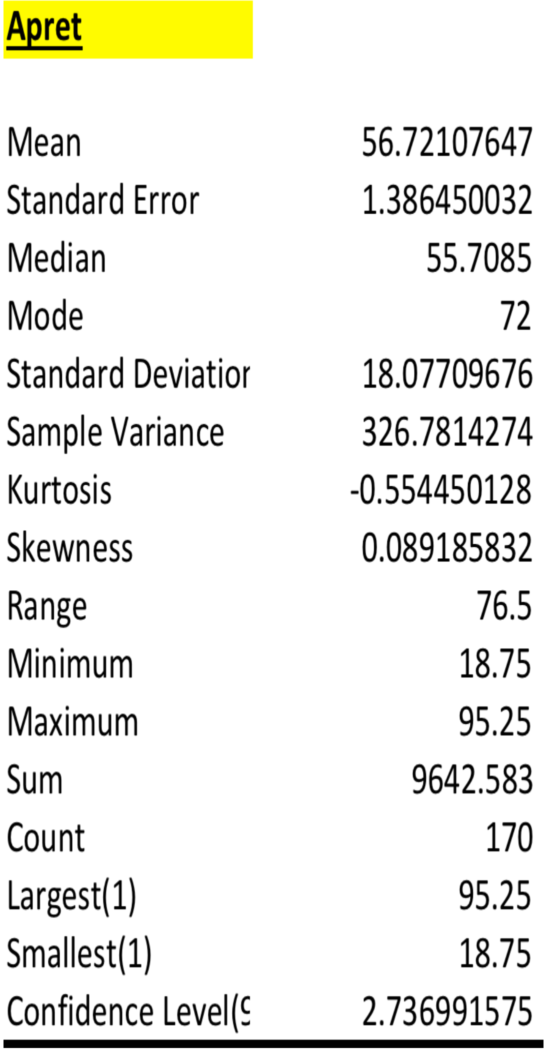
Data analytics Assignment-2

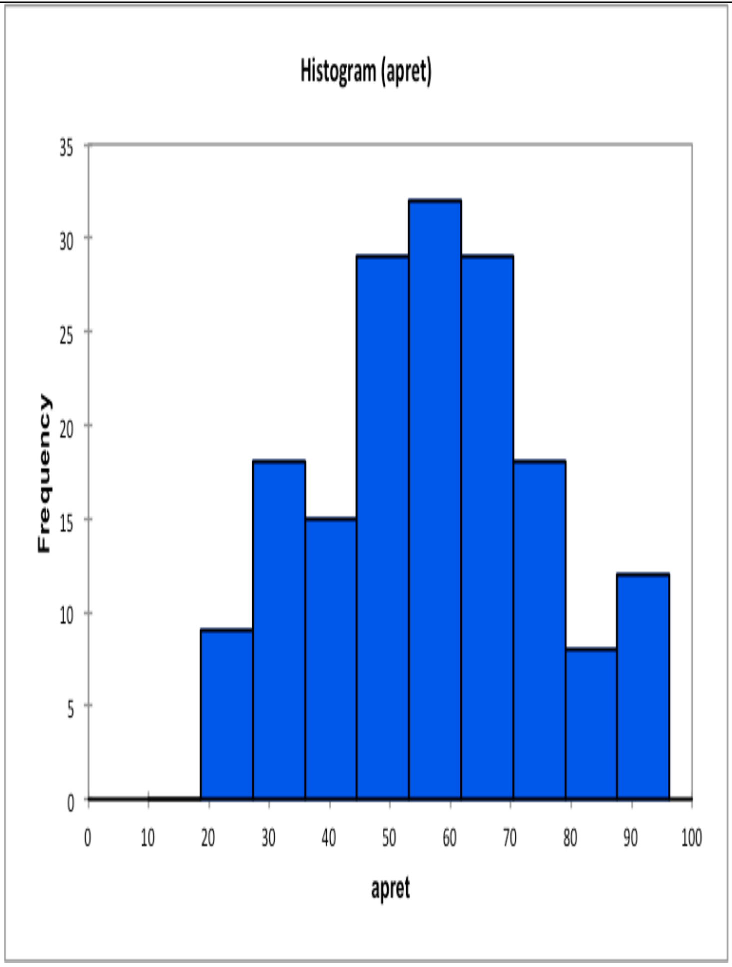
#1 Generate descriptive statistics and plot histograms for the following three columns: apret, tstsc, and salar.

Descriptive statistics

Histogram

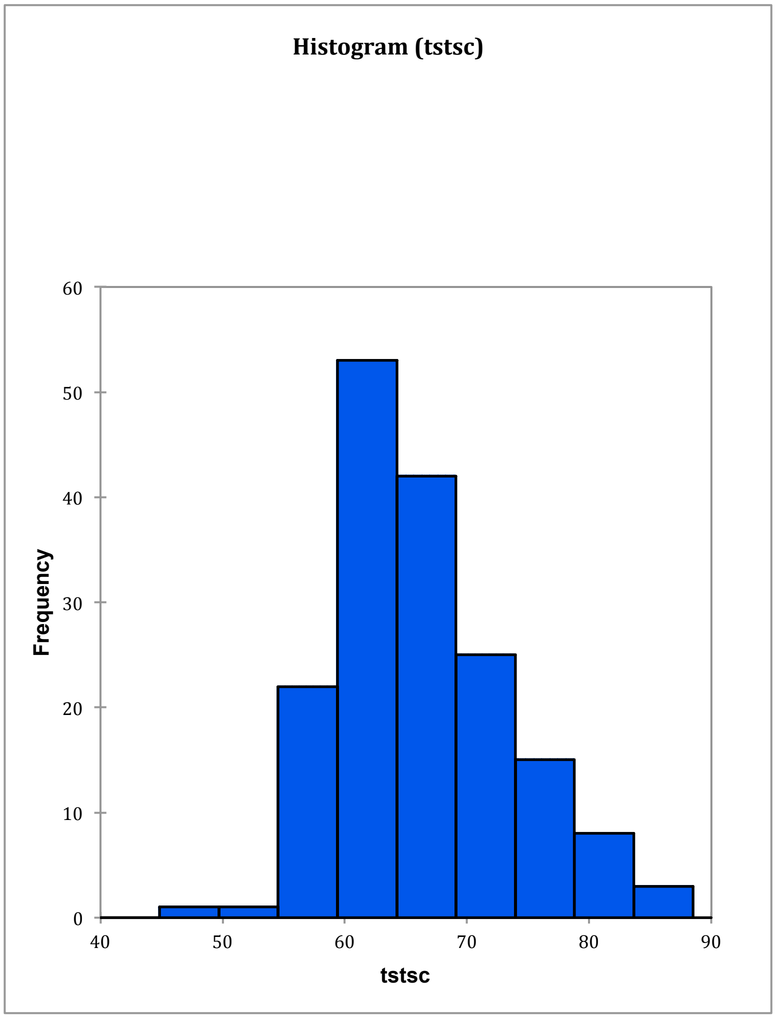


Apret	
Mean	56.72107647
Standard Error	1.386450032
Median	55.7085
Mode	72
Standard Deviatior	18.07709676
Sample Variance	326.7814274
Kurtosis	-0.554450128
Skewness	0.089185832
Range	76.5
Minimum	18.75
Maximum	95.25
Sum	9642.583
Count	170
Largest(1)	95.25
Smallest(1)	18.75
Confidence Level(|	2.736991575



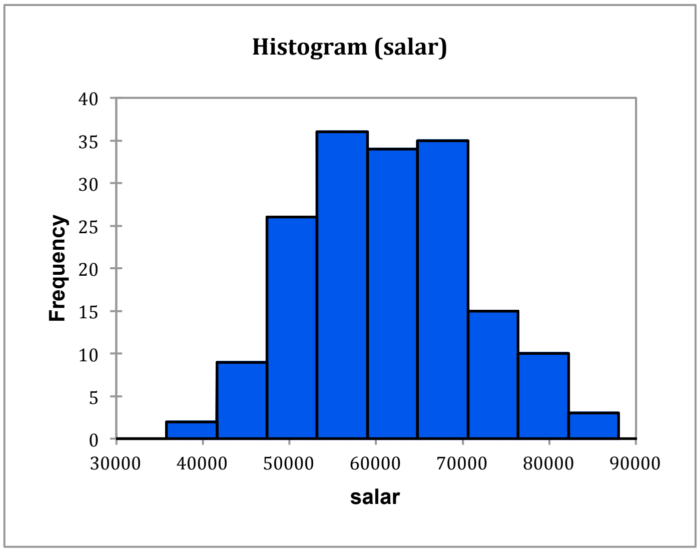
The apret data is spread across the values 18.75 to 95.25 with a range 76.5

tstsc	
Mean	66.16416471
Standard Error	0.534981569
Median	64.7815
Mode	61.111
Standard Deviation	6.975306256
Sample Variance	48.65489737
Kurtosis	0.196426383
Skewness	0.573217572
Range	39.375
Minimum	48.125
Maximum	87.5
Sum	11247.908
Count	170
Largest(1)	87.5
Smallest(1)	48.125
Confidence Level(95.0%)	1.056107333



The histogram is right skewed (mean>median)

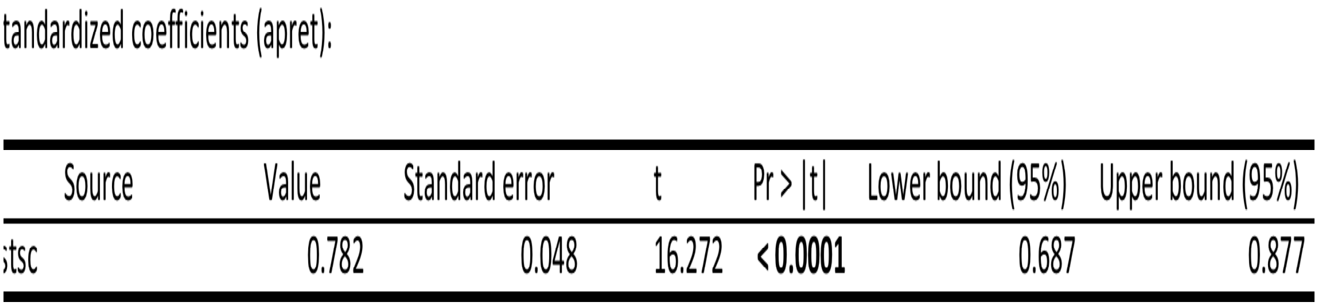
salar	
Mean	61357.64706
Standard Error	751.8394005
Median	61150
Mode	48000
Standard Deviation	9802.786457
Sample Variance	96094622.31
Kurtosis	-0.231096674
Skewness	0.257876678
Range	49260
Minimum	38640
Maximum	87900
Sum	10430800
Count	170
Largest(1)	87900
Smallest(1)	38640
Confidence Level(95.0%)	1484.206468



Analysis :

For generating the descriptive statistics and histogram, Microsoft excel was used. XLSTAT was used as the statistical tool. When we use python or R, we need to import the data and write code. But using Excel and XLSTAT, it was super easy.
Histogram helps in understanding how data is spread on the basis of certain range.

b.1) Perform linear regression of apret on tstsc



tandardized coefficients (apret):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
tsc	0.782	0.048	16.272	**< 0.0001**	0.687	0.877

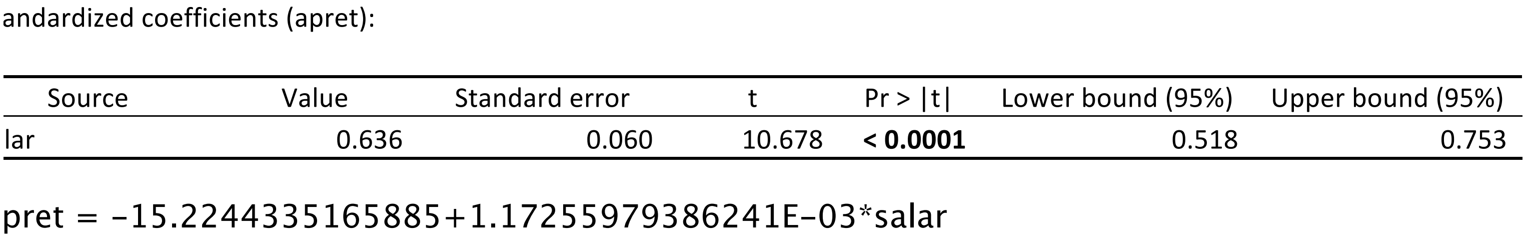
Equation of the model (apret):

apret = -77.3998900035077+2.02709377606896*tstsc

Correlation matrix:

	tstsc	apret
tstsc	1	0.782
apret	0.782	1

b.2) Perform linear regression of apret on salar



andardized coefficients (apret):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
lar	0.636	0.060	10.678	**< 0.0001**	0.518	0.753

pret = −15.2244335165885+1.17255979386241E−03*salar

Correlation matrix:

	salar	apret
salar	**1**	0.636
apret	0.636	**1**

b.3) Apret dependent on tstsc and salar

andardized coefficients (apret):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
lar	0.156	0.068	2.298	**0.023**	0.022	0.290
tsc	0.670	0.068	9.868	**< 0.0001**	0.536	0.805

apret = -75.9111069199262+2.87971945408734E-04*salar+1.73754029711285*tstsc

Correlation matrix:

	salar	tstsc	apret
salar	**1**	0.715	0.636
tstsc	0.715	**1**	0.782
apret	0.636	0.782	**1**

nalysis:

The value of apret (average retention rate) depends more on tstsc as compared to salar(salary of teachers). As we add the two variables to find value of apret, the impact of the two variables changes.

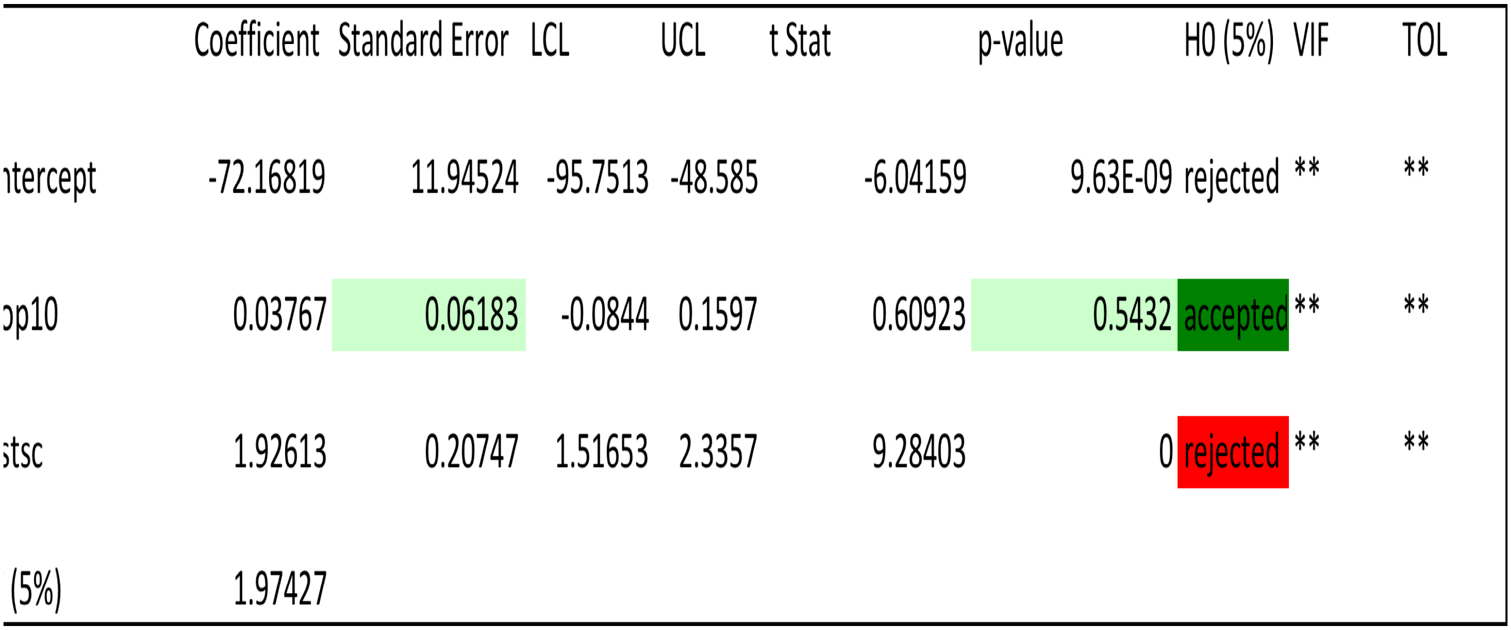
Are the data normal ?

The presence of a significant interaction indicates that the effect of one predictor variable on the response variable is different at different values of the other predictor variable.
When we have two predictor variables, the impact on the value of apret is different .
An interaction term helps to understand the fact that when we have more than one variable, the interpretation of all of the coefficients of input variables changes drastically.

Other useful observations:

ANOVA
(Analysis of
Variance)

	d.f.	SS	MS	F	p-value
Regression	2	33,835.42	16,917.71	132.07916	0
Residual	167	21,390.64	128.08766		
Total	169	55,226.06			



	Coefficient	Standard Error	LCL	UCL	t Stat	p-value	H0 (5%)	VIF	TOL
ntercept	-72.16819	11.94524	-95.7513	-48.585	-6.04159	9.63E-09	rejected	**	**
op10	0.03767	0.06183	-0.0844	0.1597	0.60923	0.5432	accepted	**	**
stsc	1.92613	0.20747	1.51653	2.3357	9.28403	0	rejected	**	**
(5%)	1.97427								

LCL - Lower limit of the 95% confidence interval
UCL - Upper limit of the 95% confidence interval

I conclude that we should have included top10 in our model, it would have provided a better insight as compared to using salary of teachers on student retention rate.