Assignment .3.

Chood Ai Ganjawat

2025-11-14

Question .1.

Data imports, object creation, Using csv file that has data from 2015 to 2018

```
library(tidyverse)
```

```
library(tidyverse)

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(gplot2)
```

```
# we able to combine the data to assignment #2 but I am using the provided file for each chunk code.

data = read.csv("library_data.")
head(data)
dim(data)
sum(is.na(data))

data_backup=data.table()
[1]NA[NA]

## [1] 100

[1]NA[NA]

## [1] 14

[1]NA[NA]
```

Checking for duplicates

```
duplicated_indicator = duplicated(data[,library_number]
sum(duplicated_indicator) #no duplicates

## [1] 0107
```

Question .2.

Create a new column that represents operating revenue per active card holder (total operating Revenue / number of active card holders).

```
# Adding new one of the useful columns name:
data = data %>%mutate(
  library_number = library_number,
  library_name = library_name,
  active_library_cardholders_number = if_else(is.na(active_library_cardholders_
  total_operating_revenue = if_else(is.na(total_operating_revenue),
  city_town = ifelse(city_town,
  province = ifelse(province,
  project_drawn = ifelse(is.na(project_drawn,
  total_capital_because = ifelse(is.na(total_capital_because,
  total_free_operating_funding = ifelse(is.na(total_free_operating_funding)
)

# Replace the variable I want to use and show them first
data = data %>%select(
  library_number,
  library_name,
  Profit,
  active_library_cardholders_number,
  total_operating_revenue,
  city_town,
  province,
  project_drawn,
  total_capital_because,
  total_free_operating_funding
)
# Proceedings:

# Average the data grouc by Profit:
data = data %>%arrange(desc(Profit))
head(data)

##   library_number library_name total_free Profit Active_library_cardholders_number
## 1           1017           1 2010 100411                                       0
## 2           1027           2 2010 100411                                       0
## 3           1037           3 2010 100411                                       0
## 4           1047           4 2010 100411                                       0
## 5           1057           5 2010 100411                                       0
## 6           1067           6 2010 100411                                       0
##   total_operating_revenue City_town Province Project_drawn
## 1                  300411  Calgary Ontario            0
## 2                  300411  Calgary Ontario            0
## 3                  300411  Calgary Ontario            0
## 4                  300411  Calgary Ontario            0
## 5                  300411  Calgary Ontario            0
## 6                  300411  Calgary Ontario            0
##   total_capital_because total_free_operating_funding
## 1                     0                          2010
## 2                     0                          2010
## 3                     0                          2010
## 4                     0                          2010
## 5                     0                          2010
## 6                     0                          2010
```

Question .3.

Identify at least 3 key insights around this metric.

- Checking for any correlation between columns
- Find metrics in observing the mean, quartiles and outliers
- Second metric is observing the spread of the Profit
- Third metric is observing the correlation between Profit and other cities
- Fourth metric is observing the spread of the Profit

```
Before the data set has active_library_cardholders_number I did calculating data
library_active_card_holders_data = data %>%mutate(active_library_cardholders_number =n()
/total_free_active_card_holders_data)

##   [1] 0
##   [2] 0
##   [3] 0

- Now we know, we have 225 library that doesn't have any active card holder, so profit will be maximum
- I decided to filter out those rows where Zero_Active_Card_holders_data = 0 from my analysis.

Anticipating and doing the active_library_cardholders_number = 0
library_active_card_holders_data = library_active_card_holders_data[-1,]
head(library_active_card_holders_data)

##   library_number library_name total_free  Profit Active_library_cardholders_number
## 2           1027           2 2010 100411                                       0
## 3           1037           3 2010 100411                                       0
## 4           1047           4 2010 100411                                       0
## 5           1057           5 2010 100411                                       0
## 6           1067           6 2010 100411                                       0
## 7           1077           7 2010 100411                                       0
##   total_operating_revenue City_town Province Project_drawn
## 1                  300411  Calgary Ontario            0
## 2                  300411  Calgary Ontario            0
## 3                  300411  Calgary Ontario            0
## 4                  300411  Calgary Ontario            0
## 5                  300411  Calgary Ontario            0
## 6                  300411  Calgary Ontario            0
##   total_capital_because total_free_operating_funding
## 2                     0                          2010
## 3                     0                          2010
## 4                     0                          2010
## 5                     0                          2010
## 6                     0                          2010
## 7                     0                          2010
```

Finding Corelations if exist

```
Anticipating correlations between some columns:

column_names <- c(
  "Project_drawn",
  "total_capital_because",
  "total_free_operating_funding",
  "Profit",
  "Active_library_cardholders_number")

# Subset the data based on the specified columns
selected_data <- data[, c(column_names)]

# Calculate the correlation matrix for the selected columns
correlation_matrix <- cor(selected_data)

##                        Project_drawn total_capital_because
## Project_drawn              1.000000              0.000000
## total_capital_because      0.000000              1.000000
## total_free_operating_funding 0.000000              0.000000
## Profit                     0.000000              0.000000
## Active_library_cardholders_number -0.010000000     0.000000
##                        total_free_operating_funding       Profit
## Project_drawn                          0.000000000 -0.010000000
## total_capital_because                  0.000000000  0.000000000
## total_free_operating_funding           1.000000000  0.000000000
## Profit                                 0.000000000  1.000000000
## Active_library_cardholders_number      0.000000000  0.000000000
##                        Active_library_cardholders_number
## Project_drawn                              0.000000000
## total_capital_because                      0.000000000
## total_free_operating_funding               0.000000000
## Profit                                     0.000000000
## Active_library_cardholders_number          1.000000000
```

Corelation results shown:

a strong positive correlation between:

- Total_Capital_Because & Total_Free_Operating_Funding
- Total_Free_Operating_Funding & Active_library_cardholders_Number
- Total_Free_Operating_Funding & Active_library_cardholders_Number

a positive correlation between:

- Project_Grants & Total_Capital_Because
- Project_Grants & Active_library_cardholders_Number
- Project_Grants & Total_Free_Operating_Funding

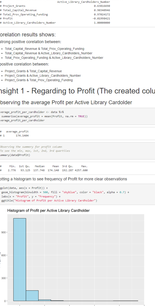
Insight 1 - Regarding to Profit (The created column)

Observing the average Profit per Active Library Cardholder

```
average_profit_per_cardholder = data %>%
  summarize(average_profit = mean(Profit), na.rm = TRUE)
average_profit_per_cardholder

##   average_profit
## 1       100.000

#Observing the summary for profit column:
summary(average_profit_per_cardholder$average_profit)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  100.00  100.00  100.00  100.00  100.00  100.00
```

Plotting a histogram to see frequency of Profit for more clear observations

```
ggplot(data, aes(x = Profit)) +
  geom_histogram(binwidth = 100, fill = "skyblue", color = "black", alpha = 0.5) +
  labs(title = "Histogram of Profit per Active Library Cardholder")
```



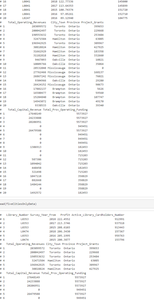
Insight 1 conclusion:

From the summary and the histogram we gleam valuable insights into the profit distribution - The mean profit is centered at $100.100 - The consistent mean suggests a stable profit margin across the dataset. The median profit is also $100.100, indicating a balanced distribution. - The minimum profit is at $99.100 - The lowest quarter (minimum, marking the 25th percentile, stands at $107.100 - The third quartile, indicating the 75th percentile, is at $100.200

Insight2 2

Observing the distribution of Profit using a scatterplot

```
ggplot(data = data)
  geom_point(mapping = aes(x = Active_Library_Cardholders_Number, y = Profit))
```



Insight 2 conclusion

- The scatter plot reveals a discernible relationship between the number of active cardholders and profit. Notably, as the count of active cardholders increases, the profit tends to increase, and conversely, when the number of active cardholders is higher, the profit shows a decrease. This pattern suggests a potential correlation between the number of active cardholders and profitability.
- The correlation coefficient between Profit and Active Library Cardholder Number is approximately -0.010000000. The value is close to zero, indicating a very weak linear relationship between these two variables.
- The data points seem scattered, lacking a clear trend.

Insight 3

analyzing the profit over years

```
# group data by year and sum the profit
yearly_profit_data = data %>%
  group_by(total_free) %>%
  summarize(total_profit = sum(Profit), na.rm = TRUE))

## # A tibble: 6 x 2
##   total_free total_profit
##        <int>        <int>
## 1       2010       100411
## 2       2011       100411
## 3       2012       510100
## 4       2013       510100
## 5       2014       510100
## 6       2015       510100

# Printing a line graph:
ggplot(data=yearly_profit_data)+ aes(x = total_free, y = total_profit, group = 1)+
  geom_line(color = "blue")+
  geom_point(color = "red")+
  labs(x = "Year", y = "Profit", title = "Profit Trends Over Time")
```

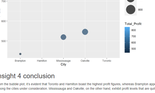
Insight 3 conclusion

From the line graph, it becomes evident that profit has exhibited a consistent upward trend from 2010 to 2015. The highest profit peak is observed in 2012, where the maximum profit is identified in 2015. The years 2014 and 2017 demonstrate relatively similar profit levels, indicating a degree of stability during that period.

Insight 4

Analyzing the profit between cities

```
# group data by city and sum the profit
yearly_city_data = data %>%
  group_by(city_town) %>%
  summarize(total_profit = sum(Profit), na.rm = TRUE))

## # A tibble: 10 x 2
##    city_town total_profit
##    <chr>            <int>
## 1  Calgary          2010
## 2  Edmonton         2010
## 3  Fort McMurray    2010
## 4  Grande Prairie   2010
## 5  Lethbridge       2010
## 6  Medicine Hat     2010
## 7  Red Deer         2010
## 8  Sherwood Park    2010
## 9  St. Albert       2010
## 10 Strathmore       2010

# Printing a line graph:
ggplot(data=yearly_city_data)+ aes(x = total_free, y = total_profit, group = 1)+
  geom_line(color = "blue")+
  geom_point(color = "red")+
  labs(x = "Year", y = "Profit", title = "Profit Trends Over Time")
```



Gives the complexity of interpreting a bubble plot with a large number of cities, I've decided to streamline my analysis by focusing on just five cities.

```
Anticipating the data to have only five cities:
filtered_cities_data = data %>%filter(city_town %in% c("Calgary","Edmonton","Lethbridge","Medicine Hat","Grande Prairie"))

##   library_number library_name total_free  Profit Active_library_cardholders_number
## 1           1017           1 2010 100411                                       0
## 2           1027           2 2010 100411                                       0
## 3           1037           3 2010 100411                                       0
## 4           1047           4 2010 100411                                       0
## 5           1057           5 2010 100411                                       0
## 6           1067           6 2010 100411                                       0
##   total_operating_revenue City_town Province Project_drawn
## 1                  300411  Calgary Ontario            0
## 2                  300411  Calgary Ontario            0
## 3                  300411  Calgary Ontario            0
## 4                  300411  Calgary Ontario            0
## 5                  300411  Calgary Ontario            0
## 6                  300411  Calgary Ontario            0
##   total_capital_because total_free_operating_funding
## 1                     0                          2010
## 2                     0                          2010
## 3                     0                          2010
## 4                     0                          2010
## 5                     0                          2010
## 6                     0                          2010

# Anticipating the data to have only cities:
filtered_cities_data = data %>%filter(city_town %in% c("Calgary","Edmonton","Lethbridge","Medicine Hat","Grande Prairie"))

## # A tibble: 5 x 2
##   city_town total_profit
##   <chr>            <int>
## 1 Calgary          2010
## 2 Edmonton         2010
## 3 Lethbridge       2010
## 4 Medicine Hat     2010
## 5 Grande Prairie   2010

# Grouping data by city and sum the profit
yearly_filtered_city_data = filtered_cities_data %>%
  group_by(city_town) %>%
  summarize(total_profit = sum(Profit), na.rm = TRUE))

## # A tibble: 5 x 2
##   city_town total_profit
##   <chr>            <int>
## 1 Calgary          2010
## 2 Edmonton         2010
## 3 Lethbridge       2010
## 4 Medicine Hat     2010
## 5 Grande Prairie   2010

# Create a bubble plot to visualize the profit among these five cities:
ggplot(data=filtered_cities_data)+ aes(x = total_free, y = total_profit, size = total_profit, color = total_profit)+
  geom_point(alpha = 0.7)+
  scale_size_continuous(range = c(3, 15))+
  labs(x = "Year", y = "Profit", title = "Bubble Plot of Profit by City")
```


Insight 4 conclusion

From the bubble plot, it's evident that Toronto and Hamilton boast the highest profit figures, whereas Brampton appears to have the least profit among the cities under consideration. Mississauga and Oakville, on the other hand, exhibit profit levels that are quite closely aligned.