Advanced Statistics

ANOVA,EDA and PCA

Submitted by:
Rahul Bilugumba Raghunath

Problem 1A:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] 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.

1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.
2. 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.
3. 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.
4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. **(Non-Graded)**

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

Answer:

The Null and Alternate Hypothesis for the One Way ANOVA for Education are:
H0: The mean Salary variable for each educational level is equal
Ha1: For at least one of the means of Salary for level of Education is different

The Null and Alternate Hypothesis for the One Way ANOVA for Occupation are:
H0: The mean Salary variable for each Occupation type is equal
Ha2: For at least one of the means of Salary for type of Occupation is different
Where Alpha = 0.05
If the p-value is < 0.05, then we reject the null hypothesis.
If the p-value is >= 0.05, then we fail to reject the null hypothesis.

2. 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
Answer :

	df	sum_sq	mean_sq	F	PR(>F)
Education	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

The null hypothesis is rejected as P value is less than 0.05

3.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.

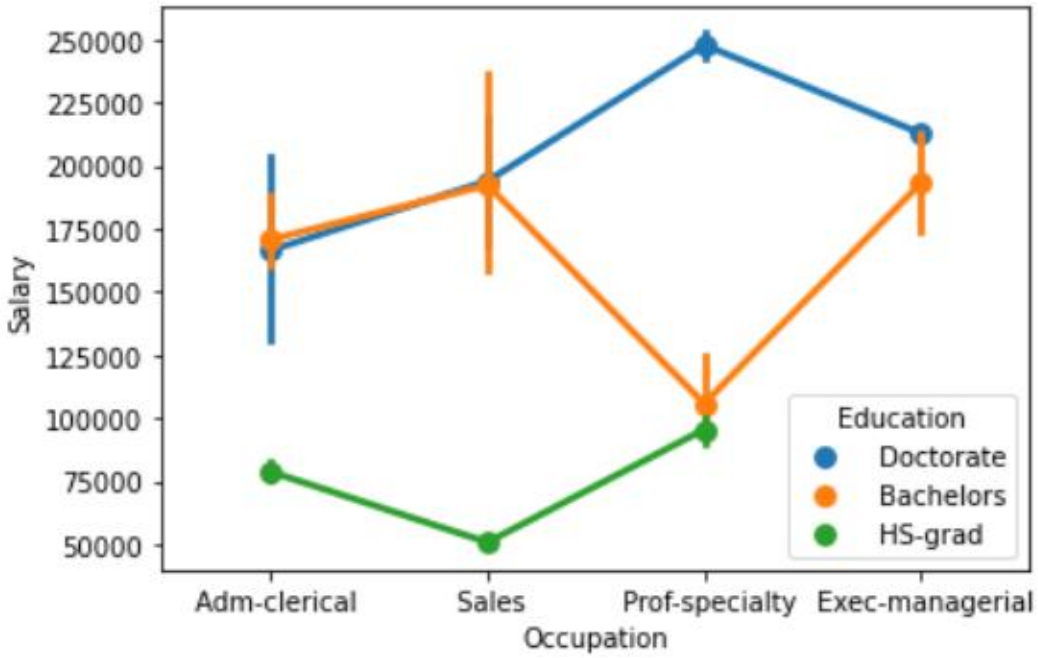
```
In [13]: ## Question 3
formula='Salary~Occupation'
model=ols(formula,df).fit()
aov_table=anova_lm(model)
print(aov_table)
```

```
                df        sum_sq       mean_sq         F    PR(>F)
Occupation     3.0  1.125878e+10  3.752928e+09  0.884144  0.458508
Residual      36.0  1.528092e+11  4.244701e+09       NaN       NaN
```

The null hypothesis is accepted as P value is greater than 0.05

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 'pointplot' function from the 'seaborn' function]**



From the point plot we can infer that , Education and Occupation are the parameters that slightly affect the Salary of the individual.

2.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?

```
                             df        sum_sq       mean_sq          F  \
C(Education)                2.0  1.026955e+11  5.134773e+10  72.211958
C(Occupation)               3.0  5.519946e+09  1.839982e+09   2.587626
C(Education):C(Occupation)  6.0  3.634909e+10  6.058182e+09   8.519815
Residual                   29.0  2.062102e+10  7.110697e+08        NaN

                               PR(>F)
C(Education)             5.466264e-12
C(Occupation)            7.211580e-02
C(Education):C(Occupation)  2.232500e-05
Residual                          NaN
```

Ho: There is no interaction between Education and Occupation on Salary

H1: There is some interaction between Education and Occupation on Salary

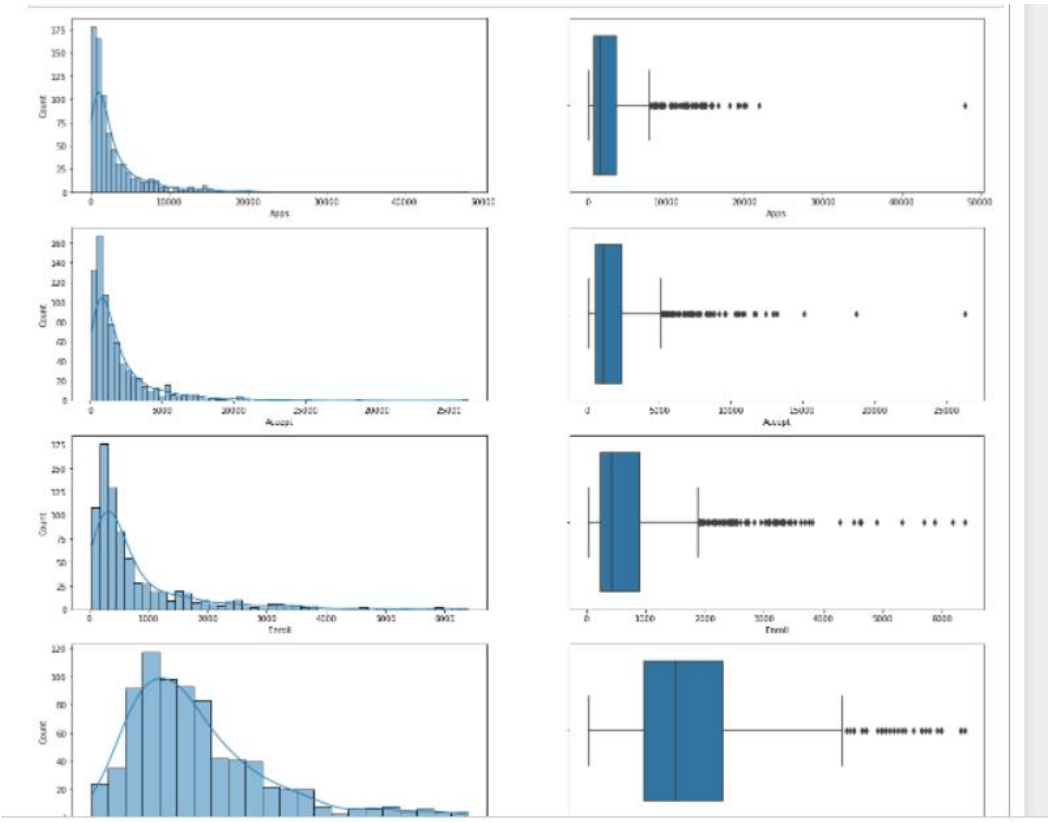
Since the P value is less than 0.05 , we reject the null hypothesis. There is some interaction between education and Occupation on salary .

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

After the ANOVA test we can conclude that Salary is slightly dependent on both Education and Salary . We cannot conclude that salary is not only dependent on one parameter (education or occupation) but dependent on both .

Problem 2 :
The dataset Education - Post 12th Standard.csv 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.

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



Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. There are outliers in the data .

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

Yes it's far vital to do scaling for PCA on this case.

Often the variables of the information set are of various scales i.e. one variable is in tens of thousands and thousands and different in handiest 100. For eg. in our information set many variables are having values in hundreds and in different simply digits. Since the information in those variables are of various scales , it's far hard to compare those variables. We are doing this for the numerical variables.

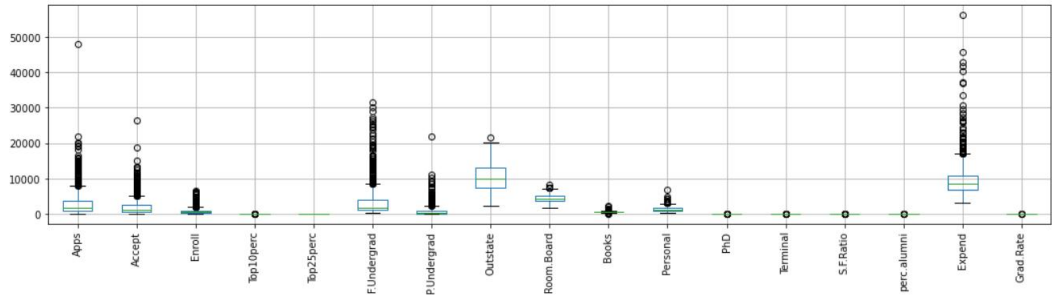
Since scaling is performed handiest on numerical values we can cast off Names Column

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

Correlation, measures each the power and path of the linear dating among variable Covariance is a degree used to decide how tons of variable alternate in tandem, It shows the path of the linear dating among variables.

Covariance is a measure of the joint variability of two random variables.

- **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]**



We can see that there are outliers present. We can see the outliers by using boxplot.

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

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.31344588, 0.22061096, 0.16779415, 0.1439785 , 0.08802464,
       0.03672545, 0.02302787])
```

The above figure represents Extracting eigen values .

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.0630921 , -0.10124906, -0.08298557,  0.03505553, -0.02414794,
        -0.06139298,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131053,  0.26781735,  0.16182677, -0.05154725, -0.10976654,
         0.10041234, -0.15855849,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.07927349,  0.26912907],
       [ 0.00574141,  0.05578609, -0.05569364, -0.39543434, -0.42653359,
        -0.04345437,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
        -0.22231102,  0.14016633,  0.20471973, -0.07938825, -0.21629741,
         0.07595812, -0.10926791],
       [-0.01623744,  0.00753468, -0.04255798, -0.0526928 ,  0.03309159,
        -0.04345423, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
        -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
        -0.29811862,  0.21616331],
       [-0.04248635, -0.01294972, -0.02769289, -0.16133207, -0.11848556,
        -0.02507636,  0.06104235,  0.10852897,  0.20974423, -0.14969203,
         0.63379006, -0.00109641, -0.02847701,  0.21925936,  0.24332116,
        -0.22658448,  0.55994394]])
```

The above figure represents Eigen Vectors

- **Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features**

```
Out[129]: array([[-1.59285540e+00, -2.19240180e+00, -1.43096371e+00, ...,
                  -7.32560596e-01,  7.91932735e+00, -4.69508066e-01],
                 [ 7.67333510e-01, -5.78829984e-01, -1.09281889e+00, ...,
                  -7.72352397e-02, -2.06832886e+00,  3.66660943e-01],
                 [-1.01073537e-01,  2.27879812e+00, -4.38092811e-01, ...,
                  -4.05641899e-04,  2.07356368e+00, -1.32891515e+00],
                 ...,
                 [-7.43975398e-01,  1.05999660e+00, -3.69613274e-01, ...,
                  -5.16021118e-01, -9.47754745e-01, -1.13217594e+00],
                 [-2.98306081e-01, -1.77137309e-01, -9.60591689e-01, ...,
                   4.68014248e-01, -2.06993738e+00,  8.39893087e-01],
                 [ 6.38443468e-01,  2.36753302e-01, -2.48276091e-01, ...,
                  -1.31749158e+00,  8.33276555e-02,  1.30731260e+00]])
```

The above figure represents PCA performed on the data .

```
In [130]: pca.explained_variance_ratio_

Out[130]: array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
                 0.04984701, 0.03558871])
```

After converting the array to a Dataframe.
We obtain

	C1	C2	C3	C4	C5	C6	C7
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944

- **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]**

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.0630921 , -0.10124906, -0.08298557,  0.03505553, -0.02414794,
        -0.06139298,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131053,  0.26781735,  0.16182677, -0.05154725, -0.10976654,
         0.10041234, -0.15855849,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.07927349,  0.26912907],
       [ 0.00574141,  0.05578609, -0.05569364, -0.39543434, -0.42653359,
        -0.04345437,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
        -0.22231102,  0.14016633,  0.20471973, -0.07938825, -0.21629741,
         0.07595812, -0.10926791],
       [-0.01623744,  0.00753468, -0.04255798, -0.0526928 ,  0.03309159,
        -0.04345423, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
        -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
        -0.29811862,  0.21616331],
       [-0.04248635, -0.01294972, -0.02769289, -0.16133207, -0.11848556,
        -0.02507636,  0.06104235,  0.10852897,  0.20974423, -0.14969203,
         0.63379006, -0.00109641, -0.02847701,  0.21925936,  0.24332116,
        -0.22658448,  0.55994394]])
```

- **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?**

```
Out[133]: array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
                 0.81657854, 0.85216726])
```

The Eigenvector is the direction of that line, while the eigenvalue is a number that tells us how the data set is spread out on the line which is an Eigenvector.

We have a dataset of 'n' predictor variables. We centered each predictor's mean and then got an n x n covariance matrix. This covariance matrix is decomposed into eigenvalues and eigenvectors. The covariance matrix is a matrix that details the correlations between different features of a random vector.The Covariance matrix measures how each variable is associated with one another. The Eigenvectors describe the directions of the spread of our data, and the Eigenvalues indicate the relative importance of these directions.

- **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]**

The cumulative percentage gives the percentage of variance accounted for by the n components. For eg. 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.

```
var
```

```
array([32. , 58.3, 65.2, 71.1, 76.6, 81.6, 85.2])
```