**Problem 1A:**

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.mygreatlearning.com/courses/61464/files/4202809/download?verifier=B8WioTvoDP8ft2YLs6N2Bej5h0qlsPoYamzkwRn5&wrap=1)] are collected and each person’s educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

 [Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

* 1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

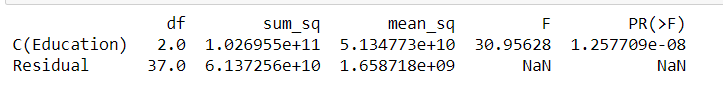
Solution: Formulation of hypothesis for conducting one-way ANOVA for education qualification w.r.t salary

* H0: Salary depend on education qualification
* Ha: Salary does not depend on education
* Confidence level = 0.05

Formulation of hypothesis for conducting one-way ANOVA for occupation w.r.t salary

* H0: Salary depend on occupation
* Ha: Salary does not depend on occupation
* Confidence level = 0.05
  1. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

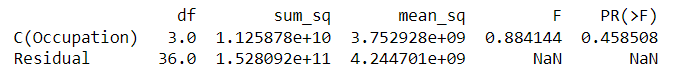
To perform one-way ANOVA for education w.r.t the variable ‘Salary’, we apply the ANOVA formula in the Jupyter notebook and run the AOV table. We get following output:



From the above table, we find that the P value is less than 0.05, hence we reject the null hypothesis.

* 1. Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

To perform one-way ANOVA for occupation w.r.t the variable ‘Salary’, we apply the ANOVA formula in the Jupyter notebook and run the AOV table. We get following output:



From the above table, we find that the P value is greater than 0.05, hence we do not reject the null hypothesis.

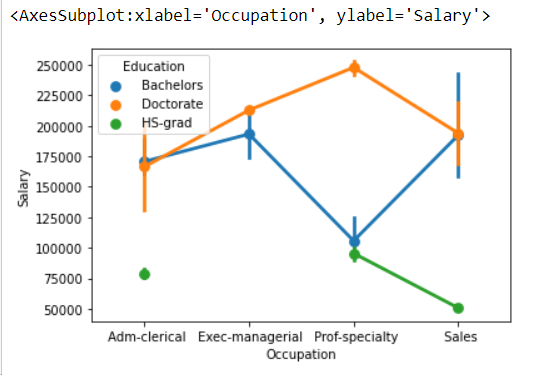
1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

**Problem 1B:**

* 1. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot. [hint: use the ‘point plot’ function from the ‘seaborn’ function]

Solution:

As seen from the below interaction plots, there seems to be moderate interaction between the two categorical variables.



Adm-clerical and sales professionals with bachelors and doctorate degrees earn almost similar salary packages.

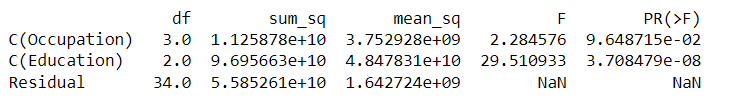
* 1. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education\*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

Formulation of hypothesis for conducting two-way ANOVA based on education and occupation w.r.t salary.

• H0: Salary depends on both categories - education and occupation

• Ha: Salary does not depend on at least one of the categories - education and occupation

• Confidence level = 0.05



Considering both education and occupation, education is a significant factor as the P value is 0.05

* 1. Explain the business implications of performing ANOVA for this particular case study.

By performing ANOVA on the given data set, we can conclude that salary is dependent on occupation.

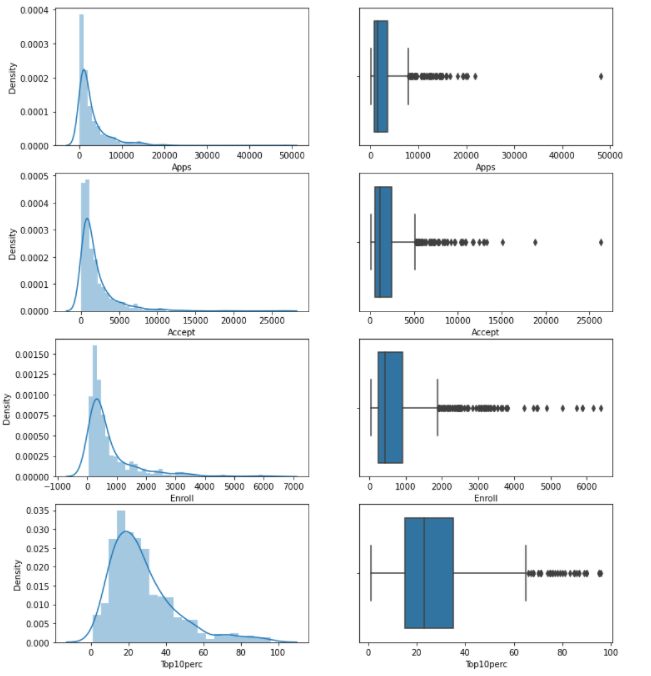
**Problem 2:**

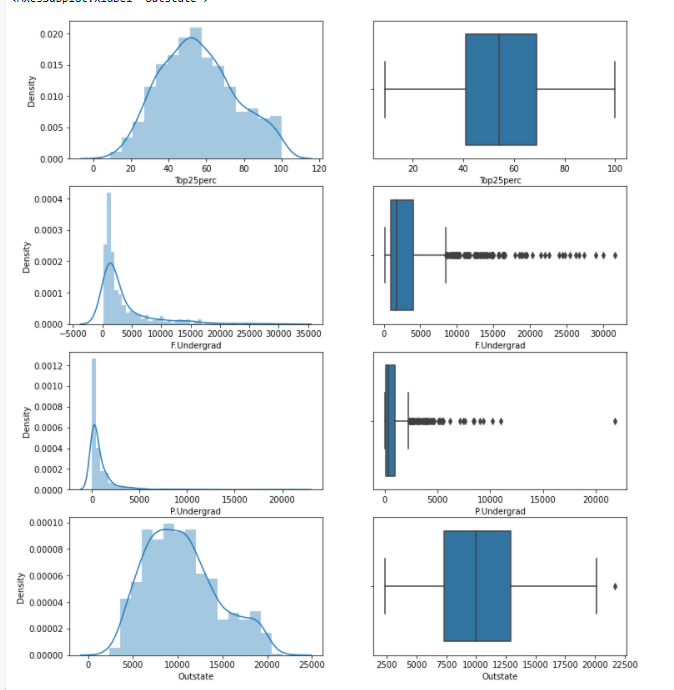
The dataset [Education - Post 12th Standard.csv](https://olympus.mygreatlearning.com/courses/61464/files/3787633/download?verifier=nD01H49pUbnjqvxvUsKQ274rHkaI518tsqSCSxtO&wrap=1) contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: [Data Dictionary.xlsx](https://olympus.mygreatlearning.com/courses/61464/files/3787632/download?verifier=SXAE7nqnvHtjzEQJ0ctUHvDff8CW9rg61cXPhYO4&wrap=1).

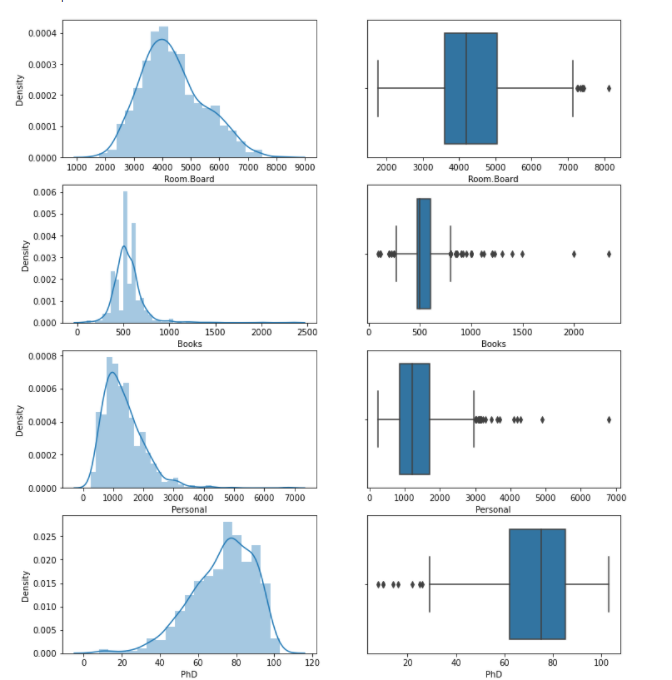
* 1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

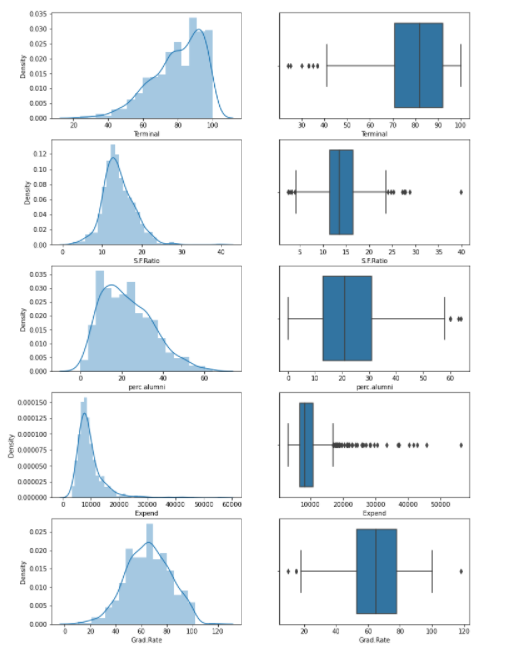
The main purpose of univariate analysis is to describe the data, summarize and finds pattern, it doesn’t deal with causes and relationships unlike regression.

1. We start with loading the dataset, checking its shape and data types of variables, shape tell us how many rows and columns we have in the data and data type tell us whether the variable is object, integer or float value. (Please refer python jupyter notebook)
2. Then we use describe function to summarize our data it tells us the mean, standard deviation, IQR, and summary of numeric columns. (Please refer python jupyter notebook)
3. Then we use distplot or density plot to check the normality. Normality means whether the data is normally distributed or not.

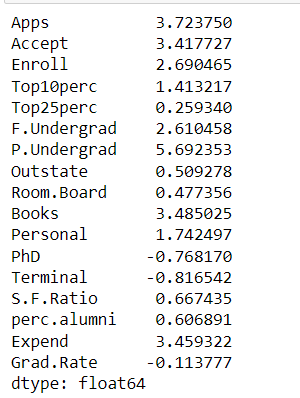




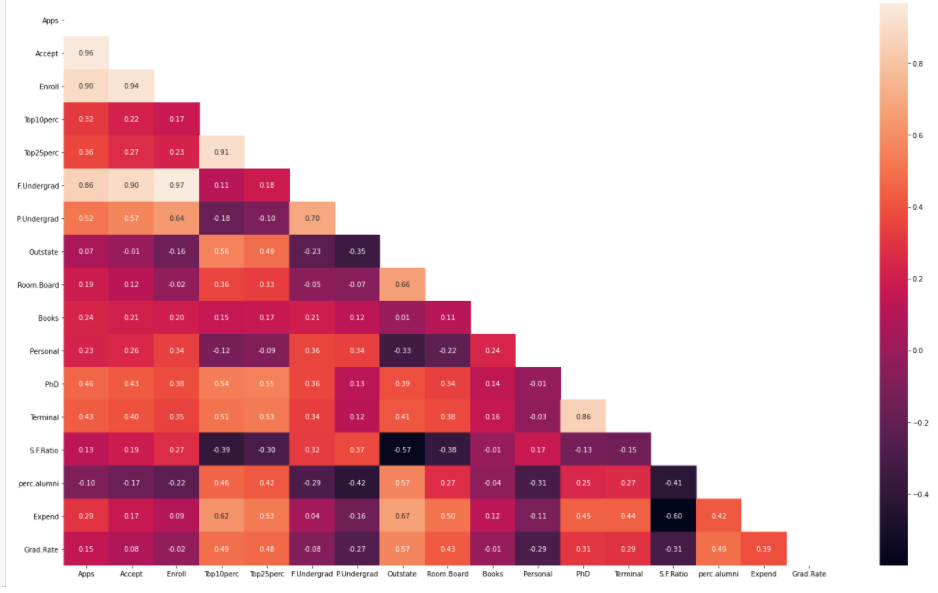




1. To understand which variable in the data set is normally distributed and which is not we use skewness. If the skewness =0, It is said to be normally distributed, if it is >0 it is left skewed and if it < 0 it is skewed towards right.



1. After that, we do the multivariate analysis like we do the correlation and heatmap

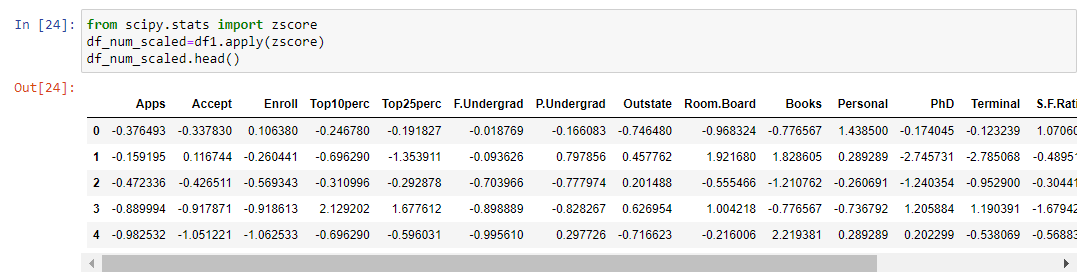


* 1. Is scaling necessary for PCA in this case? Give justification and perform scaling.

Yes, it is necessary to perform scaling for PCA. For instance, in given data set, applications and other variables are having values in thousands and few variables such as percentile is in just two digits. So, the data in these variables are of different scales, it is tough to compare these variables.

The PCA calculates a new projection of the given data set and the new axis are based on the standard deviation of the variables. So, a variable with a high standard deviation in the data set will have a higher weight for the calculation of axis than a variable with a low standard deviation. By performing scaling, we can easily compare these variables.

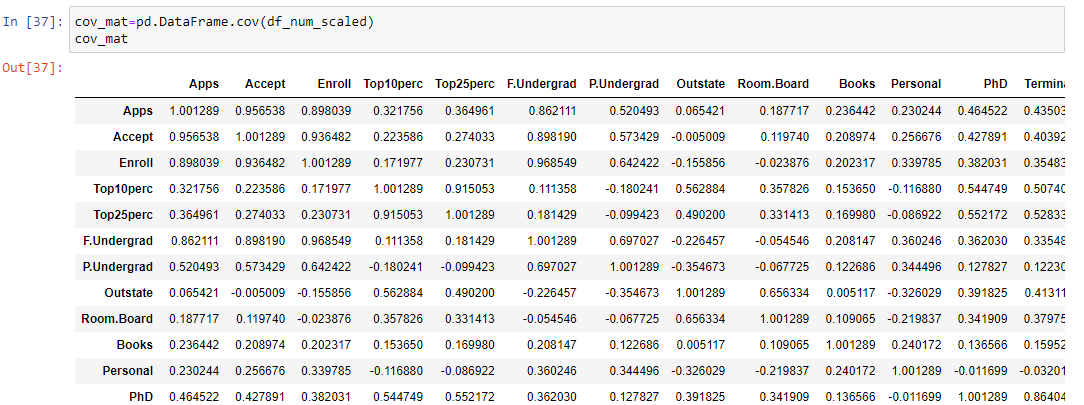
Before standardizing we need to remove the outliers which are present in the dataset.

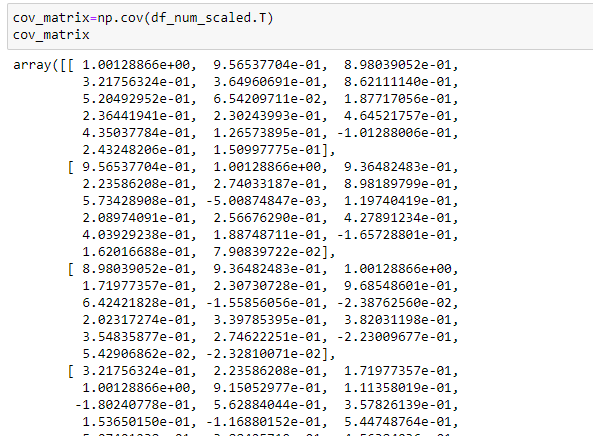


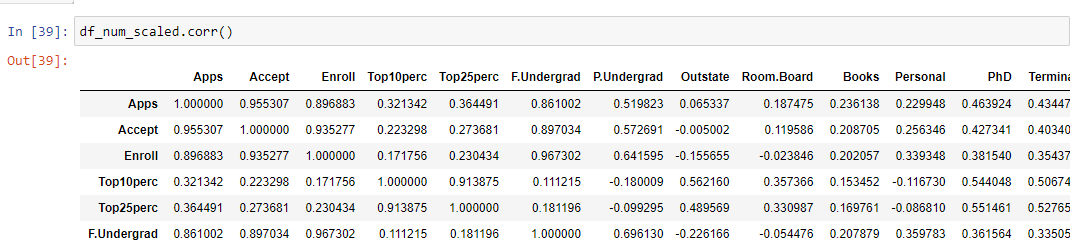
* 1. Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].

Correlation is a scaled version of covariance; note that the two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be positively correlated; when the sign is negative, the variables are said to be negatively correlated; and when the sign is 0, the variables are said to be uncorrelated.

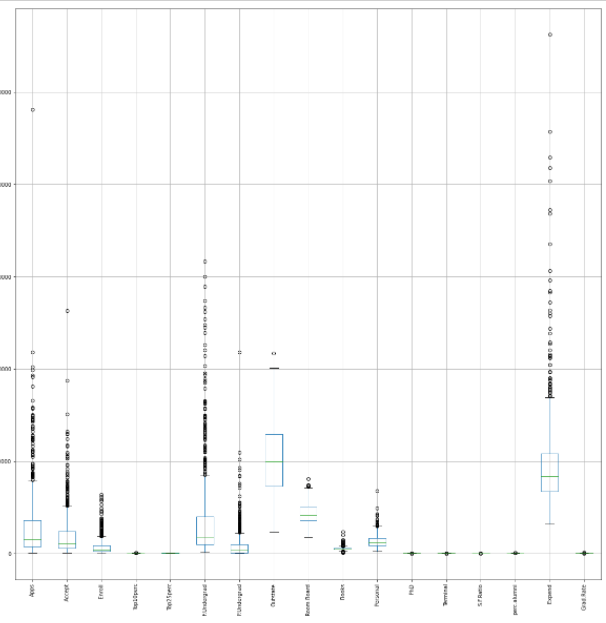
In simple sense correlation, measures both the strength and direction of the linear relationship between two variables Covariance is a measure used to determine how much two variables change in tandem. It indicates the direction of the linear relationship between variables.

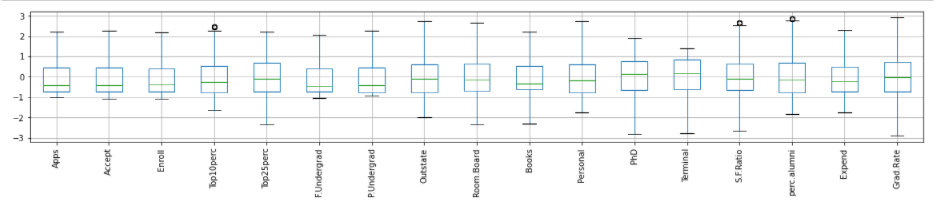






* 1. Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]
* By scaling, all variables have the same standard deviation, thus all variables have the same weight and thus resulting in PCA calculating relevant axis.
* Before scaling, we only had one variable with no outliers (top25 perc); Post scaling, we have multiple variables with negligible outliers – this is achieved by normalizing the scale of the variables.
* Please refer to jupyter notebook for reference.

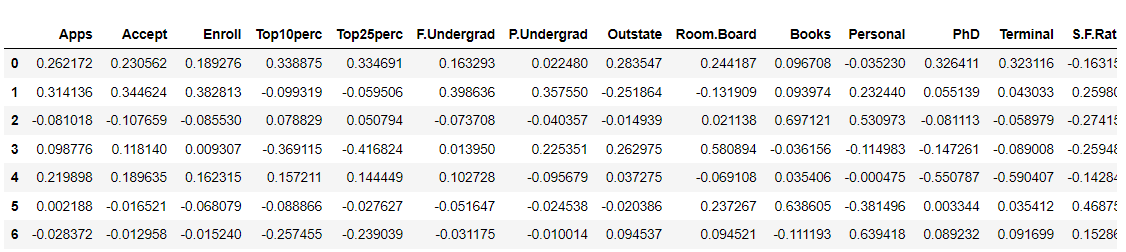




* 1. Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

Eigenvalue and Eigenmatrix are mainly used to capture key information that stored in a large matrix.

* It provides summary of large matrix.
* Performing computation on large matrix is slow and require more memory and CPU, eigenvectors and eigenvalues can improve the efficiency in computationally intensive task by reducing dimensions after ensuring of the key information is maintained.
  1. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features



* 1. Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

[-2.62 3.14 8.10 -9.87

-2.19 2.18 -2.83 -8.99

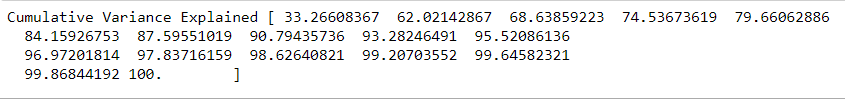
1.30 -1.56 -8.62 1.82

-5.99 8.99 8.88 5.49

5.41]

* 1. Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

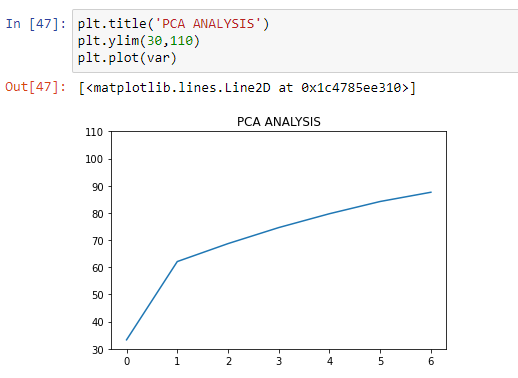
From the below screenshot of cumulative values of the eigenvalues, we can see that around 7 principal components explained over 87% of the variance. Thus, the optimum number of principal components can be 7.



Furthermore, eigenvectors indicate the direction of the principal components, we can multiply the original data by the eigenvectors to re-orient our data onto the new axes

* 1. Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [**Hint:**Write Interpretations of the Principal Components Obtained]

PCA is a statistical technique and uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. PCA also is a tool to reduce multidimensional data to lower dimensions while retaining most of the information. Principal Component Analysis (PCA) is a well-established mathematical technique for reducing the dimensionality of data, while keeping as much variation as possible.



We know that the principal components describe the amount of the total variance that can be explained by a single dimension of the data. We have generated only 7 PCA dimensions. These 7 PCA can be used for further analysis, representing more than 87% of the variance.