Applied Statistics

-Project Presentation

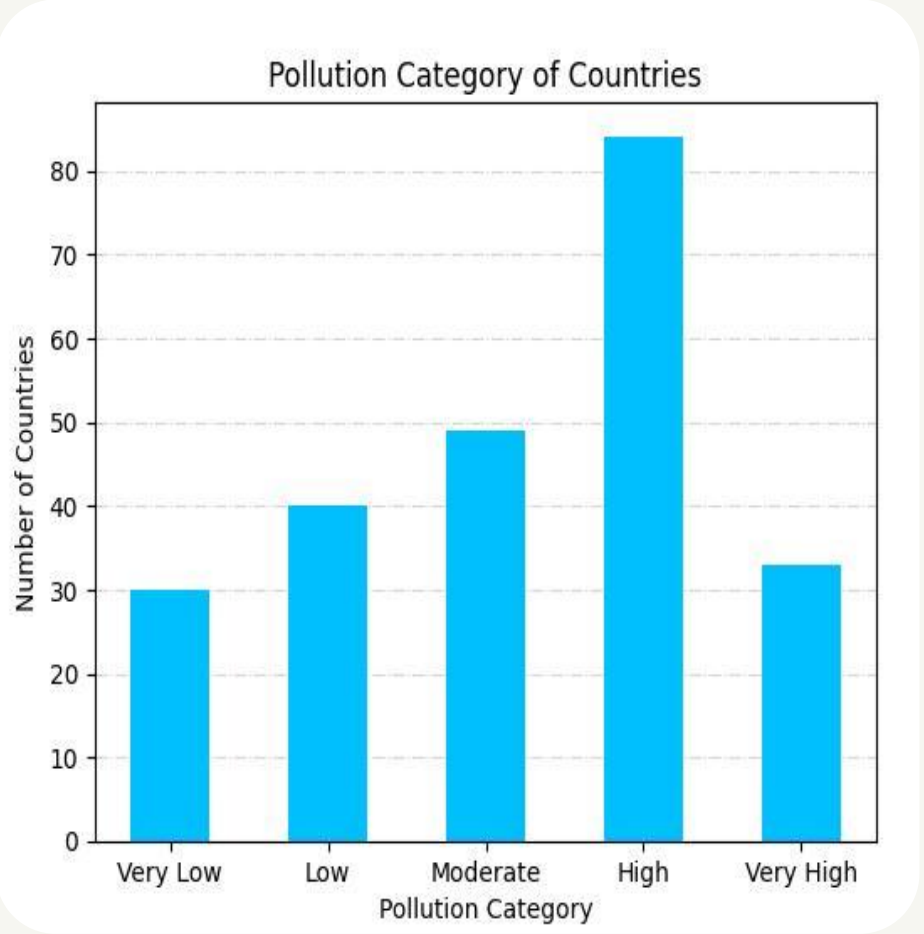
Project - 1 (Data Set Used is given below)

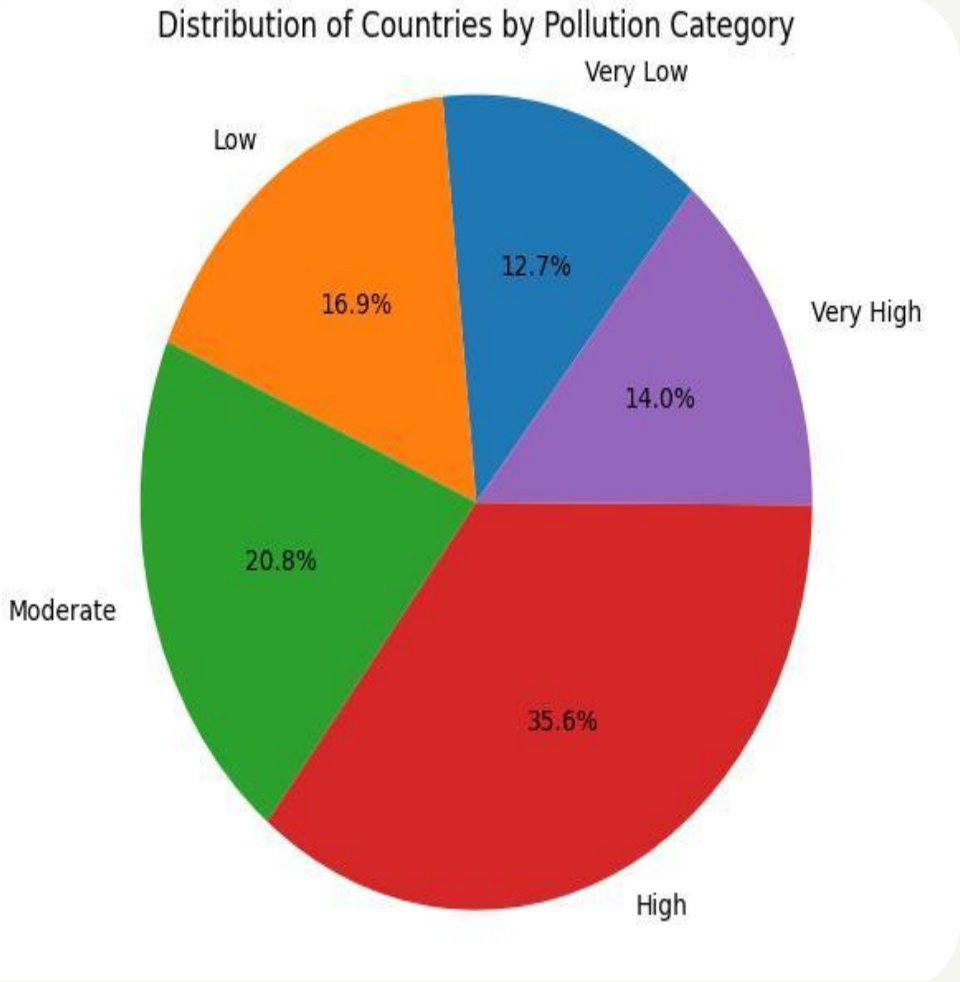
country	Purchasing Power Value	Purchasing Power Category	Safety Value	Safety Category	Health Care Value	Health Care Category	Cost of Liv
Afghanistan	32.15	Very Low	25.33	Low	24.24	Low	
Aland Islands	125.01	Very High	71.81	High	79.72	High	
Albania	42.82	Low	55.52	Moderate	48.21	Moderate	
Alderney	0	Very Low	83.79	Very High	100	Very High	
Algeria	27.6	Very Low	47.54	Moderate	54.43	Moderate	
American Samoa	0	Very Low	54.41	Moderate	0	Very Low	
Andorra	121.14	Very High	84.71	Very High	75.56	High	
Angola	224.46	Very High	33.71	Low	36.58	Low	
Anguilla	0	Very Low	75.44	High	77.08	High	
Antigua And Barbuda	0	Very Low	42.86	Moderate	30.56	Low	
Argentina	40.36	Low	36.36	Low	68	High	
Armenia	36.91	Very Low	77.81	High	58.07	Moderate	
Aruba	71.46	Moderate	69.23	High	78.7	High	
Australia	137.58	Very High	52.71	Moderate	73.35	High	
Austria	109.08	High	70.38	High	77.83	High	
Azerbaijan	41.75	Low	68.12	High	48.31	Moderate	
Bahamas	58.59	Low	43.76	Moderate	39.52	Low	
Bahrain	114.56	Very High	75.28	High	66.6	High	
Bangladesh	36.16	Very Low	38.79	Low	42.2	Moderate	
Barbados	46.93	Low	55.14	Moderate	71.88	High	
Belarus	56.24	Low	49.99	Moderate	48.5	Moderate	
Belgium	118.62	Very High	50.42	Moderate	75.76	High	
Belize	0	Very Low	47.47	Moderate	47.37	Moderate	
Benin	0	Very Low	62.16	High	51.39	Moderate	

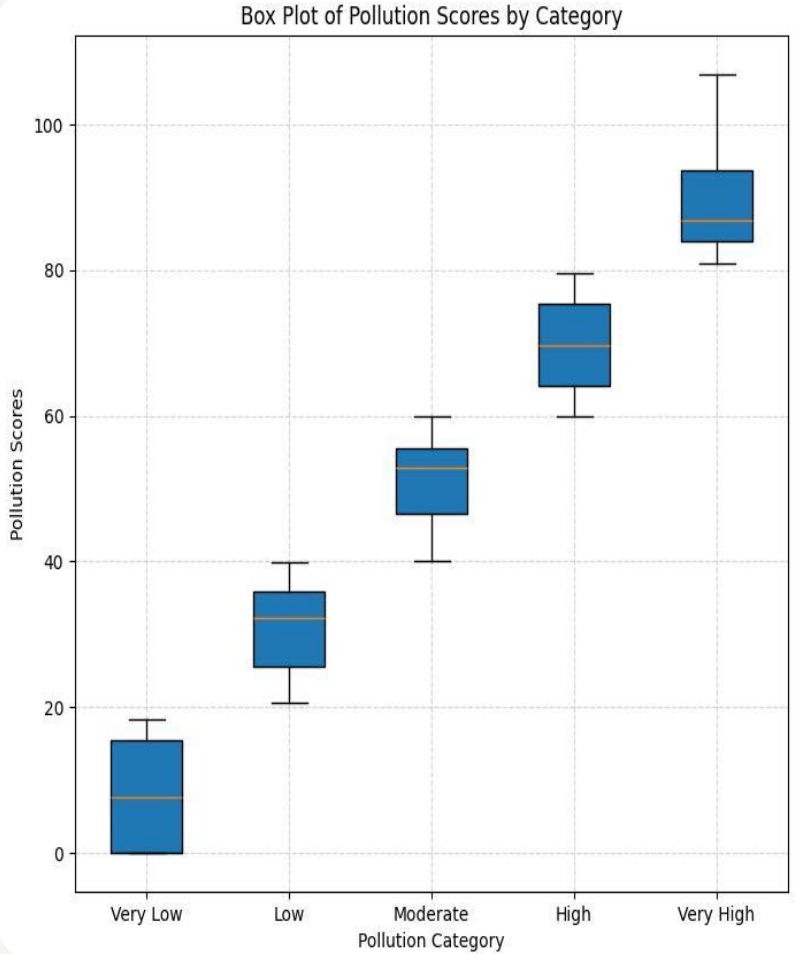
The Plots

Graphs of Pollution Category of Countries:

- This bar graph shows the distribution of countries across different pollution categories.
- We can see that a significant number of countries experience high levels of pollution.
- The pie chart and box plot of the same are shown in the next slide.




Distribution of Countries by Pollution Category
Very Low
Low
Very High
Moderate
High
12.7%
16.9%
14.0%
20.8%
35.6%


Box Plot of Pollution Scores by Category
Pollution Scores
0
20
40
60
80
100
Very Low
Low
Moderate
High
Very High
Pollution Category

Measures of Central Tendency

Africa vs. Europe

- Average Safety Value of Africa is less than that of Europe indicating more safety of European countries than African countries
- Even Median Safety value for Europe is more that of Africa indicating that the Safety standards of more than Half of Europe is higher than African countries.

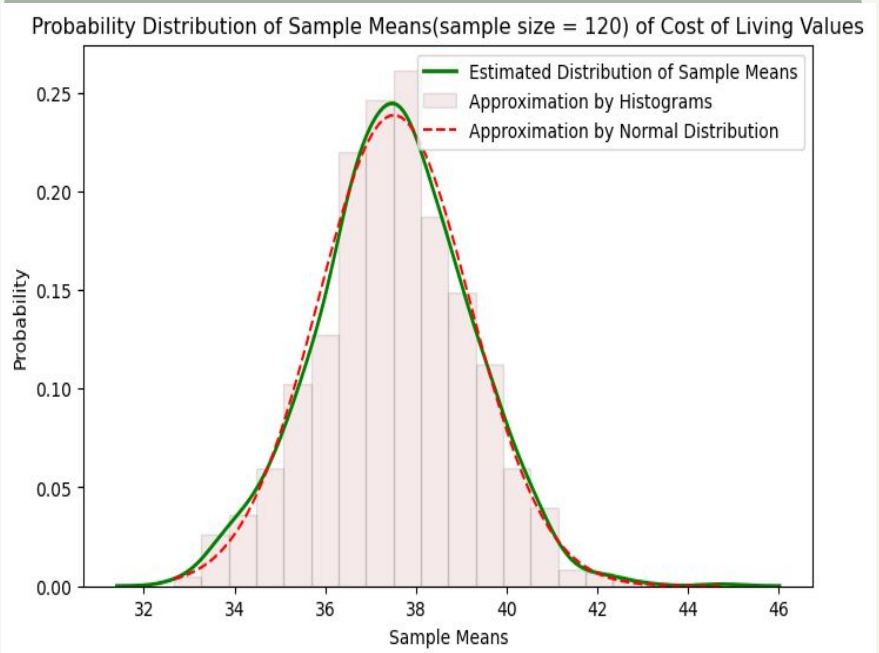
```
 Africa's Safety Values
The average is 45.737
The median is 45.795
Standard Dev is 12.893
Mean dev is 9.744
1st Quartile is 36.850
3rd Quartile is 52.935
```

```
 Europe's Safety Values
The average is 66.032
The median is 67.050
Standard Dev is 10.687
Mean dev is 8.978
1st Quartile is 57.695
3rd Quartile is 74.035
```

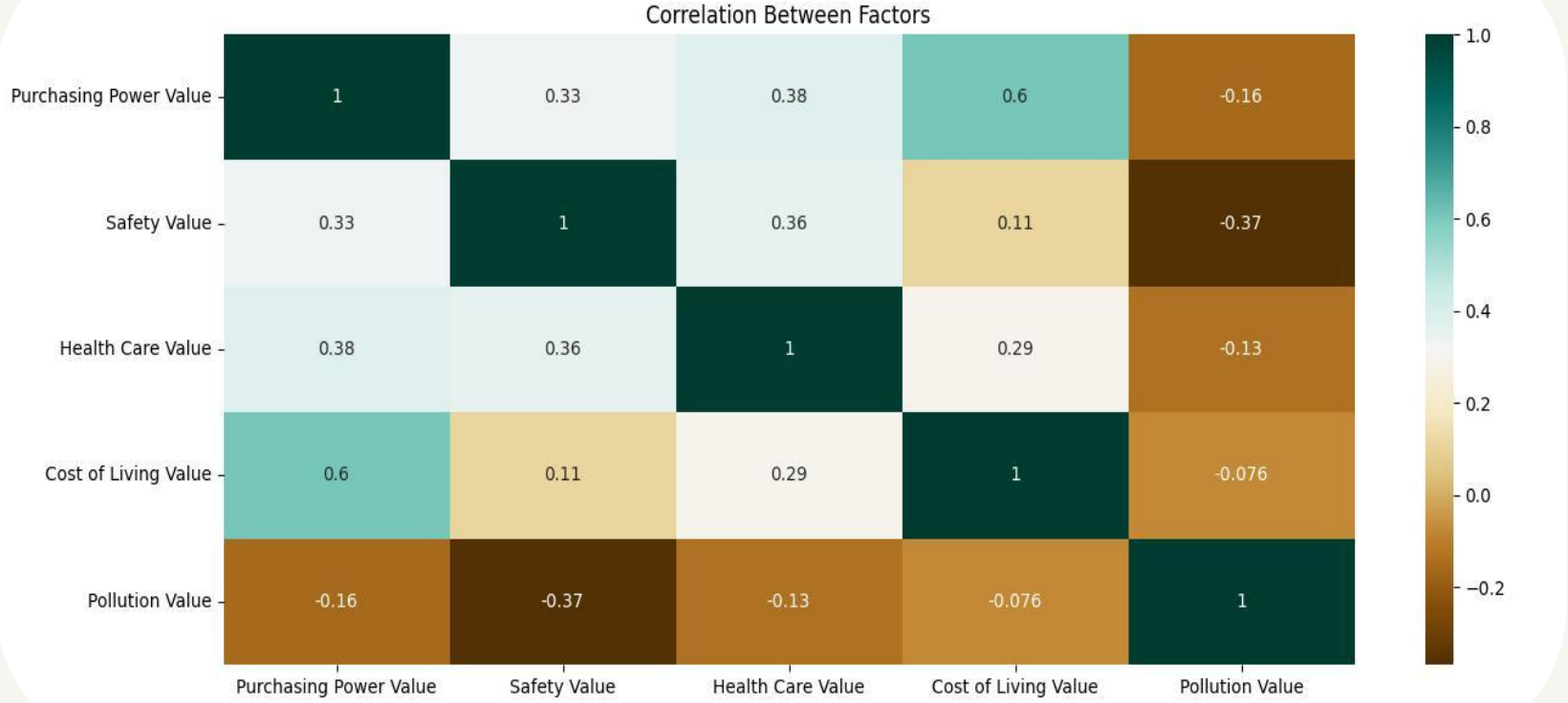
Central Limit Theorem

Taking the Cost of Living as random variable

- For this plot we are considering the sample data from the all over the countries.
- If we take the random sample n times of size 120, then the data can be approximated as normal dist., with mean around 37.



The Correlation Matrix



Project - 2 (Data Set Used is given below)

Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4	4.5	8.87	1	0.76
3	316	104	3	3	3.5	8	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.8
5	314	103	2	2	3	8.21	0	0.65
6	330	115	5	4.5	3	9.34	1	0.9
7	321	109	3	3	4	8.2	1	0.75
8	308	101	2	3	4	7.9	0	0.68
9	302	102	1	2	1.5	8	0	0.5
10	323	108	3	3.5	3	8.6	0	0.45
11	325	106	3	3.5	4	8.4	1	0.52
12	327	111	4	4	4.5	9	1	0.84
13	328	112	4	4	4.5	9.1	1	0.78
14	307	109	3	4	3	8	1	0.62
15	311	104	3	3.5	2	8.2	1	0.61
16	314	105	3	3.5	2.5	8.3	0	0.54

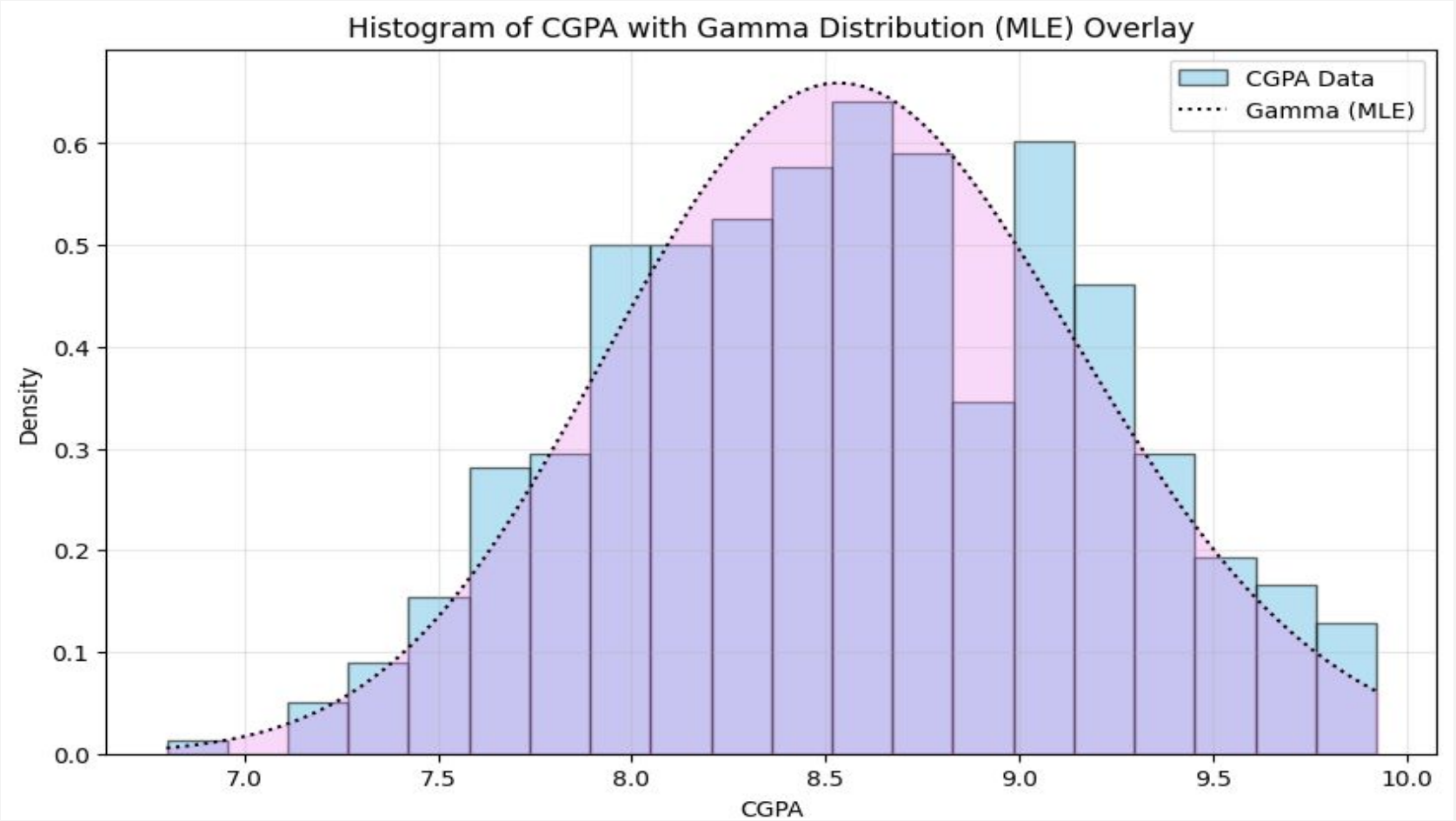
Estimating parameters for Gamma(a,b) for CGPA

- MoM estimator for a & b :

$$a_{\text{mom}} = \frac{(\bar{X})^2}{S^2} = 201.082 \qquad b_{\text{mom}} = \frac{\bar{X}}{S^2} = 23.446$$

- MLE estimator for a & b :

$$a_{\text{mle}} = \frac{\log(a_{\text{old}}) - \psi(a_{\text{old}}) - \log(\bar{X}) + \overline{\log X}}{\frac{1}{a_{\text{old}}} - \psi^{(1)}(a_{\text{old}})} = 200.169 \qquad b_{\text{mle}} = \frac{a_{\text{mle}}}{\bar{X}} = 23.339$$


Histogram of CGPA with Gamma Distribution (MLE) Overlay
CGPA Data
Gamma (MLE)
Density
CGPA
0.0
0.1
0.2
0.3
0.4
0.5
0.6
7.0
7.5
8.0
8.5
9.0
9.5
10.0

CI for Difference of means of GRE & TOEFL scores

where $\bar{x_1}, \bar{x_2}$ is sample mean of TOEFL scores and GRE scores respectively

$$\text{CI} = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}, n+m-2} \cdot s_p \sqrt{\frac{1}{n} + \frac{1}{m}} = (-210.41, -208.15)$$

$$\text{where } s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

Test Hypothesis for the H_0: p <= 1/2 vs H_1: p > 1/2

- For this Test of Hypothesis I am considering the data of the Research column. It contains the binary values {0,1} which are required for the bernoulli distribution.
- As the data contains 500 tuples (>30 condition is satisfied for CLT) we are considering the mean(x) of all the data. As it follows **N** (ρ , √(ρ(1-ρ)/n)) .
- Then, reject the null hypothesis if x > Z**α** * var + 0.5
- Else accept the null hypothesis
- To compare for the max case we will take the ρ = 0.5 (one -sided test) . And Level of significance as **α** = 0.05 .

- From the data which I have taken from (Research column) the "null hypothesis got Rejected ".
- This result simply says that the bernoulli distribution with probability (ρ) mostly satisfied with the condition ρ>0.5 .