Geospatial Analysis of Global Superstore Sales

Introduction

This project undertakes a comprehensive geospatial analysis of sales data from a global superstore. By harnessing location data, the objective is to uncover patterns, trends, and regional variations in sales, ultimately providing valuable insights for strategic decision-making.

The global superstore operates in diverse regions, each characterized by unique market dynamics. The spatial distribution of sales holds the key to identifying opportunities for targeted marketing, efficient resource allocation, and strategic expansion. The central question guiding this analysis is: How do sales vary across different geographical locations, and what factors contribute to these variations?

The dataset for this project is sourced from the Kaggle Global Superstore dataset, transformed from its original format to CSV using the pandas library. The processed CSV file serves as the foundation for the subsequent analysis, providing a consolidated repository of sales and order information.

The ‘clean_data.py’ script plays a pivotal role in data preprocessing. It addresses missing values and duplicates, transforming the date column into datetime format. Additionally, it extracts the year and month, preparing the data for insightful analysis.

The ‘run_analysis.py’ script takes center stage in the geospatial analysis, utilizing geographic coordinates to create a Folium map that visually represents the global distribution of superstore sales. Each city is marked with a color-coded indicator reflecting aggregated sales. This script also delves into regional sales analysis, examining total sales, average order values, and profit.

Exploratory Data Analysis

Correlation Matrix

Correlation analysis (Figure 1) further explores potential relationships between geographical locations and sales metrics.

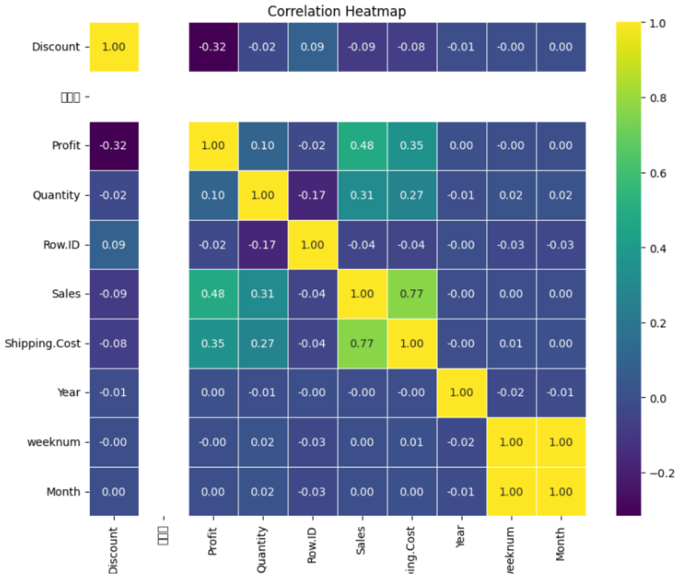


Figure 1: Correlation heatmap to display relationships between geographical locations and sales metrics

Pair Plot

A pair plot visually explores relationships between selected variables in the global superstore dataset, shedding light on potential patterns and correlations. The chosen variables – ‘Ship.Mode’, ‘Quantity’, ‘Profit’, ‘Sales’, and ‘Shipping.Cost’ – offer a comprehensive view of inter-variable dynamics. The use of hue differentiation based on shipping modes enhances the plot’s interpretability, allowing for a nuanced examination of how different shipping methods relate to other key metrics.

Observing the scatterplots and histograms, intriguing patterns emerge. Notably, the ‘Sales’ and ‘Profit’ variables exhibit a positive correlation, suggesting that higher sales volumes are associated with increased profitability. The ‘Quantity’ variable also appears positively correlated with both ‘Sales’ and ‘Profit’, reinforcing the notion that larger quantities contribute to higher sales and profitability. Interestingly, the relationship between shipping cost and the other variables is more nuanced, indicating potential interactions that warrant further investigation.

Additionally, the pair plot effectively captures the distributional characteristics of each variable. This information is crucial for understanding the dataset’s underlying structure and visibility. While scatterplots provide insights into bivariate relationships, the histograms along the diagonal offer a univariate perspective, showcasing the distribution of each variable individually.

The pair plot serves as a valuable exploratory tool, uncovering initial insights into the interplay between key variables in the global superstore dataset. This visualization lays the groundwork for more in-depth analyses and informs subsequent steps in the data exploration process.

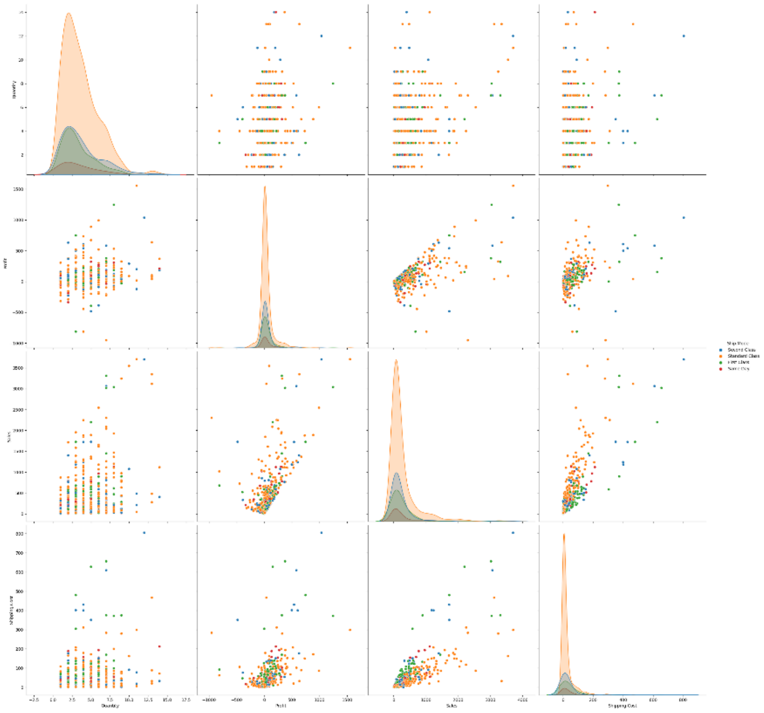


Figure 2: Pair plot for bivariate analysis

Visualizations

Bar Chart

The bar chart depicting the distribution of profits and total sales across different markets provides a comprehensive overview of the financial performance of each market within the global superstore. The x-axis represents individual markets, while the y-axis indicates the respective amounts of profit and sales. The chart distinguishes between profits and sales contributing to the total bar height.

The visualization facilitates the following key observations:

1. Profit Contribution: The upper segment of each bar represents the accumulated profits for each market, allowing stakeholders to quickly identify markets contributing the most to overall productivity. This insight is crucial for prioritizing resources and strategic initiatives in high-profit markets.
2. Sales Comparison The full height of each bar corresponds to the total sales in a particular market, providing a visual comparison of sales performance across different regions. Markets with substantial sales but relatively lower profits may prompt further analysis to understand factors influencing profit margins.
3. Market-specific Analysis: The chart enables a market-specific analysis, aiding in the identification of markets with exceptional profit generation or notable sales figures. Such insights guide decision-makers in tailoring strategies to capitalize on strengths or address challenges in specific markets.
4. Overall Financial Overview: By juxtaposing profits and sales in a single visualization, the chart delivers a comprehensive financial snapshot of the superstore’s performance across diverse markets. This information is instrumental for formulating targeted strategies, optimizing resource allocation, and aligning business objectives with market specific trends.

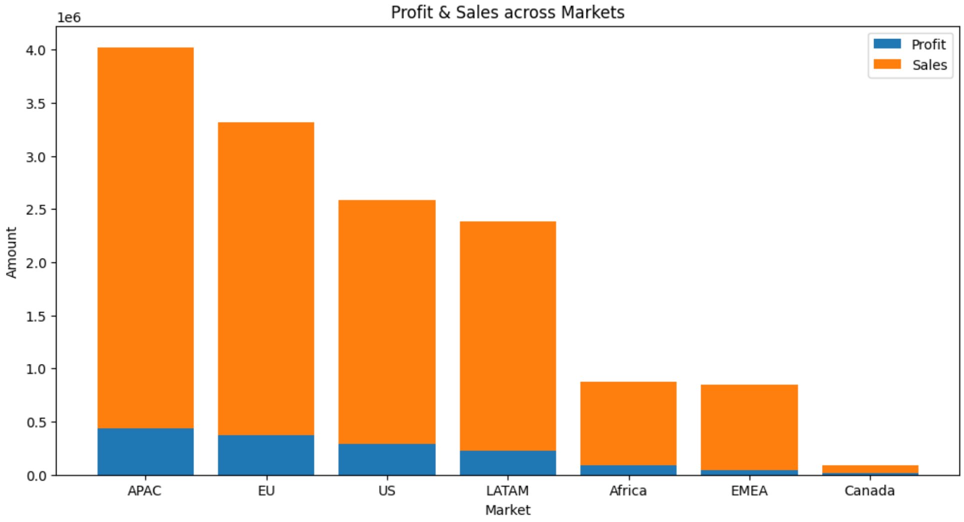


Figure 3: Bar chart to display profit and sales across markets

Line Plot

A line plot illustrates monthly sales trends segmented by different markets, providing a dynamic perspective on the global superstore's revenue patterns over time. The x-axis represents the timeline, spanning months, while the y-axis denotes the total sales amount. Each line on the plot corresponds to a specific market, distinguished by distinct colors for clarity.

The visual analysis of the plot allows us to discern several insights into the sales dynamics across markets.

1. Temporal Patterns: The plot reveals fluctuations in sales across different months, highlighting potential seasonality or periodic trends. Understanding these temporal patterns is crucial for anticipating periods of increased or decreased demand, aiding in inventory management and resource planning.
2. Market-specific Performance: By examining individual lines representing different markets, we can identify variations in performance. Markets with consistently rising sales may indicate further investigation into potential influencing factors.
3. Comparison Across Markets: The plot facilitates a direct comparison of sales performance among various markets. This information is valuable for identifying high-performing markets that may warrant focused marketing efforts or exploring untapped opportunities in markets with slower growth.
4. Overall Sales Trends: The gridlines assist in evaluating the overall sales trajectory, allowing stakeholders to gauge the general direction of sales growth or decline. This insight contributes to strategic decision-making for future planning and resource allocation.

The Monthly Sales by Market line plot offers a comprehensive view of the superstore's sales dynamics, enabling stakeholders to identify patterns, trends, and market-specific behaviors. This information is instrumental for devising targeted marketing strategies, optimizing inventory management, and making informed decisions to enhance overall business performance.

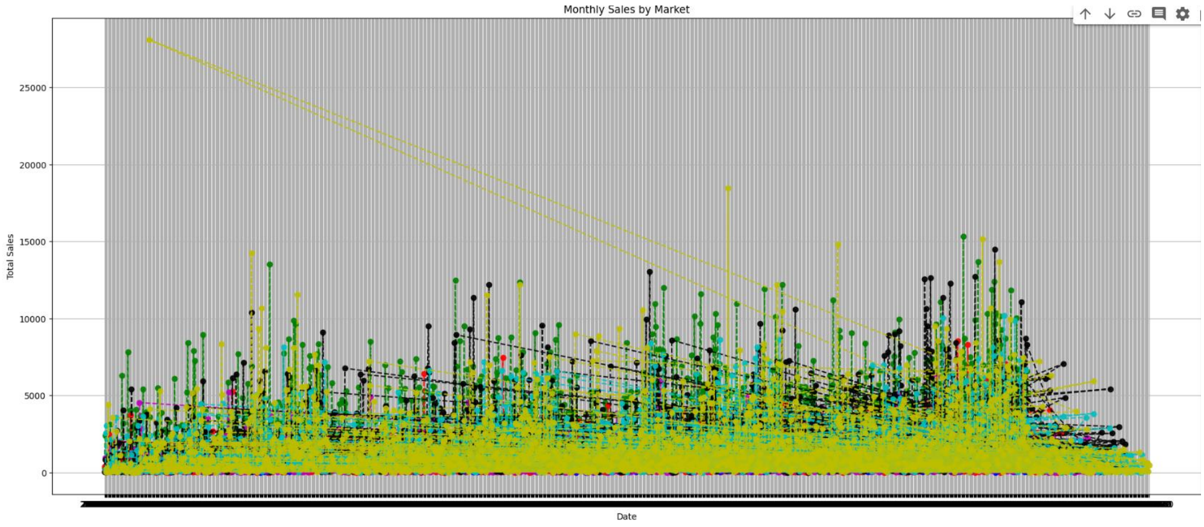


Figure 4: Line plot to show monthly sales by market

Pie Chart

The pie chart visualizes the distribution of sales across different regions in the global superstore dataset. By aggregating sales data based on geographical regions, the chart provides a clear and concise representation of the relative contribution of each region to the overall sales revenue.

The chart reveals that sales are not evenly distributed, with certain regions contributing more significantly than others. Each slice of the pie corresponds to a distinct region, and the size of each slice is proportional to the total sales generated by that region. The use of a pastel color palette enhances the visual appeal and ensures clarity in distinguishing between regions.

As depicted in the chart, the Sales by Region analysis showcases the relative importance of each geographical area in terms of revenue generation. This information is valuable for strategic decision-making, allowing stakeholders to identify high-performing regions that may warrant additional investment or targeted marketing efforts. Conversely regions with lower sales figures may require further examination to understand the factors influencing their performance.

In conclusion, the pie chart effectively communicates the sales distribution across different regions, providing a snapshot of the global superstore’s regional revenue landscape. This insight contributes to a holistic understanding of market dynamics, guiding potential strategies for resource allocation and business growth.

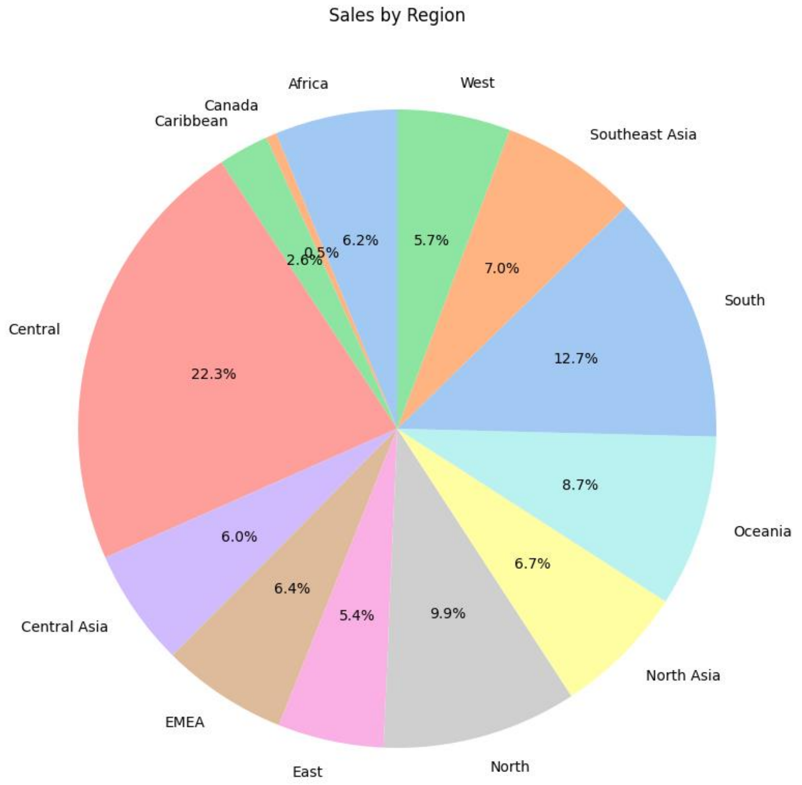
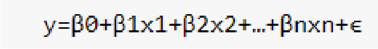


Figure 5: Distribution of sales across different regions

Modeling and Analysis

Multiple Linear Regression

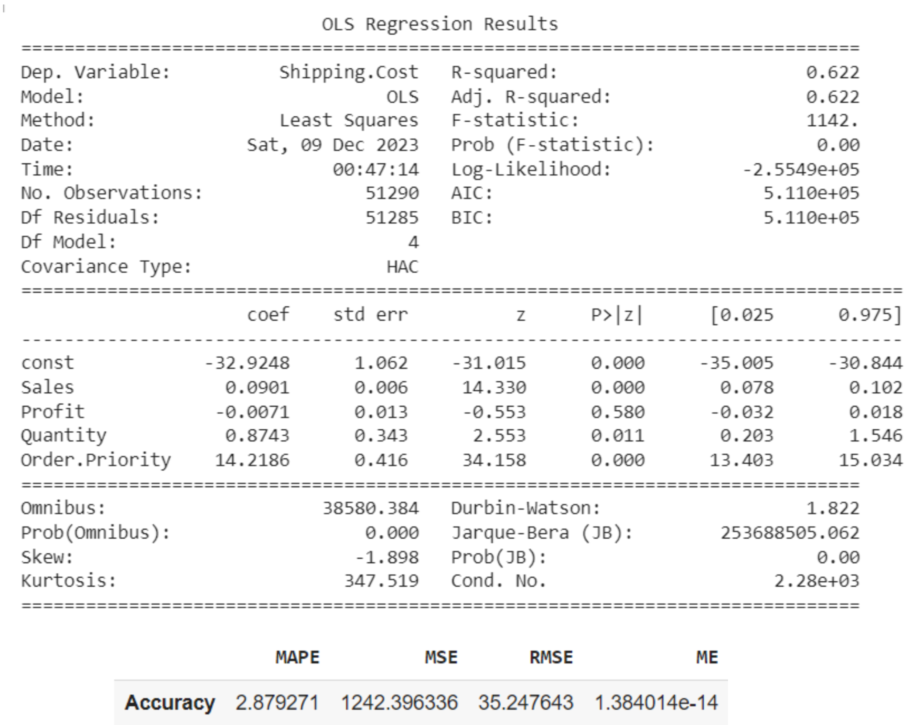
The project employs a Multiple Linear Regression model to predict shipping costs based on various features. The analysis includes a detailed examination of regression coefficients and statistical significance tests. The model performance metrics, such as Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Error (ME) provide a comprehensive evaluation of the model's accuracy.



y=β0+β1x1+β2x2+…+βnxn+ϵ

Figure 6: Multiple Linear Regression formula

The regression results showcase an R-squared value of 0.622, indicating that the model explains 62.2% of the variance in shipping costs. The coefficients reveal that Sales, Quantity, and Order Priority significantly influence shipping costs. However, Profit shows no statistical significance.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Shipping.Cost   R-squared:                       0.622
Model:                            OLS   Adj. R-squared:                  0.622
Method:                 Least Squares   F-statistic:                     1142.
Date:                Sat, 09 Dec 2023   Prob (F-statistic):               0.00
Time:                        00:47:14   Log-Likelihood:            -2.5549e+05
No. Observations:               51290   AIC:                         5.110e+05
Df Residuals:                   51285   BIC:                         5.110e+05
Df Model:                           4
Covariance Type:                  HAC
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const            -32.9248      1.062    -31.015      0.000     -35.005     -30.844
Sales              0.0901      0.006     14.330      0.000       0.078       0.102
Profit            -0.0071      0.013     -0.553      0.580      -0.032       0.018
Quantity           0.8743      0.343      2.553      0.011       0.203       1.546
Order.Priority    14.2186      0.416     34.158      0.000      13.403      15.034
==============================================================================
Omnibus:                    38580.384   Durbin-Watson:                   1.822
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        253688505.062
Skew:                          -1.898   Prob(JB):                         0.00
Kurtosis:                     347.519   Cond. No.                     2.28e+03
==============================================================================
```

	MAPE	MSE	RMSE	ME
Accuracy	2.879271	1242.396336	35.247643	1.384014e-14

Figure 7: Multiple linear regression results

Notably, the positive coefficient for sales suggests a proportional increase in shipping costs with higher sales volumes, emphasizing the operational impact of order magnitude on logistical expenses. Interestingly, the negative coefficient for profit, though statistically non-significant, hints at a potential cost-saving opportunity associated with more profitable transactions. However, caution is warranted in over-interpreting this relationship due to its lack of statistical significance. On the other hand, the positive relationship between quantity and shipping cost aligns with logistical expectations, indicating that larger quantities contribute to increased shipping expenses. The standout predictor is the order priority, as reflected by its high coefficient and statistical significance. This underscores the strategic importance of order prioritization in managing and forecasting shipping costs. The overall model, with a commendable R-squared of 0.622, captures a substantial portion of the variability in shipping costs, providing a robust foundation for strategic decision-making in logistics and supply chain management.

The performance metrics further support the model's effectiveness. The MAPE of 2.88% suggests that, on average, the model's predictions are accurate within 2.88% of the actual shipping costs. The MSE and RMSE values of 1242.40 and 35.25 respectively, reflect the model's ability to minimize prediction errors. The ME close to zero indicates a well-calibrated model, aligning predicted and actual shipping costs.

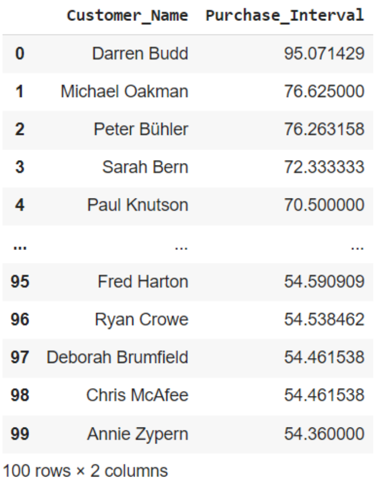
Purchase Interval Analysis

In addition to geographical and sales analyses, this project delves into understanding customer behavior through the exploration of purchase intervals, This analysis specifically focuses on identifying customers likely to churn based on extended intervals between orders.

The approach involves employing data mining methods to locate customers with prolonged purchase intervals. This is a crucial aspect of customer retention strategies, as identifying customers showing signs of extended inactivity allows for targeted promotional efforts. The primary objective is to re-engage these customers, preventing churn and encouraging continued interaction with the superstore.

The purchase intervals analysis calculates the average time intervals between consecutive orders for each customer. By determining customers with the longest intervals, the superstore can tailor promotions and incentives to rekindle interest and encourage more frequent transactions.

This data-driven method aims to optimize promotional efforts, ensuring they are directed toward customers most likely to churn due to prolonged purchase intervals. The insights gained from this analysis contribute to a proactive customer retention strategy, aligning with the broader goal of enhancing customer satisfaction and sustaining long-term relationships with the superstore.



	Customer_Name	Purchase_Interval
0	Darren Budd	95.071429
1	Michael Oakman	76.625000
2	Peter Bühler	76.263158
3	Sarah Bern	72.333333
4	Paul Knutson	70.500000
...	...	...
95	Fred Harton	54.590909
96	Ryan Crowe	54.538462
97	Deborah Brumfield	54.461538
98	Chris McAfee	54.461538
99	Annie Zypern	54.360000

100 rows × 2 columns

Figure 8: Purchase interval analysis results

Folium Map

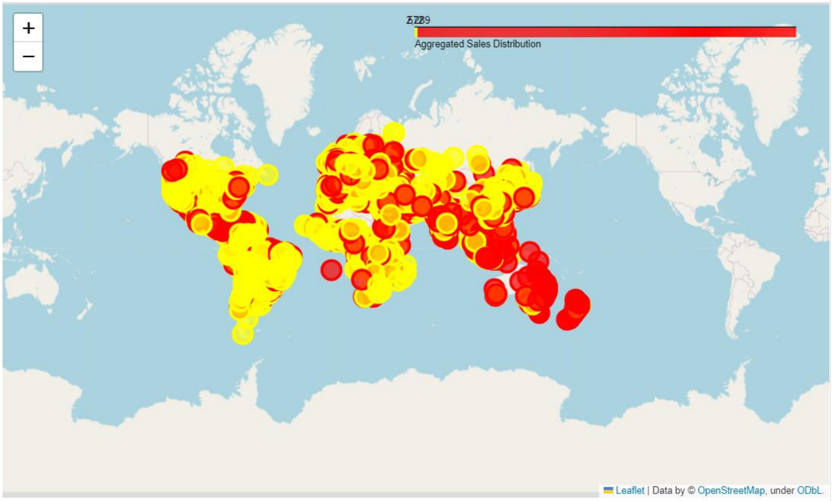
The interactive Folium map provides a captivating visual representation of aggregated sales across different cities within the global superstore network. Leveraging geographic coordinates associated with each city, this map allows stakeholders to gain valuable insights into the distribution of sales on a global scale.

Each city is marked on the map with a distinctive circular marker. The size of the markers corresponds to the aggregated sales in each city. Larger markers signify higher sales, providing an immediate visual cue to the relative sales magnitude in different locations.

The markers are color-coded based on the aggregated sales figures, employing a step colormap that transitions from yellow to orange and finally red. This color gradient aids in quickly distinguishing cities with varying sales levels. Yellow indicates lower sales, orange represents moderate sales, and red signifies cities with the highest aggregated sales.

Clicking on any marker reveals a popup with detailed information about the respective city's aggregated sales. This includes the city name and the total sales amount, offering a convenient way to access specific sales data for each location.

This map serves as a powerful tool for visually exploring and understanding the global distribution of sales within the superstore network. It enables stakeholders to identify key hubs of sales activity, spot regional variations, and inform strategic decisions related to marketing, resource allocation, and market-specific initiatives. The combination of size and color coding enhances the map's effectiveness in conveying complex sales data in an accessible and intuitive manner.


+
−
Aggregated Sales Distribution
Leaflet | Data by © OpenStreetMap, under ODbL.

This comprehensive analysis not only provides insights into the global superstore's sales landscape, but also offers a robust predictive model for shipping costs. The geographical, temporal, and predictive aspects collectively contribute to a data-driven decision-making framework, empowering strategic improvements and informed business choices.

Future Work

In future, this project can be enhanced by implementing a dynamic data integration pipeline for real-time or periodic updates. This enhancement would contribute to the project's evolution into a robust and adaptable analytics framework, continually providing value to the global superstore's strategic decision-making processes.