Intro to Visualization in Python - Exploratory data analysis - 1

One should look for what is and not what he thinks should be. (Albert Einstein)

Exploratory Data Analysis: Topic introduction

In this part of the course, we will cover the following concepts:

- Exploratory data analysis use cases
- Perform EDA on data

Module completion checklist

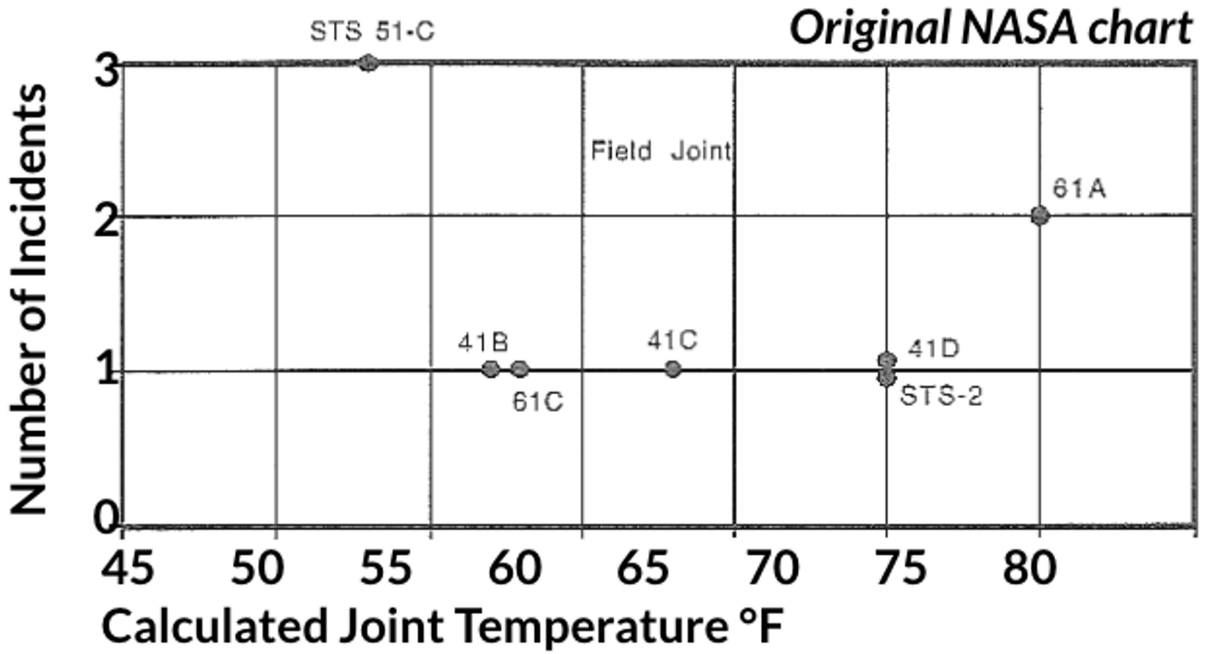
Objective	Complete
Discuss data visualization and exploratory data analysis	
Describe chart types by data and form	

The Challenger explosion example

- The 1986 **Space Shuttle Challenger explosion** is an emblematic case study of how data visualization can play an essential role in decision-making
- The explosion happened due to low temperatures that affected shuttle parts
- Edward Tufte, a visualization expert, argues that the cause of this tragedy was an unreadable format of data given to decision-makers

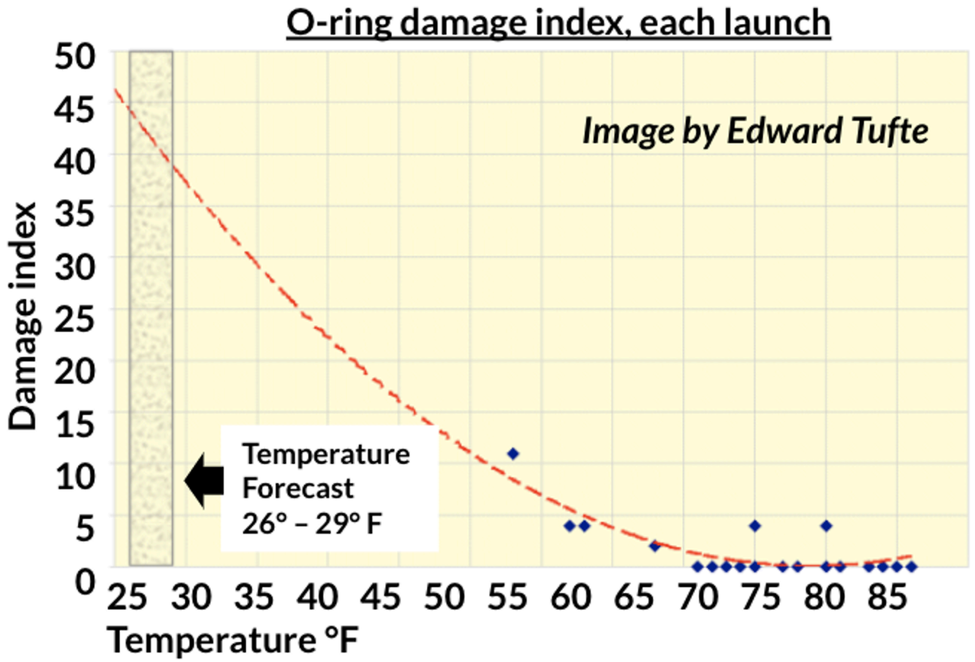
The original visualization of the temperature

- The chart below was presented to the experts at the time
- How **easy is it to interpret** the chart?



The revised visualization of the temperature

- Edward Tufte argues a better chart may have prevented disaster
- How **easy is it to interpret** the **revision** created?

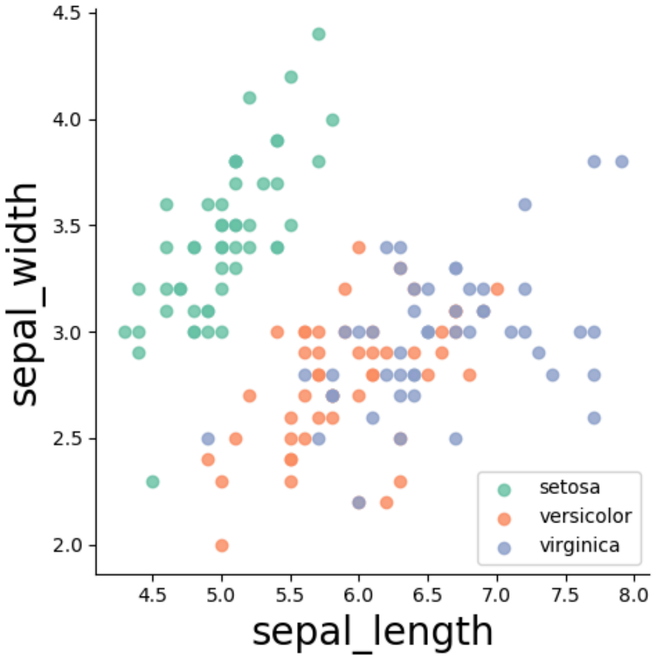


Data Visualization

- Data visualization is an attempt to make data more easily digestible by rendering it in a visual context (e.g., charting, graphing, etc.)
- We use data visualization to transform raw data into something compelling
- Data visualization is at the intersection of art and science

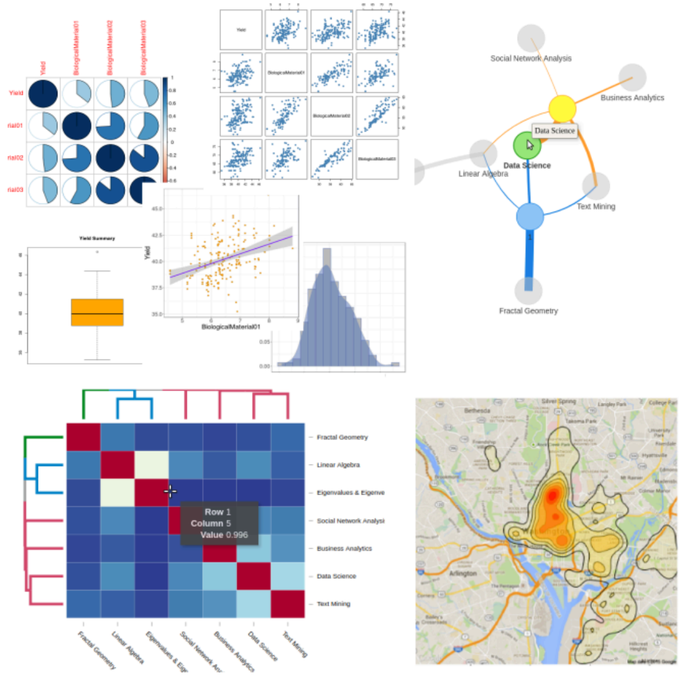
Why visualize data?

- Visual context provides **insights on patterns, trends, and correlations** that might be difficult to detect otherwise
- It is a simple way to **convey concepts** and **provide visual access to large amounts of complex data**
- Using **Python** is excellent as it has **multiple graphing libraries** with many valuable features



Why build a visualization?

- To provide **valuable, interpretable, and relevant insights**
- To give a **visual or graphical representation** of data / concepts
- To **communicate ideas**
- To provide an accessible way to **see and understand trends, outliers, and patterns** in data
- To try to **confirm a hypothesis**



Chat Activity: Explore a Dashboard

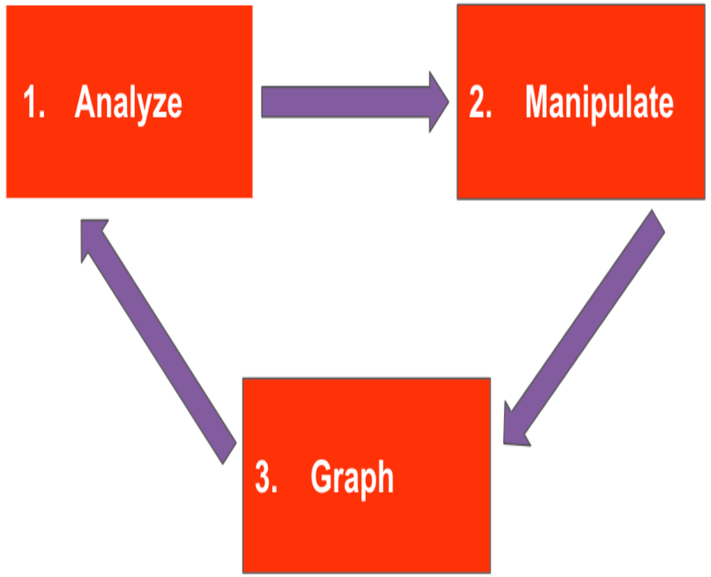
- What is a dashboard? It is a **visual display of all your data**
- Let’s assume you work at a recruitment firm, and the firm has a dashboard to track and view its KPIs (Key Performance Indicators)
- What **KPIs would you like to track** using that dashboard to help make better business decisions?
- Share your thoughts in chat

Explore a dashboard

- Take a couple minutes to explore the dashboard
- It was designed to answer various questions or user queries about the financial metrics of business operations
- You can access the dashboard from the following ***link***

Exploratory data analysis (EDA)

- Exploratory data analysis (EDA) is the process of reviewing new data to discover patterns, spot anomalies, test hypotheses, and check assumptions
- It helps to create graphs without breaking the train of thought as you explore your data
- **Visualization is an iterative process** and consists of a few steps:
  - Analyze
  - Manipulate
  - Graph
  - Repeat



Exploratory data analysis in Python

- Python is a powerful tool for EDA because the graphics tie in with the functions used to analyze data
- **What is possible using Python?**
  - Visualization tools available through multitudes of packages (e.g. `matplotlib`, `seaborn`)
  - The visualizations created are high quality graphics that can be saved as `SVG`, `PNG`, `JPEG`, `BMP`, `PDF`
  - Visualizations are often the best way to display patterns in data for printed publications
- Further, we will explore how to visualize data using Python and perform exploratory data analysis to understand and detect the patterns

Module completion checklist

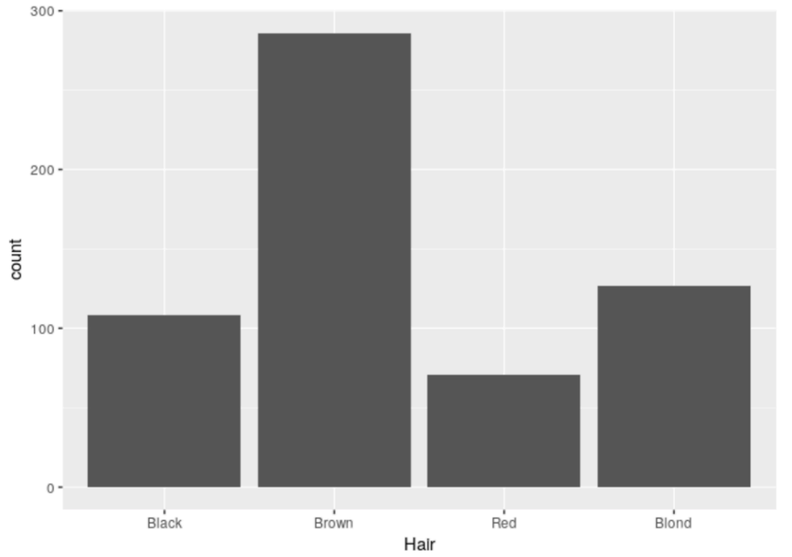
Objective	Complete
Discuss data visualization and exploratory data analysis	✔
Describe chart types by data and form	

Getting started with data viz

- Deciding on what visualization type to use will depend on the data and message you want to communicate
- Common data types include:
  - Categorical
  - Univariate
  - Bivariate
  - Time-based (trending)
  - Text
  - Geospatial

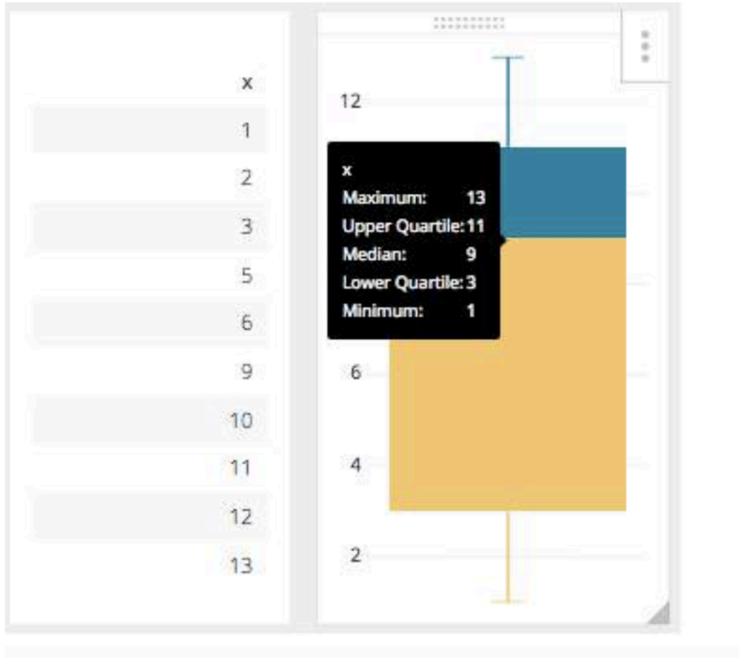
Categorical Data

- Categorical data is **non-numeric or qualitative**
- Insight: comparisons and proportions
- Chart types: vertical bar, column bar, horizontal bar, pie, bullet charts, stacked bar, and tree maps



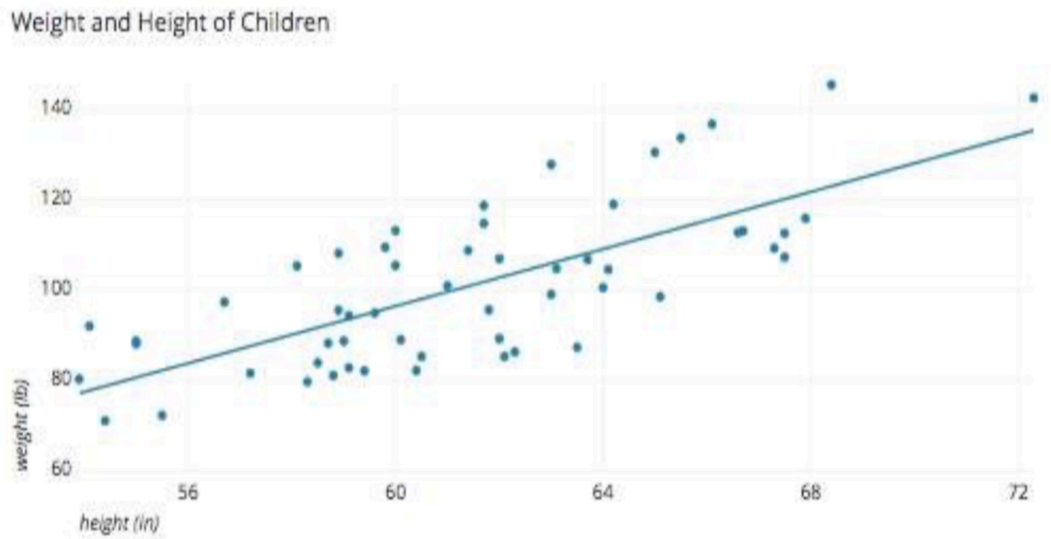
Univariate Data

- Univariate data consists of a **single numeric variable**
- Insight: distributions, proportions, and frequencies
- Chart types: histogram, density, box plots
  - Is the data normally distributed?
  - Are there any outliers?
  - Do you notice any other patterns in the data?
  - These are some of the steps for initial data exploration



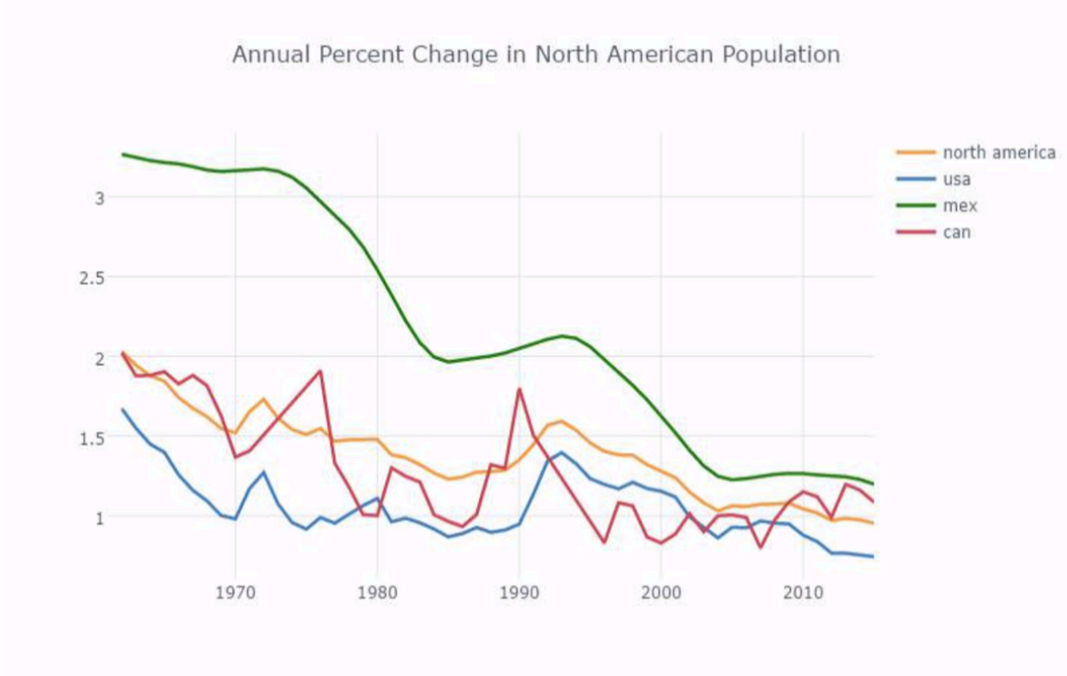
Bivariate Data

- Bivariate data consists of **two (or more) numeric variables** (i.e., weight and height)
- Insight: relationships, correlation, proportions, and frequencies
- Chart types: scatterplot, bubble, parallel, radar, bullet, and heat



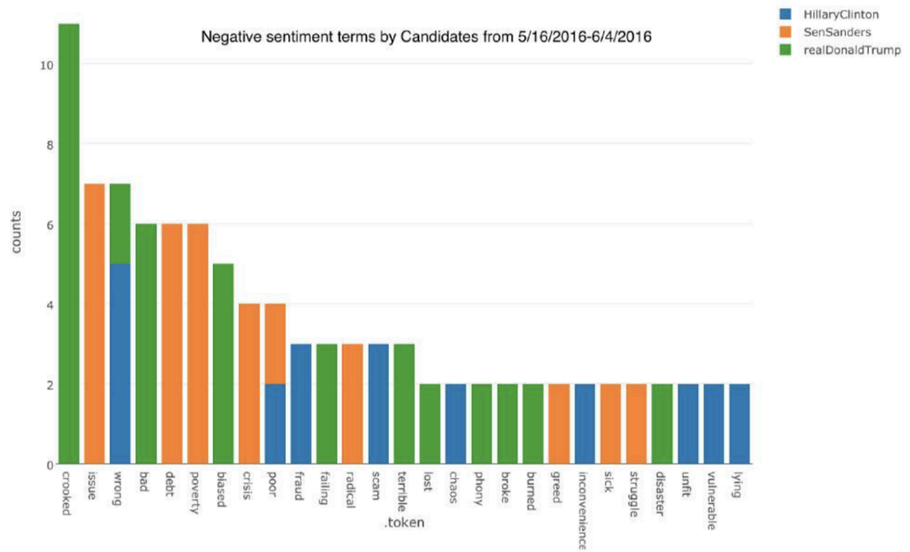
Trend Data

- Trend data includes a **time-based data** (i.e., years, months, days, hours, etc.)
- Insight: trends, comparisons, and cycles
- Chart types: line, area, bubble, vertical bar



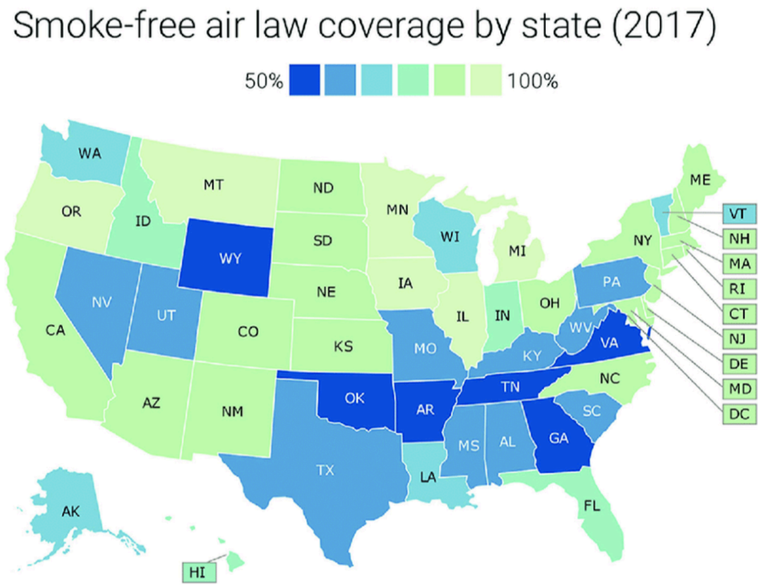
Text Data

- Text data includes **alphanumeric single words or phrases** (keywords)
- Insight: sentiment, comparisons, and frequency
- Chart types: word cloud, histogram, stacked bar chart



Geospatial Data

- Geospatial data includes **qualitative or quantitative information about specific locations**
- Insight: locations, comparisons, and trends
- Chart types: chloropleth filled map, point map, connection map, isopleth map



Common visualizations

- Let’s review when to use some of the common visualizations, including:
  - Tables
  - Bar charts
  - Line charts
  - Area charts
  - Heatmaps
  - Scatterplots

Simple text or table

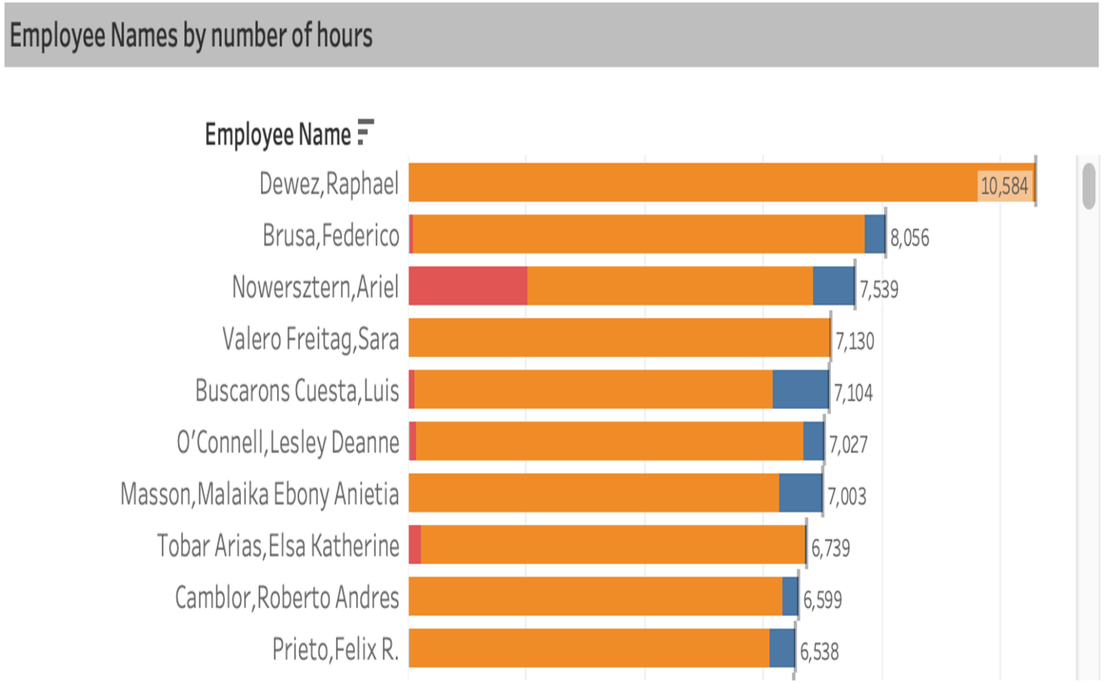
- **Simple text** is used when there is just a number or two to share. Simple text can be a great way to communicate something like:
  - *440 employees worked a total of 31,702 days, an average of 72.05 days per employee*
- **Tables** are helpful when communicating to a mixed audience or showing a few different units of measure

Top Operations based on Work Days

Index	Operation	Duration with the team	Work Days	Employee Count	Work Days/Employee
1	Portfolio Monitoring and Reporting	36	31,702	440	72.05
2	Support to Project Execution	36	21,315.375	385	55.364610390
3	Support to Project Preparation	36	21,270.25	422	50.403436019
4	Support to Fiduciary Work	36	16,639.5	204	81.566176471
5	Dialogue with public sector authoriti..	36	14,207.875	373	38.090817694
6	Technical Advisory and Quality Contr..	36	10,492	90	116.577777778
7	Economic Research (not product spe..	36	9,854.5	113	87.207964602
8	Country and Sector Programming	36	9,138.375	245	37.299489796
9	Trust Fund: Coordination and Corpor..	36	8,983.75	56	160.424107143
10	Macroeconomic monitoring	36	8,961.375	94	95.333776596
11	Strategic Outreach	36	7,495.875	249	30.103915663
12	Communications Planning and Client..	36	5,834.5	84	69.458333333
13	Support to TC (OS) Execution	36	5,731.5	161	35.599378882
14	Strengthening of Country Systems P..	36	5,293.375	111	47.688063063
15	INT Annual Research Report [Global ..	36	4,557.625	19	239.875

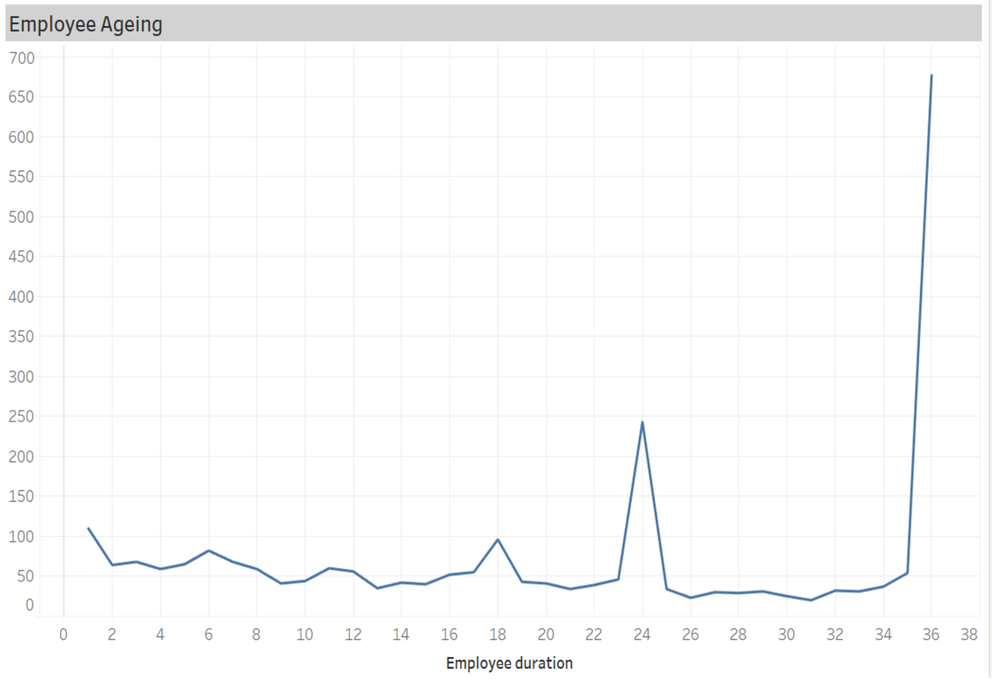
Bar Chart

- Bar charts are used to express **larger variations in data** and how individual data points relate to a whole, comparisons, and ranking
- They express quantities through a bar's length, using a common baseline (=zero)
- **Note:** when the data has lengthy names, using a horizontal bar chart will make the data easier to read



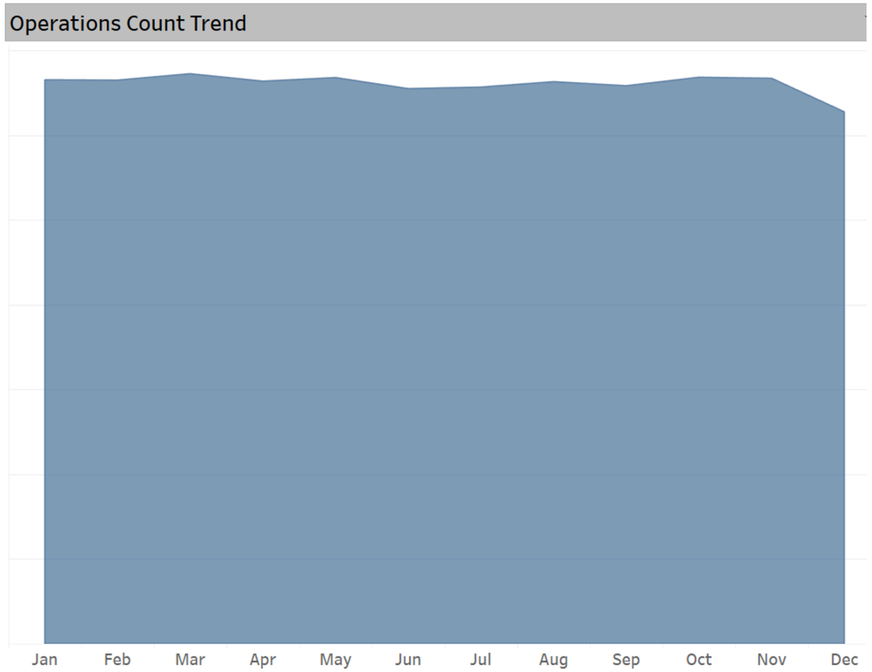
Line Chart

- Line charts are used to **plot continuous data in some unit of time**, such as days, months, quarters or years
- They can also be used to show multiple series of data
- A line graph can also represent a summary statistic, like the average and confidence level range or the point estimate of a forecast



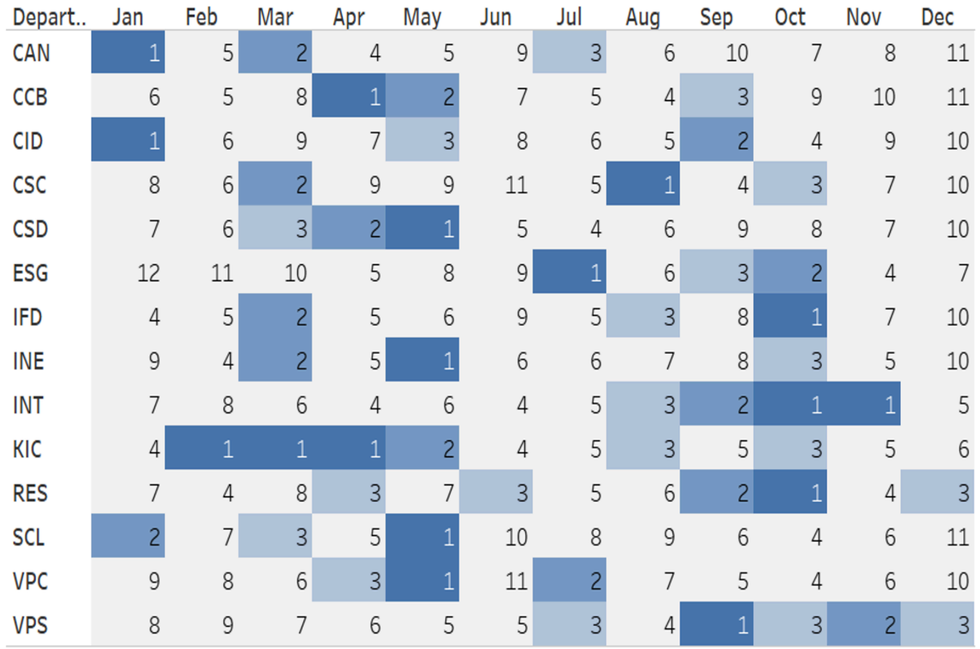
Area Chart

- Area charts are used to **summarize relationships between datasets**, how individual data points relate to a whole
- The visual at the right shows the monthly trend of active operations
- In chat, share your thoughts on how you think this visual could be improved



Heatmap

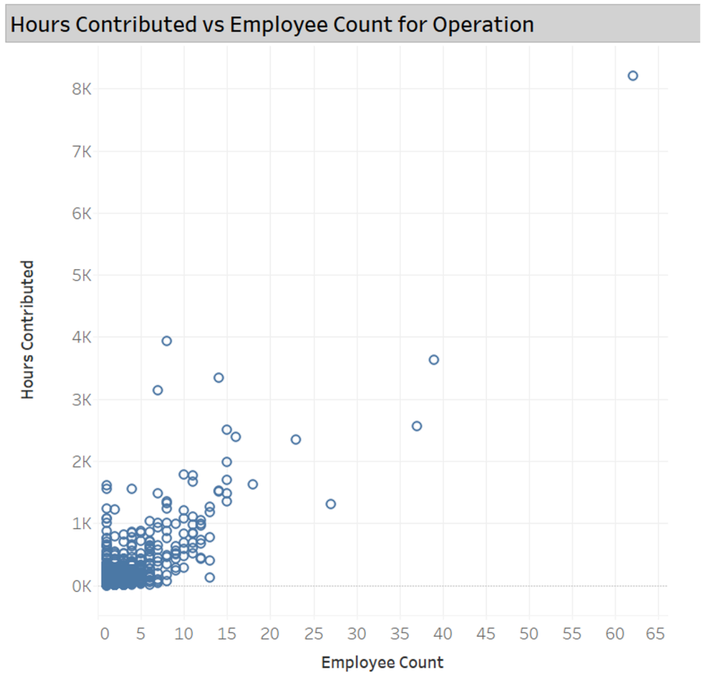
- Heatmaps **visualize data in tabular format**, using colored cells to show the relative magnitude of the numbers
- When using a heatmap, it is helpful to restrict the number of different color gradations
- The visual at the right shows the busiest months ranked by the number of operations for each department



Depart..	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
CAN	1	5	2	4	5	9	3	6	10	7	8	11
CCB	6	5	8	1	2	7	5	4	3	9	10	11
CID	1	6	9	7	3	8	6	5	2	4	9	10
CSC	8	6	2	9	9	11	5	1	4	3	7	10
CSD	7	6	3	2	1	5	4	6	9	8	7	10
ESG	12	11	10	5	8	9	1	6	3	2	4	7
IFD	4	5	2	5	6	9	5	3	8	1	7	10
INE	9	4	2	5	1	6	6	7	8	3	5	10
INT	7	8	6	4	6	4	5	3	2	1	1	5
KIC	4	1	1	1	2	4	5	3	5	3	5	6
RES	7	4	8	3	7	3	5	6	2	1	4	3
SCL	2	7	3	5	1	10	8	9	6	4	6	11
VPC	9	8	6	3	1	11	2	7	5	4	6	10
VPS	8	9	7	6	5	5	3	4	1	3	2	3

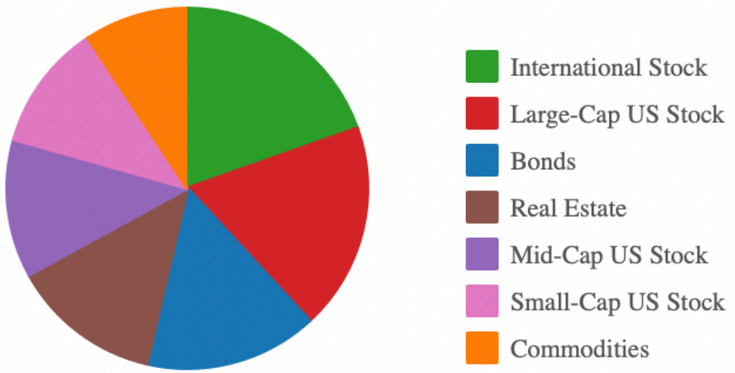
Scatterplot

- Scatterplots show the type of **relationship between two numeric variables**
- Scatterplots are often used in scientific fields and are sometimes viewed as “complicated” to understand, but there are real-world uses as well
- In chat, share your thoughts on what relationship this scatterplot represents

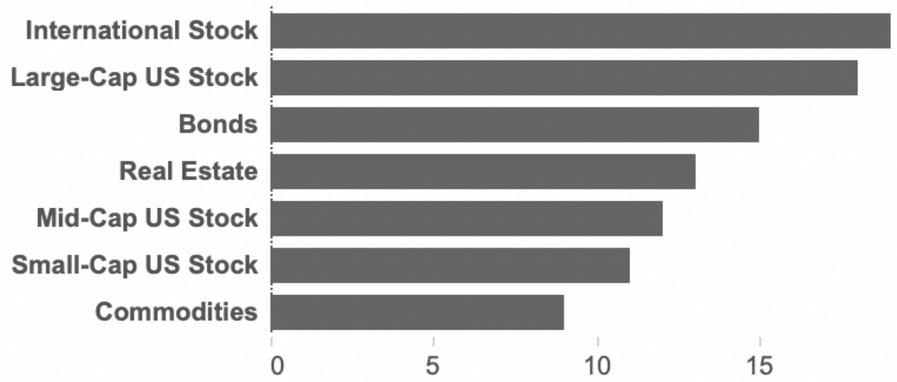


Review Quiz: Data Visualization

- Question 1: **Which graph of the two makes it easy to determine what investment has a more significant market share?**
- Share your answer in the chat



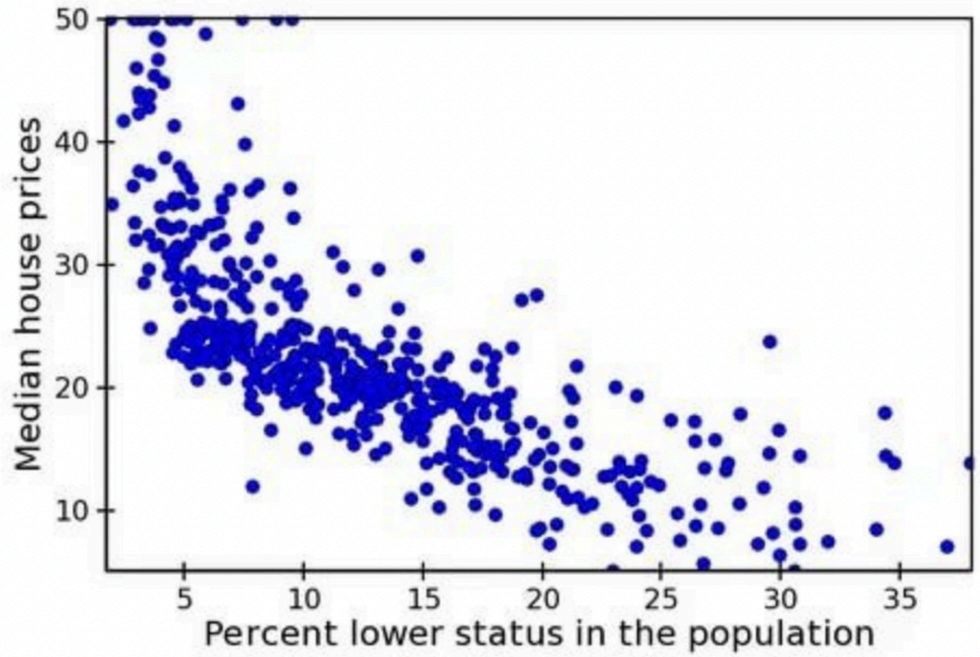
Option A



Option B

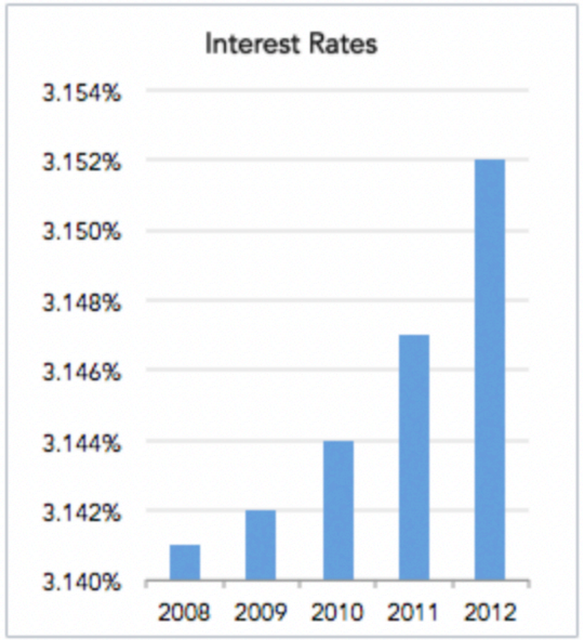
Review Quiz: Data Visualization (cont'd)

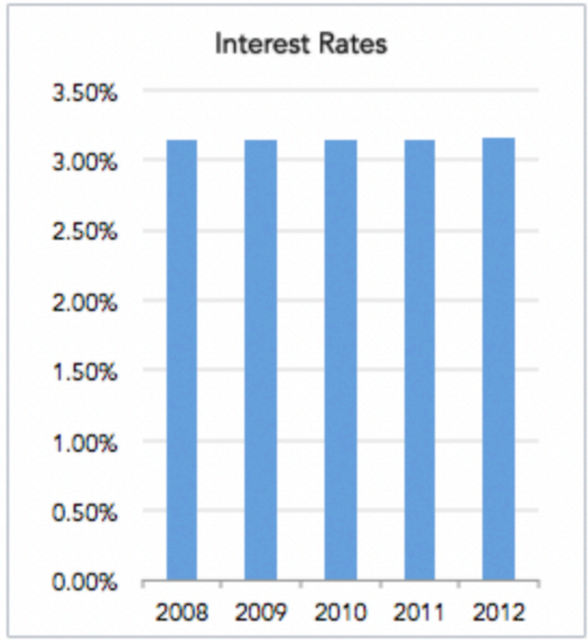
- Question 2: **What is this graph called?**
- Share your answer in the chat



Review Quiz: Data Visualization (cont'd)

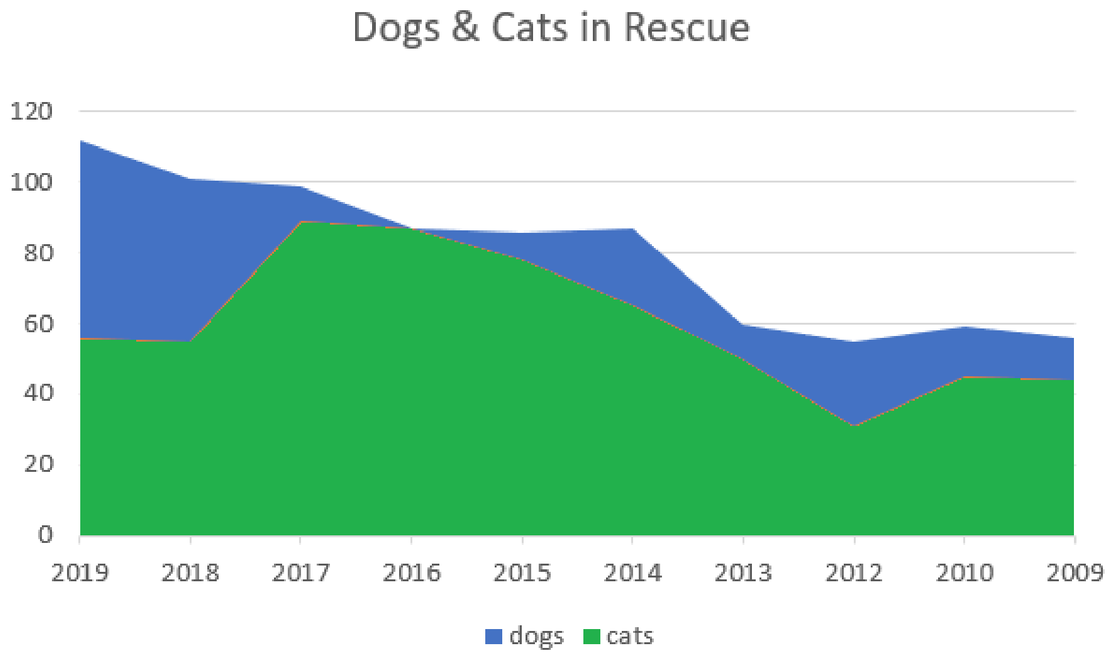
- Question 3: **Which graph represents the values accurately?Why do you think so?**
- Share your answer in the chat





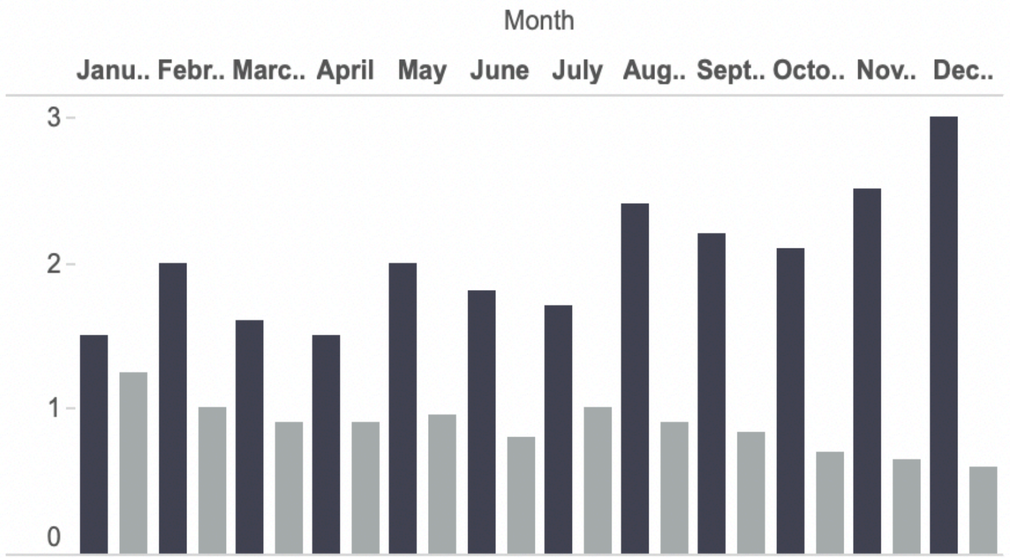
Review Quiz: Data Visualization (cont'd)

- Question 4: **What type of graph is this?**
- Share your answer in the chat

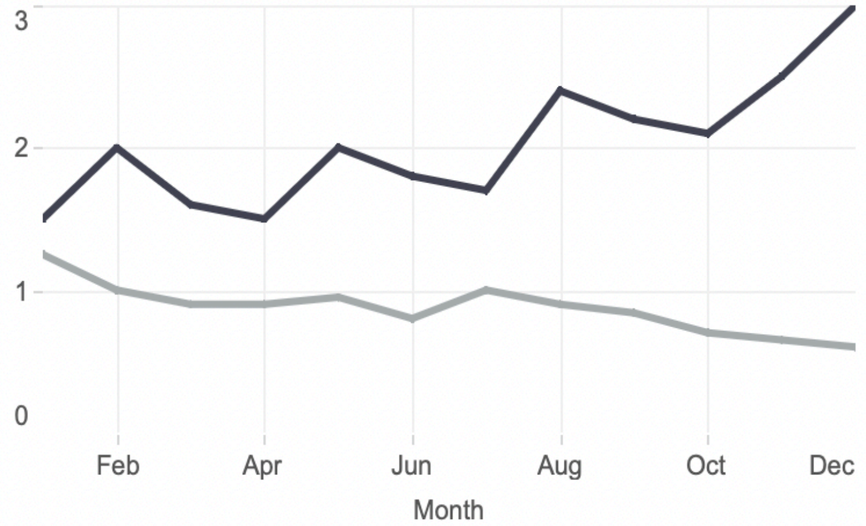


Review Quiz: Data Visualization (cont'd)

- Question 5: **Which of the following graphs focus on trends rather than individual values?**
- Share your answer in the chat



Option A



Option B

Knowledge check

Module completion checklist

Objective	Complete
Discuss data visualization and exploratory data analysis	✔
Describe chart types by data and form	✔

Congratulations on completing this module!