Project

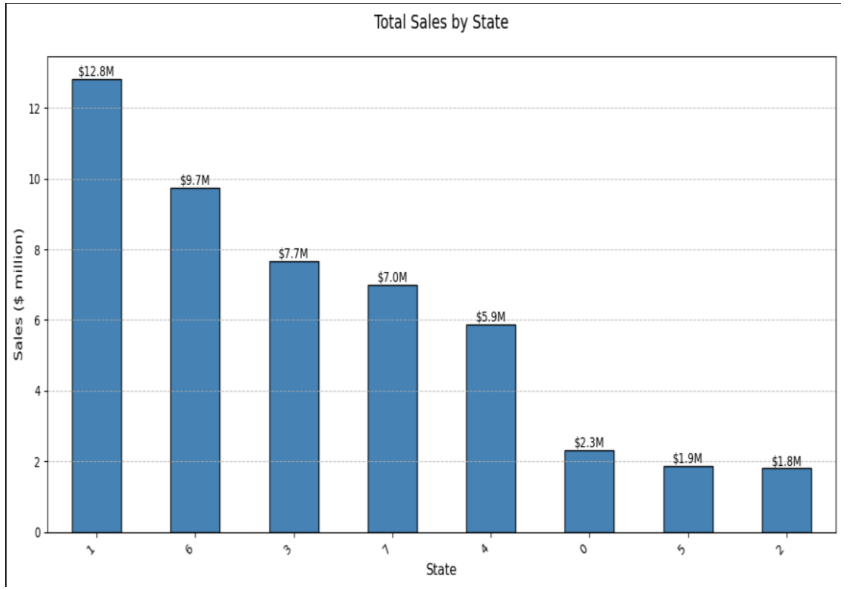
Customers' preferences

Team members :

Ahmedm Hesham ELkady	**2305277**
Arwa Ahmed Saadawy	**2305279**
Nourhan Ibrahim Abdelaleem	**2305440**

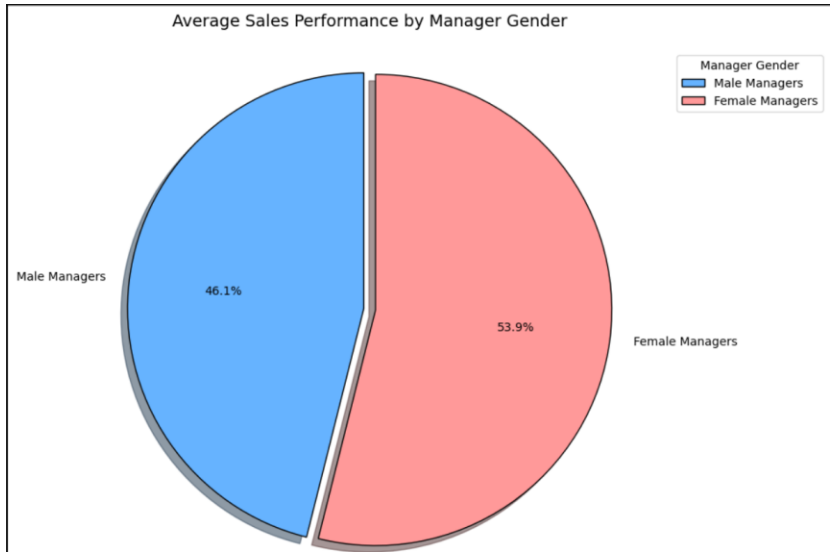
- ***Introduction***
- This project analyzes store sales data to understand customer preferences and purchasing patterns. The dataset contains information about 150 stores, including sales figures, wages, employee counts, locations, and other key variables.
- ***Data Overview***
- **Number of Stores**: 150 stores
- **Number of Variables**: 26 variables including:
  - Sales and wages data
  - Employee and location information
  - Advertising and competitor data
  - Operating hours and home delivery information
- ***Data Preprocessing Steps***
- **Missing Values Check**: No missing values found in the original data
- **Categorical Encoding**: Text variables were converted to numerical values
- **Normalization**: Min-Max Normalization applied to scale data between 0 and 1
- **Clean Data Storage**: Processed data saved in new CSV file (`cleaned_data.csv`)

- ***1. Sales by State***
- Significant variation in sales performance across states
- Top performing states identified for resource allocation
- Opportunities to improve performance
- in lower-performing regions



- ***2. Variable Relationships***
- Explored relationships between:
- Employee count vs sales volume
- Operating hours impact on performance
- Advertising expenditure effects on sales
- ***Conclusions and Recommendations***
- **Sales Distribution**: Wide performance gaps between states indicate opportunities for improvement in underperforming regions
- **Resource Allocation**: Optimize staff distribution and advertising budgets based on regional performance
- **Operating Hours**: Test different operating hours to maximize sales potential
- **Competitive Analysis**: Competitor count analysis can provide better market understanding

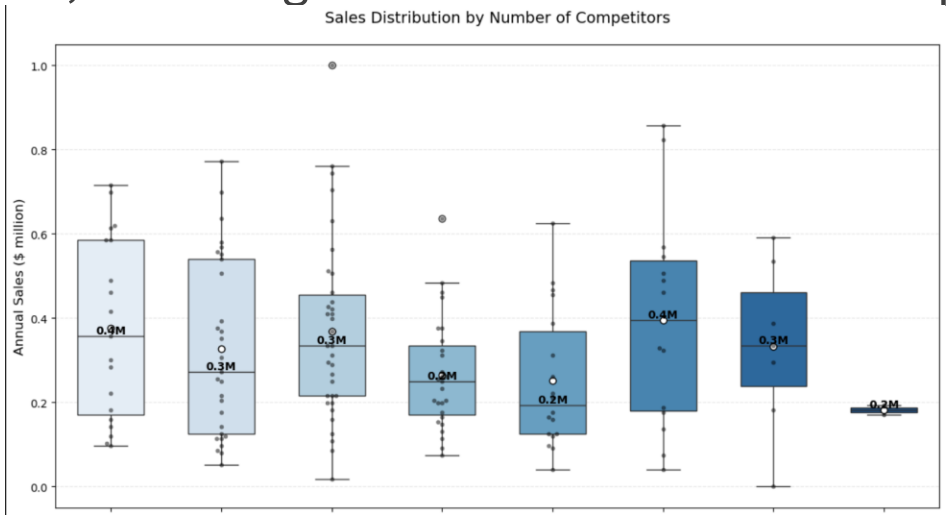
- **Executive Summary:**
  The attached image displays data on average sales performance segmented by manager gender. It shows that male managers account for 46.1% of sales, while female managers achieve 53.9%. These results indicate that female managers outperform their male counterparts in sales performance during the analyzed period.
- **Analysis:**



- **Performance of Female Managers:**
  - Female managers achieved a higher sales share of 53.9%, which may reflect their effectiveness in team leadership or sales strategy execution.
  - Further investigation into the contributing factors (e.g., leadership style, customer engagement) is recommended.
- **Performance of Male Managers:**
  - Male managers accounted for 46.1% of sales, a lower share compared to female managers.
  - Potential challenges (e.g., management approaches, resource allocation) should be examined to identify areas for improvement.

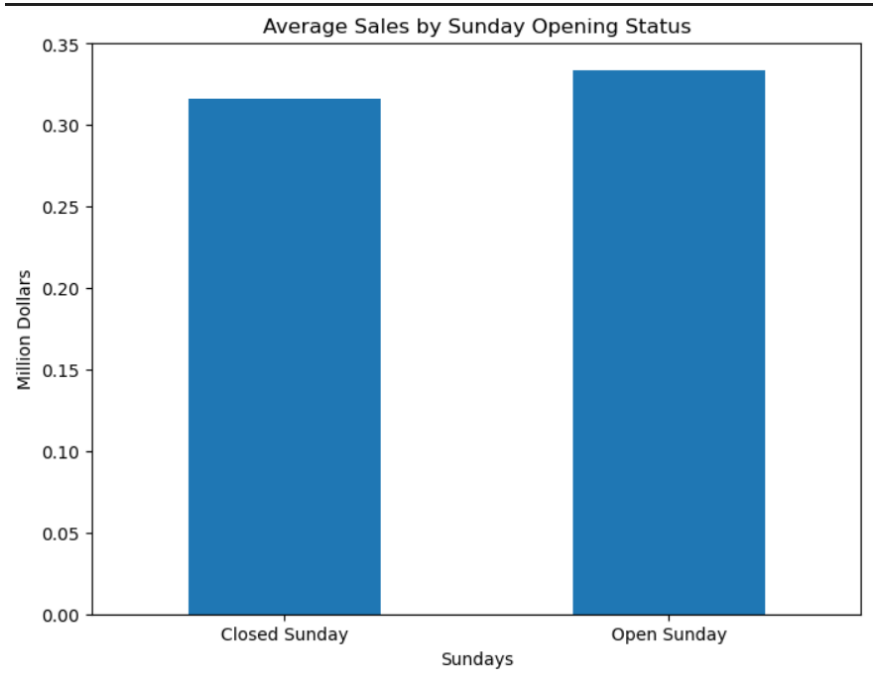
The graph illustrates annual sales (in $ millions) across different competitive environments. Key observations:

- Peak sales reached **$0.9M** in one segment, suggesting high potential in specific market conditions.
- Sales dropped to **$0.0M** in others, indicating possible market saturation or ineffective strategies.
- Most sales clustered between 0.2M–0.2*M*–**0.5M**, reflecting moderate but inconsistent performance.



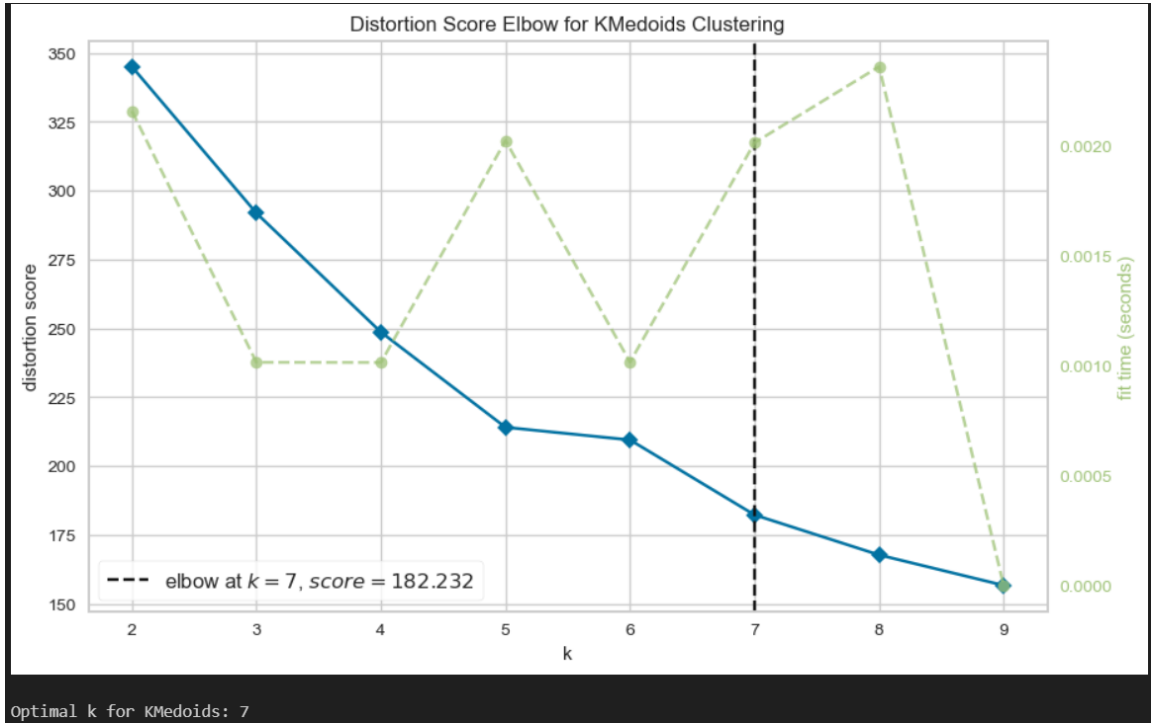
- **High-Performance Segment ($0.9M):**
  - Likely driven by factors like low competition, unique offerings, or superior customer targeting.
  - *Recommendation:* Reverse-engineer this success for replication in similar markets.
- **Zero-Sales Segments ($0.0M):**
  - May result from excessive competitors, pricing issues, or lack of differentiation.
  - *Recommendation:* Conduct competitive analysis and revise market-entry tactics.
- **Volatility (**0.2M–0.2*M*–**0.5M):**
  - Fluctuations suggest sensitivity to competitor actions or demand shifts.
  - *Recommendation:* Stabilize performance through loyalty programs or operational efficiencies.

- The data compares average sales performance between stores that operate on Sundays versus those that remain closed. Results show:
- **Open Sunday:** Achieved higher average sales (approximately **$0.30M**).
- **Closed Sunday:** Recorded significantly lower sales (around **$0.15M**).
- **Analysis**



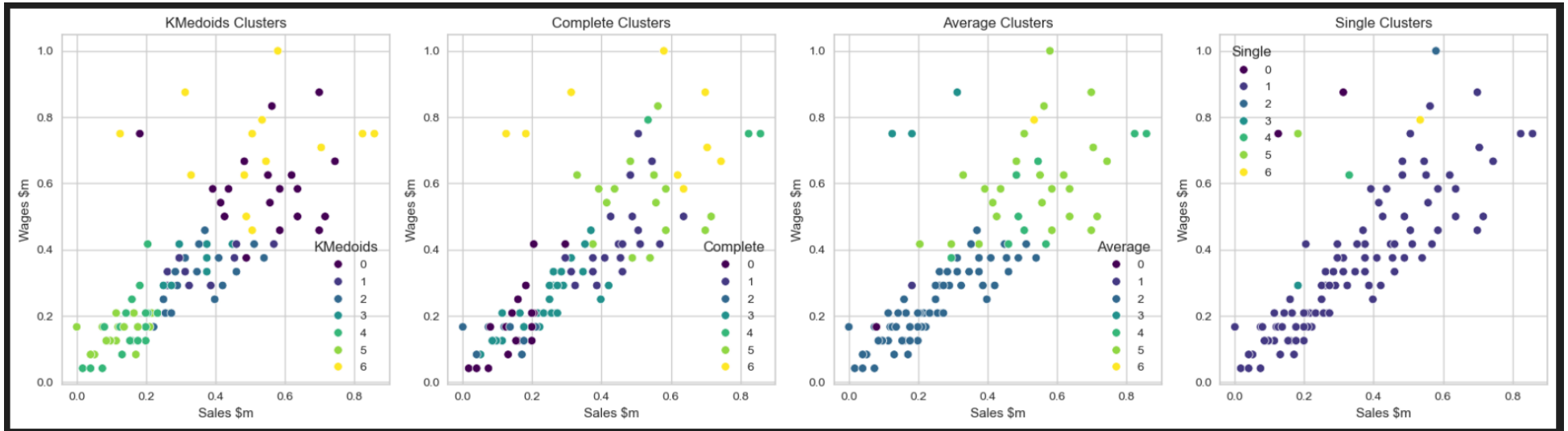
- **Revenue Opportunity:**
  - Stores open on Sundays generate **double** the sales of closed locations, highlighting Sunday as a high-potential sales day.
  - Possible drivers: Weekend shopping trends, relaxed consumer schedules, or strategic promotions.
- **Closed-Sunday Challenges:**
  - The $0.15M average suggests missed revenue from Sunday demand.
  - Potential causes: Local regulations, staffing constraints, or cultural norms affecting operations.

- The elbow method analysis for K-Medoids clustering indicates the optimal number of clusters (k) is 7, with a distortion score of 182.232 at this point.
- **Technical Analysis:**



- **Elbow Method Visualization:**
  - The plot shows the relationship between number of clusters (k) and distortion score
  - The "elbow" (point of maximum curvature) occurs at k=7, representing the optimal trade-off between model complexity and performance
- **Interpretation of k=7:**
  - This suggests the dataset naturally divides into 7 distinct groups
  - The distortion score of 182.232 represents the sum of squared distances from each point to its assigned medoid

- ***Clustering Results Analysis Using Different Algorithms***
- The image presents a comparison of four clustering algorithms applied to a dataset containing two features: **Sales ($m)** and **Wages ($m)**. The scatter plots visualize how each algorithm grouped the data points, while the evaluation table below summarizes the performance metrics for each method:
- **Algorithms Compared**:
- **KMedoids**
- **Complete Linkage**
- **Average Linkage**
- **Single Linkage**
- Each clustering result is visualized in a subplot, where data points are colored based on their assigned cluster.
- ***Evaluation Metrics:***



Clustering Evaluation Metrics:

Method	Silhouette	Calinski-Harabasz	Davies-Bouldin	Accuracy	Recall	F1-Score
KMedoids	0.166938	30.7221	1.26517	0.2	0.2	0.315488
Complete Linkage	0.250349	33.8817	1.20369	0.238095	0.238095	0.353239
Average Linkage	0.214363	18.0348	0.819592	0.00952381	0.00952381	0.0188609
Single Linkage	0.0554088	3.64166	0.624248	0.0571429	0.0571429	0.0403434

- **Silhouette Score**: Measures how similar a point is to its own cluster compared to other clusters. Higher values indicate better-defined clusters.
- **Calinski-Harabasz Index**: Evaluates cluster separation. Higher values suggest better performance.
- **Davies-Bouldin Index**: Measures cluster overlap. Lower values are better.
- **Accuracy, Recall, F1-Score**: These are used when ground truth labels are available to assess clustering quality relative to true classes.
- **Comparison Insights**:
- **Complete Linkage** achieved the highest **Silhouette** and **Calinski-Harabasz** scores, indicating well-separated and compact clusters.
- **KMedoids** provided moderate performance across all metrics and showed a good balance between accuracy and intra-cluster cohesion.
- **Average Linkage** showed relatively weaker performance, particularly in **Accuracy** and **F1-Score**.
- **Single Linkage** had the lowest scores across all metrics, suggesting poor cluster formation and overlapping clusters.
- **Conclusion**:
- Among the four methods, **Complete Linkage** is the most effective clustering algorithm for this dataset, producing the most distinct and coherent groups. On the other hand, **Single Linkage** performed the worst and is not recommended for similar data structures.

- This section compares the classification performance of two supervised learning algorithms: **Decision Tree** and **Random Forest**, based on a dataset with three target classes: **Low**, **Medium**, and **High**.

• 1. Decision Tree Results

- **Accuracy**: 0.87
- **Recall (Weighted)**: 0.87
- **F1-Score (Weighted)**: 0.86
- **Analysis**:
- The model performs very well on the **Low** and **Medium** classes.
- However, the **High** class is not correctly classified at all — with one misclassified as Medium.
- Despite this, the overall accuracy and F1-score are relatively high, showing that the model generalizes well for most categories

• 2. Random Forest Results

- **Accuracy**: 0.83
- **Recall (Weighted)**: 0.83
- **F1-Score (Weighted)**: 0.82
- **Classification Report Summary**:
- **Medium** and **High** classes are well classified with high precision and recall.
- **Low** class has poor performance (precision, recall, and F1-score are all 0.00), likely due to having only one sample in the test set, which introduces imbalanc
- **Analysis**:
- Random Forest performs slightly worse overall than Decision Tree in terms of accuracy and F1-score.
- However, it does a better job classifying the **High** class (15 correct out of 16).
- The model struggles with the **Low** class, which negatively affects the macro average metrics.

- **Conclusion**
- **Decision Tree** achieved higher overall accuracy and balanced performance across **Low** and **Medium** classes.
- **Random Forest** showed stronger performance on the **High** class but failed to classify the **Low** class due to data imbalance.
- For balanced classification tasks, Decision Tree might be preferable in this specific case. However, Random Forest could be better with more balanced data
- ***Why This Project?***
- ✓ **Business Relevance:** Directly impacts revenue and customer satisfaction.
  ✓ **Real-world Application:** Used by retail giants (Amazon, Walmart).
  ✓ **Technical Skills:** Implements clustering, classification, and visualization
- Scan QR_Code to get source code :