**A**

**PROJECT REPORT**

**PS04CSTA24**

**ON**

**“STATISTICAL ANALYSIS OF RICE PRODUCTION”**

**BY**

**GANESH DHANAWADE**

**VARSHA SHINDE**

**KIRTI SAWANT**

**SUPRIYA MANE**

**SHAILAJA SHINDE**

**PROJECT GUIDE**

**DR. MILIND B. BHATT**

**“MASTER OF STATISTICS”**

**DEPARTMENT OF STATISTICS**

**SARDAR PATEL UNIVERSITY**

**VALLBH-VIDYANAGAR**

**2020-2021**

**CERTIFICATE**

This is to certify that **MR. GANESH MAHADEV DHANAWADE** Student of Master of Science in **Statistics**, Roll No. 07 has satisfactory complete his Project work on **“STATISTICAL ANALYSIS OF RICE PRODUCTION”** for M.Sc. Statistics semester IV during the Oct2020-May2021.

Date: 10/05/2021

(Dr. Milind B. Bhatt) (Dr. Jyoti M. Divecha)

Project Guide Head and professor

**CERTIFICATE**

This is to certify that **MISS. VARSHA HANMANT SHINDE** Student of Master of Science in **Statistics**; Roll No. 35 has satisfactory complete his Project work on **“STATISTICAL ANALYSIS OF RICE PRODUCTION”** for M.Sc. Statistics semester IV during the Oct2020-May2021.

Date: 10/05/2021

(Dr. Milind B. Bhatt) (Dr. Jyoti M. Divecha)

Project Guide Head and professor

**CERTIFICATE**

This is to certify that **MISS. KIRTI ASHOK SAWANT** Student of Master of Science in **Statistics**, Roll No. 12 has satisfactory complete his Project work on **“STATISTICAL ANALYSIS OF RICE PRODUCTION”** for M.Sc. Statistics semester IV during the Oct2020-May2021.

Date: 10/05/2021

(Dr. Milind B. Bhatt) (Dr. Jyoti M. Divecha)

Project Guide Head and professor

**CERTIFICATE**

This is to certify that **MISS. SUPRIYA UTTAM MANE** Student of Master of Science in **Statistics**; Roll No. 34 has satisfactory complete his Project work on **“STATISTICAL ANALYSIS OF RICE PRODUCTION”** for M.Sc. Statistics semester IV during the Oct2020-May2021.

Date: 10/05/2021

(Dr. Milind B. Bhatt) (Dr. Jyoti M. Divecha)

Project Guide Head and professor

**CERTIFICATE**

This is to certify that **MISS. SHAILAJA GAJANAN SHINDE** Student of Master of Science in **Statistics**, Roll No. 31s has satisfactory complete his Project work on **“STATISTICAL ANALYSIS OF RICE PRODUCTION”** for M.Sc. Statistics semester IV during the Oct2020-May2021.

Date: 10/05/2021

(Dr. Milind B. Bhatt) (Dr. Jyoti M. Divecha)

Project Guide Head and professor

**CERTIFICATE**

This is to certify that **Mr. Ganesh Mahadev Dhanawade, Miss. Varsha Hanmant Shinde, Miss. Kirti Ashok Sawant, Miss. Supriya Uttam Mane & Miss. Shailaja Gajanan Shinde** Student of Master of Science in **Statistics**, Roll No. 07, 35, 12, 34, 31 respectively has satisfactory complete his Project work on **“STATISTICAL ANALYSIS OF RICE PRODUCTION”** for M.Sc. Statistics semester IV during the Oct2020-May2021.

Date: 10/05/2021

(Dr. Milind B. Bhatt) (Dr. Jyoti M. Divecha)

Project Guide Head and professor

**PROJECT CHARTER**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Project Name**: Statistical Analysis of Rice Production | | | | | | |
| **Project Guide**: M.B. Bhatt sir & Aniket Salunkhe | | | | | | |
| **Student Name**: Ganesh Mahadev Dhanawade | | | | | | |
|  | | | | | | |
| **Sr. No.** | **Month** | **Week** | **Hours Spent** | **No. of meetings** | **Work** |
| 1 | Jan | 1(External exams) | 2 hours | - | Select the Project topic |
| 2 | 3 hours | - | Discuss Project topic |
| 3 | 2 hours daily | 1 | Study Project topic |
| 4 | 2 hours daily | 1 | Study information related to project topic |
| 5 | 6 hours | 2 | Learn Basic of Python |
| 2 | Feb | 1 | 5 hours | 2 | Data understanding using python |
| 2 | 2 hours daily | 1 | Data Visualization using python |
| 3 | 6 hours | - | Check Normality of data |
| 4 | 2 hours daily | 1 | Study Non-parametric test |
| 3 | Mar | 1 | 5 hours | - | Graphical technique |
| 2 | 2 hours daily | 1 | Study Category wise data |
| 3(Internal exams) | 0 | - | No work |
| 4 | 3 hours | 2 | To check Correlation |
| 5 | 4 hours daily | 1 | Analyse PCA |
| 4 | Apr | 1 | 3 hours daily | - | Analyse factor analysis |
| 2 | 3 hours daily | - | Interpretation |
| 3 | 4 hours daily | - | Conclusion |
| 4 | 5 hours daily | 1 | Create word file |
| 5 | 5 hours daily | 2 | Project typing |
| 5 | May | 1 | 4 hours daily | - | Final Submission |

**1.**

**PROJECT CHARTER**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Project Name**: Statistical Analysis of Rice Production | | | | | |
| **Project Guide**: M.B. Bhatt sir & Aniket Salunkhe | | | | | |
| **Student Name**: Varsha Hanmant Shinde | | | | | |
|  | | | | | |
| **Sr. No.** | **Month** | **Week** | **Hours Spent** | **No. of**  **meetings** | **Work** |
| 1 | Jan | 1(External exams) | 1 hours | - | Select the Project topic |
| 2 | 3 hours | - | Discuss Project topic |
| 3 | 2 hours | 1 | Study Project topic |
| 4 | 5 hours | 1 | Study information related to project topic |
| 5 | 6 hours | 2 | Learn Basic of Python |
| 2 | Feb | 1 | 5 hours | 2 | Data understanding using python |
| 2 | 3 hours | 1 | Data Visualization using python |
| 3 | 6 hours | - | Check Normality of data |
| 4 | 2 hours | 1 | Study Non-parametric test |
| 3 | Mar | 1 | 5 hours | - | Graphical technique |
| 2 | 3 hours | 1 | Study Category wise data |
| 3(Internal exams) | 0 | - | No work |
| 4 | 2 hours daily | 2 | To check Correlation |
| 5 | 3 hours daily | 1 | Analyse PCA |
| 4 | Apr | 1 | 2 hours daily | - | Analyse factor analysis |
| 2 | 3 hours daily | - | Interpretation |
| 3 | 4 hours daily | - | Conclusion |
| 4 | 5 hours daily | 1 | Create word file |
| 5 | 5 hours daily | 2 | Project typing |
| 5 | May | 1 | 5 hours daily | - | Final Submission |

**2.**

**PROJECT CHARTER**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Project Name**: Statistical Analysis of Rice Production | | | | | |
| **Project Guide**: M.B. Bhatt sir & Aniket Salunkhe | | | | | |
| **Student Name**: Kirti Ashok Sawant | | | | | |
|  | | | | | |
| **Sr. No.** | **Month** | **Week** | **Hours Spent** | **No. of**  **meetings** | **Work** |
| 1 | Jan | 1(External exams) | 0 | - | Select the Project topic |
| 2 | 2 hours | - | Discuss Project topic |
| 3 | 3 hours | 1 | Study Project topic |
| 4 | 2 hours | 1 | Study information related to project topic |
| 5 | 2 hours | 2 | Learn Basic of Python |
| 2 | Feb | 1 | 1 hours | 2 | Data understanding using python |
| 2 | 3 hours | 1 | Data Visualization using python |
| 3 | 3 hours | - | Check Normality of data |
| 4 | 2 hours | 1 | Study Non-parametric test |
| 3 | Mar | 1 | 2 hours | - | Graphical technique |
| 2 | 1 hours | 1 | Study Category wise data |
| 3(Internal exams) | 0 | - | No work |
| 4 | 1 hours | 2 | To check Correlation |
| 5 | 3 hours | 1 | Analyse PCA |
| 4 | Apr | 1 | 2 hours | - | Analyse factor analysis |
| 2 | 2 hours | - | Interpretation |
| 3 | 4 hours | - | Conclusion |
| 4 | 0 hours | 1 | Create word file |
| 5 | 3 hours | 2 | Project typing |
| 5 | May | 1 | 3 hours | - | Final Submission |

**3.**

**PROJECT CHARTER**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Project Name**: Statistical Analysis of Rice Production | | | | | |
| **Project Guide**: M.B. Bhatt sir & Aniket Salunkhe | | | | | |
| **Student Name**: Supriya Uttam Mane | | | | | |
|  | | | | | |
| **Sr. No.** | **Month** | **Week** | **Hours Spent** | **No. of**  **meetings** | **Work** |
| 1 | Jan | 1(External exams) | Half hours | - | Select the Project topic |
| 2 | 2 hours | - | Discuss Project topic |
| 3 | 3 hours | 1 | Study Project topic |
| 4 | 2 hours | 1 | Study information related to project topic |
| 5 | 2 hours | - | Learn Basic of Python |
| 2 | Feb | 1 | 5 hours | 2 | Data understanding using python |
| 2 | 3 hours | 1 | Data Visualization using python |
| 3 | 2 hours | - | Check Normality of data |
| 4 | 4 hours | 1 | Study Non-parametric test |
| 3 | Mar | 1 | 6 hours | - | Graphical technique |
| 2 | 2 hours | 1 | Study Category wise data |
| 3(Internal exams) | 0 | - | No work |
| 4 | 3 hours | 2 | To check Correlation |
| 5 | 4 hours | 1 | Analyse PCA |
| 4 | Apr | 1 | 4 hours | - | Analyse factor analysis |
| 2 | 3 hours | - | Interpretation |
| 3 | 4 hours | - | Conclusion |
| 4 | 5 hours | 1 | Create word file |
| 5 | 3 hours | 2 | Project typing |
| 5 | May | 1 | 5 hours | - | Final Submission |

**4.**

**PROJECT CHARTER**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Project Name**: Statistical Analysis of Rice Production | | | | | |
| **Project Guide**: M.B. Bhatt sir & Aniket Salunkhe | | | | | |
| **Student Name**: Shailaja Gajanan Shinde | | | | | |
|  | | | | | |
| **Sr. No.** | **Month** | **Week** | **Hours Spent** | **No. of**  **meetings** | **Work** |
| 1 | Jan | 1(External exams) | 0 | - | Select the Project topic |
| 2 | 1 hours | - | Discuss Project topic |
| 3 | 4 hours | 1 | Study Project topic |
| 4 | 2 hours | 1 | Study information related to project topic |
| 5 | 2 hours | - | Learn Basic of Python |
| 2 | Feb | 1 | 3 hours | 2 | Data understanding using python |
| 2 | 0 hours | 1 | Data Visualization using python |
| 3 | 4 hours | - | Check Normality of data |
| 4 | 2 hours | 1 | Study Non-parametric test |
| 3 | Mar | 1 | 5 hours | - | Graphical technique |
| 2 | 6 hours | 1 | Study Category wise data |
| 3(Internal exams) | 0 | - | No work |
| 4 | 3 hours | 2 | To check Correlation |
| 5 | 6 hours | 1 | Analyse PCA |
| 4 | Apr | 1 | 4 hours | - | Analyse factor analysis |
| 2 | 3 hours | - | Interpretation |
| 3 | 5 hours | - | Conclusion |
| 4 | 3 hours | 1 | Create word file |
| 5 | 2 hours | 2 | Project typing |
| 5 | May | 1 | 3 hours | - | Final Submission |

**5.**

**PREFACE**

It is great opportunity for me to have the MASTER OF STATISTICS in SARDAR PATEL UNIVERSITY, ANAND. In the accomplishment of this degree, I am submitting a project report on “STATISTICAL ANALYSIS OF RICE PRODUCTION”. Subject to the limitation of time efforts and resources, every possible attempt has been made to study the problem deeply. The whole project is measured through the secondary data, the further analysed and interpreted and the result was obtained.

This project is presented in simple and lucid language. We would fill amply rewarded if the project would prove to be beneficial to anyone who studies it.

**ACKNOWLEDGEMENT**

This year has been an extremely informative journey for, my friend and me. I would like to extend our gratitude to **Prof. Dr. Jyoti M. Divecha** (Head of Department M.Sc. (Statistics)) for entrusting upon me these invaluable projects. The journey of the study at the department and the projects gave me immense insight into the world of analytics. I am very thankful to **Dr. Milind B. Bhatt, Dr. P.A. Patel and** **Mr. Aniket Salunkhe**. my internal guide for their incomparable affection during my project works. Documentation is heart of project, so I take opportunity to express my heartfelt thanks to all my dear friends who support and encourage my project partners and me to complete out documentation successfully. These projects have been the outcome of ideas of combination of ideas suggestions and contribution of many people. My project is dedicated to all the people whom I met, took guidance interviewed and something from them. At this occasion, I want to grab this opportunity to acknowledge our sincere thanks to all of them while submitting.

Mr. Ganesh Mahadev Dhanawade. Roll No. 07

Miss. Varsha Hanmant Shinde. Roll No. 35

Miss. Kirti Ashok Sawant. Roll No. 12

Miss. Supriya Uttam Mane. Roll No. 34

Miss. Shailaja Gajanan Shinde. Roll No. 31

Place: Anand

Date: 10/05/2021

**Project**

**on**

**“STATISTICAL ANALYSIS OF RICE PRODUCTION”**



**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Contents** | **Page No.** |
| 1. | Introduction | 17 |
| 2. | Scope of Work | 18 |
| 3. | Study Data | 19 |
| 4. | Information about Parameters | 20 |
| 5.  6.  7. | Analysis   * 1. Importing and understanding the data.   2. Graphical Representation   3. Descriptive Statistics   5.4 Shapiro-Wilk and Kolmogorov Smirnov Test  5.5 Kruskal-Wallis test  5.6 Post hoc test  5.7 Correlation Matrix / Heat map  5.8 Principal Component Regression Analysis  5.9 Factor Analysis  Conclusion  Reference | 21 to 23  24 to 28  29 to 30  31 to 34  35 to 38  39 to 41  42 to 44  45 to 55  56 to 69  70 to 73  74 |

**1.Introduction**

India is one of leading countries in the world in production of a number of crops including Rice. Rice is one of the world's most important staple food products. This statement is particularly applicable to the Asian continent where rice forms the main staple food for the majority of the population (in particular the poorer segments of society) and where farmers account for more than 90 percent of the world's total rice production.

Rice Production in Indonesia is an important part of the national economy. Indonesia is the third largest Producer of rice in the world Leading rice producers, with Paddy production in 2003 of more than 50 million tones and cultivated Area of more than 11.5 million. Since 1980, Indonesia’s National rice yield has been Highest in tropical Asia.

Recent developments in the rice sector, Indonesia is the world’s third-largest rice producer and also one of the world’s biggest rice consumers. The country’s rice area expanded from 11.4 million ha in 1995 to 13.2 million ha in 2010, which represented 24% of the total agricultural area. Rice yield increased slightly from 4.3 t/ha in 1995 to 5 t/ha in 2010. Rice is the most important food crop in the country.

Agriculture is one of the key sectors within the Indonesian economy. In the last 50 years, the sector’s share in national gross domestic product has decreased considerable, due to the rise of industrialisation and service sector. Agriculture in Indonesia started as a means to grow and provide food.

It is consisted of 19 predict variables and 1 response variable. The variables are Id, Size, Status, Varieties, Bimas, Seed, Urea, Phosphate, Pesticide, Pseed, Purea, Pphosph, Hiredlabor, Famlabor, Wage, Goutput, Noutput, Price, Region.

After randomly selecting 1026 observation and 20 variables were taken 7 significant variables to fit a generalize linear model to predict the minimum error in variables to fit a generalize linear model to predict the minimum error in variables.

**2.Scope of Work**

1. We Importing and understanding the data.

2. Data Visualization

3. Descriptive Statistics

4. To check normality of data we use Shapiro-Wilk Test and Kolmogorov-Smirnov Test.

5. To check average of Gross output corresponding to Status\_land, Varieties, Bimas & region are same or different using the Kruskal-Wallis test

6. To check which pair differ significantly using Post-hoc test.

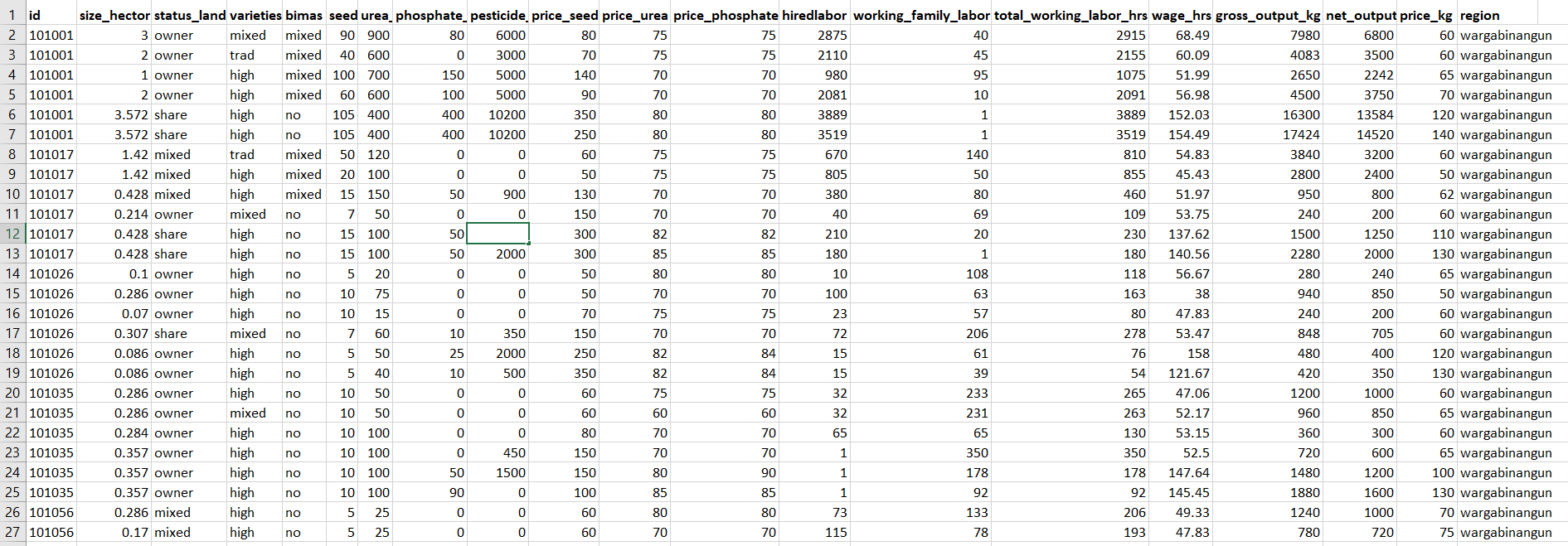
7. We check which variable positively or negatively correlated with other variable that is to see the actual relationship the variables using the correlation matrix.

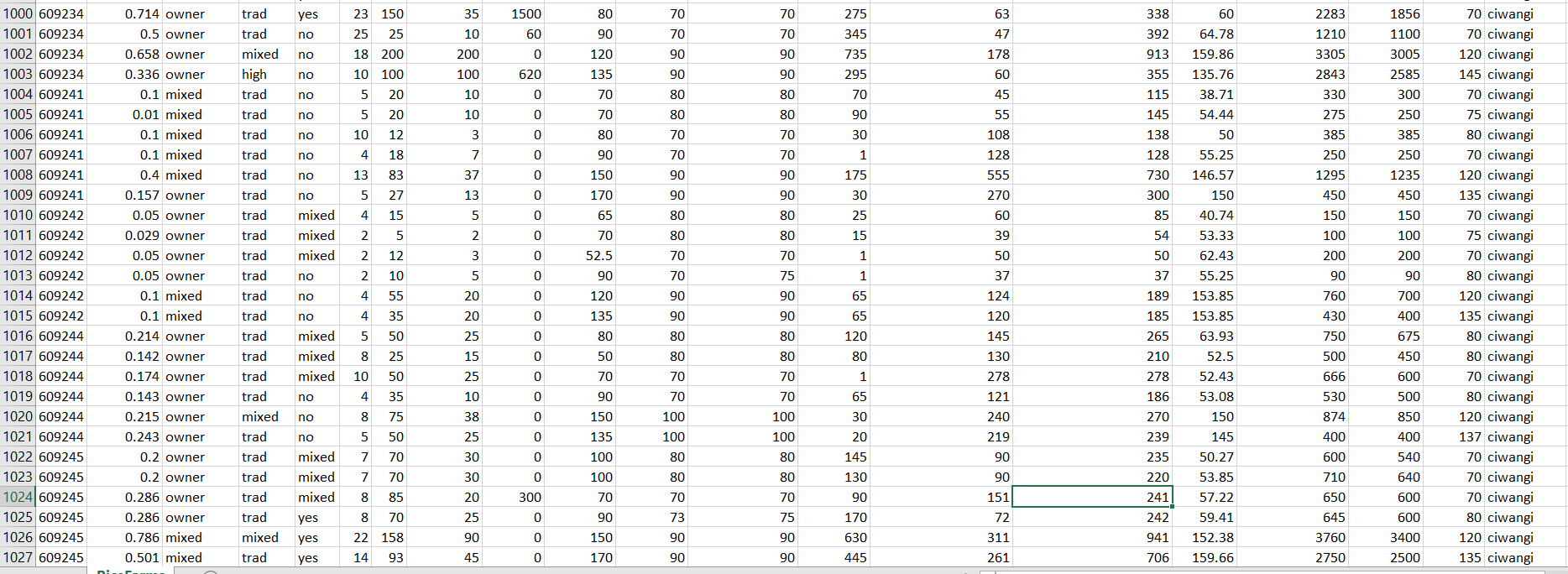
8. Using Principal component Regression Analysis we fit the model and predict the Gross output of rice Production.

9. In factor analysis, we explain the criteria and process used for deciding how many factors & which items were selected & also check how many contributions of variables that explain maximum variation in the factors.

**3.Study data**

**The dataset has been taken from Kaggle.**

****

****

**4.Information about Parameters**

Id : the farm identifier.

Size\_hector : the total area cultivated with rice, measured in hector.

Status\_land : land status, on of ‘owner’ (non-sharecroppers, owner operators or leaseholders or both), ‘share’ (sharecroppers), ‘mixed’, (mixed of the two previous status).

Varieties : one of ‘trad’ (traditional varieties), ‘high’ (high yielding varieties) and ‘mixed’ (mixed varieties).

Bimas : bimas is an intensification program; one of ‘no’ (non-bimas farmer), ‘yes’ (bimas farmer) or ‘mixed’ (part but not all of farmer’s land was registered to be in the bimas program).

Seed : seed in kilogram.

Urea : urea in kilogram.

Phosphate : phosphate in kilogram pesticide.

Pesticide : cost in Rupiah.

Pseed : cost in Rupiah per kg.

Purea : price of urea in Rupiah per kg.

Pphosph : price of phosphate in Rupiah per kg.

Hiredlabor : hired labor in hours.

Famlabor : family labor in hours total labor (excluding harvest labor).

Wage : labor wage in Rupiah per hour.

Goutput : gross output of rice in kg.

Noutput :net output, gross output minus harvesting cost (paid in terms of rice).

Price : price of rough rice in Rupiah per kg.

Region :one of ‘wargabinangun’, ‘langan’, ‘gunungwangi’, ‘malausma’, ‘sukaambit’, ‘firangi’.

**5. Analysis**

**5.1 Importing and understanding data**

**In python,**

Libraries Descriptions Used in Python

Introduction:

Python has a large collection of libraries. For Beginner’s it is easy programming language to learn and while learning it is fun, because code is easy to implement because of libraries but hard to remember. Python’s syntax is easy to learn and is of high level compared to c, c++, java etc. To update all Python Libraries in anaconda we can use command “conda update all” using command prompt. For installing or updating any library we use either “pip” or “conda” command.

1. Numpy

2. Pandas

3. Matplotlib

4. Seaborn

**1)Numpy:**

It stands for Numerical Python. Basically, Numpy is a general-Purpose array-processing package designed to efficiently manipulate large multi-dimensional arrays of arbitrary records without sacrificing too much speed for small multi-dimensional arrays

**2) Pandas:**

To work with csv files and data frames. Pandas’ data frame is a 2D data structure i.e., data is aligned in a tabular fashion in rows and columns. Basic operations perform on pandas is Creating a data frame, dealing with rows and columns, Indexing and selecting data, working with missing data, Iterating over rows and columns.

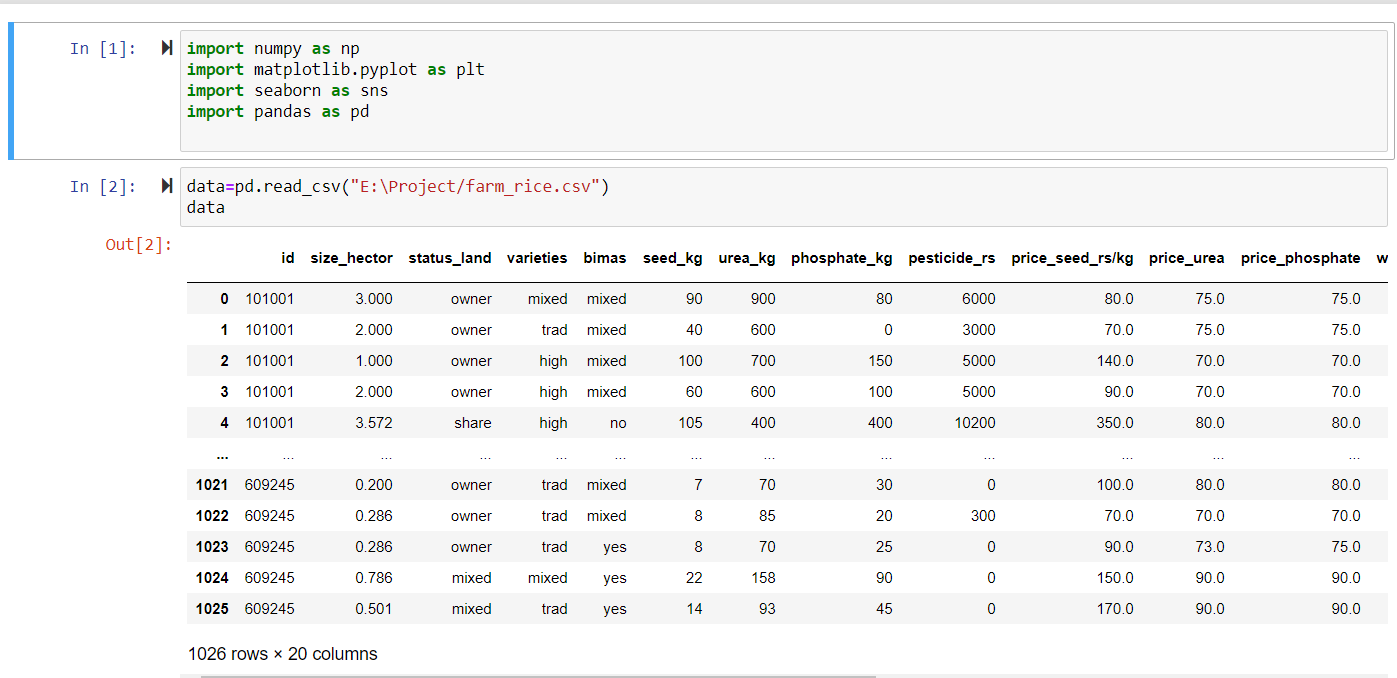
**3) Matplotlib:**

Matplotlib is a data visualization python library. Matplotlib is a plotting library for the python programming language and its numerical mathematics extension is Numpy. It is for creating 2D graphs and plots. It makes heavy use of Numpy and other extension code to provide good performance even for large arrays. Matplotlib is pre-installed in python.

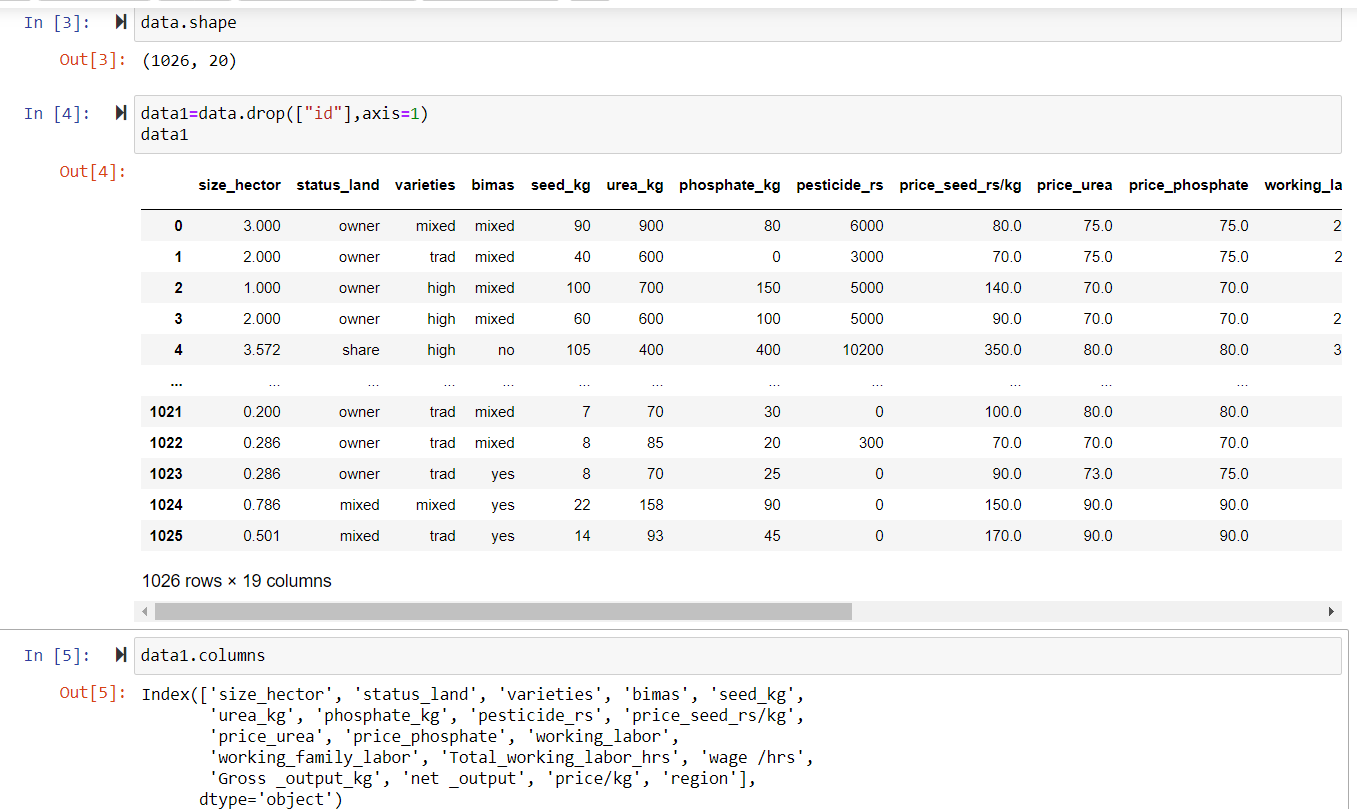
**4) Seaborn:**

It is also a data visualization python library. it is built on matplotlib so we can built it concurrently. It has simply its own library of graphs and has pleasant formatting built in. It provides high level interface for drawing attractive and informative statistical graphics.

We will be using Anaconda Jupyter notebook to work on this dataset. We will first go with importing the necessary libraries and import our data set to Jupyter notebook.

****

We can observe that the data set contain 1026 rows and 20 columns. Gross output is the columns which we going to predict and analyze.

****

Here id is an identity of farm it does not depend on gross output of rice. So here we delete the column and continue the analysis.

They also show that the parameters of data which we study.

**5.2 Graphical Representation**

Graphical Representation is a way of analysing numerical data. It exhibits the relation between data, ideas, information and concepts in a diagram. It is easy to understand and it is one of the most important learning strategies. It always depends on the type of information in a particular domain. There are different types of graphical representation.

**Pie chart:**

The “**pie chart”** is also known as “circle chart”, that divides the circular statistical graphic into sectors or slices in order to illustrate the numerical problems. Each sector denotes a proportionate part of the whole. To find out the composition of something, Pie-chart works the best at that time. In most cases, pie charts replace some other graphs like the bar graph, line plots, histograms, etc.

The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sectors of a pie chart forms a certain portion of the total(percentage). The total of all the data is equal to 360°. The total value of the pie is always 100%.

Therefore, the pie chart formula is given as **(Given Data/Total value of Data) × 360°**

Measure the angle of each slice of the pie chart and divide by 360 degrees. Now multiply the value by 100. The percentage of particular data will be calculated.

Pie charts are used to represent the proportional data or relative data in a single chart. The concept of pie slices is used to show the percentage of a particular data from the whole pie.

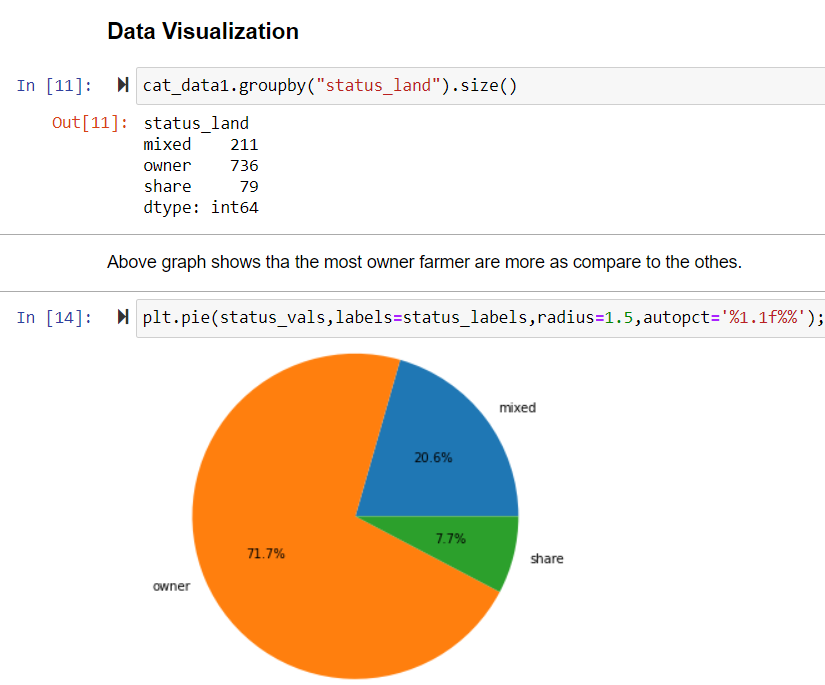
Advantages:

* The picture is simple and easy-to-understand
* Data can be represented visually as a fractional part of a whole
* It helps in providing an effective communication tool for the even uninformed audience
* Provides a data comparison for the audience at a glance to give an immediate analysis or to quickly understand information
* No need for readers to examine or measure underlying numbers themselves, which can be removed by using this chart
* To emphasize a few points you want to make, you can manipulate pieces of data in the pie chart

Disadvantages:

* It becomes less effective, if there are too many pieces of data to use
* If there are too many pieces of data. Even if you add data labels and numbers may not help here, they themselves may become crowded and hard to read
* As this chart only represents one data set, you need a series to compare multiple sets
* This may make it more difficult for readers when it comes to analyze and assimilate information quickly.

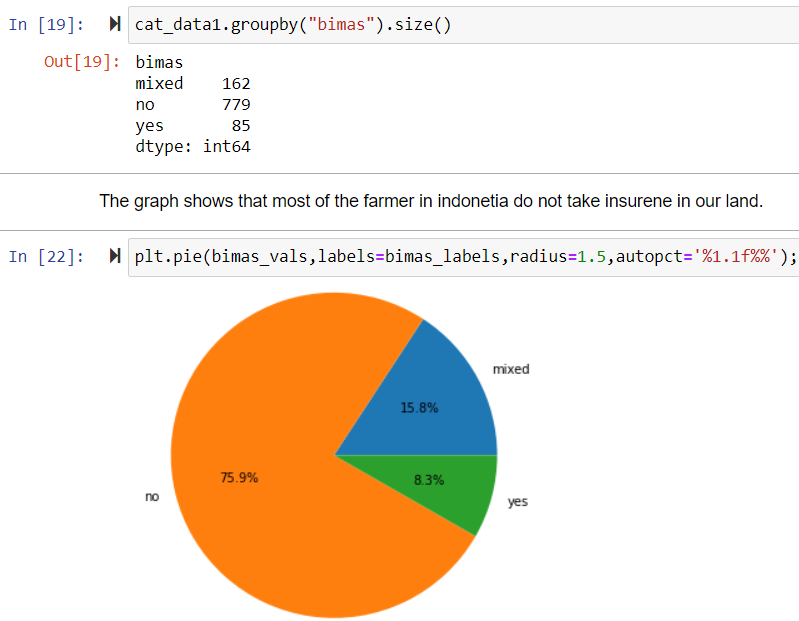
1.For Status\_land:



**Interpretation:**

From above pie plot on Indonesia status, 71.7% of farmers prefer to farm on their own and 7.7% give their land to others for farming or 20.6% of farmers do the farming themselves and give some of their farms to others for farming. This suggests that more and more farmers prefer to farm on their own.

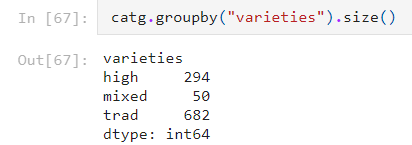
2.For Bimas:

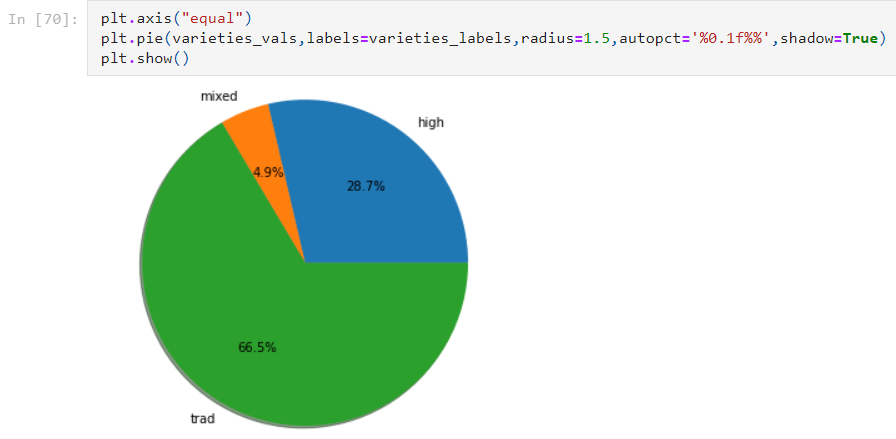


**Interpretation:**

For above pie plot we conclude that 8.3% of the population have their insurance and 75.9% of the people they do not have insurance for that Agricultural part and 15.8% they have mixed type of insurance (that means some part of agricultural land they have insurance and for remaining part of the agricultural land they don’t have their insurance) and this is called mixed. This suggest that most of the farmer’s they do not have any insurance or bimas.

3.For Varieties:

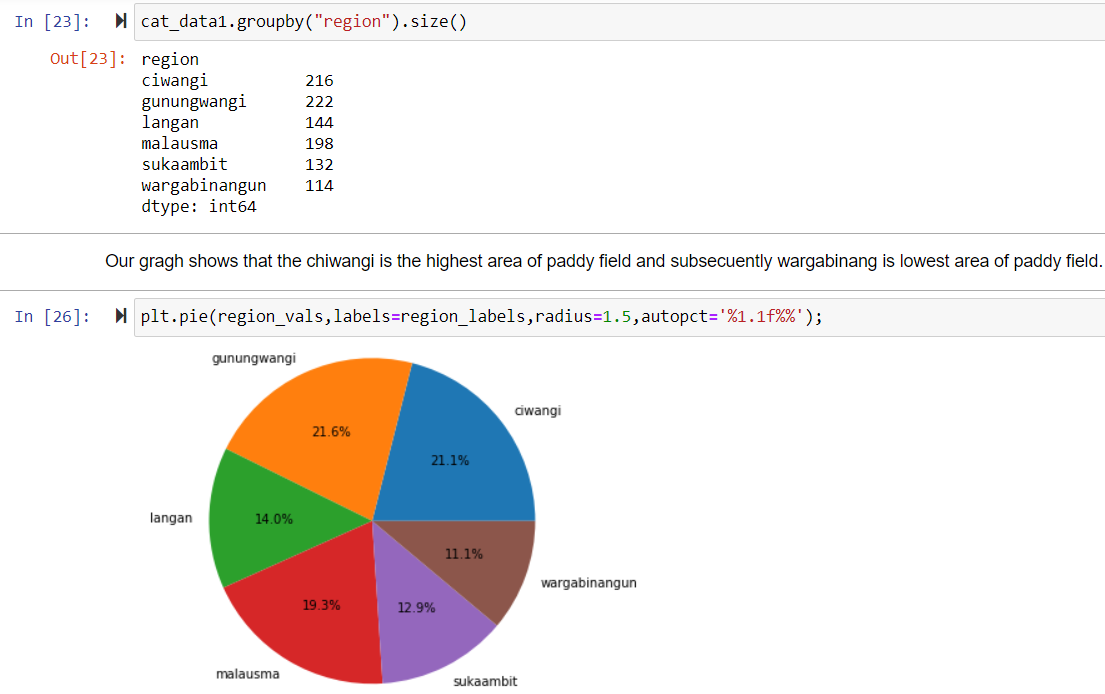




**Interpretation:**

From above pie plot, we can say that 66.5% people use the traditional varieties of seed and 28.7% people use the high yielding varieties of seed while 4.9% people use the mixed (both methods) varieties of seed. This suggests that more and more people are using traditional seeds for farming in Indonesia.

4.for region:



**Interpretation:**

From above pie plot, we can say that 21.6% of the land in Gunungwangi area is used for paddy field, while 21.1% in ciwangi, 19.3% in malausma, 14.0% in langan, 12.9% in sukaambit and 11.1% in wargabinangun. This suggest that the ciwangi is the highest area of paddy field and subsequently wargabinangun is lowest area of paddy field.

**5.3 Descriptive Statistics**

Descriptive Statistics is a branch of Statistics that aims at describing a number of features of data usually involved in a study. The main purpose of descriptive statistics is to provide a brief summary of the samples and the measures done on a particular study.

A Descriptive Statistics is a summary Statistics that quantitatively describes or summarizes features from a collection of information, while descriptive statistics is the process of using and analysis those statistics. Descriptive statistics is distinguished from inferential statistics by its aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory, and are frequently non-parametric statistics. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented. For example, in papers reporting on human subjects, typically a table is included giving the overall sample size, sample sizes in important subgroups and demographic or clinical characteristics such as the average age, the proportion of subjects of each sex, the proportion of subjects with related co-morbidities, etc.

Some measures that are commonly used to describe a data set are measures of central tendency and measures of variability or dispersion. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation, the minimum and maximum values of the variables, kurtosis and skewness.

Descriptive statistics provide simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e., summary statistics, or visual, i.e., simple to understand graphs. These summaries may either from the basis of the initial description of the data as part of a more extensive statistical analysis, or they may be sufficient in and of themselves for a particular investigation.

Descriptive Statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis and skewness.

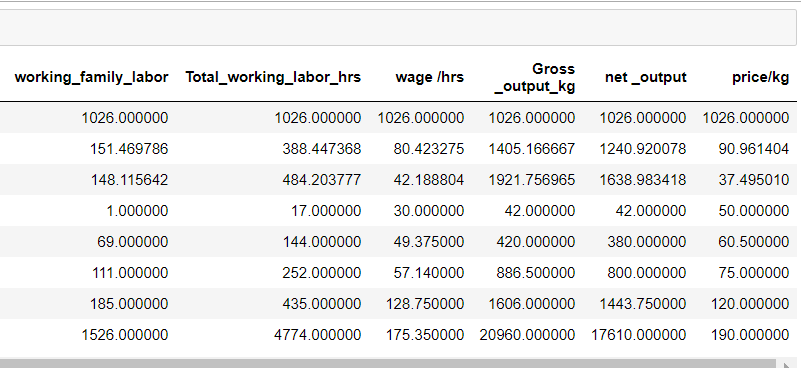
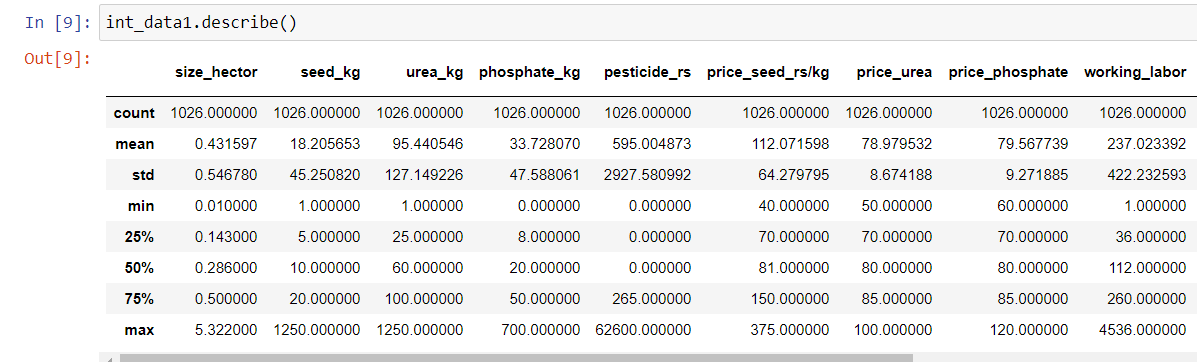
* Descriptive statistics summarizes or describes the characteristics of a data set.
* Descriptive statistics consists of two basic categories of measures: measures of central tendency and measures of variability (or spread).
* Measures of central tendency describe the center of a data set.
* Measures of variability or spread describe the dispersion of data within the set.

**Descriptive Statistics:**

**Python code:**

*df.describe()*

**Output:**

****

**Interpretation:**

**For size\_hector:**

Out of 1026 observations, 50% of the sizes of farm in hector is below 0.286 and 50% of the sizes of farm in all population is above 0.2860 hectors. therefore 25% data rise above 25%. It means most of big farm belonging to above 50% part.

**Seed\_kg:**

Out of 1026 observations, 50% of the seed in kilogram is below 10kg and 50% of the seed is above 10kg. because of most of the farm is covered above middle 50% Area.

**Gross\_Output\_kg:**

Out of 1026 observations, 50% of the total output of rice production is below 886.50 kilogram and 50% of the total output of rice production is above 886.50 kilogram.

* 1. **Shapiro-Wilk Test and Kolmogorov-Smirnov Test**

**Normality test**

**Aim** : To check normality of data we use Shapiro-Wilk Test and Kolmogorov-Smirnov Test

We know that normality assumption plays important role in Data Analysis.

We go for parametric test if data is normally distributed otherwise, we use non-parametric tests but question is How to check whether data is normally distributed or not.

There are various methods to check

1.Numerically : Shapiro-Wilk and Kolmogorov Smirnov Test

2.Graphical : Q-Q plot or Histogram

* **Shapiro-Wilk Test:**

The Shapiro-Wilk test, proposed by Shapiro in 1965, is considered the most reliable test for non-normality for small to medium sized samples by many authors. A formal hypothesis test for Null hypothesis is one sample drawn from normal distributed population.

Type of Data: Univariate, continuous data

The test statistics is defined as:

W=

Here is the sample value in ascending order,

is sample mean,

slope of observed data vs, expected normal values, normalized to constant,

constants are defined as components of the vector () = ,

where m are the expected values of the order statistics of independent and identically distributed (i. i. d) random variables sampled from the standard normal distribution (µ=0, sigma=1), and v is the covariance matrix of those statistics.

If distribution is normal, an estimate of . An estimate of if H0 true, W would =1, W<1 may be significantly different from normal ( p value determined from empirical distribution).

The Shapiro-Wilk test is a statistical test of the hypothesis that the distribution of the data as a whole deviate from a comparable normal distribution. If the test is non-significant (p> 0.05) it tells us that the distribution of the sample is not significantly different from a normal distribution.

This test for normality has been found to be the most powerful test in most situations. It is the ratio of two estimates of the variance of a normal distribution based on a random sample of n observations. The numerator is proportional to the square of the best linear estimator of the standard deviation. The denominator is the sum of square of the observations about the sample mean. The test statistic W may be written as the square of the Pearson correlation coefficient between the ordered observations and a set of weights which are used to calculate the numerator.

* + - * Shapiro-Wilk test is fairly powerful test. Not good with small samples or discrete data.
      * Shapiro-Wilk test Good power with symmetrical, short and long tails. Good with asymmetry.
      * Shapiro-Wilk test is too sensitive-you should use histogram/Q-Q plot when assessing t test /ANOVA assumptions.
* **Kolmogorov-Smirnov Test:**

The Kolmogorov-Smirnov test (K-S test) compares sample data with a fitted normal distribution to decide if a sample comes from a population with a normal distribution. Test statistics is defined as

D= ,

In Statistics, the Kolmogorov-Smirnov test is a nonparametric test of the equality of continuous, One-dimensional probability distributions that can be used to compare a sample with a reference probability distribution or to compare two samples. It is named after Andrey Kolmogorov and Nikolai Smirnov.

Where CDF is the normal cumulative distribution function. when the CDF parameters are not known a priori, the test becomes conservative and loses power. The Lilliefors correction (K-S-L test) of the Kolmogorov-Smirnov test (Lilliefors 1967) estimates the mean and standard deviation of the CDF from the data. The correction used different critical values and produces more powerful test.

This test for normality is based on the maximum difference between the observed distribution and expected cumulative-normal distribution. Since it used the sample mean and standard deviation to calculated the expected Normal distribution, the Lilliefors adjustment is used. The smaller the maximum difference the more likely that the distribution is normal.

This test has been shown to be less powerful than the other tests in most situations. It is included because of its historical popularity.

This test is used to test whether a sample comes from a specific distribution. We can use this test to determine whether a sample comes from a population which is normally distributed.

**Hypothesis:**

Null Hypothesis(H0) : Data is normally distributed

Alternative Hypothesis(H1) : Data is not normally distributed

Test of Normality:

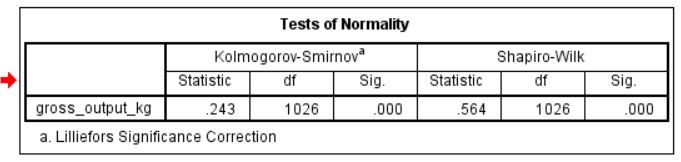
**SPSS Normality Tests:**

*Analyze>Descriptive Statistics>Explore, then Plots>Normality Tests with plots.*

**Available Test:**

Shapiro-Wilk and Kolmogorov Smirnov Normally distributed Data.

**Output in SPSS:**



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | | |
| **Status \_land** | | **Kolmogorov-Smirnova** | | | **Shapiro-Wilk** | | |
| **Statistic** | **df** | **Sig.** | **Statistic** | **df** | **Sig.** |
| **gross\_output\_kg** | **owner** | **.248** | **736** | **.000** | **.585** | **736** | **.000** |
| **share** | **.334** | **79** | **.000** | **.343** | **79** | **.000** |
| **mixed** | **.172** | **211** | **.000** | **.725** | **211** | **.000** |
| **a. Lilliefors Significance Correction** | | | | | | | |
|  | | | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | | |
| **varieties** | | **Kolmogorov-Smirnova** | | | **Shapiro-Wilk** | | |
| **Statistic** | **df** | **Sig.** | **Statistic** | **df** | **Sig.** |
| **gross\_output\_kg** | **trad** | **.171** | **682** | **.000** | **.758** | **682** | **.000** |
| **high** | **.242** | **294** | **.000** | **.620** | **294** | **.000** |
| **mixed** | **.290** | **50** | **.000** | **.575** | **50** | **.000** |
| **a. Lilliefors Significance Correction** | | | | | | | |
|  | | | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | | |
| **bimas** | | **Kolmogorov-Smirnova** | | | **Shapiro-Wilk** | | |
| **Statistic** | **df** | **Sig.** | **Statistic** | **df** | **Sig.** |
| **gross\_output\_kg** | **yes** | **.186** | **85** | **.000** | **.762** | **85** | **.000** |
| **no** | **.249** | **779** | **.000** | **.528** | **779** | **.000** |
| **mixed** | **.258** | **162** | **.000** | **.586** | **162** | **.000** |
| **a. Lilliefors Significance Correction** | | | | | | | |

**Interpretation:**

In SPSS Output above the table shows the result of both the test K.S. test as well as Shapiro-Wilk test. Here p-value is 0.000 is less than 0.05. therefore, we reject the Null Hypothesis. So, if we follow the Kolmogorov-Smrinov test we reject the Null hypothesis therefore we conclude that data is not normally distributed. Similarly, Shapiro-Wilks’s test, P-value is 0.000 is less than 0.05. therefore, we reject the Null Hypothesis and we conclude that data is not normally distributed.

From above tables we can see that data is not normally distributed for different levels of variables, hence to compare the gross output of Status\_land, varieties and bimas at different level we will use Kruskal Wallis test i.e., to compare means of Several groups.

**5.5 Kruskal-Wallis test**

**Aim**  : To Check Average of Gross output corresponding to Status\_land, Varieties, Bimas & region are same or different.

**Theory:**

Kruskal-Wallis test or H-test is the non-parametric and a valuable alternative to a one-way ANOVA test. (F-test) when the normality and equality of variance assumptions are violated.

The usual ANOVA test is used to test the equality of several Population means and is based on the fundamental assumption that the population which samples are drawn normally distributed.

This test is appropriate for use under the following circumstances:

(a)You have three or more conditions that you want to compare.

(b)Each condition is performed by a different group of participants; i.e., you have an independent-measures design with three or more conditions.

(c) The data do not meet the requirements for a parametric test. (i.e., use it if the data are not normally distributed; if the variances for the different conditions are markedly different; or if the data are measurements on an ordinal scale). If the data meet the requirements for a parametric test, it is better to use a one-way independent-measures Analysis of Variance (ANOVA) because it is more powerful than the Kruskal-Wallis test.

Procedure:

Step 1 : Define Null Hypothesis (H0) and Alternative Hypothesis (H1)

Step 2 : Rank the Sample Observation in the Combined series and

Compute Ti: sum of the ranks of the ith sample

Step 3 : Compute Test Statistics, H-value:

Under H0,

H=

Step 4 : Conclusion:

Take the tabulated values of .

If the Calculated values pf H>

We Reject H0, Otherwise, we fail to reject H0.

Notation:

K : number of samples  
ni : Size of the ith sample; i=1, 2, …, k; n=n1+n2+…+nk; Total number of observations in k samples  
Fi : Sum of the ranks of the ith Sample

Assumption:

* There are at least three independently drawn random samples
* Each sample has at least 5 observations, ni ≥ 5

**1)For region:**

**Hypothesis:**

Null Hypothesis:

H0: µ1 = µ2 = µ3 = µ4 = µ5= µ6

H0: Average of Gross output corresponding to region are same.

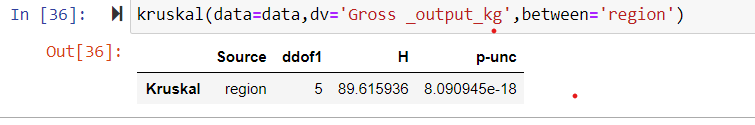
Alternative Hypothesis:

H1: µ1 ≠ µ2 ≠ µ3 ≠ µ4 ≠ µ5≠ µ6

H1: At least one average of Gross output corresponding to region are different.

In Python, Kruskal-Wallis test or H-test:

Output:



**Interpretation:**

In our output we see that p-value < 0.05 so we conclude that our given 5 region and region corresponding to gross output is not similar there will be difference between the gross output region wise.

**2)For Status\_land:**

Hypothesis:

Null Hypothesis:

H0: µ1 = µ2 = µ3

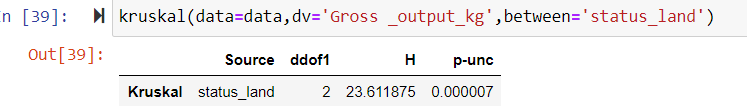
H0: Average of Gross output corresponding to Status\_land is same.

Alternative Hypothesis:

H1: µ1 ≠ µ2 ≠ µ3

H1: At least one average of Gross output corresponding to Status\_land is different.

Output:



**Interpretation:**

In our output we see that p-value<0.05 so we conclude that our given 3 status\_land and status\_land corresponding to gross output is not similar there will be difference between the gross output status\_land wise.

**3)For Varieties**

Hypothesis:

Null Hypothesis:

H0: µ1 = µ2 = µ3

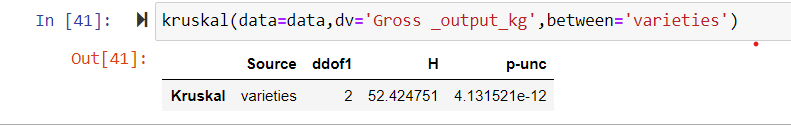
H0: Average of Gross output corresponding to Varieties is same.

Alternative Hypothesis:

H1: µ1 ≠ µ2 ≠ µ3

H1: At least one average of Gross output corresponding to Varieties is different.

Output:



**Interpretation:**

In our output we see that p-value<0.05 so we conclude that our given 3 varieties and varieties corresponding to gross output is not similar there will be difference between the gross output varieties wise.

**4)For Bimas:**

Hypothesis:

Null Hypothesis:

H0: µ1 = µ2 = µ3

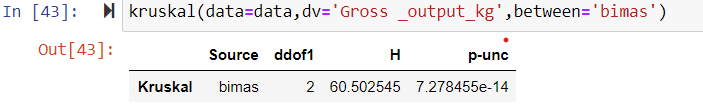
H0: Average of Gross output corresponding to Bimas is same.

Alternative Hypothesis:

H1: µ1 ≠ µ2 ≠ µ3

H1: At least one average of Gross output corresponding to Bimas is different.

Output:



**Interpretation:**

In our output we see that p-value<0.05 so we conclude that the our given 3 bimas and bimas corresponding to gross output is not similar there will be difference between the gross output bimas wise.

* 1. **Post Hoc Test**

**Aim:** To check which pair differ significantly**.**

**Theory:**

A Post hoc test is used only after we find a statistically significant result and need to determine where our differences truly came from. The term “post hoc” comes from the Latin “after the event”. There are many different Post hoc tests that have been developed and most of them will give us similar answers. we will also only focus here on the most commonly used ones. We will also discuss the concepts behind each and will not worry about calculations.

Post hoc tests are used to uncover specific differences between three or more group means when an analysis of variance (ANOVA) F test is significant.

When we get a significant F test result in an ANOVA test for a main effect of a factor with more than two levels, this tells us we reject H0. i.e., the samples are not all from populations with the same mean.

* We can use post hoc tests to tell us which groups differ from the rest.
* There are a number of tests which can be used.
* A significant F from an ANOVA tells you there is a difference somewhere between the groups
* You do not know where the difference
* To find out where the difference is, post hoc-test is are used
* There are multiple post-hoc test that are available to use
* Each will tell you where the difference is
* The Post hoc tests maintain the overall alpha level
* The difference between the tests are the assumptions of each

Commonly used Post-hoc Tests

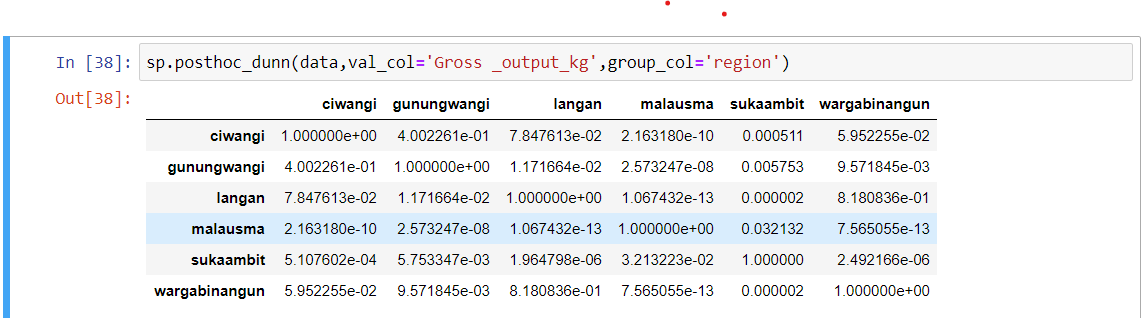
* Tukey HSD
* Scheffe
* Games-Howell
* Newman-Keuls
* Duncan
* Fisher’s LSD

**1.For region:**

In python,

Cod**e:** *sp.posthoc\_dunn(data,val\_col=’Gross\_output\_kg’,group\_col=’region’)*

Output:

****

**Interpretation:**

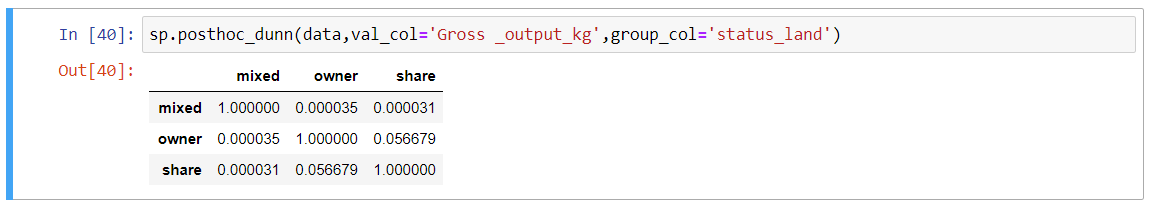
In above post hoc dunn test, pairwise comparison is done for each possible pair. P-value correspond to pair is less than 0.05 which indicates that respective pair is significantly differ.

**2.For Status\_land:**

In python,

Cod**e:** *sp.posthoc\_dunn(data,val\_col=’Gross\_output\_kg’,group\_col=’status\_land’)*

Output:



**Interpretation:**

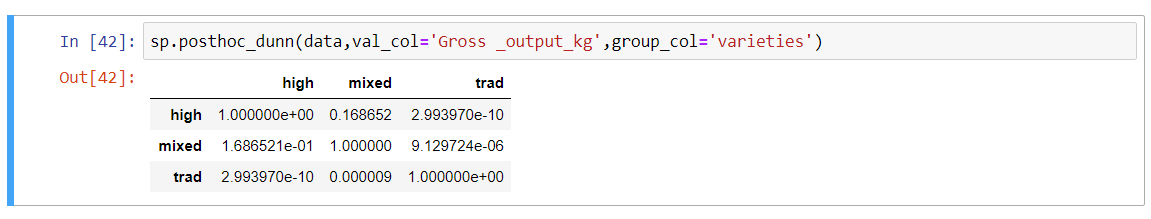
In above post hoc dunn test, pairwise comparison is done for each possible pair. P-value correspond to some pair is less than 0.05 which indicates that respective pair is significantly differ but share and owner pair have no significantly different.

**3.For Varieties:**

In python,

Cod**e:** *sp.posthoc\_dunn(data,val\_col=’Gross\_output\_kg’,group\_col=’varieties’)*

Output:



In [41]:

**Interpretation:**

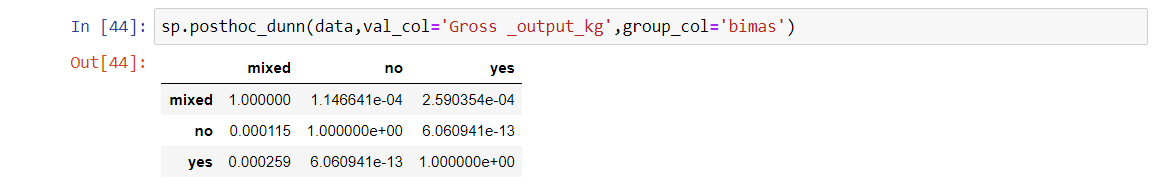
In above post hoc dunn test, pairwise comparison is done for each possible pair. P-value correspond to pair is less than 0.05 which indicates that respective pair is significantly differ.

**3.For Bimas:**

In python,

Cod**e:** *sp.posthoc\_dunn(data,val\_col=’Gross\_output\_kg’,group\_col=’Bimas’)*

Output:



**Interpretation:**

In above post hoc dunn test, pairwise comparison is done for each possible pair. P-value correspond to pair is less than 0.05 which indicates that respective pair is significantly differ.

**5.7 Correlation matrix / Heat Map**

**Aim:** To Check the relationship between two or more variables.

Theory:

Correlation matrix /Heat map:

* A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.
* Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are, that is as one increases so does the other and the closer to 1 the stronger this relationship.
* Relationship between two variables is the correlation coefficient which shows how one variable is associated with other variables. It is denoted by the Correlation Coefficient ‘r’. Which is percentage of variables in the dependent variable explained by independent variable.
* Correlation Coefficient **r=1** shows Perfectly Positive Correlated between two variables.
* Correlation Coefficient **r=-1** shows Perfectly Negative Correlated between two variables.
* Correlation Coefficient around zero means No relationship is Present between two variables

In python, I created a heatmap plot and would like to be able to explain the correlation among the variables. However, I don't understand how the relationship works and how it can be interpreted. Any explanation on how to interpret the map would be highly appreciated.

A heat map is an eye-catcher, nothing more. It gives extreme colours to extreme values so they are easily visible to the naked eye. A heat map is a data visualization technique that shows magnitude of a phenomenon as colour in two dimensions. The variation in colour may be hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. There are two fundamentally different categories of heat maps: the cluster heat map and the spatial heat map.

Heat maps originated in 2D displays of the values in a data matrix. Larger values were represented by small Dark Gray, or black squares and smaller values by lighter squares. Loua (1873) used a shading matrix to visualize social statistics across the districts of Paris. Leland Wilkinson developed the first computer program in 1994 (SYSTAT) to produce cluster heat maps with high-resolution colour graphics. The Eisen er al. display shown in the figure is a replication of the earlier SYSTAT design.

Heatmaps are used to show relationships between two variables, one plotted on each axis. Cell colourings can correspond to all manner of matrices, like a frequency count of points in each bin, or summary statistics like mean or median for a third variable.

A heatmap is a graphical representation of data that uses a system of colour-coding to represent different values. Heatmaps are used in various forms of analytics behaviour or specific webpages or webpage templates.

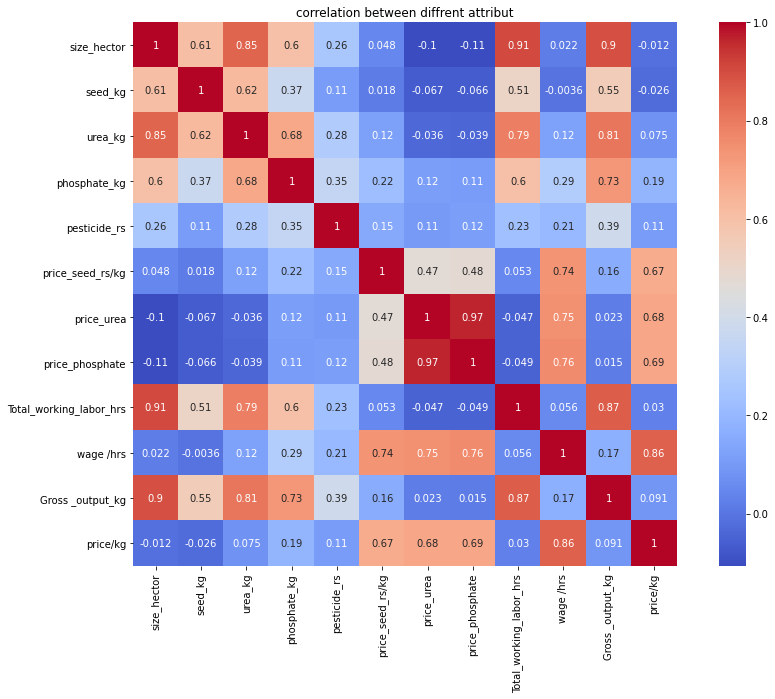
Python Code: plt.figure(figsize=(15,10))

sns.heatmap(data2.corr(),vmax=1,square=True,annot=True,cmap='coolwarm')

plt.title('correlation between different attribute')

plt.show()

Output:



**Interpretation:**

The result of correlation analysis is considered in the subsequent interpretation. A high correlation coefficient (nearly 1 or -1) means good relation between two variables and a correlation coefficient around zero means no relationship. Positive value indicates a positive relationship while negative values of r indicated an inverse relationship.

Interpretation:

* Price phosphate has highly positively correlated with price urea is (0.97)
* Gross\_output\_kg has highly positive correlated with size\_hector is (0.9)
* Gross output\_kg has highly positively correlated with total working labour hrs (0.87)
* Price\_kg has highly positively correlated with wage hrs (0.86)
* Seed\_kg has highly negatively correlated with price urea (-0.067)

From above we conclude that the gross output of rice is depend on size hectors also total working labour

**5.8 Principal Component Regression Analysis**

**Aim**: Using Principal component Regression Analysis we fit the model and predict the Gross output of rice Production.

Definition:

* It is a method of extracting important variables (in form of components) from large set of variables available in the data set.
* It is data reduction technique or data compression method

**Theory:**

* Principal component is a new set of variables that are linear combination of original variable which are uncorrelated.
* It is a dimension reduction tool that can be used to reduce a large set of variables to a small set that still contains most of information in the large set.
* It is a statistical procedure that uses an orthogonal transformation to convert a set of observation of possibly correlated variables into a set of values of linearly uncorrelated variables called principal component.
* PCA is sensitive to the relative scaling of the origin variables.
* PCA was invested in 1901 by Karl Pearson

Description:

* In Rice production analysis, twenty variables are included in the analysis. All of them are correlated and which can be reduced using PCA and FA. Relationship among sets of many interrelated variables is examined and represented in terms principal components which are orthogonal to each other.
* In PCA reduction is done in the column of data matrix so that reduced number of variables (principal components) maintain as much of original information as possible This reduction in the number of variables is helpful for comprehensive idea about the population during further research.

Objective:

* The objective of this analysis is to reduce the number of variables into few components that can explain most of the variance of the original data set

**Principal Component Analysis Model:**

Let be the elements of p-component random vector X. Assume that mean vector of X is zero and variance-covariance matrix is Σ, where Σ is real positive definite matrix.

Let us suppose that the non-zero eigen values of Σ are and corresponding eigen vectors are

For distincts (i=1, 2, …, p) a (p\*p) orthogonal matrix Γ can be formed, where

Γ = []

The Γ matrix diagonalizes Σ matrix such that

Σ = Γ λ Γ ‘ , where ᴧ = diag() = Σ Γ

Now, let us consider an orthogonal transformation of X vector to Y vector by

Y = X

Where, are the p components of Y and are called Principal Components.

**Terms and Tests used in Analysis:**

* **Correlation matrix:** A correlation matrix is a lower triangular matrix showing the simple correlations r, between all possible pairs of variables included in the analysis. The diagonal elements, which are 1, are usually omitted.
* **Eigen value:** The Eigen value represents the total variance explained by each factor.
* **Component Loadings:** The contribution of Variable in component is interpreted as component loadings.
* **Percentage variance:** It is a percentage of total variance explained by principal components.

* **Scree plot:** A scree plot is a plot of Eigen values against the number of components in order of extraction.

One of the problems of Principal component regression is how to select the number of dimensions (variables). There are several approaches

1. Proportion of variance explained. This approach is based only on input parameter space and may not be useful for wide range of problems.
2. Cross validation. How many components can predict output without adding Unstability.
3. AIC/BIC

**Test:**

* **Bartlett’s test of Sphericity:**

Bartlett’s (1951) test of Sphericity tests whether a matrix (of correlations) is significantly different from an identity matrix. The test provides probability that the correlation matrix has significant correlations among at least some of the variables in a dataset, a prerequisite for factor analysis to work. In other words, before starting with factor analysis, one needs to check whether Bartlett’s test of sphericity is significant.

* Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples.
* The Bartlett test can be used to verity that assumption.

This test tests the hypothesis that your correlation matrix is an identity matrix, which would indicate that our variables are uncorrelated and therefore unsuitable for structural detection. Small values (Less than 0.05) of significance level indicate that factor analysis may be useful with our data.

Bartlett’s test of Sphericity is used to compare your correlation matrix to the identity matrix. i.e., to check if there is a redundancy between variables that can be summarized with some factors.

For simplification we considered,

S1=Size\_hector

S2=Status\_land

S3=varieties

S4=bimas

S5=seed\_kg

S6=urea\_kg

S7=phosphate\_kg

S8=pesticide\_rs

S9=Price\_seed\_rs\_kg

S10=price\_urea

S11=Price\_phosphate

S12=total\_working\_labor\_hrs

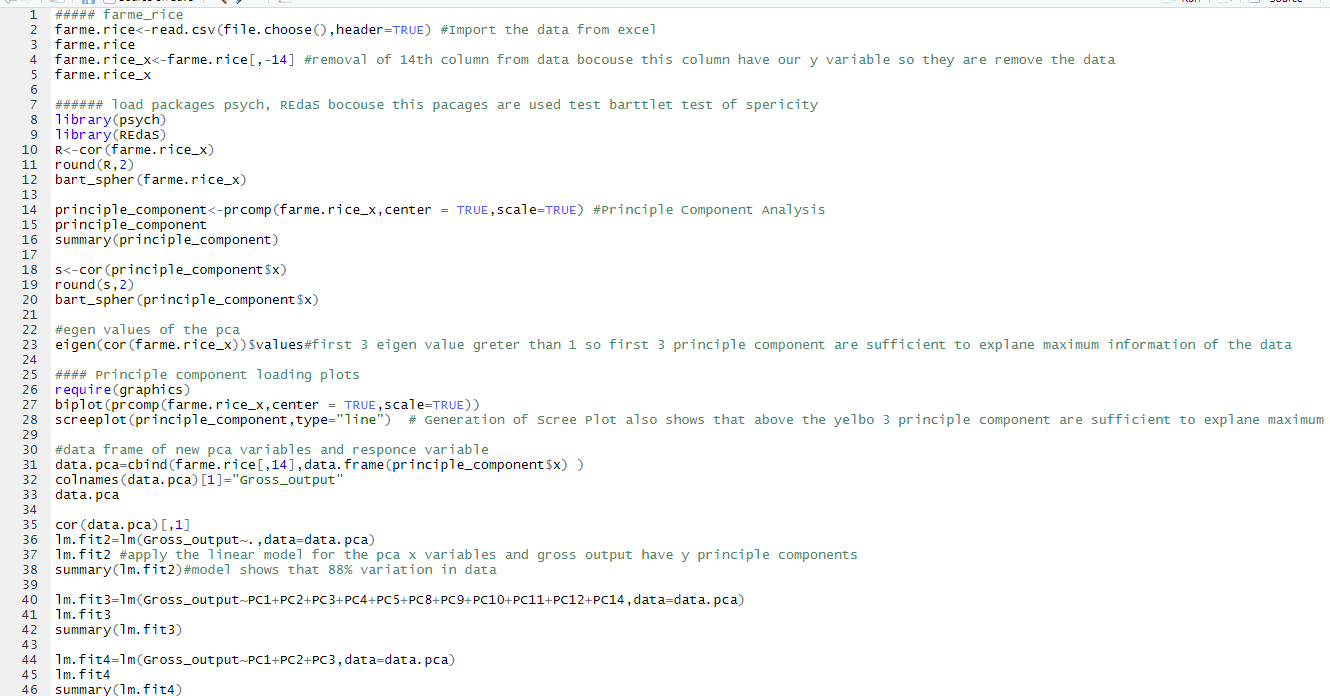
S13=Wage\_hrs

S14=Gross\_output\_kg

S15=Price\_kg

S16=region

R code for PCA:



**Output:**

**Principal Component Analysis**

The objective of this analysis is to reduce the number of variables into few components that can explain most of the variance of the original data set

R<-cor(farme.rice) # correlation matrix of data

> round(R,2)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S15 | S16 |
| S1 | 1.00 | 0.00 | 0.23 | 0.10 | 0.61 | 0.85 | 0.60 | 0.26 | 0.05 | -0.10 | -0.11 | -0.91 | 0.02 | -0.01 | -0.22 |
| S2 | 0.00 | 1.00 | -0.13 | -0.05 | -0.02 | -0.01 | 0.00 | -0.03 | -0.09 | -0.02 | -0.02 | -0.01 | -0.05 | -0.06 | 0.10 |
| S3 | 0.23 | -0.13 | 1.00 | 0.04 | 0.15 | 0.20 | 0.15 | 0.11 | 0.30 | 0.10 | 0.09 | 0.18 | 0.18 | 0.04 | -0.41 |
| S4 | 0.10 | -0.05 | 0.04 | 1.00 | 0.09 | 0.06 | -0.05 | -0.06 | -0.22 | -0.19 | -0.19 | 0.04 | -0.31 | -0.26 | -0.19 |
| S5 | 0.61 | -0.02 | 0.15 | 0.09 | 1.00 | 0.62 | 0.37 | 0.11 | 0.02 | -0.07 | -0.07 | 0.51 | 0.00 | -0.03 | -0.14 |
| S6 | 0.85 | -0.01 | 0.20 | 0.06 | 0.62 | 1.00 | 0.68 | 0.28 | 0.12 | -0.04 | -0.04 | 0.79 | 0.12 | 0.08 | -0.19 |
| S7 | 0.60 | 0.00 | 0.15 | -0.05 | 0.37 | 0.68 | 1.00 | 0.35 | 0.22 | 0.12 | 0.11 | 0.60 | 0.29 | 0.19 | -0.01 |
| S8 | 0.26 | -0.03 | 0.11 | -0.06 | 0.11 | 0.28 | 0.35 | 1.00 | 0.15 | 0.11 | 0.12 | 0.23 | 0.21 | 0.11 | -0.01 |
| S9 | 0.05 | -0.09 | 0.30 | -0.22 | 0.02 | 0.12 | 0.22 | 0.15 | 1.00 | 0.47 | 0.48 | 0.05 | 0.74 | 0.67 | -0.19 |
| S10 | -0.10 | -0.02 | 0.10 | -0.19 | -0.07 | -0.04 | 0.12 | 0.11 | 0.47 | 1.00 | 0.97 | -0.05 | 0.75 | 0.68 | 0.20 |
| S11 | -0.11 | -0.02 | 0.09 | -0.19 | -0.07 | -0.04 | 0.11 | 0.12 | 0.48 | 0.97 | 1.00 | -0.05 | 0.76 | 0.69 | 0.22 |
| S12 | 0.91 | -0.01 | 0.18 | 0.04 | 0.51 | 0.79 | 0.60 | 0.23 | 0.05 | -0.05 | -0.05 | 1.00 | 0.06 | 0.03 | -0.10 |
| S13 | 0.02 | -0.05 | 0.18 | 0.04 | 0.51 | 0.79 | 0.60 | 0.23 | 0.05 | -0.05 | -0.05 | 1.00 | 0.06 | 0.03 | -0.10 |
| S15 | -0.01 | -0.06 | 0.04 | -0.26 | -0.03 | 0.08 | 0.19 | 0.11 | 0.67 | 0.68 | 0.69 | 0.03 | 0.86 | 1.00 | 0.05 |
| S16 | -0.22 | 0.10 | -0.41 | -0.19 | -0.14 | -0.19 | -0.01 | -0.01 | -0.19 | 0.20 | 0.22 | -0.10 | 0.04 | 0.05 | 1.00 |

**Interpretation:**

We see the above correlation matrix some variables have correlated with each other. So before applying PCA we check given matrix is identity matrix or not. We use Bartlett’s test of Sphericity.

**Bartlett’s test of Sphericity:**

Hypothesis:

Null Hypothesis: Population Correlation matrix is identity

Alternative Hypothesis: Population Correlation matrix is not identity.

>bart\_spher(farme.rice)

Bartlett's Test of Sphericity

Call: bart\_spher(x = farme.rice)

X2 = 11484.131

df = 105

p-value < 2.22e-16

**Result:**

* Here p-value is less than 0.05 Hence we reject Ho i.e., Population correlation matrix is not identity matrix
* Hence significant amount of correlation is present in the data we can proceed for PCA

> summary(principle\_component)

Importance of components:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Pc1** | **Pc2** | **Pc3** | **Pc4** | **Pc5** | **Pc6** | **Pc7** | **Pc8** | **Pc9** | **Pc10** | **Pc11** | **Pc12** | **Pc13** | **Pc14** | **Pc15** |
| Standard  deviation | 2.0527 | 1.9409 | 1.2334 | 0.9874 | 0.9760 | 0.9318 | 0.8252 | 0.7530 | 0.6770 | 0.6139 | 0.4909 | 0.412 | 0.3235 | 0.2630 | 0.1816 |
| Proportion  of Variance | 0.2809 | 0.2511 | 0.1014 | 0.0650 | 0.0650 | 0.0578 | 0.0454 | 0.0378 | 0.0305 | 0.0251 | 0.0160 | 0.0113 | 0.0069 | 0.0046 | 0.0022 |
| Cumulative  Proportion | 0.2809 | 0.5321 | 0.6335 | 0.6985 | 0.7619 | 0.8198 | 0.8653 | 0.9031 | 0.9336 | 0.9587 | 0.9748 | 0.9862 | 0.9931 | 0.9978 | 1.0000 |

**Interpretation:**

In summary statistics we see that part of cumulative proportion we conclude that principal component first has explain 28% information of the data and two principal components explain 53%information of the data. And also see that three principal components explain 63% of the data. In this summary, cumulative proportion has first 6 component explain 81% information

**Correlation matrix:**

After applying PCA check correlation matrix they show that no correlation between independent variable or no problem of multicollinearity.

Call:

> s=cor(principle\_component$x)

> round(s,2)

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11 PC12 PC13 PC14 PC15

PC1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0

PC2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0

PC3 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

PC4 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0

PC5 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0

PC6 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0

PC7 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0

PC8 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0

PC9 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0

PC10 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0

PC11 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0

PC12 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0

PC13 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0

PC14 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0

PC15 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1

**Interpretation:**

Here we see that above correlation matrix is identity matrix. so, we use Principal component Analysis.

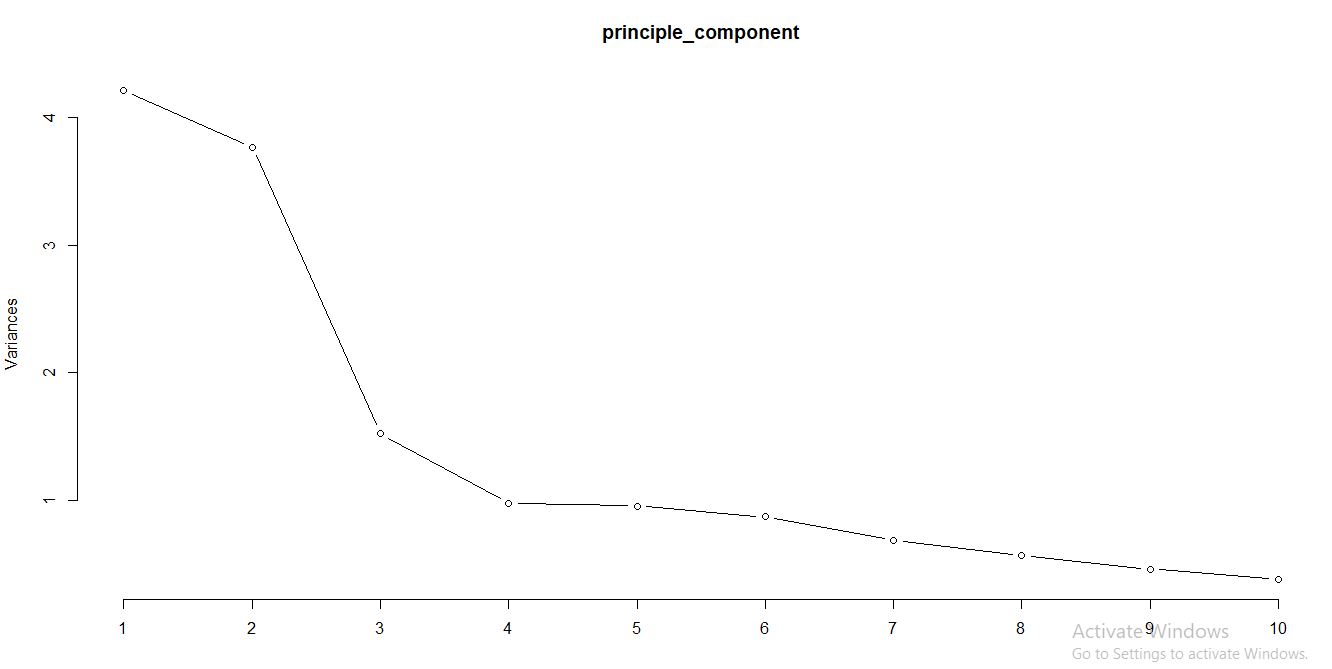
|  |
| --- |
|  |
|  |
|  |

|  |
| --- |
| > bart\_spher(principle\_component$x)#correlation matrix shows no correlation but this test have to test p value  Bartlett's Test of Sphericity  Call: bart\_spher(x = principle\_component$x)  X2 = 0  df = 105  p-value = 1 |
| **Interpretation:**  Here p-value is equal to 1 then we accept the null Hypothesis and we conclude that given Population Correlation matrix is identity matrix. |
| |  | | --- | |  | | **Eigen values:**  >eigen(cor(farme.rice\_x))$values | |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | 4.213598 | 3.767148 | 1.521397 | 0.974930 | 0.952673 | 0.868312 | 0.681005 | 0.567044 | | 0.458457 | 0.376962 | 0.241038 | 0.170535 | 0.104683 | 0.032995 |  |  | |
|  |
| **Interpretation:**  As 3 Eigen values are greater than 1. It indicates that corresponding 3 components are important to explain maximum variation in the data set. |
|  |

|  |
| --- |
|  |

**Scree plot:**

> screeplot(principle\_component,type="line")



**Interpretation:**

Scree plot shows that, above the Elbow 3 principal component are sufficient to explain maximum information.

Here we apply the linear model for the PCA x variables and gross output have y principal components, which are given below

|  |
| --- |
|  |
| |  | | --- | |  | |
| >lm.fit2=lm(Gross\_output~.,data=data.pca)  >lm.fit2  Call:  lm(formula = Gross\_output ~ ., data = data.pca)  Coefficients:   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Intercept | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | | -567.97 | 676.97 | 676.12 | -190.34 | -76.42 | -85.57 | 24.77 | 47.94 | | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 |  | | 227.20 | -87.65 | 180.77 | -456.99 | 110.97 | -487.79 | 200.13 |  |   > summary(lm.fit2)  Call:  lm(formula = Gross\_output ~ ., data = data.pca)  Residuals:  Min 1Q Median 3Q Max  -5898.1 -243.4 11.4 201.8 4469.0 Coefficients:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Estimate | Std. Error | t value | Pr(>|t|) |  | | (Intercept) | 1405.17 | 20.35 | 69.039 | <2e-16 | \*\*\* | | PC1 | -567.97 | 9.92 | -57.255 | <2e-16 | \*\*\* | | PC2 | 676.12 | 10.49 | 64.445 | <2e-16 | \*\*\* | | PC3 | -190.34 | 16.51 | -11.529 | <2e-16 | \*\*\* | | PC4 | -76.42 | 20.62 | -3.705 | 0.000222 | \*\*\* | | PC5 | -85.57 | 20.86 | -4.101 | 4.44e-05 | \*\*\* | | PC6 | 24.77 | 21.85 | 1.134 | 0.257268 |  | | PC7 | 47.94 | 24.68 | 1.943 | 0.052324 |  | | PC8 | -244.61 | 27.04 | -9.046 | <2e-16 | \*\*\* | | PC9 | 227.20 | 30.07 | 7.555 | 9.38e-14 | \*\*\* | | PC10 | -87.65 | 33.17 | -2.643 | 0.008353 | \*\* | | PC11 | 180.76 | 41.48 | 4.358 | 1.44E-05 | \*\*\* | | PC12 | -456.99 | 49.31 | -9.268 | <2e-16 | \*\*\* | | PC13 | 110.97 | 62.94 | 1.763 | 0.078170 |  | | PC14 | -487.79 | 77.40 | -6.302 | 4.38e-10 | \*\*\* | | PC15 | 200.13 | 112.10 | 1.785 | 0.074523 |  |   Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 651.9 on 1010 degrees of freedom  Multiple R-squared: 0.8866, Adjusted R-squared: 0.8849  F-statistic: 526.4 on 15 and 1010 DF, p-value: < 2.2e-16  Interpretation:  The above lm. fit2 model shows that there is an 88% variation in data. We know that if p-value is less than 0.05 then we accept the null hypothesis. |
| Now, after deleting whose individual testing t statistics in above lm. fit2 model p-value >0.05 this principal component not important for testing so, below model not important principal components excluded and apply new model for another variable in below.  therefore,  > lm.fit3=lm(Gross\_output~PC1+PC2+PC3+PC4+PC5+PC8+PC9+PC10+PC11+PC12+PC14,data=data.pca)  > lm.fit3  Call: lm(formula = Gross\_output ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC8 + PC9 + PC10 + PC11 + PC12 + PC14, data = data.pca)  Coefficients:  > summary(lm.fit3)#after deleting principle componentw r square and adjusted r square are remains same hence we conclude that  Call:  lm(formula = Gross\_output ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC8 + PC9 + PC10 + PC11 + PC12 + PC14, data = data.pca)  Residuals:  Min 1Q Median 3Q Max  -6021.1 -251.3 4.7 200.5 4506.2  Coefficients:   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Intercept | PC1 | PC2 | PC3 | PC4 | PC5 | PC8 | PC9 | PC10 | | -567.97 | 676.12 | -190.34 | -76.42 | -85.57 | -244.61 | 227.20 | -87.65 | | PC11 | PC12 | PC14 |  |  |  |  |  | | 180.77 | -456.99 | -487.79 |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Estimate | Std. Error | t value | Pr(>|t|) |  | | (Intercept) | 1405.167 | 20.427 | 69.790 | <2e-16 | \*\*\* | | PC1 | -567.974 | 9.956 | -57.048 | <2e-16 | \*\*\* | | PC2 | 676.123 | 10.529 | 64.213 | <2e-16 | \*\*\* | | PC3 | -190.337 | 16.569 | -11.488 | <2e-16 | \*\*\* | | PC4 | -76.418 | 20.698 | -3.692 | 0.000234 | \*\*\* | | PC5 | -85.568 | 20.938 | -4.087 | 4.72e-05 | \*\*\* | | PC8 | -244.613 | 27.140 | -9.013 | <2e-16 | \*\*\* | | PC9 | 227.202 | 30.183 | 7.527 | 1.14e-13 | \*\*\* | | PC10 | -87.647 | 33.286 | -2.633 | 0.008588 | \*\* | | PC11 | 180.765 | 41.626 | 4.343 | 1.55e-05 | \*\*\* | | PC12 | -456.986 | 49.489 | -9.234 | <2e-16 | \*\*\* | | PC14 | -487.786 | 77.680 | -6.279 | 5.03e-10 | \*\*\* | |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 654.3 on 1014 degrees of freedom  Multiple R-squared: 0.8853, Adjusted R-squared: 0.8841  F-statistic: 711.7 on 11 and 1014 DF, p-value: < 2.2e-16 |
| **Interpretation:**  After detecting principal components R-square and adjusted R-square are remains same hence we conclude that excluded principal component are not really important.  But above calculated eigen value and scree plot shows that first 3 components sufficient to explain the data so that 3 principal components sufficient to explain maximum information. |
| >lm.fit4=lm(Gross\_output~PC1+PC2+PC3,data=data.pca)  >lm.fit4  Call:  lm(formula = Gross\_output ~ PC1 + PC2 + PC3, data = data.pca)  Coefficient:   |  |  |  |  | | --- | --- | --- | --- | | Intercept | PC1 | PC2 | PC3 | | 1405.2 | -568.0 | 676.1 | -190.3 |   > summary(lm.fit4)  Call:  lm(formula = Gross\_output ~ PC1 + PC2 + PC3, data = data.pca)  Residuals:  Min 1Q Median 3Q Max  -7444.4 -252.0 -6.9 185.9 5536.2 |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Estimate | Std. Error | t value | Pr(>|t|) |  | | (Intercept) | 1405.17 | 23.33 | 60.24 | <2e-16 | \*\*\* | | PC1 | -567.97 | 11.37 | -49.96 | <2e-16 | \*\*\* | | PC2 | 676.12 | 12.02 | 56.23 | <2e-16 | \*\*\* | | PC3 | -190.34 | 18.92 | -10.06 | <2e-16 | \*\*\* | |
|  |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 747.2 on 1022 degrees of freedom  Multiple R-squared: 0.8493, Adjusted R-squared: 0.8488  F-statistic: 1920 on 3 and 1022 DF, p-value: < 2.2e-16  **Interpretation:**  From above model, we conclude that first three components must be explained 84% variation in the data. Hence lm. fit4 is final model for the predicting to the response variable. |
|  |
|  |
|  |
|  |

**5.9 Factor Analysis**

**Aim:** In factor analysis, we explain the criteria and process used for deciding how many factors & which items were selected & also check how many contributions of variables that explain maximum variation in the factors.

**Theory:**

**Definition:**

* A class of procedures preliminarily used for data reduction and summarization.
* It is a dimension reduction technique.
* Factor analysis allows us to look at groups of variables that tend to be correlated to each other and identify underlying the dimension that explain these correlations.

**Description:**

* In Gross Output analysis, there are large number of variables, most of which are correlated and which must be reduced to manageable level. Relationship among sets of many interrelated variables is examined and represented in terms of a few underlying factors.

**Interdependence Technique:**

Multivariate statistical technique in which the whole set of interdependent relationship is examined.

**Applications of Factor Analysis in Rice production analysis:**

* It can be used in Rice production analysis for identifying the underlying variables on which Gross output is dependent.

**Factor Analysis Model:**

* The unique factors are uncorrelated with each other and with the common factors. The common factors themselves can be expressed as

𝐹𝑖 = 𝑊 𝑖1𝑋1 + 𝑊 𝑖2𝑋2 + 𝑊 𝑖3𝑋3 + ⋯+ ;ℎ𝑒𝑟𝑒 ∑𝑊𝑖𝑗 = 1

𝐹𝑖: Estimate of ith factor

𝑊 𝑖: Weight or factor score co-efficient

k: number of variables.

* It is possible to select weights or factor score co-efficient so that the first factor explains the largest portion of the total variance. Then a second set of weights can be selected, so that the second factor accounts for most of the residual variance, subject to being uncorrelated with the first factor. The same principle could be applied to selecting additional weights for the additional factors. Thus, the factors can be estimated so that their factor scores, unlike the values of original variables, are not correlated.

**Max. No. of factors = Total no. of variables under study.**

**Terms and Tests used in Analysis:**

* **Correlation matrix:**

A correlation matrix is a lower triangular matrix showing the simple correlations r, between all possible pairs of variables included in the analysis. The diagonal elements, which are 1, are usually omitted.

* **Communality:**

It is the amount of variance, a variable shares with all other variables being considered. This is also the proportion of variance explained by the common factors.

* **Eigen value:**

The Eigen value represents the total variance explained by each factor.

* **Factor loadings:**

Factor loadings are simple correlations between the variables and the factors.

* **Factor loading plot:**

A factor loading plot is a plot of the original variables using the factor loadings as co-ordinates.

* **Factor matrix:**

A factor matrix contains the factor loadings of all the variables on the entire factor exacted.

* **Factor scores:**

Factor scored is composite score estimated for each respondent on the derived factors.

* **Percentage variance:**

It is a percentage of total variance attributed to each variable.

* **Residuals:**

Residuals are the difference between observed correlations as given in input correlation matrix and reproduced correlations as estimated from factor matrix.

* **Scree plot:**

A scree plot is a plot of Eigen values against the number of factors in order of extraction Stochastic associated with Factor Analysis.

**Tests**

* **Bartlett’s test of Sphericity:**

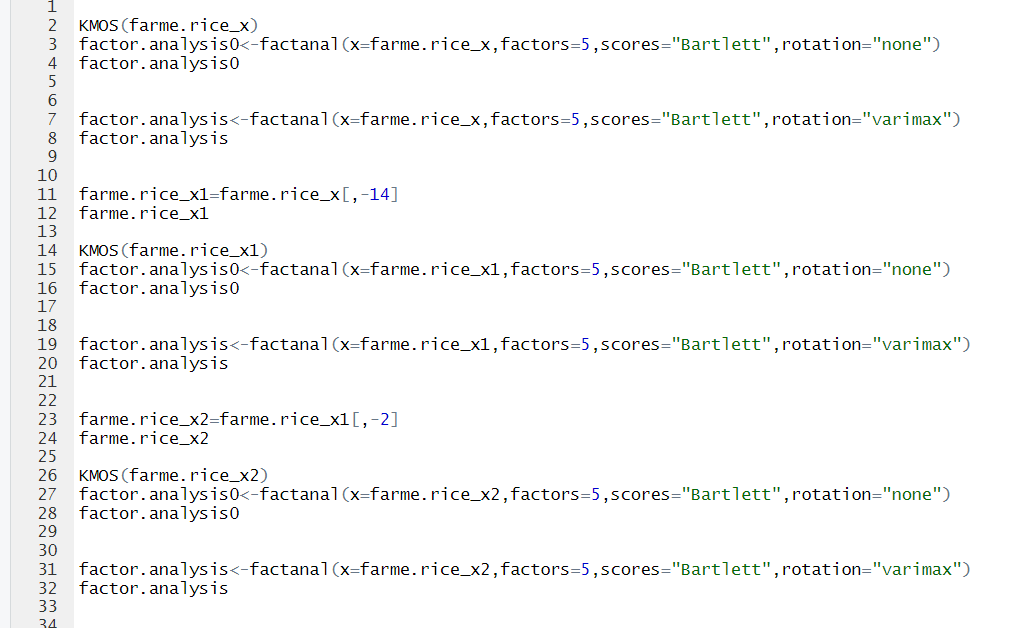
It is a test statistic used to examine the hypothesis that the variables are uncorrelated in the population. In other words, the population correlation matrix is an identity matrix.

* **Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy:**

The KMO measure of sampling adequacy is an index to examine the appropriateness of factor analysis. High values (0.5 to 1) indicate factor analysis is appropriate.

Since the value of KMO is more than 0.5 so the sample taken in the study is adequate to run the factor analysis. If the value for significance in bartlett test of sphericity is less than 0.05 so the null hypothesis i.e., all the correlation between the variables is 0 is rejected. So the correlation matrix is not an identity matrix and that is good.

**R code for Factor Analysis:**

****

**Here above factor analysis R-code can use to all varieties of seed to check variety wise analysis of gross output of rice production.**

1. Factor analysis can be used to study only traditional type of seed. Below analysis is done only when traditional seed used to farm and we check which factor is important to gross output.

When we select only traditional(trad) varieties of seed wise analysis of gross output of rice production

**Output:**

* **Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy:**

KMO-Criterion: 0.825289

KMO value = 0.825289 (>0.5).

**Interpretation**:

We know that, If KMO value is greater than 0.5 than factor analysis can be possible for this data. Here KMO is 0.82>0.5. so, we use factor analysis.

* **(i.e., measure of sample adequacy):**

>0.5 indicates that variable remains in the analysis otherwise drop variable from the analysis.

**2.** Measures of Sampling Adequacy (MSA):

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S1 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S15 |
| 0.8070 | 0.8588 | 0.9004 | 0.8149 | 0.8359 | 0.8896 | 0.8566 | 0.7434 | 0.7370 | 0.8870 | 0.8831 | 0.8211 |

**Interpretation:**

As MSA value for each variable is greater than 0.5. Hence, we conclude that each variable has some importance in factor analysis.

* **Uniqueness:**

The common variance of Xiis called communality of the variable. Therefore, the communality of Xi is the portion of total variance of XiWhich is accounted for by the common factors. The unique variance of Xi is called the uniqueness of the variable.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S1 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S15 |
| 0.073 | 0.868 | 0.285 | 0.005 | 0.331 | 0.906 | 0.171 | 0.063 | 0.020 | 0.215 | 0.090 | 0.061 |

* **Factor Loadings:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loadings: | | | | | |
|  | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
| S1 | 0.958 |  |  |  |  |
| S4 |  | -0.173 |  |  | -0.306 |
| S5 | 0.824 |  |  |  | 0.174 |
| S6 | 0.848 | 0.146 |  | 0.499 |  |
| S7 | 0.664 | 0.101 |  | 0.421 | 0.194 |
| S8 | 0.180 |  |  | 0.117 | 0.206 |
| S9 |  | 0.862 | 0.245 |  |  |
| S10 |  | 0.434 | 0.859 |  |  |
| S11 |  | 0.431 | 0.884 |  | 0.100 |
| S12 | 0.879 |  |  |  |  |
| S13 |  | 0.775 | 0.469 |  | 0.286 |
| S15 |  | 0.913 | 0.311 |  |  |

**Conclusion:**

A factor can be interpreted in terms of the variables that load high on it. Here below factors shows that which variable having maximum contribution to explain maximum variations in that factors.

* Test of the Hypothesis:

The chi square statistic is 5.86 on 5 df.

& P-value is 0.32

Conclusion: Our test indicates that p-value >0.05. so, we conclude that our 5 factors are sufficient to explain maximum variation in data.

**Factor1**

1.S1: Size\_hector

2.S5: seed\_kg

3.S6: urea\_kg

4.S7: phosphare\_kg

5.S12: total\_working\_labor\_hrs

**Conclusion:**

From this we conclude that above five variables have high loading on factor1. Hence this variable forming a one group we name this factor as farming components in rice production.

**Factor2**

1.S9: Price\_seed

2.S13: wage\_hrs

3.S15: Price\_kg

**Conclusion:**

From this we conclude that above three variables have high loading on factor2. Hence this variable forming a one group we name this factor as paid Rupees in rice production.

**Factor3**

1.S10: Price\_urea

2.S11: Price\_Phosphate

**Conclusion:**

From this we conclude that above two variables have high loading on factor3. Hence this variable forming a one group we name this factor as paid Rupees in Pestiside in rice production.

**Factor4**

1.S6: urea\_kg

2.S7: Phosphate\_kg

**Conclusion:**

From this we conclude that above two variables have high loading on factor4. Hence this variable forming a one group we name this factor as Pestiside in rice production.

**Factor5**

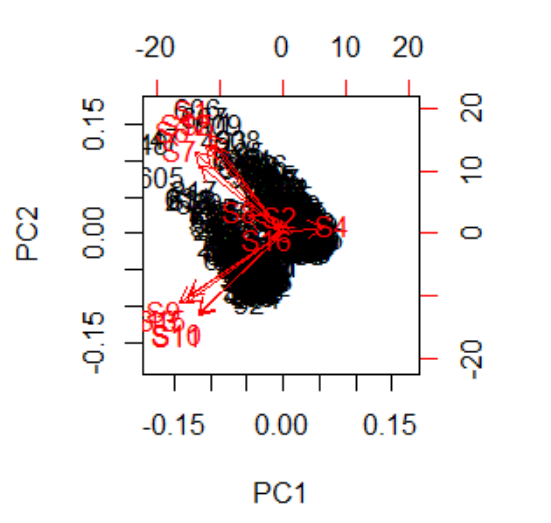
1.S8: Pesticide\_rs

2.S13: Wage\_hrs

**Conclusion:**

From this we conclude that above two variables have high loading on factor3. Hence this variable forming a one group we name this factor as Rupees paid on Pestiside in rice production.

**Biplot:**



**Interpretation:**

In above Biplot, the below arrows and its corresponding variables load principal component first and similarly above arrows and its corresponding variables load principal component two.

2. Similarly, this procedure follows the **high varieties** of seed and analysis of corresponding parameters which is important to gross output of rice production.

* **Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy:**

KMO-Criterion: 0.8003236

KMO value = 0.8003236 (>0.5).

**Interpretation**:

We know that, If KMO value is greater than 0.5 than factor analysis can be possible for this data. Here KMO is 0.80>0.5. so, we use factor analysis.

* **(i.e., measure of sample adequacy):**

>0.5 indicates that variable remains in the analysis otherwise drop variable from the analysis.

**2.** Measures of Sampling Adequacy (MSA):

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S1 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S15 | S16 |
| 0.7578 | 0.9369 | 0.8180 | 0.8732 | 0.8886 | 0.7105 | 0.8602 | 0.7286 | 0.7267 | 0.7615 | 0.8001 | 0.8647 | 0.8199 |

* **Uniqueness:**

The common variance of Xiis called communality of the variable. Therefore, the communality of Xi is the portion of total variance of XiWhich is accounted for by the common factors. The unique variance of Xi is called the uniqueness of the variable.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S1 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S15 | S16 |
| 0.079 | 0.691 | 0.476 | 0.098 | 0.419 | 0.005 | 0.437 | 0.022 | 0.019 | 0.005 | 0.022 | 0.197 | 0.700 |

* **Factor Loadings:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loadings: | | | | | |
|  | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
| S1 | 0.946 |  | -0.101 |  |  |
| S4 | 0.203 | -0.490 | -0.127 |  |  |
| S5 | 0.562 |  |  |  | 0.438 |
| S6 | 0.895 |  | -0.108 | 0.228 | 0.283 |
| S7 | 0.690 | 0.221 |  | 0.956 |  |
| S8 | 0.238 | 0.121 |  |  |  |
| S9 |  | 0.742 | 0.106 |  |  |
| S10 |  | 0.443 | 0.881 |  |  |
| S11 |  | 0.466 | 0.870 |  |  |
| S12 | 0.966 |  |  |  | -0.229 |
| S13 |  | 0.913 | 0.352 | 0.113 |  |
| S15 |  | 0.780 | 0.431 |  |  |
| S16 |  |  | 0.527 |  |  |

**Conclusion:**

A factor can be interpreted in terms of the variables that load high on it. Here below factors shows that which variable having maximum contribution to explain maximum variations in that factors.

* Test of the Hypothesis:

The chi square statistic is 41.93 on 5 df.

& P-value is 0.092

**Conclusion**: Our test indicates that p-value >0.05. so, we conclude that our 5 factors are sufficient to explain maximum variation in data

**Factor1:**

1.S1: Status\_land

2.S5: Seed\_kg

3.S6: urea\_kg

4.S7: Phosphate\_kg

5.S12: total\_working\_labor\_hrs

**Conclusion:**

From this we conclude that above five variables have high loading on factor1. Hence this variable forming a one group we name this factor as farming components in rice production.

**Factor2:**

1.S9: Price\_Seed

2.S13: Wage\_hrs

3.S15: Price\_kg

**Conclusion:**

From this we conclude that above three variables have high loading on factor2. Hence this variable forming a one group we name this factor as Rupees paid in rice production.

**Factor3:**

1.S10: Price\_urea

2.S11: Price\_Phosphate

3.S16: region

**Conclusion:**

From this we conclude that above three variables have high loading on factor3. Hence this variable farming a one group we name this factor as Rupees paid in Pestiside in rice production.

**Factor4:**

1.S7: Phosphate\_kg

**Conclusion:**

From this we conclude that above one variable has high loading on factor4. Hence this variable farming a one group we name this factor as in Pestiside in rice production.

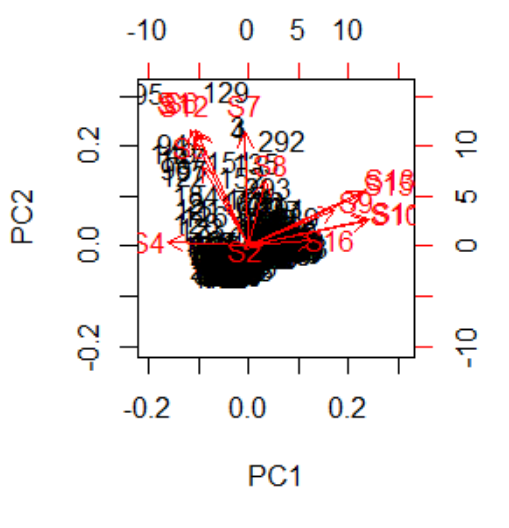
**Factor5:**

S5: Seed\_kg

**Conclusion:**

From this we conclude that above one variable has high loading on factor5. Hence this variable farming a one group we name this factor as seed in rice production.

* Biplot:



**Conclusion:**

Above biplot shows that most of the variables load the principal component two.

3. Similarly, this procedure follows the **mixed varieties** of seed and analysis of corresponding parameters which is important to gross output of rice production

Output:

* **Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy:**

KMO-Criterion: 0.8003236

KMO value = 0.7327032 (>0.5).

**Interpretation**:

We know that, If KMO value is greater than 0.5 than factor analysis can be possible for this data. Here KMO is 0.73>0.5. so, we use factor analysis.

* **(i.e. measure of sample adequacy ):**

>0.5 indicates that variable remains in the analysis otherwise drop variable from the analysis.

**2.** Measures of Sampling Adequacy (MSA):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S1 | S5 | S6 | S9 | S10 | S11 | S12 | S13 | S15 | S16 |
| 0.6776 | 0.6873 | 0.7554 | 0.7770 | 0.6701 | 0.6599 | 0.8555 | 0.7936 | 0.7323 | 0.8503 |

* **Uniqueness:**

The common variance of Xiis called communality of the variable. Therefore, the communality of Xi is the portion of total variance of XiWhich is accounted for by the common factors. The unique variance of Xi is called the uniqueness of the variable.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S1 | S5 | S6 | S9 | S10 | S11 | S12 | S13 | S15 | S16 |
| 0.005 | 0.005 | 0.156 | 0.385 | 0.05 | 0.014 | 0.039 | 0.024 | 0.038 | 0.534 |

* **Factor Loadings:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loadings: | | | | | |
|  | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
| S1 | 0.940 | 0.321 |  |  |  |
| S5 | 0.925 | 0.352 |  | -0.123 |  |
| S6 | 0.708 | 0.279 | 0.154 | 0.484 |  |
| S9 | -0.366 | 0.335 | 0.538 | -0.144 | 0.242 |
| S10 | -0.480 | 0.868 | -0.102 |  |  |
| S11 | -0.481 | 0.861 |  |  |  |
| S12 | 0.881 | 0.381 | 0.119 | 0.149 |  |
| S13 | -0.522 | 0.660 | 0.472 |  | -0.206 |
| S15 | -0.493 | 0.560 | 0.568 | -0.144 | 0.249 |
| S16 | -0.478 | 0.405 |  |  | -0.243 |

**Conclusion:**

A factor can be interpreted in terms of the variables that load high on it. Here below factors shows that which variable having maximum contribution to explain maximum variations in that factors.

**Factor1:**

1.S1: Status\_land

2.S5: Seed\_kg

3.S6: urea\_kg

4.S12: total\_working\_labor\_hrs

**Conclusion:**

From this we conclude that above four variables have high loading on factor1. Hence this variable forming a one group we name this factor as farming components in rice production.

**Factor2:**

1.S10: Price\_urea

2.S11: Price\_Phosphate

3.S13: Wage\_hrs

4.S15: Price\_kg

**Conclusion:**

From this we conclude that above four variables have high loading on factor2. Hence this variable forming a one group we name this factor as paid Rupees in rice production.

**Factor3:**

1.S9: Price\_Seed

2.S15: Price\_kg

**Conclusion:**

From this we conclude that above two variables have high loading on factor3. Hence this variable forming a one group we name this factor as Rupees paid in Pestiside in rice production.

**Factor4:**

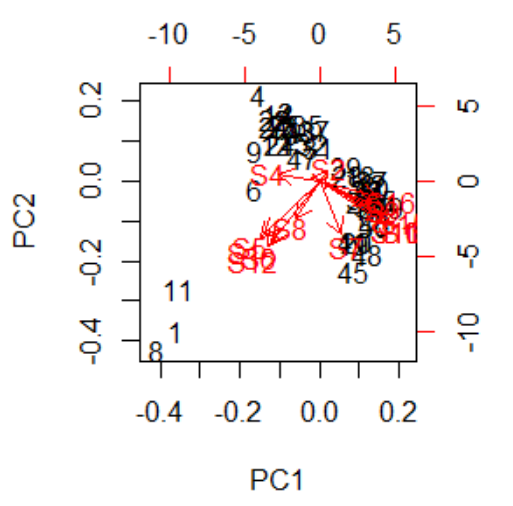
1.S6: : urea\_kg

**Conclusion:**

From this we conclude that above one variable has high loading on factor4. Hence this variable forming a one group we name this factor as Pestiside in rice production.

Here we understand that if we split data into varieties wise and Analysis of factor analysis. both the trad varieties and high varieties have same factor introduce. But mixed varieties are not having same factor introduce. Which are different from the two varieties.

**Biplot:**



**Interpretation:**

In above Biplot, the below arrows and its corresponding variables load principal component first and similarly above arrows and its corresponding variables load principal component two.

**Conclusion**

**1.Graphical Representation:**

* The pie plot on status\_land, 71.7% of farmers prefer to farm on their own and 7.7% give their land to others for farming or 20.6% of farmers do the farming themselves and give some of their farms to others for farming. This suggests that more and more farmers prefer to farm on their own.
* The pie plot on Bimas, 8.3% of the population have their insurance and 75.9% of the people they do not have insurance for that Agricultural part and 15.8% they have mixed type of insurance (that means some part of agricultural land they have insurance and for remaining part of the agricultural land they don’t have their insurance) and this is called mixed. This suggest that most of the farmer’s they do not have any insurance or bimas.
* The pie plot on Varieties, 66.5% people use the traditional varieties of seed and 28.7% people use the high yielding varieties of seed while 4.9% people use the mixed (both methods) varieties of seed. This suggests that more and more people are using traditional seeds for farming in Indonesia.
* The pie plot on region, 21.6% of the land in Gunungwangi area is used for paddy field, while 21.1% in ciwangi, 19.3% in malausma, 14.0% in langan, 12.9% in sukaambit and 11.1% in wargabinangun. This suggest that the ciwangi is the highest area of paddy field and subsequently wargabinangun is lowest area of paddy field.

**2.Descriptive Statistics:**

**Our data is not normally distributed. So, median is representing the data.**

* For size\_hector:

Out of 1026 observations, 50% of the sizes of farm in hector is below 0.286 and 50% of the sizes of farm in all population is above 0.2860 hectors. therefore 25% data rise above 25%. It means most of big farm belonging to above 50% part.

* Seed\_kg:

Out of 1026 observations, 50% of the seed in kilogram is below 10kg and 50% of the seed is above 10kg. because of most of the farm is covered above middle 50% Area.

* Gross\_Output\_kg:

Out of 1026 observations, 50% of the total output of rice production is below 886.50 kilogram and 50% of the total output of rice production is above 886.50 kilogram.

**3. Shapiro-Wilk Test and Kolmogorov-Smirnov Test**

Here we check normality of data we use Shapiro-Wilk Test and Kolmogorov-Smirnov Test

In SPSS Output, both the test K.S. test as well as Shapiro-Wilk test. Here p-value is 0.000 is less than 0.05. therefore, we reject the Null Hypothesis and we conclude that data is not normally distributed.

From above tables we can see that data is not normally distributed for different levels of variables, hence to compare the gross output of Status\_land, varieties and bimas at different level we will use Kruskal Wallis test i.e., to compare means of Several groups.

**3. Kruskal-Wallis test**

**Here we check Average of Gross output corresponding to Status\_land, Varieties, Bimas region are same or different.**

* Here given 3 status\_land (owner, share, mixed) and status\_land corresponding to gross output is not similar there will be difference between the gross output status\_land wise.
* Here given 3 varieties (trad, high, mixed) and varieties corresponding to gross output is not similar there will be difference between the gross output varieties wise.
* Here given bimas (yes, no, mixed) and bimas corresponding to gross output is not similar there will be difference between the gross output bimas wise.

**4. Post-hoc test**

**Here we check which pair differ significantly using the test.**

1.For region:

Herepairwise comparison is done for each possible pair. P-value correspond to pair is less than 0.05 which indicates that respective pair is significantly differ.

2.For Status\_land:

Here pairwise comparison is done for each possible pair. P-value correspond to some pair is less than 0.05 which indicates that respective pair is significantly differ but share and owner pair have no significantly different.

3.For Varieties:

Here pairwise comparison is done for each possible pair. P-value correspond to pair is less than 0.05 which indicates that respective pair is significantly differ.

4.For Bimas:

Here pairwise comparison is done for each possible pair. P-value correspond to pair is less than 0.05 which indicates that respective pair is significantly differ.

**5.Correlation Matrix:**

**Here we check relationship between two or more variables.**

The result of correlation analysis is considered in the subsequent interpretation. A high correlation coefficient (nearly 1 or -1) means good relation between two variables and a correlation coefficient around zero means no relationship. Positive value indicates a positive relationship while negative values of r indicated an inverse relationship.

* Price phosphate has highly positively correlated with price urea is (0.97)
* Gross\_output\_kg has highly positive correlated with size\_hector is (0.9)
* Gross output\_kg has highly positively correlated with total working labour hrs (0.87)
* Price\_kg has highly positively correlated with wage hrs (0.86)
* Seed\_kg has highly negatively correlated with price urea (-0.067)

From above we conclude that the gross output of rice is depending on size hectors also total working labour

**6. Principal component regression Analysis:**

Here Principal component Regression Analysis is used to compute important variables from data to Analyse gross output of rice production.

The below components are most important from data to Predict gross output of rice production.

Coefficient:

|  |  |  |  |
| --- | --- | --- | --- |
| Intercept | PC1 | PC2 | PC3 |
| 1405.2 | -568.0 | 676.1 | -190.3 |

Here we conclude that first three components must be explained 84% variation in the data. Hence which is the final model for the predicting to the response variable.

**7. Factor Analysis:**

Here factor analysis is performed to check the main factors which are important to determine the gross output of rice.

* **Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy:**

KMO-Criterion: 0.825289

KMO value = 0.825289 (>0.5).

Here we know that, If KMO value is greater than 0.5 than factor analysis can be possible for this data. Here KMO is 0.82>0.5. so, we use factor analysis

**Factor1:**

1.S1: Status\_land

2.S5: Seed\_kg

3.S6: urea\_kg

4.S7: Phosphate\_kg

5.S12: total\_working\_labor\_hrs

From this we conclude that above five variables have high loading on factor1. Hence this variable forming a one group we name this factor as farming components in rice production.

**Factor2:**

1.S9: Price\_Seed

2.S13: Wage\_hrs

3.S15: Price\_kg

From this we conclude that above three variables have high loading on factor2. Hence this variable forming a one group we name this factor as Rupees paid in rice production.

**Factor3:**

1.S10: Price\_urea

2.S11: Price\_Phosphate

3.S16: region

From this we conclude that above three variables have high loading on factor3. Hence this variable forming a one group we name this factor as Rupees paid in Pestiside in rice production.

**Reference**

* K. C. Bhuyan – Multivariate Analysis and its Application.
* John E. Kolassa- An introduction to Nonparametric Statistics (2020)
* Food and Agriculture Organization if the United Nations 2020

books.google.co.in

* Rice production in Indonesia as Explained by Minister of Agriculture

books.google.co.in

Above analysis will be performed using below statistical software.

1. Python (Jupyter Notebook),

(https://python.com/).

2. RStudio, (<https://rstudio.com/>)

3. SPSS

4. Microsoft Excel

Website: 1. <https://www.statisticsshowto.com>

[2. https://en.wikipedia.org](%20%20%20%20%20%20%20%20%20%20%202.%20https://en.wikipedia.org)

3. Youtube.com

Project Learning Remark:

* 1. We goy knowledge other subjects excluding syllabus
  2. We learn various techniques to analyse data
  3. Various software skills were developed

Thank you