



SALES PRODUCT DATA ANALYSIS.

Prepared By:
Versacorp Group
Workshop 1/2
14:00 - 1600



**Northumbria
University**
NEWCASTLE

ABOUT THE DATA



- This dataset is about a sales analysis of an online store.
- Sales analysis is the process of mining your data to evaluate your sales team's performance in relation to its goals. It provides insights into the top performing and failing products/services, selling issues and market opportunities, sales forecasting, and revenue-generating sales activities.
- The products sold are mostly technological products and accessories such as Laptops, iPhones, charging cables, etc.
- The dataset is available on Kaggle.
- The duration of the data collected is from January to December 2019.
- Data for each months are available in separate spreadsheets.
- Link to the data set:
<https://www.kaggle.com/datasets/knightbearr/sales-product-data>

KEY VARIABLES

- Order ID - this is exclusively used to track orders. Each order receives its own Order ID that will not be duplicated. This number can be useful to the seller when attempting to find out certain details about an order such as shipment date or status.
- Product - The Product that have been sold.
- Quantity Ordered - is the total number of items ordered.
- Price Each - The price of each products.
- Order Date - Date of request by the customer.
- Purchase Address – This is the customer's details, usually includes a PO number, which is useful in matching shipments with purchases; a shipping date; billing address; shipping address; and the request items, quantities and price.

DATA VIEW

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
2	NA	NA	NA	NA	NA	NA
3	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
4	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
5	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
6	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001
7	176562	USB-C Charging Cable	1	11.95	04/29/19 13:03	381 Wilson St, San Francisco, CA 94016
8	176563	Bose SoundSport Headphones	1	99.99	04/02/19 07:46	668 Center St, Seattle, WA 98101
9	176564	USB-C Charging Cable	1	11.95	04/12/19 10:58	790 Ridge St, Atlanta, GA 30301
10	176565	Macbook Pro Laptop	1	1700	04/24/19 10:38	915 Willow St, San Francisco, CA 94016
11	176566	Wired Headphones	1	11.99	04/08/19 14:05	83 7th St, Boston, MA 02215

PURPOSE OF THE STUDY



The goal is to isolate metrics that matter most for the company so that management can make educated decisions and boost productivity.



To get insight on data to boost sales.



To analyze sales by mining existing data.



To shed light on the company's best and worst selling items and services.

POTENTIAL INSIGHTS FROM THE DATASET



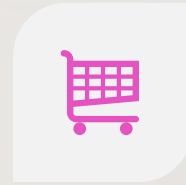
WHAT MONTH
HAD THE HIGHEST
NUMBER OF SALES?
WHAT WAS THE
TOTAL AMOUNT
EARNED DURING
THAT MONTH?



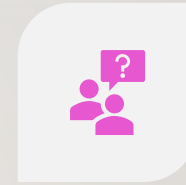
WHICH CITY HAD
THE HIGHEST SALES
VOLUME?



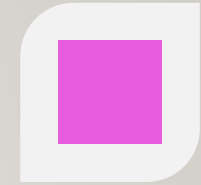
WHAT TIME SHOULD
ADVERTISEMENT BE
MADE TO INCREASE
SALES?



WHAT PRODUCTS
ARE OFTEN SOLD
TOGETHER?



WHAT PRODUCT
SOLD THE MOST,
AND WHY?



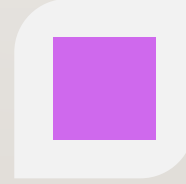
WHAT PRODUCT
HAD THE HIGHEST
SALES?



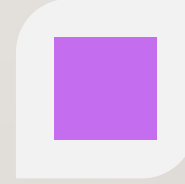
WHAT DAY OF THE
WEEK HAD THE
MOST SALES?



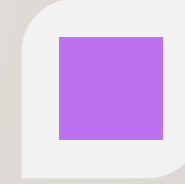
HOW MANY
PRODUCT EACH
WAS SOLD IN EACH
CITY?



SALES OF EACH
PRODUCT SOLD IN
EACH MONTH



HOW MANY ORDERS
WERE FOR EACH
PRODUCT FOR ALL
THE MONTHS?



WHICH PRODUCT
GENERATED THE
HIGHEST SALES?

DATA PREPARATION

- Installation of packages.
- Importation of Datasets.

```
1 #Installing packages
2
3 install.packages("tidyverse")
4 install.packages("lubridate")
5 install.packages("ggplot2")
6 install.packages("anytime")
7 install.packages("scales")
8 install.packages("hrbrthemes")
9 install.packages("qqman")
10 install.packages("ggpubr")
11 library(tidyverse)
12 library(lubridate)
13 library(ggplot2)
14 library(anytime)
15 library(stringr)
16 library(scales)
17 library(gridExtra)
18 library(ggthemes)
19 library(hrbrthemes)
20 library(qqman)
21 library(ggpubr)
22
23
24 #Importing Datasets
25
26 January <- read_csv("Sales_January_2019.csv")
27 February <- read_csv("Sales_February_2019.csv")
28 March <- read_csv("Sales_March_2019.csv")
29 April <- read_csv("Sales_April_2019.csv")
30 May <- read_csv("Sales_May_2019.csv")
31 June <- read_csv("Sales_June_2019.csv")
32 July <- read_csv("Sales_July_2019.csv")
33 August <- read_csv("Sales_August_2019.csv")
34 September <- read_csv("Sales_September_2019.csv")
35 October <- read_csv("Sales_October_2019.csv")
36 November <- read_csv("Sales_November_2019.csv")
37 December <- read_csv("Sales_December_2019.csv")
38
```

DATA MANIPULATION

- Merged datasets (January – December) into a single data frame using `bind_rows` function.
- Changed variable names to match R variable name structure.

```
#DATA MANIPULATION
```

```
# Marging all datasets into a single data frame
```

```
sales_2019 <- bind_rows(January,  
                        February,  
                        March,  
                        April,  
                        May,  
                        June,  
                        July,  
                        August, |  
                        September,  
                        October,  
                        November,  
                        December)
```

	Order_ID	Product	Quantity_Ordered	Price	Order_Date	Purchase_Address
1	141234	iPhone	1	700	01/22/19 21:25	944 Walnut St, Boston, MA 02215
2	141235	Lightning Charging Cable	1	14.95	01/28/19 14:15	185 Maple St, Portland, OR 97035
3	141236	Wired Headphones	2	11.99	01/17/19 13:33	538 Adams St, San Francisco, CA 94016
4	141237	27in FHD Monitor	1	149.99	01/05/19 20:33	738 10th St, Los Angeles, CA 90001
5	141238	Wired Headphones	1	11.99	01/25/19 11:59	387 10th St, Austin, TX 73301
6	141239	AAA Batteries (4-pack)	1	2.99	01/29/19 20:22	775 Willow St, San Francisco, CA 94016
7	141240	27in 4K Gaming Monitor	1	389.99	01/26/19 12:16	979 Park St, Los Angeles, CA 90001
8	141241	USB-C Charging Cable	1	11.95	01/05/19 12:04	181 6th St, San Francisco, CA 94016
9	141242	Bose SoundSport Headphones	1	99.99	01/01/19 10:30	867 Willow St, Los Angeles, CA 90001
10	141243	Apple AirPods Headphones	1	150	01/22/19 21:20	657 Johnson St, San Francisco, CA 94016
11	141244	Apple AirPods Headphones	1	150	01/07/19 11:29	492 Walnut St, San Francisco, CA 94016

```
#Changing name of Variables
```

```
sales_2019 <- sales_2019 %>%  
  rename(quantity_ordered = `quantity ordered`,  
         order_ID = `order ID`,  
         Price = `Price Each`,  
         order_Date = `order Date`,  
         Purchase_Address = `Purchase Address`)
```


DATA MANIPULATION

- Checked the datatype of each variable using stringr function.

```
#checking the datatype of the variables
```

```
str(Sales_2019)
```

```
> str(Sales_2019)
spec_tbl_df [186,850 × 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Order_ID      : chr [1:186850] "141234" "141235" "141236" "141237" ...
 $ Product       : chr [1:186850] "iPhone" "Lightning Charging Cable" "Wired
 $ Quantity_Ordered: chr [1:186850] "1" "1" "2" "1" ...
 $ Price         : chr [1:186850] "700" "14.95" "11.99" "149.99" ...
 $ Order_Date    : chr [1:186850] "01/22/19 21:25" "01/28/19 14:15" "01/17/1
 $ Purchase_Address: chr [1:186850] "944 Walnut St, Boston, MA 02215" "185 Map
 ncisco, CA 94016" "738 10th St, Los Angeles, CA 90001" ...
 - attr(*, "spec")=
 .. cols(
 ..   `Order ID` = col_character(),
 ..   Product = col_character(),
 ..   `Quantity Ordered` = col_character(),
 ..   `Price Each` = col_character(),
 ..   `Order Date` = col_character(),
 ..   `Purchase Address` = col_character()
 .. )
 - attr(*, "problems")=<externalptr>
> |
```

Product	Quantity_Ordered	Price	Order_Date	Purchase_Address	Month	Day	Year	Day_of_Week	Hour
iPhone	1	700	2019-01-22 21:25:00	944 Walnut St, Boston, MA 02215	January	22	19	Tuesday	21
Lightning Charging Cable	1	14	2019-01-28 14:15:00	185 Maple St, Portland, OR 97035	January	28	19	Monday	14
Wired Headphones	2	11	2019-01-17 13:33:00	538 Adams St, San Francisco, CA 94016	January	17	19	Thursday	13
27in FHD Monitor	1	149	2019-01-05 20:33:00	738 10th St, Los Angeles, CA 90001	January	05	19	Saturday	20
Wired Headphones	1	11	2019-01-25 11:59:00	387 10th St, Austin, TX 73301	January	25	19	Friday	11
AAA Batteries (4-pack)	1	2	2019-01-29 20:22:00	775 Willow St, San Francisco, CA 94016	January	29	19	Tuesday	20
27in 4K Gaming Monitor	1	389	2019-01-26 12:16:00	979 Park St, Los Angeles, CA 90001	January	26	19	Saturday	12
USB-C Charging Cable	1	11	2019-01-05 12:04:00	181 6th St, San Francisco, CA 94016	January	05	19	Saturday	12
Bose SoundSport Headphones	1	99	2019-01-01 10:30:00	867 Willow St, Los Angeles, CA 90001	January	01	19	Tuesday	10

- Changed variable datatype;
Order_Date – as.POSIXct
Quantity_Ordered – as.integer
Price – as.integer
- Created Month, Day, Year, Hour and Day of the week from Order_Date.

```
#Changing variable datatype of Order_Date
```

```
Sales_2019$Order_Date = as.POSIXct(Sales_2019$Order_Date, format = "%m/%d/%y %H:%M")
```

```
#Changing Quantity_Ordered and Price Data Type to integer
```

```
Sales_2019$Quantity_Ordered<- as.integer(Sales_2019$Quantity_Ordered)
```

```
Sales_2019$Price<- as.integer(Sales_2019$Price)
```

```
#Creating Month, Day, Year and Day of the week from Order_Date variable
```

```
Sales_2019$Month <- format(Sales_2019$Order_Date, "%B")
```

```
Sales_2019$Day <- format(as.Date(Sales_2019$Order_Date), "%d")
```

```
Sales_2019$Year <- format(as.Date(Sales_2019$Order_Date), "%y")
```

```
Sales_2019$Day_of_Week <- format(as.Date(Sales_2019$Order_Date), "%A")
```

```
#Creating Hour variable from Order_Date
```

```
Sales_2019$Hour = format(Sales_2019$Order_Date, "%H")
```

DATA MANIPULATION

```
#Separating Address, City and State from Purchase_address
```

```
sales_2019 <- sales_2019 %>%  
  separate(Purchase_Address, c("Address", "City", "State"), sep = ",")
```

```
#Separating State and Zipcode from State
```

```
sales_2019 <- sales_2019 %>%  
  separate(State, c("EmptySpace", "State", "zipcode"), sep = " ")
```

```
# Deleting variables|
```

```
sales_2019$EmptySpace <- NULL
```

```
sales_2019$...1 <- NULL
```

```
sales_2019$Year <- NULL
```

- Separated the Purchase_Address variable to create City, State, Zipcode.
- Deleted the irrelevant variables.

Quantity_Ordered	Price	Order_Date	Address	City	State	Zipcode	Month	Day	Day_of_Week	Hour
1	700	2019-01-22 21:25:00	944 Walnut St	Boston	MA	02215	January	22	Tuesday	21
1	14	2019-01-28 14:15:00	185 Maple St	Portland	OR	97035	January	28	Monday	14
2	11	2019-01-17 13:33:00	538 Adams St	San Francisco	CA	94016	January	17	Thursday	13
1	149	2019-01-05 20:33:00	738 10th St	Los Angeles	CA	90001	January	05	Saturday	20
1	11	2019-01-25 11:59:00	387 10th St	Austin	TX	73301	January	25	Friday	11
1	2	2019-01-29 20:22:00	775 Willow St	San Francisco	CA	94016	January	29	Tuesday	20
1	389	2019-01-26 12:16:00	979 Park St	Los Angeles	CA	90001	January	26	Saturday	12
1	11	2019-01-05 12:04:00	181 6th St	San Francisco	CA	94016	January	05	Saturday	12
1	99	2019-01-01 10:30:00	867 Willow St	Los Angeles	CA	90001	January	01	Tuesday	10
1	150	2019-01-22 21:20:00	657 Johnson St	San Francisco	CA	94016	January	22	Tuesday	21

DATA MANIPULATION

	Order_ID	Product	Quantity_Ordered	Price	Order_Date	Address	City	State	Zipcode	Month	Day	Day_of_Week
1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
6	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
7	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
12	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
16	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
17	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
19	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA
20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
21	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	Order ID	Product	NA	NA	NA	Purchase Address	NA	NA	NA	NA	NA	NA

Showing 1 to 23 of 900 entries, 13 total columns

- Filtered out the missing values.
- We have 900 entries of missing rows.
- Created new df (Clean_Sales_2019)

```
#Checking out empty columns (we have 900 entries of not completed rolls/missing values)
```

```
Missing_values <- Sales_2019 %>%
  select(Order_ID, Product, everything()) %>%
  filter(!complete.cases(.))
```

```
#Cleaned Data Set
```

```
Clean_sales_2019 <- Sales_2019 %>%
  select(Order_ID, Product, everything()) %>%
  filter(complete.cases(.))
```

DATA MANIPULATION


```
#Multiplied Quantity Ordered by Price to get Sales column
```

```
Clean_Sales_2019$Sales <- Clean_Sales_2019$Quantity_Ordered * Clean_Sales_2019$Price
```

Created new variable “Sales” by multiplying “Quantity_Ordered” by “Price”.

Quantity_Ordered	Price	Order_Date	Address	City	State	Zipcode	Month	Day	Day_of_Week	Hour	Sales
1	700	2019-01-22 21:25:00	944 Walnut St	Boston	MA	02215	January	22	Tuesday	21	700
1	14	2019-01-28 14:15:00	185 Maple St	Portland	OR	97035	January	28	Monday	14	14
2	11	2019-01-17 13:33:00	538 Adams St	San Francisco	CA	94016	January	17	Thursday	13	22
1	149	2019-01-05 20:33:00	738 10th St	Los Angeles	CA	90001	January	05	Saturday	20	149
1	11	2019-01-25 11:59:00	387 10th St	Austin	TX	73301	January	25	Friday	11	11
1	2	2019-01-29 20:22:00	775 Willow St	San Francisco	CA	94016	January	29	Tuesday	20	2
1	389	2019-01-26 12:16:00	979 Park St	Los Angeles	CA	90001	January	26	Saturday	12	389
1	11	2019-01-05 12:04:00	181 6th St	San Francisco	CA	94016	January	05	Saturday	12	11

DATA MANIPULATION

FINAL VIEW OF DATA

	Order_ID	Product	Quantity_Ordered	Price	Order_Date	Address	City	State	Zipcode	Month	Day	Day_of_Week	Hour	Sales
1	141234	iPhone	1	700	2019-01-22 21:25:00	944 Walnut St	Boston	MA	02215	January	22	Tuesday	21	700
2	141235	Lightning Charging Cable	1	14	2019-01-28 14:15:00	185 Maple St	Portland	OR	97035	January	28	Monday	14	14
3	141236	Wired Headphones	2	11	2019-01-17 13:33:00	538 Adams St	San Francisco	CA	94016	January	17	Thursday	13	22
4	141237	27in FHD Monitor	1	149	2019-01-05 20:33:00	738 10th St	Los Angeles	CA	90001	January	05	Saturday	20	149
5	141238	Wired Headphones	1	11	2019-01-25 11:59:00	387 10th St	Austin	TX	73301	January	25	Friday	11	11
6	141239	AAA Batteries (4-pack)	1	2	2019-01-29 20:22:00	775 Willow St	San Francisco	CA	94016	January	29	Tuesday	20	2
7	141240	27in 4K Gaming Monitor	1	389	2019-01-26 12:16:00	979 Park St	Los Angeles	CA	90001	January	26	Saturday	12	389
8	141241	USB-C Charging Cable	1	11	2019-01-05 12:04:00	181 6th St	San Francisco	CA	94016	January	05	Saturday	12	11
9	141242	Bose SoundSport Headphones	1	99	2019-01-01 10:30:00	867 Willow St	Los Angeles	CA	90001	January	01	Tuesday	10	99
10	141243	Apple Airpods Headphones	1	150	2019-01-22 21:20:00	657 Johnson St	San Francisco	CA	94016	January	22	Tuesday	21	150
11	141244	Apple Airpods Headphones	1	150	2019-01-07 11:29:00	492 Walnut St	San Francisco	CA	94016	January	07	Monday	11	150
12	141245	Macbook Pro Laptop	1	1700	2019-01-31 10:12:00	322 6th St	San Francisco	CA	94016	January	31	Thursday	10	1700
13	141246	AAA Batteries (4-pack)	3	2	2019-01-09 18:57:00	618 7th St	Los Angeles	CA	90001	January	09	Wednesday	18	6
14	141247	27in FHD Monitor	1	149	2019-01-25 19:19:00	512 Wilson St	San Francisco	CA	94016	January	25	Friday	19	149
15	141248	Flatscreen TV	1	300	2019-01-03 21:54:00	363 Spruce St	Austin	TX	73301	January	03	Thursday	21	300
16	141249	27in FHD Monitor	1	149	2019-01-05 17:20:00	440 Cedar St	Portland	OR	97035	January	05	Saturday	17	149
17	141250	Vareebadd Phone	1	400	2019-01-10 11:20:00	471 Center St	Los Angeles	CA	90001	January	10	Thursday	11	400
18	141251	Apple Airpods Headphones	1	150	2019-01-24 08:13:00	414 Walnut St	Boston	MA	02215	January	24	Thursday	08	150
19	141252	USB-C Charging Cable	1	11	2019-01-30 09:28:00	220 9th St	Los Angeles	CA	90001	January	30	Wednesday	09	11
20	141253	AA Batteries (4-pack)	1	3	2019-01-17 00:09:00	385 11th St	Atlanta	GA	30301	January	16	Wednesday	00	3
21	141254	AAA Batteries (4-pack)	1	2	2019-01-08 11:51:00	238 Sunset St	Seattle	WA	98101	January	08	Tuesday	11	2
22	141255	USB-C Charging Cable	1	11	2019-01-09 20:55:00	764 11th St	Los Angeles	CA	90001	January	09	Wednesday	20	11

Showing 1 to 23 of 185,950 entries, 14 total columns

The final data has 14 variables and 185,950 entries

DATA VISUALIZATION

- The data was grouped by “Month” then summarized by the sum of Sales.
- ggplot function was used to plot the graph.

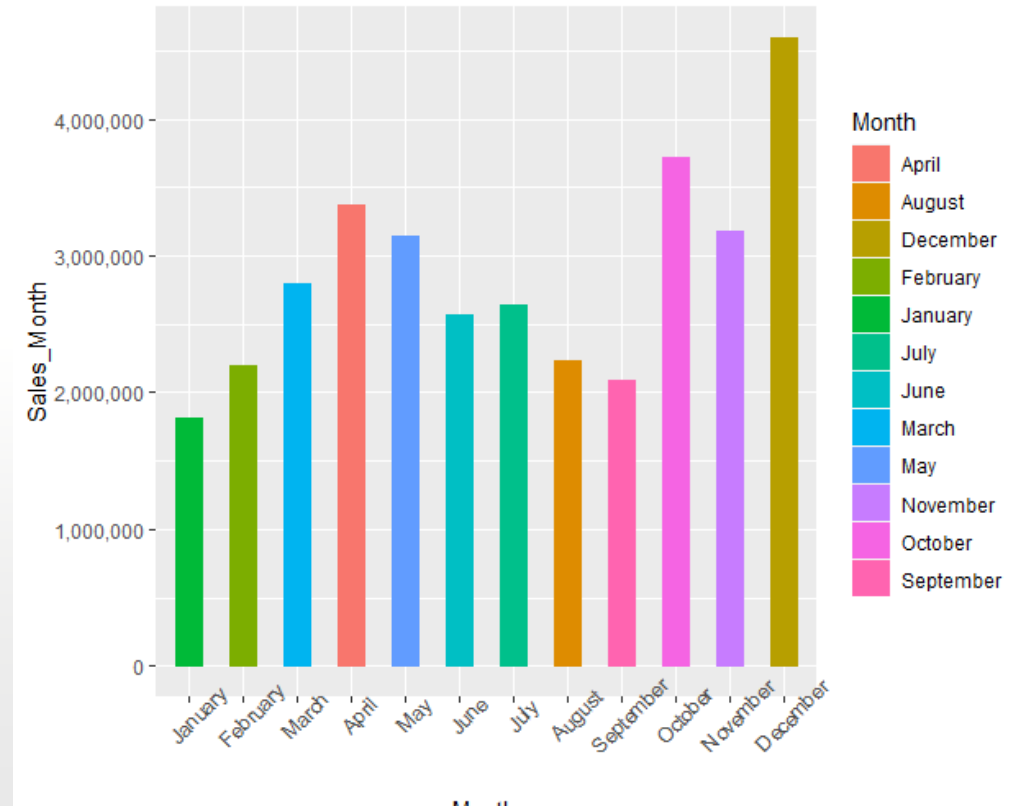
```
#DATA VISUALIZATION
```

```
#Q: what was the best month for sales? How much was earned that month?
```

```
clean_sales_2019 %>%  
  group_by(Month) %>%  
  summarise(Sales_Month = sum(Sales)) %>%  
  arrange(desc(Month)) %>%  
  ggplot(aes(x = Month, y = Sales_Month, color = Month, fill = Month))+  
  geom_col(width=0.5, position = position_dodge(width=0.5))+  
  theme(axis.text.x = element_text(angle = 45))+  
  labs(title = "what was the best month for sales? How much was earned that month?")+  
  scale_y_continuous(labels = comma)+  
  scale_x_discrete(limits = month.name)
```

WHAT MONTH HAD THE HIGHEST NUMBER OF SALES? WHAT WAS THE TOTAL AMOUNT EARNED DURING THAT MONTH?

- From the chart, December had the most sales.
- \$4,591,824 was earned during the month of December.
- The least sales was recorded in January with \$1,813,956.



DATA VISUALIZATION

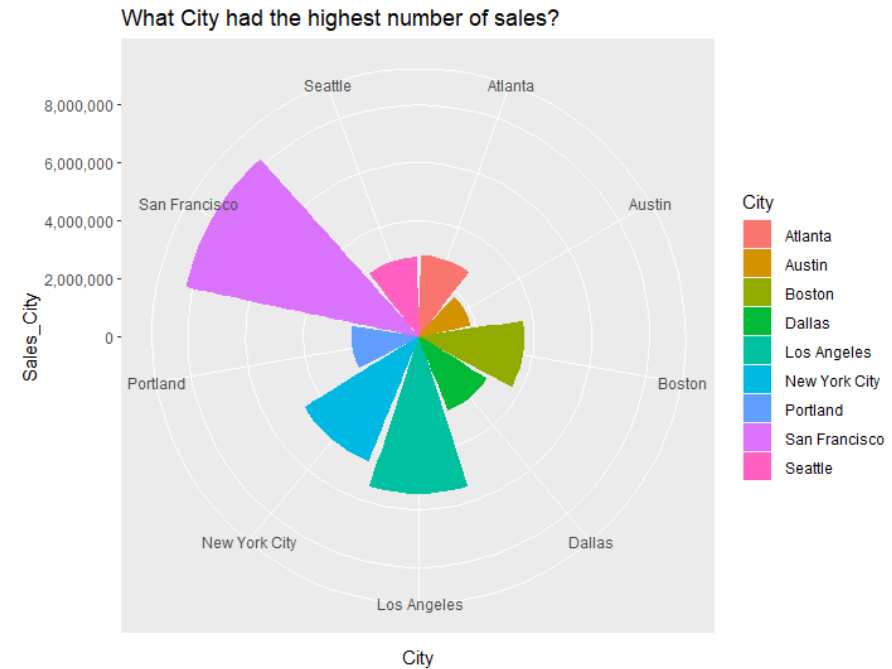
- The data was grouped by “City” then summarized by the sum of Sales.
- ggplot function was used to plot the graph.

```
#Q: what city had the highest number of sales?
```

```
clean_sales_2019 %>%  
  group_by(City) %>%  
  summarise(Sales_City = sum(Sales)) %>%  
  arrange(desc(Sales_City)) %>%  
  ggplot(aes(x = City, y = Sales_City, color = City, fill = City))+  
  geom_bar(stat="identity", position=position_dodge())+  
  coord_polar(theta = "x", start=0)+  
  labs(title = "what city had the highest number of sales?")+  
  scale_y_continuous(labels = comma)
```

WHICH CITY HAD THE HIGHEST SALES VOLUME?

- San Francisco had the highest sales with total sales reaching \$8,262,203.
- Los Angeles recorded \$5,426,973 which is the second highest sales volume.
- The least sales was recorded in Austin with \$1,811,054.



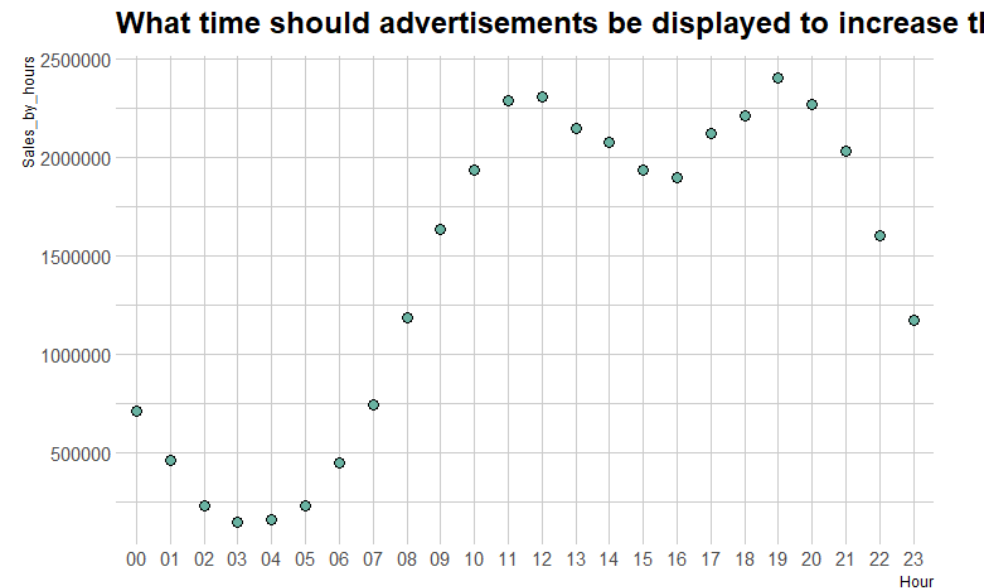
DATA VISUALIZATION

- The data was grouped by “Hour” then summarized by the sum of Sales.
- ggplot function was used to plot the graph.

```
#Q what time should advertisements be displayed to increase the likelihood of a sale?  
  
clean_sales_2019 %>%  
  group_by(Hour) %>%  
  summarise(Sales_by_hours = sum(Sales)) %>%  
  arrange(desc(Sales_by_hours)) %>%  
  tail(24) %>%  
  ggplot(aes(x=Hour, y=Sales_by_hours)) +  
  geom_line(color="black") +  
  geom_point(shape=21, color="black", fill="#69b3a2", size=3) +  
  theme_ipsum() +  
  ggtitle("what time should advertisements be displayed to increase the likelihood of a sale?")
```

WHAT TIME SHOULD ADVERTISEMENT BE MADE TO INCREASE SALES?

- From the graph, it shows that sales starts to increase around 9:00 and sales begins to drop significantly around 22:00.
- We would suggest to the organization to display advertisement between the hours 9:00 – 22:00 as this increases the likelihood of sales of the advertised products.
- Sales significantly increase during this time period because this is when people are most active.



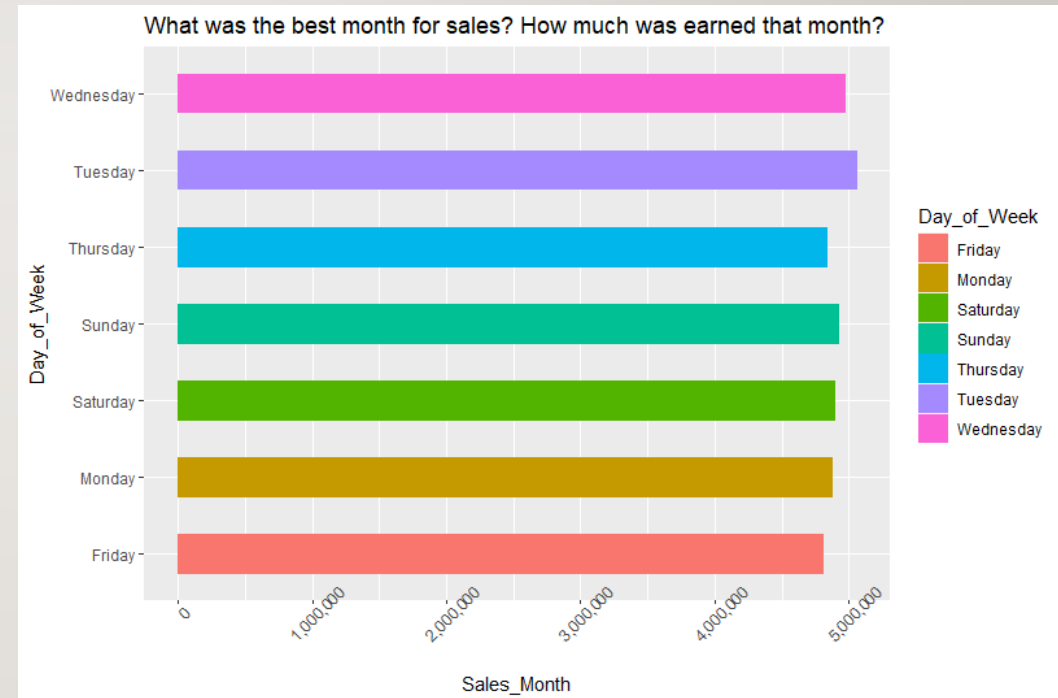
DATA VISUALIZATION

- The data was grouped by “Day_of_week” then summarized it by the sum of Sales.
- ggplot function was used to plot the graph.

```
#what Day of the week has the most sales?  
  
Clean_sales_2019 %>%  
  group_by(Day_of_week) %>%  
  summarise(Sales_Month = sum(Sales)) %>%  
  arrange(desc(Day_of_week)) %>%  
  ggplot(aes(x = Day_of_week, y = Sales_Month, color = Day_of_week, fill = Day_of_week))+  
  geom_col(width=0.5, position = position_dodge(width=0.5))+  
  theme(axis.text.x = element_text(angle = 45))+  
  labs(title = "what was the best month for sales? How much was earned that month?")+  
  scale_y_continuous(labels = comma)+  
  rotate()
```

WHAT DAY OF THE WEEK HAS THE MOST SALES?

- Tuesday has the most sales followed by Wednesday.
- However, there is no significant difference in sales during the rest of the week.



DATA VISUALIZATION

- All duplicated “Order_ID” were selected and a new dataframe called “Products_Sold_Together” was created.
- The new dataframe was grouped by “Order_ID” and mutate function was used to create “Group_Product”.
- Grouped by “Group_Product” and summarized by the number of instances it occurred.

```
#Q what products are most often sold together?  
  
Products_Sold_Together <- Clean_Sales_2019[duplicated(Clean_Sales_2019$Order_ID) | duplicated(Clean_Sales_2019$Order_ID, fromLast = TRUE), ]  
  
Products_Sold_Together <- Products_Sold_Together %>%  
  group_by(Order_ID) %>%  
  mutate('Group_Product' = paste0(Product, collapse = ", ")) %>%  
  ungroup()  
  
#####  
  
Products_Sold_Count <- Products_Sold_Together %>%  
  group_by(Group_Product) %>%  
  summarise(total_count=n(), .groups = 'drop') %>%  
  arrange(desc(total_count))
```

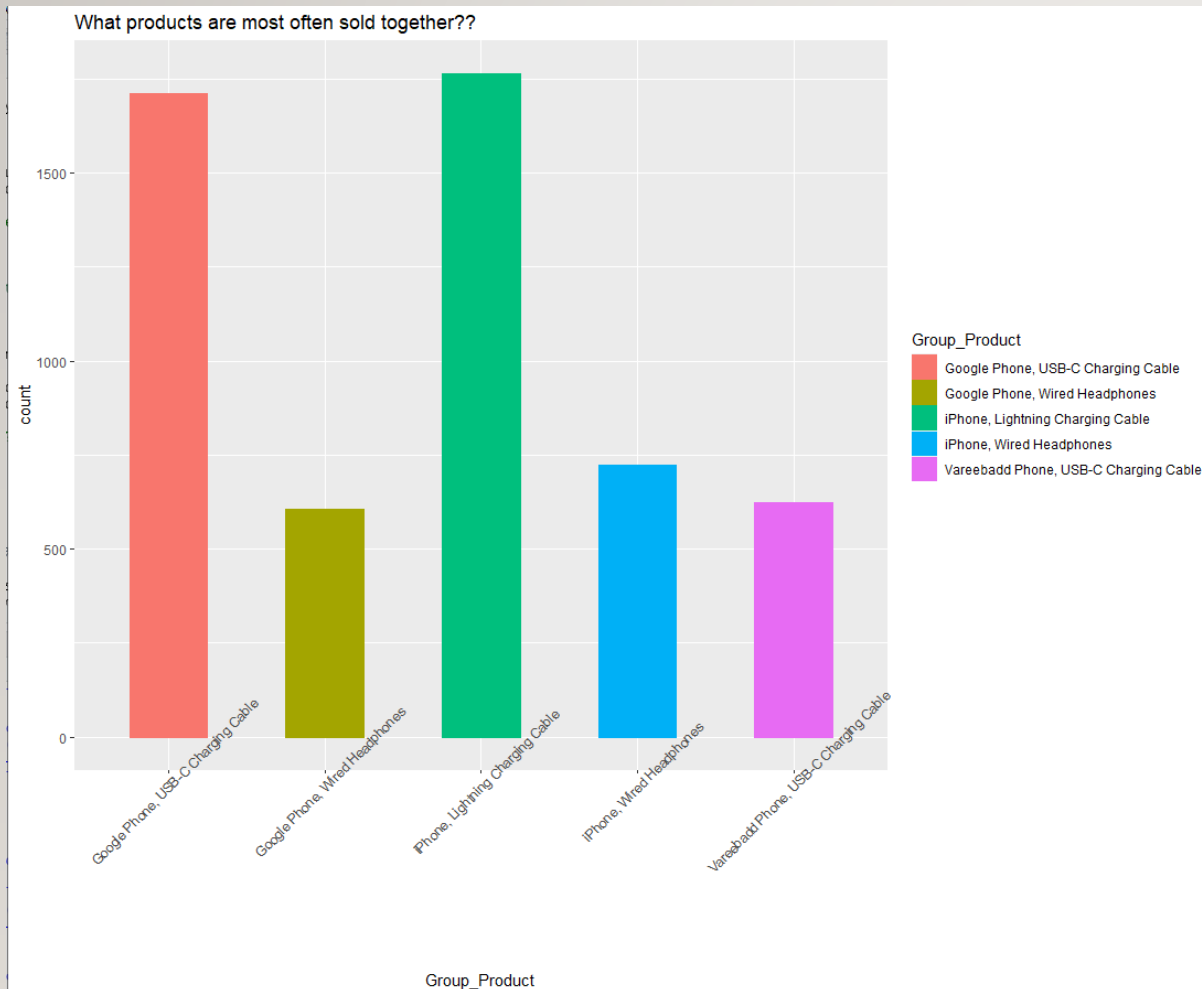
WHAT PRODUCTS ARE MOST OFTEN SOLD TOGETHER?

- iPhone and Lightning Charging Cable are mostly purchased together with 1764 orders closely followed by Google Phone and USB-C Charging Cable with 1712 orders.
- We have a total of 366 entries.
- Next slide shows a plot for the top 5.

	Group_Product	total_count
1	iPhone, Lightning Charging Cable	1764
2	Google Phone, USB-C Charging Cable	1712
3	iPhone, Wired Headphones	722
4	Vareebadd Phone, USB-C Charging Cable	624
5	Google Phone, Wired Headphones	606
6	iPhone, Apple Airpods Headphones	572
7	Google Phone, Bose SoundSport Headphones	322
8	Google Phone, USB-C Charging Cable, Wired Headphones	231
9	Vareebadd Phone, Wired Headphones	208
10	iPhone, Lightning Charging Cable, Wired Headphones	168
11	iPhone, Lightning Charging Cable, Apple Airpods Headphon...	129
12	Vareebadd Phone, Bose SoundSport Headphones	120
13	Lightning Charging Cable, USB-C Charging Cable	110
14	Lightning Charging Cable, AA Batteries (4-pack)	104
15	AA Batteries (4-pack), Lightning Charging Cable	102
16	Lightning Charging Cable, Lightning Charging Cable	102
17	AAA Batteries (4-pack), USB-C Charging Cable	100
18	AA Batteries (4-pack), AAA Batteries (4-pack)	96
19	AAA Batteries (4-pack), AAA Batteries (4-pack)	96

Showing 1 to 19 of 366 entries, 2 total columns

WHAT PRODUCTS ARE MOST OFTEN SOLD TOGETHER?



```
Plot <- Products_Sold_Together %>% group_by(Group_Product) %>%  
  summarise(count = n()) %>%  
  top_n(n = 5, wt = count)  
  
ggplot(Plot, aes(x = Group_Product, y = count, color = Group_Product, fill = Group_Product))+  
  geom_col(width=0.5, position = position_dodge(width=0.5))+  
  theme(axis.text.x = element_text(angle = 90))+  
  labs(title = "what products are most often sold together??")
```

- The data was grouped by “Group_Product” and summarized by count focusing on the top 5 (top_n(n=5)).
- ggplot was used to plot the graph.

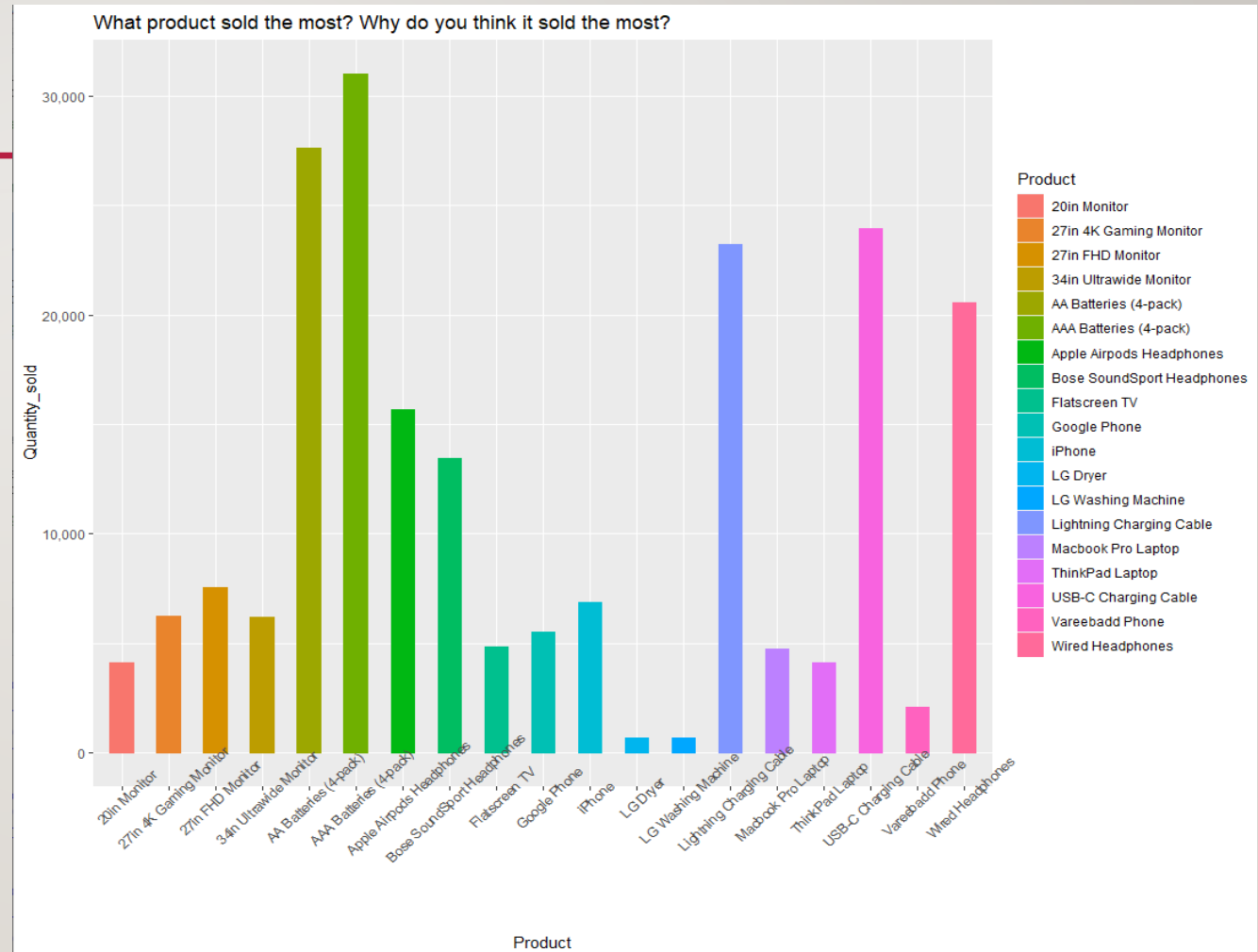
DATA VISUALIZATION

- The data was grouped by “Product” then summarized by the sum of “Quantity_Ordered”.
- ggplot function was used to plot the graph.

```
#Q what product sold the most? why do you think it sold the most?  
  
clean_sales_2019 %>%  
  group_by(Product) %>%  
  summarise(Quantity_sold = sum(Quantity_ordered)) %>%  
  arrange(desc(Quantity_sold)) %>%  
  ggplot(aes(x = Product, y = Quantity_sold, color = Product, fill = Product))+  
  geom_col(width=0.5, position = position_dodge(width=0.5))+  
  theme(axis.text.x = element_text(angle = 45))+  
  labs(title = "what product sold the most? why do you think it sold the most?")+  
  scale_y_continuous(labels = comma)  
,
```

WHAT PRODUCT SOLD THE MOST AND WHY?

- AAA Batteries (4-pack) sold the most with a total count of 31017 pieces sold.
- Most small electronic devices, like TV remote controls, kitchen timers, graphing calculators, and bathroom scales, use triple A batteries.
- AAA batteries are essential items for each household, and this might have been the reason it sold the most.
- Lightning Charging Cable, USB-C Charging Cable, and Wired Headphones are also part of the products that were mostly purchased



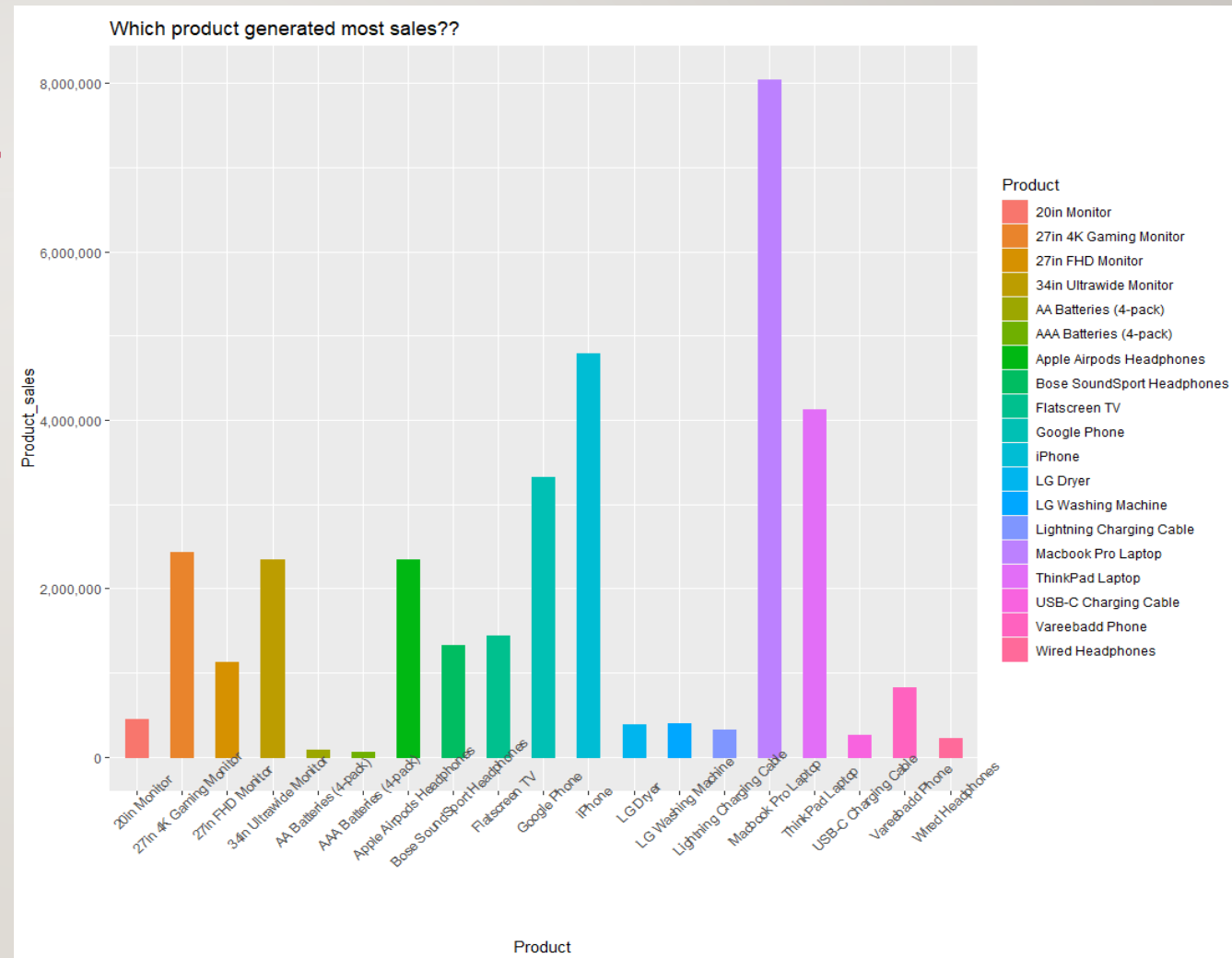
DATA VISUALIZATION

- The data was grouped by “Product” then summarized by the sum of Sales.
- ggplot function was used to plot the graph.

```
#Q which product generated most sales?  
  
clean_sales_2019 %>%  
  group_by(Product) %>%  
  summarise(Product_sales = sum(Sales)) %>%  
  arrange(desc(Product_sales)) %>%  
  ggplot(aes(x = Product, y = Product_sales, color = Product, fill = Product))+  
  geom_col(width=0.5, position = position_dodge(width=0.5))+  
  theme(axis.text.x = element_text(angle = 45))+  
  labs(title = "what product sold the most? why do you think it sold the most?")+  
  scale_y_continuous(labels = comma)  
|
```


WHAT PRODUCT HAD THE HIGHEST SALES?

- Macbook Pro Laptop generated the most sales.
- The cost of each Macbook Pro Laptop is \$1700.
- The Macbook Pro Laptop generated the most sales because they are the most expensive item, and a significant number (4728) of them were sold.



DATA VISUALIZATION

- The data was grouped by “Product” and “Month” then summarized by the sum of Quantity_Ordered.
- ggplot function with facet_wrap was used to plot the graph.

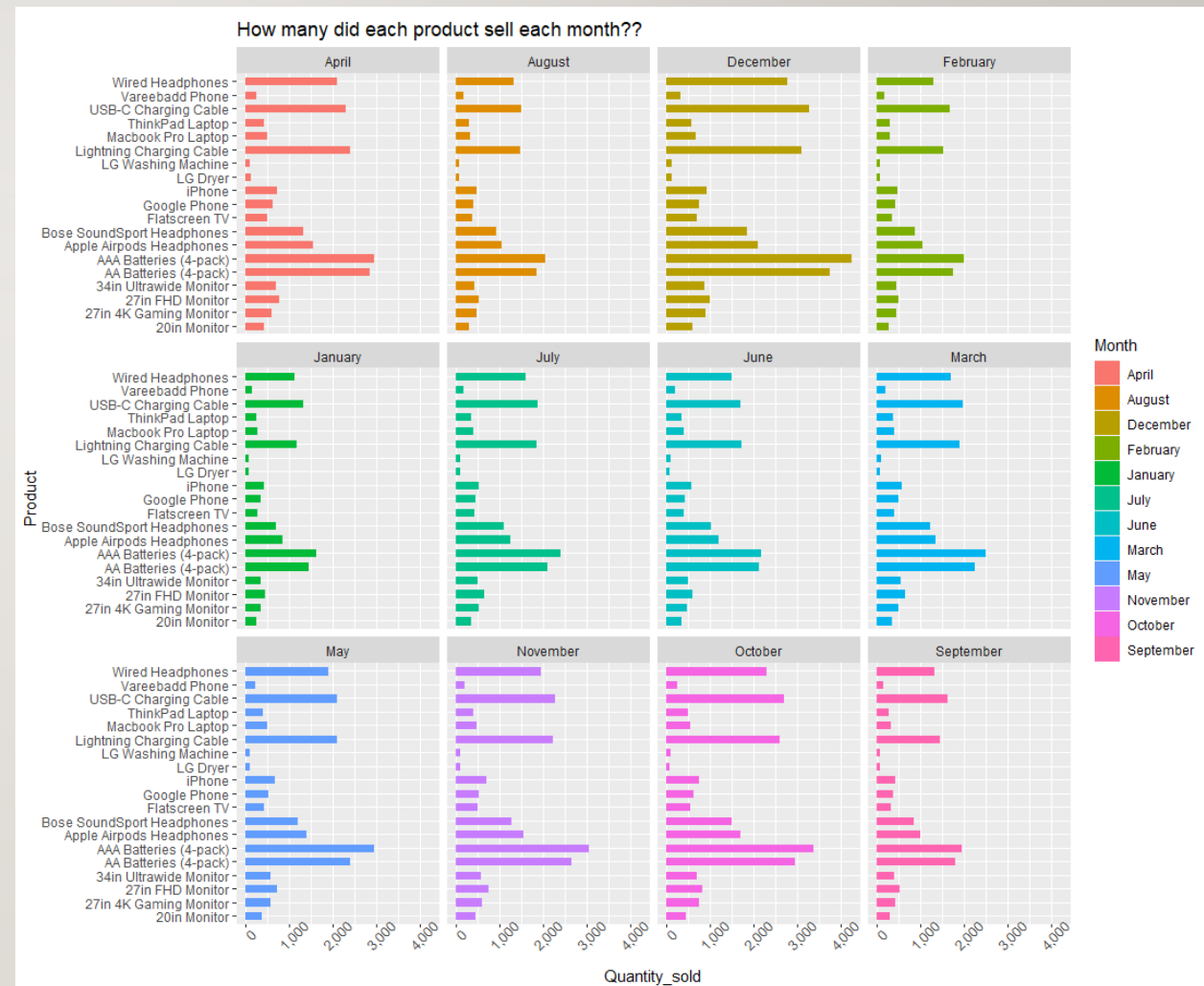
```
#How many did each product sell each month?
```

```
By_Month <- Clean_Sales_2019 %>%  
  group_by(Month, Product) %>%  
  summarise(Quantity_sold = sum(Quantity_Ordered)) %>%  
  arrange(desc(Month))
```

```
ggplot(By_Month, aes(x = Product, y= Quantity_sold, color = Month, fill = Month))+  
  geom_col(width=0.5, position = position_dodge(width=0.5))+  
  theme(axis.text.x = element_text(angle = 45))+  
  labs(title = "How many did each product sell each month??")+  
  scale_y_continuous(labels = comma)+  
  facet_wrap(~Month)+  
  rotate()
```

HOW MANY ORDERS WERE FOR EACH PRODUCT FOR ALL THE MONTHS?

- Consistently across all the months, we can see the items with lowest price were mostly sold.
- AAA Batteries (4-pack), Lightning Charging Cable, USB-C Charging Cable, and Wired Headphones are mostly ordered due to the relatively low price.



DATA VISUALIZATION

- The data was grouped by “Product” and “Month” then summarized by the sum of Sales.
- ggplot function with facet_wrap was used to plot the graph.

```
#How much did each product sell each month?
```

```
By_Sales <- clean_sales_2019 %>%  
  group_by(Month, Product) %>%  
  summarise(Sales = sum(Sales)) %>%  
  arrange(desc(Month))
```

```
ggplot(By_Sales, aes(x = Product, y= Sales, color = Month, fill = Month))+  
  geom_col(width=0.5, position = position_dodge(width=0.5))+  
  theme(axis.text.x = element_text(angle = 45))+  
  labs(title = "How much did each product sell each month??")+  
  scale_y_continuous(labels = comma)+  
  facet_wrap(~Month)+  
  rotate()
```

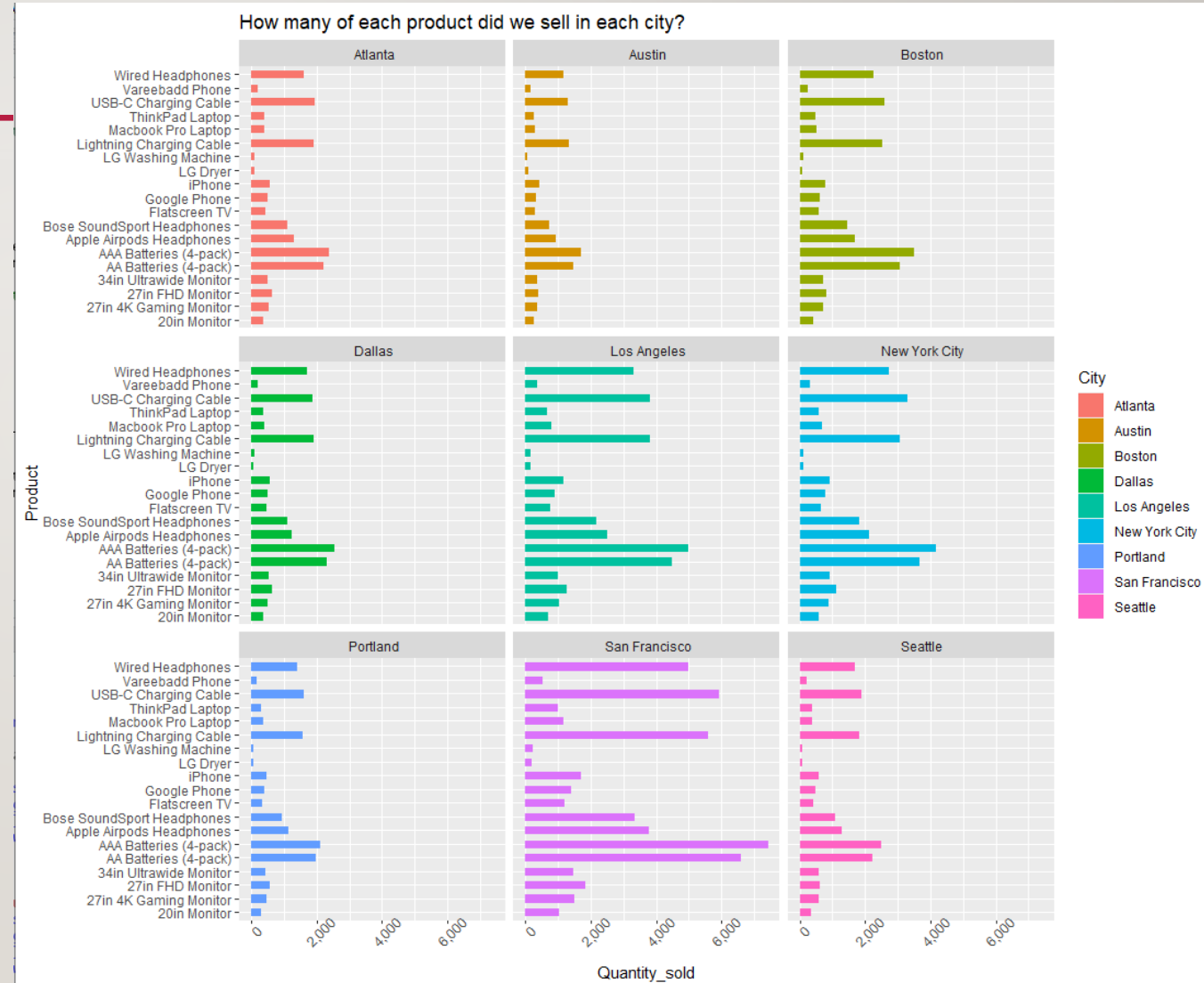

DATA VISUALIZATION

- The data was grouped by “City” and “Product” then summarized by the sum of Quantity_Ordered.
- ggplot function with facet_wrap was used to plot the graph.

```
#How many of each product did we sell in each city?
By_City <- Clean_Sales_2019 %>%
  group_by(City, Product) %>%
  summarise(Quantity_sold = sum(Quantity_Ordered)) %>%
  arrange(desc(City))

ggplot(By_City, aes(x = Product, y= Quantity_sold, color = City, fill = City))+
  geom_col(width=0.5, position = position_dodge(width=0.5))+
  theme(axis.text.x = element_text(angle = 45))+
  labs(title = "How many of each product did we sell in each city?")+
  scale_y_continuous(labels = comma)+
  facet_wrap(~City)+
  rotate()
```

- The products with the lowest price has been the mostly ordered item across all cities.



RECOMMENDATIONS AND CONCLUSION.

From the potential insights we checked using the data provided, we would recommend the following.

- To increase sales, advertisement of least selling products should be targeted between the rush hours (9:00 – 22:00).
- To get more sales, products such as LG Washing Machine, LG Dryer, Flatscreen TV, and ThinkPad Laptop etc., should be discounted as these items had lower orders.
- Promos can be targeted to cities with lower sales such as Austin.
- Promos can as well be introduced in January to increase sales turnover.
- Advertisements (such as black Friday) can also be introduced in December as we can achieve more sales.

LIMITATIONS OF THE ANALYSIS.

Due to certain variables not provided by the dataset, we could not do more in-depth analysis to determine the following;

- Highest revenue generating product based on the cost price and selling price.
- The relationship between the City with the most sales and products ordered.
- The data provided was only for a year, therefore we could not compare the sales generated by year.

PROFESSIONAL ISSUES THAT CAN AFFECT ONLINE STORES.

Cybersecurity is of paramount importance for online retailers because of the sensitive nature of customer information. A look at some of the most pressing problems

- Cybersecurity
- Competition
- Order fulfillment
- Customer experience



LEGAL ISSUES THAT CAN AFFECT ONLINE STORES.

Doing business on the web is rapidly becoming the norm. More than half of all businesses now exist only in the digital realm, with e-commerce representing the majority of their revenue. However, just because something can be done online, that doesn't imply you can avoid legal implications.

- Incorporation Problem
- Trademark Security Problem
- Copyright Protection Issue

ETHICAL ISSUES THAT CAN AFFECT ONLINE STORES.

Knowing the difference between good and wrong only scratches the surface of ethics. It all comes down to the beliefs and practices we uphold.

- Unreliable Customer Service.
- Counterfeit Products.
- Online Piracy.



THANK YOU.

