Dilirici Radu, Susan Radu
Grupa 232

Documentatie Proiect Statistica Stackloss

Stackloss dataset:

Air.Flow	Water.Temp	Acid.Conc.	stack.loss
80	27	89	42
80	27	88	37
75	25	90	37
62	24	87	28
62	22	87	18
62	23	87	18
62	24	93	19
62	24	93	20
58	23	87	15
58	18	80	14
58	18	89	14
58	17	88	13
58	18	82	11
58	19	93	12
50	18	89	8
50	18	86	7
50	19	72	8
50	19	79	8
50	20	80	9
56	20	82	15
70	20	91	15

Informatii utile:

```
[1] "Medii"
60.4285714285714
21.0952380952381
86.2857142857143
17.5238095238095
[1] "Variatii"
84.0571428571429
9.99047619047619
28.7142857142857
103.461904761905
```

```
cor(stackloss$Air.Flow, stackloss$stack.loss)  # corelatie Air.Flow, stack.loss
```

0.919663452905855

```
cor(stackloss$Water.Temp, stackloss$stack.loss)  # corelatie Water.Temp, stack.loss
```

0.875504404419447

```
cor(stackloss$Acid.Conc, stackloss$stack.loss)  # corelatie Acid.Conc, stack.loss
```

0.399829587099311

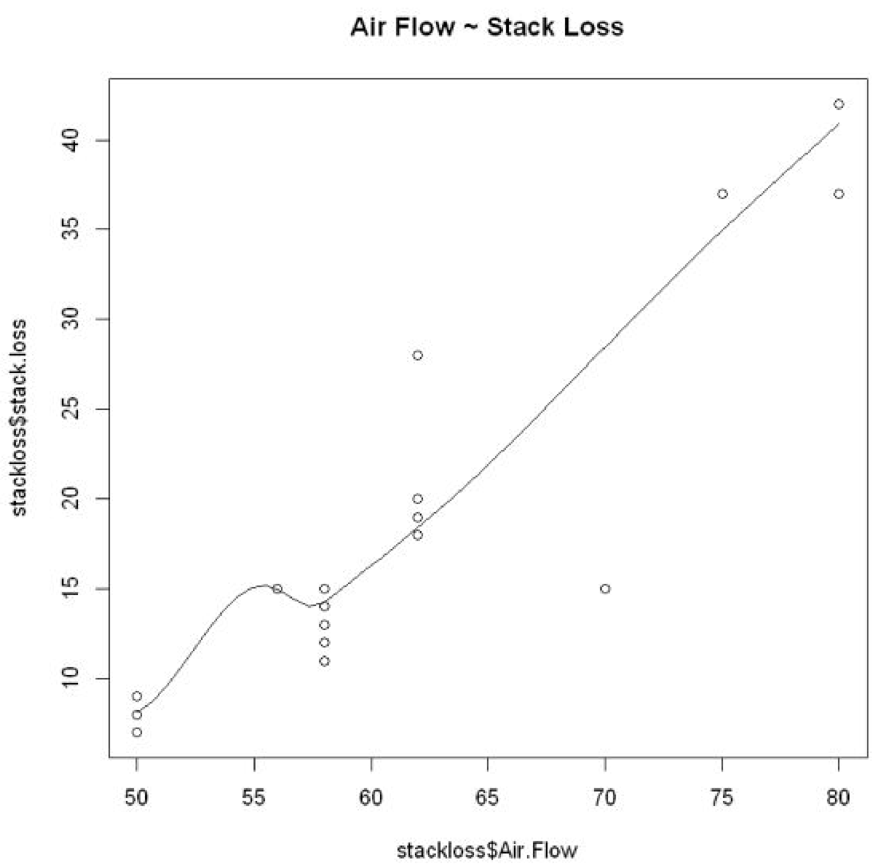
De aici putem observa ca din cele 3 variabile, stackloss este puternic corelata cu Air flow si Water Temperature, iar corelatia cu Acid Concentration e semnificativ mai mica.

Quartile:

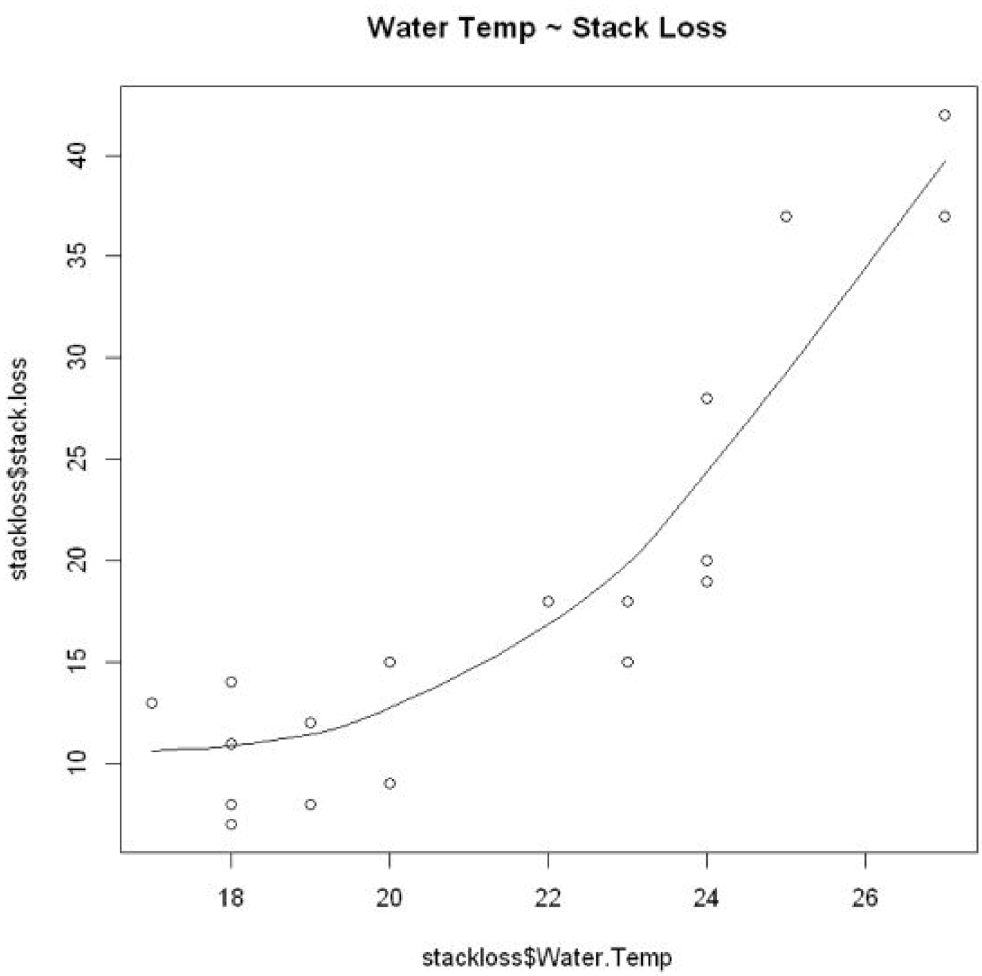
	Air flow	Water Temp.	Acid Conc.	Stackloss
0%	50	17	72	7
25%	56	18	82	11
50%	58	20	87	15
75%	62	24	89	19
100%	80	27	93	42

Incercam sa vedem mai bine relatia dintre ele cu un scatterplot:

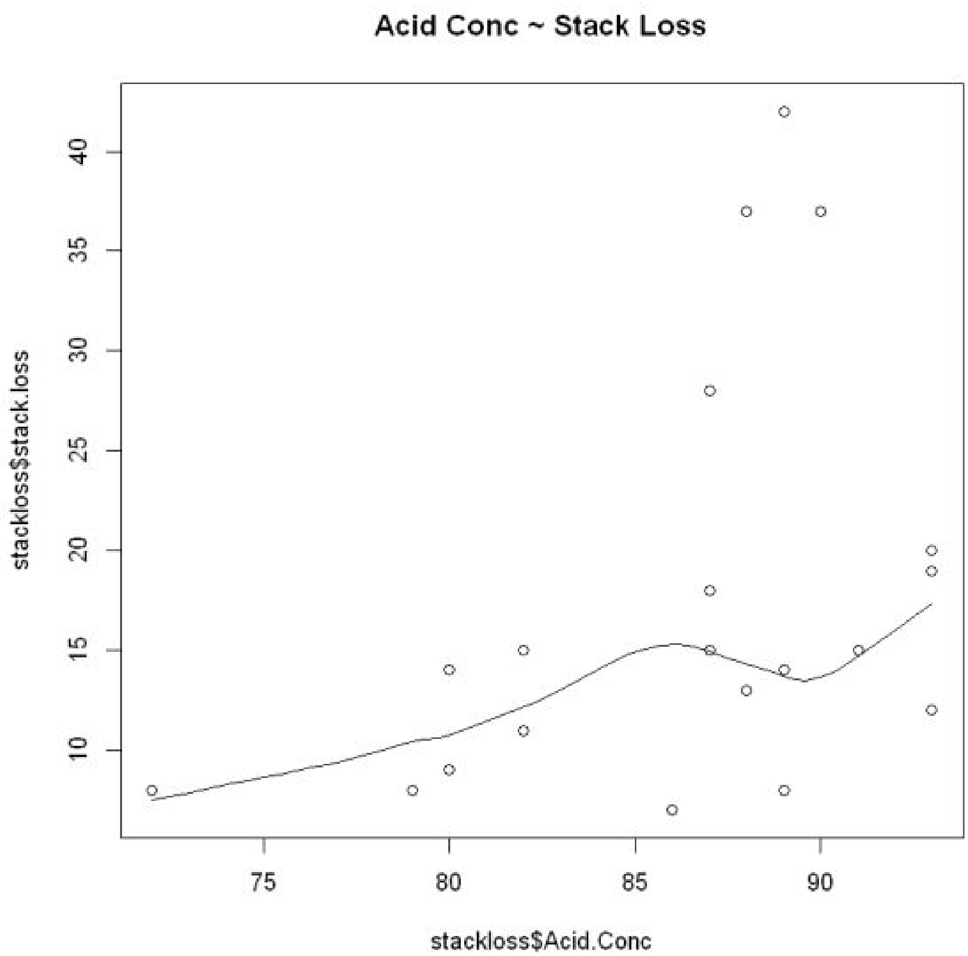
```
scatter.smooth(x=stackloss$Air.Flow, y=stackloss$stack.loss, main="Air Flow ~ Stack Loss")
```



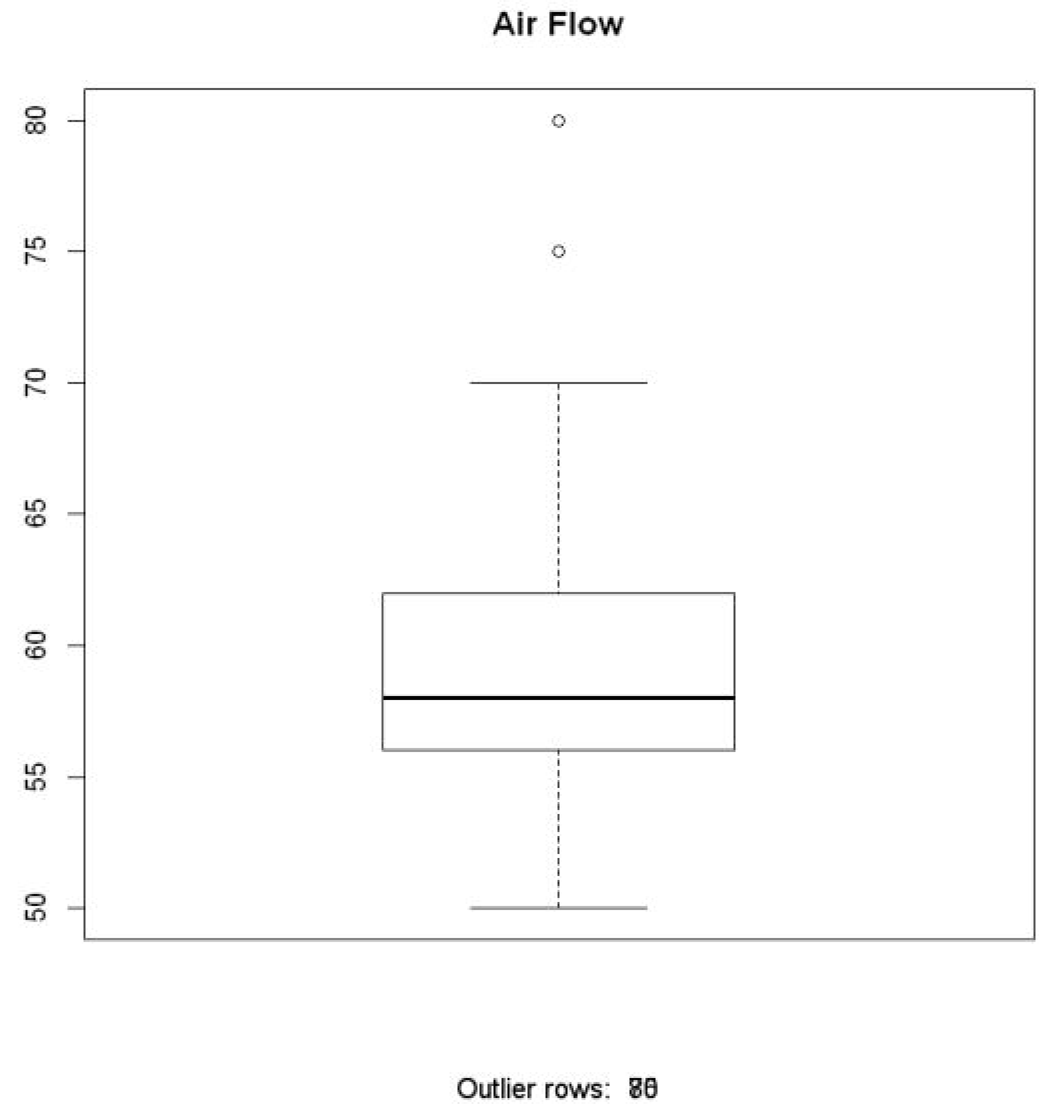
```
scatter.smooth(x=stackloss$Water.Temp, y=stackloss$stack.loss, main="Water Temp ~ Stack Loss")
```

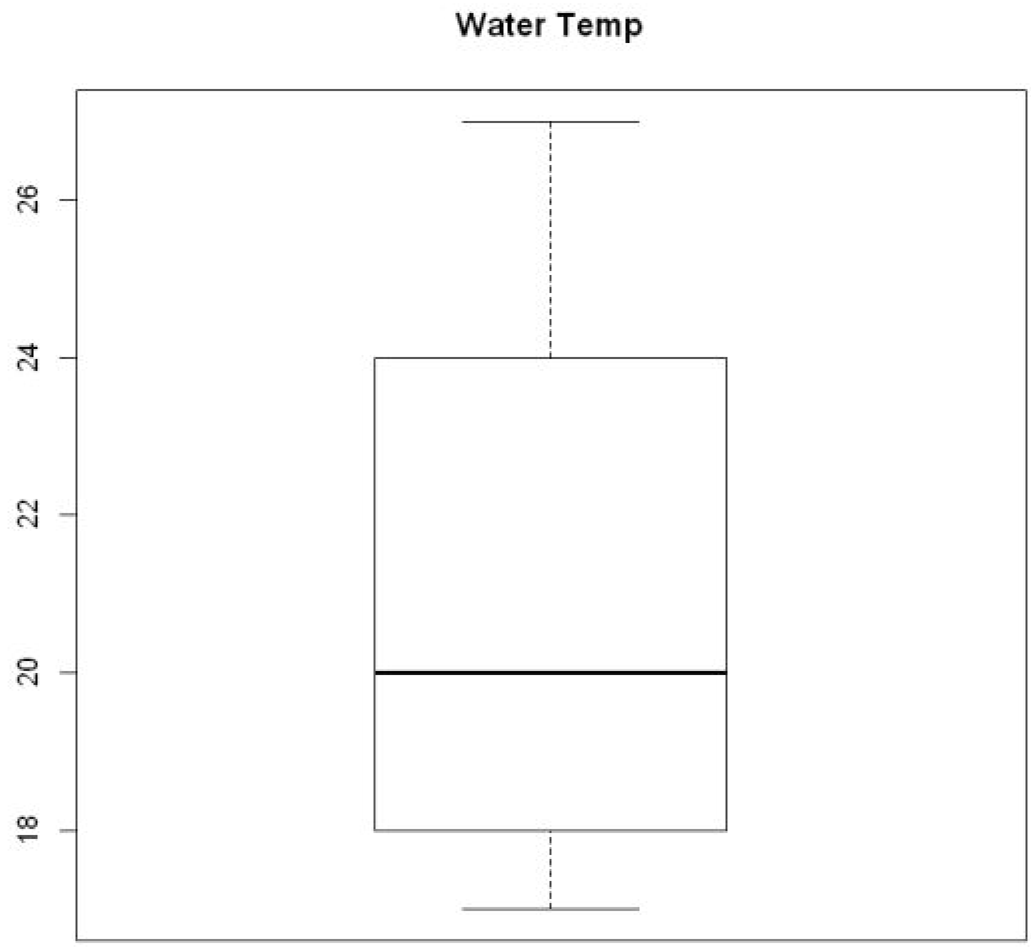


```r
scatter.smooth(x=stackloss$Acid.Conc, y=stackloss$stack.loss, main="Acid Conc ~ Stack L
oss")
```

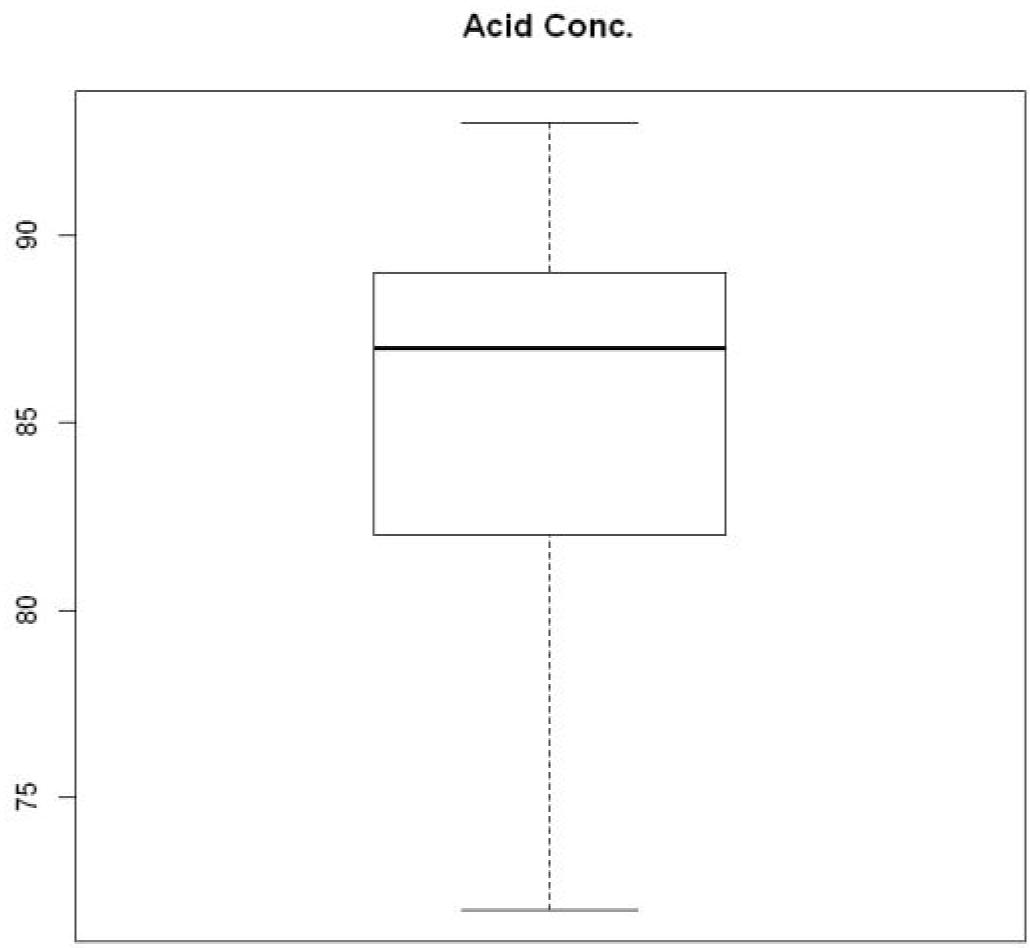


Boxplots:

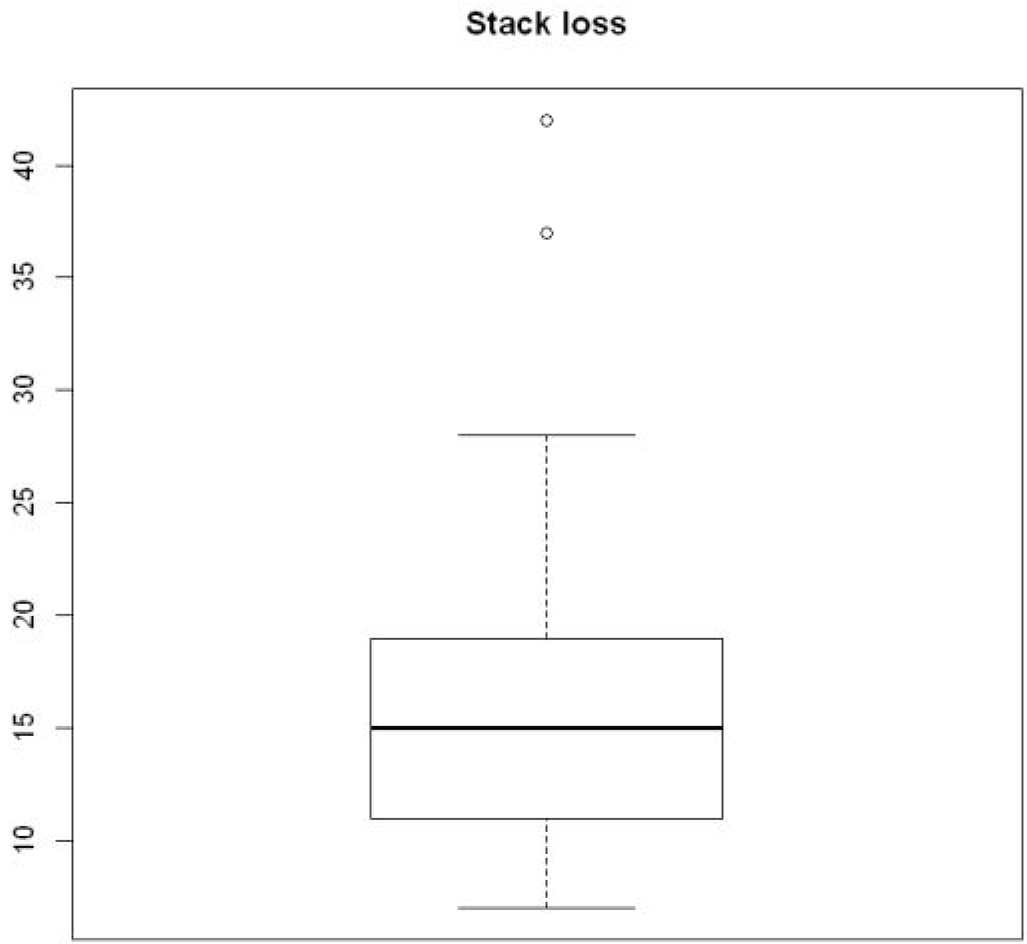




Outlier rows:



Outlier rows:



Outlier rows:

Din boxplot-uri putem observa ca Airflow si stackloss au cateva valori extreme care o sa influenteze negativ modelul de regresie, deoarece dataset-ul este mic.

In afara de Acid Concentration, celelalte doua variabile par a fi bune pentru un model de regresie cu stackloss.La scatterplot, in cazul Acid concentration, unele puncte sunt prea departate de grafic, deci nu pare sa se poata construi un model liniar util. In schimb, putem face o regresie liniara intre stackloss si Air flow. Alegem stackloss ca variabila raspuns pentru ca ne intereseaza cata substanta se pierde la fiecare experiment si ce poate influenta aceasta pierdere.

Regresie stackloss~Airflow:

```
linearMod <- lm(stack.loss ~ Air.Flow, data=stackloss)
print(linearMod)
summary(linearMod)
```

```
Call:
lm(formula = stack.loss ~ Air.Flow, data = stackloss)

Coefficients:
(Intercept)     Air.Flow
     -44.13         1.02


Call:
lm(formula = stack.loss ~ Air.Flow, data = stackloss)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2896  -1.1272  -0.0459   1.1166   8.8728

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.13202    6.10586  -7.228 7.31e-07 ***
Air.Flow      1.02031    0.09995  10.208 3.77e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.098 on 19 degrees of freedom
Multiple R-squared:  0.8458,    Adjusted R-squared:  0.8377
F-statistic: 104.2 on 1 and 19 DF,  p-value: 3.774e-09
```

Regresie stackloss~Air Flow, Water Temperature

```
linearMod2 <- lm(stack.loss ~ Water.Temp + Air.Flow, data=stackloss)
print(linearMod2)
summary(linearMod2)
```

```
Call:
lm(formula = stack.loss ~ Water.Temp + Air.Flow, data = stackloss)

Coefficients:
(Intercept)   Water.Temp     Air.Flow
   -50.3588       1.2954       0.6712


Call:
lm(formula = stack.loss ~ Water.Temp + Air.Flow, data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5290 -1.7505  0.1894  2.1156  5.6588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -50.3588     5.1383  -9.801 1.22e-08 ***
Water.Temp    1.2954     0.3675   3.525  0.00242 **
Air.Flow      0.6712     0.1267   5.298 4.90e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.239 on 18 degrees of freedom
Multiple R-squared:  0.9088,	Adjusted R-squared:  0.8986
F-statistic: 89.64 on 2 and 18 DF,  p-value: 4.382e-10
```

Dintre aceste doua modele de regresie, cel de-al doilea pare mai bun, deoarece are coeficientul R patrat 0.9088, pe cand primul il are 0.8458.

Desemenea, analizand valorile AIC si BIC la fiecare model:

```
AIC(linearMod)
AIC(linearMod2)
BIC(linearMod)
BIC(linearMod2)
```

122.737102300346

113.71438151498

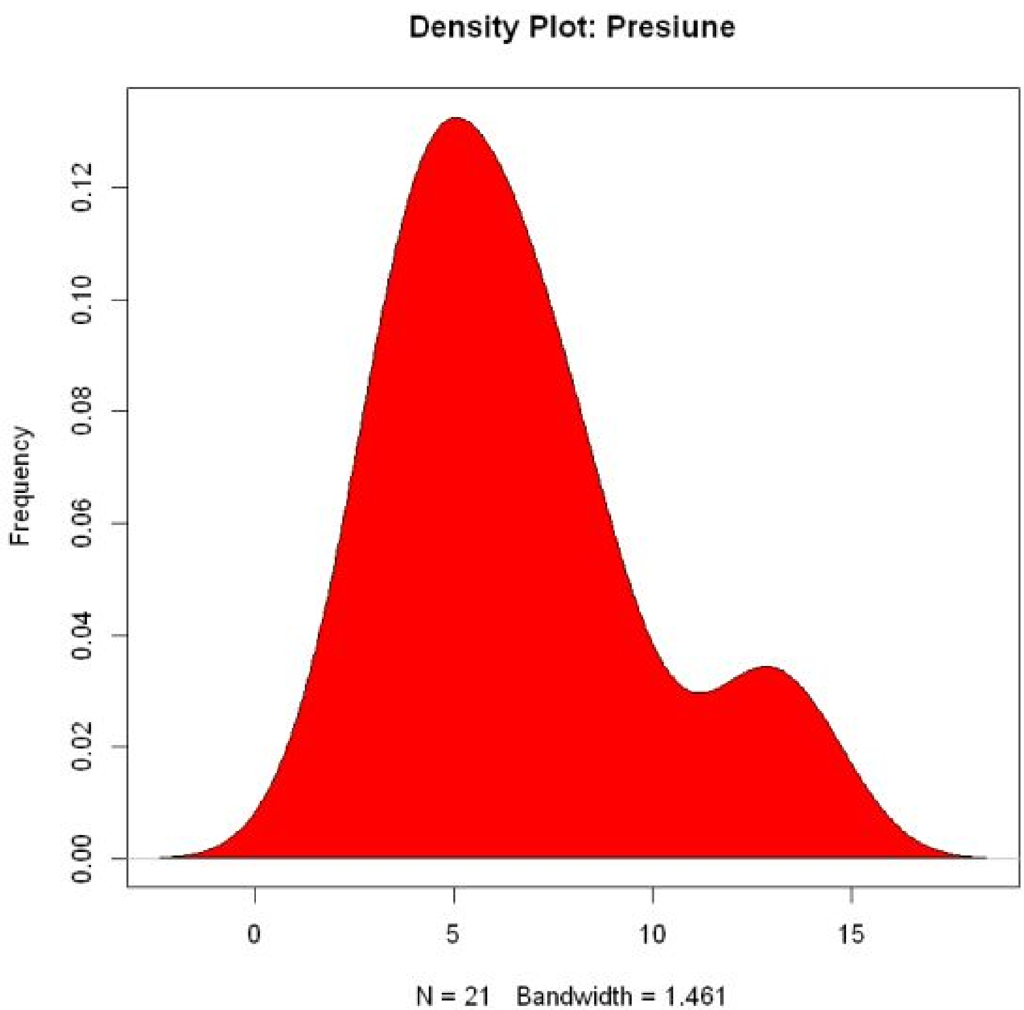
125.870669613517

117.892471265874

Putem observa ca si in cazul acesa modelul de regresie multipla este mai bun, deoarece are valorile AIC respectiv BIC mai mici.

La dataset-ul nostru, am putea adauga presiunea ca variabila:

```
set.seed(7)
p <- rpois(length(stackloss$stack.loss), 7)
plot(density(y), main="Density Plot: Presiune", ylab="Frequency")
polygon(density(y), col="red")
```



Am ales presiunea pentru ca ne-am gandit ca ar putea influenta valorile lui stackloss. Am generat-o folosind o repartitie Poisson cu lambda=7.

Am adaugat variabila la dataset-ul nostrul:

```
my.stackloss <- stackloss
my.stackloss['presiune'] = p
my.stackloss
```

Air.Flow	Water.Temp	Acid.Conc.	stack.loss	presiune
80	27	89	42	14
80	27	88	37	6
75	25	90	37	4
62	24	87	28	3
62	22	87	18	5
62	23	87	18	9
62	24	93	19	6
62	24	93	20	12
58	23	87	15	4
58	18	80	14	7
58	18	89	14	4
58	17	88	13	5
58	18	82	11	9
58	19	93	12	4
50	18	89	8	7
50	18	86	7	4
50	19	72	8	7
50	19	79	8	2
50	20	80	9	13
56	20	82	15	6
70	20	91	15	8

Am construit un model de regresie cu Water Temp. si presiune:

```
linearMod3 <- lm(stack.loss ~ Water.Temp + presiune, data=my.stackloss)
print(linearMod3)
summary(linearMod3)
```

```
Call:
lm(formula = stack.loss ~ Water.Temp + presiune, data = my.stackloss)

Coefficients:
(Intercept)   Water.Temp     presiune
   -41.7734       2.8452      -0.1091


Call:
lm(formula = stack.loss ~ Water.Temp + presiune, data = my.stackloss)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2289 -4.0665  0.1517  2.6086  8.4815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.7734     7.8083  -5.350 4.38e-05 ***
Water.Temp    2.8452     0.3772   7.543 5.61e-07 ***
presiune     -0.1091     0.3654  -0.299    0.769
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.168 on 18 degrees of freedom
Multiple R-squared:  0.7677,	Adjusted R-squared:  0.7418
F-statistic: 29.74 on 2 and 18 DF,  p-value: 1.973e-06
```

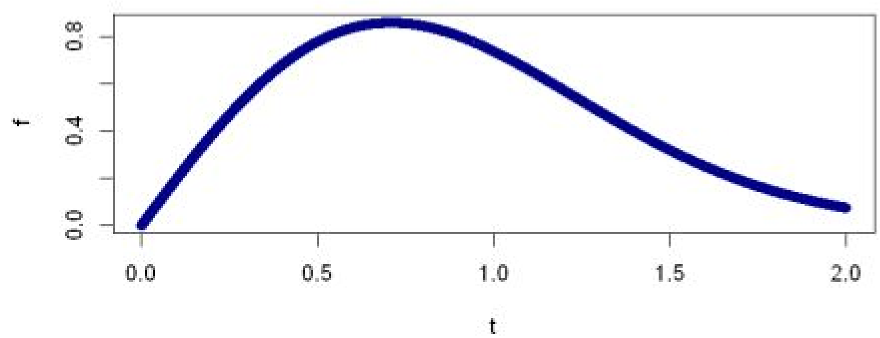
Am observat ca generarea unor valori aleatoare nu a imbunatatit cu nimic modelul. Un mare factor este faptul ca chiar daca valorile formeaza o densitate asemanatoare cu stackloss, nu sunt puse si in ordinea corespunzatoare.

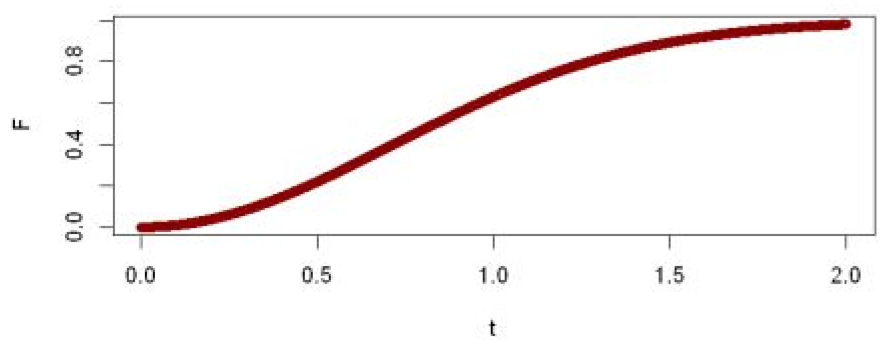
O repartitie care nu a fost prezentata la curs/laborator este repartitia Weibull

```
par(mfrow = c(2, 1))
t <- seq(0, 2, 0.0005)

set.seed(1)
f <- dweibull(t, 2)
plot(t, f, col = "darkblue")

set.seed(1)
F <- pweibull(t, 2)
plot(t, F, col = "darkred")
```





Proprietati ale functiilor:
-Nu sunt simetrice
-Nu sunt pare/impare
-f are un singur punct de maxim

Distributia Weibull este folosita in analiza de supravietuire (Survival analysis) in care se analizeaza timpul pana la defectarea/moartea unui obiect sau a unei vietati in natura.