```
## Lab_7_20122065
```

In [5]:

```
txt = read.delim('startup profits.txt',sep = ",")
txt
```

R.D.Spend	Administration	Marketing.Spend	State	Profit
165349.20	136897.80	471784.1	New York	192261.8
162597.70	151377.59	443898.5	California	191792.1
153441.51	101145.55	407934.5	Florida	191050.4
144372.41	118671.85	383199.6	New York	182902.0
142107.34	91391.77	366168.4	Florida	166187.9
131876.90	99814.71	362861.4	New York	156991.1
134615.46	147198.87	127716.8	California	156122.5
130298.13	145530.06	323876.7	Florida	155752.6
120542.52	148718.95	311613.3	New York	152211.8
123334.88	108679.17	304981.6	California	149760.0
101913.08	110594.11	229161.0	Florida	146122.0
100671.96	91790.61	249744.5	California	144259.4
93863.75	127320.38	249839.4	Florida	141585.5
91992.39	135495.07	252664.9	California	134307.4
119943.24	156547.42	256512.9	Florida	132602.6
114523.61	122616.84	261776.2	New York	129917.0

In [8]:

```
str(txt)
```

```
'data.frame':   1000 obs. of  5 variables:
 $ R.D.Spend      : num  165349 162598 153442 144372 142107 ...
 $ Administration : num  136898 151378 101146 118672 91392 ...
 $ Marketing.Spend: num  471784 443899 407935 383200 366168 ...
 $ State          : Factor w/ 3 levels "California","Florida",..: 3 1 2 3 2 3 1 2 3 1 ...
 $ Profit         : num  192262 191792 191050 182902 166188 ...
```

In [9]:

```
attach(txt)
```

In [15]:

```
cor(Profit,Administration)

```

```
0.741560268160455
```

In [16]:
```
cor(Profit,R.D.Spend)
```

```
0.945245288893763
```

In [17]:
```
cor(Profit,Marketing.Spend)
```

```
0.917270176692212
```

In [19]:
```
model= lm(Profit~Administration + R.D.Spend +Marketing.Spend, data= txt)
model
```

```
Call:
lm(formula = Profit ~ Administration + R.D.Spend + Marketing.Spend,
    data = txt)

Coefficients:
    (Intercept)   Administration        R.D.Spend  Marketing.Spend
     -7.016e+04        1.027e+00        5.539e-01        8.057e-02
```

In [20]:
```
summary(model)
```

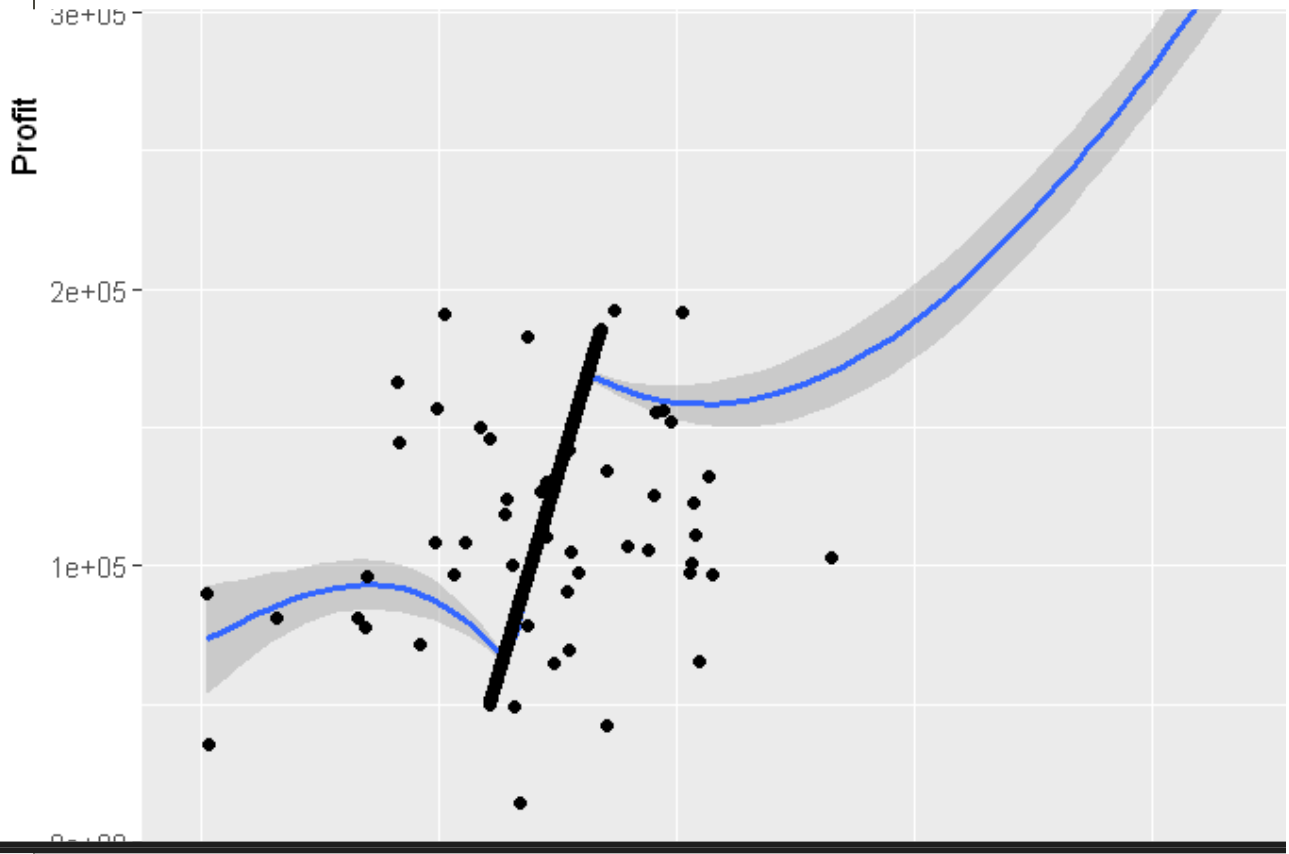
```
Call:
lm(formula = Profit ~ Administration + R.D.Spend + Marketing.Spend,
    data = txt)

Residuals:
   Min     1Q Median     3Q    Max
-60178   -605   -294    -28 161897

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.016e+04  3.967e+03  -17.69  < 2e-16 ***
Administration   1.027e+00  3.105e-02   33.07  < 2e-16 ***
R.D.Spend        5.539e-01  3.477e-02   15.93  < 2e-16 ***
Marketing.Spend  8.057e-02  1.682e-02    4.79 1.92e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9618 on 996 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9497
F-statistic:  6290 on 3 and 996 DF,  p-value: < 2.2e-16
```

In [22]:
```
1 library(ggplot2)
```

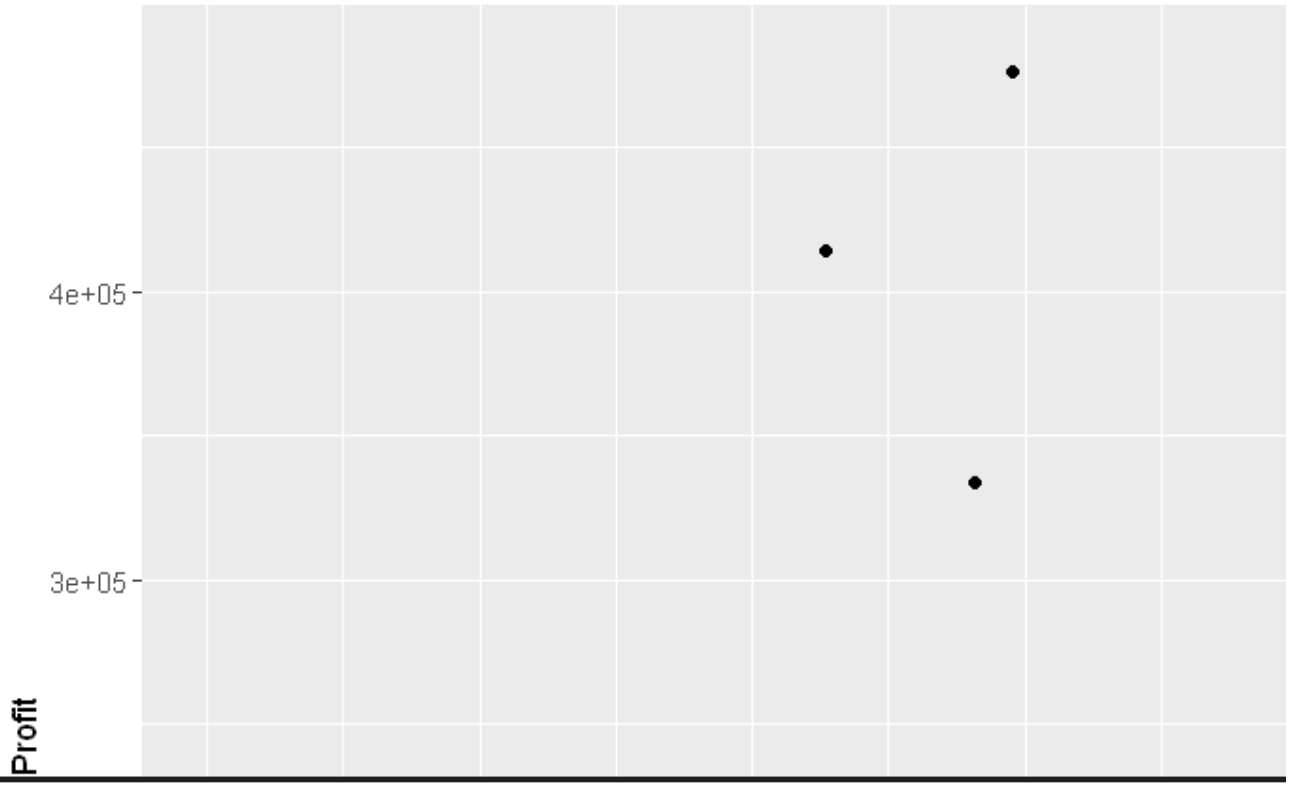
```
Warning message:
"package 'ggplot2' was built under R version 3.6.3"
```

In [23]:
```
1 ggplot(txt ,aes(x = R.D.Spend ,y = Profit))+stat_smooth()+ geom_point()
```

In [24]:
```
1 ggplot(txt ,aes(x = Administration,y = Profit))+stat_smooth()+ geom_point()
```



In [25]:
```
1 ggplot(txt ,aes(x = Marketing.Spend,y = Profit))+stat_smooth()+ geom_point()
```



In [26]:
```
library(caTools)
```

```
Warning message:
"package 'caTools' was built under R version 3.6.3"
```

In [27]:
```
library(ggplot2)
```

In [28]:
```
any(is.na(txt))
```

```
FALSE
```

In [29]:
```
library(dplyr)
```

```
Warning message:
"package 'dplyr' was built under R version 3.6.3"
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

In [30]:
```
library(ggthemes)
```

```
Warning message:
"package 'ggthemes' was built under R version 3.6.3"
```

In [31]:
```
num_cols <- sapply(txt,is.numeric)
num_cols
```

```
R.D.Spend       TRUE
Administration  TRUE
Marketing.Spend TRUE
State          FALSE
Profit          TRUE
```

In [34]:
```
cor_data <- cor(txt[,num_cols])
cor_data
```

	R.D.Spend	Administration	Marketing.Spend	Profit
R.D.Spend	1.0000000	0.5824338	0.9784066	0.9452453
Administration	0.5824338	1.0000000	0.5204649	0.7415603
Marketing.Spend	0.9784066	0.5204649	1.0000000	0.9172702
Profit	0.9452453	0.7415603	0.9172702	1.0000000

In [35]:
```
library(corrplot)
```

```
Warning message:
"package 'corrplot' was built under R version 3.6.3"corrplot 0.84 loaded
```

In [36]:
```
corrplot(cor_data,method = 'color')
```

nd

In [38]:
```
sample <- sample.split(txt$Administration, SplitRatio = 0.7)
```

In [39]:
```
train <- subset(txt,sample= TRUE)
test <- subset(txt, sample = FALSE)
```

In [41]:

```
model1 <- lm(Profit~.,train)
summary(model1)
```

```
Call:
lm(formula = Profit ~ ., data = train)

Residuals:
   Min     1Q Median     3Q    Max
-59776   -624   -302     -4 161846

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.005e+04  4.001e+03 -17.509  < 2e-16 ***
R.D.Spend        5.532e-01  3.481e-02  15.892  < 2e-16 ***
Administration   1.026e+00  3.108e-02  33.014  < 2e-16 ***
Marketing.Spend  8.109e-02  1.684e-02   4.814 1.71e-06 ***
StateFlorida    -4.464e+02  7.475e+02  -0.597    0.551
StateNew York    9.773e+01  7.395e+02   0.132    0.895
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9625 on 994 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9496
F-statistic:  3769 on 5 and 994 DF,  p-value: < 2.2e-16
```

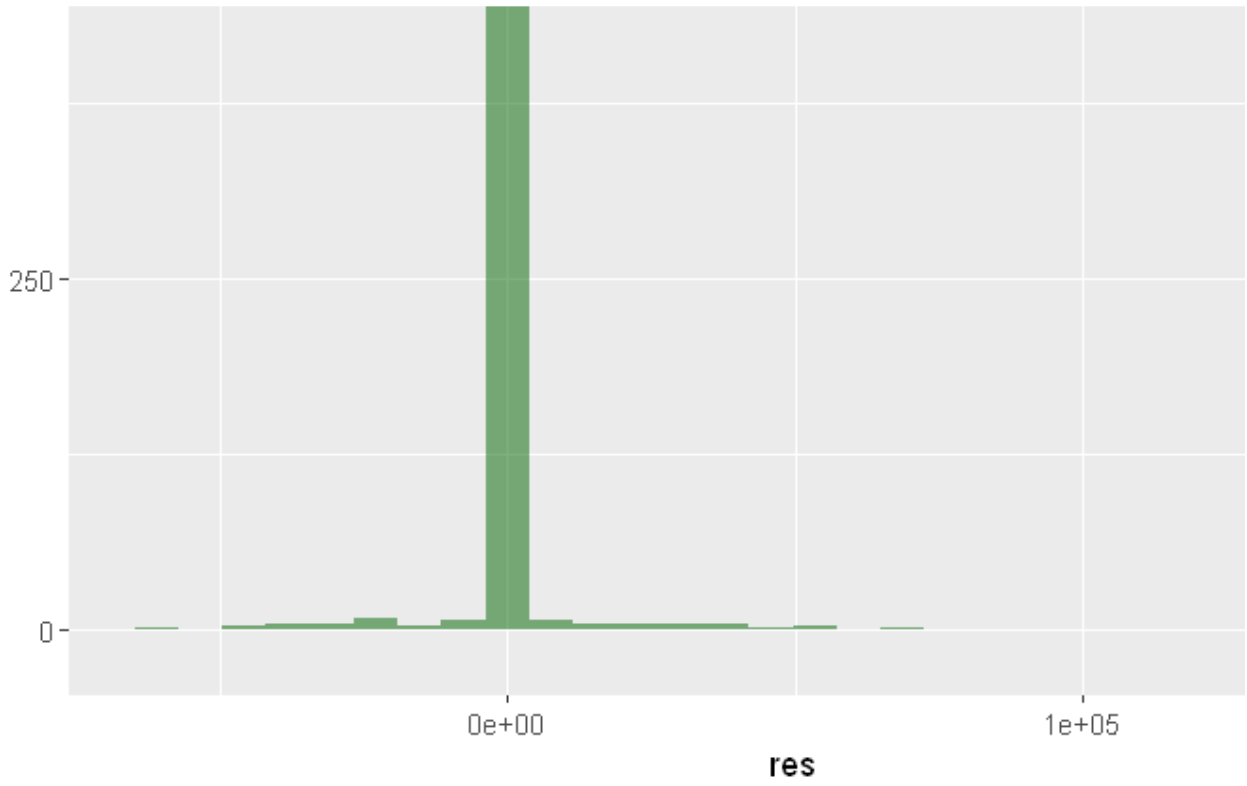
In [43]:

```
res <- residuals(model1)
res <- as.data.frame(res)
res
```

res
-7993.517
-19442.271
39793.796
20137.235
34597.273
22139.476
-9707.105
-21435.773
-22402.571
15327.181
28167.785
44173.390
9241.819
-6065.919
-44702.182
-10539.646

In [44]:

```
1 ggplot(res,aes(res))+geom_histogram(fill= 'dark green',alpha = 0.5)
2
```



In [45]:

```
1 Profit.prediction <- predict(model,test)
```

In [46]:

```
result <- cbind(Profit.prediction, test$Profit)
colnames(result)<- c('pred','real')
result <- as.data.frame(result)
result
```

pred	real
199990.74	192261.8
211085.31	191792.1
151545.80	191050.4
162522.36	182902.0
131888.76	166187.9
134602.63	156991.1
165820.22	156122.5
177520.08	155752.6
174401.90	152211.8
134308.12	149760.0
118298.95	146122.0
99965.55	144259.4
132678.01	141585.5
140261.48	134307.4
177667.48	132602.6
140255.11	129917.0

In [50]:

```
to_zero <- function(x) {if (x<0){return(0)}else{return(x)}}
```

In [48]:
```
1 result
```

pred	real
199990.74	192261.8
211085.31	191792.1
151545.80	191050.4
162522.36	182902.0
131888.76	166187.9
134602.63	156991.1
165820.22	156122.5
177520.08	155752.6
174401.90	152211.8
134308.12	149760.0
118298.95	146122.0
99965.55	144259.4
132678.01	141585.5
140261.48	134307.4
177667.48	132602.6
140255.11	129917.0

In [51]:
```
1 result$pred <- sapply(result$pred,to_zero)
2 result
```

pred	real
199990.74	192261.8
211085.31	191792.1
151545.80	191050.4
162522.36	182902.0
131888.76	166187.9
134602.63	156991.1
165820.22	156122.5
177520.08	155752.6
174401.90	152211.8
134308.12	149760.0
118298.95	146122.0
99965.55	144259.4
132678.01	141585.5
140261.48	134307.4
177667.48	132602.6
140255.11	129917.0

In [53]:
```
mse <- mean((result$real - result$pred)^2)
print(mse)
print(sqrt(mse))
```

```
[1] 90656212
[1] 9521.356
```

In [55]:
```
sse <- sum((result$pred - result$real)^2)
sse
```

```
90656212424.4765
```

In [56]:
```
sst <- sum((result$pred - mean(result$pred))^2)
sst
```

```
1741183391154.01
```

In [57]:
```
R2 <- 1-(sse/sst) # accurancy is here
R2
```

```
0.947934138997046
```