MONASH University
Information Technology

ASSESSMENT COVER SHEET

Student ID number	26389126
Given Name	MOON BYEONG
Family name	KIM

Unit Name and Code:	**FIT5145**
Campus:	**Caufield**
Assignment Title:	**Assignment3- Coding**
Name of Lecturer:	**Prof Wray Buntine**
Name of Tutor:	**Vijayalaxmi Beeravalli**
Tutorial Day and Time:	**Tuesday, 2:00 pm**
Phone Number:	**0416148198**
Email Address:	**Mbkim1@student.monash.edu**

Has any part of this assignment been previously submitted as part of another unit/course? ☐ Yes ☒ No

Due Date:	**08/05/2018**	Date Submitted:	08/05/2018

All work must be submitted by the due date. If an extension of work is granted this must be specified with the signature of the lecturer/tutor.

Extension granted until (date)____________________ Signature of lecturer/tutor ___________________________

Please note that it is your responsibility to retain copies of your assessments.

Intentional plagiarism or collusion amounts to cheating under Part 7 of the Monash University (Council) Regulations

Plagiarism: Plagiarism means taking and using another person's ideas or manner of expressing them and passing them off as one's own. For example, by failing to give appropriate acknowledgement. The material used can be from any source (staff, students or the internet, published and unpublished works).

Collusion: Collusion means unauthorised collaboration with another person on assessable written, oral or practical work and includes paying another person to complete all or part of the work.

Where there are reasonable grounds for believing that intentional plagiarism or collusion has occurred, this will be reported to the Associate Dean (Education) or delegate, who may disallow the work concerned by prohibiting assessment or refer the matter to the Faculty Discipline Panel for a hearing.

Student Statement:

- I have read the university's Student Academic Integrity Policy and Procedures.
- I understand the consequences of engaging in plagiarism and collusion as described in Part 7 of the Monash University (Council) Regulations http://adm.monash.edu/legal/legislation/statutes
- have taken proper care to safeguard this work and made all reasonable efforts to ensure it could not be copied.
- No part of this assignment has been previously submitted as part of another unit/course.
- I acknowledge and agree that the assessor of this assignment may for the purposes of assessment, reproduce the assignment and:
  - i. provide to another member of faculty and any external marker; and/or
  - ii. submit it to a text matching software; and/or
  - iii. submit it to a text matching software which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking.
- I certify that I have not plagiarised the work of others or participated in unauthorised collaboration when preparing this assignment.

Signature ....MBK............................. Date………08/05/2018……………………………

* delete (iii) if not applicable

The information on this form is collected for the primary purpose of assessing your assignment and ensuring the academic integrity requirements of the University are met. *Other purposes of collection include recording your plagiarism and collusion declaration, attending to course and administrative matters and statistical analyses. If you choose not to complete all the questions on this form it may not be possible for Monash University to assess your assignment. You have a right to access personal information that Monash University holds about you, subject to any exceptions in relevant legislation. If you wish to seek access to your personal information or inquire about the handling of your personal information, please contact the University Privacy Officer: privacyofficer@adm.monash.edu.au*

Table of Contents

Task A

A.1. The top 20 emoticons

This task is to extract the top 20 emoticons and their counts from the tweets in the msgraw_sample.txt.

The table below shows the top 20 emoticons and their counts extracted from the tweets. The most often emoticons people used in the msgraaw_sample.txt is ":-*" emoticon, 1431349 times used, among 80 potential emoticons created.

Rank	Frequency	Emoticons
1	1431349	:-*
2	91723	:3
3	82960	\o/
4	16278	:)
5	6500	:D
6	6500	:))
7	6474	D8
8	3696	^^
9	1977	T.T
10	1878	\^o^
11	1793	(*^o^*)
12	1588	=)
13	1317	<3
14	1158	xD
15	1027	:P
16	846	:-)
17	735	=3
18	731	0_0
19	726	D:
20	661	DX

<Table1. The top 20 emoticons and their counts.>

Followings are the way of extracting the top 20 emoticons and frequencies from tweets.

- Read the contents and the output is redirected to "temp1" through "cat msgraw_sample.txt >temp1" before this command is run.
- Converting whitespace characters to newline characters to tokenize each line of text of temp1.
- Converting the embedded HTML escapes for '>' and '<' back to their original format.
- Finding match line containing string through "grep -e.
- Reading the entire file and counts the number of line-endings through "wc -l".
- Printing a value and write into temp2 through "echo" and ">", and append it into variable by ">>".

- Creating and copying temp2 into “potential_emoticon.csv.
- Then reads temp2 file, then sort in descending order by number, then display the first 20 lines, then sort and count the items, and then finally write this value into “emoticon.csv”.

A.2. The word co-occurrence with emoticons

This task is to compute word co-occurrence with emoticons from the tweets in the msgraw_sample.txt. Use the 20 most frequently used emoticons extracted to find the 15 words for each emoticon that occurs most often, as shown in the figures below.
There are emoticons used with many words like :) and: D, but some emoticons are not associated with any words like \ ^ o ^. There are also emoticons that do not occur with even 15 words, like 0_0.
I found some interesting things in results below. First, there are differences in emoticons that are frequently used between English-speaking countries and non-English speaking countries. Secondly, the emoticons that are mostly used come with words like “RT”. This means they are more influential. Thirdly, the emoticons symbolically represent the words used. For example, emoticons such as <3 are bright and proactive and display words that give the impression of “love”, “happy”, “will”, “day” and “you”.

	:-*		:3		\o/		:)	
Rank	**Frequency**	**Words**	**Frequency**	**Words**	**Frequency**	**Words**	**Frequency**	**Words**
1	5	RT	163	RT	10	que	1633	RT
2	4	jilat	17	ya	10	dia	539	you
3	3	you	16	aku	8	de	445	to
4	3	day	16	a	6	o	432	a
5	3	11	14	ga	6	es	369	I
6	2	will	13	you	6	e	315	the
7	2	to	13	ada	6	RT	268	for
8	2	sisig	11	to	6	111111	255	ya
9	2	pramudina	11	me	5	do	234	me
10	2	kecup	11	i	5	a	221	111111
11	2	i	11	di	5	Bom	207	and
12	2	daahills	10	de	4	um	206	my
13	2	canned	10	cantik	4	sextafeira	172	de
14	2	basah	9	yg	4	me	169	is
15	2	a	9	the	3	viernes	169	in

<Table2. The word co-occurrence with emoticons 1>

	:D		:))		D8		^^	
Rank	Frequency	Words	Frequency	Words	Frequency	Words	Frequency	Words
1	1800	RT	149	RT	3	RT	129	RT
2	193	ya	32	na	2	Joker11297	28	you
3	163	haha	22	a	1	w	24	to
4	157	I	20	you	1	sih	23	the
5	156	a	19	to	1	pawang	22	111111
6	148	you	18	di	1	nya	21	for
7	133	di	17	ya	1	kufufu	19	I
8	129	p	17	me	1	httptco09v34HhK	18	a
9	123	aku	16	ako	1	eh	15	in
10	122	the	16	I	1	apaan	13	and
11	118	to	16	D	1	apa	12	me
12	117	yg	15	yg	1	ada	11	3
13	114	aja	15	sa	1	Pawang	10	is
14	106	me	15	111111	1	Kenapa	10	Kevinwoo91
15	105	111111	14	my	1	Ini	9	di

<Table3. The word co-occurrence with emoticons 2>

	T.T		\^o^		(*^o^*)		=)	
Rank	Frequency	Words	Frequency	Words	Frequency	Words	Frequency	Words
1	40	RT	1		3	O	42	RT
2	13	I			2	RT	27	a
3	7	ini			1	yasmineco	19	you
4	5	ya			1	w	19	dia
5	5	nonton			1	twitpiccom7cziv2	18	de
6	4	you			1	stelymac	17	I
7	4	my			1	snowyukiswing	14	the
8	4	ke			1	satoova	13	que
9	4	gak			1	rain2255	13	¡
10	3	yaa			1	poeplus	13	Bom
11	3	u			1	pet	13	111111
12	3	to			1	oo	12	to
13	3	p			1	neko	11	my
14	3	not			1	mochi819	11	me
15	3	nih			1	miffychanx	10	is

<Table4. The word co-occurrence with emoticons 3>

	<3		xD		:P		:-)	
Rank	Frequency	Words	Frequency	Words	Frequency	Words	Frequency	Words
1	179	you	63	RT	266	RT	61	RT
2	147	RT	50	de	48	you	46	to
3	127	to	40	que	48	I	46	the
4	111	I	36	no	42	to	44	a
5	109	111111	30	me	42	D	39	I
6	88	the	30	a	38	a	36	you
7	87	my	25	la	37	the	29	for
8	84	me	21	y	35	haha	27	it
9	84	love	21	en	29	aku	26	in
10	81	a	20	es	26	p	24	me
11	67	and	19	el	25	¡	24	is
12	59	for	17	to	24	ga	22	and
13	57	¡	17	lo	24	and	21	de
14	51	all	16	haha	23	ya	19	on
15	50	in	15	D	23	it	16	my

<Table5| The word co-occurrence with emoticons 4>

	=3		0_0		D:		DX	
Rank	Frequency	Words	Frequency	Words	Frequency	Words	Frequency	Words
1	1	yak	1	yall	19	de	1	you
2	1	te	1	this	18	que	1	with
3	1	rome	1	like	17	I	1	want
4	1	reis	1	drivin	11	no	1	type
5	1	posso	1	bnoo14	11	a	1	too
6	1	pariis	1	avi	10	to	1	tam4man
7	1	op	1	Tri	10	the	1	su
8	1	oi	1	NiKKiBADDDD	9	me	1	pubblicato
9	1	of			9	RT	1	photo
10	1	neem			7	you	1	order
11	1	naar			7	it	1	one
12	1	moide			7	e	1	nellalbum
13	1	mee			5	y	1	httptcotzDlycSk
14	1	je			5	was	1	httptcoiTEvUUmU
15	1	ik			5	my	1	httptcoHGbDyMhR

<Table6| The word co-occurrence with emoticons 5>

To get these outputs, create "emoworlds.sh" and calls "emoworld.py". Figure1 below is a shell script that used to output of 20 emoticons the most frequent words co-occurring with it.

- Using while loop and input the files "emoticon.csv" that divides two lists, first is counts and second is emoticons.
- Use "cut -d -f2" command to divide the lists with "space" and get the second line.
- Use the "emoworld.py" to read the "emoticon.csv file and search all tweets from the "msgraw_sample.txt" file.

- Delete all punctuations and change the space to a newline, then sort the list by “a-z” and order list with the counts and display 15 values.

```
#!/bin/bash

while read -r line
do
  emo=`echo "$line" | cut -d'   ' -f2`
  result=`python ./emoword.py $emo < msgraw_sample.txt`
  echo $result | tr -d '[:punct:]' | tr '[:space:]' '\n' | tr -s '\n' | sort | uniq -c | sort -nr | head -n 15

done < emoticon.csv
```

<Figure1. Emoticon shell script>

A.3. Findings

As mentioned above, there are differences in emoticons that are frequently used between English-speaking countries and non-English speaking countries. In English-speaking countries, mainly use playful and bright emoticon such as :), <3, ^^, TT, DX, D:, =), (* ^ o ^), o_o, : P and :-). Regardless of language, emoticons tend to feel bright. Some emoticons have the word, real-twitter user name such as “NikikiBaddd” or “mochi819”. This means they are more influential. Thirdly, the emoticons symbolically represent the words used. For example, emoticons such as <3 are bright and proactive and display words that give the impression of “love”, “happy”, “will”, “day” and “you”. Moreover, this emoticon, (* ^ o ^ *), is related to pets such as dogs or cats.

Below is the output of using emodata.py. These words show about the countries such as Canada and US, city of Jakarta or time values such as “2011”, “Nov”, “11”, “3”, “Fri”.

```
11298 11
11218 2011
11206 Nov
11203 0000
11200 Fri
6289 D
4192 RT
2524 Canada
2522 Time
2505 US
2081 Pacific
1497 3
1142 Jakarta
1133 P
1047 you
```

Task B

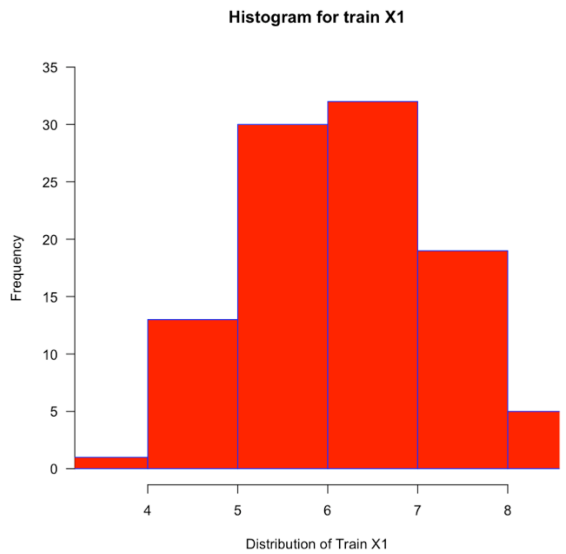
B.1. Plot histograms

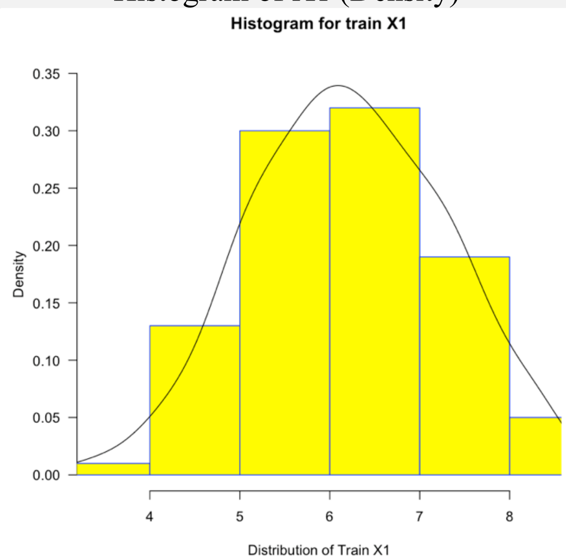
Figures below are histograms of variable X1~4 in train.csv respectively. The variable X1 and X4 are more likely samples drawn from normal distribution because they are ball-shaped curve and points are as likely to occur on one side of the average as on the other [1]. However, X1 is better normal distribution model than X4. Both have small standard deviation. X1 has similar value of median, mean and mode value, so X1 is definitely normal distribution; but, X4 has different value of mean, mode, and median. Therefore, variable X1 is most likely sample drawn from normal distribution.

1. X1

This is normal distribution.

Summary of X1	Standard deviation of X1
`Min. 1st Qu. Median Mean 3rd Qu. Max.` `3.390 5.437 6.204 6.204 6.988 8.376`	1.06431424636715
Histogram of X1 (Frequency)	Histogram of X1 (Density)



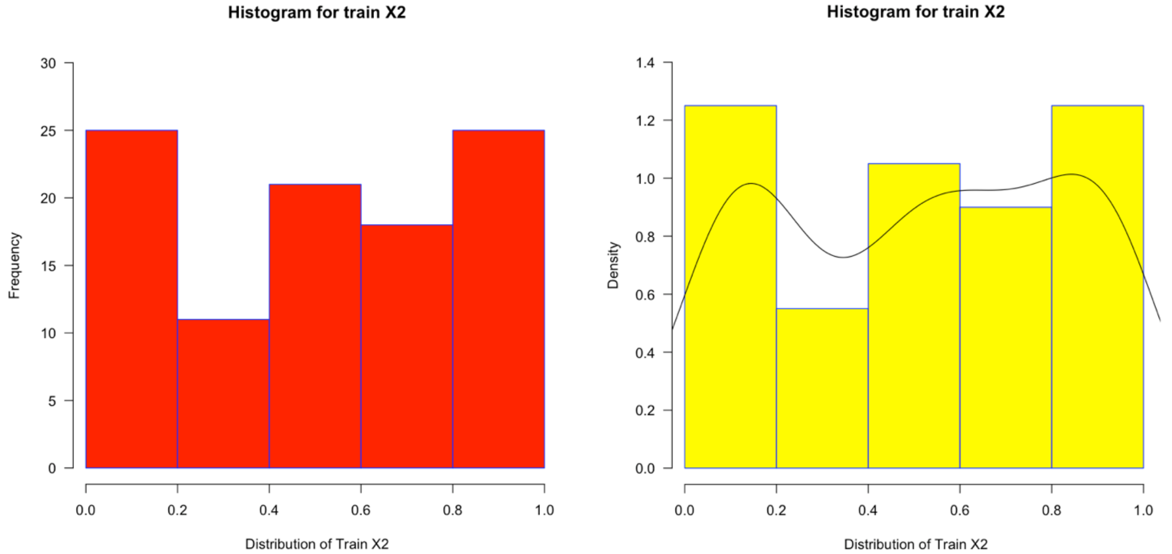


<Figure2. Histogram of train X1>

2. X2

This is not a normal distribution, it is more likely double-skewed or plateau distribution.

Summary of X2	Standard deviation of X2
Min. 1st Qu. Median Mean 3rd Qu. Max. 0.01178 0.19606 0.53263 0.51696 0.79926 0.99709	0.313443557051801
Histogram of X2 (Frequency)	Histogram of X2 (Density)

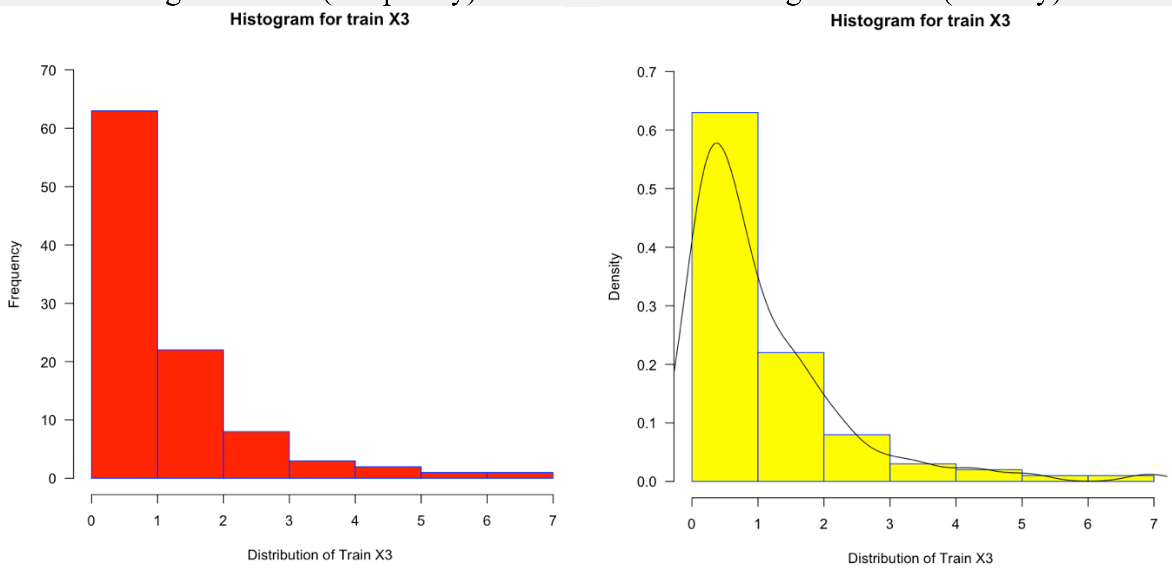


<Figure3. Histogram of train X2>

3. X3

This is right-skewed distribution. This is asymmetrical because a natural limit prevents outcomes on one side [1].

Summary of X3	Standard deviation of X3
Min. 1st Qu. Median Mean 3rd Qu. Max. 0.01178 0.19606 0.53263 0.51696 0.79926 0.99709	1.17707531834911
Histogram of X3 (Frequency)	Histogram of X3 (Density)

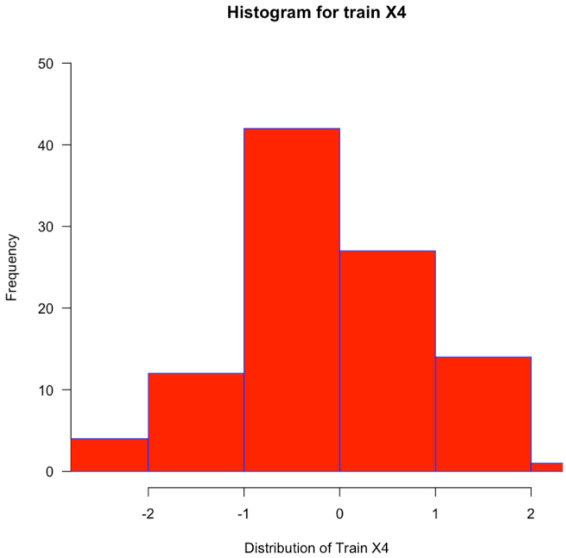


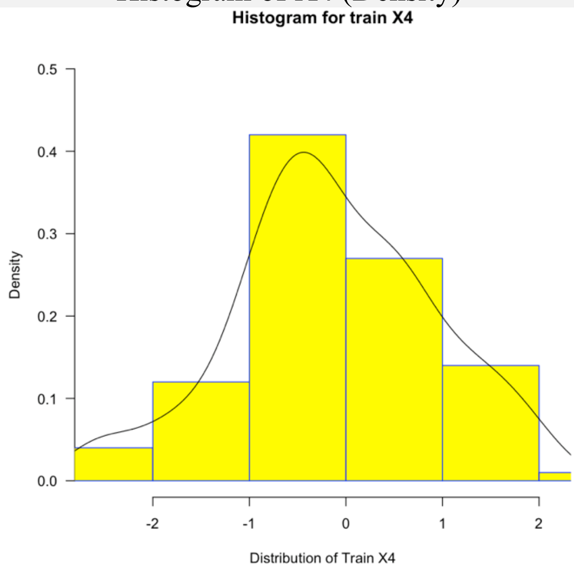
<Figure4. Histogram of train X3>

4. X.4

This is normal distribution. But not perfect normal distribution because of differences of mean, median and mode.

Summary of X4	Standard deviation of X4
Min. 1st Qu. Median Mean 3rd Qu. Max. -2.6189 -0.7669 -0.1968 -0.1340 0.5634 2.1409	1.05274800513095
Histogram of X4 (Frequency)	Histogram of X4 (Density)





<Figure5. Histogram of train X4>

B.2. Linear Regression models

Based on figures below clearly shows that Model1 (0.9402) has higher Multiple R-squared value than Model2 (0.9388). R-squared is statistical measure of how close the data are to the fitted regression line. R-squared is fairly straight-forward. 0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits data [2]. So, Multiple R-squared point of view, model1 is better fit model. However, for predict, significant is more important. Model1 has good significant values except Model1$X. Model2 has strong significant values – All ***.

1. Model1

The figure below shows the details of the linear regression of Model1 and its relationship between variables. These information use to predict Y based on X1~4. The Multiple R-squared value of Model1 is 0.9402.

```
# Linear regression of model1
model1<- lm(Y~(X1+X2+X3+X4)+1, data=train)
summary(model1)
```

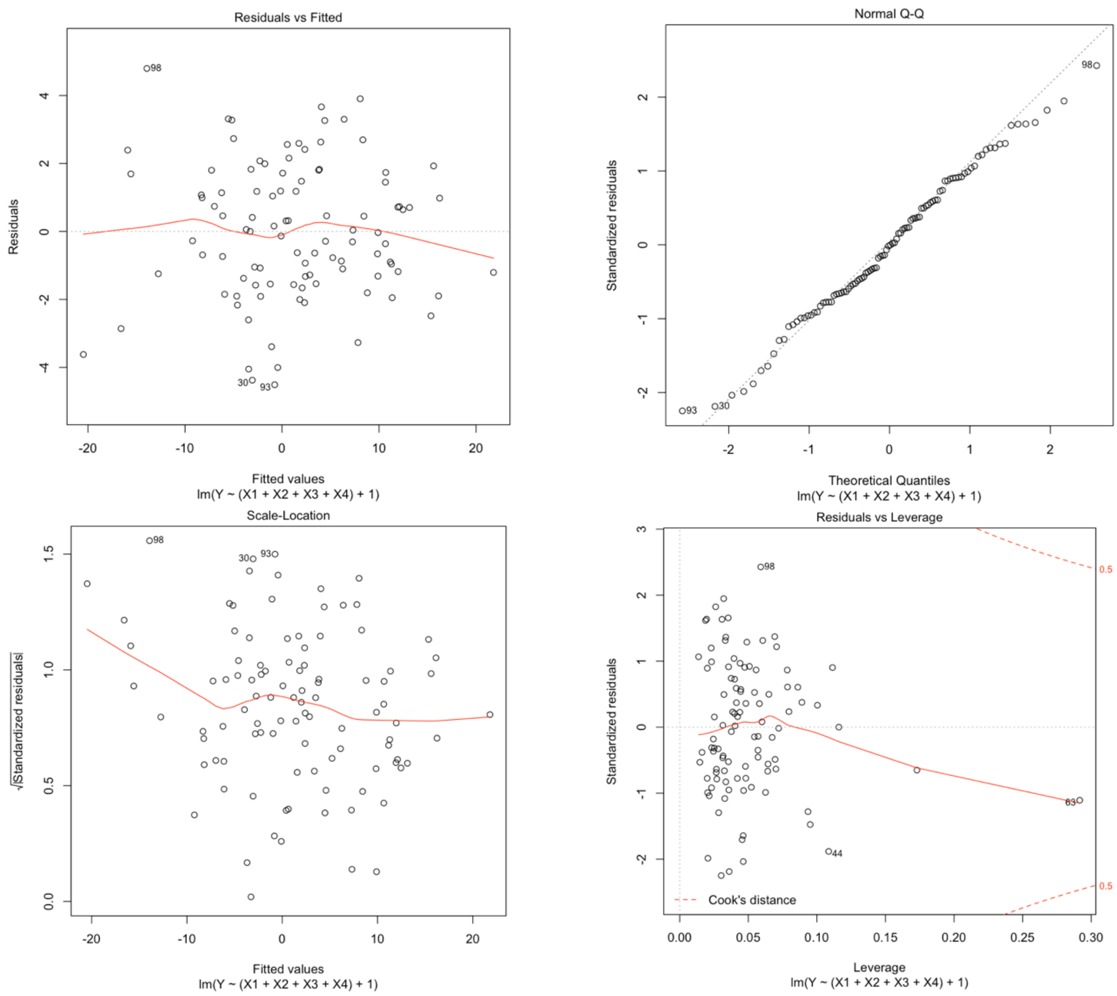
```
Call:
lm(formula = Y ~ (X1 + X2 + X3 + X4) + 1, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5110 -1.3386 -0.0158  1.5315  4.7958 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.7394     1.3259   3.575 0.000554 ***
X1           -0.2850     0.1945  -1.465 0.146156    
X2           -5.5824     0.6609  -8.447 3.42e-13 ***
X3            2.1597     0.1760  12.273  < 2e-16 ***
X4            6.9379     0.1951  35.568  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.037 on 95 degrees of freedom
Multiple R-squared:  0.9402,	Adjusted R-squared:  0.9376 
F-statistic: 373.1 on 4 and 95 DF,  p-value: < 2.2e-16
```

<Figure6. Linear regression of model1>



<Figure7. Plot of linear regression of model1>

2. Model2

The Multiple R-squared value of Model2 is 0.9388.

```
# Linear regression of model1
model2<- lm(Y~(X2+X3+X4)+1, data=train)
summary(model2)
```

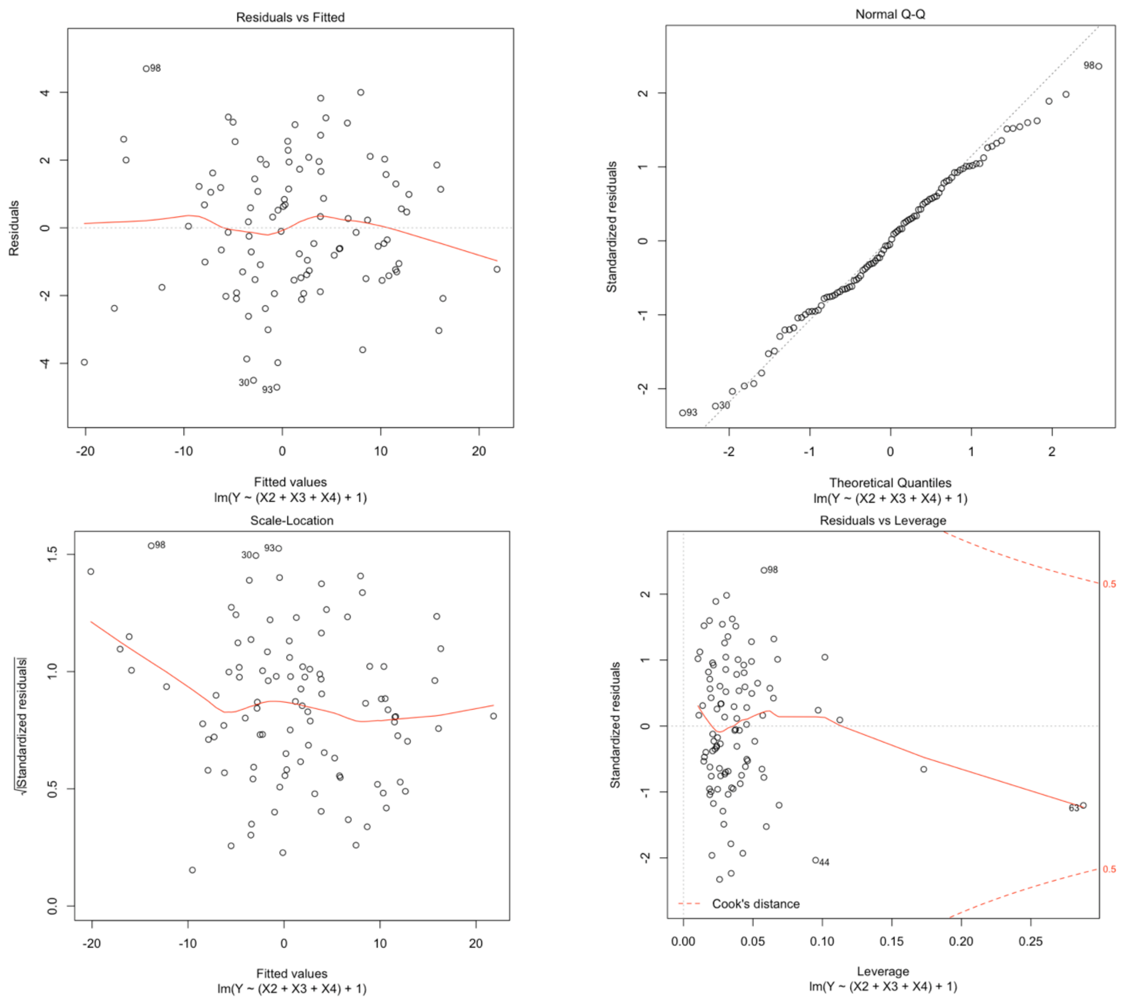
```
Call:
lm(formula = Y ~ (X2 + X3 + X4) + 1, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7054 -1.4289 -0.0285  1.5845  4.6968 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.8978     0.4247   6.823 7.96e-10 ***
X2           -5.4905     0.6618  -8.296 6.70e-13 ***
X3            2.1826     0.1763  12.378  < 2e-16 ***
X4            6.9213     0.1959  35.333  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.049 on 96 degrees of freedom
Multiple R-squared:  0.9388,	Adjusted R-squared:  0.9369 
F-statistic: 490.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

<Figure8. Linear regression of model2>



<Figure9. Plot of linear regression of model2>

3. B.3. Predict R-squared

The figure below clearly shows that Model2 has smaller MSE (2.73729977048782) than Model1(2.87093470564021). This means that Model2 is better model than Model1 because MSE close to 0 is a good model, also Model 2 has better significant values.

```
# Predict model of y based on X1~4
y = test$Y
y1 = predict(lm(formula = Y ~ X1 + X2 + X3 + X4,data = train),test)
y2 = predict(lm(formula = Y ~ X2 + X3 + X4,data=train),test)

# MSE of Model1
MSE1 = sum((y-y1)^2/length(y))
MSE1
```

2.87093470564021

```
# MSE of Model2
MSE2 = sum((y-y2)^2)/length(y)
MSE2
```

2.73729977048782

<Figure10. MSE of model1 and 2>

The higher R square are not always better. The R-squared must evaluate residual plots and other statistics because it is impossible to determine whether the coefficient estimates and predictions are biased. The R squared does not indicate whether the regression model is appropriate. Good models can have low R-squared values, and models that do not fit the data can have high R-squared values [1].

References

[1] Minitab Blog, 2013. Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Retrieved on http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

[2] ASQ. Typical Histogram Shapes and What They Mean. Retrieved on http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/histogram2.html