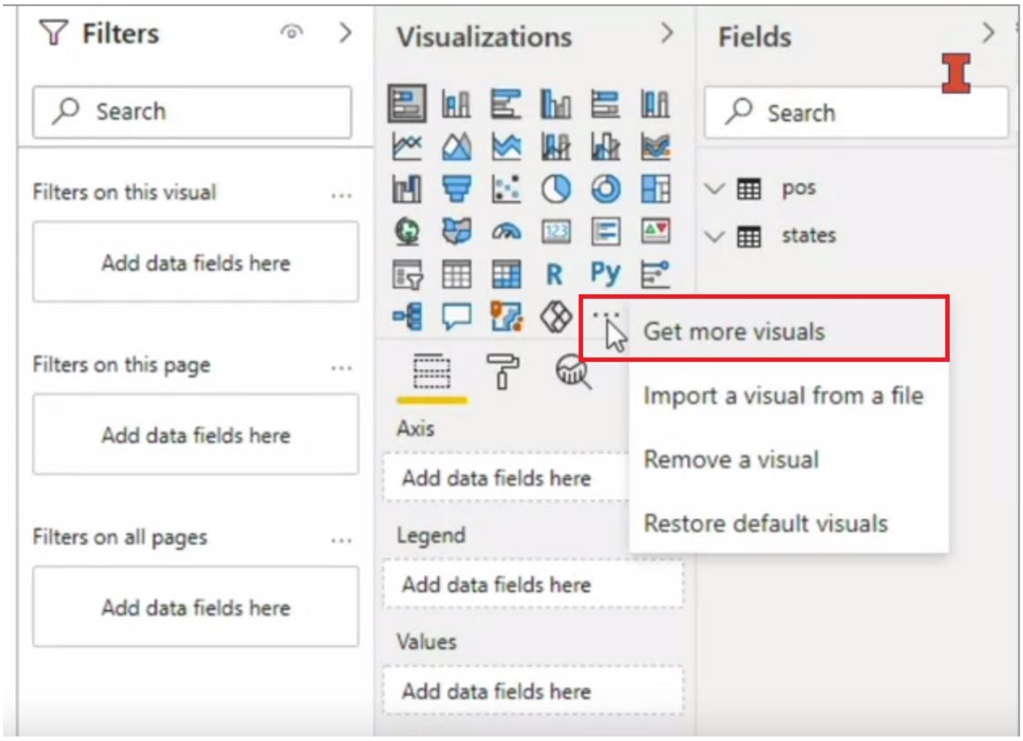
Power BI – Exploratory Data Analysis

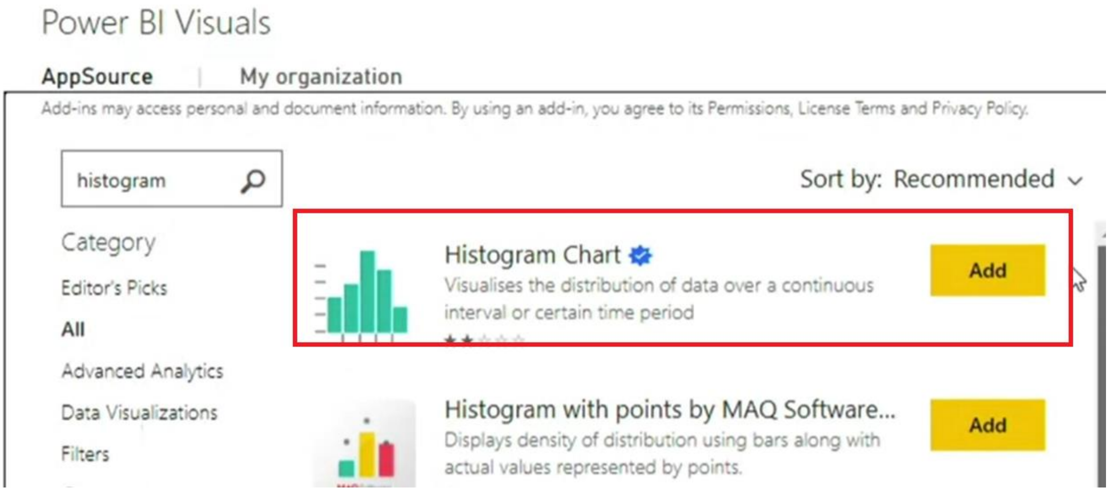
EDA 1 – Univariate Plots for Numeric Data: Histograms and Boxplots

Histograms don't come by default with Power BI. Click on the three dots and select 'Get more visuals'.

You will need to sign in using your company's email.

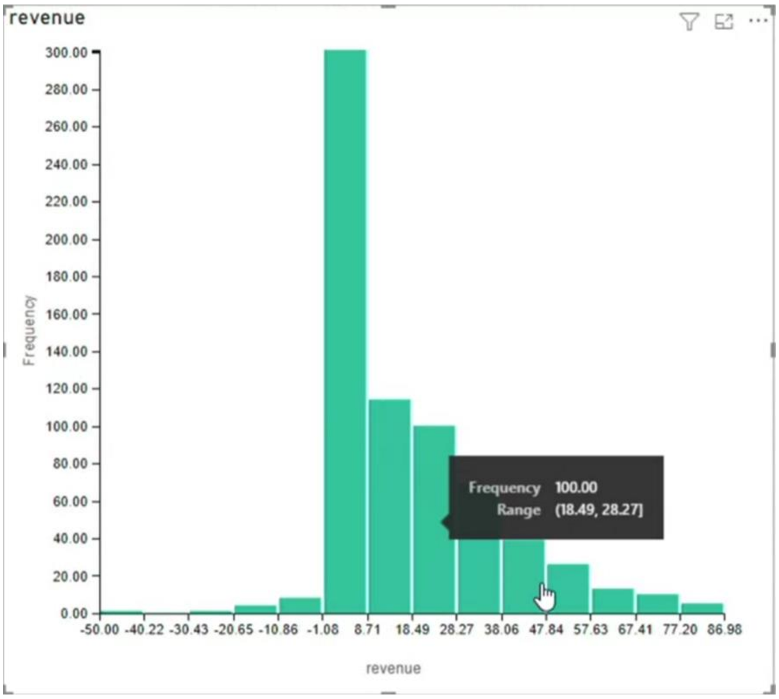


Type Histogram in the search box and select

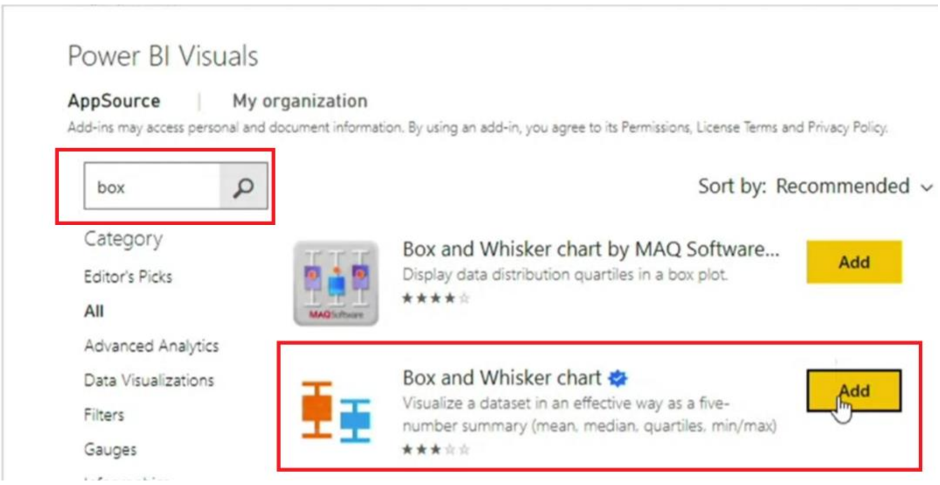


Power BI Visuals
AppSource
My organization
Add-ins may access personal and document information. By using an add-in, you agree to its Permissions, License Terms and Privacy Policy.
histogram
Sort by: Recommended
Category
Editor's Picks
All
Advanced Analytics
Data Visualizations
Filters
Histogram Chart
Visualises the distribution of data over a continuous interval or certain time period
Add
Histogram with points by MAQ Software...
Displays density of distribution using bars along with actual values represented by points.
Add

A Histogram takes a continuous variable or a column of data that is numeric and divides it into equally sized bins.



Now select a Box and Whisker plot in the same way



Power BI Visuals

AppSource | My organization

Add-ins may access personal and document information. By using an add-in, you agree to its Permissions, License Terms and Privacy Policy.

box

Sort by: Recommended

Category

Editor's Picks

All

Advanced Analytics

Data Visualizations

Filters

Gauges

Box and Whisker chart by MAQ Software...

Display data distribution quartiles in a box plot.

★★★★☆

Add

Box and Whisker chart

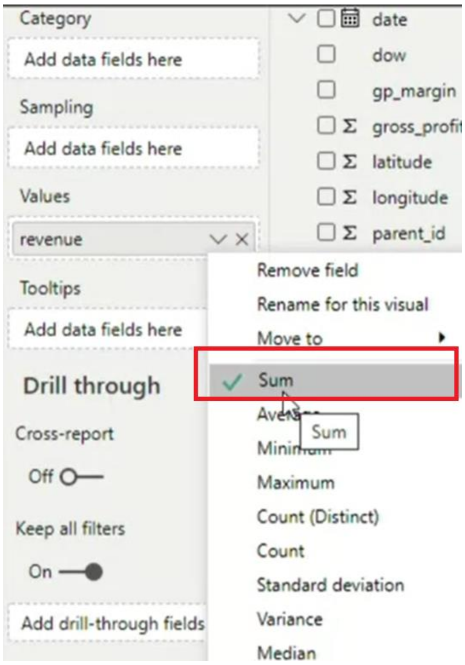
Visualize a dataset in an effective way as a five-number summary (mean, median, quartiles, min/max)

★★★☆☆

Add

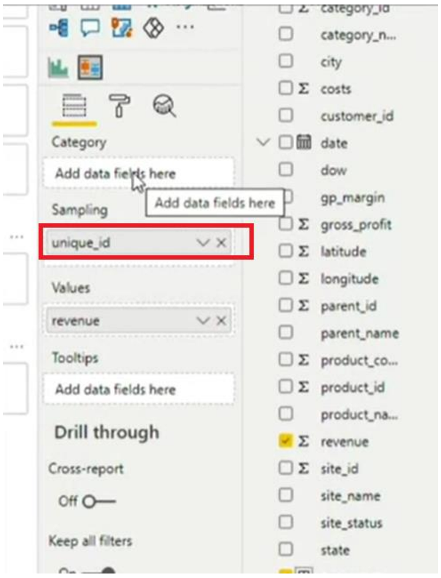
If nothing shows in the plot after dragging revenue in the axis, check if there is an aggregation.

Power BI is really good at aggregation and is summing up all of the revenue observations into one data point.

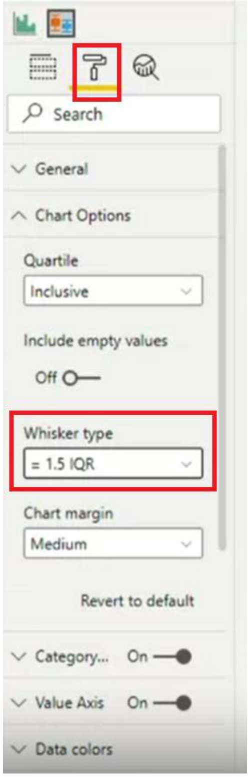


Carry out some sampling to remove the aggregation. Add the unique_id.

Add the super_parent_name to the Category field.



Now format the Box Plot by clicking on the Roller Paint brush icon, 'Chart Options' and then select 1.5 IQR as the Whisker type.

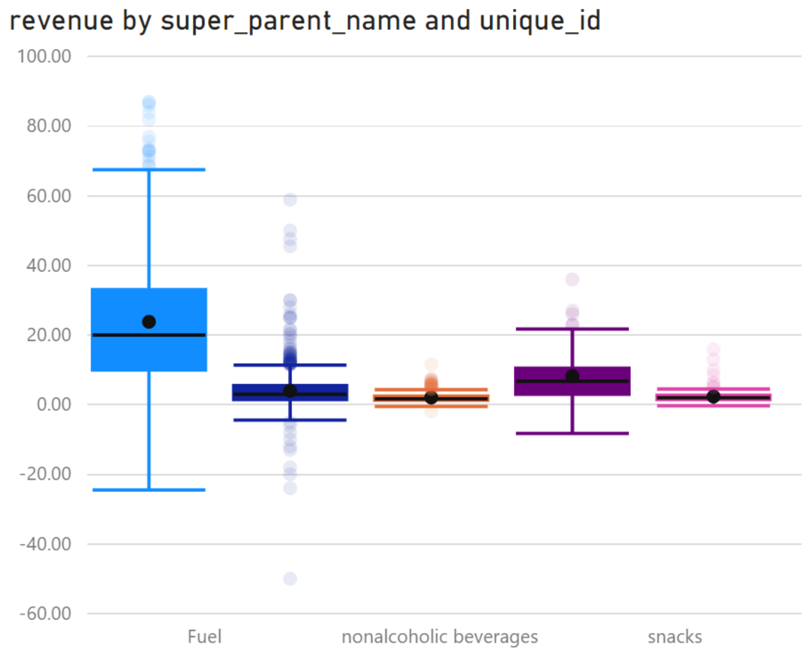


Quartile: Inclusive

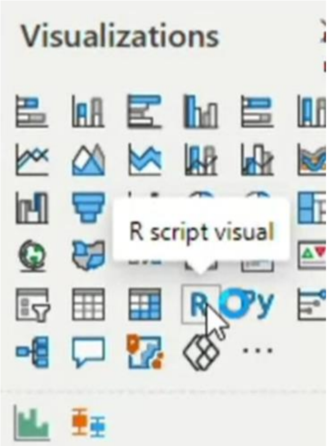
Include empty values: Off

Whisker type: ≈ 1.5 IQR

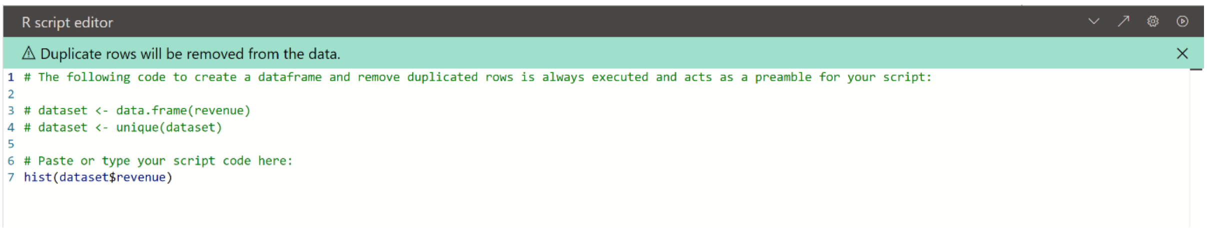
Chart margin: Medium



Now add a R Chart by clicking on R Script visual



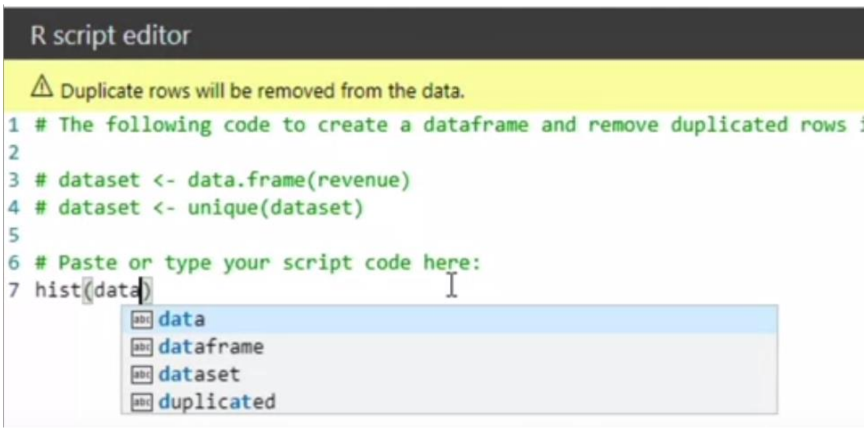
Drag and drop in revenue to the plot and it populates the R Script editor. It creates a dataset with revenue as a column.



⚠ Duplicate rows will be removed from the data.

```
# The following code to create a dataframe and remove duplicated rows is always executed and acts as a preamble for your script:

# dataset <- data.frame(revenue)
# dataset <- unique(dataset)

# Paste or type your script code here:
hist(dataset$revenue)
```

Now we can type in the R editor the type of plot that we want e.g. hist() specify dataset as parameter

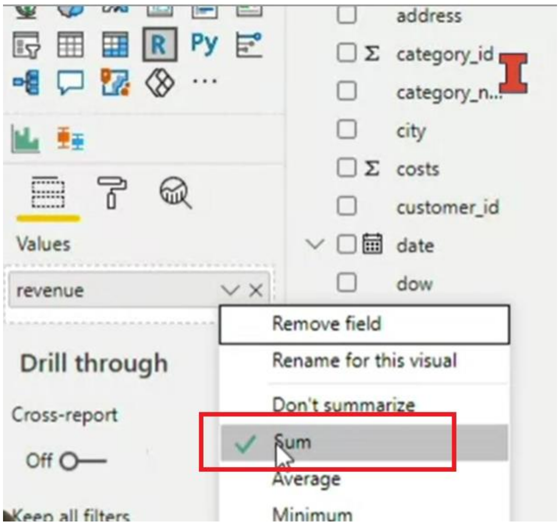


R script editor

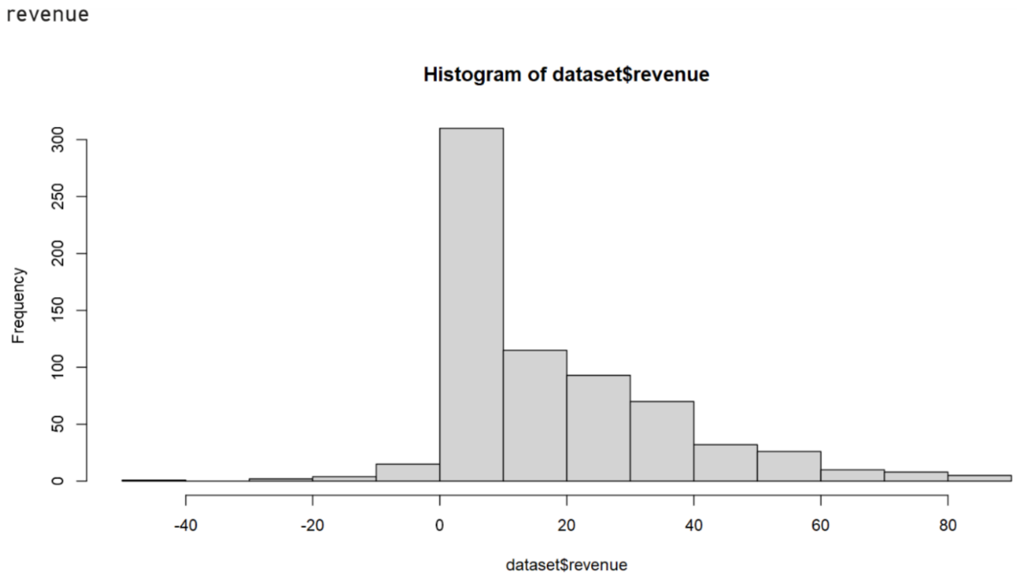
⚠ Duplicate rows will be removed from the data.

```
# The following code to create a dataframe and remove duplicated rows i

# dataset <- data.frame(revenue)
# dataset <- unique(dataset)

# Paste or type your script code here:
hist(data)
```

If there is any issue check to see if there is an aggregation, and then undo the aggregation by selecting ‘Don’t summarize’.



Now we can see the Histogram

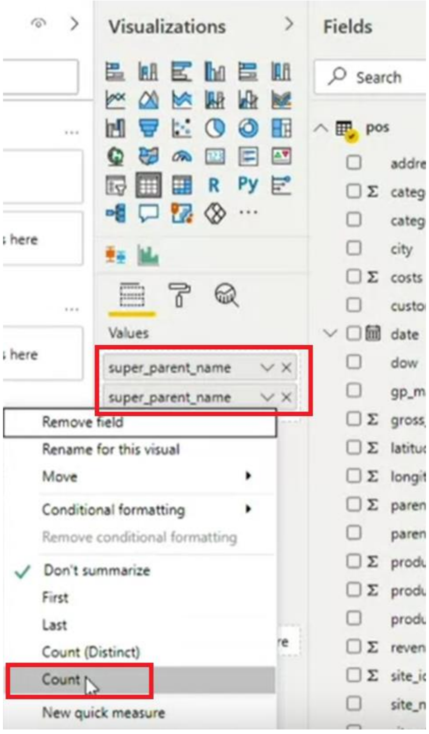


EDA 2 – Univariate, Bivariate, and Multivariate Plots with Categorical Data

Categorical variable are not numeric.

Check out the data using the 'Table' icon.

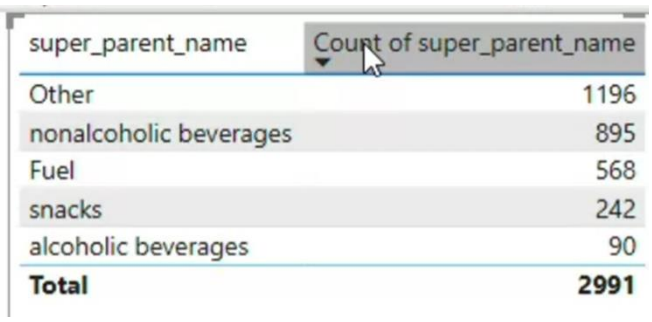
Put two super_parent_names next to each other and set the second one to Count.



This should not display the categories and their count.

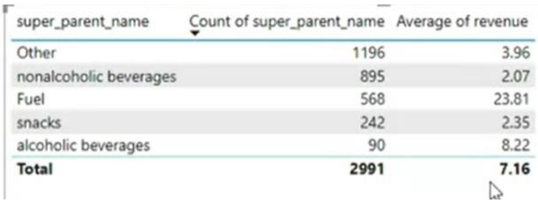
We can see that we have 90 observations that are alcoholic beverages.

We can also click on the column name to get descending order by count for example.



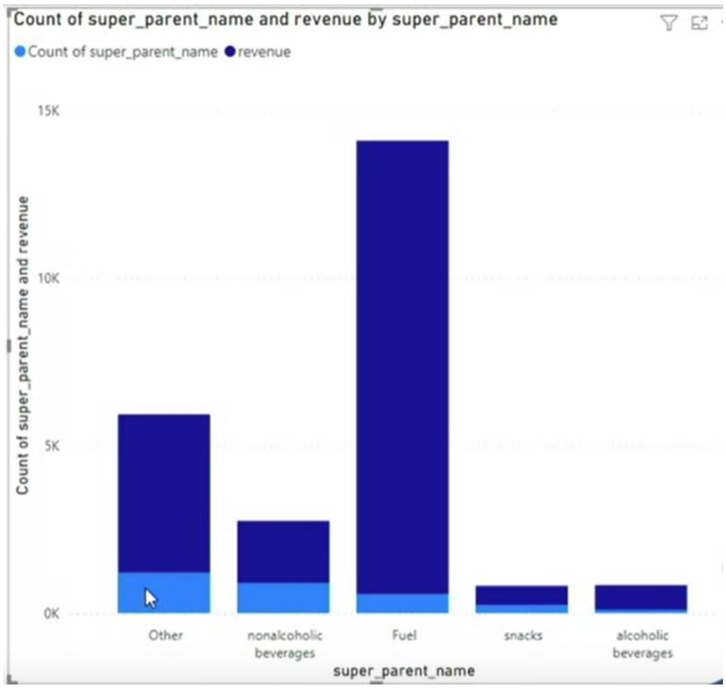
super_parent_name	Count of super_parent_name
Other	1196
nonalcoholic beverages	895
Fuel	568
snacks	242
alcoholic beverages	90
Total	**2991**

We could even drag in revenue to get a bivariate analysis.



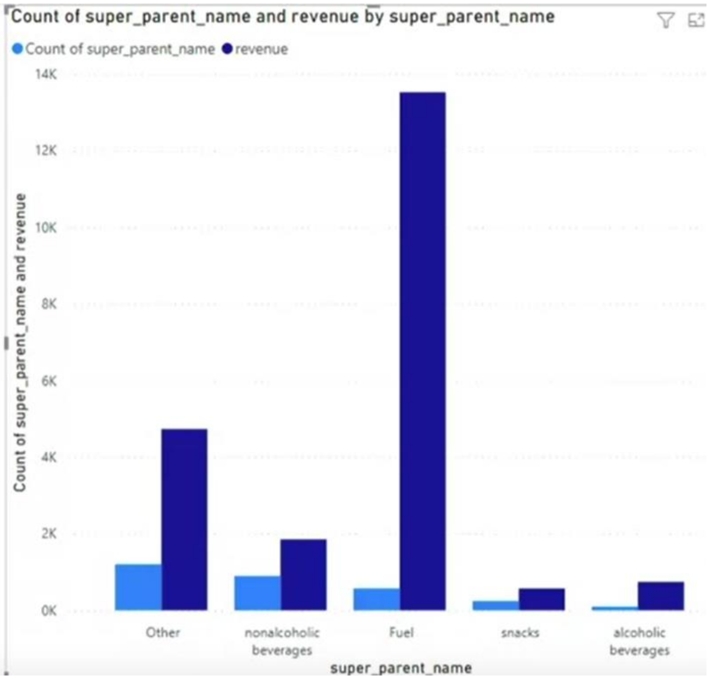
super_parent_name	Count of super_parent_name	Average of revenue
Other	1196	3.96
nonalcoholic beverages	895	2.07
Fuel	568	23.81
snacks	242	2.35
alcoholic beverages	90	8.22
Total	**2991**	**7.16**

Let’s now create a bar chart. This is a stacked bar chart showed revenue and count.

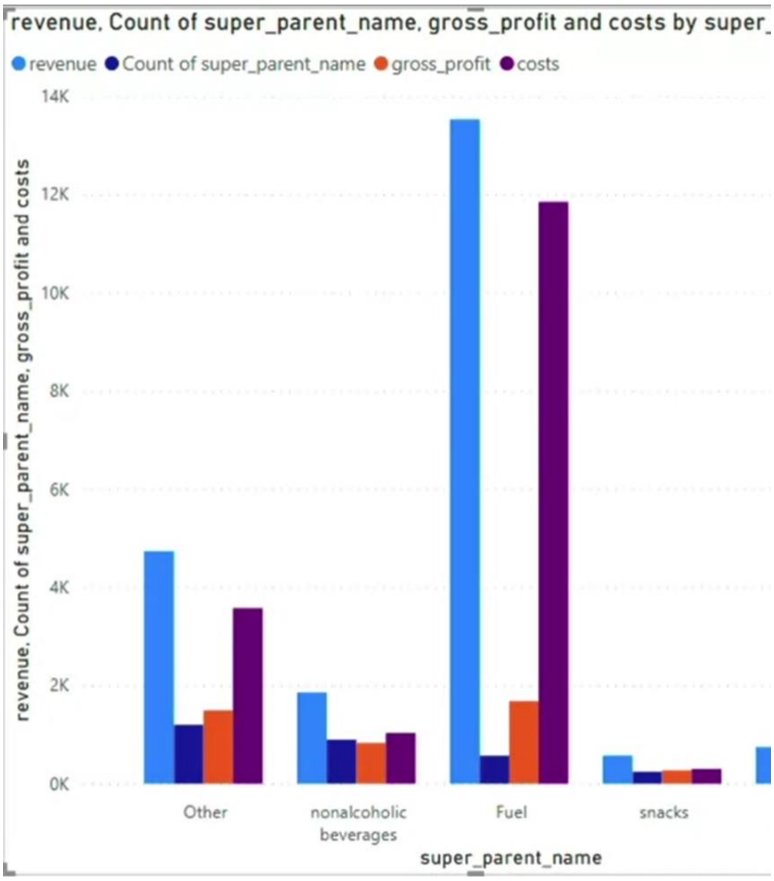


However, in this case stacking revenue and count on top of each other isn’t useful.

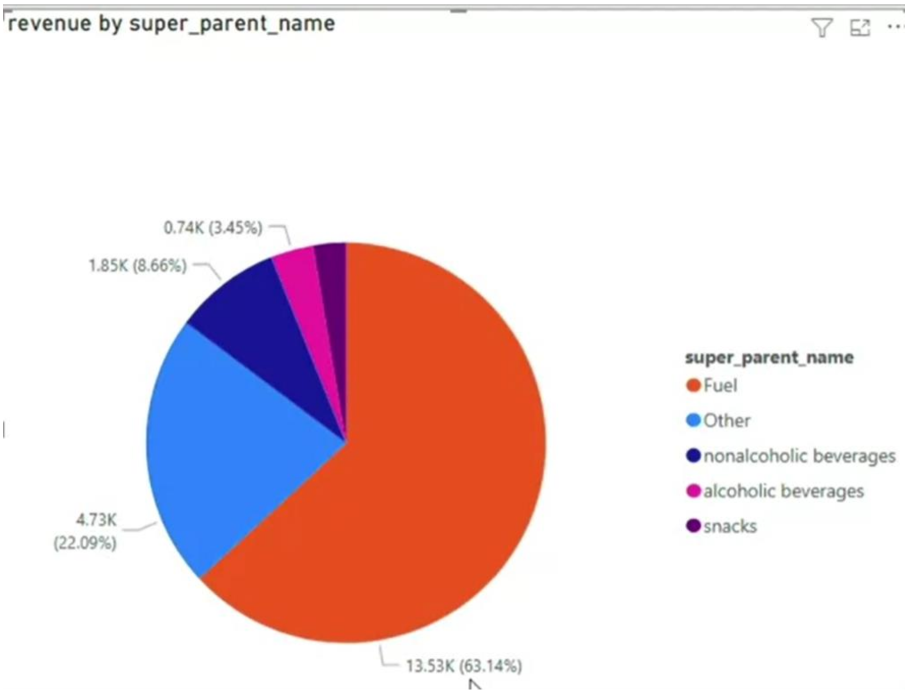
We could created a clustered bar chart instead.



We could add gross profit and costs to the chart by dragging and dropping those fields.



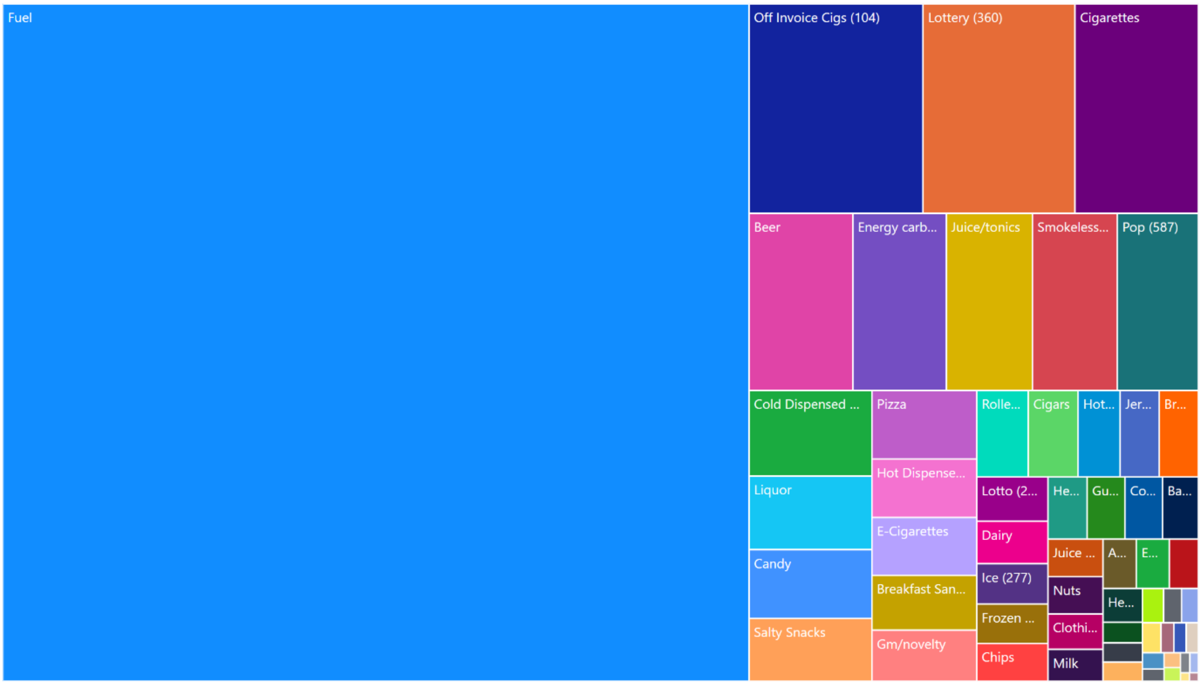
For Exploratory Data Analysis, it's useful to rotate between different types of charts. Bar charts are helpful to see relative amounts. A Pie Chart is often useful to show percentages.



We can see that Fuel make up 63.14% of the total revenue. Pie charts are very useful of 6 or less slices/categories especially if you're looking at percentages. The downside is that they don't display lots of pie slices very well. Another downside is that it is not ideal to have the legend separate from the slices of the pie. Another type of chart is a tree map. It's like a pie chart but in rectangular form.

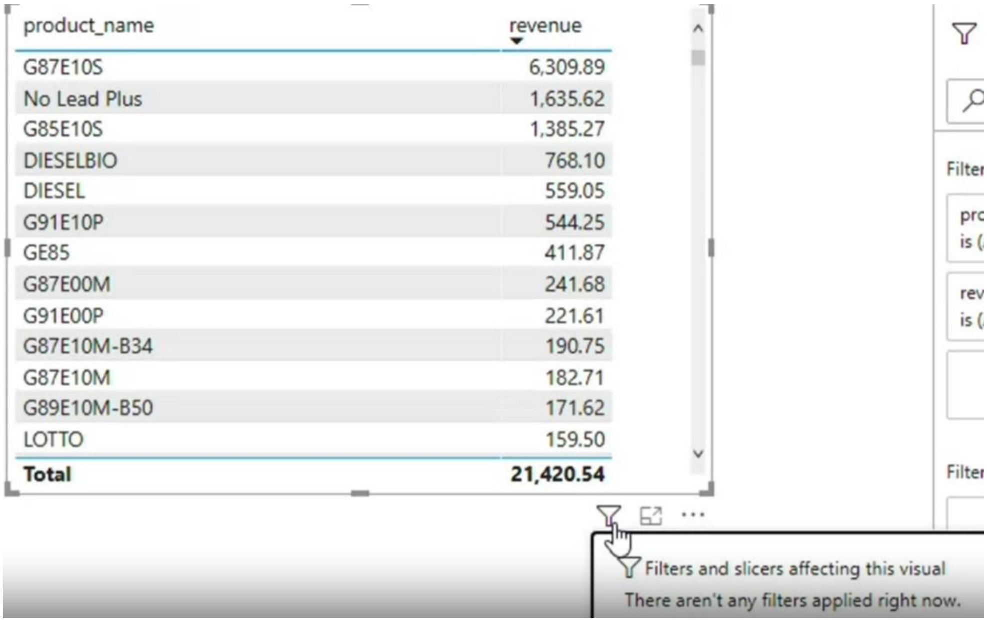
However, the pieces of the tree map have the label on them and they are also much more helpful for displaying categorical data that have more than 6 different values.

revenue by parent_name



EDA 3 – Filters, Slicers and Drill Through

In the 'Categorical' tab add a table. From the pos dataset, add in product_name and revenue. Sort it by highest to lowest revenue. Clicking on the funnel icon shows what filters are applied. Currently none.



product_name	revenue
G87E10S	6,309.89
No Lead Plus	1,635.62
G85E10S	1,385.27
DIESELBIO	768.10
DIESEL	559.05
G91E10P	544.25
GE85	411.87
G87E00M	241.68
G91E00P	221.61
G87E10M-B34	190.75
G87E10M	182.71
G89E10M-B50	171.62
LOTTO	159.50
Total	**21,420.54**

Filte

pro
is (

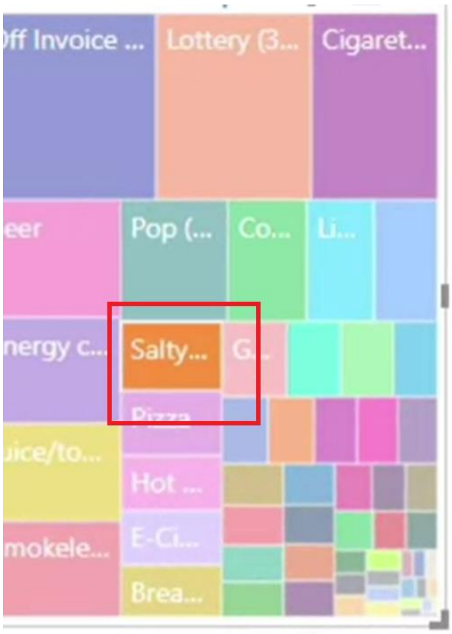
rev
is (

Filte

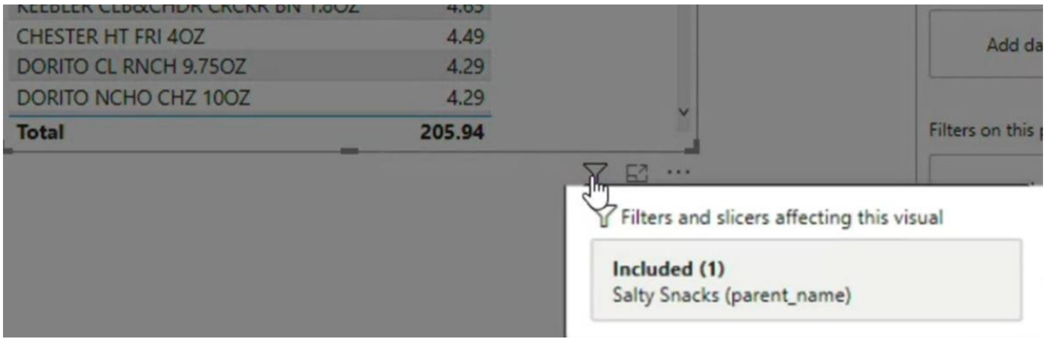
Filters and slicers affecting this visual

There aren't any filters applied right now.

We can now click on Salty snacks in the tree map.



Clicking on the funnel now shows that the table is filtered by Salty snacks.

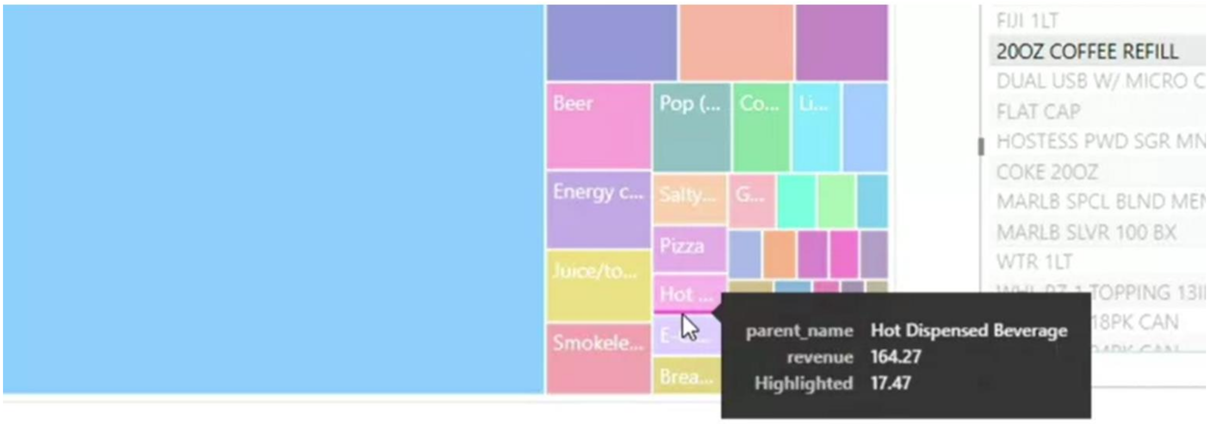


KEEBLER CLB&CHDR CRCKR BN 1.8OZ	4.65
CHESTER HT FRI 4OZ	4.49
DORITO CL RNCH 9.75OZ	4.29
DORITO NCHO CHZ 10OZ	4.29
Total	**205.94**

Filters and slicers affecting this visual

Included (1)
Salty Snacks (parent_name)

We can also select a row in the table to highlight something in the treemap



FIJI 1LT
20OZ COFFEE REFILL
DUAL USB W/ MICRO C
FLAT CAP
HOSTESS PWD SGR MN
COKE 20OZ
MARLB SPCL BLND ME
MARLB SLVR 100 BX
WTR 1LT
WHL 07 1 TOPPING 13
18PK CAN

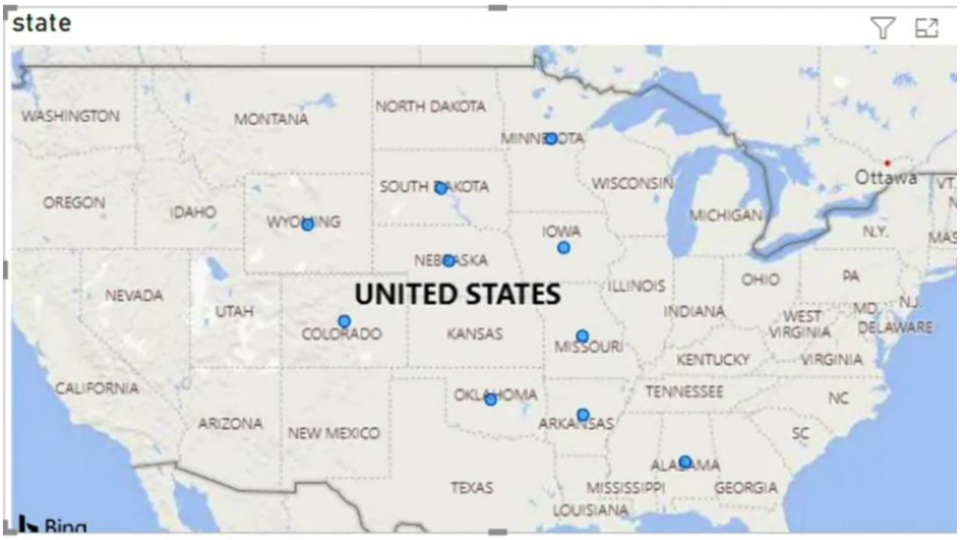
parent_name Hot Dispensed Beverage
revenue 164.27
Highlighted 17.47

We can see the context from the label.

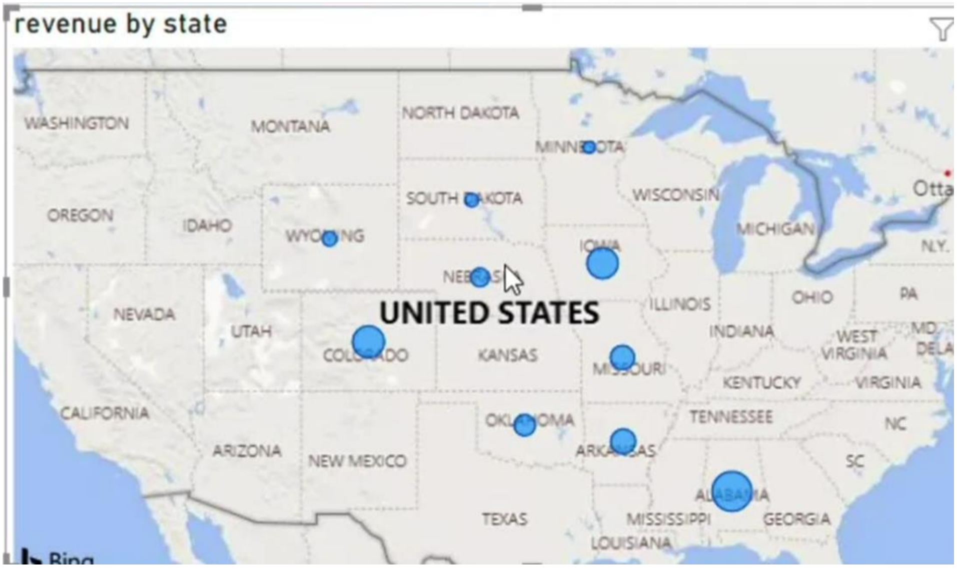
There are other ways to filter.

Now let’s add in a map.

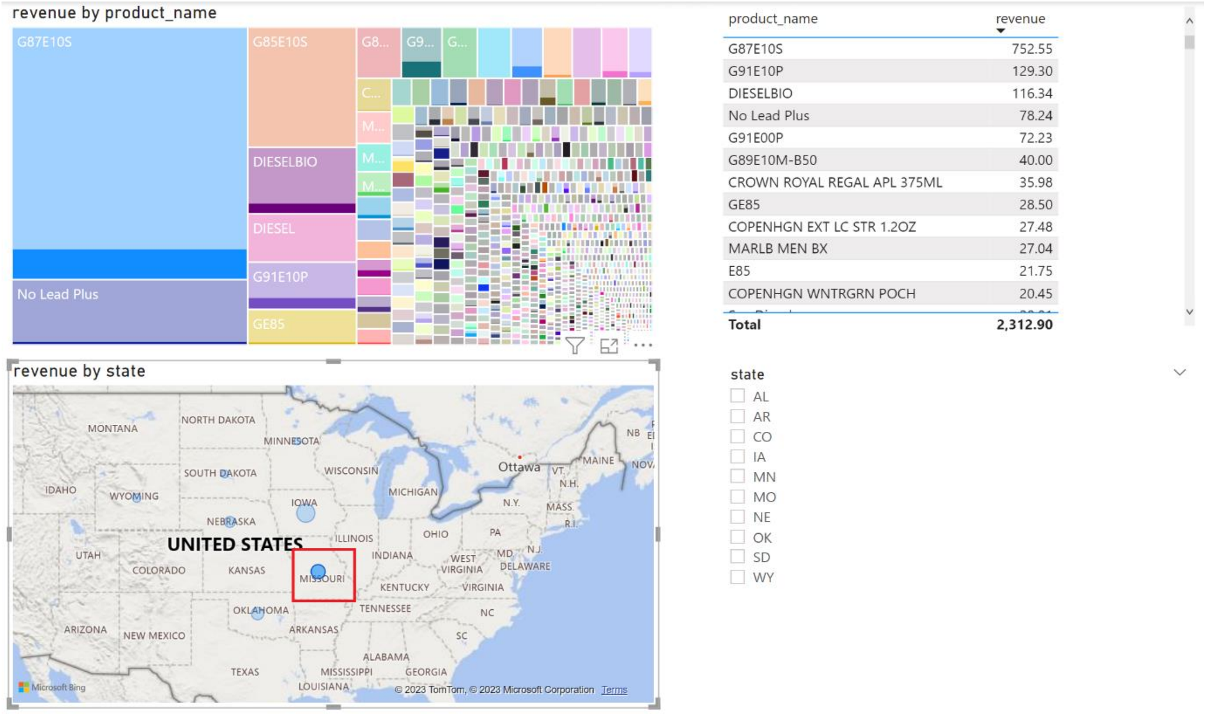
Add state into the location box



We can resize the dots according to revenue in those states



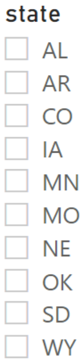
We can also filter the data by clicking on a state on the map. This also updates the tree map and the table data.



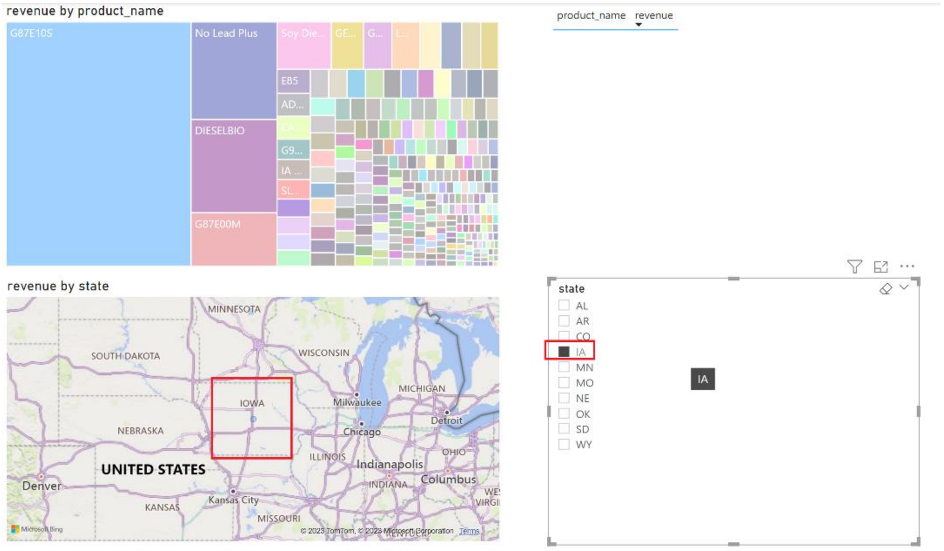
product_name	revenue
G87E10S	752.55
G91E10P	129.30
DIESELBIO	116.34
No Lead Plus	78.24
G91E00P	72.23
G89E10M-B50	40.00
CROWN ROYAL REGAL APL 375ML	35.98
GE85	28.50
COPENHGN EXT LC STR 1.2OZ	27.48
MARLB MEN BX	27.04
E85	21.75
COPENHGN WNTRGRN POCH	20.45
Total	**2,312.90**

state
□ AL
□ AR
□ CO
□ IA
□ MN
□ MO
□ NE
□ OK
□ SD
□ WY

Another way to filter the data is by using slicers. The below is a slicer on the state. This is useful if we want to publish a report for others to use.



state
□ AL
□ AR
□ CO
□ IA
□ MN
□ MO
□ NE
□ OK
□ SD
□ WY

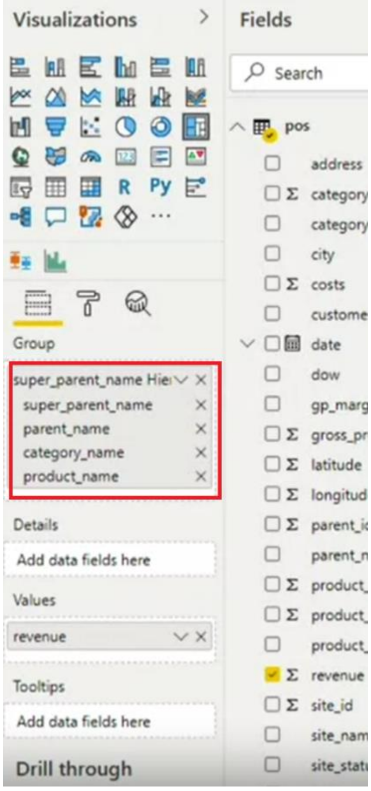


product_name	revenue

state
□ AL
□ AR
□ CO
■ IA
□ MN
□ MO
□ NE
□ OK
□ SD
□ WY

IA

In order to demonstrate the drill down functionality, we need to add the super_parent_name hierarchy to the Group.



Group

super_parent_name Hie
super_parent_name
parent_name
category_name
product_name

Details

Add data fields here

Values

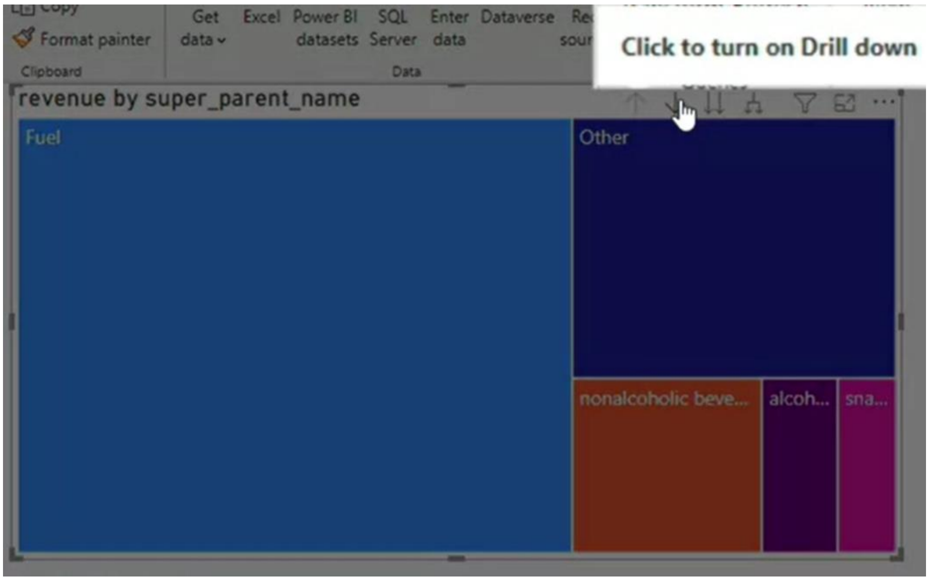
revenue

Tooltips

Add data fields here

Drill through

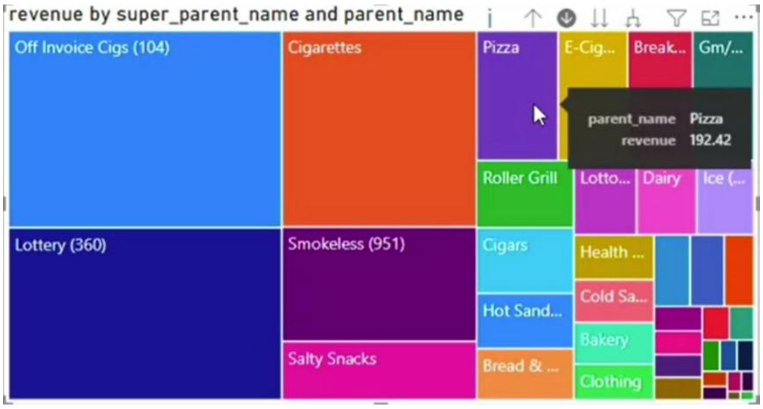
Now we have some new icons appearing on the tree map



Click to turn on Drill down

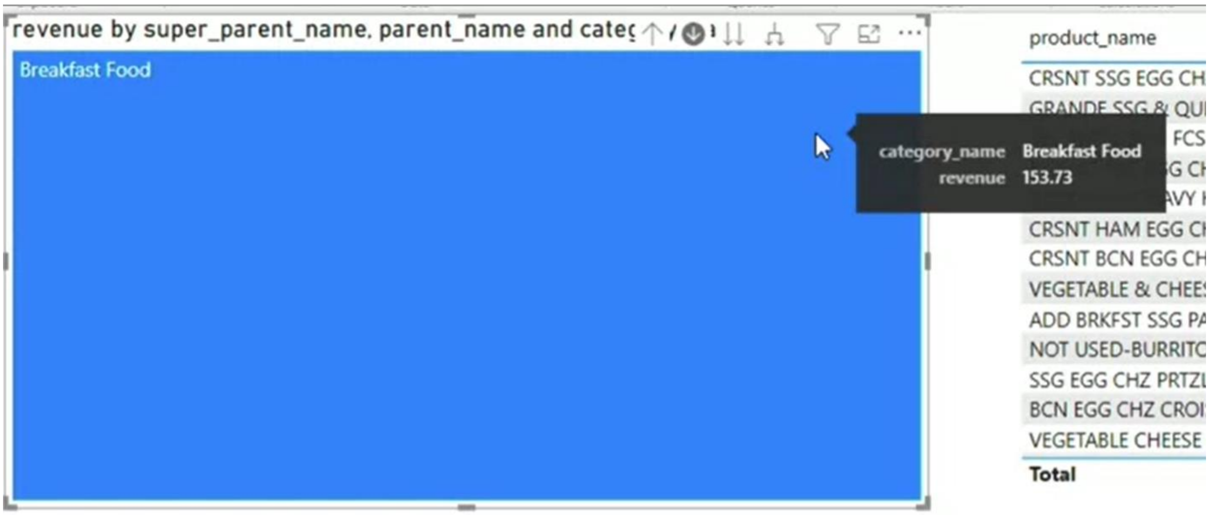
revenue by super_parent_name

If we now click on ‘Other’, we can see what makes up the total amount in the ‘Other’ category.



parent_name Pizza
revenue 192.42

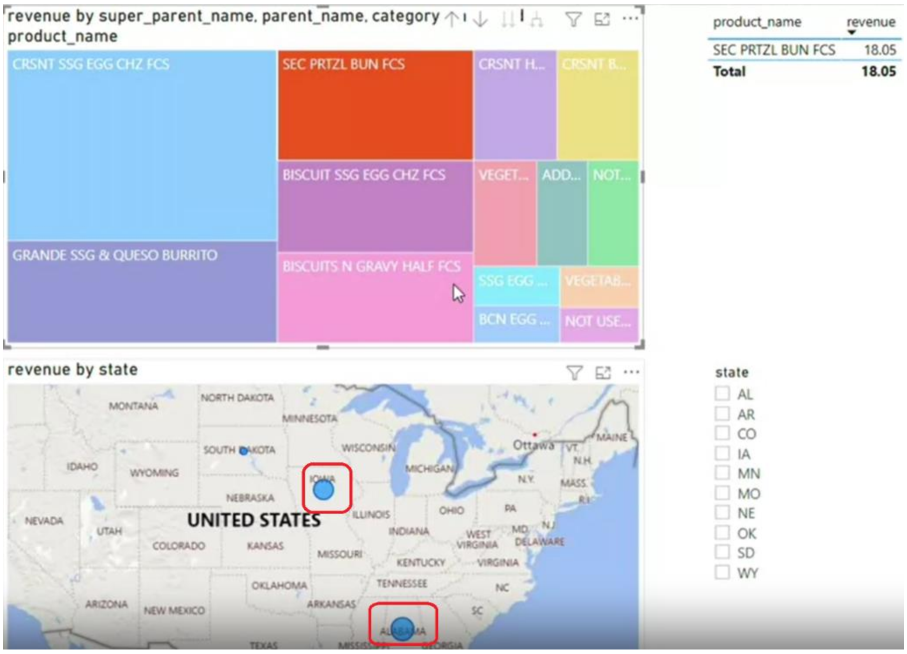
We can drill down further into Breakfast Sandwiches. The table is updated as well.



category_name Breakfast Food
revenue 153.73

product_name
CRSNT SSG EGG CH
GRANDE SSG & QU
FCS
G CH
VY I
CRSNT HAM EGG CI
CRSNT BCN EGG CH
VEGETABLE & CHEES
ADD BRKFST SSG PA
NOT USED-BURRITC
SSG EGG CHZ PRTZL
BCN EGG CHZ CROI
VEGETABLE CHEESE
Total

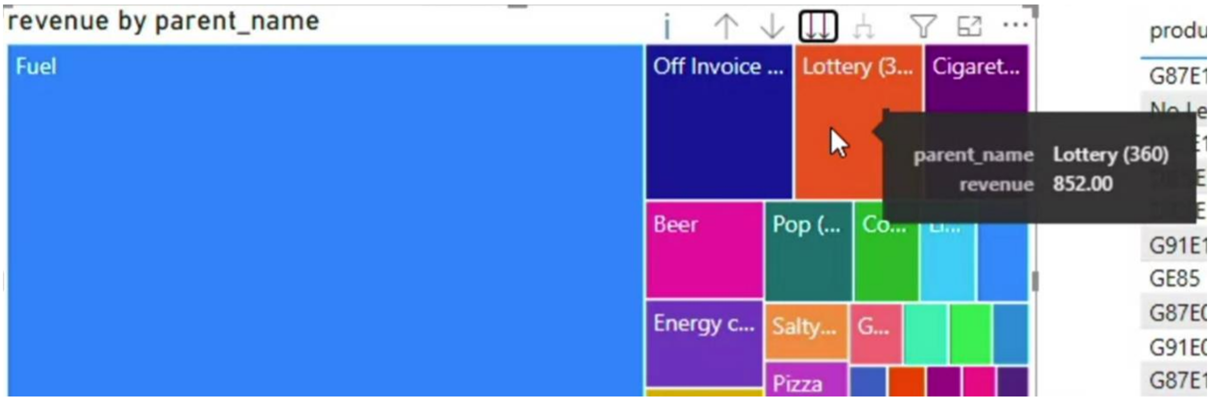
We can turn off the Drill Down and use the Tree map as a filter. We see the Pretzel buns were mostly sold in Iowa and Alabama.



product_name	revenue
SEC PRTZL BUN FCS	18.05
Total	**18.05**

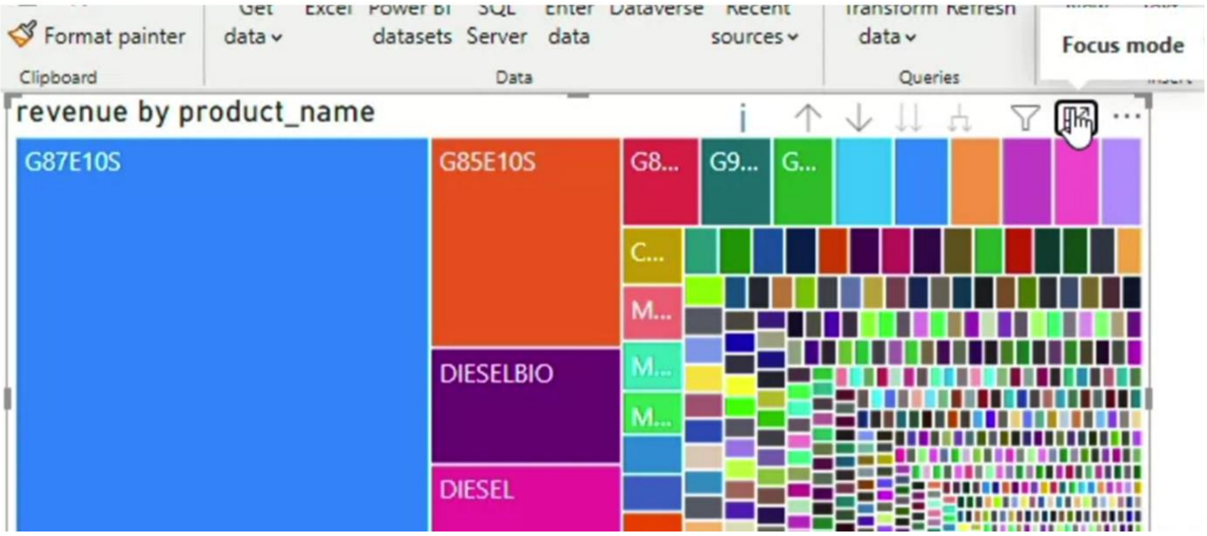
state
- ☐ AL
- ☐ AR
- ☐ CO
- ☐ IA
- ☐ MN
- ☐ MO
- ☐ NE
- ☐ OK
- ☐ SD
- ☐ WY

Another way to Drill down is clicking on the Double Arrow



parent_name Lottery (360)
revenue 852.00

We can also click on Focus mode to show more detail. However, sometimes it can be too much detail like below.

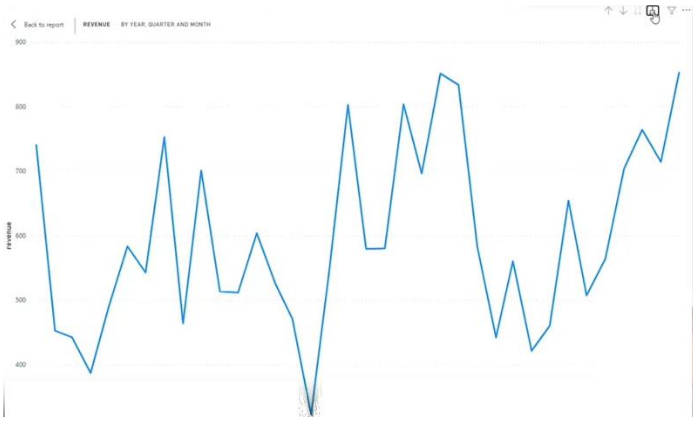


EDA 4 – Multivariate Plots: Scatter Plots and Line Plots

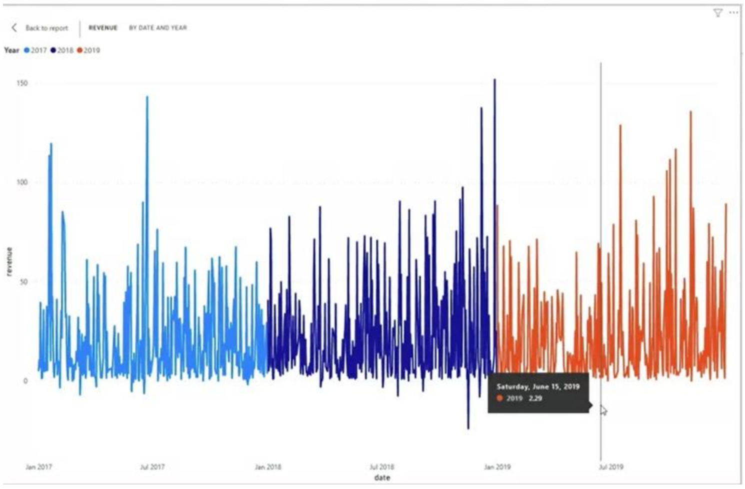
Multivariate plots look at 3 or more columns of data and these types of plots can contain a mixture of continuous and categorical data.

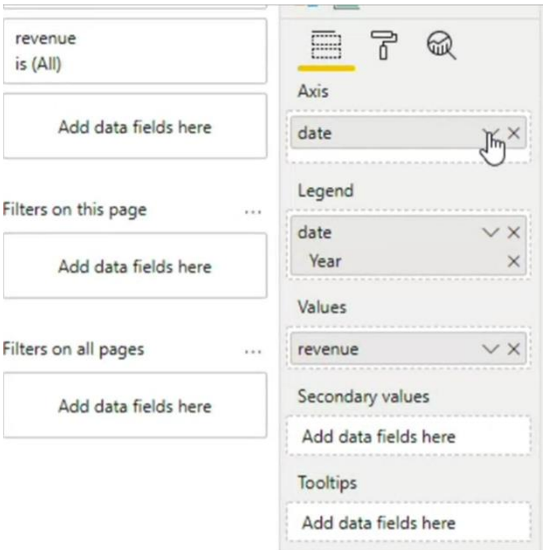
Line charts are useful for connecting points that are related to each other in terms of time.

We can also drill down to see revenue by year, quarter and month.



We could also see the daily sales coloured by date




revenue
is (All)

Add data fields here

Filters on this page ...

Add data fields here

Filters on all pages ...

Add data fields here

Axis

date

Legend

date

Year

Values

revenue

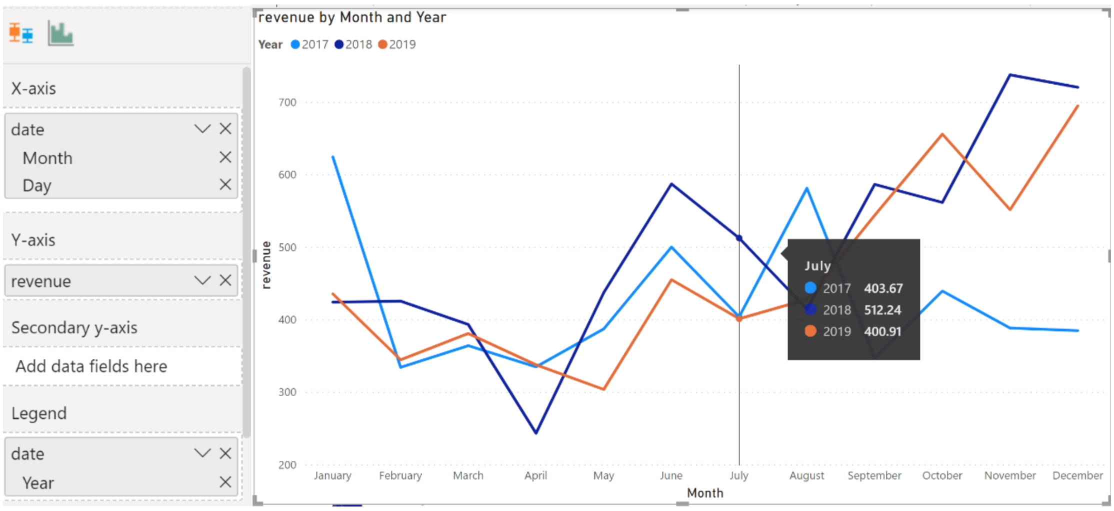
Secondary values

Add data fields here

Tooltips

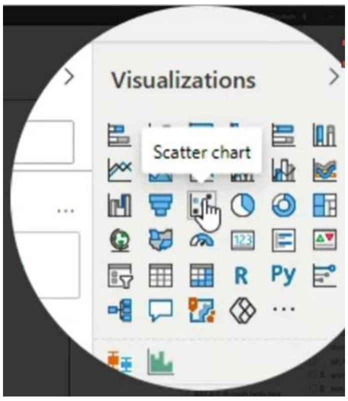
Add data fields here

We can also filter per month per year



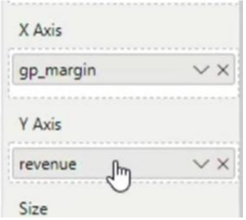
X-axis
date
Month
Day
Y-axis
revenue
Secondary y-axis
Add data fields here
Legend
date
Year

Now create a scatterplot by clicking on the Scatter chart icon



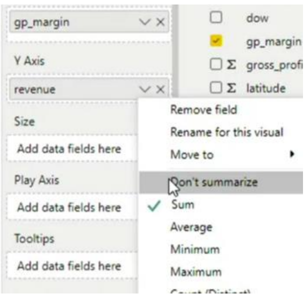
Let’s look at the relationship between gross profit margin and revenue.

The rationale being that we might want to investigate if units or line items that have a high gross profit margin also have a high revenue. Put gp_margin on the X-axis and revenue on the Y-axis.

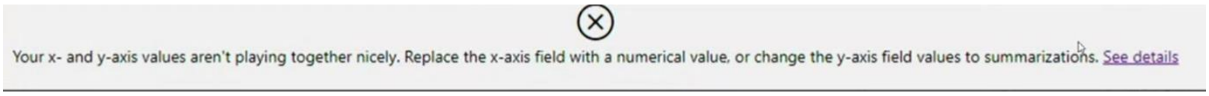


X Axis
gp_margin
Y Axis
revenue
Size

We seem to be aggregating so click on the down arrow for revenue and select ‘Don’t summarize’.



You may get the following error:

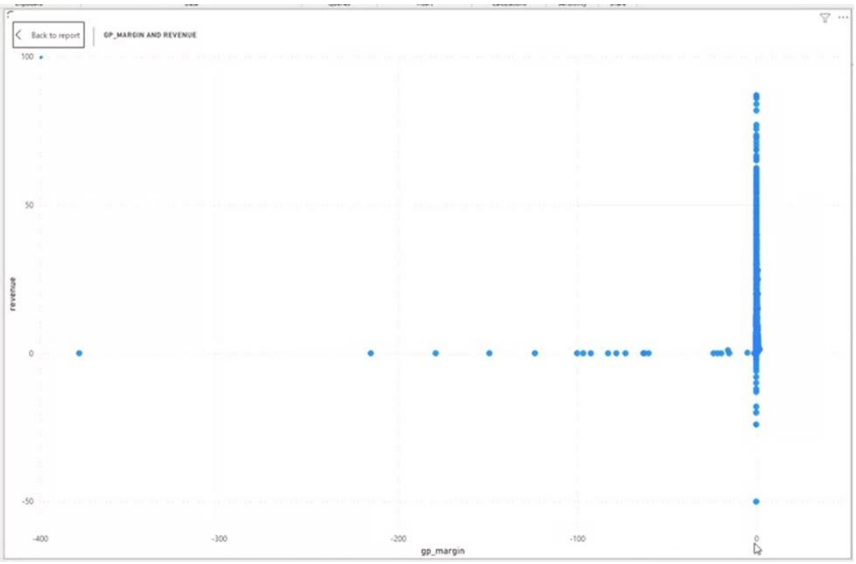


Your x- and y-axis values aren't playing together nicely. Replace the x-axis field with a numerical value, or change the y-axis field values to summarizations. See details

This is because gp_margin is currently a text data type.

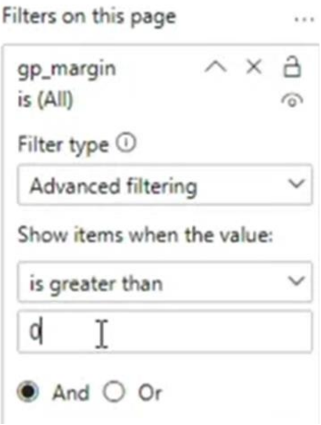
Change the data type to decimal using the Power Query Editor.

Now we get a scatter plot but it's not very informative



This is because we have a lot of negative outliers -400 -> 0.

Use a filter to remove them. Drag gp_margin to the Filter pane in the 'Filters on this page' section. We can use filters with numeric data as well. Add a greater than zero condition on the gp_margin.



Filters on this page

gp_margin
is (All)

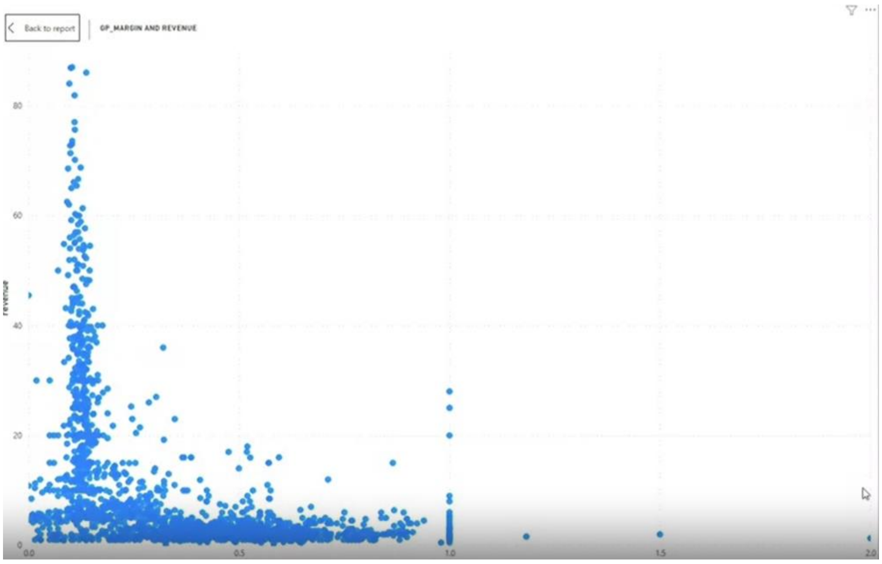
Filter type: Advanced filtering

Show items when the value:

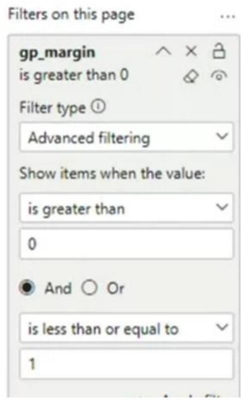
is greater than

0

● And ○ Or



Now we have a better Scatter plot but we could still remove some outliers. Add another filter to exclude point greater than 1.



Filters on this page ...

gp_margin
is greater than 0

Filter type ⓘ

Advanced filtering

Show items when the value:

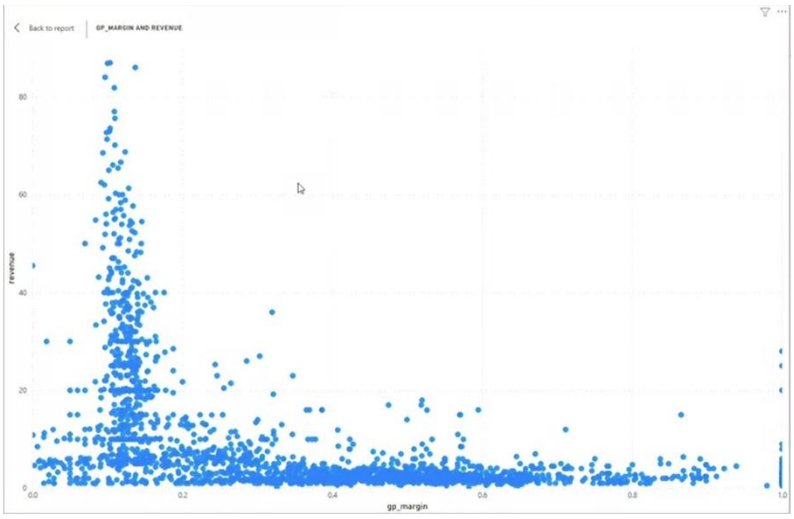
is greater than

0

● And ○ Or

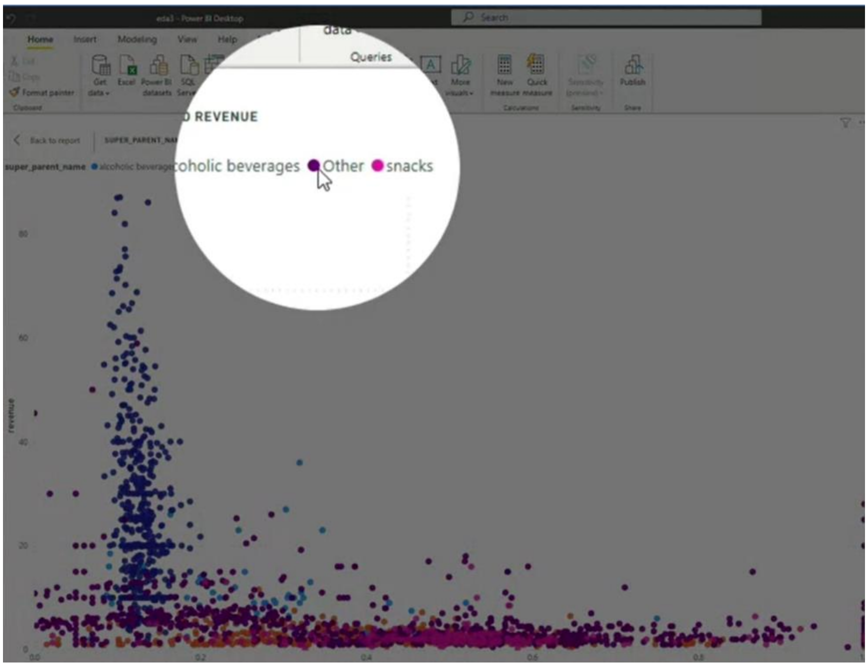
is less than or equal to

1

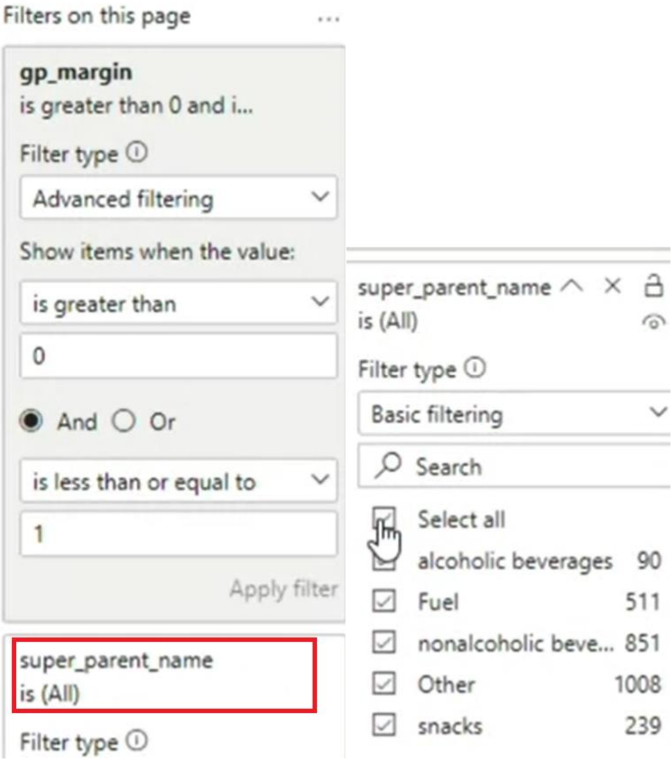


Let’s colour these points by dragging the super_parent_name to the Legend box.

We need to filter out the ‘Other’ observations as the data is too overwhelming.



We do this by dragging super_parent_name to the filter pane, selecting all and then unchecking ‘Other’.



Filters on this page ...

gp_margin
is greater than 0 and i...

Filter type ⓘ

Advanced filtering

Show items when the value:

is greater than

0

● And ○ Or

is less than or equal to

1

Apply filter

super_parent_name
is (All)

Filter type ⓘ

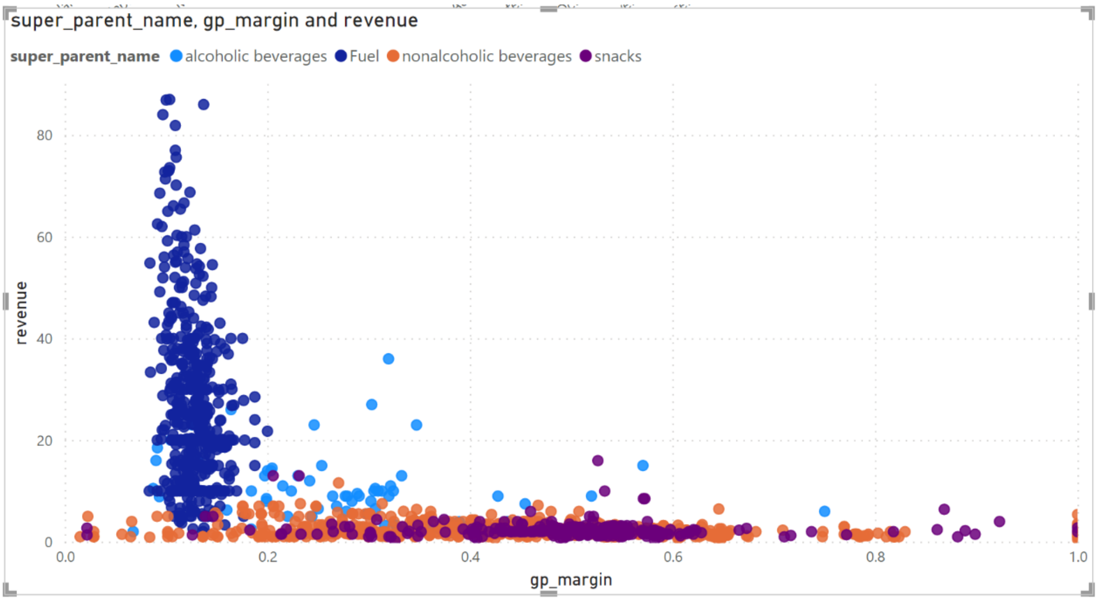
super_parent_name
is (All)

Filter type ⓘ

Basic filtering

Search

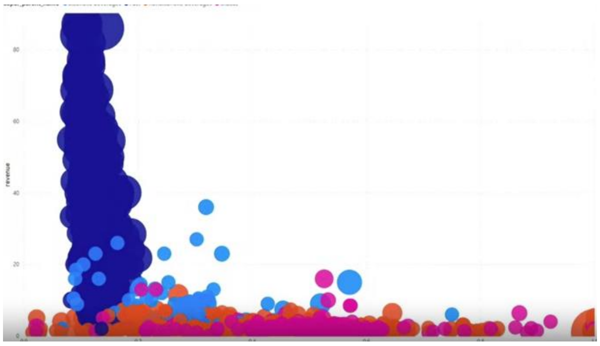
☑	Select all	
☑	alcoholic beverages	90
☑	Fuel	511
☑	nonalcoholic beve...	851
☑	Other	1008
☑	snacks	239



We can see that alcoholic beverages have often have a higher revenue and high gross profit margin.

Whereas non-alcoholic beverages and snacks have a really high gp margin but pretty low revenue.

Maybe some of these observations have high revenue because people are buying a lot of units of them. We can check this buy adding dragging units to the size box.



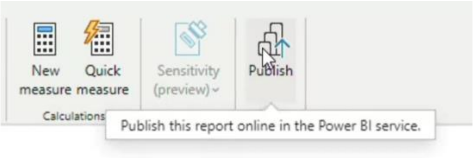
This seems to be quite overwhelming. This is because Fuel is in the gallons unit and in the teens or 20s whereas snacks are in the 1s and 2s.

One must be wary of how many multivariate plots are put on a report as it can get quite overwhelming. Not more than 1 or 2 on a tab.

Just consider how much information processing will need to occur by the people interpreting these. You could also replace one multivariate plot several univariate and bivariate plots.

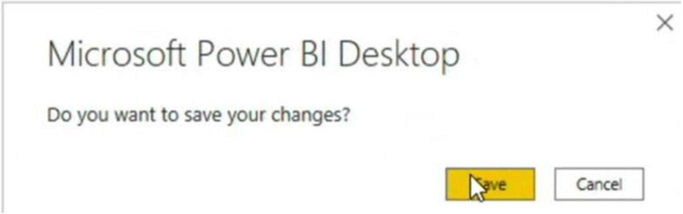
EDA 5 – Publishing a Power BI report

Go to the 'Home' tab and the Publish icon



Publish this report online in the Power BI service.

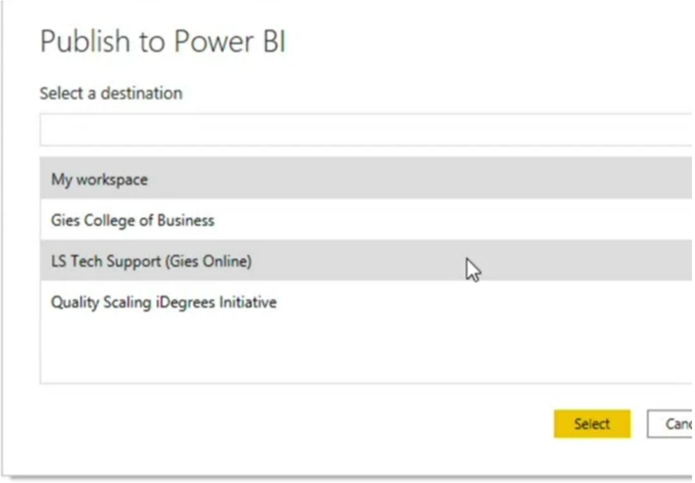
Then click 'Save'.



Microsoft Power BI Desktop

Do you want to save your changes?

Then you can choose which Workspace you want to publish to



Publish to Power BI

Select a destination

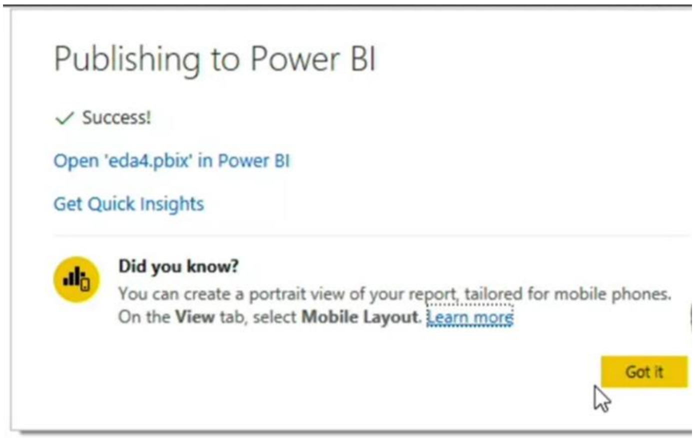
My workspace

Gies College of Business

LS Tech Support (Gies Online)

Quality Scaling iDegrees Initiative

And now it's published



Publishing to Power BI

✓ Success!

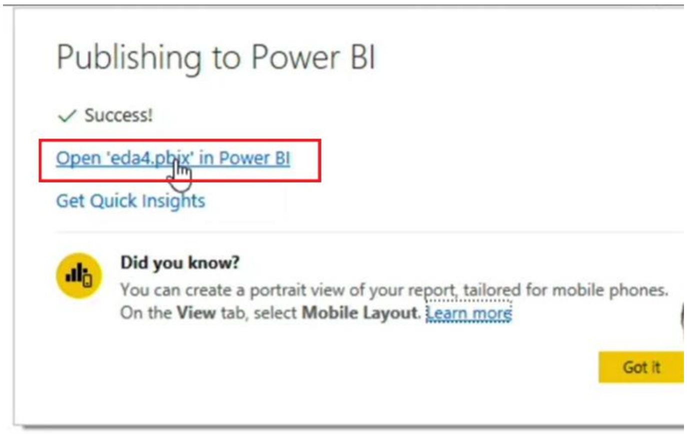
Open 'eda4.pbix' in Power BI

Get Quick Insights

Did you know?

You can create a portrait view of your report, tailored for mobile phones. On the **View** tab, select **Mobile Layout**. Learn more

Clicking on the link opens Power BI online



Publishing to Power BI

✓ Success!

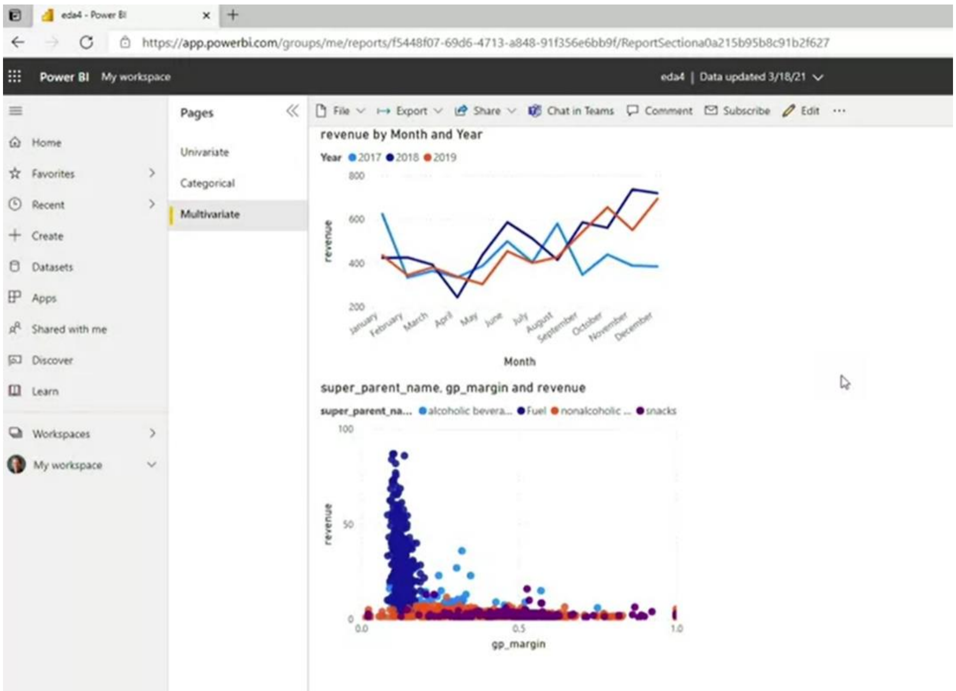
Open 'eda4.pbix' in Power BI

Get Quick Insights

Did you know?

You can create a portrait view of your report, tailored for mobile phones. On the **View** tab, select **Mobile Layout**. Learn more

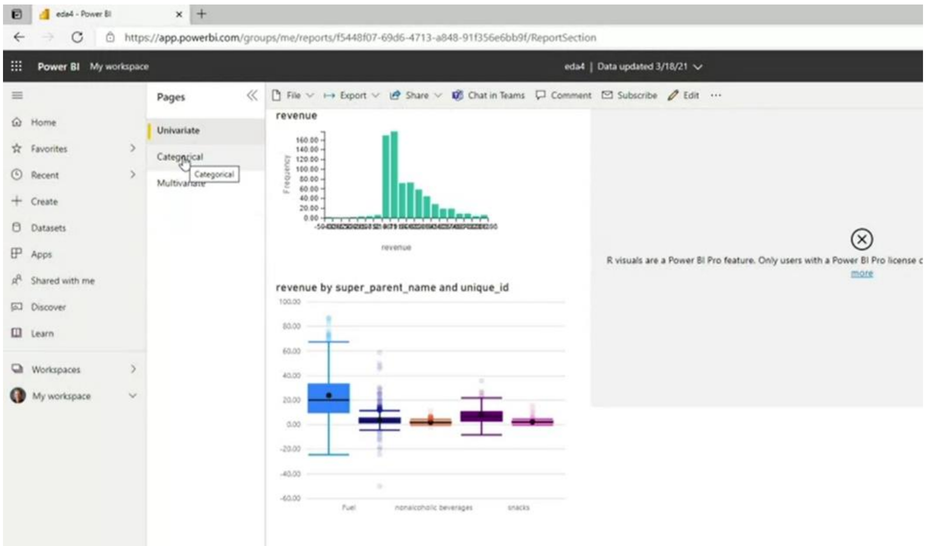
Here is the report in a personal workspace online



https://app.powerbi.com/groups/me/reports/f5448f07-69d6-4713-a848-91f356e6bb9f/ReportSectiona0a215b95b8c91b2f627

eda4 | Data updated 3/18/21

The different tabs can be accessed

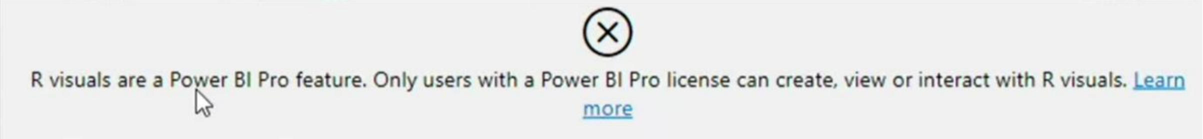


https://app.powerbi.com/groups/me/reports/f5448f07-69d6-4713-a848-91f356e6bb9f/ReportSection

eda4 | Data updated 3/18/21

R visuals are a Power BI Pro feature. Only users with a Power BI Pro license c

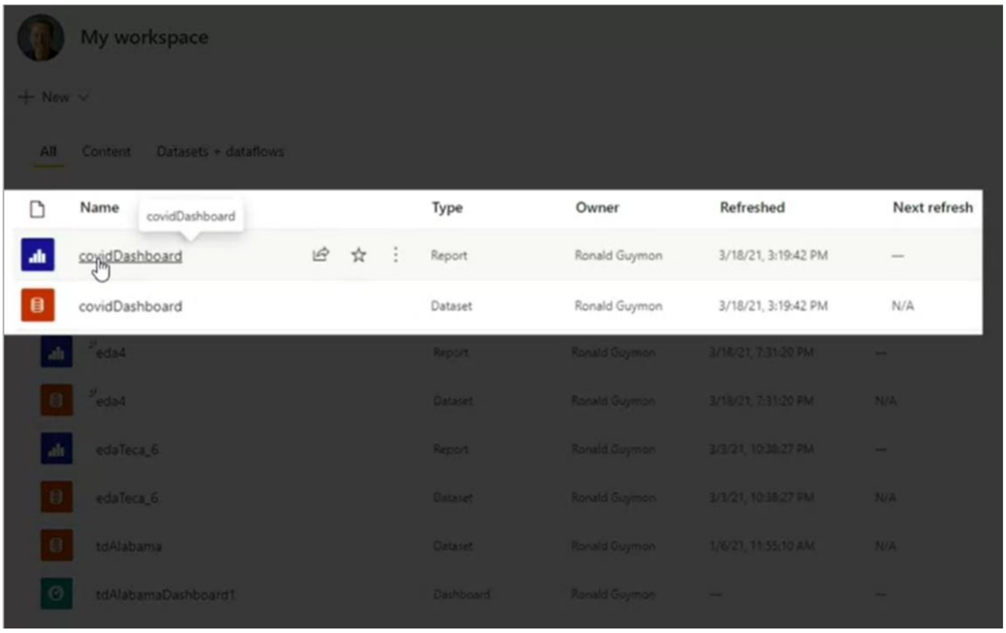
Note that R Visuals are a Power BI Pro feature and has to be paid for.



R visuals are a Power BI Pro feature. Only users with a Power BI Pro license can create, view or interact with R visuals. Learn more

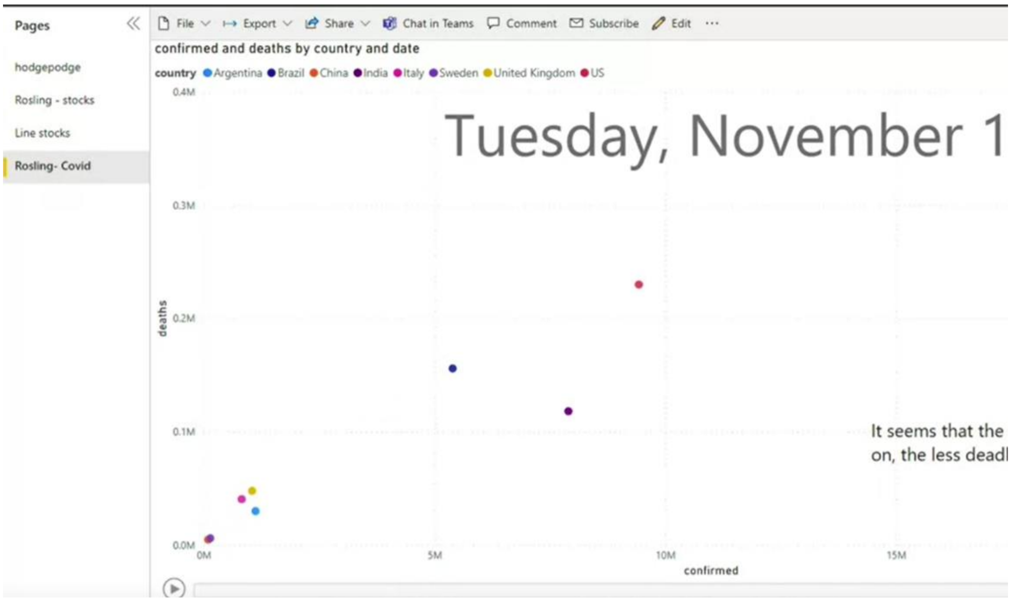
Users of the published report can apply filters and explore the data.

It is possible to view other Dashboards as well as the data that goes along with it.



Name	Type	Owner	Refreshed	Next refresh
covidDashboard	Report	Ronald Guymon	3/18/21, 3:19:42 PM	—
covidDashboard	Dataset	Ronald Guymon	3/18/21, 3:19:42 PM	N/A
eda4	Report	Ronald Guymon	3/18/21, 7:31:20 PM	—
eda4	Dataset	Ronald Guymon	3/18/21, 7:31:20 PM	N/A
edaTeca_6	Report	Ronald Guymon	3/3/21, 10:38:27 PM	—
edaTeca_6	Dataset	Ronald Guymon	3/3/21, 10:38:27 PM	N/A
tdAlabama	Dataset	Ronald Guymon	1/6/21, 11:55:10 AM	N/A
tdAlabamaDashboard1	Dashboard	Ronald Guymon	—	—

Here is a Hans-Rosling plot to show progression of Covid over time



It seems that the on, the less deadl