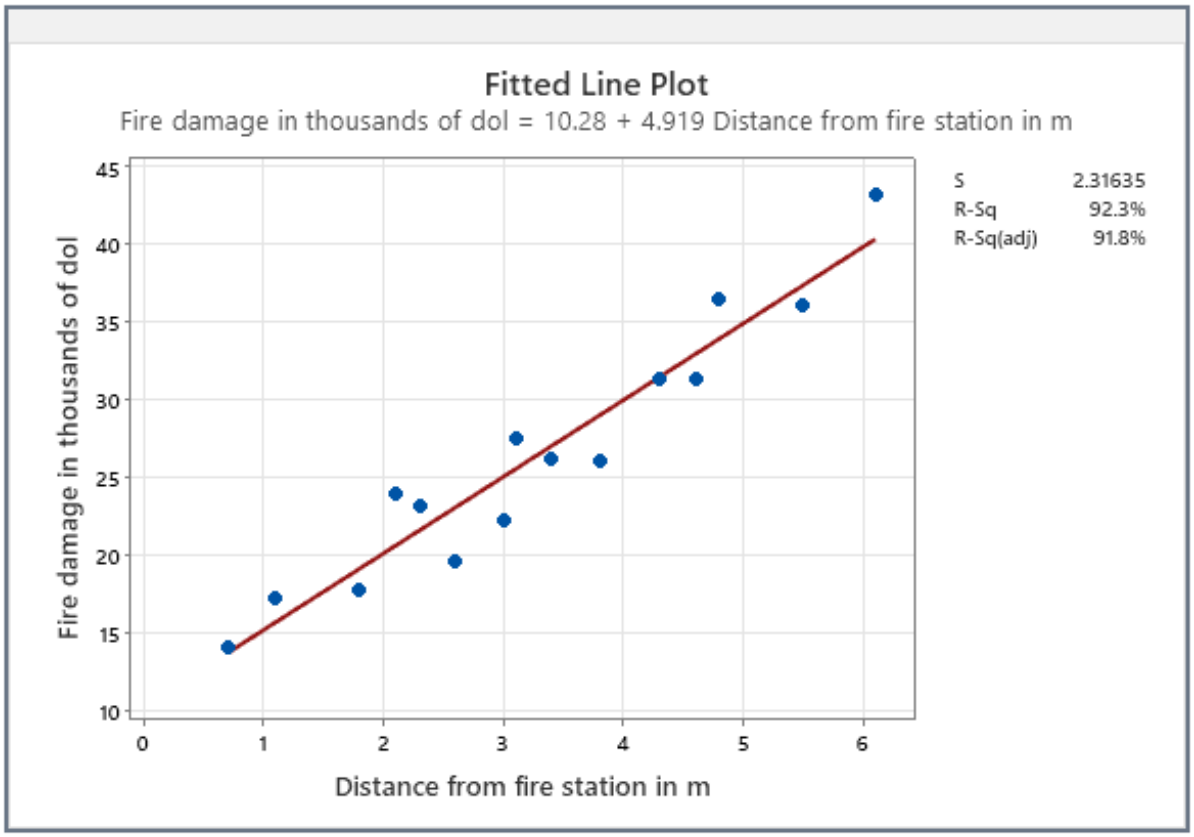
#1)

a)



b)

$$Fire\ Damage\ in\ Thousands\ of\ Dollars = 10.28 + 4.919(Distance\ from\ Fire\ Station)$$

c)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	841.766	841.766	156.89	0.000
Error	13	69.751	5.365		
Total	14	911.517			

d)

Because the relationship is a positive relationship (seen by the graph above), this is our hypothesis test:

$$H_0: \beta_1 \leq 0, H_a: \beta_1 > 0$$

A critical F-Statistic of 4.668 was found using Statkey. Because F=156.89 $>$ 4.668 (the F-statistic in our ANOVA table seen above), that means that $P < 0.05$ (also seen in the ANOVA table above). Therefore,

we can reject our null hypothesis that $\beta_1 \leq 0$ and conclude that there exists a positive linear relationship between fore damage in thousands of dollars and distance from fire station.

e)

It really isn't relevant. The only way this would be relevant is if the fire station itself was on fire, which is the only time you would be 0 miles from the fire station. This doesn't seem like a very relevant scenario.

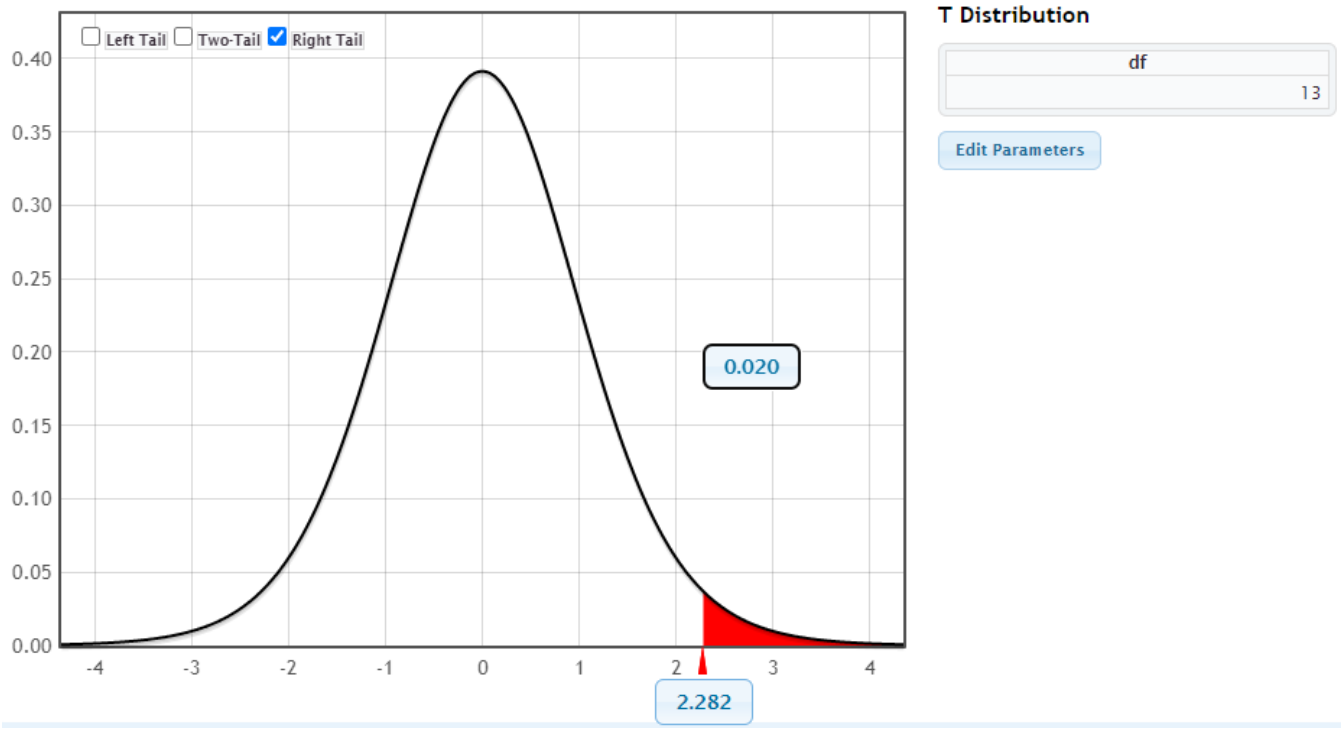
f)

The co-efficient of determination (R-Sq) is 92.3% (seen in the scatter plot on *page 1*). This means that the linear regression model, using distance from the fire station as an explanatory variable, explains 92.3% of the variability in our response variable (fire damage in thousands of dollars).

g)

$$x\ COEF\ \pm\ \mathrm{t}\ *\ (\mathrm{x\ SE\ COEF}) = 4.919\ \pm 2.282(0.3963) = (4.0146434, 5.8233566)$$

See diagrams below for how I found the SE COEF and t-statistic. My slope is inside the suggested value. This suggests that, based on our sample data and statistical analysis, there is not strong evidence to conclude that the population slope β_1 is significantly different from the estimated value. It means our linear relationship, as represented by the slope, is consistent with the data within the confidence interval.



h)

Using this equation:

$$\hat{\mu}_{y|x_0} \pm t_{\frac{a}{2},n-2}\sqrt{MS_{Res}(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}}$$

$\hat{\mu}_{y|x_0} = 10.28 + 4.919(3.5) = 27.4965$

$t_{\frac{a}{2},n-2} = 2.282$ (using StatKey, shown below)

$MS_{Res} = 5.365$ (using ANOVA table)

$n = 15$

$x_0 = 3.5$

$\bar{x} = 3.28$ (using MiniTab descriptive statistics, see output below)

$S_{XX} = 34.78$ (used a calculator to do this one)

So now we have…

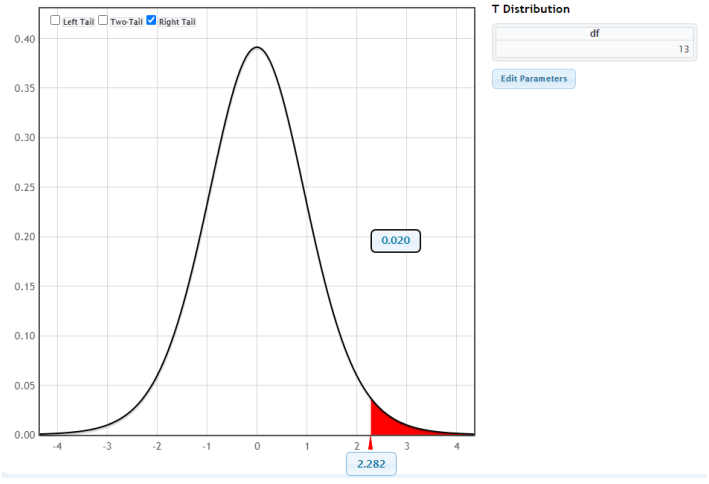
$$98\%\ CI = 27.4965 \pm 2.282\sqrt{5.365(\frac{1}{15}+\frac{(3.5-3.28)^2}{34.78}} = (26.117574, 28.875426)$$

Analysis: This means that we can be 98% confident that the true population average value for fire damage(in thousands of dollars) lies between approximately $26.1 and $26.8 when the distance from the fire station is 3.5 miles.

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Distance from fire station in m	15	0	3.280	0.407	1.576	0.700	2.100	3.100	4.600

Variable	Maximum
Distance from fire station in m	6.100



i)

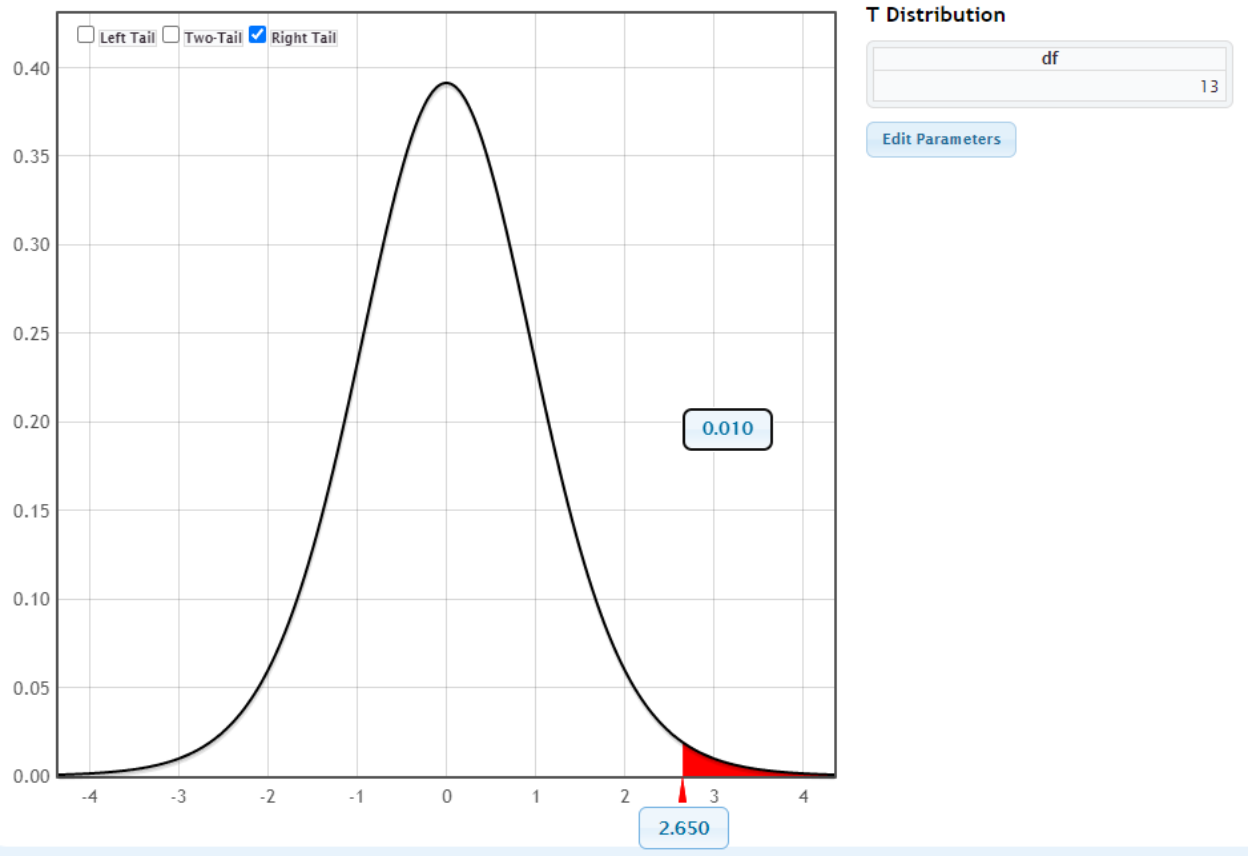
$$99\%\ PI = \hat{y}_0 \pm t_{\frac{a}{2},n-2}\sqrt{MS_{Res}(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}})}$$

All other stats shown above, now with $\hat{y}_0 = 3.5$ and using $t_{\frac{a}{2},n-2} = 2.650$ for the adjusted interval window (StatKey T-Distribution shown below).

Now we have

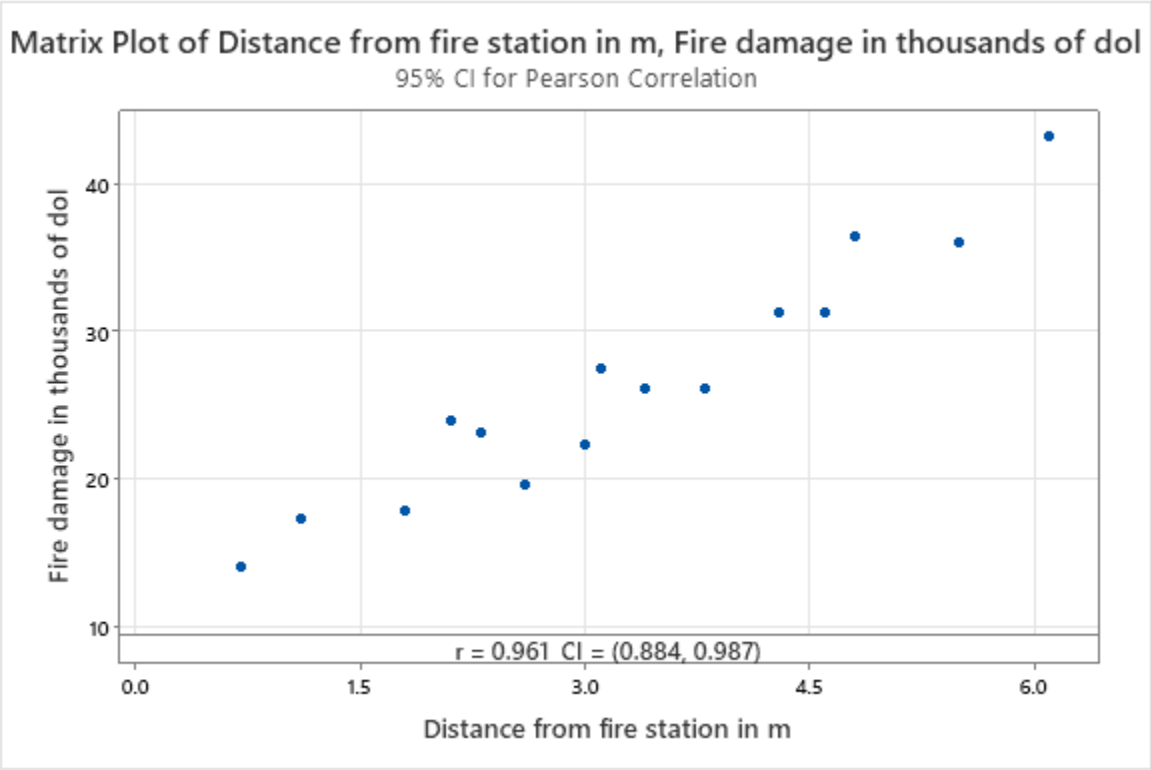
$$99\%\ PI = 3.5 \pm 2.65\sqrt{5.365(1+\frac{1}{15}+\frac{(3.5-3.28)^2}{34.78})} = (-2.843489288, 9.843489288)$$

Analysis: If we were to repeat the data collection process many times, we would expect that, if the distance from the fire station was 3.5 miles, then we could expect the damage, in thousands of dollars, would be between $0 and $9.84, because negative damages don't make sense in the context of this problem.



j)

Correlation coefficient (r) = 0.961. This quantifies the strength and direction of the linear relationship between fire damage and distance from fire station. In this instance, we can see that there is a pretty strong positive linear relationship between the two variables.



j)

$H_0: \rho \leq 0, H_a: \rho > 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.961\sqrt{15-2}}{\sqrt{1-0.961^2}} = 12.52921793$$

$\alpha = .05$

Critical Value=1.771

Because our test statistic is greater than our critical value, we can reject our null hypothesis that our population coefficient (ρ) is less than or equal to zero.

k)

$$CI = \hat{\beta}_1 \pm t_{\frac{a}{2},n-2} * SE(\hat{\beta}_1) = 4.919 \pm 2.060 * 0.393 = (4.10942, 5.72858)$$

This confidence provides a range within which we can be 97% confident that the true population coefficient lies, based on our sample data. Since the interval does not include zero, it suggests that the coefficient is statistically significant.

#2)

a)

I would propose a simple linear regression model.

b)

Yes, looking at the scatter plot it does appear that there is a relatively strong linear relationship between market value and sale price.

c)

$Sale\_Price = 1.4 + 1.4083(Market\_Val)$

d)

The y-intercept is 1.4 (thousand) dollars, or $1,400. In the context of this problem, it means that a home with a market value of $0 has a sale price of $1,400. This is (obviously) not meaningful because there aren't homes that have market values of $0 or sale prices of $1,400. (If there are, I don't want to live there!)

e)

The x coefficient, or the slope, is 1.4083. This means that for every $1,000 change in the market value of a home, there is a $1,408.30 change in the sale price of the home (I'm assuming these numbers are given in thousands). Looking at the scatterplot, I would say that this slope is useful somewhere between approximately $150k-$800k. There are only two observations greater than $800k and they are far beyond all the other datapoints. Those two data points also have larger residuals. So it would likely not be as accurate to use this coefficient to predict outside of the range of (150k,800k).

f)

$Sale\ Price = 1.4k + 1.4083k(300k) = 423.89k$

g)

Yes, there is sufficient evidence to at $\alpha = .01$ to indicate that $\beta_1$is positive. This is because the P-value for the Pearson correlation is $0.000 < 0.01$.

h)

$x\ COEF\ \pm\ \mathrm{t}\ *\ (\mathrm{x\ SE\ COEF}) = 1.4083 \pm 1.666 * (0.0369) = (1.3468246,1.4697754)$

(t found from a right tail test on statkey) We are 97% confident that the true population slope β_1 falls within the range of $(1.337821,1.478779)$. Since the slope does not include 0, we can conclude that there is evidence of a statistically significant linear relationship between sale price and market value.

i)

Increase our confidence level. For example, a confidence level of 99% yields a narrower interval:

$1.4083 \pm 2.378 * (0.0369) = (1.3205518,1.4960482)$

j)

Correlation Coefficient=0.975. This Pearson correlation coefficient is used to indicate how closely the data points cluster around a straight line. A value of 0.975 indicates a very strong positive linear relationship between sale price and market value.

k)

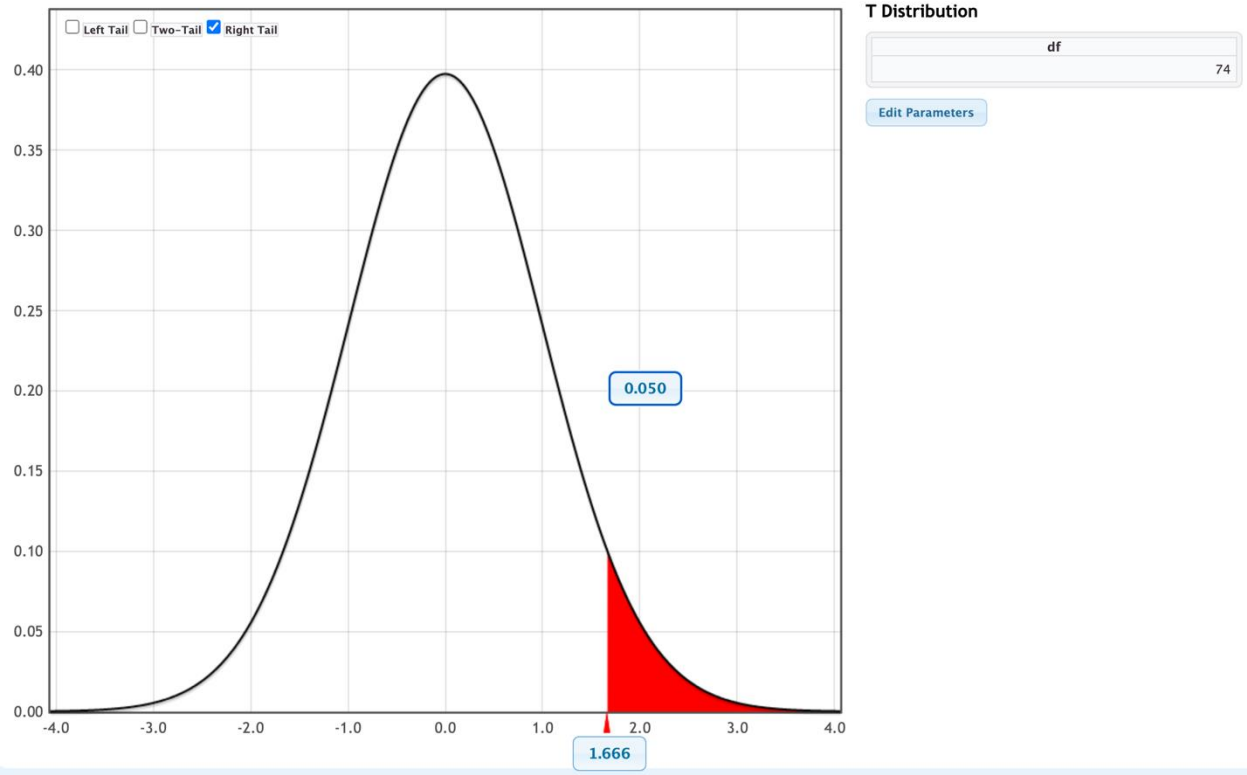
$$\text{Coefficient of determination} = \text{R} - \text{Sq} = 95.16\% = .9516$$

This means that approximately 95.16% of the variability in sale price can be explained by our linear regression model, which includes market value as the predictor variable.

l)

$H_0: r \leq 0, H_a: r > 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$



$$= \frac{.975\sqrt{76-2}}{\sqrt{1-.975^2}} = 37.7456511$$

$Critical\ Value = 1.666$ (StatKey result shown below)

Because $t > critical\ value$, we can reject our null hypothesis that $r \leq 0$ and determine that there is a direct correlation between sale price and market value.

m)

$$\text{CI} = \tanh\left(\operatorname{arctanh} r \pm \frac{Z_{\frac{a}{2}}}{\sqrt{n-3}}\right) = \tanh\left(2.184723926234 \pm \frac{1.96}{\sqrt{73}}\right)$$

$$= (0.9607313587, 0.984125967)$$

We are 95% confident that the true correlation coefficient (r) between appraised value and sale price for properties in the neighborhood falls within the range of 0.961 to 0.984.

#3)

a)

My field of study is data science, so I just used a couple examples of things that are positively and negatively correlated.

<u>Positively Correlated:</u>

Temperature and Ice Cream Sales: On warmer days, the sales of ice cream tend to increase. This is a classic example of a positive correlation between temperature and sales.

<u>Negatively Correlated:</u>

Customer Wait Time and Customer Satisfaction: Longer customer wait times are often associated with lower customer satisfaction. Reducing wait times can lead to increased satisfaction.

b)

i) The slope of the least squares line will be positive. It indicates a strong positive linear relationship between the two variables. As one variable increases, the other tends to increase.

ii) The slope of the least squares line will be negative. It signifies a strong negative linear relationship between the two variables. As one variable increases, the other tends to decrease.

iii) The slope of the least squares line will be zero. It indicates no linear relationship between the two variables. Changes in one variable do not predict or explain changes in the other.

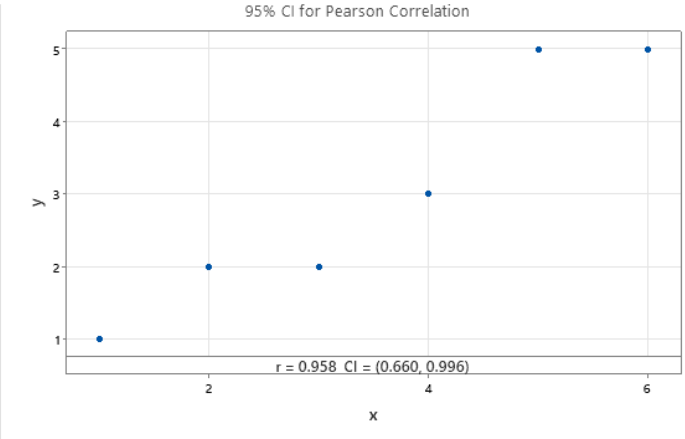
iv) The coefficient of determination (r^2)represents the proportion of variance in the dependent variable explained by the independent variable(s). The slope of the least squares line will be either positive or negative, depending on the sign of r (i.e., whether it's positive or negative). If r = 0.8, it implies that 64% of the variability in the dependent variable can be explained by the independent variable(s). The stronger the linear relationship, the more the slope deviates from zero.

c)

i) Correlation coefficient $r = 0.958$ and coefficient of determination $R^2 = .9184$, as seen by the Minitab output below.

Model Summary

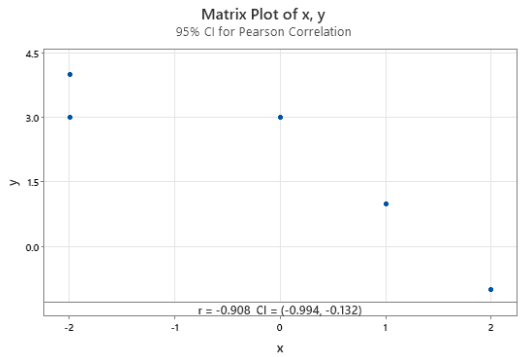
S	R-sq	R-sq(adj)	R-sq(pred)
0.534522	91.84%	89.80%	84.77%



ii) ii) Correlation coefficient $r = -0.908$ and coefficient of determination $R^2 = .8252$, as seen by the Minitab output below.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.965553	82.52%	76.69%	49.59%



d) I would expect the correlation to be positive. As population increases, it stands to reason that the crime rate would also increase.

e) It's difficult for me to answer, because I feel that there are people that are much smarter than I am with lower GPA's. But I work harder, so I have a higher GPA. So I'm not entirely sure if I would say that intelligence is correlated. But let's say, for the sake of this question, that it is correlated. I would expect it to be positive, where a higher GPA would correlate to a higher IQ.

#4)

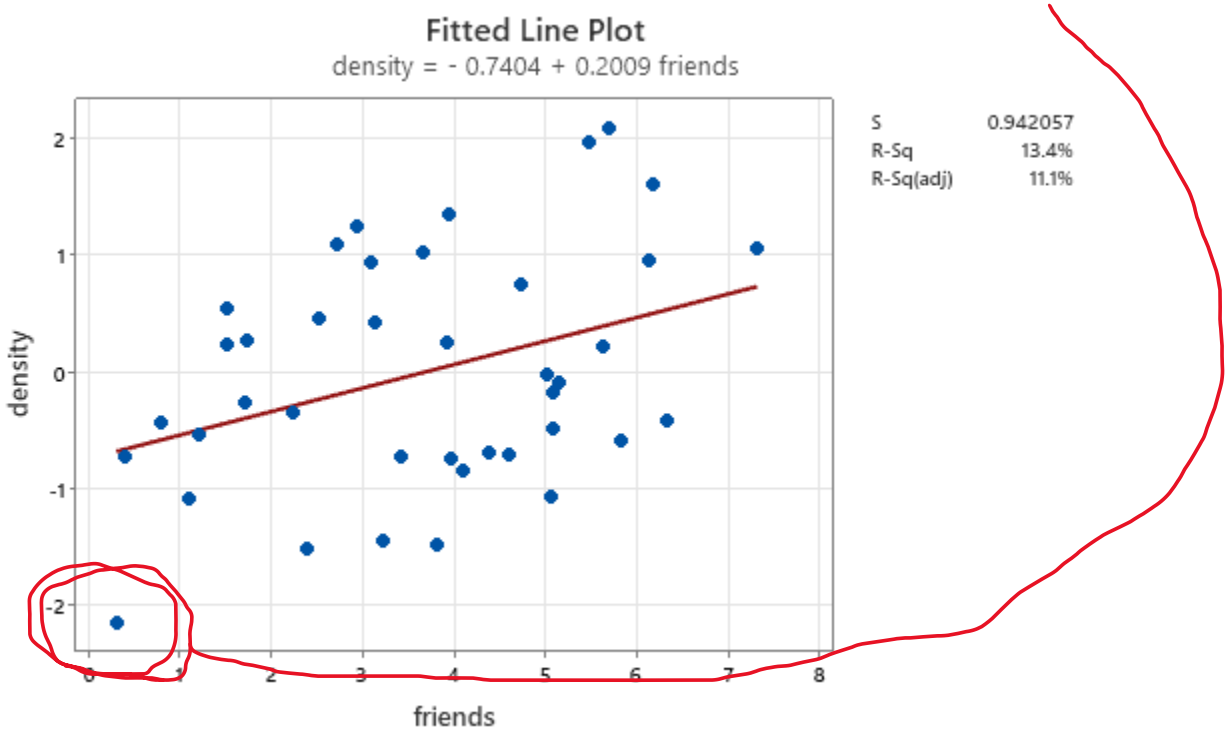
a) My null hypothesis would be that there is no association between the number of Facebook friends that a person has and their grey matter density. And my alternative hypothesis would be that there is an association between the number of Facebook friends that a person has and their grey matter density.

b) We will select 40 student volunteers who are willing to participate in the study. We will use MRI scans to measure brain density in the left middle temporal gyrus for each participant. This will be our dependent variable. The data will be collected on the number of Facebook friends (measured in units of 100 friends) for each participant, which will be our independent variable. Once we have collected our data, we will conduct a statistical analysis to examine the association between the number of Facebook friends and brain density in the left middle temporal gyrus using regression analysis.

c) Explanatory variable: Number of Facebook friends (measured in units of 100 friends).
Response Variable: Measure brain density in the left middle temporal gyrus for each participant.

d) Observational study because we are not intervening or manipulating any variables. We are observing and colling the data as it naturally occurs.

e) From the below graph, we can see that the direction of the graph is positive. The observations are very spread out, and there is likely not a very strong association between the two variables. The observation I've circled in red where friends = approximated .2 is the only unusual one.



f) $density = -0.7404 + 0.2009(friends)$
The slope is 0.2009, which implies that for every 100 friends, there is a .2009 increase in grey matter density.

g) If we do a hypothesis test on the slope to determine whether the slope we found is statistically significant from 0, we find that it is because the p-value of the slope (friends) is 0.020, which is less than 0.05. Since the slope is significantly different from 0, we can conclude that there is an association in the population.

Coefficients

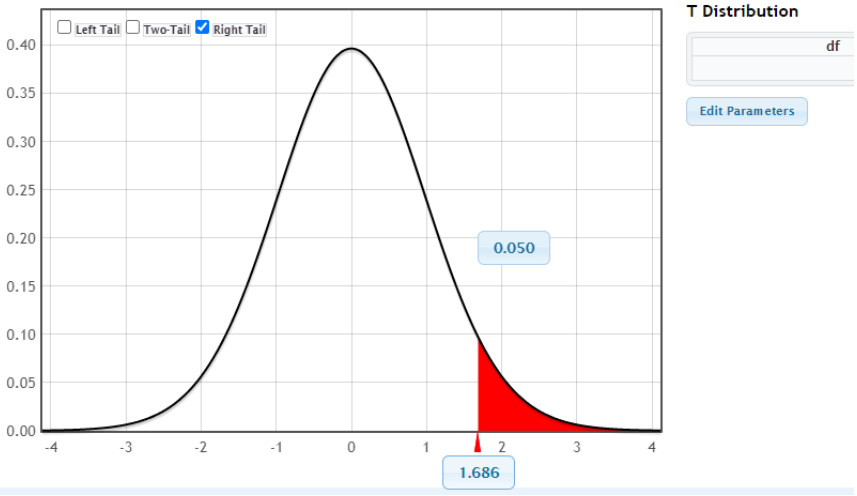
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.740	0.339	-2.18	0.035	
friends	0.2009	0.0830	2.42	0.020	1.00

h) I would use $R^2 = 13.36\%$ to calculate the effect size. Additionally, I would use the hypothesis test described above to suggest that the relationship is statistically significant.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.942057	13.36%	11.08%	4.65%

i) $t = \frac{coefficient\ of\ the\ slope}{SE\ of\ the\ slope} = \frac{0.2009}{0.0830} = 2.420481928$ (also shown in table above)
$P - value = 0.020$ (shown on table above). On a single tail test, a critical value of 1.686 is found (shown below). Since our T-Value is > 1.686, that determines a p-value of < .05, which determines that the relationship is statistically significant.



j) It measures the strength of evidence against the null hypothesis. In the context of this study, the null hypothesis is that there is no association between the number of Facebook friends and brain density in the population. The p-value quantifies the probability of obtaining the observed results if the null hypothesis were true. So, it tells us how likely it is to see the observed association between fb friends and brain density if there were no real relationship in the entire population.

k) It would be challenging to conclude that there is strong evidence of a positive association between # of FB friends and brain density in the population. This is because the R-Sq value of 13.4% suggests that the relationship, while statistically significant, explains a pretty small proportion of the variance in brain density. This small R-Sq value may not have substantial practical implications. Additionally, since this is an observational study and not a randomized experiment, we cannot imply causation.

l) $x\ COEF\ \pm\ \mathrm{t}\ *\ (\mathrm{x\ SE\ COEF}) = 0.209 \pm 1.686(0.830) = (-1.19038, 1.60838)$
We are 97% confident that the true population slope β_1 falls within the range of $(-1.19038, 1.60838)$. Because the interval does include 0, we can conclude that there is not evidence of a statistically significant linear relationship between fb friends and grey matter. This further strengthens the conclusion I came to in part k

m) I would want to randomly sample people at a shopping mall. This would include a more diverse range of different types of people in different age groups, socioeconomic backgrounds, professions, and lifestyles than sampling from university students.

n) As mentioned before, we cannot make this conclusion because it is not a randomized experiment. So we cannot imply causation.

o) No, because we used fb friends in our study to explain brain density. In order to answer that question, we would have to conduct an entirely new study where brain density was the explanatory variable and fb friends was the response variable.

p) The only thing I can think that I would do differently would be the group of people that we are pulling from. Most college students are very social and likely have a lot of Facebook friends. Also, most of them are the same age. You don't have much variance in your population in terms of demographic information. So I think it would be better to pull from somewhere like a shopping mall. It would obviously be better to do a randomized experiment, but that wouldn't really be feasible in this because we cant manipulate the number of fb friends a person has.

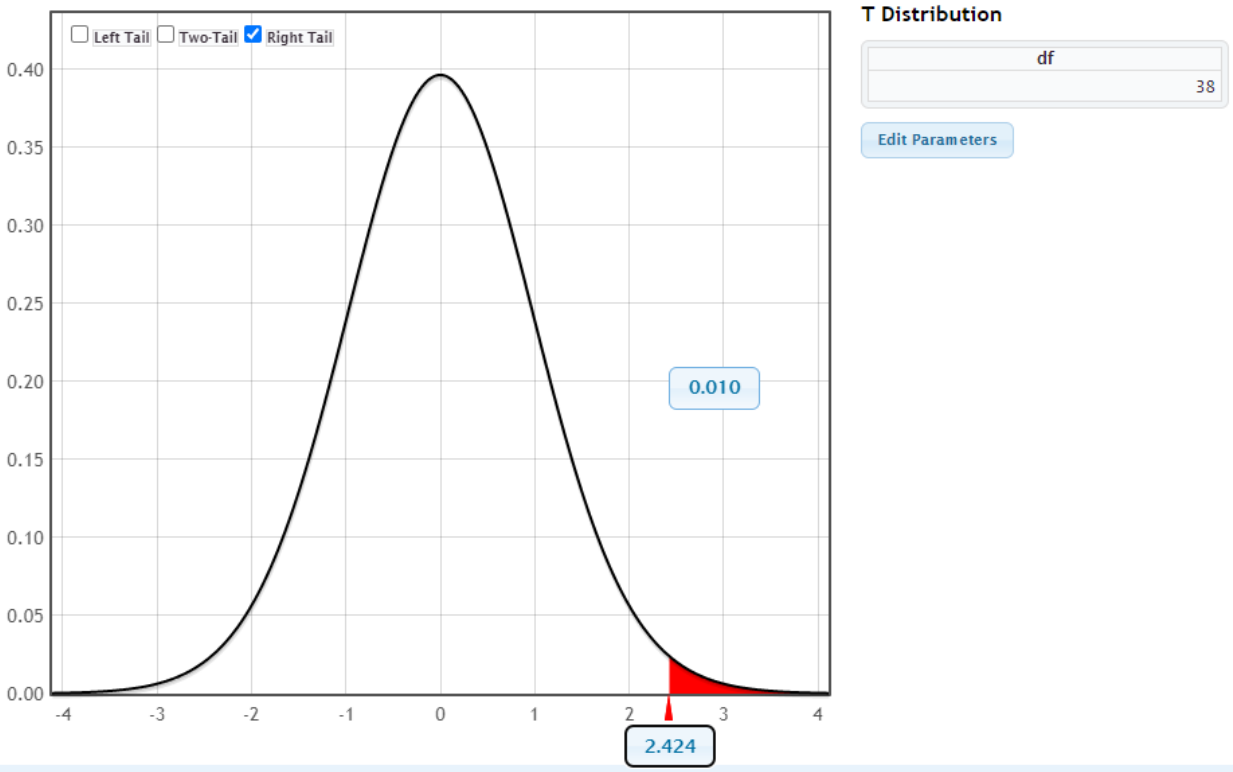
q) .366

Correlations

	friends
density	0.366

r) $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.366\sqrt{38}}{\sqrt{.866044}} = 2.424392351$
The test-statistic measures the strength and significance of the correlation between fb friends and grey matter while accounting for the sample size and provides evidence for the presence of a meaningful relationship.

s) If I plug this t-value into statkey, I can approximate a p-value of 0.010.



t) Because the scatterplot shows a relatively non-linear pattern, it may indicate a violation of the linearity assumption.

u) $\text{CI} = \tanh\left(\text{arctanh}\, r \pm \frac{Z_{\frac{a}{2}}}{\sqrt{n-3}}\right) = \tanh\left(0.383796542847 \pm \frac{1.96}{\sqrt{37}}\right)$

$= (-0.2764575306, 0.3457438908)$

We are 95% confident that the true correlation coefficient r between FB friends and grey matter density falls within the range of -0.2765 and 0.3457.

I, Ben DuPey, Can honestly say that I didn't get help from anybody for this exam, neither I communicate with anybody in class.

Ben DuPey — Signature

10/05/23 — Date

Good Luck!

Dr Sergio Loch