

Introduction to Data Visualization

Data Visualization Course – Lecture 1

Dr. Muhammad Sajjad

R.A: Imran Nawar

September 2024

Overview

➤ Data

- What is data?
- Sources of data?
- Importance of data in the modern world.

➤ Overview of data visualization

➤ Importance of data visualization in data science

- Why visualization matters in data science

➤ Storytelling with data (Communicating Insights Effectively)

➤ Common Types of Data Plots

- Bar chart, line plot, scatter plot, histogram, box plot, pie chart.

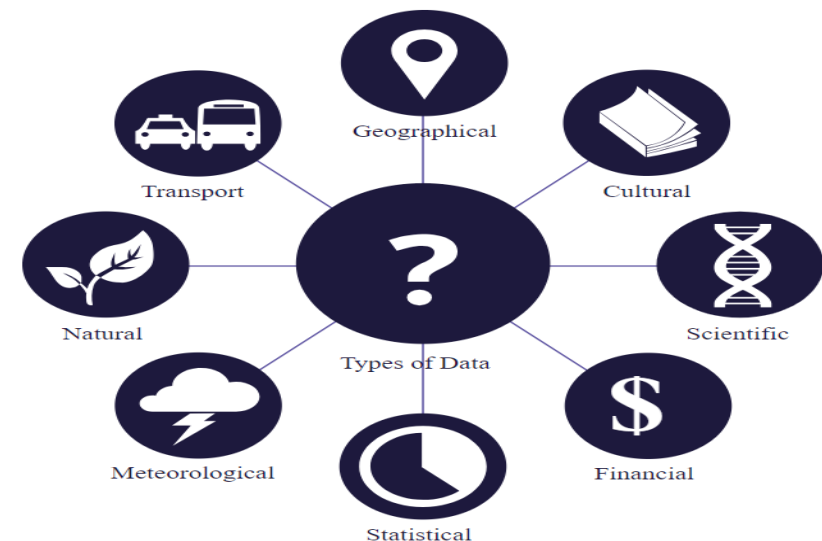
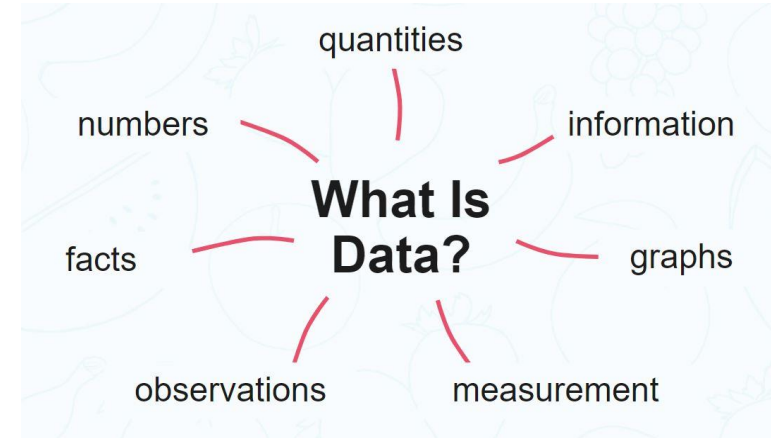
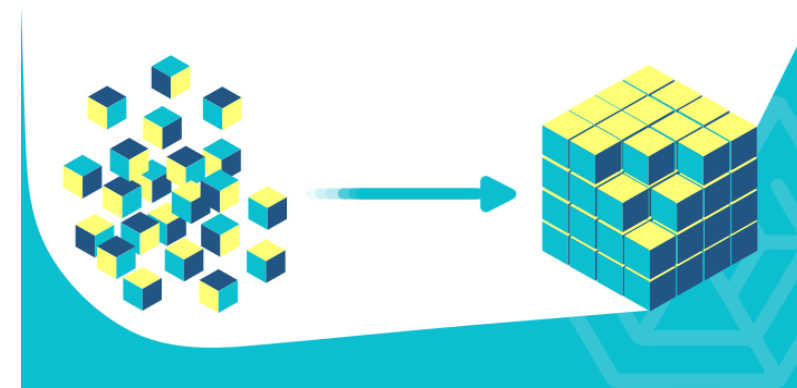
➤ Data Visualization Tools

- Overview of Python libraries for data visualization
- Matplotlib, Pandas, Seaborn, Plotly, Bokeh, Pygal, Folium.

➤ Coding: Python environment and installing libraries.

What is Data?

- Data is information in raw or unorganized form
- A collection of facts, statistics, or information that can be analyzed, processed, and interpreted to derive meaningful insights.
- It's the building block of insights and decisions
- **Data comes in various types:**
 - Numbers
 - Text
 - Images
 - Speech
 - etc.
- **Examples:**
 - Yes, Yes, No, Yes, No, Yes, No, Yes
 - 42, 63, 96, 74, 56, 86
 - None of the above data sets have any meaning until they are given a **CONTEXT** and **PROCESSED** into a useable form
- Information is data that has been processed, organized, or structured in a way that makes it meaningful, valuable and useful.



Sources of Data

1. Primary sources:

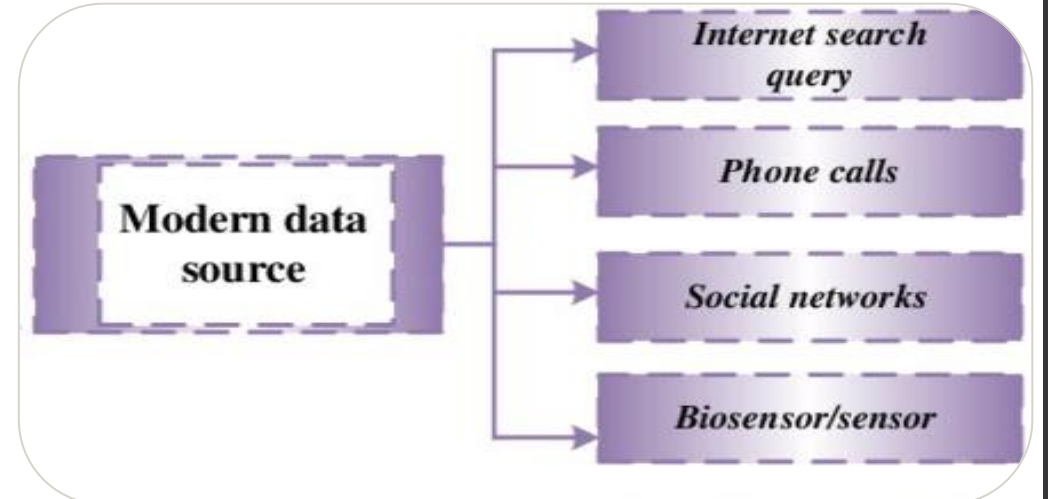
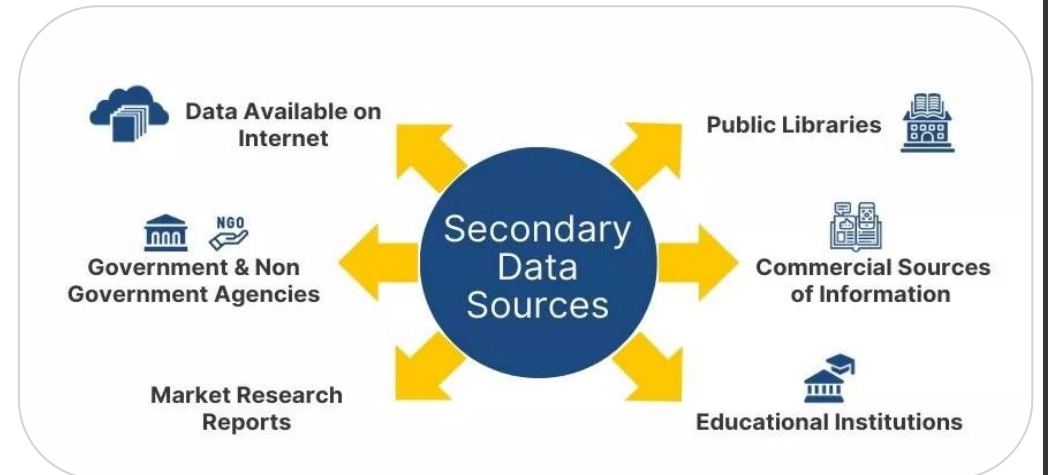
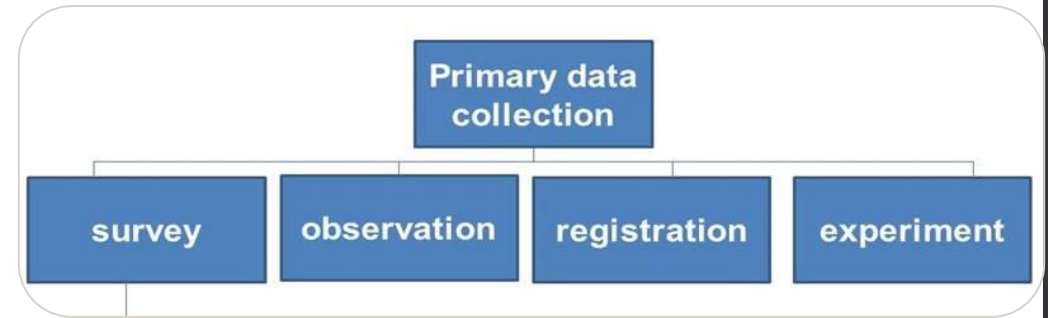
- Surveys and questionnaires
- Experiments and observations

2. Secondary sources:

- Databases: Public or private data collections.
- Government databases (e.g., census data)
- Academic research publications
- Web scraping

3. Modern Data sources:

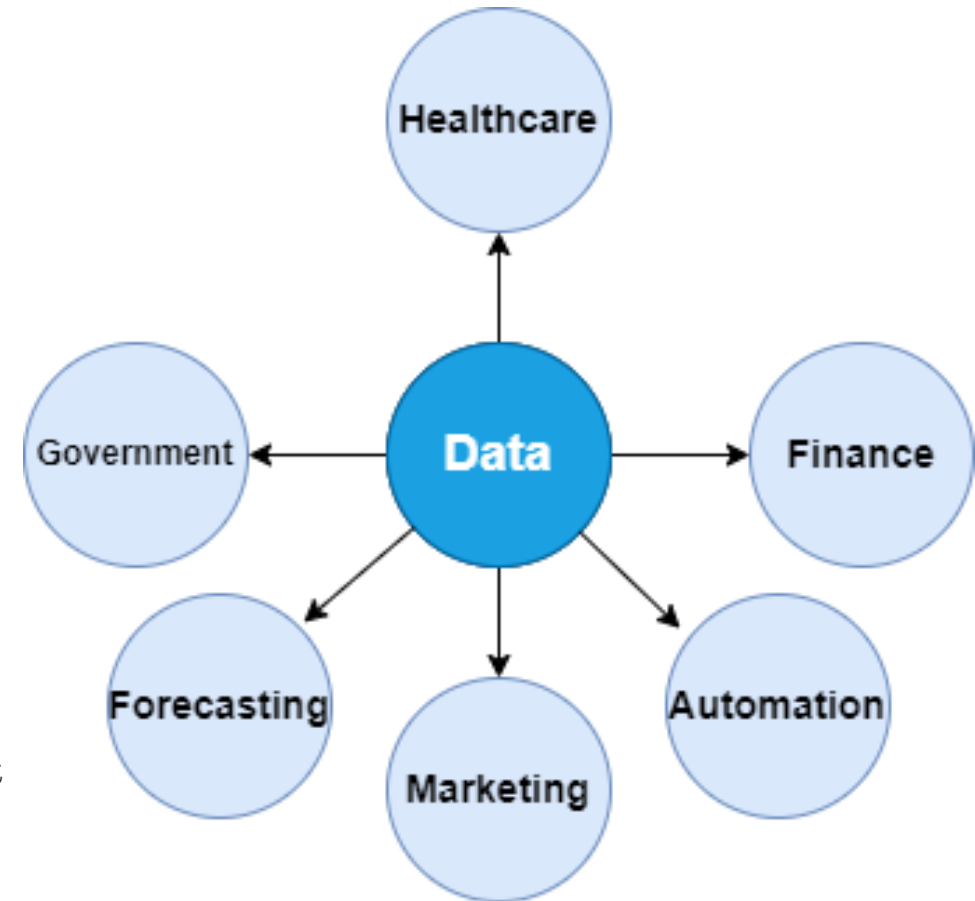
- **Big Data:** Vast datasets from varied sources.
 - User-generated content (social media, online reviews)
 - Business transactions (e-commerce, financial records)
 - Scientific experiments (large-scale research, simulations)
- Sensors and IoT devices (smart cities etc.)
- Surveillance systems (security cameras)
- Medical Data (electronic health records, medical imaging, etc.)



Importance of Data in the Modern World

Why Data Matters?

- **Data-Driven Decision Making:**
 - Essential in businesses, healthcare, finance, etc.
 - Helps make better decisions.
- **Powering AI and Machine Learning:**
 - Foundation for training models and making predictions.
 - Enables automation in various applications.
- **Innovation and Improvement:**
 - Facilitates advancements in technology and science
 - Enables performance evaluation
- **Enhancing Understanding:**
 - Provides insights into consumer behavior and market trends.
 - Helps solve complex challenges



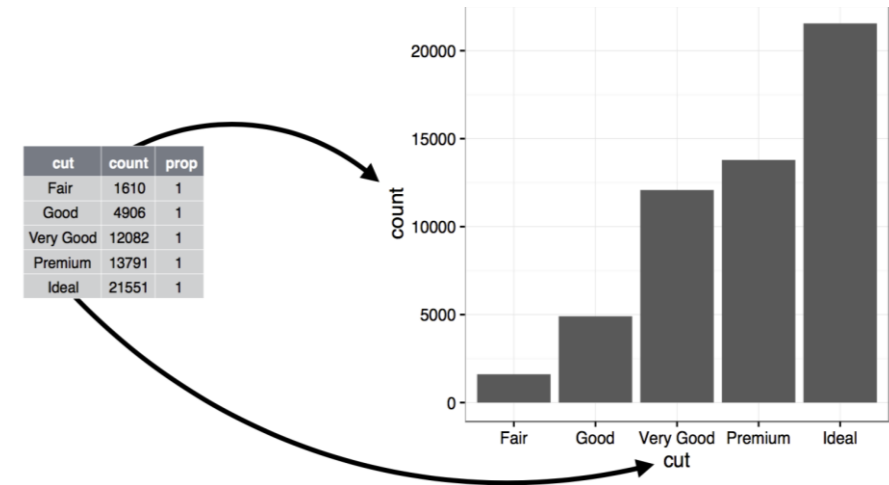
Overview of Data Visualization

What is Data Visualization?

- The graphical representation of information and data using visual elements such as charts, graphs, and maps.
- Data visualization is not limited to just the graphical representation of data. While visual representations like charts, graphs, and plots are a key part of it, data visualization also involves:
 - Simplifies complex data: Organizes large datasets for easier understanding.
 - Communicates insights: Reveals hidden trends, patterns, and correlations.
 - Facilitates decision-making: Aids in interpreting data for informed analysis.
- **Purpose:**
 - Helps in understanding complex data through visual context.
 - Visualizing data allows people to see relationships, patterns, and trends in the information you're trying to communicate.
- **Examples:**
 - **Charts:** Bar charts, pie charts, etc.
 - **Maps:** Geospatial data visualizations.
 - **Dashboards:** Interactive data summaries.

Recent trends:

- Augmented and virtual reality (AR/VR) visualizations
- Real-time and streaming data visualizations
- AI-assisted data visualization tools



Importance of Data Visualization in Data Science

1. **Trend Analysis:** Identify patterns over time.
2. **Outlier Detection:** Easily spot anomalies.
3. **Insight Communication:** Simplifies complex data for stakeholders.
4. **Decision support:** Enabling quick and informed choices based on visual insights
5. **Pattern recognition:** Identifying trends and anomalies that may be missed in raw data.

Advantages of Data Visualization

Helps in understanding data

01

Facilitates communication

02

Enables pattern recognition

03

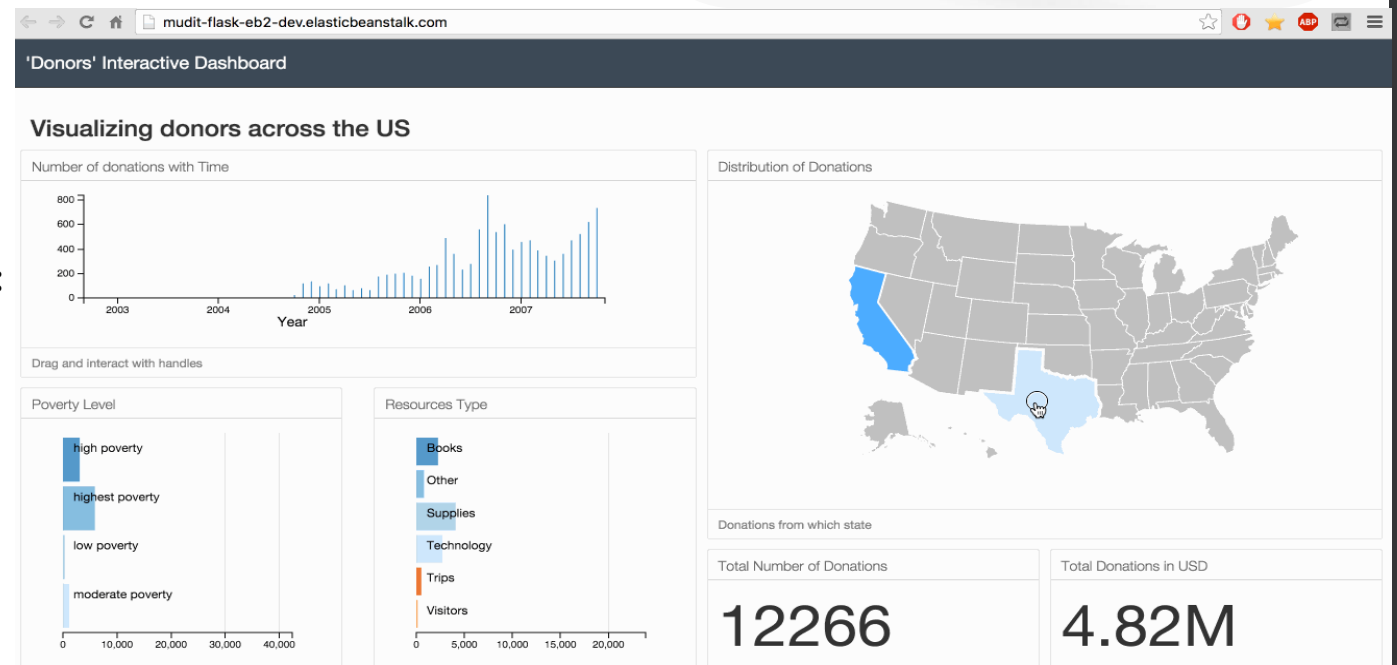
Facilitates data exploration

04

Enhances decision-making

05

Example of an interactive dashboard:



Storytelling with Data: Communicating Insights Effectively

Definition:

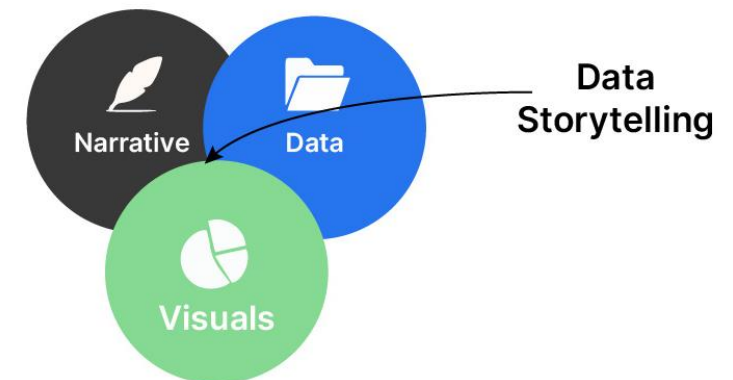
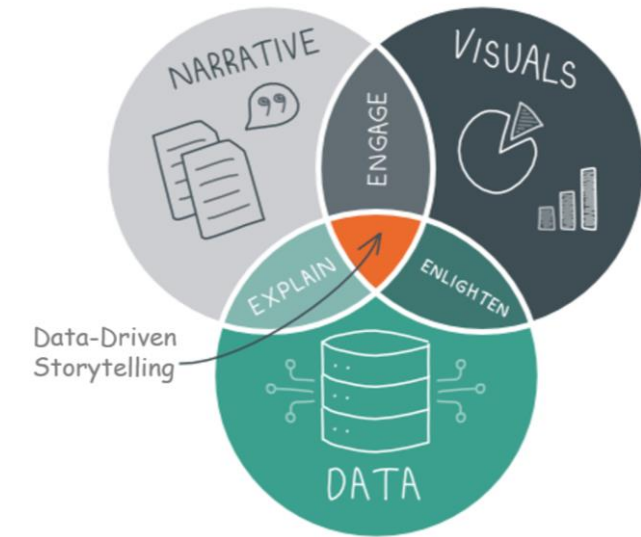
- Data storytelling is the practice of creating a compelling narrative using data to convey insights clearly and effectively.
- It goes beyond mere visualization, aiming to make data meaningful and impactful.

Purpose:

- To make complex information accessible, understandable, and memorable to diverse audiences.
- To bridge the gap between data analysis and decision-making.

Key Elements:

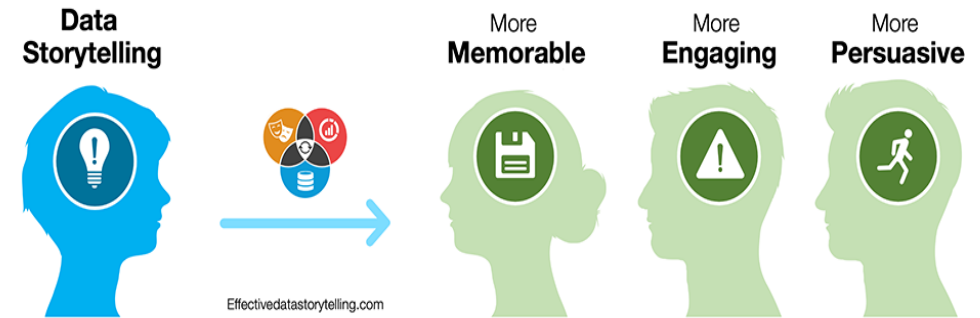
1. **Data:** The foundation of the story, providing factual content.
2. **Visuals:** Graphical representations that helps illustrate the data.
3. **Narrative:** The context and explanation that connects the data and visuals, providing meaning and insights.



The Importance of Data Storytelling

Why Data Storytelling Matters:

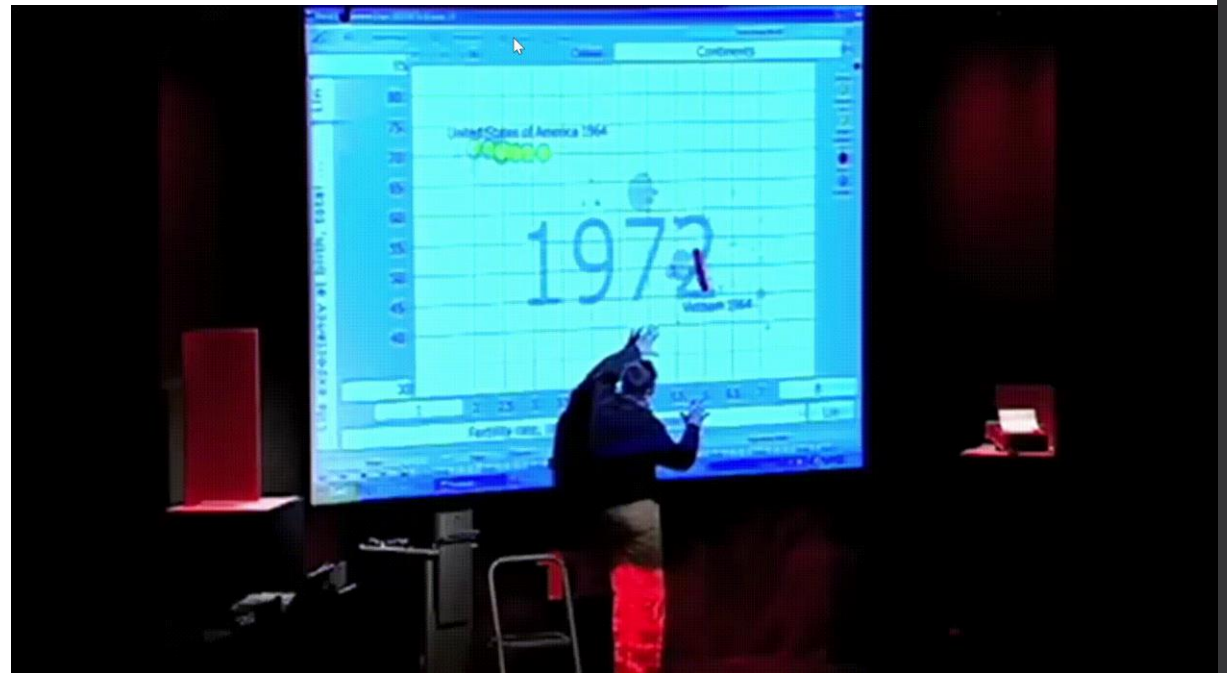
- **Enhances Understanding:** Transforms complex data into a narrative, making it easier to comprehend.
- **Engages the Audience:** Captivates and holds attention, making data more memorable.
- **Bridges Data and Decision-Making:** Makes complex information accessible, facilitating informed decisions.
- **Drives Action:** Encourages change by presenting insights in an impactful way.
- **Improves Retention:** Helps people remember data through compelling stories.



Real-World Example:

Hans Rosling's TED Talk: [Link](#)

- Lecture, “The Best Stats You’ve Ever Seen”
- Used animated bubbles to show global health trends over time.
- Made complex demographic data engaging and understandable.

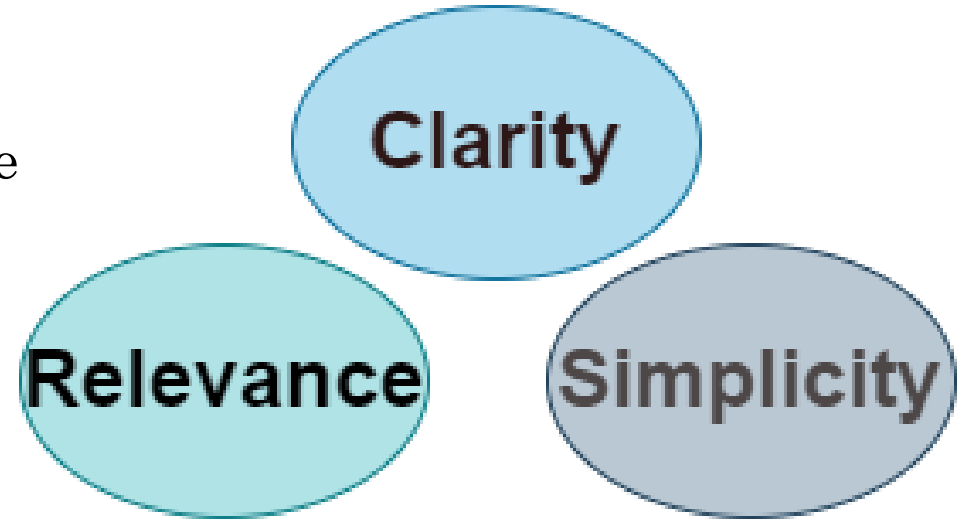


Principles of Good Data Storytelling

- **Clarity:** Make the message easy to understand
- **Simplicity:** Avoid unnecessary complexity
- **Relevance:** Focus on what matters to the audience

Steps in Data Storytelling:

1. **Collect Data:** Gather relevant data points
2. **Analyze Data:** Identify key insights
3. **Visualize Insights:** Use visual aids to convey the story



Data Storytelling vs. Data Visualization

Data Visualization:

- Graphical representation of data (charts, graphs, maps)
- Aims to present information efficiently and clearly.
- Enables quick grasp of patterns and trends

Data Story Telling:

- Uses narratives, context, and visuals to communicate insights.
- Aims to engage audience emotionally and make data memorable
- Follows a narrative structure (beginning, middle, end)

➤ Key Difference:

- Data visualization is a tool used within data storytelling.
- Data storytelling goes beyond to create a compelling narrative



Applications of Data Storytelling

- **Decision-Making**

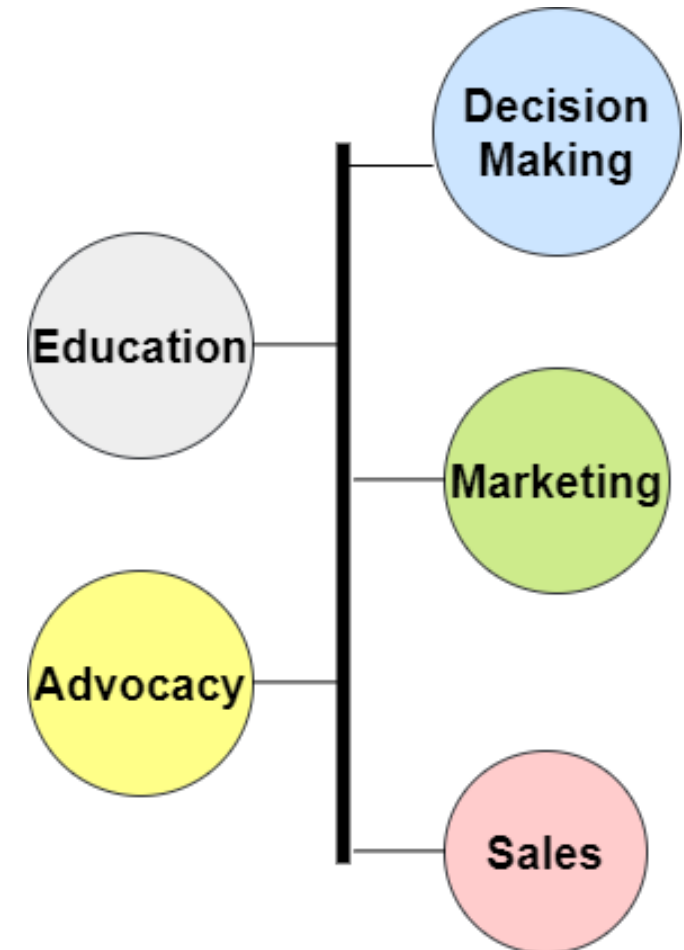
- Presenting data in a narrative format, enabling decision makers to easily grasp the significance of the data and understand its implications.

- **Business & Advocacy**

- Marketing: Communicates product value
- Sales: Demonstrates service impact
- Advocacy: Illustrates cause importance

- **Education & Training**

- Engages students and participants
- Improves grasp of key insights
- Enhances information retention

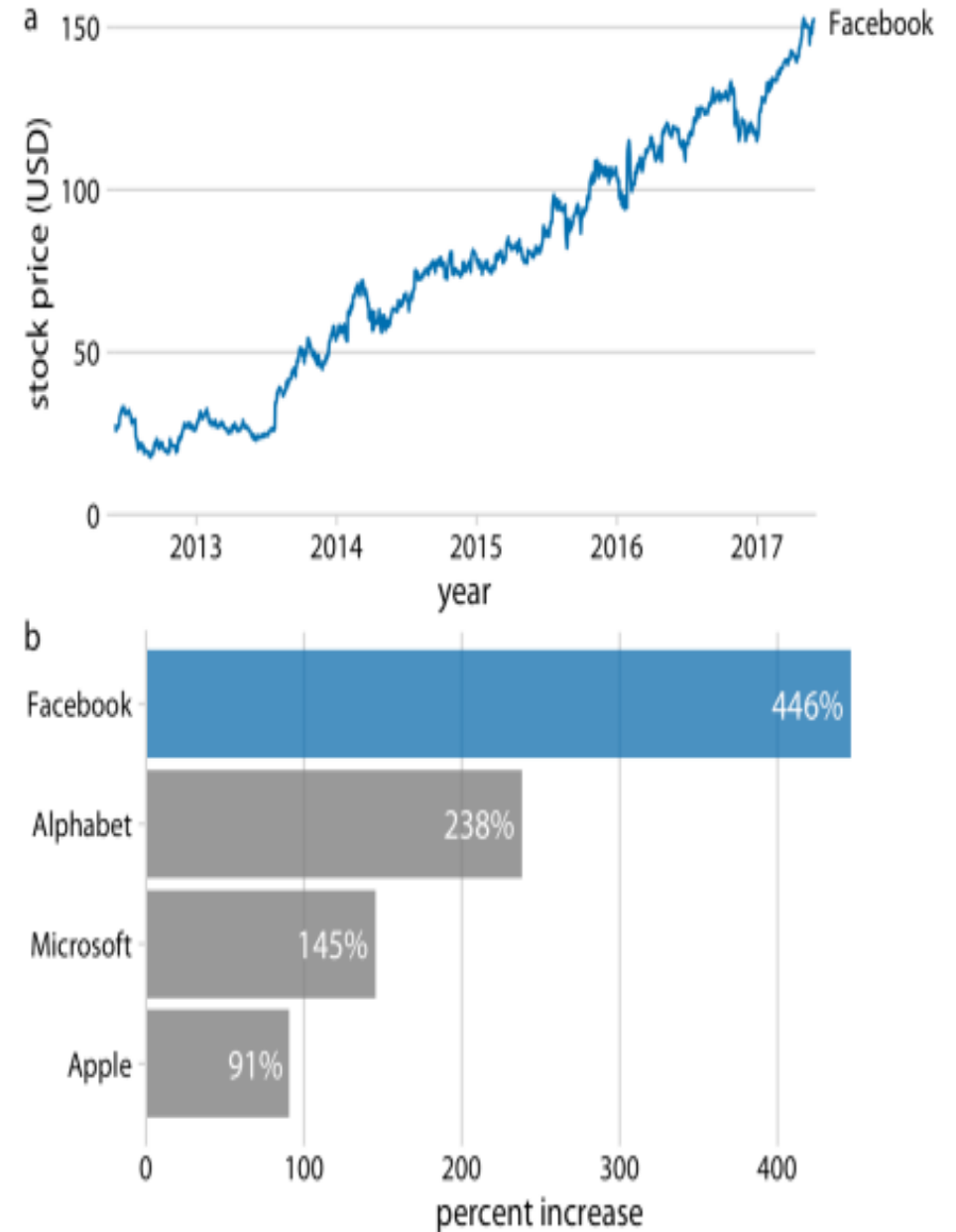


Case Studies

1. Facebook Stock Growth

Growth of Facebook stock price over a five-year interval and comparison with other tech stocks.

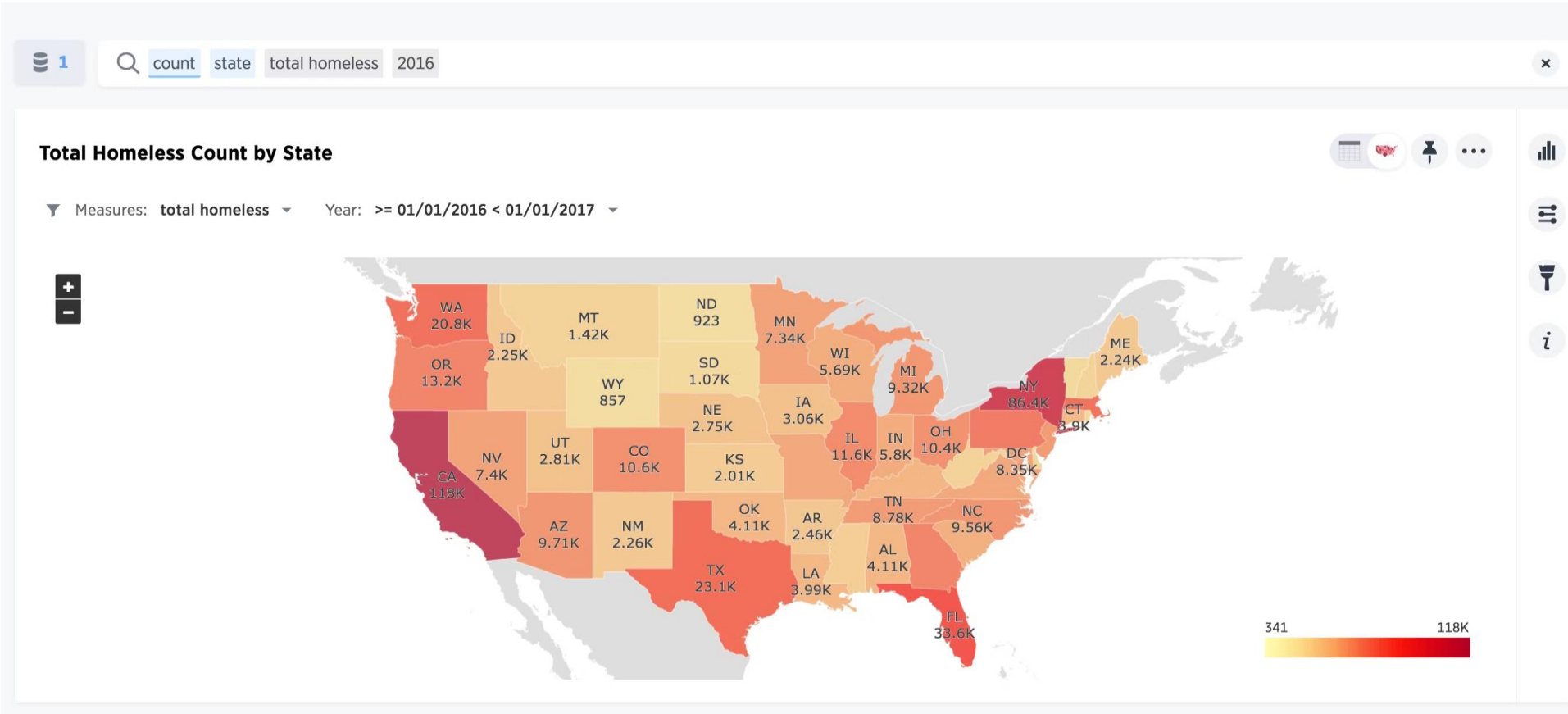
- (a) The Facebook stock price rose from around \$25/share in mid-2012 to \$150/share in mid-2017, an increase of almost 450%.
- (b) The prices of other large tech companies did not rise comparably over the same time period. Price increases ranged from around 90% to almost 240%. Data source: Yahoo! Finance.



Case Studies

2. Homeless in America

1. **Visualization:** Map of the US with color-based density indicators
2. **Impact:** Quickly shows where homelessness is most acute across the country
3. **Benefit:** Simplifies complex data for easy understanding without technical expertise

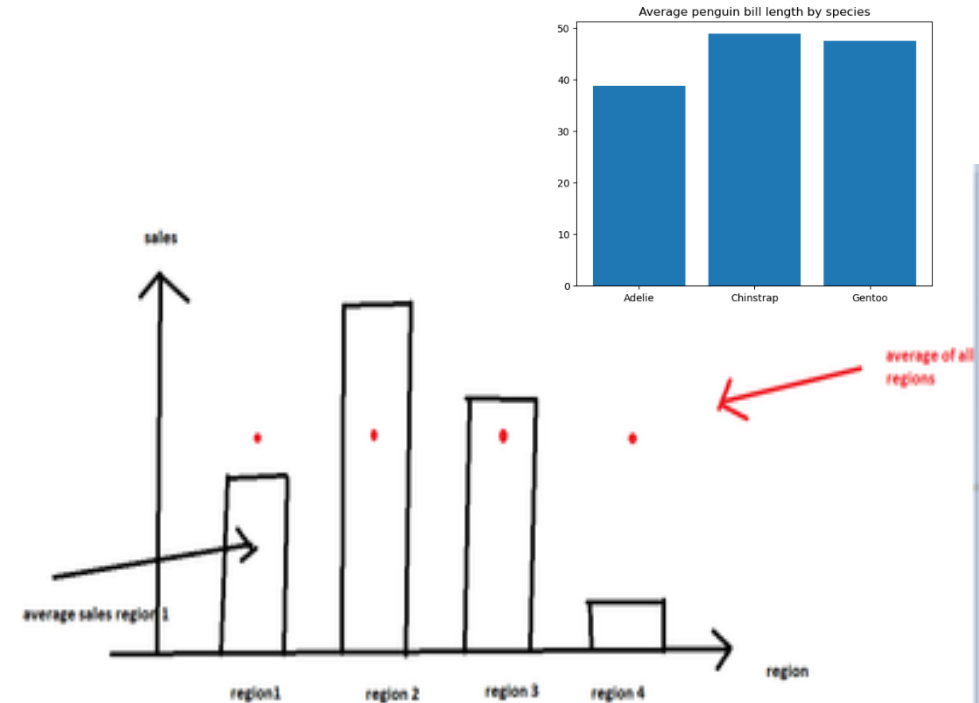


Common Types of Data Plots

Creating data plots is an essential step of exploratory data analysis

Bar Chart

- Use Case: Comparing quantities across categories.
- Example: Sales data across different regions.
- **Features:**
 - Easy to compare different categories side by side.
 - Useful for categorical data.



Line Plot

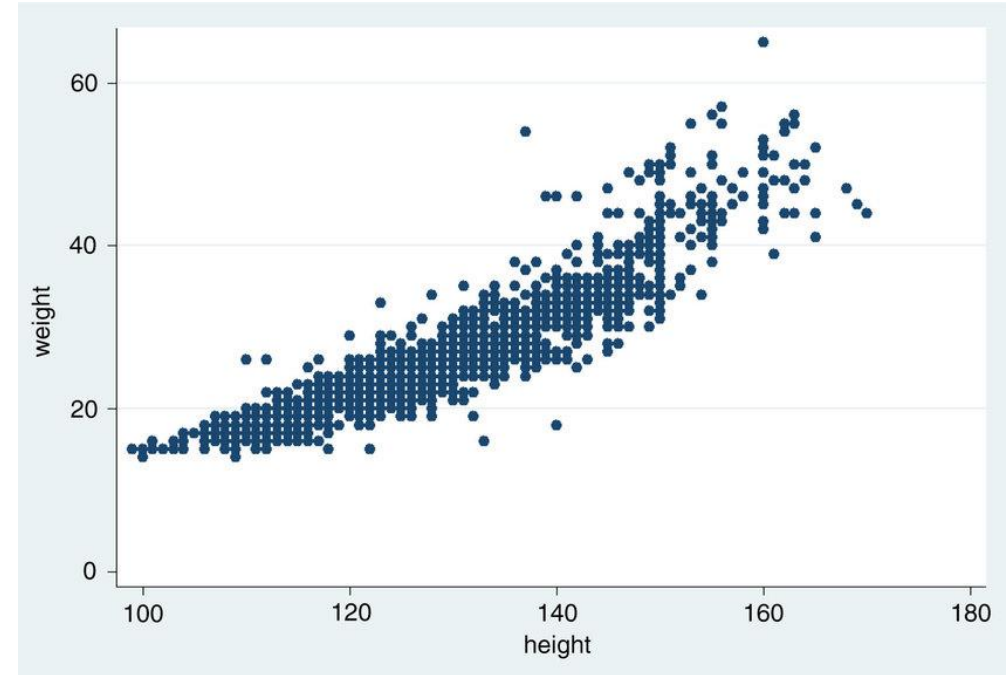
- Use Case: Show trends over time.
- Example: Stock price changes over a month.
- **Features:**
 - Illustrates continuous data and trends.
 - Ideal for time series data.



Common Types of Data Plots

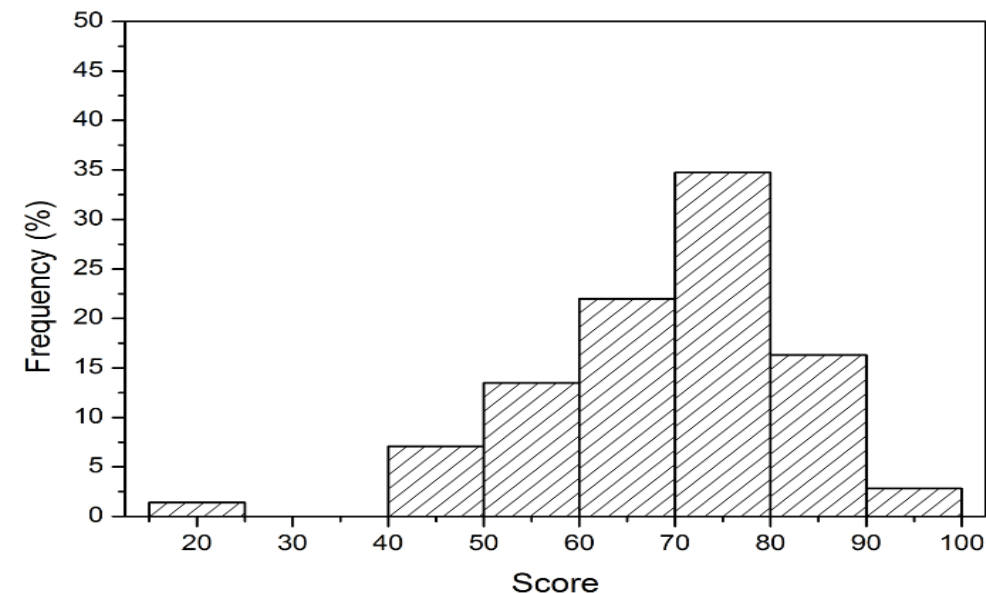
Scatter Plot

- Use Case: Show relationships between two variables
- Example: Height vs. weight in a population.
- **Features:**
 - Helps identify correlations, patterns, and outliers.
 - Useful for regression analysis.



Histogram

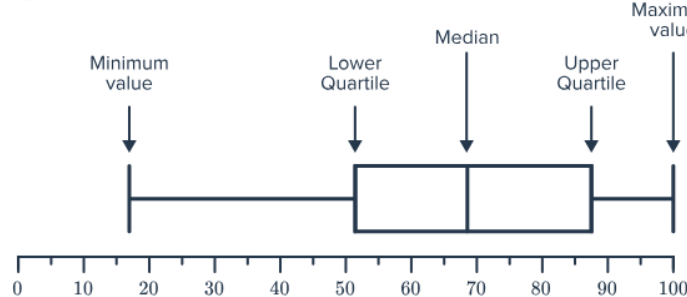
- Use Case: Display the frequency distribution of a variable.
- Example: Showing the frequency distribution of test scores in a class
- **Features:**
 - Provides insight into the spread and shape of data distribution.
 - Helps identify patterns such as skewness or normality.



Common Types of Data Plots

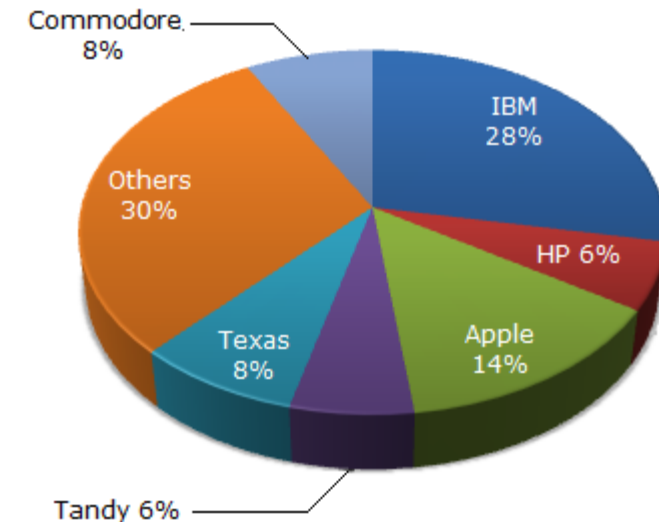
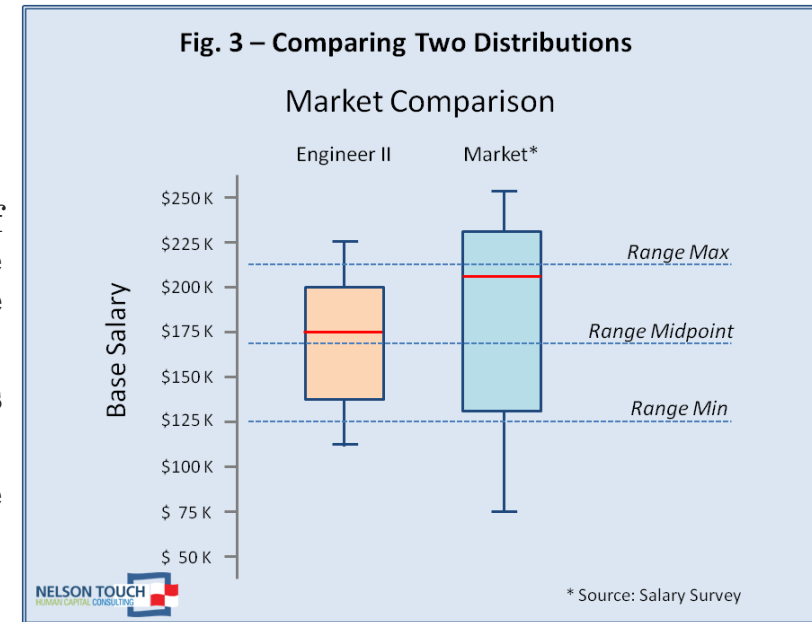
Box Plot (box-and-whisker plot)

- **Use Case:** Display the distribution and outliers of data.
- A box plot is a data plot type that shows a set of five descriptive statistics of the data: the minimum and maximum values (excluding the outliers), the median, and the first and third quartiles. Optionally, it can also show the mean value.
- Interquartile Range (IQR): The distance between the first and third quartiles ($Q3 - Q1$), representing the spread of the middle 50% of the data.
- Outliers: Data points that fall significantly outside the IQR (typically more than 1.5 times the IQR from $Q1$ or $Q3$).
- **Example:** Distribution of salaries in a company.



Pie Chart

- **Use Case:** Show proportions within a whole.
- A pie chart is a type of data visualization represented by a circle divided into sectors, where each sector corresponds to a certain category of the categorical data, and the angle of each sector reflects the proportion of that category as a part of the whole
- **Example:** Market share of companies.



The pie chart shows the distribution of New York market share by value of different computer companies in 2005.

Data Visualization Tools

Overview of Python Libraries for Data Visualization

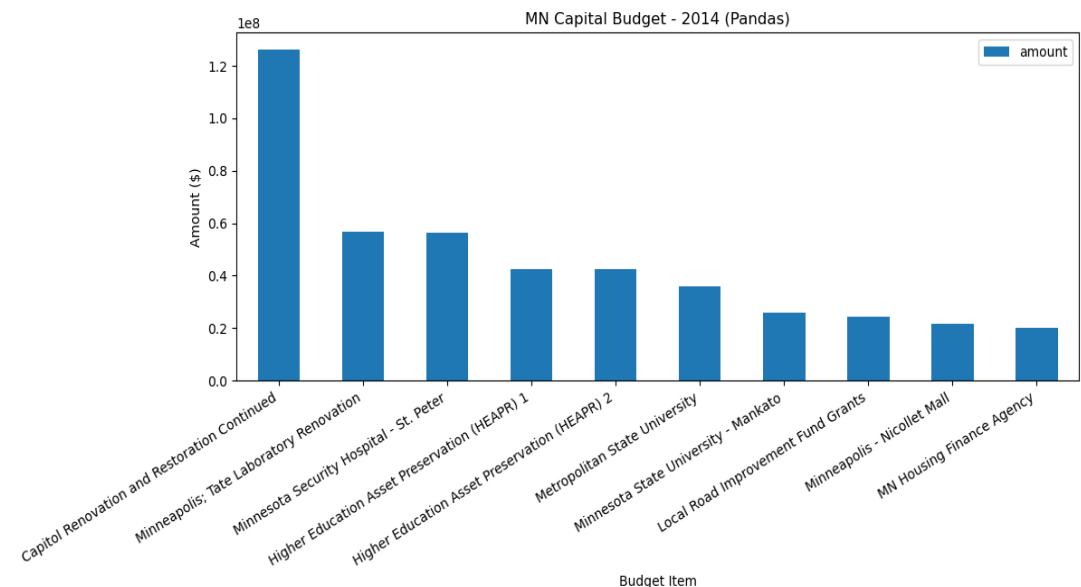
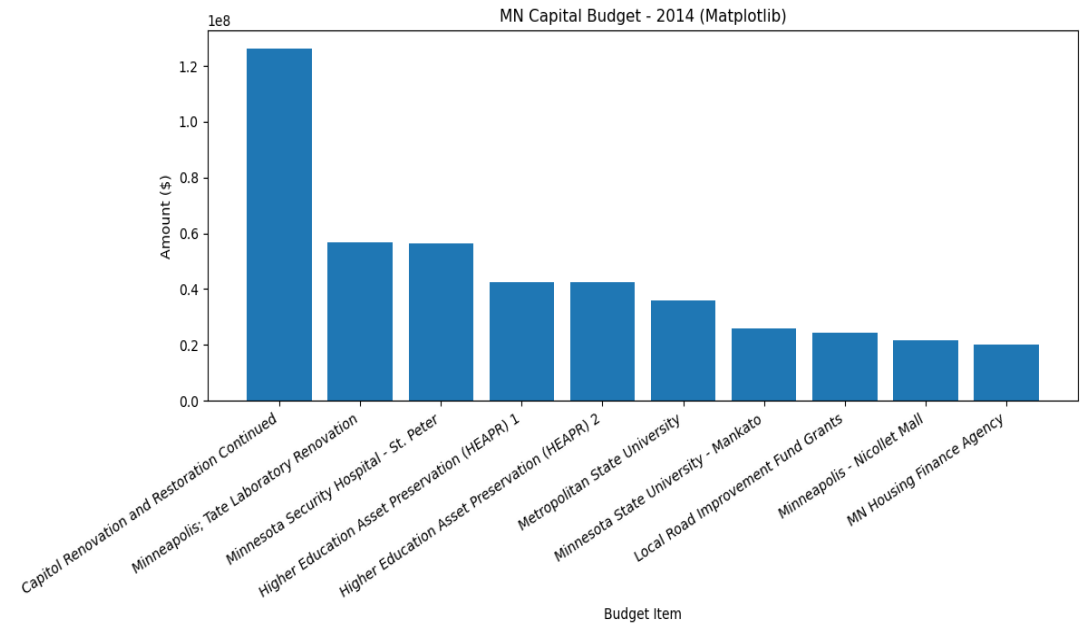
- Python has become a dominant language in data science, offering a rich ecosystem of visualization libraries.

1. Matplotlib

- Matplotlib is the grandfather of python visualization packages.
- Foundation for many other libraries.
- It is extremely powerful but with that power comes complexity.
- Best for creating publication-quality static visualizations.
- **Example:** Line plots, bar charts.

2. Pandas

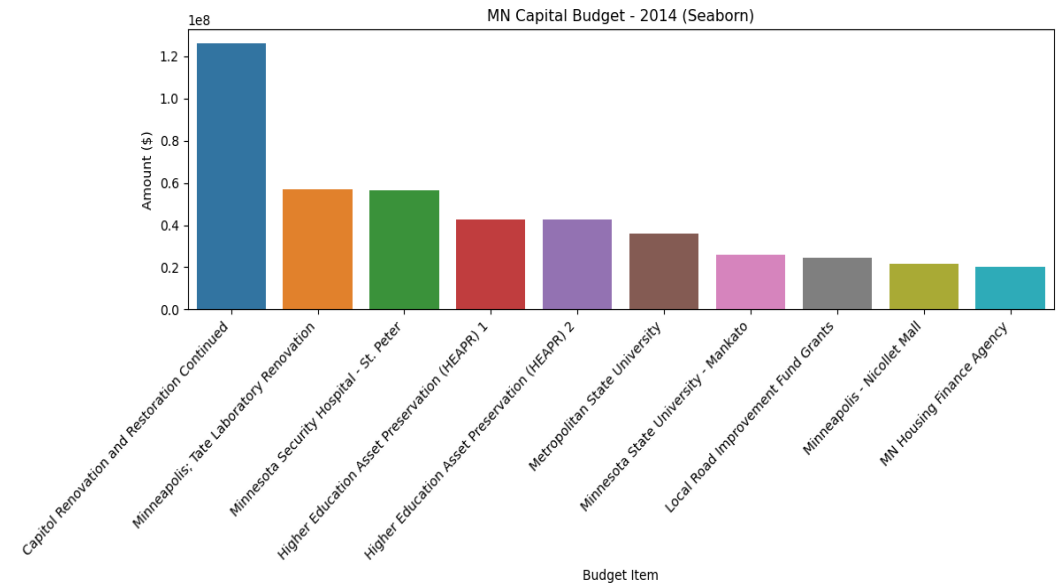
- Built-in plotting functions based on Matplotlib.
- **Features:**
 - Convenient for quick visualizations directly from DataFrames.
 - Ideal for simple, exploratory plots.
- **Use Case:** Rapid, straightforward plotting.
- **Example:** Quick plots from Dataframes.



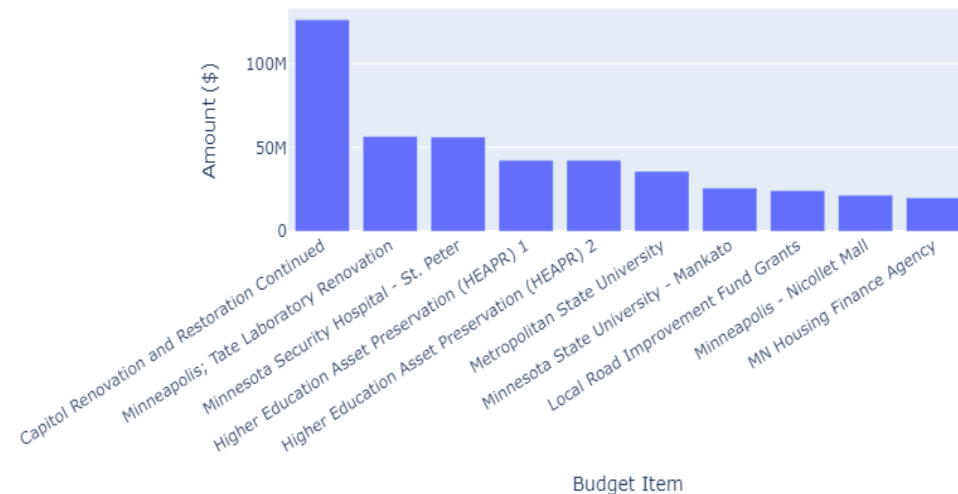
Overview of Python Libraries for Data Visualization

3. Seaborn

- Visualization library based on Matplotlib.
- **Features:**
 - Leverages matplotlib for beautiful, minimal-code charts.
 - It's default styles and color palettes are more modern and visually appealing.
 - It also has the goal of making more complicated plots simpler to create.
 - Specialized in statistical visualizations.
- **Example:** Heatmaps, pair plots.



MN Capital Budget - 2014 (Plotly)



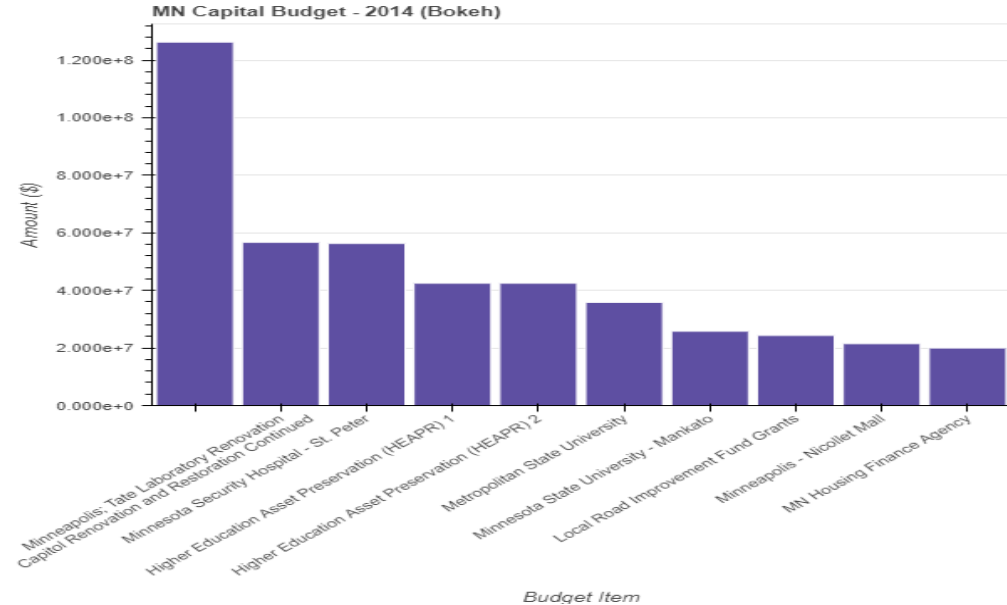
4. Plotly

- **Overview:** Creates interactive, web-based visualizations.
- **Features:**
 - Supports diverse chart types and maps.
 - Ideal for dashboards and web applications.
 - Handles real-time data and complex interactivity.
- **Example:** 3D plots, interactive dashboards.

Overview of Python Libraries for Data Visualization

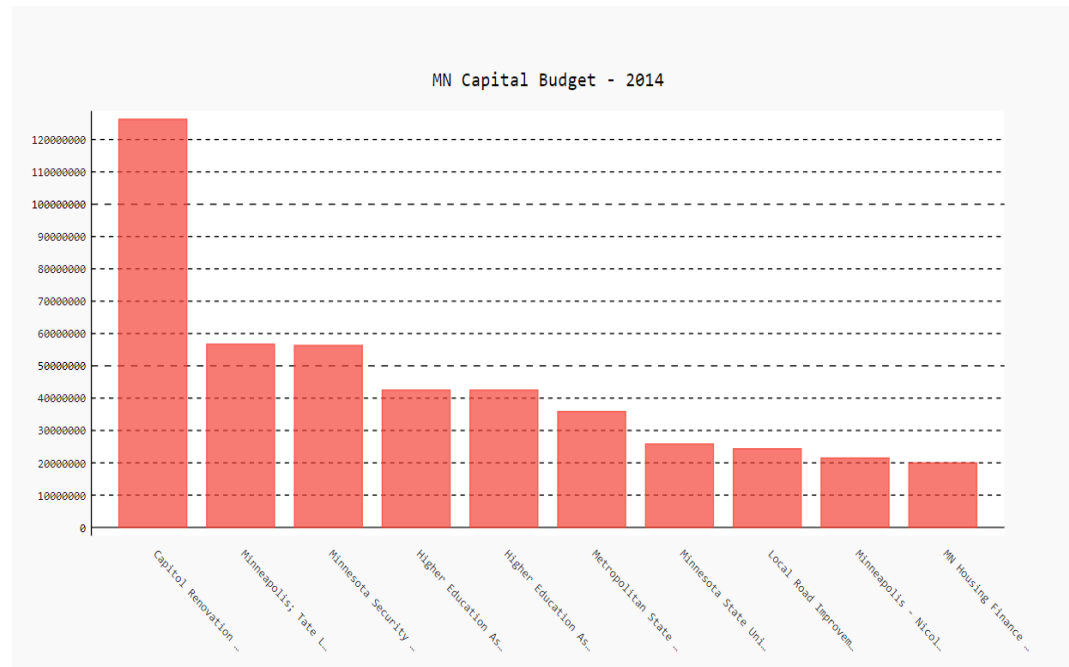
5. Bokeh

- **Overview:** Python library for interactive plots and applications.
- **Use Case:** Web-based visualizations and large datasets.
- **Features:** Supports streaming and real-time data.
- **Example:** Interactive web-based visualizations.



6. Pygal:

- **Overview:** Python library for creating SVG charts.
- **Features:** Generates interactive, high-quality vector graphics.
- **Use Case:** Ideal for producing visually appealing and scalable charts.
- **Example:** Interactive and customizable charts for web applications.

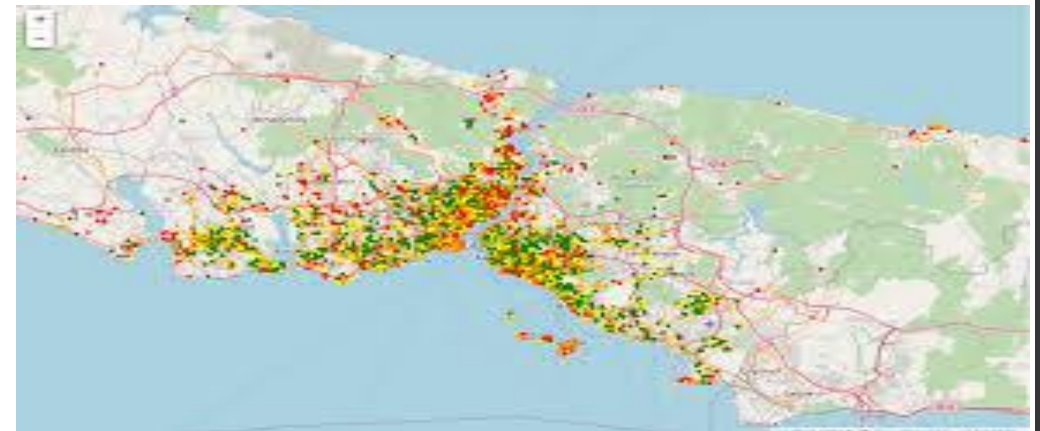
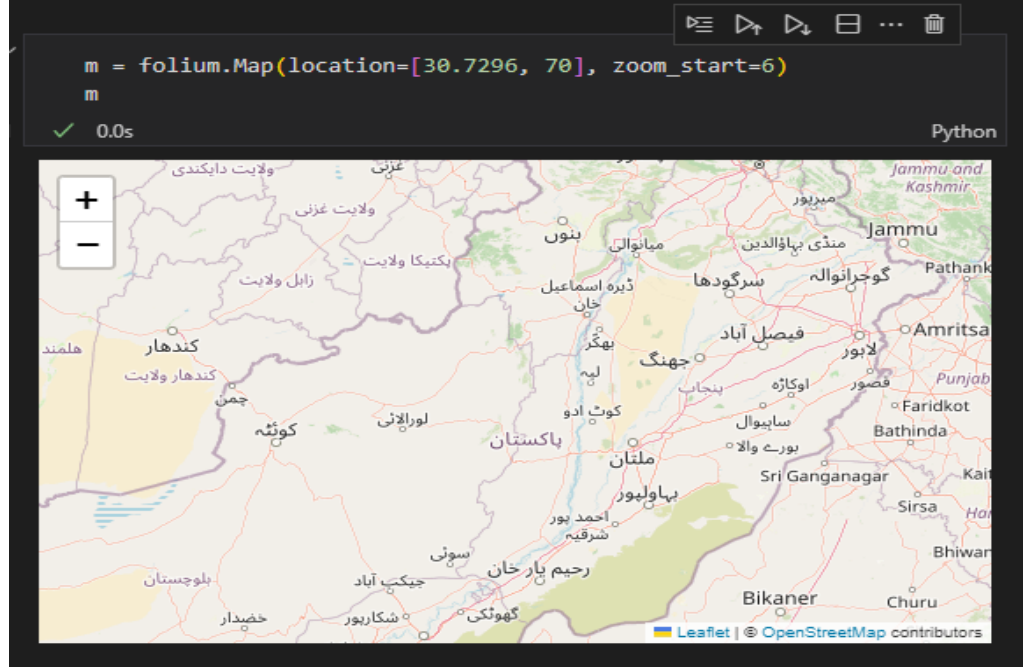


Overview of Python Libraries for Data Visualization

7. Folium

- **Overview:** Python library for interactive maps.
- **Backend:** Utilizes Leaflet.js (a popular JavaScript library for interactive maps).
- **Use Case:** Geospatial data visualization.
- **Features:** Specializes in creating interactive, mobile-friendly maps.
- **Example:** Visualizing geographic data interactively.

6. Folium (Map visualization)



Coding: Python Environment Setup

To get started with data visualization in Python, follow these steps:

1. **Install Python:** Download and install the latest version of Python from python.org.

2. Set up a virtual environment (optional but recommended):

```
python -m venv data_viz_env
data_viz_env\Scripts\activate
```

3. Set up a coding environment:

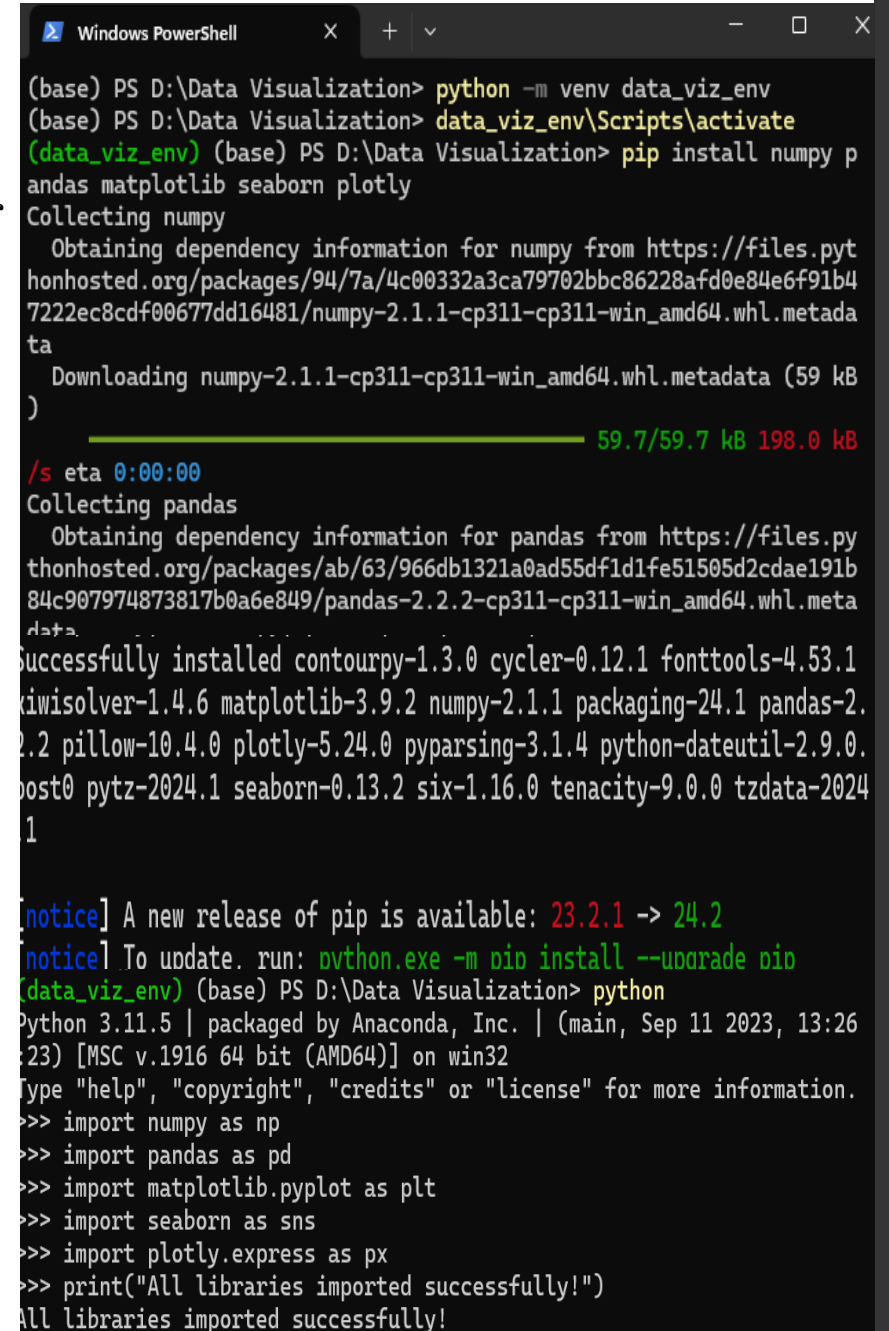
- **Jupyter Notebook:** Ideal for interactive coding.
- **VS Code / PyCharm:** Full featured IDEs.

4. Install essential libraries:

```
pip install numpy pandas matplotlib seaborn plotly
```

5. Verify installations:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
print("All libraries imported successfully!")
```



```
(base) PS D:\Data Visualization> python -m venv data_viz_env
(base) PS D:\Data Visualization> data_viz_env\Scripts\activate
(data_viz_env) (base) PS D:\Data Visualization> pip install numpy pandas matplotlib seaborn plotly
Collecting numpy
  Obtaining dependency information for numpy from https://files.pythonhosted.org/packages/94/7a/4c00332a3ca79702bbc86228afd0e84e6f91b47222ec8cdf00677dd16481/numpy-2.1.1-cp311-cp311-win_amd64.whl.metadata
  Downloading numpy-2.1.1-cp311-cp311-win_amd64.whl.metadata (59 kB)
/s eta 0:00:00
Collecting pandas
  Obtaining dependency information for pandas from https://files.pythonhosted.org/packages/ab/63/966db1321a0ad55df1d1fe51505d2cdae191b84c907974873817b0a6e849/pandas-2.2.2-cp311-cp311-win_amd64.whl.metadata
Successfully installed contourpy-1.3.0 cyclo-0.12.1 fonttools-4.53.1 kiwisolver-1.4.6 matplotlib-3.9.2 numpy-2.1.1 packaging-24.1 pandas-2.2 pillow-10.4.0 plotly-5.24.0 pyparsing-3.1.4 python-dateutil-2.9.0 pytz-2024.1 seaborn-0.13.2 six-1.16.0 tenacity-9.0.0 tzdata-2024.1
[notice] A new release of pip is available: 23.2.1 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip
(data_viz_env) (base) PS D:\Data Visualization> python
Python 3.11.5 | packaged by Anaconda, Inc. | (main, Sep 11 2023, 13:26:23) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> import numpy as np
>>> import pandas as pd
>>> import matplotlib.pyplot as plt
>>> import seaborn as sns
>>> import plotly.express as px
>>> print("All libraries imported successfully!")
All libraries imported successfully!
```


Summary and Next Steps

➤ Key Takeaways:

- Understanding the types and sources of data.
- Recognizing the importance of data visualization in data science.
- Understanding how storytelling with data enhances communication and decision-making.
- Familiarity with common types of plots and their applications.
 - Bar, line, scatter, histogram, box, pie charts.
- Introduction to essential Python libraries for data visualization.
 - Matplotlib, Pandas, Seaborn, Plotly, Bokeh, etc.

➤ Next Lecture: Visualizing Numeric and Categorical Data.



Thank You