# Assignment 4

Chood Al Qeshawi

2023-11-21

## Purpose of the Analysis

- To enhance the outcome of Ontario libraries in the 21st century, data-driven insights aim to extract insights about the metric of Revenue per cardholder.
- By identifying the relations between the data set's columns, and finding the patterns, the analysis will inform strategic decisions, and help with ensuring an efficient library system in Ontario.
- In this analysis I will be looking into data from 2015 to 2019.
- I will be comparing the Revenue per active library card holder among those years.
- I will also compare the Revenue per active library card holder among 5 cities (Toronto, Hamilton, Mississauga, Oakvill and Brampton)
- I will be looking at columns that directly affects the Revenue per card holder.
- Into, I included a Mahalanobis filter file, that has code to check for co-relation between every numeric column in the dataset.

## Step 1

- Data Input

```
rm(list=ls())

## Loading required package: tabularis

## warning: package 'tabularis' was built under R version 4.2.2

##
## Attaching packages: 'tabularis'

## The following object is masked from 'package:dplyr':
##
##     group_rows

library(tabularis)
data.set <- read.csv("library_data.csv")

data <- original_data
data <- data.set
dim(data)
head(data,1)
dim(data.set)
print(paste("Data Set", ncol(data), "Rows"))

## [1] "Data Set 100 Rows"

print(paste("Data Set", ncol(data), "Columns"))

## [1] "Data Set 100 Columns"

dim(data)
```

## Step 2 - Create a column to represent operating revenue per active card holder.

```
#Creating one of the required columns names
data <- data %>% rename(
  Library_Number = Library_Number,
  Survey_Year_From = Survey_Year_From,
  Revenue_per_cardholder = Revenue_per_cardholder,
  Total_Operating_Revenues = Total_Operating_Revenues,
  City_Town = CD-CSD City_Town,
  Province = Province,
  Project_Grants = Grants_for_Project_Projects,
  Total_Capital_Revenue = Total_Capital_Revenue,
  Total_Prov_Operating_Funding = Grants_Provincial_Operating_Funding,
  Total_spends_on_computers = Total_Expenditures_for_electronic_equipment_and_related_expenses_e.g.computers,
  Total_spends_on_utilities = Total_Expenditures_for_utilities_(costs_related_to_Library_Facility_operation_e.g._insurance_,_rent_mortgage_payments_and_utilities),
  Self_generated_revenues = Self_Generated_Revenue_e.g._Fines_Fees_sales_Fundraising_room_rentals_cafe_revenue_)
)
head(data)

#Adding the new column Revenue_per_cardholder representing revenue per active card holder
#Dividing Total_Operating_Revenues by the number Active_Cardholders_Number
#Using gsub function to remove commas in the data set
#Using as.numeric to convert the data into numeric
data <- data %>%
  Total_Operating_Revenues = as.numeric(gsub(",", "", Total_Operating_Revenues)),
  Active_Library_Cardholders_Number = as.numeric(gsub(",", "", Active_Library_Cardholders_Number),
  Revenue_per_cardholder = Total_Operating_Revenues / Active_Library_Cardholders_Number,
  Project_Grants = as.numeric(gsub(",", "", Project_Grants)),
  Total_Capital_Revenue = as.numeric(gsub(",", "", Total_Capital_Revenue)),
  Total_Prov_Operating_Funding = as.numeric(gsub(",", "", Total_Prov_Operating_Funding)),
  Total_spends_on_computers = as.numeric(gsub(",", "", Total_spends_on_computers)),
  Total_spends_on_utilities = as.numeric(gsub(",", "", Total_spends_on_utilities)),
  Self_generated_revenues = as.numeric(gsub(",", "", Self_generated_revenues))

#Select the columns I want to use and save them first
data <- data %>% select (
  Library_Number,
  Survey_Year_From,
  Revenue_per_cardholder,
  Active_Library_Cardholders_Number,
  Total_Operating_Revenues,
  City_Town,
  Province,
  Project_Grants,
  Total_Capital_Revenue,
  Total_Prov_Operating_Funding,
  Total_spends_on_computers,
  Total_spends_on_utilities,
  Self_generated_revenues
)
head(data)

#Arranging the data by Revenue_per_cardholder
data <- data %>%arrange(Revenue_per_cardholder)
#Filtering the data to only include Ontario, and Toronto, Hamilton, Mississauga, Oakville, and Brampton
filter(City_Town, Province_Territory_Name)

##   Library_Number Survey_Year_From Revenue_per_cardholder Active_Library_Cardholders_Number Total_Operating_Revenues City_Town Province Project_Grants Total_Capital_Revenue Total_Prov_Operating_Funding Total_spends_on_computers Total_spends_on_utilities Self_generated_revenues
## 1         L0170             2016             3054.012                    0                  3054012 Cobourg  Ontario             0                  0                   19160                     0                     0                     0
## 2         L0170             2017             2994.06                     0                  2994060 Cobourg  Ontario             0                  0                   19160                     0                     0                     0
## 3         L0170             2018             2954.60                     0                  2954600 Cobourg  Ontario             0                  0                   19160                     0                     0                     0
## 4         L0170             2015             2839.78                     0                  2839780 Cobourg  Ontario             0                  0                   19160                     0                     0                     0
## 5         L0852             2015             2549.52                     0                  2549520 Oakville Ontario             0                  0                   23752                     0                     0                     0
## 6         L0852             2016             2549.52                     0                  2549520 Oakville Ontario             0                  0                   23752                     0                     0                     0
```

## Step 3 - Identity at at least 3 key insights around Revenue per cardholder metric.

T-Test is observing the mean, quarters and outliers, and find out the Point Jump. 2- Second is observing the spread of the
Revenue_per_cardholder among cities.
3- For a more concise and accurate results, I will remove all the Active_Library_Cardholders_Number that equals to
0, and all the Active_Library_Cardholders_Number that has less than 500.

```
#Filter the data has the Active_Library_Cardholders_Number > 0 and counting rows
data_active_more_than_zero <- data %>%filter (Active_Library_Cardholders_Number>0)

dim(data_active_more_than_zero)

##  [1] 99 19
##  [2] 270

#Filtering out data has Active_Library_Cardholders_Number < 500
data <- data %>%filter (Active_Library_Cardholders_Number >500)
dim(data)

#Selecting the columns I want to use and save them first
#Arranging the data by Revenue_per_cardholder
data <- data %>%select (Library_Number, Survey_Year_From, Revenue_per_cardholder, Active_Library_Cardholders_Number, Total_Operating_Revenues, City_Town, Province, Project_Grants, Total_Capital_Revenue, Total_Prov_Operating_Funding, Total_spends_on_computers, Total_spends_on_utilities, Self_generated_revenues) %>%arrange(Revenue_per_cardholder)
head(data,1)

##   Library_Number Survey_Year_From Revenue_per_cardholder Active_Library_Cardholders_Number Total_Operating_Revenues City_Town Province Project_Grants Total_Capital_Revenue Total_Prov_Operating_Funding Total_spends_on_computers Total_spends_on_utilities Self_generated_revenues
## 1         L0907             2017             104.3615                  750                 78423 Welland  Ontario        15362                 0                   15062                     0                  1295                  6421
## 2         L0907             2016             102.2108                  737                 75373 Welland  Ontario        18898                 0                   15062                  1355                  1607                 21077
## 3         L0907             2015             100.4615                  750                 75462 Welland  Ontario        17377                 0                   15062                     0                  1295                  6421
## 4         L0907             2018             101.4044                  742                 75625 Welland  Ontario         2595                 0                   15062                  2856                  1477                  11462
## 5         L0030             2015             444.6614                11367                5054230 Brantford Ontario        45231             50929                 153731                 153447                 233544                 155964
```

## Main observation from the previous table is:

- L0907 library was the top at Revenue per card holder with 104.3615 in 2016
- Finding Correlations among some specific columns that I suspected a potential co-relation.

```
#Finding correlations between some columns

column_name1 <- c(
  "Revenue_per_cardholder",
  "Total_Capital_Revenue",
  "Total_Prov_Operating_Funding",
  "Project_Grants",
  "Active_Library_Cardholders_Number",
  "Total_Operating_Revenues",

#Select the data based on the specified column names
selected_data1 <- data[, column_name1]

#Calculate the correlation matrix for the selected columns
correlation_matrix1 <- cor(selected_data1)
print("Correlation Matrix:")
table_rounded1<-round(correlation_matrix,"Total", full_width = FALSE, position = "center")

##                       Project_Grants Total_Capital_Revenue Total_Prov_Operating_Funding  Revenue_per_cardholder  Active_Library_Cardholders_Number  Total_Operating_Revenues
## Project_Grants             1.000000              0.001970                     0.023147                0.150400                           0.000000                  0.000000
## Total_Capital_Revenue      0.001970              1.000000                     0.404883                0.107182                           0.404883                  0.404882
## Total_Prov_Operating_Funding 0.023147            0.404883                     1.000000                0.380438                           0.970701                  0.993617
## Revenue_per_cardholder     0.150400              0.107182                     0.380438                1.000000                           0.042782                  0.081718
## Active_Library_Cardholders_Number 0.000000       0.404883                     0.970701                0.042782                           1.000000                  0.993566
## Total_Operating_Revenues   0.000000              0.404882                     0.993617                0.081718                           0.993566                  1.000000
```

### Correlation results shows:

- strong positive correlation between
  - Total_Operating_Revenues & Revenue_per_cardholder
  - Total_Operating_Revenues & Total_Prov_Operating_Funding
  - Total_Operating_Revenues & Active_Library_Cardholders_Number
  - Total_Prov_Operating_Funding & Active_Library_Cardholders_Number
  - Total_Capital_Revenue & Active_Library_Cardholders_Number
  - Total_Prov_Operating_Funding & Active_Library_Cardholders_Number
- a positive correlation between
  - Project_Grants & Revenue_per_cardholder
  - Project_Grants & Active_Library_Cardholders_Number
  - Project_Grants & Total_Prov_Operating_Funding
  - Project_Grants & Total_Capital_Revenue

It also shows that most of the selected columns actually affect Revenue_per_cardholder, so I will be testing co-relation with other columns

To investigate if any other column has co-relation with Revenue_per_cardholder, I am testing with the following columns.

```
#Finding correlations between Revenue_per_cardholder and some other columns

column_name2 <- c(
  "Revenue_per_cardholder",
  "Active_Library_Cardholders_Number",
  "Total_Operating_Revenues",
  "Total_spends_on_computers",
  "Total_spends_on_utilities",
  "Self_generated_revenues"

#Select the data based on the specified column names
selected_data2 <- data[, column_name2]

#Calculate the correlation matrix for the selected columns
correlation_matrix2 <- cor(selected_data2)
print("Correlation Matrix:")
table_rounded2<-round(correlation_matrix,"Total", full_width = FALSE, position = "center")

##                       Revenue_per_cardholder Active_Library_Cardholders_Number Total_Operating_Revenues Total_spends_on_computers Total_spends_on_utilities Self_generated_revenues
## Revenue_per_cardholder              1.000000                          0.042782                0.081718                0.078657                0.078657                0.067809
## Active_Library_Cardholders_Number   0.042782                          1.000000                0.993566                0.557000                0.557000                0.574407
## Total_Operating_Revenues            0.081718                          0.993566                1.000000                0.552763                0.552763                0.565000
## Total_spends_on_computers           0.078657                          0.557000                0.552763                1.000000                0.944076                0.945409
## Total_spends_on_utilities           0.078657                          0.557000                0.552763                0.944076                1.000000                0.979710
## Self_generated_revenues             0.067809                          0.574407                0.565000                0.945409                0.979710                1.000000
```

### Observations:

- From the previous table, it shows that Revenue_per_cardholder doesn't have strong co-relation with the selected columns.

### However, there is a weak positive co-relation between:

- Total_spends_on_computers & Revenue_per_cardholder
- Total_spends_on_utilities & Revenue_per_cardholder
- Self_generated_revenues & Revenue_per_cardholder

## Recommendations

- For a better library performance it is advised to 1- increase the Self_generated_revenues by conducting more sales, fundraising, room rentals, and cafe revenue.
2- To increase the Revenue per cardholder, Revenue_per_cardholder can be affected slightly by both Total_spends_on_utilities,
Total_spends_on_computers. so increasing the investment in good utilities and advanced computer system effects with a slight increase with the
Revenue_per_cardholder. so spending more in these areas can improve the libraries performance.

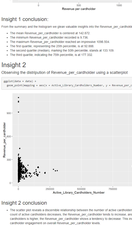
## Insight 1 - Regarding to Revenue_per_cardholder

### Observing the average Revenue_per_cardholder per Active Library Cardolder

```
average_revenue_per_cardholder <- data %>%
  summarise(average_revenue_per_cardholder = mean(Revenue_per_cardholder), na.rm = TRUE)
print(average_revenue_per_cardholder)

##   average_revenue_per_cardholder
## 1                       1.0e+03
## 2                       1.0e+03

#Observing the Insight for Revenue_per_cardholder column
#Filter the data data_size, data_100 quantities
summary(data$Revenue_per_cardholder)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   104.3 1000.0 1000.0 1000.0 1000.0 1000.0
```

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

```
ggplot(data = data) +
  geom_histogram(mapping = aes(x = Active_Library_Cardholders_Number, y = Revenue_per_cardholder))

library("Revenue_per_cardholder", y = "Frequency") +
ggtitle("Histogram of Revenue per Active Library Cardholder")
```




### Insight 1 conclusion:

From the summary and the histogram we glean valuable insights into the Revenue_per_cardholder distribution:

- The mean Revenue_per_cardholder is centered at 1e+03.
- The minimum Revenue_per_cardholder recorded is 0.136.
- The first quartile, representing the 25th percentile, is at 1000.000.
- The first quartile, representing the 25th percentile, is at 1000.000.
- The second quartile (median) marking the 50th percentile, stands at 1000.000.
- The third quartile, indicating the 75th percentile, is at 1000.000.

## Insight 2

Observing the distribution of Revenue_per_cardholder using a scatterplot

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

### Insight 2 conclusion

- The scatter plot reveals a discernible relationship between the number of active cardholders and Revenue_per_cardholder. Notably, as the count of active cardholders decreases, the Revenue_per_cardholder tends to increase, and conversely, when the number of active cardholders is higher, the Revenue_per_cardholder shows a tendency to decrease. This inverse correlation suggests a potential impact of cardholder quantity on revenue per cardholder, warranting further analysis.
- The correlation coefficient between "Revenue_per_cardholder" and "Active Library Cardholders" is approximately -0.0427782. This value is close to zero, indicating a very weak linear relationship between these two variables.
- We also notice active score outliers around the 1500000 Revenue_per_cardholder

## Insight 3

Analyzing the Revenue_per_cardholder over years

```
#Average data by year and use the Revenue_per_cardholder
group1_data_year <- data %>%
  group_by(Survey_Year_From) %>%
  summarise(total_revenue_per_cardholder = sum(Revenue_per_cardholder), na.rm = TRUE)
table_rounded1<-round(group1_data_year, "Total", full_width = FALSE, position = "center")

##   Survey_Year_From Total_revenue_per_cardholder
## 1             2015                     39893.28
## 2             2016                     32216.75
## 3             2017                     33676.21
## 4             2018                     34122.23

#Create a line series plot
ggplot(group1_data_year, data=) +geom_line(aes_string(x = "Survey_Year_From", y = "Total_Revenue_per_cardholder", group = 1)) +
  geom_point(aes_string(x = "Survey_Year_From",
  geom_point(color = "blue") +
  labs(x = "Year", y = "Revenue_per_cardholder", title = "Revenue_per_cardholder Trends Over Time")
```

### Insight 3 conclusion

From the line graph, it becomes evident that Revenue_per_cardholder has exhibited a consistent upward trend from 2015 to 2019. The highest Revenue_per_cardholder was recorded in 2015, with a subsequent decline observed in 2016. The years 2017 and 2018 demonstrate relatively similar Revenue_per_cardholder trends, indicating a degree of stability during that period.

So the question that rises here, what changed from 2015 to 2018 that affected the increase of the Revenue per card holder?

## Recommendation:

- More analysis is be conducted to find a relation between the over column and other columns
- In the file Following-up.html, I included code to check every numeric column in the data set and its co-relation with the rest of the columns.
- Surprisingly the result shows that Project Grants had no relation with any other columns.
- Conclusion that from has to be other data about these years are not included in the data set and more data is be collected and used in this analysis.

## Insight 4

Analyzing the Revenue_per_cardholder between cities

```
#Average data by city and use the Revenue_per_cardholder
group4_data_city <- data %>%
  group_by(City_Town) %>%
  summarise(total_revenue_per_cardholder = sum(Revenue_per_cardholder), na.rm = TRUE)
table_rounded4<-round(group4_data_city, "Total", full_width = FALSE, position = "center")

##   City_Town Total_Revenue_per_cardholder
## 1   Address                     101.6304
## 2      Ajax                     147.5329
## 3   Alliston                     107.1323
## 4   Amherst                      183.5146
## 5     Angus                      120.8039

#Create a bubble plot
ggplot(group4_data_city, aes(x = City_Town, y = Total_Revenue_per_cardholder, size = Total_Revenue_per_cardholder, color =
  Total_Revenue_per_cardholder)) +
  geom_point(alpha = 0.5) +
  scale_size_continuous(range = c(1, 20)) +
  labs(x = "City", y = "Revenue_per_cardholder", title = "Bubble Plot of Revenue per cardholder by City")
```




Given 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.

```
#Filtering the data to only include Toronto, Hamilton, Mississauga, Oakville, Brampton
filter(data$City_Town, data$ Active_Library_Cardholders_Number, data$ Revenue_per_cardholder, data$ Total_Operating_Revenues)

##   Library_Number Survey_Year_From Revenue_per_cardholder Active_Library_Cardholders_Number Total_Operating_Revenues City_Town Province Project_Grants Total_Capital_Revenue Total_Prov_Operating_Funding Total_spends_on_computers Total_spends_on_utilities Self_generated_revenues
## 1          L0850             2016             213.3745                   23750                 5068253 Toronto  Ontario        229842            24230488                   5373627                  571550                 1644824                 710321
## 2          L0850             2017             213.3745                   23750                 5068253 Toronto  Ontario        229842            24230488                   5373627                  571550                 1644824                 710321
## 3          L0170             2016             208.6135                    9154                 1909393 Toronto  Ontario        231484            20200931                   5371627                  556742                 1595284                 687791
## 4          L0470             2016             206.5450                   10747                 2221554 Hamilton Ontario         53393                 0                   545401                   281542                  411992                  584375
## 5          L0850             2015             205.7985                   23750                 4886425 Toronto  Ontario        229827            24270056                   5373627                  561556                 1601550                 705404

#Group data by city and use the Revenue_per_cardholder
group5$Total=sum(.)
table_rounded5<-round(group5, "Total", full_width = FALSE, position = "center")
  summarise(total_revenue_per_cardholder = sum(Revenue_per_cardholder), na.rm = TRUE)
table_rounded5<-round(group5, "Total", full_width = FALSE, position = "center")

##   City_Town Total_Revenue_per_cardholder
## 1  Brampton                     455.6152
## 2  Hamilton                     812.3200
## 3 Mississauga                   319.8210
## 4   Oakville                    545.3780
## 5   Toronto                     1404.5390

#Create a bubble plot to visualize the Revenue_per_cardholder among these five cities
ggplot(group5$Total=sum(.), aes(x = City_Town, y = Total_Revenue_per_cardholder, size = Total_Revenue_per_cardholder, color =
  Total_Revenue_per_cardholder)) +
  geom_point(alpha = 0.5) +
  scale_size_continuous(range = c(1, 20)) +
  labs(x = "City", y = "Revenue_per_cardholder", title = "Bubble Plot of Revenue per cardholder by City")
```




## Insight 4 conclusion

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

So the question that rises here, what is incommon between Toronto and Hamilton that wasn't present in the other 3 cities that affected this result.

## Recommendation:

- More analysis to be conducted to find the relation between this city column and other columns that might affect Revenue_per_cardholder
- I did this step, like the Following-up.html shows all the relations between all the numeric columns in the data set.
- Surprisingly, Project_Grants had no relation with any other columns.
- Conclusion that has to be conducted on city column and its relation with other columns is recommended.