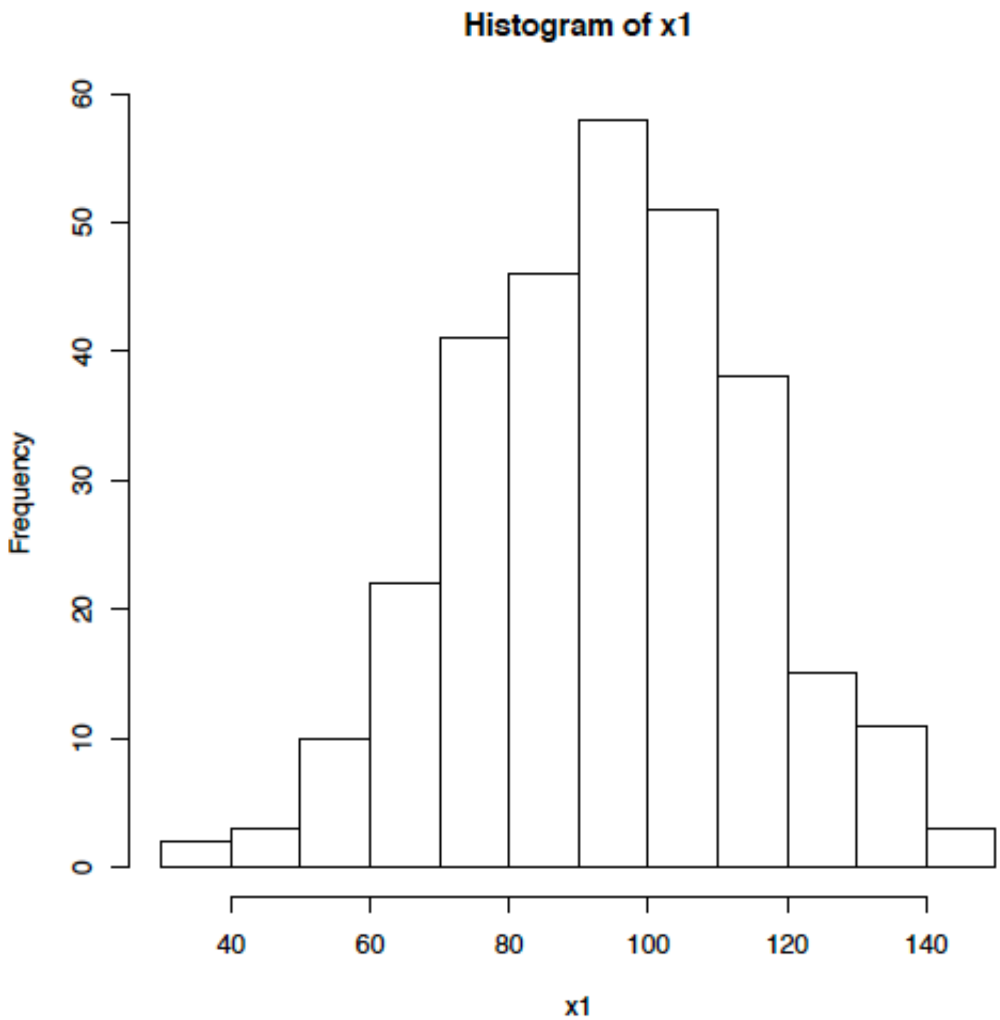
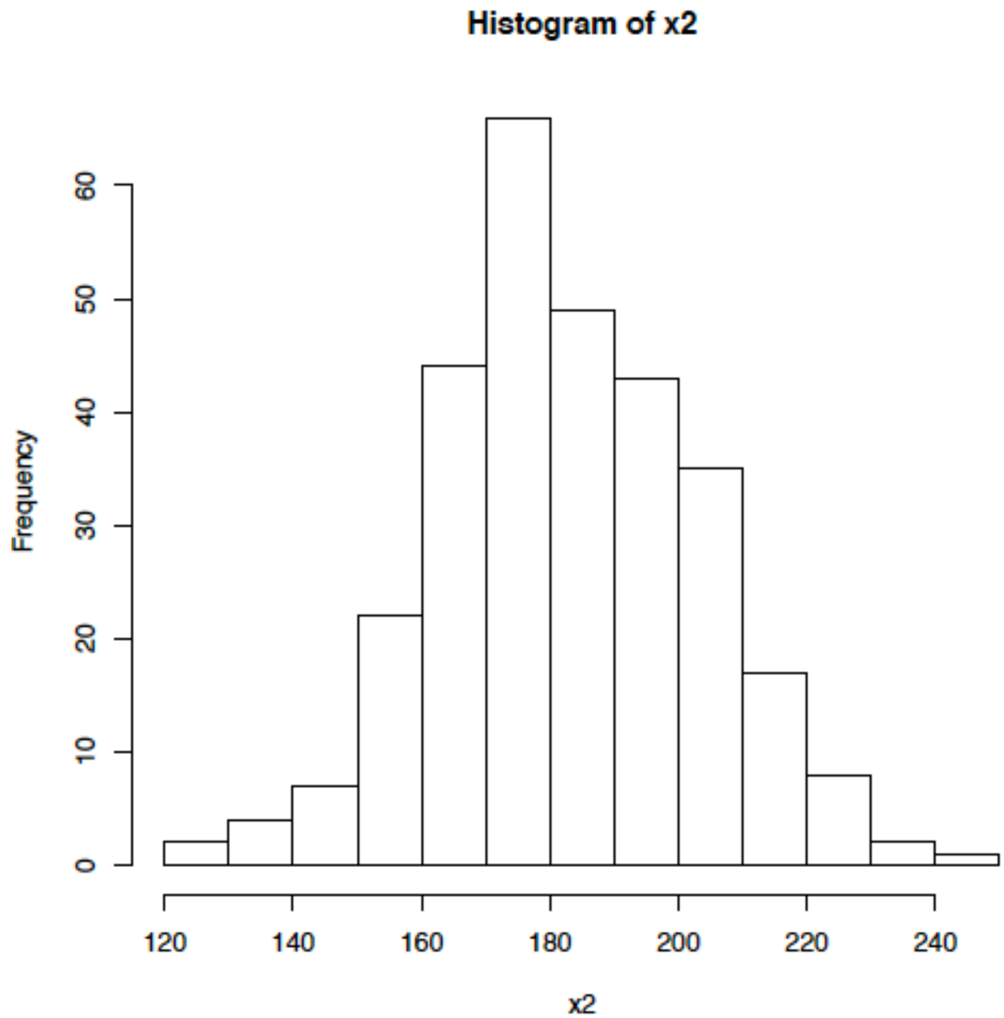
Task 1

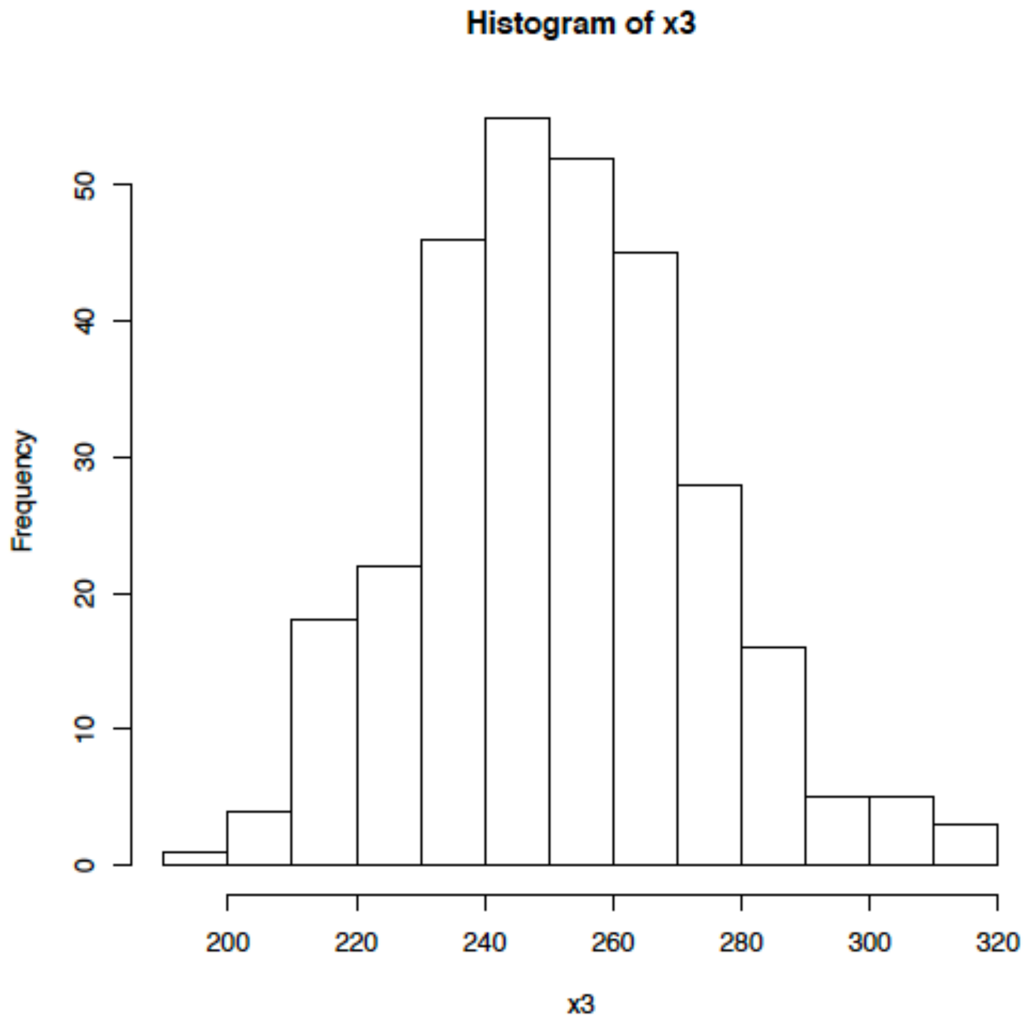
1.1

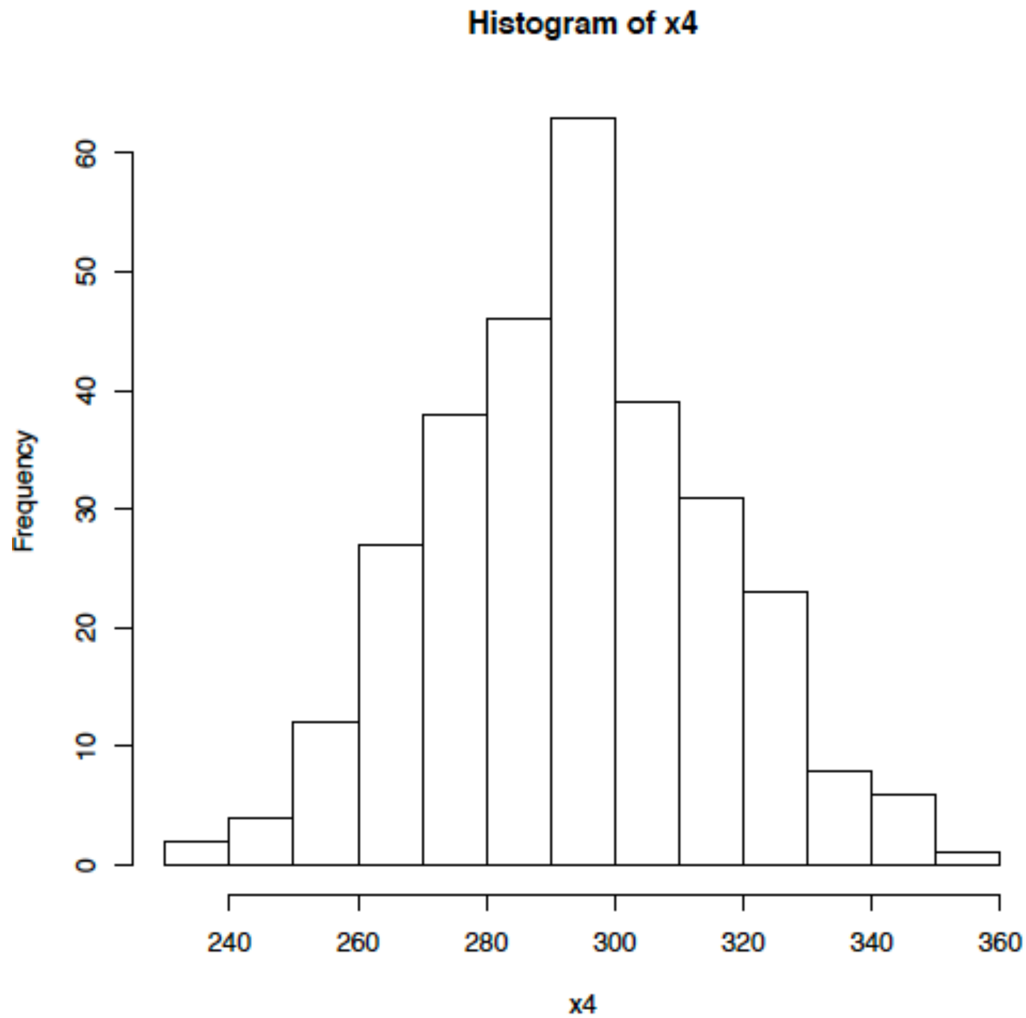
the Histogram of x1,x2,x3,x4,x5 as follows, drawn by R. the x axis is the value of the product x_i and y axis is the frenquecy.


Histogram of x1
Frequency
0
10
20
30
40
50
60
40
60
80
100
120
140
x1

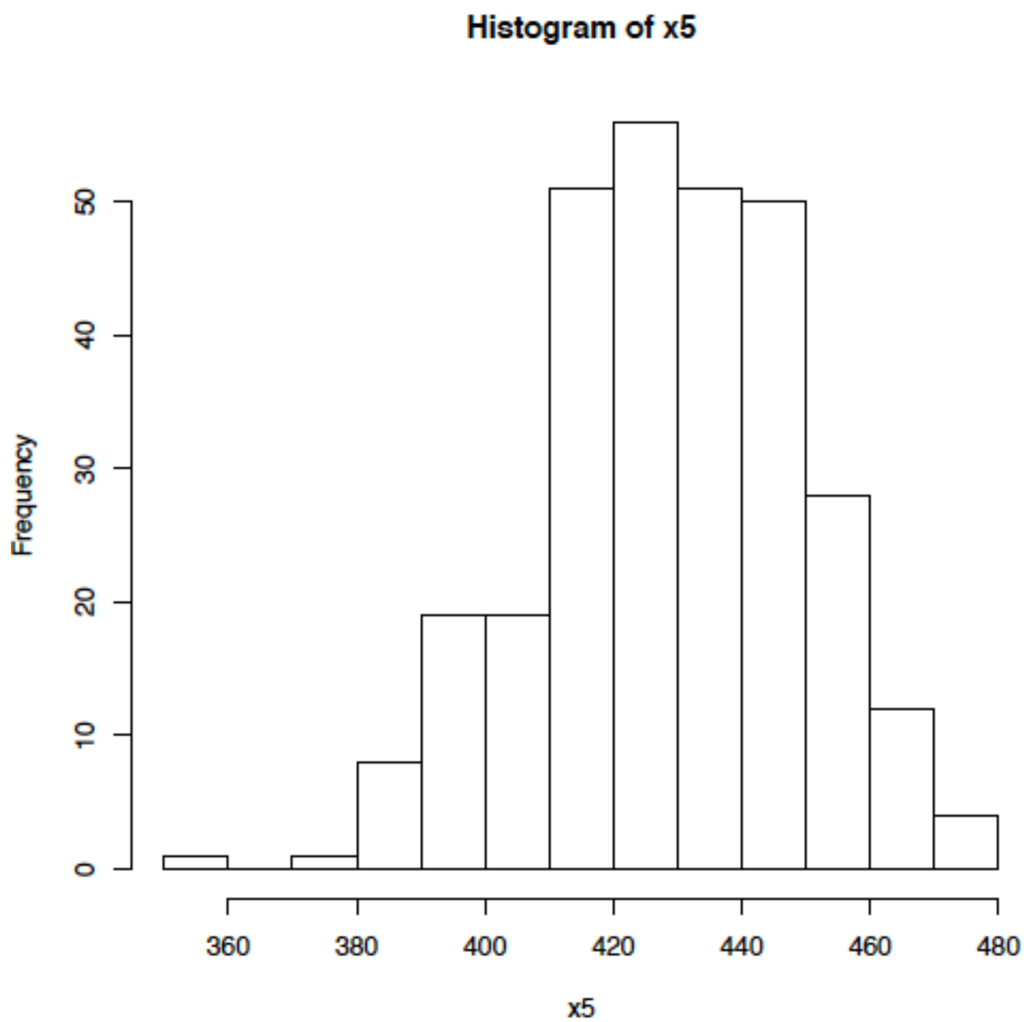


Histogram of x2
Frequency
x2
0
10
20
30
40
50
60
120
140
160
180
200
220
240


Histogram of x3
Frequency
0
10
20
30
40
50
200
220
240
260
280
300
320
x3



Histogram of x4
Frequency
0
10
20
30
40
50
60
240
260
280
300
320
340
360
x4



calculate the mean and variance of x_1 with function mean() and var() in R and gets the result

mean of x1,x2,x3,x4,x5 are 93.94838, 182.8386, 251.7409,293.4849,428.6984 perspectively.

Variance of x1,x2,x3,x4,x5 are 427.4719, 407.3511, 467.5864, 490.0639, 412.2473 perspectively.

1.2

Firstly, we use summary() to get a overall view of the sample data by its min,max, quantile, median, and mean,as shown follows:

> summary(x1)
> Min. 1st Qu. Median Mean 3rd Qu. Max.
> 37.50 79.09 94.76 93.95 108.20 145.58

> summary(x2).
> Min. 1st Qu. Median Mean 3rd Qu. Max.
> 123.3 168.7 181.1 182.8 196.1 244.5

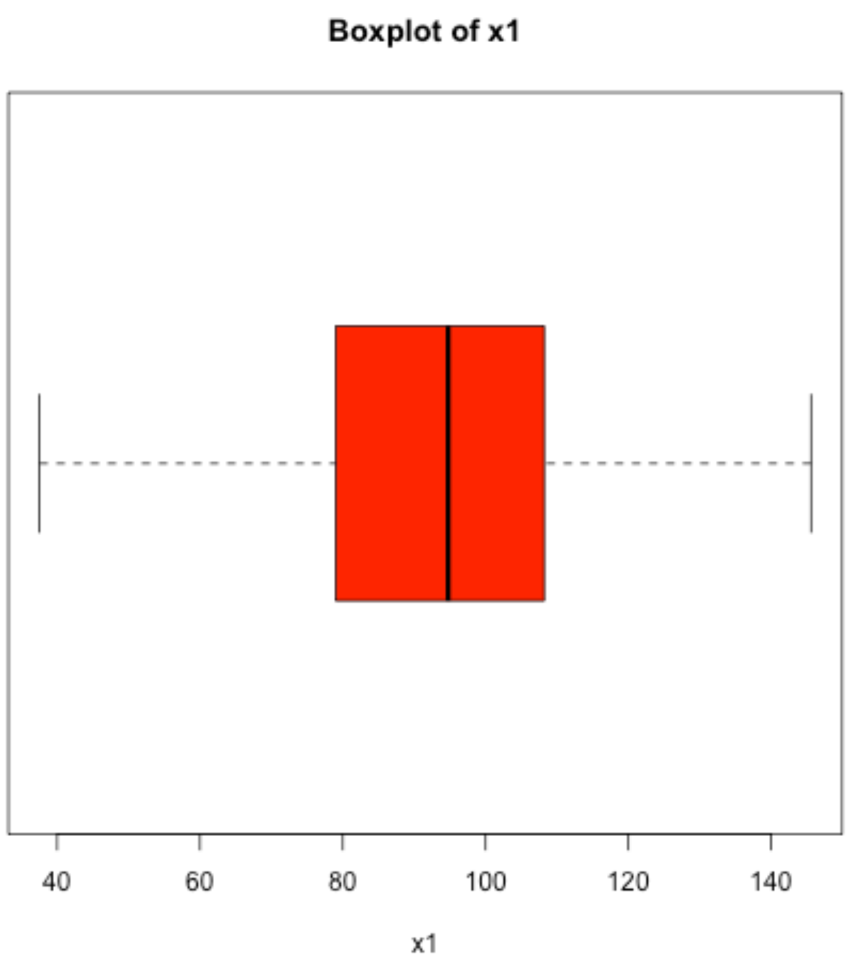
> summary(x3).
> Min. 1st Qu. Median Mean 3rd Qu. Max.
> 199.5 236.9 250.6 251.7 265.8 319.5

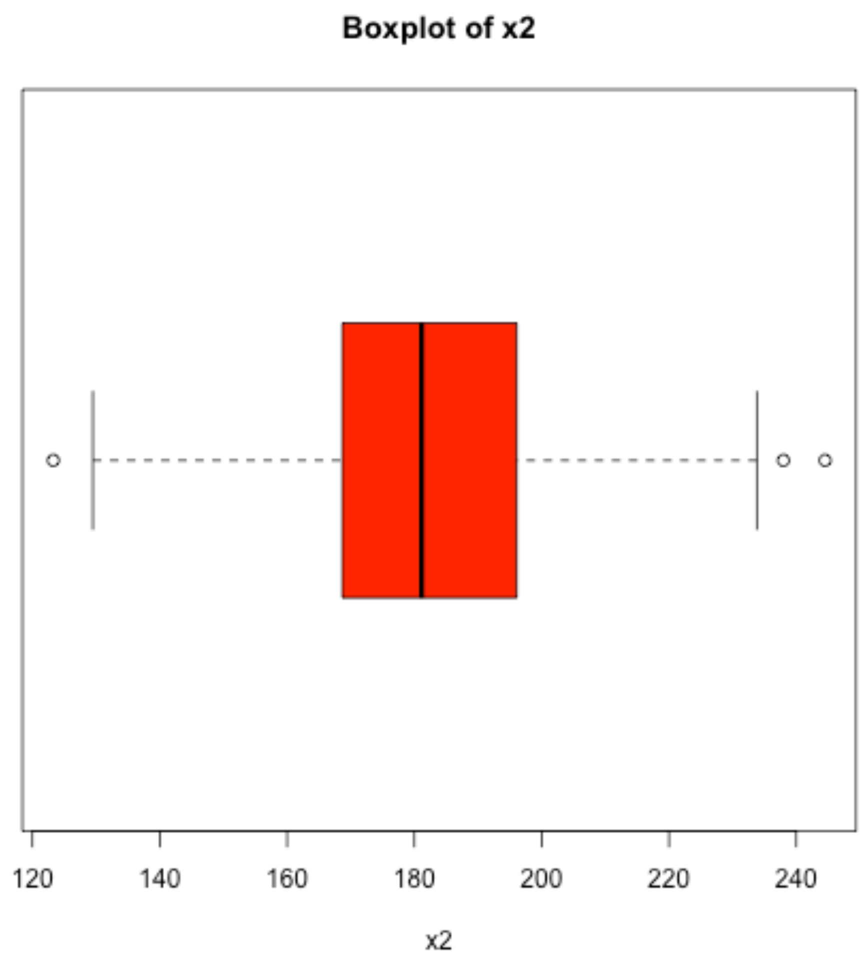
> summary(x4).
> Min. 1st Qu. Median Mean 3rd Qu. Max.
> 230.8 278.0 293.2 293.5 307.3 358.0

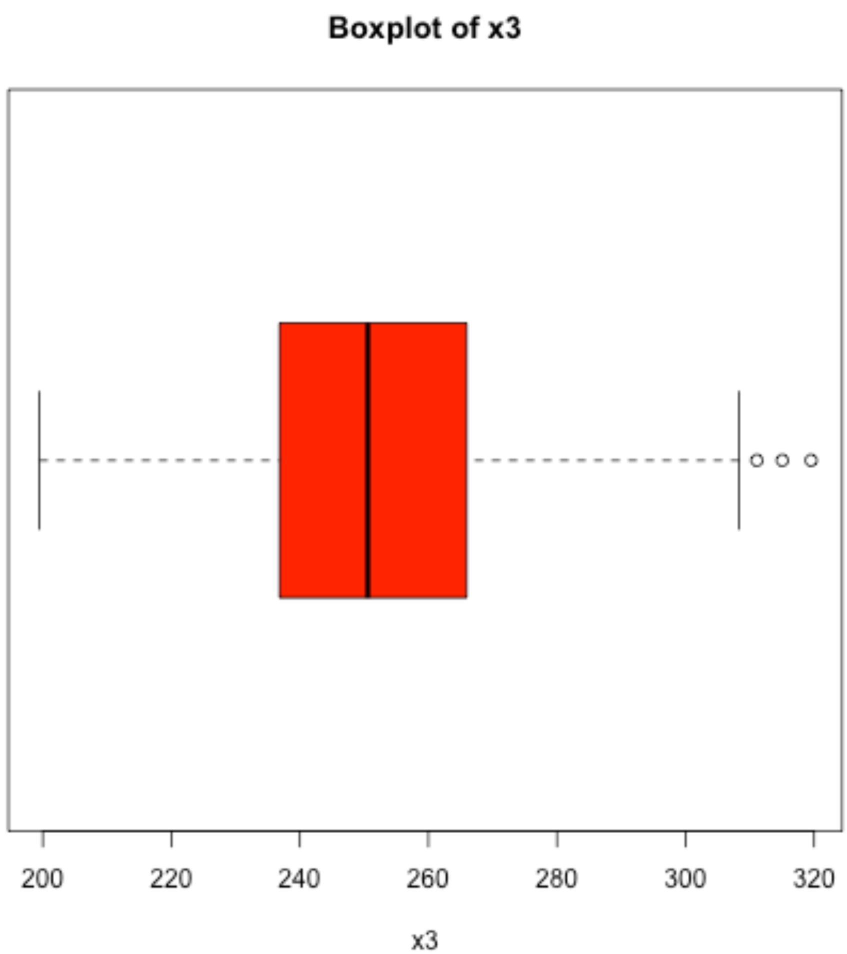
> summary(x5).
> Min. 1st Qu. Median Mean 3rd Qu. Max.
> 355.0 414.9 428.4 428.7 443.6 478.9

then we use boxplot() to remove the outlier, shown as follows: from the graph,we knows that :

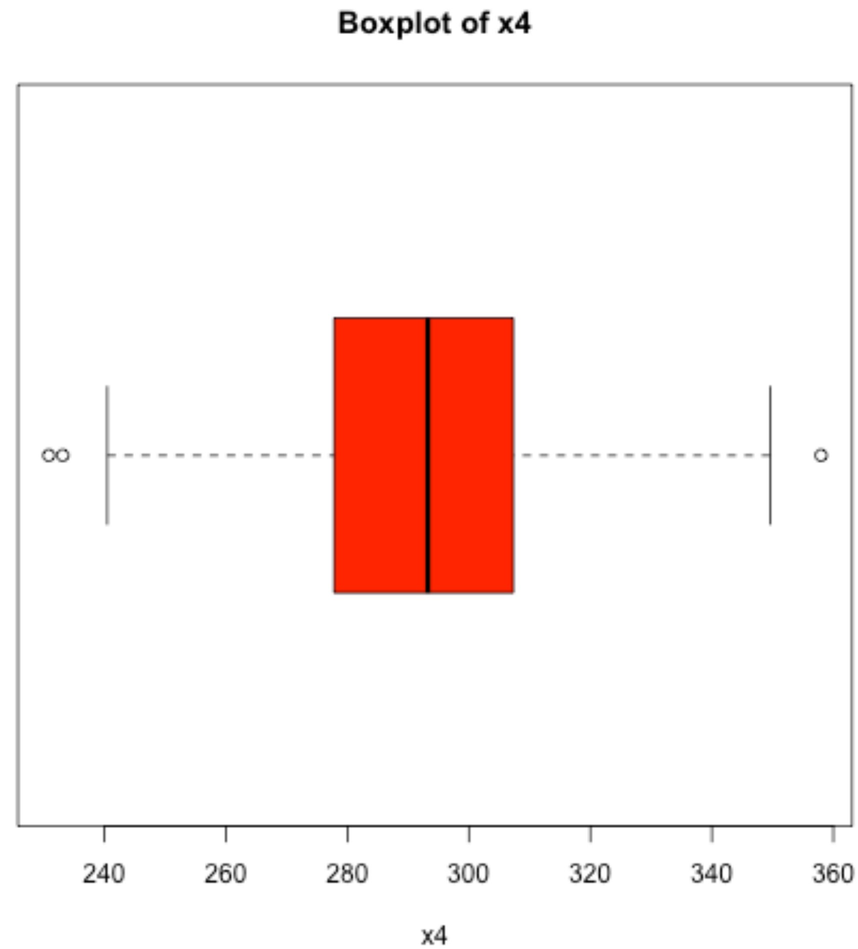
- there is no outlier removed for x1.
- there are three outliers removed for x2, among which two numbers are too larger and one is too small
- there are three outliers removed for x3, among which all are too large.
- there are three outliers removed for x4, among which two are too small and one is too large
- there is one outliers that was too small removded for x5

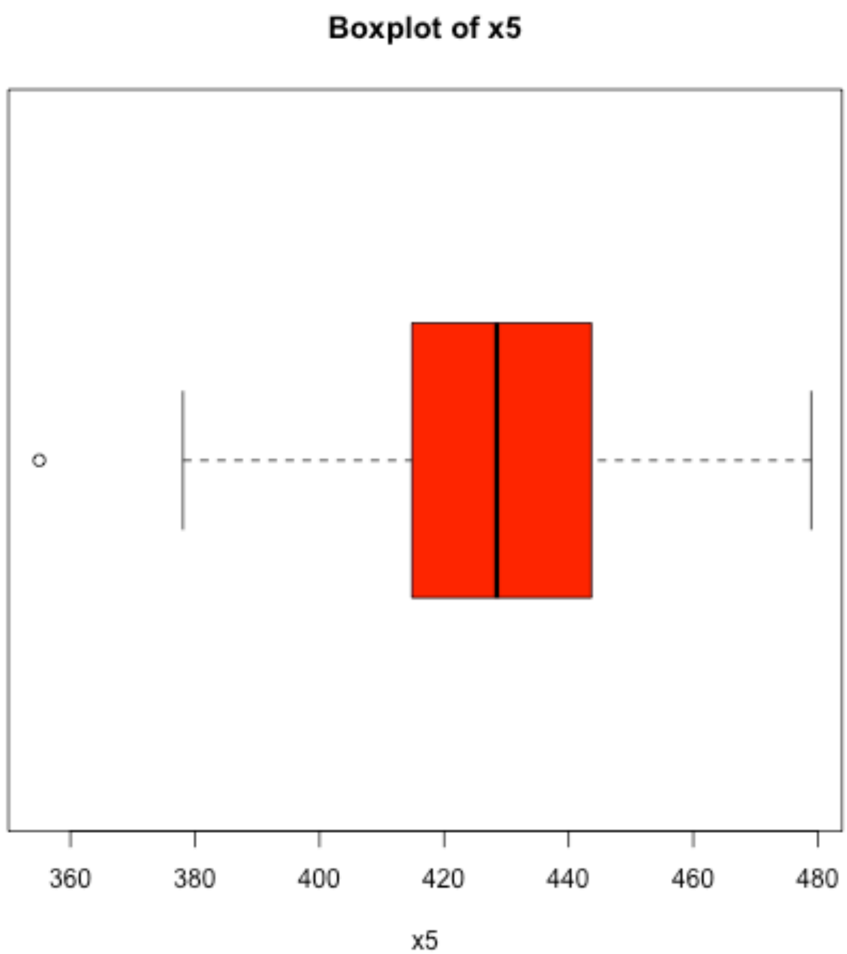



Boxplot of x2
120
140
160
180
200
220
240
x2



Boxplot of x3
200
220
240
260
280
300
320
x3


Boxplot of x4
240
260
280
300
320
340
360
x4



1.3

we use cor() fuction in R to calculate the matrix

```
> r <- cor(data)
> round(r,4)
       x1      x2     x3      x4     x5      y
x1 1.0000  0.1048 0.0627  0.0091 0.0243 0.9882
x2 0.1048  1.0000 0.0815 -0.0220 0.0273 0.1038
x3 0.0627  0.0815 1.0000  0.1618 0.0608 0.0893
x4 0.0091 -0.0220 0.1618  1.0000 0.1578 0.0501
x5 0.0243  0.0273 0.0608  0.1578 1.0000 0.0478
y  0.9882  0.1038 0.0893  0.0501 0.0478 1.0000
> round(r,3)
```

dependency conclusion:

- the diagonal value is always 1.0 because a variable is correlated with itself so that is always 1.0
- Dependency of x_i and y: we can see that the correlation between x1 and y is 0.9882, it is the strongest correlations, and the correlation between x5 and y is 0.0478 which is the weakest.
- Dependency between x_i : we can see the dependency between them are weak, so we can consider they are independent.

1.4

to check it of the histogram. we can see that and outlier doing is reasonable which is not doing too much and keep the characteristics of the variable distributions

from the correlation matrix. we can draw the preliminary conclusion that the x1 is most correlated with y, and there is no or little multicollinearity of the indepent variable as a whole. If strictly, we can do hypothesis test of the correlation beween x1 with x2 or x3 with x4 or x4 with x5 ,which the correlation is 0.1, 0.16, 0.15 perspectively to do further diagnosis.

Task 2

2.1

Y = $a_0 + a_1 x_1$ + ε.

we can use the following function to gets the $a_0 = 322.896, a_1 = -6958.084, \sigma^2 = 1074345$

> slr <- lm(y~x1).
> summary(slr).

> Residuals:
> Min 1Q Median 3Q Max.
> -1666.3 -640.4 -304.2 269.2 4768.3

> Coefficients:
> Estimate Std. Error t value Pr(>|t|)
> (Intercept) -6958.084 278.874 -24.95 <2e-16 ***.

> x1 322.896 2.899 111.37 <2e-16 ***.

> Signif. codes: 0 ‘*’ **0.001** ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Residual standard error: 1037 on 298 degrees of freedom.
> Multiple R-squared: 0.9765, Adjusted R-squared: 0.9765
> F-statistic: 1.24e+04 on 1 and 298 DF, p-value: < 2.2e-16.
>
> > (summary(slr)$sigma)**2.
> > [1] 1074345

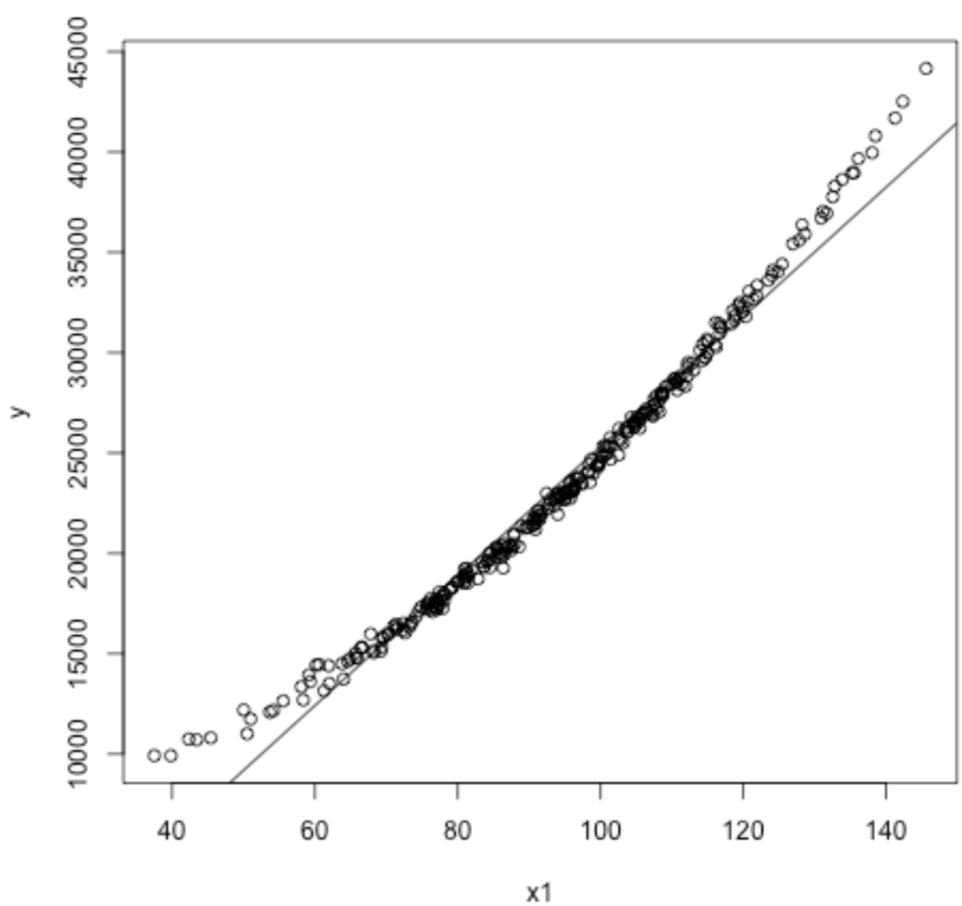
2.2

we use statistics package to calculate the P-value, R square,and F.

Conclusion: The reported p-values for both tails are 5.821915e-75 and7.099010e-245 for the intercept and slope respectively. Consequently, we reject the null hypothesis that intecept and slope are zero at 90%, 95% and 99% confidence, Hence there is a significant relationship between y and x1 in the linear regression model. the R square is 0.9765, which shows that approximately 97.65% of variation in y can be explained by x1. it also confirm that there is a significant relationship between y and x1 in the linear regression model. Similar conclusion can be drawn from F.

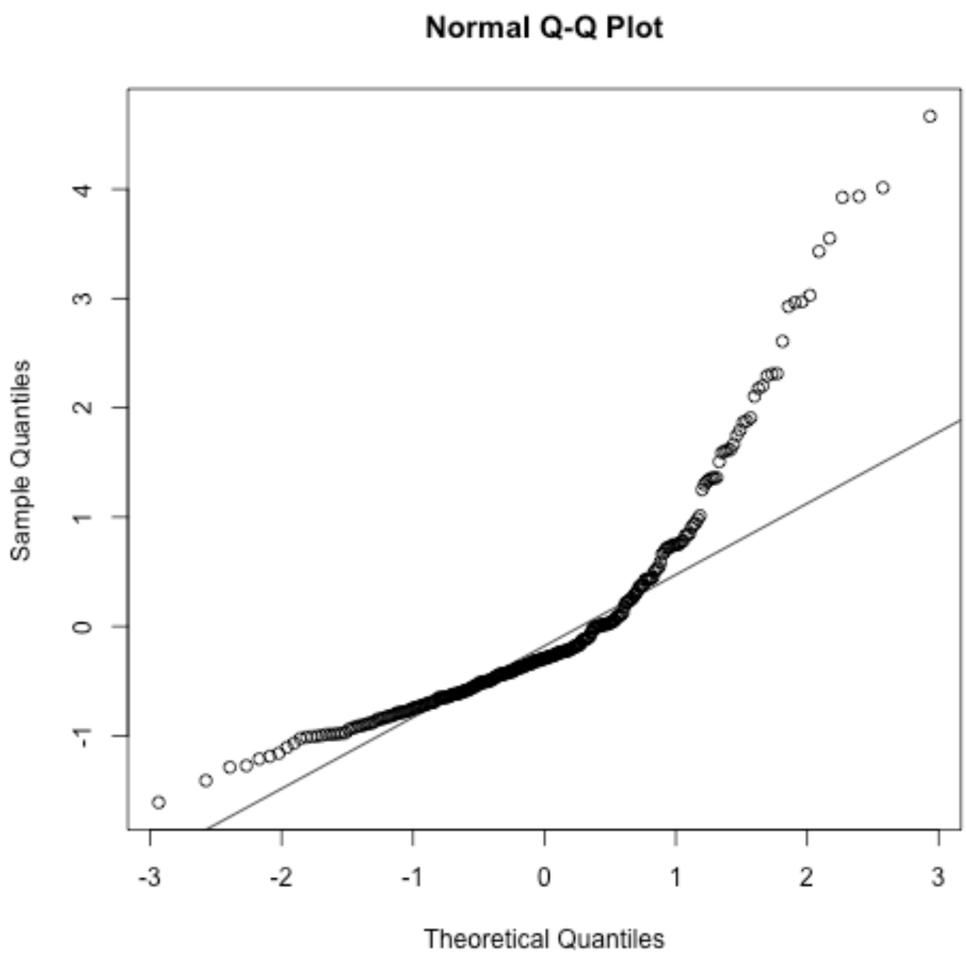
> summary(slr)$coefficients[,4].
> (Intercept) x1
> 5.821915e-75 7.099010e-245
> summary(slr)$r.squared.
> [1] 0.976539.
> summary(slr)$f.
> value numdf dendf
> 12403.94 1.00 298.00

2.3

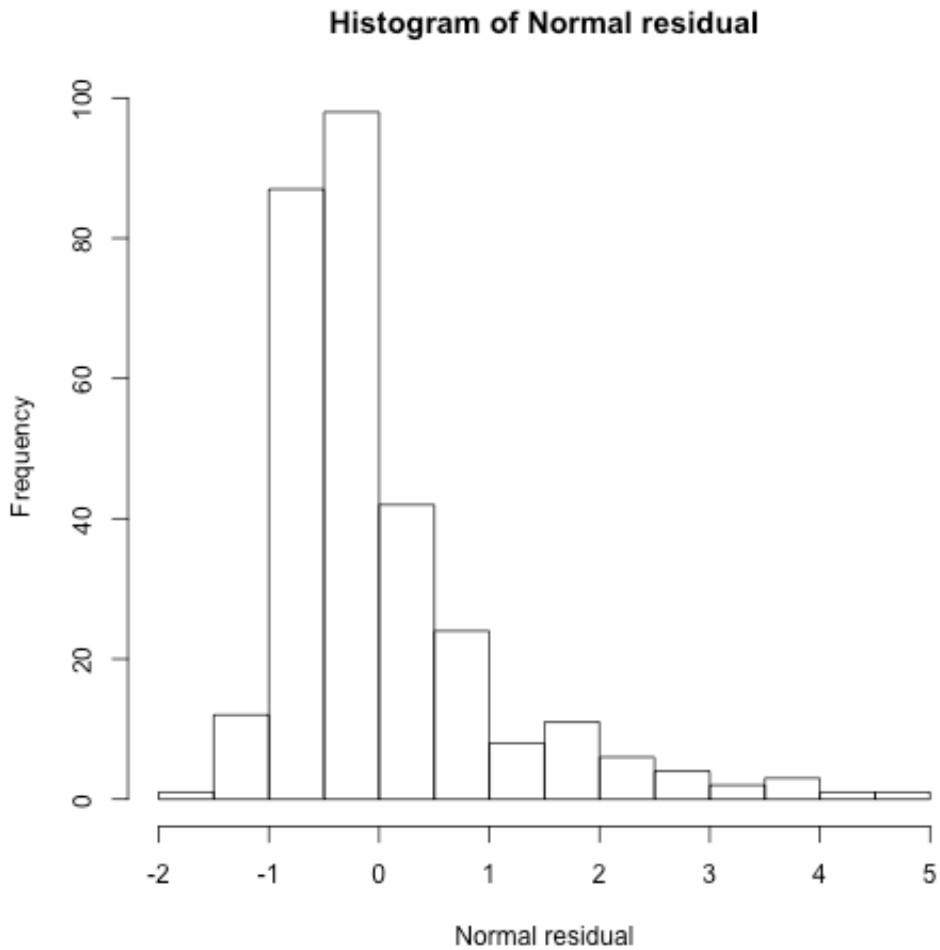


2.4

a)



Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-1
0
1
2
3
4
-3
-2
-1
0
1
2
3



we use ks.test to carry out the χ^2 test . From the reslut. we can reject null hypothesis, means we have strong confidence that residual follows the normal distribution N(0, s^2)

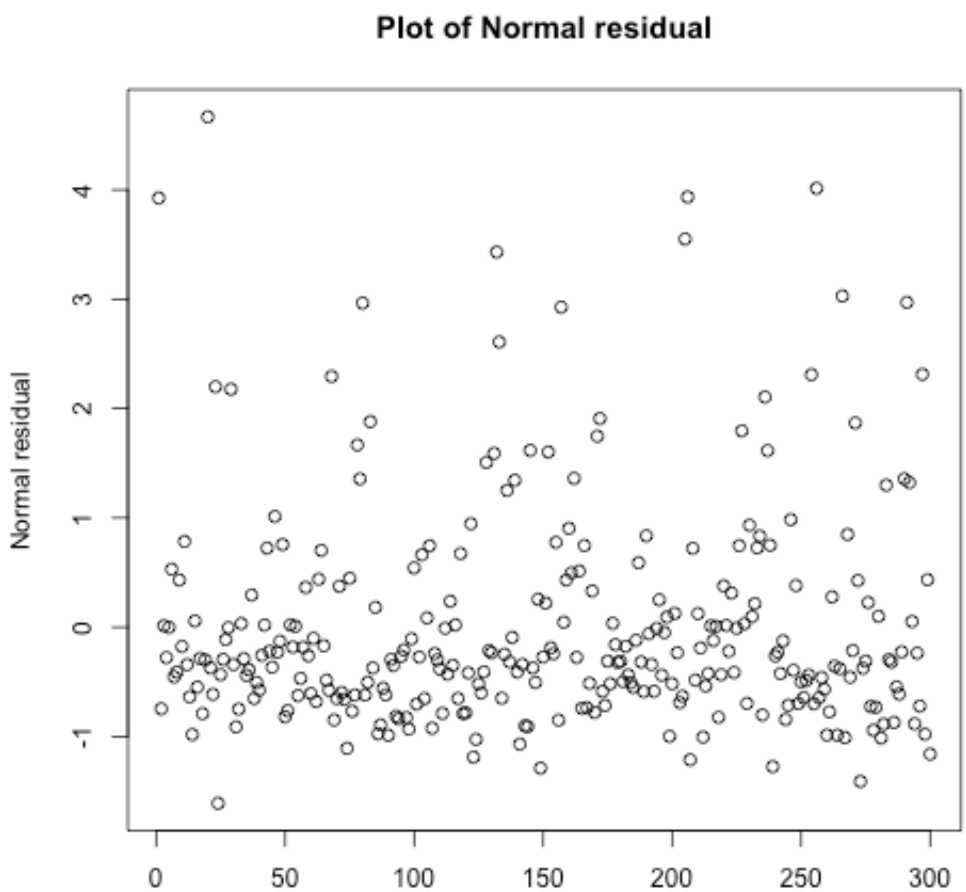
> slr.res <- resid(slr.lm).
> ks.test(slr.res,"pnorm",mean(slr.res),sd(slr.res)).

> One-sample Kolmogorov-Smirnov test.

> data: slr.res.
> D = 0.18382, p-value = 3.134e-09.
> alternative hypothesis: two-sided

b) the residual scatter plot drawn as follows.

Conclusion: residuals have no correlation trends



2.7

we use model <- lm(y~ x1+I(x1^2)) to generate the fuction of Y = $a_0 + a_1x_1 + a_2x_1^2$+ ε, to gets $a_0 = 7.328e+03, a_1 = 7.472e-02, a_2 = 1.734e+00$,

From the result the P-value of a_1 is 0.99, So we can accept null hypothesis that the slope of x_1 is 0 ,and the p_value of a_2 is smaller than 2e-16, which we reject the null hypothesis. So it has a signification for it. From the R square value 0.998. it means that about 99.8% of variation in y can be explained by x_1^2 . it gets a better result.

> model <- lm(y~ x1+I(x1^2)).
> summary(model)

> Call:
> lm(formula = y ~ x1 + I(x1^2)).

> Residuals:
> Min 1Q Median 3Q Max
> -1004.37 -213.99 26.99 210.78 842.06

> Coefficients:
> Estimate Std. Error t value Pr(>|t|)
> (Intercept) 7.328e+03 2.680e+02 27.346 <2e-16 ***.
> x1 7.472e-02 5.827e+00 0.013 0.99

> I(x1^2) 1.734e+00 3.096e-02 56.008 <2e-16 ***.

> Signif. codes: 0 ‘***’ 0.001** ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Residual standard error: 305.3 on 297 degrees of freedom.
> Multiple R-squared: 0.998, Adjusted R-squared: 0.998
> F-statistic: 7.303e+04 on 2 and 297 DF, p-value: < 2.2e-16

2.8

From Task 1, we knows that x_1 has the strongest correlation with y, So we did SLR between (y,x1), and results confirm us this conclusion. From the p value, we concluse that there a significant relationship between y and x1 in the linear regression model. the R square tells us that approximately 97.65% of variation in y can be explained by x1. We also did polynomial regression between y and x_1 and results shows that about 99.8% of variation in y can be explained by x_1^2 .

From the residuals analysis. we carried out χ^2 test and Q-Q plot and lead to the conclusion that residual follows the normal distribution N(0,σ^2). From the scatter plot of residuals, we conclude that residuals should have no trends.

Task 3

3.1

we use mlr <- lm(y~x1+x2+x3+x4+x5) fuction to get the multiple linear regression.

the coeffiecients of x1,x2,x3,x4,x5 are 3.223e+02, -3.353e-01,6.492e+00,1.071e+01, 5.698e+00 perspectively. and $\sigma^2 = 976653$

> mlr <- lm(y~x1+x2+x3+x4+x5).
> summary(mlr).

> Call:
> lm(formula = y ~ x1 + x2 + x3 + x4 + x5)

> Residuals:
> Min 1Q Median 3Q Max
> -976.5 -623.5 -348.5 216.7 4580.8

> Coefficients:
> Estimate Std. Error t value Pr(>|t|)
> (Intercept) -1.406e+04 1.511e+03 -9.303 < 2e-16 ***.**
> **x1 3.223e+02 2.784e+00 115.747 < 2e-16 *.**

> **x2 -3.353e-01 2.858e+00 -0.117 0.9067**
> **x3 6.492e+00 2.694e+00 2.410 0.0166 ***
> **x4 1.071e+01 2.648e+00 4.043 6.74e-05 ***

> x5 5.698e+00 2.854e+00 1.996 0.0468 *

> Signif. codes: 0 '**' **0.001** '' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Residual standard error: 988.3 on 294 degrees of freedom.
> Multiple R-squared: 0.979, Adjusted R-squared: 0.9786
> F-statistic: 2736 on 5 and 294 DF, p-value: < 2.2e-16.
>
> > (summary(mlr)$sigma)**2.
> > [1] 976653.8

3.2

The p-values for x1,x2,x3,x4,x5 are < 2e-16, 0.9067, 0.0166, 6.74e-05, 0.0468 perspectively. $R^2 = 0.979$ and F value is 2736 on 5 and 294 DF, the correlation matrix is shown in task 1.3. The p-value of x2 is 0.9067, it shows that we accept the null hypothesis that the slop for x2 = 0, and according to the correlation matrix. the correlation between x1,x2 is 0.1. So maybe we can try remomve x2. So redo the mulitple linear regression with (y,x1,x3,x4,x5) ,as shown before:

From the result.we knows that the R^2 is still 0.979, it shows that approxiamtely 97.9% of variantion in y can be explained by our model of x1,x3,x4,x5. same with model of x1,x2,x3,x4,x5. So we can comfirm that it is reasonable to remove x2 above.

> mlrAdjust <- lm(y~x1+x3+x4+x5).
> summary(mlrAdjust)

> Call:
> lm(formula = y ~ x1 + x3 + x4 + x5).

> Residuals:
> Min 1Q Median 3Q Max
> -981.7 -627.7 -350.2 217.8 4581.5

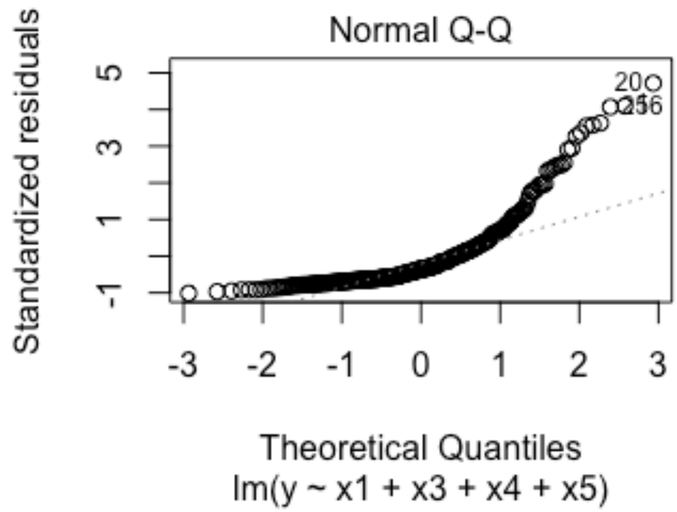
> Coefficients:
> Estimate Std. Error t value Pr(>|t|)
> (Intercept) -14108.909 1442.810 -9.779 < 2e-16 ***.
> x1 322.232 2.766 116.510 < 2e-16 *.
> x3 6.467 2.681 2.413 0.0164 *
> x4 10.720 2.642 4.058 6.35e-05 *.

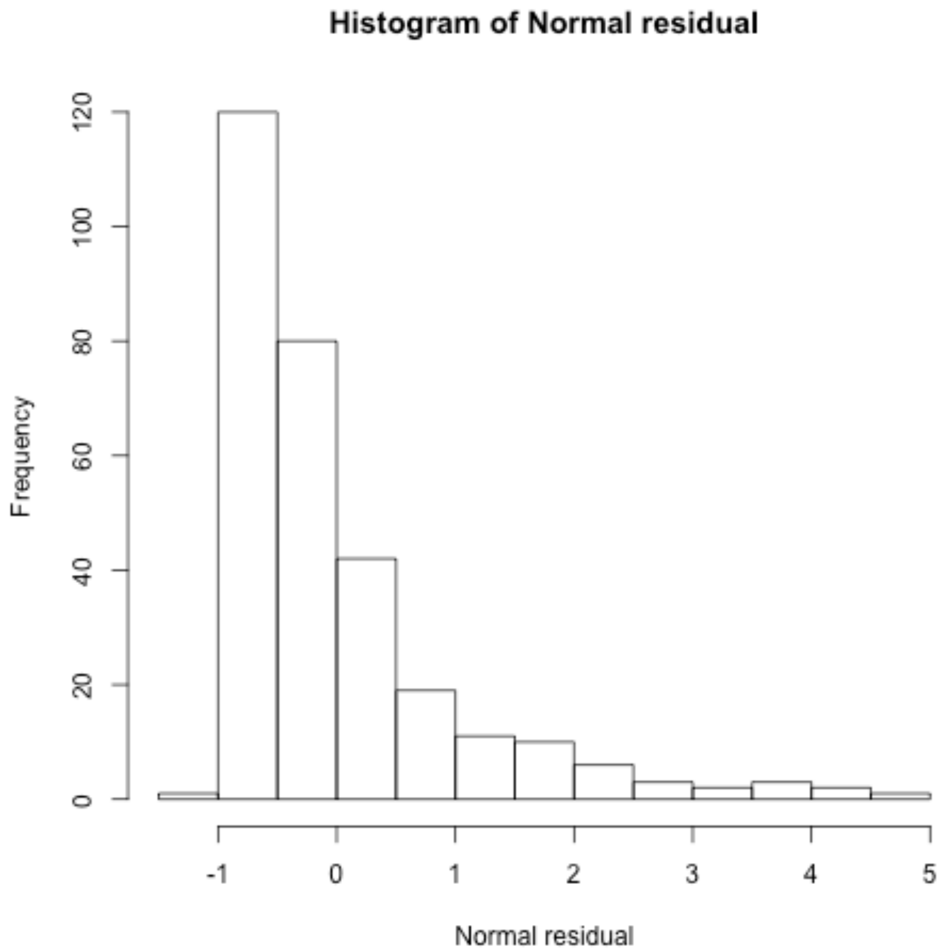
> x5 5.689 2.848 1.997 0.0467 *

> Signif. codes: 0 '**' **0.001** '' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Residual standard error: 986.6 on 295 degrees of freedom.
> Multiple R-squared: 0.979, Adjusted R-squared: 0.9787
> F-statistic: 3431 on 4 and 295 DF, p-value: < 2.2e-16

3.3





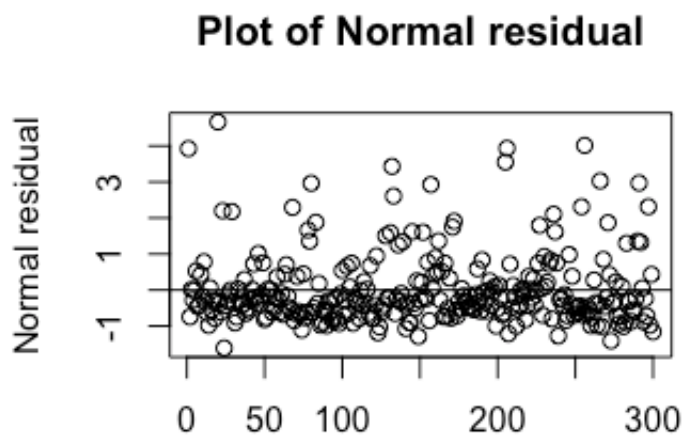
we carried out χ^2 test by the following code. From the reslut, we can reject null hypothesis, means we have strong confidence that residual follows the normal distribution N(0, s^2)

> mlrAdjust.res <- resid(mlrAdjust).
> ks.test(mlrAdjust.res,"pnorm",mean(mlrAdjust.res),sd(mlrAdjust.res))

> One-sample Kolmogorov-Smirnov test

> data: mlrAdjust.res.
> D = 0.18678, p-value = 1.621e-09.
> alternative hypothesis: two-sided

b) the residual scatter plot drawn as follows. Conclusion: residuals have no correlation trends



3.4

We did MLR for (y,x1,x2,x3,x4,x5). the R^2 is 0.979, it shows that approxiamtely 97.9% of variantion in y can be explained by our model of x1,x2,x3,x4,x5. and the p-value of x2 shows that we accept the null hypothesis that the slop for x2 = 0,which means that there is no or little signification between y and x2. so we remove x2 and redid regression for (y,x1,x3,x4,x5) ,From the result.we knows that the R^2 is still 0.979, it shows that approxiamtely 97.9% of variantion in y can be explained by our model of x1,x3,x4,x5. same with model of x1,x2,x3,x4,x5. So we can comfirm that it is reasonable to remove x2 above and x_2 has no signification on y.

From the residuals analysis,we carried out χ^2 test and Q-Q plot and lead to the conclusion that residual follows the normal distribution N(0,σ^2). From the scatter plot of residuals, we conclude that residuals have no trends.