Data Science

Table of Contents

[Table of Contents 2](#_Toc1132406218)

[1 Week 1 3](#_Toc1411414474)

[1.1 What is data science and its applications? 3](#_Toc1751824913)

[1.2 Introduction to Python for data science 3](#_Toc1694271124)

[1.3 Basic data analysis using Python libraries like NumPy and Pandas 3](#_Toc24528309)

[1.4 Mini Project: Hands-on practice with Jupyter Notebooks 3](#_Toc1680220543)

[2 Week 2 3](#_Toc1452922069)

[2.1 Why visualization is important in data science: 3](#_Toc742051232)

[2.2 Introduction to Tableau for data visualization: 4](#_Toc677156904)

[2.3 Creating basic charts and graphs in Tableau: 4](#_Toc1343156802)

[2.4 Understanding advanced visualization techniques in Tableau: 4](#_Toc1962814668)

[2.5 Hands-on practice creating visualizations using real-world datasets: 4](#_Toc533971187)

[3 Week 3 4](#_Toc1814918610)

[3.1 Overview of statistical concepts like mean, median, mode, and standard deviation: 4](#_Toc1043468189)

[3.2 Probability distributions and hypothesis testing: 4](#_Toc1670202327)

[3.3 Correlation and regression analysis: 5](#_Toc2008378672)

[3.4 Analysis of variance (ANOVA) and Chi-Square tests: 5](#_Toc770263552)

[3.4.1 Mini Project: Hands-on practice implementing statistical analysis using Python libraries like Scipy and Statsmodels: 5](#_Toc1933991835)

[4 Week 4 5](#_Toc671818381)

[4.1 What is machine learning and its applications 5](#_Toc1024261049)

[4.2 Supervised and unsupervised learning 5](#_Toc1731361376)

[4.3 Types of machine learning algorithms: linear regression, logistic regression, decision trees, and clustering 6](#_Toc986500945)

[4.4 Introduction to scikit-learn library for machine learning 6](#_Toc1812214162)

[4.5 Hands-on practice implementing basic machine learning algorithms in Python 6](#_Toc1809338209)

[5 Week 5 6](#_Toc587880131)

[5.1 Overview of data wrangling and cleaning 6](#_Toc208180277)

[5.2 Cleaning and transforming data using Python libraries like Pandas 6](#_Toc1782992785)

[5.3 Handling missing values and outliers 7](#_Toc306842363)

[5.4 Merging and reshaping data 7](#_Toc1948961987)

[5.5 Hands-on practice cleaning and transforming real-world datasets 7](#_Toc318802822)

[6 Week 6 7](#_Toc1961016842)

[6.1 Creating interactive visualizations in Tableau 7](#_Toc1076537434)

[6.2 Mapping data in Tableau 7](#_Toc1652894141)

[6.3 Working with calculated fields and parameters in Tableau 7](#_Toc2078786924)

[6.4 Best practices for designing effective dashboards in Tableau 8](#_Toc14691148)

[6.5 Hands-on practice creating advanced visualizations using real-world datasets 8](#_Toc1585727828)

[7 Week 7 8](#_Toc1983059048)

[7.1 Deep learning and neural networks 8](#_Toc2047474694)

[7.2 Natural language processing (NLP) and text analytics 8](#_Toc1957623490)

[7.3 Hands-on practice implementing advanced machine learning techniques in Python 8](#_Toc1781953739)

# Week 1

## What is data science and its applications?

* Definition of data science and its role in extracting insights from data
* Overview of data science process including data collection, cleaning, analysis, and interpretation
* Examples of real-world applications of data science in various industries such as healthcare, finance, and marketing
* **Task: Setup environment**

## Introduction to Python for data science

* Overview of Python as a programming language and its popularity in the data science community
* Introduction to Python libraries commonly used in data science such as NumPy, Pandas, Matplotlib, and Scikit-learn
* Basic syntax and programming concepts in Python such as data types, control structures, and functions
* **Task: Basic problem-solving using Python**

## Basic data analysis using Python libraries like NumPy and Pandas

* Introduction to NumPy and its functions for mathematical operations on arrays
* Introduction to Pandas and its data structures such as Series and DataFrame
* Basic data manipulation using Pandas such as selecting, filtering, and grouping data
* **Task: Hands on practice using NumPy and Pandas**

## Mini Project: Hands-on practice with Jupyter Notebooks

* Introduction to Jupyter Notebooks and its benefits for interactive data analysis and visualization
* Creating and running Python code cells in a Jupyter Notebook
* Performing basic data analysis tasks using Python libraries within Jupyter Notebooks

# Week 2

## Why visualization is important in data science:

* Helps to identify patterns and trends that may not be apparent in raw data
* Simplifies complex data for easier interpretation
* Enables better communication and collaboration of data insights
* Allows for exploration of different data scenarios and what-if analyses
* Facilitates decision-making processes based on data insights

## Introduction to Tableau for data visualization:

* Overview of Tableau and its features
* Explanation of data connection and data source
* Familiarization with the Tableau interface
* **Task: Install Tableau and setup**

## Creating basic charts and graphs in Tableau:

* Understanding the different types of charts and when to use them
* How to create bar charts, line charts, scatter plots, and pie charts
* Customizing colors, labels, and titles
* Adding filters and sorting data

## Understanding advanced visualization techniques in Tableau:

* Overview of more complex charts, such as heat maps and box plots
* How to use calculations and aggregations in visualization
* Adding reference lines and annotations to visualization
* Incorporating maps and geographical data into visualization

## Hands-on practice creating visualizations using real-world datasets:

* Using sample datasets to create basic and advanced visualizations
* Creating a dashboard with multiple visualizations
* Applying design principles to enhance visualization aesthetics
* Interpreting and communicating insights from the created visualizations.

# Week 3

## Overview of statistical concepts like mean, median, mode, and standard deviation:

* Definition and calculation of each concept
* Use cases and examples in real-world scenarios
* Differences and similarities between these concepts
* Interpretation of results and their significance
* **Task: Hands-on using Python and Pandas**

## Probability distributions and hypothesis testing:

* Types of probability distributions such as normal, binomial, and Poisson
* Hypothesis testing and its importance in statistical analysis
* Steps involved in hypothesis testing, including setting up the null and alternative hypotheses, selecting a significance level, calculating the test statistic, and making conclusions
* Confidence intervals and their relationship to hypothesis testing
* **Task: Hands on using Python and Pandas**

## Correlation and regression analysis:

* Definition and calculation of correlation coefficients such as Pearson and Spearman
* Interpretation of correlation coefficients and their significance
* Types of regression analysis including linear and logistic regression
* Use cases and examples of regression analysis
* Analysis of residuals and model fit

## Analysis of variance (ANOVA) and Chi-Square tests:

* Definition and calculation of ANOVA and Chi-Square tests
* Differences between one-way and two-way ANOVA
* Use cases and examples of ANOVA and Chi-Square tests
* Interpretation of results and their significance
* **Task: Hands on using Python and Pandas**

### Mini Project: Hands-on practice implementing statistical analysis using Python libraries like Scipy and Statsmodels:

* Introduction to Python libraries such as Scipy and Statsmodels
* Installation and setup of the required libraries
* Importing and loading data
* Performing statistical analysis using Python libraries
* Visualization of results using Matplotlib or other visualization libraries

# Week 4

## What is machine learning and its applications

* Machine learning is a subfield of artificial intelligence that involves training computer algorithms to make predictions or decisions based on data.
* Applications of machine learning include image recognition, natural language processing, recommendation systems, fraud detection, and many others.

## Supervised and unsupervised learning

* Supervised learning involves training an algorithm on labeled data, where the input features are mapped to known output values.
* Unsupervised learning involves training an algorithm on unlabeled data, where the algorithm must find patterns or structure in the data without explicit guidance.

## Types of machine learning algorithms: linear regression, logistic regression, decision trees, and clustering

* Linear regression is a supervised learning algorithm used for predicting continuous numeric values.
* Logistic regression is a supervised learning algorithm used for predicting binary or categorical outcomes.
* Decision trees are a supervised learning algorithm used for making decisions by recursively partitioning the input space into smaller regions based on the input features.
* Clustering is an unsupervised learning algorithm used for finding groups or clusters in the data based on similarity between data points.

## Introduction to scikit-learn library for machine learning

* Scikit-learn is a popular open-source library for machine learning in Python that provides a wide range of tools for data preprocessing, model selection, and model evaluation.
* Scikit-learn includes implementations of many common machine learning algorithms, as well as tools for feature selection, dimensionality reduction, and model interpretation.

## Hands-on practice implementing basic machine learning algorithms in Python

* Python is a popular programming language for machine learning due to its ease of use and the availability of many powerful libraries, including scikit-learn.
* Hands-on practice with machine learning involves working through examples of implementing basic algorithms, such as linear regression or k-means clustering, on sample datasets in Python.

# Week 5

## Overview of data wrangling and cleaning

* Understanding the importance of data cleaning and wrangling in data analysis
* Identifying common data quality issues, such as missing values, duplicates, and inconsistent formatting
* Developing a data cleaning plan and documenting the steps taken

## Cleaning and transforming data using Python libraries like Pandas

* Loading data into Pandas data frames
* Filtering and selecting data based on specific criteria
* Modifying data using functions like map(), apply(), and replace()
* Creating new columns or transforming existing ones based on calculations and operations

## Handling missing values and outliers

* Identifying missing values and outliers using Pandas functions like isna(), isnull(), and describe()
* Dealing with missing values by dropping rows/columns, filling in missing values using interpolation, or imputing values based on other data points
* Identifying outliers using techniques like z-score analysis, box plots, and scatter plots
* Handling outliers by removing them, transforming the data using logarithmic or other functions, or treating them as separate categories

## Merging and reshaping data

* Combining data from multiple sources using Pandas functions like merge() and concat()
* Reshaping data using functions like pivot(), melt(), and stack()/unstack()
* Aggregating and summarizing data using functions like groupby(), pivot\_table(), and crosstab()

## Hands-on practice cleaning and transforming real-world datasets

* Importing and cleaning real-world datasets, such as CSV files or Excel spreadsheets
* Applying the techniques learned in class to address common data quality issues
* Transforming and manipulating data to prepare it for analysis
* Communicating the results of the data cleaning and wrangling process to stakeholders.

# Week 6

## Creating interactive visualizations in Tableau

* Understanding the principles of interactivity in Tableau
* Adding and editing interactive elements such as filters, actions, and tooltips
* Customizing interactivity options to enhance the user experience
* Best practices for designing effective interactive visualizations

## Mapping data in Tableau

* Connecting to geographic data sources and creating maps
* Customizing maps with layers, labels, and tooltips
* Using mapping functions to analyze data geographically
* Best practices for designing effective maps in Tableau

## Working with calculated fields and parameters in Tableau

* Understanding the role of calculated fields and parameters in Tableau
* Creating and editing calculated fields and parameters
* Using calculations to analyze data in new ways
* Best practices for using calculated fields and parameters in Tableau

## Best practices for designing effective dashboards in Tableau

* Understanding the principles of effective dashboard design
* Choosing the right visualizations and layout for your dashboard
* Enhancing usability with interactivity and navigation options
* Best practices for designing effective and visually appealing dashboards in Tableau

## Hands-on practice creating advanced visualizations using real-world datasets

* Working with a variety of real-world datasets to create advanced visualizations
* Applying advanced Tableau techniques such as blending data sources and using Level of Detail calculations
* Best practices for creating effective and insightful visualizations in Tableau

# Week 7

## Deep learning and neural networks

* Types of neural networks: Convolutional neural networks (CNNs), Recurrent neural networks (RNNs), and Deep belief networks (DBNs).
* Activation functions: Sigmoid, ReLU, and Tanh.
* Backpropagation algorithm: The process of adjusting weights and biases in the network to minimize the error.
* Popular deep learning frameworks: TensorFlow, Keras, and PyTorch.

## Natural language processing (NLP) and text analytics

* Text preprocessing techniques: Tokenization, stemming, and stop word removal.
* NLP tasks: Sentiment analysis, named entity recognition (NER), and topic modeling.
* Machine learning models for NLP: Naive Bayes, support vector machines (SVMs), and recurrent neural networks (RNNs).
* Pretrained language models: BERT, GPT-3, and ULMFiT.

## Hands-on practice implementing advanced machine learning techniques in Python

* Exploratory data analysis (EDA): Understanding the data by visualizing and analyzing it.
* Feature engineering: Transforming raw data into features that the machine learning model can use.
* Model selection: Choosing the appropriate algorithm for the task.
* Hyperparameter tuning: Fine-tuning the model's parameters to improve its performance.
* Model evaluation: Measuring the model's accuracy and performance using metrics like precision, recall, and F1 score.