BrainDead
Revelation '23

Team Name :
643311-UBW5377T

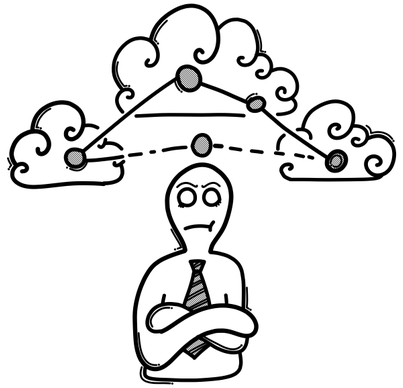
Click for Code

MANAV BALDEWA
21f1002723@ds.study.iitm.ac.in

RADHIKA JALAN
radhi.jalan@gmail.com

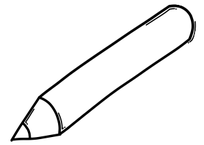
ROHAN BHATTACHARJEE
rohan.bhattacharjee.77@gmail.com

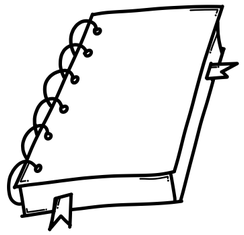
643311-UBW5377T



Problem Statement 1

1. Your main task is to analyze the factors that affect the placement and salary of students.





In this challenge, we are supposed to analyze the placement records of the students of a MBA college. The dataset includes secondary and higher secondary school percentages and specializations. It also contains degree specialization, work experience, and the salary offered to the students.

2. Analyze the dataset and derive meaningful insights from the data.

Dataset

643311-UBW5377T

Placement_Data_Full_Class.csv (19.71 kB)

Detail | Compact | Column

15 of 15 columns

sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisat...	mba_p	s
1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55	Mkt&HR	58.8	Pla
2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Pla
3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75	Mkt&Fin	57.8	Pla
4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66	Mkt&HR	59.43	Not
5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.5	Pla
6	M	55.00	Others	49.80	Others	Science	67.25	Sci&Tech	Yes	55	Mkt&Fin	51.58	Not
7	F	46.00	Others	49.20	Others	Commerce	79.00	Comm&Mgmt	No	74.28	Mkt&Fin	53.29	Not
8	M	82.00	Central	64.00	Central	Science	66.00	Sci&Tech	Yes	67	Mkt&Fin	62.14	Pla
9	M	73.00	Central	79.00	Central	Commerce	72.00	Comm&Mgmt	No	91.34	Mkt&Fin	61.29	Pla
10	M	58.00	Central	70.00	Central	Commerce	61.00	Comm&Mgmt	No	54	Mkt&Fin	52.21	Not
11	M	58.00	Central	61.00	Central	Commerce	60.00	Comm&Mgmt	Yes	62	Mkt&HR	60.85	Pla
12	M	69.60	Central	68.40	Central	Commerce	78.30	Comm&Mgmt	Yes	60	Mkt&Fin	63.7	Pla
13	F	47.00	Central	55.00	Others	Science	65.00	Comm&Mgmt	No	62	Mkt&HR	65.04	Not
14	F	77.00	Central	87.00	Central	Commerce	59.00	Comm&Mgmt	No	68	Mkt&Fin	68.63	Pla
15	M	62.00	Central	47.00	Central	Commerce	50.00	Comm&Mgmt	No	76	Mkt&HR	54.96	Not
16	F	65.00	Central	75.00	Central	Commerce	69.00	Comm&Mgmt	Yes	72	Mkt&Fin	64.66	Pla
17	M	63.00	Central	66.20	Central	Commerce	65.60	Comm&Mgmt	Yes	60	Mkt&Fin	62.54	Pla
18	F	55.00	Central	67.00	Central	Commerce	64.00	Comm&Mgmt	No	60	Mkt&Fin	67.28	Not
19	F	63.00	Central	66.00	Central	Commerce	64.00	Comm&Mgmt	No	68	Mkt&HR	64.08	Not
20	M	60.00	Others	67.00	Others	Arts	70.00	Comm&Mgmt	Yes	50.48	Mkt&Fin	77.89	Pla
21	M	62.00	Others	65.00	Others	Commerce	66.00	Comm&Mgmt	No	50	Mkt&HR	56.7	Pla
22	F	79.00	Others	76.00	Others	Commerce	85.00	Comm&Mgmt	No	95	Mkt&Fin	69.06	Pla
23	F	69.80	Others	60.80	Others	Science	72.23	Sci&Tech	No	55.53	Mkt&HR	68.81	Pla
24	F	77.40	Others	60.00	Others	Science	64.74	Sci&Tech	Yes	92	Mkt&Fin	63.62	Pla

The dataset has 215 rows and 15 columns. Each row represents a student and his/her/their corresponding data.

The columns and their description:

- sl_no: Serial Number
- gender: Gender- Male='M',Female='F'
- ssc_p: Secondary Education percentage- 10th Grade
- ssc_b: Board of Education- Central/ Others
- hsc_p: Higher Secondary Education percentage- 12th Grade
- hsc_b: Board of Education- Central/ Others
- hsc_s: Specialization in Higher Secondary Education
- degree_p: Degree Percentage
- degree_t: Under-Graduation(Degree type)- Field of degree education
- workex: Work Experience
- etest_p: Employability test percentage (conducted by the college)
- specialisation: Post Graduation(MBA)- Specialization
- mba_p: MBA percentage
- status: Status of placement- Placed/Not placed
- salary: Salary offered by corporate to candidates

643311-UBW5377T

Workflow

The problem statement has been approached with a thorough methodology and divided into actionable steps.

Does mba percentage matter in placement?

What are the factors affecting the placement of a student?

Action 1

Data Exploration through Excel and Python

Action 2

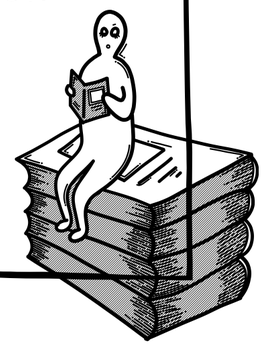
Cleaning of Data (Converting categorical values into numeric values and filling the empty spaces)

Action 3

Data Analysis using Excel and Python

Action 4

Data Visualization using Excel

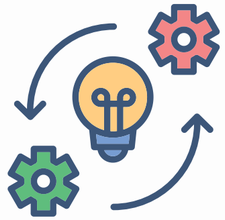


643311-UBW5377T

Data Exploration

Categorical Data

gender : {'M', 'F'}
ssc_b : {'Central','Others'}
hsc_b : {'Central','Others'}
hsc_s : {'Commerce','Arts','Science'}
degree_t:{'Sci&Tech','Comm&Mgmt','Others'}
workex : {'yes' , 'no'}
specialization : {'Mkt&Fin','Mkt&HR'}
status : {'placed', 'not placed'}



Numeric Data

ssc_p: Secondary Education percentage- 10th Grade
hsc_p: Higher Secondary Education percentage- 12th Grade
degree_p: Degree Percentage
etest_p: Employability test percentage
mba_p: MBA percentage
salary: Salary offered by corporate to candidates

Missing value

We have observed that the column "salary" has 67 missing values (for those who were not placed).

643311-UBW5377T

DATA CLEANING

Missing Values

Filling up the value of "salary" as 0 for the students who were not placed.

ACTION 2A

Making Compatible for Analysis

Converting categorical data into numerical data using LabelEncoder with the help of Python

ACTION 2B

Fragmenting Data

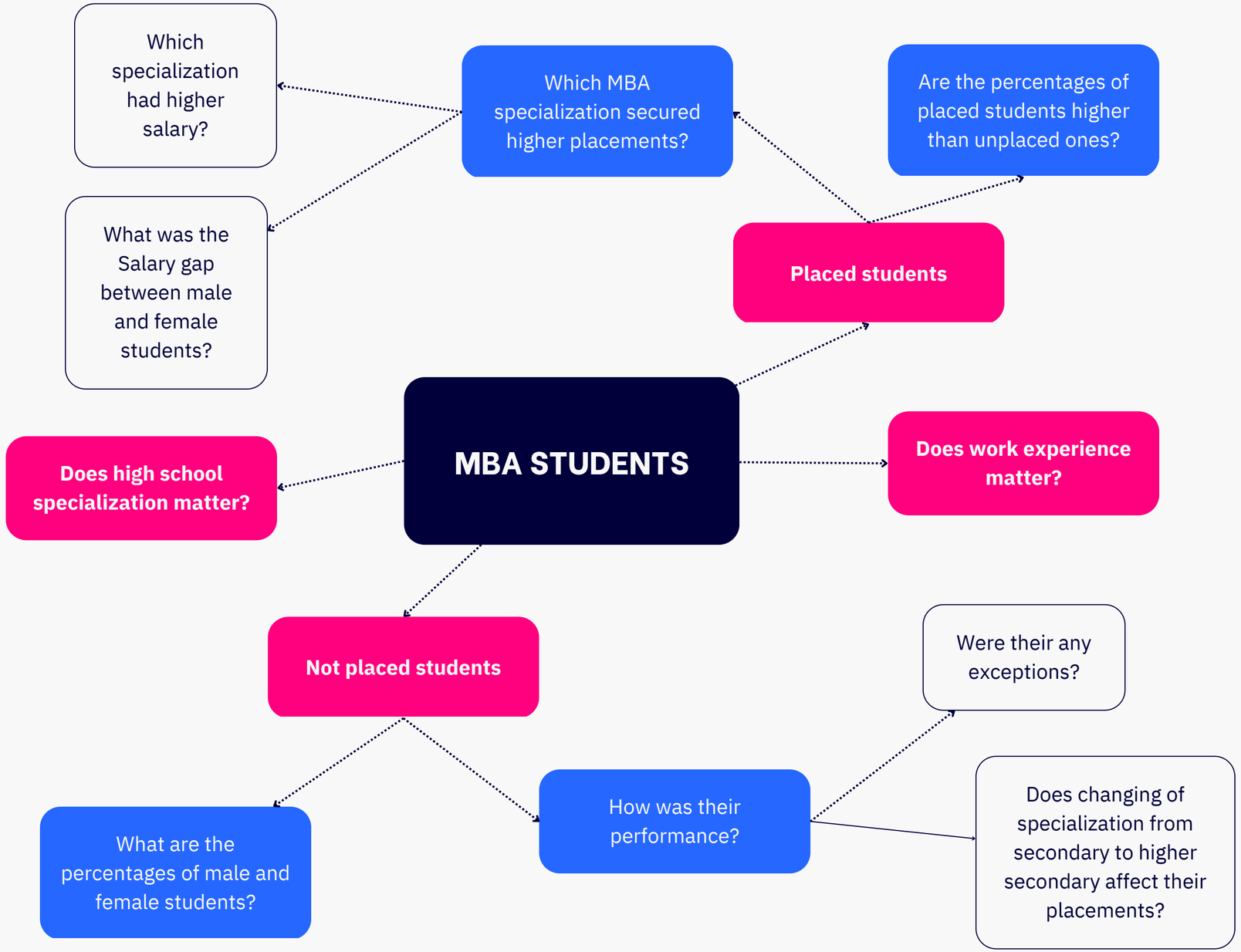
Grouping the data on the basis of the whether students got placed or not

ACTION 2C

643311-UBW5377T

Questions under Limelight

There can be many questions asked after fragmenting the dataset and we have listed a few of them through a mindmap.



643311-UBW5377T

Methods

1. Charts - Used to visualize the trends
2. Data Analysis Toolkit (Excel) - Used to extract the dependencies of features from one another
3. Pandas Dataframe - Used to read the dataset and to manipulate it
4. Label Encoders from SkLearn - Used to convert categorical data into numerical data
5. Pivot Table - Used to analyze summary statistics

643311-UBW5377T

RESULTS OF ANALYSIS

	gender_numeric	ssc_p	ssc_b_numeric	hsc_p	hsc_b_numeric	hsc_s_numeric	degree_p	degree_t_numeric	workex_numeric	etest_p	specialisation_numeric	mba_p	salary	status_numeric
gender_numeric	1													
ssc_p	-0.068968614	1												
ssc_b_numeric	0.019429113	0.116194113	1											
hsc_p	-0.021333906	0.511472102	-0.137012913	1										
hsc_b_numeric	0.065944733	0.066996092	0.605883321	-0.019548082	1									
hsc_s_numeric	0.071827029	0.236363763	0.050918616	-0.164090641	0.152227129	1								
degree_p	-0.173216618	0.538403999	0.038069911	0.434205806	0.067229329	0.137276347	1							
degree_t_numeric	0.061345136	0.2058961	0.100862783	-0.086449835	0.057959519	0.596299594	0.079316894	1						
workex_numeric	0.085152599	0.175675476	-0.040743924	0.141024862	0.038356689	0.007855772	0.122647554	0.105816191	1					
etest_p	0.084293518	0.261992695	-0.018990806	0.245112928	0.039108032	0.075642862	0.224470171	0.011508712	0.056734693	1				
specialisation_numeric	-0.106159607	-0.172536259	-0.051564759	-0.241629958	0.002232093	0.172106522	-0.218285935	0.084361417	-0.191173732	-0.236315103	1			
mba_p	-0.300531215	0.388477552	0.083120209	0.354822595	0.090201303	0.039344647	0.402363771	0.11666632	0.168811493	0.218054671	-0.105727827	1		
salary	0.143109966	0.538089713	0.034594007	0.452568776	0.011543843	0.058970368	0.408370781	0.053154812	0.298285188	0.186987685	-0.275766034	0.139822739	1	
status_numeric	0.090670398	0.607888734	0.037296512	0.491227944	0.016944537	0.033442144	0.479860993	-0.020352162	0.276059965	0.127639375	-0.25065509	0.076921649	0.865773711	1

Feature dependency with each other

We have used Excel's Data Analysis ToolKit to get the correlation table between difference features.

Applying conditional formatting

We have used conditional formatting to visually highlight important values.

What have we got?

We have got that 'Salary' and 'Status' depends on : 'ssc_p', 'hsc_p', 'degree_p' and 'work_ex'.

Any exceptions observed?

We have observed that 'mba_p do not depend on 'Salary' and 'Status'.

Data visualization

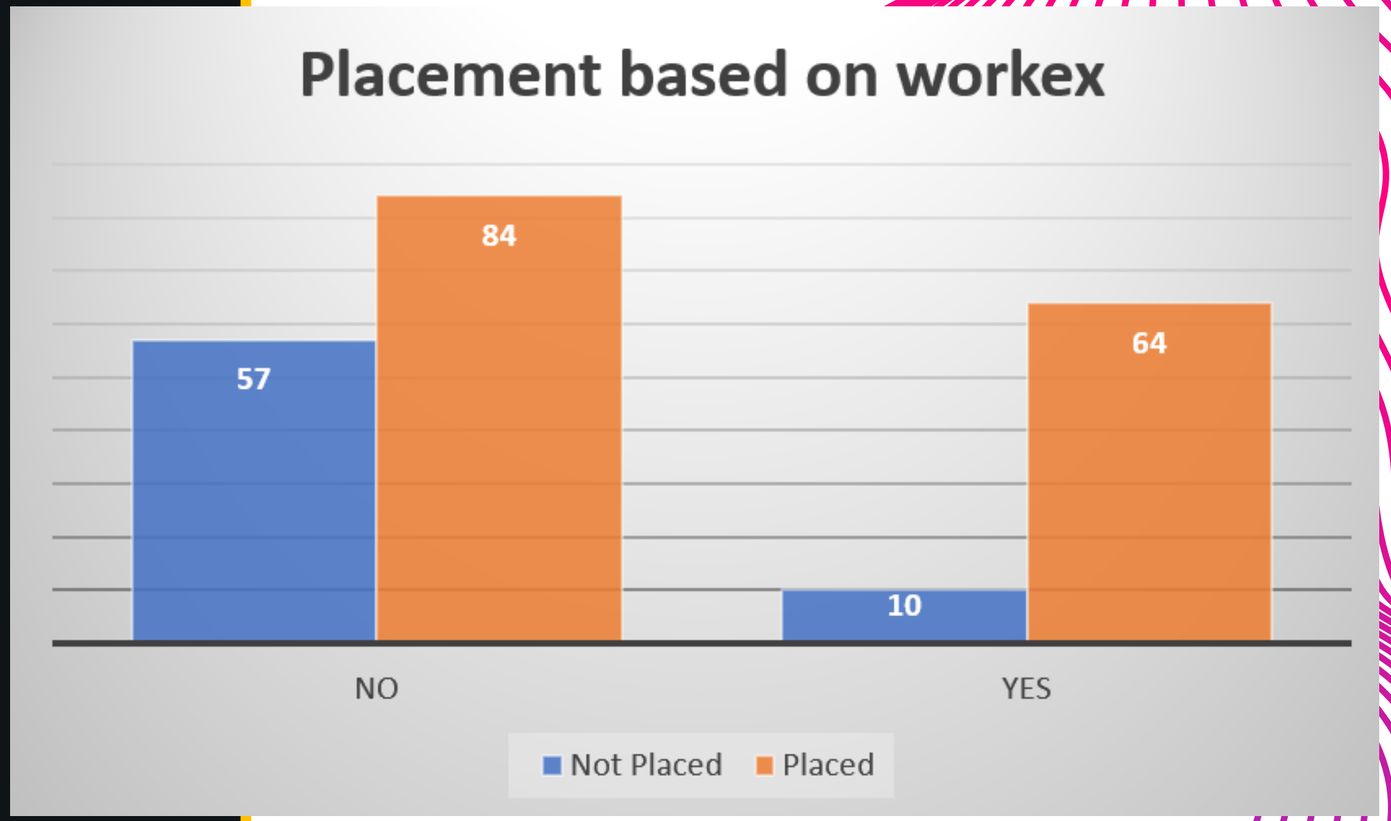
We have used excel to generate the chart and also used pivot table to summarize the statistics between 'wokex' and 'Status'.

What have we got?

We have got that 66% of the students were not work experienced and 34% students had work experience.

Any interesting facts?

We observed that among the 34% students who had work experience, 84% students were placed. While on the other hand only 60% were placed among the 66% students who did not have work experience.



Data visualization

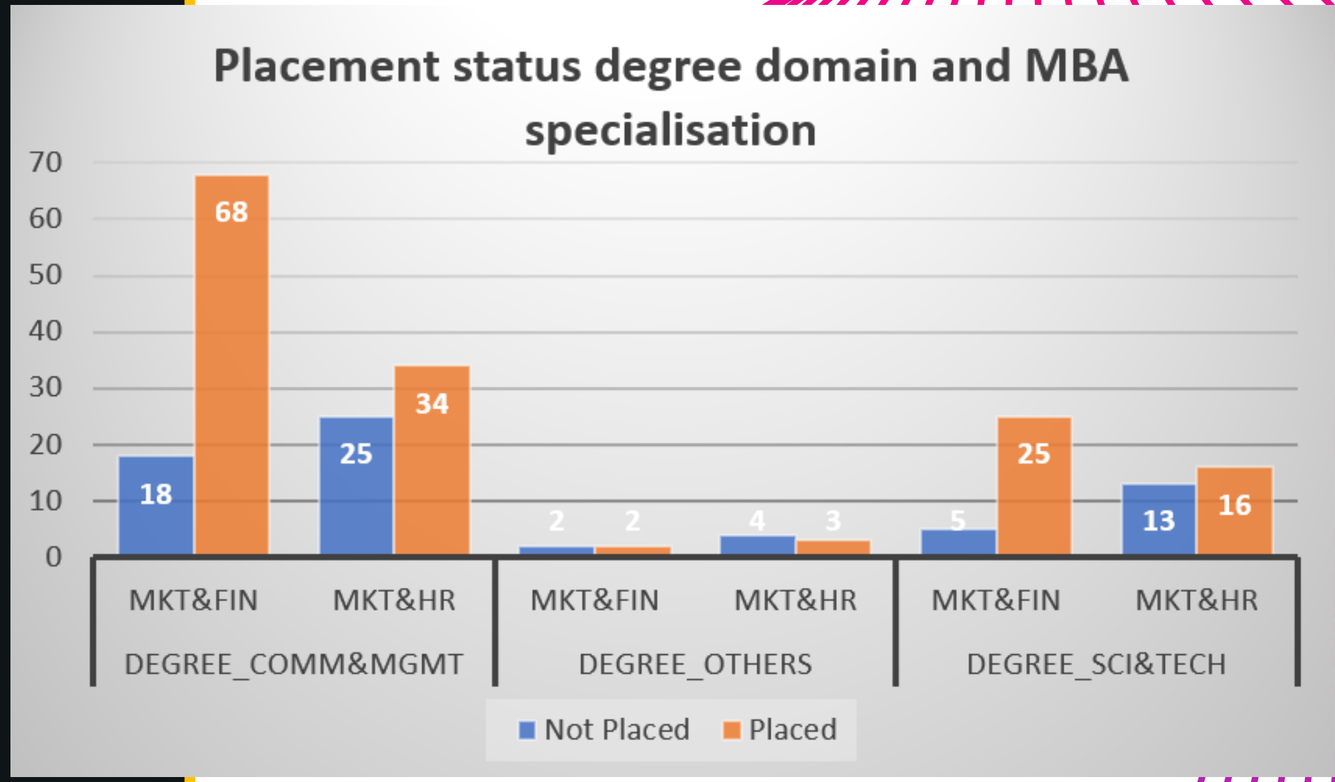
We have used excel to generate the chart and also pivot table to summarize the statistics between 'degree_t', 'specialisation', and 'Stuatus'.

What have we got?

There were more students from the commerce and management background. And there were more number of placed students in the Mkt&Fin specialisation.

Exceptions

Those students who were from 'degree_others' had poor placement drive.



Data Visualization

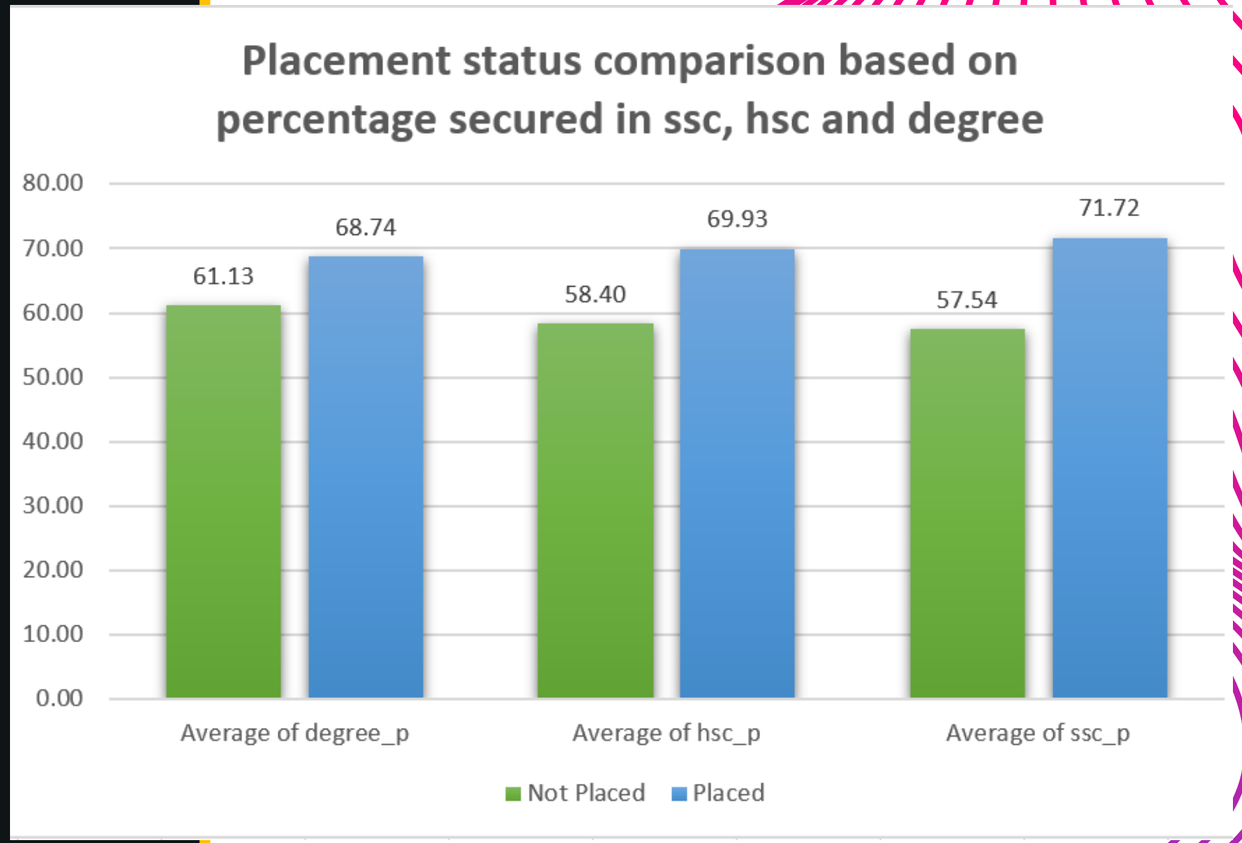
We have used excel to generate the chart and also pivot table to summarize the statistics between 'degree_p', 'hsc_p', 'ssc_p' and 'Status'

What have we got?

We have got that the average percentages of the students in degree, secondary school and higher secondary school is greater than the average percentages of the students who were not placed.

Any interesting facts?

Those who have scored in secondary school and higher secondary school have also performed well in undergraduate.



Data Visualization

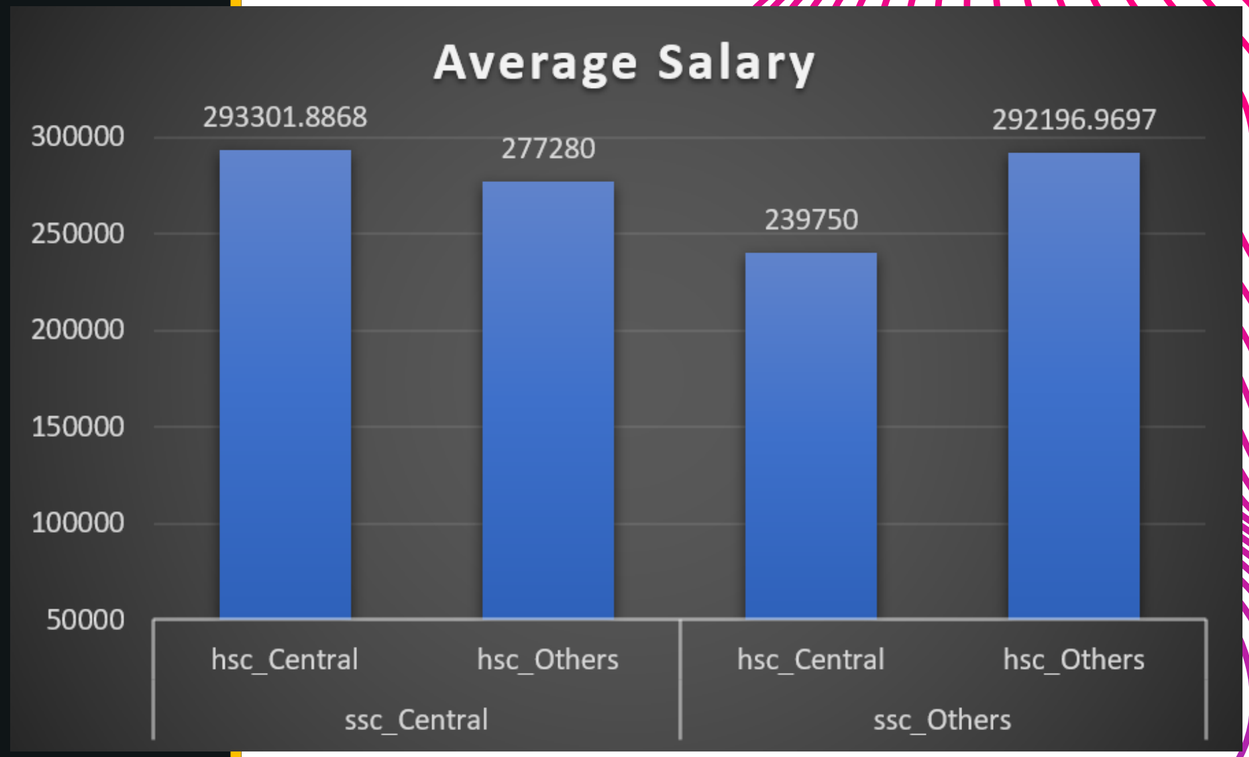
We have used excel to generate the chart and also pivot table to summarize the statistics between 'hsc_b', 'ssc_b' and 'Salary'.

What have we got?

Average salary for MBA placement drive is approximately 277,132.

Any interesting facts?

The students who has not changed there branch when transitioning from secondary school to higher secondary school have got more salary offered than those who have changed their branch.



Data Visualization

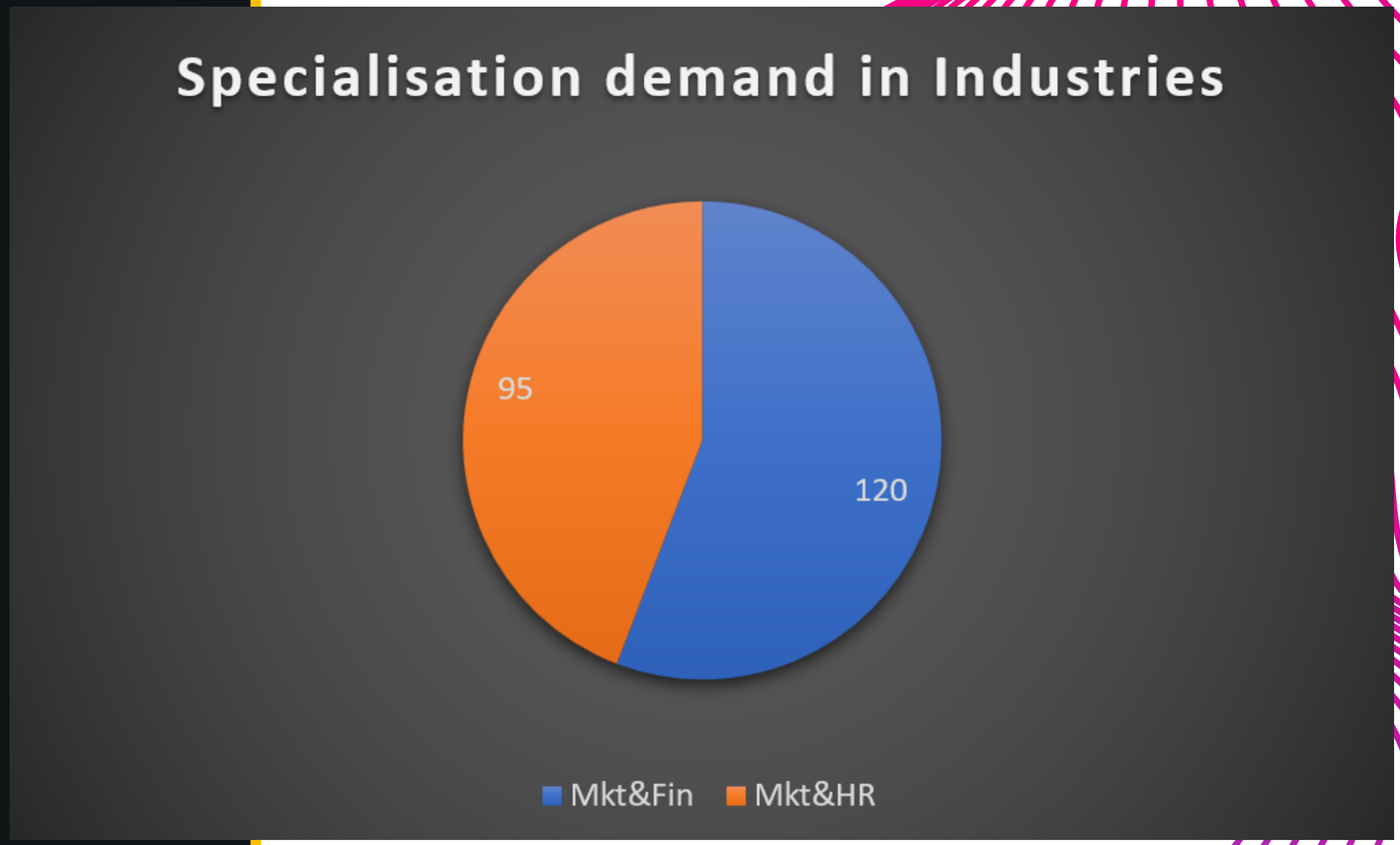
We have used excel to generate the chart and also pivot table to summarize the statistics between 'specialisation' and 'Status'.

What have we got?

We got that Marketing and Finance specialisation had a higher demand than that of Marketing and HR.

Any interesting facts?

No.



"pip install Insights"

We can observe a salary gap between males and females :

```
avg_man = males_placed['salary'].mean()
avg_woman = females_placed['salary'].mean()
print(f'Average salary of males placed : {avg_man}')
print(f'Average salary of females placed : {avg_woman}')
man_perc_more_than_woman = round(((avg_man-avg_woman)/avg_woman)*100,2)
print(f'Average Salary of makes is {man_perc_more_than_woman}% more than average salary of females')
```

```
Average salary of males placed : 298910.0
Average salary of females placed : 267291.6666667
Average Salary of makes is 11.83% more than average salary of females
```

We observe that work experience influences the salary of the students placed :

```
print (f'Average salary of the students placed having work experience: {average_workex_salary}')
print (f'Average salary of the students placed having no work experience: {average_no_workex_salary}')
perc = round (((average_workex_salary-average_no_workex_salary)/average_no_workex_salary)*100, 2)
print (f'Average salary of the students having work experience is {perc} % more than that of students having no work experience.')
```

```
Average salary of the students placed having work experience : 303265.625
Average salary of the students placed having no work experience : 277523.8095238095
Average salary of the students having work experience is 9.28% more than that of students having no work experience.
```

MBA specialization with Mkt&Fin earn 10% more than that of Mkt&HR specialization :

```
avg_hr_salary = HR['salary'].mean()
avg_fin_salary = Fin['salary'].mean()
perc = round(((avg_fin_salary-avg_hr_salary)/avg_hr_salary)*100, 2)
print (f'Average salary of students with Mkt&HR domain : {avg_hr_salary}')
print (f'Average salary of students with Mkt&Fin domain : {avg_fin_salary}')
print (f'Average salary of students having Mkt&Fin domain is {perc} % more than that of students having Mkt&HR domain.')
```

```
Average salary of students with Mkt&HR domain : 270377.358490566
Average salary of students with Mkt&Fin domain : 298852.63157894736
Average salary of students having Mkt&Fin domain is 10.53% more than that of students having Mkt&HR domain.
```

We observe that Average MBA Percentage of placed and not-placed students are approximately same so it does not affect in placement drives :

```
average_percentage_of_students_placed = MBA_placed ['mba_p'].mean()
average_percentage_of_students_unplaced = MBA_unplaced ['mba_p'].mean()
print (f'Average MBA percentage of students who were not placed : {average_percentage_of_students_unplaced}')
```

```
Average MBA percentage of students who got placed : 62.579391891891866
Average MBA percentage of students who were not placed : 61.61283582089551
```

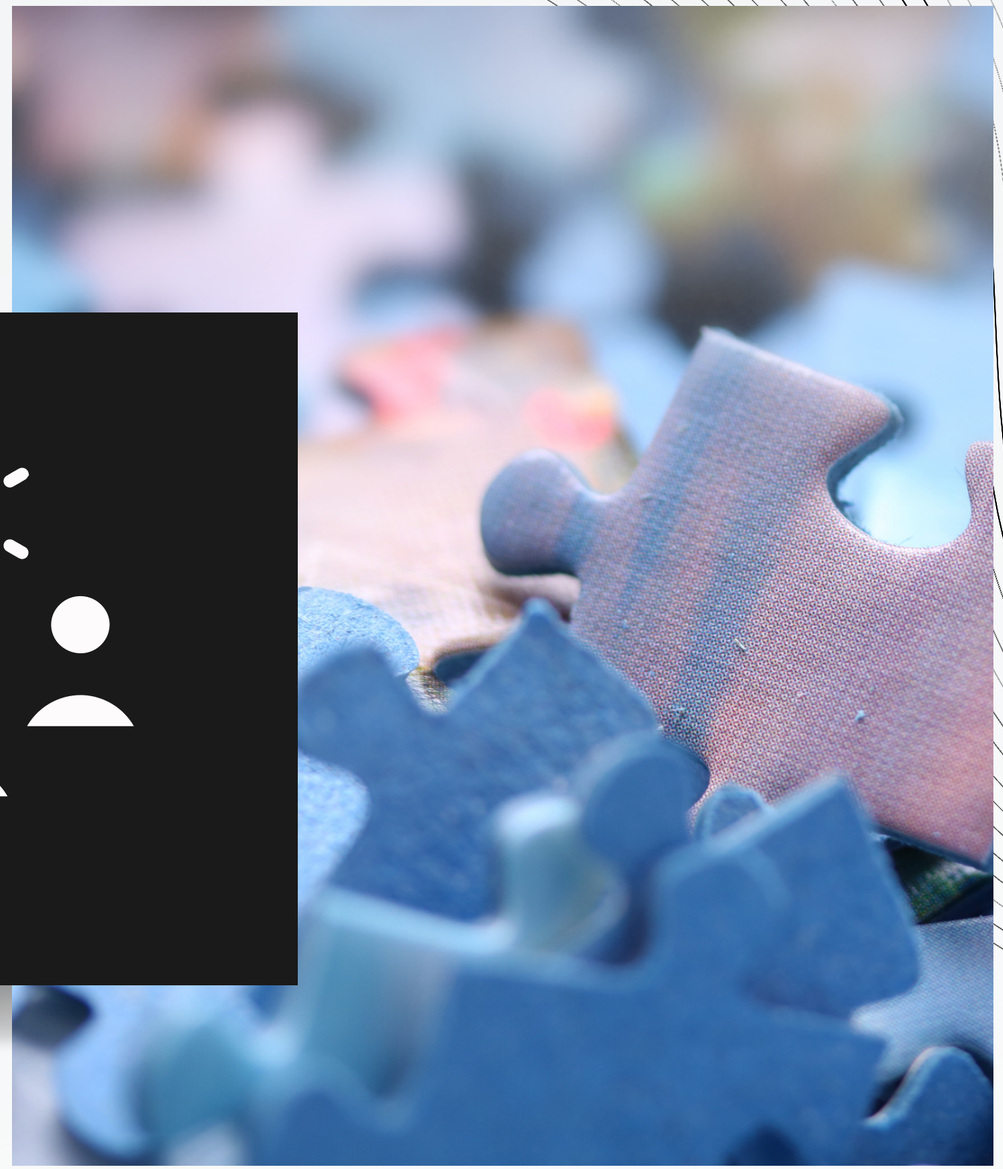
DISCUSSION

We adopted a naive approach in going through the various phases of Data Analysis and have provided some of the trends, patterns and insights in the form of visualization

Future work can be to focus on expanding the array of technologies used to provide further concrete trends in data.

643311-UBW5377T

PROBLEM STATEMENT 2



Social media platforms are widely used by individuals and organizations to express emotions, opinions, and ideas. These platforms generate vast amounts of data, which can be analyzed to gain insights into user behavior, preferences, and sentiment. Accurately classifying the sentiment of social media posts can provide valuable insights for businesses, individuals, and organizations to make informed decisions.

To accomplish this task, a customized private cartoon dataset (original images) of social media posts has been provided, which contains labels for each post's emotion category, such as happy, angry, sad, or neutral.

The task is to build and fine-tune a machine-learning model that accurately classifies social media posts into their corresponding emotion categories, using synthetic images.

643311-UBW5377T

Workflow

The problem statement has been approached with a thorough methodology and divided into actionable steps.

Action 1

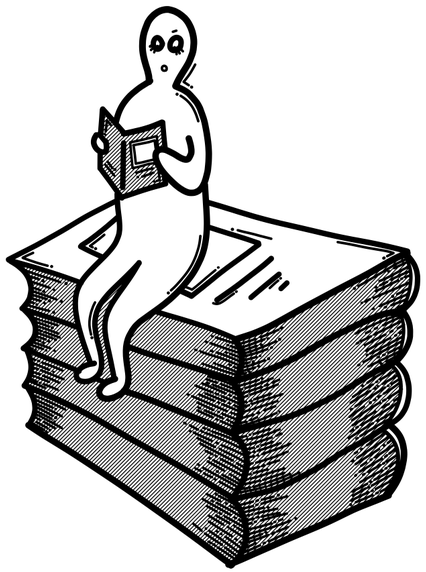
Synthetic Image Generation

Action 2

Building Schema for a Deep Learning Model

Action 3

Training the model on original and synthetic images



643311-UBW5377T

Synthetic Image Generation

Choosing an algorithm

Choosing an algorithm is crucial for the task.

- Initially, we went with GANs however due to lack of resources on free tier of Google Colab we switched to Stable Diffusion algorithm.

Getting Model

Stable Diffusion is a deep learning model efficient in text-to-image generation.

Dreambooth

It is a deep learning model that finetunes existing image generation algorithms.

We instantiate with the training and testing images and create a session

Model Training

The model is trained based on the original dataset

Testing

Dreambooth allows a web UI that helps us to generate further images in batches and create the synthetic data.

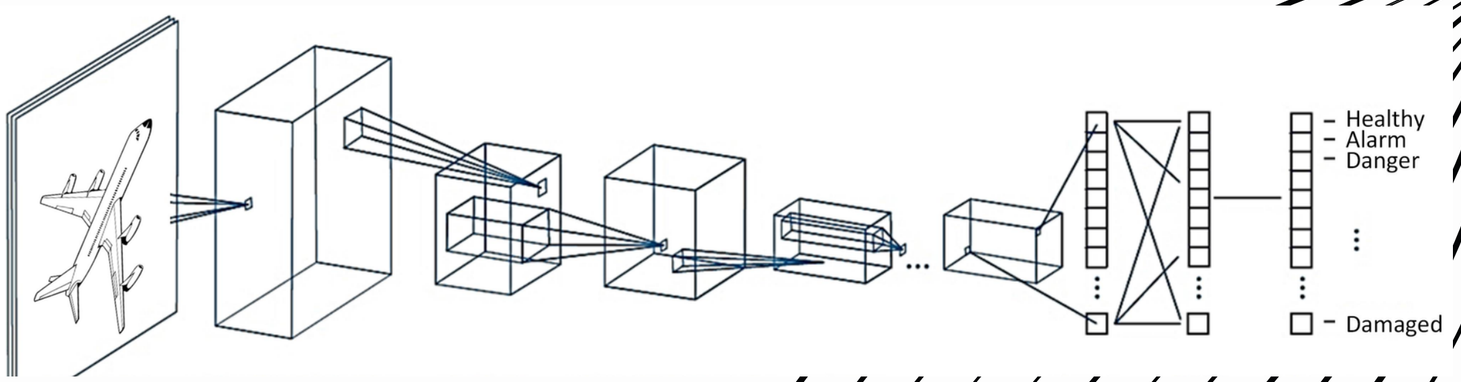
Github

643311-UBW5377T

Model Structure

We followed a naive approach in building our image classification model. We worked on a simple model with two Conv2D layers at first and evaluated it. To cope up with increasing variations in data, we expanded the layers into blocks of Conv2D, BatchNorm2D, ReLU and MaxPool2D followed by two Fully Connected Layers (fcv).

However we could not expand more layers considering the further complexities and fear of overfitting.



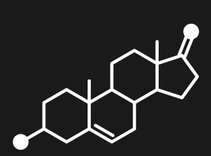

```
ConvNet(
  (conv1): Conv2d(3, 16, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (bn1): BatchNorm2d(16, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (relu1): ReLU()
  (pool1): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (conv2): Conv2d(16, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (bn2): BatchNorm2d(32, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (relu2): ReLU()
  (pool2): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (conv3): Conv2d(32, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (bn3): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (relu3): ReLU()
  (pool3): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (conv4): Conv2d(64, 128, kernel_size=(1, 1), stride=(1, 1), padding=(1, 1))
  (bn4): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (relu4): ReLU()
  (pool4): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  (fc): Linear(in_features=36992, out_features=256, bias=True)
  (fc1): Linear(in_features=256, out_features=4, bias=True)
)
```

643311-UBW5377T

metrics.io

Employing the usage of Original Images only in model training, we obtained a maximum accuracy of :

57.45%



Employing the usage of Synthetic Image only in model training, we obtained a maximum accuracy of :

59.09%

Employing the usage of both Original and Synthetic images, we obtained a maximum accuracy of:

71.86%

DISCUSSION

We built a very naive model with respect to the problem statement and could not achieve the threshold accuracy, however we have observed and thus proved that Synthetic data do play an important role in model training and evaluation.
We achieved a 25% increase in accuracy by employing synthetic data with original data.

Future works can be considered with perfecting the employed model for image classification and to work on achieving better metric values.

25%

643311-UBW5377T

Thats a wrap!

Thank you for your time

643311-UBW5377T