PSTAT 131 - HW1

Emily Lu

April 06, 2020

Predicting Algae Blooms

Background High concentrations of certain harmful algae in rivers constitute a serious ecological problem with a strong impact not only on river lifeforms, but also on water quality. Being able to monitor and perform an early forecast of algae blooms is essential to improving the quality of rivers. With the goal of addressing this prediction problem, several water samples were collected in different European rivers at different times during a period of approximately 1 year. For each water sample, different chemical properties were measured as well as the frequency of occurrence of seven harmful algae. Some other characteristics of the water collection process were also stored, such as the season of the year, the river size, and the river speed.

Goal We want to understand how these frequencies are related to certain chemical attributes of water samples as well as other characteristics of the samples (like season of the year, type of river, etc.) Data Description The data set consists of data for 200 water samples and each observation in the available datasets is in effect an aggregation of several water samples collected from the same river over a period of 3 months, during the same season of the year. Each observation contains information on 11 variables. Three of these variables are nominal and describe the season of the year when the water samples to be aggregated were collected, as well as the size and speed of the river in question. The eight remaining variables are values of different chemical parameters measured in the water samples forming the aggregation, namely: Maximum pH value, Minimum value of

~~~
    ```r
    s.d <- function(x){
        n <- length(x) # Sample size
        s2 <- sum((x - mean(x))^2)/(n-1) # sample variance
        s.d <- sqrt(s2) # sample standard deviation
        return(s.d)
    }
    ```
* Estimator of mean absolute deviation (MAD):

    $$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_{i} - \bar{x}|$$

    ```r
    mean.abs.d <- function(x){
        n <- length(x) # Sample size
        m <- sum(abs(x - mean(x)))/n # mean average deviation
        return(m)
    }
    ```
~~~

2. Construct box-plots, histograms, QQ-plots and kernel density estimates for these variables. Comment on features such as the distribution and outliers in these plots.
   When asked to construct a graph, you should always precede your graph by the R command/function that generated it properly annotated.

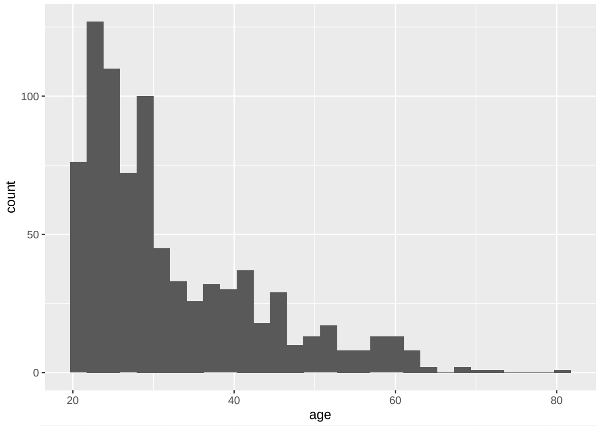
```r
library(MASS)
pima2 = rbind(Pima.tr, Pima.tr2, Pima.te)

x = pima2$age
```

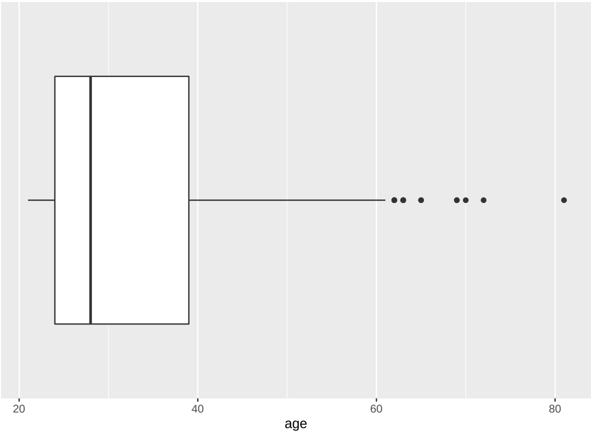
```r
var.name = 'age'
```

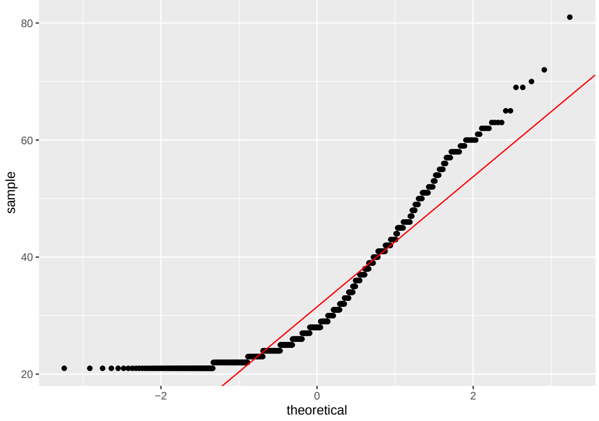
```r
library(ggplot2)

ggplot(pima2, aes(x=age)) + geom_histogram()
```

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

```r
ggplot(pima2, aes(x = factor(0), y = age)) + geom_boxplot() + xlab("") +
    scale_x_discrete(breaks = NULL) + coord_flip()

y     <- quantile(pima2$age, c(0.25, 0.75)) # Find the 1st and 3rd quartiles
x     <- qnorm( c(0.25, 0.75))         # Find the matching normal values on the x-axis
slope <- diff(y) / diff(x)             # Compute the line slope
int   <- y[1] - slope * x[1]           # Compute the line intercept

ggplot(pima2, aes(sample=age)) + stat_qq() +
    geom_abline(intercept=int, slope=slope, color='red')
```


count
age


age


sample
theoretical