**Problem 1:**

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.greatlearning.in/courses/33034/files/2327707/download?verifier=7VloqUJAA6lwETySS1bsaZsdwxufxLyDLpNxBXLB&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.]

|  |
| --- |
| **Preliminary Analysis** |
| **Shape:** We have 40 rows and 3 columns with no null values. |
| **Salary Variable Analysis:**   * Mean and the Median seem to be near. * The 25%, 50% and 75% indicate that the Response variable 'Salary' could be Normally Distributed. * Pictorially the Salary variable seems to have a near to normal distribution. * Minimum and Maximum values can be visualized along with the mean and median. * No outliers exist for Salary variable, data seems to be clean enough to proceed. |
| **Data frame information:**   * Columns: Education and Occupation are of Object type; we can infer them to be Categorical * Column: Salary is of type numeric. |
| **Factor1: Occupation**  **Occupation Levels:**   * It was identified above that we have 4 Levels of Occupation. * 13 people have Occupation Prof-Speciality * 12 people have Occupation Sales * 10 people have Occupation Adm-clerical * 5 people have Occupation Exec-managerial   **Salary Pattern for the Occupation:**  Means were different with Exec-Managerial with a higher Salary value stating that the people with this, Occupation seem to earn better and Occupation Sales seem to earn less.    **Visual comparison of means for Occupation:**   * Sales seem to have hight salary and a better median. * Prof-Speciality has wide spread of people with different salaries and it seem to be right skewed. |
| **Factor2: Education**  **Education Levels:**   * It was identified above that we have 3 Levels of Education. * 16 people were Doctorate * 15 people were Bachelors * 9 people were HS-Grad   **Salary pattern for the variable Education:**  Means were different with Doctorate with a higher Salary value stating that the people with this, Education seem to earn better, and Education HS-Grad seem to earn less.  **Visual comparison of means for Education:**   * Bachelors education has nice spread of Salaries. * Doctorate seems to have most of the data on its left, suggesting left skewness. |

|  |
| --- |
| 1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. |
|  |
| 1.2 Perform one-way ANOVA for Education with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results. |
| * The F-Statistic is approximately 31 and P Value is highly significant. * Based on the ANOVA test we, reject the null hypothesis that the three-population means are Identical. * At least for one Education mean the Salary is different from the rest. |
| 1.3 Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results. |
| * The F-Statistic is approximately 0.884 and P Value is not highly significant. * Based on the ANOVA test we, fail to reject the null hypothesis and the four-population means are Identical. Occupation does impact the Salary. |
| 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. |
| We perform Tukey test:  For Education:   * None of the means are equal as the P value is significant for the all mean comparisons. * That makes us to reject the Null Hypothesis i.e. Equality of all population means is rejected for Education.   For Occupation:   * P value is not significant for all mean comparisons. * That makes us to fail to reject the Null Hypothesis i.e. Equality of all population means is accepted for Occupation. |
| 1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot. |
| * Lines are not parallel. * Mean Salary is lowest for all Occupations with HS-Grad except for Exec-managerial Occupation. * This indicates interaction effect between Education and Occupation. |
| 1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education\*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result? |
| Hypothesis:   * Factor Education (row) effects ->   + Ho: mU1 = mU2 = mU3   + Ha: Not all mU are equal * Factor Occupation (row) effects ->   + Ho: mU1 = mU2 = mU3 = mU4   + Ha: Not all mU are equal * Interaction effects ->   + Ho: Interaction effect does not exist   + Ha: Interaction Effect exists.   Conclusion of Two-Way Anova is:   * Out of 3 hypothesis 2 are significant and 1 is not. * Significance is tested at 5% level. * We reject Null Hypothesis; Group means are equal for Education. * We reject Null Hypothesis; Equality of means at each combination of Education and Occupation. Hence there will be an interaction effect on Salary wrt to Education and Occupation. * We fail to reject the null hypothesis; Group means are equal for Occupation.   Summary:  Salary of a person is bound to be dependant on the Occupation and Occupation along with Education seem to have an interaction effect on the Salary. However, Education alone does not have any effect on Salary. |
| 1.7 Explain the business implications of performing ANOVA for this particular case study. |
| * Highly paid individuals in a corporate company seem to be in Higher Positions. Our Anova results show, a person with Higher Occupation draws more Salary than others. * Combination of Occupation and Education seem to have effect on Salary of individual, a person with good education in higher position seems to earn more Salary. * Education alone does not fetch good Salary. |

**Problem 2:**

The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/33034/files/1739362/download?verifier=e0UhTRC80UE0jBKdRBnK9E0JMZ5t8aaUWHL1KlH2&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.greatlearning.in/courses/33034/files/1739361/download?verifier=NFA4eBpTX3tEnAyjhU0od2RAbVAlkDEJXwLU9hec&wrap=1).

|  |
| --- |
| **Preliminary data analysis** |
| * Data has 777 rows and 18 Columns. * There are no Null Values in the dataset. * Other than Names attribute in the dataset, other attributes are either an int64 or float64. * There are no duplicate rows in the data set. * Data set seems to be clean and we can proceed with EDA. |

|  |
| --- |
| 2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? |
| Univariate Analysis Insights:   * Univariate analysis was performed for 16 numeric columns. * Distribution plot was used to check the normal distribution of attributes. * Box plot was used to verify outliers, median and IQR. * Apps – Number of applications received has a right skewed distribution and it had no outliers. * Accept – Number of applications accepted has similar distribution as Apps without any outliers. It specifies that almost all applications were accepted. * Enroll – Number of applications enrolled also has similar distribution as Accept and Apps. It specified that the applications submitted were enrolled without discarding any. * Top10Perc – 10% Toppers from Higher Secondary class, seems to have a right skewed distribution with a sharp decline in the toppers getting enrolled compared to Top25Perc. * Top25Perc – 25% Toppers from Higher Secondary Class, also seem to have almost normal distribution, with slight right skewedness. This mentions that almost 70% of the top 25% students were enrolled. * Number of Full-time Undergrad students, seem to fall under almost normal distribution with a right tail. Some colleges seem to have huge number of full time students causing the outliers. * Number of Part-Time Undergrad students, seem to fall under normal distribution with outliers, specifying, some colleges seem to have large number of part-time students causing the outliers. * Outstate students seem to have normal distribution, with a neat spread of out of state students in all colleges. * Most colleges seem to have good number of Class Rooms with Boards, with almost all colleges having the Rooms with boards. * Books cost seem to be high for some colleges causing the outliers. * Personal Spending for some colleges seem to be high causing outliers. * Faculties with PhD seem to have left skewed distribution, stating many colleges have less to average number of faculties with PhD Qualification. With negative outliers, this states some colleges never have a PhD faculty. * Terminal degree also follows the PhD distribution. * Student to Faculty ratio has normal distribution, stating, some colleges have better S/F ratio and other colleges are also not bad. * Good population of alumni donate, but they do donate well causing outliers. * Instructional exp per student seem to be high for some colleges causing outliers, other colleges seem not to have this. * Graduation Rate for colleges follows normal distribution, specifying enrolled students graduating out of colleges every year.     Bi Variate Analysis:   * With Pearson method for Correlation, below is understood. * Apps, Accept, Enroll, F.Undergrad attributes have strong positive correlation. * Top 10 and 25 perc columns have strong positive correlation with each other. * P.Undergrad has average correlation quotient with F.Undergrad and Enroll. Other attributes seem to weakly correlate. * Outstate does not have strong positive correlation with any, but Expend, Room.Rate, Graduate have average effect on Outstate. * Books, Personal does not have strong correlation with any. * PhD and Terminal have strong positive correlation with each other. * SF Ratio, Expend, Grad.Rate have average factors affecting the correlation but, nothing strongly affects them. * Pair plot shows above correlation in pictorial method. * Summary: There are positively correlated attributes in the dataset, we need to identify the components that impact most of the correlation. |
| 2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling. |
| Yes Scaling is mandatory, i.e. Attributes; Apps, Accept, Enroll, F.Undergrad, P.Undergrad, Expend have huge numbered outliers and the min-max data range is also huge and the medians also vary by large margin between attributes, it’s hard to visualize with this setup. It’s good to bring down all attributes to single scale (with outliers also being scaled). Below plot shows the spread of values for all attributes with outliers and min-max range. |
| 2.3 Comment on the comparison between the covariance and the correlation matrices from this data. |
| “Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. You can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values. In our case both matrices are similar. Stating Covariance and the derived Correlation are inline with each other. |
| 2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? |
| As mentioned in 2.2 Yes, there are significant outliers.  Data with outliers before scaling:    Data with Outliers after scaling with Z-Score:    Summary:   * Scaling is needed if you have extreme values for attributes. * Scaling was mandatory here, unscaled attributes have huge min max and median values, which cannot be compared side by side. However, after scaling you can see that the attributes min, max, medians lie almost near to each other and it’s easy to visualize. * After scaling was applied, Outliers were also scaled, GL asked not to remove outliers, so I did not, usually we remove outliers as per EDA process. * Z-Score scaling has scaled the outliers as well. The outliers for all attributes are placed at identical distances from IQR. * Note: I have commented code for removing outliers in the pynb, uncommenting will remove the outliers. |
| 2.5 Perform PCA and export the data of the Principal Component scores into a data frame. |
| Cumulative Variance Explained:    We can observe; top 5 components make up to ~77% of variance. We can use the top 5 and not use the others. We have reduced our dimension from 16 to 5.  Scree Plot:      Principal Component Data Frame: |
| 2.6 Extract the eigenvalues, and eigenvectors. |
| Eigen Values:  array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,  0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,  0.02302787, 0.03672545, 0.31344588, 0.08802464, 0.1439785 ,  0.16779415, 0.22061096])  Eigen Vectors:  array([[-2.48765602e-01, 3.31598227e-01, 6.30921033e-02,  -2.81310530e-01, 5.74140964e-03, 1.62374420e-02,  4.24863486e-02, 1.03090398e-01, 9.02270802e-02,  -5.25098025e-02, 3.58970400e-01, -4.59139498e-01,  4.30462074e-02, -1.33405806e-01, 8.06328039e-02,  -5.95830975e-01, 2.40709086e-02],  [-2.07601502e-01, 3.72116750e-01, 1.01249056e-01,  -2.67817346e-01, 5.57860920e-02, -7.53468452e-03,  1.29497196e-02, 5.62709623e-02, 1.77864814e-01,  -4.11400844e-02, -5.43427250e-01, 5.18568789e-01,  -5.84055850e-02, 1.45497511e-01, 3.34674281e-02,  -2.92642398e-01, -1.45102446e-01],  [-1.76303592e-01, 4.03724252e-01, 8.29855709e-02,  -1.61826771e-01, -5.56936353e-02, 4.25579803e-02,  2.76928937e-02, -5.86623552e-02, 1.28560713e-01,  -3.44879147e-02, 6.09651110e-01, 4.04318439e-01,  -6.93988831e-02, -2.95896092e-02, -8.56967180e-02,  4.44638207e-01, 1.11431545e-02],  [-3.54273947e-01, -8.24118211e-02, -3.50555339e-02,  5.15472524e-02, -3.95434345e-01, 5.26927980e-02,  1.61332069e-01, 1.22678028e-01, -3.41099863e-01,  -6.40257785e-02, -1.44986329e-01, 1.48738723e-01,  -8.10481404e-03, -6.97722522e-01, -1.07828189e-01,  -1.02303616e-03, 3.85543001e-02],  [-3.44001279e-01, -4.47786551e-02, 2.41479376e-02,  1.09766541e-01, -4.26533594e-01, -3.30915896e-02,  1.18485556e-01, 1.02491967e-01, -4.03711989e-01,  -1.45492289e-02, 8.03478445e-02, -5.18683400e-02,  -2.73128469e-01, 6.17274818e-01, 1.51742110e-01,  -2.18838802e-02, -8.93515563e-02],  [-1.54640962e-01, 4.17673774e-01, 6.13929764e-02,  -1.00412335e-01, -4.34543659e-02, 4.34542349e-02,  2.50763629e-02, -7.88896442e-02, 5.94419181e-02,  -2.08471834e-02, -4.14705279e-01, -5.60363054e-01,  -8.11578181e-02, -9.91640992e-03, -5.63728817e-02,  5.23622267e-01, 5.61767721e-02],  [-2.64425045e-02, 3.15087830e-01, -1.39681716e-01,  1.58558487e-01, 3.02385408e-01, 1.91198583e-01,  -6.10423460e-02, -5.70783816e-01, -5.60672902e-01,  2.23105808e-01, 9.01788964e-03, 5.27313042e-02,  1.00693324e-01, -2.09515982e-02, 1.92857500e-02,  -1.25997650e-01, -6.35360730e-02],  [-2.94736419e-01, -2.49643522e-01, -4.65988731e-02,  -1.31291364e-01, 2.22532003e-01, 3.00003910e-02,  -1.08528966e-01, -9.84599754e-03, 4.57332880e-03,  -1.86675363e-01, 5.08995918e-02, -1.01594830e-01,  1.43220673e-01, -3.83544794e-02, -3.40115407e-02,  1.41856014e-01, -8.23443779e-01],  [-2.49030449e-01, -1.37808883e-01, -1.48967389e-01,  -1.84995991e-01, 5.60919470e-01, -1.62755446e-01,  -2.09744235e-01, 2.21453442e-01, -2.75022548e-01,  -2.98324237e-01, 1.14639620e-03, 2.59293381e-02,  -3.59321731e-01, -3.40197083e-03, -5.84289756e-02,  6.97485854e-02, 3.54559731e-01],  [-6.47575181e-02, 5.63418434e-02, -6.77411649e-01,  -8.70892205e-02, -1.27288825e-01, -6.41054950e-01,  1.49692034e-01, -2.13293009e-01, 1.33663353e-01,  8.20292186e-02, 7.72631963e-04, -2.88282896e-03,  3.19400370e-02, 9.43887925e-03, -6.68494643e-02,  -1.14379958e-02, -2.81593679e-02],  [ 4.25285386e-02, 2.19929218e-01, -4.99721120e-01,  2.30710568e-01, -2.22311021e-01, 3.31398003e-01,  -6.33790064e-01, 2.32660840e-01, 9.44688900e-02,  -1.36027616e-01, -1.11433396e-03, 1.28904022e-02,  -1.85784733e-02, 3.09001353e-03, 2.75286207e-02,  -3.94547417e-02, -3.92640266e-02],  [-3.18312875e-01, 5.83113174e-02, 1.27028371e-01,  5.34724832e-01, 1.40166326e-01, -9.12555212e-02,  1.09641298e-03, 7.70400002e-02, 1.85181525e-01,  1.23452200e-01, 1.38133366e-02, -2.98075465e-02,  4.03723253e-02, 1.12055599e-01, -6.91126145e-01,  -1.27696382e-01, 2.32224316e-02],  [-3.17056016e-01, 4.64294477e-02, 6.60375454e-02,  5.19443019e-01, 2.04719730e-01, -1.54927646e-01,  2.84770105e-02, 1.21613297e-02, 2.54938198e-01,  8.85784627e-02, 6.20932749e-03, 2.70759809e-02,  -5.89734026e-02, -1.58909651e-01, 6.71008607e-01,  5.83134662e-02, 1.64850420e-02],  [ 1.76957895e-01, 2.46665277e-01, 2.89848401e-01,  1.61189487e-01, -7.93882496e-02, -4.87045875e-01,  -2.19259358e-01, 8.36048735e-02, -2.74544380e-01,  -4.72045249e-01, -2.22215182e-03, 2.12476294e-02,  4.45000727e-01, 2.08991284e-02, 4.13740967e-02,  1.77152700e-02, -1.10262122e-02],  [-2.05082369e-01, -2.46595274e-01, 1.46989274e-01,  -1.73142230e-02, -2.16297411e-01, 4.73400144e-02,  -2.43321156e-01, -6.78523654e-01, 2.55334907e-01,  -4.22999706e-01, -1.91869743e-02, -3.33406243e-03,  -1.30727978e-01, 8.41789410e-03, -2.71542091e-02,  -1.04088088e-01, 1.82660654e-01],  [-3.18908750e-01, -1.31689865e-01, -2.26743985e-01,  -7.92734946e-02, 7.59581203e-02, 2.98118619e-01,  2.26584481e-01, 5.41593771e-02, 4.91388809e-02,  -1.32286331e-01, -3.53098218e-02, 4.38803230e-02,  6.92088870e-01, 2.27742017e-01, 7.31225166e-02,  9.37464497e-02, 3.25982295e-01],  [-2.52315654e-01, -1.69240532e-01, 2.08064649e-01,  -2.69129066e-01, -1.09267913e-01, -2.16163313e-01,  -5.59943937e-01, 5.33553891e-03, -4.19043052e-02,  5.90271067e-01, -1.30710024e-02, 5.00844705e-03,  2.19839000e-01, 3.39433604e-03, 3.64767385e-02,  6.91969778e-02, 1.22106697e-01]]) |
| 2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). |
| array([-0.25, 0.33, 0.06, -0.28, 0.01, 0.02, 0.04, 0.1 , 0.09,  -0.05, 0.36, -0.46, 0.04, -0.13, 0.08, -0.6 , 0.02]) |
| 2.8 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? |
| Cumulative Variance Explained:  It is derived from the cumulative sum on the eigen values, below are the results.  Cumulative Variance Explained:  [ 32.0206282 58.36084263 65.26175919 71.18474841 76.67315352  81.65785448 85.21672597 88.67034731 91.78758099 94.16277251  96.00419883 97.30024023 98.28599436 99.13183669 99.64896227  99.86471628 100. ]  First 5 components of 16, contribute to ~ 77% variance. Components that explain 75% is found to be sufficient to explain the variance of the attributes in a dataset. So we go ahead with 5 Principal Components, there by, reducing the dimension from 16 to 5.  Eigen Vectors:  An Eigen Vector corresponds to a real non-zero eigen value, it points in a direction in which it is stretched by the transformation and the eigen value is factor by which it is stretched. If the eigen value is negative the direction is reversed. |
| 2.9 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] |
| Highlighted Correlations w.r.t to Principal Components:    Business Implications:   * Business can use the above PCA which concludes that the 77% variance between attributes is achieved with just 5 Principal Components. * To achieve 85% variance, we need to use 7 principal components and 10 PCs to achieve 90%. From 5 principal component onwards, more components are required to achieve a higher variance. * That said, majority of the variance can be explained with 5 PCs, so we go ahead with 5 principal components. * Business can use PCs to understand financial impact considering the major contributing factors using Principal Components. * PC1 Interpretation: Explains Business’s expenditure on an Outstation Candidate. For PC1, Loadings for “Outstate” and “Expend” seem to be more than other attributes. * PC2 Interpretation: Explains the Undergrad’s Alumni Donations for the business. For PC2, Loadings for Apps, Accept, Enroll, F.Undergrad, P.Undergrad, perc.alumni seem to be higher than other attributes. * PC3 Interpretation: Explains Student Expenditure on the Student to Faculty Ratio. Business can plan on the Student Expenditure to manage better StoF ratio. For PC3, Loadings for Books, Personal and S.F.Ratio are higher than other attributes. * PC4 Interpretation: Explains the Impact on Graduation rate w.r.t Faculty Qualification. Business can plan to hire Faculty with good credentials for a better Graduation Rate. For PC4, Loadings for PhD, Terminal and Grad.Rate are higher than other attributes. * PC5 Interpretation: Explains the Room and Board cost for the Top 10 and 25 percentage students that enrolled the Business. It can plan for as much room with boards to get more enrolments from the top scored individuals. For PC5, Loadings for Top10Perc, Top25Perc and Room.Board are higher than other attributes. * Columns after applying PCA: ['pc\_outstate\_student\_expenditure',   'pc\_undergrad\_alumni\_donations',  'pc\_faculty\_cost\_student\_expend',  'pc\_faculty\_qualification\_student\_graduation\_rate',  'pc\_room\_board\_cost\_for\_top\_students'] |