum
```

Out[23]:

	lea_name	Sum_homeschool_students
0	Franklin	169.0
1	Glades	398.0
2	Liberty	454.0
3	Lafayette	558.0
4	Hamilton	600.0
...	...	...
63	Broward	45036.0
64	Palm Beach	45531.0
65	Orange	46317.0
66	Duval	48072.0
67	Hillsborough	55389.0

68 rows × 2 columns

local education agency (lea) with the lowest enrollment

In [24]:

```
query = """SELECT *
          FROM(SELECT *
          FROM district
          WHERE state == 'FLORIDA') a
          WHERE lea_name == 'Franklin'
          """

Franklin = sqldf(query)
Franklin
```

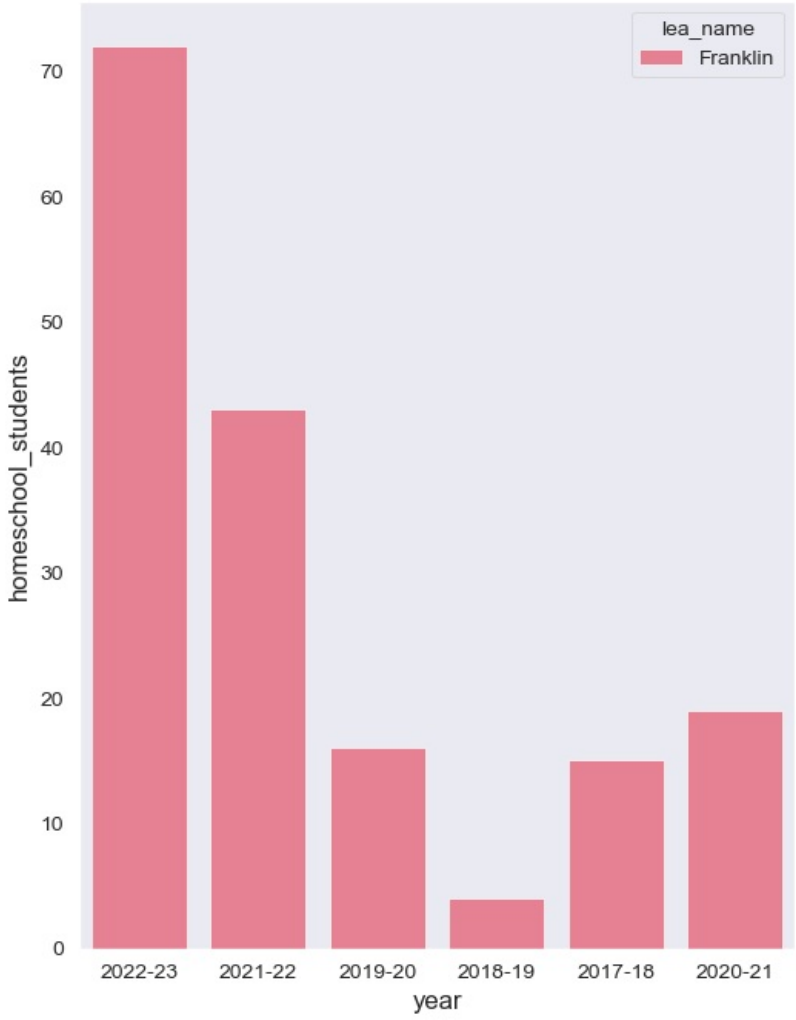
Out[24]:

	index	lea_name	lea_id	state	year	homeschool_students
0	8239	Franklin	1200570	FLORIDA	2022-23	72.0
1	8241	Franklin	1200570	FLORIDA	2021-22	43.0
2	8242	Franklin	1200570	FLORIDA	2019-20	16.0
3	8243	Franklin	1200570	FLORIDA	2018-19	4.0
4	8244	Franklin	1200570	FLORIDA	2017-18	15.0
5	8249	Franklin	1200570	FLORIDA	2020-21	19.0

In [25]:

```
import numpy as np
import matplotlib.pylab as plt
import seaborn as sns
plt.style.use('ggplot')

sns.set_style("ticks")
sns.set_style("dark")
sns.set_palette("husl")
plt.figure(figsize=(6, 8))
sns.barplot(x="year", y="homeschool_students",hue="lea_name" ,data=Franklin)
plt.show()
```



local education agency (lea) with the highest enrollment

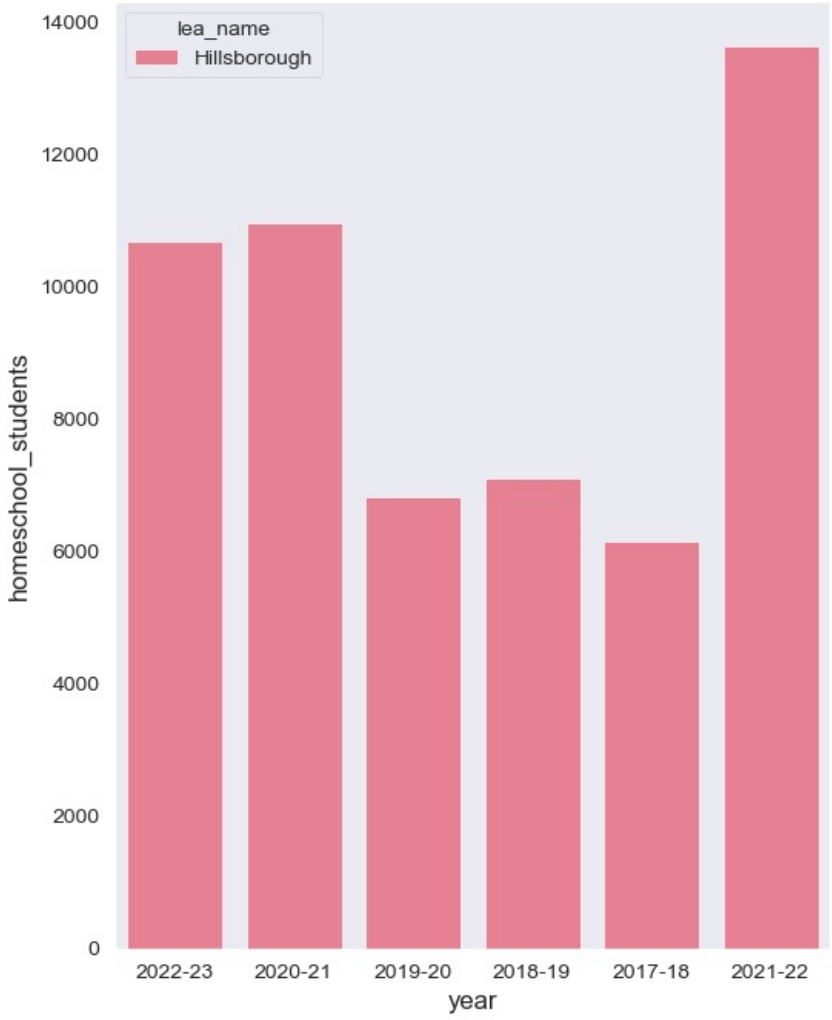
In [26]:
```
query = """SELECT *
         FROM(SELECT *
         FROM district
         WHERE state == 'FLORIDA') a
         WHERE lea_name == 'Hillsborough'
         """

Hillsborough = sqldf(query)
Hillsborough
```

Out[26]:

	index	lea_name	lea_id	state	year	homeschool_students
0	7981	Hillsborough	1200870	FLORIDA	2022-23	10680.0
1	7982	Hillsborough	1200870	FLORIDA	2020-21	10964.0
2	7983	Hillsborough	1200870	FLORIDA	2019-20	6837.0
3	7984	Hillsborough	1200870	FLORIDA	2018-19	7117.0
4	7985	Hillsborough	1200870	FLORIDA	2017-18	6150.0
5	7989	Hillsborough	1200870	FLORIDA	2021-22	13641.0

In [27]:
```
sns.set_style("ticks")
sns.set_style("dark")
sns.set_palette("husl")
plt.figure(figsize=(6, 8))
sns.barplot(x="year", y="homeschool_students",hue="lea_name" ,data=Hillsborough)
plt.show()
```



State with the lowest enrollment is District of Columbia

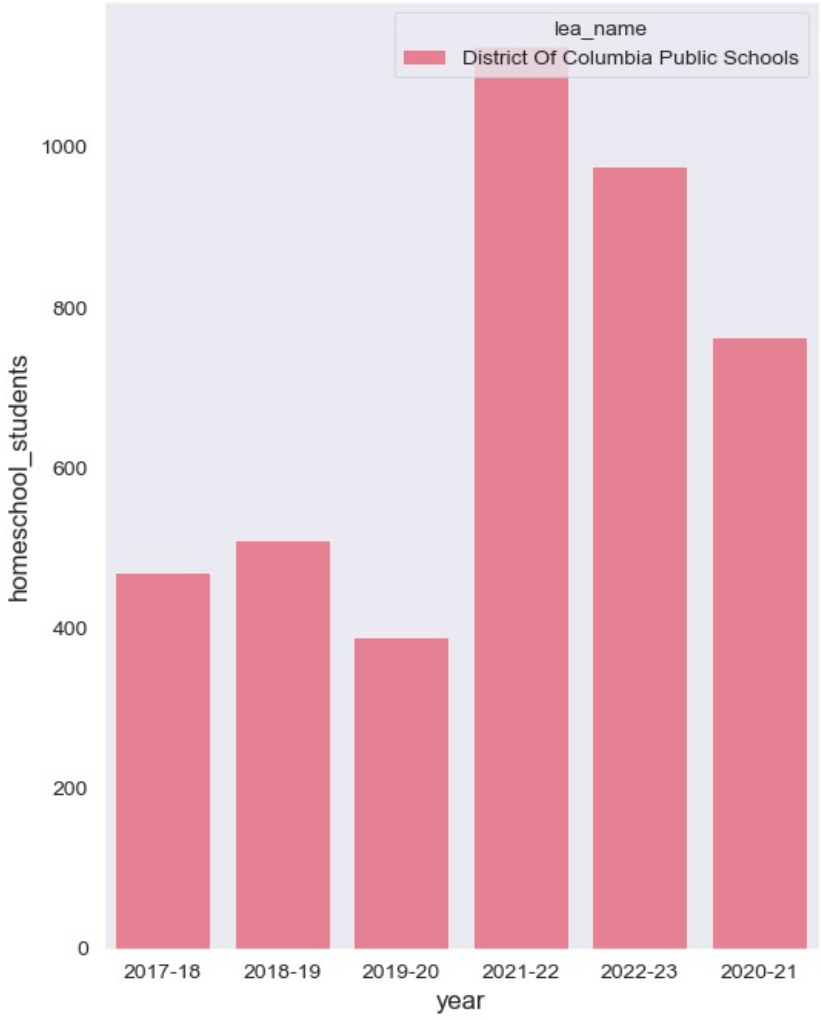
In [28]:
```
query = """SELECT *
         FROM district
         WHERE state == 'DISTRICT OF COLUMBIA'
         """

District_of_Columbia = sqldf(query)
District_of_Columbia
```

Out[28]:

	index	lea_name	lea_id	state	year	homeschool_students
0	7832	District Of Columbia Public Schools	1100030	DISTRICT OF COLUMBIA	2017-18	469.0
1	7833	District Of Columbia Public Schools	1100030	DISTRICT OF COLUMBIA	2018-19	510.0
2	7834	District Of Columbia Public Schools	1100030	DISTRICT OF COLUMBIA	2019-20	389.0
3	7835	District Of Columbia Public Schools	1100030	DISTRICT OF COLUMBIA	2021-22	1126.0
4	7836	District Of Columbia Public Schools	1100030	DISTRICT OF COLUMBIA	2022-23	977.0
5	7837	District Of Columbia Public Schools	1100030	DISTRICT OF COLUMBIA	2020-21	764.0

In [29]:
```
sns.set_style("ticks")
sns.set_style("dark")
sns.set_palette("husl")
plt.figure(figsize=(6, 8))
sns.barplot(x="year", y="homeschool_students",hue="lea_name" ,data=District_of_Columbia)
plt.show()
```



In [69]:
```
query="""SELECT *,
CASE
  WHEN  homeschool_students < 100 THEN 'Low attendance'
  WHEN homeschool_students  <= 1000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM district;
"""

Grade_homeschool_students_district = sqldf(query)
Grade_homeschool_students_district
```

Out[69]:

	index	lea_name	lea_id	state	year	homeschool_students	rate_of_attendance
0	0	Alma School District	502250	ARKANSAS	2017-18	142.0	High attendance
1	1	Mulberry/Pleasant View Bi-County Schools	510290	ARKANSAS	2021-22	71.0	Low attendance
2	2	Mulberry School District	510290	ARKANSAS	2020-21	73.0	Low attendance
3	3	Mulberry School District	510290	ARKANSAS	2019-20	54.0	Low attendance
4	4	Mulberry School District	510290	ARKANSAS	2018-19	37.0	Low attendance
...	...	...	...	...	...	...	...
37162	37162	Platte County School District #2	5603180	WYOMING	2018-19	9.0	Low attendance
37163	37163	Platte County School District #2	5603180	WYOMING	2019-20	0.0	Low attendance
37164	37164	Platte County School District #2	5603180	WYOMING	2020-21	21.0	Low attendance
37165	37165	Sublette County School District #1	5604860	WYOMING	2017-18	56.0	Low attendance
37166	37166	Washakie County School District #2	5605820	WYOMING	2017-18	0.0	Low attendance

37167 rows × 7 columns

In [72]:
```
query="""SELECT *,
CASE
  WHEN  homeschool_students < 100 THEN 'Low attendance'
  WHEN homeschool_students  <= 1000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM district
WHERE homeschool_students < 100;
```

```
"""

Grade_homeschool_students_Low_attendance = sqldf(query)
Grade_homeschool_students_Low_attendance
```

Out[72]:

	index	lea_name	lea_id	state	year	homeschool_students	rate_of_attendance
0	1	Mulberry/Pleasant View Bi-County Schools	510290	ARKANSAS	2021-22	71.0	Low attendance
1	2	Mulberry School District	510290	ARKANSAS	2020-21	73.0	Low attendance
2	3	Mulberry School District	510290	ARKANSAS	2019-20	54.0	Low attendance
3	4	Mulberry School District	510290	ARKANSAS	2018-19	37.0	Low attendance
4	5	Mulberry School District	510290	ARKANSAS	2017-18	50.0	Low attendance
...	...	...	...	...	...	...	...
28820	37162	Platte County School District #2	5603180	WYOMING	2018-19	9.0	Low attendance
28821	37163	Platte County School District #2	5603180	WYOMING	2019-20	0.0	Low attendance
28822	37164	Platte County School District #2	5603180	WYOMING	2020-21	21.0	Low attendance
28823	37165	Sublette County School District #1	5604860	WYOMING	2017-18	56.0	Low attendance
28824	37166	Washakie County School District #2	5605820	WYOMING	2017-18	0.0	Low attendance

28825 rows × 7 columns

In [73]:

```
query="""SELECT *,
CASE
  WHEN  homeschool_students < 100 THEN 'Low attendance'
  WHEN homeschool_students  <= 1000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM district
WHERE homeschool_students > 1000;
"""

Grade_homeschool_students_Highest_attendance = sqldf(query)
Grade_homeschool_students_Highest_attendance
```

Out[73]:

	index	lea_name	lea_id	state	year	homeschool_students	rate_of_attendance
0	534	Pulaski County Special School District	511850	ARKANSAS	2022-23	1111.0	Highest attendance
1	536	Pulaski County Special School District	511850	ARKANSAS	2021-22	1095.0	Highest attendance
2	545	Pulaski Co. Spec. School Dist.	511850	ARKANSAS	2020-21	1233.0	Highest attendance
3	676	Rogers School District	511970	ARKANSAS	2020-21	1044.0	Highest attendance
4	1021	Bentonville School District	503060	ARKANSAS	2022-23	1503.0	Highest attendance
...	...	...	...	...	...	...	...
621	32786	Loudoun Co Pblc Schs	5102250	VIRGINIA	2017-18	1765.0	Highest attendance
622	32787	Loudoun Co Pblc Schs	5102250	VIRGINIA	2018-19	1833.0	Highest attendance
623	32912	Puyallup School District	5306960	WASHINGTON	2020-21	1033.0	Highest attendance
624	33685	Spokane School District	5308250	WASHINGTON	2020-21	1146.0	Highest attendance
625	36777	Milwaukee School District	5509600	WISCONSIN	2022-23	1052.0	Highest attendance

626 rows × 7 columns

In [75]:

```
query="""SELECT *
FROM(SELECT *,
CASE
  WHEN  homeschool_students < 10000 THEN 'Low attendance'
  WHEN homeschool_students  <= 100000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM district) a
WHERE rate_of_attendance == 'High attendance';
"""

Grade_homeschool_students_High_attendance = sqldf(query)
Grade_homeschool_students_High_attendance
```

Out[75]:

	index	lea_name	lea_id	state	year	homeschool_students	rate_of_attendance
0	7981	Hillsborough	1200870	FLORIDA	2022-23	10680.0	High attendance
1	7982	Hillsborough	1200870	FLORIDA	2020-21	10964.0	High attendance
2	7989	Hillsborough	1200870	FLORIDA	2021-22	13641.0	High attendance
3	8153	Broward	1200180	FLORIDA	2021-22	10412.0	High attendance

In [ ]:

home_school_state

In [30]:
```
import pandas as pd
data_state = pd.read_csv('home_school_state.csv')
data_state=data_state.sort_values(by=['state'],ascending=True)
data_state
```

Out[30]:

	state	year	homeschool_students
51	ARKANSAS	2018-19	21959.0
84	ARKANSAS	2019-20	22249.0
183	ARKANSAS	2022-23	29762.0
117	ARKANSAS	2020-21	30267.0
18	ARKANSAS	2017-18	20331.0
...	...	...	...
116	WYOMING	2020-21	3884.0
83	WYOMING	2019-20	2585.0
50	WYOMING	2018-19	1797.0
17	WYOMING	2017-18	2572.0
182	WYOMING	2022-23	3769.0

198 rows × 3 columns

In [31]:
```
data_state.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 198 entries, 51 to 182
Data columns (total 3 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   state                198 non-null    object 
 1   year                 198 non-null    object 
 2   homeschool_students  192 non-null    float64
dtypes: float64(1), object(2)
memory usage: 6.2+ KB
```

In [32]:
```
data_state=data_state.dropna()
data_state
```

Out[32]:

	state	year	homeschool_students
51	ARKANSAS	2018-19	21959.0
84	ARKANSAS	2019-20	22249.0
183	ARKANSAS	2022-23	29762.0
117	ARKANSAS	2020-21	30267.0
18	ARKANSAS	2017-18	20331.0
...	...	...	...
116	WYOMING	2020-21	3884.0
83	WYOMING	2019-20	2585.0
50	WYOMING	2018-19	1797.0
17	WYOMING	2017-18	2572.0
182	WYOMING	2022-23	3769.0

192 rows × 3 columns

In [33]:
```
data_state.to_csv("data_state.csv",index=False)
```

In [34]:
```
from sqlalchemy import create_engine
import pandas as pd
df = pd.read_csv('data_state.csv')
```

```
from sqlalchemy import create_engine
engine = create_engine('postgresql://postgres:kayode@localhost:5432/Ex1')

df.to_sql("home_school_state1", engine, if_exists='replace')
```

Out[34]: 192

In [35]:
```
sql_query_state = pd.read_sql_query("SELECT * FROM home_school_state1", engine)
sql_query_state
```

Out[35]:

	index	state	year	homeschool_students
0	0	ARKANSAS	2018-19	21959.0
1	1	ARKANSAS	2019-20	22249.0
2	2	ARKANSAS	2022-23	29762.0
3	3	ARKANSAS	2020-21	30267.0
4	4	ARKANSAS	2017-18	20331.0
...	...	...	...	...
187	187	WYOMING	2020-21	3884.0
188	188	WYOMING	2019-20	2585.0
189	189	WYOMING	2018-19	1797.0
190	190	WYOMING	2017-18	2572.0
191	191	WYOMING	2022-23	3769.0

192 rows × 4 columns

In [36]:
```
from pandasql import sqldf
State = sqldf("SELECT * FROM sql_query_state")
State
```

Out[36]:

	index	state	year	homeschool_students
0	0	ARKANSAS	2018-19	21959.0
1	1	ARKANSAS	2019-20	22249.0
2	2	ARKANSAS	2022-23	29762.0
3	3	ARKANSAS	2020-21	30267.0
4	4	ARKANSAS	2017-18	20331.0
...	...	...	...	...
187	187	WYOMING	2020-21	3884.0
188	188	WYOMING	2019-20	2585.0
189	189	WYOMING	2018-19	1797.0
190	190	WYOMING	2017-18	2572.0
191	191	WYOMING	2022-23	3769.0

192 rows × 4 columns

In [37]:
```
query="""SELECT state, SUM(homeschool_students) as Homeschool_Summation
FROM State
GROUP BY state
ORDER BY Homeschool_Summation asc"""

Homeschool_Summation = sqldf(query)
Homeschool_Summation
```

Out[37]:

	state	Homeschool_Summation
0	DISTRICT OF COLUMBIA	4235.0
1	RHODE ISLAND	11860.0
2	WYOMING	18082.0
3	KANSAS	20260.0
4	VERMONT	20437.0
5	NORTH DAKOTA	21725.0
6	DELAWARE	22026.0
7	NEW HAMPSHIRE	23621.0
8	HAWAII	25797.0
9	MAINE	29690.0
10	SOUTH DAKOTA	38316.0
11	MONTANA	41983.0
12	COLORADO	59500.0
13	MASSACHUSETTS	64765.0
14	TENNESSEE	66520.0
15	NEW MEXICO	67315.0
16	NEBRASKA	67593.0
17	MISSISSIPPI	134301.0
18	MINNESOTA	144096.0
19	ARKANSAS	154773.0
20	WISCONSIN	154987.0
21	SOUTH CAROLINA	155294.0
22	LOUISIANA	160000.0
23	PENNSYLVANIA	162080.0
24	WASHINGTON	163713.0
25	KENTUCKY	188024.0
26	MARYLAND	196616.0
27	OHIO	243599.0
28	NEW YORK	243683.0
29	CALIFORNIA	249012.0
30	VIRGINIA	315343.0
31	GEORGIA	479064.0
32	FLORIDA	743022.0

In [38]:

```
query = """SELECT state, COUNT(*) total_states
FROM (SELECT *
FROM State) a
GROUP BY state
ORDER BY total_states asc
"""
total_states = sqldf(query)
total_states
```

Out[38]:

	state	total_states
0	MAINE	3
1	PENNSYLVANIA	5
2	RHODE ISLAND	5
3	TENNESSEE	5
4	ARKANSAS	6
5	CALIFORNIA	6
6	COLORADO	6
7	DELAWARE	6
8	DISTRICT OF COLUMBIA	6
9	FLORIDA	6
10	GEORGIA	6
11	HAWAII	6
12	KANSAS	6
13	KENTUCKY	6
14	LOUISIANA	6
15	MARYLAND	6
16	MASSACHUSETTS	6
17	MINNESOTA	6
18	MISSISSIPPI	6
19	MONTANA	6
20	NEBRASKA	6
21	NEW HAMPSHIRE	6
22	NEW MEXICO	6
23	NEW YORK	6
24	NORTH DAKOTA	6
25	OHIO	6
26	SOUTH CAROLINA	6
27	SOUTH DAKOTA	6
28	VERMONT	6
29	VIRGINIA	6
30	WASHINGTON	6
31	WISCONSIN	6
32	WYOMING	6

In [39]:

```
query = """SELECT d.state, COUNT(*) total_states
FROM (SELECT *
FROM State) d
GROUP BY state
HAVING COUNT(*) < 6
ORDER BY state
"""
Total_states_less_than_six= sqldf(query)
Total_states_less_than_six
```

Out[39]:

	state	total_states
0	MAINE	3
1	PENNSYLVANIA	5
2	RHODE ISLAND	5
3	TENNESSEE	5

In [40]:

```
query = """SELECT state, year,homeschool_students,
FIRST_VALUE(homeschool_students) OVER(PARTITION BY state ORDER BY homeschool_students desc) first_value
FROM State
"""
First_value = sqldf(query)
First_value
```

Out[40]:

	state	year	homeschool_students	first_value
0	ARKANSAS	2020-21	30267.0	30267.0
1	ARKANSAS	2021-22	30205.0	30267.0
2	ARKANSAS	2022-23	29762.0	30267.0
3	ARKANSAS	2019-20	22249.0	30267.0
4	ARKANSAS	2018-19	21959.0	30267.0
...	...	...	...	...
187	WYOMING	2022-23	3769.0	3884.0
188	WYOMING	2021-22	3475.0	3884.0
189	WYOMING	2019-20	2585.0	3884.0
190	WYOMING	2017-18	2572.0	3884.0
191	WYOMING	2018-19	1797.0	3884.0

192 rows × 4 columns

In [41]:

```
query = """SELECT state, SUM(state_count) state_sum
FROM(SELECT state, COUNT(*) OVER(PARTITION BY state) state_count
FROM State) a
GROUP BY state
ORDER BY state_sum
"""
State_sum = sqldf(query)
State_sum
```

Out[41]:

	state	state_sum
0	MAINE	9
1	PENNSYLVANIA	25
2	RHODE ISLAND	25
3	TENNESSEE	25
4	ARKANSAS	36
5	CALIFORNIA	36
6	COLORADO	36
7	DELAWARE	36
8	DISTRICT OF COLUMBIA	36
9	FLORIDA	36
10	GEORGIA	36
11	HAWAII	36
12	KANSAS	36
13	KENTUCKY	36
14	LOUISIANA	36
15	MARYLAND	36
16	MASSACHUSETTS	36
17	MINNESOTA	36
18	MISSISSIPPI	36
19	MONTANA	36
20	NEBRASKA	36
21	NEW HAMPSHIRE	36
22	NEW MEXICO	36
23	NEW YORK	36
24	NORTH DAKOTA	36
25	OHIO	36
26	SOUTH CAROLINA	36
27	SOUTH DAKOTA	36
28	VERMONT	36
29	VIRGINIA	36
30	WASHINGTON	36
31	WISCONSIN	36
32	WYOMING	36

In [42]:
```
query = """SELECT state, SUM(state_count) state_sum
FROM(SELECT state, COUNT(*) OVER(PARTITION BY state) state_count
FROM State) a
GROUP BY state
UNION ALL
SELECT 'TOTAL', COUNT(*)
FROM(SELECT state, COUNT(*) OVER(PARTITION BY state) state_count
FROM State) a
ORDER BY state_sum
"""
State_sum_total = sqldf(query)
State_sum_total
```

Out[42]:

	state	state_sum
0	MAINE	9
1	PENNSYLVANIA	25
2	RHODE ISLAND	25
3	TENNESSEE	25
4	ARKANSAS	36
5	CALIFORNIA	36
6	COLORADO	36
7	DELAWARE	36
8	DISTRICT OF COLUMBIA	36
9	FLORIDA	36
10	GEORGIA	36
11	HAWAII	36
12	KANSAS	36
13	KENTUCKY	36
14	LOUISIANA	36
15	MARYLAND	36
16	MASSACHUSETTS	36
17	MINNESOTA	36
18	MISSISSIPPI	36
19	MONTANA	36
20	NEBRASKA	36
21	NEW HAMPSHIRE	36
22	NEW MEXICO	36
23	NEW YORK	36
24	NORTH DAKOTA	36
25	OHIO	36
26	SOUTH CAROLINA	36
27	SOUTH DAKOTA	36
28	VERMONT	36
29	VIRGINIA	36
30	WASHINGTON	36
31	WISCONSIN	36
32	WYOMING	36
33	TOTAL	192

In [43]:
```
query="""SELECT state, SUM(homeschool_students) sum_homeschool_students
FROM(
SELECT state, homeschool_students,
RANK() OVER(PARTITION BY state ORDER BY homeschool_students desc)
FROM State)a
GROUP BY state
ORDER BY sum_homeschool_students desc"""

sum_homeschool_students = sqldf(query)
sum_homeschool_students
```

Out[43]:

	state	sum_homeschool_students
0	FLORIDA	743022.0
1	GEORGIA	479064.0
2	VIRGINIA	315343.0
3	CALIFORNIA	249012.0
4	NEW YORK	243683.0
5	OHIO	243599.0
6	MARYLAND	196616.0
7	KENTUCKY	188024.0
8	WASHINGTON	163713.0
9	PENNSYLVANIA	162080.0
10	LOUISIANA	160000.0
11	SOUTH CAROLINA	155294.0
12	WISCONSIN	154987.0
13	ARKANSAS	154773.0
14	MINNESOTA	144096.0
15	MISSISSIPPI	134301.0
16	NEBRASKA	67593.0
17	NEW MEXICO	67315.0
18	TENNESSEE	66520.0
19	MASSACHUSETTS	64765.0
20	COLORADO	59500.0
21	MONTANA	41983.0
22	SOUTH DAKOTA	38316.0
23	MAINE	29690.0
24	HAWAII	25797.0
25	NEW HAMPSHIRE	23621.0
26	DELAWARE	22026.0
27	NORTH DAKOTA	21725.0
28	VERMONT	20437.0
29	KANSAS	20260.0
30	WYOMING	18082.0
31	RHODE ISLAND	11860.0
32	DISTRICT OF COLUMBIA	4235.0

In [44]:

```
query="""SELECT state,year, homeschool_students,
CASE
  WHEN  homeschool_students < 10000 THEN 'Low attendance'
  WHEN homeschool_students  <= 100000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM State;
"""

Grade_homeschool_students = sqldf(query)
Grade_homeschool_students
```

Out[44]:

	state	year	homeschool_students	rate_of_attendance
0	ARKANSAS	2018-19	21959.0	High attendance
1	ARKANSAS	2019-20	22249.0	High attendance
2	ARKANSAS	2022-23	29762.0	High attendance
3	ARKANSAS	2020-21	30267.0	High attendance
4	ARKANSAS	2017-18	20331.0	High attendance
...	...	...	...	...
187	WYOMING	2020-21	3884.0	Low attendance
188	WYOMING	2019-20	2585.0	Low attendance
189	WYOMING	2018-19	1797.0	Low attendance
190	WYOMING	2017-18	2572.0	Low attendance
191	WYOMING	2022-23	3769.0	Low attendance

192 rows × 4 columns

In [54]:

```
query="""SELECT state,year, homeschool_students,
CASE
  WHEN  homeschool_students < 10000 THEN 'Low attendance'
  WHEN homeschool_students  <= 100000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM State
WHERE homeschool_students <= 10000;
"""

Grade_homeschool_students_less_10000 = sqldf(query)
Grade_homeschool_students_less_10000
```

Out[54]:

	state	year	homeschool_students	rate_of_attendance
0	COLORADO	2018-19	9284.0	Low attendance
1	COLORADO	2017-18	7387.0	Low attendance
2	COLORADO	2022-23	8674.0	Low attendance
3	COLORADO	2019-20	7880.0	Low attendance
4	DELAWARE	2018-19	2954.0	Low attendance
...	...	...	...	...
74	WYOMING	2020-21	3884.0	Low attendance
75	WYOMING	2019-20	2585.0	Low attendance
76	WYOMING	2018-19	1797.0	Low attendance
77	WYOMING	2017-18	2572.0	Low attendance
78	WYOMING	2022-23	3769.0	Low attendance

79 rows × 4 columns

In [66]:

```
query="""SELECT state,year, homeschool_students,
CASE
  WHEN  homeschool_students < 10000 THEN 'Low attendance'
  WHEN homeschool_students  <= 100000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM State
WHERE homeschool_students > 100000;
"""

Grade_homeschool_students_greater_than_100000 = sqldf(query)
Grade_homeschool_students_greater_than_100000
```

Out[66]:

	state	year	homeschool_students	rate_of_attendance
0	FLORIDA	2019-20	106115.0	Highest attendance
1	FLORIDA	2022-23	154289.0	Highest attendance
2	FLORIDA	2020-21	143431.0	Highest attendance
3	FLORIDA	2021-22	152109.0	Highest attendance

In [68]:

```
query="""SELECT *
FROM(SELECT state,year, homeschool_students,
CASE
  WHEN  homeschool_students < 10000 THEN 'Low attendance'
```

```
  WHEN homeschool_students  <= 100000 THEN 'High attendance'
ELSE
  'Highest attendance'
END as rate_of_attendance
FROM State) a
WHERE rate_of_attendance == 'High attendance';
"""

Grade_homeschool_students_High_attendance = sqldf(query)
Grade_homeschool_students_High_attendance
```

Out[68]:

	state	year	homeschool_students	rate_of_attendance
0	ARKANSAS	2018-19	21959.0	High attendance
1	ARKANSAS	2019-20	22249.0	High attendance
2	ARKANSAS	2022-23	29762.0	High attendance
3	ARKANSAS	2020-21	30267.0	High attendance
4	ARKANSAS	2017-18	20331.0	High attendance
...	...	...	...	...
104	WISCONSIN	2021-22	29402.0	High attendance
105	WISCONSIN	2020-21	31878.0	High attendance
106	WISCONSIN	2018-19	21577.0	High attendance
107	WISCONSIN	2017-18	21633.0	High attendance
108	WISCONSIN	2022-23	28853.0	High attendance

109 rows × 4 columns

In [45]:

```
query="""SELECT district.year
FROM district
WHERE
EXISTS
(SELECT state
FROM State
WHERE homeschool_students = district.homeschool_students)
GROUP BY year;
"""
Year = sqldf(query)
Year
```

Out[45]:

	year
0	2017-18
1	2018-19
2	2019-20
3	2020-21
4	2021-22
5	2022-23

In [50]:

```
query= """
SELECT state, homeschool_students, a.RANK
FROM(SELECT state, homeschool_students,
RANK() OVER(PARTITION BY state ORDER BY homeschool_students desc) RANK
FROM State) a
WHERE RANK= 1
ORDER BY homeschool_students
"""

homeschool_students_rank1 = sqldf(query)
homeschool_students_rank1
```

Out[50]:

	state	homeschool_students	RANK
0	DISTRICT OF COLUMBIA	1126.0	1
1	RHODE ISLAND	3396.0	1
2	WYOMING	3884.0	1
3	NORTH DAKOTA	4657.0	1
4	DELAWARE	4905.0	1
5	VERMONT	5463.0	1
6	NEW HAMPSHIRE	6114.0	1
7	HAWAII	6232.0	1
8	KANSAS	7036.0	1
9	SOUTH DAKOTA	9120.0	1
10	MONTANA	9868.0	1
11	MAINE	12044.0	1
12	NEBRASKA	14780.0	1
13	NEW MEXICO	15629.0	1
14	COLORADO	15773.0	1
15	MASSACHUSETTS	17207.0	1
16	TENNESSEE	18525.0	1
17	ARKANSAS	30267.0	1
18	MISSISSIPPI	30358.0	1
19	MINNESOTA	30955.0	1
20	WISCONSIN	31878.0	1
21	SOUTH CAROLINA	31998.0	1
22	LOUISIANA	32728.0	1
23	KENTUCKY	39544.0	1
24	WASHINGTON	39843.0	1
25	PENNSYLVANIA	42766.0	1
26	MARYLAND	44931.0	1
27	OHIO	51502.0	1
28	NEW YORK	54414.0	1
29	CALIFORNIA	59275.0	1
30	VIRGINIA	65571.0	1
31	GEORGIA	91515.0	1
32	FLORIDA	154289.0	1

In [49]:

```
query= """
SELECT state, homeschool_students, a.RANK
FROM(SELECT state, homeschool_students,
RANK() OVER(PARTITION BY state ORDER BY homeschool_students desc) RANK
FROM State) a
WHERE RANK= 6
ORDER BY homeschool_students
"""

homeschool_students_rank6 = sqldf(query)
homeschool_students_rank6
```

Out[49]:

	state	homeschool_students	RANK
0	DISTRICT OF COLUMBIA	389.0	6
1	WYOMING	1797.0	6
2	KANSAS	1947.0	6
3	VERMONT	2338.0	6
4	NORTH DAKOTA	2683.0	6
5	HAWAII	2726.0	6
6	DELAWARE	2864.0	6
7	NEW HAMPSHIRE	2875.0	6
8	SOUTH DAKOTA	4696.0	6
9	MONTANA	5390.0	6
10	COLORADO	7387.0	6
11	MASSACHUSETTS	7435.0	6
12	NEW MEXICO	8811.0	6
13	NEBRASKA	8919.0	6
14	MISSISSIPPI	18323.0	6
15	MINNESOTA	18988.0	6
16	ARKANSAS	20331.0	6
17	SOUTH CAROLINA	20611.0	6
18	WASHINGTON	20844.0	6
19	WISCONSIN	21577.0	6
20	LOUISIANA	21717.0	6
21	KENTUCKY	22998.0	6
22	CALIFORNIA	25423.0	6
23	NEW YORK	25541.0	6
24	MARYLAND	26040.0	6
25	OHIO	30923.0	6
26	VIRGINIA	43361.0	6
27	GEORGIA	70414.0	6
28	FLORIDA	89817.0	6

In [51]:

```
query= """
SELECT state, homeschool_students, a.RANK
FROM(SELECT state, homeschool_students,
RANK() OVER(PARTITION BY state ORDER BY homeschool_students desc) RANK
FROM State) a
WHERE RANK= 5
ORDER BY homeschool_students
"""

homeschool_students_rank5 = sqldf(query)
homeschool_students_rank5
```

Out[51]:

	state	homeschool_students	RANK
0	DISTRICT OF COLUMBIA	469.0	5
1	RHODE ISLAND	1648.0	5
2	KANSAS	2143.0	5
3	VERMONT	2389.0	5
4	WYOMING	2572.0	5
5	NORTH DAKOTA	2808.0	5
6	DELAWARE	2954.0	5
7	NEW HAMPSHIRE	3252.0	5
8	HAWAII	3303.0	5
9	SOUTH DAKOTA	5046.0	5
10	MONTANA	5743.0	5
11	MASSACHUSETTS	7648.0	5
12	COLORADO	7880.0	5
13	NEBRASKA	9030.0	5
14	TENNESSEE	9069.0	5
15	NEW MEXICO	9793.0	5
16	MISSISSIPPI	18828.0	5
17	MINNESOTA	19228.0	5
18	SOUTH CAROLINA	20752.0	5
19	WASHINGTON	21022.0	5
20	WISCONSIN	21633.0	5
21	ARKANSAS	21959.0	5
22	LOUISIANA	23161.0	5
23	KENTUCKY	24573.0	5
24	PENNSYLVANIA	25378.0	5
25	CALIFORNIA	26345.0	5
26	NEW YORK	26805.0	5
27	MARYLAND	27527.0	5
28	OHIO	32887.0	5
29	VIRGINIA	43505.0	5
30	GEORGIA	75111.0	5
31	FLORIDA	97261.0	5

Overall, pandasql is a useful tool for data analysts and scientists who are proficient in SQL and want to apply SQL queries to manipulate pandas DataFrames effectively.

In [ ]: