**Problem Statement**

Our client Ditto, an Insurance company, has provided Health Insurance to its customers and now needs help in analyzing whether the policyholders (customers) from the past year will also be interested in Vehicle Insurance provided by the company. An accurate data analysis will enable the company to tailor its communication strategy effectively, optimize its business model, and increase revenue.

**Dataset  
[https://drive.google.com/file/d/1RcLIB-vrZmJhJSUI93FAMlZm6qKiL6yT/view?usp=sharing](Ditto Insurance Case Study.docx)**

The dataset includes information on:

* **Demographics**: gender, age, region code type
* **Vehicles**: vehicle age, damage
* **Policy**: premium amount, sourcing channel

You can download the dataset from this.

**Concepts Tested**

* Exploratory Data Analysis (EDA)
* Imputation of Missing Values
* Working with an imbalanced dataset

**What "Good" Looks Like**

**1. Exploratory Data Analysis (EDA) (30 points)**

Examine the dataset's structure, data types, missing values, and statistical summary.

* **Problem Definition**: Clearly define the problem.
* **Observations**:
  + Dataset shape
  + Data types
  + Missing values
  + Statistical summary
* **Univariate Analysis**:
  + Distribution plots for continuous variables
  + Bar plots/count plots for categorical variables
* **Bivariate Analysis**:
  + Relationships between important variables
* **Insights**:
  + Provide insights based on EDA.

**2. Data Preprocessing (40 Points)**

Prepare the data for analysis.

* **Missing Values Imputation**:
  + Check for missing values and prepare data for Imputation.
* **Feature Engineering**:
  + Create new features like changes in premium, age groups, etc.
* **Class Imbalance Treatment**:
  + Use techniques like SMOTE to balance the dataset.
* **Standardization**:
  + Standardize numerical features.
* **Encoding**:
  + One-hot encode categorical variables.

**Evaluation Criteria**

**1. Define Problem Statement and Perform EDA (40 points)**

* Definition of the problem
* Observations on dataset shape, data types, missing values, statistical summary
* Univariate Analysis: Distribution plots for continuous variables, bar plots/count plots for categorical variables
* Bivariate Analysis: Relationships between important variables
* Insights based on EDA

**2. Data Preprocessing (60 Points)**

* Missing Values Imputation
* Feature Engineering
* Class Imbalance Treatment
* Standardization
* Encoding

**Questions to Answer During Data Analysis**

**Dataset Structure:**

1. What is the shape of the dataset?
2. What are the data types of each column?
3. Are there any missing values in the dataset? If so, how many and in which columns?

**Statistical Summary:**

1. What are the basic statistical summaries (mean, median, standard deviation) for the numerical features?
2. What is the distribution of categorical variables?

**Univariate Analysis:**

1. What are the distribution plots for continuous variables?
2. What are the bar plots/count plots for categorical variables?

**Bivariate Analysis:**

1. How do the important variables relate to each other?
2. Are there any interesting patterns or correlations between variables?

**Missing Values:**

1. How will Imputation be applied to handle missing values?

**Feature Engineering:**

1. What new features can be created from the existing data?
2. How do these new features add value to the analysis?

**Class Imbalance:**

1. How is class imbalance in the target variable addressed?
2. What techniques, such as SMOTE, can be applied to balance the dataset?

**Standardization and Encoding:**

1. How are numerical features standardized?
2. How are categorical variables one-hot encoded?

**Insights to Investigate**

**Demographic Distribution:**

1. Gender distribution across policyholders.
2. Age distribution of policyholders.
3. Distribution of policyholders across different regions (Region\_Code).

**Vehicle Information:**

1. Distribution of vehicle age among policyholders.
2. Number of policyholders with prior vehicle damage.
3. Correlation between vehicle age and vehicle damage.

**Insurance Policy Details:**

1. Distribution of Annual Premiums.
2. Popularity of different policy sales channels.
3. Average vintage (number of days since the customer became a policyholder).

**Driving License and Previously Insured:**

1. Number of policyholders with a driving license.
2. Number of policyholders who were previously insured.

**Relationships and Patterns:**

1. Relationship between age and annual premium.
2. Relationship between vehicle age and annual premium.
3. Correlation between region code and annual premium.
4. Average annual premium for previously insured vs. not previously insured policyholders.
5. Distribution of policy sales channels among male and female policyholders.

**Advanced Questions**

**Advanced Correlation Analysis:**

1. What are the Pearson and Spearman correlation coefficients between numerical features (e.g., Age, Annual Premium, Vintage)? What do these correlations suggest?
2. Are there any non-linear relationships between the variables that might be missed by standard correlation measures?

**Multivariate Analysis:**

1. How do multiple factors (e.g., Age, Gender, Region\_Code, Vehicle\_Age) together influence the annual premium?
2. Use pair plots or scatter plot matrices to visualize relationships between multiple variables.

**Interaction Effects:**

1. Are there any significant interaction effects between categorical variables on the annual premium (e.g., Gender \* Vehicle\_Age, Region\_Code \* Previously\_Insured)?
2. How do interactions between continuous and categorical variables affect the target outcome?

**Geospatial Analysis:**

1. How does the interest in vehicle insurance vary across different regions? Are there any geographical patterns or clusters?
2. Use geospatial visualization tools to map the distribution of policyholders and their interest in vehicle insurance.

**Advanced Feature Engineering:**

1. Can we create features based on domain knowledge, such as loyalty scores, risk scores, or customer value indices?
2. How do these engineered features enhance the predictive power or insights of the analysis?

**Deliverables**

* **EDA Report**: A comprehensive report of the exploratory data analysis.
* **Data Preprocessing Report**: Documentation of the preprocessing steps, including any imputation, feature engineering, class imbalance treatment, standardization, and encoding.
* **Insights and Recommendations**: Provide actionable insights and recommendations based on the analysis.