Exploratory Data Analysis – 1

Module Overview

The topics in this module are briefly described below:

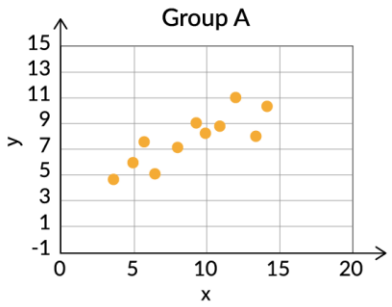
- **Introduction to Data Visualization and EDA:** This session will help you understand why we need data visualization and what its purpose is. Further, you will learn the various types of data and how to present it visually.
- **Univariate Analysis:** You will learn how to use Python's Seaborn library to create different types of visualizations, such as count plots, box plots, histograms, density plots, and strip plots.

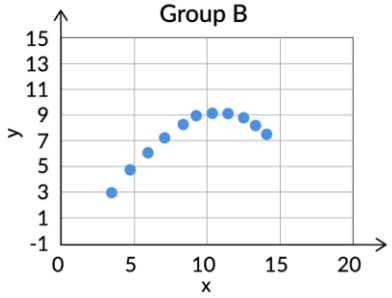
Here is what you will learn in this session:

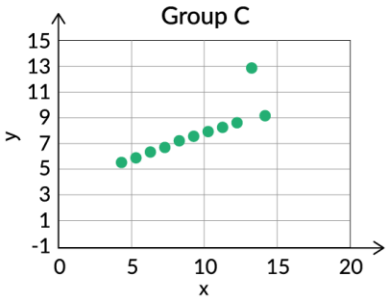
- **Data visualization:** Representing data visually helps to convey information in a simple, clear, and concise manner.

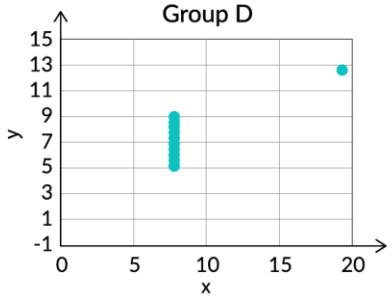
  There are two purposes of data visualization:

  - Exploration
  - Explanation
- **Tools for exploratory analysis:** Data can be visualized with a range of tools from Python to Excel. The choice of a tool depends on the end goal and the types of visualizations required.
- **Visual components:** *Coordinates*, *scales*, and *cues* are the three visual components that help us present data visually.
- **Types of data:** We can create visualizations based on the data types available. For example, we can create visualizations based on *numerical*, *categorical*, *geographic*, and *temporal* data.
- **Exploring data:** This step involves you becoming familiar with your data and exploring initial patterns, outliers, and characteristics for further investigation. Visualizations can often be a useful tool to explore data.









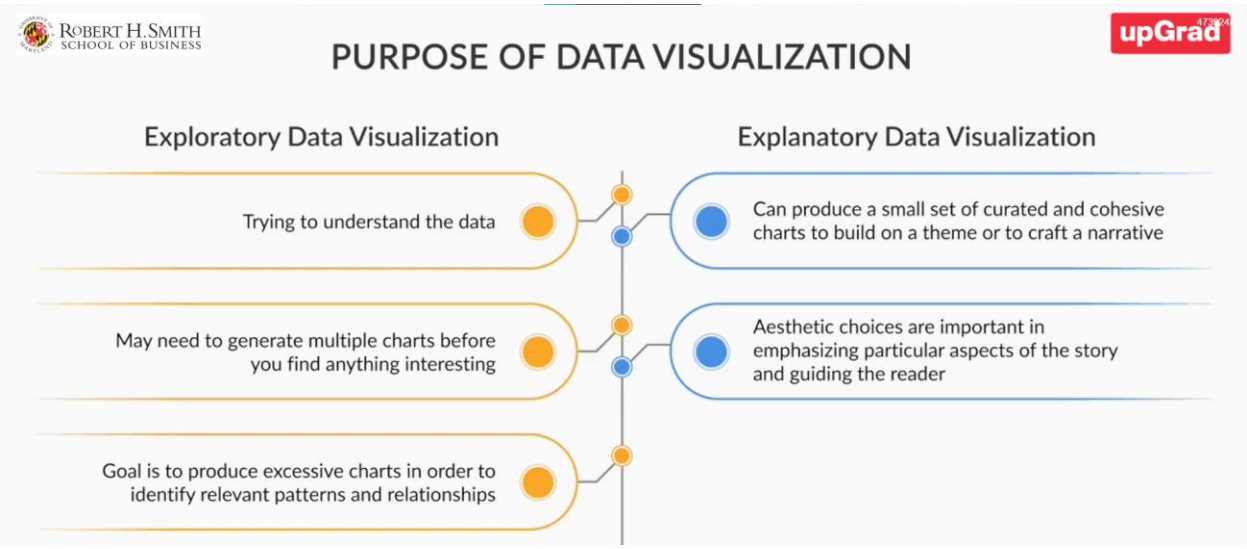
The graphs above make it easy to infer relationships such as these:

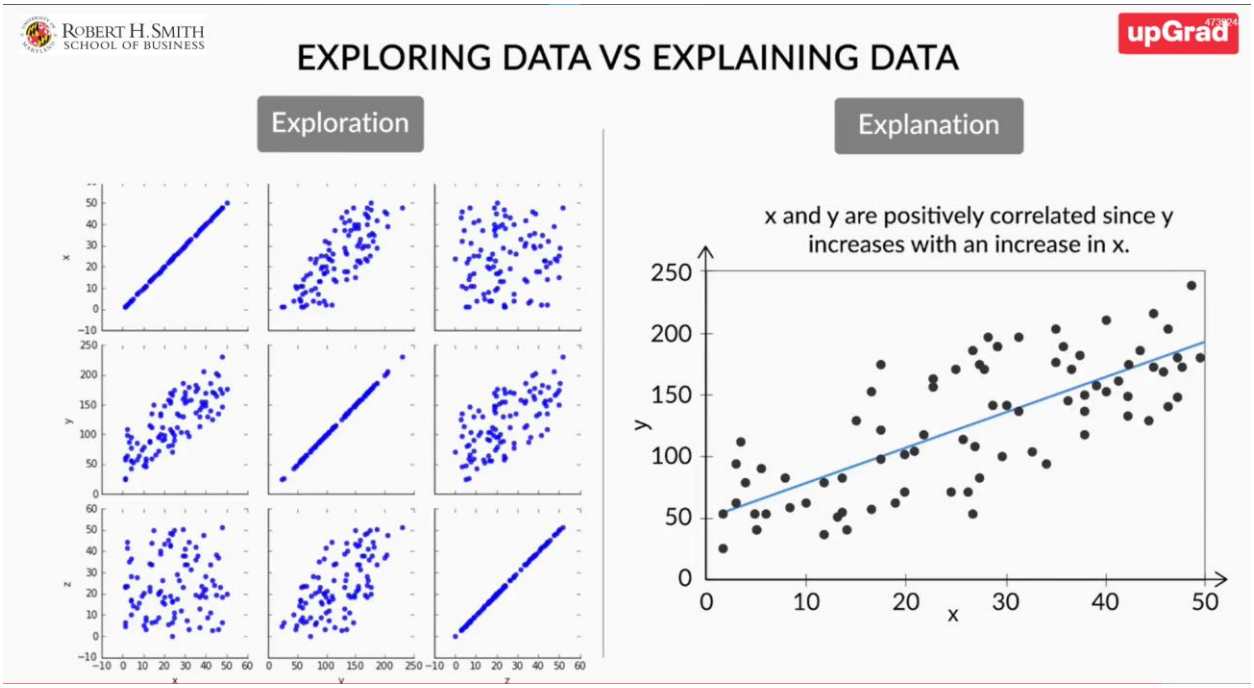
- In Group A, x and y appear to be positively correlated since y appears to increase with an increase in x.
- In Group B, x and y appear to have an inverted U-shaped relationship, with a potential maximum y value at x=11.
- In Group C, x and y appear to have a strong positive relationship, with a potential outlier at x=13.
- In Group D, y increases while x remains at a fixed value, except for a single outlier at x=19.

These relationships would not have been readily visible to us and we could not have inferred them had the information been presented only as raw numerical data, as shown below.

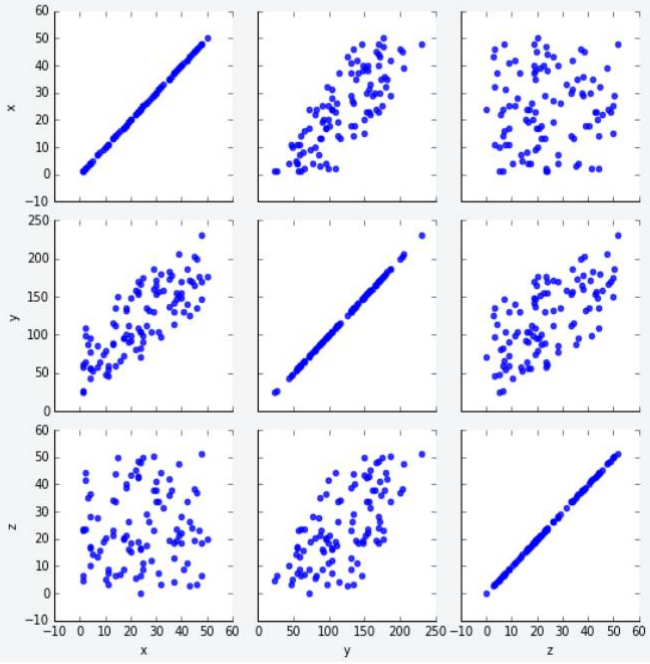
Group A		Group B		Group C		Group D	
x	y	x	y	x	y	x	y
10.00	8.04	10.00	9.14	10.00	7.46	10.00	6.58
8.00	6.95	8.00	8.14	8.00	7.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

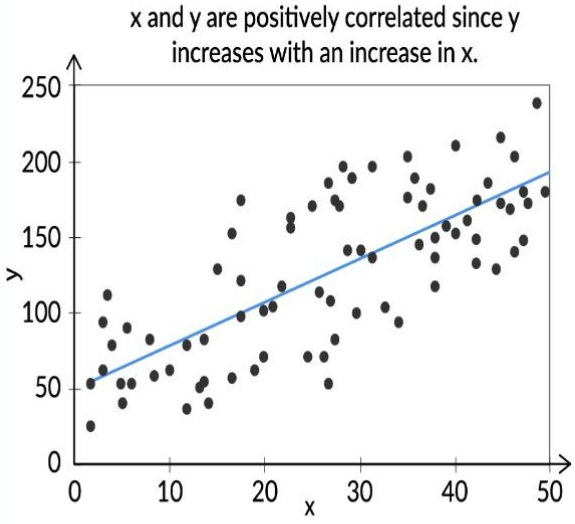
Purpose Of Data Visualization





Exploration	Explanation
• Understand data • Explore data in different ways to derive insights • Find interesting areas to investigate further • Identify relevant patterns and relationships in the data	• Convey interesting insights • Present relevant findings to the audience

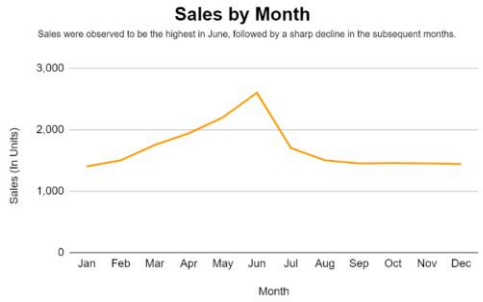




The above chart is a scatterplot matrix containing nine scatterplots. It shows the relationships among three variables: x, y, and z. This is an example of exploration since the scatterplot matrix shows the relationships among all possible pairs of the three variables in a single plot.	This scatterplot above shows a positive relationship between x and y. The title contains relevant information about the scatterplot. This is an example of explanation since the scatterplot shows only the relevant information.

Data Visualization

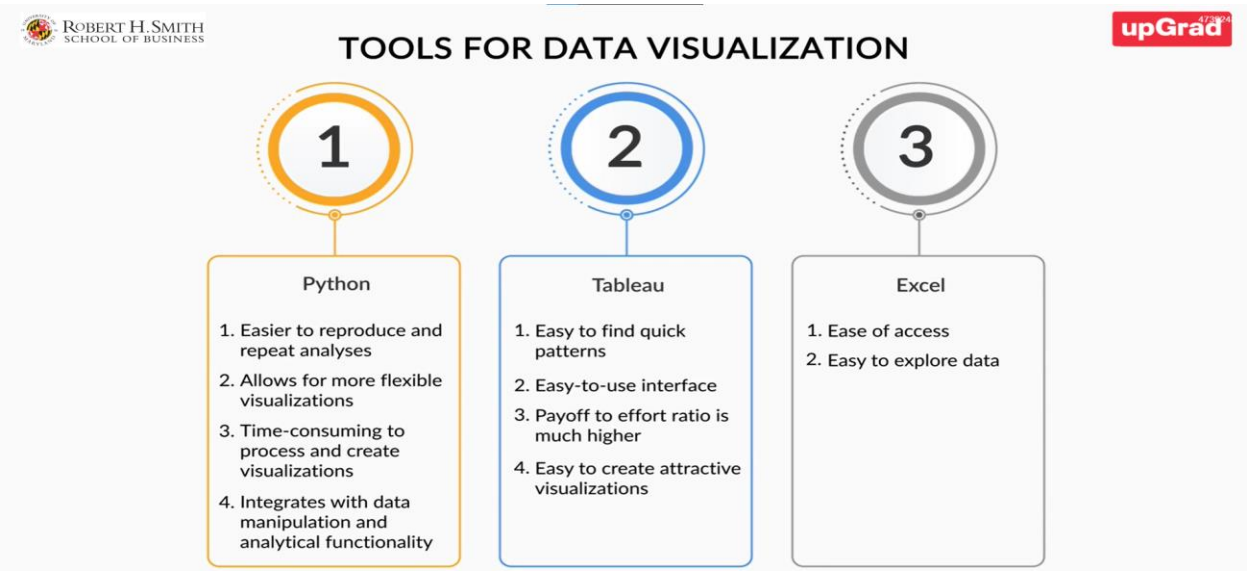
The graph below shows the monthly sales of a company for the year 2021.



Data explanation ✓ Correct

Feedback:
Correct! This chart is an insightful presentation of monthly sales and shows the relevant findings to the audience. Hence, this is an example of data explanation.

Tools for data-visualization :-

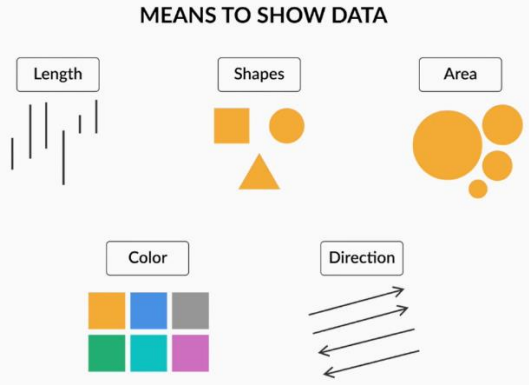


Why Use Python	Why Use Tableau	Why Use Excel
• Reproducible and repeatable analysis • More flexible visualizations than with the others • Integrates data manipulation and analysis in one tool	• Helps find quick patterns • Allows for creating attractive visualizations with little effort	• Commonly available • Helps find patterns quickly in a small data set • Allows for exploring data easily using pivot tables

Visual Components :-

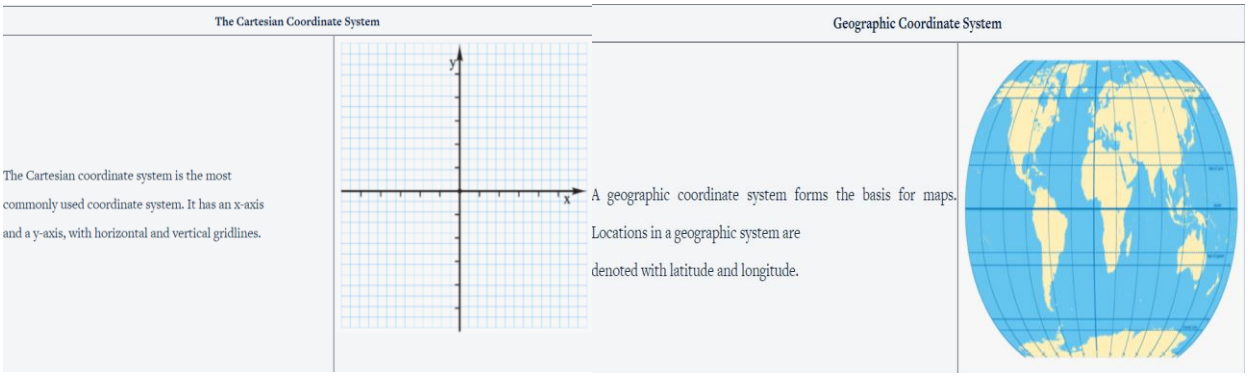
Cues

Visual cues help the audience to easily focus on the relevant parts of a visualization. The visual cues you incorporate in your visualization depend on the type of data and the type of graph on which you are plotting it. This image shows some commonly used visual cues.



Coordinates

The data in a chart can be shown on a two-dimensional (2D) plane using a coordinate system. Let's take a look at the types of coordinate systems.

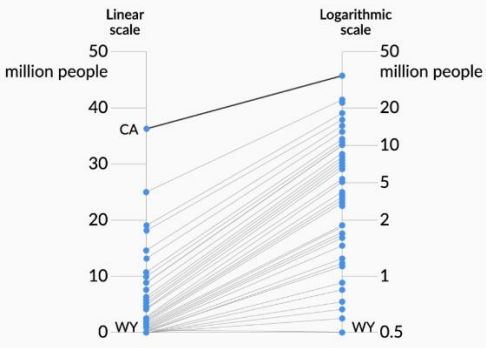


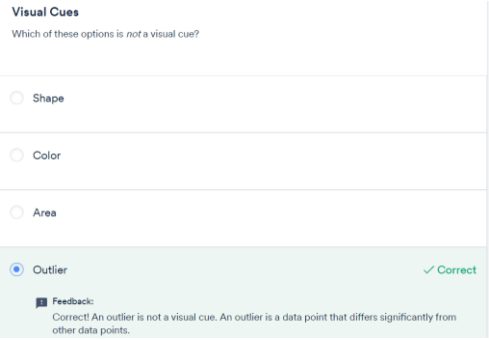
Note: There are other coordinate systems, including the polar coordinate system, but they are out of the scope of this module. You can read more about them in the resources provided under "Additional Reading" at the end of this segment.

Scale

Scale means the size of something (e.g., an object) in comparison with something else. You can, for instance, use a linear or logarithmic scale to measure the distance between the points in a Cartesian coordinate system.

You can refer to this image to recall what each of these scales means.

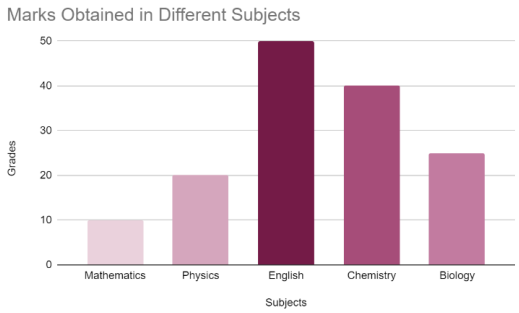


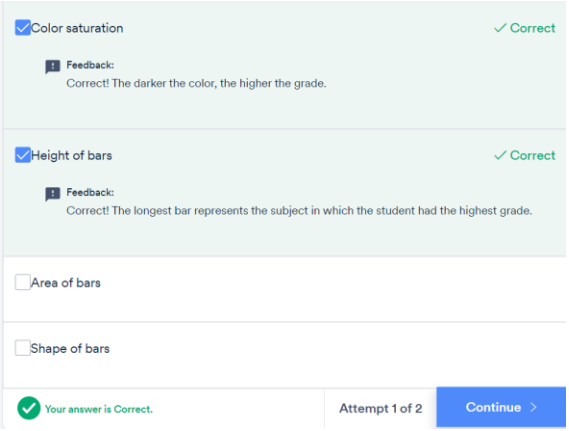


Visual Cues

The bar graph below shows the grades of a student in five subjects. Which among the following visual cues represents the grades obtained by the student in each subject in the graph?

(**Note:** More than one option may be correct.)





Types Of Data :-

The data you come across can be of four potential types:

- **Temporal:** A specific time associated with a record. Temporal data include information such as dates (e.g., *June 19, 2021*), time stamps (e.g., *Jan 4, 2021 at 12:35 pm*), and seasonality (e.g., *Winter 2021*).
- **Categorical:** A classification associated with a record. When imported, categorical data is often in the form of strings of text.
- **Geographic:** Location-related information about a record. Geographic data can be in multiple formats:
  - Strings, such as the *Region* field, which describes a specific area, (e.g., a state),
  - Numbers, such as zip codes.
- **Numerical:** A value associated with a record. Numerical data can be in multiple formats, such as integers, doubles, percentages, and currency.

Year	Channel	Region	Sales (#)	Revenue ($)
2019	Retail	California	1,000	50,000
2019	Retail	New York	650	29,250
2019	Digital	California	500	27,500
2019	Digital	New York	120	6,600
2020	Retail	California	100	5,000
2020	Retail	New York	50	2,250
2020	Digital	California	2,000	110,000
2020	Digital	New York	1,500	82,500
2021	Retail	California	800	40,000
2021	Retail	New York	700	31,500
2021	Digital	California	1,900	104,500
2021	Digital	New York	1,600	88,000

In this data set:

- ***Year*** is a temporal field because it denotes a unit of time.
- ***Channel*** is a categorical field because it denotes qualitative information.
- ***Region*** is a geographic field because it denotes a location in a geographic coordinate system.
- ***Sales*** **and** ***Revenue*** are both numerical fields as they denote the values associated with the fields.

Categorical Attributes

Consider this data set:

	car_name	price($)	date_of_sale	mileage	condition	meter_reading
0	Lamborghini	174,390	12/09/20	3.5	good	75,625
1	Lamborghini	119,780	29/07/21	4.5	better	45,400
2	Porsche	97,400	17/03/22	NaN	best	25,000
3	Honda Civic	71,432	NaN	12.0	bad	125,000

Which of these attributes are categorical? Note: More than one option may be correct.

Car_name and condition are categorical form of data from the above dataset.

Variable Type

Which of these types of variables does *date_of_sale* represent?

Temporal ✓ Correct

Feedback:
Correct! *date_of_sale* is a temporal variable since it is a specific time associated with a record.

Data Example and understanding.

Data exploration includes these levels:

- Considering interesting business questions
- Translating the business needs to data questions
- Using visualization techniques to plot data
- Identifying interesting patterns

What makes a question interesting? Here are some characteristics of an interesting business question:

- The answers should have actionable implications.
- Data visualization should reveal novel insights.

To help you understand, let's take the example data set that you saw earlier, in the previous segment.

Example

Here is the data set. Here are some potentially interesting questions that you can ask based on this data set.

Year	Channel	Region	Sales (#)	Revenue ($)
2019	Retail	California	1,000	50,000
2019	Retail	New York	650	29,250
2019	Digital	California	500	27,500
2019	Digital	New York	120	6,600
2020	Retail	California	100	5,000
2020	Retail	New York	50	2,250
2020	Digital	California	2,000	110,000
2020	Digital	New York	1,500	82,500
2021	Retail	California	800	40,000
2021	Retail	New York	700	31,500
2021	Digital	California	1,900	104,500
2021	Digital	New York	1,600	88,000

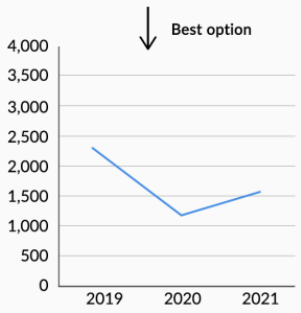
- What is the relationship between Sales and Revenue? — Not novel. Revenue is always a function of Sales.
- What are the trends for digital? — Not actionable. It's a vague question with unclear implications.
- Were sales higher in CA or NY in 2021? — Very simple. No exploratory visualization needed to get answer.
- How have digital sales changed compared to retail in the last 3 years? — Good question. The answer could determine channel investments.

Data Exploration – 1

Year	Channel	Sales(#)
2019	Retail	1000
2019	Retail	650
2019	Digital	500
2019	Digital	120
2020	Retail	100
2020	Retail	50
2020	Digital	2000
2020	Digital	1500
2021	Retail	800
2021	Retail	700
2021	Digital	1900
2021	Digital	1600

Using data types :

- Categorial (Channel)
- Temporal (Year)
- Numerical (Sales)

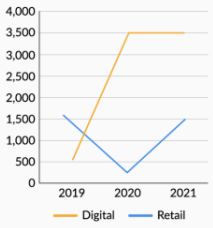


Data Exploration - 2

Year	Channel	Sales(#)
2019	Retail	1000
2019	Retail	650
2019	Digital	500
2019	Digital	120
2020	Retail	100
2020	Retail	50
2020	Digital	2000
2020	Digital	1500
2021	Retail	800
2021	Retail	700
2021	Digital	1900
2021	Digital	1600

Aggregate the data appropriately

Year	Digital	Retail
2019	620	1650
2020	3500	150
2021	3500	1500



Interpretation:
Digital sales increased sharply in 2020 and stayed strong in 2021