HO CHI MINH UNIVERSITY OF TECHNOLOGY
FACULTY OF APPLIED SCIENCE

PROBABILITY AND STATISTICS

Project: Analysis of factors affecting the profits of start-ups
Instructor: Dr. Nguyễn Tiến Dũng
Class: DTQ1

Team members:

Ordinal	Name	ID	Major
1	Trần Đặng Gia Huy	2153391	Chemical Engineering
2	Phạm Quang Trường	2153077	Mechatronics Engineering
3	Nguyễn Phan Trí Đức	2152528	Computer Science
4	Nguyễn Đặng Hoài Nam	2152181	Computer Science

Ho Chi Minh City , September 2022

WORKLOAD

	ID	Name	Major	Workload
1	2153391	Trần Đặng Gia Huy	Chemical Engineering	Choose the topic, write R code file, find dataset, do team planning, summarize and edit report, write latex
2	2153077	Phạm Quang Trường	Mechatronics Engineering	Write R code file, prepare theoretical materials, write content, edit, write latex
3	2152528	Nguyễn Phan Trí Đức	Computer Science	Write R code file, prepare theoretical materials, correct errors, write latex
4	2152181	Nguyễn Đặng Hoài Nam	Computer Science	Write R code file, write content and edit report , correct errors, write latex

Monitor's name: Trần Đặng Gia Huy. Email:huy.tranchemiengi@hcmut.edu.vn

Monitor's sign

Instructor's sign

Trần Đặng Gia Huy

Mục lục

I INTRODUCTION

1 Topic introduction and requirements

Knowing the kind of approach, strategy, mindset and direction to take to improve performance is always a daunting task, both for large organizations and small businesses alike. Getting to the decision to where to put money and what kind of data and analytics are important are essential to how successful an organization can be going forward. It's important to focus on things that will help maximize growth and profitability. Part of the process of getting to know these things is taking a look at profitability analysis.

So what is profitability analysis in the first place?

Marketing 91 defines profitability analysis as an “analysis of cost and revenue of the firm which determines whether or not the firm is profiting". Plus, profitability analysis can anticipate sales and profit potential specific to aspects of the market such as customer age groups, geographic regions, or product types.

After having finished studying PROBABILITY AND STATISTICS in semester 213 and been assigned to a project, we discussed and chose profitability analysis as a topic to do research on since we think that engineers must also be knowledgeable on the business side of their work and they may need to do some start-up research in the future. Afterwards, we collected dataset from every website that provides materials for our topic. Finally, a dataset from kaggle.com was chosen at the main material. Here is the link of the source.

The requirements for the topic is that we have to be well-learned, have a good knowledge of statistical data analysis, the ability to use statistical methods and so on.

2 Statistical methods

In this project, we are going to use Analysis of variance (ANOVA) to compare the profits between states in investment and Linear regression to assess the factors that affect the profits of the start up.

2.1 Analysis of variance (ANOVA)

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means. ANOVA was developed by the statistician Ronald Fisher. ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means. In other words, the ANOVA is used to test the difference between two or more means.

2.2 Linear regression model

A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables. Continuous predictor variables are also called covariates, and categorical predictor variables are also called factors.

II DESCRIPTION OF DATA AND RESULTS

CONTENT: ANALYSIS OF FACTORS AFFECTING THE PROFIT OF START UPS

Key variables in the dataset:

- RD spend: Expenses related to research and development
- Administration: Operating costs
- Marketing Spent: Advertising costs
- State
- Profit

Request:

- Compare Profits earned by startups in different regions
- Analysis of factors affecting the profits earned by start-ups

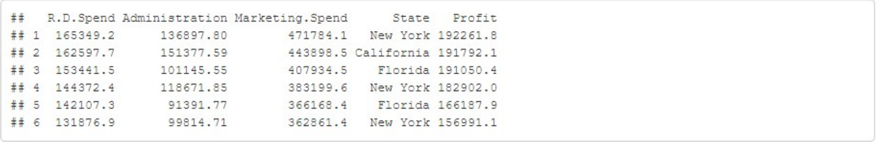
IMPLEMENT:

1 Read the data:

Input :

```
startups <- read.csv("C:/startups.csv", sep=";")
head(startups)
```

Output :



```
##   R.D.Spend Administration Marketing.Spend      State   Profit
## 1  165349.2      136897.80        471784.1   New York 192261.8
## 2  162597.7      151377.59        443898.5 California 191792.1
## 3  153441.5      101145.55        407934.5    Florida 191050.4
## 4  144372.4      118671.85        383199.6   New York 182902.0
## 5  142107.3       91391.77        366168.4    Florida 166187.9
## 6  131876.9       99814.71        362861.4   New York 156991.1
```

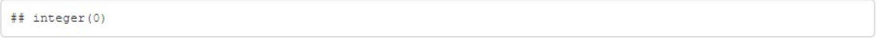
2 Data Cleaning:

Check for missing data in the file:

Input :

```
apply(is.na(startups), 2, which)
```

Output :



```
## integer(0)
```

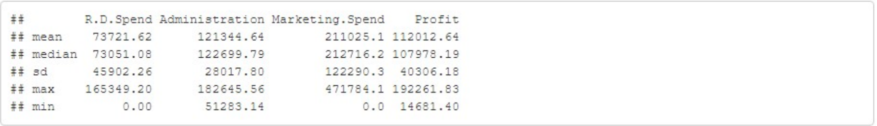
Comment: The file has no missing data.

3 Data clarification: Calculate descriptive statistics for variables "R.D.Spend","Administration","Marketing.Spend","Profit":

Input :

```
mean = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,mean)
median = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,median)
sd = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,sd)
max = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,max)
min = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2, min)
t(data.frame(mean,median,sd,max,min))
```

Output :



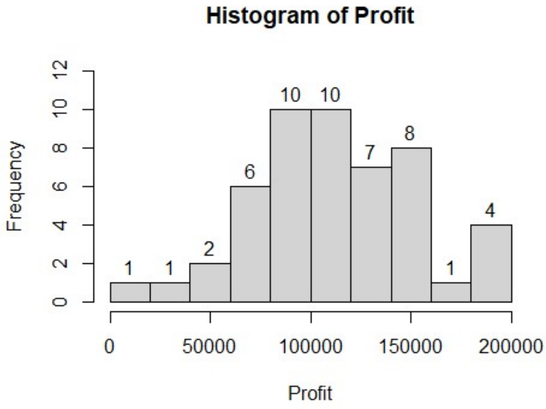
```
##        R.D.Spend Administration Marketing.Spend    Profit
## mean    73721.62      121344.64        211025.1 112012.64
## median  73051.08      122699.79        212716.2 107978.19
## sd      45902.26       28017.80        122290.3  40306.18
## max    165349.20      182645.56        471784.1 192261.83
## min         0.00       51283.14             0.0  14681.40
```

4 Plot a graph of the frequency distribution for the variable Profit:

Input :

```
hist(startups$Profit,ylim = c(0,12), xlab ="Profit", main = "Histogram of Profit", labels = T)
```

Output :

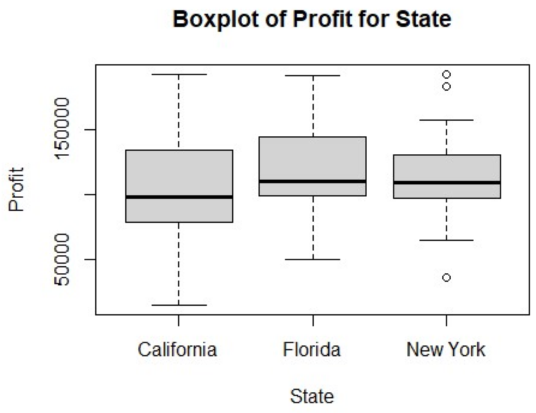


5 Plot a boxplot showing the distribution of Profit over State:

Input :

```
boxplot(Profit~State,data=startups, main = "Boxplot of Profit for State")
```

Output :



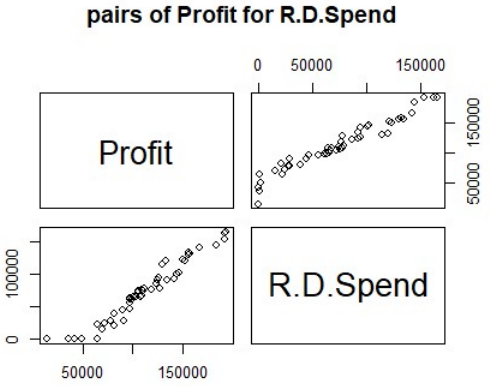
Comment: Based on the graph, we don't expect much difference in the profitability of startups in different regions.

6 Draw a scatter plot showing the distribution of Profit according to R.D.Spend, Administration, Marketing.Spend:

Input :

```
pairs(Profit ~ R.D.Spend,main = "pairs of Profit for R.D.Spend",data=startups)
```

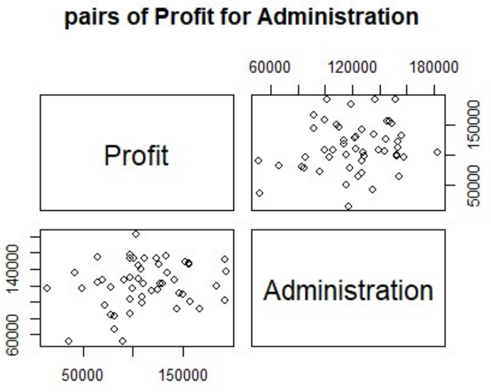
Output :



Input :

```
pairs(Profit ~ Administration,main = "pairs of Profit for Administration",data=startups)
```

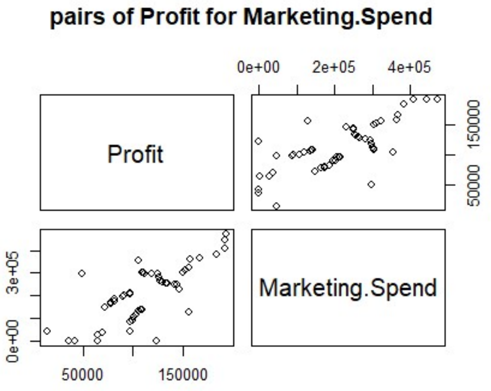
Output :



Input :

```
pairs(Profit~ Marketing.Spend,main = "pairs of Profit for Marketing.Spend",data=startups)
```

Output :



Comment: Based on the graphs, we see that R.D.Spend has a linear relationship with Profit specifically, an increasing function relationship), while Administration and Marketing.Spend do not have.

7 Building anova model:

Comparing the profits earned by startups in regions Hypothetical statement:

Assumption H0: Average profits of startups across regions are the same

Assumption H1: There are at least 2 regions where the average profits of startups are different.

Assumptions have to be made before ANOVA analysis:

- These populations are normally distributed: The average profits of startups in the regions follow a normal distribution.
- Homogeneity of Variances: The Profit Variances of Startups in different regions are equal Filter data corresponding to each region:

Input :

```r
New_York<-subset(startups,startups$State =="New York")
California<-subset(startups,startups$State =="California")
Florida<-subset(startups,startups$State =="Florida")
```

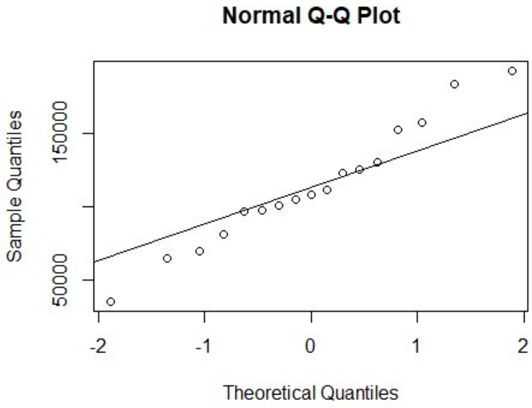
Check the normal distribution assumption:

- For the New York area:

Input :

```r
qqnorm(New_York$Profit)
qqline(New_York$Profit)
```

Output :

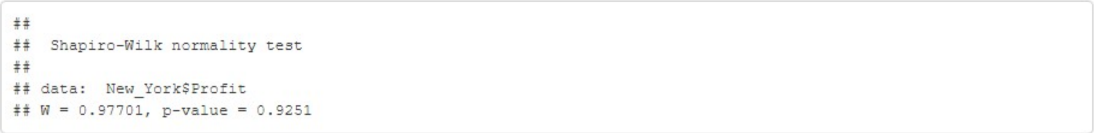


Comment: The QQ-plot shows that the observed values are mostly on the expected line of the normal distribution, so the Profit variable in the New York area follows the normal distribution. Alternatively, we can use the shapiro.test function to check:

Input :

```r
shapiro.test(New_York$Profit))
```

Output :



```
##
##  Shapiro-Wilk normality test
##
## data:  New_York$Profit
## W = 0.97701, p-value = 0.9251
```

Comment:

Assumption H0: profits in the New York area are normally distributed.

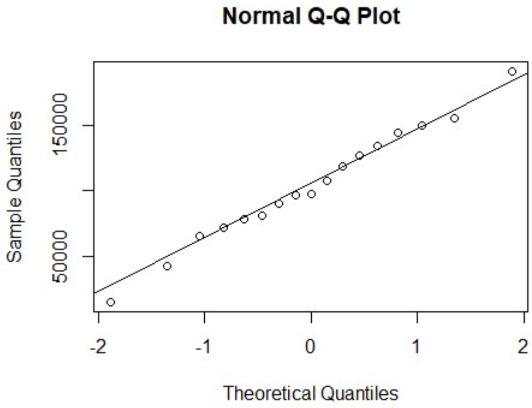
Assumption H1: profits in the New York area are not normally distributed.

Since Pr(>F) = 0.9251 > 5% significance level, we accept the hypothesis H0. So the Profit variable in the New York area follows a normal distribution.

- For the California area: **Input :**

```r
qqnorm(California$Profit)
qqline(California$Profit)
```

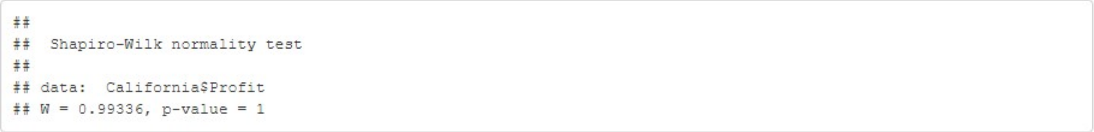
Output :



Comment: The QQ-plot shows that the observed values are mostly on the expected line of the normal distribution, so the Profit variable in the California area follows the normal distribution. Alternatively, we can use the shapiro.test function to check **Input :**

```
shapiro.test(Florida$Profit)
```

Output :



```
##
##  Shapiro-Wilk normality test
##
## data:  California$Profit
## W = 0.99336, p-value = 1
```

Comment:

Assumption H0: Profits in the Florida area are normally distributed.

Assumption H1: Profits in the Florida area are not normally distributed.

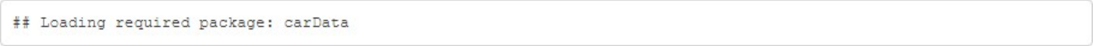
Since Pr(>F) = 0.9834 > 5% significance level, we accept hypothesis H0. So the Profit variable in the Florida area follows a normal distribution.

Using the Test for Homogeneity of Variance:

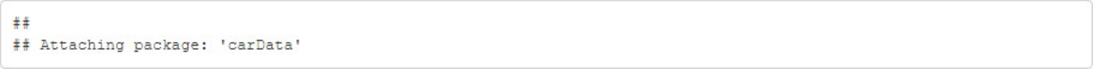
Input :

```
library(car)
```

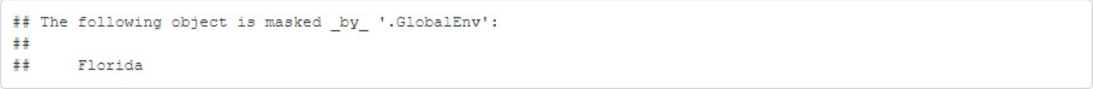
Output :



```
## Loading required package: carData
```



```
##
## Attaching package: 'carData'
```

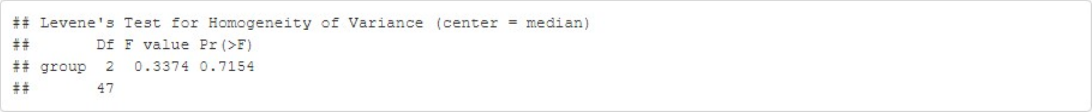


```
## The following object is masked _by_ '.GlobalEnv':
##
##     Florida
```

Input :

```
leveneTest(Profit~as.factor(State),data=startups)
```

Output :



```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.3374 0.7154
##       47
```

Comment:

Assumption H0: Returns' variance in different areas are the same.

Assumption H1: There exist 2 areas that have different returns' variance.

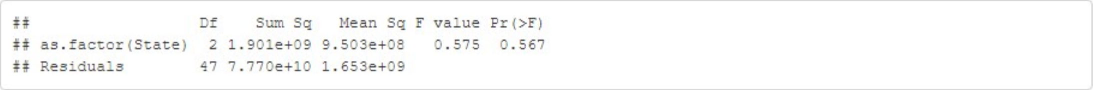
Since Pr(>F) = 0.7154 > 5% significance level, we accept hypothesis H0. Hence the returns' variance in different areas are the same.

8 Using ANOVA (Analysis of variance) for single factor:

Input :

```
anova_model<-aov(Profit~as.factor(State),data=startups)
summary(anova_model)
```

Output :



```
##                  Df    Sum Sq   Mean Sq F value Pr(>F)
## as.factor(State)  2 1.901e+09 9.503e+08   0.575  0.567
## Residuals        47 7.770e+10 1.653e+09
```

Comment: Since Pr(>F) = 0.567 > 5% significance level, we accept hypothesis H0. Hence the average return of the startups in different areas are the same.

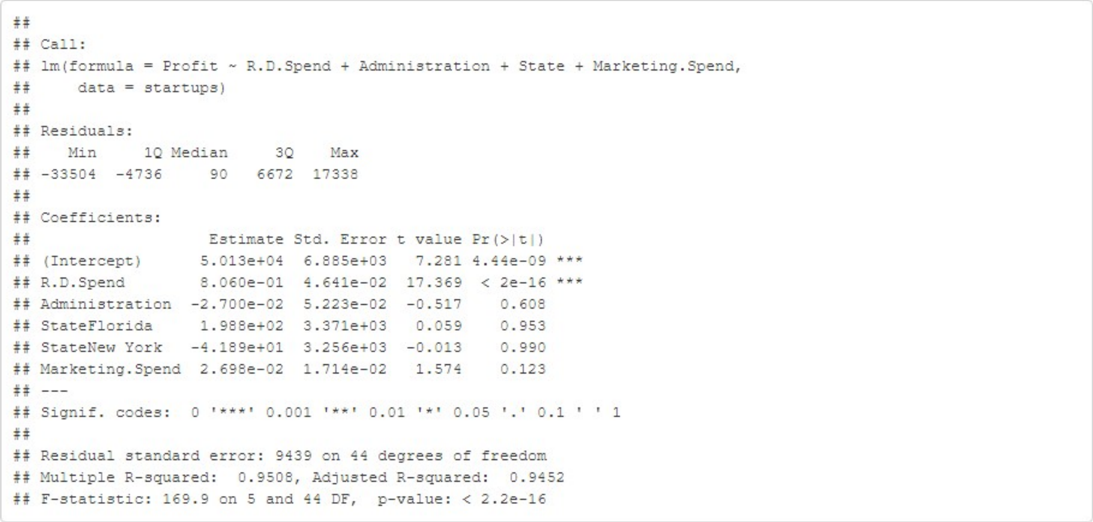
9 Building a regression model:

Analyze the factors that affect the profit of the startups. In the model: Dependent variable: Profit Independent variable: All the remainings.

Input :

```
M1 = lm(Profit ~ R.D.Spend + Administration + State + Marketing.Spend, data = startups)
summary(M1)
```

Output :



```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + Administration + State + Marketing.Spend,
##     data = startups)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33504  -4736     90   6672  17338
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.013e+04  6.885e+03   7.281 4.44e-09 ***
## R.D.Spend        8.060e-01  4.641e-02  17.369  < 2e-16 ***
## Administration  -2.700e-02  5.223e-02  -0.517    0.608
## StateFlorida     1.988e+02  3.371e+03   0.059    0.953
## StateNew York   -4.189e+01  3.256e+03  -0.013    0.990
## Marketing.Spend  2.698e-02  1.714e-02   1.574    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9439 on 44 degrees of freedom
## Multiple R-squared:  0.9508, Adjusted R-squared:  0.9452
## F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

Comment: According to the result of the linear regression model above, suppose:

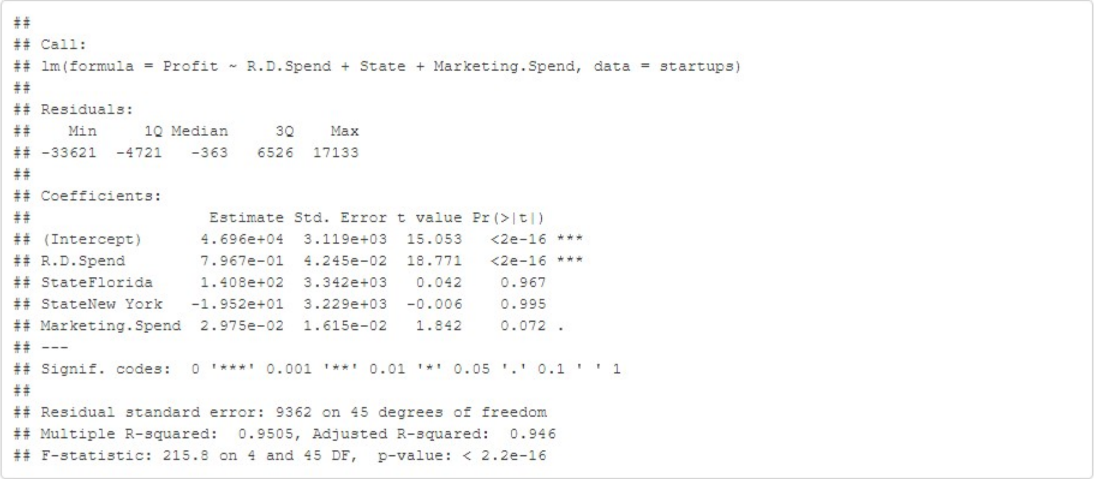
Assumption H0: The coefficients on variables don't have statistical significance.

Assumption H1: The coefficients on variables have statistical significance.

Since Pr of Administration, StateFlorida, StateNew York, Marketing.Spend are both > 5% significance level, we accept hypothesis H0. So the coefficients on those variables don't have statistical significance. We will exclude those variables from the model. Since Pr of the remaining variables are < 5% significance level, we can reject hypothesis H0. So the coefficients on those variables have statistical significance. We don't need to exclude those variables from the model. Building the second model with the exclusion of variable Administration from the first model. **Input :**

```
M2 = lm(Profit ~ R.D.Spend + State + Marketing.Spend, data = startups)
summary(M2)
```

Output :

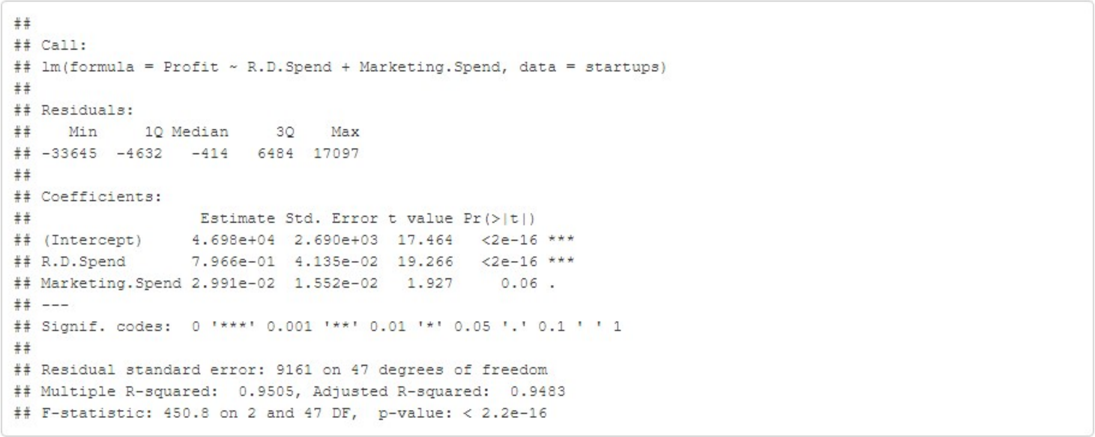


```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + State + Marketing.Spend, data = startups)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33621  -4721   -363   6526  17133
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.696e+04  3.119e+03  15.053   <2e-16 ***
## R.D.Spend        7.967e-01  4.245e-02  18.771   <2e-16 ***
## StateFlorida     1.408e+02  3.342e+03   0.042    0.967
## StateNew York   -1.952e+01  3.229e+03  -0.006    0.995
## Marketing.Spend  2.975e-02  1.615e-02   1.842    0.072 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9362 on 45 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.946
## F-statistic: 215.8 on 4 and 45 DF,  p-value: < 2.2e-16
```

Building model 3 remove variable State from model 2. **Input :**

```
M3 = lm(Profit ~ R.D.Spend + Marketing.Spend, data = startups)
summary(M3)
```

Output :



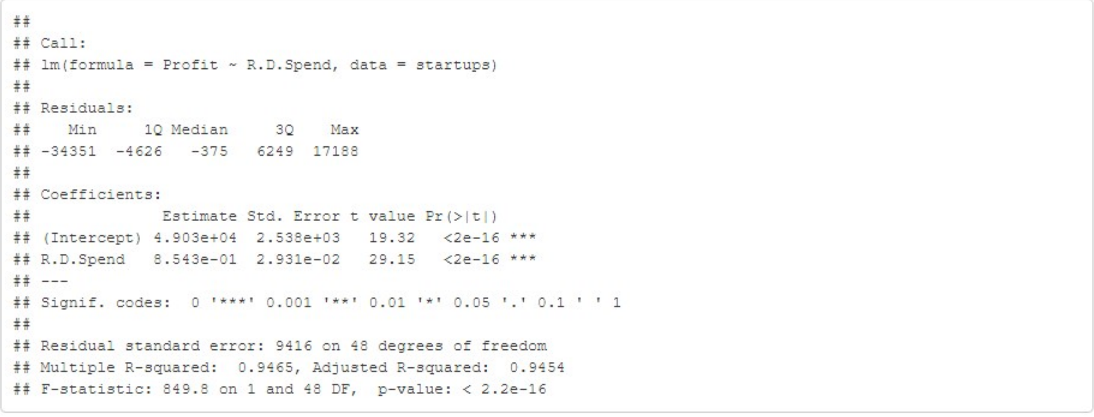
```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = startups)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33645  -4632   -414   6484  17097
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
## R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
## Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9161 on 47 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9483
## F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Building model 4 remove variable Marketing.Spend from model 3. **Input :**

```
M4 = lm(Profit ~ R.D.Spend, data = startups)
```

```
summary(M4)
```

Output :



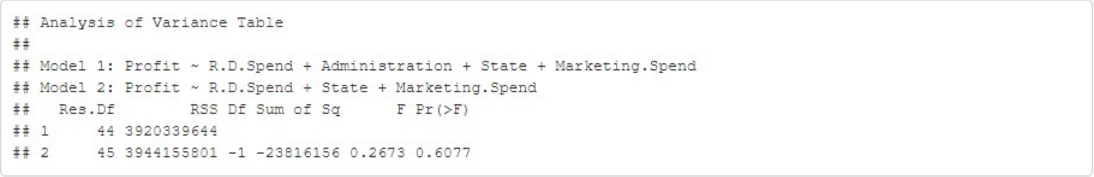
```
##
## Call:
## lm(formula = Profit ~ R.D.Spend, data = startups)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34351  -4626   -375   6249  17188
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
## R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9416 on 48 degrees of freedom
## Multiple R-squared:  0.9465, Adjusted R-squared:  0.9454
## F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

10 Comparing models:

Input :

```
anova(M1,M2)
```

Output :



```
## Analysis of Variance Table
##
## Model 1: Profit ~ R.D.Spend + Administration + State + Marketing.Spend
## Model 2: Profit ~ R.D.Spend + State + Marketing.Spend
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1     44 3920339644
## 2     45 3944155801 -1 -23816156 0.2673 0.6077
```

Comment:

Assumption H0: Model 2 is more effective than model 1

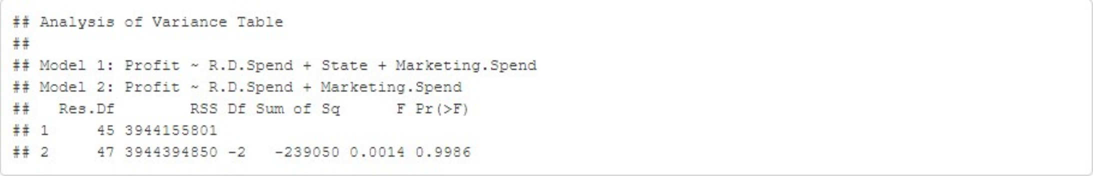
Assumption H1: Model 1 is more effective than model 2

Since the observation probability Pr $>$ 5% significance level, the hypothesis H0 can not be rejected. So model 2 is more effective than model 1.

Input :

```
anova(M2,M3)
```

Output :



```
## Analysis of Variance Table
##
## Model 1: Profit ~ R.D.Spend + State + Marketing.Spend
## Model 2: Profit ~ R.D.Spend + Marketing.Spend
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1     45 3944155801
## 2     47 3944394850 -2   -239050 0.0014 0.9986
```

Comment: Assumption H0: Model 3 is more effective than model 2

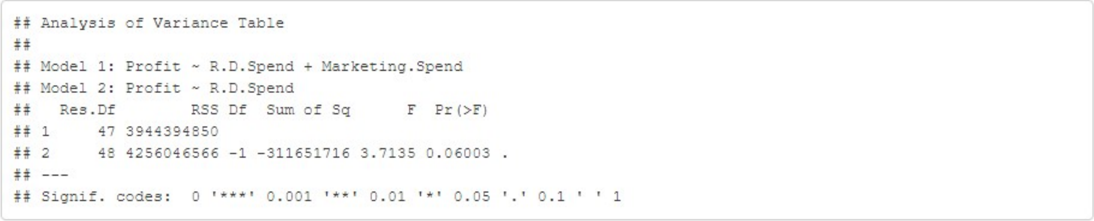
Assumption H1: Model 2 is more effective than model 3

Since the observation probability Pr $>$ 5% significance level, the hypothesis H0 can not be rejected. So model 3 is more effective than model 2.

Input :

```
anova(M3,M4)
```

Output :



```
## Analysis of Variance Table
##
## Model 1: Profit ~ R.D.Spend + Marketing.Spend
## Model 2: Profit ~ R.D.Spend
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1     47 3944394850
## 2     48 4256046566 -1 -311651716 3.7135 0.06003 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment:

Assumption H0: Model 4 is more effective than model 3

Hypothesis H1: Model 3 is more effective than model 4

Since the observation probability Pr $>$ 5% significance level, the hypothesis H0 can not be rejected. So models 4 is more effective than model 3

Conclusion: From the comparison of models, we find that model 4 is the most effective model.

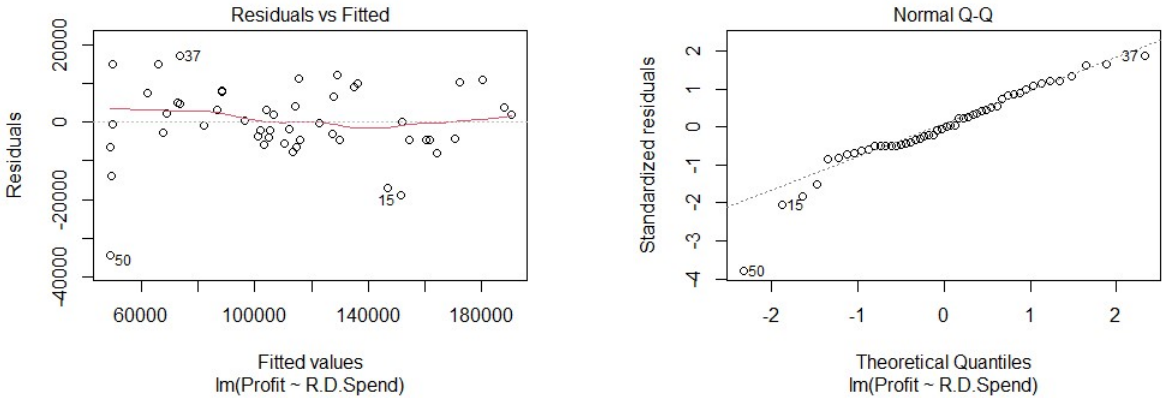
Check the assumptions in the regression model: Linearity of data: the relationship between predictor X and dependent variable Y is assumed to be linear. Errors have a mean of 0. The variance of the errors is constant. The error is normally distributed. The errors are independent of each other.

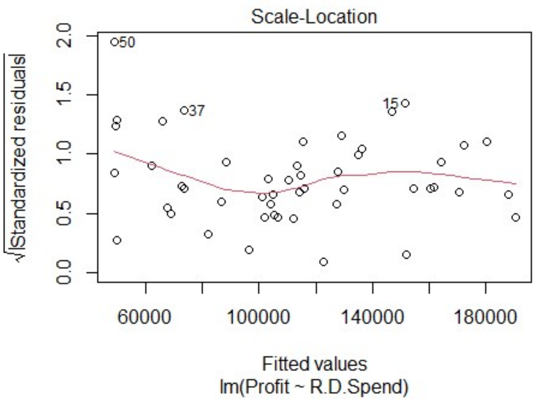
We draw graphs to test the model's assumptions:

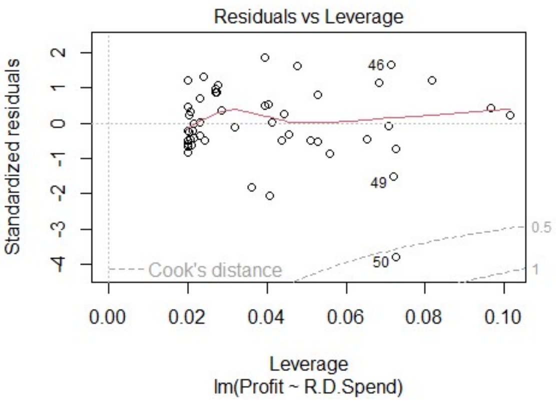
Input :

```
plot(M4)
```

Output :







Comment:

The first graph (Residuals vs Fitted) plots the predicted values with corresponding residual values (errors), used to test the assumption that the errors have zero expectation and the homogeneity of the variances. error. Based on the graph, we see that the red line is close to the y = 0 line, so it is assumed that the errors have equal expectations. The errors are not distributed randomly along the red line, so the assumption that the variance of errors is constant is not satisfied.

The second plot (Normal Q-Q) plots the normalized error values, allowing the assumption of the normal distribution of errors to be tested. Based on the graph, we see that the errors are mostly concentrated on the normal distribution expected line, so the normal distribution of the errors is assumed to be satisfied.

The third plot (Scale - Location) plots the square root of the residuals normalized with the predicted values, which is used to test the assumption that the variance of the errors is constant. Based on the graph, we see that the red line is not a horizontal line and the errors are not randomly scattered along the red line, so assuming the variance of errors is a constant, it is not satisfied.

The fourth graph (Residuals vs Leverage) allows to identify points with high influence (influential observations), if they are present in the data set. These high-impact points can be outliers, which are the ones that can have the most impact when analyzing data.

Based on the graph, we can see that the 46th, 49th, and 50th observations can be highly influential points in the data set. However, we also observe that only point 50 crosses the Cook's distance line (red dashed line Cook's distance). Therefore, a score of 50 is a highly influential point in the data set. Therefore we need to remove them when analyzing.

III FULL R CODE

```r
startups <- read.csv("C:/startups.csv", sep=";")
head(startups)
apply(is.na(startups), 2, which)
mean = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,
    mean)
median = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,
    median)
sd = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,sd)
max = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2,max)
min = apply(startups[,c("R.D.Spend","Administration","Marketing.Spend","Profit")],2, min)
t(data.frame(mean,median,sd,max,min))
hist(startups$Profit,ylim = c(0,12), xlab ="Profit", main = "Histogram of Profit", labels
     = T)
boxplot(Profit~State,data=startups, main = "Boxplot of Profit for State")
pairs(Profit ~ R.D.Spend,main = "pairs of Profit for R.D.Spend",data=startups)
pairs(Profit ~ Administration,main = "pairs of Profit for Administration",data=startups)
pairs(Profit~ Marketing.Spend,main = "pairs of Profit for Marketing.Spend",data=startups)
New_York<-subset(startups,startups$State =="New York")
California<-subset(startups,startups$State =="California")
Florida<-subset(startups,startups$State =="Florida")
qqnorm(New_York$Profit)
qqline(New_York$Profit)
shapiro.test(New_York$Profit)
qqnorm(California$Profit)
qqline(California$Profit)
shapiro.test(California$Profit)
qqnorm(Florida$Profit)
qqline(Florida$Profit)
shapiro.test(Florida$Profit)
library(car)
leveneTest(Profit~as.factor(State),data=startups)
anova_model<-aov(Profit~as.factor(State),data=startups)
summary(anova_model)
M1 = lm(Profit ~ R.D.Spend + Administration + State + Marketing.Spend, data = startups)
summary(M1)
M2 = lm(Profit ~ R.D.Spend + State + Marketing.Spend, data = startups)
summary(M2)
M3 = lm(Profit ~ R.D.Spend + Marketing.Spend, data = startups)
summary(M3)
M4 = lm(Profit ~ R.D.Spend, data = startups)
summary(M4)
anova(M1,M2)
anova(M2,M3)
anova(M3,M4)
plot(M4)
```

IV CONCLUSION

After performing the ANOVA analysis, it can be seen that the average profits of start ups from different states are equal. We can, therefore, anticipate that states where start up operate do not affect the profits. After building a regression model and analyzing factors affecting the profitability of start-ups, it can be concluded that only R.D.Spend has statistical significance,while all remaining variables don't. After excluding Administration, State and Marketing.spend, the model with only R.D.Spend show the most effective one. However, in reality, Administration, State and Marketing.spend may affect the profit because the statistical data may not be entirely random and relatively small therefore, we can yet evaluate correctly, if we can collect a bigger dataset, we can assess factors that affect the profits better.

V REFERENCE

1. Dataset from Kaggle, website
   https://www.kaggle.com/datasets/karthickveerakumar/startup-logistic-regression/code?fbclid=IwAR3t4SaYOzvyPpmh4gTYr-29BTrE28VSOkFWuGi02gkEKD2pUf4nfeZg3Co

2. Example of statistical data analysis using the R, website:
   http://www.css.cornell.edu/faculty/dgr2/_static/files/R_PDF/corregr.pdf

3. Using R for Introductory Statistics, website: http://e-learning-old.hcmut.edu.vn/pluginfile.php/2036519/mod_resource/content/0/Verzani-SimpleR.pdf

4. R Tutorial - Learn R Programming
   https://www.datamentor.io/r-programming/