Project BRD

Status	In progress

> Project Title: **Advanced Data Analytics for Insightful E-commerce Business Strategy**

Project Overview:

This project aims to utilize advanced data analytics techniques to analyze a comprehensive e-commerce dataset. The focus is on conducting exploratory data analysis, geospatial analysis, regression analysis, cluster analysis, and time-series analysis.

Motivation:

The goal is to uncover deep insights into sales patterns, customer behavior, product performance, and market trends to drive informed business decision-making.

Project Significance:

The project will enable the e-commerce business to leverage data-driven insights for optimizing their strategies across various facets of the business. The insights obtained from different analyses will help in fine-tuning marketing efforts, improving customer satisfaction, enhancing product offerings, and ultimately driving business growth and profitability. This comprehensive approach to data analysis is essential for staying competitive in the dynamic e-commerce market.

Objectives:

1. **Detailed Data Insights**: A deep understanding of the factors that drive sales and customer behavior.
2. **Strategic Recommendations**: Actionable strategies for product placement, pricing, marketing, and inventory management.
3. **Market Segmentation Strategy**: Effective customer and product segmentation strategies based on buying patterns and preferences.
4. **Sales Forecasting**: Improved ability to forecast future sales trends and prepare for seasonal fluctuations.
5. **Geographic Market Analysis**: Insights into how geographic factors influence sales and customer engagement.

Methodology:

1. **Data Preparation**: Clean and preprocess the dataset.

2. **Exploratory Analysis**: Utilize scatterplots, correlation heatmaps, pair plots, and categorical plots to explore relationships within the data.
3. **Geospatial Analysis**: If geographical data is available or can be linked, use shapefiles to visualize and analyze spatial patterns in sales and customer locations.
4. **Regression Analysis**: Use statistical models to understand how various factors like price, quantity, and store location influence total sales.
5. **Cluster Analysis**: Segment customers and products using clustering techniques to identify patterns and behaviors in different groups.
6. **Time-Series Analysis**: Analyze sales data over time to identify trends, cyclic patterns, and potential forecast models.

Key questions:

Exploratory Data Analysis Questions:

1. **What are the top-selling products, and what common characteristics do they share?**
2. **How does customer purchasing behavior vary across different demographic groups?**
3. **Are there any unexpected correlations between different product categories and sales volumes?**
4. **Which products are frequently purchased together?**

Geospatial Analysis Questions:

1. **What are the geographic patterns in sales distribution?**
2. **Do certain regions show a higher preference for specific product categories?**
3. **How do shipping distances and locations impact sales and customer satisfaction?**

Regression Analysis Questions:

1. **What factors most significantly affect the total sales price?**
2. **How does pricing affect sales volumes for different product categories?**

3. **Is there a relationship between discount levels and sales volumes?**

Cluster Analysis Questions:

1. **Can we identify distinct customer segments based on purchasing patterns?**
2. **Are there specific product clusters that indicate niche market opportunities?**
3. **How do customer segments vary in terms of product preferences and spending habits?**

Time-Series Analysis Questions:

1. **How do sales trends vary over time (daily, weekly, monthly, seasonally)?**
2. **Can we predict future sales trends based on historical data?**
3. **Are there specific times of the year when sales peak or decline significantly?**

Additional Questions:

1. **What is the impact of external factors (like holidays or economic changes) on sales trends?**
2. **How do inventory levels correlate with sales performance?**
3. **What is the customer feedback and rating pattern for top-selling products?**

Tools and Technologies:

- Data Visualization Tools - Tableau
- Statistical and Data Analysis Software - Python with Pandas, Scikit-learn, Seaborn, Matplotlib
- Geographic Information System (GIS) tools for geospatial analysis (if applicable)

Assumption & constraints:

Assumptions

1. **Data Quality and Completeness**:

- It's assumed that the data in the various CSV files (fact_table, Trans_dim, customer_dim, item_dim, store_dim, time_dim) is complete, accurate, and up-to-date. Inaccuracies or missing data can skew analysis results.

2. **Data Structure Consistency**:
   - The structure of the data (columns, data types) in each CSV file is assumed to be consistent throughout the dataset. Inconsistent data structures can lead to challenges in integration and analysis.
3. **Relevance of Data**:
   - The assumption here is that the collected data is relevant and sufficient for the intended analysis. This includes demographic information, transaction details, product information, etc.
4. **Stable Data Sources**:
   - The assumption that the sources of the data, particularly external sources, are stable and reliable over time.
5. **Compliance and Ethical Use**:
   - It's assumed that the data collection and usage comply with relevant data protection laws and ethical standards, especially concerning PII.

Constraints

1. **Technical Limitations**:
   - Constraints related to the tools and technology used (e.g., Python libraries, database systems) which might limit the types of analysis which can performed or the volume of data can be processed efficiently.
2. **Data Security and Privacy**:
   - Ensuring the security and privacy of the data, especially PII, might limit how data can be shared, accessed, or used for analysis. This includes encryption, access controls, and compliance with laws like GDPR.
3. **Data Integration Challenges**:
   - Combining data from multiple sources (various CSV files) might present challenges due to differences in data formats, scales, or conventions.
4. **Scalability Issues**:

- Handling large volumes of data especially in a growing e-commerce environment could be a constraint, depending on the existing infrastructure.

5. **Bias and Representation**:
    - The data may not be fully representative of the broader customer base or market trends, leading to potential biases in analysis results.

Data profile

To develop a basic understanding of dataset and perform an initial descriptive analysis, by profiling each variable. This process involves examining the distribution, central tendency, and dispersion of dataset,

Understanding Data

1. **Review Variables**:
    - For each key (e.g., `payment_key`, `customer_key`, `time_key`, `item_key`, `store_key`), determine the range, unique count, and any patterns.
    - For quantitative variables like `quantity`, `unit_price`, and `total_price`, calculate summary statistics (mean, median, mode, range, standard deviation).
    - Explore distributions of these variables through histograms or box plots to identify any skewness or outliers.
2. **Data Profiling**:
    - Create a data profile that includes:
        - Data type (numeric, categorical) for each variable.
        - No Missing values count.
        - Frequency distribution for categorical data (e.g., most common customer segments, product categories).
        - Temporal patterns in `time_key` (like trends over days, months, or specific seasons).
3. **Initial Observations**:
    - Note initial observations like common trends, anomalies, or interesting correlations.

- Identify potential areas for deeper analysis, such as time periods with unusually high or low sales.

Considerations of Limitations and Ethics

1. **Data Limitations**:
   - **Completeness**: Assess if the dataset comprehensively covers all relevant aspects of the e-commerce operations.
   - **Accuracy**: Determine the reliability of the data sources and any potential biases in data collection.
   - **Timeliness**: Consider whether the data is up-to-date and relevant for current market conditions.
2. **Ethical Considerations**:
   - **Privacy**: Ensure customer data is anonymized and used in compliance with privacy laws (like GDPR or CCPA).
   - **Consent**: Confirm that data was collected with proper consent, especially for customer demographic information.
   - **Bias and Fairness**: Be aware of potential biases in the data that could lead to unfair conclusions or decisions, especially in customer segmentation and product recommendations.
   - **Transparency**: Maintain transparency in how data is used and analyzed, particularly when making predictions or business decisions based on this data.
3. **Data Usage and Sharing**:
   - Consider the implications of sharing data insights or models with third parties.
   - Ensure that data sharing complies with all legal and ethical guidelines.

Star Schema Overview

The **Star Schema** is a database design ideal for data warehousing and business intelligence. It features a central **fact table** surrounded by **dimension tables**, resembling a star. This design enhances performance, simplifies queries, and is easy to understand.

Data dictionary

E-commerce Star Schema:

Our dataset uses the star schema to effectively analyze e-commerce transactions.

Fact Table:

- `fact_table.csv`: Central table containing transaction metrics and keys linking to dimension tables.

Dimension Tables:

- `Trans_dim.csv`: Transaction details like type and status.
- `customer_dim.csv`: Customer demographics and contact information.
- `item_dim.csv`: Product names, categories, and pricing.
- `store_dim.csv`: Store locations and characteristics.
- `time_dim.csv`: Detailed transaction times for trend analysis.

Data set:

This project dataset is choosen from kaggle website Kaggle LINK

Further analysis:

When embarking on an e-commerce data analysis project in Python, utilizing a star schema, we should be mindful of potential limitations such as Python's in-memory data processing constraints, performance issues with large datasets, and the lack of built-in support for concurrency and real-time processing. Python scripts may also present challenges in scalability, data integrity, complex querying, security, and maintenance compared to traditional database systems.

To overcome these issues, considering the following strategies:

- Employ libraries like Dask or save the cleaned dataset into pickles for handling large datasets that exceed available RAM.
- Integrate Python with a relational database to leverage efficient data storage and SQL querying capabilities.
- Utilize cloud-based solutions for enhanced scalability and real-time data processing tools for handling streaming data.
- Ensure robust documentation and modular code design for better maintainability.