DATA ANALYST INTERNSHIP REPORT WEEK-2

Customer Behavior & Revenue Optimization for E-Commerce Business :

Objective:
Analyze customer purchase patterns, churn behavior, product trends, and marketing performance to generate insights that improve retention, sales, and overall business performance.

Suggested Datasets (choose or combine):
Cleaned Dataset provided by Team A which is from the below link:
• E-Commerce Public Dataset (Olist)

Task 3. Customer Segmentation
• Perform RFM Analysis (Recency, Frequency, Monetary):
  ➢ Group customers into value-based segments.
• Apply K-Means Clustering or DBSCAN for unsupervised customer grouping.
• Visualize clusters using PCA or t-SNE.
• Recommend strategies per cluster (e.g., VIPs vs. low-engagement users).

Project Overview

Objective: Segment customers based on their purchasing behavior and develop targeted marketing strategies for each segment.

Data Source: E-commerce transaction data stored in master_dataset.csv.

Implementation Steps

1. Data Preparation

- Loading and cleaning transaction data
- Converting timestamps to proper datetime format
- Removing incomplete records
- Establishing a reference point ("today") for recency calculations

2. RFM Feature Extraction

The analysis uses the three standard RFM metrics:

- **Recency:** Days since customer's last purchase
- **Frequency:** Number of unique orders by the customer
- **Monetary:** Total amount spent by the customer

3. Data Preprocessing

- Filtering out zero-value customers
- Standardizing features using StandardScaler
- Preparing data for clustering algorithms

4. Clustering Approaches

Two different clustering methods are implemented for comparison:

1. **MiniBatchKMeans:**
    - Memory-efficient version of K-Means
    - Configured for large datasets with batch processing
    - Fast convergence with 4 predefined clusters
2. **HDBSCAN (Hierarchical Density-Based Spatial Clustering):**
    - Density-based clustering that can find clusters of varying shapes
    - Identifies outliers as noise points
    - Configurable minimum cluster size

5. Dimensionality Reduction for Visualization

Two techniques are used to visualize high-dimensional data:

1. **Principal Component Analysis (PCA):**
    - Linear dimensionality reduction
    - Preserves global structure and variance
    - Faster computation for large datasets
2. **t-SNE (t-Distributed Stochastic Neighbor Embedding):**
    - Non-linear dimensionality reduction
    - Better at preserving local relationships

- Reveals cluster structures not visible in PCA

6. Segment Analysis and Strategy Assignment

- Computing cluster centroids in RFM space
- Analyzing cluster characteristics
- Assigning targeted marketing strategies based on segment behavior:
  - **VIP:** Reward and upsell opportunities
  - **At-Risk:** Win-back campaigns
  - **Low-Value:** Promotional offers
  - **Mid-Value:** Loyalty programs

7. Visualization of Segments

- Scatter plots of clusters using PCA coordinates
- Alternative visualization using t-SNE
- Bar chart showing distribution of customers across marketing strategies

Business Applications

This customer segmentation system can be used to:

1. Develop personalized marketing campaigns for different customer segments
2. Allocate marketing resources more efficiently
3. Identify high-value customers for retention efforts
4. Recognize at-risk customers for proactive intervention
5. Understand customer portfolio composition and value distribution

Results and Insights

The clustering approach reveals distinct customer segments based on their purchasing behavior, allowing for:

- Identification of high-value customer segments
- Recognition of customers with potential for growth
- Early detection of customers at risk of churning
- Optimization of marketing spend across segments

Future Improvements

Potential enhancements to consider:

- Including additional behavioral features (e.g., product categories, session data)
- Time-based segmentation to track customer movement between segments
- A/B testing of marketing strategies for different segments
- Predictive modeling to anticipate segment transitions
- Customer lifetime value projections by segment

Google Colab Code Link : Week_2_Task_1

Task 4. Churn Prediction Model

• Label customers as "churned" or "active."
• Feature engineering (tenure, contract type, last purchase, complaints).
• Train and evaluate 2–3 models:
  ➢ Logistic Regression
  ➢ Random Forest / XGBoost
  ➢ Neural Network (optional)
• Evaluate with:
  ➢ Accuracy, Precision, Recall
  ➢ ROC-AUC, Confusion Matrix

Project Overview

Objective: Predict whether a customer will churn (defined as no purchase in the last 6 months) based on their purchase history and behavior.

Data Source: E-commerce transaction data stored in master_dataset.csv.

Implementation Steps

1. Data Preparation

- Loading transaction data with appropriate date parsing
- Defining churn (customers with no purchases in the last 180 days)
- Merging churn labels with the main dataset

2. Feature Engineering

The model uses the following features:

- **Tenure:** Duration between first and last purchase
- **Order Frequency:** Total number of orders per customer
- **Delivery Performance:** Average delivery time
- **Customer Satisfaction:** Number of complaints (reviews ≤ 2)
- **Purchase Behavior:** Average order value and freight costs

3. Data Preprocessing

- Handling missing values with median imputation
- Standardizing numeric features
- Train-test split (80% train, 20% test)

4. Model Training

Four different models are trained and compared:

1. **Logistic Regression:** A baseline linear model
2. **Random Forest:** An ensemble of decision trees
3. **XGBoost:** Gradient boosted decision trees
4. **Neural Network:** A simple deep learning model with:
   - 2 hidden layers (32 and 16 neurons)
   - ReLU activation
   - Dropout for regularization
   - Sigmoid output for binary classification

5. Model Evaluation

Each model is evaluated using:

- Accuracy

- Precision
- Recall
- ROC AUC Score
- Confusion Matrix
- ROC Curve Visualization

Results and Insights

The model comparison allows identification of the best performing algorithm for churn prediction. Key performance indicators focus on:

- The ability to correctly identify customers at risk of churning (recall)
- Minimizing false positives to optimize retention campaign resources (precision)
- Overall discriminative power (ROC AUC)

Business Applications

This churn prediction model can be used to:

1. Identify at-risk customers for targeted retention campaigns
2. Understand key factors contributing to customer churn
3. Optimize customer lifetime value through proactive engagement
4. Reduce customer acquisition costs by improving retention

Future Improvements

Potential enhancements to consider:

- Feature importance analysis to identify key churn drivers
- Hyperparameter tuning for model optimization
- Cost-sensitive learning to account for imbalanced classes
- Temporal validation to test model stability over time
- A/B testing of retention strategies based on model prediction.

Google Colab Code Link : Week_2_Task_2

Task 6. Business Dashboard Development

Use Power BI or Tableau to create dashboards with:

• KPIs: Total Sales, Avg. Basket Size, Churn Rate, LTV

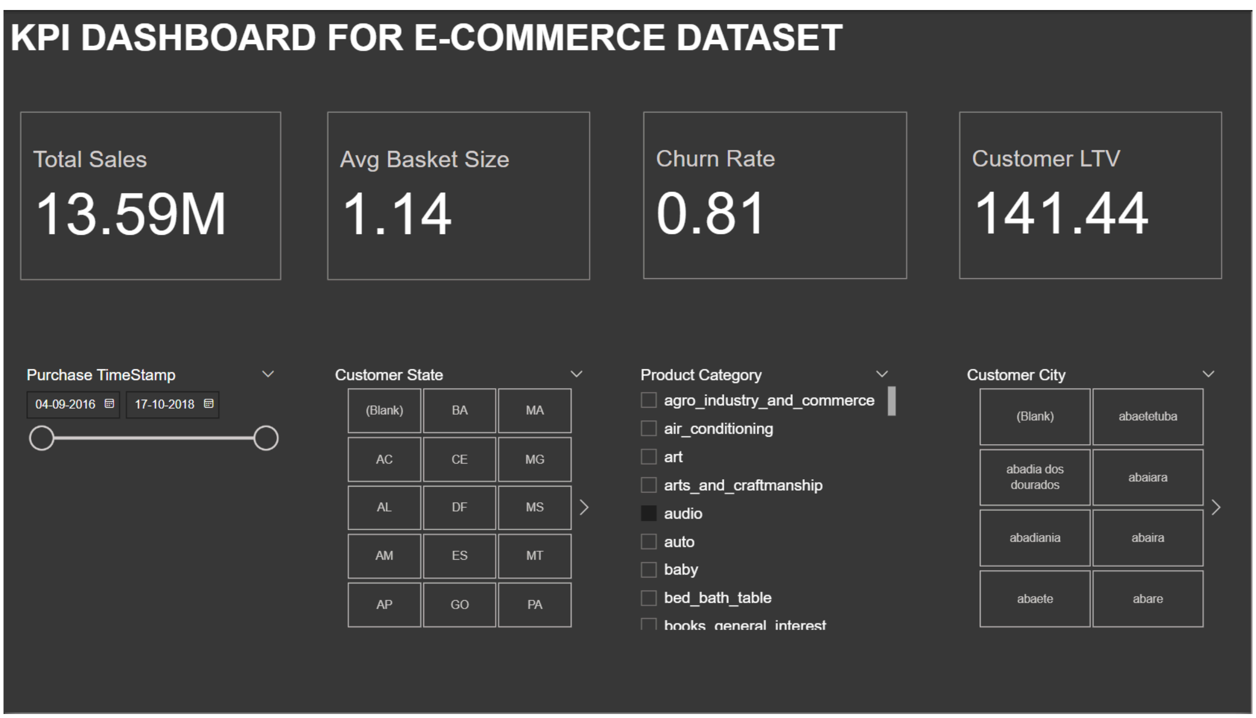
• Sales by:
  - ➢ Product category
  - ➢ Region
  - ➢ Customer segment

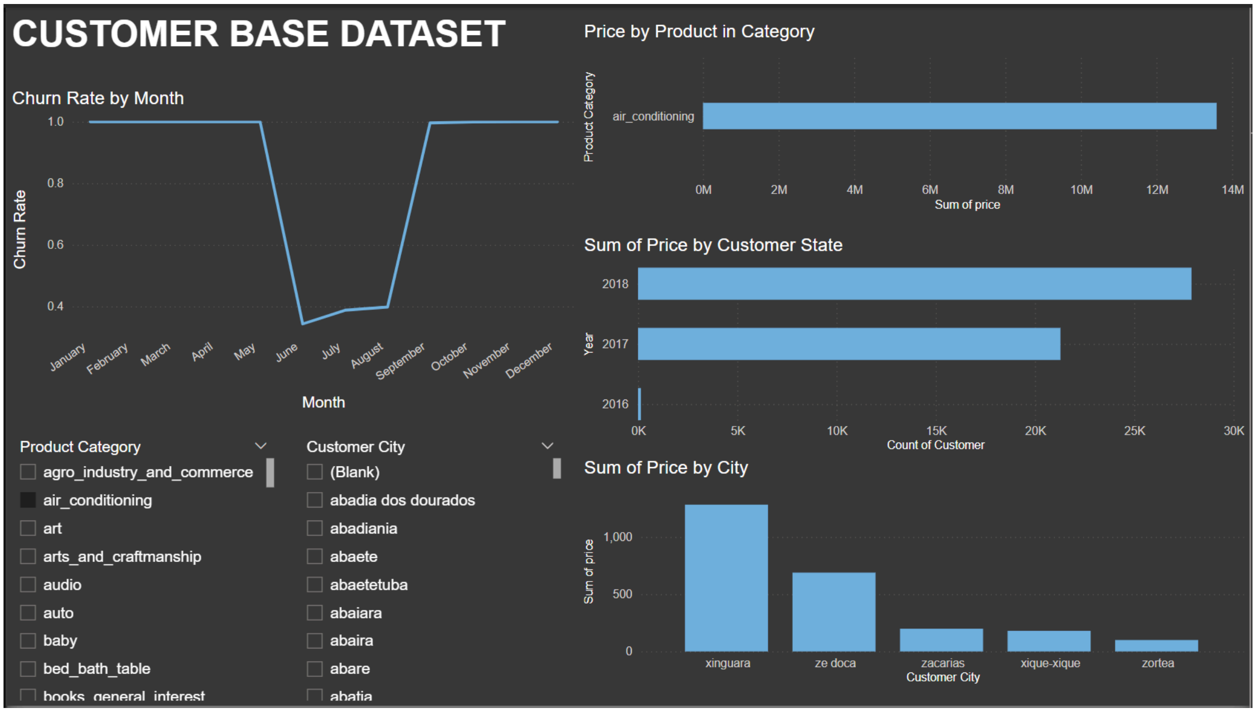
• Churn trends over time

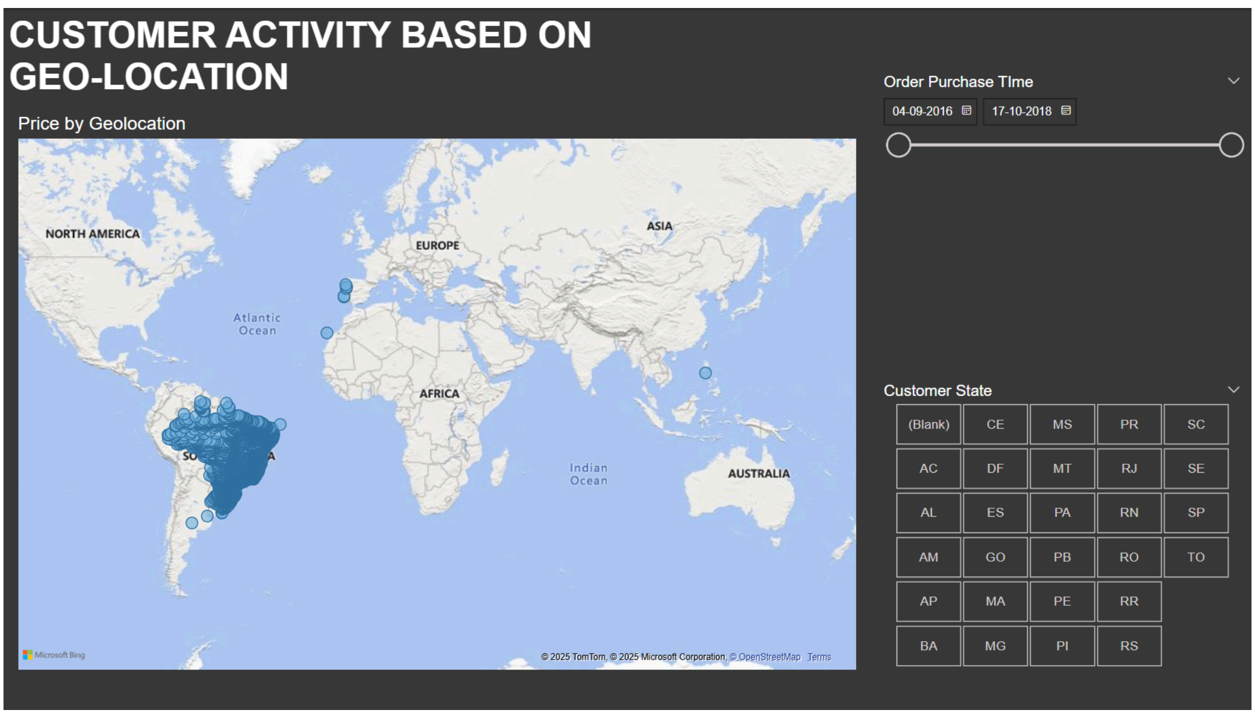
• Filters/Slicers for:
  - ➢ Time period
  - ➢ Region
  - ➢ Product type
  - ➢ Customer segment

• Geo-map of customer activity

Export to web dashboard or PDF format.







Task 7. Predictive Revenue Model

- Forecast future revenue using:
  - ➢ Time series forecasting (ARIMA, Prophet)
  - ➢ Regression models based on sales inputs

• Build scenario-based forecasts (e.g., +10% in marketing spend).

Project Overview

Objective: Forecast future revenue using historical transaction data and compare different forecasting approaches.

Data Source: E-commerce transaction data stored in master_dataset.csv.

Implementation Steps

1. Data Preparation

- Loading transaction data with appropriate date parsing
- Converting timestamps to date format for daily aggregation
- Aggregating daily revenue for time series analysis

2. Exploratory Visualization

- Plotting daily revenue trends over time
- Identifying patterns, seasonality, and potential outliers

3. Prophet Forecasting

Prophet is Facebook's time series forecasting tool that:

- Handles daily seasonality, weekly patterns, and holiday effects
- Automatically detects trend changes
- Makes 30-day forward predictions with uncertainty intervals

4. ARIMA Forecasting

The Auto-Regressive Integrated Moving Average (ARIMA) approach:

- Automatically determines optimal p, d, q parameters using auto_arima
- Creates non-seasonal time series forecasts
- Provides alternative modeling approach for comparison

5. Regression-Based Forecasting

A causal modeling approach that:

- Aggregates data monthly to reduce noise
- Uses price and freight values as predictive features
- Applies linear regression to predict payment values
- Evaluates model performance using MAE and RMSE metrics

6. Scenario Analysis

Business scenario simulation that:

- Models the impact of a 10% increase in marketing spend (simulated by increasing price)
- Compares actual, predicted, and scenario-based revenue forecasts
- Provides decision support for marketing investment decisions

Results and Insights

The multiple forecasting methods allow for:

- Cross-validation of predictions using different statistical approaches
- Identification of the most reliable forecasting technique for this dataset
- Quantification of forecast accuracy using error metrics
- Understanding of key revenue drivers through regression analysis

Business Applications

This revenue forecasting system can be used to:

1. Set realistic revenue targets for upcoming periods
2. Plan inventory and resource allocation based on expected demand
3. Simulate different business scenarios to optimize strategy
4. Identify seasonal patterns to inform marketing and promotional activities
5. Support budget planning and financial projections

Future Improvements

Potential enhancements to consider:

- Ensemble methods combining multiple forecasting approaches
- Inclusion of external variables (holidays, promotions, marketing spend)
- Implementation of deep learning models (LSTM, Transformer)
- Cross-validation techniques specific to time series data
- Hierarchical forecasting by product category or customer segment

Google Colab Code Link : Week_2_Task_3