Case Study 1: Predicting Insurance Claim Severity

Scenario

An insurance company is aiming to improve its underwriting process by better predicting the severity of claims. Historical data include claim amounts, policyholder demographics (e.g., age, gender, location), vehicle details, and previous claim history over the past five years. The goal is to develop a predictive model that can estimate claim costs accurately to support pricing decisions and risk management.

Tasks for Students

1. **Data Exploration & Preprocessing:**
    - Identify the key variables in the dataset.
    - Discuss potential data quality issues (e.g., missing values, outliers) and propose methods for cleaning and preprocessing the data.
    - Explain any transformations (such as log transformations) that might be necessary to address skewness or heteroscedasticity.
2. **Exploratory Data Analysis (EDA):**
    - Develop visualizations (e.g., histograms, scatter plots, box plots) to illustrate the distribution of claim severity and its relationship with major predictors.
    - Describe how you would identify correlations between predictors and claim amounts.
3. **Model Building:**
    - Propose a modeling strategy (e.g., multiple linear regression, generalized linear models) for predicting claim severity.
    - Specify the null hypotheses for key predictors and discuss how you would test them using appropriate statistical tests.
    - Outline how you would evaluate model performance (mention metrics like R-squared, MAE, etc.).
4. **Interpretation & Recommendations:**
    - Explain how the results of your model can inform pricing strategy and risk management.
    - Discuss potential limitations of your approach and suggest further improvements.

Case Study 2: Data Preparation and Exploratory Analysis of Insurance Customer Data

Scenario

An insurance company has collected a raw dataset containing details about its policyholders. The dataset includes the following columns:

- **Customer_ID:** Unique identifier for each customer.
- **Age:** Age of the policyholder.
- **Gender:** Categorical variable (e.g., Male, Female, Other).
- **Policy_Type:** Type of insurance policy (e.g., Auto, Home, Life).
- **Premium:** Monthly premium amount.
- **Claim_Count:** Number of claims filed in the past year.
- **Region:** Geographic region (e.g., North, South, East, West).
- **Date_Joined:** Date the policyholder joined the insurance plan.
- **Customer_Satisfaction:** Survey rating on a scale of 1–10.

The dataset may contain issues such as missing values, duplicate records, inconsistent categorical entries (e.g., "Male" vs. "M"), and potential outliers in numeric fields.

Tasks for Students

1. Data Quality Assessment & Cleaning

- **Identify Data Issues:**
    - List the potential data quality challenges you might encounter (e.g., missing values, duplicate records, inconsistent formats in categorical data, outlier values).
    - Explain why addressing these issues is critical before proceeding with any analysis.
- **Data Cleaning Procedures:**
    - Propose methods to handle missing values (e.g., removal or simple imputation with a constant or average value).
    - Describe how you would standardize inconsistent categorical entries (e.g., ensuring "Male," "M," and similar variants are unified).
    - Explain your approach to detecting and handling duplicate records and outliers.

2. Data Transformation & Preparation

- **Processing Date Fields:**
    - Outline how you would work with the Date_Joined column to create additional features such as the year, month, or the tenure (the duration since joining).
- **Standardizing and Enriching Data:**
    - Describe the steps you would take to ensure all categorical variables (like Gender and Policy_Type) are formatted consistently.
    - Suggest any simple derived features (such as categorizing customers into age groups or tenure brackets) that could enrich the analysis without invoking complex statistical methods.
- **Data Aggregation:**
    - Propose ways to summarize the data, such as calculating the average premium per region or the average claim count per policy type.
    - Explain how you would structure the dataset to facilitate an easy overview of key performance indicators for the company.

3. Exploratory Data Analysis (EDA)

- **Descriptive Summaries:**
    - List which basic descriptive information (e.g., counts, ranges, and simple averages) you would extract from each column.
    - Explain the importance of these summaries in understanding the overall dataset.
- **Insight Identification:**
    - Discuss how these visualizations and summaries could help identify key patterns or issues in the dataset (e.g., spotting regions with unusually high claim counts or low customer satisfaction).

4. Reporting & Communication

- **Document Your Process:**
    - Provide a clear, step-by-step report outlining your data cleaning, transformation, and analysis steps.
    - Use diagrams or flowcharts if necessary to illustrate your process.
- **Interpreting Findings:**
    - Summarize the main insights you discovered from your cleaned and transformed dataset.

    - Suggest actionable recommendations for the insurance company based on your findings (e.g., targeting specific customer segments for policy renewals or focusing on regions with higher premiums).

- **Presentation for a Non-Technical Audience:**
    - Outline how you would present your findings using clear visualizations and simple language.
    - Emphasize the practical implications of your insights without relying on technical statistical jargon.

Instructions for Assessment

- **Report Structure:**
  Each case study should be presented as a structured report with an introduction, methodology, analysis, and conclusions. Use clear headings and subheadings.
- **Data & Tools:**
  Students may assume access to a dataset matching the scenario description. They should mention any assumptions made about data attributes and quality.
- **Supporting Material:**
  Include diagrams, charts, or pseudocode where relevant to support your explanations.
- **Critical Analysis:**
  Emphasis should be placed on the rationale behind methodological choices and the interpretation of results, rather than on coding details alone.
- **Presentation:**
  The final submission should be clear, concise, and well-organized, reflecting both technical proficiency and the ability to communicate complex ideas effectively.