Write an R script for the following.

Information on this data set from its source, “Introduction to the Practice of Statistics (5th edition)” by Moore and McCabe: “For one vehicle equipped in this way [with a computer], miles per gallon (MPG) and miles per hour (MPH) were recorded each time the gas tank was filled and the computer was then reset.”

We are interested in the question: What is the relationship between average fuel efficiency (MPG) and speed (MPH)?

To address this question, we define the random variable that is the speed MPH during a recording period, and the variable that is the fuel efficiencies MPG during a recording period.

One could make a pretty good case that our inferences extend to this particular car, and for automotive reasons (but not statistical ones), we would expect to be able to draw similar (but not exactly the same) inferences for other cars.

R1. Put the csv file fuel.csv (with that filename) into your working directory so that it can be read into R. Read that csv file into a data frame called fuelData in R.

R2. Make a scatterplot of our observations of MPG versus those of MPH (with suitable axis labels).

R3. Add a column to the fuelData data frame that gives the base 10 logarithm of our observations of MPH.

R4. Make a scatterplot of our observations of MPG versus those of log(MPH) (with suitable axis labels, omitting units on the log transformed variable).

R5. Fit a linear model with model formula $MPG \sim \log(MPH)$, and generate a model summary of the fitted model.

R6. Make a standardized residuals versus fitted plot for this fitted model.

R7. Make a normal quantile plot of the residuals of the fitted model.

Answer the following, but don’t include any R code in your write-up.

W1. Suppose we would like to know the relationship between the average time it takes to read a passage and the image quality of that passage as displayed on a computer screen. We have the following available data which its source (Devore and Peck, Statistics, 5th edition) describes as follows: “Image quality of monitors is an important characteristic affecting, among other things, extent of eye strain and work efficiency. The article ‘Image Quality Determines Differences in Reading Performance and Perceived Image Quality with CRT
and Hard Copy Displays’ (Human Factors [1991]: 459-469) reported on an experiment in which the image quality and average time for a group of subjects to read certain passages were determined.” Image quality was on a scale from 1 to 9, and its units are points, and time was measured in seconds.

We define a random variable $T$, which is the time reported for a subject to read a specific passage (the subject is actually a group of people, and the reported time is an average, but you can ignore such complications for our purposes here). We also define a variable $Q$, which is the image quality with which a passage is displayed on a computer monitor. This data set contains 8 simultaneous observations of these two variables. You should think about the scope of inference (namely, to what population do these results apply) and whether or not causative evidence may occur here.

For reference, here is a scatterplot of our observations of $T$ versus our observations of $Q$:

![Scatterplot of T versus Q](image)

We denote the true model equation for our inference by

$$\mu[T|Q] = \beta_0 + \beta_1 Q,$$

where $\beta_0$ and $\beta_1$ are unknown.

We check the model fitting assumptions, and they seem fine, so we fit a model with model formula $T \sim Q$ to our observations. Then we check the sampling variability assumptions, and they also seem fine, so we can proceed with statistical inferences.

In R, we obtain a summary of the fitted model with this output:

```R
Call:
  lm(formula = time ~ quality, data = qualityData)

Residuals:
  Min     1Q Median     3Q    Max
```
-0.41875 -0.15191 0.07709 0.15436 0.24790

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 9.6378   | 0.6419     | 15.015  | 5.5e-06  ***|
| quality        | -0.3485  | 0.1125     | -3.096  | 0.0212   *  |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2475 on 6 degrees of freedom
Multiple R-squared: 0.6151, Adjusted R-squared: 0.5509
F-statistic: 9.588 on 1 and 6 DF, p-value: 0.02121

Reading from this output, what is the fitted model equation?

W2. This summary contains information about the results of a hypothesis test that addresses whether there is a nonzero linear relationship between $\mu[T|Q]$ and $Q$. What are the null and alternative hypotheses of this hypothesis test? (Your answer should be in terms of the true model equation coefficient(s).)

W3. Report the results of the hypothesis test from the previous question. giving the value of the test statistic, the degrees of freedom, and the $p$-value.

W4. By comparing the $p$ value to the usual significance level of 0.05, interpret the results of the hypothesis test from the previous question (in statistical language, but specific to this problem, not just in general terms).

W5. Why would reporting the associated confidence interval here probably be more useful than just reporting the results of the above hypothesis test?

W6. This data set contains measurements of the lengths and widths of 88 butter clams, both in cm. The description of this data set from https://seattlecentral.edu/qelp/sets/001/001.html is as follows. “Our sample ($n = 88$) was collected from a 100 m stretch of intertidal gravelly beach at Alki Point in west Seattle during November 1998. We did not sample randomly; rather, we picked up every shell at first, then looked for really small shells to make the range in sizes greater. We did not sample whole live butter clams (but lots of people do!), just shells of clams that had died naturally or been killed by predators (snails, humans) or environmental conditions. We assume that the shells are representative of the organism as a whole, but this assumption could be false. We also assume that shells of dead clams are similar to shells of live clams; this assumption seems reasonable. So we assume that sampling the shells of dead clams is basically the same as sampling live clams; the shells are a substitute or proxy for the live clams. Our sample of 88 butter clam shells is a very small subset of the population of live butter clams at Alki Point. Over a 100 m stretch of beach, there must be 1,000-10,000 butter clams in the intertidal zone alone, and many more out below low tide.

Students in our environmental mathematics course at Seattle Central Community College measured the lengths and widths of each shell in our sample. Our shells were disarticulated (broken apart) so that students could measure one half of each shell pair. Students
laid the shells on metric-ruled graph paper to more easily measure length and width at right angles to one another.”

We are interested in the question: *How is average length of Puget Sound butter clams related to the width of the clam?*

We design a study that uses available data from this data set. We define $L$ (short for *length*) to be the random variable whose value is the length a randomly selected clam from the population represented. Also, we define $W$ (short for *width*) to be the random variable whose value is the width a randomly selected clam from the population represented.

An individual in this study is a Puget Sound butter clam from among the population represented.

The population to which our statistical inferences extend is only the clams in our sample. However, statistical inference still seems appropriate. I invite you to think about why (and feel free to ask me about it). That isn’t an easy question in this case.

In the population represented, do the results carry strong causative weight? (That is, might the results of this study, if positive, be used in an argument that a change in a clam’s width causes a change in the average length of clams with that width.) Why or why not?

**W7.** Name two things that indicate that a log transformation is in order for both variables here, even before exploring the data.

**W8.** What if you hadn’t noticed that log transformations were needed? Let’s see what types of other indications there would be. For reference, below is a scatterplot of our observed values of $L$ versus the simultaneous observed values of $W$.

![Scatterplot of observed values of L versus W](image)

Also, here is a standardized residuals versus fitted plot for the model with formula $L \sim W$ fitted to our observations of these variables:
And here is a normal quantile plot of the residuals for the model with formula \( L \sim W \) fitted to our observations of these variables:

In these last two plots (the original scatterplot is just for reference), what indications are there that a different model is needed?

W9. Since all indications are that you should log transform both variables, you do so (using base 10 logarithms). The model fitting assumptions seem fine, so you fit the model of \( \log(L) \sim \log(W) \) to your observations of the two variables. Now the plots show below to verify the sampling variability assumptions look much better. First, for reference, here is a scatterplot of the observed values of \( \log(L) \) versus the observed values of \( \log(W) \):
Also, here is a standardized residuals versus fitted plot for the model with formula $\log(L) \sim \log(W)$ fitted to our observations of these variables:

And here is a normal quantile plot of the residuals for the model with formula $\log(L) \sim \log(W)$ fitted to our observations of these variables:
You decide to look into the outlier in the standardized residuals versus fitted plot of the residuals, and to report what you see in the normal quantile plot, but the sampling variability assumptions now appear good enough to go forward with.

You get a summary of the log transformed fitted model in R, and it gives you the following output:

```
> summary(clamsModel)
```

```
Call:
  lm(formula = logLength ~ logWidth, data = clamsData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.042127 -0.014346  0.000286  0.011349  0.101738

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.132502   0.007907  16.76  <2e-16 ***
logWidth     0.961542   0.012390  77.61  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02283 on 86 degrees of freedom
Multiple R-squared:  0.9859, Adjusted R-squared:  0.9858
F-statistic: 6023 on 1 and 86 DF,  p-value: < 2.2e-16
```

Based on this summary, what is the model equation of the log transformed fitted model? (Since this equation involves logarithms, you do not need to include units.)
W10. Show how to back transform the equation, so that it will be in terms of $L$ and $W$ rather than $\log(L)$ and $\log(W)$.

W11. Based on the back transformed fitted model equation, how do we interpret the first model coefficient in this equation?

W12. Based on the back transformed equation, how do we interpret our the second model coefficient in this equation?