2

*Project 2: FACTOR ANALYSIS REPORT*

Jasbir Singh

Madhu Sredhar

Oct 29, 2017

Table of Contents

[1 Problem Description: 1](#_Toc497137736)

[1.1 Why Factor Analysis? 1](#_Toc497137737)

[1.2 Assumptions for Factor Analysis: 1](#_Toc497137738)

[2 Technical Appendix 3](#_Toc497137739)

[2.1 Data Description: 3](#_Toc497137740)

[2.2 Data Pre-Processing & Cleaning: 4](#_Toc497137741)

[2.3 Factor Analysis: 5](#_Toc497137742)

[2.4 Conclusion: 10](#_Toc497137743)

[3 PROGRAMS 10](#_Toc497137744)

# Problem Description:

Analysis in this report is done on a commodity data to identify the patterns in the data & take a mass of data and shrinking it to a smaller data set that is more manageable and more understandable using Factor Analysis.

## Why Factor Analysis?

Factor Analysis is way to:

* Reduce the number of variables (shrinking it to a smaller data set that is more manageable and more understandable)
* Way to find hidden patterns, show how those patterns overlap and show what characteristics are seen in multiple patterns

There are two types of factor analysis:

* **Exploratory factor analysis** is if you don’t have any idea about what structure your data is or how many dimensions are in a set of variables.
* **Confirmatory Factor Analysis**is used for verification as long as you have a specific idea about what structure your data is or how many dimensions are in a set of variables.

For the given problem dataset, we will be using **Exploratory factor analysis** approach.

## Assumptions for Factor Analysis:

1. Adequate sample size: The case must be greater than the factor.
2. No perfect multicollinearity: Factor analysis is an interdependency technique.  There should not be perfect multicollinearity between the variables.
3. Homoscedasticity: Since factor analysis is a linear function of measured variables, it does not require homoscedasticity between the variables.
4. Linearity: Factor analysis is also based on linearity assumption.  Non-linear variables can also be used.  After transfer, however, it changes into linear variable.
5. Pare down to only numeric variables that are actually quantitative
6. Handle Missing value
7. Standardize remaining variables to mean zero and variance 1

In the given data set, we made sure that given dataset should abide by all these assumptions for Factor analysis, we will elaborate in throughout this report, what statistical procedure were used to achieve the required result.

**Factor1: Households**

Below are the commodities with their percentages that come under households. From the description we have identified that the import and export is evenly spread across households

|  |  |  |
| --- | --- | --- |
| **COMMODITIES** | **DESCRIPTION (HOUSEHOLDS)** | **%** |
| PSOYB\_USD | Soybeans, U.S. soybeans, Chicago Soybean futures contract (first contract forward) No. 2 yellow and par, US$ per metric ton | 93.862 |
| PTIN\_USD | Tin, standard grade, LME spot price, US$ per metric ton | 88.061 |
| PSUGAUSA\_USD | Sugar, U.S. import price, contract no.14 nearest futures position, US cents per pound (Footnote: No. 14 revised to No. 16) | 73.301 |
| PRUBB\_USD | Rubber, Singapore Commodity Exchange, No. 3 Rubber Smoked Sheets, 1st contract, US cents per pound | 93.666 |
| PSOIL\_USD | Soybean Oil, Chicago Soybean Oil Futures (first contract forward) exchange approved grades, US$ per metric ton | 93.598 |
| PSMEA\_USD | Soybean Meal, Chicago Soybean Meal Futures (first contract forward) Minimum 48 percent protein, US$ per metric ton | 87.664 |
| PSUGAISA\_USD | Sugar, Free Market, Coffee Sugar and Cocoa Exchange (CSCE) contract no.11 nearest future position, US cents per pound | 74.344 |
| PWOOLC\_USD | Wool, coarse, 23 micron, Australian Wool Exchange spot quote, US cents per kilogram | 89.258 |
| PPOIL\_USD | Palm oil, Malaysia Palm Oil Futures (first contract forward) 4-5 percent FFA, US$ per metric ton | 88.928 |
| PRICENPQ\_USD | Rice, 5 percent broken milled white rice, Thailand nominal price quote, US$ per metric ton | 84.22 |
| PSUNO\_USD | Sunflower oil, Sunflower Oil, US export price from Gulf of Mexico, US$ per metric ton | 85.218 |
| PWHEAMT\_USD | Wheat, No.1 Hard Red Winter, ordinary protein, FOB Gulf of Mexico, US$ per metric ton | 87.96 |
| PWOOLF\_USD | Wool, fine, 19 micron, Australian Wool Exchange spot quote, US cents per kilogram | 73.047 |
| PTEA\_USD | Tea, Mombasa, Kenya, Auction Price, US cents per kilogram, From July 1998,Kenya auctions, Best Pekoe Fannings. Prior, London auctions, c.i.f. U.K. warehouses | 69.844 |
| PSHRI\_USD | Shrimp, No.1 shell-on headless, 26-30 count per pound, Mexican origin, New York port, US cents per pound | 64.219 |

**Factor2: Poultry & Wood**

Below are the commodities with their percentages that come under Poultry & Wood. From the description we have identified that the export dominates imports in Poultry & Wood in global market.

|  |  |  |
| --- | --- | --- |
| **COMMODITIES** | **DESCRIPTION (Poultry & Wood)** | **%** |
| PSAWORE\_USD | Soft Sawnwood, average export price of Douglas Fir, U.S. Price, US$ per cubic meter | 96.27 |
| PSAWMAL\_USD | Hard Sawnwood, Dark Red Meranti, select and better quality, C&F U.K port, US$ per cubic meter | 82.594 |
| PPOULT\_USD | Poultry (chicken), Whole bird spot price, Ready-to-cook, whole, iced, Georgia docks, US cents per pound | 81.591 |
| POLVOIL\_USD | Poultry (chicken), Whole bird spot price, Ready-to-cook, whole, iced, Georgia docks, US cents per pound | 77.028 |
| PPORK\_USD | Olive Oil, extra virgin less than 1% free fatty acid, ex-tanker price U.K., US$ per metric ton | -73.651 |

**Factor3: Sugar and Metals**

Below are the commodities with their percentages that come under Sugar and Metals. From the description we have identified that imports hold the majority in Sugar and Metals in global market.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **COMMODITIES** | **DESCRIPTION (Sugar and Metals)** | | | **%** |
| PSUGAEEC\_USD | Sugar, European import price, CIF Europe, US cents per pound | | | 85.838 |
| PURAN\_USD | Uranium, NUEXCO, Restricted Price, Nuexco exchange spot, US$ per pound | | | 79.78 |
| PZINC\_USD | Zinc, high grade 98% pure, US$ per metric ton | | | 78.527 |
|  | |  |

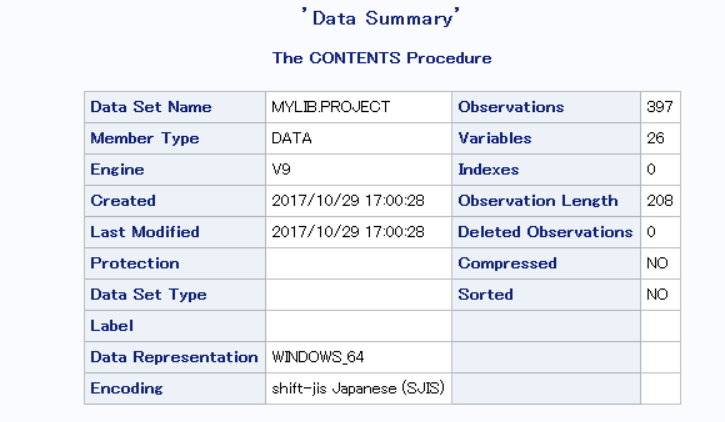
# Technical Appendix

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 observations that were collected from the global commodity market prices.

## Data Description:

Dataset has 26 Quantitative Variables and 397 observations in total

Variable “PSUGAEEC\_USD” has 137 rows of data with “n.a”.



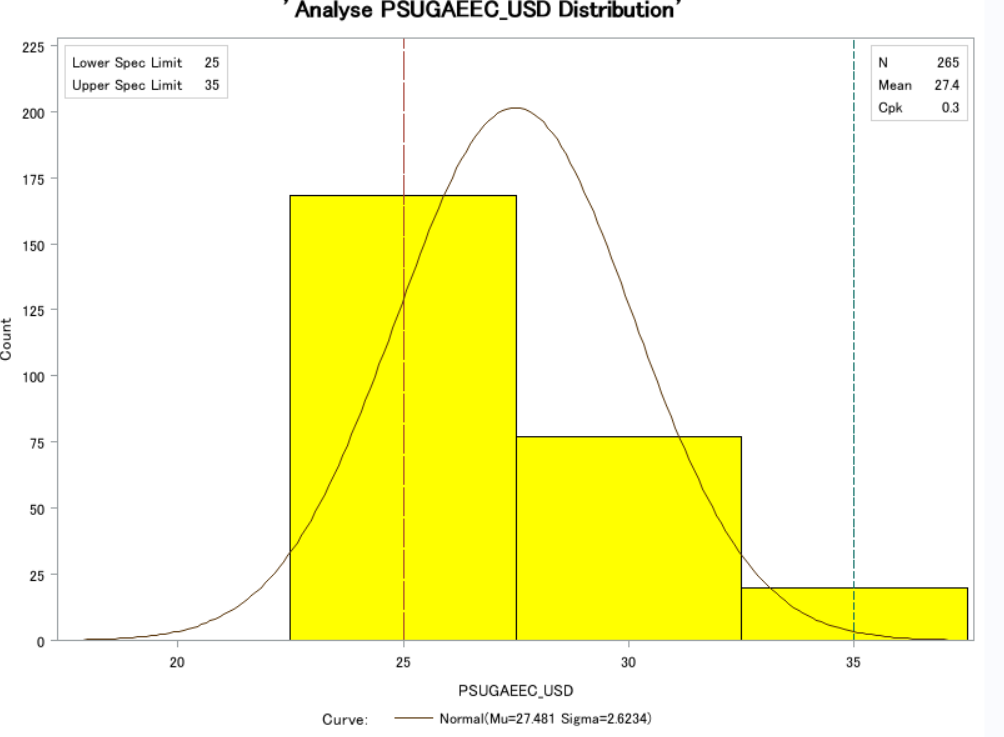
## Data Pre-Processing & Cleaning:

* To handle 137 rows of “N.A” values in the given dataset for Variable “PSUGAEEC\_USD”, I replace them all by median of the rest of the values.

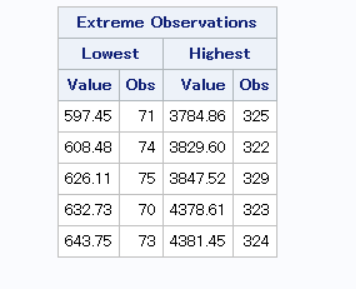
Why Median Not Mean?

If you check distribution of the variable PSUGAEEC\_USD it is very much Right skewed (as shown in the figure), so Mean is not a best representation of the data as compared to Median.

Replace all 137 rows for variable PSUGAEEC\_USD with median value = 26.83



* Standardize remaining variables to mean zero and variance 1 as per one of factor analysis assumptions.
* After analyzing correlation matrix for all the variables, it is very clear that POILAPSP\_Index is very significantly correlated with all the variables & it has very significant outliers as shown below from univariate analysis.



## Factor Analysis:

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.

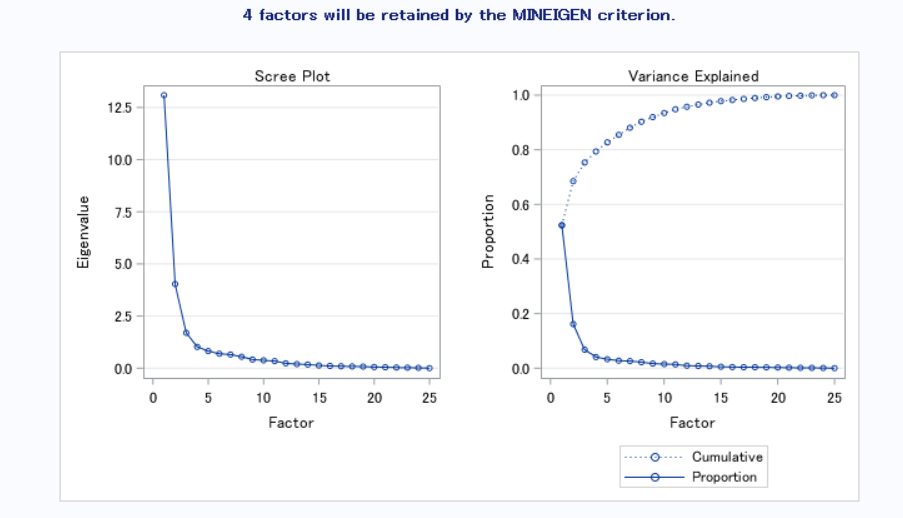
Among the various methods –Promax , Equamax & Varimax tested

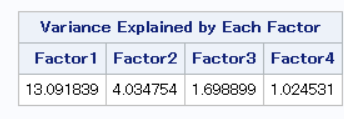
Promax rotation outperformed to the others due to greater Factor Structure (Correlations), variations explained by all the factors for all variables.

And by analyzing Path diagrams and input patterns for Promax rotation, it clearly defines all the factors that can be used represent almost all the variables in the dataset.

Below are the technical decisions relative to the factor analysis:

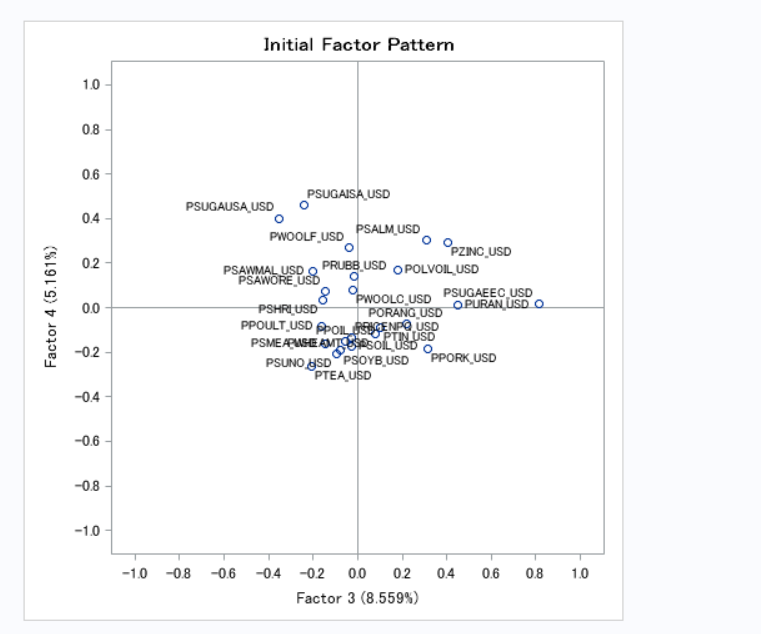
1. **Without Rotation:** Give Four Factor variables, all of them has eigen value more than 1, as visible from scree plot.



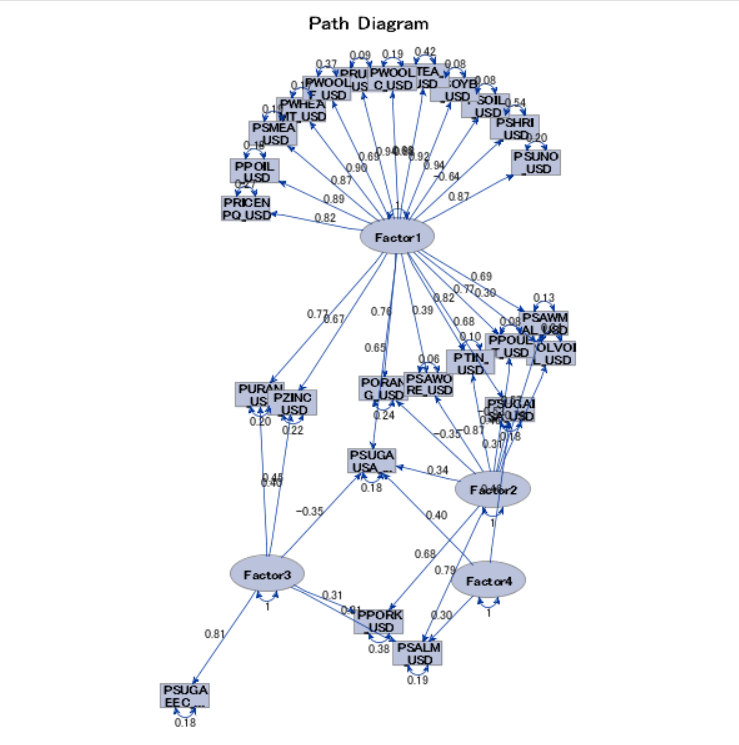


But when we refer to Initial factor pattern we are getting following plot , which clearly signifies that there is not much difference in variation explained by each for all the variables , even though they both have eigen values greater than 1.

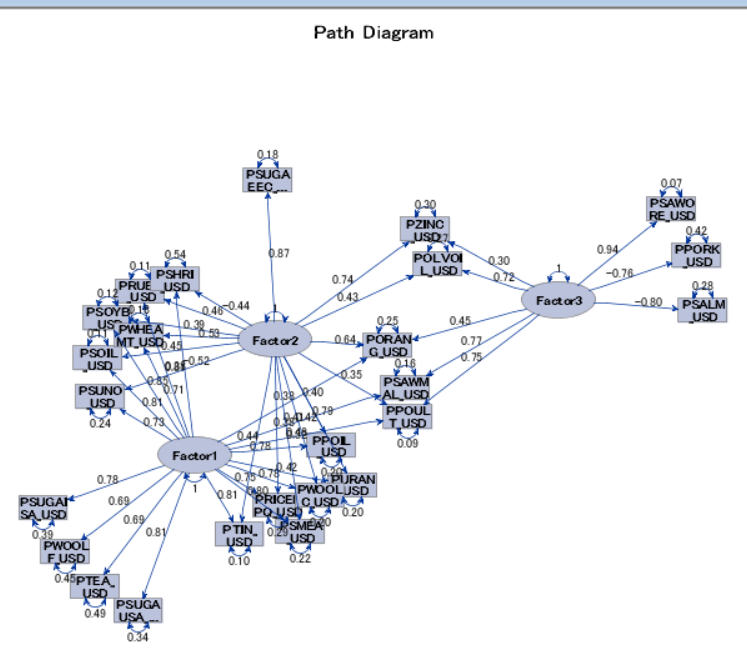
Same thing can be inferred from scree plot, shown above but not that clearly , factor 3 and 4 does not have very significant difference in eigen values.



It gets even more evident from Path diagrams as shown below: Factor 4 is explaining variance only that variables that already accounted in other factors.

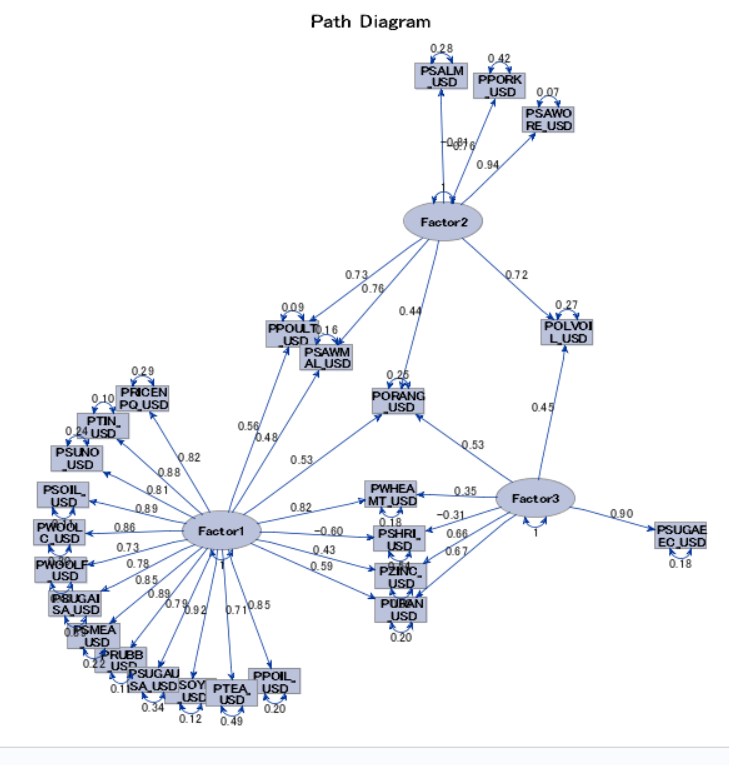


1. **With Equamax Rotation**: Variation explained by 3 Factors is significantly better than unrotated factor analysis & we can compare Factor pattern (Correlations) with rest of the rotation methods. (Refer Table 1.0 on next page). Path Diagram clearly defines factors and related variables: It seems Factor 1 and Factor 2 tries to combine same variables .



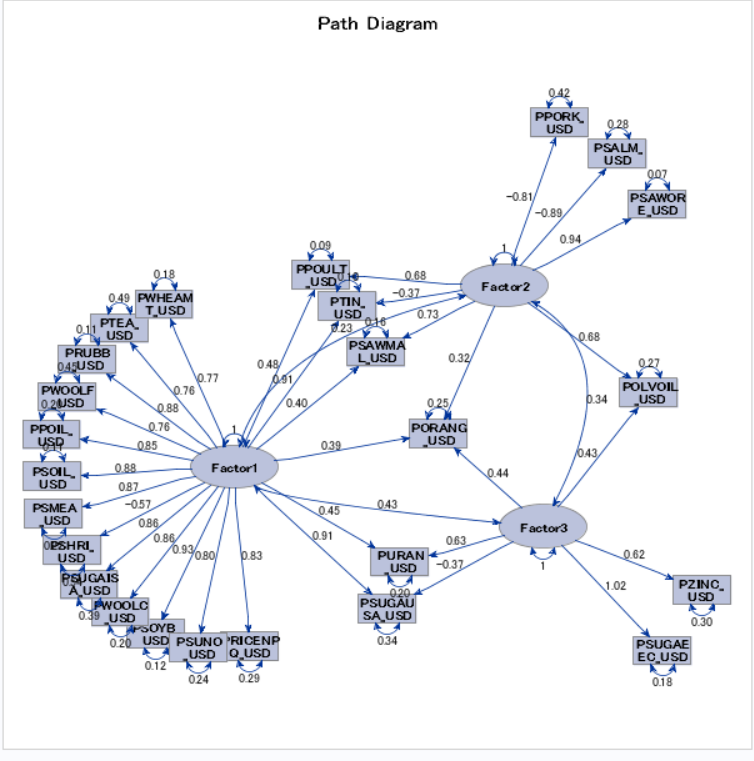
1. **With Varimax Rotation**: Variation explained by 3 Factors is significantly better than Equamax factor analysis & we can compare Factor Structure (Correlations) with rest of the rotation method. (Refer Table 1.0 on next page).

Varimax has best possible split of path diagram when compared to rest of the rotations.

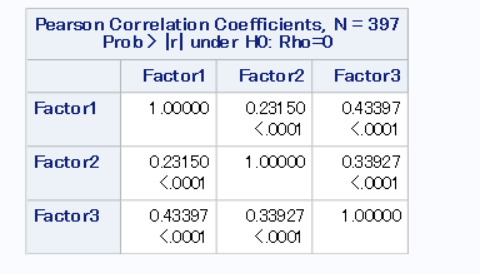


1. **With Promax Rotation**: Variation explained by 3 Factors is significantly better than all of above mentioned rotations & we can compare Factor Structure (Correlations) with rest of the rotation method. (Refer Table 1.0 on next page)

But Promax mostly used when there is significant correlation within the factors even it’s path diagram is almost same as Varimax rotation.



Correlation Factor1, Factor2, Factor3: There is significant correlation between the factors, so promax rotation is considered.



|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Equamax** | **Varimax** | **Promax** |
| Variance |  |  |  |
| Factor Pattern |  |  |  |

**Table 1.0**

## Conclusion:

After Analyzing

1. Variance explained by each factor
2. Factor Structure (Correlations)
3. Path Diagrams
4. Factors pattern

Promax rotation approach gives best factors that represents maximum variation in all the variables given in the data set.

# PROGRAMS

filename proj 'C:\1\_Jasbir\UIC\_Local\Spring 2017\SAS\SAS\_Lab\Project\project1\_data\_10.csv';

libname mylib 'C:\1\_Jasbir\UIC\_Local\Spring 2017\SAS\SAS\_Lab\Project';

**data** mylib.project;

infile proj dlm=',' missover dsd FIRSTOBS=**9**;

input POILAPSP\_index POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

IF PSUGAEEC\_USD="n.a." then PSUGAEEC\_USD=**.**; /\*Replacing n.a with . SAS understand missing vallues\*/

**run**;

**PROC** **CONTENTS** data= mylib.project;

**run**;

title 'Schematic Box Plot POILAPSP\_index';

**proc** **univariate** data=mylib.project robustscale plot; /\*Checking for Outliers\*/

var POILAPSP\_index;

**run**;

ods graphics on;

title 'Schematic Box Plot';

**proc** **univariate** data=mylib.project ; /\*Checking for Outliers\*/

hist POILAPSP\_index;

**run**;

ods graphics off;

/\*Checking for distribution of the values , to replace with mean or median\*/

title 'Analyse PSUGAEEC\_USD Distribution';

ods graphics off;

**proc** **capability** data=mylib.project noprint;

spec lsl=**25** llsl=**5** usl=**35** lusl=**3**;

histogram PSUGAEEC\_USD / normal( noprint )

midpoints = **20** to **35** by **5**

vscale = Count

cfill = yellow

nospeclegend ;

inset lsl usl / cfill=blank;

inset n mean (**25.2**) cpk (**25.2**) / cfill=blank;

**run**;

title 'Statistical Analysis';

**PROC** **MEANS** MEAN STD mode MEDIAN KURTOSIS SKEWNESS MEDIAN Q1 Q3 data= mylib.project ;

**RUN**;

/\*Replace all missing values with median and drop variable POILAPSP\_Index8u\*/

title 'Dealing with Missing Values';

**data** mylib.project\_final;

set mylib.project;

IF PSUGAEEC\_USD=**.** then PSUGAEEC\_USD=**26.83**;

drop POILAPSP\_Index;

**run**;

title 'Data Summary';

**PROC** **CONTENTS** data= mylib.project;

**run**;

/\*Correlation of all the variables in the dataset , to check if eny varaible is highly correlated \*/

title 'Variables Correlation Matrix';

**Proc** **CORR** data=mylib.project Rank out=mylib.Project\_corr ;

**run**;

title 'Standardized data for Factor analysis';

**proc** **standard** data=mylib.project\_final mean=**0** std=**1**

out=mylib.project\_Std;

**run**;

title 'Factor analysis';

ods graphics on;

**proc** **factor** data=mylib.project\_final

plots=scree;

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

**run**;

ods graphics off;

title 'Factor analysis with plots';

ods graphics on;

**proc** **factor** data=mylib.project\_final

plots=(scree initloadings pathdiagram);

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

**run**;

ods graphics off;

ods graphics on;

**proc** **factor** data=mylib.project\_final

plots=all;

**run**;

ods graphics off;

title 'Varimax Rotation';

ods graphics on;

**proc** **factor** nfactor=**3** data=mylib.project\_final out=mylib.project\_Factored rotate=varimax reorder

plots=(scree pathdiagram initloadings preloadings loadings);

**run**;

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

ods graphics off;

title 'Check for Corelation between Factors';

ods graphics on;

**proc** **corr** data=mylib.project\_Factored;

var Factor1 Factor2 Factor3 ;

**run**;

ods graphics off;

title 'Promax Rotation';

ods graphics on;

**proc** **factor** nfactor=**3** data=mylib.project\_final out=mylib.project\_Factored rotate=promax

plots=(scree pathdiagram initloadings preloadings loadings);

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

**run**;

ods graphics off;

title 'Equamax Rotation';

ods graphics on;

**proc** **factor** nfactor=**3** data=mylib.project\_final out=mylib.project\_Factored rotate=equamax

plots=(scree pathdiagram initloadings preloadings loadings);

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

**run**;

ods graphics off;

title 'Correlation between Factors After equamax';

ods graphics on;

**proc** **corr** data=mylib.project\_Factored;

var Factor1 Factor2 Factor3 ;

**run**;

ods graphics off;

title 'Equamax Plots';

**proc** **factor** data=mylib.project\_final

priors=smc msa residual

rotate=equamax

outstat=fact\_all

plots=(scree initloadings preloadings loadings);

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

**run**;

ods graphics off;

ods graphics on;

title 'Promax and all plots with reorder';

**proc** **factor** nfactors=**3** data=mylib.project\_final

priors=smc msa residual

rotate=promax reorder

outstat=fact\_all

plots=(scree initloadings preloadings loadings);

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

**run**;

ods graphics off;

ods graphics on;

title 'Varimax and all plots with reorder';

**proc** **factor** nfactors=**3** data=mylib.project\_final

priors=smc msa residual

rotate=varimax reorder

outstat=fact\_all

plots=(scree initloadings preloadings loadings preloadings(vector));

var POLVOIL\_USD PORANG\_USD PPOIL\_USD PPORK\_USD PPOULT\_USD PRICENPQ\_USD PRUBB\_USD PSALM\_USD

PSAWMAL\_USD PSAWORE\_USD PSHRI\_USD PSMEA\_USD PSOIL\_USD PSOYB\_USD PSUGAEEC\_USD PSUGAISA\_USD PSUGAUSA\_USD PSUNO\_USD

PTEA\_USD PTIN\_USD PURAN\_USD PWHEAMT\_USD PWOOLC\_USD PWOOLF\_USD PZINC\_USD ;

**run**;

ods graphics off;

ods graphics on;

**proc** **factor** data=mylib.project\_final

priors=smc msa residual

outstat=fact\_all

plots=all;

**run**;

ods graphics off;

%Let rotation=Promax; %Let nfactors=3; %let a=POLVOIL\_USD; %Let b=PORANG\_USD ; %Let c=PPOIL\_USD;

%Let d=PPORK\_USD;%Let e=PPOULT\_USD ;%Let f=PRICENPQ\_USD;%Let g=PRUBB\_USD;%Let h=PSALM\_USD; %Let i=PSAWMAL\_USD ;

%Let j=PSAWORE\_USD;%Let k=PSHRI\_USD;%Let l=PSMEA\_USD;%Let m=PSOIL\_USD;%Let n=PSOYB\_USD;%Let o=PSUGAEEC\_USD;%Let p=PSUGAISA\_USD;

%Let q=PSUGAUSA\_USD;%Let r=PSUNO\_USD;%Let s=PTEA\_USD;%Let t=PTIN\_USD;%Let u=PURAN\_USD;%Let v=PWHEAMT\_USD;

%Let x=PWOOLC\_USD; %Let y=PWOOLF\_USD; %Let z=PZINC\_USD ;

title &rotation 'and all plots with reorder';

**proc** **factor** nfactors=&nfactors data=mylib.project\_final

priors=smc msa residual

rotate=&rotation reorder

outstat=fact\_all

plots=(scree initloadings preloadings loadings);

var &a &b &c &d &e &f &g &h &i &j &k &l &m &n &o &p &q &r

&s &t &u &v &x &y &z;

**run**;

ods graphics off;

/\*Creating Macro for Factor\_Analysis\*/

**%macro** factor\_Analysis(rotation=Promax, nfactors=**3**, a=POLVOIL\_USD, b=PORANG\_USD , c=PPOIL\_USD,

d=PPORK\_USD,e=PPOULT\_USD ,f=PRICENPQ\_USD,g=PRUBB\_USD,h=PSALM\_USD, i=PSAWMAL\_USD ,

j=PSAWORE\_USD,k=PSHRI\_USD,l=PSMEA\_USD,m=PSOIL\_USD,n=PSOYB\_USD,o=PSUGAEEC\_USD,p=PSUGAISA\_USD,

q=PSUGAUSA\_USD,r=PSUNO\_USD,s=PTEA\_USD,t=PTIN\_USD,u=PURAN\_USD,v=PWHEAMT\_USD,

x=PWOOLC\_USD, y=PWOOLF\_USD, z=PZINC\_USD );

proc factor nfactors=&nfactors data=mylib.project\_final

priors=smc msa residual

rotate=&rotation reorder

outstat=fact\_all

plots=(scree initloadings preloadings loadings);

var &a &b &c &d &e &f &g &h &i &j &k &l &m &n &o &p &q &r

&s &t &u &v &x &y &z;

run;

**%mend** factor\_Analysis;

%***factor\_Analysis***;