**Projects/Assignments**

**Project: 1**

**Assignment for the Mini Project: Retail Sales Dataset**

Here is a structured assignment for the project. Each task is designed to test specific skills in data analytics.

**Objective:**

Analyze the retail dataset to uncover trends, insights, and correlations.

**Assignment Tasks:**

**1. Data Loading and Inspection**

1. Load the dataset into a Pandas DataFrame.
2. Display the first 10 rows of the dataset.
3. Print the column names and their respective data types.
4. Check for missing values and duplicates in the dataset. How would you handle them?

**Deliverables:**

* A brief summary of the dataset.
* A count of missing values and duplicate rows.

**2. Data Cleaning**

1. If there are duplicates, remove them.
2. Handle missing values appropriately (drop or impute them).
3. Rename the columns to more meaningful names, if necessary.

**Deliverables:**

* Cleaned dataset.
* Explanation of how you handled missing values and duplicates.

**3. Exploratory Data Analysis (EDA)**

1. Plot the distribution of the following discrete features:
   * Quantity
   * Category
   * Region
2. Identify the most frequently purchased product category.
3. Find the region contributing the most to sales.

**Deliverables:**

* Distribution plots.
* Insights on top product categories and regions.

**4. Timestamp Transformation**

1. Convert the Order Date column to a datetime format.
2. Create new columns:
   * Day of the week.
   * Day of the month.
   * Month.
   * Year.
   * Season (Winter, Spring, Summer, Fall).
3. Set the Order Date column as the DataFrame index.

**Deliverables:**

* Updated DataFrame with new columns.
* A few sample rows to verify the transformations.

**5. Correlation Analysis**

1. Compute the correlation matrix for numerical columns.
2. Visualize the correlation matrix using a heatmap.
3. Identify and discuss two key relationships from the heatmap.

**Deliverables:**

* Heatmap visualization.
* Observations about key correlations.

**6. DataFrame Operations**

1. Use loc to extract sales data for a specific date or range of dates.
2. Use iloc to extract the first 10 rows of the dataset.
3. Drop a column (e.g., Region) using drop(axis=1).
4. Drop the first two rows using drop(axis=0).

**Deliverables:**

* Code and results for each operation.
* Explanation of when to use loc vs iloc.

**7. Insights and Reporting**

1. Write a one-page summary of key insights from the analysis.
2. Highlight recommendations for improving sales based on your findings.

**Deliverables:**

* A well-documented report with visualizations and actionable insights.

**Submission Guidelines:**

1. Submit your code as a Jupyter Notebook or Python script.
2. Include all visualizations and clearly label them.
3. Attach your report as a separate document (PDF or Word).

**Grading Criteria:**

1. **Code Quality (20%)**: Clean, well-commented, and efficient code.
2. **Data Cleaning (20%)**: Proper handling of missing values and duplicates.
3. **EDA (20%)**: Insightful visualizations and observations.
4. **Feature Engineering (20%)**: Accurate timestamp transformations and new feature creation.
5. **Report Quality (20%)**: Clear, concise, and insightful reporting.

Projtect:2

**Project Title: Advanced Customer Churn Analysis**

**Objective:**

Analyze a telecom dataset to predict customer churn and identify factors influencing customer retention. The project includes advanced data transformations, feature engineering, and modeling.

**Assignment Tasks:**

**1. Data Loading and Initial Inspection**

1. Load the telecom dataset into a Pandas DataFrame.
2. Print the column names, data types, and descriptive statistics.
3. Check for missing values and duplicates in the dataset.
4. Summarize the target variable (Churn) distribution.

**Deliverables:**

* Overview of the dataset.
* Insights on missing values, duplicates, and churn distribution.

**2. Data Cleaning and Preprocessing**

1. Handle missing values:
   * Impute numerical columns with median or mean.
   * Impute categorical columns with the mode.
2. Remove duplicate rows, if any.
3. Convert all categorical features into numeric using one-hot encoding.

**Deliverables:**

* Cleaned dataset with imputed values.
* Updated dataset after applying one-hot encoding.

**3. Exploratory Data Analysis (EDA)**

1. Plot the distribution of the following features:
   * Tenure
   * Monthly Charges
   * Total Charges
2. Identify the top 3 features contributing to churn using visualizations.
3. Use box plots to compare churn vs. non-churn customers for numerical features.

**Deliverables:**

* Visualizations for distribution and box plots.
* Insights about churn factors.

**4. Feature Engineering**

1. Create new features from existing columns:
   * Tenure\_Group: Group tenure into categories (e.g., 0-12 months, 13-24 months, etc.).
   * Average\_Monthly\_Charges: Calculate Total Charges / Tenure.
2. Create interaction features between important numerical columns (e.g., Monthly Charges \* Tenure).
3. Perform scaling or normalization for numerical features.

**Deliverables:**

* Dataset with new features.
* Scaled or normalized numerical features.

**5. Correlation and Feature Selection**

1. Compute the correlation matrix for numerical columns.
2. Visualize the correlation using a heatmap.
3. Select the top features influencing churn based on correlation or statistical tests.

**Deliverables:**

* Heatmap of correlations.
* Final list of selected features for the model.

**6. Model Building**

1. Split the dataset into training (80%) and testing (20%) sets.
2. Build and compare the following models:
   * Logistic Regression
   * Random Forest
   * Gradient Boosting (e.g., XGBoost or LightGBM)
3. Evaluate models using:
   * Accuracy
   * Precision
   * Recall
   * F1 Score
   * ROC-AUC Score

**Deliverables:**

* Trained models.
* Performance comparison using evaluation metrics.

**7. Model Optimization**

1. Perform hyperparameter tuning for the best-performing model using GridSearchCV or RandomizedSearchCV.
2. Evaluate the optimized model on the testing set.

**Deliverables:**

* Hyperparameter-tuned model.
* Performance metrics post-tuning.

**8. Insights and Recommendations**

1. Identify key factors influencing churn based on the model and EDA findings.
2. Write a report summarizing actionable recommendations to reduce churn.

**Deliverables:**

* A detailed report with insights and recommendations.
* Visualizations highlighting key findings.

**9. Advanced Analytics (Optional)**

1. Use SHAP (SHapley Additive exPlanations) or LIME to interpret the model.
2. Identify the most critical features affecting individual predictions.

**Deliverables:**

* SHAP or LIME visualizations.
* Observations from model interpretability.

**Submission Guidelines:**

1. Submit your code as a Jupyter Notebook or Python script.
2. Include all visualizations and clearly label them.
3. Attach your report as a separate document (PDF or Word).

**Grading Criteria:**

1. **Code Quality (15%)**: Clean, well-commented, and efficient code.
2. **Data Preprocessing (20%)**: Proper handling of missing values, duplicates, and feature engineering.
3. **EDA (20%)**: Insightful visualizations and observations.
4. **Model Building (25%)**: Accurate implementation and performance evaluation.
5. **Report Quality (20%)**: Clear, concise, and actionable recommendations.