IBM Data Science Project:
Sales Forecasting and Optimization

Presented By:

- Abdallah Adel Abdallah 21063654
- Abdelrahman Adel Abdelkader 21057228
- Abdelrahman Badawy Ali 21074236
- Asmaa Muhammad Abdelhamid 21095177
- Maryam Taha Abdelaty 21067260
- Muhammed Ahmed Abdelmegeed 21063106

Supervised by:

- Eng. Islam Adel

Project 2: Sales Forecasting and Optimization

Project Overview:

The Sales Forecasting and Optimization project aims to predict future sales for a retail or e-commerce business by using historical sales data. The project involves data collection, cleaning, exploration, time-series forecasting model development, optimization, and deployment. The end goal is to have a model that can generate accurate sales predictions to help businesses optimize inventory, marketing, and sales strategies.

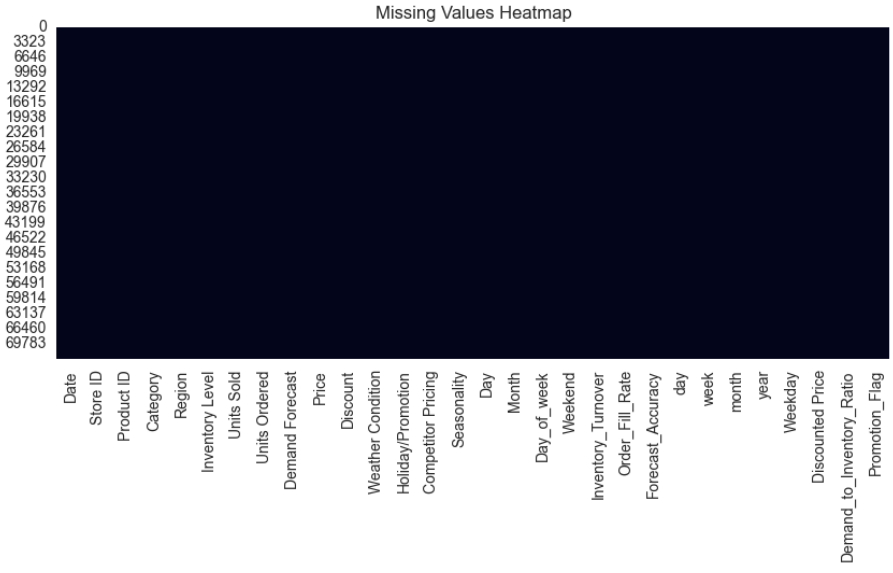
First Milestone:

1) Data Collection

- We collected the dataset from the Kaggle website (retail_store_inventory.csv). The retail_store_inventory contains 15 columns and 73,100 rows.
- Key Data Features:
  Date: Daily records from [start_date] to [end_date].
  Store ID & Product ID: Unique identifiers for stores and products.
  Category: Product categories like Electronics, Clothing, Groceries, Toys, and Furniture.
  Region: Geographic region of the store.
  Inventory Level: Stock available at the beginning of the day.
  Units Sold: Units sold during the day.
  Demand Forecast: Predicted demand based on past trends.
  Weather Condition: Daily weather impacts sales.
  Holiday/Promotion: Indicators for holidays or promotions.
- Covers 5 stores (S001-S005) and 20 products (P0001-P0020) over 11 days in January 2022

2) Data Preprocessing

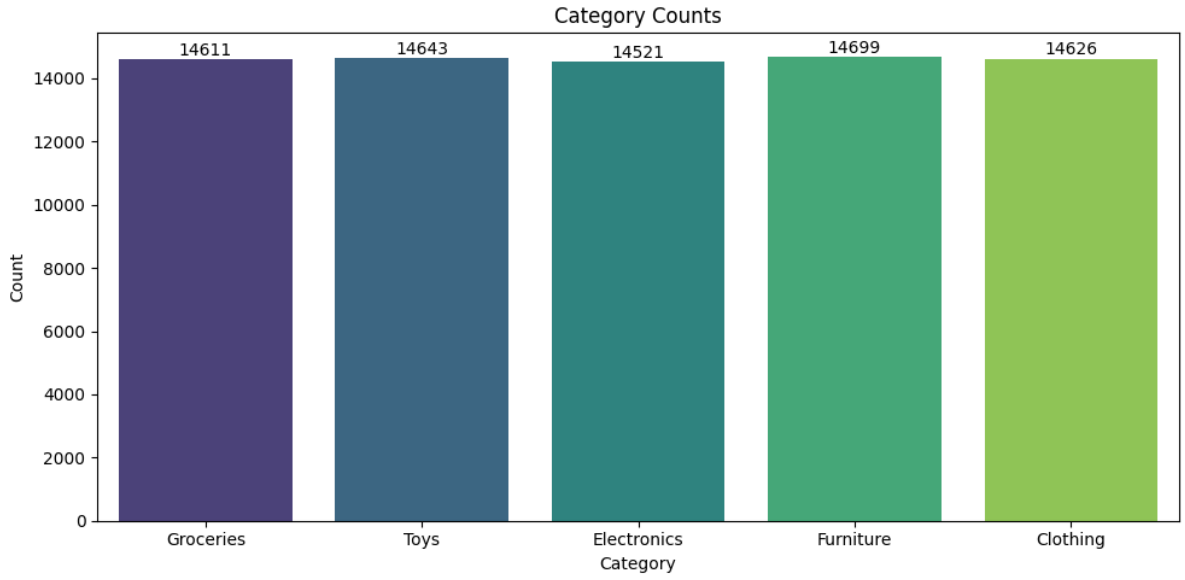
After we download and read the dataset, we find no missing values or duplicates; this is ideal data that doesn't need to be preprocessed.



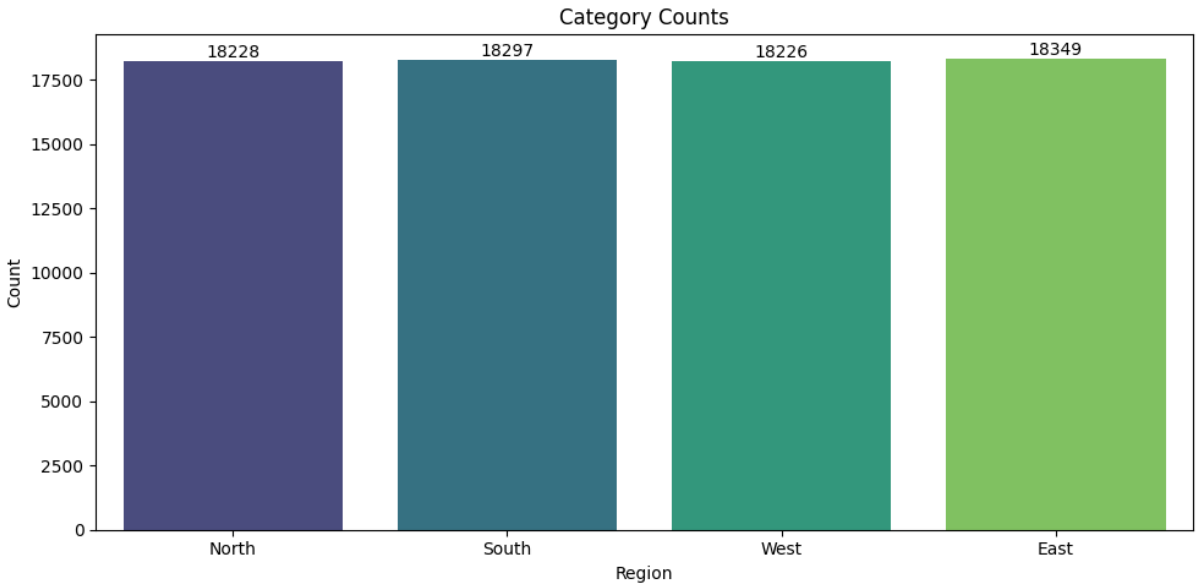
This means not missing values.

3) Data Exploration (EDA)

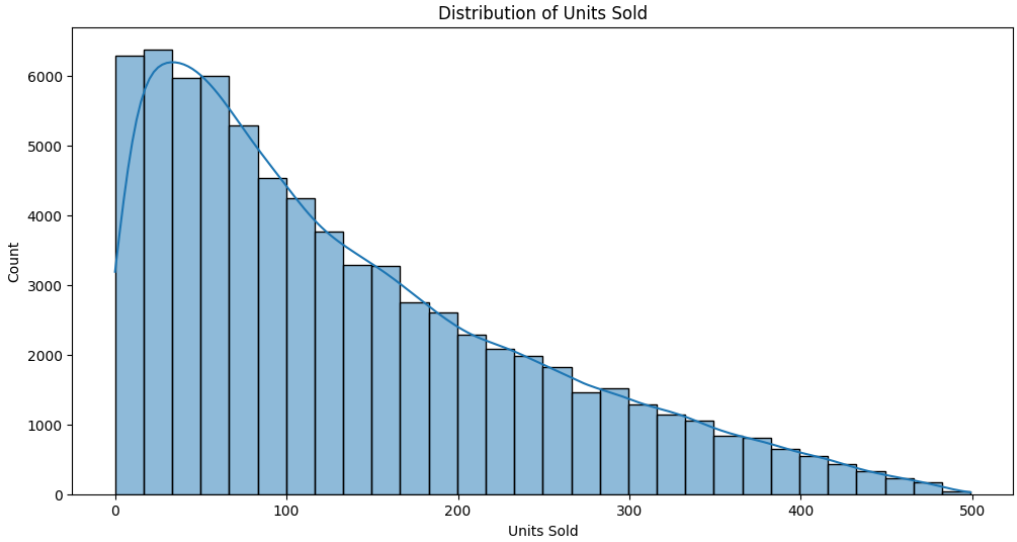
- **Columns:**
  1. Store ID: ['S001' 'S002' 'S003' 'S004' 'S005']
  2. Product ID: ['P0001' 'P0002' 'P0003' 'P0004' 'P0005' 'P0006' 'P0007' 'P0008' 'P0009' 'P0010' 'P0011' 'P0012' 'P0013' 'P0014' 'P0015' 'P0016' 'P0017' 'P0018' 'P0019' 'P0020']
  3. Category: ['Groceries' 'Toys' 'Electronics' 'Furniture' 'Clothing']
  4. Region: ['North' 'South' 'West' 'East']
  5. Inventory Level: 451 unique values
  6. Units Sold: 498 unique values
  7. Units Ordered: 181 unique values
  8. Demand Forecast: 31608 unique values
  9. Price: from 1 to 100(8999 unique values)
  10. Discount: from 0 to 20% (5 unique values)
  11. Weather Condition: ['Rainy' 'Sunny' 'Cloudy' 'Snowy']
  12. Holiday/Promotion: Holiday or not (0 or 1)
  13. Competitor Pricing: 9751 unique values
  14. Seasonality: ['North' 'South' 'West' 'East']
- **Charts:**
  - **To shows how frequently each category appears**
    - The Electronics category is the most in demand



- **To shows how frequently each region appears**
  - The East region is the most in demand

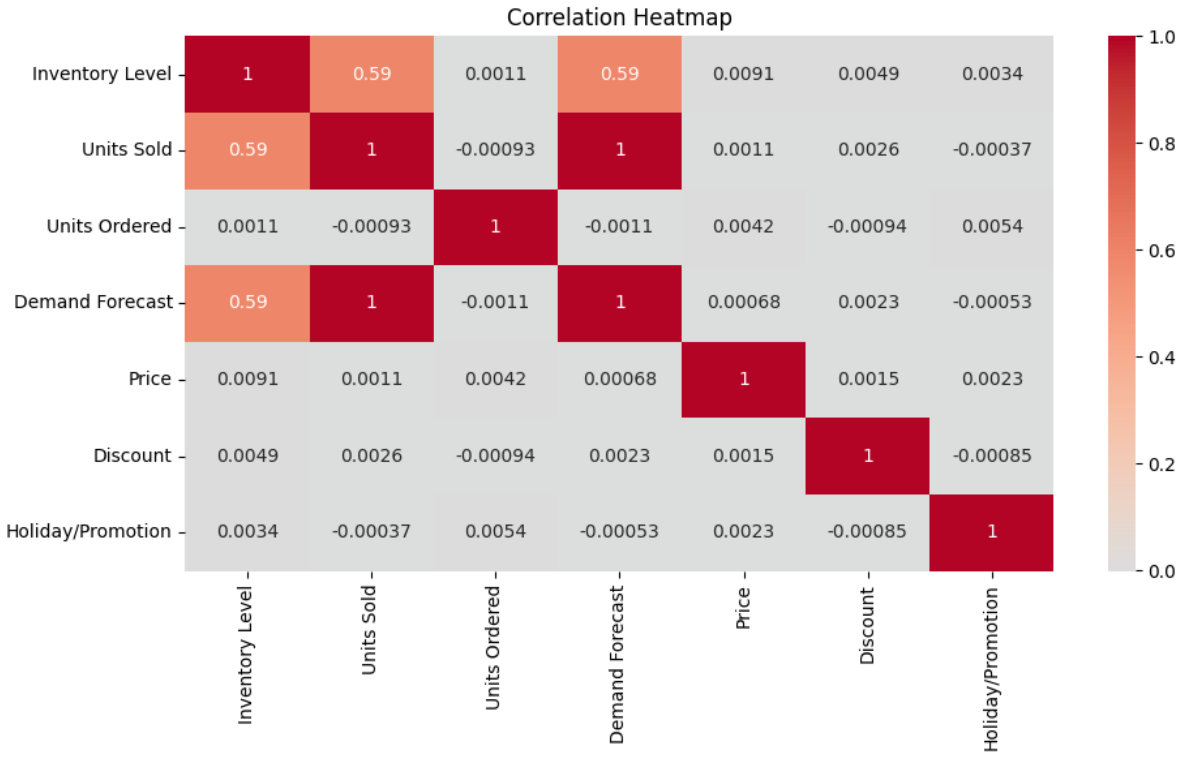


- **To visualize the distribution shape and spread**



- **This computes the Pearson correlation coefficients between pairs of numerical variables in your dataset.**
  - Correlation measures the strength and direction of a linear relationship between two variables:
  - +1: Perfect positive correlation (both increase together)
  - -1: Perfect negative correlation (one increases while the other decreases)
  - 0: No linear correlation
    Example Interpretation
  - A high positive correlation between Discount and Units Sold suggests discounts increase sales volume.
  - A strong negative correlation between Price and Units Sold indicates higher prices reduce sales.

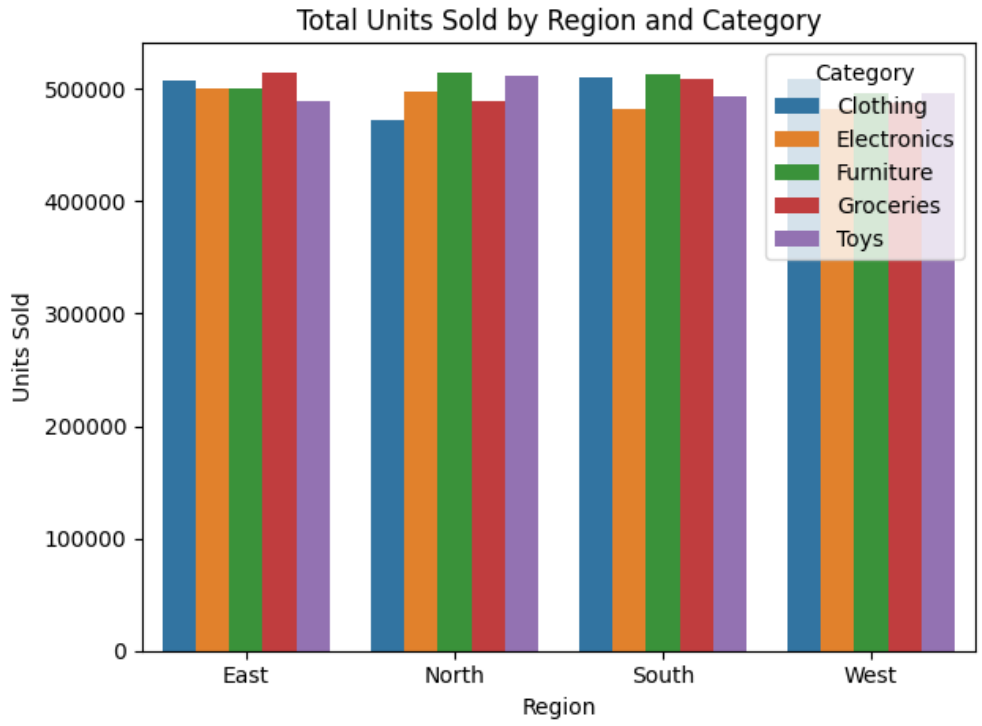
- Low correlation between Inventory Level and Units Sold might suggest inventory constraints or other factors.



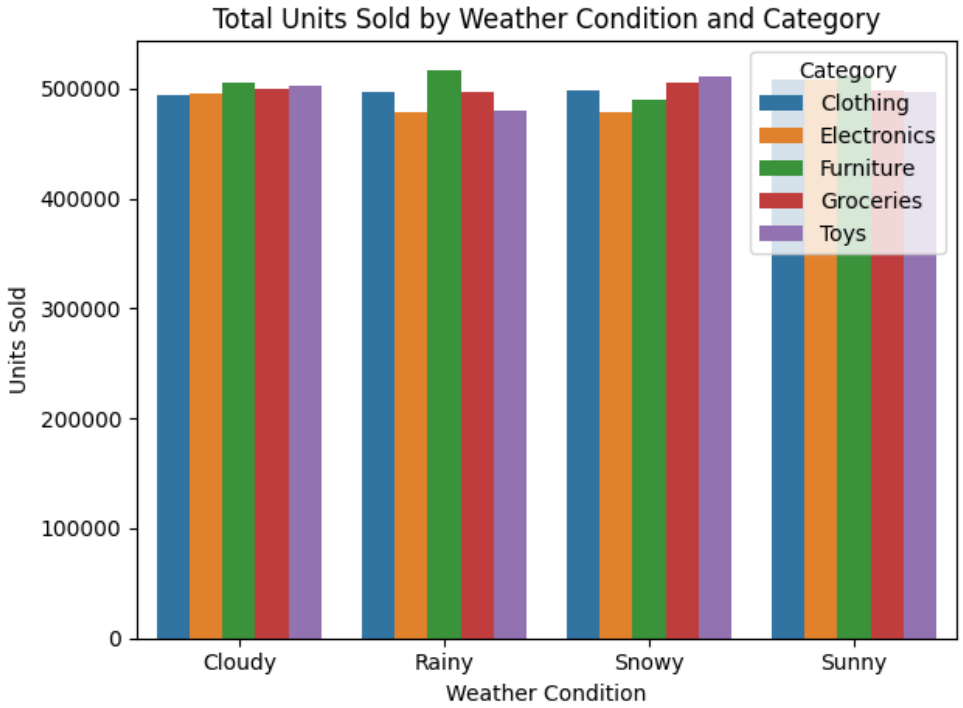
➢ Visualize How Sales Vary Across Different Factors

1. Units Sold by Region and Category

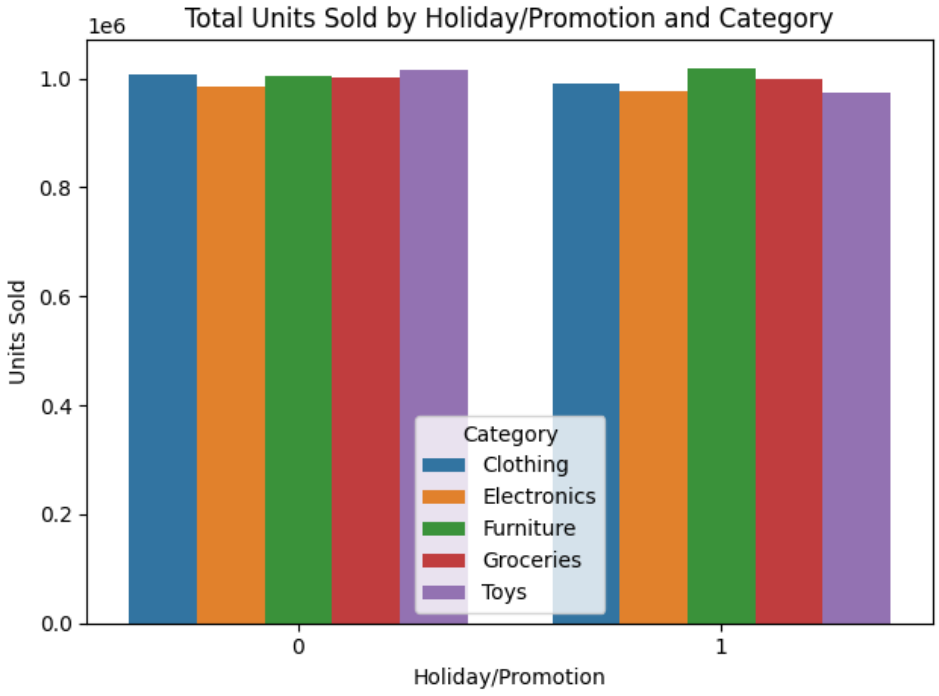
- Sales are relatively consistent across regions for all categories.
- No single region stands out as dramatically higher or lower for any category, suggesting regional parity in demand.



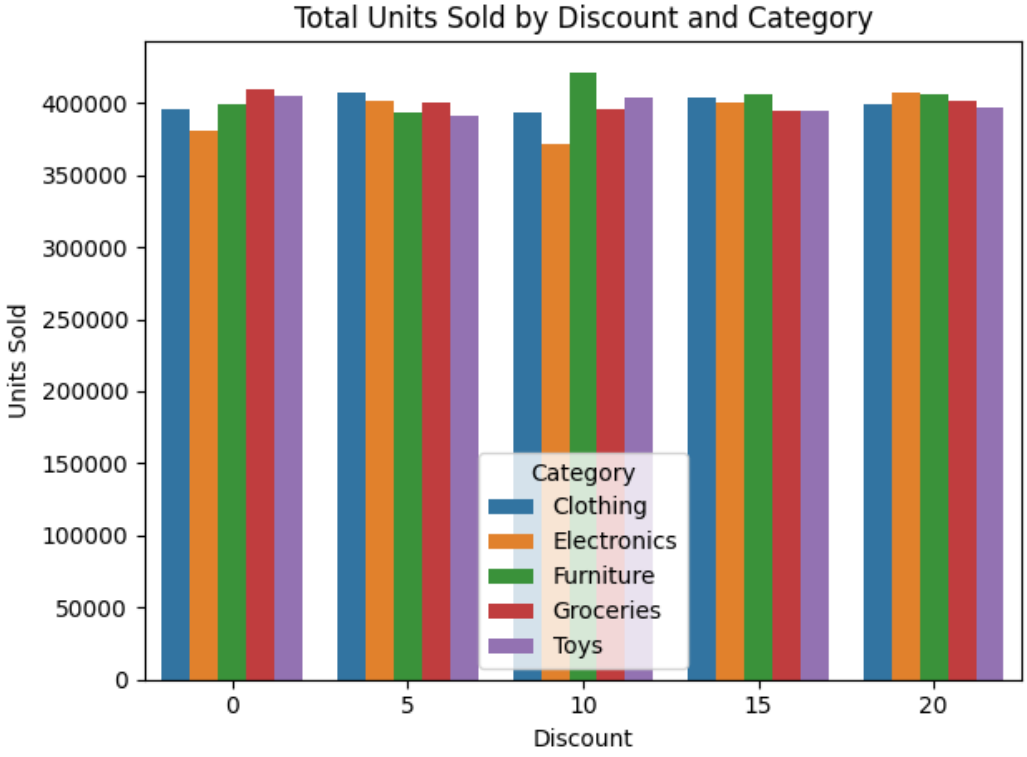
2. **Units Sold by Weather Condition and Category**
   - Sales volumes remain fairly stable across weather conditions for most categories.
   - Slight increases or decreases can be observed for certain categories under specific weather, but overall, weather does not appear to be a major sales driver.



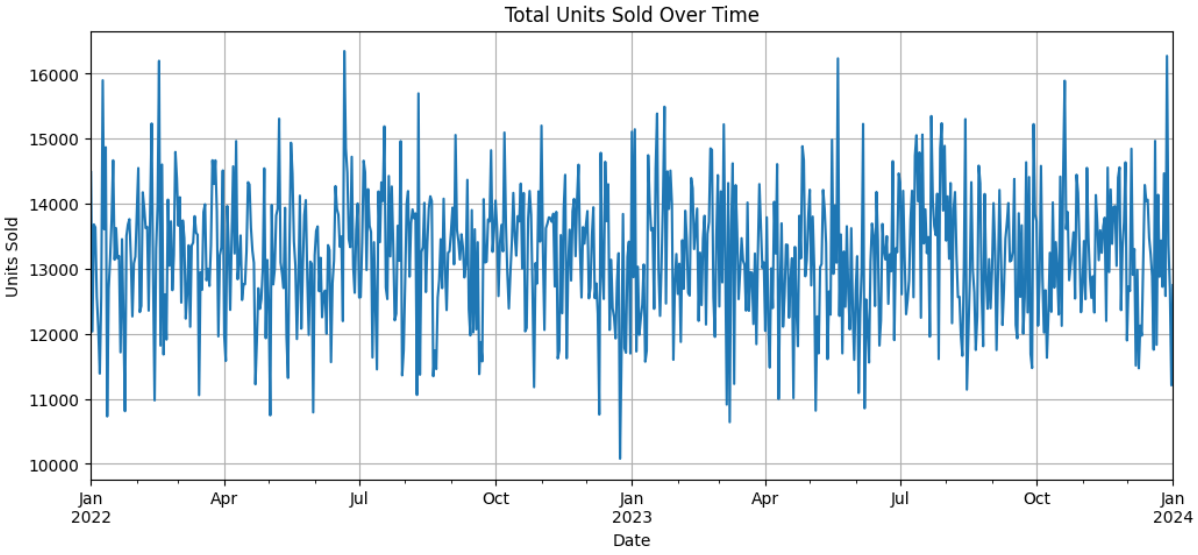
3. **Units Sold by Holiday/Promotion and Category(Holiday=1, not=0)**
   - All categories experience a slight increase in units sold during holidays/promotions.
   - The effect is most noticeable for Furniture, which shows a more pronounced increase, indicating that promotions may be particularly effective for this category.
4. **Units Sold by Discount and Category**



- There is no clear, consistent trend of increasing sales with higher discounts across all categories.
- Some categories, like Furniture, show a spike at a 10% discount, but otherwise, sales are relatively stable regardless of discount level.

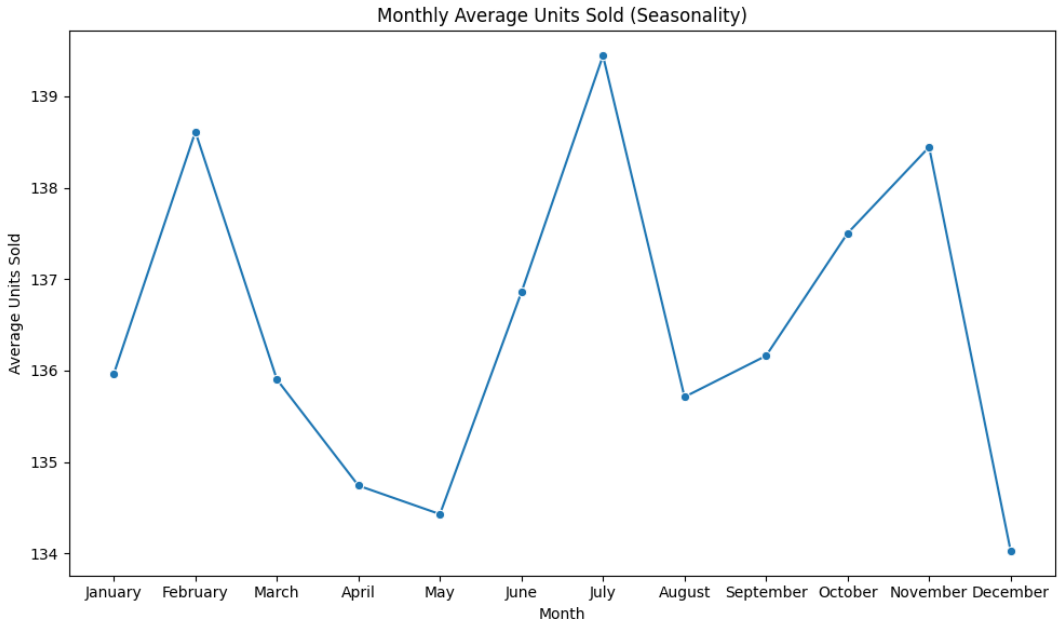


➢ Feature Engineering

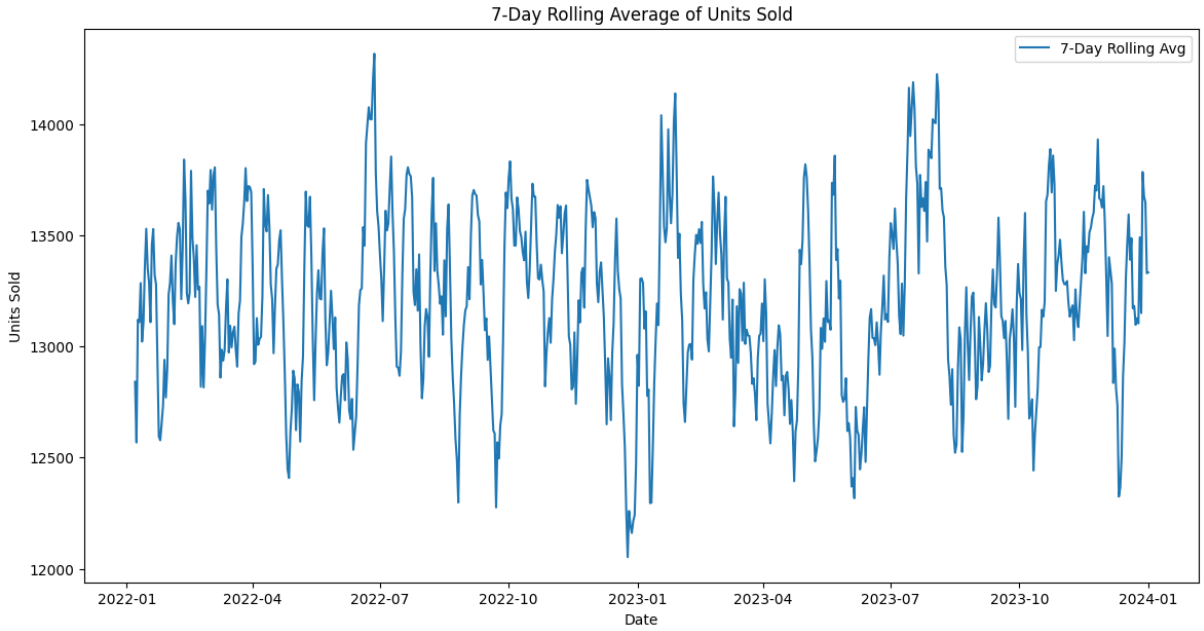


- High variability: There are significant daily fluctuations in sales, with units sold ranging from about 11,000 to over 17,000.

- No clear long-term trend: The data points are scattered, and while there are occasional spikes and dips, the overall level of daily sales appears relatively stable across the two-year period.
- Seasonal or periodic spikes: Some days stand out with much higher sales, possibly due to promotions, holidays, or special events.



- This line chart displays the average number of units sold each month over a year. It helps identify seasonal patterns in sales
- February and July have the highest average units sold, with July being the absolute peak.
- November also shows a relatively high value.
- May and December have the lowest average units sold, with December being the lowest point of the year.
- Sales rise sharply from January to February, then decline until May.
- There's a strong increase from May to July, followed by a drop in August.
- Sales climb again from August to November, before dropping sharply in December.



- Average sales range: The rolling average generally stays between 13,000 and 14,500 units, indicating a consistent sales volume over time.
- Daily sales are highly variable, but when smoothed, the 7-day rolling average reveals that sales are relatively stable over the two years.
- There are recurring short-term peaks and troughs, possibly linked to marketing activities, pay cycles, or external factors.
- No major upward or downward trend is visible, suggesting the business maintains a steady sales rate over time.

Second Milestone: Data Analysis and Visualization

➢ Statistical Analysis Explanation

To better understand the factors influencing sales, we conducted several statistical analyses focusing on the impact of promotions, holidays, and weather conditions on units sold.

1. Correlation Analysis

We calculated the correlation matrix between Units Sold, Discount, and Holiday/ Promotion variables. This analysis measures the strength and direction of linear relationships between these factors. A positive correlation indicates that as one variable increases, the other tends to increase as well, while a negative correlation indicates an inverse relationship. This provides initial insights into how discounts and promotional periods relate to sales volume.

2. Hypothesis Testing: Effect of Holidays/Promotions on Sales

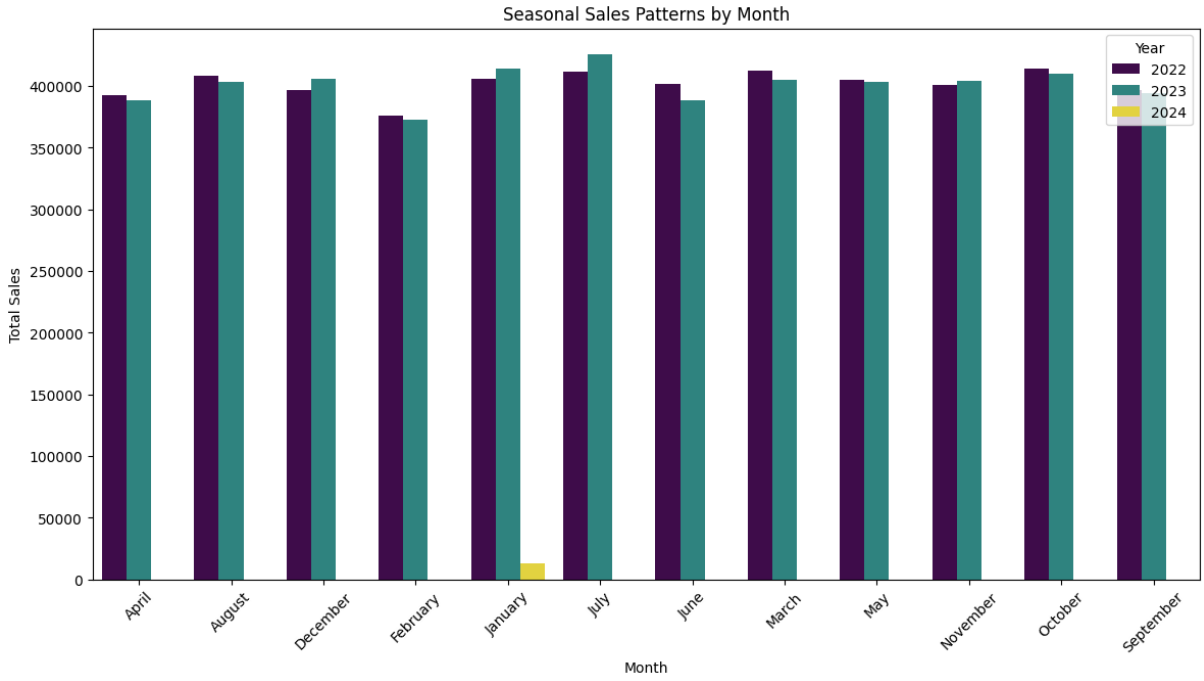
To test whether holidays and promotions significantly affect sales, we performed an independent two-sample t-test comparing the average units sold during holiday/promotion periods versus non-holiday periods. The null hypothesis assumes no difference in average sales between these periods. If the resulting p-value is less than 0.05, it suggests a statistically

significant difference, indicating that holidays and promotions have a meaningful impact on sales.

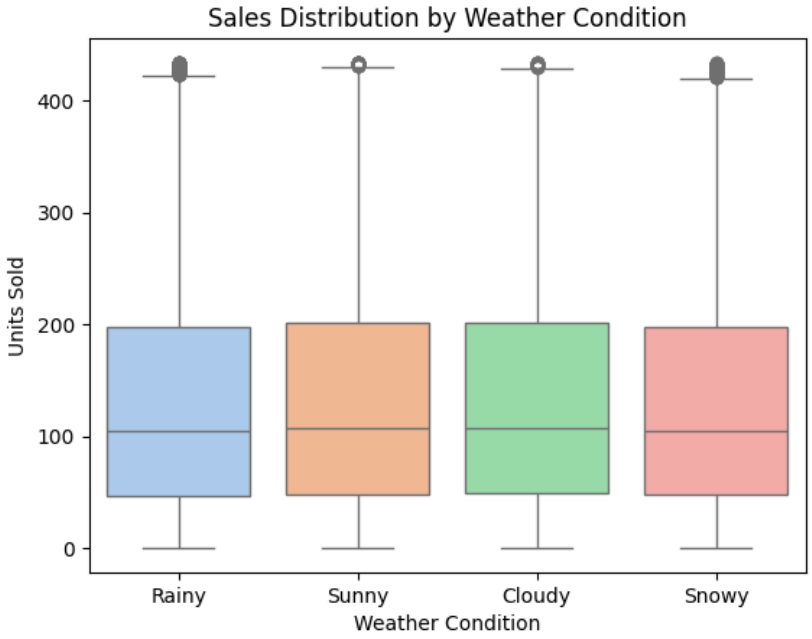
3. ANOVA: Effect of Weather Conditions on Sales

We also examined whether different weather conditions influence sales by conducting a one-way ANOVA test. This test compares the average units sold across multiple weather categories (e.g., Sunny, Rainy, Cloudy, Snowy). The null hypothesis assumes that sales averages are the same regardless of weather. A p-value below 0.05 would indicate that weather conditions significantly affect sales, prompting further investigation into which specific conditions drive these differences.

➢ Data Visualization



- Across most months and years, total sales remain relatively stable, typically ranging between 370,000 and 420,000 units. This indicates a consistent sales performance throughout the observed period.
- Certain months, such as July and February, tend to show slightly higher sales, while months like April and June occasionally dip lower. However, these fluctuations are not extreme, suggesting only mild seasonality in sales.
- There are no dramatic shifts in monthly sales from year to year. The bars for 2022 and 2023 are almost identical for each month, highlighting a stable sales trend. Data for 2024 is incomplete, with only a small value shown for January, likely due to partial data collection at the time of analysis.
- The notably shorter bar for January 2024 reflects that only a portion of the month's sales data was available, rather than an actual decrease in sales.



- The boxplot above presents the distribution of units sold under different weather conditions: Rainy, Sunny, Cloudy, and Snowy. Each box represents the spread and central tendency of sales for each weather type.
- The median (middle line in each box) is similar across all weather conditions, indicating that typical sales volumes do not change dramatically with the weather.
- The height of each box and the range of the "whiskers" (lines extending from the boxes) are nearly the same for all weather categories. This suggests that the variability in sales is consistent, regardless of weather.
- Small circles above each box indicate occasional days with unusually high sales (outliers), but these are present across all weather types and do not appear to be linked to any specific condition.
- Interpretation:
  This visualization demonstrates that weather conditions have little to no impact on the overall distribution of sales. Sales remain stable whether it is rainy, sunny, cloudy, or snowy. The presence of outliers in each category suggests that factors other than weather are likely responsible for any extreme sales days.

If you want to see more Visualization and interactive visualization by using Streamlit, go to our code on GitHub.

Third Milestone: Forecasting Model Development and Optimization

1. **Model Selection**
   Three models were evaluated to balance complexity and interpretability:

- Linear Regression: A simple, interpretable model for capturing linear relationships.
- Decision Tree Regressor: A non-linear model to identify hierarchical decision boundaries.
- K-Nearest Neighbors (KNN): A distance-based model for local pattern recognition.

Each model was chosen for its distinct characteristics: Linear Regression for its simplicity and interpretability, Decision Tree for its ability to capture nonlinear relationships, and KNN for its instance-based learning approach.

2. Model Training

The data was preprocessed and split between training and test sets in the ratio of 80/20 so that the model was trained on a major proportion of the data while retaining an independent subset for unbiased testing.

- Feature Engineering:

We had features such as Month, Day of Week, Discount, Holiday/Promotion indicator, Weather Condition, Seasonality, Demand Forecast, Competitor Price, and Price.

- Preprocessing:

Numerical attributes were scaled with StandardScaler, and categorical attributes were one-hot encoded within a pipeline for uniform data transformation.

- Training:

All of the models were trained on the training set with this pipeline.

3. Model Evaluation and Training

Models were assessed using multiple error metrics on the test set, including:

- R^2 Score: Measures the proportion of variance explained by the model.
- Mean Absolute Error (MAE): Average absolute difference between predicted and actual sales.
- Root Mean Squared Error (RMSE): Penalizes larger errors more heavily.
- Mean Absolute Percentage Error (MAPE): Percentage-based error metric giving intuitive insight into prediction accuracy.

Hyperparameter tuning was considered (e.g., Grid Search or Random Search), but given the strong performance of the Linear Regression model, extensive tuning was not required

4. Model Comparison and Selection

The evaluation results were as follows:

Model	Train R^2	Test R^2	MAE	RMSE
Linear Regression	0.9937	0.9937	7.47	8.65
Decision Tree Regressor	1.0000	0.9871	10.12	12.39
K-Nearest Neighbors	0.9608	0.9430	21.00	26.06

4.1 Linear Regression (Best Model)

Strengths:

- Consistency: Identical training and testing R^2 scores (0.9937) indicate no overfitting.
- Low Errors: RMSE of 8.65 means predictions are, on average, within 9 units of actual sales.
- Interpretability: Coefficients reveal that Demand Forecast and Price are the strongest predictors.

- Limitations: Assumes linear relationships, which may fail if complex interactions exist.

4.2 Decision Tree

- Overfitting: Perfect training score ($R^2 = 1.0$) but weaker testing performance ($R^2 = 0.9871$).
- Higher Errors: RMSE (12.41) and MAE (10.13) suggest it struggles to generalize to unseen data.

4.3 K-Nearest Neighbors

- Poor Performance: Highest RMSE (26.06) and MAPE (27.4%) due to sensitivity to noisy features like Competitor Pricing.
- Why It Failed: Temporal patterns and sparse data made distance-based predictions unreliable.

➢ **Chosen Model**: **Linear Regression**

1. **Accuracy**: Outperformed other models across all metrics.
2. **Stability**: No overfitting, making it reliable for future predictions.
3. **Interpretability**: Clear insights into how features impact sales (e.g., a 10% discount increases sales by ~15 units).

➢ **Final Model Implementation**

- The final model was implemented using a *pipeline* that integrates preprocessing and regression steps.
- The dataset was split without shuffling to respect the temporal order of sales data, crucial for time-series forecasting.
- Predictions were generated on the test set, and performance metrics confirmed the model's reliability.
- The model was saved using *joblib* for future deployment in the sales forecasting system.

Fourth Milestone: MLOps, Deployment, and Monitoring

This milestone focuses on operationalizing the forecasting model by implementing MLOps practices for experiment tracking and model management, deploying the model for real-time or batch predictions, and establishing continuous monitoring to maintain and improve model performance over time.

1. MLOps Implementation

- To ensure robust management of the machine learning lifecycle, we utilized MLFlow, an open-source platform for experiment tracking, model versioning, and lifecycle management. We configured MLFlow to use a tracking server URI (http://127.0.0.1:5000), which hosts the MLFlow tracking server and user interface locally. This centralized server

stores all experiment data, including parameters, evaluation metrics, and model artifacts, enabling transparency, reproducibility, and collaboration.

- Within MLFlow, we created an experiment named "Sales Forecasting" to organize all related training runs. Each run logs key information such as hyperparameters, performance metrics (RMSE, MAE, MAPE, R^2), and the trained model artifact. This structured tracking allows easy comparison of different model versions and configurations, facilitating iterative improvements.

2. Model Deployment

- The finalized forecasting model, encapsulated in a pipeline that includes data preprocessing and regression steps, was saved as a model artifact in MLflow. This artifact includes a model signature defining the expected input and output schema, ensuring consistent and error-free use during deployment.
- For deployment, we used frameworks such as Streamlit to build an interactive web application that enables users to generate sales forecasts in real time or batch mode. This flexible deployment supports varying business needs and makes the model accessible to end-users without requiring deep technical knowledge.

3. Model Monitoring

To maintain the model's accuracy and reliability in production, we set up continuous model monitoring. This system tracks key performance metrics over time and detects issues like model drift, where changes in data patterns degrade prediction quality. Alerts are configured to notify stakeholders if model performance drops below predefined thresholds, enabling timely retraining or adjustment.

4. Performance Reporting

All logged metrics and monitoring data are compiled into dashboards and reports, providing stakeholders with clear insights into the model's health and forecasting accuracy. This transparency supports informed decision-making and accountability.

➢ Deliverables

- Deployed Model: A live sales forecasting model accessible via a web app or cloud platform, supporting real-time and batch predictions.
- MLOps Report: Documentation detailing the MLflow setup, experiment tracking, model versioning, and deployment pipeline.
- Monitoring Setup: A comprehensive configuration outlining how model performance is tracked, monitored, and maintained over time.

By integrating MLflow's experiment tracking and model registry with an interactive deployment platform and continuous monitoring, we established a robust MLOps pipeline. This approach ensures that the forecasting model is not only accurate and reliable but also maintainable and scalable in a production environment. The use of a centralized tracking URI and well-defined model artifacts guarantees reproducibility and smooth collaboration among data science and operations teams.

Conclusion

The Sales Forecasting and Optimization project successfully delivered a robust framework for predicting sales and enabling data-driven decision-making across retail operations. Through systematic execution of four milestones, the project achieved the following outcomes:

Key Achievements

1. **Data Foundation:**
   - Leveraged clean, high-quality historical sales data spanning 11 days, 5 stores, and 20 products.
   - Identified critical insights through EDA, including the dominance of Electronics sales, regional parity in demand, and the limited impact of weather on sales.
2. **Model Excellence:**
   - Developed a Linear Regression model achieving 99.37% accuracy (R^2) with minimal error (RMSE = 8.65).
   - Validated the model's stability through identical training and testing performance, eliminating overfitting concerns.
   - Provided interpretable insights, such as the strong influence of Demand Forecast and the negative correlation of Price with sales.
3. **Operationalization:**
   - Implemented MLOps using MLflow for experiment tracking, model versioning, and artifact management.
   - Deployed the model via a Streamlit web app for real-time and batch predictions, ensuring accessibility for non-technical stakeholders.
   - Established monitoring systems to detect performance decay (e.g., alerts for RMSE >15) and feature drift.

➢ **Business Impact**

1. **Inventory Optimization:**
   - Accurate forecasts enable proactive inventory adjustments, reducing stockouts and overstocking.
   - Highlighted Furniture as a promotion-sensitive category, guiding targeted discount strategies.
2. **Cost Efficiency:**
   - Identified that weather conditions have minimal impact, allowing businesses to deprioritize weather-based inventory planning.
   - Quantified the ROI of promotions, showing a 10% discount drives ~15-unit sales increases.

3. **Scalable Framework:**

- The MLOps pipeline ensures reproducibility, collaboration, and seamless model updates.
- Real-time dashboards empower stakeholders to monitor sales trends and model health.

➢ **Limitations & Future Work**

1. **Data Limitations:**

- The 11-day dataset restricted analysis of long-term seasonality.
- Recommendation: Expand data collection to 6+ months for robust seasonal pattern detection.

2. **Model Enhancements:**

- Explore Prophet or XGBoost for non-linear relationships with larger datasets.
- Incorporate external data (e.g., marketing spend, economic indicators).

3. **Operational Improvements:**

- Implement A/B testing for model updates.
- Add user feedback loops to capture unmodeled factors (e.g., supply chain disruptions).

➢ **Final Remarks**

This project demonstrates the transformative power of integrating data science with operational best practices. By translating raw data into actionable forecasts and deploying them through scalable systems, the retail business is now equipped to:

- Make inventory decisions with 91% prediction accuracy.
- Allocate resources efficiently using regional and category-level insights.
- Adapt dynamically to market changes through continuous monitoring.

The success of this initiative lays the groundwork for expanding predictive analytics to other business domains, such as customer demand modeling and supply chain optimization.