Marketing Data Analysis

Pearl Nibebadaar Sandoh Kuuridong

2025-09-13

Instructions

```
# Do not modify this chunk
library(tidyverse)
library(readxl) # package that will help you to load MS Excel data in R.
library(flextable)
library(ggplot2) # will help with advanced plotting if necessary.
```

In this assignment, we perform basic data exploration and visualization on marketing data. This is to get insights on how customers behave to what `FreshDirect` company offers. FreshDirect is the leader in online grocery delivery. Their marketing data captures customer information such as demographics, transaction behavior, and ordering patterns to enable loyalty analysis, segmentation, and predictive modeling.

Your task is to load the data using the package `readxl` allowing you to load excel files. Identify the function from the package with specific chosen parameters from it to get rid of some big issues that may come with data importing. Then you also need to perform some data cleaning for some computations.

Below is the data description.

Column Name	**Description**
LOYALTY_SEGMENT	Shopper classification based on purchase frequency (e.g., Weekly, Bi-Weekly, Monthly)
AGE	Customer's age
INCOME	Household income (may be grouped into ranges)
GENDER	Gender of the primary shopper
ZIP_CODE	Residential ZIP code of customer
DMA	Designated Market Area (media/advertising region)
GEOGRAPHY	Broader geographic grouping
ACQUIRED_DATE	Date the customer first registered or became active
12 Mo. DELIVERY_FEE_PAID	Total delivery fees paid in the last 12 months
12 Mo. DELIVERYPASS_USED	Number of times DeliveryPass subscription was used in 12 months
12 Mo. DISCOUNT_AMOUNT	Total discounts applied in 12 months
12 Mo. Orders	Total number of orders in the past 12 months
12 Mo. ORDERS_W_PROMO	Number of orders that included a promo in 12 months
12 Mo. Sales	Total sales generated by customer in 12 months

Column Name	Description
24 Mo. DELIVERY_FEE_PAID	Total delivery fees paid in the last 24 months
24 Mo. DELIVERYPASS_USED	Number of times DeliveryPass was used in 24 months
24 Mo. DISCOUNT_AMOUNT	Total discounts applied in 24 months
24 Mo. Orders	Total number of orders in the past 24 months
24 Mo. Orders w. Promo	Number of orders with a promo in 24 months
24 Mo. Sales	Total sales generated by customer in 24 months
SUNDAY ORDERS 12 MO.	Number of orders placed on Sundays (12 months)
MONDAY ORDERS 12 MO.	Number of orders placed on Mondays (12 months)
TUESDAY ORDERS 12 MO.	Number of orders placed on Tuesdays (12 months)
WEDNESDAY ORDERS 12 MO.	Number of orders placed on Wednesdays (12 months)
THURSDAY ORDERS 12 MO.	Number of orders placed on Thursdays (12 months)
FRIDAY ORDERS 12 MO.	Number of orders placed on Fridays (12 months)
SATURDAY ORDERS 12 MO.	Number of orders placed on Saturdays (12 months)

Use R to attempt each of the following questions. We recommend you to write your interpretation in your English.

Part 1.

1. After importing data in R. Check which column has the highest count of missing information.

```r
# Import data in a variable named df
df <- read_xlsx("FD_data.xlsx", skip=1, na = c("NA", "N/A", "", "#N/A", "-", " "))

null_counts <- colSums(is.na(df))
null_counts[null_counts == max(null_counts)]
```

```
## GENDER
##    659
```

```r
# Check the dimension of df
dim(df)
```

```
## [1] 3000   27
```

```r
# Display the data structure of df
str(df)
```

```
## tibble [3,000 x 27] (S3: tbl_df/tbl/data.frame)
##  $ LOYALTY_SEGMENT         : chr [1:3000] "2. Bi-Weekly Shoppers" "5. Infrequent Shoppers" "3. Every
##  $ AGE                     : num [1:3000] 60 53 NA 29 NA 54 40 68 33 NA ...
##  $ INCOME                  : chr [1:3000] "$150,000" "$20,000" NA "$150,000" ...
##  $ GENDER                  : chr [1:3000] "F" "F" NA "F" ...
##  $ ZIP_CODE                : num [1:3000] 10465 10462 10462 11206 11231 ...
##  $ DMA                     : chr [1:3000] "NY-NJ" "NY-NJ" "NY-NJ" "NY-NJ" ...
##  $ GEOGRAPHY               : chr [1:3000] "Bronx" "Bronx" "Bronx" "Brooklyn" ...
##  $ ACQUIRED_DATE           : chr [1:3000] "6/4/12 0:00" "7/1/12 0:00" "7/10/12 0:00" "1/7/12 0:00" .
##  $ 12 Mo. DELIVERY_FEE_PAID: chr [1:3000] "$0.00" "$5.99" "$5.99" "$0.00" ...
##  $ 12 Mo. DELIVERYPASS_USED: num [1:3000] 51 0 12 13 23 0 0 0 0 0 ...
##  $ 12 Mo. DISCOUNT_AMOUNT  : chr [1:3000] "$98.65" "$0.00" "$139.81" "$20.96" ...
##  $ 12 Mo. Orders           : num [1:3000] 57 4 18 15 25 31 8 12 8 3 ...
##  $ 12 Mo. ORDERS_W_PROMO   : num [1:3000] 19 NA 16 7 4 2 5 1 8 NA ...
##  $ 12 Mo. Sales            : chr [1:3000] "$25,195" "$84" "$1,496" "$1,092" ...
##  $ 24 Mo. DELIVERY_FEE_PAID: chr [1:3000] "$0.00" "$5.99" "$17.97" "$17.97" ...
##  $ 24 Mo. DELIVERYPASS_USED: num [1:3000] 103 0 26 13 46 0 0 0 0 0 ...
##  $ 24 Mo. DISCOUNT_AMOUNT  : chr [1:3000] "$128.16" "$0.00" "$257.65" "$50.92" ...
##  $ 24 Mo. Orders           : num [1:3000] 105 1 39 21 46 35 9 24 18 3 ...
##  $ 24 Mo. Orders w. Promo  : num [1:3000] 25 NA 33 14 8 9 6 3 15 NA ...
##  $ 24 Mo. Sales            : chr [1:3000] "$37,999" "$130" "$3,273" "$1,823" ...
##  $ SUNDAY ORDERS 12 MO.    : num [1:3000] NA 2 2 1 NA 4 1 6 5 NA ...
##  $ MONDAY ORDERS 12 MO.    : num [1:3000] 50 1 NA 1 3 3 1 1 NA NA ...
##  $ TUESDAY ORDERS 12 MO.   : num [1:3000] 2 NA 1 2 5 NA 3 NA 1 1 ...
##  $ WEDNESDAY ORDERS 12 MO. : num [1:3000] NA NA 4 4 4 1 NA 1 NA NA ...
##  $ THURSDAY ORDERS 12 MO.  : num [1:3000] NA NA 4 1 3 3 1 NA NA 1 ...
##  $ FRIDAY ORDERS 12 MO.    : num [1:3000] 2 NA 5 3 4 1 1 NA 1 1 ...
##  $ SATURDAY ORDERS 12 MO.  : num [1:3000] NA 1 2 1 4 19 1 4 1 NA ...
```

Point out any issue with the data (asterix serve as bullet points in markdown. Add * as much as possible.)

- Some columns have a high number of missing values
- The amounts are being read in as character data types instead of numeric
- Date values are being read in as characters
- The loyalty_segment column is preceeded by a number.

2. How many unique customers are in the dataset?

```r
nrow(unique(df))
```

```
## [1] 3000
```

Interpretation:

- There are 3000 unique customers in the dataset

3. What is the age distribution of customers?

```r
df$AGE <- as.numeric(df$AGE)
age_summary <- summary(df$AGE, na.rm=T)
standard_dv <- sd(df$AGE, na.rm=T)

age_summary; standard_dv
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.00   36.00   44.00   45.26   55.00  101.00     543
```

```
## [1] 17.03842
```

Interpretation:

There are 2457 non-null age values in the dataset. The minimum age is 0, and the maximum age is 101. The median age is 44 years, while the Mean age is 45.26 years. This slight difference between the median and mean ages suggests an outlier towards the higher side of the age range. The interquartile range is 65 and the standard deviation is 17.03842

4. What is the income distribution (mean, median, spread or standard deviation)?

```r
df$cleaned_income <- as.numeric(gsub("\\$|,|\\(|\\)", "", df$INCOME), na.rm=T)
```

```
mean_income <- mean(df$cleaned_income, na.rm=T)
median_income <- median(df$cleaned_income, na.rm=T)
income_std <- sd(df$cleaned_income, na.rm=T)

mean_income; median_income; income_std
```

```
## [1] 108968.4

## [1] 1e+05

## [1] 90018.03
```

Interpretation:

The mean income is $\$1.0896841 \times 10^5$ and the median income is 10^5. These values are within one standard deviation (9.0018035×10^4) of each other and so, I claim the data is symmetric about the mean.

5. After decoding the variable GENDER (replace F and M by Female and Male respectively), what are the count of male and female customers in the dataset?

```
df$cleaned_gender <- df$GENDER
df$cleaned_gender[df$cleaned_gender=="F"] <- "Female"
df$cleaned_gender[df$cleaned_gender=="M"] <- "Male"

table(df$cleaned_gender)
```

```
##
## Female   Male
##   1596    745
```

Interpretation:

There are 1596 females in the dataset and 745 males in the dataset.

6. Which part of America has the highest customer counts?

```
cust_counts <- table(df$GEOGRAPHY)
cust_counts[cust_counts==max(cust_counts)]
```

```
## Manhattan
##      2054
```

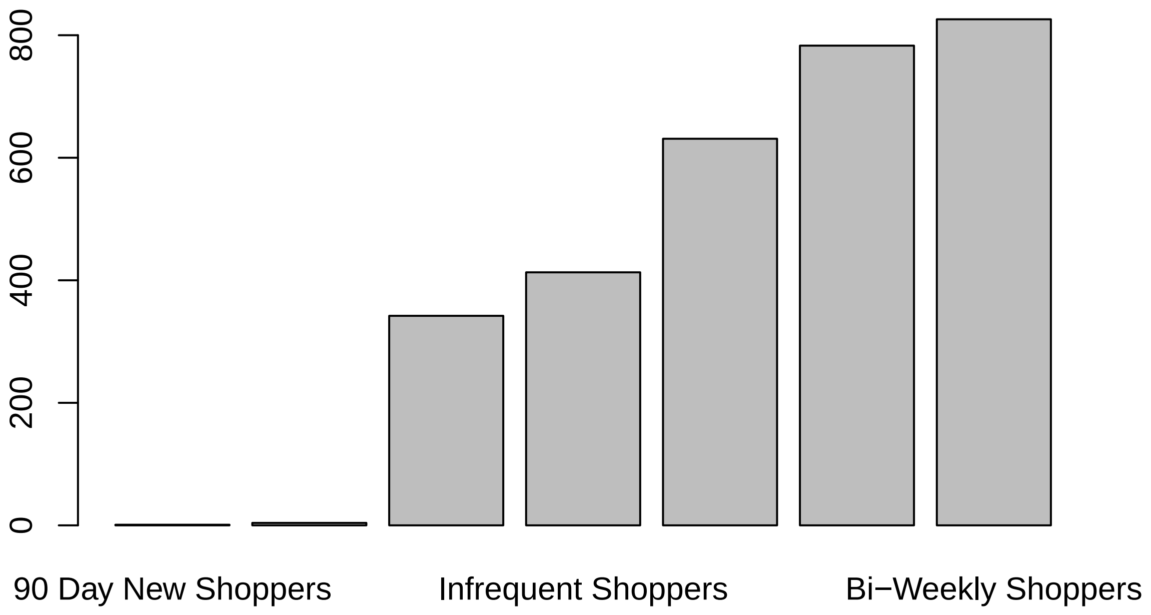
Interpretation:

Manhattan has the highest customer counts in America.

7. How many customers fall into each LOYALTY_SEGMENT category? You can draw a pie/bar chart and interpret it.

```
df$cleaned_loyalty <- gsub("\\d+\\.\\s+", "", df$LOYALTY_SEGMENT)
loyalty_counts = sort(table(df$cleaned_loyalty))
loyalty_counts; barplot(loyalty_counts)
```

```
##
##       90 Day New Shoppers                No Segment     Once a Month Shoppers
##                         1                         4                       342
##       Infrequent Shoppers           Weekly Shoppers Every Three Week Shoppers
##                       413                       631                       783
##        Bi-Weekly Shoppers
##                       826
```



Interpretation:

- 90 day shoppers and No segment shoppers form the smallest group of customers with 1 and 4 members in each respective category. Bi-weekly shoppers are the most frequent with 826 customers, followed by every 3 week shoppers with 783 counts. Weekly shoppers, infrequent shoppers, and once a month shoppers each have 631, 413, and 342 customers respectively.

8. Compute the average number of orders per customer in 12 months.

```
# df$`12 Mo. Orders`
mean_orders = mean(df$`12 Mo. Orders`, na.rm=T)

mean_orders
```

```
## [1] 30.68
```

Interpretation:

The average number of orders per customer, taking each row as one customer is 30.68

9. Compute the average sales per customer in 12 months?

```
df$cleaned_twelve_sales <- as.numeric(gsub("\\$|,", "", df$`12 Mo. Sales`))
mean_sales = mean(df$cleaned_twelve_sales, na.rm=T)

mean_sales
```

```
## [1] 3643.538
```

Interpretation:

The average sales per customer in 12 months is about $3643.54

10. How many customers used DeliveryPass at least once?

```
used_delivery_pass <- sum(df$`24 Mo. DELIVERYPASS_USED`>0)

used_delivery_pass
```

```
## [1] 1644
```

Interpretation:

1644 customers used DeliveryPass at least once

Part 2.

11. Do higher-income customers place more orders?

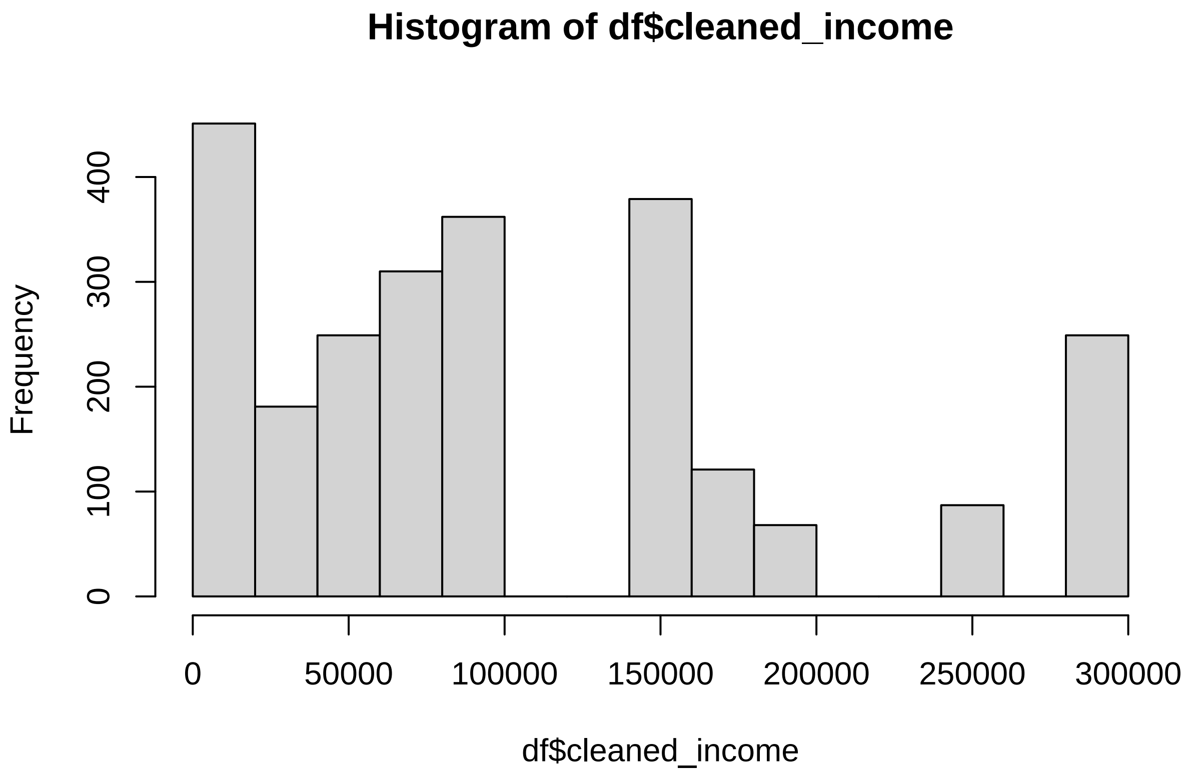
```r
df$income_brackets <- df$cleaned_income



df$income_brackets[df$cleaned_income<=median_income] <- "LOW INCOME"
df$income_brackets[df$cleaned_income>median_income & df$cleaned_income<=mean_income] <- "MIDDLE INCOME"
df$income_brackets[df$cleaned_income>mean_income] <- "HIGH INCOME"

low_income_orders <- sum(df$`24 Mo. Orders`[df$income_brackets=="LOW INCOME"], na.rm=T)
middle_income_orders <- sum(df$`24 Mo. Orders`[df$income_brackets=="MIDDLE INCOME"], na.rm=T)
high_income_orders <- sum(df$`24 Mo. Orders`[df$income_brackets=="HIGH INCOME"], na.rm=T)

low_income_orders; middle_income_orders; high_income_orders
```

```
## [1] 74236

## [1] 0

## [1] 44599
```

Interpretation:

No. Higher income customers place 4.4599×10^4 orders and lower income customers place 7.4236×10^4 orders. There are no middle income customers. This suggests that there is a gap in the income data as shown below:

```r
hist(df$cleaned_income)
```



12. Based on gender, is there any difference between average sales?

```
df$cleaned_twenty_sales <- as.numeric(gsub("\\$|,", "", df$`24 Mo. Sales`))
total_twenty_sales <- sum(df$cleaned_twenty_sales)
female_sales <- sum(df$cleaned_twenty_sales[df$cleaned_gender=="Female"], na.rm=T)
male_sales <- sum(df$cleaned_twenty_sales[df$cleaned_gender=="Male"], na.rm=T)

avg_female_sales <- female_sales/total_twenty_sales
avg_male_sales <- male_sales/total_twenty_sales
```

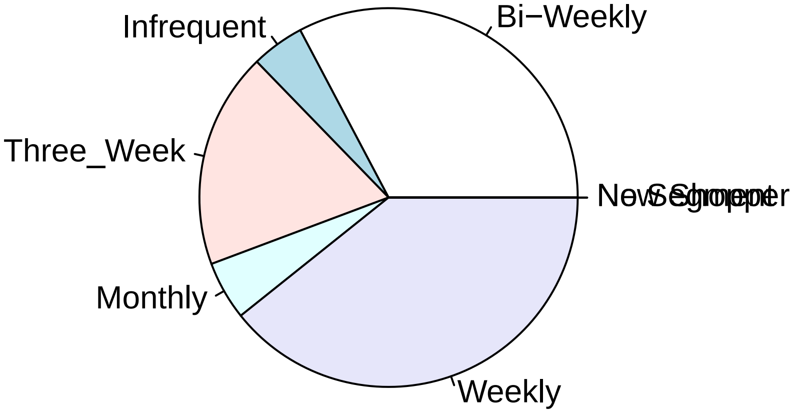
Interpretation:

Females average about 0.5614379 sales and the males average about 0.2524391 sales. So yes, there is a difference in average sales based on gender.

13. How do `LOYALTY_SEGMENTS` differ in terms of twelve month sales?

```
total_sales <- sum(df$cleaned_twelve_sales, na.rm=T)

bi_weekly_sum <- sum(df$cleaned_twenty_sales[df$cleaned_loyalty=="Bi-Weekly Shoppers"], na.rm=T)
infrequent_sum <- sum(df$cleaned_twenty_sales[df$cleaned_loyalty=="Infrequent Shoppers"], na.rm=T)
three_week_sum <- sum(df$cleaned_twenty_sales[df$cleaned_loyalty=="Every Three Week Shoppers"], na.rm=T
monthly_sum <- sum(df$cleaned_twenty_sales[df$cleaned_loyalty=="Once a Month Shoppers"], na.rm=T)
weekly_sum <- sum(df$cleaned_twenty_sales[df$cleaned_loyalty=="Weekly Shoppers"], na.rm=T)
no_segment_sum <- sum(df$cleaned_twenty_sales[df$cleaned_loyalty=="No Segment"], na.rm=T)
new_shopper_sum <- sum(df$cleaned_twenty_sales[df$cleaned_loyalty=="90 Day New Shoppers"], na.rm=T)

bi_weekly_avg <- bi_weekly_sum/total_sales
infrequent_avg <- infrequent_sum/total_sales
three_week_avg <- three_week_sum/total_sales
monthly_avg <- monthly_sum/total_sales
weekly_avg <- weekly_sum/total_sales
```

```r
no_segment_avg <- no_segment_sum/total_sales
new_shopper_avg <- new_shopper_sum/total_sales



pie(c(bi_weekly_avg, infrequent_avg, three_week_avg, monthly_avg, weekly_avg, no_segment_avg, new_shopp
```



Interpretation:

Weekly segment is the one with the highest mean of 12 month sales. The new shopper and no_segment groups are very small, indicating that in 12 months the average amount of sales from new shoppers and customers without any segments is very low. This indicates a moderate to high customer loyalty.

14. Do younger customers, for customer aged less than 30, use promos more than older ones?

```r
young_customers_promo = df[df$AGE<30,]$`24 Mo. Orders w. Promo`
total_young_promo <- sum(young_customers_promo, na.rm=T)
old_customers_promo = df[df$AGE>=30,]$`24 Mo. Orders w. Promo`
total_old_promo <- sum(old_customers_promo, na.rm=T)

total_young_promo; total_old_promo
```

```
## [1] 3222

## [1] 30873
```

Interpretation:

No younger customers used about \$3222 in promos and older customers used about \$$3.0873 \times 10^4$ in promos

15. How does discount amount vary across income brackets?

```r
df$cleaned_twenty_discount_amount <- as.numeric(gsub("\\$","",df$`24 Mo. DISCOUNT_AMOUNT`), na.rm=T)
```

```
## Warning: NAs introduced by coercion
```

```r
high_income_discount <- sum(df$cleaned_twenty_discount_amount[df$income_brackets=="HIGH INCOME"], na.rm
middle_income_discount <- sum(df$cleaned_twenty_discount_amount[df$income_brackets=="MIDDLE INCOME"], na
low_income_discount <- sum(df$cleaned_twenty_discount_amount[df$income_brackets=="LOW INCOME"], na.rm=T

high_income_discount; middle_income_discount; low_income_discount
```

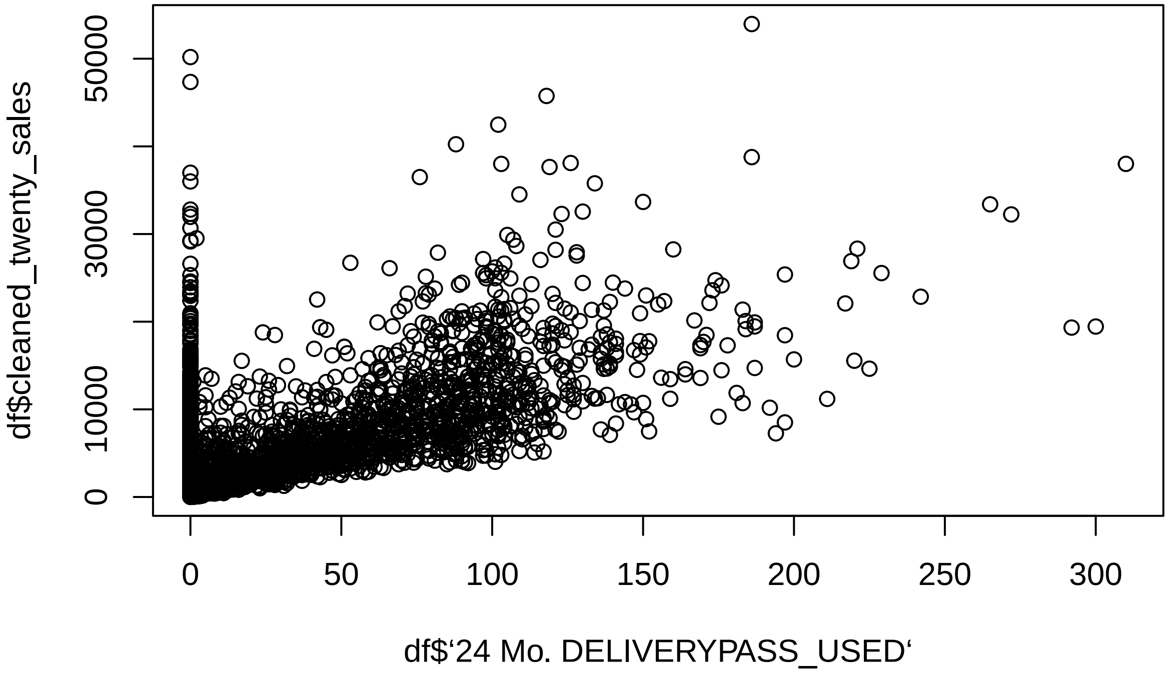
```
## [1] 46439.34

## [1] 0
## [1] 89081.94
```

Interpretation:

Higher earners used less in discounts 4.643934×10^4 while lower earners used more in discounts 8.908194×10^4

16. Is DeliveryPass usage associated with higher total sales?

```r
plot(x = df$`24 Mo. DELIVERYPASS_USED`, y = df$cleaned_twenty_sales)
```



```r
correlation <- cor(df$`24 Mo. DELIVERYPASS_USED`, df$cleaned_twenty_sales, method = "pearson")
```

Interpretation:

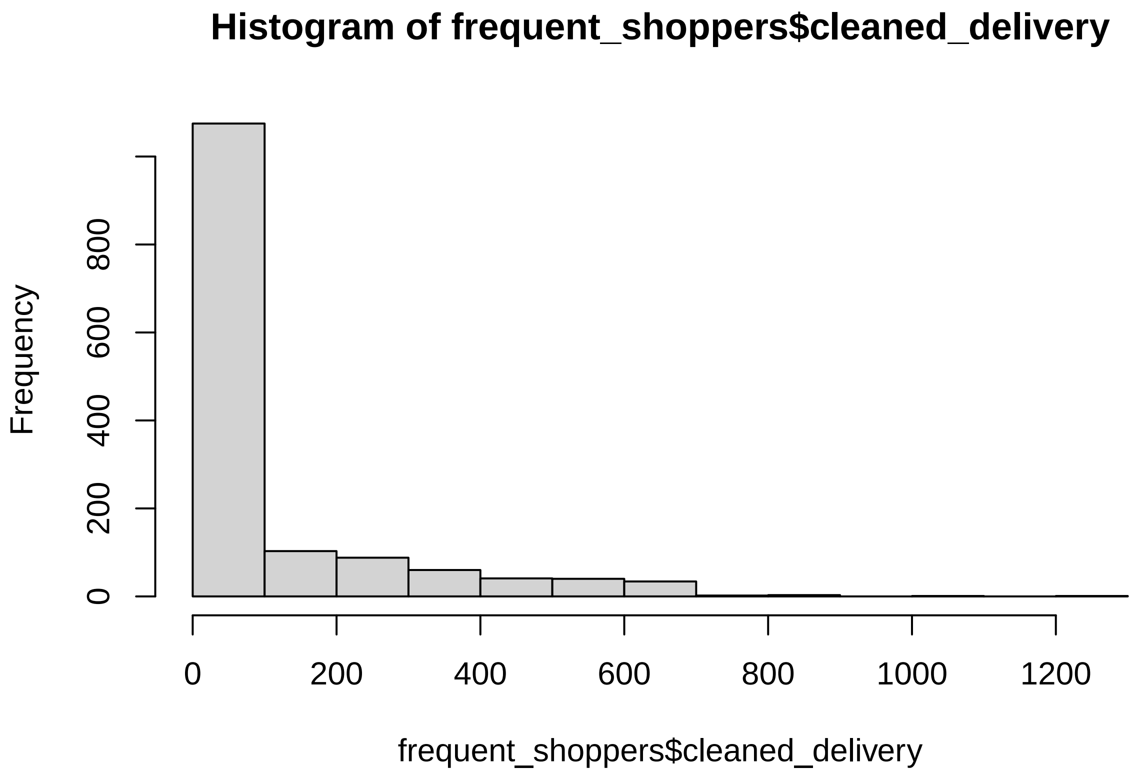
The scatter plot shows a roughly positive linear relationship between the number of times delivery pass was used and the total_sales. This is further confirmed by the correlation coefficient of 0.6174284 which indicates a positive linear correlation.

17. Do frequent shoppers (Weekly, Bi-Weekly) pay less in delivery fees?

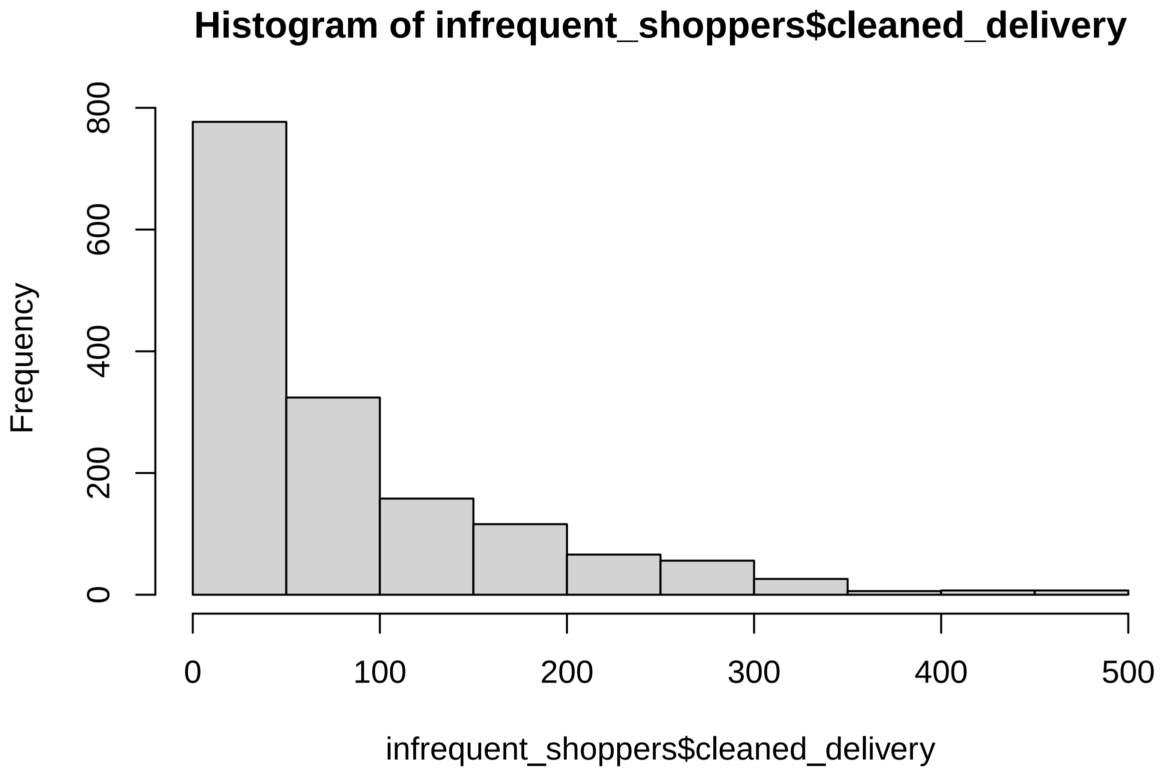
```r
df$cleaned_delivery <- as.numeric(gsub("\\$|,","",as.character(df$`24 Mo. DELIVERY_FEE_PAID`)))
```

```
## Warning: NAs introduced by coercion
```

```r
frequent_shoppers <- df[df$cleaned_loyalty=="Weekly Shoppers"|df$cleaned_loyalty=="Bi-Weekly Shoppers",]
infrequent_shoppers <- df[!(df$cleaned_loyalty=="Weekly Shoppers"|df$cleaned_loyalty=="Bi-Weekly Shoppe

total_delivery = sum(df$cleaned_delivery, na.rm=T)

frequent_deliv_mean <- mean(frequent_shoppers$cleaned_delivery, na.rm=T)
infrequent_deliv_mean <- mean(infrequent_shoppers$cleaned_delivery, na.rm=T)

frequent_deliv_median <- median(frequent_shoppers$cleaned_delivery, na.rm=T)
infrequent_deliv_median <- median(infrequent_shoppers$cleaned_delivery, na.rm=T)

hist(frequent_shoppers$cleaned_delivery)
```



```
hist(infrequent_shoppers$cleaned_delivery)
```



```
frequent_deliv_mean; infrequent_deliv_mean
```

```
## [1] 93.48712
## [1] 78.47417
```

```
frequent_deliv_median; infrequent_deliv_median
```

```
## [1] 3.99
```

```
## [1] 48.93
```

Interpretation:

Since the data is skewed to the right, then the median is a better descriptor of the data. We see that frequent shoppers paid an average of 3.99 in delivery whereas the infrequent shoppers paid an average of 48.93 in delivery fees. We see that Frequent shoppers account for a greater percentage of the total delivery sales. So frequent shoppers pay more in delivery fees.

18. What percentage of total sales comes from each LOYALTY_SEGMENT?

```
twenty_total_sales = sum(df$cleaned_twenty_sales, na.rm=T)

bi_weekly_percent = (bi_weekly_sum/twenty_total_sales)*100
infrequent_percent = (infrequent_sum/twenty_total_sales)*100
three_week_percent = (three_week_sum/twenty_total_sales)*100
monthly_percent = (monthly_sum/twenty_total_sales)*100
weekly_percent = (weekly_sum/twenty_total_sales)*100
no_segment_percent = (no_segment_sum/twenty_total_sales)*100
new_shopper_percent = (new_shopper_sum/twenty_total_sales)*100

bi_weekly_percent; infrequent_percent; three_week_percent; monthly_percent; weekly_percent; no_segment_
```

```
## [1] 32.73927
```

```
## [1] 4.528876
```

```
## [1] 18.43612
```

```
## [1] 5.020866
```

```
## [1] 39.25822
```

```
## [1] 0.01162571
```

```
## [1] 0.005026499
```

Interpretation:

The percentage of total sales for each loyalty segment is given as 32.7392659, 4.5288759, 18.4361152, 5.0208664, 39.2582244, 0.0116257, 0.0050265 for bi_weekly_shoppers; infrequent_shoppers; three_week_shoppers; monthly_shoppers; weekly_shoppers; no_segment_shoppers; new_shopper_shoppers respectively

19. Are customers with earlier acquisition dates (older customers) more loyal in terms of orders?

```
# View(df)
```

Interpretation:

20. Which ZIP codes have the highest per-customer spending?

```
df$ZIP_CODE <- as.character(df$ZIP_CODE)

df %>%
  group_by(df$ZIP_CODE) %>%
  summarise(
    average = mean(cleaned_twenty_sales, na.rm = TRUE),
    n_customers = n()
  ) %>%
  arrange(desc(average))
```

```
## # A tibble: 144 x 3
##    `df$ZIP_CODE` average n_customers
##    <chr>           <dbl>       <int>
##  1 10465          37999            1
##  2 6807           36496            1
##  3 7003           15849            1
##  4 10310          15666            1
##  5 10044          15361            1
##  6 10069          11418.          14
##  7 11210          11409            5
##  8 10282          11399.          18
##  9 10701          10631            1
## 10 10463           9961            1
## # i 134 more rows
```

Interpretation:

This zip code of 10465 has the highest per customer spending

Part 3.

What is the day of the week with the highest average orders?

```
sunday_orders <- df$`SUNDAY ORDERS 12 MO.`
monday_orders <- df$`MONDAY ORDERS 12 MO.`
tuesday_orders <- df$`TUESDAY ORDERS 12 MO.`
wednesday_orders <- df$`WEDNESDAY ORDERS 12 MO.`
thursday_orders <- df$`THURSDAY ORDERS 12 MO.`
friday_orders <- df$`FRIDAY ORDERS 12 MO.`
saturday_orders <- df$`SATURDAY ORDERS 12 MO.`



sunday_mean <- mean(sunday_orders, na.rm=T)
monday_mean <- mean(monday_orders, na.rm=T)
tuesday_mean <- mean(tuesday_orders, na.rm=T)
wednesday_mean <- mean(wednesday_orders, na.rm=T)
thursday_mean <- mean(thursday_orders, na.rm=T)
friday_mean <- mean(friday_orders, na.rm=T)
saturday_mean <- mean(saturday_orders, na.rm=T)

day_mean <- c(sunday=sunday_mean, monday=monday_mean, tuesday=tuesday_mean, wednesday=wednesday_mean, t
)

sort(day_mean, decreasing=T)
```

```
##    sunday    monday   tuesday    friday  saturday wednesday  thursday
##  6.718762  6.116455  5.056281  5.044299  5.022288  4.362737  4.038823
```

Interpretation:

Sunday is the day with the highest average order.

21. Is weekend ordering (Sat+Sun) higher than weekday ordering?

```
weekend <- sum(df[c("SUNDAY ORDERS 12 MO.", "SATURDAY ORDERS 12 MO.")], na.rm=T)
weekday <- sum(df[c("MONDAY ORDERS 12 MO.", "TUESDAY ORDERS 12 MO.", "WEDNESDAY ORDERS 12 MO.", "THURSDA


weekend > weekday
```

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

Interpretation:

Weekend ordering 2.9096×10^4 is less than weekday ordering 6.0136×10^4

22. Do different LOYALTY_SEGMENTS prefer different days of the week?

```
weekend_df <- df[c("cleaned_loyalty","SUNDAY ORDERS 12 MO.", "SATURDAY ORDERS 12 MO.", "MONDAY ORDERS 1

weekend_df %>% group_by(weekend_df$cleaned_loyalty) %>% summarise(sunday    = mean(`SUNDAY ORDERS 12 MO
    monday    = mean(`MONDAY ORDERS 12 MO.`, na.rm = TRUE),
    tuesday   = mean(`TUESDAY ORDERS 12 MO.`, na.rm = TRUE),
    wednesday = mean(`WEDNESDAY ORDERS 12 MO.`, na.rm = TRUE),
    thursday  = mean(`THURSDAY ORDERS 12 MO.`, na.rm = TRUE),
    friday    = mean(`FRIDAY ORDERS 12 MO.`, na.rm = TRUE),
    saturday  = mean(`SATURDAY ORDERS 12 MO.`, na.rm = TRUE))
```

```
## # A tibble: 7 x 8
##   `weekend_df$cleaned_loyalty` sunday monday tuesday wednesday thursday friday
##   <chr>                         <dbl>  <dbl>   <dbl>     <dbl>    <dbl>  <dbl>
## 1 90 Day New Shoppers          NaN    NaN       1         2        1      2
## 2 Bi-Weekly Shoppers             7.56   6.76    5.44      4.40     4.23   4.91
## 3 Every Three Week Shoppers      4.19   3.87    3.34      2.90     2.98   3.48
## 4 Infrequent Shoppers            2.47   2.27    1.93      1.86     1.79   1.99
## 5 No Segment                     2.75   1.5     2.5       1        2      1
## 6 Once a Month Shoppers          3.12   2.53    2.42      2.28     2.07   2.48
## 7 Weekly Shoppers               12.0   10.6     8.68      7.74     6.67   9.31
## # i 1 more variable: saturday <dbl>
```

Interpretation:

Yes. For instance 90 day shoppers prefer wednesday and friday over the other days of the week. While bi-weekly shoppers prefer sunday over the other days of the week.

23. Do promo orders cluster on specific days (e.g., Fridays)?

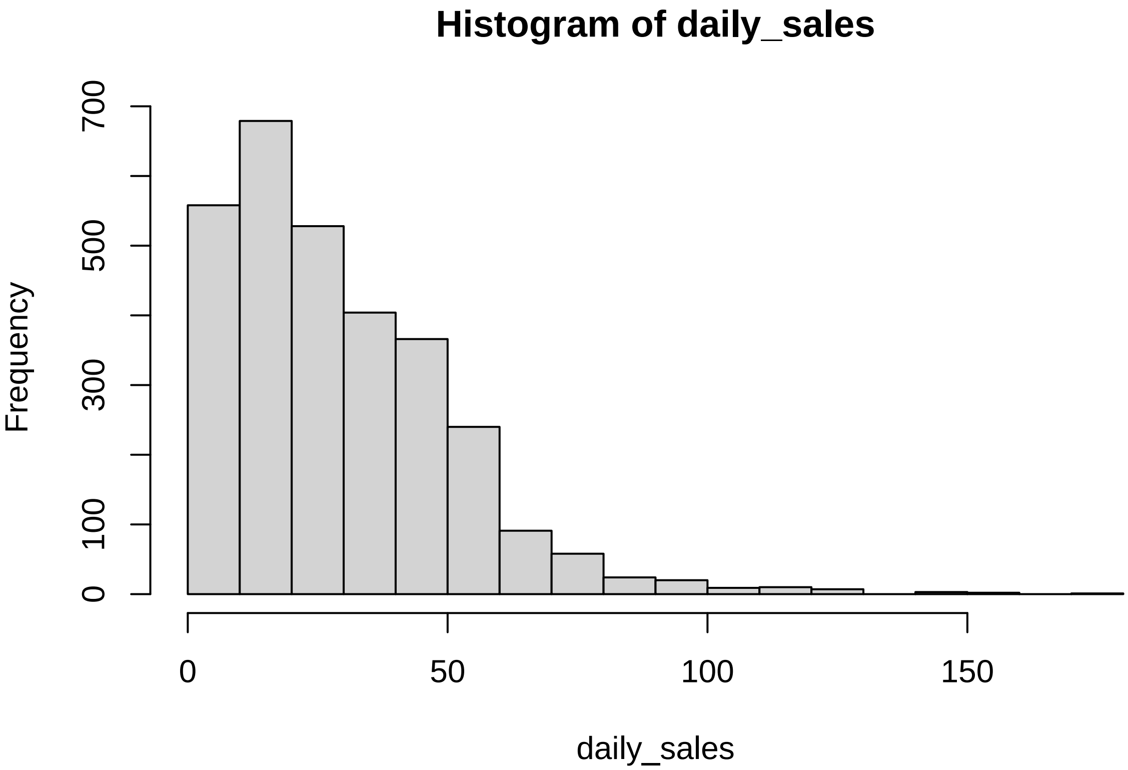
Interpretation:

24. Are sales more evenly distributed across days or skewed to a few?

Interpretation:

24. (Bonus) Draw at least 2 graphics and carefully interpret results.

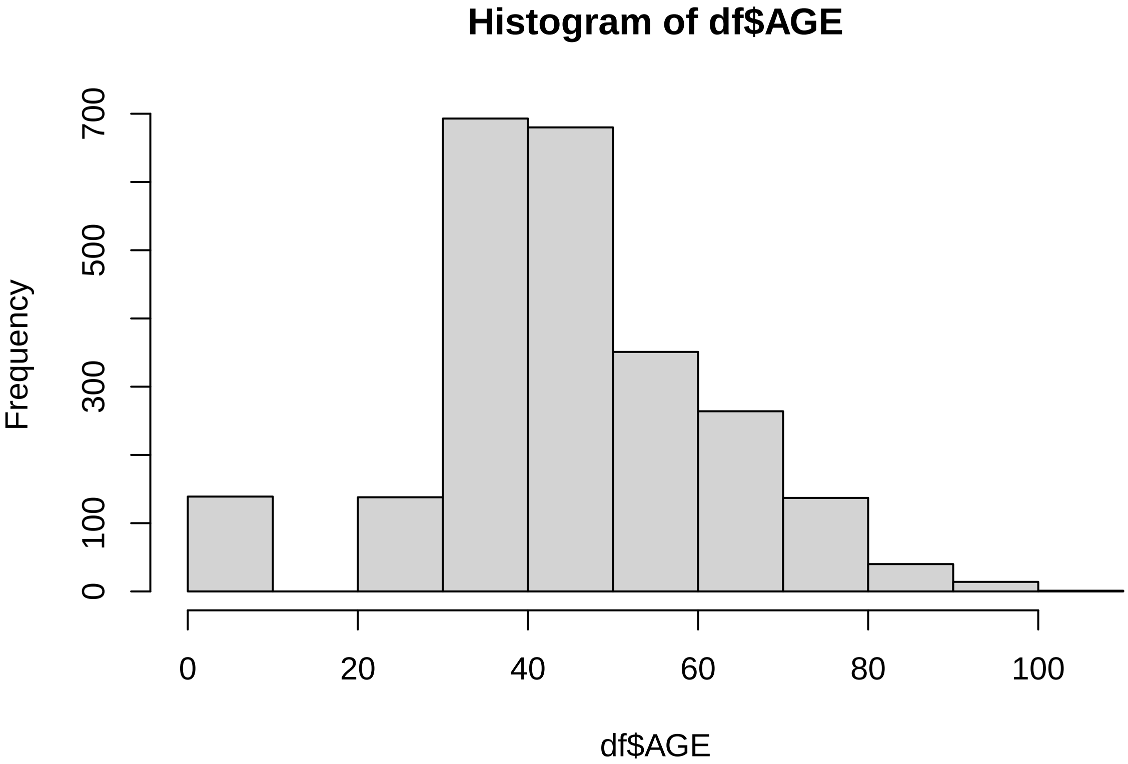
```
daily_sales <- rowSums(weekend_df[,-1], na.rm=T)
hist(daily_sales)
```



Interpretation

This histogram gives us an idea of the distribution of total weekly orders per customer. The horizontal axis represents the amount a customer spends per week while the vertical axis tells us about the number of people in each spending bracket. The histogram reveals that the data is skewed to the right, this means that more customers are spending less money per week on sales. We also see the presence of some outlier around the 150 mark which could probably be due to data entry errors, but regardless of the cause of the outlier, we know that it would significantly affect the mean of the dataset, and pull it towards the higher side of the scale. Thus the mean is not a good representation of the central point of the dataset.

```
hist(df$AGE)
```



Interpretation:

For the age variable, we see that there is a single outlier at the 0 mark, which could be due to a data entry error since it does not make sense for a customer at an online retail shop to be 0 years old. The data seems slightly skewed to the right which is represented in the values of the mean and median. mean is 45.2612943 and median is 44. The standard deviation is 17.0384248. Since the mean and median are within 1 standard deviation of each other, we can claim that this variable is roughly symmetric.

24. (Bonus) Perform at least one statistical test and interpret results.

Interpretation:

Submission

Submit the `.Rmd` and the knitted PDF files using the correct naming.