Quantitative Data Analysis - Assignment 1

Question 1

Use the NELS dataset (on educational achievement and characteristics of children) that we have used in the classes.

Use any nominal level variable and create a bar chart to present the frequency of this variable's values; report the frequency of values of the variable.

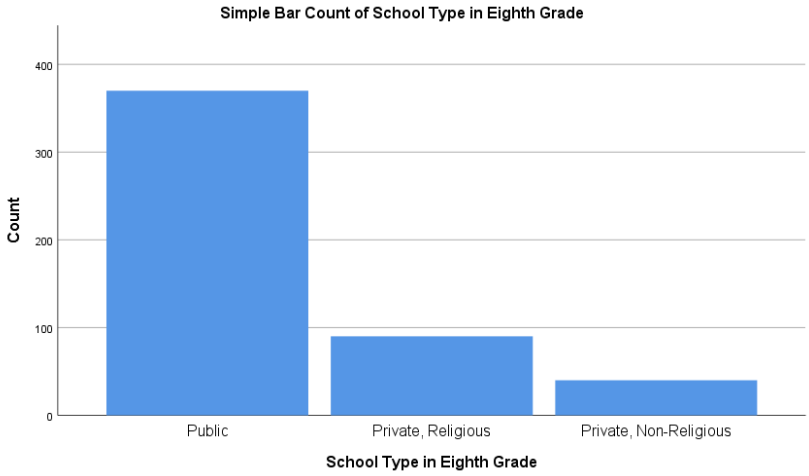
Use the socio-economic status variable as an interval/ratio variable and create histogram of the distribution of values of this variable; report the mean, median and mode, and comment on the shape of the distribution of this variable (compared to a bell shaped/normal cure).

Perform a linear transformation of the variable socio-economic status by calculating a Z score from the original socio-economic status variable.

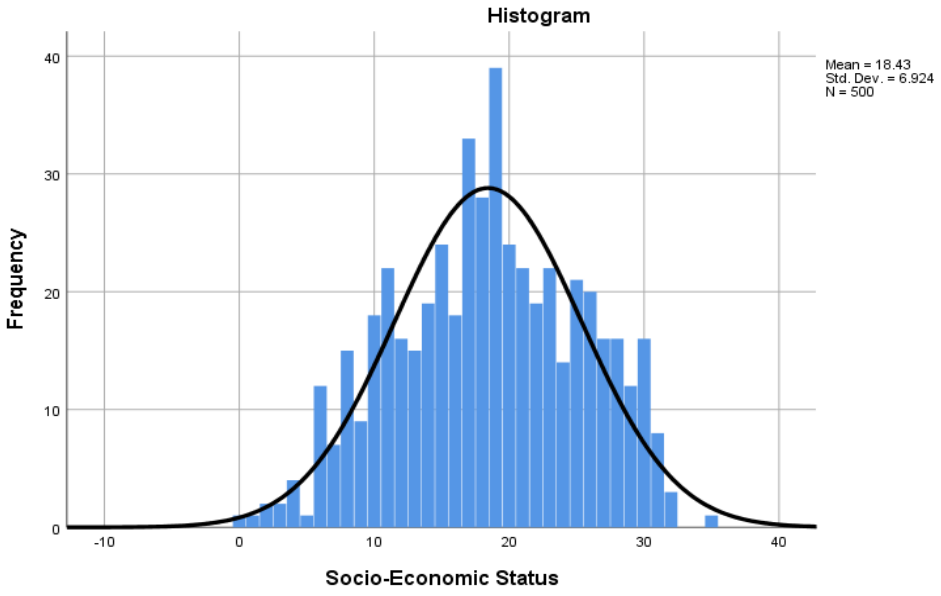
Produce a scatterplot of the relationship between the variable socio-economic status and the Z score of socio-economic status variable.

Conduct a non-linear transformation of socio-economic status variable by squaring it. Create a scatterplot of the relationship between socio-economic status and the variable socioeconomic status squared.

Answer for question 1:



The chosen nominal value is "The school; type in eight grade with "public", "private, religious" and "private, non-religious" being the options.
The Bar chart is demonstrating that in the eighth grade 380 pupils have been attending a public school, 90 have been attending a private, religious school and 30 have been attending a private, non- religious school.



Statistics

Socio-Economic Status

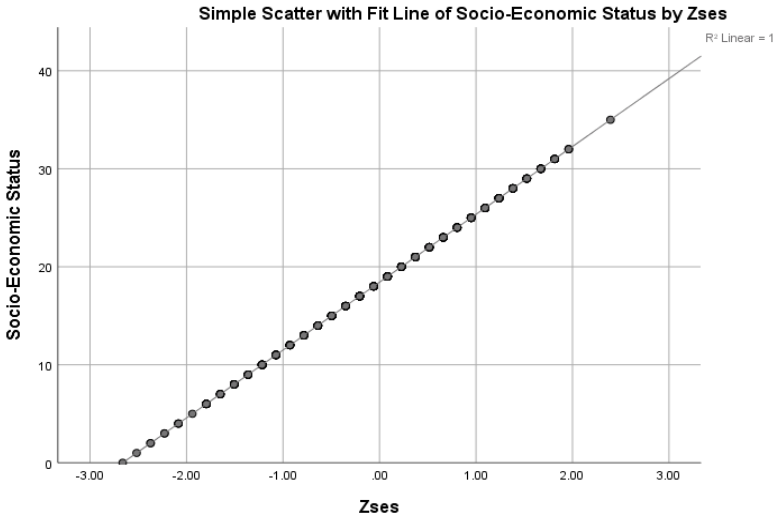
N	Valid	500
	Missing	0
Mean		18.43
Median		19.00
Mode		19
Skewness		-.112
Std. Error of Skewness		.109

The Histogram and the table above represent the data about distribution of socio-economic status of the pupils.
The data indicates that the mean value is 18.43; the median value is 19; the mode value is 19. The distribution is bell shaped which indicates that it is a normal distribution. We can also see that the distribution has a slight negative skew of -0.122 as the mean value is less than the median value.

```
COMPUTE Zses =(ses - 18.43) / 6.924.EXECUTE.
```

The command above shows the transformation of the Socio-Economic Status (ses) variable by calculating a Z score (Zses). In order to calculate the Zses, I have subtracted the mean from the ses variable score and divide the result by the standard deviation.

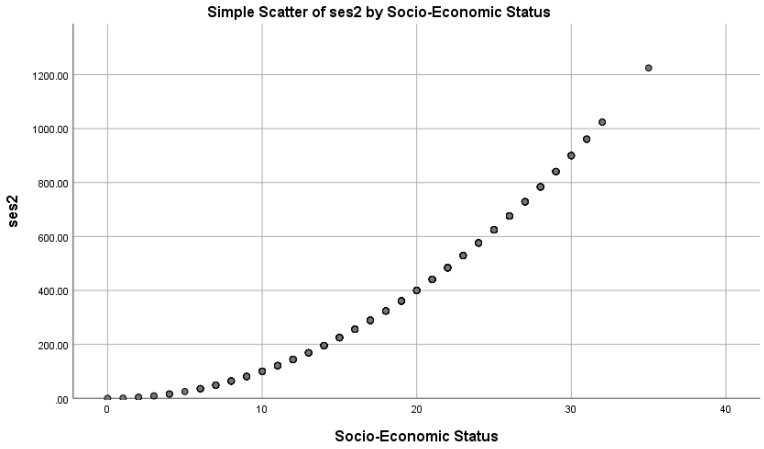
Below we can see a scatter plot of the ses and the Zses variables. The scatter plot indicates that it is a linear variable transformation.



Consequently, I have conducted a non - linear transformation of ses variable by using the command below and creating an additional column of ses2 variable.

```
COMPUTE ses2=ses ** 2.
EXECUTE.
```

To demonstrate the non-linear transformation of the variable we can construct a scatter plot of a ses and ses2 variable.



The relationship between the ses and ses2 variables shown above represents a non-linear quadratic transformation and has a different shape to the linear transformation shown earlier.

Question 2

Use the NELS dataset.

Carry out a t test for a hypothesised value of the mean for socio-economic status. Use SPSS compare means, one sample command.

Perform a two tailed test at the 0.05 level of significance, report and comment on your results.

Construct a confidence interval for the mean of socio-economic status at the 95 per cent level of confidence.

Create a scatterplot for two interval/ratio level variables (socio-economic status and another interval/ratio level variable).The second variable should be one that might be argued have a theoretical relationship with socio-economic status (for example positively or negatively correlated).

Calculate the Pearson Correlation Coefficient and perform a one tailed test of the null at the 0.05 level of significance, comment on your results.

Answer for question 2:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Socio-Economic Status	500	18.43	6.924	.310

One-Sample Test

	Test Value = 18					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Socio-Economic Status	1.402	499	.162	.434	- .17	1.04

Above is a one-sample t test for the hypothesised mean value of 18 for the socio-economic status variable. This is 0.434 from the real mean value as shown by the "Mean Difference" cell. Given that the Std. Error mean is 0.310, we deduce that the t-value for a hypothesised value of 18 is 1.402. "df" or the degree of freedom for the test is 499 as it is defined as N-1.

The significance value we obtained (0.162) is larger than 0.05 meaning that our hypothesised mean value of 18 is statistically significant. The t test also indicates that we can be 95 % certain that the actual population mean value lays in the range of 17.83 (-0.17) and 19.04 (+1.04) from our hypothesised value of 18. Knowing that the "edges" of the 95% accuracy range lay within 0.17 and 1.04 we can select a hypothesised mean value of 17.83 that gives us a 0.05 level of significance. This is the value we will then use to conduct a two-tailed test "at the 0.05 level of significance" as required.

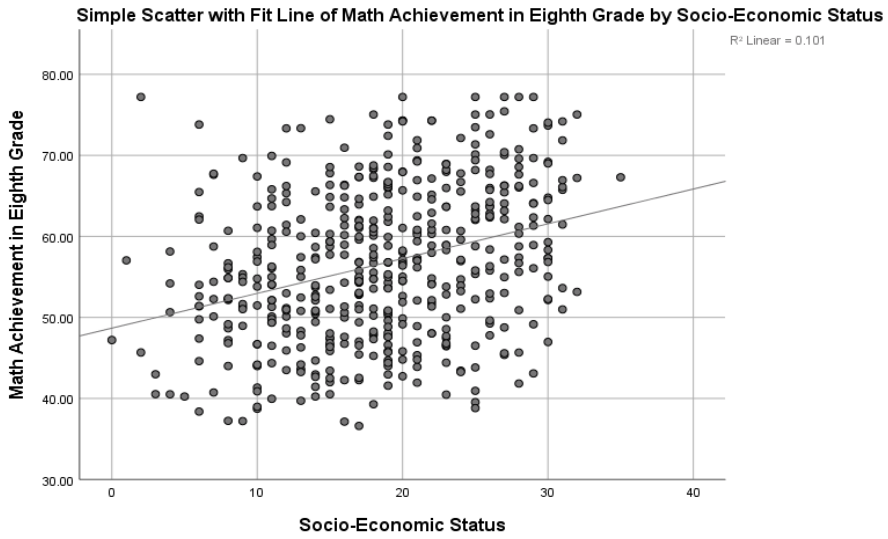
One-Sample Test

	Test Value = 17.83					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Socio-Economic Status	1.951	499	.052	.604	.00	1.21

The reason why the p value is not exactly 0.05 is due to the fact that SPSS t test calculations are done to the three decimal places. The value chosen lays around the boundary of significance and if the p value were to be below 0.05, we would be justified in assuming that the two means are statistically different.

The 95% confidence interval lays 0.00 and 1.21 away from our hypothesised value of 17.83.

In this section of the question, I would like to compare the relationship between the socio-economic status of pupils and their Maths performance in the eight grade. My hypothesis is that there is a positive corelation between these variables, meaning that the higher is the socio-economic status the higher is the maths performance in the eight grade. The scatterplot below visually demonstrates the relationship between these variables



The value of the R^2 is 0.101 which is not very high and thus the wellness of the fit is not good. However, we can investigate the Pearson Coefficient to assess the correlation.

Correlations

		Socio-Economic Status	Math Achievement in Eighth Grade
Socio-Economic Status	Pearson Correlation	1	.318**
	Sig. (1-tailed)		.000
	N	500	500
Math Achievement in Eighth Grade	Pearson Correlation	.318**	1
	Sig. (1-tailed)	.000	
	N	500	500

**. Correlation is significant at the 0.01 level (1-tailed).

Above is the Pearson Correlation Coefficient calculation for the socio-economic status and the math achievement in the eighth grade. We can see that the one tailed test obtained a p-value of 0.000 meaning that the correlation is statistically significant as it is lower than 0.05. Nonetheless the correlation coefficient between socio-economic status and math achievement in eighth grade is 0.318 which is a value well below 1, indicating to us that the correlation between these variables is weak. Thus we conclude that there is a correlation but its is a weak one.

Question 3

Use the NELS dataset.

Carry out a bivariate linear regression for the relationship between an outcome (Y) and one independent (X) variable.

Report the results of a two tailed t test on the regression coefficient for X and comment on the results.

Report the R square of the regression and comment on the results, illustrate your comments with a scatterplot of the relationship between the variables X and Y.

Calculate predicted values for the Y variable, work out the correlation between the predicted Y and the actual Y in the data, report the correlation coefficient and show the relationship in a scatterplot of the relationship between predicted Y and actual Y.

Answer for question 3:

To perform a bivariate linear regression, I would like to select the variables of Reading Achievement and Social Studies achievement in eight grade. It seems that in order for a pupil to do well in a Social studies class they would have to have a good reading achievement. Thus, the reading achievement in eight grade is a variable x and the social studies achievement in eight grade is a variable y. Here is a bivariate linear regression of the relationship.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Reading Achievement in Eighth Grade[b]	.	Enter

a. Dependent Variable: Social Studies Achievement in Eighth Grade

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.595[a]	.355	.353	7.08352

a. Predictors: (Constant), Reading Achievement in Eighth Grade

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13563.332	1	13563.332	270.314	.000[b]
	Residual	24686.740	492	50.176		
	Total	38250.072	493			

a. Dependent Variable: Social Studies Achievement in Eighth Grade

b. Predictors: (Constant), Reading Achievement in Eighth Grade

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	22.289	2.050		10.874	.000
	Reading Achievement in Eighth Grade	.594	.036	.595	16.441	.000

a. Dependent Variable: Social Studies Achievement in Eighth Grade

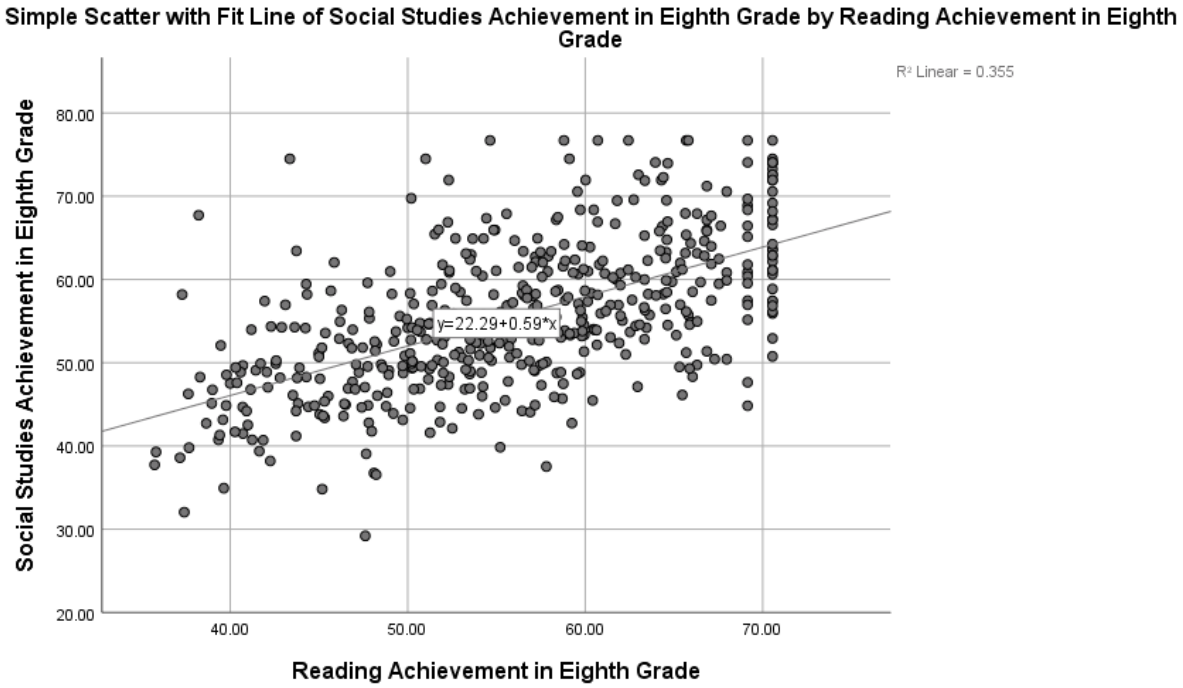
The coefficients of the line are $b_0 = 22.289$ and $b_1 = 0.594$ giving us the expression of $y = 22.289 + 0.594x$. We may primarily focus on the gradient of the line b_1 and ignore the intercept as the data at $x = 0$ is not useful to us. Thus, when performing a t test on this regression we should focus on the gradient value, i.e., $b_{1.} = 0.594$

Let us suppose the two possible values of b_1 at 0.05 statistical significance:

- $b_1 = 0$ meaning there is no correlation between x and y (Null Hypothesis)
- b_1 not equal to 0, in which case there is a correlation between x and y (Alternate Hypothesis)

The t value of b_1 is 16.441 and the p value is 0.000. Because the p value is lower than 0.05, we ought to reject the null hypothesis and assume that b_1 greater than 0, meaning that there is a correlation between x and y.

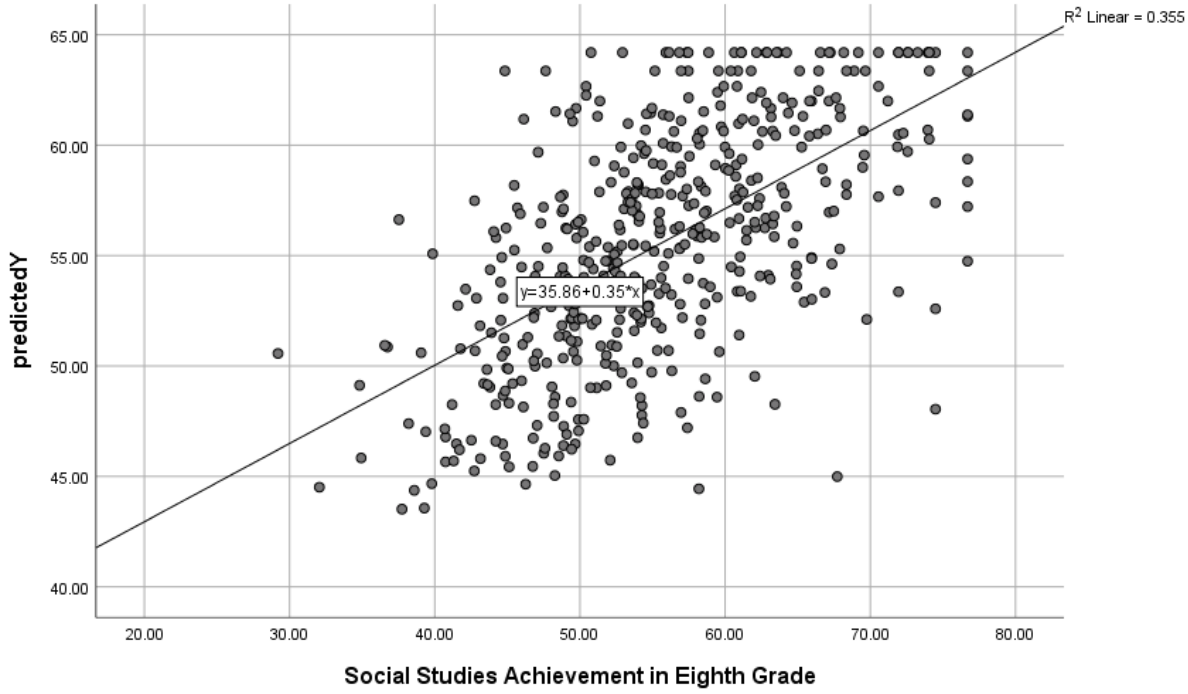
Below is a scatter plot of the relationship between x and y and the equation for the line of best fit.



The $R^2 = 0.355$ which is below 0.7 and is not close to 1. This means that despite the t test performed above we cannot insist on there being a good fit.

In order to compare the predicted and actual values of Y we may calculate a set of predicted Y values using the previous equation of $y = 22.289 + 0.594x$.

Let us now compare the predicted values of y and the actual values of y (i.e. "social studies achievement in eight grade" variable).



The R^2 is still very low meaning that the wellness of the fit between the predicted and actual values of Y is not good.

Despite this fact, below we can see the Pearson Correlation Coefficient of 0.595 meaning that the data is correlated but the predicted values do not fit the data as shown by the R^2.

Correlations

		predictedY	Social Studies Achievement in Eighth Grade
predictedY	Pearson Correlation	1	.595**
	Sig. (1-tailed)		.000
	N	500	494
Social Studies Achievement in Eighth Grade	Pearson Correlation	.595**	1
	Sig. (1-tailed)	.000	
	N	494	494

**. Correlation is significant at the 0.01 level (1-tailed).