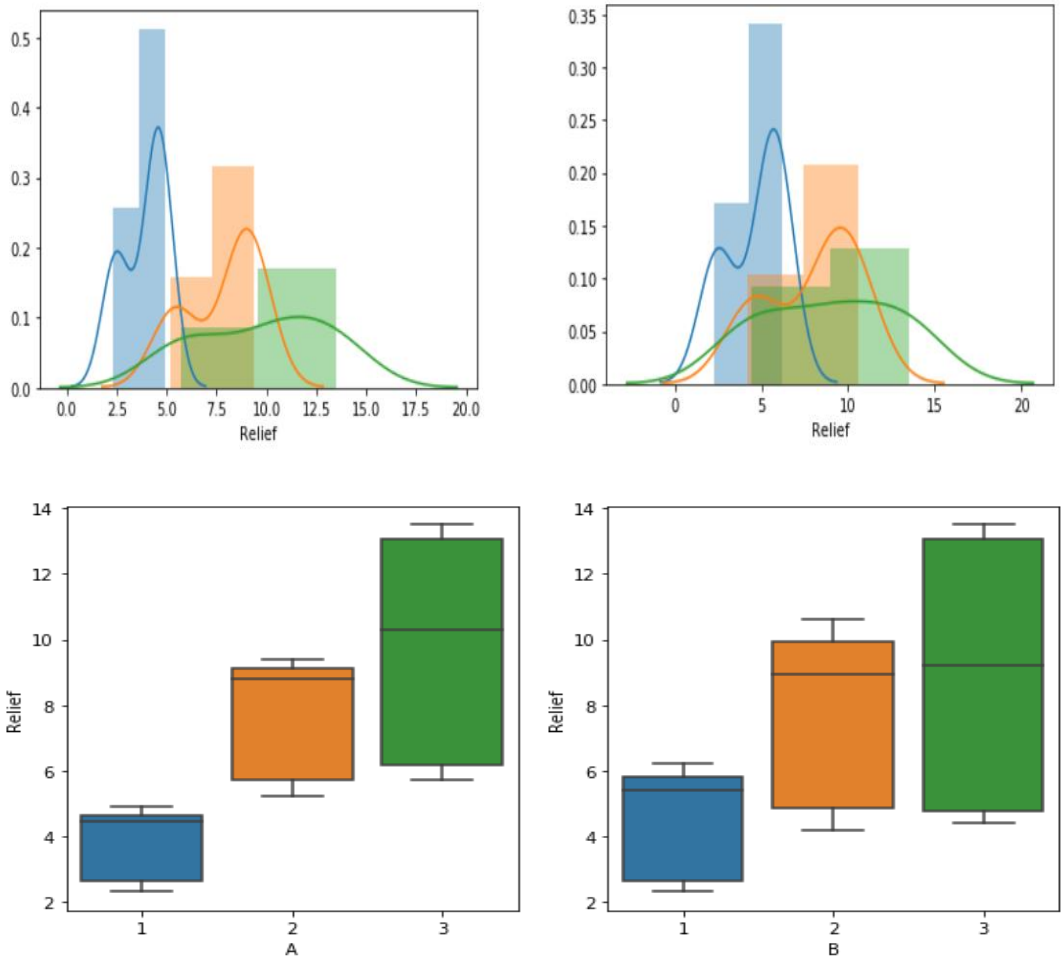
Table of Contents

ANOVA

Understanding Case study:

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 (1,2,3) 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
As per data available, there are two ingredients (A &B) which are independent variables and the Relief factor which is the dependent variable.
Before performing Anova, it is important to check two assumptions:

1) Data is normally distributed
2) Equality of Variance



1. From graph, for both ingredient 'A' and 'B', the in-between and with-in group variance seem to be away from each other.
2. From above graph there is no visible normality in the data.
3. Also from **Shapiro test**, it is found that data is not normally distributed for both A and B (working in the python file)

4. For equality of Variance, we visualised boxplot, but cannot clearly conclude equality of variance for a given ingredient.
5. Visually it sems that level 1 and level 3 of A and has similar variance as level 1 and level 3 of B respectively; level 2 of A and B does not have similarity in variance but median value seems to be same or very close.
6. For statistical check of equality of variance, we performed **Levene's test** and it holds true at a 1% level of significance. Since, p-value is greater than 0.01, so accept *Ho*. Thus, we conclude that there is homogeneity of variance across groups.
7. Since, ANOVA is robust against normality if the assumption of equality of variance is met, we can proceed for ANOVA.

1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually

Hypothesis Test for one-way ANOVA for variable A

Ho: Means of relief obtained from all the levels of ingredient A are equal : $\mu1=\mu2=\mu3$
Ha: At least one of the mean of relief obtained from different levels of ingredient A is not equal

Hypothesis Test for one-way ANOVA for variable B

Ho: Means of relief obtained from all the levels of ingredient B are equal : $\mu1=\mu2=\mu3$
Ha: At least one of the mean of relief obtained from different levels of ingredient B is not equal

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.

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.02	110.010000	23.465387	4.578242e-07
Residual	33.0	154.71	4.688182	NaN	NaN

1. The corresponding p-value is less than 0.05, so we fail to accept the null hypothesis *Ho*. So, we can say that at mean of Relief caused by ingredient 'A' is different for at least one of the levels (1,2,3) of ingredient 'A'.
2. The variance between the relief factor caused by levels of ingredient is 23.47 times the variance within the treatment. There is a variance because there are different treatments and there is variance within each treatment. So, here F value is large, we can say that chance of being more than 23.47 is only 10^{-07}

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_sq	F	PR(>F)
C(B)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

1. The corresponding p-value is less than 0.05, so we fail to accept the null hypothesis *Ho*. So, we can conclude that at least one of the mean of relief caused by different level (1,2,3) of ingredient 'B' is different.
2. The variance between the relief factor caused by levels of ingredient is 8.13 times the variance within the treatment. There is a variance because there are different treatments and there is variance within each treatment. So, here F value is large, we can say that chance of being more than 8.13 is only 0.00135.

1.4) Analyse the effects of one variable on another with the help of an interaction plot. What is the interaction between the two treatments?

fig1

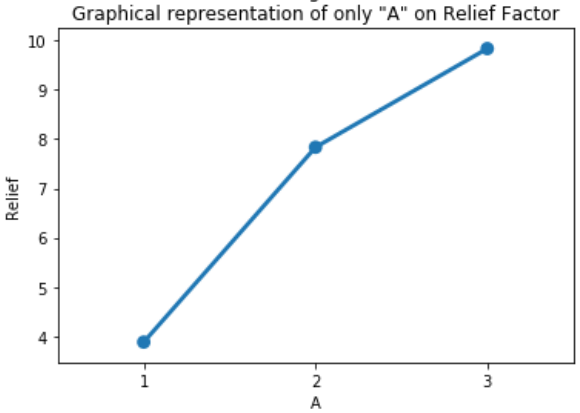
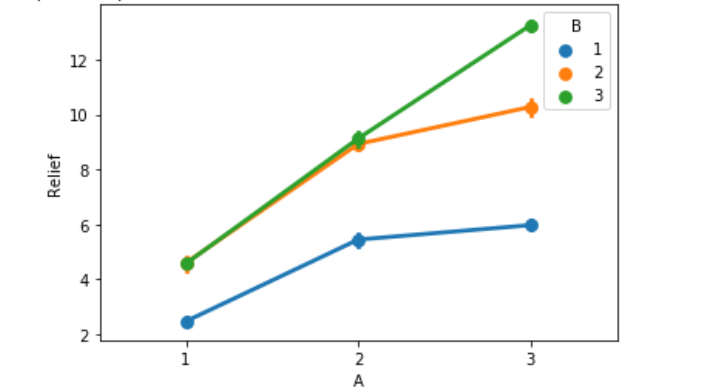


fig 2



1. From the visual in fig 2, there seems to be an interaction effect between two treatments. It means ingredients B when added to ingredient A in making the compound, it changes the Relief Factor / effect of medicine.
2. With the help of 2 graphs, we can say that ingredient 'A' has its own effect on the relief factor of hay fever but when 'B' is also added at three levels, the effect of 'A' and 'B' together changes the Relief factor from 'A' alone.
3. From fig2, when 1st level of ingredient A is mixed with 1st level of ingredient B, then least relief is felt
4. When 2nd level of ingredient A is mixed with 1st level of ingredient B, then relief factor increased and with third level of ingredient A and 1st level of B, it becomes a bit better.
5. But we can see that when combination is like A_1+ B_2 & A_1 + B_3 then the effects are better and same.
6. Same goes for combination A_2+ B_2 & A_2 + B_3, here effects are better than above combination but almost similar
7. Whereas combination of $\boldsymbol{A_3}$**+** $\boldsymbol{B_3}$ **gives the maximum Relief factor**.
8. However, this needs to be tested statistically, as visuals may not be conclusive for taking decision.

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

For two-way ANOVA, we can check the impact of ingredient 'A' and 'B' together on Relief factor with and without interaction effect

ANOVA without interaction: We will have two hypotheses for two-way ANOVA:

H0_A: For all three levels (1,2,3) of ingredient 'A', mean of 'Relief factor' is equal
H1_A: For all three levels (1,2,3) of ingredient 'A', at least one mean of 'Relief factor' is different

H0_B: For all three levels (1,2,3) of ingredient 'B', mean of 'Relief factor' is equal
H1_B: For all three levels (1,2,3) of ingredient 'B', at least one mean of 'Relief factor' is different

ANOVA table without interaction summary

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.02	110.010000	109.832850	8.514029e-15
C(B)	2.0	123.66	61.830000	61.730435	1.546749e-11
Residual	31.0	31.05	1.001613	NaN	NaN

1. From the above table, p-values of 'A' is less than .05, which means null hypothesis for 'A' is rejected. Hence, we can conclude that Relief Factor is a function of 'A' or mean of Relief Factor changes with changes in levels of 'A'
2. It can also be noted that F-stat for 'A' shows that the between-level variance of ingredient 'A' is 109.8 times that variance with-in each level of 'A'. This value is high and can be observed in fig 1 on pg 2
3. From the above table, p-values of 'B' is less than .05, which means null hypothesis for 'B' is also rejected. Hence, we can conclude that Relief Factor is a function of 'B' or mean of Relief Factor changes with changes in levels of 'B'
4. The F-stat for 'B' , however is 61.7 , showing the between-level variance of ingredient 'B' is 61.7 times that variance with-in each level of 'B'. Same can be observed in fig 2 on pg 2

ANOVA with interaction: We will have three hypotheses for two-way ANOVA and one for interaction:

H0_A: For all three levels (1,2,3) of ingredient 'A', mean of 'Relief factor' is equal
H1_A: For all three levels (1,2,3) of ingredient 'A', at least one mean of 'Relief factor' is different

H0_B: For all three levels (1,2,3) of ingredient 'B', mean of 'Relief factor' is equal
H1_B: For all three levels (1,2,3) of ingredient 'B', at least one mean of 'Relief factor' is different

H0_int: There is no interaction
H1_int: There is interaction

ANOVA table with interaction summary

	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

1. The p-value of both 'A' and 'B' goes further down when tested with interaction effect, which means there is even more evidence to reject Null Hypothesis for the two ingredients. Hence, for different levels of 'A' and different levels of 'B', mean Relief factor will be different.
2. The p-value for the interaction effect is also less than 0.05, which is enough for rejecting the Null Hypothesis, i.e. **there is an interaction effect**.
3. It also means that at least one mean among 9 different combination of different levels of ingredients A & B is different and there is interaction between variable A & B and it affects the relief level.

1.6) Mention the business implications of performing ANOVA for this particular case study.

1. The goal of studying data is to understand a treatment effect for providing relief from hay fever. A research laboratory is trying to develop a new compound for the same.
2. ANOVA test is generally used to determine the influence that independent variables have on the dependent variable in a study. Here, it is for checking influence of Variable A & B (independent) on Relief factor (dependent). ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.
3. When one-way anova performed for ingredient A & B individually, it reflected that in A there is significant difference of effect at different level of component used, B also reflected the same. But comparatively, A is showing larger difference than B.
4. Two-way anova compares the mean differences between two particular sets of treatment
5. When we are performing two-way anova we are trying to identify how relief factors are affected by ingredient A, B and interaction of A & B individually. With the help of above study, we come to know that when ingredient B is added with A, it influence the effect of medicine. That is why we can see changes in two-way anova results. The p-value comes more closer to 0 after two-way anova treatment. And when interaction is also added and then anova computed it calibrated the result. So, when we adjust the relief for interaction of A&B, interaction itself shows significant effect on relief factor and also leads to increase in the efficiency of ingredient A & B.
6. Overall, both ingredients show impact on Relief factor individually , although, ingredient A have stronger effect then B when measured individually.
7. When both are used together, both have their impact on Relief factor individually.
8. Also, the interaction effect statistics show that there is enough evidence to prove that B enhances the effect of A on Relief factor
9. Hence, from **business perspective**, the medicine shows to be more effective with A and B combined. In individual, contribution, A has stronger results on Relief factor than B , but not by too much.

EDA

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

	Working HoursWife	Wife Wage	Working Hours Husband	Husband Wage	Unemployment Rate	Family Income
count	753.000000	753.000000	753.000000	753.000000	753.000000	753.000000
mean	740.576361	1.849734	2267.270916	7.482179	8.623506	23080.594954
std	871.314216	2.419887	595.56`+6649	4.230559	3.114934	12190.202026
min	0.000000	0.000000	175.000000	0.412100	3.000000	1500.000000
25%	0.000000	0.000000	1928.000000	4.788300	7.500000	15428.000000
50%	288.000000	0.000000	2164.000000	6.975800	7.500000	20880.000000
75%	1516.000000	3.580000	2553.000000	9.166700	11.000000	28200.000000
max	4950.000000	9.980000	5010.000000	40.509000	14.000000	96000.000000

Working Hours Wife:

Mean hours of working in wives is 740.58 and is quite high from median value which indicates the right skewness in data.
Most of the data fall in the between 0 to 1612 hours which is far away from 4950, the max value. It can be looked into if there could be outliers on the upper side in this data.

Working Hours Husband:

The mean and median working hours are close values, indicating variable is more or less symmetrically distributed.
Most husband have working hours range of 1672-2863 hours which is far away from 5010 i.e the maximum value and 175 hrs i.e min value. Hence, there must be both upper and lower outliers in this variable.
It is important to note that at least 25% (i.e about 188) of husband work for more than 2550 hrs.

Wage Husband:

Where the range of wages is from 0.4 to 40, the mean wage is only 7.5, which signifies highly right skewed data.
The wage distribution for the first 75% is between 0.4 to 9.1 and for rest 25% wage distribution is between 9.1 to 40
Most of the husband fall in the wage range of 3.25 – 11.7 i.e. one STD from mean.
An interesting noting can be that three-fourth of population received one-fourth wage distribution and one-fourth population receives three-fourth wage distribution.

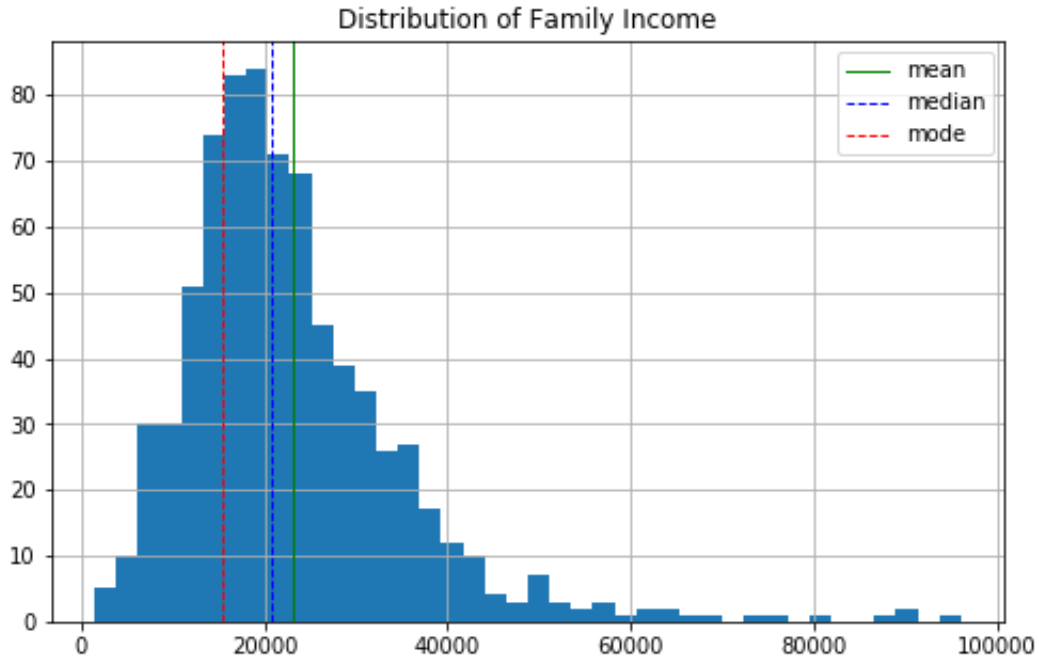
Wife Wage:

75% of the wife are not earning any wages.
Wage range for wife is very short at only 9.98 as compared to husband.
Where the range of wages start from 0-9.98, the mean wage of a wife is only 1.85
Here only 25% of the wife's are earning wages distributed from 3.58 - 9.98. So, it would not be wrong if we say that actual mean earned wages of a wife are higher than 1.85, as we can exclude the 75% (i.e. approx. 565) wife from the earning population.

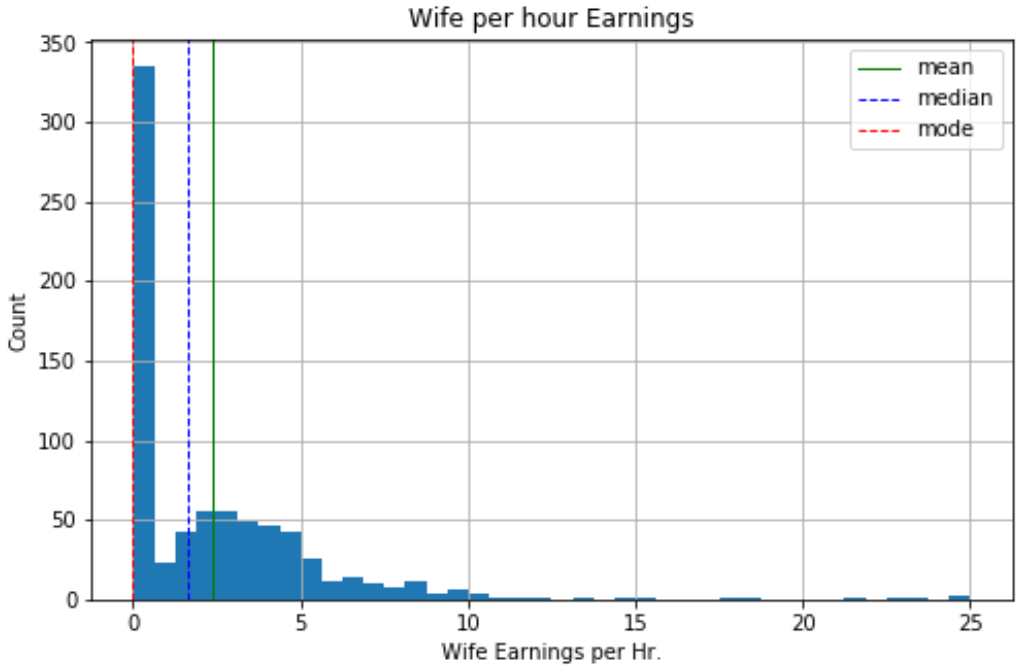
Family Income

Family income range is from 1500 – 96000, i.e. 94500, shows highly dispersed data.
Mean is 23080 and median is 20880 data is skewed to the right.
Also, 50% of families (approx 376) earn a family oncome in the range of 28200-15428, which is approx symmetrical around median value of 20880.
However, above 75% quartile, distribution of family income is dispersed widely in the from 28200 to 96000.
Above, shows that 75% of the population has a family income distribution of only 26700 (28200-1500) which is approx. 29% (26700/96000*100) of total range.
However, 25% of population shares income distribution of 67800 (96000-28200) which is 71% (67800/96000*100) of income distribution.

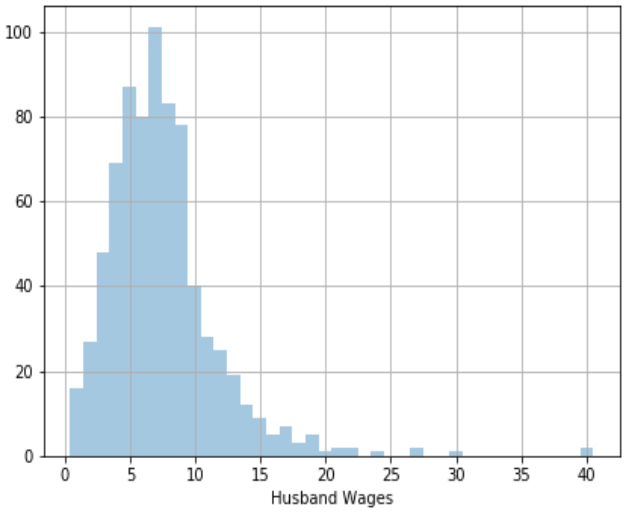
UNIVARIATE ANALYSIS (HISTOGRAM)

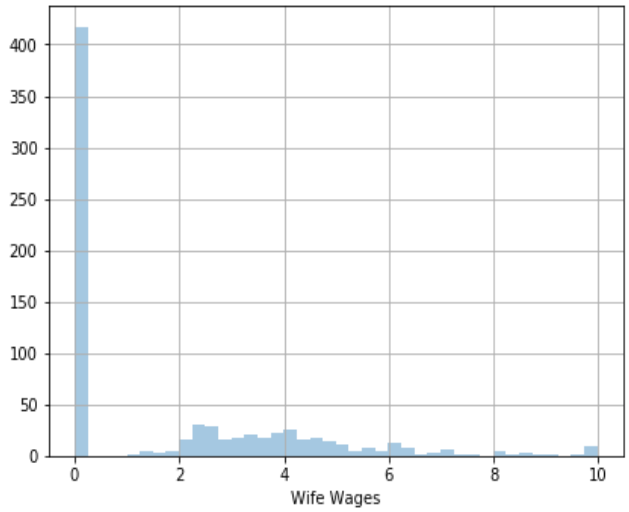


1. It can be seen that maximum number of families fall in the approx. 18000 to 20000.
2. We can see Mean is greater than median.
3. Plot shows data is highly skewed on right.
4. There are very few families with family income more than 42000.

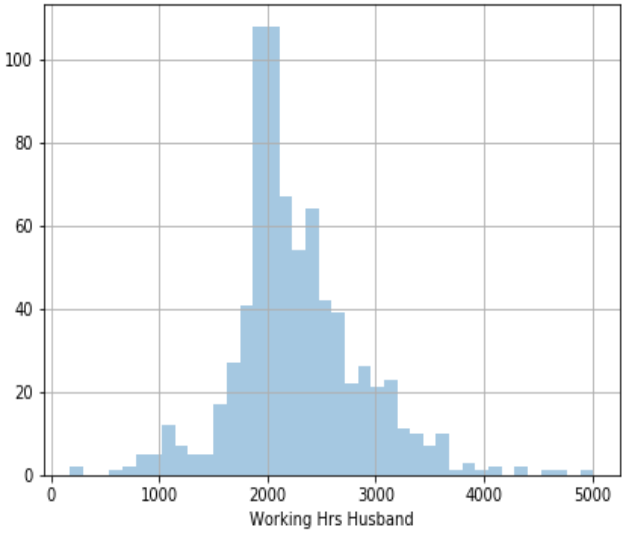


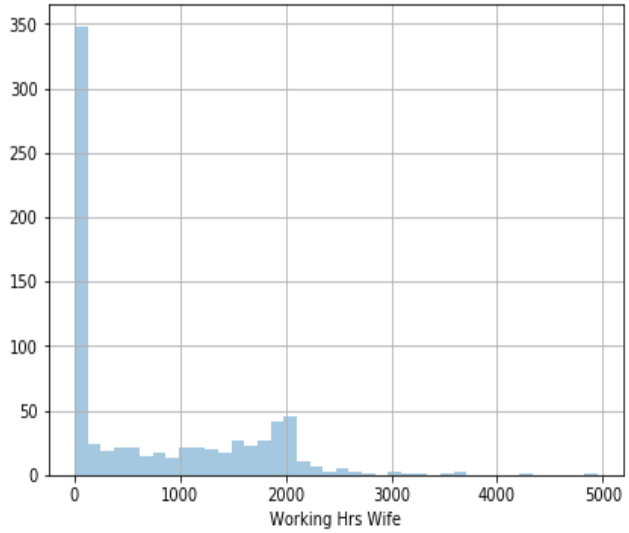
1. We can see that close to 50% wives are not earning
2. Data is skewed to the right and there are several outliers





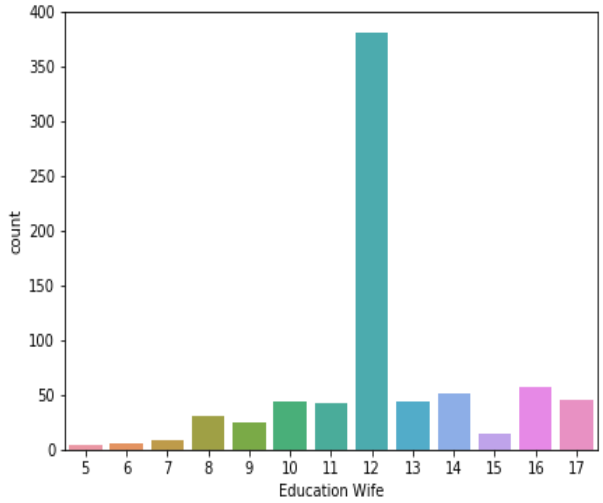
1. Comparing the wages distribution of husband and wife it is clear that the range of wages for husband is approx. four times as compared to wife
2. 100% husbands earn wages, however, more than 400 wife do not earn any wages
3. It can be seen that maximum number of husband wages fall in the range of 7-8 and the next mode can be seen at 5-6 interval, however, for wife wages the curve is quite flatter and maximum number of wife earn wage in the range of 2.5-3.0 only.

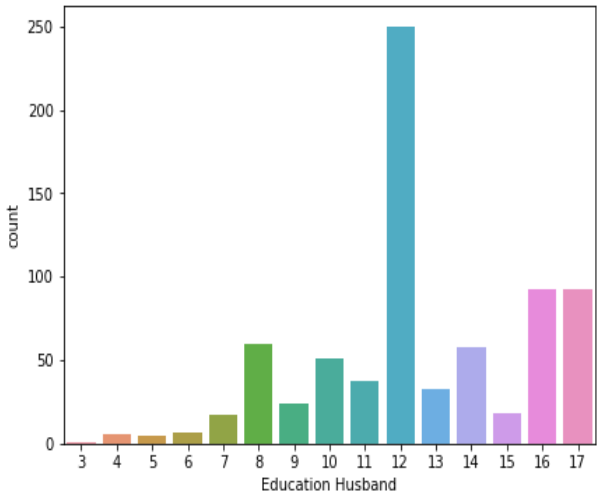




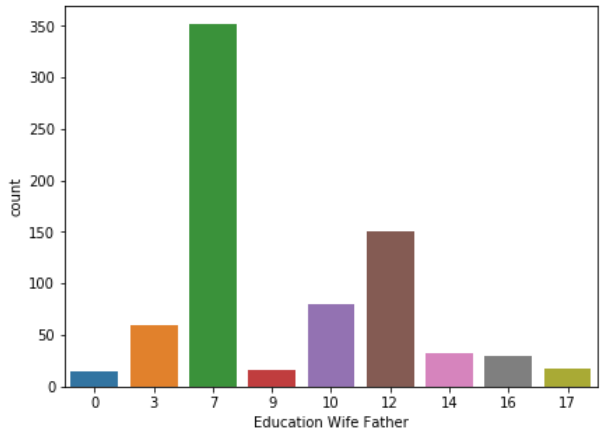
1. Comparing the working hours of husbands and wife, it is visible that close to 50% wife are not working. The maximum working hours for a wife is between 1900-2100 and that applies to approx. 50 families
2. For husband, the wage distribution is comparatively normally distributed with a peak working hour in the range of approx. 1900-2100 and is applicable to more than 150 famiies.
3. It can be noted that for both husband and wife, modal working hours are in aprox same range of 1900-2100, however, only wife of only 50 families fall in the range and husband of more than 100 families fall in this range.

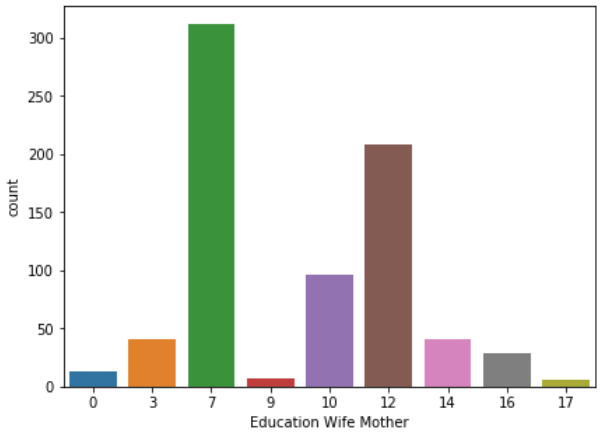
UNIVARIATE ANALYSIS (BAR GRAPHS)





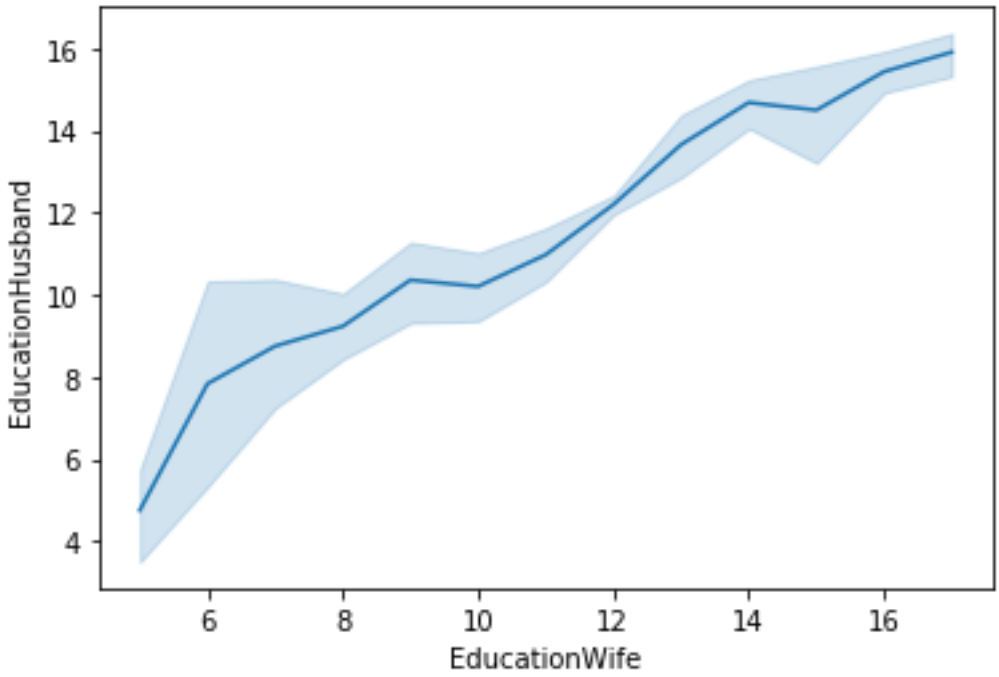
1. On the first view, it is visible that popular level of highest education is 12 years for both wife and husband.
2. More than 50% wife (about in 375 families) have an education of 12 years and good to see that more than 85% of wife are educated for 12 or more years.
3. Similarly, majority husbands (about in 250 families) also have an education of 12 years. Education years above and below are not even close to 100 for higher or lower years of education.
4. Except for 12 years, most other levels number of husbands is more than number of wives for the particular level of education.



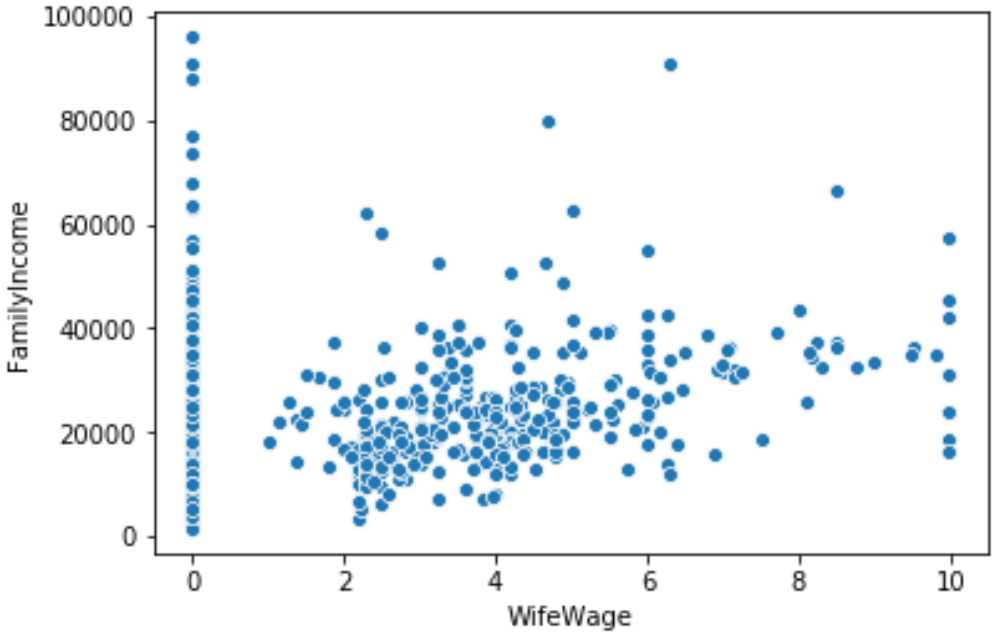


1. For wife's father and mother both, it can be noted that most popular level of education is 7 years and second most is 12 years, which can be seen in the next generation in both husband and wife trend of education.

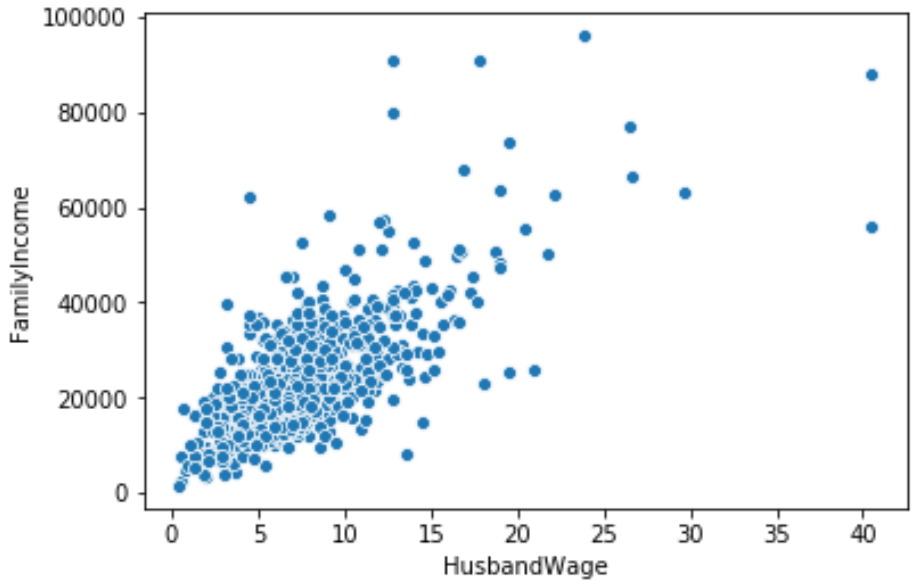
BIVARIATE ANALYSIS (SCATTER PLOTS)



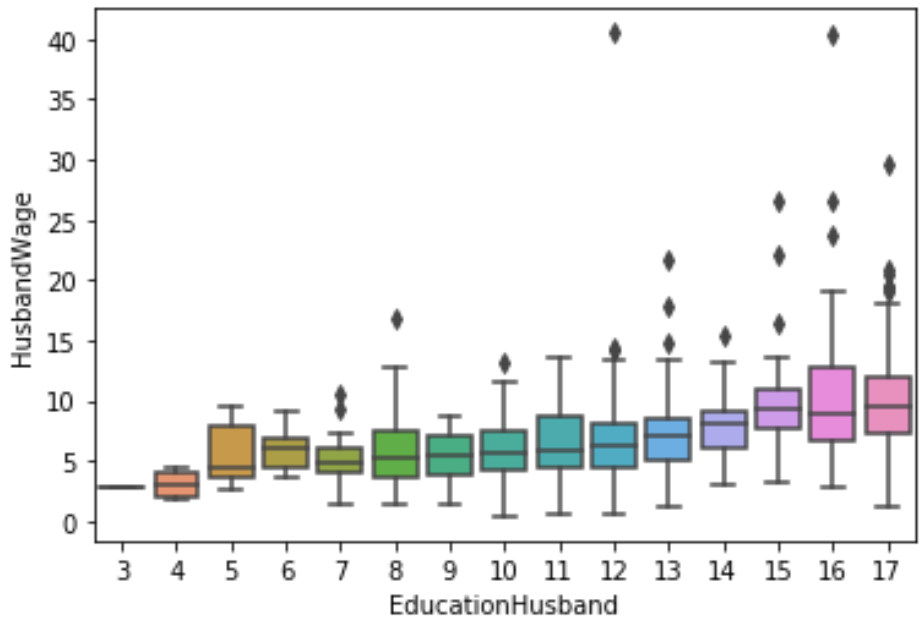
1. We see a strong correlation between husband age and wife age



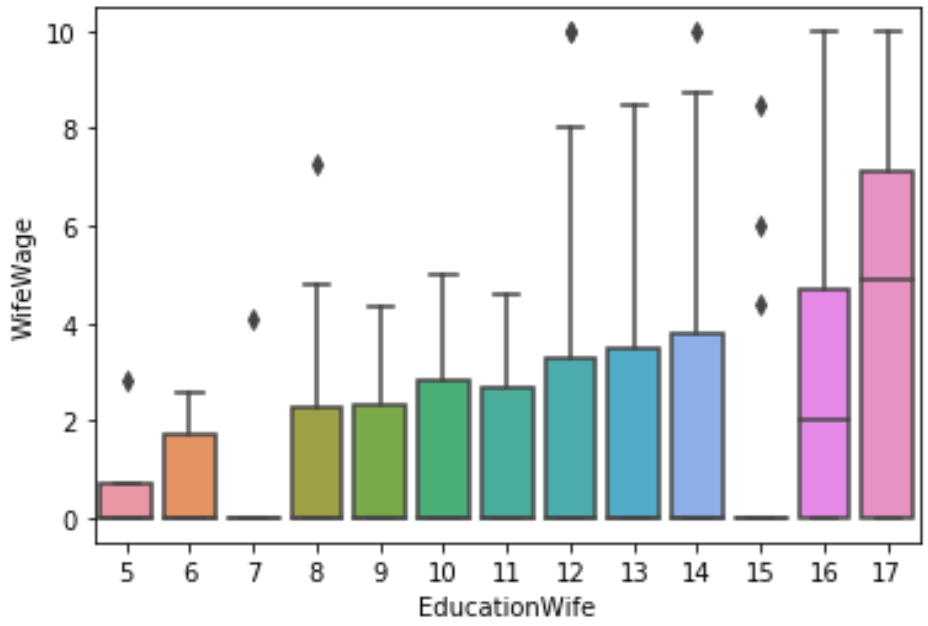
1. From the pattern, it is visible that even though wifes wage is increasing the spread of family income remain consistent. Hence, the data does not show any correlation between family income and wifes wage.



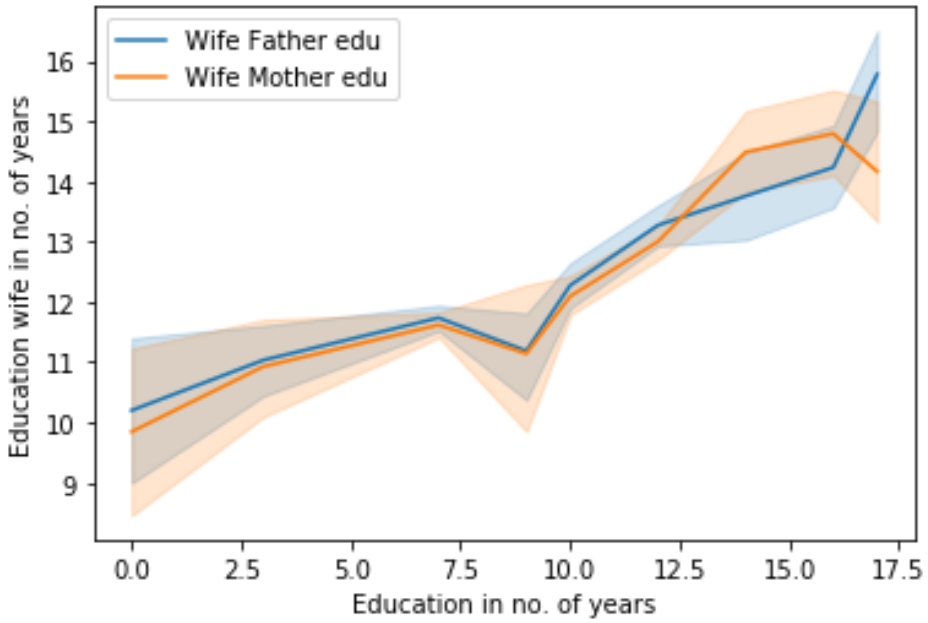
1. There is a strong positive pattern above showing an increase in family income with the increase in husband wage, showing a strong correlation between two.



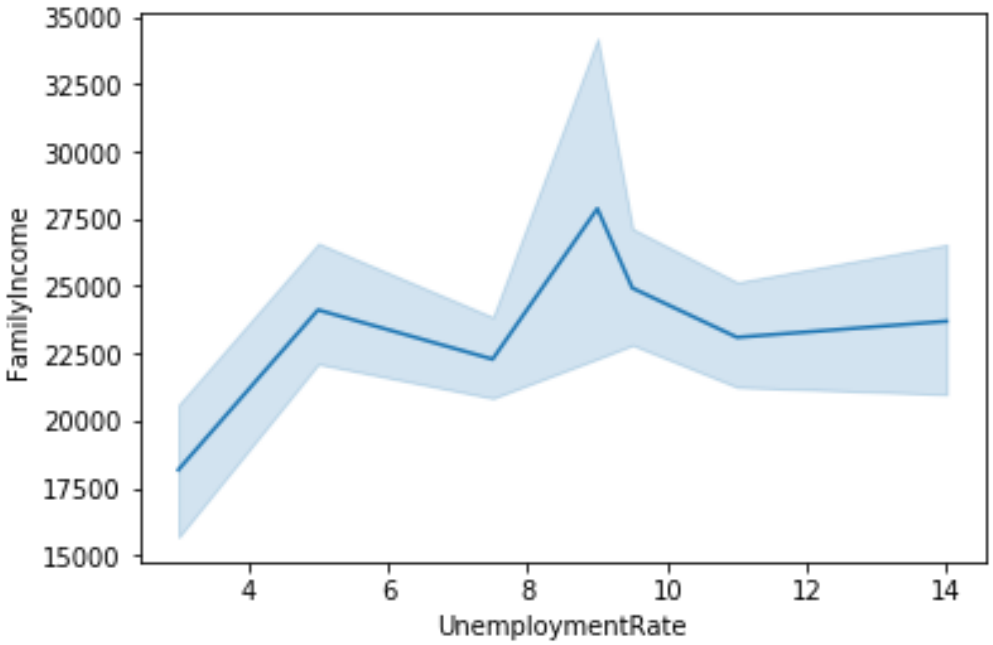
1. The wage of husband can be seen to have a increasing pattern with increasing years of education. Median wages of higher education are higher.
2. It is also visible that there are extreme upper outliers in wages, and they all are after the 12 or higher years of education, which means that having a education more than 12 or more does have impact on wages for husband.



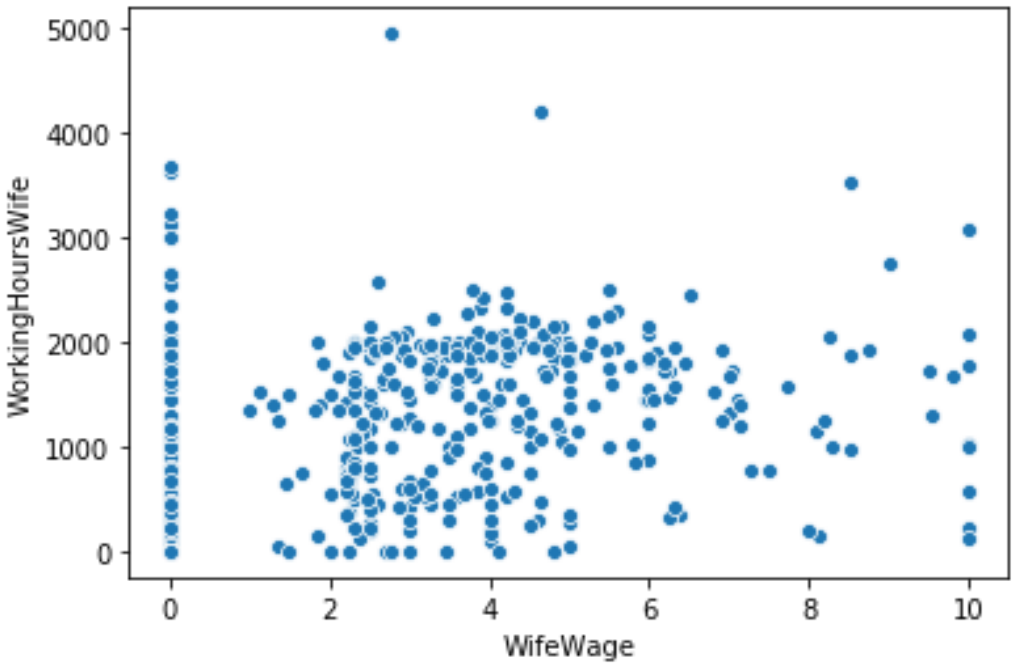
1. Wages of wife do not give any set trend on the basis of education. For most of the levels, median lies at zero, which means about 50% of wifes at most levels of education (except 16 years and 17 years) are not earning any wages.



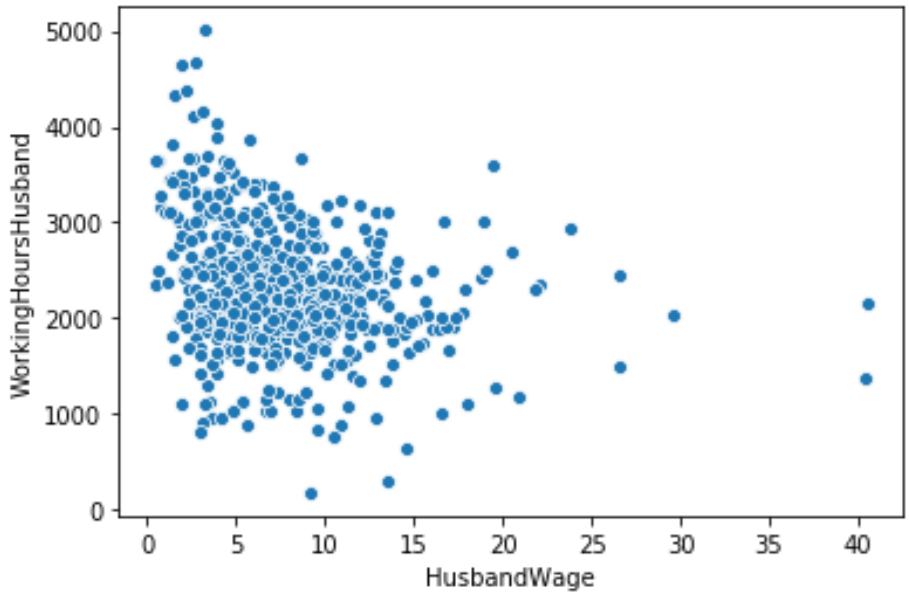
1. The pattern above is very interesting showing how the education level of wife is more or less positively related to the education level of both mother and father.
2. Also, data shows an overlapping trend on education level of mother and father.



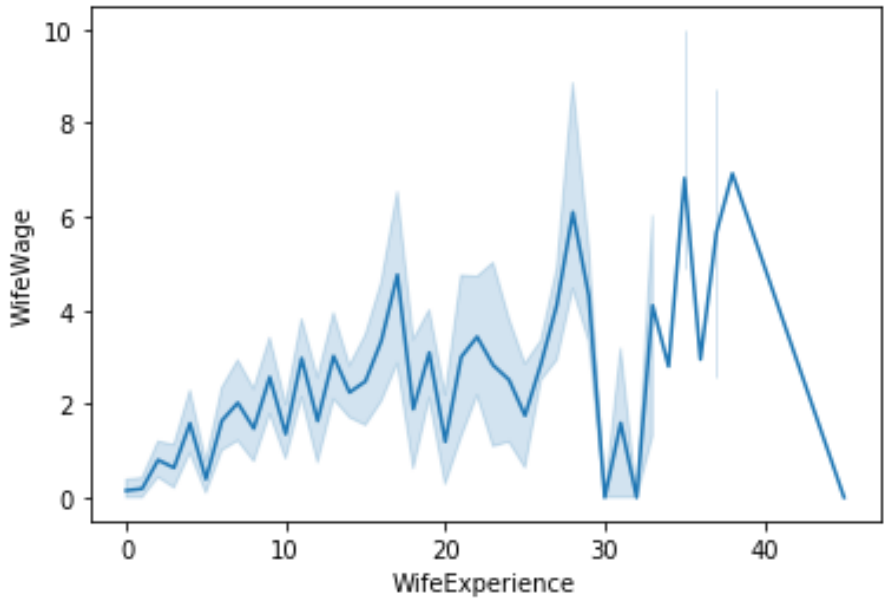
1. No impact of Unemployment Rate is visible on family income for the dataset.



1. From the data working hours of a wife do not have a conclusive relation with the working hours. It has a slight positive relation but not very effective.

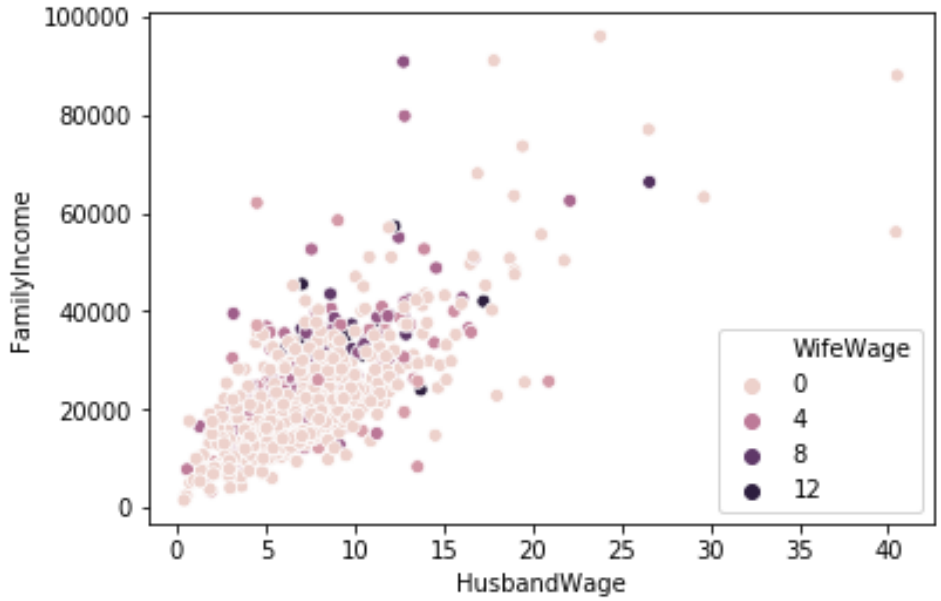


1. Contrastingly, for husband the working hours seem to be inversely related to wages. No. of hours are decreasing with increase in wage; however the impact is very low.

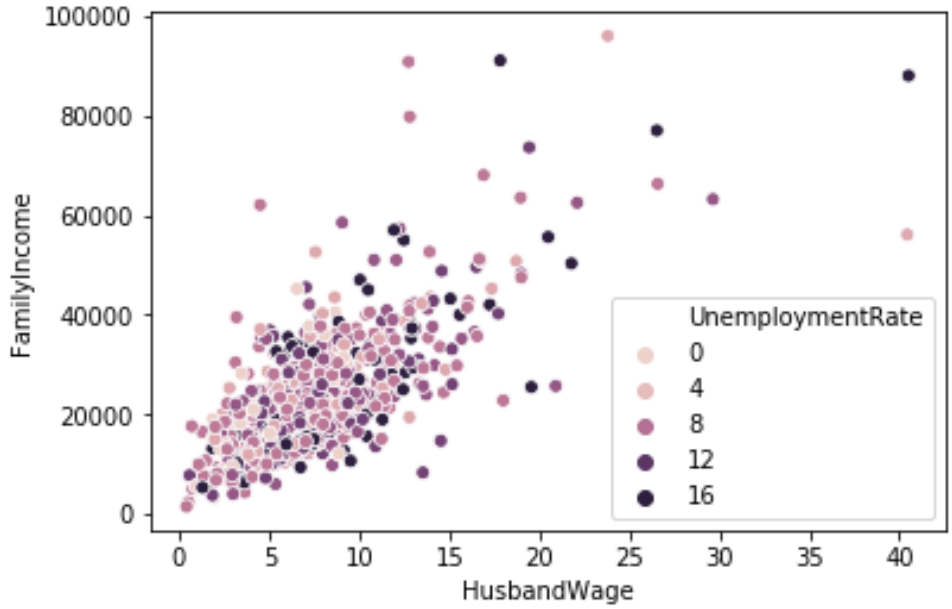


1. Surprisingly, there is no clear relationship between WifeWage and WifeExperience.

MULTIVARIATE ANALYSIS



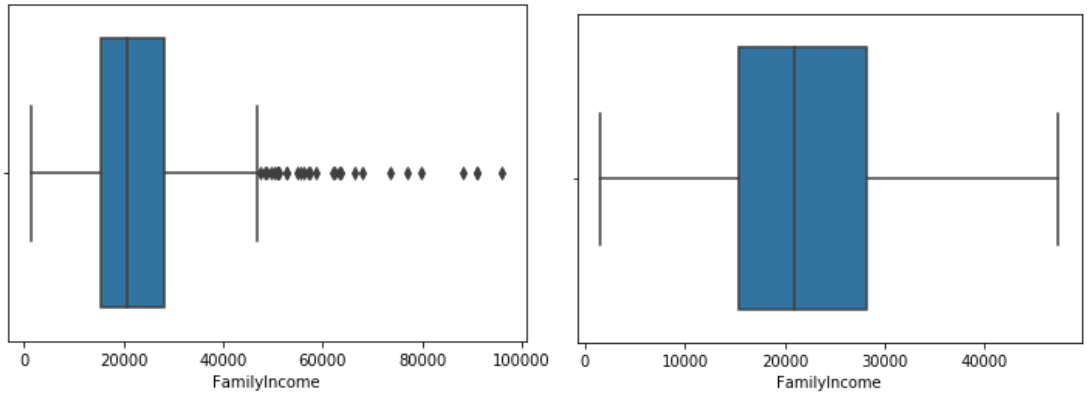
1. Plot show a strong positive correlation between husband wages and family income, and the hue shows the impact of wife income on family income.
2. It is interesting to see that only a very few numbers of wife wages effect the family income.

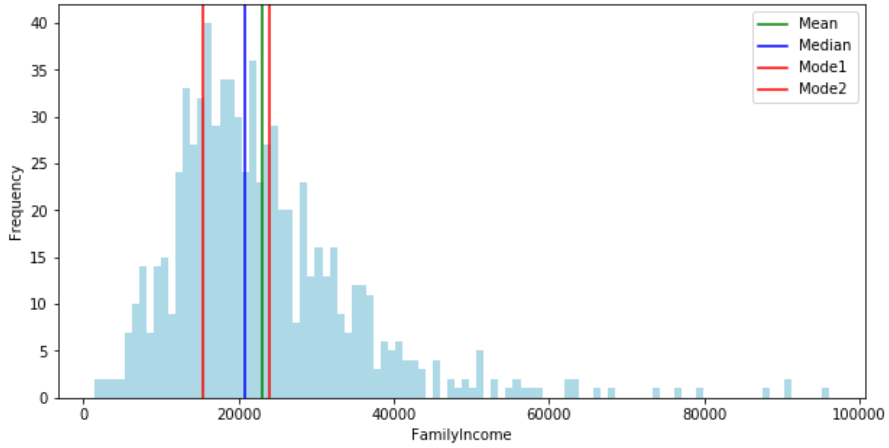


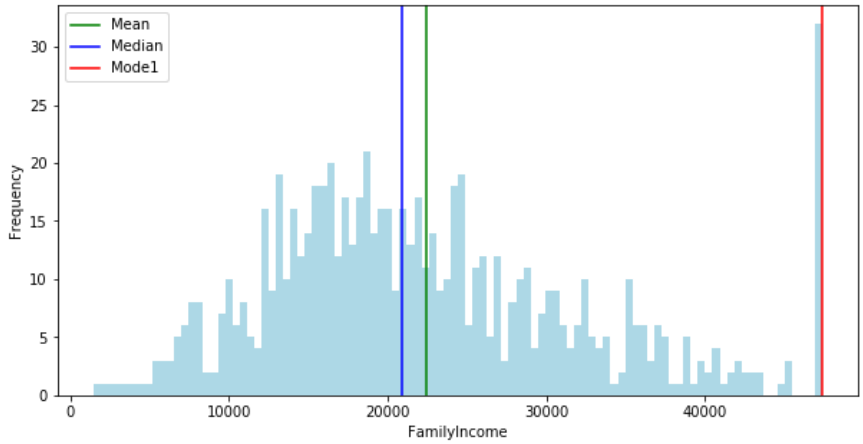
1. We plotted in above graph if EmploymentRate, has any impact on relationship between HusbandWages and FamilyIncome.
2. It can be observed there is no particular trend or distribution of EmploymentRate among HusbandWages, it can be seen all over the distribution and with different intensity level.
3. However, we see that most HusbandWages fall under lighter dots, meaning more husband wages are affected very lightly by EmploymentRate.S

OUTLIER TREATMENT BY REMOVING OULIERS

Below is the boxplot and histogram for family income showing family income distribution. We can see the data is highly skewed to the right. If we treat the outliers, by removal, we can see the change in plot no. 2, how the data becomes much more symmetrical.





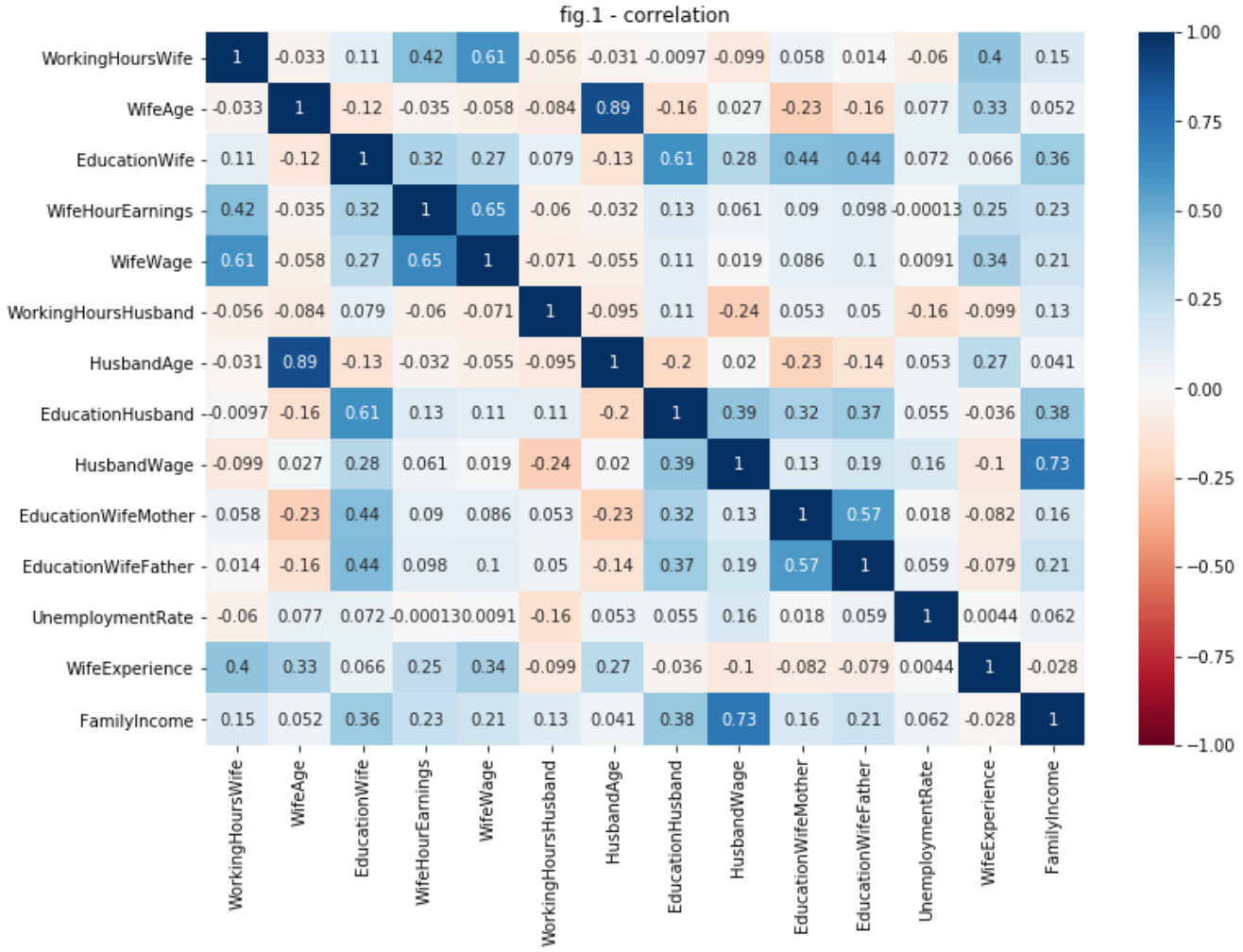


Though the data is a beautiful normal plot after outlier removal, but it can be noted that number of outliers here are too many and dropping them all will lead to loss of a lot of data. Hence, outlier removal not recommended here

COLLINEARITY

2.2) Is there evidence of multicollinearity? Showcase your analysis.

There is evidence of multicollinearity between a few variables, for some it is very strong too. But most independent variables do not show multicollinearity in the given data. Complete analysis is as follows.



From fig1 above, we can notice an additional observation between DV and IV. It is visible that only *HusbandWage and FamilyIncome* shows strong linear relationship of 0.73 and rest of the IVs show very slight correlation with FamilyIncome which is 0.38 and lower. This is an observation regarding *assumptions of Linear Regression*.

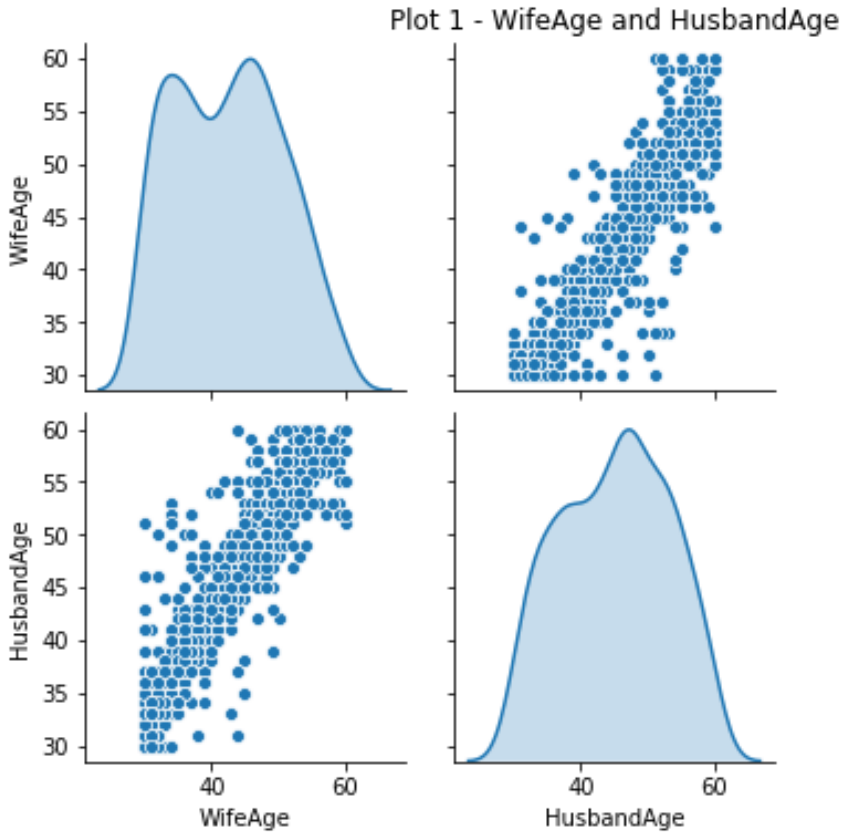
Now, we will check for multicollinearity:

In statistics, multicollinearity is a situation in which one independent variable (or we can say feature variable) can be linearly predicted from the others with a substantial degree of accuracy as they are correlated.

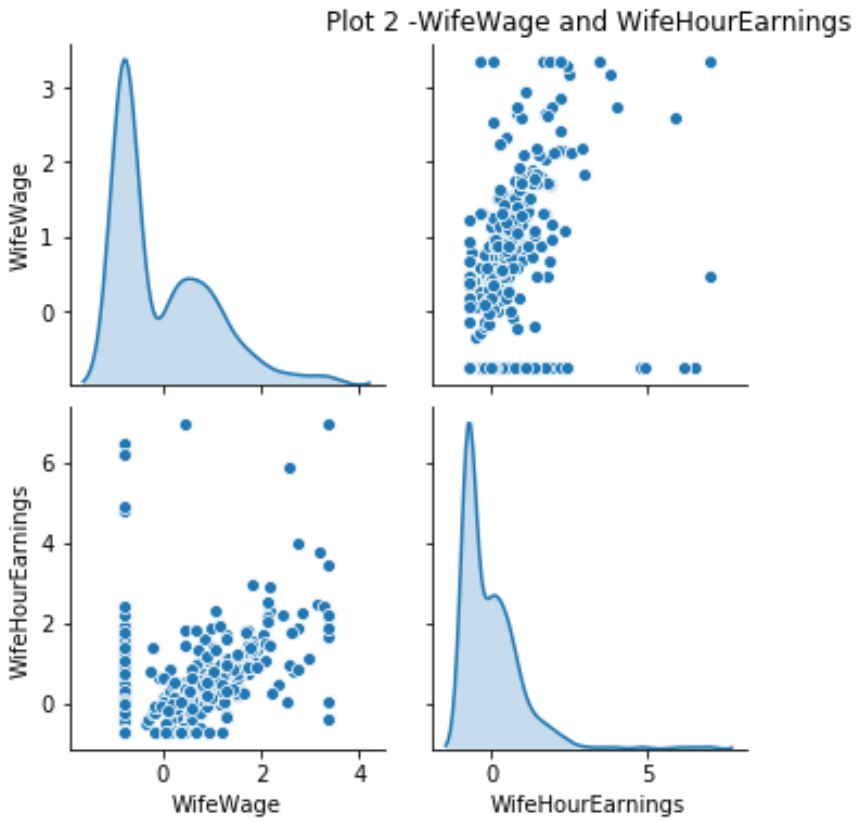
For checking the multicollinearity in data, we can refer to statistical display of values of correlation or correlation matrix or through heatmap or we can also use pair plot.

Above fig 1, we can see correlation matrix and through that matrix it can be said that there is correlation between few variables. Below is the list of few of them which shows some strong correlation:

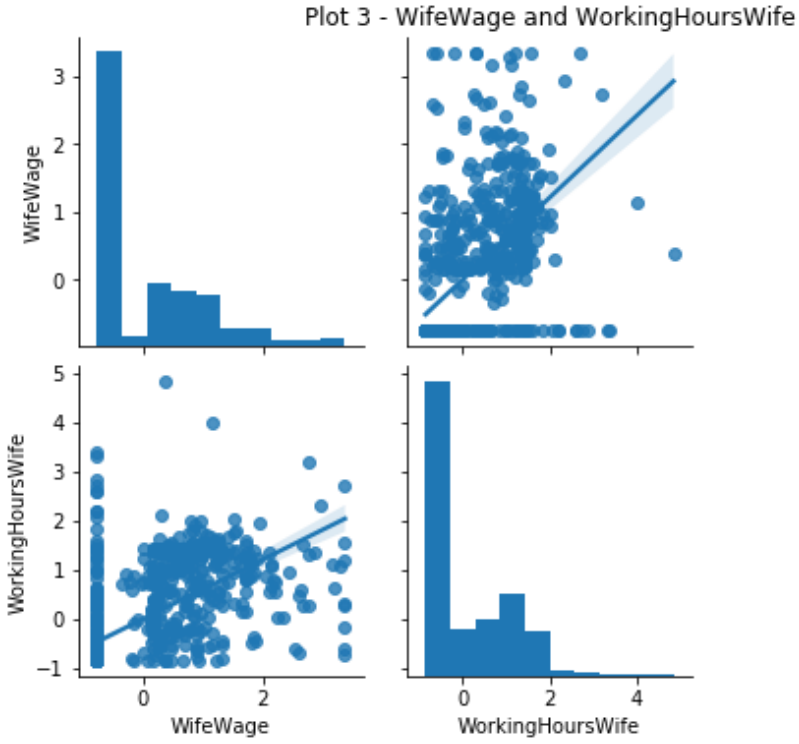
1. Wife Age & Husband Age: there is positive correlation of 0.89. It is showing good relation between the age of husband and wife.
2. Wife Wage & Wife hour Earnings: there is also positive correlation but of 0.65.
3. Wife Wage & Working Hours Wife: It is also showing positive correlation of 0.61
4. Education Wife Mother & Education Wife Father: It is not reflecting much strong correlation still of 0.57
5. Education Wife is equally correlated with Education Wife Father and Education Wife Mother. Although it's not very high but still moderate positive correlation of 0.44.
6. Wife Hour Earnings and Working Hour Wife is also showing moderate positive correlation of 0.42.
7. There are some more variables showing positive and negative correlation but those are weak.
8. So, it is clearly seen that there is multicollinearity in the Family Income data.
9. Together with the correlation, pairplot shows how the relation look like. It will be not clear if we present pairplot for whole data. So, presenting graphs of few variables mentioned above to show how their relation looks on graph and does the correlation value justifies it.



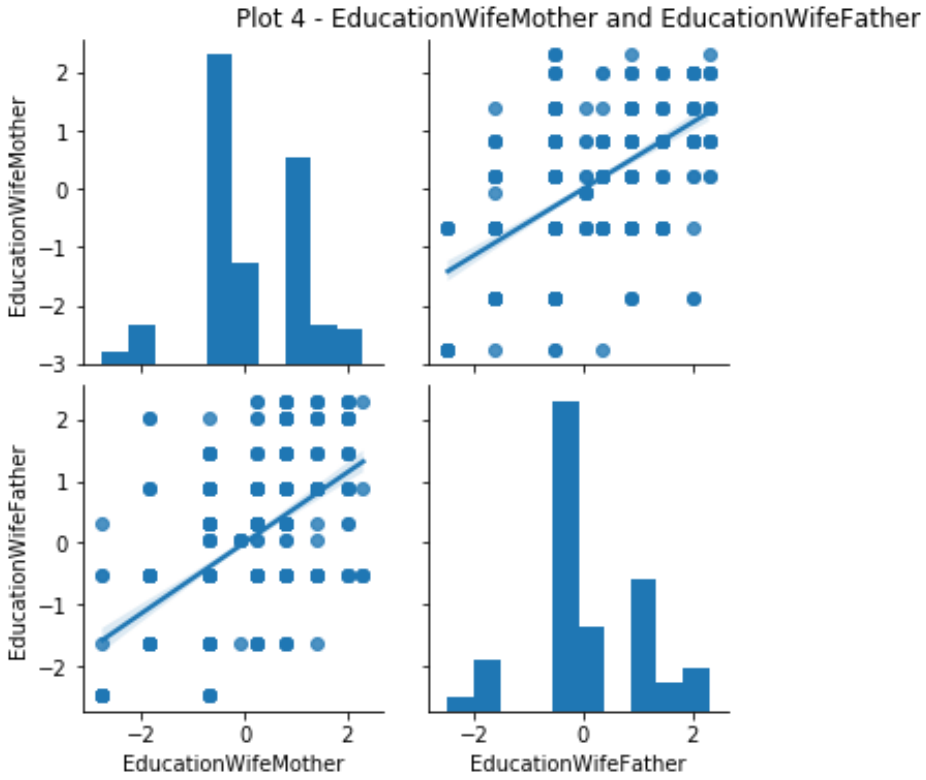
By pair plot 1, we can see high linearity between the two variables, which as per correlation matrix is .89.



Here for plot 2, we can see slightly lighter pattern of linear relation between the two feature variables (which as per matrix is 0.61) and it is not as strong as the above plot 1.



Plot 3, again, does not show strong correlation, but still some pattern is visible. As per corr matrix = 0.61.



Plot 4 shows even more dispersed pattern for which correlation coefficient is 0.57.

So, it is clearly visible that as the value of correlation starts reducing, it becomes visible in pair plot as well.

Overall, it can be said that very few variables are there to show strong collinearity as strong as 0.89. If we consider strong multicollinearity as correlation more than 0.5, there are only four such cases as plotted above. Rest most independent variables have a correlation less than 0.44

MULTIPLE LINEAR REGRESSION

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

1. Before performing multiple linear regression, we notice a great difference in unit scales among variables, hence, it is important to scale the data and standardise.
2. For scaling the data, we use standard scaler from python which standardize it based on z score. We are scaling the whole data, as values of target variable is also comparatively very large. So, if dependent variable will not be scaled then the result includes large error value. Whereas after scaling error get minimised and on the same time it will help to interpret the relation of target variable with feature variable on per unit basis.
3. So, after scaling the data, for checking that scaling has been done correctly we have checked covariance matrix and with the result we found out that covariance value is equivalent to the correlation value. Thus, it depicts that scaling is done correctly.
4. Now, for performing linear regression, data is split into X and Y as feature and target variable, respectively and apply regression model. Summary table is as below.

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.705
Model:	OLS	**Adj. R-squared:**	0.700
Method:	Least Squares	**F-statistic:**	135.9
Date:	Fri, 12 Jun 2020	**Prob (F-statistic):**	8.01e-186
Time:	11:51:42	**Log-Likelihood:**	-608.73
No. Observations:	753	**AIC:**	1245.
Df Residuals:	739	**BIC:**	1310.
Df Model:	13		
Covariance Type:	nonrobust		

5. Here R-squared value is 0.705, which we can say its generally considered a moderate effect size. R^2 of the above model is 0.705, it can be concluded that approximately 70% of the observed variation can be explained by the model's inputs.
6. Prob(F-statistics) is close to 0, it shows that reject the null hypothesis, that is all of the regression coefficients are zero. So it states that most of the regression parameters are non-zero and the regression equation does have validity in fitting the data, in other words independent variables are not purely random with respect to the dependent variable.

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.031e-17	0.020	2.52e-15	1.000	-0.039	0.039
WorkingHoursWife	0.1981	0.027	7.456	0.000	0.146	0.250
WifeAge	0.0888	0.045	1.967	0.050	0.000	0.178
EducationWife	0.0702	0.029	2.449	0.015	0.014	0.126
WifeHourEarnings	0.0827	0.027	3.060	0.002	0.030	0.136
WifeWage	0.0551	0.031	1.802	0.072	-0.005	0.115
WorkingHoursHusband	0.3312	0.022	15.356	0.000	0.289	0.374
HusbandAge	0.0182	0.044	0.412	0.680	-0.068	0.105
EducationHusband	-0.0187	0.027	-0.682	0.495	-0.072	0.035
HusbandWage	0.7934	0.024	33.699	0.000	0.747	0.840
EducationWifeMother	0.0088	0.026	0.341	0.733	-0.042	0.059
EducationWifeFather	0.0072	0.026	0.282	0.778	-0.043	0.057
UnemploymentRate	-0.0130	0.021	-0.634	0.526	-0.053	0.027
WifeExperience	-0.0712	0.024	-2.923	0.004	-0.119	-0.023

Omnibus:	404.335	Durbin-Watson:	2.073
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5410.657
Skew:	2.107	Prob(JB):	0.00
Kurtosis:	15.438	Cond. No.	5.29

7. Coefficient of Husband wage is the largest coefficient here, so we can say it is the most impacting factor on deriving family income. Its t value is also large, which says that it's standard error is less and coefficient is also large. In this feature we can say with 100% probability of being correct that the variable is having good effect, as p-value is 0.
8. Second largest coefficient is of Working Hours Husband, that is 0.3312, it's t-value is also large, so small error and the p-value is 0. So again, we can say with 100% probability of being correct that this variable is having good effect.
9. Third largest coefficient is 0.1981 that is of Working Hours Wife, it's t-value is 7.45, so small error and the p-value here is 0. So, for this as well we can say with 100% probability of being correct that this variable is having good effect.
10. Some of the variables have p-value of greater than 0.05, like Wife Wage, Husband age, Education Husband, Education Wife mother, Education Wife Father and Unemployment rate. For these variables, coefficient is also very low. So, we can say that these variables have no effect.
11. And rest variables have low coefficient, but p-value is less than 0.05, so it can be concluded they are showing impact.
12. Some have good coefficient (comparatively with other variables having greater than 0.05 p-value) in negative, like wife experience and its p-value is also less than 0.05. So, we can say it is showing inverse effect.
13. After this we have calculated y predicted and finally Root Mean Square Error.
14. RMSE is a measure of how spread out the residuals or we can also say how concentrated the data is around the line of best fit. Here **RMSE value is 0.5431**. We can conclude here that RMSE value is large for the model.

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.

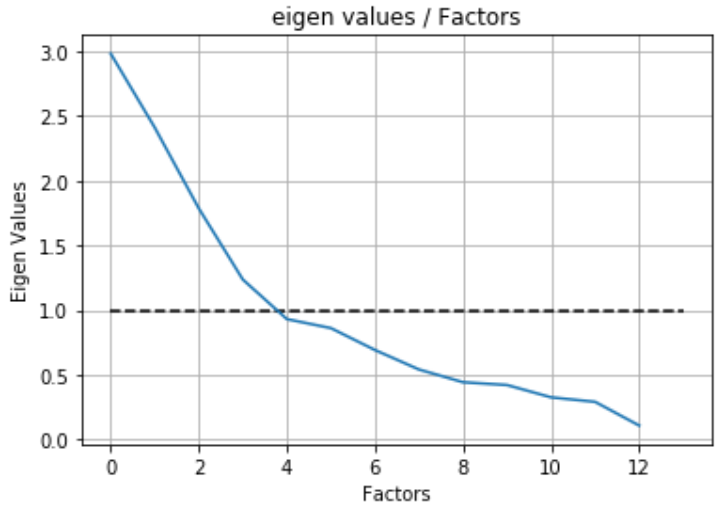
1) In the given data set, we see that number of feature variable are 13 which is a big number of independent variables to try a model. In general, PCA is useful to reduce the number of dimensions and here we will use PCA for the same.
2) Another reason for using PCA is to avoid multicollinearity. In question 2.2 above we saw there are few variables which show multicollinearity, hence, we can apply PCA and check the performance of model by removing multicollinearity.
3) PCA is a statistical technique which reduces the dimensions of the data and help us understand, plot the data with lesser dimension compared to original data. As the name says PCA helps us compute the Principal components in data. Principal components are vectors that are linearly uncorrelated and have a variance within data.
4) From the principal components top *n components are* picked which have the most variance.

We carry out PCA in a sequential manner as explained in following steps:

5) Standardise the data and build a covariance matrix: Since the data is standardised here, all the diagonal values of our covariance matrix should be 1, or the covariance matrix will be an identity matrix of 13x13 variables. This is a must before extraction of eigen values/vectors.
6) Decompose covariance matrix into eigen values and eigen vectors through linear algebra module in python. Following eigen values are extracted from our data.

Eigen Values

[2.98225435 2.41033838 1.78562682 1.2353852 0.10674915 0.92972654
 0.85994247 0.69081582 0.54007441 0.28991404 0.32473922 0.44163312
 0.42008771]



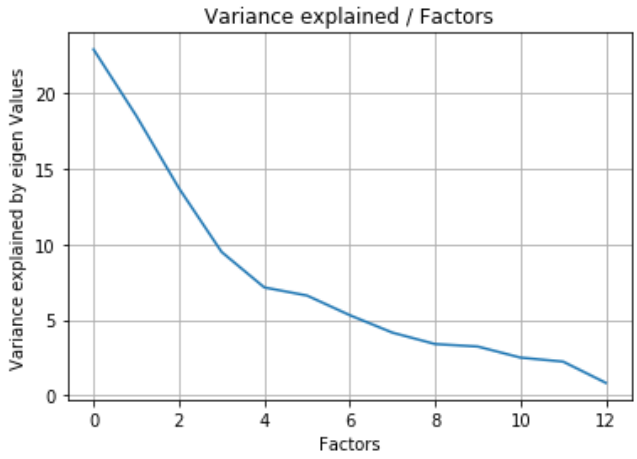
7) From the above list of eigen values and plot, according to Kaizer rule, we must shortlist all the components which can explain the variation of at least one variable out of total variables or in simple words whose eigen value is greater than 1. Here, first 4 eigen are greater than 1 and they can explain

the variation of more than 8 out of 13 variables. However, if we choose, highest 5 eigen values which includes eigen value 0.9297, it can explain more than 9 out of 13 variables' variation. Hence, we can consider highest 5 components which are close to 1 or more and check the next step.

8) Calculate variance and cumulative variance explained by eigen values:

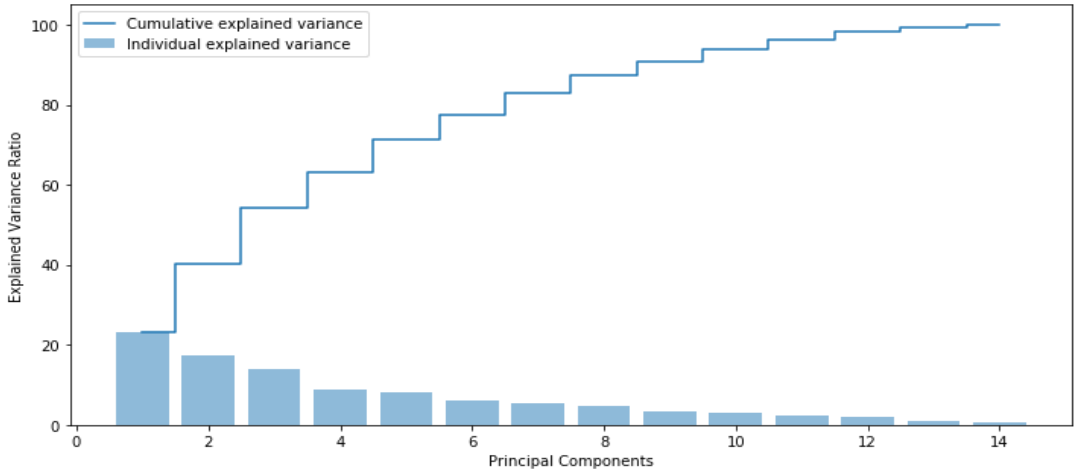
The variance explained by each of eigen values in order is:

[22.909952727776243, 18.51644154604999, 13.717349758794828, 9.490342942168786, 7.142244968694198, 6.606157297491236, 5.306910765021286, 4.148901360032477, 3.392666313185718, 3.2271525130873715, 2.4946766067414483, 2.227146383598584, 0.8200568173578473]



Cumulative Variance Explained:

[22.90995273 41.42639427 55.14374403 64.63408697 71.77633194
78.38248924 83.68940001 87.83830137 91.23096768 94.45812019
96.9527968 99.17994318 100.]



9) From the above two list of explained variance and cumulative variance and the plot between eigen values-factors, we can see that the cumulative variance explained by the selected 5 components (according to Kaizer rule) is 71.776% only

10) Extract the principal components based on variance explained: As per industry norms, it is good to choose as many Principal Components as can explain at least 75% variance. Hence, we see that if

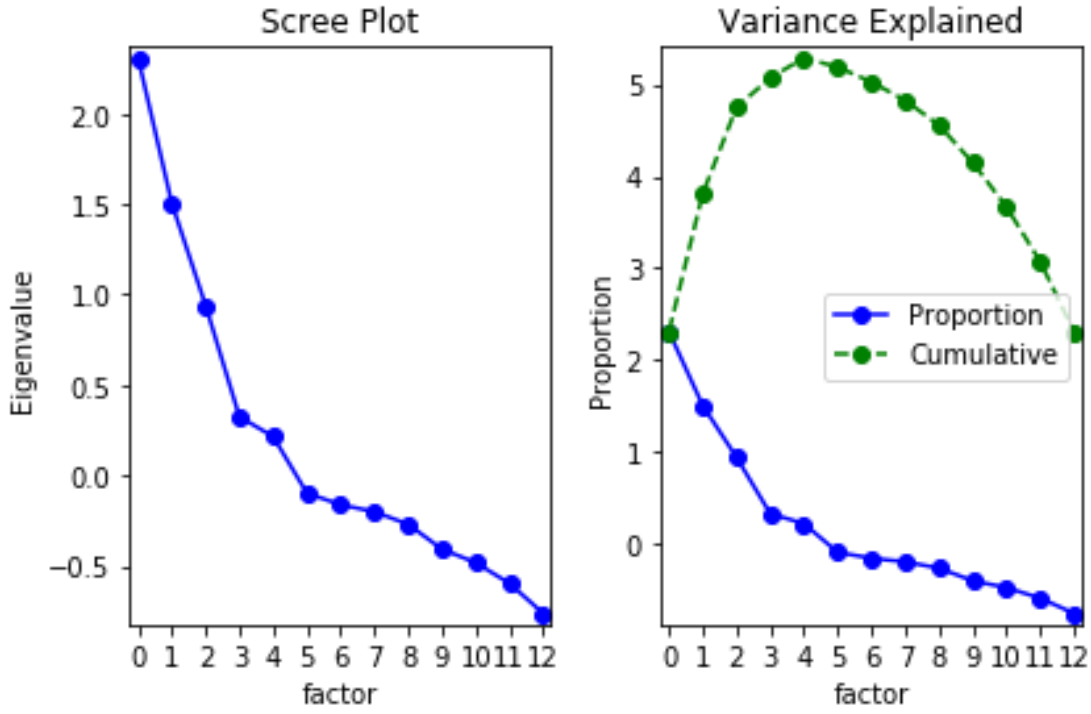
we select highest six principal components, cumulative variance explained would be 78.382%, which is a good value to be considered.

11) Hence, considering both Kaizer rule and industry norms, it deems fit to choose minimum **6 principal components which explain 78.382% of variance in the independent variables.**
12) Finally, the 6 principal components that replace the 13 features are loaded and explained in next question

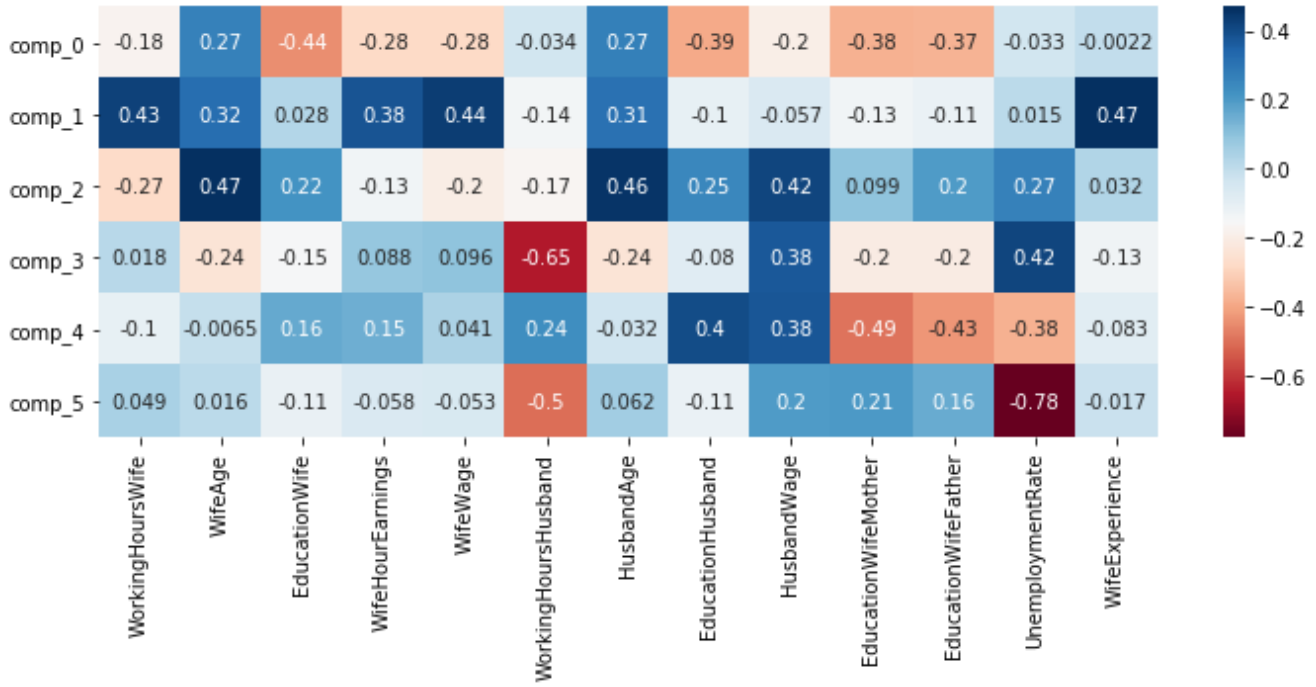
MULTIPLE LINEAR REGRESSION AFTER PCA

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

1) To perform Multiple Linear Regression, we need to follow the same process as in Q2.3, however, with extracted principal components, the number of independent variables will be reduced to six as we extracted in Q 2.4.
2) Since, we have already done the exercise of extracting PC from eigen values, we can have look on the principal components through Scree plot here



3) Now, can see the 6 Principal Components representing all 13 variables.
4) Also, the new components are not correlated to each other , hence the issue of multicollinearity , if any , can be taken care of.



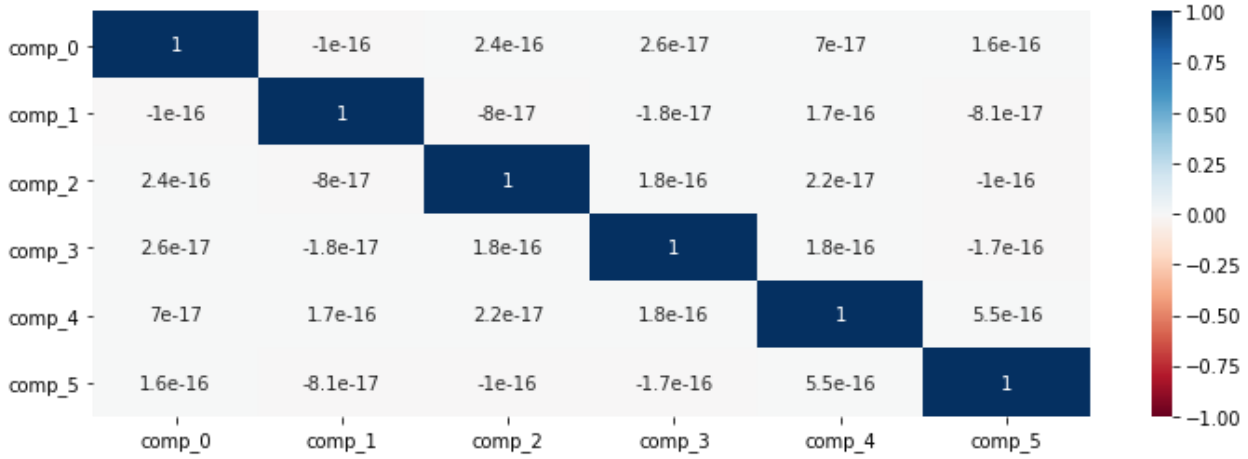
5) From the above we can explain the different component according to the loading they have for a given variable / variables.

Comp_0: represents one dimension i.e. EducationWife
Comp_1 : represents WorkingHoursWife , WifeWage, WifeHoursEarning and WifeExperience
Comp_2: represents WifeAge, HusbandAge, HusbandWage
Comp_3: represents WorkingHoursHusband
Comp_4: represents EducationHusband, EducationWifeMother and EducationWifeFather
Com_5: represents UnemploymentRate

6) We can finally extract the reduced dataset with six dimensions. Below are the top 5 rows of new dataset, which will be used for performing regression

	comp_0	comp_1	comp_2	comp_3	comp_4	comp_5
0	-0.019107	-0.001738	-0.067640	-0.020916	0.000299	0.015232
1	-0.005178	-0.015930	-0.058042	0.054033	-0.005109	-0.029012
2	-0.019851	0.019512	-0.065041	-0.041485	0.004189	0.003137
3	0.022086	-0.000003	-0.028621	-0.000746	0.006209	0.036393
4	-0.055158	-0.009212	-0.028315	0.033167	-0.031157	0.012545

7) Check correlation matrix for new dataset: As is visible that the dimensions in the new dataset have no correlation with eachother. So the multicollinearity and dimension reduction, both problems are resolved by PCA.



8) We can now perform the multiple regression on the dataset and analyse its result:

OLS Regression Results

Dep. Variable:	FamilyIncome	**R-squared:**	0.437
Model:	OLS	**Adj. R-squared:**	0.432
Method:	Least Squares	**F-statistic:**	96.37
Date:	Sat, 13 Jun 2020	**Prob (F-statistic):**	1.47e-89
Time:	17:22:29	**Log-Likelihood:**	-852.40
No. Observations:	753	**AIC:**	1719.
Df Residuals:	746	**BIC:**	1751.
Df Model:	6		
Covariance Type:	nonrobust		

9) Here, we see that R square value has significantly reduced to 0.437 from 0.705, which means that model can only explain approx. 44% variation due to the principal components chosen.
10) Also, the RMSE value of the model has increased from 0.543 to 0.751, which indicates that the dispersion of residuals has also increased after performing PCA.

MODEL EXPLANATION

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

1. Linear regression model built using the principal component as independent variable and Family income as dependent variable.
2. In this variable number is reduced to 6 principal components from 1616 dimensions.
3. *After PCA, when model is run, R-squared value of the new model has reduced to 0.437 and adjusted R-squared value decreased further to 0.432.*
4. On this basis we can say that there are some weak attributes as it decreases the adjusted R-squared and here the value of R-squared is generally considered a poor effect size. So, it can be concluded that only approximately 43% of the observed variation can be explained by the model's input.
5. Prob(F-statistics) is 10^{-89}, which is close to 0, it states that reject null hypothesis and accept alternate hypothesis that all of the regression coefficient are not zero.
6. Standard error remains constant to 0.754 for all the components.
7. Component 0 has good impact on family income but in inverse relation, it's coefficient is -11.4083 and p-value is 0.
8. Component 4 and component 2 is also showing p-value as 0 and high coefficient, so we can conclude it's also has good impact over family income.
9. Only component 5, has p-value of 0.801, which is greater than 0.05 and its coefficient is also lowest, so we can say this component has negligible effect on family income.
10. And finally, the root mean square error value is also increased upto 0.75 which is very large. This RMSE value indicates poor fit.

Tested model with more no. of components:

11. We can also observe the impact on the model as we change the no. of components selected to perform the linear regression. Following is the behaviour of model when we choose **7components**:

OLS Regression Results			
Dep. Variable:	FamilyIncome	**R-squared:**	0.456
Model:	OLS	**Adj. R-squared:**	0.451
Method:	Least Squares	**F-statistic:**	89.13
Date:	Sun, 14 Jun 2020	**Prob (F-statistic):**	4.34e-94
Time:	14:07:39	**Log-Likelihood:**	-839.40
No. Observations:	753	**AIC:**	1695.
Df Residuals:	745	**BIC:**	1732.
Df Model:	7		
Covariance Type:	nonrobust		

12. The MSE value is 0.5442 and Root Mean Square Error (RMSE) value is 0.7377118229818004
13. When we select top **8 components**, it can be seen that R-squared and adjusted R-squared value has increased considerably and the RMSE has decreased.

OLS Regression Results			
Dep. Variable:	FamilyIncome	**R-squared:**	0.645
Model:	OLS	**Adj. R-squared:**	0.642
Method:	Least Squares	**F-statistic:**	169.2
Date:	Sun, 14 Jun 2020	**Prob (F-statistic):**	8.07e-162
Time:	14:09:01	**Log-Likelihood:**	-678.19
No. Observations:	753	**AIC:**	1374.
Df Residuals:	744	**BIC:**	1416.
Df Model:	8		
Covariance Type:	nonrobust		

14. The MSE value is 0.3547 and Root Mean Square Error (RMSE) value is 0.595542893037681.
15. Thus, in conclusion we **approximate the original data** with PCA. We are likely to lose some information, but if we can minimize the information we lose, we would get better result from the model built.
16. PCA works by forming a new set of variables from the original features. It does that by maximizing the variance the new variables can account for. We approximate what we have, but the new approximation is not perfect. In practice, we can decide how many principal components we keep. The more we keep, the better approximation we get.

BUSINESS IMPLICATION

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

Business implications:

1) The data provided here is to understand how the different attributes impact the family income of this sample753 families and we are supposed to find if there can be a few chosen variables which can explain the family income trend
2) Above model built from the provided data can be used to develop a tool for designing a wage model to bring in uniformity of wages and working hours depending on employees' experience, education & working hours for wife and husband both. It will help in maintaining even salary break up which finally leads to even family income for everyone as per their experience, education and working hours. Model reflected that husband wages highly effect the family income. So, it can also be used to formulate salary for male and female employees as per the company norms
3) However, during EDA certain findings are interesting as follows: only Husband wage is an independent variable which is strongly linearly related to the family income, rest are very moderately impacting the target. Unlike husband wages, wife wages, surprisingly, do not show any conclusive relation with family income.
4) There is no other continuous variable which can strongly explain the family income.
5) Unemployment Rate variable shows no relation with the target, which may be explained by low unemployment rate or strong economy, hence people are financially secured.
6) More than 50% of wives are working but not drawing any wages is also unexplained by the data.
7) Also, more independent variables given are in reference to the wives as compared to the husbands.
8) Most of the I.V are not even linearly related to the family income variable.
9) The variables on EDA seem to have not much impact on target.
10) After performing regression also, the results were not promising.
11) After performing PCA too, results were further down.
12) It can also be recommended that certain other dimensions may be looked into which may make the model from this data stronger.

Interpretation of the models:

1. R-squared value of the new model built after performing PCA, has reduced to 0.437 and adjusted R-squared value decreased further to 0.432 comparative to previously performed model. R-squared and adjusted R-squared of previous performed model is 0.705 and 0.70. So, it can be concluded that approximately 43% of the observed variation can be explained by the new model's input whereas prior to PCA it was 70%.
2. Prob(F-statistics) is close to 0, means reject the null hypothesis. It means that all **regression coefficient are zero**. But in new model it is at 10^{-89} whereas before PCA it was 10^{-186}.
3. Standard error shows how much it is away from mean. In the new model we observed that **standard error** becomes constant for all the components at 0.754 whereas in previous model it was below 0.05.
4. Through p-value we can identify whether to reject or accept null hypothesis which is the coefficient is equal to zero (no effect). P-value less than 0.05 reject null hypothesis and it shows there is relation. In the model performed before PCA there were 13 variables and out of it, 6 has p-value of greater than 0.05 whereas in the model after PCA there were total 6 components and out of it only 1 component has p-value more than 0.05.

5. Finally, the Root Mean Square Error of previous model was 0.54 which in later model increased to 0.75. In both the cases RMSE is large, but in second it increased.
6. PCA might have led to loss of information, nevertheless, model was not very strong without PCA too