Erin Kane

MSDS670 Data Visualization

Koenig

25 June 2023

***Situation***

The analysis uses the superstore.csv dataset, which features Orders, ship dates, customer id, region, category, etc., to create visuals that display in the histogram, heatmap, and choropleth models. The Superstore dataset comes from Tableau.com as a range of different areas to investigate. It contains information about additional products, sales, and profits so the user can identify opportunities for improvement and growth.

***Research Questions***

Since the Superstore dataset has 19 different features, the research questions primarily focus on the following:

* Which state of the United States has a higher frequency of sales
* In which year were max sales
* In which month were the max sales
* Which days of the month yield higher sales
* What are the top 15 states with high sales
* What are the top 15 cities with high sales
* What is the ultimate product that is in high demand
* Who are the most frequent customers

***Data***

The Superstore.csv dataset focuses on twenty features, narrowing in on the frequency of sales, the maximum sales for day, month, and year, the top ten states, the top twenty cities, the top product purchase, and the top customer. There are 9994 different entries in this dataset, with no missing values.

***Methodology & Results***

To begin, the necessary libraries are imported into the Python code editor. The libraries used are numpy, pandas, matplotlib, seaborn, plotly, and statsmodels. These libraries allow the user to investigate the data and use the findings to create visually appealing graphics that help provide more insight into the happenings of the Superstore.

Next, Exploratory Data Analysis is applied, walking through the steps to identify the data info(). Then, to see the top of the dataset, and the shape of the dataset, use the describe() function to statistically analyze the data, then locate any missing values within the dataset—the decision to drop the postal code data column and then plot any correlations with a heatmap, for this heatmap, I chose the color palate ‘Crest’ I liked the different variant of colors.

A screenshot of a computer

Description automatically generated with medium confidence

*Figure 1 – Initial Heatmap to identify correlations in palate ‘Crest’*

Using Figure 1: Heatmap to find the correlations within the data, I chose to drop Row Id, Ship Date, Ship Mode, Customer ID, Postal Code, Order ID, Profit, and Discount from the future set. Once those features were dropped, I saved this as a new dataset.

I continued to press forward on the previously mentioned questions and analyzed the ‘Country’ Column. I narrowed it down to only United States to identify the different states and cities within those parameters. I was able to locate the top fifteen cities:

1. New York
2. Los Angeles
3. Philadelphia
4. San Francisco
5. Seattle
6. Houston
7. Chicago
8. Columbus
9. Sand Diego
10. Springfield
11. Dallas
12. Jacksonville
13. Detroit
14. Neward
15. Richmond

Next, I found the most frequent customers. After identifying the most frequent customer, I evaluated the most frequently purchased product. To do this, I had to create a unique category column. Then I used this new category to generate The Distribution of Sales in Each Category for most purchased products. I used a barplot in the palate, ‘ice fire.’

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

*Figure 2 – Barplot of Distribution of Sales in each Category.*

Taking this a step further, I concentrated on the product subcategories with high sales. I continued with the same color scheme and produced the following:

A screen shot of a graph

Description automatically generated with low confidence

*Figure 3 – Distribution of Sales in each Subcategory.*

Moving through the data and the different data visualizations available, I could determine the sales in each segment, a statewide sales distribution, identify the state with the highest sales, and the complete sales distribution. This is a pretty neat visual, using a distribution plot, in the color blue, with x label Range of Sales, y label No of Sales count, and Title The Distribution of Sales.

A picture containing text, screenshot, plot, number

Description automatically generated

*Figure 4 – The Distribution of Sales*

***Conclusion***

In summary, the Superstore.csv dataset was a fantastic dataset to work with, and it allowed me to practice my work in visualizations. I could identify and find all the answers to my questions, work with Matplotlib, and plotly to provide exciting and relevant information displayed in heatmaps, bar plots, and choropleth models.

***References***

Seaborn heatmap color templates.

<https://seaborn.pydata.org/generated/seaborn.heatmap.html>

<https://pythonbasics.org/seaborn-heatmap/>