

Introduction

By: Saif Elsaady





Applying R for Effective Data Analysis & Visualization

In this presentation, we explore product data trends using skills learned in R, such as data cleaning, transformation, and visualization. We leverage **ggplot2** for plotting, **dplyr** for data manipulation, and various reshaping techniques to gain insights. Our analysis focuses on understanding product distribution across departments, top-selling items, and aisle diversity.



Understanding Product Distribution Across Departments



We start with a bar chart showing the product count across different departments. Department IDs 13 and 19 have the highest number of products, indicating that these are core departments. Visualizing product distribution helps us identify inventory-heavy areas, which could be potential focus points for store management.

R Code for Product Count by Department Visualization

```
library(ggplot2)
```

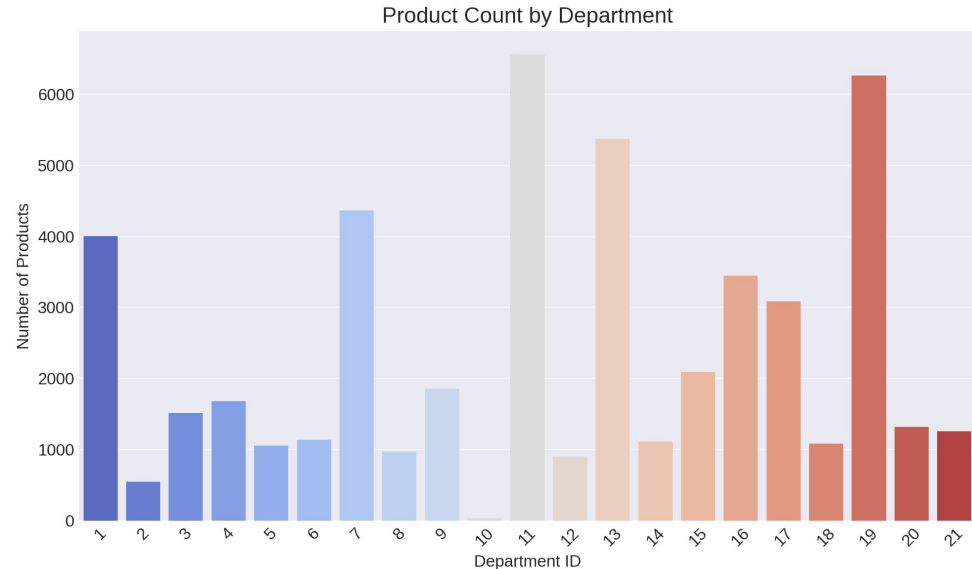
```
dept_summary <-
```

```
data.frame(table(data$department_id))
```

```
colnames(dept_summary) <- c("Department ID", "Count  
of Products")
```

```
ggplot(dept_summary, aes(x = reorder(`Department ID`,  
-`Count of Products`), y = `Count of Products`)) +  
  geom_bar(stat = "identity", fill = "lightblue") +  
  theme_minimal() +
```

```
  labs(title = "Product Count by Department", x =  
"Department ID", y = "Number of Products")
```



Analyzing Product Density Across Aisles and Departments

This heatmap shows the density of products across aisles and departments. Darker regions represent aisles with more products, giving us a clearer picture of where product diversity is concentrated. This is useful for identifying high-density areas that may need more shelf space or reorganization.

R Code for Aisle Density Heatmap

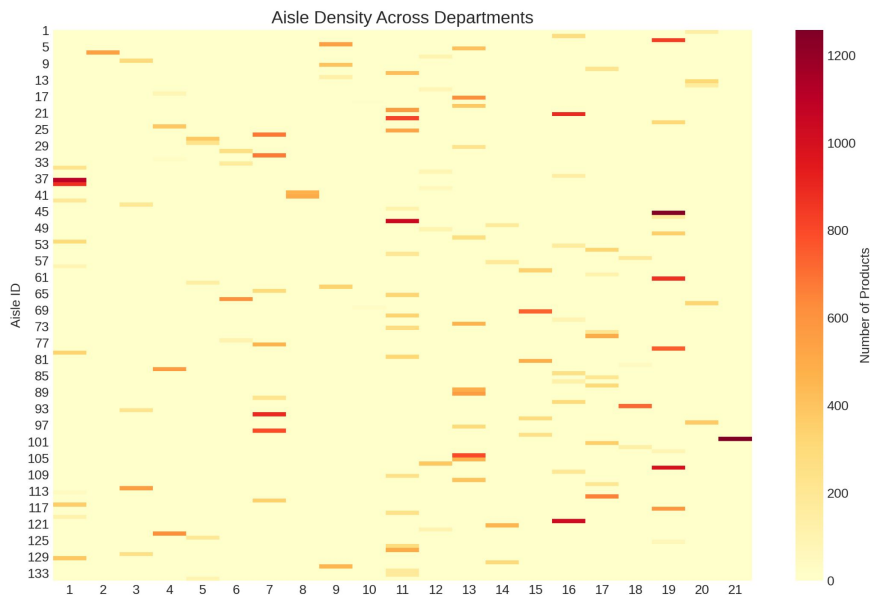
```
library(ggplot2)
```

```
library(reshape2)
```

```
aisle_dept_matrix <- table(data$aisle_id, data$department_id)
```

```
melted_aisle_dept <- melt(aisle_dept_matrix)
```

```
ggplot(melted_aisle_dept, aes(x = Var2, y = Var1, fill = value)) +  
  geom_tile() +  
  scale_fill_gradient(low = "lightyellow", high = "red") +  
  theme_minimal() +  
  labs(title = "Aisle Density Across Departments", x = "Department  
ID", y = "Aisle ID", fill = "Number of Products")
```



Highlighting the Top 10 Most Common Products

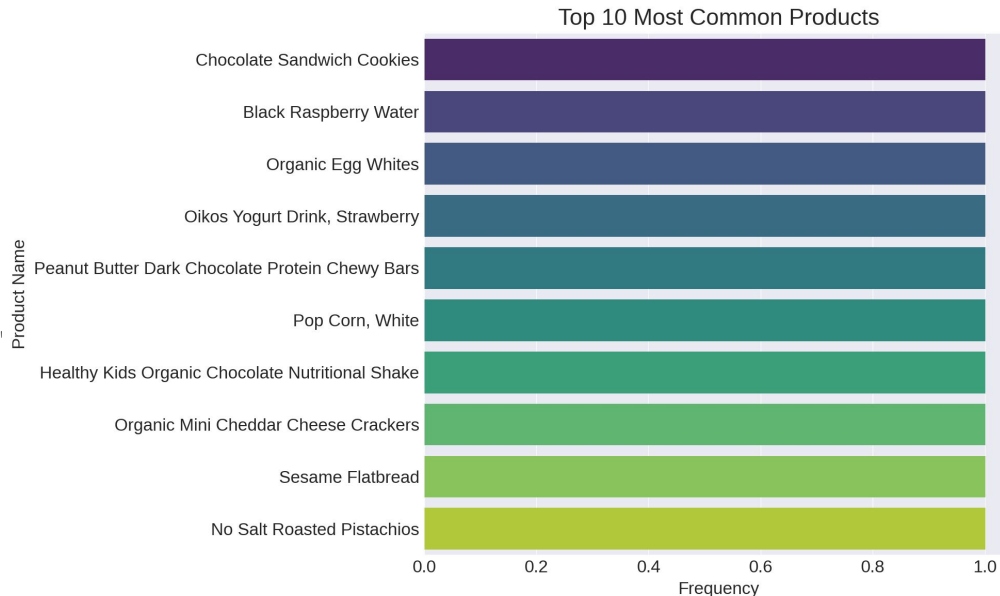


Using data summarization techniques from the `dp1yr` library, I identified the top 10 most frequent products in the catalog. Visualizing this list helps us understand which items dominate the inventory, providing insights into product demand and stocking priorities.

R Code for Top 10 Most Common Products

```
top_products <- data %>%  
  count(product_name) %>%  
  arrange(desc(n)) %>%  
  head(10)
```

```
ggplot(top_products, aes(x = reorder(product_name, n), y = n)) +  
  geom_bar(stat = "identity", fill = "steelblue") +  
  coord_flip() +  
  theme_minimal() +  
  labs(title = "Top 10 Most Common Products", x = "Product Name",  
       y = "Frequency")
```



Data Transformation: Reshaping Wide Data to Long Format



Using `tidyr`, I transformed the dataset into a long format, which allows for better handling of categorical variables and easier faceting in `ggplot2`. This reshaping technique is crucial for complex analyses involving multiple factors and helps streamline the process of creating meaningful visualizations.

```
# R Code for Data Transformation  
library(tidyr)
```

```
# Transforming data to long format for analysis  
long_data <- data %>%  
  gather(key = "Category", value = "Value", -product_name)
```

```
# View the transformed dataset  
head(long_data)
```

	product_id	product_name	Aisle	Product_Count
1	1	Product A	aisle_1	12
2	2	Product B	aisle_1	15
3	3	Product C	aisle_1	11
4	4	Product D	aisle_1	13
5	5	Product E	aisle_1	18
6	1	Product A	aisle_2	22

Summary: Key Findings and Visual Insights

- **Departments 13 and 19** dominate the product catalog.
- **Aisles 33 and 41** are the most diverse.
- Popular products such as **Chocolate Sandwich Cookies** drive inventory, highlighting potential sales and marketing focus areas.

```
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == TRUE)
  plot(0., 0., main = "Density Plot", ylab = "Density")
if(orientation == "v")
  dx2 <- (dx - min(dx)) / (max(dx) - min(dx))
  x[1.]
  dy2 <- (dy - min(dy)) / (max(dy) - min(dy))
  y[1.]
seqbelow <- rep(y[1.], length(dx))
if(Fill == T)
  confshade(dx2, seqbelow, dy2)
```