Interactive Correlation Visualization for Store Sales Dataset

Abdallah Waked, Diego Bernal Olaya, Xavier Merino Mino

INFO-6151 Data Visualization for Machine Learning

Dr. Magdalene Ravichandran

Apr 3, 2025

Contents

[preprocessing methods 3](#_Toc194522857)

[Technical Classification of Libraries Used 4](#_Toc194522858)

[Chart 1. Distribution of Store Area 5](#_Toc194522859)

[Chart 5](#_Toc194522860)

[Insights and Interpretations 5](#_Toc194522861)

[Analysis and Visualization Techniques 5](#_Toc194522862)

[Machine Learning Techniques Used 5](#_Toc194522863)

[Chart 2. Relationship between Store Area and Daily Customer Count (Scatter Plot) 6](#_Toc194522864)

[Chart 6](#_Toc194522865)

[Insights and Interpretations 6](#_Toc194522866)

[Analysis and Visualization Techniques 6](#_Toc194522867)

[Machine Learning Techniques Used 6](#_Toc194522868)

[Chart 3. Distribution of Items Available 7](#_Toc194522869)

[Chart 7](#_Toc194522870)

[Insights and Interpretations 7](#_Toc194522871)

[Analysis and Visualization Techniques 7](#_Toc194522872)

[Machine Learning Techniques Used 7](#_Toc194522873)

[Chart 4. Relationship between Daily Customer Count and Store Sales 8](#_Toc194522874)

[Chart 8](#_Toc194522875)

[Insights and Interpretations 8](#_Toc194522876)

[Analysis and Visualization Techniques 8](#_Toc194522877)

[Machine Learning Techniques Used 8](#_Toc194522878)

[Chart 5. Distribution of Store Sales 9](#_Toc194522879)

[Chart 9](#_Toc194522880)

[Insights and Interpretations 9](#_Toc194522881)

[Analysis and Visualization Techniques 9](#_Toc194522882)

[Machine Learning Techniques Used 9](#_Toc194522883)

[Chart 6. Average Daily Customer Count by Store Area 10](#_Toc194522884)

[Chart 10](#_Toc194522885)

[Insights and Interpretations 10](#_Toc194522886)

[Analysis and Visualization Techniques 10](#_Toc194522887)

[Machine Learning Techniques Used 10](#_Toc194522888)

[Chart 7. Store Sales Trend Over Time 11](#_Toc194522889)

[Chart 11](#_Toc194522890)

[Insights and Interpretations 12](#_Toc194522891)

[Analysis and Visualization Techniques 12](#_Toc194522892)

[Machine Learning Techniques Used 12](#_Toc194522893)

[Chart 8. Correlation Matrix Heatmap 13](#_Toc194522894)

[Chart 13](#_Toc194522895)

[Insights and Interpretations 14](#_Toc194522896)

[Analysis and Visualization Techniques 14](#_Toc194522897)

[Machine Learning Techniques Used 14](#_Toc194522898)

# preprocessing methods

* Outlier Removal
* Date Conversion and Indexing
* Feature Engineering (Time Index, Rolling Mean)
* Data Normalization (Scaling)
* Numeric Column Filtering

# Technical Classification of Libraries Used

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classification of Libraries | Charts | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Data Analysis |  |  |  |  |  |  |  |  |
| pandas |  |  |  |  |  |  |  |  |
| numpy |  |  |  |  |  |  |  |  |
| |  | | --- | | statsmodels.tsa.seasonal  .seasonal\_decompose |  |  | | --- | |  | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Visualization |  |  |  |  |  |  |  |  |
| plotly |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Interactive Dashboard |  |  |  |  |  |  |  |  |
| streamlit |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Machine Learning Algorithm |  |  |  |  |  |  |  |  |
| |  | | --- | | statsmodels.tsa.arima.model.ARIMA |  |  | | --- | |  | |  |  |  |  |  |  |  |  |
| r2\_score |  |  |  |  |  |  |  |  |
| sklearn.linear\_model.LinearRegression |  |  |  |  |  |  |  |  |
| Evaluation Metrics |  |  |  |  |  |  |  |  |
| sklearn.metrics.mean\_squared\_error |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Data Preprocessing |  |  |  |  |  |  |  |  |
| sklearn.preprocessing.StandardScaler |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

# Chart 1. Distribution of Store Area

* Use ML algorithms to analyse the distribution of store areas.
* Visualize the distribution using a histogram to show the frequency of different store area sizes.

## Chart

## Insights and Interpretations

## Analysis and Visualization Techniques

## Machine Learning Techniques Used

# Chart 2. Relationship between Store Area and Daily Customer Count (Scatter Plot)

* Perform predictive analysis to understand the relationship between store area and daily customer count using ML regression techniques.
* Develop a predictive model to estimate the daily customer count based on the store area.
* Create a scatter plot visualization to depict the relationship between store area and daily customer count.

## Chart

## Insights and Interpretations

## Analysis and Visualization Techniques

## Machine Learning Techniques Used

# Chart 3. Distribution of Items Available

* Aggregate stores by the number of items available using ML techniques.
* Visualize the distribution using a bar chart, where each bar represents a range of items available and the height represents the number of stores within each range.

## Chart

## Insights and Interpretations

## Analysis and Visualization Techniques

## Machine Learning Techniques Used

# Chart 4. Relationship between Daily Customer Count and Store Sales

* Analyse the relationship between daily customer count and store sales using machine learning regression techniques.
* Develop a predictive model to estimate store sales based on the daily customer count.
* Visualize the relationship using a line chart, where daily customer count is on the x-axis and store sales is on the y-axis.

## Chart

## Insights and Interpretations

## Analysis and Visualization Techniques

## Machine Learning Techniques Used

# Chart 5. Distribution of Store Sales

* Analyse the distribution of store sales using machine learning techniques.
* Visualize the distribution using a box plot to show the median, quartiles, and outliers of store sales.

## Chart

## Insights and Interpretations

## Analysis and Visualization Techniques

## Machine Learning Techniques Used

# Chart 6. Average Daily Customer Count by Store Area

* Aggregate the daily customer count by different ranges of store areas using machine learning techniques.
* Visualize the average daily customer count for each range using a bar chart, where each bar represents a range of store areas.

## Chart

## Insights and Interpretations

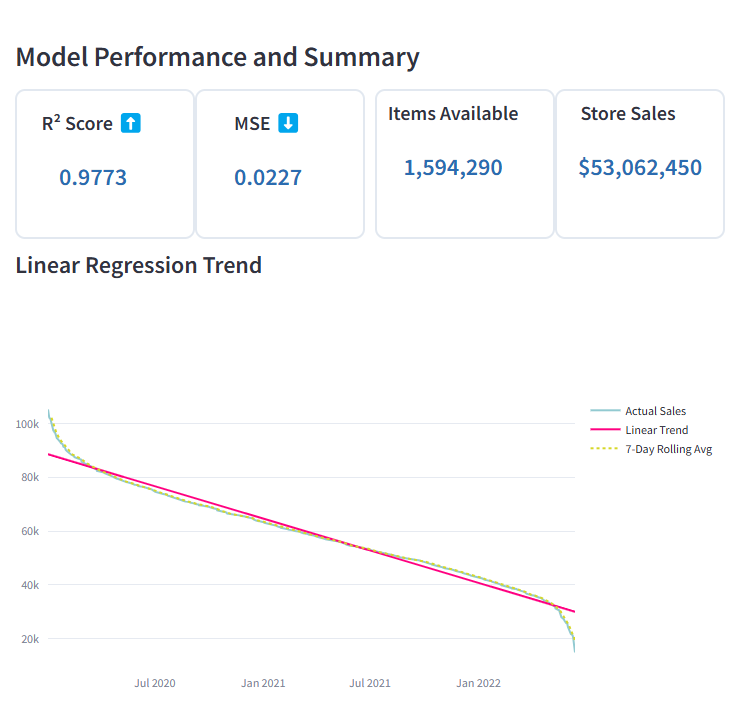
## Analysis and Visualization Techniques

## Machine Learning Techniques Used

# Chart 7. Store Sales Trend Over Time

* Analyse the trend of store sales over time using machine learning time series analysis techniques.
* Visualize the trend using a line chart, where the x-axis represents time (e.g., months or years) and the y-axis represents store sales.

## Chart



## Insights and Interpretations

* The sales trend shows a clear decline over time.
* The R² score (0.9773) and low MSE (0.0227) show that the trendline fits the data very well.
* The decomposition reveals:
  + Trend: Steady sales decrease.
  + Seasonality: Regular monthly patterns.
  + Residual: Random small fluctuations.
* The ARIMA forecast predicts sales will keep dropping in the next 30 days, from $14,321.99 down to $7,852.94.

## Analysis and Visualization Techniques

* Cleaned the data and removed outliers.
* Added a Time\_Index to track sales over time.
* Used Linear Regression to analyze the trend.
* Decomposed the time series to show trend, seasonality, and noise.
* Forecasted future sales using the ARIMA model (1,1,1).
* Visualized everything using interactive line charts and summary cards in Streamlit.

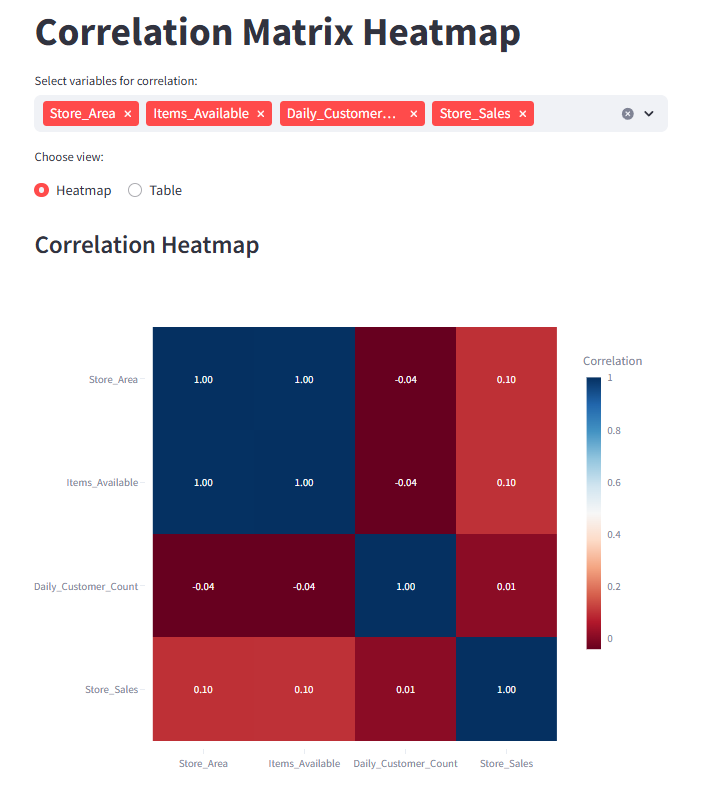
## Machine Learning Techniques Used

* Linear Regression: To show the sales trend.
* ARIMA: To forecast future sales.
* StandardScaler: To normalize the data before applying the model.

# Chart 8. Correlation Matrix Heatmap

* Generate a correlation matrix using machine learning techniques to explore the relationships between different variables (store area, items available, daily customer count, store sales).
* Visualize the correlation matrix using a heatmap, where each cell represents the correlation coefficient between two variables.

## Chart



## Insights and Interpretations

* There is a perfect correlation (1.00) between Store Area and Items Available → These two features are directly related and may be duplicate.
* Store Area and Store Sales have a very weak positive correlation (0.10) → Bigger area does not strongly affect sales.
* Daily Customer Count shows almost no correlation with sales (0.01), suggesting customer count alone does not explain sales.
* Most variables have low or no strong correlation with each other.

## Analysis and Visualization Techniques

* Selected numeric columns for correlation analysis.
* Calculated Pearson correlation between variables using Pandas .corr().
* Built an interactive Streamlit app:
  + Allows users to select variables dynamically.
  + Provides two view options: Heatmap and Table.
* Visualized the correlation matrix using:
  + Plotly heatmap with color gradient to highlight strong/weak relationships.
  + Styled correlation table with color-coded values for better readability.

## Machine Learning Techniques Used

* Pearson Correlation Coefficient: A statistical method used to measure linear relationships between variables.
* Multiselect and Visualization: Although no predictive ML model was used here, the correlation matrix is a core step in ML data analysis to detect feature relationships and multicollinearity.