Data Analytics 101: E-commerce Performance Analysis

Capstone Project Assignment Guide

Project Overview

In this capstone project, you will analyze an e-commerce marketplace's performance to help the business understand their customer behavior, sales patterns, and develop predictive models for future sales. This project will integrate all the key concepts you've learned throughout the course, from data analysis to predictive modeling.

Dataset Information

You will work with the E-commerce Public Dataset. This rich dataset includes:

- Customer information
- Order details and status
- Product information
- Payment details
- Review ratings
- Geographic data

Phase 1: Understanding and Preparing Your Data (Part 1)

Part A: Initial Data Exploration

Your first task is to understand the structure and content of your dataset:

1. Data Loading and Overview
    - Load each CSV file from the dataset
    - Examine the structure of each table
    - Create a data dictionary for your project
    - Document the relationships between different tables
2. Data Quality Assessment
    - Identify missing values in each column
    - Check for duplicate entries
    - Verify data types of each column
    - Look for potential outliers
    - Document any anomalies you find

Part B: Data Integration

Develop a strategy to combine relevant tables:

1. Planning Your Integration
    - Decide which tables you need for your analysis
    - Plan your join strategy
    - Identify potential challenges in merging data
2. Creating Your Analytical Dataset
    - Combine tables using appropriate joins
    - Ensure no unintended data loss
    - Verify the accuracy of your merged dataset

Phase 2: Data Cleaning and Feature Engineering (Part 2)

1. Data Cleaning Strategy
    - Handle missing values with appropriate techniques
    - Remove or handle duplicate entries
    - Convert data types as needed
    - Handle outliers appropriately
    - Document all cleaning decisions
2. Feature Engineering
    - Create time-based features from order dates
    - Calculate customer-centric metrics
    - Develop product-based features
    - Generate geographical insights
    - Create any additional relevant features
3. Data Validation
    - Verify the integrity of your cleaned dataset
    - Ensure all engineered features are correctly calculated
    - Document any assumptions made

Phase 3: Exploratory Data Analysis (Part 3)

1. Temporal Analysis
    - Analyze daily, weekly, and monthly sales patterns
    - Identify seasonal trends
    - Study delivery time patterns
    - Examine payment patterns over time
2. Customer Analysis
    - Analyze customer geographic distribution
    - Study purchase frequency patterns
    - Examine customer spending habits
    - Analyze the impact of reviews on future purchases
3. Product Analysis
    - Identify top-selling products and categories
    - Analyze price distributions

    - Study product category relationships
    - Examine product return rates

Phase 4: Statistical Analysis and Modeling (Part 4)

1. Predictive Modeling
    - Pick one prediction problem where you can use a regression technique to predict (ex. Order value per customer , days to order etc). Document the hypothesis etc.,
    - Choose your target variable.
    - Select appropriate features for modeling
    - Split your data into training and testing sets
    - Build and evaluate multiple linear regression models
    - Document your model selection process
2. Model Evaluation
    - Assess model assumptions
    - Calculate and interpret key metrics (R^2, RMSE, MAE)
    - Analyze feature importance
    - Validate your model on test data
    - Document model limitations

Deliverables

1. Project Code
    - Well-commented Python scripts or notebooks
    - Organized and reproducible code structure
    - Requirements.txt file
2. Technical Documentation
    - Data processing methodology
    - Feature engineering decisions
    - Model development process
    - Analysis of results
3. Business Report
    - Executive summary
    - Key findings and insights
    - Actionable recommendations
    - Visual representations of key metrics
4. Presentation
    - 10-minute presentation
    - Key insights and recommendations
    - Visual supports
    - Business impact assessment

Evaluation Criteria

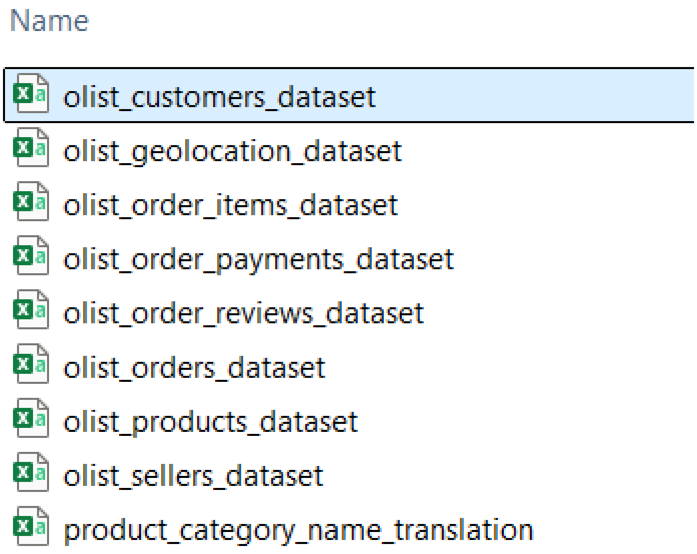
- Code Quality and Documentation (25%)
- Analysis Depth and Creativity (25%)
- Analytical Rigor (25%)
- Business Insights and Recommendations (25%)

Technical Requirements

- All analysis must be reproducible
- Code must be well-commented and formatted
- Git repository with regular commits

Data Description

1. Files



Name

olist_customers_dataset

olist_geolocation_dataset

olist_order_items_dataset

olist_order_payments_dataset

olist_order_reviews_dataset

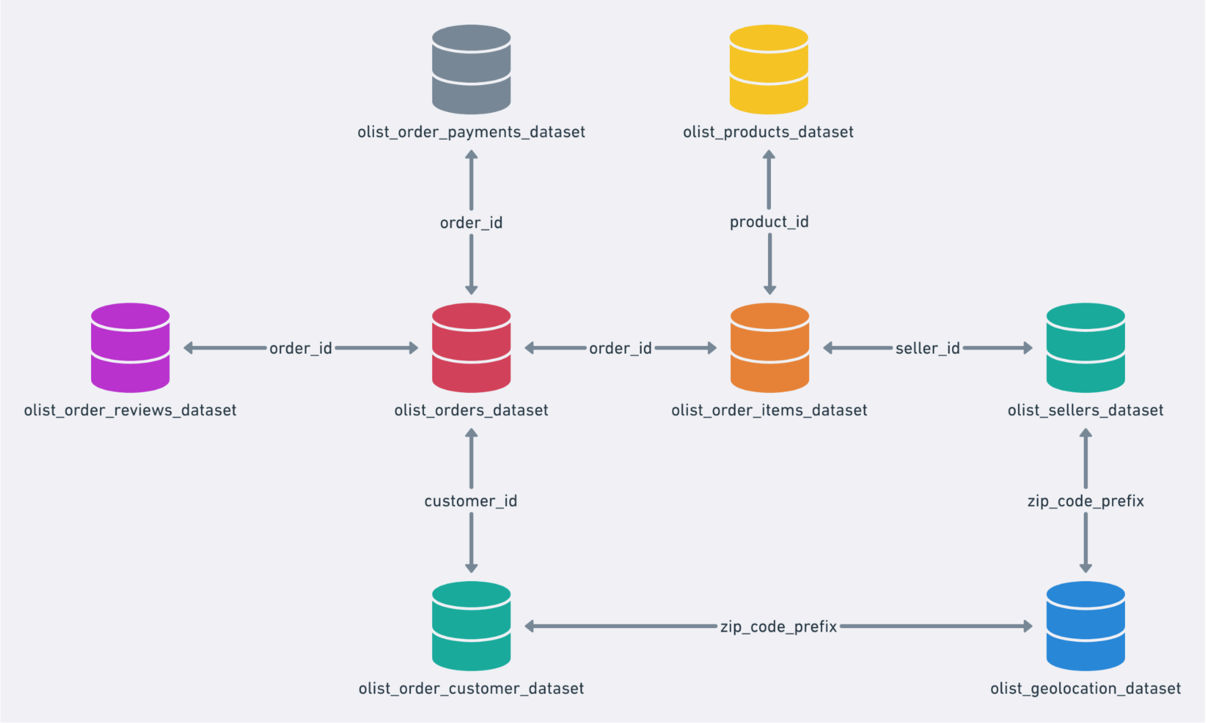
olist_orders_dataset

olist_products_dataset

olist_sellers_dataset

product_category_name_translation

2. Relationships


olist_order_payments_dataset
olist_products_dataset
order_id
product_id
olist_order_reviews_dataset
order_id
olist_orders_dataset
order_id
olist_order_items_dataset
seller_id
olist_sellers_dataset
customer_id
zip_code_prefix
olist_order_customer_dataset
zip_code_prefix
olist_geolocation_dataset