MG212-01 - DATA ANALYSIS - Fall 2020

HOMEWORK # 3 – Hospital Discharge Data – Due Friday October 9th (11:59 pm)

YOUR NAME: Daniel Gillinger

1. Open the Minitab file called **"HW3 - 2014 Hospital Discharges (Data File)"** that accompanies this document on Blackboard / ASSIGNMENTS, which should automatically open in Minitab.

2. Look at the data in the file:
    a. How many variables does it have? ____9_____

    b. How many total records does it include? ______79______

    c. In general, what does this dataset tell us? What is it reporting to you?

This data is telling us how many patients were discharged for a set of different diseases, in 2014. It gives the average length, average cost, and percent male for a discharge for each disease.

PLEASE ANSWER THE FOLLOWING:

3. Right click on the column header "Discharges", and select "Sort Columns" → "Entire Worksheet". [This is similar to the column sorting feature in Excel.]

    a. Which medical condition "CCS Category Description" had the MOST discharges, and how many did it have in 2014?

    __Congestive heart failure, nonhypertension__________Discharges: 901,425___

    b. Which medical condition "CCS Category Description" had the FEWEST discharges, and how many did it have in 2014?

    ___Female infertility________________________________Discharges: 60__

4. Right click on the column header "Avg Length of Stay, days", and select "Sort Columns" → "Entire Worksheet". [This is similar to the sorting feature in Excel.]

    a. Which medical condition "CCS Category Description" had the HIGHEST average length of stay in days, and what was this average length of stay in 2014?

    ____Heart valve disorders__________________________Avg. Length of Stay: 7.92__

    b. Which medical condition "CCS Category Description" had the LOWEST average length of stay in days, and what was this average length of stay in 2014?

Prolapse of female genital organs Avg. Length of Stay: 1.73

5. Right click on the column header “Avg Charges $”, and select “Sort Columns” → “Entire Worksheet”. [This is similar to the sorting feature in Excel.]

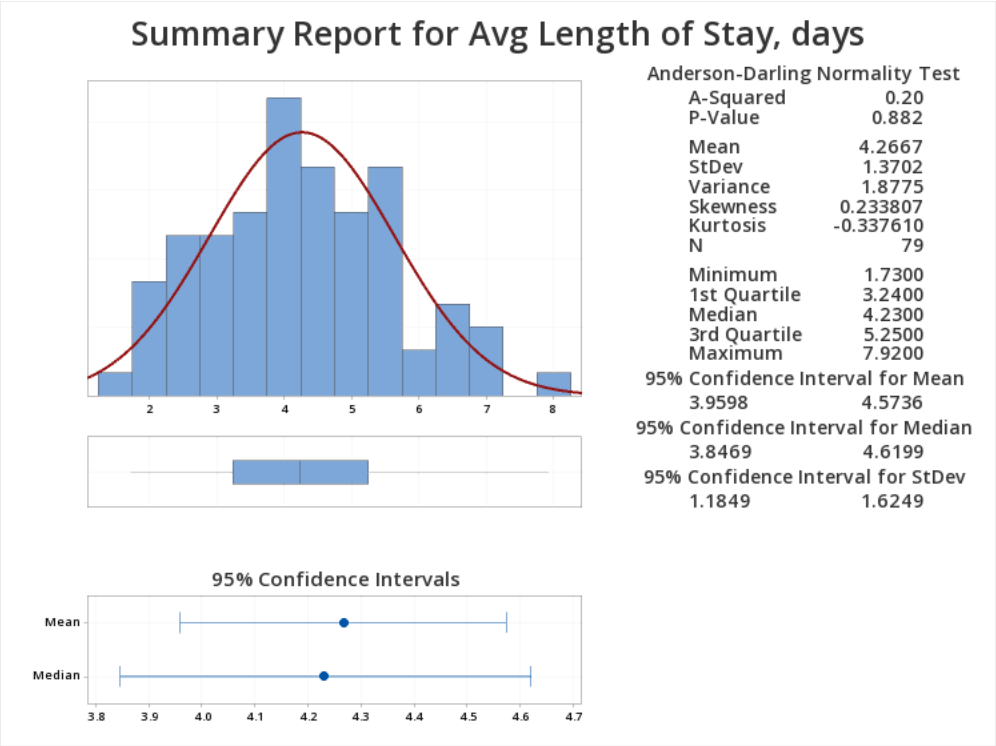
   a. Which medical condition “CCS Category Description” had the HIGHEST average charge, and what was this average charge in 2014?

      Heart valve disorders Avg. Charges: $169,965

   b. Which medical condition “CCS Category Description” had the LOWEST average charge, and what was this average charge in 2014?

      Acute and chronic tonsillitis Avg. Charges: $19,197

6. Generate the Graphical Summary for the “Avg Length of Stay, days” variable. Copy and paste the graphical summary below:



Summary Report for Avg Length of Stay, days

Anderson-Darling Normality Test

A-Squared	0.20
P-Value	0.882
Mean	4.2667
StDev	1.3702
Variance	1.8775
Skewness	0.233807
Kurtosis	-0.337610
N	79
Minimum	1.7300
1st Quartile	3.2400
Median	4.2300
3rd Quartile	5.2500
Maximum	7.9200

95% Confidence Interval for Mean

3.9598 4.5736

95% Confidence Interval for Median

3.8469 4.6199

95% Confidence Interval for StDev

1.1849 1.6249

Answer the following using the graphical summary:

a. What’s the mean value? 4.27 What’s the IQR? 2.01

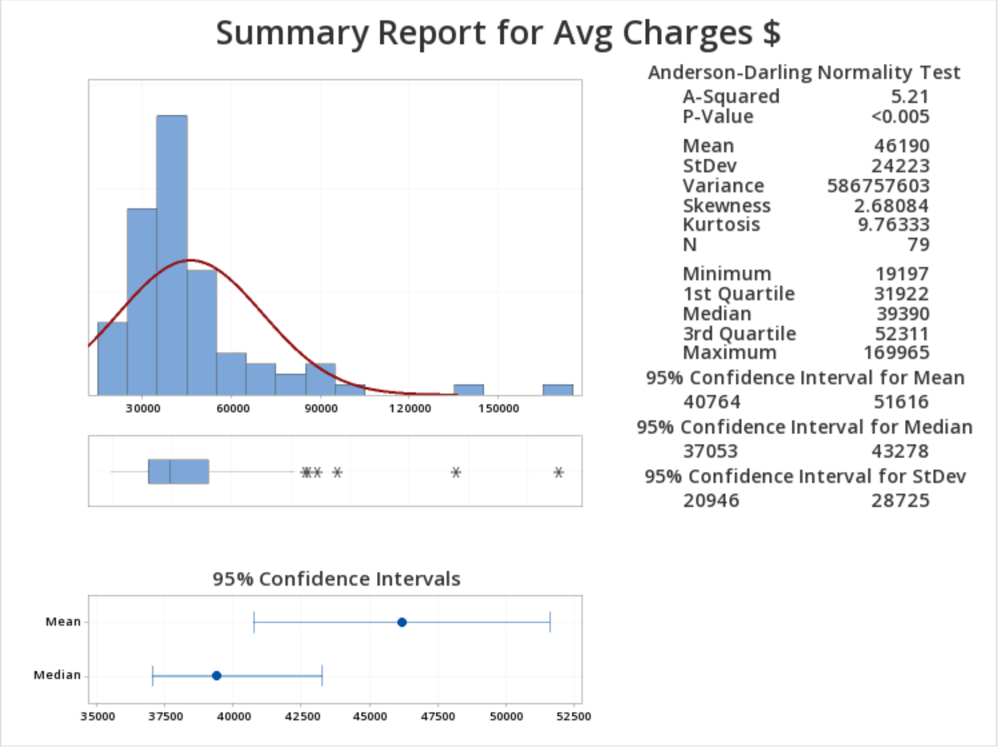
b. Can we say with confidence that these data are normally distributed? Why or why not?
The data is normally distributed. Both Kurtosis and Skew are fairly close to zero, and the minitab normality test shows that the data is fairly consistent.

c. These data are: highly skewed moderately skewed approximately symmetric

Why?
Approximately skewed because it has a right skew of 0.233, which is within -0.5 and 0.5.

d. Do there appear to be any outliers in the "Average Length of Stay, days" values? Explain.
No. The boxplot doesn't show that there are any outliers. Both ends are what bring down the normality of the data, but no so much so that they would be considered outliers.

7. Generate the Graphical Summary for the “Avg Charges $” variable. Copy and paste the graphical summary below:



Summary Report for Avg Charges $

```
Anderson-Darling Normality Test
   A-Squared                  5.21
   P-Value                  <0.005

   Mean                      46190
   StDev                     24223
   Variance              586757603
   Skewness                2.68084
   Kurtosis                9.76333
   N                            79

   Minimum                   19197
   1st Quartile              31922
   Median                    39390
   3rd Quartile              52311
   Maximum                  169965
95% Confidence Interval for Mean
   40764                     51616
95% Confidence Interval for Median
   37053                     43278
95% Confidence Interval for StDev
   20946                     28725
```

Answer the following using the graphical summary:

a. What’s the mean value? __46190______ What’s the IQR? _____20389_________

b. Can we say with confidence that these data are normally distributed? Why or why not?
No, the data is not very normal. It has a high skew, especially due to a number of outliers and high kurtosis.

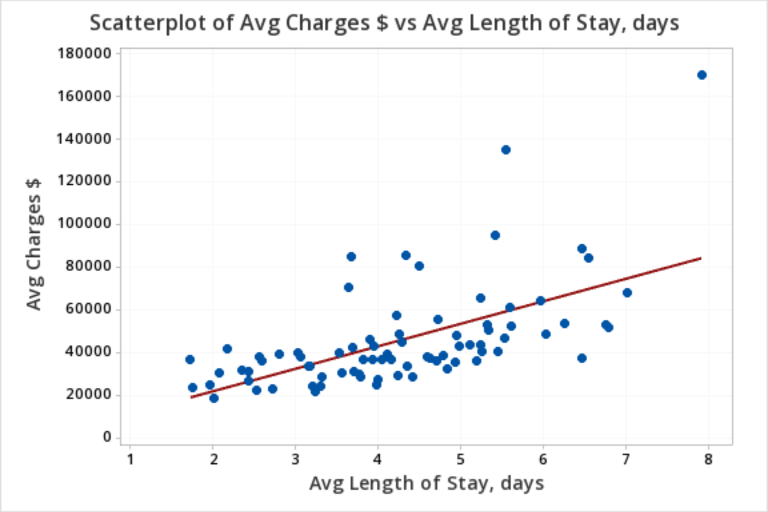
c. These data are: highly skewed moderately skewed approximately symmetric

Why?
Highly skewed because it is 2.68, which is higher than 1.

d. Do there appear to be any outliers in the “Avg Charges $” values? Explain.
Yes, the asterisks in the box plot show the outliers.

8. Generally speaking, what would you expect the relationship between the variables "Avg Length of Stay, days" and "Avg Charges $" to be? In other words, as either one of these values goes up or down for a particular medical condition, what could we reasonably expect the other to do?
   I would expect a positive correlation. The greater the length of stay, the greater the charge.

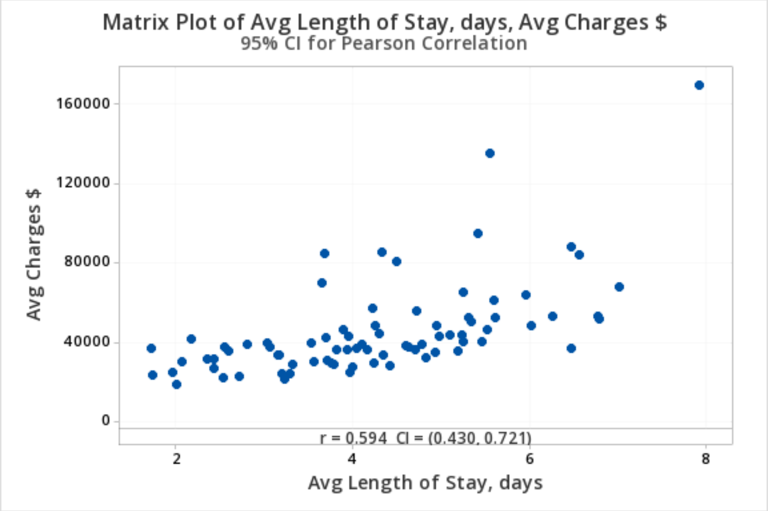
9. Generate the Scatterplot with "Avg Length of Stay, days" on the x-axis, and "Avg Charges $" on the y-axis. Copy and paste the graph below.

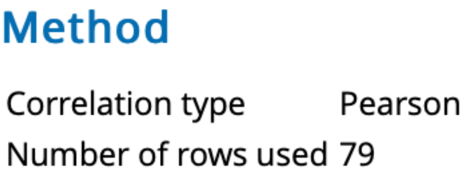


Does this scatterplot seem to support your answer in question # 8 above? Explain.

Yes it does. The data is spread out, but the regression line still shows a clear positive correlation.

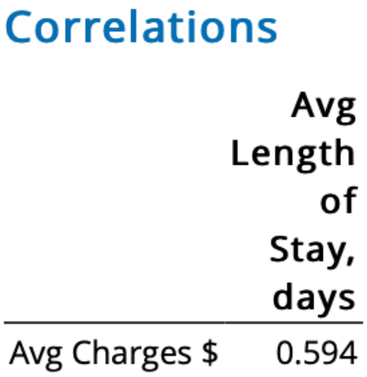
10. Generate the Correlation (Basic Statistics) plot for “Avg Length of Stay” and “Avg Charges”, and copy and paste the graph below.





Method

Correlation type	Pearson
Number of rows used	79



Correlations

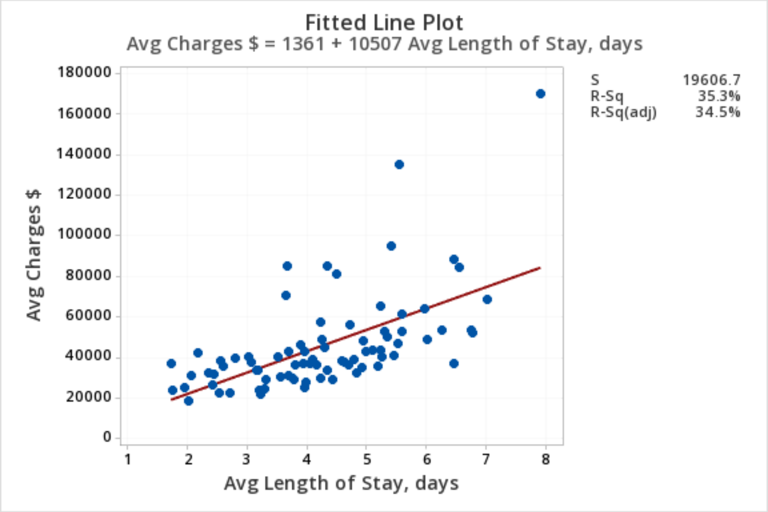
	Avg Length of Stay, days
Avg Charges $	0.594

a. What is the Coefficient of Correlation (r)?

____0.594__________________

b. Are these variables positively or negatively / strongly or weakly correlated?
The variables are positively correlated. The data is between weak and strong correlation, but a little closer to a strong correlation.

11. Generate the Stat/Regression/Fitted Line Plot with “Avg Charges” as the Response (Y-axis) variable, and “Avg Length of Stay” as the Predictor (X-axis) variable. Copy and paste the “Fitted Line Plot” below.



The regression equation is
Avg Charges $ = 1361 + 10507 Avg Length of Stay, days

Model Summary

S	R-sq	R-sq(adj)
19606.7	35.32%	34.48%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1.61666E+10	1.61666E+10	42.05	0.000
Error	77	2.96005E+10	3.84422E+08		
Total	78	4.57671E+10			

a. What is the regression equation that was calculated?
Y = b + mx
Avg. Charges = 1361 + 10507*(Avg. Length of Stay)

b. An insurance company is trying to figure out how much they are going to be billed by the hospital for certain charges. They ask you what the expected charges are for a medical condition that would keep a patient in the hospital for 5 days. You can tell them confidently that based on the data you have just analyzed, the expected charge for a patient’s 5-day stay is:

$ 53,896

c. How did you come up with the estimate for the insurance company in (b)?
I inputted 5 days into the regression equation, and the output was the cost.

d. In this particular dataset, how much of the variance (what %) in the variable "Average Charges" is explained by the variable "Average Length of Stay"?
____35.32%_______

12. In your opinion, what other factors might influence the amount charged?
The type of disease, and therefore the type of procedure done, would probably be the biggest factor in cost. Additionally, if there were multiple factors or pre-existing conditions, both could raise the cost.