Data Translation Challenge

Tai Nguyen

```
library(ggpubr)
```

```
Loading required package: ggplot2
```

```
library(ggplot2)
library(ggforce)
library(ggalt)
```

```
Registered S3 methods overwritten by 'ggalt':
  method                  from
  grid.draw.absoluteGrob  ggplot2
  grobHeight.absoluteGrob ggplot2
  grobWidth.absoluteGrob  ggplot2
  grobX.absoluteGrob      ggplot2
  grobY.absoluteGrob      ggplot2
```

```
library(dplyr)
```

```

Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ lubridate 1.9.2     ✔ tibble    3.2.1
✔ purrr     1.0.1     ✔ tidyr     1.3.0
✔ readr     2.1.4
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(rio)
library(scales)
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

1. Insert data

```
setwd("D:/Seattle University/Spring 2023 (03.2023 - 06.2023)/BUAN 4220/R_Assignment/")
load("sales_data.Rdata")
zip_info <- read_csv("zip_info.csv")
```

```
Rows: 10 Columns: 13
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (13): ZIP, TotalPopulation, MedianHHIncome, PCIncome, MedianAge, Race_Wh...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2. Exploring data

```
sales$ZIP <- as.numeric(sales$ZIP)
sales$Quantity <- as.numeric(sales$Quantity)
sales$PriceEach <- as.numeric(sales$PriceEach)
product <- unique(sales$Product)
 print(product)
```

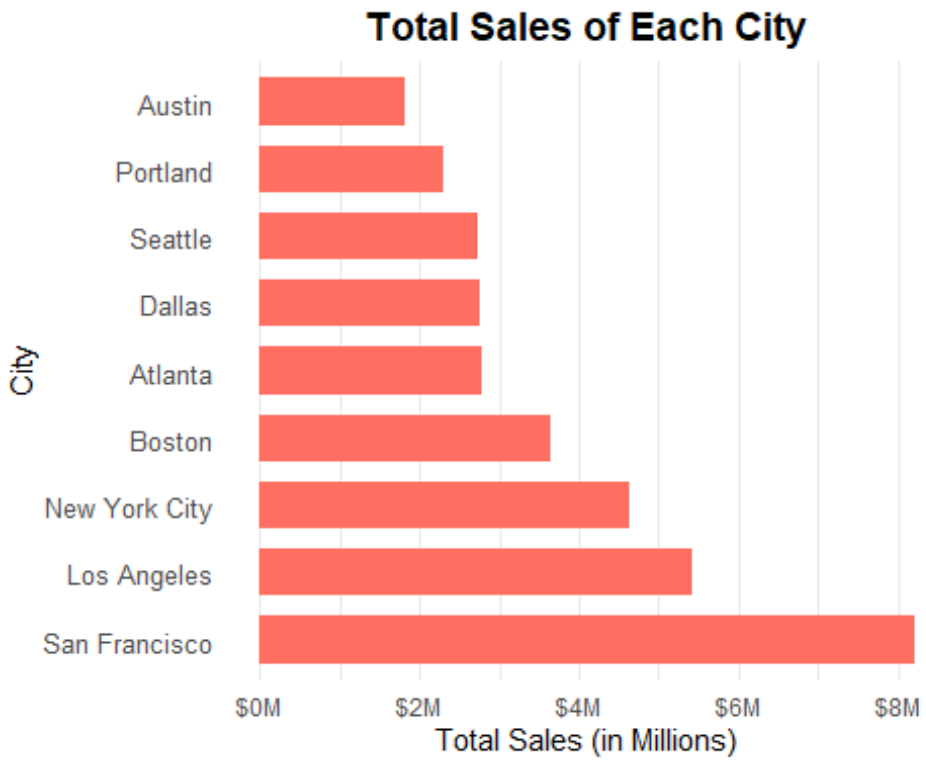
```
 [1] "USB-C Charging Cable"       "Bose SoundSport Headphones"
 [3] "Google Phone"               "Wired Headphones"
 [5] "Macbook Pro Laptop"         "Lightning Charging Cable"
 [7] "27in 4K Gaming Monitor"     "AA Batteries (4-pack)"
 [9] "Apple Airpods Headphones"   "AAA Batteries (4-pack)"
[11] "iPhone"                     "Flatscreen TV"
[13] "27in FHD Monitor"           "20in Monitor"
[15] "LG Dryer"                   "ThinkPad Laptop"
[17] "Vareebadd Phone"            "LG Washing Machine"
[19] "34in Ultrawide Monitor"
```

3. Graph 1

Purpose:

I want to compare the total sales between all the city in order to show what is the most sales city and what is the least. From this point, I can make a story whether about Marketing Plan or Average Income to be a facts for the difference between these cities. From this point, I am as Business Analytics, can either find out the problems and solve it.

```
library(ggpubr)
library(ggplot2)
library(ggforce)
library(ggalt)
library(dplyr)
library(tidyverse)
library(rio)
library(scales)

sales %>%
  group_by(City) %>%
  summarize(SumOrder = sum(PriceEach)) %>%
  ggplot(aes(x = SumOrder, y = reorder(City, -SumOrder))) +
  geom_col(fill = "#FF6F61", width = 0.7) +  # Customizing bar color and width
  scale_x_continuous(labels = dollar_format(scale = 1/1000000, suffix = 'M')) +
  labs(x = 'Total Sales (in Millions)', y = 'City', title = 'Total Sales of Each City') +
  theme_minimal() +  # Using a minimal theme
  theme(axis.text.y = element_text(size = 10),  # Customizing font size for y-axis text
        plot.title = element_text(face = "bold", size = 14, hjust = 0.5),  # Customizing title font
        panel.grid.major.y = element_blank())  # Removing vertical grid lines
```



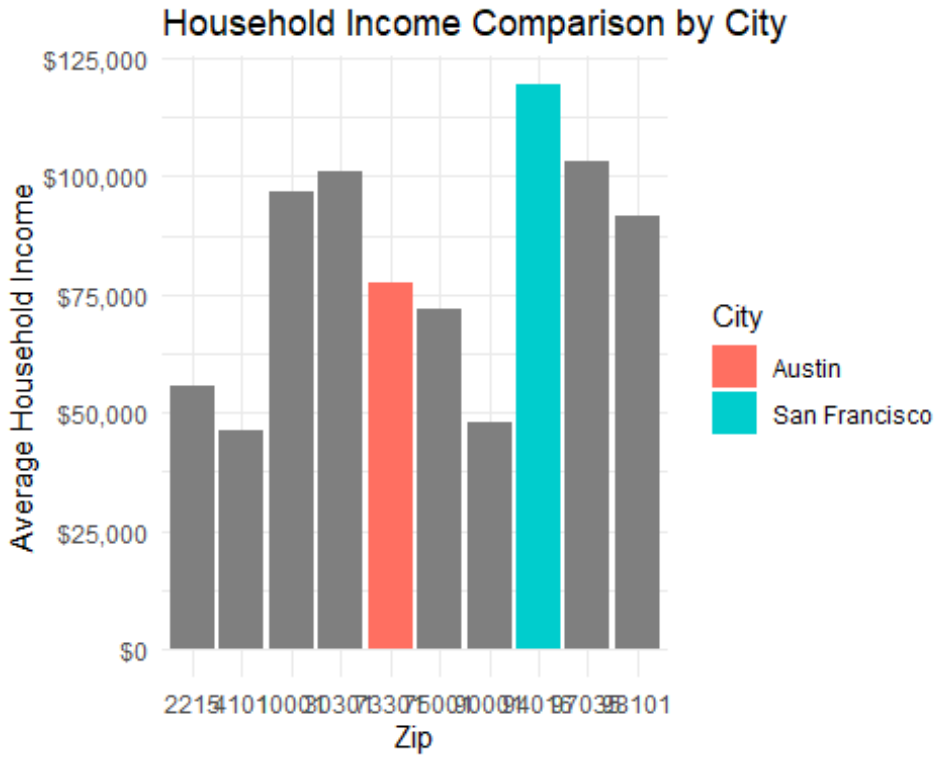
4. Graph 2:

Purpose:

From the previous graph, I already showed the total sales in every city and it is clear that Austin is the city that have the least sales and San Francisco is the top 1 city that have the highest sales. So, I want to dig deeper to see whether the Average Household Income between these 2 cities has a big gap or not.

```
library(ggpubr)
library(ggplot2)
library(ggforce)
library(ggalt)
library(dplyr)
library(tidyverse)
library(rio)
library(scales)
new_data <- merge(sales, zip_info, by = "ZIP")

graph4 <- new_data %>%
  select(ZIP, Product, PriceEach, City, MedianHHIncome) %>%
  group_by(ZIP, Product, City, MedianHHIncome) %>%
  summarise(TotalPrice = sum(PriceEach)) %>%
  arrange(MedianHHIncome) %>%
```

```
  mutate(ZIP = factor(ZIP),
         MedianHHIncome = as.numeric(as.character(MedianHHIncome)))

`summarise()` has grouped output by 'ZIP', 'Product', 'City'. You can override
using the `.groups` argument.

ggplot(graph4, aes(x = ZIP, y = MedianHHIncome, fill = City)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("Austin" = "#FF6F61", "San Francisco" = "cyan3", other_value = "gray"))
+
  scale_y_continuous(labels = dollar_format()) +
  labs(x = "Zip", y = "Average Household Income", title = "Household Income Comparison by City") +
  theme_minimal()
```



From the graph, I can tell that the Average Household Income could be a problem to cause the big gap between these cities's sales - almost double $75,000 when comparing with $120,000

5. Graph 3

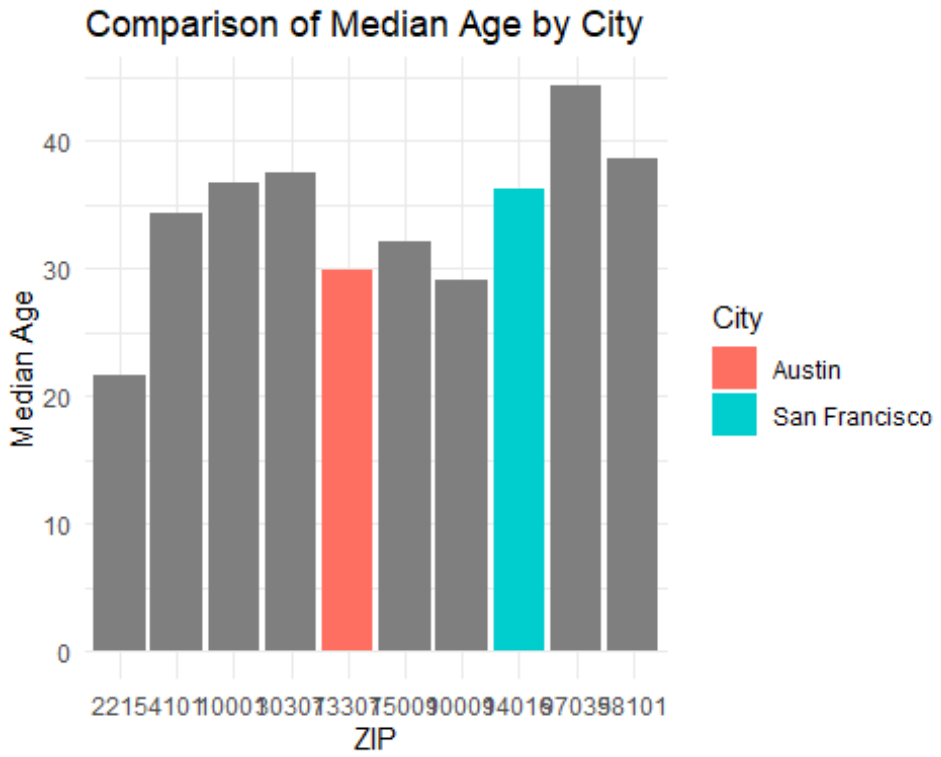
Purpose:

In this graph, I want to see if the age between Austin and San Francisco is another factor for the big gap of sales between these 2 cities.

```
graph3 <- new_data %>%
  select(ZIP, Product, PriceEach, City, MedianHHIncome, MedianAge) %>%
  group_by(ZIP, Product, City, MedianHHIncome, MedianAge) %>%
  summarise(TotalPrice = sum(PriceEach)) %>%
  arrange(MedianHHIncome) %>%
  mutate(ZIP = factor(ZIP),
         MedianHHIncome = as.numeric(as.character(MedianHHIncome)))
```

```
`summarise()` has grouped output by 'ZIP', 'Product', 'City', 'MedianHHIncome'.
You can override using the `.groups` argument.
```

```
ggplot(graph3, aes(x = ZIP, y = MedianAge, fill = City)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("Austin" = "#FF6F61", "San Francisco" = "cyan3", other_value = "gray"))
+
  labs(x = "ZIP", y = "Median Age", title = "Comparison of Median Age by City") +
  theme_minimal()
```



However, as we can see, the median age between these 2 cities are not a big difference. But, I can tell that when people get older, they will have more things to buy for their life and family if they have wife and children. So, I can consider age can be a slightly problem in this situation.

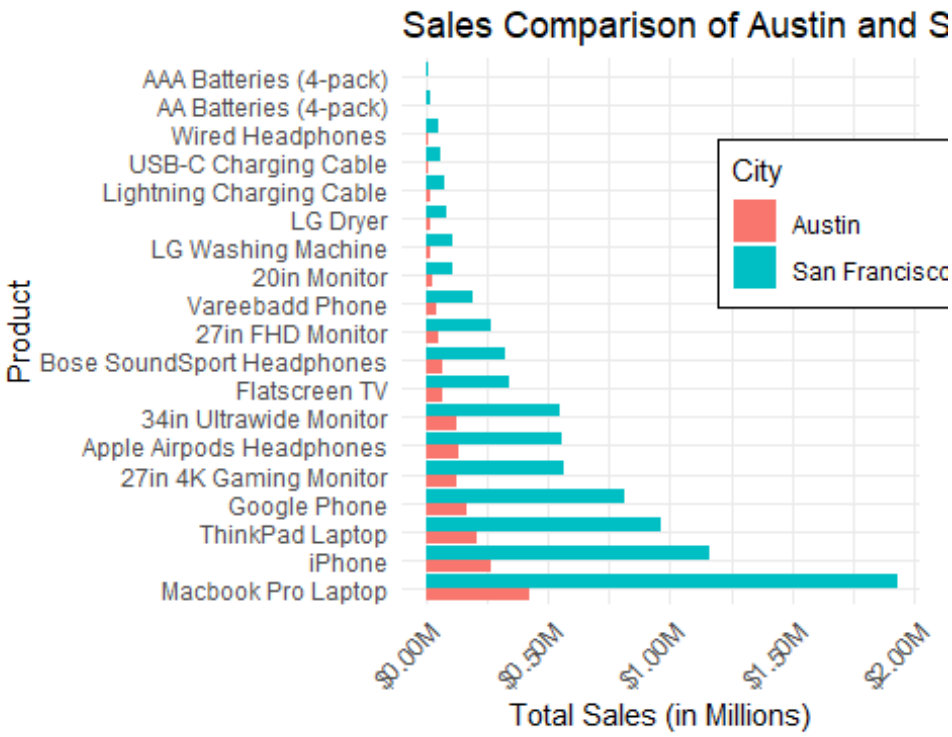
So, after graph 2 and 3, I can tell that the problem that cause the big gap sales between these 2 cities are the income. So, if Amazon want to increase the sales in Austin, their Marketing team should have more promotion in this city.

6. Graph 4

Purpose:

In the previous graph, I know that the highest sale city is San Francisco, and the least sale city is Austin. So, in this graph, I want to breakdown the most sales item in each city to see if they have the same trend or same item that have the most sales. From that point, I can show the Marketing team whether apply the Marketing plan on San Francisco city to Austin to increase the sales for the next year.

```
library(ggplot2)
library(scales)

# Group by product and city, calculate the total sales for each combination
product_sales <- sales %>%
  group_by(Product, City) %>%
  summarize(SumOrder = sum(PriceEach)) %>%
  ungroup()

`summarise()` has grouped output by 'Product'. You can override using the
`.groups` argument.

# Get the city with the highest sales for each product
top_city <- product_sales %>%
  group_by(Product) %>%
  filter(SumOrder == max(SumOrder)) %>%
  ungroup()

# Get the city with the least sales for each product
bottom_city <- product_sales %>%
  group_by(Product) %>%
  filter(SumOrder == min(SumOrder)) %>%
  ungroup()

# Combine the top and bottom cities data
combined_data <- bind_rows(bottom_city %>% mutate(City = "Austin"),
                           top_city %>% mutate(City = "San Francisco"))

# Plot the data
ggplot(combined_data, aes(x = SumOrder, y = reorder(Product, -SumOrder), fill = City)) +
  geom_col(position = "dodge") +
  scale_x_continuous(labels = dollar_format(scale = 1/1000000, suffix = 'M')) +
  labs(x = 'Total Sales (in Millions)', y = 'Product', title = 'Sales Comparison of Austin and San Francisco')
+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = c(0.85, 0.7),  # Adjust the legend position inside the graph
        legend.background = element_rect(color = "black"),  # Add a border to the legend
        plot.margin = margin(10, 10, 10, 10))
```



As we can see, the trend of every items in these cities are mostly the same. So, I believe that the Marketing team in this situation applied the same plan on the same products in each city.

7. Graph 5:

Purpose:

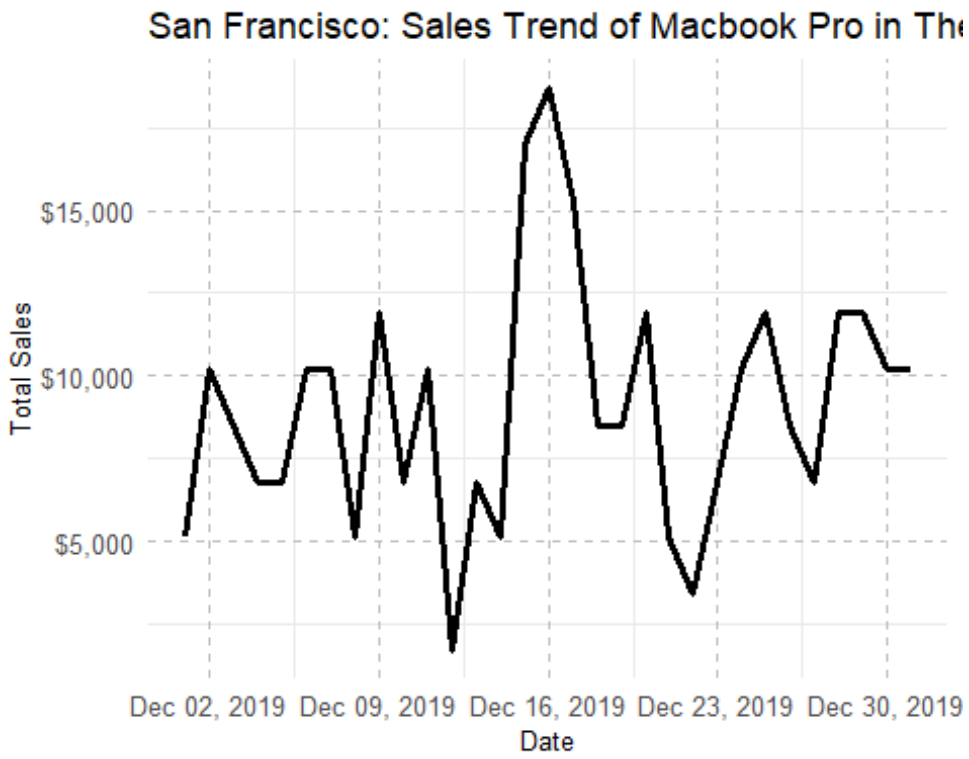
In this graph, I want to see what is the highest volume month that the best seller item into. From the graph, I know that the highest volume month is December for Macbook Pro items and the day that have the highest volume is on December 16, 2019. From this, I can assume that during this time, Macbook Pro should have very good promotion, or the Marketing team have the best decision on promoting the best sellers in the right time, especially near the Christmas and New Year Holiday

```
library(ggplot2)
library(scales)

# Group by product, city, and date, calculate the total sales for each combination
product_sales2 <- sales %>%
  group_by(Product, City, Date) %>%
  summarize(SumOrder = sum(PriceEach)) %>%
  ungroup()
```

```
`summarise()` has grouped output by 'Product', 'City'. You can override using
the `.groups` argument.
```

```
# Filter the data for the most sold item in the city
most_sold_item <- top_city %>%
  filter(City == "San Francisco") %>%
  slice_max(order_by = SumOrder, n = 1)

# Filter the data for the most sold item in the city and the highest volume month
most_sold_item_data <- product_sales2 %>%
  filter(Product == most_sold_item$Product, City == most_sold_item$City) %>%
  mutate(YearMonth = format(Date, "%Y-%m")) %>%
  group_by(YearMonth) %>%
  summarize(SumOrder = sum(SumOrder)) %>%
  ungroup() %>%
  arrange(desc(SumOrder)) %>%
  slice(1)

# Filter the original data to include only the highest volume month
most_sold_item_data <- product_sales2 %>%
  filter(Product == most_sold_item$Product, City == most_sold_item$City, format(Date, "%Y-%m") ==
most_sold_item_data$YearMonth)

# Convert the 'Date' column to a date format
most_sold_item_data$Date <- as.Date(most_sold_item_data$Date, format = "%Y-%b-%d")

# Plot the data with a clearer appearance
ggplot(most_sold_item_data, aes(x = Date, y = SumOrder, fill = City)) +
  geom_line(size = 1.2) +
  scale_x_date(date_labels = "%b %d, %Y", date_breaks = "1 week") +
  scale_y_continuous(labels = dollar_format()) +
  labs(x = 'Date', y = 'Total Sales', title = "San Francisco: Sales Trend of Macbook Pro in The Highest
Volume Month") +
  theme_minimal() +
  theme(axis.title = element_text(size = 10),
        axis.text = element_text(size = 10),
        panel.grid.major = element_line(color = "gray", linetype = "dashed"))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```


San Francisco: Sales Trend of Macbook Pro in The
$15,000
$10,000
$5,000
Total Sales
Dec 02, 2019
Dec 09, 2019
Dec 16, 2019
Dec 23, 2019
Dec 30, 2019
Date