

## Crop Production in Asia

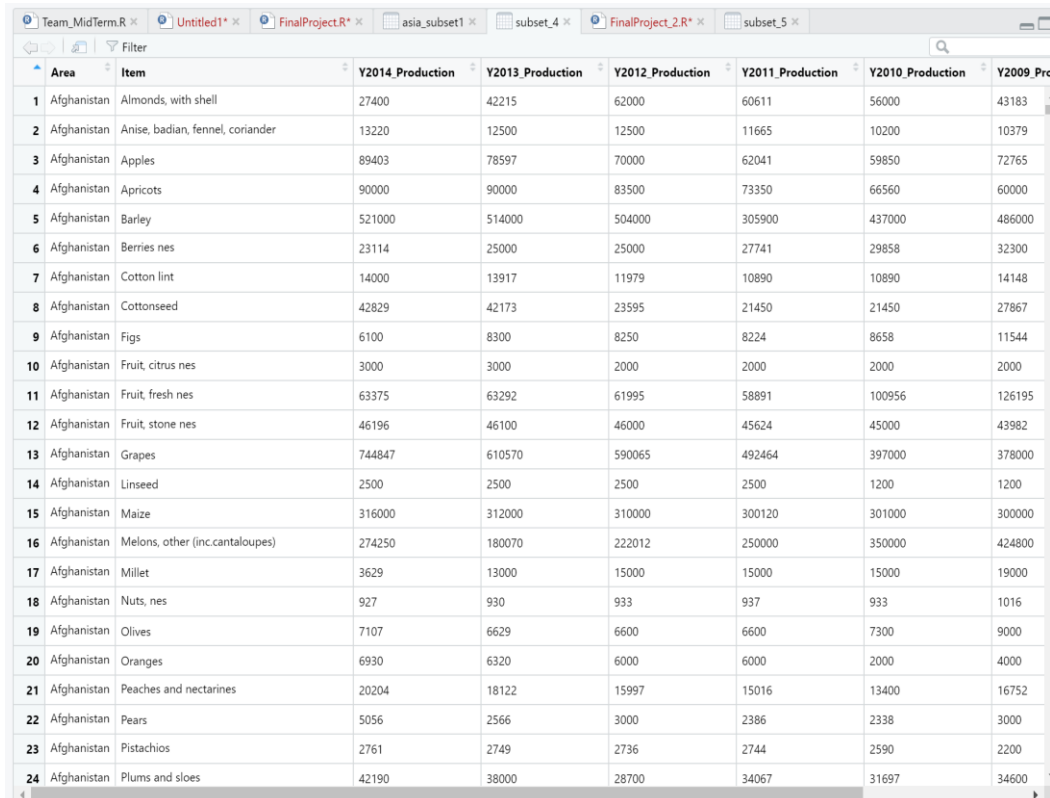
### Introduction

The dataset we have chosen deals with crop production data for Asia from the year 1961 to 2014. this data set is taken from DataWorld which is collected by the Food and Agriculture Organization of the United Nations (FAO). The main objective was to predict crop production in 2014 considering the previous five years of production data for various countries for all the crops in Asia. And also check for a same pattern in other continents like Africa and Europe. We have used R and Tableau for data exploration, visualization and prediction

### Data Set Cleaning

This data set consists of 115 variables with over 10,000 + records. This data included the Production, yield, and area harvested for about **173 different crops** in 51 countries from the year 1961 to 2014 in Asia. Here we have considered the Production as the unit for analysis. So, in order to predict the Production in 2014 for various available crops, we need Crop Production data from 2008-2013 which is the past 5 years data. Also, on deep analysis of the dataset, we could understand that there a lot of NULL values present. And these Null values are because certain countries were not able to produce few crops because of weather conditions, consumption of the place and other varied factors which we cannot assume. Hence, the removal of this data doesn't have an impact on the output.

First, a subset created to access the data only for the years from 2008-2014 for all the countries and respective crops. Later using the **pivot\_wider()** function, the subset is further divided in order to get the production data. This final subset includes the data related to production in tonnes for each crop in each country from 2008-2014. The below screenshot gives a clear view of the subset. This reduces to 9 variables and 3500+ records.



	Area	Item	Y2014_Production	Y2013_Production	Y2012_Production	Y2011_Production	Y2010_Production	Y2009_Pro
1	Afghanistan	Almonds, with shell	27400	42215	62000	60611	56000	43183
2	Afghanistan	Anise, badian, fennel, coriander	13220	12500	12500	11665	10200	10379
3	Afghanistan	Apples	89403	78597	70000	62041	59850	72765
4	Afghanistan	Apricots	90000	90000	83500	73350	66560	60000
5	Afghanistan	Barley	521000	514000	504000	305900	437000	486000
6	Afghanistan	Berries nes	23114	25000	25000	27741	29858	32300
7	Afghanistan	Cotton lint	14000	13917	11979	10890	10890	14148
8	Afghanistan	Cottonseed	42829	42173	23595	21450	21450	27867
9	Afghanistan	Figs	6100	8300	8250	8224	8658	11544
10	Afghanistan	Fruit, citrus nes	3000	3000	2000	2000	2000	2000
11	Afghanistan	Fruit, fresh nes	63375	63292	61995	58891	100956	126195
12	Afghanistan	Fruit, stone nes	46196	46100	46000	45624	45000	43982
13	Afghanistan	Grapes	744847	610570	590065	492464	397000	378000
14	Afghanistan	Linseed	2500	2500	2500	2500	1200	1200
15	Afghanistan	Maize	316000	312000	310000	300120	301000	300000
16	Afghanistan	Melons, other (inc.cantaloupes)	274250	180070	222012	250000	350000	424800
17	Afghanistan	Millet	3629	13000	15000	15000	15000	19000
18	Afghanistan	Nuts, nes	927	930	933	937	933	1016
19	Afghanistan	Olives	7107	6629	6600	6600	7300	9000
20	Afghanistan	Oranges	6930	6320	6000	6000	2000	4000
21	Afghanistan	Peaches and nectarines	20204	18122	15997	15016	13400	16752
22	Afghanistan	Pears	5056	2566	3000	2386	2338	3000
23	Afghanistan	Pistachios	2761	2749	2736	2744	2590	2200
24	Afghanistan	Plums and sloes	42190	38000	28700	34067	31697	34600

Figure 1: Asia dataset

## Research Questions:

1. Predicting the 2014 Crop Production based on Previous years from 2008 to 2013.
2. Comparing the values with the original Crop production for the year 2014.
3. To predict and understand the maximum crop which will be harvested in 2014.

## Data Visualizations

The below tableau visualization represents the Area and Production from 2008-2014. We can depict that China has the highest production and India has the second highest production. Both of these countries has consistency in its production rate over the years.

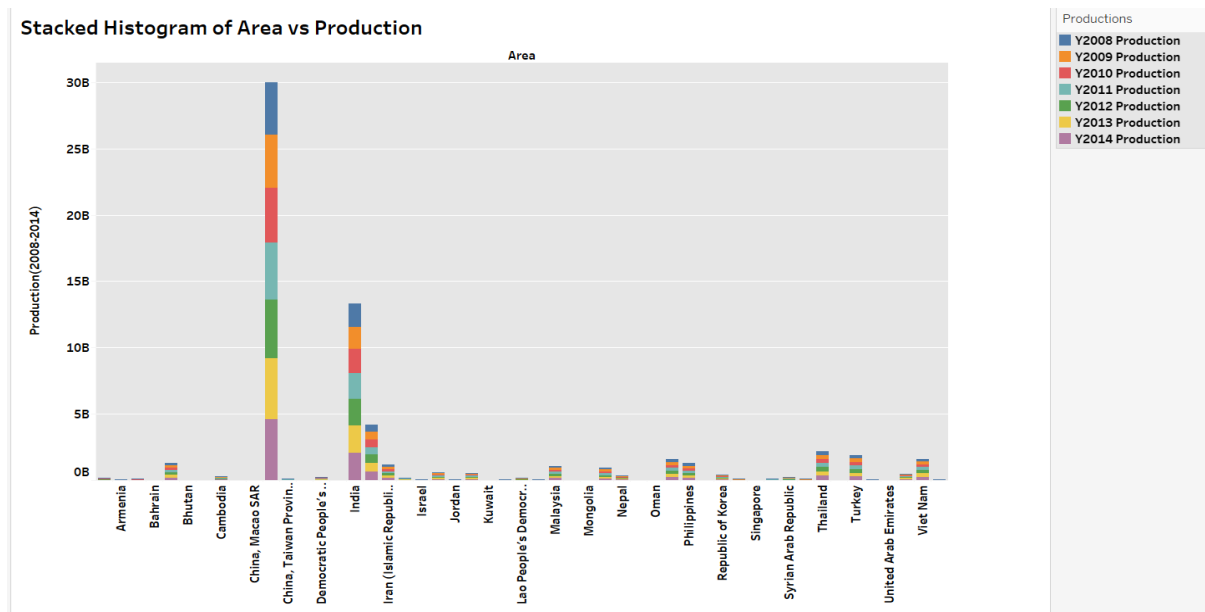


Figure 2: Visualisations of Production (2014) and Area

The below tableau graph is plotted for the year 2014's Area harvested, Production, Yield in the form of a grid with respect to a few items from the dataset. We can observe that cereals have a high harvesting area and highest production value. While tomatoes have high yield among all other items. Figure 4 represents the scatter plot which clearly shows that all the production values from 2008-2014 are related linearly to each other.

### Barplot of Items vs years of Harvest, Production and Yield

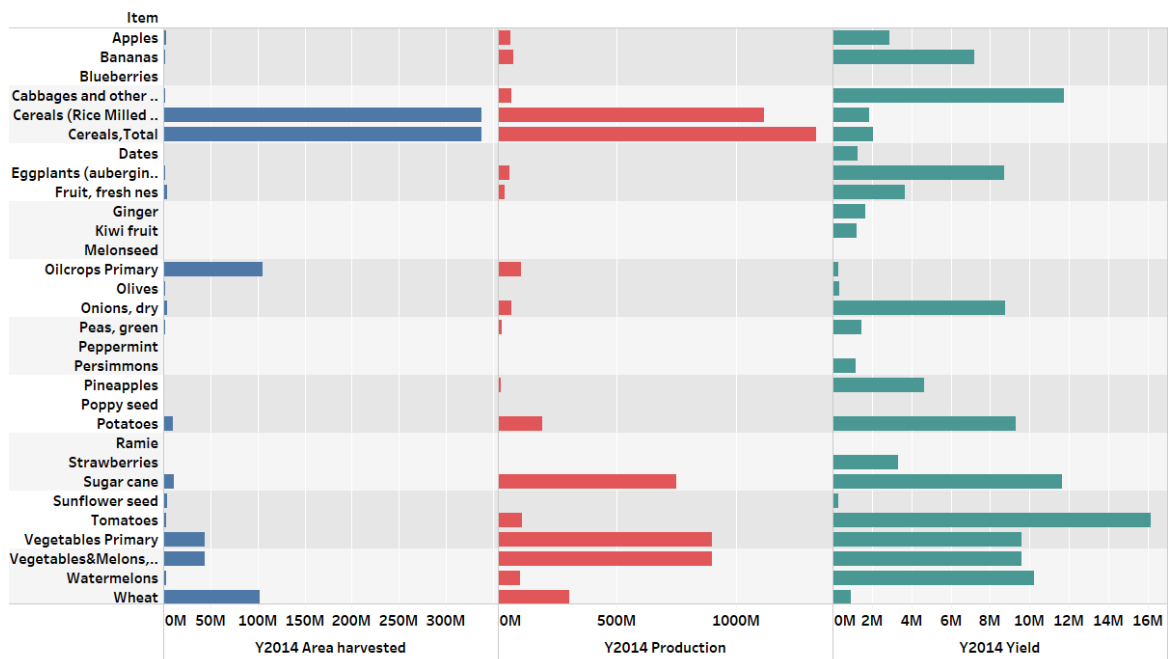


Figure 3: Visualizations of Items vs Years of Harvest, Production, and Yield

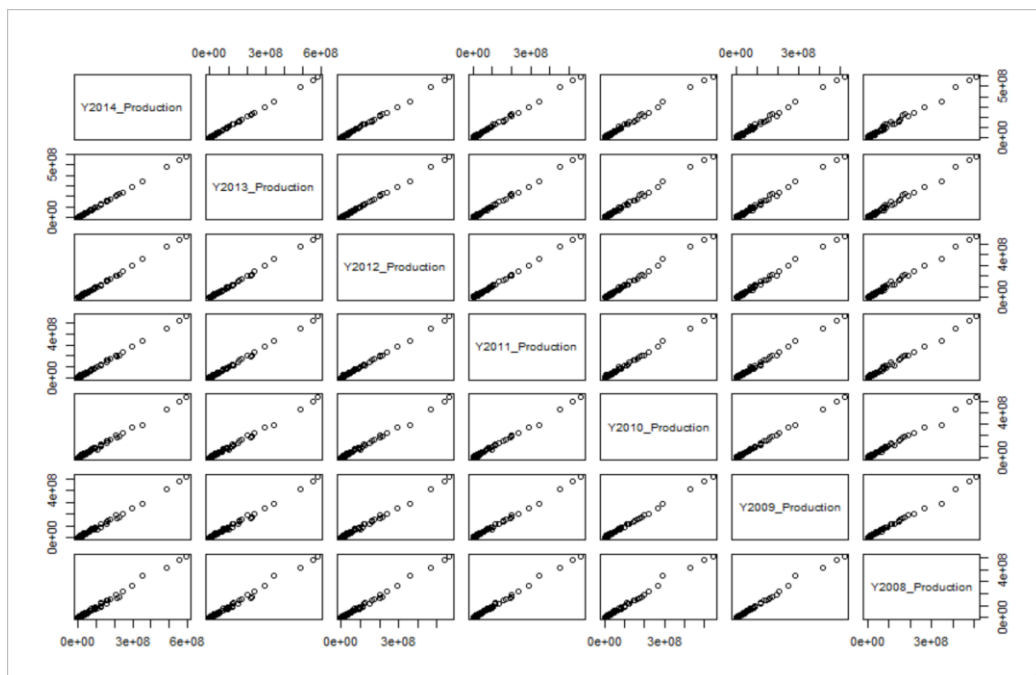


Figure 4: Scatter plot

### Data Predictions

We have used Linear regression and Regression Tree methods to predict the year 2014 production for the three continents. We have chosen these particular as the values in our data set are continuous.

We have used clustering method to depict the patterns among the production data with to area.

## Linear Regression

For the above correlation plot between the productions, we could understand that the Linear Regression is the best model to predict the 2014 Production. Here we have done both Single and Multiple Linear Regression over the production data. Here in this model, we have trained the model using 80% of the data and later we tested the model over the 20% data. Initially, we only considered the 2013 Production data to predict the 2014 production. Later, as a part of multiple linear regression, we included the production values from 2008-2013. We could see that the adjusted R square is 0.99 and the P-value is less than 0.05 which means we reject the Null Hypothesis Testing implies that the variables are dependent on each other. From the below screenshot, we could see that all the Production values of the years 2013,2012 and 2008 are significant variables for predicting the 2014 values.

```
Call:
lm(formula = Y2014_Production ~ Y2013_Production + Y2012_Production +
    Y2011_Production + Y2010_Production + Y2009_Production +
    Y2008_Production, data = trainingData)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-5330184  -12989   -11762   -7268  4919488
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.177e+04  7.593e+03   1.550   0.1213
Y2013_Production  7.454e-01  1.433e-02  52.027  <2e-16 ***
Y2012_Production  3.569e-01  1.836e-02  19.444  <2e-16 ***
Y2011_Production -8.447e-03  1.374e-02  -0.615   0.5387
Y2010_Production -2.580e-02  1.395e-02  -1.850   0.0645 .
Y2009_Production  1.786e-02  1.320e-02   1.354   0.1760
Y2008_Production -8.111e-02  9.534e-03  -8.507  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 380200 on 2573 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9996
F-statistic: 1.016e+06 on 6 and 2573 DF,  p-value: < 2.2e-16
```

Figure 5

## Regression Plots:

The residual vs the fitted values plot gives us information about the outliers and non-linearity between the parameters. Clearly, the residual values are close to zero, which means that the chosen variables are appropriate for the model and our prediction value is proper. Also, the accuracy of the model is 99%.

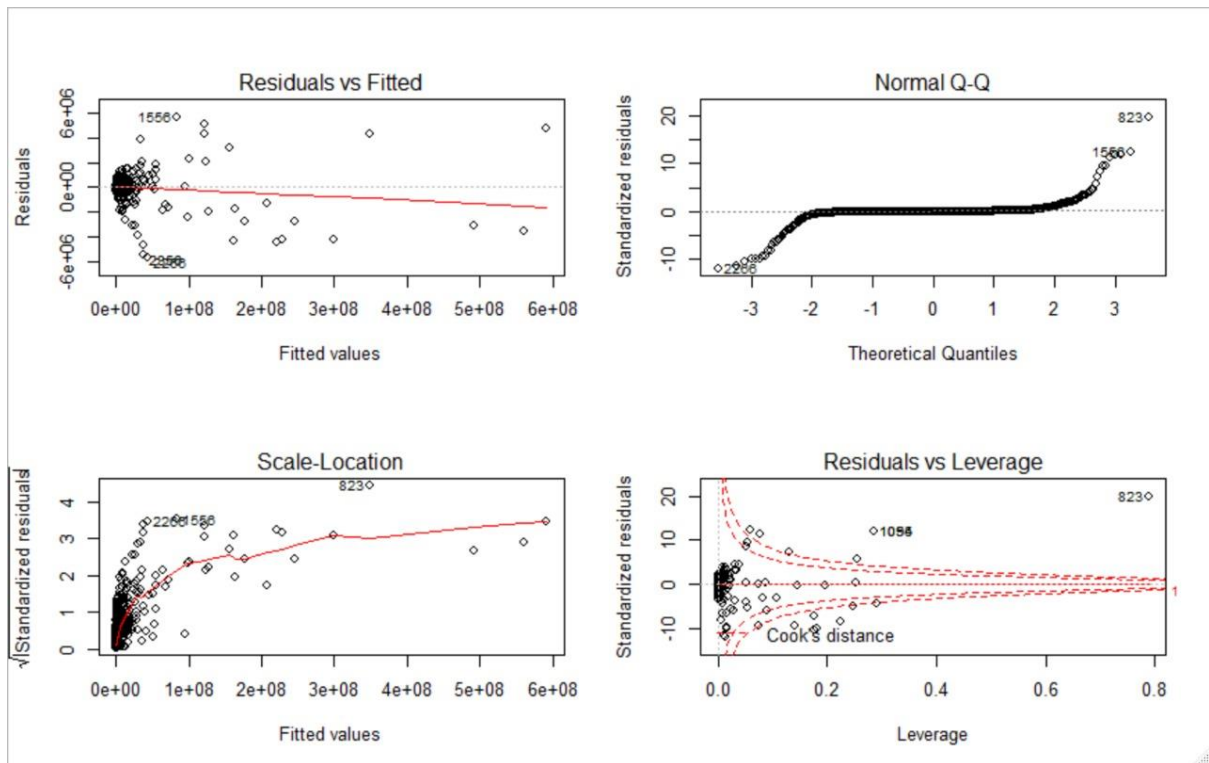


Figure 6

From the scatter plot, we see that the model line is linear which shows that there exists a finite relation between the dependent and predictor variables.

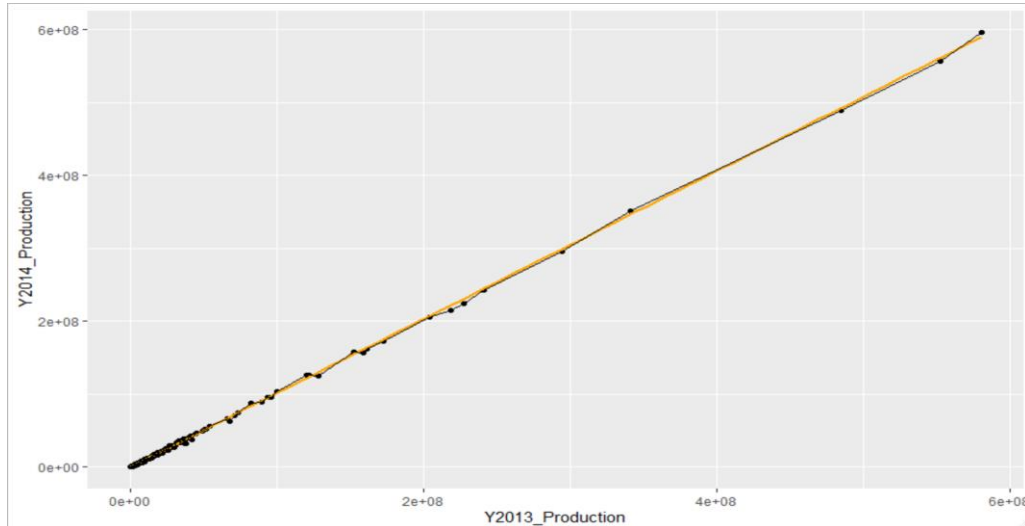


Figure 7: Plot Between Production(2014) and Production(2013)

## Regression Tree

We wanted to see the best variable, which can predict the values of the year 2014. So, we have tried two different models. Single dependent variable and multi dependent variable. Initially, we have tried the regression tree with a single dependent variable by taking the variable of production values of the year 2013 to predict 2014 production values. We repeated the same with the year 2012 to predict the production values of 2014.

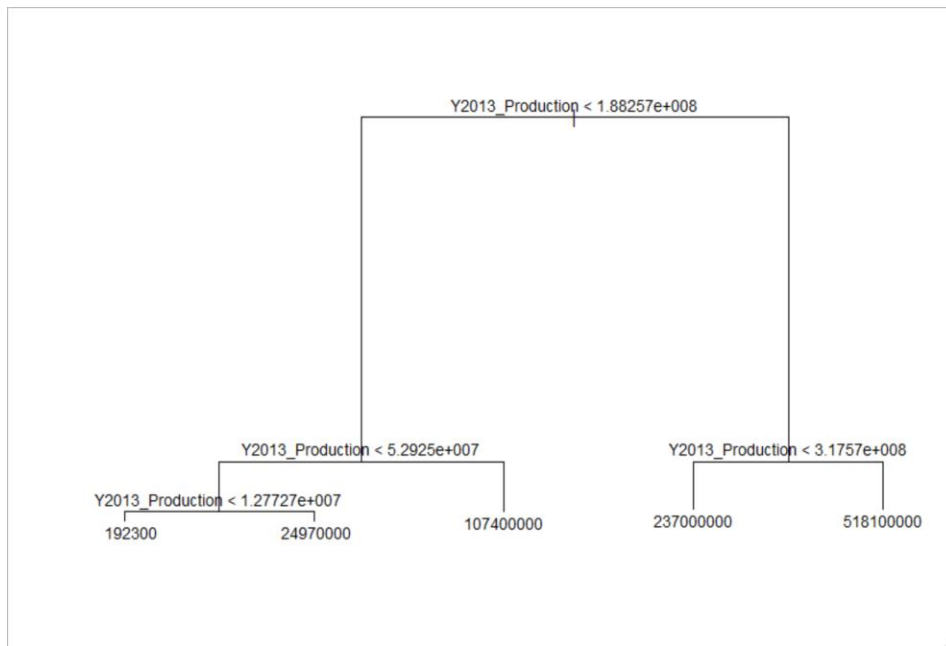


Figure 8: Regression tree (2013 production values are the best to predict production values of 2014)

Then we started considering multi dependent variables, the production values of the years 2008 to 2013. To check the best production value to predict production values of 2014.

#### #Regression Trees

#We install the Tree package for the Regression Tree

```
tree_item=tree(Y2014_Production~Y2012_Production+Y2011_Production+Y2010_Production)
summary(tree_item)
plot(tree_item)
text(tree_item,pretty=0,cex=0.9)
tree_item
```

#Cross validation plot for REgression tree

```
cv_treeitem<-cv.tree(tree_item)
plot(cv_treeitem)
```

```
plot(cv_treeitem$size,cv_treeitem$dev,type="b")
plot(cv_treeitem$k,cv_treeitem$dev,type="b")
```

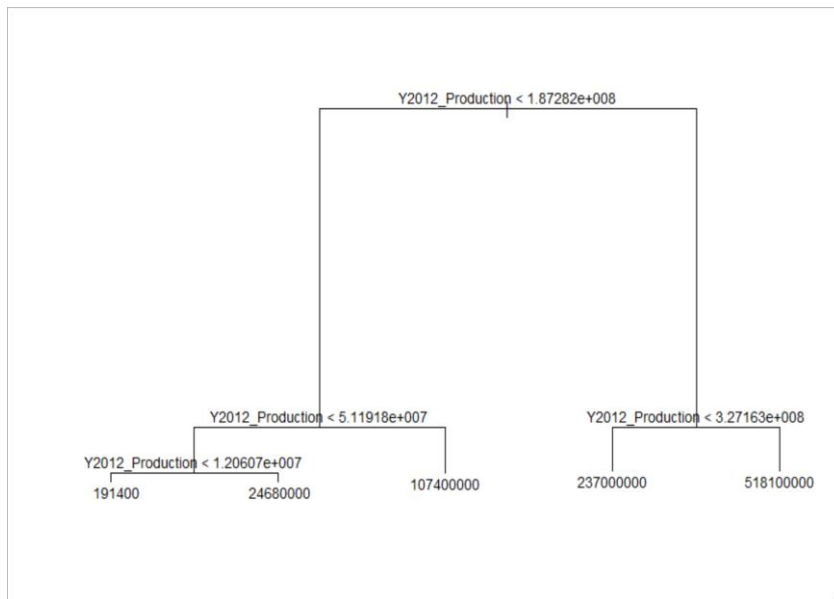


Figure 9: Regression tree (2012 production values are the best to predict production values of 2014 when 2013 is not considered)

Then we have taken another set of variables just to observe if we can find any pattern. This time we have taken the production values of the years 2008 to 2012 to the best values to predict the production values of the year 2014. We have observed a pattern, the best values to predict the production values of upcoming years are the preceding year's values (For example, the best values to predict 2014 values are the 2013 values).

## K Means Clustering

We have used two different subsets for the clustering method. The first subset includes the production values of the years 2008 - 2014. A size of 3 clusters was given and the resultant Kmeans values were plotted.

```
#Now creating another subset to include Area Code (Region) and item code (\
subset_5<-pivot_wider(data=asia_subset1,names_from = Element,values_from =
subset_5[is.na(subset_5)] <- 0
subset_5<-pivot_wider(data=asia_subset1,names_from = Element,values_from =
subset_5=na.omit(subset_5)

cropTrain1 <- select(subset_5, -Area, -Item)
cropTrain1
ggscatmat(cropTrain1)

RNGkind(sample.kind="Rounding") # To obtain same results as RFE
set.seed(200)
cropTrain1 = scale(cropTrain1)
head(cropTrain1)
crop1 <- kmeans(x=cropTrain1, centers=3)
crop1

plot.kmeans(crop1, data=cropTrain1)

fviz_cluster(crop1, data = cropTrain1)
```

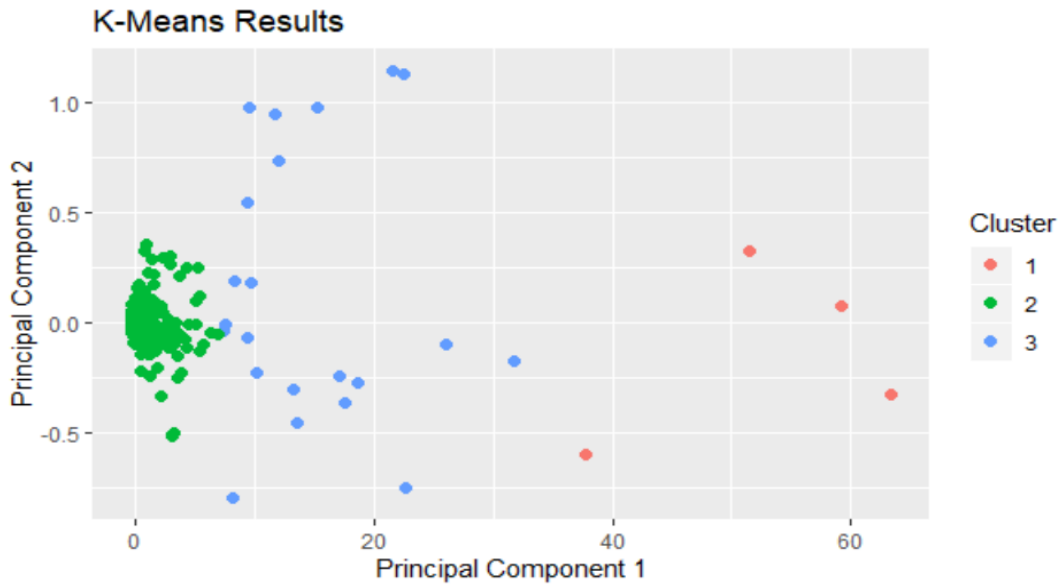


Figure 10: Kmeans plot with just the production values

For better visualizations, fviz\_cluster method was used, and the resultant plot was this:

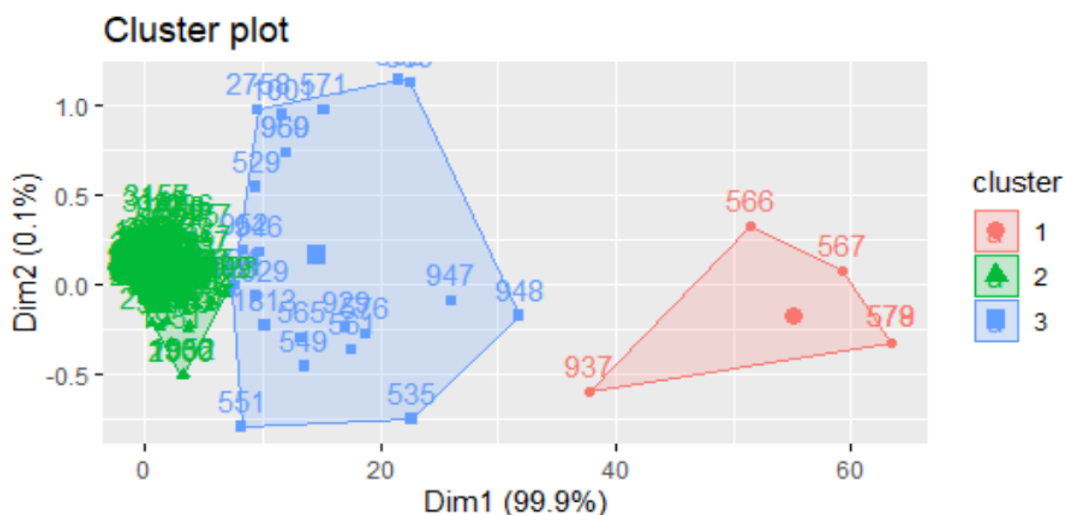


Figure 11: Kmeans plot with fviz\_cluster method

In the second subset, we have included the Area code(various regions were assigned different unique values) and Item code (various items were given different unique values). Even in this subset, we have given the cluster size as 3 and plotted two different graphs. Similarly, we have followed all the above procedures for two other continents (Africa & Europe).

### Comparison among different Continents

The same prediction of 2014 production data is performed for the continents **Africa** and **Europe** to see and observe if there is a specific pattern. We could understand from the results that the Linear regression model is the best for prediction which gives an accuracy of 99% for other continents data as well. Also, 2013 is the most significant year for predicting the production values in 2014 for respective crops.



## **Challenges**

The data is humongous. There were a lot of null values in the dataset. The reason for that is, not all countries produce all the different crops. Most of the countries are suitable for certain crops and they produce those crops. There are a lot of terminologies that we had trouble understanding. We had to spend a lot of time understanding the metrics and we also had to do some manual calculations.

## **Conclusion**

We have different methods to predict the production values of the various years but, mainly for the year 2014. The best model according to our analysis is the Linear Regression model where we got a 99% accuracy. We conclude that the Linear Regression model is the best method for the production values for the year 2014. This method is best for this dataset. Various R packages are used for data predictions and used Tableau for data visualizations.

## **Contributions**

We have used 3 different analysis methods to answer the research questions. Cleaning of the data sets is done together. Linear regression is done by Chandana. Clustering is done by Anirudh. Regression Tree it is done by Ramya. Since we are 3 members on the team we have the analysis on 3 different continents. Analysis of Asia is done by Chandana. Analysis of Europe is done by Anirudh. Analysis of Africa is done by Ramya. The documentation is done together.

## **References**

- [1] RStudio, Inc. ,RStudio Desktop [Computer software], 2019. Version 1.2.5001.  
Retrieved from <https://rstudio.com>
- [2] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.. Retrieved from <https://cloud.r-project.org/web/packages/ggplot2/index.html>
- [3] Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>
- [4] Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. <https://CRAN.R-project.org/package=rpart>
- [5] Tableau Desktop Professional Edition [Computer Software], (2019). Version 2019.2.3. Retrieved from <https://www.tableau.com>
- [6] Schloerke,Crowley, "CRAN - Package GGally," 17 05 2018. [Online]. Available: <https://cran.r-project.org/web/packages/GGally/index.html>.
- [7] Maechler, Rousseeuw, Struyf , "CRAN - Package Cluster," 19 06 2019. [Online]. Available: <https://cran.r-project.org/web/packages/cluster/index.html>.
- [8] Kassambara, Mundt , "CRAN - Package Factoextra," 05 12 2019. [Online]. Available: <https://cran.r-project.org/web/packages/factoextra/index.html>.
- [9] G. K. C. Lander, "CRAN - Package Useful," 08 10 2018. [Online]. Available: <https://cran.r-project.org/web/packages/useful/index.html>.
- [10] S. L. Wei, "CRAN - Package corrplot," 16 10 2017. [Online]. Available: <https://cran.r-project.org/web/packages/corrplot/index.html>.

[11] Hurley, "CRAN - Package gclus," 07 01 2019. [Online]. Available: <https://cran.r-project.org/web/packages/gclus/index.html>.

[12] Ripley, "CRAN - Package tree," 26 04 2019. [Online]. Available: <https://cran.r-project.org/web/packages/tree/index.html>.

[13] FAO), F. a. (n.d.). Crop Production. Retrieved from Data World: <https://data.world/agriculture/crop-production>

[14] Chakravarthi V. Narasimhan, T. R. (n.d.). britannica. Retrieved from britannica: <https://www.britannica.com/place/Asia/Agriculture>

[15] Nations, F. a. (n.d.). FAOSTAT. Retrieved from FAOSTAT: <http://www.fao.org/faostat/en/#definitions>

## **Appendix:**

```
#install.packages("corrplot")
#install.packages("rpart")
#install.packages("rpart.plot")
#install.packages("gclus")
#install.packages("tree")
#install.packages("GGally")
#install.packages("factoextra")
#install.packages("useful")
library(tidyverse)
library(ggplot2)
library(ggplot2)
library(corrplot)
library(rpart) #used for regression and the various residual plots
library(rpart.plot)
library(gclus)
library(tree) #used for generating the regression tree
library(GGally) #used for the kmeans clustering
library(cluster)
library(factoextra) #to include the fvizcluster function
library(useful) # by Jared Lander

asiacrop <- read_csv(file="C:/Users/chand/Documents/ChandanaNarla/chandana/CourseWork/Sem_1/Assignments/Stat515/FinalProject/Production_Crops_E_Asia.csv")

#asiacrop=read.csv(file = "Production_Crops_E_Asia.csv",header = TRUE)
#asiacrop
summary(asiacrop)
sum(is.na(asiacrop))

#Here we need the data from 2008 to 2013 for predicting the crops production yield in 2014
```

```
#hence considering the required column for analysis
asia_subset1<-asiacrop[,c(1:7,102,104,106,108,110,112,114)]
#datacleaning
sum(is.na(asia_subset1))
#omitting the null values as the mean of the values is large
asia_subset1<-na.omit(asia_subset1)
#check whether the null values are removed
sum(is.na(asia_subset1))

#Subsetting the data in terms of production
subset_4<-
pivot_wider(data=asia_subset1,names_from = Element,values_from = c(Y2014,Y2013
,Y2012,Y2011,Y2010,Y2009,Y2008))%>%select(c(Area,Item,Y2014_Production,Y2013_P
roduction,Y2012_Production,Y2011_Production,Y2010_Production,Y2009_Production,
Y2008_Production))
subset_4=na.omit(subset_4)

#Pairscorrelation plot
#cparis(subset_4)
pairs(subset_4[3:9],colour='blue')

#Linear Regression Model 1 for predicting the Production of Year 2014 using ye
ar 2013
set.seed(100)
trainingIndex1<- sample(1:nrow(subset_4),0.8*nrow(subset_4))
trainingData1<-subset_4[trainingIndex1,]
testData1<-subset_4[-trainingIndex1,]

lmModel1<-lm(Y2014_Production~Y2013_Production, data=trainingData1)
#predict_2 = predict(lmModel1,data.frame(Y2013_Production = 4223))
ElementPredict1<- predict(lmModel1,testData1)
summary(lmModel1)
summary(ElementPredict1)

actual_Predict<-
data.frame(cbind(actuals=testData1$Y2014_Production,predicteds=ElementPredict1
))
corr_accuracy<-cor(actual_Predict)

#Linear Regression Model 2 for predicting the Production of Year 2014 using ye
ars from 2013-2008
trainingIndex<- sample(1:nrow(subset_4),0.8*nrow(subset_4))
trainingData<-subset_4[trainingIndex,]
testData<-subset_4[-trainingIndex,]
```

```
lmModel2<-
lm(Y2014_Production~Y2013_Production+Y2012_Production+Y2011_Production+Y2010_P
roduction+Y2009_Production+Y2008_Production, data=trainingData)
#predict_2 = predict(lmModel2,data.frame(Y2013_Production = 4223,Y2013_Product
ion,Y2013_Production,Y2013_Production,Y2013_Production))
ElementPredict2<- predict(lmModel2,testData)
summary(lmModel2)
par(mfrow=c(2,2))
plot(lmModel2)

actual_Predict1<-
data.frame(cbind(actuals=testData$Y2014_Production,predicteds=ElementPredict2)
)
corr_accuracy1<-cor(actual_Predict1)

#Plot showing the Linear relationship between the predictors.
ggplot(data=subset_4,aes(x=Y2013_Production,y=Y2014_Production))+geom_point()+
stat_smooth(method = 'lm',col="orange")+
  geom_line()

#Regression Trees

#We install the Tree package for the Regression Tree
tree_item=tree(Y2014_Production~Y2012_Production+Y2011_Production+Y2010_Produc
tion+Y2009_Production+Y2008_Production,subset_4,subset=trainingIndex)
summary(tree_item)
plot(tree_item)
text(tree_item,pretty=0,cex=0.9)
tree_item

#Cross validation plot for REgression tree
cv_treeitem<-cv.tree(tree_item)
plot(cv_treeitem)

plot(cv_treeitem$size,cv_treeitem$dev,type="b")
plot(cv_treeitem$k,cv_treeitem$dev,type="b")

#KMeans Clustering

cropTrain <- select(subset_4, -Area, -Item)
cropTrain

ggscatmat(cropTrain)

RNGkind(sample.kind="Rounding") # To obtain same results as RFE
set.seed(100)
cropTrain = scale(cropTrain)
head(cropTrain)
```

```
crop <- kmeans(x=cropTrain, centers=3)
# crop

plot.kmeans(crop, data=cropTrain)

fviz_cluster(crop, data = cropTrain)

#Now creating another subset to include Area Code (Region) and item code (Various items)
subset_5<-
pivot_wider(data=asia_subset1,names_from = Element,values_from = c(Y2014,Y2013,Y2012,Y2011,Y2010,Y2009,Y2008))%>%select(c('Area Code',Area,'Item Code',Item,Y2014_Production,Y2013_Production,Y2012_Production,Y2011_Production,Y2010_Production,Y2009_Production,Y2008_Production))
subset_5=na.omit(subset_5)

cropTrain1 <- select(subset_5, -Area, -Item)
cropTrain1
ggscatmat(cropTrain1)

RNGkind(sample.kind="Rounding") # To obtain same results as RFE
set.seed(200)
cropTrain1 = scale(cropTrain1)
head(cropTrain1)
crop1 <- kmeans(x=cropTrain1, centers=3)
crop1
plot.kmeans(crop1, data=cropTrain1)
fviz_cluster(crop1, data = cropTrain1)

#Other Continents comparision

##### AFRICA #####
#####

africacrop <- read_csv(file="C:/Users/chand/Documents/ChandanaNarla/chandana/CourseWork/Sem_1/Assignments/Stat515/FinalProject/Production_Crops_E_Africa.csv")
#africacrop=read_csv(file = "Production_Crops_E_Africa.csv",header = TRUE)
#africacrop
summary(africacrop)
sum(is.na(africacrop))

#Here we need the data from 2008 to 2013 for predicting the crops production yield in 2014
#hence considering the required column for analysis
```

```
africa_subset1<-africacrop[,c(1:7,102,104,106,108,110,112,114)]
#datacleaning
sum(is.na(africa_subset1))
#omitting the null values as the mean of the values is large
africa_subset1<-na.omit(africa_subset1)
#check whether the null values are removed
sum(is.na(africa_subset1))

#Subsetting the data in terms of production
africa_subset_4<-
pivot_wider(data=africa_subset1,names_from = Element,values_from = c(Y2014,Y20
13,Y2012,Y2011,Y2010,Y2009,Y2008))%>%select(c(Area,Item,Y2014_Production,Y2013
_Production,Y2012_Production,Y2011_Production,Y2010_Production,Y2009_Production,Y2008_Production))
africa_subset_4=na.omit(africa_subset_4)

#Linear Regression Model 1 for predicting the Production of Year 2014 using ye
ar 2013
set.seed(100)
trainingIndex3<- sample(1:nrow(africa_subset_4),0.8*nrow(africa_subset_4))
trainingData3<-africa_subset_4[trainingIndex3,]
testData3<-africa_subset_4[-trainingIndex3,]

lmModel3<-lm(Y2014_Production~Y2013_Production, data=trainingData3)
ElementPredict3<- predict(lmModel3,testData3)
summary(lmModel3)
summary(ElementPredict3)

actual_Predict<-
data.frame(cbind(actuals=testData3$Y2014_Production,predicteds=ElementPredict3
))
corr_accuracy<-cor(actual_Predict)

#Linear Regression Model 2 for predicting the Production of Year 2014 using ye
ars from 2013-2008
trainingIndex4<- sample(1:nrow(africa_subset_4),0.8*nrow(africa_subset_4))
trainingData4<-subset_4[trainingIndex4,]
testData4<-africa_subset_4[-trainingIndex4,]
lmModel4<-
lm(Y2014_Production~Y2013_Production+Y2012_Production+Y2011_Production+Y2010_P
roduction+Y2009_Production+Y2008_Production, data=trainingData)
ElementPredict4<- predict(lmModel4,testData4)
summary(lmModel4)
par(mfrow=c(2,2))
plot(lmModel4)
```

```
actual_Predict4<-
data.frame(cbind(actuals=testData4$Y2014_Production,predicteds=ElementPredict4
))
corr_accuracy4<-cor(actual_Predict4)

#Plot showing the Linear relationship between the predictors.
ggplot(data=africa_subset_4,aes(x=Y2013_Production,y=Y2014_Production))+geom_p
oint()+stat_smooth(method = 'lm',col="orange")+
  geom_line()

#Regression Trees

#We install the Tree package for the Regression Tree
tree_item1=tree(Y2014_Production~Y2013_Production+Y2012_Production+Y2011_Produ
ction+Y2010_Production+Y2009_Production+Y2008_Production,subset_4,subset=train
ingIndex)
summary(tree_item1)
plot(tree_item1)
text(tree_item1,pretty=0,cex=0.9)
tree_item1

#KMeans Clustering

cropTrain2 <- select(africa_subset_4, -Area, -Item)
cropTrain2

ggscatmat(cropTrain2)

RNGkind(sample.kind="Rounding") # To obtain same results as RFE
set.seed(100)
cropTrain2 = scale(cropTrain2)
head(cropTrain)
crop2 <- kmeans(x=cropTrain2, centers=3)
crop2
plot.kmeans(crop2, data=cropTrain2)
fviz_cluster(crop2, data = cropTrain2)
#Now creating another subset to include Area Code (Region) and item code (Vari
ous items)
subset_6<-
pivot_wider(data=africa_subset1,names_from = Element,values_from = c(Y2014,Y20
13,Y2012,Y2011,Y2010,Y2009,Y2008))%>%select(c('Area Code',Area,'Item Code',Ite
m,Y2014_Production,Y2013_Production,Y2012_Production,Y2011_Production,Y2010_Pr
oduction,Y2009_Production,Y2008_Production))
subset_6=na.omit(subset_6)

cropTrain3 <- select(subset_5, -Area, -Item)
cropTrain3
```

```
ggscatmat(cropTrain3)

RNGkind(sample.kind="Rounding") # To obtain same results as RFE
set.seed(200)
cropTrain3 = scale(cropTrain3)
head(cropTrain3)
crop3 <- kmeans(x=cropTrain3, centers=3)
crop3

plot.kmeans(crop3, data=cropTrain3)

fviz_cluster(crop3, data = cropTrain3)

##### EUROPE #####
#####

Europecrop <- read_csv(file="C:/Users/chand/Documents/ChandanaNarla/chandana/C
ourseWork/Sem_1/Assignments/Stat515/FinalProject/Production_Crops_E_Europe.csv
")
#Europecrop
summary(Europecrop)
sum(is.na(Europecrop))

#Here we need the data from 2008 to 2013 for predicting the crops production y
eild in 2014
#hence considering the required column for analysis
Europe_subset1<-Europecrop[,c(1:7,102,104,106,108,110,112,114)]
#datacleaning
sum(is.na(Europe_subset1))
#omitting the null values as the mean of the values is large
Europe_subset1<-na.omit(Europe_subset1)
#check whether the null values are removed
sum(is.na(Europe_subset1))

#Subsetting the data in terms of production
Europecrop_subset_4<-
pivot_wider(data=Europe_subset1,names_from = Element,values_from = c(Y2014,Y20
13,Y2012,Y2011,Y2010,Y2009,Y2008))%>%select(c(Area,Item,Y2014_Production,Y2013
_Production,Y2012_Production,Y2011_Production,Y2010_Production,Y2009_Production
,Y2008_Production))
Europecrop_subset_4=na.omit(Europecrop_subset_4)

#Linear Regression Model 1 for predicting the Production of Year 2014 using ye
ar 2013
set.seed(100)
```



```
trainingIndex5<- sample(1:nrow(Europecrop_subset_4),0.8*nrow(Europecrop_subset_4))
trainingData5<-Europecrop_subset_4[trainingIndex5,]
testData5<-Europecrop_subset_4[-trainingIndex1,]

lmModel5<-lm(Y2014_Production~Y2013_Production, data=trainingData5)
ElementPredict5<- predict(lmModel5,testData5)
summary(lmModel5)
summary(ElementPredict5)

#Linear Regression Model 2 for predicting the Production of Year 2014 using years from 2013-2008
trainingIndex6<- sample(1:nrow(Europecrop_subset_4),0.8*nrow(Europecrop_subset_4))
trainingData6<-Europecrop_subset_4[trainingIndex6,]
testData6<-Europecrop_subset_4[-trainingIndex6,]
lmModel6<-
lm(Y2014_Production~Y2013_Production+Y2012_Production+Y2011_Production+Y2010_Production+Y2009_Production+Y2008_Production, data=trainingData)
ElementPredict6<- predict(lmModel6,testData6)
summary(lmModel6)
par(mfrow=c(2,2))
plot(lmModel6)

actual_Predict6<-
data.frame(cbind(actuals=testData6$Y2014_Production,predicteds=ElementPredict6))
corr_accuracy6<-cor(actual_Predict6)

#Plot showing the Linear relationship between the predictors.
ggplot(data=Europecrop_subset_4,aes(x=Y2013_Production,y=Y2014_Production))+geom_point()+stat_smooth(method = 'lm',col="orange")+
  geom_line()

#Regression Trees

#We install the Tree package for the Regression Tree
tree_item2=tree(Y2014_Production~Y2012_Production+Y2011_Production+Y2010_Production+Y2009_Production+Y2008_Production,subset_4,subset=trainingIndex)
summary(tree_item2)
plot(tree_item2)
text(tree_item2,pretty=0,cex=0.9)
tree_item2

#KMeans Clustering

cropTrain4 <- select(Europecrop_subset_4, -Area, -Item)
cropTrain4
```

```
ggscatmat(cropTrain4)

RNGkind(sample.kind="Rounding") # To obtain same results as RFE
set.seed(100)
cropTrain4 = scale(cropTrain4)
head(cropTrain4)
crop4 <- kmeans(x=cropTrain4, centers=3)
crop4

plot.kmeans(crop4, data=cropTrain4)

fviz_cluster(crop4, data = cropTrain4)

#Now creating another subset to include Area Code (Region) and item code (Various items)
subset_6<-
pivot_wider(data=Europe_subset1,names_from = Element,values_from = c(Y2014,Y2013,Y2012,Y2011,Y2010,Y2009,Y2008))%>%select(c('Area Code',Area,'Item Code',Item,Y2014_Production,Y2013_Production,Y2012_Production,Y2011_Production,Y2010_Production,Y2009_Production,Y2008_Production))
subset_6=na.omit(subset_6)

cropTrain5 <- select(subset_6, -Area, -Item)
cropTrain5
ggscatmat(cropTrain5)

RNGkind(sample.kind="Rounding") # To obtain same results as RFE
set.seed(200)
cropTrain5 = scale(cropTrain5)
head(cropTrain5)
crop5 <- kmeans(x=cropTrain5, centers=3)
crop5

plot.kmeans(crop5, data=cropTrain5)

fviz_cluster(crop5, data = cropTrain5)
```

**Team: Anirudh Tunuguntla, Chandana Narla, Ramya Sri Sonar**