In [1]:

```
# Problem 1
library(caretEnsemble)
library(RColorBrewer)
library(tm)
library(datarium)
library(leaps)
library(glmnet)
library(pls)
library(gam)
library(splines)
library(MVA)
library(nortest)
library(mvnormtest)
library(pastecs)
library(mvtnorm)
library(igraph)
library(dplyr)
library(ggplot2)
library(ggraph)
library(caret)
library(car)
library(mlbench)
library(tidyverse)
library(MASS)
library(ISLR)
library(psych)
library(faraway)
library(pls)
library(Matrix)
library(stats)
library(biotools)
library(faraway)
```

```
Loading required package: NLP

Loading required package: Matrix

Loaded glmnet 4.1-2


Attaching package: ‘pls’


The following object is masked from ‘package:stats’:

    loadings


Loading required package: splines

Loading required package: foreach

Loaded gam 1.20


Loading required package: HSAUR2

Loading required package: tools
```

```
Attaching package: ‘igraph’


The following objects are masked from ‘package:stats’:

    decompose, spectrum


The following object is masked from ‘package:base’:

    union



Attaching package: ‘dplyr’


The following objects are masked from ‘package:igraph’:

    as_data_frame, groups, union


The following objects are masked from ‘package:pastecs’:

    first, last


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union



Attaching package: ‘ggplot2’


The following object is masked from ‘package:NLP’:

    annotate


The following object is masked from ‘package:caretEnsemble’:

    autoplot


Loading required package: lattice


Attaching package: ‘caret’


The following object is masked from ‘package:pls’:
```

```
    R2

Loading required package: carData

Attaching package: ‘car’

The following object is masked from ‘package:dplyr’:

    recode

── Attaching packages ──────────────────────────────────── tidyverse 1.3.1 ──

✓ tibble  3.1.3     ✓ purrr   0.3.4
✓ tidyr   1.1.3     ✓ stringr 1.4.0
✓ readr   2.0.1     ✓ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
x purrr::accumulate()     masks foreach::accumulate()
x ggplot2::annotate()     masks NLP::annotate()
x tibble::as_data_frame() masks dplyr::as_data_frame(), igraph::as_data_frame()
x ggplot2::autoplot()     masks caretEnsemble::autoplot()
x purrr::compose()        masks igraph::compose()
x tidyr::crossing()       masks igraph::crossing()
x tidyr::expand()         masks Matrix::expand()
x tidyr::extract()        masks pastecs::extract()
x dplyr::filter()         masks stats::filter()
x dplyr::first()          masks pastecs::first()
x dplyr::groups()         masks igraph::groups()
x dplyr::lag()            masks stats::lag()
x dplyr::last()           masks pastecs::last()
x purrr::lift()           masks caret::lift()
x tidyr::pack()           masks Matrix::pack()
x car::recode()           masks dplyr::recode()
x purrr::simplify()       masks igraph::simplify()
x purrr::some()           masks car::some()
x tidyr::unpack()         masks Matrix::unpack()
x purrr::when()           masks foreach::when()

Attaching package: ‘MASS’

The following object is masked from ‘package:dplyr’:

    select


Attaching package: ‘psych’

The following object is masked from ‘package:car’:

    logit

The following objects are masked from ‘package:ggplot2’:
```

```
    %+%, alpha


Attaching package: ‘faraway’


The following object is masked from ‘package:psych’:

    logit


The following objects are masked from ‘package:car’:

    logit, vif


The following object is masked from ‘package:lattice’:

    melanoma


The following objects are masked from ‘package:HSAUR2’:

    epilepsy, toenail


---
biotools version 4.2
```

In [2]:
```
attach(swiss)
head(swiss)
```

A data.frame: 6 × 6

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6

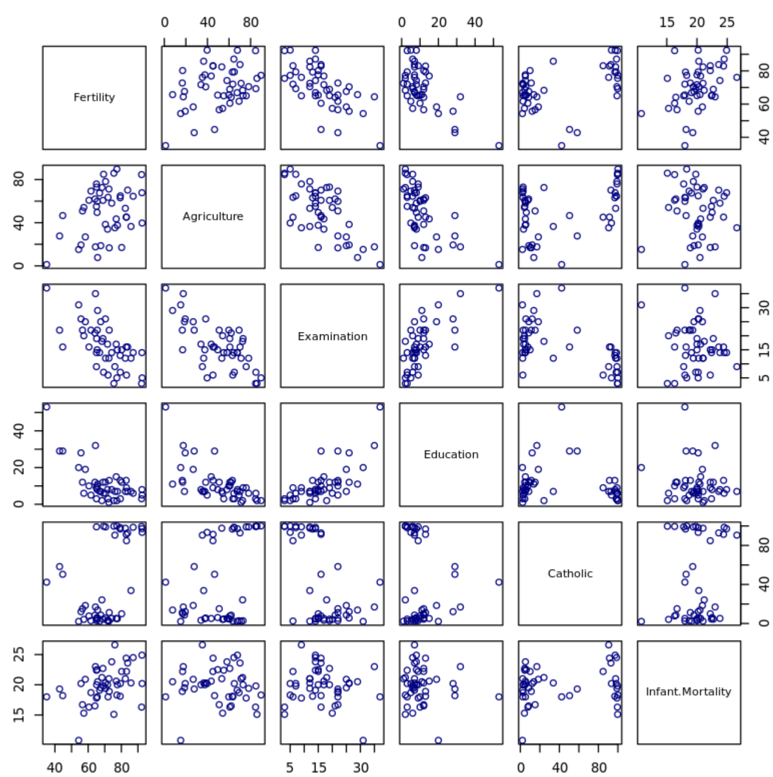
In [3]:
```
model_01 <- lm(Fertility~., data=swiss)
model_01
summary(model_01)
```

```
Call:
lm(formula = Fertility ~ ., data = swiss)

Coefficients:
```

```
     (Intercept)       Agriculture       Examination         Education 
         66.9152           -0.1721           -0.2580           -0.8709 
        Catholic  Infant.Mortality 
          0.1041            1.0770 
Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2743  -5.2617   0.5032   4.1198  15.3213 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
Examination      -0.25801    0.25388  -1.016  0.31546    
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 ** 
Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671 
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Agriculture, Education, Catholic, Infant.Mortality are statistically significant.Write the equation for the Linear Formula based on your linear regression: Yi = b0 + b1xi1 + b2xi2 + ... + bkxik + ei where Y is the outcome or response, b0 is the intercept, bj represents the estimated coefficient of the j predictor and e is the random error term that cannot be explained by the model: Fertility = 66.9152 + (-0.1721)*Agriculture + (-0.2580)*Examination + (-0.8709)*Education + (0.1041)*Catholic + (1.0770)*Infant.Mortality

In [4]:
```
pairs(swiss, col='navy')
```

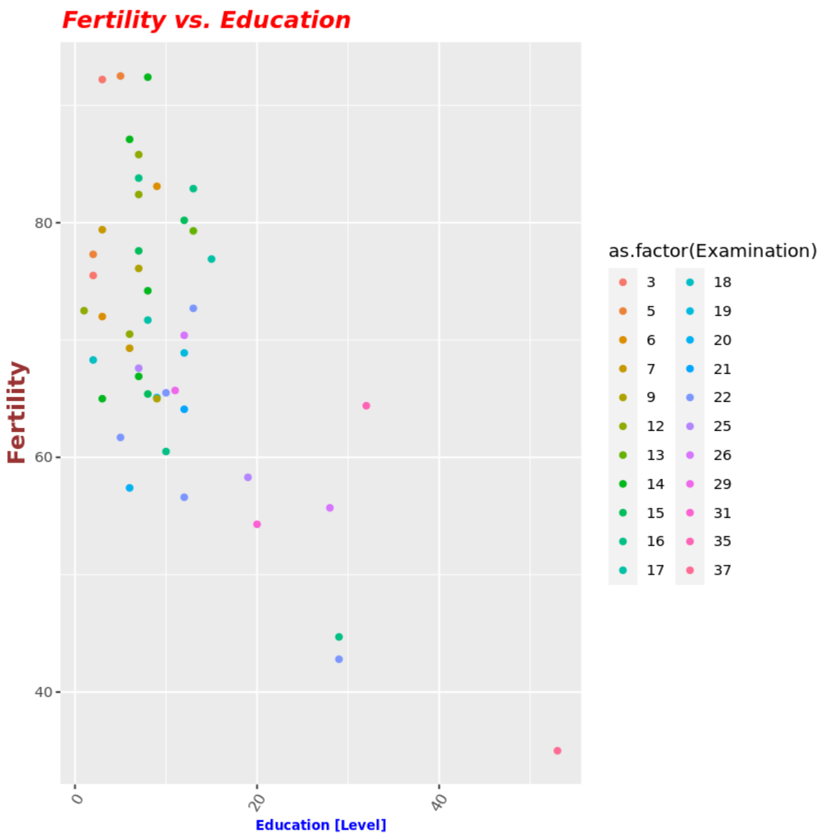


In [5]:

```
options(digits=3)
```

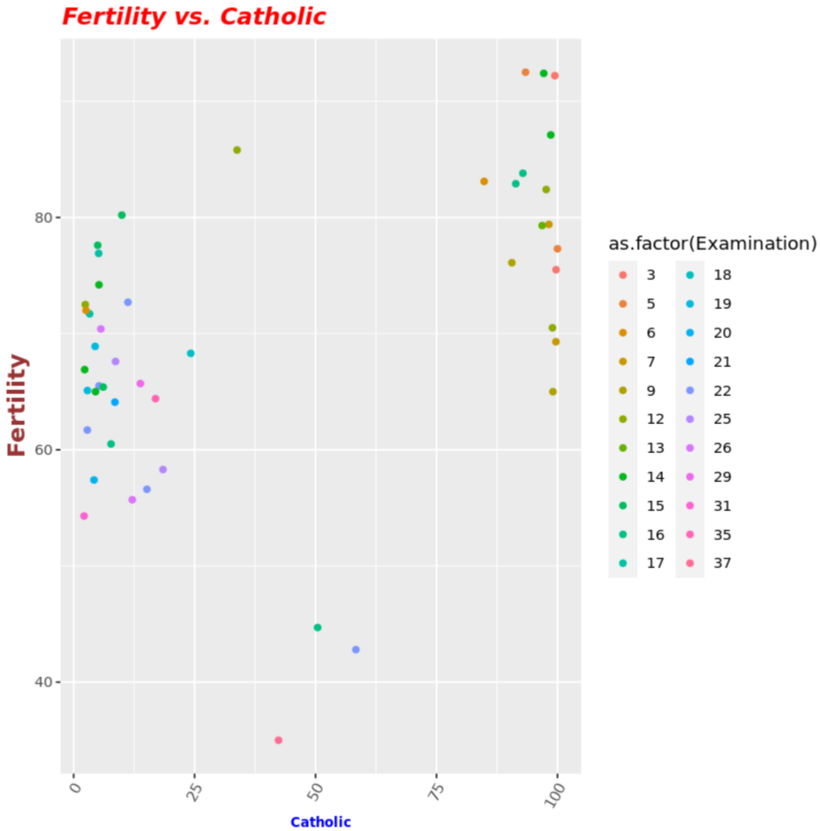
In [6]:

```
ggplot(swiss, aes(x=Education, y=Fertility, col= as.factor(Examination))) +geom_point()+g
  xlab("Education [Level]") + ylab("Fertility")+
  theme(
    plot.title = element_text(color="red", size=14, face="bold.italic"),
    axis.title.x = element_text(color="blue", size=8, face="bold"),
    axis.title.y = element_text(color="#993333", size=14, face="bold")
  )+ theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



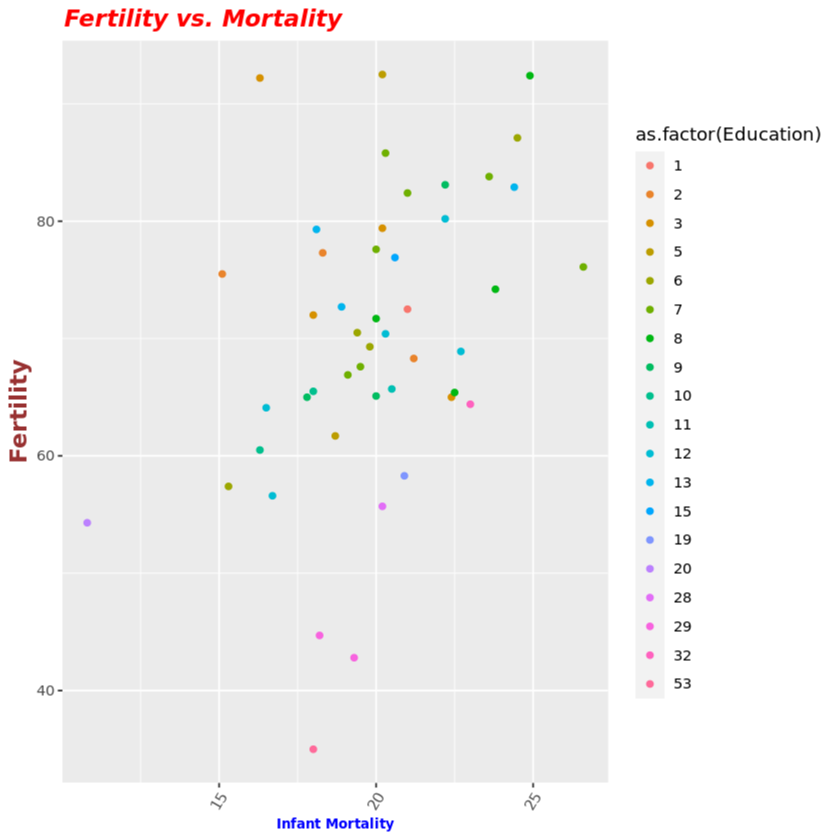
In [7]:

```
ggplot(swiss, aes(x=Catholic, y=Fertility, col= as.factor(Examination))) +geom_point()+gg
  xlab("Catholic") + ylab("Fertility")+
  theme(
    plot.title = element_text(color="red", size=14, face="bold.italic"),
    axis.title.x = element_text(color="blue", size=8, face="bold"),
    axis.title.y = element_text(color="#993333", size=14, face="bold")
  )+ theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



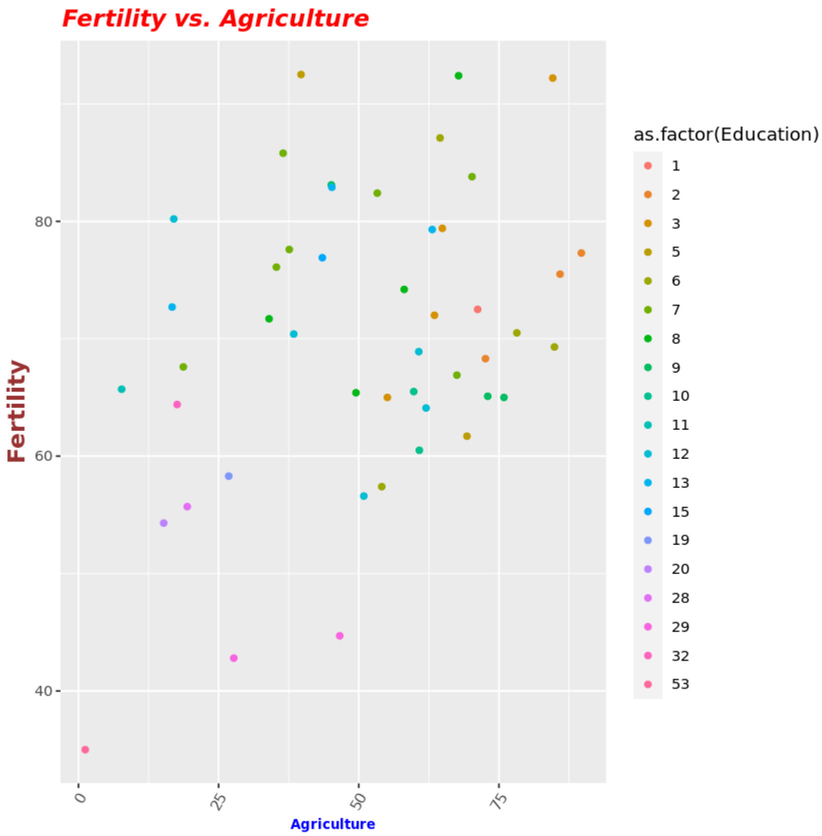
In [8]:

```
ggplot(swiss, aes(x=Infant.Mortality, y=Fertility, col= as.factor(Education))) +geom_poin
  xlab("Infant Mortality") + ylab("Fertility")+
  theme(
    plot.title = element_text(color="red", size=14, face="bold.italic"),
    axis.title.x = element_text(color="blue", size=8, face="bold"),
    axis.title.y = element_text(color="#993333", size=14, face="bold")
  )+ theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



In [9]:

```
ggplot(swiss, aes(x=Agriculture, y=Fertility, col= as.factor(Education))) +geom_point()+g
  xlab("Agriculture") + ylab("Fertility")+
  theme(
    plot.title = element_text(color="red", size=14, face="bold.italic"),
    axis.title.x = element_text(color="blue", size=8, face="bold"),
    axis.title.y = element_text(color="#993333", size=14, face="bold")
  )+ theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



In [10]:

```
# Coefficians:
round(coef(summary(model_01)),5)
```

A matrix: 6 × 4 of type dbl

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.915	10.7060	6.25	0.00000
Agriculture	-0.172	0.0703	-2.45	0.01873
Examination	-0.258	0.2539	-1.02	0.31546
Education	-0.871	0.1830	-4.76	0.00002
Catholic	0.104	0.0353	2.95	0.00519
Infant.Mortality	1.077	0.3817	2.82	0.00734

In [11]:

```
# Confidence Intervals
confint(model_01)
```

A matrix: 6 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	45.2939	88.5365
Agriculture	-0.3141	-0.0301
Examination	-0.7707	0.2547
Education	-1.2406	-0.5013
Catholic	0.0329	0.1753

	2.5 %	97.5 %
Infant.Mortality	0.3061	1.8479

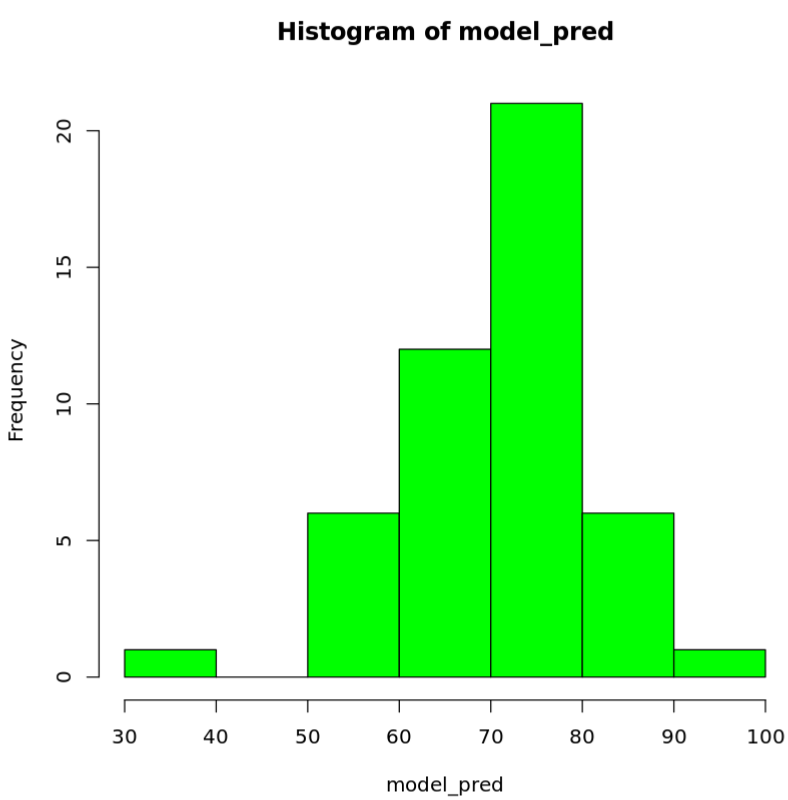
In [12]:
```
#Anova
anova(model_01)
```

A anova: 6 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Agriculture	1	895	894.8	17.43	1.52e-04
Examination	1	2210	2210.4	43.05	6.88e-08
Education	1	892	891.8	17.37	1.55e-04
Catholic	1	667	667.1	12.99	8.39e-04
Infant.Mortality	1	409	408.8	7.96	7.34e-03
Residuals	41	2105	51.3	NA	NA

In [13]:
```
# Visualizing Predictions
model_pred <- predict(model_01)
```

In [14]:
```
hist(model_pred, col='green')
```



In [15]:
```
# problem 2
```

```
attach(rock)
```

In [16]:

```
head(rock, 10)
```

A data.frame: 10 × 4

	area	peri	shape	perm
	<int>	<dbl>	<dbl>	<dbl>
1	4990	2792	0.0903	6.3
2	7002	3893	0.1486	6.3
3	7558	3931	0.1833	6.3
4	7352	3869	0.1171	6.3
5	7943	3949	0.1224	17.1
6	7979	4010	0.1670	17.1
7	9333	4346	0.1897	17.1
8	8209	4345	0.1641	17.1
9	8393	3682	0.2037	119.0
10	6425	3099	0.1624	119.0

In [17]:

```
model_02 <- lm(perm~., data=rock)
model_02
summary(model_02)
```

```
Call:
lm(formula = perm ~ ., data = rock)

Coefficients:
(Intercept)         area         peri        shape  
   485.6180       0.0913      -0.3440     899.0693  
Call:
lm(formula = perm ~ ., data = rock)

Residuals:
   Min     1Q Median     3Q    Max 
-750.3  -59.6   10.7  100.3  620.9 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 485.6180   158.4083    3.07  0.00371 ** 
area          0.0913     0.0250    3.65  0.00068 ***
peri         -0.3440     0.0511   -6.73  2.8e-08 ***
shape       899.0693   506.9510    1.77  0.08307 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246 on 44 degrees of freedom
Multiple R-squared:  0.704,	Adjusted R-squared:  0.684 
F-statistic:   35 on 3 and 44 DF,  p-value: 1.03e-11
```

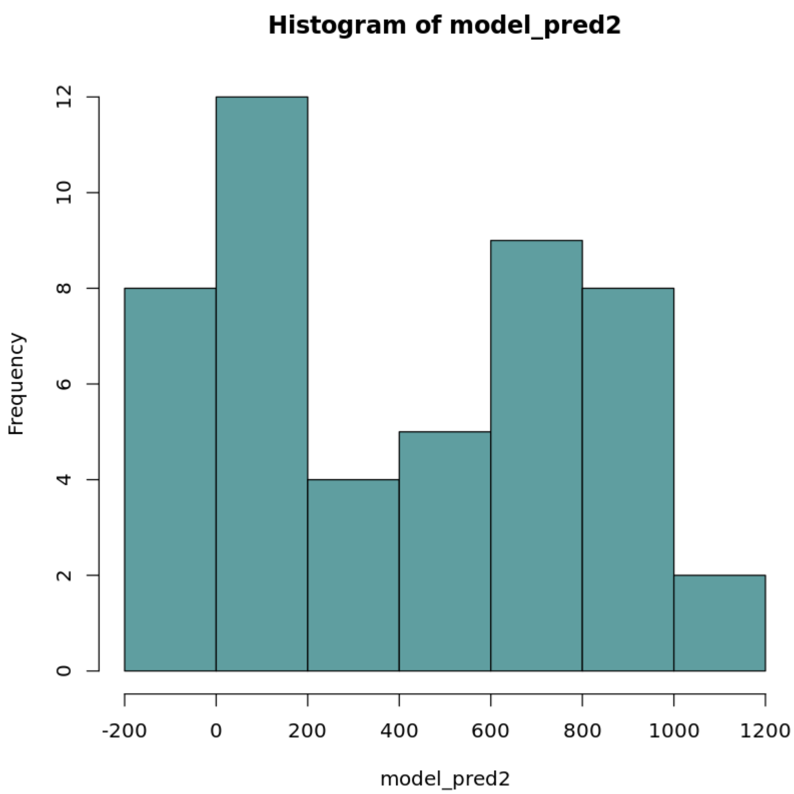
area, peri are statistically significant. Write the equation for the Linear Formula based on your linear regression: Yi = b0 + b1xi1 + b2xi2 + ... + bkxik + ei where Y is the outcome or response, b0 is the intercept, bj represents the estimated coefficient of the j predictor and e is the random error term that cannot be explained by the model: perm = 485.6180 + (0.0913)*area + (-0.3440)*peri + (899.0693)*shape

In [19]:

```
# Visualizing Predictions
model_pred2 <- predict(model_02)
```

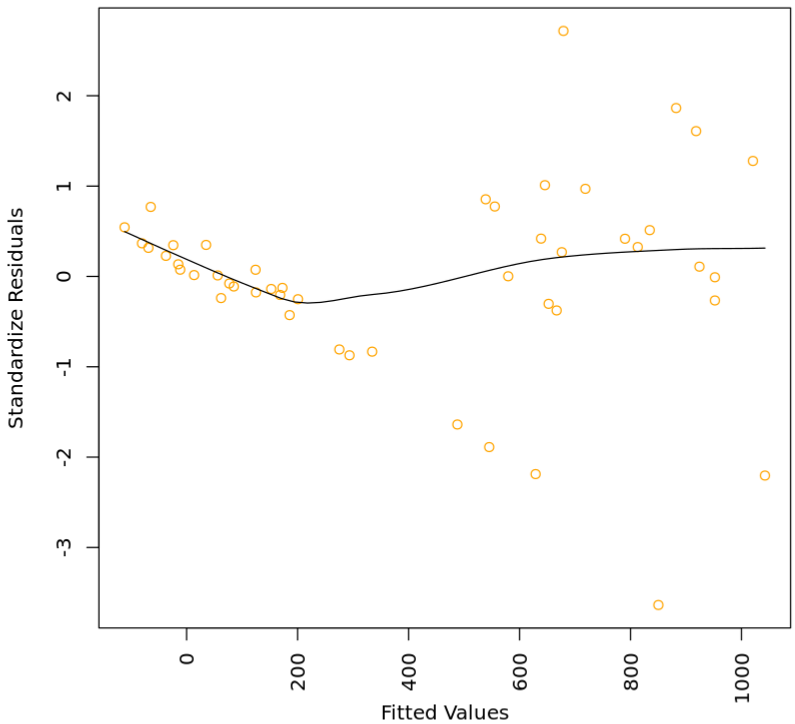
In [20]:

```
hist(model_pred2, col='cadetblue')
```

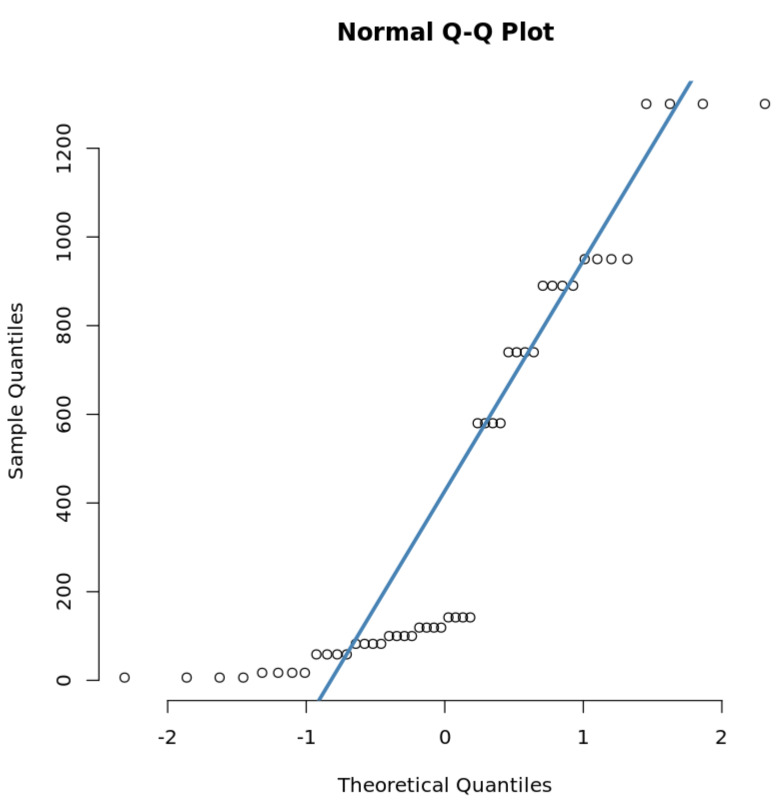


In [21]:

```
# Plot residuals against the fitted values
scatter.smooth(rstandard(model_02) ~ fitted(model_02), col="orange",
               las=3, ylab="Standardize Residuals", xlab="Fitted Values")
```

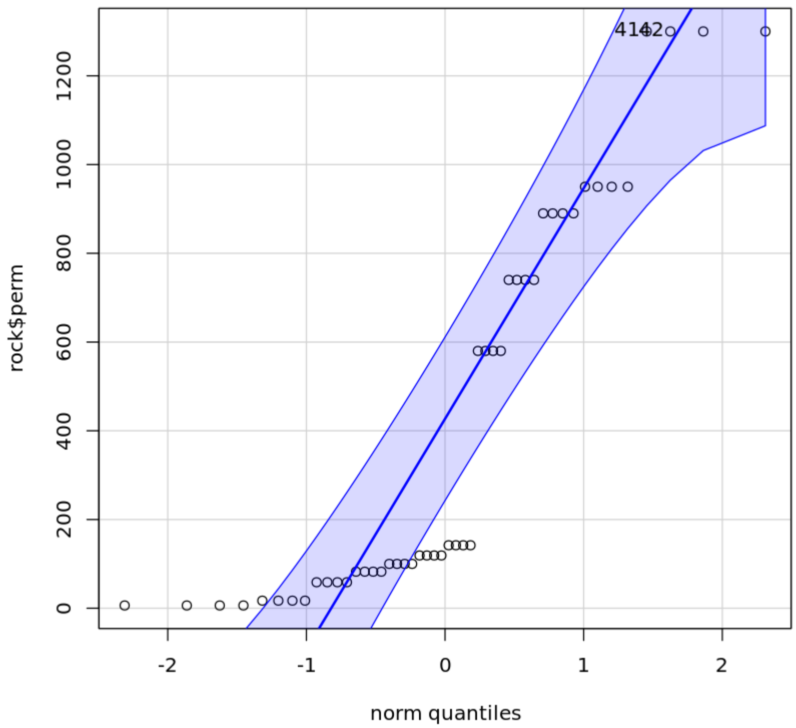


In [22]:
```
qqnorm(rock$perm, pch = 1, frame = FALSE)
qqline(rock$perm, col = "steelblue", lwd = 3)
```



In [43]:
```
qqPlot(rock$perm)
```

41 · 42



In [44]:
```
# Shapiro-Wilk normality test
shapiro.test(rock$perm)
```

```
	Shapiro-Wilk normality test

data:  rock$perm
W = 0.8, p-value = 2e-06
```

The result is significant, so we can't assume the normality.

In [23]:
```
# Problem 3
attach(prostate)
head(prostate, 10)
```

A data.frame: 10 × 9

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
	<dbl>	<dbl>	<int>	<dbl>	<int>	<dbl>	<int>	<int>	<dbl>
1	-0.580	2.77	50	-1.386	0	-1.39	6	0	-0.431
2	-0.994	3.32	58	-1.386	0	-1.39	6	0	-0.163
3	-0.511	2.69	74	-1.386	0	-1.39	7	20	-0.163
4	-1.204	3.28	58	-1.386	0	-1.39	6	0	-0.163
5	0.751	3.43	62	-1.386	0	-1.39	6	0	0.372
6	-1.050	3.23	50	-1.386	0	-1.39	6	0	0.765
7	0.737	3.47	64	0.615	0	-1.39	6	0	0.765

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
	<dbl>	<dbl>	<int>	<dbl>	<int>	<dbl>	<int>	<int>	<dbl>
8	0.693	3.54	58	1.537	0	-1.39	6	0	0.854
9	-0.777	3.54	47	-1.386	0	-1.39	6	0	1.047
10	0.223	3.24	63	-1.386	0	-1.39	6	0	1.047

In [24]:

```
str(prostate)
```

```
'data.frame':	97 obs. of  9 variables:
 $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
 $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
 $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
 $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
 $ svi    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
 $ gleason: int  6 6 7 6 6 6 6 6 6 6 ...
 $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
 $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

In [25]:

```
model_03 <- lm(lpsa~., data=prostate)
summary(model_03)
model_03
```

```
Call:
lm(formula = lpsa ~ ., data = prostate)

Residuals:
   Min     1Q Median     3Q    Max 
-1.733 -0.371 -0.017  0.414  1.638 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.66934    1.29639    0.52   0.6069    
lcavol       0.58702    0.08792    6.68  2.1e-09 ***
lweight      0.45447    0.17001    2.67   0.0090 ** 
age         -0.01964    0.01117   -1.76   0.0823 .  
lbph         0.10705    0.05845    1.83   0.0704 .  
svi          0.76616    0.24431    3.14   0.0023 ** 
lcp         -0.10547    0.09101   -1.16   0.2496    
gleason      0.04514    0.15746    0.29   0.7750    
pgg45        0.00453    0.00442    1.02   0.3089    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.708 on 88 degrees of freedom
Multiple R-squared:  0.655,	Adjusted R-squared:  0.623 
F-statistic: 20.9 on 8 and 88 DF,  p-value: <2e-16
Call:
lm(formula = lpsa ~ ., data = prostate)

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi  
    0.66934      0.58702      0.45447     -0.01964      0.10705      0.76616  
        lcp      gleason        pgg45  
   -0.10547      0.04514      0.00453  
```

Write the equation for the Linear Formula based on your linear regression: lpsa = 0.66934 + (0.58702)*lcavol + (0.45447)*lweight + (-0.01964)*age + (0.10705)*lbph + (0.76616)*svi + (-0.10547)*lcp + (0.04514)*gleason + (0.00453)*pgg45

In [26]:
```
#install.packages("broom")
```

In [27]:
```
library(broom)
```

In [28]:
```
glance(model_03)
```

A tibble: 1 × 12

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nol
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int
0.655	0.623	0.708	20.9	2.24e-17	8	-99.5	219	245	44.2	88	[illegible]

AIC: 219

In [29]:
```
#intercept-only model
intercept_only <- lm(lpsa ~ 1, data=prostate)
```

In [30]:
```
summary(intercept_only)
```

```
Call:
lm(formula = lpsa ~ 1, data = prostate)

Residuals:
   Min     1Q Median     3Q    Max 
-2.909 -0.747  0.113  0.578  3.104 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.478      0.117    21.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.15 on 96 degrees of freedom
```

In [31]:
```
summary(model_03)
```

```
Call:
lm(formula = lpsa ~ ., data = prostate)

Residuals:
   Min     1Q Median     3Q    Max 
-1.733 -0.371 -0.017  0.414  1.638 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.66934    1.29639    0.52   0.6069    
lcavol       0.58702    0.08792    6.68  2.1e-09 ***
lweight      0.45447    0.17001    2.67   0.0090 ** 
```

```
age         -0.01964    0.01117   -1.76   0.0823 .
lbph         0.10705    0.05845    1.83   0.0704 .
svi          0.76616    0.24431    3.14   0.0023 **
lcp         -0.10547    0.09101   -1.16   0.2496
gleason      0.04514    0.15746    0.29   0.7750
pgg45        0.00453    0.00442    1.02   0.3089
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.708 on 88 degrees of freedom
Multiple R-squared:  0.655,     Adjusted R-squared:  0.623
F-statistic: 20.9 on 8 and 88 DF,  p-value: <2e-16
```

In [32]:

```
library(leaps)
```

In [33]:

```
tmp<-regsubsets(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data=
```

In [34]:

```
all_mods <- summary(tmp)[[1]]
all_mods <- lapply(1:nrow(all_mods), function(x) as.formula(paste("lpsa~", paste(names(wh
head(all_mods)
```

```
[[1]]
lpsa ~ lweight
<environment: 0x55f4a3acf328>

[[2]]
lpsa ~ gleason
<environment: 0x55f4a3afc7d8>

[[3]]
lpsa ~ age
<environment: 0x55f4a3f17be8>

[[4]]
lpsa ~ lcavol
<environment: 0x55f4a3f1d2e0>

[[5]]
lpsa ~ pgg45
<environment: 0x55f4a3f1eba8>

[[6]]
lpsa ~ svi
<environment: 0x55f4a3f204a8>
```

In [35]:

```
all.lm<-lapply(all_mods, lm, prostate)
sapply(all.lm, extractAIC)[2,]
```

17.8403457601089 · 16.6408170305833 · 28.006697193729 · -44.366078653343 · 11.7826711146296 · -6.65782447556234 · 27.6495252412358 · -3.92583179091702 · -52.6904253337696 · -43.0527238826421 · -42.3702561403725 · -17.8838448393945 · -15.3592868880694 · -10.4232624343575 · -6.20523897865511 · -3.3474374953463 · -3.33377248373142 · -0.287461302801033 · 13.061121338133 · 13.4232063244145 · 3.38235056095283 · 16.0267917916439 · 19.7493207716217 · 19.388094642313 ·

18.0394827436608 · -43.4525362930821 · 28.341807251681 · -45.2027510167193 ·
-47.8977727777257 · -51.3969674345817 · -9.59050077083949 · -3.25896911979357 ·
11.1928995091139 · -12.2907091829577 · -12.4683284903337 · -6.72451193997565 ·
-60.6762056888869 · -54.326904033922 · -52.2356458797727 · -51.9347344862093 ·
-51.4626251128488 · -52.1314434906208 · -49.6635919252833 · -46.3588661589353 ·
-49.4035512307305 · -41.0625618571322 · -43.4253480399745 · -41.7141641715896 ·
-46.5886614362062 · -43.2842367171758 · -41.4626098943508 · -23.8409696718923 ·
-21.5360430605254 · -17.8411016524908 · -23.3416856432358 · -15.9288962326842 ·
-15.2421226030914 · -15.0499881950001 · -15.6709400918703 · -13.8946649944391 ·
-13.3788195794766 · -12.2440740150112 · -8.95775483955692 · -8.9128412426551 ·
-11.5198122150304 · -10.5412036847367 · -8.12207645193694 · -5.75331817354464 ·
-2.22002930667959 · -1.60760148733998 · -4.93294794765438 · 1.38348649605743 ·
-2.12088308295128 · 1.71106743326993 · 0.387371494401829 · -43.2515805952965 ·
12.5293637454592 · 14.788913077323 · -49.6572594436937 · 5.14460753656548 ·
5.38071399740246 · 13.1898848220978 · 17.9685747297009 · 21.3694060419506 ·
-47.2354297376101 · -48.3247867958338 · -50.1802419296138 · -57.6755366306678 ·
-11.3525981581593 · -14.2309496050154 · -5.62705228581085 · -17.6458955544812 ·
-61.3517930886 · -59.8473565603895 · -59.4967272155732 · -58.8064444667279 ·
-59.4691019941421 · -54.1834070636418 · -53.2219667831581 · -55.7541801588237 ·
-52.3541589409133 · -52.4343303033505 · -51.5510358769055 · -50.9705359238407 ·
-51.5716613526586 · -51.4564128933176 · -51.082609577189 · -50.9403964145822 ·
-48.1766508529333 · -56.7352317311376 · -47.6640338605949 · -48.1993454251334 ·
-47.6611887057573 · -45.33037518216 · -46.632498138417 · -45.3069495610285 ·
-48.1925745081722 · -41.4879832175907 · -39.7143108956047 · -41.4681416625508 ·
-47.6776555547438 · -45.8919463982768 · -41.3155121621542 · -23.5022895255518 ·
-23.2343865085841 · -24.854287729997 · -21.8440013010698 · -20.7266567509754 ·
-19.7229901194455 · -21.7301458467555 · -22.4632855982972 · -21.6455654510084 ·
-17.0907217783893 · -15.870517905797 · -13.3825204620357 · -13.5506962175729 ·
-13.2901353054889 · -13.0992223766509 · -13.8538612311614 · -13.9589434776963 ·
-13.7507798472544 · -11.8973979530945 · -10.8881813317575 · -10.2733754825463 ·
-15.671523197529 · -7.34671766848627 · -10.1280954248545 · -12.2566148402494 ·
-3.81111303276786 · -49.1876551137943 · -3.90900939466594 · -0.371343031079395 ·
-3.68638696862941 · 1.8643460432163 · 3.34929567435453 · 2.38637593559684 ·
-46.5025266376337 · -56.0573933749993 · 14.5289434379672 · 7.1384987631457 ·
-56.0210830737545 · -16.2209395544565 · -60.0918816805667 · -61.374391959973 ·
-59.7963505494484 · -59.5654916283568 · -59.2927371332246 · -58.7974114230059 ·
-58.7163722934164 · -57.8845793108511 · -57.6437327537032 · -57.9365314145183 ·
-54.0093450207768 · -54.3142541633977 · -52.1836554001214 · -52.1955412902215 ·
-51.3875835564675 · -51.3288649235875 · -51.2819730855795 · -54.0521383819151 ·
-54.9594935137087 · -50.4581387885706 · -51.0389724517166 · -50.1779133377495 ·
-50.0229013385949 · -55.1894533535734 · -46.1864406601784 · -47.2000606165167 ·
-55.372808278637 · -46.2097153302048 · -47.2572314285858 · -44.12497460419 ·
-44.7979568724914 · -45.9053954280281 · -39.516460632686 · -45.7451518567485 ·
-24.0321773788631 · -22.5449191874223 · -21.5933071141939 · -21.3101146100732 ·

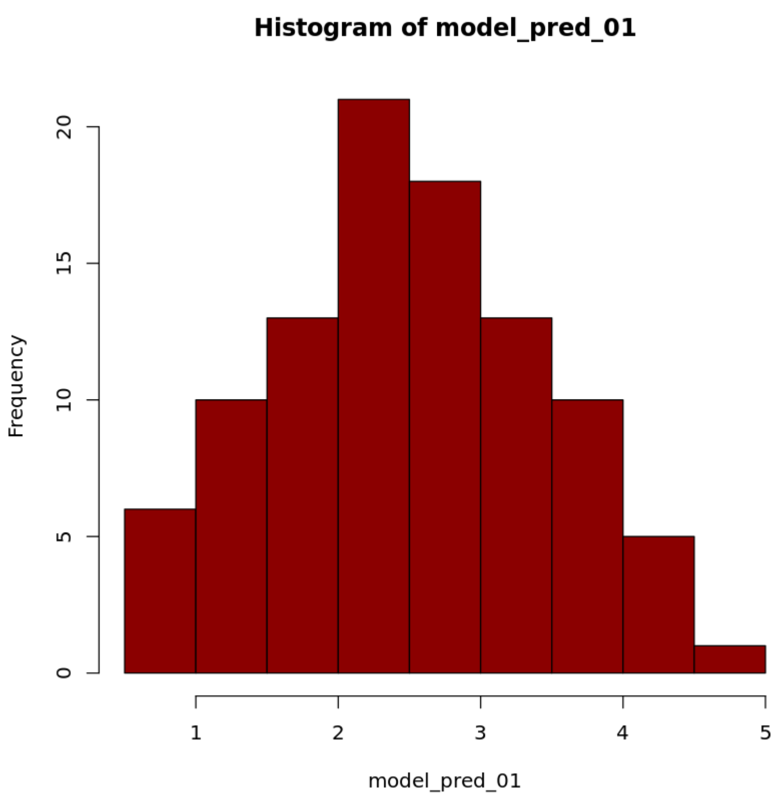
-22.8613827072383 · -23.0442983257433 · -19.2481122481577 · -20.7761080381279 · -21.1654782606323 · -20.1152949191056 · -15.0909772594281 · -15.1033011849478 · -12.1197419847596 · -12.1358226494728 · -11.5189087933118 · -11.9712528307677 · -11.6241666795313 · -8.89273362518049 · -14.221106961061 · -54.9744185236867 · -1.94653818254626 · 3.84286509154306 · -60.3508883154857 · -60.7886748146441 · -58.9592530733739 · -58.5595867412226 · -59.6502156873116 · -58.1024498778683 · -58.3005948987251 · -57.3941869665546 · -57.4637683576411 · -56.7631114858812 · -52.0179981138804 · -52.0372835270785 · -53.7450505862972 · -52.9969278260493 · -50.1957376924529 · -50.0325502264086 · -49.4424626991591 · -53.3823026195116 · -54.7460107686086 · -45.2657509131698 · -43.9686043929414 · -22.5581282964948 · -22.0322792454369 · -20.8718976892494 · -21.0644193667541 · -19.5758339418333 · -13.1033127625729 · -10.3448260725615 · -59.1738945601312 · -60.2312951721691 · -58.8526540957516 · -56.9751533323352 · -56.6928302310116 · -50.0455238843363 · -52.7488121205749 · -20.5671934435049 · -58.3218423151846

In [36]:

```
# Visualizing Predictions
model_pred_01 <- predict(model_03)
```

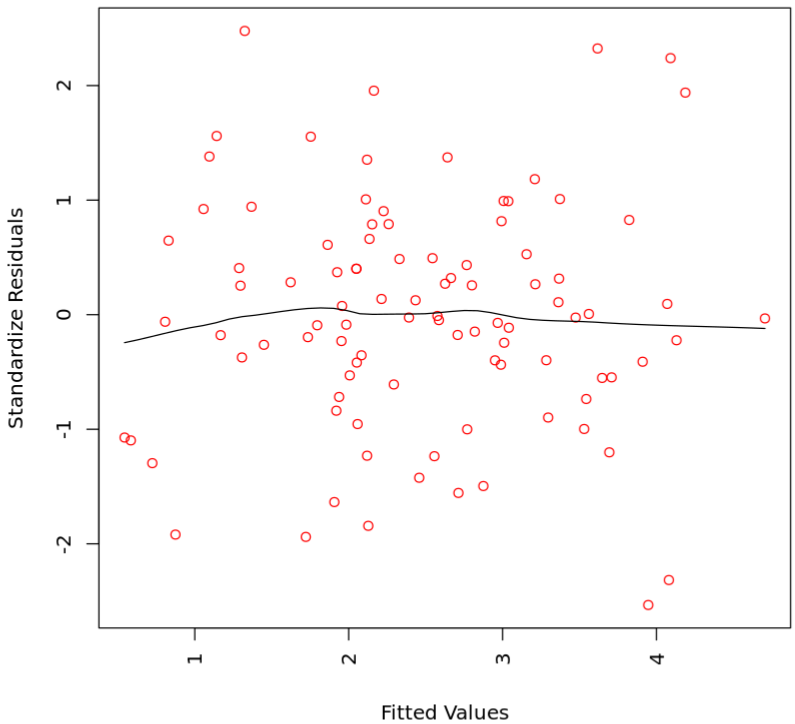
In [37]:

```
hist(model_pred_01, col='darkred')
```

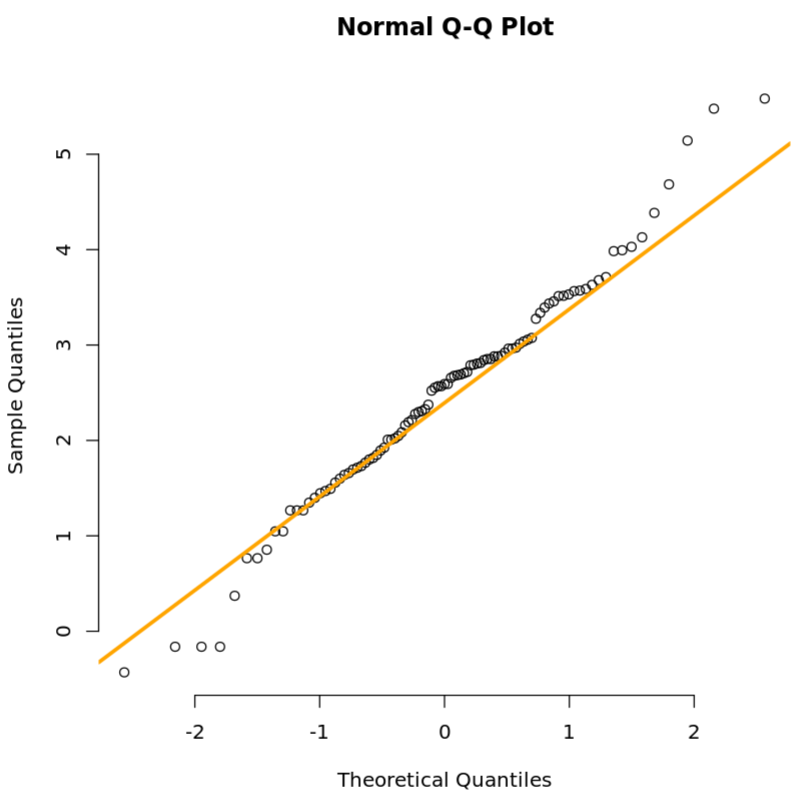


In [39]:

```
# Plot residuals against the fitted values
scatter.smooth(rstandard(model_03) ~ fitted(model_03), col="red",
               las=3, ylab="Standardize Residuals", xlab="Fitted Values")
```



In [40]:
```
qqnorm(prostate$lpsa, pch = 1, frame = FALSE)
qqline(prostate$lpsa, col = "orange", lwd = 3)
```

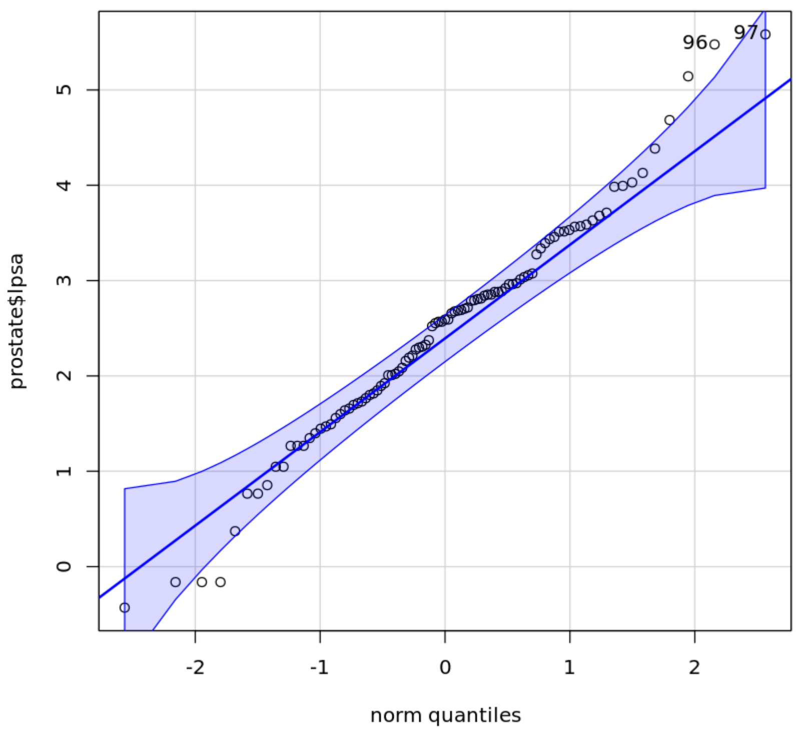


More or less points are closer to the straight line.

In [41]:

```
qqPlot(prostate$lpsa)
```

97 · 96



qqPlot function provides better visualization compared to previous one. With few exceptions, all points fall approximately along this straight line, so we can assume normality.

In [45]:
```
# Shapiro-Wilk normality test
shapiro.test(prostate$lpsa)
```

```
	Shapiro-Wilk normality test

data:  prostate$lpsa
W = 1, p-value = 0.3
```

The result is not significant, so we can assume the normality.

In [47]:
```
#Problem #4
data01 <- read.csv('Utilityweather-2.csv')
head(data01, 10)
```

A data.frame: 10 × 1

	Date	Value	AVG_TEMP	AVG_WIND	AVG_HUMID	CLOUD	PRESSURE	changeAVG_TEMP	changeA
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	1-Jan-14	4730	25.7	3.5	68.7	47.5	1027	0.0	
2	2-Jan-14	5603	24.5	7.7	81.9	87.1	1014	-1.2	
3	3-Jan-14	5898	10.9	11.3	63.1	37.3	1021	-13.6	

	Date	Value	AVG_TEMP	AVG_WIND	AVG_HUMID	CLOUD	PRESSURE	changeAVG_TEMP	changeA
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
4	4-Jan-14	5563	11.9	4.2	65.9	6.1	1030	1.0	
5	5-Jan-14	5108	22.6	3.3	85.4	54.5	1021	10.7	
6	6-Jan-14	5514	34.1	14.4	71.0	69.4	1004	11.5	
7	7-Jan-14	6616	3.1	14.9	47.7	11.8	1024	-31.0	
8	8-Jan-14	6326	12.4	5.3	55.4	38.3	1033	9.3	
9	9-Jan-14	5739	23.6	2.2	61.5	33.3	1036	11.2	
10	10-Jan-14	5311	30.3	2.7	84.6	89.7	1029	6.7	

In [48]:

```
str(data01)
```

```
'data.frame':	2102 obs. of  15 variables:
 $ Date           : chr  "1-Jan-14" "2-Jan-14" "3-Jan-14" "4-Jan-14" ...
 $ Value          : num  4730 5603 5898 5563 5108 ...
 $ AVG_TEMP       : num  25.7 24.5 10.9 11.9 22.6 34.1 3.1 12.4 23.6 30.3 ...
 $ AVG_WIND       : num  3.5 7.7 11.3 4.2 3.3 14.4 14.9 5.3 2.2 2.7 ...
 $ AVG_HUMID      : num  68.7 81.9 63.1 65.9 85.4 71 47.7 55.4 61.5 84.6 ...
 $ CLOUD          : num  47.5 87.1 37.3 6.1 54.5 69.4 11.8 38.3 33.3 89.7 ...
 $ PRESSURE       : num  1027 1014 1021 1030 1021 ...
 $ changeAVG_TEMP : num  0 -1.2 -13.6 1 10.7 11.5 -31 9.3 11.2 6.7 ...
 $ changeAVG_WIND : num  0 4.2 3.6 -7.1 -0.9 11.1 0.5 -9.6 -3.1 0.5 ...
 $ changeAVG_HUMID: num  0 13.2 -18.8 2.8 19.5 -14.4 -23.3 7.7 6.1 23.1 ...
 $ changeCLOUD    : num  0 39.6 -49.8 -31.2 48.4 14.9 -57.6 26.5 -5 56.4 ...
 $ changePRESSURE : num  0 -12.7 7.1 8.4 -9 -17.1 20 9.3 2.9 -6.3 ...
 $ Month          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Week           : int  1 1 1 1 2 2 2 2 2 2 ...
 $ Weeday         : int  4 5 6 7 1 2 3 4 5 6 ...
```

In [66]:

```
data02 <- head(data01, 90)
head(data02)
```

A data.frame: 6 × 15

	Date	Value	AVG_TEMP	AVG_WIND	AVG_HUMID	CLOUD	PRESSURE	changeAVG_TEMP	changeAV
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	1-Jan-14	4730	25.7	3.5	68.7	47.5	1027	0.0	
2	2-Jan-14	5603	24.5	7.7	81.9	87.1	1014	-1.2	
3	3-Jan-14	5898	10.9	11.3	63.1	37.3	1021	-13.6	

	Date	Value	AVG_TEMP	AVG_WIND	AVG_HUMID	CLOUD	PRESSURE	changeAVG_TEMP	changeAV
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
4	4-Jan-14	5563	11.9	4.2	65.9	6.1	1030	1.0	
5	5-Jan-14	5108	22.6	3.3	85.4	54.5	1021	10.7	
6	6-Jan-14	5514	34.1	14.4	71.0	69.4	1004	11.5	

In [67]:

```
# Explore the dataset (Utilityweather.csv) - perform exploratory analysis
glimpse(data02)# Number of observations (rows), variables, and a head of the first cases.
```

```
Rows: 90
Columns: 15
$ Date           <chr> "1-Jan-14", "2-Jan-14", "3-Jan-14", "4-Jan-14", "5-Jan…
$ Value          <dbl> 4730, 5603, 5898, 5563, 5108, 5514, 6616, 6326, 5739, …
$ AVG_TEMP       <dbl> 25.7, 24.5, 10.9, 11.9, 22.6, 34.1, 3.1, 12.4, 23.6, 3…
$ AVG_WIND       <dbl> 3.5, 7.7, 11.3, 4.2, 3.3, 14.4, 14.9, 5.3, 2.2, 2.7, 6…
$ AVG_HUMID      <dbl> 68.7, 81.9, 63.1, 65.9, 85.4, 71.0, 47.7, 55.4, 61.5, …
$ CLOUD          <dbl> 47.5, 87.1, 37.3, 6.1, 54.5, 69.4, 11.8, 38.3, 33.3, 8…
$ PRESSURE       <dbl> 1027, 1014, 1021, 1030, 1021, 1004, 1024, 1033, 1036, …
$ changeAVG_TEMP <dbl> 0.0, -1.2, -13.6, 1.0, 10.7, 11.5, -31.0, 9.3, 11.2, 6…
$ changeAVG_WIND <dbl> 0.0, 4.2, 3.6, -7.1, -0.9, 11.1, 0.5, -9.6, -3.1, 0.5,…
$ changeAVG_HUMID <dbl> 0.0, 13.2, -18.8, 2.8, 19.5, -14.4, -23.3, 7.7, 6.1, 2…
$ changeCLOUD    <dbl> 0.0, 39.6, -49.8, -31.2, 48.4, 14.9, -57.6, 26.5, -5.0…
$ changePRESSURE <dbl> 0.0, -12.7, 7.1, 8.4, -9.0, -17.1, 20.0, 9.3, 2.9, -6.…
$ Month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ Week           <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, …
$ Weeday         <int> 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, …
```

In [54]:

```
#install.packages("funModeling")
```

```
Installing package into '/home/mladenoffj/R_libs'
(as 'lib' is unspecified)

also installing the dependencies 'bitops', 'checkmate', 'gtools', 'caTools', 'Formula', 'htmlTable', 'gplots', 'Hmisc', 'ROCR', 'pander', 'entropy'
```

In [55]:

```
library(tidyverse)
library(funModeling)
```

```
Loading required package: Hmisc

Loading required package: survival


Attaching package: 'survival'


The following objects are masked from 'package:faraway':
```

```
    rats, solder


The following object is masked from ‘package:caret’:

    cluster


Loading required package: Formula


Attaching package: ‘Hmisc’


The following object is masked from ‘package:psych’:

    describe


The following objects are masked from ‘package:dplyr’:

    src, summarize


The following objects are masked from ‘package:base’:

    format.pval, units


funModeling v.1.9.4 :)
Examples and tutorials at livebook.datascienceheroes.com
 / Now in Spanish: librovivodecienciadedatos.ai
```

In [68]:

```
# Check the metrics about data types, zeros, infinite numbers, and missing values
status(data02)
```

A data.frame: 15 × 9

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
	<chr>	<int>	<dbl>	<int>	<dbl>	<int>	<dbl>	<chr>	<int>
Date	Date	0	0.0000	0	0	0	0	character	90
Value	Value	0	0.0000	0	0	0	0	numeric	90
AVG_TEMP	AVG_TEMP	0	0.0000	0	0	0	0	numeric	85
AVG_WIND	AVG_WIND	0	0.0000	0	0	0	0	numeric	60
AVG_HUMID	AVG_HUMID	0	0.0000	0	0	0	0	numeric	75
CLOUD	CLOUD	0	0.0000	0	0	0	0	numeric	83
PRESSURE	PRESSURE	0	0.0000	0	0	0	0	numeric	78
changeAVG_TEMP	changeAVG_TEMP	1	0.0111	0	0	0	0	numeric	81
changeAVG_WIND	changeAVG_WIND	2	0.0222	0	0	0	0	numeric	63
changeAVG_HUMID	changeAVG_HUMID	3	0.0333	0	0	0	0	numeric	79
changeCLOUD	changeCLOUD	1	0.0111	0	0	0	0	numeric	87

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
	<chr>	<int>	<dbl>	<int>	<dbl>	<int>	<dbl>	<chr>	<int>
changePRESSURE	changePRESSURE	1	0.0111	0	0	0	0	numeric	83
Month	Month	0	0.0000	0	0	0	0	integer	3
Week	Week	0	0.0000	0	0	0	0	integer	14
Weeday	Weeday	0	0.0000	0	0	0	0	integer	7

In [69]:

```
# Profiling the Data Input
data01_status=df_status(data02, print_results = F)
data01_status
```

A data.frame: 15 × 9

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
<chr>	<int>	<dbl>	<int>	<dbl>	<int>	<dbl>	<chr>	<int>
Date	0	0.00	0	0	0	0	character	90
Value	0	0.00	0	0	0	0	numeric	90
AVG_TEMP	0	0.00	0	0	0	0	numeric	85
AVG_WIND	0	0.00	0	0	0	0	numeric	60
AVG_HUMID	0	0.00	0	0	0	0	numeric	75
CLOUD	0	0.00	0	0	0	0	numeric	83
PRESSURE	0	0.00	0	0	0	0	numeric	78
changeAVG_TEMP	1	1.11	0	0	0	0	numeric	81
changeAVG_WIND	2	2.22	0	0	0	0	numeric	63
changeAVG_HUMID	3	3.33	0	0	0	0	numeric	79
changeCLOUD	1	1.11	0	0	0	0	numeric	87
changePRESSURE	1	1.11	0	0	0	0	numeric	83
Month	0	0.00	0	0	0	0	integer	3
Week	0	0.00	0	0	0	0	integer	14
Weeday	0	0.00	0	0	0	0	integer	7

In [70]:

```
freq(data=data02, input = c('changeAVG_TEMP','changeAVG_WIND', 'changeAVG_HUMID', 'change
```

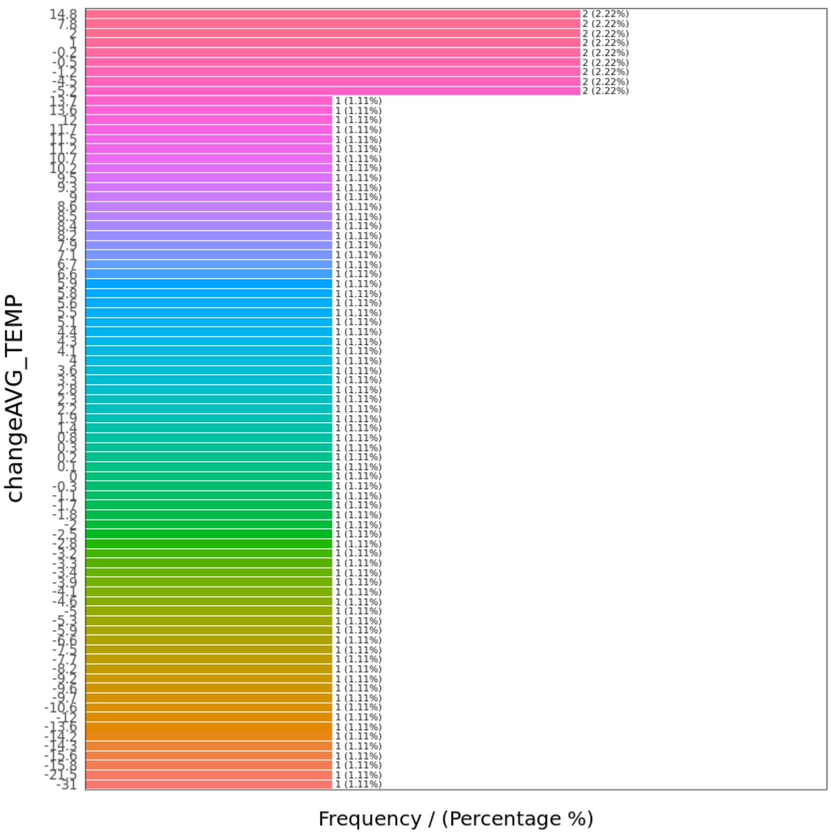
```
Warning message:
“`guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.”
   changeAVG_TEMP frequency percentage cumulative_perc
1            -5.2         2       2.22            2.22
2            -4.5         2       2.22            4.44
3            -1.2         2       2.22            6.66
4            -0.5         2       2.22            8.88
5            -0.2         2       2.22           11.10
6               1         2       2.22           13.32
```

7	2	2	2.22	15.54
8	7.8	2	2.22	17.76
9	14.8	2	2.22	19.98
10	-31	1	1.11	21.09
11	-21.5	1	1.11	22.20
12	-15.8	1	1.11	23.31
13	-15.6	1	1.11	24.42
14	-14.3	1	1.11	25.53
15	-14.2	1	1.11	26.64
16	-13.6	1	1.11	27.75
17	-12	1	1.11	28.86
18	-10.6	1	1.11	29.97
19	-9.7	1	1.11	31.08
20	-9.6	1	1.11	32.19
21	-9.2	1	1.11	33.30
22	-8.2	1	1.11	34.41
23	-7.7	1	1.11	35.52
24	-7.5	1	1.11	36.63
25	-6.6	1	1.11	37.74
26	-5.9	1	1.11	38.85
27	-5.3	1	1.11	39.96
28	-5	1	1.11	41.07
29	-4.6	1	1.11	42.18
30	-4.1	1	1.11	43.29
31	-3.9	1	1.11	44.40
32	-3.4	1	1.11	45.51
33	-3.3	1	1.11	46.62
34	-3.2	1	1.11	47.73
35	-2.8	1	1.11	48.84
36	-2.5	1	1.11	49.95
37	-2	1	1.11	51.06
38	-1.8	1	1.11	52.17
39	-1.7	1	1.11	53.28
40	-1.1	1	1.11	54.39
41	-0.3	1	1.11	55.50
42	0	1	1.11	56.61
43	0.1	1	1.11	57.72
44	0.2	1	1.11	58.83
45	0.3	1	1.11	59.94
46	0.8	1	1.11	61.05
47	1.4	1	1.11	62.16
48	1.9	1	1.11	63.27
49	2.2	1	1.11	64.38
50	2.3	1	1.11	65.49
51	2.8	1	1.11	66.60
52	3.3	1	1.11	67.71
53	3.6	1	1.11	68.82
54	4	1	1.11	69.93
55	4.1	1	1.11	71.04
56	4.3	1	1.11	72.15
57	4.4	1	1.11	73.26
58	5.1	1	1.11	74.37
59	5.5	1	1.11	75.48
60	5.6	1	1.11	76.59
61	5.8	1	1.11	77.70
62	5.9	1	1.11	78.81
63	6.6	1	1.11	79.92
64	6.7	1	1.11	81.03
65	7.1	1	1.11	82.14
66	7.9	1	1.11	83.25
67	8.2	1	1.11	84.36

```
68            8.4         1       1.11            85.47
69            8.5         1       1.11            86.58
70            8.6         1       1.11            87.69
71              9         1       1.11            88.80
72            9.3         1       1.11            89.91
73            9.5         1       1.11            91.02
74           10.2         1       1.11            92.13
75           10.7         1       1.11            93.24
76           11.2         1       1.11            94.35
77           11.5         1       1.11            95.46
78           11.7         1       1.11            96.57
79             12         1       1.11            97.68
80           13.6         1       1.11            98.79
81           13.7         1       1.11           100.00

Warning message:
“`guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.”
   changeAVG_WIND frequency percentage cumulative_perc
1             0.1         4       4.44             4.44
2            -0.9         3       3.33             7.77
3            -7.1         2       2.22             9.99
4            -3.5         2       2.22            12.21
5            -3.1         2       2.22            14.43
6              -3         2       2.22            16.65
7            -2.2         2       2.22            18.87
8            -1.1         2       2.22            21.09
9              -1         2       2.22            23.31
10           -0.7         2       2.22            25.53
11           -0.6         2       2.22            27.75
12           -0.5         2       2.22            29.97
13              0         2       2.22            32.19
14            0.2         2       2.22            34.41
15            0.4         2       2.22            36.63
16            0.5         2       2.22            38.85
17            0.7         2       2.22            41.07
18            0.9         2       2.22            43.29
19              1         2       2.22            45.51
20            1.1         2       2.22            47.73
21            2.6         2       2.22            49.95
22            2.7         2       2.22            52.17
23            3.6         2       2.22            54.39
24            4.1         2       2.22            56.61
25          -10.6         1       1.11            57.72
26           -9.6         1       1.11            58.83
27           -7.3         1       1.11            59.94
28           -6.2         1       1.11            61.05
29           -6.1         1       1.11            62.16
30           -5.2         1       1.11            63.27
31             -5         1       1.11            64.38
32           -3.6         1       1.11            65.49
33           -3.3         1       1.11            66.60
34           -2.9         1       1.11            67.71
35           -2.8         1       1.11            68.82
36           -2.7         1       1.11            69.93
37           -2.5         1       1.11            71.04
38           -2.3         1       1.11            72.15
39             -2         1       1.11            73.26
40           -1.7         1       1.11            74.37
41           -1.5         1       1.11            75.48
42           -0.4         1       1.11            76.59
```

43	-0.2	1	1.11	77.70
44	0.3	1	1.11	78.81
45	0.8	1	1.11	79.92
46	1.4	1	1.11	81.03
47	1.9	1	1.11	82.14
48	2.3	1	1.11	83.25
49	2.4	1	1.11	84.36
50	2.8	1	1.11	85.47
51	3.5	1	1.11	86.58
52	4.2	1	1.11	87.69
53	4.4	1	1.11	88.80
54	4.7	1	1.11	89.91
55	4.9	1	1.11	91.02
56	5.3	1	1.11	92.13
57	5.4	1	1.11	93.24
58	6.8	1	1.11	94.35
59	7.3	1	1.11	95.46
60	7.5	1	1.11	96.57
61	10.4	1	1.11	97.68
62	10.9	1	1.11	98.79
63	11.1	1	1.11	100.00

```
Warning message:
“`guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.”
```

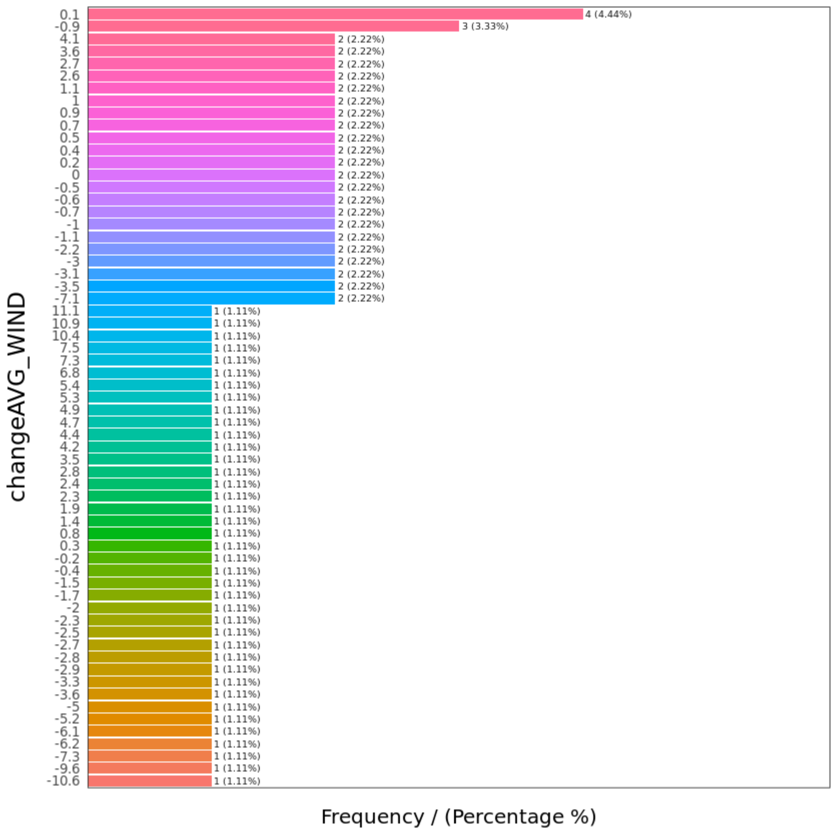


	changeAVG_HUMID	frequency	percentage	cumulative_perc
1	-1.3	3	3.33	3.33
2	0	3	3.33	6.66
3	6.1	3	3.33	9.99
4	-18.8	2	2.22	12.21
5	-15.6	2	2.22	14.43
6	-4.9	2	2.22	16.65
7	2.8	2	2.22	18.87
8	5.3	2	2.22	21.09
9	-34.8	1	1.11	22.20

10	-31	1	1.11	23.31
11	-26.4	1	1.11	24.42
12	-26.2	1	1.11	25.53
13	-23.3	1	1.11	26.64
14	-22.6	1	1.11	27.75
15	-20.7	1	1.11	28.86
16	-19.1	1	1.11	29.97
17	-17.2	1	1.11	31.08
18	-15.8	1	1.11	32.19
19	-14.7	1	1.11	33.30
20	-14.6	1	1.11	34.41
21	-14.4	1	1.11	35.52
22	-12.9	1	1.11	36.63
23	-12.4	1	1.11	37.74
24	-10.4	1	1.11	38.85
25	-10.2	1	1.11	39.96
26	-10	1	1.11	41.07
27	-9.6	1	1.11	42.18
28	-9.4	1	1.11	43.29
29	-8.7	1	1.11	44.40
30	-8.4	1	1.11	45.51
31	-8.3	1	1.11	46.62
32	-6.9	1	1.11	47.73
33	-5.7	1	1.11	48.84
34	-5.2	1	1.11	49.95
35	-4.8	1	1.11	51.06
36	-4.7	1	1.11	52.17
37	-3.5	1	1.11	53.28
38	-2.7	1	1.11	54.39
39	-1.7	1	1.11	55.50
40	-1.6	1	1.11	56.61
41	-1.4	1	1.11	57.72
42	-1	1	1.11	58.83
43	-0.6	1	1.11	59.94
44	-0.3	1	1.11	61.05
45	-0.1	1	1.11	62.16
46	0.5	1	1.11	63.27
47	1	1	1.11	64.38
48	1.7	1	1.11	65.49
49	3.9	1	1.11	66.60
50	4.2	1	1.11	67.71
51	4.9	1	1.11	68.82
52	5	1	1.11	69.93
53	5.7	1	1.11	71.04
54	6.7	1	1.11	72.15
55	7.7	1	1.11	73.26
56	8.2	1	1.11	74.37
57	8.9	1	1.11	75.48
58	9.6	1	1.11	76.59
59	10.4	1	1.11	77.70
60	10.6	1	1.11	78.81
61	11.6	1	1.11	79.92
62	12.4	1	1.11	81.03
63	12.7	1	1.11	82.14
64	13.2	1	1.11	83.25
65	14.4	1	1.11	84.36
66	14.7	1	1.11	85.47
67	15.5	1	1.11	86.58
68	15.6	1	1.11	87.69
69	16.6	1	1.11	88.80
70	18.2	1	1.11	89.91

71	19.5	1	1.11	91.02
72	20.2	1	1.11	92.13
73	21.1	1	1.11	93.24
74	22.7	1	1.11	94.35
75	23.1	1	1.11	95.46
76	25.4	1	1.11	96.57
77	26.4	1	1.11	97.68
78	32.5	1	1.11	98.79
79	37	1	1.11	100.00

```
Warning message:
“`guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.”
```

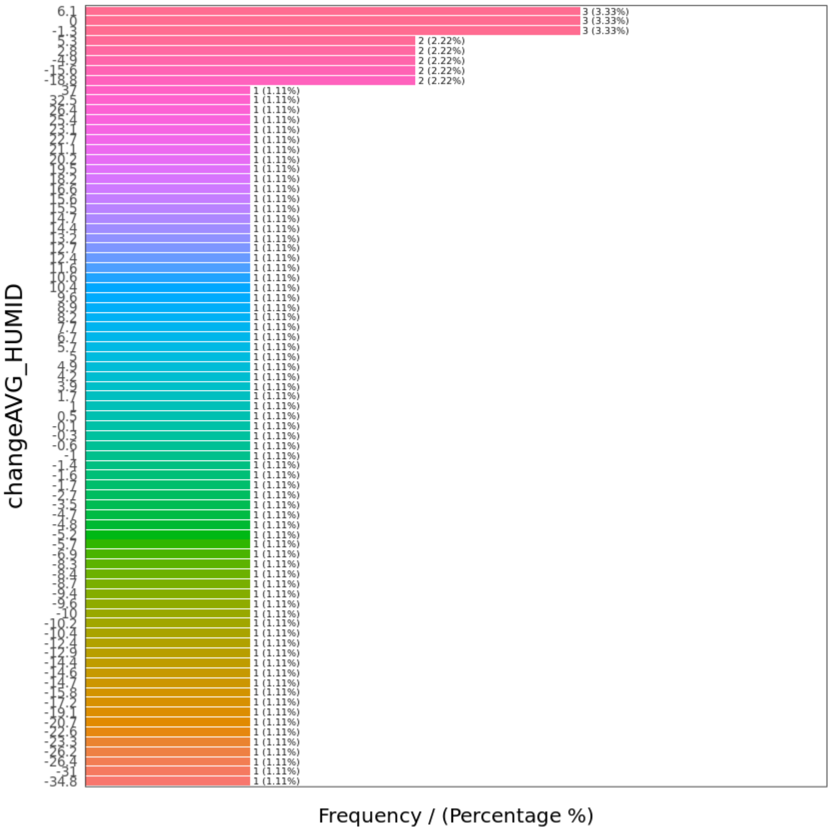


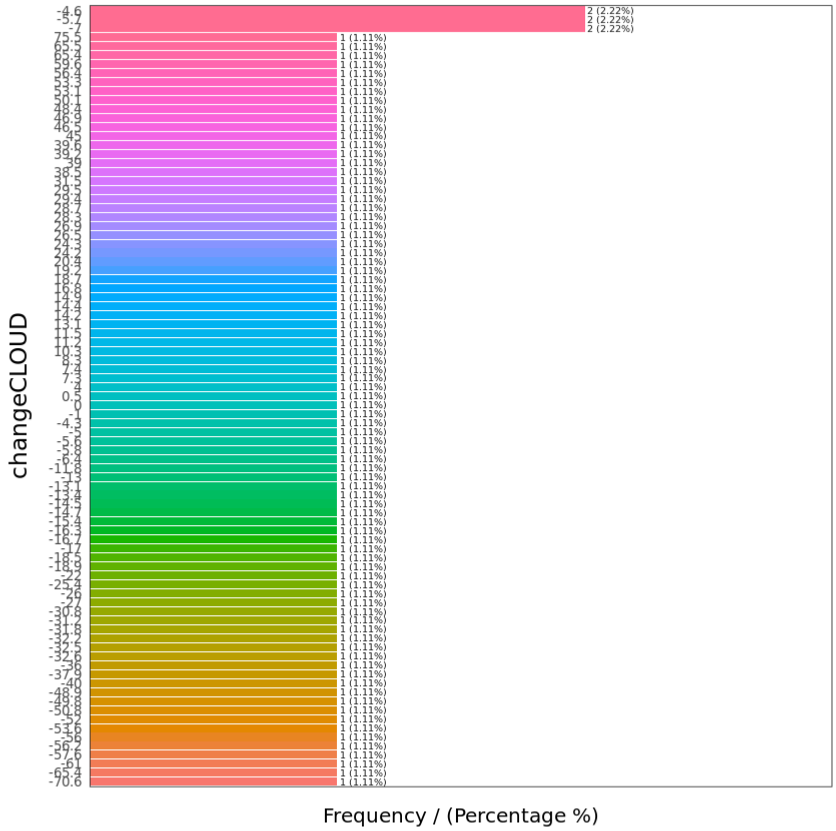
	changeCLOUD	frequency	percentage	cumulative_perc
1	-7	2	2.22	2.22
2	-5.7	2	2.22	4.44
3	-4.6	2	2.22	6.66
4	-70.6	1	1.11	7.77
5	-65.4	1	1.11	8.88
6	-61	1	1.11	9.99
7	-57.6	1	1.11	11.10
8	-56.2	1	1.11	12.21
9	-56	1	1.11	13.32
10	-53.6	1	1.11	14.43
11	-52	1	1.11	15.54
12	-50.8	1	1.11	16.65
13	-49.8	1	1.11	17.76
14	-48.9	1	1.11	18.87
15	-40	1	1.11	19.98
16	-37.9	1	1.11	21.09
17	-36	1	1.11	22.20
18	-32.6	1	1.11	23.31
19	-32.5	1	1.11	24.42
20	-32.2	1	1.11	25.53
21	-31.8	1	1.11	26.64

22	-31.2	1	1.11	27.75
23	-30.8	1	1.11	28.86
24	-27	1	1.11	29.97
25	-26	1	1.11	31.08
26	-25.4	1	1.11	32.19
27	-22	1	1.11	33.30
28	-18.9	1	1.11	34.41
29	-18.5	1	1.11	35.52
30	-17	1	1.11	36.63
31	-16.7	1	1.11	37.74
32	-16.3	1	1.11	38.85
33	-15.4	1	1.11	39.96
34	-14.7	1	1.11	41.07
35	-14.5	1	1.11	42.18
36	-13.4	1	1.11	43.29
37	-13.1	1	1.11	44.40
38	-13	1	1.11	45.51
39	-11.8	1	1.11	46.62
40	-6.4	1	1.11	47.73
41	-5.8	1	1.11	48.84
42	-5.6	1	1.11	49.95
43	-5	1	1.11	51.06
44	-4.3	1	1.11	52.17
45	-1	1	1.11	53.28
46	0	1	1.11	54.39
47	0.5	1	1.11	55.50
48	4	1	1.11	56.61
49	7.3	1	1.11	57.72
50	7.4	1	1.11	58.83
51	8.3	1	1.11	59.94
52	10.3	1	1.11	61.05
53	11.2	1	1.11	62.16
54	11.5	1	1.11	63.27
55	13.1	1	1.11	64.38
56	14.2	1	1.11	65.49
57	14.4	1	1.11	66.60
58	14.9	1	1.11	67.71
59	16.8	1	1.11	68.82
60	18.7	1	1.11	69.93
61	19.2	1	1.11	71.04
62	20.4	1	1.11	72.15
63	24.2	1	1.11	73.26
64	24.3	1	1.11	74.37
65	26.5	1	1.11	75.48
66	26.9	1	1.11	76.59
67	28.3	1	1.11	77.70
68	28.7	1	1.11	78.81
69	29.4	1	1.11	79.92
70	29.5	1	1.11	81.03
71	31.5	1	1.11	82.14
72	38.5	1	1.11	83.25
73	39	1	1.11	84.36
74	39.2	1	1.11	85.47
75	39.6	1	1.11	86.58
76	45	1	1.11	87.69
77	46.5	1	1.11	88.80
78	46.9	1	1.11	89.91
79	48.4	1	1.11	91.02
80	50.1	1	1.11	92.13
81	53.1	1	1.11	93.24
82	53.3	1	1.11	94.35

83	56.4	1	1.11	95.46
84	59.6	1	1.11	96.57
85	65.4	1	1.11	97.68
86	65.5	1	1.11	98.79
87	75.5	1	1.11	100.00

```
Warning message:
“`guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.”
```

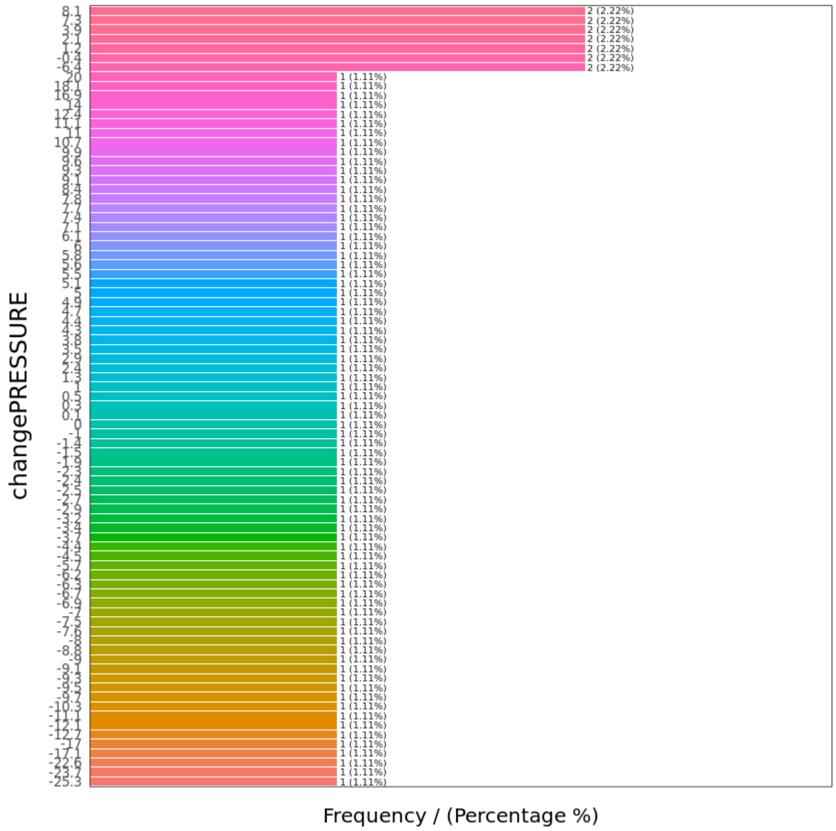




	changePRESSURE	frequency	percentage	cumulative_perc
1	-6.4	2	2.22	2.22
2	-0.4	2	2.22	4.44
3	1.2	2	2.22	6.66
4	2.1	2	2.22	8.88
5	3.9	2	2.22	11.10
6	7.3	2	2.22	13.32
7	8.1	2	2.22	15.54
8	-25.3	1	1.11	16.65
9	-23.7	1	1.11	17.76
10	-22.6	1	1.11	18.87
11	-17.1	1	1.11	19.98
12	-17	1	1.11	21.09
13	-12.7	1	1.11	22.20
14	-12.1	1	1.11	23.31
15	-11.1	1	1.11	24.42
16	-10.3	1	1.11	25.53
17	-9.7	1	1.11	26.64
18	-9.5	1	1.11	27.75
19	-9.3	1	1.11	28.86
20	-9.1	1	1.11	29.97
21	-9	1	1.11	31.08
22	-8.8	1	1.11	32.19
23	-8	1	1.11	33.30
24	-7.6	1	1.11	34.41
25	-7.5	1	1.11	35.52
26	-7	1	1.11	36.63
27	-6.9	1	1.11	37.74
28	-6.7	1	1.11	38.85
29	-6.3	1	1.11	39.96
30	-6.2	1	1.11	41.07
31	-5.7	1	1.11	42.18
32	-4.5	1	1.11	43.29
33	-4.4	1	1.11	44.40

34	-3.7	1	1.11	45.51
35	-3.4	1	1.11	46.62
36	-3.2	1	1.11	47.73
37	-2.9	1	1.11	48.84
38	-2.7	1	1.11	49.95
39	-2.5	1	1.11	51.06
40	-2.4	1	1.11	52.17
41	-2.3	1	1.11	53.28
42	-1.9	1	1.11	54.39
43	-1.5	1	1.11	55.50
44	-1.4	1	1.11	56.61
45	-1	1	1.11	57.72
46	0	1	1.11	58.83
47	0.1	1	1.11	59.94
48	0.3	1	1.11	61.05
49	0.5	1	1.11	62.16
50	1	1	1.11	63.27
51	1.3	1	1.11	64.38
52	2.4	1	1.11	65.49
53	2.9	1	1.11	66.60
54	3.5	1	1.11	67.71
55	3.8	1	1.11	68.82
56	4.3	1	1.11	69.93
57	4.4	1	1.11	71.04
58	4.7	1	1.11	72.15
59	4.9	1	1.11	73.26
60	5	1	1.11	74.37
61	5.1	1	1.11	75.48
62	5.5	1	1.11	76.59
63	5.6	1	1.11	77.70
64	5.8	1	1.11	78.81
65	6	1	1.11	79.92
66	6.1	1	1.11	81.03
67	7.1	1	1.11	82.14
68	7.4	1	1.11	83.25
69	7.7	1	1.11	84.36
70	7.8	1	1.11	85.47
71	8.4	1	1.11	86.58
72	9.1	1	1.11	87.69
73	9.3	1	1.11	88.80
74	9.6	1	1.11	89.91
75	9.9	1	1.11	91.02
76	10.7	1	1.11	92.13
77	11	1	1.11	93.24
78	11.1	1	1.11	94.35
79	12.4	1	1.11	95.46
80	14	1	1.11	96.57
81	16.9	1	1.11	97.68
82	18.1	1	1.11	98.79
83	20	1	1.11	100.00

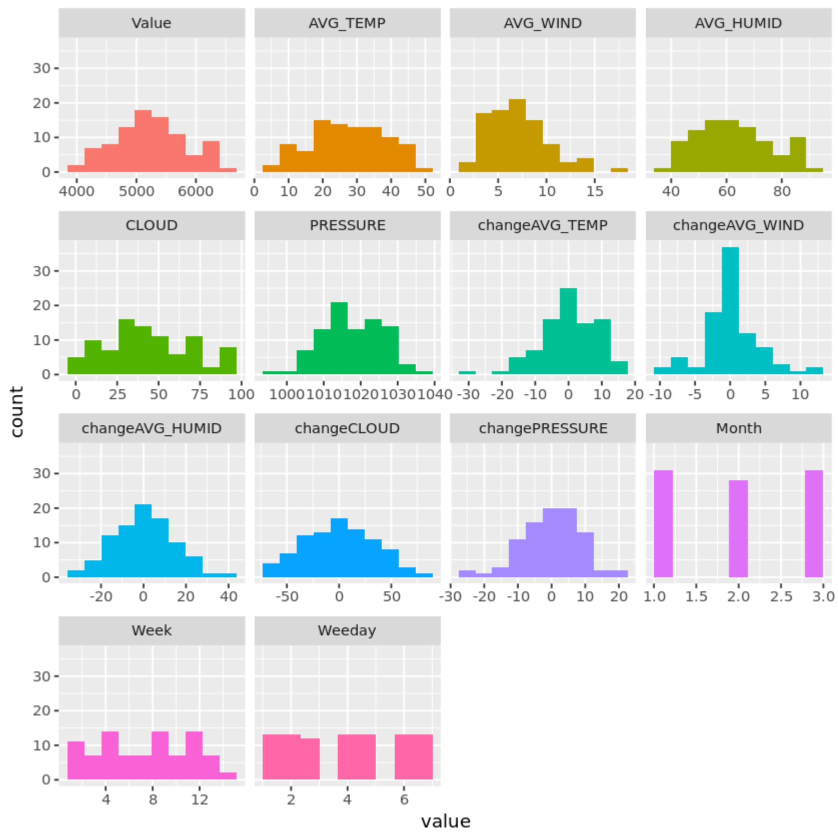
'Variables processed: changeAVG_TEMP, changeAVG_WIND, changeAVG_HUMID, changeCLOUD, changePRESSURE'



In [71]:

```
# Analyzing numerical variables
plot_num(data02) # Graphically
```

```
Warning message:
“`guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.”
```



In [73]:
```
data02_prof=profiling_num(data02) # Quantitatively
data02_prof
```

A data.frame: 14 × 16

variable	mean	std_dev	variation_coef	p_01	p_05	p_25	p_50	p_75	p_95
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Value	5226.3614	591.229	1.13e-01	4063.802	4291.4	4763.48	5233.94	5596.51	6226.32
AVG_TEMP	27.6278	10.543	3.82e-01	5.681	11.1	19.77	27.35	35.50	44.34
AVG_WIND	7.0400	3.218	4.57e-01	2.378	2.8	4.53	6.60	8.80	13.28
AVG_HUMID	63.1678	13.605	2.15e-01	41.047	44.4	52.73	62.35	72.38	87.90
CLOUD	44.0400	25.495	5.79e-01	0.645	5.5	26.38	39.55	61.12	89.25
PRESSURE	1017.5967	8.313	8.17e-03	998.939	1003.9	1012.32	1017.15	1023.65	1029.60
changeAVG_TEMP	0.2000	8.395	4.20e+01	-22.545	-14.3	-4.50	0.25	6.42	11.86
changeAVG_WIND	0.1122	3.985	3.55e+01	-9.710	-6.7	-2.20	0.10	2.38	7.07
changeAVG_HUMID	-0.0833	14.258	-1.71e+02	-31.418	-23.0	-9.55	-0.20	8.73	22.92
changeCLOUD	-0.1089	34.434	-3.16e+02	-65.972	-56.1	-24.55	-4.60	25.95	55.00
changePRESSURE	-0.1300	8.853	-6.81e+01	-23.876	-15.1	-6.38	0.40	5.95	11.81
Month	2.0000	0.835	4.17e-01	1.000	1.0	1.00	2.00	3.00	3.00
Week	7.3556	3.752	5.10e-01	1.000	2.0	4.00	7.00	10.75	13.00
Weeday	4.0111	2.020	5.03e-01	1.000	1.0	2.00	4.00	6.00	7.00

In [78]:
```
data02_prof %>% select(variable, variation_coef, range_98)
```

```
Error in select(., variable, variation_coef, range_98): unused arguments (variable, varia
tion_coef, range_98)
Traceback:

1. data02_prof %>% select(variable, variation_coef, range_98)
```

In [79]:
```
library(Hmisc)
```

In [80]:
```
# Analyzing numerical and categorical at the same time
describe(data02)
```

```
data02

 15  Variables      90  Observations
--------------------------------------------------------------------------------
Date
       n  missing distinct 
      90        0       90 
```

```
lowest : 1-Feb-14  1-Jan-14  1-Mar-14  10-Feb-14 10-Jan-14
highest: 8-Jan-14  8-Mar-14  9-Feb-14  9-Jan-14  9-Mar-14
--------------------------------------------------------------------------------
Value
       n  missing distinct     Info     Mean      Gmd      .05      .10
      90        0       90        1     5226      678     4291     4427
     .25      .50      .75      .90      .95
    4763     5234     5597     6162     6226

lowest : 4056 4065 4168 4199 4231, highest: 6230 6245 6309 6326 6616
--------------------------------------------------------------------------------
AVG_TEMP
       n  missing distinct     Info     Mean      Gmd      .05      .10
      90        0       85        1    27.63    12.16    11.12    12.26
     .25      .50      .75      .90      .95
   19.77    27.35    35.50    41.85    44.34

lowest :  3.1  6.0  9.8 10.4 10.9, highest: 44.7 45.1 45.2 46.6 47.7
--------------------------------------------------------------------------------
AVG_WIND
       n  missing distinct     Info     Mean      Gmd      .05      .10
      90        0       60        1     7.04    3.581    2.800    3.300
     .25      .50      .75      .90      .95
   4.525    6.600    8.800   11.300   13.285

lowest :  2.2  2.4  2.5  2.7  2.8, highest: 13.6 13.8 14.4 14.9 18.0
--------------------------------------------------------------------------------
AVG_HUMID
       n  missing distinct     Info     Mean      Gmd      .05      .10
      90        0       75        1    63.17     15.6    44.43    45.89
     .25      .50      .75      .90      .95
   52.73    62.35    72.38    84.61    87.90

lowest : 39.0 41.3 41.4 43.5 43.8, highest: 85.4 87.4 87.9 88.7 94.2
--------------------------------------------------------------------------------
CLOUD
       n  missing distinct     Info     Mean      Gmd      .05      .10
      90        0       83        1    44.04    29.32    5.495   11.720
     .25      .50      .75      .90      .95
  26.375   39.550   61.125   81.110   89.250

lowest :  0.2  0.7  1.8  4.2  5.0, highest: 88.7 89.7 90.1 91.3 92.2
--------------------------------------------------------------------------------
PRESSURE
       n  missing distinct     Info     Mean      Gmd      .05      .10
      90        0       78        1     1018    9.503     1004     1007
     .25      .50      .75      .90      .95
    1012     1017     1024     1028     1030

lowest :  994 1000 1004 1004 1004, highest: 1030 1032 1033 1034 1036
--------------------------------------------------------------------------------
changeAVG_TEMP
       n  missing distinct     Info     Mean      Gmd      .05      .10
      90        0       81        1      0.2    9.314  -14.255   -9.790
     .25      .50      .75      .90      .95
  -4.500    0.250    6.425   10.250   11.865

lowest : -31.0 -21.5 -15.8 -15.6 -14.3, highest:  11.7  12.0  13.6  13.7  14.8
--------------------------------------------------------------------------------
changeAVG_WIND
```

```
       n  missing distinct     Info     Mean      Gmd      .05      .10 
      90        0       63        1   0.1122    4.349   -6.695   -3.740 
     .25      .50      .75      .90      .95 
  -2.200    0.100    2.375    4.720    7.075 

lowest : -10.6  -9.6  -7.3  -7.1  -6.2, highest:   7.3   7.5  10.4  10.9  11.1
--------------------------------------------------------------------------------
changeAVG_HUMID 
       n  missing distinct     Info     Mean      Gmd      .05      .10 
      90        0       79        1 -0.08333    16.16  -22.985  -18.800 
     .25      .50      .75      .90      .95 
  -9.550   -0.200    8.725   18.330   22.920 

lowest : -34.8 -31.0 -26.4 -26.2 -23.3, highest:  23.1  25.4  26.4  32.5  37.0
--------------------------------------------------------------------------------
changeCLOUD 
       n  missing distinct     Info     Mean      Gmd      .05      .10 
      90        0       87        1  -0.1089    39.58   -56.11   -49.90 
     .25      .50      .75      .90      .95 
  -24.55    -4.60    25.95    47.05    55.00 

lowest : -70.6 -65.4 -61.0 -57.6 -56.2, highest:  56.4  59.6  65.4  65.5  75.5
--------------------------------------------------------------------------------
changePRESSURE 
       n  missing distinct     Info     Mean      Gmd      .05      .10 
      90        0       83        1    -0.13    9.916  -15.065   -9.760 
     .25      .50      .75      .90      .95 
  -6.375    0.400    5.950    9.630   11.815 

lowest : -25.3 -23.7 -22.6 -17.1 -17.0, highest:  12.4  14.0  16.9  18.1  20.0
--------------------------------------------------------------------------------
Month 
       n  missing distinct     Info     Mean      Gmd 
      90        0        3    0.888        2   0.9134 
                            
Value          1     2     3
Frequency     31    28    31
Proportion 0.344 0.311 0.344
--------------------------------------------------------------------------------
Week 
       n  missing distinct     Info     Mean      Gmd      .05      .10 
      90        0       14    0.994    7.356    4.343     2.00     2.00 
     .25      .50      .75      .90      .95 
    4.00     7.00    10.75    12.10    13.00 

lowest :  1  2  3  4  5, highest: 10 11 12 13 14
                                                                            
Value          1     2     3     4     5     6     7     8     9    10    11
Frequency      4     7     7     7     7     7     7     7     7     7     7
Proportion 0.044 0.078 0.078 0.078 0.078 0.078 0.078 0.078 0.078 0.078 0.078
                            
Value         12    13    14
Frequency      7     7     2
Proportion 0.078 0.078 0.022
--------------------------------------------------------------------------------
Weeday 
       n  missing distinct     Info     Mean      Gmd 
      90        0        7     0.98    4.011    2.321 

lowest : 1 2 3 4 5, highest: 3 4 5 6 7
```

```
Value          1     2     3     4     5     6     7
Frequency     13    13    12    13    13    13    13
Proportion 0.144 0.144 0.133 0.144 0.144 0.144 0.144
--------------------------------------------------------------------------------
```

In [82]:

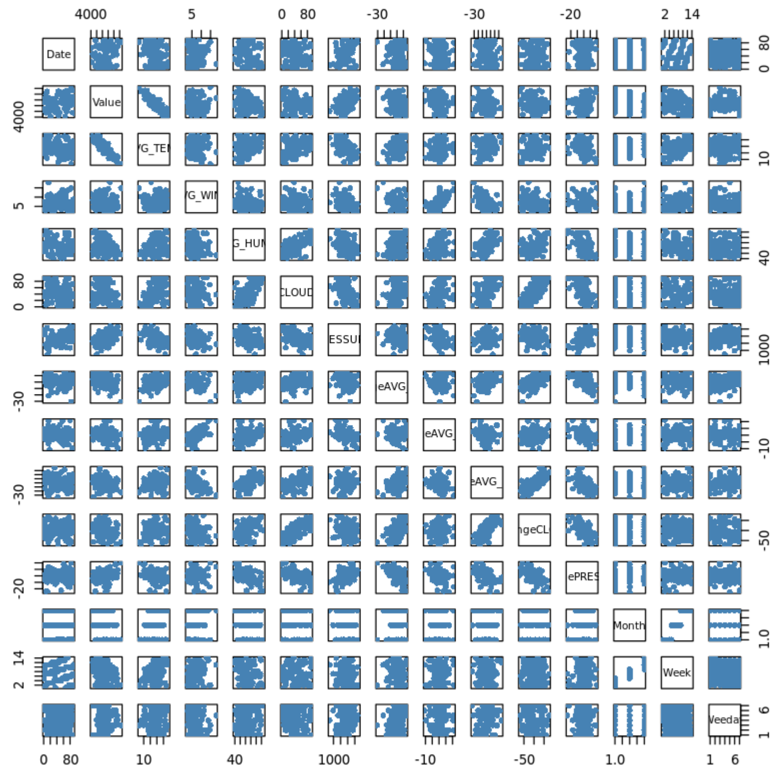
```
# Fit a GLM regressor - which variables are appropriate for this model?
model_data02 <- lm(AVG_TEMP~changeAVG_TEMP + changeAVG_WIND + changeAVG_HUMID + changeCLO
summary(model_data02)
model_data02
```

```
Call:
lm(formula = AVG_TEMP ~ changeAVG_TEMP + changeAVG_WIND + changeAVG_HUMID + 
    changeCLOUD + changePRESSURE, data = data02)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.836  -6.919  -0.623   7.587  16.670 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      27.4672     1.0039   27.36   <2e-16 ***
changeAVG_TEMP    0.4292     0.1671    2.57    0.012 *  
changeAVG_WIND    0.3843     0.3381    1.14    0.259    
changeAVG_HUMID   0.0101     0.1208    0.08    0.933    
changeCLOUD      -0.0239     0.0482   -0.50    0.622    
changePRESSURE   -0.2297     0.2043   -1.12    0.264    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.51 on 84 degrees of freedom
Multiple R-squared:  0.231,	Adjusted R-squared:  0.186 
F-statistic: 5.06 on 5 and 84 DF,  p-value: 0.000424
Call:
lm(formula = AVG_TEMP ~ changeAVG_TEMP + changeAVG_WIND + changeAVG_HUMID + 
    changeCLOUD + changePRESSURE, data = data02)

Coefficients:
    (Intercept)   changeAVG_TEMP   changeAVG_WIND  changeAVG_HUMID  
        27.4672           0.4292           0.3843           0.0101  
    changeCLOUD   changePRESSURE  
        -0.0239          -0.2297  
```

AVG_TEMP = (27.4672) + (0.4292)*changeAVG_TEMP + (0.3843)*changeAVG_WIND + (0.0101)*changeAVG_HUMID + (-0.0239)*changeCLOUD + (-0.2297)*changePRESSURE

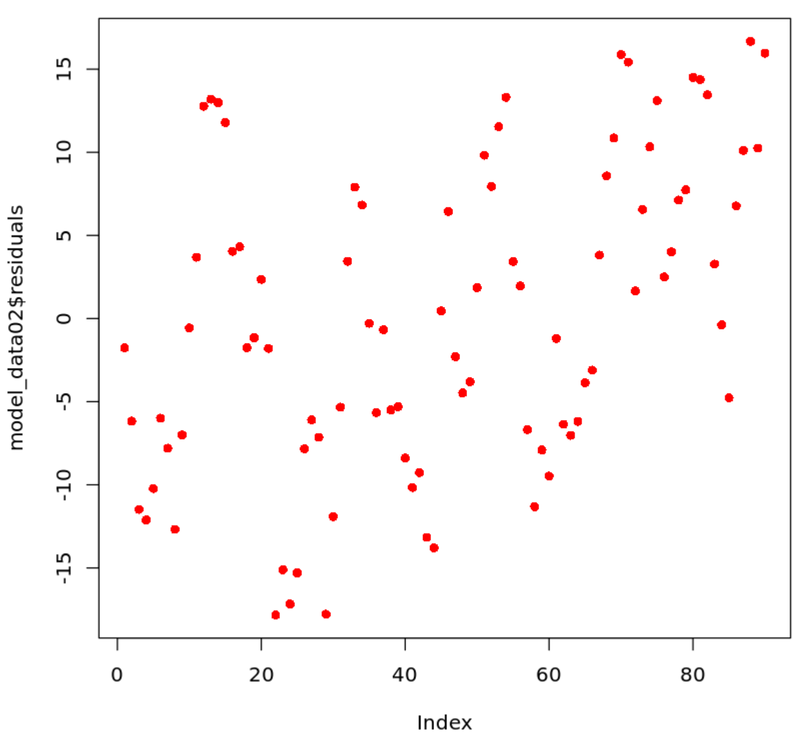
In [88]:

```
plot(data02, pch = 16, col = "steelblue")
abline(data02)
```



In [86]:

```
#Discuss if a linear regressor is appropriate for this problem?
plot(model_data02$residuals, pch = 16, col = "red")
```



From the plot of the residuals it is clear that they look random. Otherwise means that maybe there is a hidden pattern that the linear model is not considering. On the plot above there are no clear patterns presented, so the linear regression is appropriate for this model.

In [ ]: