THE STATE UNIVERSITY OF ZANZIBAR (SUZA)

SCHOOL OF COMPUTING, COMMUNICATION, AND MEDIA STUDIES

DEPARTMENT OF COMPUTER SCIENCE AND IT

CS 3202 – DATA MINING AND BUSINESS INTELLIGENCE

ASSIGNMENT

INSTRUCTIONS:

- Organize yourselves into groups of seven students.

- Prepare a demo for presentation of your work.

- Deadline for this assignment is July 16, 2025.

- This assignment takes 20% of your coursework marks.

- Submit the following:

  - SQL scripts or Machine Learning model depending on the question
  - A short report (3-5 pages) containing the findings or outputs of the experiment.

Question One

Conduct sales trend analysis and forecasting using the AdventureWorks (https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver17&tabs=ssms) or any other retail dataset. The following tasks should be accomplished:

(i) Extract monthly sales for the last 2 years.
(ii) Calculate 3-month and 6-month moving averages.
(iii) Determine the month-over-month growth rate.
(iv) Predict sales for the next 2 months based on trend.
(v) Provide a short report on seasonal patterns or anomalies.

Question 2

Follow the link provided below to access customer transaction data available on Kaggle (https://www.kaggle.com/datasets/bhanupratapbiswas/customer-lifetime-value-analytics-case-study?) or any other relevant dataset. Use the data to conduct a customer segmentation and lifetime values analysis. The following tasks should be accomplished:

(i) Segment customers into 4 categories (e.g., Platinum, Gold, Silver, Bronze) based on total purchase amount and frequency.
(ii) Calculate average purchase per segment.
(iii) Identify top 5% high-value customers.
(iv) Propose one marketing action for each segment.

Question 3

Conduct product basket analysis using the dataset available on the following link (https://medium.com/%40samratjain/explained-market-basket-analysis-using-sql-a7434f30e649) . The following tasks should be completed:

(i) Identify at least 10 frequent item pairs using SQL JOINs.
(ii) Calculate support and confidence for each pair.
(iii) Recommend which pairs could be bundled together.
(iv) Visualize the result in a table simulating association rules.

Question 4

Follow the link to access the dataset from (https://www.kaggle.com/datasets/blastchar/telco-customer-churn?) or any appropriate dataset. Use the dataset to perform customer churn risk analysis. The following tasks should be completed:

(v) Define churn as number of purchase in the last 6 months.
(vi) Create a list of churned and active customers.
(vii) Compare churned customers with active ones in terms of:

- Average purchase value
- Number of purchases
- Last purchase date
- Suggest 3 most suitable features that can be used to predict churn.

Question 5

Use the data available from this link https://databank.worldbank.org/source/world-development-indicators to conduct data preprocessing through the following tasks:

(i) Detect and fix missing values.
(ii) Normalize inconsistent text data (e.g., names, addresses).
(iii) Remove duplicates with a clear query.
(iv) Document 3 types of data quality issues encountered and solved.

Question 6

Use SQL server to define and write queries to support a manager's dashboard with the data available from the following link https://github.com/Microsoft/sql-server-samples/releases/tag/adventureworks. Perform the following tasks:

(i) Top 5 sales regions by revenue
(ii) Total profit this year vs. last year
(iii) Top performing sales representatives
(iv) Average delivery time per region
(v) Customer satisfaction (if available)

Question 7

Using product return and complaints data available from this link https://data.world/shirleywilliams/customer-returns-and-complaints-data to conduct the following:

(i) Identify customers with highest return frequency.
(ii) Analyze which products have the most returns.
(iii) Detect complaint keywords (if textual data exists).
(iv) Suggest improvements in product/service based on insights.

Question 8

Profitability Analysis by Product Line

Use revenue and cost data:

(i) Compute total profit per product and category.
(ii) Identify low-selling but high-margin products.
(iii) Identify high-selling but low-margin products.
(iv) Recommend which products should be promoted or retired.

Question 9

Use the data available from https://data.world/markbradbourne/superstore-sales to do the following tasks:

(i) Identify abnormal transactions (very high or very low).
(ii) Detect purchases made outside business hours.
(iii) Find customers with a suspiciously high number of refunds.
(iv) Flag these anomalies using SQL.

Question 10

Conduct analysis of geographic sales distribution using the data from the following link https://data.world/markbradbourne/superstore-sales. Complete the following tasks based on customer address or location:

(i) Aggregate total sales by region/state/country.
(ii) Identify regions with the highest and lowest sales.
(iii) Compute average order value by region.
(iv) Make a short proposal for regional sales expansion.

Question 11

From a transactional dataset available on this link
https://data.world/markbradbourne/superstore-sales

(i) Create dimension tables (Date, Product, Customer, Store).
(ii) Create a fact table for sales.
(iii) Populate with appropriate sample data from source tables.
(iv) Write queries to perform:
(v) Monthly sales
(vi) Sales by product category and store
(vii) Comparison of sales between two months

Question 12

Track the changes in customer behavior based on the dataset available on this link https://data.world/uci/online-retail-ii-data-set

(i) Extract monthly purchase data per customer.
(ii) Detect changes in spending habits over 6 months.
(iii) Identify dormant customers and newly active customers.
(iv) Suggest 3 customer retention strategies based on findings.

Question 13

Conduct business intelligence using the data available from the following link: https://data.world/markbradbourne/superstore-sales

(i) Write a series of SQL queries to help answer a business decision question such as:
(ii) Should we open a new branch in Region X?
(iii) Is it time to increase the price of Product Y?
(iv) Which customers are ready for upselling?
(v) Justify your decision using extracted insights and KPIs.