Project Synopsis

Project Title: Predicting Customer Spending in E-commerce Using Linear Regression

Objective and Scope:

To determine whether optimizing the mobile app experience or the website experience would have a greater impact on customer spending and engagement for the e-commerce company.

This objective will be achieved by analyzing the relationship between various customer attributes (Avg. Session Length, Time on App, Time on Website, Length of Membership) and customer spending. By understanding which platform (mobile app or website) has a stronger correlation with customer behavior and spending, the company can prioritize its efforts accordingly.

Process Description:

1. **Data Acquisition:** Gather relevant customer data from the e-commerce platform, including purchase history, demographics, browsing behavior, and engagement metrics.
2. **Data Preprocessing:** Clean and prepare the data by handling missing values, outliers, and inconsistencies.
3. **Exploratory Data Analysis (EDA):** Conduct a thorough analysis to identify potential relationships between customer characteristics and spending patterns.
4. **Feature Engineering:** Create new features or transform existing ones to improve model performance.
5. **Model Selection and Training:** Choose linear regression as the appropriate model and train it on the prepared dataset.
6. **Model Evaluation:** Assess the model's accuracy using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.
7. **Deployment:** Integrate the trained model into the e-commerce platform for real-time predictions.

Flow Chart :

Start

1. **Import Libraries:**
   - pandas (data manipulation)
   - matplotlib.pyplot (visualization)
   - seaborn (visualization)
2. **Read Data:**
   - Load "Ecommerce Customers.csv" dataset into a pandas DataFrame named "customers".
3. **Data Exploration:**
   - Display the first few rows using `customers.head()`.
   - Get data type and missing value information using `customers.info()`.
   - Summarize data statistics using `customers.describe()`.
   - Create visualizations to explore relationships between features and target variable:
     - Joint plots for "Time on Website" and "Yearly Amount Spent"
     - Joint plots for "Time on App" and "Yearly Amount Spent"
     - Pairplot for all features with scatter plots

        - Linear model plot for "Length of Membership" and "Yearly Amount Spent"

4. **Data Preparation:**
    - Check for missing values (not shown in your code).
    - Handle missing values if necessary (imputation, deletion).
    - Split data into features (X) and target variable (y):
        - X: DataFrame containing "Avg. Session Length", "Time on App", "Time on Website", "Length of Membership"
        - y: Series containing "Yearly Amount Spent"
    - Split data into training and testing sets using `train_test_split` (70% training, 30% testing, random state set to 42 for reproducibility).
5. **Model Building:**
    - Create a linear regression model using `LinearRegression` from scikit-learn.
    - Train the model on the training data (`X_train`, `y_train`).
6. **Model Evaluation:**
    - Get model coefficients using `lm.coef_` to understand feature importance.
    - Calculate R-squared using `lm.score(X, y)` to assess model performance.
    - Create a DataFrame (`cdf`) to display coefficients with corresponding feature names.
    - Make predictions on the testing data using `lm.predict(X_test).`
    - Visualize actual vs. predicted values using a scatter plot.
    - Calculate evaluation metrics:
        - Mean Absolute Error (MAE)
        - Mean Squared Error (MSE)
        - Root Mean Squared Error (RMSE) using `mean_absolute_error`, `mean_squared_error`, and mathematical operations.
    - Analyze model residuals:
        - Create a distribution plot of residuals using `sns.distplot`.
        - Check normality of residuals using a Q-Q plot with `scipy.stats.probplot`.

Resources and Limitations:

Resources:

- **Hardware:** A computer with sufficient processing power and memory.
- **Software:** Python programming language, data science libraries (Pandas, NumPy, Scikit-learn), and visualization tools (Matplotlib, Seaborn).
- **Data:** The e-commerce dataset containing customer information.

Limitations:

- **Linear Assumption:** Linear regression assumes a linear relationship between features and the target variable. Non-linear relationships might not be captured accurately.
- **Data Quality:** The quality of the collected data can significantly impact model performance.
- **Overfitting:** The model may become overly complex and fit the training data too well, leading to poor generalization.

Conclusion:

It can be tricky to interpret the information in this analysis. According to the model, the most significant factor for clients is not the time spent on the app or website, but their length of membership. However, of the two predictors (desktop vs app), the app has the strongest influence by far. In fact, the time spent on the desktop website does not seem to have any correlation at all! In other words, according to the data, the amount of time that the customer spends on the desktop website has almost nothing to do with the amount of money they will spend. We could interpret this in two different ways. Firstly, this could mean that the desktop website needs more work to make its visitors buy more. Secondly, it could mean that people tend to be more influenced by mobile applications of online stores than by desktop websites. So maybe efforts should be directed towards taking advantage of this fact. Indeed, the interpretation of this information requires expertise in the online marketing sphere. Our analysis and our model, however, does a very good job in weighting the predictors importance.

References:

1. **Kaggle Dataset:** Linear Regression E-commerce Dataset (kaggle.com)
2. **Python Libraries:**
    - Pandas: https://pandas.pydata.org/docs/getting_started/install.html
    - NumPy:https://numpy.org/
    - Scikit-learn: https://scikit-learn.org/
    - Matplotlib: https://matplotlib.org/
    - Seaborn: https://seaborn.pydata.org/

The students should be able to answer the following set of questions at the time of submission of the synopsis (if asked):

1. Why did you choose this project topic?
2. What data sources will you use for your project?
3. What is the scope of your project?
4. What are the expected outcomes of your project?
5. What ethical considerations are relevant to your project?