# Project 1: FACTOR ANALYSIS REPORT

Name: Anjali Sodani

Oct 31, 2017

# Table of Contents

## 1. Problem Description:

The data given has 26 observed, correlated variables which make it difficult to handle the data because of the repetition of the data. Therefore, we want to reduce the number of variables into smaller factors/themes and detect structure in the relationships between the variables.

Analysis in this report is done on a commodity data to identify the themes (i.e. factors) in the data. We used Factor Analysis technique:

- To take mass data and shrink it to smaller data that is more understandable and manageable
- Reduce the number of variables
- To analyze and find the relationship structure among the variables
- Factor analysis uses mathematical procedures to simplify the interrelated measures to discover patterns in a set of variables

## 2. Executive Summary:

According to my analysis, I have identified 3 factors/themes to classify my 25 variables. The three factors I have identified are Metals, Energy and agriculture, Households.

As you observe below, through this factor analysis technique we are now able to show how strongly our dataset variables are associated with each of the three factors. These associations with the factors can vary from -100% to 100%. Numbers having larger absolute values indicate a stronger association with that particular factor.

We may see some variables in the factors which doesn't make logical sense in the factor but if we see the mathematical interpretation they have high correlation but the factor as well as with the other variables in the factor.

## Factor1: Metals

Below are the commodities with their percentages that come under Metals. From the description we have identified that the import and export is evenly spread across metals and all the variables are more than 70% related to the factor metals. Also, we see that there are two variables PFISH_USD and PROIL_USD which doesn't related to category metals but from the correlation table we see that these two variables are highly correlated with all the other variables in the factor.

| COMMODITIES | DESCRIPTION (Metals) | % |
|---|---|---|
| PCOPP_USD | Copper, grade A cathode, LME spot price, CIF European ports, US$ per metric ton | 95% |
| PNICK_USD | Nickel, melting grade, LME spot price, CIF European ports, US$ per metric ton | 91% |
| PALUM_USD | Aluminum, 99.5% minimum purity, LME spot price, CIF UK ports, US$ per metric ton | 89% |
| PLEAD_USD | Lead, 99.97% pure, LME spot price, CIF European Ports, US$ per metric ton | 85% |
| PIORECR_USD | China import Iron Ore Fines 62% FE spot (CFR Tianjin port), US dollars per metric ton | 72% |
| PFISH_USD | Fishmeal, Peru Fish meal/pellets 65% protein, CIF, US$ per metric ton | 72% |
| PROIL_USD | Rapeseed oil, crude, fob Rotterdam, US$ per metric ton | 72% |

## Factor2: Energy and agriculture

Below are the commodities with their percentages that come under Energy and agriculture. From the description we have identified that the variables are 60%-80% related to factor

| COMMODITIES | DESCRIPTION (Energy and Agriculture) | % |
|---|---|---|
| PLOGSK_USD | Hard Logs, Best quality Malaysian meranti, import price Japan, US$ per cubic meter | 76% |
| PCOALAU_USD | Coal, Australian thermal coal, 12,000- btu/pound, less than 1% sulfur, 14% ash, FOB Newcastle/Port Kembla, US$ per metric ton | 66% |
| POILAPSP_USD | Crude Oil (petroleum), Price index, 2005 = 100, simple average of three spot prices; Dated Brent, West Texas Intermediate, and the Dubai Fateh | 64% |
| PBARL_USD | Barley, Canadian no.1 Western Barley, spot price, US$ per metric ton | 66% |
| PMAIZMT_USD | Maize (corn), U.S. No.2 Yellow, FOB Gulf of Mexico, U.S. price, US$ per metric ton | 63% |
| PCOFFOTM_USD | Coffee, Other Mild Arabicas, International Coffee Organization New York cash price, ex-dock New York, US cents per pound | 60% |
| PCOCO_USD | Cocoa beans, International Cocoa Organization cash price, CIF US and European ports, US$ per metric ton | 74% |
| PBANSOP_USD | Bananas, Central American and Ecuador, FOB U.S. Ports, US$ per metric ton | 73% |
| PHIDE_USD | Hides, Heavy native steers, over 53 pounds, wholesale dealer's price, US, Chicago, fob Shipping Point, US cents per pound | 69% |

## Factor3: Households

Below are the commodities with their percentages that come under households.

| COMMODITIES | DESCRIPTION (Households) | % |
|---|---|---|
| PNGASEU_USD | Natural Gas, Russian Natural Gas border price in Germany, US$ per thousands of cubic meters of gas | 85% |

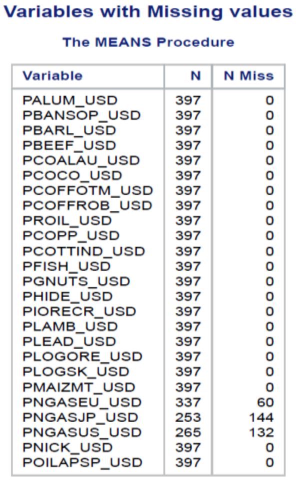
| | | |
|---|---|---|
| PNGASUS_USD | Natural Gas, Natural Gas spot price at the Henry Hub terminal in Louisiana, US$ per thousands of cubic meters of gas | 83% |
| PNGASJP_USD | Natural Gas, Indonesian Liquefied Natural Gas in Japan, US$ per cubic meter of liquid | 78% |
| PBEEF_USD | Beef, Australian and New Zealand 85% lean fores, CIF U.S. import price, US cents per pound | 84% |
| PLAMB_USD | Lamb, frozen carcass Smithfield London, US cents per pound | 65% |
| PGNUTS_USD | Groundnuts (peanuts), 40/50 (40 to 50 count per ounce), cif Argentina, US$ per metric ton | 81% |
| PCOFFROB_USD | Coffee, Robusta, International Coffee Organization New York cash price, ex-dock New York, US cents per pound | 72% |
| PLOGORE_USD | Soft Logs, Average Export price from the U.S. for Douglas Fir, US$ per cubic meter | 67% |
| PCOTTIND_USD | Cotton, Cotton Outlook 'A Index', Middling 1-3/32 inch staple, CIF Liverpool, US cents per pound | 66% |

# 3. Technical Appendix

This appendix introduces the various technical issues that were encountered during the factor analysis of the commodities data which is represented above. The analysis was based on a set of 397 records that were collected from the global commodity market prices.

## 3.1 Data Description:

There were several issues with the data initially, I had a total 144 missing records which makes 36% of the total 397 records, so we cannot simply delete them. So, I dealt with them by imputing the values by mean/median. Fig 1.1 shows that the description of total records(N) and the corresponding missing values by N Miss.



Variables with Missing values

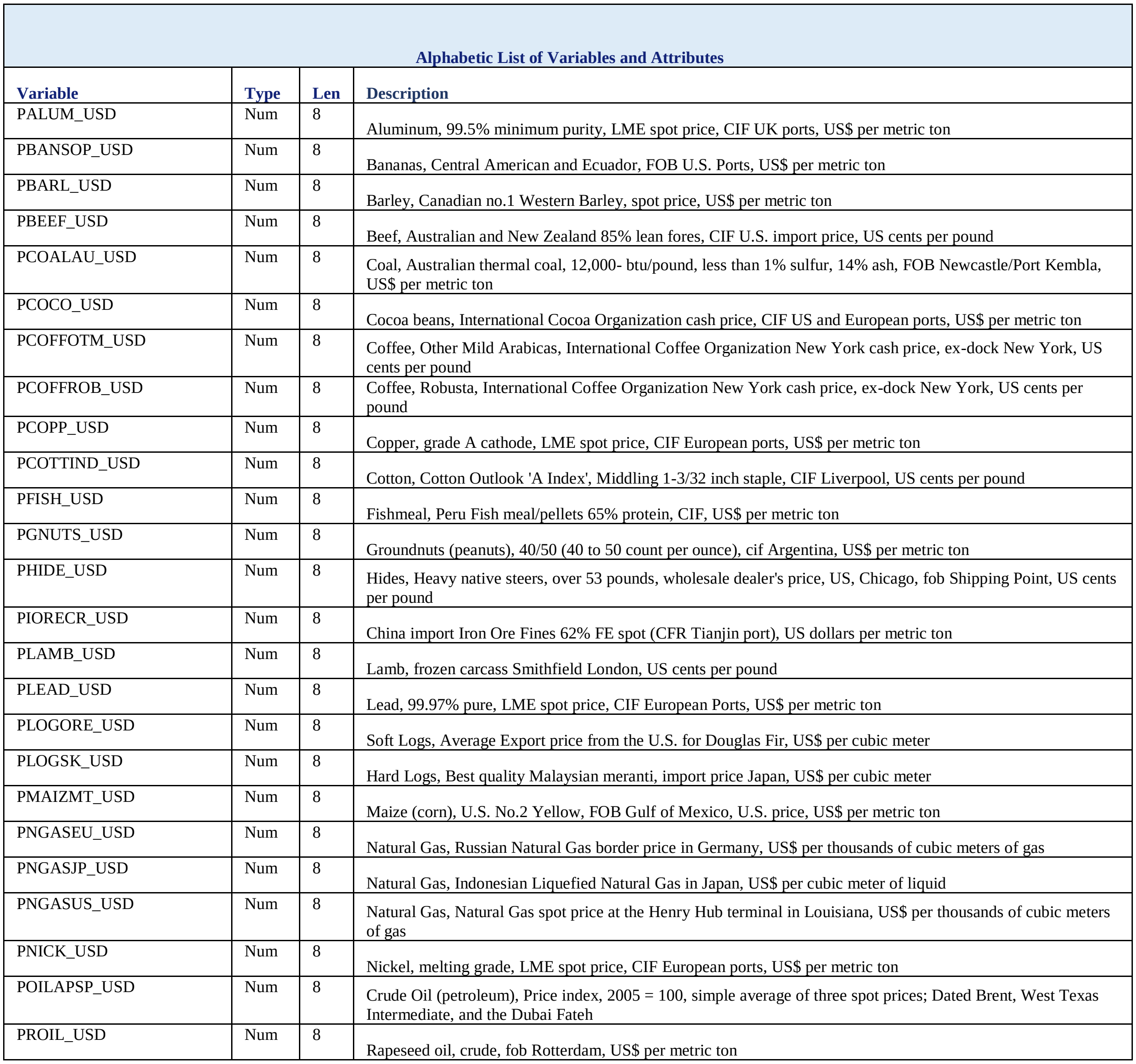
The MEANS Procedure

| Variable | N | N Miss |
|---|---|---|
| PALUM_USD | 397 | 0 |
| PBANSOP_USD | 397 | 0 |
| PBARL_USD | 397 | 0 |
| PBEEF_USD | 397 | 0 |
| PCOALAU_USD | 397 | 0 |
| PCOCO_USD | 397 | 0 |
| PCOFFOTM_USD | 397 | 0 |
| PCOFFROB_USD | 397 | 0 |
| PROIL_USD | 397 | 0 |
| PCOPP_USD | 397 | 0 |
| PCOTTIND_USD | 397 | 0 |
| PFISH_USD | 397 | 0 |
| PGNUTS_USD | 397 | 0 |
| PHIDE_USD | 397 | 0 |
| PIORECR_USD | 397 | 0 |
| PLAMB_USD | 397 | 0 |
| PLEAD_USD | 397 | 0 |
| PLOGORE_USD | 397 | 0 |
| PLOGSK_USD | 397 | 0 |
| PMAIZMT_USD | 397 | 0 |
| PNGASEU_USD | 337 | 60 |
| PNGASJP_USD | 253 | 144 |
| PNGASUS_USD | 265 | 132 |
| PNICK_USD | 397 | 0 |
| POILAPSP_USD | 397 | 0 |

**Fig 1.1**

My dataset had 25 numeric variables which represented the US dollars of the commodity. The description of all the dataset is shown below in the Fig 1.2

where we use the content procedure to get the description of all the data and column Description is attached as an explanatory column to give the detail description of what the variable represents.



Alphabetic List of Variables and Attributes

| Variable | Type | Len | Description |
|---|---|---|---|
| PALUM_USD | Num | 8 | Aluminum, 99.5% minimum purity, LME spot price, CIF UK ports, US$ per metric ton |
| PBANSOP_USD | Num | 8 | Bananas, Central American and Ecuador, FOB U.S. Ports, US$ per metric ton |
| PBARL_USD | Num | 8 | Barley, Canadian no.1 Western Barley, spot price, US$ per metric ton |
| PBEEF_USD | Num | 8 | Beef, Australian and New Zealand 85% lean fores, CIF U.S. import price, US cents per pound |
| PCOALAU_USD | Num | 8 | Coal, Australian thermal coal, 12,000- btu/pound, less than 1% sulfur, 14% ash, FOB Newcastle/Port Kembla, US$ per metric ton |
| PCOCO_USD | Num | 8 | Cocoa beans, International Cocoa Organization cash price, CIF US and European ports, US$ per metric ton |
| PCOFFOTM_USD | Num | 8 | Coffee, Other Mild Arabicas, International Coffee Organization New York cash price, ex-dock New York, US cents per pound |
| PCOFFROB_USD | Num | 8 | Coffee, Robusta, International Coffee Organization New York cash price, ex-dock New York, US cents per pound |
| PCOPP_USD | Num | 8 | Copper, grade A cathode, LME spot price, CIF European ports, US$ per metric ton |
| PCOTTIND_USD | Num | 8 | Cotton, Cotton Outlook 'A Index', Middling 1-3/32 inch staple, CIF Liverpool, US cents per pound |
| PFISH_USD | Num | 8 | Fishmeal, Peru Fish meal/pellets 65% protein, CIF, US$ per metric ton |
| PGNUTS_USD | Num | 8 | Groundnuts (peanuts), 40/50 (40 to 50 count per ounce), cif Argentina, US$ per metric ton |
| PHIDE_USD | Num | 8 | Hides, Heavy native steers, over 53 pounds, wholesale dealer's price, US, Chicago, fob Shipping Point, US cents per pound |
| PIORECR_USD | Num | 8 | China import Iron Ore Fines 62% FE spot (CFR Tianjin port), US dollars per metric ton |
| PLAMB_USD | Num | 8 | Lamb, frozen carcass Smithfield London, US cents per pound |
| PLEAD_USD | Num | 8 | Lead, 99.97% pure, LME spot price, CIF European Ports, US$ per metric ton |
| PLOGORE_USD | Num | 8 | Soft Logs, Average Export price from the U.S. for Douglas Fir, US$ per cubic meter |
| PLOGSK_USD | Num | 8 | Hard Logs, Best quality Malaysian meranti, import price Japan, US$ per cubic meter |
| PMAIZMT_USD | Num | 8 | Maize (corn), U.S. No.2 Yellow, FOB Gulf of Mexico, U.S. price, US$ per metric ton |
| PNGASEU_USD | Num | 8 | Natural Gas, Russian Natural Gas border price in Germany, US$ per thousands of cubic meters of gas |
| PNGASJP_USD | Num | 8 | Natural Gas, Indonesian Liquefied Natural Gas in Japan, US$ per cubic meter of liquid |
| PNGASUS_USD | Num | 8 | Natural Gas, Natural Gas spot price at the Henry Hub terminal in Louisiana, US$ per thousands of cubic meters of gas |
| PNICK_USD | Num | 8 | Nickel, melting grade, LME spot price, CIF European ports, US$ per metric ton |
| POILAPSP_USD | Num | 8 | Crude Oil (petroleum), Price index, 2005 = 100, simple average of three spot prices; Dated Brent, West Texas Intermediate, and the Dubai Fateh |
| PROIL_USD | Num | 8 | Rapeseed oil, crude, fob Rotterdam, US$ per metric ton |

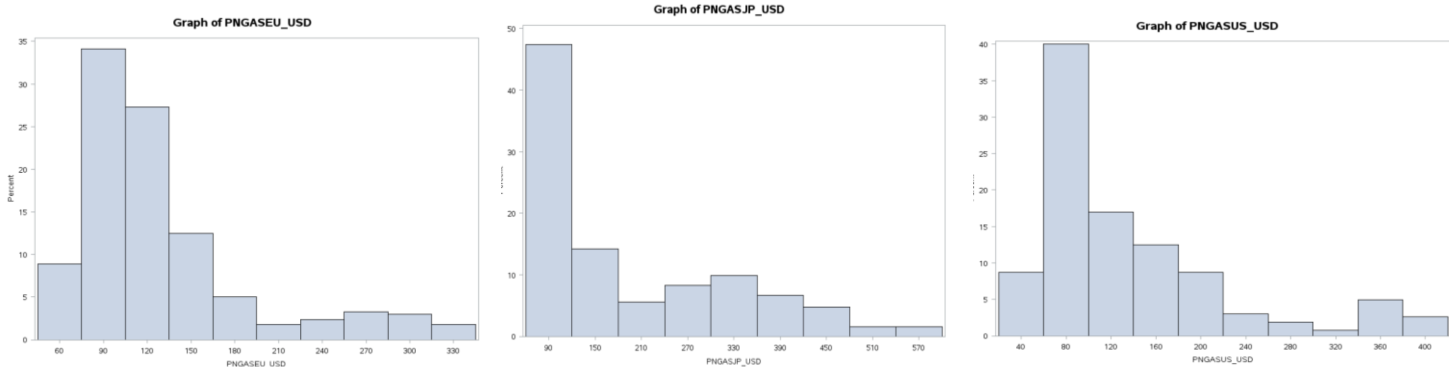
**Fig 1.2**

## 3.2 Imputation:

We perform imputation by assuming a multivariate normal distribution for the data. We use the Univariate procedure to get the summary statistics of each variable with missing values.

From the figure 1.3 below, we see the data is skewed on the right since the mean is greater than the median and we see the histogram is not symmetric for all the 3 variables (All-Natural gases) therefore we will use median to replace our missing values.

**Fig 1.3**



## 3.3 Standardization:

Before we do our analysis, it is important for all the variables to be of the same dimension. Therefore, we standardize the dataset to have a zero mean and unit variance.

## 3.4 Correlation:

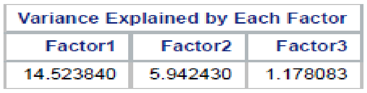
After performing correlation, we see that the variables are highly correlated with each other and therefore, factor analysis is a right approach.

## 3.5 Methods Adopted:

The factor analysis was performed with assigning variance as one to each variable. The number of factors was initially selected by retaining only those factors with an eigenvalue greater than one. Also, we perform the scree plot test to retain the number of factors. In my case, I see there are only three factors former performing both the test and we do further rotation to maximize the variance (variability) of the factor.

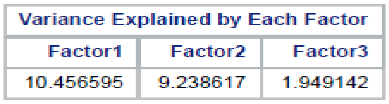
Among the various orthogonal (Varimax, factor parsimax) and oblique(Promax) methods tested, Factor parsimax rotation outperformed to the others due to better distribution of variance among the three factors. Below are the technical decisions relative to the factor analysis:

**No Rotation**: As we can see from the "Variance explained by each factor" table that factor 1 has a very large variance as compared to the other two and therefore we would need rotation.



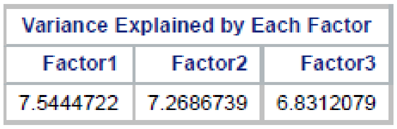
| Variance Explained by Each Factor | | |
|---|---|---|
| Factor1 | Factor2 | Factor3 |
| 14.523840 | 5.942430 | 1.178083 |

**Varimax Rotation:** From the factor loadings table in varimax rotation we see that there are 4 complex variables i.e. the 2 or more factors have very minor difference in their values. Also, we see that the variance is so less for the third factor which gives us only one variable in the third factor. Such classification where there's one variable for a factor is not considered a good factor analysis.

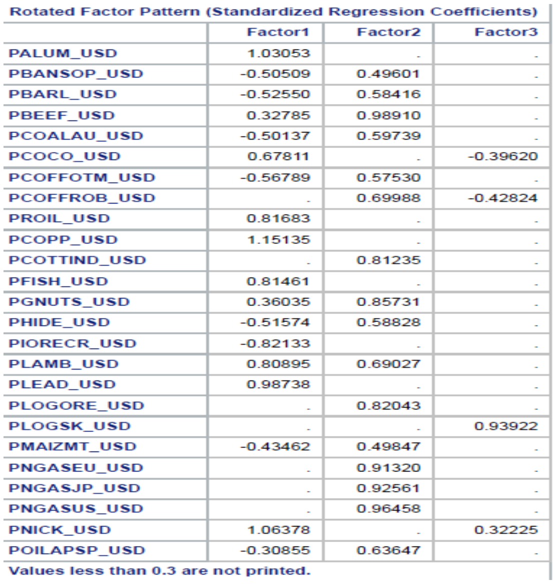


| Variance Explained by Each Factor | | |
|---|---|---|
| Factor1 | Factor2 | Factor3 |
| 10.456595 | 9.238617 | 1.949142 |

**Factor Parsimax Rotation:** This method gives the most appropriate factors with equal distribution of variance and variables among the factors as we can see from the figure below.



| Variance Explained by Each Factor | | |
|---|---|---|
| Factor1 | Factor2 | Factor3 |
| 7.5444722 | 7.2686739 | 6.8312079 |



| Rotated Factor Pattern (Standardized Regression Coefficients) | | | |
|---|---|---|---|
| | Factor1 | Factor2 | Factor3 |
| PALUM_USD | 1.03053 | . | . |
| PBANSOP_USD | -0.50509 | 0.49601 | . |
| PBARL_USD | -0.52550 | 0.58416 | . |
| PBEEF_USD | 0.32785 | 0.98910 | . |
| PCOALAU_USD | -0.50137 | 0.59739 | . |
| PCOCO_USD | 0.67811 | . | -0.39620 |
| PCOFFOTM_USD | -0.56789 | 0.57530 | . |
| PCOFFROB_USD | . | 0.69988 | -0.42824 |
| PROIL_USD | 0.81683 | . | . |
| PCOPP_USD | 1.15135 | . | . |
| PCOTTIND_USD | . | 0.81235 | . |
| PFISH_USD | 0.81461 | . | . |
| PGNUTS_USD | 0.36035 | 0.85731 | . |
| PHIDE_USD | -0.51574 | 0.58828 | . |
| PIORECR_USD | -0.82133 | . | . |
| PLAMB_USD | 0.80895 | 0.69027 | . |
| PLEAD_USD | 0.98738 | . | . |
| PLOGORE_USD | . | 0.82043 | . |
| PLOGSK_USD | . | . | 0.93922 |
| PMAIZMT_USD | -0.43462 | 0.49847 | . |
| PNGASEU_USD | . | 0.91320 | . |
| PNGASJP_USD | . | 0.92561 | . |
| PNGASUS_USD | . | 0.96458 | . |
| PNICK_USD | 1.06378 | . | 0.32225 |
| POILAPSP_USD | -0.30855 | 0.63647 | . |

Values less than 0.3 are not printed.

**Promax Rotation:** This is an Oblique rotation method which is used when the factors have high correlation. This rotation is not suitable because it is outputting loadings greater than 100% for some variables plus it also gives only one variable for the fatcor3 which doesn't gives a good clustering. Therefore, we will not move ahead with rotation.

# 4.PROGRAMS

```
libname project1 "/home/asodan20/Project 1";

filename dataset "/home/asodan20/Project 1/project1_data_29.csv";

/*Inputting the variables*/

data project1.commodities;

 infile dataset dlm=',' firstobs= 9;

 input   PALUM_USD          PBANSOP_USD       PBARL_USD PBEEF_USD  PCOALAU_USD
         PCOCO_USD  PCOFFOTM_USD    PCOFFROB_USD       PROIL_USD  PCOPP_USD
         PCOTTIND_USD    PFISH_USD     PGNUTS_USD          PHIDE_USD  PIORECR_USD
         PLAMB_USDPLEAD_USD  PLOGORE_USD     PLOGSK_USD          PMAIZMT_USD
         PNGASEU_USD       PNGASJP_USD PNGASUS_USD     PNICK_USD  POILAPSP_USD;

 run;

proc contents data = project1.commodities;

   title "Contents of the given file";

run;

proc means data = project1.commodities;

   title" Summary of the given file";

run;
```

**/*Deal with the missing values*/**

```
proc means data=project1.commodities n nmiss;
title "Variables with Missing values";
run;
* Variables PNGASEU_USD,PNGASJP_USD,PNGASUS_USD has missing values;

* For the variable PNGASEU_USD, we will do the following analysis;
title 'Graph of PNGASEU_USD';
ods graphics off;
proc univariate data=project1.commodities;
   histogram PNGASEU_USD;
   var PNGASEU_USD;
run;
* Since the graph is not symmetric we will use median
 instead of mean to replace the misisng values
 therefore replacing the missing values with 109.5500;

* For the variable PNGASJP_USD, we will do the following analysis;
title 'Graph of PNGASJP_USD ';
ods graphics off;
proc univariate data=project1.commodities;
   histogram PNGASJP_USD;
   var PNGASJP_USD;
run;
* Again the graph is not symmetric we will use median
 value  128.8000 for replacing the missing values;

* For the variable PNGASUS_USD, we will do the following analysis;
```

```
title 'Graph of PNGASUS_USD ';
ods graphics off;
proc univariate data=project1.commodities;
   histogram PNGASUS_USD;
   var PNGASUS_USD;
run;
* Again the graph is not symmetric we will use median
 value 101.4100 for replacing the missing values;
*Replacing all the missing values with medians;
data project1.NewCommodities;
  set project1.Commodities;
  if PNGASEU_USD = 'n.a.' then PNGASEU_USD = 109.5500;
  if PNGASJP_USD = 'n.a.' then PNGASJP_USD = 128.8000;
  if PNGASUS_USD = 'n.a.' then PNGASUS_USD = 101.4100;
run;
proc print data =project1.newcommodities;run;
/*Standardise data*/
proc standard data=project1.NewCommodities
 out=project1.standarddata mean=0 std=1;
 var _numeric_;
run;
proc print data = project1.standarddata;
* correlation between the variables;
proc corr data=project1.standarddata;
 run;
 * check results with/without priors";
title "checking results after priors";
proc factor data=project1.standarddata rotate=varimax scree;
```

```
priors smc;

run;


* Factor Analysis without rotation;

 ods graphics on;

 proc factor data = project1.standarddata

 plot = SCREE reorder;

   TITLE "Factor Analysis without rotation";

   var PALUM_USD   PBANSOP_USD       PBARL_USD PBEEF_USD  PCOALAU_USD
          PCOCO_USD  PCOFFOTM_USD   PCOFFROB_USD     PROIL_USD  PCOPP_USD
          PCOTTIND_USD  PFISH_USD      PGNUTS_USD        PHIDE_USD  PIORECR_USD
          PLAMB_USDPLEAD_USD  PLOGORE_USD      PLOGSK_USD        PMAIZMT_USD
          PNGASEU_USD       PNGASJP_US  PNGASUS_USD     PNICK_USD  POILAPSP_USD;

 RUN;

ods graphics off;

* As we can see that without rotation, the number of factors= 3 since the eigenvalues of

3 variables >1  and from the scree plot also we can see that.

Therefore, we will retain three factors;


* With Rotation ;

%let rotation = varimax;

ods graphics on;

PROC FACTOR DATA= project1.standarddata

          ROTATE=&rotation

          reorder;

TITLE "Factor Analysis with &rotation Rotation";

RUN;

ods graphics off;


* Running the Oblique rotation
```

```
*Promax rotation;

%let rotation1= promax;

ods graphics on;

PROC FACTOR DATA= project1.standarddata

            ROTATE=&rotation1  fuzz=0.3

             ;

   TITLE "Factor Analysis with &rotation1 Rotation";

RUN;

ods graphics off;

*Rotation = factorparsimax;

%let rotation2=Factorparsimax;

ods graphics on;

PROC FACTOR DATA= project1.standarddata

            Plot=SCREE

            ROTATE=&rotation2

            reorder

           ;

   TITLE "Factor Analysis with &rotation2 Rotation";

RUN;

ods graphics off;
```