Problem 1A

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person’s educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

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

- The shape of the data is (40, 3) meaning the dataset has 40 rows and 3 columns.
- The dataset has no missing values or null values.
- The column ‘Education’ and ‘Occupation’ is of object data type while the column ‘Salary’ is of integer data type.

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

One-way ANOVA (Education)

Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.

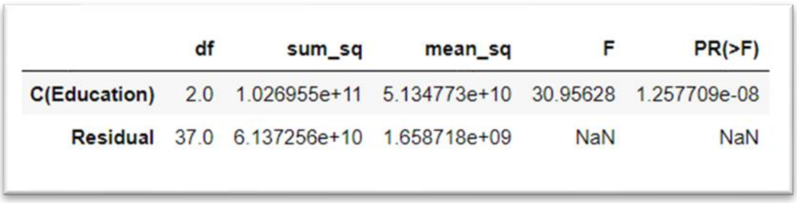
One-way ANOVA (Occupation)

Null Hypothesis $H0$: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of occupation.

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

Table 1: One-way ANOVA for 'Education'

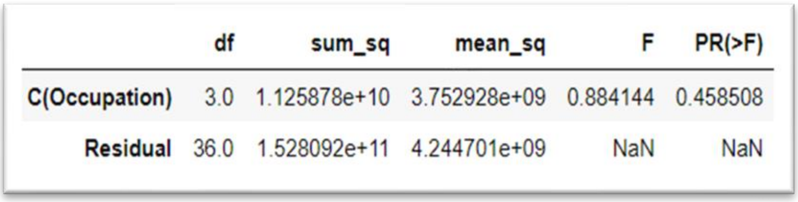


	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

From table 1, we can see that the p-value is 1.257709e-08. The p-value is lesser than the significance level (alpha = 0.05). Therefore the **null hypothesis ($H0$) is rejected**. From this result, we can conclude that there is difference in the mean salaries for at least one category of education.

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

Table 2: One-way ANOVA for 'Occupation'



	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

From table 2, we can see that the p-value is 0.458508. The p-value is greater than the significance level (alpha = 0.05). Therefore **we fail to reject the null hypothesis ($H0$) (we accept the null hypothesis)**. From this result, we can conclude that there is no difference in the mean salaries across the 4 categories of occupation.

1.4 If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. **(Non-Graded)**

Problem 1B

1.5 What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot. [Hint: use the 'pointplot' function from the 'seaborn' function]

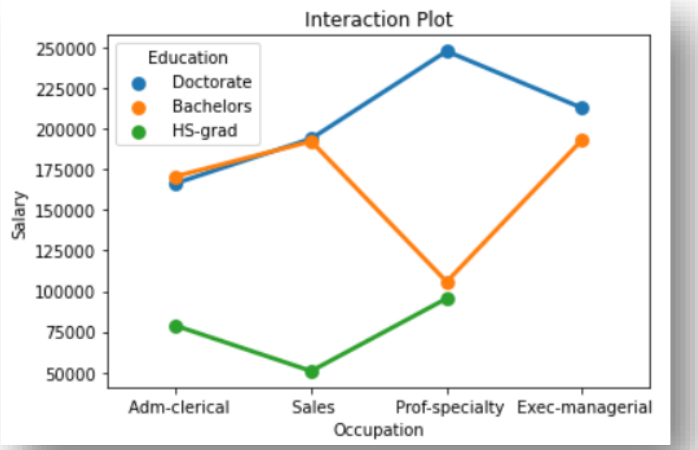


Figure 1: Interaction plot

The interaction of one variable on the other (Education and Occupation) is analysed with the help of an interaction plot. Figure 1 shows the interaction plot.

From figure 1, we can see that there is a significant amount of interaction between the categorical variables (Education and Occupation).

Observations:

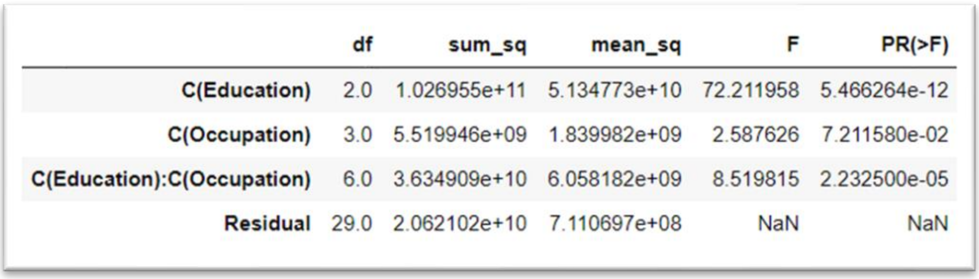
- People with Bachelors and Doctorate education earn more than the people with HS-grad education.
- People with HS-grad education are able to take up Adm-clerical, sales and prof-specialty as occupations but not Exec-managerial.
- People with Bachelors and Doctorate education are able to take up occupations in all 4 categories (Adm-clerical, sales, prof-specialty and Exec-managerial).
- People with sales occupation and HS-grad education earn the least out of all.
- People with prof-specialty occupation and doctorate education earn the most out of all.
- People with education as bachelors or doctorate and occupation as Adm-clerical and sales almost earn the same salaries (salaries ranging from 170000–190000) and higher than people with education as HS-grad.
- People with Bachelors education and Prof-specialty occupation earn lesser than other occupation with bachelor education.
- There are no people with education as HS -grad who hold Exec-managerial occupation.
- People with education as HS -Grad earn the lesser salary than people with Bachelors and Doctorate education.
- Bachelors education people with sales occupation earn higher than bachelors education people with prof-specialty occupation while doctorate education people with sales occupation earn lesser than doctorate education people with prof-specialty occupation.
- Bachelors education people with prof-specialty occupation earn lesser than bachelors education people with exec-managerial occupation whereas doctorate education people with prof-specialty occupation earn higher than doctorate education people with exec-managerial occupation.
- Among the people with HS-grad education, people with prof-specialty occupation earn the highest.
- Among the people with bachelors education, people with sales and exec-managerial occupation earn the highest.
- Among the people with doctorate education, people with prof-specialty occupation earn the highest.

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

Null Hypothesis $H0$: The effect of the variable ‘Education’ on the mean ‘Salary’ does not depend on the effect of the other independent variable ‘Occupation’. There is no interaction effect between the 2 independent variables ‘Education’ and ‘Occupation’.

Alternate Hypothesis $H1$: There is an interaction between the two independent variables ‘Education’ and ‘Occupation’ on the mean ‘Salary’.

Table 3: Two-way ANOVA table



	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

From table 3, we can see that there is a significant amount of interaction between the variables ‘Education’ and ‘Occupation’.

The p-value (of the interaction between ‘Education’ and ‘Occupation’) is 2.232500e-05. The p-value is lesser than the significance level (alpha = 0.05). Therefore the **null hypothesis ($H0$) is rejected**. From this result, we can conclude that there is an interaction effect between ‘Education’ and ‘Occupation’ on the mean salary.

From the two way ANOVA method and the interaction plot showing in figure 1, we see that education along with occupation results in higher and better salaries among the people. It is clearly seen that people with Doctorate education earn 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.

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

Performing ANOVA for this case study gives us insights about the data. As explained in question 1.6, we can see that there is an interaction effect between ‘Education’ and ‘Occupation’ on the mean salary. This is also seen in the interaction plot shown in figure 1. It is seen that education along with occupation results in higher and better salaries among the people. It is clearly seen that people with Doctorate education earn the maximum salaries and people with education HS-grad earn the least.

Hence we can conclude that education plays a very important role in the salary of a person. Therefore the business implication is that we have to encourage more people to study so that they get better jobs and thus better salaries. People having good jobs and earning well means people are able to afford a better lifestyle. Earning a higher salary means they are also able to pay better taxes which indirectly improves the standards to the city. Therefore education has to be promoted and more students should be encouraged to study.