WOMEN WHO CODE®
/data-science

Build a Data Science Project from Scratch - SESSION 2

📅 JUNE 16
10 PM ET

Kirthikka Devi Venkataram
Product Manager

womenwhocode.com/datascience/events

Session Agenda

- Why is EDA & Visualization important?
- EDA Tools & Libraries
- Visual Representations
- Other Visualization Tools
- Best practices

Why is EDA important?

EDA - Exploratory Data Analysis & Visualization

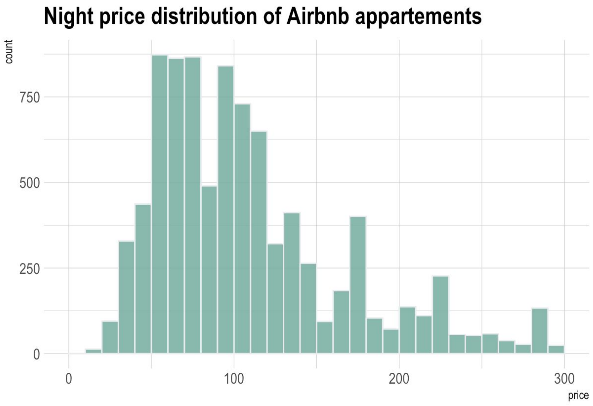
- Explore and get a sense of the size, shape and structure of the data
- Uncover trends and important insights such as outliers, anomalies
- Presents an understanding of underlying **PATTERNS**, **DISTRIBUTION** and **RELATIONSHIP** between variables
- Insights from EDA informs the next steps in a Data Science project

EDA Tools & Libraries 📊

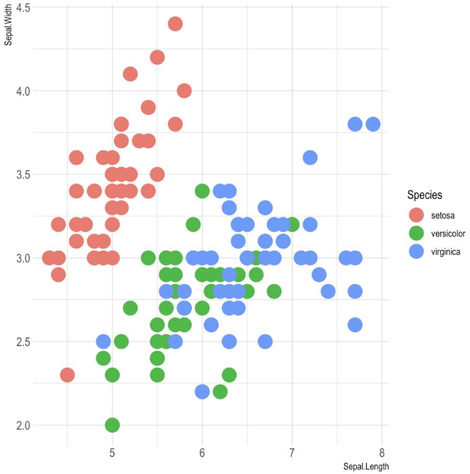
- ***Pandas*** - a library for data manipulation and analysis of structured data
- ***NumPy*** - a library offering efficient multi-dimensional array operations and mathematical functions
- ***SciPy*** - a library for statistical analysis, optimization, and integration.
- ***Matplotlib*** - a 2D plotting library for creating static, interactive, and animated visualizations in Python.
- ***Seaborn*** - A data visualization library built on top of Matplotlib, for drawing attractive and informative plots.
- ***Plotly*** - an interactive graphing library, enabling the creation of visually appealing, interactive, and web-based charts and plots.
- ***SciKit-Learn*** - a machine learning library with built-in tools for preprocessing, feature selection, and dimensionality reduction.

Visual Representations

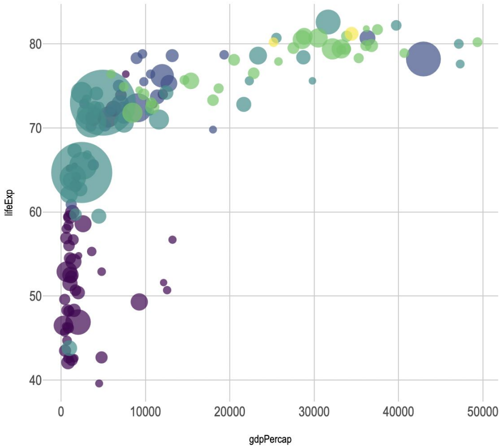
Basic Plots



Histograms



Scatterplots

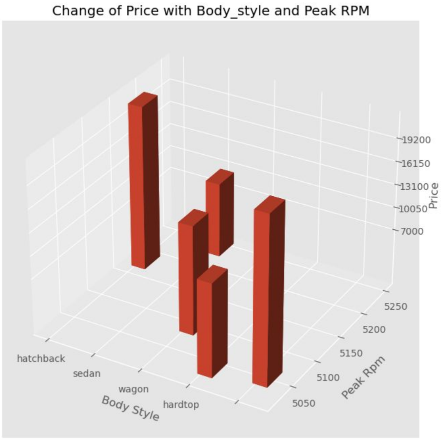


Bubble Charts

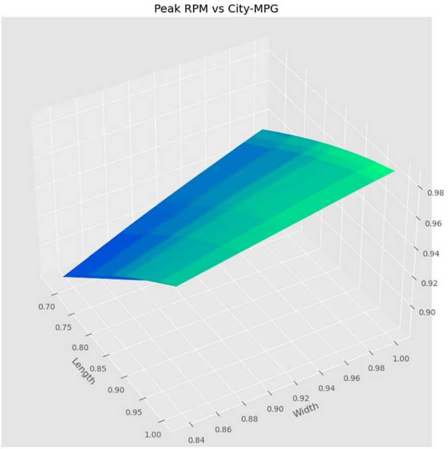
Image Source: https://www.data-to-viz.com

Visual Representations

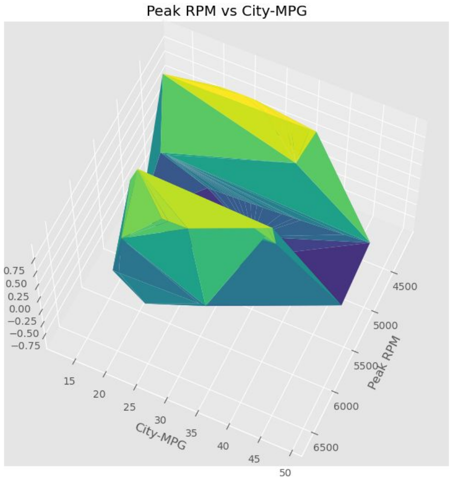
Advanced Plots



Bar Plot



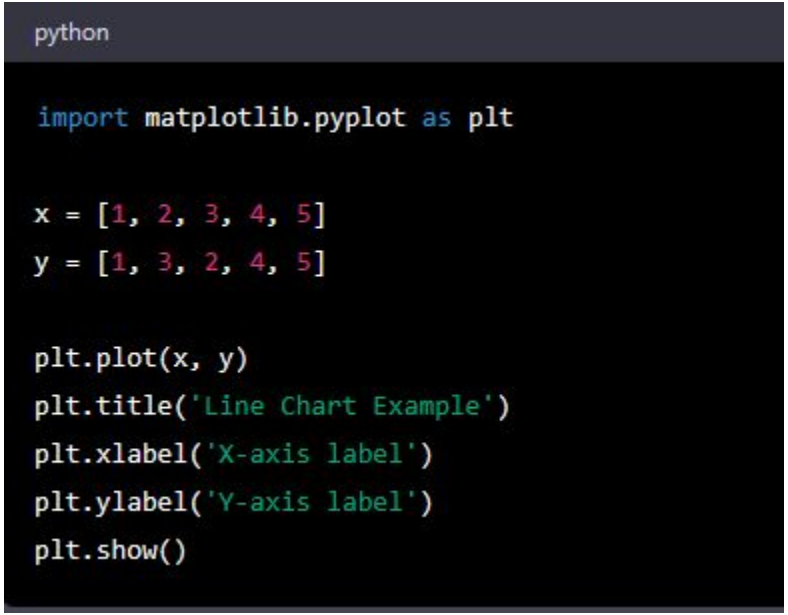
Surface Plot



Tri-Surf Plot

Image Source: https://www.regenerativetoday.com

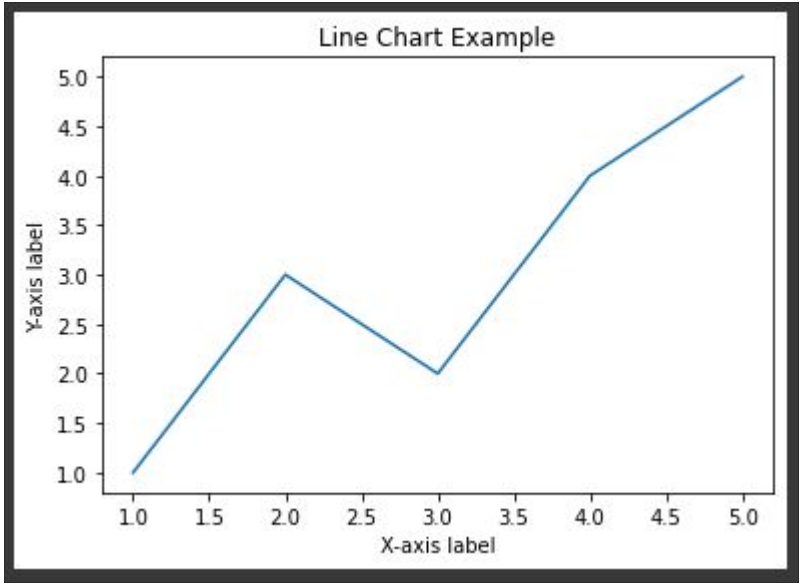
CODE Snippets



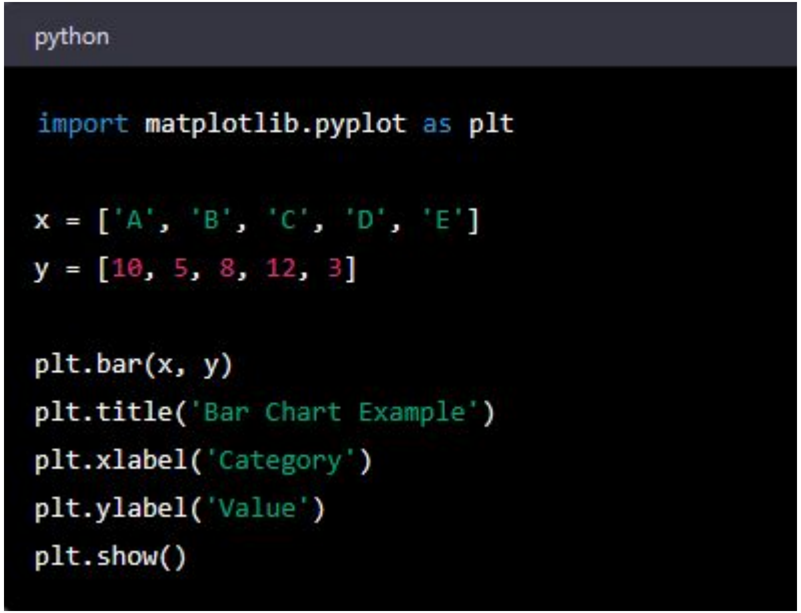
```python
import matplotlib.pyplot as plt

x = [1, 2, 3, 4, 5]
y = [1, 3, 2, 4, 5]

plt.plot(x, y)
plt.title('Line Chart Example')
plt.xlabel('X-axis label')
plt.ylabel('Y-axis label')
plt.show()
```



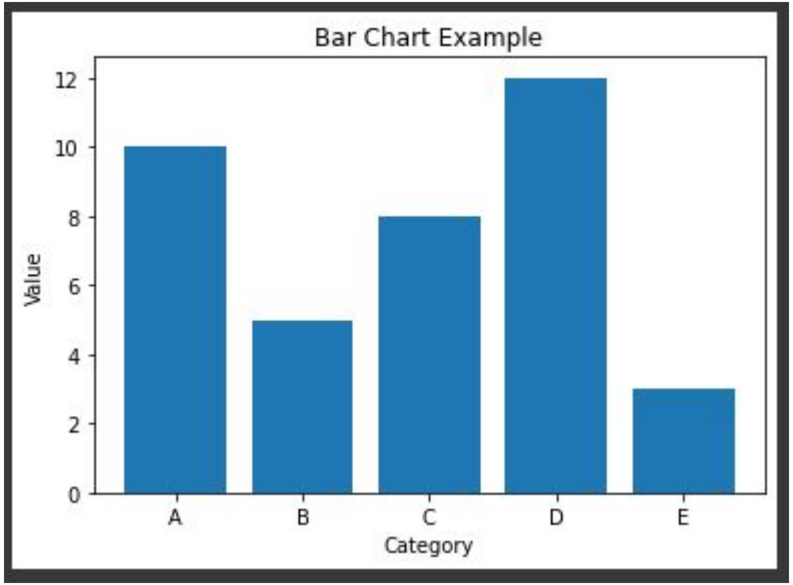
CODE Snippets



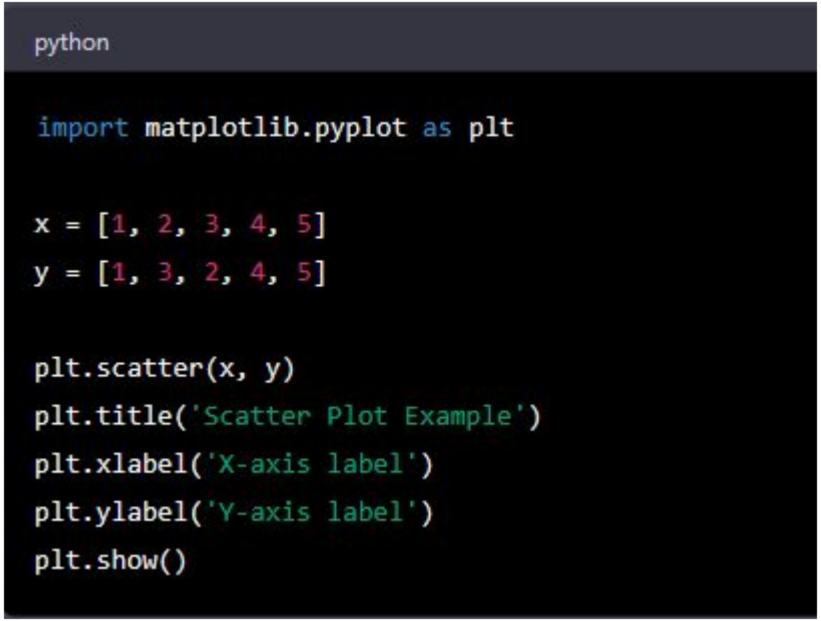
```python
import matplotlib.pyplot as plt

x = ['A', 'B', 'C', 'D', 'E']
y = [10, 5, 8, 12, 3]

plt.bar(x, y)
plt.title('Bar Chart Example')
plt.xlabel('Category')
plt.ylabel('Value')
plt.show()
```



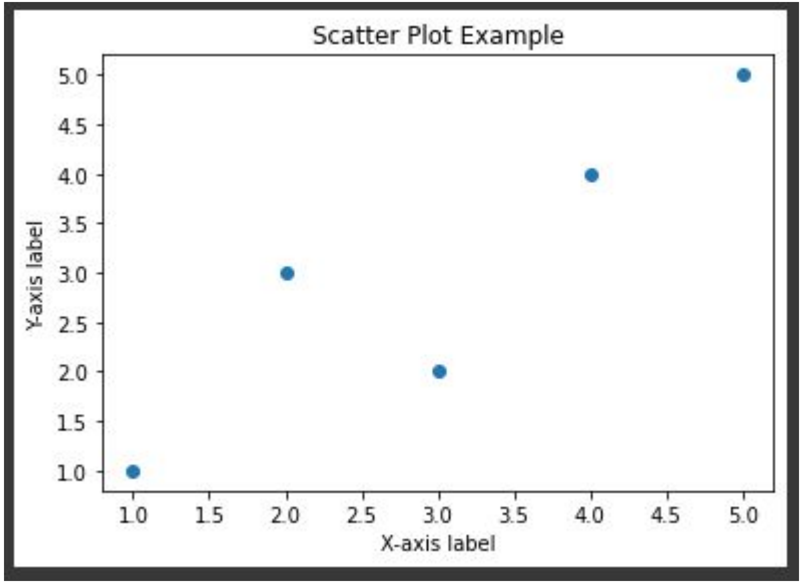
CODE Snippets



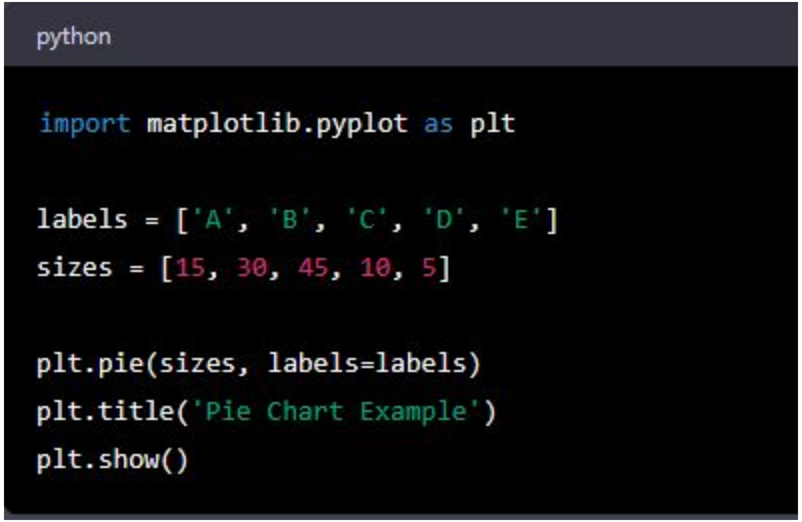
```python
import matplotlib.pyplot as plt

x = [1, 2, 3, 4, 5]
y = [1, 3, 2, 4, 5]

plt.scatter(x, y)
plt.title('Scatter Plot Example')
plt.xlabel('X-axis label')
plt.ylabel('Y-axis label')
plt.show()
```



CODE Snippets

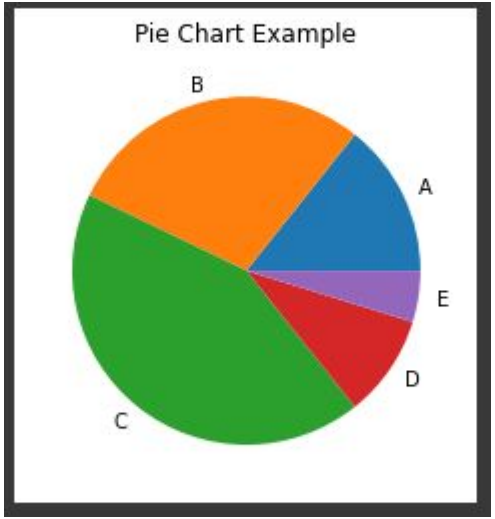


python

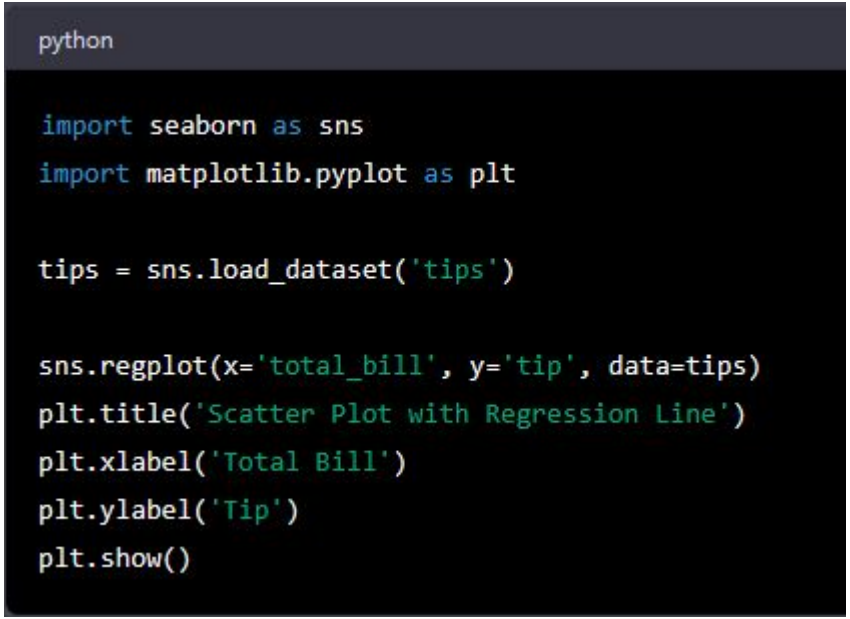
```python
import matplotlib.pyplot as plt

labels = ['A', 'B', 'C', 'D', 'E']
sizes = [15, 30, 45, 10, 5]

plt.pie(sizes, labels=labels)
plt.title('Pie Chart Example')
plt.show()
```



CODE Snippets

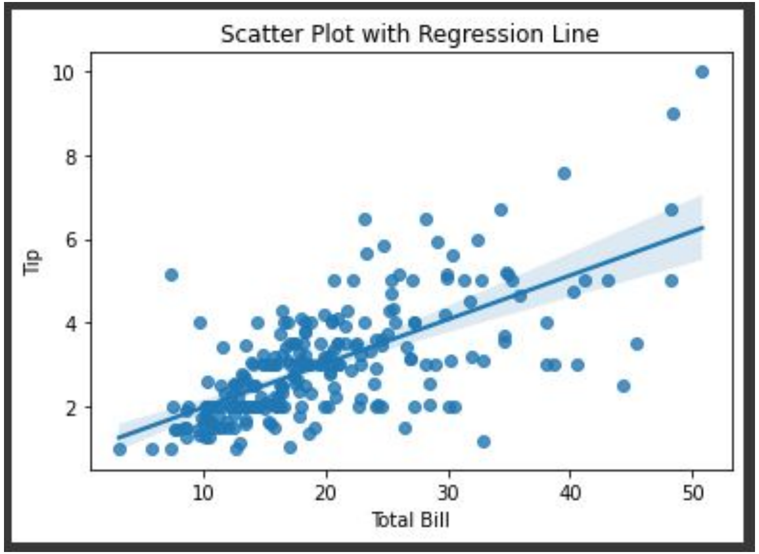


python

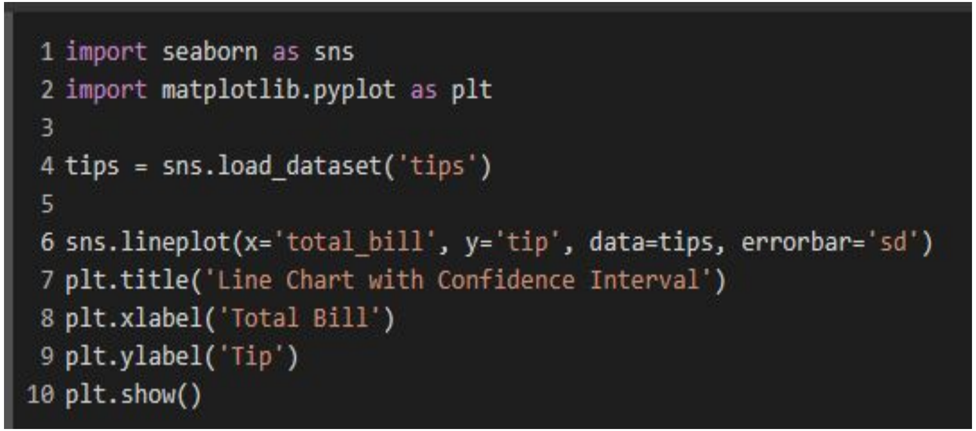
```python
import seaborn as sns
import matplotlib.pyplot as plt

tips = sns.load_dataset('tips')

sns.regplot(x='total_bill', y='tip', data=tips)
plt.title('Scatter Plot with Regression Line')
plt.xlabel('Total Bill')
plt.ylabel('Tip')
plt.show()
```



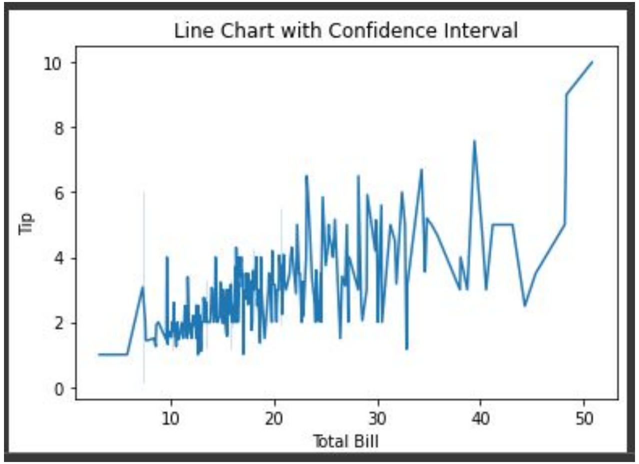
CODE Snippets



```python
import seaborn as sns
import matplotlib.pyplot as plt

tips = sns.load_dataset('tips')

sns.lineplot(x='total_bill', y='tip', data=tips, errorbar='sd')
plt.title('Line Chart with Confidence Interval')
plt.xlabel('Total Bill')
plt.ylabel('Tip')
plt.show()
```



Other Visualization Tools

- ***Tableau:*** One of the most widely used data visualization tools, it offers interactive visualization solutions
- ***Power BI:*** Microsoft's easy-to-use data visualization tool, is available for both on-premise installation and deployment on the cloud infrastructure
- ***Dundas BI:*** offers highly-customizable data visualizations with interactive scorecards, maps, gauges, and charts,
- ***JupyteR:*** one of the top-rated data visualization tools that enable users to create and share documents containing visualizations, equations and live code
- ***Google Charts:*** coded with SVG and HTML5, is famed for its capability to produce graphical and pictorial data visualizations.
- ***ZoHo:*** a comprehensive data visualization tool allow quick creation and sharing of extensive reports in minutes

Best Practices

- ***Keep it Simple:*** Data overload can quickly lead to confusion, so it’s important to only include the most important information. Avoid distracting elements.
- ***Annotation:*** Add explanatory or descriptive information to enhance clarity and draw attention to important insights and trends.
- ***Labelling:*** Labels should be clear and concise and they should accurately describe the data that is being represented.
- ***Colours:*** Use colors to highlight important trends, improve readability and provide context. Avoid overuse and inappropriate use of colors.
- ***Visual Hierarchy:*** Direct the viewer’s attention to the most important information through the use of size, color and position
- ***Data Points:*** Choose data points that accurately represent the underlying data. Be conservative and avoid clutter and confusion

Join us on Slack to ask questions and keep the discussion going!

Use the channel:

#build-a-ds-project