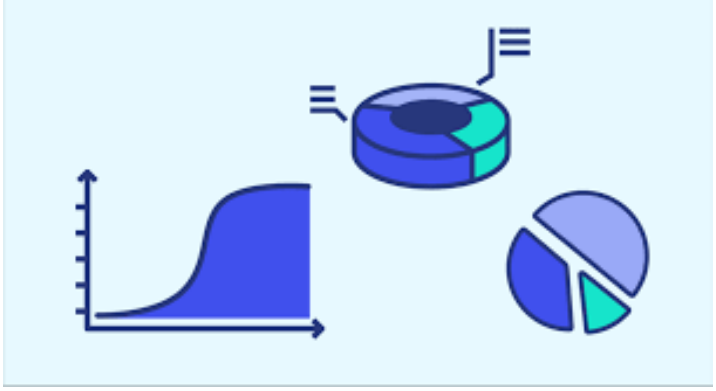
Exploratory Data Analysis of Placements of MBA Students



Report By:- Aditi Ranakoti, Sunita Yadav, Yash Gupta
BSC.(H) STATISTICS 6th Semester

Acknowledgement

We would like to express our sincere gratitude to our professor, Dr.Veena Budharaja for providing us with the opportunity to work on this project centered around Analysis of variation and variance amongst the data. Working on this project allowed us to bridge the gap between theory and practical applications. We gained a deeper appreciation for how concepts of design of experiment, EDA hypothesis testing , univariate and bivariate analysis can be used in various fields to get deeper insights from the raw real world dataset. Another significant takeaway from this project was the importance of effective collaboration and communication. Working on this project as a team was an enlightening and enriching learning experience.

We would like to thank her for her guidance, support, and expertise throughout this endeavor. We sincerely hope that you will consider providing more opportunities for us to do so. We would be thrilled to participate in any future projects you may have in mind.

Abstract

This project explores the placement outcomes of 215 MBA students through comprehensive analysis of a dataset comprising 14 key features. The dataset encompasses student details including gender, academic performance in 10th and 12th grades, specialization in MBA, work experience, placement status, and salary among others. Notably, students coming from diverse educational backgrounds, having completed their 10th and 12th grades either from central boards or other boards, and pursued degrees across three distinct streams: Science & Technology, Commerce & Management, and Arts.

Utilizing statistical tool like Spss, Excel and Python for Exploratory Data Analysis (EDA) and statistical tests, we investigated various factors influencing student placements. Our findings reveal intriguing insights into the dynamics of MBA placements, shedding light on correlations between academic performance, specialization choices, and placement success. Additionally, our analysis identifies distinct trends across different educational backgrounds and streams, offering valuable implications for both students and educational institutions.

Significantly, out of the 215 students included in the dataset, 67 were not placed, emphasizing the importance of understanding the underlying factors contributing to placement outcomes.

Further analysis revealed:
Average Package: We found that the average salary package offered to placed students is- $ 288656.0
Placement Rate: The overall placement rate is 69%
MBA in Marketing & HR: 55.8%
MBA in Marketing & Finance: 71.2%

By leveraging EDA techniques and statistical tests, this study contributes to the body of knowledge surrounding MBA placements, providing actionable insights for stakeholders in the field of education and recruitment."

Table of Contents

INTRODUCTION

Background:

The placement outcomes of MBA students serve as crucial benchmarks of academic success and future career trajectories. Identifying and understanding the factors influencing these outcomes can provide valuable insights for educational institutions, recruiters, and students. This project aims to dissect the placement dynamics of 215 MBA students through an exhaustive analysis of a multifaceted dataset.

Dataset Overview:

The dataset encompasses 14 key features, capturing a spectrum of dimensions from the students' profiles, including:

- Gender
- Academic performance in 10th and 12th grades
- Specialization in MBA
- Work experience
- Placement status
- Salary, among others

The students hail from diverse educational backgrounds, having completed their 10th and 12th grades either from central boards or other boards. Additionally, they have pursued degrees across three distinct streams: Science & Technology, Commerce & Management, and Arts.

Objectives:

The objectives of this analysis include:

Univariate Analysis: To profile the demographic distribution, academic performance, and specialization choices using pie charts and distribution plots.
Bivariate Analysis: To explore relationships between various attributes and placement outcomes, encompassing:
Independence of Attribute Tests: Assessing the independence of various categorical columns.
ANOVA Tests: Investigating the effects of different factors on placement outcomes.
Randomized Block Design (R.B.D): Evaluating the impact of work experience on placement status while controlling for degree streams and academic performance.

Descriptive and Inferential Statistics: To derive insights on:
Salary distribution among placed students.
Placement rates across different specializations and academic performance brackets.
Logistic Regression Modeling: To fit a logistic regression model using SPSS for classifying placed vs. unplaced students, exploring the predictive power of various factors on placement success.

This study employs a multi-layered analytical approach to decipher the intricate placement landscape for MBA students:

Python for EDA: Utilized Python for preliminary univariate and bivariate analysis, generating descriptive statistics, and visualizing distributions using pie charts, histplots, and kde plots.

SPSS for Logistic Regression: Employed SPSS software to fit a logistic regression model, facilitating classification and predictive analysis of placement outcomes.

Excel for Descriptive Statistics: Utilized Excel for calculating descriptive statistics, generating insights,performing tests and facilitating data visualization through box plots.

By integrating a diverse array of analytical techniques and tools, this study offers a holistic understanding of the multifactorial influences shaping MBA placement outcomes, providing actionable insights and strategic implications for stakeholders in the education and recruitment sectors.

Python code link:

https://drive.google.com/drive/folders/19_y3l7ClisvHmwOCiYfneX8Q1jiq42O0?usp=drive_link

RELATED LITERATURE

- **Descriptive Statistics:-**

Descriptive statistics refers to a set of statistical techniques used to summarize and describe the main features of a dataset. These techniques help in understanding the characteristics of the data and in making sense of the information it contains without necessarily making any conclusions or inferences about a larger population.

- **Mean**: The mean, also known as the arithmetic average, is a measure of central tendency that represents the sum of all values in a dataset divided by the total number of values.

  Formula: For a dataset with n values$x_1, x_2, \ldots, x_n$, the mean (μ) is

  $$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  The mean is commonly used to describe the central value of a dataset.

- **Median**: The median is the middle value in a dataset when the values are arranged in ascending or descending order. If the dataset has an odd number of values, the median is the middle value. If the dataset has an even number of values, the median is the average of the two middle values.

- **Variance:** Variance is commonly used in statistics and data analysis to quantify the degree of variability or dispersion in a dataset. It is often accompanied by the standard deviation, which is simply the square root of the variance.

- **Skewness:** Skewness measures the asymmetry of the distribution. A positive skewness indicates that the distribution is skewed to the right (longer right tail), while a negative skewness indicates that the distribution is skewed to the left (longer left tail.
- **Kurtosis:** Kurtosis measures the "tailedness" or peakedness of the distribution relative to the normal distribution. A higher kurtosis value indicates heavier tails and a sharper peak compared to the normal distribution.

  $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

  Python dataframe functions:

  Dataframe_name.describe()

  Dataframe_name.mean()

  Dataframe_name.median()

  Dataframe_name.var()

o Graphical Representation:-

- **Histogram**: A histogram is a graphical representation of the distribution of numerical data. It consists of a series of adjacent bars, where the area of each bar represents the frequency or proportion of observations falling into a specific interval. Histograms are commonly used to visualize the frequency distribution of continuous data and to identify patterns such as skewness, central tendency, and variability.

- **Bar Plot**: A bar plot is a graphical representation of categorical data using rectangular bars. The length or height of each bar represents the frequency, count, or proportion of data points in each category.

- **Box Plot (Box-and-Whisker Plot):** A box plot is a graphical representation of the distribution of numerical data through five summary statistics: minimum, first quartile (Q_1), median (Q_2), third quartile (Q_3), and maximum. The box represents the interquartile range (IQR), while the whiskers extend to the minimum and maximum values within a certain range.

Box plots are useful for visualizing the central tendency, spread, and variability of numerical data. They also help in identifying outliers and comparing the distributions of different groups or variables.

- **Pie Chart**: A pie chart is a circular statistical graphic divided into parts to illustrate numerical proportions. Each part represents a proportionate part of the whole, with the size of the part corresponding to the magnitude of the proportion it represents.

Pie charts are commonly used to visualize the composition of a whole or to compare the relative sizes of different categories within a single variable.

- **KDE PLOT(Kernel Density Estimation):**A KDE plot represents the probability density function of a dataset by approximating it with a sum of kernel functions, typically Gaussian (normal) kernels centered at each data point.
- KDE plots are useful for visualising the distribution of continuous data and identifying patterns such as multimodality, skewness, or outliers.

- **Q-Q Plot:** A Q-Q plot is a graphical tool used to assess whether a given dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset to the quantiles of a theoretical distribution, allowing for visual examination of how closely the data matches the expected distribution.

  Interpretation:

  1) **Perfect Normality:** If the dataset perfectly follows the theoretical distribution (e.g., normal distribution), the points on the Q-Q plot will fall along a straight line. This indicates that the empirical quantiles match the theoretical quantiles, suggesting a close fit to the assumed distribution.

2) **Deviation from Normality**: If the points deviate from the straight line, it suggests that the dataset deviates from the assumed distribution. Points curving upwards indicate heavier tails in the dataset compared to the theoretical distribution, while points curving downwards suggest lighter tails.S-shaped patterns suggest skewness in the dataset relative to the theoretical distribution.

- **Independence of Attributes Test:**In a Chi-square test of independence, the aim is to determine whether there is a statistically significant association between two categorical variables. It assumes that the observations are independent of each other.

  - **Procedure of Chi-square test:-**

    1) **Formulate Hypotheses:** Start with the null hypothesis (H_o) that there is no association between the two variables. The alternative hypothesis (H_1) is that there is an association between the variables.
    2) **Create Contingency Table**: Organize the data into a contingency table. This table displays the frequency counts of observations for each combination of the two categorical variables.
    3) **Calculate Expected Frequencies:** Under the null hypothesis, calculate the expected frequencies for each cell in the contingency table. This is done using the formula: (row total * column total) / grand total.
    4) **Compute the Chi-square Statistic**:

    $$\chi 2=\sum\frac{(O-E)^2}{E}$$

    Where:

    O = Observed frequency

    E = Expected frequency

    5) **Determine Degrees of Freedom**: The degrees of freedom (df) for a Chi-square test of independence is given as:

    $$df=(r-1)\times(c-1)$$

    Where:

    r = Number of rows in the contingency table

    c = Number of columns in the contingency table

    6) **Find Critical Value or P-value**: Based on the calculated Chi-square statistic and degrees of freedom, you can find the critical value from the Chi-square distribution table or directly compute the p-value.
    7) **Make a Decision**: If the p-value is less than the chosen significance level (usually 0.05), you reject the null hypothesis and conclude that there is a significant association between the variables. If the p-value is greater than the significance level, you fail to reject the null hypothesis.

○ For Randomized Block Design (RBD):

The model for RBD is $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

where ,

$Y_{ij} = Observation\ from\ ith\ treatment\ of\ the\ j\ th\ block$

$\mu = general\ mean\ effect$

$\alpha_i = additional\ effect\ due\ to\ ith\ treatment$

$\beta_j = additional\ effect\ due\ to\ jth\ block$

And the assumptions are:

Y_{ij}'s are independent of each other.

- $\varepsilon_{ij} \sim N(0, \sigma^2)$ are also independent.
- All the effects are additive in nature.

For RBD we have ANOVA table as follows:

Source Of Variation	d.f.	SS	MSS	F-ratio
Treatments	t-1	SST	MST	$F = \frac{MST}{MSE} \sim F_{(t-1,(r-1)(t-1))}$
Blocks	r-1			$F = \frac{MSB}{MSE} \sim F_{(r-1,(r-1)(t-1))}$
Error	(t-1)(r-1)	SSE	MSE	
Total	rt-1	TSS		

Where,

$$SST = \sum_{i=1}^{t} r\,(\bar{Y}_{i.} - \bar{Y}_{..})^2\ ,$$

$$SSE = \sum_{i}\sum_{j}\left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\right)^2\ ,SSB = \sum_{j=1}^{r} t\left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2\ \ and$$

$$TSS = \sum_i \sum_j \left(Y_{ij} - Y_{..}\right)^2$$

As well as $TSS = SST + SSB + SSE$.

Under the hypothesis that

$H_{\alpha 0}$: $\alpha_i = 0\ \forall\ i$ (i.e. there is no effect due to the treatments)

$H_{\alpha 1}$: $atleast\ one\ \ \alpha_i \neq 0$

Procedure:

Organize Data: Arrange your data in a table format with columns for the blocking factor, treatment, and response variable.

Conduct Analysis of Variance (ANOVA):

Select the range of your data---> Go to the "Data" tab and click on "Data Analysis"---> Choose "ANOVA: Single Factor" and click "OK".

In the "Input Range" box, select the range of your response variable.
In the "Grouped By" box, select the range of your blocking factor.,Click "OK"

Interpret Results: Look for the "F" value and its associated p-value in the ANOVA output. If the p-value is less than your chosen significance level (e.g., 0.05), you reject the null hypothesis, indicating that there is a significant difference between treatment means.

Post-hoc Tests (Optional): If your ANOVA results indicate a significant difference between treatment means, you may want to perform post-hoc tests to determine which specific treatments differ significantly from each other.

- **Logistic regression** is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors.

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

Procedure in Spss:

Data Preparation:

- Load your dataset into SPSS.
- Make sure your categorical independent variable is coded appropriately (usually as a nominal variable with numerical codes).
- Ensure your outcome variable is binary (e.g., 0 and 1).

Running Logistic Regression:

- Go to Analyze > Regression > Binary Logistic > continue > ok
- Transfer your outcome variable to the Dependent box.
- Transfer your independent variables (both categorical and numerical) to the Covariates box.
- Setting Up Categorical Variables:
- If your categorical variable is not recognized as nominal, you need to specify it as such:
- Click on the Categorical button.
- Move your categorical independent variable to the Categorical Covariates box.

Interpreting the Output:

Look at the Model Summary section to see the goodness-of-fit statistics like the -2 Log likelihood, Cox & Snell R Square, and Nagelkerke R Square.

Check the Variables in the Equation table to see which variables are significant predictors of the outcome. Look for the Sig. column to identify statistically significant predictors (typically, you'd look for a significance level below 0.05).

Examine the Omnibus Tests of Model Coefficients table for the overall significance of the model.

The logistic function is given by:

$\text{logit}(p)=\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_K$

where:

- P is the probability of the evnt occuring
- $\text{logit(p)} = \log(\frac{p}{1-p})$ is the log-odds of the event occuring.
- $\beta_0, \beta_1, \ldots \ldots, \beta_k$ are the coefficients of logistic regression model.
- $X1, X2, .. Xk \; are \; the \; independent \; variables$

To obtain the fitted probabilities (i.e., predicted probabilities) from the logistic regression model, you can use the following formula:

$$\hat{p} = \frac{1}{1 + e^{-logit(p)}}$$

The Nagelkerke R-squared is a measure of the proportion of variance explained by the model, similar to R-squared in linear regression. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data.

- **Sampling** is a process used in statistics and research to select a subset of individuals or items from a larger population. This subset, known as a sample, is used to make inferences or draw conclusions about the population from which it was drawn. Sampling is essential when it is impractical or impossible to study the entire population.

There are various sampling methods, each with its own advantages, disadvantages, and suitability for different types of studies:-

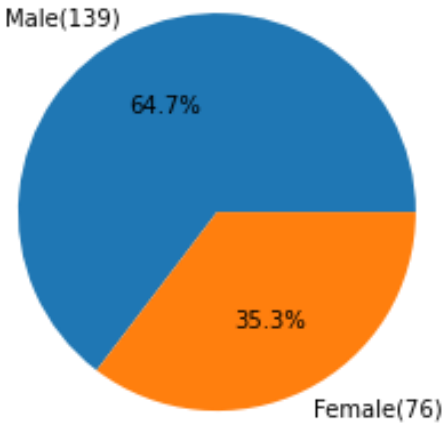
- **Simple Random Sampling**: In simple random sampling, each member of the population has an equal chance of being selected, and every possible sample of a given size has an equal chance of being chosen. This method is straightforward and unbiased but may be impractical for large populations.
- **Stratified Sampling:** Stratified sampling involves dividing the population into homogeneous subgroups or strata based on certain characteristics (e.g., age, gender, income), and then taking a random sample from each stratum. This method ensures representation from each subgroup and can increase the precision of estimates.

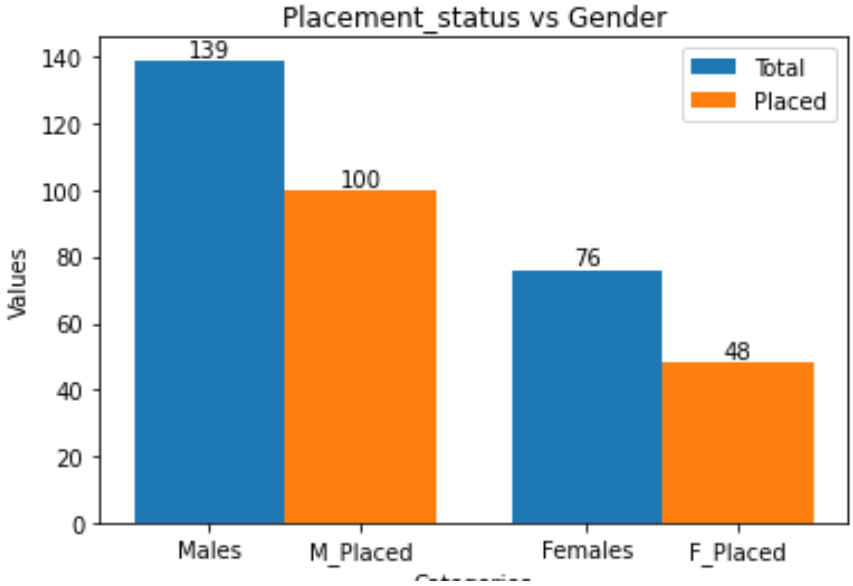
Other types are *Systematic Sampling, Cluster Sampling, Snowball Sampling, Convenience Sampling*

E.D.A- Analysis

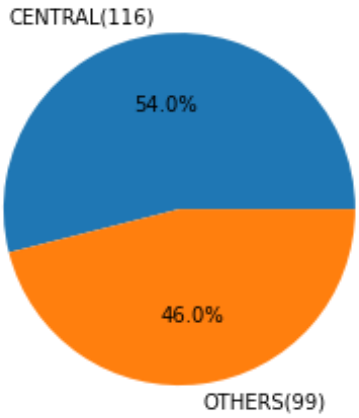
Exploratory Data Analysis (EDA) is a critical initial step in the data analysis process that focuses on summarizing the main characteristics of a dataset, often employing visual methods. It aims to understand the underlying patterns, relationships, anomalies, and insights within the data, thereby providing a foundation for more in-depth analyses and modeling..

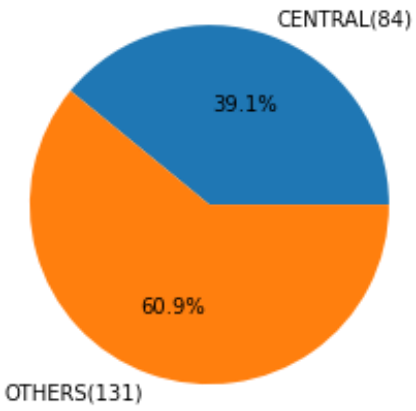
Male and Female Ratio In the Class:





Insights on Educational Background of Students:





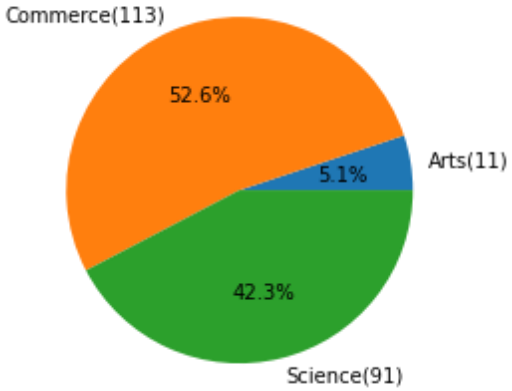
(i) (ii)

(i) Shows the proportion of students who were in central/other board in class 10th (SSC Board)

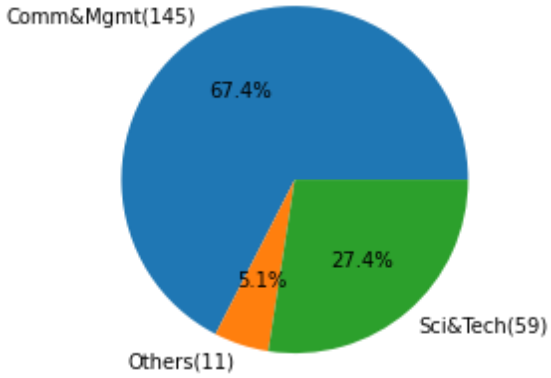
(ii) Shows the proportion of students who were in central/other board in class 12th (HSC Board)

Insight: A significant drop in the no. of students who pursued education from Central Board is seen moving from 10th to 12th standard (No proper reason can be explained for that.)

12th Board Stream of MBA Graduates:

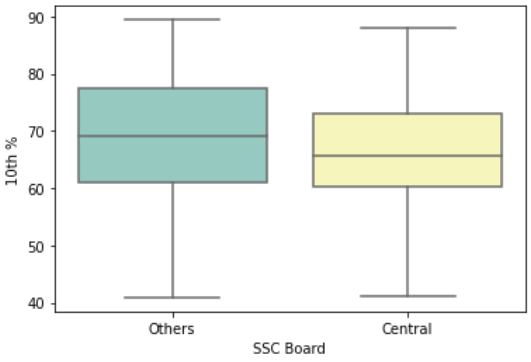


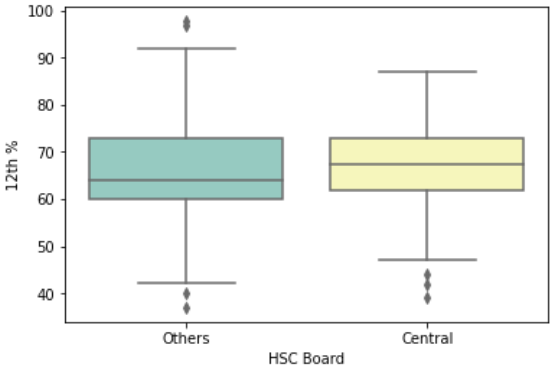
Degree Background of MBA Graduates:



Insight: MBA is the hot choice among the commerce graduates. And a lot of students changed their stream from science to commerce.

Percentages of students in class 10th (SSC) and 12th :



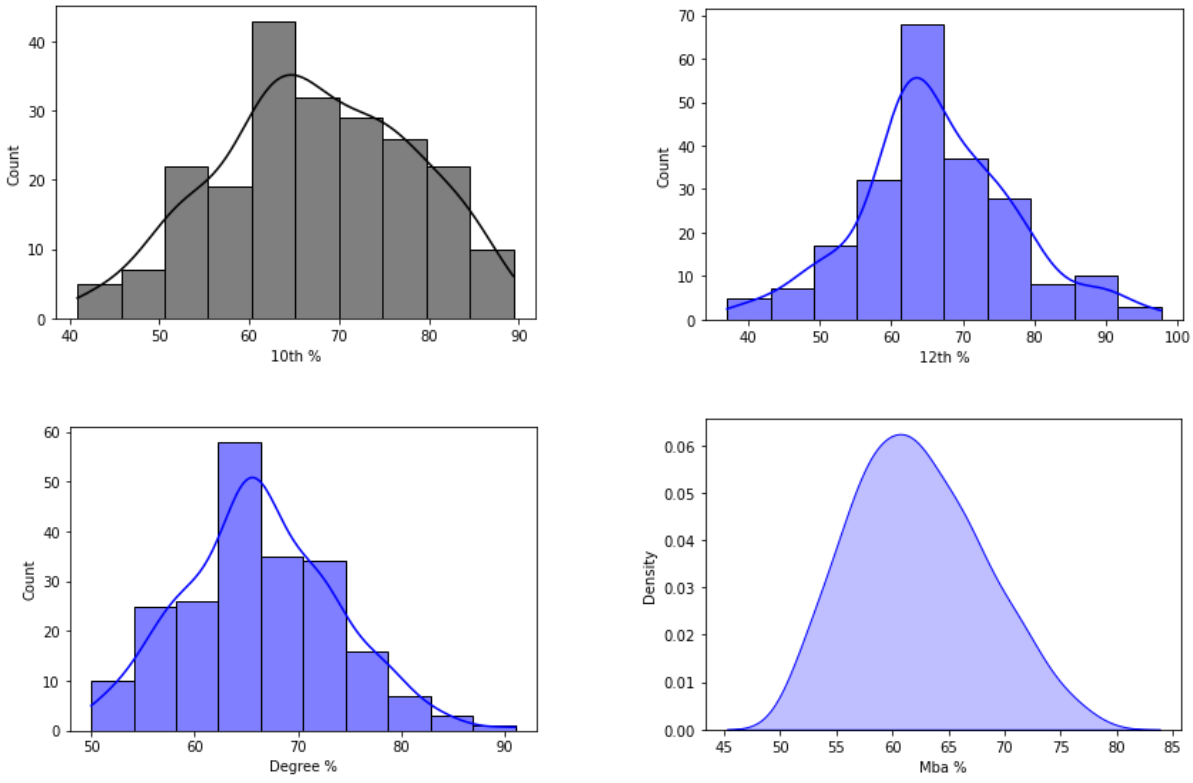


mean	67.303395
std	10.827205
min	40.890000
25%	60.600000
50%	67.000000
75%	75.700000
max	89.400000

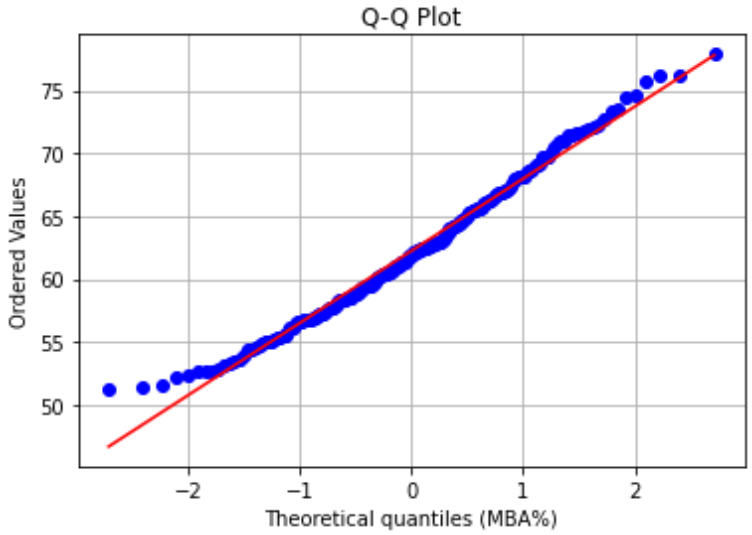
mean	66.333163
std	10.897509
min	37.000000
25%	60.900000
50%	65.000000
75%	73.000000
max	97.700000

Insight: Only 25% student scored above 75 % in class 10th and 12th
Variability: There is considerable variability or diversity in the class percentages. The data points are not tightly clustered around the mean; instead, they are more spread out, indicating less consistency across the dataset.

"Distribution of Percentage Scores Across Educational Phases: KDE + Histogram Analysis"



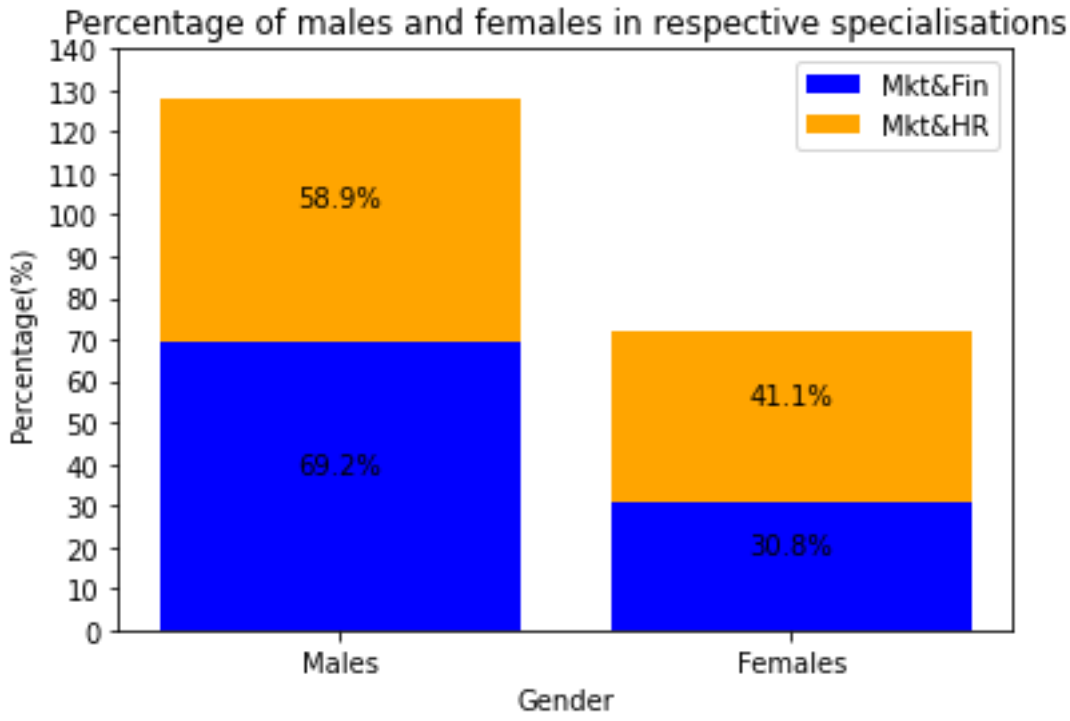
Insight: Student's percentage data is approximately Normal. (KS-test rejects the normality assumption as there are outliers in our data, but not due to sampling mistakes).



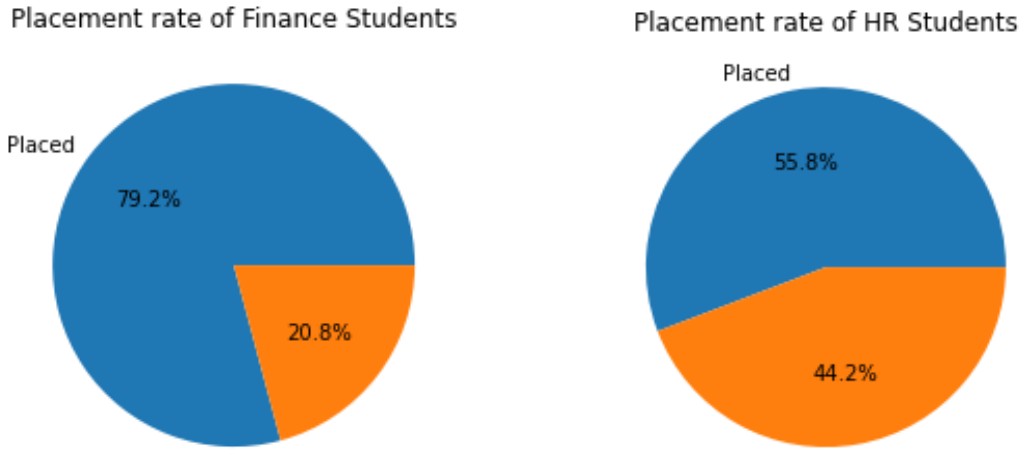
"Deviation near the end tells us the deviation from normality and s-sahped structure tells us that distribution is skewed compared to normal distribution"

Barplot showing the distribution of students across two specialisations:

1) **Mkt&Fin**
2) **Mkt &HR**

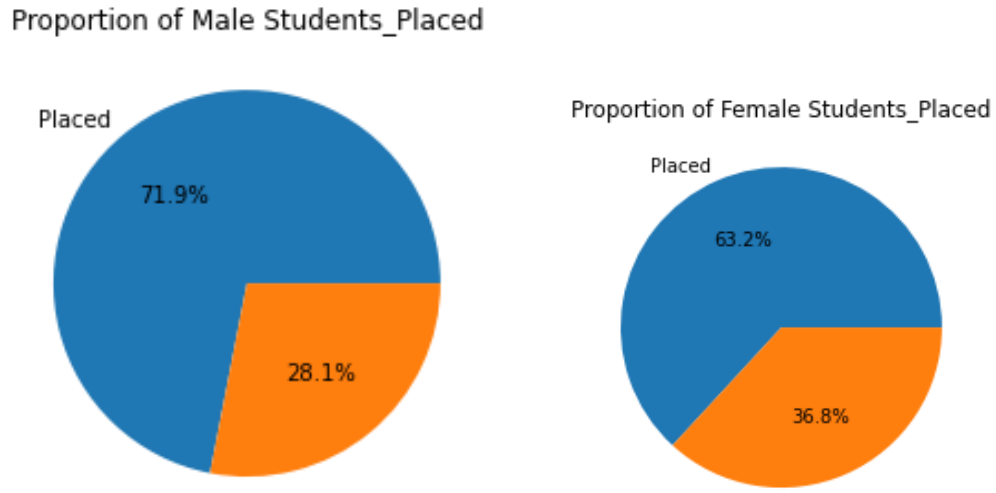


Placement rate across two different specialisations:



Insight: We see a higher placement rate in Mkt&Finance Dept. (**Overall Placement rate is: 69%).** 95 students from Mkt&Fin and 53 students from Mkt&HR were placed.

Placement rate in Males and Females:



Insight: Male students have a better placement rate as compared to female students
Relative placement success = $\left(\frac{100}{76} - \frac{48}{76}\right) * 100$ = ***68.42% higher of Male Students***

Academic Excellence and Placement Success: Examining the Impact of High Academic Performance on Placement Rates

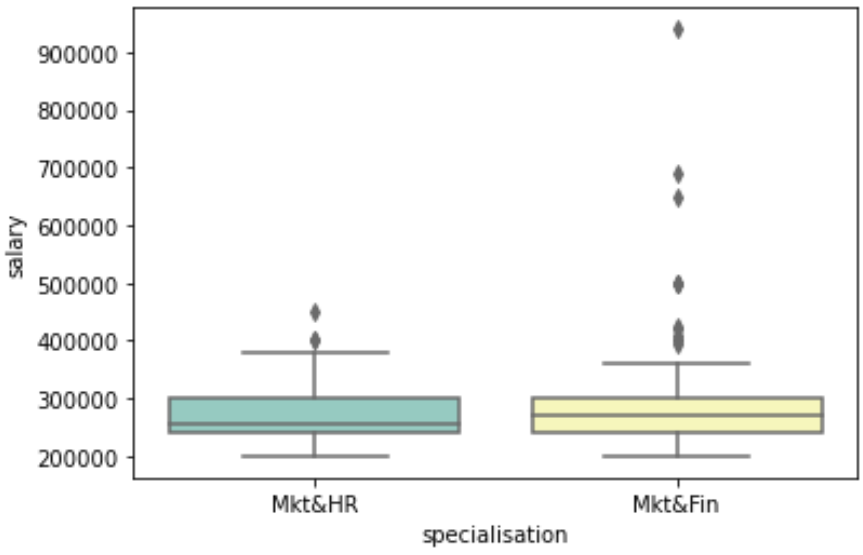
37 students who have scored above median percentage in their academic career at all levels of their education i.e. 10th, 12th, Degree and MBA have

100% placement rate irrespective of their work_experience

Code: toppers=*data[(data['10th %']> data['10th %'].median()) & (data['12th %']> data['12th %'].median())&(data['Mba %']> data['Mba %'].median())&(data['Degree %']> data['Degree %'].median())]*

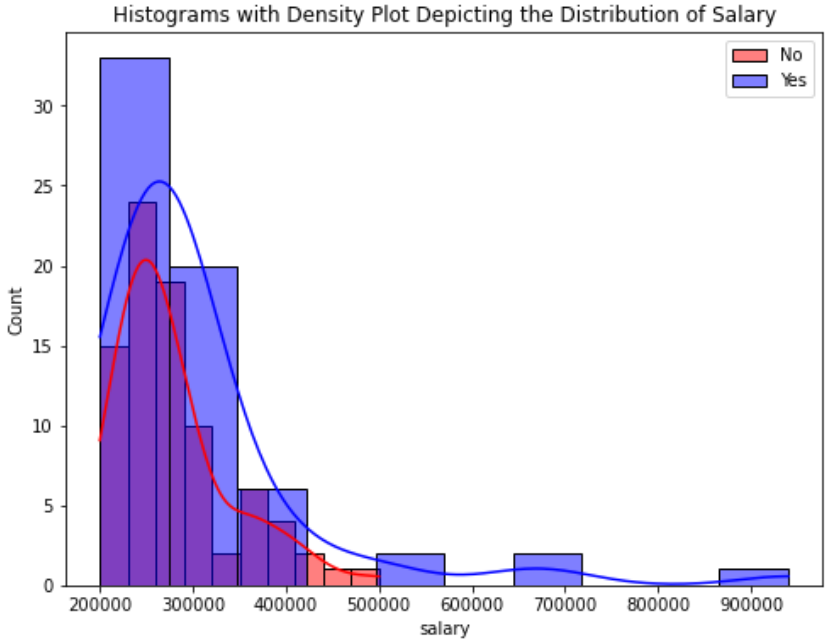
Salary Analysis of MBA Students: Exploring Distribution and Trends Through Visualization

Insight: The above boxplot depicts that the median salary of students who were placed is somewher near 270000.0 per month Packages above 4.0 per month are grabbed only by only 6 students **in the batch which are reflected by outliers. Width of the box plot hint at presence of high variabilty in salary of students, and that our distribution is skewed.**



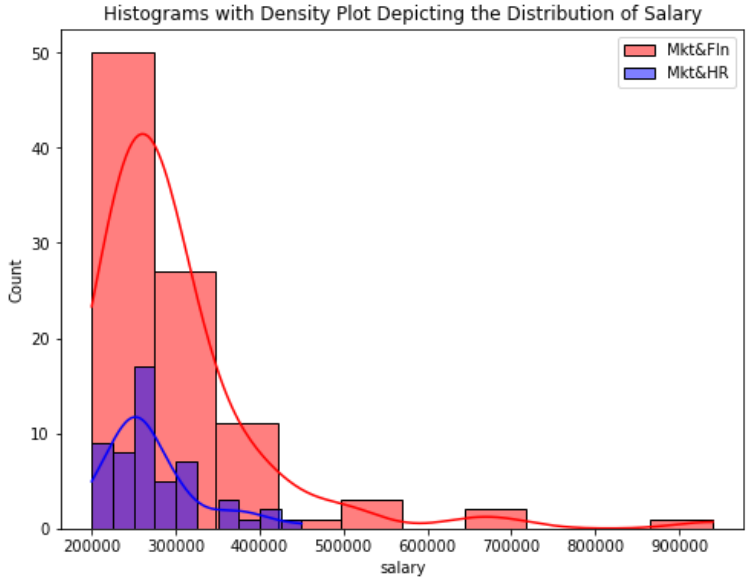
Insight: Box plot tells us that the variabilty in salary in both the specialisation group is same and there are outliers present in Mkr&Fin department i.e. few students got an exceptionally higher package than most of class.

Salary vs work exp:



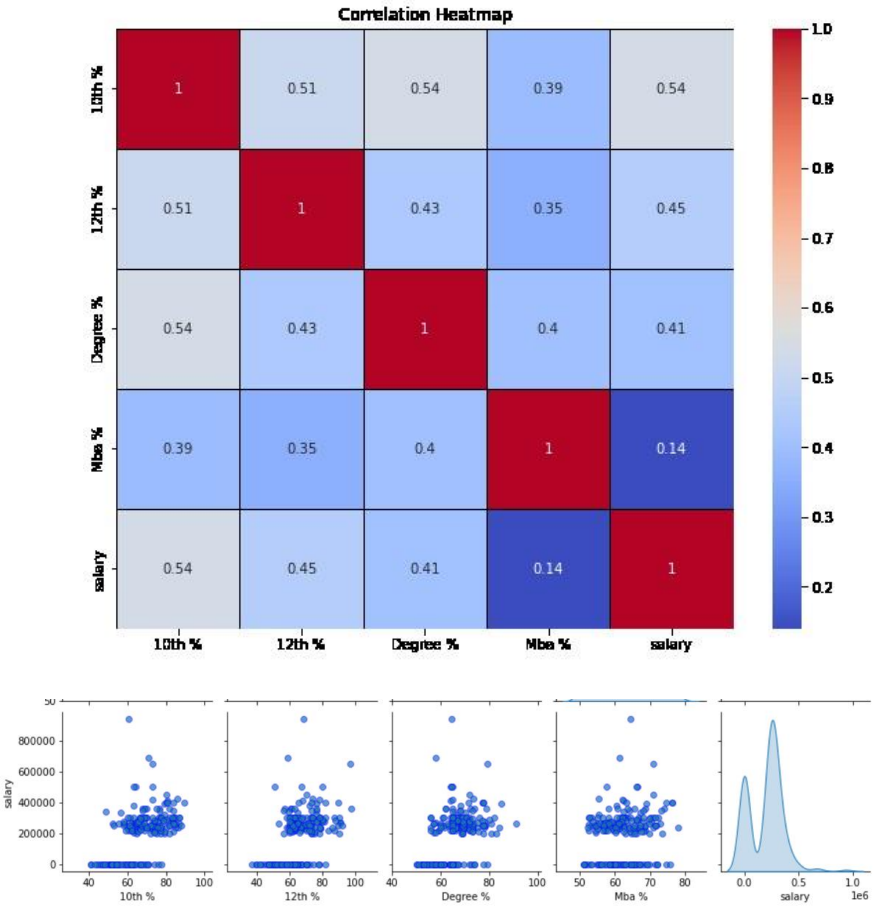
Insight: ***Higher packages are grabbed by students with experience, but that is not significantly large. Thus we can say that salary package of students are not dependent on Work Exp.(confirmed by the anova test in further section)***

Salary vs specialisations:



Insight: Packages above 450 K are grabbed only by students of Mkt&Finance Dept.(No:6**)**
A right-skewed curve of salary distribution indicates that the majority of salaries are concentrated towards the lower end of the scale in both the specialisation.This implies that there are relatively fewer individuals earning higher salaries compared to those earning lower salaries**.**
Skewness: 3.570 suggests moderate positive skewness, indicating that the distribution is skewed to the right
Kurtosis: 18.544 suggests heavy-tailedness and a sharper peak compared to the normal distribution

Correlation Analysis: Relationship Between 10th, 12th, Mba % and Salary column



Insight: The above correlation heatmap and scatterplot hints moderate correlation between the academic performance and salary of MBA students, which may also be the factor in Placement of them.(Going forward we have fitted a logistic model using the same insight)

Average Package = $ 288656.0
Median Package = $ 265000.0
Maximum Package = $ 940000.0.
Minimum Package = $ 200000.0
Maximum Package in Mkt&Fin = $ 940000.0
Maximum Package in Mkt&HR= $ 450000.0
Maximum Package among Females = $ 650000.0
Maximum Package among Males = $ 940000
Standard Deviation = $ 93457.45

Other Placement Statistics:

1) Probaility of a student getting placement given that he has work experience is**: 90.54%.**

2) Probaility of a student getting placement given that he has no work experience is**: 59.57%.**

3) **57** out of **67** students who didn't got placement were without work experience.

4) Placement rate of females in Mkt&Hr is **75.68%** which is greater than the placement rate of females in Mkt&Fin i.e. **51.28%**

Statistical Tests- Analysis

Randomized Block Design : To test the effect of work experience on the placement of MBA students controlling the variation from degree stream

Treatments: Work Experience (Yes or No)

Blocks: Degree Stream (Sci&Tech, Comm&Mgmt, Others)

Observations: Percentages for each combination is taken as the experiment obsevation.

	Degree Stream		
Work Exp	Comm&Mgmt	Sci&Tech	Others
Yes	0.866666667	0.88	0.75
No	0.63	0.558824	0.285714

Under the hypothesis that
$H_{\alpha 0}$: $\alpha_i = 0 \; \forall \; i$ (i.e. there is no effect due to the work experience)
$H_{\alpha 1}$: $atleast\ one\ \ \alpha_i \neq 0$

On the basis of the theory we get the ANOVA table as:

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Treatments	0.174124502	1	0.174124502	26.30515899	0.035976	18.51282
Blocks	0.063053468	2	0.031526734	4.762774563	0.173528	19
Error	0.01323881	2	0.006619405			
Total	0.25041678	5				

Hence , on the basis of p-value **as $p - value < 0.05$ for treatments so we can conclude that at 5% level of significance** the null hypothesis will be rejected i.e. there is significant difference between treatments , which implies that ***work experience has significant effect on placement***..

Independence Of Attribute tests:

For testing independence of 2 attributes each divided into 2 classes we form a 2*2 contingency table as:

		A		
		A1	A2	
B	B1	a	b	a+b
	B2	c	d	c+d
		a+c	b+d	N

Under the hypothesis of independence of attributes,

$$E(a) = \frac{(a+b)(a+c)}{N}$$

$$E(b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) = \frac{(a+c)(c+d)}{N}$$

$$E(d) = \frac{(b+d)(c+d)}{N}$$

Hence,

$$\chi^2 = \frac{[a - E(a)]^2}{E(a)} + \frac{[b - E(b)]^2}{E(b)} + \frac{[c - E(c)]^2}{E(c)} + \frac{[d - E(d)]^2}{E(d)}$$

And tabulated $\chi^2_{0.05}$ for (2-1)(2-1)=1 d.f. is 3.841.

Procedure:

Organize Data: Create a contingency table with rows and columns representing categories of two variables, and cell counts.

Calculate Expected Frequencies: Compute expected frequencies for each cell assuming independence of variables.

Calculate Chi-square Statistic: Sum the squared differences between observed and expected frequencies, divided by expected frequencies.

Determine Degrees of Freedom: Calculate degrees of freedom based on the number of rows and columns in the contingency table.

Perform Hypothesis Test: Compare calculated Chi-square statistic with critical Chi-square value at chosen significance level.

Interpret Results: Reject null hypothesis if calculated Chi-square statistic exceeds critical value, indicating significant association between variables.

Work experience and status

Hence for our data the contingency table with respect to work experience and status is:

		Status	
		Placed	Not placed
work exp	Yes	64	10
	No	84	57

On applying the formula we get our calculates $\chi^2 = 16.38496$. This implies that as calculated $\chi^2 >$ tabulated $\chi^2_{0.05}$ therefore we will reject H_0 i.e.***work experience enhances the chances of placement of a student (we used this insight to futher fit our logistic model).***

Gender and specialisation

Hence for our data the contingency table with respect to gender and specialisation is:

		Specialisation	
		Mkt&Fin	Mkt&HR
Gender	Male	83	56
	Female	37	39

On applying the formula we get our calculates $\chi^2 = 2.42302$. This implies that as calculated $\chi^2 <$ tabulated $\chi^2_{0.05}$ therefore we will accept H_0 i.e. *gender and specialisation are independent.*

Degree stream and specialisation

Hence for our data the contingency table with respect to gender and specialisation is:

		Degree		
		Comm&Mgmt	Sci&Tech	Others
Specialisation	Mkt&Fin	86	30	4
	Mkt&HR	59	29	7

On applying the formula we get our calculates $\chi^2 = 2.996252$. This implies that as calculated $\chi^2 <$ tabulated $\chi^2_{0.05}$ i.e. 5.991 , therefore we will accept H_0 i.e. *degree stream and specialisation are independent.*

Specialisation and placed

Hence for our data the contingency table with respect to specialisation and status is:

		Specialisation	
		Mkt&Fin	Mkt&HR
Status	Placed	95	53
	Not Placed	25	42

On applying the formula we get our calculates χ^2 =13.50801. This implies that as calculated χ^2 > tabulated $\chi^2_{0.05}$ therefore we will reject H_0 i.e. **there is some association between MBA specialisation and *Placement* *status that hints that maybe a particular specialisation enhances chances of Placement.***

Degree stream and placed

		Degree		
		Comm&Mgmt	Sci&Tech	Others
status	placed	102	41	5
	Not placed	43	18	6

On applying the formula we get our calculates $\chi^2 = 2.969043$. This implies that as calculated χ^2 < tabulated $\chi^2_{0.05}$ i.e. 5.991 , therefore we will accept H_0 i.e. ***there is no edge to the student of a specific degree on his/her placement.***

Degree Stream and 12th board stream

Hence for our data the contingency table with respect to specialisation and status is:

		Degree		
		Comm&Mgmt	Sci&Tech	Others
hsc	Science	29	56	6
	commerce	109	3	1
	arts	7	0	4

On applying the formula we get our calculates $\chi^2 = 123.4133$. This implies that as calculated χ^2 > tabulated $\chi^2_{0.05}$ i.e. 9.488 , therefore we will reject H_0 i.e.**stream of students in degree and 12th grade have some association in them.*Such as 109 out of 113 commerce students chose Comm&Mgmt as their degree option***

Regression Analysis:

1) Fitting of Logistic regression Model (taking academic performance as indicator variable)

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	215	100.0
	Missing Cases	0	.0
	Total	215	100.0
Unselected Cases		0	.0
Total		215	100.0

Dependent Variable Encoding

Original Value	Internal Value
Not placed	0
Placed	1

Block 0: Beginning block

Classification Table[a,b]

			Predicted		
			VAR00009		Percentage
	Observed		Not placed	Placed	Correct
Step 0	VAR00009	Not placed	0	67	.0
		Placed	0	148	100.0
	Overall Percentage				68.8

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.793	.147	28.968	1	.000	2.209

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Per_10	79.449	1	.000
		Per_12	51.881	1	.000
		Per_deg	49.507	1	.000
		Per_mba	1.272	1	.259
	Overall Statistics		106.920	4	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	138.362	4	.000
	Block	138.362	4	.000
	Model	138.362	4	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	128.409[a]	.475	.668

Calssification Table

			Predicted		
			VAR00009		Percentage
	Observed		Not placed	Placed	Correct
Step 1	VAR00009	Not placed	46	21	68.7
		Placed	13	135	91.2
	Overall Percentage				84.2

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Per_10	.178	.034	27.038	1	.000	1.195
	Per_12	.107	.032	11.174	1	.001	1.113
	Per_deg	.129	.043	9.015	1	.003	1.138
	Per_mba	-.190	.046	17.196	1	.000	.827
	Constant	-14.202	3.362	17.848	1	.000	.000

Interpretation:

The logistic model is given by:

logit(p)= -14.202 + 0.178X_1 +0.107X_2+0.129X_3-0.190X_4

where:

$$X_1 = 10th\ \%,\ X_2 = 12th\ \%, X_3 = Degree\ \%, X_4 = Mba\ \%\ of\ student$$

Nagelkerke R Square value of 66.8% tell us that our the model explains a large proportion of the variance in the outcome variable, indicating a better fit.

2) Fitting of Logistic regression Model (taking work_exp as indicator variable)

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	215	100.0
	Missing Cases	0	.0
	Total	215	100.0
Unselected Cases		0	.0
Total		215	100.0

Dependent Variable Encoding

Original Value	Internal Value
Not Placed	0
Placed	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
Work_exp	No Exp	141	1.000
	Exp	74	.000

Block 0: Beginning Block

Classification Table[a,b]

			Predicted Status		
	Observed		Not Placed	Placed	Percentage Correct
Step 0	Status	Not Placed	0	67	.0
		Placed	0	148	100.0
	Overall Percentage				68.8

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.793	.147	28.968	1	.000	2.209

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Per_10	79.449	1	.000
		Per_12	51.881	1	.000
		Per_deg	49.507	1	.000
		Per_mba	1.272	1	.259

	Work_exp(1)	16.385	1	.000
Overall Statistics		114.096	5	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	154.300	5	.000
	Block	154.300	5	.000
	Model	154.300	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	112.471[a]	.512	.720

Classification Table[a]

			Predicted Status		Percentage Correct
	Observed		Not Placed	Placed	
Step 1	Status	Not Placed	50	17	74.6
		Placed	10	138	93.2
	Overall Percentage				87.4

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Per_10	.192	.038	26.097	1	.000	1.212
	Per_12	.119	.035	11.664	1	.001	1.127
	Per_deg	.150	.048	9.859	1	.002	1.162
	Per_mba	-.228	.050	20.932	1	.000	.796
	Work_exp(1)	-2.267	.647	12.292	1	.000	.104
	Constant	-13.222	3.881	11.606	1	.001	.000

a. Variable(s) entered on step 1: Per_10, Per_12, Per_deg, Per_mba, Work_exp.

Interpretation:

The logistic model is given by:

logit(p)= -13.222 + 0.192X_1 + 0.119X_2 −0.150X_3 − 0.228X_4 − 2.267X_5

where:

$X_1 = 10th\ \%,\ X_2 = 12th\ \%, X_3 = Degree\ \%, X_4 = Mba\ \%\ of\ student$

$X_5 = Work\_\text{exp}\ (0\ or\ 1)$

Nagelkerke R Square value of 72% tell us that our the model explains a large proportion of the variance in the outcome variable, indicating a better fit,

Normality test (of distribution curves) :

Python Code:

```
from scipy.stats import kstest
# Perform Kolmogorov-Smirnov test
statistic, p_value = kstest(data['Degree %'], 'norm')
print("Kolmogorov-Smirnov Test:")
print("Statistic:", statistic)
print("P-value:", p_value)

# Interpret the results
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: Data is not normally distributed.")
else:
    print("Fail to reject the null hypothesis: Data is normally distributed.")
```

Result:

Kolmogorov-Smirnov Test:
Statistic: 1.0
P-value: 0.0
Reject the null hypothesis: Data is not normally distributed.

Interpretation:

Outliers in dataset as depicted by Box_plot resulted in regection of null hypothesis i.e our data is not normally distributed.

Application of sampling methods:

The section intends to show that the variance of the sample mean obtained through stratified sampling is smaller than the variance of the sample mean obtained through simple random sampling without replacement (SRSWOR). Also it shows that $\hat{y}_{st}$ is an unbiased estimate of population mean.

Sample mean of Stratified sample = $\hat{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i$

We have divided the population into two strata on the basis of specialisations

Total population size: 148

Size of stratum 1 (Mkt&Fin) N_1 = 95

Size of stratum 2 (Mkt&HR) N_2 =53

Sample size n = 40

$$V(\bar{y}_{st})_{Prop} = \frac{N-n}{nN} . \frac{\sum_{i=1}^{k} N_i S_i^2}{N}$$

$$V(\bar{y})_{SRSWOR} = \left(\frac{N-n}{nN}\right) S^2$$

Where $S^2 = \frac{1}{N-1}[\sum_i N_i \sigma_i^2 + \sum_i N_i \bar{Y}_i^2 - N * \bar{Y}^2]$

Calculation table:

Specialisation	N_i	$\bar{Y}_i$	σ_i	$N_i * \sigma_i^2$	$N_i * \bar{Y}_i^2$	$S_i^2 = (\frac{N}{N-1}) * \sigma_i^2$	$N_i * S_i^2$
Mkt&Fin	95	298852.6	107619.6	1.10029E+12	8.48473E+12	11660767250	1.10777E+12
Mkt&HR	53	270377.4	54264.98	1.56068E+11	3.87451E+12	2964719679	1.5713E+11
Total	148	569230	161884.6	1.25636E+12	**1.23592E+13**	14625486929	**1.2649E+12**

Using table we get,

$V(\bar{y}_{st})_{Prop}$	155918470.9
S^2	8792435835
$V(\bar{y})_{SRSWOR}$	169568405.4

Hence we have verified that,

$$V(\bar{y}_{st})_{Prop} \leq V(\bar{y})_{SRSWOR}$$

Also we have shown that $\hat{y}_{st}$ is an unbiased estimate of population mean.

Conclusion & Recommendations

The comprehensive analysis conducted in this study has provided valuable insights into various factors influencing the placement outcomes of MBA students. Key findings include the observation that MBA is a popular choice among commerce graduates, with notable trends such as a higher placement rate in Marketing and Finance specialization. Additionally, academic excellence has shown a strong correlation with placement success, as evidenced by the 100% placement rate among students scoring above the median percentage across academic levels. The logistic regression models fitted further confirm the significant impact of academic performance and work experience on placement outcomes. Furthermore, statistical tests such as the Randomized Block Design (RBD) and independence of attribute tests have contributed to our understanding of the factors affecting placement. The application of sampling techniques has verified the efficiency of different sampling procedures. Overall, these findings offer valuable insights for students, educators, and recruiters in the field of MBA placements.

Recommendations:

Based on the insights gained from this study, several recommendations can be made to MBA students to enhance their placement prospects. Firstly, prioritizing academic excellence across all levels of education, from secondary school to MBA, can significantly increase the likelihood of securing placements. Students should strive to maintain high academic performance, as evidenced by the 100% placement rate among those scoring above the median percentage. Additionally, while work experience is valuable, particularly in securing higher salary packages, it is not the sole determinant of placement success. Therefore, students should focus on acquiring relevant skills through internships, projects, and extracurricular activities to complement their academic achievements. Furthermore, students should consider the specialization they pursue carefully, taking into account placement rates and industry demand. The higher placement rate observed in Marketing and Finance specialization suggests promising career prospects in these fields

Limitations and Future Directions:

Despite the insightful findings obtained in this study, it is important to acknowledge certain limitations that may have influenced the results. Firstly, the reliance on a single dataset limits the generalizability of our findings to other MBA cohorts or institutions. Additionally, the retrospective nature of the data restricts our ability to establish causal relationships between variables. Methodological constraints such as the presence of outliers and the assumption of normality has restricted us from applying t-test and anova.

References and Link:

1. https://www.investopedia.com/terms/stratified_random_sampling.asp
2. https://study.com/learn/lesson/randomized-block-design-experimentexample.html#:~:text=Randomized%20Block%20Design%20Example,place%20of%20the%20real%20drug.
3. https://home.iitk.ac.in/~shalab/sampling/chapter4-sampling-stratified-sampling.pdf
4. *Fundamentals of Applied Statistics by S. C. Gupta, V. K. Kapoor*
5. *Fundamentals of Mathematical Statistics by S. C. Gupta, V. K.* Kapoor
6. **https://www.geeksforgeeks.org/ml-kolmogorov-smirnov-test/**
7. https://sites.education.miami.edu/statsu/wp-content/uploads/sites/4/2020/07/Logistic-Regression-Webinar.pdf
8. https://www.jmp.com/en_in/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html#:~:text=The%20Chi%2Dsquare%20test%20of%20independence%20checks%20whether%20two%20variables,idea%20is%20plausible%20or%20not.