Exploratory Data Analysis

Course Description:

Exploratory data analysis is one of the crucial steps in the data analysis process. It's the first thing that you'd do before you start with modelling your data. It provides all the necessary context to develop a deeper understanding of the data you're dealing with and help create an appropriate model and interpret the results correctly.

In a typical project life-cycle, upwards of 50% of the time is spent on procuring, cleaning, and exploring the data.

Learning Outcomes:

After the end of this course, the participants will be able to:

- Understand the importance of EDA
- Work with Pandas library, create and manipulate data frames
- Understand various functions in the Pandas library
- Gain an overview of Seaborne package
- Gain an overview of matplotlib package

Pedagogy:

The course is a mixture of classroom lectures, in-class lab exercises, quizzes, take-home exercises and mini-projects. Jupyter notebook will be the medium of teaching python.

Course Content with day wise breakup (16 hours of class-room + 8 hours of lab exercises):

Starter Kits

- Introduction to Statistics
- Population vs Sample
- Univariate and Multivariate statistics
- Types of variables – Categorical and Continuous
- Measures of central tendency - Mean, median, mode
- Measures of dispersion – Quartiles, percentiles, Standard deviation, variance, coefficient of variation

DAY - 1

1. What is statistics , sampling, population and how EDA will take part in it
2. Types of data - cross sectional data , time series data and panel data
3. Variables and types of variables
4. Explain in detail about Categorical and Continuous
5 . detailed notes about ordinal, interval, nominal and ratio with example
6. Different types of source data and how to read it with examples - dat,txt, excel, csv,tsv, URL files
7. Central tendency - Mean , median, mode . How to calculate each. Pro's and con's of each
8. measure of dispersion - variance, std deviation, range, quartiles and percentiles

read a data, plot a box plot ,explain about IQR - highlight the formula

identify outliers

9. Shape of data - Skewness(right and left skew - where the mean and median will be), kurtosis
10 Important concepts -
Correlation :
brief about correlation
correlation matrix
corr plot - pair plot, scatter plot, heat map

covariance:

difference between correlation and covariance

DAY - 2

1. Brief about univariate – and example for univariate. This intro should be given before the dataset

2. Descriptive stats with EDA – please find my python notes attached

3. Introduction to distribution – normal distribution. Definition, math and conditions for normal distribution

4. Central limit theorem.

5. Univariate analysis (PDF, CDF, Boxplot, Violin plots, Distribution plots) – all these charts have to be covered for univariate
6. Multivariate analysis (pair plot, corrplot, heatmap, multivariate scatter plot, grouped box plot)
7. Data transformation – shared reference link

Detail out more on

i. how to handle outliers
ii. How to handle missing values – data imputation
iii. Data imbalance – brief about over sampling and under sampling. We can take only on the concepts more in details
iv. How to check if the given data falls under normal distribution, if not how to change it.
v. Scaling and normalization – theory part is missing. What is the formula behind it how the data gets converted into normalized or standardized?
vi. Other types of data transformation available

DAY - 3

Bivariate

- Feature to feature relationships
- Correlation and Frequency tables
- Seasonality and looking at trended data
- Multi variate analysis

DAY - 4

Wrangling

- Various ways of treating missing value's / Missing value Treatment?

- Various ways of outlier treatment?
- Data Imbalance treatment
- Feature engineering, Introduction to Test and Train