Faculty of Engineering
Computer Engineering Department

Phase 1: Project Proposal

Big Data

Semester Project

Presented by

Name	Section	BN
Mohamed Yehia Alahmad	2	14
Moustafa Mohammed Elsayed	2	23
Mostafa Hani Mostafa	2	24
Mennatallah Ahmed Moustafa	2	25

Project Idea – Predicting Product Sales in E-Commerce

This project aims to predict the **weekly sales count** of products on an online shopping website. The system will provide accurate sales forecasts by looking at factors like price, ratings, stock levels, and vendor performance. This will help vendors better manage their inventory and marketing planning.

To get better insights we will consider using our dataset in 2 formats:

1. Dealing with weekly sales count column as continuous value → Regression algorithms
2. Categorize weekly sales count into high, medium, or low-performing → Classification problem

Dataset

The **BaSalam dataset** (products file) contains detailed information on products from the BaSalam online marketplace, including attributes such as product names, prices, sales counts, ratings, stock levels, and vendor details.

The most important features:

- **Numerical**: price, rating_average, stock, vendor_score, preparation days.
- **Categorical**: categoryTitle, isFreeShipping, vendor_status_title.

Dataset link: BaSalam (products) **Size**: 1.4GB

Number of Rows: 2.41M (1.6M after ignoring null rows for all useful columns)

Number of Columns: 43 (around 21 useful features after removing irrelevant columns)

Exploratory Data Analysis (EDA)

1. **Handle Missing Data**: Fill or remove missing values.
2. **Detect Outliers**: Use box plots or Z-scores to find outliers.
3. **Check Correlations**: Identify relationships between features
4. **Dimensionality Reduction**: Apply PCA to simplify the data.
5. **Visualize Data**: Use bar charts, histograms, and scatter plots for better understanding

Predictive Analysis Techniques

1. **Linear Regression** (using MapReduce)
2. **Logistic Regression** (using MapReduce)
3. **Decision Trees, Random Forest**
4. **XGBoost**, **SVM**
5. **K-Nearest Neighbors** (K-NN)

Descriptive Analysis Techniques

1. **K-Means Clustering**: Group products by features like price, stock, sales, and vendor score to identify high-demand products and categories.
2. **Association Rules**: For example: (Low Prices and high Rating) implies (Increased Sales). Products with lower prices and high rates tend to have higher sales, offering insights for pricing other decisions.

Key Insights

#	Insight	Columns Used
1	Analyze the relationship between sales_count_week, stock, and price. -> Determine if high sales and low stock levels lead to price adjustments. -> Determine if price affects sales (Are cheaper products selling more?)	stock, price, sales_count_week
2	Investigate how free shipping affects product sales. -> Compare the sales count of products with and without free shipping to see if it's a major selling point.	isFreeShipping, sales_count_week
3	Analyze product sales by category. Which categories are performing best? Does a particular category consistently outperform others in terms of sales?	categoryTitle, sales_count_week
4	Analyze how the gap between primary price (original price) and new price (discounted price) influences sales performance.	price, primaryPrice, sales_count_week
5	Check how products with variations perform compared to those without. Are customized or multi-option products more popular?	has_variation, sales_count_week
6	Determine which vendors receive the highest customer satisfaction ratings. Are there specific vendors that consistently receive positive feedback from customers?	rating_average, rating_count, vendor_name
7	Investigate whether product weight affects the likelihood of offering free shipping or impacts the total sales (heavier items could be harder to ship).	weight, isFreeShipping
8	Identify the vendors with the highest sales volumes. Are there particular vendors that consistently achieve top sales performance?	vendor_name, sales_count_week
9	Analyze how different free shipping thresholds (vendor_freeShippingToIran and vendor_freeShippingToSameCity) affect sales_count_week.	vendor_freeShippingToIran, vendor_freeShippingToSameCity, sales_count_week
10	Analyze the vendor city influence on sales_count_week (Is being in a certain city makes the vendor perform better?)	vendor_owner_city, sales_count_week
11	Identify the products with the highest sales counts per week to understand what products customers are buying most frequently	sales_count_week, name
12	Analyze how having delivery (has_delivery) affects sales_count_week. Determine if products that offer delivery options experience higher sales	has_delivery, sales_count_week,
13	Analyze how preparation time affects the products sales (Do products with less preparation time perform better?)	preparationDays, sales_count_week
14	Analyze how vendor_has_delivery and vendor_status_title (available vs. not available) affects sales_count_week. Vendors offering delivery might cause increased sales.	vendor_has_delivery, vendor_status_title, sales_count_week