Assignment 2

Jenni Simon
09-116-005

Exercise 1 Figure 1 shows a plot of the dependent variable "Wage" against the independent variable "Education" (in years). We can observe a roughly linear relationship between the variables (depicted with a red line). Note that outliers have been removed from the data in a pre-processing step, as suggested in Exercise 1.

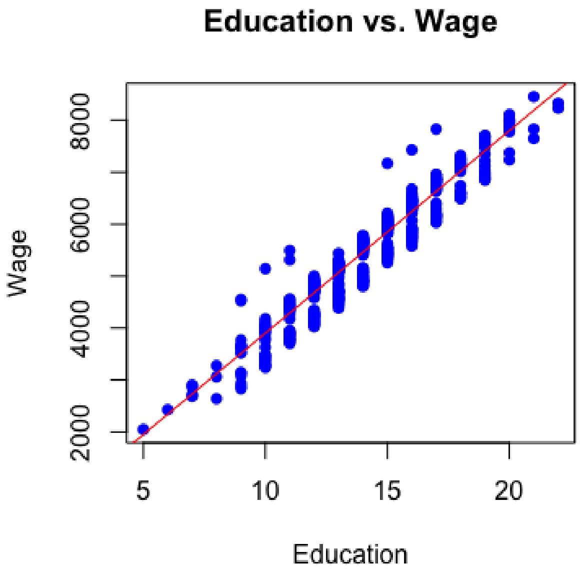

Fig. 1: Plot showing years of education vs. wage. Possible relationship is depicted with a red line.

Exercise 2 Figure 2 shows the same data with additional color indication of gender. We can clearly observe how men (blue) tend to have a higher wage at the same amount of education. Again we observe a linear dependency between wage and education for both classes "men" and "women".

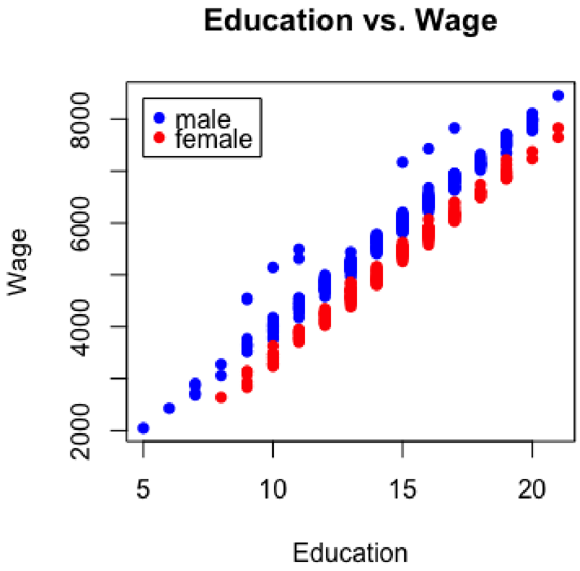

Fig. 2: Plot showing years of education vs. wage for men (blue) and women (red).

Exercise 3 Table 1 shows the minimum, mean, median, maximum, 1st- and 2nd quantile as well as the standard deviation of the dataset after preprocessing. The supplied dataset contained a negative time-value and a NaN which have both been removed before the computation. Fixing the wrong sign would arguably have also been a sensible solution, as the value seems to agree with the other observations.

Tab. 1: Summary of the dataset Mean20

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.
6.850	6.968	7.010	7.008	7.072	7.120	0.075

Exercise 4 We can test the hypothesis $H_0 : \mu = 7.05$ against $H_1 : \mu \neq 7.05$ using Student's two-sided t-test. As the dataset is rather small, we choose a

significance-level of $\alpha = 0.05$. Table 2 shows the results of this test. We observe a very low p-value (smaller than α) and therefore reject H_0 in favour of H_1.

Tab. 2: Result of two-sided t-test on pre-processed data.

t-value	95% confidence interval	p-value	Conclusion
-2.499	[6.973, 7.043]	0.0218	H_1

If we repeat the test on the original, not pre-processed data we obtain the results shown in table 3. We observe that the outliers have a large impact on the test results. Based on the relatively large p-value, we would accept H_0 in this case.

Tab. 3: Result of two-sided t-test on original data.

t-value	95% confidence interval	p-value	Conclusion
-1.063	[4.948, 7.733]	0.3006	H_0

Exercise 5 In this case we test the hypothesis $H_0 : \mu = 7.05$ against $H_1 : \mu > 7.05$ using the one-sided version of Student's t-test. The results of this test are shown in table 4. In this case we would clearly accept H_0. However, I would highly doubt Mary's claim with such a result (especially the extremely high p-value).

Tab. 4: Result of one-sided t-test on pre-processed data.

t-value	95% confidence interval	p-value	Conclusion
-1.063	[6.979, $+\infty$]	0.9891	H_0