DUE DATE: FRIDAY, APRIL 29, 2022 by 11:59 PM on Gradescope
LATE DUE DATE: WEDNESDAY, MAY 4, 2022 by 11:59 PM
on Gradescope with penalty of 5 percentage points per day late.

Directions:

- Submit your project on Gradescope under "Final Project" under the correct version number. You only need to submit to one version (your assigned one).
- Project must be typed for full credit. Write your answers in the R Script (top-left hand corner) area in R Studio. Save your R Script as a .R file. Save any graphs you generate. Submit your .R file and graphs to Gradescope.
- Add appropriate amounts of white space to make your responses easy to read.
- Report your code for each part, and provide your graphs. Answer any additional questions. Clearly indicate which question you are answering.
- Answers without code will not receive full credit. Code without answers will not receive full credit.
- Provide appropriate commentary where needed.

Assignment Details:

- Work on this project INDIVIDUALLY. You may receive assistance (but not solutions) from Teaching Assistants and the Instructors. You are allowed to reference Blackboard, MyOpenMath, Gradescope, your notes, the course textbook, R documentation, and previous projects / homework for assistance.
- You cannot work in teams. You cannot work side-by-side, you cannot submit someone else's work (partial or complete) as your own. The University's policy is available here: `https://dos.uic.edu/conductforstudents.shtml`
- In particular, note that you are guilty of academic dishonesty if you extend or receive any kind of unauthorized assistance.
- Absolutely no transfer of program code or files between students is permitted (paper or electronic), and you may not solicit advice or solutions from family, friends, online forums, or websites including, but not limited to Chegg. Asking for solutions or viewing solutions on third-party websites is not allowed.
- Other examples of academic dishonesty include emailing your program or files to another student, copying-pasting code from the internet, working in a group, and allowing a tutor, TA, or another individual to write an answer for you.
- Academic dishonesty is unacceptable, and penalties range from a letter grade drop to expulsion from the university. Cases are handled via the official student conduct process described at `https://dos.uic.edu/conductforstudents.shtml`.

Notes:

- Projects will not be accepted via email.
- Points will be deducted if project not submitted to the correct version.

Project Goal:

Practice working with real data. Investigating variables. Calculating confidence intervals. Performing a hypothesis test.

Dataset Descriptions:

The data set *energydata_381.csv* contains data regarding Energy Usage (Gas: Peoples Natural Gas, and Electric: ComEd) in Chicago from several years ago. The data set consists of the following variables:

- **ID** - Unique location identifier
- **COMMUNITY.AREA.NAME** - Name of the Chicago community
- **BUILDING_SUBTYPE** - Building Sub-Type (6): Single Family, Multi < 7, Multi 7+, Commercial, Industrial, Municipal
- **TOTAL.KWH** - Total 2010 kWh from ComEd accounts
- **THERMS.TOTAL.SQFT** - Total square footage associated with the natural gas energy usage for Kilowatt Hours in 2010 according to Cook County Assessor Records
- **KWH.MEAN.2010** - Average Total KWHs for 2010
- **THERM.MEAN.2010** - Average Total Therms for 2010
- **THERMS.SQFT.MEAN.2010** - Average Therms per square foot in 2010
- **Age_Group** - Newer if the `AVERAGE.BUILDING.AGE` is less than 25, Middle1 if the `AVERAGE.BUILDING.AGE` is at least 25 and less than 50, Middle2 if the `AVERAGE.BUILDING.AGE` is at least 50 and less than 75, Middle3 if the `AVERAGE.BUILDING.AGE` is at least 75 and less than 100, Ancient if the `AVERAGE.BUILDING.AGE` is at least 100.

Dataset File:

The dataset may be found as a .csv file. It is named energydata_381.csv. There is a header in the dataset.

Project Grading (144 points):

Required Files Submitted: 5 points

Task 1: 3 points

Task 2: 8 points

Task 3: 8 points

Task 4: 9 points

Task 5: 5 points

Task 6: 7 points

Task 7: 9 points

Task 8: 5 points

Task 9: 7 points

Task 10: 16 points

Task 11: 18 points

Task 12: 15 points

Task 13: 29 points

a) 3 points

b / c) 10 points

d / e / f / g / h) 5 points

i / j) 11 points

Hints:

- To import your dataset, see R Project 5 (C1).
- To reference specific columns in your dataset, see R Project 5 (C1). You will need code similar to `cars$Overall_MPG`.
- For summary statistics, see R Project 5 (C3).
- To create a histogram, see R Project 5 (C4).
- To create a boxplot, see the Document Camera Notes from Section 8.2 / Chapter 1, page 16.
- For the Shapiro-Wilk Test, see R Project 5 (C5).
- To make conclusions about confidence intervals, see Section 9.8 Document Camera Notes page 4.
- To create confidence intervals, see the appropriate sections from the weekly content. All applicable R code needed is contained in the Document Camera Notes.
- To perform a hypothesis test, see the appropriate sections from the weekly content. All applicable R code needed is contained in the Document Camera Notes.

Tasks:

Task 1 Import your dataset into `R` and save it. There is a header in the dataset. Provide your code.

Task 2 Summary Statistics for `TOTAL.KWH`.

(a) Use the `summary()` function in `R` to find summary statistics (Minimum, Q1, Median, Mean, Q3, Max) for the Total KWH.

(b) Find the variance.

(c) Find the standard deviation.

(d) How many values are there for this variable?

For the above,

- Report your code.
- State the values that you calculated from your code.

Task 3 Summary Statistics for `THERMS.SQFT.MEAN.2010`.

(a) Use the `summary()` function in `R` to find summary statistics (Minimum, Q1, Median, Mean, Q3, Max) for the average therms per square foot.

(b) Find the variance.

(c) Find the standard deviation.

(d) How many values are there for this variable?

For the above,

- Report your code.
- State the values that you calculated from your code.

Task 4 Histogram for `TOTAL.KWH`.

Make a histogram of the Total KWH with frequency on the y-axis, where the intervals are *left closed, right open.* Set the `breaks` to be the values $0, 25000, 50000, \ldots, 225000$. Make the limits on the y-axis go from 0 to 800.

(a) Report your code.

(b) Upload your plot to Gradescope.

(c) Describe your histogram.

- State whether it is relatively symmetric or not.
- State whether it is unimodal, bimodal, or multimodal.

Task 5 Boxplot for `TOTAL.KWH`.

Make a boxplot of the Total KWH. Set the y-axis limits to be from 0 to 210000. Add the title "Boxplot of Total KWH". Are there outliers present?

(a) Report your code.

(b) Upload your plot to Gradescope.

(c) Answer whether there are outliers or not.

Task 6 Shapiro-Wilk Test for `TOTAL.KWH`.

Run the Shapiro-Wilk Test for the Total KWH. Does it appear that this variable is normally distributed?

(a) Report your code.

(b) Provide your test results as a comment within your code.

(c) One of the result values that you obtained is a p-value. Assume that if your p-value $<$ 0.05, the data is not normally distributed. Based on what you see, do you think that your data is normally distributed? Why or why not?

(d) Does your decision in Part C match what you are seeing with your histogram from **Task 4**? Why or why not?

Task 7 Histogram for `THERMS.SQFT.MEAN.2010`.

Make a histogram for the average therms per square foot with frequency on the y-axis, where the intervals are *left closed, right open.* Set the `breaks` to be the values $900, 950, 1000, 1050, \ldots, 1500$. Make the limits on the y-axis go from 0 to 850.

(a) Report your code.

(b) Upload your plot to Gradescope.

(c) Describe your histogram.

- State whether it is relatively symmetric or not.
- State whether it is unimodal, bimodal, or multimodal.

Task 8 Boxplot for `THERMS.SQFT.MEAN.2010`.

Make a boxplot for the average therms per square foot. Set the y-axis limits to be from 900 to 1500. Add the title "Boxplot of Average Therms Per Sqft". Are there outliers present?

(a) Report your code.

(b) Upload your plot to Gradescope.

(c) Answer whether there are outliers or not.

Task 9 Shapiro-Wilk Test for `THERMS.SQFT.MEAN.2010`.

Run the Shapiro-Wilk Test for the average therms per square foot. Does it appear that this variable is normally distributed?

(a) Report your code.

(b) Provide your test results as a comment within your code.

(c) One of the result values that you obtained is a p-value. Assume that if your $p\text{-value} < 0.05$, the data is not normally distributed. Based on what you see, do you think that your data is normally distributed? Why or why not?

(d) Does your decision in Part C match what you are seeing with your histogram from **Task 7**? Why or why not?

Task 10 We want to compare the average KWH Mean in 2010 for the Commercial Building Subtype to the average KWH Mean in 2010 for the Single Family Building Subtype. Create a 98.2% confidence interval for $\mu_{Commercial} - \mu_{Single}$, assuming equal variances.

(a) We need to first split the dataset into two vector that you can use for analysis. To do this, you may use the below code. You will need to change the name of the dataset to correspond to how you named your dataset in **Task 1**.

```
Comm <- datasetname$KWH.MEAN.2010[datasetname$BUILDING_SUBTYPE == "Commercial"]
SFH <- datasetname$KWH.MEAN.2010[datasetname$BUILDING_SUBTYPE == "Single Family"]
```

(b) Report your code.

(c) Provide your results as a comment within your code.

(d) State the parameter the confidence interval is for.

(e) Write down the confidence interval.

(f) Write an interpretation of your confidence interval (We are xx% confident ...).

(g) Suppose we are interested in whether there is a difference of 2000 between the two building types ($\mu_{Commercial} - \mu_{Single} = 2000$). Is there evidence that this is true? Why or why not? Your answer should reference your confidence interval.

Task 11 Create a 97.3% confidence interval for the proportion of buildings that are less than 25 years old ("Newer").

(a) To help you determine the number of buildings that are less than 25 years old ("Newer"), and the total number of buildings, copy / paste / run the below code in R. You will need to change the name of the dataset to correspond to how you named your dataset in **Task 1**.

```
addmargins(table(datasetname$Age_Group))
```

(b) Check the success / failure condition. Report the expected number of successes and the expected number of failures. Based on this information, can we use the Normal Distribution to approximate the confidence interval?

(c) Find the confidence interval by using the large sample option without a continuity correction. Report your code.

(d) Provide your results as a comment within your code.

(e) State the parameter the confidence interval is for.

(f) Write down the confidence interval.

(g) Write an interpretation of your confidence interval (We are xx% confident...).

Task 12 Create a 92.6% confidence interval for the variance of `THERMS.SQFT.MEAN.2010`.

(a) Report your code.

(b) Provide your results as a comment within your code.

(c) State the parameter the confidence interval is for.

(d) Write down the confidence interval.

(e) Write an interpretation of your confidence interval (We are xx% confident...).

(f) What assumption did we need to make to be able to construct this confidence interval? Do you think that this assumption is met? You should reference an earlier Task from this project to answer this question.

Task 13 One might be concerned that the population mean of `TOTAL.KWH` is greater than 89100. Conduct a hypothesis test at the 6.2% significance level to determine if this is the case.

(a) What condition(s) must you satisfy to perform this hypothesis test? Do you think the condition(s) is(are) met? Why or why not?

(b) State the hypotheses.

(c) Report your code.

(d) Provide your results as a comment within your code.

(e) State the test statistic value.

(f) State the p-value.

(g) State your decision (Reject / Do Not Reject) based on the p-value and the significance level.

(h) State your conclusion.

(i) Suppose you wanted to find the critical region for this test. State the critical value, and the state the critical region. Include all code required to obtain these values. Do NOT use a table or your calculator.

(j) Would you make the same decision based on the critical region that you did with your p-value? Why or why not?