FUTURE VISION™
COMPUTER INSTITUTE

PROBLEM STATEMENT

Income Exploratory Data Analysis

PROBLEM STATEMENT

The goal of this analysis is to explore the relationships between various demographic, social, and employment-related attributes of individuals and their income levels.

The dataset contains information about individuals, such as their age, workclass, education, marital status, occupation, race, gender, hours worked per week, and native country.

The primary objective is to understand patterns in income distribution, identify factors that influence income, and predict whether an individual earns above or below 50K annually.

The analysis aims to identify key predictors for income and provide insights on the demographic, social, and professional factors that influence an individual's income level.

BUSINESS PROBLEM OVERVIEW

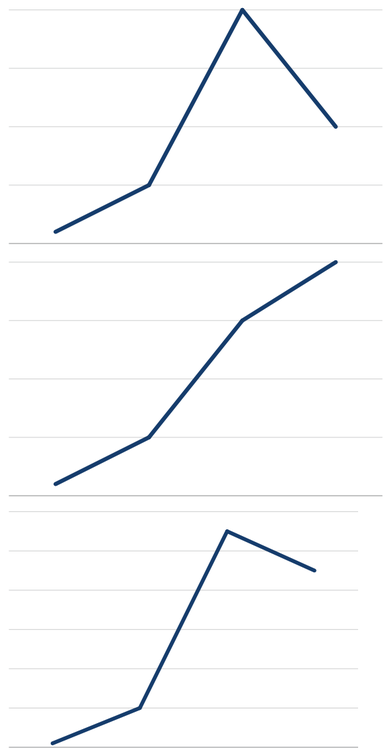
In today's competitive job market, understanding the factors that influence an individual's income can be invaluable for businesses, policymakers, and job seekers alike. For businesses, understanding income distribution helps with talent acquisition, wage benchmarking, and employee retention strategies.

Policymakers can use insights to develop targeted economic programs and initiatives. Additionally, individuals looking to improve their career prospects can benefit from knowing which factors (e.g., education, experience, work class) have the greatest impact on earnings.

The dataset provided includes detailed information about individuals' demographic and employment characteristics, such as age, education, occupation, work class, gender, and hours worked per week, as well as whether they earn more or less than $50K annually.

By analyzing this data, businesses and organizations can identify trends, make data-driven decisions, and improve strategies around hiring, salary offerings, and workforce development.

By leveraging this data, businesses can optimize workforce management, improve employee satisfaction, and contribute to broader societal goals like reducing wage inequality.



In this context the "Income" is the target variable that indicates whether an individual earns more or less than 50,000 annually. This is typically used to classify individuals into two income groups: those earning above 50K and those earning $50K or less.

The income variable is crucial in determining economic outcomes and is used to understand the relationship between various demographic, educational, and professional characteristics and income levels. The Income column serves as the target variable for predictive modeling. It is used to:

1. Classify individuals: Predict whether an individual's income falls above or below the $50K threshold based on features like age, education level, work class, occupation, gender, etc.
2. Analyze income disparities: Investigate how different factors (e.g., education, occupation, race, gender) affect the likelihood of earning a higher income.
3. Guide business and policy decisions: Understanding income levels helps organizations design compensation strategies, set salary benchmarks, and identify trends that could inform workforce development, hiring practices, or policy-making around income equality.

UNDERSTANDING & DEFINING DATASET

PROJECT PIPELINE

The project pipeline can be briefly summarized in the following steps:

- Data Understanding: Here, we need to load the data and understand the features present in it. This would help us choose the features that we will need for your final model.

- Exploratory data analytics (EDA): Normally, in this step, we need to perform univariate and bivariate analyses of the data, followed by feature transformations, if necessary. For the current data set, because Gaussian variables are used, we do not need to perform Z-scaling. However, you can check if there is any skewness in the data and try to mitigate it, as it might cause problems during the model-building phase.

FUTURE VISION™
COMPUTER INSTITUTE

THANK YOU