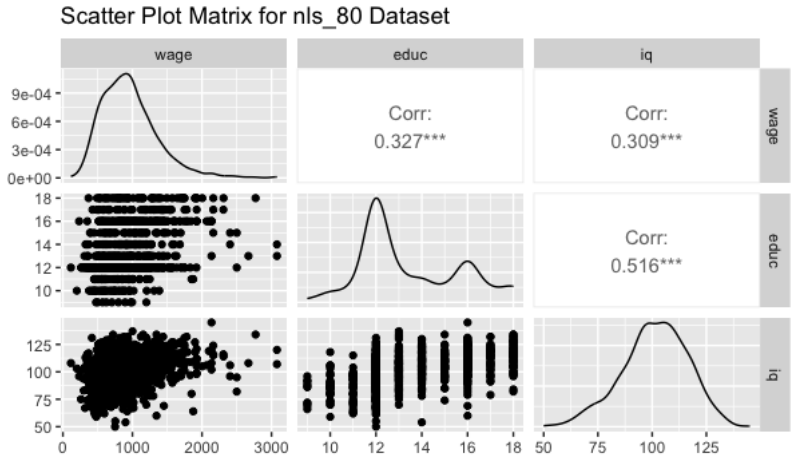
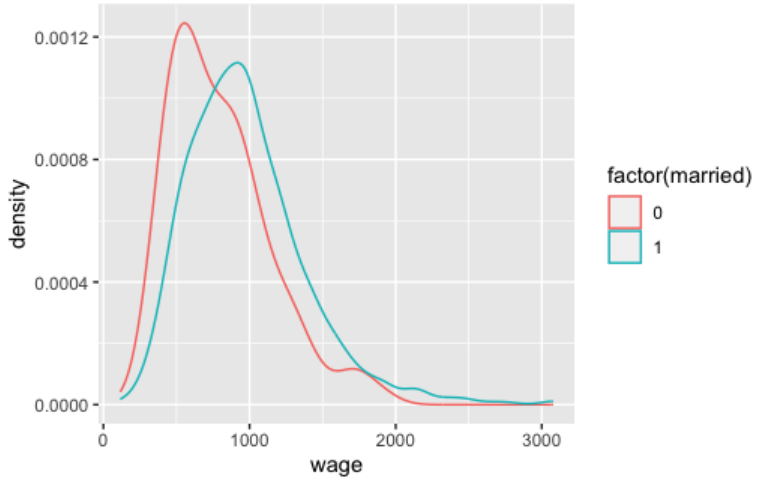
Question 2:

1. scatterplot matrix for the wage, education, and IQ.



we can see IQ and education have positive correlation with wage, the correlation between IQ and education has stronger correlation than others. IQ is normal distribution which mean value is around 100, and wage is right skewed distribution.

2. The figure below is the density curves of wages on married or not.



3. Table about mean, standard deviation, min, and max of wages by South and Urban.

	Wages of south	Wages of urban
Mean	880.24	1008.24
SD	427.43	412.93
Min	200.00	115.00
Max	3078.00	3078.00

4.

(1). Run the regression of the wage on education here is the results below:

	wage		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	146.95	−5.56 — 299.47	0.059
educ	60.21	49.04 — 71.39	**<0.001**
Observations	935		
R^2 / R^2 adjusted	0.107 / 0.106		

Problems I find in this regression, the intercept of p-value > 0.05, so the intercept is not significant, and the r-square is too low as 0.1, so I decided to drop intercept and make reg_2 which has no intercept for the regression.

(2) the results for no intercept regression.

	wage		
Predictors	*Estimates*	*CI*	*p*
educ	70.84	69.04 — 72.64	**<0.001**
Observations	935		
R^2 / R^2 adjusted	0.865 / 0.864		

Estimator beta is 70.84, and the p-value < 0.05, r-squared value is 0.864 which is much bigger than the results of reg_1, which means this linear model is more fitted the true data and could explain more about variance in wages.

5. Run another regression of wage on education and IQ. The regression of wage on education and IQ with no intercept.

	wage		
Predictors	*Estimates*	*CI*	*p*
educ	38.25	26.55 — 49.95	**<0.001**
iq	4.39	2.83 — 5.95	**<0.001**
Observations	935		
R^2 / R^2 adjusted	0.869 / 0.869		

6. Between 2 regressions, the coefficients of wage on educ is different, when we add one more variable like IQ, which IQ could have correlation with education, and we know it from the scatter plot matrix.it cause the estimator is smaller than the first simple linear regression just between educ and wage. and it may cause multicollinearity and affects the coefficients, and we may not trust the p-values. For those reasons I make a variance inflation factor test for these 2 factors. The VIF results shows the value of education and IQ are bigger than 10, that means when we run a regression, we should not put them together in a regression.