Report For Assignment 4:

Introduction:

Our project was made to analyze supermarket product sales containing 200 records using machine learning techniques discussed in class. We made our project using python as the main development language. We implemented a K-means clustering algorithm which segments products based on specific features which include price and sales volume. Furthermore, we applied a regression model which predicted monthly profits, it also compares the performance of linear and polynomial models.

Data Processing:

We created a python script named preprocessing.py which handles the missing values, outliers and scaling problems we ran into.

Missing value handling approach and justification: We found that the columns price and units_sold had missing values, so we decided to take the numerical data approach as it is more robust on outliers.

Outlier detection method and treatment: In order to detect outliers we use the Interquartile range method, for the treatment we found that removing records was not ideal so we decided on capping meaning that any value exceeding the upper bound was set to the upper bound, and values below the lower bound became the new lower bound.

Normalization approach and reason: For the normalization we decided to use the Min-Max Normalization to scale all numerical features to a range. Our reasoning for using this approach is the fact that normalization requires k-clustering because it relies on euclidean distance. Without scaling, large values would be too dominant in the distance calculations over smaller features which can lead to bias.

Preprocessing summary statistics: We Successfully imputed 5% of the records, when it came to outliers we were able to successfully cap extreme values price and profit. As far scaling all the features we used included price, cost, units_sold, promotion_frequency, shelf_level.

K-means Clustering Analysis:

Implementation approach: We implemented the K-means algorithm in the kmeans.py without using sklearn, the process works by randomly selecting k data points as initial centroids, next they calculate the euclidean distance from every point to every centroid and assigned points to the nearest cluster.

Elbow method results and optimal k selection: We ran the algorithm from k=2 to k=8 and calculated the within-cluster sum of squares. We can see a heavy drop from k=2 to k=3 following a decrease in the rate, the most optimal k would be 3 for clusters.

Cluster analysis with statistics:

Cluster interpretation and naming: Using k=3 we got

Cluster Name	Avg Price	Avg Units Sold	Avg Profit	Characteristics
Budget Movers (Cluster 0)	$6.24	664	$1,483	Low price, high volume items.
Premium Niche (Cluster 1)	$20.57	125	$1,357	High price, low volume, luxury items.
Star Products (Cluster 2)	$20.31	618	$1,568	High price and high volume (High earners).

Business insights from clustering: The budget movers drive the foot traffic and it is good to keep a high stock of them to prevent losing profit, the premium niche has high margins but low frequencies, the goal would be to market it to the niche group who buys them rather than promoting them everywhere. And for the star products would be to have their own shelf and create special deals for them.

Regression Analysis:
Models chosen and why: We decided to use linear regression chosen as a baseline for the simplicity, and interpretability, polynomial regression chosen to capture non-linear relationships.
Training process: We split the data split into 70% training 30% training sets
Performance comparison table:

Model	MSE (Lower is Better)	MAE (Lower is Better)
Linear Regression	414,986	500.85
Polynomial Regression (Deg 2)	91,896	234.13

Best model selection and justification: The polynomial regression model is far better as it reduced the mean absolute error by over 50% compared to the linear model. The justification for this is given by the fact that profit is calculated by price - cost * units, this is a multiplication problem and linear regression tends to struggle with this a lot more than polynomial regression.

Conclusion:

Key findings: Products naturally group into three categories: budget, premium, and stars, we also found the profit is not linear; the connection between the price and volume is key to predicting profit, and polynomial regression significantly outperforms linear regression in this dataset.

Limitations: The dataset is small, which increases the risk of overfitting, especially for polynomial models.

Potential improvements: We can test other regression models like random forest or Gradient Boosting

AI tool usage summary: We utilized different methods and different AI agents, Gemini Pro and Open AI and perplexity for coding functions and help solving bugs in the deployment stages.