| **Project Title** | **🛒 Amazon India: A Decade of Sales Analytics 📈🇮🇳** |
| --- | --- |
| **Skills take away From This Project** | **Python • Pandas • Matplotlib • Seaborn • Data Cleaning • SQL • PowerBI/Streamlit • Business Intelligence • Statistical Analysis** |
| **Domain** | **E-Commerce Analytics** |

****

## **🎯 Problem Statement**

Build a comprehensive e-commerce analytics platform using Amazon India's 10-year transactional data (2015-2025) to create an end-to-end data pipeline from raw messy data to professional business intelligence dashboards. The platform should deliver:

● 🧹 **Advanced data cleaning** on realistic messy e-commerce data with 25% data quality issue

● 📊 **Professional EDA visualizations** with 20 comprehensive analytical plots

● 🗄️ **SQL database integration** for cleaned data storage and dashboard connectivity

● 📈 **Interactive dashboard creation** using PowerBI or Streamlit with 25-30 business-focused analytical charts

● 💡 **Strategic business insights** for data-driven decision making

## **💼 Business Use Cases**

### **1. 🏢 E-Commerce Platform Management**

● Revenue trend analysis and growth forecasting across decade

● Product category performance optimization and inventory planning

● Customer segmentation strategies for targeted marketing campaigns

● Geographic expansion analysis across Indian tier cities

### **2. 📊 Business Intelligence & Strategic Planning**

● Executive dashboard creation for C-level decision making

● KPI monitoring and performance tracking across regions

● Seasonal pattern analysis for promotional campaign timing

● Market penetration insights and competitive positioning

### **3. 🎯 Digital Marketing & Customer Analytics**

● Customer behavior analysis for personalization engines

● Festival sales impact measurement and ROI optimization

● Payment method evolution tracking for fintech partnerships

● Prime membership value analysis and growth strategies

### **4. 💳 Financial & Operational Excellence**

● Revenue optimization through pricing strategy analysis

● Delivery performance tracking across geographic regions

● Return rate analysis and quality improvement initiatives

● Cost structure analysis and operational efficiency gains

### **5. 🎓 Educational & Portfolio Development**

● Real-world data science project for professional portfolio

● Advanced Pandas operations on large datasets (Almost 1M records)

● Professional data visualization and storytelling techniques

● End-to-end business intelligence development skills

## 

## 

**🏗 Approach**

### **Key Development Steps**

#### **1. 🗂️ Dataset Understanding & Analysis**

○ Analyze comprehensive e-commerce dataset with almost 1M transactions

○ Understand product catalog with 2000+ items across multiple categories

○ Explore realistic data quality issues including missing values, format inconsistencies

○ Study Indian e-commerce market evolution from 2015-2025

#### **2. 🧹 Advanced Data Cleaning Pipeline**

○ Handle missing values using strategic imputation techniques

○ Standardize inconsistent date formats, price formats, and categorical data

○ Clean geographic data and resolve city name variations

○ Remove duplicates and handle outlier values appropriately

#### **3. 📈 Exploratory Data Analysis (EDA)**

○ Comprehensive statistical analysis with 20 visualization challenges

○ Revenue trends, seasonality patterns, and growth analysis

○ Customer segmentation and behavioral insights

○ Product performance and category analysis

#### **4. 🗄️ SQL Database Integration**

○ Design optimized database schema for analytics

○ Create tables for transactions, products, and customers

○ Implement data loading and validation procedures

○ Build SQL queries for dashboard connectivity

#### 

#### **5. 📊 Interactive Dashboard Development**

○ Multi-page analytics application using PowerBI or Streamlit

○ 25-30 comprehensive business intelligence visualizations

○ Interactive filtering and drill-down capabilities

○ Executive summary reports with actionable insights

## 

## **🧮 Data Cleaning Practice Questions (10 Complete Challenges)**

### **Question 1**

Your dataset contains order\_date in multiple formats: 'DD/MM/YYYY', 'DD-MM-YY', 'YYYY-MM-DD', and some invalid entries like '32/13/2020'. Clean and standardize all dates to 'YYYY-MM-DD' format, handling invalid dates appropriately.

### **Question 2**

The original\_price\_inr column contains mixed data types: numeric values, text with '₹' symbols, comma separators ('₹1,25,000'), and some entries like 'Price on Request'. Clean this column to contain only numeric values in Indian Rupees.

### **Question 3**

Customer ratings appear in various formats: '5.0', '4 stars', '3/5', '2.5/5.0', and some missing values. Standardize all ratings to numeric scale 1.0-5.0, handling inconsistent formats and missing values strategically.

### **Question 4**

The customer\_city column has inconsistent naming: 'Bangalore/Bengaluru', 'Mumbai/Bombay', 'Delhi/New Delhi', along with spelling errors and case variations. Standardize all city names and handle geographical variations.

### **Question 5**

Boolean columns (is\_prime\_member, is\_prime\_eligible, is\_festival\_sale) contain mixed values: True/False, Yes/No, 1/0, Y/N, and some missing entries. Convert all boolean columns to consistent True/False format.

### **Question 6**

Product categories have variations: 'Electronics/Electronic/ELECTRONICS/Electronics & Accessories'. Standardize category names across the dataset and ensure consistent naming conventions.

### **Question 7**

The delivery\_days column contains negative values, text entries like 'Same Day', '1-2 days', and some unrealistic values like 50 days. Clean this column to contain only valid numeric delivery days.

### **Question 8**

Identify and handle duplicate transactions where the same customer, product, date, and amount appear multiple times. Some duplicates are genuine (bulk orders) while others are data errors. Develop a strategy to distinguish and handle both cases.

### **Question 9**

The dataset contains outlier prices where some products show prices 100x higher than expected due to data entry errors (decimal point issues). Identify and correct these outliers using statistical methods and domain knowledge.

### **Question 10**

Payment methods contain inconsistent naming: 'UPI/PhonePe/GooglePay', 'Credit Card/CREDIT\_CARD/CC', 'Cash on Delivery/COD/C.O.D'. Standardize payment method categories and create a clean categorical hierarchy.

## 

## 

## **📊 Exploratory Data Analysis (EDA) Questions (20 Complete Visualizations)**

### **Question 1**

Create a comprehensive revenue trend analysis showing yearly revenue growth from 2015-2025. Include percentage growth rates, trend lines, and highlight key growth periods with annotations.

### **Question 2**

Analyze seasonal patterns in sales data. Create monthly sales heatmaps and identify peak selling months. Compare seasonal trends across different years and categories.

### **Question 3**

Build a customer segmentation analysis using RFM (Recency, Frequency, Monetary) methodology. Create scatter plots and segment customers into meaningful groups with actionable insights.

### **Question 4**

Visualize the evolution of payment methods from 2015-2025. Show the rise of UPI, decline of COD, and create stacked area charts to demonstrate market share changes over time.

### **Question 5**

Perform category-wise performance analysis. Create treemaps, bar charts, and pie charts showing revenue contribution, growth rates, and market share for each product category.

### **Question 6**

Analyze Prime membership impact on customer behavior. Compare average order values, order frequency, and category preferences between Prime and non-Prime customers using multiple visualization types.

### **Question 7**

Create geographic analysis of sales performance across Indian cities and states. Build choropleth maps and bar charts showing revenue density and growth patterns by tier (Metro/Tier1/Tier2/Rural).

### **Question 8**

Study festival sales impact using before/during/after analysis. Visualize revenue spikes during Diwali, Prime Day, and other festivals with detailed time series analysis.

### **Question 9**

Analyze customer age group behavior and preferences. Create demographic analysis with category preferences, spending patterns, and shopping frequency across different age segments.

### **Question 10**

Build price vs demand analysis using scatter plots and correlation matrices. Analyze how pricing strategies affect sales volumes across different categories and customer segments.

### **Question 11**

Create delivery performance analysis showing delivery days distribution, on-time performance, and customer satisfaction correlation with delivery speed across different cities and customer tiers.

### **Question 12**

Analyze return patterns and customer satisfaction using return rates, reasons, and correlation with product ratings, prices, and categories through multiple visualization techniques.

### **Question 13**

Study brand performance and market share evolution. Create brand comparison charts, market share trends, and competitive positioning analysis across different categories.

### **Question 14**

Build customer lifetime value (CLV) analysis using cohort analysis, retention curves, and CLV distribution across different customer segments and acquisition years.

### **Question 15**

Analyze discount and promotional effectiveness. Create discount impact analysis showing correlation between discount percentages, sales volumes, and revenue across categories and time periods.

### **Question 16**

Study product rating patterns and their impact on sales. Analyze rating distributions, correlation with sales performance, and identify patterns across categories and price ranges.

### **Question 17**

Create customer journey analysis showing purchase frequency patterns, category transitions, and customer evolution from first purchase to loyal customers using flow diagrams and transition matrices.

### **Question 18**

Analyze inventory and product lifecycle patterns. Study product launch success, decline phases, and category evolution over the decade with detailed trend analysis.

### **Question 19**

Build competitive pricing analysis comparing brand positioning, price ranges, and market penetration strategies across different product categories using box plots and competitive matrices.

### **Question 20**

Create a comprehensive business health dashboard combining key metrics like revenue growth, customer acquisition, retention rates, and operational efficiency using multi-panel visualizations with executive summary insights.

## **🗄️ SQL Database Integration & Storage**

### 

### **Database Schema Design**

Create optimized database structure for analytics and dashboard connectivity:

**Tables to Create:**

* transactions - Main transaction data with proper indexing
* products - Product catalog with category hierarchies
* customers - Customer master data with segmentation
* time\_dimension - Date dimension table for time-based analysis

### **SQL Implementation Steps**

1. **Database Setup**: Create database and establish connections
2. **Table Creation**: Design optimized schemas with proper data types
3. **Data Loading**: Bulk insert cleaned data with validation
4. **Indexing**: Create indexes on frequently queried columns
5. **Dashboard Connection**: Configure database connectivity for PowerBI/Streamlit

### **Key SQL Operations Required**

* Data loading and validation procedures
* Aggregation queries for dashboard KPIs
* Complex joins for multi-table analysis
* Performance optimization for large datasets
* Connection setup for visualization tools

## **📈 Dashboard Development Questions (30 Comprehensive Challenges)**

### **Executive Dashboard (Questions 1-5)**

### **Question 1**

Create an Executive Summary Dashboard showing key business metrics: Total Revenue, Growth Rate, Active Customers, Average Order Value, and Top Performing Categories. Include year-over-year comparisons and trend indicators.

### **Question 2**

Build a Real-time Business Performance Monitor displaying current month performance vs targets, revenue run-rate, customer acquisition metrics, and key operational indicators with alerts for underperformance.

### **Question 3**

Design a Strategic Overview Dashboard showing market share analysis, competitive positioning, geographic expansion metrics, and business health indicators for C-level decision making.

### **Question 4**

Create a Financial Performance Dashboard with revenue breakdown by categories, profit margin analysis, cost structure visualization, and financial forecasting models.

### **Question 5**

Build a Growth Analytics Dashboard tracking customer growth, market penetration, product portfolio expansion, and strategic initiative performance with predictive insights.

### **Revenue Analytics (Questions 6-10)**

### **Question 6**

Design a Revenue Trend Analysis Dashboard showing monthly/quarterly/yearly revenue patterns, growth rates, seasonal variations, and revenue forecasting with interactive time period selection.

### **Question 7**

Create a Category Performance Dashboard analyzing revenue contribution, growth trends, market share changes, and category-wise profitability with drill-down capabilities.

### **Question 8**

Build a Geographic Revenue Analysis showing state-wise and city-wise performance, tier-wise growth patterns, and market penetration opportunities with interactive maps.

### **Question 9**

Design a Festival Sales Analytics Dashboard tracking festival period performance, campaign effectiveness, promotional impact, and seasonal revenue optimization insights.

### **Question 10**

Create a Price Optimization Dashboard analyzing price elasticity, discount effectiveness, competitive pricing, and revenue impact of pricing strategies.

### **Customer Analytics (Questions 11-15)**

### **Question 11**

Build a Customer Segmentation Dashboard using RFM analysis, behavioral segmentation, lifetime value analysis, and targeted marketing recommendations with interactive customer profiles.

### **Question 12**

Design a Customer Journey Analytics Dashboard tracking acquisition channels, purchase patterns, category transitions, and customer evolution from first-time to loyal customers.

### **Question 13**

Create a Prime Membership Analytics Dashboard analyzing Prime vs non-Prime behavior, membership value analysis, retention rates, and Prime-specific business insights.

### **Question 14**

Build a Customer Retention Dashboard showing cohort analysis, churn prediction, retention strategies effectiveness, and customer lifecycle management insights.

### **Question 15**

Design a Demographics & Behavior Dashboard analyzing age group preferences, spending patterns, geographic behaviors, and targeted marketing opportunities.

### **Product & Inventory Analytics (Questions 16-20)**

### **Question 16**

Create a Product Performance Dashboard ranking products by revenue, units sold, ratings, and return rates with category-wise analysis and product lifecycle tracking.

### **Question 17**

Build a Brand Analytics Dashboard comparing brand performance, market share evolution, customer preferences, and competitive positioning across categories.

### **Question 18**

Design an Inventory Optimization Dashboard analyzing product demand patterns, seasonal trends, inventory turnover, and demand forecasting for better inventory management.

### **Question 19**

Create a Product Rating & Review Dashboard analyzing rating distributions, review sentiment, correlation with sales, and product quality insights.

### **Question 20**

Build a New Product Launch Dashboard tracking launch performance, market acceptance, competitive analysis, and success metrics for product development teams.

### **Operations & Logistics (Questions 21-25)**

### **Question 21**

Design a Delivery Performance Dashboard analyzing delivery times, on-time delivery rates, geographic performance variations, and operational efficiency metrics.

### **Question 22**

Create a Payment Analytics Dashboard showing payment method preferences, transaction success rates, payment trends evolution, and financial partnership insights.

### **Question 23**

Build a Return & Cancellation Dashboard analyzing return rates, return reasons, cost impact, and quality improvement opportunities with category-wise breakdown.

### **Question 24**

Design a Customer Service Dashboard tracking customer satisfaction scores, complaint categories, resolution times, and service quality improvements.

### **Question 25**

Create a Supply Chain Dashboard monitoring supplier performance, delivery reliability, cost analysis, and vendor management insights.

### **Advanced Analytics (Questions 26-30)**

### **Question 26**

Build a Predictive Analytics Dashboard with sales forecasting, customer churn prediction, demand planning, and business scenario analysis using advanced statistical models.

### **Question 27**

Design a Market Intelligence Dashboard analyzing competitor tracking, market trends, pricing intelligence, and strategic positioning insights.

### **Question 28**

Create a Cross-selling & Upselling Dashboard identifying product associations, recommendation effectiveness, bundle opportunities, and revenue optimization strategies.

### **Question 29**

Build a Seasonal Planning Dashboard for inventory planning, promotional calendar, resource allocation, and seasonal business optimization.

### **Question 30**

### Design a Business Intelligence Command Center integrating all key metrics, automated alerts, performance monitoring, and strategic decision support tools.

## 

## **🎯 Expected Results**

### **📌 Technical Deliverables**

**Dashboard Features (PowerBI/Streamlit):**

* 25-30 comprehensive business intelligence visualizations
* Interactive filtering and drill-down capabilities
* Real-time data connectivity and refresh capabilities
* Mobile-responsive design for executive access
* Automated alert systems for key performance indicators

**Data Pipeline Components:**

* Complete data cleaning and validation procedures
* Optimized SQL database with proper indexing
* Automated data loading and refresh mechanisms
* Data quality monitoring and reporting systems
* Scalable architecture for future data growth

### 

### **📊 Business Intelligence Outputs**

* Executive summary dashboards for strategic decision making
* Operational dashboards for daily business monitoring
* Customer analytics for marketing and retention strategies
* Product performance insights for inventory optimization
* Financial analytics for revenue management and forecasting

**Technical Tags:**

Python Streamlit SQL Database Pandas PowerBI requests Data Analytics Business Intelligence E-commerce Analytics Statistical Analysis Data Visualization

**📊 Data Set:**

[Dataset](https://drive.google.com/drive/folders/1ZHB4x8nZHuXmyDlwujWtbOaxiMHWf-3-?usp=drive_link)

## **📊 Dataset Overview**

The Amazon India dataset contains almost **1,000,000 transactions** spanning **2015-2025** with realistic e-commerce patterns and data quality challenges designed for learning.

**Core Files:**

* amazon\_india\_complete\_2015\_2025.csv - Complete transaction dataset ( Almost 1M records)
* amazon\_india\_products\_catalog.csv - Product master data (2000+ products)
* amazon\_india\_{year}.csv - Year-wise transaction files (2015-2025)

### **Key Features & Columns**

### **Transaction Data (45+ columns):**

* **Identifiers**: transaction\_id, customer\_id, product\_id
* **Temporal**: order\_date, order\_month, order\_quarter, order\_year
* **Product**: product\_name, category, subcategory, brand, product\_rating
* **Pricing**: original\_price\_inr, discount\_percent, final\_amount\_inr, delivery\_charges
* **Customer**: customer\_city, customer\_state, age\_group, is\_prime\_member
* **Operations**: payment\_method, delivery\_days, return\_status, customer\_rating
* **Business**: is\_festival\_sale, festival\_name, customer\_spending\_tier

**Product Catalog (12+ columns):**

* product\_id, product\_name, category, subcategory, brand, base\_price\_2015
* weight\_kg, rating, is\_prime\_eligible, launch\_year, model
* **Data Characteristics**
* **Geographic Coverage**: 30+ Indian cities (Metro to Rural)
* **Product Categories**: 8 major categories, 25+ subcategories
* **Brands**: 100+ realistic Indian and global brands
* **Time Evolution**: Realistic price inflation, payment method adoption (UPI growth)
* **Data Quality Issues**: 25% intentional quality issues for cleaning practice
* **Business Patterns**: Festival seasons, Prime membership growth, customer behavior evolution

**Project Deliverables:**

* **Source Code**: Complete Python data pipeline and analysis scripts
* **Database Schema**: SQL table structure optimized for analytics
* **Dashboard Files**: PowerBI (.pbix) or Streamlit application files
* **Documentation**: Project setup instructions and data dictionary
* **Requirements**: Dependencies list (pandas, matplotlib, seaborn, sqlalchemy, streamlit/powerbi)
* **Analytics Reports**: Complete EDA insights and business recommendations
* **Cleaned Dataset**: Production-ready processed data files

**Project Guidelines:**

* **Coding Standards**: Follow PEP 8 Python style guidelines
* **Modularity**: Separate functions for different operations
* **Documentation**: Clear comments and docstrings for all functions
* **Version Control**: Use Git for code versioning and collaboration
* **Consistency:** Standardize visuals, colors, and layouts

**Timeline:**

The project should be completed and submitted **within 10 days** from the date it is assigned.

**PROJECT DOUBT CLARIFICATION SESSION ( PROJECT AND CLASS DOUBTS)**

**About Session:** The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.

**Note: Book the slot at least before 12:00 Pm on the same day**

**Timing: Monday-Saturday (3:30PM to 4:30PM)**

**Booking link :**[**https://forms.gle/XC553oSbMJ2Gcfug**](https://forms.gle/XC553oSbMJ2Gcfug9)**LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)**

**About Session:** The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.

**Note: This form will Open only on Saturday (after 2 PM ) and Sunday on Every Week**

**Timing:**

**For DS and AIML**

**Monday-Saturday (05:30PM to 07:00PM)**

**Booking link :** [**https://forms.gle/1m2Gsro41fLtZurRA**](https://forms.gle/1m2Gsro41fLtZurRA)

**References:**

| TOPIC | LINK |
| --- | --- |
| **Project Live Evaluation** | [Project Live Evaluation](https://docs.google.com/document/u/0/d/1QisLD2kqDWFZJG2oDknKn2eMGi-Xq8oFPgA7UWSbcIQ/edit) |
| **EDA Guide** | [Exploratory Data Analysis (EDA) Guide](https://docs.google.com/document/d/1tHiTU1X9UwXSLySpJ-FVCohlf_8xpXwa75vlK9S6wl8/edit?usp=sharing) |
| **Capstone Explanation Guideline** | [Capstone Explanation Guideline](https://docs.google.com/document/d/1gbhLvJYY7J73lu1g9c6C9LRJvYemiDOdRDAEMe632w8/edit) |
| **GitHub Reference** | [How to Use GitHub.pptx](https://docs.google.com/presentation/d/1XHCbgUOqbcXNUyQ87vTlKdKRgAbBxtkA/edit?usp=sharing&ouid=109735616107417446342&rtpof=true&sd=true) |
| **Power BI tutorial:** | [**PPB - The Ultimate Power BI**](https://docs.google.com/document/d/1Nuo_m1Lpc7Uy2SLutfI1SRMR25LMi2hfhGD7jIMliog/edit?tab=t.0#heading=h.ikp6b6btf6qt) |
| **POWERBI\_MYSQL\_CONN\_ISSUe** | [**POWERBI\_MYSQL\_CONN\_ISSUE**](https://docs.google.com/document/d/1Otn_3g32a9DDZhyVhnUzwWhIi689B6AQ0tHrMNIYmTM/edit?tab=t.0)  [**.NET\_Connector(8.0.32)Error\_Fix**](https://docs.google.com/document/d/1mKrPSBlPtefwvFkFZKv0NzpWjgbK0Eha1vFMMew6HLU/edit?tab=t.0) |
| **Project Orientation (English)** | [Project Orientation Session\_Amazon India: A Decade of Sales Analytics\_AIML-C-WD-E-B20 & DS-C-WD-E-B - 2025/09/04 10:45 GMT+05:30 - Recording](https://drive.google.com/file/d/1oLPb0yEmVAFeyWAnvmDnRXjgh10gb9EJ/view?usp=sharing) |
| **Project Orientation (Tamil)** |  |
| **STREAMLIT RECORDING (English)** | [Special session for STREAMLIT(11/08/2024)](https://docs.google.com/document/d/1aR3pUZFlCi8gicpF6aPHPESeFdOtGMlfob5PckresZk/edit?usp=sharing) |
| **STREAMLIT DOCUMENTATION** | [Install Streamlit](https://docs.streamlit.io/get-started/installation) |
| **HOW TO ESTABLISH SQL CONNECTION:** | [PYTHONSQLCODE\_TAMIL.ipynb](https://drive.google.com/file/d/1ZdgP0OgpxamIKJzAPspZkrExD0JvZb-A/view?usp=sharing) |
| **Project Excellence Series [Python + SQL] (English)** | [**Project Excellence Series: Guided Learning & Problem Solving [Python & SQL](English)**](https://docs.google.com/document/d/1hj9s5C9mylUEQR7E9kcfuKyICr3sQ9T7YDuE3LnRZYQ/edit?usp=sharing) |
| **Project Excellence Series [Python + SQL] (Tamil)** | [**Project Excellence Series: Guided Learning & Problem Solving [Python & SQL](Tamil)**](https://docs.google.com/document/d/1MiStEEh6KHpoCvdpD0yTtNpphEbp02j4SVbQrGBQqx8/edit?usp=sharing) |
| **Project Excellence Series [EDA] (English)** | [**Project Excellence Series: Guided Learning & Problem Solving [EDA](English)**](https://docs.google.com/document/d/1kbAZAtGhABWK5S7IAWXs3qiQ8PZybqCpv67S0DwppJ0/edit?usp=sharing) |
| **Project Excellence Series [EDA] (Tamil)** | [**Project Excellence Series: Guided Learning & Problem Solving [EDA](Tamil)**](https://docs.google.com/document/d/120-XJMD8ol-WBt3GaoSsdkdQD--i-TcS5t3N4Ur9LhQ/edit?usp=sharing) |

| **Created By:** | **Verified By:** | **Approved By:** |
| --- | --- | --- |
| **Subhash Govindharaj** | [**Shadiya P P**](mailto:shadiya@guvi.in) | [**Nehlath Harmain**](mailto:nehlath@guvi.in) |