Madeline Kuehn

Reflection #2

1. **Case Study:** The movie industry is a competitive business. More than 50 studios produce hundreds of new movies for theater release each year, and the financial success of each movie varies considerably. The opening weekend gross sales ($millions), the total gross sales ($ millions), the number of theaters the movie was shown in, and the number of weeks the movie was in release are common variables used to measure the success of a movie. Data on the top 100 grossing movies released in 2016 (Box Office Mojo website) are contained in an Excel file called Movie Data.xlsx.

a.

1. Descriptive statistics for each of the four variables along with a brief discussion of what the descriptive statistics tell us about the movie industry.

Opening Gross Sales ($millions)					
Measures of Center		**Measures of Variation**		**5 Number Summary**	
Mean	27.51	Range	169.2	Minimum	0.07
Trimmed Mean (5%)	24.00	Variance	703.08	Q1	12.97
Median	19.08	Standard Deviation	26.52	Q3	32.06
Mode	None	Coefficient of Variation	96.37	Maximum	169.19
				IQR	19.09

The opening gross sales clearly has some high outliers, since Q3 is 32.06 and the Maximum is 169.19. This is also evident in the fact that the trimmed mean drops when losing the 5% of data on each end. It is also evident that this data is right skewed, as the median is smaller than the mean. Most opening gross sales are around 20-24 million but can be as big as 169 million and as small as 0.07. The CoV is 96%, showing the data is very spread out.

Total Gross Sales ($millions)					
Measures of Center		**Measures of Variation**		**5 Number Summary**	
Mean	90.47	Range	351.87	Minimum	29.14
Trimmed Mean (5%)	82.00	Variance	4640.97	Q1	39.35
Median	72.4	Standard Deviation	68.12	Q3	107.08

Mode	35.06, 37.3	Coefficient of Variation	75.3	Maximum	381.01
				IQR	67.72

Similarly to the opening gross sales, the total gross sales has high outliers, as seen by the trimmed mean dropping to 82. The Coefficient of Variation is a little smaller, but the data is still spread out fairly drastically from the mean. This data, like the opening gross sales, is right skewed, so we know that there are a few movies increasing the mean over the median.

Number of Theaters					
Measures of Center		**Measures of Variation**		**5 Number Summary**	
Mean	3114.35	Range	3337	Minimum	1038
Trimmed Mean (5%)	3141.78	Variance	373065	Q1	2849.25
Median	3102.5	Standard Deviation	610.79	Q3	3553.25
Mode	2904, 2950, 3440, 3555,....	Coefficient of Variation	19.61	Maximum	4375
**Minitab said there were more than 5 modes, but just showed the lowest 4				IQR	704

This data set is symmetric with the data, though it is slightly left skewed. We know that is symmetric because the mean and median are very similar. Because the trimmed mean increased, that tells us that any outliers that this data set has are low values, or movies that didn't show in many theaters. This dataset also has a CoV of about 20%, giving another reason for this dataset being the most symmetric. Most movies showed in about 3,100 theaters.

Weeks in Release					
Measures of Center		**Measures of Variation**		**5 Number Summary**	
Mean	14.58	Range	169.2	Minimum	6
Trimmed Mean (5%)	14.24	Variance	25.50	Q1	11.25
Median	14.5	Standard Deviation	5.04	Q3	17
Mode	16	Coefficient of Variation	34.63	Maximum	43

				IQR	5.75

Just like the number of theaters, data set is symmetric with the data. It is slightly right skewed. Again, we know that is symmetric because the mean and median are very similar. Because the trimmed mean decreased, that tells us that any outliers that this data set has are high values, or movies that showed for a long time in theaters. Most movies showed for about 14-15 weeks, though there was one movie that showed for a whopping 43 weeks.

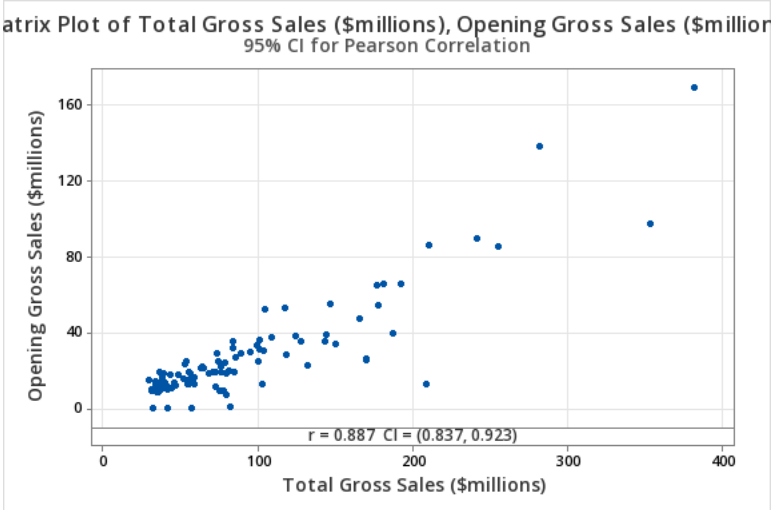
2. What movie, if any, should be considered high-performance outliers? Explain.

To determine what I considered "high-performance outliers" I used the 1.5IQR test on both the Opening Gross Sales and on the Total Gross Sales. Movies that were outliers for both variables I considered to be high-performance outliers, because they did very well, grossing far above the mean of both variables.

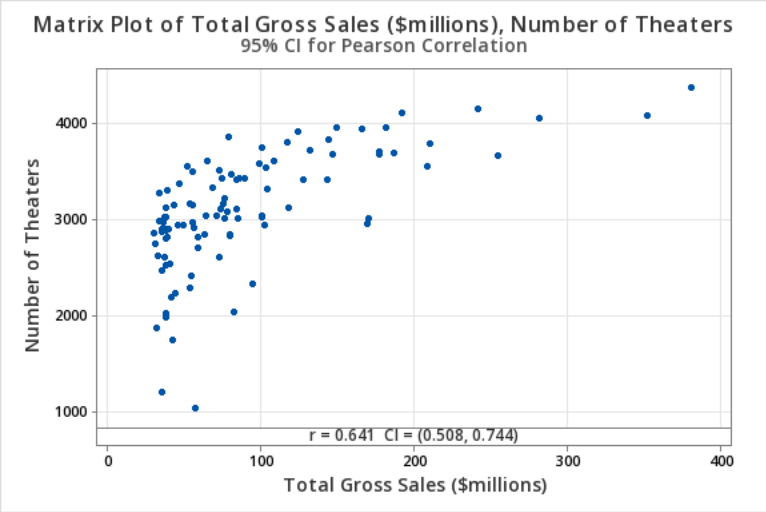
These movies were:

Movies or Motion Pictures	**Opening Gross Sales ($millions)**	**Total Gross Sales ($millions)**
Fast Five	86.2	209.84
Pirates of the Caribbean: On Stranger Tides	90.15	241.07
The Hangover Part II	85.95	254.46
The Twilight Saga: Breaking Dawn Part 1	138.12	281.29
Transformers: Dark of the Moon	97.85	352.39
Harry Potter and the Deathly Hallows Part 2	169.19	381.01

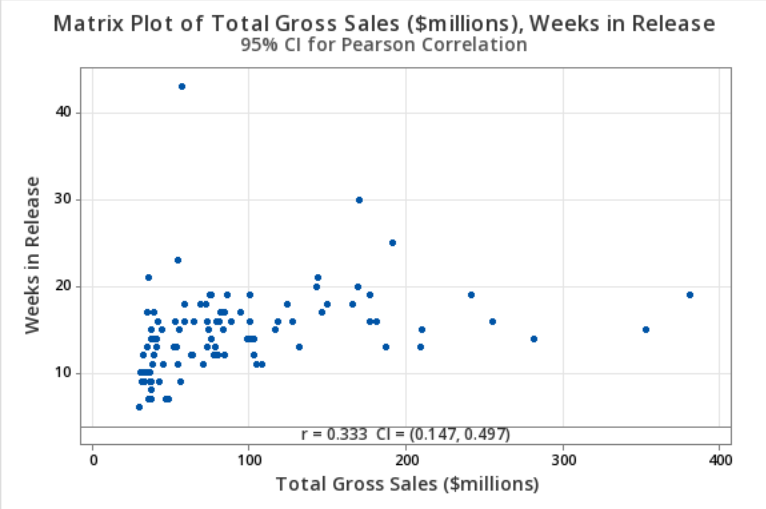
3. Use (a) scatter diagrams and (b) descriptive statistics to show the relationship between Total Gross Sales and each of the other variables. Interpret your results.



This was the scatter plot for Total Gross vs. Opening Gross. Because the Correlation Coefficient is 0.887, that means these two variables are very closely related. This makes sense, because typically movies that have a good opening perform well the entire time they are in theaters.



This is the scatter plot for Theaters vs Total Gross. This data had a correlation coefficient of 0.641, meaning that these variables are not as related, though there is a slight relationship to gross sales vs number of theaters. This makes sense to me because if a movie does well, it typically is in more theaters, but just because a movie is in lots of theaters does not guarantee that it will perform well.



The scatter plot of weeks in release vs total gross sales clearly has non-correlated data. The correlation coefficient is 0.333, which is very low. This means that how many weeks a movie is in release for does not correspond to the total gross sales.

2. Assuming that MSUM is conducting a survey on MSUM's Student Union activities. One of the questions on the survey asks students what they think of the quality of work the Student Union is doing on our campus. The possible responses are excellent, good, fair, and poor. Another question on the survey asks students how they feel about a proposed

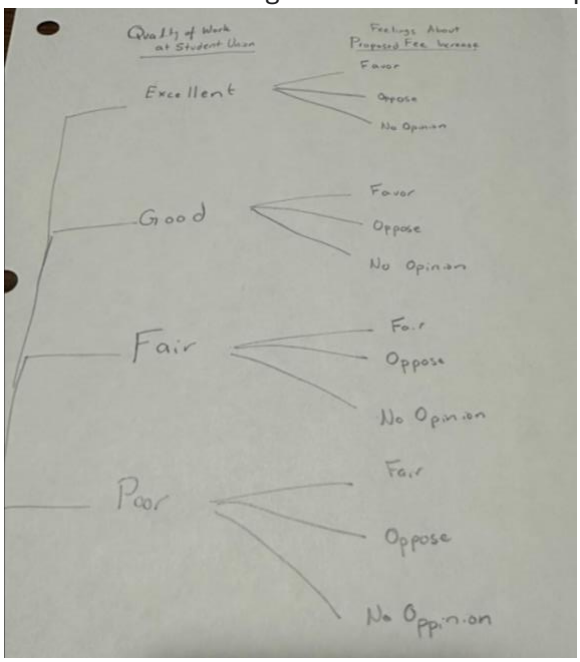
fee increase to help fund the cost of building a new student dormitory. The possible responses to this question are in favor, oppose, and no opinion.

a. Develop the sample space for someone who is responding to both questions.

There are 12 possible responses for someone responding to both questions.

S = {(Excellent, Favor), (Excellent, Oppose), (Excellent, No Opinion), (Good, Favor), (Good, Oppose), (Good, No Opinion), (Fair, Favor), (Fair, Oppose), (Fair, No Opinion), (Poor, Favor), (Poor, Oppose), (Poor, No Opinion)}

b. Show how a tree diagram can be used to display the outcomes listed in part a.



c. State the probability for each outcome in the sample space based on what you know about the way classical/theoretical probabilities are assigned.

Each outcome in the above sample space has a probability of 0.083, or $\frac{1}{12}$, since there are 12 items in the sample space. We know that each outcome is equally likely to occur.

d. Show that your assignment of probability in part c. satisfies the basic requirement for assigning probabilities. Provide justification for your response(s).

We know that there are 12 items in the sample space, and each outcome has a $\frac{1}{12}$ probability for being chosen. By Probability Law #4, the probability of each item in the sample space needs to add together to be one, or P(ExcellentFavor)+P(ExcellentOppose)+P(ExcellentNoOpinion)+P(GoodFavor)+…P(All other items)…+P(PoorNoOpinion)=1. Since $P(S) = \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = 1$, this assignment satisfies this requirement.

3. Assuming that three math Professors, Professors X, Y, and Z in the Mathematics Department have taken turns in constructing the comprehensive exam that every statistics student must pass for graduation. Because each Professor has extremely different views, it would be useful for students to know who wrote the exam questions so they can slant their preparation accordingly. Assume that Professor X has written the exam 20% of the time. Professor Y 30% of the time, and Professor Z the rest of the time. Professor X has asked a question on Bayes' Theorem 40% of the time, Professor Y has asked a question on that 30% of the time, and Professor Z, 20% of the time. If there is a Bayes' Theorem question on the exam, what is the probability that Professor Y did not write the exam?

   Use this problem to demonstrate how you would teach your high school students to understand the use of a tree diagram solving conditional probability problems that involve Bayes' theorem. Include the outline of the tree diagram and solve the problem.