

PRAC2

Bernardo Columba - Paula Borao

15/12/2021

Conjunto de datos de demanda de bicicletas compartidas de Seúl

El conjunto de datos contiene el recuento de bicicletas públicas alquiladas cada hora en el sistema de bicicletas públicas de Seúl con los datos meteorológicos y la información de vacaciones correspondientes.

Fuente: <http://data.seoul.go.kr/> SOUTH KOREA PUBLIC HOLIDAYS. URL: publicholidays.go.kr

Repositorio: <https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand#>

1. Familiarizándose con los datos

Actualmente, las bicicletas de alquiler se introducen en muchas ciudades urbanas para mejorar la comodidad de la movilidad. Es importante que la bicicleta de alquiler esté disponible y sea accesible al público en el momento adecuado, ya que reduce el tiempo de espera. Con el tiempo, proporcionar a la ciudad un suministro estable de bicicletas de alquiler se convierte en una gran preocupación. La parte crucial es la predicción del recuento de bicicletas necesario a cada hora para el suministro estable de bicicletas de alquiler. El conjunto de datos contiene información meteorológica (temperatura, humedad, velocidad del viento, visibilidad, punto de rocío, radiación solar, nevadas, precipitaciones), el número de bicicletas alquiladas por hora e información sobre la fecha.

Con estos sistemas de bicicletas compartidas, las personas alquilan una bicicleta en un lugar y la devuelven a un lugar diferente o al mismo según sea necesario. Las personas pueden alquilar una bicicleta mediante membresía (en su mayoría usuarios habituales) o bajo demanda (en su mayoría usuarios ocasionales). Este proceso está controlado por una red de quioscos automatizados en toda la ciudad.

2. Información sobre los atributos

1. Date -> Fecha: año-mes-día
2. Rented Bike Count -> Recuento_de_bicicletas_alquiladas: recuento de bicicletas alquiladas cada hora
3. Hour -> Hora: Hora del día
4. Temperature(°C) -> Temperatura: Temperatura en grados Celsius
5. Humidity(%) -> Humedad_porcentaje: Humedad en porcentaje
6. Wind speed (m/s) -> Velocidad_viento: Velocidad del viento (m/s)
7. Visibility (10m) -> Visibilidad: Visibilidad a 10 m
8. Dew point temperature(°C) -> Temperatura_punto_rocio: Temperatura del punto de rocío en Celsius
9. Solar Radiation (MJ/m2) -> Radiación_solar: Radiación solar en (MJ/m2)
10. Rainfall(mm) -> Precipitaciones: Precipitaciones en mm
11. Snowfall (cm) -> Nevada: Nevada en cm
12. Seasons -> Temporadas:
 - a. invierno

- b. primavera
 - c. verano
 - d. otoño
13. Holiday-> Vacaciones: a.Vacaciones b.No vacaciones
14. Functioning Day -> Dia_laboral - NoFunc (horas no funcionales), diversión (horas funcionales)
- a. Si
 - b. No

3. Carga del archivo de datos

```
#Cargo la base de datos de formato csv, separados por coma ','
data <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-databases/00560/SeoulBikeData.csv',st
#Permite acceder a cada columna directamente
attach(data)
```

4. Revisión de los atributos del Dataset

Renombro las columnas con nombres mas comprensibles

```
names(data) = c("Fecha", "Recuento_de_bicicletas_alquiladas", "Hora", "Temperatura", "Humedad_porcentaje",
#Permite acceder a cada columna directamente
attach(data)
#Verifico la dimensión del juego de datos
dim(data)

## [1] 8760 14

# Vemos la estructura del dataset
str(data)

## 'data.frame': 8760 obs. of 14 variables:
## $ Fecha : chr "01/12/2017" "01/12/2017" "01/12/2017" "01/12/2017" ...
## $ Recuento_de_bicicletas_alquiladas: int 254 204 173 107 78 100 181 460 930 490 ...
## $ Hora : int 0 1 2 3 4 5 6 7 8 9 ...
## $ Temperatura : num -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
## $ Humedad_porcentaje : int 37 38 39 40 36 37 35 38 37 27 ...
## $ Velocidad_viento : num 2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
## $ Visibilidad : int 2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 ...
## $ Temperatura_punto_rocio : num -17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8 -22 ...
## $ Radiación_solar : num 0 0 0 0 0 0 0 0.01 0.23 ...
## $ Precipitaciones : num 0 0 0 0 0 0 0 0 0 ...
## $ Nevada : num 0 0 0 0 0 0 0 0 0 ...
## $ Temporadas : chr "Winter" "Winter" "Winter" "Winter" ...
## $ Vacaciones : chr "No Holiday" "No Holiday" "No Holiday" "No Holiday" ...
## $ Dia_laboral : chr "Yes" "Yes" "Yes" "Yes" ...
```

```
# Estadística básica de los atributos del dataset
summary(data)
```

```
##      Fecha          Recuento_de_bicicletas_alquiladas      Hora
##  Length:8760      Min.   : 0.0                  Min.   : 0.00
##  Class :character  1st Qu.: 191.0                1st Qu.: 5.75
##  Mode  :character  Median : 504.5               Median :11.50
##                  Mean   : 704.6               Mean   :11.50
##                  3rd Qu.:1065.2              3rd Qu.:17.25
##                  Max.   :3556.0              Max.   :23.00
##      Temperatura    Humedad_porcentaje  Velocidad_viento  Visibilidad
##  Min.   :-17.80      Min.   : 0.00      Min.   :0.000      Min.   : 27
##  1st Qu.:  3.50      1st Qu.:42.00      1st Qu.:0.900      1st Qu.: 940
##  Median : 13.70     Median :57.00      Median :1.500      Median :1698
##  Mean   : 12.88     Mean   :58.23      Mean   :1.725      Mean   :1437
##  3rd Qu.: 22.50     3rd Qu.:74.00      3rd Qu.:2.300      3rd Qu.:2000
##  Max.   : 39.40     Max.   :98.00      Max.   :7.400      Max.   :2000
##      Temperatura_punto_rocio Radiación_solar  Precipitaciones      Nevada
##  Min.   :-30.600     Min.   :0.0000      Min.   : 0.00000      Min.   :0.000000
##  1st Qu.: -4.700     1st Qu.:0.0000      1st Qu.: 0.00000      1st Qu.:0.000000
##  Median :  5.100     Median :0.0100      Median : 0.00000      Median :0.000000
##  Mean   :  4.074     Mean   :0.5691      Mean   : 0.1487      Mean   :0.07507
##  3rd Qu.: 14.800     3rd Qu.:0.9300      3rd Qu.: 0.00000      3rd Qu.:0.000000
##  Max.   : 27.200     Max.   :3.5200      Max.   :35.0000      Max.   :8.80000
##      Temporadas        Vacaciones        Dia_laboral
##  Length:8760        Length:8760        Length:8760
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
```

4.1. Tratamiento de los atributos

Retiro (en caso de haberlo) los espacios en blanco de los atributos categóricos

```
#Recorro el juego de datos
for(i in 1:nrow(data)) {
  #Verifico si la columna es de tipo caracter
  if(is.character(data$i)){
    #Retiro los espacios de inicio y fin de cada columna tipo caracter
    data$i <- trimws(data$i)
  }
}

#Paso la Fecha a formato Date de R
data$Fecha=as.Date(data$Fecha,format="%d/%m/%Y")
```

4.2. Estadísticas de valores vacíos

```
# Estadísticas de valores vacíos  
colSums(data=="")
```

```
##                               Fecha Recuento_de_bicicletas_alquiladas  
##                               NA                         0  
##                               Hora                      Temperatura  
##                               0                          0  
##             Humedad_porcentaje          Velocidad_viento  
##                               0                          0  
##             Visibilidad            Temperatura_punto_rocio  
##                               0                          0  
##             Radiación_solar          Precipitaciones  
##                               0                          0  
##             Nevada                  Temporadas  
##                               0                          0  
##             Vacaciones              Dia_laboral  
##                               0                          0
```

```
colSums(is.na(data))
```

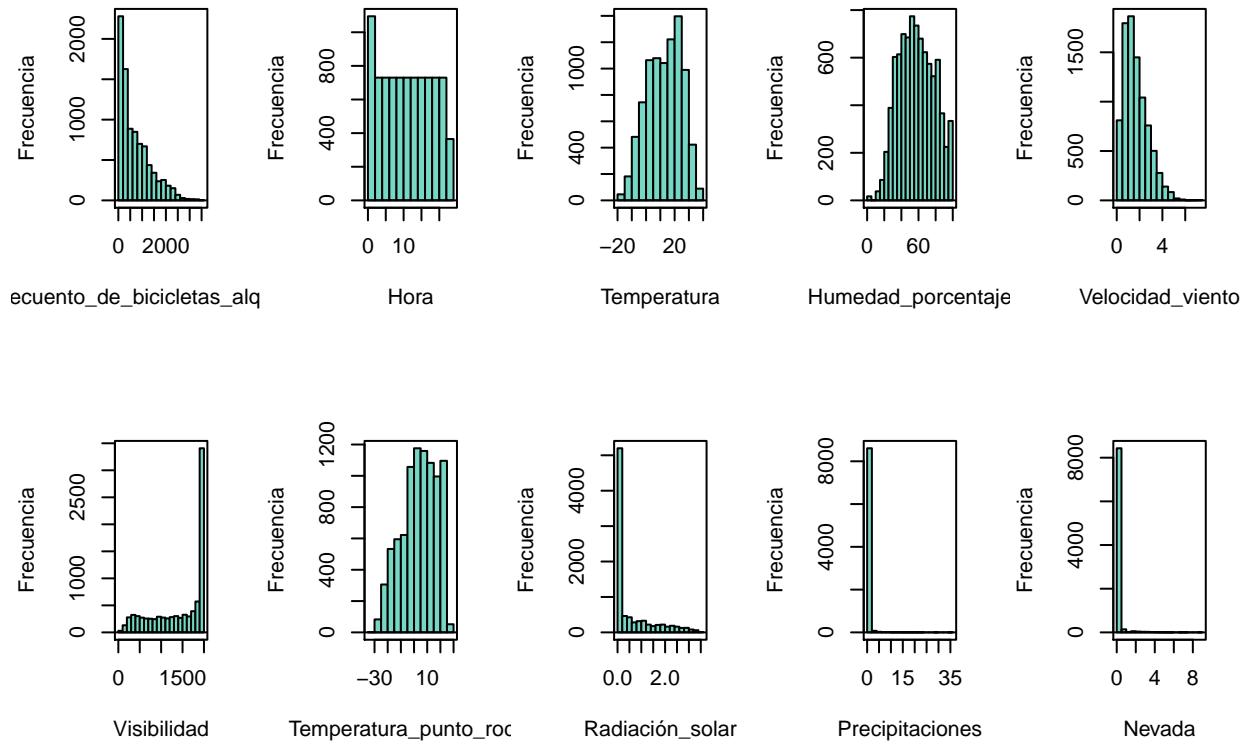
```
##                               Fecha Recuento_de_bicicletas_alquiladas  
##                               0                         0  
##                               Hora                      Temperatura  
##                               0                          0  
##             Humedad_porcentaje          Velocidad_viento  
##                               0                          0  
##             Visibilidad            Temperatura_punto_rocio  
##                               0                          0  
##             Radiación_solar          Precipitaciones  
##                               0                          0  
##             Nevada                  Temporadas  
##                               0                          0  
##             Vacaciones              Dia_laboral  
##                               0                          0
```

No se encuentran valores vacíos, nulos o desconocidos

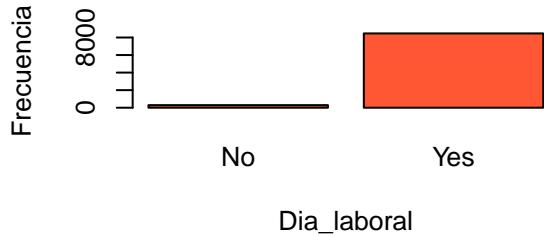
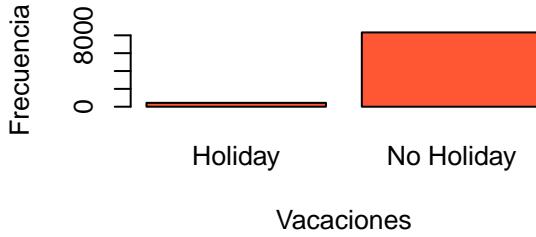
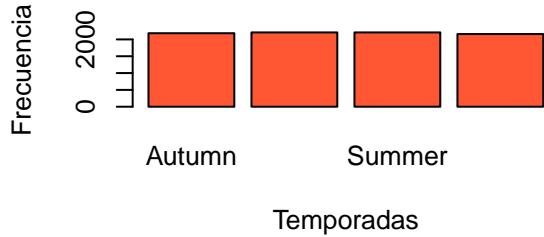
4.3. Análisis descriptivo visual

Se representa de forma visual las variables del conjunto de datos y las distribuciones de sus valores

```
#Visualización de la frecuencia de las variables  
#Variables cuantitativas  
par(mfrow=c(2,5))  
#Recorro el juego de datos  
for (i in 1:ncol(data)){  
  #Verifco si la columna es de tipo numérica  
  if (is.numeric(data[,i])){  
    with(data, Hist(data[,i], scale="frequency", ylab="Frecuencia", xlab=colnames(data)[i], breaks="Stur")  
  }  
}
```



```
#Variables cualitativas
par(mfrow=c(2,2))
#Re corro el juego de datos
for (i in 1:ncol(data)){
  #Verifico si la columna es de tipo caracter
  if (is.character(data[,i])){
    with(data, Barplot(data[,i], xlab=colnames(data)[i], ylab="Frecuencia", col="#FF5733"))
  }
}
```



4.3.1 Análisis de los Histogramas y los Diagramas de Barras Atributos Cuantitativos Histogramas

La máxima frecuencia de bicicletas alquiladas está dentro del rango de [0, 250].

Las bicicletas permanecen alquiladas en mayor rango de [0, 2] horas.

A mayor temperatura mayor número de bicicletas alquiladas, pero si pasa los $30^{\circ}C$ el alquiler de bicicletas disminuye.

Similar ocurre con la humedad a mayor humedad mayor frecuencia de bicicletas alquiladas, pero cuando se sobrepasa la humedad del 60 el alquiler de bicicletas empieza a disminuir lentamente.

Si la velocidad del viento aumenta, la frecuencia de bicicletas alquiladas disminuye.

Se alquilan mayor número de bicicletas cuando la visibilidad es mayor.

Mientras menor sea la radiación solar mayor es el número de bicicletas alquiladas.

El mayor número de bicicletas alquiladas solo se da cuando las precipitaciones son bajas.

Se puede observar también que en gran porcentaje solo se alquilan bicicletas si no hay nevada o muy poca.

Atributos Cualitativos Diagramas de barras

Se observa mayor número de bicicletas alquiladas se dan cuando no hay días festivos.

4.3.2 Hipótesis Iniciales

Con esta información puedo realizarme las primeras preguntas sobre el set de datos y realizar las primeras hipótesis que se contestarán en los siguientes apartados:

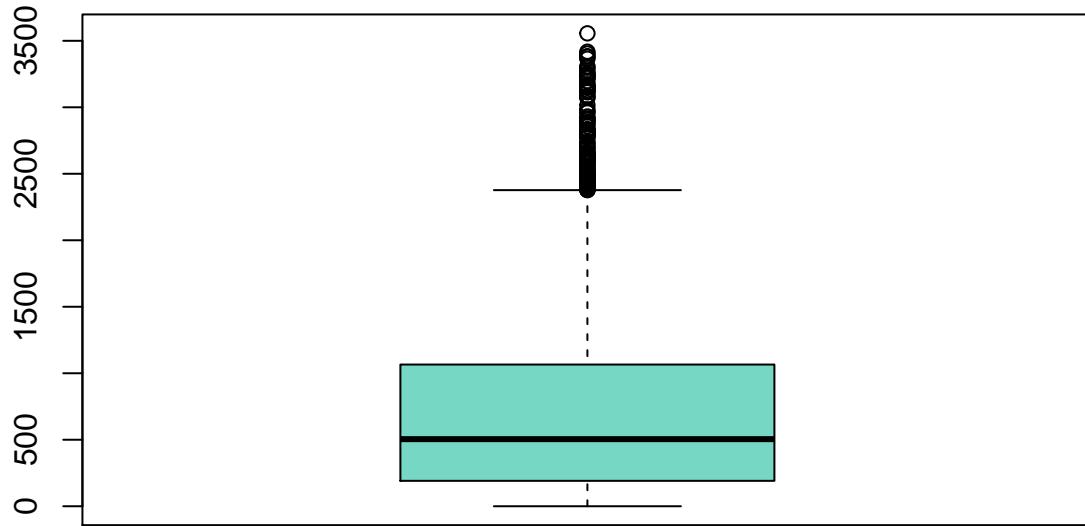
1. ¿El alquiler de bicicletas es mayor en la hora de ingreso y salida del horario laboral, es decir a las 07:00 y a las 17:00?
2. ¿Cuando se realizan mas alquileres de bicicletas entre semana o fines de semana?
3. ¿Que tempordas prefieren los usuarios alquilar bicicletas? ¿Prefieren los días soleados o los días de invierno?
4. ¿Los alquileres de bicicleta son mayores en días festivos o en días normales?
5. ¿En el transcurso de los años el alquiler de bicicleta aumenta o disminuye?
6. ¿Depende el número de alquiler de bicicletas de la radiación solar, es decir mientras mayor es la radiación solar menor es el número de usuarios o viceversa?

4.4 Outliers

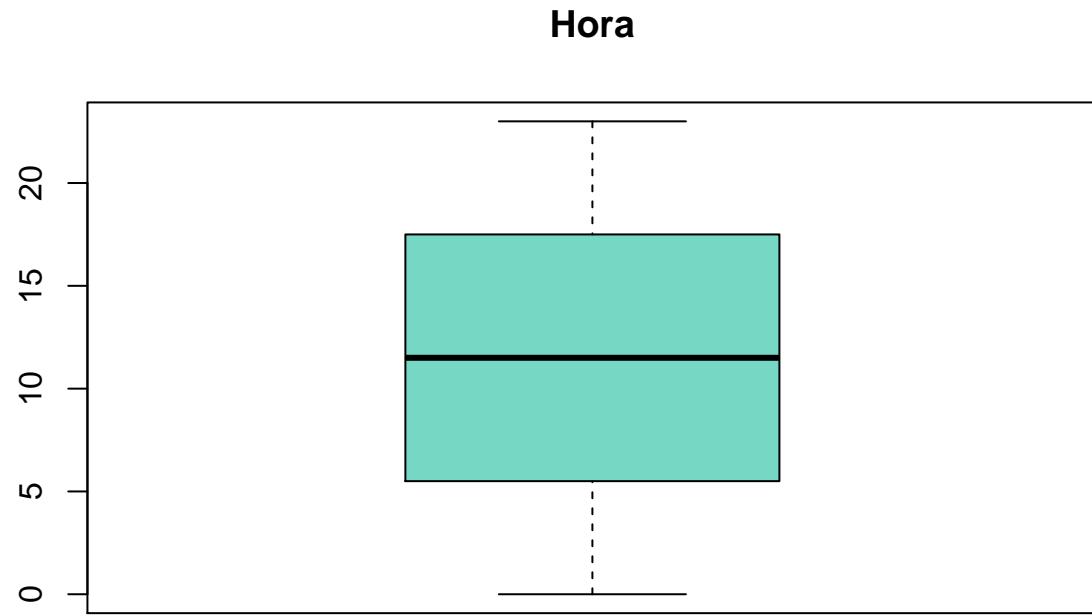
Visualización de cajas de las variables cuantitativas

```
# Genero un bucle para graficar las cajas por cada una de las columnas cuantitativas del juego de datos
for (i in 1:(ncol(data))){
  if (is.numeric(data[,i])){
    boxplot(data[,i], main = colnames(data)[i], col=(c("#76D7C4")))
    #Valores atípicos
    print('Valores atípicos')
    elementos <- boxplot.stats(data[,i])$out #muestra los valores atípicos
    print(elementos)
    longitud <- length(elementos) #Cantidad de valores atípicos
    print('Cantidad de valores atípicos')
    print(longitud)
  }
}
```

Recuento_de_bicicletas_alquiladas

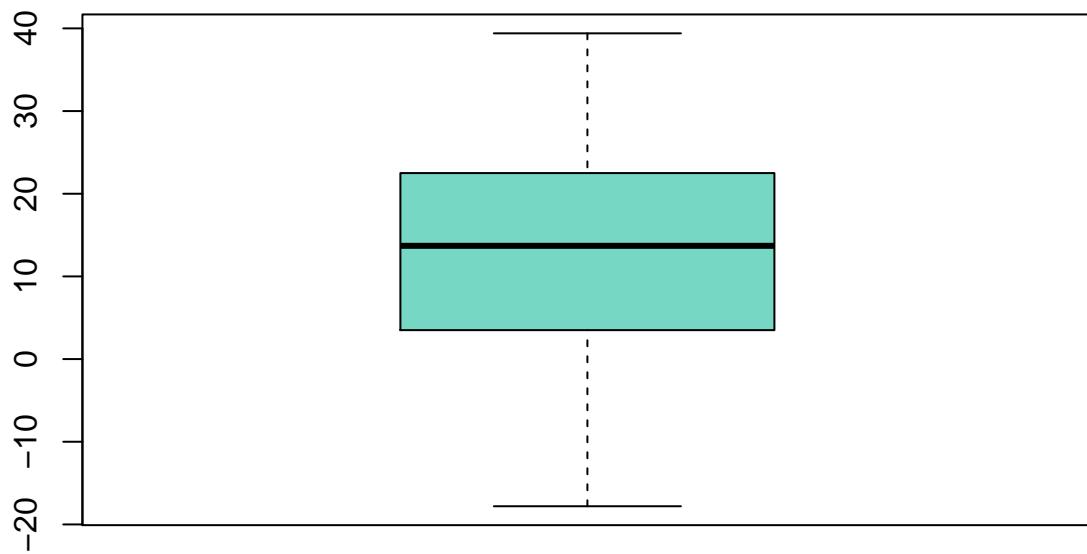


```
## [1] "Valores atípicos"
## [1] 2401 2402 2401 2404 2692 2807 2574 2577 2558 2661 2392 3130 2405 2701 2379
## [16] 2410 2906 2915 2479 2439 2403 3069 2450 3123 2454 2825 2916 3245 2656 3251
## [31] 2650 3119 2534 3088 2505 2514 2460 3380 2788 2508 3227 2615 2404 3221 2649
## [46] 3309 2797 2476 2431 2495 3404 2873 2579 2505 2474 2664 2479 2404 2962 2460
## [61] 2891 2984 2441 2383 2451 2475 2497 2435 2474 2481 2429 2515 2430 3556 2809
## [76] 2525 2378 2440 3384 2741 2519 3418 2811 2556 2436 3365 2732 2493 2387 3238
## [91] 2779 2456 3172 2487 3113 2602 2965 2598 2398 3080 2637 2415 2594 3196 2557
## [106] 2931 3016 2884 2636 2770 2692 2640 2419 2405 2416 2481 2797 2628 2830 2528
## [121] 2836 3166 2491 3160 2468 2497 3154 3298 2518 3222 2455 2391 3256 2443 3146
## [136] 2826 3277 2489 3154 2422 2432 2397 2857 2400 2787 2514 2378 2618 2635 2716
## [151] 2445 2499 2631 2613 2415 2612 2632
## [1] "Cantidad de valores atípicos"
## [1] 157
```



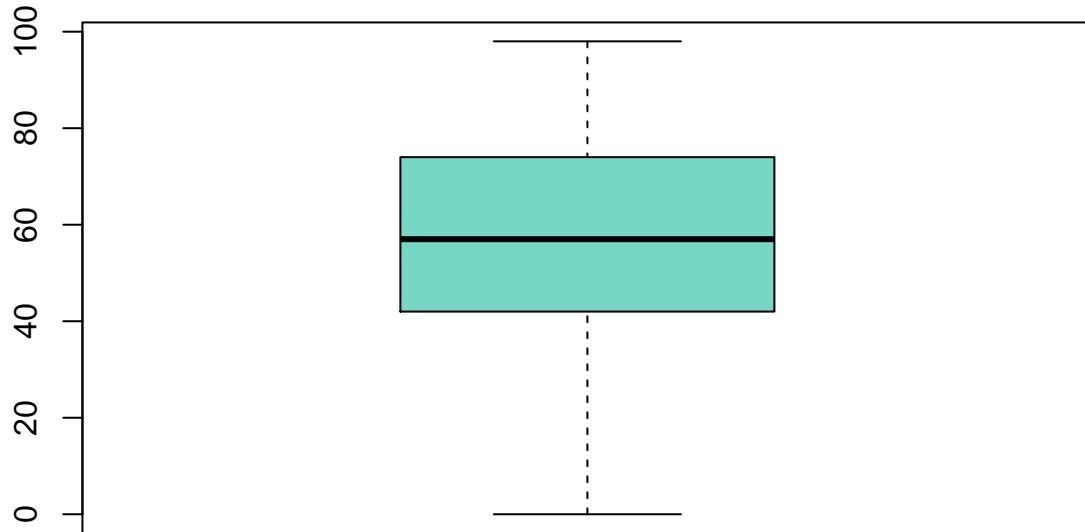
```
## [1] "Valores atípicos"  
## integer(0)  
## [1] "Cantidad de valores atípicos"  
## [1] 0
```

Temperatura



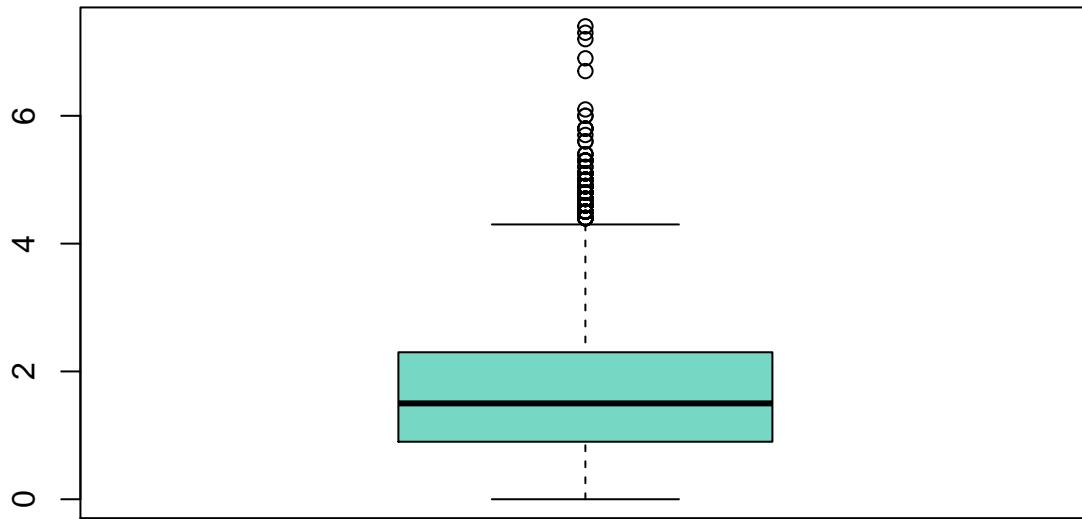
```
## [1] "Valores atípicos"  
## numeric(0)  
## [1] "Cantidad de valores atípicos"  
## [1] 0
```

Humedad_porcentaje



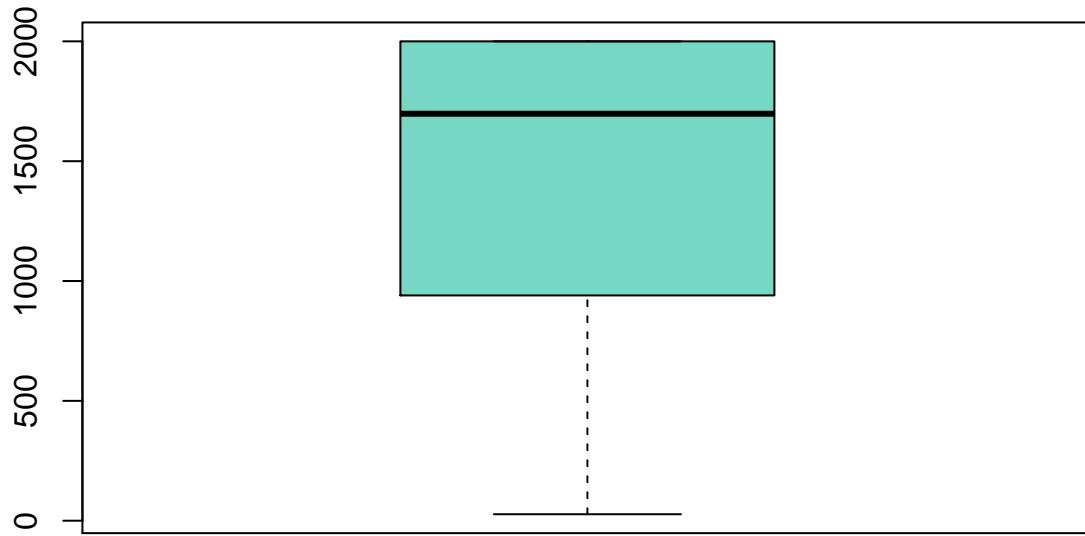
```
## [1] "Valores atípicos"  
## integer(0)  
## [1] "Cantidad de valores atípicos"  
## [1] 0
```

Velocidad_viento



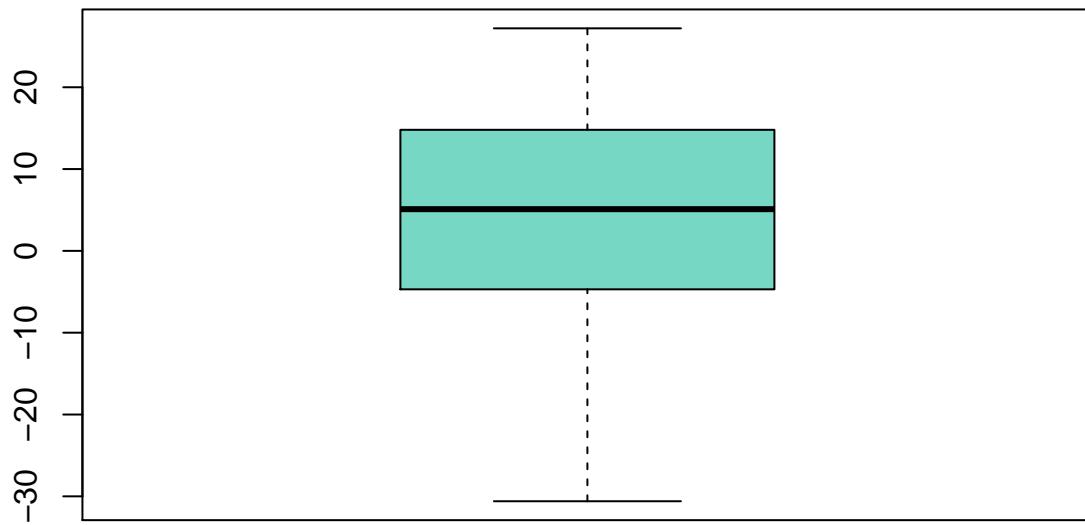
```
## [1] "Valores atípicos"
## [1] 4.5 4.8 5.4 4.5 5.8 4.7 5.3 4.5 4.5 4.7 5.1 4.5 4.6 5.1 4.5 5.0 4.4 4.9
## [19] 4.7 5.0 4.7 4.5 4.4 6.7 4.6 4.6 5.0 4.4 4.5 4.6 4.5 4.5 4.7 5.3 4.8 4.5
## [37] 4.5 4.6 6.0 4.7 5.2 4.6 4.5 4.5 4.8 4.6 4.6 4.4 4.8 4.9 4.7 4.6 4.4 4.5
## [55] 4.4 4.7 4.5 4.9 5.0 5.6 5.1 5.8 4.4 4.5 4.4 5.3 4.6 4.4 5.0 4.9 5.3 4.6
## [73] 4.7 6.0 4.5 4.8 5.2 4.7 4.7 4.7 4.5 4.4 4.6 4.4 5.3 4.6 4.9 4.9 4.9 4.8
## [91] 4.5 4.7 5.0 4.7 5.3 4.6 5.1 7.4 4.8 5.8 5.6 7.2 5.8 6.1 7.3 5.4 4.6 4.7
## [109] 4.4 4.6 5.1 4.9 4.8 4.4 4.7 4.7 4.8 4.8 4.6 4.9 5.1 4.6 4.7 4.5 4.8 5.3
## [127] 4.9 5.0 4.4 4.7 4.5 4.6 4.7 4.9 4.5 6.9 4.5 4.9 5.0 4.8 5.0 5.1 5.0 4.7
## [145] 4.4 4.6 4.4 4.5 4.7 5.0 5.0 4.9 5.0 5.4 4.9 4.6 4.4 5.7 4.9 4.7 5.3
## [1] "Cantidad de valores atípicos"
## [1] 161
```

Visibilidad



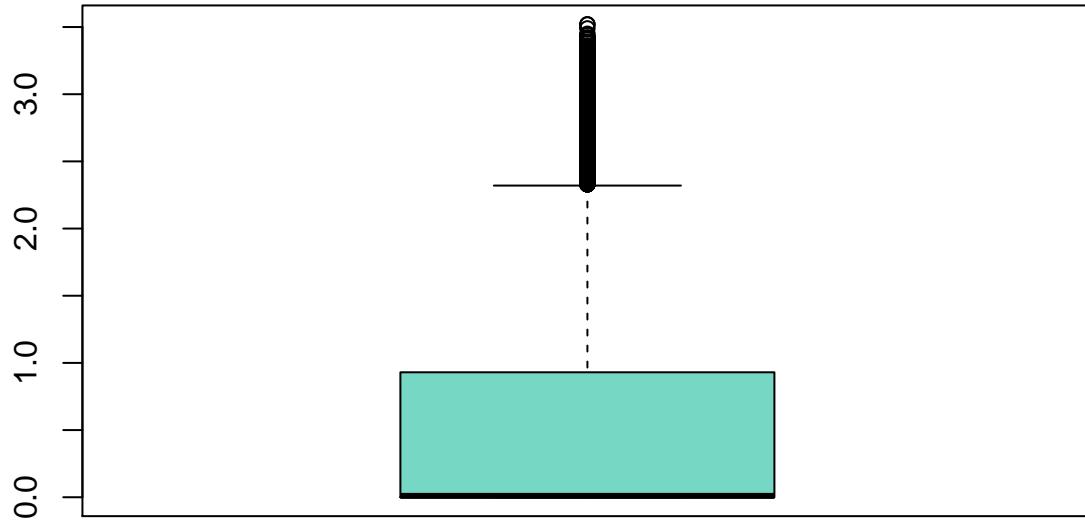
```
## [1] "Valores atípicos"  
## integer(0)  
## [1] "Cantidad de valores atípicos"  
## [1] 0
```

Temperatura_punto_rocio



```
## [1] "Valores atípicos"  
## numeric(0)  
## [1] "Cantidad de valores atípicos"  
## [1] 0
```

Radiación_solar



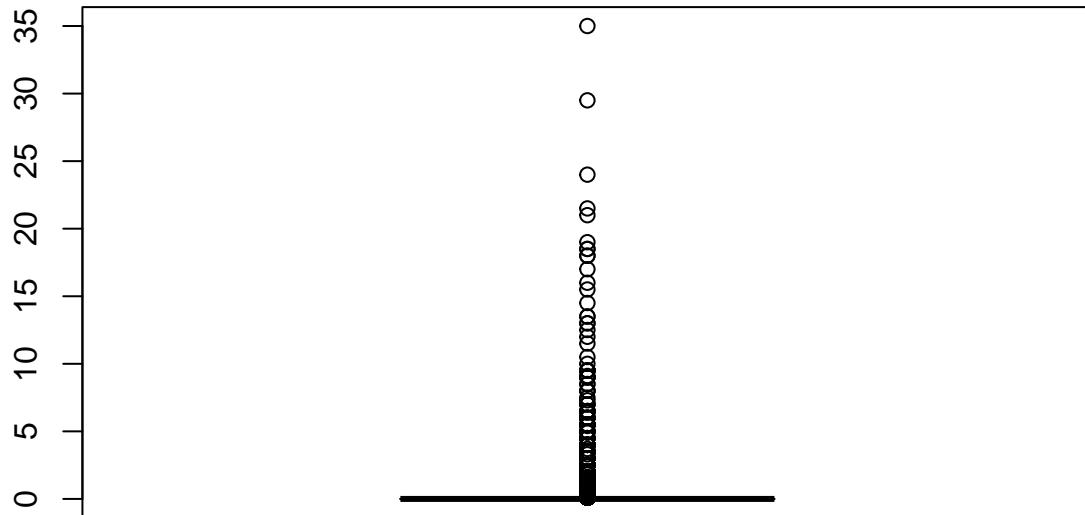
```
## [1] "Valores atípicos"
##   [1] 2.35 2.46 2.39 2.49 2.36 2.50 2.52 2.35 2.48 2.43 2.49 2.95 2.86 2.54 2.61
##  [16] 2.83 2.74 2.48 2.56 2.75 2.66 2.33 2.46 2.45 2.55 2.65 2.47 2.66 2.53 2.47
##  [31] 2.67 2.69 2.34 2.44 2.60 2.43 2.66 2.49 2.38 2.78 2.99 2.91 2.62 2.51 2.52
##  [46] 2.36 2.59 2.33 2.70 2.91 2.82 2.51 2.72 2.87 2.71 2.41 2.48 2.42 2.63 2.75
##  [61] 2.62 2.65 2.80 2.70 2.44 2.54 2.37 2.73 2.87 2.75 2.40 2.41 2.53 2.57 2.83
##  [76] 3.24 3.04 2.79 2.98 2.99 2.49 2.93 3.06 2.99 2.60 2.59 2.94 2.77 2.75 3.05
##  [91] 3.32 3.16 2.90 2.39 2.68 3.07 3.20 3.13 2.86 2.39 2.52 2.71 2.67 3.10 3.21
## [106] 3.11 2.80 2.66 3.03 3.16 3.01 2.80 2.82 3.13 3.27 3.17 2.88 2.52 2.79 2.89
## [121] 2.83 2.53 2.51 2.88 3.03 2.92 2.59 2.85 3.19 3.31 3.29 2.98 2.50 2.74 3.08
## [136] 3.18 3.12 2.75 2.55 2.67 2.84 2.66 2.51 2.89 2.99 2.91 2.66 2.84 3.18 3.25
## [151] 3.22 2.96 2.45 2.79 2.79 3.35 2.93 2.79 2.65 2.96 2.87 2.47 2.40 2.93 3.28
## [166] 3.39 3.31 3.00 2.49 2.56 2.74 2.51 2.57 2.72 2.33 3.05 2.35 2.46 2.89 3.15
## [181] 2.46 2.41 2.88 3.21 3.34 3.10 2.88 2.38 2.38 2.84 3.17 3.26 3.22 2.87 2.41
## [196] 2.53 3.08 3.17 2.52 2.40 2.38 2.74 2.67 2.70 2.88 3.14 2.51 2.61 3.12 3.42
## [211] 3.52 3.39 3.07 2.54 2.63 3.11 3.39 3.49 3.11 2.63 3.13 3.44 3.52 3.41 3.11
## [226] 2.57 2.44 2.92 3.20 3.28 3.18 2.91 2.42 2.53 3.01 3.28 3.36 2.92 2.45 2.92
## [241] 3.18 3.29 3.14 2.82 2.46 2.90 3.24 3.33 3.21 2.50 2.76 2.94 2.92 2.55 2.44
## [256] 2.45 2.89 3.20 3.34 2.97 2.90 2.41 2.45 2.99 3.26 3.30 3.25 2.94 2.46 2.52
## [271] 2.98 3.29 3.42 3.32 3.00 2.58 2.83 3.29 3.18 3.20 2.95 2.48 2.37 2.79 2.61
## [286] 3.03 2.36 2.49 2.91 3.23 2.82 2.37 2.84 3.16 3.21 3.01 2.79 2.36 2.87 3.19
## [301] 3.26 3.11 2.63 2.87 2.35 2.42 2.88 2.84 2.38 2.53 2.45 2.67 2.56 3.18 2.45
## [316] 2.95 3.07 3.42 3.23 2.42 2.46 2.96 3.27 3.36 3.26 2.89 2.58 2.40 2.84 2.92
## [331] 3.10 2.63 2.87 2.51 2.59 2.33 3.38 3.36 3.24 3.04 2.55 2.54 3.01 3.26 3.42
## [346] 3.14 2.96 2.62 2.57 3.02 3.21 3.17 3.21 2.57 2.70 2.93 2.88 2.65 2.41 2.33
## [361] 2.82 3.15 3.24 3.12 2.85 2.38 2.90 2.91 2.71 2.38 2.37 2.62 2.47 2.36 2.72
```

```

## [376] 2.54 3.11 2.99 2.61 2.75 2.46 2.51 2.48 2.95 3.26 3.41 3.33 2.98 2.33 2.59
## [391] 2.93 3.33 3.24 2.58 2.64 2.67 2.34 3.28 3.13 2.81 2.34 2.83 3.09 3.09 3.23
## [406] 2.95 2.51 2.36 2.74 2.52 2.63 2.93 3.01 2.55 2.77 2.96 2.96 2.62 2.90 2.52
## [421] 2.38 2.92 3.23 3.36 3.26 3.04 2.55 2.84 3.21 2.50 2.67 2.88 2.47 2.72 3.14
## [436] 3.21 3.17 2.71 2.41 2.74 3.05 3.24 3.05 2.46 2.53 2.94 2.48 2.86 2.40 2.57
## [451] 2.50 2.89 2.33 2.60 2.33 2.48 2.34 2.80 2.86 3.12 2.46 2.81 2.77 2.80 3.05
## [466] 3.12 2.82 2.42 2.58 2.56 2.53 2.71 2.56 2.47 2.35 2.33 2.68 3.05 2.67 2.37
## [481] 2.93 2.44 2.72 2.37 2.33 2.51 2.59 2.34 2.84 2.57 2.37 2.40 2.87 2.72 2.73
## [496] 2.50 2.68 2.68 2.58 2.73 2.46 2.71 3.19 3.45 3.17 2.80 2.36 2.34 2.85 3.17
## [511] 3.30 3.21 2.90 2.39 2.82 3.14 3.24 2.80 2.98 3.10 2.73 2.45 2.72 2.63 2.95
## [526] 3.10 2.87 2.46 2.71 2.68 2.73 2.50 2.65 2.99 3.10 3.00 2.72 2.63 2.96 2.88
## [541] 2.37 2.62 2.75 2.94 2.39 2.40 2.70 3.02 2.87 2.67 2.38 2.80 2.92 2.79 2.65
## [556] 2.62 2.85 2.37 2.59 2.62 2.93 3.06 2.96 2.65 2.61 2.89 3.12 3.09 2.90 2.59
## [571] 2.92 2.56 2.67 2.60 2.64 2.98 2.55 2.91 2.60 2.54 2.55 2.83 2.82 2.80 2.50
## [586] 2.57 2.76 2.42 2.62 2.39 2.55 2.85 2.93 2.78 2.42 2.47 2.78 2.90 2.77 2.38
## [601] 2.50 2.84 2.52 2.42 2.73 2.74 2.36 2.66 2.75 2.62 2.74 2.48 2.35 2.67 2.71
## [616] 2.56 2.53 2.64 2.51 2.36 2.44 2.34 2.40 2.58 2.56 2.55 2.59 2.43 2.45 2.53
## [631] 2.38 2.37 2.45 2.48 2.40 2.39 2.43 2.37 2.45 2.33 2.40
## [1] "Cantidad de valores atípicos"
## [1] 641

```

Precipitaciones



```

## [1] "Valores atípicos"
## [1] 0.5 1.0 2.5 0.1 0.1 0.2 0.3 0.7 2.5 1.6 0.3 0.4 1.6 1.1 0.1
## [16] 0.1 0.1 6.4 9.5 3.5 0.5 0.3 0.2 0.2 1.0 0.9 0.5 0.3 0.2 0.2
## [31] 0.2 3.3 1.4 1.5 0.1 0.4 0.4 3.7 4.5 9.5 9.0 2.0 0.1 0.4 0.1

```

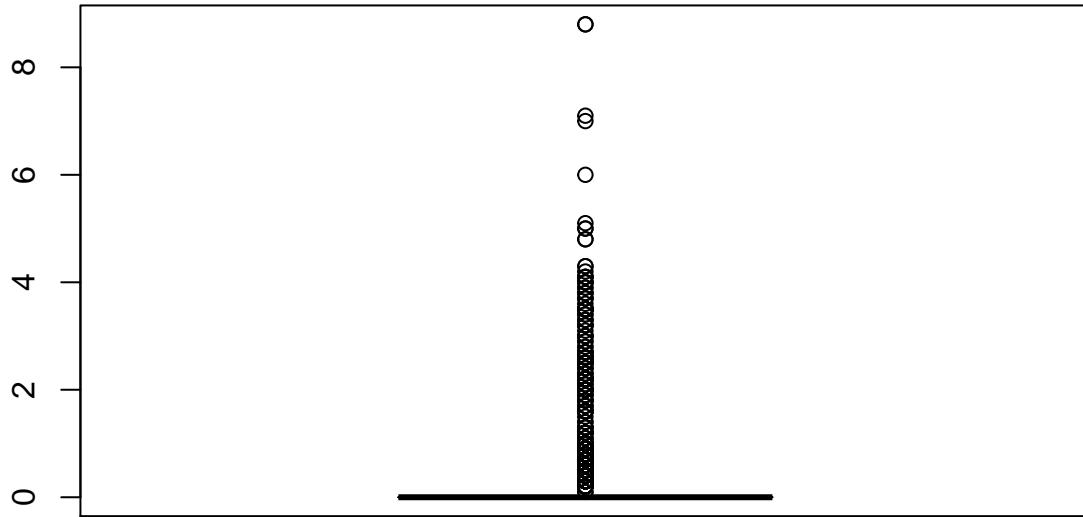
```

## [46] 0.2 0.2 2.0 9.0 1.5 2.0 0.5 2.0 1.5 0.5 8.0 17.0 2.0 0.5 0.3
## [61] 0.2 0.5 0.5 0.5 0.1 3.5 7.0 3.0 3.0 2.0 0.1 0.4 1.5 2.0 0.5
## [76] 0.5 1.0 1.0 1.5 1.0 0.5 0.5 0.5 1.5 2.0 2.0 0.5 0.1 0.1 0.1
## [91] 1.9 1.0 0.5 1.5 3.0 0.2 0.8 1.0 2.0 1.0 1.5 1.5 0.5 0.5 0.3
## [106] 0.2 1.0 0.5 0.5 1.5 2.5 4.0 3.0 2.0 0.1 0.9 0.5 3.5 2.5 7.0
## [121] 6.0 5.0 8.0 4.0 1.5 5.5 2.5 4.0 1.5 1.0 0.5 1.0 1.5 1.0 1.5
## [136] 0.1 2.4 0.5 0.4 0.1 1.0 3.5 4.0 0.5 0.5 1.0 1.0 0.1 0.4 0.5
## [151] 0.5 1.5 1.5 2.0 3.0 2.5 2.0 1.0 1.5 3.5 3.0 0.4 0.1 1.0 3.0
## [166] 2.5 3.5 2.0 3.0 2.0 3.0 1.5 1.5 2.0 2.0 1.0 0.5 1.5 1.0 0.5
## [181] 0.2 0.1 0.2 0.2 0.8 2.0 4.0 1.0 35.0 0.5 0.5 0.5 0.5 0.5 1.0
## [196] 2.5 19.0 13.5 10.0 2.0 2.5 1.5 0.5 2.5 1.5 12.5 4.0 6.0 2.5 1.0
## [211] 1.0 0.5 2.0 1.0 0.5 0.5 0.5 0.5 0.1 0.4 1.0 2.0 1.5 0.5 0.5
## [226] 0.5 2.5 3.5 3.0 3.0 0.5 1.0 0.4 0.1 0.5 0.5 3.5 1.0 1.5 13.5
## [241] 11.5 1.5 1.0 0.1 5.4 10.5 0.5 0.1 0.4 1.5 3.5 6.5 9.5 16.0 5.0
## [256] 14.5 2.5 4.5 1.0 2.0 1.0 1.5 0.5 1.5 0.1 0.4 0.5 0.1 4.9 21.5
## [271] 0.1 0.2 0.1 0.1 1.5 4.5 0.1 6.4 3.5 0.5 4.0 0.1 0.4 0.5 1.0
## [286] 1.5 12.0 18.5 7.5 3.5 1.0 5.5 1.0 1.5 3.5 5.5 6.0 3.5 3.5 5.0
## [301] 2.5 3.0 2.0 4.5 4.0 0.5 1.0 0.5 1.5 3.0 1.0 8.0 0.5 1.5 24.0
## [316] 1.0 2.5 0.5 0.5 0.1 0.5 0.2 0.3 0.5 0.5 1.0 1.0 0.5 0.5 1.0
## [331] 1.5 0.5 1.0 5.5 8.5 6.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5
## [346] 0.5 0.5 0.2 7.3 0.4 1.6 0.5 4.0 2.0 0.4 1.1 0.5 1.0 4.0 1.0
## [361] 0.1 0.9 0.1 0.4 1.0 1.5 2.0 0.1 0.1 0.1 0.1 0.1 1.5 0.5 0.5
## [376] 0.2 0.8 1.0 1.0 0.5 0.5 1.5 0.5 1.5 0.1 0.1 1.3 0.5 1.5 1.0
## [391] 1.5 3.5 3.0 1.0 2.0 0.5 0.5 2.5 0.5 1.0 0.5 0.5 2.0 9.5 4.0
## [406] 0.5 2.5 2.5 18.5 7.0 15.5 29.5 0.5 3.5 1.0 0.5 1.5 1.5 6.5 21.0
## [421] 5.5 4.5 1.0 0.5 6.5 1.5 0.5 0.5 0.1 2.0 0.5 3.5 13.0 13.0 1.0
## [436] 0.5 0.5 0.5 1.0 0.5 0.5 0.1 0.1 0.2 0.3 0.5 1.0 0.1 0.1 0.1
## [451] 0.5 1.5 0.5 1.0 0.5 0.5 0.5 1.5 2.5 0.5 1.5 5.0 4.0 1.0 1.0
## [466] 0.5 0.5 6.5 0.3 1.2 1.5 1.5 2.0 4.0 2.5 4.0 6.0 5.5 1.0 1.5
## [481] 1.0 1.5 1.5 0.5 1.0 1.0 1.5 2.0 2.5 4.5 8.5 9.5 9.0 6.0 6.0
## [496] 3.5 1.0 0.5 4.5 2.5 5.0 0.5 0.5 1.0 1.0 1.5 5.5 1.0 1.5 2.5
## [511] 1.5 0.5 1.0 0.5 0.5 1.5 2.0 5.5 9.5 9.0 18.0 18.0 0.5 0.5 0.2
## [526] 1.0 9.1 1.8

## [1] "Cantidad de valores atípicos"
## [1] 528

```

Nevada



```
## [1] "Valores atípicos"
## [1] 0.1 0.3 0.4 0.4 0.4 0.4 0.4 0.3 0.2 0.2 0.2 0.2 0.3 1.0 1.0 1.0 1.0 1.0 1.0 0.9
## [19] 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.8 0.8 0.7 0.6 0.3 0.7 2.7 4.0 4.1 4.3 4.3
## [37] 3.9 3.1 2.1 1.4 1.0 0.5 0.4 0.2 0.5 0.7 0.8 1.3 1.8 2.0 4.0 4.8 4.8 5.1
## [55] 5.0 4.2 3.2 2.7 2.7 2.7 2.7 2.6 2.6 2.6 2.6 2.6 2.6 2.6 2.6 2.6 2.6 2.6 2.6
## [73] 2.6 2.5 2.5 2.5 2.3 2.2 2.2 2.1 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0
## [91] 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 2.0 1.9 1.9 1.7 1.7 1.7 1.6 1.6 2.2 2.2
## [109] 2.2 2.4 2.4 2.3 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.1 2.0 1.9
## [127] 1.8 1.6 1.0 0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.7 0.7 0.7 0.7
## [145] 0.7 0.7 0.6 0.5 0.2 0.6 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.1 1.1 1.0 0.7
## [163] 0.5 1.0 1.0 1.0 1.0 1.0 1.0 0.9 0.9 0.9 0.9 0.8 0.7 0.6 0.5 0.5 0.4
## [181] 0.3 0.2 0.2 0.5 0.8 0.8 0.8 0.8 0.8 0.8 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.4 0.3
## [199] 0.3 0.2 0.2 0.4 0.6 0.6 0.4 0.2 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
## [217] 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 0.9 0.9 0.9 0.9 0.9
## [235] 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.8 0.7
## [253] 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.6 0.6
## [271] 0.6 0.6 0.6 0.6 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5
## [289] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.4 0.4 0.4 0.3 0.3 0.3 0.3 0.3 0.3 0.3
## [307] 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.2 0.6
## [325] 2.0 3.0 3.3 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.5 3.4 3.2 3.0
## [343] 3.0 2.8 2.5 2.2 1.8 1.8 1.6 1.6 1.6 1.6 1.6 1.6 1.6 1.6 1.6 1.6 1.6 1.6 1.6 1.6
## [361] 1.6 1.6 1.6 1.6 1.5 1.4 1.3 1.0 0.5 0.1 0.2 0.3 0.3 0.3 0.3 0.3 0.2 0.2
## [379] 0.4 4.0 4.1 4.1 3.9 3.8 3.8 3.8 3.7 3.7 3.7 3.7 3.4 1.7 0.6 1.3 4.0 7.1 8.8
## [397] 8.8 7.0 6.0 5.0 4.1 3.6 3.5 3.5 3.3 3.3 3.2 3.2 3.0 3.0 2.9 2.9 2.8 2.6
## [415] 2.5 2.5 2.5 2.5 2.4 2.3 2.2 1.8 1.3 1.1 0.8 0.4 0.4 0.4 0.4 0.4 0.4
## [433] 0.4 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.2
```

```

## [1] "Cantidad de valores atípicos"
## [1] 443

print(paste('El porcentaje de outliers es', 1930/8760*100, '%'))

```

```

## [1] "El porcentaje de outliers es 22.0319634703196 %"

```

El porcentaje de outliers representa el 22 del total de datos, un porcentaje medio alto dependiendo del análisis futuro decidiré si considerar estos outliers o no, por el momento se mantendrán.

5. Comprobación de la normalidad

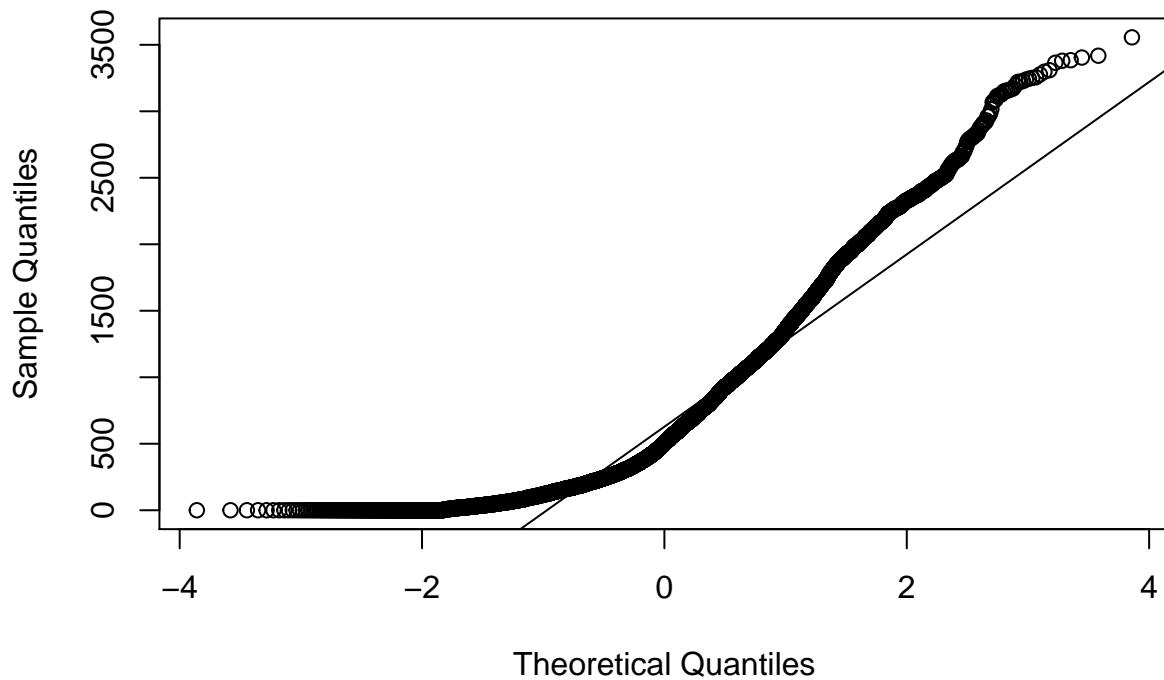
```

#Verifico normalidad
for (i in 1:(ncol(data))){
  if (is.numeric(data[,i])){
    print(colnames(data)[i])
    skewness(data[,i]) #Para tratar de ser un poco más exacto voy a utilizar un método numérico para el
    print(agostino.test(data[,i])) #D'Agostino skewness test
    qqnorm(data[,i])
    qqline(data[,i])
  }
}

## [1] "Recuento_de_bicicletas_alquiladas"
## 
## D'Agostino skewness test
## 
## data: data[, i]
## skew = 1.1532, z = 35.7712, p-value < 2.2e-16
## alternative hypothesis: data have a skewness

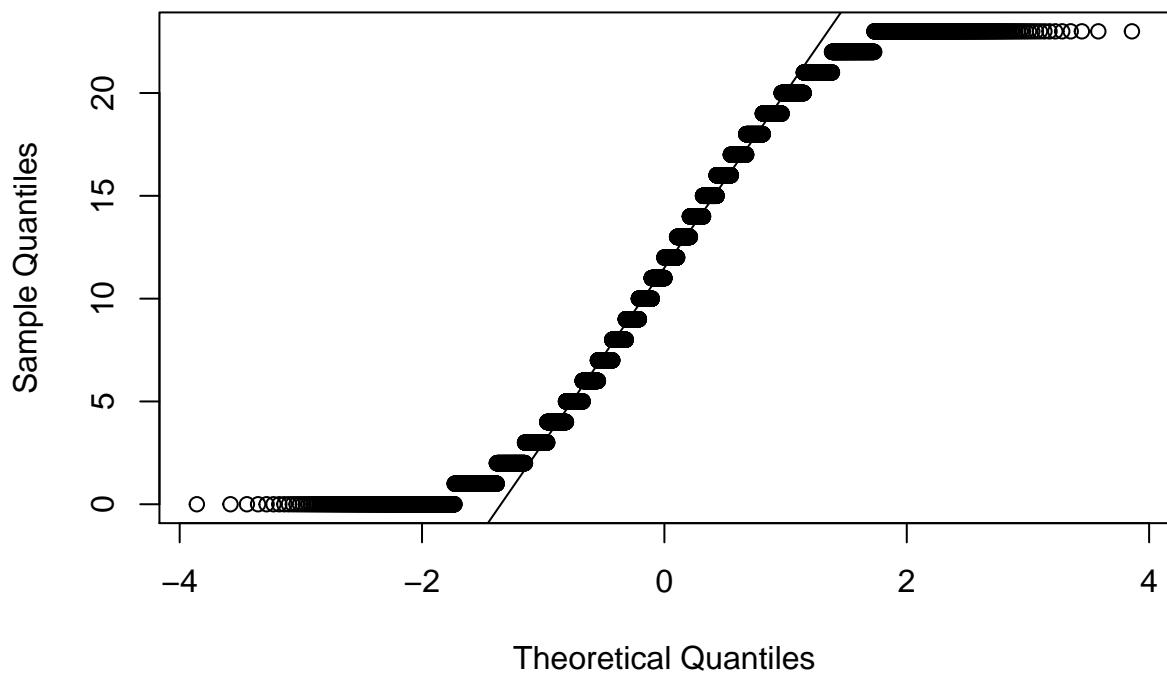
```

Normal Q-Q Plot



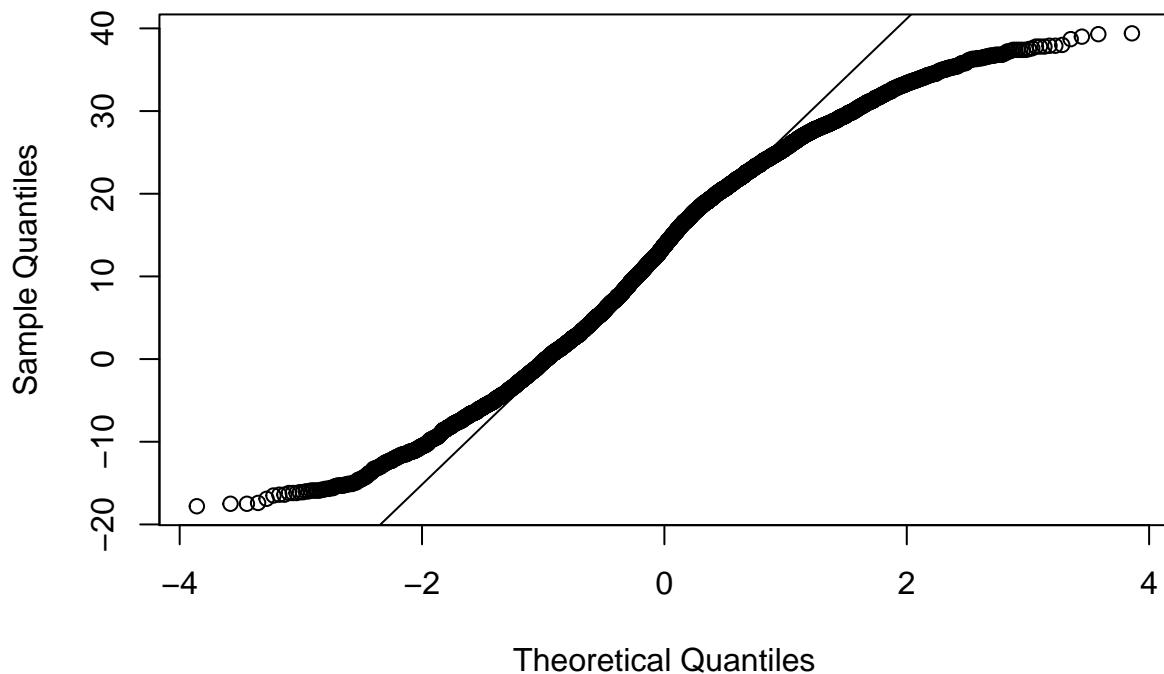
```
## [1] "Hora"
##
## D'Agostino skewness test
##
## data: data[, i]
## skew = 0, z = 0, p-value = 1
## alternative hypothesis: data have a skewness
```

Normal Q-Q Plot

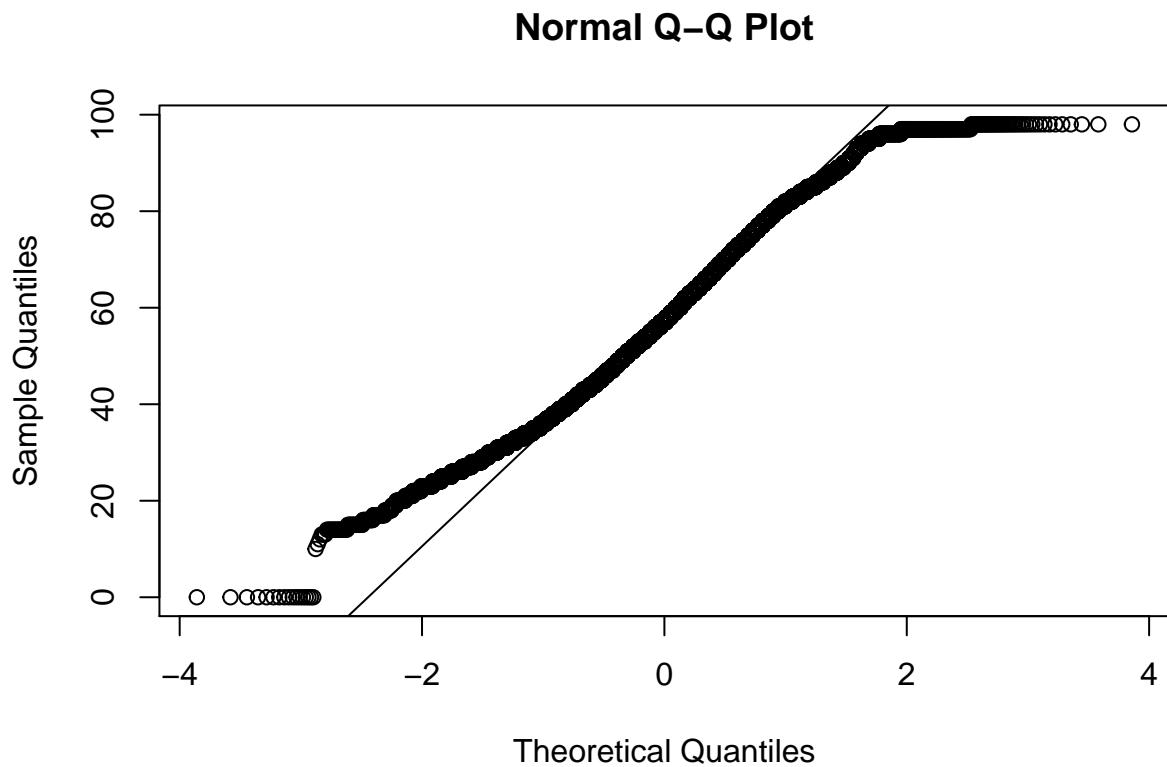


```
## [1] "Temperatura"
##
## D'Agostino skewness test
##
## data: data[, i]
## skew = -0.19829, z = -7.51074, p-value = 5.884e-14
## alternative hypothesis: data have a skewness
```

Normal Q-Q Plot

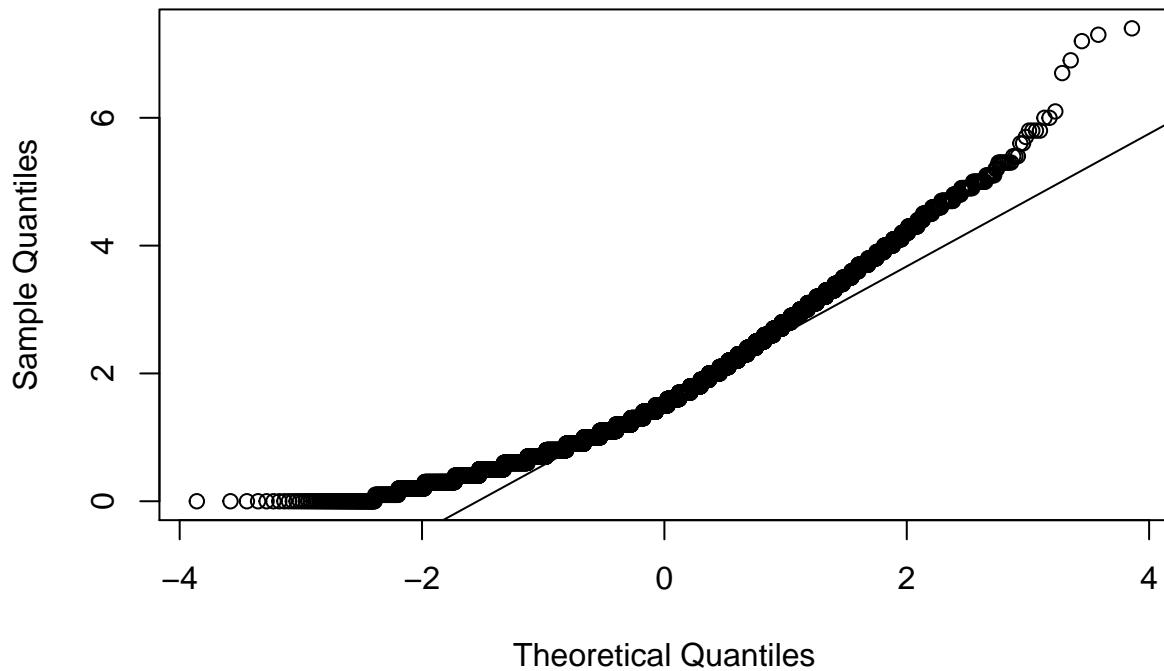


```
## [1] "Humedad_porcentaje"
##
## D'Agostino skewness test
##
## data: data[, i]
## skew = 0.059569, z = 2.276053, p-value = 0.02284
## alternative hypothesis: data have a skewness
```

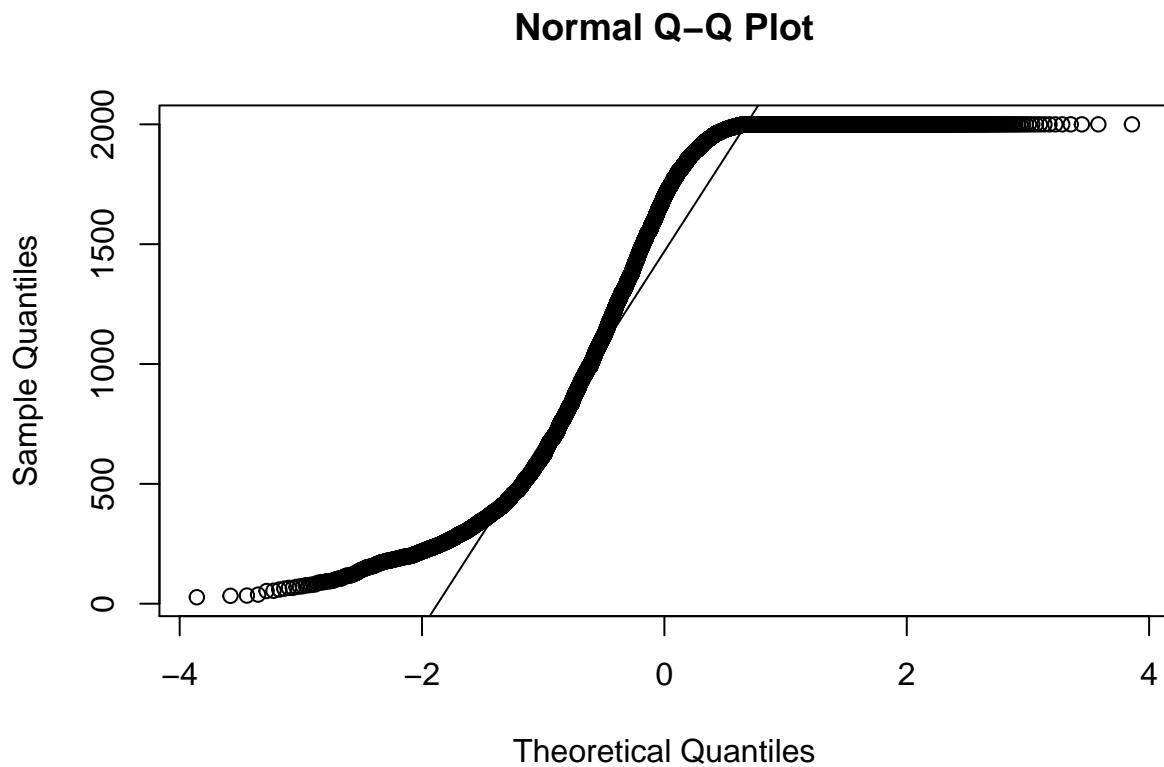


```
## [1] "Velocidad_viento"
##
## D'Agostino skewness test
##
## data: data[, i]
## skew = 0.8908, z = 29.4935, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Normal Q-Q Plot

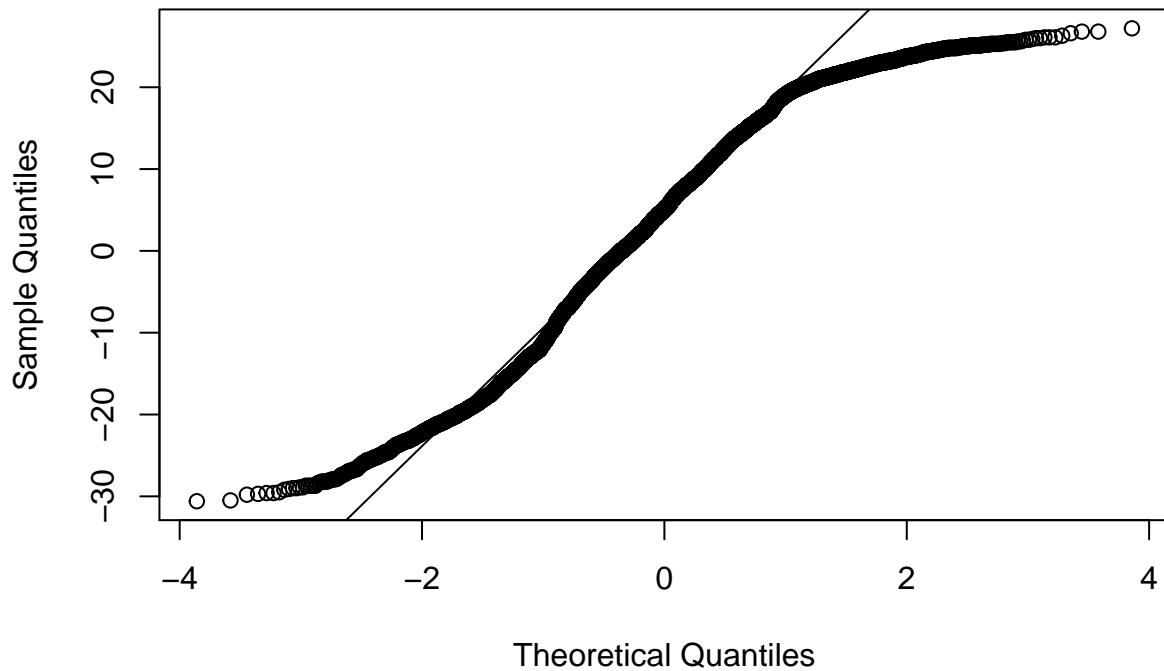


```
## [1] "Visibilidad"
##
## D'Agostino skewness test
##
## data: data[, i]
## skew = -0.70167, z = -24.30914, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```



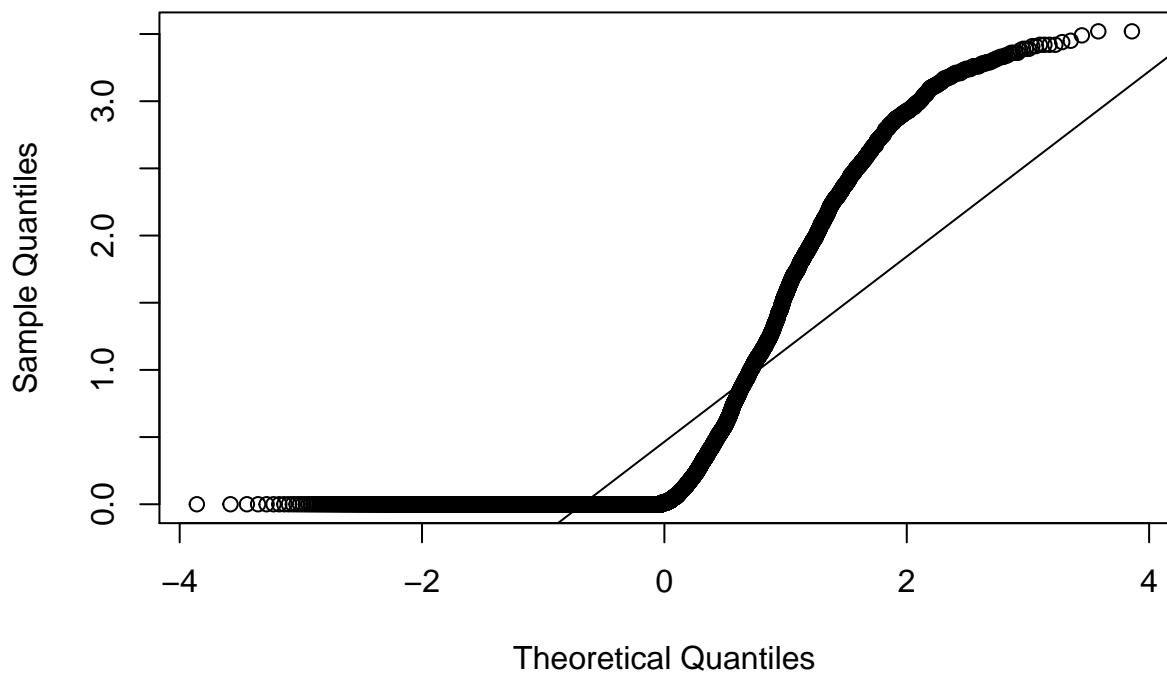
```
## [1] "Temperatura_punto_rocio"
##
## D'Agostino skewness test
##
## data: data[, i]
## skew = -0.36724, z = -13.61010, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Normal Q-Q Plot



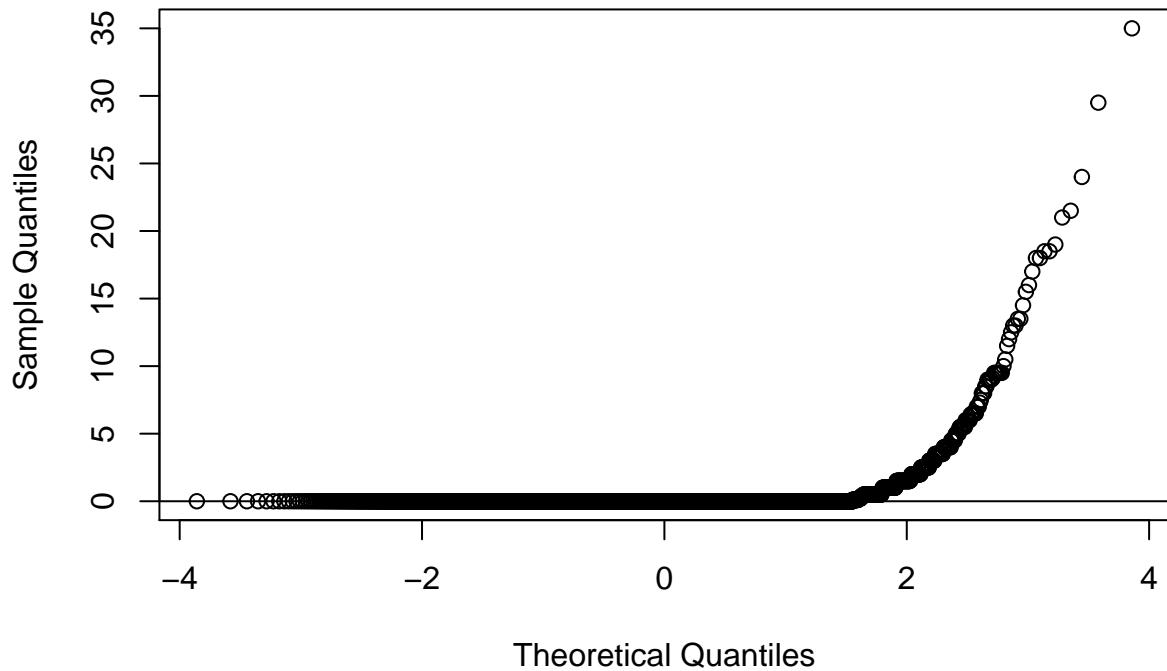
```
## [1] "Radiación_solar"
##
## D'Agostino skewness test
##
## data: data[, i]
## skew = 1.5038, z = 42.8147, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Normal Q-Q Plot



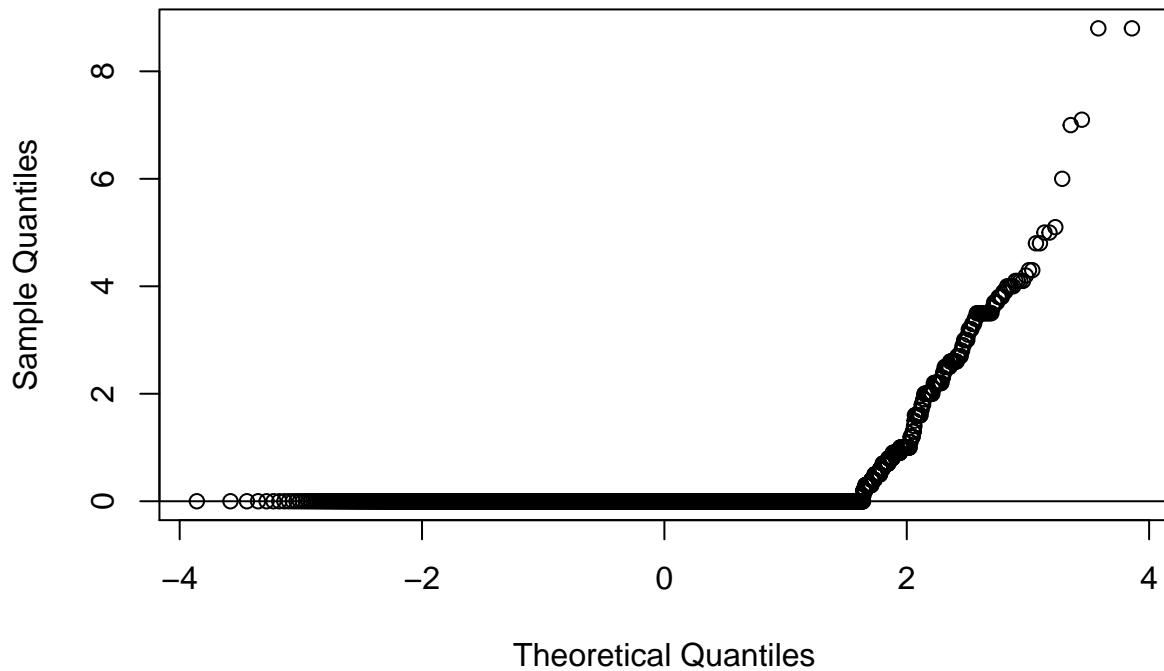
```
## [1] "Precipitaciones"  
##  
## D'Agostino skewness test  
##  
## data: data[, i]  
## skew = 14.531, z = 111.668, p-value < 2.2e-16  
## alternative hypothesis: data have a skewness
```

Normal Q-Q Plot



```
## [1] "Nevada"  
##  
## D'Agostino skewness test  
##  
## data: data[, i]  
## skew = 8.4394, z = 94.7281, p-value < 2.2e-16  
## alternative hypothesis: data have a skewness
```

Normal Q-Q Plot



Segundo método para verificar Normalidad

```
alpha = 0.05
col.names = colnames(data)
for (i in 1:ncol(data)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(data[,i]) | is.numeric(data[,i])) {
    p_val = ad.test(data[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(data) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}

## Variables que no siguen una distribución normal:
## Recuento_de_bicicletas_alquiladas, Hora,
## Temperatura, Humedad_porcentaje, Velocidad_viento,
## Visibilidad, Temperatura_punto_rocio, Radiación_solar,
## Precipitaciones, Nevada,
```

5.1 Análisis de normalidad

Con el método gráfico se puede comprobar que la muestra sigue una distribución normal, los puntos se distribuyen a lo largo de la línea.

Para tratar de ser un poco más exacto voy a utilizar un método numérico para el test de normalidad.

El punto de corte que se suele utilizar es $P = 0.05$:

1. Recuento_de_bicicletas_alquiladas $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
2. Hora $p_{valor} > 0.05$, p-value es mayor que 0.05, acepto la hipótesis nula, es decir: la distribución es normal
3. Temperatura $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
4. Humedad_porcentaje $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
5. Velocidad_viento $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
6. Visibilidad_punto $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
7. Temperatura_punto_rocio $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
8. Radiación_solar $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
9. Precipitaciones $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal
10. Nevada $p_{valor} < 0.05$, p-value es menor que 0.05, rechazo la hipótesis nula, es decir: la distribución no es normal

6. Comprobación de la homogeneidad de la varianza

```
fligner.test(Recuento_de_bicicletas_alquiladas ~ Hora, data = data)
```

```
##  
## Fligner-Killeen test of homogeneity of variances  
##  
## data: Recuento_de_bicicletas_alquiladas by Hora  
## Fligner-Killeen:med chi-squared = 3517.7, df = 23, p-value < 2.2e-16
```

```
fligner.test(Recuento_de_bicicletas_alquiladas ~ Temperatura, data = data)
```

```
##  
## Fligner-Killeen test of homogeneity of variances  
##  
## data: Recuento_de_bicicletas_alquiladas by Temperatura  
## Fligner-Killeen:med chi-squared = 2622.8, df = 545, p-value < 2.2e-16
```

```
fligner.test(Recuento_de_bicicletas_alquiladas ~ Humedad_porcentaje, data = data)
```

```

## 
## Fligner-Killeen test of homogeneity of variances
## 
## data: Recuento_de_bicicletas_alquiladas by Humedad_porcentaje
## Fligner-Killeen:med chi-squared = 853.15, df = 89, p-value < 2.2e-16

fligner.test(Recuento_de_bicicletas_alquiladas ~ Velocidad_viento, data = data)

## 
## Fligner-Killeen test of homogeneity of variances
## 
## data: Recuento_de_bicicletas_alquiladas by Velocidad_viento
## Fligner-Killeen:med chi-squared = 582.17, df = 64, p-value < 2.2e-16

fligner.test(Recuento_de_bicicletas_alquiladas ~ Visibilidad, data = data)

## 
## Fligner-Killeen test of homogeneity of variances
## 
## data: Recuento_de_bicicletas_alquiladas by Visibilidad
## Fligner-Killeen:med chi-squared = 2344.9, df = 1788, p-value < 2.2e-16

fligner.test(Recuento_de_bicicletas_alquiladas ~ Temperatura_punto_rocio, data = data)

## 
## Fligner-Killeen test of homogeneity of variances
## 
## data: Recuento_de_bicicletas_alquiladas by Temperatura_punto_rocio
## Fligner-Killeen:med chi-squared = 1994.3, df = 555, p-value < 2.2e-16

fligner.test(Recuento_de_bicicletas_alquiladas ~ Radiacion_solar, data = data)

## 
## Fligner-Killeen test of homogeneity of variances
## 
## data: Recuento_de_bicicletas_alquiladas by Radiacion_solar
## Fligner-Killeen:med chi-squared = 897.2, df = 344, p-value < 2.2e-16

fligner.test(Recuento_de_bicicletas_alquiladas ~ Precipitaciones, data = data)

## 
## Fligner-Killeen test of homogeneity of variances
## 
## data: Recuento_de_bicicletas_alquiladas by Precipitaciones
## Fligner-Killeen:med chi-squared = 567.62, df = 60, p-value < 2.2e-16

fligner.test(Recuento_de_bicicletas_alquiladas ~ Nevada, data = data)

```

```

## 
## Fligner-Killeen test of homogeneity of variances
## 
## data: Recuento_de_bicicletas_alquiladas by Nevada
## Fligner-Killeen:med chi-squared = 541.44, df = 50, p-value < 2.2e-16

```

En ninguno de los casos de comprueba Homogeneidad de la varianza.

Pero podemos concluir revisando los gráficos también:

A la vista del gráfico de residuos en relación a los valores ajustados, no se observa ningún patrón especial, por lo que tanto podemos asumir que se cumple homocedasticidad. Existen algunos valores extremos que pueden alterar la homocedasticidad, pero asumimos que se cumple en general.

Por otro lado el Q_Q plot, muestra que los datos no se ajustan exactamente a una distribución normal para los valores inferiores o superiores de la gráfica, pero sí para los valores centrales. Por tanto, existen algunos valores extremos que pueden afectar a la condición de normalidad.

7. Pruebas Estadísticas

7.1 Correlaciones

Realizo el cálculo y la visualización para comprender la correlación de los atributos en caso de que exista alguna correlación entre ellos.

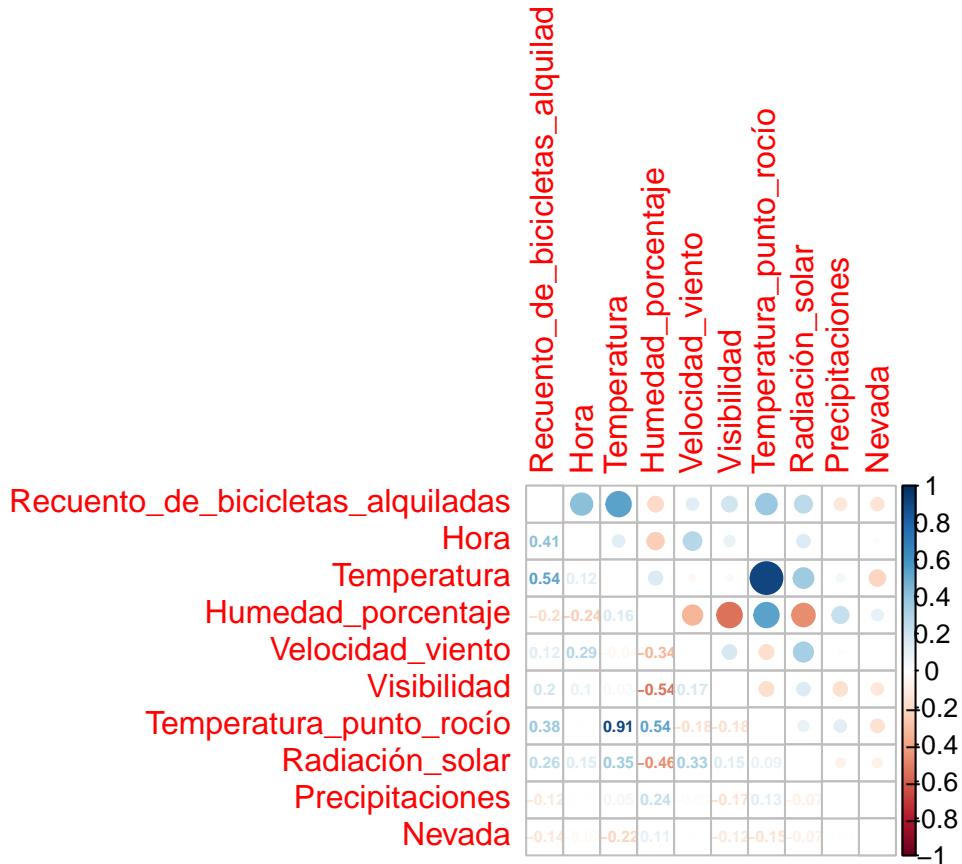
```

corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
data_num <- data.frame(data['Recuento_de_bicicletas_alquiladas'], data['Hora'], data['Temperatura'], data['Velocidad'])

corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
# Calcular el coeficiente de correlación para cada variable cuantitativa
for (i in 1:(ncol(data_num) - 1)) {
  if (is.integer(data_num[,i]) | is.numeric(data_num[,i])) {
    spearman_test = cor.test(data_num[,i],
    data_num[,length(data_num)],
    method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value
    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(data_num)[i]
  }
}

MC <- cor(data_num) #permite ejecutar una matriz de correlación
corrplot.mixed(MC, tl.pos = "lt", number.cex = 0.5) #Visualizo la correlación de los atributos

```



```
print(corr_matrix) #Imprimo los datos de correlación
```

```
##                               estimate      p-value
## Recuento_de_bicicletas_alquiladas -0.220760280 3.698167e-97
## Hora                            -0.031921068 2.808249e-03
## Temperatura                      -0.307068545 1.248150e-190
## Humedad_porcentaje                0.050255090 2.527225e-06
## Velocidad_viento                  0.029232240 6.215663e-03
## Visibilidad                       -0.074217543 3.526043e-12
## Temperatura_punto_rocio          -0.249373747 2.689327e-124
## Radiacion_solar                  -0.076845148 5.936353e-13
## Precipitaciones                   0.001896111 8.591620e-01
```

Se observa que los atributos en general que tienen una correlación muy baja a excepción entre Temperatura_punto_rocio con Temperatura con un valor de correlación alto 0.91

```
prop.table(table(data$Temporadas)) #Reviso los Valores dentro de las estaciones en la temporada
```

```
##
##      Autumn     Spring    Summer   Winter
## 0.2493151 0.2520548 0.2520548 0.2465753
```

```
#Convierbo las variables discretas en factores (Temporada, Vacaciones, Dia_laboral, Hora)
data$Temporadas=as.factor(data$Temporadas)
data$Vacaciones=as.factor(data$Vacaciones)
data$Dia_funcional=as.factor(data$Dia_laboral)
data$Hora=as.factor(data$Hora)
```

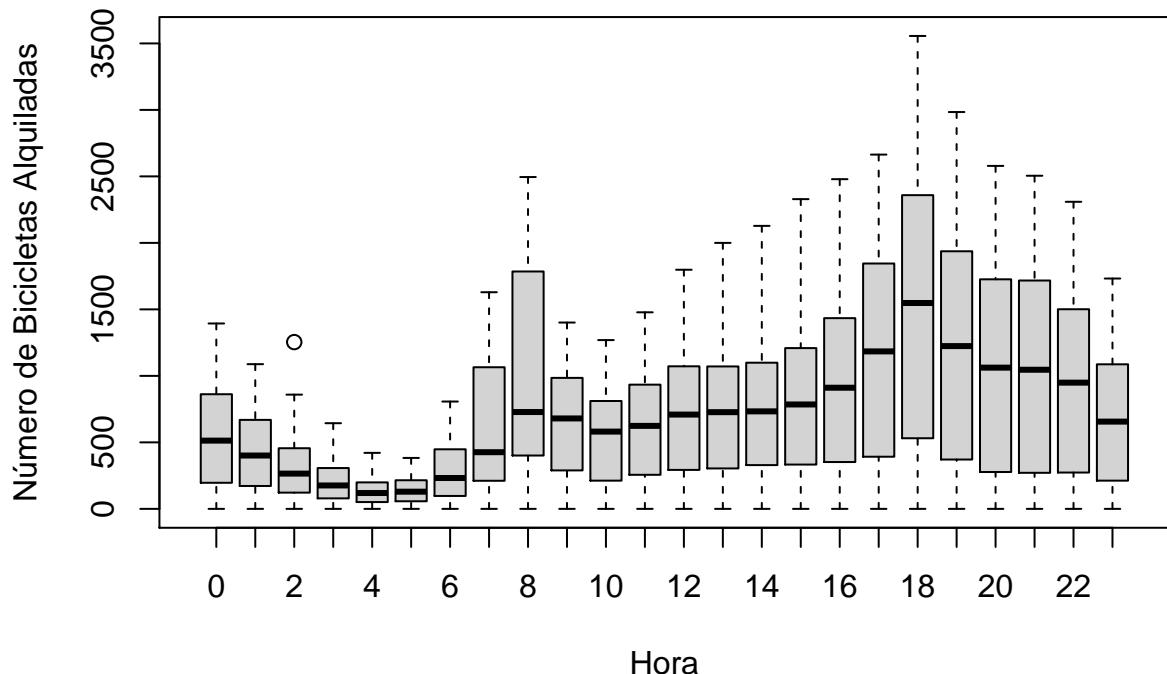
8. Prueba de hipótesis (Análisis Multivariante)

Tendencia Horaria:

Veo la tendencia horaria del recuento durante horas y verifico si la hipótesis es correcta o no.

Separo el conjunto de datos de prueba y entrenamiento.

```
#Demanda de uso compartido de bicicletas, pronosticar la demanda de alquiler de bicicletas
boxplot(data$Recuento_de_bicicletas_alquiladas~data$Hora, xlab="Hora", ylab="Número de Bicicletas Alquiladas")
```



Tendencia de la demanda de bicicletas en horas del día

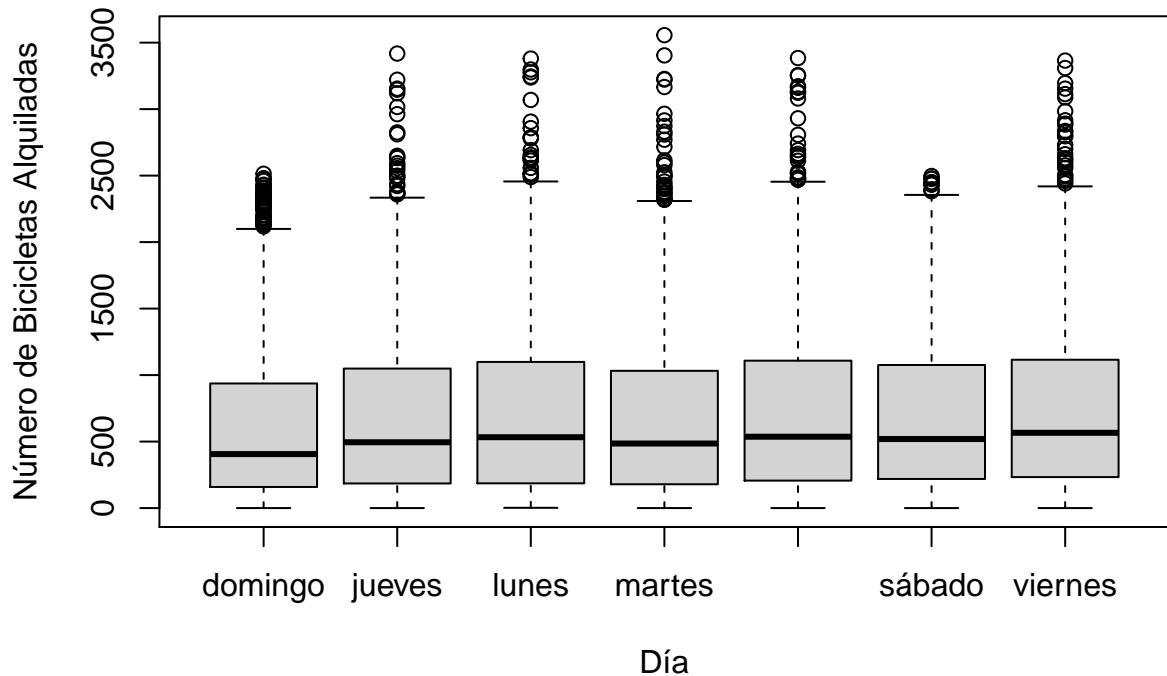
Se pueden observar Rangos en los que hay mayor demanda de bicicletas:

1. En el rango [7-9] horas y de [17-19] horas hay una alta demanda de alquiler de bicicletas.
2. En el rango [0-6] horas y de [20-23] horas hay una baja demanda de alquiler de bicicletas.
3. En el rango de [10-16] horas se mantiene un promedio constante en la demanda.

Tendencia Diaria:

El gráfico muestra la demanda de alquiler de bicicletas durante días de la semana.

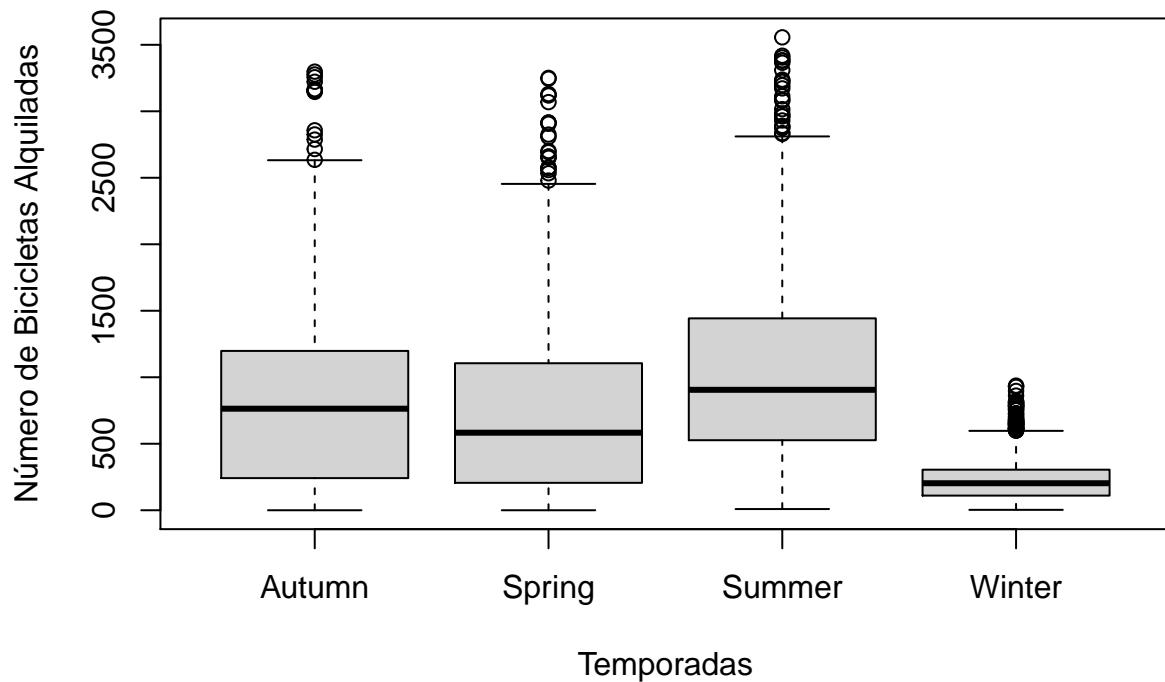
```
#Paso las fechas a nombres de días  
dias<-weekdays(data$Fecha)  
data$dias=dias #Agrego la columna días al Dataframe data  
boxplot(data$Recuento_de_bicicletas_alquiladas~data$dias,xlab="Día", ylab="Número de Bicicletas Alquiladas")
```



Tendencia de la demanda de bicicletas en los días de la semana, prácticamente no varía

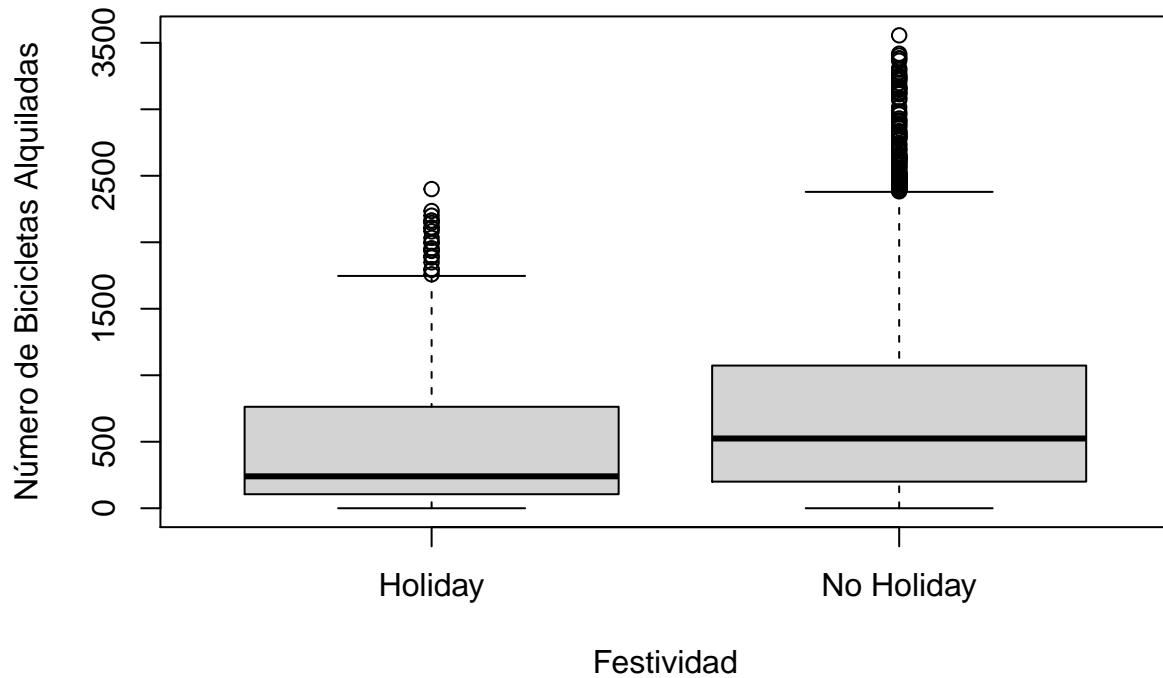
Tendencia Temporada:

```
boxplot(data$Recuento_de_bicicletas_alquiladas~data$Temporadas,xlab="Temporadas", ylab="Número de Bicicletas Alquiladas")
```



Tendencia de la demanda de bicicletas en temporadas, en verano se tiene mayor demanda de Bicicletas
Tendencia Días Festivos:

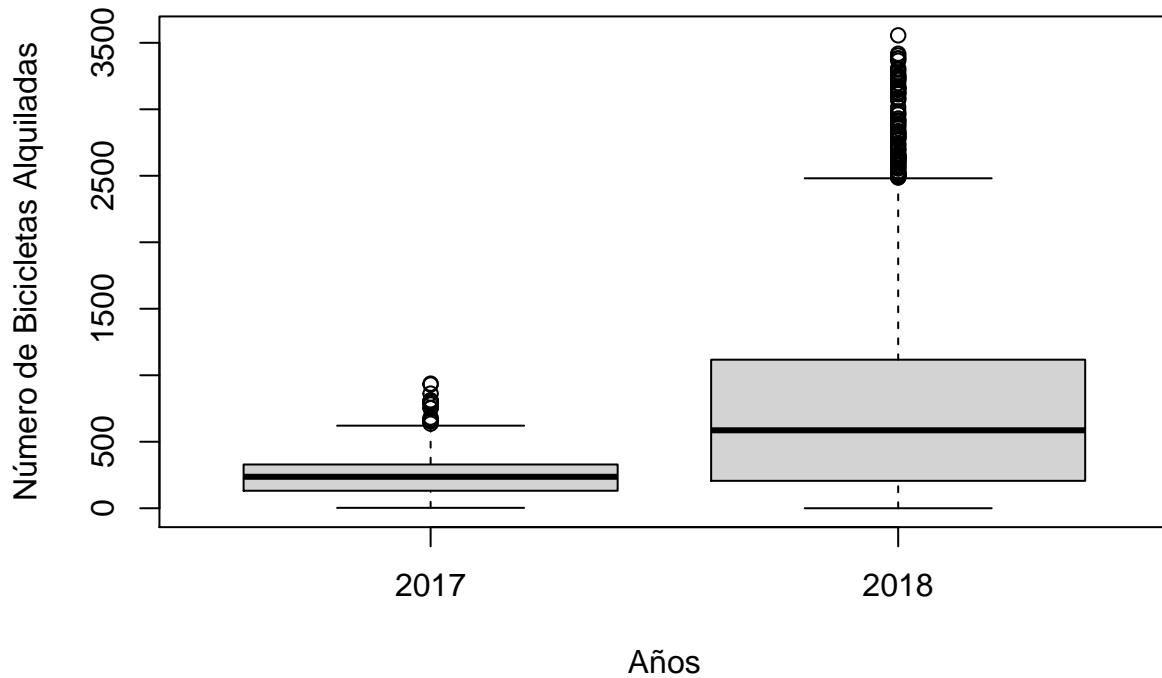
```
boxplot(data$Recuento_de_bicicletas_alquiladas ~ data$Vacaciones, xlab="Festividad", ylab="Número de Bicicletas")
```



Tendencia de la demanda de bicicletas en días festivos, en días no festivos la tendencia de alquiler de bicicletas es mayor

Tendencia Años:

```
data$anio <- format(data$Fecha, "%Y") #Capturo solo el año del formato fecha
data$anio=as.factor(data$anio) #Agrego al data frame un nuevo atributo año como factor
boxplot(data$Recuento_de_bicicletas_alquiladas~data$anio,xlab="Años", ylab="Número de Bicicletas Alquiladas")
```



Tendencia de la demanda de bicicletas por años, en el 2018 aumenta la demanda de alquiler de bicicletas

8.1 Análisis de las relaciones mediante hipótesis

¿Existe alguna relación entre la radiación solar alta (mayor a la media) y la baja (menor a la media) demanda del alquiler de bicicletas?

Para contestar estas preguntas voy a sumir normalidad de los atributos.

H_0 (Hipótesis Nula): La variable Radiación_solar alta (mayor a la media) y el bajo Recuento_de_bicicletas_alquiladas (menor a la media) son independientes

H_1 (Hipótesis Alternativa): Existe una relación de dependencia entre las variables

$\alpha = 0.05$ para un Nivel de Confianza del 95%

```
#Realizo la comparación de los datos con la media y genero nuevas columnas
data$lower.Recuento_de_bicicletas_alquiladas <- (data$Recuento_de_bicicletas_alquiladas < mean(data$Recuento_de_bicicletas_alquiladas))
data$upper.Radiación_solar <- (data$Radiación_solar > mean(data$Radiación_solar))
#Codifico los valores como 0 y 1.
data$lower.Recuento_de_bicicletas_alquiladas <- ifelse( data$lower.Recuento_de_bicicletas_alquiladas==TRUE, 1, 0)
data$upper.Radiación_solar <- ifelse( data$upper.Radiación_solar==TRUE, 1, 0)
#Utilizo la función R chisq.test
table <- table( data$lower.Recuento_de_bicicletas_alquiladas, data$upper.Radiación_solar)
print(table) #Imprimo en pantalla la tabla
```

```

0     1
0 1774 1748
1 4233 1005

```

```
chisq.test(table,correct=FALSE) #Valores de la función chisq
```

```

Pearson's Chi-squared test

data: table
X-squared = 905.74, df = 1, p-value < 2.2e-16

pvalor = 5.54593372106752e - 199
pvalor < α

```

Dado que p_{valor} es $>$ que α , rechazo H_0 . Rechazamos la hipótesis nula y concluyo que existe relación entre la radiación alta y la baja demanda de alquiler de bicicletas.

8.2 Modelo de regresión lineal univariante

Modelo lineal: Recuento_de_bicicletas_alquiladas ~ Precipitaciones

Se estima por mínimos cuadrados ordinarios un modelo lineal que explique la variable Recuento_de_bicicletas_alquiladas en función de las Precipitaciones

```
Model.8.2<- lm(Recuento_de_bicicletas_alquiladas~Precipitaciones, data=data )
summary(Model.8.2)
```

```

Call:
lm(formula = Recuento_de_bicicletas_alquiladas ~ Precipitaciones,
    data = data)

```

Residuals:

Min	1Q	Median	3Q	Max
-715.1	-506.1	-195.1	353.9	2840.9

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	715.064	6.899	103.66	<2e-16 ***
Precipitaciones	-70.362	6.063	-11.61	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 640.1 on 8758 degrees of freedom

Multiple R-squared: 0.01515, Adjusted R-squared: 0.01503

F-statistic: 134.7 on 1 and 8758 DF, p-value: < 2.2e-16

```
corr = cor(x = Recuento_de_bicicletas_alquiladas, y = Precipitaciones, method = "pearson")
```

A la vista de los resultados, no existe una relación lineal positiva entre ambas variables. Se observa que el coeficiente de determinación ajustado es: $R^2_{ajustado} = 0.0150347477888589$. Es decir, el modelo de regresión lineal explica el 1.5% de la varianza del recuento de bicicletas.

Si se calcula el coeficiente de correlación obtenemos un valor de -0.123074.

8.3 Modelo de regresión lineal múltiple (regresores cuantitativos)

Modelo lineal: Recuento_de_bicicletas_alquiladas ~ Precipitaciones + Visibilidad

Se estima por mínimos cuadrados ordinarios un modelo lineal que explique la variable Recuento_de_bicicletas_alquiladas en función de las Precipitaciones y la Visibilidad. Se procederá a evaluar la bondad de ajuste a través del coeficiente de determinación ajustado y se verá si el modelo mejora.

```
#Ajusto el modelo de regresión múltiple:  
Model.8.3<- lm(Recuento_de_bicicletas_alquiladas~Precipitaciones+Visibilidad, data=data)  
vif(Model.8.3)
```

```
variable      gvif  
1: Precipitaciones 1.028912  
2:     Visibilidad 1.028912  
  
cor(data$Visibilidad, data$Precipitaciones)  
  
[1] -0.1676292
```

```
summary( Model.8.3)
```

```
Call:  
lm(formula = Recuento_de_bicicletas_alquiladas ~ Precipitaciones +  
    Visibilidad, data = data)  
  
Residuals:  
    Min      1Q Median      3Q      Max  
-822.2 -478.4 -174.7  343.1 2870.0  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 432.40144   17.62154  24.538 <2e-16 ***  
Precipitaciones -52.74654    6.04655  -8.723 <2e-16 ***  
Visibilidad     0.19490    0.01121   17.380 <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 629.4 on 8757 degrees of freedom  
Multiple R-squared:  0.04799, Adjusted R-squared:  0.04777  
F-statistic: 220.7 on 2 and 8757 DF, p-value: < 2.2e-16
```

A la vista de los resultados, no existe una relación lineal positiva entre ambas variables. Se observa que el coeficiente de determinación ajustado es: $R^2_{ajustado} = 0.0477681593676631$. Es decir, el modelo de regresión lineal explica el 4.77% de la varianza del recuento de bicicletas. No hay evidencia de mejora del modelo.

8.4 Generación y Comparación del $R^2_{ajustado}$ para elegir el mejor modelo

```
#Ajusto varios modelos de regresión múltiple:
Model.8.4.1<- lm(Recuento_de_bicicletas_alquiladas~Precipitaciones+Visibilidad+Temperatura+Humedad_porcentaje+Velocidad_viento+Radiación_solar+Temporadas+Vacaciones)
Model.8.4.2<- lm(Recuento_de_bicicletas_alquiladas~Hora+Temperatura+Humedad_porcentaje+Velocidad_viento+Radiación_solar+Temporadas+Vacaciones)
Model.8.4.3<- lm(Recuento_de_bicicletas_alquiladas~Velocidad_viento+Visibilidad+Radiación_solar+Temperatura+Humedad_porcentaje+Velocidad_viento+Radiación_solar+Temporadas+Vacaciones)
Model.8.4.4<- lm(Recuento_de_bicicletas_alquiladas~Temperatura_punto_rocio+Humedad_porcentaje+Visibilidad+Radiación_solar+Temporadas+Vacaciones)
Model.8.4.5<- lm(Recuento_de_bicicletas_alquiladas~Nevada+Hora+Temperatura+Precipitaciones+Temporadas+Vacaciones)
```

```
# Tabla con los coeficientes de determinación de cada modelo
tabla.coeficientes <- matrix(c(1, summary(Model.8.4.1)$r.squared, 2, summary(Model.8.4.2)$r.squared, 3, summary(Model.8.4.3)$r.squared, 4, summary(Model.8.4.4)$r.squared, 5, summary(Model.8.4.5)$r.squared), nrow=5, byrow=TRUE)
mejor_modelo <- summary(Model.8.4.2)$r.squared
colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes
```

Modelo	R ²
[1,]	0.4663020
[2,]	0.6619954
[3,]	0.3902831
[4,]	0.3799671
[5,]	0.5747639

8.5 Predicciones del valor con el mejor modelo

El mejor modelo es el segundo modelo con un valor $R^2_{ajustado} = 0.6619954$

```
newdata <- data.frame(
  Hora = '7',
  Temperatura = 15,
  Humedad_porcentaje = 40,
  Velocidad_viento = 2,
  Temperatura_punto_rocio = 10,
  Nevada = 2,
  Temporadas = 'Summer',
  Dia_laboral = 'Yes',
  Precipitaciones = 10,
  Vacaciones = 'No Holiday',
  Radiación_solar = 2.0
)
# Predecir el precio
pred <- predict(Model.8.4.2, newdata)
```

Con los datos para realizar la predicción se alquilarían en total 731.0798361 bicicletas

Exporto los datos del dataset a un archivo csv

```
write.csv(data,file="SeoulBikeData_final.csv",row.names=TRUE)
```

Conclusiones

Se ha logrado realizar el análisis del dataset con datos y resultados muy interesantes. Se han logrado contestar preguntas iniciales así como obtener información adicional que se detalla dentro de todo el análisis con datos detallados, además se han construido varios modelos de regresión lineal para poder obtener una predicción al final eligiendo el que tiene mejor $R^2_{ajustado}$, y se logra realizar la predicción de cuantas bicicletas se alquilarán con los datos que se proporcionan como predicción.