AFRICAN FOOD PRODUCTION AND SUPPLY.

Africa is the second-largest continent in the world but yet we still have a high food crisis. Two dataset in pdf format was given to carry out the analkysis of food shortage in Africa. The datasets, one with the title AFRICA FOOD PRODUVTION and the other Africa food supply. This dataset first had to be converted to csv using (tabula-py module) so I can parse it correctly, then used pandas module to read the csv datasets. The cleaning of the dataset involved checking for missing values and duplicates

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Import the required Module
import tabula
# Read a PDF File
df = tabula.read_pdf("c:\\users\\blessing.onumaegbu\\Desktop\\African Food Production.pdf", encoding="latin-1", pages="all")
# convert PDF into CSV
tabula.convert_into("c:\\users\\blessing.onumaegbu\\Desktop\\African Food Production.pdf",
                    "c:\\users\\blessing.onumaegbu\\Desktop\\African Food Production.csv", output_format="csv", pages='all')
print(df)
```

Figure 1: Converting the African food production pdf to csv format

```
df1 = tabula.read_pdf("c:\\users\\blessing.onumaegbu\\Desktop\\African Food Supply.pdf", encoding="latin-1", pages="all")
# convert PDF into CSV
tabula.convert_into("c:\\users\\blessing.onumaegbu\\Desktop\\African Food Supply.pdf",
                    "c:\\users\\blessing.onumaegbu\\Desktop\\African Food Supply.csv", output_format="csv", pages='all')
print(df1)
```

Figure 2: Converting the African food supply pdf to csv format

DATA EXPLORATION

Pd.read_csv was used to read and parse the csv file

jupyter Untitled5 Last Checkpoint: Last Thursday at 12:55 (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code Voilà

In [3]:
```
df= pd.read_csv("c:\\users\\blessing.onumaegbu\\Desktop\\African Food Production.csv")
df.head()
```

Out[3]:

	Country	Item	Year	Value
0	Algeria	Wheat and products	2004	2731
1	Algeria	Wheat and products	2005	2415
2	Algeria	Wheat and products	2006	2688
3	Algeria	Wheat and products	2007	2319
4	Algeria	Wheat and products	2008	1111

In [4]:
```
df.shape
```

Out[4]: (23110, 4)

Figure 3: Africa Food Production data

The African food production data has a total of 4(four) columns, and 23110 entries, this was gotten using **.info**

```
In [5]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23110 entries, 0 to 23109
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   Country  23110 non-null  object
 1   Item     23110 non-null  object
 2   Year     23110 non-null  int64 
 3   Value    23110 non-null  int64 
dtypes: int64(2), object(2)
```

Figure 4: Africa Food Production info

and **.describe was used to get the mean, median, quartiles, min and max in the Value Column of the dataset**

```
In [15]: df.describe()
```

Out[15]:

	Year	Value
count	23110.000000	23110.000000
mean	2008.498269	327.785201
std	2.871740	1607.940343
min	2004.000000	0.000000
25%	2006.000000	3.000000
50%	2008.000000	18.000000
75%	2011.000000	108.000000
max	2013.000000	54000.000000

Figure 5: Africa Food Production Statistical info

While going through the info of the dataset, the **Year** column was an integer instead of a datetime.

```
import datetime as dt
from datetime import date
```

```
df['Year'] = pd.to_datetime(df['Year'])
df1['Year'] = pd.to_datetime(df1['Year'])
```

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 449 entries, 0 to 448
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Country  449 non-null    object
 1   Year     449 non-null    datetime64[ns]
 2   Value    449 non-null    int64
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 10.6+ KB
```

Figure 6: conversion of the year column d-type and Africa Food Supply info

The **Value** column in each dataset was renamed.

```
df.rename(columns ={"Value": "Value in kt"},
                    inplace=True)
df1.rename(columns={"Value": "Value in Kcal/(PersonDay)"},
                    inplace= True)
```

```
df1.head()
```

	Country	Year	Value in Kcal/(PersonDay)
0	Algeria	2004	2987
1	Algeria	2005	2958
2	Algeria	2006	3047
3	Algeria	2007	3041
4	Algeria	2008	3048

```
df.isnull().sum()
```

```
Country        0
Item           0
Year           0
Value in kt    0
dtype: int64
```

Figure 7: Rename of the Value column

Then the African food production is grouped

```
In [23]: df_group=df.groupby(["Country","Year","Item"])["Value in kt"].sum("Value in kt")
         print(df_group)

Country   Year  Item
Algeria   2004  Apples and products       165
                Bananas                    0
                Barley and products     1212
                Beans                      2
                Beer                     110
                                         ...
Zimbabwe  2013  Tea (including mate)       19
                Tomatoes and products      24
                Vegetables, Other         203
                Wheat and products         25
                Wine                        2
Name: Value in kt, Length: 23110, dtype: int64
```

Figure 8: grouping of the common variables

Checking the statistical measurement of both datasets, the value in Production is less than the value in supply with more outliers in the Production dataset.

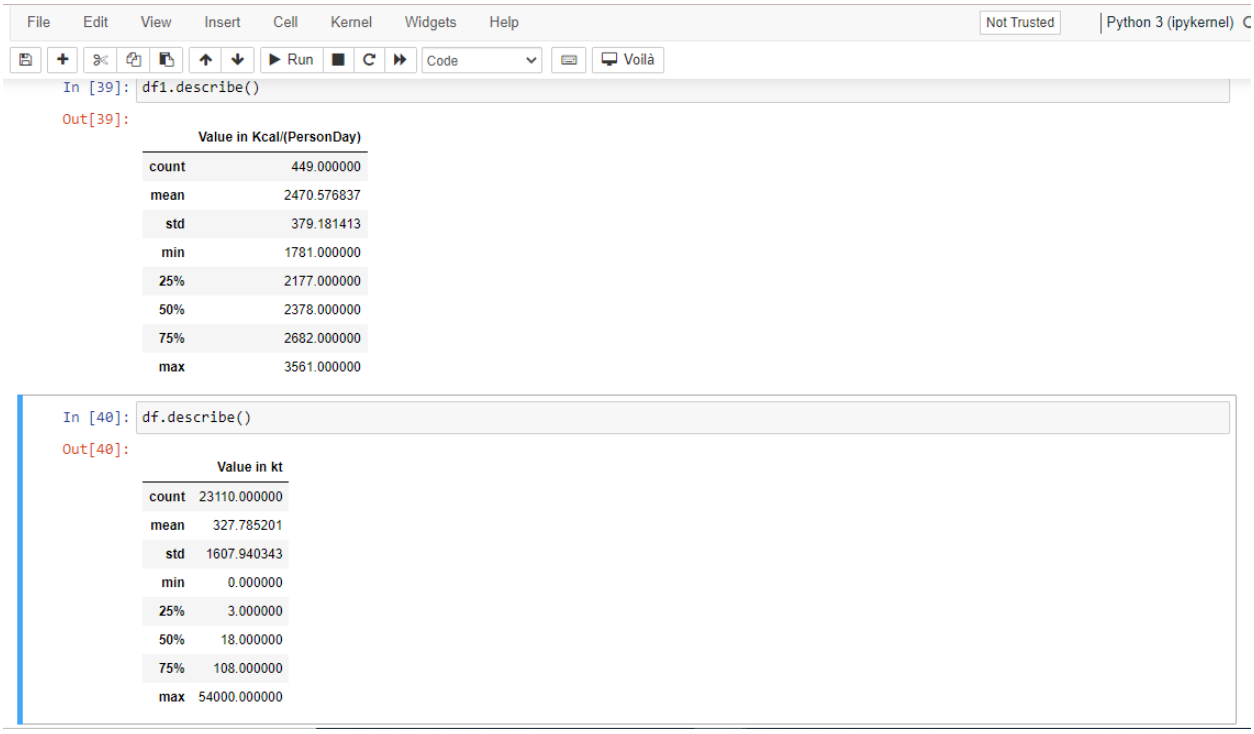

Figure 7: Statistical measures of the Value column in both datasets

RESULT

From the data set given we could see that Nigeria, Egypt and South Africa were the top food-producing countries from 2004 to 2013, but Egypt, Morocco and Tunisia had the highest food supply. If more of the countries could produce more, there would be no food shortage in Africa