Problem Statement

Recruiting and retaining drivers is seen by industry watchers as a tough battle for Ola. Churn among drivers is high and it's very easy for drivers to stop working for the service on the fly or jump to Uber depending on the rates. As the companies get bigger, the high churn could become a bigger problem. To find new drivers, Ola is casting a wide net, including people who don't have cars for jobs. But this acquisition is really costly. Losing drivers frequently impacts the morale of the organization and acquiring new drivers is more expensive than retaining existing ones.

You are working as a data scientist with the Analytics Department of Ola, focused on driver team attrition. You are provided with the monthly information for a segment of drivers for 2019 and 2020 and tasked to predict whether a driver will be leaving the company or not based on their attributes like:

Data Dictionary Overview

- **MMMM-YY**: Reporting Date (Monthly)
- **Driver_ID**: Unique ID for drivers
- **Age**: Age of the driver
- **Gender**: Gender of the driver (Male: 0, Female: 1)
- **City**: City Code of the driver
- **Education_Level**: Education level (0 for 10+, 1 for 12+, 2 for graduate)
- **Income**: Monthly average Income of the driver
- **Date Of Joining**: Joining date for the driver
- **LastWorkingDate**: Last date of working for the driver
- **Joining Designation**: Designation of the driver at joining
- **Grade**: Grade of the driver at reporting
- **Total Business Value**: Total business value acquired (negative indicates cancellation/refund or car EMI adjustments)
- **Quarterly Rating**: Quarterly rating of the driver (1-5, higher is better)

Dataset:-
https://drive.google.com/file/d/10_njv9Lj2dgatw2j5KNBRc0aP6rIhHDp/view?usp=sharing

Analysis Guide

1. Define Problem Statement and Perform Exploratory Data Analysis (10 points)

- **Definition of Problem**:
    - o Understand the challenge of driver attrition and its impact on Ola.
- **Data Exploration**:
    - o Check data shape, data types, and convert categorical attributes if necessary.
    - o Detect missing values and prepare for simple imputation.

o Perform statistical summary to understand data distribution.

2. Data Preprocessing (50 Points)

- **Simple Imputation**:
  - o Handle missing values using mean, median, or mode for numerical features.
- **Feature Engineering**:
  - o Create a target variable indicating whether a driver has left the company based on `LastWorkingDate.`
  - o Generate additional features:
    - Calculate age of each driver based on `Date Of Joining.`
    - Determine if quarterly rating has increased (1 if yes, 0 if no).
    - Identify if monthly income has increased (1 if yes, 0 if no).
- **Class Imbalance Treatment**:
  - o Check for class imbalance in the target variable.
  - o Address imbalance using techniques like oversampling, undersampling, or synthetic data generation if necessary.
- **Standardization**:
  - o Standardize numerical features to ensure they are on the same scale.
- **Encoding**:
  - o Perform one-hot encoding for categorical variables like `City`, `Education_Level`, and `Joining Designation.`

4. Actionable Insights & Recommendations (10 Points)

- **Provide actionable insights based on the analysis**:
  - o Identify key factors influencing driver attrition.
  - o Recommend strategies to improve driver retention.

Question Distribution

Basic Level Questions (10 points)

1. **Data Structure and Overview**:
   - o What is the structure of the dataset (number of rows and columns)?
   - o What are the data types of each column?
   - o Are there any missing values in the dataset? If so, which columns are affected?
2. **Descriptive Statistics**:
   - o What are the basic statistics (mean, median, standard deviation) for numerical features like `Age`, `Income`, `Total Business Value`, and `Quarterly Rating`?
   - o How many unique drivers are there in the dataset?
3. **Temporal Analysis**:

o How many drivers joined and left each month?
　　o Can we determine the average tenure of drivers in the dataset?

Intermediate Level Questions (30 points)

4. **Feature Engineering**:
　　o How can we create a target variable to indicate whether a driver has left the company based on `LastWorkingDate`?
　　o What additional features can we extract from `Dateofjoining`, such as tenure or duration of employment?
5. **Exploratory Data Analysis (EDA)**:
　　o What are the distributions of `Age`, `Income`, and `Total Business Value`?
　　o How does `Quarterly Rating` vary across different drivers and time periods?
　　o Are there any trends or patterns in the monthly income or business value acquired?
6. **Missing Values Handling**:
　　o How should missing values in `LastWorkingDate` be treated, considering it indicates whether a driver has left?

Advanced Level Questions (60 points)

7. **Correlation and Relationships**:
　　o Is there a correlation between `Age` and `Income`?
　　o How do `Education_Level` and `City` affect `Total Business Value`?
　　o Are drivers with higher `Quarterly Rating` more likely to stay longer?
8. **Predictive Analysis(Optional)** :
　　o Can we predict which drivers are likely to leave based on their demographic and performance attributes?
　　o What machine learning techniques could be applied to predict driver attrition?
9. **Recommendations**:
　　o Based on the analysis, what strategies can Ola implement to improve driver retention?
　　o Are there specific demographic groups or performance metrics that require targeted interventions?

Actionable Insights & Recommendations (10 Points)

● Provide actionable insights based on the analysis:
　　o Identify key factors influencing driver attrition.
　　o Recommend strategies to improve driver retention.

Submission Requirements

- Convert Jupyter notebook to PDF format and upload to a public Google Drive.
- Include insights, recommendations, and relevant images/graphs in the document.