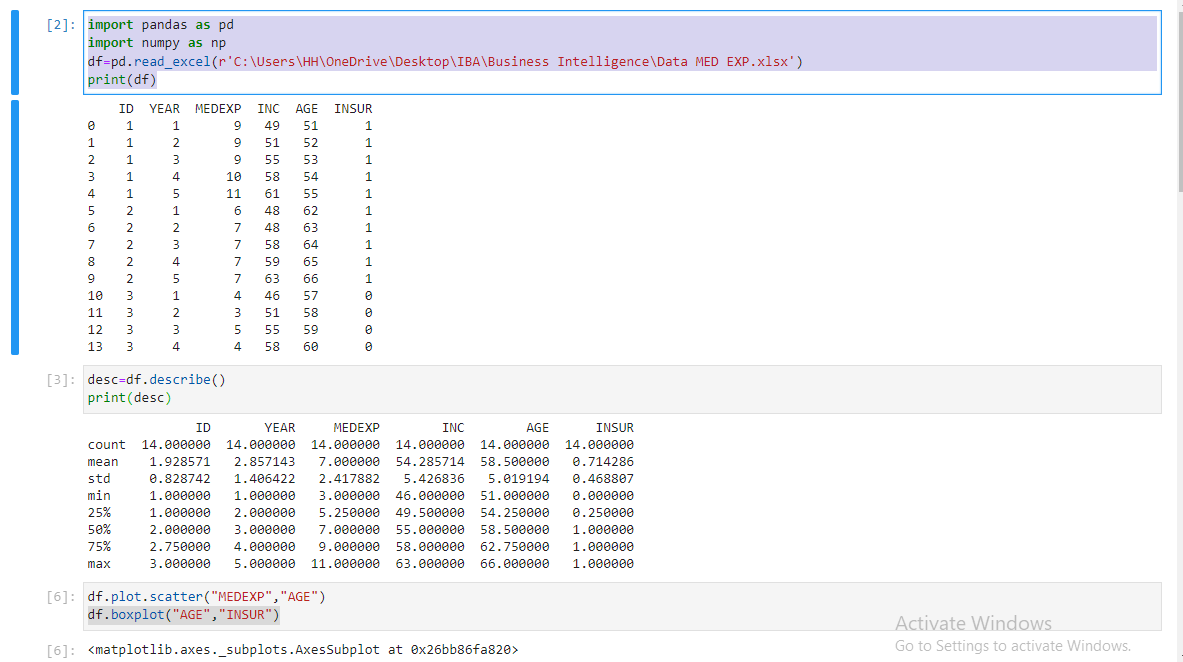
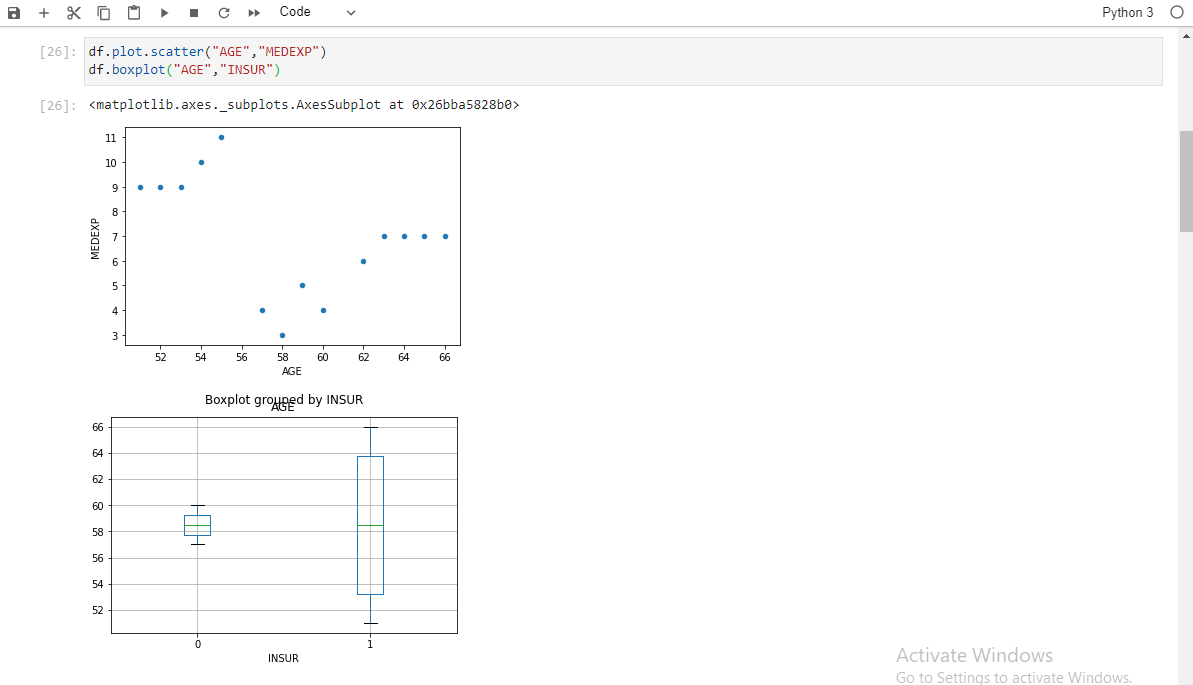
**Shaheryar Ali – 23841 – Business Intelligence**

**TASK 2**

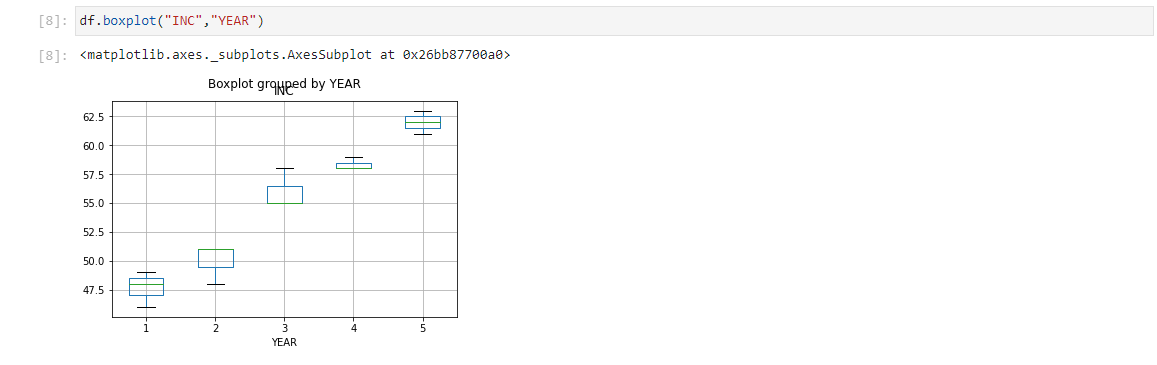
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2.** |  |  |  |  |  |  |  |  |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.637521114 |  |  |  |  |  |  |  |
| R Square | 0.406433171 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.298511929 |  |  |  |  |  |  |  |
| Standard Error | 2.025094364 |  |  |  |  |  |  |  |
| Observations | 14 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 2 | 30.88892097 | 15.44446049 | 3.766016442 | 0.056765299 |  |  |  |
| Residual | 11 | 45.11107903 | 4.101007184 |  |  |  |  |  |
| Total | 13 | 76 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 11.57347496 | 7.630033237 | 1.516831526 | 0.157508891 | -5.220114967 | 28.36706488 | -5.220114967 | 28.36706488 |
| INC | 0.216002001 | 0.108174282 | 1.996796248 | 0.071193595 | -0.022087989 | 0.454091992 | -0.022087989 | 0.454091992 |
| AGE | -0.278620477 | 0.116959835 | -2.382189374 | 0.036360106 | -0.536047339 | -0.021193615 | -0.536047339 | -0.021193615 |
|  |  |  |  |  |  |  |  |  |
| **3.** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **Multiple R** | It shows that the medical expense is positively correlated with Income and Age and 0.638 means that a change of 1 in variables would effect the Medical expense upto 0.638. | | | | | | |  |
|  |
|  |
|  |  |  |  |  |  |  |  |  |
| **R Square** | The strength of correlation is decent, it means that the 40.64% data is closely fitted to the regression line, ideally a r square value over 0.7 is considered strongly correlated and values below 0.4 is considered weakly correlated. Since the figure at hand of 0.406 is slightly above 0.4 we can consider it decent but not strong. | | | | | | |  |
|  |
|  |
|  |
|  |  |  |  |  |  |  |  |  |
| **p-values** | The P-value of Age is below 0.05 meaning it is significant and is a stronger factor for medical expense than Income since the P-value of income is above 0.05 meaning it is insignificant and is not a proper base. However, if we take them of the factor of being close to zero then both are resonably close to zero since both are less than 0.1 | | | | | | |  |
|  |
|  |
|  |
|  |  |  |  |  |  |  |  |  |
| **T-stat** | Income | it is slightly less than 2 meaning it fails the test | | | |  |  |  |
| Age | It is above 2 meaning it signifies the factor of age being correlated to medical expense | | | | | |  |
|  |  |  |  |  |  |  |  |  |
| **S.E < I Estimate I** | Income | TRUE | Passes the test | |  |  |  |  |
|  | Age | TRUE | Passes the test | |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **4.** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| The scenario presents two x variables Age and Income. According to the multipl linear regression analysis of the data of 14 individuals it can be observed that Age is a stronger and significant factor / variable since it has a closer to zero p-value, a standard error less than its estimate and lastly a T-stat value of greater than 2. Whereas the other factor Income is relatively not significant since it only passes the Standard Error being less than estimate test and its P value is above 0.05 and is a bit far from zero, lastly its T-stat is also slightly less than two, hence it is not a significant factor. | | | | | | | |  |
|  |
|  |
|  |
|  |
|  |
|  |  |  |  |  |  |  |  |  |
| **5.** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Model | MED EXP = 11.573+(0.2160\*Inc)-(0.2786\*Age) | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Age | 71 |  |  |  |  |  |  |  |
| Income | 40 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| MED Expense | 0.4324 |  |  |  |  |  |  |  |

**TASK 3**

****

****

The category of people without insurance have a much smaller range from only 57 to 60 years whereas people who have insurance range from lower than 52 years to 66 years, however the mean age of both such people is almost the same.

****

There is a definite inclination of years to income as the years of people increase their income increases as well and the results will be definitely significant if p value testing would be carried out using linear regression.

**TASK 4**

1. Null hypothesis is a statement that refers to declining or proving that there is no association between two or more variables, it is a hopeless statement with a equal sign usually. Null hypothesis is refered as Ho (H not).

Alternate hypothesis is a statement that refers to a difference or proving that there might be an association between two or more variable, it is a hopeful statement with a greater than, less than or not equal to sign. Alternate hypothesis is referred as Ha.

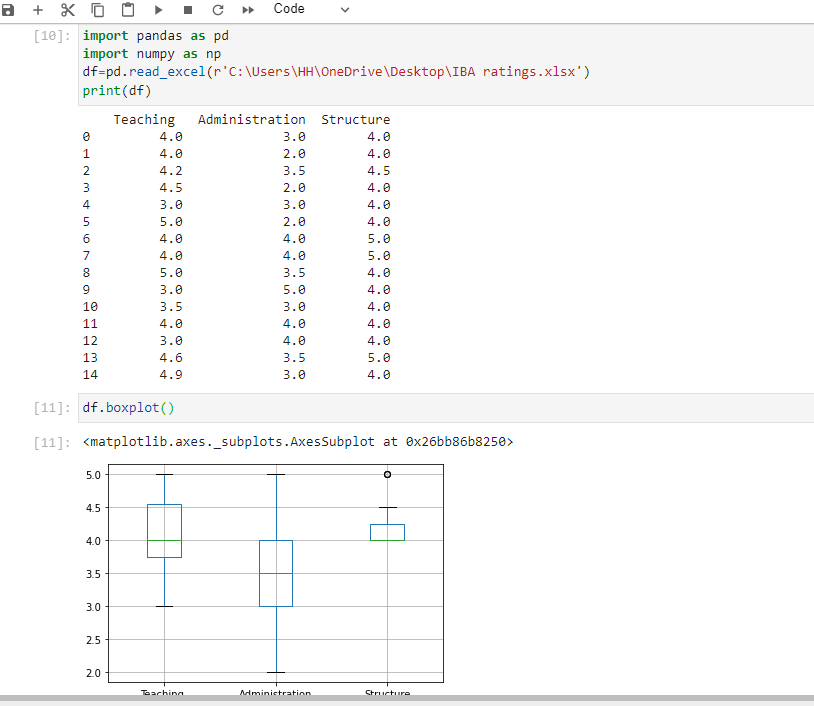
1. There can be a total of 3 possibilities where:
2. There are two quantitative variables, in such scenarios scatter diagram / plot can be used along with correlation and regression as tests to verify the possible association between them. Example of such variable could be age and income.
3. There are two qualitative variables, in such scenarios cross tabulation could be used along with Chi-square test of association to verify its association. Example could be gender and work sector.
4. There could be one qualitative and one quantitative variable, in such scenarios Box plot could be drawn and T-test / Z-test could be used depending on size of data. Anova is another possible way to test the association. Example could be gender and age of individuals.
5. Null Hypothesis would be Ho = There is no association between marital status and satisfaction from salaries.

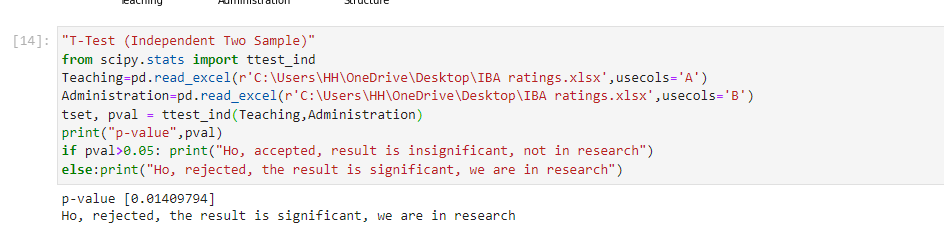
Whereas Ha = There is association between marital status and satisfaction from salaries.

**Comment**

Given our p-value of 0.689 it is greater than 0.05 we have to accept the null hypothesis, marking the result as insignificant, not being in research hereon and accept that there is no association between marital status and satisfaction from salary.

**TASK – 5**





Comments:

The results of the box plot show that there is a clear difference in means of the teaching compared with administration, whereas the teaching scores are skewed to the higher side near 5, administration scores are far spread from 2 till 5. However only on the basis of mean and its distribution it is difficult to say whether any correlation exists.

Further, as far as the T-Test results are concerned the results show that there is correlation and since the p-value is below 0.05 it is significant and research can be made on the fact that their means are correlated.