Using Regression Analysis for Predicting Energy Consumption in Dubai Police

Data Analysis

4.1 Data Preparation

First of all, we need to add all bills for all quarters from 2017 to 2021 in one sheet (a total of 17,873 rows). Then, we need to merge all the required data. This was done using the "left_join" function in R. Facilities list will be joined with bills data sheet based on "contract number" and "Division" to get two more columns (DepartmentOrPS and GHQ Department). Then, we will join capita and area based on the DepartmentOrPS column.

4.2 Data Preprocessing

4.2.1 Data Cleaning

The cleaning process was conducted in Microsoft Excel and R Studio. The bills received from DEWA contain many duplicates, zero values, Na's, unknown locations, and unnecessary accounts. I followed these cleaning steps before we start joining the three datasets together:

Below are the cleaning steps for the bill's datasheet:

Microsoft Excel:

- Compile all year's bills together (monthly basis from January to December)
- Prepare datasheet for locations with their meter numbers (Contract Account)
- Add a new column for years to distinguish which year each consumption is assigned.
- Remove the units from the Consumption unit and Consumption Amount columns. Units are affecting the variable type in R so it's difficult to deal with them as numeric. As we have a column (Division) describing each row if it is assigned to "Electricity" or "Water" keeping the units is useless.

R Studio:

- Removing NAs, blanks, and zero values in Bills dataset.

```
                Division                        X Collective.Contract.Acc          Contract.Account
                       0                        0                        0                         0
Contract.Account.Nam          Calendar.month          Consumption.Unit        Consumption.Amount
                       0                        0                      210                       210
          Tax.Amount          Tax.base.amount                    Year
                    3653                     3653                       0
```

- Removing NAs and zero values in the "Consumption.Unit" column.

There are 210 NA values and 1160 Zero values in the Consumption.unit. As all the study is based on this variable so it is not logical to keep zero values in this column. As it is impossible

to have zero consumption for a working meter unless they closed a building for at least 1 month. And even if it does happen, we will not need this information for the study.

- Replacing NA in Tax.Amount with zero.

Some values in the "Tax.Amount" column are NAs because the tax policy in UAE starts on 2018. So, the data of this column for year 2017 will be NA.

- Replacing NA in Tax.base.amount by Consumption.Amount plus Tax.Amount for year 2017

- Changing the type of features

Convert the month column from character to factor.

Convert Consumption.Unit, Consumption.Amount, Tax.Amount, and Tax.base.Amount from character to numeric.

The output of the bill's dataset:

After the cleaning steps, we have 16,502 insights and 11 columns.

```
```{r}
dim(data)
```

[1] 16502   11
```

4.2.2 Data Preparation

First, we need to unify the names of columns that will be joined based on. To join bills with facilities list based on "Contract.Account" and "Division" columns. So, the column name in the Facilities list file should be changed from "contract account number" to "Contract.Account". Same with Area and Capita file, the two files will be joined based on the "DepartmentOrPS" column in the bills file which is the same as "Facilities" in Area and Capita file. The column name has been changed to "DepartmentOrPS" to be easily joined. Also, some facilities' names are not unified in bills and (Area and Capita) files as GHQ is named in GHQ Campus. This is also needed to be fixed before joining the process tale place. Use the left.join() and join() function to merge the three datasheets in one dataset as shown below.

```
names(Facilities.List)[2] <- "Contract.Account"
names(Facilities.List)[4] <- "Division"
data <- left_join(Bills, Facilities.List, by=c("Contract.Account","Division"))
```

```
# Join both datasets by department or PS column
base1 <- join(data, area_capita, by = "DepartmentOrPS")
```

- Joining the three datasets don to add all of the following columns:

- Add a new column describing which Department or Police Station this contract number is assigned to, this column is called “DepartmentOrPS”.
- Add a new column for the exact location in General Head Quarters called “GHQ.Department”.
- Add a column for the Area of each location called “Area (meter square)”.
- Add a column for Capita number of employees called “Capita”.

- Remove all rows that are not assigned to any PS or Department (extra account number received in the bill (might be private or excepted from the payment or unneeded for the study).

```
  {r}
# Filter data having empty rows / drop null values
data <- data %>%
  filter (DepartmentOrPS != "")

dim(data)
```

```
[1] 11773    10
```

- Remove unneeded columns. Like, Collective.Contract.Acc and Contract.Account.Nam columns are not useful factors for our study. This information was added to the bill for payment sakes that will not be affected by our study.

- Remove locations that not have Area and Capita values. We will be left with 27 Departments and Police Stations.

```
base1 <- base1 %>%
  filter ( Capita!= 0 ,  `Area (meter square)`!= 0)

unique(base1$DepartmentOrPS)
```

```
 [1] "Awir Horse Stables"                          "GHQ"
 [3] "K9"                                          "Al Faqqa Police Station"
 [5] "Punitive and Correctional Establishments"    "Jabal Ali Police Station"
 [7] "General Department of Transport and Rescue"  "Lahbab Police station"
 [9] "Barsha Police Station"                       "Naif Police Station"
[11] "Al Rashdiyah Police Station"                 "Qusais Warehouses"
[13] "Rowaiyah Shooting Range"                     "Dubai Police Academy"
[15] "Al Wasl Protective Security and Emergency"   "Officers Club"
[17] "Port Police Station"                         "Al Riffa Police Station"
[19] "Traffic Department Deira"                    "Airport Security"
[21] "Barsha Traffic Dept"                         "Bur Dubai Police Station"
[23] "Qusais Police Station"                       "Moraqabat Police Station"
[25] "Nad Alsheba Police Station"                  "Hor Al Anz Protective Security and Emergency"
[27] "Hatta Police Station"
```

- Removing duplicated rows:

Duplicated rows are not needed in the model so they should be removed from the dataset. There are no duplicated rows as the left_join function was used. When the "join" function was used to join meters and bills datasets, there were 1387 which was removed automatically by using "left_join".

We will end up having 11,155 rows and 12 columns.

```
[1] 11155    12
```

- Convert the Dataset to be quarterly by assigning each quarter to their month and take the summation of consumption unit and amount. The below code was used to generate the datasheets for electricity and water data.

```
#add quarter column
# Consumption Units By Year and Quarter
base1$Calendar.month <- as.character(base1$Calendar.month)
base1$Quarter[base1$Calendar.month %in% c(1,2,3)] <- "Q1"
base1$Quarter[base1$Calendar.month %in% c(4,5,6)] <- "Q2"
base1$Quarter[base1$Calendar.month %in% c(7,8,9)] <- "Q3"
base1$Quarter[base1$Calendar.month %in% c(10,11,12)] <- "Q4"
base_quarter_E <- base1 %>%
  filter(Division == "Electricity") %>%
  group_by(Year, Quarter, Capita, `Area (meter square)`, DepartmentOrPS) %>%
  summarise(Consumption = sum(Consumption.Unit) , Amount = sum(Consumption.Amount))
base_quarter_W <- base1 %>%
  filter(Division == "Water") %>%
  group_by(Year, Quarter, Capita, `Area (meter square)`, DepartmentOrPS) %>%
  summarise(Consumption = sum(Consumption.Unit) , Amount = sum(Consumption.Amount))
```

4.3 Data Quality Dimensions

To check data quality, we need to look at its six dimensions. The explanation of each is mentioned below:

1. **Completeness:** the dataset has missing values and is set as NA or blanks as well as zero consumption values. Facilities area and capita values are not provided. To overcome any incompleteness in the datasets, we removed these facilities from the model.
2. **Conformity:** The facilities list received from the properties department names the facility different than the table provided for Area and Capita for each facility. This needs to be unified before joining the two datasets together.
3. **Consistency:** To make sure the bill amount (Tax.base.Amount) column received from DEWA is equal to Consumption.amount plus Tax.amount. A new column was created to check if they are consistent. The result was that our dataset is consistent.
4. **Accuracy:** The dataset was taken directly from the concerned department in Dubai Police. This makes us sure that it is accurate. However, the Temperature column added was for the temperature detected in Dubai City not specifying the year. If it has been found for the exact location and each year and quarter it will be more accurate.
5. **Duplicates:** Duplicates can't be applied to each column. As the dataset contains the consumption data for the same "Contract Numbers" repeated for different years and moth. The only duplicates we should avoid is having the same electricity/water consumption duplicated for the same facility "Contract Number" recorded in the same year and month. As this case was checked, the datasets do not contain any duplicates.
6. **Integrity:** The dataset is integrated and connected very well. When we joined the three datasets together based on specific attributes that confirm that all attributes are related and connected.

4.4 Feature Engineering

Area Rank (discretization)

To have a good visualization in comparing departments and police stations to each other based on the similar parameter (Area). A new feature was added, “Area Rank” explained in the table below.

Table 7: Area Rank discretization details.

Area_Rank	Area (meter square)	DepartmentOrPS
Extremly High Area Facilities	100,000 meters square - 200,000 meters square	GHQ
		Dubai Police Academy
		Punitive and Correctional Establishments
High Area Facilities	30,000 meters square - 42,000 meters square	Al Wasl Protective Security and Emergency
		Barsha Police Station
		Officers Club
		General Department of Transport and Rescue
Medium Area Facilities	10,000 meters square - 30,000 meters square	Naif Police Station
		K9
		Nad Alsheba Police Station
		Al Rashdiyah Police Station
		Al Riffa Police Station
		Moraqabat Police Station
		Port Police Station
		Traffic Department Deira
		Barsha Traffic Dept
Small Area Facilities	5,000 meters square - 10,000 meters square	Bur Dubai Police Station
		Qusais Police Station
		Airport Security
		Awir Horse Stables
Very Small Area Facilities	0 meters square - 5,000 meters square	Hor Al Anz Protective Security and Emergency
		Lahbab Police station
		Hatta Police Station
		Al Faqqa Police Station
		Qusais Horse Stables
		Qusais Warehouses
		Jabal Ali Police Station
		Rowaiyah Shooting Range

This was done using mutate () function as shown here:

```
#discretize area
base_quarter_E <-
base_quarter_E %>%
    mutate(Area_Rank = case_when( `Area (meter square)` >= 100000 ~ "Extremly High Area Facilities", `Area (meter square)` >= 30000 &  `Area
(meter square)` <= 42000 ~ "High Area Facilities" , `Area (meter square)` >= 10000 &  `Area (meter square)` < 30000 ~ "Medium Area Facilities" ,
`Area (meter square)` >= 5000 &  `Area (meter square)` < 10000 ~ "Small Area Facilities" , `Area (meter square)` < 5000 ~ "Very Small Area
Facilities"
                       ))
```

Consumption per m square

A new attribute was added for each electricity and water consumption which is as below:
Consumption.per.area = Consumption / Area (meter square)
As Consumption is either in kWh or Gallons.
This attribute will give us a better comparison criterion between buildings per building area.

Consumption per Capita

A new attribute was added for each electricity and water consumption which is as below:
Consumption.per.capita = Consumption / Capita (number of employees)
As Consumption is either in kWh or Gallons
This attribute will give us a better comparison criterion between departments as it is per employee number.

Quarter

Add a column that describes each month related to each quarter. As the study is to predict the quarter consumption so the dataset must be built quarterly before input it in the model. The code was shown in the previous section. The column will be described as below:

Table 8: Quarter Column description.

Calendar month	Quarter
1 or 2 or 3	Q1
4 or 5 or 6	Q2
7 or 8 or 9	Q3
10 or 11 or 12	Q4

Temperature

One more column was added to describe the temperature and study the effect of the temperature on the consumption. The temperature of each quarter is "Temperature_Q". The average Temperature in Dubai City was taken for each month as shown below.

Table 9: The average monthly temperature in Dubai (What is weather like in Dubai, Dubai, AE (worldweatheronline.com)).

Calendar month	Temperature
1	21
2	22.5
3	25
4	28.5
5	32.5
6	34.5
7	36
8	36
9	34.5
10	31.5
11	27
12	23.5

Then by taking the average of each quarter we calculated the temperature quarterly as shown below:

Table 10: Quarterly temperature column description.

Quarter	Temperature_Q
Q1	22.83
Q2	31.83
Q3	35.50
Q4	27.33

Using the "join" function. This column was joined to our data set based on the "Quarter" Column respectively.

```
Adding temperature column for quarterly data
```{r}
Temperature_Q <- read_excel("Temp_Q.xlsx")
base_quarter_E <- join(base_quarter_E, Temperature_Q, by = "Quarter")
```
```

4.5 Correlation between attributes

The Correlogram is a graph of the correlation matrix. This is commonly used to highlight the most connected variables in a data set or data table. We can reorder the correlation matrix based on the degree of relationship between the variables.
Negative correlations are shown by a red scale, whereas positive correlations are represented by a blue scale.
The correlation matrix:

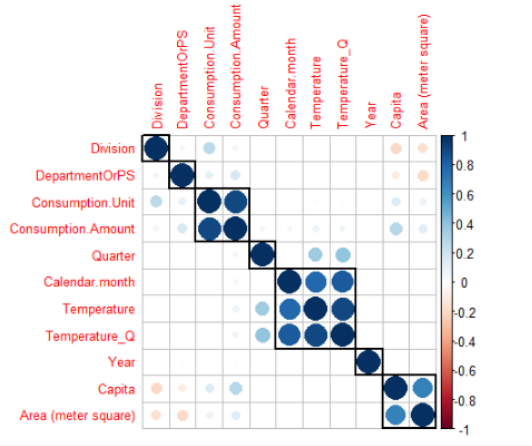

Figure 4: Correlation Matrix.

The tax amount and tax base amount were excluded as they are a function of consumption amount. “Consumption.Unit” is the variable we concerned on. As we can see it is highly correlated with Consumption.amount which is logical as the consumption amount is (consumption unit × tariff rate (constant)). So, it is logical to have them perfectly correlated. It is positively correlated with Division, Capita, DepartmentOrPS, and Area respectively from high to low. It is almost zero correlated with “Year”, “Calendar month”, “Quarter”, “Temperature”, and “Temperature_Q.

4.6 Variable Dictionary

Table 11: Variables Dictionary.

Column	**Data Type**	**Explanation**
Division	Categorical	It is Electricity for electricity consumption and Water for water consumption tuple.
Contract.Account	Character	The electricity and water contract number assigned to each consumption and facility.
Quarter	Categorical	The quarter of the year for the observed consumption. Its "Q1" for the first quarter of the year (month 1,2,3), "Q2" for the second quarter of the year (month 4,5,6), "Q3" for the third quarter of the year (month 7,8,9), "Q4" for the last quarter of the year (month 10,11,12).
Consumption.Unit	Numeric	The quarterly electricity consumption (kwh) for Division "Electricity" and water consumption (gallons) for Division "Water".
Consumption.Amount	Numeric	The quarterly consumption amount in AED received in the bill. It is equal to consumption unit multiply by tariff rate.
Tax.Amount	Numeric	Tax amount (5%) in AED.
Tax.base.amount	Numeric	The total paid amount, consumption amount plus the tax.
Year	Integer	The year where the consumption detected in (2017,2018,2019,2020,2021).
DepartmentOrPS	Categorical	The names of facilities, departments and police stations (27 unique values).
GHQ.Department	Categorical	The specific department name in General Headquarters.
Capita	Numeric	The number of employees in each facility.
Area (meter square)	Numeric	The area in meter squares for each facility.
Consumption.per.capita	Numeric	The quarterly consumption of electricity and water per number of employees.
Consumption.per.Area	Numeric	The quarterly consumption of electricity and water per meter squares.
Area Rank	Categorical	• The area rank for each facility as the following: • "Extremely High Area Facilities" (100,000 meters square - 200,000 meters square) • "High Area Facilities" (30,000 meters square - 42,000 meters square) • "Medium Area Facilities" (10,000 meters square - 30,000 meters square) • "Small Area Facilities" (5,000 meters square - 10,000 meters square) • "Very Small Area Facilities" (0 meters square - 5,000 meters square)
Temperature_Q	Numeric	The temperature in Dubai city in each quarter (Celsius).

1.7 Data Exploration and Visualization

In this section, a set of graphs have been drawn separately for electricity and water consumption to understand the data better before moving toward the modeling part. The plots were done using R Studio and Tableau. This plot is established to have an overview of what type of questions/information we will try to explore in this section, following is a set of questions that are devised in this case.

- Average consumption units by category overall.
- Total Consumption units by year for electricity and water.
- Month-wise comparison of consumption units for electricity and water.
- The quarter-wise trend of consumption for each year for both water and electricity.
- A trend of consumption units over the past few years.
- Consumption units by GHQ Department and DepartmentOrPS for both water and electricity.
- Consumption per capita and Consumption per area plots.
- A comparison of consumption based on each area rank.
- A sample of quarterly report of one department showing the percentage of saving/loss.
- The temperature affects consumption.

Starting with the analysis part, the pie charts shown below tell us about the overall consumption units and consumption amount by category. It is evident that that the average consumption units for water is much higher as compared to electricity, however, the average consumption amounts for both water and electricity are quite close with the consumption amount for water still higher as compared to electricity.

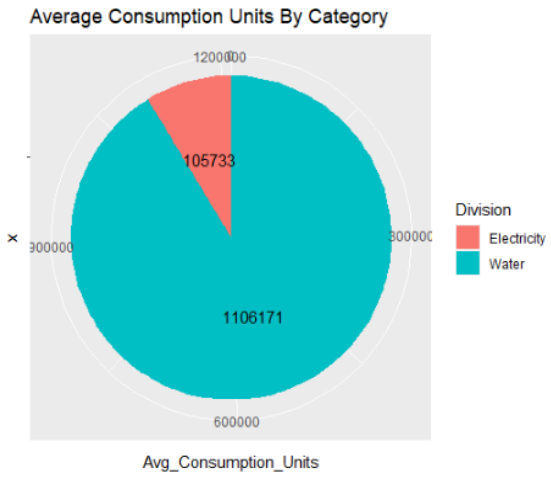

Figure 6: Average Consumption by Division.

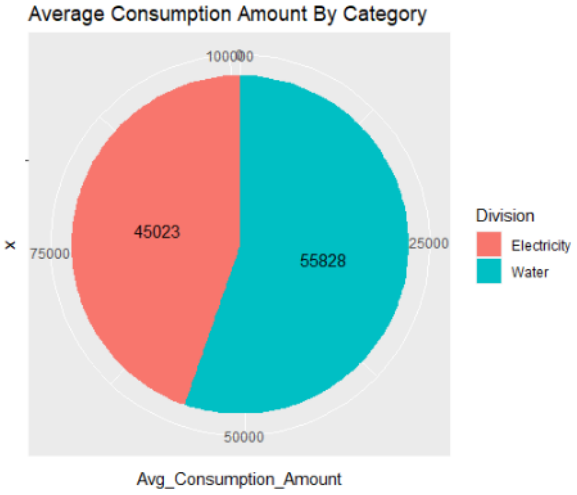

Figure 5: Average Consumption Amount by Division.

The second plot attached below, shows the comparison of total consumption units by year for both water and electricity. Figure 7, shows the consumption units by year for electricity while Figure 8 shows the consumption units by year for water. We can see that the consumption units for electricity are highest in year 2019 among all and lowest in year 2020. This might be due to covid-19 pandemic where most of the departments were closed and employees working remotely from home. Meanwhile the consumption for water is highest in year 2020 and lowest in year 2021. The water consumption units are not affected by the covid-19 pandemic in this case.

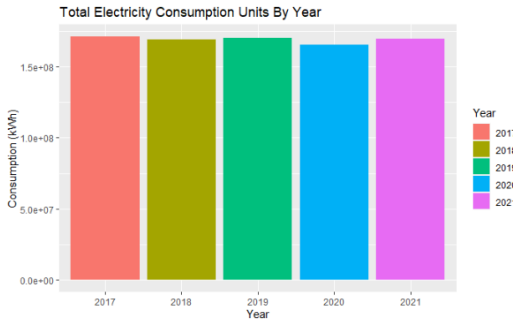

Figure 7: Electricity Consumption by Year.

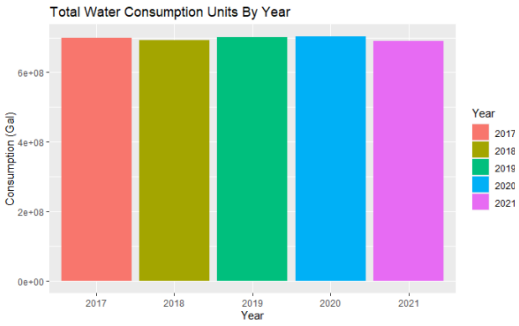

Figure 8: Water Consumption by Year.

Next, we will explore the total consumption units by month to see which month consumes the highest units of electricity and water. We can see from below attached bar graphs that the consumption units for electricity are recorded lowest in 2nd and 3rd month of each year respectively. The highest consumption is recorded in 8th and 9th month which is August and September, the peaks of summer season. Same trend is seen for water as well, the consumption is highest in 8th month and lowest in 3rd month. This shows us that the time of year has a strong effect on consumption of units.

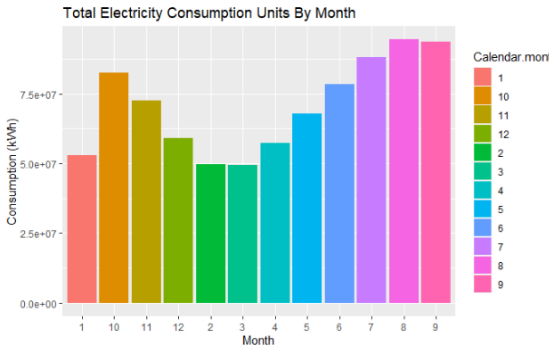

Figure 9: Electriciy Consumption by Month.

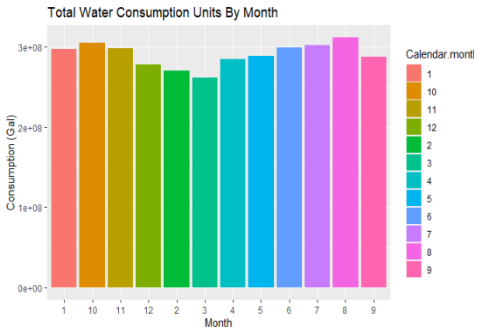

Figure 10: Water Consumption by Month.

While looking at the quarter-wise comparison of consumption units for each year for electricity, we can see that in almost all the years except 2019, the third quarter of the year records the highest consumption of electricity. This information could be seen in the graphs attached below:

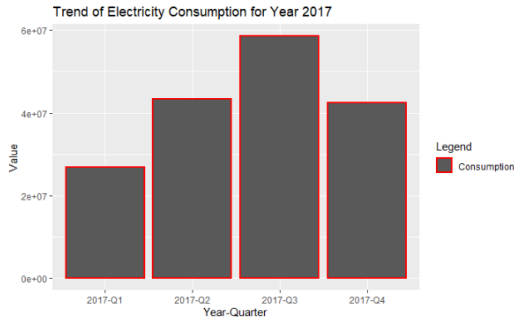

Figure 11: Electricity consumption quarterly for year 2017.

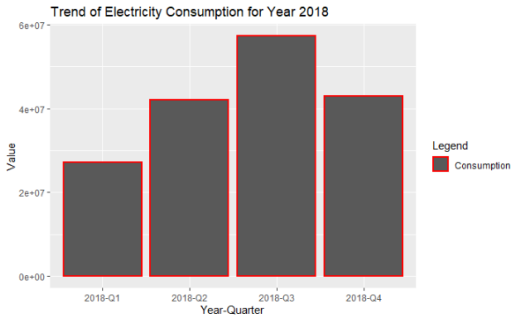

Figure 12: Electricity consumption quarterly for year 2018.

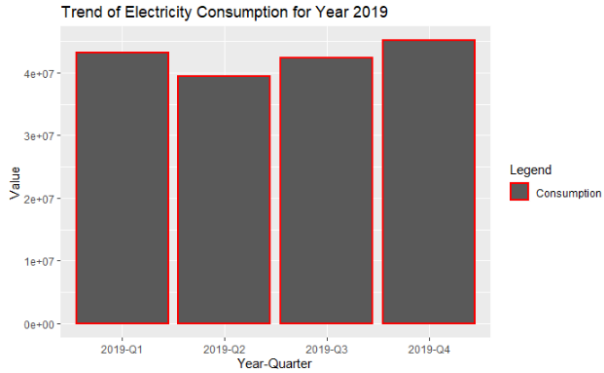

Figure 13: Electricity consumption quarterly for year 2019.

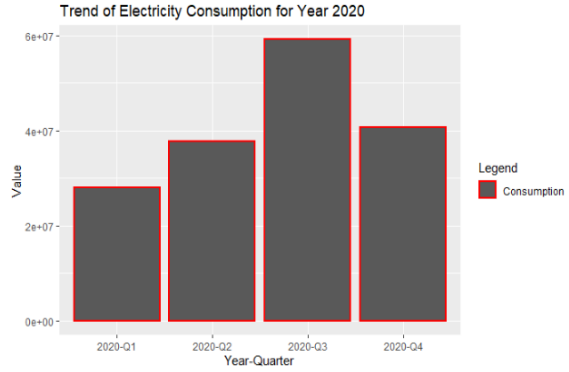

Figure 14: Electricity consumption quarterly for year 2020.

Same graphs are generated for water consumption as well and are attached below. Same trend as observed above could be seen here as well except for year 2019. For 2019, the consumptions are abrupt.

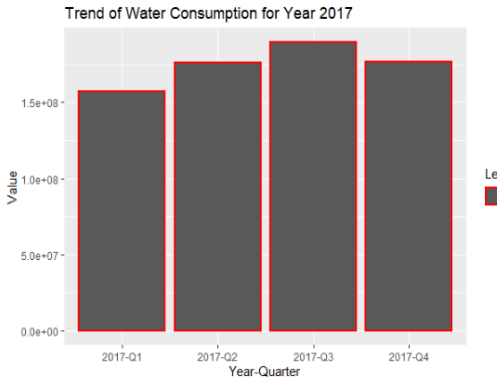

Figure 15: Water consumption quarterly for year 2017.

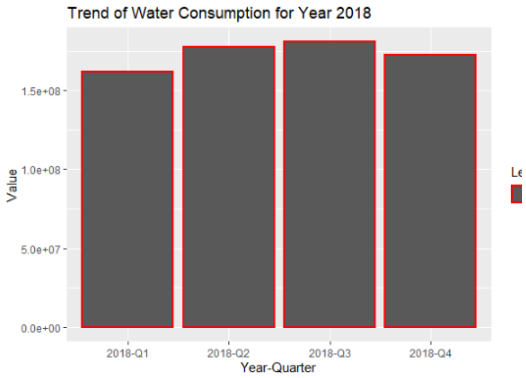

Figure 16: Water consumption quarterly for year 2018.

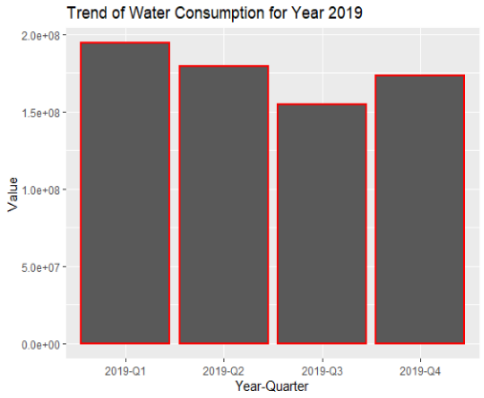

Figure 17: Water consumption quarterly for year 2019.

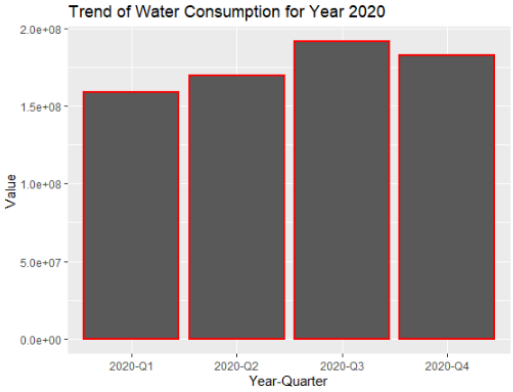

Figure 18: Water consumption quarterly for year 2020.

A graph showing the overall trend from 2017-2021 is attached below for both the categories which are electricity and water consumption. From the trend for electricity consumption, we can see that the lowest consumptions are recorded in first quarter of 2017 and the highest is recorded in quarter 3rd of year 2020. Same for the consumption units of water, the highest consumption for water is recorded in 1st quarter 2019 and lowest in 3rd quarter in 2019.

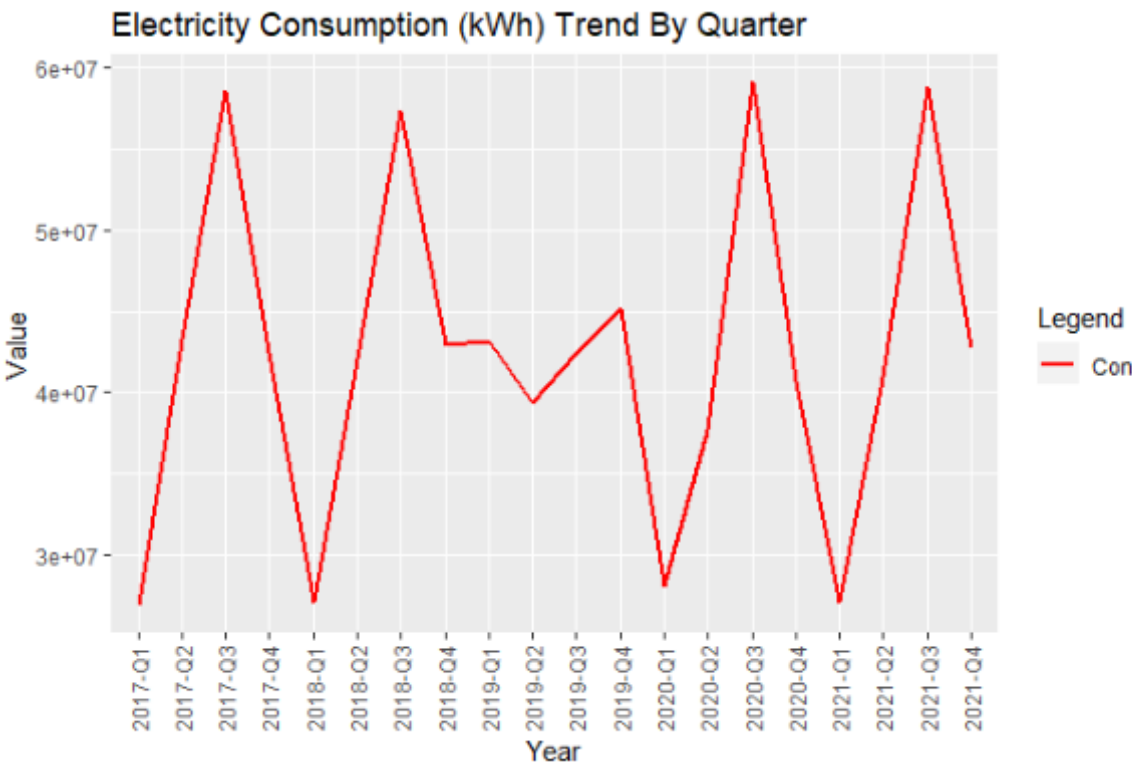

Figure 19: Electricity Consumption trend.

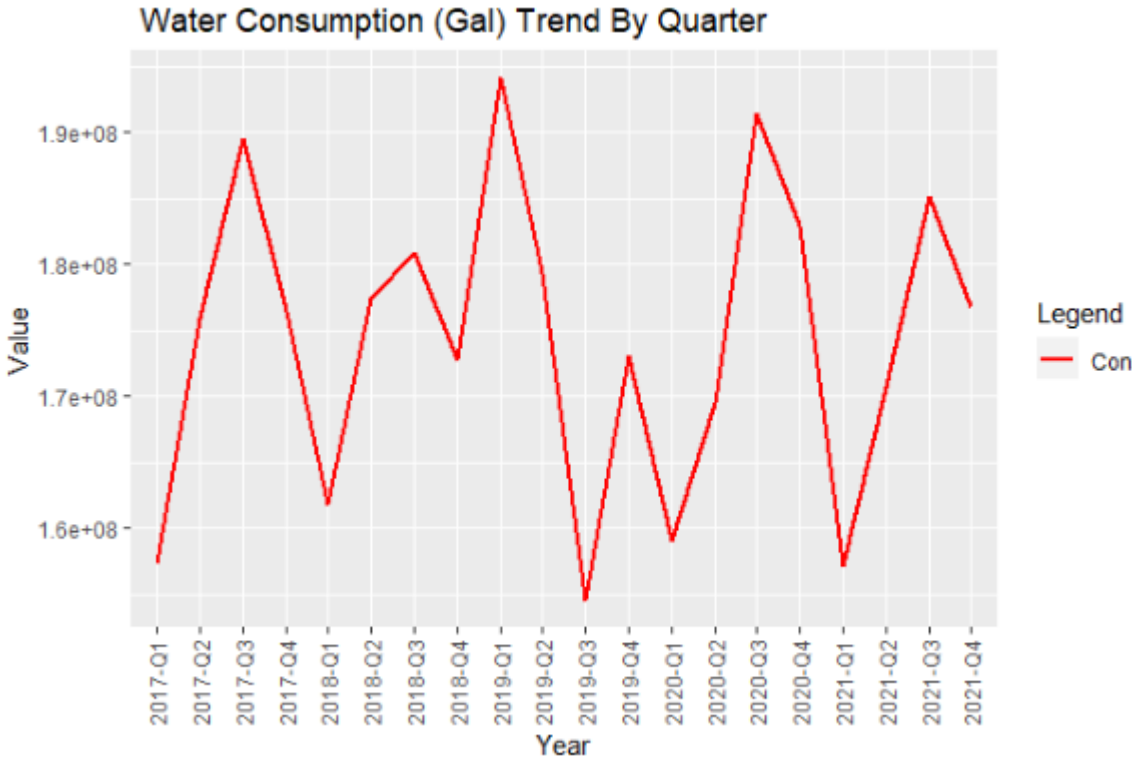

Figure 20: Water Consumption trend.

Below, two other bar graphs are created to show the department wise consumption for water and electricity which are attached below. It can be seen that the consumption of electricity for Forensic department and HQ building is highest while it is lowest for Decision making new department. Similarly, the consumption of water is highest in GHQ shared and lowest in explosive department.

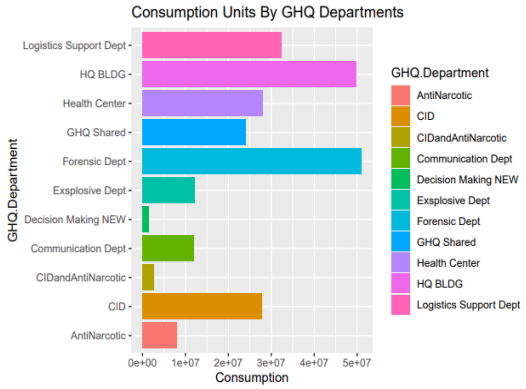

Figure 21: Electricity Consumption for each GHQ Department.

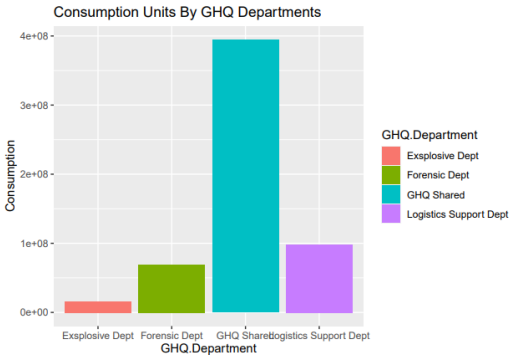

Figure 22: Water Consumption for each GHQ Department.

In the below plots, the total electricity consumption per area and capita based on departments or police stations for the total period of study (2017-2021) is shown. As we can see the highest consumption per area is conducted for Al Faqqa Police Station and the lowest for Dubai Police Academy. On the other hand, the highest consumption per capita was detected for Qusias Warehouses and the lowest for Airport Security.

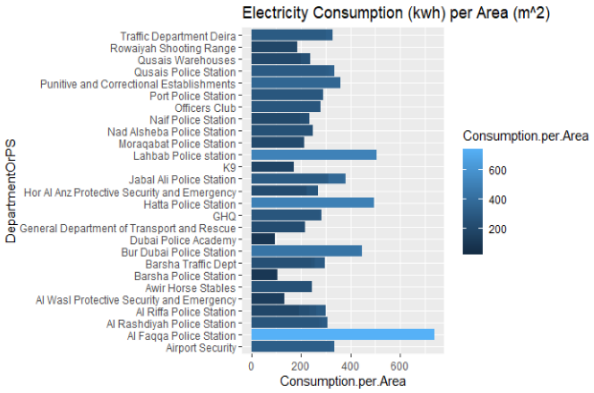

Figure 23: Electricity Consumption per Area for all departments and PSs.

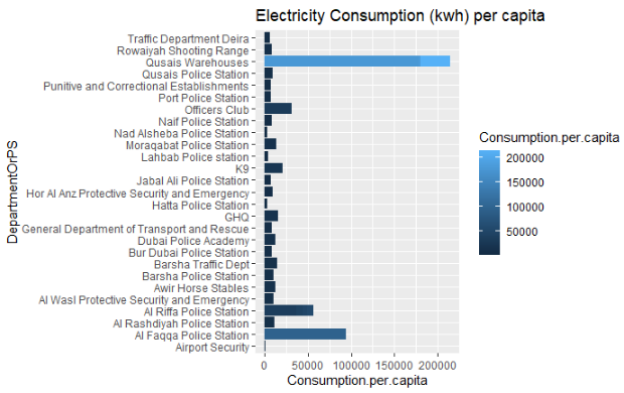

Figure 24: Electricity Consumption per Capita for all departments and PSs.

Same for water consumption, as we can see the highest consumption per area is conducted for Rawiyah Shooting Range, however the lowest for Al Riffa Police Station. In the other hand, the highest consumption per capita was detected for Qusias Warehouses again and the lowest for Airport Security, General Department of Transport and Rescue, and Lehbab Police Station.

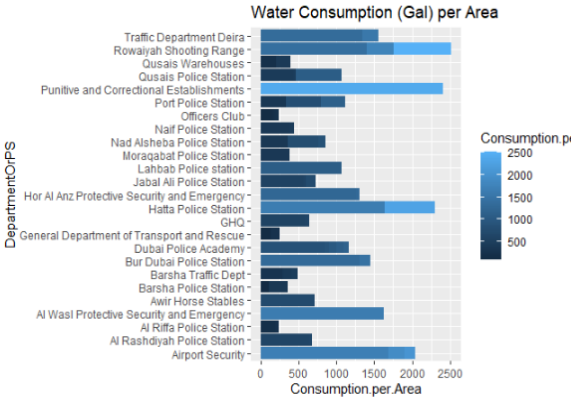

Figure 25: Water Consumption per Area for all departments and PSs.

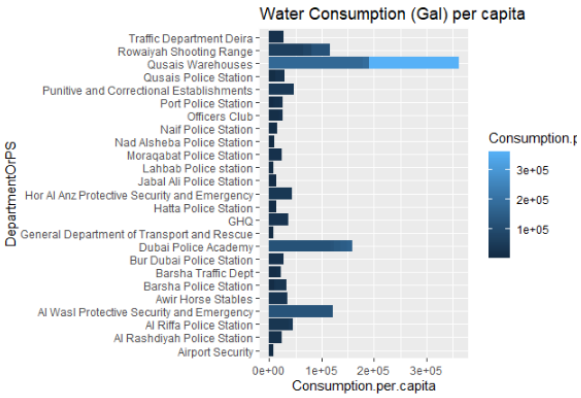

Figure 26: Water Consumption per Capita for all departments and PSs.

Moreover, the below plots showing electricity and water consumption for each department based on their area ranks. For electricity plot, GHQ campus showed the highest consumption in extremely high area facilities section, General Department of Transport and Rescue in high area facilities, Officers Club in medium area facilities, Bur Dubai Police Station in small area facilities, and Jabal Ali Police Station in very small area facilities. However, for water consumption Punitive and Correctional Establishments showed the highest consumption in extremely high area facilities section, Al Wasl Protective Security and Emergency in high area facilities, Traffic Department Deira in medium area facilities, Airport Security in small area facilities, and Hor Al Anz Protective Security and Emergency in very small area facilities. This concludes that these facilities need to be focused on, to detect the reasons of their high consumptions compared with the other locations in the same rank.

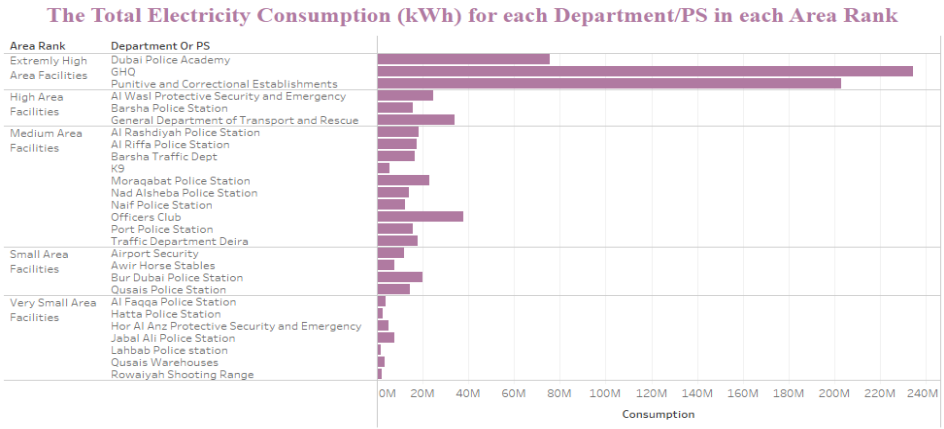

Figure 27: Electricity Consumption of Departments and PSs based on their Area Rank.

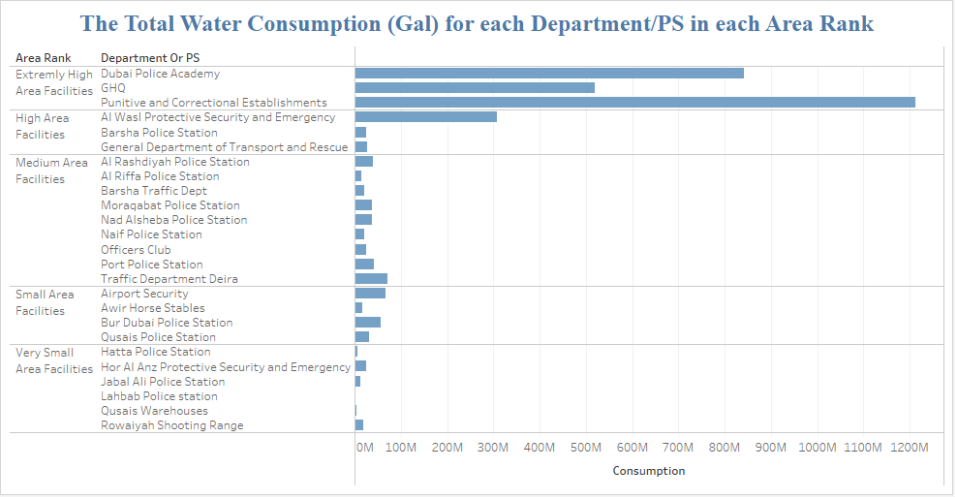

Figure 28: Water Consumption of Departments and PSs based on their Area Rank.

The below plots show the trend of temperature in each Quarter with electricity and water consumption. For electricity, we can see that the trend of consumption is same as temperature for year 2017. However, the consumption is the highest in all years except year 2019 in quarter 3 where the temperature is the highest as well. Same for water consumption plot.

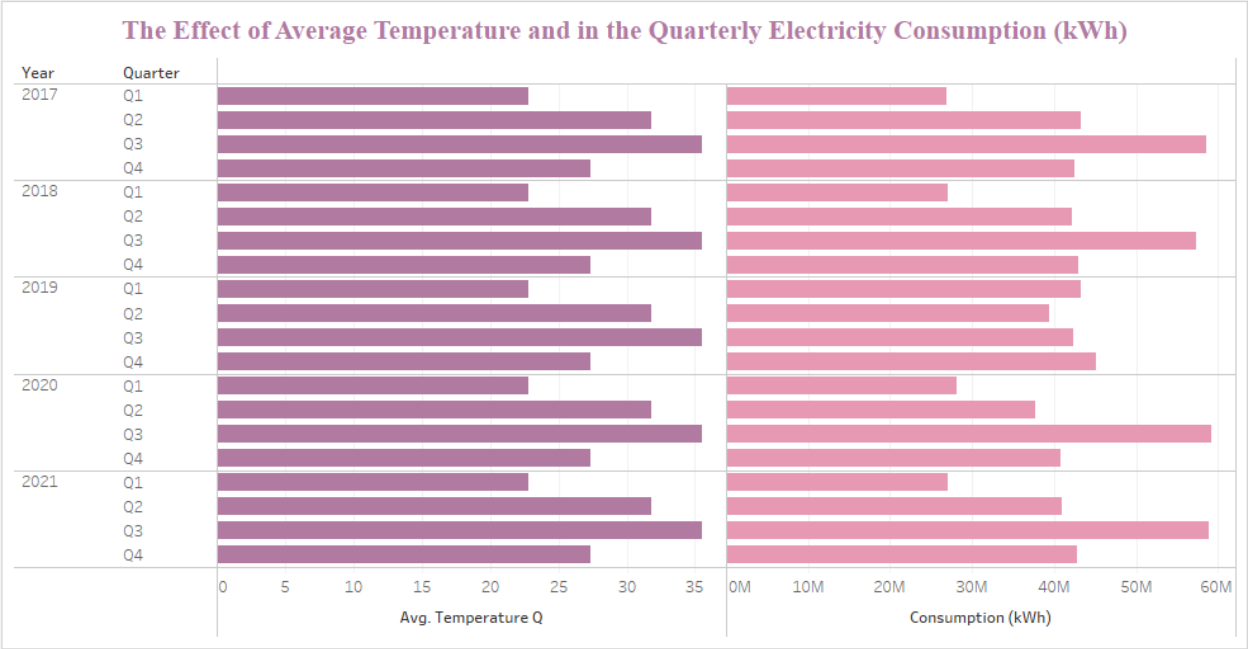

Figure 29: The trend of temperature and electricity consumption each quarter.

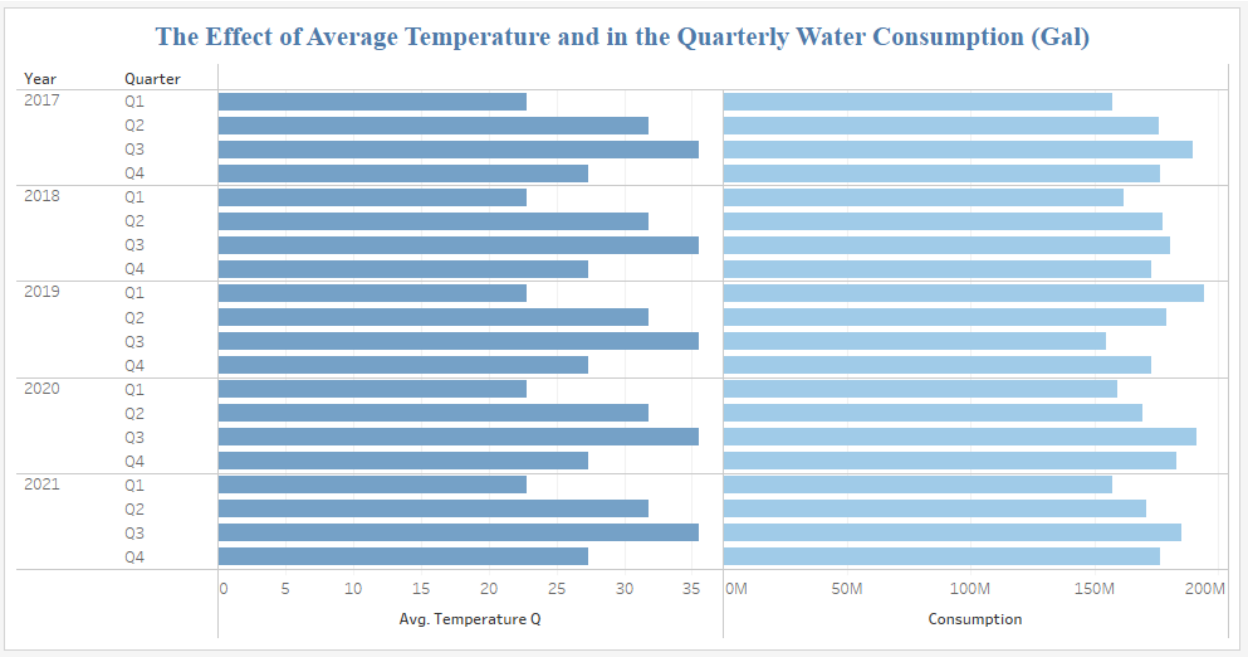

Figure 30: The trend of temperature and water consumption each quarter.

The plots here show a sample for logical comparison per a specific Quarter (Q1) and a facility (Al Riffa Police Station. In electricity consumption, first quarter of year 2019 got the highest and 2021 got the lowest electricity consumption. In the other hand, for water consumption in Q1, year 2020 detected the highest water consumption and 2018 the lowest.

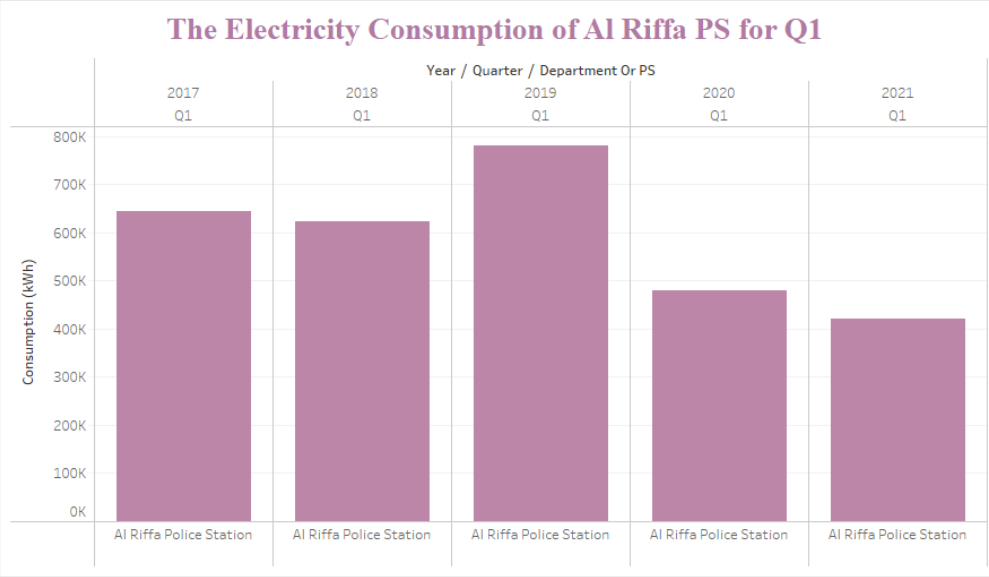

Figure 31: A quarterly report of electricity consumption for Al Riffa PS for the first Quarter.

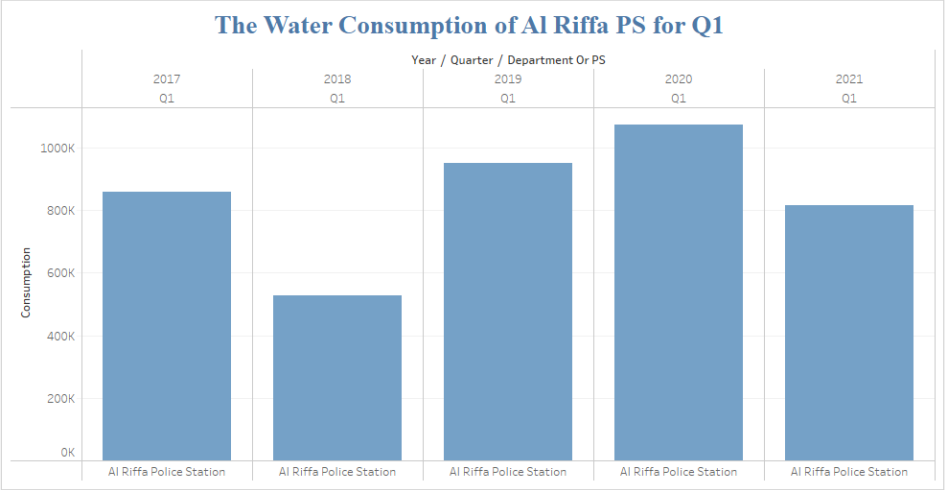

Figure 32: A quarterly report of water consumption for Al Riffa PS for the first Quarter.

This plot below represents the quarterly report that should be published for each department and police station (here is a sample for Airport Security) showing the percentage of saving and loss.

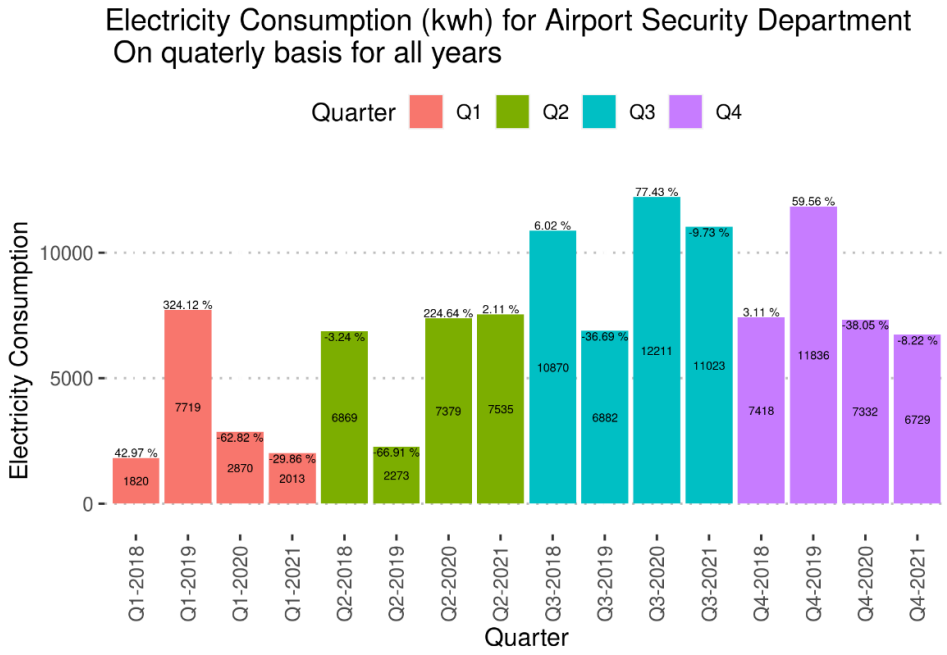

Figure 33: Electricity Consumption for Airport Security Department for each quarter.

The below plots show the total electricity consumption quarterly based on area rank. We can see the high difference in the consumption of extremely high area facilities compared with other ranks. This means if we need to see a noticeable saving in electricity bills. It is enough to concentrate on reducing consumption in extremely high area facilities.

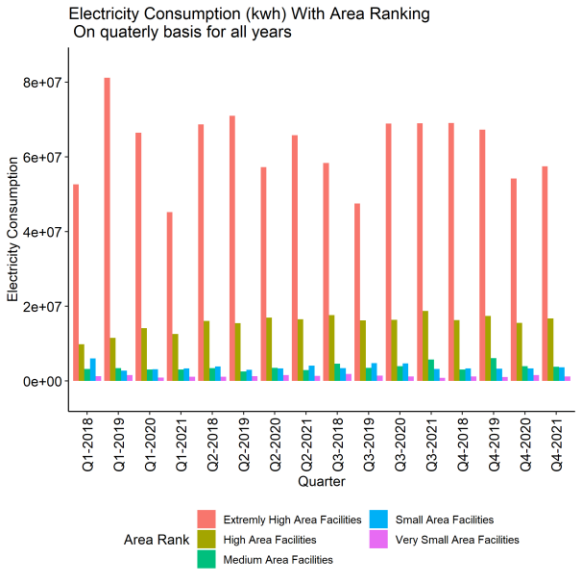

Figure 34: Quarterly Electricity Consumption based on Area Ranking.

4.8 Data Separation

The train-test split procedure is used to evaluate the performance of machine learning algorithms that make predictions on data that was not used to train the model. It's a fast and easy way for comparing the performance of various machine learning algorithms for your predictive modeling task.

We end up testing and training our model on the same data if we don't divide the dataset into training and testing sets. When we test on the same data that we used to train our model, we usually receive decent results. However, this does not imply that the model will perform as well on data that has not been observed. In the domain of machine learning, this is known as overfitting.

As our dataset is ready now it’s ready to run the model. However, before that, data separation is needed. The dataset will be splatted in a split ratio of 70:30 which means 70% of the data will be training set and 30% of the dataset will be testing set. The following code is used for this process:

```
library(caTools)
set.seed(123)
index<-sample.split(base_electricity$Consumption,SplitRatio=0.70)
Train<- subset(base_electricity,index==TRUE)  #357  observations
Test<-subset(base_electricity,index==FALSE)  #161 observations
```

4.9 Data Modeling

This project will study the dataset in Multiple linear regression model and ARIMA model. The two models will be discussed based on the following performance evaluation metrics:

- ME:

The mean error (ME) is calculated by adding the variances and dividing the result by n.

Equation 1: Mean Error

$$ME = \frac{Sum\ of\ All\ Errors}{Number\ of\ Observations}$$

- RMSE:

The Root Mean Squared Error (RMSE) is an unusual measurement, but extremely useful one. it is calculated by taking the square root of the average of summation of all squared error.

Equation 2: Root Mean Squared Error.

$$\text{RMSE}= \sqrt{\frac{1}{n}\sum e_t^2}$$

- MAE:

The Mean Absolute Error (MAE) is an excellent metric for determining forecast accuracy. It is calculated by finding the mean of the absolute error.

Equation 3: Mean Absolute Error.

$$\text{MAE} = \frac{1}{n}\sum |e_t|$$

- MPE:

This metric refers to the Mean Percentage Error which is equal to the average of percentage errors by which model projections differ from actual values of the quantity being forecasted in statistics.

Equation 4: Mean Percentage Error

$$\text{MPE} = \frac{100\%}{n}\sum_{t-1}^{n}\frac{a_t - f_t}{a_t}$$

a_t is the actual value, f_t is the forecasted value, and n is the number of times the variable will be forecasted.

- MAPE:

One of the most prominent methods for determining forecast accuracy is the Mean Absolute Percentage Error (MAPE). MAPE is the total of all absolute errors divided by the demand (each period separately).

Equation 5: Mean Absolute Percentage Error

$$\text{MAPE} = \frac{1}{n}\sum\frac{|e_t|}{d_t}$$

- MASE:

The (one-period-ahead) forecast error divided by the average forecast error of the naive method yields the mean absolute scaled error (MASE) of a single data point. The MASE can be used to assess forecast methods on a single series as well as between series to compare forecast accuracy.

Equation 6: Mean Absolute Scaled Error

$$\text{MASE} = \frac{1}{n}\sum_{t=1}^{n}|q(t)|$$

4.9.1 Multiple Linear Regression Model

Multiple linear regression (MLR) is a statistical technique that predicts the result of an independent variable by combining numerous dependent variables. The linear relationship between explanatory (independent) and response (dependent) variables is proposed to be represented using multiple linear regression.

In this step, I have used a regression model to build a machine learning model which will predict the consumption units based on the input data for electricity. As I'm trying to predict a continuous value in this case, hence linear regression model is one of the best choices in this case. The model is a multiple regression model as it will involve multiple predictors to predict a response variable. The summary of the model couldn't be pasted in this case here because our predictors are large in number and we made two models, one for electricity and one for water. However, I will discuss the significance of the model and the performance of our model. Two models will be done, one using the electricity dataset and the other using the water dataset. As combing them together will be not logical because the unit of the out dependent variable (Consumption) is not identical.

4.9.1.1 Electricity

In this step, we have used a regression model to build a machine learning model which will predict the consumption units based on the input data for electricity. For electricity, we got a huge number of significant variables which have a value less than the significance level alpha = 0.05. The R square of our model is 99.82%, which means that the model was able to predict 99.82% variability in the dataset. Moreover, in simple terms, we can say that the regression model was able to cover 98.9% data points in the dataset. The overall p-value (less than 2.2×10^{-16}) of the model is less than significance level alpha = 0.05, hence we can say that the overall model is significant. The AIC (9915.891) and the residuals of the model (131800) are high.

```
Call:
lm(formula = Consumption ~ ., data = Train)

Residuals:
    Min      1Q  Median      3Q     Max
-477939  -21112    6575   37603 1062943

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.604e+07  9.837e+06  -1.631  0.10382
Year                   7.952e+03  4.872e+03   1.632  0.10354
Quarter               -8.855e+03  6.678e+03  -1.326  0.18564
DepartmentOrPS        -4.077e+03  8.918e+02  -4.572 6.62e-06 ***
Capita                 3.424e+01  1.066e+01   3.213  0.00143 **
`Area (meter square)`  1.721e+00  2.936e-01   5.862 1.02e-08 ***
Amount                 2.271e+00  1.505e-02 150.970  < 2e-16 ***
Temperature_Q          1.664e+03  1.603e+03   1.038  0.30007
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 131800 on 367 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9982
F-statistic: 2.89e+04 on 7 and 367 DF,  p-value: < 2.2e-16
```

The multiple linear equation is:

Equation 7: Electricity Linear Regression

$$Consumption = -1.604\times10^{7} + 7.952\times10^{3}\ Year - 8.855\times10^{3}\ Quarter - 4.077\times10^{3}\ DepartmentOrPS + 3.424\times10^{1}\ Capita + 1.721\ Area + 2.271\ Amount + 1.664\times10^{3}\ Temperature\_Q$$

The residuals show a linear trend, the points deviate at the tails, indicating that the data were not normally distributed, the condition of equal variance is not violated because the scale location line is straight enough, and a few of influential observations have been identified. Concluding this section of modeling, the model is a better-fitted model when it comes to the R Square as the predictors were able to explain almost 98.9% variability in the dataset.

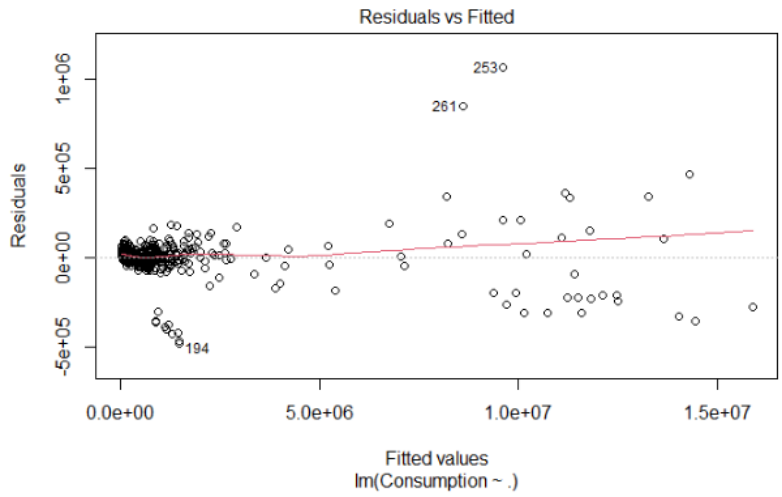

Figure 35 : Residuals versus fitted values for electricity linear regression model.

In the first plot “Residuals vs Fitted”, the red line (which is a scatterplot smoother, displaying the average value of the residuals at each value of the fitted value) is almost flat. This indicates that the residuals have a detectable linear tendency. Furthermore, across the whole range of fitted values, the residuals appear to be unequally varied. There's evidence of non-constant variation.

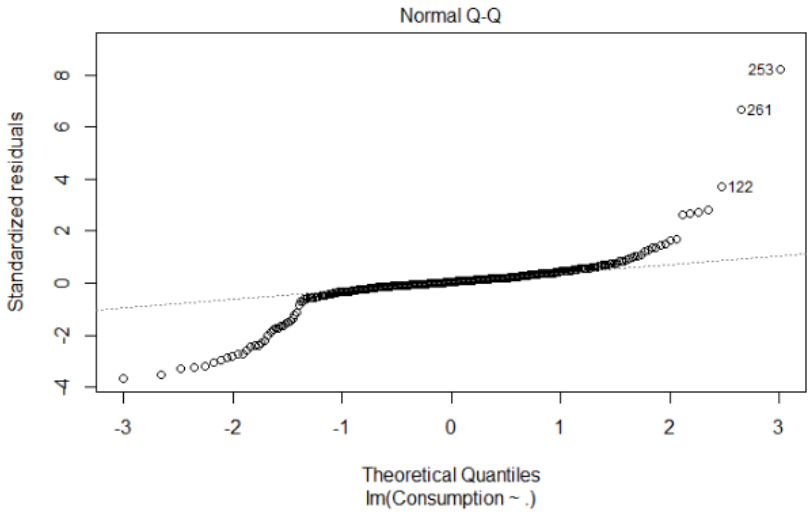

Figure 36: QQ plot for electricity linear regression model.

A standard QQ plot is good if the residuals are properly lined up on the straight dashed line. In both the higher and lower tails of the QQ plot, the residuals depart from the diagonal line. We can see that the tails are 'heavier' (have higher values) than we would predict based on the

typical modeling assumptions. This is represented by the points making a "steeper" line than the diagonal.

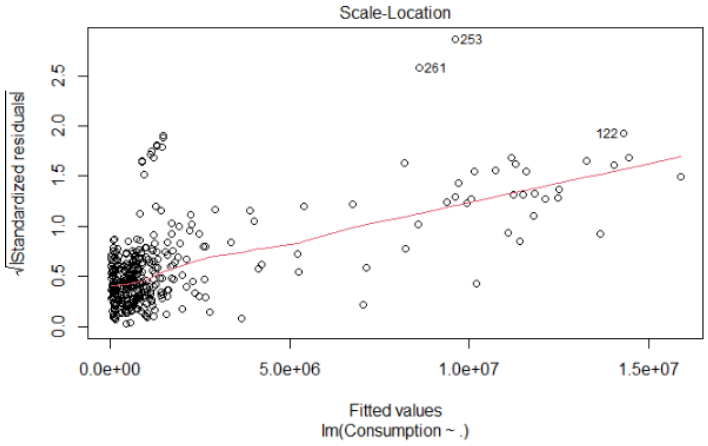

Figure 37: Spread-Location graph for electricity linear regression model.

The Spread-Location graph reveals if residuals are distributed evenly across predictor ranges. This is how you can test the equal variance assumption (homoscedasticity). If you notice a horizontal line with evenly (randomly) spaced points, that's a positive sign. The scale location plot indicates some non-linearity, although the dispersion of magnitudes appears to be greatest in the fitted values close to 0 and less than 3×10^6, lower in the fitted values greater.

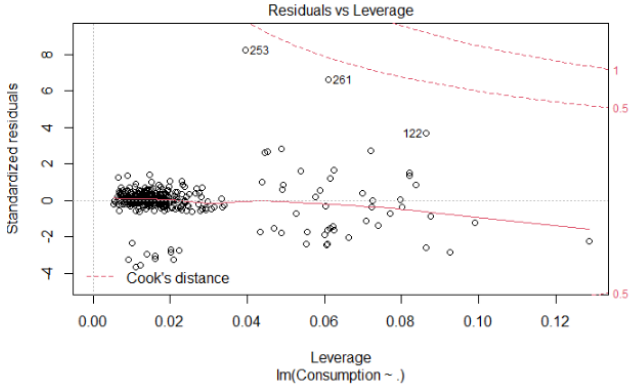

Figure 38: Residuals versus leverage plot for electricity linear regression model.

Residuals versus leverage plot is unlike the others as patterns are irrelevant. Outlying values in the upper or lower right corner should be avoided. Cases can have an impact on a regression line at such points. Outside of a dashed line, Cook's distance. When cases go outside of the Cook's distance, the regression findings are influenced. If we exclude certain instances, the regression results will improve ("Understanding Diagnostic Plots for Linear Regression Analysis", 2015). In this model, the residuals appear to be concentrated on the left. #253 and #261 could be

recognized as the influential observation by the plot as it's neat to cook distance but still it is fine as it's in the acceptable range. Although, if I exclude these two observations from the analysis, the slope coefficient and R2 will change positively.

Error Table:

Table 12 : Error table for electricity linear regression model.

Department.or.Police.Station <chr>	Year <int>	Quarter <dbl>	Actual.electricity.consumption <dbl>	Predicted.electricity.consumption <int>	Error <int>
Al Rashdiyah Police Station	2021	1	500166.8	517971	17804
Al Rashdiyah Police Station	2021	2	896389.6	909946	13556
Al Rashdiyah Police Station	2021	3	1256534.2	1258940	2405
Airport Security	2021	2	603000.0	671126	68126
Al Faqqa Police Station	2021	1	179377.6	195322	15944
Al Faqqa Police Station	2021	2	265601.6	280420	14818
Al Faqqa Police Station	2021	3	368334.4	378003	9668
Naif Police Station	2021	4	570665.6	580192	9526

Table 11 shows the actual and predicted electricity consumption (Quarterly) as well as the error. It shows a good result as the error is not that high compared with consumption values.

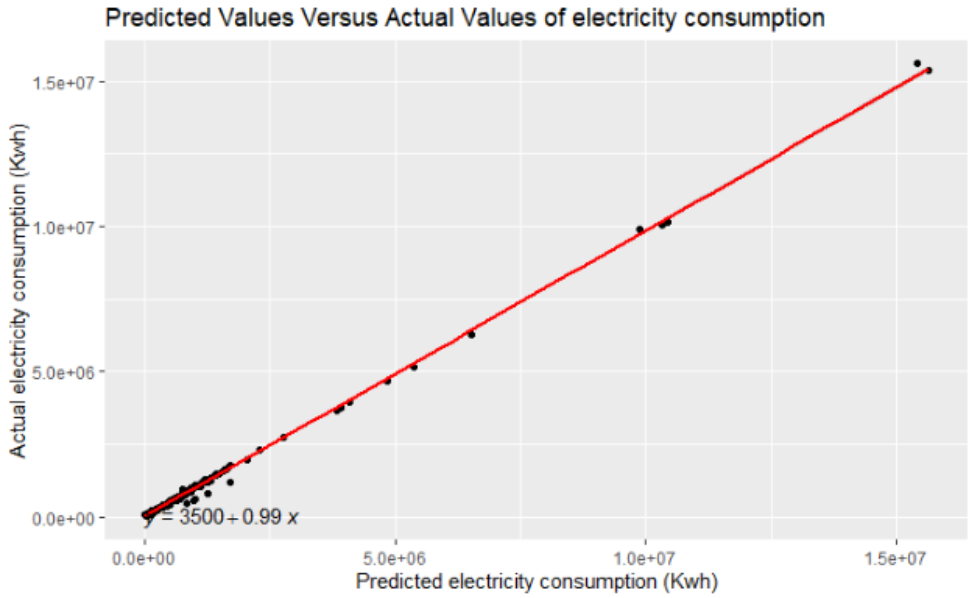

Figure 39: Predicted versus actual values for electricity linear regression model.

The model was used to verify the accuracy of the proposed multiple linear regression model in examining the link between predicted and actual consumption. As shown almost all the actual values are fitted in the predicted line which gives us a good result for the model.

4.9.1.2 Water

The same was done for water consumption prediction, we got a huge number of significant variables which have a value less than the significance level alpha = 0.05. The R square of our model is 99.86%, which means that the model was able to predict 99.86% variability in the dataset. Moreover, in simple terms, we can say that the regression model was able to cover 98.9% data points in the dataset. The overall p-value (is less than 2.2×10^{-16}) of the model is less than significance level alpha = 0.05, hence we can say that the overall model is significant. The AIC (10231.53) and the residuals of the model (554200) are high.

```
Call:
lm(formula = Consumption ~ ., data = Train_W)

Residuals:
     Min       1Q   Median       3Q      Max
-2030206   -80512    -7525    69308  7106751

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -8.779e+07  4.249e+07  -2.066   0.0396 *
Year                   4.344e+04  2.104e+04   2.064   0.0397 *
Quarter               -2.753e+04  2.909e+04  -0.946   0.3446
DepartmentOrPS         9.413e+02  3.946e+03   0.239   0.8116
Capita                 1.307e+02  4.646e+01   2.812   0.0052 **
`Area (meter square)`  1.298e+00  1.093e+00   1.188   0.2358
Amount                 1.950e+01  1.007e-01 193.640   <2e-16 ***
Temperature_Q          4.058e+03  6.802e+03   0.597   0.5512
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 554200 on 341 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9986
F-statistic: 3.444e+04 on 7 and 341 DF,  p-value: < 2.2e-16
```

The multiple linear equation is:

Equation 8: Water Linear Regression

$$Consumption = -8.779\times10^{7} + 4.344\times10^{4}\ Year - 2.753\times10^{4}\ Quarter + 9.413\times10^{2}\ DepartmentOrPS + 1.307\times10^{2}\ Capita + 1.298\ Area + 19.5\ Amount + 4.058\times10^{3}\ Temperature\_Q$$

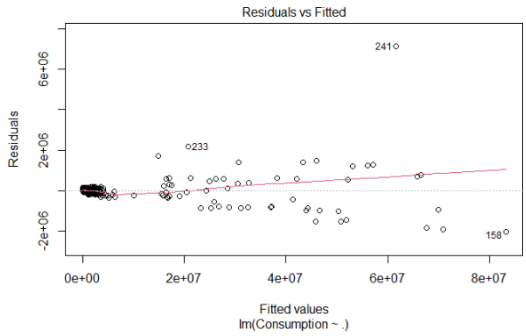

Figure 40: Residuals versus fitted values for water linear regression model.

We can see that residuals appear to be unequally varied as in electricity consumption. There's evidence of non-constant variation.

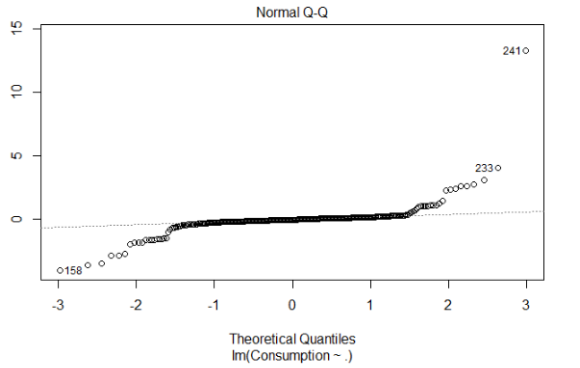

Figure 41: QQ plot for water linear regression model.

We can see that the tails are 'heavier' than we would predict based on the typical modeling assumptions. This is represented by the points making a "steeper" line than the diagonal.

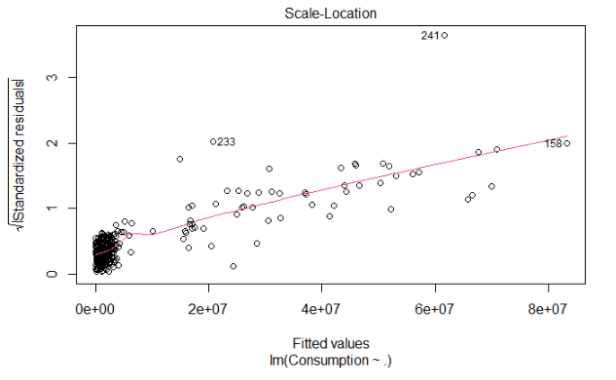

Figure 42: Spread-Location graph for water linear regression model.

The scale location plot indicates non-linearity, although the dispersion of magnitudes appears to be greatest in the lowest fitted values and almost nothing for higher values. As it shows a steep angle which means it's not horizontal at all.

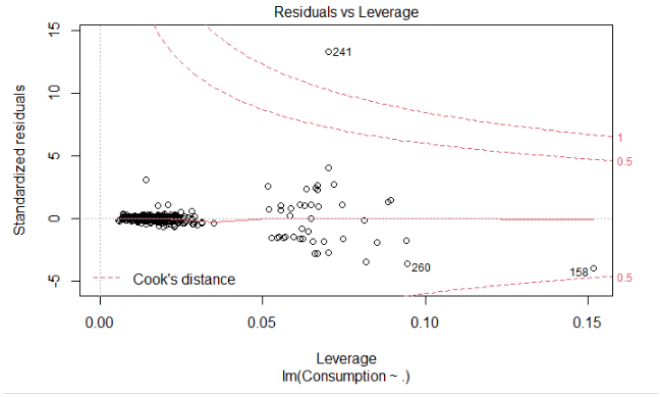

Figure 43: Residuals versus leverage plot for water linear regression model.

In this model, the residuals appear to be concentrated from leverage 0.01 to almost 0.04 was recognized as an influential observation by the plot. Observation 241 is an outlying value in this model.

Error Table:

Table 13: Error table for water linear regression model.

Department.or.Police.Station <chr>	Year <int>	Quarter <dbl>	Actual.Water.consumption <dbl>	Predicted.Water.consumption <int>	Error <int>
Qusais Horse Stables	2021	3	4667520	4701455	33935
Bur Dubai Police Station	2021	1	2648580	2722530	73950
Bur Dubai Police Station	2021	3	3423420	3474448	51028
Bur Dubai Police Station	2021	4	3063940	3063238	702
Port Police Station	2021	4	869880	923875	53995
GHQ	2021	4	26235000	26119560	115440
Nad Alsheba Police Station	2021	4	1109020	1216231	107211

A table was built to show the (Quarterly) actual and predicted water consumption as well as the error. A sample of the water consumption data is presented in Table 13. It doesn't show bad results as the error is not that high compared with consumption values.

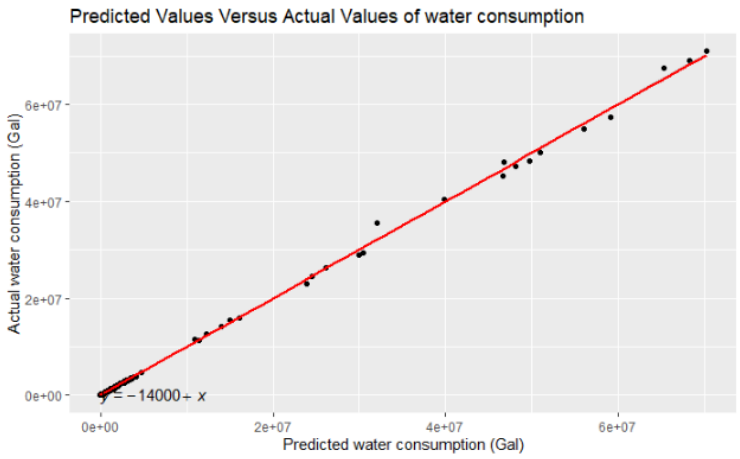

Figure 44: Predicted versus actual values for water linear regression model.

Here as well it showed almost all the actual water consumption values are fitted in the predicted line which gives us a good result for the model.

4.9.1.4 Accuracy Table

Table 14: Accuracy Table for electricity and water models.

Accuracy.Table	Electricity.model	Water.model
ME	0.000000e+00	0.000000e+00
RMSE	1.303853e+05	5.477769e+05
MAE	6.707942e+04	2.103196e+05
MPE	2.463195e+00	1.752813e+00
MAPE	9.976469e+00	8.440297e+00
MASE	3.625120e-02	2.250670e-02

The accuracy parameters was calculated using accuracy() function in R for both models based on the test dataset. The table above shows a summary of the two models results of the six measurements. I will discuss here the RMSE and MAPE values. For the electricity regression model the RMSE and MAPE are 130385.3 kWh and 11.89% respectively. In the other hand, for the water model they are equal to 547776.9 Gallons and 9.53% respectively. both models gave us a good result.

All of these outputs provide information about your model and data. The existing model may not be the most effective approach to comprehend the information we have. That's why I build a time series model to check if we will get a better result.

4.9.2 ARIMA Model

ARIMA stands for autoregressive integrated moving average, and it's a statistical analysis model that employs time series data to better understand the data set or anticipate future trends.

An autoregressive integrated moving average model determines how powerful one dependent variable is in contrast to other variables associated. The model's goal is to predict future value of the independent variable by analyzing differences between values in a series rather than actual values.

Components of the ARIMA model:

- Autoregression AR: A model that displays a changing variable regressing on its own lagged.
- Integrated (I): denotes the differencing of raw observations to allow the time series to stabilize.
- Moving average (MA): A moving average model applied to lagged observations incorporates the dependency between an observation and a residual error.

A univariate ARIMA time series model for water and electricity consumption was done. As we have chosen to forecast each department forecast, it is wiser to only consider a single variable-based ARIMA model.

The steps to prepare data for the ARIMA model:

As I'm considering groups, I prepared data differently than a single ARIMA model. The first step is to convert data to time series, for which I have used the (ts) function. As data is collected from 2017 to 2022, and data is aggregated quarterly, I only need to provide frequency and year to create a date column.

As for turning data into a time series format, there are many ways to create time series from data. However, the most convenient way is to use the ts() function in R. The ts function is like this:

ts(data, start_date, end_date, frequency)

```
listed_ts <- lapply(listed,
                    function(x) ts(x[["Consumption"]], start = c(2017, 1), frequency = 4) )

dat <- do.call(cbind, listed_ts)

train <- window(dat, start = 2017, end = 2020.9)
test <- window(dat, start = 2021, end = 2022)
```

As you can see, in this format, we can create time-series data easily.

For quarterly data, we will do this:

ts(data, start_date, end_date, freq=4), which indicates that there will be 4 observations for each year. After converting to time series, now I need to concatenate all-time series row-wise, so I can apply time series more efficiently and fast way. Then, Split data into train and test. Train data contains data from year 2017 to 2020 data and test data is 2022 data. All as shown in the screenshot above.

```
# 1.Own functions for forecasting
FORECASTING_FUNCTION_ARIMA <- function(z, hrz = 4) {
  timeseries <- msts(z, start = 2017, seasonal.periods = 4)
  forecast <- auto.arima(timeseries)
  #ic = c("bic")
}
FORECASTING_LIST_ARIMA <- lapply(X = train, FORECASTING_FUNCTION_ARIMA)

ACCURACY_ARIMA <- Map(function(x, y) accuracy(forecast(x, h = 4),
        x = test[, y]), FORECASTING_LIST_ARIMA, seq_len(ncol(test)))


# Plot forecasts and data
lapply(FORECASTING_LIST_ARIMA, function(x) plot(forecast(x, h = 4)))
```

After this step, I create my function, which first considers seasonality in data (msts function) and then passes the series to auto.arima() function, which is a time series model. The benefit of using auto.arima is that it will automatically infer the auto-regressive and moving average part of a series and we don't need to find specific order for every series, which is impractical when considering different groups. Then, apply ARIMA function to all our groups. For that, I use the lapply() function, which splits data into groups and then applies the required function. Now that the ARIMA model is built, we can forecast future values and then look at the accuracy table for each group.

4.9.2.1 Results

4.9.2.1.1 Electricity Consumption ARIMA Model

When we forecast with time series models, we receive three values for each observation: Mean, Low, and high. The Mean value is the actual forecast. The low is the confidence interval of that forecast doing a downward trend, and the high is the confidence interval of that forecast going upward. We got a result for each department and police station. As shown for example in protective security emergency department electricity consumption model.

```
$`Al Wasl Protective Security and Emergency`
$`Al Wasl Protective Security and Emergency`$mean
          Qtr1      Qtr2      Qtr3      Qtr4
2021  584176.2  913777.2 1649981.5 1140666.2

$`Al Wasl Protective Security and Emergency`$lower
                80%        95%
2021 Q1   37236.31 -252296.17
2021 Q2  366837.31   77304.83
2021 Q3 1103041.61  813509.13
2021 Q4  593726.31  304193.83

$`Al Wasl Protective Security and Emergency`$upper
            80%     95%
2021 Q1 1131116 1420649
2021 Q2 1460717 1750250
2021 Q3 2196921 2486454
2021 Q4 1687606 1977139
```

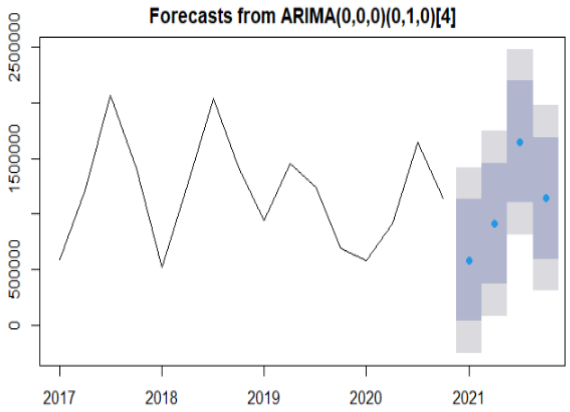

Figure 45: Protective Security Emergency Department Electricity Consumption ARIMA Model forecast output.

The forecast shows that for most of the groups, there is no clear seasonality or trend, which results in constant mean predictions for these groups (as ARIMA models are mean models). However, some groups average changes each quarter, so we can also see predictions accommodate that seasonality. The forecasts accuracy depends on past values as its essence of time series model). In plots, we can see that some locations have cyclical pattern and some locations have no clear trend. This can also be seen in forecasts. When ARIMA found clear trend plus seasonality. the forecasts are good. But when observations have no pattern, the predictions follow same mean for all forecasts. The visualizations with no seasonal pattern and no clear trend, results in poor predictions.

The below table shows a sample for some location of the output of actual 2021 electricity consumption versus forecasted one. (The full table in Appendix 2)

Table 15: Actual and Forecasted Electricity Consumption of each department.

Department <chr>	Date <fctr>	Actual <dbl>	Forecast <dbl>
Airport Security	2021 Q1	359000.0	595497.20
Airport Security	2021 Q2	603000.0	595497.20
Airport Security	2021 Q3	815000.0	595497.20
Airport Security	2021 Q4	565000.0	595497.20
Al Faqqa Police Station	2021 Q1	179377.6	164534.27
Al Faqqa Police Station	2021 Q2	265601.6	271763.07
Al Faqqa Police Station	2021 Q3	368334.4	393742.27
Al Faqqa Police Station	2021 Q4	227204.8	263812.67
Al Rashdiyah Police Station	2021 Q1	500166.8	935465.55
Al Rashdiyah Police Station	2021 Q2	896389.6	935465.55

1-10 of 108 rows Previous 1 2 3 4 5 6 ... 11 Next

The accuracy table (test dataset):

Table 16: Accuracy table for ARIMA Model using Electricity dataset.

X1	ME	RMSE	MAE	MPE	MAPE	MASE
Airport Security	-9997.2	162094.7	123500	-10.7743	24.86287	0.692621
Al Faqqa Police Station	-13333.5	23685.35	20755.13	-4.26381	8.401261	0.397936
Al Rashdiyah Police Station	-62174.5	275316.5	222708.8	-19.2983	32.07427	0.946821
Al Riffa Police Station	122374.3	130350.3	122374.3	18.62394	18.62394	0.607273
Al Wasl Protective Security and Emergency	257388.9	302571.3	257832.8	16.49376	16.56988	0.761672
Awir Horse Stables	18513.6	196850.4	162481.2	-35.8172	65.30971	0.648833
Barsha Police Station	116626.9	139544.5	116626.9	10.8076	10.8076	0.800051
Barsha Traffic Dept	-47175	94098.79	72075	-6.2182	9.614732	0.507183
Bur Dubai Police Station	4789.175	315711.5	239366	-11.4122	28.58718	0.785473
Dubai Police Academy	17524.7	242972.2	210626.5	-1.05575	7.078043	0.4416
General Department of Transport and Rescue	90213.56	633711.2	483849.2	-10.5519	33.64712	0.891438
GHQ	317897.7	488301.4	437258.9	2.279393	3.644494	0.280705
Hatta Police Station	28103	29708.24	28103	11.37193	11.37193	2.290601
Hor Al Anz Protective Security and Emergency	28559.7	34924.25	28559.7	14.17428	14.17428	1.070166
Jabal Ali Police Station	-65999.4	107802.8	97301.57	-28.3468	35.15951	1.12214
Lahbab Police station	7435	41055.75	35060	-15.6129	46.87774	0.79047
Moraqabat Police Station	-47860.5	447211.6	372934.4	-15.2408	31.63697	0.647359
Nad Alsheba Police Station	37809.2	53501.88	37809.2	4.184984	4.184984	0.160408
Naif Police Station	14976	67744.99	62477.6	4.029646	11.59976	0.720109
Officers Club	105389.9	552249.2	442396.6	-3.23456	24.88511	0.794511
Port Police Station	13641.6	36492.87	24116.4	2.005363	2.904569	0.335214
Punitive and Correctional Establishments	183877.2	595027.1	515411.6	2.395666	4.710647	0.307627
Qusais Police Station	-47518.5	233881.8	194878.9	-19.8144	34.03138	0.871139
Qusais Warehouses	98834.9	118602	98834.9	33.81088	33.81088	1.378979
Rowaiyah Shooting Range	6323.45	17986.7	14392.7	3.447398	16.24954	0.367868
Traffic Department Deira	-65116	128666.2	90818.75	-6.85111	9.807892	0.298014

I will discuss 3 Metrics which are considered standard: MSE, RMSE, and MAPE. The metrics are difference between actual and predicted values in terms of Mean Square Error and Mean percentage error (MAPE).

If I take Al Faqqa Police Station as example (Electricity consumption prediction)

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	22.76526	54413.77	38256.25	-8.627517	22.754437	0.7334835	0.2834065	NA
Test set	-13333.46667	23685.35	20755.13	-4.263807	8.401261	0.3979362	0.2586741	0.1957359

The RMSE metrics shows that for training set, error margin was 54413.77 units (predictions can be off with this difference (+/-), and for test set, this metric is 23685.35 which is much better.

Same is true for MSE, which is similar to RMSE. As for MAPE it shows percentage difference between predictions and actual values. In this case the training set has larger MAPE (22.75) than test set (8.4).

So, it means that predictions are not that bad this will be considered small error margin. Some will be larger and others will be even smaller than this. The actual and prediction values was shown in previous section. The average MAPE calculated for all location in the model to be 20.02.

4.9.2.1.2 Water Consumption ARIMA Model

Same steps were done for water consumption dataset. The actual and forecasted consumption table is in Appendix 3. Here is the accuracy table (Average MAPE = 36.9164):

Table 17: Accuracy table for ARIMA Model using Water dataset.

X1	ME	RMSE	MAE	MPE	MAPE	MASE
Airport Security	-157433.65	1373734.99	1044350.91	-13.51	34.72	0.65
Al Rashdiyah Police Station	-10419.94	190278.19	183865.00	-1.43	9.29	0.25
Al Riffa Police Station	-36019.19	83591.94	72710.00	-6.21	10.75	0.30
Al Wasl Protective Security and Emergency	1179090.00	1589061.35	1378190.00	7.53	8.77	1.23
Awir Horse Stables	302445.00	321598.71	302445.00	25.68	25.68	2.04
Barsha Police Station	-485430.00	514676.49	485430.00	-31.11	31.11	0.33
Barsha Traffic Dept	-278048.60	500714.55	416059.82	-69.43	80.12	0.95
Bur Dubai Police Station	403865.00	488301.47	403865.00	12.57	12.57	0.99
Dubai Police Academy	-5200525.00	6521624.80	5200525.00	-15.94	15.94	0.78
General Department of Transport and Rescue	-355657.69	385046.18	355657.69	-39.46	39.46	0.54
GHQ	901197.08	1453849.07	1149841.04	3.10	4.18	0.84
Hatta Police Station	-1035705.00	1165528.98	1035705.00	-201.02	201.02	3.64
Hor Al Anz Protective Security and Emergency	-310475.00	324684.62	310475.00	-21.07	21.07	0.78
Jabal Ali Police Station	1919.06	217190.55	179660.47	-9.99	28.36	0.80
Lahbab Police station	90426.25	112106.57	90426.25	37.21	37.21	1.80
Moraqabat Police Station	105700.33	159805.05	144168.19	3.98	5.64	0.18
Nad Alsheba Police Station	-895588.76	903911.37	895588.76	-83.82	83.82	1.12
Naif Police Station	314270.00	340678.26	314270.00	27.93	27.93	1.24
Officers Club	-331760.00	340243.18	331760.00	-35.53	35.53	1.07
Port Police Station	618915.00	985050.12	823075.00	72.59	91.19	1.09
Punitive and Correctional Establishments	-362340.00	1787738.65	1584110.00	-0.59	2.30	0.14
Qusais Police Station	10010.00	88093.04	81400.00	0.26	7.98	0.11
Qusais Warehouses	-81309.00	83120.20	81309.00	-30.81	30.81	0.37
Rowaiyah Shooting Range	85690.00	95336.06	85690.00	10.24	10.24	0.27
Traffic Department Deira	-2378860.00	2456183.11	2378860.00	-67.22	67.22	3.15

Conclusion

5.1 Summary

The summary of the two models. Showing some model accuracy measures. All measures show a lower value in MLR model comparing with ARIMA model. This proves that the performance of Linear Regression model is better.

Table 18: Summary table of MLR and ARIMA model.

Measures	Electricity		Water	
	MLR	ARIMA	MLR	ARIMA
RMSE	130385.3	202594.9	547776.9	899285.9
MAE	67079.42	167131.5	210319.6	773177.5
MPE	2.46	-2.4025	1.753	-17.042
MAPE	9.98	20.0232	8.44	36.9164
MASE	0.03625	0.7006	0.02251	0.9864

5.2 Conclusion

Annually, the electricity and water bills cost Dubai Police over 100 million dirhams which is a critical issue that needs to be monitored and resolved. In fact, savings achieved in the utility's bills will add extra money to other budgets allocated for other various operations. Analyzing consumption data using Machine learning algorithms will help in reducing the consumption in the future. I have used two different regression models to predict electricity and water consumption and they are Multiple Linear Regression and ARIMA model. As appeared in the actual and predicted consumption tables in both models, MLR model resulted in a better value which are very close the actual values. Moving to the accuracy measurements, using electricity data, the MAPE result is 11.88 for MLR and 20.79 in ARIMA model. However, using water data, the value of MAPE is 9.53 for MLR and 36.92 for ARIMA model. To conclude, Multiple Linear Regression model results in a lower MAPE which means higher prediction performance. The poorer performance of the ARIMA model is due to the fact that external factors account for a considerable amount of the fluctuation in monthly electric energy use, which univariate forecasting approaches cannot capture.

5.3 Recommendations

This study can be used by Dubai Police, specifically the energy conservation department, to place remedial measures to reduce the consumption of the highest consuming facilities that have been identified. In addition, factoring in the temperature parameter in the model, Dubai Police should focus on spreading more awareness with regards to the consumer behavior during the summer season as the consumption peaks. Finally, this study findings can facilitate a baseline for Dubai Police to adopt clean solar energy into its facilities, starting with the highest consuming facilities.

5.4 Future Works

Although for this research work the dataset was somehow limited due to confidentiality reasons as stated earlier, being an employee at the energy conservation department in Dubai Police, I will be able to access the full dataset and make a detailed study containing all operational, capita, and area data. As well as, to add a multivariant ARIMA model to the study and compare it with the existing models.