CYBR 520

Chapter 1: Exploratory Data Analysis Fundamentals

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

WestVirginiaUniversity.
JOHN CHAMBERS COLLEGE OF
BUSINESS AND ECONOMICS

TOPICS

- 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

Objectives

In this chapter, we are going to learn and revise the following topics:

1. Understanding data science
2. The significance of EDA
3. Making sense of data
4. Comparing EDA with classical and Bayesian analysis
5. Software tools available for EDA
6. Getting started with EDA

Exploratory Data Analysis Fundamentals

- We are surrounded with data
- *Data* encompasses a collection of discrete objects, numbers, words, events, facts, measurements, observations, or even descriptions of things.
  - every event or process occurring in several disciplines, including biology, economics, engineering, marketing, and others

Exploratory Data Analysis Fundamentals

- Processing such data elicits useful *information* and processing such information generates useful knowledge.
- But an important question is: how can we generate meaningful and useful information from such data?

Exploratory Data Analysis Fundamentals

- An answer to this question is EDA.
- EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures.
- In this chapter, we are going to discuss the steps involved in performing top-notch exploratory data analysis and get our hands dirty using some open source databases.

UNDERSTANDING DATA SCIENCE

The Essentials of Data Science

- Data science is currently a hot topic, and it's evolving rapidly.
- The role of data scientists is changing – it's not just about building models anymore.
- Data science is at its peak in terms of popularity and demand.
- Skills required for data scientists are constantly evolving.
- We now expect data scientists to not only build models but also explain and utilize results for business intelligence

Skillset for Data Scientists

- People often ask, "What skills do I need to become a top-notch data scientist?"
- Contrary to popular belief, you don't necessarily need a Ph.D. in data science.
- Data science requires cross-disciplinary knowledge in computer science, data, statistics, and mathematics.

Phases of Data Analysis

1. Data Requirements
2. Data Collection
3. Data Processing
4. Data Cleaning
5. Exploratory Data Analysis
6. Modeling and Algorithms
7. Data Product and Communication

1. Data Requirements

- It is important to comprehend what type of data is required for the organization to be collected, curated, and stored.
- All of these data points are required to correctly tackle the problem (e.g., health, cyber, etc…)
- Hence, these are mandatory requirements for the application.
  - It is required to categorize the data, numerical or categorical, and the format of storage and dissemination.

2. Data Collection

- Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company.
- As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

3. 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

4. Data cleaning

- Preprocessed data is not yet ready for detailed analysis.
  - It must undergo several transformations to ensure data quality.
- Data cleaning is a vital stage that involves:
  - Checking for incompleteness
  - Identifying and removing duplicates
  - Detecting errors
  - Handling missing values
- Identifying data anomalies requires analytical techniques.
- Data cleaning's approach varies depending on the dataset types.
  - For instance, quantitative data cleaning may involve outlier detection methods.

5. Exploratory Data Analysis

- The stage where we actually start to understand the message contained in the data.
- Several types of data transformation techniques might be required during the process of exploration.
- We will cover descriptive statistics in-depth in *Section 2*, Chapter 5, *Descriptive Statistics*, to understand the mathematical foundation behind descriptive statistics.

6. Modeling and Algorithms

- In data science, models and mathematical formulas represent relationships among variables.
  - These models involve variables that depend on others to cause an event.
- Models describe relationships between <u>independent</u> and <u>dependent</u> variables.
- Inferential statistics quantifies these relationships.

Example

- Total Price of Pens (Total) = Unit Price (UnitPrice) * Quantity Bought (Quantity)

=> Model: Total = UnitPrice * Quantity

- Dependent Variable: Total Price (Depends on Unit Price)

- Independent Variable: Unit Price

7. Data Product

- Software that uses data
- Gives outputs
- Uses feedback to control things
  - Intrusion Detection System (IDS)
  - Inputs: Network Traffic Data
  - Output: Alerts on Suspicious Activities

8. **Communication**

- This stage deals with disseminating the results to end stakeholders to use the result for *business intelligence*.
- One of the most notable steps in this stage is data visualization.
- Visualization deals with information relay techniques such as tables, charts, summary diagrams, and bar charts to show the analyzed result.

THE SIGNIFICANCE OF EDA

The significance of EDA

- Different fields rely on electronic databases to store data.
- Effective decision-making requires extracting insights from large datasets.
- Computer programs are essential for analyzing datasets with numerous data points.
- Data mining is the process of extracting valuable insights from data.

The significance of EDA

- Exploratory Data Analysis (EDA) is a crucial first step in data mining.
- EDA helps visualize data, create hypotheses, and understand its content.
- EDA uncovers ground truth without making assumptions.
- Key components of EDA include data summarization, statistical analysis, and visualization.
- Python offers powerful tools for EDA, such as `pandas`, `scipy`, `matplotlib`, and `plotly`.

Steps in EDA

1. Problem definition
2. Data preparation
3. Data analysis
4. Development and representation of the results

1. Problem definition

- Before trying to extract useful insight from the data, it is essential to define the business problem to be solved.
- The problem definition works as the driving force for a data analysis plan execution.

1. Problem definition- Tasks

- Defining the main objective of the analysis
- Defining the main deliverables
- Outlining the main roles and responsibilities
- Obtaining the current status of the data,
- Defining the timetable
- Performing cost/benefit analysis.
- Based on such a problem definition, an execution plan can be created.

2. Data preparation

- Define the sources of data
- Define data schemas and tables
- Understand the main characteristics of the data
- Clean the dataset
- Delete non-relevant datasets
- Transform the data
- Divide the data into required chunks for analysis.

3. Data analysis

- This is one of the most crucial steps that deals with descriptive statistics and analysis of the data.
  - Summarize the data
  - Finding the hidden correlation and relationships among the data
  - Developing predictive models
    - Evaluating the models and calculating the accuracies.
  - Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

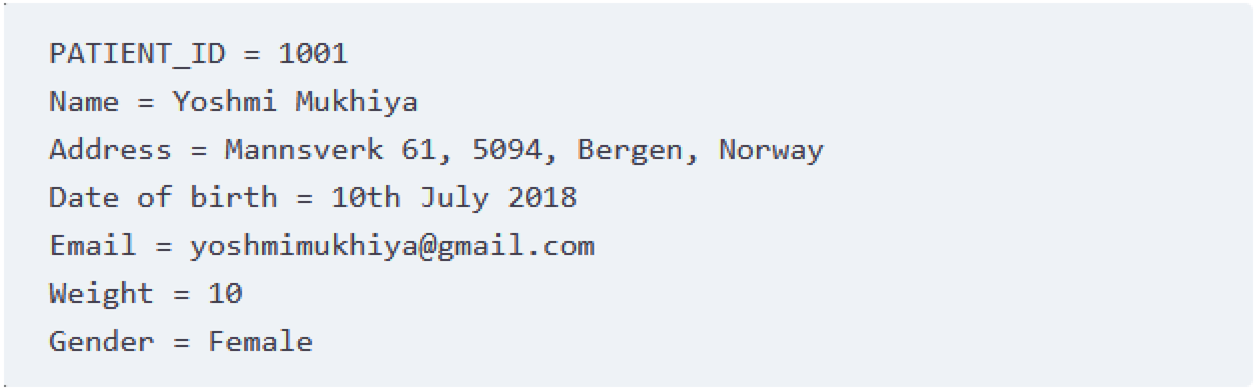
4. Development and representation of the results

- Presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams.
- The result analyzed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA.
- Most of the graphical analysis techniques include scattering plots, character plots, histograms, box plots, residual plots, mean plots, and others. We will explore several types of graphical representation in

Making sense of data

- It is crucial to identify the type of data under analysis.
- For example, medical researchers store patients' data, universities store students' and teachers' data, and real estate industries storehouse and building datasets.
- A dataset contains many observations about a particular object.

Making sense of data

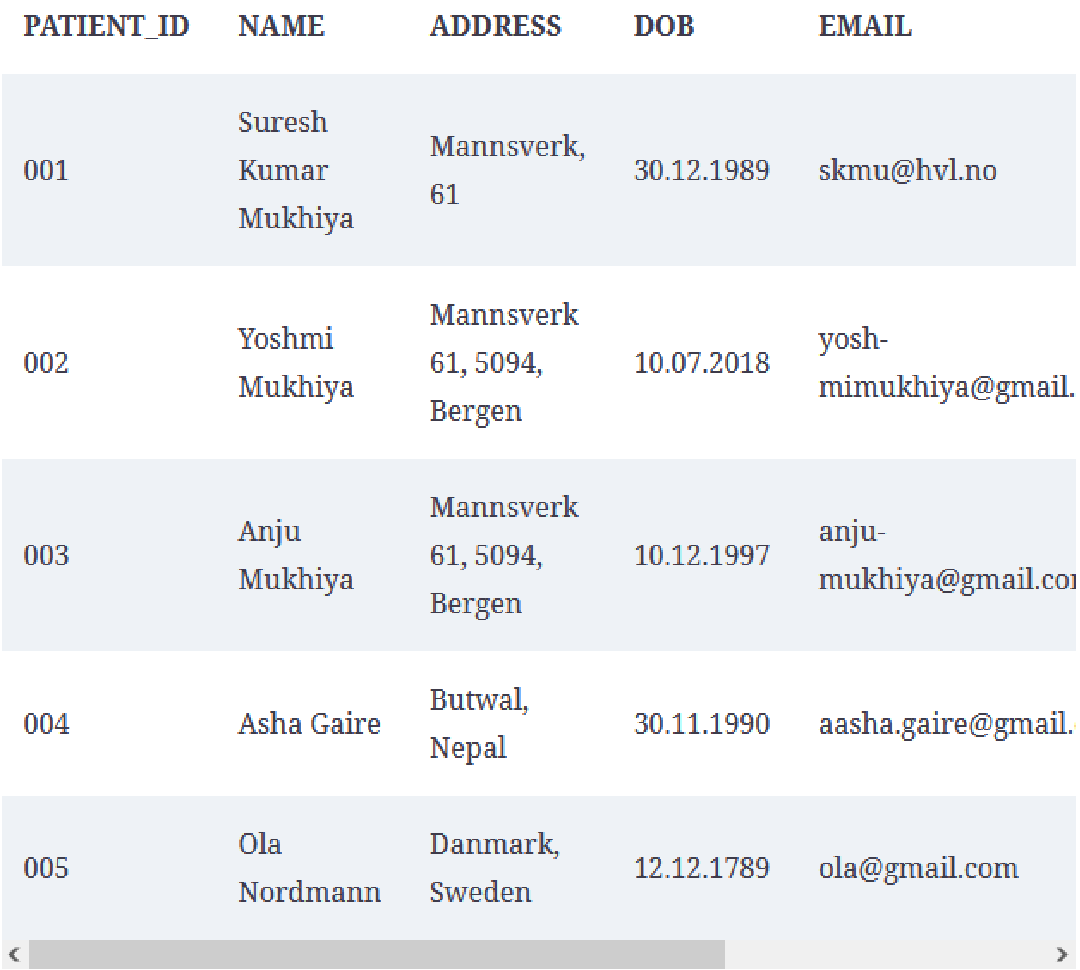


```
PATIENT_ID = 1001
Name = Yoshmi Mukhiya
Address = Mannsverk 61, 5094, Bergen, Norway
Date of birth = 10th July 2018
Email = yoshmimukhiya@gmail.com
Weight = 10
Gender = Female
```

- For instance, a dataset about patients in a hospital can contain many observations. A patient can be described by a *patient identifier (ID), name, address, weight, date of birth, address, email,* and *gender*.
- Each of these features that describes a patient is a variable. Each observation can have a specific value for each of these variables. For example, a patient can have the information provided above:

Making sense of data

- These datasets are stored in hospitals and are presented for analysis. Most of this data is stored in some sort of database management system in tables/schema.
- An example of a table for storing patient information is shown in the Figure



PATIENT_ID	NAME	ADDRESS	DOB	EMAIL
001	Suresh Kumar Mukhiya	Mannsverk, 61	30.12.1989	skmu@hvl.no
002	Yoshmi Mukhiya	Mannsverk 61, 5094, Bergen	10.07.2018	yosh-mimukhiya@gmail.
003	Anju Mukhiya	Mannsverk 61, 5094, Bergen	10.12.1997	anju-mukhiya@gmail.co
004	Asha Gaire	Butwal, Nepal	30.11.1990	aasha.gaire@gmail.
005	Ola Nordmann	Danmark, Sweden	12.12.1789	ola@gmail.com

Making sense of data

- To summarize the preceding table, there are four observations (001, 002, 003, 004, 005).
- Each observation describes variables (PatientID, name, address, dob, email, gender, and weight).
- Most of the dataset broadly falls into two groups:
    1. Numerical data
    2. Categorical data

1. Numerical data

- This data has a sense of measurement involved in it;
  - a person's age, height, weight, blood pressure, heart rate, temperature, number of teeth, number of bones, and the number of family members.
- This data is often referred to as **quantitative data** in statistics. The numerical dataset can be either:
  a) discrete
  b) continuous

1.a. 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.

1.a. Discrete Data/ Example

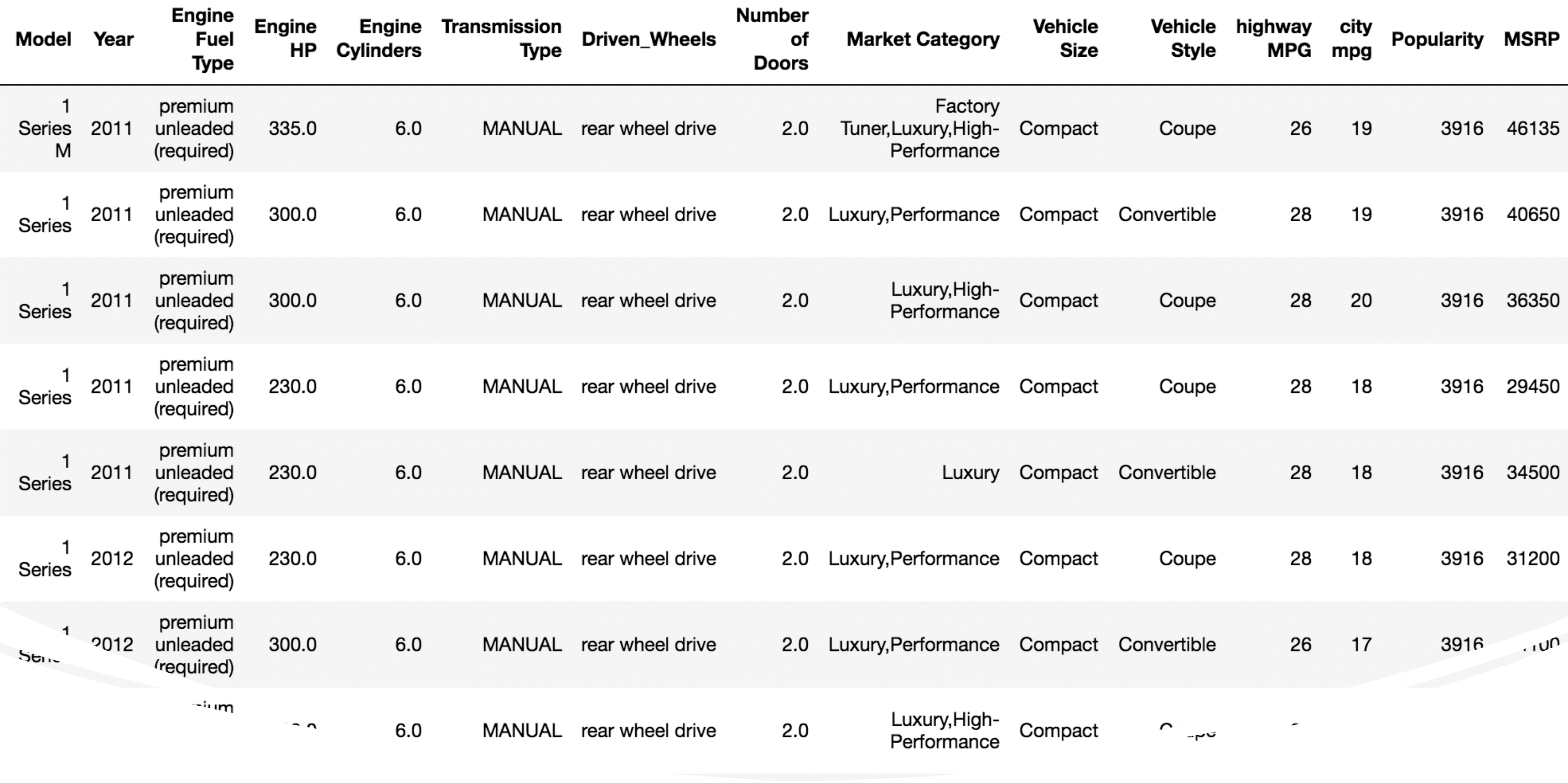
- The discrete variable takes a fixed number of distinct values.
- For example, the Country variable can have values such as Nepal, India, Norway, and Japan. It is fixed.
- The Rank variable of a student in a classroom can take values from 1, 2, 3, 4, 5, and so on.

1.b. 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.
- Continuous data can follow two scales:
  - Interval measure of scale
  - Ratio measure of scale

1.b. Continuous data / Example

- What is the temperature of your city today?
- Can we be finite?
- Similarly, the weight variable in the previous section is a continuous variable.



Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg	Popularity	MSRP
1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916	46135
1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	28	19	3916	40650
1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916	36350
1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916	29450
1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertible	28	18	3916	34500
1 Series	2012	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916	31200
1 Seri	2012	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	26	17	3916	[illegible]
		[illegible]	[illegible]	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	[illegible]				

Question

- Given the car dataset (available on GitHub), identify the data type of each column.

2. Categorical data

- This data is often referred to as **qualitative datasets** in statistics.
- This type of data represents the characteristics of an object;
  - Gender
  - Marital status
  - Type of address
  - Categories of the movies

2. Categorical data

- A variable describing categorical data is referred to as a **categorical variable**.
- These types of variables can have one of a limited number of values.
  - – Also known as enumerated types or enumerations of variables.
- Two types of categorical variables:
  a) Binary categorical
  b) Polytomous variables

2.a. Binary Categorical

- A binary categorical variable can take exactly two values and is also referred to as a **dichotomous variable**.
  - For example, when you create an experiment, the result is either success or failure.
  - Hence, results can be understood as a **binary categorical variable**.

2.a. Polytomous variables

- Categorical variables that can take more than two possible values.
  - For example, marital status can have several values, such as annulled, divorced, interlocutory, legally separated, married, polygamous, never married, domestic partners, unmarried, widowed, domestic partner, and unknown.
  - Since marital status can take more than two possible values, it is a **polytomous variable.**

2. Categorical data/ Examples

- Gender (Male, Female, Other, or Unknown)
- Marital Status (Annulled, Divorced, Interlocutory, Legally Separated, Married, Polygamous, Never Married, Domestic Partner, Unmarried, Widowed, or Unknown)
- Movie genres (Action, Adventure, Comedy, Crime, Drama, Fantasy, Historical, Horror, Mystery, Philosophical, Political, Romance, Saga, Satire, Science Fiction, Social, Thriller, Urban, or Western)
- Blood type (A, B, AB, or O)
- Types of drugs (Stimulants, Depressants, Hallucinogens, Dissociatives, Opioids, Inhalants, or Cannabis)

- Most of the categorical datasets follow either
  1. Nominal measurement scales
  2. Ordinal measurement scales

Measurement scales

- There are four different types of measurement scales described in statistics:

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Nominal Measurement Scale

- These are practiced for labeling variables without any quantitative value.
- The scales are generally referred to as **labels**.
  - And these scales are mutually exclusive and do not carry any numerical importance.
- Order is irrelevant!

Nominal Measurement Scale

- Nominal scales are considered qualitative scales and the measurements that are taken using qualitative scales are considered **qualitative data**.
  - Do not be confused, they are not quantitative
- If, someone uses numbers as labels in the nominal measurement sense, they have no concrete numerical value or meaning.
  - No form of arithmetic calculation can be made on nominal measures.

Nominal Measurement Scale/ Example

- **What is your gender?**
  - Male
  - Female
  - Third gender/Non-binary
  - I prefer not to answer
  - Other
- **Other examples include the following:**
  - The languages that are spoken in a particular country
  - Biological species
  - Parts of speech in grammar (noun, pronoun, adjective, and so on)
  - Taxonomic ranks in biology (Archea, Bacteria, and Eukarya)

Why do we care about Measurements Scales?

- You might be thinking *why should you care about whether data is nominal or ordinal? Should we not just start loading the data and begin our analysis?* Well, we could.
- But think about this: you have a dataset, and you want to analyze it.
- How will you decide whether you can make a pie chart, bar chart, or histogram?
- Are you getting my point?

What can we do with Nominal dataset?

- **Frequency** is the rate at which a label occurs over a period of time within the dataset.
- **Proportion** can be calculated by dividing the frequency by the total number of events.
- Then, you could compute the **percentage** of each proportion.
- And to **visualize** the nominal dataset, you can use either a pie chart or a bar chart.

Ordinal Measurement Scale

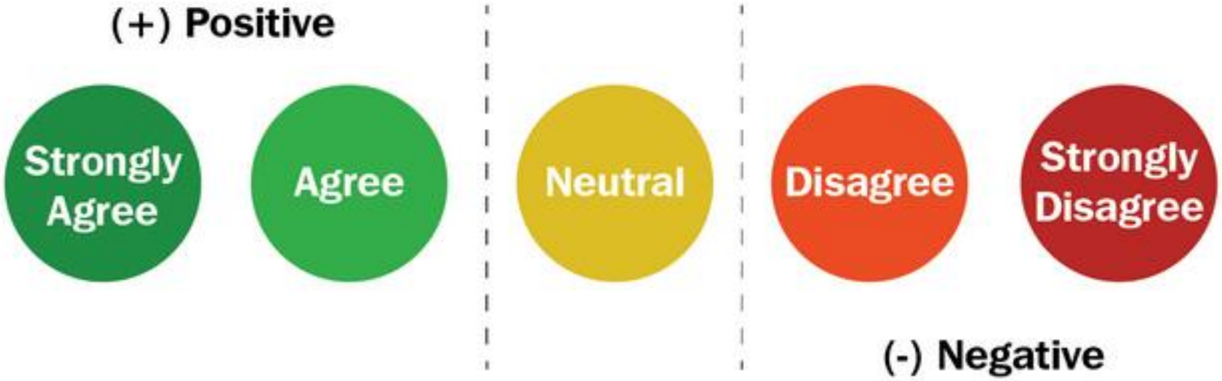
- The main difference in the ordinal and nominal scale is the **order**.
- In ordinal scales, the order of the values is a significant factor.
- To make it easier, consider ordinal scales as an order of ranking (1st, 2nd, 3rd, 4th, and so on).
- The **median** item is allowed as the measure of central tendency; however, the **average** is not permitted.
- **Likert scale**

Likert scale

- A Likert scale is a widely used method for measuring attitudes, opinions, and perceptions in surveys and research.
- It provides respondents with a range of response options to express their level of agreement or disagreement with a statement.

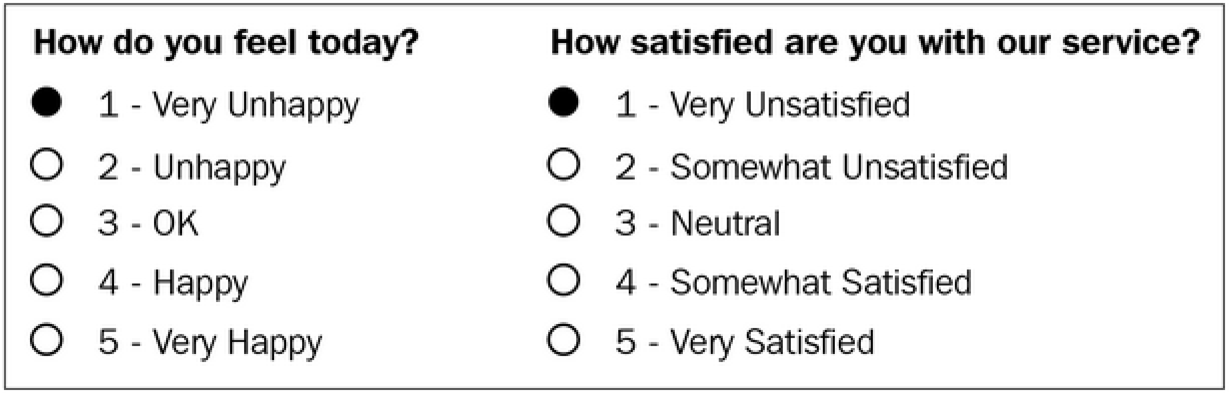
Likert scale- Example1

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



Likert scale- Example2

- *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

Interval Measurements Scale

- In interval scales, both the order and exact differences between the values are significant. Interval scales are widely used in statistics, for example, in the *measure of central tendencies—mean, median, mode, and standard deviations.*
- Examples include location in Cartesian coordinates and direction measured in degrees from magnetic north.
- The mean, median, and mode are allowed on interval data.

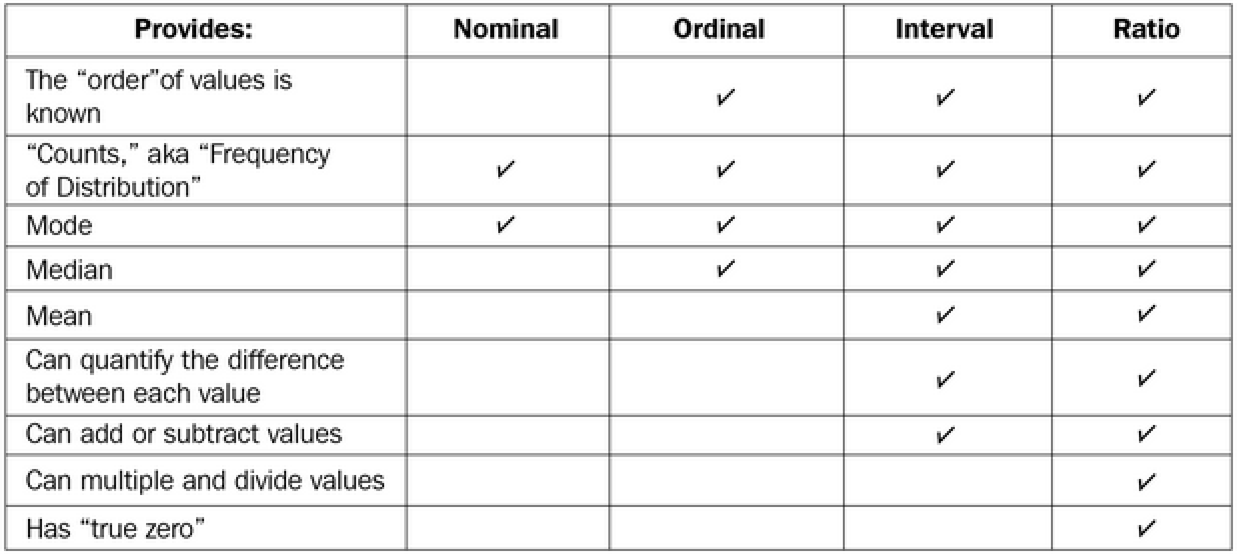
Ratio Measurements Scale

- Ratio scales contain order, exact values, and absolute zero, which makes it possible to be used in descriptive and inferential statistics.
- These scales provide numerous possibilities for statistical analysis.
- Mathematical operations, the measure of central tendencies, and the **measure of dispersion** and **coefficient of variatio**n can also be computed from such scales.

Ratio Measurements Scale

- Examples include a measure of energy, mass, length, duration, electrical energy, plan angle, and volume.

Scales/ Summary



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”				✓

Comparing EDA with classical and Bayesian analysis

- There are several approaches to data analysis.
- The most popular ones that are relevant to us are the following:

1. **Classical data analysis**
2. **Exploratory data analysis approach (EDA)***
3. **Bayesian data analysis approach**

Classical data analysis

- For the classical data analysis approach, the problem definition and data collection step are followed by model development, which is followed by analysis and result communication.

Exploratory data analysis approach

- For the EDA approach, it follows the same approach as classical data analysis except the model imposition and the data analysis steps are swapped.
- The main focus is on the data, its structure, outliers, models, and visualizations.
- Generally, in EDA, we do not impose any deterministic or probabilistic models on the data.

Bayesian data analysis approach

- The Bayesian approach incorporates prior probability distribution knowledge into the analysis steps as shown in the following diagram.
- Well, simply put, prior probability distribution of any quantity expresses the belief about that particular quantity before considering some evidence.

Software tools available for EDA

- **Python**: This is an open source programming language widely used in data analysis, data mining, and data science (https://www.python.org/). For this book, we will be using Python.
- **R programming language**: R is an open source programming language that is widely utilized in statistical computation and graphical data analysis (https://www.r-project.org).
- **Weka**: This is an open source data mining package that involves several EDA tools and algorithms (https://www.cs.waikato.ac.nz/ml/weka/).
- **KNIME**: This is an open source tool for data analysis and is based on Eclipse (https://www.knime.com/).

Python libraries for EDA

- NumPy
- Pandas
- Matplotlib
- SciPy

NumPy

- Create arrays with NumPy, copy arrays, and divide arrays
- Perform different operations on NumPy arrays
- Understand array selections, advanced indexing, and expanding
- Working with multi-dimensional arrays
- Linear algebraic functions and built-in NumPy functions

pandas

- Understand and create DataFrame objects
- Sub-setting data and indexing data
- Arithmetic functions, and mapping with pandas
- Managing index
- Building style for visual analysis