Exploratory Data Analysis using R for the Tetuan Dataset

An assignment presented as part of the MSc:
Data Science and Machine Learning

Course: Programming Tools and Technologies for Data Science

School of Electrical and Computer Engineering

Author: Apostolos Moustakis

Student number:

Email:

Instructors:

Table of Contents

List of Figures

1 Introduction

The purpose of this assignment is to perform exploratory data analysis with the use of R in the dataset Tetuan City power consumption.csv. This dataset contains 52.416 rows of data that describe the power consumption of Tetuan, a city located in the north of Morocco. Tetuan is divided into three different areas-zones that are powered from three different source stations respectively (Quads, Smir and Boussafou). Each row of the dataset contains information about the exact date and time of the measurement (time window of ten minutes), the temperature, the humidity, the speed of the wind and the power consumption for each of the three zones. The aim is to explore how these features (time and weather conditions) affect the power consumption in each zone.

2 Data Preprocessing

The first part of the process is importing the dataset using the R library data.table and checking if the types of the variables are correct [1]. The variable DateTime is of type character, which is not considered correct as this variable contains information about the date (m/d/y) and the time (h/m) that the data were observed and stored. Therefore, I decided to create new variables by extracting information that is contained in this variable. The created variables, directly from the variable DateTime, are the day, the month, the year, the hour and the minutes. An important note is that the variable year could be discarded since the dataset contains information only for the year 2017. Furthermore, from the variable DateTime, I created the variables quarter, season, weekday and yearday in order to use them in my analysis. The variable quarter represents the four quarters in a year (Q1: January, February and March, Q2: April, May and June, Q3: July, August and September and Q4: October, November and December), the variable season represents the four seasons (Winter, Spring, Summer, Fall), the variable weekday represents the day of the week (Sunday to Saturday, assuming that the first day of a week is Sunday) and the variable yearday represents the day of the year (1 to 364, as there are no data for the day 31/12). After the creation of the variables I dropped the column DateTime as it is no longer needed. Lastly, as part of the data preprocessing I decided to rename the variables Zone 1 Power Consumption, Zone 2 Power Consumption and Zone 3 Power Consumption, in order to be shorter. The R code produced for the preprocess of data is presented below.

```r
library(data.table)
data<-fread("c:/Tetuan.csv") #read the data table
dim(data) #dim of the table: 52.416 rows and 7 features
head(data) # first 6 rows
str(data) #find the type of every variable
library(lubridate)

#create new features
data$day <- as.numeric(format(as.Date(data$DateTime,format="%m/%d/%Y %H:%M"), format = "%d"))
data$month <- as.numeric(format(as.Date(data$DateTime,format="%m/%d/%Y %H:%M"), format = "%m")
    )
data$year <- as.numeric(format(as.Date(data$DateTime,format="%m/%d/%Y %H:%M"), format = "%Y"))
data$hour <- as.numeric(format(as.POSIXct(data$DateTime,format="%m/%d/%Y %H:%M"), format = "%H
    "))
data$minutes <- as.numeric(format(as.POSIXct(data$DateTime,format="%m/%d/%Y %H:%M"), format =
    "%M"))
data$quarter <- lubridate::quarter(data$month)
data$weekday <- wday(format(as.POSIXct(data$DateTime,format="%m/%d/%Y %H:%M"))) #week starts
    on Sunday
data$yearday <- yday(format(as.POSIXct(data$DateTime,format="%m/%d/%Y %H:%M")))
data$season <- 1 #winter: December, January, February
data$season[data$month>2&data$month<6] <- 2 #spring: March, April, May
data$season[data$month>5&data$month<9] <- 3 #summer: June, July, August
data$season[data$month>8&data$month<12] <- 4 #autumn: September, October, November

data <- data[,-1] #drop the column DateTime as it is not needed anymore

library("dplyr")
data <- data %>% #rename columns in order to be shorter
       rename("Z1 PC" = "Zone 1 Power Consumption",
              "Z2 PC" = "Zone 2  Power Consumption",
              "Z3 PC" = "Zone 3  Power Consumption")
```

R Code Snippet 1: Data Preprocessing

3 Exploratory Data Analysis (EDA)

3.1 Correlation Matrices

The exploratory data analysis starts with the creation of the correlation matrix that displays the correlation coefficients of the variables that are tested [2] [3]. In other words, the correlation matrix depicts the single correlation between a variable and all the other variables that exist in the dataset. The correlation takes values ranging from minus one to plus one, with values closer to the edges denoting high correlation (negative or positive respectively) and values closer to zero denoting no correlation [3]. Below I present two correlation matrices for the Tetuan dataset that are created with the library ggplot2.

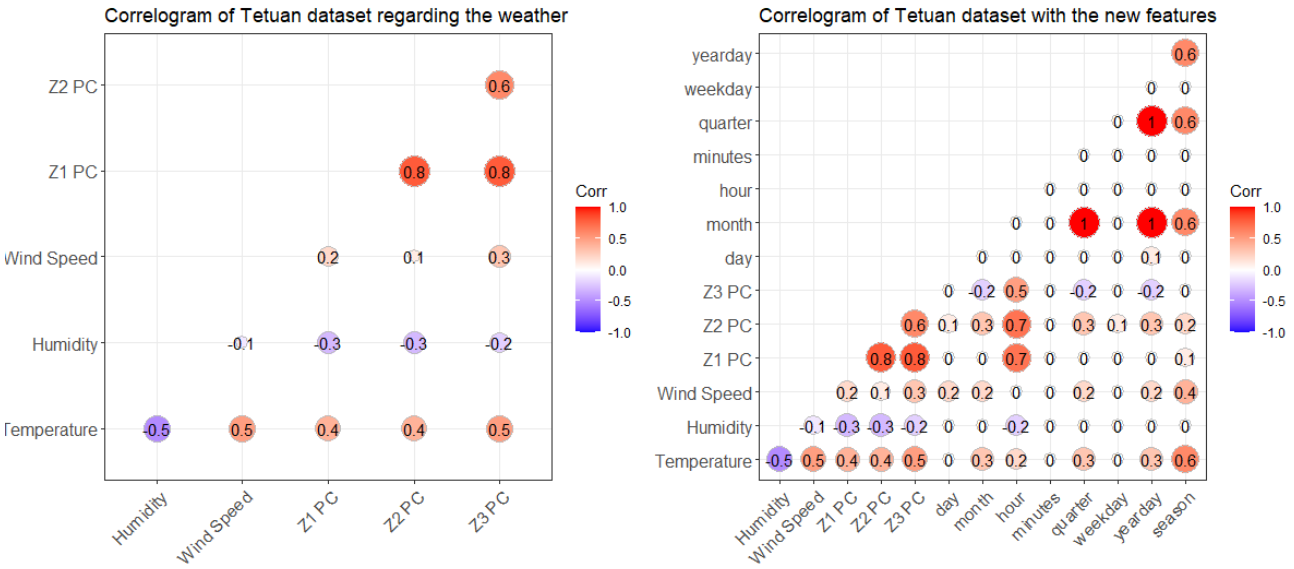

Figure 1: Correlograms of the features of Tetuan dataset

The first correlation diagram contains only the variables that refer to weather conditions. Initially, the variable DateTime cannot be used as it is of type character and needs modification. As I mentioned the interest relies on how the variables-features affect the power consumption in each zone. In this diagram one can deduce that the temperature is positively correlated with each zone to some extent (T-Z1: 0.4, T-Z2: 0.4, T-Z3: 0.5). The correlation is moderate and is almost the same for every zone. An example of what is happening can be that when the temperature rises, the power consumption in each zone also rises because citizens use the air-conditioning more, which is something that will be examined later. Furthermore, the wind speed is slightly positively correlated with each zone (WS-Z1: 0.2, WS-Z2: 0.1, WS-Z3: 0.3) and the humidity is slightly negative correlated with each zone (WS-H: -0.3, WS-H: -0.3, WS-H: -0.2), but somewhat more than the wind speed. Lastly, the zones are highly correlated with each other (Z1-Z2: 0.8, Z1-Z3: 0.8, Z2-Z3: 0.6). The correlation is not perfect, which means that there are some differences in the patterns of power consumption among the zones, but it is very strong, which means that generally the zones follow the same patterns (i.e. when the power consumption increases for zone 1 it also increases for zones 2 and 3).

The second diagram contains the additional variables that were created from the variable DateTime and refer to the date and time of the observation. In this diagram it is important to notice the strong positive correlation between the hour and each zone (h-Z1:0.7, h-Z2:0.7, h-Z3: 0.5), which is something that will be examined furthermore later. All the other added variables do not seem to present significant correlation with the power consumption in each zone, with some of them presenting only a slight correlation to some of the zones (for instance month and quarter). The R code for the creation of the correlation matrices is presented below.

```
#Corellation matrix with ggplot2
library(ggplot2)
library(ggcorrplot)
library("gridExtra") #place diagrams together
corr1<-round(cor(data[,-7:-15]), 1) #no time data
g1 <- ggcorrplot(corr1, method = "circle",
             lab=TRUE,
      lab_size = 4,
      title="Correlogram of Tetuan dataset regarding the weather",
      type = "lower",
      ggtheme=theme_bw)

corr2<-round(cor(data[,-9]), 1) #include new features - ignore the year (always 2017)
```

```r
g2 <- ggcorrplot(corr2, method = "circle",
            lab=TRUE,
      lab_size = 4,
      title="Correlogram of Tetuan dataset with the new features",
      type = "lower",
      ggtheme=theme_bw)

grid.arrange(g1, g2, ncol = 2, nrow=1)
```

R Code Snippet 2: Creation of Correlation Matrices for the Tetuan Dataset

3.2 Weather Analysis

After the creation of the Correlogram regarding the weather conditions (temperature, humidity and wind speed) some initial findings were deduced. In this section I will try to explore further how these features affect the power consumption in each zone, with a starting point being the creation of the histograms of the weather conditions in order to present how they are distributed [4]. The histograms of occurring temperature, humidity and wind speed are presented below along with the respective R code.

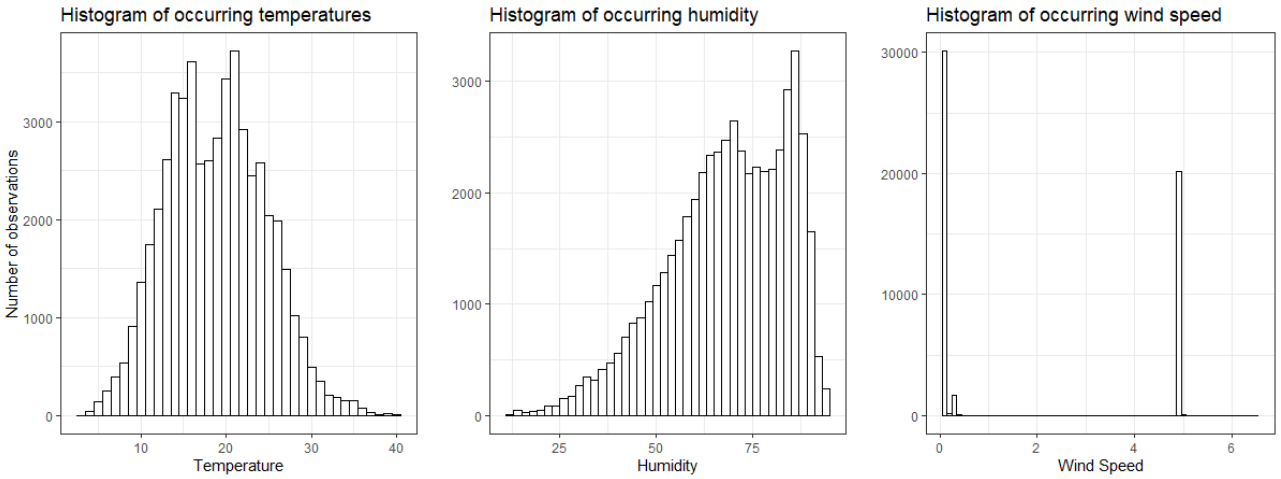

Figure 2: Histograms of Temperature, Humidity and Wind Speed

```r
theme_set(theme_bw()) #overall theme of all the diagrams

histogram <- function(name,data_x,binwidth) {  #Function for producing histograms

return(ggplot(data = data, aes(x = data_x)) +
       geom_histogram(color="black", fill="white",binwidth=binwidth)+
       labs(title = paste("Histogram of", name),
            y = "Number of observations",
        x = name))
}

grid.arrange(histogram("Temperature",data$Temperature,1),histogram("Humidity",data$Humidity,2)
    ,
histogram("Wind Speed",data$`Wind Speed`,0.1), ncol = 3, nrow=1)
```

R Code Snippet 3: Creation of Histograms for Temperature, Humidity and Wind Speed

To begin with I will examine more the relationship between the temperature and the power consumption in each zone. One approach that I will try first is creating a new variable that categorizes the temperature from very low to very high according to the temperatures that were observed in the respective histogram. Therefore, temperatures are categorized as very low (temperature<10), low (10<temperature<16), medium (16<temperature<22), high (22<temperature<28) and very high (temperature>28). Since the temperature is now a categorical variable and the power consumption of each zone is a quantitative variable, I will create boxplot diagrams for each zone [5]. The created boxplot diagrams along with the respective R code are presented below.

```r
#Function for producing boxplots

boxplot <- function(zone, d, data_x, data_y, name) {

```

```r
return(ggplot(d, aes(x = data_x, y = data_y, fill=data_x)) +
       geom_boxplot(varwidth=TRUE) +
       labs(title = zone,
       x = name,
       y = "Power consumption(KW)")+
       theme(legend.position = "none"))
}
#Creation of the variable temp_new

data$temp_new<-cut(data$Temperature, breaks=c(-Inf,10,16,22,28, Inf), labels = c("very low","
    low","medium","high","very high"))

d1 <- data[,.(temp = temp_new, power = `Z1 PC`)]
p1 <- boxplot("Zone 1", d1, d1$temp, d1$power, "Temperature")

d2 <- data[,.(temp = temp_new, power = `Z2 PC`)]
p2 <- boxplot("Zone 2", d2, d2$temp, d2$power, "Temperature")

d3 <- data[,.(temp = temp_new, power = `Z3 PC`)]
p3 <- boxplot("Zone 3", d3, d3$temp, d3$power, "Temperature")

grid.arrange(p1, p2, p3, ncol = 3, nrow=1)
```

R Code Snippet 4: Creation of boxplot diagrams for Temperature vs Power Consumption of each zone

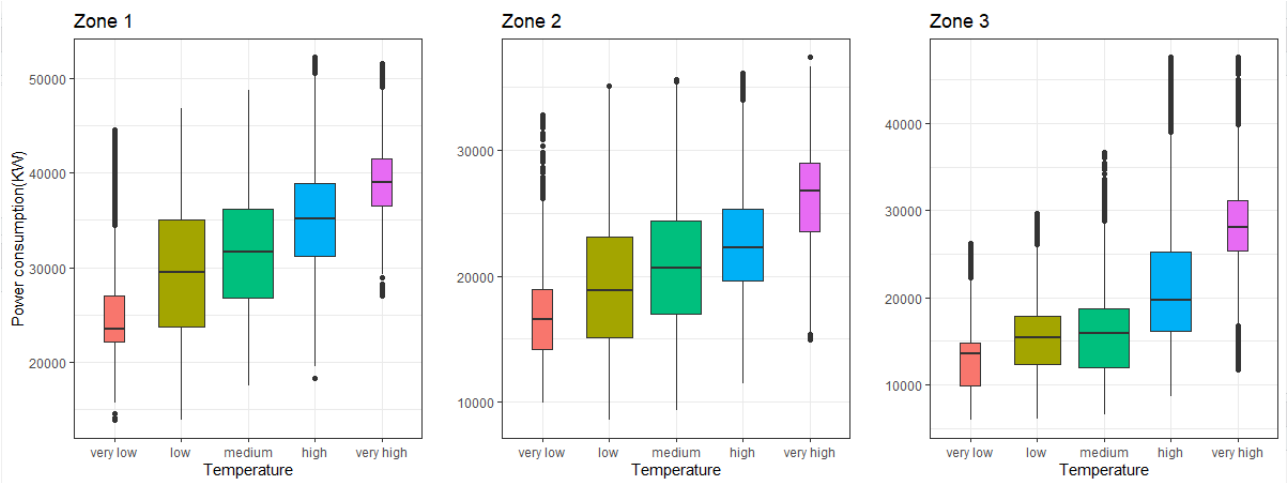

Figure 3: Boxplot Diagrams: Temperature Categories vs Power Consumption of each Zone

For the creation of the boxplot diagrams the argument=TRUE was used, which adjusts the width of the rectangles to be proportional to the number of observations of the specific level and the argument fill=temp_new, which colors the inside of the rectangles with the levels of the created variable temp_new [5]. By observing the diagrams, it is obvious that generally the higher the temperature the higher the power consumption in all the zones. For all the zones, the difference in the power consumption is immense between the very high and very low temperatures, while the power consumption seems to increase from every temperature level to the next, which validates the fact that temperature was shown to be positively correlated with each zone at the correlogram. However, the correlation was moderate for every zone, which may relies on the facts that generally the lengths of the rectangles are large and many outliers are detected, especially in zone 3.

Another approach to explore how temperature affects the power consumption in each zone is the creation of scatter plots, since both the variables are quantitative. However, the number of data is immense as the measurements occur every ten minutes. For this reason, I will calculate the average temperature and the average power consumption in each zone for every day of the year (variable yearday). Therefore, the scatter plots will show the average temperature in the x axis and the average power consumption of each zone in the y axis, while the points will represent the days of the year, which are 364 since there is no information for the last day of the year in the dataset. These scatter plots are presented below along with the respective R code.

```r
#Average Temperature vs Average Power Consumption: Scatter plots
#Fuction for producing the scatterplots

scatterplot <- function(zone, d, data_x, data_y, name) {

return(ggplot(d, aes(x=data_x, y=data_y)) +
          geom_point(color="blue") +
```

```
          labs(title=zone,
          x = paste("Average", name, "per day"),
          y = "Average Power Consumption per day(KW) ")+
          geom_jitter(width=0.5, color="blue"))
}
d1 <- data[,.(avgtemp=mean(`Temperature`), avgZ1=mean(`Z1 PC`)), by=.(yearday)]
p1 <- scatterplot("Zone 1", d1, d1$avgtemp, d1$avgZ1, "Temperature")

d2 <- data[,.(avgtemp=mean(`Temperature`), avgZ2=mean(`Z2 PC`)), by=.(yearday)]
p2 <- scatterplot("Zone 2", d2, d2$avgtemp, d2$avgZ2, "Temperature")

d3 <- data[,.(avgtemp=mean(`Temperature`), avgZ3=mean(`Z3 PC`)), by=.(yearday)]
p3 <- scatterplot("Zone 3", d3, d3$avgtemp, d3$avgZ3, "Temperature")

grid.arrange(p1, p2, p3, ncol = 2, nrow=2)
```

R Code Snippet5: Scatterplot diagrams for Average Temperature VS Average Power Consumption of each zone

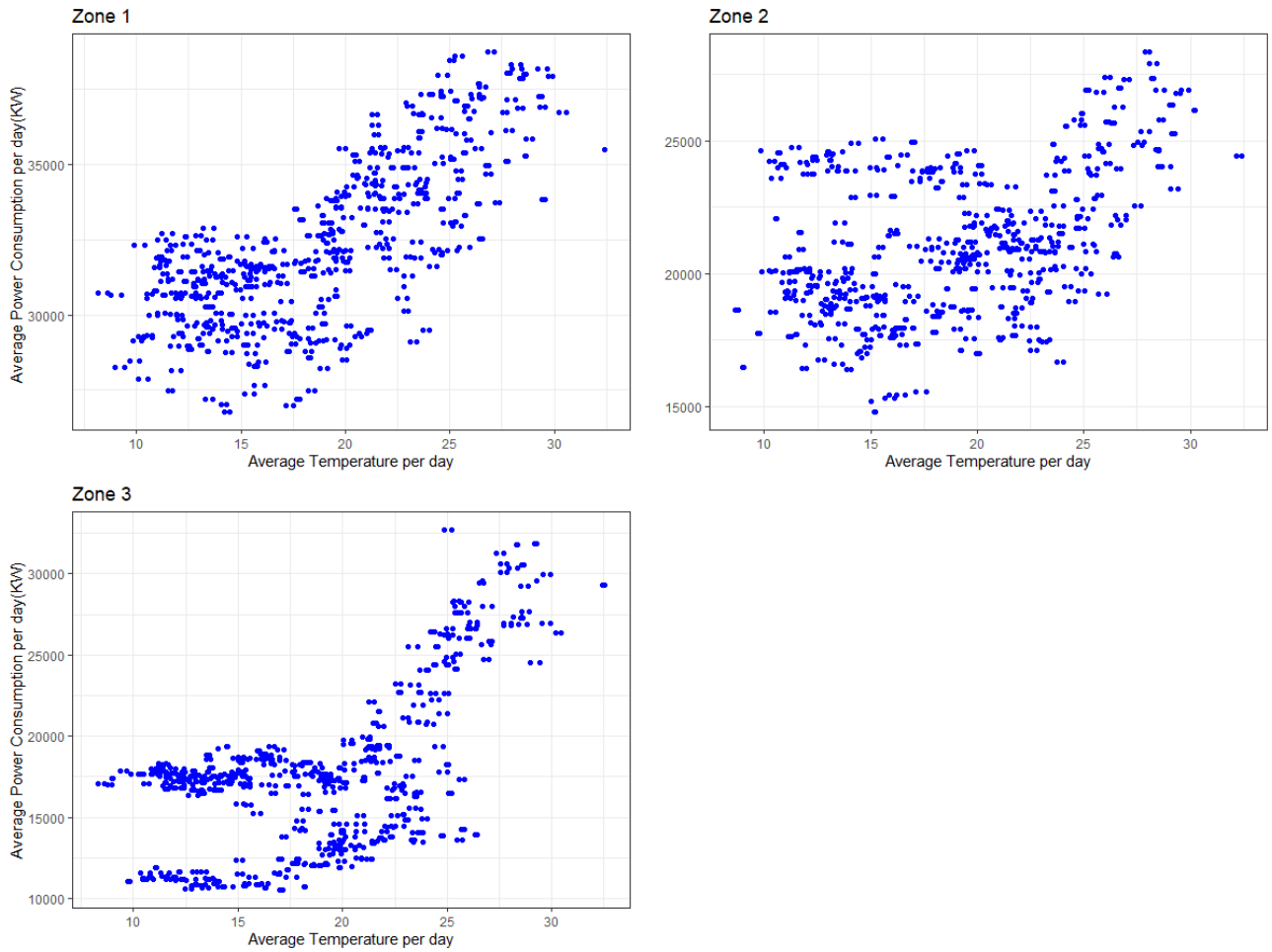

Figure 4: Scatterplot Diagrams: Average Temperature per day vs Average Power Consumption of each Zone per day

Again from all the diagrams one can notice that generally the higher the average temperature the higher the average power consumption of that day for all the zones. Of course the same conclusion as before is made. However, by using the average temperature of each day some information is lost regarding very low and very high temperatures that occur during a day. This is also the case for the average power consumption. Furthermore, from these specific diagrams one can conclude that the moderate positive correlation probably relies on the fact that when the average temperature is approximately in the range (10, 20) no increase in the power consumption is detected (the points seem to be distributed randomly around 30.000 in zone 1, randomly between 15.000 and 25.000 in zone 2 and randomly either close to 17.500 or close to 11.000 for zone 3). The increase is detected for higher average temperatures.

The same analysis will now take place for the humidity. However, I will not create scatterplots with the use of the average humidity per day because as I already described valuable information is lost. I will create a new variable that categorizes humidity from low to very high according to the humidity measurements that

were observed in the respective histogram. Therefore, humidity is categorized as low (humidity<40), medium (40<humidity<60), high (60<humidity<80) and very high (humidity>80). This time I decided to use four categories instead of five due to the humidity's distribution. The created boxplot diagrams along with the respective R code are presented below.

```r
#Creation of the variable humidity_new

data$humidity_new<-cut(data$Humidity, breaks=c(-Inf,40,60,80, Inf), labels = c("low","medium",
    "high","very high"))

#Creation of the boxplots using the function boxplot I created

d1 <- data[,.(hum = humidity_new, power = `Z1 PC`)]
p1 <- boxplot("Zone 1", d1, d1$hum, d1$power, "Humidity")

d2 <- data[,.(hum = humidity_new, power = `Z2 PC`)]
p2 <- boxplot("Zone 2", d2, d2$hum, d2$power, "Humidity")

d3 <- data[,.(hum = humidity_new, power = `Z3 PC`)]
p3 <- boxplot("Zone 3", d3, d3$hum, d3$power, "Humidity")

grid.arrange(p1, p2, p3, ncol = 3, nrow=1)
```

R Code Snippet 6: Creation of boxplot diagrams for Humidity vs Power Consumption of each zone

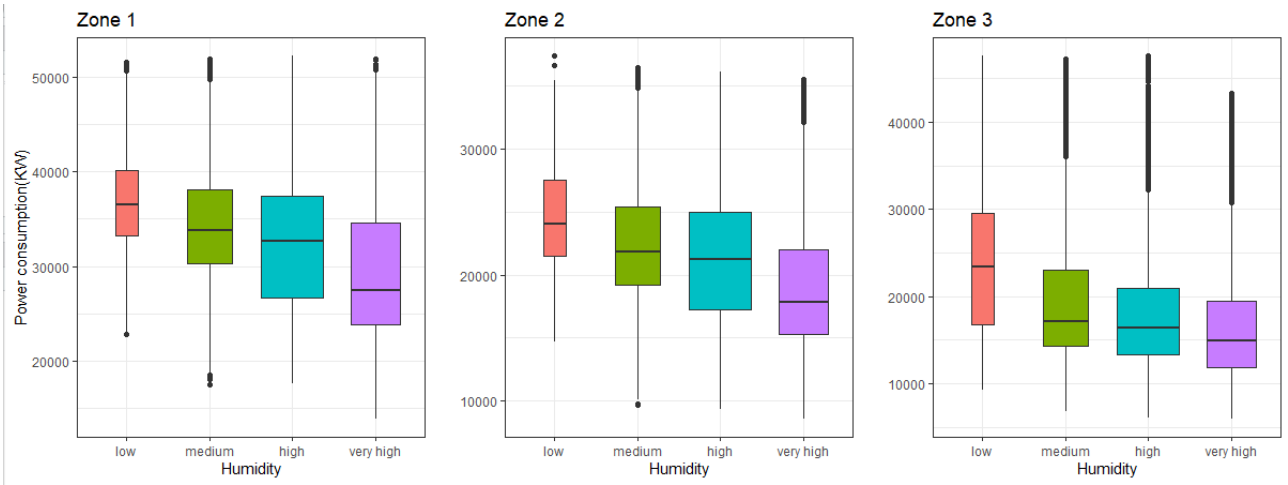

Figure 5: Boxplot Diagrams: Humidity Categories vs Power Consumption of each Zone

By observing the diagrams, one can conclude that generally the higher the humidity the lower the power consumption for all the zones. This was also the case in the correlogram where the humidity was found to be negative correlated with the power consumption in every zone. However, the correlation was slight which may relies on the facts that generally the lengths of the rectangles are large, the differences among the levels are small and many outliers are detected, especially in zone 3 that presents the slightest negative correlation with the power consumption, which is -0.2.

Last but not least, I will perform the same analysis for the wind speed. By observing the respective histogram, the wind speed is either close to zero or close to five and thus the variable that will be created will only have the levels low and high respectively. Scatterplots, in this case, not only hide information but alter the results as during a day very low or very high wind speed can occur and thus the average does not provide an accurate representation of the data. The R code for the creation of the boxplots and the boxplot diagrams are presented below.

```r
#Creation of the variable wind_new

data$wind_new<-cut(data$`Wind Speed`, breaks=c(-Inf, 2, Inf), labels = c("low", "high"))

#Creation of the boxplots using the function boxplot I created

d1 <- data[,.(wind = wind_new, power = `Z1 PC`)]
p1 <- boxplot("Zone 1", d1, d1$wind, d1$power, "Wind Speed")

d2 <- data[,.(wind = wind_new, power = `Z2 PC`)]
p2 <- boxplot("Zone 2", d2, d2$wind, d2$power, "Wind Speed")

```

```r
d3 <- data[,.(wind = wind_new, power = `Z3 PC`)]
p3 <- boxplot("Zone 3", d3, d3$wind, d3$power, "Wind Speed")

grid.arrange(p1, p2, p3, ncol = 3, nrow=1)
```

R Code Snippet 7: Creation of boxplot diagrams for Wind Speed vs Power Consumption of each zone

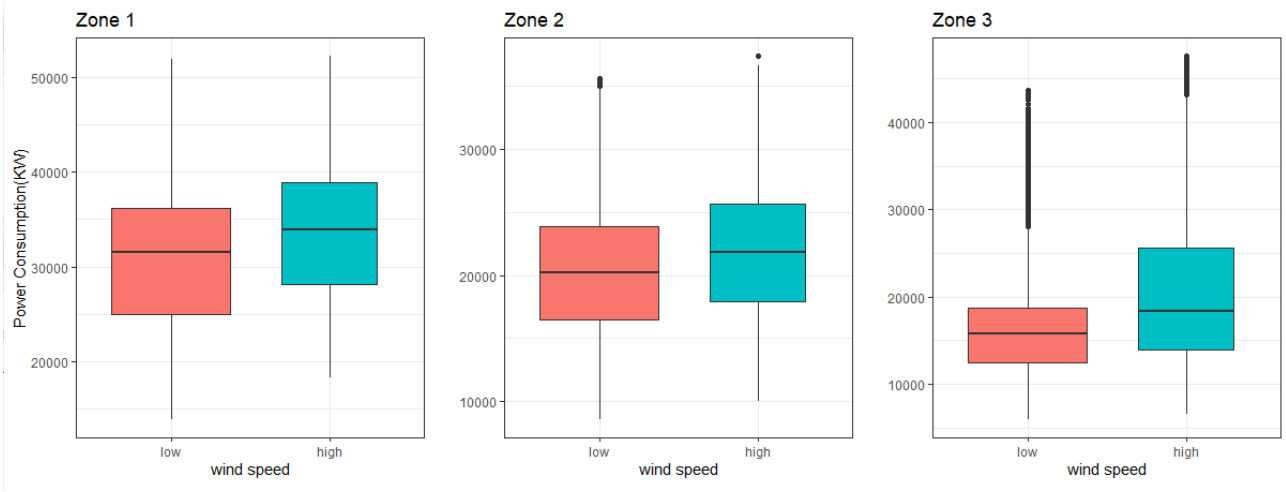

Figure 6: Boxplot Diagrams: Wind speed Categories vs Power Consumption of each Zone

In all the three zones when the wind speed is high the power consumption is generally higher. However, this correlation is slight as differences between the wind speed levels are small. The difference is greater in zone 3 where the length of the rectangle that refers to high wind speed is larger than the one that refers to the low wind speed. This explains why the correlation for zone 3 was found to be 0.3 during the creation of the correlogram, despite the fact that outliers are detected at the rectangle that refers to low wind speed.

3.3 Time Analysis

The time analysis starts with the variable hour. As I mentioned during the creation of the second correlogram, the hour has a strong correlation with the power consumption of each zone. In order to investigate more about this correlation I created boxplot diagrams that depict the power consumption of each zone in the y-axis and the 24 hours of a day in the x-axis (hour 0 belongs to midnight). The R code for the creation of the boxplots and the boxplots are presented below.

```r
#Boxplots for hour vs power consumption

p1 <- ggplot(data, aes(x = as.factor(hour), y =`Z1 PC`, fill=hour)) +
                    geom_boxplot() +
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
           labs(title = "Power Consumption of Zone 1 for every hour", #hour 0 belongs to 24:00
                    x = "hour",
                    y = "Power Consumption(KW)")+
           theme(legend.position = "none")

p2 <- ggplot(data, aes(x = as.factor(hour), y =`Z2 PC`, fill=hour)) +
                    geom_boxplot() +
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
           labs(title = "Power Consumption of Zone 2 for every hour", #hour 0 belongs to 24:00
                    x = "hour",
                    y = "")+
           theme(legend.position = "none")

p3 <- ggplot(data, aes(x = as.factor(hour), y =`Z3 PC`, fill=hour)) +
                    geom_boxplot() +
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
           labs(title = "Power Consumption of Zone 3 for every hour", #hour 0 belongs to 24:00
                    x = "hour",
                    y = "Power Consumption(KW)")+
           theme(legend.position = "none")
grid.arrange(p1, p2, p3, ncol = 2, nrow=2)
```

R Code Snippet 8: Creation of boxplot diagrams for hour vs Power Consumption of each zone

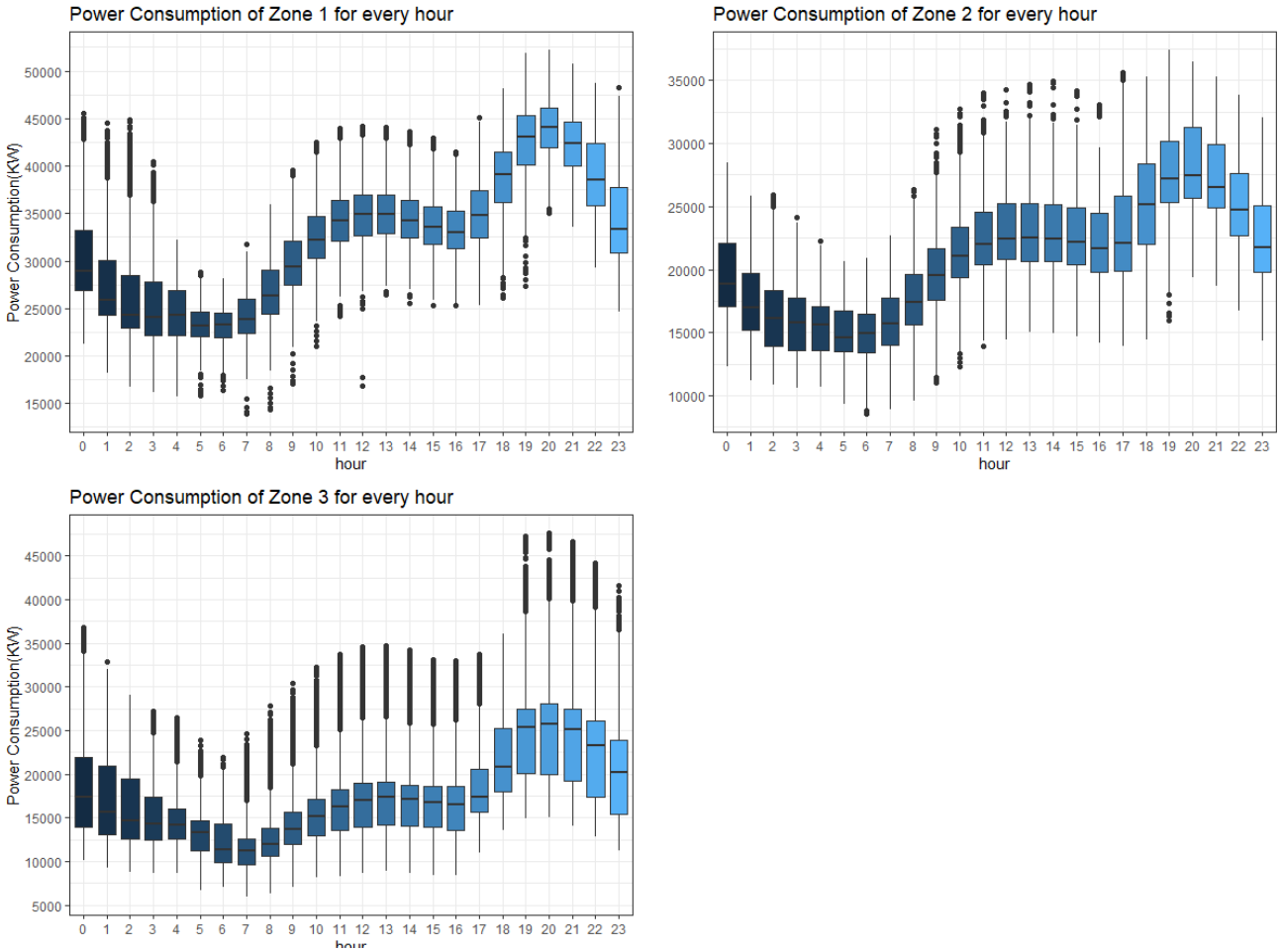

Figure 7: Boxplot Diagrams: Hour vs Power Consumption of each Zone

The correlation between the hour and the power consumption is obvious, as the power consumption tends to follow specific patterns during the 24 hours of a day. For all the zones the patterns that are detected are the same, despite the fact that the average power consumption of zone 1 is greater than that of the other two zones. More specifically, the hour with the greater power consumption is 20:00 (I am always referring to the boxplot and the middle quartile). From 20:00 the power consumption decreases until 5:00 for zones 1 and 2 and 7:00 for zone 3, then increases slightly until 12:00, stays mostly stable until 16:00 and increases considerably until 20:00. Generally the power consumption is very high during the evening hours (18:00-22:00), very low during the early morning hours (4:00-8:00) and moderate during the noon hours (12:00-16:00).

I will continue the analysis with the variable month, which according to the correlogram has no correlation with zone 1, a slight positive correlation with zone 2 and a slight negative correlation with zone 3. In order to explore more about how the month affects the power consumption in each zone I decided to create a time series diagram that depicts how the average power consumption evolves over time during the year 2017, calculated on a monthly basis. The time series diagram along with the R code for the creation of it are presented below.

```
theme_set(theme_bw()) #the theme
colors <- c("Zone 1" = "red", "Zone 2" = "green", "Zone 3" = "blue") #the colors of the lines
time_series<-data[, .(avgZ1=mean(`Z1 PC`), avgZ2=mean(`Z2 PC`), avgZ3=mean(`Z3 PC`)), by=month
    ]
ggplot(time_series, aes(x=month)) +
geom_line(aes(y=avgZ1, color="Zone 1"), size=.8) +
geom_line(aes(y=avgZ2, color="Zone 2"), size=.8)+
geom_line(aes(y=avgZ3, color="Zone 3"), size=.8)+
labs(title="Average Power Consumption of each zone by month",
x="Month",
y="Average Power Consumption",
color ="Legend")+
scale_color_manual(name="", values = colors)+
scale_x_discrete(name ="Month", limits=1:12) +
ylim(5000, 40000)
```

R Code Snippet 9: Creation of a time series diagram: months vs Power Consumption of each zone

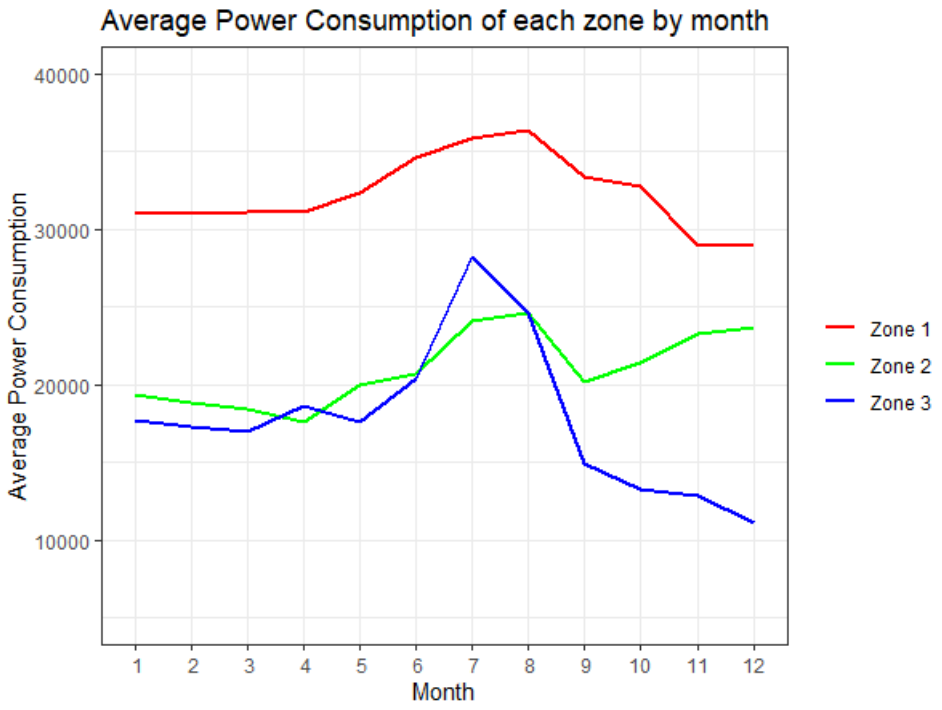

Figure 8: Time series diagram: Months vs Average Power Consumption of each Zone

Many conclusions can be drawn from this diagram. As it is already known, the average power consumption of zone 1 is always greater than zones 2 and 3 (Generally the total average power consumption for the year 2017 of every zone is respectively: 32344.97, 21042.51, 17835.41). For zone 1 the average power consumption remains almost stable during the first five months, increases a little during the summer months until August and then decreases a little until December, where it returns very close to its initial value. For zone 2 the average power consumption has some ups and downs, as it decreases very little until April, increases until August, decreases a bit in September and then increases again until December. However, the average consumption of zone 2 in December is higher than the initial value. Last but not least, for zone 3 the average power consumption has some small ups and downs until June, increases a lot in July and the decreases a lot until December, where it becomes way smaller than the initial value. Generally in this diagram there are two key points. The first one is that the average power consumption increases during the summer months for all the zones and the second one is that in December the average power consumption of zone 2 is increased and approaches zone 1 while the average power consumption of zone 3 is very decreased.

Since with the analysis of the variable month I found out that during the summer months the power consumption is higher, I will try to validate this outcome by introducing a new variable that I called season. With the variable season I divide the dataset into the fourth known seasons (winter, spring, summer and autumn) and try to explore how it affects the power consumption. For this purpose I present the following Lollipop diagrams of the average power consumption of each zone per season along with the respective R code.

```
d1 <- data[,.(avgZ1=mean(`Z1 PC`)), by=.(season)]
p1 <- ggplot(d1, aes(x = season, y = avgZ1, color =season)) +
                    geom_point(size=3)+
           geom_segment(aes(x=season, xend=season, y=0, yend=avgZ1))+
           scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = c(0, 40000)) +
           labs(title= "Zone 1",
           x = "Season",
                    y = "Average Power Consumption (KW)")+
           theme(legend.position = "none")

d2 <- data[,.(avgZ2=mean(`Z2 PC`)), by=.(season)]
p2 <- ggplot(d2, aes(x = season, y = avgZ2, color =season)) +
                    geom_point(size=3)+
           geom_segment(aes(x=season, xend=season, y=0, yend=avgZ2))+
           scale_y_continuous(breaks = seq(0, 40000, by = 5000),limits = c(0, 40000)) +
           labs(title= "Zone 2",
           x = "Season",
                    y = "")+
           theme(legend.position = "none")
```

```r

d3 <- data[,.(avgZ3=mean(`Z3 PC`)), by=.(season)]
p3 <- ggplot(d3, aes(x = season, y = avgZ3, color =season)) +
                    geom_point(size=3)+
           geom_segment(aes(x=season, xend=season, y=0, yend=avgZ3))+
           scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = c(0, 40000)) +
           labs(title= "Zone 3",
           x = "Season",
                    y = "")+
           theme(legend.position = "none")
grid.arrange(p1, p2, p3, ncol = 3, nrow=1)
```

R Code Snippet 10: Creation of lollipop charts for season vs Average Power Consumption of each zone

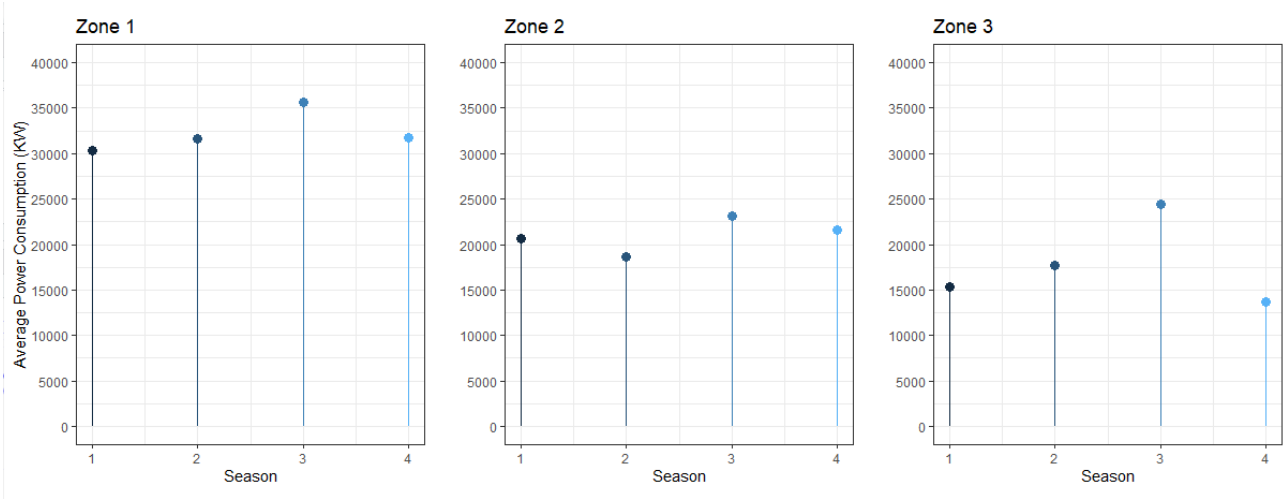

Figure 9: Lollipop charts: Seasons vs Average power consumption of each zone

Indeed for all the three zones there is higher average power consumption during the summer season (season 3). In addition, the average power consumption of zone 3 is very low during the autumn season (season 4), as expected due to the decline that was described previously. Moreover, it is important to note that the same analysis could be done for the variable quarter, mainly for statistical purposes, where the results are almost the same as the difference with the season variable is a month. Lastly, instead of lollipop charts that include the average power consumption of each zone, boxplot diagrams could be created with the power consumption of each zone.

As the next step of the time analysis I will further explore if the day of the week affects the power consumption by creating boxplots. The day of the week appeared to have 0 correlation with zones 1 and 3 and only 0.1 with zone 2. Therefore, I expect the created boxplots to show no difference in the power consumption regarding the day of the week. The diagrams and the respective R code are presented below.

```r
#boxplots of the weekday vs consumption
p1 <- ggplot(data, aes(x = as.factor(weekday), y =`Z1 PC`, fill=weekday)) +
                    geom_boxplot() +
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
           labs(title = "Zone 1",
                    x = "Day of the week", #1 is Sunday
                    y = "Power Consumption(KW)")+
           theme(legend.position = "none")

p2 <- ggplot(data, aes(x = as.factor(weekday), y =`Z2 PC`, fill=weekday)) +
                    geom_boxplot() +
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
           labs(title = "Zone 2",
                    x = "Day of the week",
                    y = "")+
           theme(legend.position = "none")

p3 <- ggplot(data, aes(x = as.factor(weekday), y =`Z3 PC`, fill=weekday)) +
                    geom_boxplot() +
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
           labs(title = "Zone 3",
                    x = "Day of the week",
                    y = "Power Consumption(KW)")+
           theme(legend.position = "none")
```

```r
grid.arrange(p1, p2, p3, ncol = 3, nrow=2)
```

R Code Snippet 11: Creation of boxplot Diagrams: Day of the week vs Power Consumption

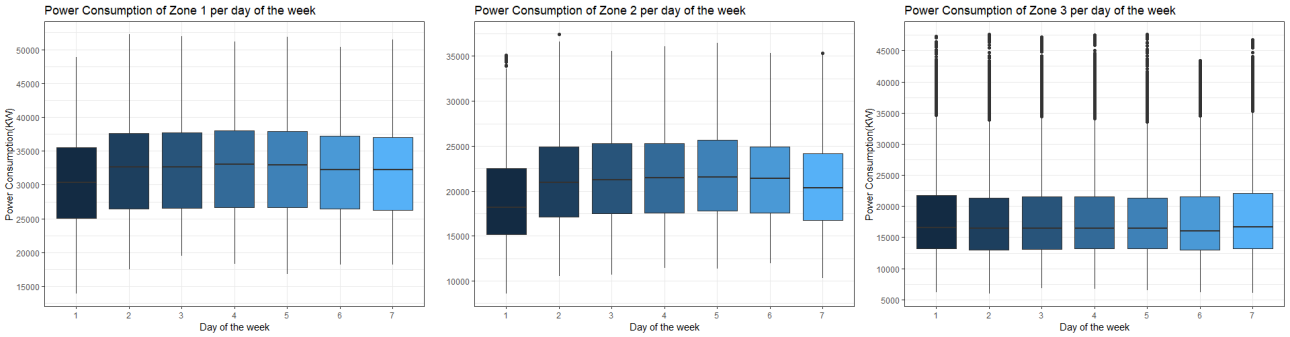

Figure 10: Boxplot Diagrams: Day of the week vs Power consumption of each zone

Indeed for zone 3 the power consumption is almost the same for every day of the week, with some outliers detected. This is also the case for zones 1 and 2, with an exception on Sundays (day 1), where the consumption seems sightly lower than the other days of the week. Moreover, in zones 1 and 2 Saturdays (day 7) also present a bit lower power consumption than the other days of the week except Sundays.

Consequently, in order to examine whether working days and weekends affect the power consumption in each zone I created a new variable, which I named work. This variable takes the values "working day" and "weekend" respectively, determined from the variable weekday (Sunday to Saturday: 1-7). Below I present three lollipop charts that show how the weekend and working days affect the average consumption in each zone on a yearly basis along with the respective R code.

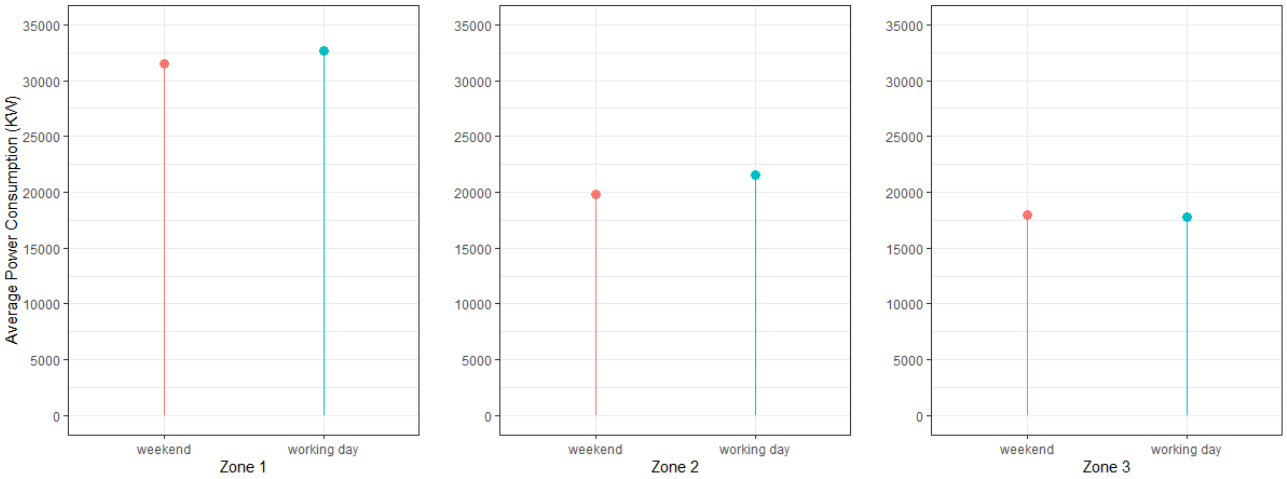

Figure 11: Lollipop charts: Working days and weekends vs Average power consumption of each zone

```r
#working days VS weekends: Lollipop diagrams
data$work <- "working day" #days from Monday to Friday (2 to 6)
data$work[data$weekday==1|data$weekday==7] <- "weekend" #Saturday&Sunday

diagram_work1 <- data[,.(avgZ1=mean(`Z1 PC`)), by=.(work)]
d1 <- ggplot(diagram_work1, aes(x = work, y = avgZ1, color =work)) +
                  geom_point(size=3)+
           geom_segment(aes(x=work, xend=work, y=0, yend=avgZ1))+
           scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = c(0, 35000)) +
           labs(x = "Zone 1",
                  y = "Average Power Consumption (KW)")+
           theme(legend.position = "none")

diagram_work2 <- data[,.(avgZ2=mean(`Z2 PC`)), by=.(work)]
d2 <- ggplot(diagram_work2, aes(x = work, y = avgZ2, color = work)) +
                  geom_point(size=3)+
           geom_segment(aes(x=work, xend=work, y=0, yend=avgZ2))+
           scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = c(0, 35000)) +
```

```
            labs(x = "Zone 2",y ="")+
            theme(legend.position = "none")

diagram_work3 <- data[,.(avgZ3=mean(`Z3 PC`)), by=.(work)]
d3 <- ggplot(diagram_work3, aes(x = work, y = avgZ3, color=work)) +
                    geom_point(size=3)+
            geom_segment(aes(x=work, xend=work, y=0, yend=avgZ3))+
            scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = tc(0, 35000)) +
            labs(x = "Zone 3", y = "")+
            theme(legend.position = "none")
grid.arrange(d1, d2, d3, ncol = 3, nrow=1)
```

R Code Snippet 12: Creation of lollipop charts for working days vs Average Power Consumption of each zone

Looking at the diagrams, one can assume that in Zones 1 and 2 the average power consumption is slightly greater on working days that it is on weekends. In Zone 1 the difference is approximately 1.100 KW and in Zone 2 the difference is approximately 1.700 KW. In Zone 3 the power consumption is almost the same between weekends and working days, as expected. It is important to note that the same observations also occur monthly. For Zones 1 and 2 the average power consumption is always greater on working days than it is on weekends in every month (expect June for Zone 1), while this is not the case for Zone 3 where the power consumption is sometimes greater on weekends and sometimes greater on working days, but mostly the same. These conclusions were drawn from the grouped bar plots that are presented below and are made on a monthly basis.

Figure 12: Grouped Barplot Diagrams:Weekend and Working Days vs Average power consumption in each Zone on a monthly basis

To sum up, it seems that on working days the average power consumption of zones 1 and 2 is a bit greater than it is on weekends, which may rely on the fact that companies are closed on weekends and therefore they do not consume electricity. However, it is difficult to determine why this difference occurs without further data, but this analysis is a good starting point. The R code for the creation of the grouped barplots is presented below.

```
#Wokring days VS weekends: Grouped barplots montlhy
d1 <- data[,.(avgZ1=mean(`Z1 PC`)), by=.(month, work)]
plot1 <- ggplot(d1, aes(x = as.factor(month), y = avgZ1, fill=work)) +
                    geom_bar(stat='identity', width=.5, position = "dodge") +
            labs(x = "Month",
                    y = "Average Power Consumption(KW)")+
            scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = c(0, 37000))+
            theme(legend.position = "bottom")

d2 <- data[,.(avgZ2=mean(`Z2 PC`)), by=.(month, work)]
plot2 <- ggplot(d2, aes(x = as.factor(month), y = avgZ2, fill=work)) +
                    geom_bar(stat='identity', width=.5, position = "dodge") +
            labs(x = "Month",
            y = "")+
            scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = c(0, 37000))+
            theme(legend.position = "bottom")

d3 <- data[,.(avgZ3=mean(`Z3 PC`)), by=.(month, work)]
plot3 <- ggplot(d3, aes(x = as.factor(month), y = avgZ3, fill=work)) +
                    geom_bar(stat='identity', width=.5, position = "dodge") +
            labs(x = "Month",
```

```
            y = "")+
            scale_y_continuous(breaks = seq(0, 40000, by = 5000), limits = c(0, 37000))+
                    theme(legend.position = "bottom")
grid.arrange(plot1, plot2, plot3, ncol = 3, nrow=1)
```

R Code Snippet 13: Creation of grouped barplots: Weekend and Working Days vs Average power consumption in each Zone on a monthly basis

As the last part of the analysis, I will try to explore how certain holiday periods affect the power consumption of each zone. Specifically, my analysis will focus on the Ramadan, as the majority of the people of Tetouan are Muslims [6]. In 2017 the Ramadan took place from 27th of May until 25th of June [7]. I will examine how Ramadan affects the power consumption of each zone by comparing the power consumption during the last five days of it with the power consumption during the first five days after it. Since weather can also influence the power consumption during a span of 10 days, I decided to include the temperature (temp_new) in this analysis too. The best diagram to depict the relationship among one quantitative and two categorical variables is a boxplot, which in this case will include the 10 days in the x-axis, the power consumption of each zone in the y-axis and the temperature "as a color" of each rectangle [5]. A vertical line will also be added between the 25th and 26th of June to denote the end of Ramadan. The produced boxplots along with the R code are presented below.

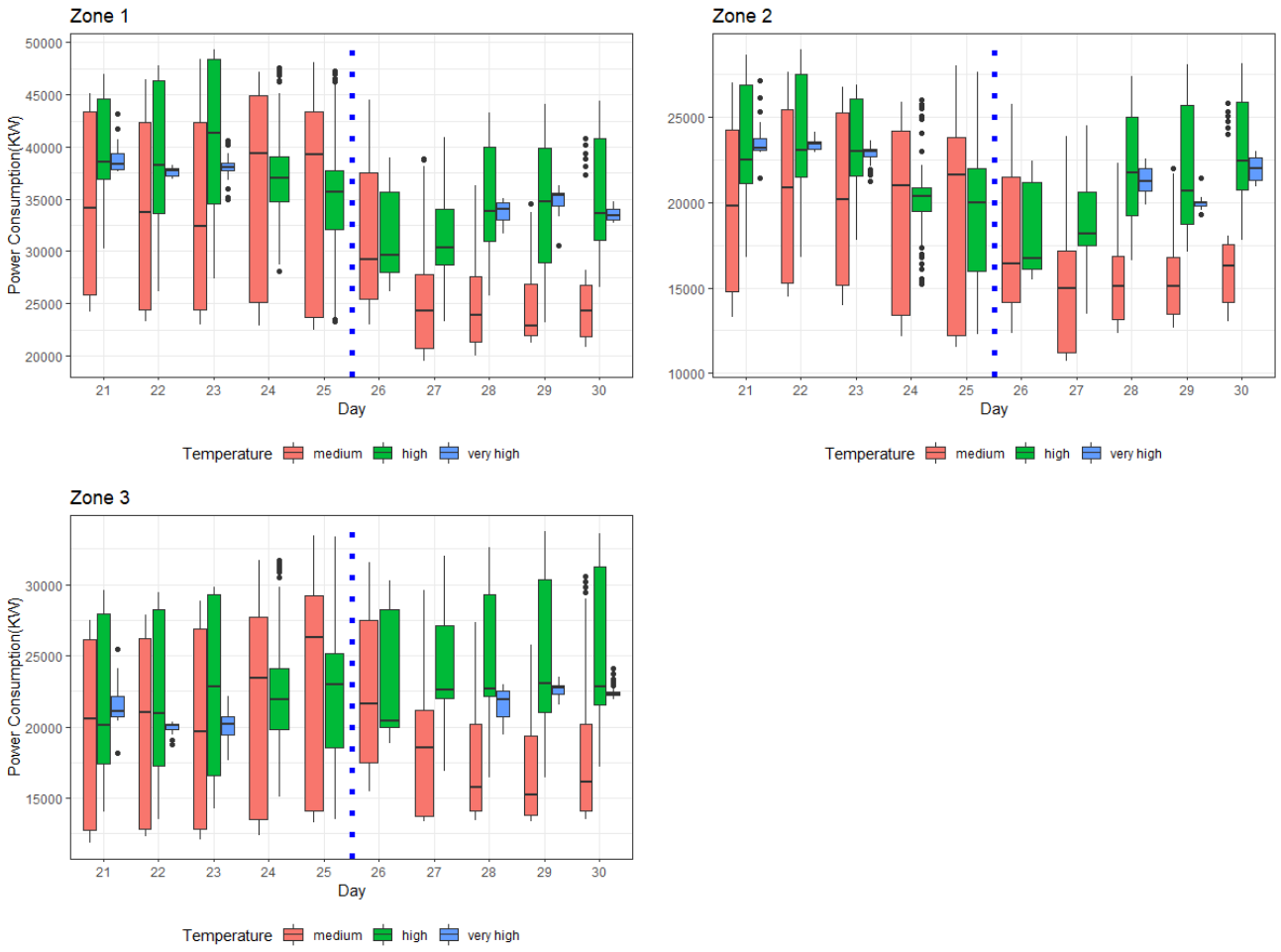

Figure 13: Boxplot Diagrams: Five last days of Ramadan and first five days after Ramadan vs Power Consumption of each Zone (colors assigned by temperature)

```
#Boxplots Power Consumption for Ramadan days

data_Ramadan <- data[data$yearday>=172&data$yearday<=181]  #period 21/6 to 30/6

p1 <- ggplot(data_Ramadan, aes(x = as.factor(day), y =`Z1 PC`, fill=temp_new)) +
                    geom_boxplot() +
            geom_vline(xintercept = 5.5, linetype="dotted",
                    color = "blue", size=2)+
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
            labs(title = "Zone 1",
```

```
                    x = "Day",
                    y = "Power Consumption(KW)")+
            labs(fill="Temperature")+
            theme(legend.position = "bottom")

p2 <- ggplot(data_Ramadan, aes(x = as.factor(day), y =`Z2 PC`, fill=temp_new)) +
                    geom_boxplot() +
            geom_vline(xintercept = 5.5, linetype="dotted",
                    color = "blue", size=2)+
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
            labs(title = "Zone 2",
                    x = "Day",
                    y = "")+
            labs(fill="Temperature")+
            theme(legend.position = "bottom")

p3 <- ggplot(data_Ramadan, aes(x = as.factor(day), y =`Z3 PC`, fill=temp_new)) +
                    geom_boxplot() +
            geom_vline(xintercept = 5.5, linetype="dotted",
                    color = "blue", size=2)+
                    scale_y_continuous(breaks = seq(0, 50000, by = 5000))+
            labs(title = "Zone 3",
                    x = "Day",
                    y = "Power Consumption(KW)")+
            labs(fill="Temperature")+
            theme(legend.position = "bottom")

grid.arrange(p1, p2, p3, ncol = 2, nrow=2)
```

R Code Snippet 14: Creation of the boxplots: Ramadan days vs Power Consumption, colored by temperature

Looking at the boxplots diagrams, the difference between the 25th of June (end of Ramadan) and the 26th of June (first day after Ramadan) is significant for all the zones. During the Ramadan the power consumption is generally higher, no matter the temperatures that occur. After the Ramadan the power consumption is generally lower and the temperature plays a more vital role. As it was mentioned in the weather analysis the temperature has a positive correlation with the power consumption, which means that when the temperature rises the power consumption generally increases. This is very clearly seen after Ramadan, but during Ramadan it seems that the temperature does not play such a significant role. Therefore, other factors determine the high power consumption, with the customs and traditions of Ramadan being the prime suspect. Several claims have been made from electricity companies in Muslim countries that during the Ramadan days (and especially the last ten days) the electricity consumption is very high compared to other days, which is distinctly depicted in the Tetouan dataset as well [8][9].

4 Summary and further research

According to the analysis presented in this report, time and weather conditions certainly affect the power consumption in each zone. The hour, regarding the time, and the temperature, regarding the weather conditions, were found to most predominantly affect the power consumption in each zone individually. Furthermore, time and weather conditions can be combined for a more thorough analysis, as it happened during the Ramadan analysis where the day and the temperature were used to examine how the power consumption in each zone was affected. In cases like this, having external information can be used in order to explore if and how the data are affected and if certain patterns appear.

With all the conclusions in the report being made a next step of the analysis could be a forecasting on how the power consumption will move the next year with the use of Machine Learning. Since all the information in the dataset refer to the year 2017 this forecasting should have been made at the end of that year for the year 2018. However, having predictions for the year 2018 and collecting the actual data can further improve the Machine Learning algorithm and produce more accurate results. Accurate predictions in the power consumption for forthcoming years can help Tetuan in many aspects, as for instance in the prevention of energy shortcomings in certain periods of time.

5 References

1. Fouskakis, D. *Introduction to Data Tables* http://www.math.ntua.gr/~fouskakis/Programming_R/Slides/3.pdf.
2. Fouskakis, D. *Common ggplot visualizations* http://www.math.ntua.gr/~fouskakis/Programming_R/Slides/5.pdf.
3. Bock, T. *What is a Correlation Matrix* https://www.displayr.com/what-is-a-correlation-matrix/.
4. Fouskakis, D. *Introduction to ggplot* http://www.math.ntua.gr/~fouskakis/Programming_R/Slides/4.pdf.
5. Fouskakis, D. *Data Analysis using R: 2nd Edition* ISBN: 978-618-5495-31-2 (Tsotras, 2021).
6. Britannica. *Tétouan* https://www.britannica.com/place/Tetouan.
7. CalendarDate.com. *Ramadan 2017* https://www.calendardate.com/ramadan_2017.htm.
8. Afifa, L. *Ramadan; Electricity Consumption Increases 5 percent, PLN Says* https://en.tempo.co/read/1204406/ramadan-electricity-consumption-increases-5-percent-pln-says.
9. Zalif, Z. *Price hikes due to increased use of electricity during Ramadan: STELCO* https://raajje.mv/57692.

6 Appendix

All the diagrams that were presented in this assignment were created with the library ggplot2 that does not support the creation of 3D diagrams. A 3D diagram is necessary when the relationship of three quantitative variables is tested [5]. For instance one can further explore the relationship among the temperature, the humidity and the power consumption (all of them are quantitative variables) since temperature is negative correlated with humidity and both temperature and humidity are correlated with the power consumption in each zone. Since the number of data is immense, as described in the report, the average temperature per day, the average humidity per day and the average consumption per day could be used for the creation of the diagram. The diagram will be created with the library plotly. Alternatively other libraries or the package gg3D could be used that extends the library ggplot2 in order to produce 3D plots. The diagram will contain the average consumption per day in the x-axis, the average humidity per day in the y-axis and the average temperature per day in the z-axis. The diagram for Zone 1 is presented below along with the respective R code.

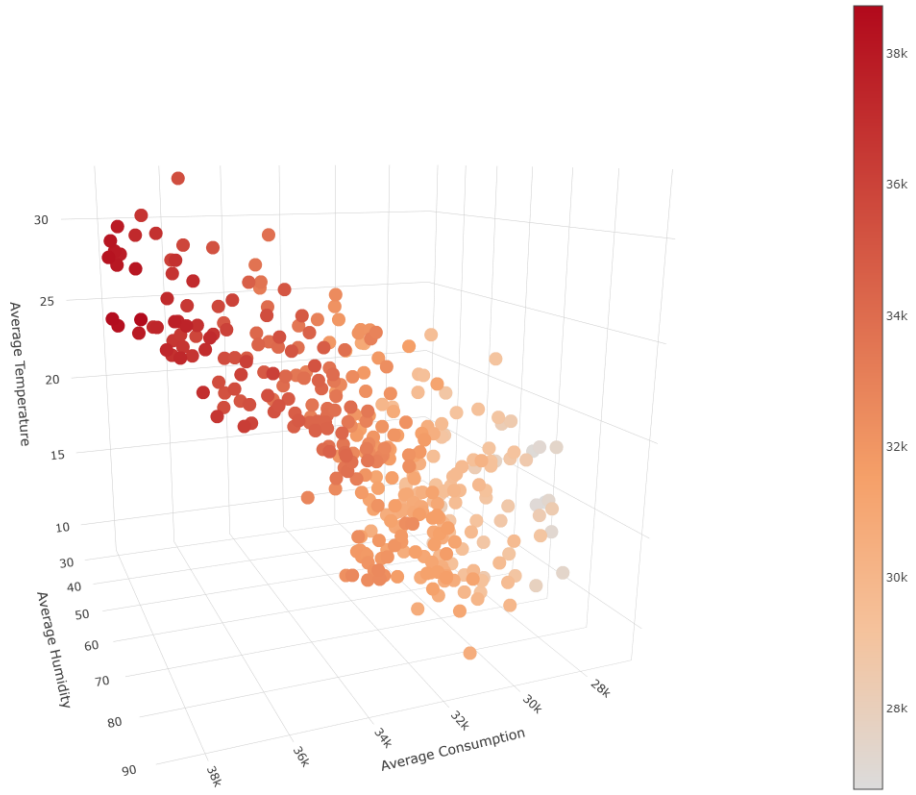

Figure 14: 3D Diagram: Average Consumption per day vs Average Humidity per day and Average Temperature per day for Zone 1

```r
#3D Diagram
library(plotly)
d <- data[,.(avgtemp=mean(`Temperature`), avgZ1=mean(`Z1 PC`), avghum=mean(`Humidity`)), by=.(
    yearday)]
fig <- plot_ly(d, x = ~avgZ1, y = ~avghum, z = ~avgtemp,
               marker = list(color = ~avgZ1, showscale = TRUE))
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(title = 'Average Consumption'),
                                   yaxis = list(title = 'Average Humidity'),
                                   zaxis = list(title = 'Average Temperature')),
                      annotations = list(
                        x = 1.13,
                        y = 1.05,
                        xref = 'paper',
                        yref = 'paper',
                        showarrow = FALSE
                        ))
fig
```

R Code Snippet 15: Creation of 3D Diagram: Average Consumption per day vs Average Humidity per day and Average Temperature per day

From the diagram one can deduce that the points, which reflect the 364 days, are distributed diagonally in the 3D space. A point in the upper left corner represents a day where the average power consumption is high, the average temperature is high and the average humidity is low, while a point in the lower right corner represents a day where the average power consumption is low, the average temperature is low and the average humidity is high. Consequently, a conclusion from this diagram is that generally the higher the average consumption, the higher the average temperature and the lower the average humidity, which was expected due to the correlation between each of the two variables. The respective diagrams for zones 2 and 3 present similar results.

One last important notice is that an alternative approach to explore the relationship among the temperature, the humidity and the power consumption, based on what is discussed during the assignment, would be to use the created variables temp_new and humidity_new. Now since two variables are categorical and one variable is quantitative, a boxplot diagram could be created with the temp_new in the x-axis, the power consumption in the y-axis and the humidity_new "as a color" of each rectangle. The created boxplot diagram for Zone 1 along with the respective R code are presented below. The results of this diagram are matching with the previous 3D plot (with more details) while similar results occur for zones 2 and 3.

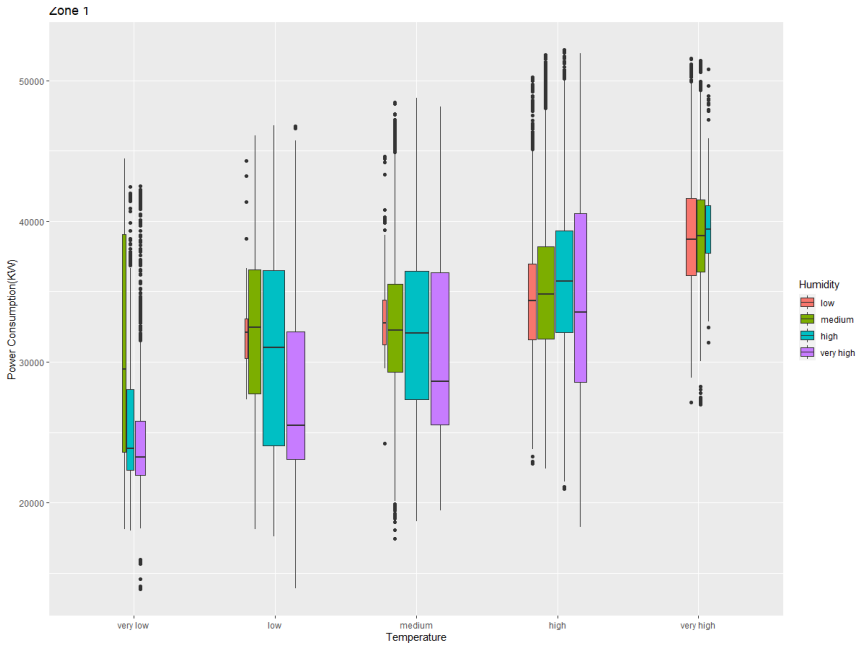

Figure 15: Boxplot Diagram: Temperature vs Power Consumption for Zone 1 (colors assigned by humidity)

```
ggplot(data, aes(x = temp_new, y = `Z1 PC`, fill=humidity_new)) +
     geom_boxplot(varwidth=TRUE) +
     labs(title = "Zone 1",
          x = "Temperature",
          y = "Power Consumption(KW)")+
     labs(fill="Humidity")
```

R Code Snippet 16: Creation of Boxplot Diagram: Temperature vs Power Consumption, colored by humidity for Zone 1