DataFrames and CSVs DataFrame as `df1`

```
-- Isolate out diabetes studies

SELECT *
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE topic = 'Diabetes';
```

	Y...		Loca...	LocationDesc	D...		Question		DataValueUnit	DataValueType
0	2019	2019	AR	Arkansas	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
1	2019	2019	ID	Idaho	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
2	2019	2019	IA	Iowa	NVSS	Diabetes	Diabetes mortality among all people, underly...	null	Number	Number
3	2019	2019	KS	Kansas	NVSS	Diabetes	Diabetes mortality among all people, underly...	null	Number	Number
4	2019	2019	NE	Nebraska	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
5	2019	2019	NE	Nebraska	NVSS	Diabetes	Diabetes mortality among all people, underly...	null	Number	Number
6	2019	2019	NJ	New Jersey	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
7	2019	2019	NC	North Carolina	NVSS	Diabetes	Diabetes mortality among all people, underly...	null	Number	Number
8	2019	2019	OK	Oklahoma	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
9	2019	2019	RI	Rhode Island	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
10	2019	2019	US	United States	NVSS	Diabetes	Diabetes mortality among all people, underly...	null	Number	Number
11	2019	2019	VI	Virgin Islands	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
12	2020	2020	RI	Rhode Island	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
13	2020	2020	VT	Vermont	NVSS	Diabetes	Diabetes mortality among all people, underly...	null	Number	Number
14	2020	2020	WV	West Virginia	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence
15	2021	2021	AL	Alabama	BRFSS	Diabetes	Diabetes among adults	null	%	Age-adjusted Prevalence
16	2021	2021	IN	Indiana	BRFSS	Diabetes	Diabetes among adults	null	%	Crude Prevalence

Rows: 2,941 ⚠ truncated from 17,318 rows

DataFrames and CSVs DataFrame as `df`

```
-- Let's see what questions study participants were asked about their diabetes

SELECT
    DISTINCT(Question)
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE topic = 'Diabetes';
```

index	Question
0	Diabetes mortality among all people, underlying or contributing cause
1	Gestational diabetes among women with a recent live birth
2	Diabetes among adults
3	Diabetic ketoacidosis mortality among all people, underlying or contributing cause

Rows: 4

DataFrames and CSVs DataFrame as `df2`

```
-- Okay, I think that the question "Diabetes among adults" is the most broad and will give us the most insight into U.S. diabetes rates
    --From this, lets filter for only by that question and see what the study was

SELECT
    DataSource
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE Question = 'Diabetes among adults'
GROUP BY DataSource;
```

index	DataSource
0	BRFSS

Rows: 1

DataFrames and CSVs DataFrame as `df3`

```
--Okay, there is one ongoing study - let's group by year and see how long we have results for

SELECT
    YearStart,
    YearEnd
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE Question = 'Diabetes among adults'
GROUP BY YearStart,YearEnd;
```

index	YearStart	YearEnd
0	2020	
1	2022	
2	2019	
3	2021	

Rows: 4

DataFrames and CSVs DataFrame as `df4`

```
-- Okay, each year has it's own set of data, this means that we can disregard yearstart and focus on year end column
    --Next I want to check the datavaluetype for this study to see whether the groups can be combined and represent the same type of data

SELECT
    DataValueType
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE Question = 'Diabetes among adults'
GROUP BY DataValueType;
```

index	DataValueType
0	Crude Prevalence
1	Age-adjusted Prevalence

Rows: 2

DataFrames and CSVs DataFrame as `df6`

```sql
-- I want to check if dataalt and datavvalues vary for the group I want to investigate

SELECT *
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE DataValue != DataValueAlt AND Question = 'Diabetes among adults';
```

Your query ran successfully but returned no results.

DataFrames and CSVs DataFrame as `df7`

```sql
--Okay, I don't want to include any adjusted values and only want to focus on crude numbers
    -- I also noticed there are stratifications for the reports so let's see what those are and how to further group the data

SELECT
    DISTINCT StratificationCategory1,
    StratificationCategory2,
    StratificationCategory3
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE Question = 'Diabetes among adults' AND DataValueType LIKE 'Crude%';
```

	StratificationCategory1	StratificationCategory2	StratificationCategory3
0	Sex	null	null
1	Race/Ethnicity	null	null
2	Age	null	null
3	Overall	null	null

Rows: 4

DataFrames and CSVs DataFrame as `df5`

```
-- We want overall data so we will select only those values
    --Now I want to select responses with non-null values and group by state and year and rank to see the best and worst over time

SELECT
    RANK() OVER (PARTITION BY YearEnd ORDER BY ROUND(AVG(DataValue),2) DESC) AS rank_within_year,
    YearEnd AS study_year,
    LocationDesc,
    LocationAbbr,
    ROUND(AVG(DataValue),2) AS avg_percent_diabetes
FROM 'U.S._Chronic_Disease_Indicators.csv'
WHERE Question = 'Diabetes among adults' AND DataValueType
        LIKE 'Crude%' AND StratificationCategory1 = 'Overall' AND DataValue IS NOT NULL
        AND LocationDesc NOT LIKE 'Uni%'
GROUP BY LocationDesc, YearEnd, LocationAbbr
ORDER BY YearEnd DESC, rank_within_year;
```

	rank_within...	s...	LocationDesc	Loca...	avg_percent_diab...
0	1	2022	Guam	GU	21.6
1	2	2022	Puerto Rico	PR	17.7
2	3	2022	West Virginia	WV	17.4
3	4	2022	Virgin Islands	VI	15.9
4	5	2022	Arkansas	AR	15.7
5	6	2022	Alabama	AL	15.5
6	7	2022	Mississippi	MS	15.3
7	8	2022	Kentucky	KY	14.8
8	8	2022	Tennessee	TN	14.8
9	10	2022	Louisiana	LA	14.7
10	11	2022	Texas	TX	13.9
11	11	2022	Delaware	DE	13.9
12	13	2022	Oklahoma	OK	13.3
13	14	2022	Ohio	OH	13.1
14	15	2022	South Carolina	SC	12.9
15	16	2022	Virginia	VA	12.8
16	17	2022	Indiana	IN	12.7

Rows: 212