1. What are the other ways to handle multicollinearity?

In bivariate analysis, multicollinearity typically refers to a situation where two predictor variables are highly correlated, which can distort the results and interpretation of the analysis. Here are some ways to handle multicollinearity in bivariate analysis:

**1. Remove One of the Correlated Variables**

* If two variables are highly correlated, consider removing one of them. This is the simplest and most straightforward approach.
* Choose the variable that is less theoretically important or has less predictive power.

**2. Combine the Correlated Variables**

* Create a new variable by combining the two correlated variables (e.g., taking the average, sum, or difference).
* This can help reduce multicollinearity while retaining the information from both variables.

**3. Use Principal Component Analysis (PCA)**

* PCA transforms the original correlated variables into a set of uncorrelated components.
* These components can then be used in the analysis instead of the original variables.

**4. Ridge Regression**

* Ridge regression is a technique that introduces a small bias (via a penalty term) to reduce the variance of the coefficient estimates.
* This can help mitigate the effects of multicollinearity.

**5. Partial Least Squares (PLS) Regression**

* PLS regression is a technique that projects the predictors into a new space where they are uncorrelated.
* It is particularly useful when the number of predictors is large relative to the number of observations.

**6. Centering or Standardizing Variables**

* Centering (subtracting the mean) or standardizing (subtracting the mean and dividing by the standard deviation) the variables can sometimes reduce multicollinearity.
* This is particularly useful when the variables are on different scales.

**7. Increase Sample Size**

* Increasing the sample size can sometimes help reduce the impact of multicollinearity, as the estimates become more stable with more data.

**8. Use Domain Knowledge**

* Sometimes, domain knowledge can help in deciding which variable to keep or how to combine variables in a meaningful way.

**9. Regularization Techniques**

* Techniques like Lasso regression can also be used to handle multicollinearity by shrinking the coefficients of the correlated variables.

**10. Check for Interaction Terms**

* Sometimes, multicollinearity can be a result of interaction terms. Ensure that interaction terms are appropriately included or excluded based on the context.

**11. Use Variance Inflation Factor (VIF)**

* Calculate the VIF for each variable to quantify the extent of multicollinearity. A VIF greater than 10 is often considered indicative of significant multicollinearity.
* Based on the VIF, you can decide which variables to remove or combine.

**12. Use Correlation Matrix**

* Examine the correlation matrix to identify pairs of variables with high correlation coefficients.
* This can help you decide which variables to address in your analysis.

**13. Bayesian Regression**

* Bayesian regression can incorporate prior information, which can help in situations where multicollinearity is present.

**14. Elastic Net Regression**

* Elastic Net combines the penalties of Ridge and Lasso regression, which can be effective in handling multicollinearity.

**15. Cross-Validation**

* Use cross-validation to assess the impact of multicollinearity on the predictive performance of your model.
* This can help you decide whether multicollinearity is a significant issue in your analysis.

Each of these methods has its own advantages and limitations, and the choice of method depends on the specific context and goals of your analysis.

1. Explain Variance Inflation Factor function ?

Bivariate ipython note book file has been prepared & given line by line explanations.

1. ANOVA analysis examples

One Way Classification (One Independent variable)

Problem statement -1:

There are three main brands of mobile. A set of 100 mobiles are examined and founded to be allocated among four groups A, B, C and D and three brands as shown below table,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Brands** | **Groups** | | | |
| **Battery** | **Charging** | **RAM** | **Front Camera** |
| Apple iPhone | 4,500mAh | 30W | 8GB | 12MP |
| Samsung | 5,000mAh | 45W | 12GB | 40MP |
| OnePlus | 6,000mAh | 80W | 16GB | 32MP |

Are there any significant differences between the brands preference?

Problem statement -2:

There are three main brands of Cars. A set of 100 cars are examined and founded to be allocated among four groups A, B, C and D and three brands as shown below table,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Brands** | **Groups** | | | |
| **Top Speed** | **Wheelbase (mm)** | **Boot Space** | **Mileage (km/l)** |
| Mercedes-Benz E-Class | 250 km/h | 2,939 mm | 540 L | 12-16 km/l |
| BMW 5 Series | 250 km/h | 2,975 mm | 530 L | 12-16 km/l |
| Audi A6 | 250 km/h | 2,924 mm | 530 L | 11-15 km/l |

Are there any significant differences between the brands preference?

Problem statement -3:

The following table shows the flying kms of four brands of Airlines.

|  |  |  |  |
| --- | --- | --- | --- |
| **Brands** | **Average Daily Flights** | **Average Distance per Flight (km)** | **Estimated Daily Kilometers Flown (km)** |
| Emirates | 3,600 | 5400 | 19,440,000 |
| Qatar Airways | 1600 | 4800 | 7,680,000 |
| Air India | 500 | 4200 | 2,100,000 |
| Indigo | 1400 | 4400 | 5,1800,000 |

Perform an analysis of variance test the homogeneity of mean hours of four brands of flights.

Two Way Classification (Two Independent variable)

Problem statement -4:

Five cricket players, each played five one day international cricket series for a certain tournament and observe the number of days each player has scored. The results are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cricket Players | Scores taken in Cricket Tournament | | | | |
| **1** | **2** | **3** | **4** | **5** |
| MS Dhoni | 45 | 23 | 42 | 28 | 53 |
| Virat Kholi | 52 | 75 | 85 | 42 | 40 |
| Rohit Sharma | 32 | 36 | 45 | 12 | 35 |
| KL Rahul | 62 | 75 | 30 | 41 | 50 |

Discuss the difference between 1). Cricket Players 2). Scores

Problem statement -5:

The following table gives monthly deliveries of Swiggy firm in three cities by its four delivery boys.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cities | Food Delivery boys | | | |
| **1** | **2** | **3** | **4** |
| Chennai | 55 | 63 | 72 | 78 |
| Bangalore | 62 | 75 | 85 | 62 |
| Mumbai | 72 | 86 | 65 | 52 |

Setup the analysis of variance table and test whether there is any significant difference

1). Between food deliveries by the firm delivery boys and

2). Between food deliveries in the three cities.

1. Implement Two way ANOVA analysis?

Bivariate ipython note book file has been updated & given what we have rejected and accepted hypothesis statements.

1. Bivariate Final assignment

1)Replace the NaN values with correct value. And justify why you have chosen the same.

At First, In order to handle missing values or null values or NaN (Not a Number) values exists in a dataset, there are four options given below from which we have to choose the best optimized one.

In the provided dataset ‘Placement.csv’ file, we could find that there is a column ‘salary’ has many missing values which means empty and no values filled. So now we are in the position to make a correction over the column by choosing any of the below four suitable option and get the dataset ready for further processing.

1). Replace the missing value with central tendency such as Mean, Median or Mode.

2). Entire Row Delete

3). With respect to the problem, we have to replace the missing values

4). We can consider this as a semi supervised, create a model, so that we can predict missing values.

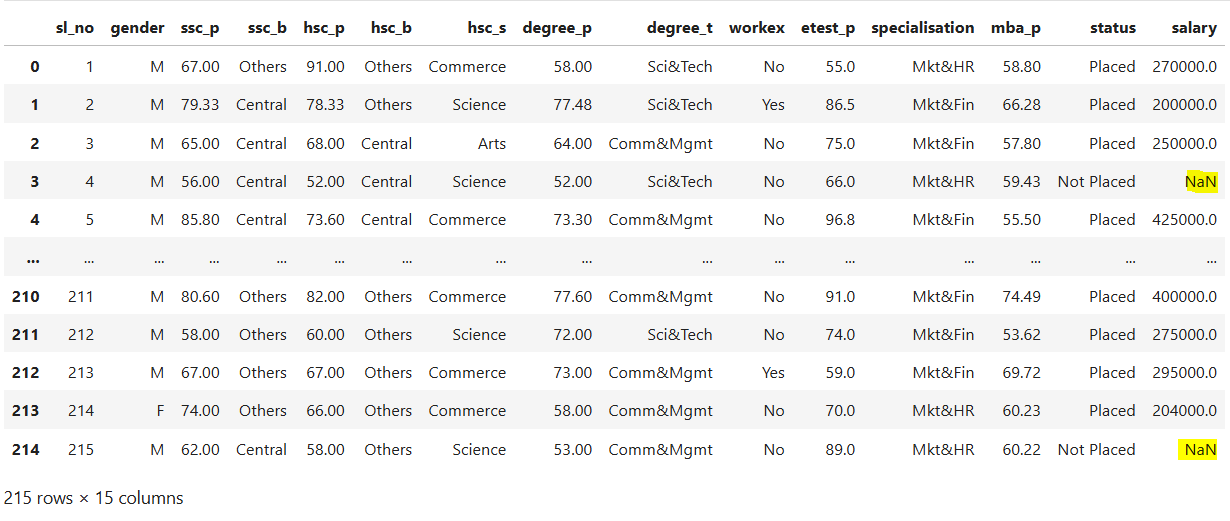
**Justification and reason for elimination of other options:**

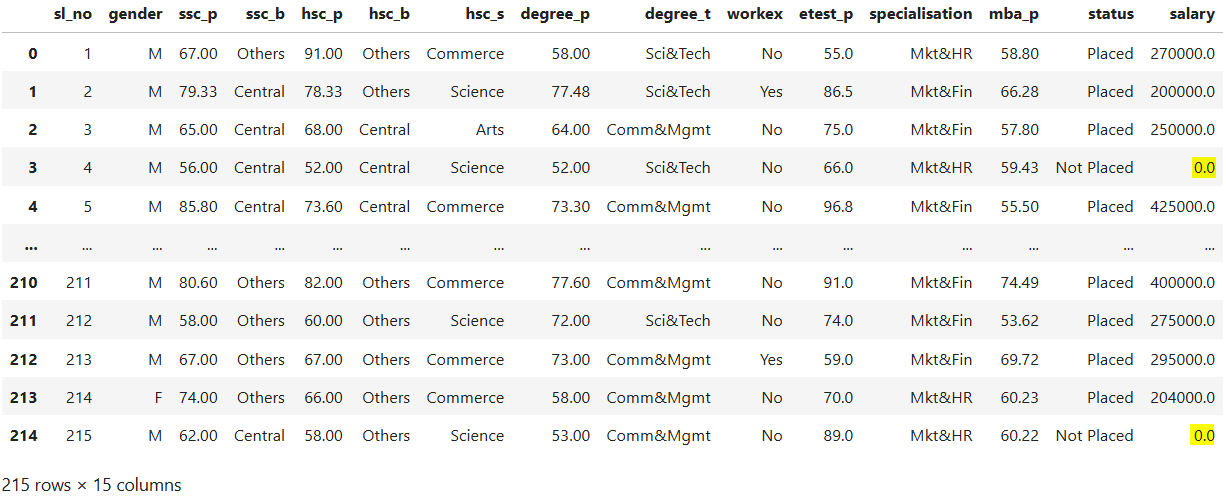
After we review the dataset and compared the columns ‘status’ and ‘salary’ to each other, we found that who ever has not placed on job would not get a salary. If we choose to replace it with any of the central tendency values or by creating model for predicting the salary values which would leads to change the meaning of the dataset or causes data manipulation. So, the first one and the fourth one is not the correct choices.

Here the ‘Salary’ column has not filled with any values for 67 rows out of 215 rows, Incase we choose to delete these entire rows, then we need to delete the whole 67 rows which causes a huge data losses in the dataset and it results incorrect final outcome of the processed model. So, this option has been eliminated.

By analyzing the dataset, **We can fill the ‘salary’ column with ‘0’ values that will NOT only the correct replacement BUT ALSO save the actual originality or meaning of the dataset**. So, this option has been selected as the best choice.

**Dataset Before replacement of Nan values:**



**Dataset After replacement of Nan values:**

2)How many of them are not placed? **DONE**

67 candidates are not placed.

3)Find the reason for non-placement from the dataset?

In ‘Placement.csv’ file dataset, there are eight qualitative columns and five quantitative columns.

As per my analysis over the dataset,Below five Quantitative columns are the primary deciding factors and eligibility criteria for the candidates to become placed or not placed.

* 1. sslc pass mark
  2. hsc pass mark
  3. degree pass mark
  4. entrance test pass mark
  5. mba pass mark

4)What kind of relation between salary and mba\_p? **DONE**

5)Which specialization is getting minimum salary? **DONE**

6). How many of them getting above 500000 salary? **DONE**

7)Test the Analysis of Variance between etest\_p and mba\_p at signifance level 5%.(Make decision using Hypothesis Testing)? **DONE**

8)Test the similarity between the degree\_t(Sci&Tech) and specialisa tion(Mkt&HR) with respect to salary at significance level of 5%.(Make decision using Hypothesis Testing) ? **DONE**

9)Convert the normal distribution to standard normal distribution for salary column?

10)What is the probability Density Function of the salary range from 700000 to 900000? **DONE**

11)Test the similarity between the degree\_t(Sci&Tech) with respect to etest\_p and mba\_p at significance level of 5%. (Make decision using Hypothesis Testing) **DONE**

12)Which parameter is highly correlated with salary?

Variance Inflation factor is a parameter where its value would be greater than 5 which indicates that it is highly correlated with all other columns (Salary column exhibits linear pattern with all the columns).

13) plot any useful graph and explain it. **DONE**