DATA ANALYSIS ASSIGNMENT FOR BEAMCO TRAINEES

Project 1: House Pricing Data Collection and Analysis Activity (Woji and Environs, Port Harcourt)

Objective: Collect, clean, and analyze house pricing data from Woji and surrounding areas in Port Harcourt. Build a model to predict house prices based on key factors like number of rooms, size, location, and facilities.

Steps:

1) Data Collection:
   a) Visit real estate websites or local listings to collect house pricing data from Woji and nearby neighborhoods (e.g., Elelewo, Rumuokwurushi, etc.).
   b) Data should include features like house price, number of bedrooms, bathrooms, lot size, location, and house condition, age of house.

2) Data Preprocessing:
   a) Clean the dataset to handle missing values and outliers.
   b) Feature engineering (e.g., creating categories for house size, proximity to amenities, etc.).

3) Exploratory Data Analysis (EDA):
   a) Conduct EDA to understand trends and patterns in the data.
   b) Visualize relationships between house prices and features (e.g., number of rooms vs. price).

4) Modelling:
   a) Apply Multiple Linear Regression to predict house prices.
   b) Evaluate the model using metrics like R-squared and Mean Squared Error (MSE). 5) Conclusion & Reporting:
   a) Report findings, discussing which factors influence house pricing the most.
   b) Present visualizations and model performance

Project 2: Web Scraping Konga for Product Information by Categories and Prices

Objective: The goal of this project is to scrape product data from Konga's e-commerce platform. The scraped data should include product categories, current prices, discounted prices (if available), and formal (original) prices.

Steps:

1. Web Scraping Setup:
   a. Use Python libraries such as BeautifulSoup, Requests, or Scrapy to scrape product data from Konga.
   b. Identify product URLs for different categories (e.g., Electronics, Fashion, Groceries, etc.) on Konga.

2. Data to Scrape:
   a. Product name
   b. Category
   c. Current price
   d. Discounted price (if any)
   e. Formal price (original price before discount)

3. Data Preprocessing:
   a. Clean the scraped data by handling missing or incorrect values.
   b. Organize the data into a structured format (e.g., pandas DataFrame) with columns like:
      i. Product Name
      ii. Category
      iii. Price iv. Discounted Price
      iv. Formal Price

4. Exploratory Data Analysis (EDA):
   a. Analyze the average prices of products across different categories.
   b. Visualize the percentage of products with discounts across categories.
   c. Identify the range of discounts being offered (e.g., 10-20%, 20-30%).

5. Conclusion & Reporting:
   a. Provide insights on which categories have the most discounts.
   b. Offer recommendations based on price trends and discount rates for customers or businesses.

Project 3: Customer Churn Prediction with Bank Data and Dashboard Creation

Objective: Develop a predictive model to identify customers likely to churn using a dataset with two sheets: Account Info and Customer Info and create an interactive dashboard to visualize the results.

Project Description:

In this project, trainees are expected to work with a bank churn dataset containing two sheets that need to be merged. They will preprocess the data, perform exploratory data analysis (EDA), build a machine learning model to predict customer churn, and create a dashboard to visualize their findings.

1. Data Exploration and Preprocessing
   a. Merging Data:
      i. Combine the Account Info and Customer Info datasets using Customer ID.
      ii. Handle missing values and correct data errors.

   b. Feature Engineering:
      i. Create new features based on the existing data, such as tenure or total account balance.
      ii. Standardize numerical features if necessary.

   c. Train-Test Split:
      i. Split the dataset into training and testing sets.

2. Exploratory Data Analysis (EDA)
   a. Visualizing Data:
      i. Plot key features to visualize their relationships with churn (e.g., churn rate by age, balance, tenure).
      ii. Examine distributions and correlations.

   b. Analyzing Churn Patterns:
      i. Identify patterns or trends related to customer churn.

3. Machine Learning Model Development
   a. Model Building:
      i. Develop and evaluate two models:
         - **Logistic Regression:** Start with a basic logistic regression model.

- **Random Forest Classifier:** Build a random forest model to capture more complex patterns.

b. Model Evaluation:

i. Use accuracy to evaluate model performance. ii. Review the confusion matrix to assess how well the model is performing in predicting churn.

4. Model Interpretation

a. Feature Importance:

i. For the Random Forest model, check which features are most important in predicting churn.

b. Model Coefficients:

i. For the Logistic Regression model, examine the coefficients to understand the impact of each feature.

5. Dashboard Creation

a. Design the Dashboard:

i. Use a tool like Tableau, Power BI, Excel.

ii. Include visualizations such as:

- **Churn Prediction Summary:** Display overall churn rates, accuracy of the model, and key features influencing churn.
- **Feature Analysis:** Interactive plots showing the relationship between features and churn.
- **Model Performance:** Confusion matrix and accuracy metrics.

b. Interactive Elements:

i. Add filters to view churn predictions by different categories (e.g., age ranges, account balance). ii. Provide options to visualize data from both Logistic Regression and Random Forest models.

6. Model Comparison and Final Steps

a. Compare Models:

i. Compare the Logistic Regression and Random Forest models based on accuracy, confusion matrix results, and feature importance.

b. Final Report:

i. Write a report summarizing data preprocessing, EDA, model development, and evaluation. ii. Include visualizations, key findings from the model results, and insights from the dashboard.

c. Dashboard Submission:

i. Ensure the dashboard is interactive and provides meaningful insights.

ii. ii. Submit both the dashboard and the final report.

Mr. Chukwuemeka

Project 4: Patient Demographics and Encounter Trends Analysis with Dashboard

Objective: Analyze patient demographics and encounter trends to uncover patterns and insights and present these findings through an interactive dashboard.
Steps:

1. Data Collection:
- Collect data on patient demographics (age, gender, race, marital status) and encounter details (encounter class, reason code, payer coverage).
- Include additional fields such as Encounter Start Time, Encounter Stop Time, and Total Claim Cost.

2. Data Preprocessing:
- Clean the dataset to address missing values and ensure consistency.
- Transform data into a structured format suitable for analysis (e.g., pandas DataFrame).

3. Exploratory Data Analysis (EDA):
- Demographic Analysis: Analyze the distribution of demographics and their relationships with encounter types and costs.
- Trend Visualization: Create visualizations to show trends in encounter reasons and costs across different demographic groups.
- Pattern Identification: Identify significant patterns or anomalies in the encounter data.

4. Trend Analysis:
- Examine trends in encounter types and reasons over time.
- Assess the impact of payer coverage on encounter characteristics and costs.

5. Dashboard Creation:

Design the Dashboard:
- Use a tool like Tableau, Power BI, or Excel.
- Include interactive visualizations such as:
- Demographic Overview: Display demographics distribution and its correlation with encounter trends.
- Encounter Trends: Show trends in encounter types, reasons, and costs over time.
- Payer Coverage Analysis: Visualize the impact of different payer coverages on encounter characteristics.

- Interactive Elements:
- Add filters for demographics, encounter types, and time periods to enable detailed analysis.
- Provide options to drill down into specific categories or time frames.

6. Conclusion & Reporting:
   - Summarize key trends and patterns found in the data.
   - Present insights from the dashboard, including significant findings and trends.
   - Provide actionable recommendations based on the analysis.