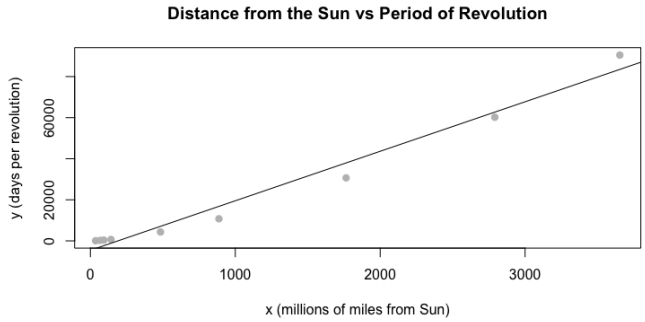
MATH 140 **Name** _____ **Date:**

R Activity: Least Squares Regression

The Scene: The following table lists the average distance from the sun (in millions of miles), and period of revolution (Period) around the sun (in Earth days) of the nine planets in the solar system (including Pluto!). Below the table is a scatterplot of the data, produced in RStudio, which includes the least-squares regression line.

	planet	distance (x)	period (y)	Predicted revolution ($\hat{y}$)	residual ($y - \hat{y}$)
1	Mercury	36	88		
2	Venus	67	225		
3	Earth	93	365		
4	Mars	142	687		
5	Jupiter	484	4332		
6	Saturn	887	10760		
7	Uranus	1765	30684		
8	Neptune	2791	60188		
9	Pluto	3654	90467		



---

1. First, create a data frame in RStudio for the first three columns in the table above. Hint: You can copy and paste code from the **regression tutorial** found on our course resource page (activities tab). This codes create a data frame called "planets." Once you see "planets" in your environment tab, write "Got it!" below and move on to the next problem.

2. Recreate in your RStudio session the scatter plot that you see above. Once you've got a nice scatter plot (figure out how to label axes, and add color if you like...) write "Got it!" below and move on to the next problem.

3. In a sentence or two describe the relationship between a planet's distance from the Sun and its period of revolution.

4. Using RStudio determine the correlation coefficient r for these two variables, and also determine r^2. Does the (close to 1) value of r^2 seem to indicate a strong linear relationship between distance and period of revolution? Explain in a sentence or two.

5. Use RStudio to determine the least-squares regression model for these data. Use this code (you can copy and paste from the tutorial page) to generate lots of useful information about the least-squares line

   ```
   fit = lm(data=planets, period~distance)
   ```

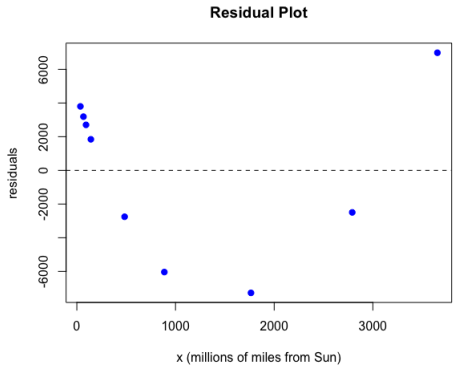
   Then run this code to get the intercept and slope of the least-squares regression line:

   ```
   fit$coefficients
   ```

   Write down the equation of the least-squares regression line here.

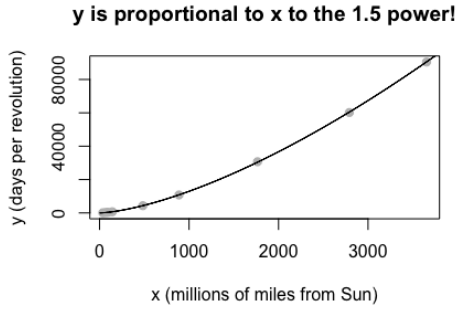
6. Use your `fit` model in RStudio to obtain the predicted periods ($\hat{y}$) for each of the 9 planets using the best fit line, as well as the residual values for the 9 planets. See the tutorial for the appropriate commands. Record your results in the last 2 columns of the table at the start of this worksheet.

7. Below is the residual plot. Does this plot reveal any pattern? What does this tell us about the goodness of a linear model for the relationship between period and distance? Explain.



When a straight line is a reasonable model, the residual plot should reveal a seemingly random scattering of points. When the residual plot reveals a pattern of some kind, as is the case here, a non-linear model would fit the data better.

8. It turns out that a planet's period (y) is not a *linear* function of its distance from the Sun (x), but rather y is proportional to $x^{1.5}$! This power, 1.5, reveals itself when one finds the least-squares regression line of the log of the data. The plot below shows the original data with the curve $\hat{y} = 0.41x^{1.5}$.



For Mars and Neptune, predict y from x using the equation $\hat{y} = 0.41x^{1.5}$.

(a) Period of revolution for Mars, as predicted by the polynomial:

(b) Period of revolution for Neptune, as predicted by the polynomial: