Final Project

Mehmet Cihan Sakman

14 05 2021

İzmir Büyükşehir Belediyesi

R Final Project

Abstract

The scope of this project is the enviromental analyze of Izmir from the three datasets generated by the Izmir Municipality (IBB Acik Veri Portali https://acikveri.bizizmir.com/dataset). In this project we will try to analyze the

Water Consumption and Production and The First Subscription Water Connection Realization Times. In the Water Consumption part, the difference between water consumption in Summer and Winter try to clarify. The most water consumer districts will be handled and try to figure out the main reasons for water consumption in these districts. In the Water Consumption part, the difference between water consumption in Summer and Winter try to clarify. The most water consumer districts have been handled and try to figure out the main reasons for water consumption in these districts. In the Water Production part, the difference between the years of water production by sources try to be handled and the question of from which year the water production started do increase answered by statistically. On the other hand, the water production differences between the sources try to be handled, and the question of 'Is there a specific difference in water production between the sources' tries to be answered. In the last part, the project tries to find if there is a specific difference in the average water-binding time by years and highlight the average subscription time if it's more than a month or not.

Imports

```
library(ggplot2)
library('stringr')
library(dplyr)
library(ggpubr)
library(moments)
```

Loading Dataset

Change the column names

```
colnames(water.consumption) <-
c("year","month","district","neighborhood","user_count","avg_consumption")
colnames(water.production) <- c("year","month","source","amount_m3")
colnames(water.binding) <-
c("year","district","binding_time_day","sub_time_day","explore_time_day","col
lection_time_day","petition_count")
```

EXPLANATION OF THE COLUMNS

1. **WATER CONSUMPTION**
   - *year:* Information about year
   - *month:* Information about month
   - *district:* Information about district
   - *neighborhood:* Information about neighborhood

- *user_count:* Information about subscriber count
- *avg_consumption:* The average water consumption in m^3

2. **WATER PRODUCTION**
   - *year:* Information about year
   - *month:* Information about month
   - *source:* Information about the source of water production
   - *amount_m3:* The water production in m^3
3. **WATER BINDING**
   - *year:* Information about year
   - *district:* Information about district
   - *binding_time_day:* First water subscription average binding time (days)
   - *sub_time_day:* Total average duration of first water subscription (days)
   - *explore_time_day:* First water subscription average discovery time (days)
   - *collection_time_day:* Average tax collection time for first water subscription (days)
   - *petition_count:* Number of subscriber petitions applied on the same day

---

Preprocessing and descriptive stats

Let's see some of data

- Summarize all three data

```
## [1] "Summary of Water Consumption"

##       year          month          district          neighborhood
##  Min.   :2020   Min.   : 1.000   Length:15112       Length:15112
##  1st Qu.:2020   1st Qu.: 2.000   Class :character   Class :character
##  Median :2020   Median : 6.000   Mode  :character   Mode  :character
##  Mean   :2020   Mean   : 5.921
##  3rd Qu.:2020   3rd Qu.: 9.000
##  Max.   :2021   Max.   :12.000
##    user_count    avg_consumption
##  Min.   :    1   Length:15112
##  1st Qu.:   75   Class :character
##  Median :  275   Mode  :character
##  Mean   : 1025
##  3rd Qu.: 1269
##  Max.   :12633

## [1] "Summary of Water Production"

##       year          month           source            amount_m3
##  Min.   :2009   Min.   : 1.000   Length:1668        Min.   :      0
##  1st Qu.:2012   1st Qu.: 3.000   Class :character   1st Qu.:  98109
##  Median :2015   Median : 6.000   Mode  :character   Median : 467908
##  Mean   :2015   Mean   : 6.413                      Mean   :1568540
```

```
##  3rd Qu.:2018   3rd Qu.: 9.000                       3rd Qu.:2571146
##  Max.   :2021   Max.   :12.000                       Max.   :9786100
```

```
## [1] "Summary of Water Binding"
```

```
##       year        district          binding_time_day   sub_time_day
##  Min.   :2018   Length:96          Length:96          Length:96
##  1st Qu.:2018   Class :character   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2019
##  3rd Qu.:2020
##  Max.   :2020
##  explore_time_day   collection_time_day petition_count
##  Length:96          Length:96           Min.   :   81.0
##  Class :character   Class :character    1st Qu.:  453.5
##  Mode  :character   Mode  :character    Median : 1469.0
##                                         Mean   : 2214.1
##                                         3rd Qu.: 2791.5
##                                         Max.   :15265.0
```

There are some mistakes in the data.

1. In **Water Consumption**, *'year'* and *'month'* columns' type should be String, *'avg_consumpiton'* column kept as String and numbers are seperated by **','**. We'll replace these ',' with '.' and convert the type into **float**.

2. In **Water Production**, *'year'* and *'month'* columns' type should be String, *'amount_m3* should kept as **float**.

3. In **Water Binding**, *'year'* and *'month'* columns' type should be String. All the numeric columns kept as String and numbers are seperated by **','**. We'll replace these ',' with '.' and convert the type into **float**

4. For all data we convert *'character'* type variable into *'factor'*

```
water.consumption$avg_consumption <-
str_replace_all(water.consumption$avg_consumption , ',', '.')

paste("Edit for Water Consumption")

## [1] "Edit for Water Consumption"

water.consumption$avg_consumption =
as.numeric(as.character(water.consumption$avg_consumption))
water.consumption$year = as.character(as.numeric(water.consumption$year))
water.consumption$month = as.character(as.numeric(water.consumption$month))

names <- c('year' ,'month', 'district','neighborhood')
water.consumption[,names] <- lapply(water.consumption[,names] , factor)
paste("Summary for Water Production after Edit")

## [1] "Summary for Water Production after Edit"
```

```
sapply(water.consumption, summary)

## $year
##  2020  2021
## 12464  2648
##
## $month
##    1   10   11   12    2    3    4    5    6    7    8    9
## 2358 1221 1151 1207 2379 1316  255  403 1274 1203 1139 1206
##
## $district
##      ALİAĞA     BALÇOVA    BAYINDIR    BAYRAKLI     BERGAMA      BEYDAĞ
##         386         114         686         341        1237         291
##     BORNOVA        BUCA       ÇEŞME       ÇİĞLİ      DİKİLİ        FOÇA
##         640         632         333         367         249         204
##    GAZİEMİR  GÜZELBAHÇE  KARABAĞLAR   KARABURUN   KARŞIYAKA   KEMALPAŞA
##         212         150         735         212         371         635
##       KINIK       KİRAZ       KONAK    MENDERES     MENEMEN   NARLIDERE
##         431         558        1387         537         757         148
##      ÖDEMİŞ SEFERİHİSAR      SELÇUK        TİRE     TORBALI        URLA
##        1104         277         193         721         732         472
##
## $neighborhood
##                ATATÜRK             CUMHURİYET                  FATİH
##                    260                    205                    118
##                   YENİ                  İNÖNÜ                  ZAFER
##                    104                     93                     92
##               HÜRRİYET               BARBAROS                YENİKÖY
##                     68                     65                     58
##                   YALI               KURTULUŞ           BAHÇELİEVLER
##                     56                     53                     52
##               İSTİKLAL                 BOZKÖY               ERTUĞRUL
##                     52                     49                     49
##                 OVACIK               GAZİPAŞA                DEREKÖY
##                     49                     42                     41
##           FEVZİ ÇAKMAK                MALTEPE             MİTHATPAŞA
##                     41                     41                     41
##            UĞUR MUMCU                 ÇAMLIK                   GAZİ
##                     41                     38                     38
##            MİMAR SİNAN MUSTAFA KEMAL ATATÜRK               ESENTEPE
##                     38                     38                     37
##                 GÖLCÜK                  TURAN                IŞIKLAR
##                     36                     36                     34
##               MENDERES                AZİZİYE                 YENİCE
##                     34                     32                     31
##               YEŞİLKÖY               EĞRİDERE                 MERKEZ
##                     31                     30                     30
##              YENİŞEHİR               YİĞİTLER            İSMET İNÖNÜ
##                     30                     30                     29
##              İSMETPAŞA                   ONUR                ORTAKÖY
```

```
##                    29                    29                    29
##             ÖRNEKKÖY              ALTINTAŞ                 ILICA
##                    29                    28                    28
##               İSKELE             KASIMPAŞA           KAZIM DİRİK
##                    28                    28                    28
##            KEMALPAŞA                KÜLTÜR                 BARIŞ
##                    28                    28                    27
##              İNKILAP                 IRMAK               KALABAK
##                    27                    27                    27
##        KEMAL ATATÜRK                  ORTA               SİTELER
##                    27                    27                    27
##                 TUNA                YILDIZ               29 EKİM
##                    27                    27                    26
##            AKINCILAR          BAHRİYE ÜÇOK                ÇAĞDAŞ
##                    26                    26                    26
##          DEĞİRMENDERE             FEVZİPAŞA                 HUZUR
##                    26                    26                    26
##        MUSTAFA KEMAL               PAYAMLI                SELÇUK
##                    26                    26                    26
##                SEVGİ               YENİGÜN              YEŞİLOVA
##                    26                    26                    26
##           YUNUS EMRE              19 MAYIS                AKTEPE
##                    26                    25                    25
##             BAHARİYE                CUMALI               ÇAMTEPE
##                    25                    25                    25
##                ÇINAR             DEMİRCİLİ            DUMLUPINAR
##                    25                    25                    25
##    GAZİ MUSTAFA KEMAL                İZKENT             KARŞIYAKA
##                    25                    25                    25
##       KAZIM KARABEKİR               SAKARYA               UMURBEY
##                    25                    25                    25
##                 YAKA                 YAYLA              YENİKENT
##                    25                    25                    25
##            ZEYTİNLİK                 BİRGİ               DUATEPE
##                    25                    24                    24
##                GÜNEY              İHSANİYE              KARAKUYU
##                    24                    24                    24
##             KURUDERE               KUYUCAK              ŞEHİTLER
##                    24                    24                    24
##              (Other)
##                 11309
##
## $user_count
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1      75     275    1025    1269   12633
##
## $avg_consumption
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.05    8.40    9.97   11.60   12.68  603.00
```

```r
###############

paste("Edit for Water Production")

## [1] "Edit for Water Production"

water.production$amount_m3 =
as.numeric(as.character(water.production$amount_m3))
water.production$year = as.character(as.numeric(water.production$year))
water.production$month = as.character(as.numeric(water.production$month))

names <- c('year' ,'month', 'source')
water.production[,names] <- lapply(water.production[,names] , factor)
paste("Summary for Water Production after Edit")

## [1] "Summary for Water Production after Edit"

sapply(water.production, summary)

## $year
## 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##  120  120  132  120  132  144  156  156  156  156  141  108   27
##
## $month
##   1  10  11  12   2   3   4   5   6   7   8   9
## 147 136 134 134 154 136 136 136 136 136 147 136
##
## $source
##         Alaçatı Kutlu Aktaş Barajı                      Balçova Barajı
##                                 71                                 148
##            Buca ve Sarnıç Kuyuları                     Göksu Kuyuları
##                                148                                 148
##                      Gördes Barajı                   Güzelhisar Barajı
##                                124                                 109
##                Halkapınar Kuyuları       Menemen - Çavuşköy Kuyuları
##                                148                                 148
## Ödemiş İçme Suyu Arıtma Tesisleri                 Pınarbaşı Kuyuları
##                                 49                                 148
##                   Sarıkız Kuyuları                     Tahtalı Barajı
##                                148                                 148
##                      Ürkmez Barajı
##                                131
##
## $amount_m3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   98109  467908 1568540 2571146 9786100

#############
paste("Edit for Water Binding")

## [1] "Edit for Water Binding"
```

```
water.binding$binding_time_day <-
str_replace_all(water.binding$binding_time_day , ',', '.')
water.binding$sub_time_day <- str_replace_all(water.binding$sub_time_day ,
',', '.')
water.binding$explore_time_day <-
str_replace_all(water.binding$explore_time_day , ',', '.')
water.binding$collection_time_day <-
str_replace_all(water.binding$collection_time_day , ',', '.')

water.binding$binding_time_day =
as.numeric(as.character(water.binding$binding_time_day))
water.binding$sub_time_day =
as.numeric(as.character(water.binding$sub_time_day))
water.binding$explore_time_day =
as.numeric(as.character(water.binding$explore_time_day))
water.binding$collection_time_day =
as.numeric(as.character(water.binding$collection_time_day))

names <- c('year' ,'district')
water.binding[,names] <- lapply(water.binding[,names] , factor)

paste("Summary for Water Binding after Edit")

## [1] "Summary for Water Binding after Edit"

head(water.binding,20)

##    year   district binding_time_day sub_time_day explore_time_day
## 1  2018    ALAÇATI            24.62        47.40            11.34
## 2  2018     ALİAĞA            16.45        34.71             7.51
## 3  2018   BAYINDIR            15.32        31.24            10.27
## 4  2018   BAYRAKLI            29.30        67.25            27.93
## 5  2018    BERGAMA            16.08        28.88             3.36
## 6  2018     BEYDAĞ             5.78        22.93             1.72
## 7  2018    BORNOVA            59.08       138.75            62.75
## 8  2018       BUCA             7.29        43.81            26.77
## 9  2018     DİKİLİ             8.23        41.12            21.31
## 10 2018       FOÇA             8.11        29.06            12.43
## 11 2018 KARABAĞLAR            27.58        61.24            23.16
## 12 2018  KARABURUN            39.79        67.17             9.49
## 13 2018  KARŞIYAKA            21.21        42.77            15.01
## 14 2018  KEMALPAŞA             8.02        45.54            27.56
## 15 2018      KINIK            25.51        65.47             7.07
## 16 2018      KONAK            28.26        53.70            14.46
## 17 2018      KİRAZ             9.46        18.70             4.20
## 18 2018   MENDERES            36.25        64.28            18.07
## 19 2018    MENEMEN             8.60        24.97             8.27
## 20 2018   MORDOĞAN            36.59        58.98             7.51
##    collection_time_day petition_count
## 1                11.44            279
```

```
## 2                10.74          1762
## 3                 5.66           349
## 4                10.02          2156
## 5                 9.45          1216
## 6                15.43           115
## 7                16.91         11099
## 8                 9.75         15265
## 9                11.57          1135
## 10                8.52          1109
## 11               10.50          8813
## 12               17.89           320
## 13                6.55          7330
## 14                9.96          1930
## 15               32.88           452
## 16               10.98          3130
## 17                5.04           218
## 18                9.96          2616
## 19                8.11          3016
## 20               14.88           187
```

```
sapply(water.binding, summary)
```

```
## $year
## 2018 2019 2020 
##   32   32   32 
## 
## $district
##     ALAÇATI      ALİAĞA    BAYINDIR    BAYRAKLI     BERGAMA      BEYDAĞ 
##           3           3           3           3           3           3 
##     BORNOVA        BUCA    ÇANDARLI       ÇEŞME       ÇİĞLİ      DİKİLİ 
##           3           3           3           3           3           3 
##        FOÇA  KARABAĞLAR   KARABURUN   KARŞIYAKA   KEMALPAŞA       KINIK 
##           3           3           3           3           3           3 
##       KİRAZ       KONAK    MENDERES     MENEMEN    MORDOĞAN   NARLIDERE 
##           3           3           3           3           3           3 
##      ÖDEMİŞ SEFERİHİSAR      SELÇUK        TİRE     TORBALI        URLA 
##           3           3           3           3           3           3 
##      ÜRKMEZ   YENİŞEHİR 
##           3           3 
## 
## $binding_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.070   7.838  16.015  18.094  25.495  64.070 
## 
## $sub_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.58   22.39   35.19   39.75   48.80  138.75 
## 
## $explore_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    1.72    6.16    9.04   12.42   15.22   62.75
##
## $collection_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.290   5.630   8.310   9.233  10.582  32.880
##
## $petition_count
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    81.0   453.5  1469.0  2214.1  2791.5 15265.0
```

WATER CONSUMPTION

Dataset provided by the Izmir Municipality (IBB Acik Veri Portali https://acikveri.bizizmir.com/dataset). The scope of the data is water consumption of 2020 and 2021 of Izmir. Due to this project is running in 2021 June there are limited data about 2021 and we only access the first three months' information of 2021. Dataset provides us that the water consumption data of each of 30 districts and over than1000 neighborhoods. Water consumption of each neighborhood has been given as average consumption out of total users in meter cube. The dataset can be provided by the following link.(https://acikveri.bizizmir.com/dataset/yillik-mahalle-bazli-su-tuketimi/resource/bbe126b7-7a6e-4ae6-9b76-f9462970abc9)

Let's take the Water Consumption of 2020 and show the most 5 consumer districts in Izmir

- Subset the values which are belong the 2020 and summarize the data.
- Create a new feature that refers total consumption rather than average consumption

Note: Due to values are huge for total consumption we'll take the (m^3) / (10^3)

```
consumption.2020 = subset(water.consumption, subset = water.consumption$year
== 2020)
#Created a new feature called consumption which refers total consumption
consumption.2020$consumption_in_month <-
paste((consumption.2020$avg_consumption * consumption.2020$user_count) /
10**3 )
consumption.2020$consumption_in_month =
as.numeric(as.character(consumption.2020$consumption_in_month))
head(consumption.2020,10)
```

```
##    year month district      neighborhood user_count avg_consumption
## 1  2020     1   ALİAĞA           ATATÜRK       2612            9.47
## 2  2020     1   ALİAĞA       AŞAĞIŞAKRAN         89            7.16
## 3  2020     1   ALİAĞA B.HAYRETTİN PAŞA        387            9.08
## 4  2020     1   ALİAĞA         BAHÇEDERE         25            7.83
## 5  2020     1   ALİAĞA            BOZKÖY        170            9.33
## 6  2020     1   ALİAĞA             FATİH        544            8.37
## 7  2020     1   ALİAĞA        GÜZELHİSAR        182            8.48
## 8  2020     1   ALİAĞA        HACIÖMERLİ        307            7.93
```

```
## 9  2020     1   ALİAĞA      HOROZGEDİĞİ         84           15.04
## 10 2020     1   ALİAĞA          KALABAK        136            7.30
##    consumption_in_month
## 1              24.73564
## 2               0.63724
## 3               3.51396
## 4               0.19575
## 5               1.58610
## 6               4.55328
## 7               1.54336
## 8               2.43451
## 9               1.26336
## 10              0.99280
```

```
summary(consumption.2020)
```

```
##    year           month            district        neighborhood
##  2020:12464   6      :1274   KONAK      :1154   ATATÜRK    :  211
##  2021:    0   10     :1221   BERGAMA    : 999   CUMHURİYET:  167
##               12     :1207   ÖDEMİŞ     : 920   FATİH      :   97
##               9      :1206   MENEMEN    : 624   YENİ       :   83
##               7      :1203   KARABAĞLAR: 609   İNÖNÜ      :   77
##               2      :1192   TORBALI    : 603   ZAFER      :   76
##               (Other):5161   (Other)    :7555   (Other)    :11753
##    user_count    avg_consumption  consumption_in_month
##  Min.   :    1   Min.   :  0.05   Min.   :  0.00005
##  1st Qu.:   76   1st Qu.:  8.48   1st Qu.:  0.92532
##  Median :  275   Median : 10.18   Median :  3.59887
##  Mean   : 1030   Mean   : 11.97   Mean   : 10.32297
##  3rd Qu.: 1283   3rd Qu.: 13.20   3rd Qu.: 12.96420
##  Max.   :12633   Max.   :603.00   Max.   :106.27610
##
```

```
# unique(consumption.2020[c("district")])
# unique(consumption.2020[c("neighborhood")])
unique(water.production[c("source")])
```

```
##                                source
## 1                      Tahtalı Barajı
## 2                      Balçova Barajı
## 3                    Sarıkız Kuyuları
## 4          Menemen - Çavuşköy Kuyuları
## 5                  Halkapınar Kuyuları
## 6                   Pınarbaşı Kuyuları
## 7              Buca ve Sarnıç Kuyuları
## 8                       Gördes Barajı
## 9                      Göksu Kuyuları
## 156                     Ürkmez Barajı
## 164         Alaçatı Kutlu Aktaş Barajı
## 268                 Güzelhisar Barajı
## 276 Ödemiş İçme Suyu Arıtma Tesisleri
```

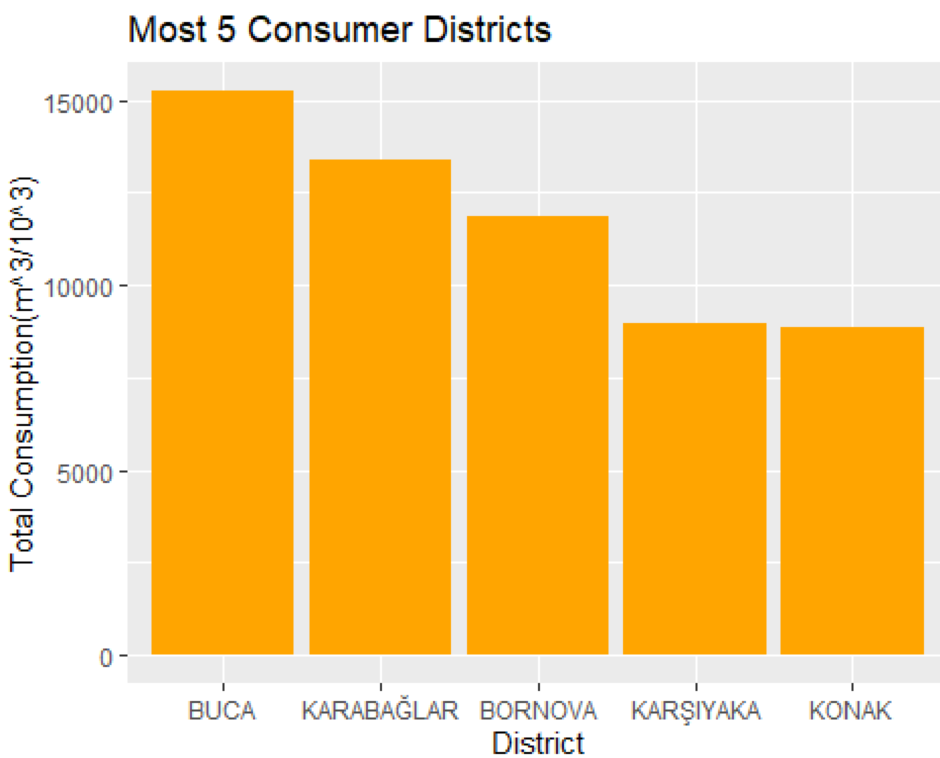
- Show the most consumer districts.

```
#We can see the most consumer districts in 2020 in Izmir.
df1 <- consumption.2020 %>%
  group_by(district) %>%
  summarize(total_consumption = sum(consumption_in_month))

df1 <- df1 %>%
  arrange(desc(total_consumption))

df1 %>% tbl_df %>% print(n=40)

## # A tibble: 30 x 2
##    district    total_consumption
##    <fct>                   <dbl>
##  1 BUCA                   15258.
##  2 KARABAĞLAR             13386.
##  3 BORNOVA                11897.
##  4 KARŞIYAKA               8956.
##  5 KONAK                   8845.
##  6 BAYRAKLI                7662.
##  7 TORBALI                 5977.
##  8 ÇİĞLİ                   5910.
##  9 MENEMEN                 5328.
## 10 ÖDEMİŞ                  5013.
## 11 GAZİEMİR                3590.
## 12 MENDERES                3496.
## 13 KEMALPAŞA               3188.
## 14 ALİAĞA                  2992.
## 15 BERGAMA                 2980.
## 16 ÇEŞME                   2632.
## 17 TİRE                    2534.
## 18 URLA                    2461.
## 19 SEFERİHİSAR             2235.
## 20 BALÇOVA                 2223.
## 21 NARLIDERE               1757.
## 22 FOÇA                    1471.
## 23 DİKİLİ                  1445.
## 24 KİRAZ                   1425.
## 25 BAYINDIR                1424.
## 26 SELÇUK                  1313.
## 27 GÜZELBAHÇE              1195.
## 28 KINIK                    899.
## 29 KARABURUN                719.
## 30 BEYDAĞ                   453.

df1 <- df1 %>% slice_max(total_consumption, n = 5)
```

- Top 5 consumer districts in Izmir

```
p<-ggplot(data=df1, aes(x=reorder(district, -total_consumption),
y=total_consumption)) +
```

```r
  geom_bar(stat="identity", fill="orange")

p <- p + labs(title = "Most 5 Consumer Districts", x='District', y='Total Consumption(m^3/10^3)')
p
```



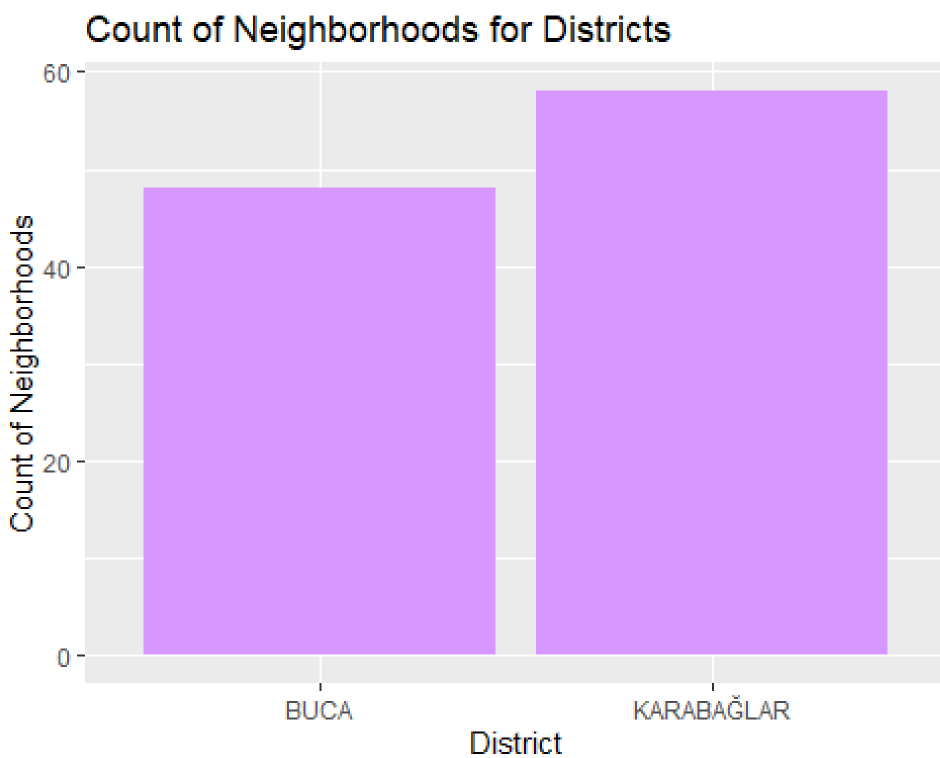
Let's take the most two consumer in 2020 and compare them(Buca and Karabağlar)

```r
top.two.consumer = subset(consumption.2020, subset = consumption.2020$district == "BUCA" | consumption.2020$district == "KARABAĞLAR" )
summary(top.two.consumer)

##    year          month            district          neighborhood
##  2020:1128   10      :106   KARABAĞLAR:609   BARIŞ       :  22
##  2021:   0   11      :106   BUCA      :519   ADATEPE     :  12
##               2      :106   ALİAĞA    :  0   ATATÜRK     :  12
##               6      :106   BALÇOVA   :  0   AYDOĞDU     :  12
##               7      :106   BAYINDIR  :  0   BASIN SİTESİ:  12
##               8      :106   BAYRAKLI  :  0   BOZYAKA     :  12
##               (Other):492   (Other)   :  0   (Other)     :1046
##    user_count      avg_consumption  consumption_in_month
##  Min.   :   1.0   Min.   : 0.360   Min.   : 0.00036
##  1st Qu.: 767.5   1st Qu.: 8.770   1st Qu.: 8.57076
##  Median :2341.0   Median : 9.705   Median :22.15396
##  Mean   :2600.5   Mean   :10.059   Mean   :25.39427
##  3rd Qu.:3760.5   3rd Qu.:10.880   3rd Qu.:35.96159
```

```
##   Max.   :8753.0   Max.   :27.110   Max.   :96.10596
##

## Be sure about taking the Buca and Karabağlar
unique(top.two.consumer[c("district")])

##       district
## 282       BUCA
## 390 KARABAĞLAR
```

- Let's see how many neighborhoods does districts have

```
p <- top.two.consumer %>%
     group_by(district) %>%
     summarise(n=n_distinct(neighborhood)) %>%
     ggplot(., aes(x=district, y=n)) +
            geom_bar(stat='identity',fill="#d896ff")


p + labs(title = "Count of Neighborhoods for Districts", x='District',
y='Count of Neighborhoods')
```



Note: It seems that Buca has less neighborhoods but the water consuming in Buca higher than Karabağlar.

- Let's see the mean water consumption for each month

```
##        districts
## months         BUCA    KARABAĞLAR
```

```
##     1  29.20295708 21.698277719
##     2  29.64263354 19.792049310
##     3  24.03825222 13.681620741
##     4   0.04007923  0.008109167
##     5   0.02998367  0.012565909
##     6  34.44375604 25.728495517
##     7  34.32146375 25.614077069
##     8  34.40346667 25.750499310
##     9  34.77165979 26.139553793
##     10 34.10051854 25.534839310
##     11 32.31859167 24.168855690
##     12 32.79559298 24.002251897
```

Note: It seems that there are some mistakes in data for 4th and 5th months.

Is water consumption correlated with user count of districts?

- Show the most crowded districts.

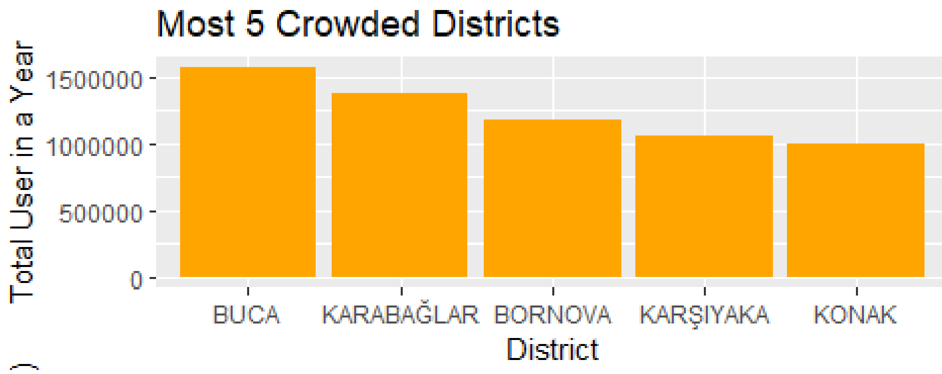
```
#We can see the most consumer districts in 2020 in Izmir.
df <- consumption.2020 %>%
  group_by(district) %>%
  summarize(total_user = sum(user_count))

df <- df %>%
  arrange(desc(total_user))

df %>% tbl_df %>% print(n=40)

## # A tibble: 30 x 2
##    district    total_user
##    <fct>            <int>
##  1 BUCA           1566378
##  2 KARABAĞLAR     1367009
##  3 BORNOVA        1176314
##  4 KARŞIYAKA      1056154
##  5 KONAK           994702
##  6 BAYRAKLI        764765
##  7 ÇİĞLİ           626773
##  8 TORBALI         555687
##  9 MENEMEN         498353
## 10 ÖDEMİŞ          433360
## 11 GAZİEMİR        339891
## 12 MENDERES        330299
## 13 KEMALPAŞA       301729
## 14 BERGAMA         296682
## 15 ALİAĞA          277443
## 16 BALÇOVA         260789
## 17 TİRE            240884
## 18 SEFERİHİSAR     224667
## 19 URLA            218302
```

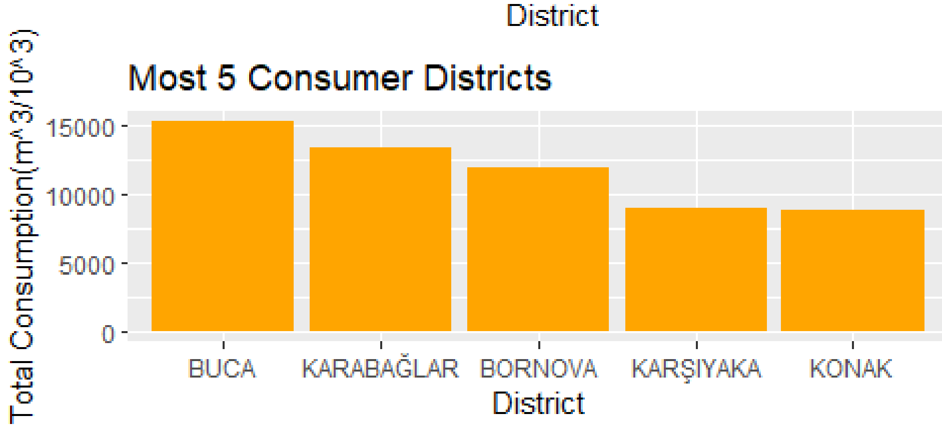
```
## 20 ÇEŞME           192978
## 21 NARLIDERE       175695
## 22 DİKİLİ          163803
## 23 FOÇA            152255
## 24 SELÇUK          125464
## 25 BAYINDIR        114784
## 26 GÜZELBAHÇE      103145
## 27 KİRAZ            92367
## 28 KINIK            74085
## 29 KARABURUN        68105
## 30 BEYDAĞ           39565
```

```
df <- df %>% slice_max(total_user, n = 5)
```

- Take a closer look at most crowded districts and most consumer districts.

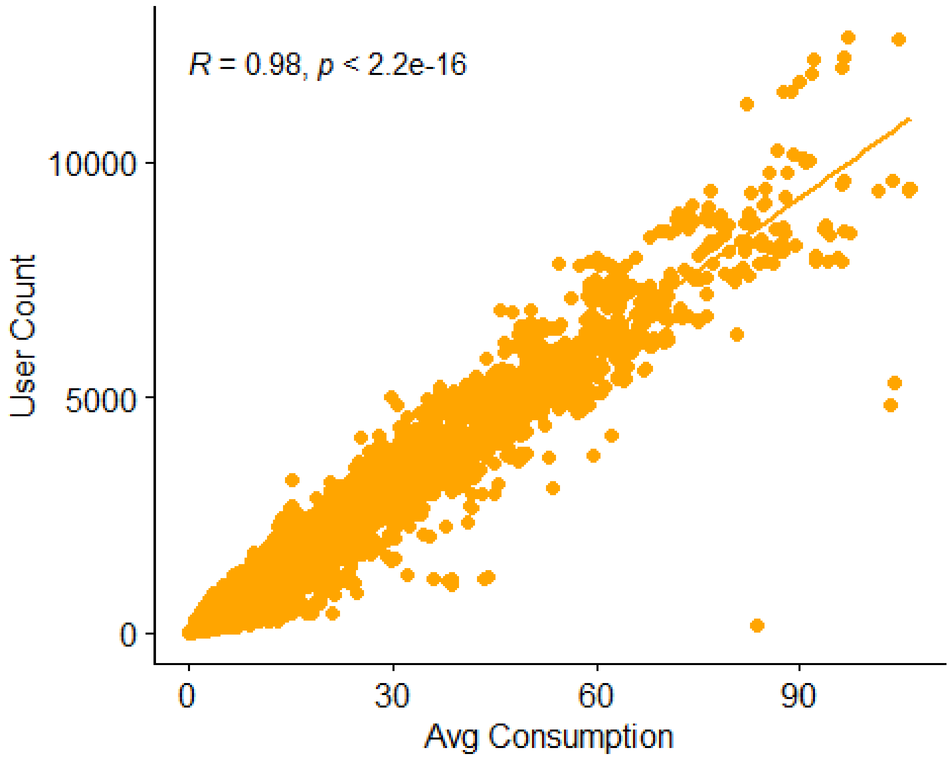
```
p1<-ggplot(data=df, aes(x=reorder(district, -total_user), y=total_user)) +
  geom_bar(stat="identity", fill="orange")

p1 <- p1 + labs(title = "Most 5 Crowded Districts", x='District', y='Total
User in a Year')



p2<-ggplot(data=df1, aes(x=reorder(district, -total_consumption),
y=total_consumption)) +
  geom_bar(stat="identity", fill="orange")

p2 <- p2 + labs(title = "Most 5 Consumer Districts", x='District', y='Total
Consumption(m^3/10^3)')

ggarrange(p1, p2, ncol =1, nrow = 2)
```





- Let's see how consumption changes with the number of users.

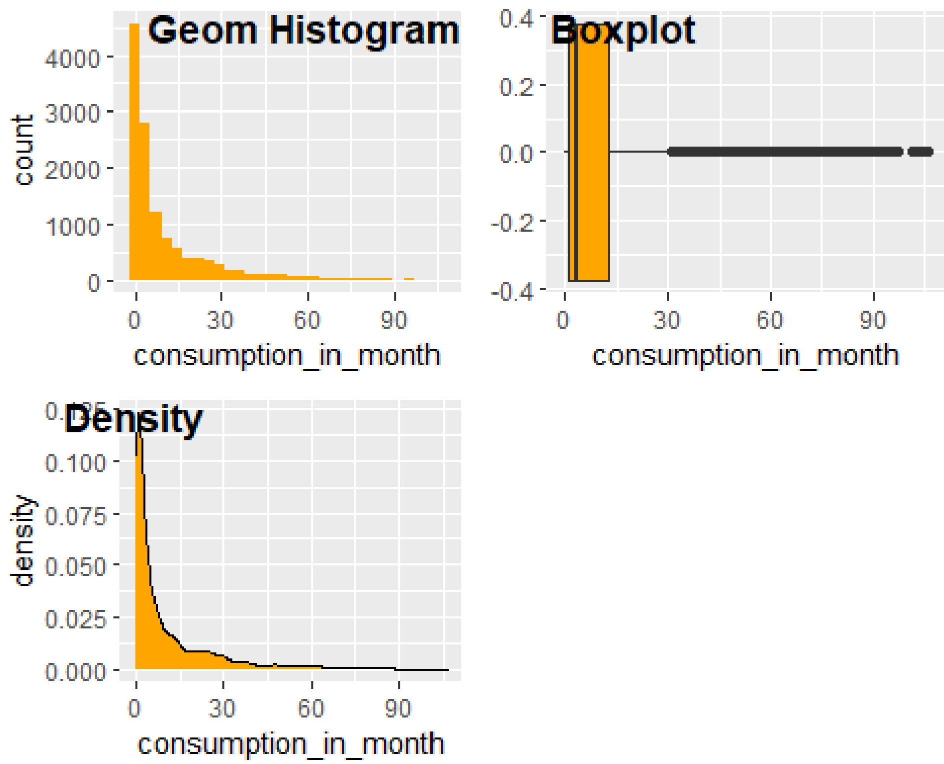
```
ggscatter(consumption.2020, x = "consumption_in_month", y = "user_count",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Avg Consumption", ylab = "User Count", color="orange")

## `geom_smooth()` using formula 'y ~ x'
```




```
cor(consumption.2020$consumption_in_month,consumption.2020$user_count)

## [1] 0.9797726
```

Note: We can clearly see from the above plot there is a highly positive linear correlation between user counts and water consumption. Now let's visualize the data and see the shape of it.

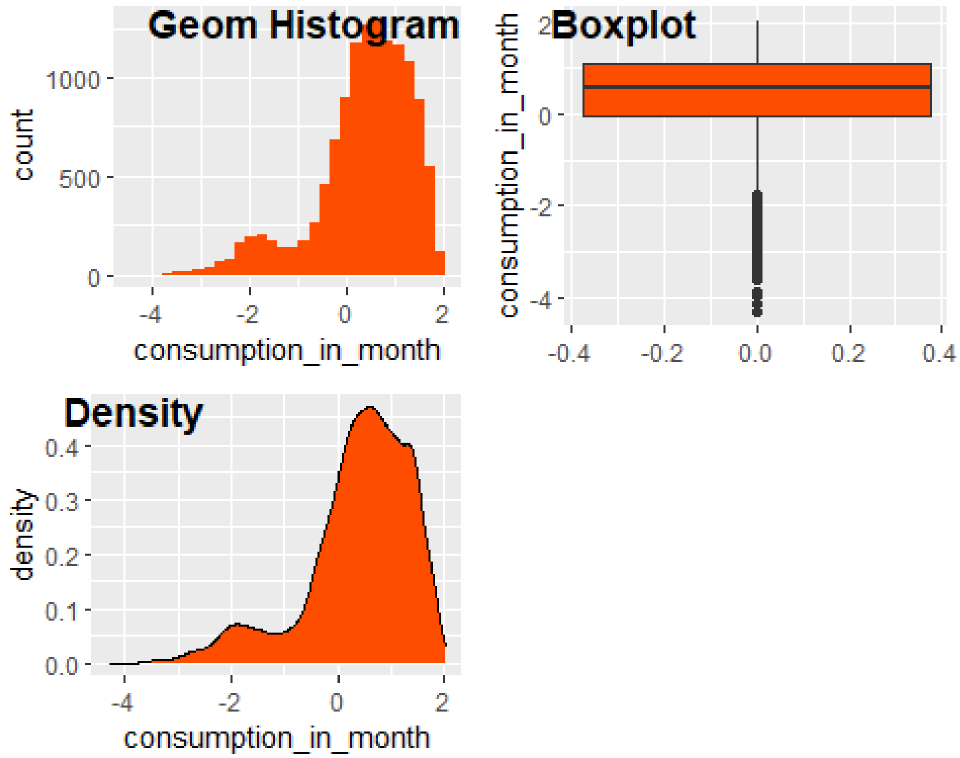
```
base.plot <- ggplot(consumption.2020, aes(x = consumption_in_month)) +
  xlab("consumption_in_month")


p1 <- base.plot + geom_histogram(fill="orange")
p2 <- base.plot + geom_boxplot(fill="orange")
p3 <- base.plot + geom_density(fill="orange")


ggarrange(p1, p2, p3, ncol =2, nrow = 2, labels = c( "Geom
Histogram","Boxplot","Density"))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It seems that we have a right skewed data.

- We'll have log transformation for transform right skewed data to normally distributed data.

```
consumption.2020.transform <- transform(consumption.2020,
            consumption_in_month = log10(consumption_in_month))



base.plot <- ggplot(consumption.2020.transform, aes(x =
consumption_in_month)) +
  xlab("consumption_in_month")


p1 <- base.plot + geom_histogram(fill="#ff4d00")
p2 <-ggplot(consumption.2020.transform, aes(y = consumption_in_month)) +
geom_boxplot(fill="#ff4d00")
p3 <- base.plot + geom_density(fill="#ff4d00")


ggarrange(p1, p2, p3, ncol =2, nrow = 2, labels = c( "Geom
Histogram","Boxplot","Density"))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It seems that log transformation didn't work for transform the data to normally distributed.

From now on we'll assume that our data is normally distributed

Confidence Intervals

Let's take the Water Consumption of 2020 and compare the consumption based on winter and summer

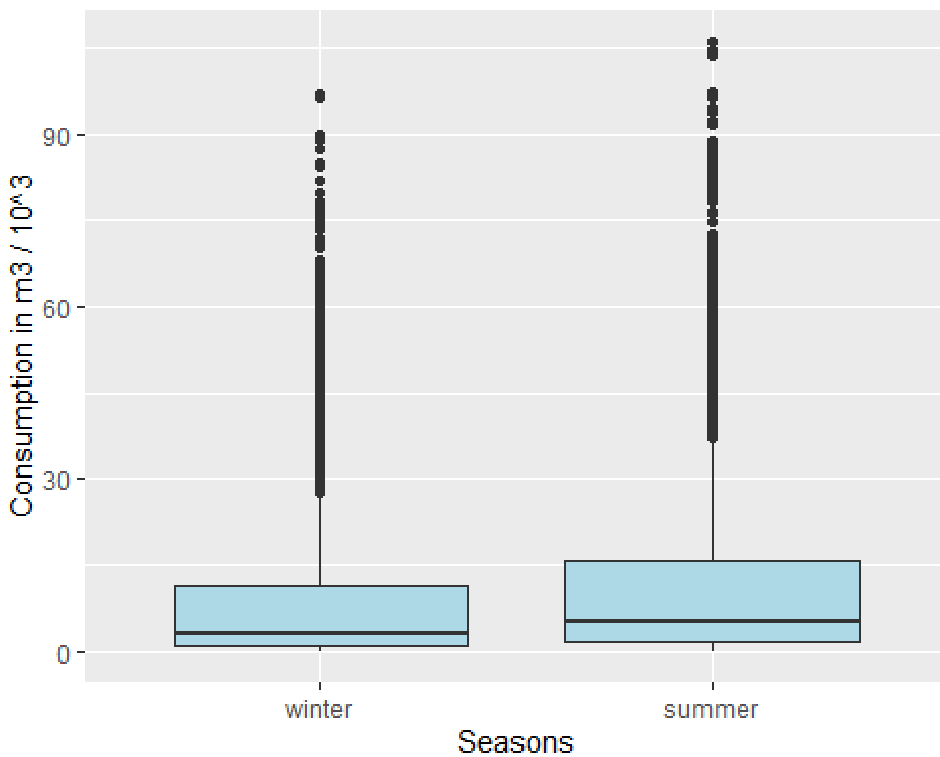
- First we'll categorize the months and just looking at the winter and summer

```
consumption.2020$month = as.numeric(as.character(consumption.2020$month))
consumption.2020$seasons = cut(consumption.2020$month,c(0,2,5,8,11,12))
levels(consumption.2020$seasons) =
c("winter","spring","summer","autumn","winter")

#subset the winter and summer
winter.vs.summer = subset(consumption.2020, subset = consumption.2020$seasons
== "winter" | consumption.2020$seasons == "summer" )
summary(winter.vs.summer)
```

```
##    year          month           district        neighborhood
##  2020:7204   Min.   : 1.000   KONAK  : 666    ATATÜRK    : 114
##  2021:   0   1st Qu.: 2.000   BERGAMA: 599    CUMHURİYET:  90
##              Median : 6.000   ÖDEMİŞ : 568    FATİH      :  54
##              Mean   : 6.001   TİRE   : 368    YENİ       :  48
##              3rd Qu.: 8.000   MENEMEN: 365    İNÖNÜ      :  42
##              Max.   :12.000   TORBALI: 360    ZAFER      :  42
##                               (Other):4278    (Other)    :6814
```

```
##    user_count       avg_consumption  consumption_in_month   seasons
##  Min.   :     1.0   Min.    :  0.11   Min.    :  0.00011    winter:3588
##  1st Qu.:    94.0   1st Qu.:  8.46   1st Qu.:  1.18191    spring:   0
##  Median :   309.5   Median : 10.09   Median :  3.96838    summer:3616
##  Mean   : 1088.4   Mean    : 11.93   Mean    : 10.84167    autumn:   0
##  3rd Qu.: 1401.2   3rd Qu.: 13.00   3rd Qu.: 13.68867
##  Max.   :12633.0   Max.    :603.00   Max.    :106.07188
##
```

```
qplot(x = seasons, y = consumption_in_month,
      geom = "boxplot", data = winter.vs.summer,
      xlab = "Seasons",
      ylab = "Consumption in m3 / 10^3",
      fill = I("lightblue"))
```



The boxplot suggest that winter is associated with less water consumption. There are lots of outliers because of our data is right skewed.

Let's compute a summary table and see it better.

```
aggregate(consumption_in_month ~ seasons,
          data = winter.vs.summer,
          FUN = function(x) {c(mean = mean(x), sd = sd(x))})
```

```
##   seasons consumption_in_month.mean consumption_in_month.sd
## 1  winter                  9.478024               14.601354
## 2  summer                 12.194762               16.600695
```

Now the summary table shows the difference significantly We shouldn't forget that the consumption_in_month represents (m^3 / 10^3). It means that we should multiply the values with 1000 to access real m^3 values of water consumption.

```
#Mean for consumption_in_month(m^3 / 1000)
mean(winter.vs.summer$consumption_in_month)

## [1] 10.84167
```

We assume that we don't have an information about the population varience and we'll use the t-score for calculating Confidence Intervals

Let's take the winter and summer one by one and compare their Confidence Intervals

```
winter = subset(winter.vs.summer, subset = winter.vs.summer$seasons ==
"winter")
summer = subset(winter.vs.summer, subset = winter.vs.summer$seasons ==
"summer")


#Confidence Interval of Winter consumption_in_month(m^3 / 1000)
winter.CI <- t.test(winter$consumption_in_month)$conf.int
winter.CI

## [1] 9.000097 9.955951
## attr(,"conf.level")
## [1] 0.95

#Confidence Interval of Summer consumption_in_month(m^3 / 1000)
summer <- t.test(summer$consumption_in_month)$conf.int
summer

## [1] 11.65350 12.73602
## attr(,"conf.level")
## [1] 0.95
```

Note: We took the 0.95 Confidence Intervals of water consumption in Summer and Winter. And we can see that the confidence intervals are not overlapping and Water consumption in Summer and Winter are significantly different from each ohters.

Before going for a t-test let's check if the variance of two groups are equal with F-test.

```
res.ftest <- var.test(consumption_in_month ~ seasons, data =
winter.vs.summer)
res.ftest

##
##  F test to compare two variances
##
## data:  consumption_in_month by seasons
## F = 0.77363, num df = 3587, denom df = 3615, p-value = 1.489e-14
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.7247052 0.8258662
## sample estimates:
## ratio of variances
##          0.7736307
```

p-value of F-test smaller than alpha(0.05) that mean variances of two sets of data are different. But we'll assume that the variences are equal.

Now let's test our hypothesis using Welch Two Sample t-test

Ho: Mu winter = Mu summer and **H1:** Mu winter != Mu summer

```
winter.summer.t.test <- with(winter.vs.summer,
t.test(x=consumption_in_month[seasons=="summer"],
                     y=consumption_in_month[seasons=="winter"]))
winter.summer.t.test

##
##  Welch Two Sample t-test
##
## data:  consumption_in_month[seasons == "summer"] and
consumption_in_month[seasons == "winter"]
## t = 7.3768, df = 7099.9, p-value = 1.806e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.994794 3.438682
## sample estimates:
## mean of x mean of y
## 12.194762  9.478024

# Calculate difference in means between smoking and nonsmoking groups
winter.summer.diff <- round(((winter.summer.t.test$estimate[1])*1000) -
((winter.summer.t.test$estimate[2])*1000), 1)

# Confidence level as a %
conf.level <- attr(winter.summer.t.test$conf.int, "conf.level") * 100
```

Results: Our study finds that water consumptions are on average 2716.7m^3 higher in the Summer compared to the Winter (t-statistic 7.38, p=0, 95% CI [2, 3.4]m^3 / 10^3)

Conclusion: Our p-value is less than 0.05(alpha) and also our confidence interval doesn't include 0. That's mean Water Consumption in Winter and Summer significantly different from each others. And we can reject that Mu winter is equal to Mu summer(H0)

- Now let's look at the mean water consumption of Buca in 2020 is more than 25(25000m3) or not.

Summarize the water consumption.

```r
summary(consumption.2020$consumption_in_month)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00005   0.92532   3.59887  10.32297  12.96420 106.27610

paste("It seems that mean consumption for 2020 is 10.32(10320m3)")

## [1] "It seems that mean consumption for 2020 is 10.32(10320m3)"

consumption.2020.buca = subset(consumption.2020, subset =
consumption.2020$district == "BUCA")

paste("Print the confidence interval for Buca's total consumption")

## [1] "Print the confidence interval for Buca's total consumption"

buca.CI <- t.test(consumption.2020.buca$consumption_in_month)$conf.int
buca.CI

## [1] 27.25011 31.54903
## attr(,"conf.level")
## [1] 0.95
```

It seems that sample means of Buca water consumption is 95% between the 27.25(27250m^3) and 31.54(31540m^3).

Let’s test the hypothesis and see the result statistically.

- **H0:** Mu buca_water_cons <= 25(25000m^{3}3), **H1:** Mu buca_water_cons > 25(25000m^{3}3)

```r
buca.cons.t.test<- t.test(consumption.2020.buca$consumption_in_month,
alternative = c("greater"), mu = 25, conf.level = 0.95)
buca.cons.t.test

##
##  One Sample t-test
##
## data:  consumption.2020.buca$consumption_in_month
## t = 4.0211, df = 518, p-value = 3.327e-05
## alternative hypothesis: true mean is greater than 25
## 95 percent confidence interval:
##  27.59668      Inf
## sample estimates:
## mean of x
##  29.39957

conf.level <- attr(buca.cons.t.test$conf.int, "conf.level") * 100
```

Result: Our study finds that water consumption for Buca in 2020 are on average, (t-statistic 4.02, p=0, CI: 95% CI [27.6,]m3)

Conclusion: the p-value is almost zero and it's less than alpha(0.05). That means we can reject H0 which says that the average mean of water consumption of Buca in 2020 is less or equal to 25000m3. In other words, we can conclude that there is strong evidence that means of water consumption of Buca in 2020 is greater than 25000m3.

From the most 5 crowded districts Buca and Karabağlar are the districts which has most users. Lets test them if the mean user count of Buca more than Karabağlar or not. First visualize the histogram and check the shape of data.

```
buca.karabag = subset(consumption.2020, subset = (district == "BUCA" |
district == "KARABAĞLAR") & user_count>0 )
summary(buca.karabag)
```

```
##     year          month              district         neighborhood
##  2020:1128   Min.   : 1.000   KARABAĞLAR:609   BARIŞ        :  22
##  2021:   0   1st Qu.: 3.000   BUCA      :519   ADATEPE      :  12
##              Median : 7.000   ALİAĞA    :  0   ATATÜRK      :  12
##              Mean   : 6.773   BALÇOVA   :  0   AYDOĞDU      :  12
##              3rd Qu.:10.000   BAYINDIR  :  0   BASIN SİTESİ:  12
##              Max.   :12.000   BAYRAKLI  :  0   BOZYAKA      :  12
##                               (Other)   :  0   (Other)      :1046
##    user_count     avg_consumption  consumption_in_month   seasons
##  Min.   :   1.0   Min.   : 0.360   Min.   : 0.00036     winter:316
##  1st Qu.: 767.5   1st Qu.: 8.770   1st Qu.: 8.57076     spring:176
##  Median :2341.0   Median : 9.705   Median :22.15396     summer:318
##  Mean   :2600.5   Mean   :10.059   Mean   :25.39427     autumn:318
##  3rd Qu.:3760.5   3rd Qu.:10.880   3rd Qu.:35.96159
##  Max.   :8753.0   Max.   :27.110   Max.   :96.10596
##
```

```
paste("Let's seperate them and see the shapes better")
```

```
## [1] "Let's seperate them and see the shapes better"
```

```
buca = subset(buca.karabag, subset = buca.karabag$district == "BUCA")
karabag = subset(buca.karabag, subset = buca.karabag$district ==
"KARABAĞLAR")
paste("Summary for Buca")
```

```
## [1] "Summary for Buca"
```

```
summary(buca)
```

```
##     year         month            district      neighborhood    user_count
##  2020:519   Min.   : 1.00   BUCA     :519   ADATEPE : 12   Min.   :   1.0
##  2021:  0   1st Qu.: 3.00   ALİAĞA   :  0   ATATÜRK : 12   1st Qu.: 737.5
##             Median : 7.00   BALÇOVA  :  0   AYDOĞDU : 12   Median :2458.0
##             Mean   : 6.73   BAYINDIR:  0   ÇAMLIK  : 12   Mean   :3018.1
##             3rd Qu.:10.00   BAYRAKLI:  0   EFELER  : 12   3rd Qu.:5186.5
##             Max.   :12.00   BERGAMA  :  0   HÜRRİYET: 12   Max.   :8028.0
```

```
##                             (Other) :  0   (Other) :447
##  avg_consumption consumption_in_month   seasons
##  Min.   : 1.08   Min.   : 0.00208     winter:143
##  1st Qu.: 8.89   1st Qu.: 7.21311     spring: 88
##  Median : 9.68   Median :22.64314     summer:144
##  Mean   :10.10   Mean   :29.39957     autumn:144
##  3rd Qu.:10.78   3rd Qu.:51.84913
##  Max.   :27.11   Max.   :96.10596
##
```

```
paste("Summary of Karabağ")
```

```
## [1] "Summary of Karabağ"
```

```
summary(karabag)
```

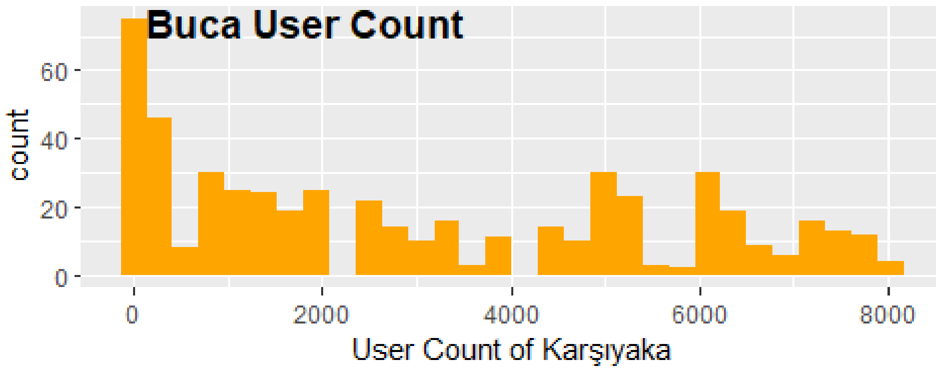
```
##    year          month             district             neighborhood
##  2020:609   Min.   : 1.00   KARABAĞLAR:609   BASIN SİTESİ    : 12
##  2021:  0   1st Qu.: 3.00   ALİAĞA    :  0   BOZYAKA         : 12
##             Median : 7.00   BALÇOVA   :  0   POLİGON         : 12
##             Mean   : 6.81   BAYINDIR  :  0   REFET BELE      : 12
##             3rd Qu.:10.00   BAYRAKLI  :  0   ALİ FUAT CEBESOY: 11
##             Max.   :12.00   BERGAMA   :  0   BARIŞ           : 11
##                             (Other)   :  0   (Other)         :539
##    user_count   avg_consumption consumption_in_month   seasons
##  Min.   :   1   Min.   : 0.36   Min.   : 0.00036     winter:173
##  1st Qu.: 881   1st Qu.: 8.63   1st Qu.:10.44981     spring: 88
##  Median :2064   Median : 9.76   Median :21.74220     summer:174
##  Mean   :2245   Mean   :10.02   Mean   :21.98089     autumn:174
##  3rd Qu.:3132   3rd Qu.:11.00   3rd Qu.:30.88900
##  Max.   :8753   Max.   :25.54   Max.   :83.06597
##
```

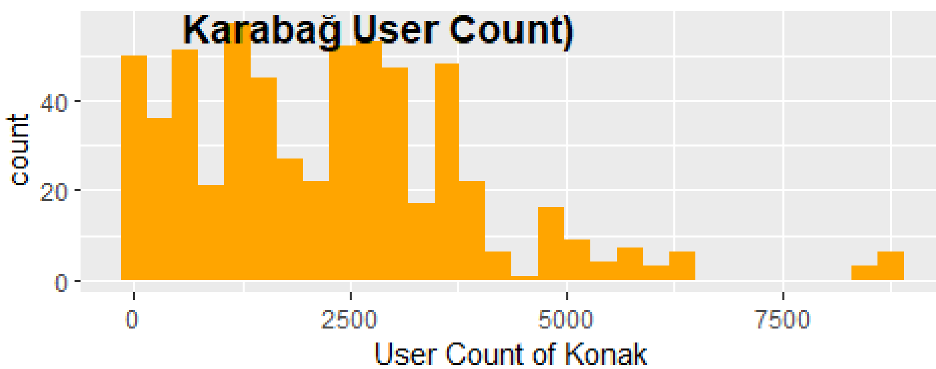
```
paste("Draw the histgorams and see the shape of them")
```

```
## [1] "Draw the histgorams and see the shape of them"
```

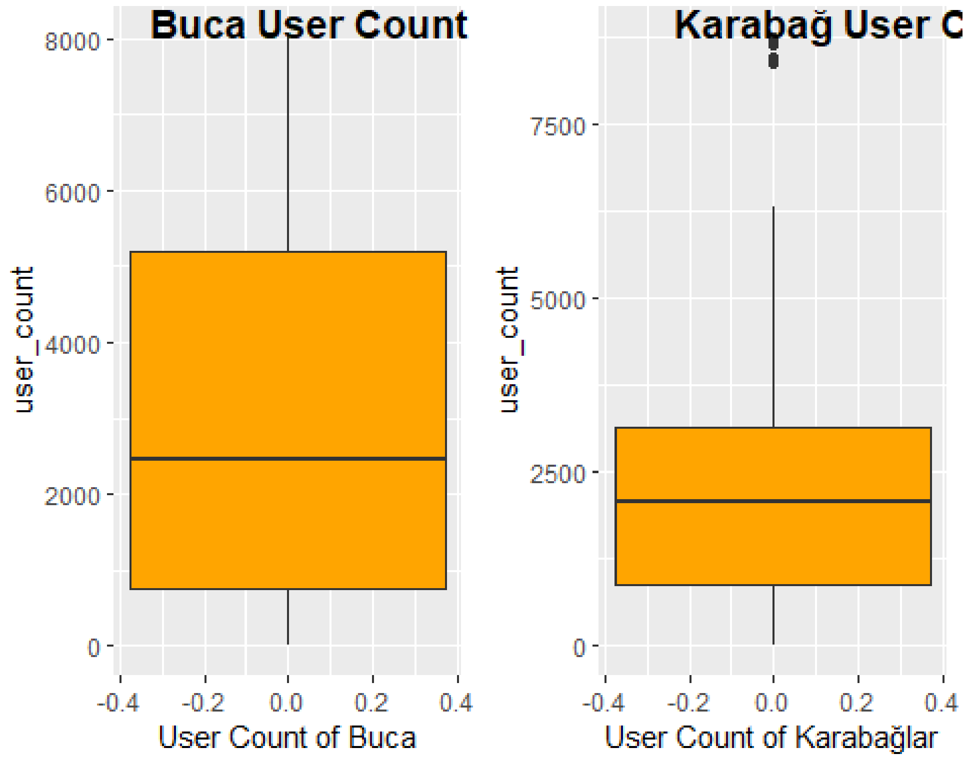
```
p1 <- ggplot(buca, aes(x = user_count)) +
  xlab("User Count of Karşıyaka")
# Violin plot
p1 <- p1 + geom_histogram(fill="orange")


p2 <- ggplot(karabag, aes(x = user_count)) +
  xlab("User Count of Konak")
# Violin plot
p2 <- p2 + geom_histogram(fill="orange")

ggarrange(p1, p2, ncol =1, nrow = 2, labels = c( "Buca User Count","Karabağ
User Count)"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```






```
paste("Check for the outliers")

## [1] "Check for the outliers"

p1 <- ggplot(buca, aes(y = user_count)) +
  xlab("User Count of Buca")
# Violin plot
p1 <- p1 + geom_boxplot(fill="orange")


p2 <- ggplot(karabag, aes(y = user_count)) +
  xlab("User Count of Karabağlar")
# Violin plot
p2 <- p2 + geom_boxplot(fill="orange")

ggarrange(p1, p2, ncol =2, nrow = 1, labels = c( "Buca User Count","Karabağ
User Count)"))
```



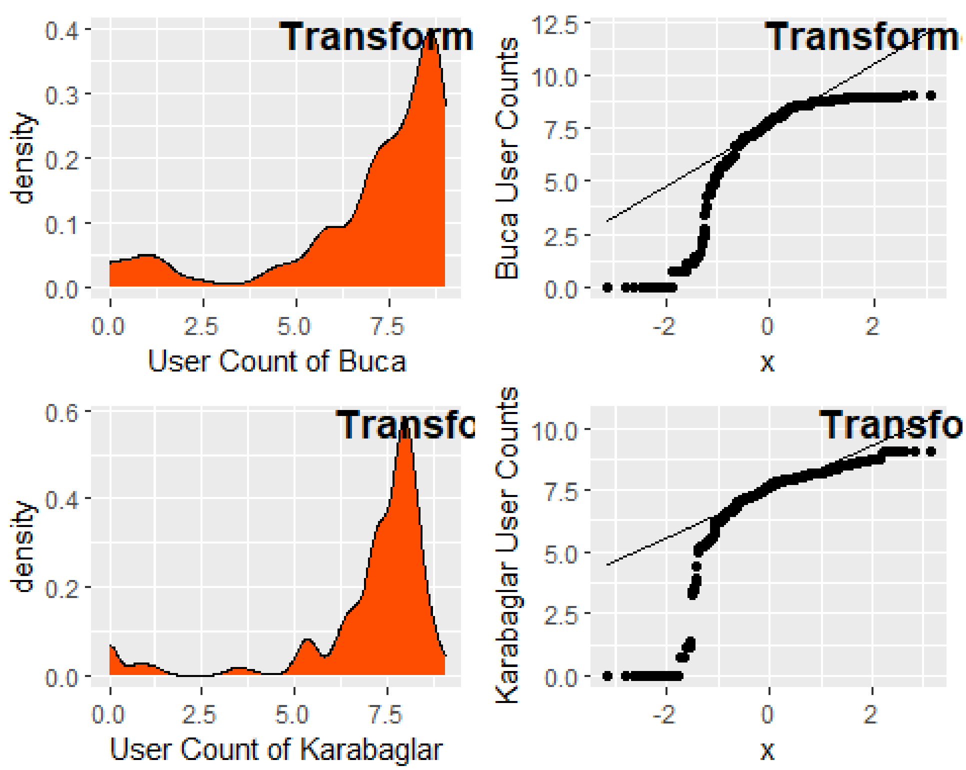

```
with(buca, shapiro.test(user_count))# p < 0.05

##
##  Shapiro-Wilk normality test
##
## data:  user_count
## W = 0.89796, p-value < 2.2e-16

with(karabag, shapiro.test(user_count))# p < 0.05

##
##  Shapiro-Wilk normality test
##
## data:  user_count
## W = 0.92899, p-value < 2.2e-16
```

From the boxplot, we can see that the Buca has a wider boxplot. That means the variance is bigger for Buca. Also, there are some outliers for Karabağlar but despite that, the shape of data seems more symmetric than Buca.

It seems that both data is right skewed. Let's try log transformation and look data again.

```
buca.log <- transform(buca,
            user_count = log(user_count))

karabag.log <- transform(karabag,
            user_count = log(user_count))
```

```r
p1 <- ggplot(buca.log, aes(x = user_count)) +
  xlab("User Count of Buca")
# Violin plot
p1 <- p1 + geom_density(fill="#ff4d00")


p2 <- ggplot(karabag.log, aes(x = user_count)) +
  xlab("User Count of Karabaglar")
# Violin plot
p2 <- p2 + geom_density(fill="#ff4d00")


p3 <- ggplot(buca.log, aes(sample = user_count)) + ylab("Buca User Counts") +
  stat_qq() +
  stat_qq_line()

p4 <- ggplot(karabag.log, aes(sample = user_count)) + ylab("Karabaglar User
Counts") +
  stat_qq() +
  stat_qq_line()


ggarrange(p1, p3, p2, p4, ncol =2, nrow = 2, labels = c( "Transformed Buca
User Count","Transformed Buca User Count","Transformed Karabaglar User
Count","Transformed Karabaglar User Count"))
```



```r
with(buca.log, shapiro.test(user_count))# p < 0.05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  user_count
## W = 0.75224, p-value < 2.2e-16
```

```r
with(karabag.log, shapiro.test(user_count))# p < 0.05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  user_count
## W = 0.66962, p-value < 2.2e-16
```

It seems that we failed to normally distribute our data and from now on we'll assume that data normally distributed. Now take a look at the confidence intervals of Buca's and Karabaglar's user counts.

```r
buca.CI <- t.test(buca$user_count)$conf.int
buca.CI
```

```
## [1] 2797.537 3238.602
## attr(,"conf.level")
## [1] 0.95
```

```r
karabag.CI <- t.test(karabag$user_count)$conf.int
karabag.CI
```

```
## [1] 2110.756 2378.600
## attr(,"conf.level")
## [1] 0.95
```

It seems that these two districts are significantly different from each other and Buca's mean of user counts more than Karabağlar. Let's test them and see the difference better. Before test it we'll first have F-test to see if their variance are same or not.

```r
res.ftest <- var.test(buca$user_count,karabag$user_count)
res.ftest
```

```
##
##  F test to compare two variances
##
## data:  buca$user_count and karabag$user_count
## F = 2.3093, num df = 518, denom df = 608, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.957396 2.727610
## sample estimates:
## ratio of variances
##           2.309342
```

p-value of F-test smaller than alpha(0.05) that mean variances of two sets of data are different. But we'll assume that the variences are equal.

H0: Mu user_count_Buca <= Mu user_count_Karabag, **H1:**Mu user_count_Buca > Mu user_count_Karabag

```
buca.karabag.t.test <- t.test(buca$user_count, karabag$user_count, var.equal
= TRUE, alternative = "greater")
buca.karabag.t.test

##
##  Two Sample t-test
##
## data:  buca$user_count and karabag$user_count
## t = 6.0773, df = 1126, p-value = 8.351e-10
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  563.8969      Inf
## sample estimates:
## mean of x mean of y
##  3018.069  2244.678

# Calculate difference in means between smoking and nonsmoking groups
user.count.diff <- round(buca.karabag.t.test$estimate[1] -
buca.karabag.t.test$estimate[2], 1)

# Confidence level as a %
conf.level <- attr(buca.karabag.t.test$conf.int, "conf.level") * 100
```

Results: Our study finds that user counts are on average 773.4person higher in the Buca compared to the Karabağlar (t-statistic 6.08, p=0, 95% CI [563.9,]person)

Conclusion: Our p-value is too close to 0 and smaller than alpha(0.05) therefore we can reject that average user count of Buca less than Karabağlar. In other words, average user count of Buca higher than Karabağlar.

ANOVA

- There are 10 neighborhoods in Narlıdere. Let's see if the mean consumptions of these neighborhoods are same or not in 2020.

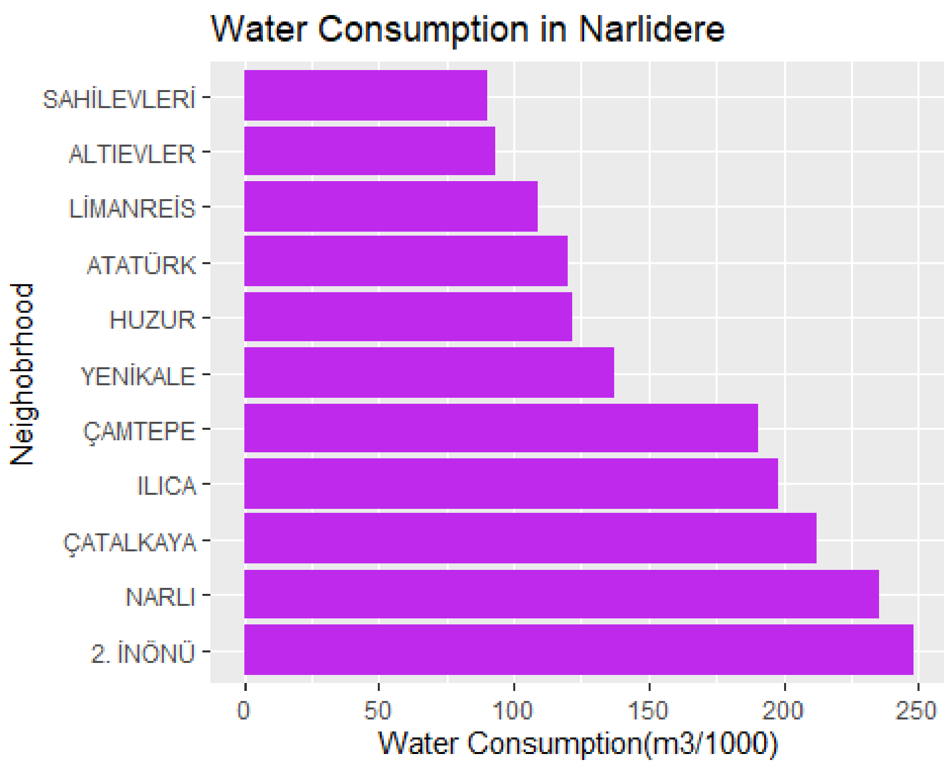
-Summarize the data

```
narlidere = subset(consumption.2020, subset = (district == "NARLIDERE") &
(consumption_in_month > 0))
summary(narlidere)

##    year          month              district        neighborhood
user_count
##  2020:124   Min.   : 1.000   NARLIDERE:124   2. İNÖNÜ   :12    Min.   :
1
##  2021:  0   1st Qu.: 3.000   ALİAĞA   :  0   NARLI      :12    1st Qu.:
```

```
547
##              Median : 7.000   BALÇOVA   :  0   SAHİLEVLERİ:12    Median
:1476
##              Mean   : 6.645   BAYINDIR :  0   YENİKALE   :12    Mean
:1417
##              3rd Qu.:10.000   BAYRAKLI :  0   ALTIEVLER  :11    3rd
Qu.:2199
##              Max.   :12.000   BERGAMA  :  0   ATATÜRK    :11    Max.
:2802
##                               (Other)  :  0   (Other)    :54
##  avg_consumption  consumption_in_month   seasons
##  Min.   : 1.050   Min.   : 0.00105     winter:33
##  1st Qu.: 8.850   1st Qu.:10.07747     spring:25
##  Median : 9.775   Median :13.19917     summer:33
##  Mean   :11.407   Mean   :14.16963     autumn:33
##  3rd Qu.:11.783   3rd Qu.:19.82099
##  Max.   :81.150   Max.   :29.07366
##
```

We can see that the mean consumption in Narlidere is 14.16(14160m^3). Let's see it for the neighborhoods.

```
p1<-ggplot(data=narlidere, aes(y=reorder(neighborhood, -
consumption_in_month), x=consumption_in_month)) +
  geom_bar(stat="identity", fill="#be29ec")

p1 <- p1 + labs(title = "Water Consumption in Narlidere", x='Water
Consumption(m3/1000)', y='Neighobrhood')

p1
```

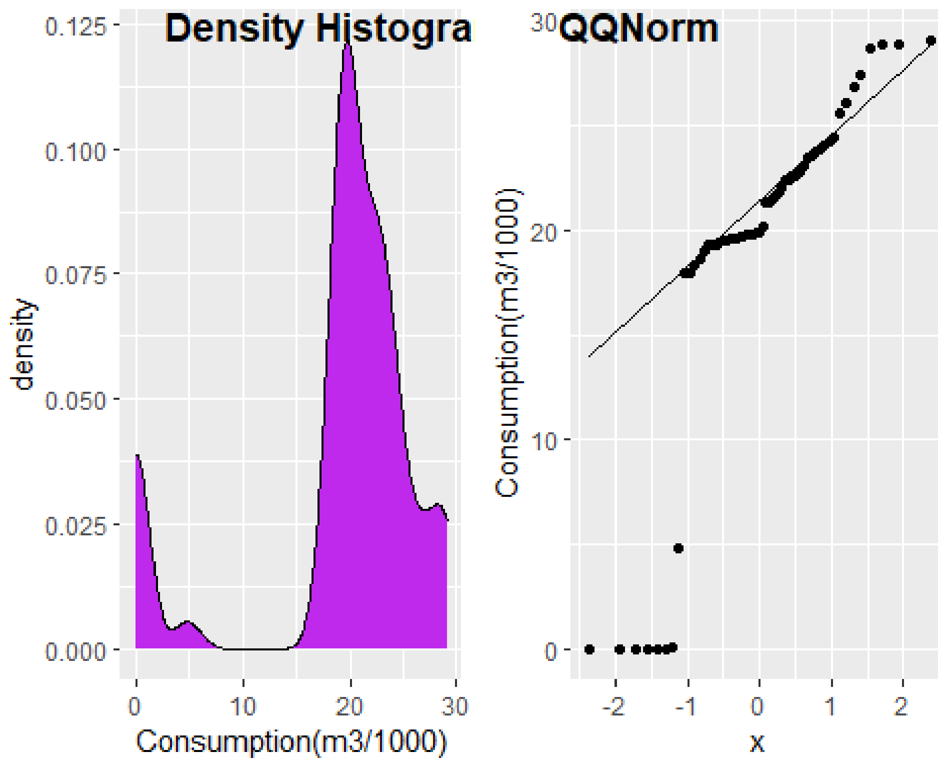


It seems that there is a significant differences between most 5 consumer neighboors and rest. Therefore we'll just have the ANOVA test for first five most consumer districts.

```
narlidere.most5 = subset(narlidere, subset = (neighborhood == "2. İNÖNÜ" |
neighborhood == "NARLI" |
                                                neighborhood == "ÇATALKAYA"|
neighborhood == "ILICA" | neighborhood == "ÇAMTEPE"))


summary(narlidere.most5)

##    year         month              district       neighborhood   user_count
##  2020:57   Min.   : 1.000   NARLIDERE:57   2. İNÖNÜ :12    Min.   :   1
##  2021: 0   1st Qu.: 3.000   ALİAĞA   : 0   NARLI    :12    1st Qu.:2133
##            Median : 7.000   BALÇOVA  : 0   ÇAMTEPE  :11    Median :2226
##            Mean   : 6.614   BAYINDIR : 0   ÇATALKAYA:11    Mean   :2030
##            3rd Qu.:10.000   BAYRAKLI : 0   ILICA    :11    3rd Qu.:2441
##            Max.   :12.000   BERGAMA  : 0   1.KADRİYE: 0    Max.   :2802
##                             (Other)  : 0   (Other)  : 0
##  avg_consumption  consumption_in_month   seasons
##  Min.   : 1.050   Min.   : 0.00105     winter:15
##  1st Qu.: 8.620   1st Qu.:19.28576     spring:12
##  Median : 9.340   Median :19.89496     summer:15
##  Mean   : 9.115   Mean   :19.03658     autumn:15
##  3rd Qu.:10.040   3rd Qu.:23.50353
##  Max.   :11.940   Max.   :29.07366
##
```

Let's see the shape of data.

```
p1 <- ggplot(narlidere.most5, aes(x = consumption_in_month)) +
  xlab("Consumption(m3/1000)")
# Violin plot
p1 <- p1 + geom_density(fill="#be29ec")


p2 <- ggplot(narlidere.most5, aes(sample = consumption_in_month)) +
ylab("Consumption(m3/1000)") +
  stat_qq() +
  stat_qq_line()

ggarrange(p1, p2, ncol =2, nrow = 1, labels = c( "Density
Histogram","QQNorm"))
```




```
with(narlidere.most5, shapiro.test(consumption_in_month))# p < 0.05

##
##  Shapiro-Wilk normality test
##
## data:  consumption_in_month
## W = 0.74712, p-value = 1.478e-08
```

As we can see from the above plots the data is not normally distributed. Also p-value obtained from Shapiro test is almost 0 and less than alpha(0.05) therefore we reject that the data normally distributed. Try to transform data with the log transformation.

```
narlidere.most5.log <- transform(narlidere.most5,
            consumption_in_month = log10(consumption_in_month))

with(narlidere.most5.log, shapiro.test(consumption_in_month))# p < 0.05

##
##  Shapiro-Wilk normality test
##
## data:  consumption_in_month
## W = 0.46624, p-value = 4.356e-13
```

After log transformation the p-value from Shapiro test still less than 0.05 therefore we again reject that the data is normally distributed. From now on we'll assume that the data normally distributed.

Now we can compute the One-way ANOVA test with the original data.

H0: All the five means of water consumptions of neighborhoods are same, **H1:** At least one differ.

```
# Compute the analysis of variance
res.aov <- aov(consumption_in_month ~ neighborhood, data = narlidere.most5)
summary(res.aov)

##              Df Sum Sq Mean Sq F value Pr(>F)
## neighborhood  4     80   19.94   0.291  0.882
## Residuals    52   3562   68.50

TukeyHSD(res.aov)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = consumption_in_month ~ neighborhood, data =
narlidere.most5)
##
## $neighborhood
##                            diff        lwr       upr     p adj
## ÇAMTEPE-2. İNÖNÜ   -3.3192395 -13.081791  6.443312 0.8712977
## ÇATALKAYA-2. İNÖNÜ -1.3334586 -11.096010  8.429093 0.9951524
## ILICA-2. İNÖNÜ     -2.6558668 -12.418418  7.106685 0.9384177
## NARLI-2. İNÖNÜ     -1.0569158 -10.604880  8.491048 0.9978551
## ÇATALKAYA-ÇAMTEPE   1.9857809  -7.986742 11.958304 0.9798341
## ILICA-ÇAMTEPE       0.6633727  -9.309150 10.635896 0.9997114
## NARLI-ÇAMTEPE       2.2623237  -7.500228 12.024875 0.9649664
## ILICA-ÇATALKAYA    -1.3224082 -11.294931  8.650115 0.9956754
## NARLI-ÇATALKAYA     0.2765428  -9.486009 10.039094 0.9999904
## NARLI-ILICA         1.5989510  -8.163600 11.361502 0.9902956
```

Conclusion: From the above table, we can see the p-value of the ANOVA test is 0.882 which greater than alpha(0.05). Then, we can not reject H0 which refers to five neighborhoods in

Narlidere on average water consumption are equal. In other words, there is no strong evidence at least one differs. Also, we can see it better from the Tukey test. All the bounds of pairs include 0. That means there is no significant difference between the average water consumption of these pairs.

WATER BINDING

Dataset provided by the Izmir Municipality (IBB Acik Veri Portali https://acikveri.bizizmir.com/dataset). The scope of the data is First Subscription Water Connection Realization Times in Izmir between 2018 and 2020. Dataset provides us the; average water-binding time(days), average subscription time(days), average water exploring time(days), average tax collection time(days), and petition counts for the water-binding by 32 districts. The dataset can be provided by the following link.(https://acikveri.bizizmir.com/dataset/ilk-abonelik-su-ve-kanal-baglama-gerceklesme-sureleri/resource/56ff9ac3-cddd-407c-8794-bb0c541bfdf4)

- Let's see the data again

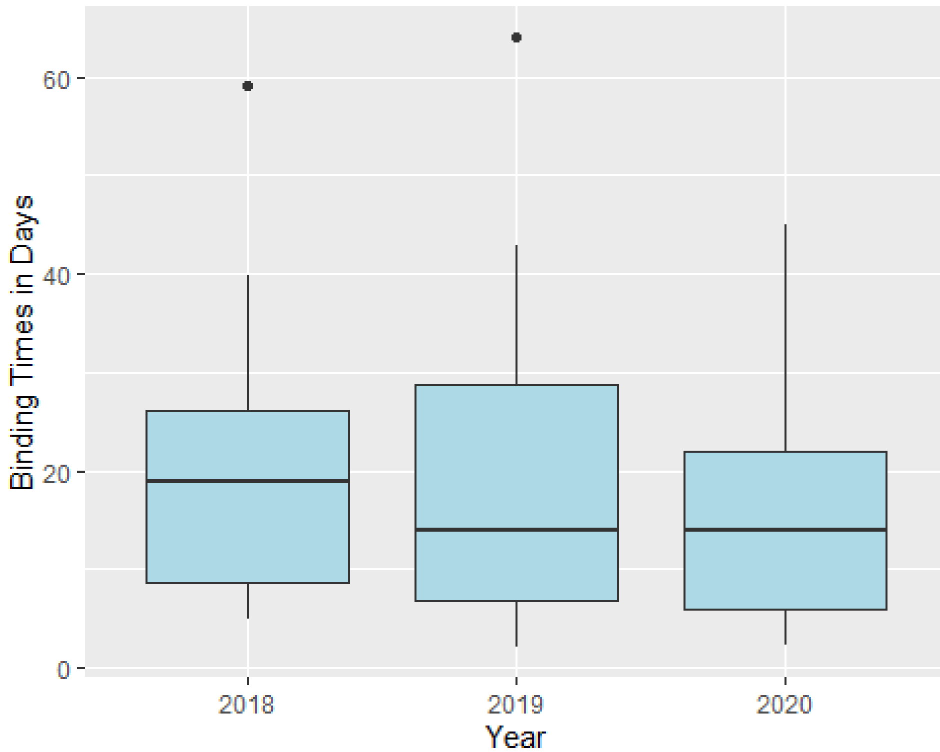
```
head(water.binding,10)

##    year district binding_time_day sub_time_day explore_time_day
## 1  2018  ALAÇATI            24.62        47.40            11.34
## 2  2018   ALİAĞA            16.45        34.71             7.51
## 3  2018 BAYINDIR            15.32        31.24            10.27
## 4  2018 BAYRAKLI            29.30        67.25            27.93
## 5  2018  BERGAMA            16.08        28.88             3.36
## 6  2018   BEYDAĞ             5.78        22.93             1.72
## 7  2018  BORNOVA            59.08       138.75            62.75
## 8  2018     BUCA             7.29        43.81            26.77
## 9  2018   DİKİLİ             8.23        41.12            21.31
## 10 2018     FOÇA             8.11        29.06            12.43
##    collection_time_day petition_count
## 1                11.44            279
## 2                10.74           1762
## 3                 5.66            349
## 4                10.02           2156
## 5                 9.45           1216
## 6                15.43            115
## 7                16.91          11099
## 8                 9.75          15265
## 9                11.57           1135
## 10                8.52           1109

sapply(water.binding, summary)

## $year
## 2018 2019 2020
```

```
##   32   32   32
##
## $district
##      ALAÇATI       ALİAĞA     BAYINDIR     BAYRAKLI      BERGAMA       BEYDAĞ
##            3            3            3            3            3            3
##      BORNOVA         BUCA     ÇANDARLI        ÇEŞME        ÇİĞLİ       DİKİLİ
##            3            3            3            3            3            3
##         FOÇA   KARABAĞLAR    KARABURUN    KARŞIYAKA    KEMALPAŞA        KINIK
##            3            3            3            3            3            3
##        KİRAZ        KONAK     MENDERES      MENEMEN     MORDOĞAN    NARLIDERE
##            3            3            3            3            3            3
##       ÖDEMİŞ  SEFERİHİSAR       SELÇUK         TİRE      TORBALI         URLA
##            3            3            3            3            3            3
##       ÜRKMEZ    YENİŞEHİR
##            3            3
##
## $binding_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.070   7.838  16.015  18.094  25.495  64.070
##
## $sub_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.58   22.39   35.19   39.75   48.80  138.75
##
## $explore_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.72    6.16    9.04   12.42   15.22   62.75
##
## $collection_time_day
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.290   5.630   8.310   9.233  10.582  32.880
##
## $petition_count
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    81.0   453.5  1469.0  2214.1  2791.5 15265.0
```

Visualize the data

- Let's take a closer look to binding time. We can explore the shape of the data and also check for the outliers

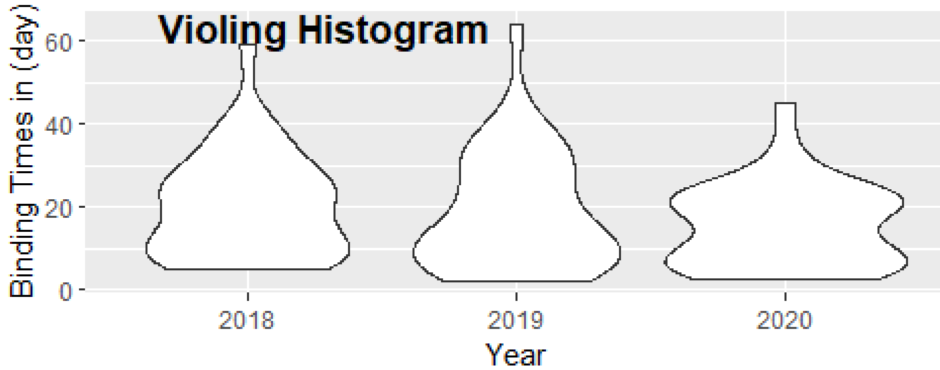
```
#boxplot(birthwt$birthwt.grams~birthwt$hypertension)


qplot(x =year, y = binding_time_day,
      geom = "boxplot", data = water.binding,
      xlab = "Year",
      ylab = "Binding Times in Days",
      fill = I("lightblue"))
```

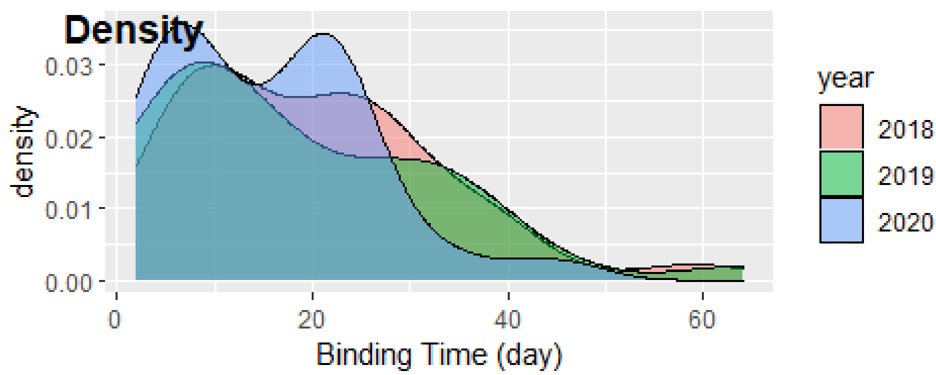


We can see the above boxplot explain the binding times for years. It suggest that in 2020 binding requires less time than other years and it's more symetric than others and narrower than others that is mean that standard deviation of 2020 is a bit smaller than others(Binding time doesn't change a lot). Also in 2019 it seems that the median is also same with in 2020 but the boxplot is wider that is mean the binding time changes much.

- Also plot the violin plot and density functions to see the difference better.

```
base.plot <- ggplot(water.binding, aes(x = year, y = binding_time_day)) +
  xlab("Year") +
  ylab("Binding Times in (day)")
# Violin plot
p1 <- base.plot + geom_violin()



#Density functions
base.plot <- ggplot(water.binding, aes(x = binding_time_day)) +
  xlab("Binding Time (day)")
p2 <- base.plot + geom_density(aes(fill = year), alpha = 0.5)

ggarrange(p1, p2, ncol =1, nrow = 2, labels = c( "Violing
Histogram","Density"))
```

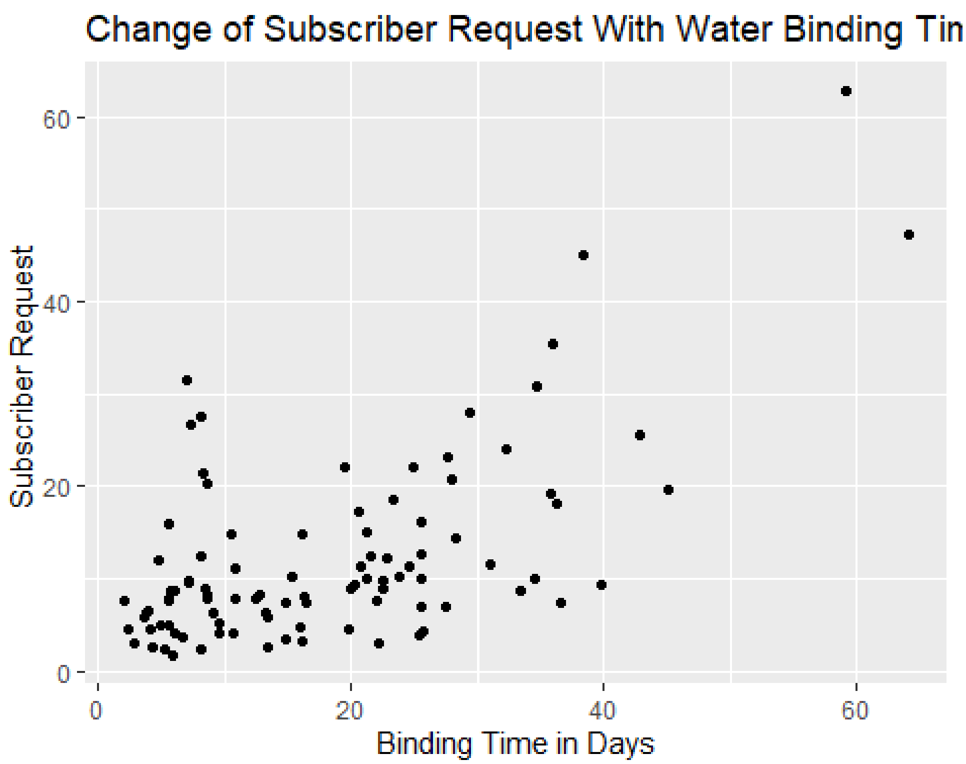




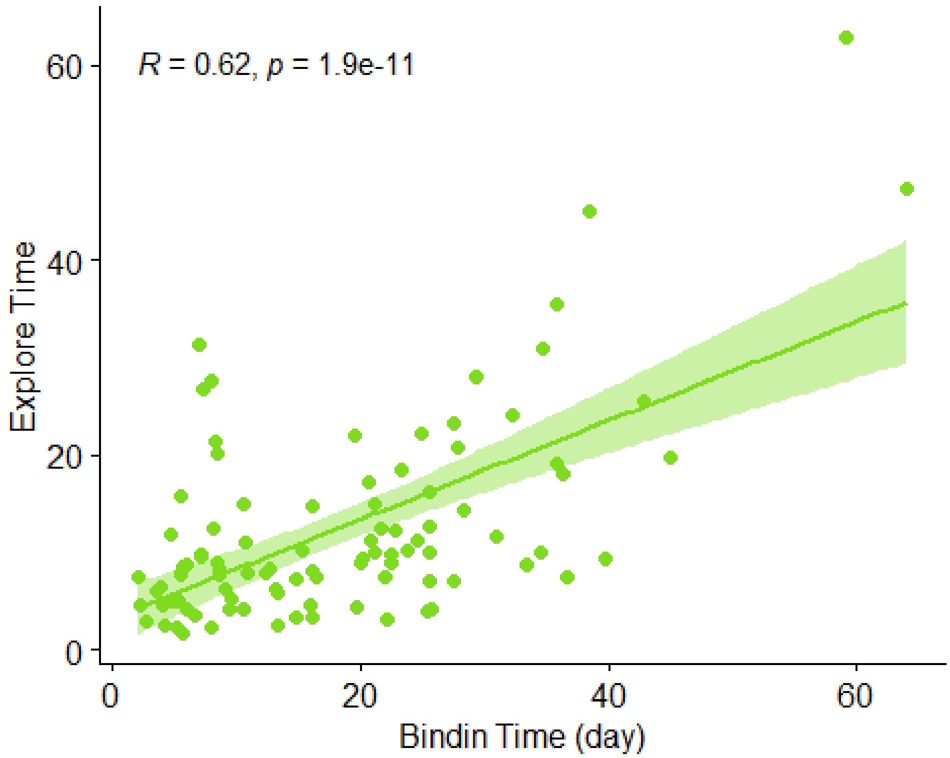
As we can see from above plots our three data are right skewed but the it suggest us that the 2020 has the less binding time.

- Let's discover the correlation between binding time and exploring time.

```
p <- ggplot(data=water.binding, aes(x=binding_time_day, y=explore_time_day))
+ geom_point()
p + labs(title = "Change of Subscriber Request With Water Binding Time",
x='Binding Time in Days', y='Subscriber Request ')
```



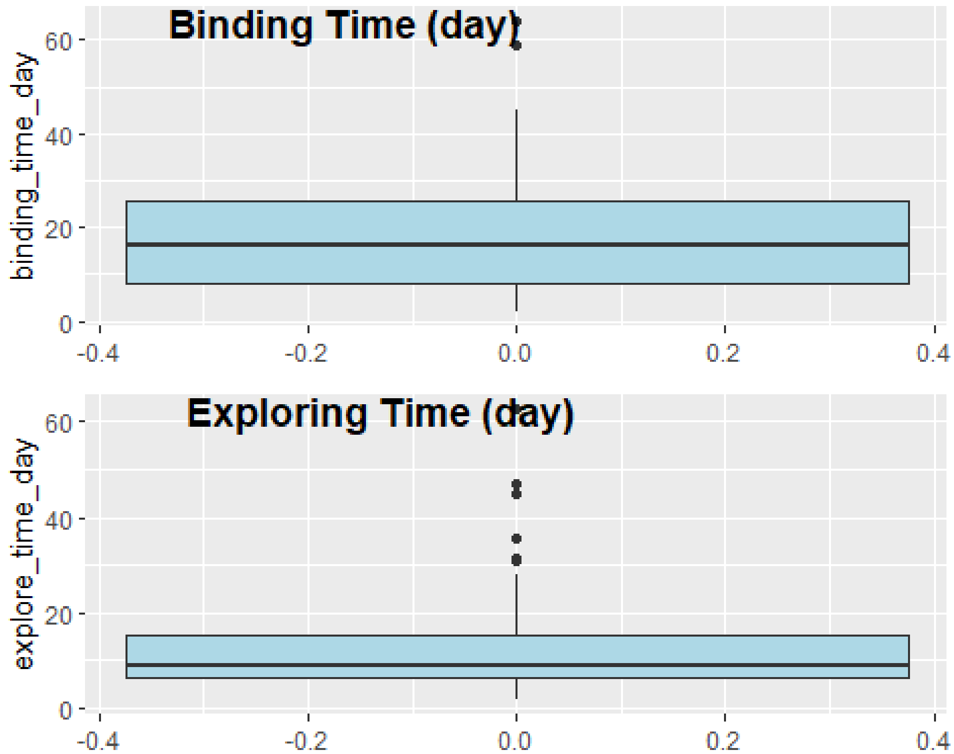

```
ggscatter(water.binding, x = "binding_time_day", y = "explore_time_day",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Bindin Time (day)", ylab = "Explore Time", color="#7ddc1f",
fill="blue")

## `geom_smooth()` using formula 'y ~ x'
```



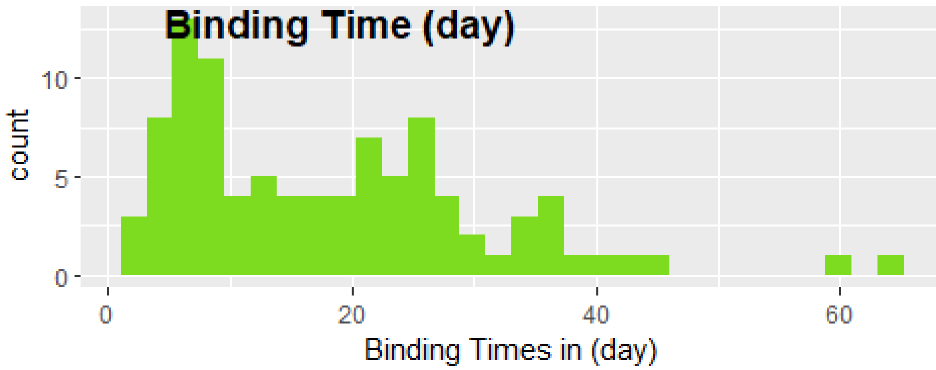
From above plots the Pearson coefficient correlation is 0.62. We can not say there is a high correlation but also can not say there is no correlation between them.

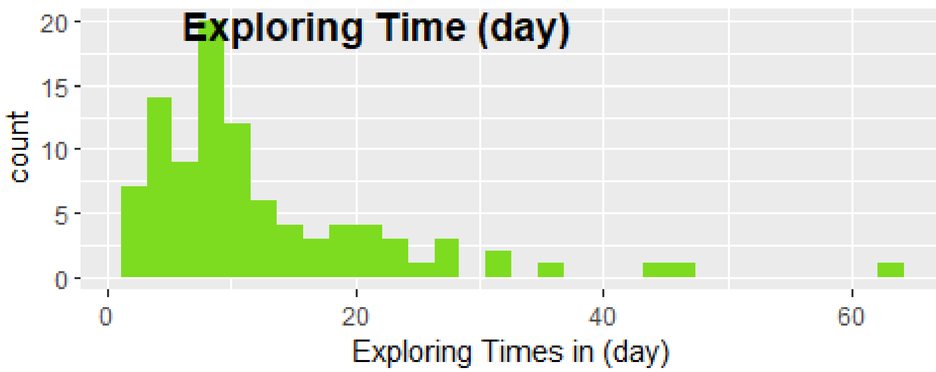
Let's viusalize the histograms and try to see the shape of Binding Time and Exploring Time

```
p1 <- ggplot(water.binding, aes(y = binding_time_day))
# Violin plot
p1 <- p1 + geom_boxplot(fill=I("lightblue"))


p2 <- ggplot(water.binding, aes(y = explore_time_day))
# Violin plot
p2 <- p2 + geom_boxplot(fill=I("lightblue"))

ggarrange(p1, p2, ncol =1, nrow = 2, labels = c( "Binding Time
(day)","Exploring Time (day)"))
```




```
p1 <- ggplot(water.binding, aes(x = binding_time_day)) +
  xlab("Binding Times in (day)")
# Violin plot
p1 <- p1 + geom_histogram(fill="#7ddc1f")


p2 <- ggplot(water.binding, aes(x = explore_time_day)) +
  xlab("Exploring Times in (day)")
# Violin plot
p2 <- p2 + geom_histogram(fill="#7ddc1f")

ggarrange(p1, p2, ncol =1, nrow = 2, labels = c( "Binding Time
(day)","Exploring Time (day)"))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

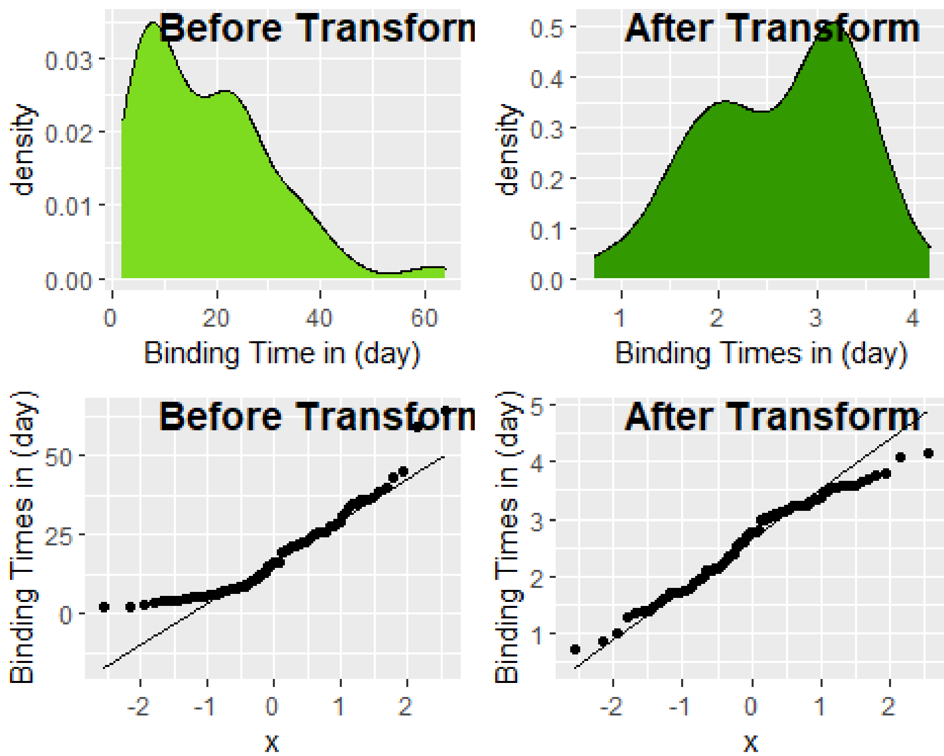




It seems that both *Binding Time* and *Exploring Time* are right-skewed. Let's try to transform them with log transform then visualize them again

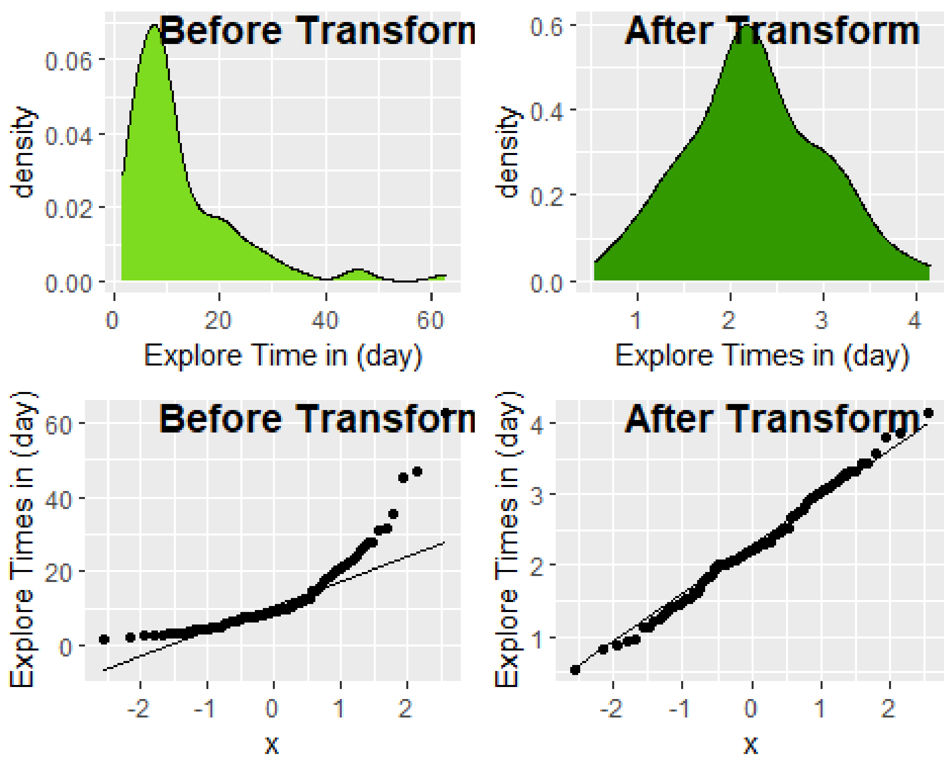
```
water.binding.log <- transform(water.binding,
            binding_time_day = log(binding_time_day),
            explore_time_day = log(explore_time_day)

            )

p1 <- ggplot(water.binding, aes(x = binding_time_day)) +
  xlab("Binding Time in (day)")
# Violin plot
p1 <- p1 + geom_density(fill="#7ddc1f")


p2 <- ggplot(water.binding.log, aes(x = binding_time_day)) +
  xlab("Binding Times in (day)")
# Violin plot
p2 <- p2 + geom_density(fill="#339900")

p3 <- ggplot(water.binding, aes(sample = binding_time_day)) + ylab("Binding
Times in (day)") +
  stat_qq() +
  stat_qq_line()

p4 <- ggplot(water.binding.log, aes(sample = binding_time_day)) +
```

```
ylab("Binding Times in (day)") +
  stat_qq() +
  stat_qq_line()

ggarrange(p1, p2, p3, p4, ncol =2, nrow = 2, labels = c( "Before
Transform","After Transform",
                                                "Before Transform","After
Transform"))
```




```
shapiro.test(water.binding.log$binding_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  water.binding.log$binding_time_day
## W = 0.96836, p-value = 0.02017

p1 <- ggplot(water.binding, aes(x = explore_time_day)) +
  xlab("Explore Time in (day)")
# Violin plot
p1 <- p1 + geom_density(fill="#7ddc1f")


p2 <- ggplot(water.binding.log, aes(x = explore_time_day)) +
  xlab("Explore Times in (day)")
# Violin plot
p2 <- p2 + geom_density(fill="#339900")
```

```
p3 <- ggplot(water.binding, aes(sample = explore_time_day)) + ylab("Explore
Times in (day)") +
  stat_qq() +
  stat_qq_line()

p4 <- ggplot(water.binding.log, aes(sample = explore_time_day)) +
ylab("Explore Times in (day)") +
  stat_qq() +
  stat_qq_line()

ggarrange(p1, p2, p3, p4, ncol =2, nrow = 2, labels = c( "Before
Transform","After Transform",
                                                 "Before Transform","After
Transform"))
```

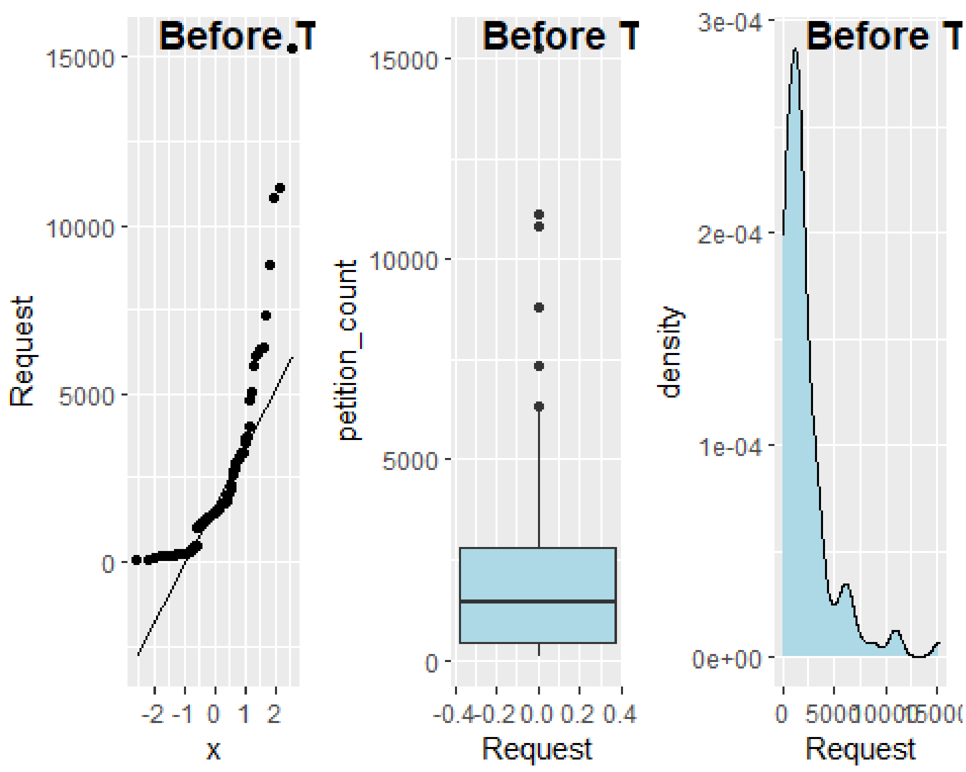



```
shapiro.test(water.binding.log$explore_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  water.binding.log$explore_time_day
## W = 0.99244, p-value = 0.8686
```

It seems that after transformation *Explore Time* value became more normally distributed and *Binding Time* seems still skewed but we'll assume that both data are normally distributed.

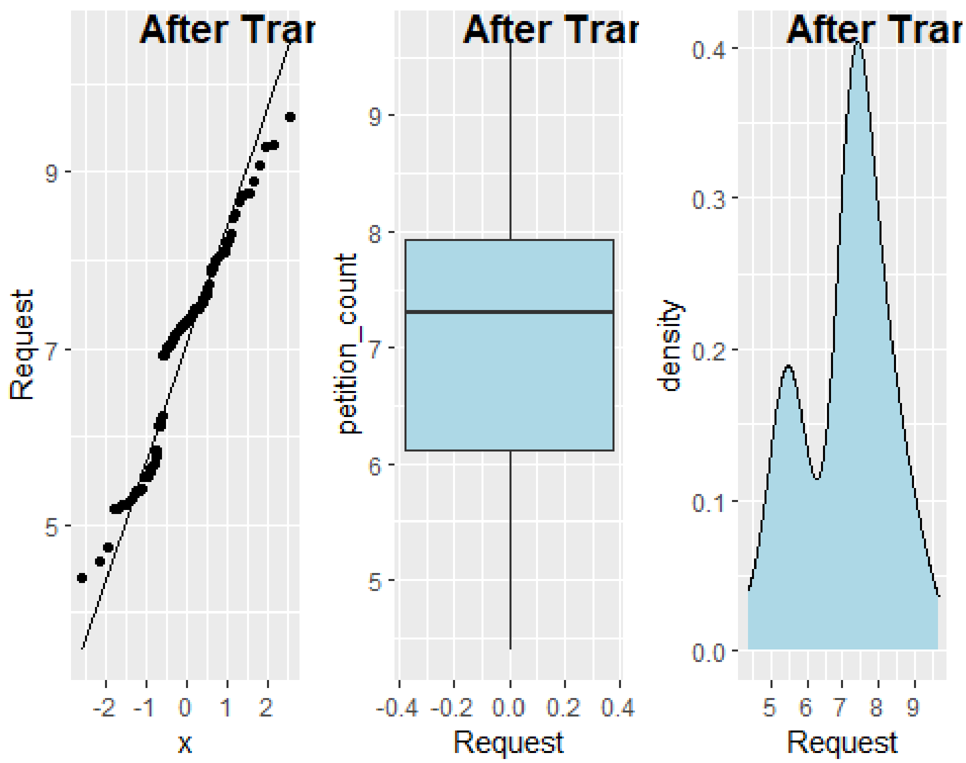
- Let's take a look at subscriber petition in last three years

```
p1 <- ggplot(water.binding, aes(sample = petition_count)) + ylab("Request") +
  stat_qq() +
  stat_qq_line()

p2 <- ggplot(water.binding, aes(y = petition_count)) +
  xlab("Request")
# Violin plot
p2 <- p2 + geom_boxplot(fill=I("lightblue"))

p3 <- ggplot(water.binding, aes(x = petition_count)) +
  xlab("Request")
# Violin plot
p3 <- p3 + geom_density(fill=I("lightblue"))

ggarrange(p1, p2, p3, ncol =3, nrow = 1, labels = c( "Before
Transform","Before Transform",
                                                    "Before Transform"))
```




```
petition.log <- transform(water.binding,
            petition_count = log(petition_count))


p1 <- ggplot(petition.log, aes(sample = petition_count)) + ylab("Request") +
  stat_qq() +
  stat_qq_line()
```

```r
p2 <- ggplot(petition.log, aes(y = petition_count)) +
  xlab("Request")
# Violin plot
p2 <- p2 + geom_boxplot(fill=I("lightblue"))

p3 <- ggplot(petition.log, aes(x = petition_count)) +
  xlab("Request")
# Violin plot
p3 <- p3 + geom_density(fill=I("lightblue"))

ggarrange(p1, p2, p3, ncol =3, nrow = 1, labels = c( "After Transform","After
Transform",
                                                  "After Transform"))
```




```r
shapiro.test(petition.log$petition_count)

##
##  Shapiro-Wilk normality test
##
## data:  petition.log$petition_count
## W = 0.95203, p-value = 0.001482
```

Before the transformation our data was right skewed and we try to take log transformation and make it normally distributed but it seems that doesn't work. After that point we'll assume that our data is normally distributed.

- Let’s compare the confidence interval of subscriber petition in 2019 and 2020

```
petition.log.2019 <- subset(petition.log, subset = year == "2019")
petition.log.2020 <- subset(petition.log, subset = year == "2020")



petition.log.2019 <- subset(petition.log, subset = year == "2019")


#Confidence Interval of petition of 2019
petition.log.2019.CI <- t.test(petition.log.2019$petition_count)$conf.int
petition.log.2019.CI

## [1] 6.647058 7.530969
## attr(,"conf.level")
## [1] 0.95

#Confidence Interval of petition of 2020
petition.log.2020.CI <- t.test(petition.log.2020$petition_count)$conf.int
petition.log.2020.CI

## [1] 6.548838 7.355479
## attr(,"conf.level")
## [1] 0.95
```

It seems that there is significant difference between petitions in 2019 and 2020. Lets test it and see it better.

- We’ll assume that we don’t have an information about sample standard deviation then we’ll use t-test.

- Mu petition2019 = Mu petition2020 and H1: Mu petition2019 != Mu petition2020

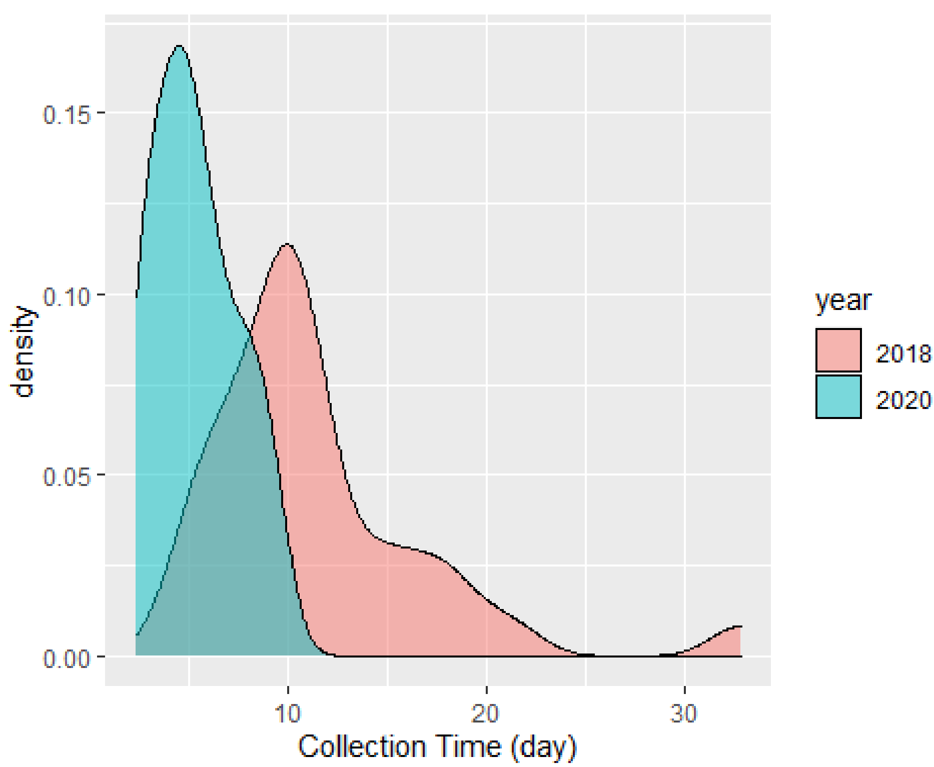
```
with(petition.log, t.test(x=petition_count[year=="2019"],
                          y=petition_count[year=="2020"]))

##
##  Welch Two Sample t-test
##
## data:  petition_count[year == "2019"] and petition_count[year == "2020"]
## t = 0.4665, df = 61.488, p-value = 0.6425
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4496731  0.7233825
## sample estimates:
## mean of x mean of y
##  7.089013  6.952159
```

According to the Welch Two Sample t-test, our p-value is 0.6425, and also the 95 percent confidence interval contains 0. That means we can not reject that the true

difference in means is equal to 0. In other words, we can not say that the petition counts in 2020 and 2019 are significantly different from each other.

- **Is The tax collection time in 2018 takes more time than 2020? Let's test it.**
- First visualize the data and see the shape of it.

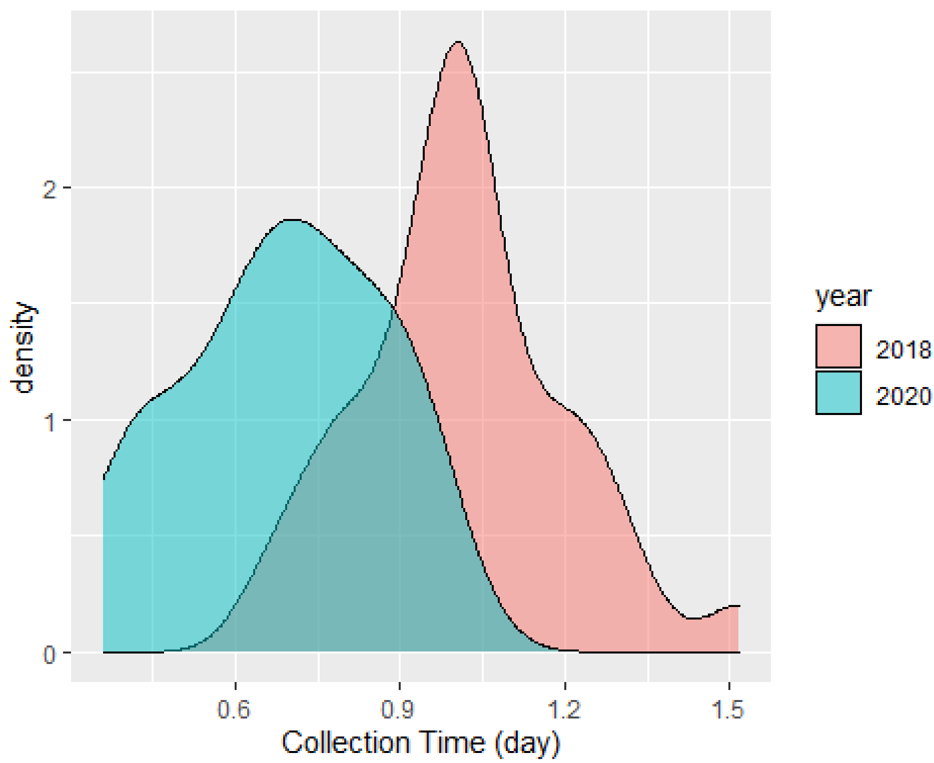
```
collection2020.2018 = subset(water.binding, subset = year != "2019")



#Density functions
base.plot <- ggplot(collection2020.2018, aes(x = collection_time_day)) +
  xlab("Collection Time (day)")
p <- base.plot + geom_density(aes(fill = year), alpha = 0.5)

p
```



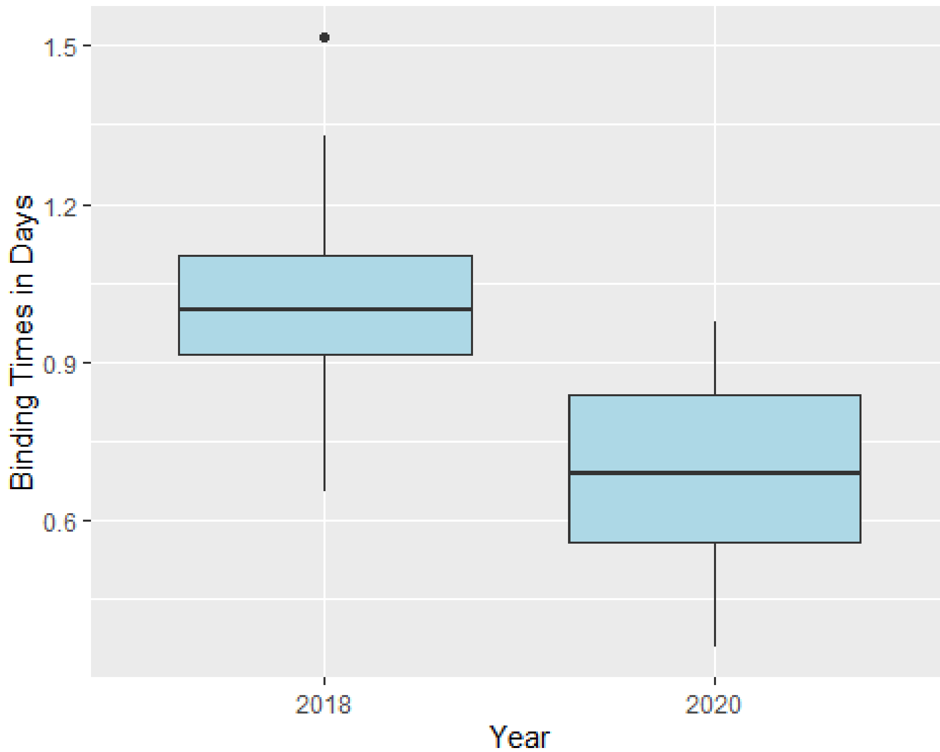

```
shapiro.test(collection2020.2018$collection_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  collection2020.2018$collection_time_day
## W = 0.82997, p-value = 4.253e-07
```

It seems that our both data right skewed let's try log transformation and make them normally distributed.

```
collection2020.2018.log <- transform(collection2020.2018,
            collection_time_day = log10(collection_time_day))


#Density functions
base.plot <- ggplot(collection2020.2018.log, aes(x = collection_time_day)) +
  xlab("Collection Time (day)")
p <- base.plot + geom_density(aes(fill = year), alpha = 0.5)

p
```




```
qplot(x =year, y = collection_time_day,
      geom = "boxplot", data = collection2020.2018.log,
      xlab = "Year",
      ylab = "Binding Times in Days",
      fill = I("lightblue"))
```




```
shapiro.test(collection2020.2018.log$collection_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  collection2020.2018.log$collection_time_day
## W = 0.98817, p-value = 0.8006
```

Now it looks better and also box plot suggest that tax collection time in 2020 takes less time than 2018. Now separate the data by year and apply the Shapiro Normality test.

```
collection2020.log = subset(collection2020.2018.log, subset = year == "2020")
collection2018.log = subset(collection2020.2018.log, subset = year == "2018")

shapiro.test(collection2020.log$collection_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  collection2020.log$collection_time_day
## W = 0.95863, p-value = 0.2519

shapiro.test(collection2018.log$collection_time_day)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  collection2018.log$collection_time_day
## W = 0.97759, p-value = 0.7272
```

Shapiro test also shows that our two sets are normally distributed We can have the two-sample t-test now.

- We'll assume that we don't have an information about sample standard deviation and assume that there is no significant difference between variances, and we'll use t-test.

- **H0:** Mu collectiontime2018 <= Mu collectiontime2020 and **H1:** Mu collectiontime2018 > Mu collectiontime2020

```
collection.time.2020.2018.t.test <-
t.test(collection2018.log$collection_time_day,
collection2020.log$collection_time_day, var.equal = TRUE, alternative =
"greater")
collection.time.2020.2018.t.test

##
##  Two Sample t-test
##
## data:  collection2018.log$collection_time_day and
collection2020.log$collection_time_day
## t = 7.0413, df = 62, p-value = 9.118e-10
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2479398       Inf
## sample estimates:
## mean of x mean of y
## 1.0144094 0.6893947

collection.time.2020.2018.diff <-
round((collection.time.2020.2018.t.test$estimate[1] -
collection.time.2020.2018.t.test$estimate[2]), 1)

# Confidence level as a %
conf.level <- attr(collection.time.2020.2018.t.test$conf.int, "conf.level") *
100

mean.of.transformed.data <- mean(collection2020.2018.log$collection_time_day)
mean.of.transformed.data

## [1] 0.851902
```

Results: Our study finds that tax collection time are on average in 2018 takes 8.6053745hours more than 2020 (t-statistic 7.04, p=0, 95% CI [0.2,]

The p-value is almost zero and less than alpha(0.05). Therefore we can reject the H0 which refers collectiontime2018 <= Mu collectiontime2020. In other words we reject

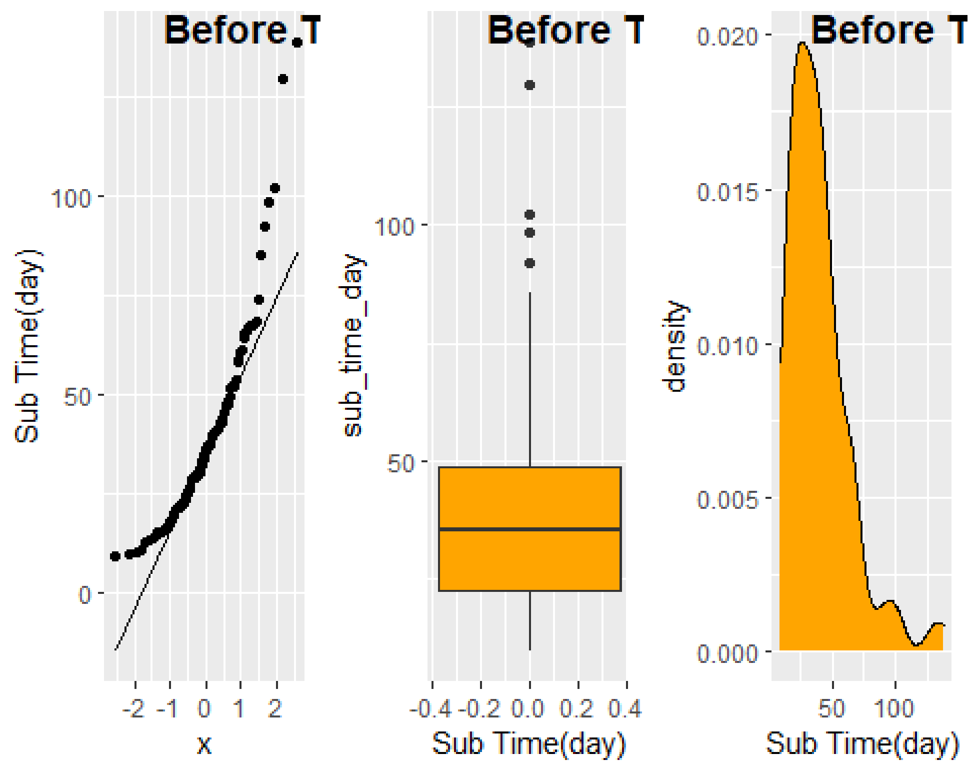
that tax collection time in 2018 takes less time than 2020. We statistically show that tax collection time is shorter in 2020.

- **Does subscriber's subscription time longer than 30 days in average?**
- First summarize the subscription time of subscribers.

```
summary(water.binding$sub_time_day)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.58   22.39   35.19   39.75   48.80  138.75
```

It seems that mean is greater than 30 days. Now let's look at the shape of data.

```
p1 <- ggplot(water.binding, aes(sample = sub_time_day)) + ylab("Sub
Time(day)") +
  stat_qq() +
  stat_qq_line()

p2 <- ggplot(water.binding, aes(y = sub_time_day)) +
  xlab("Sub Time(day)")
p2 <- p2 + geom_boxplot(fill=I("orange"))

p3 <- ggplot(water.binding, aes(x = sub_time_day)) +
  xlab("Sub Time(day)")

p3 <- p3 + geom_density(fill=I("orange"))

ggarrange(p1, p2, p3, ncol =3, nrow = 1, labels = c( "Before
Transform","Before Transform",
                                                    "Before Transform"))
```

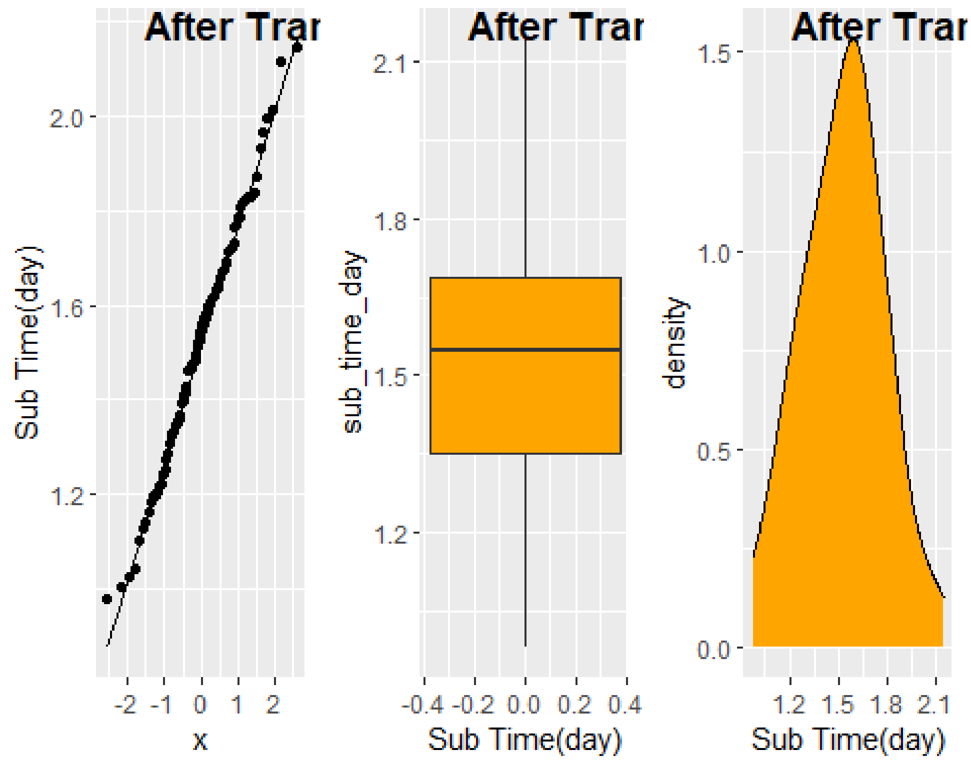



```
shapiro.test(water.binding$sub_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  water.binding$sub_time_day
## W = 0.86577, p-value = 7.658e-08
```

As we can see from plots and also in Shapiro we reject that our data is normally distributed. Let’s try log transformation.

```
sub.time.log <- transform(water.binding,
            sub_time_day = log10(sub_time_day))

p1 <- ggplot(sub.time.log, aes(sample = sub_time_day)) + ylab("Sub
Time(day)") +
  stat_qq() +
  stat_qq_line()

p2 <- ggplot(sub.time.log, aes(y = sub_time_day)) +
  xlab("Sub Time(day)")
p2 <- p2 + geom_boxplot(fill=I("orange"))

p3 <- ggplot(sub.time.log, aes(x = sub_time_day)) +
  xlab("Sub Time(day)")

p3 <- p3 + geom_density(fill=I("orange"))
```

```
ggarrange(p1, p2, p3, ncol =3, nrow = 1, labels = c( "After Transform","After
Transform",
                                                     "After Transform"))
```




```
shapiro.test(sub.time.log$sub_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  sub.time.log$sub_time_day
## W = 0.99188, p-value = 0.8319
```

It seems that after transformed the data it perfectly normally distributed.

- We'll assume that we don't have an information about sample standard deviation and assume that there is no significant difference between variances, and we'll use t-test.

-**H0:** Mu sub_time_day <= 30 and **H1:** Mu sub_time_day > 30

```
sub.time.t.test<- t.test(sub.time.log$sub_time_day, alternative =
c("greater"), mu = log10(30), conf.level = 0.95)
sub.time.t.test

##
##  One Sample t-test
##
```

```
## data:  sub.time.log$sub_time_day
## t = 1.991, df = 95, p-value = 0.02468
## alternative hypothesis: true mean is greater than 1.477121
## 95 percent confidence interval:
##  1.485577      Inf
## sample estimates:
## mean of x
##   1.52815
```

```
# Confidence level as a %
conf.level <- attr(collection.time.2020.2018.t.test$conf.int, "conf.level") *
100

mean.of.transformed.data <- mean(sub.time.log$sub_time_day)
round(sub.time.t.test$p.value,3)
```

```
## [1] 0.025
```

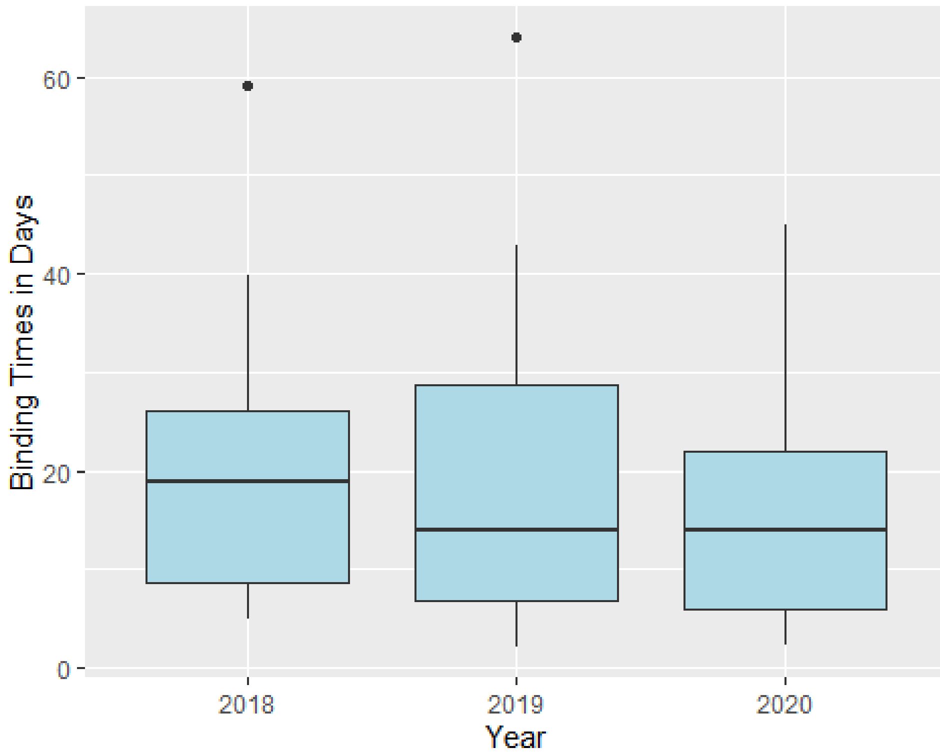
Results: Our study finds that (t-statistic 1.99, p=0.025, 95% CI [1.5,], mean of data is found 34 days.

Conclusion: The p-value is 0.024 and it's less than alpha(0.05). Therefore we reject the H0 which refers Mu sub_time_day <= 30 days. In other words we reject that subscriber's subscription time less than 30 days on average We statistically show that subscriber's subscription time is longer than 30 days.

ANOVA TEST

We were visuilazed and talk about the water binding times of 2020, 2019 and 2018. Let's visualize the boxplots again and test them if the average binding times of these three years are equal or not.

```
#boxplot(birthwt$birthwt.grams~birthwt$hypertension)


qplot(x =year, y = binding_time_day,
      geom = "boxplot", data = water.binding,
      xlab = "Year",
      ylab = "Binding Times in Days",
      fill = I("lightblue"))
```



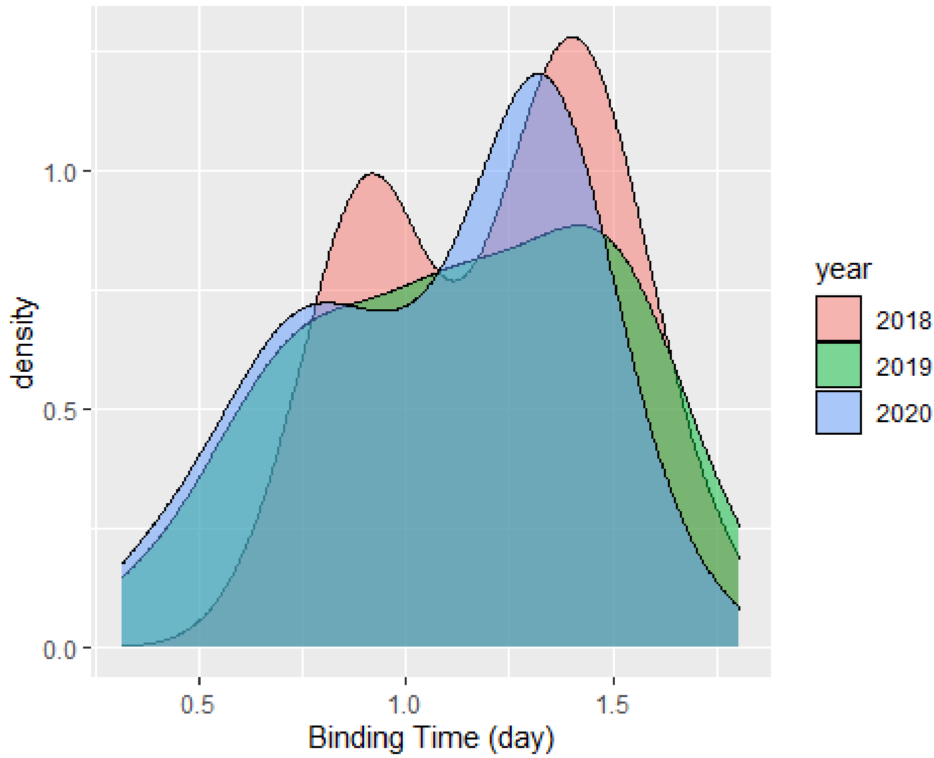
Let's summarize the water binding times(day) for years and check for Shapiro Normality Test.

```
group_by(water.binding, year) %>%
  summarise(
    count = n(),
    mean = mean(binding_time_day, na.rm = TRUE),
    sd = sd(binding_time_day, na.rm = TRUE)
  )

## # A tibble: 3 x 4
##   year  count  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 2018     32  20.2  12.7
## 2 2019     32  18.7  14.5
## 3 2020     32  15.5  10.2

shapiro.test(water.binding$binding_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  water.binding$binding_time_day
## W = 0.90884, p-value = 5.606e-06
```

It seems that there is some differences between the average binding times of years. Also the p-value of Shapiro Test less than alpha(0.05) then we reject that the data normally distributed.

Let's take the log of data and try to transform it and check the normality for each year.

```
water.binding.log <- transform(water.binding,
            binding_time_day = log10(binding_time_day))


water.binding.2020= subset(water.binding.log, subset = year == "2020")
water.binding.2019 = subset(water.binding.log, subset = year == "2019")
water.binding.2018 = subset(water.binding.log, subset = year == "2018")

shapiro.test(water.binding.2020$binding_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  water.binding.2020$binding_time_day
## W = 0.93601, p-value = 0.05773

shapiro.test(water.binding.2019$binding_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  water.binding.2019$binding_time_day
## W = 0.96603, p-value = 0.3977

shapiro.test(water.binding.2018$binding_time_day)

##
##  Shapiro-Wilk normality test
##
## data:  water.binding.2018$binding_time_day
## W = 0.94246, p-value = 0.08794

#Density functions
p <- ggplot(water.binding.log, aes(x = binding_time_day)) +
  xlab("Binding Time (day)")
p <- p + geom_density(aes(fill = year), alpha = 0.5)

p
```



After transformation each year looks like normally distributed. We can conclude them from the Shapiro test. All the p-values are greater than alpha(0.05) then we can not reject that data normally distributed. Now we can compute the One-way ANOVA test.

H0: Mu bind_time_2020 = Mu bind_time_2019 = Mu bind_time_2018, **H1:** At least one differs.

```
# Compute the analysis of variance
res.aov <- aov(binding_time_day ~ year, data = water.binding.log)
summary(res.aov)

##             Df Sum Sq Mean Sq F value Pr(>F)
## year         2  0.311  0.1557   1.375  0.258
## Residuals   93 10.534  0.1133
```

Conclusion: From the above table, we can see the p-value of the ANOVA test is 0.258 which greater than alpha(0.05). Then, we can not reject H0 which refers to three years of average binding time(day) are equal. In other words, there is no strong evidence at least one differs.

WATER PRODUCTION

Dataset provided by the Izmir Municipality (IBB Acik Veri Portali https://acikveri.bizizmir.com/dataset). The scope of the data is water production of the Water Sources in Izmir between 2009 and 2021. We have limited information about 2021 because of that this project is running in 2021 June. Dataset provides us that the total water production of 13 different Water Sources in meter cube for each month. Dateset can be provided by the following link. (https://acikveri.bizizmir.com/dataset/su-uretiminin-aylara-ve-kaynaklara-gore-dagilimi/resource/e3f93f98-38a3-41d9-b89e-a7d1d378475b)

- Let's see the data again

```
head(water.production,10)

##    year month                      source amount_m3
## 1  2021     2              Tahtalı Barajı   5818100
## 2  2021     2              Balçova Barajı         0
## 3  2021     2            Sarıkız Kuyuları         0
## 4  2021     2 Menemen - Çavuşköy Kuyuları    886348
## 5  2021     2         Halkapınar Kuyuları   2492021
## 6  2021     2          Pınarbaşı Kuyuları     58721
## 7  2021     2      Buca ve Sarnıç Kuyuları     68651
## 8  2021     2               Gördes Barajı   3292558
## 9  2021     2              Göksu Kuyuları   3877762
## 10 2021     2              Tahtalı Barajı   5818100

sapply(water.production, summary)

## $year
## 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##  120  120  132  120  132  144  156  156  156  156  141  108   27
##
## $month
##   1  10  11  12   2   3   4   5   6   7   8   9
## 147 136 134 134 154 136 136 136 136 136 147 136
##
## $source
##        Alaçatı Kutlu Aktaş Barajı                    Balçova Barajı
##                                71                               148
##           Buca ve Sarnıç Kuyuları                    Göksu Kuyuları
##                               148                               148
##                     Gördes Barajı                 Güzelhisar Barajı
##                               124                               109
##               Halkapınar Kuyuları       Menemen - Çavuşköy Kuyuları
##                               148                               148
## Ödemiş İçme Suyu Arıtma Tesisleri                Pınarbaşı Kuyuları
##                                49                               148
##                  Sarıkız Kuyuları                    Tahtalı Barajı
##                               148                               148
```

```
##                      Ürkmez Barajı
##                                131
##
## $amount_m3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   98109  467908 1568540 2571146 9786100
```

Note: There are no sufficiant information about the water production of 2021. Therefore we'll dismiss it.

```
water.production = subset(water.production, subset = water.production$year !=
"2021")
sapply(water.production, summary)
```

```
## $year
## 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##  120  120  132  120  132  144  156  156  156  156  141  108    0
##
## $month
##   1  10  11  12   2   3   4   5   6   7   8   9
## 138 136 134 134 136 136 136 136 136 136 147 136
##
## $source
##         Alaçatı Kutlu Aktaş Barajı                      Balçova Barajı
##                                 71                                 145
##            Buca ve Sarnıç Kuyuları                     Göksu Kuyuları
##                                145                                 145
##                      Gördes Barajı                   Güzelhisar Barajı
##                                121                                 109
##                 Halkapınar Kuyuları        Menemen - Çavuşköy Kuyuları
##                                145                                 145
## Ödemiş İçme Suyu Arıtma Tesisleri                 Pınarbaşı Kuyuları
##                                 49                                 145
##                   Sarıkız Kuyuları                      Tahtalı Barajı
##                                145                                 145
##                      Ürkmez Barajı
##                                131
##
## $amount_m3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   99479  466340 1563175 2561120 9786100
```

- Take a look at the most productive Sources

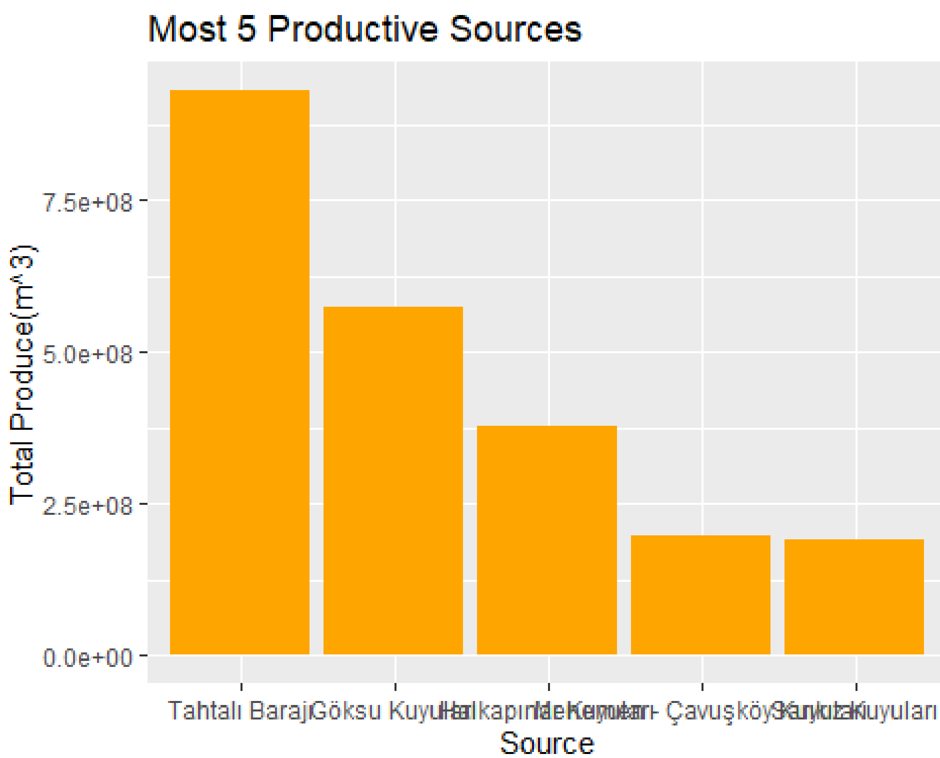
```
#We can see the most consumer districts in 2020 in Izmir.
df1 <- water.production %>%
  group_by(source) %>%
  summarize(total_amount = sum(amount_m3))

df1 <- df1 %>%
  arrange(desc(total_amount))
```

```r
df1 %>% tbl_df %>% print(n=40)
```

```
## # A tibble: 13 x 2
##    source                            total_amount
##    <fct>                                    <dbl>
##  1 Tahtalı Barajı                       930452445
##  2 Göksu Kuyuları                       575342231
##  3 Halkapınar Kuyuları                  378593273
##  4 Menemen - Çavuşköy Kuyuları          195378337
##  5 Sarıkız Kuyuları                     188932250
##  6 Gördes Barajı                        137755577
##  7 Balçova Barajı                        63903955
##  8 Alaçatı Kutlu Aktaş Barajı            30424481
##  9 Pınarbaşı Kuyuları                    17687851
## 10 Ürkmez Barajı                         14939942
## 11 Güzelhisar Barajı                     14627134
## 12 Ödemiş İçme Suyu Arıtma Tesisleri      9434313
## 13 Buca ve Sarnıç Kuyuları                7697653
```

- Top 5 productive sources in Izmir

```r
df1 <- df1 %>% slice_max(total_amount, n = 5)
p<-ggplot(data=df1, aes(x=reorder(source, -total_amount), y=total_amount)) +
  geom_bar(stat="identity", fill="orange")

p <- p + labs(title = "Most 5 Productive Sources", x='Source', y='Total
Produce(m^3)')
p
```



It seems that the most productive source is Tahtali Baraji.

Let's take a look how the water production changes year by yea in Tahtali Baraji.

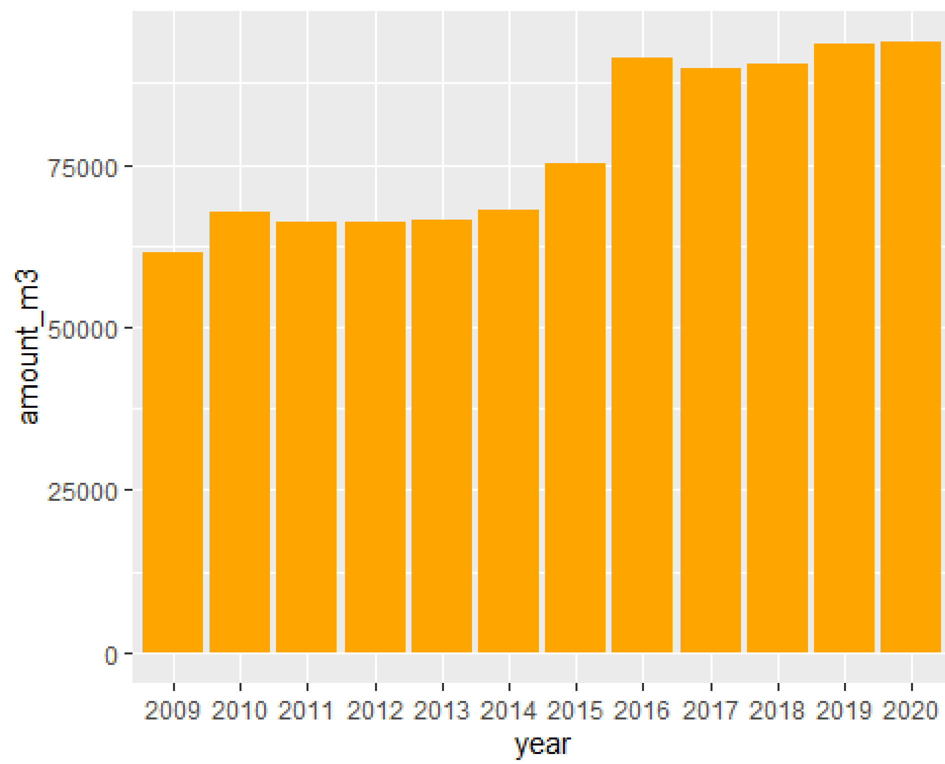
-First take the Tahtali Baraji's data.

Note: amount_m3 divided by 1000 to see the numbers better.

```
water.production.tahtali = subset(water.production, subset =
water.production$source == "Tahtalı Barajı")
water.production.tahtali$amount_m3 = water.production.tahtali$amount_m3 /
10**3
sapply(water.production.tahtali, summary)

## $year
## 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##   12   12   12   12   12   12   12   12   12   12   13   12    0
##
## $month
##  1 10 11 12  2  3  4  5  6  7  8  9
## 12 12 12 12 12 12 12 12 12 12 13 12
##
## $source
##         Alaçatı Kutlu Aktaş Barajı                     Balçova Barajı
##                                  0                                  0
##            Buca ve Sarnıç Kuyuları                     Göksu Kuyuları
##                                  0                                  0
##                      Gördes Barajı                  Güzelhisar Barajı
```

```
##                           0                           0
##           Halkapınar Kuyuları       Menemen - Çavuşköy Kuyuları
##                           0                           0
## Ödemiş İçme Suyu Arıtma Tesisleri            Pınarbaşı Kuyuları
##                           0                           0
##              Sarıkız Kuyuları                Tahtalı Barajı
##                           0                         145
##                Ürkmez Barajı
##                           0
##
## $amount_m3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3811    5178    6559    6417    7399    9786
```

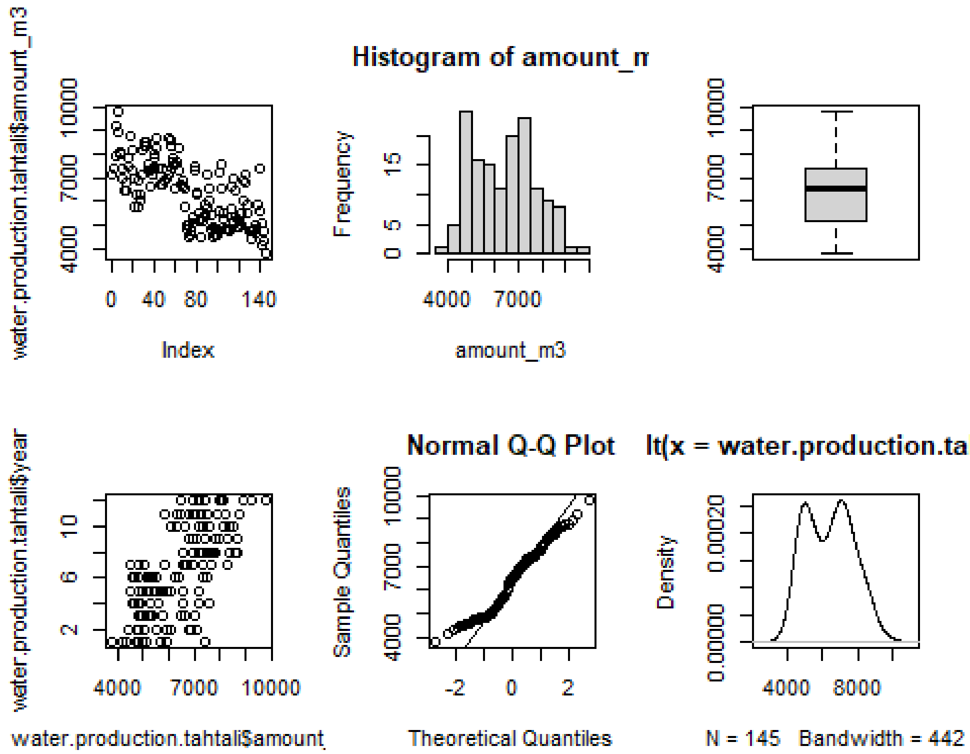
```
p<-ggplot(data=water.production.tahtali, aes(x=year, y=amount_m3)) +
  geom_bar(stat="identity", fill="orange")

p
```



Histograms of Water Production in Tahtali

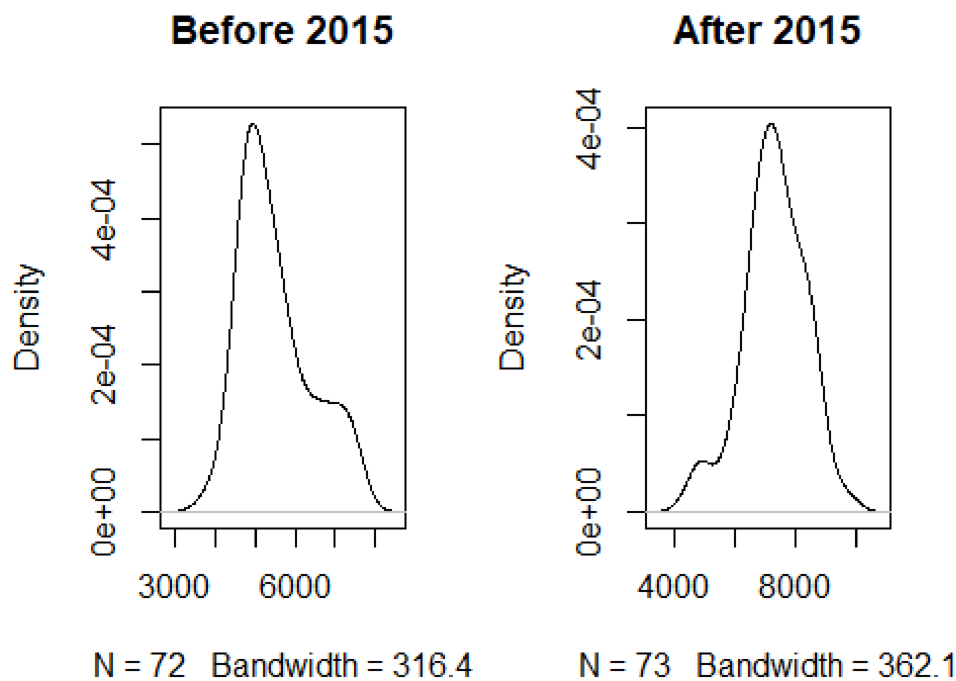
```
par(mfrow = c(2,3))
plot(water.production.tahtali$amount_m3)
with(water.production.tahtali, hist(amount_m3))
with(water.production.tahtali, boxplot(amount_m3))
plot(water.production.tahtali$amount_m3, water.production.tahtali$year)
qqnorm(water.production.tahtali$amount_m3)
```

```
qqline(water.production.tahtali$amount_m3)
plot(density(water.production.tahtali$amount_m3))
```

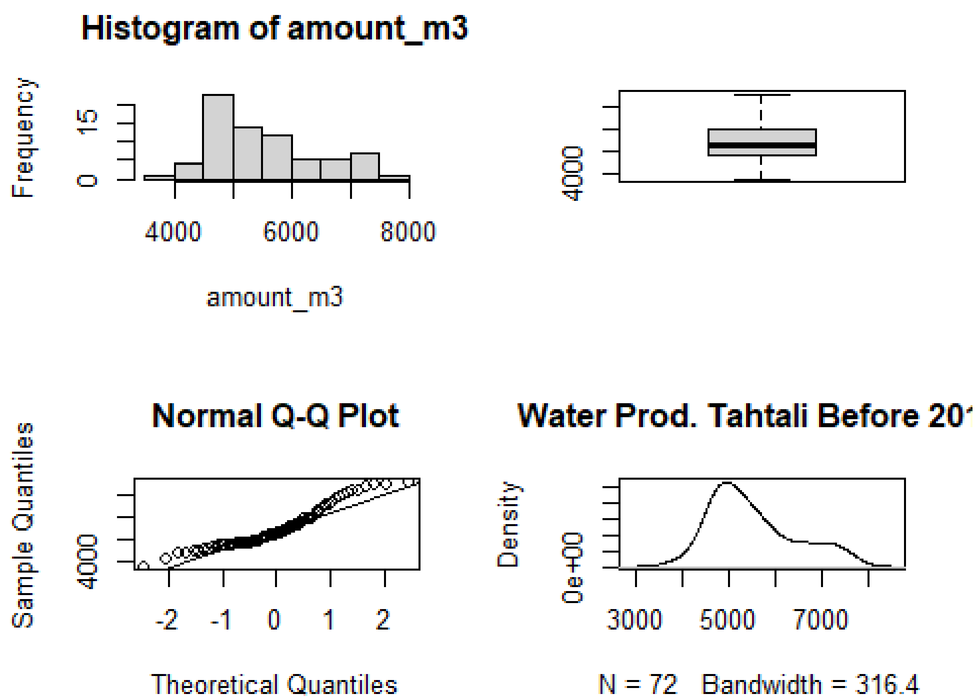


Note: It seems that we have bimodal distribution. It means that there should be different classes. Therefore we'll divide our data into two groups: *Water Consumpiton Before 2015* and *Water Consumption After 2015* and then we'll plot again.

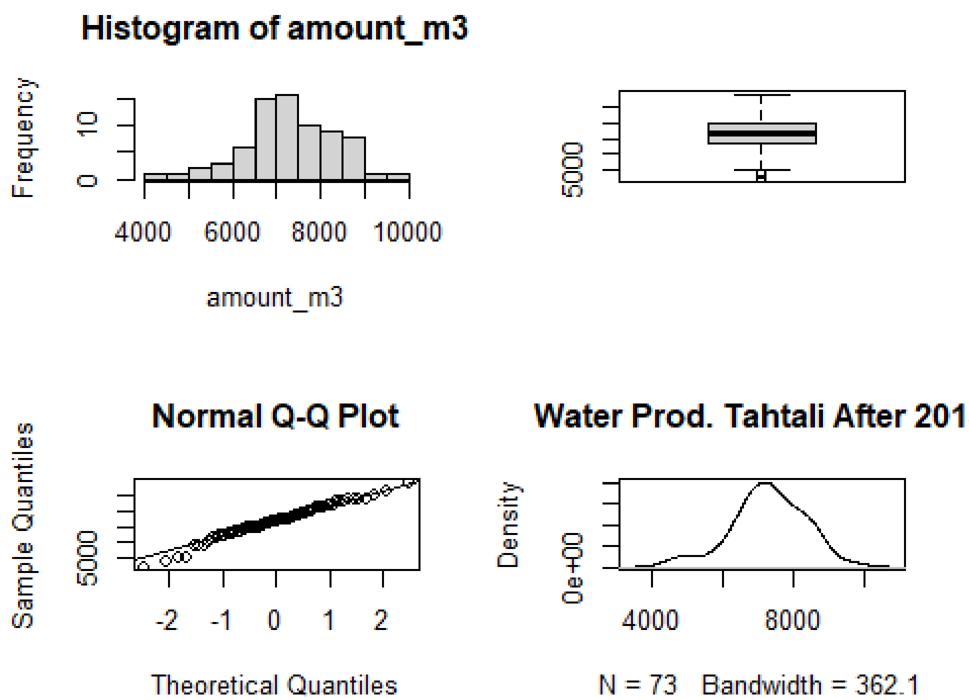
```
water.production.tahtali$year =
as.numeric(as.character(water.production.tahtali$year))
water.production.tahtali$seasons =
cut(water.production.tahtali$year,c(2008,2014,2020))
levels(water.production.tahtali$seasons) = c("before_2015","after_2015")

#Divide data into two different classes. before.2015 and after.2015
tahtali.before.2015 =subset(water.production.tahtali, subset =
water.production.tahtali$seasons == "before_2015")
tahtali.after.2015 =subset(water.production.tahtali, subset =
water.production.tahtali$seasons == "after_2015")


par(mfrow= c(1,2))
plot(density(tahtali.before.2015$amount_m3), main = "Before 2015")
plot(density(tahtali.after.2015$amount_m3), main="After 2015")
```




```
par(mfrow = c(2,2))
with(tahtali.before.2015, hist(amount_m3),main = "Water Prod. Tahtali Before 2015")
with(tahtali.before.2015, boxplot(amount_m3),main = "Water Prod. Tahtali Before 2015")
qqnorm(tahtali.before.2015$amount_m3)
qqline(tahtali.before.2015$amount_m3)
plot(density(tahtali.before.2015$amount_m3),main = "Water Prod. Tahtali Before 2015")
```




```
shapiro.test(tahtali.before.2015$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  tahtali.before.2015$amount_m3
## W = 0.92553, p-value = 0.0003814

par(mfrow = c(2,2)) # Display plots in a single 2 x 2 figure
with(tahtali.after.2015, hist(amount_m3),main = "Water Prod. Tahtali After
2015")
with(tahtali.after.2015, boxplot(amount_m3),main = "Water Prod. Tahtali After
2015")
qqnorm(tahtali.after.2015$amount_m3)
qqline(tahtali.after.2015$amount_m3)
plot(density(tahtali.after.2015$amount_m3),main = "Water Prod. Tahtali After
2015")
```




```
shapiro.test(tahtali.after.2015$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  tahtali.after.2015$amount_m3
## W = 0.98236, p-value = 0.3997
```

Note: After that point *Water Production in Tahtali After 2015* seems normally distributed by the above plots and also from the Shapiro Test. Our p value > 0.05 and that's mean we can not reject that data are not significantly different from normal distribution, but *Water Production in Tahtali Before 2015* doesn't look like normally distributed and we'll assume that both of them are normally distributed.

Let's compare the Confidence Intervals.

- Our confidince coefficient is 0.95 it's mean that our sample means are appear between Confidince Interval by 95%.

- Confidence Interval of Water Production Before 2015 in Tahtali Baraji

```
before.2015.CI <- t.test(tahtali.before.2015$amount_m3)$conf.int
before.2015.CI

## [1] 5291.162 5717.806
## attr(,"conf.level")
## [1] 0.95
```

For the Water Production in Tahtali Baraji before 2015, 95% of the sample means between the 5291.162 and 5717.806. If we transformed the data again that's mean %95 percentage of the sample means between 5291000.162 m^3 and 5717000.806 m^3 water.

- Confidence Interval of Water Production After 2015 in Tahtali Baraji.

```
after.2015.CI <- t.test(tahtali.after.2015$amount_m3)$conf.int
after.2015.CI

## [1] 7075.987 7557.701
## attr(,"conf.level")
## [1] 0.95
```

For the Water Production in Tahtali Baraji after 2015, 95% of the sample means between the 7075.987 7557.701. If we transformed the data again that's mean %95 percentage of the sample means between 7075000.987 m^3 and 7557000.701 m^3 water.

Note: Confidence intervals suggest that the Water Production between *Before 2015* and *After 2015* are significantly different from each others.

Before going for a t-test let's check if the variance of two groups are equal with F-test.

```
res.ftest <- var.test(amount_m3 ~ seasons, data = water.production.tahtali)
res.ftest

## 
##  F test to compare two variances
## 
## data:  amount_m3 by seasons
## F = 0.77331, num df = 71, denom df = 72, p-value = 0.2798
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4848107 1.2344438
## sample estimates:
## ratio of variances 
##          0.7733108

paste("p-value of F-test greater than alpha(0.05) that means there is no significant difference between the variances of the two sets. Therefore we can continue with the t-test")

## [1] "p-value of F-test greater than alpha(0.05) that means there is no significant difference between the variances of the two sets. Therefore we can continue with the t-test"
```

Now let's test our hypothesis using Welch Two Sample t-test

Ho: Mu before_2015 = Mu after_2015 and **H1:** Mu before_2015 != Mu after_2015

```
tahtali.t.test <- with(water.production.tahtali,
t.test(x=amount_m3[seasons=="after_2015"],
                     y=amount_m3[seasons=="before_2015"]))

before.after.2015.diff <- round((tahtali.t.test$estimate[1] -
tahtali.t.test$estimate[2]), 1)

# Confidence level as a %
conf.level <- attr(tahtali.t.test$conf.int, "conf.level") * 100
```

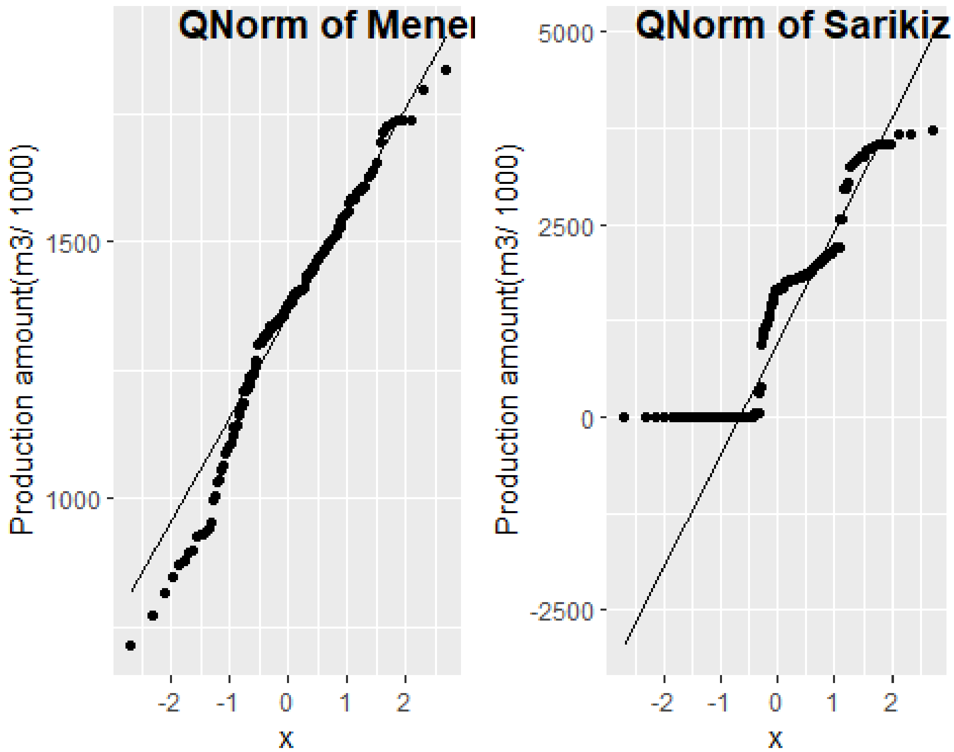
Results: Our study finds that water productions are on average 1812.4$m^3/10^3$ higher After 2015 compared to the Before 2015 (t-statistic 11.23, p=0, 95% CI [1493.3, 2131.4]m^3 / 10^3)

Our p-value is less than 0.05(alpha) and also our confidence interval doesn't include 0. That's mean Water Production after 2015 and before 2015 are significantly different from each others. And we can reject that Mu before_2015 is equal to Mu after_2015(H0). Also we can conclude that from the Confidence Intervals, Tahtali Baraji is much more productive after 2015.

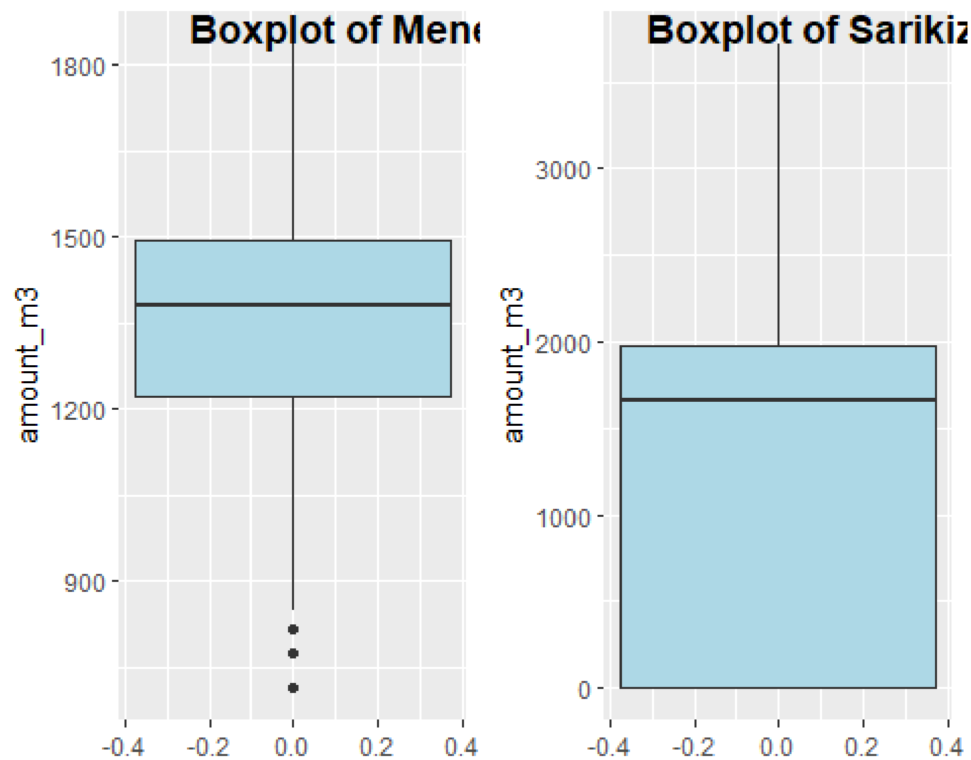
Menemen - Çavuşköy Kuyuları and Sarıkız Kuyuları seems to produce almost same. Let's discover if our hypothesis is statistically true or not.

- First visualize the data and check for the normality.

```
production.menemen = subset(water.production, subset =
water.production$source == "Menemen - Çavuşköy Kuyuları")
production.sarikiz = subset(water.production, subset =
water.production$source == "Sarıkız Kuyuları")

production.menemen$amount_m3 = production.menemen$amount_m3 / 10**3
production.sarikiz$amount_m3 = production.sarikiz$amount_m3 / 10**3

p1 <- ggplot(production.menemen, aes(sample = amount_m3)) + ylab("Production
amount(m3/ 1000)") +
  stat_qq() +
  stat_qq_line()

p2 <- ggplot(production.sarikiz, aes(sample = amount_m3)) + ylab("Production
amount(m3/ 1000)") +
  stat_qq() +
  stat_qq_line()



ggarrange(p1, p2, ncol =2, nrow = 1, labels = c( "QNorm of Menemen","QNorm of
Sarikiz"))
```




```
p1 <- ggplot(production.menemen, aes(y = amount_m3))
p1 <- p1 + geom_boxplot(fill=I("lightblue"))

p2 <- ggplot(production.sarikiz, aes(y = amount_m3))
p2 <- p2 + geom_boxplot(fill=I("lightblue"))


ggarrange(p1, p2, ncol =2, nrow = 1, labels = c( "Boxplot of
Menemen","Boxplot of Sarikiz"))
```

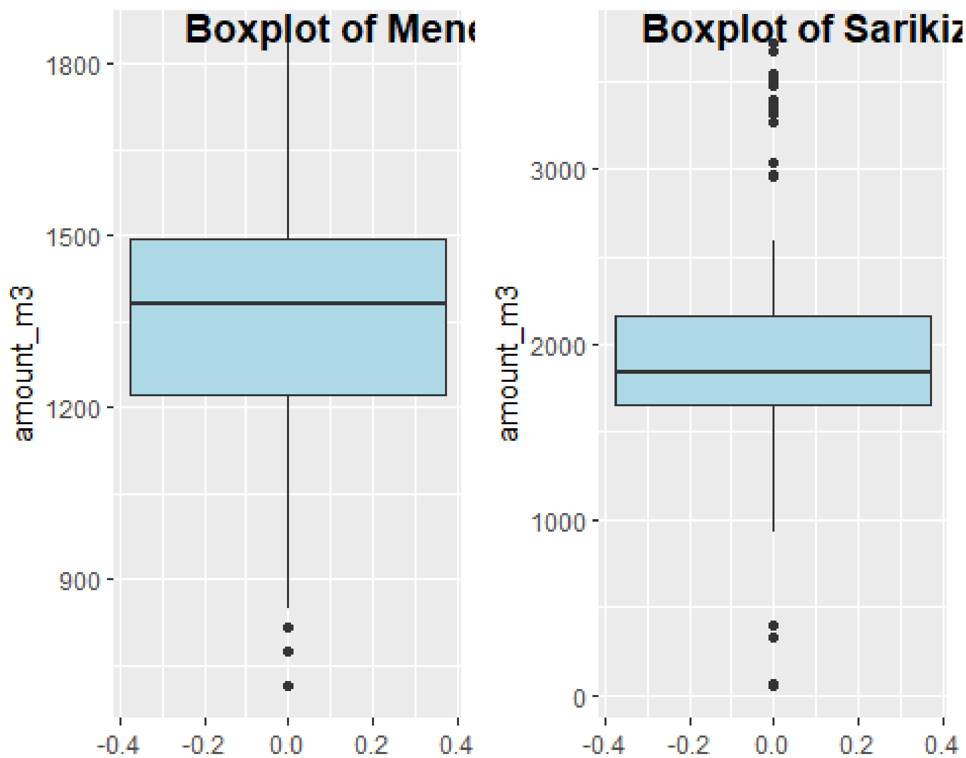



```
shapiro.test(production.menemen$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  production.menemen$amount_m3
## W = 0.96884, p-value = 0.002192

shapiro.test(production.sarikiz$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  production.sarikiz$amount_m3
## W = 0.85975, p-value = 2.008e-10
```

Note: It seems that in Sarikiz data there are lot's of outliers and missing information. We'll just take the positive values for amount_m3 and plot the boxplot again.

```
production.sarikiz = subset(production.sarikiz, subset =
production.sarikiz$amount_m3 > 0)
p1 <- ggplot(production.menemen, aes(y = amount_m3))
p1 <- p1 + geom_boxplot(fill=I("lightblue"))

p2 <- ggplot(production.sarikiz, aes(y = amount_m3))
p2 <- p2 + geom_boxplot(fill=I("lightblue"))
```

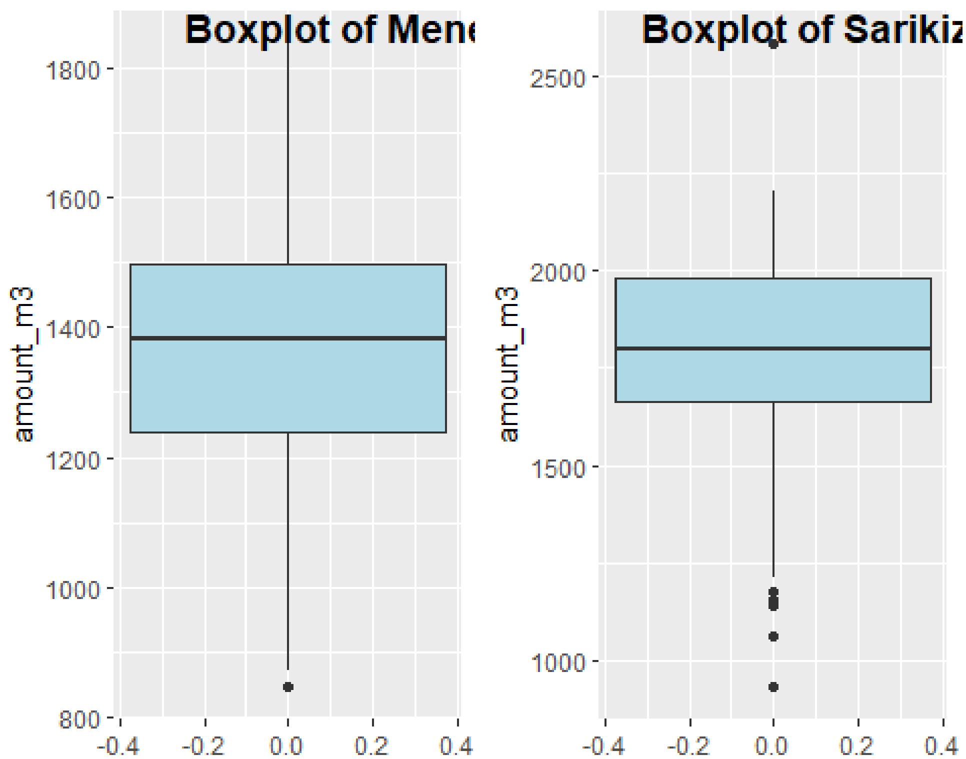
```
ggarrange(p1, p2, ncol =2, nrow = 1, labels = c( "Boxplot of 
Menemen","Boxplot of Sarikiz"))
```



There are still outliers for Menemen and Sarikiz. We'll remove the outliers and plot the data again. We'll use the Iner Quartile Range for detecting outliers. And an outlier would be a point below [Q1- (1.5)IQR] or above [Q3+(1.5)IQR].

```
Q.menemen <- quantile(production.menemen$amount_m3, probs=c(.25, .75), na.rm 
= FALSE)
Q.sarikiz <- quantile(production.sarikiz$amount_m3, probs=c(.25, .75), na.rm 
= FALSE)

iqr.menemen <- IQR(production.menemen$amount_m3)
iqr.sarikiz <- IQR(production.sarikiz$amount_m3)

low.menemen<- Q.menemen[1]-1.5*iqr.menemen # Lower Range of Menemen

up.sarikiz <-  Q.sarikiz[2]+1.5*iqr.sarikiz # Upper Range of Sarikiz
low.sarikiz<- Q.sarikiz[1]-1.5*iqr.sarikiz # Lower Range of Sarikiz


production.menemen.outlier <- subset(production.menemen, subset = 
production.menemen$amount_m3 > low.menemen)
production.sarikiz.outlier <- subset(production.sarikiz, subset = 
production.sarikiz$amount_m3 < up.sarikiz & production.sarikiz$amount_m3 >
```

```
low.sarikiz)


p1 <- ggplot(production.menemen.outlier, aes(y = amount_m3))
p1 <- p1 + geom_boxplot(fill=I("lightblue"))

p2 <- ggplot(production.sarikiz.outlier, aes(y = amount_m3))
p2 <- p2 + geom_boxplot(fill=I("lightblue"))


ggarrange(p1, p2, ncol =2, nrow = 1, labels = c( "Boxplot of
Menemen","Boxplot of Sarikiz"))
```



We almost remove all the outliers and the above boxplot suggests that Menemen Kuyulari produce less water than Sarikiz Kuyulari. Now we can check for normality test.

From Shapiro Test our p-values for two data is less than alpha(0.05) than we reject that our sets of data normally distributed. Let's try square root transformation.

```
shapiro.test(production.menemen.outlier$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  production.menemen.outlier$amount_m3
## W = 0.97305, p-value = 0.006617
```

```
shapiro.test(production.sarikiz.outlier$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  production.sarikiz.outlier$amount_m3
## W = 0.95466, p-value = 0.01195

production.menemen.log <- transform(production.menemen.outlier,
            amount_m3 = log(amount_m3))

production.sarikiz.log <- transform(production.sarikiz.outlier,
            amount_m3 = log(amount_m3))

shapiro.test(production.menemen.log$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  production.menemen.log$amount_m3
## W = 0.93969, p-value = 8.605e-06

shapiro.test(production.sarikiz.log$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  production.sarikiz.log$amount_m3
## W = 0.90931, p-value = 8.244e-05
```

As we can see after transformation p-values for Shapiro normality test became almost zero and we reject that our data normally distributed. From now on we'll assume that our data normally distributed.

- We'll assume that we don't have an information about sample standard deviation and assume that there is no significant difference between variances, and we'll use t-test.

- **H0:** Mu consumptionSarikiz = Mu consumptionMenemen and **H1:** Mu consumptionSarikiz != Mu consumptionMenemen

```
sarikiz.menemen.t.test <- t.test(x=production.sarikiz.outlier$amount_m3,
                     y=production.sarikiz.outlier$amount_m3, var.equal =
TRUE)

cons.sarikiz.menemen.diff <- round((sarikiz.menemen.t.test$estimate[1] -
sarikiz.menemen.t.test$estimate[2]), 1)

# Confidence level as a %
conf.level <- attr(sarikiz.menemen.t.test$conf.int, "conf.level") * 100
```

Results: Our study finds that water productions are on average 0m^3 between the Sarıkız Kuyuları and Menemen - Çavuşköy Kuyuları (t-statistic 0, p=1, 95% CI [-105.7, 105.7]m^3 / 10^3)

Our p-value is 1(greater than alpha) and Confidence interval includes 0. That's mean we can not reject that there is no significant difference in average water production between Sarıkız Kuyuları and Menemen - Çavuşköy Kuyuları. In other words, we can not say the average water production in Sarıkız Kuyuları and Menemen - Çavuşköy Kuyuları are significantly different from each other.

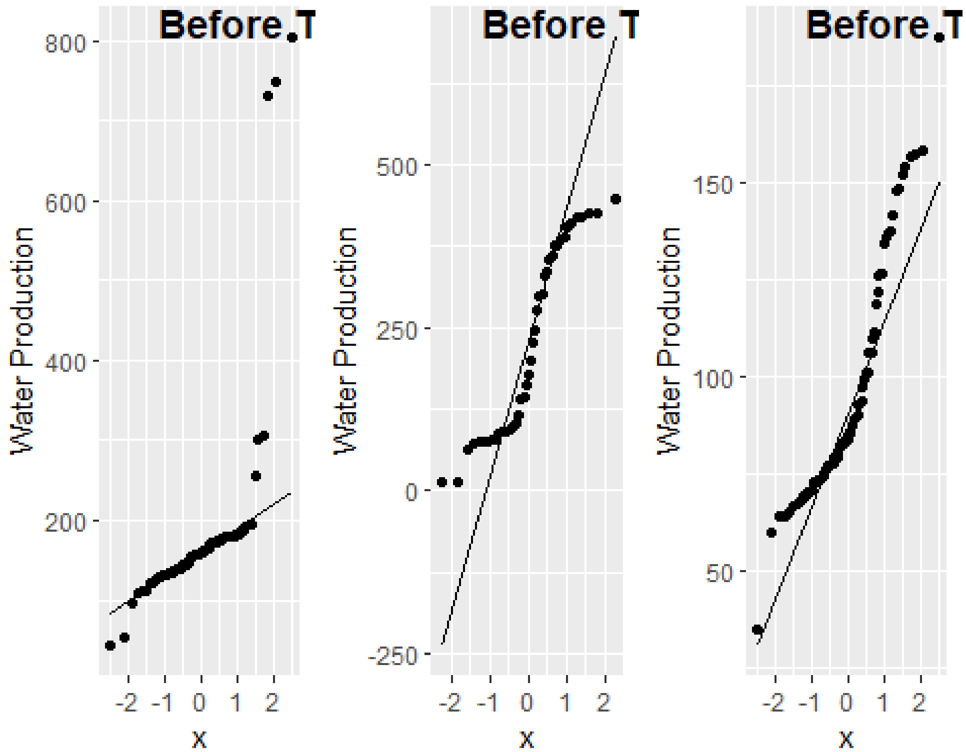
ANOVA

- Let's compare the less productive three source and see if the average production of each are equal or not.
- List the production of the sources by descending order.

```
df1 <- water.production %>%
  group_by(source) %>%
  summarize(total_amount = sum(amount_m3))

df1 <- df1 %>%
  arrange(desc(total_amount))

df1 %>% tbl_df %>% print(n=15)

## # A tibble: 13 x 2
##    source                              total_amount
##    <fct>                                      <dbl>
##  1 Tahtalı Barajı                         930452445
##  2 Göksu Kuyuları                         575342231
##  3 Halkapınar Kuyuları                    378593273
##  4 Menemen - Çavuşköy Kuyuları            195378337
##  5 Sarıkız Kuyuları                       188932250
##  6 Gördes Barajı                          137755577
##  7 Balçova Barajı                          63903955
##  8 Alaçatı Kutlu Aktaş Barajı              30424481
##  9 Pınarbaşı Kuyuları                      17687851
## 10 Ürkmez Barajı                           14939942
## 11 Güzelhisar Barajı                       14627134
## 12 Ödemiş İçme Suyu Arıtma Tesisleri        9434313
## 13 Buca ve Sarnıç Kuyuları                  7697653
```

As we can see from above the less productive three sources are Güzelhisar Barajı, Ödemiş İçme Suyu Arıtma Tesisleri and Buca ve Sarnıç Kuyuları. Now visualize them and see if the data normally distributed or not. Before seperating the groups amount_m3 will be divided by 10^3 for observing the values better.

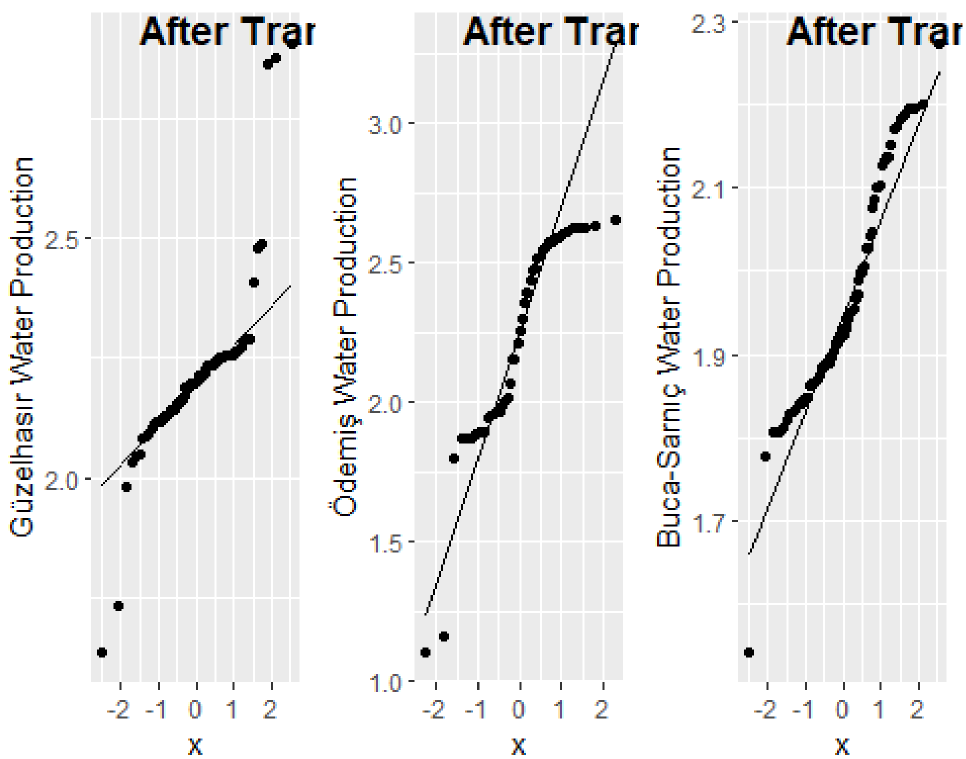
```r
water.production.1 <- water.production
water.production.1$amount_m3 <- (water.production$amount_m3) / 1000

water.cons.guzelhasir= subset(water.production.1, subset = source ==
"Güzelhisar Barajı" & amount_m3 > 0)
water.cons.odemis = subset(water.production.1, subset = source == "Ödemiş
İçme Suyu Arıtma Tesisleri" & amount_m3 > 0)
water.cons.buca.sarnic = subset(water.production.1, subset = source == "Buca
ve Sarnıç Kuyuları" & amount_m3 > 0)

p1 <- ggplot(water.cons.guzelhasir, aes(sample = amount_m3)) + ylab("Water
Production") +
  stat_qq() +
  stat_qq_line()

p2 <- ggplot(water.cons.odemis, aes(sample = amount_m3)) + ylab("Water
Production") +
  stat_qq() +
  stat_qq_line()

p3 <- ggplot(water.cons.buca.sarnic, aes(sample = amount_m3)) + ylab("Water
Production") +
  stat_qq() +
  stat_qq_line()

ggarrange(p1, p2, p3, ncol =3, nrow = 1, labels = c( "Before
Transform","Before Transform",
                                                  "Before Transform"))
```




```
shapiro.test(water.cons.guzelhasir$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  water.cons.guzelhasir$amount_m3
## W = 0.4452, p-value = 6.075e-16

shapiro.test(water.cons.odemis$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  water.cons.odemis$amount_m3
## W = 0.87554, p-value = 0.0002455

shapiro.test(water.cons.buca.sarnic$amount_m3)

##
##  Shapiro-Wilk normality test
##
## data:  water.cons.buca.sarnic$amount_m3
## W = 0.88857, p-value = 3.676e-06
```

From above plots and the Shapiro test we can conclude that three set of data are not normally distributed. Let's transform them and check the normality again.

```
water.cons.guzelhasir.log <- transform(water.cons.guzelhasir,
            amount_m3 = log10(amount_m3))
water.cons.odemis.log <- transform(water.cons.odemis,
            amount_m3 = log10(amount_m3))
water.cons.buca.sarnic.log <- transform(water.cons.buca.sarnic,
            amount_m3 = log10(amount_m3))

p1 <- ggplot(water.cons.guzelhasir.log, aes(sample = amount_m3)) +
ylab("Güzelhasır Water Production") +
  stat_qq() +
  stat_qq_line()

p2 <- ggplot(water.cons.odemis.log, aes(sample = amount_m3)) + ylab("Ödemiş
Water Production") +
  stat_qq() +
  stat_qq_line()

p3 <- ggplot(water.cons.buca.sarnic.log, aes(sample = amount_m3)) +
ylab("Buca-Sarnıç Water Production") +
  stat_qq() +
  stat_qq_line()

ggarrange(p1, p2, p3, ncol =3, nrow = 1, labels = c( "After Transform","After
Transform",
                                                 "After Transform"))
```




```
shapiro.test(water.cons.guzelhasir.log$amount_m3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  water.cons.guzelhasir.log$amount_m3
## W = 0.73836, p-value = 1.043e-10
```

```
shapiro.test(water.cons.odemis.log$amount_m3)
```

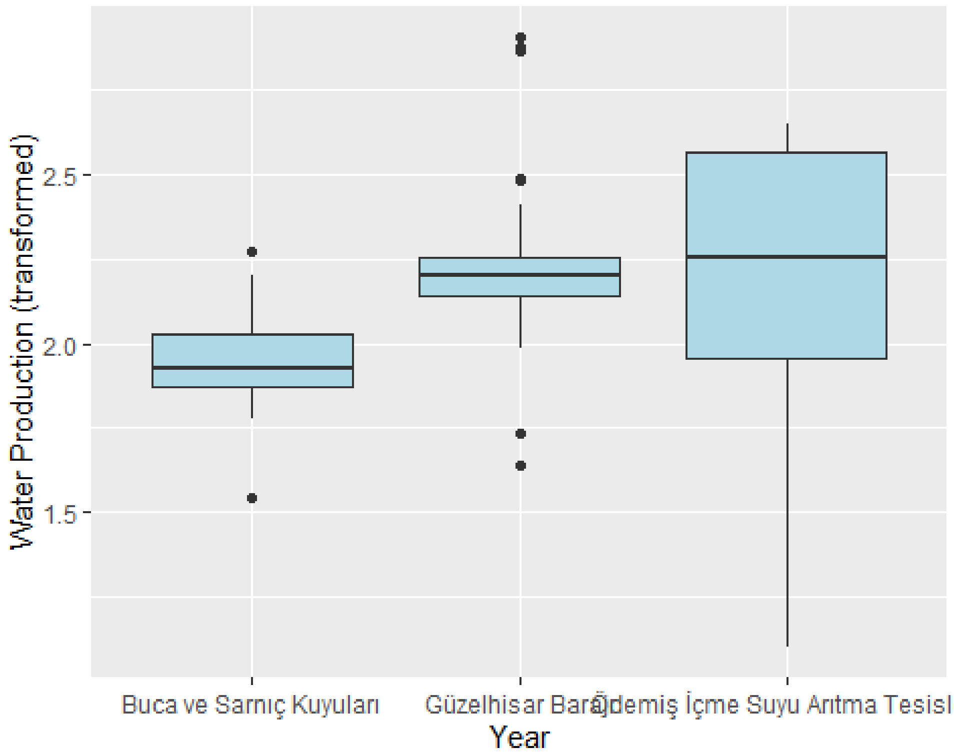
```
##
##  Shapiro-Wilk normality test
##
## data:  water.cons.odemis.log$amount_m3
## W = 0.87129, p-value = 0.0001879
```

```
shapiro.test(water.cons.buca.sarnic.log$amount_m3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  water.cons.buca.sarnic.log$amount_m3
## W = 0.93909, p-value = 0.0007916
```

After log transformassion the data still not normal and from now on we'll assume that the data normally distributed. Let's plot the boxplot before compute the ANOVA test.

```
water.production.1 <- subset(water.production.1, subset =  (source ==
"Güzelhisar Barajı" | source == "Buca ve Sarnıç Kuyuları"
                                                                    | source
== "Ödemiş İçme Suyu Arıtma Tesisleri") & amount_m3 > 0)

water.production.1.log <- transform(water.production.1,
            amount_m3 = log10(amount_m3))

qplot(x =source, y = amount_m3,
      geom = "boxplot", data = water.production.1.log,
      xlab = "Year",
      ylab = "Water Production (transformed)",
      fill = I("lightblue"))
```



Boxplot suggest us that mean water production between Buca and Güzelhasır differs but we can not say the same for others.

Now let's compute the One-way ANOVA test.

H0: Mu cons.guzelhasir = Mu cons.odemis = Mu cons.buca.sarnic, **H1:** At least one differs.

```
# Compute the analysis of variance
res.aov <- aov(amount_m3 ~ source, data = water.production.1.log)
# Summary of the analysis
summary(res.aov)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## source        2  3.144  1.5719   32.32 6.63e-13 ***
## Residuals   202  9.824  0.0486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result: The p-value obtained from ANOVA is almost 0 and less than alpha(0.05) therefore we reject that the average water production of three sources is the same. In other words, there is strong evidence that at least one differs.

- Let's compute the Tukey Test and see which are differs.

```
TukeyHSD(res.aov)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
```

```
## Fit: aov(formula = amount_m3 ~ source, data = water.production.1.log)
##
## $source
##                                                                    diff
## Güzelhisar Barajı-Buca ve Sarnıç Kuyuları                   0.252092586
## Ödemiş İçme Suyu Arıtma Tesisleri-Buca ve Sarnıç Kuyuları 0.255566942
## Ödemiş İçme Suyu Arıtma Tesisleri-Güzelhisar Barajı       0.003474356
##                                                                     lwr
upr
## Güzelhisar Barajı-Buca ve Sarnıç Kuyuları                    0.17027190
0.3339133
## Ödemiş İçme Suyu Arıtma Tesisleri-Buca ve Sarnıç Kuyuları  0.15731871
0.3538152
## Ödemiş İçme Suyu Arıtma Tesisleri-Güzelhisar Barajı       -0.09477387
0.1017226
##                                                                 p adj
## Güzelhisar Barajı-Buca ve Sarnıç Kuyuları                   0.0000000
## Ödemiş İçme Suyu Arıtma Tesisleri-Buca ve Sarnıç Kuyuları 0.0000000
## Ödemiş İçme Suyu Arıtma Tesisleri-Güzelhisar Barajı       0.9961638
```

Result: The lower and upper bound of:

- Güzelhisar Barajı-Buca ve Sarnıç Kuyuları: (0.172, 0.333)
- Ödemiş İçme Suyu Arıtma Tesisleri-Buca ve Sarnıç Kuyuları: (0.157, 0.353)
- Ödemiş İçme Suyu Arıtma Tesisleri-Güzelhisar Barajı: (-0.094, 0.101)

Conclusion: From the above results we can conclude that:

- The interval of Güzelhisar Barajı-Buca ve Sarnıç Kuyuları doesn't contain 0 therefore there is strong evidence that these two differ.
- Ödemiş İçme Suyu Arıtma Tesisleri-Buca ve Sarnıç Kuyuları: doesn't contain 0 therefore there is strong evidence that these two differ.
- Ödemiş İçme Suyu Arıtma Tesisleri-Güzelhisar Barajı: contains 0 therefore there is strong evidence that these two do not differ.

Discussion

Water Consumption

- From the statistical analysis, the average water consumption in Summer is higher than in Winter. If there wouldn't be missing pieces of information for the 4th and 5th months this comparison could have been included for four seasons and the authorizes could take their precautions for the seasons.
- The average Water Consumption in the most crowded district Buca is over than 25000m3 and the average water consumption in Izmir is just 10000m3 almost half of consumption in Buca.

Water Production

- Our study statistically proves that after 2015, the most productive source Tahtali Baraji became much more productive and produce 1812400 m^3 more water on average than before 2015. This statistic could be beneficial for the improvement of other sources too.

- The least productive three source; Güzel Hisar Baraji, Buca ve Sarnic Kuyulari and Ödemiş İçme Suyu Arıtma Tesisleri has different water production on average by years. These sources may improve by the authorizes and could be more beneficial for the people in İzmir and neighbors of it.

Water Binding

- Our study proves that the average tax collection time improved by each year and taxes are collected from users faster. This system could be continued and improved in the next years better.

- The average water subscription time in Izmir is calculated as 34 days for each year. This statistic could be improved by the authorizes with try to develop their sources and using them sufficiently for the people and keep them as subscriber longer.