Diwali Sales Dataset

Dhanachote_W

2023-11-06

Introduction

Hi, my name is Dhanachote. I'm working on a mini-project to analyze and practice the R language, build my data visualization skills, and create a report using Rmarkdown. For this project, I'm using the 'Diwali Sales' dataset sourced from Kaggle, which you can find at the link below. Therefore, I cleaned the dataset in Excel before importing it into R programming.

Reference

Diwali Sales Dataset Source from Kaggle

Table

- Part I: Questions and Data Visualization
  - Question 1: What is the most popular product category during Diwali sales?
  - Question 2: Which customer group has the highest total purchase amount?
  - Question 3: Which state has the highest total customer purchase amount?
  - Question 4 : State / Zone

    - Question 4.1: Count the number of orders for each zone and calculate the average order size for each zone.
    - Question 4.2: Calculate the total number of orders across all zones.
    - Question 4.3: For each zone, identify the product category with the highest number of orders.
    - Question 4.4: Identify the product category with the highest total sales amount in the each zone.
  - Question 5: Generation
    - Question 5.1: Which customer has spent the most money? To which generation do they belong?
    - Question 5.2: Calculate the average order amount for Generation Z customers.
    - Question 5.3: Determine the number of male and female customers
    - Question 5.4: Identify the number of customers belonging to each generation.
  - Question 6: Identify the percentage amount for each occupation.
  - Question 7: What is the total amount for each age group?
- Part II: Machine Learning

Part I Questions and Data Visualization

In this part, I asking a serveral questions to understanding in the dataset. I use library such as `tidyverse` ,`readr`, `dplyr`, `ggplot2`, and `caret` before query the dataset.

Install packages and download library

```
## Library
library(readr)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(caret)
```

```
## import read.csv()

diwali_sales <- tibble(read.csv("Diwali_Sales_Data_main.csv"))
```

Import dataset from `.csv`

```
## check NA

diwali_sales %>%
  complete.cases()%>%
  mean()
```

Check NA

```
## [1] 0
```

I use `complete.cases()` to check for missing values and then use `mean()` to calculate the logical mean (TRUE = 1, FALSE = 0). If the mean value is 0, it means that the dataset is complete and has no missing values.

Tip: You can check logical values by using the `as.logical()` function. For example, `as.logical(0)` evaluates to FALSE, while 1 evaluates to TRUE.

```
## select column

ds_cl <- diwali_sales %>%
  select(User_ID,
         Cust_name,
         Product_ID,
         Product_Category,
         Gender,
         Age,
         Marital_Status,
         State,
         Zone,
         Occupation,
         Orders,
         Amount) %>%
  drop_na()
```

Prepare the dataset

Question 1: What is the most popular product category during Diwali sales?

```
q1 <- diwali_sales %>%
  group_by(Product_Category) %>%
  summarise(Total_amount = sum(Orders * Amount)) %>%
  arrange(desc(Total_amount)) %>%
  head(5)

print(q1)
```

```
## # A tibble: 5 x 2
##   Product_Category      Total_amount
##   <chr>                        <dbl>
## 1 Food                     83591272.
## 2 Clothing & Apparel       41164094
## 3 Electronics & Gadgets    39315276
## 4 Footwear & Shoes         38731504.
## 5 Furniture                13660130.
```

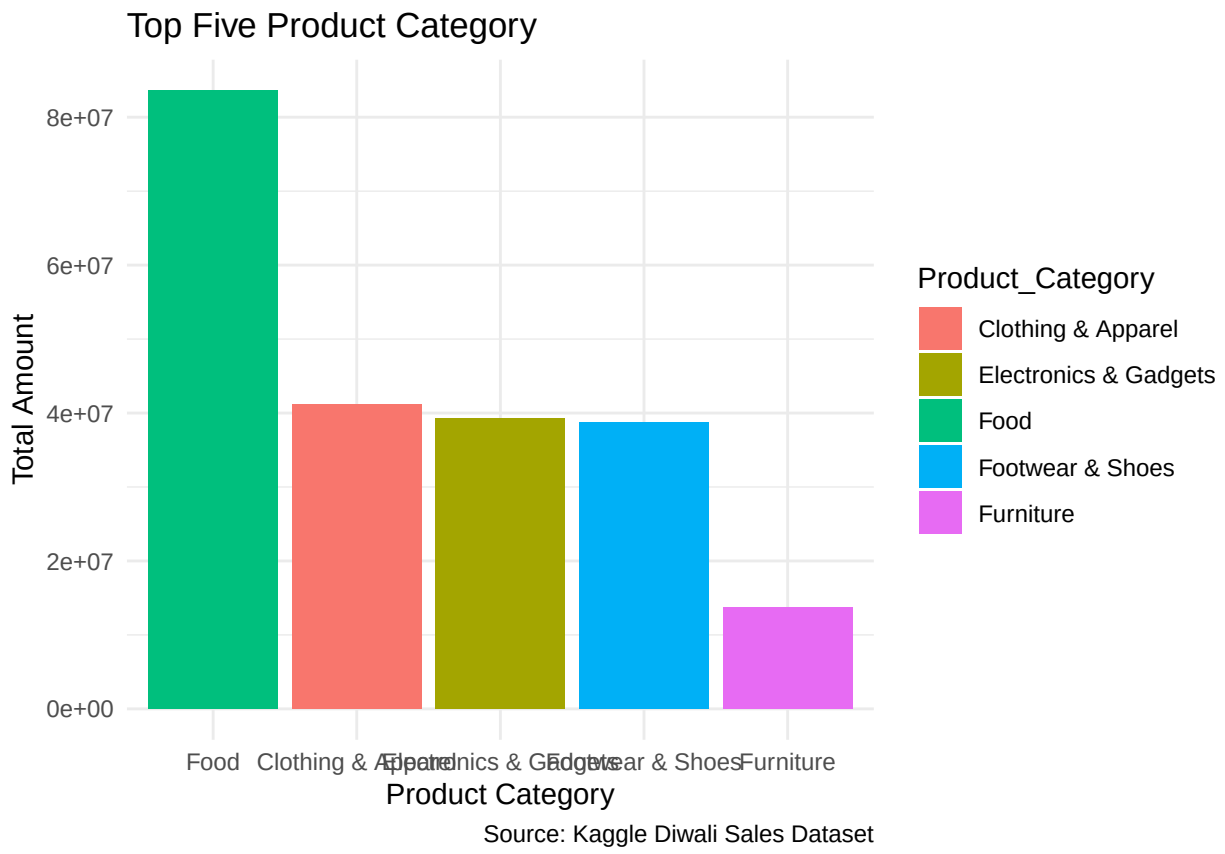
```r
message("The most product category total amount is: ",
        head(q1$Product_Category,1),
        " "  ,
        "and",
        " ",
        "The total amount is: $ ",
        head(q1$Total_amount,1))
```

```
## The most product category total amount is: Food and The total amount is: $ 83591271.5
```

```r
ds_cl %>%
  group_by(Product_Category) %>%
  summarise(
    total_amount = sum(Amount * Orders)
  ) %>%
  top_n(5) %>%
  arrange(desc(total_amount)) %>%
  ggplot(aes(reorder(Product_Category, -total_amount), total_amount, fill = Product_Category)) +
  geom_col() +
  theme_minimal() +
  labs(
    title = "Top Five Product Category",
    x = "Product Category",
    y = "Total Amount",
    caption = "Source: Kaggle Diwali Sales Dataset"
  )
```

Plot 1:

```
## Selecting by total_amount
```



Question 2: Which customer has the highest total purchase amount?

```
## Question2 : Who is the most spending amount

q2 <- diwali_sales %>%
  group_by(Cust_name) %>%
  summarise(
    Total_amount = sum(Orders * Amount)
  ) %>%
  arrange(desc(Total_amount)) %>%
  head(5)

print(q2)
```

```
## # A tibble: 5 x 2
##   Cust_name Total_amount
##   <chr>            <dbl>
## 1 Vishakha        972730
## 2 Alejandro       718053
## 3 Vasudev         698923
## 4 Sudevi          686455
## 5 Lalita          683635
```

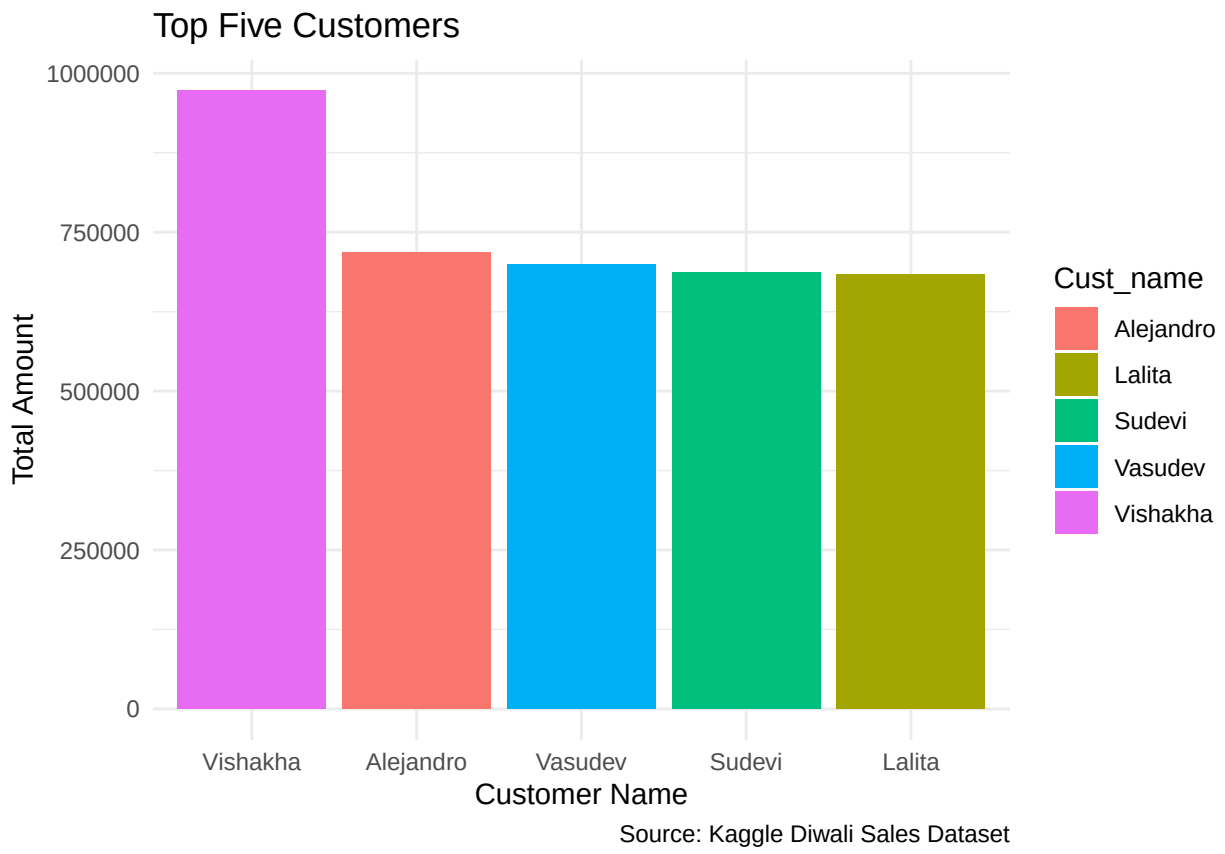
```r
message(head(q2$Cust_name,1),
        " ",
        "is the most spending amount in $ ",
        head(q2$Total_amount,1),
        " product category")
```

```
## Vishakha is the most spending amount in $ 972730 product category
```

```r
ds_cl %>%
  group_by(Cust_name) %>%
  summarise(
    total_amount = sum(Amount * Orders)
  ) %>%
  top_n(5) %>%
  ggplot(aes(reorder(Cust_name, -total_amount), total_amount, fill = Cust_name)) +
  geom_col() +
  theme_minimal() +
  labs(
    title = "Top Five Customers",
    x = "Customer Name",
    y = "Total Amount",
    caption = "Source: Kaggle Diwali Sales Dataset"
  )
```

Plot 2

```
## Selecting by total_amount
```



Question 3: Which state has the highest total customer purchase amount?

```
q3 <- diwali_sales%>%
    select(Cust_name,
           Gender,
           State,
           Zone,
           Orders,
           Amount) %>%
    mutate(Total_amount = Orders * Amount) %>%
    filter(Total_amount > 50000) %>%
    arrange(desc(Total_amount)) %>%
    head(10)

print(q3)
```

```
## # A tibble: 10 x 7
##    Cust_name Gender State            Zone     Orders Amount Total_amount
##    <chr>     <chr>  <chr>            <chr>     <int>  <dbl>        <dbl>
##  1 Balk      F      Uttar Pradesh    Central       4 23841        95364
##  2 Ginny     F      Andhra Pradesh   Southern      4 23800.       95200.
```

```
##  3 Vasudev   M      Andhra Pradesh   Southern      4 23718        94872
##  4 Ellis     F      Andhra Pradesh   Southern      4 23546        94184
##  5 Mahima    F      Andhra Pradesh   Southern      4 23451        93804
##  6 Daniels   F      Andhra Pradesh   Southern      4 23302        93208
##  7 Mike      M      Himachal Pradesh Northern      4 23267        93068
##  8 Dean      F      Andhra Pradesh   Southern      4 23252        93008
##  9 Zypern    M      Andhra Pradesh   Southern      4 23239        92956
## 10 Abhijit   F      Andhra Pradesh   Southern      4 23066        92264
```

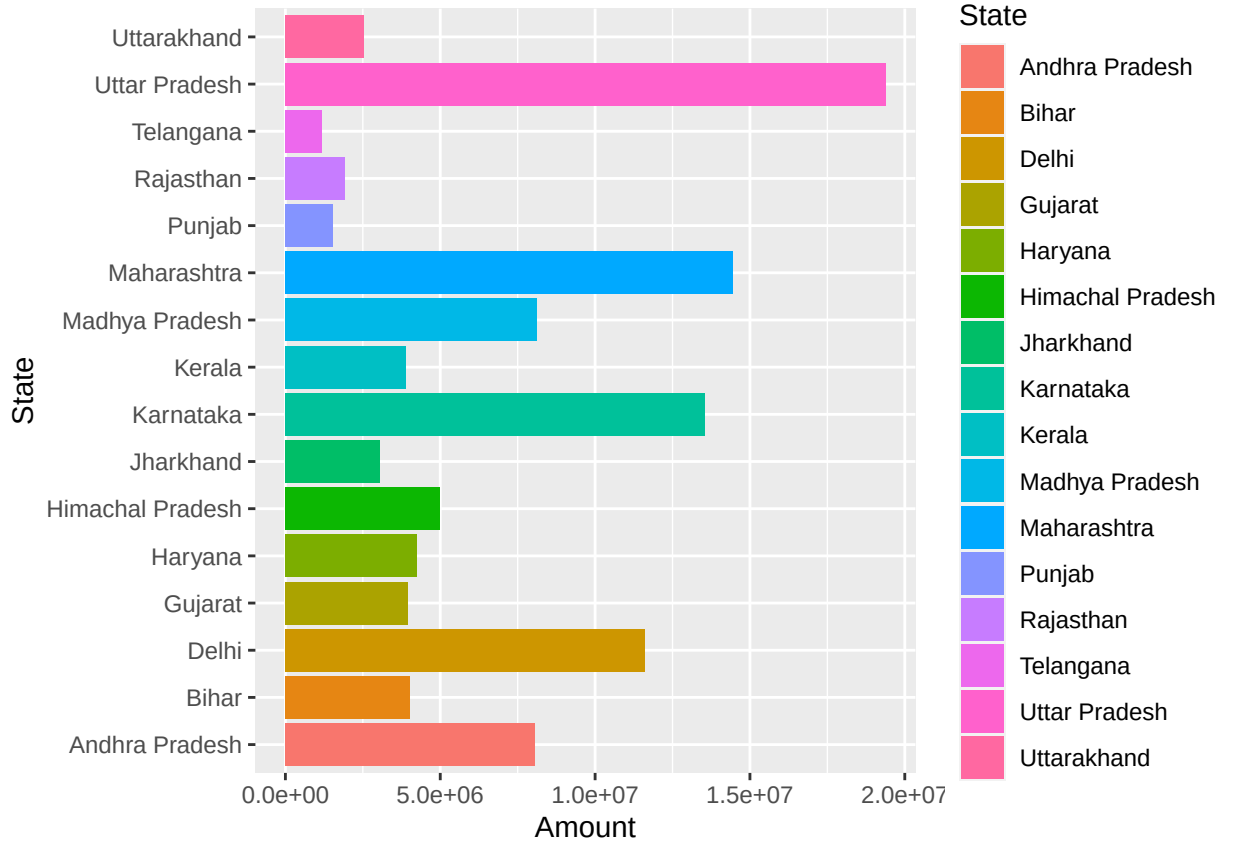
```r
message("Which state is the most user spending is: ",
        head(q3$State,1))
```

```
## Which state is the most user spending is: Uttar Pradesh
```

Question 4 : State / Zone

In this question, I provided a chart that explains which state has the highest total amount before starting the questions.

```r
## Plot 4: Relationship with State and Amount

ds_cl %>%
  group_by(State) %>%
ggplot(data = ds_cl,
       mapping = aes(x = Amount, y = State, fill = State)) +
  geom_col()
```




```
# count zone
diwali_sales %>%
  count(Zone)
```

Question 4.1: Count the number of orders for each zone and calculate the average order size for each zone.

```
## # A tibble: 5 x 2
##   Zone          n
##   <chr>     <int>
## 1 Central    4296
## 2 Eastern     814
## 3 Northern   1491
## 4 Southern   2695
## 5 Western    1955
```

```
avg_orders <- mean(diwali_sales$Orders)

  ## total and percentage avg orders

diwali_sales %>%
  select(Zone,
```

```
        Product_Category,
        Orders,
        Amount) %>%
  filter(Orders >= avg_orders) %>%
  group_by(Zone) %>%
  summarise(total_avg_orders = n()) %>%
  mutate(pct_avg_orders =  c(total_avg_orders / sum(total_avg_orders)) * 100) %>%
  arrange(desc(total_avg_orders))
```

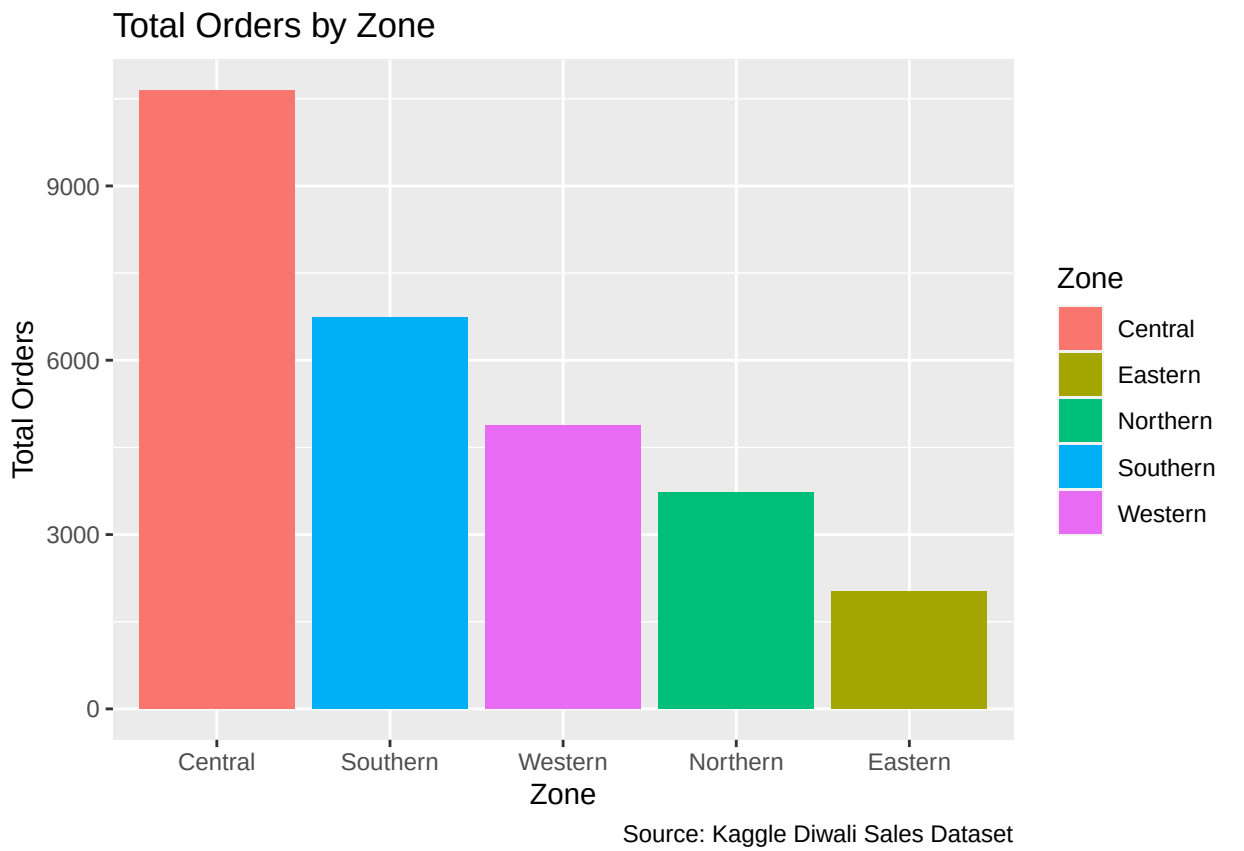
```
## # A tibble: 5 x 3
##   Zone     total_avg_orders pct_avg_orders
##   <chr>               <int>          <dbl>
## 1 Central              2087          37.7
## 2 Southern             1347          24.3
## 3 Western               989          17.8
## 4 Northern              722          13.0
## 5 Eastern               398           7.18
```

```
  # total orders
  diwali_sales %>%
    select(Zone,
           Product_Category,
           Orders) %>%
    group_by(Zone) %>%
    summarise(total_orders = sum(Orders))
```

Question 4.2: Calculate the total number of orders across all zones.

```
## # A tibble: 5 x 2
##   Zone     total_orders
##   <chr>           <int>
## 1 Central         10640
## 2 Eastern          2015
## 3 Northern         3727
## 4 Southern         6744
## 5 Western          4881
```

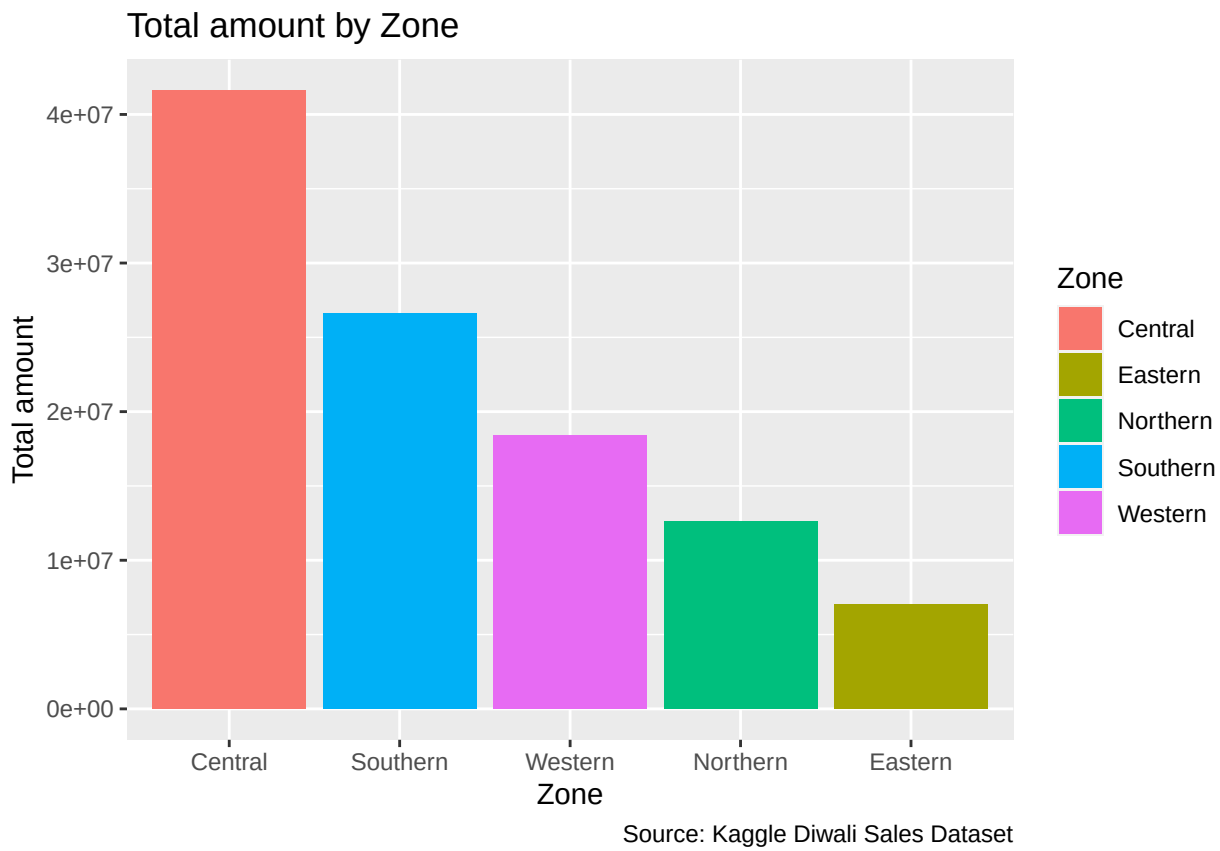
```
## Plot 4.2:  Calculate the total number of orders across all zones.

ds_cl %>%
  group_by(Zone) %>%
  summarise(total_orders = sum(Orders)) %>%
  arrange(desc(total_orders)) %>%
  ggplot(mapping = aes(x = reorder(Zone, -total_orders), y = total_orders, fill = Zone)) +
  geom_col() +
  labs(
    title = "Total Orders by Zone",
    x = "Zone",
    y = "Total Orders",
    caption =  "Source: Kaggle Diwali Sales Dataset"
  )
```



Question 4.3: For each zone, identify the product category with the highest number of orders.

The summary of this question involves visualizing the highest amount of orders by zone.

```
## Plot 4.3:  For each zone, identify the product category with the highest number of orders.

ds_cl %>%
  group_by(Zone) %>%
  summarise(total_amount = sum(Amount)) %>%
  arrange(desc(total_amount)) %>%
  ggplot(
         mapping = aes(x = reorder(Zone, -total_amount), y = total_amount, fill = Zone )) +
  geom_col() +
  labs(
    title = "Total amount by Zone",
    x = "Zone",
    y = "Total amount",
    caption = "Source: Kaggle Diwali Sales Dataset"
  )
```



Source: Kaggle Diwali Sales Dataset

Central

```
## Central

  diwali_sales %>%
    select(Zone,
           Product_Category,
           Orders,
           Amount) %>%
    group_by(Product_Category) %>%
    filter(Zone == "Central") %>%
    summarize(Total_order = sum(Orders)) %>%
    arrange(desc(Total_order)) %>%
    head(2)
```

```
## # A tibble: 2 x 2
##   Product_Category   Total_order
##   <chr>                    <int>
## 1 Food                      2452
## 2 Clothing & Apparel        2243
```

Western

```
## Western

  diwali_sales %>%
```

```
    select(Zone,
           Product_Category,
           Orders,
           Amount) %>%
    group_by(Product_Category) %>%
    filter(Zone == "Western") %>%
    summarize(Total_order = sum(Orders)) %>%
    arrange(desc(Total_order)) %>%
    head(2)
```

```
## # A tibble: 2 x 2
##   Product_Category   Total_order
##   <chr>                    <int>
## 1 Food                      1435
## 2 Clothing & Apparel        1118
```

Southern

```
## Southern

  diwali_sales %>%
    select(Zone,
           Product_Category,
           Orders,
           Amount) %>%
    group_by(Product_Category) %>%
    filter(Zone == "Southern") %>%
    summarize(Total_order = sum(Orders)) %>%
    arrange(desc(Total_order)) %>%
    head(2)
```

```
## # A tibble: 2 x 2
##   Product_Category   Total_order
##   <chr>                    <int>
## 1 Clothing & Apparel        1449
## 2 Food                      1313
```

Eastern

```
## Eastern

  diwali_sales %>%
    select(Zone,
           Product_Category,
           Orders,
           Amount) %>%
    group_by(Product_Category) %>%
    filter(Zone == "Eastern") %>%
    summarize(Total_order = sum(Orders)) %>%
    arrange(desc(Total_order)) %>%
    head(2)
```

```
## # A tibble: 2 x 2
##   Product_Category      Total_order
##   <chr>                       <int>
## 1 Clothing & Apparel            721
## 2 Electronics & Gadgets         552
```

Northern

```
## Northern

  diwali_sales %>%
    select(Zone,
           Product_Category,
           Orders,
           Amount) %>%
    group_by(Product_Category) %>%
    filter(Zone == "Northern") %>%
    summarize(Total_order = sum(Orders)) %>%
    arrange(desc(Total_order)) %>%
    head(2)
```

```
## # A tibble: 2 x 2
##   Product_Category      Total_order
##   <chr>                       <int>
## 1 Electronics & Gadgets        1316
## 2 Clothing & Apparel           1103
```

```
# which product category is the most spent in Central?

diwali_sales %>%
  select(Zone,
         Product_Category,
         Amount) %>%
  group_by(Product_Category) %>%
  filter(Zone == "Central") %>%
  summarize(Total_Amount = sum(Amount)) %>%
  arrange(desc(Total_Amount)) %>%
  head(10)
```

Question 4.4: Identify the product category with the highest total sales amount each zone.

```
## # A tibble: 10 x 2
##    Product_Category      Total_Amount
##    <chr>                        <dbl>
##  1 Food                    13685560.
##  2 Footwear & Shoes         8468991.
##  3 Clothing & Apparel       5744202
##  4 Electronics & Gadgets    3907342
##  5 Furniture                2067178
##  6 Beauty                   1420386
```

```
##  7 Games & Toys              1382910
##  8 Stationery                1266360.
##  9 Sports Products           1023434
## 10 Decor                      532002
```

```
  # which product category is the most spent in Western?

diwali_sales %>%
  select(Zone, Product_Category, Amount) %>%
  group_by(Product_Category) %>%
  filter(Zone == "Western") %>%
  summarize(Total_Amount = sum(Amount)) %>%
  arrange(desc(Total_Amount)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Product_Category      Total_Amount
##    <chr>                        <dbl>
##  1 Food                       7764072
##  2 Clothing & Apparel         2698534
##  3 Electronics & Gadgets      2230270
##  4 Sports Products            1041627
##  5 Games & Toys                979280
##  6 Furniture                   966517
##  7 Footwear & Shoes            929332
##  8 Household items             377449
##  9 Tupperware                  362751
## 10 Stationery                  246502
```

```
  # which product category is the most spent in Southern?

diwali_sales %>%
  select(Zone, Product_Category, Amount) %>%
  group_by(Product_Category) %>%
  filter(Zone == "Southern") %>%
  summarize(Total_Amount = sum(Amount)) %>%
  arrange(desc(Total_Amount)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Product_Category      Total_Amount
##    <chr>                        <dbl>
##  1 Food                       7326461
##  2 Footwear & Shoes           5436504
##  3 Electronics & Gadgets      3882459
##  4 Clothing & Apparel         3649326
##  5 Furniture                  1295949.
##  6 Auto                       1237453.
##  7 Games & Toys               1001066
##  8 Sports Products             845462
##  9 Household items             526144
## 10 Books                       478963
```

```r
  # which product category is the most spent in Eastern?

diwali_sales %>%
  select(Zone, Product_Category, Amount) %>%
  group_by(Product_Category) %>%
  filter(Zone == "Eastern") %>%
  summarize(Total_Amount = sum(Amount)) %>%
  arrange(desc(Total_Amount)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Product_Category      Total_Amount
##    <chr>                        <dbl>
##  1 Food                       2209777
##  2 Clothing & Apparel         1772370
##  3 Electronics & Gadgets      1632873
##  4 Footwear & Shoes            315519
##  5 Games & Toys                303982
##  6 Furniture                   302380
##  7 Sports Products             236266
##  8 Auto                         79632
##  9 Tupperware                   54008
## 10 Household items              48448
```

```r
  # which product category is the most spent in Northern?

diwali_sales %>%
  select(Zone, Product_Category, Amount) %>%
  group_by(Product_Category) %>%
  filter(Zone == "Northern") %>%
  summarize(Total_Amount = sum(Amount)) %>%
  arrange(desc(Total_Amount)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    Product_Category      Total_Amount
##    <chr>                        <dbl>
##  1 Electronics & Gadgets      3990902
##  2 Food                       2948013
##  3 Clothing & Apparel         2630587
##  4 Furniture                   808028
##  5 Games & Toys                664456
##  6 Sports Products             489144
##  7 Footwear & Shoes            424863
##  8 Books                       140896
##  9 Tupperware                   98820
## 10 Household items              94974
```

Question 5: Generation

Since the dataset does not have a Generation column, I need to create one.

```r
## create generation table

diwali_sales_gen <- diwali_sales %>%
                select(User_ID,
                       Cust_name,
                       Product_ID,
                       Gender,
                       Age.Group,
                       Age,
                       Marital_Status,
                       Marital_Status,
                       State,
                       Zone,
                       Occupation,
                       Product_Category,
                       Orders,
                       Amount) %>%
                mutate(Generation = ifelse(Age < 26, "Gen Z",
                                    ifelse(Age < 42, "Millennials",
                                    ifelse(Age < 58, "Gen X",
                                    ifelse(Age < 68, "Baby Boomer", "Silent Generation")))))
```

```r
## Question 5.1: Which customer has spent the most money? To which generation do they belong?

diwali_sales_gen %>%
  select(User_ID,
         Cust_name,
         Gender,
         Age,
         Orders,
         Amount,
         Occupation,
         Generation) %>%
  mutate(total_amount = Orders * Amount)%>%
  arrange(desc(total_amount)) %>%
  head(10)
```

Question 5.1: Which customer has spent the most money? To which generation do they belong?

```
## # A tibble: 10 x 9
##    User_ID Cust_name Gender   Age Orders Amount Occupation      Generation
##      <int> <chr>     <chr>  <int>  <int>  <dbl> <chr>           <chr>
##  1 1001132 Balk      F         25      4 23841  Lawyer          Gen Z
##  2 1003650 Ginny     F         26      4 23800. Media           Millennials
##  3 1001680 Vasudev   M         26      4 23718  Automobile      Millennials
##  4 1000113 Ellis     F         19      4 23546  Govt            Gen Z
##  5 1004736 Mahima    F         25      4 23451  Banking         Gen Z
##  6 1004505 Daniels   F         55      4 23302  Healthcare      Gen X
##  7 1002520 Mike      M         72      4 23267  Media           Silent Generati~
```

```
##  8 1003111 Dean      F         25      4 23252  Banking         Gen Z
##  9 1001182 Zypern    M         16      4 23239  Food Processing Gen Z
## 10 1001726 Abhijit   F         32      4 23066  Retail          Millennials
## # i 1 more variable: total_amount <dbl>
```

```
## Avg amount gen z

diwali_sales_gen %>%
  group_by(Generation) %>%
  filter(Generation == "Gen Z") %>%
  summarize(
    total_amount = sum(Amount),
    avg_amount = mean(Amount),
    percentage_amount = (sum(Amount) / sum(diwali_sales_gen$Amount)) * 100
  )
```

Question 5.2: Calculate the average order amount for Generation Z customers.

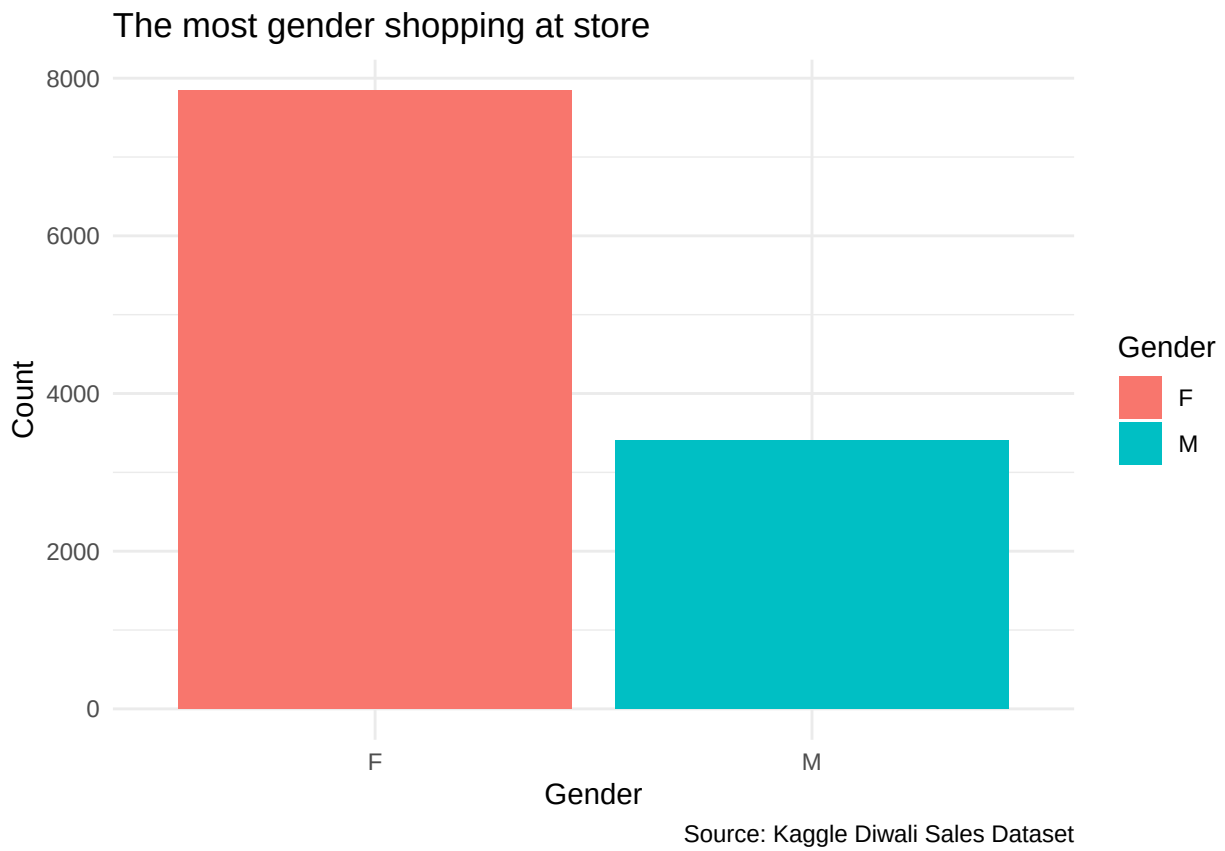
```
## # A tibble: 1 x 4
##   Generation total_amount avg_amount percentage_amount
##   <chr>             <dbl>      <dbl>             <dbl>
## 1 Gen Z          19940385      9168.              18.8
```

```
  ## How many Male and Female in Gender?

diwali_sales %>%
  count(Gender)
```

Question 5.3: Determine the number of male and female customers

```
## # A tibble: 2 x 2
##   Gender     n
##   <chr>  <int>
## 1 F       7842
## 2 M       3409
```

```
## Plot 3: Which gender are the most come to shopping at store

ggplot(data = ds_cl,
       mapping = aes(x = Gender, fill = Gender)) +
  geom_bar() +
  theme_minimal() +
  labs(
    title = "The most gender shopping at store",
    x = "Gender",
    y = "Count",
    caption = "Source: Kaggle Diwali Sales Dataset"
  )
```



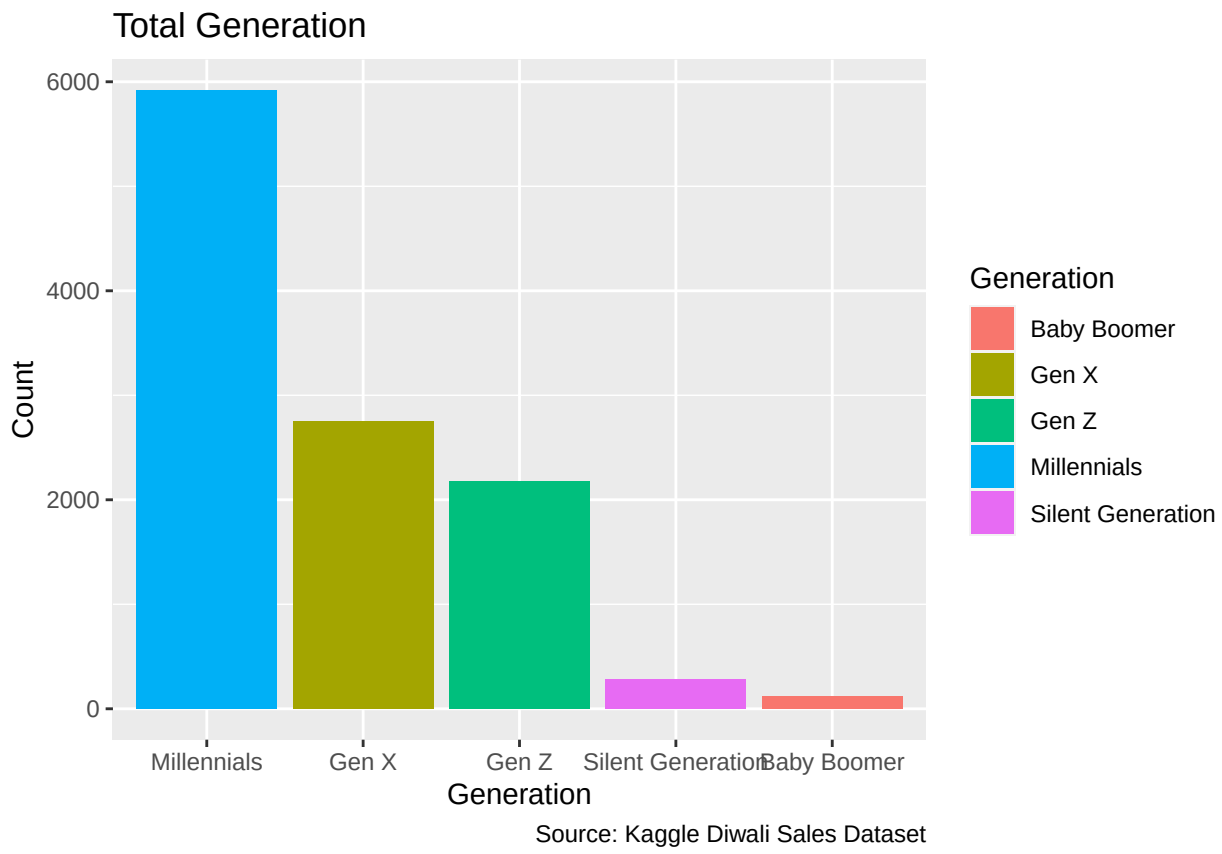

```
  ## How many Generation in Customer?

diwali_sales_gen %>%
  count(Generation)
```

Question 5.4: Identify the number of customers belonging to each generation.

```
## # A tibble: 5 x 2
##   Generation            n
##   <chr>             <int>
## 1 Baby Boomer         123
## 2 Gen X              2754
## 3 Gen Z              2175
## 4 Millennials        5916
## 5 Silent Generation   283
```

```
## Plot 5: Total Generation

diwali_sales_gen%>%
  group_by(Generation) %>%
  count() %>%
  ggplot(mapping = aes(x = reorder(Generation, -n), y = n, fill = Generation))+
  geom_col() +
```

```
  labs(
    title = "Total Generation",
    x = "Generation",
    y = "Count",
    caption = "Source: Kaggle Diwali Sales Dataset"
  )
```

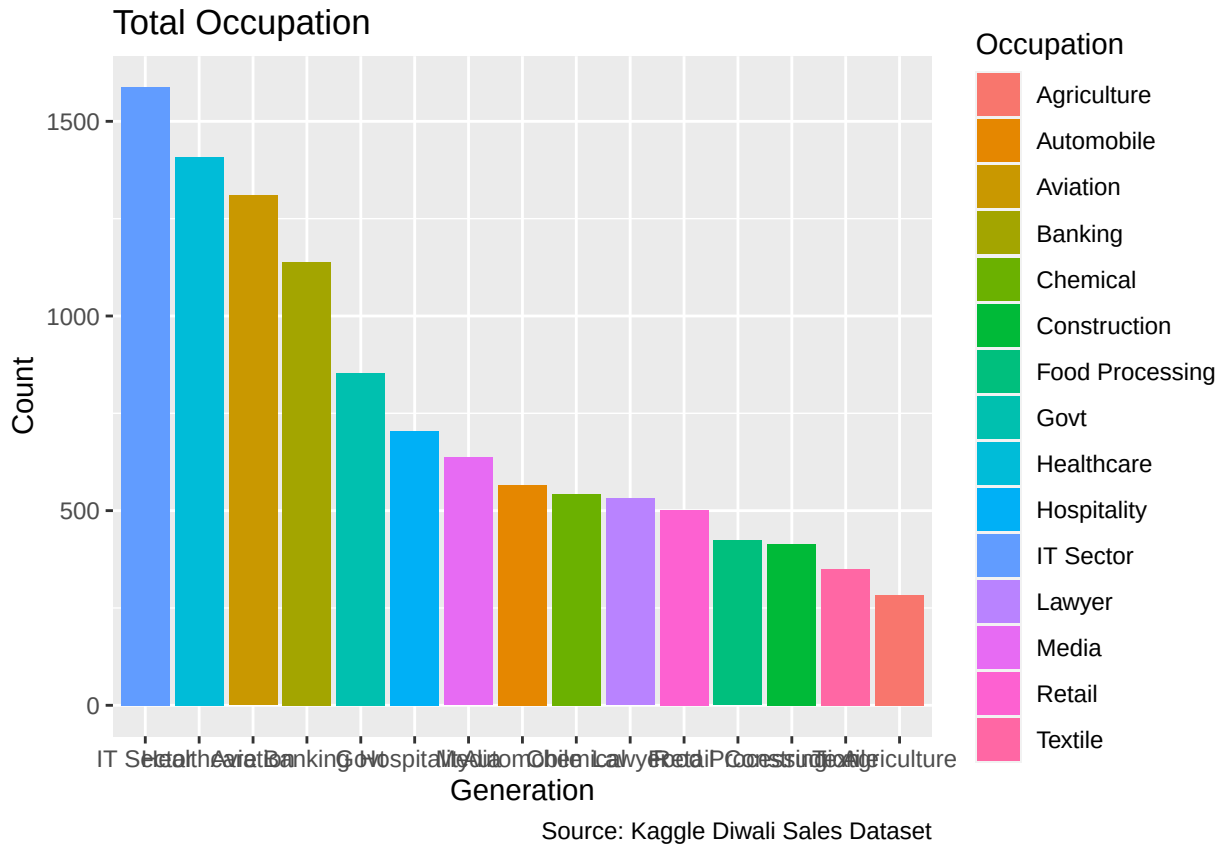


Occupationa

Before moving on to the next question, I provided the total occupation count from the dataset.

```
## Plot 6: Total Occupation

diwali_sales_gen %>%
  group_by(Occupation) %>%
  count() %>%
  ggplot(mapping = aes(x = reorder(Occupation, -n), y = n, fill = Occupation))+
  geom_col() +
  labs(
    title = "Total Occupation",
    x = "Generation",
    y = "Count",
```

```
    caption = "Source: Kaggle Diwali Sales Dataset"
  )
```



Question 6: Identify the percentage amount for each occupation.

```
## Question 6: Identify the percentage amount for each occupation.

avg_amount <- mean(diwali_sales$Amount)

diwali_sales %>%
  select(Occupation,
         Orders,
         Amount) %>%
  group_by(Occupation) %>%
  summarise(total_avg_amount = n(),
            pct_amount = (sum(Amount) / sum(diwali_sales_gen$Amount)) *100) %>%
  arrange(desc(total_avg_amount))
```

```
## # A tibble: 15 x 3
##    Occupation      total_avg_amount pct_amount
##    <chr>                      <int>      <dbl>
##  1 IT Sector                   1588       13.9
##  2 Healthcare                  1408       12.3
```

```
##  3 Aviation                    1310      11.9
##  4 Banking                     1139      10.1
##  5 Govt                         854       8.02
##  6 Hospitality                  705       6.00
##  7 Media                        637       5.93
##  8 Automobile                   566       5.05
##  9 Chemical                     542       4.99
## 10 Lawyer                       531       4.69
## 11 Retail                       501       4.50
## 12 Food Processing              423       3.83
## 13 Construction                 414       3.39
## 14 Textile                      350       3.02
## 15 Agriculture                  283       2.44
```

Age Group

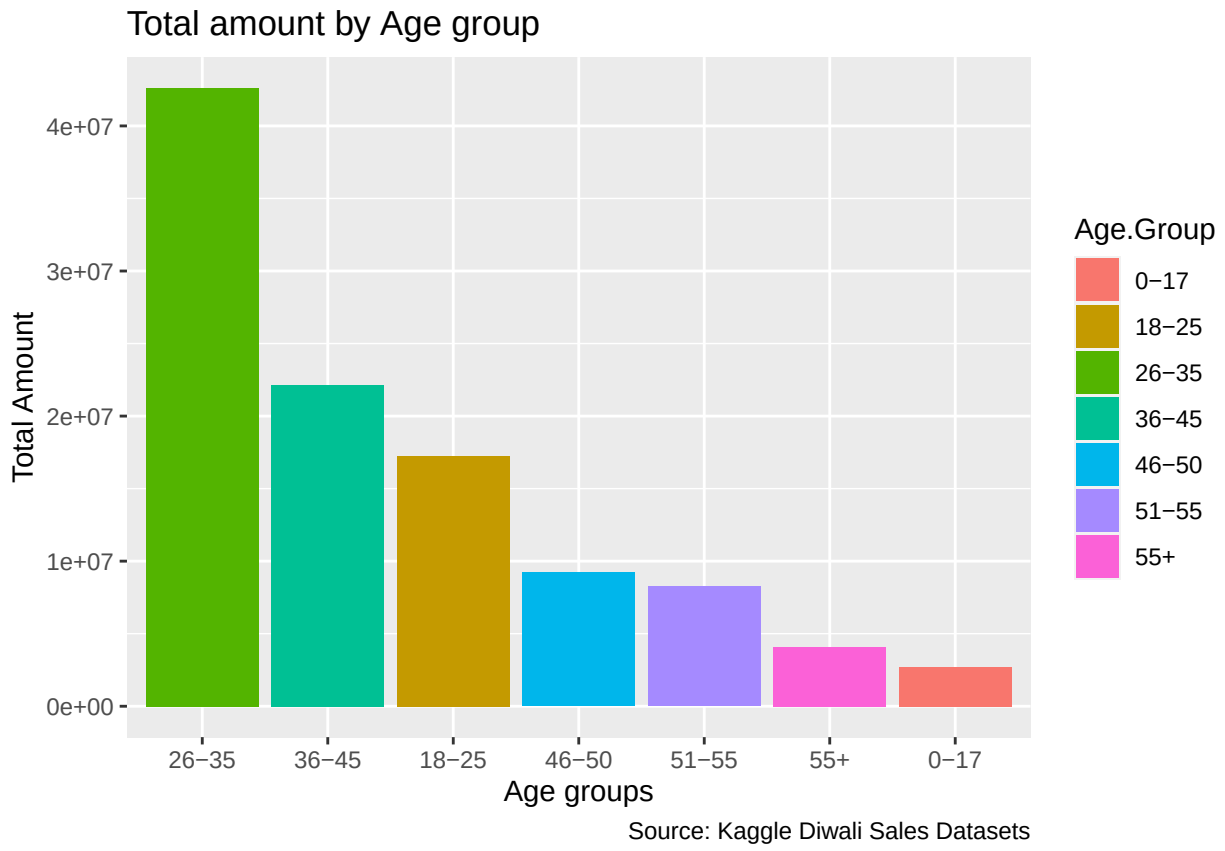
Question 7: What is the total amount for each age group?

```
## Question 7: What is the total amount for each age group?

diwali_sales %>%
  select(Age.Group,
         Product_Category,
         Amount,
         Orders) %>%
  group_by(Age.Group) %>%
  summarise(total_amount = sum(Amount),
            pct_amount = (sum(Amount) / sum(diwali_sales_gen$Amount)) * 100) %>%
  arrange(desc(total_amount))
```

```
## # A tibble: 7 x 3
##   Age.Group total_amount pct_amount
##   <chr>            <dbl>      <dbl>
## 1 26-35        42613444.      40.1
## 2 36-45        22144995.      20.8
## 3 18-25        17240732       16.2
## 4 46-50         9207844        8.67
## 5 51-55         8261477        7.78
## 6 55+           4080987        3.84
## 7 0-17          2699653        2.54
```

```
## Plot 7: What is the total amount for each age group?

diwali_sales %>%
  group_by(Age.Group) %>%
  summarise(total_amount = sum(Amount)) %>%
  arrange(desc(total_amount)) %>%
  ggplot(mapping = aes( x= reorder(Age.Group, -total_amount), y = total_amount, fill = Age.Group)) +
  geom_col() +
```

```
labs(
  title = "Total amount by Age group",
  x = "Age groups",
  y = "Total Amount",
  caption = "Source: Kaggle Diwali Sales Datasets"
)
```



Summary

In summary, the most prominent product category in the Diwali Sales dataset is food. The dataset shows that in the state of Uttar Pradesh, there are a lot of customers buying during Diwali. In terms of zones, the Central zone has the highest total amount. Lastly, the most common customer age range is 26-35, representing Generation Z. By the way, I've noticed that females spend more than males in this dataset.

Part II: Machine Learning

What is Machine Learing

Machine learning is a type of artificial intelligence (AI) focused on building computer systems that learn from data.

Before delving into machine learning. I am using machine learning to predict the amount of sales during Diwali. In this section, I have broken down the topic to make it understandable for everyone, demonstrating how I code by following these steps:

- Install package
- Import dataset
- Prepare data
- Split data
- Train & Test data
- Scoring
- Evaluate model

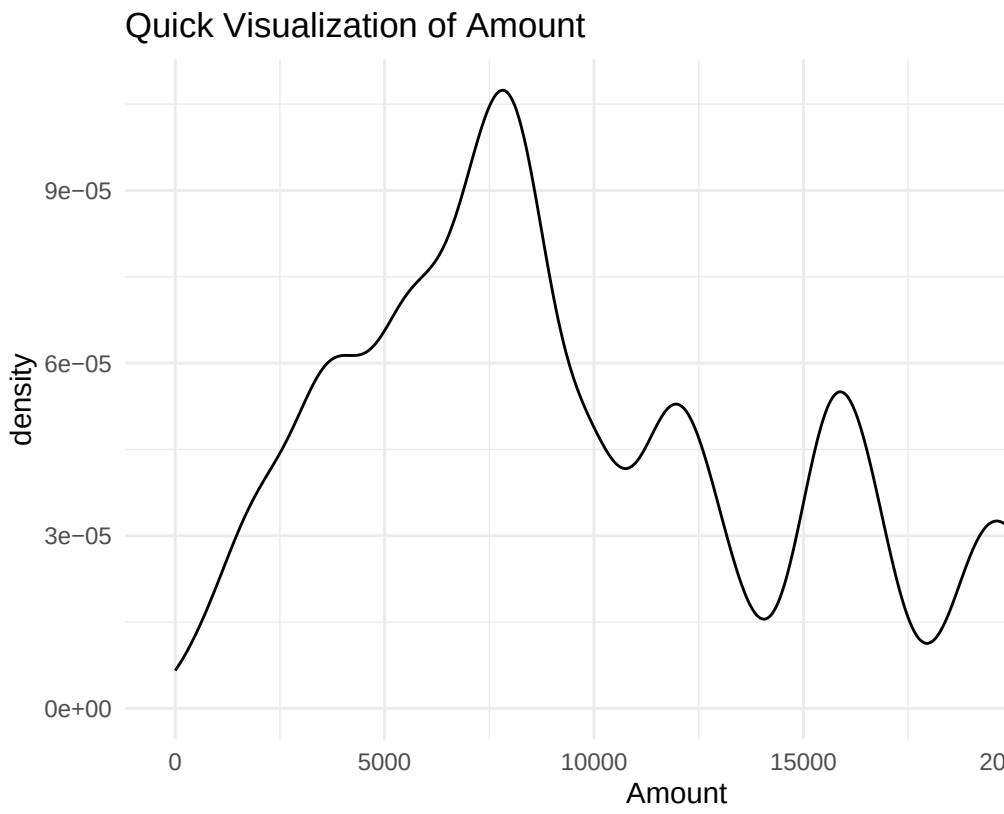
```
## install.packages
install.packages("readr")
install.packages("dplyr")
install.packages("tidyverse")
install.packages("ggplot2")

## Library
library(readr)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(caret)

## import read.csv()

diwali_sales <- tibble(read.csv("Diwali_Sales_Data_main.csv"))

## select column

ds_cl <- diwali_sales %>%
  select(User_ID,
         Cust_name,
         Product_ID,
         Product_Category,
         Gender,
         Age,
         Marital_Status,
         State,
         Zone,
         Occupation,
         Orders,
         Amount) %>%
  drop_na()



## check NA
```

```
ds_cl %>%
  complete.cases() %>%
  mean()
```

Install package and Import dataset

```
## [1] 1
```

```
ggplot(ds_cl, aes(Amount)) +
  geom_density() +
  theme_minimal() +
  labs(
    title = "Quick Visualization of Amount",
    x = "Amount",
    caption = "Source: Kaggle Diwali Sales Datasets"
  )
```



Quick Visualization of Amount

```r
## split data

sp_data <- function(ds_cl, train_size = 0.7) {
  set.seed(42)
  n <- nrow(diwali_sales)
  id <- sample(n, size = n * train_size)
  tr_df <- ds_cl[id,]
  ts_df <- ds_cl[-id,]
  list(train = tr_df,
       test = ts_df)
}


## prep data

prep_data <- sp_data(ds_cl)
tr_df <- prep_data[[1]]
ts_df <- prep_data[[2]]
```

Split and Prep Data

```r
## train model

set.seed(42)
lm_model <- train(Amount ~ Orders ,
                  data = tr_df,
                  method = "lm")


## view model

lm_model
```

Train data

```
## Linear Regression
##
## 7875 samples
##    1 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 7875, 7875, 7875, 7875, 7875, 7875, ...
## Resampling results:
##
##   RMSE     Rsquared      MAE
##   5225.28  0.0005284117  4312.628
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
# score <- prediction

p <- predict(lm_model, newdata = ts_df)
```

Scoring

```r
## evaluate
  # mae mse rmse
  ## create function

cal_mae <- function(actual, pred) {
  error <- actual - pred
  return(mean(abs(error)))
}

cal_mse <- function(actual, pred) {
  error <- actual - pred
  mean(error ** 2)
}

cal_rmse <- function(actual, pred) {
  error <- actual - pred
  sqrt(mean(error ** 2))
}


# check result

cal_mae(ts_df$Amount, p)
```

Evaluate

```
## [1] 4272.277
```

```r
cal_mse(ts_df$Amount, p)
```

```
## [1] 26999848
```

```r
cal_rmse(ts_df$Amount, p)
```

```
## [1] 5196.138
```

```r
## Summary

lm_model$finalModel %>%
  summary()
```

Summary

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9597  -3992  -1337   3247  14498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9676.97     144.54  66.949   <2e-16 ***
## Orders        -80.27      53.05  -1.513     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5243 on 7873 degrees of freedom
## Multiple R-squared:  0.0002907,  Adjusted R-squared:  0.0001637
## F-statistic: 2.289 on 1 and 7873 DF,  p-value: 0.1303
```