DESCRIPTIVE STATISTICS

Enock Bereka

2025-08-19

Load Libraries

```r
library(dplyr)       # data manipulation
library(ggplot2)     # visualization
library(psych)       # descriptive statistics
library(moments)     # skewness, kurtosis
library(tidyr)       # data wrangling
library(corrplot)    # correlation plots
library(tidyverse)
library(flextable)
library(ggcorrplot)
```

Load Data

```r
data <- read.csv("C:/Users/ADMIN/Desktop/Data Science/Datasets/Business
Intelligence/Sample - Superstore.csv")
```

Inspect Data

```r
h = head(data, 10)        # first few rows
flextable(h) %>%
  autofit() %>%
  theme_box() %>%
  color(part = "header", color = "white") %>%
  bg(part = "header", bg = "steelblue") %>%
  bold(part = "header")
```

Order.Date	Ship.Date	Ship.Mode	Segment	Country	City	State	Discount	Profit
11/8/2016	11/11/2016	Second Class	Consumer	United States	Henderson	Kentucky	0.00	41.9136
11/8/2016	11/11/2016	Second Class	Consumer	United States	Henderson	Kentucky	0.00	219.5820
6/12/2016	6/16/2016	Second Class	Corporate	United States	Los Angeles	California	0.00	6.8714
10/11/2015	10/18/2015	Standard Class	Consumer	United States	Fort Lauderdale	Florida	0.45	-383.0310
10/11/201	10/18/201	Standard	Consum	United	Fort	Florida	0.20	2.5164

Order.Date	Ship.Date	Ship.Mode	Segment	Country	City	State	Discount	Profit
5	5	Class	er	States	Lauderdale			
6/9/2014	6/14/2014	Standard Class	Consumer	United States	Los Angeles	California	0.00	14.1694
6/9/2014	6/14/2014	Standard Class	Consumer	United States	Los Angeles	California	0.00	1.9656
6/9/2014	6/14/2014	Standard Class	Consumer	United States	Los Angeles	California	0.20	90.7152
6/9/2014	6/14/2014	Standard Class	Consumer	United States	Los Angeles	California	0.20	5.7825
6/9/2014	6/14/2014	Standard Class	Consumer	United States	Los Angeles	California	0.00	34.4700

```
glimpse(data)         # structure of dataset
```

```
## Rows: 9,994
## Columns: 21
## $ Row.ID        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16, 1…
## $ Order.ID      <chr> "CA-2016-152156", "CA-2016-152156", "CA-2016-
138688", "U…
## $ Order.Date    <chr> "11/8/2016", "11/8/2016", "6/12/2016", "10/11/2015",
"10…
## $ Ship.Date     <chr> "11/11/2016", "11/11/2016", "6/16/2016",
"10/18/2015", "…
## $ Ship.Mode     <chr> "Second Class", "Second Class", "Second Class",
"Standar…
## $ Customer.ID   <chr> "CG-12520", "CG-12520", "DV-13045", "SO-20335", "SO-
2033…
## $ Customer.Name <chr> "Claire Gute", "Claire Gute", "Darrin Van Huff",
"Sean O…
## $ Segment       <chr> "Consumer", "Consumer", "Corporate", "Consumer",
"Consum…
## $ Country       <chr> "United States", "United States", "United States",
"Unit…
## $ City          <chr> "Henderson", "Henderson", "Los Angeles", "Fort
Lauderdal…
## $ State         <chr> "Kentucky", "Kentucky", "California", "Florida",
"Florid…
## $ Postal.Code   <int> 42420, 42420, 90036, 33311, 33311, 90032, 90032,
90032, …
## $ Region        <chr> "South", "South", "West", "South", "South", "West",
"Wes…
## $ Product.ID    <chr> "FUR-BO-10001798", "FUR-CH-10000454", "OFF-LA-
10000240",…
```

```
## $ Category      <chr> "Furniture", "Furniture", "Office Supplies",
"Furniture"…
## $ Sub.Category  <chr> "Bookcases", "Chairs", "Labels", "Tables",
"Storage", "F…
## $ Product.Name  <chr> "Bush Somerset Collection Bookcase", "Hon Deluxe
Fabric …
## $ Sales         <dbl> 261.9600, 731.9400, 14.6200, 957.5775, 22.3680,
48.8600,…
## $ Quantity      <int> 2, 3, 2, 5, 2, 7, 4, 6, 3, 5, 9, 4, 3, 3, 5, 3, 6,
2, 2,…
## $ Discount      <dbl> 0.00, 0.00, 0.00, 0.45, 0.20, 0.00, 0.00, 0.20,
0.20, 0.…
## $ Profit        <dbl> 41.9136, 219.5820, 6.8714, -383.0310, 2.5164,
14.1694, 1…

data$Row.ID = NULL
data$Postal.Code = NULL
```

Interpretation:

- head() shows the first rows for a preview

- glimpse() shows data types (numeric, factor, character)

Summary Statistics

```
num_data <- data %>% select_if(is.numeric)
summary(num_data)

##      Sales              Quantity        Discount          Profit
##  Min.   :    0.444   Min.   : 1.00   Min.   :0.0000   Min.   :-6599.978
##  1st Qu.:   17.280   1st Qu.: 2.00   1st Qu.:0.0000   1st Qu.:    1.729
##  Median :   54.490   Median : 3.00   Median :0.2000   Median :    8.666
##  Mean   :  229.858   Mean   : 3.79   Mean   :0.1562   Mean   :   28.657
##  3rd Qu.:  209.940   3rd Qu.: 5.00   3rd Qu.:0.2000   3rd Qu.:   29.364
##  Max.   :22638.480   Max.   :14.00   Max.   :0.8000   Max.   : 8399.976
```

Measures of Central Tendency

Sales

```
mean(data$Sales, na.rm = TRUE)

## [1] 229.858

median(data$Sales, na.rm = TRUE)

## [1] 54.49
```

Mode function

```r
get_mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
get_mode(data$Sales)

## [1] 12.96
```

Interpretation:

mean → average sales

median → middle sales (less affected by outliers)

mode → most frequent sales value

unique(x)

Extracts all the unique values in the vector x.

Example: if x = c(2, 3, 2, 5), then unique(x) = c(2, 3, 5).

match(x, ux)

Finds the position of each element of x in the list of unique values.

Example: for x = c(2, 3, 2, 5) and ux = c(2, 3, 5), we get match(x, ux) = c(1, 2, 1, 3).

tabulate(match(x, ux))

Counts how many times each unique value appears.

Example: tabulate(c(1, 2, 1, 3)) = c(2, 1, 1)

(meaning: value 2 appears 2 times, 3 appears 1 time, 5 appears 1 time).

which.max(...)

Finds the position of the most frequent value.

Example: which.max(c(2, 1, 1)) = 1 (the first element has the maximum count).

ux[...]

Returns the actual value from the unique values corresponding to the max count.

Example: ux[1] = 2, which is the mode.

Measures of Dispersion

```r
range(data$Sales, na.rm = TRUE)
```

```
## [1]     0.444 22638.480
var(data$Sales, na.rm = TRUE)
## [1] 388434.5
sd(data$Sales, na.rm = TRUE)
## [1] 623.2451
IQR(data$Sales, na.rm = TRUE)
## [1] 192.66
```

Coefficient of Variation

```
sd(data$Sales, na.rm = TRUE) / mean(data$Sales, na.rm = TRUE)
## [1] 2.711435
```

Interpretation:

Range → spread between smallest & largest sales

Variance/SD → average deviation from mean

IQR → spread of middle 50% of data

CV → relative variability (SD as % of mean)

Distribution Shape

```
skewness(data$Sales, na.rm = TRUE)
## [1] 12.97081
kurtosis(data$Sales, na.rm = TRUE)
## [1] 308.1584
```

Interpretation:

Skewness > 0 → right-skewed (long tail on right)

Skewness < 0 → left-skewed

Kurtosis > 3 → leptokurtic (peaked), < 3 → platykurtic (flat)

Frequency Distributions (Categorical Variables)

```
table(data$Category)
```

```
## 
##       Furniture Office Supplies      Technology 
##            2121            6026            1847 
```

```r
prop.table(table(data$Category)) * 100
```

```
## 
##       Furniture Office Supplies      Technology 
##        21.22273        60.29618        18.48109 
```

Interpretation:

Table → counts per category

Prop.table → percentage distribution

Cross-tabulations

Category vs Region

```r
table(data$Category, data$Region)
```

```
## 
##                    Central East South West
##   Furniture            481  601   332  707
##   Office Supplies     1422 1712   995 1897
##   Technology           420  535   293  599
```

```r
prop.table(table(data$Category, data$Region), margin=2) * 100
```

```
## 
##                    Central     East    South     West
##   Furniture       20.70598 21.10253 20.49383 22.07306
##   Office Supplies 61.21395 60.11236 61.41975 59.22573
##   Technology      18.08007 18.78511 18.08642 18.70122
```

Interpretation:

Cross-tab shows how categories are distributed across regions

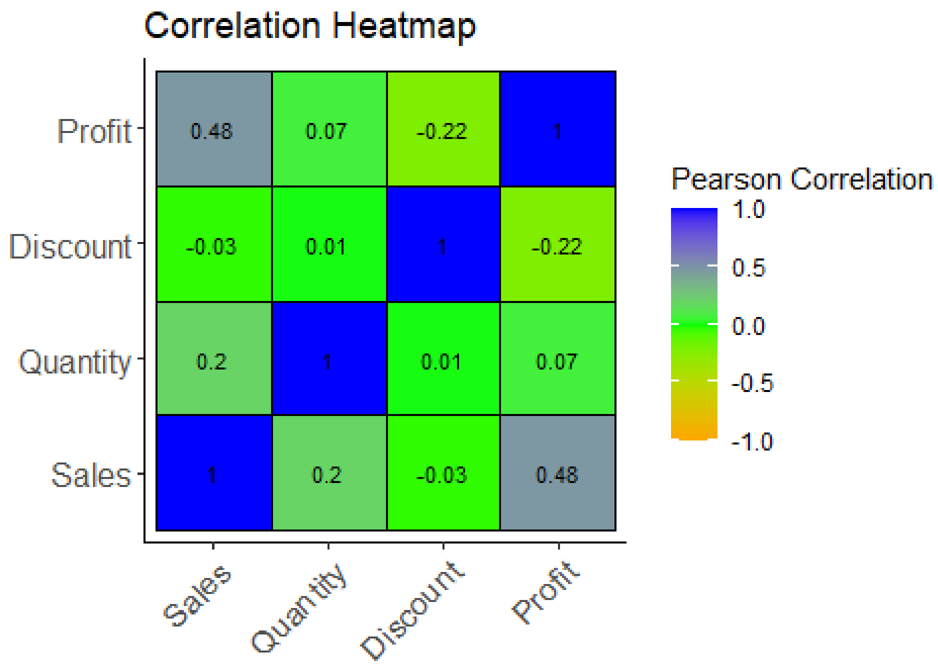
Proportions help in comparing relative distribution

Correlation & Association

```r
num_data <- data %>% select_if(is.numeric)
cor_matrix <- cor(num_data, use="pairwise.complete.obs")
```

Correlation heatmap

```r
ggcorrplot(cor_matrix,title = "Correlation Heatmap",lab_col = "black",
           lab = TRUE, legend.title = "Pearson Correlation",
           lab_size = 3, ggtheme = theme_classic(),
           outline.color = "black",
           colors = c("orange", "green", "blue"))
```



Interpretation:

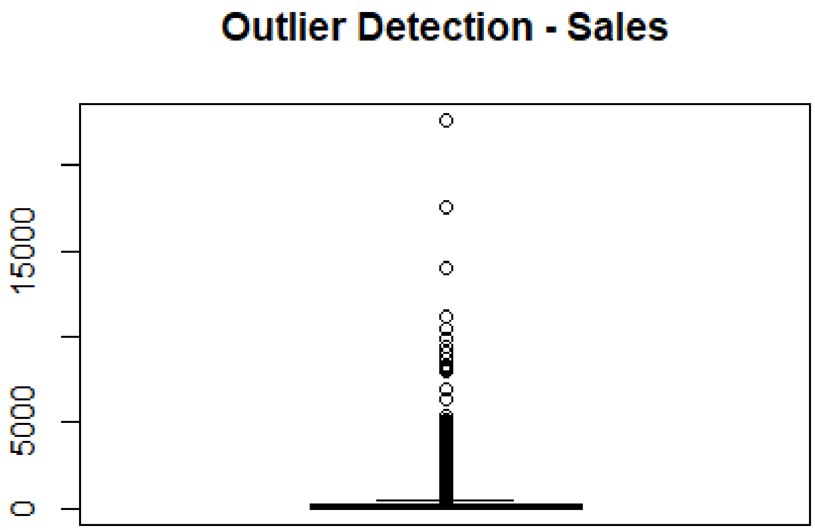
Correlation ranges from -1 (perfect negative) to +1 (perfect positive)

Helps identify relationships between Sales, Profit, Discount, etc.

Outlier Detection

Boxplot method

```
boxplot(data$Sales, main="Outlier Detection - Sales")
```



Z-score method

```
z_scores <- scale(data$Sales)
which(abs(z_scores) > 3)

##   [1]   28  166  252  263  264  319  354  400  488  510  516  684  978
995 1002
##  [16] 1086 1156 1247 1439 1455 1645 1792 1804 1806 2183 2419 2493 2506
2568 2624
##  [31] 2625 2698 2849 3012 3056 3071 3274 3281 3444 3570 3581 3984 3987
4094 4099
##  [46] 4129 4191 4219 4278 4298 4620 4866 4882 5007 5127 5171 5199 5301
5321 5531
##  [61] 5563 5627 5711 5885 5918 5991 6011 6015 6100 6102 6117 6210 6341
6426 6521
##  [76] 6535 6536 6621 6627 6818 6827 6869 6885 6902 7174 7244 7281 7475
7488 7580
##  [91] 7584 7667 7684 7773 7819 7915 7938 8101 8103 8154 8205 8237 8272
8313 8425
## [106] 8469 8489 8681 8700 8750 8859 8893 8991 9040 9057 9166 9271 9413
9426 9640
## [121] 9650 9661 9742 9775 9858 9930 9949
```

Interpretation:

Boxplot highlights outliers as points beyond whiskers

Z-scores > |3| indicate extreme outliers

Summary Table for Numerical Variables

Select numeric columns

```r
num_data <- data %>% select_if(is.numeric)
```

Compute summary statistics

```r
num_summary <- num_data %>%
  summarise_all(list(
    Mean = ~mean(., na.rm=TRUE),
    Median = ~median(., na.rm=TRUE),
    SD = ~sd(., na.rm=TRUE),
    Min = ~min(., na.rm=TRUE),
    Max = ~max(., na.rm=TRUE),
    Skewness = ~skewness(., na.rm=TRUE),
    Kurtosis = ~kurtosis(., na.rm=TRUE)
  )) %>%
  pivot_longer(cols = everything(),
               names_to = c("Variable", ".value"),
               names_sep = "_")
```

Convert to flextable

```r
flextable(num_summary) %>%
  autofit() %>%
  theme_box() %>%
  color(part = "header", color = "white") %>%
  bg(part = "header", bg = "steelblue") %>%
  bold(part = "header")
```

Variable	Mean	Median	SD	Min	Max	Skewness	Kurtosis
Sales	229.8580008	54.4900	623.245101	0.444	22,638.480	12.970805	308.158427
Quantity	3.7895737	3.0000	2.225110	1.000	14.000	1.278353	4.990293
Discount	0.1562027	0.2000	0.206452	0.000	0.800	1.684042	5.407740
Profit	28.6568963	8.6665	234.260108	-6,599.978	8,399.976	7.560297	399.989229

Frequency Table for Categorical Variables

Category variable

```r
cat_summary <- data %>%
  group_by(Category) %>%
  summarise(Count = n(),
            Percent = n()/nrow(data)*100)

flextable(cat_summary) %>%
  autofit() %>%
  theme_box() %>%
  color(part = "header", color = "white") %>%
  bg(part = "header", bg = "steelblue") %>%
  bold(part = "header")
```

Category	Count	Percent
Furniture	2,121	21.22273
Office Supplies	6,026	60.29618
Technology	1,847	18.48109

Cross-tabulation with Flextable

Category vs Region

```r
cross_tab <- table(data$Category, data$Region) %>% as.data.frame()

## Rename columns
colnames(cross_tab) <- c("Category", "Region", "Count")

## Add percentages
cross_tab <- cross_tab %>%
  group_by(Region) %>%
  mutate(Percent = Count / sum(Count) * 100)

flextable(cross_tab) %>%
  autofit() %>%
  theme_box() %>%
  color(part = "header", color = "white") %>%
  bg(part = "header", bg = "steelblue") %>%
  bold(part = "header")
```

Category	Region	Count	Percent
Furniture	Central	481	20.70598

Category	Region	Count	Percent
Office Supplies	Central	1,422	61.21395
Technology	Central	420	18.08007
Furniture	East	601	21.10253
Office Supplies	East	1,712	60.11236
Technology	East	535	18.78511
Furniture	South	332	20.49383
Office Supplies	South	995	61.41975
Technology	South	293	18.08642
Furniture	West	707	22.07306
Office Supplies	West	1,897	59.22573
Technology	West	599	18.70122