CIS 4560 Term Project Tutorial

Authors: Alexander Castellanos, Armando Perez, Eric Peralta Erick Robles, Steve Hwang, Vincent Cheung

Instructor: Jongwook Woo

Date: 05/15/2023

Lab Tutorial

Pain Pills Data Analysis in Hive

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Connect to Hadoop Cluster remotely
- Load Pain Pills Data into Hadoop Clusters
- Create a staging table in Beeline
- Check and verify the data
- Load the clean data into PainPills table
- Check and verify the data again
- Generating Top 10 Reports
- Import Hadoop File to MS Power BI Desktop

Platform Spec

- Oracle Linux Server
- CPU Speed: 1995 MHz
- # of CPU cores: 8
- # of nodes: 3
- Total Memory Size: 58 GB

Step 1: Connect to Hadoop Cluster remotely

You need to remote to your Hadoop Clusters using the ***ssh*** command from the Git Bash terminal as follows:

```
$ ssh username@ipaddress
```

Step 2: Load Pain Pills Data into Hadoop Clusters

You can download the data files using the ***wget*** command from the terminal as follows:

```
$ wget https://github.com/vcheung621/cis4560/raw/main/arcos-southern-ca-
itemized.zip
```

Once you download the data file, please proceed with the commands below to create a temporary directory (arcos) and unzip the zip file into the directory.

```
$ mkdir arcos

$ mv arcos-southern-ca-itemized.zip arcos

$ cd arcos/

$ unzip arcos-southern-ca-itemized.zip
```

After you unzip all the CSV files, the below commands will create an HDFS directory (PainPillsFiles) and put all the CVS files into it.

```
$ hdfs dfs -mkdir PainPillsFiles

$ hdfs dfs -put *.csv PainPillsFiles

$ hdfs dfs -ls PainPillsFiles
```

Step 3: Create a staging table in Beeline

The following Hive statement creates an external staging table (painpills_stage). External tables preserve the data in the original file format while allowing Hive to perform queries against the data within the file.

NOTE: You have to replace the user name **<username>** to your username.

```
USE your_databasename;

--drop the table painpills_stage
DROP TABLE IF EXISTS painpills_stage;

--create the painpills staging table on comma-separated data
CREATE EXTERNAL TABLE IF NOT EXISTS painpills_stage(
REPORTER_DEA_NO STRING,
REPORTER_BUS_ACT STRING,
REPORTER_NAME STRING,
REPORTER_ADDL_CO_INFO STRING,
REPORTER_ADDRESS1 STRING,
REPORTER_ADDRESS2 STRING,
REPORTER_CITY STRING,
REPORTER_STATE STRING,
REPORTER_ZIP BIGINT,
REPORTER_COUNTY STRING,
BUYER_DEA_NO STRING,
BUYER_BUS_ACT STRING,
BUYER_NAME STRING,
BUYER_ADDL_CO_INFO STRING,
BUYER_ADDRESS1 STRING,
BUYER_ADDRESS2 STRING,
BUYER_CITY STRING,
BUYER_STATE STRING,
BUYER_ZIP BIGINT,
BUYER_COUNTY STRING,
TRANSACTION_CODE STRING,
DRUG_CODE BIGINT,
NDC_NO STRING,
DRUG_NAME STRING,
QUANTITY BIGINT,
UNIT STRING,
ACTION_INDICATOR STRING,
ORDER_FORM_NO STRING,
CORRECTION_NO STRING,
STRENGTH STRING,
TRANSACTION_DATE STRING,
CALC_BASE_WT_IN_GM FLOAT,
DOSAGE_UNIT BIGINT,
TRANSACTION_ID BIGINT,
Product_Name STRING,
Ingredient_Name STRING,
Measure STRING,
MME_Conversion_Factor FLOAT,
Combined_Labeler_Name STRING,
Revised_Company_Name STRING,
Reporter_family STRING,
dos_str STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
   "separatorChar" = "\,",
   "quoteChar"     = "\"",
   "escapeChar"    = "\\"
)
STORED AS TEXTFILE LOCATION '/user/<username>/PainPillsFiles'
TBLPROPERTIES ('skip.header.line.count'='1');
```

Step 4: Check and verify the data

After constructing the table, we will examine and validate the data. The queries provided below will determine the total data count. It is essential to confirm the absence of any corrupt data. One method to validate data involves verifying if the column data aligns with the CSV files, selecting the DRUG_NAME column as an example because we know it contains only two unique values in the source CSV file. Additionally, we can inspect the zip code column (buyer_zip) for any letters and examine the manufacturer column (combined_labeler_name) for null values.

```
select count(*) from painpills_stage;
+-----------+
|    _c0    |
+-----------+
| 9571662   |
+-----------+
```

```
select count(*) from painpills_stage where drug_name not in
('HYDROCODONE','OXYCODONE');
+-------+
| _c0   |
+-------+
| 507   |
+-------+
```

```
select count(*) from painpills_stage where drug_name in
('HYDROCODONE','OXYCODONE');
+-----------+
|    _c0    |
+----------+
| 9571155  |
+----------+
```

```
SELECT COUNT(*) AS count_letters
FROM painpills_stage
WHERE LENGTH(regexp_extract(buyer_zip, '[a-zA-Z]', 0)) > 0;
+----------------+
| count_letters  |
+----------------+
| 507            |
+----------------+
```

```
SELECT COUNT(*) AS null_count
FROM painpills_stage
WHERE Combined_Labeler_Name = 'null';
+-------------+
| null_count  |
+-------------+
| 11430       |
+-------------+
```

Step 5: Load the clean data into PainPills table

Having identified 11,937 (507+11,430) instances of corrupt data through the previous query, we will proceed to clean this data. Additionally, the original dataset contains numerous columns that are not required for our analysis. We will selectively choose the relevant columns. The following statement will generate a new table called PAINPILLS, consisting of clean data and the essential columns.

```
--use beeline
DROP TABLE IF EXISTS painpills;

--create the painpills table on comma-separated data
CREATE TABLE painpills
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
   "separatorChar" = "\,",
   "quoteChar"     = "\"",
   "escapeChar"    = "\\"
)
STORED AS TEXTFILE LOCATION '/user/vcheung4/PainPills'
AS
SELECT buyer_dea_no AS dea_no,
buyer_name as pharmacy,
buyer_addl_co_info as addl_co_info,
buyer_address1 as address1,
buyer_address2 as address2,
buyer_city as city,
buyer_state as state,
buyer_zip as zip,
buyer_county as county,
drug_name,
quantity,
TO_DATE(from_unixtime(unix_timestamp(transaction_date,'MMddyyyy'),'yyyy-MM-
dd')) AS transaction_date,
calc_base_wt_in_gm,
dosage_unit as number_of_pills,
transaction_id,
product_name,
ingredient_name,
combined_labeler_name as manufacturer,
revised_company_name as distributor,
dos_str
FROM painpills_stage where drug_name in ('OXYCODONE','HYDROCODONE') AND
Combined_Labeler_Name <> 'null';
```

Step 6: Check and verify the data again

Now check and verify the data again to see any dirty data.

```
select count(*) from painpills;
+----------+
|   _c0    |
+----------+
| 9559725  |
+----------+
```

```
select count(*) from painpills where drug_name not in
('HYDROCODONE','OXYCODONE');
+------+
| _c0  |
+------+
| 0    |
+------+
```

```
select count(*) from painpills where drug_name in
('HYDROCODONE','OXYCODONE');
+----------+
|   _c0    |
+----------+
| 9559725  |
+----------+
SELECT COUNT(*) AS count_letters
FROM painpills
WHERE LENGTH(regexp_extract(zip, '[a-zA-Z]', 0)) > 0;
+-----------------+
| count_letters  |
+-----------------+
| 0               |
+-----------------+
```

```
SELECT COUNT(*) AS null_count
FROM painpills
WHERE manufacturer = 'null';
+-------------+
| null_count  |
+-------------+
| 0           |
+-------------+
```

Above are the expected results. The DRUG_NAME column only contains two distinct values. The zip column contains no letter characters. The manufacturer column contains no null values. The new total record count is 9559725 (9571662 – 11937).

Step 7: Generating Top 10 Reports

Now you can create a top 10 distributors' report by executing the following:

```
select distributor, format_number(sum(number_of_pills),0) AS total_pills,
round(sum(number_of_pills)/(select sum(number_of_pills) from painpills) *
100,1) as percentage from painpills group by distributor order by percentage
desc limit 10;

+---------------------------+--------------+-------------+
|        distributor        | total_pills  | percentage  |
+---------------------------+--------------+-------------+
| AmerisourceBergen Drug    | 988,807,325  | 19.5        |
| McKesson Corporation      | 875,541,900  | 17.3        |
| CVS                       | 583,582,700  | 11.5        |
| Walgreen Co               | 468,470,760  | 9.2         |
| Cardinal Health           | 430,580,165  | 8.5         |
| Thrifty Payless Inc       | 363,882,100  | 7.2         |
| Kaiser Permanente         | 290,377,930  | 5.7         |
| H. D. Smith               | 227,268,410  | 4.5         |
| Wal-Mart                  | 149,561,300  | 3.0         |
| Valley Wholesale Drug Co  | 99,650,110   | 2.0         |
+---------------------------+--------------+-------------+
```

You can create a top 10 manufacturers' report by executing the following:

```
select manufacturer, format_number(sum(number_of_pills),0) AS total_pills,
round(sum(number_of_pills)/(select sum(number_of_pills) from painpills) *
100,1) as percentage from painpills group by manufacturer order by percentage
desc limit 10;

+------------------------------+----------------+--------------+
|         manufacturer         |  total_pills   | percentage   |
+------------------------------+----------------+--------------+
| SpecGx LLC                   | 1,687,218,718  | 33.3         |
| Actavis Pharma, Inc.         | 1,573,661,563  | 31.1         |
| Par Pharmaceutical           | 978,929,948    | 19.3         |
| Amneal Pharmaceuticals LLC   | 179,715,626    | 3.5          |
| Purdue Pharma LP             | 151,651,496    | 3.0          |
| Kaiser Foundation Hospitals  | 128,272,830    | 2.5          |
| AbbVie Inc.                  | 42,803,604     | 0.8          |
| KVK-Tech, Inc.               | 42,003,700     | 0.8          |
| Dispensing Solutions Inc.    | 24,254,380     | 0.5          |
| Bryant Ranch Prepack         | 26,838,261     | 0.5          |
+------------------------------+----------------+--------------+
```

You can create a top 10 pharmacies' report by executing the following:

```
select pharmacy, format_number(sum(number_of_pills),0) AS total_pills,
round(sum(number_of_pills)/(select sum(number_of_pills) from painpills) *
100,1) as percentage from painpills group by pharmacy order by percentage
desc limit 10;

+---------------------------------------+--------------+--------------+
|               pharmacy                | total_pills  | percentage   |
+---------------------------------------+--------------+--------------+
| GARFIELD BEACH CVS, L.L.C.            | 805,190,641  | 15.9         |
| WALGREEN CO.                          | 508,510,910  | 10.0         |
| THRIFTY PAYLESS INC.                  | 500,215,840  | 9.9          |
| KAISER FOUNDATION HLTH PLN            | 222,259,950  | 4.4          |
| LONGS DRUG STORES CALIFORNIA, L.L.C.  | 166,237,150  | 3.3          |
| THE VONS COMPANIES INC                | 119,895,670  | 2.4          |
| COSTCO WHOLESALE CORPORATION          | 120,475,210  | 2.4          |
| OPTUMRX                               | 92,022,350   | 1.8          |
| NEW ALBERTSON'S, INC.                 | 84,835,050   | 1.7          |
| TARGET STORES A DIV.OF TARGET CORP.   | 64,813,340   | 1.3          |
+---------------------------------------+--------------+--------------+
```

You can create a top 10 products' report by executing the following:

```
select product_name, format_number(sum(number_of_pills),0) AS total_pills,
round(sum(number_of_pills)/(select sum(number_of_pills) from painpills) *
100,1) as percentage from painpills group by product_name order by percentage
desc limit 10;

+---------------------------------------+--------------+-------------+
|             product_name              | total_pills  | percentage  |
+---------------------------------------+--------------+-------------+
| HYDROCODONE BIT/ACETAMINOPHEN 5MG/50  | 572,361,010  | 11.3        |
| HYDROCODONE BIT. 10MG/ACETAMINOPHEN   | 429,918,585  | 8.5         |
| HYDROCODONE BIT 5MG/ACETAMINOPHEN 50  | 400,853,017  | 7.9         |
| HYDROCODONE.BIT./ACET.,10MG & 325MG/  | 316,910,150  | 6.3         |
| HYDROCODONE BITARTRATE 7.5MG/ACETAMI  | 302,069,604  | 6.0         |
| HYDROCODONE BIT/ACETA 10MG/325MG USP  | 266,708,730  | 5.3         |
| HYDROCODO.BIT/APAP 7.5MG/750MG USP T  | 248,234,042  | 4.9         |
| HYDROCODONE.BIT. & ACETA  5MG & 500M  | 220,160,616  | 4.3         |
| OXYCODONE HCL/ACETAMINOPHEN 5MG/325M  | 176,668,000  | 3.5         |
| HYDROCODONE BIT./ACETA 10MG/325MG TA  | 124,347,687  | 2.5         |
+---------------------------------------+--------------+-------------+
```

Since we have 11 Hadoop data files, we must merge them into one. Execute the below command to combine and output into one text file.

```
hdfs dfs -getmerge -nl PainPills/* output.csv
```

On your PC with git bash, you can remotely download the output file "output.csv" to your PC to visualize it using MS PowerBI.

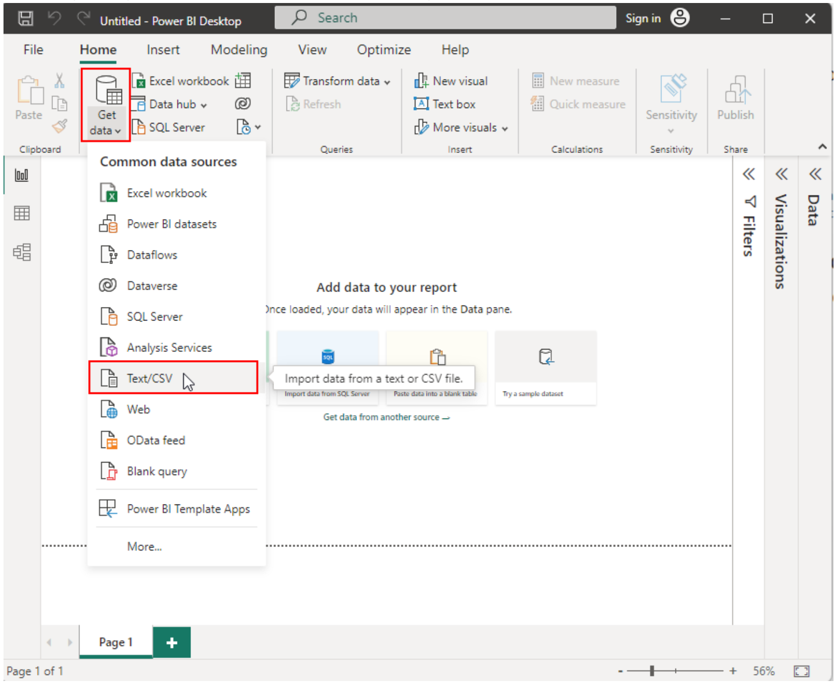
Note: You must replace the user name <username> with your username. Also, you may need to download the MS PowerBI Desktop version.

```
scp <username>@xxx.xxx.xxx.xxx:/home/<username>/output.csv .
```

Step 8: Import Hadoop File to MS Power BI Desktop

Open your MS Power BI Desktop at your local computer.

1. Open your MS Power BI Desktop and click on "Get data" and then click on "Text/CSV".

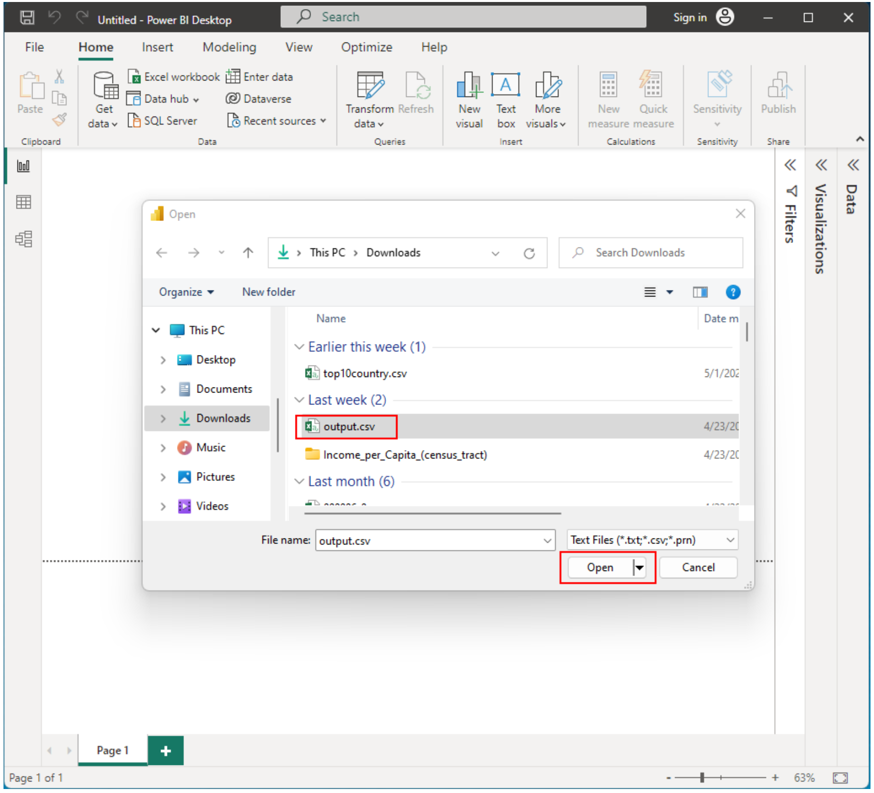


Add data to your report

)nce loaded, your data will appear in the Data pane.

Import data from a text or CSV file.

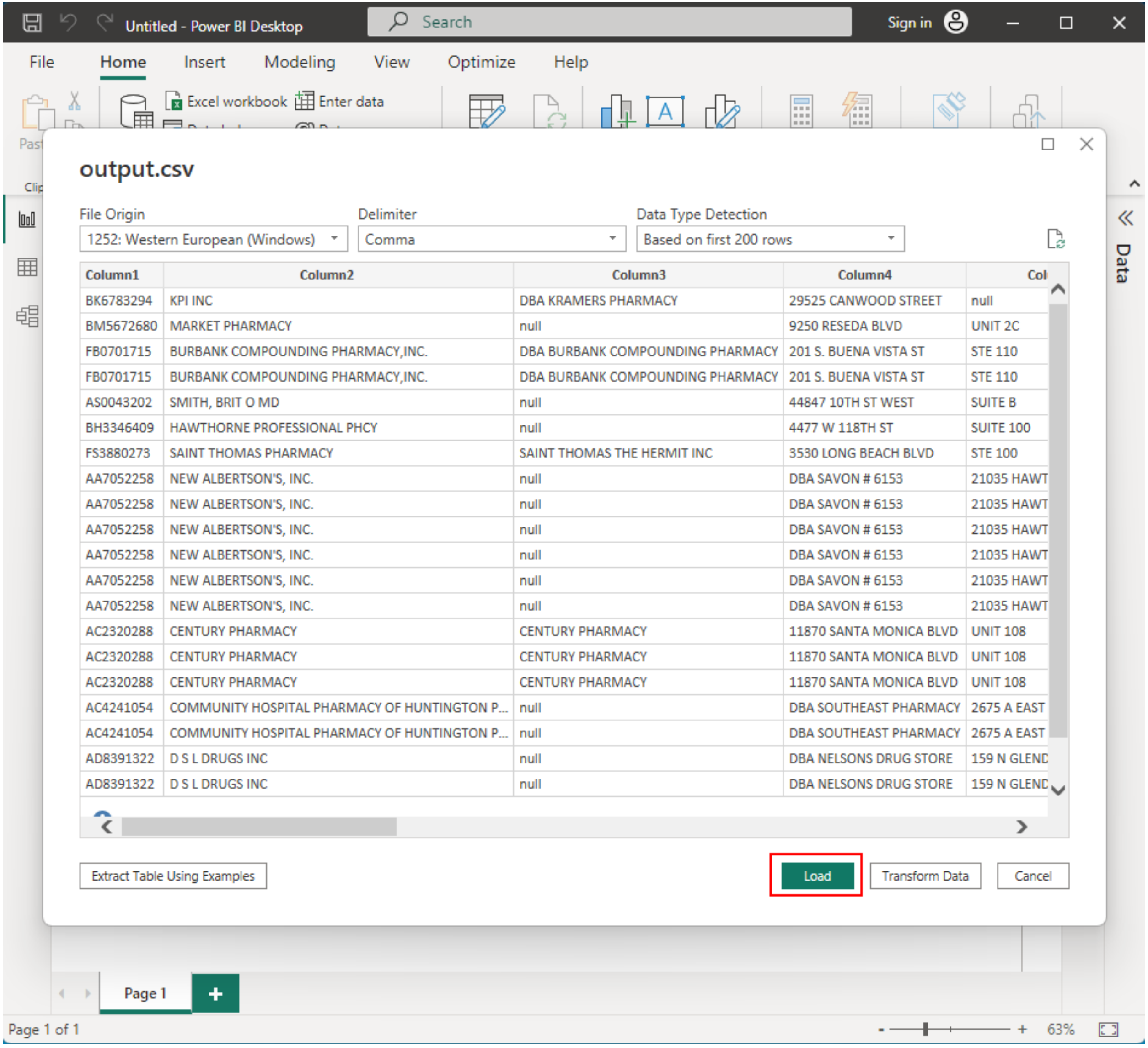
2. Browse the output file (output.csv) and click "Open".



File name: output.csv

Text Files (*.txt;*.csv;*.prn)

3. Click "Load". This process may take a while for the import.



output.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

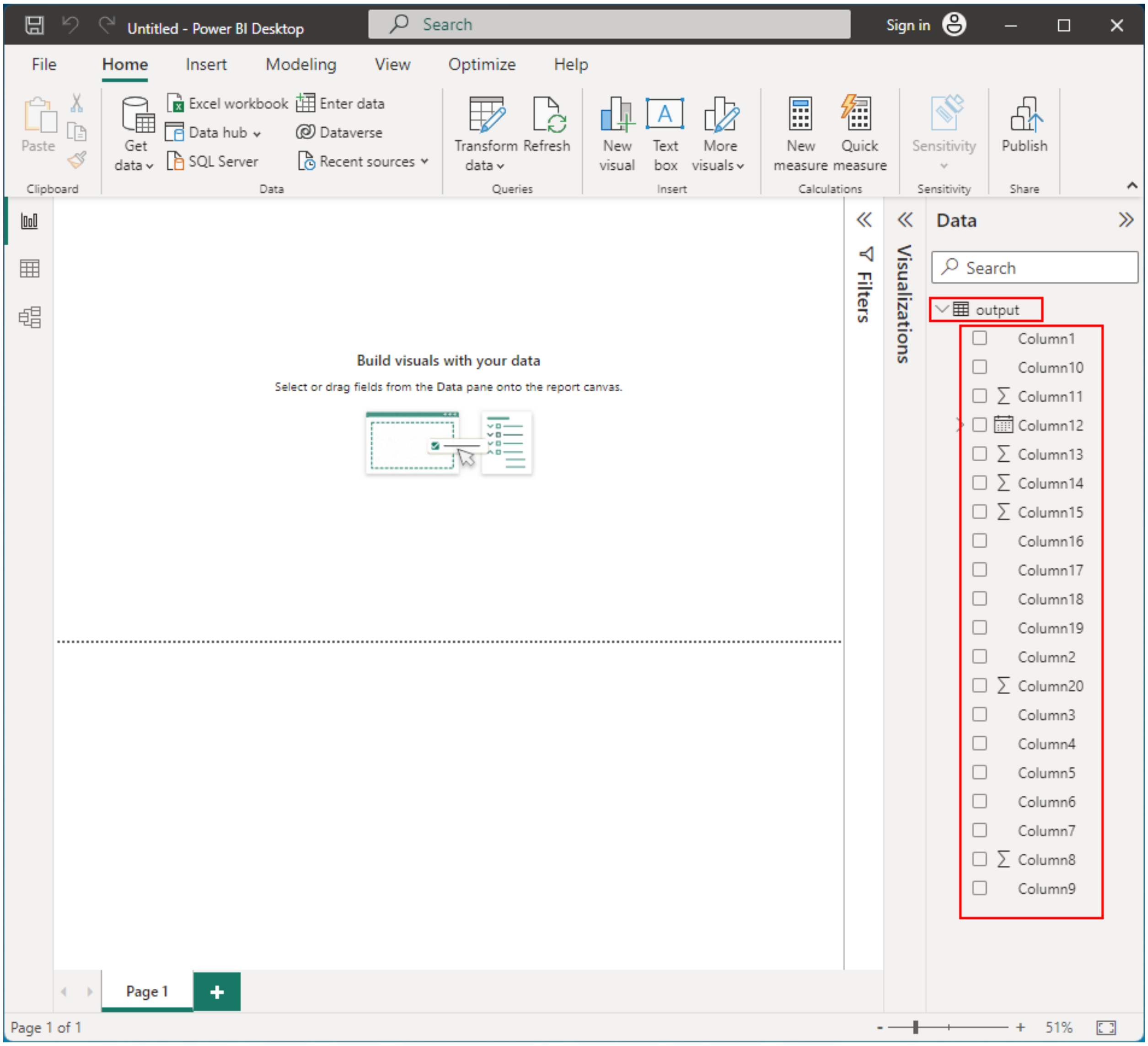
Data Type Detection: Based on first 200 rows

Column1	Column2	Column3	Column4	Col
BK6783294	KPI INC	DBA KRAMERS PHARMACY	29525 CANWOOD STREET	null
BM5672680	MARKET PHARMACY	null	9250 RESEDA BLVD	UNIT 2C
FB0701715	BURBANK COMPOUNDING PHARMACY,INC.	DBA BURBANK COMPOUNDING PHARMACY	201 S. BUENA VISTA ST	STE 110
FB0701715	BURBANK COMPOUNDING PHARMACY,INC.	DBA BURBANK COMPOUNDING PHARMACY	201 S. BUENA VISTA ST	STE 110
AS0043202	SMITH, BRIT O MD	null	44847 10TH ST WEST	SUITE B
BH3346409	HAWTHORNE PROFESSIONAL PHCY	null	4477 W 118TH ST	SUITE 100
FS3880273	SAINT THOMAS PHARMACY	SAINT THOMAS THE HERMIT INC	3530 LONG BEACH BLVD	STE 100
AA7052258	NEW ALBERTSON'S, INC.	null	DBA SAVON # 6153	21035 HAWT
AA7052258	NEW ALBERTSON'S, INC.	null	DBA SAVON # 6153	21035 HAWT
AA7052258	NEW ALBERTSON'S, INC.	null	DBA SAVON # 6153	21035 HAWT
AA7052258	NEW ALBERTSON'S, INC.	null	DBA SAVON # 6153	21035 HAWT
AA7052258	NEW ALBERTSON'S, INC.	null	DBA SAVON # 6153	21035 HAWT
AA7052258	NEW ALBERTSON'S, INC.	null	DBA SAVON # 6153	21035 HAWT
AC2320288	CENTURY PHARMACY	CENTURY PHARMACY	11870 SANTA MONICA BLVD	UNIT 108
AC2320288	CENTURY PHARMACY	CENTURY PHARMACY	11870 SANTA MONICA BLVD	UNIT 108
AC2320288	CENTURY PHARMACY	CENTURY PHARMACY	11870 SANTA MONICA BLVD	UNIT 108
AC4241054	COMMUNITY HOSPITAL PHARMACY OF HUNTINGTON P...	null	DBA SOUTHEAST PHARMACY	2675 A EAST
AC4241054	COMMUNITY HOSPITAL PHARMACY OF HUNTINGTON P...	null	DBA SOUTHEAST PHARMACY	2675 A EAST
AD8391322	D S L DRUGS INC	null	DBA NELSONS DRUG STORE	159 N GLEND
AD8391322	D S L DRUGS INC	null	DBA NELSONS DRUG STORE	159 N GLEND

4. Expand the output under the "Data" section and right click to rename the column names as follow:

- Column1 as dea_no,
- Column2 as pharmacy,
- Column3 as addl_co_info,
- Column4 as address1,
- Column5 as address2,
- Column6 as city,
- Column7 as state,
- Column8 as zip,
- Column9 as county,
- Column10 as drug_name,
- Column11 as quantity,
- Column12 as transaction_date,

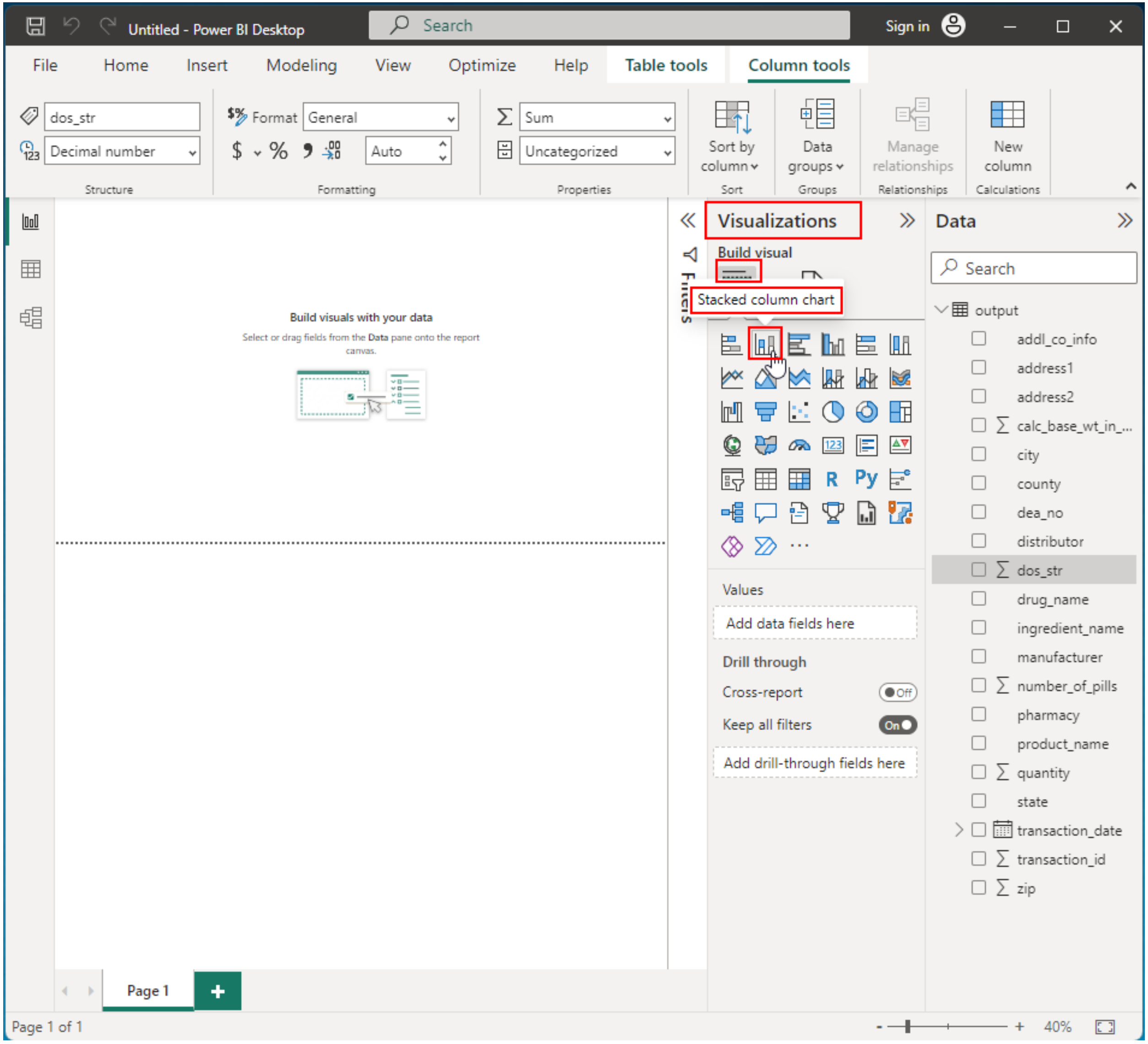
- Column13 as calc_base_wt_in_gm,
- Column14 as number_of_pills,
- Column15 as transaction_id,
- Column16 as product_name,
- Column17 as ingredient_name,
- Column18 as manufacturer,
- Column19 as distributor,
- Column20 as dos_str



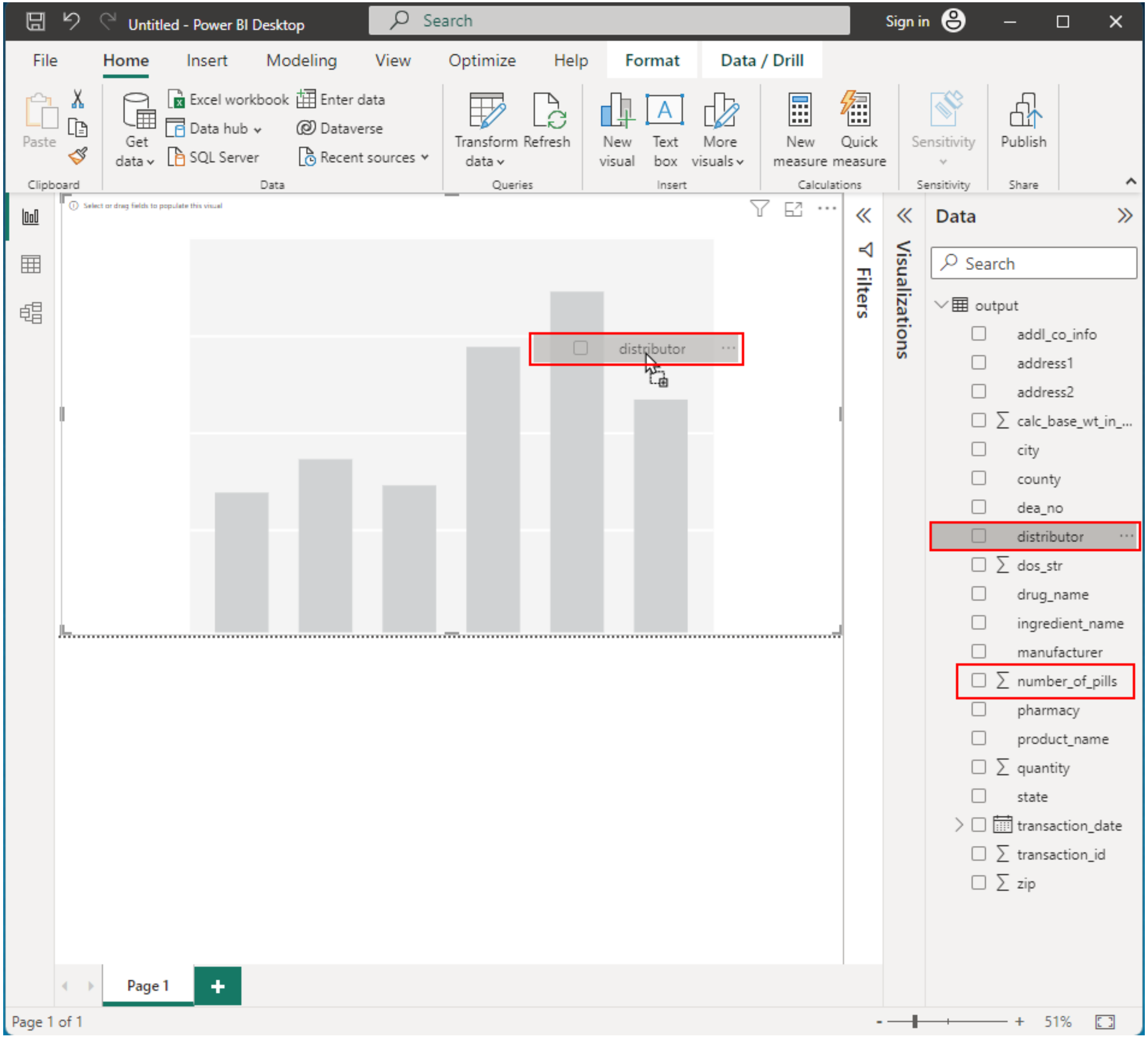
Build visuals with your data

Select or drag fields from the Data pane onto the report canvas.

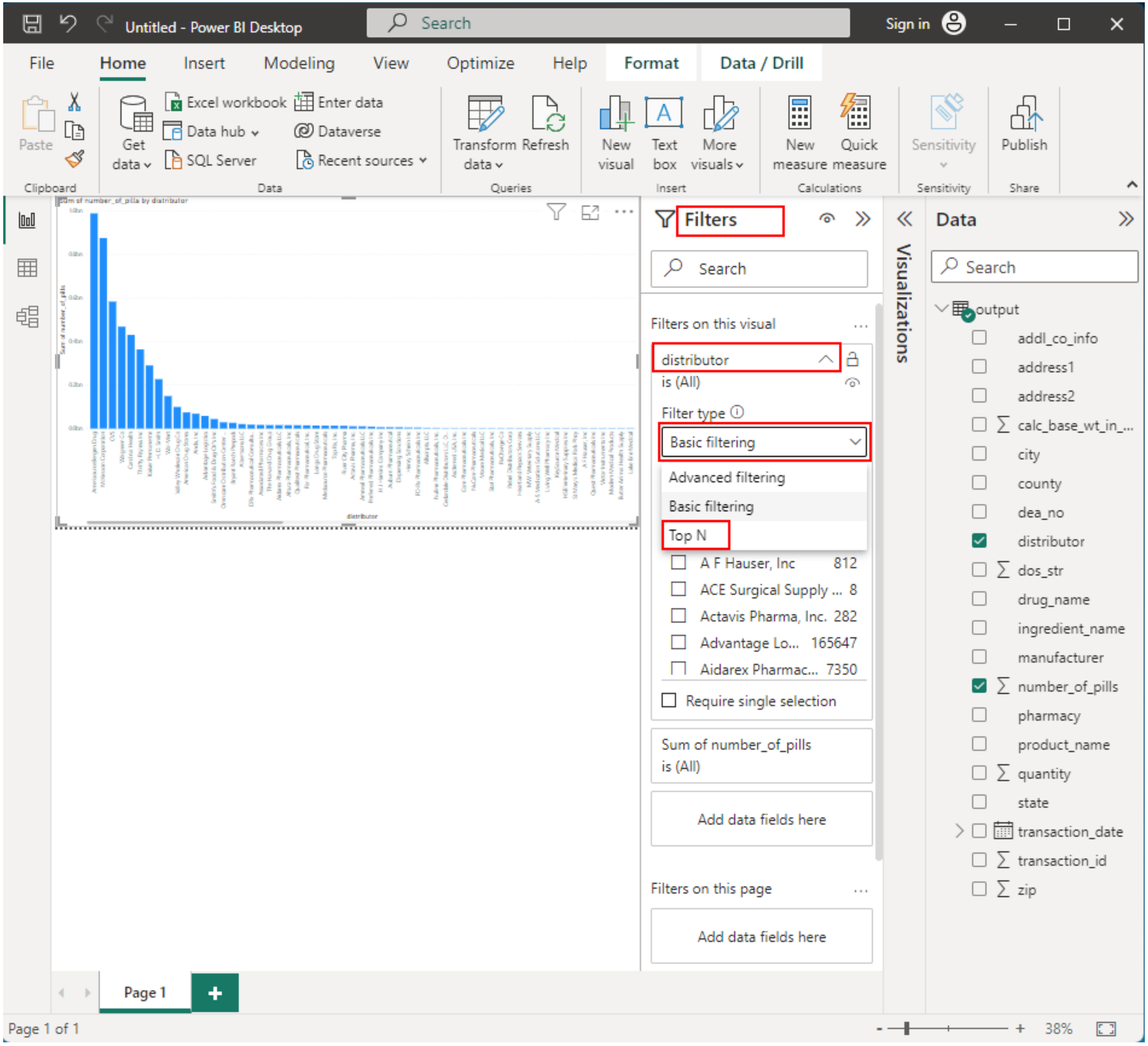
5. Click the "Stack column chart" under "Visualizations" -> "Add data to your visual"



6. Drag the "distributor" and the "number_of_pills" fields to the "Stack column chart"



7. Click on "Top N" under "Filters" -> "distributor" -> "Basic filtering"



Filters on this visual

distributor
is (All)

Filter type ⓘ
Basic filtering

- Advanced filtering
- Basic filtering
- Top N

□ A F Hauser, Inc	812
□ ACE Surgical Supply ...	8
□ Actavis Pharma, Inc.	282
□ Advantage Lo...	165647
□ Aidarex Pharmac...	7350

□ Require single selection

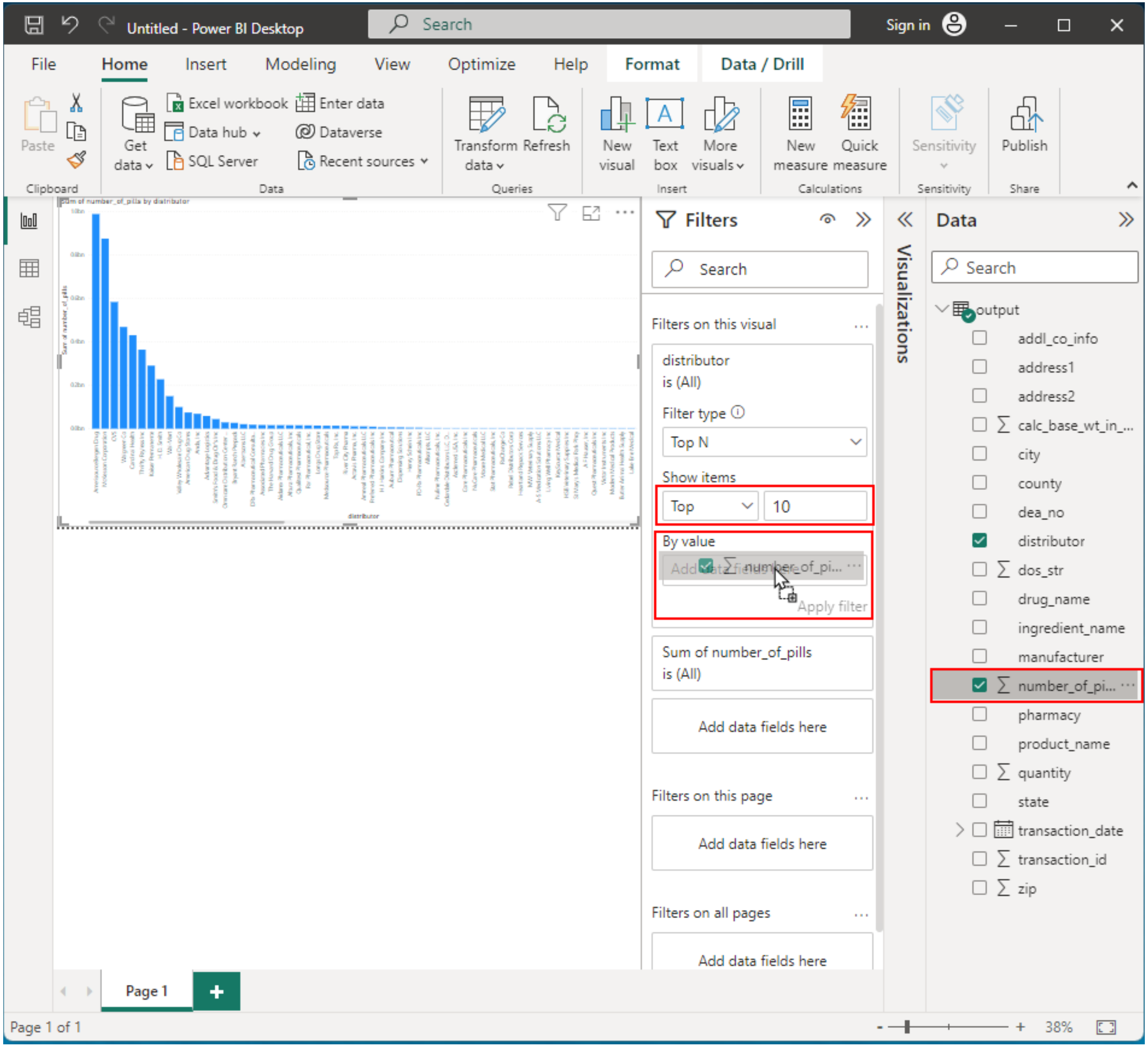
Sum of number_of_pills
is (All)

Add data fields here

Filters on this page

Add data fields here

8. Enter "10" next to the "Top" dropdown and drag the "number_of_pills" field to the "By value" box.



distributor
is (All)

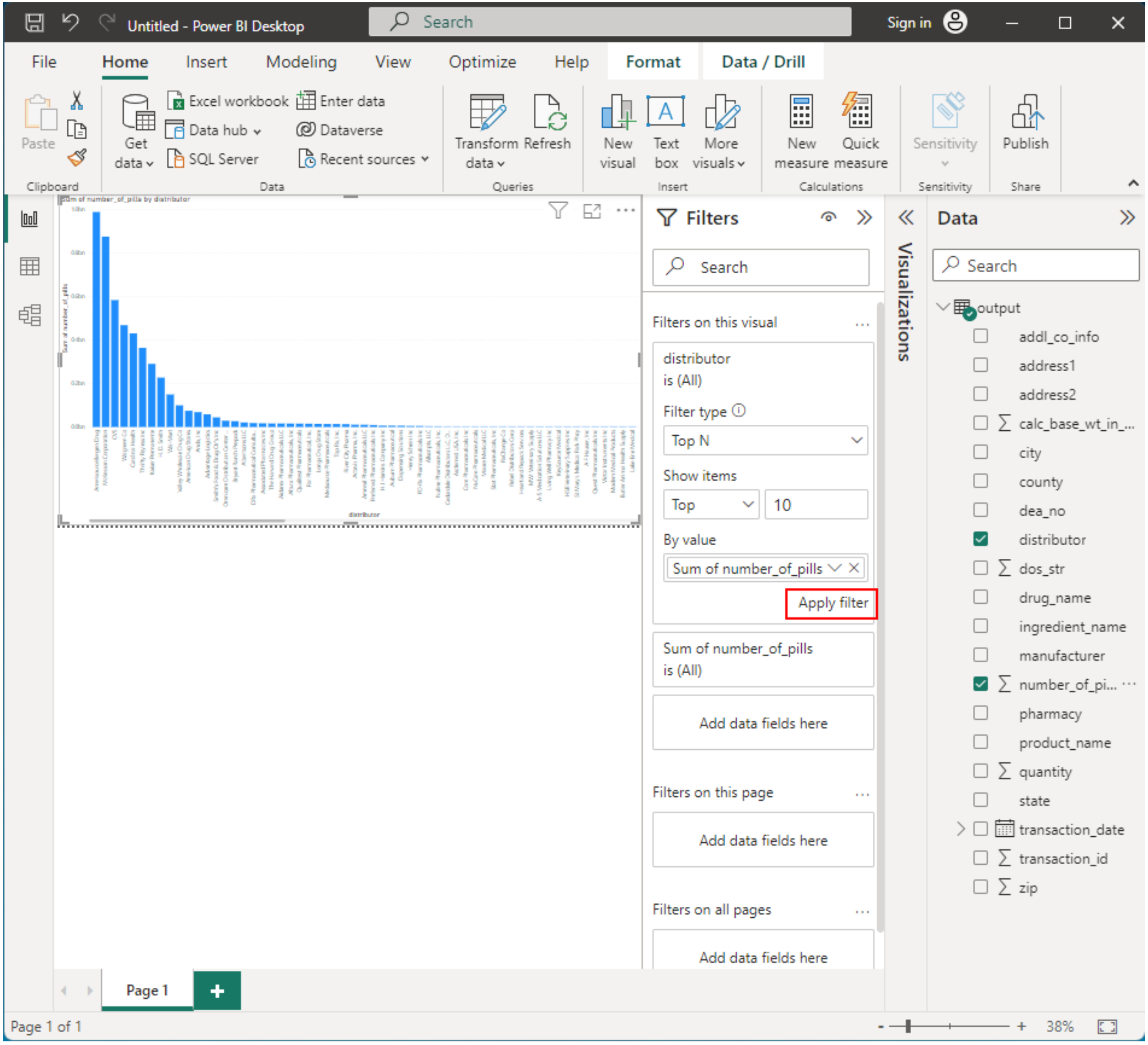
Filter type: Top N

Show items: Top 10

By value: number_of_pi...

Apply filter

9. Click the "Apply filter" button



Filters

Search

Filters on this visual

distributor
is (All)

Filter type: Top N

Show items: Top 10

By value: Sum of number_of_pills

Apply filter

Sum of number_of_pills
is (All)

Add data fields here

Filters on this page

Add data fields here

Filters on all pages

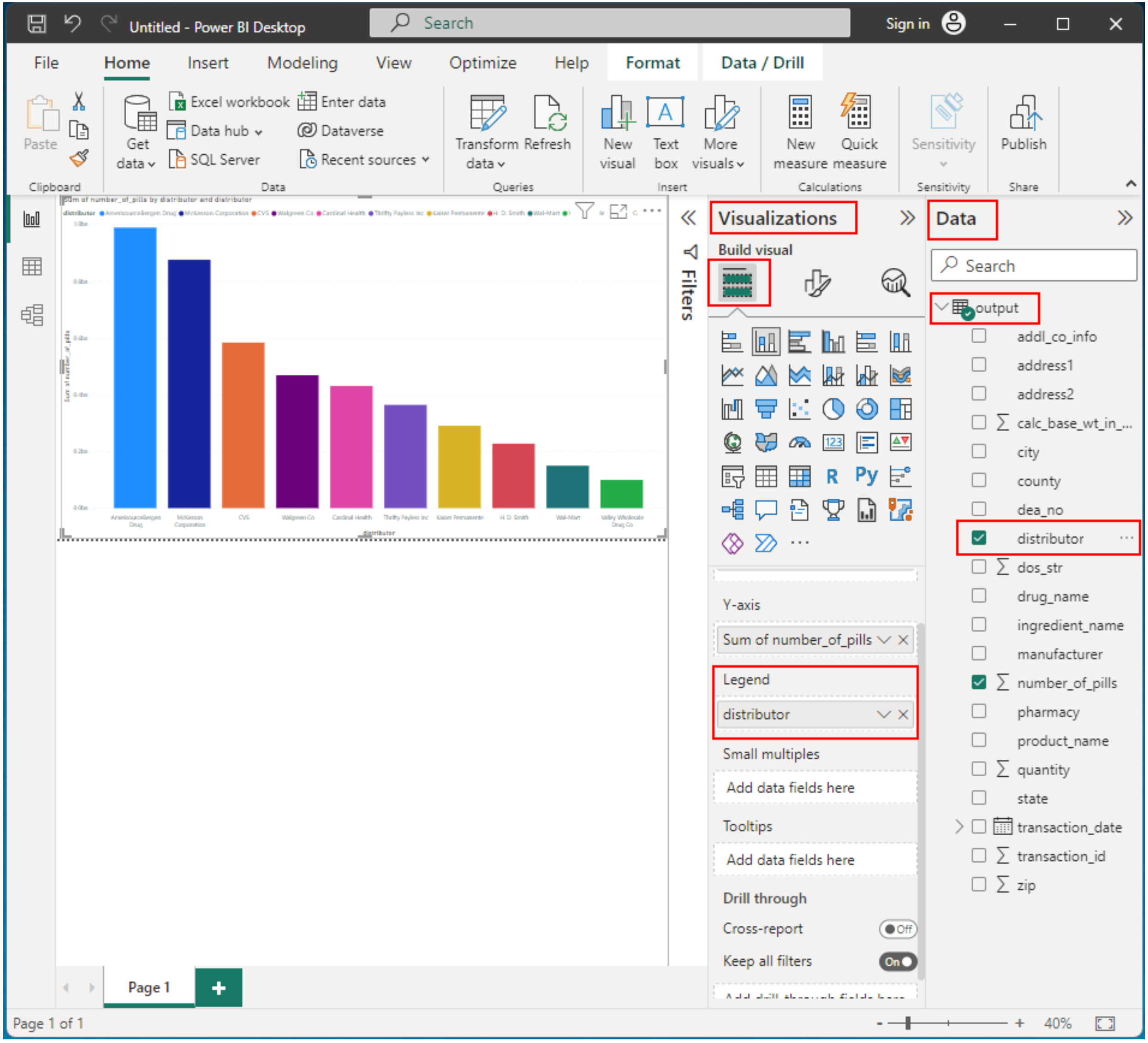
Add data fields here

Data

Search

output
- □ addl_co_info
- □ address1
- □ address2
- □ ∑ calc_base_wt_in_...
- □ city
- □ county
- □ dea_no
- ☑ distributor
- □ ∑ dos_str
- □ drug_name
- □ ingredient_name
- □ manufacturer
- ☑ ∑ number_of_pi...
- □ pharmacy
- □ product_name
- □ ∑ quantity
- □ state
- □ transaction_date
- □ ∑ transaction_id
- □ ∑ zip

10. Drag the "distributor" field to the "Legend" under "Visualizations" -> "Add data to your visual"



output
- ☐ addl_co_info
- ☐ address1
- ☐ address2
- ☐ ∑ calc_base_wt_in_...
- ☐ city
- ☐ county
- ☐ dea_no
- ☑ distributor
- ☐ ∑ dos_str
- ☐ drug_name
- ☐ ingredient_name
- ☐ manufacturer
- ☑ ∑ number_of_pills
- ☐ pharmacy
- ☐ product_name
- ☐ ∑ quantity
- ☐ state
- ☐ transaction_date
- ☐ ∑ transaction_id
- ☐ ∑ zip

Y-axis: Sum of number_of_pills

Legend: distributor

Small multiples: Add data fields here

Tooltips: Add data fields here

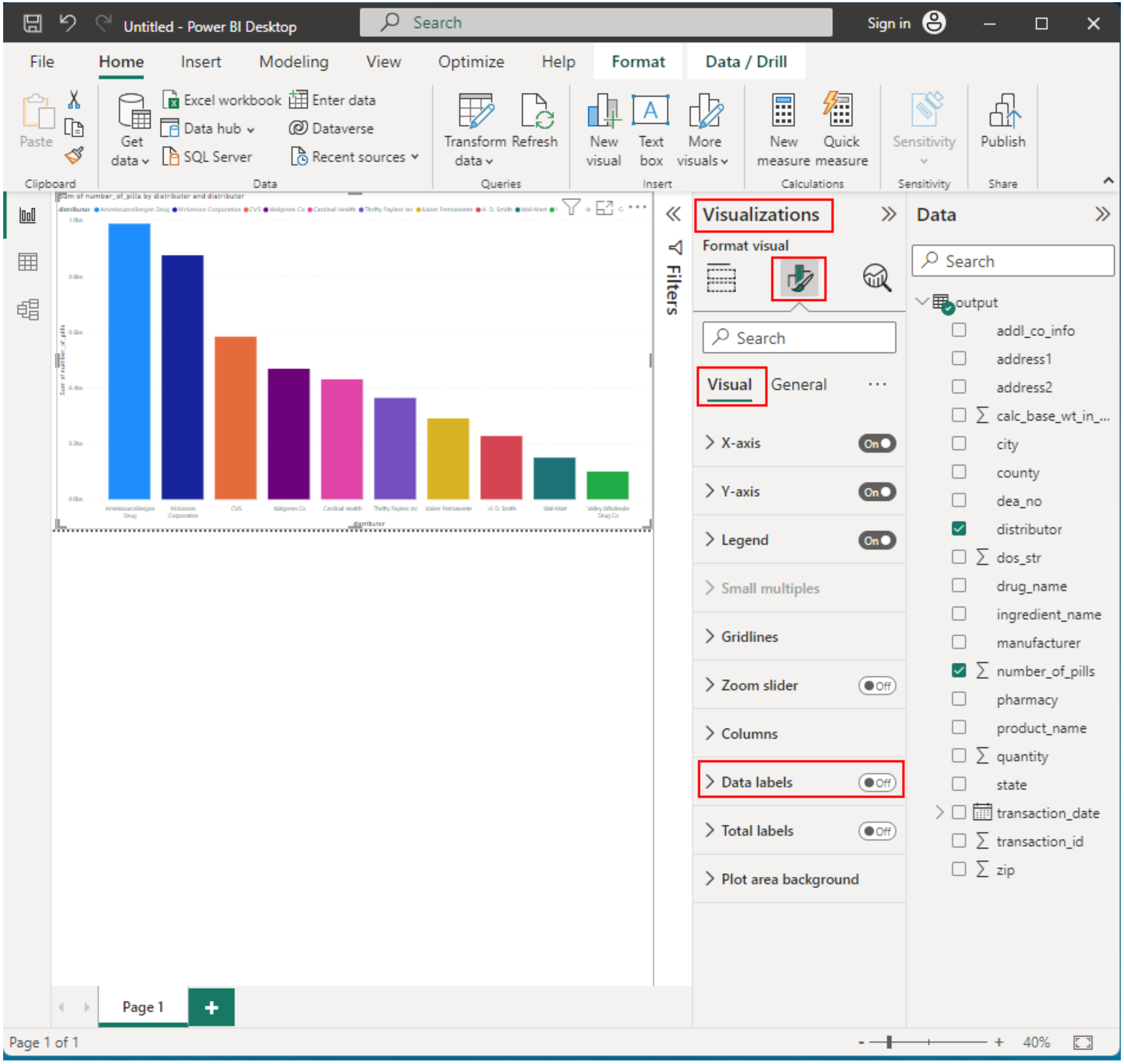
Drill through

Cross-report: Off

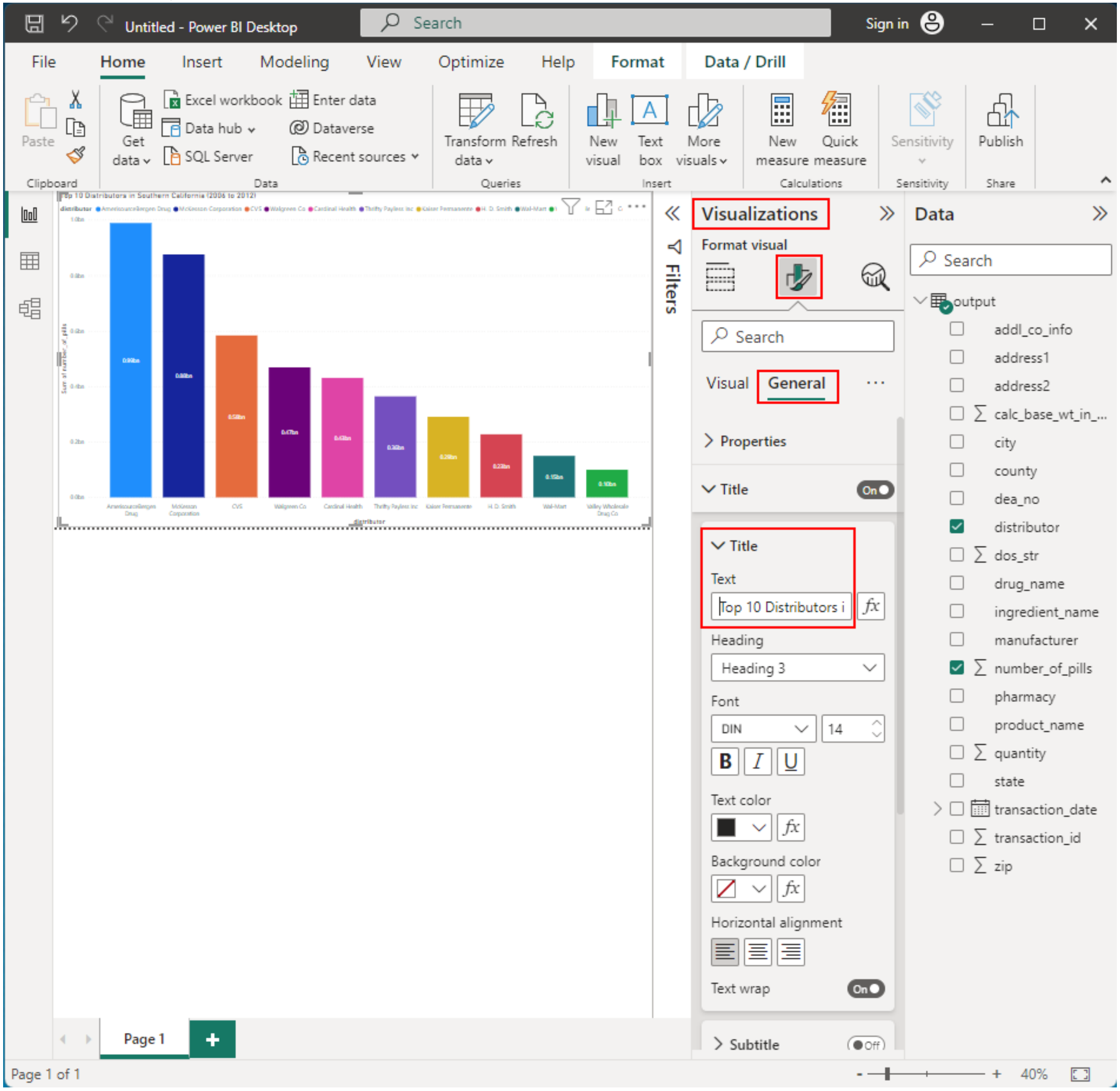
Keep all filters: On

Add drill-through fields here

11. Turn on the "Data labels" under "Visualizations" -> "Format your visual" -> "Visual"



12. Change the title to "Top 10 Distributors in Southern California (2006 to 2012)" under "Visualizations" -> "Format your visual" -> "General"



Title On

Title

Text: Top 10 Distributors i

Heading: Heading 3

Font: DIN 14

B I U

Text color

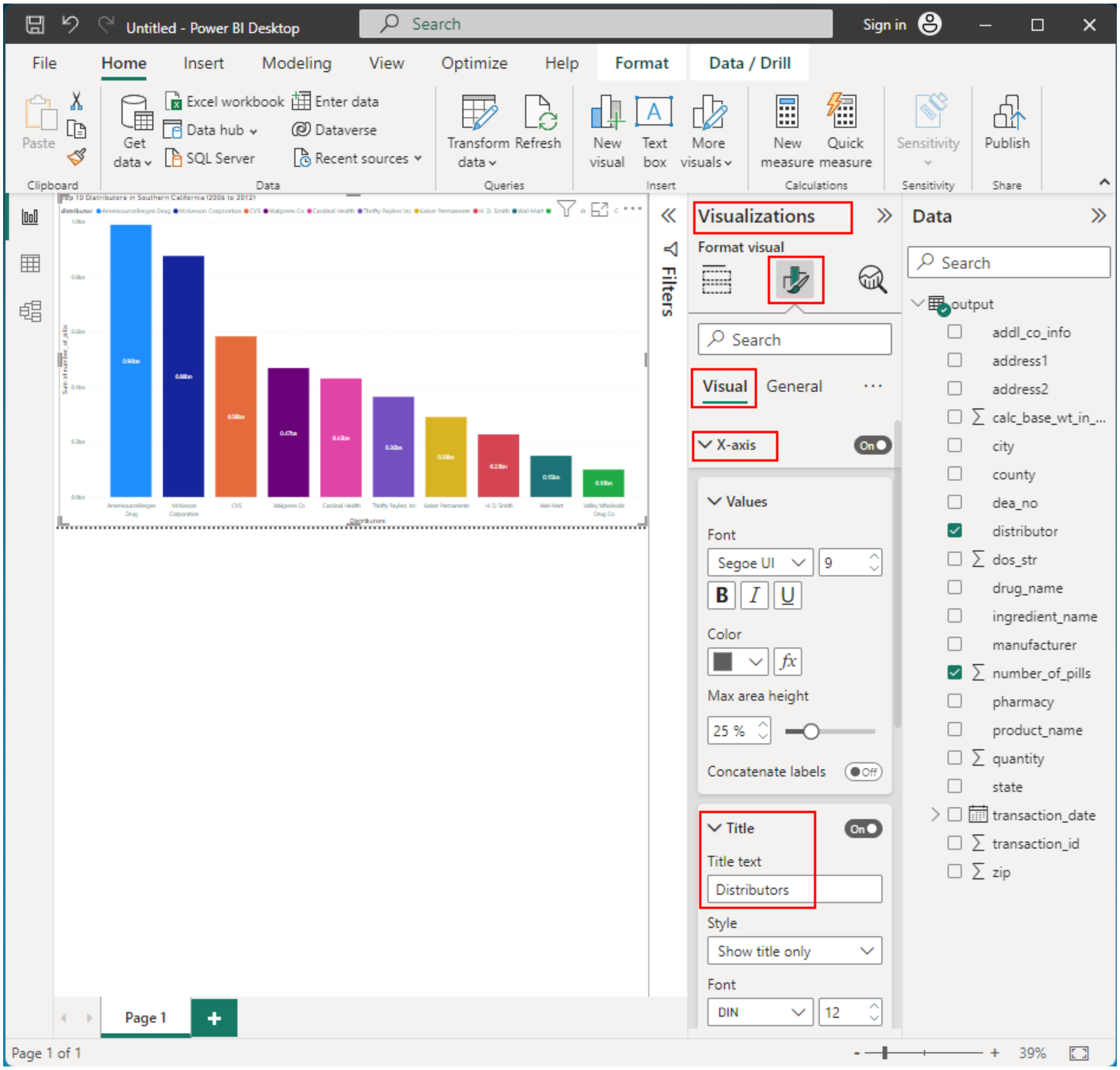
Background color

Horizontal alignment

Text wrap On

Subtitle Off

13. Change the X-axis title to "Distributors" under "Visualizations" -> "Format your visual" -> "Visual" -> "X-axis" -> "Title"



X-axis: On

Values

Font: Segoe UI, 9

B I U

Color

Max area height: 25 %

Concatenate labels: Off

Title: On

Title text: Distributors

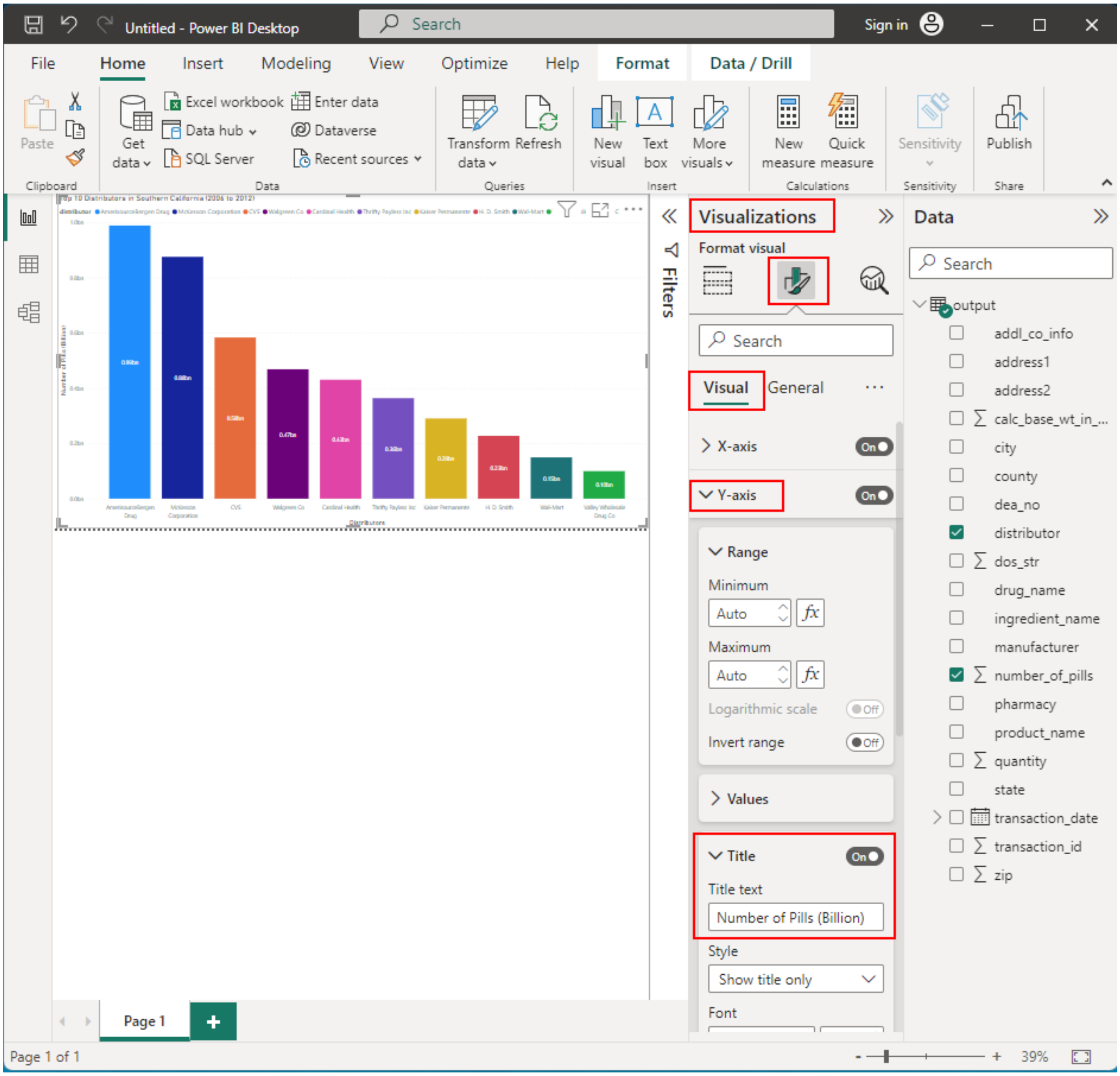
Style: Show title only

Font: DIN, 12

output

- ☐ addl_co_info
- ☐ address1
- ☐ address2
- ☐ ∑ calc_base_wt_in_...
- ☐ city
- ☐ county
- ☐ dea_no
- ☑ distributor
- ☐ ∑ dos_str
- ☐ drug_name
- ☐ ingredient_name
- ☐ manufacturer
- ☑ ∑ number_of_pills
- ☐ pharmacy
- ☐ product_name
- ☐ ∑ quantity
- ☐ state
- ☐ transaction_date
- ☐ ∑ transaction_id
- ☐ ∑ zip

14. Change the Y-axis title to "Number of Pills (Billion)" under "Visualizations" -> "Format your visual" -> "Visual" -> "Y-axis" -> "Title"



X-axis: On

Y-axis: On

Minimum: Auto

Maximum: Auto

Logarithmic scale: Off

Invert range: Off

Title: On

Title text: Number of Pills (Billion)

Style: Show title only

☑ distributor

☑ number_of_pills

15. Repeat steps 6 to 14 to create a Top 10 chart for manufacturers, pharmacies, and products.

References

16. Data Source: https://www.washingtonpost.com/graphics/2019/investigations/dea-pain-pill-database/

17. Github: https://github.com/vcheung621/cis4560

18. References: https://www.kaggle.com/datasets/paultimothymooney/pain-pills-in-the-usa