

[milan@datajoin.net](mailto:milan@datajoin.net)

<http://datajoin.net>

DataJoin

# Correlate company data with external data

[Document subtitle]

Milan Patel

5-2-2025

## Table of

## Contents

Overview .....	3
What should be matched (correlated) .....	4
Performance measures (dimensions) categories .....	4
Financial Measure .....	4
Example Operational performance measure .....	5
Example external data about society (region) in which your company operates .....	5
Examples of open-source data.....	5
Basic concept of comparing data at the same level .....	7
Example matching (compare, correlate, etc.) sales by county and year and GDP by county and year .....	7
Example matching (compare, correlate, etc.) of sales by county and year and population by county.....	8
Example matching (compare, correlate, etc.) of sales by state and year and GDP by county and year .....	9
Example matching (compare, correlate, etc.) of sales by state and year and population by county .....	10
Rank / Scale / Percentile .....	11
GDP and Unemployment rank comparison.....	11
Sales and population rank comparison .....	12
Trend over time.....	13
Knowledge based on history vs hypothesis .....	14
Principal Component Analysis .....	14
ETL (extract, transform, and load) Steps for external data.....	14
Task1 Use API to get data in Json, xml, or csv format.....	14

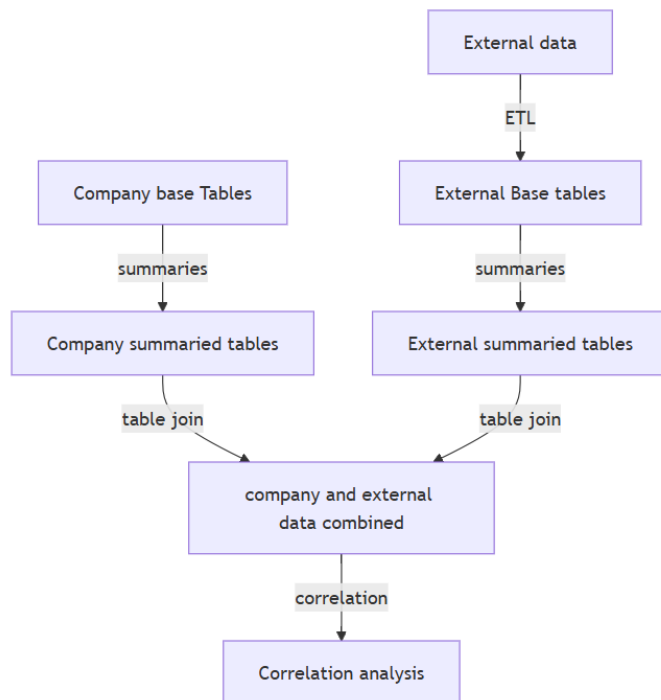
Task2 Convert Json or xml format to csv format.....	15
Task3 Copy and or add attributes.....	15
Task4 Transform values in attributes .....	15
Task5 Upload base table .....	15
Task6 Create summary table with rank/ percentile / scaled column .....	15
Python code to new table with sum and rank .....	16
Task 7 Join table to study correlation .....	16

## Overview

**Problem:** Business intelligence (BI) or customer relationship management (CRM) software provides data on regional sales but lacks comprehensive details about the region, which is necessary to understand variations in sales by region. Furthermore, when traveling, house hunting, or starting a business in a specific region, it is essential to obtain more in-depth information about that area.

**Solution:** Provide comprehensive regional details, including crime rates, education levels, environmental conditions, and local economic factors. This information can be integrated with BI (Business Intelligence) or CRM (Customer Relationship Management) data to identify potential correlations or discrepancies.

DataJoin.net provides in-depth education and consultation on matching CRM data with regional open-source information.



<https://github.com/milan888-design/correlate-data>

## What should be matched (correlated)

A company's performance measures should be compared / matched / correlated with regional demographics and other measurements (GDP, Unemployment, etc.)

## Performance measures (dimensions) categories

As per data warehouse star schema, the fact table can be associated with multiple dimensions. Each dimension has a hierarchy that can be used to drill down to find the value at a specific level.

### Examples of dimensions:

Organization  
 Product category  
 Region  
 Time  
 Financial  
 Customer satisfaction  
 Operation

Each of the above performances can have hierarchy as below.

Example of Time hierarchy: Year – Quarter - Month

Region hierarchy: Country – State - County (USA - New Jersey - Monmouth)

## Standard performance measures in Data warehouse or Business Intelligence

### Financial Measure

Operating Cost
Profit
Product Cost
Sales
Turnover
Liability
Assets

Return On Investment
----------------------

### Example Operational performance measure

Quality
Price
Volume
Variety
Delivery
Customer Satisfaction Measure

There are numerous other operational performance measures.

### Example external data about society (region) in which your company operates

Population

GDP

Income

Unemployment rate

There are many more data available by region (country, state, county, city, zip code)

### Examples of open-source data

[https://en.wikipedia.org/wiki/List\\_of\\_federal\\_agencies\\_in\\_the\\_United\\_States](https://en.wikipedia.org/wiki/List_of_federal_agencies_in_the_United_States)

United States Congress

Federal judiciary of the United States

Executive Office of the President

United States Department of Agriculture (USDA)

United States Department of Commerce

United States Department of Defense (DOD)

United States Department of Education

United States Department of Energy

United States Department of Health and Human Services

United States Department of Homeland Security

United States Department of Housing and Urban Development



CopyRight 2025

<http://datajoin.net>

[milan@datajoin.net](mailto:milan@datajoin.net)

United States Department of the Interior (DOI)  
 United States Department of Justice  
 United States Department of Labor (DOL)  
 United States Department of State (DOS)  
 United States Department of Transportation  
 United States Department of the Treasury  
 United States Department of Veterans Affairs  
 Independent agencies and government-owned corporations  
 Joint programs and interagency agencies  
 Special Inspector General Office  
 Quasi-official agencies

<https://www.usa.gov/federal-agencies/e#current-letter>

**Example results from the above URL for starting with “B”.**

Bureau of Economic Analysis (BEA)  
 Bureau of Engraving and Printing (BEP)  
 Bureau of Indian Affairs (BIA)  
 Bureau of Industry and Security (BIS)  
 Bureau of International Labor Affairs (ILAB)  
 Bureau of Justice Statistics (BJS)  
 Bureau of Labor Statistics (BLS)

**Example results from the above URL for starting with “C”.**

Capitol Police  
 Capitol Visitor Center  
 Career, Technical, and Adult Education, Office of (OCTAE)  
 Census Bureau  
 Center for Food Safety and Applied Nutrition (CFSAN)  
 Center for Nutrition Policy and Promotion (CNPP)  
 Center for Parent Information and Resources (CPIR)  
 Centers for Disease Control and Prevention (CDC)  
 Etc.

The above is partial list there are other open sources listed at the above URL.

## Basic concept of comparing data at the same level

Example matching (compare, correlate, etc.) sales by county and year and GDP by county and year

**sales summary by state, county and year**

product_type	state	county_state	year	thousands of \$
laptop	CA	Los Angeles, CA	2023	200
laptop	IL	Cook, IL	2023	300
laptop	TX	Harris, TX	2023	500
laptop	NY	New York, NY	2023	600
laptop	AZ	Maricopa, AZ	2023	100
laptop	CA	Santa Clara, CA	2023	250
laptop	CA	Orange, CA	2023	600
laptop	WA	King, WA	2023	600
laptop	CA	San Diego, CA	2023	200
laptop	TX	Dallas, TX	2023	400

**GDP by state and county**

state	county_state	year	GDP in thousands of \$
CA	Los Angeles, CA	2023	756659481
IL	Cook, IL	2023	396470899
TX	Harris, TX	2023	357130669
NY	New York, NY	2023	343437442
AZ	Maricopa, AZ	2023	312350417
CA	Santa Clara, CA	2023	283522548
CA	Orange, CA	2023	278760587
WA	King, WA	2023	277642267
CA	San Diego, CA	2023	258725373
TX	Dallas, TX	2023	207533772

**sales summary by state, county and year matched with GDP**

product_type	state	county_state	year	thousands of \$	GDP in thousands of \$
laptop	CA	Los Angeles, CA	2023	200	756659481
laptop	IL	Cook, IL	2023	300	396470899
laptop	TX	Harris, TX	2023	500	357130669
laptop	NY	New York, NY	2023	600	343437442
laptop	AZ	Maricopa, AZ	2023	100	312350417
laptop	CA	Santa Clara, CA	2023	250	283522548
laptop	CA	Orange, CA	2023	600	278760587
laptop	WA	King, WA	2023	600	277642267
laptop	CA	San Diego, CA	2023	200	258725373
laptop	TX	Dallas, TX	2023	400	207533772

If you have sales of a product type by state, county and year  
and

if you have prepared GDP (gross domestic product) by state, county and year  
then only, comparison is meaningful.

**One can explore the question: Is my laptop sales correlated with GDP at county level?**





Example matching (compare, correlate, etc.) of sales by county and year and population by county

**sales summary**

product_type	state	county_state	year	thousands of \$
laptop	CA	Los Angeles, CA	2023	200
laptop	IL	Cook, IL	2023	300
laptop	TX	Harris, TX	2023	500
laptop	NY	New York, NY	2023	600
laptop	AZ	Maricopa, AZ	2023	100
laptop	CA	Santa Clara, CA	2023	250
laptop	CA	Orange, CA	2023	600
laptop	WA	King, WA	2023	600
laptop	CA	San Diego, CA	2023	200
laptop	TX	Dallas, TX	2023	400

**population by state and county**

state	county_state	population
CA	Los Angeles County_California	9848406
IL	Cook County_Illinois	5185812
TX	Harris County_Texas	4758579
AZ	Maricopa County_Arizona	4491987
CA	San Diego County_California	3282782
CA	Orange County_California	3164063
FL	Miami-Dade County_Florida	2685296
NY	Kings County_New York	2646306
TX	Dallas County_Texas	2603816
CA	Riverside County_California	2449909

**sales summary**

product_type	state	county_state	year	thousands of \$	population
laptop	CA	Los Angeles, CA	2023	200	9848406
laptop	IL	Cook, IL	2023	300	5185812
laptop	TX	Harris, TX	2023	500	4758579
laptop	NY	New York, NY	2023	600	2646306
laptop	AZ	Maricopa, AZ	2023	100	4491987
laptop	CA	Orange, CA	2023	600	3164063
laptop	CA	San Diego, CA	2023	200	3282782
laptop	TX	Dallas, TX	2023	400	2603816

If you have sales of a product type by state, county and year  
and

if you have prepared population by state and county

then only, comparison is meaningful assuming population is approximately same for number of years.

**One can explore the question: Is my laptop sales correlated with population at county level?**

Example matching (compare, correlate, etc.) of sales by state and year and GDP by county and year

**sales summary by state and year**

product_type	state	year	thousands of \$
laptop	CA	2023	2000
laptop	IL	2023	500
laptop	TX	2023	1200
laptop	NY	2023	1300
laptop	AZ	2023	1400

**gdp by state by year**

state	year	gdp in thousands of \$
CA	2023	7566594810
IL	2023	3964708990
TX	2023	3571306690
NY	2023	3434374420
AZ	2023	3123504170

**sales summary by state and year matched with GDP**

product_type	state	year	thousands of \$	gdp in thousands of \$
laptop	CA	2023	2000	7566594810
laptop	IL	2023	500	3964708990
laptop	TX	2023	1200	3571306690
laptop	NY	2023	1300	3434374420
laptop	AZ	2023	1400	3123504170

If you have sales of a product type by state and year  
and

you must prepare GDP (gross domestic product) by state and year  
then only, comparison is meaningful

**One can explore the question: Is my laptop sales correlated with GDP at state level?**

Example matching (compare, correlate, etc.) of sales by state and year and population by county

sales summary by state and year

product_type	state	year	thousands of \$
laptop	CA	2023	2000
laptop	IL	2023	500
laptop	TX	2023	1200
laptop	NY	2023	1300
laptop	AZ	2023	1400

population by state

state	population
CA	98484060
IL	51858120
TX	47585790
NY	26463060
AZ	44919870

sales summary by state and year matched with population

product_type	state	year	thousands of \$	population
laptop	CA	2023	2000	98484060
laptop	IL	2023	500	51858120
laptop	TX	2023	1200	47585790
laptop	NY	2023	1300	26463060
laptop	AZ	2023	1400	44919870

If you have sales of a product type by state and year  
and

you have prepared population by state and county

then only, comparison is meaningful assuming population is approximately same

**One can explore the question: Is my laptop sales correlated with population at state level?**

## Rank / Scale / Percentile

### GDP and Unemployment rank comparison

Absolute value of a state's GDP is useful. However, it is also necessary to know the rank of state among all other states. The rank is also known as percentile. In the following example, Texas is 64 rank or percentile in GDP in scale of 0 to 100. It is easier to think in terms of scale of 0 to 10, or 0 to 100. It is a simple scale for comparison vs GDP is \$4305454 million.

#### GDP by state

state name	GDP in MM	GDP rank
California	6686820	100
Texas	4305454	64
New York	3645422	54.02
Florida	2659164	39.12
Illinois	1781221	25.84
Pennsylvania	1623020	23.45
Ohio	1444748	20.76
Washington	1395673	20.02
Georgia	1392275	19.96
New Jersey	1349103	19.31

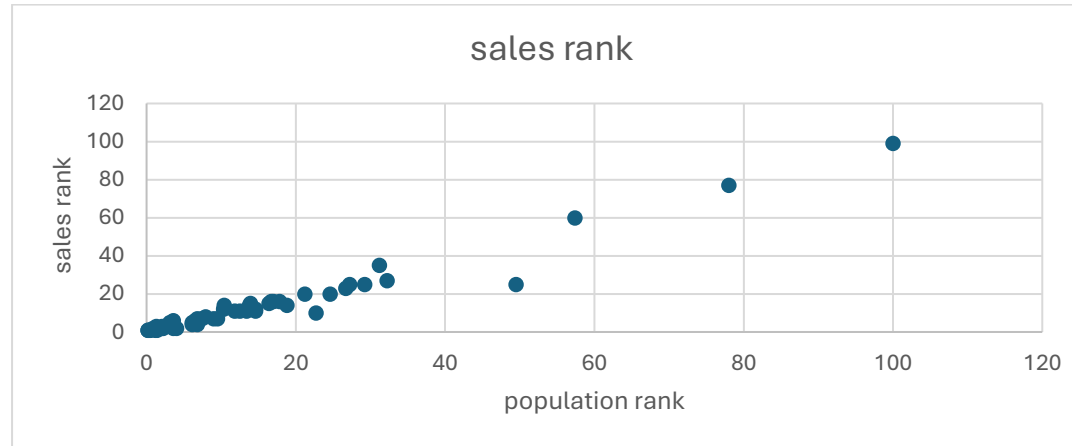
#### Unemployment by state

state name	unemployment rate	unemployment rank
District of Columbia	4.2	100
Nevada	4.1	100
California	4	90
Illinois	4	90
Kentucky	3.7	80
Washington	3.7	80
New Jersey	3.6	80
Alaska	3.4	70
Ohio	3.4	70
Rhode Island	3.4	70

Once you have GDP in rank and unemployment in rank for states, then, you can answer the following questions:  
Are the top 10 (rank/ percentile) states in GDP are also top 10 in unemployment.

## Sales and population rank comparison

state_name	population	rank	sales\$ in thousands	sales rank	difference in rank
Alabama	5108468	11.8	2200	11	0.8
Alaska	733406	0.4	200	1	-0.6
Arizona	7431344	17.8	3200	16	1.8
Arkansas	3067732	6.5	1200	6	0.5
California	38965193	100	19800	99	1
Colorado	5877610	13.8	2800	14	-0.2
Connecticut	3617176	7.9	1600	8	-0.1
Delaware	1031890	1.2	200	1	0.2
District of Col	678972	0.2	200	1	-0.8
Florida	22610726	57.4	12000	60	-2.6
Georgia	11029227	27.2	5000	25	2.2
Hawaii	1435138	2.2	400	2	0.2
Idaho	1964726	3.6	1200	6	-2.4
Illinois	12549689	31.2	7000	35	-3.8
Indiana	6862199	16.4	3000	15	1.4
Iowa	3207004	6.8	1400	7	-0.2
Kansas	2940546	6.1	1000	5	1.1
Kentucky	4526154	10.3	2400	12	-1.7
Louisiana	4573749	10.4	2800	14	-3.6
Maine	1395722	2.1	600	3	-0.9
Maryland	6180253	14.6	2400	12	2.6
Massachuset	7001399	16.7	3200	16	0.7
Michigan	10037261	24.6	4000	20	4.6
Minnesota	5737915	13.4	2200	11	2.4
Mississippi	2939690	6.1	800	4	2.1
Missouri	6196156	14.6	2200	11	3.6
Montana	1132812	1.4	200	1	0.4
Nebraska	1978379	3.6	400	2	1.6
Nevada	3194176	6.8	1000	5	1.8
New Hampsh	1402054	2.1	400	2	0.1
New Jersey	9290841	22.7	2000	10	12.7
New Mexico	2114371	4	400	2	2
New York	19571216	49.5	5000	25	24.5
North Carolin	10835491	26.7	4600	23	3.7
North Dakota	783926	0.5	200	1	-0.5
Ohio	11785935	29.2	5000	25	4.2
Oklahoma	4053824	9	1400	7	2
Oregon	4233358	9.5	1400	7	2.5
Pennsylvania	12961683	32.2	5400	27	5.2
Puerto Rico	3205691	6.8	800	4	2.8
Rhode Island	1095962	1.3	600	3	-1.7
South Carolin	5373555	12.5	2200	11	1.5
South Dakota	919318	0.9	400	2	-1.1
Tennessee	7126489	17	3200	16	1
Texas	30503301	78	15400	77	1
Utah	3417734	7.4	1400	7	0.4
Vermont	647464	0.2	200	1	-0.8
Virginia	8715698	21.2	4000	20	1.2
Washington	7812880	18.8	2800	14	4.8
West Virginia	1770071	3.1	1000	5	-1.9
Wisconsin	5910955	13.9	3000	15	-1.1

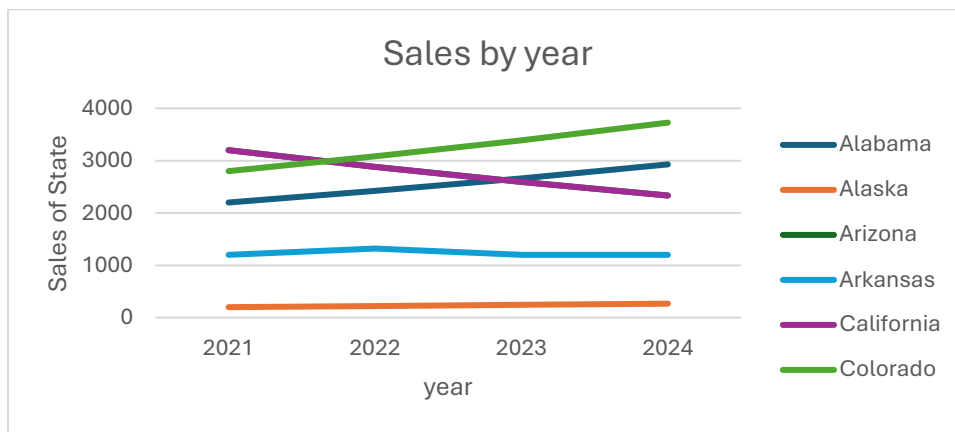


Use XY chart to see if two ranked variables are correlated. If two variables are correlated, then, the XY chart will be straight line at 45 degrees. Also, difference in large rank value means that state sales and state population are not correlated.

## Trend over time

This is classic analysis for any measurement. One can compare two more variables against time.

state_name	Alabama	Alaska	Arizona	Arkansas	California	Colorado
2021	2200	200	3200	1200	19800	2800
2022	2420	220	2880	1320	17820	3080
2023	2662	242	2592	1200	16038	3388
2024	2928.2	266.2	2332.8	1200	14434.2	3726.8



## Example Sales by state by year

If your sales depend on population, then, you need to find out why a region's sales are down when the population is up.

## Knowledge based on history vs hypothesis

### Historical knowledge scenario:

It is possible that a company knows that the sales are directly correlated to population when all other factors are in steady state. If the population is increasing but the sales are down means, the customers are not preferring your product or salespeople are not doing the job in that region.

### Hypothesis:

Assume or hypothesize that sales are correlated to population. Then, check the trend over time for population and sales, and then reject or do not reject the hypothesis.

To compare your sales or cost data, it needs to be summarized by time period (year, quarter, month) and region Country, state, county, zip code. A company's CRM or ERP data and external data must be transformed into proper format so that they can be correlated.

## Principal Component Analysis

Each of the regional (state, county, city, etc.) measurements such as GDP, population, unemployment, income level, education level can be considered as independent variables and company sales can be dependent variables. Principal Component analysis can show contribution by variables towards the sales. Use `AI_ML_intro-to-pca-t-sne-umap.py` for principal component analysis.

## ETL (extract, transform, and load) Steps for external data

### Task1 Use API to get data in Json, xml, or csv format

A government agency may have its website and API such as BEA (Bureau of Economic Analysis), BLS (Bureau of Labor Statistics). A government agency may be using common website and or API to publish their data. (<https://dev.socrata.com/> or <https://docs.ckan.org/en/2.8/api/>). Contact [milan@datajoin.net](mailto:milan@datajoin.net) for python code.

## Task2 Convert Json or xml format to csv format

Some API provide data in Json or xml format and not in csv format. Thus, it is necessary to convert to csv file. CSV is more suitable for data operation in batch using python. If data is received in CSV format, then, this task is not necessary. Contact [milan@datajoin.net](mailto:milan@datajoin.net) for python code.

## Task3 Copy and or add attributes

Necessity to copy an attribute (field or column). The original column is available just in case the data operation did not work as expected. It may also be necessary to add default value columns such as Country or Units of measurement. Contact [milan@datajoin.net](mailto:milan@datajoin.net) for python code.

## Task4 Transform values in attributes

There are at least three types of transformations:

Format change for date

Change label using lookup values

Split column in two or more if county and city are in the same column. For example, Newark NJ. Contact [milan@datajoin.net](mailto:milan@datajoin.net) for python code.

## Task5 Upload base table

It is necessary to have this base table available to create two or more summary tables depending upon need. For example, one summary table can be for state and another summary table can be for State and County combination.

Use `datajoin_generic_task5createtable.py` to create database table directly from csv file.

## Task6 Create summary table with rank/ percentile / scaled column

Summary tables are necessary to join them for correlation. Use `datajoin_generic_task6sum_scale_v3.py` to create summary table.

There is multiple python program for each of the above tasks depending upon API and output of a step.

Refer Agents (workflow) to carry out the above tasks using python (<https://github.com/milan888-design/ai-agents-workflow>)



## Python code to new table with sum and rank

[datajoin\\_generic\\_task6sum\\_scale\\_v3.py](#)

```
#for sales table
#arg1='sales_by_county_state_year'
#arg2='sales_by_county_state_year_summary'
#arg3='sales_in_thousand'
#arg4='select state,year,sum(cast(sales_in_thousand as numeric(10,1))) as sales_in_thousand from sales_by_county_state_year group by
state,year'
#arg5='postgresql://postgres:pass@localhost:5432/test_correlate'
#arg6='postgresql://postgres:pass@localhost:5432/test_correlate'

#for gdp table
arg1='gdp_by_county_state_year'
arg2='gdp_by_county_state_year_summary'
arg3='gdp_in_thousand'
arg4='select state,year,sum(cast(gdp_in_thousand as numeric(10,1))) as gdp_in_thousand from gdp_by_county_state_year group by
state,year'
arg5='postgresql://postgres:pass@localhost:5432/test_correlate'
arg6='postgresql://postgres:pass@localhost:5432/test_correlate'
```

## Task 7 Join table to study correlation

Use sum scale query for test\_correlate db.sql to join sales table and gdp table.

--join sales and gdp tables

```
SELECT
salessum.state
,salessum.year
,salessum.sales_in_thousand
,salessum.sales_in_thousand_scaled
,gdpsum.gdp_in_thousand
,gdpsum.gdp_in_thousand_scaled
```

```

FROM sales_by_county_state_year_summary salessum
,gdp_by_county_state_year_summary gdpsum
WHERE
salessum.state=gdpsum.state
and salessum.year=gdpsum.year

```

Result from the above SQL.

state	year	sales_in_thousand	sales_in_thousand_scaled	gdp_in_thousand	gdp_in_thousand_scaled
AZ	2023	100	0	312350417	2.67
CA	2023	1250	100	1577667989	100
IL	2023	300	17.39	396470899	9.14
NY	2023	600	43.48	343437442	5.06
TX	2023	900	69.57	564664441	22.08
WA	2023	600	43.48	277642267	0