# Statistical exploration of 'Bike Sharing Dataset', Washington D.C. 2011/2012

Gianluca La Malfa

Venice, July 2022

=============================================================

## Introduction

This is the final project of the 'Data & Knowledge' course of the minor in Computer and Data Science presented by Gianluca La Malfa at the Ca'Foscari University of Venice. The course aimed to build knowledge on the use of statistical methods with R.

### Objectives of the project

The objective of this project is to statistically explore the 'Bike Sharing Dataset' dataset using some of the most used R packages such as tidyverse, ggplot2 and kableExtra. These packages will be used to spot trends by calculating statistical indicators and building charts which can help to better understand the dataset. At the end of the first analysis, a regression analysis will be built to better understand the found trends.

### Dataset description:

The dataset represents the daily data of bike sharing in Washington D.C. during the years 2011 and 2012.

Dictionary:

```
- instant: record index
- dteday : date
- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month ( 1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from
  http://dchr.dc.gov/page/holiday-schedule)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
+ weathersit :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered
      clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered
```

### Source:

https://data.world/uci/bike-sharing-dataset

=============================================

# Analysis

## Setup of the environment

Import packages.

```
library(tidyverse) # analyse data
```

```
## Warning: il pacchetto 'tidyverse' è stato creato con R versione 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.9
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.1.3
```

```
## Warning: il pacchetto 'dplyr' è stato creato con R versione 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2) # visulise data
library(kableExtra) # make tables
```

```
## Warning: il pacchetto 'kableExtra' è stato creato con R versione 4.1.3
```

```
##
## Caricamento pacchetto: 'kableExtra'
## Il seguente oggetto è mascherato da ’package:dplyr’:
##
##     group_rows
```

```
library(ggridges) # make chart with gradient areas
```

```
## Warning: il pacchetto ’ggridges’ è stato creato con R versione 4.1.3
```

```
library(zoo) # change date format
```

```
## Warning: il pacchetto ’zoo’ è stato creato con R versione 4.1.3
```

```
##
## Caricamento pacchetto: ’zoo’
```

```
## I seguenti oggetti sono mascherati da ’package:base’:
##
##     as.Date, as.Date.numeric
```

```
library(ggpmisc) # polynomial regression
```

```
## Warning: il pacchetto ’ggpmisc’ è stato creato con R versione 4.1.3
```

```
## Caricamento del pacchetto richiesto: ggpp
```

```
## Warning: il pacchetto ’ggpp’ è stato creato con R versione 4.1.3
```

```
##
## Caricamento pacchetto: ’ggpp’
```

```
## Il seguente oggetto è mascherato da ’package:ggplot2’:
##
##     annotate
```

```
## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "packedMatrix" of class "replValueSp"; definition not updated
```

```
## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "packedMatrix" of class "mMatrix"; definition not updated
```

Upload the file.

```
day <- read.csv(
  'C:/Users/user/Documents/RStudio repository/uci-bike-sharing-dataset/day.csv'
  )
```

## First explorations and manipulations

Observe the first rows of the dataset to better understand how it is structured.

```
head(day)
```

```
##   instant     dteday season yr mnth holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1       0       6          0          2
## 2       2 2011-01-02      1  0    1       0       0          0          2
## 3       3 2011-01-03      1  0    1       0       1          1          1
## 4       4 2011-01-04      1  0    1       0       2          1          1
## 5       5 2011-01-05      1  0    1       0       3          1          1
## 6       6 2011-01-06      1  0    1       0       4          1          1
##       temp    atemp      hum windspeed casual registered  cnt
## 1 0.344167 0.363625 0.805833 0.1604460    331        654  985
## 2 0.363478 0.353739 0.696087 0.2485390    131        670  801
## 3 0.196364 0.189405 0.437273 0.2483090    120       1229 1349
## 4 0.200000 0.212122 0.590435 0.1602960    108       1454 1562
## 5 0.226957 0.229270 0.436957 0.1869000     82       1518 1600
## 6 0.204348 0.233209 0.518261 0.0895652     88       1518 1606
```

Change the date variable ‘dteday’ format from string to date, and denormalize the temperature variable ‘temp’ to a Celsius unit of measurement.

```
bsh <- day %>%
  mutate(dteday = as.Date(dteday, format="%Y-%m-%d"), temp = temp*41)

#bsh$dteday <- as.Date(bsh$dteday, format="%Y-%m-%d")
```

Calculate some statistics to better understand the distribution of the variables included in the dataset.

```
sommario <- bsh %>%
  select(dteday, holiday, weathersit, temp, atemp, hum,  windspeed, casual, registered,
         cnt)
summary(sommario)
```

```
##      dteday              holiday          weathersit         temp
##  Min.   :2011-01-01   Min.   :0.00000   Min.   :1.000   Min.   : 2.424
##  1st Qu.:2011-07-02   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:13.820
##  Median :2012-01-01   Median :0.00000   Median :1.000   Median :20.432
##  Mean   :2012-01-01   Mean   :0.02873   Mean   :1.395   Mean   :20.311
##  3rd Qu.:2012-07-01   3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:26.872
##  Max.   :2012-12-31   Max.   :1.00000   Max.   :3.000   Max.   :35.328
##      atemp              hum           windspeed           casual
##  Min.   :0.07907   Min.   :0.0000   Min.   :0.02239   Min.   :   2.0
##  1st Qu.:0.33784   1st Qu.:0.5200   1st Qu.:0.13495   1st Qu.: 315.5
##  Median :0.48673   Median :0.6267   Median :0.18097   Median : 713.0
##  Mean   :0.47435   Mean   :0.6279   Mean   :0.19049   Mean   : 848.2
##  3rd Qu.:0.60860   3rd Qu.:0.7302   3rd Qu.:0.23321   3rd Qu.:1096.0
##  Max.   :0.84090   Max.   :0.9725   Max.   :0.50746   Max.   :3410.0
##    registered        cnt
##  Min.   :  20   Min.   :  22
##  1st Qu.:2497   1st Qu.:3152
##  Median :3662   Median :4548
##  Mean   :3656   Mean   :4504
##  3rd Qu.:4776   3rd Qu.:5956
##  Max.   :6946   Max.   :8714
```

Create a table more specific table with statistics about variables of interest.

```r
# Create a dataframe with statistics per field
tot<-  summarise(bsh,
    Mean = round(mean(cnt, na.rm=T), 0),
    Variance = round(var(cnt, na.rm = T), 0),
    StdDev = round(sd(cnt, na.rm = T), 0),
    CV = round(StdDev/Mean, 2),
    IQR = round(IQR(cnt, na.rm = T), 0)
    )

reg<-  summarise(bsh,
    Mean = round(mean(registered, na.rm=T), 0),
    Variance = round(var(registered, na.rm = T), 0),
    StdDev = round(sd(registered, na.rm = T), 0),
    CV = round(StdDev/Mean, 2),
    IQR = round(IQR(registered, na.rm = T), 0)
  )

nreg <-  summarise(bsh,
    Mean = round(mean(casual, na.rm=T), 0),
    Variance = round(var(casual, na.rm = T), 0),
    StdDev = round(sd(casual, na.rm = T), 0),
    CV = round(StdDev/Mean, 2),
    IQR = round(IQR(casual, na.rm = T), 0)
  )

tempa<-  summarise(bsh,
    Mean = round(mean(temp, na.rm=T), 2),
    Variance = round(var(temp, na.rm = T),  2),
    StdDev = round(sd(temp, na.rm = T), 2),
    CV = round(StdDev/Mean, 2),
    IQR = round(IQR(temp, na.rm = T), 2)
  )

# Unite the data frames
newtab <- bind_rows('Total'=tot,
          'Registered'=reg,
          'Unregistered'=nreg,
          'Temperature'=tempa,
          .id= "")

# Add scaling colour to data frame for table
newtab[1:3,2:6]<-lapply(newtab[1:3,2:6], function(x) {
  cell_spec(x, color = spec_color(x, end = 0.9))
  })

# Create table
kbl(newtab, booktabs = T, escape = F, align = "c", caption = "<b>Table 1.</b>
    Distribution and variability of rented bikes by user category and temperature.", digits = 2) %>%
kable_styling(bootstrap_options = "hover", full_width = F, position = "left") %>%
column_spec(1, background = "#D3D3D3")#, bold=T)
```

The unregistered users' coefficient of variation is double that of registered users. Probably because registered

Table 1: <b>Table 1.</b> Distribution and variability of rented bikes by user category and temperature.

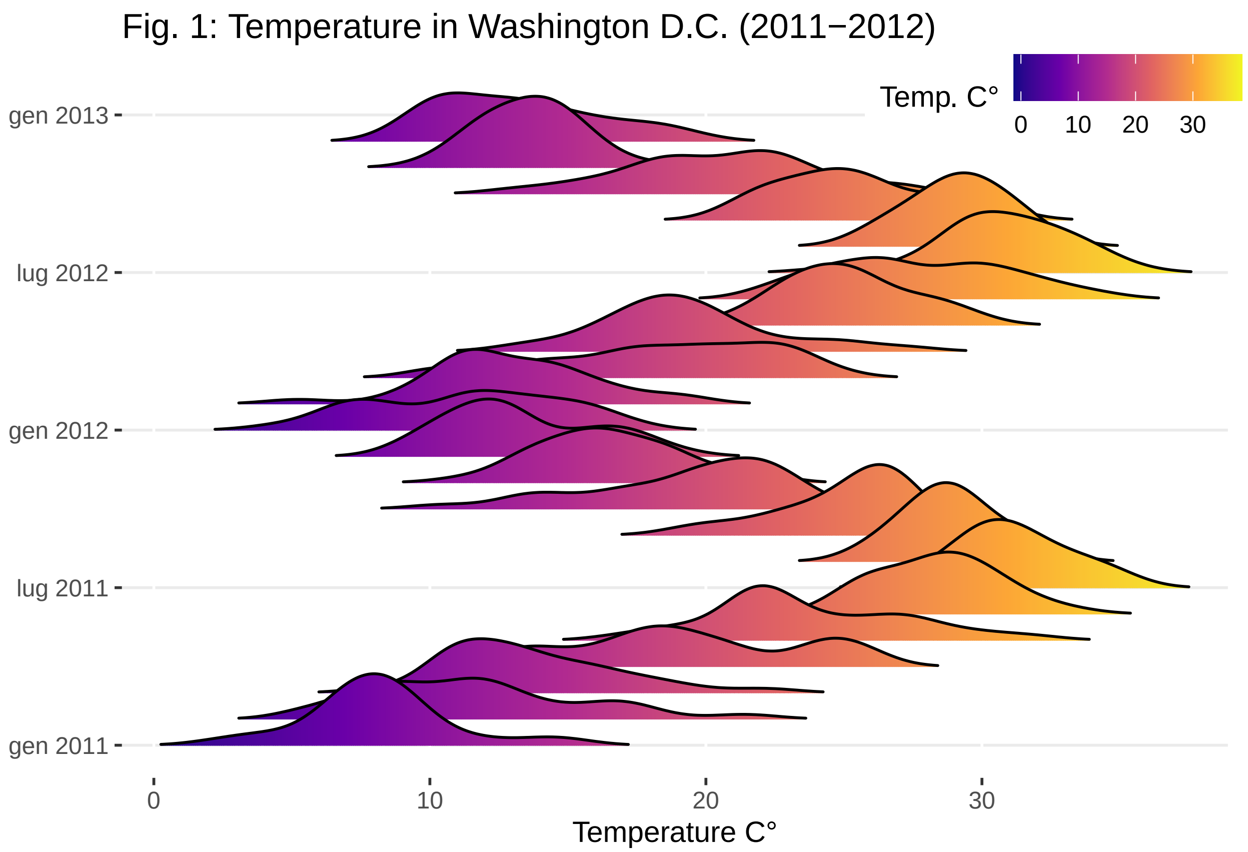
| | Mean | Variance | StdDev | CV | IQR |
|---|---|---|---|---|---|
| Total | 4504 | 3752788 | 1937 | 0.43 | 2804 |
| Registered | 3656 | 2434400 | 1560 | 0.43 | 2280 |
| Unregistered | 848 | 471450 | 687 | 0.81 | 780 |
| Temperature | 20.31 | 56.33 | 7.51 | 0.37 | 13.05 |

users have more incentives to rent more often.

Visualise the distribution of temperature per month.

```
bsh$monthyear <- as.yearmon(bsh$dteday, "%b %Y")

ggplot(bsh, aes(x = temp, y=monthyear, group=monthyear, fill = stat(x))) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_fill_viridis_c(name = "Temp. C°", option = "C") +
  coord_cartesian(clip = "off") +
  labs(x="Temperature C°",
       title = 'Fig. 1: Temperature in Washington D.C. (2011-2012)') +
  theme(axis.title.y = element_blank(),
        panel.background = element_rect(fill = NA, colour = NA),
        panel.grid.major.y = element_line(colour = "grey92"),
        legend.position = c(0.85, 0.95),
        legend.direction="horizontal")
```

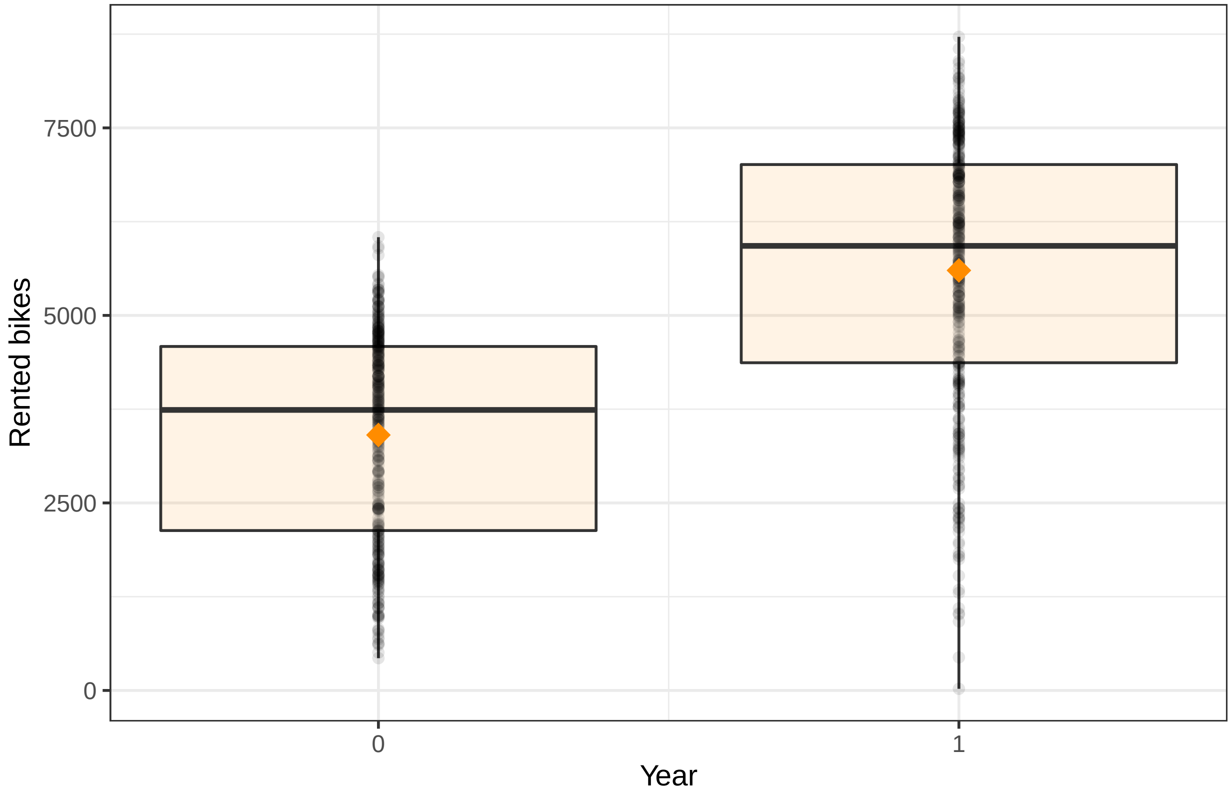
```
## Picking joint bandwidth of 1.19
```



Visualise the distribution of rented bikes per year (0=2011, 1=2012).

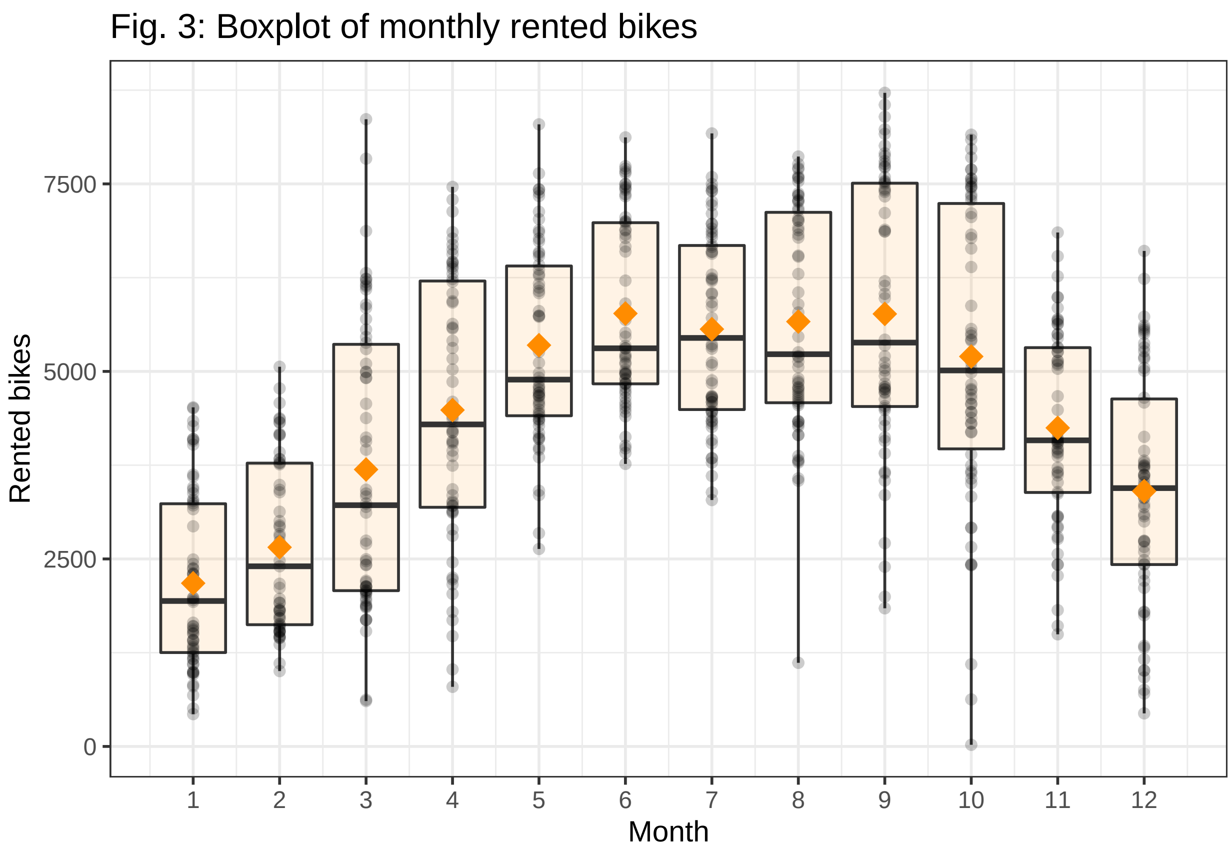
```
ggplot(data = bsh, aes(x = yr, y = cnt, group = yr)) +
  geom_boxplot(fill = "darkorange", alpha=0.1, coef = 10) +
  geom_jitter(width = 0, height = 0, alpha = 0.1, col = "black") +
  stat_summary(fun = "mean", geom = "point", col = "darkorange", pch = 18, size = 4) +
  labs(x="Year", y = "Rented bikes",
       title = "Fig. 2: Boxplot of yearly rented bikes (0=2011, 1=2012)")+
  scale_x_continuous(breaks = 0:2)+
  theme_bw()
```

Fig. 2: Boxplot of yearly rented bikes (0=2011, 1=2012)



Visualise the distribution of rented bikes per month.

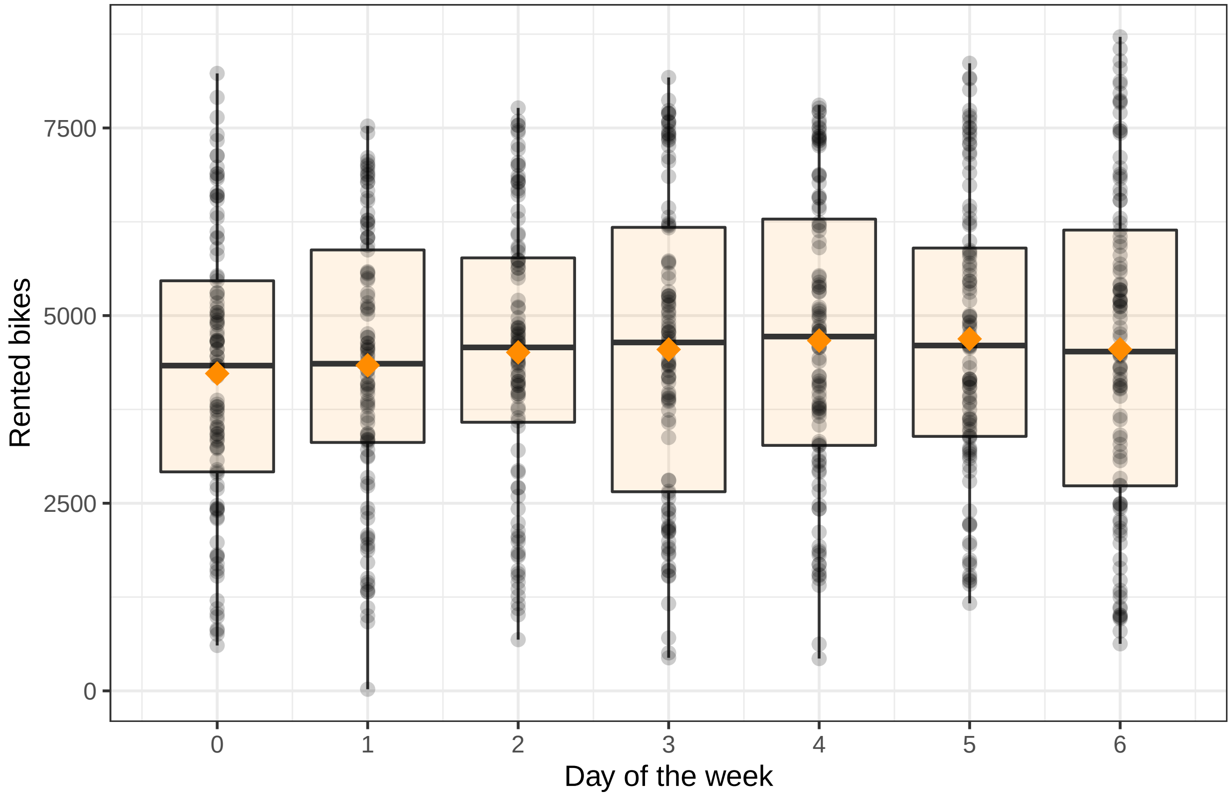
```
ggplot(data = bsh, aes(x = mnth, y = cnt, group = mnth)) +
  geom_boxplot(fill = "darkorange", alpha=0.1, coef = 10) +
  geom_jitter(width = 0, height = 0, alpha = 0.2, col = "black") +
  stat_summary(fun = "mean", geom = "point", col = "darkorange", pch = 18, size = 4) +
  labs(x="Month", y = "Rented bikes",
       title = "Fig. 3: Boxplot of monthly rented bikes")+
  scale_x_continuous(breaks = 1:12)+
  theme_bw()
```



Visualise the distribution of rented bikes per day of the week (0=Monday).

```
ggplot(data = bsh, aes(x = weekday, y = cnt, group = weekday)) +
  geom_boxplot(fill = "darkorange", alpha=0.1, coef = 10) +
  geom_jitter(width = 0, height = 0, alpha = 0.2, col = "black", size=2) +
  stat_summary(fun = "mean", geom = "point", col = "darkorange", pch = 18, size = 4) +
  labs(x = "Day of the week", y = "Rented bikes",
       title = "Fig. 4: Boxplot of rented bikes by day of the week (0=Monday)") +
  scale_x_continuous(breaks = 0:12)+
  theme_bw()
```

Fig. 4: Boxplot of rented bikes by day of the week (0=Monday)



Calculate and make a table with the average temperature and the share of bikes rented per month by the different user categories.

```
# Create a data frame containing the monthly per cent of totals of variables of interest
bsharing <- bsh %>%
  mutate(total = sum(cnt),
         total_registered = sum(registered),
         total_unregistered = sum(casual)) %>%
  group_by(mnth) %>%
  summarise(aggr_monthly = sum(cnt),
            aggr_monthly_registered = sum(registered),
            aggr_monthly_unregistered = sum(casual),
            mean_temp = mean(temp)) %>%
  mutate(total = sum(aggr_monthly),
         perc.o.t_total = aggr_monthly/total*100,
         total_registered = sum(aggr_monthly_registered),
         perc.o.t_registered = aggr_monthly_registered/total_registered*100,
         total_unregistered = sum(aggr_monthly_unregistered),
         perc.o.t_unregistered = aggr_monthly_unregistered/total_unregistered*100) %>%
  summarise("Month"=mnth,
            "Average temperature"=round(mean_temp,2),
            "Perc. total"=round(perc.o.t_total,2),
            "Perc. registered"=round(perc.o.t_registered,2),
            "Perc. unregistered"=round(perc.o.t_unregistered,2))

bsharing[3:5]<-lapply(bsharing[3:5], function(x) {
```

Table 2: <b>Table 2.</b> Percentage of total rented bikes per user category by month and average temperature.

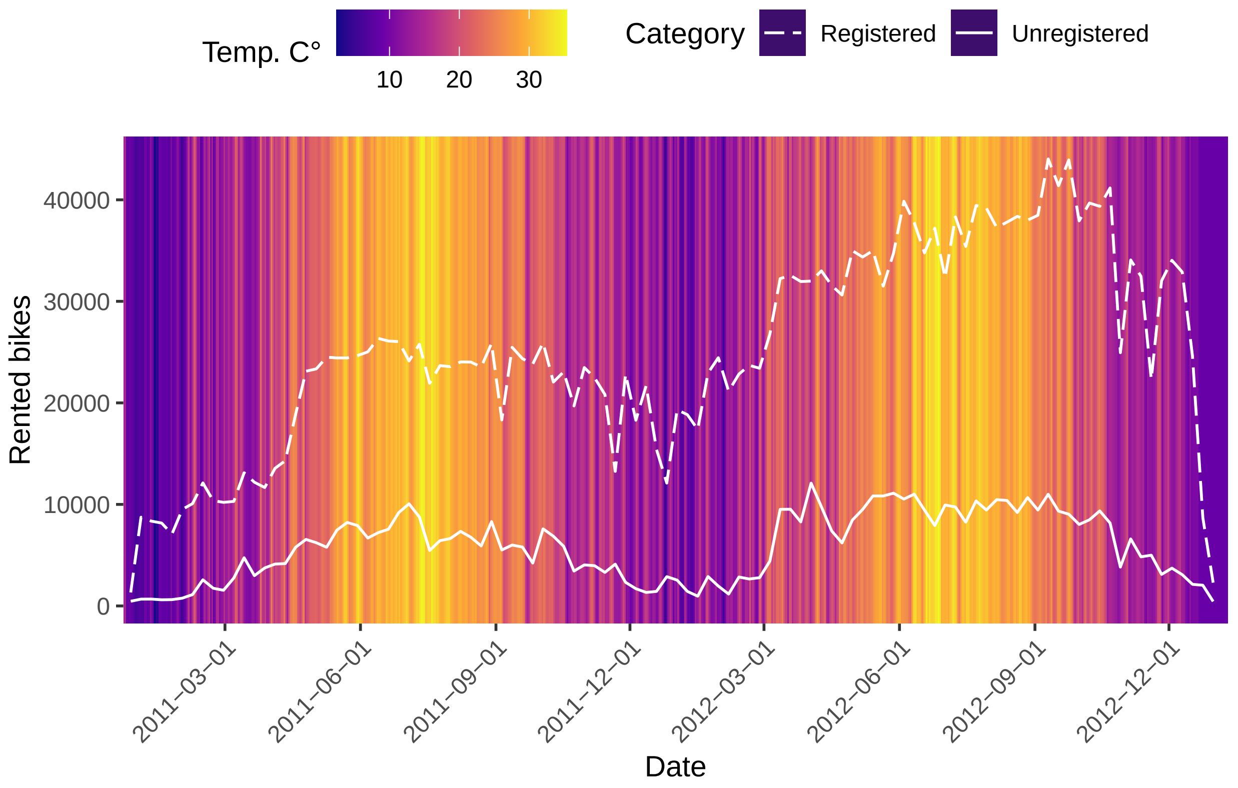
| Month | Average temperature | Perc. total | Perc. registered | Perc. unregistered |
|---|---|---|---|---|
| **1** | 9.69 | 4.1 | 4.6 | 1.94 |
| **2** | 12.27 | 4.6 | 5.1 | 2.41 |
| **3** | 16.01 | 6.95 | 6.9 | 7.17 |
| **4** | 19.27 | 8.17 | 7.79 | 9.81 |
| **5** | 24.39 | 10.07 | 9.59 | 12.14 |
| **6** | 28.05 | 10.52 | 10.19 | 11.92 |
| **7** | 30.97 | 10.48 | 9.98 | 12.61 |
| **8** | 29.05 | 10.67 | 10.44 | 11.62 |
| **9** | 25.28 | 10.51 | 10.31 | 11.34 |
| **10** | 19.89 | 9.79 | 9.83 | 9.64 |
| **11** | 15.14 | 7.74 | 8.17 | 5.9 |
| **12** | 13.29 | 6.41 | 7.08 | 3.5 |

```
  cell_spec(x, color = spec_color(x, end = 0.9))
  })


kable(bsharing, escape = F,
          caption = "<b>Table 2.</b> Percentage of total rented bikes per user category by month and ave
      ) %>%
kable_styling(bootstrap_options = "hover", full_width = F, position = "left") %>%
kable_classic_2(full_width = F) %>%
column_spec(1, background = "#D3D3D3", bold=T) %>%
column_spec(2,
            background = spec_color(bsharing$"Average temperature",
                                    end = 0.9,
                                    option = "A"),
            color="white")
```

Visualise a time series of the weekly average temperature and number of rented bikes by user category.

```
weekly <- bsh %>%
  mutate(Week = as.Date(cut(dteday, breaks = "week")))%>%
  group_by(Week) %>%
  summarise('Unregistered' = sum(casual), 'Registered' = sum(registered)) %>%
  gather(key = "Category", value = "value", -Week)


ggplot(data = weekly) +
  geom_rect(data= bsh, aes(xmin=dteday-10,xmax=dteday+10,ymin=Inf,ymax=-Inf,
      fill=temp)) +
  scale_fill_viridis_c(name = "Temp. C°", option = "C") +
  geom_line(aes(x = Week, y = value, color = Category, linetype = Category)) +
  scale_x_date(date_breaks = "3 month", expand = c(0, 0)) +
  scale_linetype_manual(values=c("longdash", "solid")) +
  scale_color_manual(values=c('white', 'white')) +
  labs(x = "Date",
       y = "Rented bikes",
       title = "Fig. 5: Average temperature and rented bikes per user category by week") +
  theme(axis.text.x = element_text(angle =45, hjust = 1),
        legend.key = element_rect(fill = "#3D0E6C"),
        legend.position = "top")
```

Fig. 5: Average temperature and rented bikes per user category by week

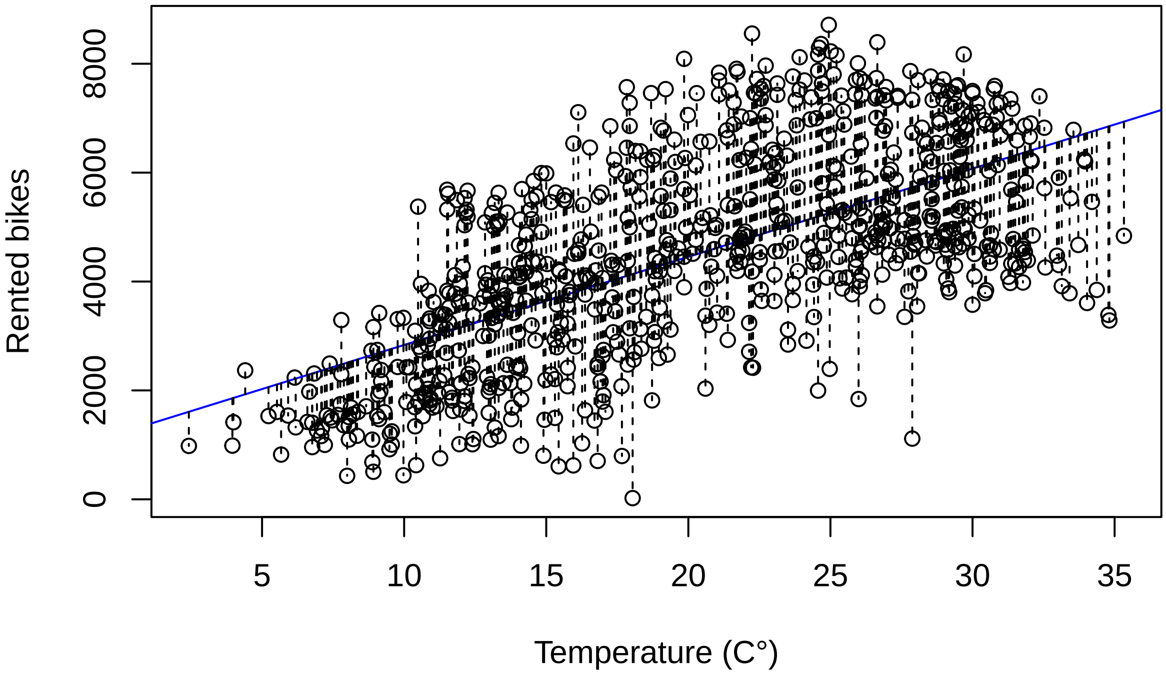


## Regression analysis

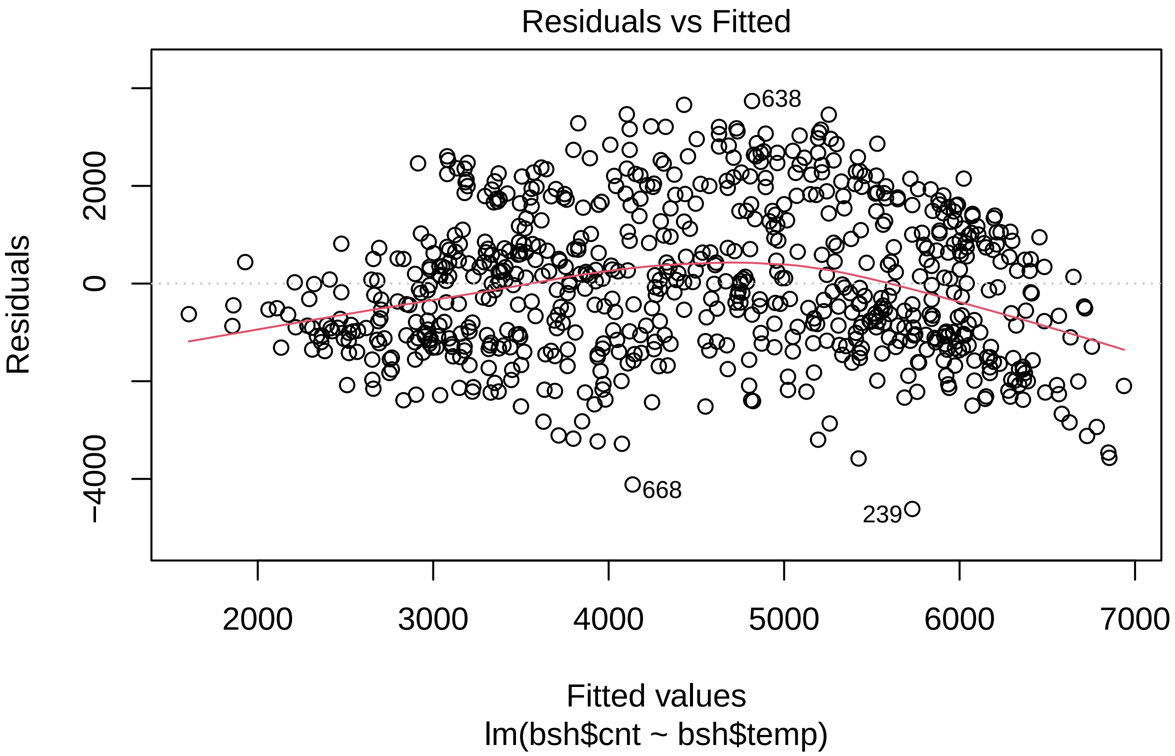
Try a linear model to fit the data.

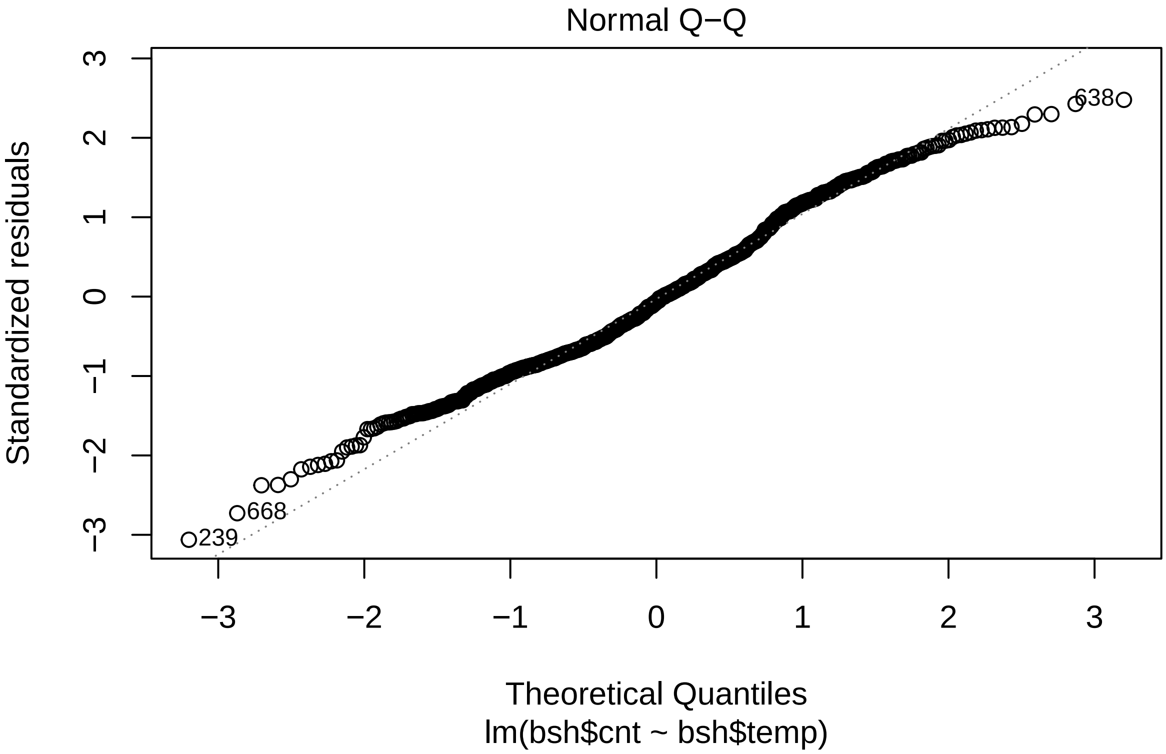
```
plot(bsh$temp, bsh$cnt, xlab= "Temperature (C°)", ylab = "Rented bikes")
retta=lm(bsh$cnt ~ bsh$temp)
abline(retta, col="blue")
segments(bsh$temp, fitted(retta), bsh$temp, bsh$cnt, lty=2)
title(main="Fig. 6: Linear regression with segments")
```

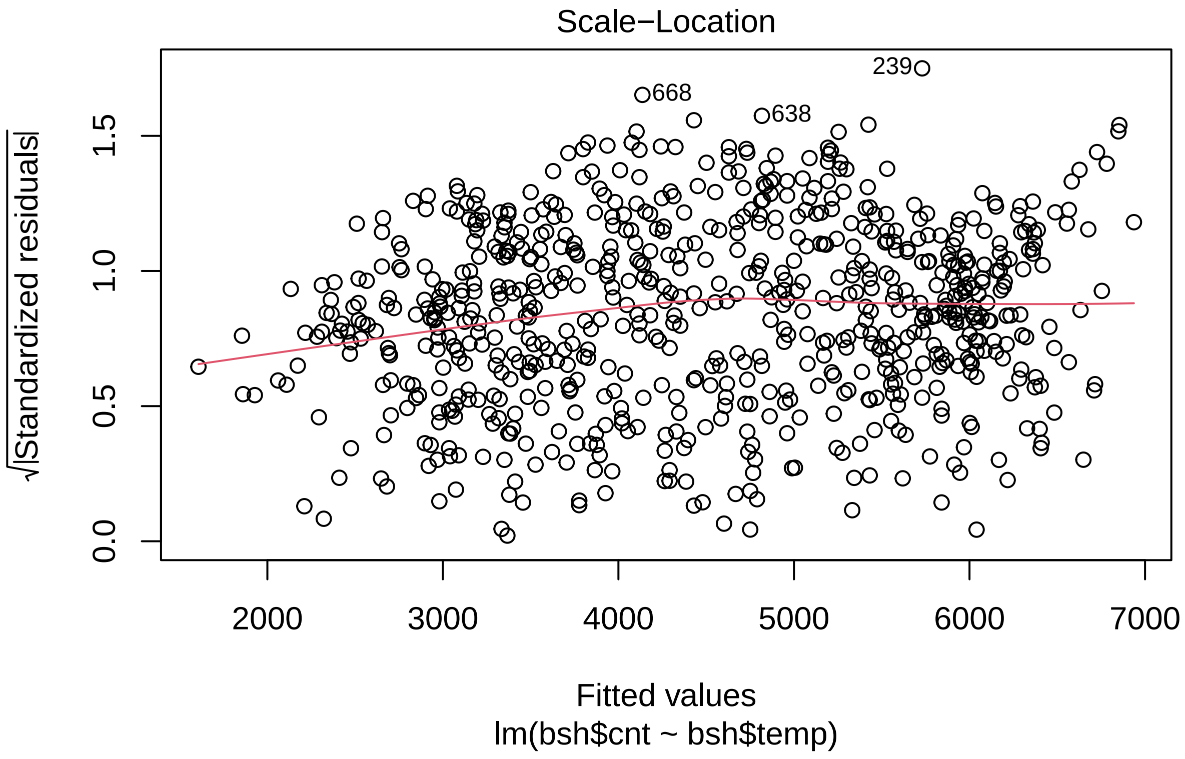
## Fig. 6: Linear regression with segments

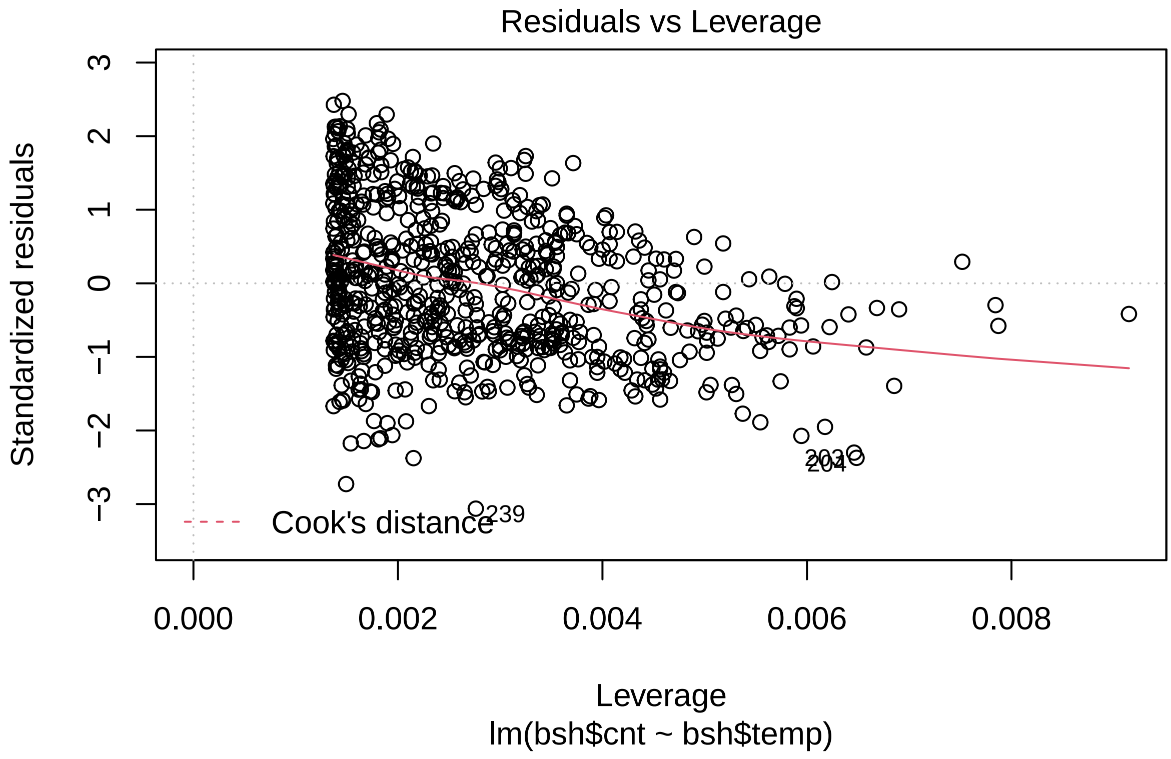


```
plot(retta)
```


Residuals vs Fitted
Residuals
2000
0
−4000
638
668
239
2000
3000
4000
5000
6000
7000
Fitted values
lm(bsh$cnt ~ bsh$temp)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
638
668
239
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(bsh$cnt ~ bsh$temp)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2000
3000
4000
5000
6000
7000
Fitted values
lm(bsh$cnt ~ bsh$temp)
239
668
638

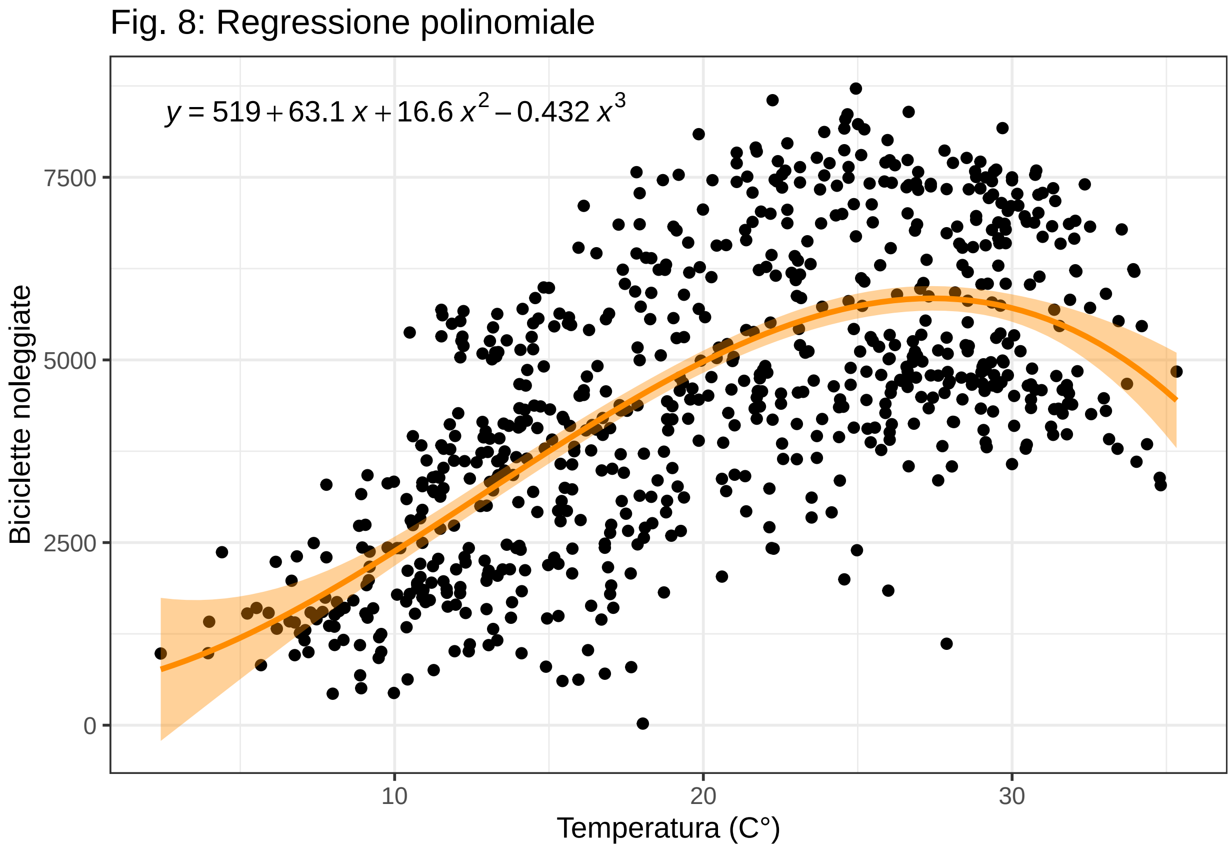


```
summary(retta)
```

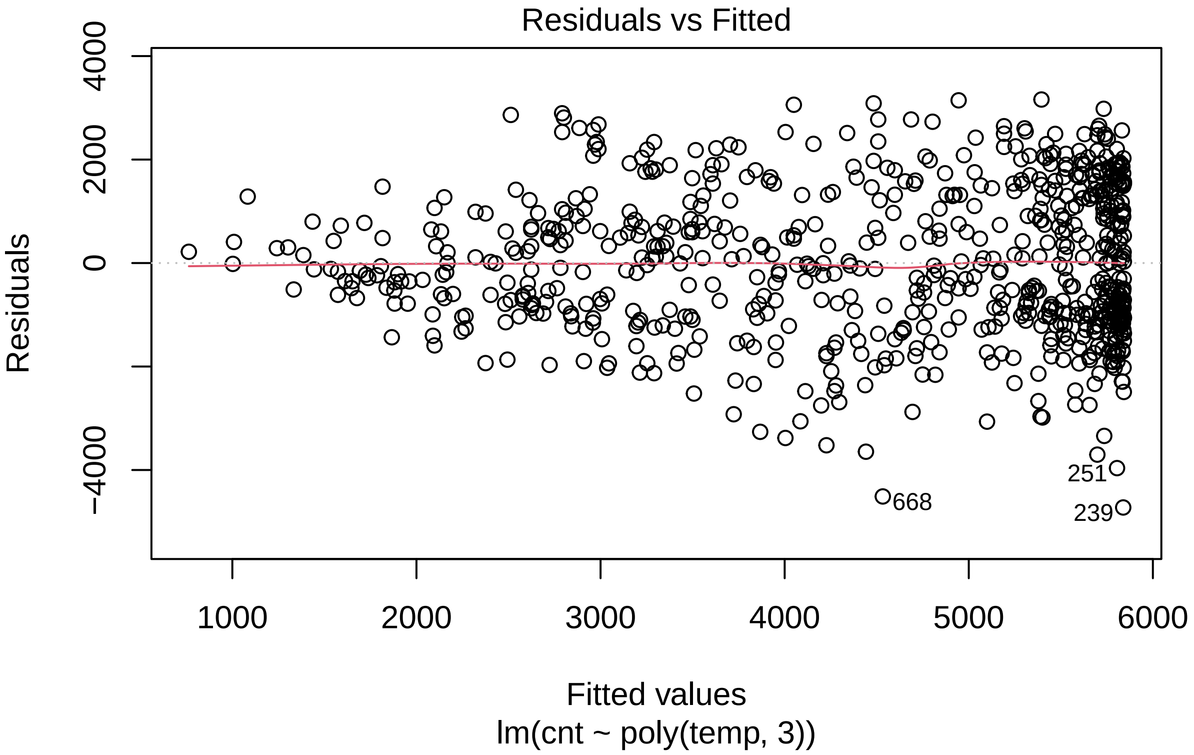
```
##
## Call:
## lm(formula = bsh$cnt ~ bsh$temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4615.3 -1134.9  -104.4  1044.3  3737.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1214.642    161.164   7.537 1.43e-13 ***
## bsh$temp     161.969      7.444  21.759  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1509 on 729 degrees of freedom
## Multiple R-squared:  0.3937, Adjusted R-squared:  0.3929
## F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

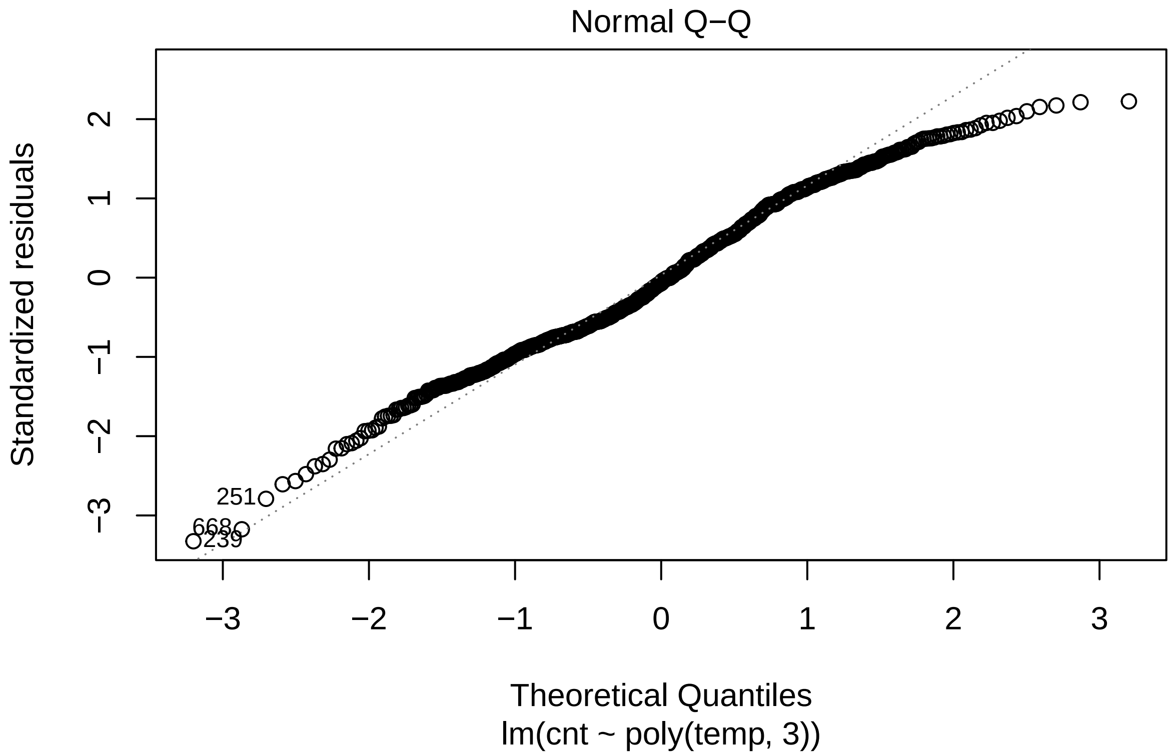
The shape of the scatterplot suggests a polynomial model can better fit the data.

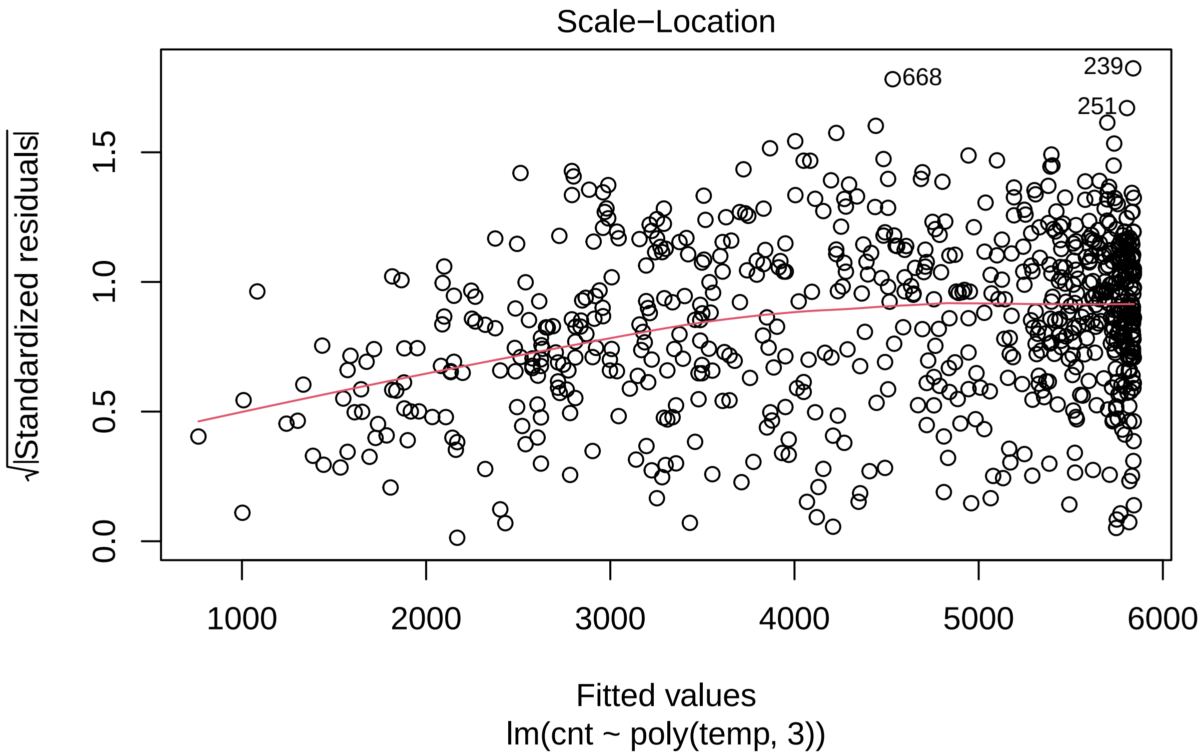
```
ggplot(bsh, aes(temp, cnt)) +
geom_point() +
theme_bw() +
stat_smooth(method = "lm",
            formula = y ~ poly(x, 3),
            color = "darkorange", fill = "darkorange") +
labs(x = "Temperatura (C°)", y = "Biciclette noleggiate",
     title = "Fig. 8: Regressione polinomiale") +
stat_poly_eq(formula = y ~ poly(x, 3, raw = TRUE),
               aes(label = after_stat(eq.label)))
```

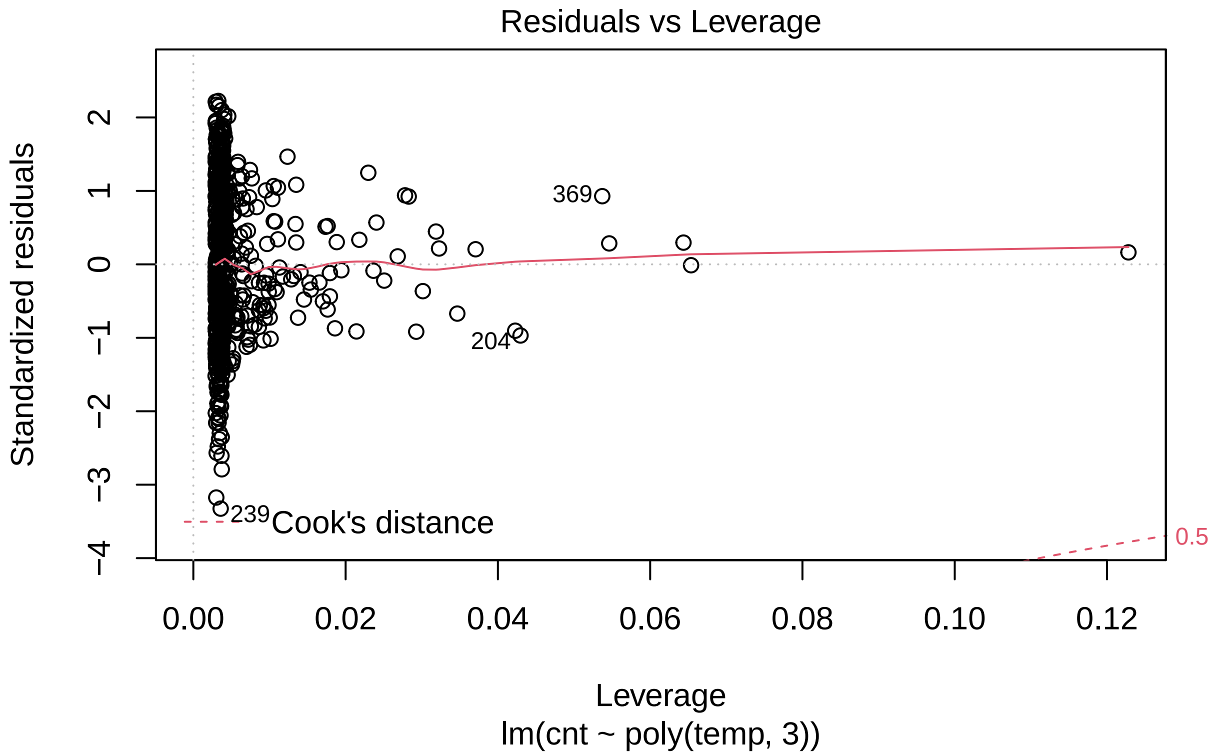


```
poly_mod <- lm(cnt ~ poly(temp, 3),
               data = bsh)

plot(poly_mod)
```


Residuals vs Fitted
Residuals
4000
2000
0
−4000
668
251
239
1000
2000
3000
4000
5000
6000
Fitted values
lm(cnt ~ poly(temp, 3))


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
251
668
239
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(cnt ~ poly(temp, 3))


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1000
2000
3000
4000
5000
6000
Fitted values
lm(cnt ~ poly(temp, 3))
668
239
251



```
summary(poly_mod)
```

```
##
## Call:
## lm(formula = cnt ~ poly(temp, 3), data = bsh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4724.0 -1034.4   -99.6  1130.1  3160.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4504.35      52.63  85.584  < 2e-16 ***
## poly(temp, 3)1  32843.39    1422.98  23.081  < 2e-16 ***
## poly(temp, 3)2 -12759.76    1422.98  -8.967  < 2e-16 ***
## poly(temp, 3)3  -5094.44    1422.98  -3.580 0.000366 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1423 on 727 degrees of freedom
## Multiple R-squared:  0.4627, Adjusted R-squared:  0.4604
## F-statistic: 208.6 on 3 and 727 DF,  p-value: < 2.2e-16
```

# Conclusions

In this project have been used some of the most important R packages to make a statistical analysis of the ‘Bike Sharing Dataset’. It has been explored with descriptive statistics and the use of visualisations. Some trends and features inherent to the dataset have shown up and it has been decided to study the relationship between temperature and daily rented bikes. To do it two regression models have been built, one linear and one polynomial. The second has performed better suggesting that temperature explains 45% of the variation of daily rented bikes.