FINAL PROJECT REPORT

DATA ANALYSIS ASSIGNMENT

By : **HARSHIT GAUR**
MASTER OF PROFESSIONAL STUDIES IN ANALYTICS
ALY 6000 : INTRODUCTION TO ANALYTICS
OCTOBER 30, 2021

To : **PROF. RICHARD HE**

ABSTRACT

Data analytics is a discipline focused on extracting insights from data. It comprises the processes, tools and techniques of data analysis and management, including the collection, organization, and storage of data. The chief aim of data analytics is to apply statistical analysis and technologies on data to find trends and solve problems.

Sales analytics is the practice of generating insights from sales data, trends, and metrics to set targets and forecast future sales performance. The best practice for sales analytics is to closely tie all activities to determine revenue outcomes and set objectives for your sales team.

Analysis should focus on improvement and developing a strategy for improving your sales performance in both the short- and long-term.

It provides insights about the top performing and underperforming products/services, the problems in selling and market opportunities, sales forecasting, and sales activities that generate revenue.

INTRODUCTION

It is strongly encouraged to find and choose a data set in an area where one is more interested and is personally motivated to explore about.

Initially, I decided to go with a dataset that will help me hone the skills that I have learned so far in this course and go even beyond it by acquiring more knowledge and broadening my prowess of the analytical skills and R programming language. Then, I came across a dataset which not only interested me very much but also gave me various ideas to implement using that data set.

I have a sweet tooth and the dataset which I chose is about a bakery and its sales of various items throughout the years 2006 to 2019. The dataset contains various numerical and categorical data. It has 5,114 data points with 9 features related to dates, days, promotion types, quantities of cakes, quantities of pies, cookies, and more. I decided to put my sweet tooth to better use and up-skill myself in the analytical, visualization, and programmatical aspects of the domain.

The features available in the data set are -

S. No.	Feature	Dictionary
1.	ID	Identification number of the records.
2.	Date	Date of the data point of which the sales of the items were recorded.
3.	weekday	Day of the week
4.	cakes	Quantity of cakes sold on a particular date
5.	pies	Quantity of pies sold on a particular date
6.	cookies	Quantity of cookies sold on a particular date
7.	smoothies	Quantity of smoothies sold on a particular date
8.	coffee	Quantity of coffee sold on a particular date
9.	promotion	Type of Promotion (if promotion was applicable for that particular date or not)

Table 1: Features of the data set with their dictionary.

From the structure of the data set, the features, their types, and values can be determined.

```
> str(dataSet)
'data.frame':   5117 obs. of  9 variables:
 $ X         : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Date      : chr  "01/01/06" "02/01/06" "03/01/06" "04/01/06" ...
 $ weekday   : chr  "Sunday" "Monday" "Tuesday" "Wednesday" ...
 $ cakes     : int  45 48 1 4 4 40 10 20 22 37 ...
 $ pies      : int  41 18 40 10 44 27 21 58 5 43 ...
 $ cookies   : int  50 18 99 15 67 104 10 59 28 76 ...
 $ smoothies: int  19 44 41 58 71 39 26 74 61 72 ...
 $ coffee    : int  73 5 8 95 20 76 5 68 54 67 ...
 $ promotion: chr  "promotion" "none" "none" "none" ...
```

Figure 1: Structure of the data set.

DATA PRE-PROCESSING AND CLEANING

The first blush of the data set portrayed the data to be clean & useful, but after an **Initial Data Analysis (IDA)** process assisted with some graph visualizations, I found out that the data set is clean with no outliers present in it.

But, data cleaning is an important & necessary factor in the data analysis process. As the famous quote says - *"Garbage In, Garbage Out."*

In order to utilize and hone my data cleaning skills, I added some data points randomly to the data set with :

- Missing values
- Duplicate values
- Outliers

After this step, I proceeded with the now needed & necessary step of data cleaning using RStudio application on this data set to wrangle and eliminate all the garbage added manually above.

	X	Date	weekday	cakes	pies	cookies	smoothies	coffee	promotion
5111	5110	29/12/19	Sunday	3	47	84	68	92	promotion
5112	5111	30/12/19	Monday	12	25	17	48	13	none
5113	5112	31/12/19	Tuesday	47	10	70	25	96	promotion
5114	5113	01/01/20	Sunday	151	152	174	148	150	promotion
5115	5114	02/01/20	Sunday	164	147	177	200	163	promotion
5116	5115	03/01/20	Monday	170	156	150	168	161	none
5117	5116	04/01/20	Tuesday	146	159	155	163	154	promotion
5118	5117	05/01/20	Sunday	*NA*	*NA*	221	*NA*	*NA*	promotion
5119	5118	06/01/20	Sunday	*NA*	*NA*	276	*NA*	*NA*	
5120	5119	07/01/20	Monday	*NA*	*NA*	*NA*	*NA*	*NA*	none
5121	5120	08/01/20	Tuesday	*NA*	199	178	189	*NA*	promotion
5122	5121	08/01/20	Tuesday	*NA*	199	178	189	*NA*	promotion

Figure 2: Missing values, NA values, Duplicate values in the data set.

a. The Figure 1 is a snapshot of some of the data points with missing values and NA values. The records of the dates (**Date - 05/01/20, 06/01/20, 07/01/20, etc.**) have *NA or empty values* in the features like cakes, pies, cookies, and more.

   An operation was performed on the data set to omit the records with NA or missing values.

```
> # Removing 'NA, Missing Values' from the data set.
> dataSet <- na.omit(dataSet)
```

Figure 3: Eliminating records with NA, missing values from the data set.

b. It also contains some *duplicate* records as well. The records of the dates (**Date - 08/01/20**) are present twice in the data set which needs to be handled properly.

   Another operation was performed on the data set to eliminate duplicate records.

```
> # Remove duplicated rows based on a feature 'Date' as the
  data is distinct based on dates.
> dataSet <- dataSet %>% distinct(Date, .keep_all = TRUE)
```

Figure 4: Eliminating duplicate records from the data set.

c. Few records of outlier were also present in the data set which have not been removed in order to plot the boxplot and showcase the visualization of these outliers. The records of the dates (**Date - 01/01/20, 02/01/20, etc**.) contain outliers for the features present in the data set respectively.

d. The data set do not contain any kind of unbalanced features.

e. The identification number feature (ID) has not been removed as we're not going to use any of the learning algorithms and models to implement here.

   If we had used any classification of regression models to fit the training data set and predict the testing data set, we would have eliminated the features with :

   i. Unbalanced data, as they would have hampered the fitting and prediction of the model.
   ii. Missing data
   iii. Identification numbers data, as they would be insignificant to the fitting.

EXPLORATORY DATA ANALYSIS

The features of the data set can be summarised to calculate the statistics -

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
X	5117	2558	1477.295	0	1279	3837	5116
cakes	5117	25.13	14.788	1	12	38	170
pies	5117	30.287	17.205	1	16	45	159
cookies	5117	60.069	34.359	1	30	90	177
smoothies	5117	37.861	21.525	1	19	56	200
coffee	5117	49.919	28.601	1	25	74	163
promotion	5117						
... none	2451	47.9%					
... promotion	2666	52.1%					

Figure 5: Summary of the data set.

```
     Date                weekday
Length:5117         Length:5117
Class :character    Class :character
Mode  :character    Mode  :character
```

Figure 6: Summary of the data set. (Date, weekday)

```
 promotion
Length:5117
Class :character
Mode  :character
```

Figure 7: Summary of the data set. (promotion)

The following observations can be made using the statistics found in summary of the data set -

1. The lowest quantities sold by the bakery store of each item in the time span of 2006 to 2019 is **One (1)**.

2. The **mean** of the quantities of *cakes* sold is around **25.13** with a **standard deviation** of **14.79** and **quartiles value (Lower Quartile - 12, Higher Quartile - 38).**
   From the observations, we can anticipate that the data points around **maximum value (170)** can be *outliers* to the feature.

3. The **mean** of the quantities of *pies* sold is around **30.29** with a **standard deviation** of **17.20** and **quartiles value (Lower Quartile - 16, Higher Quartile - 45).**
   From the observations, we can anticipate that the data points around **maximum value (159)** can be *outliers* to the feature.

4. The **mean** of the quantities of *cookies* sold is around **60.1** with a **standard deviation** of **34.36** and **quartiles value (Lower Quartile - 30, Higher Quartile - 90).**
   From the observations, we can anticipate that the data points around **maximum value (177)** can be *outliers* to the feature.

5. The **mean** of the quantities of *smoothies* sold is around **37.86** with a **standard deviation** of **21.52** and **quartiles value (Lower Quartile - 19, Higher Quartile - 56).**
   From the observations, we can anticipate that the data points around **maximum value (200)** can be *outliers* to the feature.

6. The **mean** of the quantities of *coffee* sold is around **49.92** with a **standard deviation** of **28.60** and **quartiles value (Lower Quartile - 25, Higher Quartile - 74).**
   From the observations, we can anticipate that the data points around **maximum value (163)** can be *outliers* to the feature.

7. The promotion feature has 2 values - *none* and *promotion.*
   They are distributed in a balanced way and hold **47.9%** and **52.1%** respectively amongst themselves.

8. The feature *Date* contains the date values formatted as %d/%m/%Y.
   It is an unique feature in the data set with distinct values and duplicate records can be found using this feature.

9. The feature *weekday* contains the days of the week in it.

Some of the records in the data set -

	X	Date	weekday	cakes	pies	cookies	smoothies	coffee	promotion
1	0	01/01/06	Sunday	45	41	50	19	73	promotion
2	1	02/01/06	Monday	48	18	18	44	5	none
3	2	03/01/06	Tuesday	1	40	99	41	8	none
4	3	04/01/06	Wednesday	4	10	15	58	95	none
5	4	05/01/06	Thursday	4	44	67	71	20	promotion
6	5	06/01/06	Friday	40	27	104	39	76	promotion
7	6	07/01/06	Saturday	10	21	10	26	5	none
8	7	08/01/06	Sunday	20	58	59	74	68	promotion
9	8	09/01/06	Monday	22	5	28	61	54	none
10	9	10/01/06	Tuesday	37	43	76	72	67	promotion

Figure 8: Example of records in the data set.

Total Items Sold vs Week Days : 2006 to 2019

The below plot can help in visualising the total number of bakery items sold days wise in the time span of 2006 to 2019.

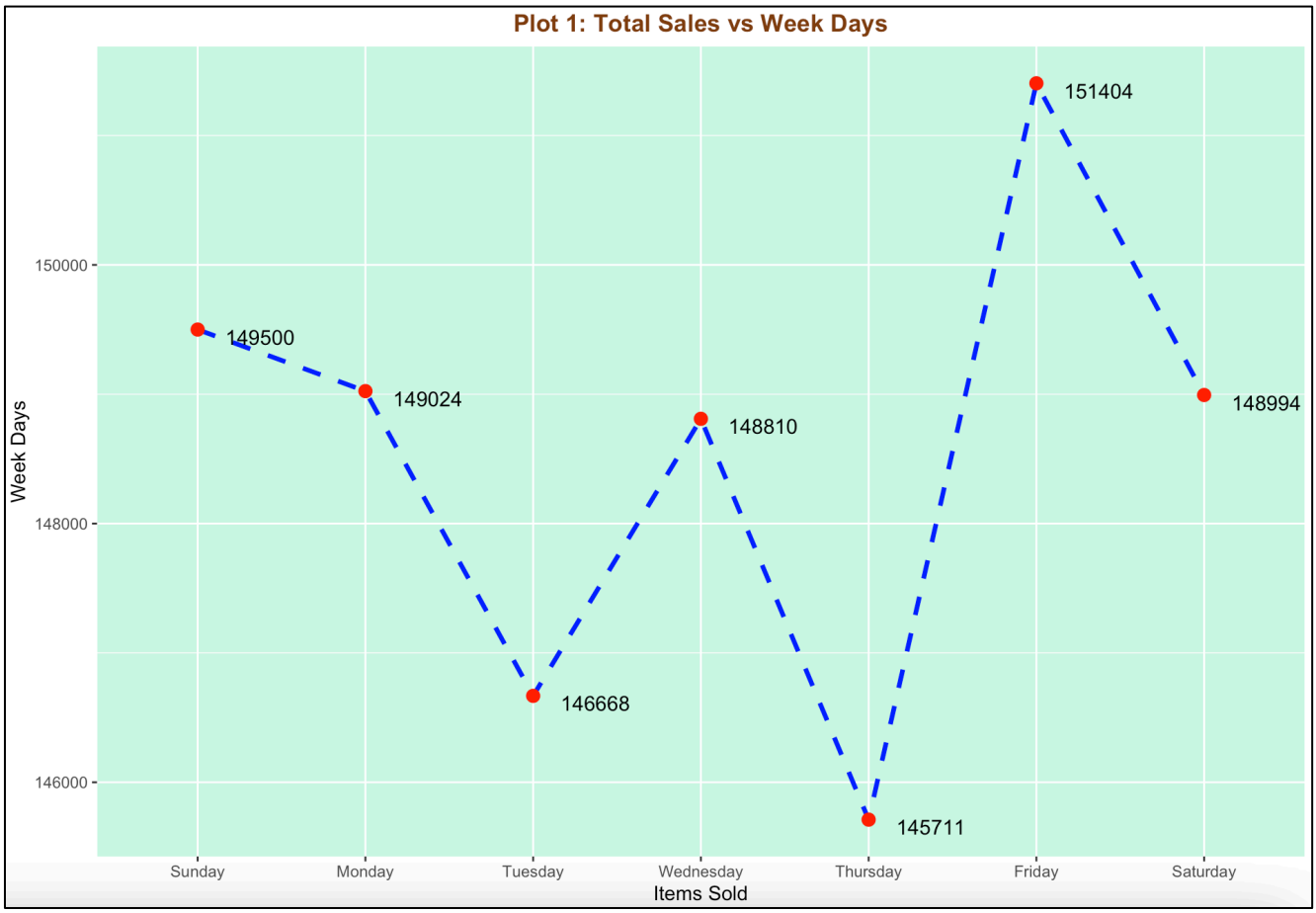

Figure 9: Plot of Total Items Sold vs Week Days

a. We can figure out that the **most quantities of bakery items & goods were sold on Friday** with the value of **1,51,404.**

b. The **least quantities of bakery items & goods have been sold on Thursday** with the value of **1,45,711.**

c. On the rest of the days in a week, the quantities of bakery items & goods sold have been around **1,48,000**.

Bakery Items vs Quantity of Items Sold : 2006 to 2019

The below plot can help in visualising the quantities of all the bakery items, differentiated with their types, sold in the time span of 2006 to 2019.

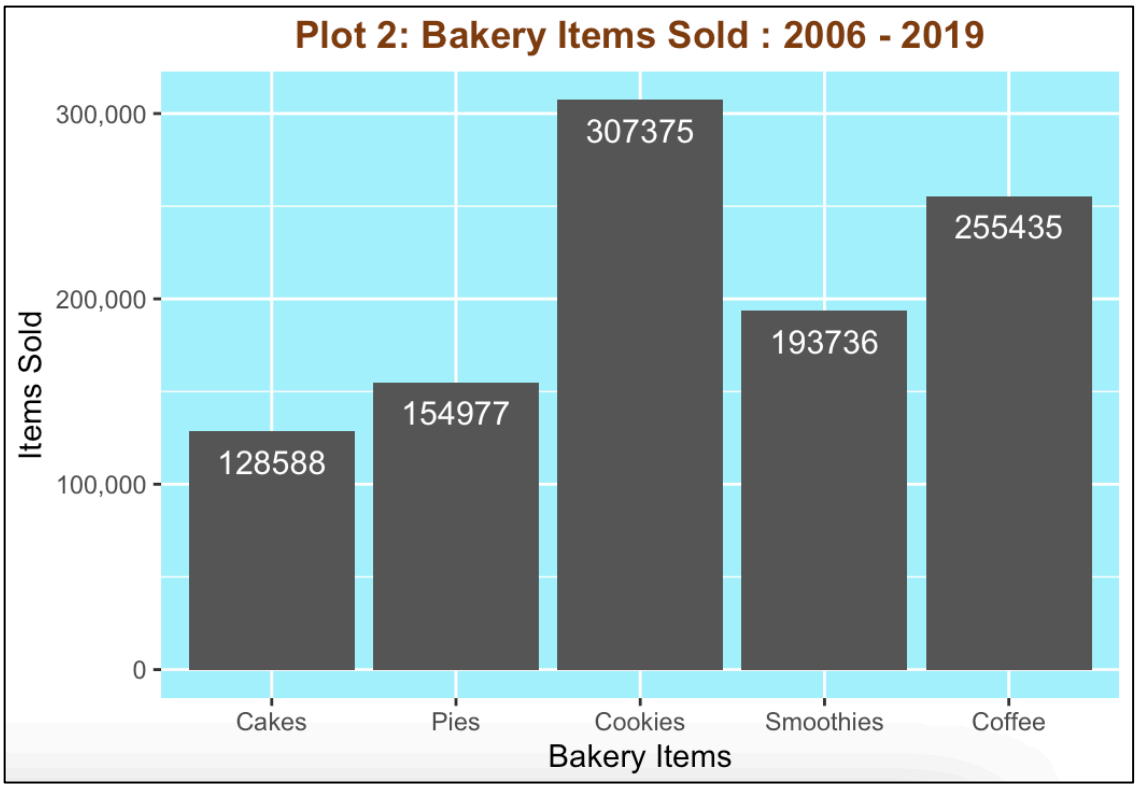

Figure 10: Plot of Bakery Items vs Quantities Sold

a. The **highest number of bakery items sold** by the bakery store in the time span of 2006 to 2019 are **COOKIES** with the value of **3,07,375.**

b. The **least number of bakery items sold** by the bakery store in the time span of 2006 to 2019 are **CAKES** with the value of **1,28,588.**

Outliers :

Some of the feature (Cakes, Pies, Cookies, Smoothies, Coffee) have few data points in them which can be considered as outliers. They have extraordinary high values when compared against the mean and standard deviation values and even the quartiles values.

Multiple Box Plots for Each Bakery Items ~ Promotion Type : 2006 to 2019

The below plot can help in visualising the mean, lower quartile, and higher quartile values of all the bakery items in the time span of 2006 to 2019. The boxplots have been arranged in a single grid for all the features to enable us analyse them properly. They have been differentiated with promotion type as well in order to understand the significance of promoting bakery items.

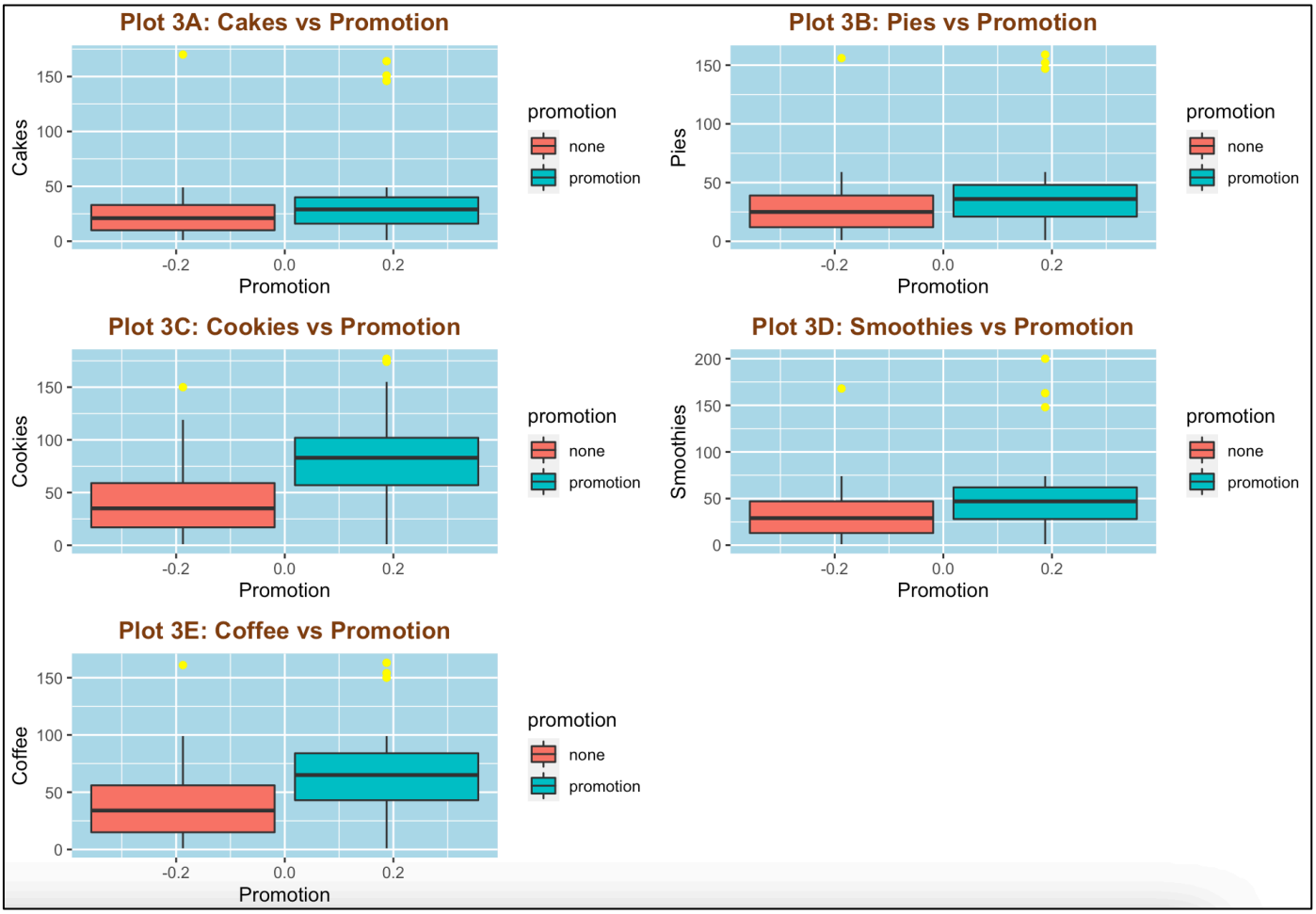

Figure 11: Box Plot of Bakery Items vs Promotion Types

a. There are some **outliers** present in the features (*Cakes, Pies, Cookies, Smoothies, Coffee*) which are marked as **yellow dots.**

b. Since, the box plots are distinguished using **promotion types** (*None, Promotion*) which means either a promotion of items was held on a particular date or not.
We can figure out that the bakery items were sold in higher quantities when they were promoted by the bakery store when compared with the number of quantities sold when not promoted.

Few questions that arises from the data set :

a. Can we perform some analysis on decline or incline of bakery items sold in the years 2006-19?
b. Is there a pattern between the bakery items sold yearly?

Total Bakery Items Sold ~ Years : 2006 to 2019

We have introduced a new feature called '**Total**' to analyse the pattern of all the bakery items sold every year from 2006 to the year 2019. The below plot can help in visualising the pattern of sales of the bakery store in a better way in order to understand the yearly sales properly.

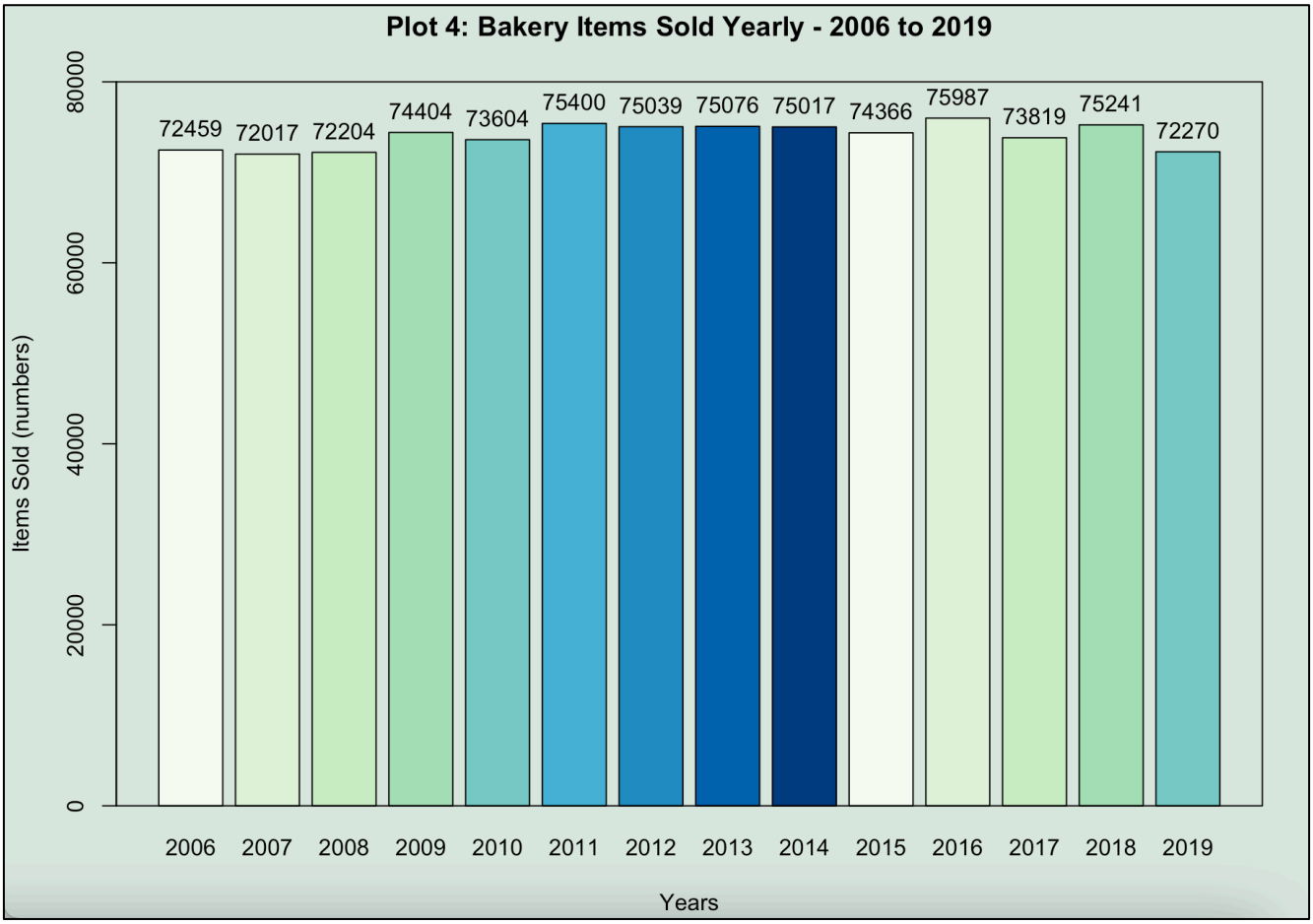

Figure 12: Plot of Total Bakery Items Yearly

a. It can be observed that the yearly sales of the bakery store is in the range of **72,000** to **76,000**.

b. The below figure shows the summary of a new variable introduced containing aggregate summation of some features. Mean, Standard Deviation, Minimum, Maximum, Quartiles values have been calculated and can be referred below in the screenshot :

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
SumCakes	14	9139.786	164.533	8903	9045	9264	9441
SumPies	14	11025.929	414.672	10300	10787	11313.25	11880
SumCookies	14	21908.5	794.792	20835	21524.75	22416.25	23406
SumSmoothies	14	13789.786	469.082	13146	13428.75	14036	14740
SumCoffee	14	18200.5	406.357	17226	17980	18567.75	18630
Total	14	74064.5	1349.468	72017	72745.25	75066.75	75987

Figure 13: Summary of the new variable containing aggregate sums of each feature

Mean Sales of Bakery Items ~ Years : 2006 to 2019

We have introduced a new feature called '**Mean**' to analyse the pattern of decline/incline (decrease/increase) in the sales of all the bakery items sold every year from 2006 to the year 2019. The below plot can help in visualising the pattern of sales of the bakery store in a better way in order to understand the yearly sales properly.

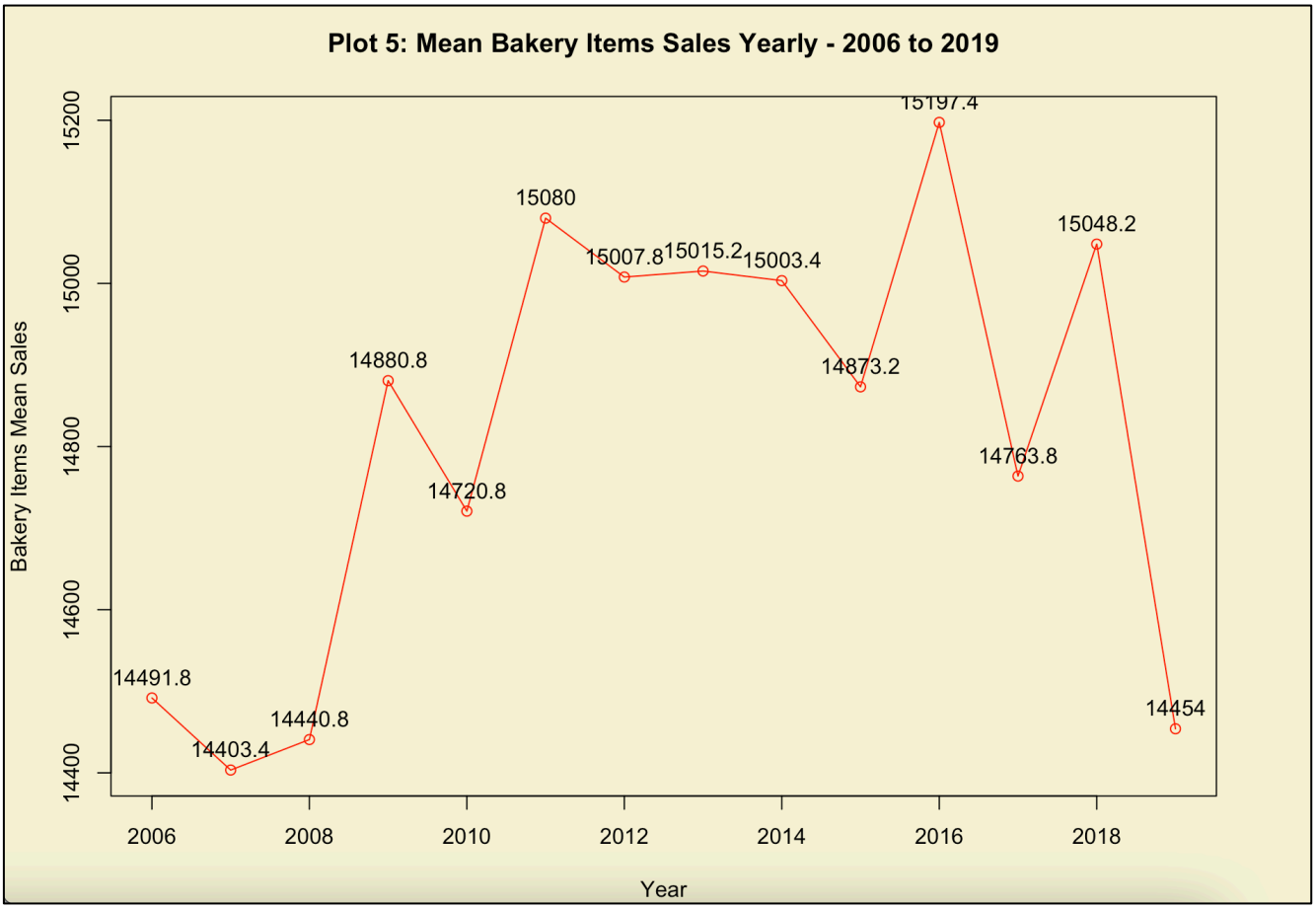

Figure 14: Plot of Total Bakery Items Yearly

a. It can be observed that the yearly sales of the bakery store picked **up pace and increased from the year 2007 till 2011 at a high pace** with a little decline in between the timeline. The sales figure reached from a mean of **14403.4** to **15080**.

b. The mean values somewhat represent the whole population. But, in our case, we already have plotted total sales and mean sales, and can analyse that the **highest sales of bakery items of the bakery store was in 2016. The lowest sales of bakery items of the bakery store was in 2007.**

c. The below figure shows the summary of a new variable introduced containing aggregate summation of some features. Mean, Standard Deviation, Minimum, Maximum, Quartiles values have been calculated and can be referred below in the screenshot :

Mean	14	14812.9	269.894	14403.4	14549.05	15013.35	15197.4

Figure 15: Summary of the new variable containing aggregate sums of each feature

CONCLUSION

The dataset of the bakery store has provided various insights about the sales of the store and its pattern over the years period of 2006 to 2019. We performed initial data analysis, exploratory data analysis, calculated various statistics, plotted several visualisation graphs in order to understand the analysis properly. The below points can be inferred from the analysis :

- The customers of the bakery store make the **highest number of purchases** of bakery items and goods on **Friday**.
- The customers make the **least number of purchases** of bakery items and goods on **Thursday**.
- **COOKIES** are the most purchased item by the customers amongst all the bakery items and goods of the bakery store. **3,07,375** cookies have been sold by the store from 2006 to 2019 which is *triple of the number of cakes* sold and *double of the number of pies* sold by the store in the same time.
- **CAKES** are the least purchased item by the customers amongst all the bakery items and goods sold by the bakery store.
- We can also infer that the bakery items and goods were sold in higher quantities when they were promoted by the store compared against when they were not promoted.
- We can also figure out that the sales of the bakery store declined from the year 2018 to 2019 drastically.

Some Follow-up Questions :

1. Since promotion has assisted the bakery store to sell more bakery items, what kind of promotions were made?
   We do not have any kind of information related to promotions and it would help us to analyse the promotion patterns and the kinds of it.
2. The data set does not contain any information related to prices of every item for each year. If the prices data would have been available, it would help the analysis to take more approaches towards the price aspect of each item and would make the patterns of purchases better understandable.
3. Since, the data set has abundant data with around 5,200 records in it and without any unbalanced features, we can apply learning algorithms to it. The classification and regression models can fit the training sets and predict the testing sets. We can tune our models for this data set properly and predict the future sales of the bakery store from them.

BIBLIOGRAPHY

1. *Bakery_Sales_Data_2006_19*. (2020, July 27). Kaggle. https://www.kaggle.com/sanu12300/bakery-sales-data-2006-19

2. Vicente, S., & Inbar, G. (2018, October 19). *Identify and Remove Duplicate Data in R*. Datanovia. https://www.datanovia.com/en/lessons/identify-and-remove-duplicate-data-in-r/

3. *A Grammar of Data Manipulation*. (2021). Dplyr. https://dplyr.tidyverse.org/

4. *ggplot2 - Essentials - Easy Guides - Wiki - STHDA*. (2021). GGPLOT2 | STHDA. http://www.sthda.com/english/wiki/ggplot2-essentials

5. *Home - RDocumentation*. (2021). Functions in R - Documentation. https://www.rdocumentation.org/

6. *Data Cleanup: Remove NA rows in R*. (2020, December 17). ProgrammingR. https://www.programmingr.com/examples/remove-na-rows-in-r/

7. Zeisig, K. (2021, October 29). *12 sales metrics to kick-start your sales analytics*. Klipfolio.Com. https://www.klipfolio.com/blog/sales-analytics-12-metrics

8. Olavsrud, T. (2021, February 8). *What is data analytics? Analyzing and managing data for decisions*. CIO. https://www.cio.com/article/3606151/what-is-data-analytics-analyzing-and-managing-data-for-decisions.html

9. Holtz, Y. (2021). *The R Graph Gallery – Help and inspiration for R charts*. The R Graph Gallery. https://www.r-graph-gallery.com/index.html

APPENDIX

```
#---------------------- GAUR_FinalProject_RScript ----------------------#

# Installing the packages.
install.packages("reshape")
install.packages("gridExtra")
install.packages("vtable")

# Importing the packages.
library("dplyr")
library("tidyr")
library("plyr")
library("tidyverse")
library("RColorBrewer")
library("plotrix")
library("scales")
library("ggplot2")
library("data.table")
library("reshape")
library("gridExtra")
library("vtable")

# STEP 2: Import Bakery's 'dataSet.csv' data set
# Note: Change the working directory as per the file's location.
setwd("/Users/HarshitGaur/Documents/Northeastern University/MPS Analytics/ALY 6000/Assignment 6/Bakery Sales/")
dataSet <- read.csv("bakery_sales.csv", header = TRUE)

# Display the data set.
View(dataSet)
View(tail(dataSet, 12))

# Remove duplicated rows based on a feature 'Date' as the data is distinct based on dates.
dataSet <- dataSet %>% distinct(Date, .keep_all = TRUE)

# Removing 'NA, Missing Values' from the data set.
dataSet <- na.omit(dataSet)

# Print the structure of 'dataSet.csv' data set
str(dataSet)
# Print the summary of 'dataSet.csv' data set
summary(dataSet)
st(dataSet)

# Mutate to sum up the quantities of all the items.
dataSet <- dataSet %>% mutate(Total_Sales = rowSums(.[4:8]))

# Summation of 'Total Sales' according to Days.
daysTable <- dataSet %>% group_by(weekday) %>% dplyr::summarise(Frequency = sum(Total_Sales))
daysTable <- data.frame((daysTable))

# Re-ordering the data frame based on 'Weekdays'
weekdaysList <- c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")
daysTable$weekday <- factor(daysTable$weekday, levels = weekdaysList)
daysTable <- daysTable[order(daysTable$weekday), ]


# Plot a Line plot of 'Days Sales'
# ---------------- Plot 1: Total Items Sold vs Week Days ---------------------- #
par(mar = c(4, 10, 8, 8) + 0.1)
ggplot(daysTable, aes(x = weekday, y = Frequency, group = 1)) +
  geom_line(linetype="dashed", color="blue", size=1.2) +
  geom_point(color="red", size=3) +
  theme(panel.background = element_rect("#c9f5e4")) +
  geom_text(aes(y = Frequency, label = Frequency), hjust = -0.4, vjust = 1, size = 4) +
  labs(
    title = "Plot 1: Total Sales vs Week Days",
    x = "Items Sold",
    y = "Week Days"
```

```
  ) + theme(plot.title = element_text(hjust = 0.5, colour = "#7F3D17", face = "bold"))
################ Plot 1: Total Items Sold vs Week Days {End} ##################

# Creating Data Frame with Summation of 'Cakes, Pies, Cookies, Smoothies, Coffee'
sumDf <- data.frame(
  "Cakes" = sum(dataSet$cakes),
  "Pies" = sum(dataSet$pies),
  "Cookies" = sum(dataSet$cookies),
  "Smoothies" = sum(dataSet$smoothies),
  "Coffee" = sum(dataSet$coffee)
)

# Melting the data frame for proper assignment.
sumDf <- melt(sumDf)

# Rename the columns of the data frame.
colnames(sumDf) <- c("BakeryItems", "ItemsSold")

# Plot a Bar plot of 'Bakery Items Sales Count'
# ---------------- Plot 2: Bakery Items Sales Count ------------------- #
par(mar = c(2, 6, 3, 2) + 0.1)
ggplot(sumDf)+
  geom_bar(
    mapping = aes(
      x = BakeryItems,
      y = ItemsSold
    ),
    stat="identity"
  ) +
  theme(panel.background = element_rect("#a7f4fc")) +
  labs(
    title = "Plot 2: Bakery Items Sold : 2006 - 2019",
    x = "Bakery Items",
    y = "Items Sold"
  ) +
  theme(plot.title = element_text(hjust = 0.5, colour = "#7F3D17", face = "bold")) +
  scale_y_continuous(labels = comma) +
  geom_text(aes(x = BakeryItems, y = ItemsSold, label = ItemsSold), vjust=1.8, color = "white", size=4)

################ Plot 2: Bakery Items Sold : 2006 - 2019 {End} ##################

bxp1 <- ggplot(dataSet) +
  geom_boxplot(aes(x = cakes, fill = promotion), outlier.color = "YELLOW") +
  coord_flip() +
  theme(panel.background = element_rect("LIGHTBLUE")) +
  labs(
    title = "Plot 3A: Cakes vs Promotion",
    y = "Promotion",
    x = "Cakes"
  ) +
  theme(plot.title = element_text(hjust = 0.5, colour = "#7F3D17", face = "bold"))

bxp2 <- ggplot(dataSet) +
  geom_boxplot(aes(x = pies, fill = promotion), outlier.color = "YELLOW") +
  coord_flip() +
  theme(panel.background = element_rect("LIGHTBLUE")) +
  labs(
    title = "Plot 3B: Pies vs Promotion",
    y = "Promotion",
    x = "Pies"
  ) +
  theme(plot.title = element_text(hjust = 0.5, colour = "#7F3D17", face = "bold"))

bxp3 <- ggplot(dataSet) +
  geom_boxplot(aes(x = cookies, fill = promotion), outlier.color = "YELLOW") +
  coord_flip() +
  theme(panel.background = element_rect("LIGHTBLUE")) +
  labs(
    title = "Plot 3C: Cookies vs Promotion",
    y = "Promotion",
    x = "Cookies"
```

```
  ) + theme(plot.title = element_text(hjust = 0.5, colour = "#7F3D17", face = "bold"))
bxp4 <- ggplot(dataSet) +
  geom_boxplot(aes(x = smoothies, fill = promotion), outlier.color = "YELLOW") +
  coord_flip() +
  theme(panel.background = element_rect("LIGHTBLUE")) +
  labs(
    title = "Plot 3D: Smoothies vs Promotion",
    y = "Promotion",
    x = "Smoothies"
  ) +
  theme(plot.title = element_text(hjust = 0.5, colour = "#7F3D17", face = "bold"))

bxp5 <- ggplot(dataSet) +
  geom_boxplot(aes(x = coffee, fill = promotion), outlier.color = "YELLOW") +
  coord_flip() +
  theme(panel.background = element_rect("LIGHTBLUE")) +
  labs(
    title = "Plot 3E: Coffee vs Promotion",
    y = "Promotion",
    x = "Coffee"
  ) +
  theme(plot.title = element_text(hjust = 0.5, colour = "#7F3D17", face = "bold"))

grid.arrange(bxp1, bxp2, bxp3, bxp4, bxp5)


################ Plot 3: Box Plot {End} ##################

# Format the 'Date' feature for Year.
dateFormat <- as.POSIXct(dataSet$Date, format = "%Y-%m-%d")
dataSet$Year <- format(dateFormat, format = "%Y")

# Make new data frame with grouping of summation of features.
yearlySumDf <- dataSet %>% group_by(Year) %>% dplyr::summarise(
                                        SumCakes = sum(cakes),
                                        SumPies = sum(pies),
                                        SumCookies = sum(cookies),
                                        SumSmoothies = sum(smoothies),
                                        SumCoffee = sum(coffee)
                                      )

yearlySumDf$Total <- yearlySumDf$SumCakes + yearlySumDf$SumPies + yearlySumDf$SumCookies + yearlySumDf$SumSmoothies
+ yearlySumDf$SumCoffee

#Adding Background
par(bg = '#dae6e0')

par(mar = c(4, 4, 4, 4) + 0.1)
brp <- barplot(yearlySumDf$Total, names.arg = yearlySumDf$Year, main = "Plot 4: Bakery Items Sold Yearly - 2006 to 2019",
          col = brewer.pal(9, "GnBu"), ylim = range(pretty(c(0, yearlySumDf$Total))), xlab = "Years", ylab = "Items Sold (numbers)")
text(x = brp, y = yearlySumDf$Total, labels = yearlySumDf$Total, pos = 3, col = "black")
box()

################ Plot 4: Bakery Items Sold Yearly - 2006 to 2019 {End} ##################

#Adding Background
par(bg = '#f5f3d5')

yearlySumDf$Mean <- (yearlySumDf$Total) / 5
msPlt <- plot(x = yearlySumDf$Year, y = yearlySumDf$Mean, type = "o", col = "RED", xlab = "Year", ylab = "Bakery Items Mean
Sales", main = "Plot 5: Mean Bakery Items Sales Yearly - 2006 to 2019", names.arg = yearlySumDf$Year)
text(x = yearlySumDf$Year, y = yearlySumDf$Mean, labels = yearlySumDf$Mean, pos = 3, col = "BLACK")

################ Plot 5: Bakery Items Mean Sales Yearly - 2006 to 2019 {End} ##################
```