INFSCI 2725 Data Analytics

Assignment 3

Zheng Liu E-mail:zhengliu@pitt.edu

Luda Wang E-mail:luw20@pitt.edu

Kewei Li E-mail:kel137@pitt.edu

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

The data description is below:

	spend	apret	top10	rejr
count	170.00	170.00	170.00	170.00
mean	10974.51	56.72	38.46	30.65
std	5500.07	18.08	23.41	17.10
min	4125.00	18.75	8.00	0.00
25%	7371.75	45.37	22.00	19.17
50%	9265.00	55.71	30.00	27.39
75%	12838.00	68.69	49.50	36.81
max	35863.00	95.25	98.00	84.07
	tstsc	pacc	strat	salar
count	170.00	170.00	170.00	170.00
mean	66.16	43.17	16.09	61357.65
std	6.98	13.11	4.01	9802.79
min	48.13	8.96	7.20	38640.00
25%	61.11	33.90	13.40	54650.00
50%	64.78	40.85	16.00	61150.00
75%	70.45	51.77	18.58	67100.00
max	87.50	76.25	29.20	87900.00

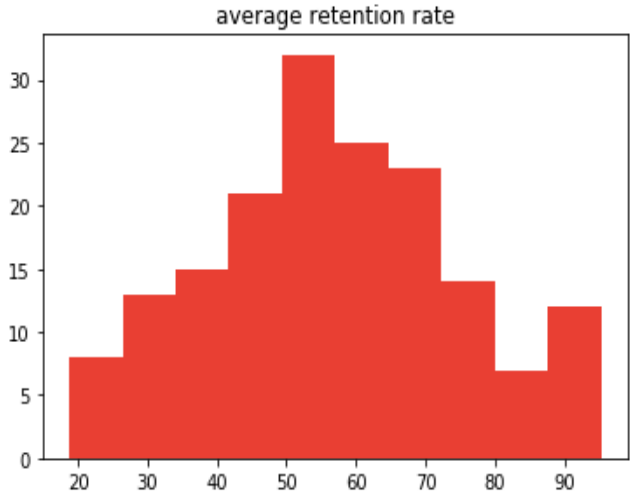
We performed the norm test for all data. The result table is below.
Based on the P-value, we can tell the apret,strat and salar is not norm data.

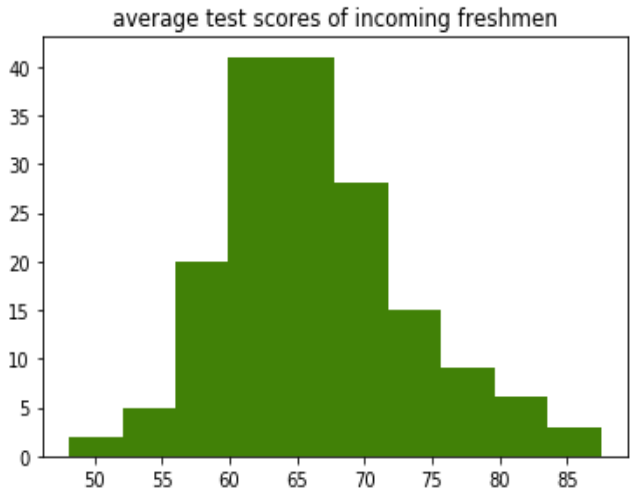
	P-value	if P-value<0.05
spend	1.58E-15	TRUE
apret	0.137658975	FALSE
top10	1.53E-05	TRUE
rejr	1.36E-06	TRUE
tstsc	0.010140806	TRUE
pacc	0.043632866	TRUE
strat	0.077490657	FALSE
salar	0.322014371	FALSE

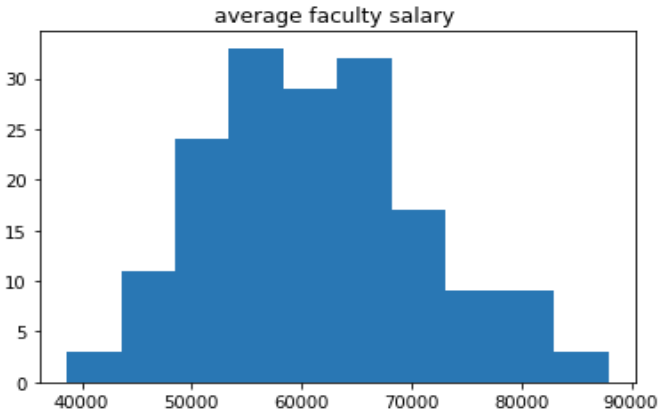
We also perform the correlation with APRET. The table is below.

	Correlation
spend	0.601231173
apret	1
top10	0.642464456
rejr	0.514957973
tstsc	0.782183117
pacc	-0.302833887
strat	-0.458311427
salar	0.635851731

Feature with the most linear relation like is TSTSC. The least is PACC. The three histograms are below:







2) Perform linear regression of apret on tstsc and salar separately and then of apret on both tstsc and salar.

a. Regression with TSTSC

The function of APRET(Y) on TSTSC(X1) is :

Y=-77.3999+2.0271*X1

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-77.3999	8.288	-9.339	0.000	-93.762	-61.038
tstsc	2.0271	0.125	16.272	0.000	1.781	2.273

The p-value is good for coefficient.

b. Regression with SALAR

The function of APRET(Y) on SALAR(X1) is:

Y=-15.2244+0.0012*X1

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-15.2244	6.823	-2.231	0.027	-28.693	-1.755
salar	0.0012	0.000	10.678	0.000	0.001	0.001

The p-value is good for coefficient.

c. Regression with TSTSC and SALAR

The function of APRET(Y) on SALAR(X1) and TSTSC(X2) is :

$$Y=-75.9111+0.0003*X1+1.7375*X2$$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.9111	8.210	-9.246	0.000	-92.119	-59.703
salar	0.0003	0.000	2.298	0.023	4.06e-05	0.001
tstsc	1.7375	0.176	9.868	0.000	1.390	2.085

The p-values are good for both coefficients.

3) And we have some interesting observations:

a. By using all the combinations of features to perform the linear regression with APRET, we find the best combination to predict the APRET by choosing the largest Adjusted R-Squared. The combination is TSTSC, PACC and STRAT. The regression result by using this combination is below.
The function is:

$$APRET=-45.7292+1.8229*TSTSC-0.2387*PACC-0.4884*STRAT$$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-45.7292	11.711	-3.905	0.000	-68.850	-22.608
tstsc	1.8229	0.135	13.509	0.000	1.557	2.089
pacc	-0.2387	0.064	-3.722	0.000	-0.365	-0.112
strat	-0.4884	0.234	-2.089	0.038	-0.950	-0.027

All of the P-values are below 0.05, which means that all the coefficients are significantly different from zero.

b. We used the Python package “apyori” to perform the frequent item-set selection. For do the frequent item-set selection, we processed the data first. We separated each variable by its range into low, medium and high. We use the “apriori” function to get the most frequent item-set with three variables. The support is over 0.35.
The results are :

item-set	support
'spend-low', 'pacc-medium', 'top10-low'	0.38
'spend-low', 'top10-low', 'rejr-low'	0.39
'salar-medium', 'spend-low', 'tstsc-medium'	0.36
'spend-low', 'top10-low', 'strat-medium'	0.44