2021 | By: Aditya Rishi

Salary and Education

[Statistical analysis of two datasets]

Earn and Learn

Advanced statistics project

Contents

Business challenge 1

1.1 State the null and the alternative hypothesis for conducting one-way ANOVA for both Education and Occupation individually 10

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. 14

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. 17

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. 24

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. 28

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? 30

1.7 Explain the business implications of performing ANOVA for this particular case study. 38

Business challenge 2

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 41

2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling. 71

2.3 Comment on the comparison between the covariance and the correlation matrices from this data. [on scaled data] 74

2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? 75

2.5 Extract the eigenvalues and eigenvectors. [print both] 76

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features.....77

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). 84

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? 84

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] 84

INDEX

Tables

6

Fig 1. Salary data head

Fig 2 Salary data tail

Fig 3 Salary data info

Fig 4 Salary data info

12

Fig 14 Salary education data slicing

Fig 14 Separated levels truncated

14

Fig 16 1-w anova edu

19

Fig 24 occu data sliced, truncated

Fig 25 1-w Anova occu

24

Fig 29 Tukey hsd education

26

Fig 31 Tukey hsd occupation

28

Fig 33 anova interaction

30

Fig 36 anova two-way table

31

Fig 37 model sum 2-way

32

Fig 38 model sum continued

Fig 39 ID data

33

Fig 40 grand mean edu

Fig 41 grand mean occu

Fig 42 mean sal combo

36

Fig 45 crosstab

41

Fig 53 EDA first look

42

Fig 54 EDA 5 pt

61

Fig 87 correlation

64

Fig 89 high accept ratio

Fig 90 low accept ratio

70

Fig 102 variances of var

71

Fig 103 dropping name datasets

72

Fig 105 std dev scaled

73

Fig 106 correlation matrix

Fig 107 covariance matrix

75

Fig 110 create cov mat

Fig 111 eigen vals vecs

78

Fig 114 apply PCA

Fig 115 loading feature

79

Fig 116 principal components

Fig 117 loaded dataset

80

Fig 118 Factor loaded dataset

82

Fig 120 dataset with 6 pcs

Fig 121 names readded

83

Fig 123 cum variances

Charts

7

Fig 5 Nullity graph

8

Fig 5 Salary subplots

9

Fig 8, 9 ,19,11 Salary boxplot, quantiles, descriptive stats, skewness etc. for salary

10

Fig 12 Education countplots

11

Fig 13 Salary distplot

14

Fig 17 Salary dist. 3 levels

15

Fig 18 Sal wrt edu box

16

Fig 19 edu pointplot

17

Fig 20 Occupation countplots

18

Fig 21 kdeplot avg sal occupation

19

Fig 23 Salary data for occupation

Fig 22 avg sal by occupation

21

Fig 26 Salary dist 4 levels occu

22

Fig 27 Salary wrt occu 4 levels

23

Fig 28 Salary on occu pointplot

25

Fig 30 threshold edu

27

Fig 32 threshold occu

28

Fig 34 Ineractionplot edu

29

Fig 35 interactionplot occupation

34

Fig 43 interaction boxplots

35

Fig 44 group sizes

36

Fig 46 fam-wise 2-way hsd

37

Fig 47 residual fitted val, qq

38

Fig 48 Sig of inter

39

Fig 49 Sig of occupation

Fig 50 resid val hist

40

Fig 51 Sig of edu

Fig 52, dash

43

Fig 55 EDA outlier

47

Fig 56 Apps count

Fig 57 Apps box

48

Fig 58 Accept count

Fig 59 Enroll

49

Fig 60 Enroll box

50

Fig 62 Top 10 box

Fig 61 Top10 count

51

Fig 63 Top 25 count

Fig 64 Top 25 box

52

Fig 65 f-under count

Fig 66 F under box

53

Fig 67 p under count

Fig 68 p-under box

54

Fig 69 outstate count

Fig 70 outstate box

55

Fig 71 room count

Fig 72 room board

56

Fig 78 books

Fig 79 personal count

57

Fig 80 Phd count

Fig 81 Terminal count

58

Fig 82 SF ratio count

Fig 83 alum box

59

Fig 84 expend count

Fig 85 Grad_Rate

60

Fig 86 pairplot

62

Fig 88 heatmap

14

Fig 91 pairplot combine apps accept

Fig 92 Apps enroll scatter

66

Fig 93 accept enroll scatter

Fig 94 F-under apps scatter

67

Fig 95 F_under accept

Fig 96 F under enroll

68

Fig 97 T25 T10 scatter

Fig 98 T10 expend

69

Fig 99 out room scatter

Fig 100 Outstate expend

70

Fig 101 Phd terminal scatter

74

Fig 108 before outlier box

Fig 109 after outlier box

77

Fig 112 Scree plot

Fig 113 Scree cumulative

81

Fig 119 loading plot

83

Fig 122 scatter pcs

85

Fig 124 rectangle heat map

Executive summary

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.

Introduction

In the course before, we have looked at t-tests, which are good for telling the difference between two groups, like people with or without dogs, families below the poverty line and families above it, petri dishes of cells that are treated with chemicals and those that aren't. But the world isn't always so binary, and so it likes to compare more than two groups, things like the country of origin, job title, ethnicity, or skills. So, presented with the business problem of comparing between three levels of education and four of occupation, it is a perfect situation for asking **ANOVA** if they can determine significantly how much a person earns. Of course, they do, you'd say, but it's safer to test your assumptions.

ANOVA, short for the **An**alysis **o**f **Va**riance, is a statistical tool for getting useful insights from the raw data by an experiment of applying different treatments to it, to know which one works. For this reason, it is also called a designed test that makes model parameters work better performance with pre-processing-stage analytics. Don't be confused by the name. The Analysis of Variance is a comparison of means (two or more). Consider it an extension of the two-sample t test. With a collection of libraries and modules for statistical analysis, computer language Python is a well-equipped laboratory for performing this randomized, control, statistical experiment that has a variety of uses in the industry.

All ANOVA can tell us is whether two or more data samples are similar or not. Which specific design contributes to that similarity or difference, it is not for ANOVA to say. We compare if the means of the collected samples are the same, significantly. Take three sets of any under-trial medicines for example. The pharmaceutical company would want to know which drug or combination would work. ANOVA analyzes the variations between and within groups of a dataset called population. Ronald Fisher created this test in 1918.

Data description

Salary data

1. **Salary:** The wages of different individuals (just numbers, units unspecified, could be rupees or dollars).
2. **Education:** Three levels (High-school graduate, Bachelors, and Doctorate)
3. **Occupation:** Four levels (Administrative and clerical, Sales, Professional or specialty, and Executive or managerial)

Data sample

First five rows

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Fig 1. Salary data head

Last five rows

	Education	Occupation	Salary
35	Bachelors	Exec-managerial	173935
36	Bachelors	Exec-managerial	212448
37	Bachelors	Exec-managerial	173664
38	Bachelors	Exec-managerial	212760
39	Doctorate	Exec-managerial	212781

Fig 2 Salary data tail

Indexed from zero, it's a dataset of three columns or variables, titled Education, Occupation, and Salary. The Education variable holds the names of degrees or qualifications, Occupation holds job roles, while Salary holds the numeric pay. Salary is our response variable in the ANOVA test. The other two are categorical, factor variables.

Exploratory data analysis

Types of variables in the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

Fig 3 Salary data info

Three columns, 40 entries, one integer and two object types. For the sake of ANOVA, we'll change the object data types to categorical. The integer or dependent variable of this experiment is fine.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     category
 1   Occupation  40 non-null     category
 2   Salary      40 non-null     int64
dtypes: category(2), int64(1)
memory usage: 864.0 bytes
```

Fig 4 Salary data info

The change is carried out. Education and Occupation now have a data type called category, so these are categorical variables now, transformed for computation.

Checking for missing values

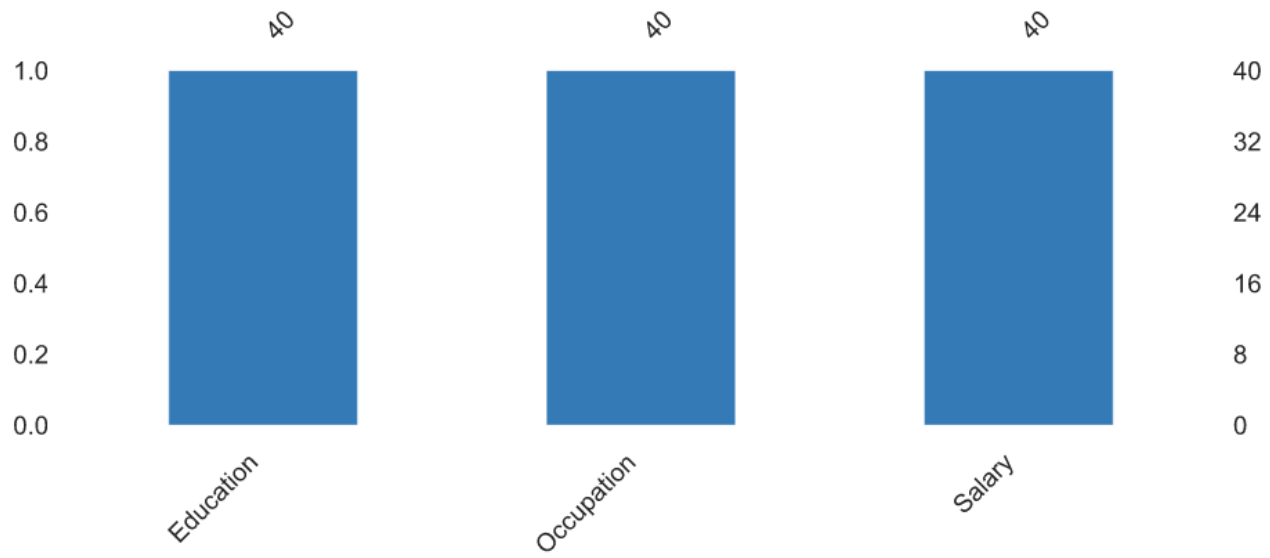


Fig 5 Nullity graph

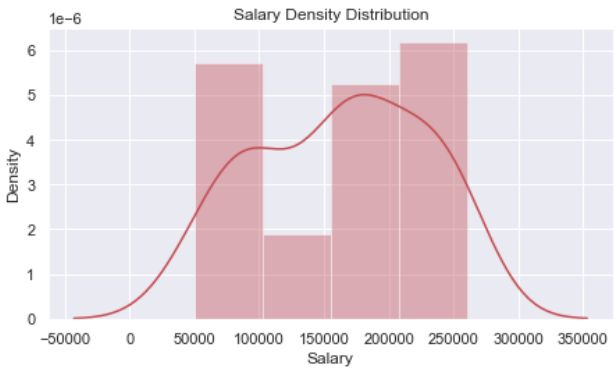
A simple visualization of nullity by column.

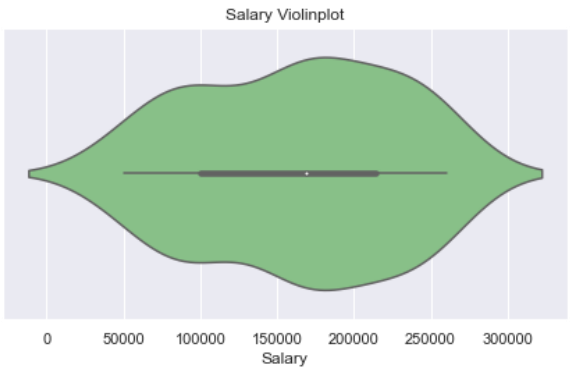
```
Education    40 non-null
Occupation   40 non-null
Salary       40 non-null
```

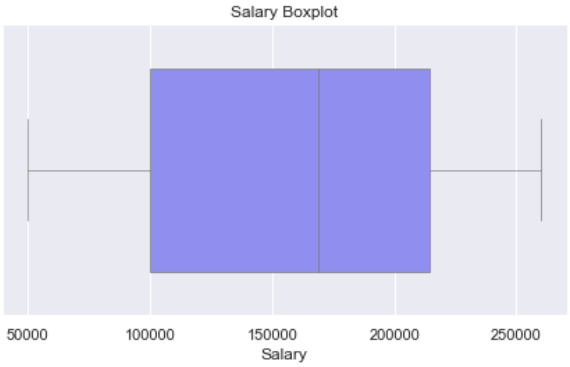
The graph and the output show no missing values, so the dataset looks clean enough.

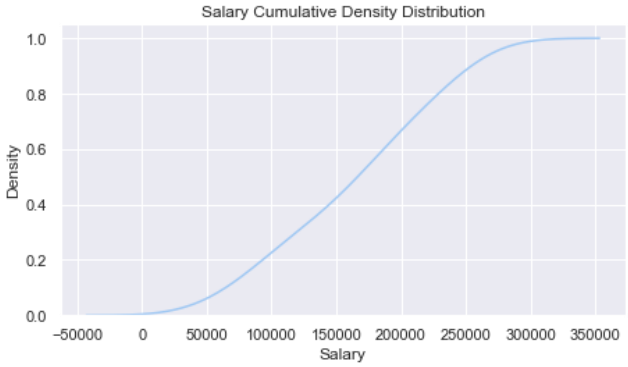
Dependent variable

Salary









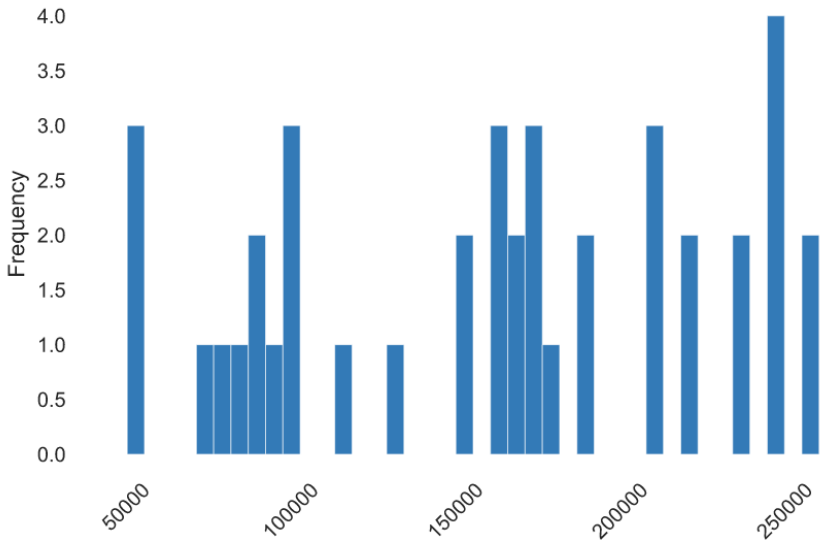
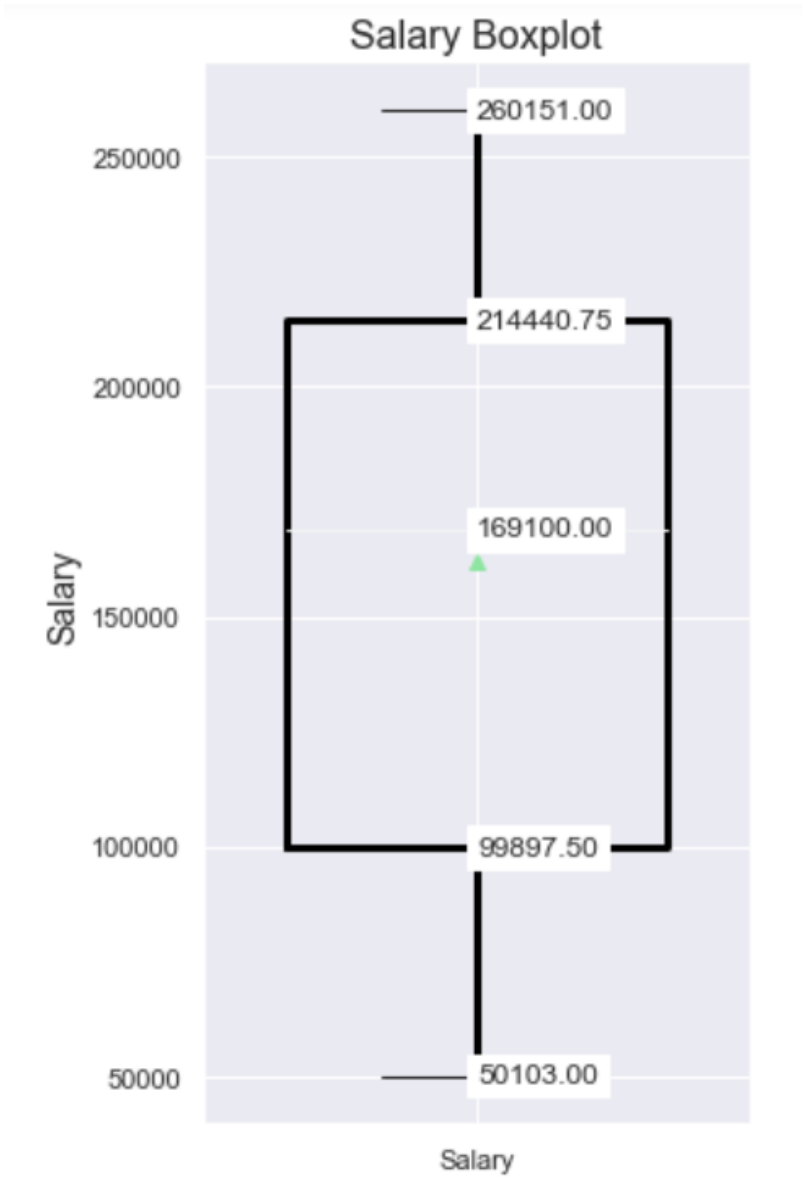


Fig 5 Salary subplots



Quantile statistics	
Minimum	50103
5-th percentile	52136
Q1	99897.5
median	169100
Q3	214440.75
95-th percentile	249613.9
Maximum	260151
Range	210048
Interquartile range (IQR)	114543.25

```
Total number of outliers in Salary distribution: 0.
```

Descriptive statistics	
Standard deviation	64860.40751
Coefficient of variation (CV)	0.3999115681
Kurtosis	-1.162113649
Mean	162186.875
Median Absolute Deviation (MAD)	59694.5

Skewness	-0.1731179243
Sum	6487475
Variance	4206872462
Monotocity	Not monotonic

Fig 8, 9 ,19,11 Salary boxplot, quantiles, descriptive stats, skewness etc. for salary

1.1 State the null and the alternative hypothesis for conducting one-way ANOVA for both Education and Occupation individually

Education, first independent variable

Let's look at the Education variable for a moment. This plot output is from the sns.countplot() function of Python.

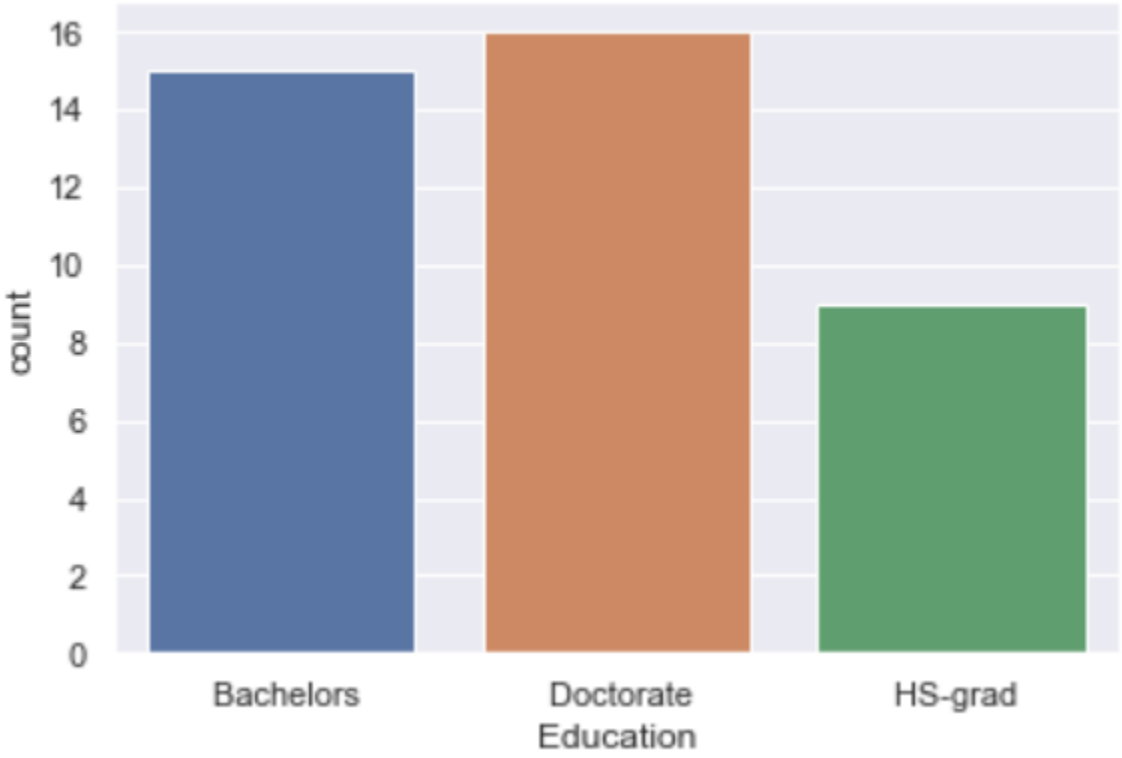
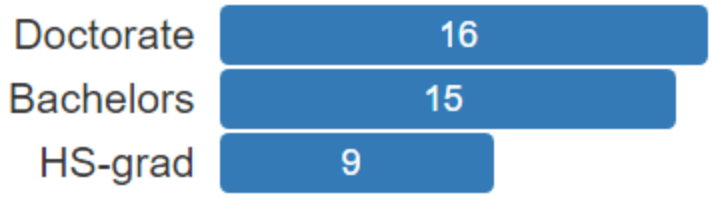
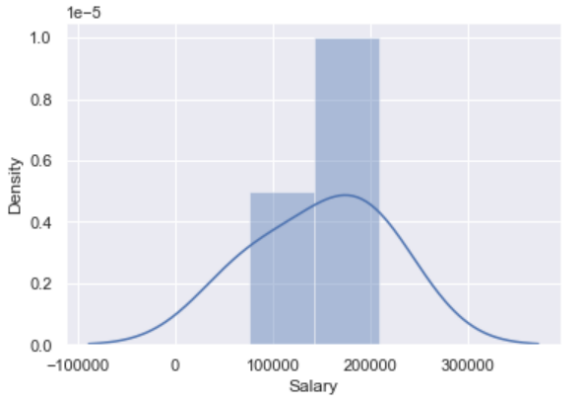


Fig 12 Education countplots



The three level are Bachelors, Doctorate, and HS-grad, out of which the maximum count is of the Doctorates, followed by Bachelor-degree holders and the high-school graduates.

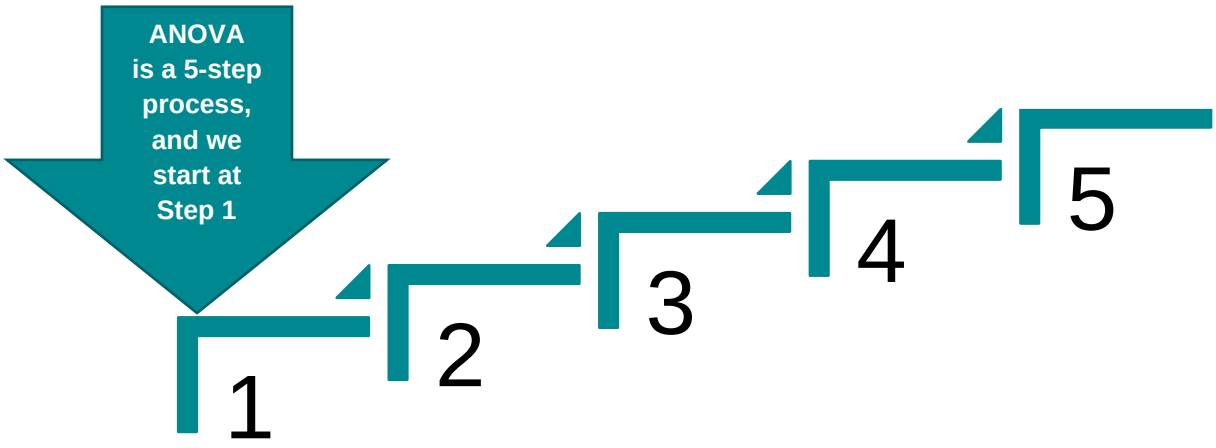


A plot to show how the pay for the Education data is distributed

Fig 13 Salary distplot

The three level are Bachelors, Doctorate, and HS-grad, out of which the maximum count is of the Doctorates, followed by Bachelor-degree holders and the high-school graduates.

Step 1. State null hypotheses and determine the level of significance



	Education	Occupation
Null Hypothesis	**H0: μ1 = μ2 = μ3** The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-grad)	**H0: μ1 = μ2 = μ3 = μ4** The mean salary is the same across all the 4 categories of occupations (Adm-clerical, Exec-managerial, Prof-specialty, and Sales)
Alternative hypothesis	**H1: Means are not all equal (α=0.05)** The mean salary is different for at least one category of education	**H1: Means are not all equal (α=0.05)** The mean salary is different for at least one category of occupation

Before we perform one-way ANOVA on Education, we must sort the column to observe stacked data and massage it to turn it into wide format from long format.

	Education	Occupation	Salary
19	Bachelors	Prof-specialty	90135
37	Bachelors	Exec-managerial	173664
36	Bachelors	Exec-managerial	212448
35	Bachelors	Exec-managerial	173935
25	Bachelors	Sales	260151
24	Bachelors	Sales	167431
23	Bachelors	Sales	191712
22	Bachelors	Sales	149909
21	Bachelors	Prof-specialty	133696
20	Bachelors	Prof-specialty	100135
38	Bachelors	Exec-managerial	212760
18	Bachelors	Prof-specialty	99185
17	Bachelors	Adm-clerical	188729
16	Bachelors	Adm-clerical	162494
15	Bachelors	Adm-clerical	160910
0	Doctorate	Adm-clerical	153197
13	Doctorate	Prof-specialty	248871
1	Doctorate	Adm-clerical	115945

Earlier stacked, the treatments or the levels become clear now

	HS-grad	Bachelors	Doctorate
0	50103.0	90135.0	153197
1	52242.0	173664.0	248871
2	75333.0	212448.0	115945
3	77743.0	173935.0	175935
4	83203.0	260151.0	220754
5	90456.0	167431.0	257345
6	100678.0	191712.0	170769
7	95469.0	149909.0	219420
8	50122.0	133696.0	212781
9	NaN	100135.0	160540
10	NaN	212760.0	180934
11	NaN	99185.0	248156
12	NaN	188729.0	247724
13	NaN	162494.0	249207
14	NaN	160910.0	235334
15	NaN	NaN	237920

Fig 14 Salary education data slicing

	HS-grad	Bachelors	Doctorate
0	50103.0	90135.0	153197
1	52242.0	173664.0	248871
2	75333.0	212448.0	115945
3	77743.0	173935.0	175935
4	83203.0	260151.0	220754
5	90456.0	167431.0	257345
6	100678.0	191712.0	170769
7	95469.0	149909.0	219420
8	50122.0	133696.0	212781

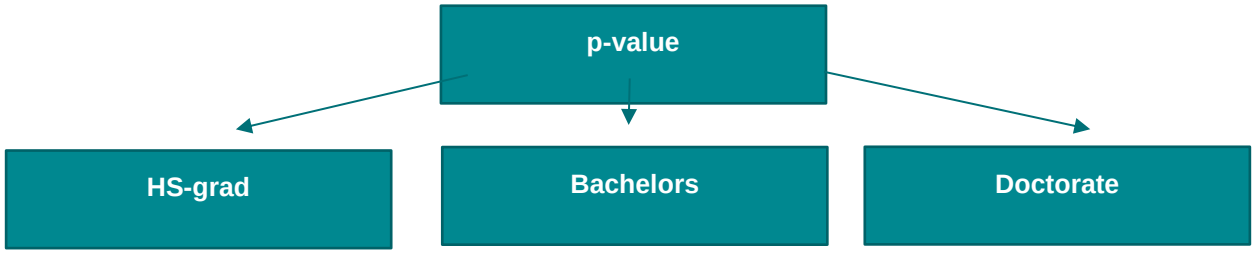
The separated levels are truncated to the same number of observations

Fig 14 Separated levels truncated

Step 2. Select the appropriate test statistic

Shapiro-Wilk's test of normality

This data is assumed to be normal but Shapiro-Wilk will give us the p-value for Bachelors, Doctorate, and HS-grad to make the ANOVA table. Bachelors seems the most dominant treatment and HS-grad the least, but we will test it.



0.1783432960510254 0.9611656069755554 0.726436972618103

Levene's test

```
LeveneResult(statistic=1.5908261877690124, pvalue=0.22450429424334328)
```

Its test statistic and p-value prove the homogeneity of variances for all the treatment levels. The null hypothesis is that the variances are equal (an assumption for the ANOVA). In this case, as the p-value is high at 0.224, so we fail to reject null hypothesis and hold our assumption that variance in the three groups is equal

Step 3. Set up the decision rule

The appropriate critical value can be found in a table of probabilities for the F distribution. In order to determine the critical value of F we need degrees of freedom, df1=k-1 and df2=N-k. Reject H0 if F > F-critical.

Step 4. Compute the test statistic.

The steps we have just done.

Step 5. Do the test. Decide whether to reject or fail to reject the null hypothesis

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.

One-way ANOVA on Education

An algorithm formula fitted into ordinary least squares is put into a model to calculate the sums of squares, the means sums of squares and the p-value etc. Data = Model + Error.

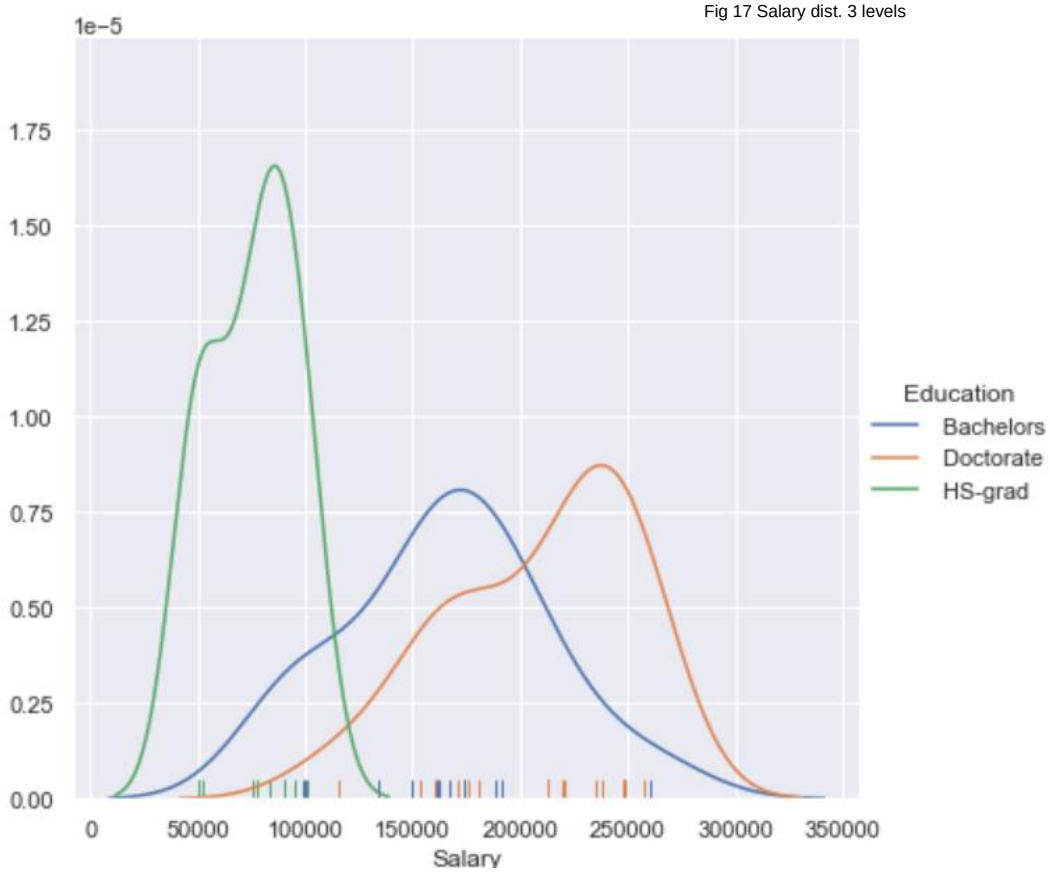
	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Fig 16 1-w anova edu

Since the p-value, **1.257709e-08**, is less than the significance level (0.05 or 95% confidence level), we can **reject the null hypothesis** and say that there is a difference in the mean salaries of different educational-qualification groups, which we will also show by a FacetGrid graph next.

Salary distributions for 3 levels of education

Fig 17 Salary dist. 3 levels



For the given problem, the sum of squares due to the factor Education (SSB) is 1.0269 and the sum of squares due to error (SSW) is 6.1372. The total sum of squares (SST) for the data is (1.0269+6.1372=7.1640). Since the factor has 3 levels, DF corresponding to Education is 3 –1 = 2. Total DF is 40–1 = 39. Hence DF due to error is 39 –2 = 37.

The mean sum of squares is obtained by dividing the sums of squares by the corresponding DF. The value of the F-statistic is approximately 31 and the p-value is highly significant, being extremely lower than the alpha level of 0.05. Based on the ANOVA test we, therefore, reject the null hypothesis that the three population means are identical. For at least one qualification, the mean salary is different from the rest. Residuals are defined as the difference between the observed values and the expected values.

All the data spreads are shifted and, so, not only their means and medians don't overlap but also their means don't come from the same population. ANOVA is a good test for catching the differences underlying a seemingly normal distribution. To study Salary with respect to Education at three levels, we also make a boxplot.

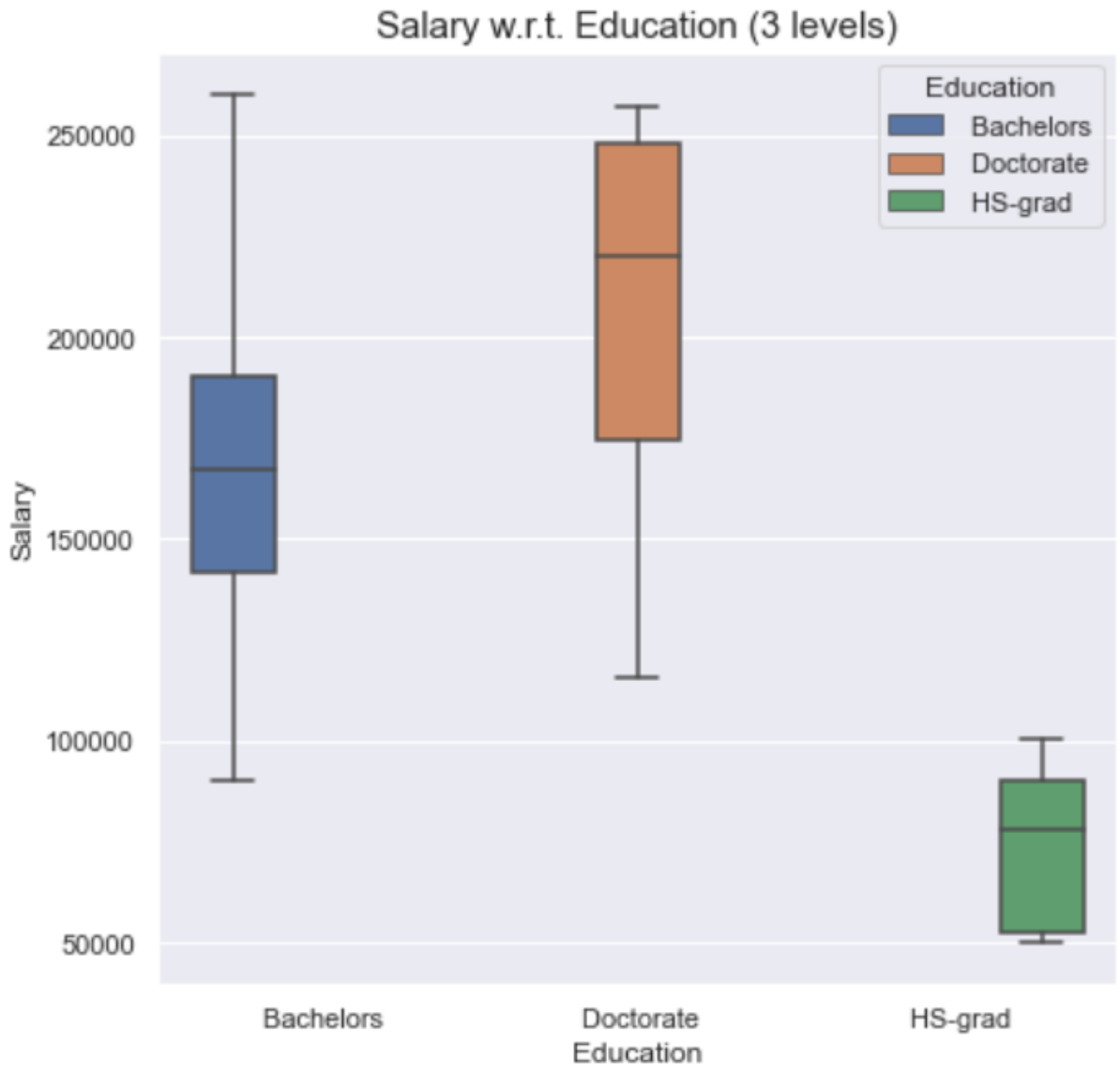
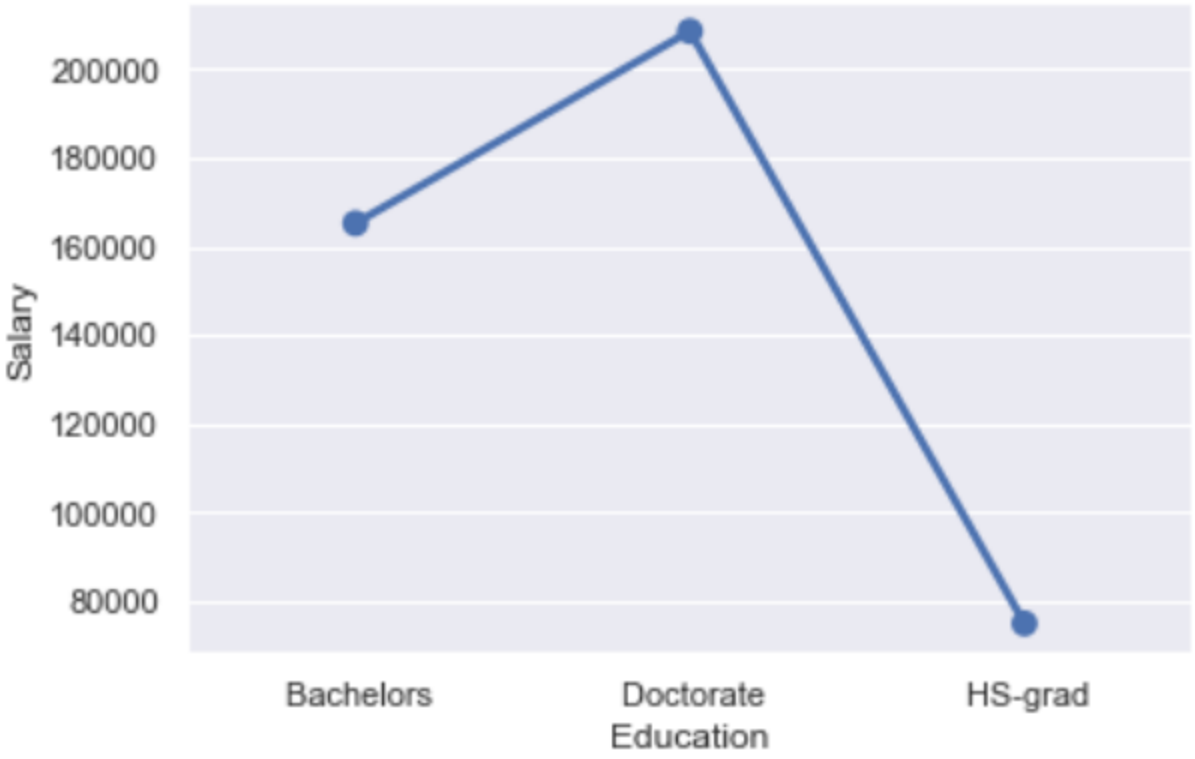


Fig 18 Sal wrt edu box

For the doctorate-degree holders, the mean and median salaries are higher than the wages of the other two groups. See just the box, not the spread of the data (the whiskers on the box plot). The salaries for high-school graduates are the least, so perhaps they need to upgrade their skills to be able to make a better living. Check out this pointplot.

Fig 19 edu pointplot



The interpretation of this pointplot is that there is a difference in the average salaries of the three kinds of degree-holders but while the gap is 40,000 (dollars or rupees) between the Bachelors and Doctorate groups, it is 1.20 lakh in contrast between high-school graduates (HS-grad) and Doctorate, so the high-school graduates have more catching up to do. So, for a well-paying career, our recommendation will be to pick up a doctorate degree. While it might be expensive in comparison and you might get into the job market much later, the money you'll earn later will more than make up for it. Suggest high-school graduates to do bachelors (get into college), and advise the college pass-outs to join a research programme to earn the skills that pay. Of course, you can also grind your way to success. The more you learn, the more you earn. It's good if you start earning after school but don't assume that the college and university degrees don't have much to teach us and they are a waste of years. One can make a good life out of a college degree, these lines show, but it doesn't harm to aim higher for a greater reward. Figures don't lie.

Once the null hypothesis of equality of means is rejected, the next natural question is to find out which mean(s) is different from the rest. Before we answer that question, we need to first check whether Salary is dependent on Occupation.

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

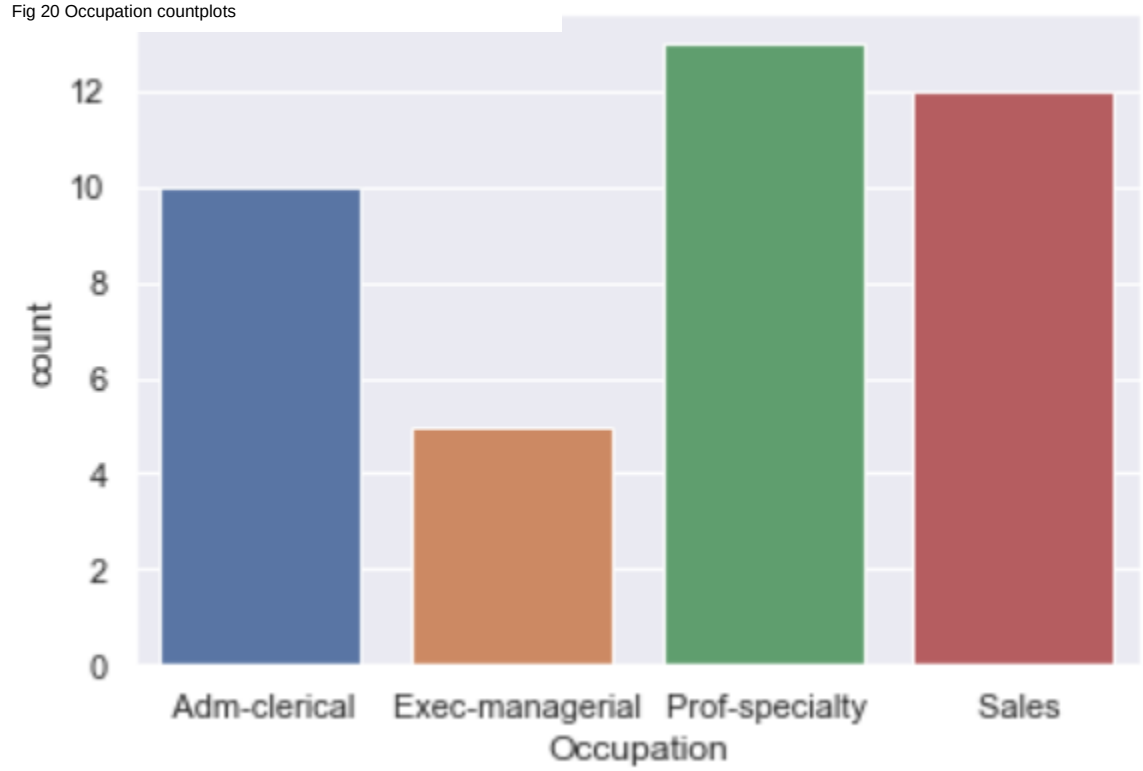
Occupation, second independent variable

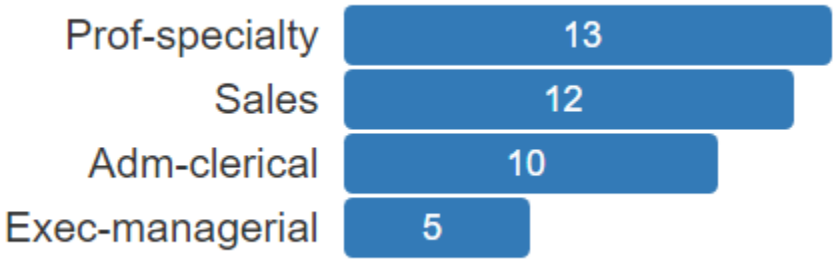
Step 1. State null hypotheses and determine the level of significance

Null and alternative hypothesis stated already, **α=0.05**, as also stated**.**

It is better to perform an exploratory data analysis before the ANOVA test.

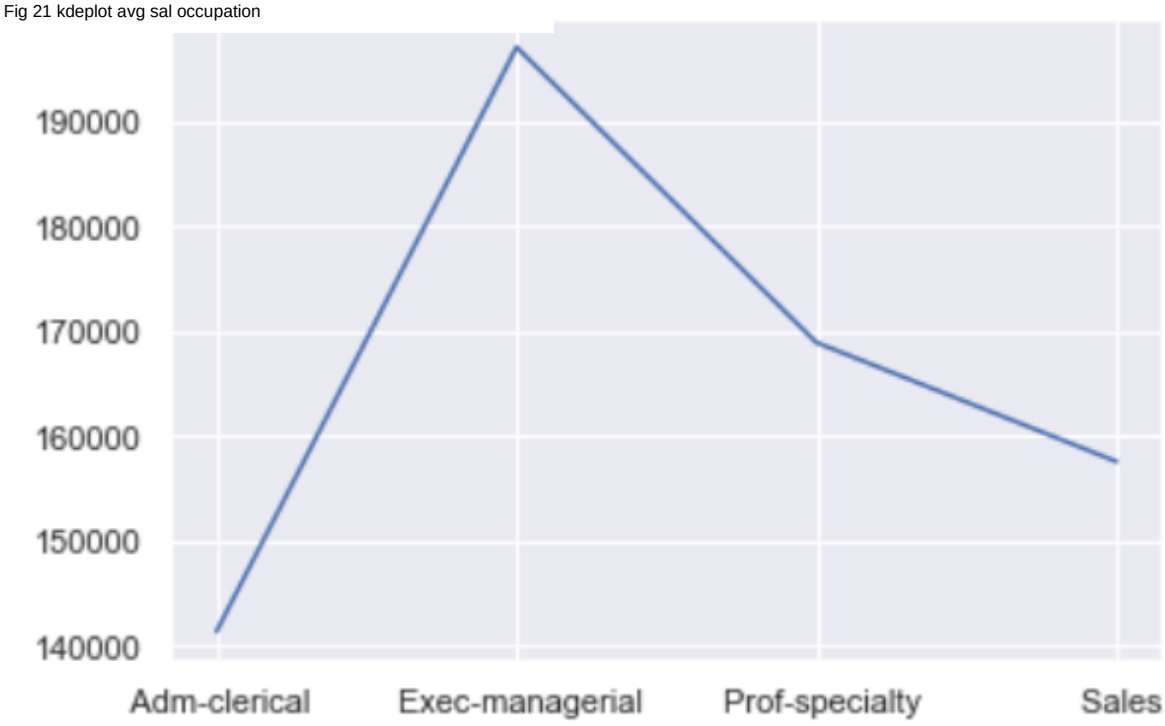
Fig 20 Occupation countplots





It's quite an unbalanced stack that'll need to be melted before it is ready for the Anova test. Professors and specialists are in the lead already but we might be on the edge of finding something quite interesting in the dataset.

Fig 21 kdeplot avg sal occupation



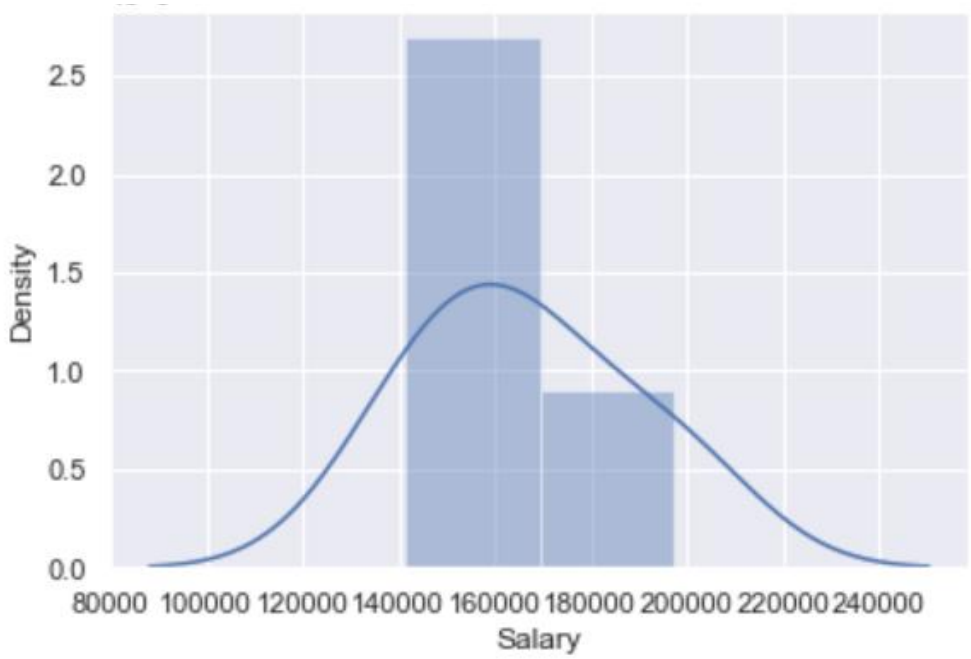
The kdeplot shows business executives and managers to be taking home the fattest average pay packets, leaving those in sales grumpy and those desk bounders of the administrative and clerical block the grumpiest. Surprise indeed. While we saw that the doctorate degree was the most rewarding, the same is not true for the jobs of professorship and specialists. The pay wonderland belongs to the managers, it seems, by this graph. People with doctorates could also be getting into this kind of a decision-making job. Meanwhile to be sure, we check the data for disproving the null hypothesis that the salaries are not different and why would there be any difference in the means.

Here is the **average salary by occupation**, followed by its distribution plot.

Occupation	
Adm-clerical	141424.300000
Exec-managerial	197117.600000
Prof-specialty	168953.153846
Sales	157604.416667

Fig 22 avg sal by occupation

Fig 23 Salary data for occupation



On the surface, the **salary data for occupation** seems almost normally distributed but we will check.

A glimpse of how the **data** was **melted and massaged** to reveal the treatments. It was truncated to five rows for the equality of observations. It did not affect the final values.

	Adm-clerical	Exec-managerial	Prof-specialty	Sales
0	153197.0	173664.0	248871	149909.0
1	115945.0	212448.0	95469	170769.0
2	175935.0	173935.0	100678	219420.0
3	220754.0	212760.0	90456	237920.0
4	83203.0	212781.0	248156	180934.0
5	77743.0	NaN	257345	52242.0
6	75333.0	NaN	90135	50103.0
7	188729.0	NaN	235334	260151.0
8	162494.0	NaN	100135	167431.0
9	160910.0	NaN	249207	191712.0
10	NaN	NaN	247724	160540.0
11	NaN	NaN	133696	50122.0
12	NaN	NaN	99185	NaN

	Adm-clerical	Exec-managerial	Prof-specialty	Sales
0	153197.0	173664.0	248871	149909.0
1	115945.0	212448.0	95469	170769.0
2	175935.0	173935.0	100678	219420.0
3	220754.0	212760.0	90456	237920.0
4	83203.0	212781.0	248156	180934.0

Fig 24 occu data sliced, truncated

other the three level are Bachelors, Doctorate, and HS-grad, out of which the maximum count is of the Doctorates, followed by Bachelor-degree holders and the high-school graduates.

Step 2. Select the appropriate test statistic

F and P-values

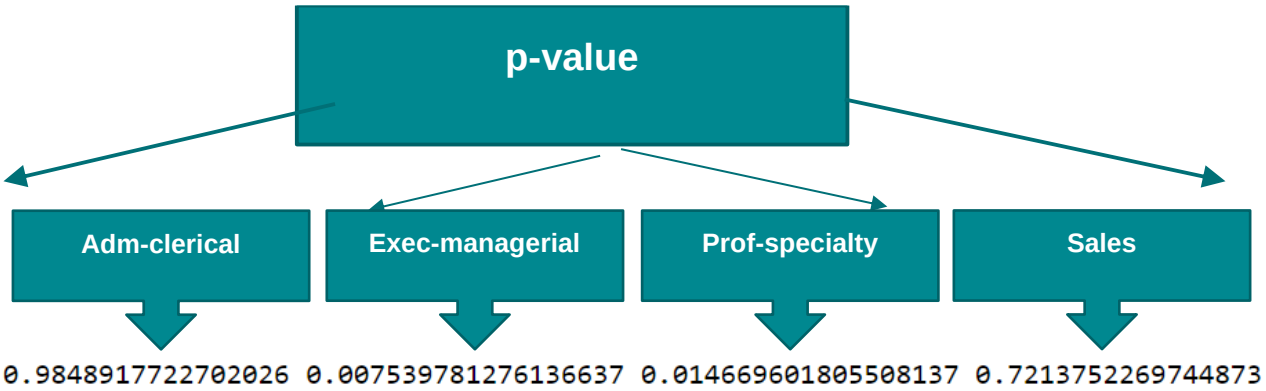
Step 3. Set up the decision rule

The appropriate critical value can be found in a table of probabilities for the F distribution. In order to determine the critical value of F we need degrees of freedom, df1=k-1 and df2=N-k. Reject H0 if F > F-critical.

Step 4. Compute the test statistic.

The steps we are about to do.

Shapiro-Wilk's test



Levene's test

```
LeveneResult(statistic=0.9966049847648979, pvalue=0.41967870017126496)
```

As the p-value is high at 0.419, we fail to reject the null hypothesis that the variance in four groups is equal

Step 5. Do the test. Decide whether to reject or fail to reject the null hypothesis

One-way ANOVA on Occupation

An algorithm formula fitted into ordinary least squares is put into a model to calculate the sums of squares, the means sums of squares and the p-value etc. Data = Model + Error.

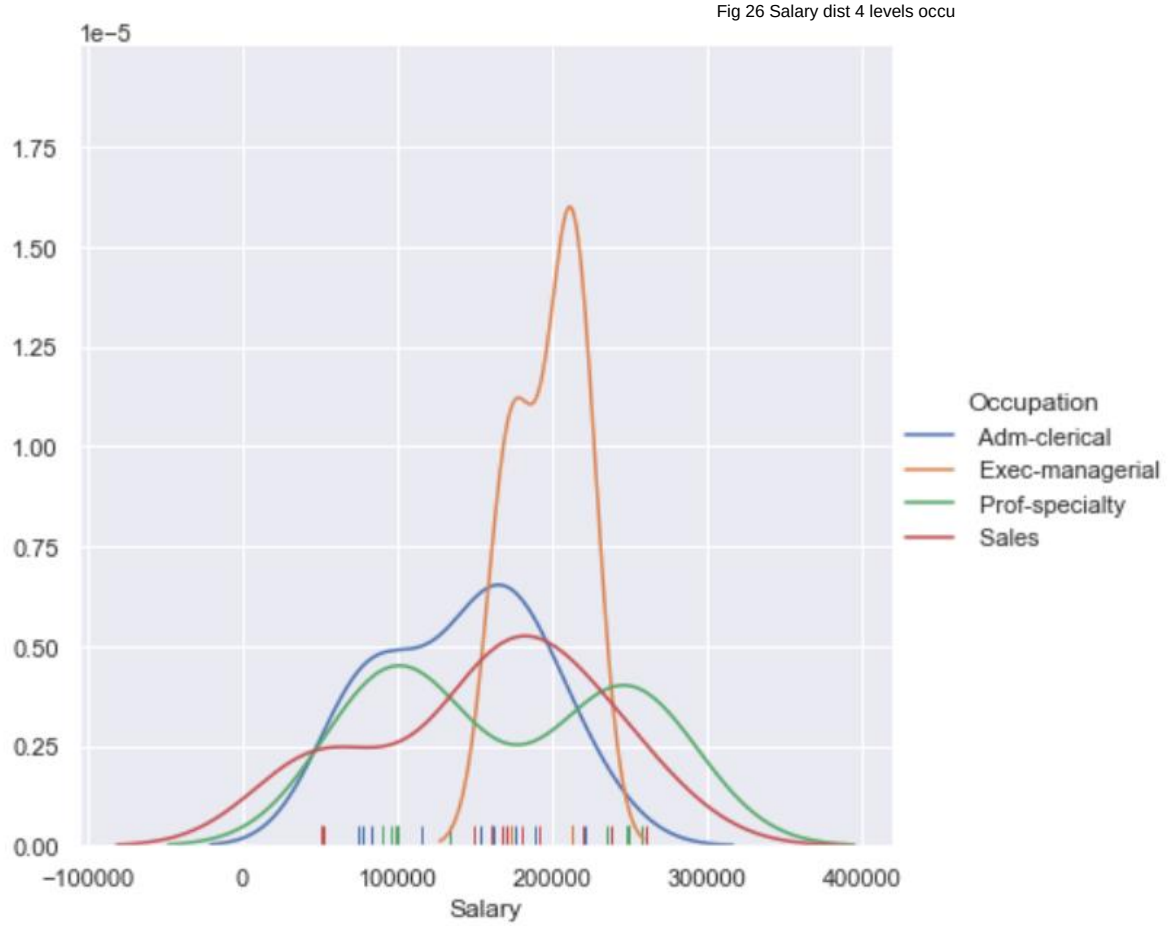
Fig 25 1-w Anova occu

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

Since the p-value, **0.45**, is more than the significance level (0.05 or 95% confidence level), we **fail to reject the null hypothesis** and, so there, there is not much difference in the mean salaries of different professional groups, which we will also show by a FacetGrid graph next. If p is high, null will fly.

Salary distributions for 4 levels of Occupation

Fig 26 Salary dist 4 levels occu



Interpretation

For the given problem, the sum of squares due to the factor Education (SSB) is 13.06045 and the sum of squares due to error (SSW) is 15.14562. The total sum of squares (SST) for the data is (13.06045 +15.14562=28.20607). Since the factor has 4 levels, DF corresponding to Education is 4 –1 = 3. Total DF is 40–1 = 39. Hence DF due to error is 39 –3 = 36.

The mean sum of squares is obtained by dividing the sums of squares by the corresponding DF. The value of the F-statistic is approximately 0.88 and the p-value of 0.45 is higher than the alpha level of 0.05. Based on the ANOVA test we, therefore, fail to reject the null hypothesis that the four population means are identical. They don't seem to be much different or shifted. Residuals are defined as the difference between the observed values and the expected values.

All the data spreads are one on top of the other and, so their means and medians kind of overlap and their means must be coming from the same population. ANOVA is a good test for catching the differences underlying a seemingly normal distribution. To study Salary with respect to Occupation at four levels, we also make a boxplot.

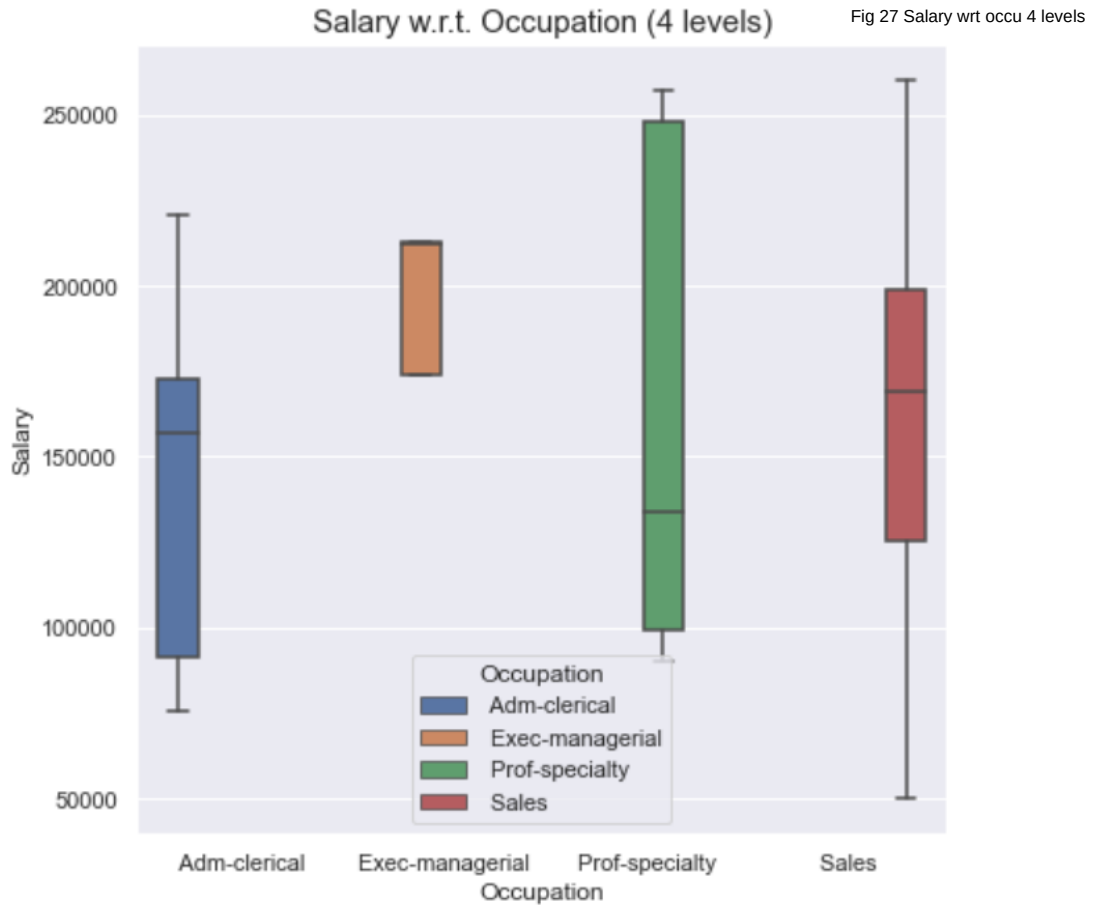
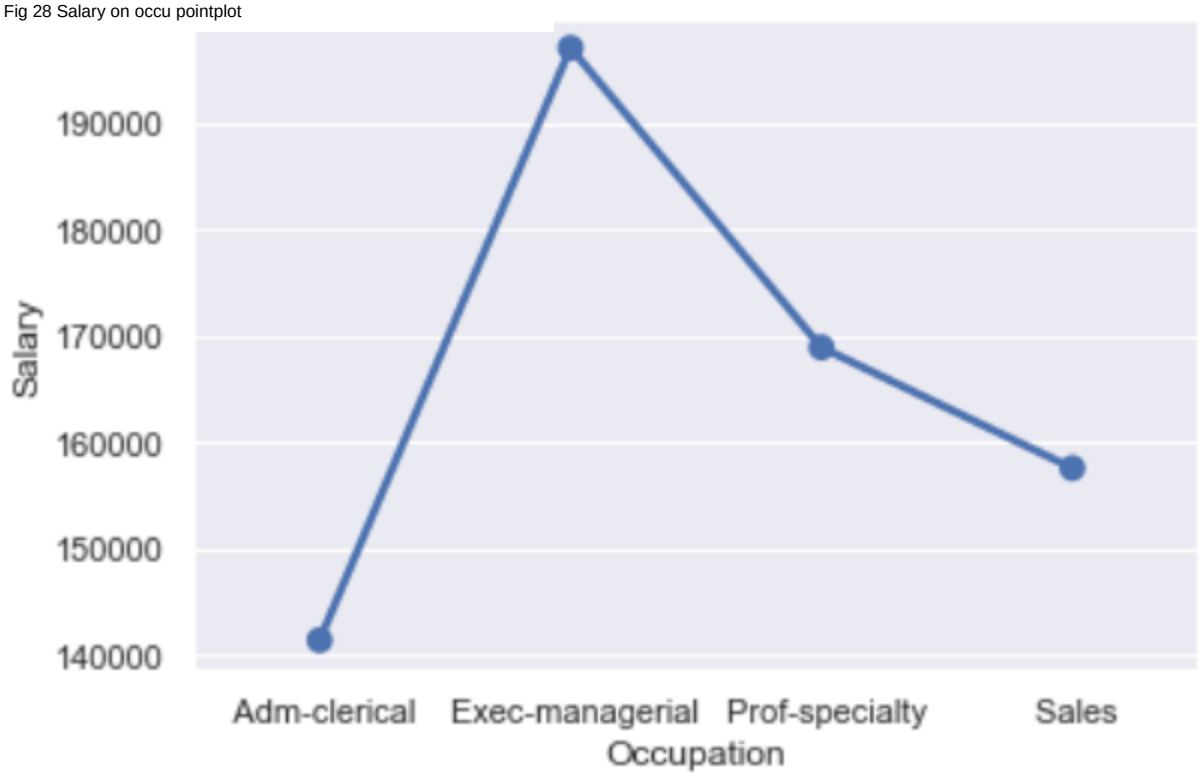


Fig 27 Salary wrt occu 4 levels

For the executive managers, the mean and median salaries are higher than the wages of the other three groups. The boxes are almost neck and neck. Like a photo-finish race. See just the box, not the spread of the data (the whiskers on the box plot). The salaries for Administrative-clerical workers seem to be the least, so perhaps they need to work to get into one of the other three to be able to make a better living. Salaries for sales cover the broadest range. Check out this pointplot also.

Fig 28 Salary on occu pointplot



The interpretation of this pointplot is that the apparent difference in the average salaries of the four kinds of professions is insignificant in statistical terms. The gap is 20,000 (dollars or rupees) between the Exec-managerial and Prof-specialty groups, while it is 50,000 in contrast between Adm-clerical and the executives and managers, so while they might be comfortable in their career, our recommendation will be to get into the executive-managerial role, if the aim is to make more money. While it might be full of more stress and risk, there are some rewards. Between professors and salesmen, there's little to choose but the former still make more. Employers are paying more for decision-making jobs. Wages seem to distributed fairly in this set of population.

Once the null hypothesis holds, we don't do post hoc test, generally, because we don't expect to find and distinct pairs of means. But this data can surprise us, so we'll check. We are 61% sure that the at least one of the means is different but we are not 95% sure, at least not now. To check, we do something called the post hoc Tukey's test. Post hoc means after the fact. While ANOVA tells us that one of the means is different, it doesn't tell us which particular mean or which particular combination of factors is different. Tukey will do that for us.

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.

Multiple comparison tests for X_1: Education

In order to identify for which educational qualification, the mean salary differs from the other groups, the hypotheses may be stated as:-

H_0: All pairs of group means are equal against
H_a: At least one group mean is different from the rest.

In this case, as there are only 3 pairs to be considered, we may write the null and alternative hypothesis as:-

$H0$:$\mu1=\mu2$ and $\mu1=\mu3$ and $\mu2=\mu3$ against

Ha:$\mu1\neq\mu2$ or $\mu1\neq\mu3$ or$\mu2\neq\mu3$, respectively, where $\mu1$represents mean salaries when education is HS-grad, $\mu2$ represents mean salaries when education is Bachelors, and $\mu3$ is the same for Doctorate.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7541.1439	79006.9894	True
Bachelors	HS-grad	-90114.1556	0.001	-132035.1958	-48193.1153	True
Doctorate	HS-grad	-133388.2222	0.001	-174815.0876	-91961.3569	True

Fig 29 Tukey hsd education

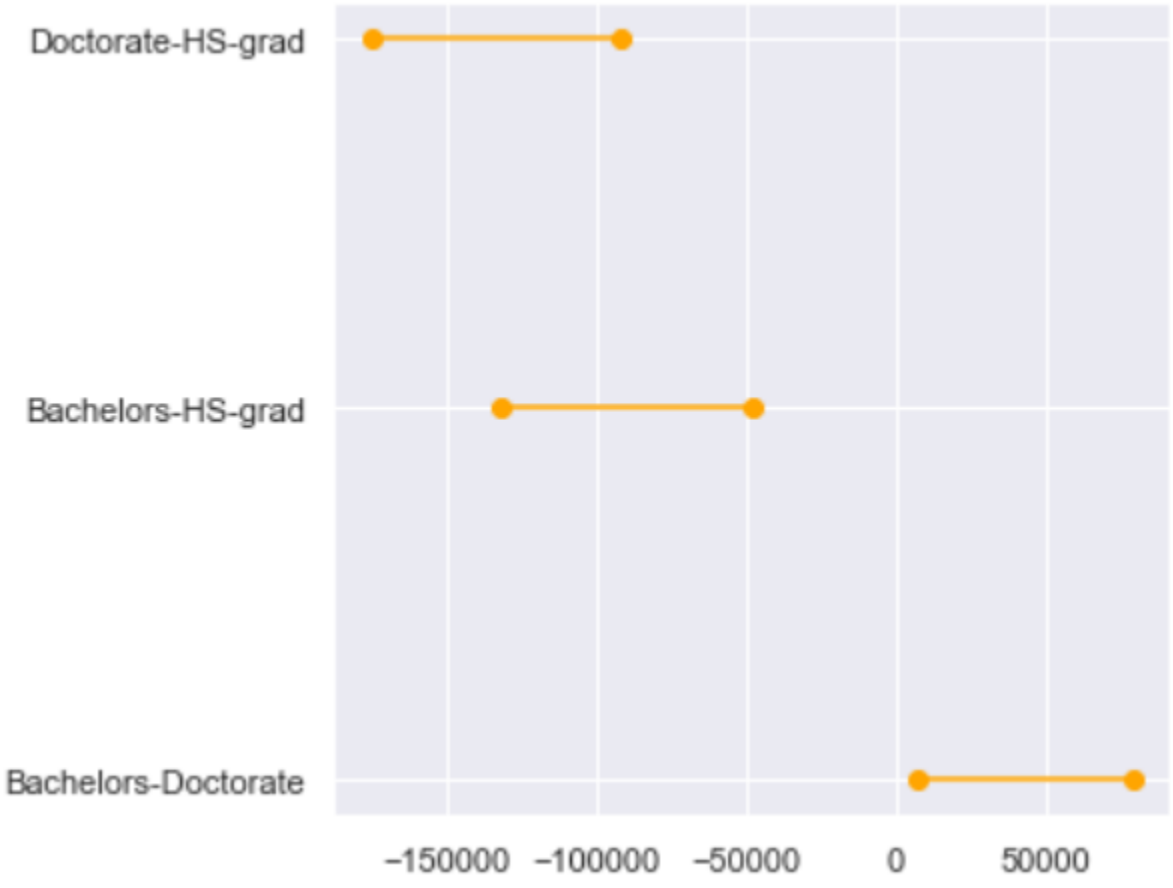
P-value is significant for comparing mean levels of salary for the pair Bachelors-HS-grad and Doctorate-HS-grad, and even for Bachelors-Doctorate. For all pairs, the null hypothesis of equality of all population means is rejected. It is now clear that mean salaries for Bachelors, Doctorate, and HS-grad are all different significantly. The numerical values of the differences being positive, mean salary for Bachelors is significantly lower than that for Doctorate. The numerical values of the differences being negative, mean salary for Bachelors and Doctorate are higher significantly than that for HS-grad. This same observation is borne out by the threshold plot, coming up on the next page.

Threshold plot

The closer it is to zero, the more significant the value. The closer the bars, the more similar means the associated pairs have. The wider apart they are, the more different from the rest these combinations are.

Zero is the threshold point, hence the name threshold plot for family-wise test.

Fig 30 threshold edu



For the graphical representation of pair-wise comparisons from Tukey's HSD for Bachelors, HS-grad, and Doctorate, the confidence intervals' not containing 0 is for the difference between all three pairs indicates that the population means of these pairs of educational qualification are different. From the values of the pairwise differences, it may also be concluded.

Multiple comparison tests for X_2: Occupation

In order to identify for which Occupation, the mean salary is different from others, the hypotheses may be stated as:

H_0: All pairs of group means are equal against

H_a: At least one group mean is different from the rest.

We may also rewrite the null and alternative hypotheses as:-

H_0: $\mu_i=\mu_j$ against H_a: $\mu_i \neq \mu_j$, for all $i \neq j$, i,j=1,2,3,4.

Subscript 1 represents mean of Adm-clerical, 2 Exec-managerial, 3 Prof-specialty, and 4 Sales

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

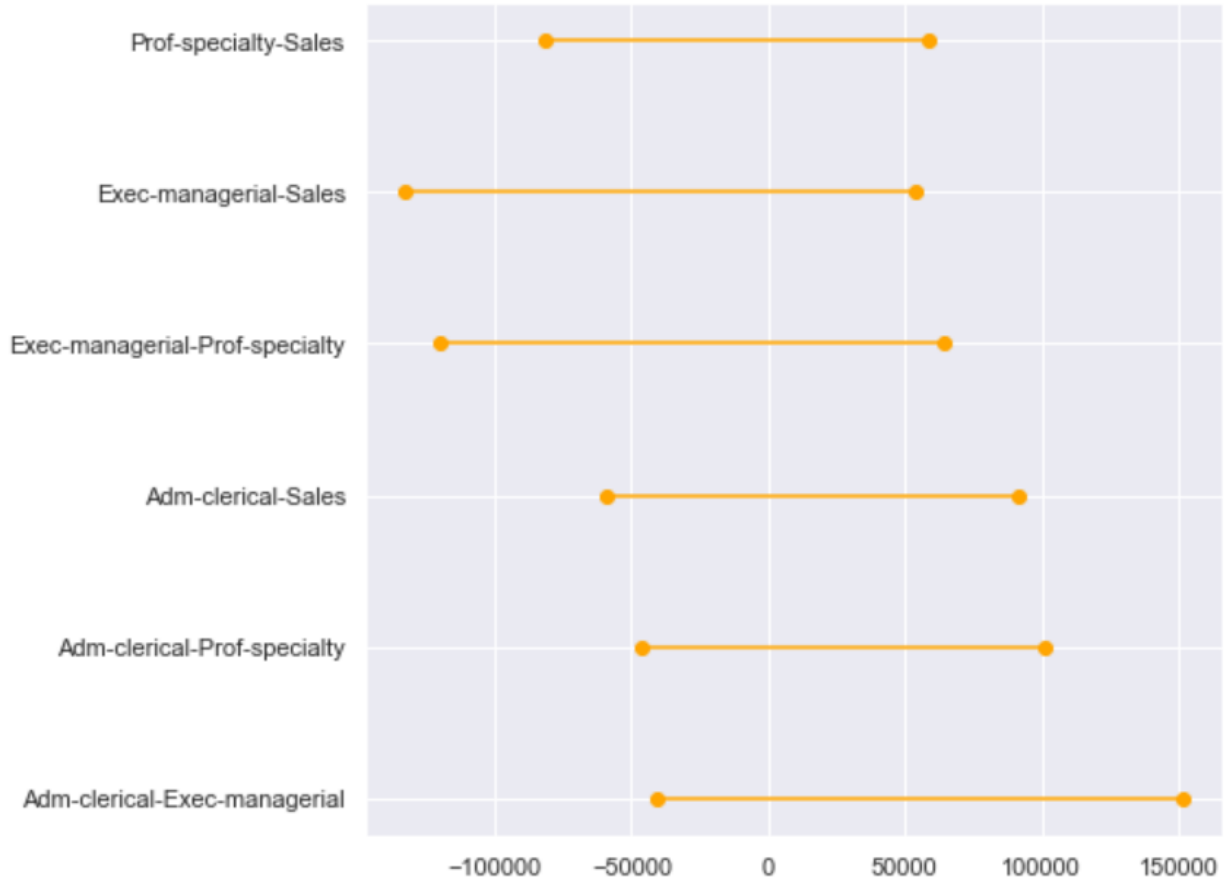
Fig 31 Tukey hsd occupation

As fail to reject the null hypotheses for each of the six pairs, It is clear from the above table that there is no significant difference in the mean salary when comparisons are done with six combinations. The same is also borne out by the threshold plot, coming up next.

Threshold plot

The evidence is clear for these pairs. Their means seem to be coming from the same population. The zero line cuts through all of these lines, another proof that these combinations are not much different.

Fig 32 threshold occu

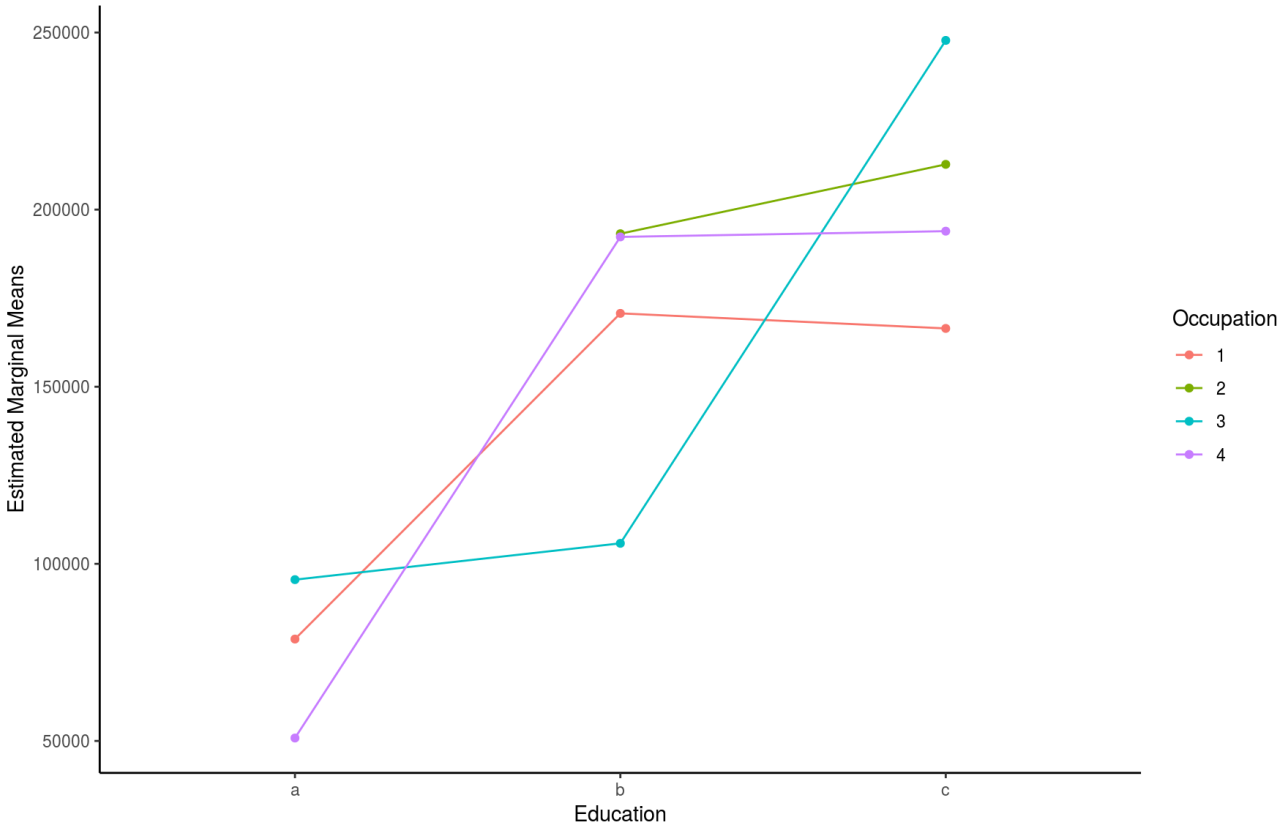


1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Fig 33 anova interaction

Interaction plot



Legend

a=HS-grad

b=Bachelors

c=Doctorate

Legend-1

1=Adm-clerical

2=Exec-managerial

3=Prof-specialty

4=Sales, Readings=Salary

Fig 34 Ineractionplot edu

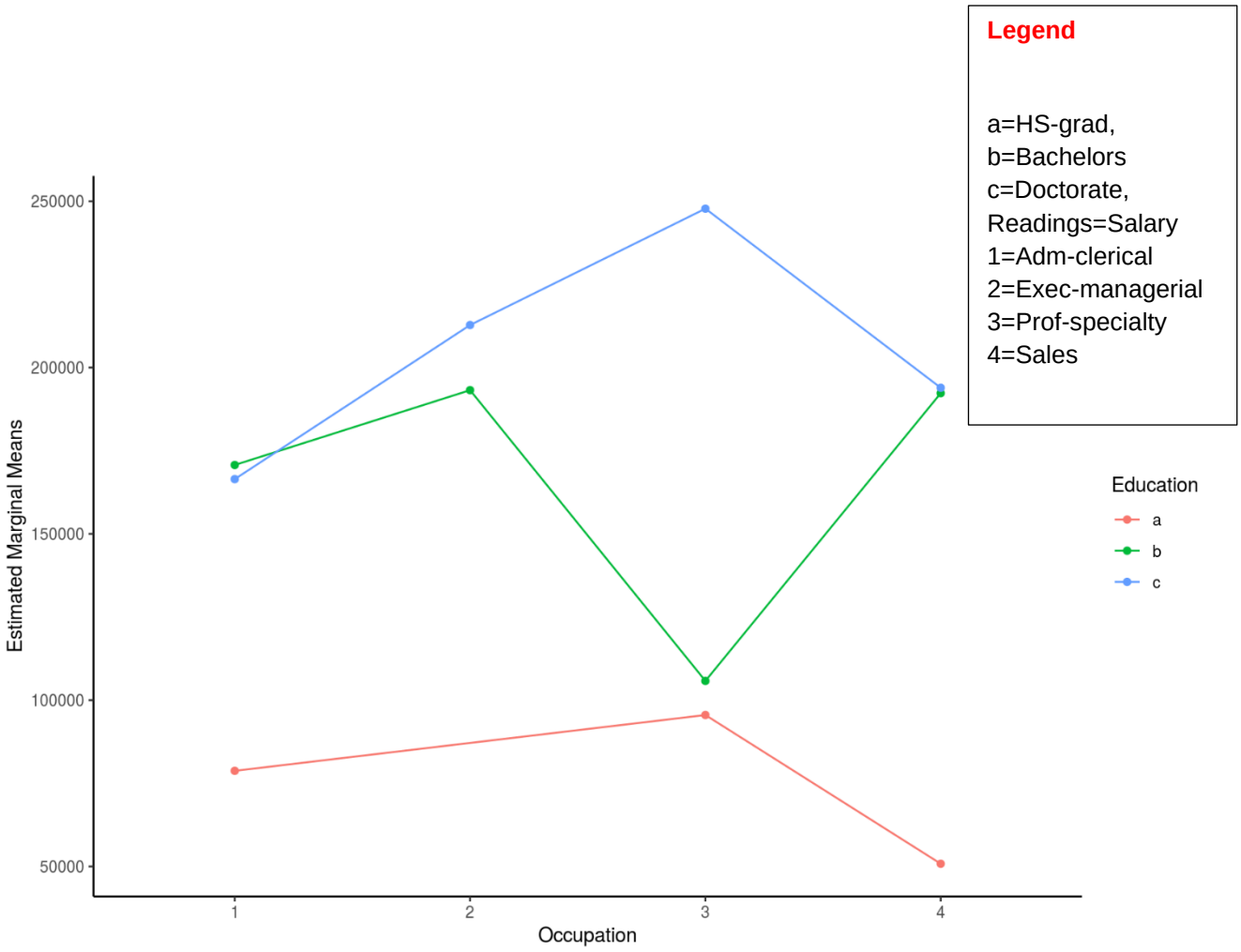


Fig 35 interactionplot occupation

Interpretation

The interaction plot shows that there is significant amount of interaction between the categorical variables, Education and Occupation. The following are some of the observations from the interaction plot:

• People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.

• People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries(salaries ranging from 170000 – 190000).

• People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.

• People with education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. A reversal in this part of the plot

• Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.

• Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.

• Adm clerical people with education as HS-grad earn the lowest salaries when compared to people with education as Bachelors or Doctorate.

• Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum

• People with education as HS -Grad earn the minimum salaries. • There are no people with education as HS -grad who hold Exec-managerial occupation.

• People with education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries.

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?

Two-way ANOVA on Education and Occupation, with salary as variable

$H0$: The effect of the independent variable 'education' on the mean 'salary' does not depend on the effect of the other independent variable 'occupation' (i. e. there is no interaction effect between the 2 independent variables, education and occupation). $H1$: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean Salary.

	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

Fig 36 anova two-way table

From the table, we see that there is a significant amount of interaction between the variables, Education and Occupation. As p value = 2.232500e-05 is lesser than the significance level (alpha = 0.05), we reject the null hypothesis. Thus, we see that there is an interaction effect between education and occupation on the mean salary.

Model summary

OLS Regression Results

Fig 37 model sum 2-way

Dep. Variable:	Salary	**R-squared:**	0.874
Model:	OLS	**Adj. R-squared:**	0.831
Method:	Least Squares	**F-statistic:**	20.17
Date:	Sat, 15 May 2021	**Prob (F-statistic):**	1.82e-10
Time:	05:50:03	**Log-Likelihood:**	-457.97
No. Observations:	40	**AIC:**	937.9
Df Residuals:	29	**BIC:**	956.5
Df Model:	10		
Covariance Type:	nonrobust		

Fig 38 model sum continued

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.707e+05	1.54e+04	11.088	0.000	1.39e+05	2.02e+05
C(Education)[T. Doctorate]	-4253.2500	2.04e+04	-0.209	0.836	-4.59e+04	3.74e+04
C(Education)[T. HS-grad]	-9.195e+04	2.18e+04	-4.223	0.000	-1.36e+05	-4.74e+04
C(Occupation)[T. Exec-managerial]	2.249e+04	2.04e+04	1.104	0.279	-1.92e+04	6.41e+04
C(Occupation)[T. Prof-specialty]	-6.492e+04	2.04e+04	-3.188	0.003	-1.07e+05	-2.33e+04
C(Occupation)[T. Sales]	2.159e+04	2.04e+04	1.060	0.298	-2.01e+04	6.32e+04
C(Education)[T. Doctorate]:C(Occupation)[T. Exec-managerial]	2.383e+04	3.61e+04	0.660	0.514	-5e+04	9.77e+04
C(Education)[T. HS-grad]:C(Occupation)[T. Exec-managerial]	3.206e-11	8.6e-12	3.728	0.001	1.45e-11	4.97e-11
C(Education)[T. Doctorate]:C(Occupation)[T. Prof-specialty]	1.462e+05	2.67e+04	5.484	0.000	9.17e+04	2.01e+05
C(Education)[T. HS-grad]:C(Occupation)[T. Prof-specialty]	8.17e+04	2.98e+04	2.740	0.010	2.07e+04	1.43e+05
C(Education)[T. Doctorate]:C(Occupation)[T. Sales]	5869.1000	2.71e+04	0.217	0.830	-4.96e+04	6.13e+04
C(Education)[T. HS-grad]:C(Occupation)[T. Sales]	-4.953e+04	2.98e+04	-1.661	0.107	-1.11e+05	1.14e+04

Omnibus:	7.354	**Durbin-Watson:**	2.514
Prob(Omnibus):	0.025	**Jarque-Bera (JB):**	6.580
Skew:	0.686	**Prob(JB):**	0.0373
Kurtosis:	4.437	**Cond. No.**	2.08e+17

ID added

	Education	IDN	Occupation	Salary
0	Doctorate	1	Adm-clerical	153197
1	Doctorate	2	Adm-clerical	115945
2	Doctorate	3	Adm-clerical	175935
3	Doctorate	4	Adm-clerical	220754
4	Doctorate	5	Sales	170769

Fig 39 ID data

Grand mean, Education

Fig 40 grand mean edu

Grand Mean	162186.875		
Education	Bachelors	Doctorate	HS-grad
Salary	165152.93	208427.0	75038.78
IDN	15.00	16.0	9.00

Grand mean, Occupation

Fig 41 grand mean occu

Grand Mean	162186.875			
Occupation	Adm-clerical	Exec-managerial	Prof-specialty	Sales
Salary	141424.3	197117.6	168953.15	157604.42
IDN	10.0	5.0	13.00	12.00

Mean salary by combinations

Fig 42 mean sal combo

		Salary	IDN
Occupation	Education		
Adm-clerical	Bachelors	170711.00	3
	Doctorate	166457.75	4
	HS-grad	78759.67	3
Exec-managerial	Bachelors	193201.75	4
	Doctorate	212781.00	1
	HS-grad	NaN	0
Prof-specialty	Bachelors	105787.75	4
	Doctorate	247772.83	6
	HS-grad	95534.33	3
Sales	Bachelors	192300.75	4
	Doctorate	193916.60	5
	HS-grad	50822.33	3

Boxplots of interaction, on Education, Occupation

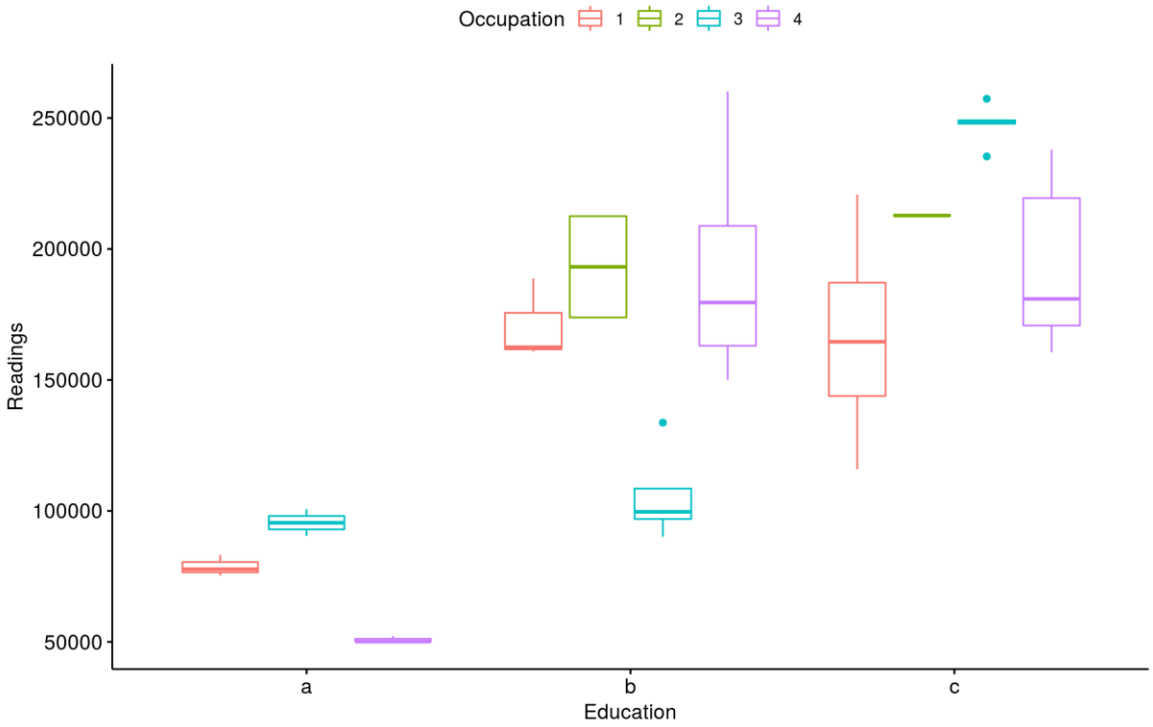
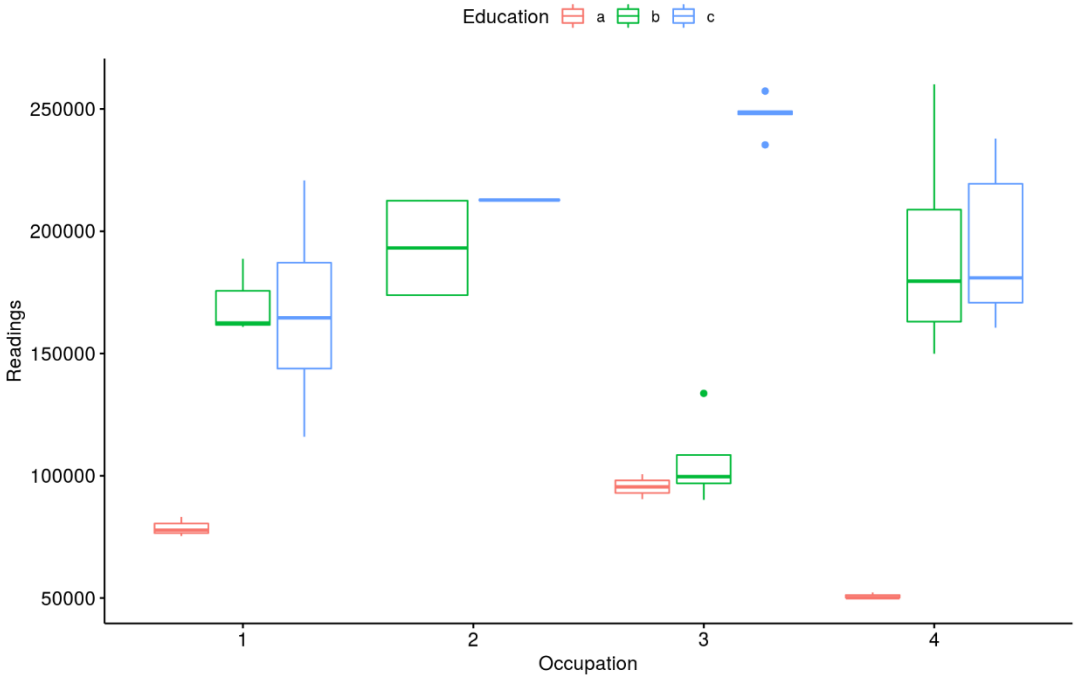


Fig 43 interaction boxplots



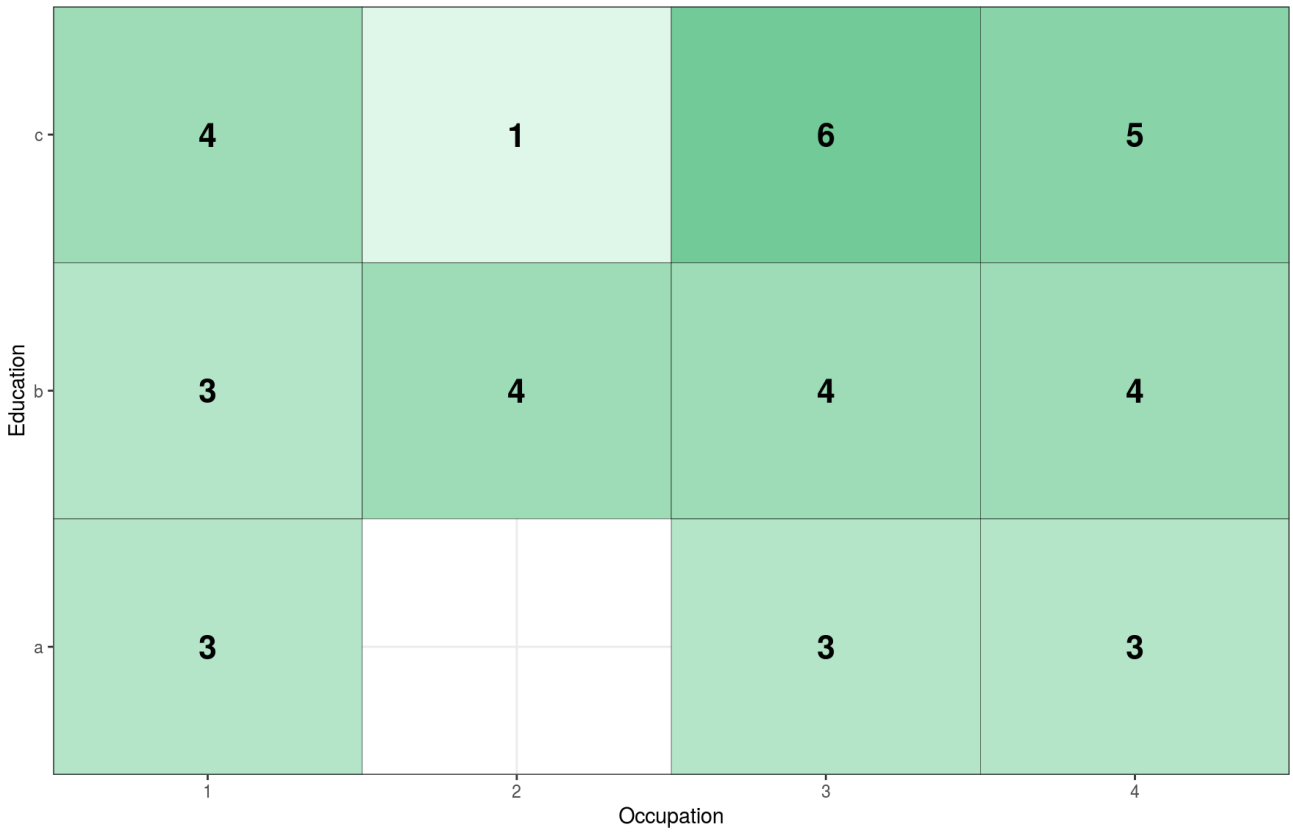
Legend

a=HS-grad, b=Bachelors c=Doctorate, Readings=Salary
1=Adm-clerical 2=Exec-managerial 3=Prof-specialty 4=Sales

Multi-comparison post-hoc Tukey's test for combination factors

Group sizes

Fig 44 group sizes



Legend

a=HS-grad, b=Bachelors c=Doctorate, Readings=Salary
1=Adm-clerical 2=Exec-managerial 3=Prof-specialty 4=Sales

Crosstab

Occupation Education	Adm-clerical	Exec-managerial	Prof-specialty	Sales
Bachelors	3	4	4	4
Doctorate	4	1	6	5
HS-grad	3	0	3	3

Fig 45 crosstab

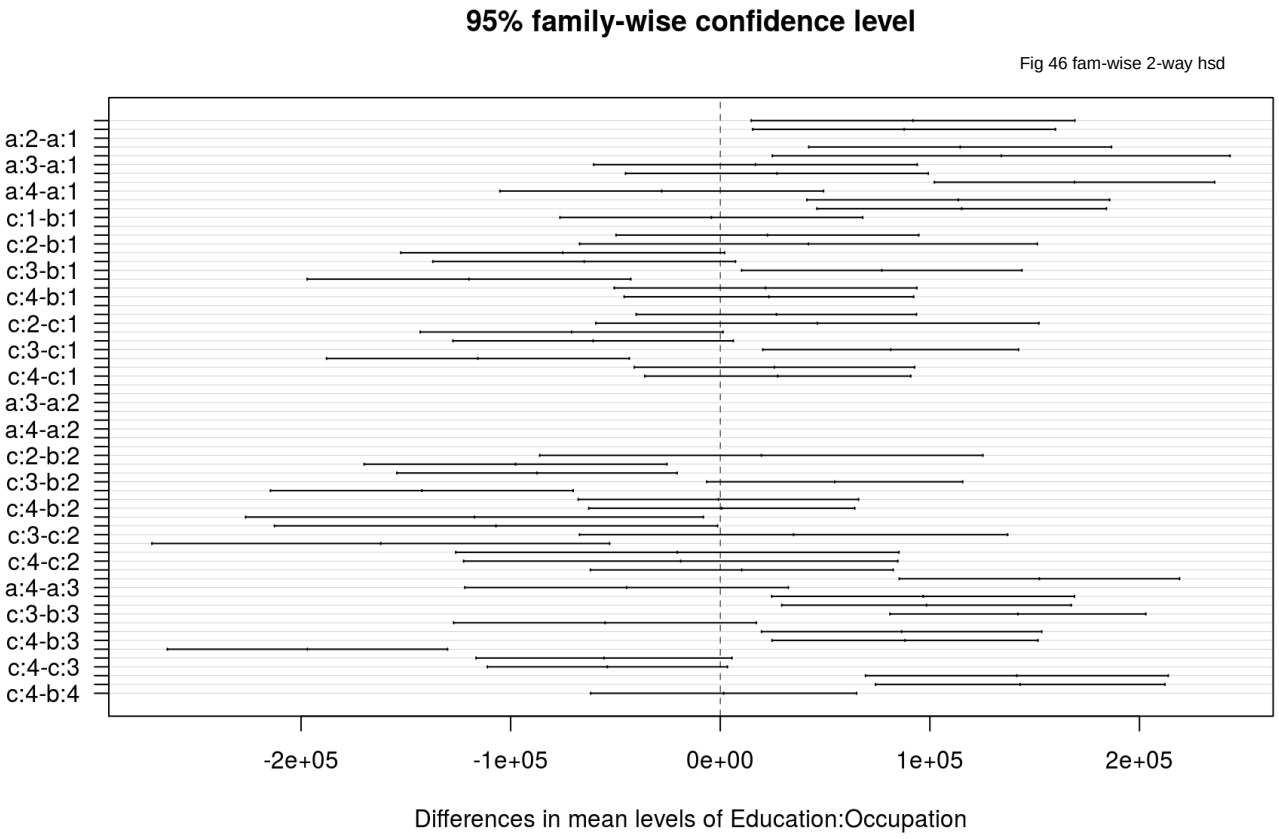


Fig 46 fam-wise 2-way hsd

a=HS-grad, b=Bachelors c=Doctorate
1=Adm-clerical 2=Exec-managerial 3=Prof-specialty 4=Sales

Significant two-way combinations are marked on the left.

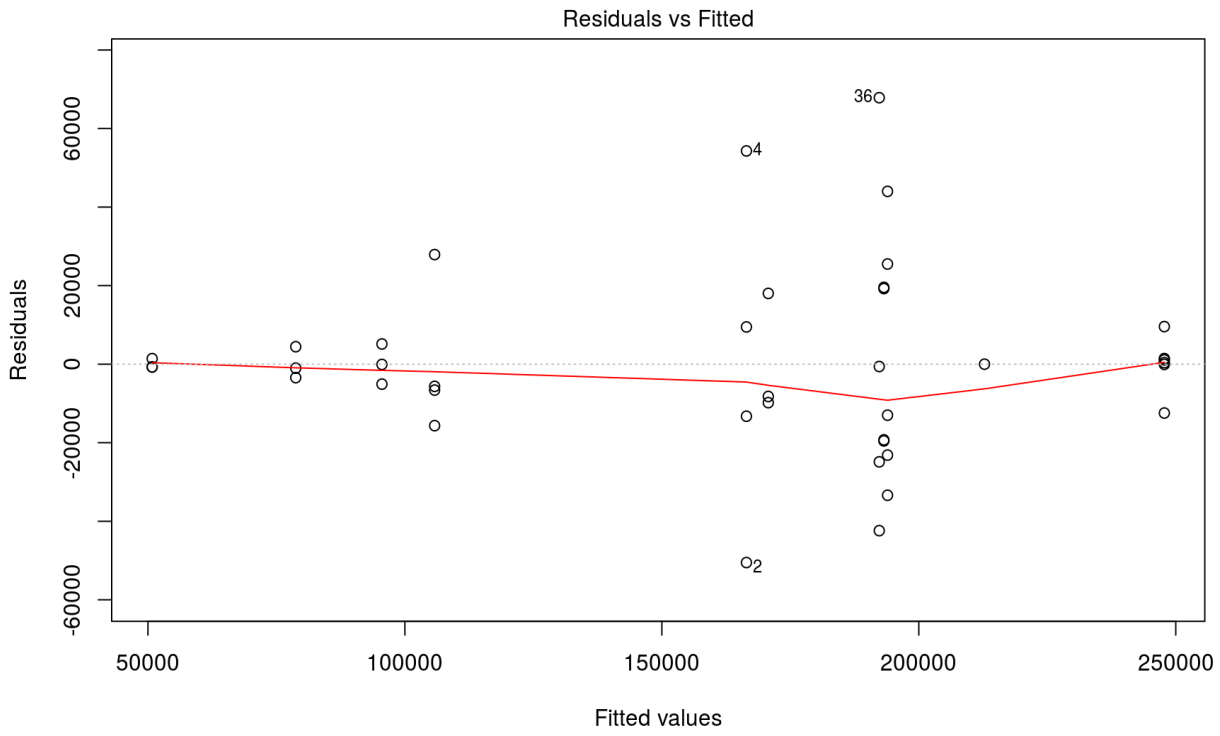
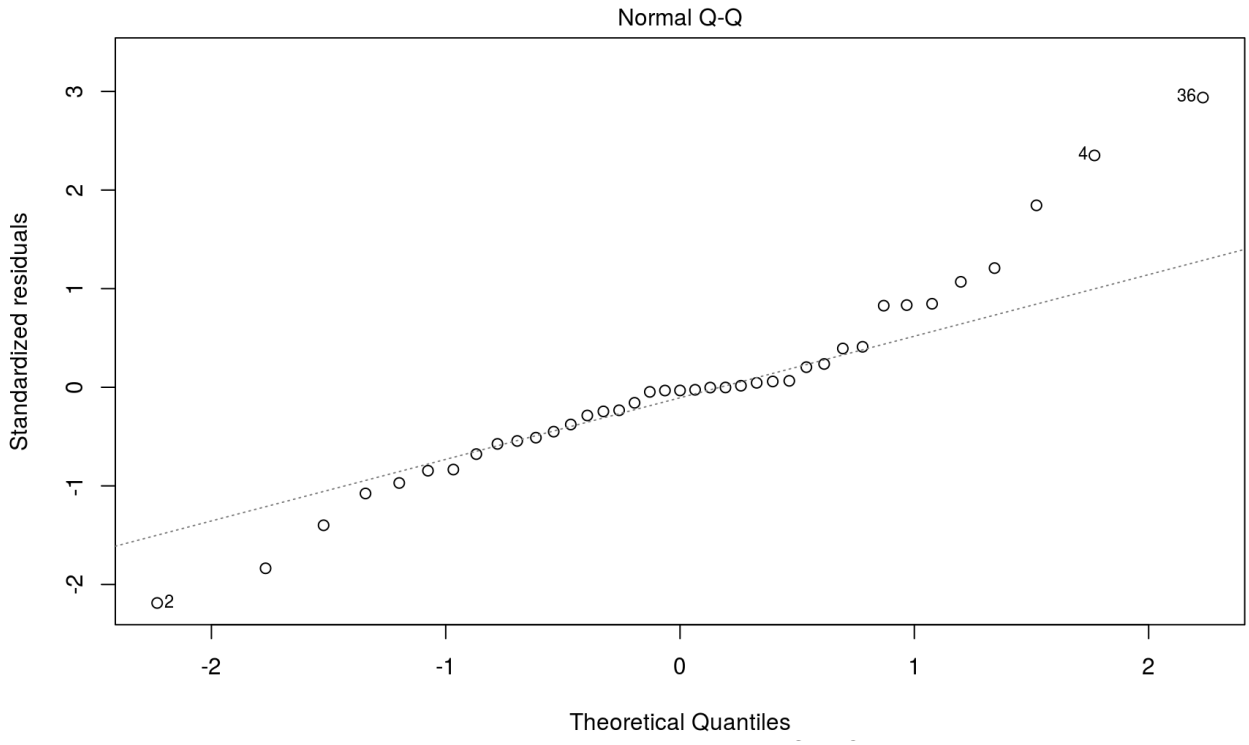


Fig 47 residual fitted val, qq

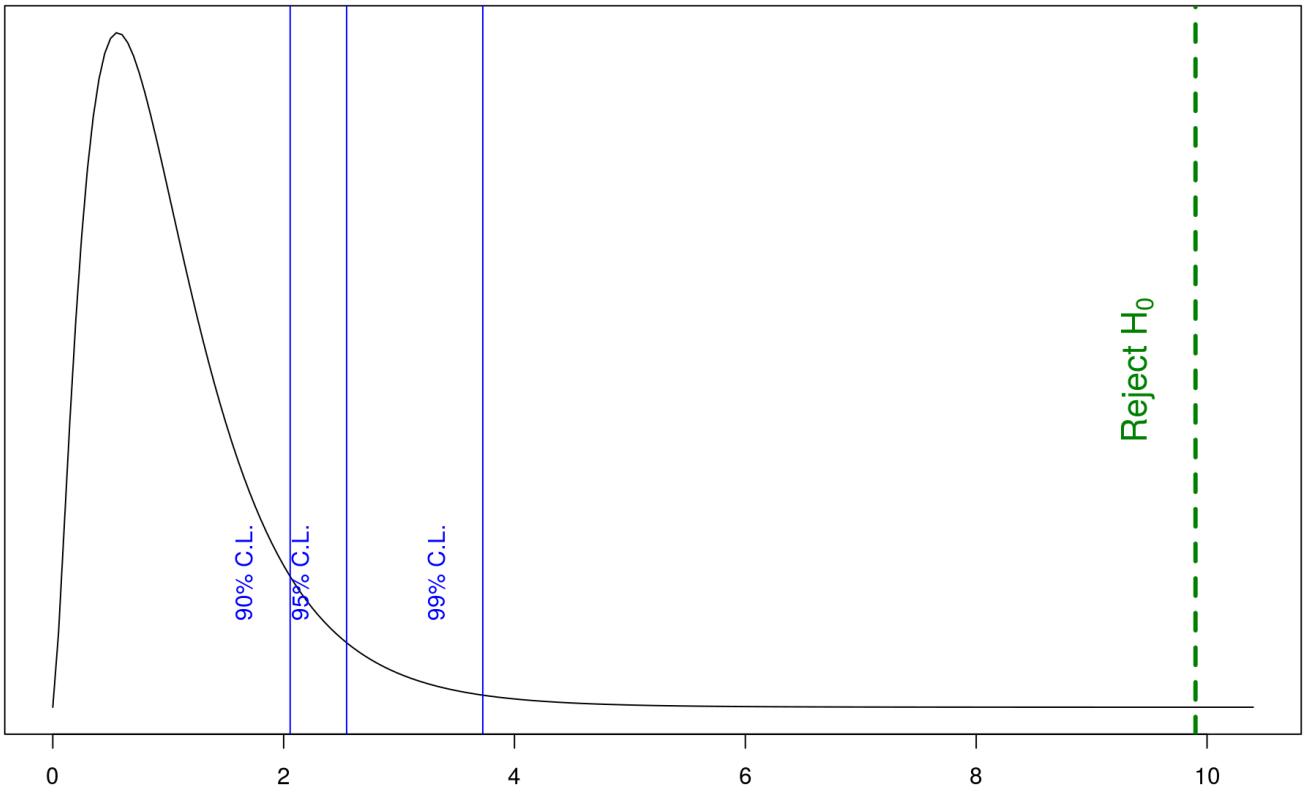


1.7 Explain the business implications of performing ANOVA for this particular case study.

From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the people. It is clearly seen that people with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least. Thus, we can conclude that Salary is dependent on educational qualifications and occupation. Explained in detail earlier.

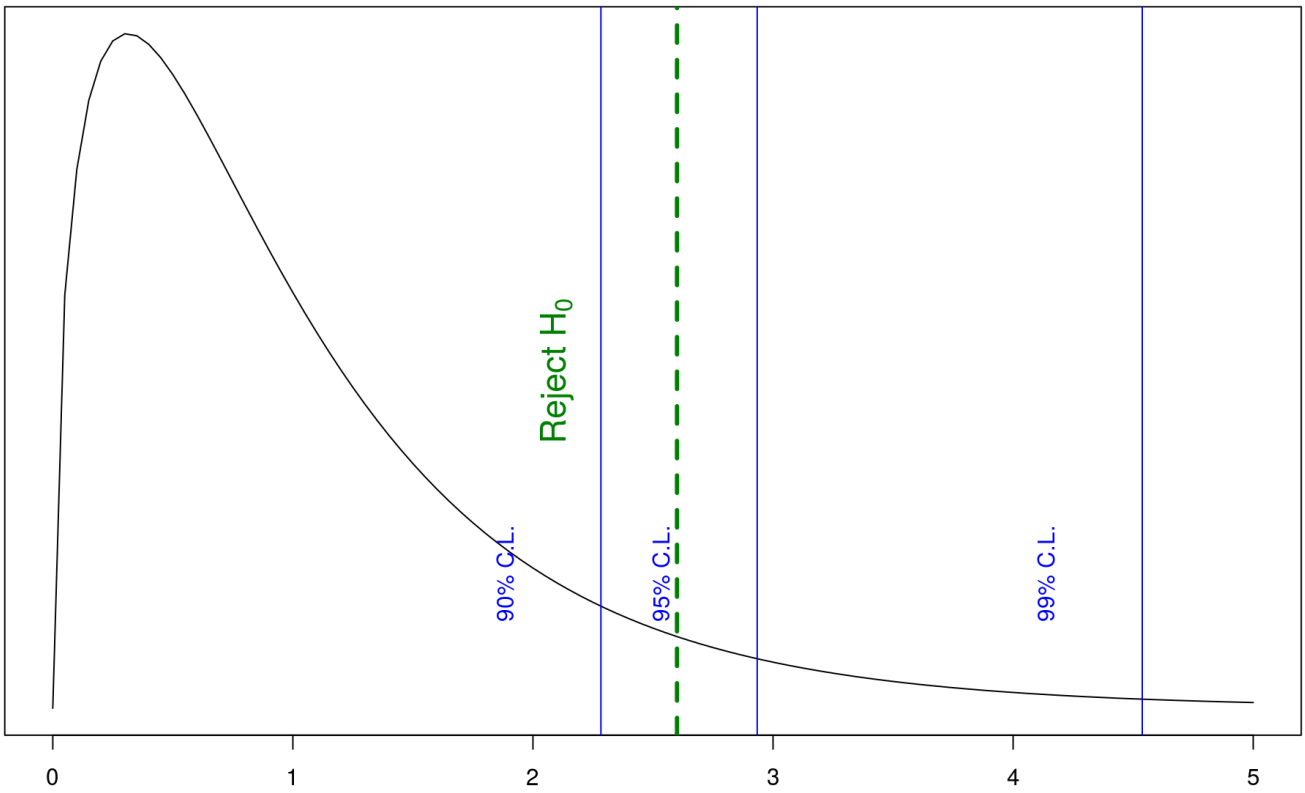
Significance of interaction

Fig 48 Sig of inter



Significance of Occupation

Fig 49 Sig of occupation



Residual values histogram

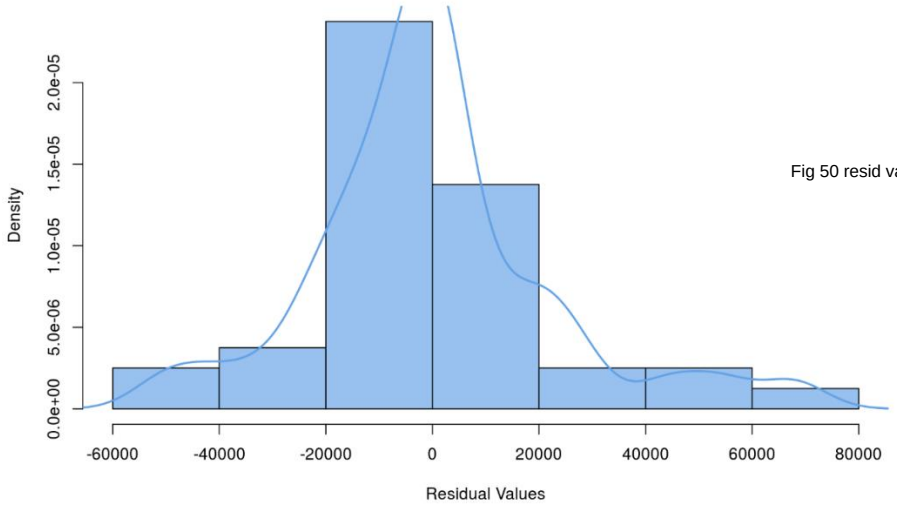


Fig 50 resid val hist

Significance of education

Fig 51 Sig of edu

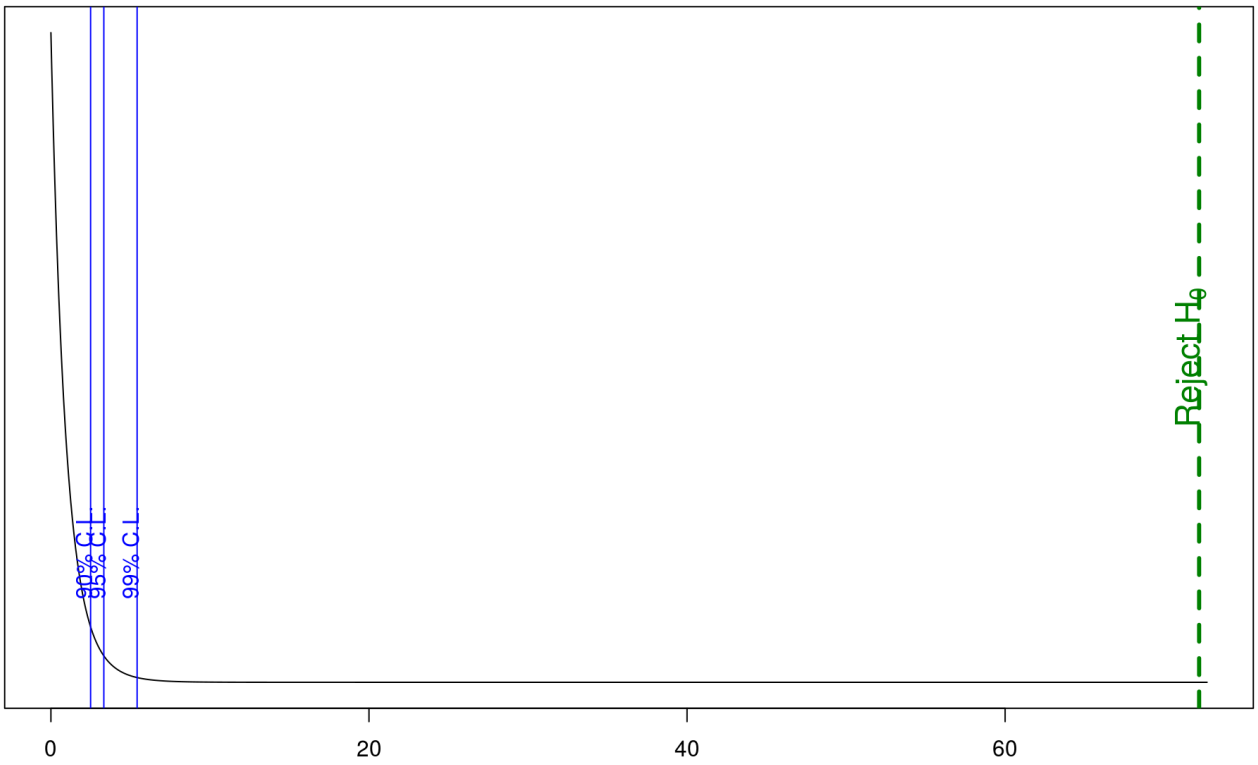


Fig 52, dash

ANOVA Model

$$y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijt}$$

where $i = 1, 2, 3,$

and $j = 1, 2, 3, 4.$

The fitted $\mu = 162186.9$ is the sample mean.

The Coefficients of Education

i	α_i
1	-87148.1
2	2966.1
3	46240.1

The Coefficients of Occupation

j	β_j
1	-20762.6
2	34930.7
3	6766.3
4	-4582.5

The Interaction Terms

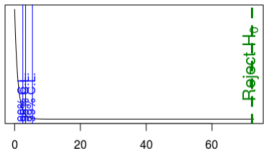
γ_{ij}	1	2	3	4
1	24483.5	NaN	13729.3	-19633.9
2	26320.6	-6881.9	-66131.5	31730.3
3	-21206.6	-30576.7	32579.6	-9927.9

ANOVA F-tests

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Education	2	102695466736	51347733368	72.2	**0.000**
Occupation	3	5519946053	1839982018	2.6	**0.072**
Education:Occupation	5	35231592718	7046318544	9.9	**0.000**
Residuals	29	20621020503	711069672	NA	**NA**

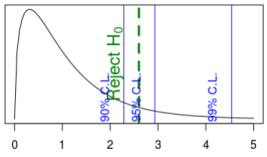
Significance of Education

H_0: Education is NOT significant



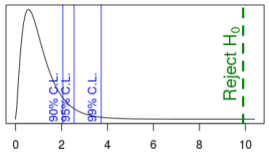
Significance of Occupation

H_0: Occupation is NOT significant



Significance of Interaction

H_0: The interaction is NOT significant



87.4% R square

62.6% of variability explained by Education

3.36% of variability explained by Occupation

21.5% of variability explained by Interaction

83.1% Adjusted R square

Main Results

✓ **Education is significant with a probability of 99%**
✓ **Occupation is significant with a probability of 90%**
✓ **The interaction between Education and Occupation is significant with a probability of 99%**

– **The model explains 83.1% of the variation**

Using statistical calculator

Problem 2:

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.

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

The following are our observations after initial exploration of the data:

Fig 53 EDA first look

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
772	Worcester State College	2197	1515	543	4	26	3089	2029	6797	3900	500	1200	60	60	21.0
773	Xavier University	1959	1805	695	24	47	2849	1107	11520	4960	600	1250	73	75	13.3
774	Xavier University of Louisiana	2097	1915	695	34	61	2793	166	6900	4200	617	781	67	75	14.4
775	Yale University	10705	2453	1317	95	99	5217	83	19840	6510	630	2115	96	96	5.8
776	York College of Pennsylvania	2989	1855	691	28	63	2988	1726	4990	3560	500	1250	75	75	18.1

• The dataset consists of 777 rows and 18 columns.

• The ‘Names’ field is an object data type, and all the remaining 17 fields are numeric fields.

• The field ‘S.F. Ratio’ is a float data type and the remaining 16 numeric fields are integer data type. • There are no missing values in the data.

5-number summary

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Fig 54 EDA 5 pt

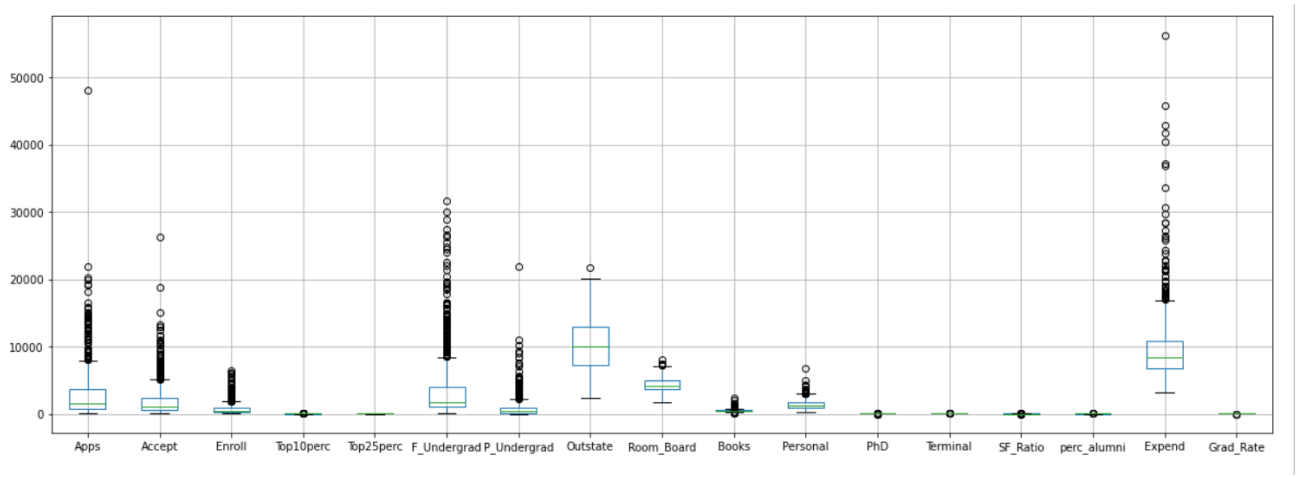
• Also, there are no bad data which is seen from the output of the ‘info’ command.

• There are no duplicate records in the data set.

Outliers

• There are lots of outliers in the data as evident from the boxplots seen in Figure 1. (No outlier treatment is being done as per the instructions-FAQ in the question) Figure 1- Boxplots for all the attributes in the data set Treating Anomalies

Fig 55 EDA outlier



• From the description of the data, we see us that there are a few anomalies in the data. The graduation rate Grad. Rate has a data with value 118 and percentage of faculty with PhDs have a value 103 which are anomalies as the upper limit can be only 100, being percentages. These two values have been imputed with the median. Data Visualization- Univariate Analysis Insights from the Data Apps: Number of applications a college/university receives

• The number of applications received by a college ranges from 81 to 48094.

• It is seen that Rutgers at New Brunswick receives the maximum number of applications (48094) and Christendom College receives the least (81).

• The mean number of applications is 3001 and the median is 1558.

• The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure. The below figure shows the presence of outliers in this attribute. Accept: Number of applications a college/university accepts

• The number of applications accepted by a college ranges from 72 to 26330.

• It is seen that Rutgers at New Brunswick accepts the maximum number of applications

• (26330) and Christendom College accepts the least (79).

• The mean number of applications accepted is 2018 and the median is 1110.

• The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure. The below figure shows the presence of outliers in this attribute. Enroll: Number of students who enroll

• The number of students who enroll ranges from 35 to 6392.

• Texas A&M Univ. at College Station has maximum enrollment with 6392 students and Capitol College has the least with 35 students.

• The mean number of students enrolled is 779.97 and the median is 434. The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure. The below boxplot shows the presence of outliers in this attribute. Top 10 perc : Percentage of new students from top 10% of Higher Secondary class

• The percentage of new students from top 10% of Higher Secondary class ranges from 1 to 96.

• Massachusetts Institute of Technology has maximum percentage of new students from top 10% of Higher Secondary class with 96% and three colleges as seen in the below table have the least with 1%.

• The mean percentage of new students from top 10% of Higher Secondary class is 27% and the median is 23%.

• The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure. The below boxplot shows the presence of outliers in this attribute. Top25perc: Percentage of new students from top 25% of Higher Secondary class

• The percentage of new students from top 25% of Higher Secondary class ranges from 9 to 100.

• There are 7 universities/colleges which have the maximum percentage (100%) of new students from top 25% of Higher Secondary class and Huron University has the least with 9% as can be seen below.

• The mean percentage of new students from top 25% of Higher Secondary class is 55.7966% and the median is 54%.

• The mean is very close to the median indicating that the distribution is almost normal as seen in the below figure. The attribute has no outliers as seen in the below boxplot.F_Undergrad: Number of full-time undergraduate students

• The number of full time under graduate students ranges from 139 to 31643.

• Texas A&M University at College Station has maximum number of full time undergraduate students with 31643 students and Christendom College has the least with 139 students.

• The mean number of full-time undergraduate students is 3699 and the median is 1707.

• The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure. The below boxplot shows the presence of outliers in this attribute. P_Undergrad: Number of part-time undergraduate students

• The Number of part-time undergraduate students ranges from 1 to 21836.

• University of Minnesota Twin cities has the maximum number of part time undergraduate students with 21836 students and the below 4 colleges have the least with 1 student.

• The mean number of part time undergraduate students is 855.30 and the median is 353.

• The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure. The below boxplot shows the outliers in the data. Outstate: Number of students for whom the particular college or university is Out-of-state tuition

• The number of students for whom the particular college or university is Out-of-state tuition ranges from 2340 to 21700.

• Bennington College has the maximum number of students for whom the particular college or university is Out-of-state tuition with 21700 students and Brigham Young University at Provo has the least with 2340 students.

• The mean number of students for whom the particular college or university is Out-of-state tuition is 10440.67 and the median is 9990.

• The mean is quite close to the median indicating that the distribution is almost normal as seen in the below figure. This is also evident from the below boxplot which has only one outlier.

Room Board: Cost of Room and board

• The cost of room and board ranges from $1780 to $8124.

• Barnard College has the maximum cost of room and board ($ 8124) and North Carolina A & T State University has the least cost of room and board ($1780).

• The mean cost of room and board is $4357.53 and the median is $ 4200.

• The mean is quite close to the median indicating that the distribution is almost normal with a shorter right tail as seen in the below figure. The boxplot below shows the presence of a few outliers. Books: Estimated book costs for a student

• The estimated book costs for a student ranges from $96 to $2340.

• Center for Creative Studies has the maximum estimated book costs ($ 2340) for a student and Appalachian State University has the least book costs for a student ($ 96).

• The mean estimated book cost is $549.38 and the median is $500.

• The mean is greater than the median indicating that the distribution is right skewed as seen in the below figure. The below box plot shows the presence of outliers in the data. Personal: Estimated personal spending for a student

• The estimated personal spending for a student ranges from $250 to $6800.

• Saint Louis University has the maximum estimated personal spending for a student ($6800) and Benedictine College has the least ($ 250)

• The mean estimated personal spending for a student is $ 1340.64 and the median is $1200

• The mean is greater than the median indicating that the distribution is right skewed as seen in the below figure. The boxplot below shows the outliers in the data. PhD: Percentage of faculties with Ph.D.'s

• The percentage of faculties with Ph.D.'s ranges from 8% to 100%.

• Three colleges have the maximum percentage of faculties with Ph.D.'s (100%) and Center for Creative Studies has the least (8%). The mean percentage of faculties with Ph.D.'s is 72.62% and the median is 75%.

The mean is lesser than the median indicating that the distribution is left skewed as seen in the below figure. The boxplot below shows the presence of outliers. Terminal: Percentage of faculties with terminal degree

• The percentage of faculties with terminal degree ranges from 24% to 100%.

• The below universities/colleges have the maximum percentage of faculties with terminal Degree (100%) and Salem-Teikyo University has the least (24%).The mean percentage of faculties with terminal degree is 79.70 % and the median is 92%. The mean is lesser than the median indicating that the distribution is left skewed as seen in the below figure. The below boxplot shows the presence of outliers. S.F. Ratio: Student/faculty ratio

• The student faculty ratio ranges from 2.5 to 39.8.

• Indiana Wesleyan University has the maximum student faculty ratio (39.8) and University of Charleston has the least (2.5).

• The mean student faculty ratio is 14.09 and the median is 13.6.

• The mean is quite close to the median indicating that the distribution is almost normal with short right tail as seen below. The below boxplot shows the presence of outliers. perc.alumni: Percentage of alumni who donate

• The percentage of alumni who donate ranges from 0% to 64%

• Williams College has the maximum percentage of alumni who donate (64%) and 2 colleges as given in the below table have the least (0%). The mean percentage of alumni who donate is 22.74% and the median is 21% . The mean is quite close to the median indicating that the distribution is almost normal with a short right tail as seen below. The below plot shows the outliers. Expend: The Instructional expenditure per student

• The instructional expenditure per student ranges from $3186 to $56233.

• Johns Hopkins University has the maximum instructional expenditure per student ($56233) and Jamestown College has the least ($3186).

• The mean instructional expenditure per student is $9660.17 and the median is $ 8377.

• The mean is greater than the median indicating that the distribution is right skewed as seen in the below figure. The below figure shows the presence of outliers. Grad.Rate: Graduation rate

• The graduation rate ranges from 10% to 100%.

• The computed table gives the universities with the maximum graduation rate a score of 100% and Tex as Southern University the least score of 10%.

• The mean graduation rate is 65.40% and the median is 65%.

• The mean is almost equal to the median indicating that the distribution is normal as seen in the below figure. The outliers are seen in the boxplot.

Univariate Analysis

Fig 56 Apps count

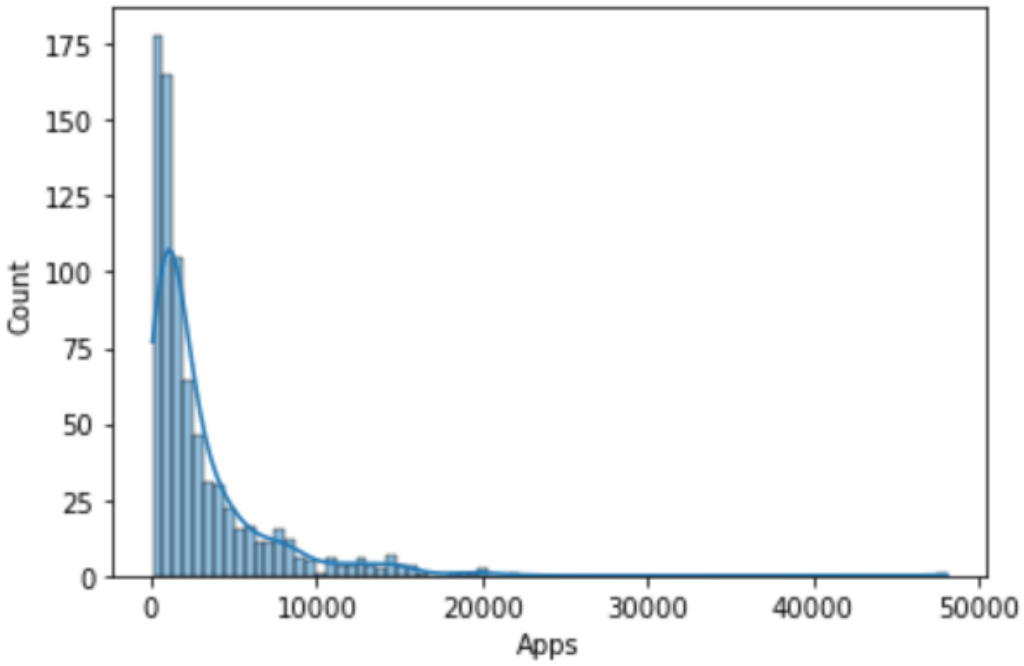


Fig 57 Apps box

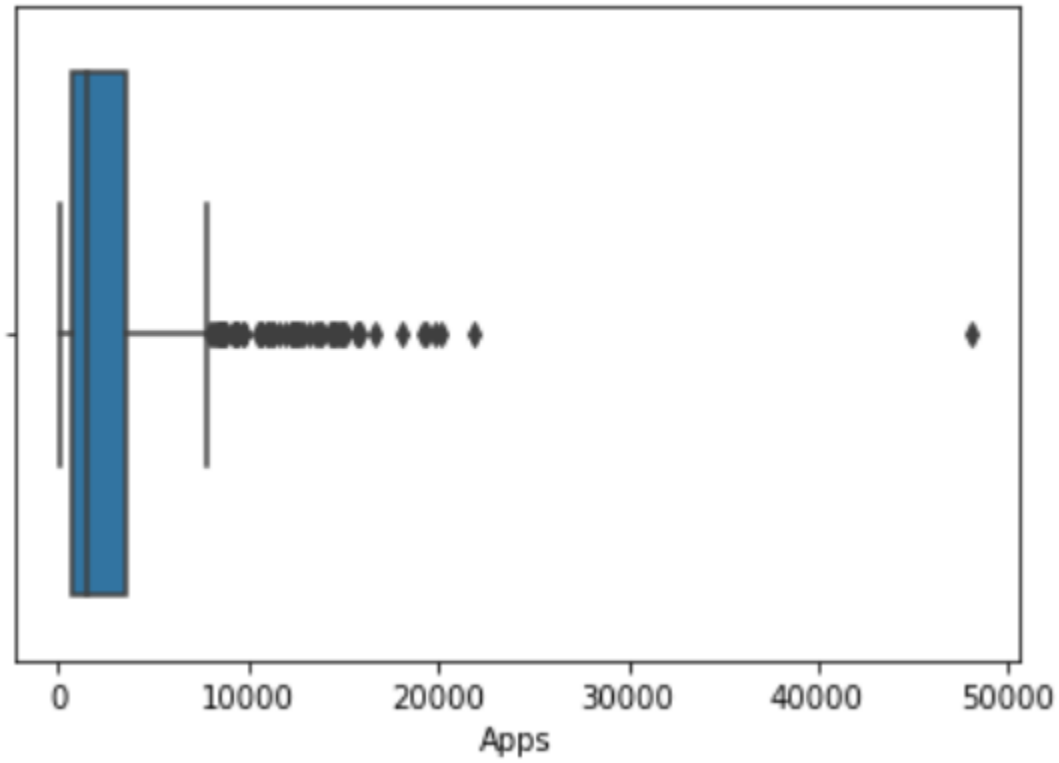


Fig 58 Accept count

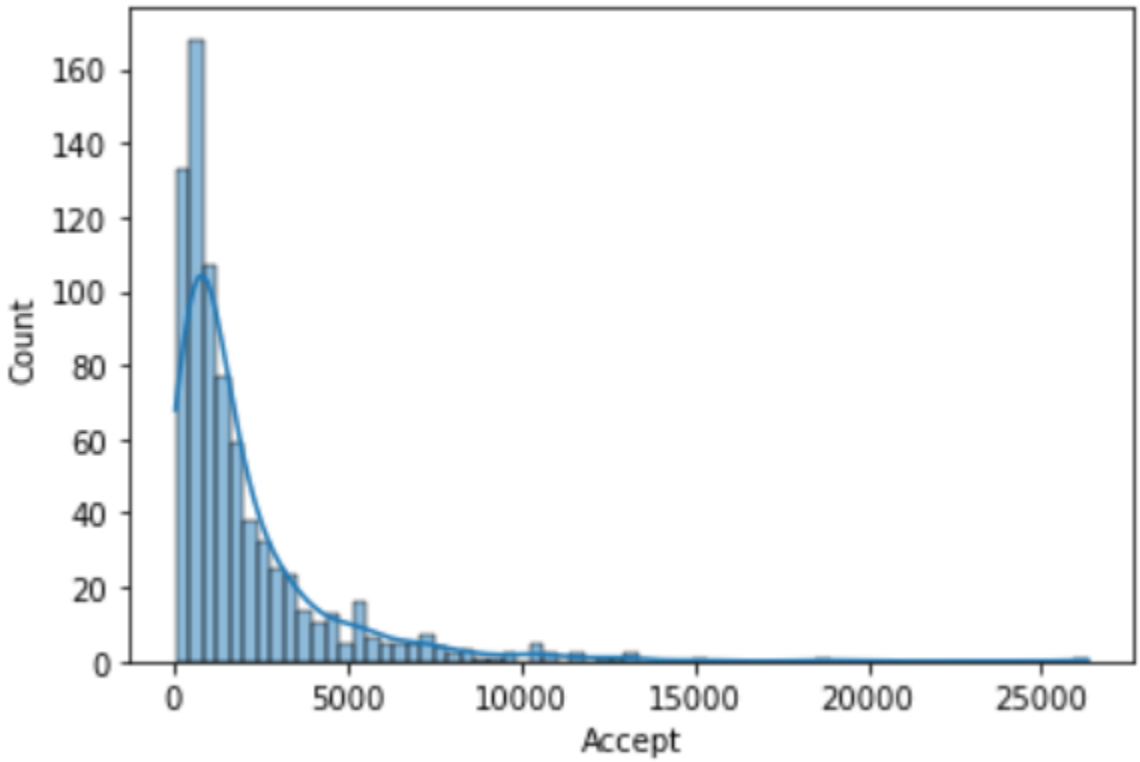


Fig 59 Enroll

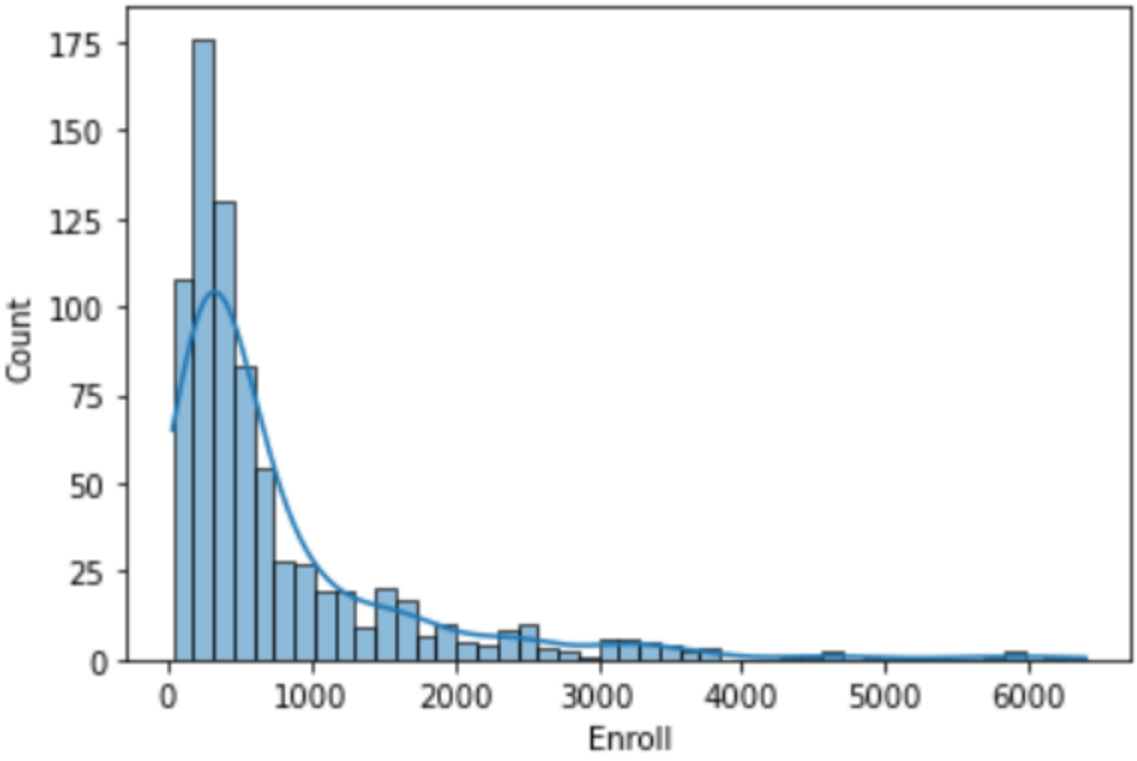


Fig 60 Enroll box

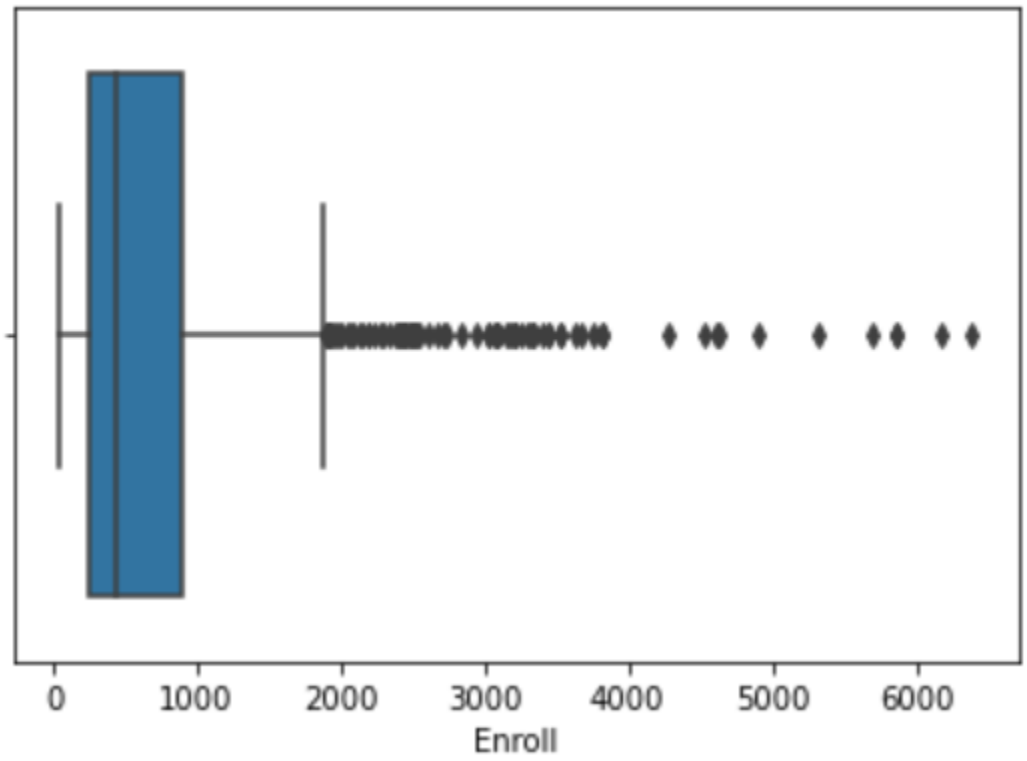


Fig 61 Top10 count

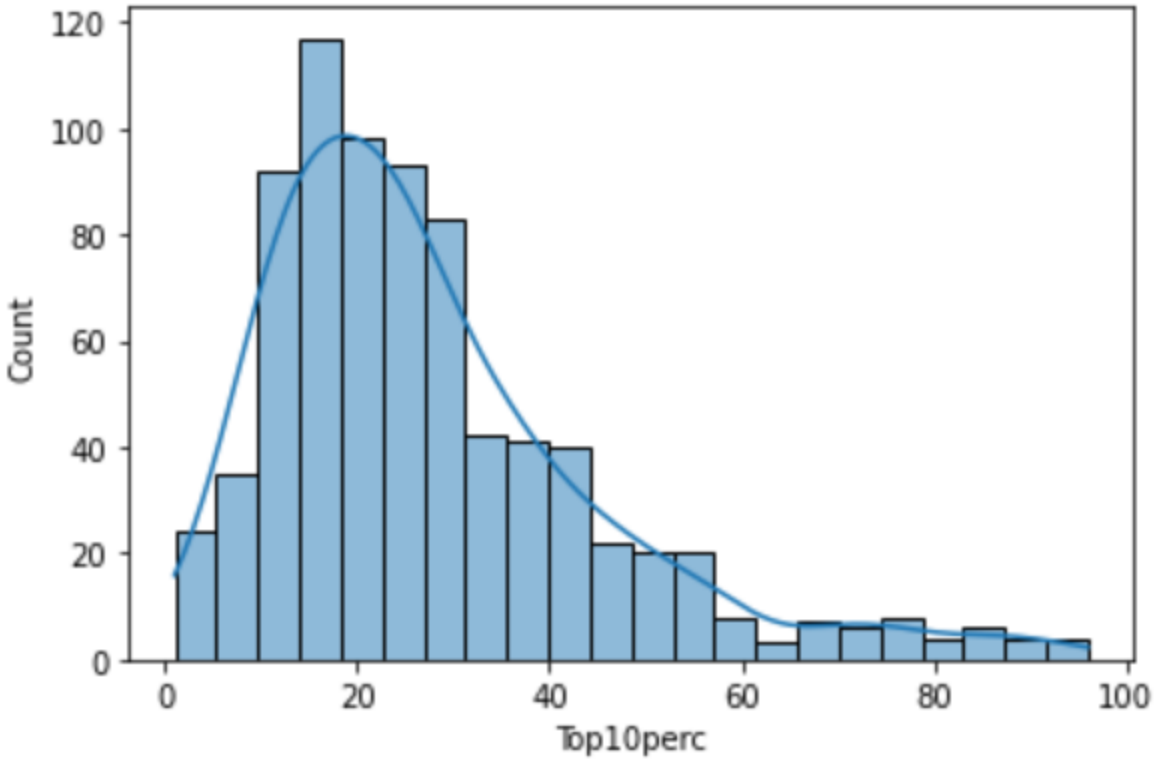


Fig 62 Top 10 box

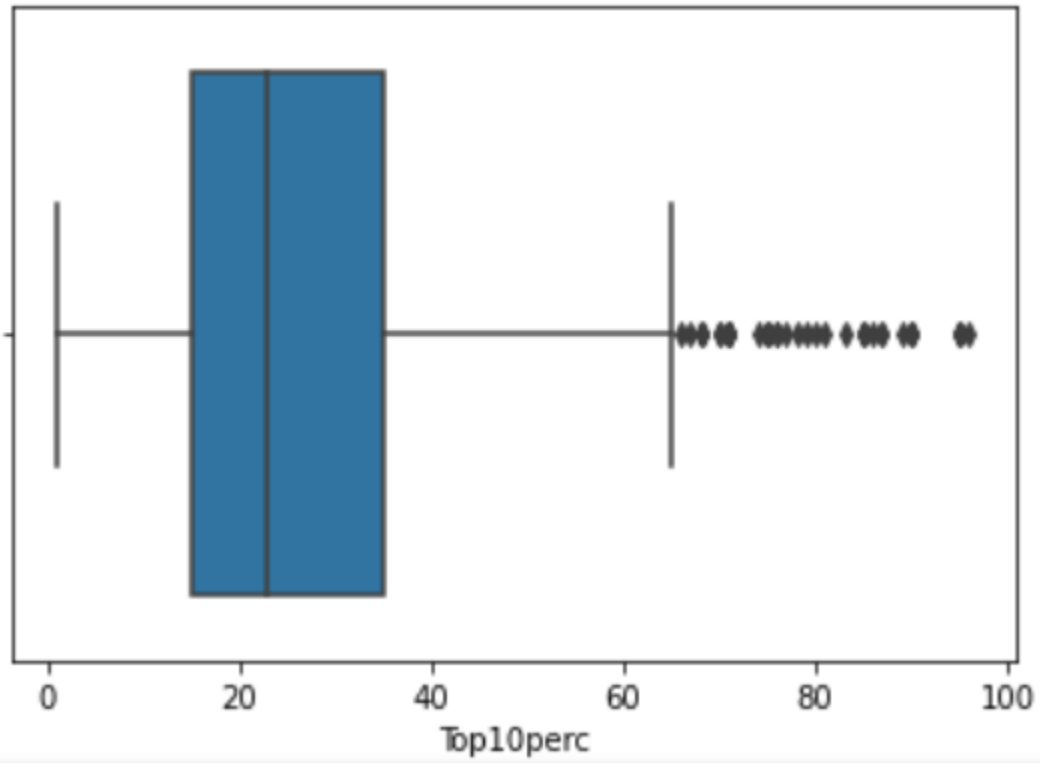


Fig 63 Top 25 count

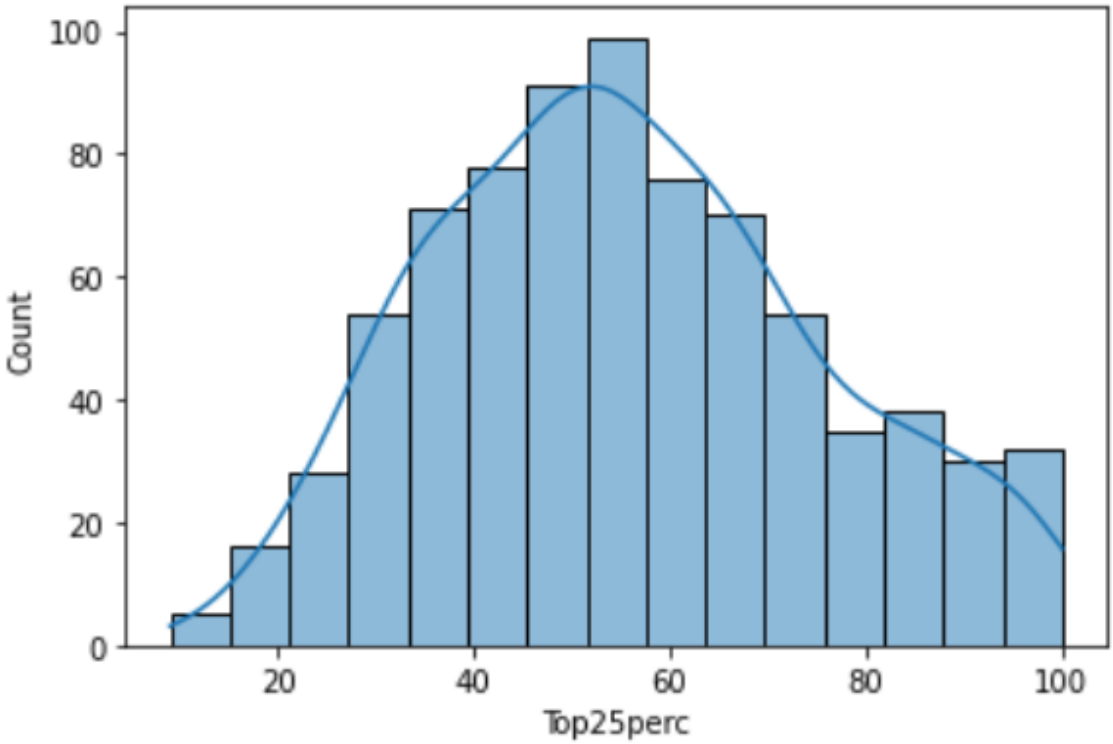


Fig 64 Top 25 box

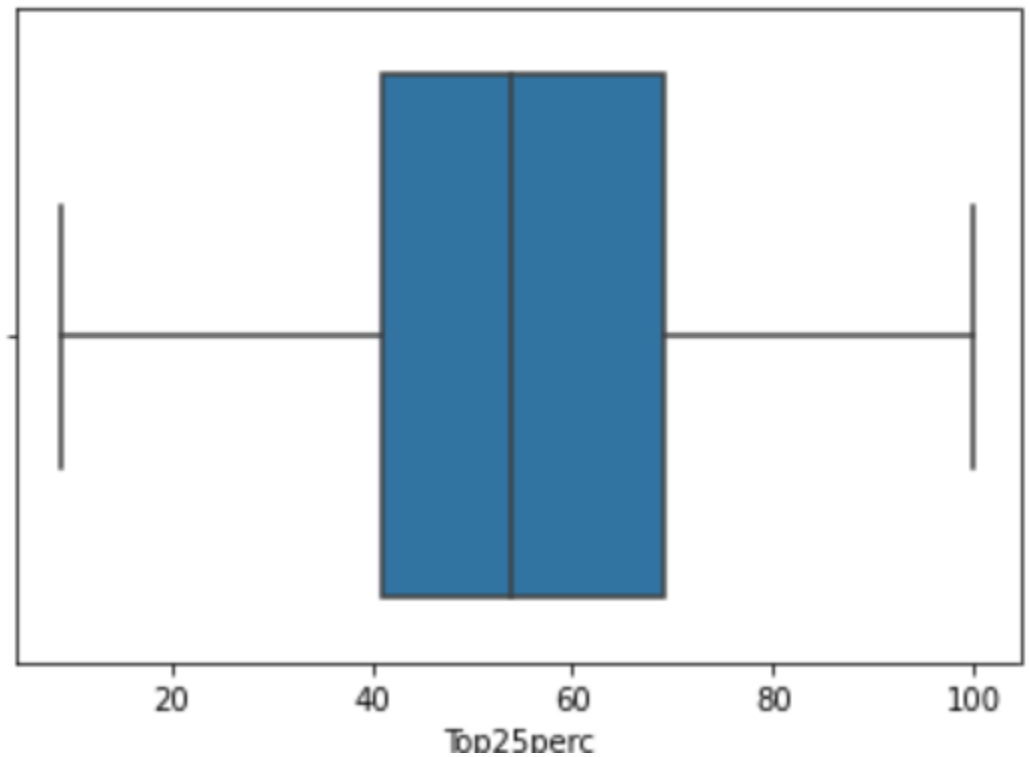


Fig 65 f-under count

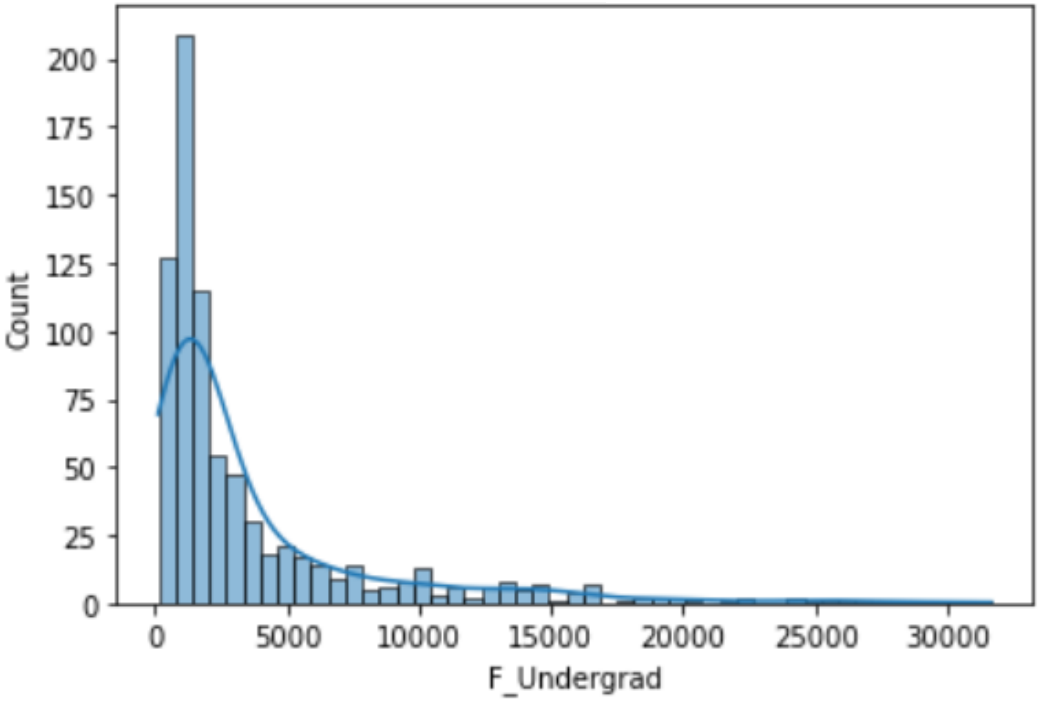


Fig 66 F under box

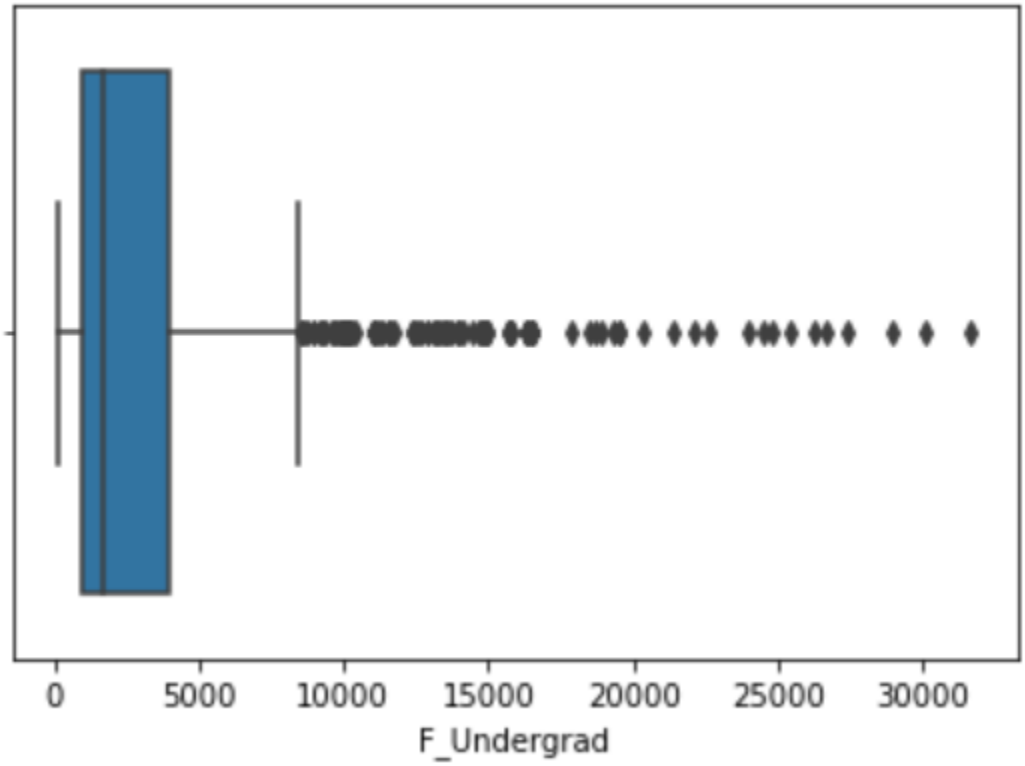


Fig 67 p under count

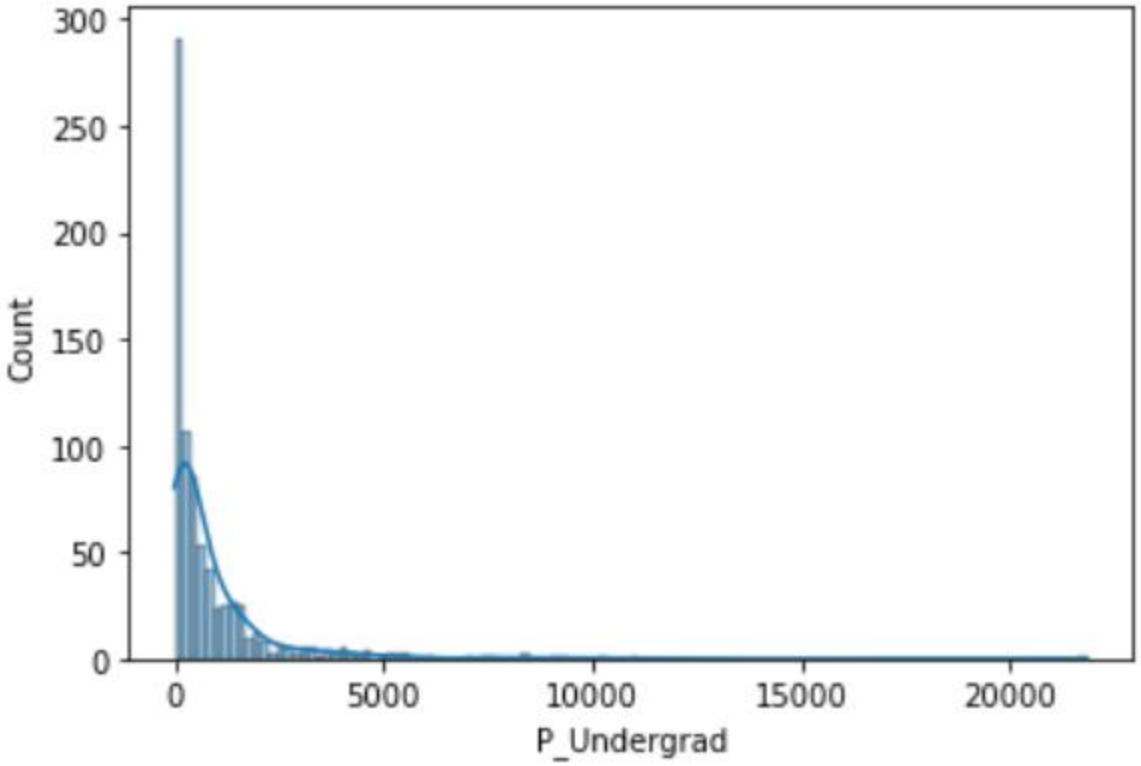


Fig 68 p-under box

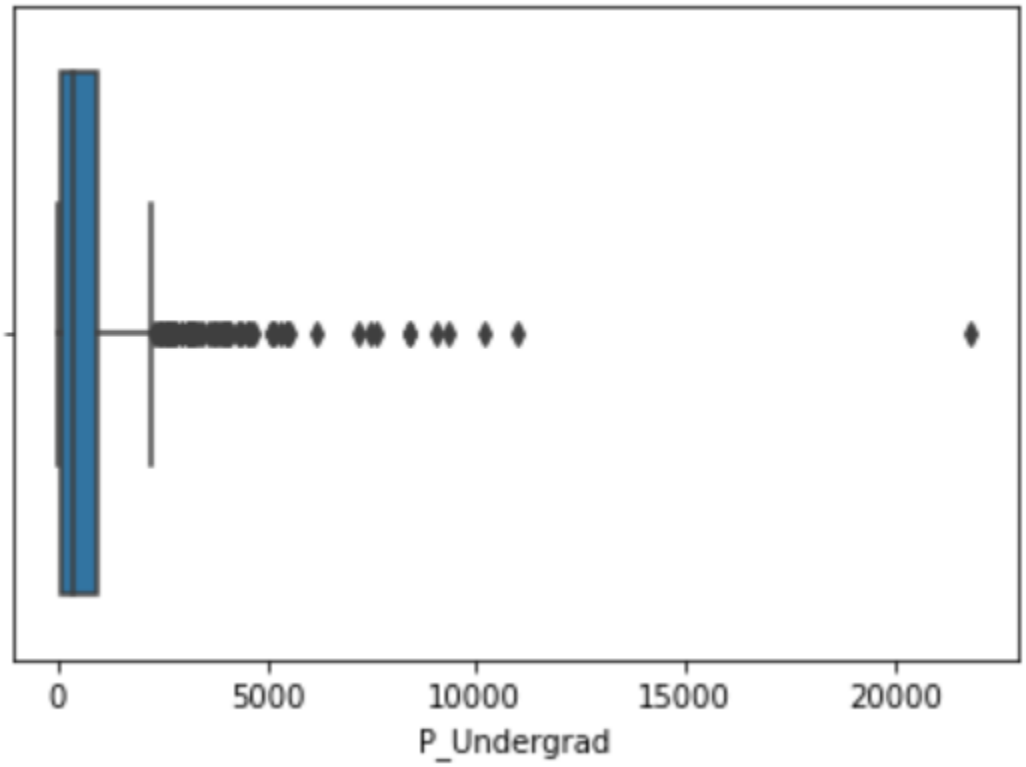


Fig 69 outstate count

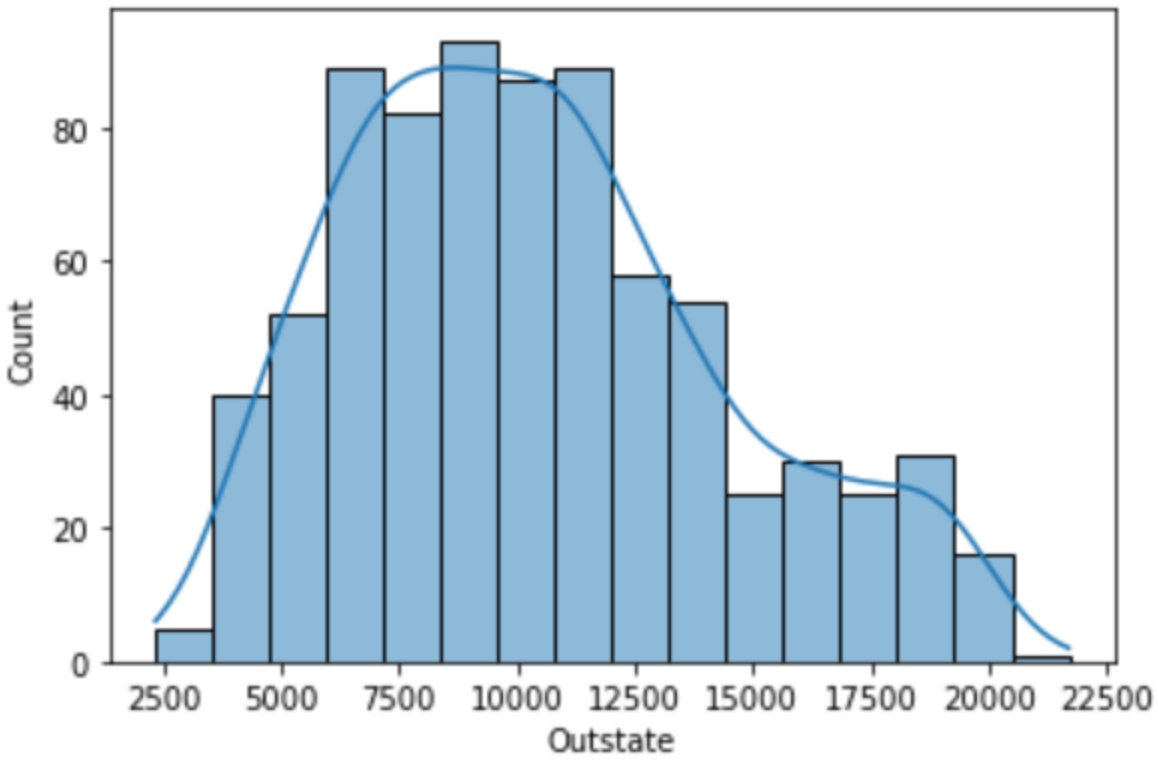


Fig 70 outstate box

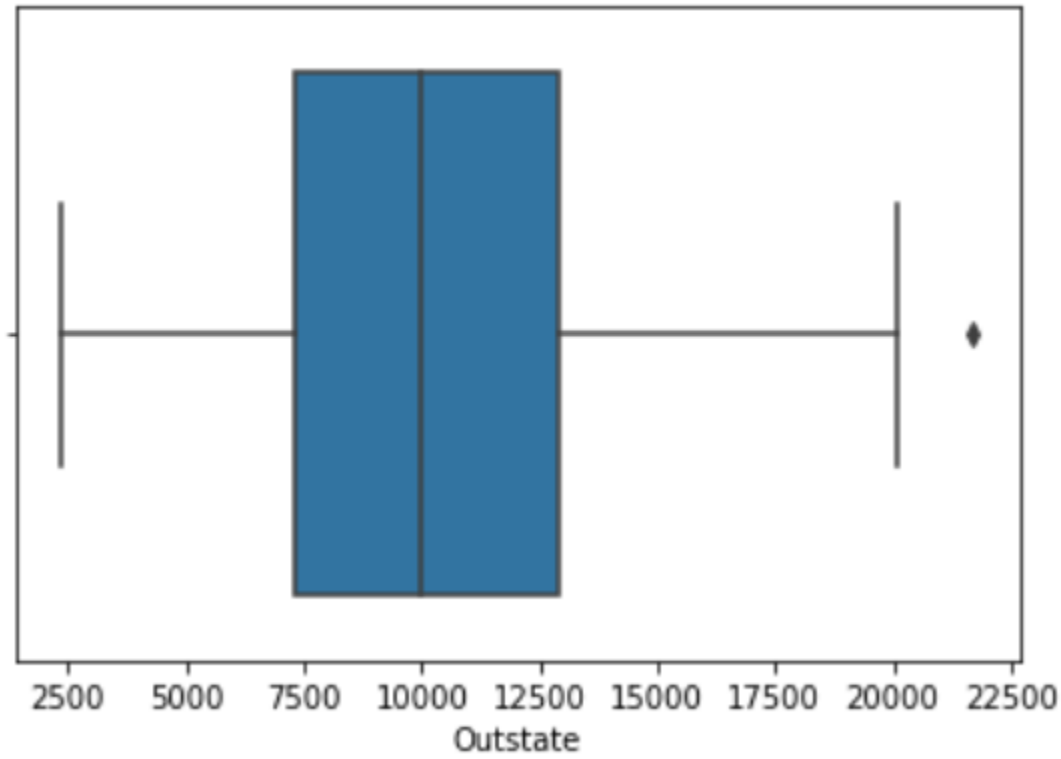


Fig 71 room count

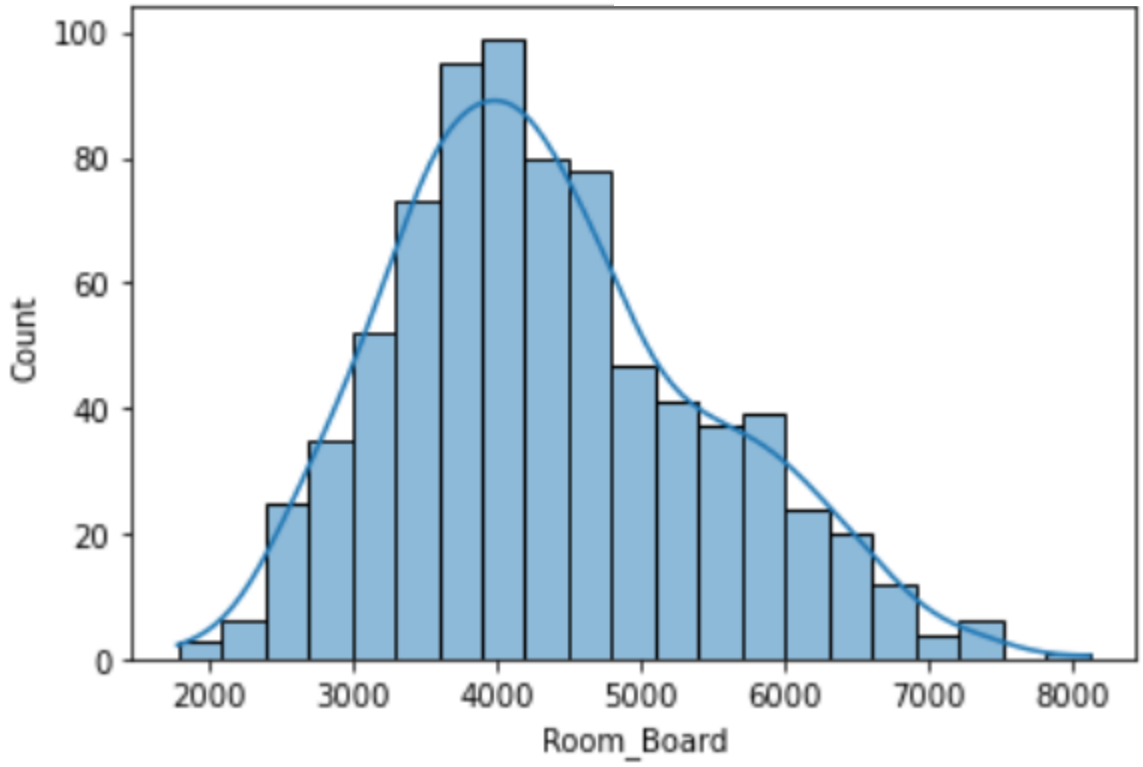


Fig 72 room board

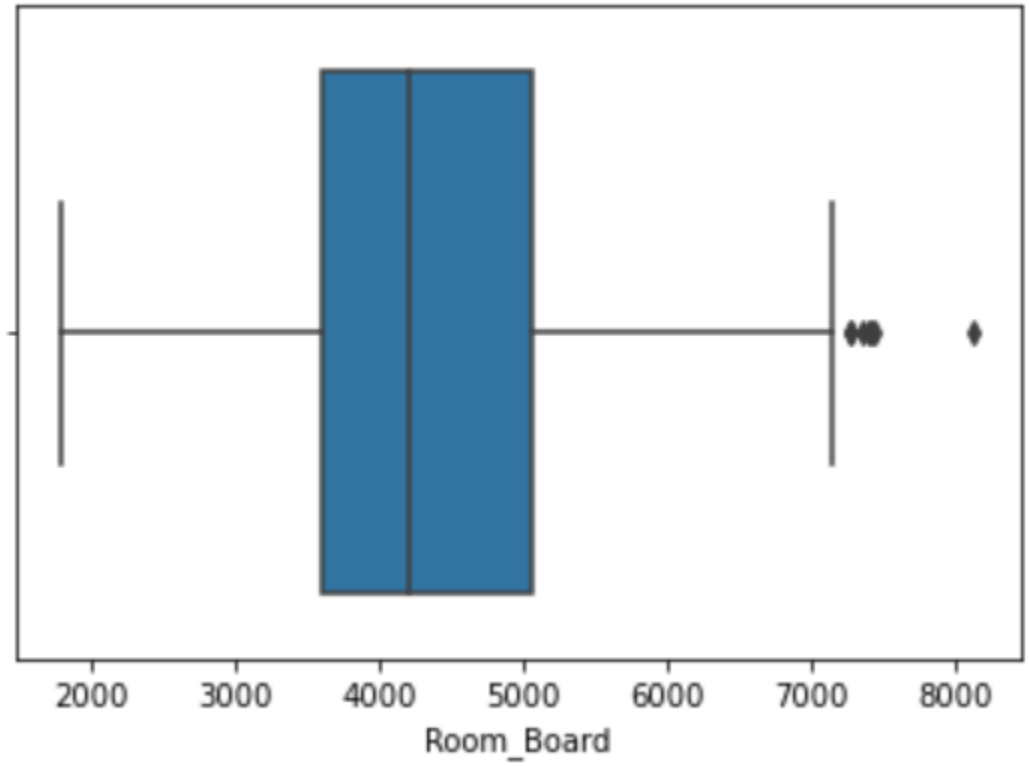


Fig 78 books

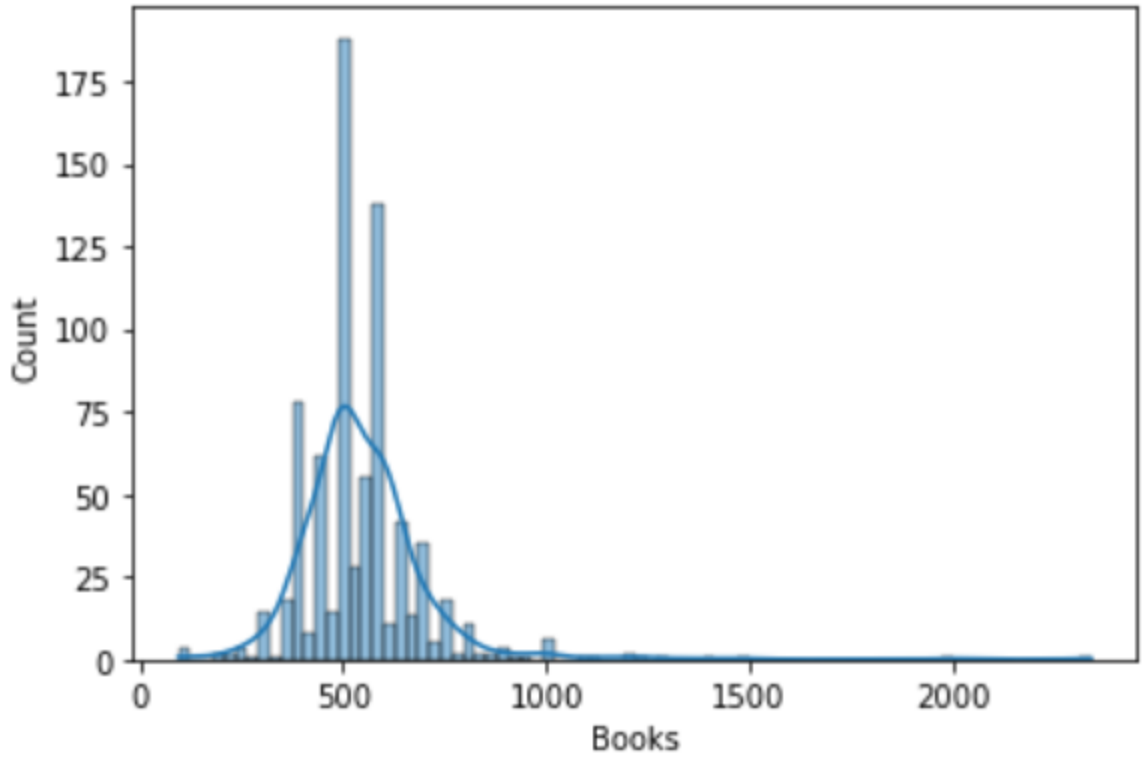


Fig 79 personal count

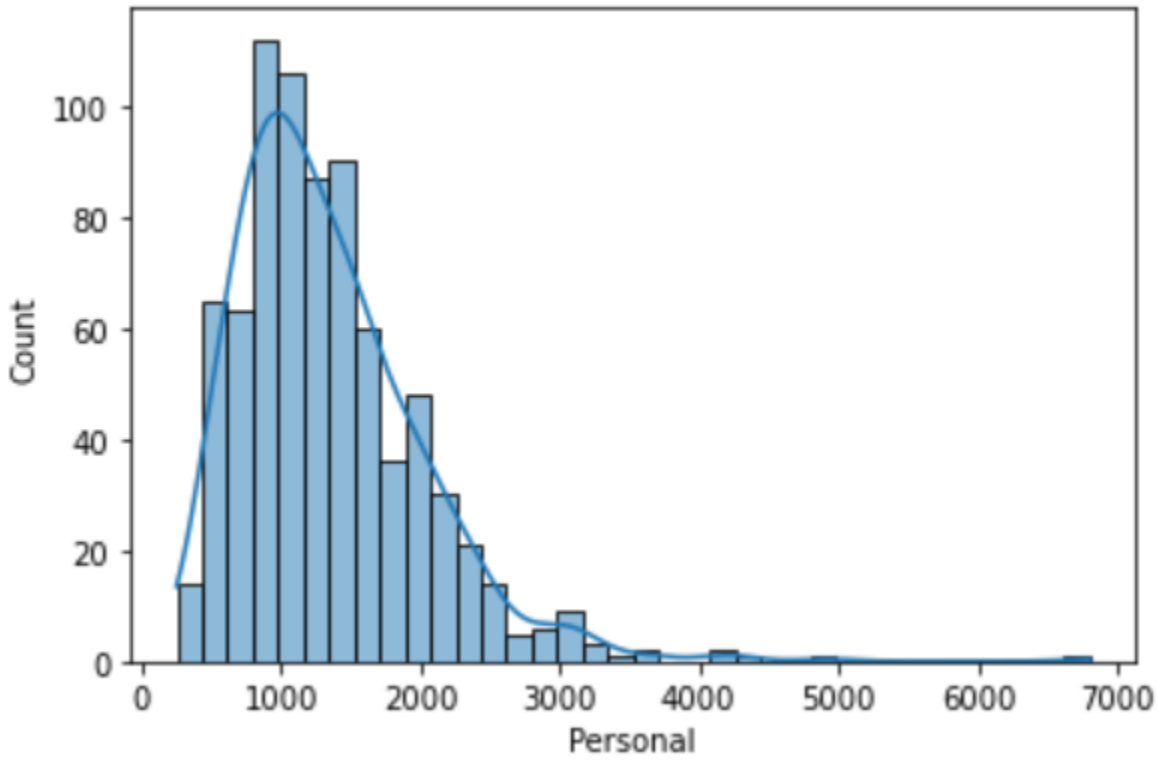


Fig 80 Phd count

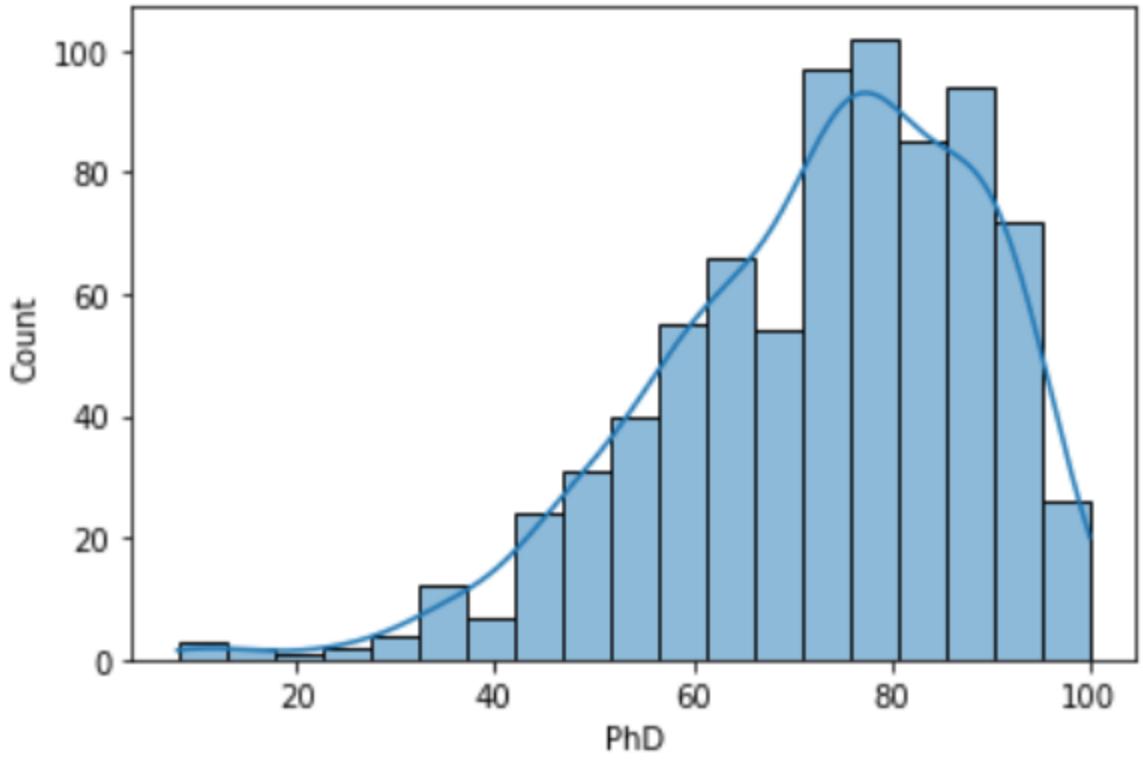


Fig 81 Terminal count

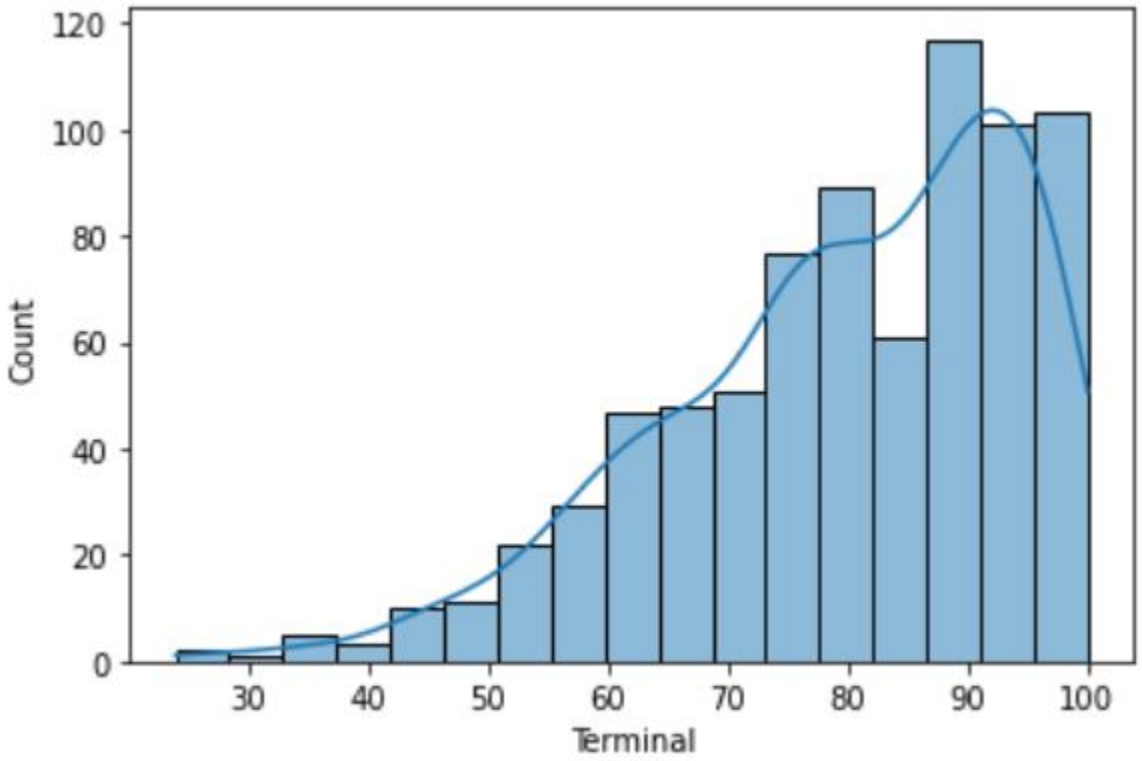


Fig 82 SF ratio count

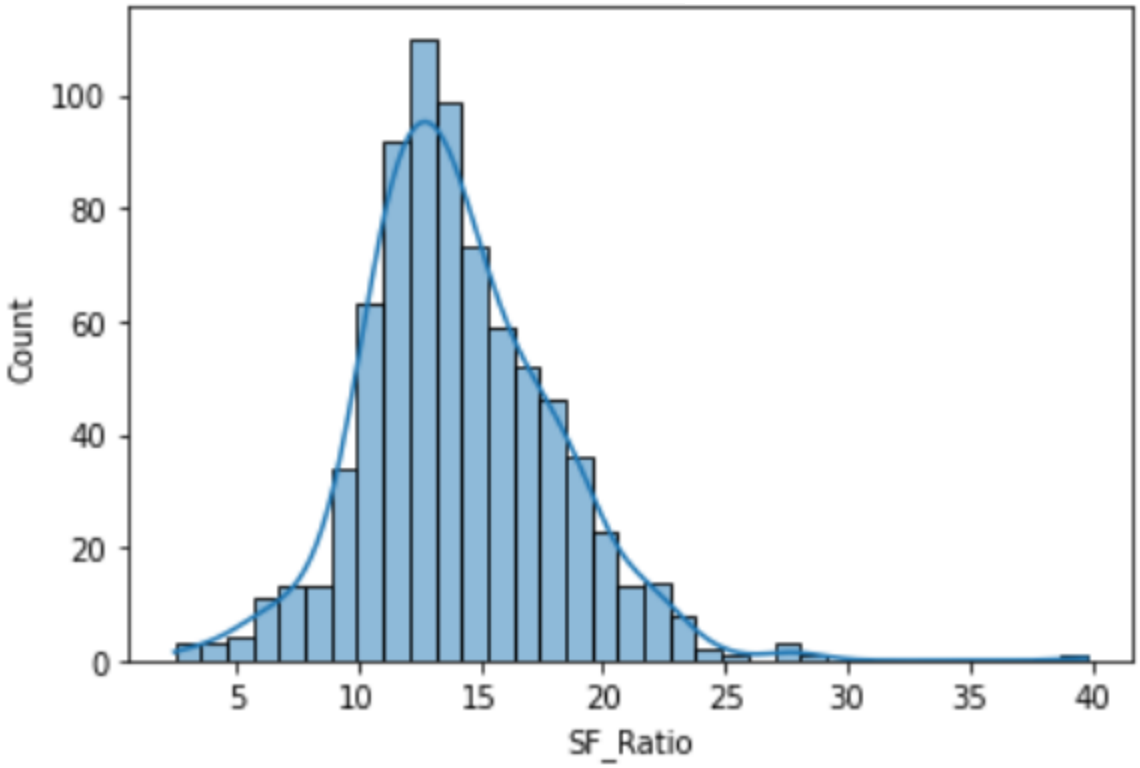


Fig 83 alum box

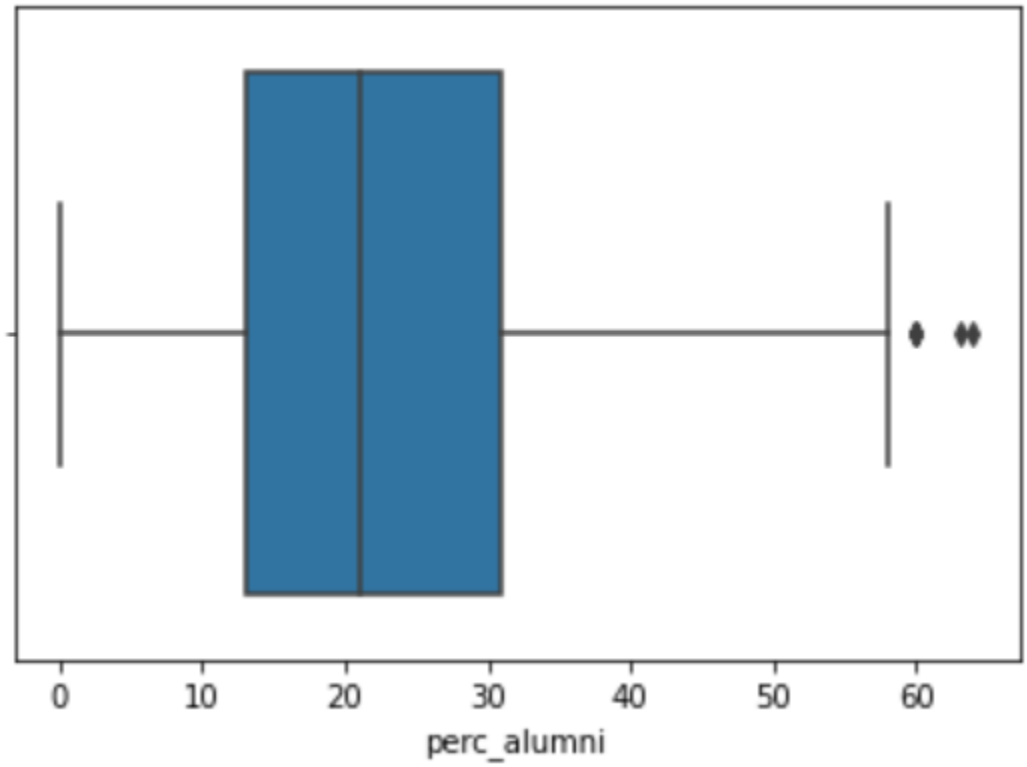


Fig 84 expend count

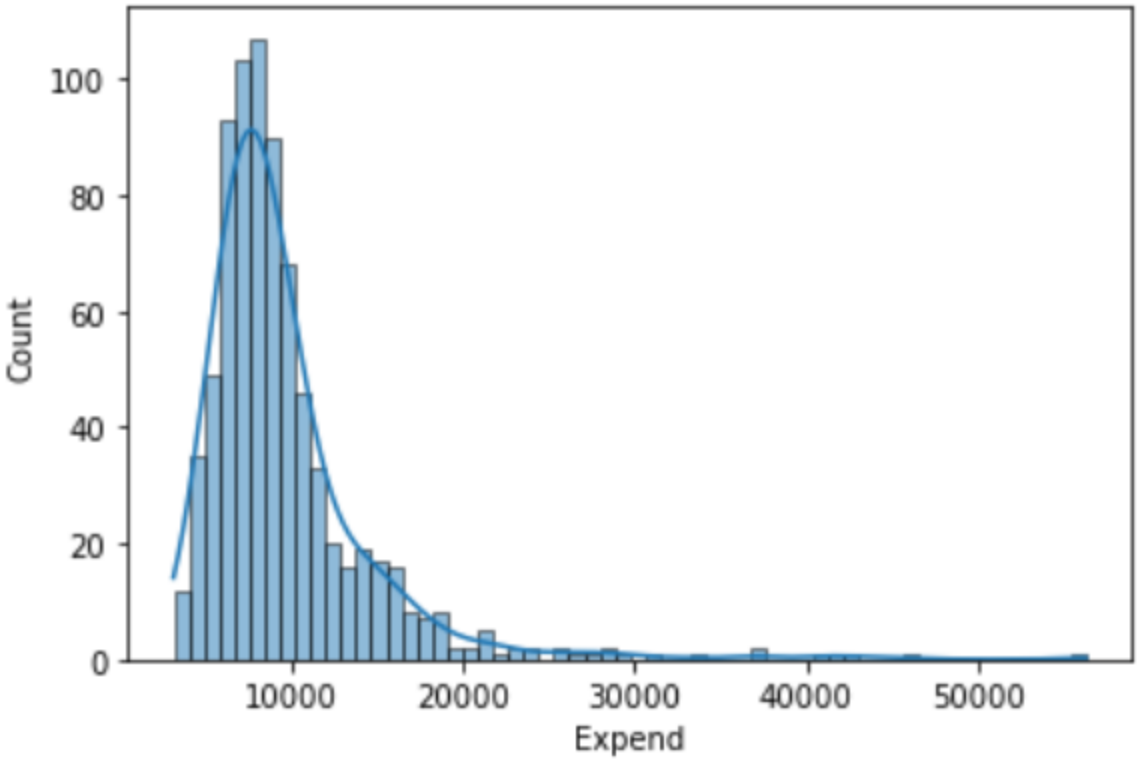
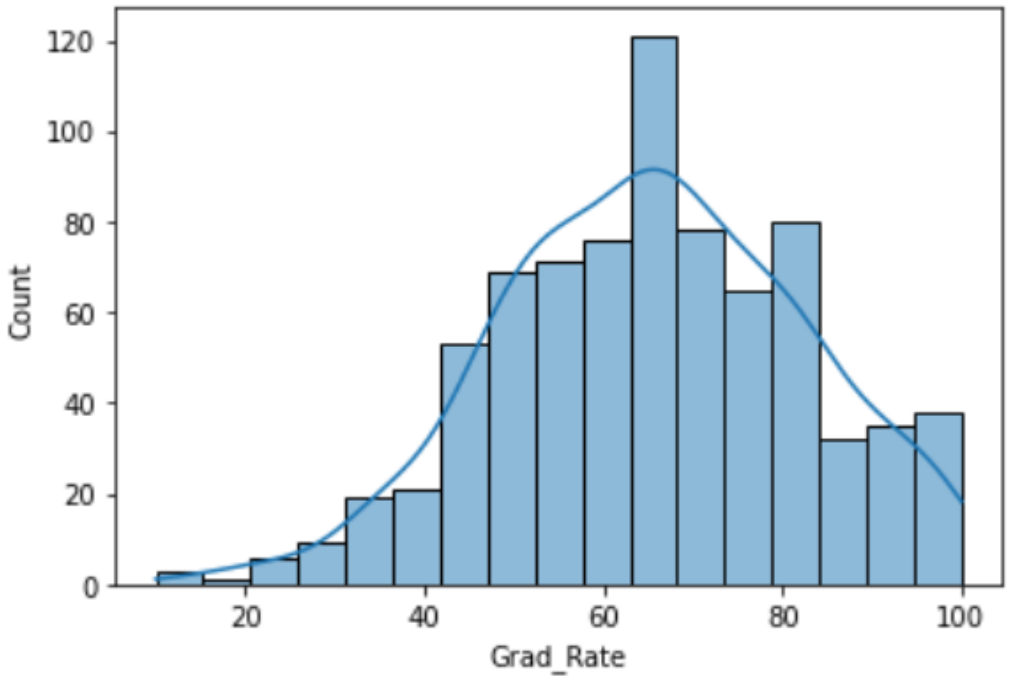


Fig 85 Grad_Rate

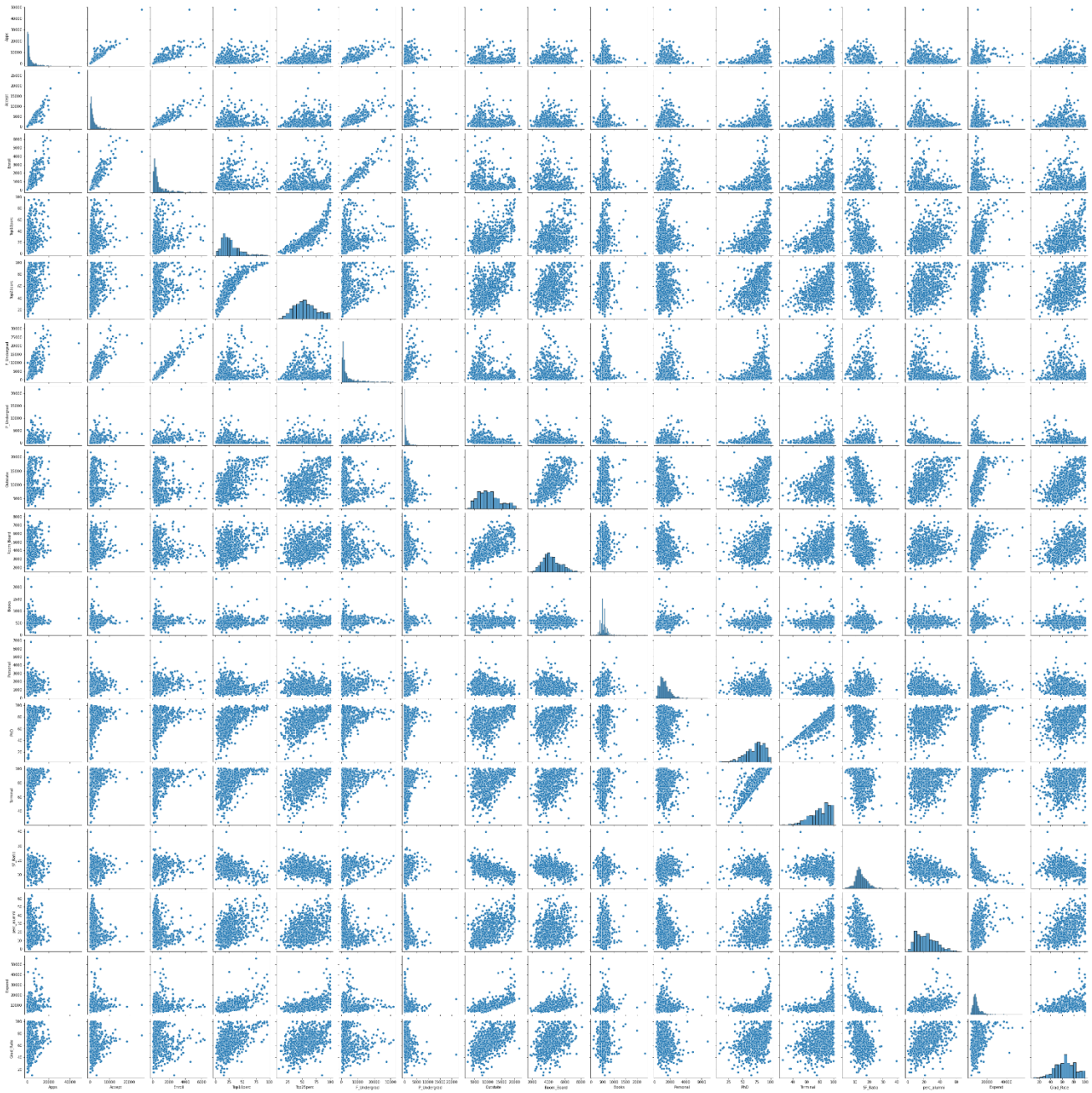


Multivariate & Bivariate Analysis

The below heat map gives the correlation between the 17 variables. The below plot gives the pairwise scatterplot between the variables. Pair plot of the 17 variables. The significant plots out of these have been presented below.

Pairplot

Fig 86 pairplot



• From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.94) between the number of applications received by a university and the number of applications accepted.

Strong correlations found

From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.85) between the number of applications received by a university and the number of students enrolled.

	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Ter
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.392980	0.3
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.357938	0.3
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.333485	0.3
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.533707	0.4
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.548058	0.5
F_Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.320183	0.3
P_Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.150495	0.1
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.386905	0.4
Room_Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.332429	0.3
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026286	0.0
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.008701	-0.0
PhD	0.392980	0.357938	0.333485	0.533707	0.548058	0.320183	0.150495	0.386905	0.332429	0.026286	-0.008701	1.000000	0.8
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.850226	1.0
SF_Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.132672	-0.1
perc_alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.250767	0.2
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.435099	0.4
Grad_Rate	0.146775	0.065414	-0.021388	0.502212	0.484388	-0.077034	-0.256350	0.575820	0.425769	-0.000159	-0.266018	0.322921	0.3

Fig 87 correlation

Heat map

• From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.91) between the number of applications accepted by a university and the number of students enrolled.

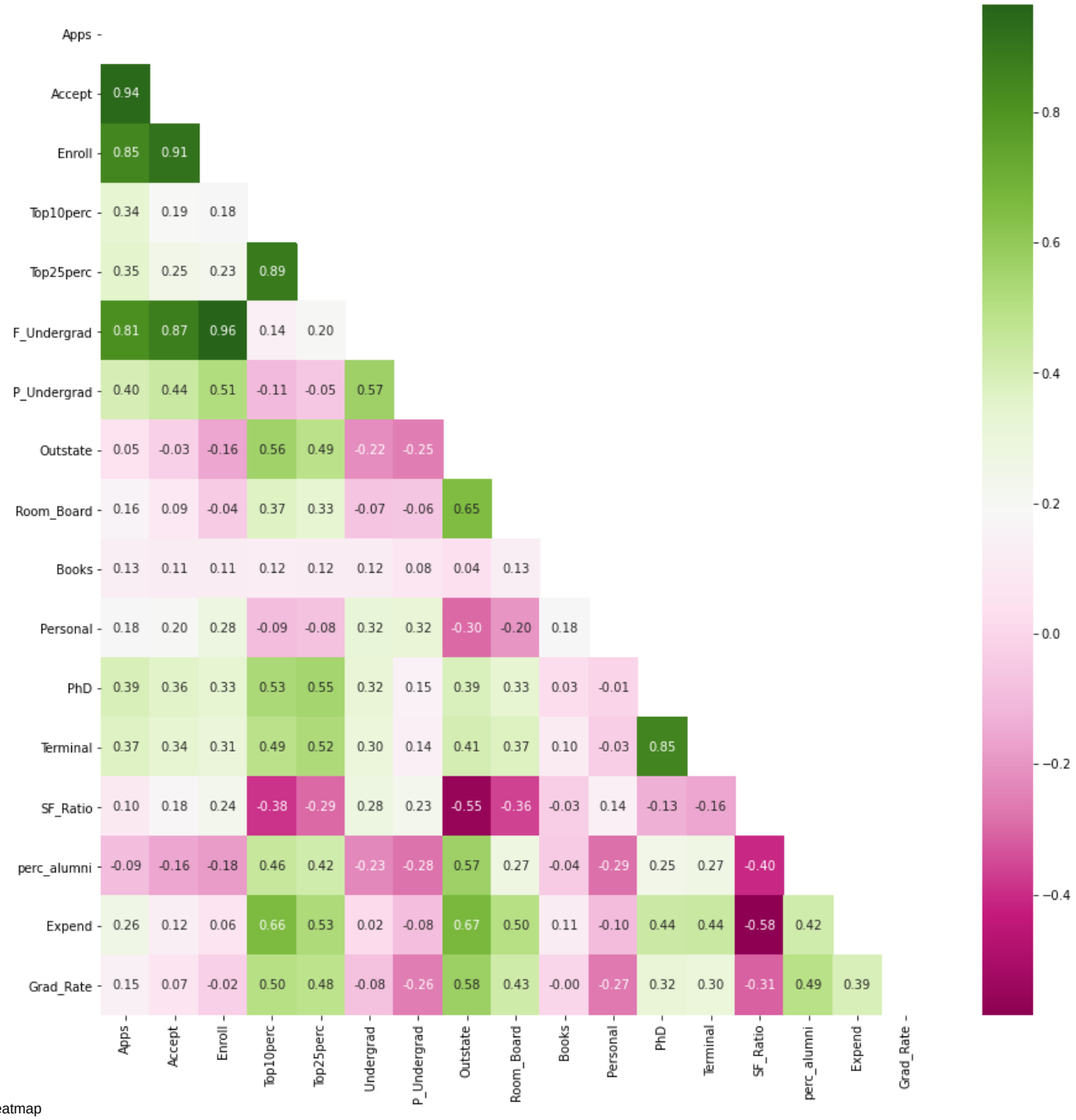


Fig 88 heatmap

• From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.81) between the number of full time undergraduate students and the number of applications received by a university.

• From the heat map and the above scatter plot, we find a very strong positive correlation (**correlation coefficient = 0.87**) between the number of full-time undergraduate students and the number of applications accepted by a university.

• From the heat map and the above scatter plot, we find a very strong positive correlation (**correlation coefficient = 0.96**) between the number of full-time undergraduate students and the number of students enrolled in a university.

• From the heat map and the above scatter plot, we find a very strong positive correlation (**correlation coefficient = 0.89**) between the percentage of new students from top 10% of Higher Secondary class and percentage of new students from top 25% of Higher Secondary class.

• From the heat map and the above scatter plot, we find a moderately positive correlation (**correlation coefficient = 0.66**) between the percentage of new students from top 10% of Higher Secondary class and the instructional expenditure per student.

• From the heat map and the above scatter plot, we find a moderately positive correlation (**correlation coefficient = 0.65**) between the cost of room and board and the number of students for whom the particular college or university is Out-of-state tuition

• From the heat map and the above scatter plot, we find a moderately positive correlation (**correlation coefficient = 0.67**) between the number of students for whom the particular college or university is Out-of-state tuition and the instructional expenditure per student.

• From the heat map and the above scatter plot, we find a very strong positive correlation (**correlation coefficient = 0.85**) between the percentage of faculties with Ph.D.'s and percentage of faculties with terminal degree.

• We can calculate the acceptance ratio or acceptance rate of a university/college by dividing the number of accepted applications divided by the total number of applications received.

• There are 6 universities/colleges with a 100% acceptance rate and **Princeton** University has the least acceptance rate, of **15.45%,** as seen from the below tables.

High acceptance ratio

These colleges must be easy to get into

	Names	Acceptance Ratio
729	Wayne State College	100.00
192	Emporia State University	100.00
355	Mayville State University	100.00
535	Southwest Baptist University	100.00
697	University of Wisconsin-Superior	100.00
368	MidAmerica Nazarene College	100.00
25	Arkansas Tech University	99.71
538	Southwestern Adventist College	99.07
452	Pikeville College	99.01
391	Mount Marty College	98.92

Fig 89 high accept ratio

Low acceptance

Elite institutes, for sure, have tougher aptitude tests for admission

	Names	Acceptance Ratio
763	Williams College	29.74
144	Columbia University	28.57
174	Duke University	28.23
158	Dartmouth College	26.47
221	Georgetown University	25.92
70	Brown University	25.73
16	Amherst College	23.06
775	Yale University	22.91
250	Harvard University	15.61
459	Princeton University	15.45

Fig 90 low accept ratio

Pairplot combinations

Fig 91 pairplot combine apps accept

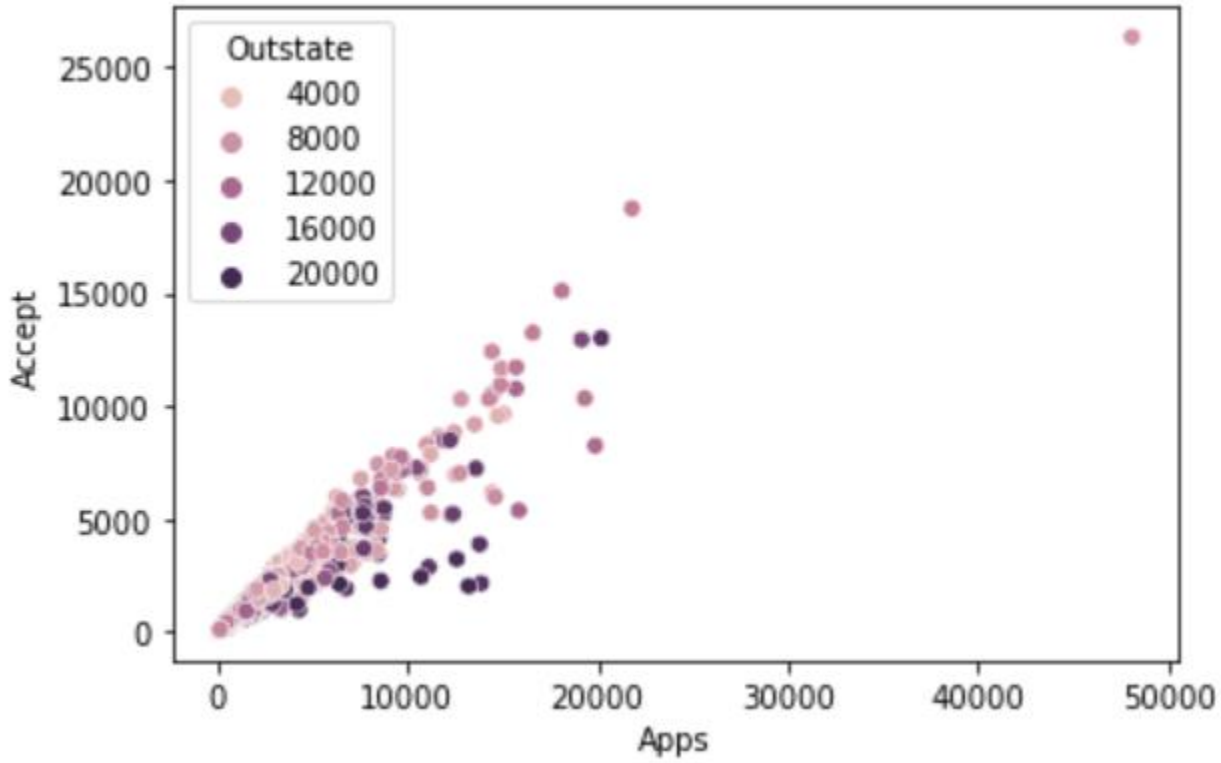


Fig 92 Apps enroll scatter

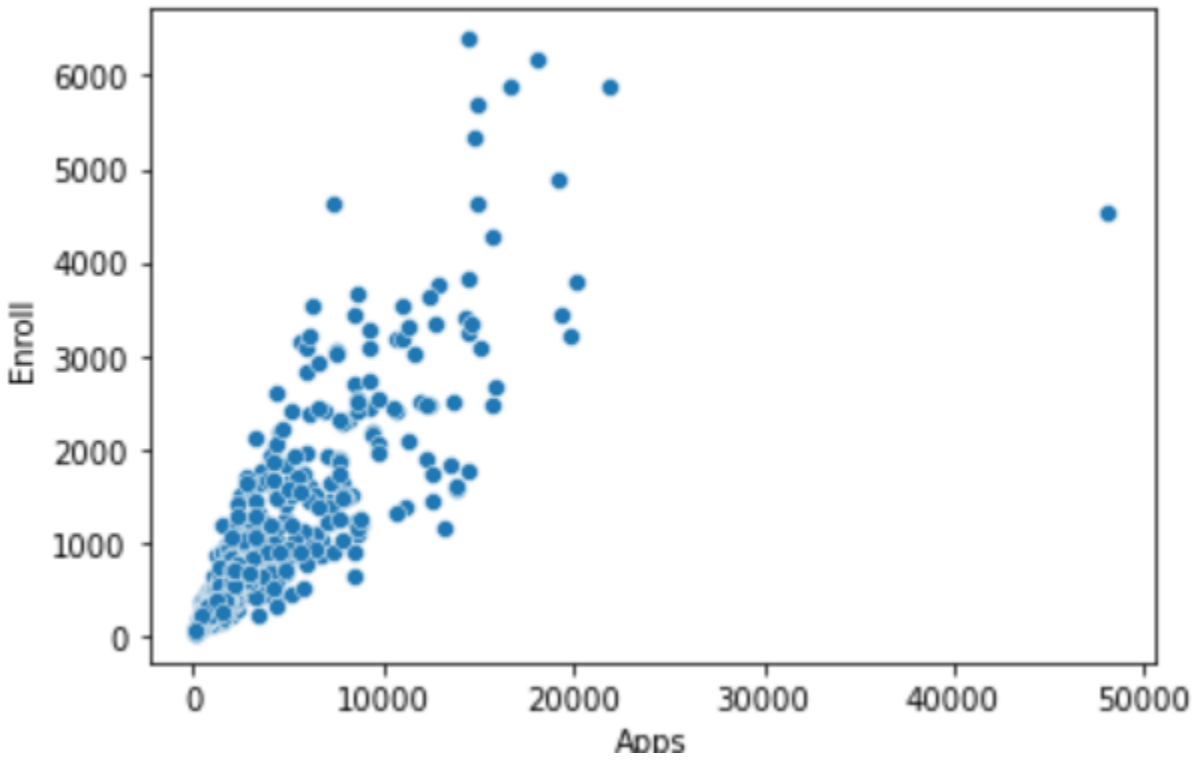


Fig 93 accept enroll scatter

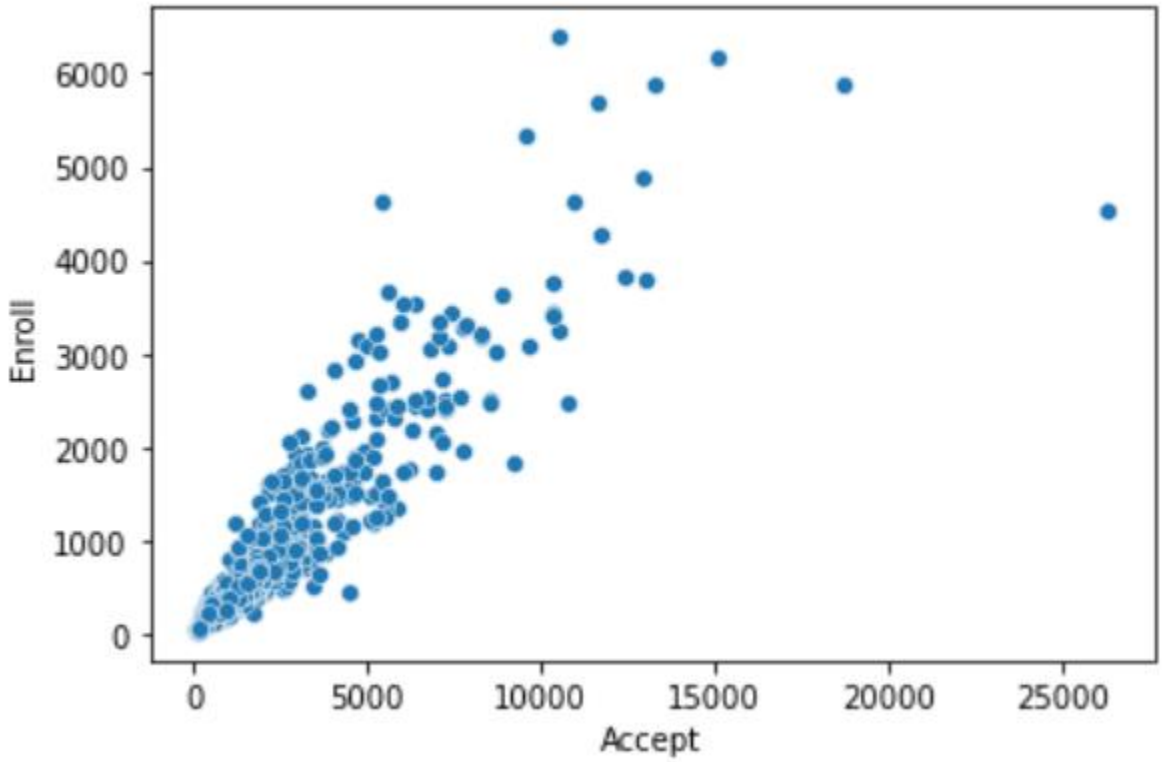


Fig 94 F-under apps scatter

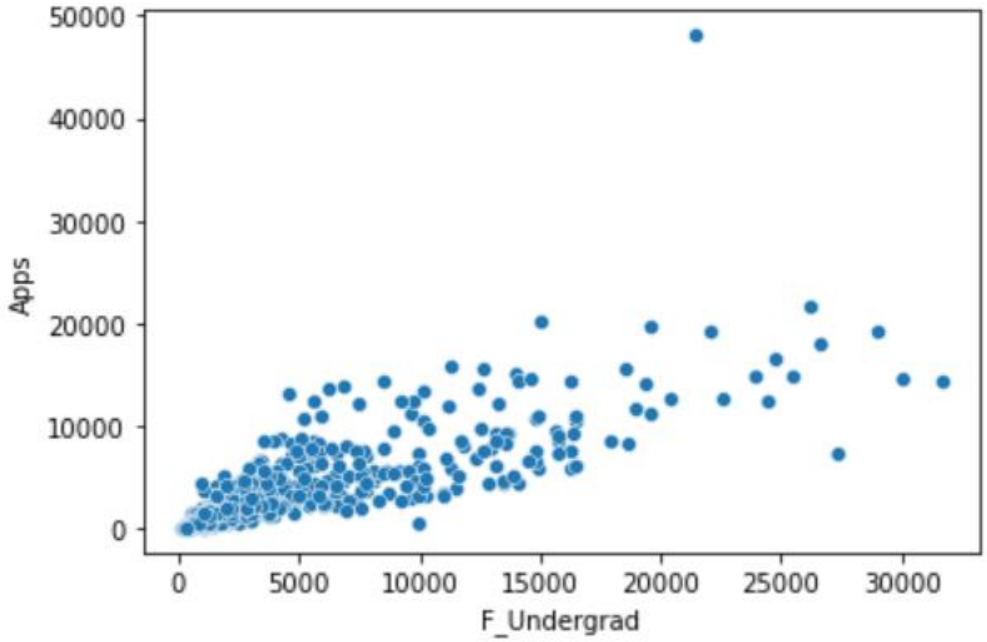


Fig 95 F_under accept

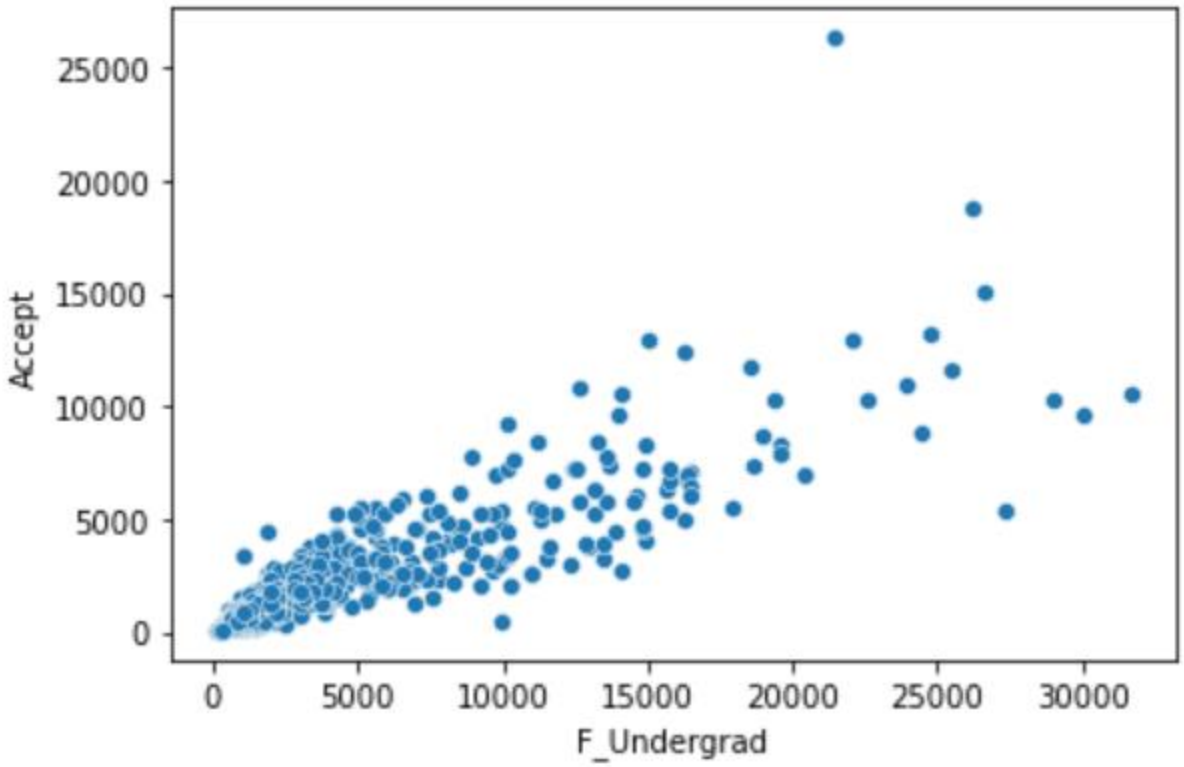


Fig 96 F under enroll

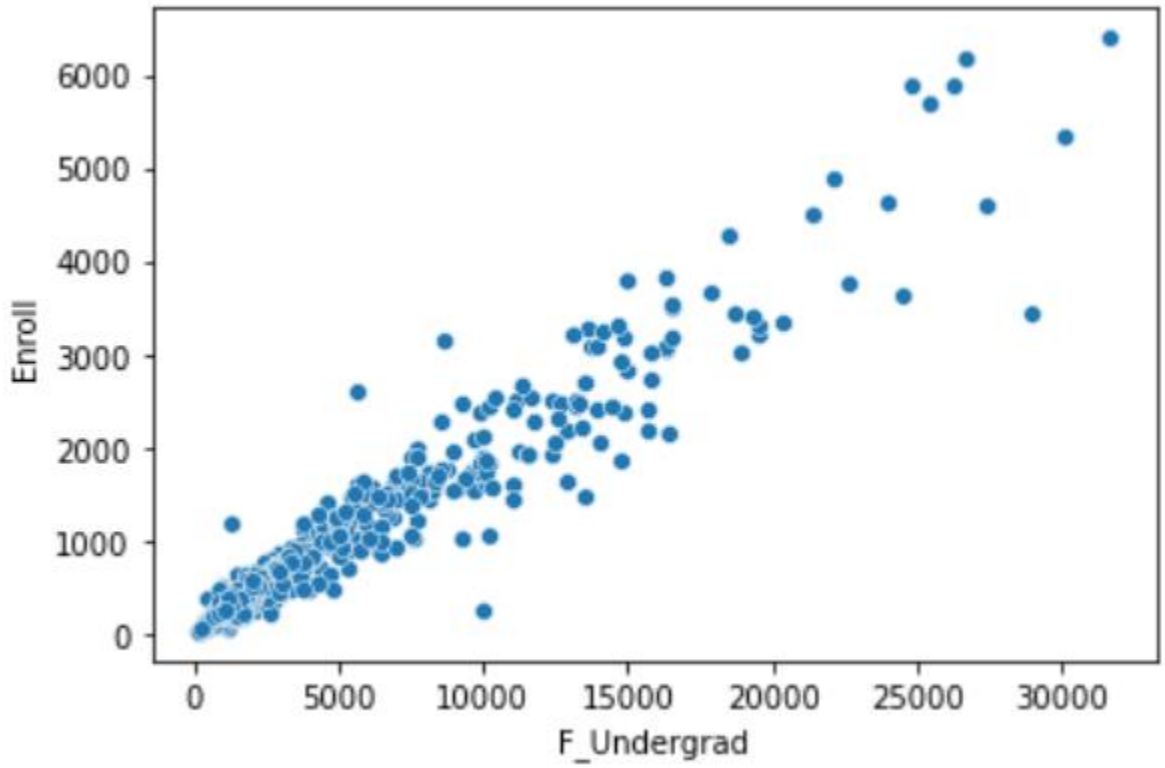


Fig 97 T25 T10 scatter

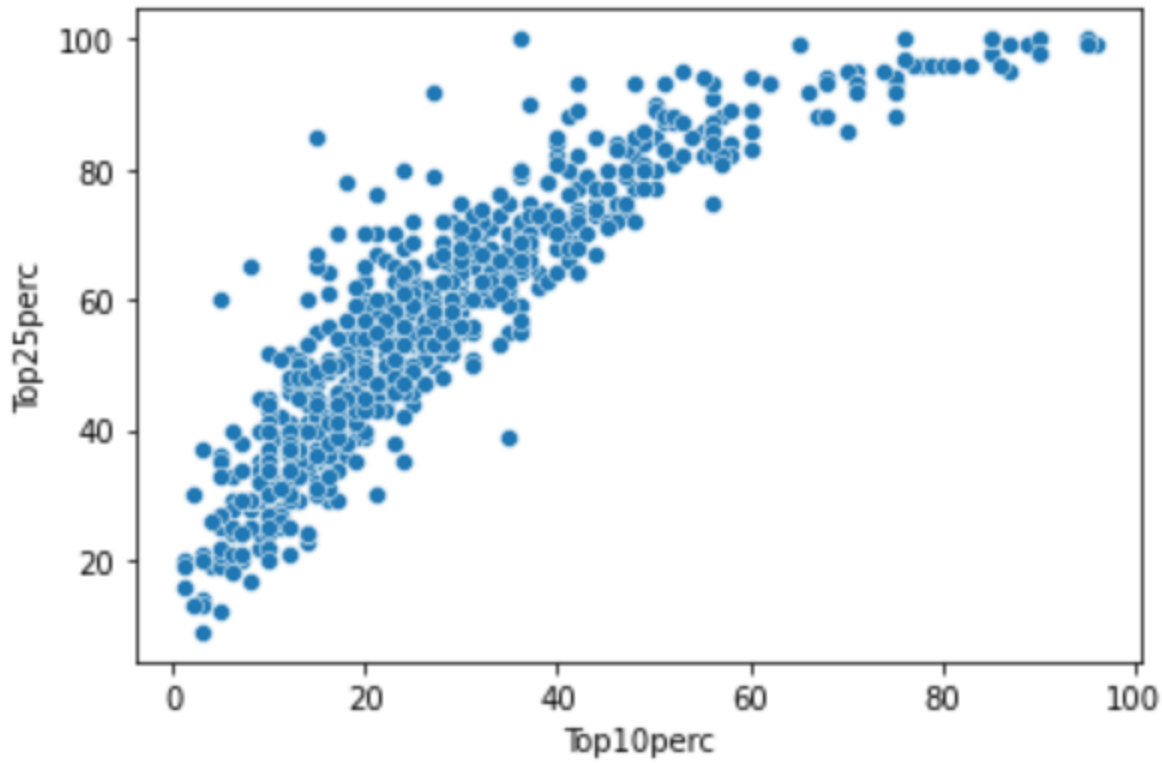


Fig 98 T10 expend

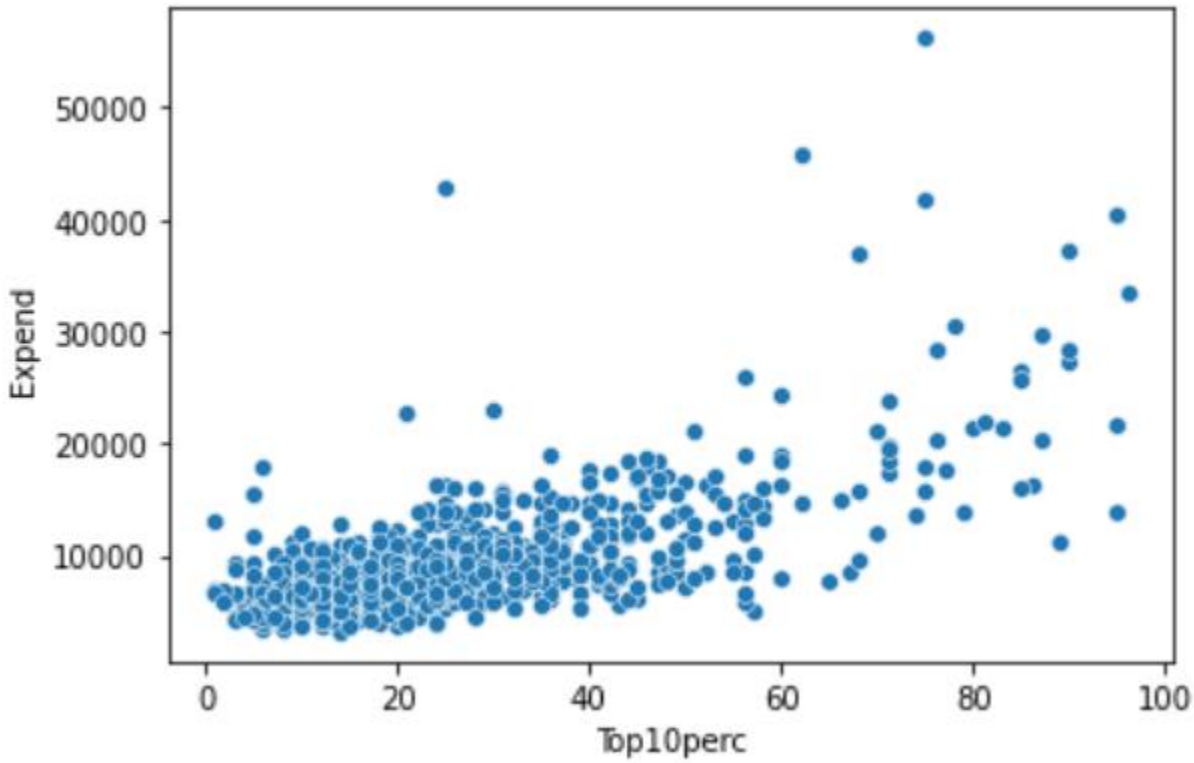


Fig 99 out room scatter

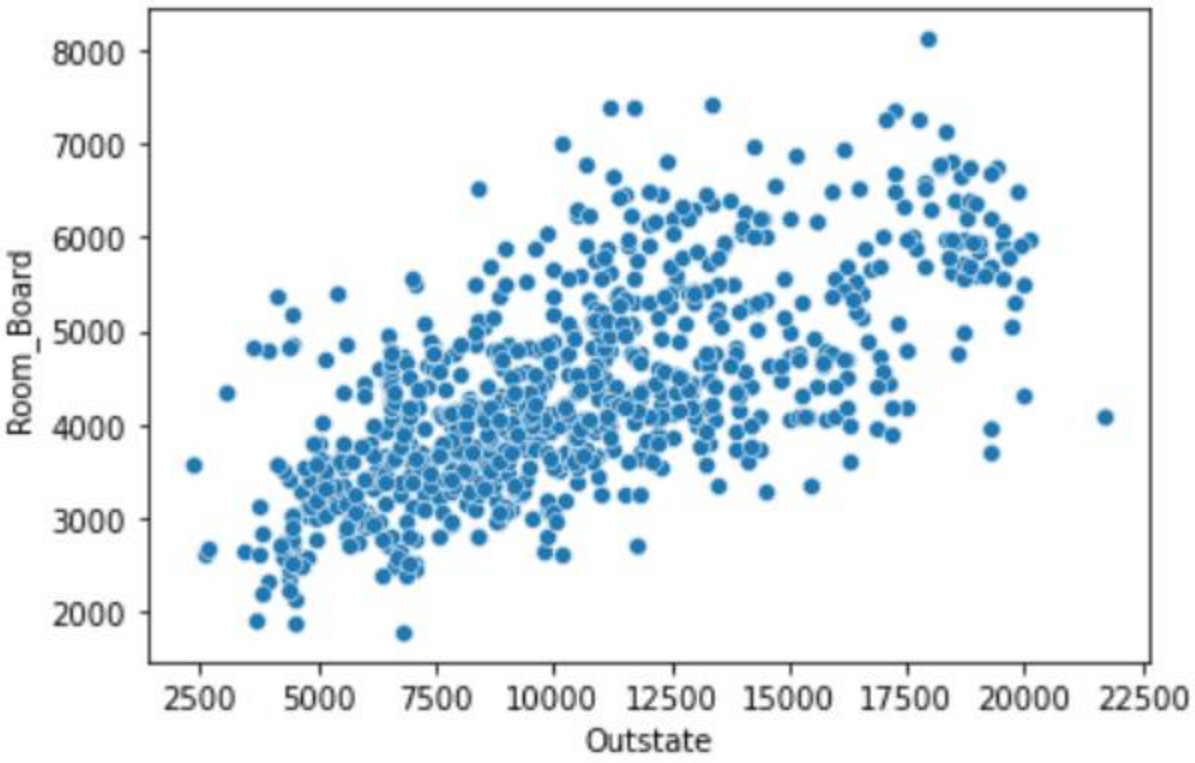


Fig 100 Outstate expend

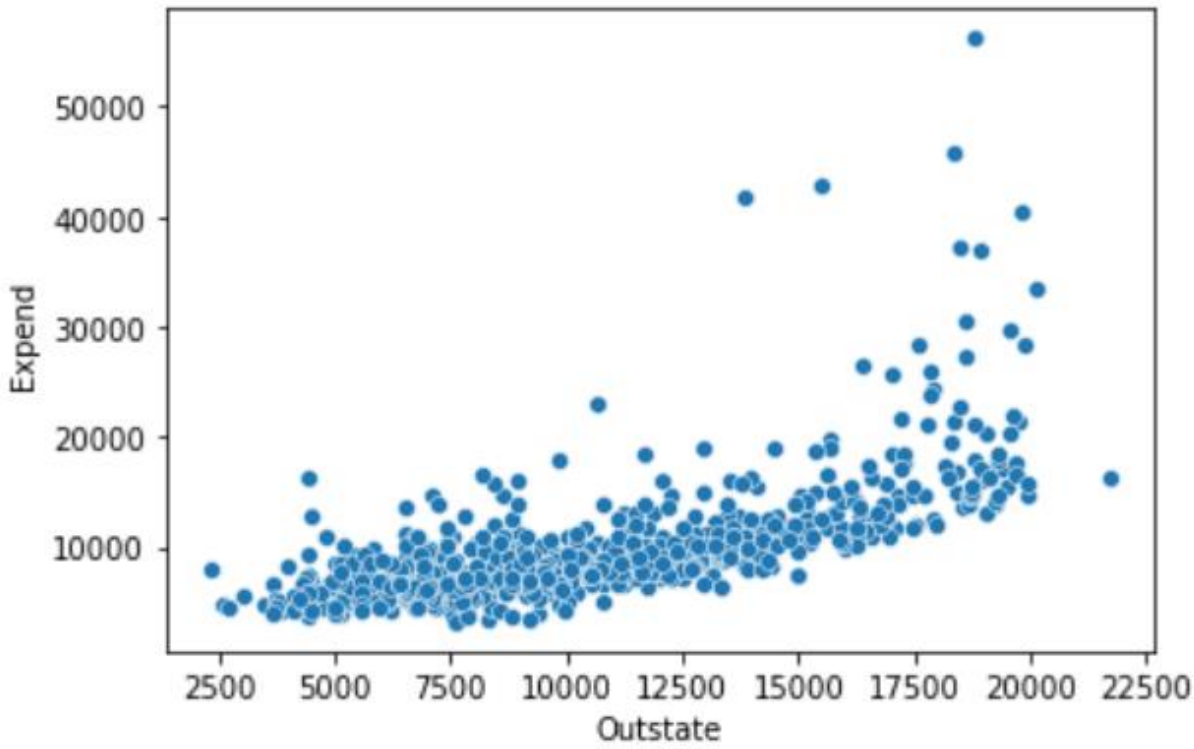
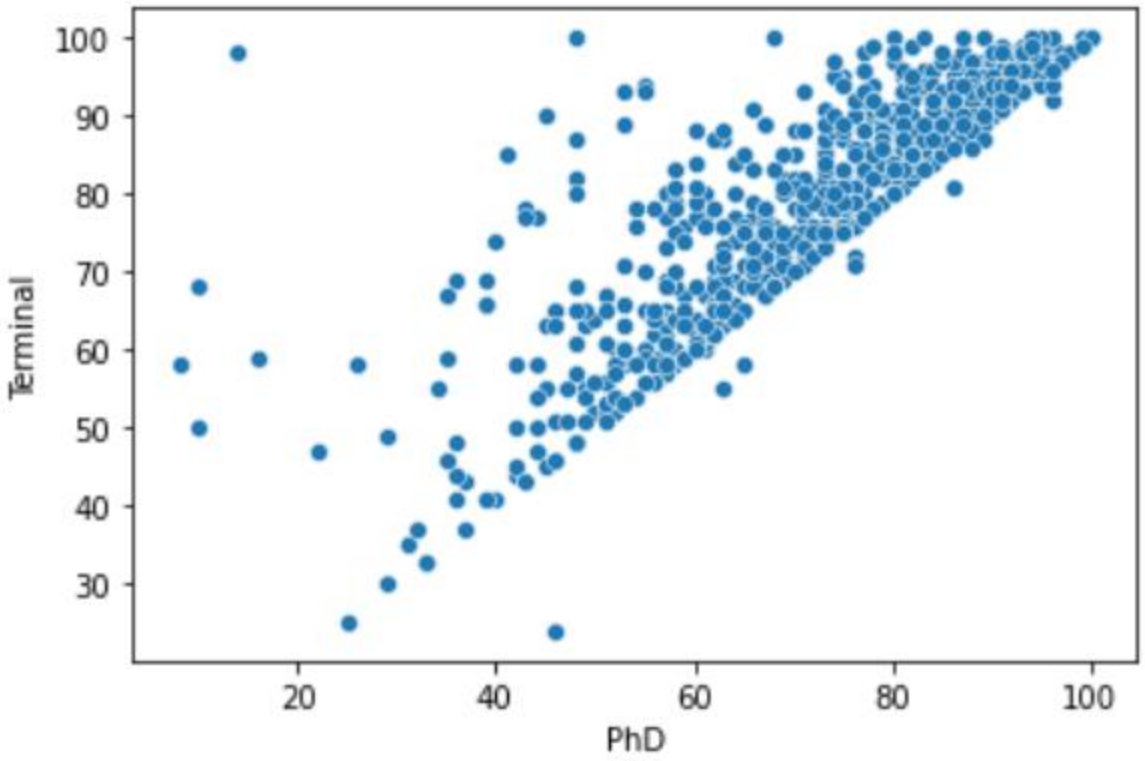


Fig 101 Phd terminal scatter



The variances of the numeric variables

Apps	1.497846e+07
Accept	6.007960e+06
Enroll	8.633684e+05
Top10perc	3.111825e+02
Top25perc	3.922292e+02
F_Undergrad	2.352658e+07
P_Undergrad	2.317799e+06
Outstate	1.618466e+07
Room_Board	1.202743e+06
Books	2.725978e+04
Personal	4.584258e+05
PhD	2.654282e+02
Terminal	2.167478e+02
SF_Ratio	1.566853e+01
perc_alumni	1.535567e+02
Expend	2.726687e+07
Grad_Rate	2.915125e+02

Fig 102 variances of var

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

Solution: Yes. Scaling is necessary for PCA in this case. Since, the sample variances of the original variables show differences by large order of magnitude, variables need to be normalized (as seen in the below table). Scaling ensures that the attribute means are all 0 and variances 1. A snapshot of the scaled data is seen in the below table.

Dropping Names column in order to extract only numerical variables

	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Terminal	SF_Ratio	perc_alumni
0	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12
1	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16
2	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30
3	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37
4	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2
5	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11
6	353	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26
7	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37
8	1038	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23
9	582	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15

	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Terminal
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.161177	-0.115729
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.679375	-3.378176
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.205308	-0.931341
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.190052	1.175657
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.207340	-0.523535
5	-0.624307	-0.628611	-0.669812	0.592287	0.313426	-0.623421	-0.535212	0.760947	-0.932970	-0.299280	-0.983753	-0.345435	-0.455567
6	-0.684808	-0.685356	-0.729043	-0.598931	-0.545505	-0.677472	-0.410988	0.708713	1.243144	-0.299280	0.235515	1.067213	0.903786
7	-0.285088	-0.121984	-0.313353	0.535563	0.616579	-0.434450	-0.541127	0.852479	0.427443	-0.602312	-0.725120	1.005793	1.379560
8	-0.507700	-0.481644	-0.595505	0.138490	0.363952	-0.562562	-0.361036	1.282036	0.038754	-1.511408	-1.242385	0.391599	0.292077
9	-0.625600	-0.620854	-0.654735	-0.372032	-0.596031	-0.598459	-0.510893	0.006798	-0.891911	0.670422	0.678885	-2.003761	-2.630532

SF_Ratio	perc_alumni	Expend	Grad_Rate
1.013776	-0.867574	-0.501910	-0.316192
-0.477704	-0.544572	0.166110	-0.550621
-0.300749	0.585935	-0.177290	-0.667836
-1.615274	1.151188	1.792851	-0.374800
-0.553542	-1.675079	0.241803	-2.953517
-1.185526	-0.948325	0.012806	-0.609228
-0.654660	0.262933	-0.153145	-0.140371
-0.098515	1.151188	0.350074	0.445701
-0.705218	0.020681	0.380160	0.855952
-0.654660	-0.625323	-0.128233	-0.785050

Fig 103 dropping name datasets

Standard deviation of the scaled data = 1

```
Apps            1.000644
Accept          1.000644
Enroll          1.000644
Top10perc       1.000644
Top25perc       1.000644
F_Undergrad     1.000644
P_Undergrad     1.000644
Outstate        1.000644
Room_Board      1.000644
Books           1.000644
Personal        1.000644
PhD             1.000644
Terminal        1.000644
SF_Ratio        1.000644
perc_alumni     1.000644
Expend          1.000644
Grad_Rate       1.000644
dtype: float64
```

The mean of the scaled data = 0

```
Apps            6.355797e-17
Accept          6.774575e-17
Enroll         -5.249269e-17
Top10perc      -2.753232e-17
Top25perc      -1.546739e-16
F_Undergrad    -1.661405e-16
P_Undergrad    -3.029180e-17
Outstate        6.515595e-17
Room_Board      3.570717e-16
Books          -2.192583e-16
Personal        4.765243e-17
PhD             1.767855e-16
Terminal       -4.481615e-16
SF_Ratio       -2.057556e-17
perc_alumni    -6.022638e-17
Expend          1.213101e-16
Grad_Rate      -3.349244e-16
```

Fig 105 std dev scaled

2.3 Comment on the comparison between the covariance and the correlation matrices from this data.

For a scaled data, the covariance matrix and the correlation matrix are the same. In this case too, we find that the covariance matrix reduces to the correlation matrix after scaling. A snapshot of the two matrices are given below.

Correlation matrix

Fig 106 correlation matrix

	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	\
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	
F_Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	
P_Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	
Room_Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	
PhD	0.392980	0.357938	0.333485	0.533707	0.548058	0.320183	
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	
SF_Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	
perc_alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	
Grad_Rate	0.146775	0.065414	-0.021388	0.502212	0.484388	-0.077034	

```
Covariance Matrix
%s [[ 1.00128866e+00  9.44666359e-01  8.47913316e-01  3.39270321e-01
   3.52093041e-01  8.15540181e-01  3.98777500e-01  5.02236717e-02
   1.65151509e-01  1.32729421e-01  1.78961168e-01  3.93486554e-01
   3.69967622e-01  9.57562670e-02 -9.03421565e-02  2.59926503e-01
   1.46963658e-01]
 [ 9.44666359e-01  1.00128866e+00  9.12811453e-01  1.92694926e-01
   2.47794654e-01  8.75349854e-01  4.41839380e-01 -2.57877355e-02
   9.10157685e-02  1.13671647e-01  2.01247673e-01  3.58398858e-01
   3.38018401e-01  1.76456113e-01 -1.60196038e-01  1.24877730e-01
   6.54987779e-02]
 [ 8.47913316e-01  9.12811453e-01  1.00128866e+00  1.81527154e-01
   2.27037304e-01  9.65882744e-01  5.13729774e-01 -1.55677702e-01
  -4.02835287e-02  1.12856137e-01  2.81291483e-01  3.33915060e-01
   3.08671332e-01  2.37577072e-01 -1.81027112e-01  6.42519204e-02
  -2.14155645e-02]
 [ 3.39270321e-01  1.92694926e-01  1.81527154e-01  1.00128866e+00
   8.93144451e-01  1.41470801e-01 -1.05492050e-01  5.63055197e-01
   3.71959090e-01  1.19011599e-01 -9.34366503e-02  5.34394970e-01
   4.91767929e-01 -3.85370484e-01  4.56072227e-01  6.61765100e-01
   5.02859166e-01]
```

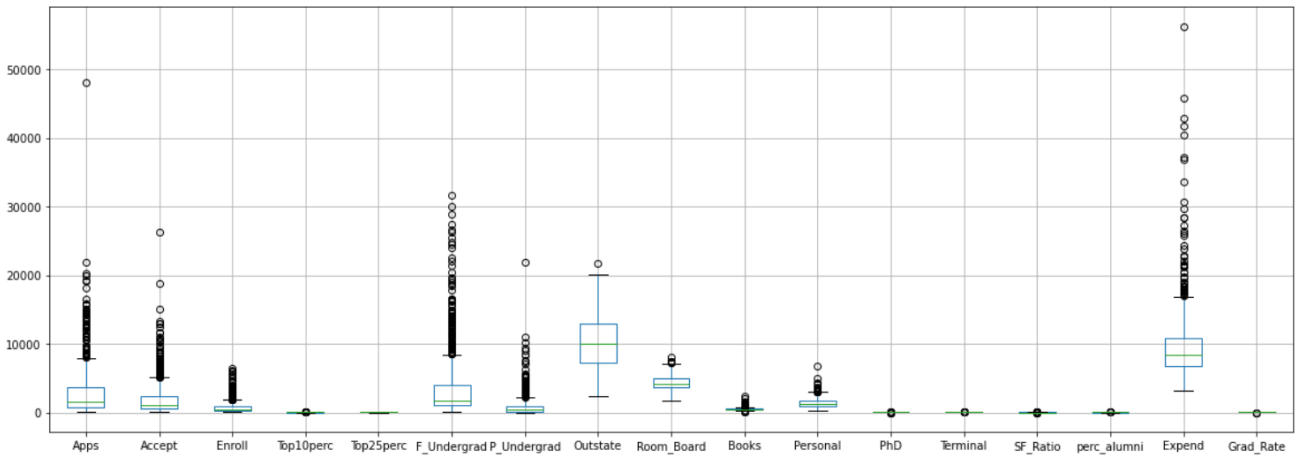
Fig 107 covariance matrix

2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

Boxplot before scaling of data Boxplot after scaling of data. The above two plots show the boxplots before and after scaling. From, the 2 boxplots, we see the presence of outliers in the data. Normalisation of data does not remove the outliers, but only the range of the data changes. More boxes are now visible.

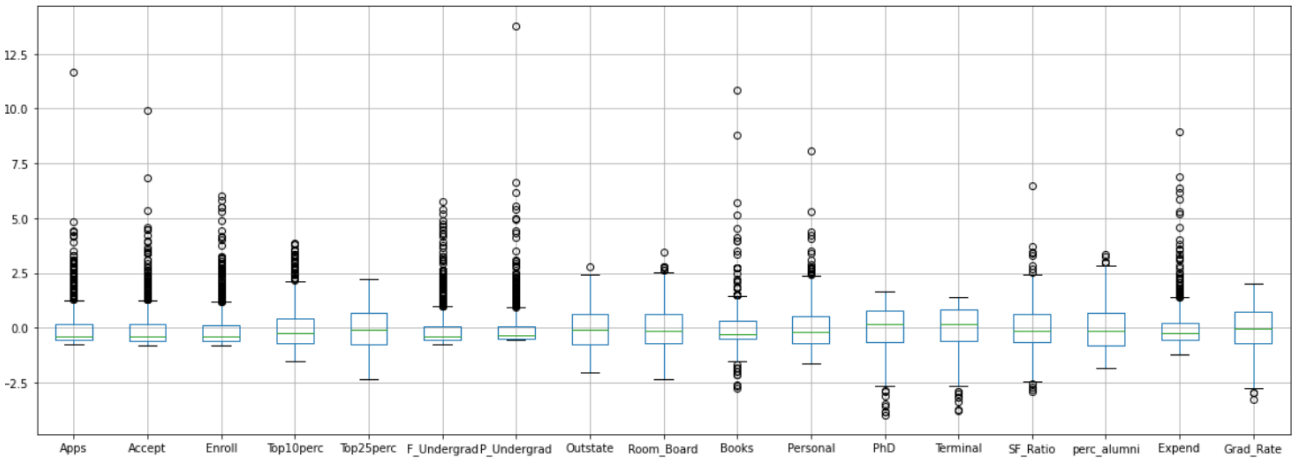
Before

Fig 108 before outlier box



After

Fig 109 after outlier box



2.5 Perform PCA and export the data of the Principal Component scores into a data frame. .

Preliminary tests…

Step 1.

Bartletts Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

- H0: All variables in the data are uncorrelated
- Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is atleast one pair of variables in the data which are correlated, hence PCA is recommended.

P_value:-

```
0.0
```

KMO Test

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

KMO model:-

```
0.8131776428730154
```

Step 1

Create the covariance Matrix

Fig 110 create cov mat

```
Covariance Matrix
%s [[ 1.00128866e+00  9.44666359e-01  8.47913316e-01  3.39270321e-01
   3.52093041e-01  8.15540181e-01  3.98777500e-01  5.02236717e-02
   1.65151509e-01  1.32729421e-01  1.78961168e-01  3.93486554e-01
   3.69967622e-01  9.57562670e-02 -9.03421565e-02  2.59926503e-01
   1.46963658e-01]
 [ 9.44666359e-01  1.00128866e+00  9.12811453e-01  1.92694926e-01
   2.47794654e-01  8.75349854e-01  4.41839380e-01 -2.57877355e-02
   9.10157685e-02  1.13671647e-01  2.01247673e-01  3.58398858e-01
   3.38018401e-01  1.76456113e-01 -1.60196038e-01  1.24877730e-01
   6.54987779e-02]
 [ 8.47913316e-01  9.12811453e-01  1.00128866e+00  1.81527154e-01
   2.27037304e-01  9.65882744e-01  5.13729774e-01 -1.55677702e-01
  -4.02835287e-02  1.12856137e-01  2.81291483e-01  3.33915060e-01
   3.08671332e-01  2.37577072e-01 -1.81027112e-01  6.42519204e-02
  -2.14155645e-02]
```

2.6 Extract the eigenvalues, and eigenvectors.

Step 2

Get eigen values and eigen vectors

Fig 111 eigen vals vecs

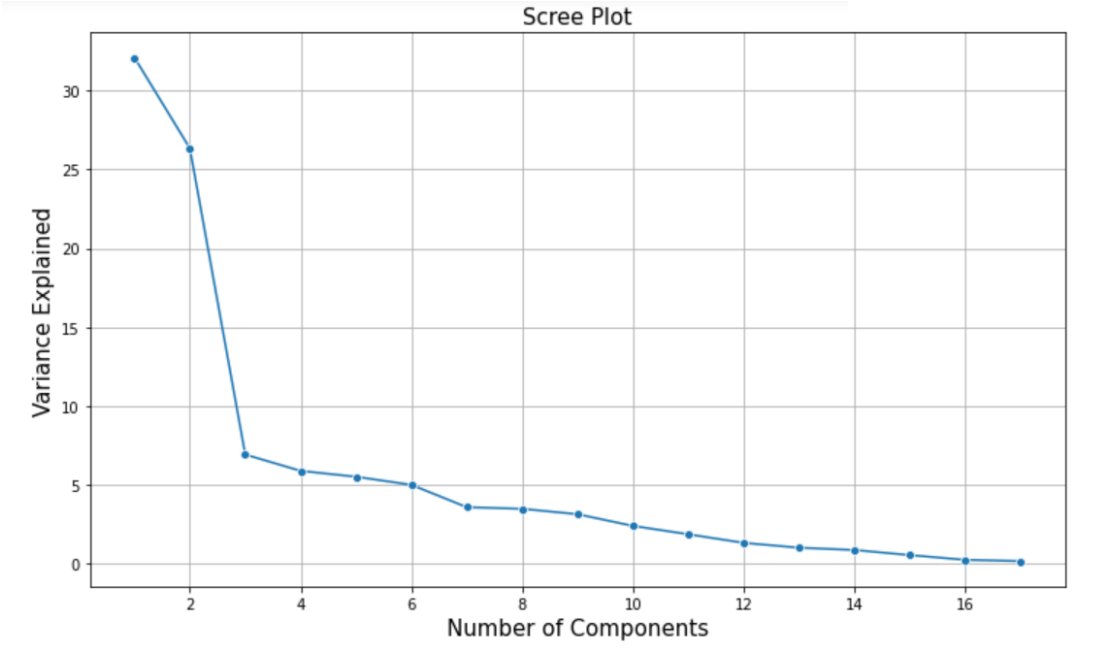
```
Eigen Values
%s [5.46384062 4.4841809  1.17453449 0.99793705 0.93465879 0.84734458
0.60586408 0.58783511 0.53014189 0.40354672 0.02300969 0.03671503
0.3131535  0.08806661 0.14299858 0.16777512 0.22030447]


Eigen Vectors
 %s [[-2.47532537e-01  3.32426877e-01  5.96777039e-02 -2.85096690e-01
   1.70019622e-04  1.22811932e-02  3.06596836e-02  1.03577277e-01
   8.93297613e-02 -5.07359465e-02  3.59723152e-01 -4.59049537e-01
   4.31741045e-02 -1.32822763e-01 -6.60495278e-02 -5.97045093e-01
  -2.33859080e-02]
 [-2.06299756e-01  3.72875058e-01  9.77593828e-02 -2.71787742e-01
   5.06165527e-02 -1.12771266e-02  2.81197218e-03  5.55282328e-02
   1.76710689e-01 -4.05496164e-02 -5.44054193e-01  5.17547918e-01
  -5.86528967e-02  1.44994783e-01 -2.62518280e-02 -2.92920428e-01
   1.46702734e-01]
```

Step 3

View Scree Plot to identify the number of components to be built

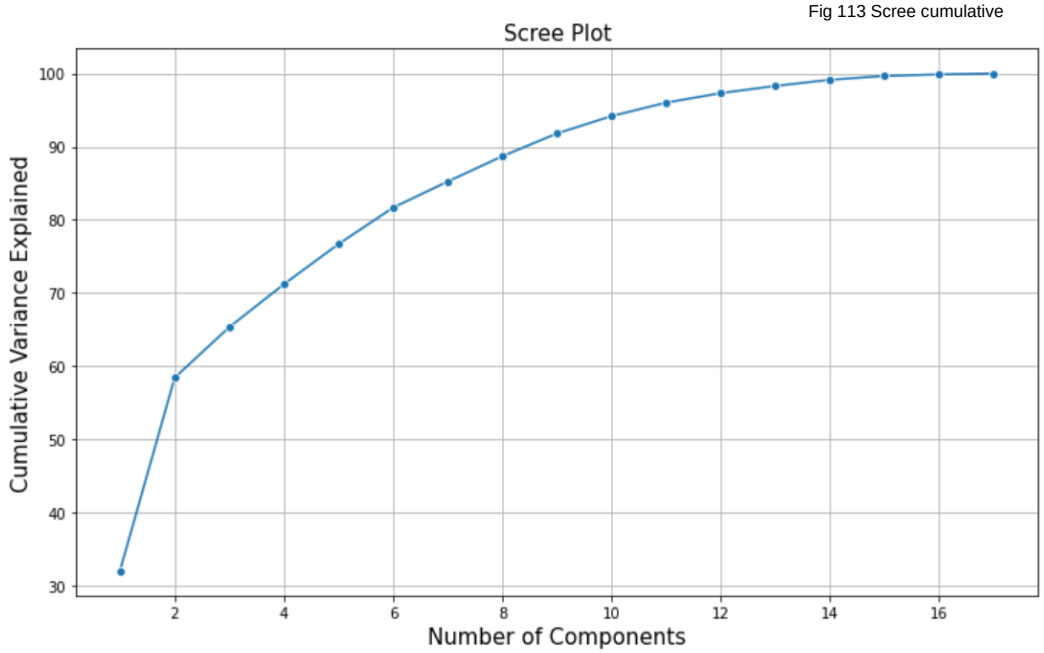
```
Cumulative Variance Explained [ 32.0988744   58.44246116  65.34259565  71.20525863  76.69617559
  81.67414057  85.23346016  88.68686349  91.80133169  94.17208083
  96.01178946  97.30603027  98.29167258  99.13175814  99.64913029
  99.86482309 100.        ]
```

Fig 112 Scree plot



Scree plot for cumulative variance explained

Fig 113 Scree cumulative



Step 4

Apply PCA for the number of decided components to get the loadings and component output

```
array([[-1.59178367e+00, -2.19726733e+00, -1.43106151e+00, ...,
        -7.35654959e-01,  7.92729370e+00, -4.57851961e-01],
       [ 7.62095222e-01, -5.82570548e-01, -1.09617695e+00, ...,
        -7.85324092e-02, -2.04892376e+00,  3.62098269e-01],
       [-1.16125321e-01,  2.31716748e+00, -4.34091528e-01, ...,
         1.60923954e-03,  2.08175826e+00, -1.33525267e+00],
       [-9.49497305e-01,  3.59195182e+00,  7.02472205e-01, ...,
         7.26782076e-02,  8.43396721e-01, -1.78424190e-01],
       [-7.24679783e-01,  9.84844680e-01, -3.79119636e-01, ...,
        -5.17643471e-01, -9.60284109e-01, -1.13325642e+00],
       [-3.13274583e-01, -1.26557398e-01, -9.47274775e-01, ...,
         4.72161888e-01, -2.06846796e+00,  8.32020244e-01]])
```

Fig 114 apply PCA

```
array([5.46384062, 4.4841809 , 1.17453449, 0.99793705, 0.93465879,
       0.84734458])
```

Loading of each feature on the components

```
array([[ 2.47532537e-01,  2.06299756e-01,  1.75138842e-01,
         3.53990557e-01,  3.43702467e-01,  1.53527590e-01,
         2.57859287e-02,  2.94965994e-01,  2.48896885e-01,
         6.42827785e-02, -4.25981738e-02,  3.19579232e-01,
         3.16776477e-01, -1.77164491e-01,  2.05418014e-01,
         3.18605544e-01,  2.55626868e-01],
```

Fig 115 loading feature

The below gives the cumulative variance

PCA explained variance ratio

```
array([0.32098874, 0.26343587, 0.06900134, 0.05862663, 0.05490917,
       0.04977965])
```

6 Principal components were generated with a cumulative variance of 81%. We went for 6 PCs since as a general rule, 80-20 is taken.

```
array([ 0.24753254,  0.20629976,  0.17513884,  0.35399056,  0.34370247,
        0.15352759,  0.02578593,  0.29496599,  0.24889688,  0.06428278,
       -0.04259817,  0.31957923,  0.31677648, -0.17716449,  0.20541801,
        0.31860554,  0.25562687])
```

Fig 116 principal components

PCA-loaded dataset

	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Terminal
0	0.247533	0.206300	0.175139	0.353991	0.343702	0.153528	0.025786	0.294966	0.248897	0.064283	-0.042598	0.319579	0.316776
1	0.332427	0.372875	0.404251	-0.081508	-0.043914	0.418089	0.315119	-0.248748	-0.136928	0.056607	0.219634	0.058854	0.047001
2	-0.059678	-0.097759	-0.081358	0.034279	-0.025509	-0.060563	0.137967	0.048149	0.152144	0.679694	0.495300	-0.131903	-0.070972
3	0.285097	0.271788	0.163457	-0.055180	-0.115855	0.101292	-0.158951	0.136572	0.191576	0.071380	-0.249033	-0.529195	-0.518145
4	0.000170	0.050617	-0.058798	-0.394019	-0.423913	-0.045408	0.306058	0.220140	0.556677	-0.131995	-0.217224	0.150814	0.214616
5	-0.012281	0.011277	-0.040120	-0.054194	0.030869	-0.041727	-0.193339	-0.026773	0.167220	0.640262	-0.337698	0.083092	0.149282

SF_Ratio	perc_alumni	Expend	Grad_Rate
-0.177164	0.205418	0.318606	0.255627
0.246077	-0.246024	-0.130727	-0.168686
-0.291240	-0.147090	0.227919	-0.205241
-0.168468	0.016115	0.086067	0.243113
-0.077059	-0.215815	0.074993	-0.115718
0.485844	-0.047150	-0.297458	0.215657

Fig 117 loaded dataset

Factor-loaded dataset

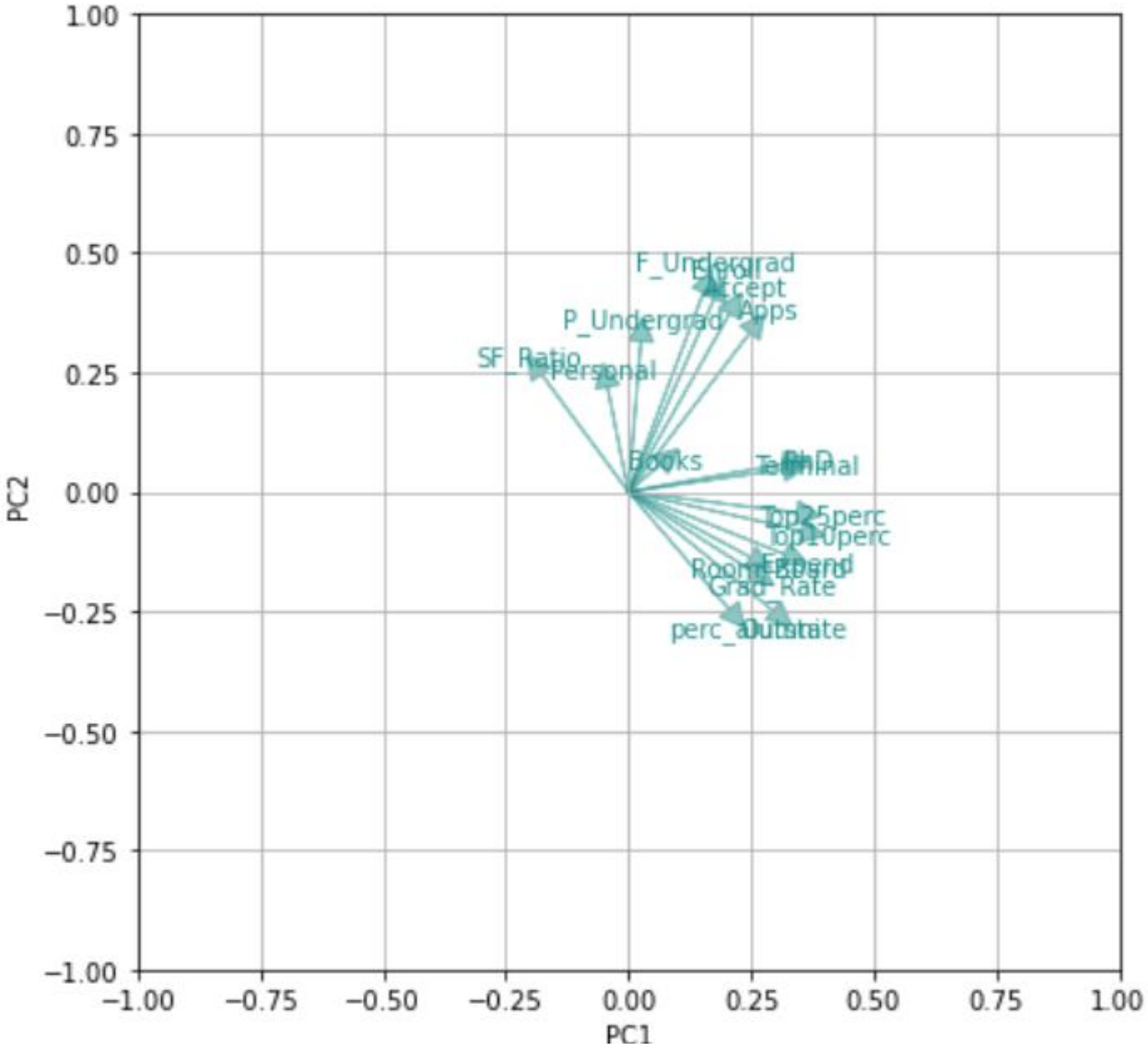
Fig 118 Factor loaded dataset

	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.578604	0.703943	-0.064676	0.284802	0.000164	-0.011305
Accept	0.482223	0.789596	-0.105948	0.271507	0.048935	0.010381
Enroll	0.409385	0.856036	-0.088172	0.163288	-0.056844	-0.036931
Top10perc	0.827448	-0.172600	0.037151	-0.055123	-0.380929	-0.049887
Top25perc	0.803400	-0.092992	-0.027645	-0.115735	-0.409830	0.028415
F_Undergrad	0.358869	0.885341	-0.065636	0.101188	-0.043899	-0.038411
P_Undergrad	0.060274	0.667293	0.149523	-0.158787	0.295890	-0.177971
Outstate	0.689479	-0.526746	0.052182	0.136431	0.212826	-0.024645
Room_Board	0.581793	-0.289957	0.164887	0.191378	0.538183	0.153929
Books	0.150260	0.119870	0.736624	0.071306	-0.127610	0.589370
Personal	-0.099573	0.465095	0.536786	-0.248776	-0.210007	-0.310855
PhD	0.747012	0.124629	-0.142951	-0.528649	0.145804	0.076488
Terminal	0.740461	0.099528	-0.076916	-0.517610	0.207486	0.137416
SF_Ratio	-0.414120	0.521089	-0.315634	-0.168294	-0.074499	0.447226
perc_alumni	0.480162	-0.520977	-0.159410	0.016098	-0.208645	-0.043402
Expend	0.744736	-0.276827	0.247009	0.085978	0.072501	-0.273814
Grad_Rate	0.597524	-0.357208	-0.222432	0.242863	-0.111874	0.198515

Loading plot

Shows how strongly each characteristic influences a principal component

Fig 119 loading plot



Dataset with 6 principal components

Fig 120 dataset with 6 pcs

	PC1	PC2	PC3	PC4	PC5	PC6
0	-1.591784	0.762095	-0.116125	-0.949497	-0.724680	-0.313275
1	-2.197267	-0.582571	2.317167	3.591952	0.984845	-0.126557
2	-1.431062	-1.096177	-0.434092	0.702472	-0.379120	-0.947275
3	2.860558	-2.623302	0.128983	-1.248905	-0.152182	-1.073626
4	-2.221481	0.018421	2.378961	-1.069785	0.706131	-0.007122

Names readded

	PC1	PC2	PC3	PC4	PC5	PC6	Names
0	-1.591784	0.762095	-0.116125	-0.949497	-0.724680	-0.313275	Abilene Christian University
1	-2.197267	-0.582571	2.317167	3.591952	0.984845	-0.126557	Adelphi University
2	-1.431062	-1.096177	-0.434092	0.702472	-0.379120	-0.947275	Adrian College
3	2.860558	-2.623302	0.128983	-1.248905	-0.152182	-1.073626	Agnes Scott College
4	-2.221481	0.018421	2.378961	-1.069785	0.706131	-0.007122	Alaska Pacific University
...	...	...	...	...	...	...	...
772	-3.334116	1.212538	-0.382427	0.133569	0.774943	0.318165	Worcester State College
773	0.207196	-0.687600	0.056045	0.541095	0.361400	0.380097	Xavier University
774	-0.735655	-0.078532	0.001609	0.072678	-0.517643	0.472162	Xavier University of Louisiana
775	7.927294	-2.048924	2.081758	0.843397	-0.960284	-2.068468	Yale University
776	-0.457852	0.362098	-1.335253	-0.178424	-1.133256	0.832020	York College of Pennsylvania

777 rows × 7 columns

Fig 121 names readded

Scatter plot of PC components

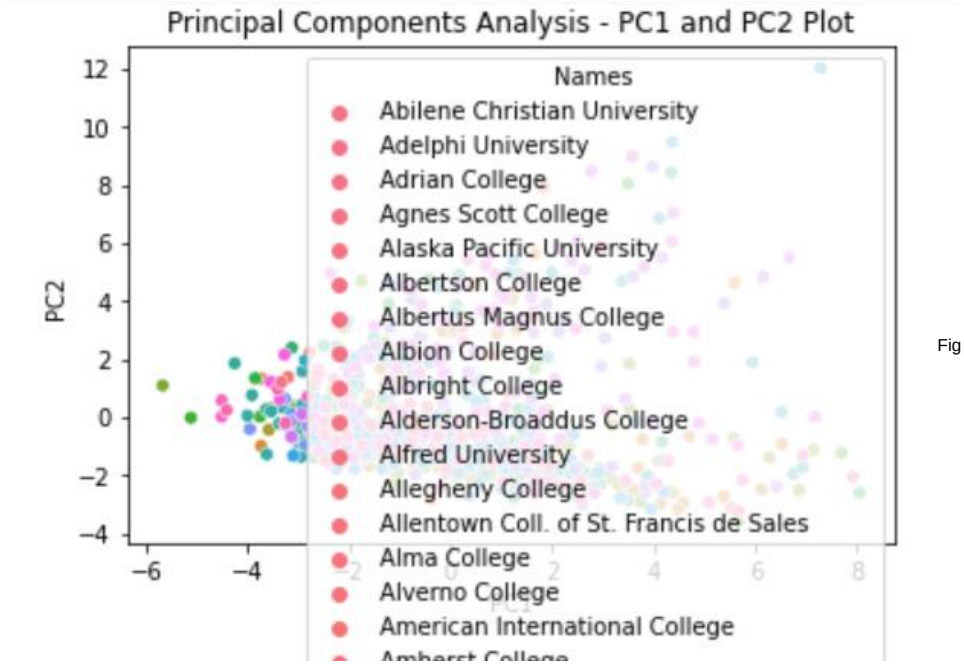


Fig 122 scatter pcs

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

The first PC is given by 0.25*Apps + 0.21*Accept + 0.18*Enroll + 0.36*Top10perc + 0.34*Top25perc + 0.15*F_Undergrad + 0.03*P_Undergrad + 0.29*Outstate + 0.25*Room_Board+ 0.06*Books - 0.04*Personal + 0.32*PhD + 0.32*Terminal - 0.18*SF_Ratio + 0.21*perc_alumni + 0.32* Expend + 0.26*Grad_Rate

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?

These are the cumulative variance explained.

Fig 123 cum variances

```
Cumulative Variance Explained [ 32.0988744   58.44246116  65.34259565  71.20525863  76.69617559
  81.67414057  85.23346016  88.68686349  91.80133169  94.17208083
  96.01178946  97.30603027  98.29167258  99.13175814  99.64913029
  99.86482309 100.        ]
```

As a general rule 80-20 is taken, for choosing the number of principal components which are chosen from the cumulative variance explained. Here, we see that 81% is achieved after the 6th Eigen value, hence 6 principal components have been chosen. The Eigenvectors determine the directions of the new attribute space, and the eigenvalues determine their magnitude. As can be seen in the PCA, the components of the eigen vectors determine the PCs.

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]

The dataset had too many attributes/variables(17, excluding the Names field). Performing PCA reduced the dimensionality of this large data set by transforming the set of attributes into a smaller one that still contained most of the information in the large set. Smaller data sets are easier to analyse and faster for ML algorithms without extraneous attributes to process.

This has been simple to achieve in this data set as the data set contained a large number of correlated variables and PCA is a powerful tool which reduces this multicollinearity. Thus, in this case study, PCA has reduced the number of attributes of the data set, at the same time has retained as much information is possible. In this data set, using the information on the Eigen values, Eigen vectors and cumulative variance explained, the 6 PCS out of the 17 have been identified. Since, choosing 6 PCS has captured 81% of the variance and information in the original data set.

As a general rule 80-20 is taken, for choosing the number of principal components which are chosen from the cumulative variance explained. Here, we see that 81% is achieved after the 6th Eigen value, hence 6 principal components have been chosen. The below gives the screenshot of the 6 PCs data frame.

The explicit form of the first PC is given by 0.25*Apps + 0.21*Accept + 0.18*Enroll + 0.36*Top10perc + 0.34*Top25perc + 0.15*F_Undergrad + 0.03*P_Undergrad + 0.29*Outstate + 0.25*Room_Board+ 0.06*Books - 0.04*Personal + 0.32*PhD + 0.32*Terminal - 0.18*SF_Ratio + 0.21*perc_alumni + 0.32* Expend + 0.26*Grad_Rate

The above coefficients, 0.25, 0.21…. indicate the weights associated with each of these 17 attributes which make up the first PC. Similarly, the other coefficients or the weights can be got from the PC data frame shown above.

Further analysis

Using a component loading on a heat map, the features that have maximum loading across the components can be identified. For each feature, the maximum loading value across the components can be found and the same can be marked with the help of rectangular box as seen in the below plot. Features marked with rectangular red box are the ones having the maximum loading on the respective component.

These marked features are used to decide the context that the component represents. Using the components additional rules can be derived and analyzed. Unsupervised learning like clustering can further be applied on the data to segment the universities/colleges based on the components created and further analyzed.
Component Loading on a Heat Map

Let's identify which features have maximum loading across the components.

- We will first plot the component loading on a heatmap.
- For each feature, we find the maximum loading value across the components and mark the same with help of rectangular box.
- Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents

Rectangle heat map

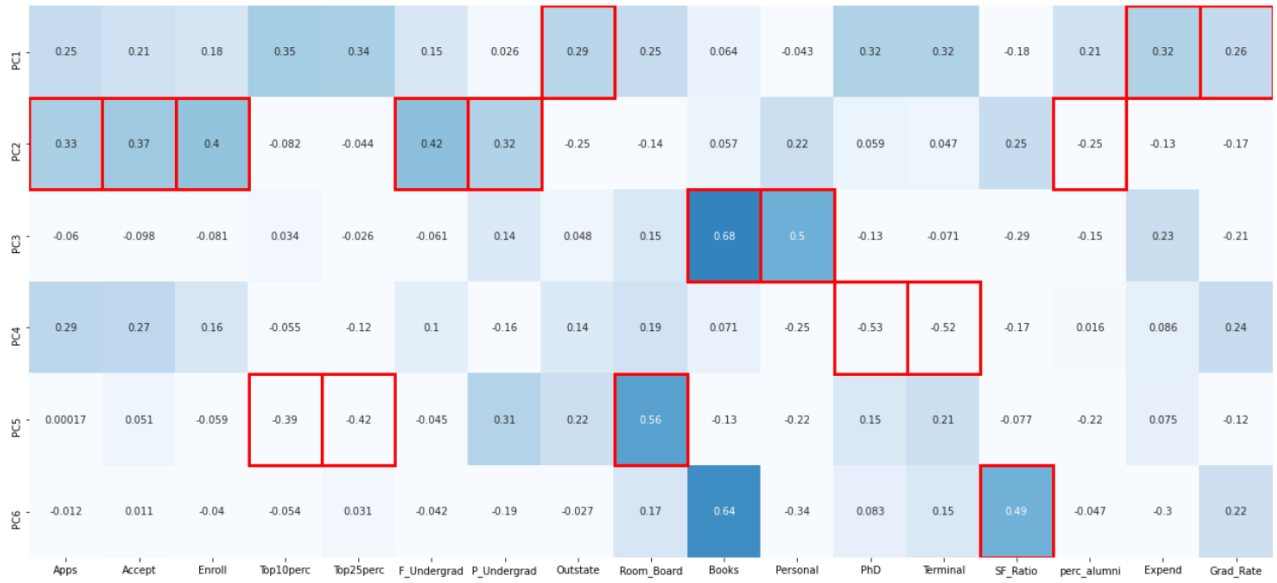


Fig 124 rectangle heat map