INFS 5116 Data Visualisation
Practical Exercise 2
Create your own graphics using R

Enna

Contents

Student name: Huining Huang

Student ID number: 110412586

Introduction

The Dominick's Finer Foods dataset is a rich collection of store-level scanner data spanning over seven years and approximately 98 million observations. The focus is on utilizing various visualization techniques to harness the scope and potential of Big Data for generating meaningful research questions. The beer data from the Dominick's dataset is used here. The visualizations are categorized into elementary, intermediate, and overall-level question visualizations, each serving unique analytical purposes.

This study integrates various visual variables such as position, size, color value, texture, color hue, orientation, and shape to produce nuanced visual representations. The key aim is to go beyond mere data presentation to facilitate data exploration, inform decision-making, and improve communication. Through these visualizations, we aim to unlock insights into product pricing, profitability, and other performance metrics that can be essential for strategic planning and operational efficiency.

Graphic 1

- Store Frequency Across Different Store Segments

- Level of Analysis: Elementary

-

Design information

Type of Graphic: Bar Chart

Invariant: The graphic focuses on representing the frequency of stores across different segments. The x-axis consistently shows 'Store IDs', and the y-axis shows 'Frequency'.

Components: Three separate bar plots for different store subsets

Imposition: Vertical layout for the bars, aligned in a single-column, three-row format for easy comparison across store segments.

Planar Variables: X-axis is 'Store', Y-axis is 'Frequency'.

Retinal Variables: Height of bars (Quantitative), Color (Nominal - to differentiate between store IDs within each subset using viridis palette).

Gestalt Principles: Proximity - bars within the same subset plot are closer to each other to imply belonging to the same category. Similarity - use of a consistent color scale across plots and subsets to denote the same type of data (Frequency).

Bertin's Schemata: The scale and axis remain invariant across the different plots for uniformity and easier cross-comparison.

Graphic 1

Questions of interest and conclusion

Questions: 1. Which stores have the highest frequency in each segment? 2. Are there stores with particularly low frequencies that require attention? 3. Does the frequency distribution indicate any particular trends across store IDs? 4. Are there stores that are over or underrepresented in the dataset?

Conclusion:

The bar charts provide a detailed view of store frequencies across various segments. They identify stores with high frequencies, serving as a key input for allocating resources or running targeted promotions. The charts also flag stores with low frequencies, signaling the need for investigative action. Despite its simplicity, the design leverages Gestalt principles to aid in cognitive processing. However, for enhanced comparative analysis across segments, data labels for low-frequency stores and a unified y-axis could be advantageous.

Graphic 2

- Top and Bottom 10 Stores by sales

- Level of Analysis: Intermediate-Level

Design information

Type of Graphic:
Combined Bar Plots

Invariant:
The graphic aims to compare sales performance between the top 10 and bottom 10 stores, segmented by Price Tier. The y-axis represents Total Sales, and the x-axis represents 'STORE'.

Components:
Two separate bar plots juxtaposed horizontally. One plot represents the top 10 stores , and the other represents the bottom 10 stores.

Imposition:
Vertical bars, horizontally juxtaposed layouts for direct top-bottom comparison.

Planar Variables:
X-axis is 'Store', Y-axis is 'MOVE' (Total Sales).

Retinal Variables:
Height of bars (Quantitative - 'MOVE'), Color fill (Nominal - 'Price_Tier').
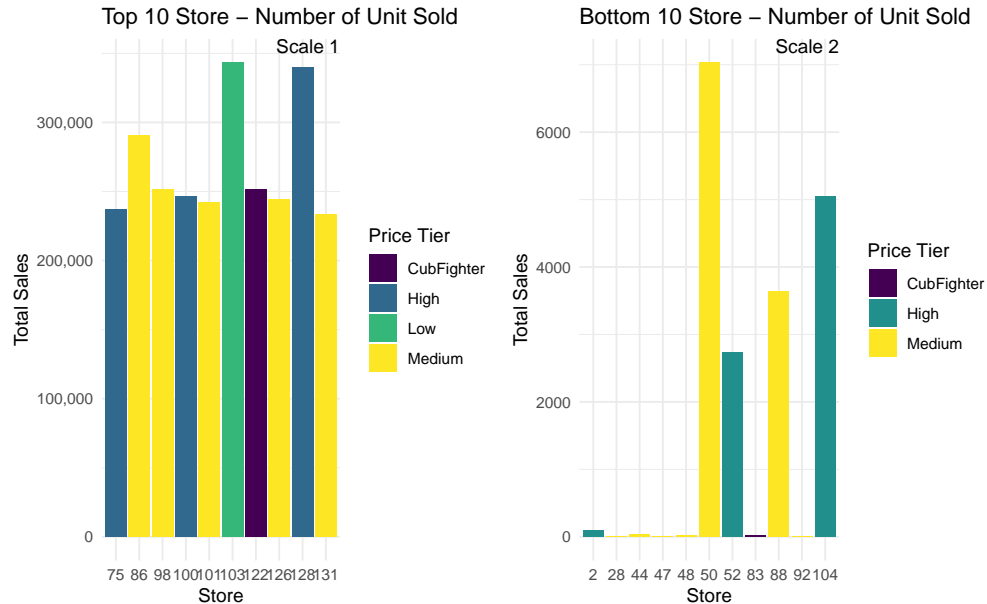
Gestalt Principles:
Proximity - Bars within the same plot are closer to each other, indicating they belong to the same category (either top 10 or bottom 10 stores).
Similarity - Use of the same color scale across the top and bottom plots to indicate Price Tier.

Bertin's Schemata:
The scale and axis remain invariant between the two plots for easier comparison.

Graphic 2

Questions of interest and conclusion

Questions:

1. Which stores are performing the best and worst in terms of unit sales?
2. How does the Price Tier correlate with total sales for each store?
3. Which Price Tier contributes the most to the sales in top-performing stores?
4. Is there a pattern in Price Tier sales distribution among the bottom-performing stores?
5. Are there any anomalies in the sales patterns of individual stores that warrant further investigation?

Conclusion: The combined bar plots adeptly facilitate the comparison of sales performance between the top 10 and bottom 10 stores, segmented by Price Tier. The layout and choice of retinal variables enable complex yet straightforward analyses. However, the graphic could be enhanced by standardizing the y-axis or explicitly noting the differing scales. This additional layer would mitigate potential misinterpretations, thereby making the visualization more robust for strategic decisions.

Graphic 3

- 100% Stacked Histogram of Gross Margin by Price Tier

- Level of Analysis: Intermediate-Level

Design information

Type of Graphic:
100% Stacked Histogram

Invariant:
The visualization aims to represent the proportion of gross margin (PROFIT) across different Price Tiers. The x-axis encodes binned profit, and the y-axis represents the proportion within each bin for various Price Tiers.

Components:
Single-component, portraying the distribution of gross margin by Price Tier using vertical bars. The data binning is used to simplify the complexity of the data structure but the data requires further cleaning and organization.

Imposition:
Vertical bars, stacked to represent the proportion of each Price Tier within each profit bin. The x-axis labels could benefit from better formatting or a larger font size for easier reading.

Planar Variables:
X-axis is 'binned_profit', Y-axis is 'prop' (proportion).

Retinal Variables:
Height of bars (Quantitative - proportion), Color fill (Nominal - Price Tier).

Gestalt Principles:
Proximity - Bars within the same profit bin are stacked together, indicating they belong to the same profit range.
Similarity - Use of the same color palette across all bins to indicate Price Tier.

Bertin's Schemata:
The scale for proportion remains consistent across the visualization, aiding in easier comparison.

Graphic 3
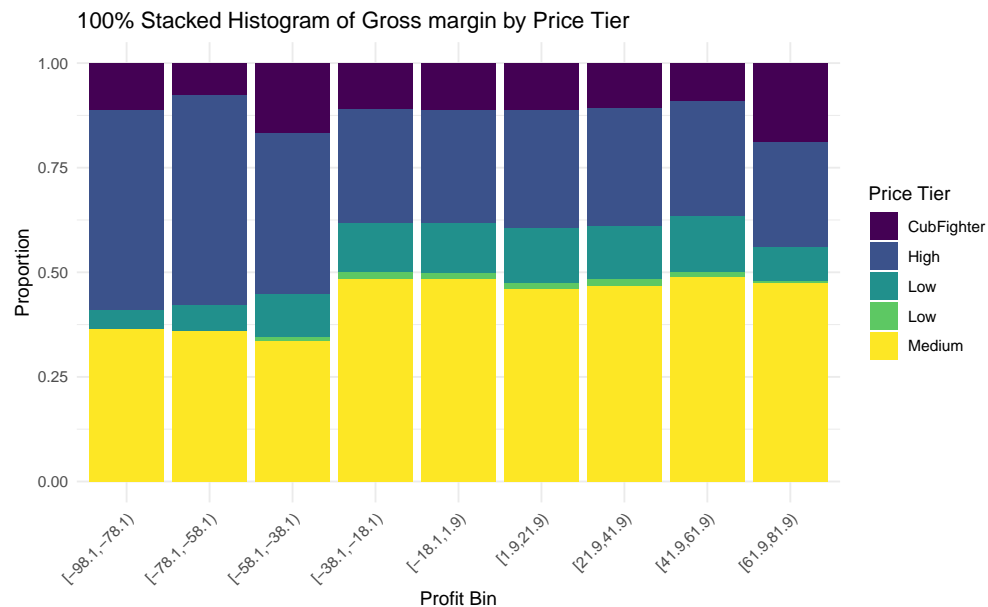
Questions of interest and conclusion

Questions:

1. What is the proportion of high-profit products within each Price Tier?

2. Which Price Tier has a higher percentage of low-profit or negative-profit items?
3. Is the distribution of profit similar across different Price Tiers?

Conclusion:

The 100% Stacked Histogram effectively portrays the distribution of gross margin across various Price Tiers. The vertical bars and color palettes are chosen to represent both proportion and Price Tier, enhancing interpretability even for those with color vision deficiencies. While the graphic is an effective tool for identifying Price Tiers requiring further investigation, enhancements in readability and potential interactive features could make it even more user-friendly and informative.

Graphic 4

- Treemap of Gross Margin by Price Tier and Year
- Level of Analysis: Overall

Design information

Type of Graphic:
Treemap

Invariant:
The visualization focuses on displaying the hierarchical and part-to-whole relationships of 'PROFIT' across 'Price_Tier' and 'year'.

Components:
Single-component, illustrating the structure of the 'PROFIT' metric by 'Price_Tier' and nested by 'year'.

Imposition:
Rectangles are spatially positioned and enclosed to represent hierarchy, size, and part-to-whole relationships. It employs a well-selected color palette and spatial positioning to guide the eye through the data.

Planar Variables:
Spatial positioning of rectangles (indicating hierarchy and part-to-whole relationships).

Retinal Variables:
Area of rectangles (Quantitative - 'PROFIT'), Color fill (Nominal - 'Price_Tier').

Gestalt Principles:
Enclosure - Rectangles are enclosed within larger rectangles to indicate hierarchy.
Proximity - Rectangles representing the same 'Price_Tier' or 'year' are positioned close to each other.

Bertin's Schemata:
Consistency in color palette and spatial positioning enhances the ability to quickly grasp hierarchical relationships.

Graphic 4

Treemap of Gross margin by Price Tier and Year
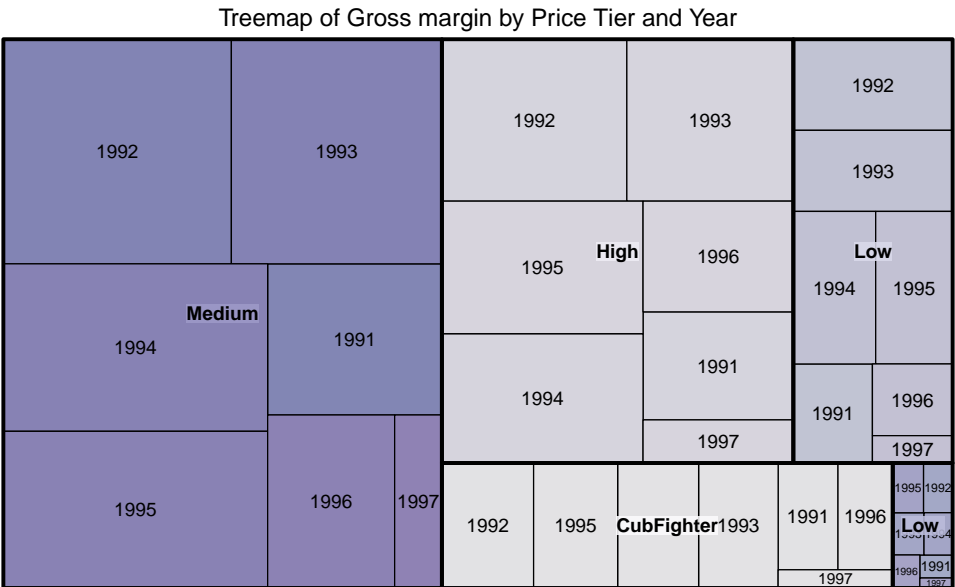
Questions of interest and conclusion

Questions:

1. Which 'Price_Tier' consistently has higher 'PROFIT' across years?
2. Is there a year where a particular 'Price_Tier' experienced a significant spike or drop in 'PROFIT'?
3. How does 'PROFIT' distribution within a 'Price_Tier' change over years?

Conclusion:

The treemap excels in visualizing the complex hierarchical and part-to-whole relationships of 'PROFIT' across 'Price_Tier' and 'year'. The use of spatial positioning and color differentiation aids in quick visual comprehension. While the design is efficient in conveying longitudinal profitability trends, the addition of in-rectangle labels could provide even more immediate clarity. No data integrity issues were observed, making this a reliable tool for longitudinal profitability analysis.

Graphic 5

- Calendar Heat Map: Retail Price by Weekday, Month, and Year
- Level of Analysis: Overall

Design information

Type of Graphic:
Calendar Heatmap

Invariant:
The heatmap focuses on visualizing time-series data of 'PRICE' across 'year', 'month', and 'weekday'.

Components:
Single-component, where the color intensity of the grid represents the 'PRICE' metric, segmented by 'year', 'month', and 'weekday'.

Imposition:
A grid-based layout structured by time dimensions ('year' and 'month') facilitates time-series analysis.

Planar Variables:
Grid layout segmented by 'year', 'month', and 'weekday'.

Retinal Variables:
Color intensity (Quantitative - 'PRICE').

Gestalt Principles:
Proximity and Continuity - Cells for the same 'year' or 'month' are grouped together, facilitating perception of temporal trends.

Bertin's Schemata:
The planar segmentation allows for an easy comparison and analysis of price fluctuations across time dimensions.

Graphic 5

Calendar Heat Map of Retail Price by Weekday and Month
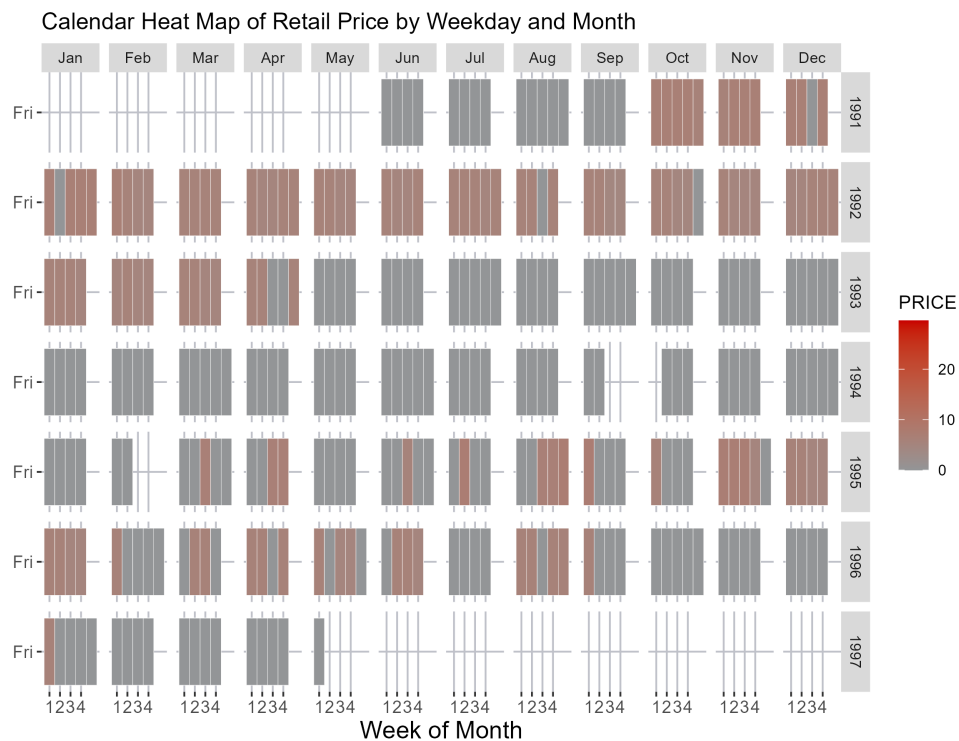
Questions of Interest and Conclusion

Questions:

1. How does retail 'PRICE' vary within a week for each month and year?
2. Are there any patterns or anomalies in 'PRICE' across weekdays?
3. Is there a noticeable seasonal variation in 'PRICE'?

Conclusion:

The calendar heatmap is a robust tool for exploring retail 'PRICE' variations across multiple time dimensions, offering insights into daily, monthly, and yearly trends. The grid layout, segmented by 'year', 'month', and 'weekday', is well-implemented. However, the visualization could gain in interpretability by employing a more perceptually uniform color palette and by adding labels for months and weekdays. In its current form, the heatmap already serves as a valuable asset for detecting temporal price trends and anomalies.

Graphic 6

- Heatmap of Profit by Month

- Level of Analysis: Intermediate-Level

Design Information

Type of Graphic:
Heatmap

Invariant:
The heatmap is an effective tool for identifying the relationship between 'PROFIT' and 'monthf'. It uses color density as a metric for frequency, simplifying complex data for quick interpretation.

Components:
Single-component heatmap where the x-axis represents 'monthf', the y-axis represents 'PROFIT', and the color density represents the frequency of data points.

Imposition:
The use of geom_bin2d() for binning creates a 2D grid that simplifies complex data into identifiable zones.

Planar Variables:
Grid layout partitioned by 'monthf' (x-axis) and 'PROFIT' (y-axis).

Retinal Variables:
Color density (Quantitative - frequency of data points).

Gestalt Principles:
Similarity and Proximity - Areas of similar color density are grouped together, facilitating the identification of trends or patterns.

Bertin's Schemata:
The grid layout allows for direct comparison across both axes, making it easier to identify correlations between 'PROFIT' and 'monthf'.

Graphic 6

Heatmap of Profit by Month
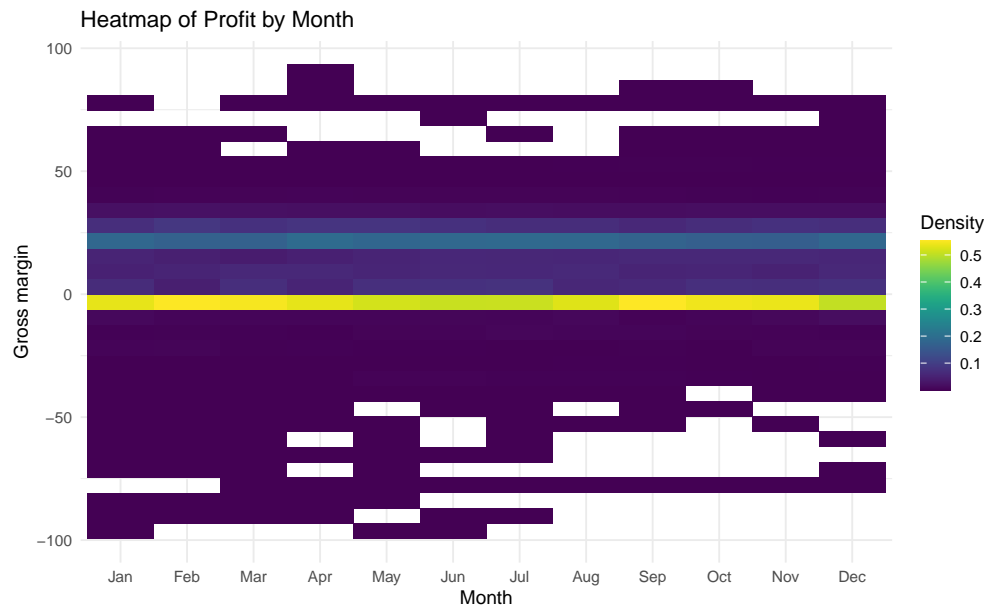
Questions of Interest and Conclusion

Questions:

1. Which months have the highest density of higher-profit items?
2. Is there any month where low or negative-profit items are more frequent?
3. Are there any significant patterns that can be observed when profits are mapped against months?

Conclusion:

The heatmap provides a compact yet insightful view of how 'PROFIT' varies across different months. Employing a Viridis color palette and data binning, it transforms complex data into easily interpretable patterns. However, the addition of a color legend and more granular axis labels could further enhance the visualization's utility. Despite these minor points for improvement, the heatmap remains a powerful tool for scrutinizing monthly trends in profitability, contributing valuable perspectives to business analytics.

References

Chiera , BA & Korolkiewicz, MW 2017, 'Visualizing Big Data: Everything Old is New Again', in FPG Marquez & B Lev (eds), Big Data Management, Springer, Netherlands, Ch. 1, pp. 1-27.

Credicts

- data source: https://www.chicagobooth.edu/research/kilts/research-data/dominicks
- Github-Copliot
- Gpt-4
- Grammarly
- vscode-R-extension