INDIAN STATISTICAL INSTITUTE

POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS (PGDBA): 2019–20

Course: STATISTICAL STRUCTURES IN DATA

Assignment 1

Name – Suman Pal
Roll number – 19BM6JP22

- *Use R to solve the problems.*
- *All relevant R programming code should be submitted for evaluation, together with a properly-formulated report* **(as a pdf document),** *containing intermediate and final outputs (including plots, if any),* ***with explanation wherever required, and comments if asked for.***
- *The data used should also be submitted as a .txt or .csv file.*
- *Submission should be emailed to* pamita@isical.ac.in*and* arijitpynestat@gmail.com. *The e-mail should have as its subject* **SSD Assignment no. I** *followed by your roll number.*
- *Extra credit will be given for individual effort.*
- *There will be penalty for copying.*
- *Submissions received after the deadline will be summarily rejected*

.

Identify a dataset consisting of a minimum of 1000 (raw) observations on at least two discrete-valued and two continuous-valued variables.

---

1. **Write a short paragraph describing the dataset, including the significance of the individual variables. Do not forget to mention the source of the data, providing appropriate links.**

Description of Dataset:

This dataset is from the streets of Ottawa Ontario Canada. From the years 2010 to 2018. The dataset is the number of vehicles (as an integer) observed before the first bike is observed that passed specific measurement locations in a given interval for various days. Each column represents a measurement location. The counters are considered accurate to within a range of +0%, -5% of the vehicles that cross over the sensing section of the pathway or bike lane.

Variables:

Variables	Significance
location_name	Name of the location
location_id	Unique location ID
count	Count of vehicle before 1st bike is observed
day	Day no. starting from 2010
day_of_year	Day no. starting from a given year
day_of_week	Day no. starting from a week
MaxTemp	Maximum Temperature
MeanTemp	Mean Temperature

MinTemp	Minimum Temperature
SnowonGrndcm	Snow on Ground (in cm)
TotalPrecipmm	Total Precipitation (in mm)
TotalRainmm	Total Rain (in mm)
TotalSnowcm	Total snow on Ground (in cm)

Summary of Data:

```
> summary(data)                                   ## Summary of dataset ##
 location_name  location_id         count              day          day_of_year      day_of_week         MaxTemp
 OBVW    :406   Min.    : 2.000   Min.    : 1.000   Min.    :  27   Min.    :  0.0   Min.    :0.000   Min.    :-24.500
 ORPY    :360   1st Qu.: 8.000   1st Qu.: 2.000   1st Qu.:1468   1st Qu.: 21.0   1st Qu.:1.000   1st Qu.: -7.475
 ALEX    :230   Median : 8.000   Median : 3.000   Median :2192   Median : 46.0   Median :3.000   Median : -3.000
 OYNG    :125   Mean   : 7.954   Mean   : 4.035   Mean   :1979   Mean   :101.1   Mean   :2.952   Mean   : -3.126
 SOMO    :120   3rd Qu.:10.000   3rd Qu.: 6.000   3rd Qu.:2588   3rd Qu.: 83.0   3rd Qu.:5.000   3rd Qu.:  1.700
 CRTZ    : 61   Max.    :13.000   Max.    :10.000   Max.    :3286   Max.    :365.0   Max.    :6.000   Max.    : 16.000
 (Other):108
    MeanTemp           MinTemp        SnowonGrndcm    TotalPrecipmm     TotalRainmm       TotalSnowcm
 Min.    :-26.800   Min.    :-29.20   Min.    : 0.00   Min.    : 0.00   Min.    : 0.0000   Min.    : 0.000
 1st Qu.:-12.700   1st Qu.:-17.48   1st Qu.:10.00   1st Qu.: 0.00   1st Qu.: 0.0000   1st Qu.: 0.000
 Median : -7.000   Median :-11.30   Median :20.00   Median : 0.00   Median : 0.0000   Median : 0.000
 Mean   : -7.438   Mean   :-11.72   Mean   :20.26   Mean   : 2.19   Mean   : 0.7993   Mean   : 1.607
 3rd Qu.: -2.300   3rd Qu.: -6.00   3rd Qu.:29.00   3rd Qu.: 2.00   3rd Qu.: 0.0000   3rd Qu.: 1.575
 Max.    :  9.400   Max.    :  4.90   Max.    :66.00   Max.    :38.20   Max.    :27.8000   Max.    :37.000
```

Source :-

Kaggle Data – Ottawa Bike Detection

Dataset:

ml_friendly_bike_detection.csv

Link for Reference:

https://www.kaggle.com/samuellara/ottawa-bike-detection

---

2. For one of the discrete variables (say, X) as well as one of the continuous variables (say, Y), carry out the following exercises:

(i). Provide the frequency distribution in a tabular form.

<a> Discrete Variable:

We have following discrete variables namely location name, location id, count, day, day of year and day of week.
We can get frequency distribution by running following command in R.

Code:

FrequencytableX=table(data$count)

where count is name of discrete Variable. Output of each frequency distribution is mentioned below.

```
> FrequencytableX

  1   2   3   4   5   6   7   8   9  10
321 246 166 167 103 109  69  83  65  81
```

<b> Continuous Variable:

We have MaxTemp, MeanTemp, MinTemp, SnowonGrndcm, Totalprecipmm, TotalRainmm, TotalSnowcm as continuous variables in our dataset. Frequency table for continuous variable can be generated in R using following commands.

Code:

histogramy=hist(data$MaxTemp)

FrequencytableY=data.frame("Class Intervals" = as.character(paste(histogramy$breaks[1:9],histogramy$breaks[2:10], sep= "-")),"Frequency"= histogramy$counts)

Here, we have constructed the frequency table for Maximum Temperature.

```
> FrequencytableY
  Class.Intervals Frequency
1         -25--20        14
2         -20--15        64
3         -15--10       141
4          -10--5       355
5            -5-0       335
6             0-5       351
7            5-10       132
8           10-15        16
9           15-20         2
```

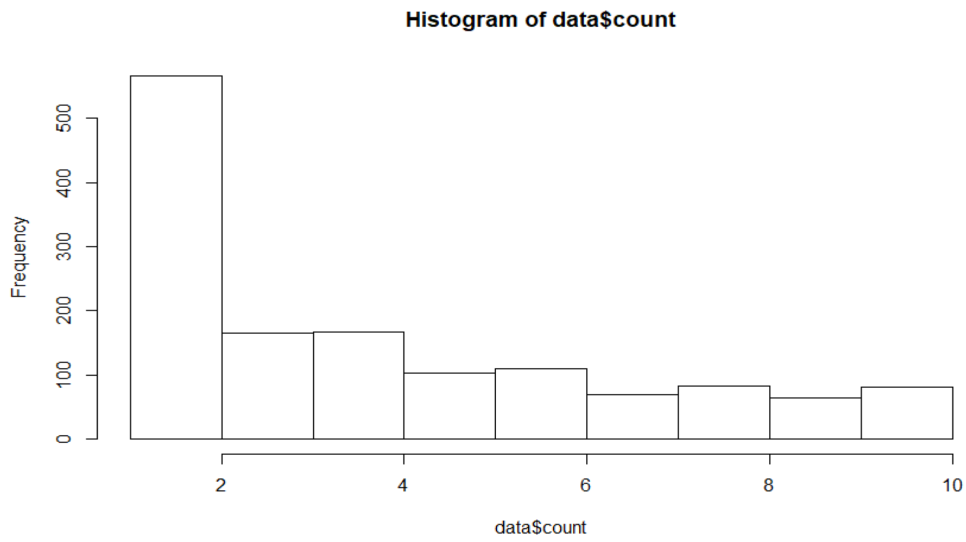
(*ii*). ***Plot the histogram, clearly mentioning what (in-built) method you have used for determining the bin-width.***

<a> For Discrete Variable:

For Discrete Variable histogram is same as bar graph where x axis denotes category and y axis will denote frequency.

Code:

histogramx = hist(data$count)

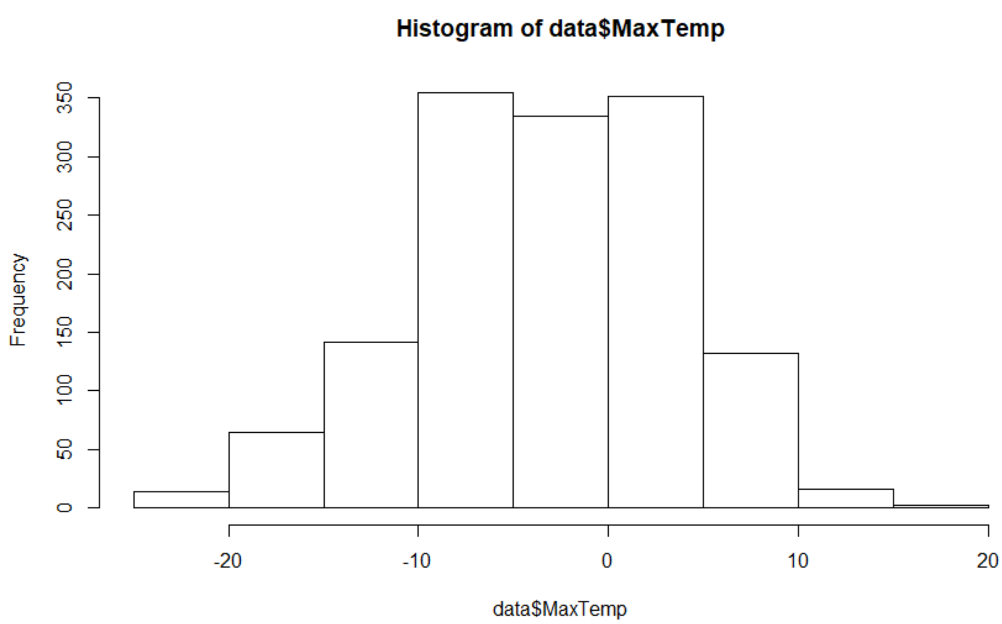


<b> For Continuous Variable:

We can create histogram using hist function.

Code:

```
histogramy=hist(data$MaxTemp)
```



(iii)Compute appropriate measures of central tendency, dispersion, skewness and kurtosis.

Central Tendency – Mean, Median, Mode
Dispersion – Standard Deviation
Skewness and kurtosis

<a> For Discrete Variable:

All the measures mentioned above can be calculated by using the following R code:

Code:

library(psych)
describe(data$count)

```
> describe(data$count)
   vars    n mean  sd median trimmed  mad min max range skew kurtosis   se
X1    1 1410 4.04 2.8      3    3.72 2.97   1  10     9  0.7    -0.69 0.07
```

<b> For Continuous Variable:

All the measures mentioned above can be calculated by using the following R code:

Code:

library(psych)
describe(data$MaxTemp)

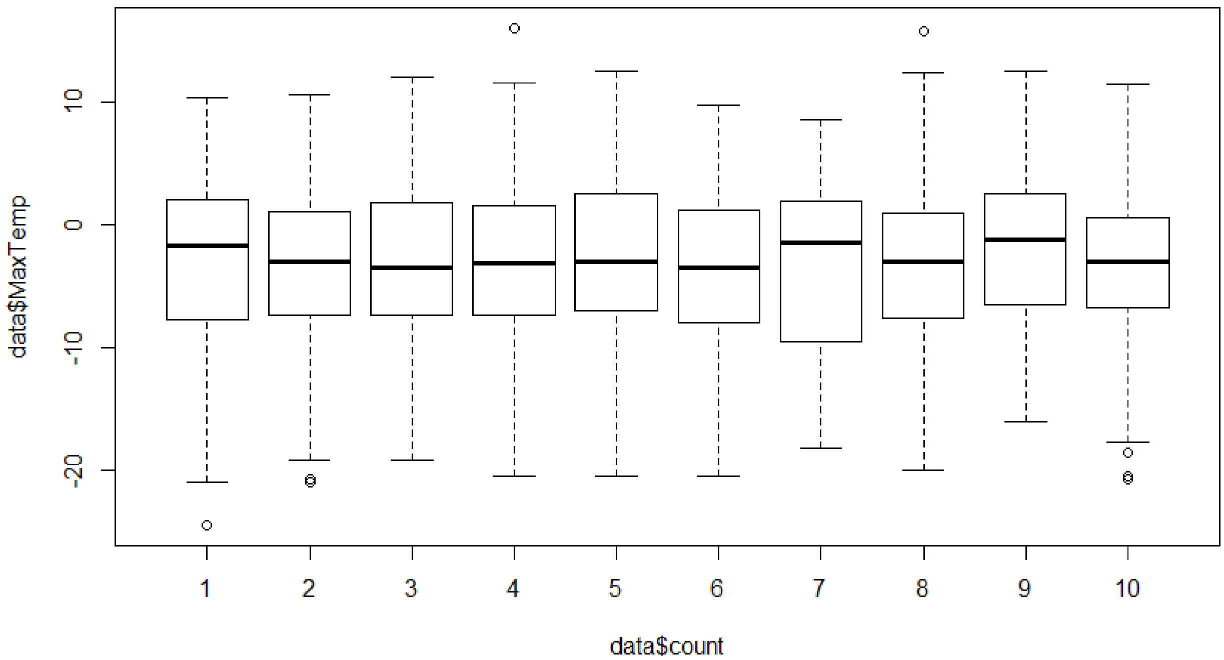
```
> describe(data$MaxTemp)
   vars    n  mean  sd median trimmed  mad   min max range  skew kurtosis   se
X1    1 1410 -3.13 6.7     -3   -2.91 6.67 -24.5  16  40.5 -0.25    -0.31 0.18
```

(iv). Generate the box-and-whisker plots for data on the two variables in a single plot.

Box plots are drawn for Maximum Temperature for different count number. R code is given as follows.

Code:

boxplot(data$MaxTemp ~ data$count , data = data)



(v). Comment on the properties of the distribution of the two variables on the basis of your observations from the outputs in (ii)-(iv).

<Discrete Variable>

Discrete variable count i.e. count of vehicles observed before the 1st bike is observed in a given time interval is likely to follow geometric distribution with parameter ‘p’ which is the probability of observing a bike. The nature of histogram confirms the same.

<Continuous Variable>

Temperature being a natural phenomenon is likely to follow a normal distribution with some mean and standard deviation, which is evident from the histogram as well.

3. **Based on your observations regarding the distribution of X and Y, fit**

<a> Discrete Variable

(i). an appropriate probability distribution to the data X:

Variable – Count
Distribution – Geometric Distribution

(a) justification for your choice of the probability distribution(s);

Discrete variable count i.e count of vehicles observed before the 1st bike is observed in a given time interval is likely to follow geometric distribution with parameter 'p' which is the probability of observing a bike. The nature of histogram confirms the same.
Assuming that the probability of observing bike remains constant over time and the occurrence of vehicles or bikes are independent of one another we can safely fit geometric distribution.

(b) a table with the observed and expected frequencies in two columns;

```
> FrequencytableX

  1   2   3   4   5   6   7   8   9  10
321 246 166 167 103 109  69  83  65  81
> expectedfreq
 [1] 224.40565 179.84059 144.12577 115.50361  92.56557  74.18283  59.45074  47.64433  38.18257  30.59983
```

Count	Observed	Calculated
1	321	224
2	246	180
3	166	144
4	167	116
5	103	93
6	109	74
7	69	59
8	83	48
9	65	38
10	81	31

<b>Continuous Variable

(i). an appropriate probability distribution to the data YY;

Variable – Maximum Temperature
Potential Distribution – Normal, Logistic Distribution.

(a) justification for your choice of the probability distribution(s);

Temperature being a natural phenomenon is likely to follow a normal distribution with some mean and standard deviation, which is evident from the histogram as well. Since the shape of logistic distribution is similar to that of normal distribution, we may try to fit both and see which gives a better result.

(b) a table with the observed and expected frequencies in two columns

Observed frequency table is calculated as shown in question 2.

Observed Frequency Table:

```
> FrequencytableY
  Class.Intervals Frequency
1         -25--20        14
2         -20--15        64
3         -15--10       141
4          -10--5       355
5            -5-0       335
6             0-5       351
7            5-10       132
8           10-15        16
9           15-20         2
```

Expected Frequency Table:

Calculated Frequency Distribution assuming Normal and Logistic Distribution

Code:

Normal Distribution:

```
> num_of_samples = 1410
> y <- rnorm(num_of_samples, mean =   normy$estimate[1], sd=  normy$estimate[2]  )
> breaks = seq(-20,20,by=5)
> temp.cut = cut(y,breaks,right = F)
> temp.freq = table(temp.cut)
> cbind(temp.freq)
          temp.freq
[-20,-15)        37
[-15,-10)       157
[-10,-5)        345
[-5,0)          425
[0,5)           275
[5,10)          124
[10,15)          24
[15,20)           7
```

Logistic Distribution:

```
> num_of_samples = 1410
> z = rlogis(num_of_samples, location =logisy$estimate[1], scale =  logisy$estimate[2])
> breaks = seq(-20,20,by=5)
> temp.cut = cut(z,breaks,right = F)
> temp.freq = table(temp.cut)
> cbind(temp.freq)
          temp.freq
[-20,-15)        52
[-15,-10)       127
[-10,-5)        312
[-5,0)          465
[0,5)           271
[5,10)          108
[10,15)          41
[15,20)           6
```

On comparing the expected frequencies of both the Normal as well as the Logistic distribution with that of the observed frequencies, one could probably guess that Normal fits with the data better.

(c) the outcome of a goodness-of-fit test;

<a> Discrete Variable

Chi square test is used here to check the goodness of fit.

Code:

```
> FrequencytableX=table( data$count )
> xx=FrequencytableX
> E_geom = dgeom( 1:10,prob  =  pfitX$estimate )
> expectedfreq = (sum(xx) *E_geom)
> chisquarev = sum((xx-expectedfreq)^2/(expectedfreq))
> pchisq ( chisquarev , df=8 )
[1] 1
```

The p value for chi square test is coming to be pretty high and thus we can safely say that the geometric distribution fits the data well.

<b> Continuous Variable

Kolmogorov-Smirnov test used to check goodness of fit.

Normal Distribution:

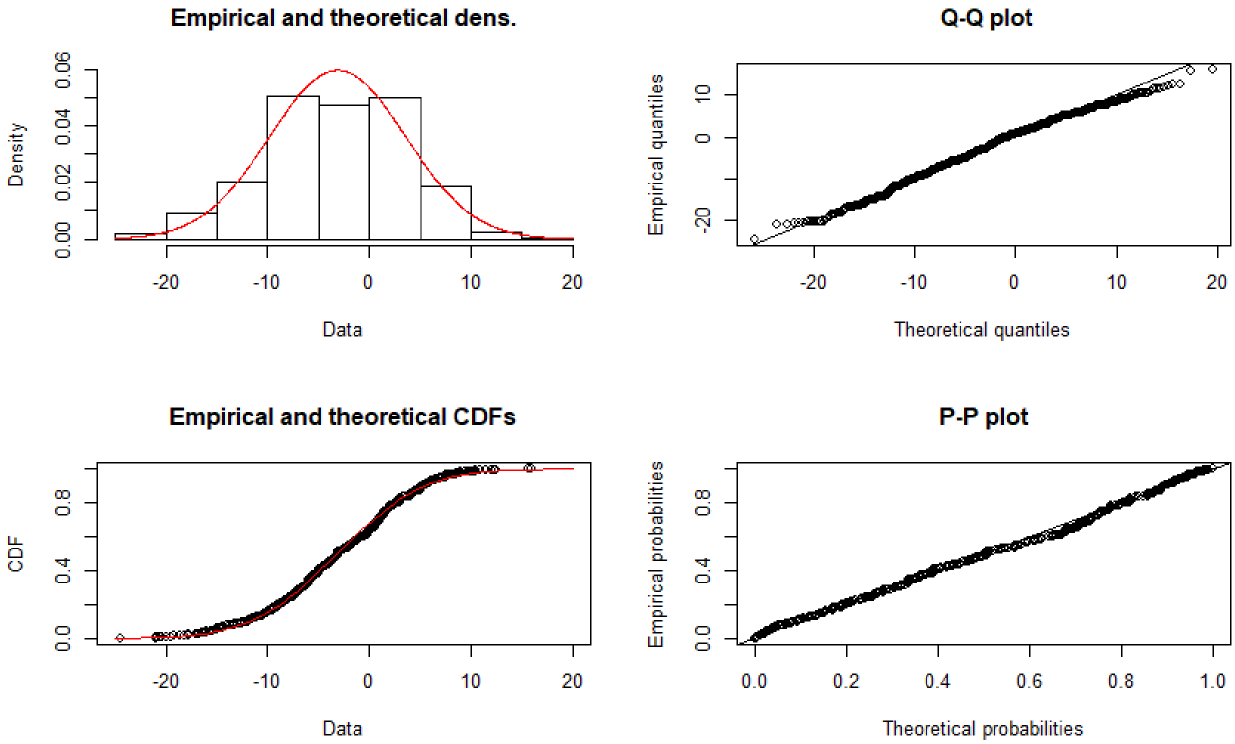
Code:

```
num_of_samples = 1410
y <- rnorm(num_of_samples, mean =   normy$estimate[1], sd=  normy$estimate[2]  )
result = ks.test(data$MaxTemp, y)
```

```
> result

        Two-sample Kolmogorov-Smirnov test

data:  data$MaxTemp and y
D = 0.044681, p-value = 0.1198
alternative hypothesis: two-sided
```



Logistic Distribution:

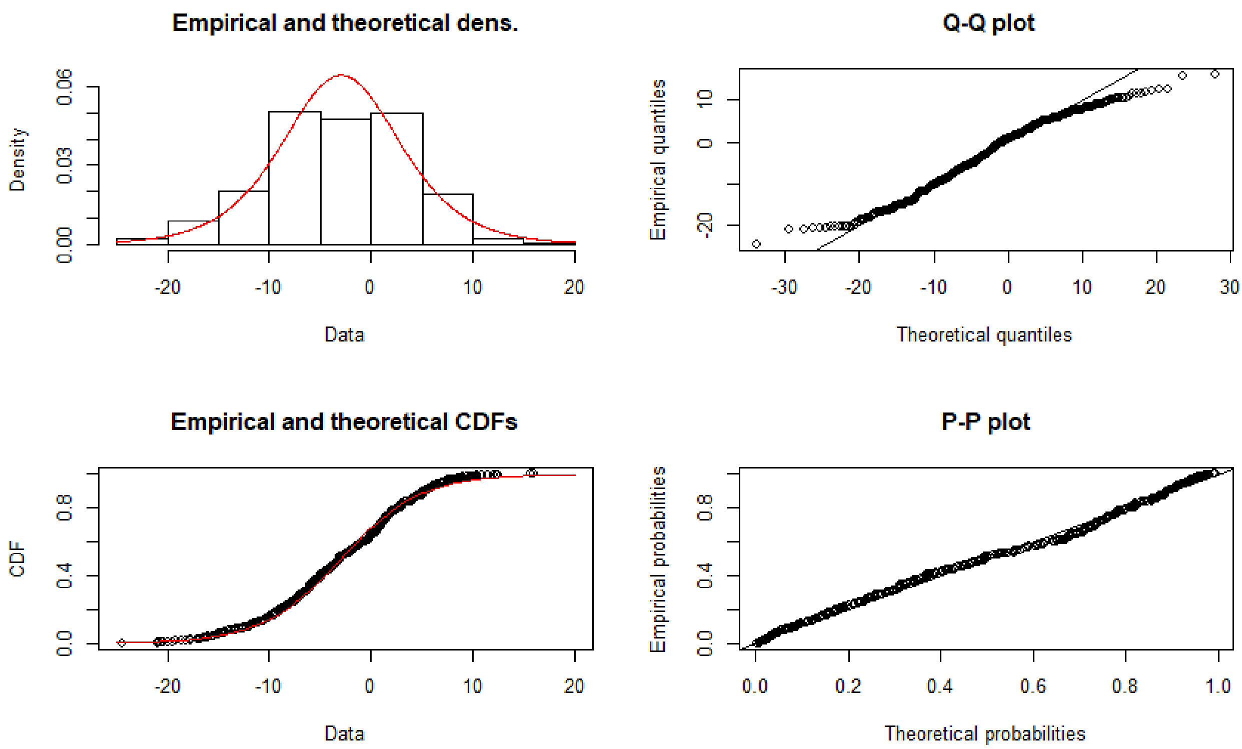
Code:

```
num_of_samples = 1410
y <- rnorm(num_of_samples, mean =   normy$estimate[1], sd=  normy$estimate[2]  )
result = ks.test(data$MaxTemp, y)
```

```
> result

        Two-sample Kolmogorov-Smirnov test

data:  data$MaxTemp and z
D = 0.065248, p-value = 0.004943
alternative hypothesis: two-sided
```



(d) on the basis of (c), explain which of the two probability distributions fitted to the data on Y is more appropriate.

Here – Y is MaxTemp
Distribution under consideration – Normal, Logistic

Distribution	p- value (ks test)
Normal Distribution	0.1198
Logistic Distribution	0.004943

From output of above ks test it is evident that normal distribution is better fit than logistic distribution to Maximum Temperature data. The p value of ks test for logistic distribution is too small and hence we fail to claim that it follows the given distribution.

4. Select any two continuous variables from your dataset, say X_1 and X_2, and carry out the following exercises:

   (i). Fit a regression line of X_1on X_2 by the method of least squares and plot it on the scatterplot for the data.
   (ii). Perform an appropriate F-test to assess the validity of the linear regression model, providing the associated ANOVA table.
   (iii). Determine the coefficient of determination (R^2) for the fitted model.
   (iv). Obtain the residual and studentized residual plots for the problem.
   (v). Compute Cooke's distances for observations that appear to be unusual, on the basis of their studentized residual values and hence identify influential points, if any.

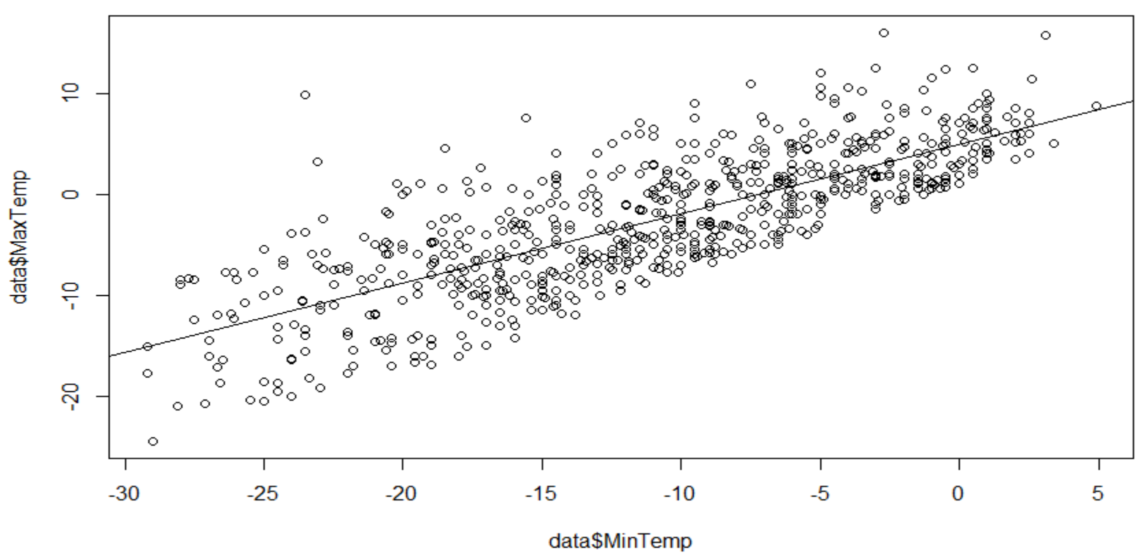
   For each of the five exercises (i)-(v), provide clear and concise comments regarding the observed outcomes.

(i). Fit a regression line of X_1on X_2 by the method of least squares and plot it on the scatterplot for the data.

Here,
X1 – MaxTemp i.e. Maximum Temperature
X2 – MinTemp i.e. Minimum Temperature

Code:

```
yonx=lm(data$MaxTemp ~ data$MinTemp)
plot(data$MinTemp, data$MaxTemp)
abline(lm(data$MaxTemp ~ data$MinTemp
```



The scatter plot of the two variable does suggest that there is a strong linear association between the two variables.

(ii). ***Perform an appropriate F-test to assess the validity of the linear regression model, providing the associated ANOVA table.***

The ANOVA table gives the summary of the F-test and it can be performed by the following R code:

Code:

```
> anova(yonx)
Analysis of Variance Table

Response: data$MaxTemp
               Df Sum Sq Mean Sq F value    Pr(>F)
data$MinTemp    1  38883   38883  2251.6 < 2.2e-16 ***
Residuals    1408  24315      17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistic is pretty large and hence the associated p value is very small. So we fail to accept the null hypothesis that there is no linear association between the two concerned variables, i.e. there exists a linear relationship between the two variables.

(iii). Determine the coefficient of determination (R^2) for the fitted model.

A large value of coefficient of determination suggests that there is a linear association between the two variables. It can be performed via the following R code:

Code:

```
> summary(yonx)$r.squared
[1] 0.6152579
```

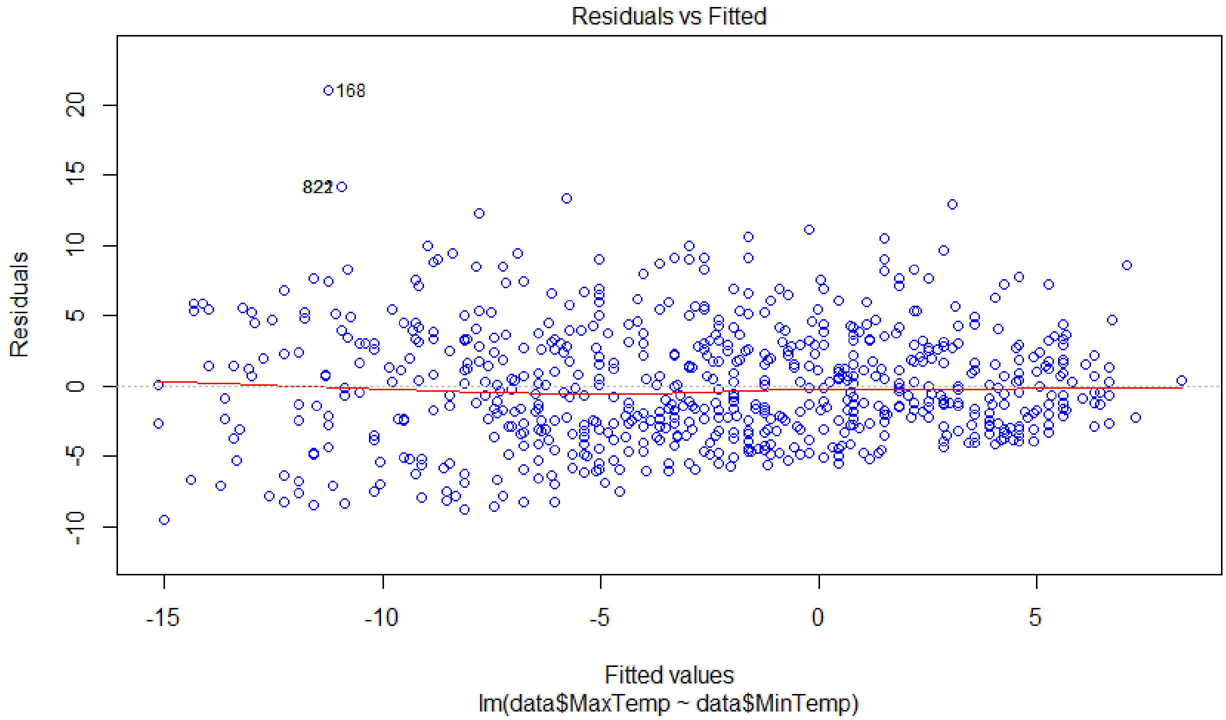
The R squared value is large as well pointing to the same direction that the variables have linear relationship among themselves.

(iv). Obtain the residual and studentized residual plots for the problem.

Residuals Plot:

Code:

plot(yonx, which=1, col=c("blue"))



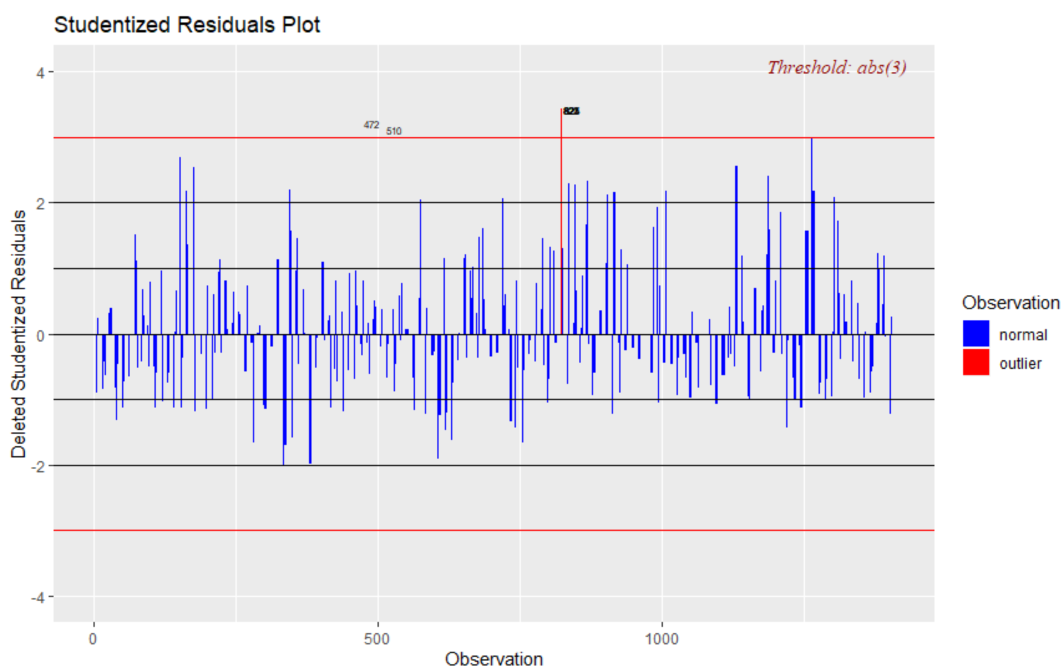
The residual plot is homoscedastic and residuals are evenly spread out around 0, which shows that the fitted line is a good fit.

Studentized Residuals Plot:

Code:

library(olsrr)

ols_plot_resid_stud(yonx)



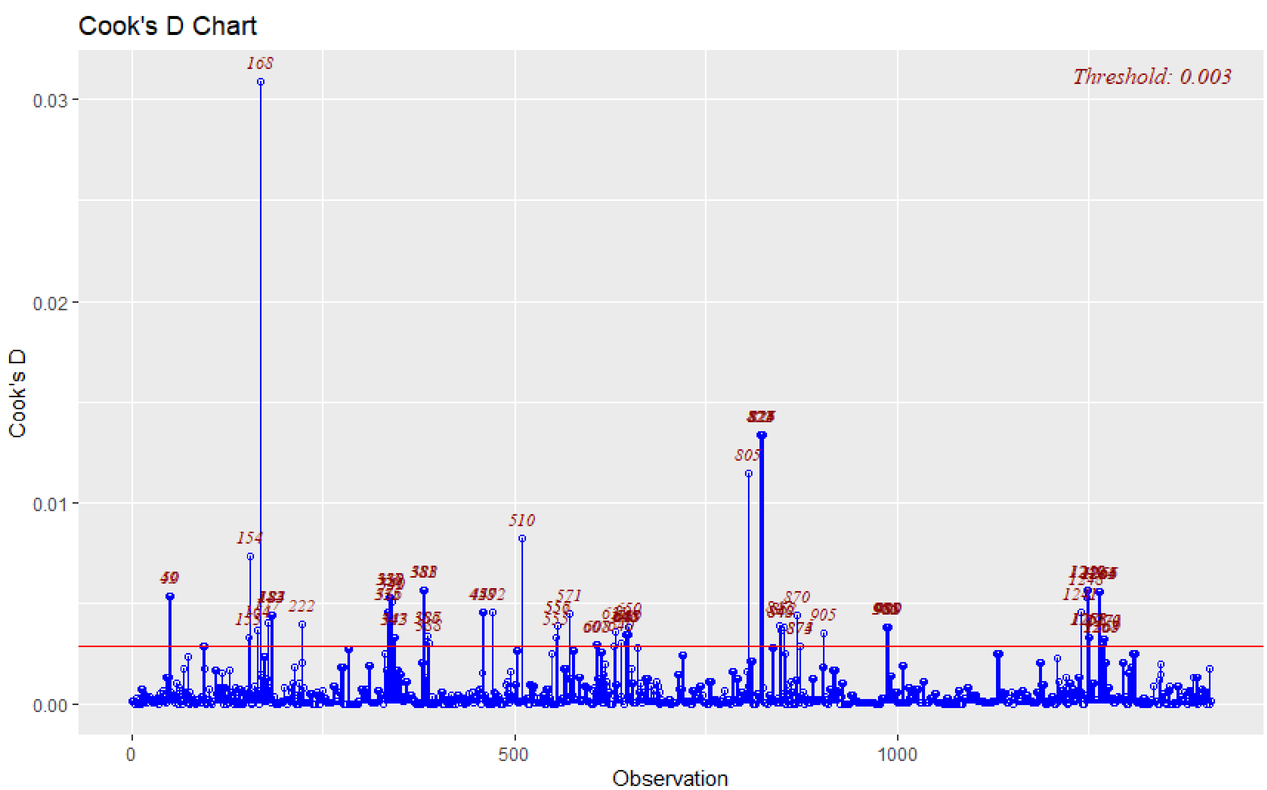
The Studentized Residuals Plot shows that there is one outlier which is just above the threshold value. We might get a better fit excluding this point.

(v). Compute Cooke's distances for observations that appear to be unusual, on the basis of their studentized residual values and hence identify influential points, if any.

We have one point as an outlier observed from the Studentized Residuals plot. Cook's Distance can be found out using the following R code:

Code:

```
library(olsrr)
ols_plot_cooksd_chart(yonx)
```



Though here we see there are many points above the threshold value, but the threshold value is set at a very low value of 0.003. The outlier observed from Studentized Residuals Plot have a Cook's Distance less than 0.015 and hence we can't regard it as an influential point.