Coffee Shop Sales Analysis and Prediction

Dataset Details

This project analyzes coffee shop sales data, leveraging datasets from Kaggle. It aims to uncover customer preferences, sales trends, and predictive insights. The datasets include:

- coffee_id.csv: Contains coffee product details, ratings, and roasters.
- coffee_clean.csv: Preprocessed dataset with attributes such as roast types, regions, and coffee types.
- coffee.csv: Comprehensive web-scraped data including reviews, aromas, and other sensory attributes.

Dataset Schema

Coffee_id.csv

- slug: Product identifier (String).
- name: Product name (String).
- roaster: Coffee roaster (String).
- rating: Product rating (Float).
- review_date: Review date (DateTime)

Coffee_clean.csv

- roast types: (E.g., medium-light, medium, dark).
- regions: Africa, South America, etc.
- type attributes: Organic, Fair Trade, Decaffeinated, etc.

coffee.csv

- all_text: Web-scraped descriptions.
- rating: Float value for customer ratings.
- review_date: Date of the review.
- roast: Specific roast profile.
- flavor: Sensory details like aroma and aftertaste.

Phases of the Project

Phase 1: Data Exploration and Quality Check

Objective

- Understand the structure of the datasets and identify data quality issues.

Tasks

1. Review the Dataset Structure:
    - Explore relationships between columns across datasets.
    - Verify unique identifiers (slug, name, review_date).
2. **Identify Data Issues:**
    - Missing values in critical fields like rating or review_date.
    - Duplicate entries in slug or name.
    - Inconsistencies in categorical values (e.g., roast types or regions).

Deliverable

A report detailing dataset schema, potential data issues, and proposed solutions.

Phase 2: Data Cleaning and Transformation

Objective

Prepare a clean and structured dataset for analysis.

Tasks

1. Handle Missing Data:
    - Impute missing ratings using averages by roaster or region.
    - Drop rows with incomplete critical fields if imputation is unfeasible.
2. Resolve Duplicates:
    - Deduplicate based on slug and review_date.
3. Standardize and Enrich:
    - Normalize ratings to a consistent scale.
    - Add derived columns like Year and Month from review_date.
    - Calculate average rating per region and roast type.

Deliverable

A clean, transformed dataset with documentation of the steps performed.

Phase 3: Business Questions and Insights

Objective

Analyze the dataset to answer key business questions.

Tasks

1. Popular Products:
    - Identify top-rated products and most-reviewed roasters.
2. Seasonal Insights:
    - Analyze ratings and reviews by season or year.
3. Regional Performance:
    - Compare average ratings across regions.
4. Roast Preference Analysis:
    - Determine which roast types are most popular and their distribution by region.
5. Correlation Analysis:
    - Examine relationships between sensory attributes (e.g., aroma and flavor).

Deliverable

A report summarizing key insights and answers to the business questions.

Phase 4: Advanced Reporting and Visualization

Objective

Visualize insights and build predictive models for future sales.

Tasks

1. Interactive Dashboard:
    - Showcase product and regional performance using Tableau/Power BI.
    - Highlight trends such as seasonal ratings or roast preferences.
2. Predictive Model:
    - Train a regression model to predict ratings based on attributes like roast type and region.
    - Use decision trees to classify products into popularity tiers.

Deliverable

- Interactive dashboard with filters for regions, roasts, and years.
- Model results showing predicted ratings or trends.

Phase 5: Recommendations and Business Strategy

Objective

Provide actionable strategies based on insights and predictions.

Tasks

1. Product Recommendations:
    - Identify attributes of top-rated products for potential replication.
2. Regional Focus:
    - Suggest regions or roasters to prioritize based on performance.
3. Marketing Campaigns:
    - Target popular roast types and regions for promotions.

Deliverable

A business strategy report detailing actionable recommendations for optimizing sales and product offerings.

Expected Outputs:

1. **Cleaned Dataset**: A final dataset ready for analysis.
2. **SQL or Python Code**: Code used for data cleaning, analysis, and reporting.
3. **Business Insights Report**: A detailed summary of key insights and answers to the business questions.
4. **Tableau/Power BI Dashboard**: An interactive dashboard showcasing key trends, sales patterns, and predictive analysis.
5. **Business Memo**: A memo with actionable recommendations for improving the coffee shop's sales and operations.