Internal Assessment

Module 2: Data Diagnostics and Predictive Module

Name 1	Souvik Samanta
Roll Number 1	19PGDM064
Name 2	Ritu Mittal
Roll Number 2	19PGDM052

==========

Question 1:
Final output is –

```
summary(data)
    ï..Year     Production.of.wheat..MT. Amount.of.rainfall Qulaity.of.Soil  Quality.of.ferti
lizer
 Min.   :1960   Min.   :  9854           Min.   :300.0      Min.   : 1.000   Min.   : 1.000
 1st Qu.:1975   1st Qu.: 25991           1st Qu.:421.0      1st Qu.: 3.000   1st Qu.: 2.000
 Median :1990   Median : 53410           Median :495.0      Median : 6.000   Median : 5.500
 Mean   :1990   Mean   : 51472           Mean   :490.6      Mean   : 5.345   Mean   : 5.224
 3rd Qu.:2004   3rd Qu.: 72309           3rd Qu.:578.0      3rd Qu.: 8.000   3rd Qu.: 8.000
 Max.   :2019   Max.   :102190           Max.   :647.0      Max.   :10.000   Max.   :10.000
                                         NA's   :3          NA's   :2        NA's   :2
      Ind
 Min.   :0.00
 1st Qu.:1.00
 Median :1.00
 Mean   :0.95
 3rd Qu.:1.00
 Max.   :1.00
```

So the missing values have been treated for production of wheat

After kNN the summary is –

```
summary(data)
    ï..Year     Production.of.wheat..MT. Amount.of.rainfall Qulaity.of.Soil  Quality.of.fertilizer
 Min.   :1960   Min.   :  9854           Min.   :300.0      Min.   : 1.000   Min.   : 1.000
 1st Qu.:1975   1st Qu.: 25991           1st Qu.:417.5      1st Qu.: 3.000   1st Qu.: 2.000
 Median :1990   Median : 53410           Median :492.5      Median : 6.000   Median : 5.000
 Mean   :1990   Mean   : 51472           Mean   :487.9      Mean   : 5.317   Mean   : 5.217
 3rd Qu.:2004   3rd Qu.: 72309           3rd Qu.:575.8      3rd Qu.: 8.000   3rd Qu.: 8.000
 Max.   :2019   Max.   :102190           Max.   :647.0      Max.   :10.000   Max.   :10.000

>
```

All missing values have been treated.

Question 2:

We have run separate simple linear model for all other variables to calculate R-squared value. And thus the best related parameter came as : Year

After building the multiple linear model, the equation came as –

Production of wheat in MT = -3116000 + 1593year - 1.743Amount of rainfall-322.9Qulaity of Soil -37.23Quality of fertilizer

And, For Year = 2020, Amount of rainfall=585, Qulaity of Soil=6.5, Quality of fertilizer=7 ;
Production of wheet in MT = -3116000 + 1593*2020 - 1.743*585 - 322.9*6.5 - 37.23*7
= 98480.88MT

Assumptions :
The regression has five key assumptions:

Linear relationship
Multivariate normality
No or little multicollinearity
No auto-correlation
Homoscedasticity