Do a qq plot of each of the 2 cent. vars. in the data from hw-lect1. By R. Describe/Interpret the results.

Note: If you find out that there is not much you can say about the qq plot, it may be that your data is not appropriate. This will be your first chance to correct the error, because later you will be doing more hw problems using your data.

So, see me, if you are not sure.

The results will vary across students, but if you see a "straight" qq plot, then that's evidence that your data came from a normal distr. with \( \mu \) equal to the y-intercept, and \( \sigma \) = slope of the qq plot.

If your qq plot looks curved, then try doing the qq plot of the log or \( \sqrt{\text{of the data. If that makes the qq plot "straight" then use the log or } \sqrt{\text{of the data for all future work.}} \)
Make a scatterplot of the 2 cont. vars collected in hw-list 2. Describe the relationship.

Note: Same comment as above. i.e. This is another chance to make sure your data are appropriate for the kinds of analysis we will do later. Let me know if you are not sure if your data is appropriate.

Again the results will vary. If you see a "linear" pattern, then it's good. If you see a pattern, but it's nonlinear, then try taking log or \( \sqrt{ } \) of one or both axes. If that gives you a "linear" pattern, then use the log or \( \sqrt{ } \) of the data in all future work. If you see no pattern at all, you should find a different data set, because you will get more out of it when we apply future methods to the data.
I gave you a formula that defines \( r \).
(The book gives you 2 others on p. 110.)

a) Start from the formula I gave you, and show that it is equal to

\[
r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]

\[\text{[I]}\]

Since \( S_x \) and \( S_y \) have no \( i \) index, they are constants.

Defn of \( S_x, S_y \).

\[r = \frac{1}{n-1} \sum_i \left( \frac{x_i - \bar{x}}{S_x} \right) \left( \frac{y_i - \bar{y}}{S_y} \right)\]

\[= \frac{1}{\sqrt{n-1} \sum (x_i - \bar{x})^2} \cdot \frac{1}{\sqrt{n-1} \sum (y_i - \bar{y})^2} \cdot \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})\]

\[= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}\]
b) Start from (1), and show that it is equal to \( r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \).

with \( S_{xx}, S_{yy}, S_{xy} \) defined on p.110

Expand the products in (1):

\[
\begin{align*}
\hat{r} &= \frac{\sum (x_i y_i) - \bar{x} \sum y_i - \bar{y} \sum x_i + \bar{x} \bar{y} \sum 1}{\sqrt{\sum x_i^2 - 2 \bar{x} \sum x_i + \bar{x}^2} \sqrt{\sum y_i^2 - 2 \bar{y} \sum y_i + \bar{y}^2}} \\
&= \frac{\sum x_i y_i - \bar{x} \sum y_i - \bar{y} \sum x_i + \bar{x} \bar{y} \sum 1}{\sqrt{\sum x_i^2 - 2 \bar{x} \sum x_i + (\bar{x})^2} \sqrt{\sum y_i^2 - 2 \bar{y} \sum y_i + (\bar{y})^2}} \\
&= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum x_i^2 - n(\bar{x})^2}} \sqrt{\sum y_i^2 - n(\bar{y})^2} \\
&= \frac{\sum x_i y_i - \frac{1}{n} (\sum x_i)(\sum y_i)}{\sqrt{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}} \sqrt{\sum y_i^2 - \frac{1}{n} (\sum y_i)^2} \\
&= S_{xy} / \sqrt{S_{xx} S_{yy}}.
\end{align*}
\]
Suppose in case of data on $x$ and $y$ fall exactly on the line $y = mx + b$. Compute the value of $r$.

Hint: In any of the formulas for $r$, eliminate all $y$ in favor of $x$.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{S_x} \right) \left( \frac{y_i - \bar{y}}{S_y} \right)$$

$$S_y^2 = m^2 S_x^2 \Rightarrow S_y = \pm m S_x$$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{S_x} \right) \left( \frac{m x_i + b - m \bar{x} - b}{m S_x} \right)$$

$$= \frac{m}{S_x^2} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{m}{S_y} \frac{S_x^2}{S_x^2} = \pm 1$$

Note that the magnitude of $m$, the slope, cancels out (if it's not 0, $\infty$). So, the magnitude of the slope is irrelevant to $r$.

**HW 12-1: Show that** $$\frac{\partial}{\partial \hat{\alpha}} \text{MSE} \Big|_{\hat{\alpha}, \hat{\beta}} = 0$$ **implies** $$\bar{y} = \hat{\alpha} - \hat{\beta} \bar{x} = 0$$

$$\text{MSE}(\hat{\alpha}, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

$$\frac{\partial}{\partial \alpha} \text{MSE}(\hat{\alpha}, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Chain rule:

$$2 (y_i - \hat{\alpha} - \hat{\beta} x_i) \cdot \frac{\partial}{\partial \alpha} (-\hat{\alpha})$$

$$= \frac{2}{n} \sum_{i=1}^{n} \left( y_i - \hat{\alpha} - \hat{\beta} x_i \right)$$

$$= \frac{2}{n} \left( \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\alpha} \frac{\sum_{i=1}^{n} x_i}{n} - \hat{\beta} \frac{\sum_{i=1}^{n} x_i}{n} \right)$$

$$= 2 \left( \bar{y} - \hat{\alpha} - \hat{\beta} \bar{x} \right)$$

$$\frac{\partial}{\partial \alpha} \text{MSE} \Big|_{\hat{\alpha}, \hat{\beta}} = 0 \Rightarrow \bar{y} - \hat{\alpha} - \hat{\beta} \bar{x} = 0$$
MOE = c
(29.8,33.2,33.7,35.3,35.5,36.1,36.2,36.3,37.5,37.7,38.7,38.8,39.6,41.0,42.8,42.8,43.5,45.6,46.0,46.9,48.0,49.3,51.7,62.6,69.8,79.5,80.0)

Strength = c
(5.9,7.2,7.3,6.3,8.1,6.8,7.0,7.6,6.8,6.5,7.0,6.3,7.9,9.0,8.2,8.7,7.8,9.7,7.4,7.7,9.7,7.8,7.7,11.6,11.3,11.8,10.7)

# a)
plot(MOE, Strength)  # There is a decent linear association.

# b)
boxplot(MOE, Strength)  # No comment.

# c)
qqnorm(MOE)  # Neither distribution is truly Normal, and so,
qqnorm(Strength)  # technically we should "fix it" before proceeding.

# d)

zx = (MOE - mean(MOE))/sd(MOE)
zy = (Strength - mean(Strength))/sd(Strength)
sum(zx*zy)/(length(MOE)-1)  # 0.859, i.e., decent correlation

# e)
cor(MOE,Strength)  # same as "by hand" above.

# f)

numerator = mean(MOE*Strength) - mean(MOE)*mean(Strength)
denominator = mean(MOE^2) - (mean(MOE))^2
beta = numerator/denominator
beta  # 0.1075
alpha = mean(Strength) - beta*mean(MOE)
alpha  # 3.295

# g)
# For every unit change in MOE, on the average Strength changes by 0.1075.
# When MOE = 0, Strength is expected to be about 3.295.
# But note that MOE = 0 is really an extrapolation, because the range of
# MOE is 29 to 80.
range(MOE)

# h)
alpha + beta*39.0  # 7.48

# i)
yhat = alpha + beta*MOE
sum((Strength - yhat)^2)  # 18.735, i.e., the smallest possible SSE.
Suppose data on $x$ and $y$ fall on a straight line $y_i = b + mx_i$.

If we perform a linear fit $y = \alpha + \beta x$ to this data, what is the value of the OLS estimate of $\beta$?

\[
\hat{\beta} = \frac{\bar{x} \overline{y} - \bar{y} \bar{x}}{\bar{x}^2 - \bar{x}} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}}{\bar{x}^2 - \bar{x}} = \frac{\bar{x} \sum (y_i - \bar{y})}{\bar{x}^2 - \bar{x}} = \frac{\bar{x} \sum (b + mx_i) - \bar{x} \sum (b + mx_i)}{\bar{x}^2 - \bar{x}} = \frac{\bar{x} \sum m x_i - \bar{x} \sum b - \bar{x} \sum x_i}{\bar{x}^2 - \bar{x}} = \frac{b \bar{x} + m \bar{x}^2 - b \bar{x} - m \bar{x}^2}{\bar{x}^2 - \bar{x}} = m
\]

Note that this result looks obvious but it's not. In fact, when the data do not fall on a straight line, the $\hat{\beta}$ of the fit is not what you may think is the best fit to the data. You will see this in Lab.

Prove that the OLS fit goes through the point $(\bar{x}, \bar{y})$.

At $x = \bar{x}$, we have
\[
\hat{y} = \bar{y} + \hat{\beta} \bar{x} = \bar{y} - \hat{\beta} \bar{x} + \hat{\beta} \bar{x} = \bar{y}
\]

I.e. The OLS fit always goes thru the point $(\bar{x}, \bar{y})$.

Show that $\hat{\beta}$ as defined by $\hat{\beta} = \frac{\bar{x} \overline{y} - \bar{y} \bar{x}}{\bar{x}^2 - \bar{x}}$ or $\frac{\sum x y}{\sum x^2}$

can be written as $\hat{\beta} = r \frac{S_y}{S_x}$ where $S_x$ = sample std. dev. of $x$;

\[
\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{S_{xx}} \frac{S_{yy}}{S_{yy}} = \frac{S_{xy}}{S_{xx} S_{yy}}S_{yy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}} \sqrt{S_{xx} S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \sqrt{\frac{n-1}{n-1} \frac{S_y^2}{S_x}} = \sqrt{\frac{S_y^2}{S_x}} \sqrt{r}
\]