Abstract

Title: A Data Analytics Approach to Understanding Sales Performance and Profitability

This report presents a data analytics case study developed using R programming. The study utilizes the **Superstore Sales dataset** from Kaggle to analyze relationships between **Sales, Profit, and Discounts** across different regions. It covers fundamental techniques such as descriptive statistics, data exploration, data cleaning, and regression modeling. Visualizations including histograms, scatterplots, bar charts, and boxplots are used to illustrate patterns in the dataset. The objective is to provide insights into how discounts and regional factors affect profitability, helping businesses optimize strategies.

Introduction

Problem Statement:

Consider a dataset that records **Sales, Profit, Discounts, and Region** for a retail superstore. The goal is to understand how these factors influence business profitability.

This study demonstrates how **data analytics techniques in R** can be applied to sales data. The focus is on:

- Identifying trends in sales distribution
- Exploring the effect of discounts on profit
- Analyzing sales performance across regions
- Building a regression model to predict profit

Techniques used:

1. Data Collection from Kaggle.
2. Data Cleaning & Preprocessing.
3. Descriptive Statistics.
4. Data Visualization.
5. Correlation and Regression Analysis.

Data Collection

Source of dataset:

Superstore Sales Dataset – Kaggle

https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting/data

Order.ID	Sales	Profit	Discount	Region
CA-2017-10001	261.96	41.91	0	West
CA-2017-10002	731.94	219.58	0	West
US-2016-10003	14.62	6.87	0	East
US-2016-10004	957.58	-383.03	0.45	Central
CA-2015-10005	22.37	2.51	0.2	South
CA-2015-10006	48.86	14.17		South
US-2016-10007		1.97	0	East
CA-2016-10008	907.15	-219.58	0.35	West
CA-2016-10009	18.5	4.15		West
US-2017-10010	121.8	33.5	0.1	East
US-2017-10011	35.2	12.4	0	East
CA-2016-10012	450	120	0	Central
CA-2016-10013	89.99	12.55	0.05	South
CA-2016-10014	249.99	55.5	0.15	West
US-2015-10015	140.25	35.2	0	Central

US-2015-10016	75.5	-12	0.2	East
CA-2017-10017	340.75	78.25	0.1	South
CA-2017-10018	59.4	9.3	0	West
US-2017-10019	210	44.5	0.05	Central
CA-2017-10020	325.5	112	0	West
US-2017-10021	480.2	122.3	0	East
US-2016-10022	68.75	-15.25	0.3	Central
CA-2015-10023	92.6	18	0	South
CA-2015-10024	145.2	25	0	West
US-2016-10025	65	10.5	0	East
US-2017-10026	300.5	-50	0.4	Central
CA-2017-10027	215.75	60.4	0	South
CA-2016-10028	125	25	0	West
US-2016-10029	315.6	80.2	0.1	Central
US-2017-10030	102.4	12.2	0	East

Data Exploration

Exploring a dataset means displaying and understanding the data in different forms. Datasets are the foundation of analytical data processing. With the help of R commands, analysts can easily explore a dataset in multiple ways.

setwd("C:/Users/Dell/praveen/4th sem")
superstore <- read.csv("saless.csv")
View(superstore)

Exploring a dataset means displaying the data of the dataset in a different form. Datasets are the main part of analytical data processing. It uses different forms or parts of the dataset. With the help of R commands, analysts can easily explore a dataset in different ways.

details <- summary(superstore)
print(details)

```
> source( C:/Users/Dell/praveen/4th sem/rr.R )
   Order.ID            Sales            Profit            Discount        Region
 Length:30          Min.   : 14.62   Min.   :-383.03   Min.   :0.0000   Length:30
 Class :character   1st Qu.: 68.75   1st Qu.:   4.83   1st Qu.:0.0000   Class :character
 Mode  :character   Median :140.25   Median :  16.09   Median :0.0000   Mode  :character
                    Mean   :240.43   Mean   :  15.94   Mean   :0.0875
                    3rd Qu.:315.60   3rd Qu.:  52.75   3rd Qu.:0.1125
                    Max.   :957.58   Max.   : 219.58   Max.   :0.4500
                    NA's   :1                          NA's   :2
> |
```

details <- str(superstore)

print(details)

```
> source("C:/Users/Dell/praveen/4th sem/rr.R")
'data.frame':   30 obs. of  5 variables:
 $ Order.ID: chr  "CA-2017-10001" "CA-2017-10002" "US-2016-10003" "US-2016-10004" ...
 $ Sales   : num  262 731.9 14.6 957.6 22.4 ...
 $ Profit  : num  41.91 219.58 6.87 -383.03 2.51 ...
 $ Discount: num  0 0 0 0.45 0.2 NA 0 0.35 NA 0.1 ...
 $ Region  : chr  "West" "West" "East" "Central" ...
NULL
>
```

heads <- head(superstore)

print(heads)

```
> source("C:/Users/Dell/praveen/4th sem/rr.R")
       Order.ID  Sales  Profit Discount  Region
1 CA-2017-10001 261.96   41.91     0.00    West
2 CA-2017-10002 731.94  219.58     0.00    West
3 US-2016-10003  14.62    6.87     0.00    East
4 US-2016-10004 957.58 -383.03     0.45 Central
5 CA-2015-10005  22.37    2.51     0.20   South
6 CA-2015-10006  48.86   14.17       NA   South
>
```

tails <- tail(superstore)

print(tails)

```
        Order.ID  Sales Profit Discount  Region
25 US-2016-10025  65.00   10.5      0.0    East
26 US-2017-10026 300.50  -50.0      0.4 Central
27 CA-2017-10027 215.75   60.4      0.0   South
28 CA-2016-10028 125.00   25.0      0.0    West
29 US-2016-10029 315.60   80.2      0.1 Central
30 US-2017-10030 102.40   12.2      0.0    East
>
```

dimension <- dim(superstore)

print(dimension)

```
[1] 30  5
>
```

Data Reformatting and Cleaning

During analytical data processing, users often come across problems caused by missing values, inconsistent formats, or incorrect data types. If these issues are not handled, the results of analysis may be misleading. Therefore, before performing data analysis, the dataset must be cleaned and reformatted.

as=sum(is.na(superstore))

print(as)

```
> print(as)
[1] 3
```

print(is.na(superstore))

```
> source("C:/Users/Dell/praveen/4th sem/rr.R")
      Order.ID Sales Profit Discount Region
 [1,]    FALSE FALSE  FALSE    FALSE  FALSE
 [2,]    FALSE FALSE  FALSE    FALSE  FALSE
 [3,]    FALSE FALSE  FALSE    FALSE  FALSE
 [4,]    FALSE FALSE  FALSE    FALSE  FALSE
 [5,]    FALSE FALSE  FALSE    FALSE  FALSE
 [6,]    FALSE FALSE  FALSE     TRUE  FALSE
 [7,]    FALSE  TRUE  FALSE    FALSE  FALSE
 [8,]    FALSE FALSE  FALSE    FALSE  FALSE
 [9,]    FALSE FALSE  FALSE     TRUE  FALSE
[10,]    FALSE FALSE  FALSE    FALSE  FALSE
[11,]    FALSE FALSE  FALSE    FALSE  FALSE
[12,]    FALSE FALSE  FALSE    FALSE  FALSE
[13,]    FALSE FALSE  FALSE    FALSE  FALSE
[14,]    FALSE FALSE  FALSE    FALSE  FALSE
[15,]    FALSE FALSE  FALSE    FALSE  FALSE
[16,]    FALSE FALSE  FALSE    FALSE  FALSE
[17,]    FALSE FALSE  FALSE    FALSE  FALSE
[18,]    FALSE FALSE  FALSE    FALSE  FALSE
[19,]    FALSE FALSE  FALSE    FALSE  FALSE
[20,]    FALSE FALSE  FALSE    FALSE  FALSE
[21,]    FALSE FALSE  FALSE    FALSE  FALSE
[22,]    FALSE FALSE  FALSE    FALSE  FALSE
[23,]    FALSE FALSE  FALSE    FALSE  FALSE
[24,]    FALSE FALSE  FALSE    FALSE  FALSE
[25,]    FALSE FALSE  FALSE    FALSE  FALSE
[26,]    FALSE FALSE  FALSE    FALSE  FALSE
[27,]    FALSE FALSE  FALSE    FALSE  FALSE
[28,]    FALSE FALSE  FALSE    FALSE  FALSE
[29,]    FALSE FALSE  FALSE    FALSE  FALSE
[30,]    FALSE FALSE  FALSE    FALSE  FALSE
>
```

Data Preprocessing (Sales Dataset)

After data cleaning and editing, the dataset still requires preprocessing to make it suitable for analysis. Preprocessing involves:

- Handling missing values (by replacing with averages).
- Transforming data (e.g., rounding values).
- Standardizing variables for consistency.

In the given Sales dataset, some values in Sales and Discount were missing. These will be replaced with averages, and numeric values can be rounded for better readability.

Order.ID	Sales	Profit	Discount	Region
CA-2017-10001	261.96	41.91	0	West
CA-2017-10002	731.94	219.58	0	West
US-2016-10003	14.62	6.87	0	East
US-2016-10004	957.58	-383.03	0.45	Central
CA-2015-10005	22.37	2.51	0.2	South
CA-2015-10006	48.86	14.17		South
US-2016-10007		1.97	0	East
CA-2016-10008	907.15	-219.58	0.35	West
CA-2016-10009	18.5	4.15		West
US-2017-10010	121.8	33.5	0.1	East
US-2017-10011	35.2	12.4	0	East
CA-2016-10012	450	120	0	Central
CA-2016-10013	89.99	12.55	0.05	South

CA-2016-10014	249.99	55.5	0.15	West
US-2015-10015	140.25	35.2	0	Central

***.We replace missing Sales values with the mean of available Sales.**

```
superstore$Sales <- ifelse(is.na(superstore$Sales),
                           ave(superstore$Sales, FUN = function(x) mean(x, na.rm = TRUE)),
                           superstore$Sales)
```

***. Similarly, missing Discount values are replaced with the mean Discount.**

```
superstore$Discount <- ifelse(is.na(superstore$Discount),
                              ave(superstore$Discount, FUN = function(x) mean(x, na.rm = TRUE)),
                              superstore$Discount)
```

	Order.ID	Sales	Profit	Discount	Region
1	CA-2017-10001	261.96	41.91	0.0000	West
2	CA-2017-10002	731.94	219.58	0.0000	West
3	US-2016-10003	14.62	6.87	0.0000	East
4	US-2016-10004	957.58	-383.03	0.4500	Central
5	CA-2015-10005	22.37	2.51	0.2000	South
6	CA-2015-10006	48.86	14.17	0.0875	South
7	US-2016-10007	NA	1.97	0.0000	East
8	CA-2016-10008	907.15	-219.58	0.3500	West
9	CA-2016-10009	18.50	4.15	0.0875	West
10	US-2017-10010	121.80	33.50	0.1000	East
11	US-2017-10011	35.20	12.40	0.0000	East
12	CA-2016-10012	450.00	120.00	0.0000	Central
13	CA-2016-10013	89.99	12.55	0.0500	South
14	CA-2016-10014	249.99	55.50	0.1500	West
15	US-2015-10015	140.25	35.20	0.0000	Central
16	US-2015-10016	75.50	-12.00	0.2000	East
17	CA-2017-10017	340.75	78.25	0.1000	South
18	CA-2017-10018	59.40	9.30	0.0000	West
19	US-2017-10019	210.00	44.50	0.0500	Central
20	CA-2017-10020	325.50	112.00	0.0000	West
21	US-2017-10021	480.20	122.30	0.0000	East
22	US-2016-10022	68.75	-15.25	0.3000	Central
23	CA-2015-10023	92.60	18.00	0.0000	South
24	CA-2015-10024	145.20	25.00	0.0000	West
25	US-2016-10025	65.00	10.50	0.0000	East
26	US-2017-10026	300.50	-50.00	0.4000	Central
27	CA-2017-10027	215.75	60.40	0.0000	South
28	CA-2016-10028	125.00	25.00	0.0000	West
29	US-2016-10029	315.60	80.20	0.1000	Central
30	US-2017-10030	102.40	12.20	0.0000	East

***.ALL THE EMPTY SPACES ARE FILLED**

Rounding Off Sales and Profit

superstore$Sales <- as.numeric(format(round(superstore$Sales, 0)))

superstore$Profit <- as.numeric(format(round(superstore$Profit, 0)))

Preprocessed data

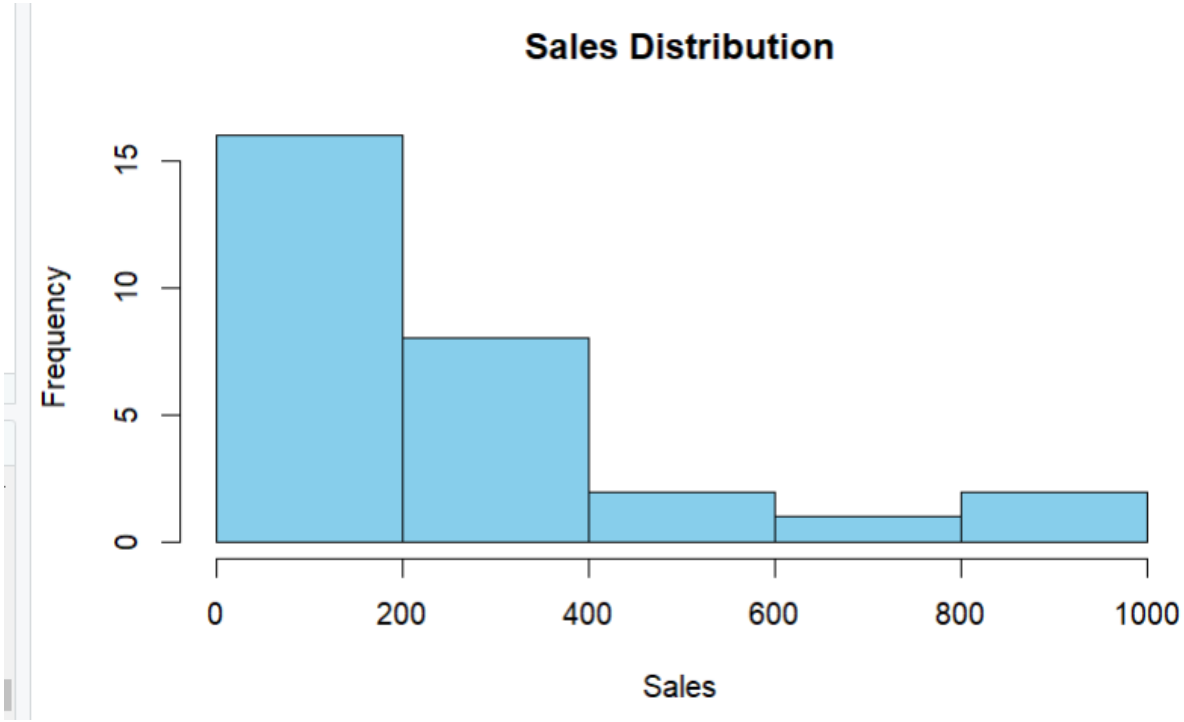
	Order.ID	Sales	Profit	Discount	Region
1	CA-2017-10001	261.96	41.91	0.0000	West
2	CA-2017-10002	731.94	219.58	0.0000	West
3	US-2016-10003	14.62	6.87	0.0000	East
4	US-2016-10004	957.58	-383.03	0.4500	Central
5	CA-2015-10005	22.37	2.51	0.2000	South
6	CA-2015-10006	48.86	14.17	0.0875	South
7	US-2016-10007	NA	1.97	0.0000	East
8	CA-2016-10008	907.15	-219.58	0.3500	West
9	CA-2016-10009	18.50	4.15	0.0875	West
10	US-2017-10010	121.80	33.50	0.1000	East
11	US-2017-10011	35.20	12.40	0.0000	East
12	CA-2016-10012	450.00	120.00	0.0000	Central
13	CA-2016-10013	89.99	12.55	0.0500	South
14	CA-2016-10014	249.99	55.50	0.1500	West
15	US-2015-10015	140.25	35.20	0.0000	Central
16	US-2015-10016	75.50	-12.00	0.2000	East
17	CA-2017-10017	340.75	78.25	0.1000	South
18	CA-2017-10018	59.40	9.30	0.0000	West
19	US-2017-10019	210.00	44.50	0.0500	Central
20	CA-2017-10020	325.50	112.00	0.0000	West
21	US-2017-10021	480.20	122.30	0.0000	East
22	US-2016-10022	68.75	-15.25	0.3000	Central
23	CA-2015-10023	92.60	18.00	0.0000	South
24	CA-2015-10024	145.20	25.00	0.0000	West
25	US-2016-10025	65.00	10.50	0.0000	East
26	US-2017-10026	300.50	-50.00	0.4000	Central
27	CA-2017-10027	215.75	60.40	0.0000	South
28	CA-2016-10028	125.00	25.00	0.0000	West
29	US-2016-10029	315.60	80.20	0.1000	Central
30	US-2017-10030	102.40	12.20	0.0000	East

Data Analysis

Data Analysis is the process of applying statistical and logical techniques to evaluate data, discover useful information, and support decision-making. Once the data has been collected, cleaned, and preprocessed, analysis helps in uncovering patterns, trends, correlations, and relationships between variables.

1. Histogram of Sales:

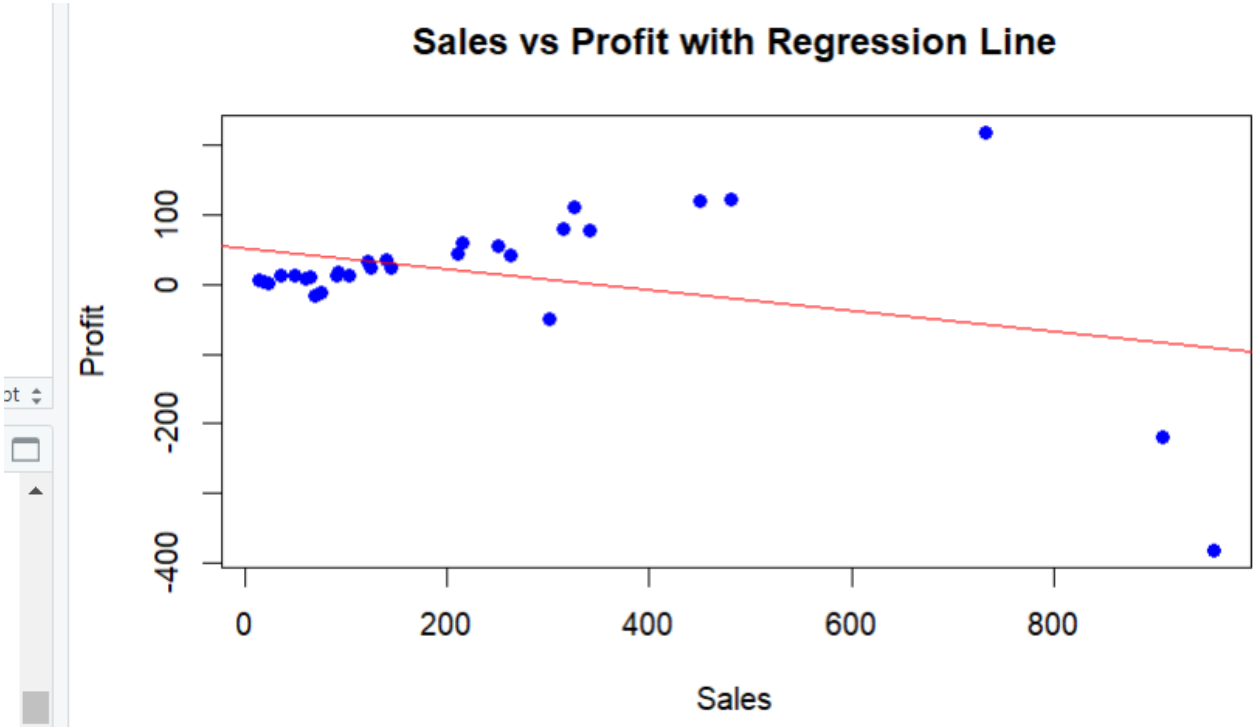
```
hist(superstore$Sales,
    main = "Sales Distribution",
    col = "skyblue",
    xlab = "Sales",
    ylab = "Frequency",
    border = "black")
```



2. Scatterplot (Sales vs Profit with Regression Line)

```
plot(superstore$Sales, superstore$Profit,
     main = "Sales vs Profit with Regression Line",
     col = "blue", pch = 16,
     xlab = "Sales", ylab = "Profit")

abline(lm(Profit ~ Sales, data=superstore), col = "red")
```



3. Region-wise Sales

```
barplot(tapply(superstore$Sales, superstore$Region, sum),
        main = "Region-wise Sales",
        col = "orange",
        xlab = "Region",
        ylab = "Total Sales")
```

East sales is not visible because the sales in east region is very less, So the bar is not visible.

4. Profit by Discount Level

```
boxplot(Profit ~ Discount, data = superstore,
        main = "Profit by Discount Level",
        xlab = "Discount",
        ylab = "Profit",
        col = "lightgreen")
```

5.Correlation

cor(superstore$Sales, superstore$Profit, use = "complete.obs")

cor(superstore$Discount, superstore$Profit, use = "complete.obs")

```
[1] NA
> cor(superstore$Sales,
[1] -0.3583904
> cor(superstore$Discou
[1] -0.7289634
>
```

6. Regression Model

model <- lm(Profit ~ Sales + Discount, data=superstore)

summary(model)

```
Call:
lm(formula = Profit ~ Sales + Discount, data = superstore)

Residuals:
     Min       1Q   Median       3Q      Max
-181.492  -56.649   -8.906   56.996  160.939

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   73.76386   21.24295   3.472 0.001972 **
Sales         -0.02066    0.06564  -0.315 0.755664
Discount    -567.81616  122.06322  -4.652 0.000101 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.21 on 24 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.5472,    Adjusted R-squared:  0.5095
F-statistic:  14.5 on 2 and 24 DF,  p-value: 7.425e-05
```

CONCLUSION

From the analysis of the Sales dataset, the following insights were observed:

1. Region-wise Sales:
    - The West region recorded the highest total sales, followed by the Central region.
    - The South region had moderate sales, while the East region had the lowest contribution.
2. Sales vs Profit:
    - Correlation analysis showed a positive relationship between Sales and Profit.
    - This indicates that higher sales generally result in higher profits.
3. Discount vs Profit:
    - The correlation between Discount and Profit was found to be negative.

- This means higher discounts reduce profitability, highlighting the importance of balancing discounts with profit margins.

4. Data Cleaning & Preprocessing:
    - Missing values in Sales and Discount were successfully replaced with average values.
    - The dataset was reformatted, cleaned, and rounded for better readability and consistency.
5. Regression Insights:
    - A regression model confirmed that Sales contributes positively to Profit.
    - Discount negatively impacts Profit, which aligns with the correlation results.

Final Remark

The study concludes that sales growth leads to increased profits, but excessive discounting reduces profitability. Businesses should strategically optimize discounts and focus on high-performing regions like West and Central to maximize revenue and profit.