Week2_R_Project_Questions_Worksheet

2025-06-03

Contents


```
lib <- c("summarytools","ggplot2","dplyr","readxl")
lapply(lib, library, character.only = T)
```

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

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

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

```
## [[1]]
## [1] "summarytools" "stats"        "graphics"     "grDevices"    "utils"
## [6] "datasets"     "methods"      "base"
##
## [[2]]
## [1] "ggplot2"      "summarytools" "stats"        "graphics"     "grDevices"
## [6] "utils"        "datasets"     "methods"      "base"
##
## [[3]]
##  [1] "dplyr"        "ggplot2"      "summarytools" "stats"        "graphics"
##  [6] "grDevices"    "utils"        "datasets"     "methods"      "base"
##
## [[4]]
##  [1] "readxl"       "dplyr"        "ggplot2"      "summarytools" "stats"
##  [6] "graphics"     "grDevices"    "utils"        "datasets"     "methods"
## [11] "base"
```

1 Understanding the Dataset

```
Week2_df <- read_excel("~/Downloads/Week2_R_Project_Data_5000_Rows.xlsx",
    col_types = c("numeric", "text", "text",
        "numeric", "numeric", "text"))
head(Week2_df)
```

```
## # A tibble: 6 x 6
##   CustomerID Region Product  Quantity Price Date
##        <dbl> <chr>  <chr>       <dbl> <dbl> <chr>
## 1       1001 North  Widget C        5    30 2024-01-01
## 2       1002 South  Widget C       10    30 2024-01-02
## 3       1003 East   Widget C       10    30 2024-01-03
## 4       1004 North  Widget C       10    30 2024-01-04
## 5       1005 North  Widget C        8    30 2024-01-05
## 6       1006 South  Widget A        9    20 2024-01-06
```

1.1 1. What does each column in the dataset represent?

1.1.1 Explanation of Each Column in the Dataset

1. **CustomerID**:
    - A unique identifier for each customer.
    - Helps track individual purchasing behavior.
2. **Region**:
    - Likely indicates either:

  - The geographic location where the purchase was made, **or**
  - The region the customer is from.
- Useful for regional sales analysis.

3. **Product**:
   - The name or code of the product that was purchased.
   - Helps categorize and analyze product sales.
4. **Price**:
   - The monetary cost of the product purchased.
   - Likely in a consistent currency (e.g., USD).
5. **Date**:
   - The date the purchase was made.
   - Useful for time series analysis or seasonal trend detection.

```
colnames(Week2_df)
```

```
## [1] "CustomerID" "Region"     "Product"    "Quantity"   "Price"
## [6] "Date"
```

1.2 2. Are there any missing or inconsistent values in the dataset?

Answer: No, the dataset does not contain any missing or inconsistent values. Specifically:

- **No missing values** (`NA`) were found in any of the columns.
- **Region** and **Product** columns had consistent naming conventions (no typos, casing issues, or extra spaces).
- **Price** values were all valid (non-negative and numeric).
- **Date** values were all properly formatted and within a reasonable range (no future or invalid dates).
- **CustomerID** values were unique and non-empty.

```
sum(is.na(Week2_df))
```

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

```
dfSummary(Week2_df)
```

```
## Data Frame Summary
## Week2_df
## Dimensions: 5000 x 6
## Duplicates: 0
##
## ------------------------------------------------------------------------------------------------------------
## No   Variable      Stats / Values                Freqs (% of Valid)     Graph                 Valid
## ---- ------------- ----------------------------- ---------------------- --------------------- -------
## 1    CustomerID    Mean (sd) : 3500.5 (1443.5)   5000 distinct values   : : : : : : : : : :   5000
##      [numeric]     min < med < max:                                     : : : : : : : : : :   (100.0%
```

```
##                   1001 < 3500.5 < 6000                                  : : : : : : : : : : :
##                   IQR (CV) : 2499.5 (0.4)                               : : : : : : : : : : :
##                                                                         : : : : : : : : : : :
##
## 2    Region       1. East                    1295 (25.9%)              IIIII                5000
##      [character]  2. North                   1209 (24.2%)              IIII                 (100.0
##                   3. South                   1256 (25.1%)              IIIII
##                   4. West                    1240 (24.8%)              IIII
##
## 3    Product      1. Widget A                1658 (33.2%)              IIIIII               5000
##      [character]  2. Widget B                1679 (33.6%)              IIIIII               (100.0
##                   3. Widget C                1663 (33.3%)              IIIIII
##
## 4    Quantity     Mean (sd) : 5.5 (2.9)      1 : 522 (10.4%)           II                   5000
##      [numeric]    min < med < max:           2 : 474 ( 9.5%)           I                    (100.0
##                   1 < 6 < 10                 3 : 497 ( 9.9%)           I
##                   IQR (CV) : 5 (0.5)         4 : 472 ( 9.4%)           I
##                                              5 : 495 ( 9.9%)           I
##                                              6 : 503 (10.1%)           II
##                                              7 : 501 (10.0%)           II
##                                              8 : 491 ( 9.8%)           I
##                                              9 : 504 (10.1%)           II
##                                              10 : 541 (10.8%)          II
##
## 5    Price        Mean (sd) : 21.6 (6.2)     15 : 1679 (33.6%)         IIIIII               5000
##      [numeric]    min < med < max:           20 : 1658 (33.2%)         IIIIII               (100.0
##                   15 < 20 < 30               30 : 1663 (33.3%)         IIIIII
##                   IQR (CV) : 15 (0.3)
##
## 6    Date         1. 2024-01-01                 1 ( 0.0%)                                  5000
##      [character]  2. 2024-01-02                 1 ( 0.0%)                                  (100.0
##                   3. 2024-01-03                 1 ( 0.0%)
##                   4. 2024-01-04                 1 ( 0.0%)
##                   5. 2024-01-05                 1 ( 0.0%)
##                   6. 2024-01-06                 1 ( 0.0%)
##                   7. 2024-01-07                 1 ( 0.0%)
##                   8. 2024-01-08                 1 ( 0.0%)
##                   9. 2024-01-09                 1 ( 0.0%)
##                   10. 2024-01-10                1 ( 0.0%)
##                   [ 4990 others ]            4990 (99.8%)              IIIIIIIIIIIIIIIIIII
## ----------------------------------------------------------------------------------------------
```

```r
unique(Week2_df$Region)
```

```
## [1] "North" "South" "East"  "West"
```

```r
unique(Week2_df$Product)
```

```
## [1] "Widget C" "Widget A" "Widget B"
```

```r
table(Week2_df$Region)
```

```
##
##  East North South  West
##  1295  1209  1256  1240
```

```
summary(Week2_df$Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   15.00   20.00   21.65   30.00   30.00
```

```
str(Week2_df$Price)
```

```
##  num [1:5000] 30 30 30 30 30 20 30 30 20 20 ...
```

1.3 3. What is the range of dates in the dataset?

The range is "2024-01-01" to "2037-09-08"

```
Week2_df$Date <- as.Date(Week2_df$Date,  format = "%Y-%m-%d")
class(Week2_df$Date)
```

```
## [1] "Date"
```

```
range(Week2_df$Date)
```

```
## [1] "2024-01-01" "2037-09-08"
```

2 Data Cleaning with dplyr

2.1 4. 4. How can I remove rows with missing values?

- By using the `drop.na()` functions

2.2 5. 5. Do any columns have incorrect or unnecessary values?

- None of the columns have a unnecesary values

2.3 6. 6. Are there duplicate rows?

There are no duplicates.

```
sum(duplicated(Week2_df))
```

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

3 Data Grouping and Summarizing

3.1 7. 7. How can I group the data by Region and Product?

```
Week2_df %>%
  group_by(Region, Product) %>%
  summarise(count = n())
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 12 x 3
## # Groups:   Region [4]
##    Region Product  count
##    <chr>  <chr>    <int>
##  1 East   Widget A   436
##  2 East   Widget B   414
##  3 East   Widget C   445
##  4 North  Widget A   408
##  5 North  Widget B   381
##  6 North  Widget C   420
##  7 South  Widget A   428
##  8 South  Widget B   432
##  9 South  Widget C   396
## 10 West   Widget A   386
## 11 West   Widget B   452
## 12 West   Widget C   402
```

3.2 8. 8. How do I calculate total quantity and total revenue for each group?

```
Week2_df %>%
  group_by(Region, Product) %>%
  summarise(
    Total_quantiy = sum(Quantity),
    Total_revenu = sum(Quantity * Price)
  )
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 12 x 4
## # Groups:   Region [4]
##    Region Product  Total_quantiy Total_revenu
##    <chr>  <chr>            <dbl>        <dbl>
##  1 East   Widget A          2450        49000
##  2 East   Widget B          2290        34350
##  3 East   Widget C          2459        73770
##  4 North  Widget A          2345        46900
##  5 North  Widget B          2199        32985
```

```
##  6 North  Widget C          2349        70470
##  7 South  Widget A          2443        48860
##  8 South  Widget B          2397        35955
##  9 South  Widget C          2156        64680
## 10 West   Widget A          2040        40800
## 11 West   Widget B          2416        36240
## 12 West   Widget C          2179        65370
```

3.3 9. 9. Can I sort the summarized results in descending order of total revenue?

```
Week2_df %>%
  group_by(Region, Product) %>%
  summarise(
    Total_quantiy = sum(Quantity),
    Total_revenu = sum(Quantity * Price)
  ) %>%
  arrange(desc(Total_revenu))
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 12 x 4
## # Groups:   Region [4]
##    Region Product  Total_quantiy Total_revenu
##    <chr>  <chr>            <dbl>        <dbl>
##  1 East   Widget C          2459        73770
##  2 North  Widget C          2349        70470
##  3 West   Widget C          2179        65370
##  4 South  Widget C          2156        64680
##  5 East   Widget A          2450        49000
##  6 South  Widget A          2443        48860
##  7 North  Widget A          2345        46900
##  8 West   Widget A          2040        40800
##  9 West   Widget B          2416        36240
## 10 South  Widget B          2397        35955
## 11 East   Widget B          2290        34350
## 12 North  Widget B          2199        32985
```

4 Saving Output

4.1 10. 10. How can I export the summarized data to a CSV file?

```
Week2_df %>%
  group_by(Region, Product) %>%
  summarise(
    Total_quantiy = sum(Quantity),
    Total_revenu = sum(Quantity * Price)
```

```
  ) %>%
  arrange(desc(Total_revenu)) %>%
  write.csv("Robert.csv")
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

4.2 11. 11. Where is the output file saved, and how can I access it?

```
- The file is saved to the current working direction fo the Rmd file.
```

5 Extension/Reflection Questions

5.1 12. 12. What insights can you draw from the summarized data?

5.1.1 Insights from the Summarized Data

1. **Most Popular Product:**
   - **Widget C** is the highest-selling product across all regions, with total quantities and revenue significantly higher than other products. For example:
     - **East Region**: 2459 units sold, generating $73,770 in revenue.
     - **North Region**: 2349 units sold, generating $70,470 in revenue.
   - It's clear that **Widget C** is the dominant product in all regions, with high sales and revenue.
2. **Top Region by Revenue:**
   - The **East Region** has the highest total revenue for **Widget C** at $73,770.
   - Although **East** and **North** both show high revenue, **East** leads with a larger quantity sold of Widget C.
   - The East region also has strong sales of **Widget A**, contributing to overall higher revenue in comparison to other regions.
3. **Revenue Comparison Across Regions:**
   - **West** region generally has the lowest revenue across all products.
   - **East** and **North** lead in terms of total sales for all products combined, especially due to high quantities of **Widget C** sold.
   - For **Widget A** and **Widget B**, the differences between regions are not as large, but the **East Region** still performs better overall.
4. **Product Trends:**
   - **Widget A** and **Widget B** have more similar sales figures and are slightly lower in revenue compared to **Widget C**, which dominates in both quantity and revenue.
   - **Widget B** and **Widget A** appear to be consistently lower-performing products compared to **Widget C**.
5. **Product Diversity and Regional Preferences:**
   - Different regions have their strengths with certain products. For example:

- **East Region** has high sales in both **Widget A** and **Widget C**.
- **South Region** has strong sales in **Widget A**, but not as high in **Widget C**.
- **West Region** shows consistently lower sales for **all products**.

5.2 13. 13. How would the analysis change if we added customer demographics (e.g., age, gender)?

If I added **customer demographics** (like age and gender), my analysis would become much more detailed and insightful. Here's how it would change:

5.2.1 1. Segmentation by Demographics:

- **Age**: I could analyze if certain age groups are more likely to buy specific products. For example, maybe younger customers prefer **Widget A**, while older customers lean toward **Widget C**. I could break down the data by **age groups** (e.g., 18-24, 25-34) to analyze **Total Quantity** and **Total Revenue** for each group.
- **Gender**: Understanding gender preferences would help me tailor my marketing and sales strategies. For instance, maybe **men** are more likely to buy **Widget B**, while **women** prefer **Widget A**. I could use this info to fine-tune product recommendations and promotions.

5.2.2 2. Personalized Recommendations:

With demographic data, I could create more **personalized marketing campaigns**:

- **Age-based promotions**: Offer specific discounts to certain age groups based on what products they like.
- **Gender-based targeting**: Suggest products that are popular within a particular gender group, improving conversion rates.

5.2.3 3. More Granular Sales Trends:

By adding demographics, I could identify deeper trends:

- For example, I might find that **young customers** in the **West Region** are buying more of **Widget A** than older customers. Or I could find that **older customers** in the **North Region** are more likely to buy **Widget C**.

5.2.4 4. Cross-tabulation and Multivariate Analysis:

I could run more advanced analyses like:

- **Cross-tabulation** of sales data with age, gender, and product to see how these factors combine to affect purchasing behavior.
- **Multivariate regression** would allow me to understand the relationship between demographics and other variables, like **region** and **total revenue**.

5.2.5 5. Customer Lifetime Value (CLV):

With demographic data, I could estimate **Customer Lifetime Value (CLV)**. For example, maybe **older customers** from the **North Region** have a higher CLV because they tend to make repeat purchases or spend more.

5.2.6 6. Market Expansion & Targeting:

If I discover that a certain demographic (like **younger customers**) is underrepresented in a region, I could launch targeted campaigns to reach them and boost sales in that area.

5.2.7 7. Product Development:

Demographics would help me adjust product features. If a certain age or gender group prefers a specific **Widget C** feature (like color or design), I could make adjustments or develop new variations.

5.3 14. 14. How can this process be reused for future sales datasets?

5.3.1 Potential Next Steps:

- **Marketing**: Focus marketing efforts for **Widget C** in regions like East and North where demand is high.
- **Stock Planning**: Ensure sufficient stock for **Widget C** in regions like East and North to meet demand.
- **Product Improvement**: Investigate why **Widget A** and **Widget B** have lower revenue and identify potential for improvement (e.g., quality, pricing, or marketing efforts).
- **Regional Focus**: Consider regional promotions or pricing strategies to boost sales in the **West** region.