Project 2

Mohammad Zahid Chowdhury

2025-03-09

Introduction: The goal of this assignment is to practice in preparing different datasets for downstream analysis work.I have chosen 3 data sets, for example, Data Set 1 is students score, Data Set 2 is Sales Data and Data Set 3 is Water Consumption And Cost (2013 - Feb 2025).These three data sets are examples of wide data and for this project and I will try to make the data more tidy and then go to conduct the analysis.

Loading required packages:

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2
```

```
## Warning: package 'readr' was built under R version 4.4.2
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
## Warning: package 'stringr' was built under R version 4.4.2
```

```
## -- 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
```

```r
library(dplyr)
library(ggplot2)
```

Data Set 1: Students Score

Read the dataset:

```
students_score <- read.csv("https://raw.githubusercontent.com/zahid607/Project-2/refs/heads/main/Studen

head(students_score)
```

```
##          Name Age                             Scores       Address
## 1    John Doe  25              Math: 80, Science: 85  123 Main St.
## 2  Jane Smith  30              Math: 92, Science: 88   456 Oak St.
## 3 Sarah White  22              Math: 75, History: 80  789 Pine St.
## 4   Bob Brown  28              Math: 85, Science: 90
## 5 Carol Green  26 Math: 78, Science: 80, History: 85 101 Maple St.
## 6  Alice Blue  NA              Math: 90, History: 85   202 Elm St.
```

Columns name of the data set:

```
colnames(students_score)
```

```
## [1] "Name"    "Age"     "Scores"  "Address"
```

Transform the data:

```
# Separate the Scores column into Math, Science, and History
students_score <- students_score %>%
  separate(Scores, into = c("Math", "Science", "History"), sep = ", ", remove = FALSE)
```

```
## Warning: Expected 3 pieces. Missing pieces filled with ‘NA‘ in 5 rows [1, 2, 3,
## 4, 6].
```

```
# Check the cleaned data
head(students_score)
```

```
##          Name Age                             Scores     Math     Science
## 1    John Doe  25              Math: 80, Science: 85 Math: 80 Science: 85
## 2  Jane Smith  30              Math: 92, Science: 88 Math: 92 Science: 88
## 3 Sarah White  22              Math: 75, History: 80 Math: 75 History: 80
## 4   Bob Brown  28              Math: 85, Science: 90 Math: 85 Science: 90
## 5 Carol Green  26 Math: 78, Science: 80, History: 85 Math: 78 Science: 80
## 6  Alice Blue  NA              Math: 90, History: 85 Math: 90 History: 85
##       History       Address
## 1        <NA>  123 Main St.
## 2        <NA>   456 Oak St.
## 3        <NA>  789 Pine St.
## 4        <NA>
## 5 History: 85 101 Maple St.
## 6        <NA>   202 Elm St.
```

Reshape the column & the "Math:", "Science:", "History:" text, leaving only the numeric values:

```
students_score <- students_score %>%
  mutate(
    Math = as.numeric(gsub("Math: ", "", Math)),
    Science = as.numeric(gsub("Science: ", "", Science)),
    History = as.numeric(gsub("History: ", "", History))
  )
```

```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘Science = as.numeric(gsub("Science: ", "", Science))‘.
## Caused by warning:
## ! NAs introduced by coercion
```

```
head(students_score)
```

```
##          Name Age                             Scores Math Science History
## 1    John Doe  25              Math: 80, Science: 85   80      85      NA
## 2  Jane Smith  30              Math: 92, Science: 88   92      88      NA
## 3 Sarah White  22              Math: 75, History: 80   75      NA      NA
## 4   Bob Brown  28              Math: 85, Science: 90   85      90      NA
## 5 Carol Green  26 Math: 78, Science: 80, History: 85   78      80      85
## 6  Alice Blue  NA              Math: 90, History: 85   90      NA      NA
##         Address
## 1  123 Main St.
## 2   456 Oak St.
## 3  789 Pine St.
## 4
## 5 101 Maple St.
## 6   202 Elm St.
```

Number of Missing Data:

```
missing_data <- colSums(is.na(students_score))
print(missing_data)
```

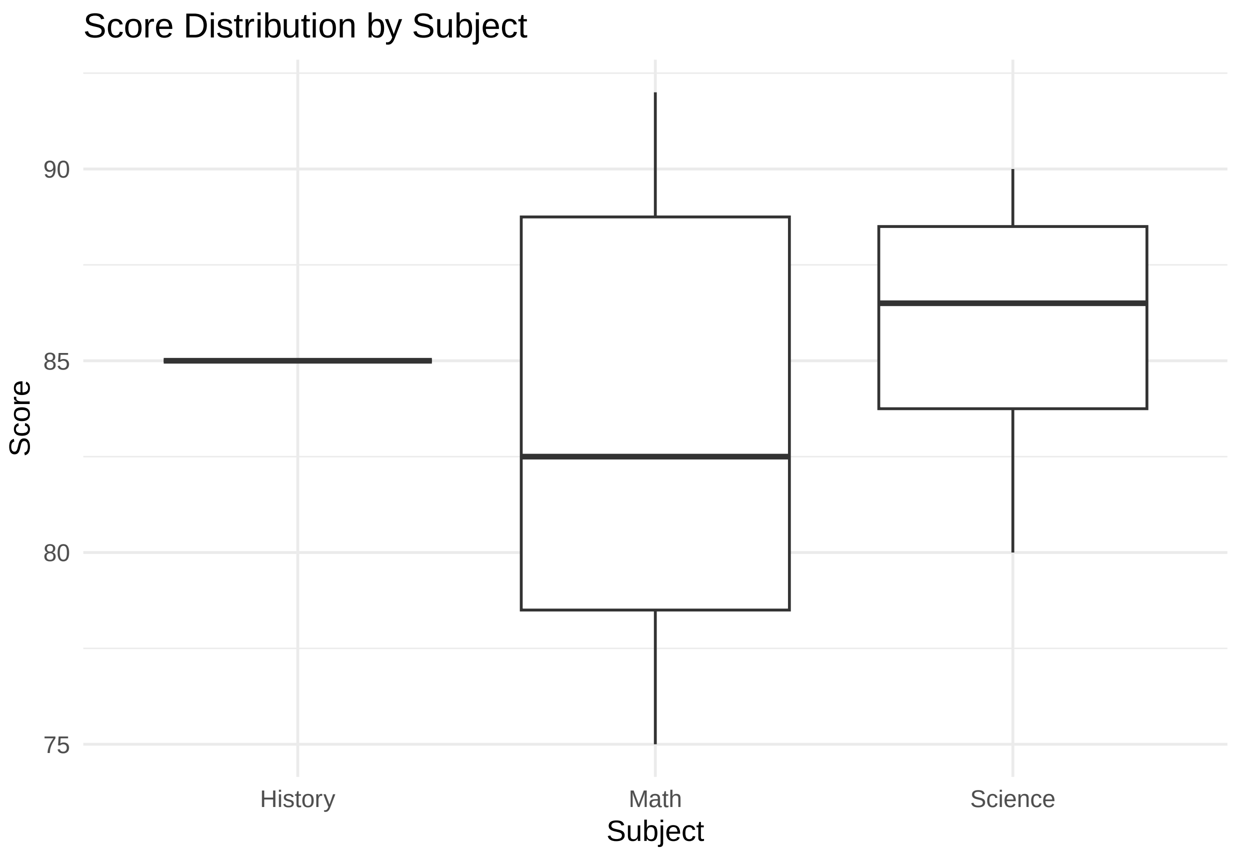
```
##    Name     Age  Scores    Math Science History Address
##       0       1       0       0       2       5       0
```

Boxplots to compare the distribution of scores in different subjects and any outlier detection.

```
# Reshape data for plotting
students_score <- students_score %>%
  pivot_longer(cols = c("Math", "Science", "History"),
               names_to = "Subject", values_to = "Score")

ggplot(students_score, aes(x = Subject, y = Score)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Score Distribution by Subject")
```

```
## Warning: Removed 7 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



Average Score of Students.

```
students_score_avg <- students_score %>%
  group_by(Name) %>%
  summarise(avg_score = mean(Score, na.rm = TRUE)) %>%
  arrange(desc(avg_score))

head(students_score_avg)
```

```
## # A tibble: 6 x 2
##   Name        avg_score
##   <chr>           <dbl>
## 1 Alice Blue       90
## 2 Jane Smith       90
## 3 Bob Brown        87.5
## 4 John Doe         82.5
## 5 Carol Green      81
## 6 Sarah White      75
```

Data Set 2: Sales Data

```
Sales_Data <- read.csv("https://raw.githubusercontent.com/zahid607/Project-2/refs/heads/main/Sales%20Da

head(Sales_Data)
```

```
##   Employee.Name Jan_Sales Feb_Sales Mar_Sales       Region Department
## 1      John Doe       500       600       700 North Region      Sales
## 2    Jane Smith       300       400       500 South Region      Sales
## 3   Sarah White       400       450       500  East Region  Marketing
## 4     Bob Brown       600       700       750  West Region      Sales
## 5   Carol Green       350       450       400 North Region         HR
## 6    Alice Blue       200       250       300 South Region  Marketing
```

Columns Names:

```
colnames(Sales_Data)
```

```
## [1] "Employee.Name" "Jan_Sales"     "Feb_Sales"     "Mar_Sales"
## [5] "Region"        "Department"
```

Creating the untidy Sales data using data.frame directly

```
Sales_Data <- data.frame(
  Employee.Name = c("John Doe", "Jane Smith", "Sarah White", "Bob Brown", "Carol Green", "Alice Blue"),
  Jan_Sales = c(500, 300, 400, 600, 350, 200),
  Feb_Sales = c(600, 400, 450, 700, 450, 250),
  Mar_Sales = c(700, 500, 500, 750, 400, 300),
  Region = c("North Region", "South Region", "East Region", "West Region", "North Region", "South Regio
  Department = c("Sales", "Sales", "Marketing", "Sales", "HR", "Marketing")
)

# Use pivot_longer to reshape the data from wide to long format
Sales_Data <- Sales_Data %>%
  pivot_longer(cols = starts_with("Jan_Sales"):starts_with("Mar_Sales"),
               names_to = "Month",
               values_to = "Sales",
               names_prefix = "([A-Za-z]+)_")  # Removing the prefix (Jan_, Feb_, Mar_)

# View the tidy data
head(Sales_Data)
```

```
## # A tibble: 6 x 5
##   Employee.Name Region       Department Month Sales
##   <chr>         <chr>        <chr>      <chr> <dbl>
## 1 John Doe      North Region Sales      Sales   500
## 2 John Doe      North Region Sales      Sales   600
## 3 John Doe      North Region Sales      Sales   700
## 4 Jane Smith    South Region Sales      Sales   300
## 5 Jane Smith    South Region Sales      Sales   400
## 6 Jane Smith    South Region Sales      Sales   500
```

Summary Statistics:

```
summary(Sales_Data$Sales)
```

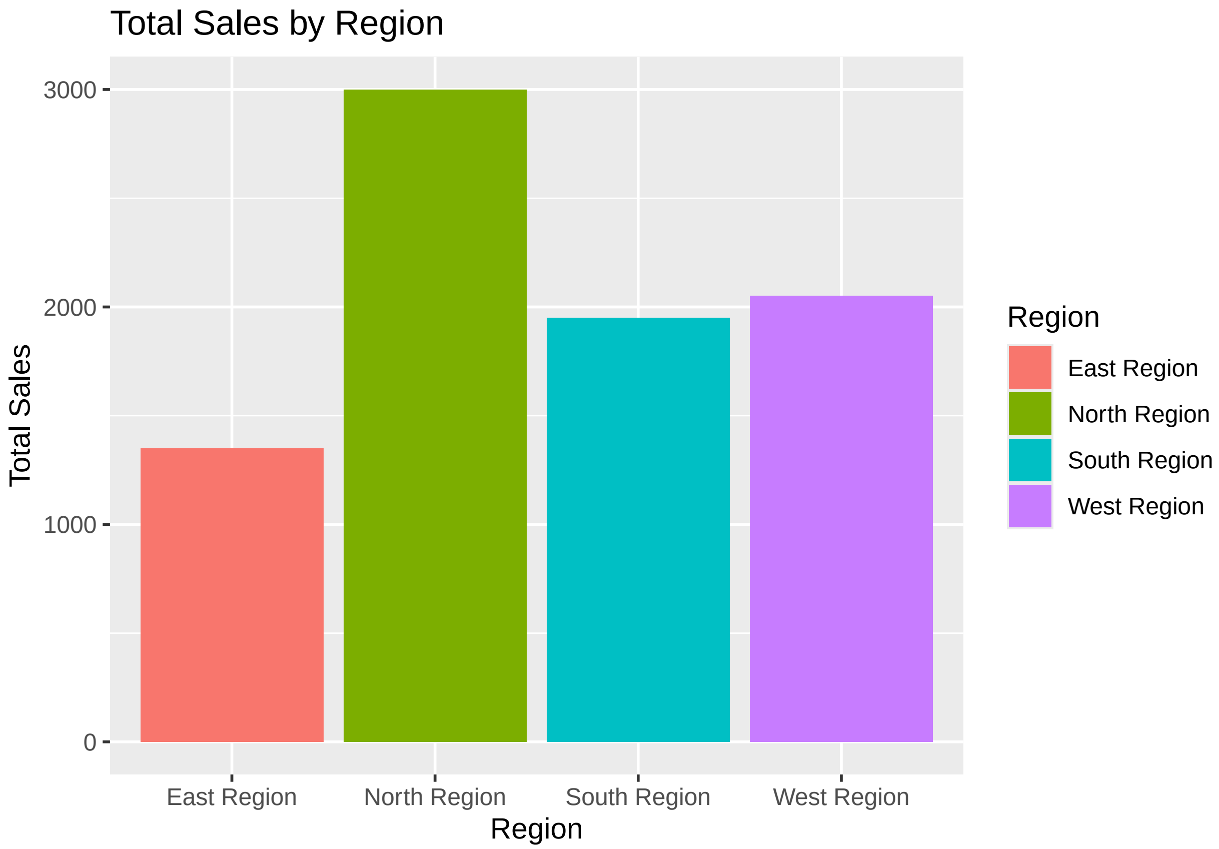
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   362.5   450.0   463.9   575.0   750.0
```

```
Sales_Data %>%
  group_by(Department) %>%
  summarise(
    Total_Sales = sum(Sales, na.rm = TRUE),
    Average_Sales = mean(Sales, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 3
##   Department Total_Sales Average_Sales
##   <chr>            <dbl>         <dbl>
## 1 HR                1200          400
## 2 Marketing         2100          350
## 3 Sales             5050          561.
```

Bar diagram of total sales by region:

```
Sales_Data %>%
  group_by(Region) %>%
  summarise(Total_Sales = sum(Sales, na.rm = TRUE)) %>%
  ggplot(aes(x = Region, y = Total_Sales, fill = Region)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Sales by Region", x = "Region", y = "Total Sales")
```



Data Set 3: Water Consumption And Cost (2013 - Feb 2025)

Read the data set:

```
water_data<-read.csv("https://raw.githubusercontent.com/zahid607/Project-2/refs/heads/main/Water_Consum

head(water_data)
```

```
##   Development.Name  Borough  Account.Name                  Location Meter.AMR
## 1    HOWARD AVENUE BROOKLYN HOWARD AVENUE                    BLD 02       AMR
## 2     BAISLEY PARK   QUEENS  BAISLEY PARK                    BLD 09       AMR
## 3     BAISLEY PARK   QUEENS  BAISLEY PARK                    BLD 09       AMR
## 4     BAISLEY PARK   QUEENS  BAISLEY PARK                    BLD 09       AMR
## 5     BAISLEY PARK   QUEENS  BAISLEY PARK                    BLD 09       AMR
## 6         BAY VIEW BROOKLYN      BAY VIEW BLD 25 - Community Center      NONE
##        Meter.Scope TDS.. EDP RC.Code     Funding.Source        AMP..
## 1                   339 782 K033900            FEDERAL NY005013510P
## 2           BLD 09    91 240 Q009100            FEDERAL NY005010910P
## 3           BLD 09    91 240 Q009100            FEDERAL NY005010910P
## 4           BLD 09    91 240 Q009100            FEDERAL NY005010910P
## 5           BLD 09    91 240 Q009100            FEDERAL NY005010910P
## 6 Community Center    92 670 K209200 MIXED FINANCE/LLC1 NY005020920P
##                  Vendor.Name UMIS.BILL.ID Revenue.Month Service.Start.Date
## 1 NEW YORK CITY WATER BOARD      8870656       2020-04          3/23/2020
## 2 NEW YORK CITY WATER BOARD      8562430       2020-01         12/23/2019
## 3 NEW YORK CITY WATER BOARD      8667039       2020-02          1/26/2020
## 4 NEW YORK CITY WATER BOARD      8759719       2020-03          2/24/2020
## 5 NEW YORK CITY WATER BOARD      8870760       2020-04          3/23/2020
## 6 NEW YORK CITY WATER BOARD      8560969       2020-01         12/23/2019
##   Service.End.Date X..days Meter.Number Estimated Current.Charges
## 1        4/23/2020      31    E11310572         N         2945.22
## 2        1/26/2020      34    K13060723         N          196.35
## 3        2/24/2020      29    K13060723         N          258.35
## 4        3/23/2020      28    K13060723         N          217.02
## 5        4/23/2020      31    K13060723         N          103.34
## 6        1/26/2020      34    E17250205         N           72.34
##               Rate.Class Bill.Analyzed Consumption..HCF. Water.Sewer.Charges
## 1 Basic Water and Sewer           Yes               285             2945.22
## 2 Basic Water and Sewer           Yes                19              196.35
## 3 Basic Water and Sewer           Yes                25              258.35
## 4 Basic Water and Sewer           Yes                21              217.02
## 5 Basic Water and Sewer           Yes                10              103.34
## 6 Basic Water and Sewer           Yes                 7               72.34
##   Other.Charges
## 1             0
## 2             0
## 3             0
## 4             0
## 5             0
## 6             0
```

Clean column names and remove rows with missing data:

```
water_data<-water_data%>%
  rename()

#  Remove rows with missing data
water_data_clean <- na.omit(water_data)
```

Handle Missing Values & Check for missing values:

```
colSums(is.na(water_data_clean))
```

```
##    Development.Name             Borough        Account.Name            Location
##                   0                   0                   0                   0
##           Meter.AMR         Meter.Scope               TDS..                 EDP
##                   0                   0                   0                   0
##             RC.Code      Funding.Source               AMP..         Vendor.Name
##                   0                   0                   0                   0
##        UMIS.BILL.ID       Revenue.Month  Service.Start.Date    Service.End.Date
##                   0                   0                   0                   0
##             X..days        Meter.Number           Estimated     Current.Charges
##                   0                   0                   0                   0
##          Rate.Class       Bill.Analyzed   Consumption..HCF. Water.Sewer.Charges
##                   0                   0                   0                   0
##       Other.Charges
##                   0
```

Filtering water data:

```
# Store the original number of rows
original_rows <- nrow(water_data_clean)

# Apply the filtering step
water_data_clean <- water_data_clean %>%
  filter(Current.Charges >= 0, Consumption..HCF. >= 0)

# Store the new number of rows after filtering
filtered_rows <- nrow(water_data_clean)

# Check if any rows were removed
if (original_rows == filtered_rows) {
  print("No outliers in the dataset")
} else {
  print("Outliers were removed from the dataset")
}
```

```
## [1] "Outliers were removed from the dataset"
```

Summarize the total water consumption for each borough.

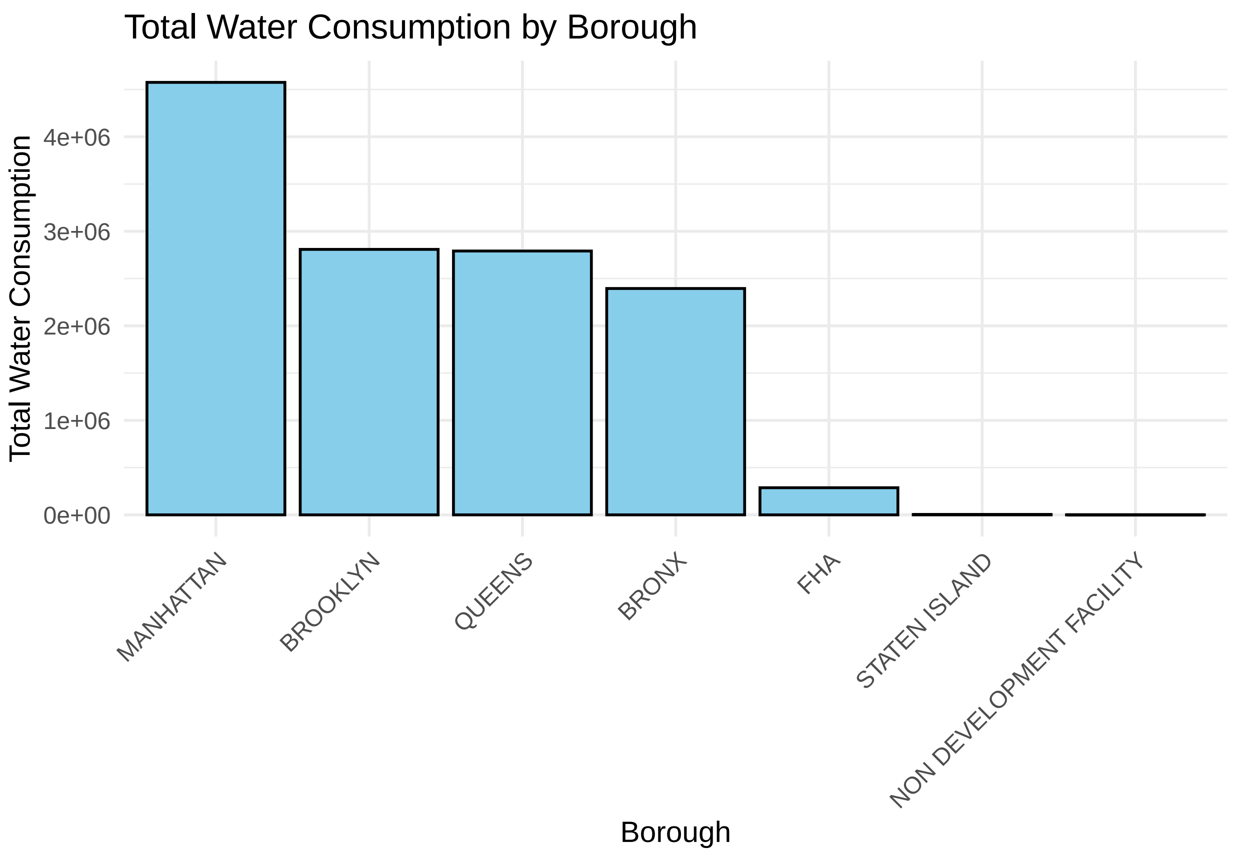
```
borough_consumption <- water_data_clean %>%
  group_by(Borough) %>%
  summarize(Total_Consumption = sum(Consumption..HCF., na.rm = TRUE))

# View the result
print(borough_consumption)
```

```
## # A tibble: 7 x 2
##   Borough                  Total_Consumption
##   <chr>                                <dbl>
## 1 BRONX                          2394508.
## 2 BROOKLYN                       2808310.
## 3 FHA                             286882
## 4 MANHATTAN                      4575288.
## 5 NON DEVELOPMENT FACILITY           986
## 6 QUEENS                         2791315
## 7 STATEN ISLAND                     3960
```

Create a bar plot to visualize water consumption by borough

```
ggplot(borough_consumption, aes(x = reorder(Borough, -Total_Consumption), y = Total_Consumption)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Total Water Consumption by Borough",
       x = "Borough",
       y = "Total Water Consumption") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Conclusion: In this analysis, I worked with 3 differents data sets containing different information about students perfomance, employee perfomance and water consumption.These datasets offered a wide variety of information that I could clean, transform, and analyze to draw meaningful insights.