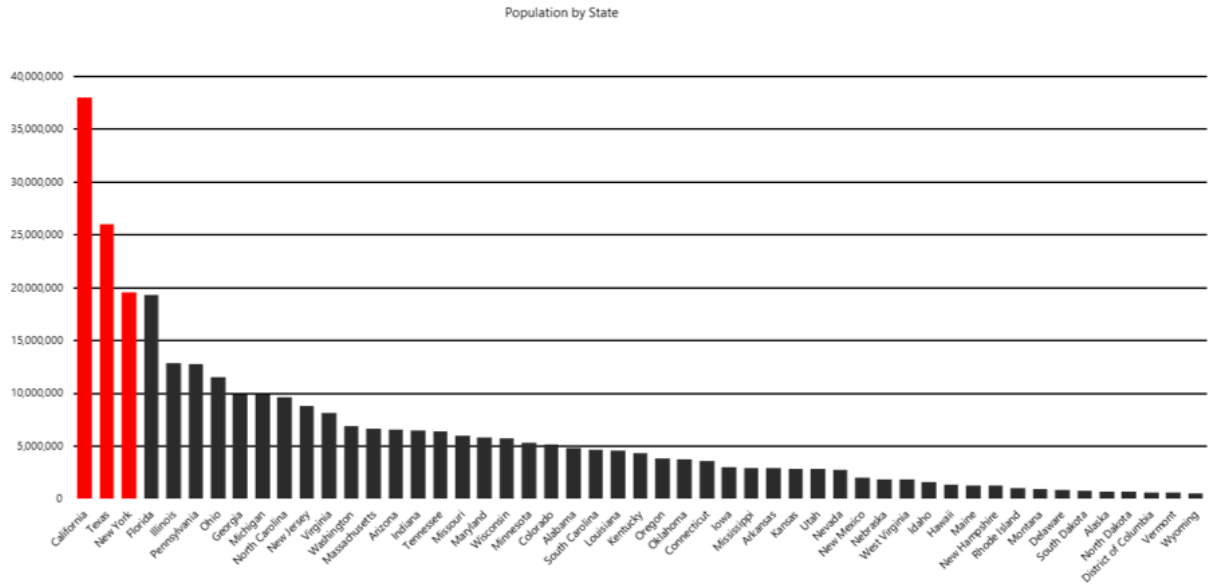
Understand skewed data and space usage

Note:

You are not required to complete the processes, tasks, activities, or steps presented in this example. The various samples provided are for illustrative purposes only and it's likely that if you try this out you will encounter issues in your system.

In simple terms, data skew is an over-represented value. Imagine that you have assigned 50 tax examiners to audit tax returns, one examiner for each US state. The Wyoming examiner, because the population there is small, has little to do. In California, however, the examiner is kept very busy because of the state's large population.



Data-skew problem example

In our scenario, the data is unevenly distributed across all tax examiners, which means that some examiners must work more than others. In your job, you frequently experience situations like the tax-examiner example here. In more technical terms, one vertex gets much more data than its peers, a condition that makes the vertex work more than the others and that eventually slows down an entire job. What's worse, the job might fail because vertices might have, for example, a 5-hour runtime limitation and a 6-GB memory limitation.

A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.

```
1
2
```

```
-- Find data skew for a distributed table
DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales');
```

Another aspect of data storage in Azure Synapse dedicated SQL pools is to monitor the table data space usage and observe its relationship with different table distribution types. Additionally, it is helpful to know the number of rows and the storage space used for indexing. Below is a list of System Dynamic Management Views (DMVs) that you can use to dig for the information. During the next exercise you will create a view using these DMVs to get a better view of the data.

Table Name	Description
sys.schemas	All schemas in the database.
sys.tables	All tables in the database.
sys.indexes	All indexes in the database.
sys.columns	All columns in the database.
sys.pdw_table_mappings	Maps each table to local tables on physical nodes and distributions.
sys.pdw_nodes_tables	Contains information on each local table in each distribution.
sys.pdw_table_distribution_properties	Holds distribution information for tables (the type of distribution tables have).
sys.pdw_column_distribution_properties	Holds distribution information for columns. Filtered to include only columns used to distribute their parent tables (distribution_ordinal = 1).
sys.pdw_distributions	Holds information about the distributions from the SQL pool.
sys.dm_pdw_nodes	Holds information about the nodes from the SQL pool. Filtered to include only compute nodes (type = COMPUTE).
sys.dm_pdw_nodes_db_partition_stats	Returns page and row-count information for every partition in the current database.