Sardar Patel Institute of Technology, Mumbai

Department of Electronics and Telecommunication Engineering

T.E. Sem-V (2018-2019)

ETL54-Statistical Computational Laboratory

Lab-8: Data Visualization and Matrix Computation

Name: SHUBHAM PARULEKAR Roll No: 53

Part-I: Data Visualization

Objective: To create a range of graphs to summarize your data and results

Outcomes:

1. **To create boxplot, scatter plots, including correlation plots**
2. **To create line graphs, pie charts and bar charts**
3. **To save graphs as files on disk (png, jpg etc)**
4. **To choose the right type of chart for your specific objectives and how to implement it in R using ggplot2.**

System Requirements: Ubuntu OS with R and RStudio installed and ggplot2, Python, Pandas, Matplotlib, seaborn, Plotly etc.

Introduction to Visualization:

Data visualization is an art of how to turn numbers into useful knowledge. [1]

Graphs are a powerful way to present your data and results in a concise manner. Whatever kind of data you have, there is a way to illustrate it graphically. A graph is more readily understandable than words and numbers, and producing good graphs is a vital skill. Some graphs are also useful in examining data so that you can gain some idea of patterns that may exist; this can direct you toward the correct statistical analysis.

Selecting the Right Chart Type:

There are four basic presentation types:

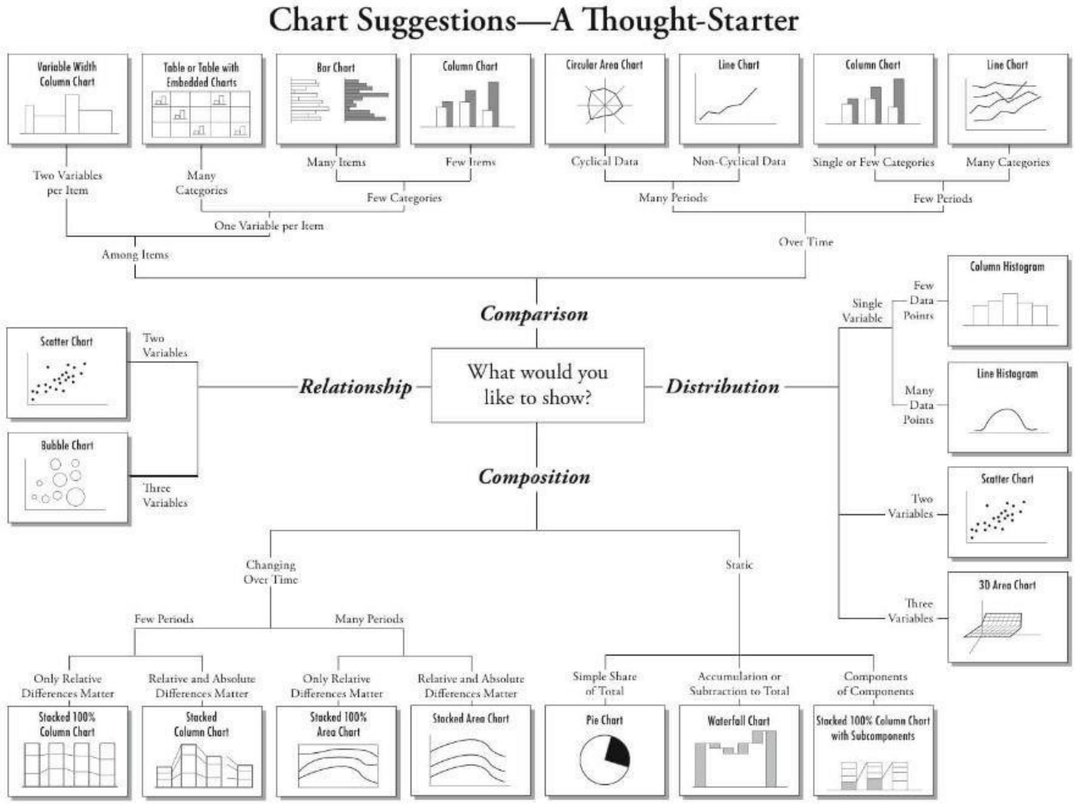
1. **Comparison**
2. **Composition**
3. **Distribution**

4. **Relationship**

To determine which amongst these is best suited for your data, We suggest you should answer a few questions like,[1]

- **How many variables do you want to show in a single chart?**
- **How many data points will you display for each variable?**
- **Will you display values over a period of time, or among items or groups?**

Do refer the following picture and understand how select a right chart type.



[Courtesy: Dr. Andrew Abela]

In your day-to-day activities, you’ll come across the below listed 7 charts most of the time.

1. **Scatter Plot**
2. **Histogram**
3. **Bar & Stack Bar Chart**
4. **Box Plot**
5. **Area Chart**
6. **Heat Map**
7. **Correlogram**

[1] Introduction to Data Visualization in Python- [1 hr]

https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97fbed

Perform this lab using the KDDCUP99 Intrusion Dataset

Refer the Lab6: Anomaly Detection using Machine Learning

Download the datasets (CSVs): i. Train ii. Test

Additional:

[2] Pre-reading material and understand (for 60 minutes)

Refer the following website;

https://www.r-bloggers.com/7-visualizations-you-should-learn-in-r/

[3] Read and perform laboratory on data visualization with R Graphics [2 hr]

Refer the following website:

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

Laboratory Session:

Procedure:

1. **Open RStudio**
2. **Go to RConsole (>)**

Practice the following 8 categories of plots (graphs):

1. Correlation
2. Deviation
3. Ranking
4. Distribution
5. Composition
6. Change
7. Groups
8. Spatial

Describe following with respect to data visualization:

1. When to use scatter plot, histogram, bar and stack charts, box plot, Area chart?

A) Scatter plot:

Scatter plot or scattergram is a type of diagram that uses Cartesian coordinates to illustrate values of two common variables for a data set. In this case, the data is represented as a collection of points. It is used in the case of more than one variable which are neither similar nor ordered.

Histogram:

The histogram is widely applied as a representation of numerical data distribution. Each bar in the histogram represents the data distributed in a single category, a continuous range of data or frequencies for a specific data point. It is generally used in cases having a single variable which isn't ordered.

Stacked bars:

In case of stacked bars, parts of the data are adjusted or stacked (horizontal bars, vertical bars or columns) representing the whole amount of data broken down into sub-amounts. Equivalent sections in each bar are colored similarly.

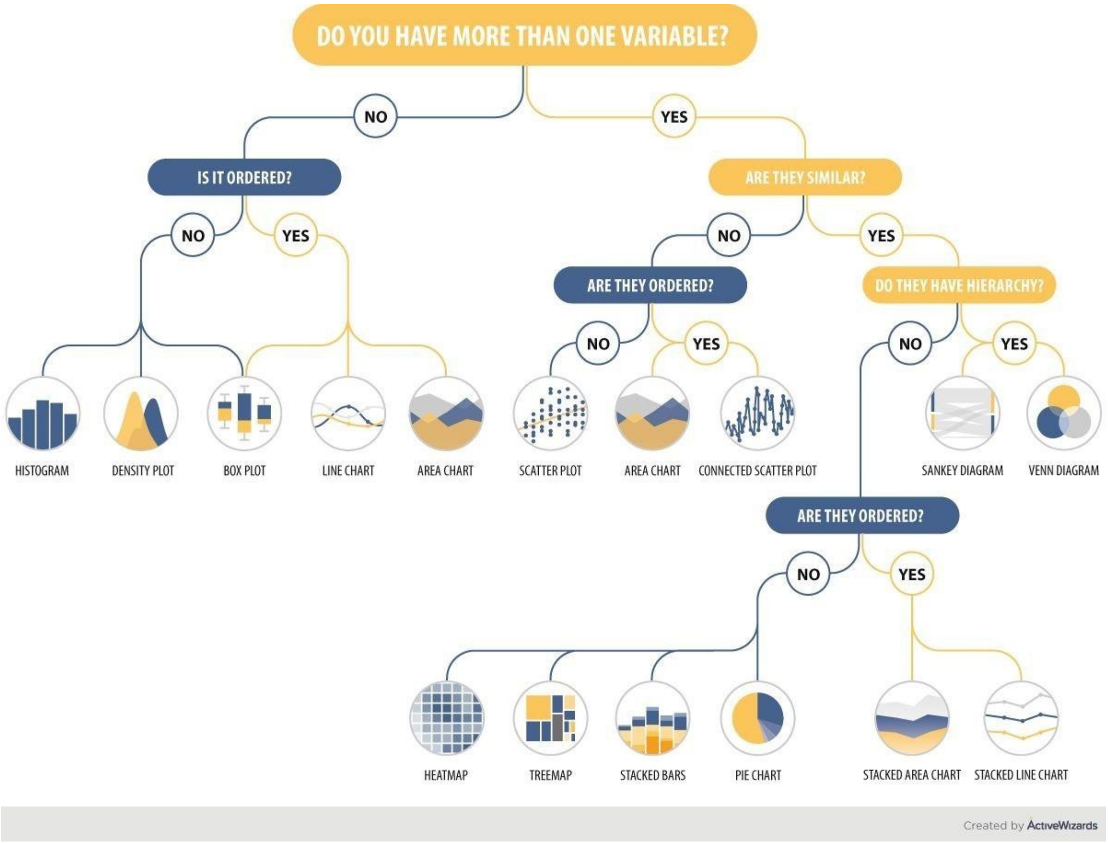
It is used in the case of more than one variable which are similar but do not have a hierarchy and aren't ordered.

Box plot:

Box plot is usually used to depict the groups of numerical data with the help of their quartiles. Box plot often has the whiskers extended vertically to illustrate variability outside the quartiles. It is used in cases having one variable and it may or may not be ordered.

Area chart:

This type of chart is based on the line chart. Thus, their functions are quite similar. An area chart is used to illustrate quantitative data graphically by plotting the data points and connecting them into line segments. It is used in cases having one ordered variable.



2. What is Heatmap?

A) Heatmap is such a graphical representation of data where individual values within the matrix are depicted as colors. Larger values are represented by dark pixels, where smaller values are pictured with lighter colors.

3.What is correlogram?

A) A correlogram (also called Auto Correlation Function ACF Plot or Autocorrelation plot) is a visual way to show serial correlation in data that changes over time (i.e. time series data). Serial correlation (also called autocorrelation) is where an error at one point in time travels to a subsequent point in time. For example, you might overestimate the value of your stock market investments for the first quarter, leading to an overestimate of values for following quarters. Correlograms can give you a good idea of whether or not pairs of data show autocorrelation.

3. Run this script in R script editor and describe:

require(ggmap)

```
require(ggplot2)
from <- 'Mumbai'
to <- 'Thane'
route_df <- route(from, to, structure = 'route', mode = 'driving')
qmap('Mumbai', zoom = 10) +
  geom_path(
    aes(x = lon, y = lat), colour = 'red', size = 1.5,
    data = route_df, lineend = 'round')
mapdist(from, to)
```

Code:

Posted on google classroom.

Conclusion:

1. The type of plot to be chosen for a particular set of data depends on the number of variables to be displayed in the chart, the number of data points to be displayed for each variable and whether the values are to be displayed over a period of time, or among items or groups.
2. For given KDD dataset, a correlation heat map is plotted to get the relationship between all variables. A perfect negative correlation is represented by the value -1, a 0 indicates no correlation, and a +1 indicates a perfect positive correlation.

References:

[1] 7 Visualizations You Should Learn in R

https://www.r-bloggers.com/7-visualizations-you-should-learn-in-r/

[2] Top 50 ggplot2 Visualizations

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

Part-II: Matrix Computation

Objective: To carry out matrix computation

Outcomes:

5. To create vectors and matrices
6. To extract elements from a matrix
7. To use and describe the general information commands with respect to matrix.
8. To carry out matrix operations
9. To find the eigenvalues and eigenvectors

System Requirements: Ubuntu OS with R and RStudio installed, Python etc

Introduction to Linear Algebra:

Procedure:

1. Refer the following for Matrix Computation in Python

Matrix Arithmetics under NumPy and Python

https://www.python-course.eu/matrix_arithmetic.php

2. Gentle Introduction to Eigenvalues and Eigenvectors for Machine Learning

https://machinelearningmastery.com/introduction-to-eigendecomposition-eigenvalues-and-eigenvectors/

[3] Refer the [1] and [2] pdf files provided and complete the lab.

Describe the following with respect to matrix computation

1. List the general information commands used in R for matrix

A) The basic syntax for creating a matrix in R is −

```
matrix(data, nrow, ncol, byrow, dimnames)
```

Following is the description of the parameters used −

- data is the input vector which becomes the data elements of the matrix.
- nrow is the number of rows to be created.
- ncol is the number of columns to be created.
- byrow is a logical clue. If TRUE then the input vector elements are arranged by row.

- dimname is the names assigned to the rows and columns.

2. Describe the importance of matrix computation.

A) **Matrices** are **used** to describe linear equations, keep track of the coefficients of linear transformations and to record data that depend on two parameters. **Matrices** can be added, multiplied, and decomposed in various ways, making them a key concept in linear algebra and **matrix** theory.

CODE:Posted on google classroom.

Conclusion:

Various operations like matrix and vector creation, extraction of elements from matrix, dot product and eigenvalue and eigenvector calculation were performed.

References:

[1]Examples of Using R with Linear Algebra by S. K. Hyde [pdf]

[2]Linear algebra in R by Søren Højsgaard [pdf]

[3]https://www.statmethods.net/advstats/matrix.html

[4] Hands-On Matrix Algebra Using R by Hrishikesh D Vinod, World Scientific