Northeastern University
College *of* Professional Studies

INTERMEDIATE ANALYTICS

ALY 6015, CRN 21454

PROFESSOR ROY WADA

MODULE 6 – ASSIGNMENT

FINAL PROJECT REPORT

SUBMITTED BY:

RICHA RAMBHIA, NUID: 001523295

SIDDHI BOROLE, NUID: 002104252

ALYSSA GESUALDI, NUID: 002128070

Dataset for Project:

Food and Agriculture Data of Africa

Dataset chosen for Analysis:

Crop Processing Dataset
Crop Production Dataset

Project Title:

Analysis of Food and Agriculture Data of Africa

Introduction

The project deals with performing the initial analysis of the food and agriculture data of Africa which contains different datasets of the crops for analysis in the country. The task is to perform the exploratory data analysis which gives the information about the statistical values of the dataset for further analysis. Data visualizations and analytical methods are implemented in order to obtain results about the different datasets of the food and agriculture data of Africa.

The analysis of food and agriculture data of Africa is done on the crop production dataset and crop processing dataset which has the details about the crops that are produced and are processing each year in Africa. This analysis is helpful in determining the crop production that takes place every year in Africa based on the region where the subset analysis of the data is performed. The descriptive analysis of the dataset gives the statistical analysis which helps in determining the parameters of the dataset along with the statistical values as the dataset has a large number of parameters.

Analytical methods performed would be helpful in answering the business questions related to the crop production and crop processing data of Africa as the correlation between the parameters of the dataset is known in order to understand the relationship between them further which it would be considered for regression analysis and testing methods. The analysis mainly focuses on the crop production in each country for the year which would also help in predicting the production of crops for the coming years in Africa. This analysis thus would be beneficial in order to determine the growth of crop production and crop processing further helping the agriculture and food industry of Africa.

The dataset was considered for the subset analysis where the data was altered for creating subsets of region based on the country helping in the analysis of crop production within the region. The analytical methods applied after performing EDA, data visualization, and subset analysis are regression model analysis, correlation matrix and plots, and chi-square testing. The results of the analytical methods would thus help in answering the business questions which are further discussed in the project.

The following are the business questions that are considered during the project in order to analyze the answer them with the help of the exploratory data analysis and analytical methods.

Business Questions

1. Is there a correlation between the highest processed crop and country of Africa?
2. Is there a specific year that a certain crop was processed in the extreme amount?
3. How much amount of crop is processed in Africa for the various crops that are produced each year?
4. Is there a correlation between the crop processed and the amount of crop produced?
5. What would be the crop production in Africa in the years to come?

Dataset Description

The crop production dataset contains the data about the different types of crops produced in Africa which has **3187 rows** of data. Similar to this, the crop processing dataset contains information about the country and the production tonnes in the year. This has **2916 rows** of data and **23 data variables**. These datasets would thus help to perform the exploratory data analysis, data visualization and perform different analytical methods for the same which would answer the business questions and help in the improvement of the food and agriculture data of Africa.

Exploratory Data Analysis

The first task of the project is to create subset of the data for which the subset analysis is performed, and the summary statistics is developed for the same which gives an overview of the subset data generated. Since the dataset has the country parameter, the subset for the same is created which would have a new parameter into consideration as the region. The region of the country is analyzed for the crop production data and crop processing data and thus analysis based on the region can be considered rather than analyzing the data based on each country. This would further help in analyzing the food and agriculture data of Africa specifically to the region. The net production of crops was computed and visualizations for the same were plotted which gave a better understanding of the dataset. The summary statistics of the dataset is displayed which gives the values of the statistics of the data.

Subset Analysis:

Processed Crop by Region:

We have separated the data on processed crops into the regions of Africa by North, South, West, Central, East using the associated 55 countries within the data set. Note that there was some missing data in some of the associated process values which were replaced with the value of 0 to not skew the mean or max as this was the focus of our data

analyzation. The value of total tonnes within the tables below is evaluated as the total tonnes per country processed. We will focus on the three crops Cotton Lint Production, Oil Soybean Production Tonnes, Bales of Barley Production, and the sum of processed crops in tonnes.

Northern Africa: First, we analyzed the data with the Northern region of Africa with **371 observations** from the associated 7 countries. With this we were able to see that there was no processing of Oil Palm, Palm Kernels, and Oil Coconut Copra.

Southern Africa: We next analyzed the data with the Southern region of Africa with **530 observatio**ns from the associated 10 countries. With this we were able to see that there was no processing of Oil Olive Virgin, Oil in Seed, Oil Safflower.

Central Africa: We next analyzed the data with the Central region of Africa with **477 observations** from the associated 9 countries. With this we were able to see that there was no processing of Oil Olive Virgin, Oil Rapeseed, Oil Safflower, Margarine Short, Oil Sunflower, Wine Oil in Seed, and Oil Maize.

East Africa: We next analyzed the data with the East region of Africa with **741 observations** from the associated 14 countries. With this we were able to see that there was no processing of Oil Olive Virgin.

West Africa: We next analyzed the data with the West region of Africa with **795 observations** from the associated 15 countries. With this we were able to see that there was no processing of Oil Olive Virgin.

```
#1.SUBSET ANALYSIS:


#EDA on the subset data


#subset by Regions
north <- subset(codebook, region == "North", select=c(BB_PT:sum))
descripnorth<--psych ::describe(north)
write_xlsx(descripnorth,"descripnorth.xlsx")


south <- subset(codebook, region == "South", select=c(BB_PT:sum))
descripsouth<--psych ::describe(south)
write_xlsx(descripsouth,"descripsouth.xlsx")


west <- subset(codebook, region == "West", select=c(BB_PT:sum))
descripwest<--psych ::describe(west)
write_xlsx(descripwest,"descripwest.xlsx")


east <- subset(codebook, region == "East", select=c(BB_PT:sum))
descripeast<--psych ::describe(east)
write_xlsx(descripeast,"descripeast.xlsx")


central <- subset(codebook, region == "Central", select=c(BB_PT:sum))
descripcentral<--psych ::describe(central)
write_xlsx(descripcentral,"descripcentral.xlsx")
```

The output of the subset analysis is as shown in the below output.

Output: Statistical Analysis of the Crop Production by Region

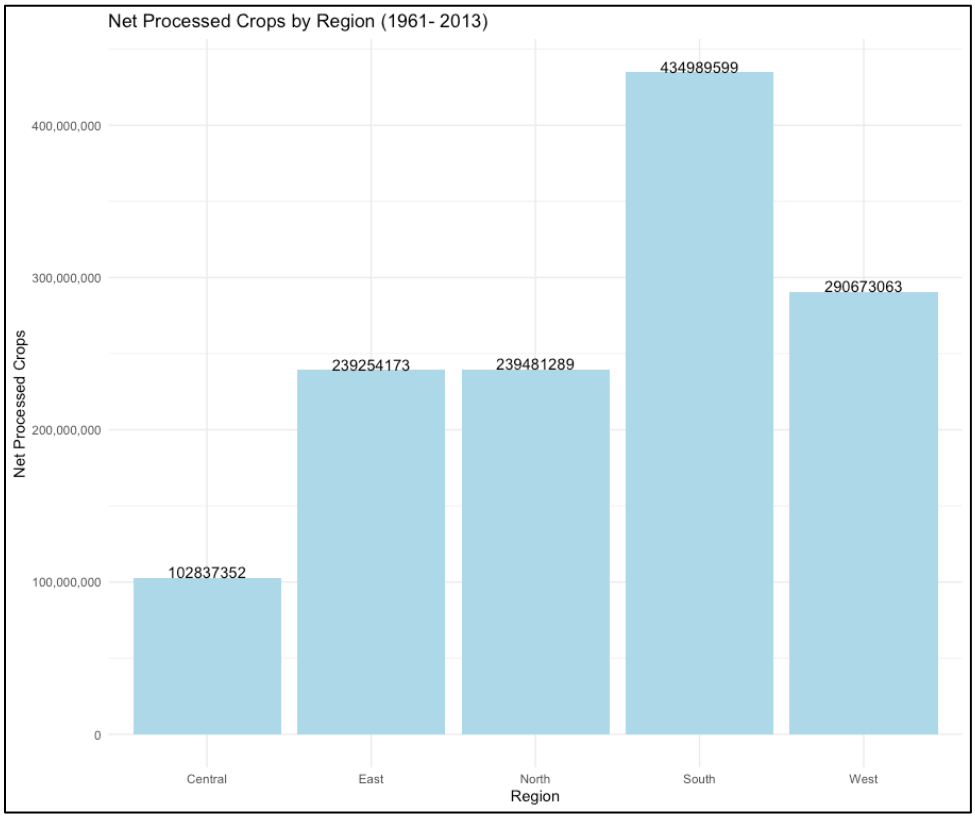
The output below shows the statistical analysis of the crop production by region which depicts the crop type for each region for the various statistical methods like mean, median, minimum, and maximum values. As observed, the mean of the barley production in Northern Africa is **40115.45** and the maximum value for the same is **284000**. Similarly, the statistical values of the crop production type in the various regions have been computed.

Crop Type	Northern Africa	Southern Africa	Central Africa	East Africa	West Africa
Mean Bales of Barley Production	40115.45	224373.73	76652.75	59336.50	74506.48
Max Bales of Barley Production	284000	3150000	750000	842856.15	2650000
Mean Cotton Lint Production	50710.48	16000.26	13419.35	18041.25	28725
Max Cotton Lint Production	541450	128000	104000	270548	283300
Mean Oil SoyBean Production Tonnes	10859.43	2798.35	26.71	562.13	107.53
Max Oil SoyBean Production Tonnes	308300	97700	859.69	26300	3600
Mean Total tonnes processed	645502.13	820735.09	215591.93	322444.98	365626.49
Max Total tonnes processed	3726900	8384374.78	1560040	1551924.33	5826387.28

The data visualization for the subset data for the crop production by region is as shown below.

Output: Net Processed Crops by Region

```r
#visualization

#plot1
Plot1= ggplot(newregionsum, aes(x=Group.1, y=x))+
  ggtitle("Net Processed Crops by Region (1961- 2013)") +
  ylab("Net Processed Crops") + xlab("Region") +
  geom_bar(stat='identity', fill= "lightblue") +
  geom_text(aes(label=round(x,digits=0) , vjust = 0)) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
Plot1
```

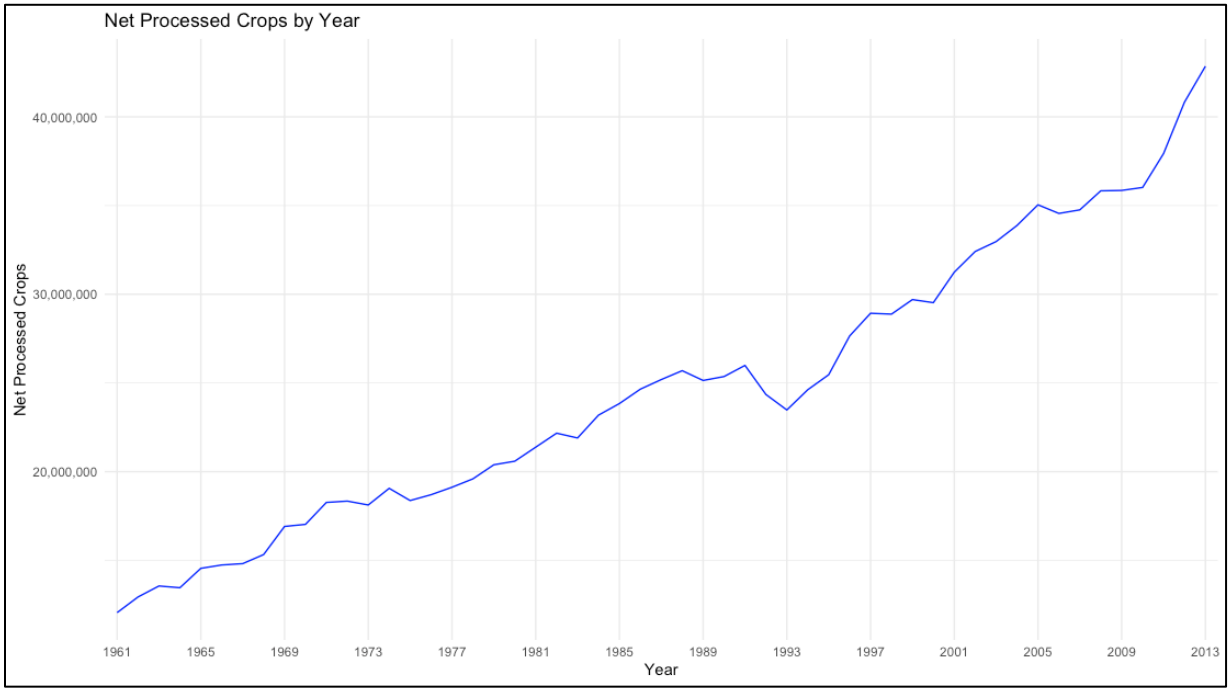


By separating the dataset into the five regions of Africa we were able to analyze the net total of processed crops per region. With the visual aid, we were able to see that Southern Africa processes the greatest amount in tonnes.

Output: Net Processed Crops by Year

We analyzed the net of processed crops by year for the entirety of Africa. We can see the positive relationship that occurs between year and the net processed crops at more than double in 2013 from the value of in 1961.

```r
#visualization

#plot2
Plot2= ggplot(yearsum, aes(x=Group.1, y=x))+
  ggtitle("Net Processed Crops by Year") +
  ylab("Net Processed Crops") + xlab("Year") +
  geom_line(stat='identity', color="blue") +
  scale_x_discrete(limits=c(1961,1965,1969,1973,1977,1981,1985,1989,1993,1997,2001,2005,2009,2013)) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
Plot2
```



Data cleaning was performed on the crop processed dataset after which the exploratory data analysis for statistical and descriptive analysis data is performed in order to generate a statistical table for the same. The describe function, skim function, and stargazer function is used for the descriptive analysis of the crop processing dataset. The output of the descriptive and statistical analysis of the dataset performed is as shown below.

Output: Describing the Crop Processed dataset (Number of observations = 2915)

The describe function gives an overview of the statistical values of the parameters of the dataset where the number of observations for the dataset is 2915 and the mean, standard deviation, minimum and maximum values, range, and skew are displayed for the various parameters of the data.

```
#descriptive analysis

#using the describe function to describe the dataset
#descriptive statistics of the dataset
describe(final_crop_processed_data, na.rm = TRUE, interp=FALSE, skew = TRUE, ranges = TRUE, trim=.1,
         type=3, check=TRUE, fast=NULL, quant=NULL, IQR=FALSE, omit=FALSE)

#describing the dataset
describeData(final_crop_processed_data,head=4,tail=4)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
Country*	1	2915	28.00	15.88	28.00	28.00	20.76	1	55.00	54.00	0.00
Region*	2	2915	3.15	1.47	3.00	3.18	1.48	1	5.00	4.00	-0.05
BB_PT	3	2915	93867.79	302641.97	16500.00	33013.87	24462.90	0	3150000.00	3150000.00	6.70
CL_PT	4	2915	23985.50	60160.04	1647.00	9628.71	2441.84	0	541450.00	541450.00	4.77
CS_PT	5	2915	40502.89	100929.43	3000.00	16695.34	4447.80	0	961000.00	961000.00	4.94
MS_PT	6	2915	4349.22	19490.20	0.00	185.68	0.00	0	200880.00	200880.00	6.11
M_PT	7	2915	45617.71	118144.84	3500.00	17872.09	5189.10	0	915000.00	915000.00	4.71
OGN_PT	8	2915	15047.75	49368.51	1648.57	4901.91	2444.17	0	630000.00	630000.00	7.02
OOV_PT	9	2915	3874.69	20528.80	0.00	0.00	0.00	0	310000.00	310000.00	7.87
OR_PT	10	2915	733.58	7260.54	0.00	0.00	0.00	0	354000.00	354000.00	40.02
OS_PT	11	2915	44.62	364.30	0.00	0.00	0.00	0	4275.75	4275.75	8.98
OSUN_PT	12	2915	4607.05	26248.78	0.00	266.93	0.00	0	340700.00	340700.00	8.73
SRC_PT	13	2915	131430.96	334391.44	11720.00	51987.76	17376.07	0	2914000.00	2914000.00	4.68
W_PT	14	2915	21489.77	120489.57	0.00	73.26	0.00	0	1500000.00	1500000.00	7.06
OCS_PT	15	2915	5333.11	16168.99	156.00	1796.71	231.29	0	208000.00	208000.00	6.64
OP_PT	16	2915	28575.00	113002.49	0.00	4636.53	0.00	0	1330000.00	1330000.00	6.70
OPK_PT	17	2915	6001.80	37935.31	0.00	709.68	0.00	0	573666.00	573666.00	11.72
PK_PT	18	2915	17132.74	86839.08	0.00	3264.69	0.00	0	1230000.00	1230000.00	10.16
OCC_PT	19	2915	1776.46	5477.16	0.00	406.19	0.00	0	47360.00	47360.00	4.68
OSB_PT	20	2915	2067.69	14544.32	0.00	6.44	0.00	0	308300.00	308300.00	14.68
OS_PT.1	21	2915	1552.48	7171.71	0.00	34.09	0.00	0	92086.00	92086.00	6.95
OIS_PT	22	2915	460.48	3147.60	0.00	0.03	0.00	0	59500.00	59500.00	11.27
oilmaize_production_tonnes	23	2915	2003.52	9409.56	0.00	0.00	0.00	0	80700.00	80700.00	6.11
Sum	24	2915	448451.28	945253.83	146444.92	242562.84	208446.27	0	8384374.78	8384374.78	4.69
Year	25	2915	1987.00	15.30	1987.00	1987.00	19.27	1961	2013.00	52.00	0.00

The descriptive statistics of the crop processing data shows the minimum, median, mean, and maximum values of the selected variables from the dataset as there is a large number of parameters in the dataset and thus filtering of the dataset to select the required parameters is effective.

Output: Descriptive Statistics of Crop Processing Data

Table: Descriptive Statistics of Crop Processing Data

	BalesProd	CottonLintProd	CottonSeedProd	OilNutProd	OilSunProd	SugarProd	WineProd	OilCottonsProd	OilPalmProd	OilSoybeanProd	OilSeasmeProd
Min.	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.00
1st Qu.	260.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.00
Median	16500.00	1647.00	3000.00	1648.57	0.00	11720.0	0.00	156.00	0.0	0.00	0.00
Mean	93867.79	23985.50	40502.89	15047.75	4607.05	131431.0	21489.77	5333.11	28575.0	2067.69	1552.48
3rd Qu.	64173.50	16850.00	30000.00	10000.00	18.94	102189.0	0.00	3063.50	6000.0	0.00	0.00
Max.	3150000.00	541450.00	961000.00	630000.00	340700.00	2914000.0	1500000.00	208000.00	1330000.0	308300.00	92086.00
	302641.97	60160.04	100929.43	49368.51	26248.78	334391.4	120489.57	16168.99	113002.5	14544.32	7171.71

The summary statistics of the crop processed dataset is as follows which gives the summary analysis of the dataset using the skim and stargazer function.

```r
#summary statistics
skim(final_crop_processed_data)


#analyzing statistical results
stargazer(final_crop_processed_data,
          title = "Descriptive Analysis of the Crop Processed Data",
          header = FALSE,
          single.row = TRUE)


#importing stargazer to table
stargazer(final_crop_processed_data,  type="html", out="CropProcessedData_SummaryOutput.html")
```

Output: Statistical Analysis of the dataset

Statistic	N	Mean	St. Dev.	Min	Max
BB_PT	2,915	93,867.790	302,642.000	0.000	3,150,000.000
CL_PT	2,915	23,985.500	60,160.040	0.000	541,450.000
CS_PT	2,915	40,502.890	100,929.400	0.000	961,000.000
MS_PT	2,915	4,349.215	19,490.200	0.000	200,880.000
M_PT	2,915	45,617.710	118,144.800	0.000	915,000.000
OGN_PT	2,915	15,047.750	49,368.510	0.000	630,000.000
OOV_PT	2,915	3,874.691	20,528.790	0	310,000
OR_PT	2,915	733.583	7,260.538	0.000	354,000.000
OS_PT	2,915	44.620	364.302	0.000	4,275.750
OSUN_PT	2,915	4,607.051	26,248.780	0.000	340,700.000
SRC_PT	2,915	131,431.000	334,391.400	0.000	2,914,000.000
W_PT	2,915	21,489.770	120,489.600	0.000	1,500,000.000
OCS_PT	2,915	5,333.109	16,168.990	0.000	208,000.000
OP_PT	2,915	28,575.000	113,002.500	0.000	1,330,000.000
OPK_PT	2,915	6,001.796	37,935.310	0.000	573,666.000
PK_PT	2,915	17,132.740	86,839.090	0	1,230,000
OCC_PT	2,915	1,776.456	5,477.161	0.000	47,360.000
OSB_PT	2,915	2,067.686	14,544.320	0.000	308,300.000
OS_PT.1	2,915	1,552.476	7,171.711	0.000	92,086.000
OIS_PT	2,915	460.480	3,147.601	0.000	59,500.000
oilmaize_production_tonnes	2,915	2,003.525	9,409.559	0.000	80,700.000
Sum	2,915	448,451.300	945,253.800	0.000	8,384,375.000
Year	2,915	1,987.000	15.300	1,961	2,013

The statistical analysis of the crop processing dataset is as displayed above which gives the values of the number of observations, mean, standard deviation, minimum, and maximum value for all the parameters of the data. This helps in identifying and analyzing each parameter based on comparing its statistical value with each other for further analysis. Here, we observe that the number of observations is **2915** for the crop processing dataset.

In order to better understand and analyze the data, new variables for the dataset are created for further analysis and visualization of the dataset.

Data Visualization:

Data visualization was performed on the dataset which is one of the analytical methods used to analyze and identify the results. The graphs plotted for the crop processing dataset and subset dataset is as shown below.

Output: R Code for Data Visualization

```
#data visualization

#Data Analysis of Crop Processing Data through visualization

#graph 1: Bales of Barley Production within the region
barley<-ggplot(final_crop_processed_data,aes(y=BB_PT,x=Region)) +
  geom_bar(width=0.2,stat="identity", fill="lightblue3")+
  labs(x="Region")+
  labs(y= "Bales of Barley Production")+
  ggtitle("Bales of Barley Production within the Region")

#graph 2: Cotton Production within the region
cotton<-ggplot(final_crop_processed_data,aes(y=CL_PT,x=Region)) +
  geom_bar(width=0.2,stat="identity", fill="steelblue3")+
  labs(x="Region")+
  labs(y= "Cotton Lint Production")+
  ggtitle("Cotton Lint Production within the Region")

#graph 3: Wine Production Tonnes within the region
wine<-ggplot(final_crop_processed_data,aes(y=W_PT,x=Region)) +
  geom_bar(width=0.2,stat="identity", fill="cadetblue4")+
  labs(x="Region")+
  labs(y= "Wine Production Tonnes")+
  ggtitle("Wine Production Tonnes within the Region")

#graph 4: Sugar Production Tonnes within the region
sugar<-ggplot(final_crop_processed_data,aes(y=SRC_PT,x=Region)) +
  geom_bar(width=0.2,stat="identity", fill="deepskyblue2")+
  labs(x="Region")+
  labs(y= "Sugar Production Tonnes")+
  ggtitle("Sugar Production Tonnes within the Region")

ggarrange(barley, cotton, wine, sugar, ncol=1, nrow=4)
```

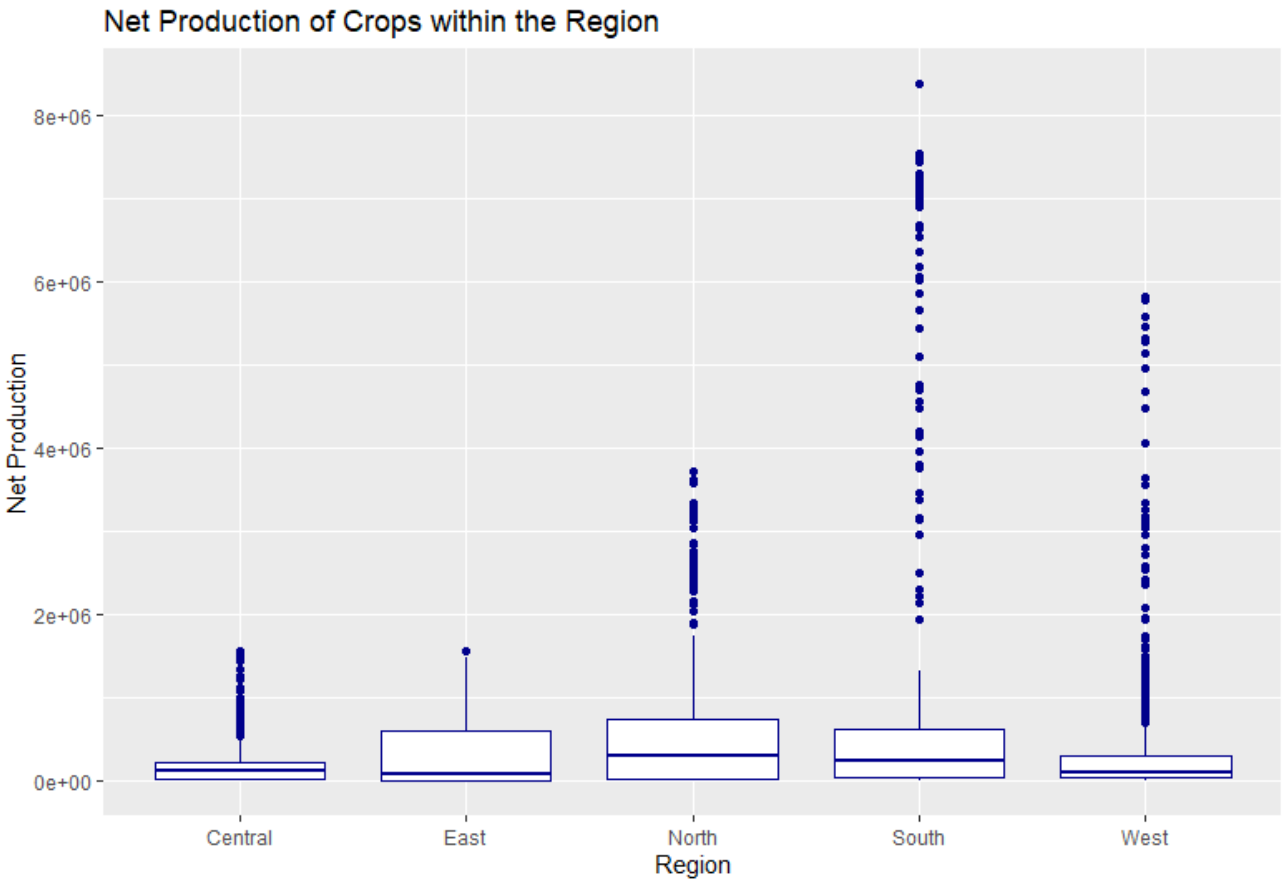
Output: Analysis of Crop Production in the Region

The above visual represents the bar plot for the various crops produced within the Region. The crops considered and selected for the analysis with respect to the region are bales of barley production, cotton lint production, wine production tonnes, and sugar production tonnes where we can analyze the different crop production and compare them within the region. As observed from the visual, the barley production in south of region has the highest crop production whereas the wine production in the central and west region has the lowest crop production in Africa. The sugar and wine production in the south region has the highest crop production, and thus from the analysis we can conclude that the crop production lowest in the regions should be considered for improvement. Also, the cotton lint production in the west region is having the highest crop production in Africa.

Output: Net Production of Crops within the Region

```
#graph 5: Net Production of crops within the region
ggplot(final_crop_processed_data,aes(y=Sum,x=Region)) +
  geom_boxplot(color = "darkblue")+
  labs(x="Region")+
  labs(y= "Net Production")+
  ggtitle("Net Production of Crops within the Region")
```

The boxplot below shows the visual representation of the net production of crops within the regions of Africa, and it can be observed the net production of the crops which is the total production of crops within the regions of Africa is highest in the northern and southern regions of Africa and lesser in the central region as compared to others.



Analytical Methods

The next step was to perform the analytical methods on the dataset and identify the results using regression analysis, correlation matrix and plot, chi-square test, ridge regression methods, or lasso regression methods. The analytical methods that were used and applied on the crop processing dataset and the subset data in order to generate report and analyze the results are **Correlation plot, Regression Model analysis, and Chi-Square Test** which are as follows.

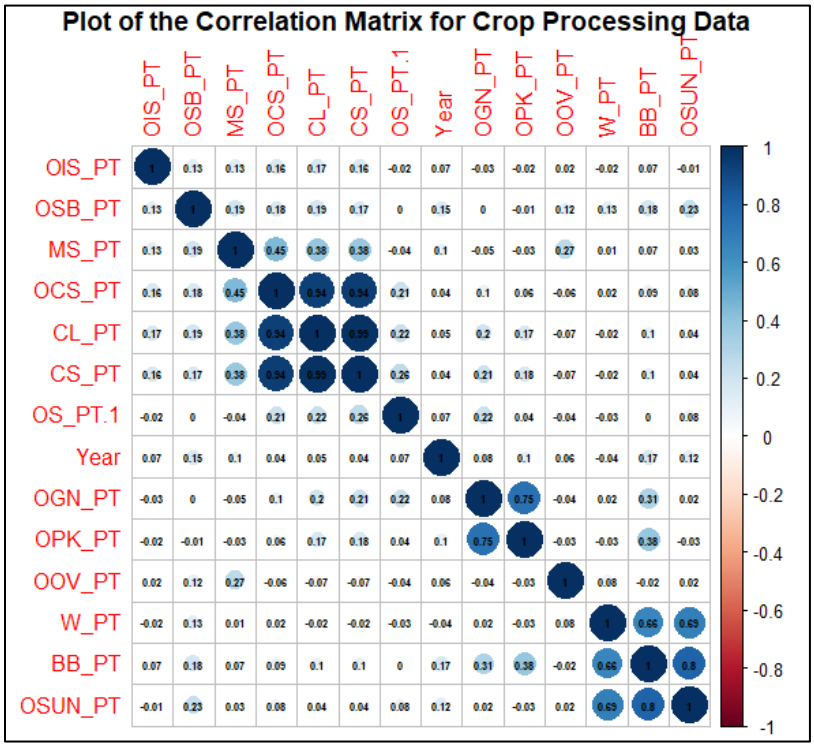
Correlation Matrix and Correlation Plot:

Correlation matrix and plot helps in identifying and understanding the relationship between the two variables of the dataset. The crop processing dataset and subset of the data generated was used in order to plot the correlation plot which helped in understanding the correlation between the various parameters of the dataset, in turn helping for further analysis during the regression model building. The plot of the correlation matrix generated is as follows.

```r
#a. ANALYTICAL METHOD - CORRELATION

#correlation matrix
correlation_data <- final_crop_processed_data %>% dplyr::select (-Country, -Region, -Sum, -M_PT, -OR_PT,
                                                                 -OS_PT, -SRC_PT, -OP_PT, -PK_PT, -OCC_PT,
                                                                 -oilmaize_production_tonnes)
colnames(correlation_data)

correlation_data <- as.data.frame(sapply(correlation_data, as.numeric))
correlation_matrix <- cor(correlation_data)
correlation_matrix


#Plot of the correlation matrix
corrplot(correlation_matrix, order="hclust", title="Plot of the Correlation Matrix for Crop Processing Data",
         number.cex=(0.50), addCoef.col = "black", mar=c(0,0,1,0))
```

Output: Plot of the Correlation Matrix

The plot of the correlation matrix gives an overview of the relationship between the parameters of the dataset. It can be noted that the darker shade of color depicts a strong linear relationship between the variables whereas the lighter shade of color depicts a weak relationship between the variables of the dataset. The values here are between -1 and 1 where 1 represents a positive relationship between the variables and -1 represents a negative relationship between the variables. Thus, it can be observed that cotton lint production and cotton seed production have the highest correlation between them as the correlation value obtained is 0.99. Thus, there is a strong linear relationship between the cotton lint and cotton seed production of Africa.



Regression Model Analysis:

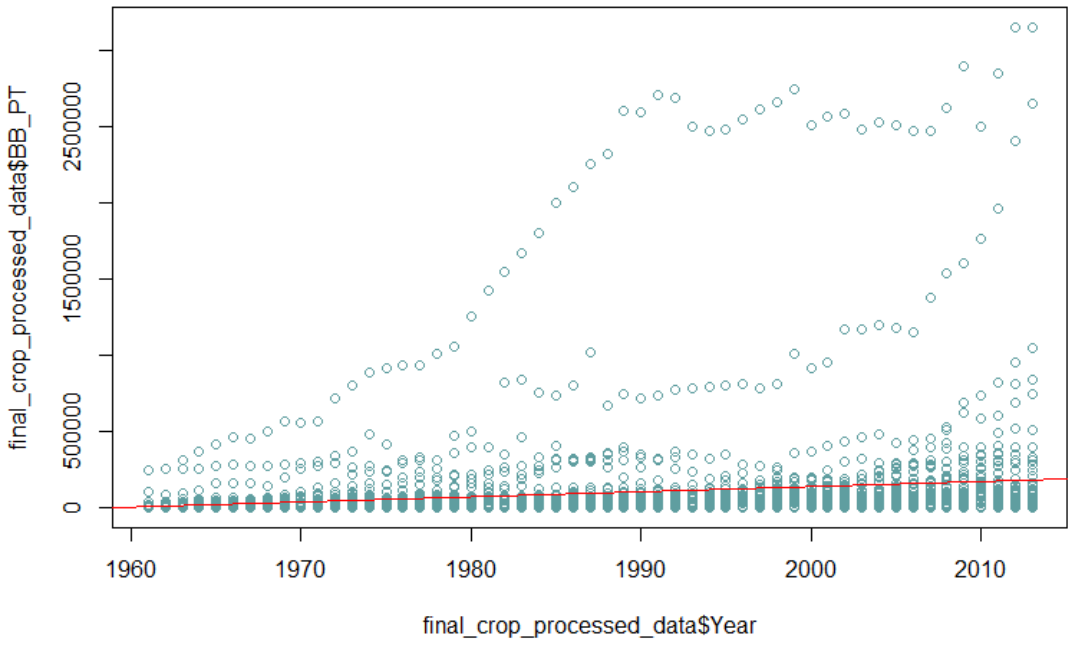
The next analytical method applied to the crop dataset is the regression model analysis where the linear regression model is built and applied in order to analyze the results obtained. The regression model was applied to a set of selected crops based on the year and region for the various countries of Africa. The correlation value is computed in order to understand the relation between the two parameters of the data. The output obtained for the selected crops and the subset data in the regression model analysis is as follows.

Output: R Code – Regression Analysis Model

```
#b. ANALYTICAL METHOD - REGRESSION ANALYSIS


#correlation value for BARLEY production in the year
correlation_value_barleyproduction <- cor(final_crop_processed_data$Year, final_crop_processed_data$BB_PT)
correlation_value_barleyproduction

#linear regression model for BARLEY production in the year
plot(final_crop_processed_data$Year, final_crop_processed_data$BB_PT, col = "cadetblue")
regression_model_1 <- lm(BB_PT ~ Year, data = final_crop_processed_data)
regression_model_1
abline(regression_model_1, col = "red")
summary(regression_model_1)

#regression table for BARLEY production in the year
tbl_regression(regression_model_1)


#correlation value for COTTON production in the year
correlation_value_cottonproduction <- cor(final_crop_processed_data$Year, final_crop_processed_data$CL_PT)
correlation_value_cottonproduction

#linear regression model for COTTON production in the year
plot(final_crop_processed_data$Year, final_crop_processed_data$CL_PT, col = "lightblue")
regression_model_2 <- lm(CL_PT ~ Year, data = final_crop_processed_data)
regression_model_2
abline(regression_model_2, col = "red")
summary(regression_model_2)

#regression table for COTTON production in the year
tbl_regression(regression_model_2)
```

Output: Regression Model for Barley Production in the Year

a. Plot of the regression model



The plot represents the visual of the linear regression model built for the year and the barley production crop of the dataset along with the regression line in the graph.

b. Summary of regression model

The summary of the regression model is as represented below, and it is observed that the p-value is less than the significance level value of 0.05, thus we reject the null hypothesis which is the barley crop production within each year. The adjusted r-squared value is **0.027** which is lower and thus can be concluded that the model is not the best fit regression model.

```
Call:
lm(formula = BB_PT ~ Year, data = final_crop_processed_data)

Residuals:
    Min      1Q  Median      3Q     Max
-179305  -93868  -48423  -11575 2973981

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6435527.7   718142.2  -8.961   <2e-16 ***
Year            3286.1      361.4   9.092   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 298500 on 2913 degrees of freedom
Multiple R-squared:  0.0276,    Adjusted R-squared:  0.02726
F-statistic: 82.67 on 1 and 2913 DF,  p-value: < 2.2e-16
```

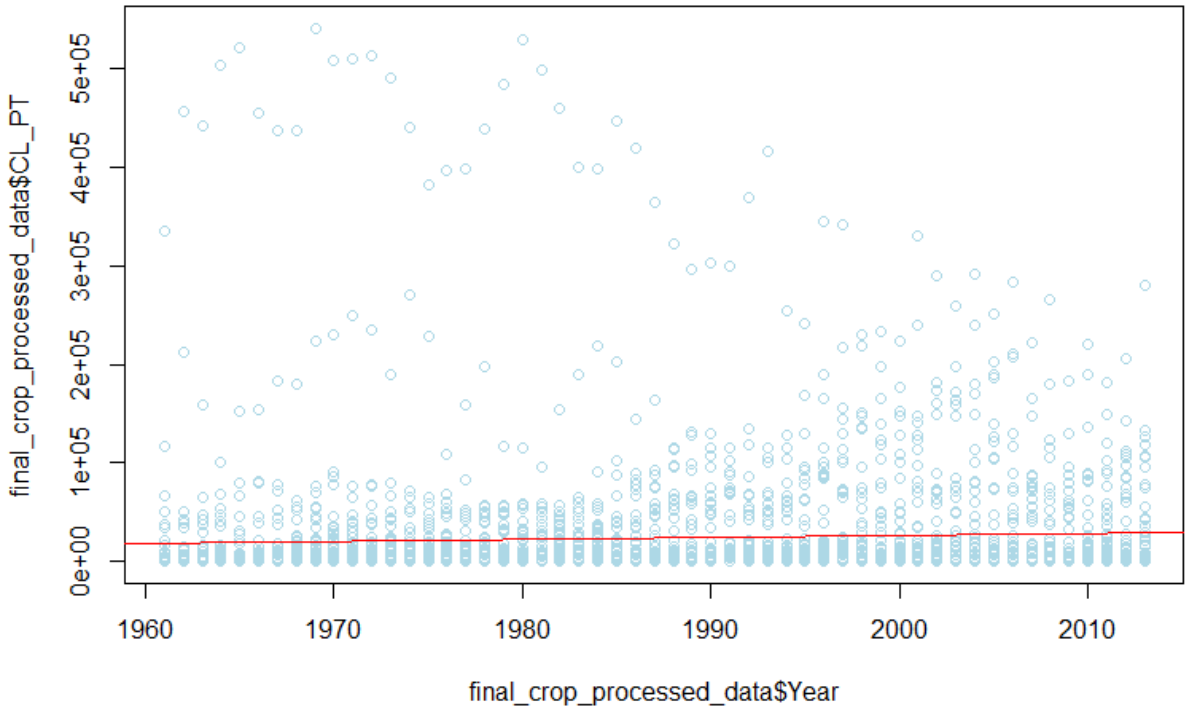
c. Regression Table

Characteristic	Beta	95% CI[1]	p-value
Year	3,286	2,577, 3,995	<0.001

[1] CI = Confidence Interval

Output: Regression Model for Cotton Production in the Year

a. Plot of the regression model



The plot represents the visual of the linear regression model built for the year and the cotton production crop of the dataset along with the regression line in the graph.

b. <u>Summary of regression model</u>

```
Call:
lm(formula = CL_PT ~ Year, data = final_crop_processed_data)

Residuals:
   Min     1Q Median     3Q    Max
-29565 -23991 -19479  -6522 521327

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -402444.65  144550.08  -2.784   0.0054 **
Year            214.61      72.75   2.950   0.0032 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60080 on 2913 degrees of freedom
Multiple R-squared:  0.002979,  Adjusted R-squared:  0.002637
F-statistic: 8.703 on 1 and 2913 DF,  p-value: 0.003202
```

The summary of the regression model is as represented above, and it is observed that the p-value is less than the significance level value of 0.05, thus we reject the null hypothesis which is the cotton crop production within each year. The adjusted r-squared value is **0.026** which is lower and thus can be concluded that the model is not the best fit regression model.

c. Regression Table

Characteristic	Beta	95% CI[1]	p-value
Year	215	72, 357	0.003

[1] CI = Confidence Interval

Output: Regression Model Analysis for Cotton Production in the year

Multiple R-Squared	0.002979
Adjusted R-Squared	0.002637
F-statistic	8.703
DF	2913
P-Value	0.003202

Regression Analysis on Subset Data:

The regression model analysis for the subset data of the various crop production is as follows.

```
#REGRESSION ANALYSIS ON SUBSET DATA

#linear regression model for OIL GROUND NUT production tonnes in the region
regression_model_3 <- lm(OGN_PT ~ Region, data = final_crop_processed_data)
regression_model_3
summary(regression_model_3)

#regression table for OIL GROUND NUT production tonnes in the region
tbl_regression(regression_model_3)


#linear regression model for OIL PALM production tonnes in the region
regression_model_4 <- lm(OP_PT ~ Region, data = final_crop_processed_data)
regression_model_4
summary(regression_model_4)

#regression table for OIL PALM production tonnes in the region
tbl_regression(regression_model_4)
```

Output: Regression Model for Oil Ground Nut Production Tonnes in the Region

a. Summary of regression model

The summary of the regression model is as represented below for the subset of the data based on the region, and it is observed that the p-value is less than the significance level value of 0.05, thus we reject the null hypothesis which is the oil ground nut crop

production in the region. The adjusted r-squared value is **0.07** which is lower and thus can be concluded that the model is not the best fit regression model.

```
Call:
lm(formula = OGN_PT ~ Region, data = final_crop_processed_data)

Residuals:
   Min     1Q Median     3Q    Max
-37077  -8271  -6834  -1693 592923

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8271.3     2174.9   3.803 0.000146 ***
RegionEast    -913.4     2787.7  -0.328 0.743185
RegionNorth  -4971.3     3288.1  -1.512 0.130673
RegionSouth  -1178.7     2997.9  -0.393 0.694214
RegionWest   28805.3     2751.1  10.471  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47500 on 2910 degrees of freedom
Multiple R-squared:  0.07551,   Adjusted R-squared:  0.07424
F-statistic: 59.42 on 4 and 2910 DF,  p-value: < 2.2e-16
```

b. <u>Regression Table</u>

Characteristic	Beta	95% CI[1]	p-value
Region			
Central	—	—	
East	-913	-6,379, 4,553	0.7
North	-4,971	-11,419, 1,476	0.13
South	-1,179	-7,057, 4,699	0.7
West	28,805	23,411, 34,200	<0.001

[1] CI = Confidence Interval

Output: Regression Model Analysis for Oil Ground Nut Production Tonnes in Region

Multiple R-Squared	0.07551
Adjusted R-Squared	0.07424
F-statistic	59.42
DF	2910
P-Value	0.00000000000000022

Output: Regression Model for Oil Palm Production Tonnes in the Region

a. Summary of regression model

```
Call:
lm(formula = OP_PT ~ Region, data = final_crop_processed_data)

Residuals:
    Min      1Q  Median      3Q     Max
 -78611  -36394   -4270       0 1251389

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    38111       4947   7.703 1.81e-14 ***
RegionEast    -37627       6341  -5.934 3.31e-09 ***
RegionNorth   -38111       7480  -5.095 3.71e-07 ***
RegionSouth   -33841       6820  -4.962 7.36e-07 ***
RegionWest     40500       6258   6.472 1.13e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 108100 on 2910 degrees of freedom
Multiple R-squared:  0.08694,   Adjusted R-squared:  0.08569
F-statistic: 69.28 on 4 and 2910 DF,  p-value: < 2.2e-16
```

The summary of the regression model is as represented below for the subset of the data based on the region, and it is observed that the p-value is less than the significance level value of 0.05, thus we reject the null hypothesis which is the oil ground nut crop production in the region. The adjusted r-squared value is **0.085** which is lower and thus can be concluded that the model is not the best fit regression model but is better as compared to the other regression models built.

b. Regression Table

Characteristic	Beta	95% CI[1]	p-value
Region			
Central	—	—	
East	-37,627	-50,061, -25,193	<0.001
North	-38,111	-52,777, -23,445	<0.001
South	-33,841	-47,212, -20,469	<0.001
West	40,500	28,229, 52,770	<0.001

[1] CI = Confidence Interval

Chi-Square Test:

The last analytical method applied to the dataset is the chi-square test for the net production of the crop in the region. The parameters considered here for the test was the net production of the crops rather than the individual crops for the region. The output obtained for the chi-square test is as follows.

Output: Chi-Square Test

```
#c. ANALYTICAL METHOD - CHI-SQUARE TESTING


#chi square test for net production of crop in the region
crop.data <- data.frame(final_crop_processed_data$Region, final_crop_processed_data$Sum)
crop.data = table(final_crop_processed_data$Region, final_crop_processed_data$Sum)

print(chisq.test(crop.data))
```

Output: Chi-Square test for net production of crop in the region

```
        Pearson's Chi-squared test

data:  crop.data
X-squared = 11362, df = 10732, p-value = 1.202e-05
```

The output of the chi-square test obtained shows that the p-value of the crop data is less than the significance level value of 0.05, and thus we reject the null hypothesis which indicates that the net production within the region is to be considered in the years to come.

Conclusion

The crop production and crop processing dataset of Africa was analyzed in order to understand the food and agricultural parameters of the countries and regions of Africa. The exploratory data analysis was performed which gave an overview of the statistical values of the dataset that helped in understanding the parameters of the data. Subset analysis was performed where subset of the data was created based on the regions and data visualizations for the net production of crop was plotted. This gave a better understanding of the crop productions within the region.

The descriptive analysis table generated gave the statistical values of the data for the region in the year that gave an overview of which crops produced are high or less in amount within the countries of Africa. The stargazer and describe function were used in order to generate the statistical tables of the data and generate reports for the same. After performing the EDA on the subset data, data visualization was performed where the visual representation of the parameters of data were generated in order to compare the crop production within each region of Africa that gives an overall understanding where the crop production is more and where the crop production needs attention on.

The different analytical methods applied to the dataset helped in answering the above mentioned business questions for the crop produced data in Africa. The different analytical methods used were correlation matrix and plot, regression analysis model, and chi-square test for the dataset of Africa. The correlation plot helped in determining the relationship between the two parameters of the dataset where it was found that the cotton lint and cotton seed production has a strong correlation relation with the correlation value of 0.99 which is the highest as compared to others. The regression model built for the parameters of the dataset and the subset of the data gave an understanding of the p-value and helped in answering the questions related to the crop production, and since the p-value was less than 0.05, the hypothesis was rejected. The adjusted r-squared value was less, and thus it was observed that the model was not a best fit model for the parameters of the dataset. The regression table generated for the built models depicted the p-value and confidence interval of the model and the results were analyzed. Lastly, chi-square test was performed for the net production crops of the regions in Africa where the p-value obtained was less than the significance value of 0.05 and thus the hypothesis was rejected.

The crop production and crop processing dataset was thus analyzed based on the exploratory data analysis and the analytical methods helped in analyzing the results for the same.

References

Z. (2021f, July 27). How to Subset a Data Frame in R (4 Examples). Statology. Retrieved March 17, 2022, from https://www.statology.org/subset-data-frame-in-r/

GeeksforGeeks. (2021e, December 25). Data Visualization in R. Retrieved March 17, 2022, from https://www.geeksforgeeks.org/data-visualization-in-r/

Facer, C. (2020, December 3). How to Create a Correlation Matrix in R. Displayr. Retrieved March 18, 2022, from https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/#:%7E:text=There%20are%20several%20packages%20available%20for%20visualizing%20a,matrix%20as%20the%20data%20input%20to%20the%20function.

Bevans, R. (2020c, December 14). A step-by-step guide to linear regression in R. Scribbr. Retrieved March 19, 2022, from https://www.scribbr.com/statistics/linear-regression-in-r/

Team, D. (2021, August 25). Chi-Square Test in R | Explore the Examples and Essential concepts! DataFlair. Retrieved March 19, 2022, from https://data-flair.training/blogs/chi-square-test-in-r/