PART I: Descriptive statistics and probability for data analysis

DESCRIPTION:

In this section, at first, we will cover the ***descriptive statistics*** including central tendency (e.g. mean, median) and measures of spread (dispersion) (standard deviation, variance, range, etc.). We will introduce the most important Python libraries used for computing and visualizing descriptive statistics – `NumPy` (as part of `SciPy`), `Matplotlib`, and `Seaborn`. You will learn, ***through exploring a real-life dataset***, how to create various exploratory charts and tables. We will also show how the popular Python library `Pandas` have built-in methods to perform basic statistical analysis.

In the second part, we will cover concepts of probability and distributions which lie at the heart of ***inferential statistics***, the main engine behind majority of data science tasks. You will learn how to simulate probabilistic events using simple Python programming. Next, we will cover in detail, various discrete and continuous probability distributions – Bernoulli, Binomial, Normal (Gaussian), Uniform, Poisson, etc. and how to generate such data using Python.

Almost all data science problems have a phase of ***exploratory data analysis*** which significantly aids the later modeling or machine learning phase. We will demonstrate, through an analytics exercise, how the above-mentioned statistical techniques can be used to prepare a stage for advanced machine learning.

1. Why statistics is foundation of data science
2. Central tendency and dispersion measures
3. Bivariate statistics, scatterplot, and correlation coefficient
4. The concept of probability
5. Discrete and continuous probability distributions
6. Bayes' rule and how it is used in data science
7. Exploratory data analysis (EDA) and how it powers data science

PART II: Inferential and Bayesian statistics for data science

DESCRIPTION:

One of the fundamental problem of data science is to reach conclusions that extend beyond the immediate data alone and this is accomplished with the techniques of inferential statistics. In this section, we equip you with the necessary theoretical and programmatic concepts and tools to perform that task. You will learn the concept of p-values and how it can be used in the context of hypothesis testing. Next, we cover some popular statistical tests such as t-test, ANOVA, chi-square test, normality test, etc. and show how they can be performed using Python libraries.

Next, we review the Bayes' rule, as an advanced concept of computing probability, and expand on how it can be used for statistical inference. You will learn, through real-life problem-solving example, why Bayesian inference can be particularly suitable for modern data science tasks where we have to continually update our models with new data.

1. What is estimation in statistics
2. Concept of p-values
3. t-test, ANOVA, Chi-square test
4. Bayes' rule and how to use it for probability computation
5. Application example of Bayesian inference using Python

PART III: Statistical methods as used in practical data science and ML

DESCRIPTION:

In this section, we will show practical examples of using statistical modeling techniques in solving data science problems and machine learning tasks. Specifically, we will cover – simple and multi-variate linear regression, logistic regression, Naïve Bayes classification, and Maximum Likelihood Estimation.

You will learn the probabilistic aspects of the solution of a ***linear regression*** problem and how to compute confidence levels in your prediction. We will illustrate how to use various Python libraries and methods, including the most popular machine learning package `Scikit-learn`, to solve a linear regression problem with a real-life dataset.

Logistic regression is a mixed machine learning technique in the sense that it employs the mechanics of regression but is primarily used for classification problems where probabilistic outputs ae desired. We will show the technique and application of logistic regression for a classification problem with US income data.

Next, we will analyze the concept behind the ***Naïve Bayes*** and show the key assumption which is needed to make the computation tractable and gives it the name '*Naïve*'. You will learn practical application of the algorithm for spam email filtering and perform a hands-on programming exercise with Scikit-learn.

Maximum likelihood estimation (MLE) is a general-purpose probabilistic technique used in unsupervised machine learning such as clustering. We will show example of ***k-means clustering*** which is a special case of MLE and how it can be used in data analytics task of product/market segmentation.

1. Linear regression with practical example
2. Linear regression as a statistical inference problem, advanced linear regression topics
3. Logistic regression as a classification algorithm, case study with the US income data
4. Naïve Bayes concept and practical application – spam filtering
5. MLE and k-means clustering using market segmentation example