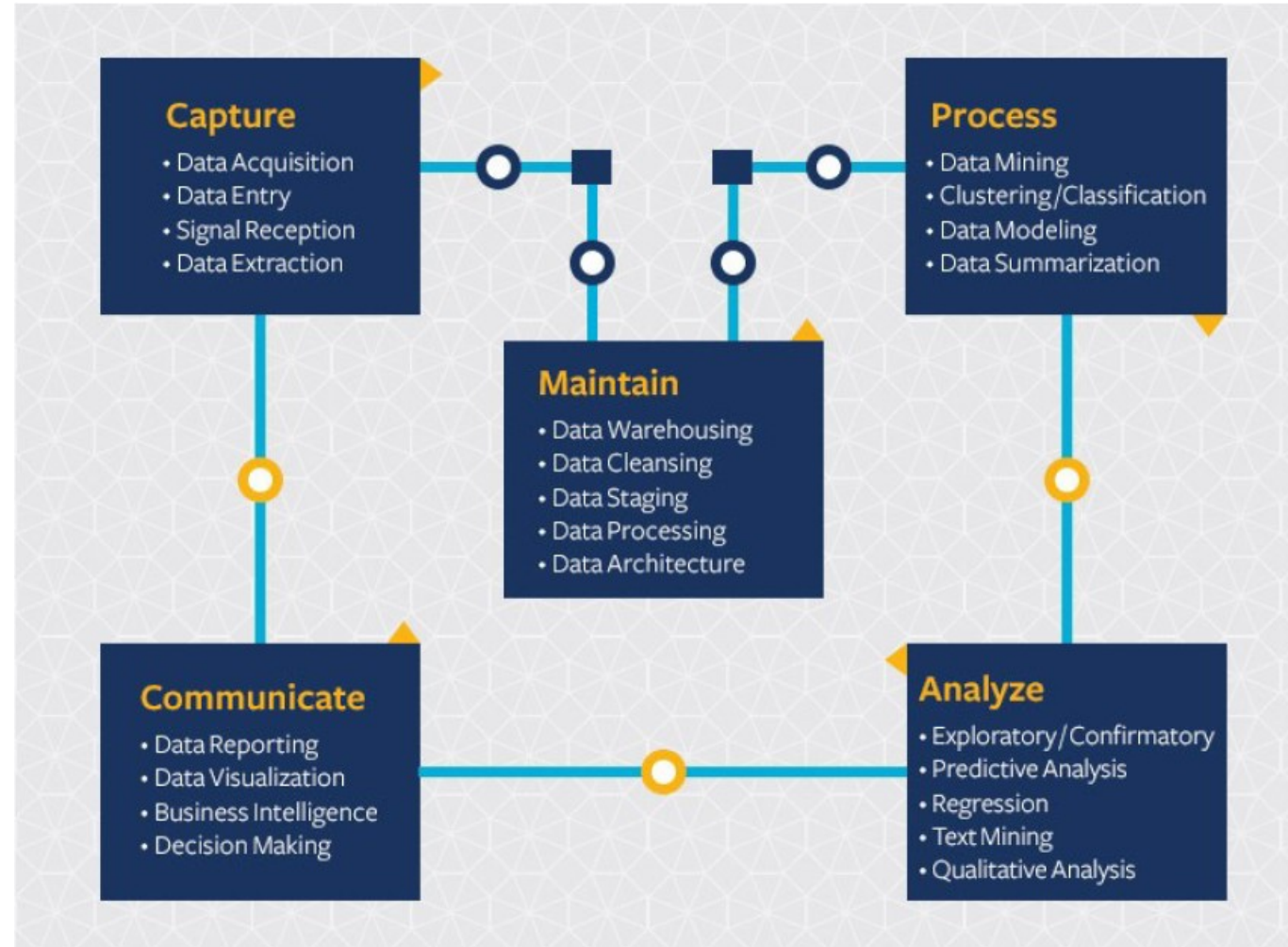

Data Engineering Introduction




The Data Revolution

- The convergence of a number of technologies and trends has led to a rapid change in the importance of data
 - Cloud Computing
 - The proliferation of mobile devices, IOT and other sources of data
 - Big Data
- Data Engineering is the art of building and maintaining systems to collect, clean, translate, organise and shape this data
- Ultimately the value of data is in being able to analyse and derive actionable insights. Decisions and actions may be made by humans or machines.

The Science Process

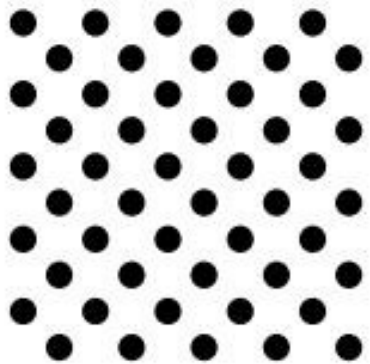




“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it...

The Challenges in Big Data: The four V's

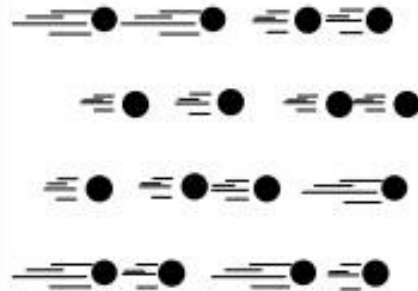
Volume



Data at Rest

Terabytes to exabytes of existing data to process

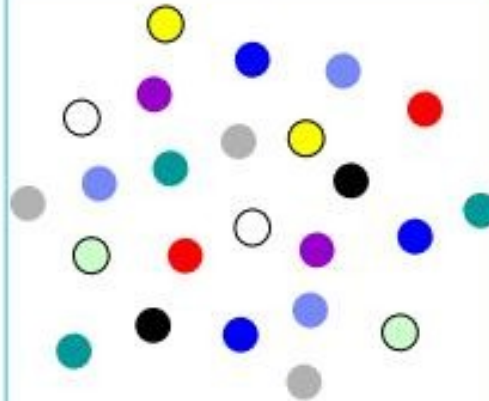
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

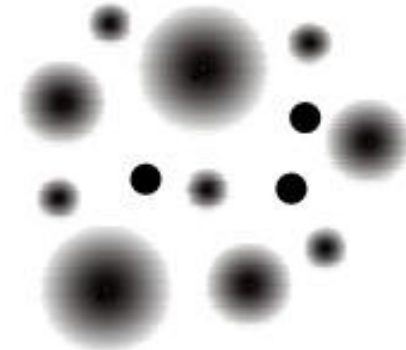
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

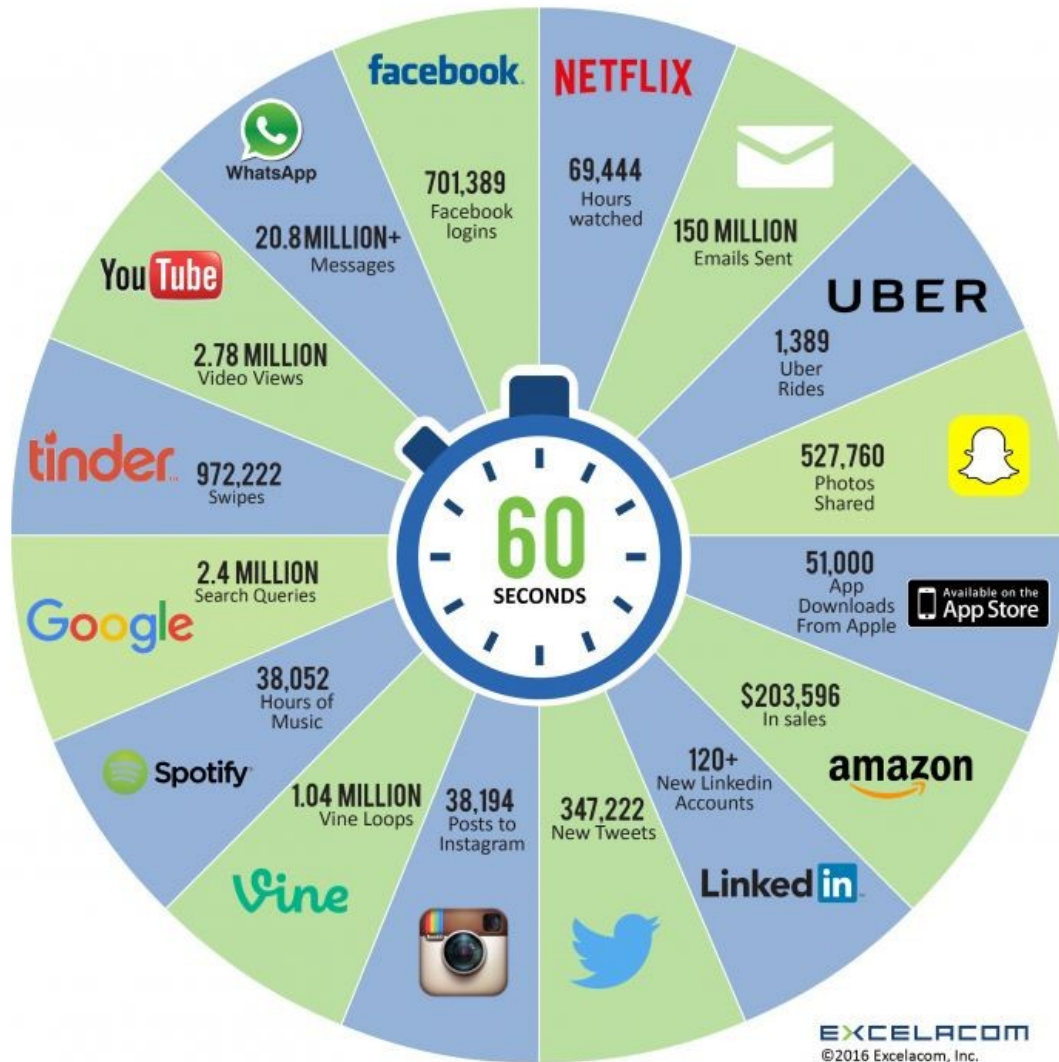
Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

2016 What happens in an INTERNET MINUTE?



The Scale of Big Data

90% Of today's data has been created in the last two years

Every day we create 2.5 quintillion bytes of data or enough to fill 10 million Blu-ray discs

Most companies in the US have over 100 terabytes (100,000 gigabytes) of data stored

40 zettabytes (40 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005, and the equivalent of 5,200 gigabytes of data for every man, woman and child on Earth

2019 *This Is What Happens In An Internet Minute*



Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000 ¹ bytes	1,000 bytes
megabyte (MB)	1000 ² bytes	1,000,000 bytes
gigabyte (GB)	1000 ³ bytes	1,000,000.000 bytes
terabyte (TB)	1000 ⁴ bytes	1,000,000,000,000 bytes
petabyte (PB)	1000 ⁵ bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000 ⁶ bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

FORTNITE

A background image from the game Fortnite showing several characters in a battle royale setting. In the foreground, a character with a large tomato head and a mustache is holding a large black assault rifle. In the background, three other characters are running and shooting. The scene is set in a hilly, mountainous landscape under a blue sky with clouds.

Realtime Ingestion

- Fortnite processes 92 million events a minute and sees its data grow 2 petabytes a month
- Up to 5000 kinesis shards

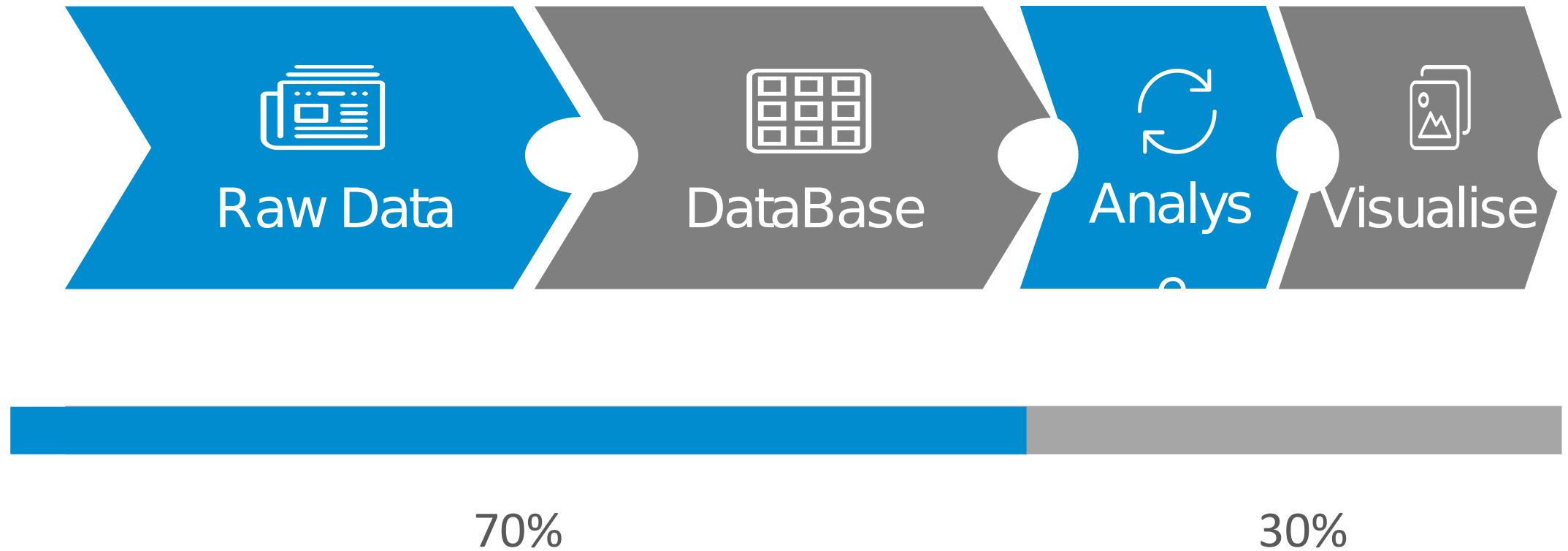
Data Warehouse

- 14 Petabytes
- 2 PB/month growth rate

"We use the data for everything from ARPU to game analysis and improvements"

<https://www.zdnet.com/article/how-fortnite-approaches-analytics-cloud-to-analyze-petabytes-of-game-data/>

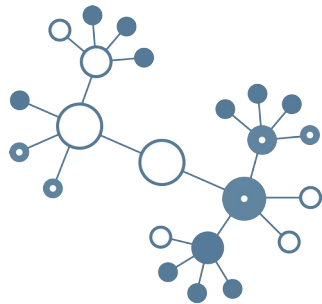
The Data Science Process



Data Visualization

What is Data Visualisation?

Regardless of industry or size, all types of businesses are using data visualisation to help make sense of their data



**1. Faster
decision making**



**2. Trend
identification**



**3. Interacting
with data**



**4. Data story-
telling**

What is Data Visualization

- A 2013 report by Aberdeen Group found that “at organizations that use visual discovery tools, 48 percent of BI users are able to find the information they need without the help of IT staff.” Without visual discovery, the rate drops to a mere 23 percent.
- *“A well-crafted, thoughtful visualization makes the light bulb go off. You just don’t get that with a spreadsheet.”* —Dana Zuber, Wells Fargo
- With visual discovery you enter a Cycle of Visual Analysis:
You get data, view the data, ask and answer questions, and repeat. Each time, your inquiry deepens along with your insights. You may drill down, drill up, or drill across. You may bring in new data. You may create view after view as your visualization speeds and extends your thinking.

Questions to Ask Yourself

- What is the context?
- How do I best convey my message?
- Who is viewing the data?
- How are they viewing the data?
- Is it easy to process the data?
 - What details do you need to present and what can you leave out
- Is it easy to see the conclusions?
- Am I using the right chart or table?

Are you using the Right Kind of Chart?

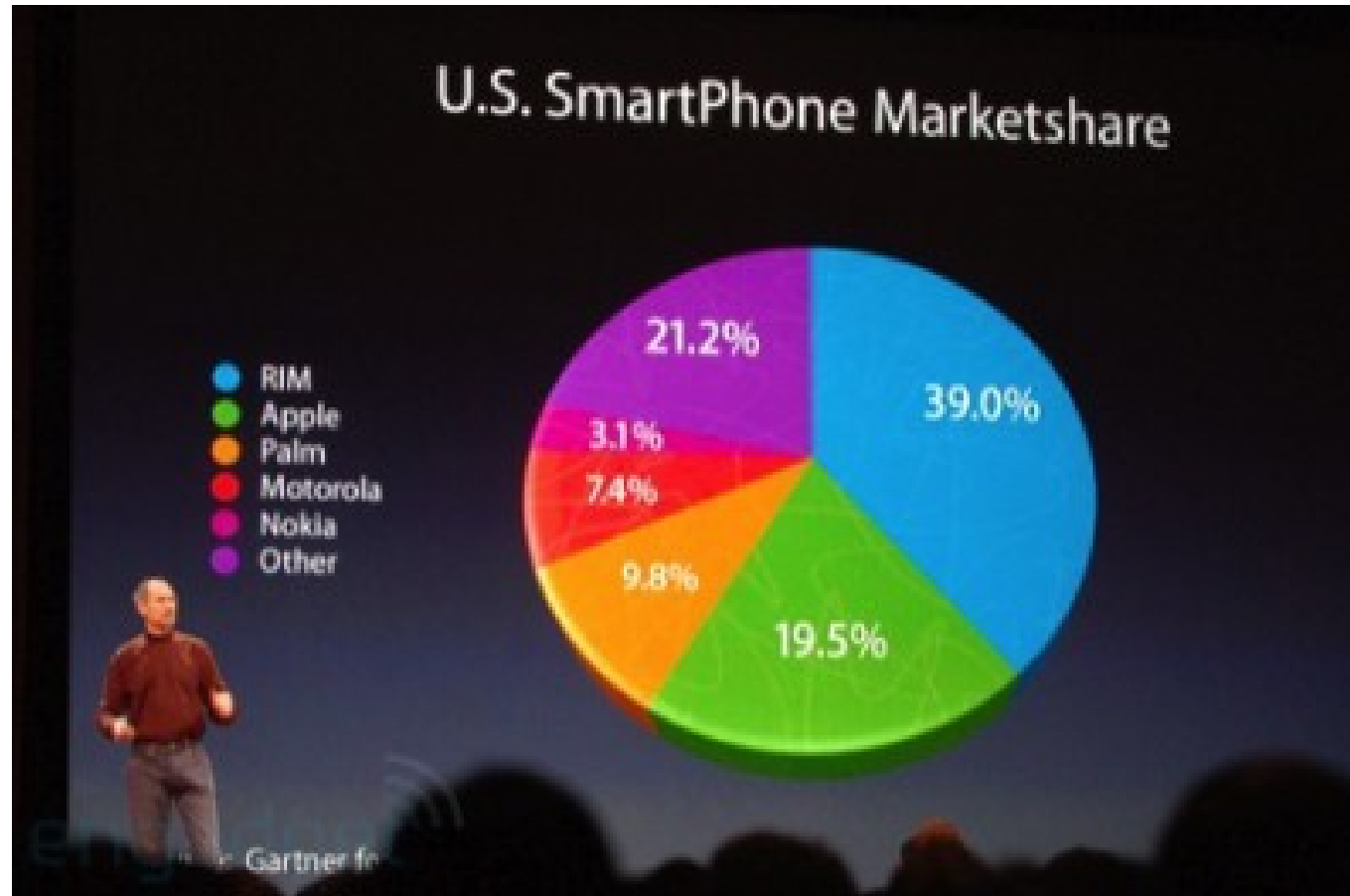
Do

1. Understand the business need
2. Keep your audience in mind
3. Add clear labels
4. Remove distracting chart elements

Do Not

1. Use a chart just because it looks good – if you have to explain what a chart is displaying, it's not working
2. Clutter the chart with unnecessary features
3. Add too much text
4. Use too many colours

What is wrong with this?



(presented by Steve Jobs at Engadget 2008 <http://www.engadget.com/2008/01/15/live-from-macworld-2008-steve-jobs-keynote/>)

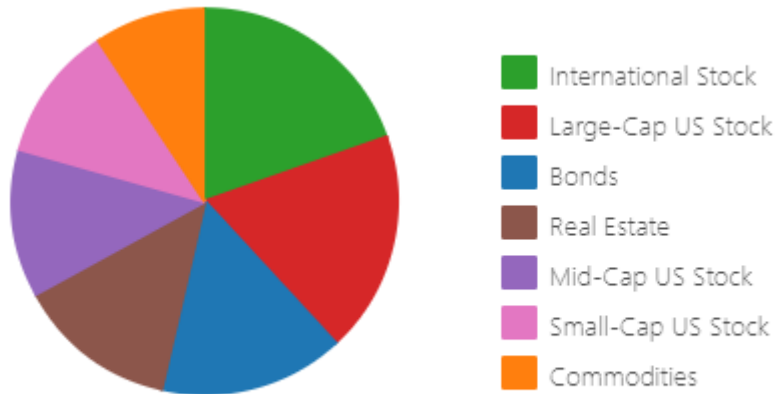
Visualisation References

- O' Reilly : Fundamentals of Data Visualization
 - <https://clauswilke.com/dataviz/>
- Visual Vocabulary Lookup
 - <https://ft-interactive.github.io/visual-vocabulary/>
- Data Visualization Article (towardsdatascience.com)
 - <https://towardsdatascience.com/the-art-and-science-of-data-visualization-6f9d706d673e>

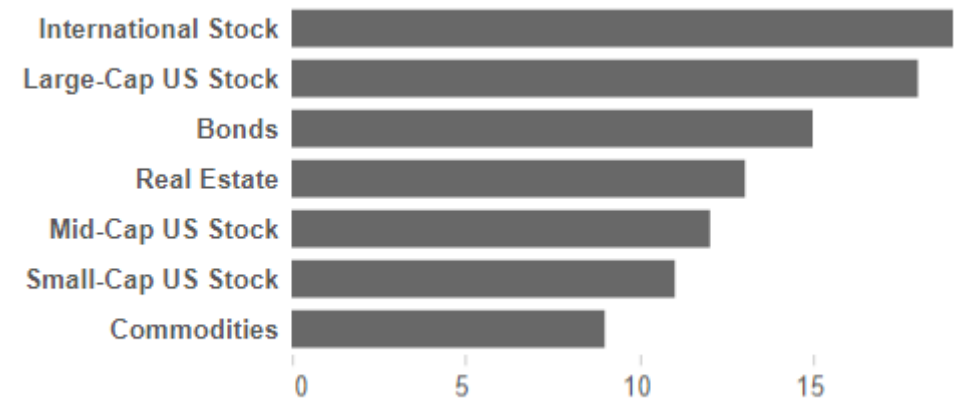
Data Visualizations: **Quiz**

Round 1 of 10

Which graph makes it easier to determine which investment has greater market share?



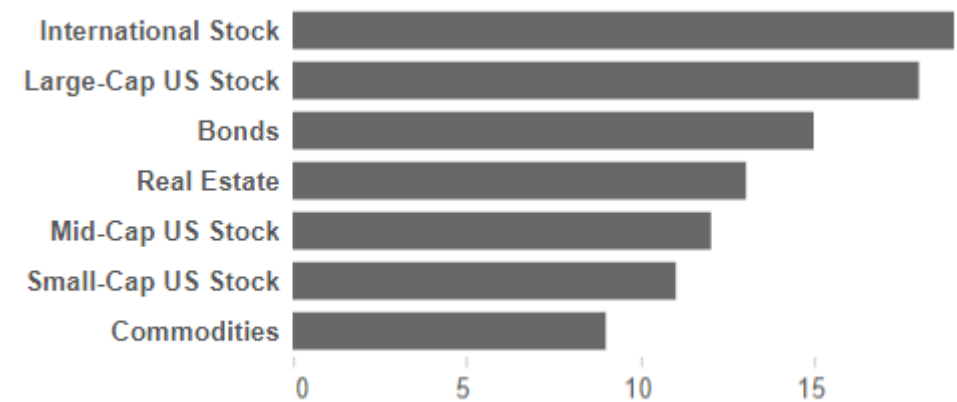
Option A



Option B

Round 1 of 10 - result

It is much easier to compare and clearly see the best performing stock



Option B

Round 2 of 10

Which of these two tables is easier to read?

Region	Revenue	% of Total Revenue	Expenses	Profit
Western US	\$58,753,092.00	15.98%	\$25,725,650.00	\$33,027,442.00
Europe	\$86,671,628.00	23.58%	\$48,859,689.00	\$37,811,939.00
Eastern US	\$61,165,954.00	16.64%	\$34,294,598.00	\$26,871,356.00
Canada	\$83,650,773.00	22.76%	\$29,785,749.00	\$53,865,024.00
Asia	\$77,324,523.00	21.04%	\$43,587,987.00	\$33,736,536.00
Total	#####	100.00%	#####	#####

Option A

Region	Revenue	% of Total Revenue	Expenses	Profit
Western ..	58,753,092	16.0%	25,725,650	33,027,442
Europe	86,671,628	23.6%	48,859,689	37,811,939
Eastern US	61,165,954	16.6%	34,294,598	26,871,356
Canada	83,650,773	22.8%	29,785,749	53,865,024
Asia	77,324,523	21.0%	43,587,987	33,736,536
Total	367,565,970	100.0%	182,253,673	185,312,297

Option B

Round 2 of 10 - result

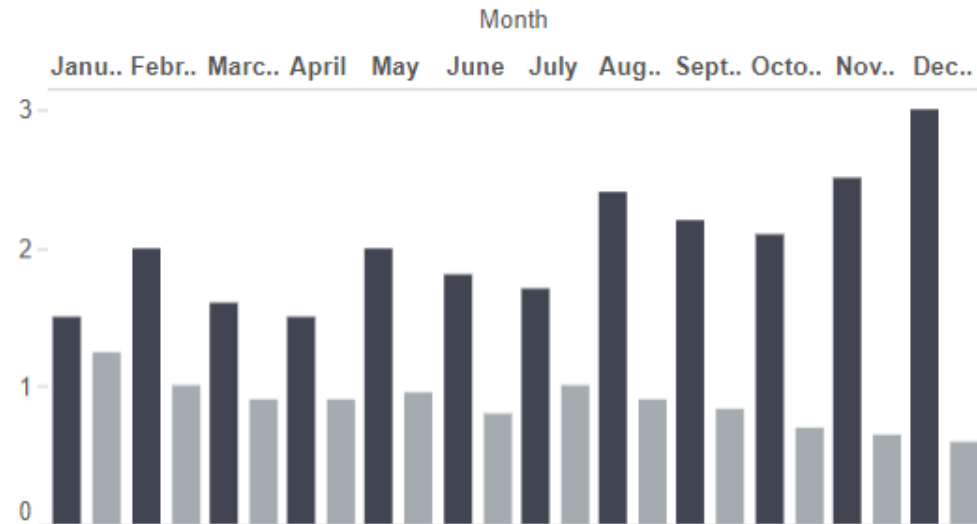
Some of the data in Option A could not be seen and the darker rows were not highlighting key information

Region	Revenue	% of Total Revenue	Expenses	Profit
Western ..	58,753,092	16.0%	25,725,650	33,027,442
Europe	86,671,628	23.6%	48,859,689	37,811,939
Eastern US	61,165,954	16.6%	34,294,598	26,871,356
Canada	83,650,773	22.8%	29,785,749	53,865,024
Asia	77,324,523	21.0%	43,587,987	33,736,536
Total	367,565,970	100.0%	182,253,673	185,312,297

Option B

Round 3 of 10

Which graph helps you identify a trend most easily?



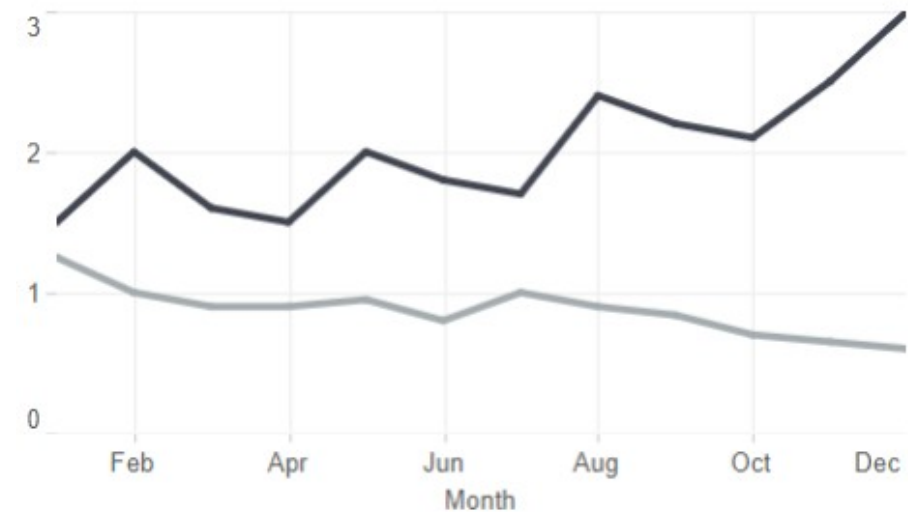
Option A



Option B

Round 3 of 10 - result

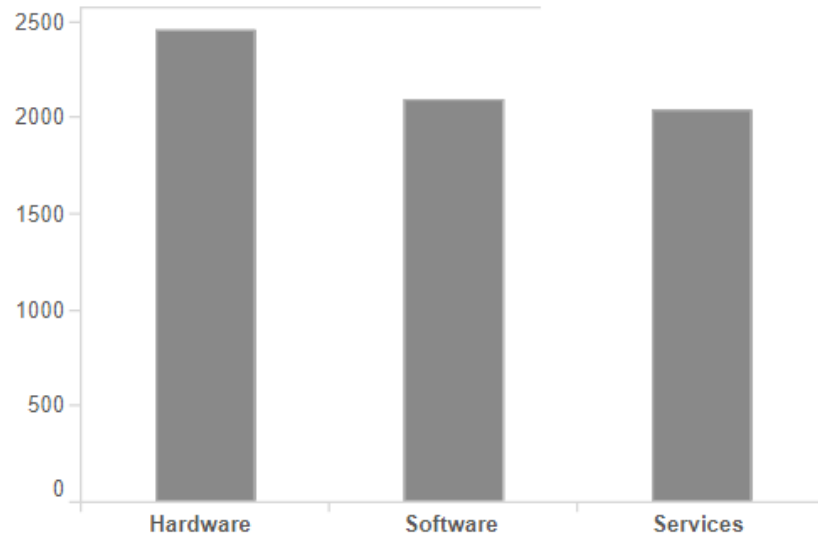
Without needing to read the detail in the chart you can easily see one category is increasing and one is decreasing



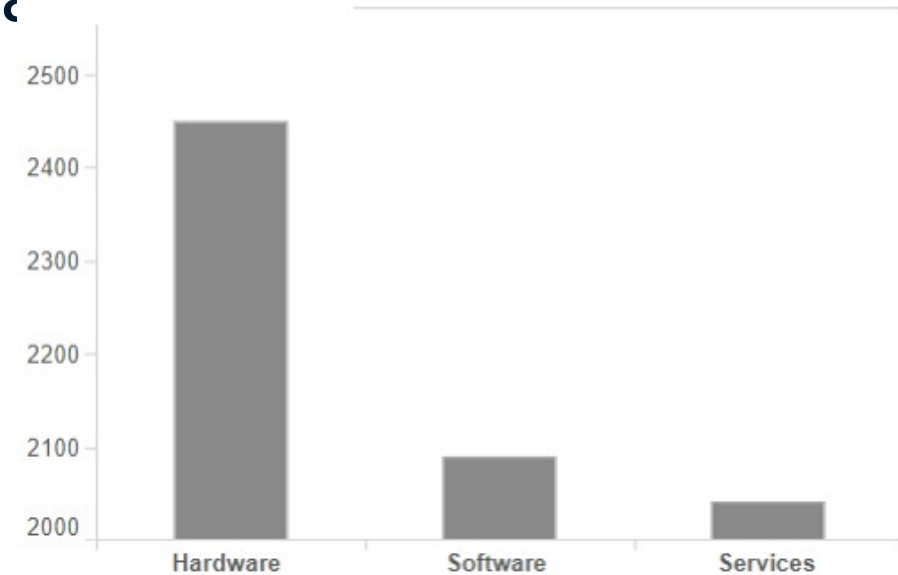
Option B

Round 4 of 10

Both graphs display the same data, which chart do you believe displays the data best without bias?



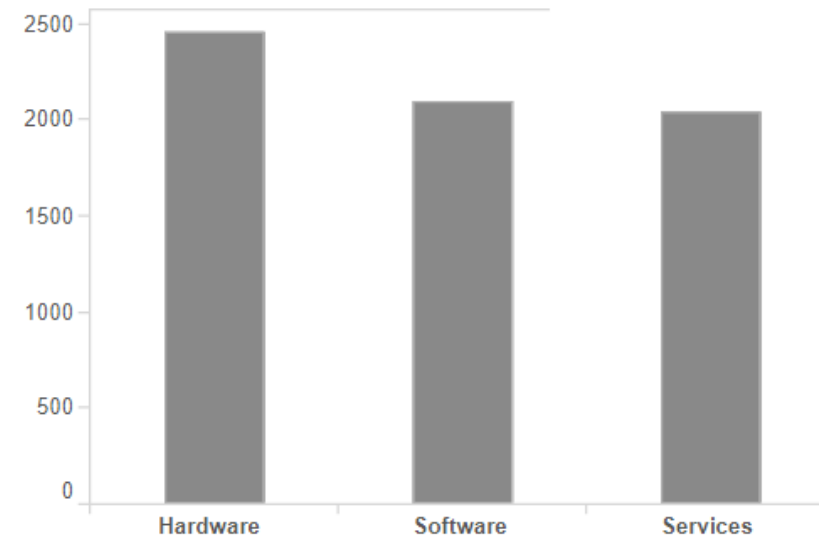
Option A



Option B

Round 4 of 10 - result

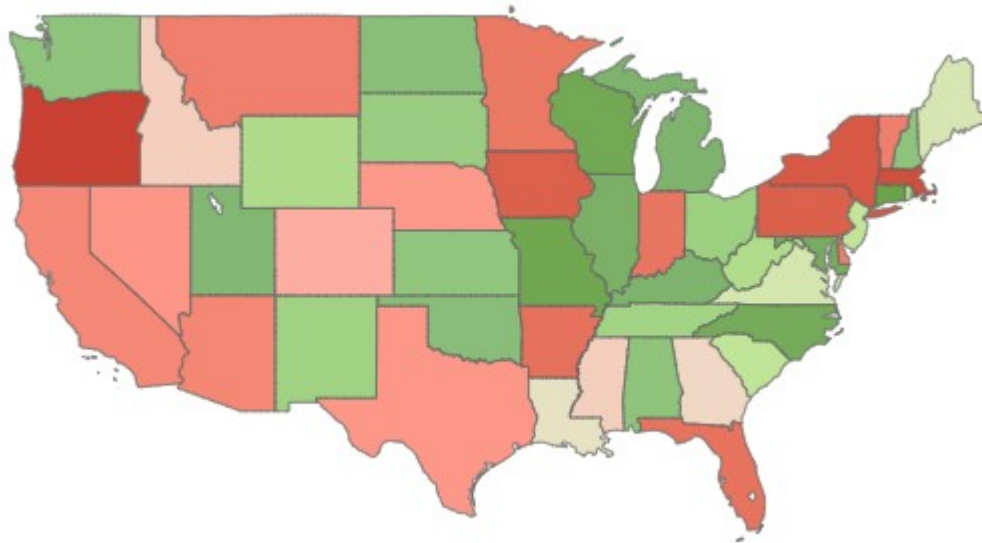
User tend to assume an axis starts at zero. By changing the scale of the axis, Option B skews the data, making the difference between the bars seem much larger



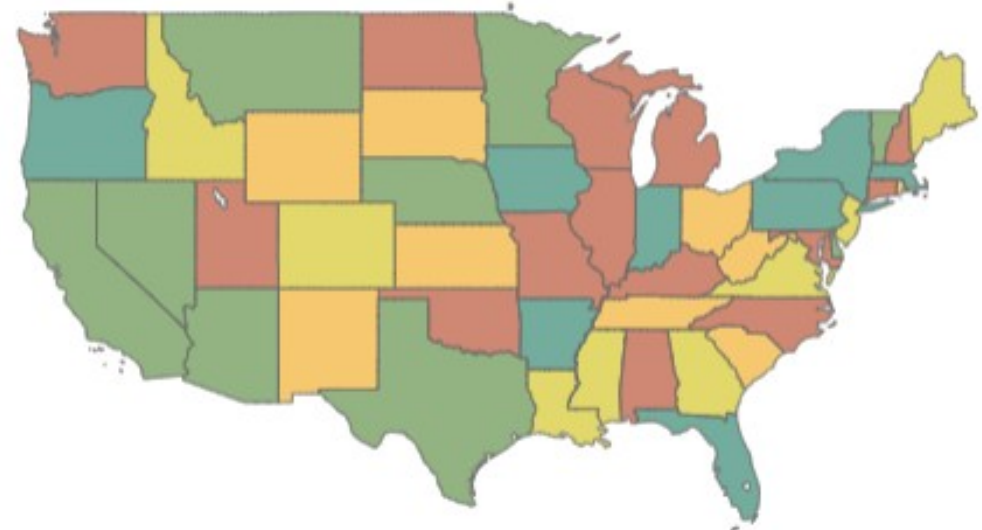
Option A

Round 5 of 10

Which graph makes it easier to identify the states with positive values?



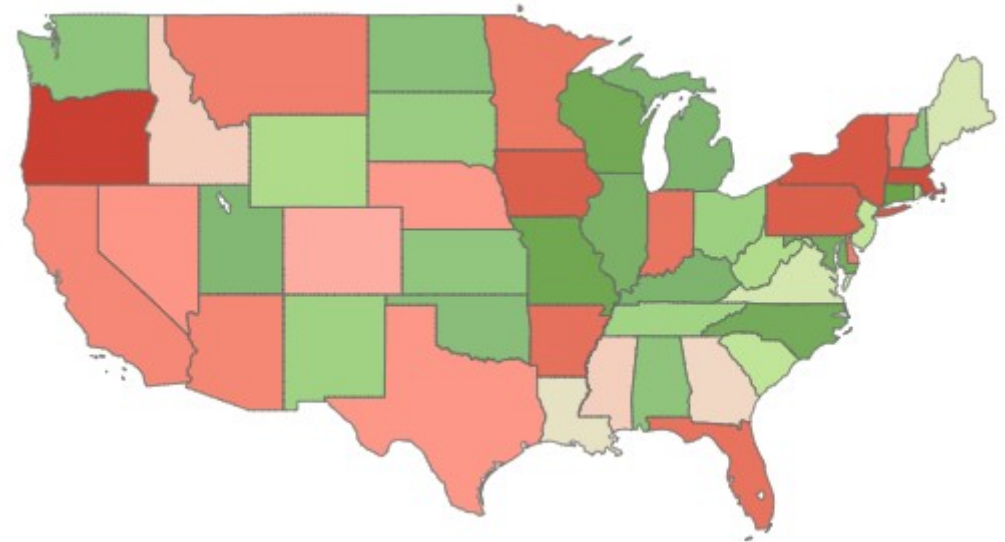
Option A



Option B

Round 5 of 10 - result

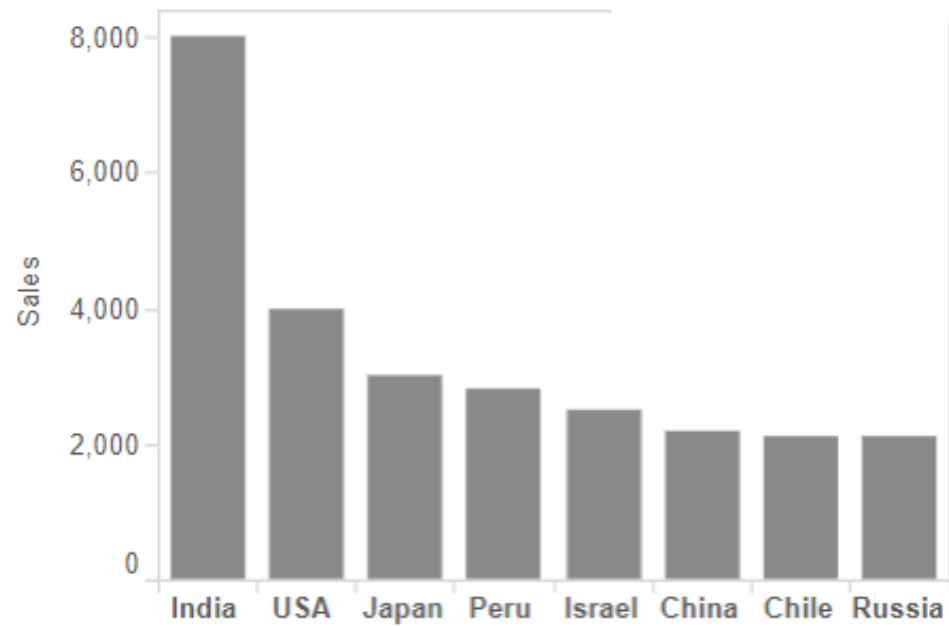
Option A's simpler colour scheme makes identifying positive and negative values much easier



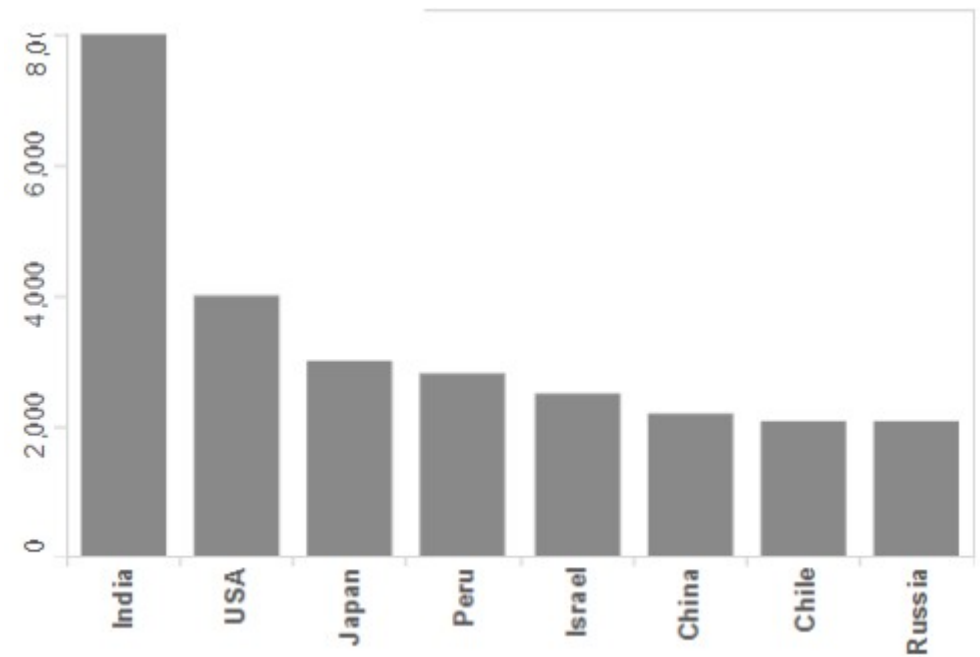
Option A

Round 6 of 10

Which graph is easier to read?



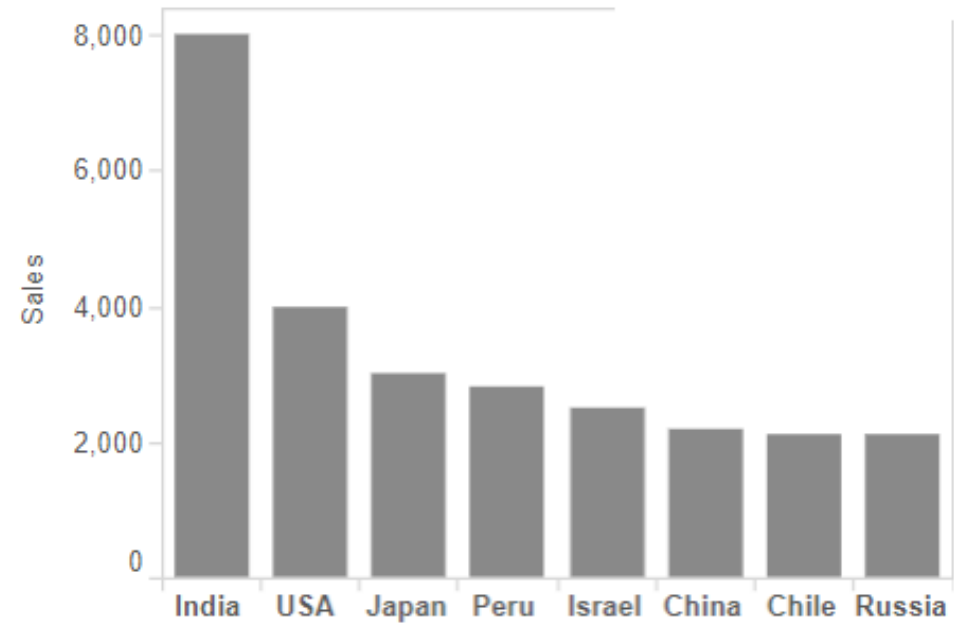
Option A



Option B

Round 6 of 10 - result

Horizontal data labels
are much easier to read



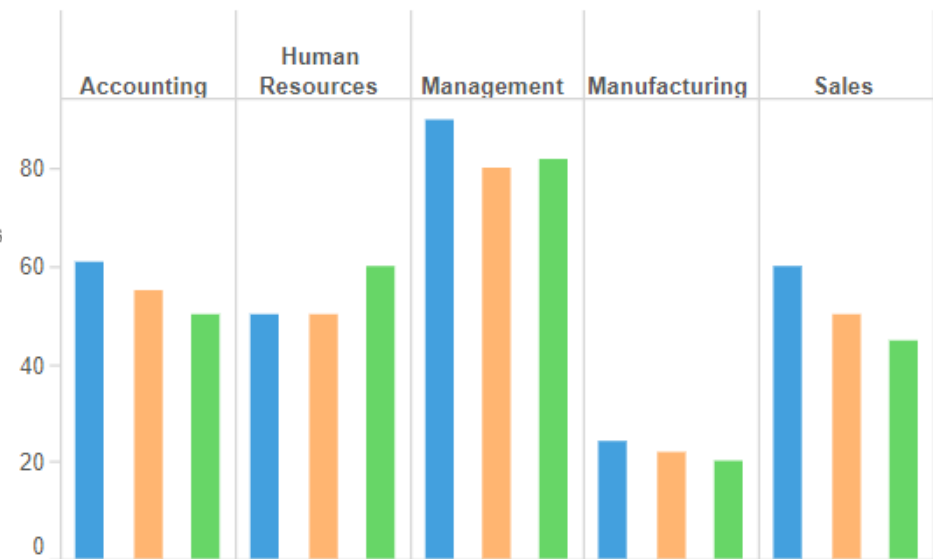
Option A

Round 7 of 10

Which graph is easier to view?



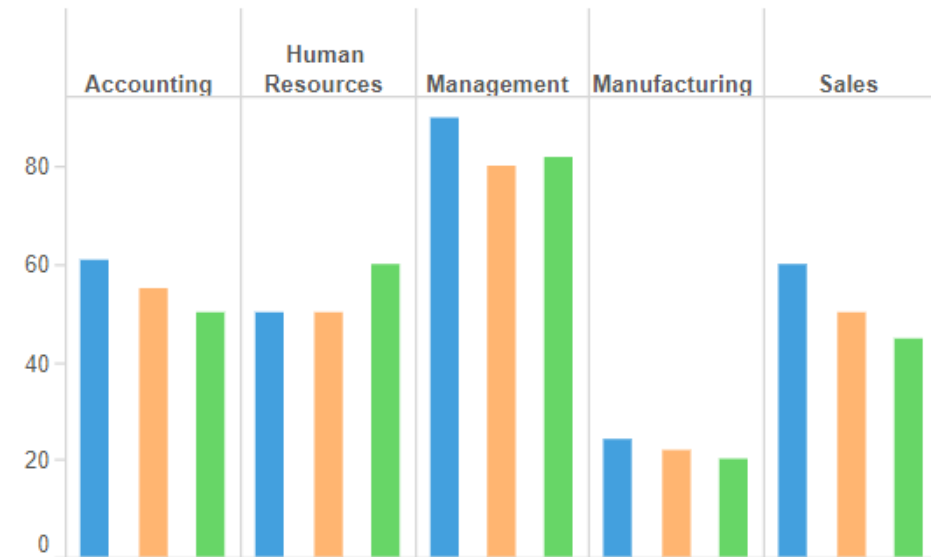
Option A



Option B

Round 7 of 10 - result

Background colour can draw a user's attention but can also make the chart more difficult to read – keep it simple



Option B

Round 8 of 10

Which table helps you identify areas of poor performance the fastest?

Region	Revenue	Expenses	Profit
Asia	\$55,850	\$24,770	\$31,080
Canada	\$68,379	\$26,228	\$42,151
Europe	\$54,300	\$46,645	\$7,655
USA	\$74,411	\$17,573	\$56,838

Option A

Region	Revenue	Expenses	Profit
Asia	55,850	24,770	31,080
Canada	68,379	26,228	42,151
Europe	54,300	46,645	7,655
USA	74,411	17,573	56,838

Option B

Round 8 of 10 - result

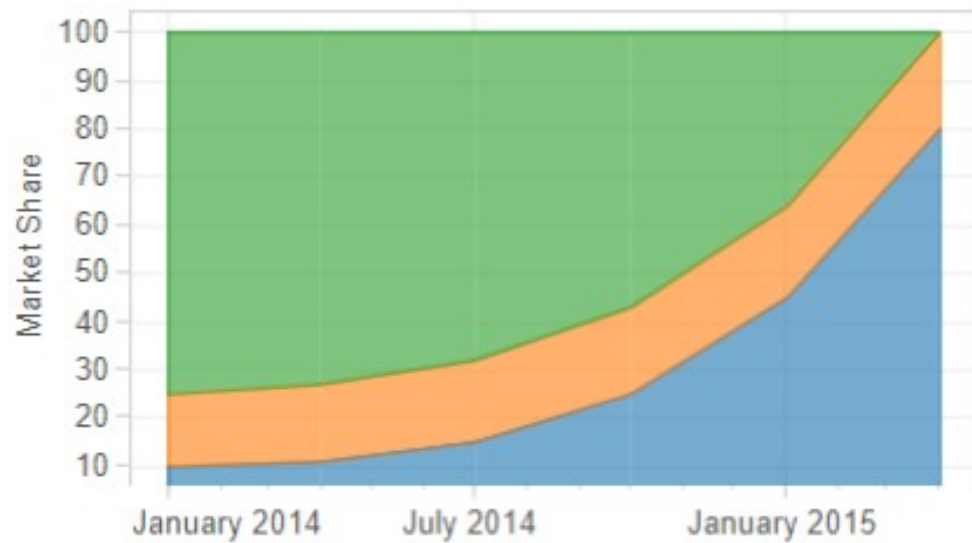
The additional colours are not needed. In option A you can very easily identify poor performance

Region	Revenue	Expenses	Profit
Asia	\$55,850	\$24,770	\$31,080
Canada	\$68,379	\$26,228	\$42,151
Europe	\$54,300	\$46,645	\$7,655
USA	\$74,411	\$17,573	\$56,838

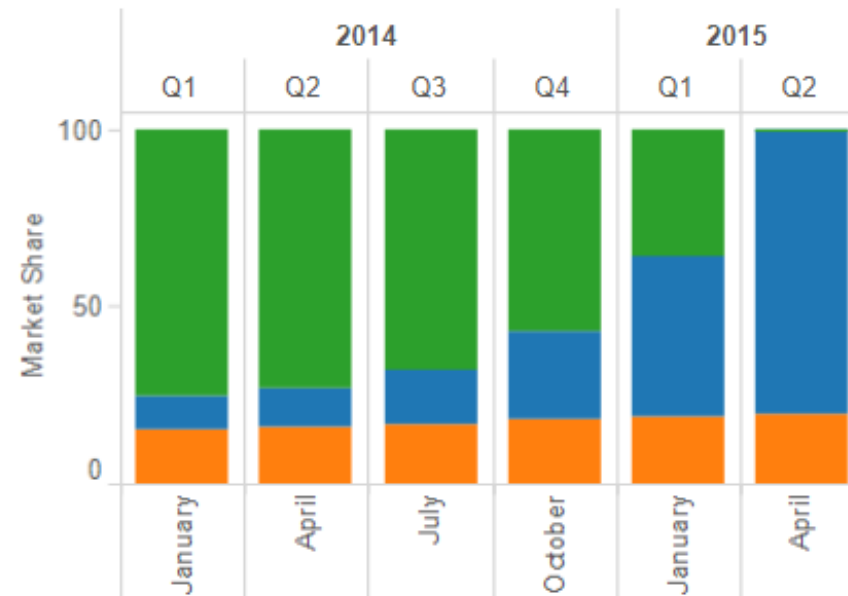
Option A

Round 9 of 10

Which graph gives a clearer picture of relative change in values?



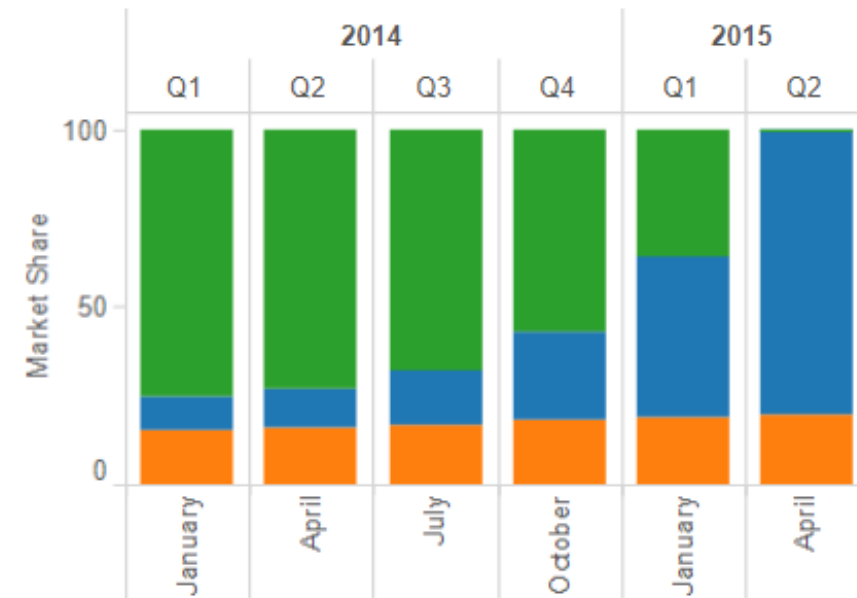
Option A



Option B

Round 9 of 10 - result

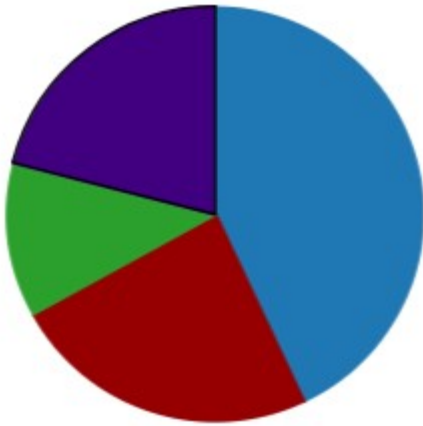
It is easier to compare the proportions of each category



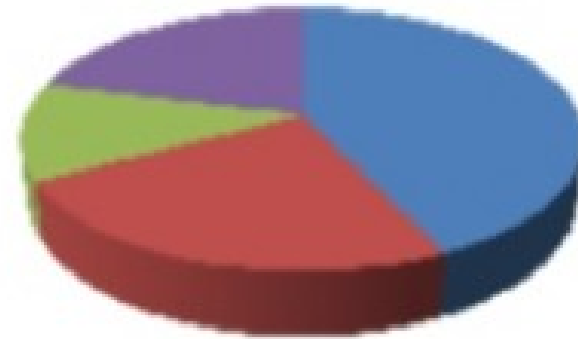
Option B

Round 10 of 10

Which graph is easier to read?



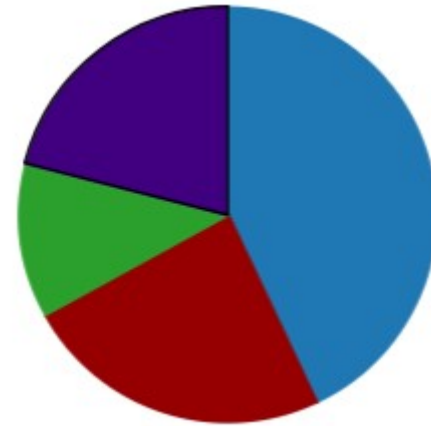
Option A



Option B

Round 10 of 10 - result

3D pie charts makes it much more difficult for the human eye to determine proportion sizes

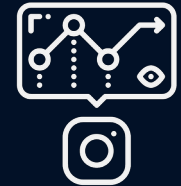


Option A

Thankyou

DATA VISUALISATION WITH POWER BI

- POWER BI ESSENTIALS



-
- WHAT IS POWER BI?
 - WHY CHOOSE POWER BI?
 - WHAT CAN YOU DO WITH IT?

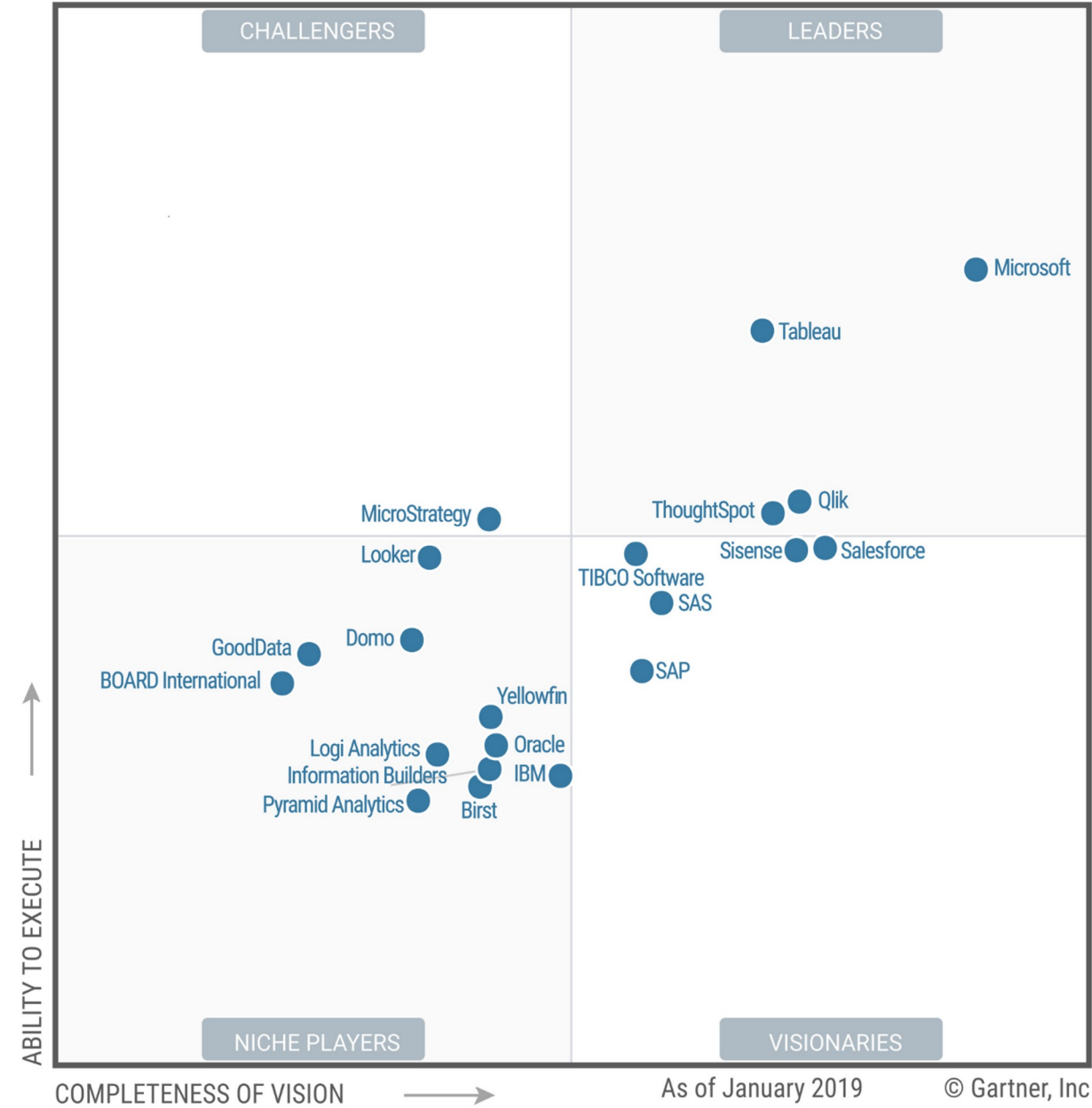
POWER BI – THE TOP 3 QUESTIONS

WHAT IS POWER BI?

- Business Intelligence Visualisation Tool from Microsoft
- Competes with other data visualisation products such as Tableau and Qlik
- Easily import from many data sources
- Data preparation, modelling & visualisation through graphics and reports



Figure 1. Magic Quadrant for Analytics and Business Intelligence Platforms



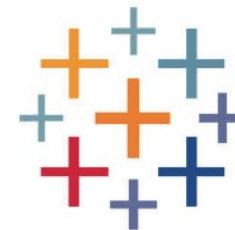
Source: Gartner (February 2019)

WHY CHOOSE POWER BI?

- According to Gartner it is the leading Business Intelligence Tool.
- Over 97% of Fortune 500 companies use Power BI.
- Some of the main advantages include cost & integration with Office365
- Microsoft are throwing vast resources at PowerBI – new features arrive every month.



Power BI



WHAT CAN YOU DO WITH IT?



HR ANALYTICS



A full end-to-end analytics and reporting solution – built on Microsoft Power BI – to help you discover useful HR insights for strategic decision making.

It enables HR professionals to make data-driven decisions to attract, manage, and retain employees, which improves ROI, and helps leaders make decisions to create better work environments and maximize employee productivity. It has a major impact on the bottom-line when used effectively.

You can see your data from any source such as Workday, ADP, Oracle HCM, ...

We can help you build and customize these reports, or you can get the **Power BI source file (PBIX)** with instructions to do it yourself (DIY).



1. Summary Dashboard



6. Salary Analysis



11. Termination Analysis



2. Diversity



7. Departments



12. Turnover Analysis



3. Historical & Trends



8. Performance Rating



13. Monthly Analysis



4. Employees Overview



9. Training

CONTACT US TO GET THIS PACKAGE



5. Employee Details



10. Absenteeism

version 1.2
contact info@agile-analytics.com.au



[Click here to open example](#)

What Can Be
Created With
Power BI?

Take a look at the examples here to get some context
of what can be achieved with Power BI

<https://community.powerbi.com/t5/Data-Stories-Gallery/bd-p/DataStoriesGallery>

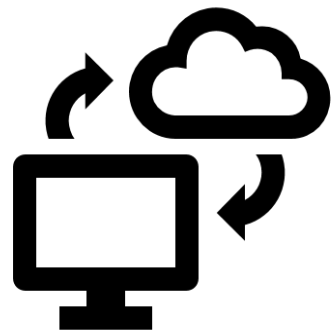
POWER BI CORE ELEMENTS

DESKTOP VS SERVICE



- **PowerBI Desktop**

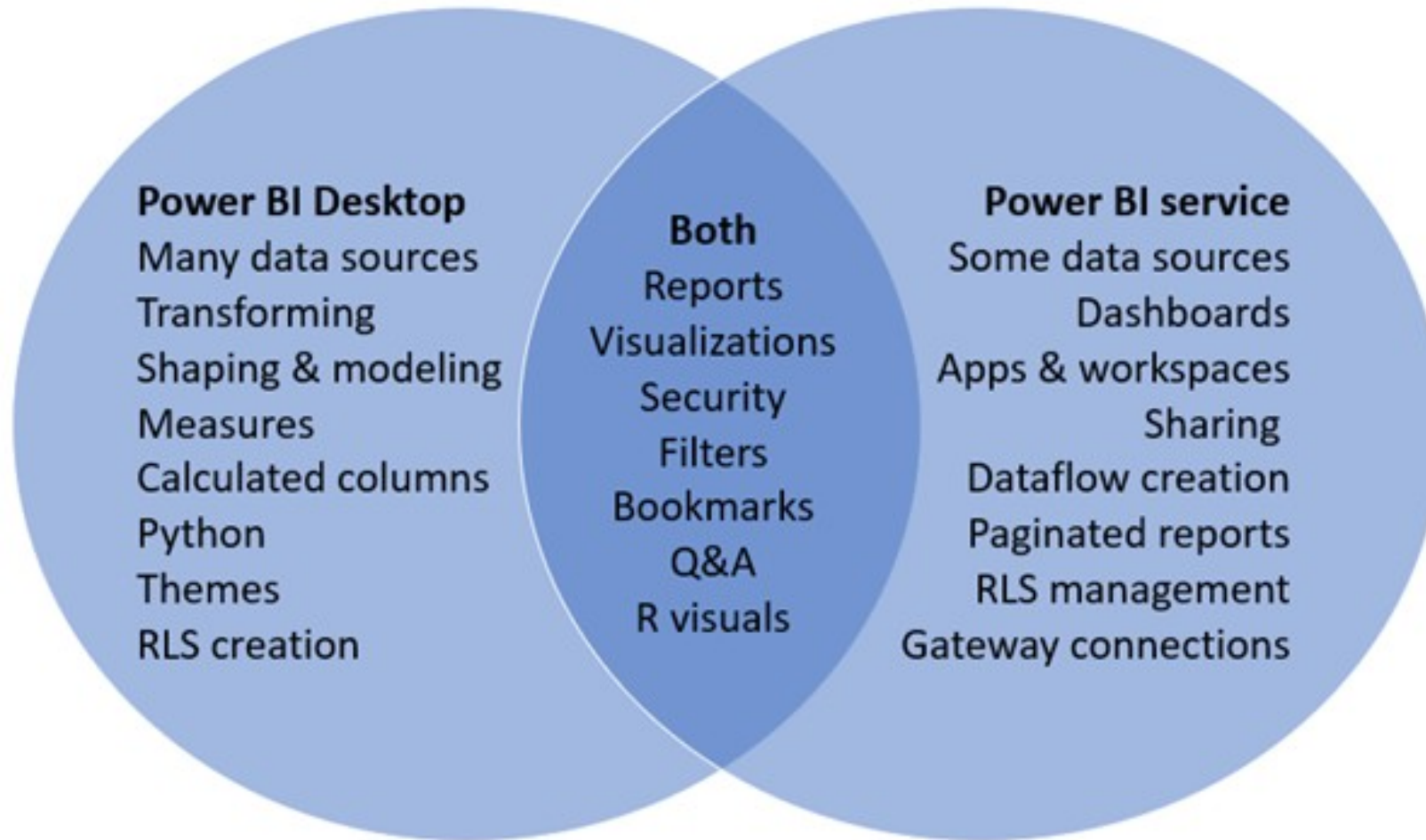
- Usually used to create reports
- Import, shape, visualise data
- Export Data & Static Visualisations



- **PowerBI Service** – This is where a license is needed!

- Usually used for sharing of reports on the web
- Basic report creation & editing features
- Cloud based or on premise

POWER BI DESKTOP VS SERVICE



EXCEL VS POWER BI

- Excel was built for a world where
 - Datasets were small < 1M rows
 - Real-time information wasn't needed
 - Collaboration wasn't as important
 - Data came from a single source, had a single format, stored in a single location
- In many ways these tools complement each other
- PowerBI is more geared towards **communicating** insights



EXCEL VS POWER BI

- Pros:
 - Report viewers can consume raw data for themselves and draw their own conclusions
- Cons:
 - Report viewers can consume raw data for themselves and draw their own conclusions



STATIC VS INTERACTIVE REPORTS

- **PowerBI Interactivity**
 - Can import live data feeds
 - Easy to combine data from multiple sources
 - New Data can be automatically incorporated into the report at set intervals e.g. monthly
 - Interactive exploration with slicers and filters

STATIC VS INTERACTIVE REPORTS

PROS & CONS OF STATIC REPORTS

- **PROS:**

- It's easier to control the narrative
- Don't tempt readers to arrive at incorrect conclusions

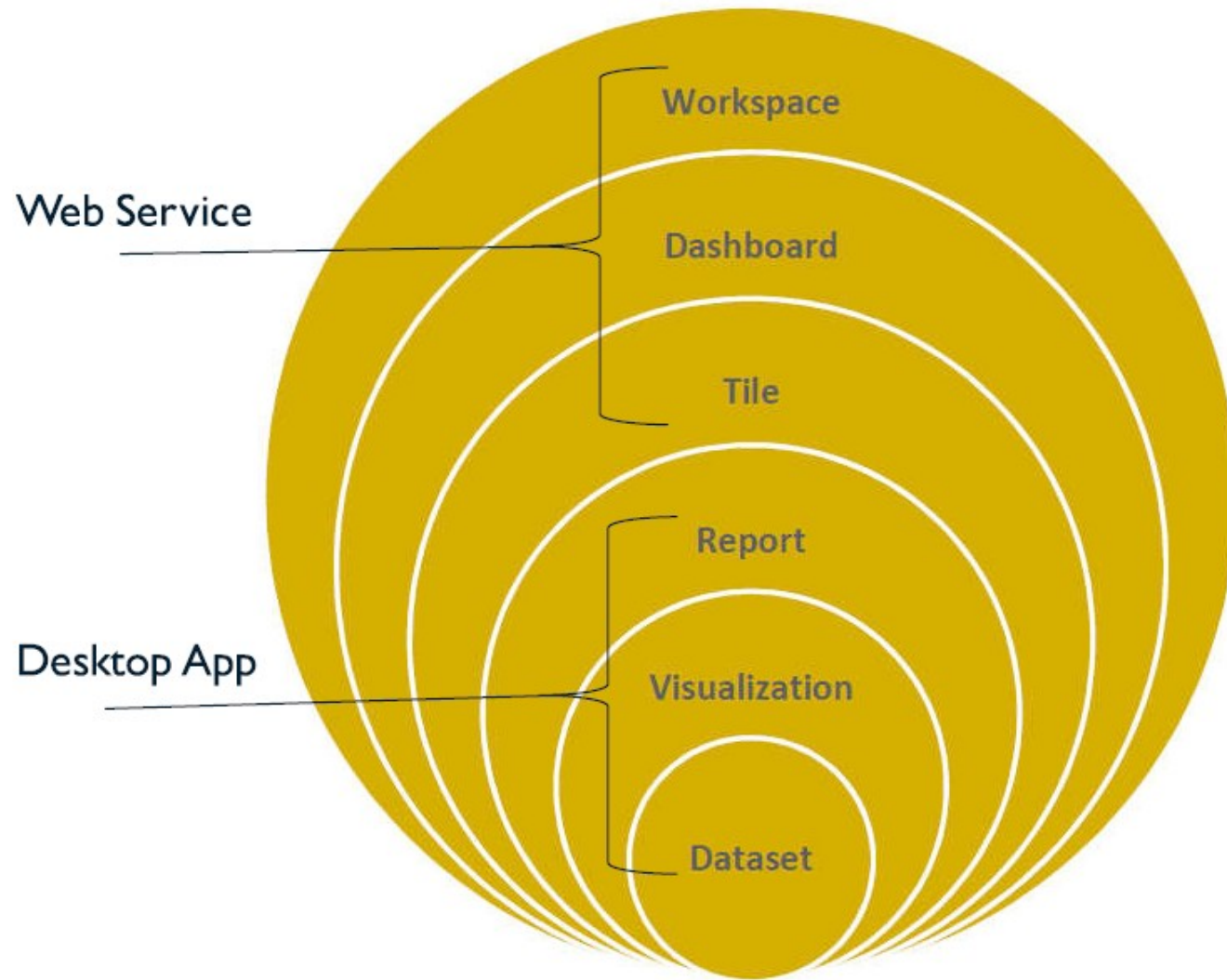
- **CONS:**

- The quantity of information conveyed can be limited
- With large amounts of data it can be difficult to display at the readers preferred resolution
- Data must be manually updated

STATIC VS INTERACTIVE REPORTS

HYBRID APPROACH

- **Interactive** reports are only for analysts, who want ALL of the information
- **Static** reports are exported by analysts as required for publishing e.g. To Word, Powerpoint, Excel



POWER BI TERMINOLOGY CLARIFICATION

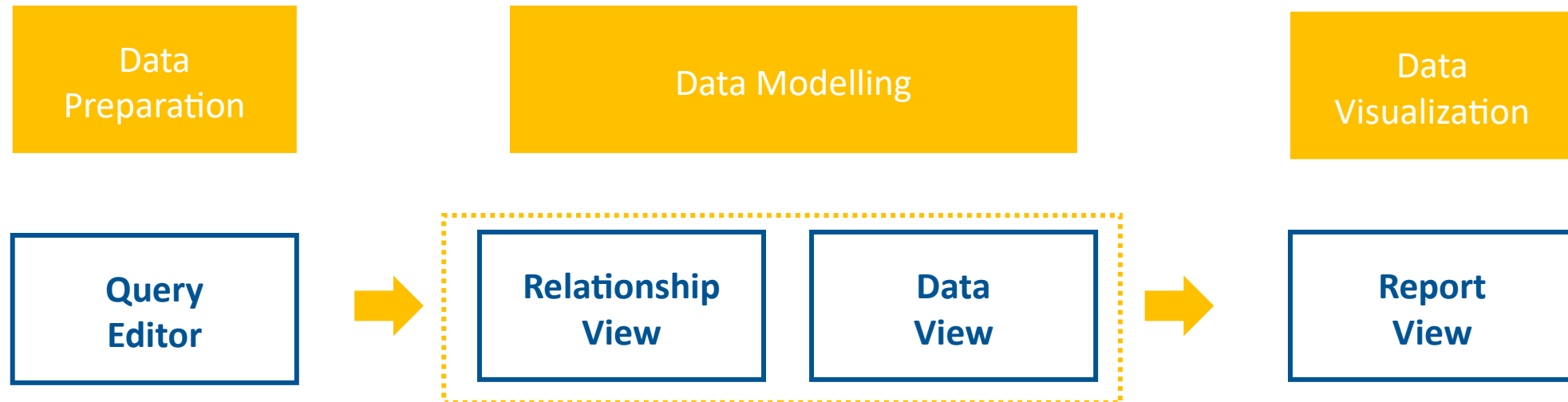
Report

- Multi-perspective view into a dataset with one or more visuals that represent different findings and insights from that dataset
- Based on a single data set
- Interact with the report through
 - Filtering
 - Slicing
 - Exporting
- Changing the report does not affect the underlying data

Data Views in Power Bi

- **Report/Visualization** view – where you use queries you create to build compelling visualizations, arranged as you want them to appear, and with multiple pages, that you can share with others
- **Data** view – see the data in your report in data model format, where you can add measures, create new columns, and manage relationships
- **Relationships** view – get a graphical representation of the relationships that have been established in your data model, and manage or modify them as needed.

Workflow of Power BI





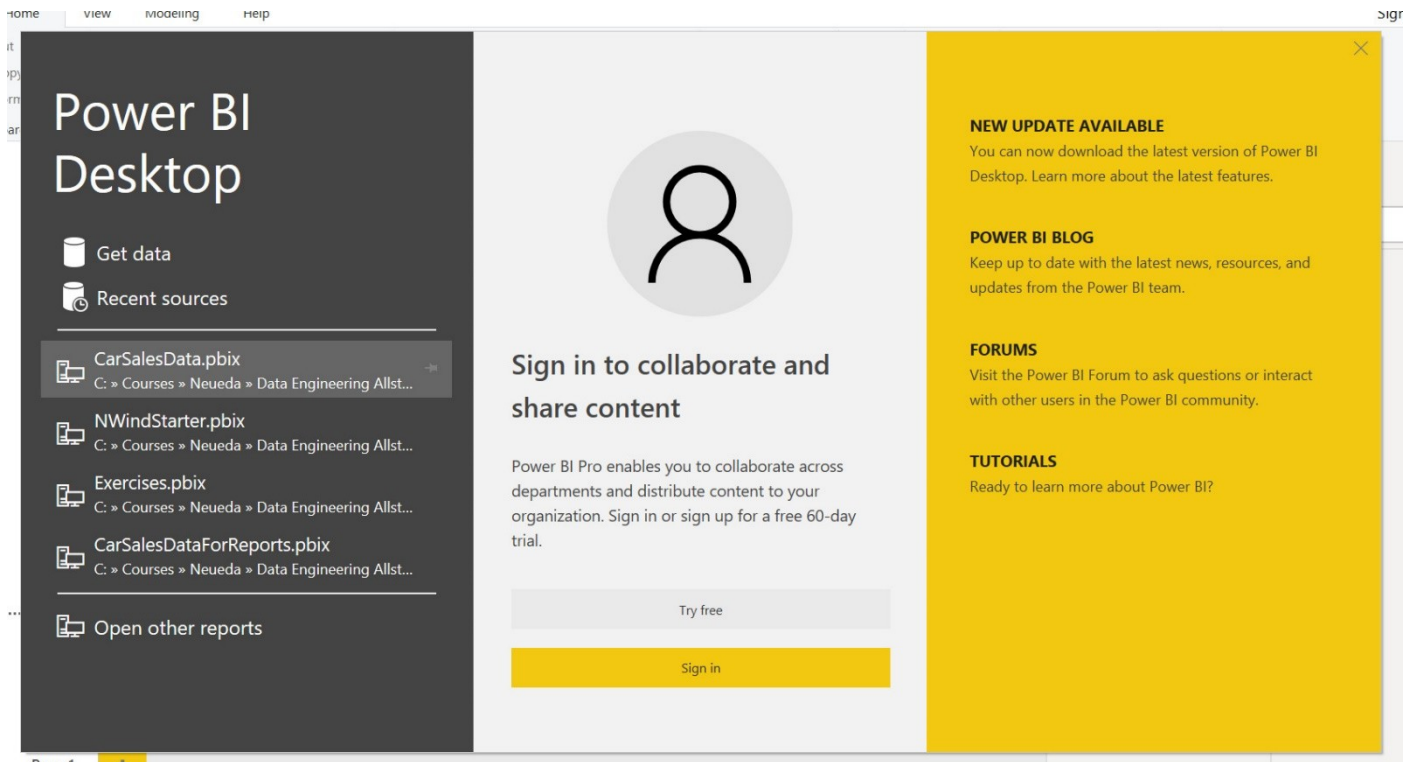
LET'S CREATE OUR FIRST REPORT!



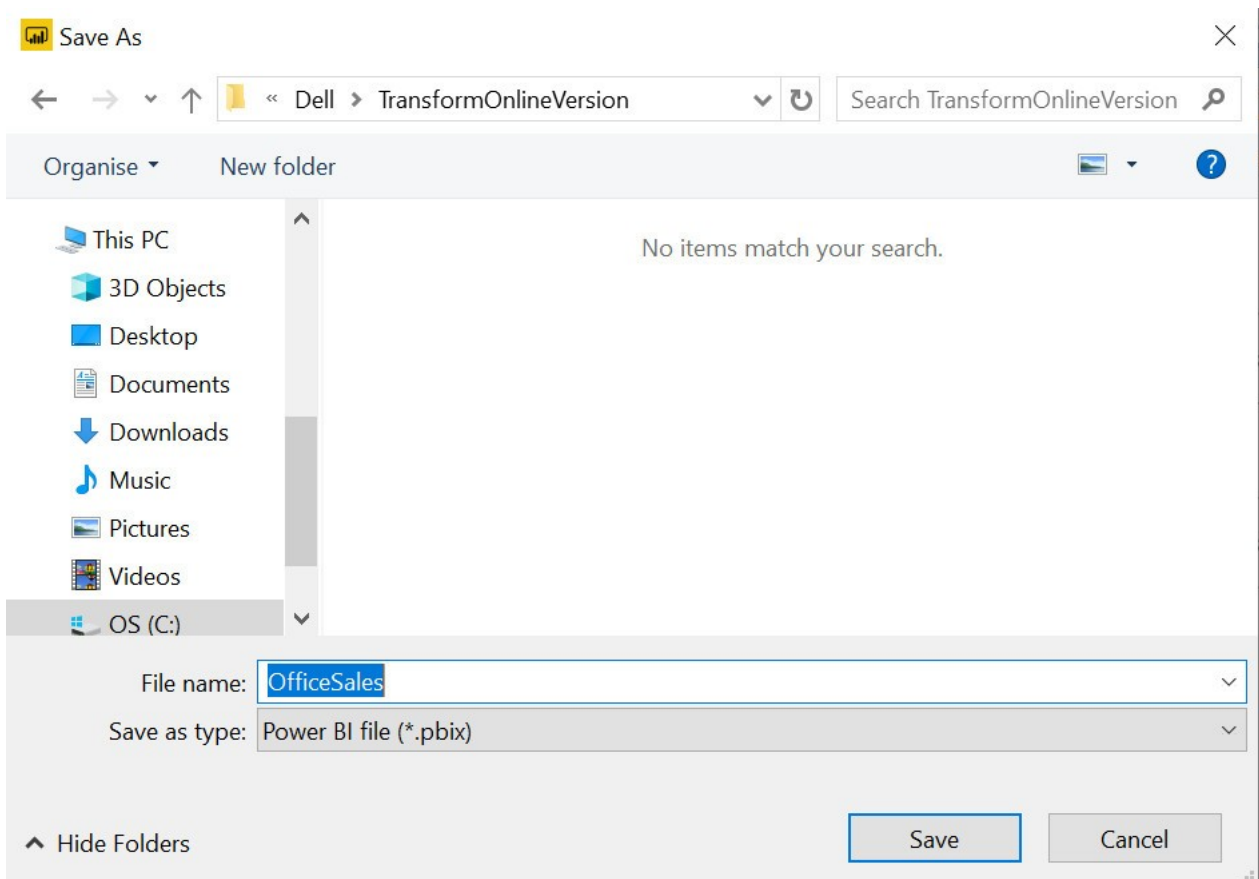
QUESTIONS

Using Power BI Desktop – Data Model and Visualizations

1. Open the Power BI Desktop Application.
2. You can close the splash screen that is displayed.

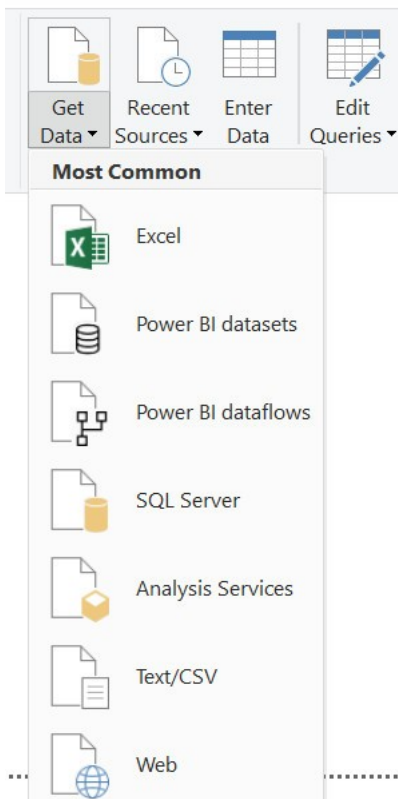


3. Like Microsoft Office Applications (Excel, Word etc) Power BI Desktop has a ribbon, the Home tab is displayed when the application is started. Choose **File – Save As** and save the file as OfficeSales in a suitable location on your computer.

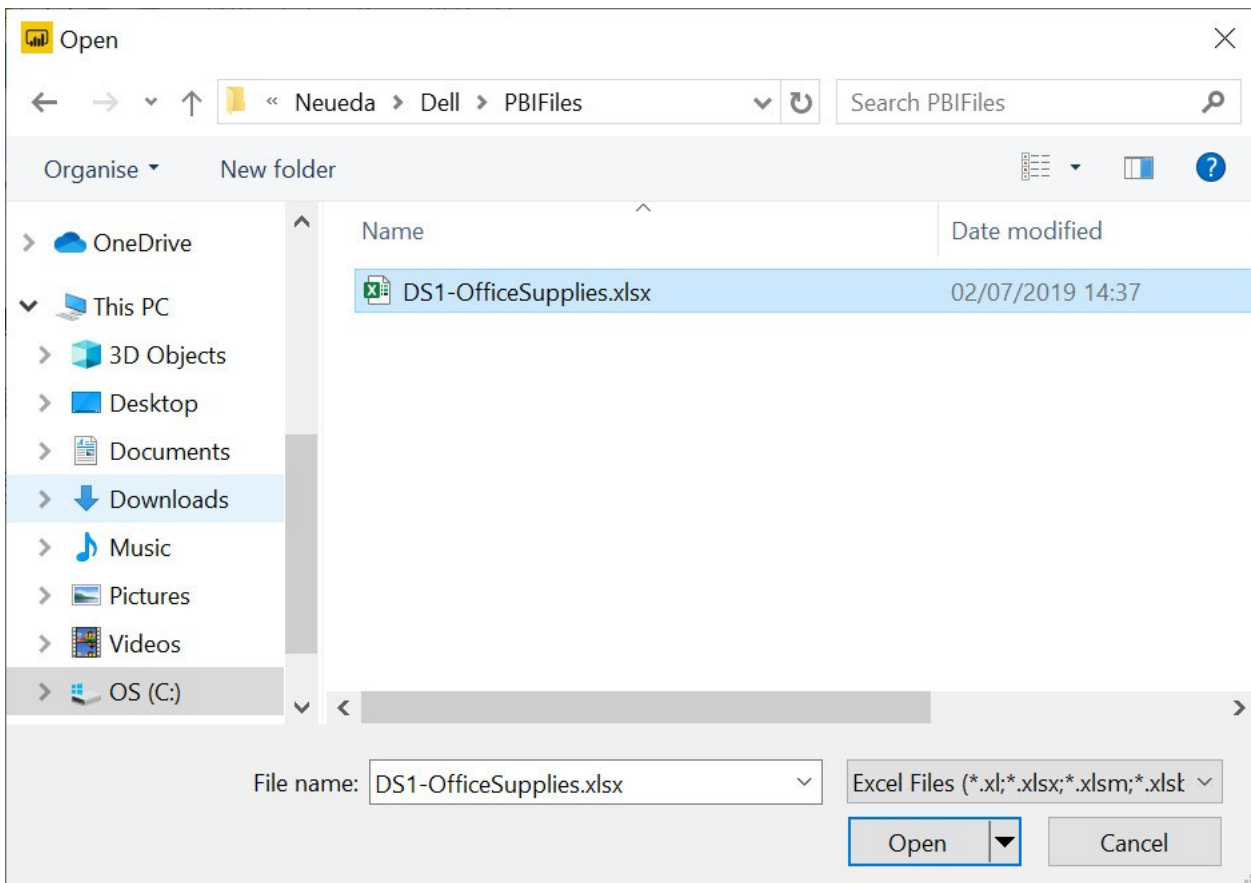


Using Power BI Desktop – Data Model and Visualizations

- Power BI files are saved with a .pbix extension.
- Using Power BI Desktop we can import data from databases, the internet and files. We will use an Excel file. Choose Excel from the **Get Data** dropdown button on the **Home** tab on the ribbon.

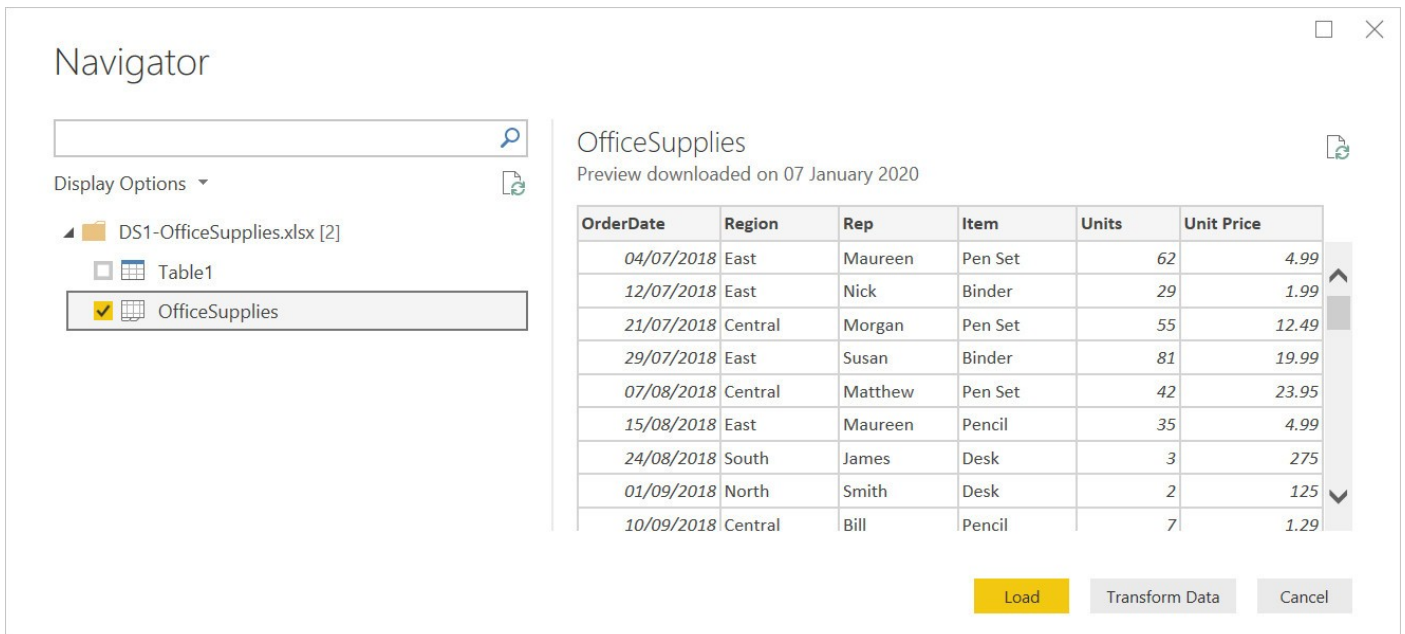


- Select the DS1-OfficeSupplies file and click **Open**.



Using Power BI Desktop – Data Model and Visualizations

7. A **Navigator** window opens to allow you to select the data that you want to load into Power BI. Select the checkbox next to **Office Supplies**. You will then get a preview of the data. Click **Load** to load the data into Power BI Desktop.

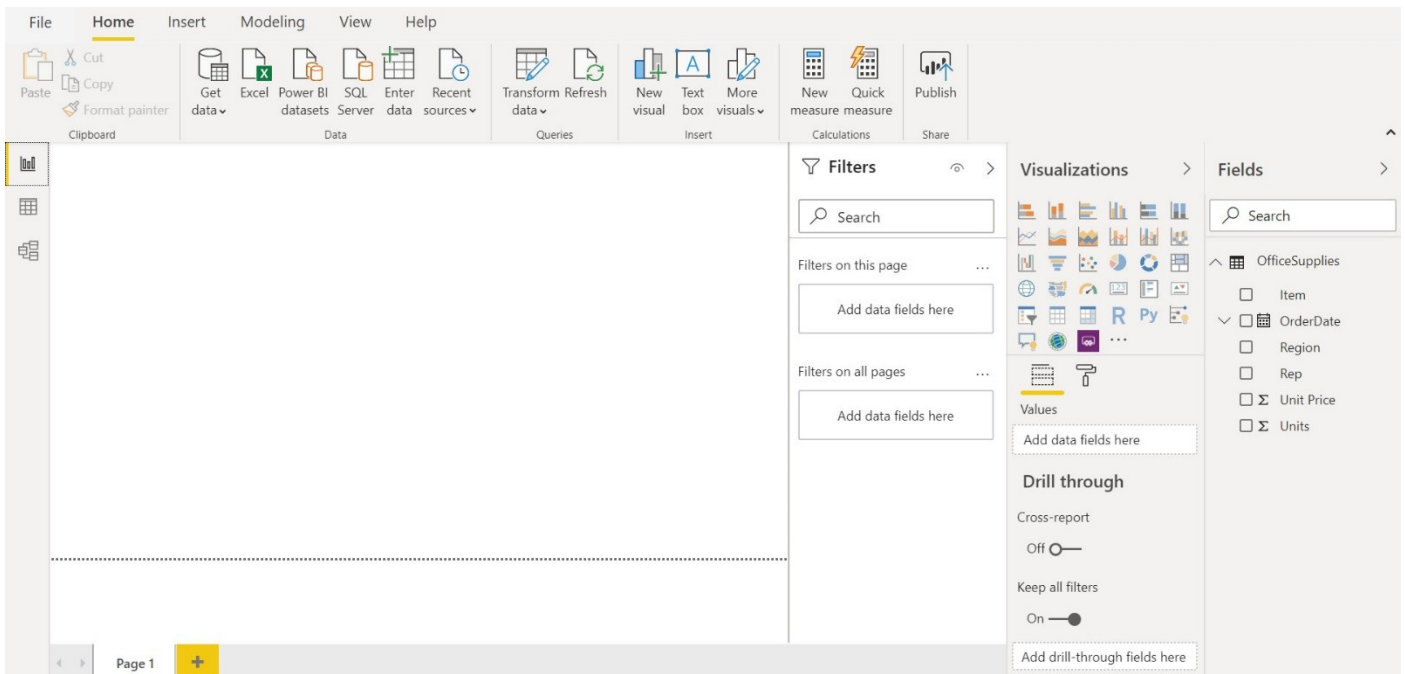


The Navigator window displays the 'OfficeSupplies.xlsx' file. Under 'Display Options', the 'OfficeSupplies' table is selected. The preview shows data downloaded on 07 January 2020.

OrderDate	Region	Rep	Item	Units	Unit Price
04/07/2018	East	Maureen	Pen Set	62	4.99
12/07/2018	East	Nick	Binder	29	1.99
21/07/2018	Central	Morgan	Pen Set	55	12.49
29/07/2018	East	Susan	Binder	81	19.99
07/08/2018	Central	Matthew	Pen Set	42	23.95
15/08/2018	East	Maureen	Pencil	35	4.99
24/08/2018	South	James	Desk	3	275
01/09/2018	North	Smith	Desk	2	125
10/09/2018	Central	Bill	Pencil	7	1.29

Buttons at the bottom: Load, Transform Data, Cancel.

8. In a short period the data will be loaded. Under **Fields** you should see the **OfficeSupplies** table.



The Power BI Desktop interface shows the 'Fields' pane on the right. The 'OfficeSupplies' table is loaded, and its fields are listed: Item, OrderDate, Region, Rep, Unit Price, and Units. The 'Values' section is empty, and the 'Drill through' section is set to 'Off'.

Using Power BI Desktop – Data Model and Visualizations

9. Here is a description of the OfficeSupplies dataset.

Field	Data Type	Description
Item	Text	The item that was sold
OrderDate	Date	The date of the order
Region	Text	The sales region
Rep	Text	The sales rep who made the sale
UnitPrice	Decimal Number	The price of the item
Units	Whole Number	The number of units sold

10. There are three views in Power BI Desktop that can be accessed on a panel on the left. We use Report view to add and customize visualizations. Data view is used to see the data and the results of any calculated columns. If we have multiple data sets or tables in the model, we can create relationships in Model view.

11. Fields and Calculated Columns can be used in Visualizations. To create a Calculated Column select Data view and click the **New Column** button.

File Home Help **Table tools**

Name OfficeSupplies

Mark as date table Calendars

Manage relationships Relationships

New measure Quick measure New column New table

Structure

Report View

Data View

Model View

OrderDate	Region	Rep	Item	Units	Unit Price
04 July 2018	East	Maureen	Pen Set	62	4.99
12 July 2018	East	Nick	Binder	29	1.99
21 July 2018	Central	Morgan	Pen Set	55	12.49
29 July 2018	East	Susan	Binder	81	19.99
07 August 2018	Central	Matthew	Pen Set	42	23.95
15 August 2018	East	Maureen	Pencil	35	4.99
24 August 2018	South	James	Desk	3	275
01 September 2018	North	Smith	Desk	2	125
10 September 2018	Central	Bill	Pencil	7	1.29
18 September 2018	East	Maureen	Pen Set	16	15.99
27 September 2018	South	James	Pen	76	1.99

Using Power BI Desktop – Data Model and Visualizations

12. Change the formula in the formula bar to **Revenue = [Units] * [Unit Price]** and press **Enter** or click on the tick as you would do in Excel. Formulas in Power BI are **DAX** formulas or expressions. DAX stands for Data Analysis Expressions. DAX also includes a rich set of functions (300 plus) that can be used in formulas, many are the same as Excel (SUM, AVERAGE, MAX, MIN etc).

File	Home	Help	Table tools	Column tools
Name	Revenue	\$% Format	General	Σ Summarization Sum
Data type	Decimal number	\$ % ¢ 1000 Auto		Data category Unca
Structure		Formatting		Propertie

1	Revenue = [Units] * [Unit Price]
---	----------------------------------

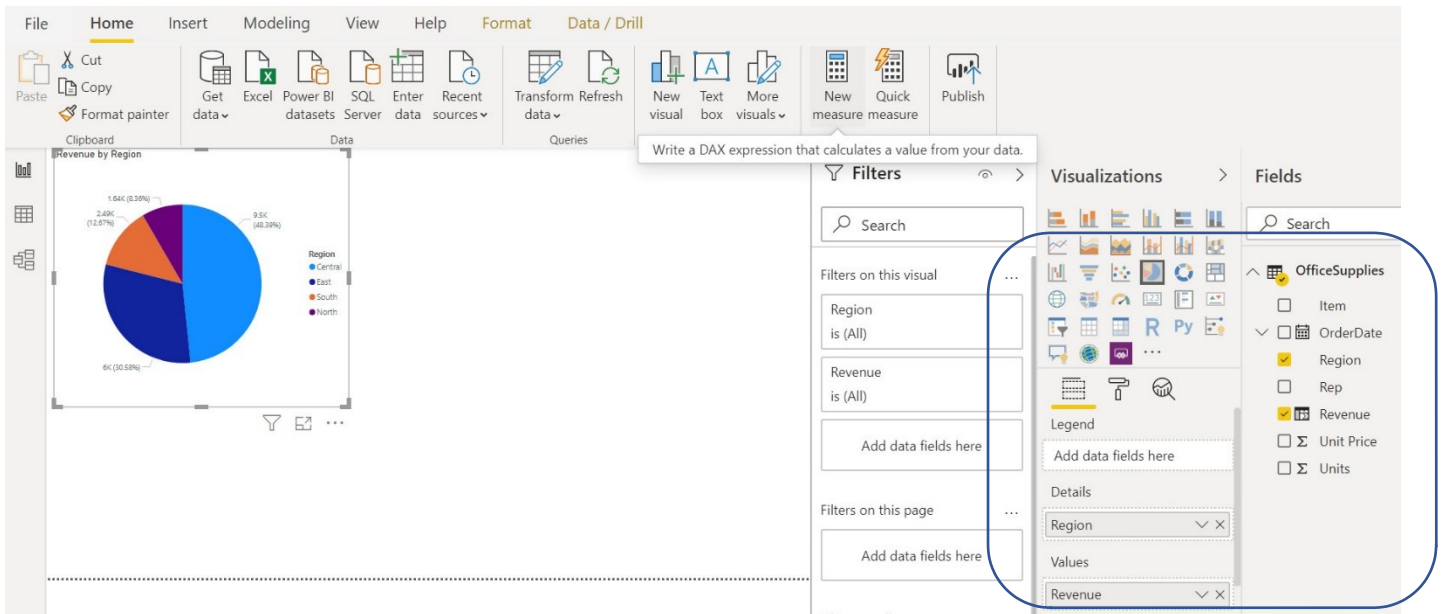
OrderDate	Region	Rep	Item	Units	Unit Price	Revenue
04 July 2018	East	Maureen	Pen Set	62	4.99	309.38
12 July 2018	East	Nick	Binder	29	1.99	57.71
21 July 2018	Central	Morgan	Pen Set	55	12.49	686.95
29 July 2018	East	Susan	Binder	81	19.99	1619.19
07 August 2018	Central	Matthew	Pen Set	42	23.95	1005.9
15 August 2018	East	Maureen	Pencil	35	4.99	174.65
24 August 2018	South	James	Desk	3	275	825
01 September 2018	North	Smith	Desk	2	125	250
10 September 2018	Central	Bill	Pencil	7	1.29	9.03

13. Now we will create our first visualization. In **Report** view, click the Pie chart icon on the Visualizations panel.

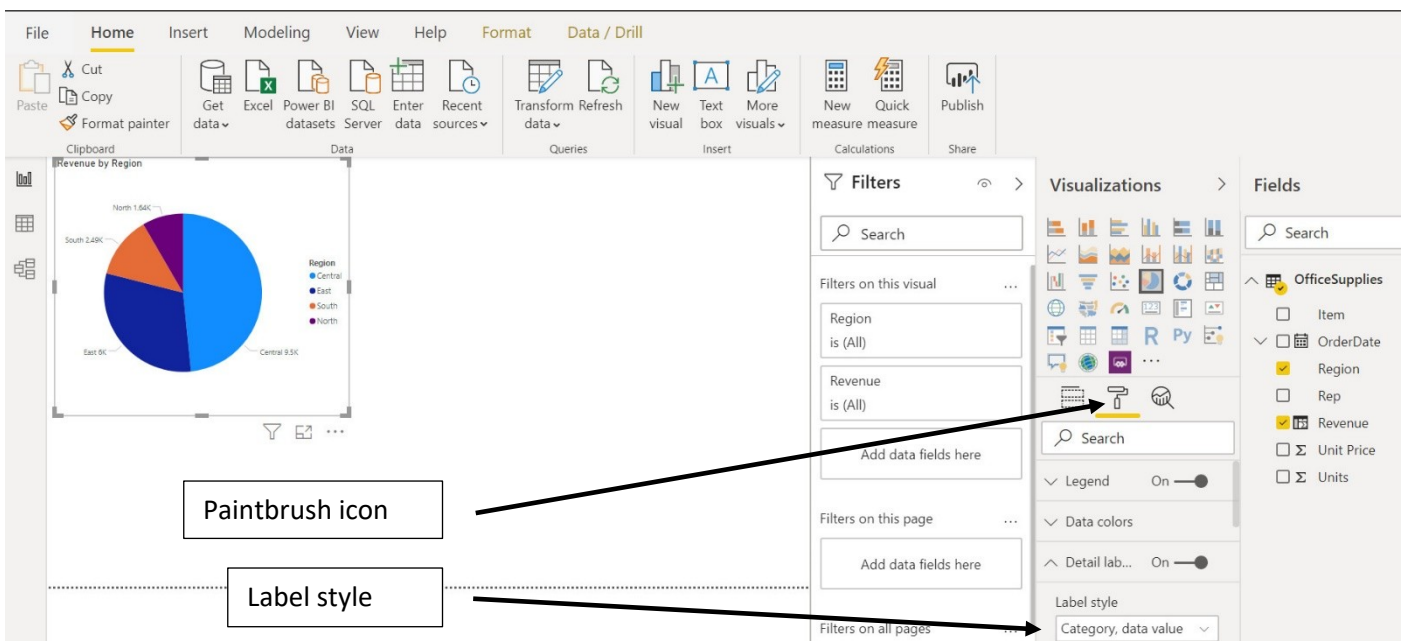
The screenshot shows the Power BI Desktop interface. The top ribbon includes File, Home, Insert, Modeling, View, and Help. The Home ribbon has options like Paste, Copy, Format painter, Get data, Excel, Power BI datasets, SQL Server, Enter data, Recent sources, Transform data, Refresh data, New visual, Text box, More visuals, New measure, Quick measure, and Publish. The main area is labeled 'Report View'. On the right, the 'Visualizations' panel is open, showing a search bar and a grid of chart icons. A pie chart icon is highlighted with a red box and labeled 'Pie Chart Icon'. Below the chart icons are sections for 'Filters on this page', 'Filters on all pages', 'Values', and 'Drill through'. The 'Fields' panel on the far right shows a list of fields: OfficeSupplies, Item, OrderDate, Region, Rep, Revenue, Unit Price, and Units. A red arrow points from the 'Pie Chart Icon' label to the pie chart icon in the Visualizations panel. Another red arrow points from the 'Report View' label to the large empty area on the left.

Using Power BI Desktop – Data Model and Visualizations

14. You will now have a pie chart visualization ready to be configured. With the pie chart visualization selected drag the **Region** field and drop it into the **Details** section. Then drag the **Revenue** calculated field into the **Values** section. Use this screenshot for guidance.



15. The pie chart can be customized. Under the Visualizations icons there is a **Paintbrush** icon that lets you format the selected visualization. Click the **Paintbrush** icon, expand **Detail labels**, then under **Label style** select the **Category, data value** option. You can also experiment with other options and turn the **Legend** on and off.



Using Power BI Desktop – Data Model and Visualizations

16. Your instructor will demonstrate creating other **visualizations** and **slicers** (filters) after which you can try creating the following charts and slicers.

Pie chart of units sold by rep

Stacked bar chart of units sold by region

Stacked bar chart of units sold by rep

Line chart for units sold over time

Line chart comparing units sold over time with unit price over time

Slicer to filter by region

Slicer to filter by rep