Exploratory Data Analysis and Feature Engineering

Course Outcomes

At the end of the course students will be able to –

1. Explain exploratory data analysis and visualization techniques.
2. Explain the theoretical foundation of Hypothesis Testing and Analysis of Variance.
3. Explain different methods of feature selection.
4. Explain how to Reduce feature space in a dataset.

Exploratory Data Analysis and Feature Engineering

1. Fundamentals of Exploratory Data Analysis
2. Hypothesis Testing and Analysis of Variance
3. Exploratory Data Analysis
4. Feature Construction and Feature Selection
5. Feature Transformations
6. Feature Learning

Text and Reference Books

Text Books

1. Suresh Kumar Mukhiya, Usman Ahmed, “Hands-On Exploratory Data Analysis with Python”, Packt Publishing, ISBN 978-1-78953-725-3
2. Sinan Ozdemir, Divya Susarla, “Feature Engineering Made Easy”, Packt Publishing, ISBN 978-1-78728-760-0
3. Howard J .Seltman, “Experimental Design and Analysis”, http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf
4. Max Kuhn , Kjell Johnson, “Feature Engineering and Selection: A Practical Approach for Predictive Models” 1st Edition, Chapman & Hall/CRC Data Science Series, ISBN 13-978-1-138-07922-9

Reference Books

1. John W. Tukey, “Exploratory Data Analysis1st Edition”, Pearson Education, ISBN 0134995457, 9780134995458

Fundamentals of Exploratory Data Analysis

What is Data?

- Data encompasses a collection of discrete objects, numbers, words, events, facts, measurements, observations, or even descriptions of things.
- Data is collected in the form of numbers, text, pictures, videos, objects, audio, and other entities.
- Processing data provides a great deal of information.
- Data is collected and stored by every event or process occurring.
- Processing data elicits useful information and processing such information generates useful knowledge
- An important question is:

  How can we generate meaningful and useful information from such data?

- Answer is **Exploratory Data Analysis (EDA).**

EDA

- EDA is a process of
  - ✓ Examining the available dataset to discover patterns,
  - ✓ Spot anomalies,
  - ✓ Test hypotheses, and
  - ✓ Check assumptions using statistical measures.
- Primary aim of EDA is to examine what data can tell us before actually going through formal modeling or hypothesis formulation.
- **John Tuckey** promoted EDA to statisticians
  - ✓ To examine and discover the data and
  - ✓ Create newer hypotheses that could be used for the development of a newer approach in data collection and experimentations.

Fundamentals of Exploratory Data Analysis

Topics in Unit-I

- Understanding data science
- The significance of EDA
- Making sense of data
- Comparing EDA with classical and Bayesian analysis
- Software tools available for EDA
- Getting started with EDA

Visual Aids for EDA

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

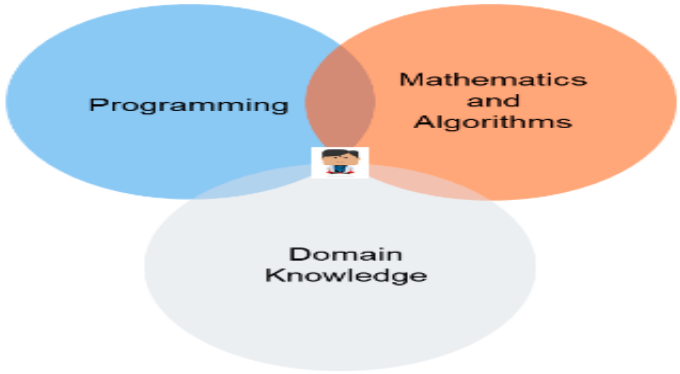
Understanding Data Science

- **Data science** is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

- Data science is related to data mining, deep learning and big data.

Understanding Data Science Cont..

- **Data science** is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.
- **Data science** practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce **artificial intelligence (AI) systems** to perform tasks that ordinarily require human intelligence.
- Systems generate insights which analysts and business users can translate into tangible business value.



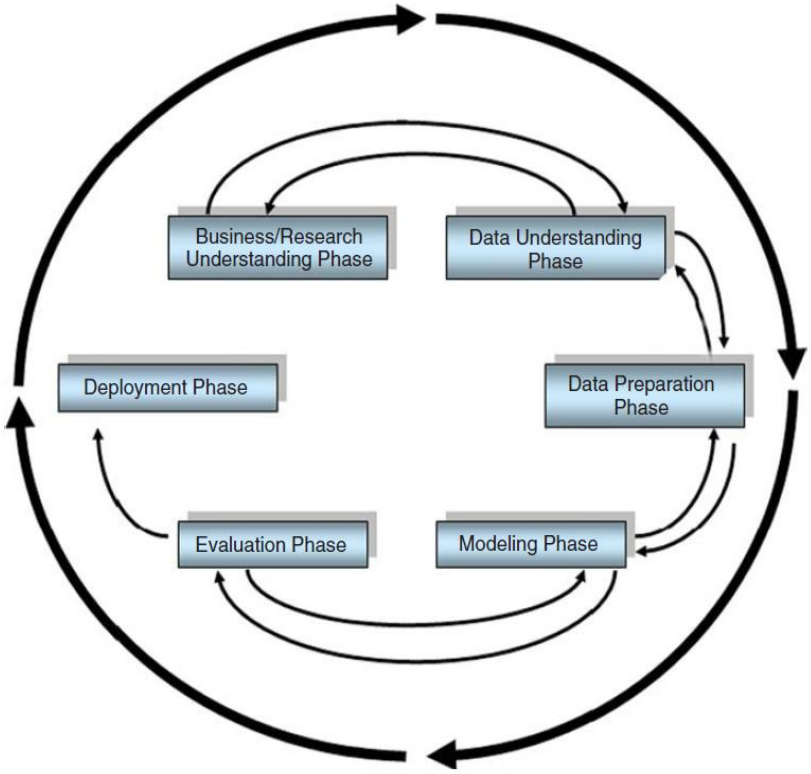
Understanding Data Science Cont..

- Data science is at the peak of its hype.
- Skills for data scientists are changing.
- Data scientists are required to build a performant model, explain the results obtained and use result for business intelligence.
- Data science involves cross-disciplinary knowledge from computer science, data, statistics, and mathematics.
- Data science and machine learning provides the basis for business growth, cost and risk reduction and even new business model creation.

Understanding Data Science Cont..

Phases of Data Analysis:

- ❑Data requirements,
- ❑Data collection,
- ❑Data processing,
- ❑Data cleaning,
- ❑Exploratory Data Analysis (EDA),
- ❑Modeling and Algorithms,
- ❑Data product and
- ❑Communication.



CRISP-DM - Iterative, Adaptive

- Similar to **CRoss-Industry Standard Process for data mining (CRISP)** framework in data mining.

Understanding Data Science Cont..

Phases of Data Analysis: Data Requirements

- Various sources of data for an organization.
- Comprehend what type of data is required for the organization to be collected, curated, and stored.
- For e.g., application tracking sleeping pattern of patients suffering from dementia.
- Requires several types of sensors' data storage, such as sleep data, heart rate from the patient, electro-dermal activities, and user activities pattern.
- Required to correctly diagnose the mental state of the person (mandatory requirements).
- Categorize the data, numerical or categorical, and the format of storage and dissemination.

Understanding Data Science Cont..

Phases of Data Analysis: Data Collection

- Data collected from several sources must be stored in the correct format.
- Transferred to the right information technology personnel.
- Data can be collected from several objects on several events using different types of sensors and storage tools.

Understanding Data Science Cont..

Phases of Data Analysis: Data Processing

- Preprocessing involves the process of pre-curating the dataset before actual analysis.
- **Common tasks involve**
  - ✓ Correctly exporting the dataset,
  - ✓ Placing them under the right tables,
  - ✓ Structuring them, and
  - ✓ Exporting them in the correct format.

Understanding Data Science Cont..

Phases of Data Analysis: Data Cleaning

- Preprocessed data is still not ready for detailed analysis.
- It must be correctly transformed for an incompleteness check, duplicates check, error check, and missing value check.
- **Tasks performed are**
  - Matching the correct record,
  - Finding inaccuracies in the dataset,
  - Understanding the overall data quality,
  - Removing duplicate items, and
  - Filling in the missing values, and outlier detection
- **Data cleaning is dependent on types of data under study.**

Understanding Data Science Cont..

Phases of Data Analysis: Exploratory Data Analysis (EDA)

- Exploratory data analysis, is the stage where we actually start to understand message contained in data.
- Need several types of data transformation techniques
- Part of this course is dedicated to tasks involved in exploratory data analysis.

Understanding Data Science Cont..

Phases of Data Analysis: Modeling and Algorithm

- Model always describes the relationship between independent and dependent variables.
- Inferential statistics deals with quantifying relationships between particular variables.
- Example of inferential statistics - Regression analysis.

Understanding Data Science Cont..

Phases of Data Analysis: Data Product

- Any computer software that uses data as inputs, produces outputs, and provides feedback based on the output to control the environment is referred to as a **data product**.
- Based on a model developed during data analysis, for example, a recommendation model.
- Recommendation model inputs user purchase history and recommends a related item that user is highly likely to buy.

Understanding Data Science Cont..

Phases of Data Analysis: Communication

- This stage deals with disseminating results to end stakeholders to use result for business intelligence.
- Most notable steps - data visualization.
- Visualization deals with information relay techniques such as tables, charts, summary diagrams, and bar charts to show analyzed result.

The significance of EDA

- Different fields of science, economics, engineering, and marketing accumulate and store data primarily in electronic databases.
- Exploratory data analysis - first exercise in data mining
- EDA allows to visualize data to understand it as well as to create hypotheses for further analysis.
- Centers around creating a synopsis of data or insights for the next steps in a data mining project
- EDA actually reveals ground truth about content
- EDA can understand what type of modeling and hypotheses can be created.
- Key components of exploratory data
    1. analysis include summarizing data,
    2. statistical analysis, and
    3. visualization of data
- Python provides expert tools for EDA.

The significance of EDA Cont…

Steps in EDA

- Problem definition
- Data preparation
- Data analysis
- Development and representation of the results

The significance of EDA Cont…

Steps in EDA : Problem definition

- Define business problem to be solved.
- The main tasks involved in problem definition are
  - Defining main objective of the analysis,
  - Defining main deliverables,
  - Outlining main roles and responsibilities,
  - Obtaining current status of the data,
  - Defining timetable, and
  - Cost/benefit analysis.
- Create execution plan

The significance of EDA Cont…

Steps in EDA : Data preparation

- This step involves methods for preparing the dataset before actual analysis.
  - Define sources of data, data schemas, and tables,
  - Understand main characteristics of data,
  - Clean dataset,
  - Delete non-relevant datasets,
  - Transform data, and
  - Divide data into required chunks for analysis.

The significance of EDA Cont…

Steps in EDA : Data Analysis

- Most crucial steps deals with descriptive statistics and analysis of data.
- The main tasks involve summarizing data,
- Finding hidden correlation and relationships among data,
- Developing predictive models,
- Evaluating models, and calculating accuracies.
- Some of Techniques for data summarization - summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, grouping, and mathematical models.

The significance of EDA Cont…

Steps in EDA : Development & representation of Results

- Involves presenting the dataset to target audience (graphs, summary tables, maps, and diagrams).
- Results should be interpretable by business stakeholders,
- Most of the graphical analysis techniques include scattering plots, character plots, histograms, box plots, residual plots, mean plots, and others.

Making sense of Data

- Need to identify type of data under analysis.
- Different disciplines store different kinds of data for different purposes.
  - medical researchers store patients' data,
  - universities store students' and teachers' data, and
  - real estate industries store house and building datasets.
- A dataset contains many observations about a particular object.
- Dataset about patients can contain many observations.
  - patient identifier (ID),
  - name,
  - address,
  - weight,
  - date of birth,
  - address,
  - email, and
  - gender.

Making sense of Data Cont…

- Each of these features that describes a patient is a variable.
- Each observation can have a specific value for each of these variables
- Data stored in database management system in tables/schema.

PATIENT_ID	NAME	ADDRESS	DOB	EMAIL	Gender	WEIGHT
001	Suresh Kumar Mukhiya	Mannsverk, 61	30.12.1989	skmu@hvl.no	Male	68
002	Yoshmi Mukhiya	Mannsverk 61, 5094, Bergen	10.07.2018	yoshmimukhiya@gmail.com	Female	1
003	Anju Mukhiya	Mannsverk 61, 5094, Bergen	10.12.1997	anjumukhiya@gmail.com	Female	24
004	Asha Gaire	Butwal, Nepal	30.11.1990	aasha.gaire@gmail.com	Female	23
005	Ola Nordmann	Danmark, Sweden	12.12.1789	ola@gmail.com	Male	75

Making sense of Data Cont…

- Most of dataset variables broadly falls into two groups—
  - Numerical data and
  - Categorical data.

Making sense of Data Cont…

Numerical Data

- This data has a sense of measurement involved in it.
- For example, a person's age, height, weight, blood pressure
- This data is often referred to as **quantitative data** in statistics.
- The numerical dataset can be either
  - Discrete or
  - Continuous types.

Making sense of Data Cont…

Discrete data

- This is data that is countable and its values can be listed out.
- For example, if we flip a coin, the number of heads in 200 coin flips can take values from 0 to 200 (finite) cases.
- A variable that represents a discrete dataset is referred to as a discrete variable.
- The discrete variable takes a fixed number of distinct values.
- For example,
  - Country variable can have values such as Nepal, India, Norway, and Japan.
  - Rank variable of a student in a classroom can take values from 1, 2, 3, 4, 5, and so on.

Making sense of Data Cont…

Continuous Data

- A variable that can have an infinite number of numerical values within a specific range is classified as **continuous data.**
- A variable describing continuous data is a continuous variable.
- For example, temperature of your city today
- Similarly, weight of car variable is a continuous variable.

Making sense of Data Cont…

Categorical Data

- This type of data represents the characteristics of an object.
- For example, gender, marital status, type of address, or categories of the movies.
- This data is often referred to as **qualitative datasets** in statistics.
- Variable describing categorical data is referred as categorical variable.
- Types of categorical variables:
  - A binary categorical variable can take exactly two values, referred to as a **dichotomous variable**.
  - **Polytomous variables** are categorical variables that can take more than two possible values. E.g. marital status

Measurement Scales

- **Scales of measurement** refer to ways in which variables/numbers are defined and categorized.
- Each scale of measurement has certain properties which in turn determines appropriateness for use of certain statistical analyses.
- Four types of measurement in Statistics
  - Nominal,
  - Ordinal,
  - Interval, and
  - Ratio

Measurement Scales - Nominal

- Used for labeling variables without any quantitative value.
- Scales are generally referred to as **labels**.
- Mutually exclusive and do not carry any numerical importance.
- **Example**:

What is your gender?

- Male
- Female
- Third gender/Non-binary
- I prefer not to answer
- Other

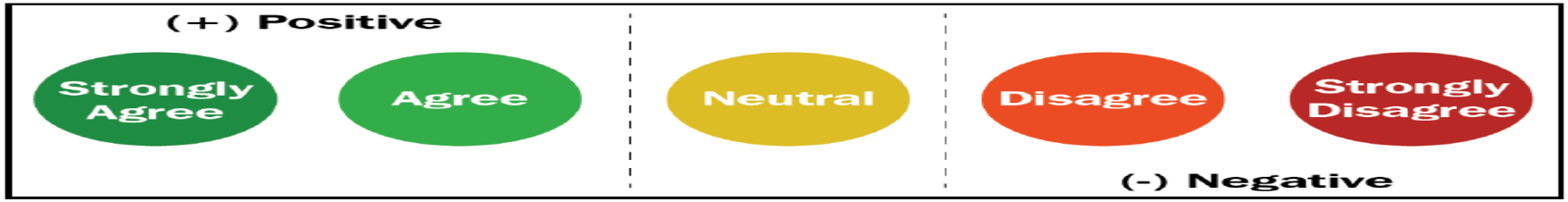
Why should you care about whether data is nominal or ordinal?

- For nominal dataset, you can certainly know-
  - ✓ **Frequency** is the rate at which a label occurs over a period of time
  - ✓ **Proportion** - calculated by dividing frequency by total no. of events.
  - ✓ Could compute the **percentage** of each proportion.
  - ✓ **Visualize** using pie chart or a bar chart.

Measurement Scales - Ordinal

- Main difference in ordinal and nominal scale is the order.
- Order of the values is a significant factor.
- **Likert scale**, uses a variation of an ordinal scale?
- Likert scale:

  WordPress is making content managers' lives easier. How do you feel about this statement?



How do you feel today?

- ● 1 - Very Unhappy
- ○ 2 - Unhappy
- ○ 3 - OK
- ○ 4 - Happy
- ○ 5 - Very Happy

How satisfied are you with our service?

- ● 1 - Very Unsatisfied
- ○ 2 - Somewhat Unsatisfied
- ○ 3 - Neutral
- ○ 4 - Somewhat Satisfied
- ○ 5 - Very Satisfied

- Median is allowed as measure of central tendency
- Average is not permitted

Measurement Scales - Interval

- Both order and exact differences between values are significant.
- Variables that have familiar, constant, and computable differences are classified using the Interval scale.
- Interval scales are widely used in statistics, E.g, in the *measure of central tendencies—mean, median, mode, and standard deviations.*
- The mean, median, and mode are allowed on interval data.
- Interval' indicates 'distance between two entities
- Examples include location in Cartesian coordinates and direction measured in degrees from magnetic north.
- Fahrenheit temperature scale – 80 degrees is always higher than 50 degrees and difference between these two temperatures is the same as the difference between 70 degrees and 40 degrees.

Measurement Scales - Interval

- Both order and exact differences between values are significant.
- Variables that have familiar, constant, and computable differences are classified using the Interval scale.
- Interval scales are widely used in statistics, E.g, in the *measure of central tendencies—mean, median, mode, and standard deviations.*
- The mean, median, and mode are allowed on interval data.
- Interval Scale Examples
  - ✓ Examples include location in Cartesian coordinates and direction measured in degrees from magnetic north.
  - ✓ temperature scale (values are already established, constant, and measurable)
  - ✓ Time and Calendar years

Measurement Scales - Interval

Want an easy guide to the interval scale and its data?

- Interval scale is preferred to nominal scale or ordinal scale because the latter two are qualitative scales. The interval scale is quantitative in the sense that it can quantify the difference between values.
- Interval data can be discrete with whole numbers like 8 degrees, 4 years, 2 months, etc., or continuous with fractional numbers like 12.2 degrees, 3.5 weeks or 4.2 miles.
- You can subtract values between two variables that help understand the difference between two variables.
- Interval measurement allows you to calculate mean and median of variables.
- Interval data is especially useful in business, social, and scientific analysis and strategy because it is straightforward and quantitative.
- Preferred scale in statistics because can assign a numerical value to any arbitrary assessment, such as feelings and sentiments.

Measurement Scales - Ratio

- Ratio scales contain order, exact values, and absolute zero.
- Used in descriptive and inferential statistics.
- These scales provide numerous possibilities for statistical analysis.
- Mathematical operations, measure of central tendencies, and measure of dispersion and coefficient of variation can also be computed from such scales.
- Examples include a measure of energy, mass, length, duration, electrical energy, plan angle, and volume.

Measurement Scales

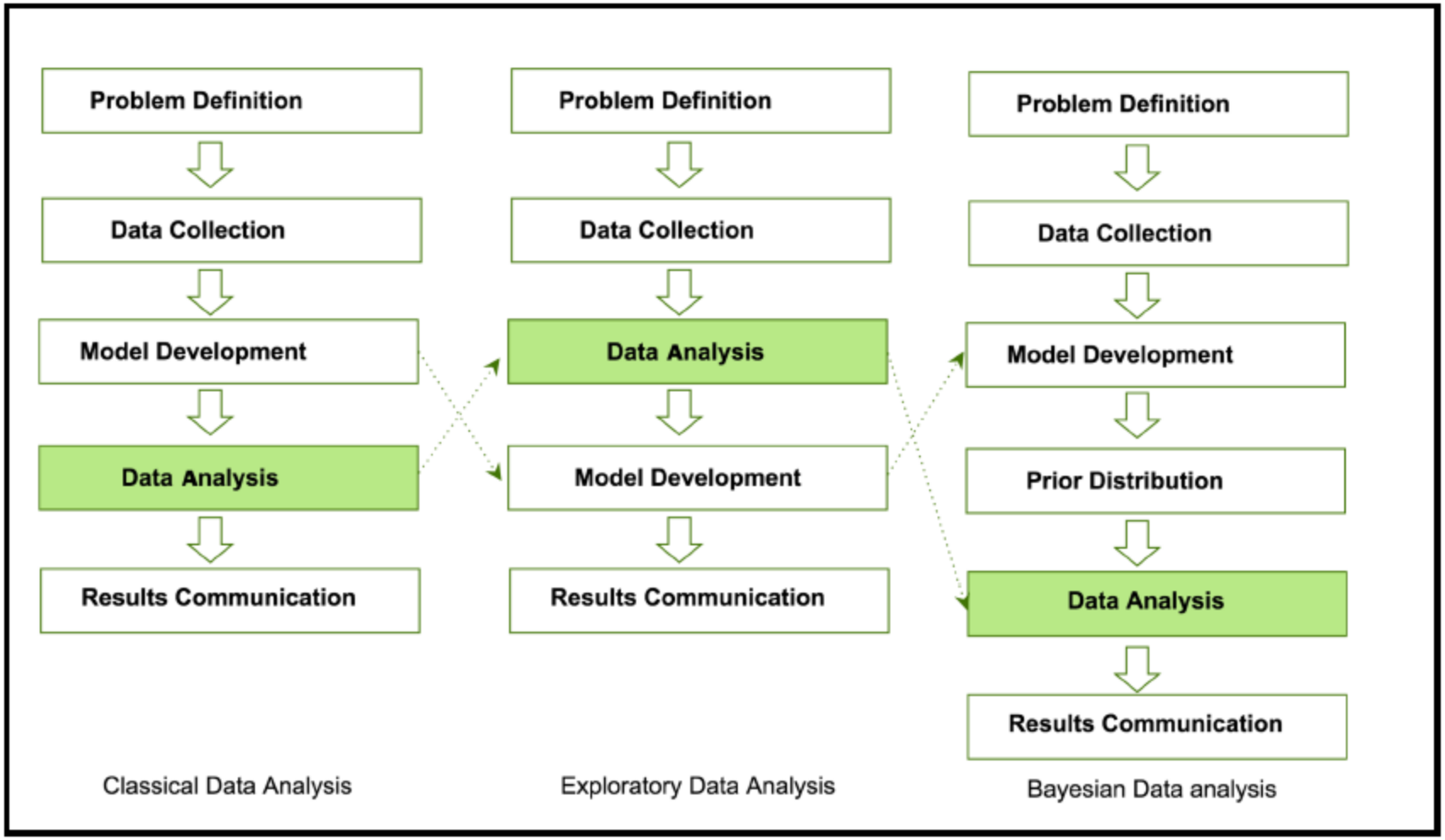
Data types and Scale measures

Provides:	Nominal	Ordinal	Interval	Ratio
The "order"of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓

Approaches to Data Analysis

1. **Classical Data Analysis**
2. **Exploratory Data Analysis**
3. **Bayesian Data Analysis**

Comparing EDA with classical and Bayesian analysis



Software tools available for EDA

- **Python**:
  - Open source programming language
  - Widely used in data analysis, data mining, and data science
- **R programming language**:
  - Open source programming language
  - Widely utilized in statistical computation and graphical data analysis
- **WEKA**:
  - Open source data mining package
  - Several EDA tools and algorithms
- **KNIME**:
  - Open source tool for data analysis
  - Based on Eclipse

Visual Aids for Exploratory Data Analysis

Visual Aids for Exploratory Data Analysis

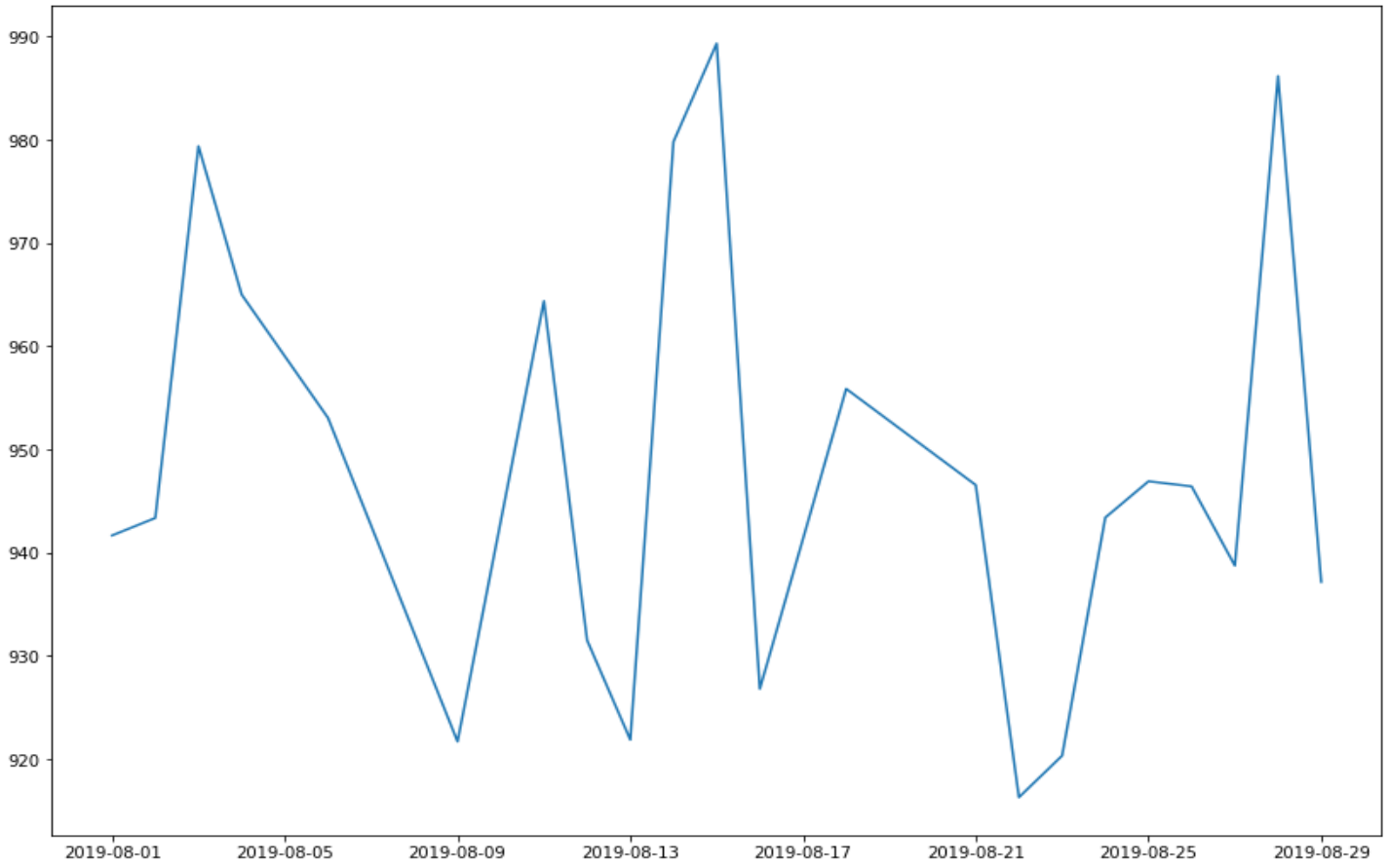
- Important goals of Data Scientist
  1. Extract knowledge from the data and
  2. Present the data to stakeholders.
- Presenting results to stakeholders is very complex
- Visual aids are very useful tools.

Visual Aids for Exploratory Data Analysis

- Line chart
- Bar chart
- Scatter plot
- Area plot and stacked plot
- Pie chart
- Table chart
- Polar chart
- Histogram
- Lollipop chart

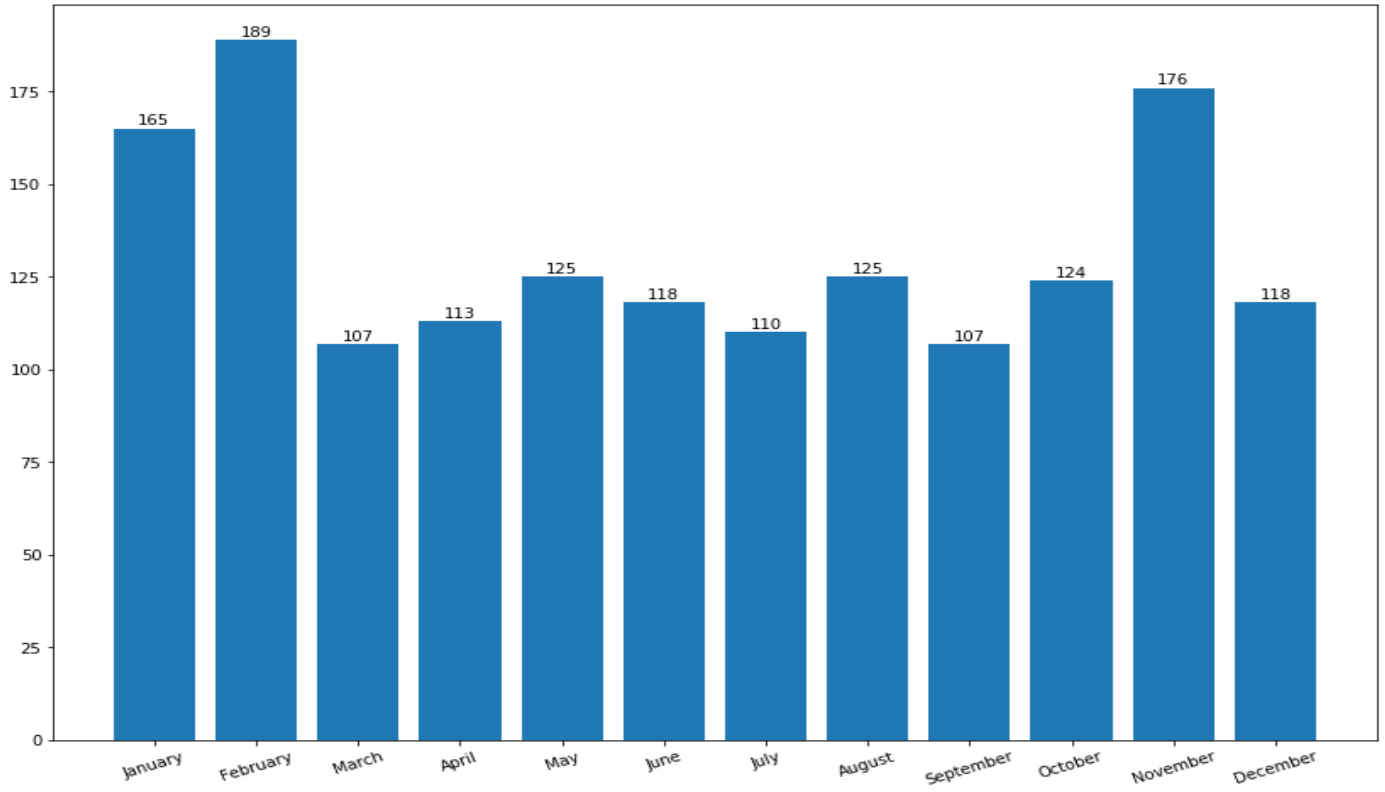
Line Chart

A **line chart** or **line plot** or **line graph** or **curve chart** is a type of chart which displays information as a series of data points called 'markers' connected by **straight line segment**



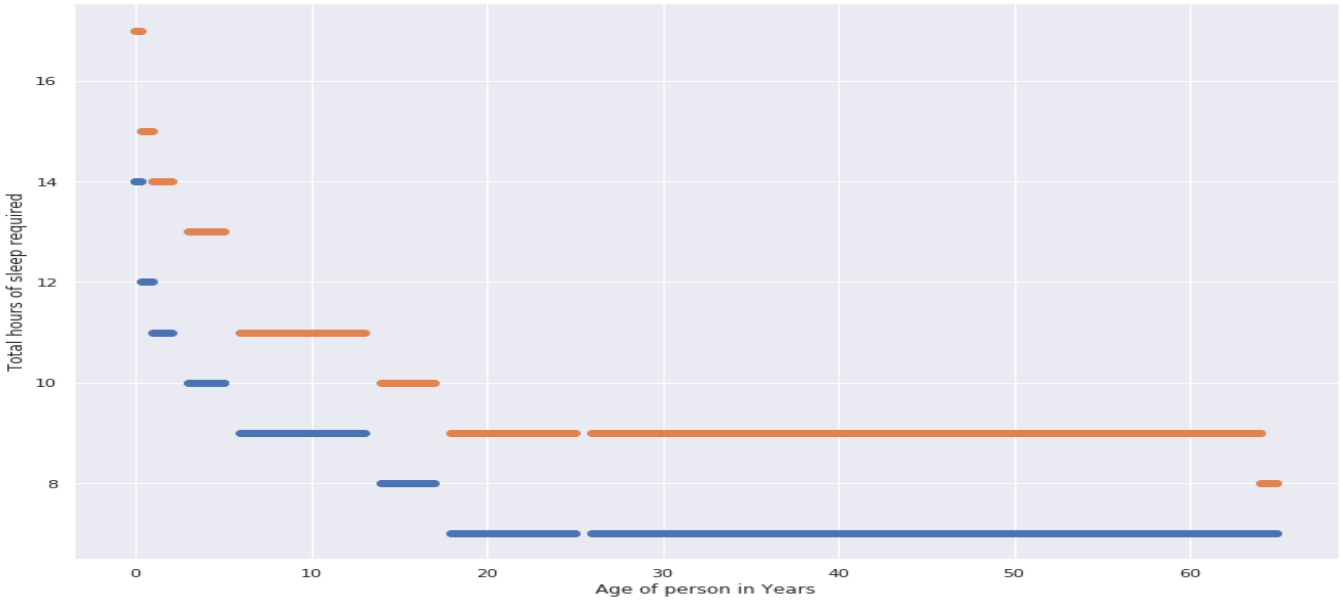
Bar Charts

- Most common types of visualization
- Bars can be drawn horizontally or vertically
- Bar charts are very convenient when the changes are large

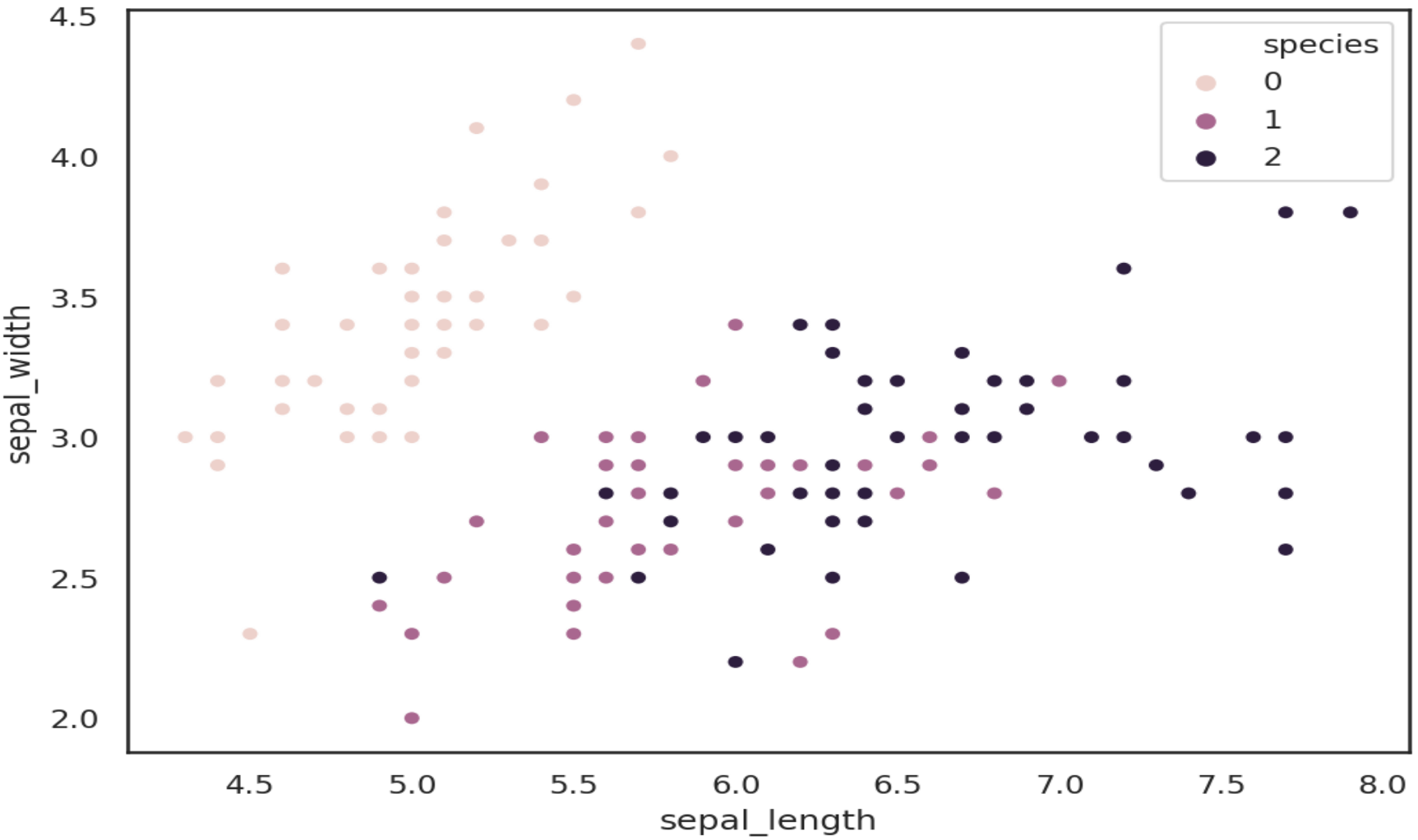


Scatter Plot

- Called as scatter graphs, scatter charts, scattergrams, and scatter diagrams.
- Use a **Cartesian coordinates system** to display values
- When should we use a scatter plot?
  - ✓ When one continuous variable is dependent on another variable
  - ✓ When both continuous variables are independent
- Used when we need to show relationship between two variables
- Referred as correlation plots



Scatter Plot

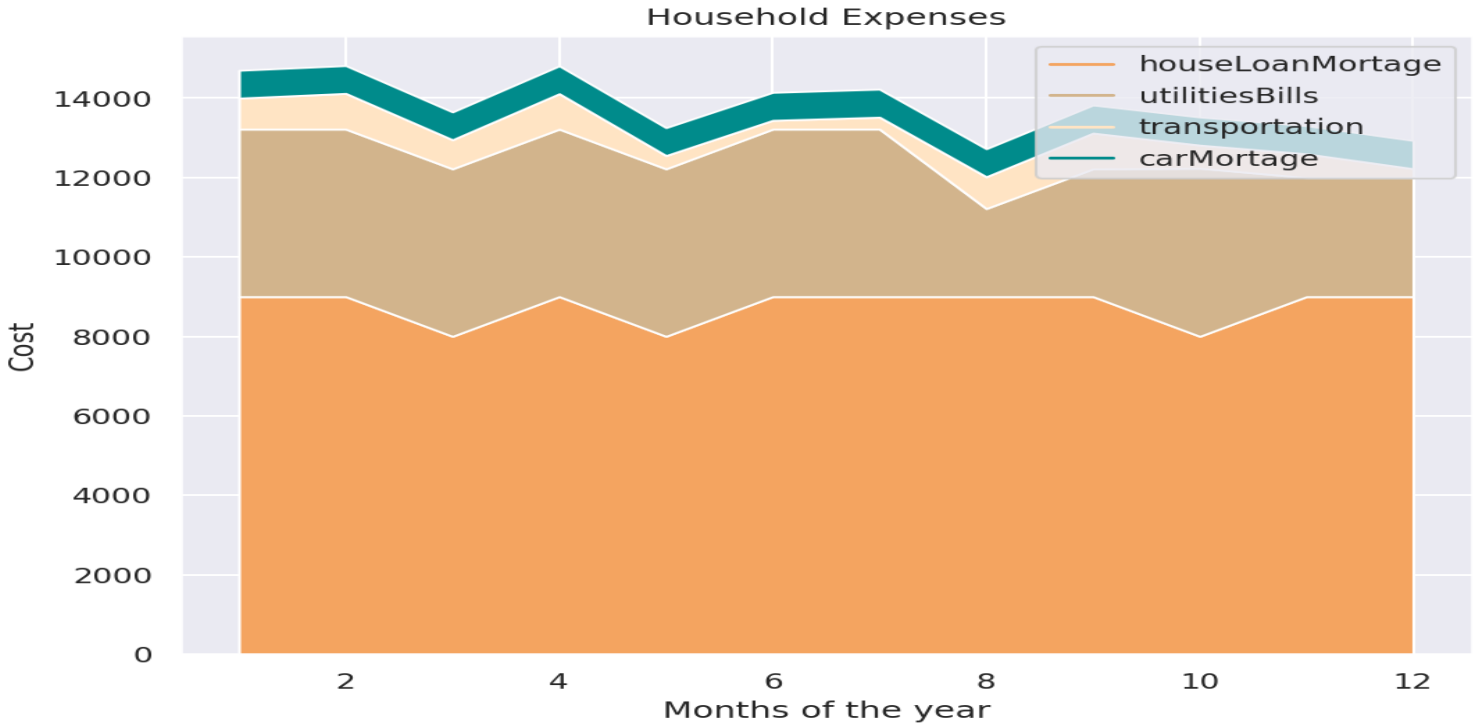


Bubble Chart

- Manifestation of scatter plot where each data point on graph is shown as a bubble.
- Each bubble can be illustrated with a different color, size, and appearance.
-

Area Plot and Stacked Plot

- Stacked plot owes its name to the fact that it represents the area under a line plot and that several such plots can be stacked on top of one another, giving the feeling of a stack.
- Useful when want to visualize **cumulative effect** of multiple variables being plotted on the *y* axis.



- House mortgage loan is largest expense since area under the curve is largest.

Pie Chart

- More interesting types of data visualization graphs
- Purpose of the pie chart is to communicate proportions

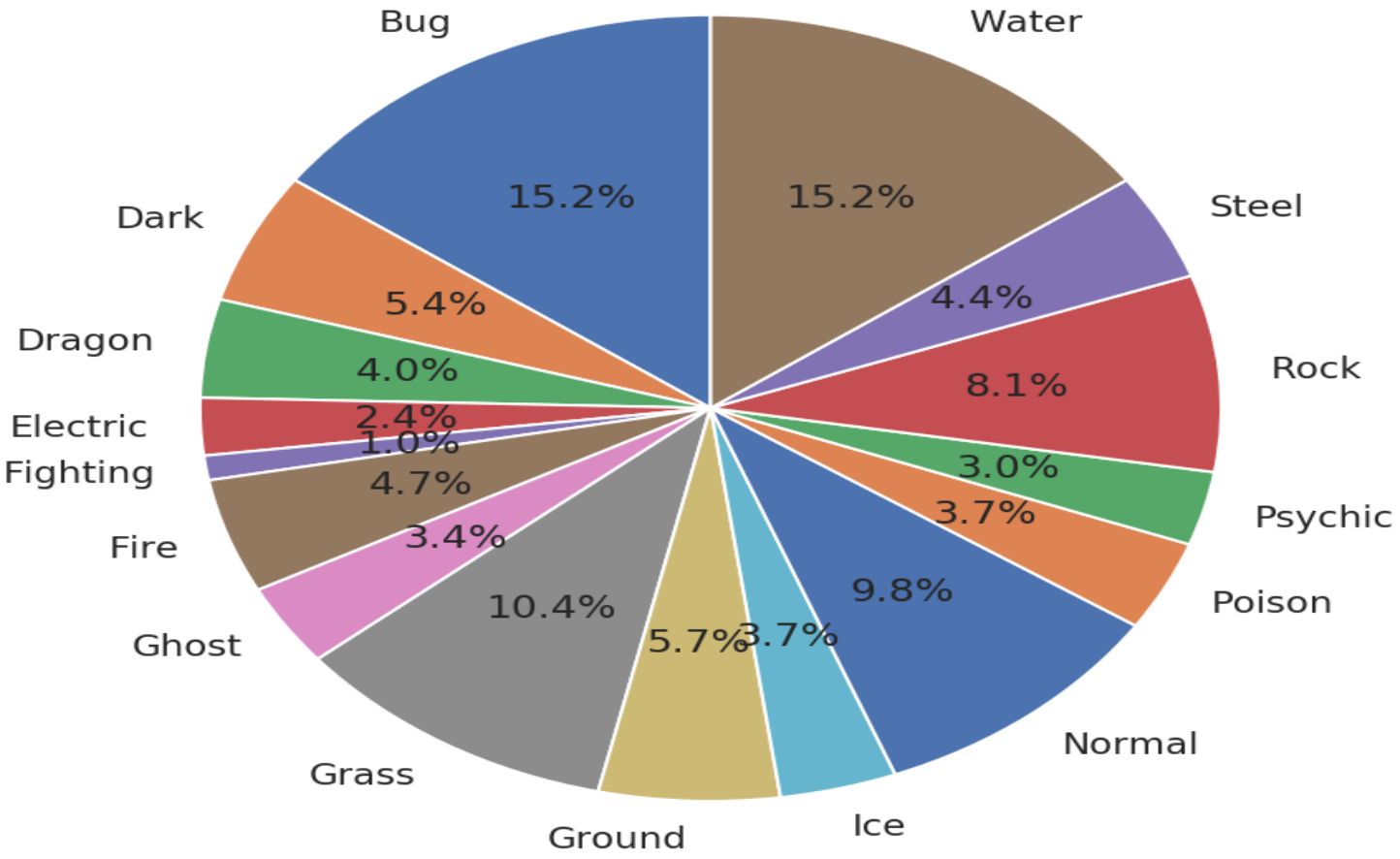
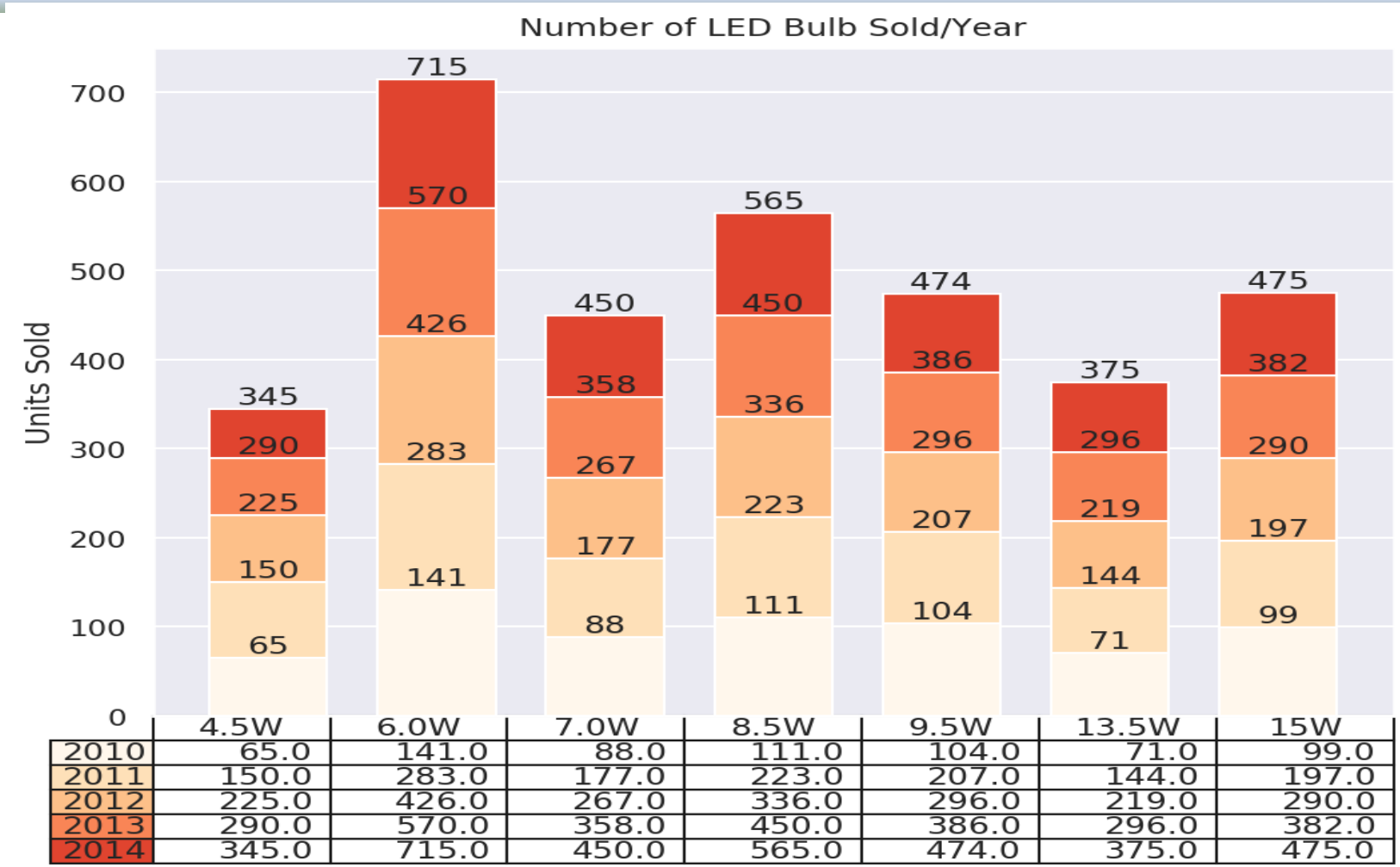


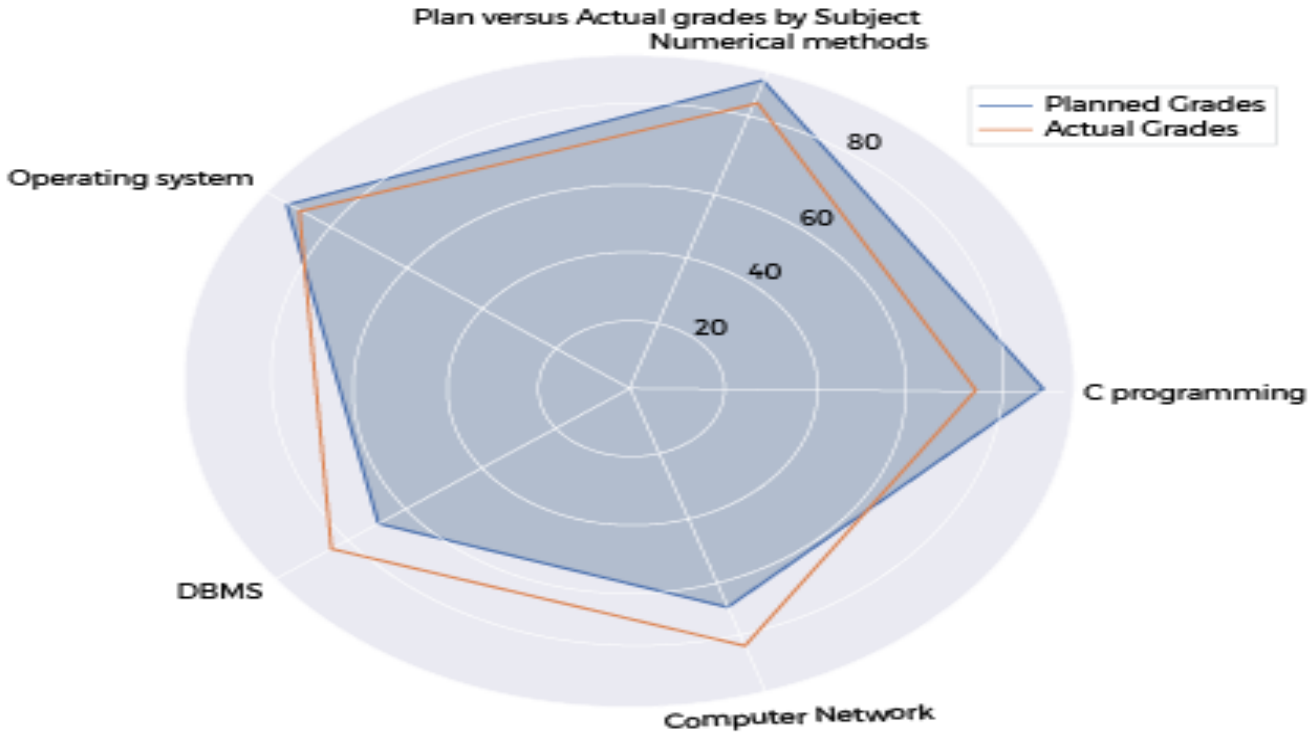
Table Chart



	4.5W	6.0W	7.0W	8.5W	9.5W	13.5W	15W
2010	65.0	141.0	88.0	111.0	104.0	71.0	99.0
2011	150.0	283.0	177.0	223.0	207.0	144.0	197.0
2012	225.0	426.0	267.0	336.0	296.0	219.0	290.0
2013	290.0	570.0	358.0	450.0	386.0	296.0	382.0
2014	345.0	715.0	450.0	565.0	474.0	375.0	475.0

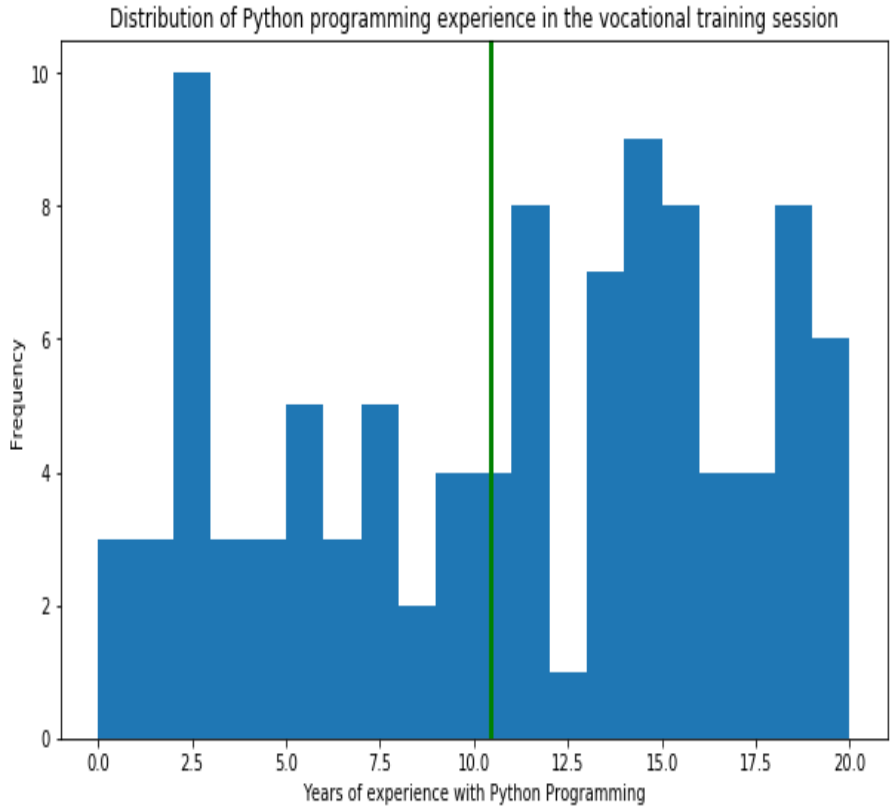
Polar Chart

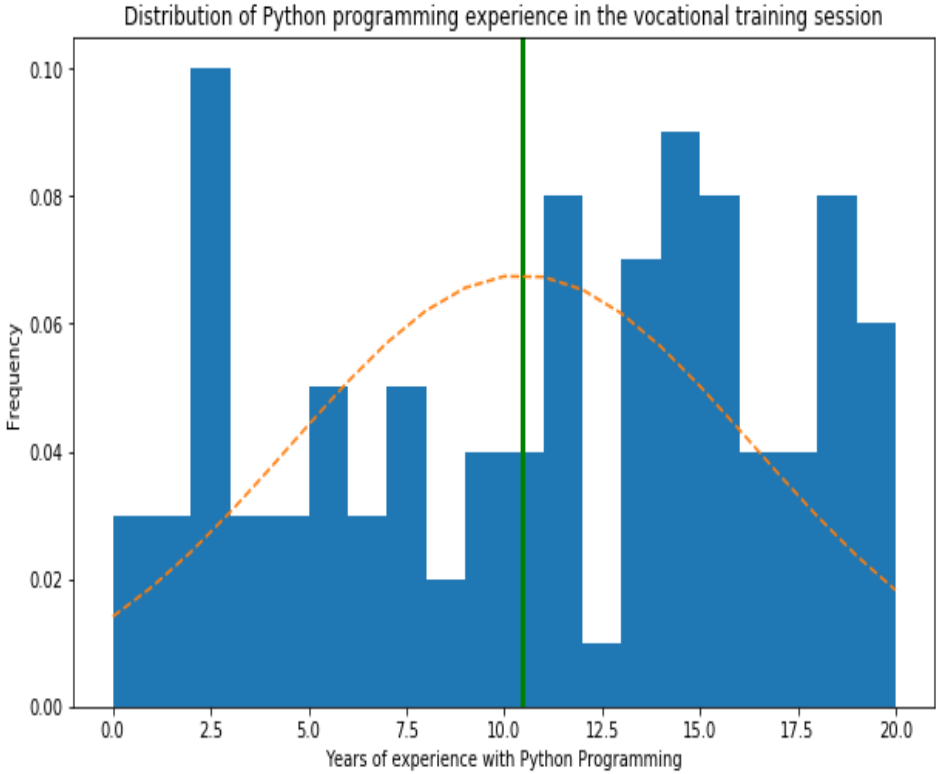
- Polar chart is a diagram that is plotted on a polar axis
- Referred to as a spider web plot
- Planned and actual grades by subject can easily be distinguished



Histogram

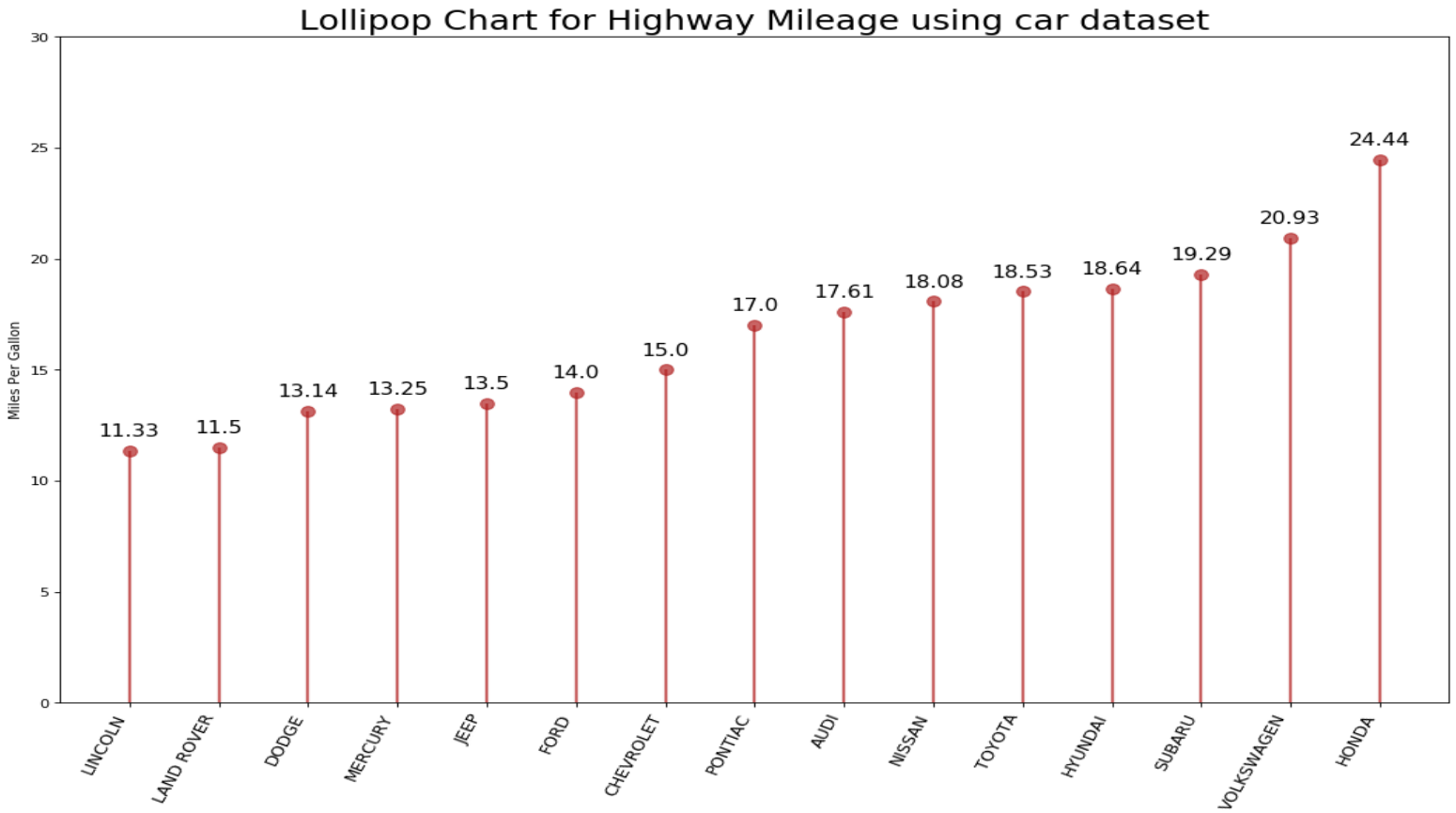
- Used to depict distribution of any continuous variable
- Very popular in statistical analysis





Lollipop Chart

- Can be used to display ranking in the data
- Similar to an ordered bar chart.



Choosing the Best Chart

- No standard that defines which chart to use to visualize data
- **Some Guidelines**
  - ✓ what type of data – continuous variables -> histogram
  - ✓ Want to show ranking, an ordered bar chart would be a good choice.
  - ✓ Choose chart that effectively conveys right and relevant meaning of the data without actually distorting the facts
  - ✓ Simplicity is best - better to draw a simple chart
  - ✓ Choose a diagram that does not overload the audience with information

Choosing the Best Chart - 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

Choosing the Best Chart - charts based on the purposes

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

Choosing the Best Chart - charts based on the purposes

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

Summary

- Studied most fundamental theory behind data analysis and Exploratory Data Analysis (EDA).
- EDA is one of most prominent steps in data analysis.
- A dataset contains many observations about a particular object.
- Four types of data measurement scales: nominal, ordinal, interval, and ratio.
- Presenting results to stakeholders is very complex.
- Visual aids are widely used.

Thank You !!!