**ADVANCED STATISTICS**

**A**yyalasomayajula Aditya

**1. A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv**

|  |  |
| --- | --- |
| **1.1 State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables** | |
| **‘A’ and ‘B’ individually.** |  |

**Hypothesis for ingredient A:**

Null Hypothesis – 𝐻**o:** Means of all the 3 levels in ingredient A are equal i.e the mean hours of Relief ('Relief') resulted when the patients were observed with each of the 3 different levels of ingredient A are equal. i.e When ingredient A is added at, either of the 3 levels, to the compound, it does not make any difference on the relief time for the Hay fever.

Alternate Hypothesis – 𝐻𝑎:At least 1 level of ingredient A gives significantly more hours of relief than the other 2 levels of ingredient A. The mean hours of Relief ('Relief') for at least one level of ingredient A are unequal.

# 𝐍𝐮𝐥𝐥-𝐡𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬

𝐻0: 𝜇𝐴1 = μA2 =μA3

𝐻0: The means between various levels are equal.

# 𝐀𝐥𝐭𝐞𝐫𝐧𝐚𝐭𝐢𝐯𝐞-𝐡𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬

𝐻a: 𝜇𝐴1 ≠ μA2 =μA3

(or)

𝐻a: μA1 = μA2 ≠ μA3

(or)

𝐻a: 𝜇𝐴1 = μA3 ≠ μA2

(or)

𝐻a: μA1 ≠ μA2≠ μA3

𝐻a: The means of at least one level of ingredient A are unequal.

**Hypothesis for ingredient B:**

Null Hypothesis – 𝐻**o:** Means of all the 3 levels in ingredient B are equal i.e the mean hours of Relief ('Relief') resulted when the patients were observed with each of the 3 different levels of ingredient B are equal. i.e When ingredient B is added at, either of the 3 levels, to the compound, it does not make any difference on the relief time for the Hay fever.

Alternate Hypothesis – 𝐻𝑎:At least 1 level of ingredient A gives significantly more hours of relief than the other 2 levels of ingredient B. The mean hours of Relief ('Relief') for at least one level of ingredient B are unequal.

# 𝐍𝐮𝐥𝐥-𝐡𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ~

𝐻0: 𝜇B1 = μB2 =μB3

𝐻0: The means between various levels are equal.

# 𝐀𝐥𝐭𝐞𝐫𝐧𝐚𝐭𝐢𝐯𝐞-𝐡𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢s

𝐻a: 𝜇B1 ≠ μB2 =μB3

(or)

𝐻a: μB1 = μB2 ≠ μB3

(or)

𝐻a: 𝜇B1 = μB3 ≠ μB2

(or)

𝐻a: μB1 ≠ μB2≠ μB3

𝐻a: The means of at least one level of ingredient B are unequal.

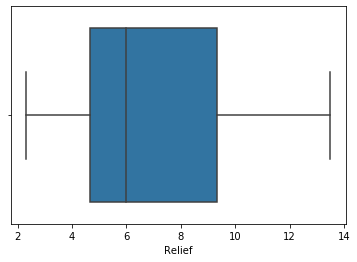
**Data Setup to build the model:**

To Build an ANOVA model, the population data should be homogeneous and normally distributed.

* In the given data, **Relief** is the dependant variable and needs to be continuous. The data for Relief is as needed to build the ANOVA model.
* Data for Ingredients A, B and volunteers should be categorical to be able to build the ANOVA model for analysis. Hence the conversion of data types of the A, B and Volunteer to category.
* Derived data counts for all the various individual factors, and the data records for the various levels of ingredients and volunteers used for the treatments is homogeneous.

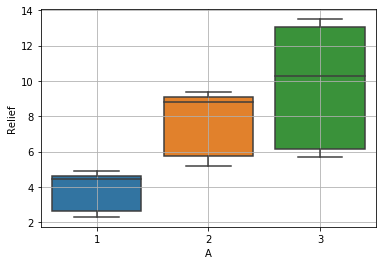
|  |  |  |
| --- | --- | --- |
| A | B | Volunteer |
| 3 12  2 12  1 12  Name: A, dtype: int64 | 3 12  2 12  1 12  Name: B, dtype: int64 | 4 9  3 9  2 9  1 9  Name: Volunteer, dtype  : int64 |

**Checking for any outliers in the hours of relief.**



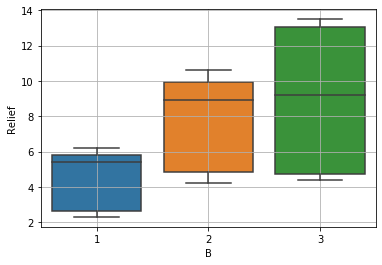
The box plot below clearly illustrates that the data is normally distributed and has no outliers.

**Checking the distribution of data for Ingredient A**



* There are no outliers within the data.
* Considering the individual data for level 1 and 2 of ingredient A, half of the respective volunteers are getting approximately the same hours of relief, while the hours of relief are vastly differing in case of level 3 of ingredient A.
* The data shows that there is significant difference in the hours of relief across various levels of ingredient A which can be confirmed by performing an ANOVA analysis on the data.
* The median number of hours of relief is highest when level 3 of ingredient A is used.

**Checking the distribution of data for Ingredient B**



* There are no outliers within the data.
* Considering levels of Ingredient - A constant, change in level of Ingredient - B from 1 to 2 gives significant change in hours of relief, while change from level 2 to 3 is giving significantly less change in hours of relief.
* The data shows that there is significant difference in the hours of relief across various levels of ingredient B which can be confirmed by performing an ANOVA test on the data.

|  |  |
| --- | --- |
| **1.2) Perform one-way ANOVA for variable ‘A’ with respect to the variable ‘Relief’. State** | |
| **whether the Null Hypothesis is accepted or rejected based on the ANOVA results.** |  |

**Results of the ANOVA test thus performed:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | Sum-sq | Mean-sq | F-Statistic | PR(>F) |
| C(A) | 2.0 | 220.02 | 110.010000 | 23.465387 | 4.578242e-07 |
| Residual | 33.0 | 154.71 | 4.688182 | NaN | NaN |

Since the p-value obtained, **4.578242e-07,** from ANOVA test for ingredient 'A' is less than α (0.05). We conclude that there is at least one level of ingredient A that gives more relief from hay fever than the other two levels as we can clearly see that the means of the 3 levels are different.

***If p - value is less than alpha, we have evidence to reject the null hypothesis since p value < Level of significance.***

***If p-value is greater than alpha-value, we fail to reject the null hypothesis since p value > Level of significance.***

Since p-value < alpha, here we reject the null hypothesis and accept the alternate hypothesis.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1.3) Perform one-way ANOVA for variable ‘B’ with respect to the variable ‘Relief’. State** | | | | | | | | |
| **whether the Null Hypothesis is accepted or rejected based on the ANOVA results.** | | | | | | |  | |
|  | df | sum\_sq | mean\_s | F-Stat | PR(>F) | |
| C(B) | 2.0 | 123.66 | 61.830000 | 8.126777 | 0.00135 | |
| Residual | 33.0 | 251.07 | 7.608182 | NaN | NaN | |

**Interpretation**: Since the p – value, **0.00135** obtained from ANOVA analysis for ingredient 'B' is less than α (0.05). We conclude that there is at least one level of ingredient B that gives more relief from hay fever than the other two levels as we can clearly see that the means of the 3 levels is different.

**1.4**

**)**

**Analyse the effects of one variable on another with the help of an interaction plot.**

**W**

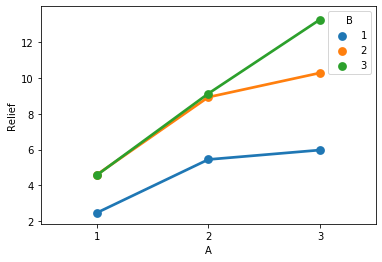
**hat is the interaction between the two treatments? [hint: use the ‘point**

**-**

**plot’ function**

**from the ‘seaborn’ function**

**]**



**Interpretation:** Since both lines are non-parallel lines, it shows that there is interaction effect between the 2 variables. This interaction effect indicates that the relationship between ingredient A and Relief depends on the value of ingredient 'B'. For instance: The data holds that the maximum hours of relief can be obtained, if one uses level 3 of Ingredient 'A' and level 3 of ingredient 'B' in the compound. Also, hours of relief do not differ much when ingredient A is interacted with level 2 and 3 variations for ingredient B.

|  |  |
| --- | --- |
| **1.5) Perform a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’)** | |
| **with the variable 'Relief' and state your results** |  |

**Forming Hypothesis:**

𝐻**o: There is no interaction between the 2 ingredients A and B i.e all combinations of various levels of A and B give the same hours of relief.**

𝐻𝑎**: At least one of the combinations of A and B levels gives more hours of relief than the other combinations.**

|  |
| --- |
| df sum\_sq mean\_sq F PR(>F) |
| C(A) 2.0 220.020 110.010000 1827.858462 1.514043e-29 |
| C(B) 2.0 123.660 61.830000 1027.329231 3.348751e-26 |
| C(A):C(B) 4.0 29.425 7.356250 122.226923 6.972083e-17 |
| Residual 27.0 1.625 0.060185 NaN NaN |

**Interpretation:**

Based on the p-values at a significance level( α ) of 0.05

Since the p-value, **6.972083e-17 -** obtained on interactions of A and B, is less than α, we reject the Null Hypothesis. Hence, we can conclude for certain that there is at least one combination among the 9 different combinations of the compound that gives more hours of relief from hay fever than the 8 other combinations.

**1.6) Mention the business implications of performing ANOVA for this particular case study. Business Implications:**

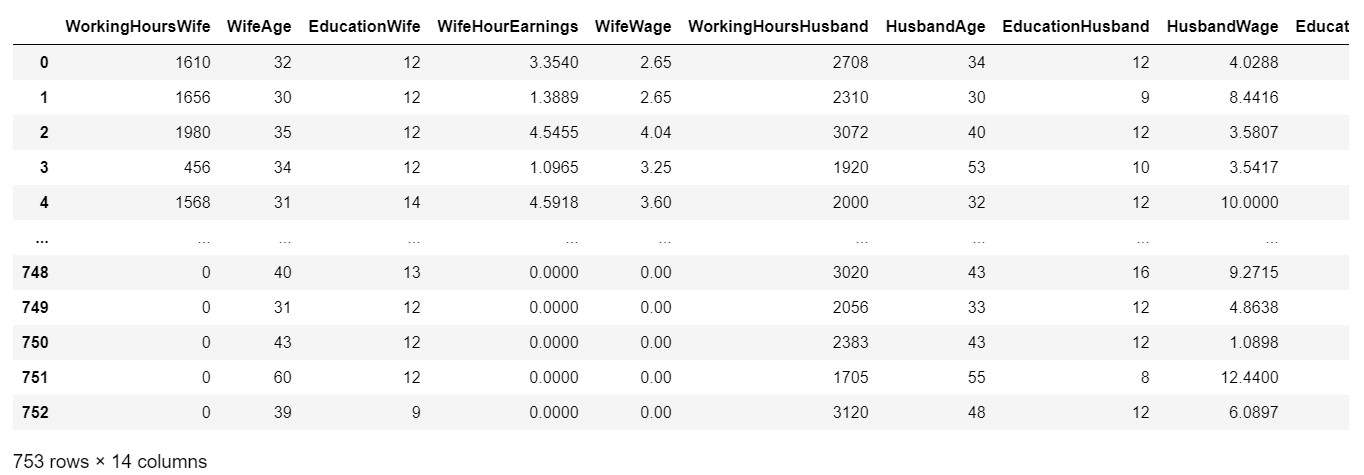
* The p-value for A is **1.514043e-29**(**<** α (0.05)), which indicates that the various levels of A added to the compound lead to different hours of relief.
* The p-value for B is **3.348751e-26**(**<** α (0.05)), which indicates that the various levels of B added to the compound lead to different hours of relief.
* The p-value for the interaction between A\*B is **6.972083e-17**(**<** α (0.05)), which implies that the relationship between A and Relief depends on the value of B. Because the interaction effect between A and B is statistically significant, you cannot interpret the effects of A and B without considering the interaction effect.
* The hours of relief are the least when the compound is made out of A1 and B1
* There is a significant increase in the hours of relief when A1 is constant, but B is changed from B1 to B2. But for the same ingredient A1, the difference in the hours of relief remains almost constant between B2 and B3.
* The central tendencies of the data shows that there exists a significant increase in the hours of relief for combo A1B1 to A2B1. Again, the increase in hours from the compound A2B1 to A2B2 is significant but, the difference in hours of relief for compound A2B2 to A2B3 is very less significant.
* The means and central tendencies definitely ascertain that given the volunteer conditions, the compound that gives the most hours of relief in significant number of cases is level 3 of ingredient A and level 3 of ingredient B(A3B3).

**Problem 2:**

**A company performed a survey to understand the income of households in various neighbourhoods of a country. The data dictionary is also present. You can access the data dictionary from the following file** [**Income\_Data Dictionary.**](https://olympus.greatlearning.in/courses/10942/files/1795019/download?verifier=7nqaUrC6TYHnoc85d0rFcbiLswmDSri5RNMneGF1&wrap=1) **Please refer to the following data set to solve the problem** [**Income.csv.**](https://olympus.greatlearning.in/courses/10942/files/1795018/download?verifier=EyyhNXTMqnZJ5D6UlZQXoxPMltY0j2gQZHTvVzok&wrap=1) **['FamilyIncome' is the target variable]**

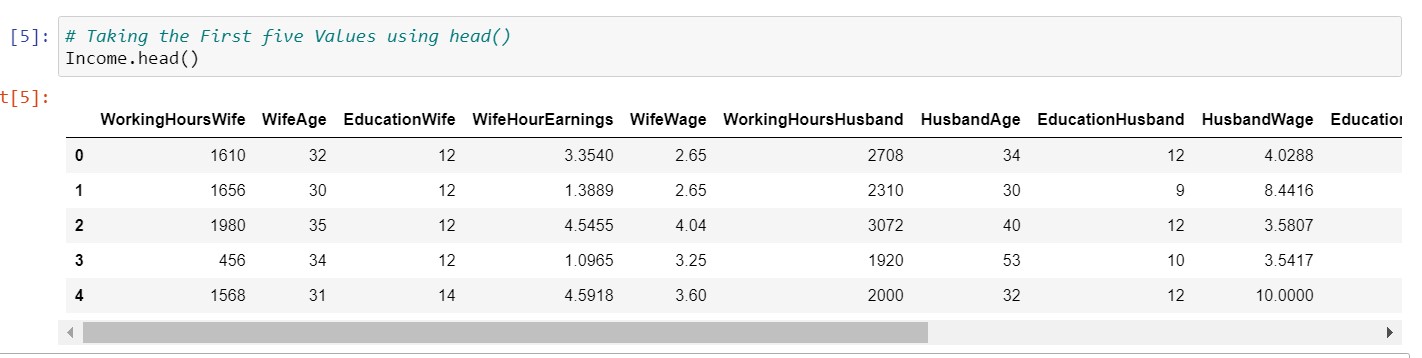
**2.1) Perform exploratory data analysis on the dataset. Showcase some charts, graphs.**

 **EDA**

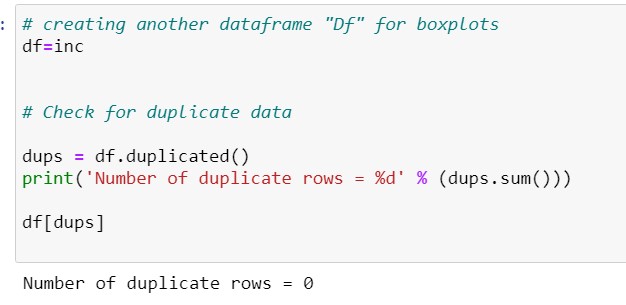
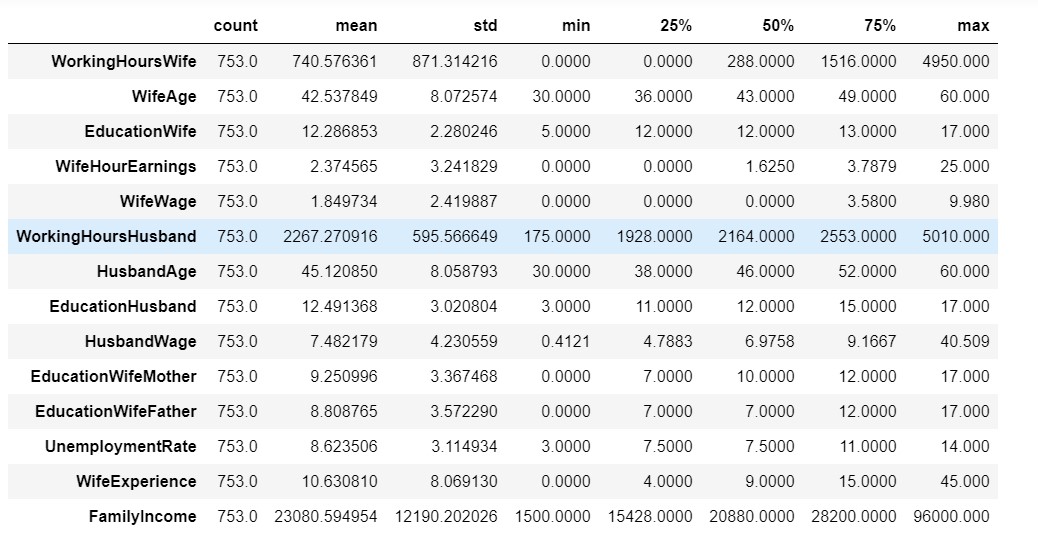


**Getting the complete data.**

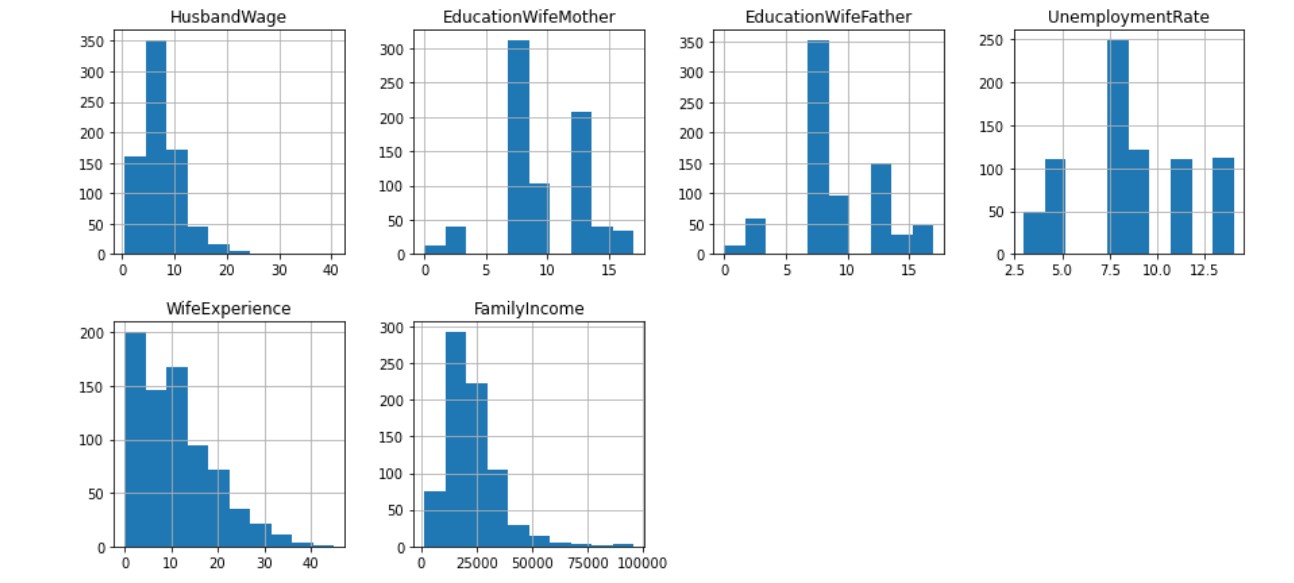
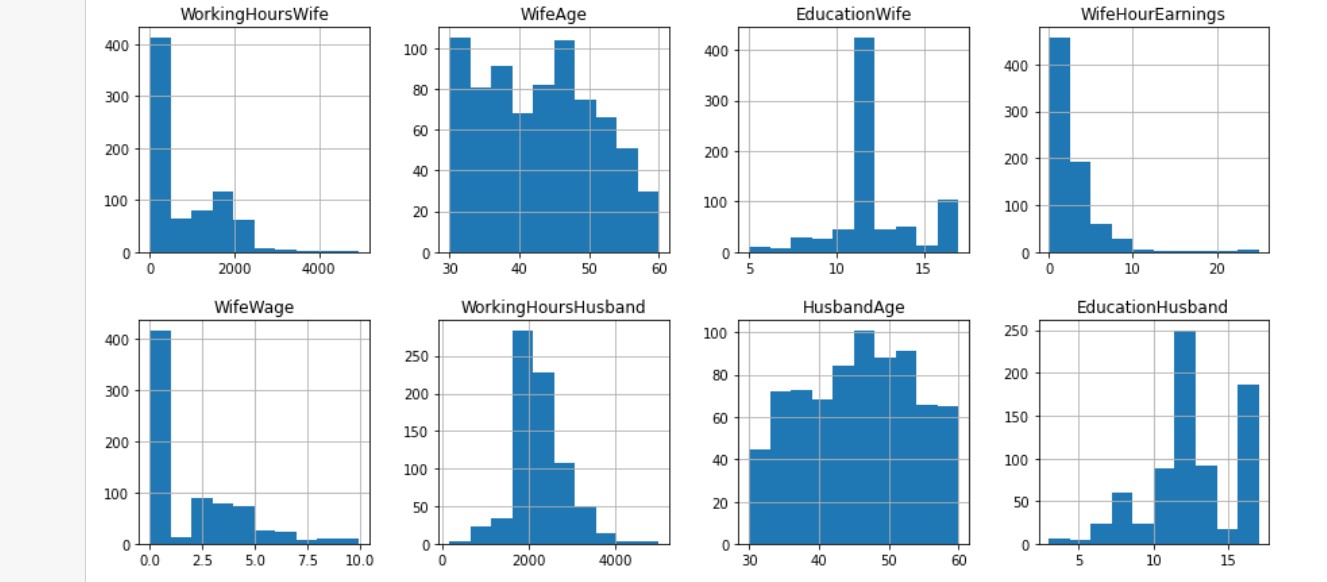
* + - **Header of the data**



* + - **Getting The summary of the dataset**

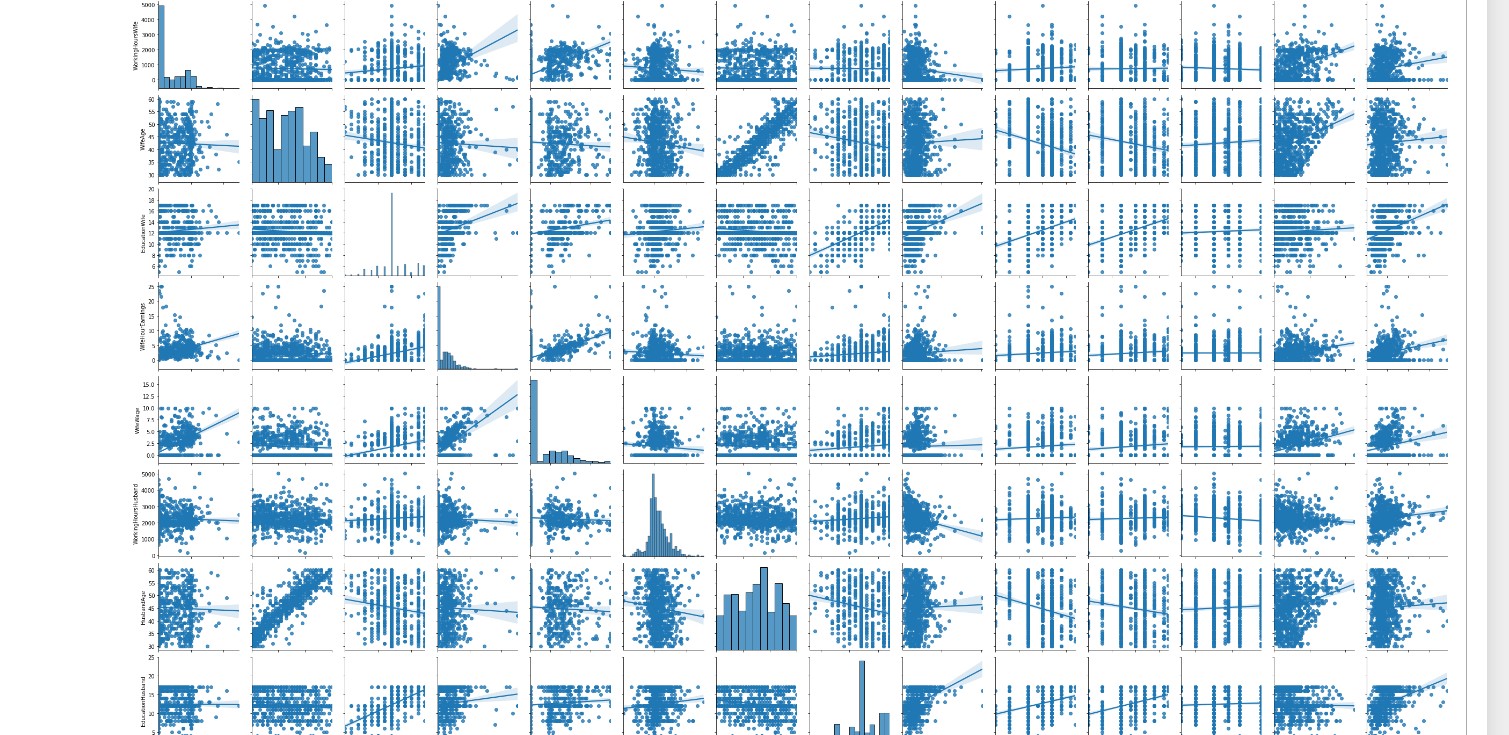


* + - **Histogram of the variables:**



**Wife'age** is multimodal. **WifeExperience** , **Wife-hour-Earnings**, **Husband-wage** **family's** **income** is right skewed **husband-age**, **working-Hours-Husband** is normally distributed.

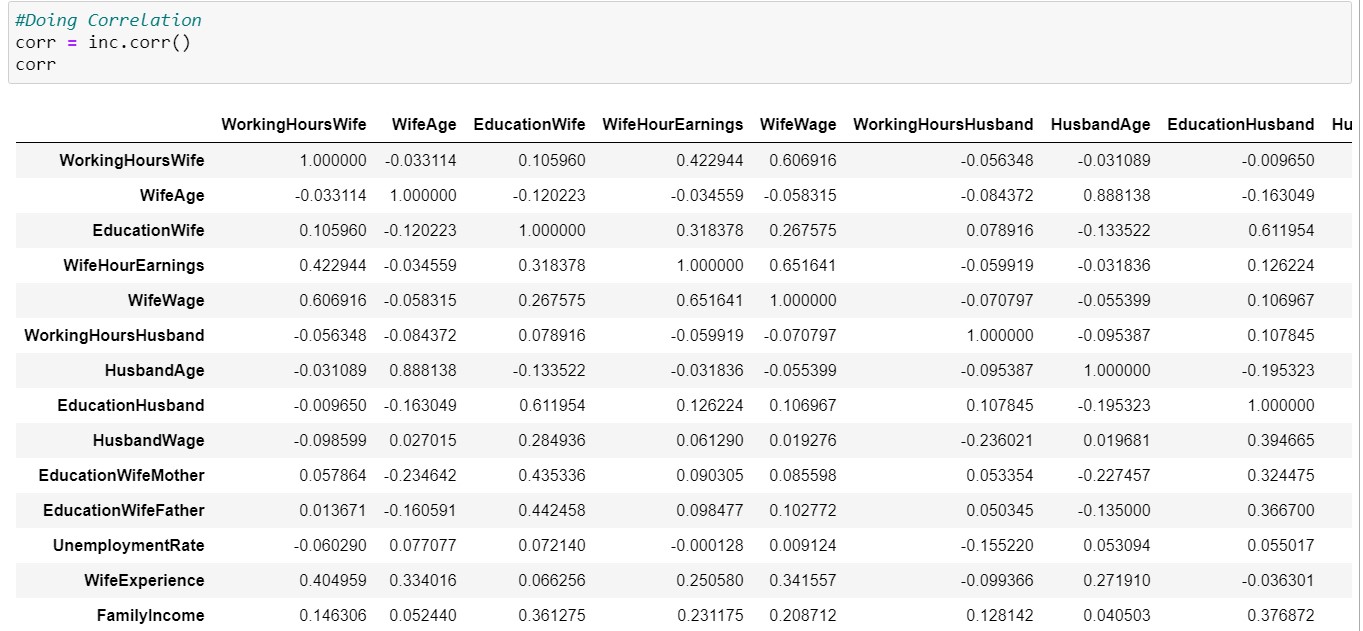
* + - **Pair-Plot of the dataset**



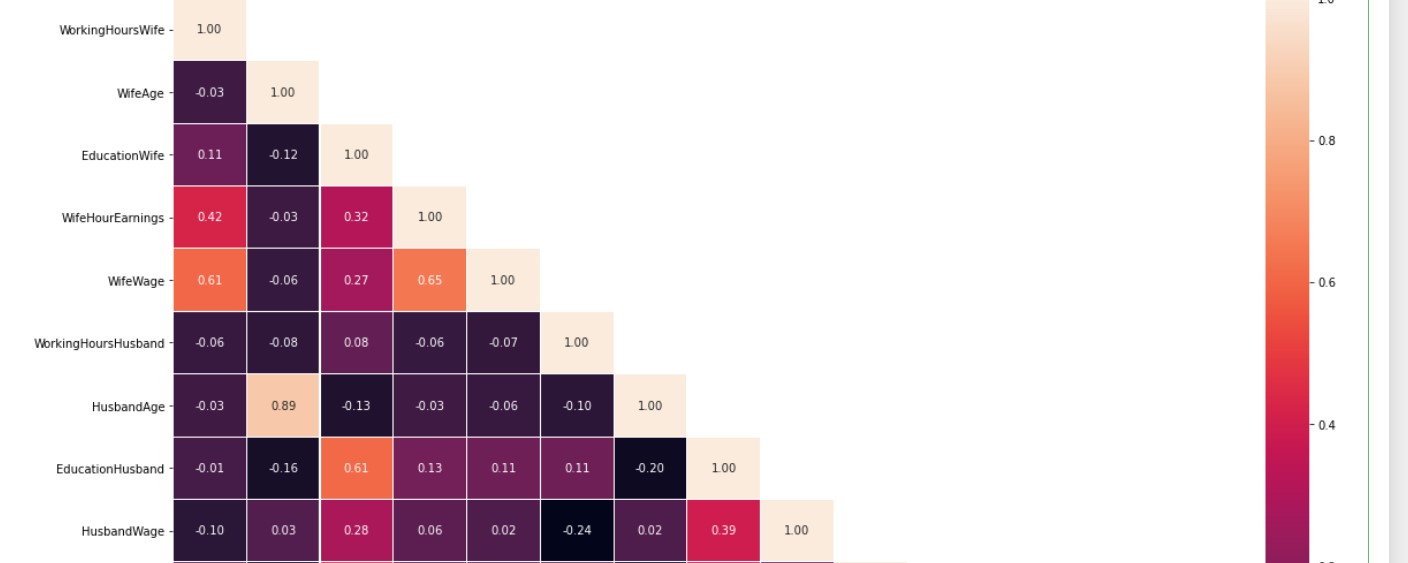
* + - This dataset is Income dataset containing 753 records of various factors and variables conatining of Wife's demographics with respect to job, education and Family.
    - There are no duplicates neither null values for any variable.
    - There are 381 wives who have 12 years of education and 103 wives with more than or equivalent 16-17 years of education and 186 husbands have education equivalent to 16-17 years of education and this combination is resulting into most of the outliers in family income.
    - There are insights regarding educated women who have educated parents choosing educated husbands who have high wages which is resulting high family income, as they are mostly employed and also leading some of the most high value outliers.
    - Upon doing exploration data analysis of income dataset we have observed family income is affected by various variables like wife's wage, experience and working hours along with husband's wages, education as the well educated parents choose well educated spouses for their children leading to high standard life.

**2.2) Is there evidence of multi-collinearity? Showcase your analysis**

* **Doing Corelation**



* **Now checking Heat-map**



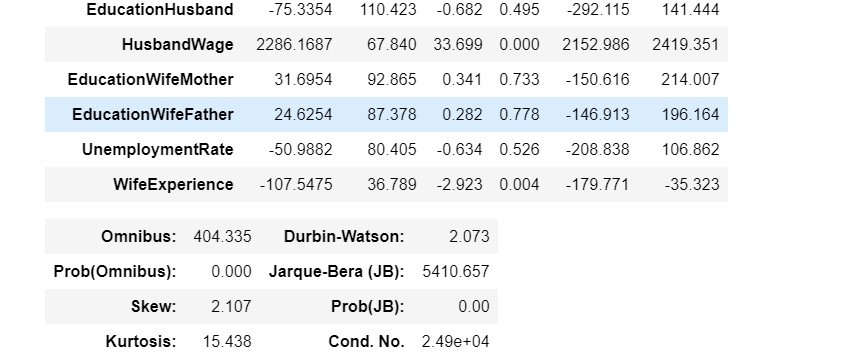
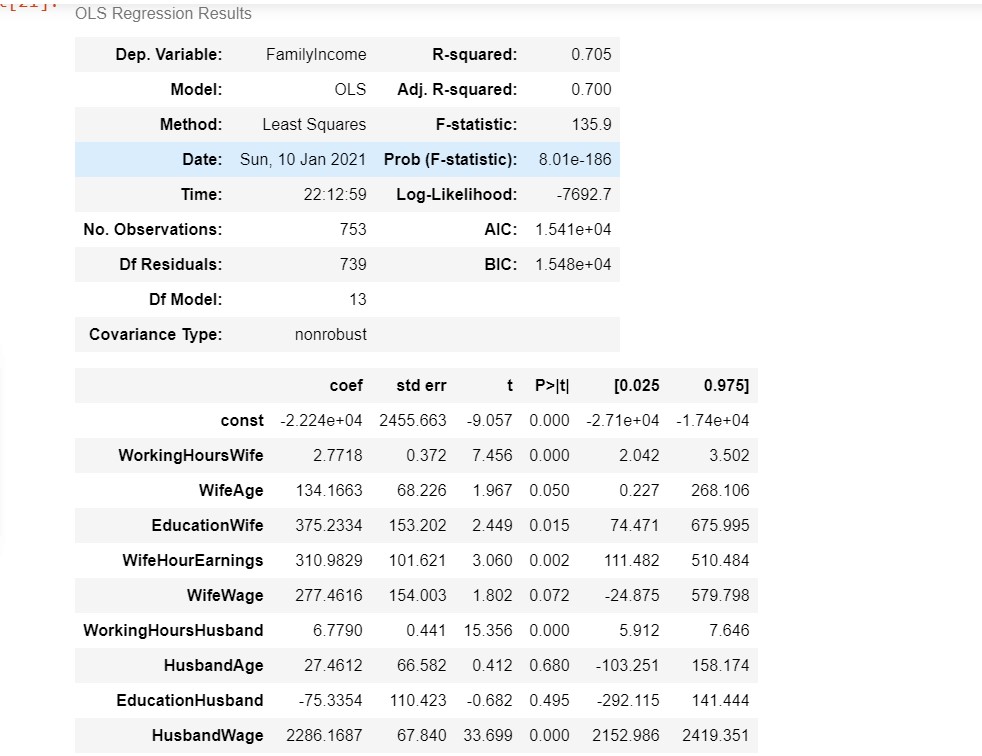
By observing the heat map and pair-plot we see that there is evidence of multi-collinearity within the independent variables.

In this dataset, following predictors for **Family** **Income** are strongly correlated: **Husband**-**Age** and

**Wife-Age.** **Wife-wage** and **working-hours wife**. **Education-Wife-Mother** and **Education**-**WifeFather** . **Education-Husband** and **Education-Wife**

**2.3) Perform Multiple Linear Regression (using the 'statsmodels' library) and comment on the model thus built.**

* **After Performing Multi-linear-Regression**



* The R\_square value : 0.705 this model is quite good as we getting this R-square value F-statistic = 135.9.

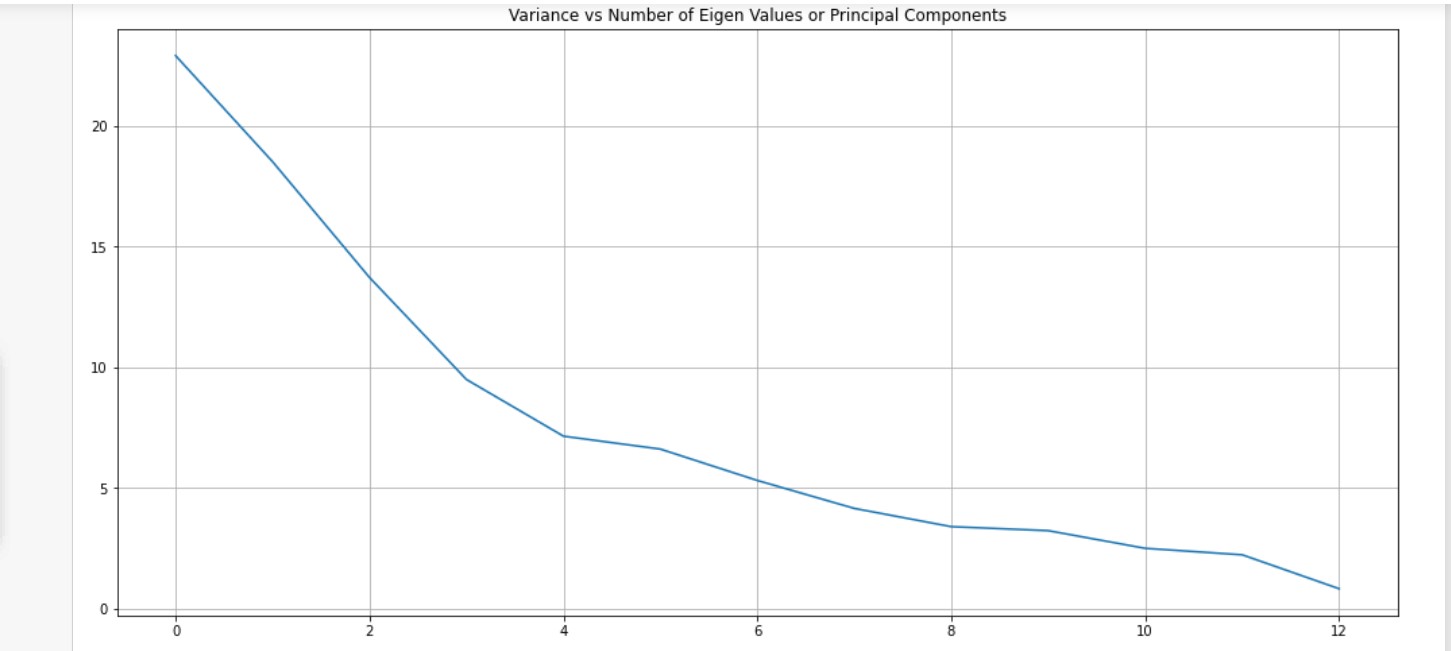
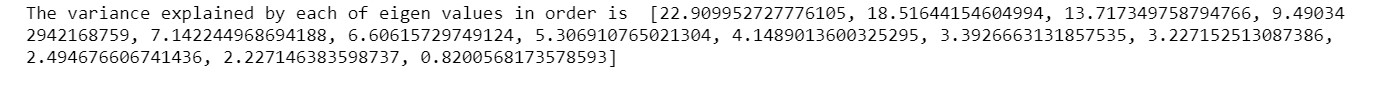
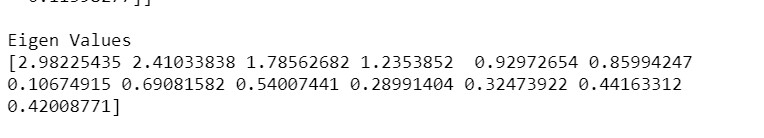
* The Root Mean Square Error (RMSE) of the model is 6615.60216353618
* This model is flawed as we saw the multi-collinearity in the data which will lead false results hence we should proceed with PCA.

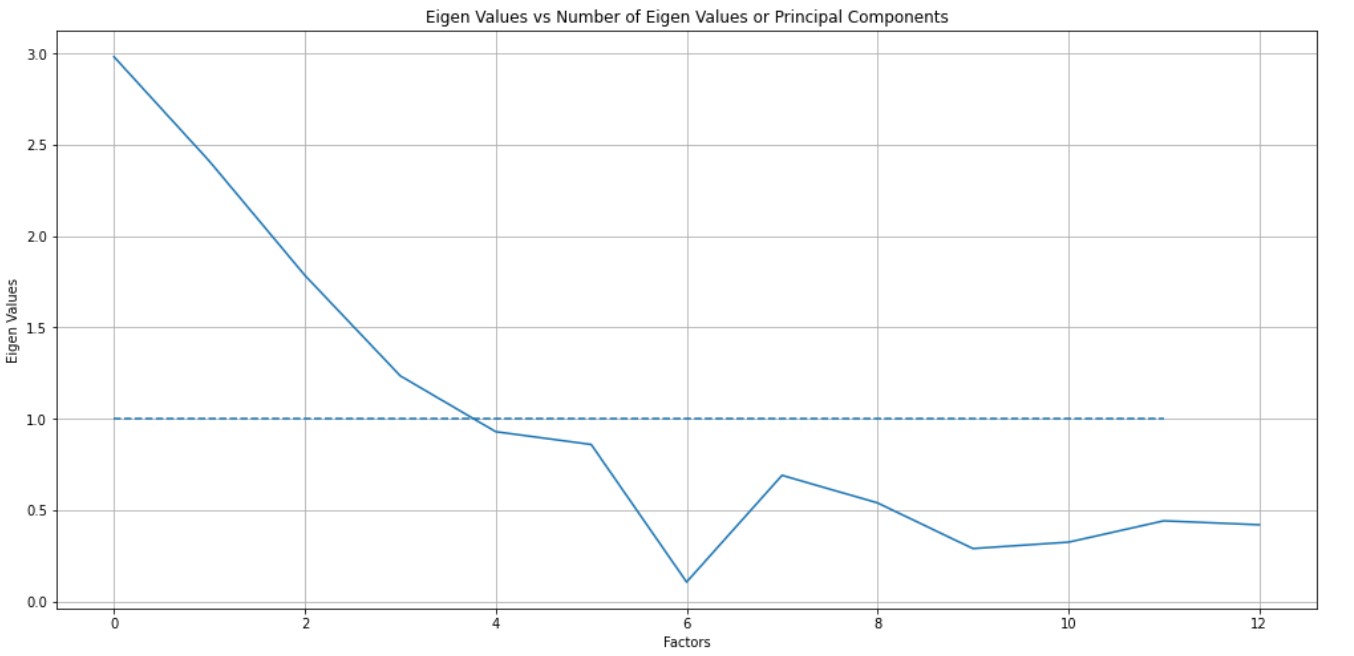
**2.4) Perform Principal Component Analysis (on the predictor variables) and extract the Principal Components. Comment on the reason behind choosing the number of Principal Components.**

* Performing Principal Component Analysis on the data and reduce the predictor variables to a suitable number of dimensions.
* Scaling of the data

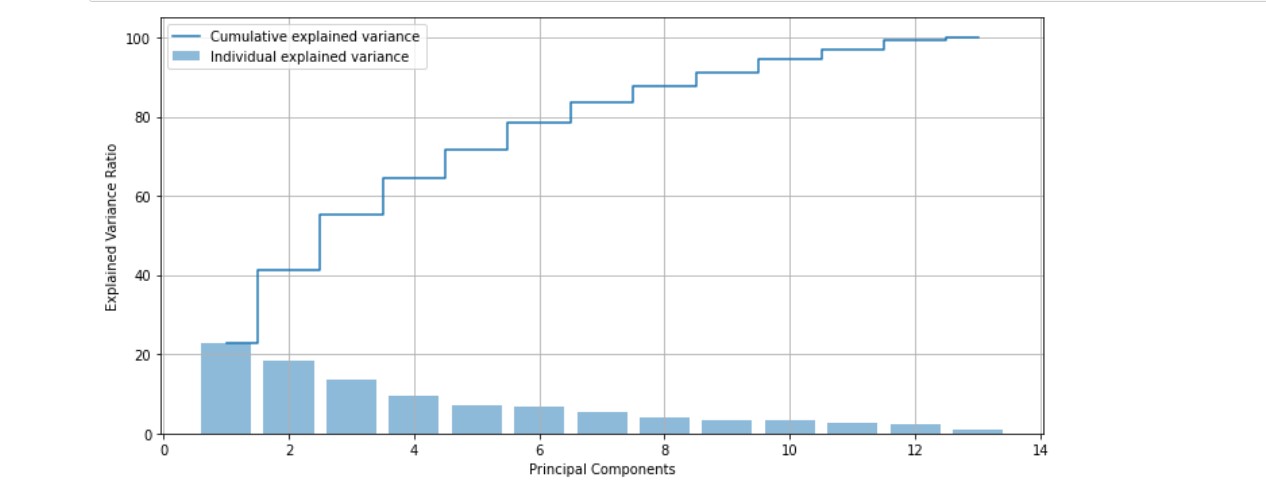


* Eigen Values

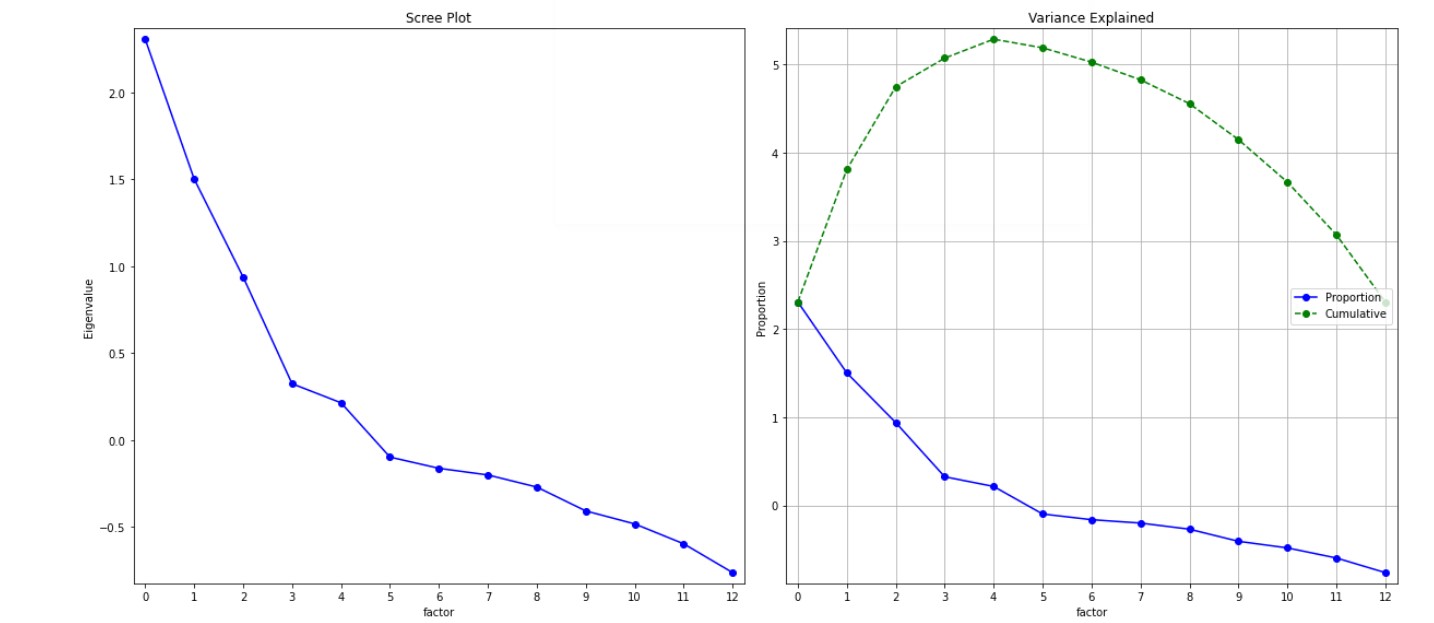




**Eigen Values vs Number of Eigen Values or Principal Components**

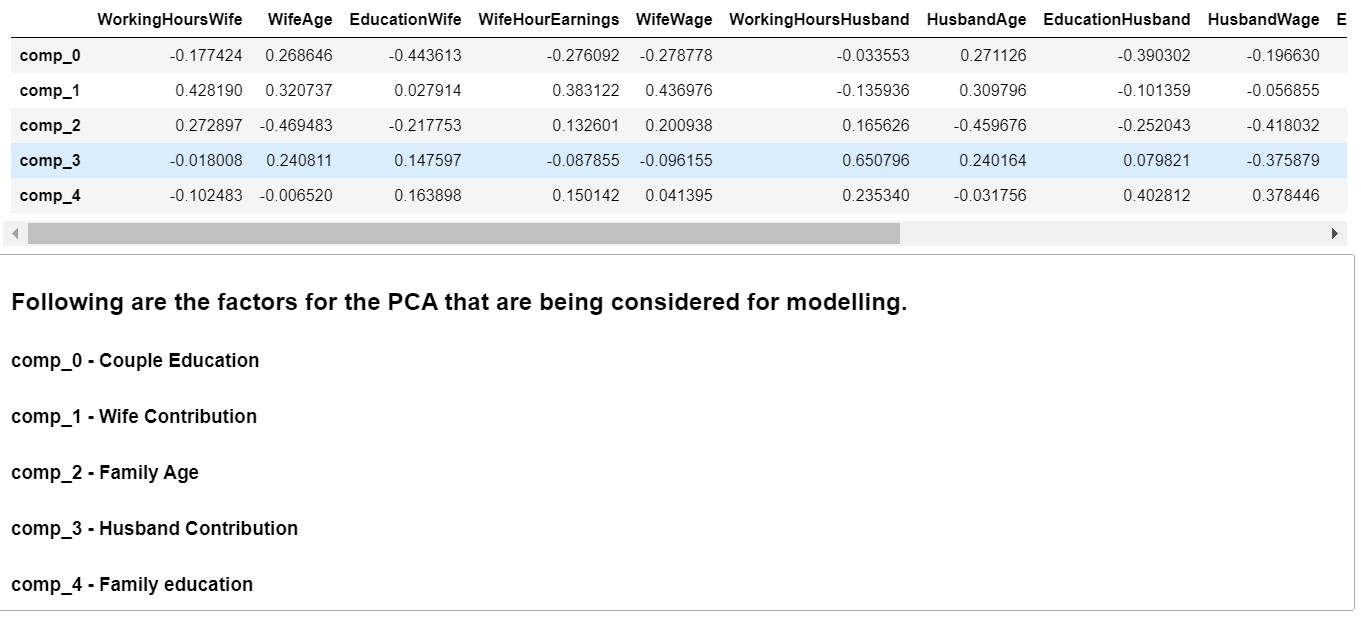


**Step Plot of cumulative explained area**

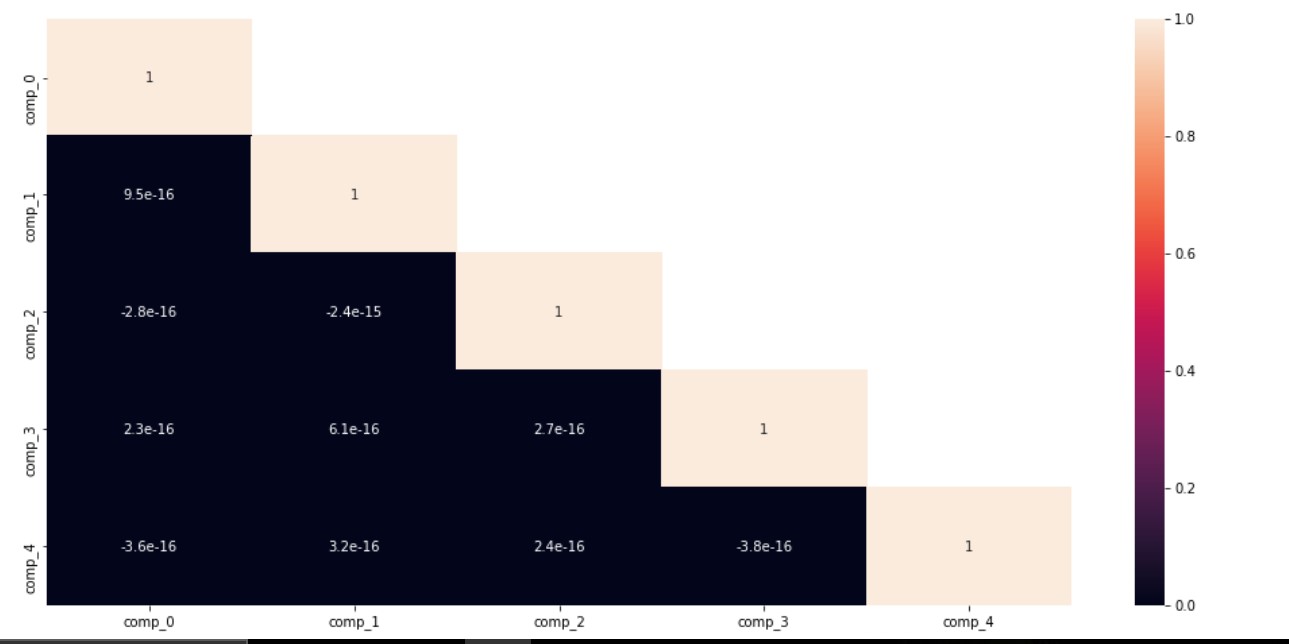


**Factors considerted : 5**

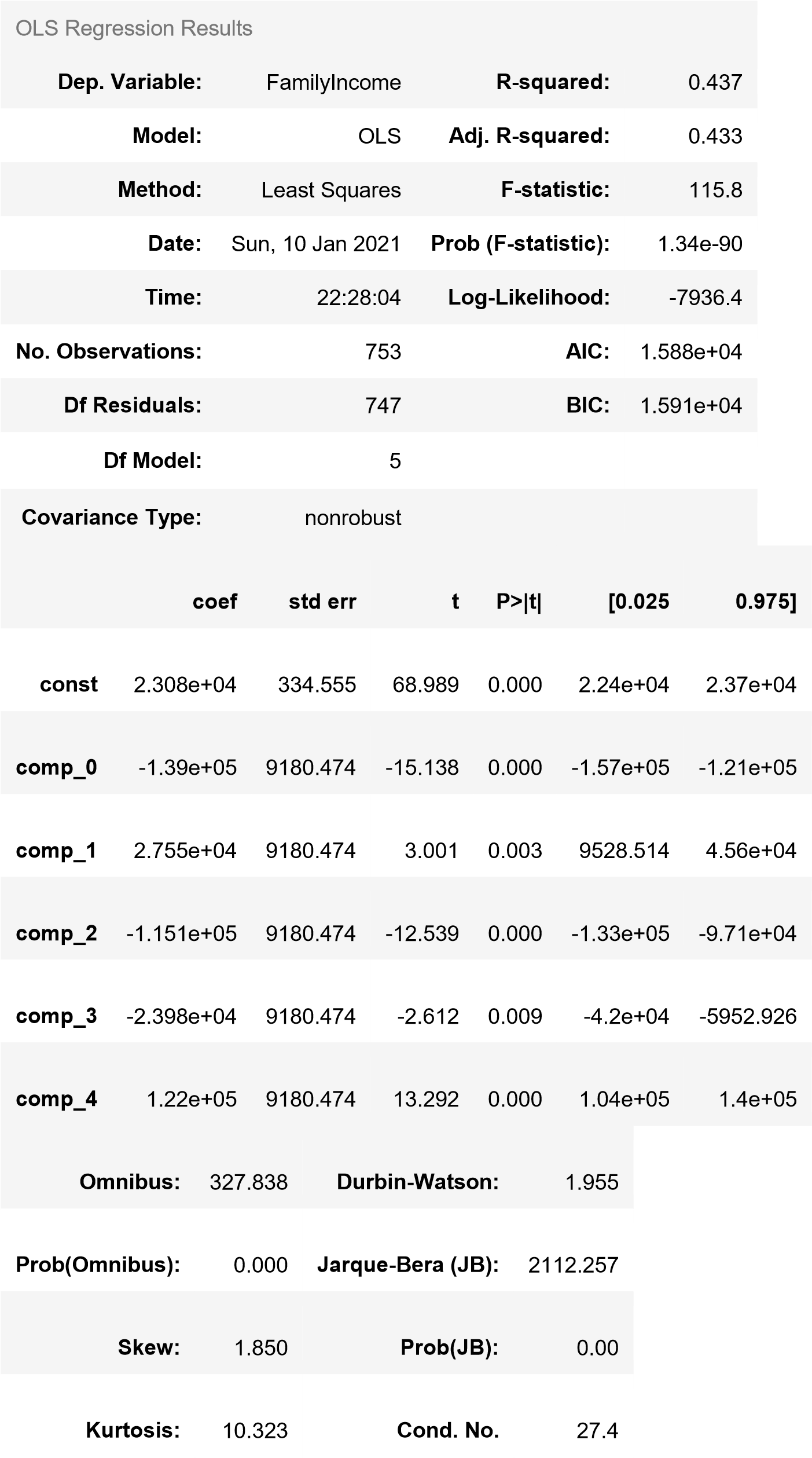
**As per scree plot's golden rule we should consider 3 factors for PCA but we are taking 5 factors because 3 factors are not covering most of the variance as compared to 5 factors.**



**Heat map of the factors considered for the PCA**



|  |  |
| --- | --- |
| **2.5) Perform Multiple Linear Regression with 'FamilyIncome' as the dependent variable** | |
| **and the Principal Components extracted as the independent variables.** |  |



**The Rsq value is 0.43660742366773275**

**Interpretation :**

*comp\_0 - Couple Education comp\_1 - Wife Contribution comp\_2 - Family Age comp\_3 - Husband Contribution comp\_4 - Family education*

*Wife Contribution and Husband Contribution are one of the principal factors affecting the family income. Second comes other factors which are family education and Couple education.*

*R-square value is 0.437 and the adjusted R\_square value is 0.433 is consequence of the outliers present in the factors responsible for this model leading to this R-Square value.*

|  |  |
| --- | --- |
| **2.6) Comment on the Model thus built using the Principal Components and with** | |
| **'FamilyIncome'.** |  |

The RMSE values thus obtained :-

*The Root Mean Square Error (RMSE) of the model with all the predictor variables is*

***6615.602163536179***

*The Root Mean Square Error (RMSE) of the model whose predictor variables have be en reduced by PCA is* ***9143.824784099732***

**Analysis:** After the PCA the adjusted R-Square value went down to 0.433 which is little low as compared to standards which implies model has not improved which means we have to im prove the model more as there may be some other factors which are missing like husband's ex perience, Husband-hours-earning etc and also because of the outliers present.

**2.7) Mention the business implication and interpretation of the models.**

**Business Implications:**

* The factors are affecting the family income is education of every member in the family and wages which in turn tell you that higher the education higher the family income and better life upon consi deration of outliers in family income.
* Those with education below graduation are in middle class income range as this implies with the p arent's education and also this has been observed educated people choose educated spouses and hus band wages are directly related to family income saying husbands are the bread winners for most o f the families and next comes the wives.
* Also, parents of Wives are contributing somewhat or the other by their education.

**\*\*\*\*\*\*\*\*\***