Roadmap for Supermarket Sales Analysis Project (Jun. 05. 2024)

Step 1: Project Planning and Setup

1. **Define Objectives:**
   - **Understand the relationship between sales and customer rating**
     - **Higher ratings lead to increased sales, and vice versa**
     - **Word of Mouth and recommendations that lead to improvement and innovation**
     - **Tools: R, Excel…etc. analysis**

2. **Gather Requirements**
   - **Datasets: https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales**
   - **Data Cleaning: Missing values, correct inconsistencies or errors in data**
   - **Data Transformation: Convert categorical variables to numerical values (if needed)**

3. **Data Analysis**
   - **Statistics: Calculate necessary statistics (mean, median, range…etc)**
   - **Visual Analysis: Histograms, boxplots, and scatterplots to find the correlation to understand distributions and relationships**
   - **Tools: R & Excel**

4. **Statistical Analaysis**
   - **Correlation analysis: Calculate correlation coefficients between Total sales and Customer Rating**
   - **Regression analysis: Try linear regression to model relationships between Total sales and Customer Rating**

#Welcome to My First Practice Project

I am a rising sophomore who recently completed an Introduction to Data class, focusing on learning the R programming language.
To expand my knowledge further, I am undertaking a practice project using R. This project aims to find the correlation between sales and customer ratings in a supermarket.

Before we dive into the analysis, we need to load the necessary packages and import the dataset. Let's get started!

First, we need to load the necessary packages to ensure we have all the tools required for data manipulation and visualization. Specifically, we will load the "tidyverse" and "readr" packages. The "tidyverse" pacakge includes "dplyr" for data manipulation, and "readr" is used for reading CSV files.

```
> library(tidyverse)
> library(readr)
> data <-read.csv("/Users/hazel/Desktop/Projects/supermarket_sales.csv",header=T,row.names=1)
> data
> summary(data)
```

```
    Branch             City          Customer.type        Gender          Product.line        Unit.price       Quantity        Tax.5.
 Length:1000        Length:1000        Length:1000        Length:1000        Length:1000        Min.   :10.08   Min.   : 1.00   Min.   : 0.5085
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:32.88   1st Qu.: 3.00   1st Qu.: 5.9249
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :55.23   Median : 5.00   Median :12.0880
                                                                             Mean   :55.67   Mean   : 5.51   Mean   :15.3794
                                                                             3rd Qu.:77.94   3rd Qu.: 8.00   3rd Qu.:22.4453
                                                                             Max.   :99.96   Max.   :10.00   Max.   :49.6500
     Total              Date               Time             Payment              cogs        gross.margin.percentage  gross.income
 Min.   :  10.68   Length:1000        Length:1000        Length:1000        Min.   : 10.17   Min.   :4.762            Min.   : 0.5085
 1st Qu.: 124.42   Class :character   Class :character   Class :character   1st Qu.:118.50   1st Qu.:4.762            1st Qu.: 5.9249
 Median : 253.85   Mode  :character   Mode  :character   Mode  :character   Median :241.76   Median :4.762            Median :12.0880
 Mean   : 322.97                                                           Mean   :307.59   Mean   :4.762            Mean   :15.3794
 3rd Qu.: 471.35                                                           3rd Qu.:448.90   3rd Qu.:4.762            3rd Qu.:22.4453
 Max.   :1042.65                                                           Max.   :993.00   Max.   :4.762            Max.   :49.6500
     Rating
 Min.   : 4.000
 1st Qu.: 5.500
 Median : 7.000
 Mean   : 6.973
 3rd Qu.: 8.500
 Max.   :10.000
```

We will use the "summary(data)" function to obtain basic statistics for each column, making the dataset easier to understand. Since more columns have vague values, such as "Branch" and "City". To find more specific values we will convert categorical columns to factors using "as.factor" function. Additionally we must check for any missing values using "sum(is.na(data))".

```
> sum(is.na(data))
> [1] 0
> data$Branch <-as.factor(data$Branch)
> data$City <-as.factor(data$City)
> data$Customer.type <-as.factor(data$Customer.type)
> data$Gender <-as.factor(data$Gender)
> data$Product.line <-as.factor(data$Product.line)
> data$Date <-as.factor(data$Date)
> data$Time <-as.factor(data$Time)
> data$Payment <-as.factor(data$Payment)
> summary(data)
```

```
 Branch         City     Customer.type    Gender                  Product.line   Unit.price       Quantity         Tax.5.
 A:340   Mandalay :332   Member:501     Female:501   Electronic accessories:170   Min.   :10.08   Min.   : 1.00   Min.   : 0.5085
 B:332   Naypyitaw:328   Normal:499     Male  :499   Fashion accessories    :178   1st Qu.:32.88   1st Qu.: 3.00   1st Qu.: 5.9249
 C:328   Yangon   :340                               Food and beverages     :174   Median :55.23   Median : 5.00   Median :12.0880
                                                    Health and beauty      :152   Mean   :55.67   Mean   : 5.51   Mean   :15.3794
                                                    Home and lifestyle     :160   3rd Qu.:77.94   3rd Qu.: 8.00   3rd Qu.:22.4453
                                                    Sports and travel      :166   Max.   :99.96   Max.   :10.00   Max.   :49.6500

     Total                Date          Time           Payment        cogs        gross.margin.percentage  gross.income        Rating
 Min.   :  10.68   2/7/2019 : 20   14:42  :  7   Cash       :344   Min.   : 10.17   Min.   :4.762           Min.   : 0.5085   Min.   : 4.000
 1st Qu.: 124.42   2/15/2019: 19   19:48  :  7   Credit card:311   1st Qu.:118.50   1st Qu.:4.762           1st Qu.: 5.9249   1st Qu.: 5.500
 Median : 253.85   1/8/2019 : 18   17:38  :  6   Ewallet    :345   Median :241.76   Median :4.762           Median :12.0880   Median : 7.000
 Mean   : 322.97   3/14/2019: 18   10:11  :  5                     Mean   :307.59   Mean   :4.762           Mean   :15.3794   Mean   : 6.973
 3rd Qu.: 471.35   3/2/2019 : 18   11:40  :  5                     3rd Qu.:448.90   3rd Qu.:4.762           3rd Qu.:22.4453   3rd Qu.: 8.500
 Max.   :1042.65   1/23/2019: 17   11:51  :  5                     Max.   :993.00   Max.   :4.762           Max.   :49.6500   Max.   :10.000
                   (Other)  :890   (Other):965
```

We can clearly see the difference in the previous data; it is now more understandable as we can identify the exact numbers for each column.

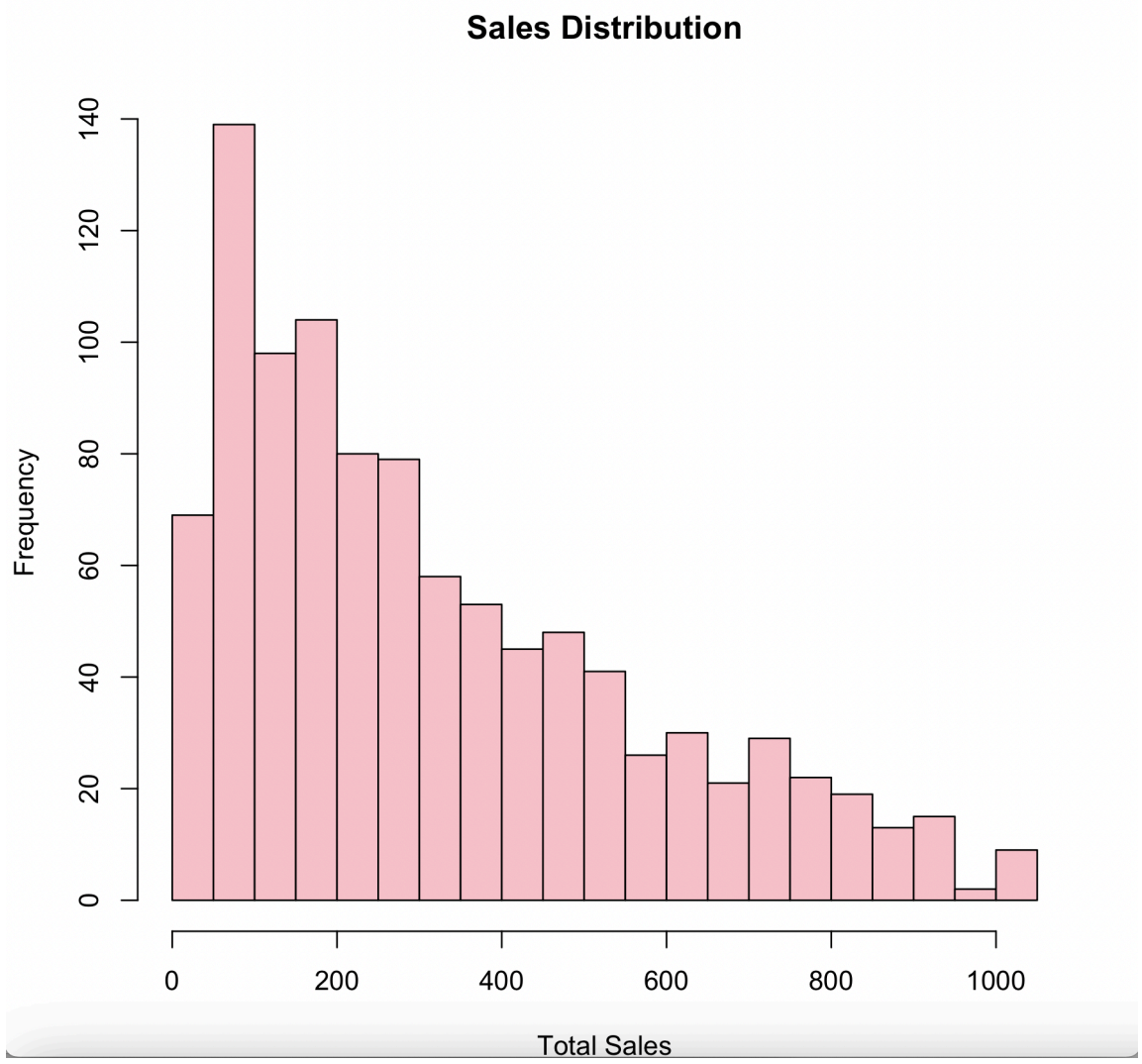
To obtain descriptive statistics, we need to get basic summary statistics for the columns of interest.

“Total” for total sales and “Rating” for customer ratings

```
> summary(data$Total)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.68  124.42  253.85  322.97  471.35 1042.65
> summary(data$Rating)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   5.500   7.000   6.973   8.500  10.000
```

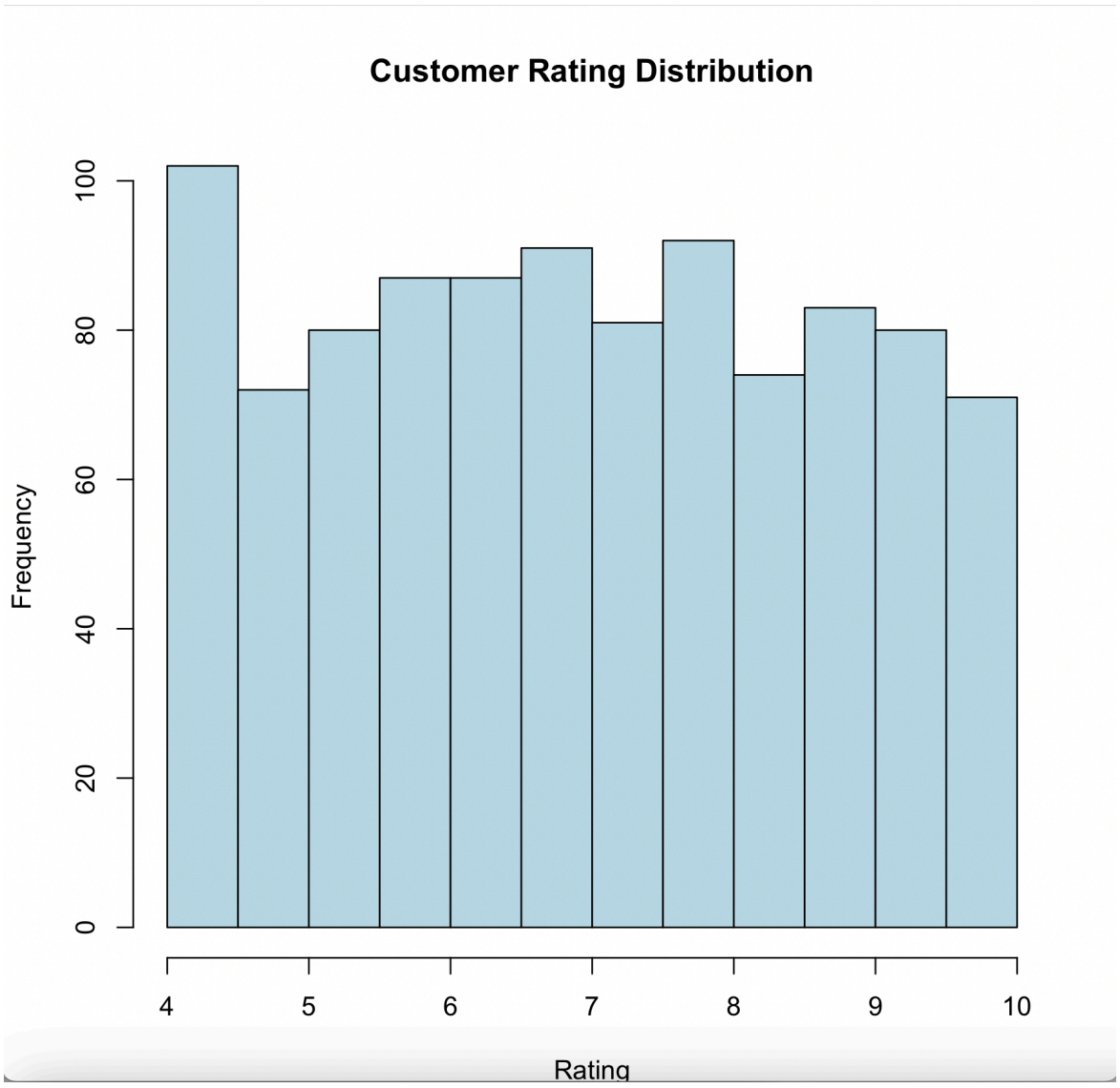
By examining these summary values, I can predict that the graph for “Total” will be positively skewed (right-skewed) because the mean value is higher than the median value. Conversely, for “Rating”. I expect it to be either negatively skewed or almost symmetric since the mean value is lower than the median value, but the values are quite close. Let’s verify this by creating a histogram.

```
> hist(data$Total, breaks=20, col="pink", main="Sales Distribution", xlab="Total Sales",
ylab="Frequency")
```



The histogram confirms that the distribution of total sales is right-skewed, as the mean value is higher than the median value. This indicates that a few higher sales values are pulling the mean upward. This skewness suggests that there could be some large sales outliers or a small number of very high sales transactions affecting the average, with the majority of sales transactions being lower than the mean.

```
> hist(data$Rating, breaks=20, col="lightblue", main="Customer Rating Distribution", xlab="Rating",
ylab="Frequency")
```



For customer ratings, the mean value is lower by 0.1 than the median value, indicating a slightly left-skewed or nearly symmetrical distribution. This shows that the distribution of customer ratings is fairly balanced, with a slight inclination towards higher ratings. The small difference between the mean and median (0.1) suggests that the ratings are quire evenly distributed without significant outliers.

Now, let's examine both "Total Sales" and "Customer Ratings" together by creating a scatter plot. Visualizing the data helps us understand distributions and spot any anomalies or tends.

```
> plot(data$Rating, data$Total, main="Sales vs Customer Rating", xlab="Customer Rating", ylab="Total Sales", pch=19, col="lightgreen")
```

The scatterplot reveals no clear relationships between customer ratings and total sales. The points are scattered across the plot, indicating that ratings do not have a strong positive or negative correlation with sales. No discernible pattern or trend is suggesting that higher ratings correlate with higher sales or vice versa.

To further analyze this, we will conduct a correlation analysis to measure the strength and direction of the linear relationship between sales and customer ratings.

```
> correlation <-cor(data$Total, data$Rating)
> print(correlation)
[1] -0.0364417
```

The correlation analysis shows a negative correlation value, indicating that higher total sales are associated with lower customer ratings, and vice versa. This could be due to several reasons. High sales might be drive by aggressive marketing or promotions, which could attract many customers but could lead to dissatisfaction if product or service quality is compromised to meet high demand.

For example, when a company launches a new product with significant marketing, sales might rise, but due to the rush to meet demand, product quality could suffer, resulting in negative reviews and lower customer ratings. Hence, while sales are high, customer ratings drop, resulting in a negative correlation.

Next, we will perform a regression analysis to model the relationship between customer ratings and sales, helping us understand how one variable affects the other.

```
> model <-lm(Total ~ Rating, data=data)
> summary(model)

Call:
lm(formula = Total ~ Rating, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-317.9 -198.6  -67.9  149.8  725.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  359.322     32.502  11.056   <2e-16 ***
Rating        -5.214      4.526  -1.152     0.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245.8 on 998 degrees of freedom
Multiple R-squared:  0.001328,  Adjusted R-squared:  0.0003273
F-statistic: 1.327 on 1 and 998 DF,  p-value: 0.2496
```

As these results do not provide clear evidence that customer ratings affect or are related to total sales, I decided to plot the regression line. Performing a regression analysis helps us determine the magnitude and direction of changes in the response variable when the explanatory variable changes. Since these results do not provide clear evidence that customer ratings affect or are related to total sales, I decided to plot the regression line. Performing a regression analysis helps us identify how much and in what direction the response variable changes when the explanatory variable changes.

```
> plot(data$Rating, data$Total, main="Sales vs Customer Rating", xlab="Customer Rating", ylab="Total
Sales", pch=19, col="blue")
> abline(lm(Total ~ Rating, data=data), col="red")
```

The linear regression results indicate a weak relationship between total sales and customer ratings, closer to no significant relationship. This shows that total sales tend to decrease slightly as customer ratings increase, or vice versa. However, since the relationship is weak, this trend is not very strong or consistent. This small inverse association indicates that customer ratings alone are not a reliable predictor of total sales.

In conclusion, the analysis shows a weak and negative correlation between total sales and customer ratings in the supermarket dataset. This suggests that while there is a small inverse relationship, it is not strong enough to use customer ratings as a reliable predictor of total sales. Further investigation into other factors affecting sales and customer satisfaction might be necessary to understand the dynamics fully.