Statistical Learning - Basic Statistics

Basic Statistics - Topics

- Outline
- Why Statistics and Big data
- Statistics – methods
- Classical definition and Types of stats
- Some vital terms in stats
- Sources and types of data and datasets
- Data objects, attributes and its types
- Descriptive Statistics outline
- Data and Histogram
- Central tendency and 3 Ms

Basic Statistics - Topics(Contd.)

- Measure of dispersion, Range, IQR
- Descriptive statistics – Standard Deviation
- Coefficient of variation
- The Empirical Rule and Chebyshev Rule
- The Five Number Summary and the Boxplot
- Quantile Plot and Scatter Plot
- Data Visualization
- Visually Evaluating Correlation
- Summary

Case Study

Why Statistics and Big data

Why statistics is so important

The significant events triggered the current meteoric growth in the use of analytical decision making and Statistics is central to all of them.

Event 1

- Technological developments, Revolution of Internet and Social Networks, data generated from mobile phones produce large amount of data from which insights will be shifted.
- The discovery of pattern and trends from these data for organizations will pay the way for improving profitability, understanding customer expectations so that they can gain competitive advantage in the market place.

Why Statistics and Big data (Contd.)

greatlearning

Event 2

- Advances in enormous computing power to effectively process and analyse massive amounts of data.
- Sophisticated and faster algorithms for solving problems.
- Data visualization for Business intelligence and artificial intelligence.

Event 3

- Large data storage capability.
- Parallel and cloud computing have enabled business to solve large scale problems.

Why Statistics and Big data (Contd.)

Big Data

- A set of data that cannot be managed, processed, or analysed with traditional software/algorithms within a reasonable amount of time.

- Big data revolves around

  Volume, Velocity, Variety, Value, Veracity

Statistics - Methods

Classification

- Classification techniques helps in segmenting the customers into appropriate groups based on key characteristics.
- For example, using appropriate statistical model, an organization could easily segment the customers into Long term customers, medium term customers, and Brand switchers.
- Classification helps professionals understand the customer behaviour and position their products and brands using appropriate strategies.

Statistics - Methods

Pattern Recognition

- "A picture is worth thousand words" and it reveals hidden pattern in the data that could be leveraged be retail professionals. Pattern recognition techniques include *Histogram, Box Plot, Scatter plot and other visual analytics.*

- For example, histogram drawn for income of a particular class of customers may reveal a symmetrical bell curve pattern or may be left or right skewed.

- Relationship between age and expenditure could be captured using a scatter plot.

Statistics - Methods

Association

- *Association* analysis helps in determining which of the items go together. Association rules include a set of analytics that focuses on discovering relationships.

- In this context, market basket analysis refers to an association rule that generates the probability for an outcome.

- Association rules can be adapted by organizations to store layout, items bundling, discount and sales promotion decisions, and cross selling among others.

Statistics - Methods

Predictive Modeling

- Both customer segmentation as well as identifying and targeting most profitable customers can be facilitated by predictive models.
- Regression can be used for predicting the amount of expenditure on a particular product based on input variables income, age and gender.
- Organizations can leverage on other advanced models that comprise Logistic Regression, and neural networks for predicting a target variable as well as classifying and predicting into which group the consumer belongs to.

Classical definition and Types of stats

- “By statistics, we mean methods specially adopted to the elucidation of quantitative data affected to a marked extent by multiplicity of causes”.

- It is interesting to see what *Thomas Davenport* means by Business Analytics and note the similarities and dissimilarities between the two.

- Business Analytics (BA) can be defined as the broad use of data and quantitative analysis for decision making within organizations.

Classical definition and Types of stats

Types of statistics

- **Descriptive statistics** is concerned with Data summarization, Graphs/Charts, and tables
- **Inferential statistics** is a method used to talk about a Population parameter from a sample

Some vital terms in stats

Population, Parameter, Sample, Statistic

- A **population** is the universe of possible data for a specified object.
- A **parameter** is a numerical value associated with a population.
- A **sample** is a selection of observations from a population
- A **Statistic** is a numerical value associated with an observed sample.

Sources and Types of data

greatlearning

Data Sources

Primary Data are collected by organization itself for a particular purpose. The benefits of primary data are that they fits the needs exactly and reliable.

Secondary Data are collected by other organizations or for other purposes. Any data, which are not collected by the organization for the specified purpose, are secondary data.

Sources and Types of data

greatlearning

Types of Data

- **Qualitative** data are non numeric in nature and cannot be measured.
- **Quantitative** data are numerical in nature and can be measured and can be classified into two: discrete and continuous.
- **Discrete** type can take only certain values, and there are discontinuities between values.
- **Continuous** type can take any value within a specific interval.

Sources and Types of data

Types of datasets

1. Record
2. Graph and network
3. Ordered
4. Spatial, image and Multimedia

Data objects, attributes and its types

greatlearning

Data objects

- Data sets are made up of data objects.
- A data object represents an entity.
- Data objects are described by attributes.
- Examples:
  - sale database : customers, sales
  - medical database: patients, treatments

Data objects, attributes and its types

greatlearning

Attributes

- Attribute: a data field, representing a characteristic or feature of a data object.
  Example: customer_ID, name, address

Types:

- Nominal
- Binary
- Ordinal
- numeric

Data objects, attributes and its types

greatlearning

Attribute types

1. Nominal: categories, states or “names of things”
   - Hair_color = {black, brown, grey}
   - marital status, occupation, ID
2. Binary
   - Symmetric binary
   - Asymmetric binary
3. Ordinal

Descriptive Statistics outline

greatlearning

Outline

1. Raw data
2. Frequency Distribution - histograms
3. Cumulative frequency distribution
4. Measures of central tendency
5. Mean, median, mode
6. Measures of dispersion
7. Range, IQR, standard deviation, coefficient of variation
8. normal distribution
9. Five number summary, boxplots, QQ plots
10. . Visualization: scatter plot matrix
11. Correlation analysis

Data and Histogram

greatlearning

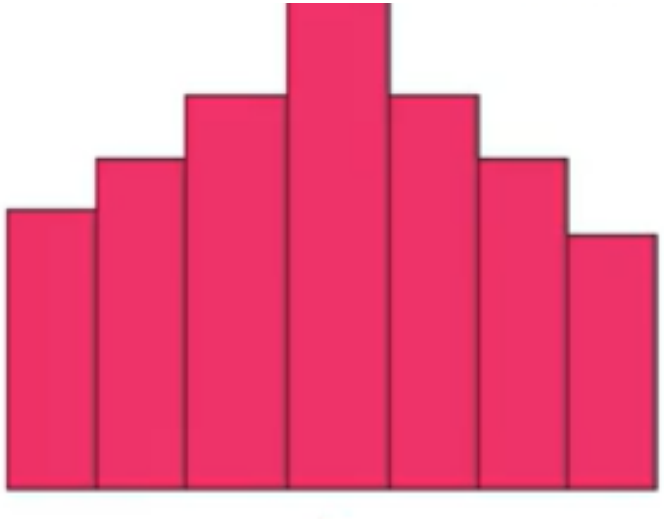
Raw data

Raw data represent numbers and facts in the original format in which the data have been collected. We need to convert the raw data into information for decision making.

Frequency distribution

Frequency distribution focuses on classifying raw data into information. It is widely used data reduction technique in descriptive statistics.

Histogram



1. Histogram is a snapshot of the frequency distribution.
2. histogram depicts the pattern of the distribution emerging from the characteristic being measured.

Central tendency and 3 Ms

Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. Such a value is called a measure of “Central tendency”

Mean

The statistical mean refers to the mean or average that is used to derive the central tendency of the data.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \lambda + X_n}{n}$$

Median

The middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for original data, in which values are ranked relative to each other but are not measured absolutely.

$$Median = \left(\frac{n+1}{n}\right)^{th} term$$

greatlearning

Mode

The most frequent value in the data set. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.

$$Mode = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

Where,

l = Lower Boundary of modal class

$h = size\ of\ model\ class$

$f_m = Frequency\ corresponding\ to\ modal\ class$

$f_1 = Frequency\ preceding\ to\ modal\ class$

$f_2 = Frequency\ proceeding\ to\ modal\ class$

Measure of dispersion, Range, IQR

greatlearning

Measures of dispersion

Measure of dispersion indicate how large the spread of distribution in around the central tendency.

Range

Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in dataset.

range = X(maximum) - X (minimum)

IQR

- Q1 divides the values between minimum and Q2 into 2 equal half. In other words Q1 is that value which has 25% values below it and rest above.

IQR = Q3 - Q1

Descriptive statistics – Standard Deviation

- Interpreting variance (a squared term) is not intuitive. Instead we under root it to get Standard deviation which has the same units as variable.

- Standard deviation, is a measure of average spread i.e., on an average what is the difference between any data point and the central value of the variable.

$$\sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Coefficient of variation

Coefficient variation is defined as ratio of standard deviation to mean.

In symbolic form,

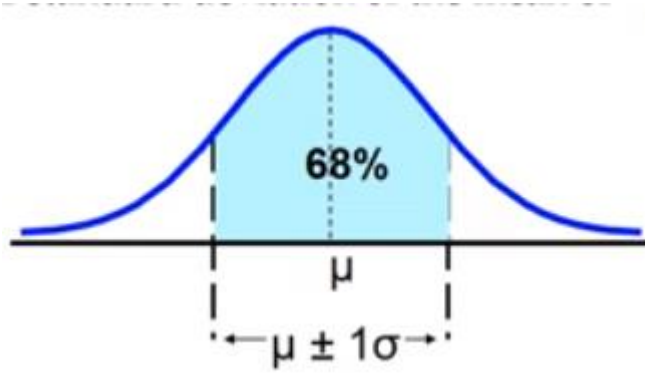
$$CV = \frac{S}{\bar{X}} \text{ for the sample data and } = \frac{\sigma}{\mu} \text{ for the population}$$

The Empirical Rule and Chebyshev Rule

greatlearning

The Empirical rule

- The empirical rule approximates the variation of data in a bell-shaped distribution.
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean.



The Empirical Rule and Chebyshev Rule

greatlearning

Chebyshev Rule

- Regardless of how the data are distributed, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for $k > 1$)

- For Example, when k=2, at least 75% of the values of any data set will be within $\mu \pm 2\sigma$

The Five Number Summary

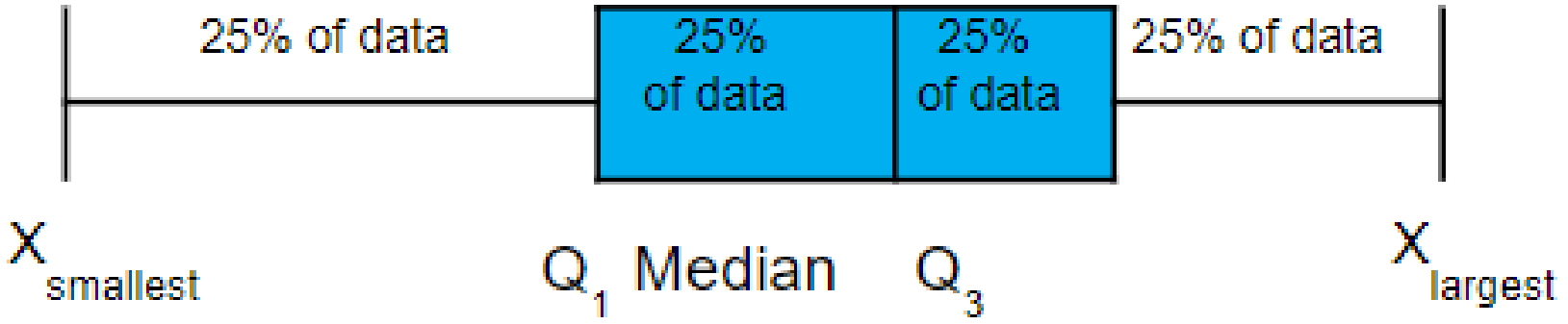
The five numbers that help describe the center, spread and shape of data are:

- $x_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $x_{largest}$

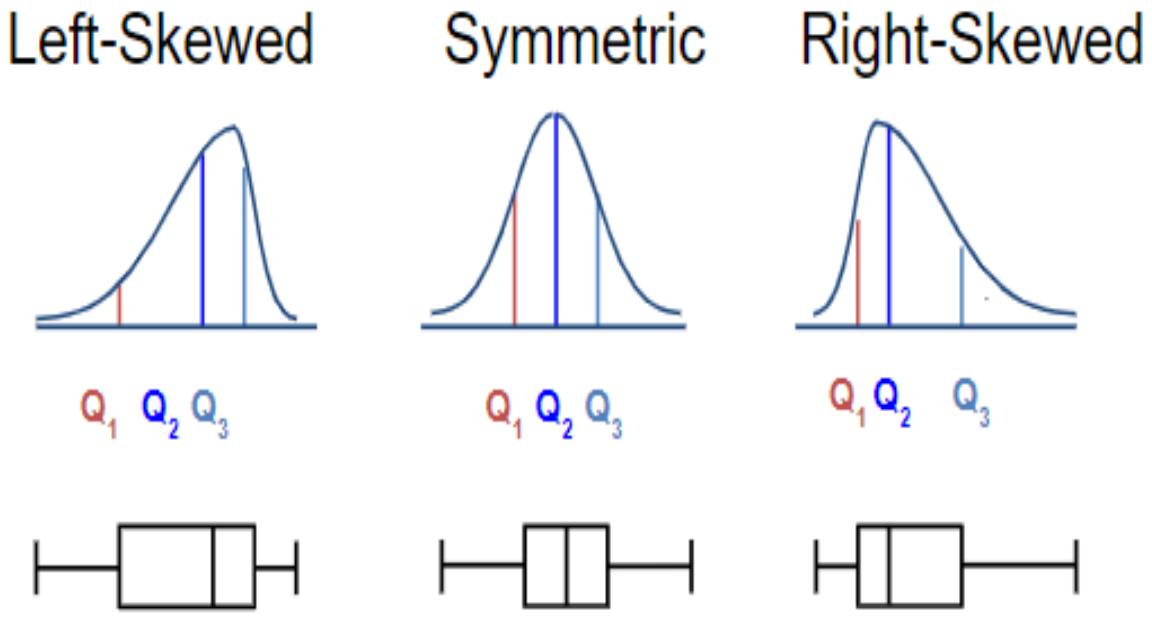
Five Number Summary and The Boxplot

- The Boxplot: A Graphical display of the data based on the five-number summary:

Example:



Distribution Shape and the Boxplot



Graphic Displays of Basic Statistical Descriptions

greatlearning

Boxplot: graphic display of five-number summary

Histogram: x-axis are values, y-axis repres. frequencies

Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$

Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another

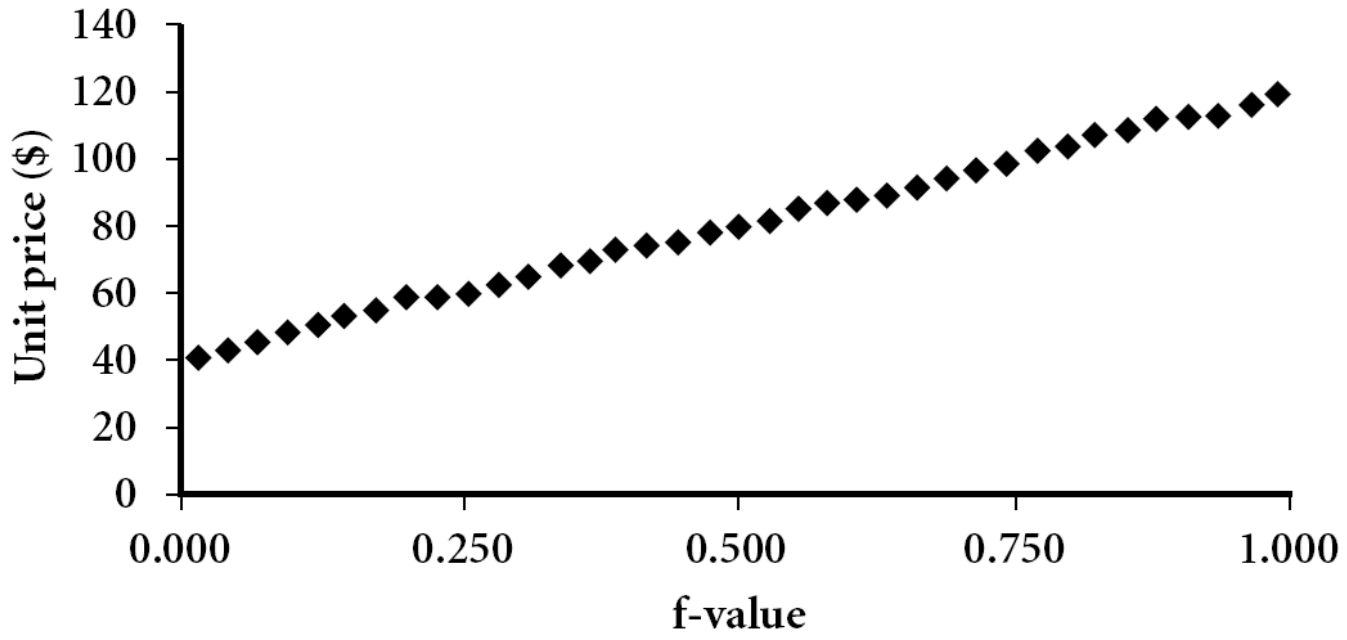
Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

Quantile Plot

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

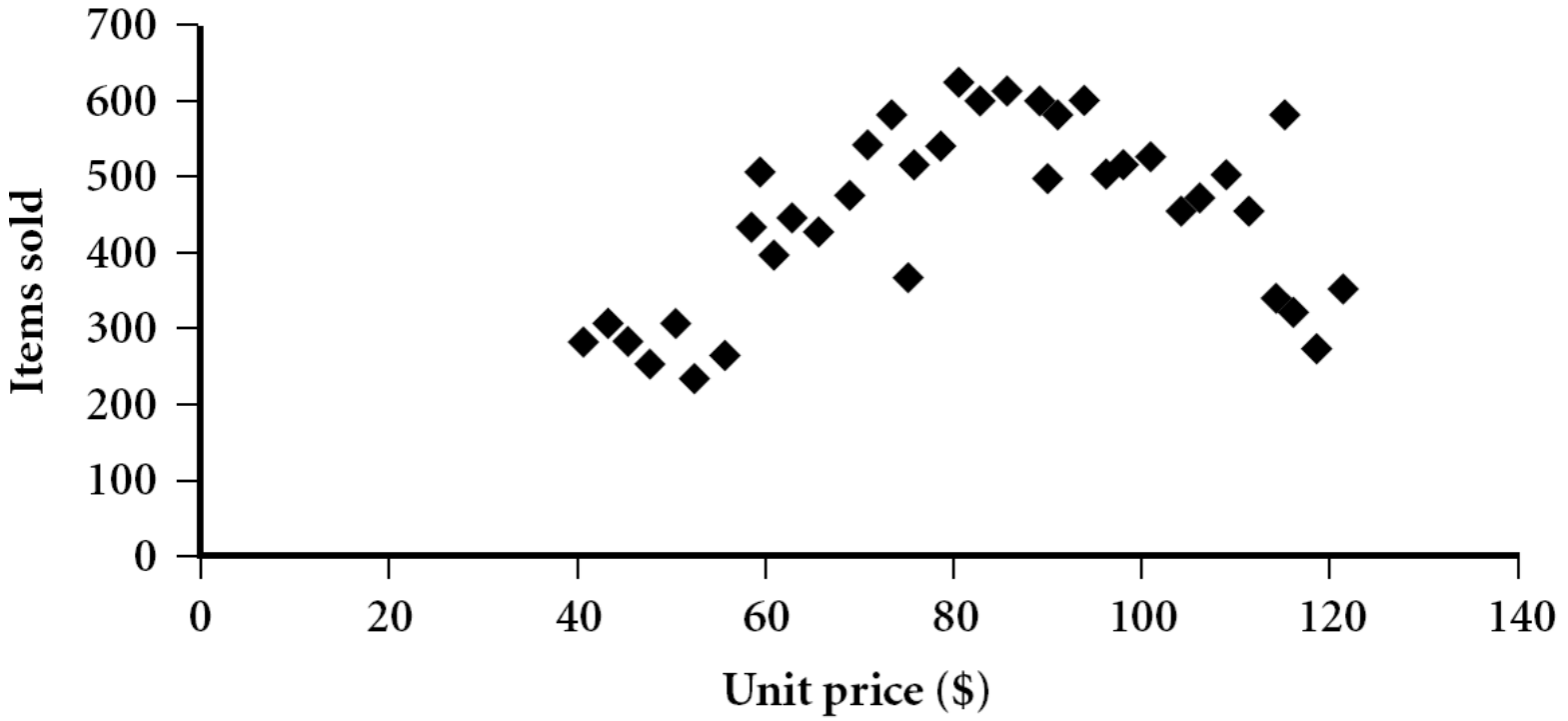
Plots **quantile** information

For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



Scatter Plot

Provides a first look at bivariate data to see clusters of points, outliers, etc
Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Data Visualization

greatlearning

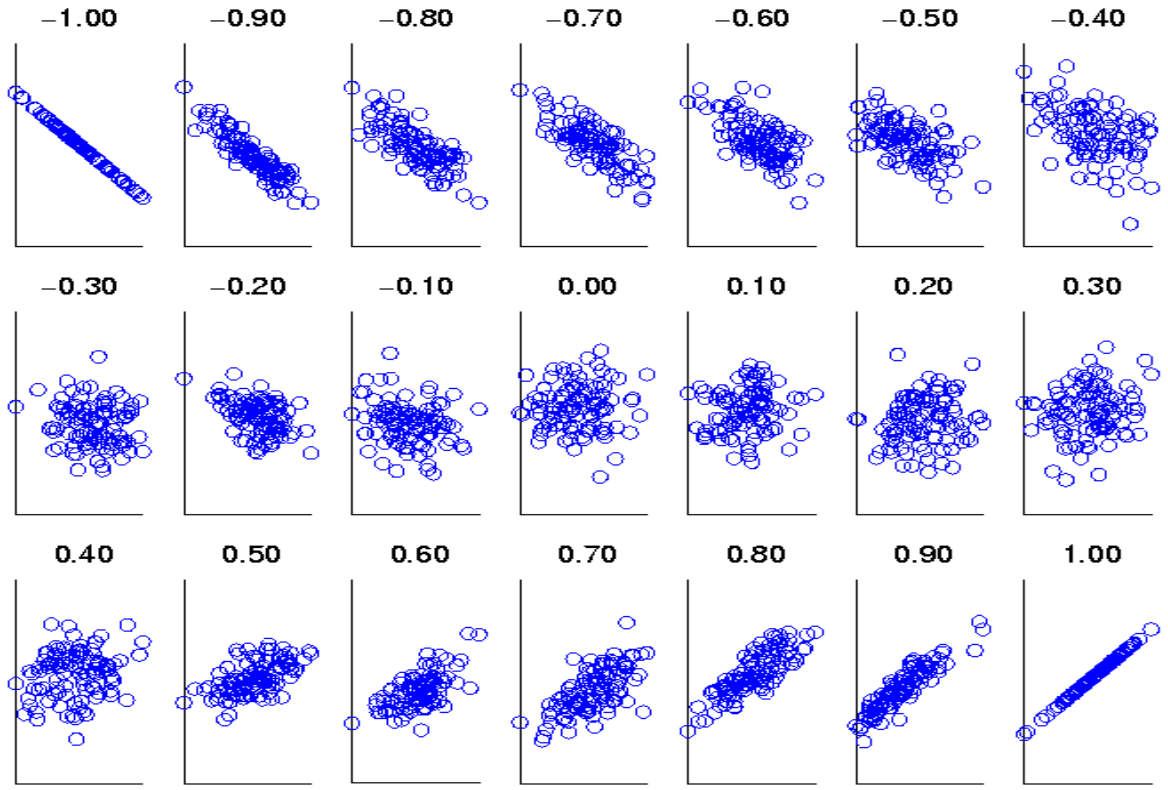
Why data visualization?

- Gain insight into an information space by mapping data onto graphical primitives
- Provide qualitative overview of large data sets
- Search for patterns, trends, structure, irregularities, relationships among data
- Help find interesting regions and suitable parameters for further quantitative analysis
- Provide a visual proof of computer representations derived

Categorization of visualization methods:

- Pixel-oriented visualization techniques
- Geometric projection visualization techniques
- Icon-based visualization techniques
- Hierarchical visualization techniques
- Visualizing complex data and relations

Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1

Summary

- Histograms
- Measures of central tendency: mean, mode, median
- Measures of dispersion: range, IQR, variance, std deviation, coefficient of variation.
- Normal distribution, Chebyshev Rule.
- Five number summary, boxplots, QQ plots, Quantile plot, scatter plot.
- Visualization: scatter plot matrix, parallel coordinates.
- Correlation analysis.

Case Study

greatlearning

Churn in Telecom's dataset-

The dataset is about telecom industry which tells about the number of customers who churned the service. It consists of 3333 observations having 21 variables. We have to predict which customer is going to churn the service.

Dataset -

The dataset contains State, Account Length, area code, phone number, international plan, voice mail plan and important variables like call charges, international call charges, customer service calls, etc.

For Reference: https://www.kaggle.com/blastchar/telco-customer-churn

Steps Followed

Objective - Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs.

Steps-

- Import pandas, numpy, seaborn, matplotlib.pyplot packages
- Get the data
- Calculate histogram for minutes per day spent by customers.
- How do we categorize the churner and the non-churner for the time spent on day calls?
- Find the number of customers who did opt for voicemail plan.
- Create a boxplot for a categorical variable and continuous variable.
- How to pivot information using python for categorical values?
- Understand the correlation between all variables.
- Plot a distplot to check total night calls.
- Calculate area wise churner or non-churner

greatlearning

HappyLearning

Questions?