Problem Understanding

The main goal of this project was to group customers into segments and predict how much they'll spend in the future for an online retail business.
Doing this helps the business personalize marketing, optimize spending, and boost return on investment (ROI) by focusing on the right customers.

Data Preprocessing

Here’s the step-by-step of how we got the data ready:

1. Data Quality Checks:
    - Imported the dataset and fixed any missing or duplicate records.
    - Removed any transactions that showed negative quantities (which didn't make sense).
    - Created a TotalPrice feature by multiplying Quantity and UnitPrice.
2. Feature Engineering:
    - Built RFM features:
        - Recency: How recently a customer purchased.
        - Frequency: How often they purchased.
        - Monetary: How much they spent.
    - Added Average Order Value for deeper customer insight.
3. Scaling and Dimensionality Reduction:
    - Standardized feature values using StandardScaler to bring everything onto the same scale.
    - Used PCA (Principal Component Analysis) to shrink the data into 2 key components for easier visualization and faster processing:

python

CopyEdit

```
pca = PCA(n_components=2)

X_pca = pca.fit_transform(X_scaled)

rfm['PCA1'], rfm['PCA2'] = X_pca[:, 0], X_pca[:, 1]
```

- This made the customer clusters easy to visualize and interpret.

4. Customer Segmentation:
    - Applied K-Means Clustering to the RFM features.
    - Visualized the groups using the PCA components — and the segments were nicely separated!
5. Preventing Overfitting with Regularization:
    - Built both Ridge and Lasso regression models to predict how much customers might spend in the future.
    - Evaluated models using a custom evaluate_model function calculating MAE, RMSE, and R^2 scores.

Model Comparison and Selection

Here's a snapshot of the model results:

Model	MAE	RMSE	R² Score
Ridge	{ridge_mae}	{ridge_rmse}	{ridge_r2}
Lasso	{lasso_mae}	{lasso_rmse}	{lasso_r2}

- Ridge Regression performed better overall, with slightly lower errors and a better R² score.
- Lasso Regression was still valuable because it pushed less important features' coefficients to zero — great for feature selection — but it lagged a little behind Ridge in prediction accuracy.

Clustering and Visualization Insights

The PCA scatter plot after clustering showed clear, distinct groups of customers.
This proved that using PCA for dimensionality reduction made analysis easier without losing important information.

Conclusion:

- Combining RFM analysis, PCA, K-Means clustering, and Ridge regression gave powerful insights into customer behavior.
- Ridge Regression was chosen as the best model for purchase amount prediction because of its balance between performance and overfitting control.

Recommendations:

1. Use K-Means segments to design highly targeted and personalized marketing campaigns.
2. Deploy Ridge Regression to improve sales forecasting and optimize inventory planning.
3. Leverage PCA for regular data health checks and faster model updates.
4. Integrate predictions into CRM systems to prioritize high-value customers and reduce churn.
5. Explore more clustering methods like DBSCAN or Agglomerative Clustering to find even better customer groups in future analyses.