**Prerequisites**

* The learners should have a basic understanding of Python.
* They should know about libraries that are used in data science such as NumPy, Pandas, Matplotlib, and Seaborn.
* They should also be aware of basic data manipulation techniques like loading, indexing and slicing.
* They should have knowledge about basic data visualisation techniques such as bar graphs, histograms, and box plots.
* They should know how to perform basic data preprocessing steps like data sourcing and data cleaning.
* Finally, they should know how to perform univariate and bivariate analysis.

**Session Overview**

In this 180-min-long live session, learners will participate in an interactive session on data preprocessing and exploratory data analysis. The session will cover various exploratory data analysis techniques, differentiating between univariate and bivariate analysis, as well as providing an application of data preprocessing techniques. The instructor will guide learners through practical problem-solving, focusing on reinforcing concepts of preliminary study of the data for initial findings.

Learners are encouraged to code along with the SME for better understanding. The dataset to be used for this session can be found [here](https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists) (Data Scientist Jobs). The final segment of the session will be reserved for clarifying doubts and reinforcing best practices.

**NOTE 1:** The SME is free to use any other relevant dataset of their choice. Ensure the same is shared with the learners at the start of the session. Other suggested datasets: [Used Cars](https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data), [House Prices](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset)

**NOTE 2:** While creating plots, the SME and the learners are encouraged to use both Matplotlib and Seaborn libraries.

**Learning Outcomes**

By the end of this session, learners will be able to

* Perform data preprocessing based on the business objective and the raw data received to prepare it for analysis
* Study different types of variables, whether numerical or categorical, and draw preliminary conclusions from the same
* Study numerical variables by grouping them based on a categorical variable for comparison between categories
* Examine the relationships between two or more variables (numerical vs numerical, categorical vs numerical, categorical vs categorical), and draw inferences from the same

**Brief Agenda**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Section** | **Title** | **Comments** | **Duration (mins)** | **Mode** |
| **I** | **Introduction** | Discuss the session's objectives and the flow of the session | 5 mins | - |
| **II** | **Data Preprocessing and Univariate Analysis** | You have been provided with the raw dataset.  Tasks to be done:  - Dataset loading and description  - Analysis of different columns and variables  - Handling null and missing values (if any)  - Format/standardise data (if required)  - Filter/delete rows and columns (if necessary)  - Exploration of categorical variables for univariate analysis  - Exploration of numerical variables for univariate analysis  Analysis involves creating suitable plots and drawing inferences from the same. | 80 mins | Demonstration |
| **III** | **Bivariate and Multivariate Analysis** | Use the cleaned dataset from the previous section.  Tasks to be done:  - Performing bivariate analysis using two variables (both categorical and numerical)  - Performing multivariate analysis using more than two variables (including correlation maps)  - Derive any new metrics (if feasible)  Analysis involves creating suitable plots and drawing inferences from the same.  Discuss the strategic business decisions that can be taken based on the analysis. | 80 mins | Demonstration |
| **IV** | **Q&A / Closing** | Final Q&A, troubleshoot issues  Open the floor for learner’s queries and address any doubts  Recap key takeaways on the importance and practical application of data preprocessing and EDA | 15 mins | - |

**Conclusion**

By the end of this session, learners will be more aware of the importance of data preprocessing and exploratory data analysis on various types of variables (such as numerical, ordered categorical, and unordered categorical). They will draw inferences from the initial findings of the data and ensure these inferences are relevant in making business decisions.