**Introduction**

The chosen data-set for this project is “**The Great American Coffee Taste Test**” from TidyTuesday (Week 21, 2024). In October 2023, “world champion barista” James Hoffmann and coffee company Cometeer held the “Great American Coffee Taste Test” on YouTube, during which viewers were asked to fill out a survey about 4 coffees they ordered from Cometeer for the tasting. Based on the research conducted by the NIH: National Institutes of Health, which found that 75% of U.S. adults over the age of 20 drink coffee, with 49% drinking it daily, coffee consumption is clearly a significant part of American culture and daily life. This widespread prevalence of coffee consumption makes it an ideal subject for data analysis, as it offers insights into behavior patterns that affect a large portion of the population.

Understanding the coffee consumption habits of individuals offers valuable insights into consumer behavior, lifestyle preferences, and spending patterns. Coffee is an integral part of daily life for many people around the world, and studying why people choose to drink coffee, their preferred roast, where they typically consume it, and how much they spend can reveal a lot about personal tastes, habits, and economic factors. These factors could also provide valuable information for businesses in the coffee industry, helping them tailor their products and marketing strategies more effectively. People from different age groups or genders may gravitate toward specific coffee types (e.g., espresso, cappuccino, black coffee) and may prefer different levels of customization when it comes to additives. By exploring how age and gender impact these preferences, we can gain insight into how demographic characteristics shape coffee consumption patterns.

**Question 1: Why do people choose to drink coffee, what roast preference are common, where do they typically drink it, and how much do they spend on coffee overall?**

**Introduction**

We are interested in this question because analyzing coffee consumption habits provides valuable insights into consumer behavior, lifestyle choices, and spending trends. Coffee plays a significant role in the daily routines of people worldwide, and examining why individuals drink coffee, their preferred roast type, typical consumption locations, and spending amounts can shed light on personal preferences, habits, and economic influences. These insights could also be beneficial for businesses in the coffee industry, enabling them to better tailor their products and marketing strategies.

To explore this question, the essential attributes in the coffee dataset—total\_spend, why\_drink, where\_drink, and roast\_level—will offer a comprehensive view of coffee consumption patterns, revealing motivations, preferences, spending behavior, and lifestyle tendencies. We plan to use bar plots and box plots to visualize categorical data, such as reasons for drinking coffee (why\_drink) and coffee roast preferences (roast\_level), as well as to display the spending distribution by location (where\_drink).

**Approach**

In order to answer our first question we choose a bar plot and heatmap.

The first part of question 1 focuses on understanding why people choose to drink coffee and identifying common roast preferences. In this study, we aim to illustrate relationships between multiple variables in a single plot. We selected a bar plot for this purpose due to its simplicity, flexibility, and ability to clearly highlight differences across categories. A bar plot offers a straightforward yet powerful way to present complex information in an organized and easily understandable format.

To address the second part of question 1, we used a heatmap. A heatmap is a valuable data visualization tool that employs color gradients to depict the intensity or magnitude of values within a dataset. This visual approach makes it easy to recognize complex data patterns and identify areas of high or low concentration. Here, we aim to discover the most popular locations for coffee consumption and how much people typically spend at each location.

**Analysis**

Question No.1- Part 1

Code

```{r}

# Remove rows with NA values in the important columns

coffee\_sankey\_clean <- na.omit(coffee\_survey[, c("age", "gender", "favorite", "additions")])

View(coffee\_sankey\_clean)

```

```{r}

coffee\_Q1 <- coffee\_survey[, c("why\_drink", "roast\_level", "total\_spend", "where\_drink")]

```

```{r}

na.omit(coffee\_Q1)

```

```{r}

coffee\_Q1 <- na.omit(coffee\_Q1[, c("why\_drink", "roast\_level", "total\_spend", "where\_drink")])

```

```{r}

View(coffee\_Q1)

```

```{r}

library(dplyr)

library(tidyr)

library(ggplot2)

# Calculate total percentages for each reason

reason\_counts <- coffee\_Q1 %>%

mutate(

tastes\_good = grepl("It tastes good", why\_drink),

ritual = grepl("I need the ritual", why\_drink),

caffeine = grepl("I need the caffeine", why\_drink),

To\_Poop = grepl("It makes me go to the bathroom", why\_drink)

) %>%

pivot\_longer(cols = tastes\_good:To\_Poop, names\_to = "reason", values\_to = "selected") %>%

filter(selected) %>%

count(reason) %>%

mutate(percentage = (n / nrow(coffee\_Q1)) \* 100)

# Calculate percentages for each roast level within each reason

roast\_level\_counts <- coffee\_Q1 %>%

mutate(

tastes\_good = grepl("It tastes good", why\_drink),

ritual = grepl("I need the ritual", why\_drink),

caffeine = grepl("I need the caffeine", why\_drink),

To\_Poop = grepl("It makes me go to the bathroom", why\_drink)

) %>%

pivot\_longer(cols = tastes\_good:To\_Poop, names\_to = "reason", values\_to = "selected") %>%

filter(selected) %>%

filter(roast\_level %in% c("Light", "Medium", "Dark")) %>%

count(reason, roast\_level) %>%

group\_by(reason) %>%

mutate(percentage = (n / sum(n)) \* 100)

# Define roast colors

roast\_colors <- c("Light" = "wheat", "Medium" = "wheat4", "Dark" = "chocolate4")

# Plot original bars for reasons

p1 <- ggplot(reason\_counts, aes(x = reorder(reason, -percentage), y = percentage)) +

geom\_bar(stat = "identity", fill = "skyblue") +

labs(

title = "Percentage of Reasons for Drinking Coffee (with Roast Level Breakdown)",

x = "Reason",

y = "Percentage (%)"

) +

scale\_y\_continuous(breaks = seq(0, 100, 10), labels = scales::percent\_format(scale = 1)) +

theme\_minimal() +

coord\_flip()

# Add the roast level bars just above the original bars

p2 <- ggplot(roast\_level\_counts, aes(x = reorder(reason, +percentage), y = percentage, fill = roast\_level)) +

geom\_bar(stat = "identity", position = "stack", width = 0.4) + # Narrow bars for roast levels

scale\_fill\_manual(values = roast\_colors) +

theme\_minimal() +

coord\_flip() +

theme(axis.title.x = element\_blank(), axis.text.x = element\_blank(), axis.ticks.x = element\_blank())

# Combine the two plots by overlaying them

p1 + geom\_bar(data = roast\_level\_counts, aes(y = percentage, fill = roast\_level),

stat = "identity", position = position\_dodge(width = 1), width = 0.4) +

theme\_minimal() +

scale\_fill\_manual(values = roast\_colors) +

theme(legend.position = "bottom")

**Part-2**

**Code**

**```{r}**

**coffee\_survey <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2024/2024-05-14/coffee\_survey.csv')**

**head(coffee\_survey)**

**```**

**```{r}**

**coffee\_Q1 <- na.omit(coffee\_survey[, c("why\_drink", "roast\_level", "total\_spend", "where\_drink")])**

**coffee\_Q1**

**```**

**```{r}**

**# remove columns with NA**

**coffee\_Q1$roast\_level[coffee\_Q1$roast\_level == "Blonde"] <- NA**

**coffee\_Q1$roast\_level[coffee\_Q1$roast\_level == "Italian"] <- NA**

**coffee\_Q1$roast\_level[coffee\_Q1$roast\_level == "Nordic"] <- NA**

**coffee\_Q1$roast\_level[coffee\_Q1$roast\_level == "French"] <- NA**

**coffee\_Q1 <- coffee\_Q1[!is.na(coffee\_Q1$roast\_level), ]**

**coffee\_Q1**

**```**

**```{r}**

**library(tidyr)**

**# Separate each reason into individual rows**

**Q1\_separated <- coffee\_Q1 %>%**

**separate\_rows(why\_drink, sep = ",\\s\*") %>%**

**separate\_rows(where\_drink, sep = ",\\s\*")**

**Q1\_separated**

**```**

**```{r}**

**# Count occurrences of each total\_spend range per location**

**data\_Q1\_part2 <- Q1\_separated %>%**

**count(where\_drink, total\_spend)**

**```**

**```{r}**

**library(dplyr)**

**# Remove "None of these" from the where\_drink column**

**data\_Q1\_part3 <- data\_Q1\_part2 %>%**

**filter(where\_drink != "None of these")**

**View(data\_Q1\_part3)**

**```**

**```{r}**

**a <- ggplot(data\_Q1\_part3, aes(x = where\_drink, y = total\_spend, fill = n)) +**

**geom\_tile(color = "white") +**

**geom\_text(aes(label = n), color = "black") +**

**scale\_fill\_gradient(low = "lightyellow", high = "darkred", name = "Count") +**

**labs(title = "Distribution of Spending by Location",**

**x = "Where Drink",**

**y = "Total Spend Range") +**

**theme\_minimal() +**

**theme(panel.grid.major = element\_blank(),**

**panel.grid.minor = element\_blank())**

**a**

**#ggsave("heatmap.png",plot = a, width = 10, height = 6, dpi = 300)**

**```**

```**Discussion**

We can effectively address our first question—why people consume coffee and which roast preferences are the most popular—using a bar plot. The "why\_drink" column consists of four categories: taste good, caffeine, ritual, and to poop. More than 35% of respondents enjoy light or medium coffee, while only about 15% prefer dark coffee. The data visualization analysis indicates that a significant 94% of people drink coffee primarily for its taste. Within this group, 50% prefer light roast, and 40% opt for medium roast, with just 10% choosing dark coffee. This trend suggests that light or medium roasts are favored over dark coffee. Additionally, around 53-56% of people consume coffee as a ritual or for the caffeine boost, with these preferences also leaning towards light or medium flavors. Interestingly, only 13% drink coffee for the purpose of aiding digestion.

In our next analysis, we focus on where people typically enjoy their coffee and how much they spend. To explore this, we created a heatmap comparing "where\_drink" with total spending ranges. The heatmap reveals that individuals prefer to have their coffee at home, more so than at the office, cafes, or on the go. The results also show that those who drink coffee at home generally spend between $20 and $40. Conversely, the most popular choice for coffee consumption is while on the go, with people willing to spend between $20 and $80 for their coffee.

**Question No.2: How do age and gender impact preferences for various types of coffee and the use of additives?**

**Introduction**

We are particularly interested in this question because age and gender are significant demographic factors that often affect consumer behavior, especially regarding preferences for various types of coffee and the use of additives such as sugar, milk, and cream. Individuals from different age groups or genders may favor specific coffee types (like espresso, cappuccino, or black coffee) and may have varying preferences for customization regarding additives. By investigating how age and gender influence these preferences, we can gain valuable insights into how demographic characteristics shape coffee consumption patterns.

To tackle this question, the key components of the dataset include four columns: ‘favorite,’ ‘age,’ ‘gender,’ and ‘additions.’ These elements will facilitate a comprehensive analysis of how coffee preferences (reflected in the “favorite” variable) differ across various age groups and between genders, helping to reveal meaningful patterns in consumer behavior.

**Approach**

We approach the second question with a circular bar plot and a bubble plot.

A circular bar plot is a variation of the traditional bar plot where bars are arranged in a circular layout, making it useful for representing cyclical or categorical data. This format allows for an aesthetically engaging display of information, especially when comparing values across multiple categories or groups. This plot is best suited for data that can be clearly understood in a circular format and may not be ideal for precise value comparisons.

bubbleplot

**Analysis**

**Code**

**```{r}**

**# Define color palette for different ages or genders**

**age\_palette <- c("#3B2A1E", "#E3B89A", "#D9C9B0", "#BFA7A1", "#D2B48C", "#A57B52", "#8B4513", "#4B3D33")**

**# Circular bar plot with fill by age**

**ggplot(Q2\_aggregated, aes(x = favorite, y = proportion, fill = age)) +**

**geom\_bar(stat = "identity", position = "stack", width = 0.9) +**

**coord\_polar(start = 0) +**

**scale\_fill\_manual(values = age\_palette, name = "Age") +**

**labs(title = "Preferences for Coffee type by age",**

**x = "", y = "") +**

**theme\_minimal() +**

**theme(**

**axis.text.x = element\_text(size = 8, color = "black", face = "bold"),**

**axis.title = element\_blank(),**

**panel.grid = element\_blank(),**

**legend.position = "right"**

**)**

Part 1-

A diagram of coffee type

Description automatically generated

Part – 2

Code

q <- ggplot(Q2\_aggregated, aes(x = age, y = additions, size = count, color = gender)) +

geom\_point(alpha = 1.5) +

scale\_color\_manual(values = c("Female" = "darkred", "Male" = "#D2B48C", "Non-binary" = "navy"))+

scale\_size\_continuous(range = c(5, 20)) +

labs(title = "Bubble Plot of Age, Addition, and Gender",

x = "Age Group",

y = "Additives",

size = "Count")+

theme(axis.text.y = element\_text(size = 10, face = "bold"),

axis.text.x = element\_text(size = 5, face = "bold"),

legend.text = element\_text(size = 10), # Font size for legend text

legend.title = element\_text(size = 12)) # Font size for legend title)

q

A screenshot of a graph

Description automatically generated

Discussion

"Regular drip coffee" and "Blended drinks (e.g., Frappuccino)" are popular across multiple age groups, including older adults (represented by darker shades). For instance, individuals aged 55-64 show a strong preference for these types, as indicated by the darker sections on those bars. On the other hand, "Iced coffee" and "Espresso" are preferred by younger age groups, with lighter-colored bars representing those in the 18-24 and 25-34 age ranges. Interestingly, some coffee types, like "Mocha" and "Cortado," have relatively low popularity across all age groups, as indicated by the short bar lengths. This analysis highlights that older individuals tend to prefer "Regular drip coffee" and "Blended drinks," while younger demographics lean more towards "Iced coffee" and "Espresso," reflecting generational differences in coffee preferences.The bubble plot provides an analysis of coffee additive preferences across different age groups and genders. Larger bubbles indicate higher counts of individuals who prefer a particular additive. The most popular additive choices include milk and black coffee, especially among males (shown in light tan) and non-binary individuals (in blue) within the 25–34 and 35–44 age groups. Sugar or sweeteners and dairy alternatives are less commonly preferred, with a few notable cases among younger and older age groups. Females (represented by red) show more diverse preferences across additives but with generally lower counts. This visualization highlights age and gender trends in coffee customization preferences, particularly the tendency of younger adults to prefer black coffee and milk as additives.

The bubble plot provides an analysis of coffee additive preferences across different age groups and genders. Larger bubbles indicate higher counts of individuals who prefer a particular additive. The most popular additive choices include milk and black coffee, especially among males (shown in light tan) and non-binary individuals (in blue) within the 25–34 and 35–44 age groups. Sugar or sweeteners and dairy alternatives are less commonly preferred, with a few notable cases among younger and older age groups. Females (represented by red) show more diverse preferences across additives but with generally lower counts. This visualization highlights age and gender trends in coffee customization preferences, particularly the tendency of younger adults to prefer black coffee and milk as additives.

The bubble plot provides an analysis of coffee additive preferences across different age groups and genders. Larger bubbles indicate higher counts of individuals who prefer a particular additive. The most popular additive choices include milk and black coffee, especially among males (shown in light tan) and non-binary individuals (in blue) within the 25–34 and 35–44 age groups. Sugar or sweeteners and dairy alternatives are less commonly preferred, with a few notable cases among younger and older age groups. Females (represented by red) show more diverse preferences across additives but with generally lower counts. This visualization highlights age and gender trends in coffee customization preferences, particularly the tendency of younger adults to prefer black coffee and milk as additives.