|  |
| --- |
| GREATLEARNING |
| ADVANCE STATISTICS |
| **PROJECT REPORT** |
|  |
| **BY,** |
| **RAGAVEDHNI K R** |

|  |
| --- |
|  |

1. A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments.
   1. **State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually. [both statement and statistical form like Ho=mu, Ha>mu]**

**Formulating the Hypothesis for One-way ANOVA for the ingredient** **A:**

* Statement:

H0 = the population means of all the three levels in the ingredient A are same.

H1 = the population mean of at least one of the level in the ingredient A is different.

* Statistical form:

H0: µ1 = µ2 = µ3

H1: not all the means µ1, µ2, µ3 are equal.

**Formulating the Hypothesis for One-way ANOVA for the ingredient** **B:**

* Statement:

H0 = the population means of all the three levels in the ingredient B are same.

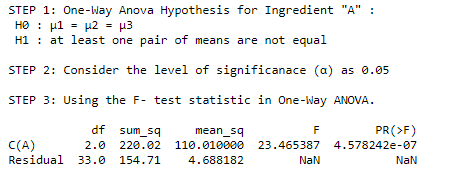
H1 = the population mean of at least one of the level in the ingredient B is different.

* Statistical form:

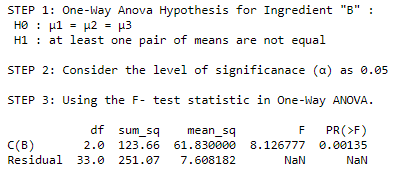
H0: µ1 = µ2 = µ3

H1: not all the means µ1, µ2, µ3 are equal.

* 1. **Perform one-way ANOVA for variable ‘A’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.**



* The p-value of the ingredient A (4.578242e-07) is less than the level of significance; we can reject the Null Hypothesis.
* Hence, we have evidence that there is difference in the mean of three levels in Ingredient A and not all of the means are same.
  1. **Perform one-way ANOVA for variable ‘B’ with respect to the variable ‘Relief’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.**

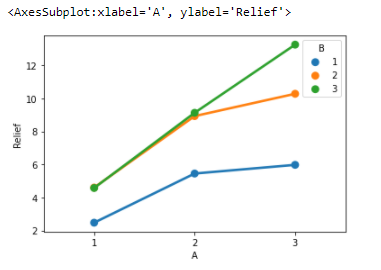


* The p-value of the ingredient B is less than the level of significance; we can reject the Null Hypothesis.
* Hence, we have evidence that there is difference in the mean of three levels in the

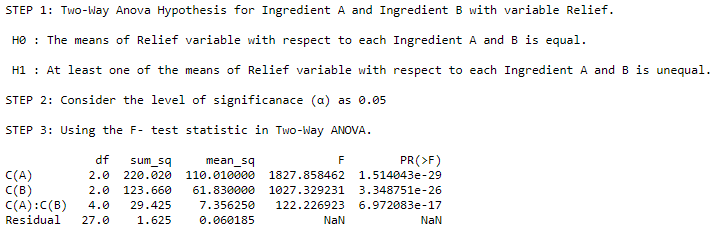
Ingredient B and not all of the means are same.

* 1. **Analyze the effects of one variable on another with the help of an interaction plot.**

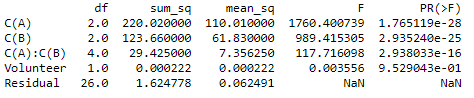
**What is the interaction between the two treatments?**



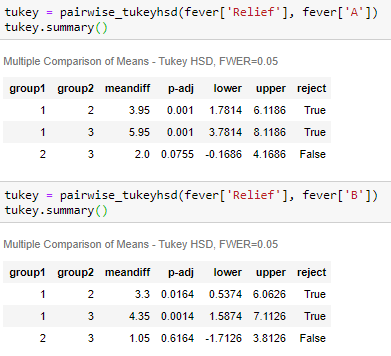
* The point plot shows the interactions of the one categorical variable across different levels of another categorical variable.
* From the above point plot, it is evident that both the Ingredient 'A' and Ingredient 'B' are interacting with each other across three different groups/levels, because the distances between the means across the three levels are not the same.
* The group 1 of Ingredient A is very closely interacted with group 2 and group 3 of Ingredient B.
* The group 2 of Ingredient A is slightly interacting with group 2 and group 3 of Ingredient B.
  1. **Perform a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’ along with their interaction 'A\*B') with the variable 'Relief' and state your results.**



* In the Two-Way ANOVA with interaction effect, the p-value for the Ingredient A (P=1.514043e-29) and the p-value Ingredient B (P=3.348751e-26) both are significant (>0.05).
* The interaction between the Ingredient A and Ingredient B (P=6.972083e-17) is less than the level of significance of 0.05.
* Because the interaction effect between Ingredient A and Ingredient B is statistically significant, you cannot interpret the main effects without considering the interaction effect.
* Hence, we can reject the Null Hypothesis, stating that there is significant interaction between the levels of Ingredient A and Ingredient B.
  1. **Mention the business implications of performing ANOVA for this particular case study.**
* ANOVA is used to find the differences among the group means in the data by considering the variation within the data and variation between/among the data.
* The sample size is 36 and in 9 clusters, by performing the ANOVA with F- statistic.
* The ingredients A and B have individual effect with Relief treatment.
* There is interaction effect of the ingredients with Relief treatment.
* The Randomized variable Volunteer is significant with the ingredients A and B and their interaction with Relief.



* Using the pairwise\_tukeyhsd() function, we can see that the comparison of the means of the groups in the ingredient A with Relief is significant, and leads to the rejection of the Null hypothesis
* The comparison of the means of the groups in the ingredient B with Relief variable is significant as their p-values are less than 0.05.
* We can see that all pairwise comparisons across the different groups in ingredient A and ingredient B are significant with their adjusted p-values less than 0.05.

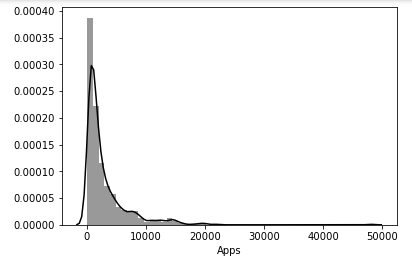
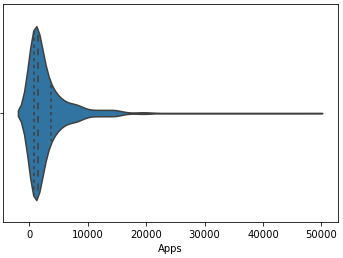


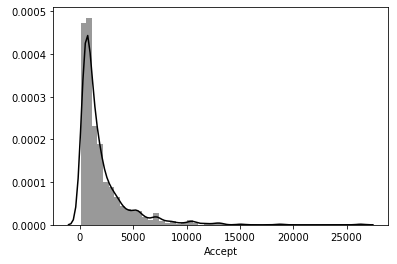
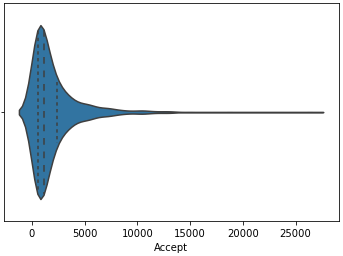
1. **The dataset**[**Education - Post 12th Standard.csv**](https://olympus.greatlearning.in/courses/13136/files/909246/download?verifier=Y3WOpNpJY1AbH2kCyKIXlRg6H0qpJ36fAFHDUTzE&wrap=1)**is a dataset that contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file:**[**Data Dictionary.xlsx**](https://olympus.greatlearning.in/courses/13136/files/909245/download?verifier=Y0n0S1jSl5A0YPz6YfEn45RIpUaDpnRkjfTFDKsi&wrap=1)**.**
   1. **Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented**.

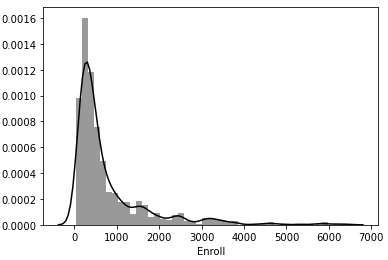
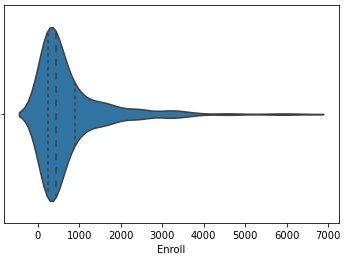
**UNIVARIATE ANALYSIS:**

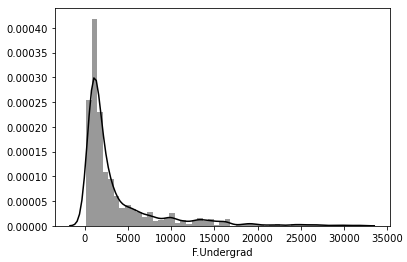
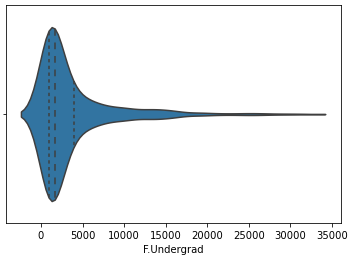
* The plots provide information about the distribution of the observations in the single data variable.
* Here we consider the distplot and violinplot (from sns package) to know the distribution of the individual variables from the dataset.
* The right skewed (or positively skewed) distribution has large occurrence in the left side and few in the right side.
* The left skewed (or negatively skewed) distribution has large occurrence in the right side and few in the left side.
* The symmetric distribution is the bell-shaped or normal distribution.
* Most of variables like Apps, Accept, Enroll, F.Undergrad, P.Undergrad, Books are higly Right Skewed.
* While the variables Personal, Top10perc, S.F.Ratio, Expend, Perc. Alumini are moderately Right Skewed.
* The variables PhD. And Terminal is Left Skewed.
* The variables Top25perc, Outstate, Room Board, Grad. Rate shows almost symmetric distribution.

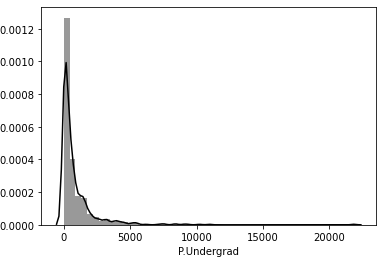
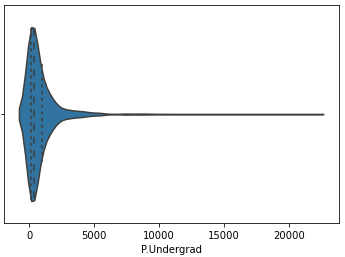
**HIGHLY RIGHT SKEWED VARIABLES:**

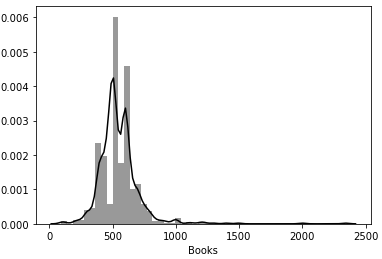
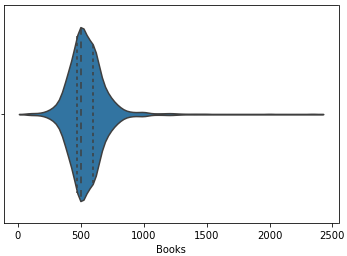
 

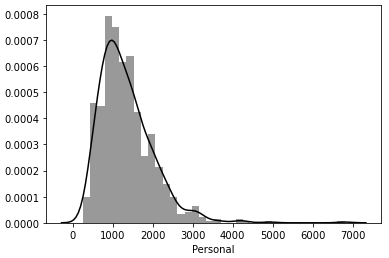
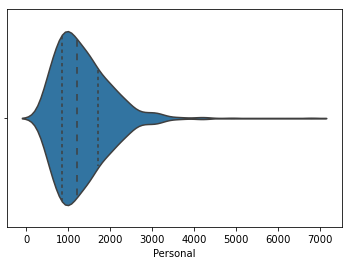
 

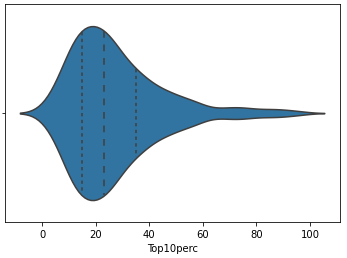
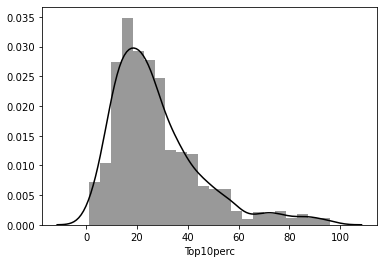
 

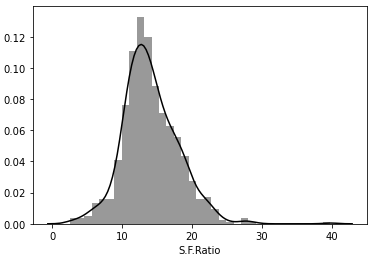
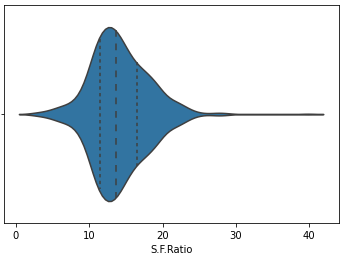
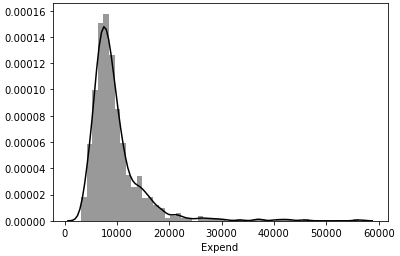
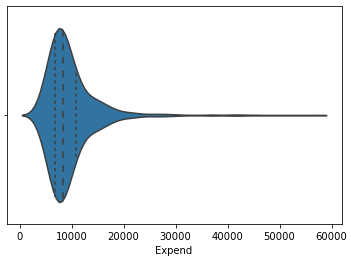
 

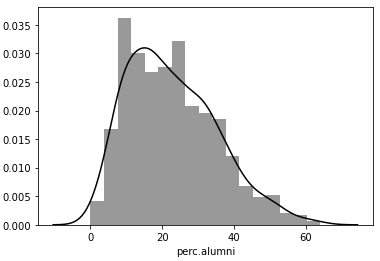
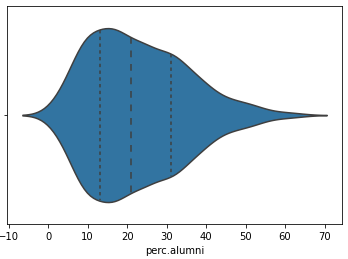
 

**MODERATELY RIGHT SKEWED VARIABLES:**

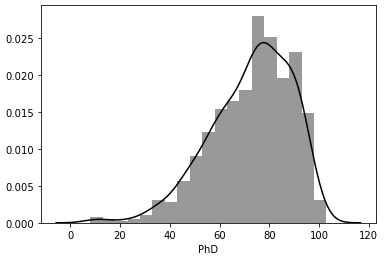
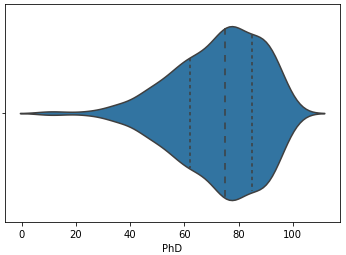
 

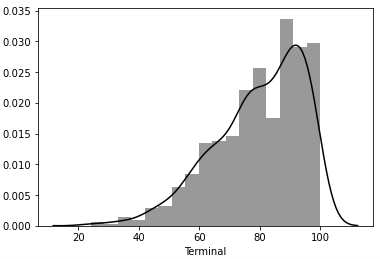
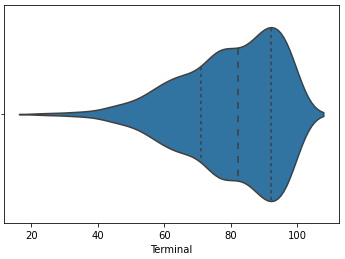


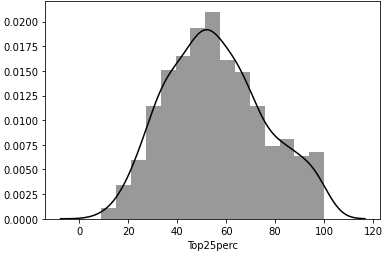
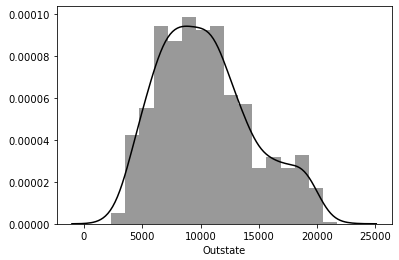
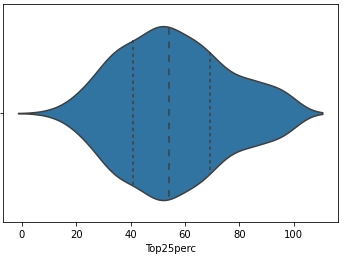
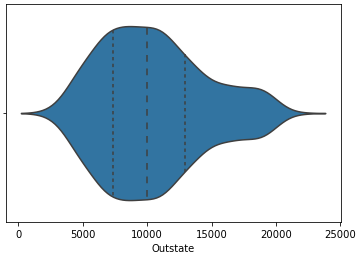
 

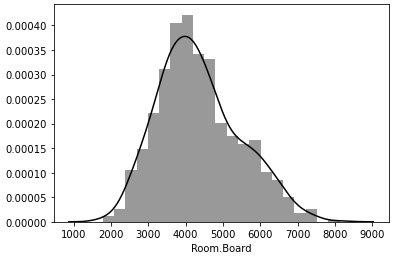
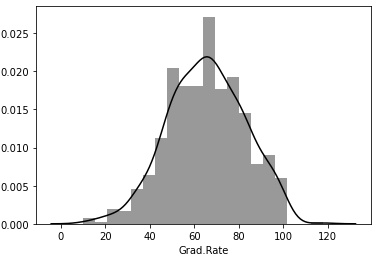
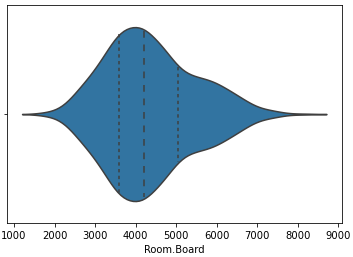
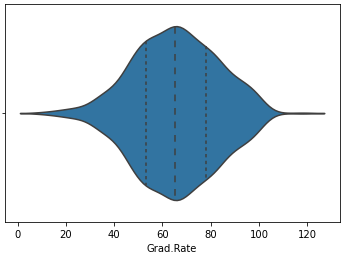
**LEFT SKEWED VARIABLES:**

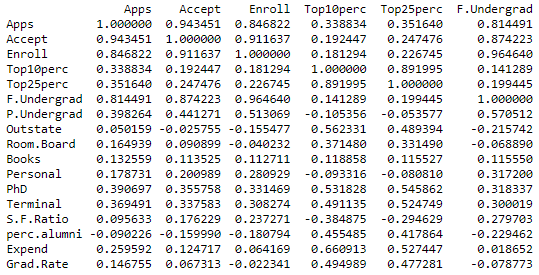
**NEARLY SYMMETRIC VARIABLE:**

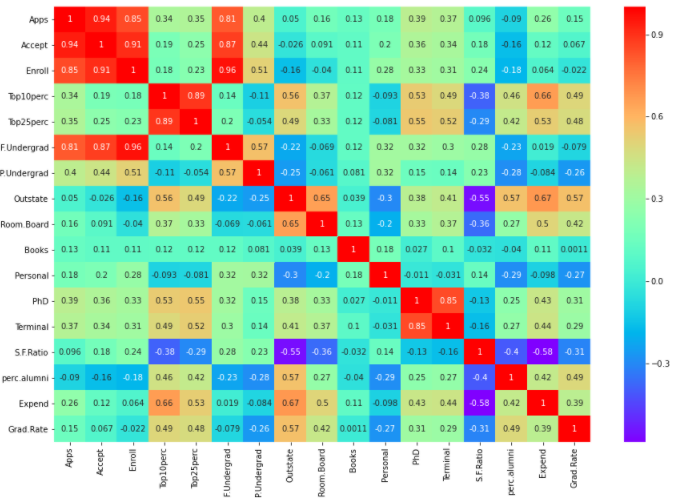
  

**MULTIVARIATE ANALYSIS:**

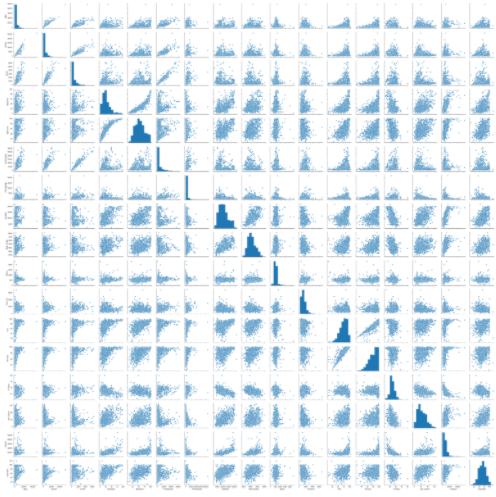
* The pairplot (from sns package) is used for visualizing the pair wise relationship across entire dataframe.
* The correlation across the variables can be found using corr() function in a matrix form.



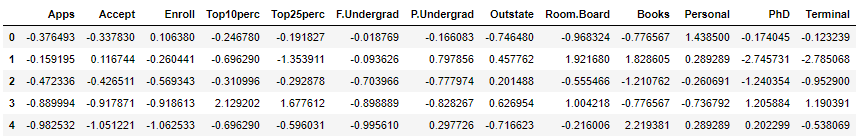
* To indicate and visualize the clusters within the data using the heatmap (from sns package).
* There is high correlation between the number of students who scored the top 10% and top 25% in the Higher Secondary class, as the top 25% will include top 10% scores of students.
* We can see that the colleges accepted more number of applications for the full-time undergraduate students.
* The high number of students has been enrolled for the full-time undergraduate students.
* There is high correlation between the number of applications received and the number of applications accepted, as most of the received applications have been accepted by the colleges.
* Most of the students got enrolled in the college whose applications have been accepted.
* Maximum number of the faculties has their terminal degree in Ph.D.



* The pair wise relationship across the variables can be seen using pairplot function.
* There is a linear relationship between Number of applications received, accepted and Number of students got enrolled in the college.
* Between the Terminal degree and Ph.D., we can see huge relation.
* The enrolled students are mostly Full-time undergraduates.



* 1. **Scale the variables and write the inference for using the type of scaling function for this case study.**
* The dataset does not have any null values and missing values, but it has outliers.
* The Principal Component Analysis (PCA) is sensitive to the outliers and can lead to misleading conclusion in the presence of outliers.
* Hence the outliers are treated using the IQR score method before scaling the data.
* The various scaling techniques are Min-Max scaler, StandardScaler (Z-Score) and Log Transformation.
* The variables are scaled using the Standardization technique (commonly called Z-Score normalization) which has the property of standard normal distribution with mean as 0 and standard deviation as 1.
* The PCA can be misleading with dataset having the variables with different units of measurement and variables having same units but lying in different ranges, as PCA considers variable with maximum unit and variables in maximum range.
* Standardization has to be applied to center the data before applying the PCA, as the First Principal Component must include maximum variance of the data.
* The scaled data as a dataframe can be seen for some variables as below:



* 1. **Comment on the comparison between covariance and the correlation matrix after scaling.**

**Covariance:**

* To determine the direction of the relationship between two variables.
* A positive value indicates that the two variables move in the same direction.
* A negative value indicates that the two variables move in the opposite direction to each other.

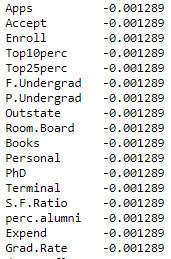
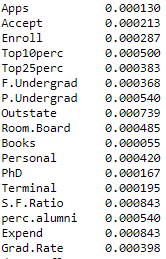
**Correlation:**

* To determine the direction and strength of the relationship between two variables.
* A value closer to -1 or +1 means a strong correlation.
* A value close to zero means weak correlation.

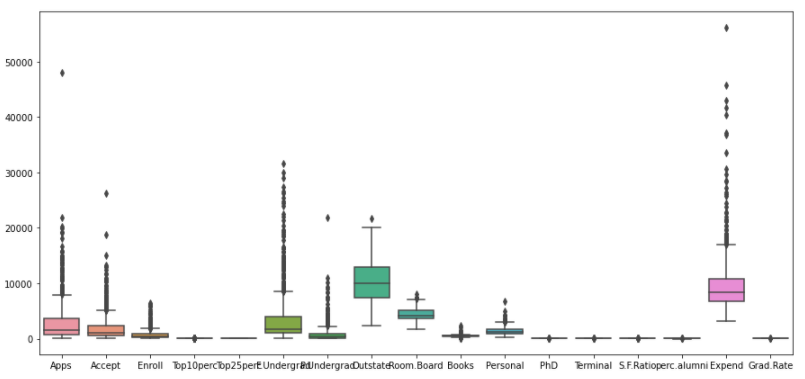
**Comparison between Covariance and Correlation:**

* Correlation is a function of the covariance.
* Covariance is when the 2 variables vary with each other, while correlation is to find whether the change in one variable results in the change in another variable.
* The correlation matrix is the covariance matrix for the standardized variables.
* The Correlation remains unaffected by the change in the scale or dimension.
* Hence the scaled covariance matrix will be same (or almost same) as the correlation matrix.
* In our dataset, we can see that as the data is scaled so there isn't much difference between the correlation matrix and covariance matrix.
* We could see a change from the third decimal place between both the matrices.
* Example: Covariance between the variables 'Apps' and 'Accept' is 0.956538 and Correlation between the variables 'Apps' and 'Accept' is 0.955307.

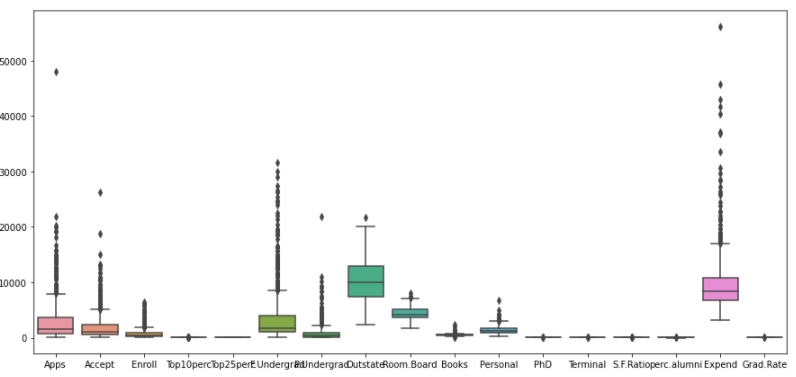
MINIMUM DIFFERENCE: MAXIMUM DIFFERENCE:

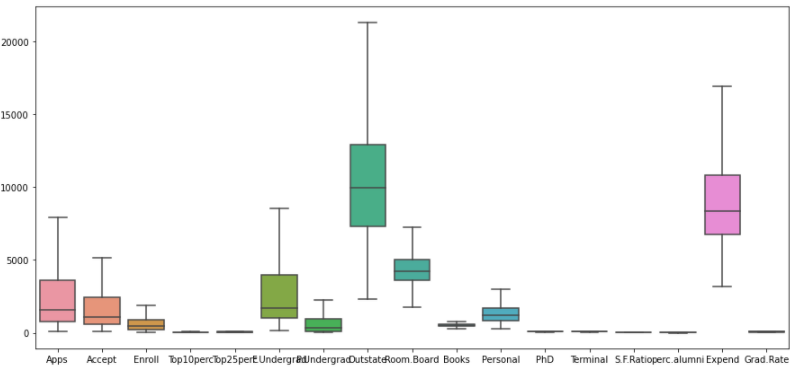
* 1. **Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.**
* The dataset has outliers in almost all the variables; we can know the outliers in the data while we club all the variables together as they depend on each other and project the outliers clearly as seen below.



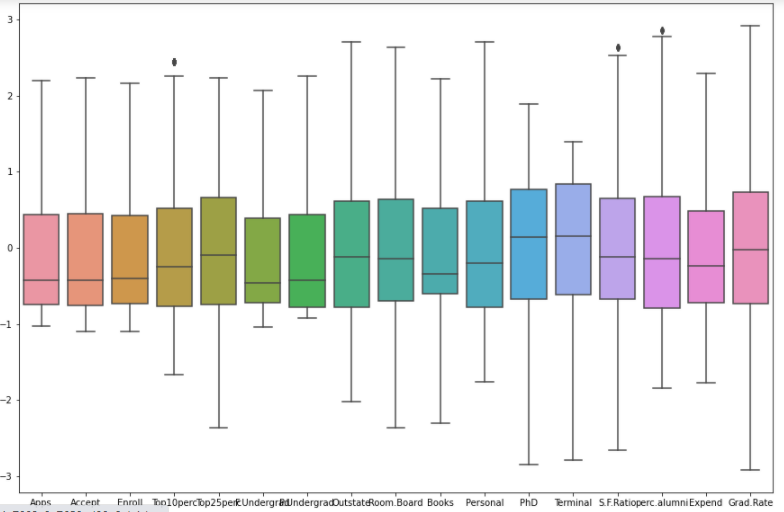
* The Z-score scaling is applied on the data and checked for the outliers, we can see that the outliers are not treated by applying scaling on the data as seen below.



* The outliers are treated by the IQR score method, by replacing the outliers either by the lower whiskers or upper whiskers of that variable based on the position of the outliers.

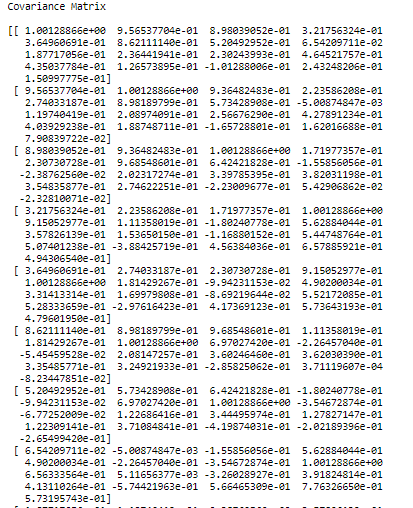


* Now scaling is applied on the data after treating the outliers, we can see that there is countable number of outliers in the scaled data.

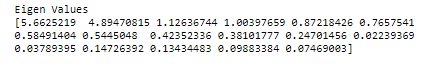


* 1. **Build the covariance matrix, eigenvalues, and eigenvector.**

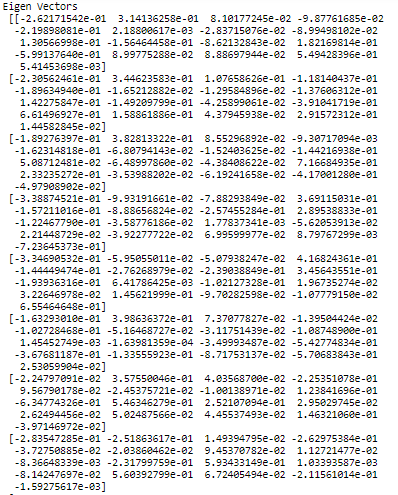
**Covariance matrix:**



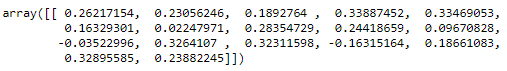
**Eigen Values:**



**Eigen Vectors:**

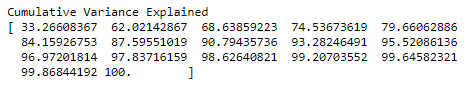


* 1. **Write the explicit form of the first PC (in terms of Eigen Vectors).**
* By matrix multiplication of the original data with the eigenvectors, we get an equation that can be used to evaluate the first principal component.
* PC1 = EV[0][0]\* Col1 + EV[0][1]\*Col2+…..+ EV[0][9]\*Cov10 where EV is Eigen Vector.
* To know the explicit form of first PC, as it holds the maximum coverage of the variance across the columns.
* This can be achieved by applying the PCA model on the scaled data frame.
* Considering the number of PCA components as 1 and apply the fit\_transform function on the scaled data.
* Reading the PCA components, we get the explicit form of the first PC as shown below.

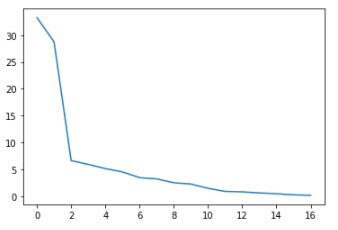


* 1. **Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame.**

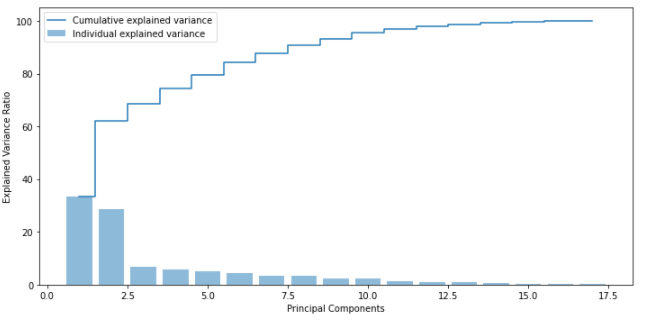
**CUMULATIVE VALUES OF EIGEN VALUES:**



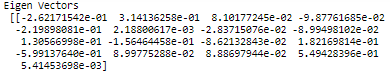
* The Cumulative percentage value gives the percentage of variance accounted for the n components. For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components.
* It helps in deciding the number of components by selecting the components which explained the high variance.
* In the above array we see that the first feature explains 33.3% of the variance within our data set while the first two explain 62.1 and so on.
* If we employ 8 features we capture nearly 90% of the variance within the dataset, thus we gain very little by implementing an additional feature (think of this as diminishing marginal return on total variance explained).



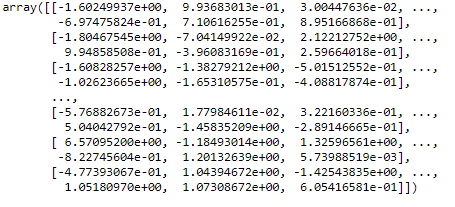
* Consider the number of components until the line gets flat in the Scree plot.
* Below is the plot for viewing the Individual Explained Variance and Cumulative Explained Variance for the Principal Components across the Explained Variance Ratio.



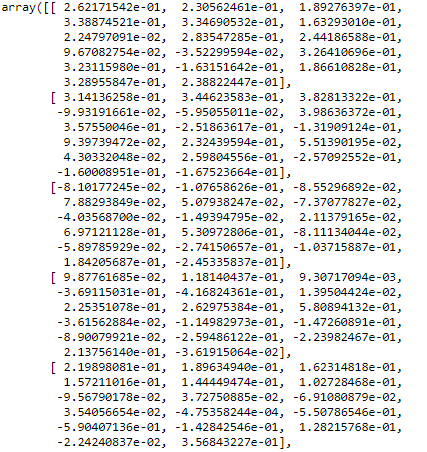
* Eigen Vectors Indicate the amount of coverage of variance of the Principal components with respect to each column in the dataframe.



**Performing PCA:**

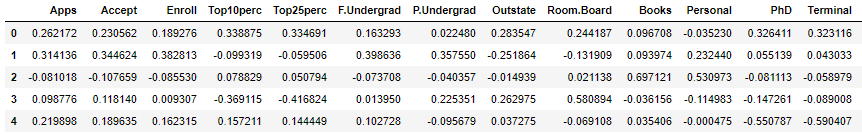


**PCA Components:**



**Exporting data of PCA scores into data frame:**

* Create a dataframe with the pca components as values and use the columns from the scaled dataframe.



* 1. **Mention the business implication of using the Principal Component Analysis for this case study.**

**Business Implications of using PCA in general:**

* 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.
* This PCA can only be done on continuous variables.
* The given dataset is checked with null values, missing values and outliers and outliers are treated with IQR scale method.
* The scaling is made using Z-Score normalization, so we get both the covariance and correlation matrix as same.
* The eigen vectors and eigen values are calculated.
* The PCA is performed and we take 8 components with 90% variation captured.

**Business Implications of using PCA for this case study:**

* This heatmap and the color bar basically represent the correlation between the various feature and the principal component itself
* The first Principal component looks more related to the variables - Top10perc, Top25perc, PhD, Terminal, Expend.
* As we know already that the variables Top10perc and Top25perc are highly correlated, also the variables PhD and Terminal are closely related.
* So the Component 1 picks the closely related variables.
* Component 2 looks more related to Books and Personal - We can label it as Expenditure property
* The Component 3 looks very closely related to Books.
* Depending on relationship, we could go ahead and label relationship with features.

