HW4

Salamot Fakoya

2024-10-13

```r
# Importing Necessary Libraries:
library(ggplot2)
library(tidyr)
library(UsingR)
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Loading required package: Hmisc
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

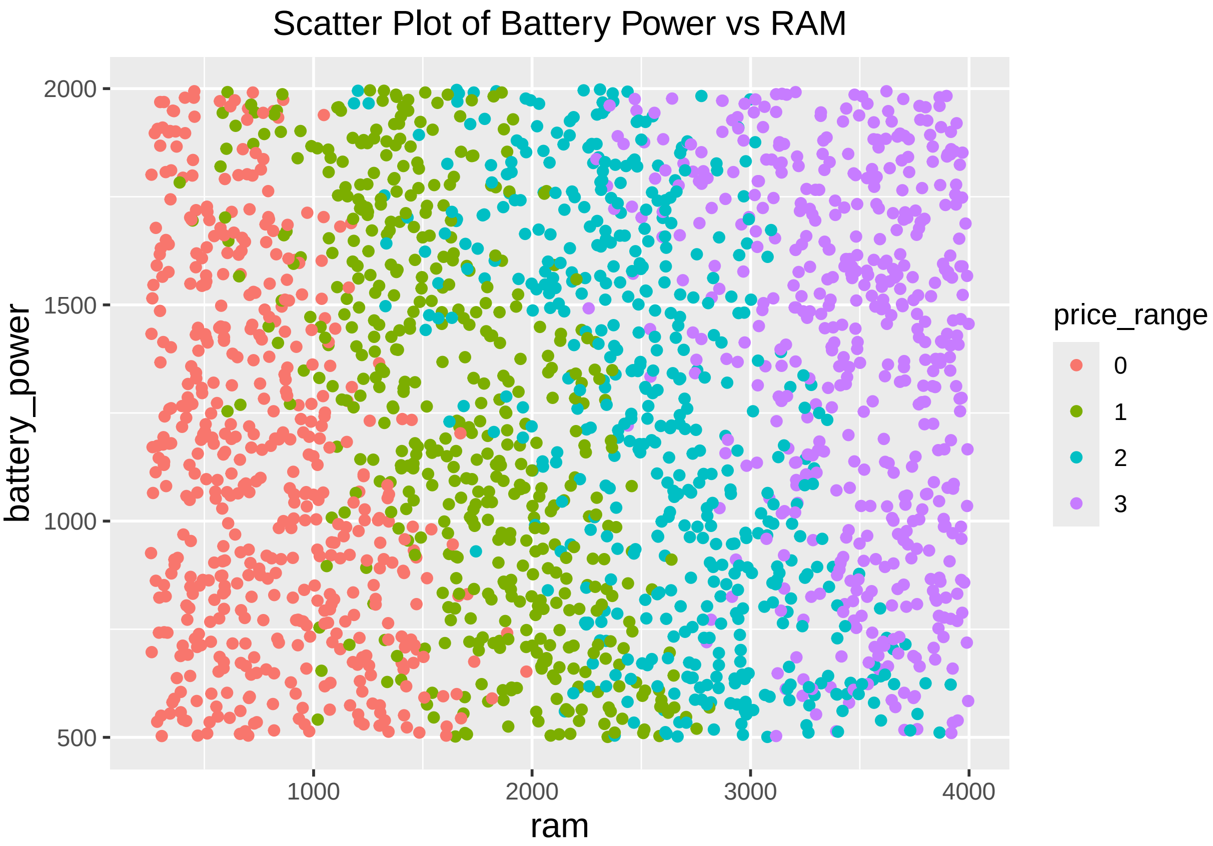
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Get Path to folder:
path <- "C:\\Users\\fakoy\\OneDrive - Houston Community College\\UTD_Courses\\Fall2024\\Data_Analysis_wi
setwd(path)
```

Problem 1

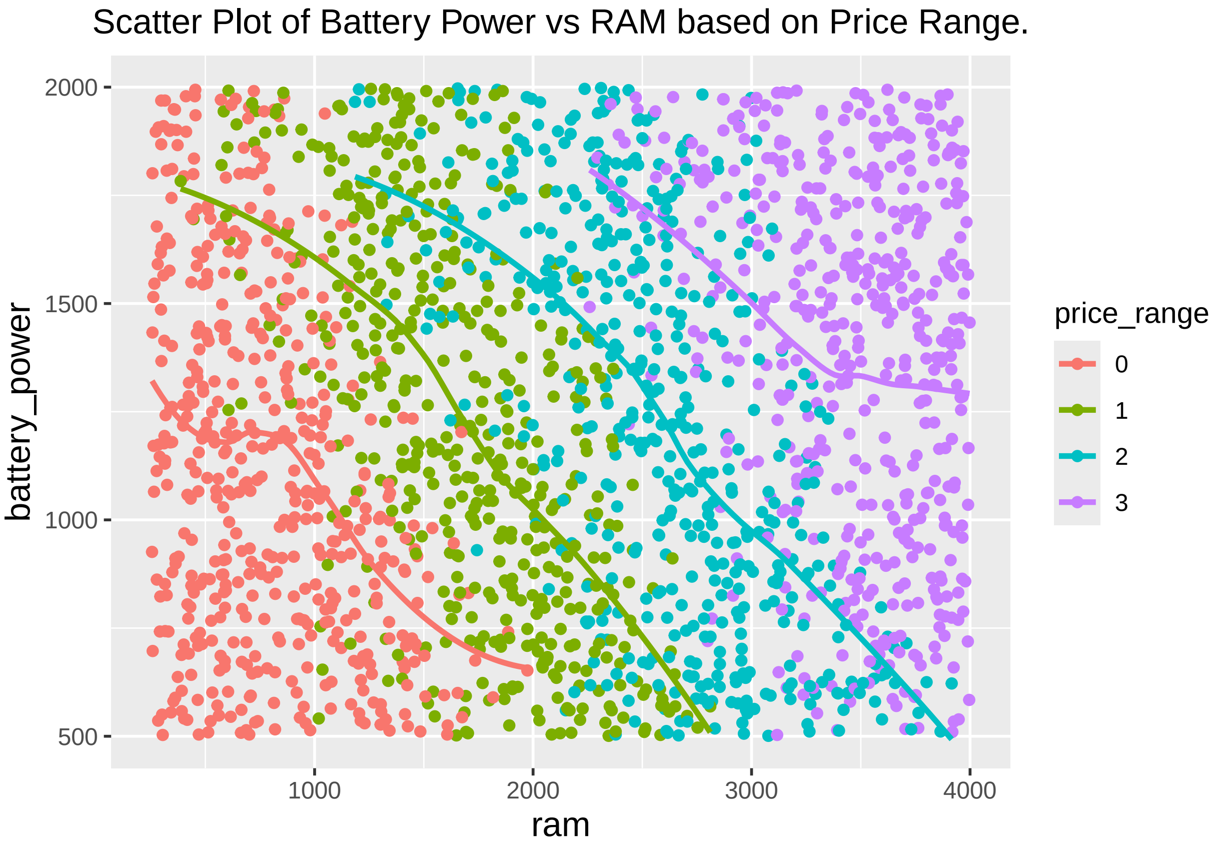
(a.)

```r
# Read the file:
df <- read.csv(file = "./HW4_Data/train.csv", header = T)

# 1a: scatterplot between battery_power vs ram
df$price_range <- factor(df$price_range)
base_plot <- df %>%
  ggplot(
    data = .,
    mapping =
      aes(x = ram, color = price_range)
  )

scatter_plot <- base_plot +
  geom_point(mapping = aes(y = battery_power)) +
  ggtitle("Scatter Plot of Battery Power vs RAM") +
  theme(
    plot.title = element_text(hjust = 0.5, size = rel(1.2)),
    axis.title = element_text(size = rel(1.2)))

scatter_plot
```



(b)

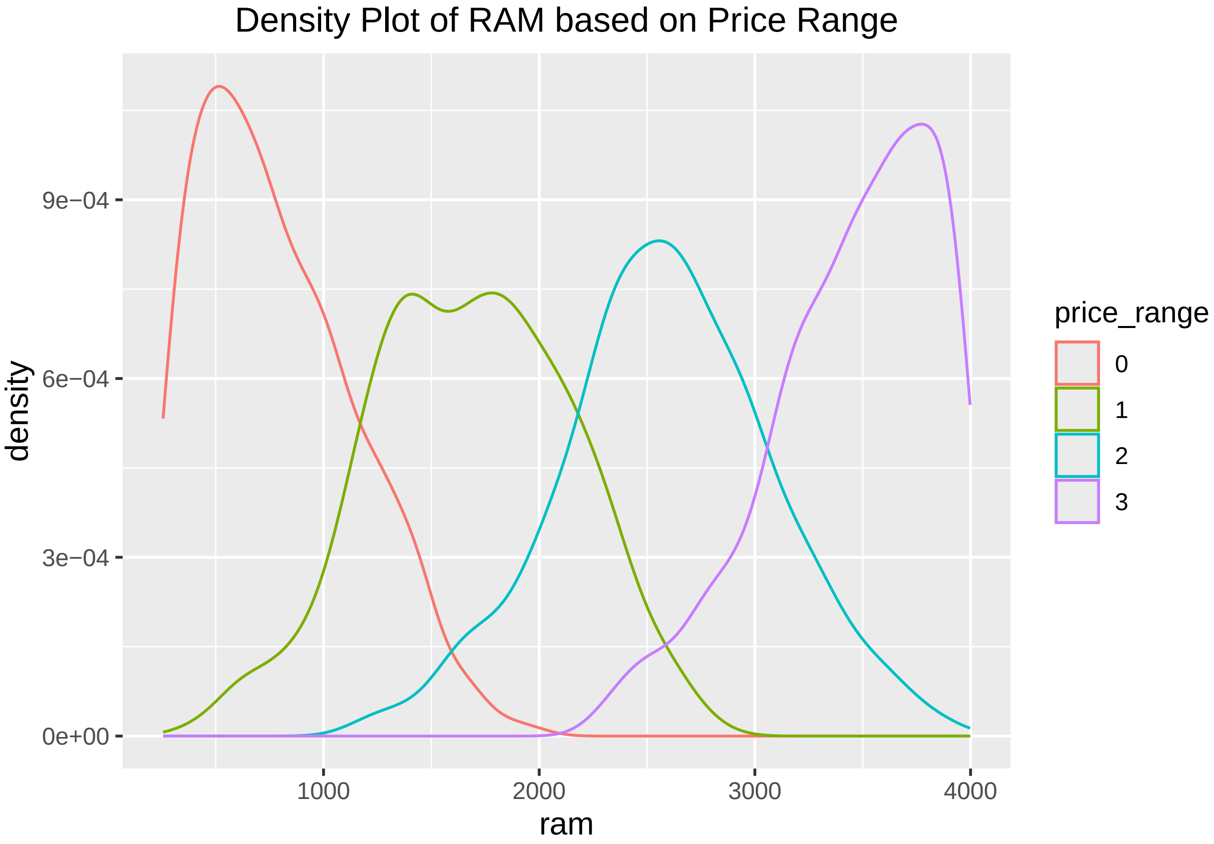
```r
# 1b: Recreate plot from (a), and add trend lines for each price separately
scatter_plot_w_trend <- scatter_plot +
  geom_smooth(
    mapping = aes(y = battery_power),
    method = "loess",
    formula = "y ~ x",
    fill = NA
  ) +
  coord_cartesian(
    ylim =
      c(
        round(min(df$battery_power), 2),
        round(max(df$battery_power), 2)
      )
  ) +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.location = "panel"
  ) +
  ggtitle("Scatter Plot of Battery Power vs RAM based on Price Range.")
scatter_plot_w_trend
```



(c)

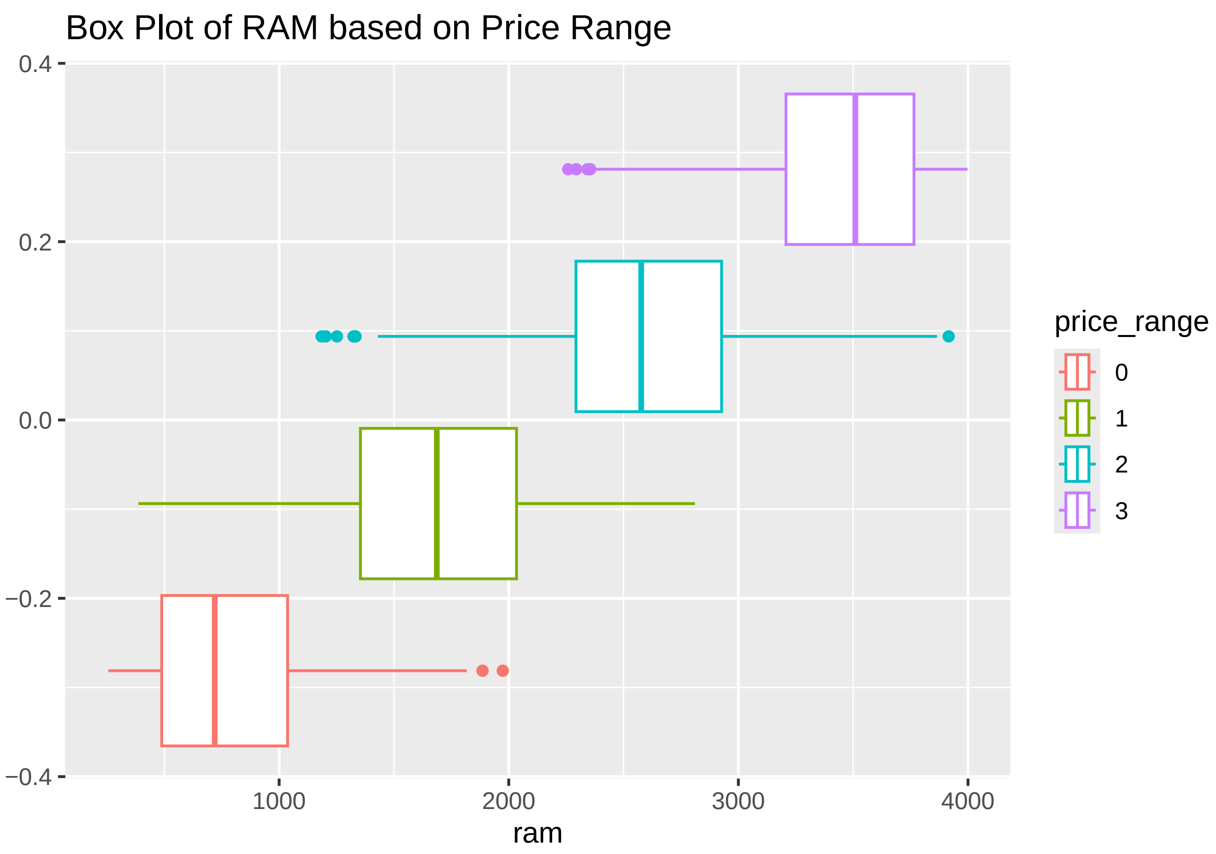
```r
# 1c: density curves in one plot
density_curve <- base_plot +
  geom_density() +
  theme(
    plot.title = element_text(hjust = 0.5, size = rel(1.2)),
    axis.title = element_text(size = rel(1.1))
  ) +
  ggtitle("Density Plot of RAM based on Price Range")
```

```
density_curve
```



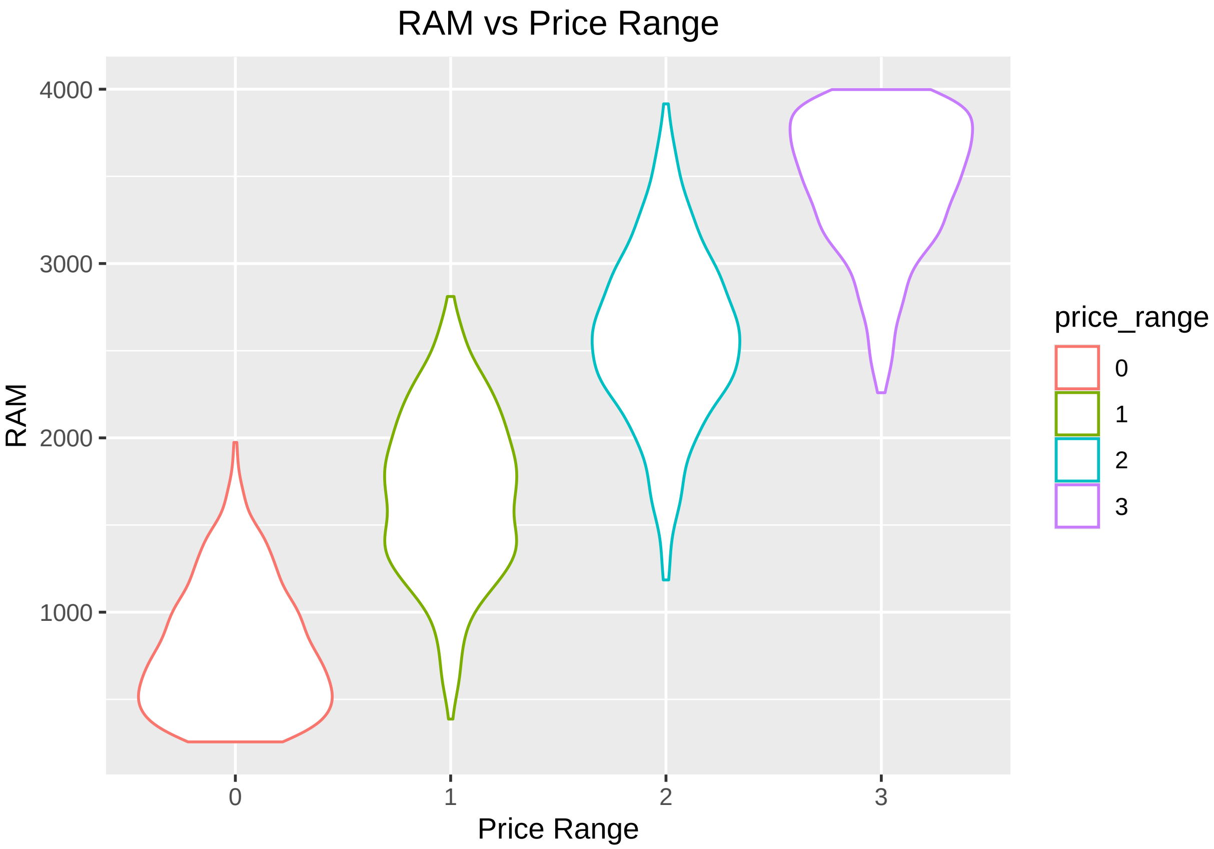
(d)

```
# 1d: boxplots in one plot
boxplots <- base_plot +
  geom_boxplot() +
  ggtitle("Box Plot of RAM based on Price Range")
boxplots
```



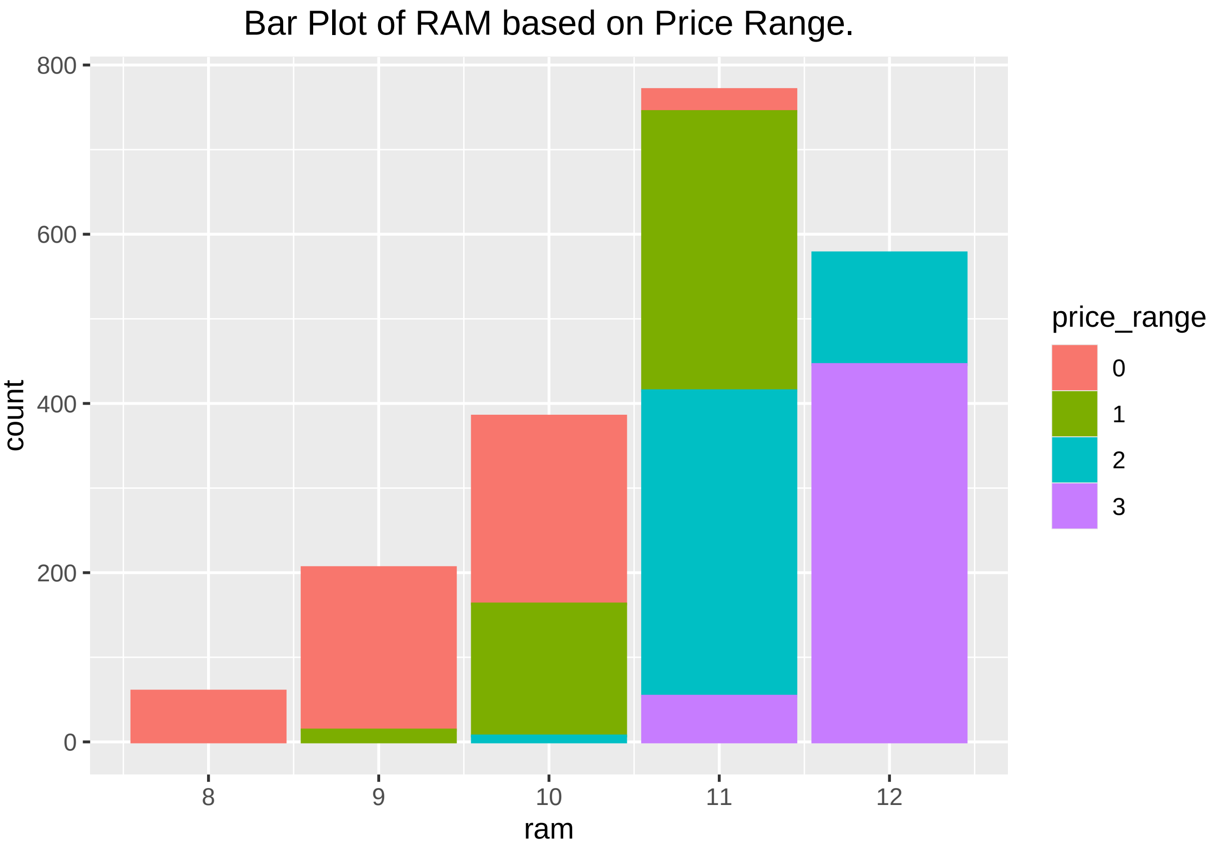
(e)

```
# 1e: violin plot
violin <- base_plot +
  geom_violin(mapping = aes(x = price_range, y = ram)) +
  xlab("Price Range") +
  ylab("RAM") +
  ggtitle("RAM vs Price Range") +
  theme(plot.title = element_text(hjust = 0.5))

violin
```



(f)

```
# 1f: stacked bar plot
bar_plot <- base_plot +
  geom_bar(
    mapping = aes(
        x = round(log(ram, 2)),
        fill = price_range),
    position = "stack") +
  ggtitle("Bar Plot of RAM based on Price Range.") +
  theme(plot.title = element_text(hjust = 0.5))

bar_plot
```



Problem 2

(a)

```
# Get data:
df <- UScereal

# 2a: replace levels of the factor
# variable mfr to their full names
levels(df$mfr) <- c(
  "General Mills",
  "Kellogs",
  "Nabisco",
  "Post",
  "Quaker Oats",
  "Ralston Purina")
```

(b)

```
# 2b: turn variable shelf to a factor variable

df <- df %>%
  mutate(shelf = factor(shelf))
```

(c)

```
# 2c: Create new variable Product for the product name
rows <- rownames(df)
df <- df %>%
  mutate(product = rows)
```

(d)

```
# d: Calculate the Pearson Correlation coefficient between calories and each seven nutrition facts

pearson <- lapply(
  df[
    ,
    c(
      "protein",
      "fat",
      "sodium",
      "fibre",
      "carbo",
      "sugars",
      "potassium"
    )
  ],
  FUN = cor,
  x = df$calories,
  method = "pearson"
) %>%
  data.frame()

rownames(pearson) <- "calories"
table <- knitr::kable(pearson,
  caption = "Pearson Coeffiecients",
  align = "lcccc"
)
```

‘

Table 1: Pearson Coeffiecients

	protein	fat	sodium	fibre	carbo	sugars	potassium
calories	0.7060105	0.5901757	0.5286552	0.3882179	0.7887227	0.4952942	0.4765955

‘

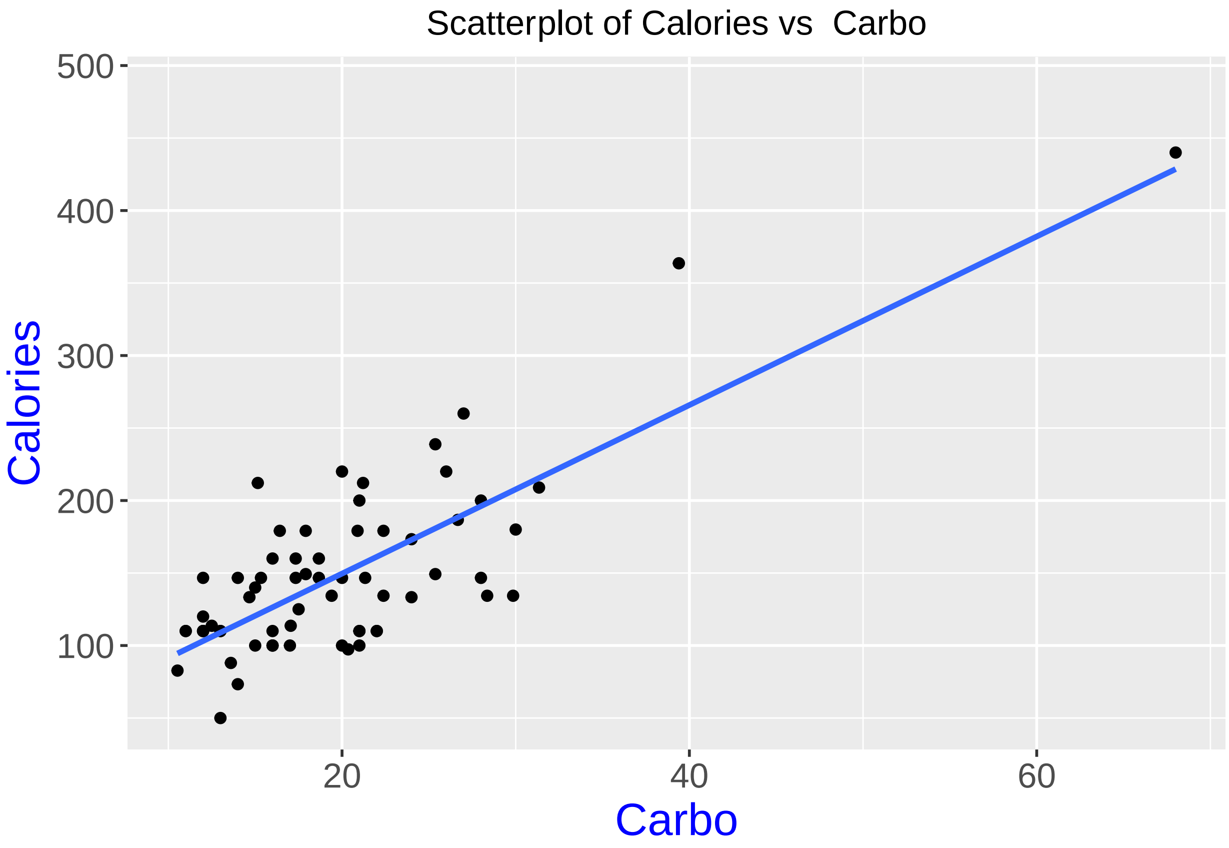
(e)

```
# 2e: make a bar plot of the resulting correlations in (a)
# and arrange the nutrition facts in decreasing order in terms of
#  their correlation with calories
sorted_pearson <- pearson %>%
  t() %>%
  data.frame() %>%
  arrange(desc(calories))

# which nutrition fact has the highest values: ans: carbo
max_pearson <- c(pearson[which(max(pearson) == pearson)])
```

The nutrition fact that has the highest values is `carbo` with value of `0.788722682963849`

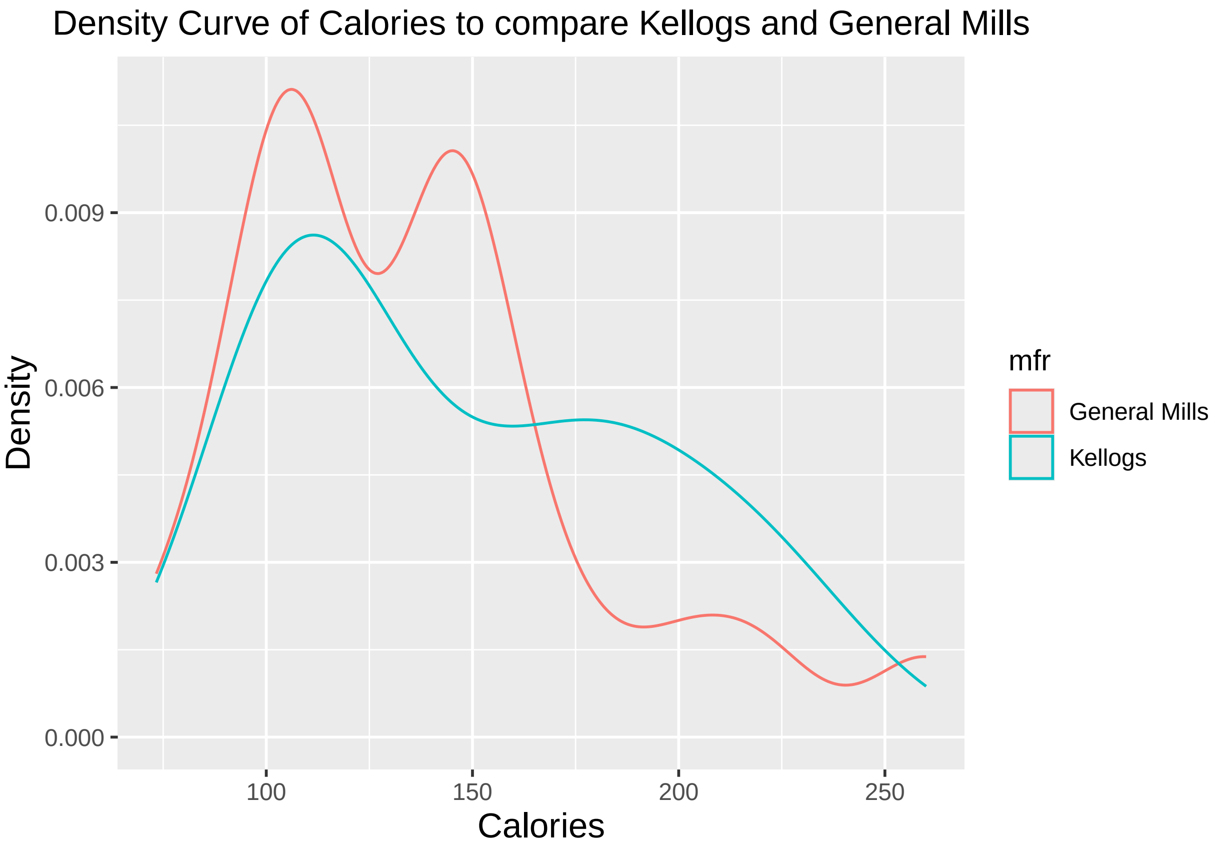
(f)

```r
# 2f: scatter plot where y represents calories and x represents
# the nutrition fact with the largest pearson correlation
# coefficient to calories

scatter_plot_w_trend <- ggplot(
  data = df,
  mapping = aes(
    x =
      df[, names(max_pearson)],
    y = calories)) +
  geom_point() +
  geom_smooth(
    method = "lm",
    formula = "y ~ x",
    fill = NA) +
  ggtitle(paste("Scatterplot of Calories vs ",
    capitalize(names(max_pearson)))) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.title = element_text(
      colour = "BLUE",
      size = rel(1.5)),
    axis.text = element_text(size = rel(1.2))) +
  labs(
    x = capitalize(names(max_pearson)),
    y = "Calories")

scatter_plot_w_trend
```



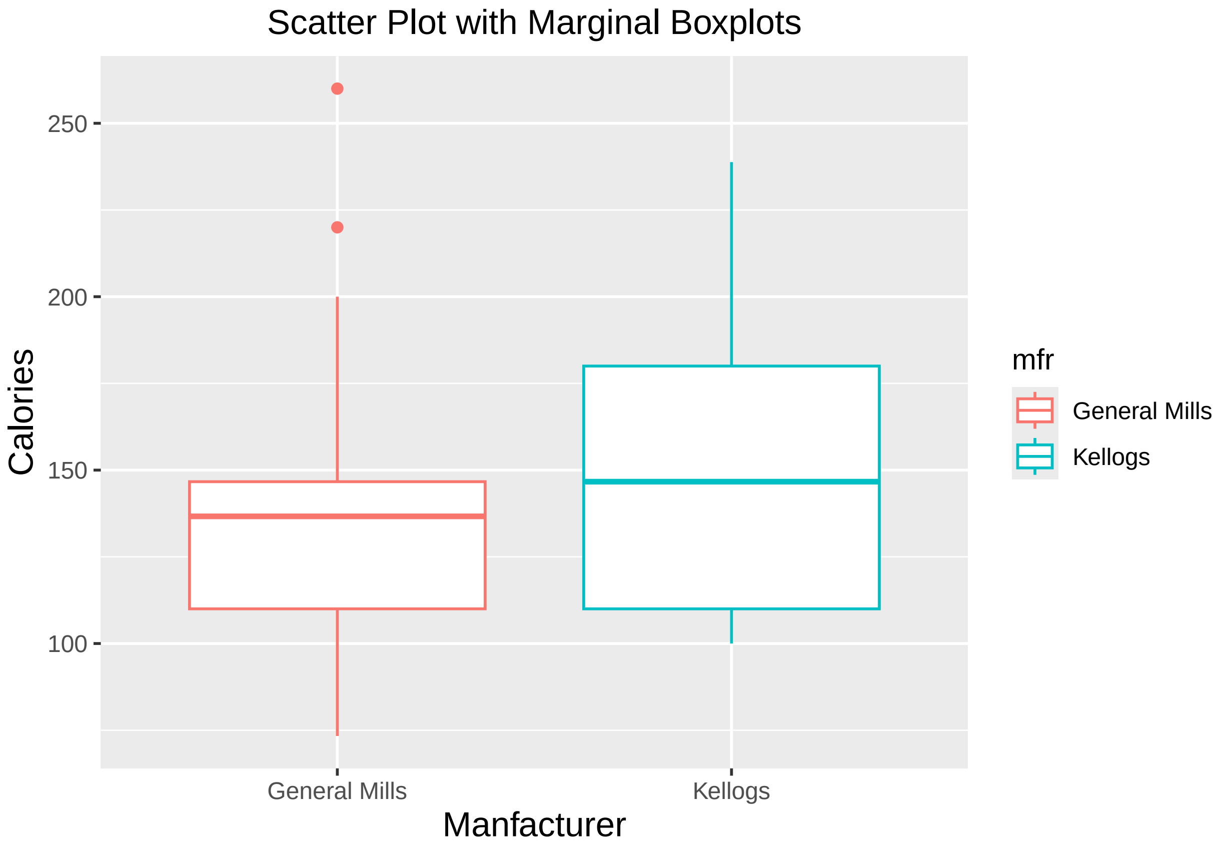
(g)

```
# 2g: Plot a density curve of Calories
density_curve <- df %>%
  filter(mfr %in% c("General Mills", "Kellogs")) %>%
  ggplot(
    data = .,
    mapping = aes(x = calories, color = mfr)
  ) +
  geom_density() +
  ggtitle("Density Curve of Calories to compare Kellogs and General Mills") +
  theme(
    plot.title = element_text(hjust = 0.5, size = rel(1.2)),
    axis.title = element_text(size = rel(1.2))
  ) +
  labs(
    x = "Calories",
    y = "Density"
  )
density_curve
```

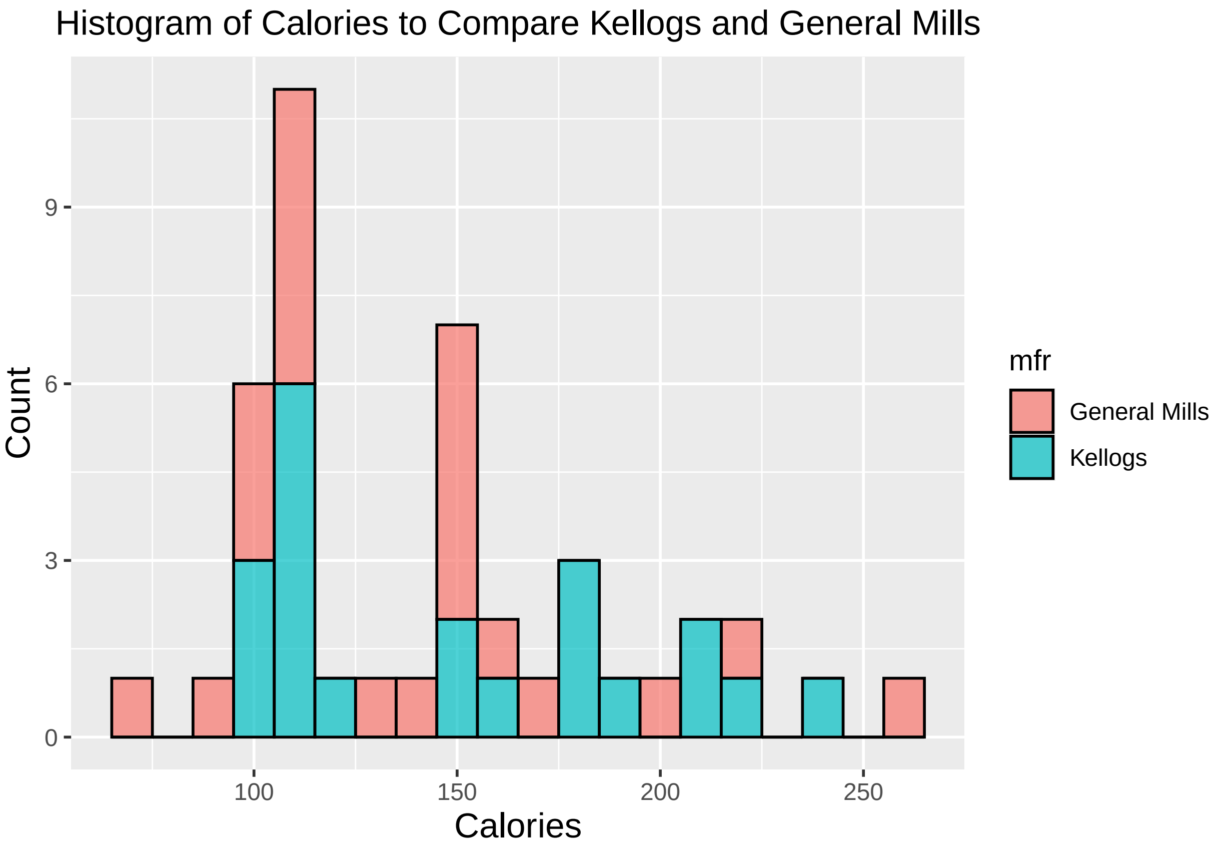


(h)

```
# 2h: Plot Calories and Manufactures in a Histogram Plot
# to see the distribution of calories for General Mills and Kellogs:
histogram_plot <- df %>%
  filter(mfr %in% c("General Mills", "Kellogs")) %>%
  ggplot(mapping = aes(
    x = calories,
    fill = mfr
  )) +
  geom_histogram(
    binwidth = 10,
    position = "stack",
    alpha = 0.7, color = "black"
  ) + # Adjust binwidth as needed
  ggtitle("Histogram of Calories to Compare Kellogs and General Mills") +
  theme(
    plot.title = element_text(hjust = 0.5, size = rel(1.2)),
    axis.title = element_text(size = rel(1.2))
  ) +
  labs(
    x = "Calories",
    y = "Count"
  )

# Create a Box plot
box_plot <- df %>%
  filter(mfr %in% c("General Mills", "Kellogs")) %>%
  ggplot(mapping = aes(x = mfr, y = calories, color = mfr)) +
  geom_boxplot() +
  ggtitle("Scatter Plot with Marginal Boxplots") +
  theme(
    plot.title = element_text(hjust = 0.5, size = rel(1.2)),
    axis.title = element_text(size = rel(1.2))
  ) +
  labs(x = "Manfacturer", y = "Calories")

# Display the plot
print(box_plot)
```

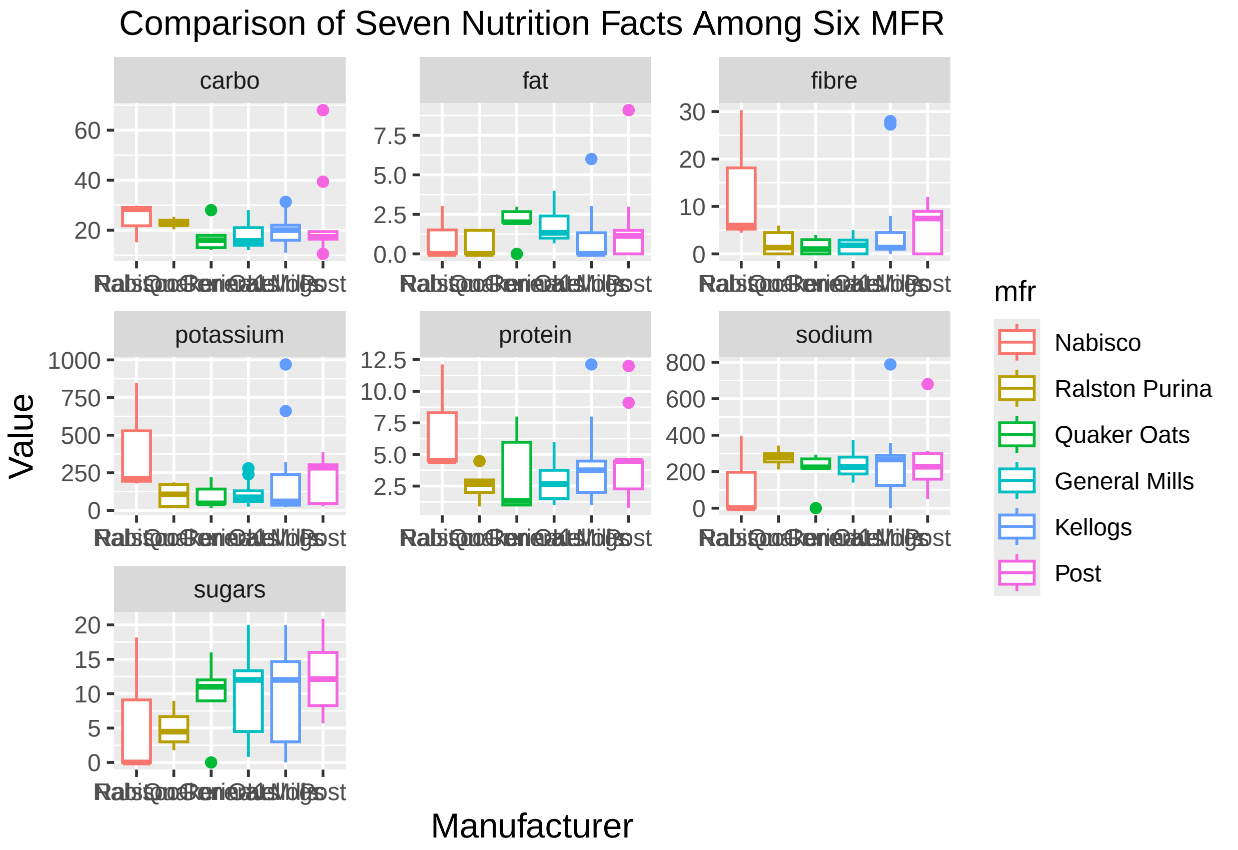


```
print(histogram_plot)
```



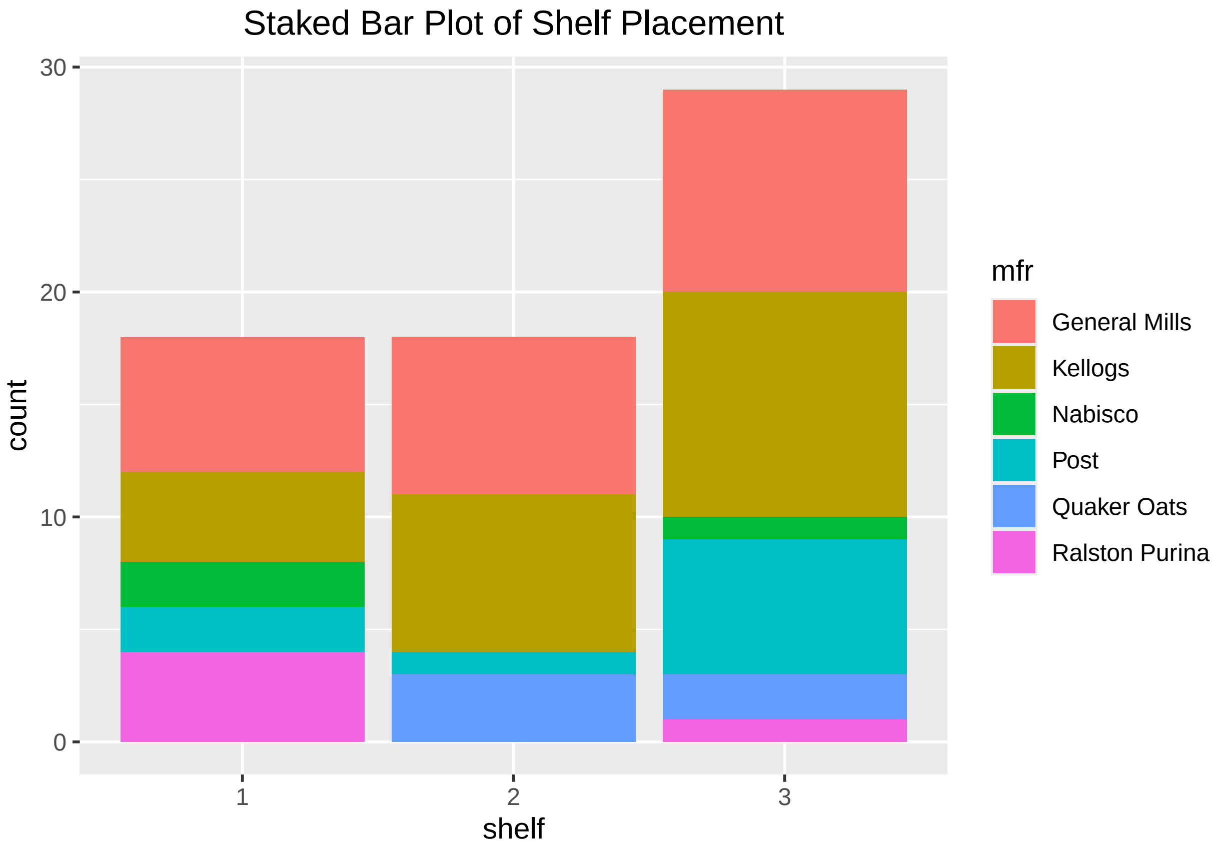
(i)

```
# 2i: Seven side-by-side Boxplots to compare each of the seven nutrition facts among the six mfr:

long_df <- df %>%
  select(mfr, calories, protein, fat, sodium, fibre, carbo, sugars, potassium) %>%
  pivot_longer(cols = c(protein, fat, sodium, fibre, carbo, sugars, potassium),
    names_to = "Seven_Nutri_Facts",
    values_to = "value")

# To order the boxplots according to the median value,
median_values <- long_df %>%
  group_by(mfr) %>%
  summarise(median_value = median(value)) %>%
  arrange(median_value)

# We can re-order the mfr factor levels based on
# the median values calculated
long_df$mfr <- factor(long_df$mfr, levels = median_values$mfr)

# Create the boxplot with the reordered mfr levels
boxplot_nutrition <- ggplot(long_df,
  aes(x = mfr, y = value, color = mfr)) +
  geom_boxplot() +
  facet_wrap(~Seven_Nutri_Facts, scales = "free") +
  theme(plot.title = element_text(hjust = 0.5, size = rel(1.2)),
    axis.title = element_text(size = rel(1.2))) +
  labs(title = "Comparison of Seven Nutrition Facts Among Six MFR",
    x = "Manufacturer",
    y = "Value")

boxplot_nutrition
```



(j)

```
# 2j : Stacked Bar plot to show the relationship between manufacturer
# and shelf placement:
stacked_barplot <- df %>%
  select(mfr, shelf) %>%
  ggplot(
    data = .,
    mapping = aes(
      x = shelf,
      fill = mfr),
    color = "black"
  ) +
  geom_bar(position = "stack") +
  ggtitle("Staked Bar Plot of Shelf Placement") +
  theme(plot.title = element_text(hjust = 0.5))

stacked_barplot
```


Staked Bar Plot of Shelf Placement
30
20
10
0
count
1
2
3
shelf
mfr
General Mills
Kellogs
Nabisco
Post
Quaker Oats
Ralston Purina