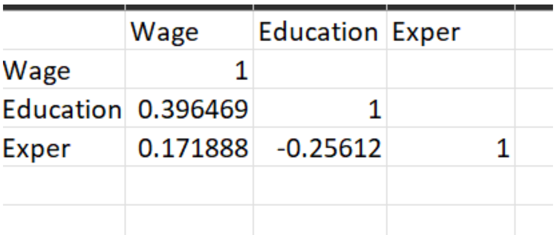
Question 1:



	Wage	Education	Exper
Wage	1		
Education	0.396469	1	
Exper	0.171888	-0.25612	1

Correlation indicates strength and direction of the relationships between variables.

The correlation coefficient between "wage" and "education" is 0.396469. This is a moderate positive i.e., as the value of education increases, wages is likely to increase, but it is not a very strong relationship.

The correlation coefficient between "wage" and "exper" is 0.1718. This is a weak positive correlation between wage and experience i.e., as experience increases, wages tend to increase slightly, but the relationship is not very strong. This is highly unlikely as per the real-world scenario.

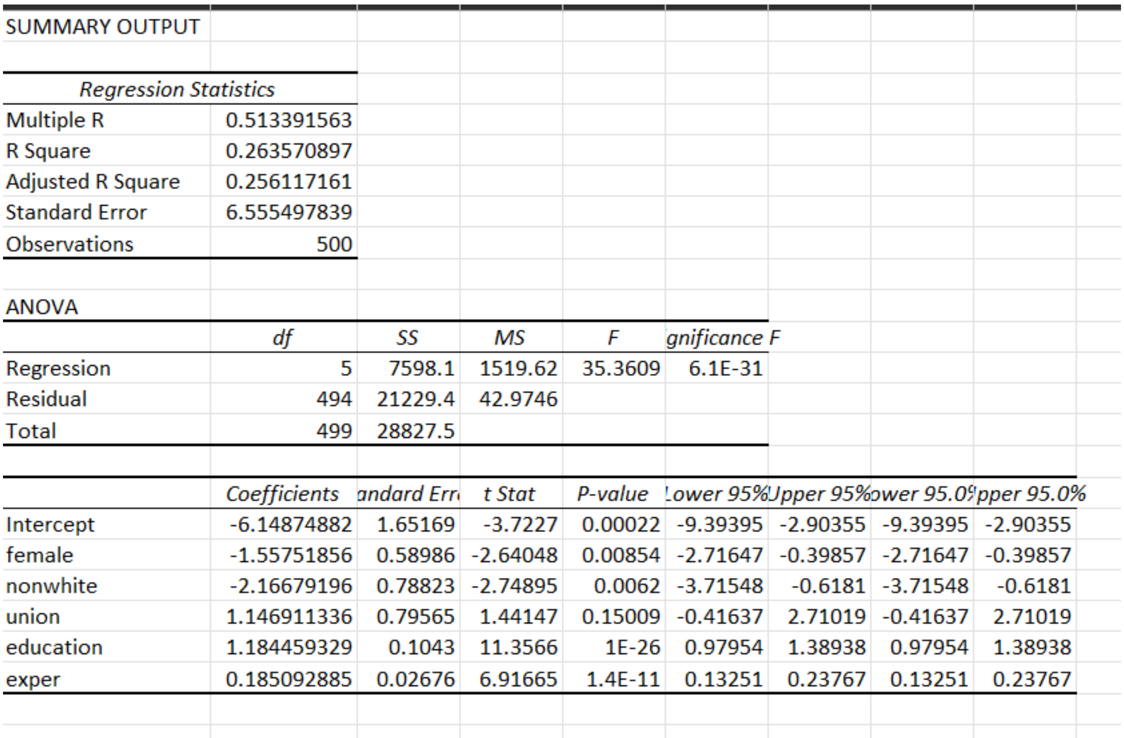
The correlation coefficient between "education" and "exper" is -0.25612. This is a weak negative correlation between education and experience, suggesting that as education levels increases, there is a tendency for experience levels to decrease slightly.

Question 2:

The Regression Equation with Wage as dependent variable and all other variables except Age as independent variable is given by:

Wage $=B_o + B_1 * female + B_2 * nonwhite + B_3 * union + B_4 * education + B_5 * exper$

Each coefficient represents the expected change in the wage for an additional unit change in the respective explanatory variable holding other variables constant.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.513391563
R Square	0.263570897
Adjusted R Square	0.256117161
Standard Error	6.555497839
Observations	500

ANOVA

	df	SS	MS	F	gnificance F
Regression	5	7598.1	1519.62	35.3609	6.1E-31
Residual	494	21229.4	42.9746		
Total	499	28827.5			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-6.14874882	1.65169	-3.7227	0.00022	-9.39395	-2.90355	-9.39395	-2.90355
female	-1.55751856	0.58986	-2.64048	0.00854	-2.71647	-0.39857	-2.71647	-0.39857
nonwhite	-2.16679196	0.78823	-2.74895	0.0062	-3.71548	-0.6181	-3.71548	-0.6181
union	1.146911336	0.79565	1.44147	0.15009	-0.41637	2.71019	-0.41637	2.71019
education	1.184459329	0.1043	11.3566	1E-26	0.97954	1.38938	0.97954	1.38938
exper	0.185092885	0.02676	6.91665	1.4E-11	0.13251	0.23767	0.13251	0.23767

Wage $= B_o + B_1 * female + B_2 * nonwhite + B_3 * union + B_4 * education + B_5 * exper$

This equation can be interpreted as that:

Wage $= B_o$ means that wage for Male with white race, not a union member , with no education and no work experience .

Wage = $B_o + B_1$ means that wage for Female with white race, not a union member, with no education and no work experience.

Wage = $B_o + B_1 + B_2$ means that wage for Female with non-white race, not a union member, with no education and no work experience.

So on, Wage = $B_o + B_1 + B_2 + B_3 + B_4 * education + B_5 * exper$ means that wage for Female with non-white race, a union member, with some education and with some work experience.

Regression Equation 1:

Wage = $-6.1487488180495 - 1.55751856352247 * female - 2.16679195919089 * nonwhite + 1.14691133598611 * union + 1.18445932909071 * education + 0.185092885466996 * exper$

Question 3:

Explanation of R^2 , F-test and individual p values of the coefficients with respect to Regression Equation 1.

1. R^2 (Coefficient of Determination):- The R^2 value is 0.263570897062628 suggesting that approx. 26% of the variance in the wage can be explained by the predictor variables used in the model.

2. F-test: - F- test in the regression model asses the overall significance of the model. F statistics of the given model is 35.3609119003031. It measures the mean square for regression (MS) by the mean square for the residual (MS). The significance F value is 6.14758498146384E-31 (very close to zero). This represents the p-value associated with the F-statistic. The extremely small p-value indicates that the null hypothesis is being rejected.

3. The p-values of each coefficient provide information about the statistical significance of the individual predictor variables in relation to the dependent variable (wage).

    1. Intercept: The p-value for the intercept term is 0.000219786453906055 indicating that the intercept term is statistically significant. So, when all the predictor variables are zero, average wage is significantly different from zero.
    2. Female: The p-value for the coefficient of the "female" variable is 0.00854084818489423. This suggests that the variable is statistically significant i.e., female has a significant effect on the wage.
    3. Nonwhite: The p-value for the coefficient of the "nonwhite" variable is 0.00619767412399708 less than the threshold 0.05. Therefore, the "nonwhite" variable is statistically significant in explaining the wage in this model.
    4. Union: The p-value for the coefficient of the "union" variable is 0.150085539750413. greater than 0.05. Thus, the "union" variable is not statistically significant, and hence has no significant effect on the wage when controlling for other variables.
    5. Education: The p-value for the coefficient of the "education" variable is 1.02362951041274E-26, suggesting that the "education" variable is statistically significant even after accounting for other variables.
    6. Exper: The p-value for the coefficient of the "exper" variable is 1.43769300093448E-11,implying that the "exper" variable is statistically significant i.e., the experience has a significant effect on the wage, even when considering other variables.

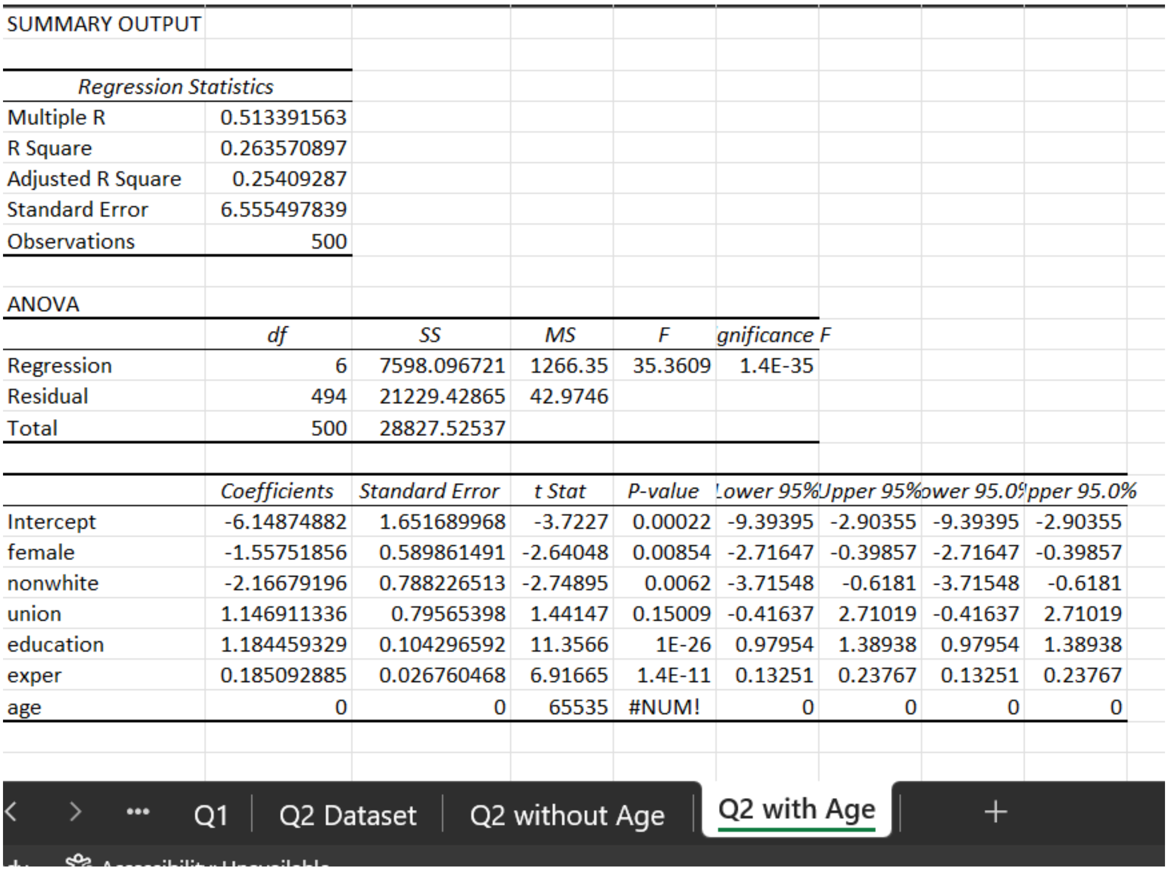
Question 4:

The Regression Equation with Wage as dependent variable and all other variables including Age as independent variable is given by:

Wage $=B_o + B_1 * female + B_2 * nonwhite + B_3 * union + B_4 * education + B_5 * exper + B_6 * age$

Each coefficient represents the expected change in the wage for an additional unit change in the respective explanatory variable holding other variables constant.

Including age in the model gave below result:-


SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.513391563
R Square	0.263570897
Adjusted R Square	0.25409287
Standard Error	6.555497839
Observations	500

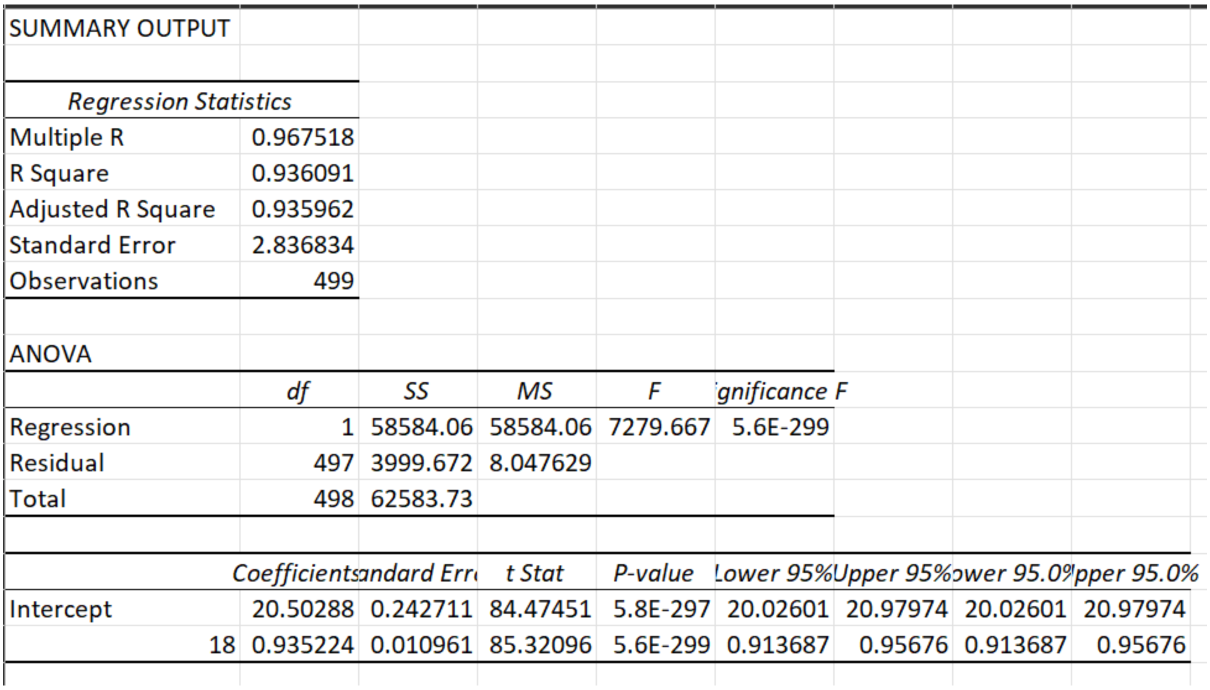
ANOVA

	df	SS	MS	F	gnificance F
Regression	6	7598.096721	1266.35	35.3609	1.4E-35
Residual	494	21229.42865	42.9746		
Total	500	28827.52537			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-6.14874882	1.651689968	-3.7227	0.00022	-9.39395	-2.90355	-9.39395	-2.90355
female	-1.55751856	0.589861491	-2.64048	0.00854	-2.71647	-0.39857	-2.71647	-0.39857
nonwhite	-2.16679196	0.788226513	-2.74895	0.0062	-3.71548	-0.6181	-3.71548	-0.6181
union	1.146911336	0.79565398	1.44147	0.15009	-0.41637	2.71019	-0.41637	2.71019
education	1.184459329	0.104296592	11.3566	1E-26	0.97954	1.38938	0.97954	1.38938
exper	0.185092885	0.026760468	6.91665	1.4E-11	0.13251	0.23767	0.13251	0.23767
age	0	0	65535	#NUM!	0	0	0	0

Here, the age coefficients is not provided indicating some error in the data. When checked whether there is some linearity dependency within variables, found that age and exper are dependent.

Running the regression between the two variables with age being the dependent variable found that exper has significant p value suggesting significant effect on age variable . Also, scatterplot is checked to check the collinearity between these two variables.

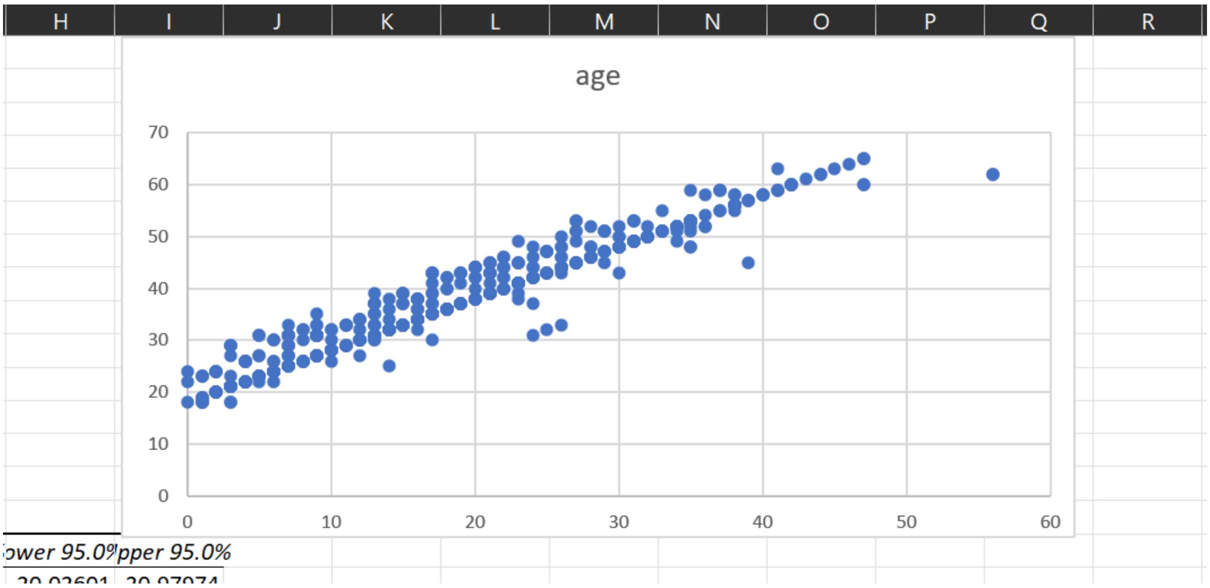

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.967518
R Square	0.936091
Adjusted R Square	0.935962
Standard Error	2.836834
Observations	499

ANOVA

	df	SS	MS	F	gnificance F
Regression	1	58584.06	58584.06	7279.667	5.6E-299
Residual	497	3999.672	8.047629		
Total	498	62583.73			

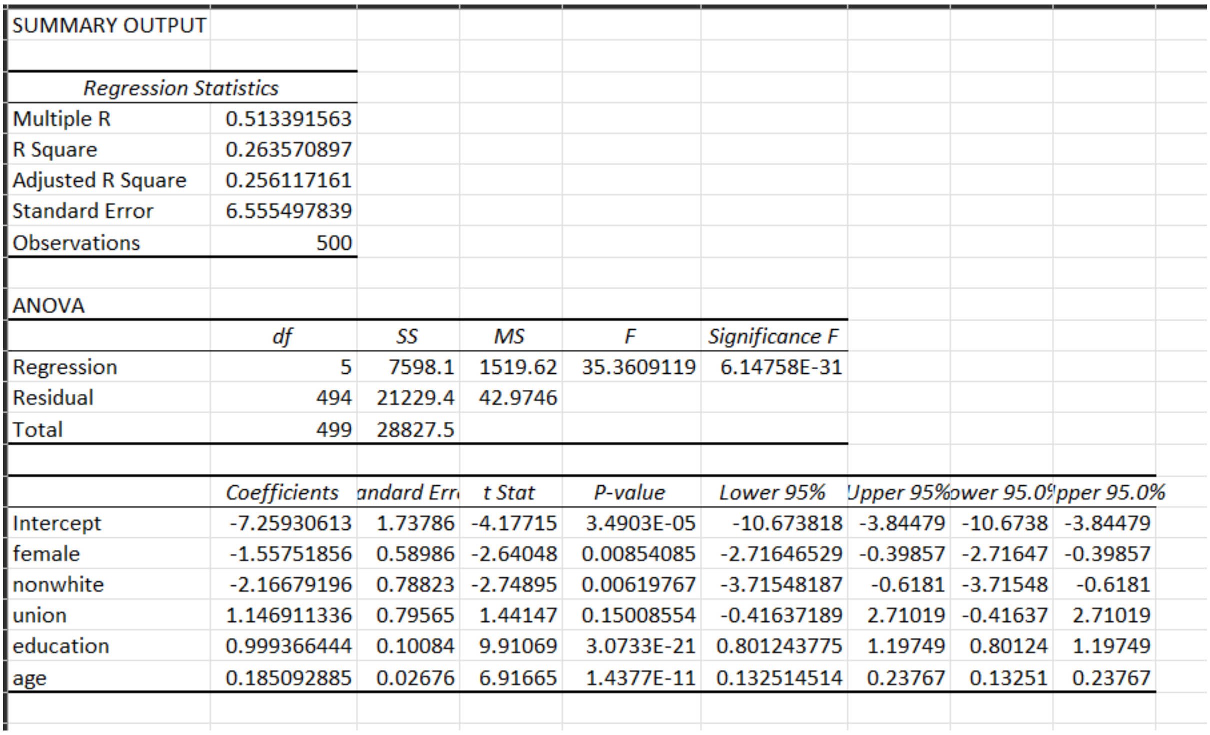
	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	20.50288	0.242711	84.47451	5.8E-297	20.02601	20.97974	20.02601	20.97974
18	0.935224	0.010961	85.32096	5.6E-299	0.913687	0.95676	0.913687	0.95676



ower 95.0%	pper 95.0%
20.02601	20.97974

From scatterplot, it appears that both the variables have positive linear relationship between age and exper variable, suggesting that as age increases exper tends to increase. This suggests a linear dependence between the two variables.

Now, removing exper variable and running the regression model provided same Regression statistics with slight increase in the Adjusted R square. Also, noted that there has been slight decrease in the coefficients of intercept and education but rest all the variables' coefficient remained the same. Also, coefficient of exper and age came out to be same.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.513391563
R Square	0.263570897
Adjusted R Square	0.256117161
Standard Error	6.555497839
Observations	500

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	7598.1	1519.62	35.3609119	6.14758E-31
Residual	494	21229.4	42.9746		
Total	499	28827.5			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-7.25930613	1.73786	-4.17715	3.4903E-05	-10.673818	-3.84479	-10.6738	-3.84479
female	-1.55751856	0.58986	-2.64048	0.00854085	-2.71646529	-0.39857	-2.71647	-0.39857
nonwhite	-2.16679196	0.78823	-2.74895	0.00619767	-3.71548187	-0.6181	-3.71548	-0.6181
union	1.146911336	0.79565	1.44147	0.15008554	-0.41637189	2.71019	-0.41637	2.71019
education	0.999366444	0.10084	9.91069	3.0733E-21	0.801243775	1.19749	0.80124	1.19749
age	0.185092885	0.02676	6.91665	1.4377E-11	0.132514514	0.23767	0.13251	0.23767

Explanation of R^2 , F-test and individual p values of the coefficients with respect to Regression Equation 1.

1. R^2 (Coefficient of Determination):- The R^2 value is 0.263570897062629 suggesting that approx. 26% of the variance in the wage can be explained by the predictor variables used in the model.

2. F-test: - F- test in the regression model asses the overall significance of the model. F statistics of the given model is 35.3609119003032. It measures the mean square for regression (MS) by the mean square for the residual (MS). The significance F value is 6.1475849814622E-31 (very close to zero). This represents the p-value associated with the F-statistic. The extremely small p-value indicates that the null hypothesis is being rejected.

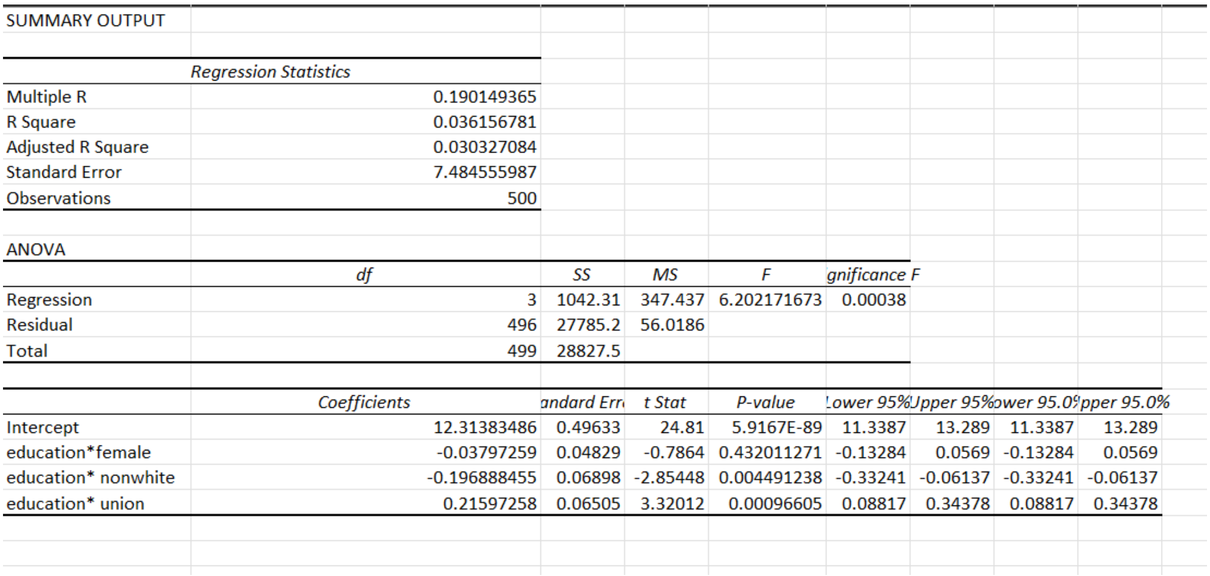
3. Adjusted R^2 : The increase in the adjusted R square indicates that the additional independent variable in the model is contributing to a better fit of data.

4. The p-values of each coefficient provide information about the statistical significance of the individual predictor variables in relation to the dependent variable (wage).

    1. Intercept: The p-value for the intercept term is 0.000219786453906055 indicating that the intercept term is statistically significant. So, when all the predictor variables are zero, average wage is significantly different from zero.
    2. Female: The p-value for the coefficient of the "female" variable is 0.00854084818489423. This suggests that the variable is statistically significant i.e., female has a significant effect on the wage.
    3. Nonwhite: The p-value for the coefficient of the "nonwhite" variable is 0.00619767412399708 greater than the threshold 0.05. Therefore, the "nonwhite" variable is statistically significant in explaining the wage in this model. We reject the null hypothesis because nonwhite has significant effect on wage while considering other variables.
    4. Union: The p-value for the coefficient of the "union" variable is 0.150085539750413. greater than 0.05. Thus, the "union" variable is not statistically significant, and hence has no significant effect on the wage when controlling for other variables.
    5. Education: The p-value for the coefficient of the "education" variable is 3.0732846770329E-21, suggesting that the "education" variable is statistically significant even after accounting for other variables.
    6. Age: The p-value for the coefficient of the "age" variable 1.43769300093438E-11, implying that the "age" variable is statistically significant i.e., the experience has a significant effect on the wage, even when considering other variables.

Question 5:

The interaction effect between "education" and each of the variables "female," "nonwhite," and "union.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.190149365
R Square	0.036156781
Adjusted R Square	0.030327084
Standard Error	7.484555987
Observations	500

ANOVA

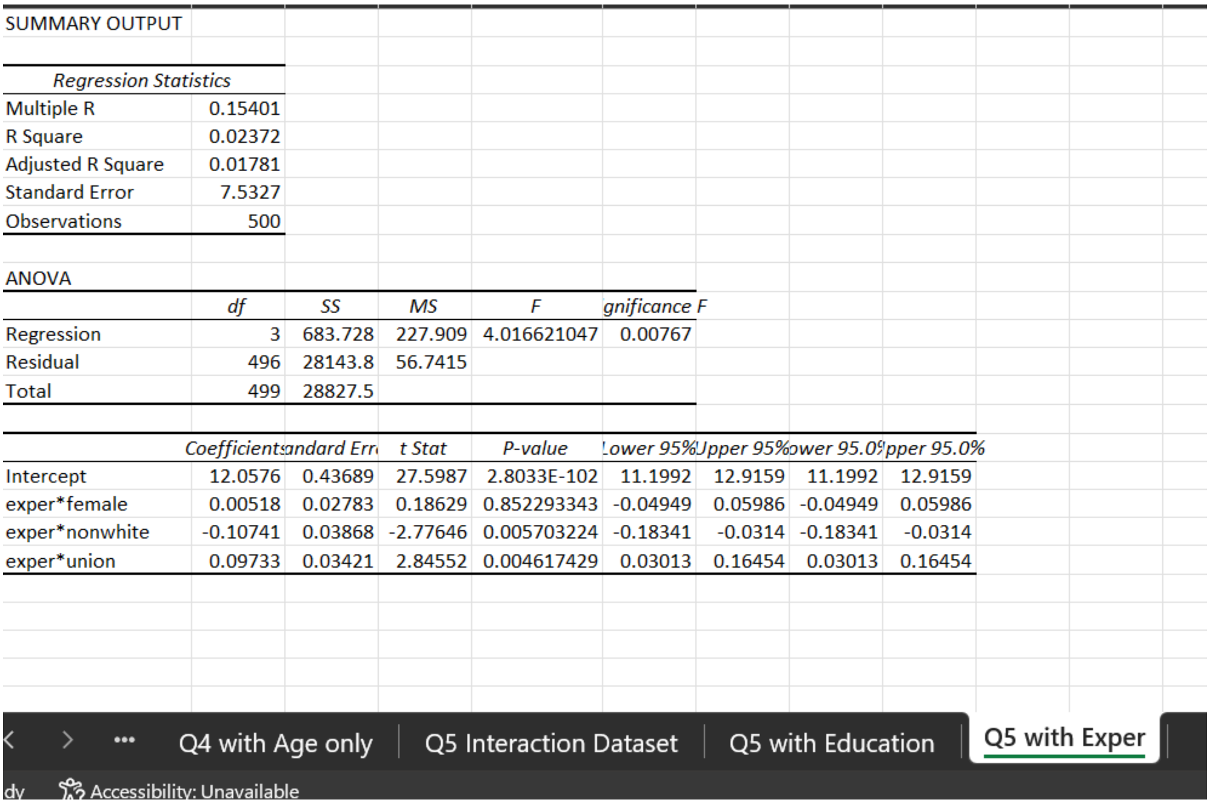
	df	SS	MS	F	gnificance F
Regression	3	1042.31	347.437	6.202171673	0.00038
Residual	496	27785.2	56.0186		
Total	499	28827.5			

	Coefficients	andard Erro	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	12.31383486	0.49633	24.81	5.9167E-89	11.3387	13.289	11.3387	13.289
education*female	-0.03797259	0.04829	-0.7864	0.432011271	-0.13284	0.0569	-0.13284	0.0569
education* nonwhite	-0.196888455	0.06898	-2.85448	0.004491238	-0.33241	-0.06137	-0.33241	-0.06137
education* union	0.21597258	0.06505	3.32012	0.00096605	0.08817	0.34378	0.08817	0.34378

Yes, there is an interaction effect in the regression model.

1. education*female: The coefficient -0.0379725904240176 indicates that the effect of education on the wage variable differs depending on the gender (female or male). The negative coefficient suggests that the relationship between education and the wage is weaker for females compared to males. However, the p-value (0.432011271254887) is not significant, indicating that this interaction effect is not statistically significant.
2. education*nonwhite: The coefficient is -0.19688845505998. This suggests that the effect of education on the wage variable differs based on the nonwhite or white background. However, the p-value (0.00449123826573105) is significant, indicating that this interaction effect is statistically significant.
3. education*union: The coefficient is 0.215972579585415. This suggests that the effect of education on the wage variable differs depending on whether the individual is a member or not. The positive coefficient indicates that the relationship between education and the wage is stronger for a union member.

The interaction effect between "exper" and each of the variables "female," "nonwhite," and "union.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.15401
R Square	0.02372
Adjusted R Square	0.01781
Standard Error	7.5327
Observations	500

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	683.728	227.909	4.016621047	0.00767
Residual	496	28143.8	56.7415		
Total	499	28827.5			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	12.0576	0.43689	27.5987	2.8033E-102	11.1992	12.9159	11.1992	12.9159
exper*female	0.00518	0.02783	0.18629	0.852293343	-0.04949	0.05986	-0.04949	0.05986
exper*nonwhite	-0.10741	0.03868	-2.77646	0.005703224	-0.18341	-0.0314	-0.18341	-0.0314
exper*union	0.09733	0.03421	2.84552	0.004617429	0.03013	0.16454	0.03013	0.16454

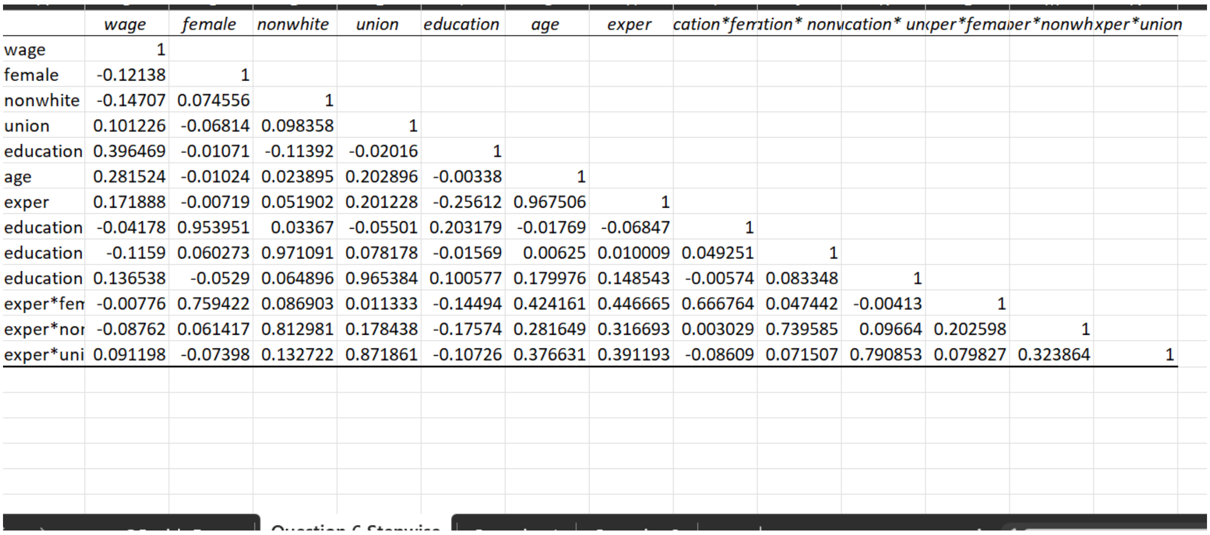
1. exper*female: The coefficient 0.00518424776135004 suggests that the effect of "exper" on the wage variable differs depending on the gender of the individual. The positive coefficient indicates that the relationship between "exper" and the wage is stronger for females compared to males. However, the p-value (0.852293342985486) is not significant, indicating that this interaction effect is not statistically significant.
2. exper*nonwhite: The coefficient is -0.10740695073777. This suggests that the effect of "exper" on the wage variable differs depending on whether the individual is white or nonwhite. However, the p-value (0.0057032241803729) is significant, indicating that this interaction effect is statistically significant.
3. exper*union: The coefficient is 0.0973310257886072. This suggests that the effect of "exper" on the age variable differs depending on whether the individual is a union member or not. The positive coefficient indicates that the relationship between "exper" and the wage is stronger for union members compared to non-union members.

Question 6:

Stepwise method is an approach to select most relevant variables in the regression model. It is a combination of forward selection and backward elimination steps. Considering wage as the dependent variable and all the other variables including the interaction variables, below model is prepared using stepwise approach.

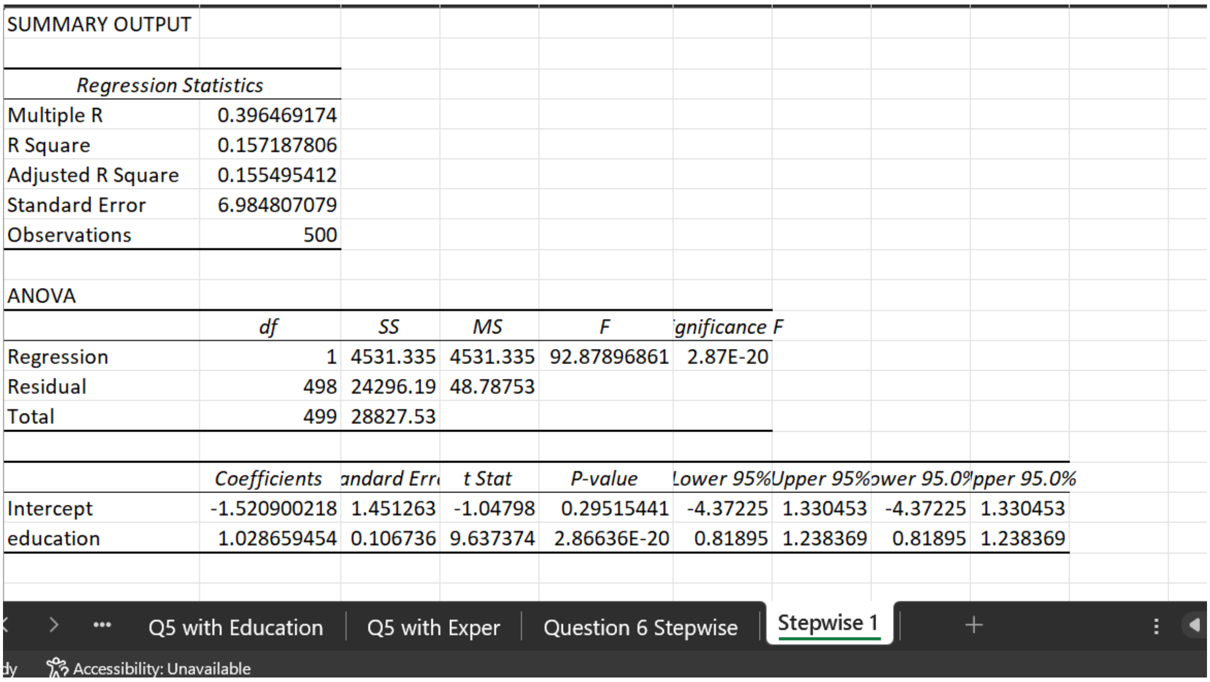
Steps included:

1. Correlation between wage and all other variables in calculated.



	wage	female	nonwhite	union	education	age	exper	cation*fem	ation* nonw	cation* un	per*fema	er*nonwh	xper*union
wage	1												
female	-0.12138	1											
nonwhite	-0.14707	0.074556	1										
union	0.101226	-0.06814	0.098358	1									
education	0.396469	-0.01071	-0.11392	-0.02016	1								
age	0.281524	-0.01024	0.023895	0.202896	-0.00338	1							
exper	0.171888	-0.00719	0.051902	0.201228	-0.25612	0.967506	1						
education	-0.04178	0.953951	0.03367	-0.05501	0.203179	-0.01769	-0.06847	1					
education	-0.1159	0.060273	0.971091	0.078178	-0.01569	0.00625	0.010009	0.049251	1				
education	0.136538	-0.0529	0.064896	0.965384	0.100577	0.179976	0.148543	-0.00574	0.083348	1			
exper*fen	-0.00776	0.759422	0.086903	0.011333	-0.14494	0.424161	0.446665	0.666764	0.047442	-0.00413	1		
exper*nor	-0.08762	0.061417	0.812981	0.178438	-0.17574	0.281649	0.316693	0.003029	0.739585	0.09664	0.202598	1	
exper*uni	0.091198	-0.07398	0.132722	0.871861	-0.10726	0.376631	0.391193	-0.08609	0.071507	0.790853	0.079827	0.323864	1

2. Criterion Selection: Criteria to enter a variable in the model is p-value less than 0.05 (p-in) and criteria to remove a variable is p-value is 0.1 (p-out).
3. Excel Sheet “Stepwise 1” contains regression output of outcome variable wage and explanatory variable education (as education and wage had the highest correlation). The explanatory variable satisfies the criterion of p-in and hence is a good fit as per the criterion.



SUMMARY OUTPUT

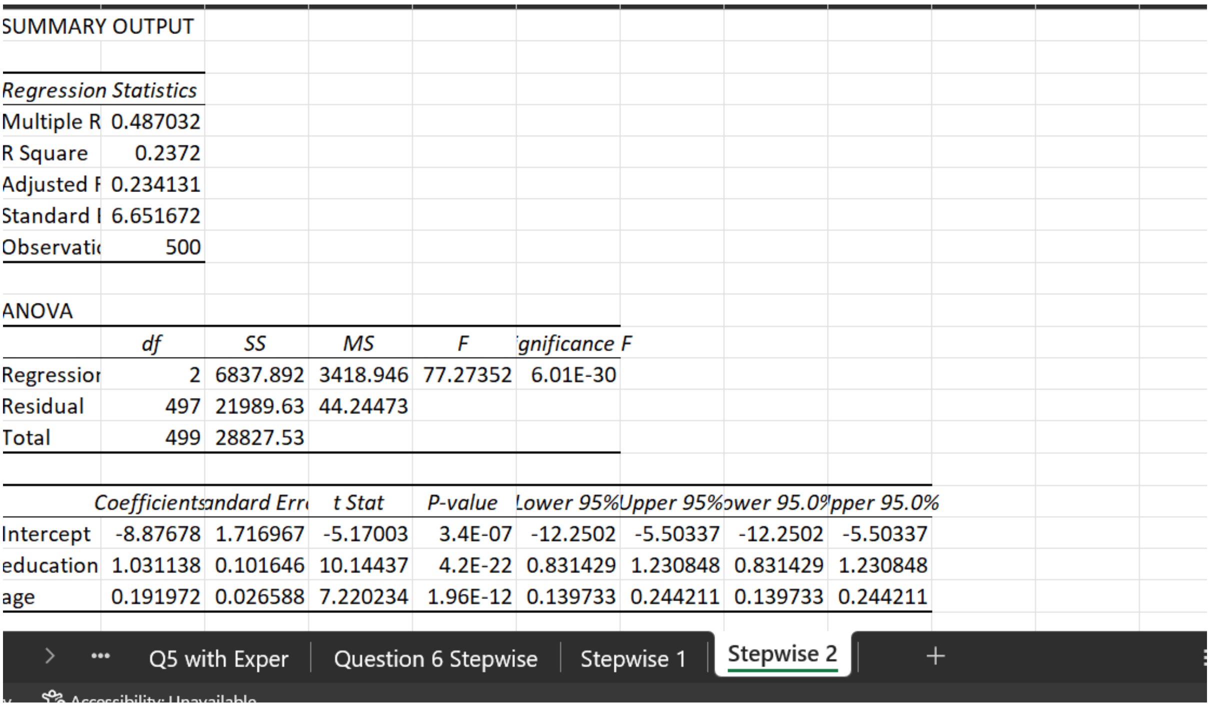
Regression Statistics	
Multiple R	0.396469174
R Square	0.157187806
Adjusted R Square	0.155495412
Standard Error	6.984807079
Observations	500

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	4531.335	4531.335	92.87896861	2.87E-20
Residual	498	24296.19	48.78753		
Total	499	28827.53			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.520900218	1.451263	-1.04798	0.29515441	-4.37225	1.330453	-4.37225	1.330453
education	1.028659454	0.106736	9.637374	2.86636E-20	0.81895	1.238369	0.81895	1.238369

4. The next highest correlated independent variable “age” is included in the model and both the independent variable education ad age satisfies the p-value criterion. Also, there has been increase in the Adjusted R value denoting inclusion of age is a good fit in the model.



SUMMARY OUTPUT

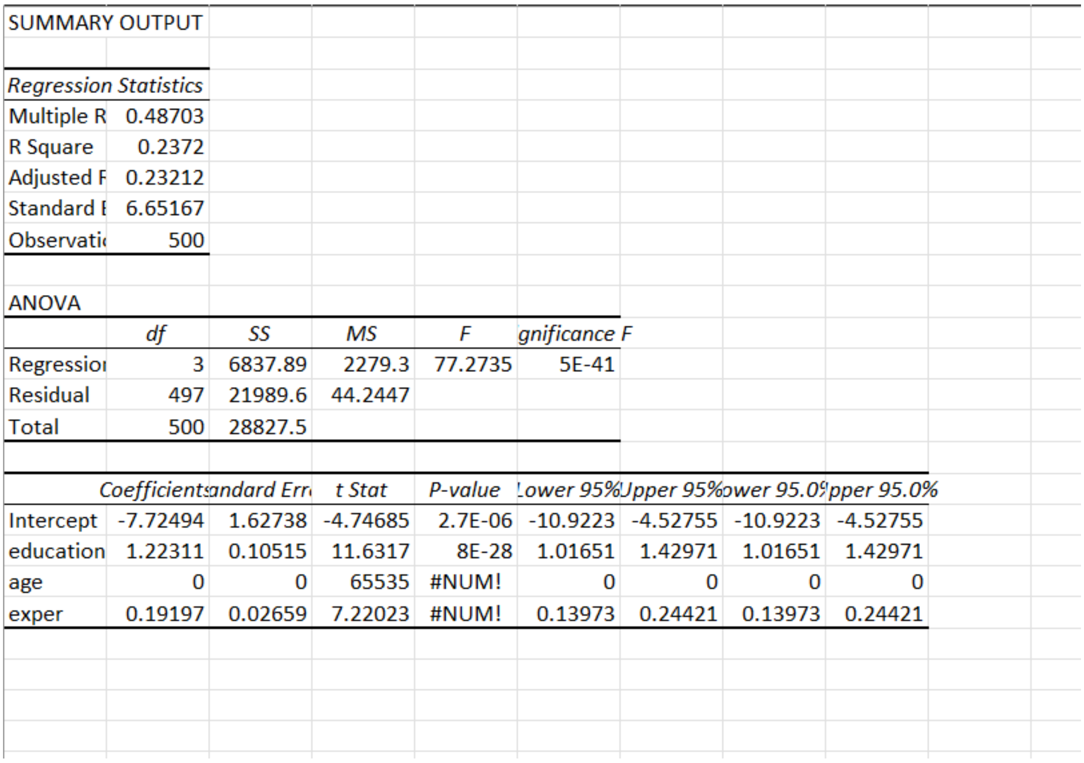
Regression Statistics	
Multiple R	0.487032
R Square	0.2372
Adjusted F	0.234131
Standard I	6.651672
Observati	500

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	2	6837.892	3418.946	77.27352	6.01E-30
Residual	497	21989.63	44.24473		
Total	499	28827.53			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-8.87678	1.716967	-5.17003	3.4E-07	-12.2502	-5.50337	-12.2502	-5.50337
education	1.031138	0.101646	10.14437	4.2E-22	0.831429	1.230848	0.831429	1.230848
age	0.191972	0.026588	7.220234	1.96E-12	0.139733	0.244211	0.139733	0.244211

5. Next highly correlated independent variable is exper. But as we know age and exper are linearly dependent variable, we are getting error in the regression report. So, I have checked the regression report excluding age and have found that inclusion of either age or exper provides the same regression output.



SUMMARY OUTPUT

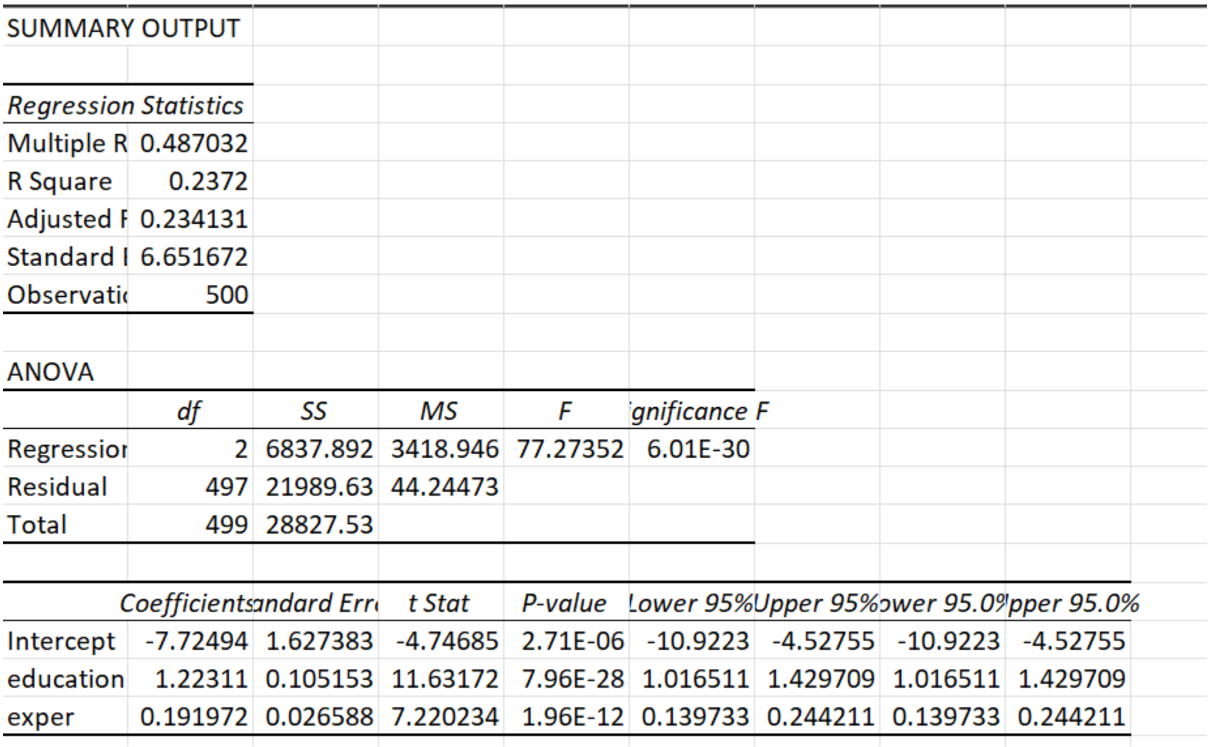
Regression Statistics	
Multiple R	0.48703
R Square	0.2372
Adjusted F	0.23212
Standard E	6.65167
Observati	500

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	3	6837.89	2279.3	77.2735	5E-41
Residual	497	21989.6	44.2447		
Total	500	28827.5			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-7.72494	1.62738	-4.74685	2.7E-06	-10.9223	-4.52755	-10.9223	-4.52755
education	1.22311	0.10515	11.6317	8E-28	1.01651	1.42971	1.01651	1.42971
age	0	0	65535	#NUM!	0	0	0	0
exper	0.19197	0.02659	7.22023	#NUM!	0.13973	0.24421	0.13973	0.24421

Regression output excluding age and including exper.



SUMMARY OUTPUT

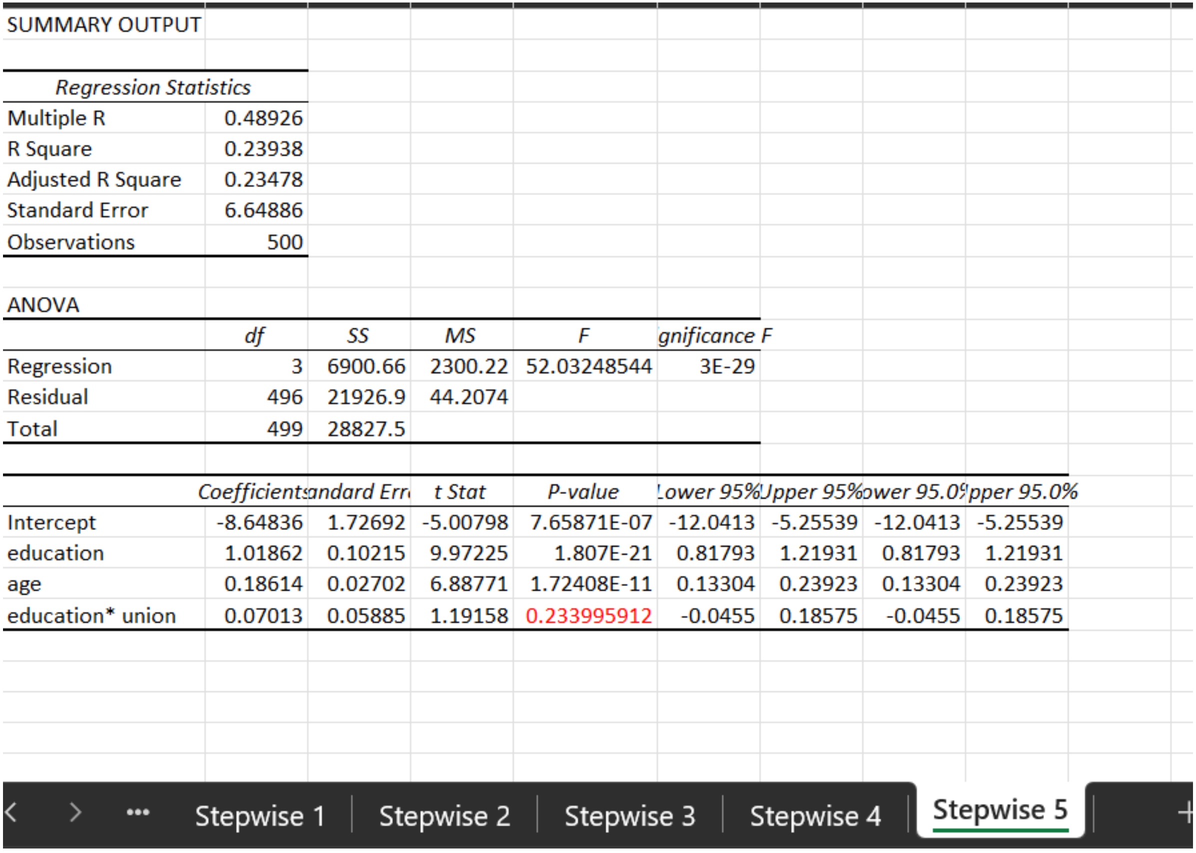
Regression Statistics	
Multiple R	0.487032
R Square	0.2372
Adjusted F	0.234131
Standard I	6.651672
Observatic	500

ANOVA

	df	SS	MS	F	gnificance F
Regressior	2	6837.892	3418.946	77.27352	6.01E-30
Residual	497	21989.63	44.24473		
Total	499	28827.53			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-7.72494	1.627383	-4.74685	2.71E-06	-10.9223	-4.52755	-10.9223	-4.52755
education	1.22311	0.105153	11.63172	7.96E-28	1.016511	1.429709	1.016511	1.429709
exper	0.191972	0.026588	7.220234	1.96E-12	0.139733	0.244211	0.139733	0.244211

6. Testing the model with next highly correlated variable “education*union”. The p- value of this variable is greater than p-in hence cannot be included in the model. Now testing with next highly correlated variable.



SUMMARY OUTPUT

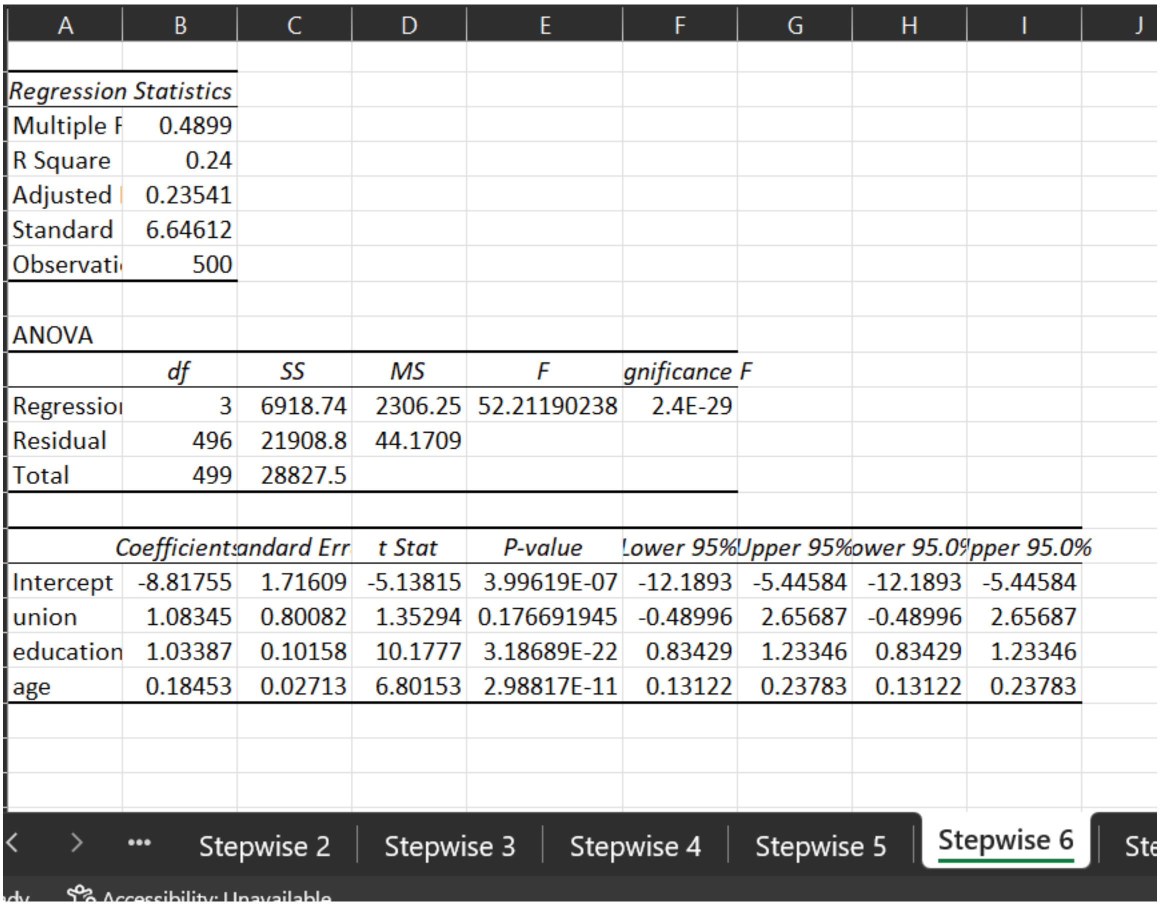
Regression Statistics	
Multiple R	0.48926
R Square	0.23938
Adjusted R Square	0.23478
Standard Error	6.64886
Observations	500

ANOVA

	df	SS	MS	F	gnificance F
Regression	3	6900.66	2300.22	52.03248544	3E-29
Residual	496	21926.9	44.2074		
Total	499	28827.5			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-8.64836	1.72692	-5.00798	7.65871E-07	-12.0413	-5.25539	-12.0413	-5.25539
education	1.01862	0.10215	9.97225	1.807E-21	0.81793	1.21931	0.81793	1.21931
age	0.18614	0.02702	6.88771	1.72408E-11	0.13304	0.23923	0.13304	0.23923
education* union	0.07013	0.05885	1.19158	0.233995912	-0.0455	0.18575	-0.0455	0.18575

7. For below independent variables the p-value p-in criteria didn’t satisfy. Hence were not included in the model. Also, inclusion of these variable did not make any other previously added variable insignificant i.e., p value > pout. So, no variable was not excluded.
    1. Union



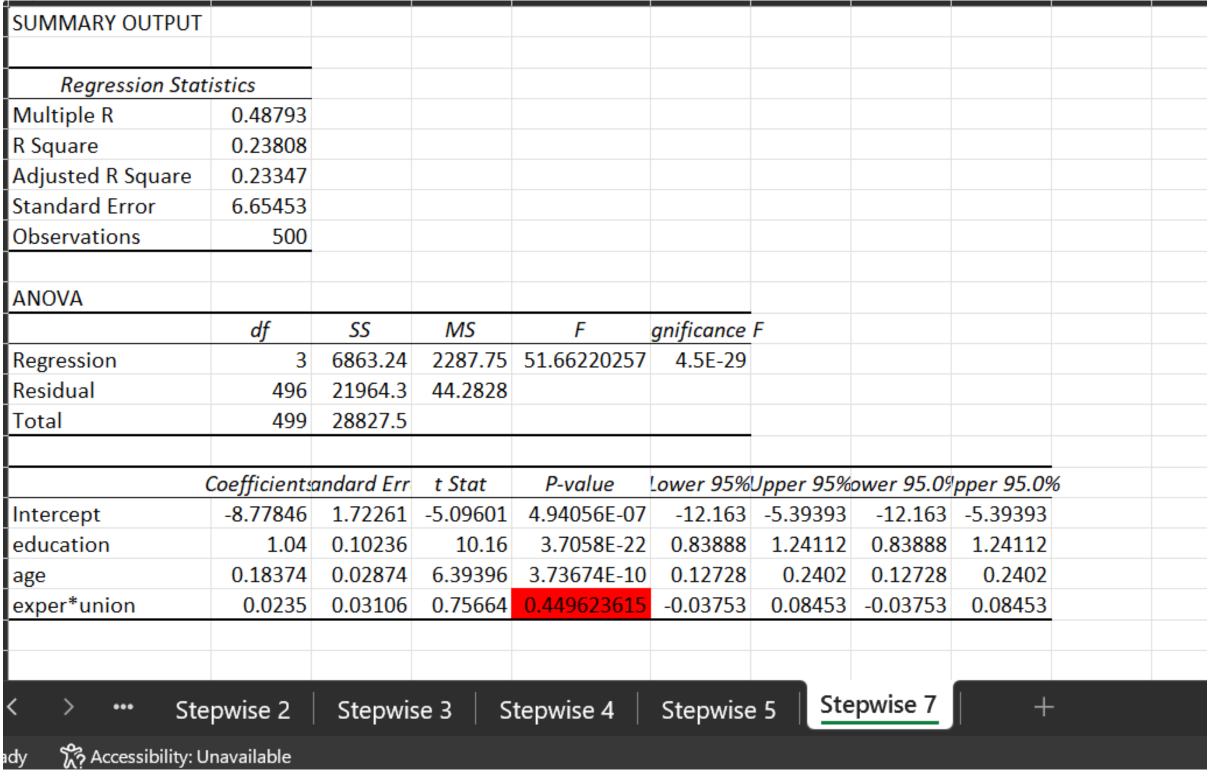
Regression Statistics	
Multiple F	0.4899
R Square	0.24
Adjusted	0.23541
Standard	6.64612
Observati	500

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	3	6918.74	2306.25	52.21190238	2.4E-29
Residual	496	21908.8	44.1709		
Total	499	28827.5			

	Coefficient	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-8.81755	1.71609	-5.13815	3.99619E-07	-12.1893	-5.44584	-12.1893	-5.44584
union	1.08345	0.80082	1.35294	0.176691945	-0.48996	2.65687	-0.48996	2.65687
education	1.03387	0.10158	10.1777	3.18689E-22	0.83429	1.23346	0.83429	1.23346
age	0.18453	0.02713	6.80153	2.98817E-11	0.13122	0.23783	0.13122	0.23783

2. exper*union



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.48793
R Square	0.23808
Adjusted R Square	0.23347
Standard Error	6.65453
Observations	500

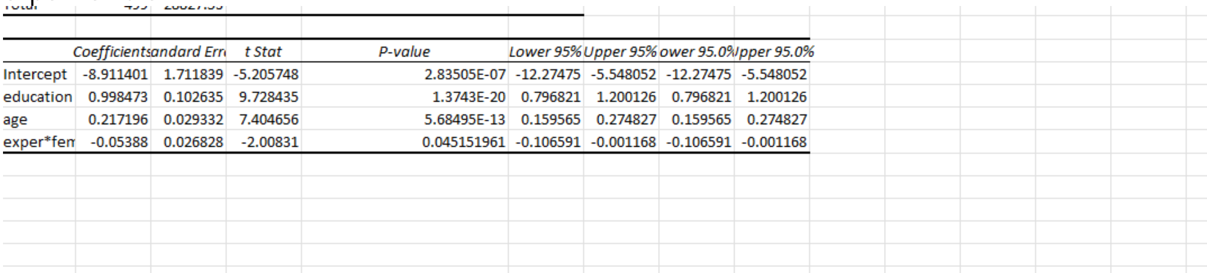
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	6863.24	2287.75	51.66220257	4.5E-29
Residual	496	21964.3	44.2828		
Total	499	28827.5			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-8.77846	1.72261	-5.09601	4.94056E-07	-12.163	-5.39393	-12.163	-5.39393
education	1.04	0.10236	10.16	3.7058E-22	0.83888	1.24112	0.83888	1.24112
age	0.18374	0.02874	6.39396	3.73674E-10	0.12728	0.2402	0.12728	0.2402
exper*union	0.0235	0.03106	0.75664	0.449623615	-0.03753	0.08453	-0.03753	0.08453

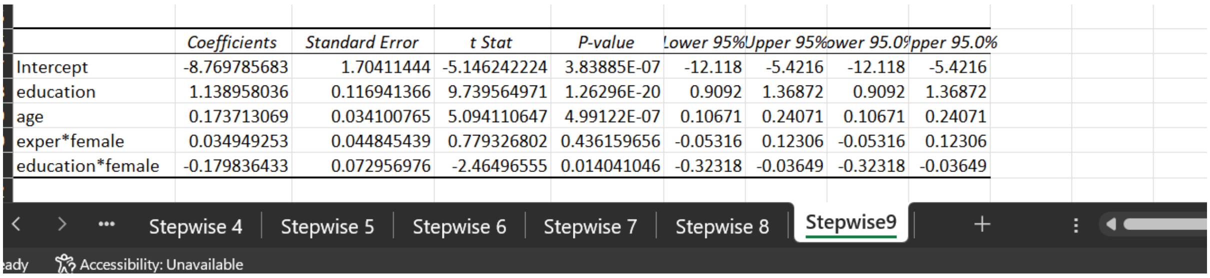
8. For below independent variables the p-value p-in criteria satisfy. Hence were included in the model. Also, inclusion of these variable did not make any other previously added variable insignificant i.e., p value > pout. So, no variable was not excluded. Also, inclusion of these variables increased Adjusted R value proving these variables are good fit in the model.

1. exper*female



	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-8.911401	1.711839	-5.205748	2.83505E-07	-12.27475	-5.548052	-12.27475	-5.548052
education	0.998473	0.102635	9.728435	1.3743E-20	0.796821	1.200126	0.796821	1.200126
age	0.217196	0.029332	7.404656	5.68495E-13	0.159565	0.274827	0.159565	0.274827
exper*fen	-0.05388	0.026828	-2.00831	0.045151961	-0.106591	-0.001168	-0.106591	-0.001168

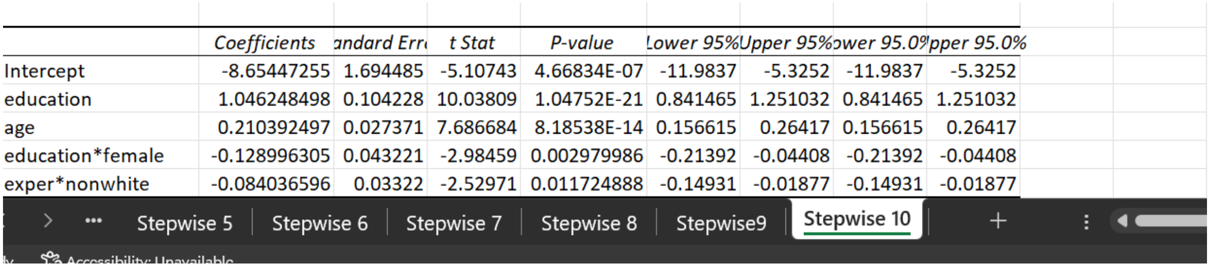
2. education*female



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-8.769785683	1.70411444	-5.146242224	3.83885E-07	-12.118	-5.4216	-12.118	-5.4216
education	1.138958036	0.116941366	9.739564971	1.26296E-20	0.9092	1.36872	0.9092	1.36872
age	0.173713069	0.034100765	5.094110647	4.99122E-07	0.10671	0.24071	0.10671	0.24071
exper*female	0.034949253	0.044845439	0.779326802	0.436159656	-0.05316	0.12306	-0.05316	0.12306
education*female	-0.179836433	0.072956976	-2.46496555	0.014041046	-0.32318	-0.03649	-0.32318	-0.03649

But the inclusion of education* female variable in the model increase the p value of exper*female > pout. So, excluding exper*female from the model.

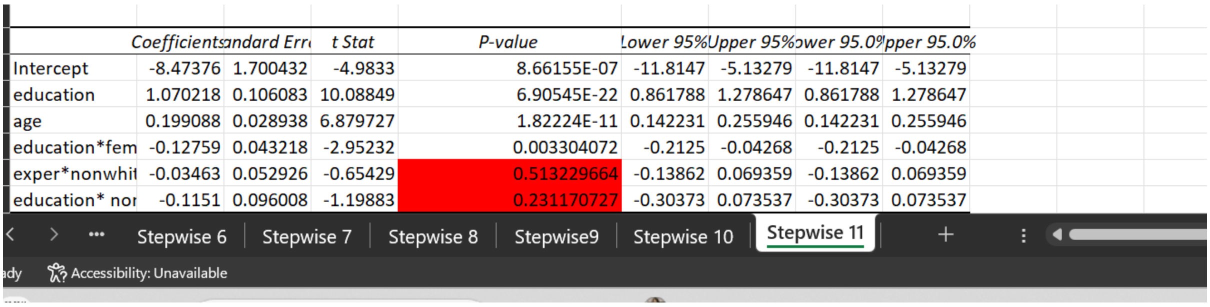
3. Exper*nonwhite



	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-8.65447255	1.694485	-5.10743	4.66834E-07	-11.9837	-5.3252	-11.9837	-5.3252
education	1.046248498	0.104228	10.03809	1.04752E-21	0.841465	1.251032	0.841465	1.251032
age	0.210392497	0.027371	7.686684	8.18538E-14	0.156615	0.26417	0.156615	0.26417
education*female	-0.128996305	0.043221	-2.98459	0.002979986	-0.21392	-0.04408	-0.21392	-0.04408
exper*nonwhite	-0.084036596	0.03322	-2.52971	0.011724888	-0.14931	-0.01877	-0.14931	-0.01877

9. Education*nonwhite:

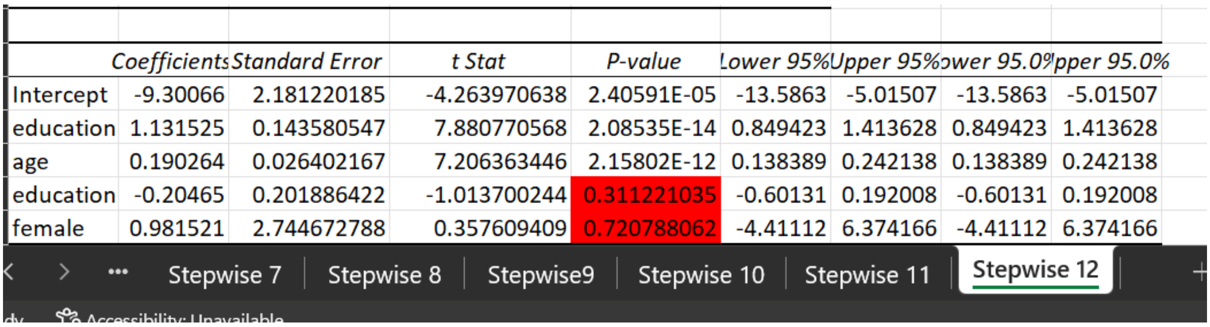
Inclusion of education*nonwhite variable increased the p value of exper*nonwhite > pout. Hence, removing the variable from model. Also, education*nonwhite pvalue > pin so the variable cannot be included in the model.



	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-8.47376	1.700432	-4.9833	8.66155E-07	-11.8147	-5.13279	-11.8147	-5.13279
education	1.070218	0.106083	10.08849	6.90545E-22	0.861788	1.278647	0.861788	1.278647
age	0.199088	0.028938	6.879727	1.82224E-11	0.142231	0.255946	0.142231	0.255946
education*fem	-0.12759	0.043218	-2.95232	0.003304072	-0.2125	-0.04268	-0.2125	-0.04268
exper*nonwhit	-0.03463	0.052926	-0.65429	0.513229664	-0.13862	0.069359	-0.13862	0.069359
education* nor	-0.1151	0.096008	-1.19883	0.231170727	-0.30373	0.073537	-0.30373	0.073537

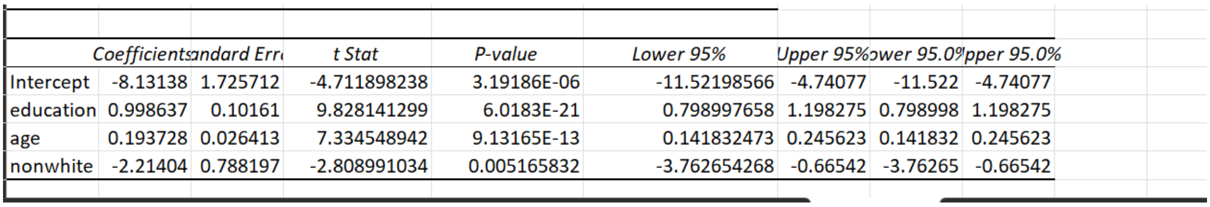
10. Female:

Inclusion of female variable increased the p value of education*female > pout. Hence, removing the variable from model. Also, female pvalue > pin so the variable cannot be included in the model.



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-9.30066	2.181220185	-4.263970638	2.40591E-05	-13.5863	-5.01507	-13.5863	-5.01507
education	1.131525	0.143580547	7.880770568	2.08535E-14	0.849423	1.413628	0.849423	1.413628
age	0.190264	0.026402167	7.206363446	2.15802E-12	0.138389	0.242138	0.138389	0.242138
education	-0.20465	0.201886422	-1.013700244	0.311221035	-0.60131	0.192008	-0.60131	0.192008
female	0.981521	2.744672788	0.357609409	0.720788062	-4.41112	6.374166	-4.41112	6.374166

11. Nonwhite:
    Non white p value < pin and also the inclusion of non white variable does not affect existing variable in the model.



	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-8.13138	1.725712	-4.711898238	3.19186E-06	-11.52198566	-4.74077	-11.522	-4.74077
education	0.998637	0.10161	9.828141299	6.0183E-21	0.798997658	1.198275	0.798998	1.198275
age	0.193728	0.026413	7.334548942	9.13165E-13	0.141832473	0.245623	0.141832	0.245623
nonwhite	-2.21404	0.788197	-2.808991034	0.005165832	-3.762654268	-0.66542	-3.76265	-0.66542

The Final regression model is:

Wage = $-8.1313788983491 + 0.998636574402803 * education + 0.193727704140893 * age — 2.21403793976649 * nonwhite$

Question 7:

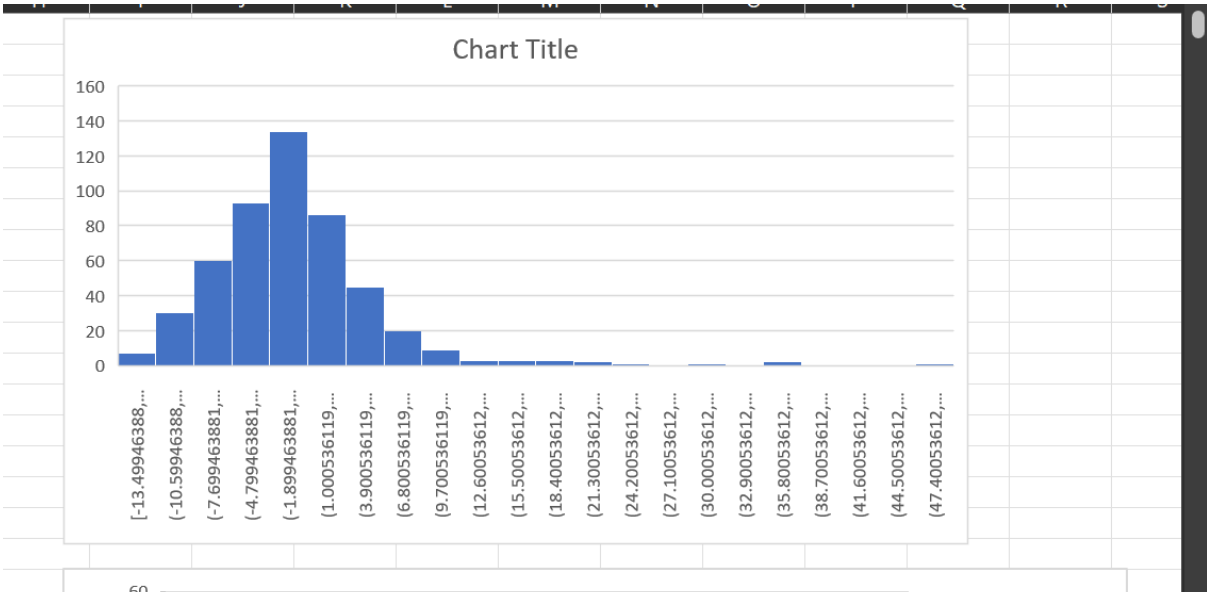
Interpretation of above variables

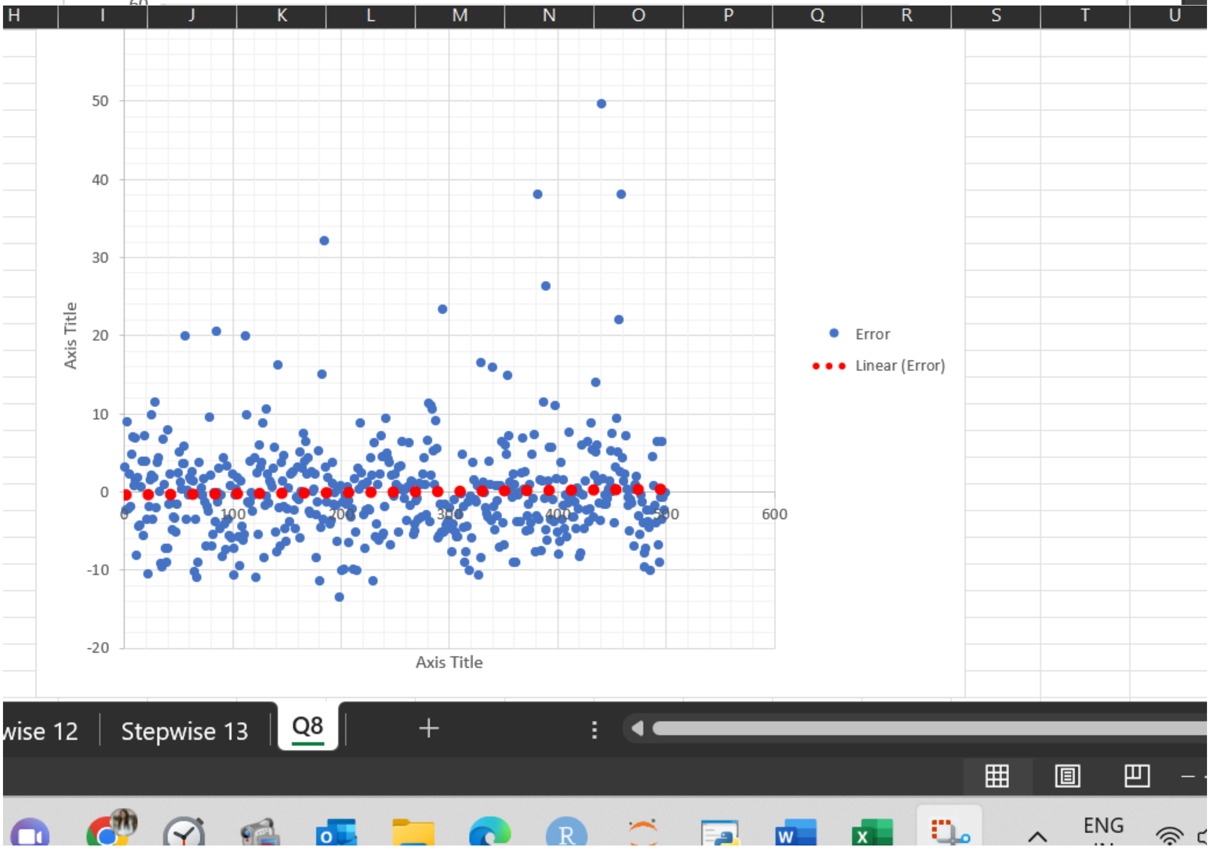
1. Intercept: It defines that wage of individual with no education , just born and white race will get wage -8.1313788983491 dollars less. The estimated wage in dollars when all the dependent variable is zero. The interpretation of intercept will be not meaningful in this context.
2. Education: The coefficient for "education" is 0. 998636574402803, holding all other variables constant. This indicates that additional a unit increase in education is associated with a 0.998636574402803 dollars increase in the wage variable. This coefficient suggests a positive relationship between education and the wage variable.
3. Age: The coefficient for "age" is 0.193727704140893. It suggests that an additional increase in age is associated with a 0.193727704140893 dollars increase in the dependent variable, holding other constant.
4. Nonwhite: The coefficient for nonwhite is $— 2.21403793976649$. This indicates the average difference in the predicted wage in dollars when the individual belongs to non white race i.e., the wage for non white will be $— 2.21403793976649$ dollars less than the white race individual.

Question 8:

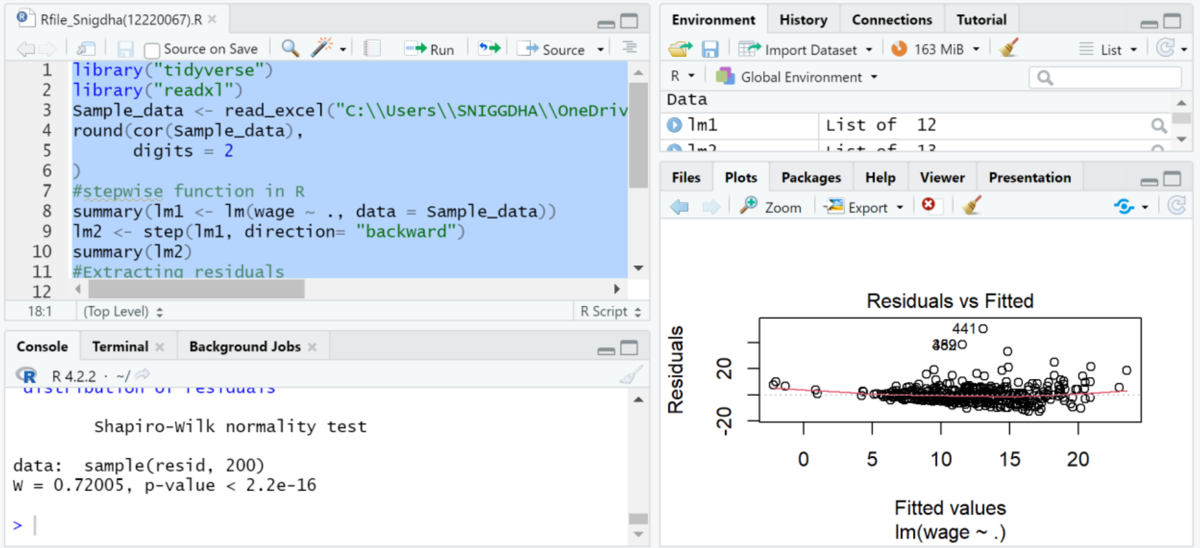
Main Assumptions:

1. Linear Assumption is correct.
2. E[$\varepsilon|X$] =0
3. Homoskedasticity: Var E[$\varepsilon|X$] = σ_ε^2
4. Corr [$\varepsilon_i, \varepsilon_j$] = 0 for all i not equal to j.
5. Normality of errors: $\varepsilon|X \sim N\ (\ 0, \sigma_\varepsilon^2\ )$





Using Shapiro-Wilk normality test



```
library("tidyverse")
library("readxl")
Sample_data <- read_excel("C:\\Users\\SNIGGDHA\\OneDriv
round(cor(Sample_data),
      digits = 2
)
#stepwise function in R
summary(lm1 <- lm(wage ~ ., data = Sample_data))
lm2 <- step(lm1, direction= "backward")
summary(lm2)
#Extracting residuals
```

```
        Shapiro-Wilk normality test

data:  sample(resid, 200)
W = 0.72005, p-value < 2.2e-16

>
```

As per the Main Assumptions Mean error is not zero and by visualisation it is right skewed, hence not normal. Also, Residual scatter plot shows no pattern with few outliers.

Confirming the same with Shapiro-Wilk normality test

Since the p-value is extremely small 2.2e-16, much smaller than the significance level 0.05, so we reject the null hypothesis. Thus, we conclude that the data does not follow a normal distribution.

Question 9:

Regression equation is:

Wage = $-8.1313788983491 + 0.998636574402803 * education + \ 0.193727704140893 * age — 2.21403793976649 * nonwhite$

The regression equation obtained from Stepwise Method, does not include gender/female variable. So, the wage variable does not have any significant effect of the gender/ female variable. Hence, there is no gender bias in the wage rate.

Question 10:

Regression equation is:

Wage = $-8.1313788983491 + 0.998636574402803 * education + \ 0.193727704140893 * age — 2.21403793976649 * nonwhite$

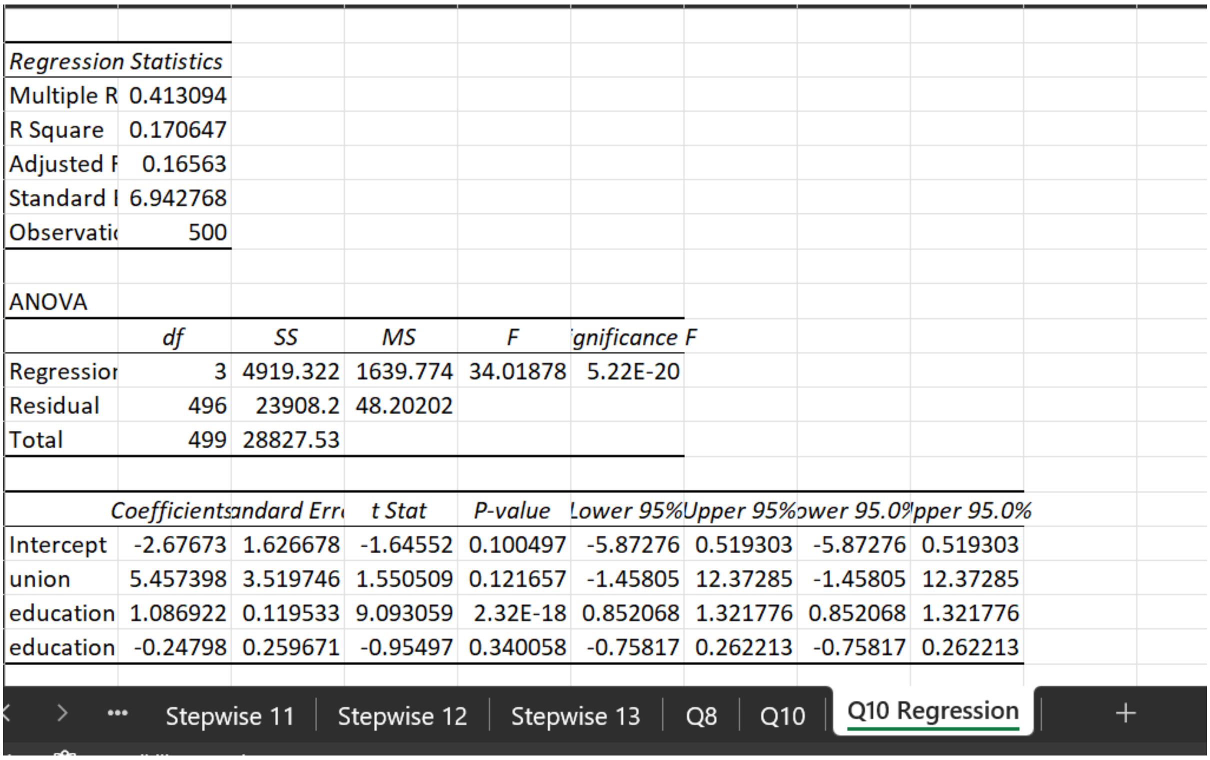
The regression equation obtained from Stepwise Method, does include union variable but includes education variable. As per the equation obtained wage rate is not dependent on the union variable but dependent on the education variable, which affects the wage in dollars by 0. 998636574402803 dollars with every additional unit increase in education. Thus, there is no advantage of union membership to those with higher education.

Question 10:

If this question is not dependent on Regression Equation2.

Regression equation used for this is:

Wage = $-2.6767261576117 + 5.457397975786 * union + \ 1.08692227644397 * education - -0.24797861176055 * (education * union)$



Regression Statistics	
Multiple R	0.413094
R Square	0.170647
Adjusted F	0.16563
Standard I	6.942768
Observati	500

ANOVA

	df	SS	MS	F	ignificance F
Regressior	3	4919.322	1639.774	34.01878	5.22E-20
Residual	496	23908.2	48.20202		
Total	499	28827.53			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-2.67673	1.626678	-1.64552	0.100497	-5.87276	0.519303	-5.87276	0.519303
union	5.457398	3.519746	1.550509	0.121657	-1.45805	12.37285	-1.45805	12.37285
education	1.086922	0.119533	9.093059	2.32E-18	0.852068	1.321776	0.852068	1.321776
education	-0.24798	0.259671	-0.95497	0.340058	-0.75817	0.262213	-0.75817	0.262213

The coefficient for the "union" variable is 5.457397975786, with a p-value of 0.1216, indicating union membership has no statistically significant positive effect on wages.

The coefficient for the "education" variable is 1.08692227644397, with a p-value of 2.31968064083756E-18, indicating education has highly a statistically significant positive effect on wages holding other variables constant. Individuals with education can expect, on average, a 1.08692227644397 dollars increase in their wages compared to people with no education.

The coefficient for the interaction term "education*union" is -0. 24797861176055, with a p-value of 0.340058. This indicates that the interaction between education and union membership is not statistically significant at conventional levels ($p < 0.05$). Hence, there is no advantage of union membership to those with higher education.