Final Project

Due: Monday, 4/12/21, OR Wednesday, 4/14/21

GENERAL INFORMATION

OBJECTIVE: This **group** project is designed to test your understanding statistics and of R. Group sizes will be between 3 – 4 people for a total of 10 groups in the class. Groups will be assigned randomly. You and your group mates will:

1. Collect data or select a data set to analyze. If you don't know where to start, go here: https://www.data.gov/
2. Topic for approval: **Due Friday, 3/12/21**
3. Perform all prompts and analysis below.
4. Submit your 1) data set, 2) R-code, and 3) presentation: **Due by 11:59 pm on Sunday, 4/11/21**
5. Showcase findings to the class using a PowerPoint presentation: **Monday, 4/12/21, or Wednesday, 4/14/21**

PRESENTATION GUIDELINES

- Every member of the group is required to present at least 1 element.
- Presentations should not exceed 10 minutes.
- **You will use R to create all diagrams, tables, graphs, and computations.**

REPORT QUESTIONS AND ANALYSES REQUIRED

Using your data set, complete the following:

1. **[55 points]** Choose a variable from your data set to:
   a. *[25 points]* Find the mean, median, mode, variance, and standard deviation *s* of this *sample*.
   b. *[10 points]* Using 10 bins, create a histogram for this variable. Explain what the spread looks like.
   c. *[10 points]* Create a boxplot for this variable. Explain what the spread looks like.
   d. [*10 points*] Choose one other graphically display for this variable, such as a pie chart, bar graph, or line graph. Explain why you choose to display the data using this visualization.

2. **[60 points]** Using the variable above, find another variable in your data set that you think may have an effect:
   a. *[6 points]* Create a scatterplot and clearly label your variables on the y-axis (dependent) and on the x-axis (independent).
   b. *[24 points]* Draw in a trend line, find and clearly label the equation for the line of regression. Then, using your equation for the trendline:
      i. Approximate the dependent variable (y) value by choosing a value for the independent variable (x)
      ii. Approximate the independent variable (x) value by choosing a value for the dependent variable (y)
   c. *[10 points]* Calculate the correlation coefficient, *R.* Explain what this means for your model.
   d. *[10 points]* Calculate the coefficient of determination, R^2. Explain what this means for your model.
   e. *[10 points]* Create a Normal QQ-Plot for dependent variable. Explain what this plot means for your data.

3. **[30 points]** Assuming the data is normally distributed, perform an estimation of the Mean using a 95% confidence level by:
   a. Dependent Variable (y) from questions 1 and 2:
      i. *[5 points]* Creating a Margin of Error.
      ii. *[5 points]* Creating the confidence interval.
      iii. *[5 points]* Interpret for the confidence interval, specifically in the context of this problem.
   b. Independent Variable (x) from question 2:
      i. *[5points]* Creating a Margin of Error.
      ii. *[5 points]* Creating the confidence interval.
      iii. *[5 points]* Interpret the confidence interval, specifically in the context of this problem.

4. **[5 points]** Conclusion:
   a. What computational methods that you learned in MATH 2100 helped to complete this project?
   b. Describe one another way you could have analyzed this data.
   Note: Do not list another software you could have used. You should explain another variable or variables that you could add to your dataset to make it more interesting.