Walmart Store Sales Data Project

In R Programming Language

By
Ayush Thapa

Objective: To analyse Walmart store sales data to practise R programming skills, including data manipulation, statistical analysis, and data visualization.

Dataset: File - `Walmart.csv`

1: Setting Up the Environment

Installing the packages:

```
install.packages("tidyverse")
install.packages("ggplot2")
install.packages("summarytools")
```

Loading the packages:

```
library(tidyverse)
library(ggplot2)
library(summarytools)
```

Reading the dataset by loading the CSV file from the local drive path using `read.csv()`:

```
walmart.csv <-
read.csv("/Users/akheil/Downloads/Just_IT -
Data Bootcamp/R/Walmart.csv")
```

Previewing the first few rows using 'head()':

```
head(walmart.csv)
```

Console | Terminal | Background Jobs

R 4.4.1 · ~/Downloads/Just_It – Data Bootcamp/R/

```
> walmart.csv <- read.csv("/Users/akheil/Downloads/Just_It - Data Bootcamp/R/Walmart.csv")
> head(walmart.csv)
  Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      CPI Unemployment
1     1 05-02-2010      1643691            0       42.31      2.572 211.0964        8.106
2     1 12-02-2010      1641957            1       38.51      2.548 211.2422        8.106
3     1 19-02-2010      1611968            0       39.93      2.514 211.2891        8.106
4     1 26-02-2010      1409728            0       46.63      2.561 211.3196        8.106
5     1 05-03-2010      1554807            0       46.50      2.625 211.3501        8.106
6     1 12-03-2010      1439542            0       57.79      2.667 211.3806        8.106
>
```

This function shows the first six rows from the ‘walmart.csv’ data frame.

2: Data Exploration

Summarizing the dataset:

```
dfSummary(walmart.csv)
```

Console | Terminal | Background Jobs

R 4.4.1 · ~/Downloads/Just_It – Data Bootcamp/R/

```
> dfSummary(walmart.csv)
Data Frame Summary
walmart.csv
Dimensions: 6435 x 8
Duplicates: 0

----------------------------------------------------------------------------------------------------------------
--------------
No   Variable        Stats / Values                    Freqs (% of Valid)     Graph                   Valid
Missing
---- -------------- ------------------------------- --------------------- ---------------------- ------
---- ---------
1    Store           Mean (sd) : 23 (13)               45 distinct values     : : : : : : : : :      6435
0
     [integer]       min < med < max:                                         : : : : : : : : :      (100.
0%)   (0.0%)
                     1 < 23 < 45                                              : : : : : : : : :
                     IQR (CV) : 22 (0.6)                                      : : : : : : : : :
                                                                              : : : : : : : : :

2    Date            1. 01-04-2011                       45 ( 0.7%)                                  6435
0
     [character]     2. 01-06-2012                       45 ( 0.7%)                                  (100.
0%)   (0.0%)
                     3. 01-07-2011                       45 ( 0.7%)
                     4. 01-10-2010                       45 ( 0.7%)
                     5. 02-03-2012                       45 ( 0.7%)
                     6. 02-04-2010                       45 ( 0.7%)
                     7. 02-07-2010                       45 ( 0.7%)
                     8. 02-09-2011                       45 ( 0.7%)
                     9. 02-12-2011                       45 ( 0.7%)
                     10. 03-02-2012                      45 ( 0.7%)
                     [ 133 others ]                    5985 (93.0%)          IIIIIIIIIIIIIIIIIII
```

Checking for missing values:

```
summary(is.na(walmart.csv))
```

Console | Terminal | Background Jobs

R 4.4.1 · ~/Downloads/Just_It – Data Bootcamp/R/

```
> summary(is.na(walmart.csv))
   Store            Date          Weekly_Sales     Holiday_Flag     Temperature      Fuel_Price
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:6435      FALSE:6435      FALSE:6435      FALSE:6435      FALSE:6435      FALSE:6435
    CPI          Unemployment
 Mode :logical   Mode :logical
 FALSE:6435      FALSE:6435
>
```

Console Terminal Ba

R 4.4.1 · ~/Download

```
> dim(walmart.csv)
[1] 6435    8
> nrow(walmart.csv)
[1] 6435
> ncol(walmart.csv)
[1] 8
>
```

There are no missing values in the ‘walmart.csv’ data frame as all the 6,435 entries have some values.

3: Statistical Analysis

<u>Descriptive Statistics</u>:

- `mean(walmart.csv$Weekly_Sales, na.rm = TRUE)`
- `median(walmart.csv$Weekly_Sales, na.rm = TRUE)`
- `sd(walmart.csv$Weekly_Sales, na.rm = TRUE)`

These three separate lines of codes calculate **Mean**, **Median** and **Standard deviation** individually.

Or,

```
descr(walmart.csv)
```

This one code calculates **all** descriptive statistics in one go.

Console Terminal Background Jobs

R 4.4.1 · ~/Downloads/Just_It – Data Bootcamp/R/

```
> mean(walmart.csv$Weekly_Sales, na.rm = TRUE)
[1] 1046965
> median(walmart.csv$Weekly_Sales, na.rm = TRUE)
[1] 960746
> sd(walmart.csv$Weekly_Sales, na.rm = TRUE)
[1] 564366.6
>
```

Console | Terminal | Background Jobs

R 4.4.1 · ~/Downloads/Just_It – Data Bootcamp/R/

```
> descr(walmart.csv)
Non-numerical variable(s) ignored: Date
Descriptive Statistics
walmart.csv
N: 6435
```

	CPI	Fuel_Price	Holiday_Flag	Store	Temperature	Unemployment
Mean	171.58	3.36	0.07	23.00	60.66	8.00
Std.Dev	39.36	0.46	0.26	12.99	18.44	1.88
Min	126.06	2.47	0.00	1.00	-2.06	3.88
Q1	131.74	2.93	0.00	12.00	47.43	6.89
Median	182.62	3.44	0.00	23.00	62.67	7.87
Q3	212.75	3.73	0.00	34.00	74.95	8.62
Max	227.23	4.47	1.00	45.00	100.14	14.31
MAD	60.88	0.56	0.00	16.31	20.30	1.29
IQR	81.01	0.80	0.00	22.00	27.48	1.73
CV	0.23	0.14	3.65	0.56	0.30	0.23
Skewness	0.06	-0.10	3.37	0.00	-0.34	1.19
SE.Skewness	0.03	0.03	0.03	0.03	0.03	0.03
Kurtosis	-1.84	-1.18	9.37	-1.20	-0.61	2.63
N.Valid	6435.00	6435.00	6435.00	6435.00	6435.00	6435.00
Pct.Valid	100.00	100.00	100.00	100.00	100.00	100.00

Table: Table continues below

	Weekly_Sales
Mean	1046964.88
Std.Dev	564366.62
Min	209986.25

	Weekly_Sales
Mean	1046964.88
Std.Dev	564366.62
Min	209986.25
Q1	552985.34
Median	960746.04
Q3	1420405.41
Max	3818686.45
MAD	631596.11
IQR	866808.55
CV	0.54
Skewness	0.67
SE.Skewness	0.03
Kurtosis	0.05
N.Valid	6435.00
Pct.Valid	100.00

Key

```
>
```

Correlation Analysis

- Creating a correlation matrix between key metrics: a) Weekly_Sales, b) Temperature, and c) Fuel_Price using the following piece of code:

```
cor(walmart.csv[, c("Weekly_Sales",
"Temperature", "Fuel_Price")], use =
"complete.obs")
```

Console Terminal Background Jobs

R 4.4.1 · ~/Downloads/Just_It – Data Bootcamp/R/

```
> #   Correlation Analysis:
>
> cor(walmart.csv[, c("Weekly_Sales", "Temperature", "Fuel_Price")], use = "complete.obs")
             Weekly_Sales Temperature  Fuel_Price
Weekly_Sales  1.000000000 -0.06381001 0.009463786
Temperature  -0.063810013  1.00000000 0.144981806
Fuel_Price    0.009463786  0.14498181 1.000000000
>
```

<u>Correlation between</u>:

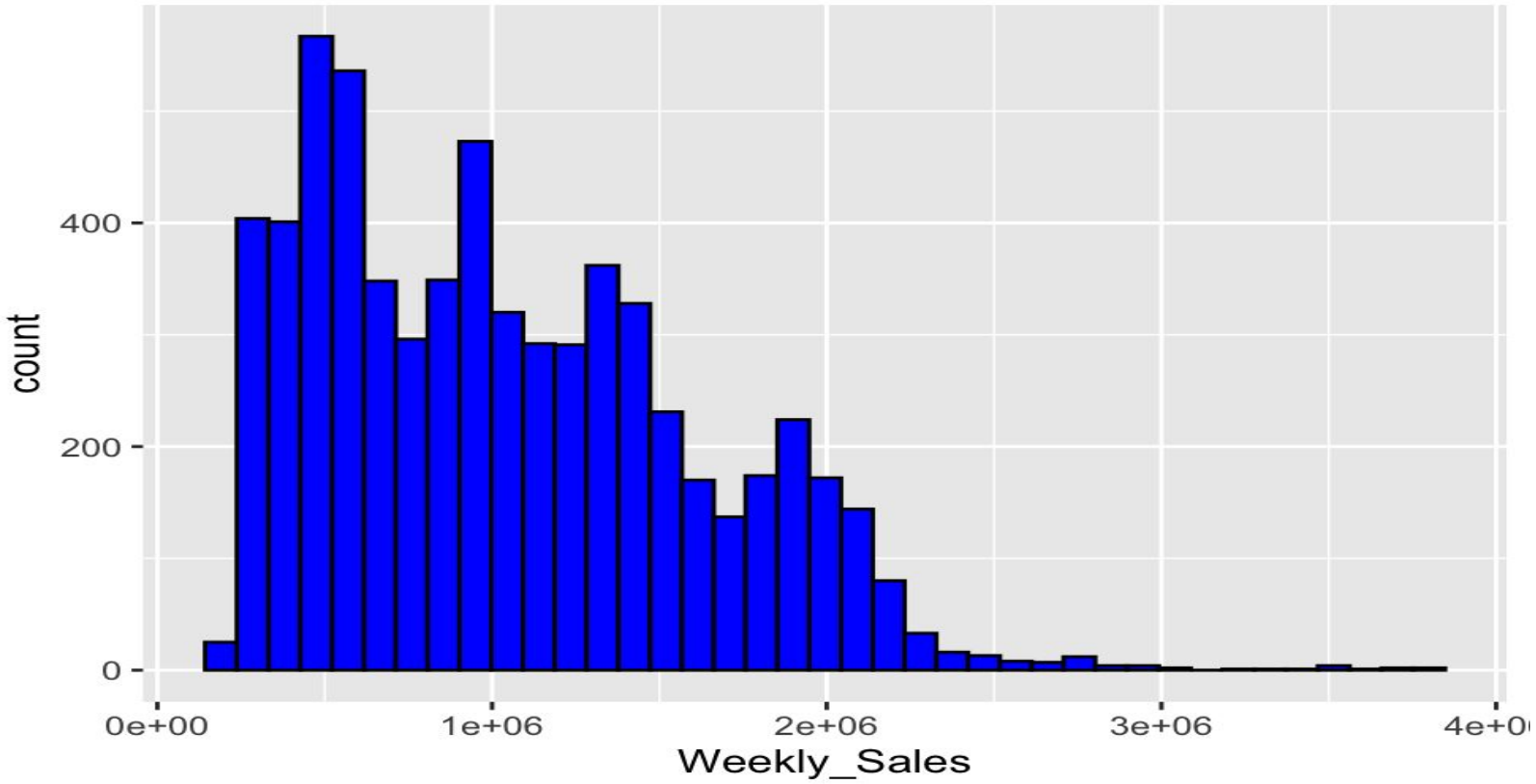
- Weekly_Sales and Temperature ≈ -0.06
- Temperature and Fuel_Price ≈ 0.14
- Fuel_Price and Weekly_Sales ≈ 0.01

They don’t have any correlation with each other because their values are almost 0.

4: Data Visualization

1. <u>Histogram for Weekly Sales</u>:

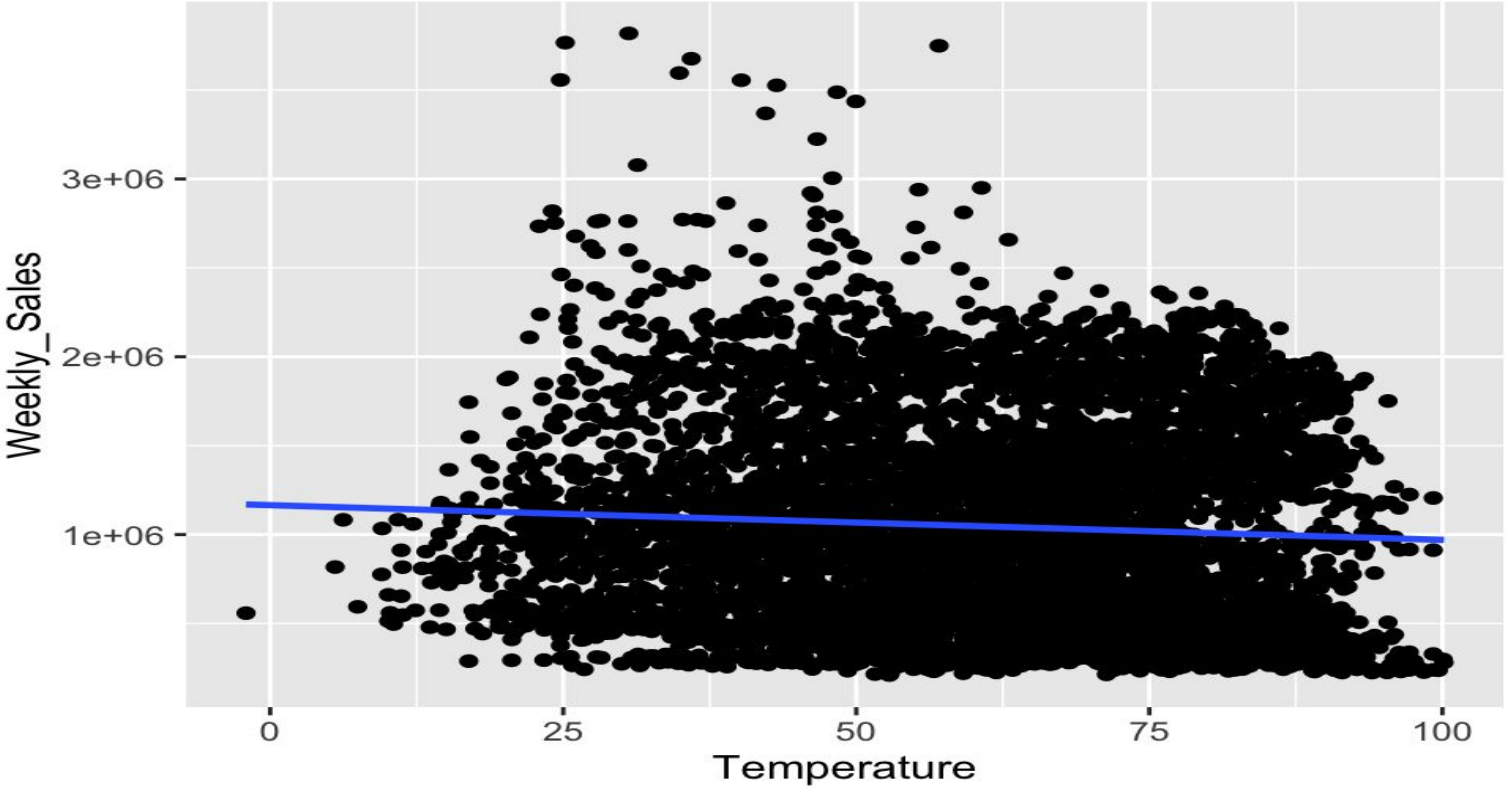
```
ggplot(walmart.csv, aes(x = Weekly_
Sales)) + geom_histogram(binwidth =
95000, fill = "blue", color = "black")
```



➔ This histogram shows that in general there were more number of lower Weekly sales than the higher Weekly sales.

2. Scatter Plot for 'Temperature' Vs. 'Weekly Sales':

```
ggplot(walmart.csv, aes(x =
Temperature, y = Weekly_Sales))+
geom_point() + geom_smooth(method =
"lm", se = FALSE)
```

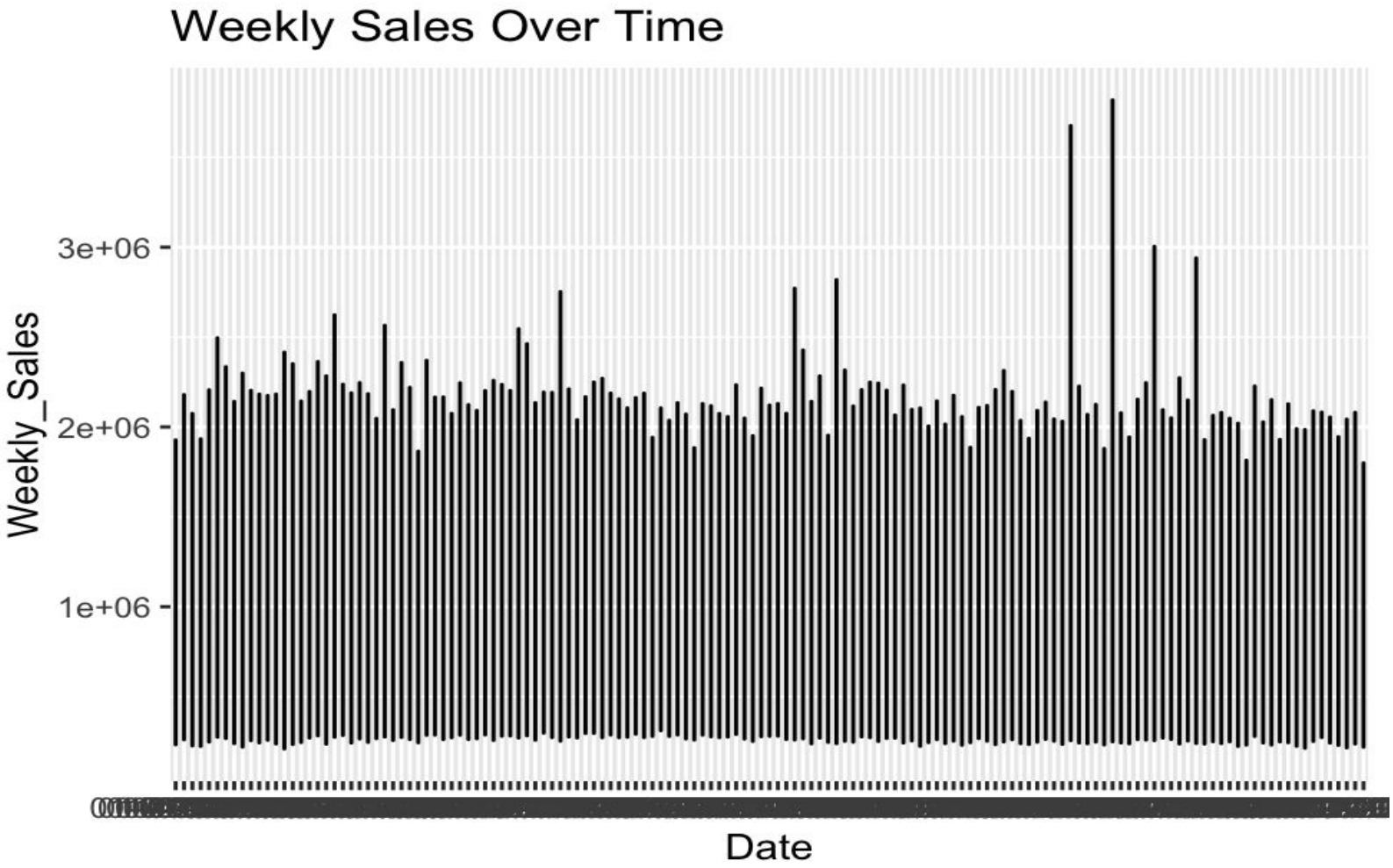


➔ There is no apparent relationship between Weekly_Sales and Temperature here.

3. (**Optional**) Time Series Plot:

Creating a time series plot for `Weekly_Sales` over time.

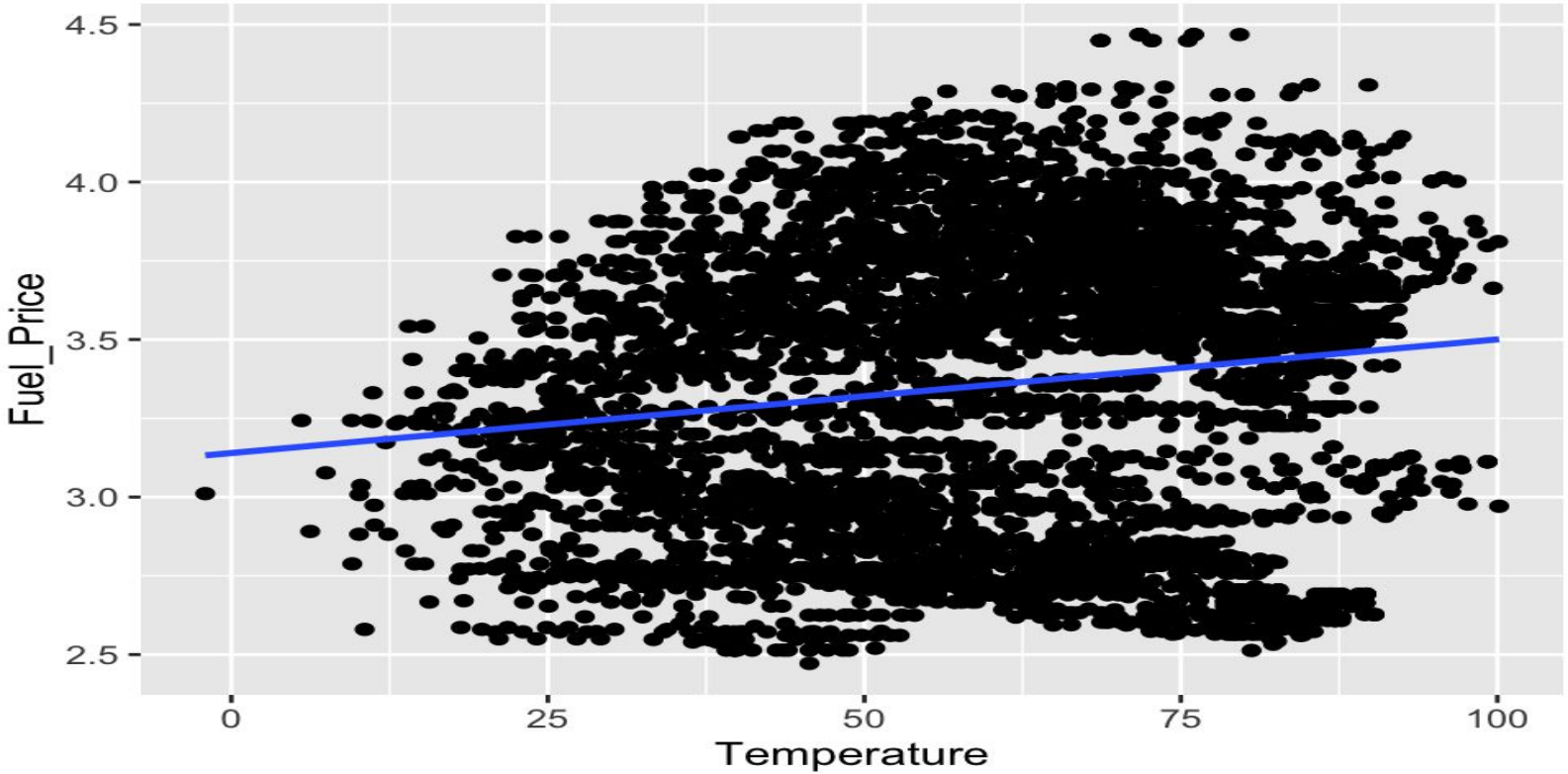
```
ggplot(walmart.csv, aes(x = Date, y =
Weekly_Sales)) + geom_line() +
labs(title = "Weekly Sales Over Time")
```


Weekly Sales Over Time
3e+06
2e+06
1e+06
Weekly_Sales
Date

Extra

4. Scatter Plot for 'Temperature' Vs. 'Fuel Price':

```
ggplot(walmart.csv, aes(x =
Temperature, y = Fuel_Price)) + geom
_point() + geom_smooth(method = "lm",
se = FALSE)
```



➔ No relationship between Temperature and Fuel_Price: data points are scattered everywhere.

- **Creating a separate sub-table called 'Store _ Weekly_ Sales' that shows the Average of Weekly Sales from each Store (ie. Store no. from 1 to 45):**

Here, I've assumed that the column 'Store' shows different stores that are assigned as a store number.

```
Store_Weekly_Sales <- walmart.csv %>%
group_by(Store) %>% summarize(Avg_
Weekly_Sales = mean(Weekly_Sales))
```

'R' Project.R | Store_Weekly_Sales

Filter

	Store	Avg_Weekly_Sales
1	1	1555264.4
2	2	1925751.3
3	3	402704.4
4	4	2094713.0
5	5	318011.8
6	6	1564728.2
7	7	570617.3
8	8	908749.5
9	9	543980.6
10	10	1899424.6
11	11	1356383.1
12	12	1009001.6
13	13	2003620.3
14	14	2020978.4
15	15	623312.5
16	16	519247.7

Showing 1 to 16 of 45 entries, 2 total columns

5. Line plot displaying the Average of Weekly Sales in each Store (ie. Store no. from 1 to 45):

This graph will be plotted from the previous table (previous slide).

```
ggplot(data = Store_Weekly_Sales, aes(x =
Store, y = Avg_Weekly_Sales, group = 1)) +
geom_line() + geom _point() + labs(title =
"Average Weekly_Sales by Store", x = "Store
(Store no. from 1 to 45)", y = "Average
Weekly Sales") + theme minimal()
```

➔ This chart shows Store no. 20 having one of the highest Average weekly sales. On the other hand, Store no. 5 seems it's not doing much good.

To Conclude

This project was a good practice for extracting, manipulating, analysing and visualising data in R programming language.

Having that said, there were some columns such as Holiday_Flag, CPI etc. which I didn't understand quite well. The Weekly_Sales column was also a bit confusing because it was not clearly stated whether it denotes the amount of sales made by the store on weekly basis or the number of sales in the store on weekly basis.

But with help of provided set of instructions (which by the way was very clear and on point), I was able to complete the project successfully. Thank you!