Ecommerce Analysis using Linear Regression

1. Project Overview

This project performs an analysis on an ecommerce company's dataset using **Linear Regression** to determine whether the company should focus its efforts on enhancing its **mobile app experience** or **website performance**. The study evaluates multiple factors such as **Average Session Length, Time on App, Time on Website, and Length of Membership** to predict **Yearly Amount Spent** by customers.

2. Data Retrieval & Preprocessing

2.1 Dataset

The dataset consists of 500 entries, containing:

- **Customer Information** (Email, Address, Avatar)
- **Behavioral Metrics** (Avg. Session Length, Time on App, Time on Website, Length of Membership)
- **Target Variable** (Yearly Amount Spent)

2.2 Libraries Used

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
```

2.3 Data Exploration

- **ECDF (Empirical Cumulative Distribution Function)** plotted for `Yearly Amount Spent`.
- **Correlation Analysis** to determine the relationships between variables.
- **Pairplots & Jointplots** to visualize the impact of features on the target variable.
- **Heatmaps** to check multi-collinearity between features.

3. Hypothesis & Initial Findings

3.1 Hypothesis

The assumption is that **Time on App** and **Time on Website** are key drivers of **Yearly Amount Spent**.

3.2 Key Observations

- **Time on Website** shows very low correlation with `Yearly Amount Spent` (`-0.0026`).
- **Time on App** has a significant positive correlation (`0.4993`).
- **Length of Membership** is highly correlated (`0.81`), suggesting that retaining customers increases spending over time.

4. Model Development

4.1 Train-Test Split

Features used for regression:

```
X = customers[['Avg. Session Length', 'Time on App', 'Time on Website',
'Length of Membership']]
y = customers['Yearly Amount Spent']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=101)
```

4.2 Model Training

A linear regression model is trained using Scikit-Learn's `LinearRegression()` class.

```
lm = LinearRegression()
lm.fit(X_train, y_train)
```

4.3 Model Coefficients

```
lm.coef_
```

Feature	Coefficient
Avg. Session Length	25.98
Time on App	38.59
Time on Website	0.19
Length of Membership	61.28

- **Length of Membership** has the strongest impact on spending.
- **Time on App** is more impactful than **Time on Website**.

5. Model Evaluation

5.1 Predictions

```
predictions = lm.predict(X_test)
```

5.2 Performance Metrics

```
print('R^2:', metrics.explained_variance_score(y_test, predictions))
print('MAE:', metrics.mean_absolute_error(y_test,predictions))
print('MSE:', metrics.mean_squared_error(y_test,predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
```

Metric	Value
R^2	0.989
MAE	7.22
MSE	79.81
RMSE	8.93

5.3 Residuals Analysis

- Residual plot shows **normal distribution**, indicating a good fit.

6. Insights & Business Recommendations

- **Length of Membership** has the highest influence on spending → **Customer Retention** should be a key focus.
- **Time on App** impacts spending significantly more than **Time on Website**.
- Investing in **app experience** may provide better ROI compared to improving the website.
- Conduct **economic feasibility study** before investing in platform improvements.

7. Conclusion

A **linear regression model** was successfully built to analyze ecommerce customer behavior. The findings suggest that **app development** should be prioritized over website improvements, but **customer retention strategies** should also be heavily considered due to the strong impact of **Length of Membership** on spending.