CYBR 520

Chapter 2: Visual Aids for EDA

Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data.

PLEASE REFER TO MODULE 2 ON GITHUB TO OBTAIN THE CODES OR REFER TO THE BOOK'S GITHUB REPO

TOPICS

- Line chart
- Bar chart
- Scatter plot
- Area plot and stacked plot
- Pie chart
- Table chart
- Polar chart
- Histogram
- Lollipop chart
- Choosing the best chart
- Other libraries to explore

Why Data Visualization

- A picture is worth a thousand words.
- As data scientists, two important goals in our work would be to extract knowledge from the data and to present the data to stakeholders.
- Presenting results to stakeholders is very complex in the sense that our audience may not have enough technical know-how to understand programming jargon and other technicalities.
  - Hence, visual aids are very useful tools.

Steps to create Visuals in Python

- Import libraries needed (i.e., pandas, seaborn, and matplotlib installed.
- Import / create data/ datasets
  - Local csv, text, etc...
  - Database (local/ online)
  - Online sources
- Plot graph
- Display/ save

Line Chart

- A line chart is a type of graph that displays data points connected by straight lines. It is commonly used to visualize trends, changes over time, and continuous data sets.
- Use line charts when you want to:
  - Show trends or patterns in data.
  - Display data that has a clear progression over time.
  - Compare multiple data series with a common x-axis (e.g., time, categories).
  - Emphasize the continuity and relationship between data points.

Bar Charts

- A bar chart is a graphical representation of data using rectangular bars.
- It is used to compare categories or show discrete data values.
- Use bar charts when you want to:
  - Compare quantities or values among different categories.
  - Show data that is not continuous and doesn't have a natural progression.
  - Emphasize differences between categories.
  - Visualize data in a straightforward and easy-to-understand manner.

Scatter Plots

- A scatter plot is a graphical representation of data points on a two-dimensional plane, where each point represents an observation with two numeric variables. It is used to visualize relationships, patterns, and trends between variables.
- Use scatter plots when you want to:
  - Explore the relationship between two continuous variables.
  - Identify correlations or trends in data.
  - Detect outliers or clusters.
  - Assess the distribution and spread of data points.

Bubble Chart

- A bubble chart is a variation of a scatter plot where data points are represented as bubbles, and the size or color of each bubble encodes a third dimension of data. It is used to visualize three numeric variables simultaneously.
- Use bubble charts when you want to:
  - Display relationships between three continuous variables.
  - Highlight the magnitude of a third variable through bubble size or color.
  - Provide a clear visualization of multivariate data.

Area and Stacked Plots

- Area plots and stacked plots are types of charts that display data with filled areas under curves. They are used to show the cumulative contribution of multiple data series or parts to a whole.
- Use area and stacked plots when you want to:
  - Visualize how multiple data series contribute to a whole over time.
  - Show the composition of a whole broken down into its parts.
  - Highlight changes in the distribution of data over time or categories.

Pie Charts

- A pie chart is a circular graph divided into slices, where each slice represents a portion or percentage of a whole. It is used to display the composition of a data set and highlight the relative sizes of different categories.

- Use pie charts when you want to:
  - Show the distribution of categories within a whole.
  - Emphasize the relative proportions of different parts.
  - Present data with a limited number of categories (typically less than 5-7) to ensure clarity.
  - Make it easy for viewers to grasp the relative sizes visually.

Table chart

- A table chart is a visual representation of data organized in rows and columns. It presents data in a structured and tabular format, making it easy to compare values, relationships, and details.

- Use table charts when you want to:
  - Display and compare individual data points or values.
  - Present data with many categories or dimensions.
  - Provide precise numerical information and maintain data integrity.
  - Offer a detailed view of data without the need for visual encoding.

Polar chart

- A polar chart, also known as a radar chart or spider chart, is a circular graph with multiple axes radiating from a central point. It is used to display multivariate data in a way that reveals patterns, strengths, and weaknesses across different variables.
- Use polar charts when you want to:
  - Compare multiple variables across different categories or data points.
  - Highlight the strengths and weaknesses of each category relative to the variables.
  - Visualize data with cyclical or periodic patterns.
  - Present data with a wide range of variables that need to be compared simultaneously.

Histogram

- A histogram is a graphical representation of the distribution of a dataset. It is used to visualize the frequency or count of data points within predefined intervals or bins.
- Use histograms when you want to:
  - Understand the distribution and central tendencies of a dataset.
  - Identify patterns, peaks, or outliers in data.
  - Explore the spread and variability of numerical data.
  - Visualize the frequency of data points within specific intervals.

Lollipop Charts

- A lollipop chart is a hybrid graph that combines elements of a bar chart and a line chart. It is used to display and compare data points, emphasizing their individual values, while also showing their position on a scale.
- Use lollipop charts when you want to:
  - Highlight individual data points or values.
  - Compare data points across different categories or groups.
  - Emphasize the ranking or order of data points.
  - Show both the values and their relative positions on a scale.

Choosing the best chart

The following table shows the different types of charts based on the purposes:

Purpose	Charts
Show correlation	Scatter plot Correlogram Pairwise plot Jittering with strip plot Counts plot Marginal histogram Scatter plot with a line of best fit Bubble plot with circling
Show deviation	Area chart Diverging bars Diverging texts Diverging dot plot Diverging lollipop plot with markers
Show distribution	Histogram for continuous variable Histogram for categorical variable Density plot Categorical plots Density curves with histogram Population pyramid Violin plot Joy plot Distributed dot plot Box plot
Show composition	Waffle chart Pie chart Treemap Bar chart
Show change	Time series plot Time series with peaks and troughs annotated Autocorrelation plot Cross-correlation plot Multiple time series Plotting with different scales using the secondary *y* axis Stacked area chart Seasonal plot Calendar heat map Area chart unstacked
Show groups	Dendrogram Cluster plot Andrews curve Parallel coordinates
Show ranking	Ordered bar chart Lollipop chart Dot plot Slope plot Dumbbell plot

Choosing the Best Chart - Consider Your Data

- **Understanding Your Data**
  - Examine your dataset to determine the type and nature of the data.
  - Ask yourself: Is the data numerical or categorical? Is it continuous or discrete?
- **Data Variables**
  - Identify the variables you want to visualize.
  - Decide whether you need to compare, show distribution, relationships, or trends.

Choosing the Best Chart - Data Types

- **Numerical Data**
  - For numerical data:
    - Use histograms for distribution.
    - Employ scatter plots for relationships.
    - Consider line charts for trends over time.
- **Categorical Data**
  - For categorical data:
    - Choose bar charts for comparisons.
    - Opt for pie charts when displaying parts of a whole.

Handling Data Quantity

- Consider the amount of data you have.
- If dealing with a few data points, go for detailed charts like scatter plots or lollipop charts.
- For large datasets, use aggregated or summarized charts like bar charts or histograms.

Ensuring Clarity

- Think about the message you want to convey.
- Use simple and familiar charts for straightforward messages.
- Employ more complex charts like polar or lollipop charts when emphasizing specific points or relationships.

Context and Audience

- Reflect on where and how your chart will be presented.
- Consider your audience's familiarity with chart types.
- Choose a chart that aligns with the context and is easily understandable to your audience.

Remember that choosing the best chart or graph depends on a careful analysis of your data, your message, and your audience's needs. There is no one-size-fits-all solution, so use these guidelines to make informed decisions about visualizing your data effectively.

Other libraries to explore

- Ploty (https://plot.ly/python/): This is a web-application-based toolkit for visualization. Its API for Jupyter Notebook and other applications makes it very powerful to represent 2D and 3D charts.
- Ggplot (http://ggplot.yhathq.com/): This is a Python implementation based on the Grammar of Graphics library from the R programming language.
- Altair (https://altair-viz.github.io/): This is built on the top of the powerful Vega-Lite visualization grammar and follows very declarative statistical visualization library techniques. In addition to that, it has a very descriptive and simple API.