Final Project Submission
CE 93 Engineering Data Analysis

Project Description

In this submission, you will continue your exploratory data analysis (EDA) using the same data sets that you have selected. You will implement techniques covered in the lectures and the labs. An EDA serves as an initial investigation to summarize key characteristics of data sets using numerical and graphical summaries before performing advanced statistical modeling. This process is crucial in any data analysis task and involves testing hypotheses and conducting preliminary statistical modeling.

This submission begins with the same sections from the first submission and includes additional sections related to topics covered later in the course. Ensure that any feedback received from the initial submission is incorporated. This final report is worth 200 points (20% of your final grade). You will submit a .ipynb file and a PDF version of your notebook. Your .ipynb and PDF submissions should resemble a report, incorporating sections with text, code, and output.

Project Instructions

You will work in groups of two (same group as Project 1). Only one team member should submit both the .ipynb file and the PDF version. Detailed instructions on group submission are available on the final page.

Everything should be completed in a single Jupyter notebook. This includes any text or paragraphs (using Markdown cells) and any code (use Code cells). The text of the document (i.e., Markdown cells) should provide a narrative structure around your code/output. All presented results must have corresponding code. Any answers, results, or plots given without the corresponding code that generated the result will not be considered. Furthermore, all code within your notebook must work correctly. Please do not include any extraneous code or code which produces error messages.

Once you finish working on your notebook, follow these steps:

- Run all cells by clicking `Run` > `Run All Cells`
- Make sure that all your visuals and answers show up and there are no errors
- Export a PDF file of your notebook by clicking `File` > `Save and Export Notebook As...` > `Webpdf`
- Submit the exported PDF file of your notebook to bCourses
- Download the Jupyter notebook file (.ipynb) by clicking `File` > `Download` and submit it to Gradescope
- Submit both files by the due date. No late submissions will be accepted. The documents will be graded jointly, so they must be consistent (i.e., do not change the Jupyter notebook without also updating the PDF document). Failure to submit both files per the instructions will result in a 0.

It is your responsibility to check both the Jupyter notebook and the exported PDF for consistency and lack of errors. Check the exported PDF file to ensure that it has an acceptable format, and that all outputs appear and are legible. You will receive no credits for any missing outputs, even if this was due to problems with exporting the PDF file. Don't wait until the due date to run your notebook and export it as a PDF file. Export a PDF file regularly after every section you complete and check for any issues.

Rubric

1. Introduction (10 points)

- Write a narrative paragraph or two describing the data sets, the variables they contain, and their data type. Expand on potential associations (relationships between the data sets) you expect, if any. (5 pts)
- Load the data into Jupyter, show a few rows, and **output** the number of measurements. (4 pts)
- Indicate whether you obtained your own data sets or if you used one of the data sets we provided. If you used one of the data sets we provided, mention its number, e.g., `CE93_04`. (1 pt)

2. Summary Statistics (25 pts)

- Compute and **output** two measures of central tendency for each data set. (2 pts)
- Compute and **output** three measures of variability for each data set. (3 pts)
- Discuss the characteristics of both data sets based on the numerical summaries. (6 pts)
- Discuss the best measure of central tendency and best measure of variability for each data set. (4 pts)
- Create a new variable that is a function of at lease one of the data sets. For example, this could be a unit conversion, or some quantity that is calculated from the data sets. Show the equation of the new variable in a markdown cell using the $ symbol (refer to EQUATIONS in 'Markdown_Guide' in bCourses). (2 pts)
- Compute and **output** one measure of central tendency for the new variable. Use the newly created variable. Do not convert the numerical summaries you computed above. (2 pts)
- Compute and **output** one measure of variability for the new variable. Use the newly created variable. Do not convert the numerical summaries you computed above. (2 pts)
- Discuss whether the numerical summaries of the new variable can be computed by converting the numerical summaries from the first two parts using the full equation you defined or part of it. (4 pts)

3. Visualizations (25 pts)

- For only one of the data sets, show **three** effective, polished plots of your choice. (18 pts)
  - The plots need to be different. For example, a frequency histogram and a relative frequency histogram count as a single plot because both are histograms.
  - These should be univariate plots (i.e., one variable per plot).
  - Each plot should have a title, axes labels, units, etc.
  - Change **at least two** default plotting parameters per plot (e.g., number of bins, line color, etc.).
  - Provide a supporting paragraph for each plot describing the data characteristics.
- Create a scatter plot of the two data sets. (7 pts)
  - The plot should have a title, axes labels, units, etc.
  - Change **at least two** default plotting parameters (e.g., line color, marker style, etc.).
  - Provide a supporting paragraph describing the relationships/trends that are apparent. Discuss what these trends suggest about the data sets.

4. Independence (10 pts)

- Compute and **output two** measures of dependence/independence between the two data sets. (4 pts)
- Provide a supporting paragraph discussing the values, the strength of the dependence/independence between the two data sets, and whether it makes sense. (4 pts)
- Discuss potential factors that could contribute to the observed dependence/independence. (2 pts)

5. Distributions (15 pts)

Perform the tasks below only on one of the data sets.

- Plot a density histogram of the data set and show on the **same** plot the theoretical density function of a normal distribution. The parameters of the distribution should be estimated from the data set. Add a legend showing the parameters of the plotted normal distribution. (5 pts)
- Generate a Q-Q plot comparing the data set against a normal distribution. (5 pts)
- Discuss the suitability of the normal distribution and identify the range of values where it may or may not be appropriate. Justify your observations with supporting reasons. (5 pts)

6. Confidence Interval Estimation for Mean Using Bootstrapping (25 pts)

Perform the tasks below only on the data set you used in 5. Distributions.

- Recompute and **output** the mean of the data set. Assume the data set represents the population. (1 pt)
- Select a random sample using Python with a size equal to **10**. Save it as `sample`. **So that your sample does not change every time you rerun Python, include** `random.seed(93)` **at the beginning of the code cell where you select the random sample.** (1 pt)
- **Output all** the values of the sample. (2 pts)
- Compute and **output** a point estimate for the population mean based on the sample. (2 pts)
- Using 5000 bootstrapped samples from `sample`, calculate and **output** a two-sided 95% confidence interval for the population mean. Interpret the confidence interval. (5 pts)
- Plot a frequency histogram of the bootstrapped means and show the confidence interval. (5 pts)
- Discuss whether the confidence interval includes the population mean and explain why/why not. (4 pts)
- How would the confidence interval based on a sample of size 40 compare to the one you calculated? Do not recalculate the interval, only discuss and provide a clear and complete explanation. (5 pts)

7. Hypothesis Testing for Mean (30 pts)

Perform the tasks below using the same `sample` you previously selected.

- Define hypotheses of the form: $H_0 : \mu = a$ and $H_1 : \mu \neq a$, where a is a constant of your choosing. Explain the rationale behind your choice of the value for a. Show the hypotheses in a markdown cell using the $ symbol (refer to the labs). (3 pts)
- Assuming that the underlying population is normal and that you only have `sample`, test your hypotheses using Python functions. **Output** the test statistic and the p-value of your test. (5 pts)
- Using a 0.05 significance level, state the conclusion and discuss the results. (4 pts)
- Choose another significance level that would change your conclusion compared to the 0.05 level. Clearly mention the significance level and the conclusion. (4 pts)
- Based on the bootstrapped confidence interval you computed in Section 6, what is the conclusion of the hypothesis test you defined above? Explain. (4 pts)
- Compare your conclusion based on the p-value at the 5% significance level with that based on bootstrapped confidence intervals. Which method is more appropriate for hypothesis testing in this case: existing Python functions or confidence intervals from bootstrapping? Discuss and explain why. (5 pts)
- Discuss the adjustments you would make to your procedure and the Python function if the sample size was 40. Predict if the p-value would decrease, increase, or if it would be impossible to determine. Do not redo the test, only discuss and provide a clear and complete explanation. (5 pts)

8. Linear Regression (30 pts)

- Run a linear regression model between the two original data sets. Clearly state which is the independent variable and which is the dependent variable and why. (2 pts)
- **Output** the slope and the intercept of the linear regression. (4 pts)
- Plot your data sets along with the fitted regression line in the same figure. (5 pts)
- Using Python functions, calculate and **output** the coefficient of determination. Interpret its value and discuss the performance of the model. (7 pts)
- Assess all the assumptions of the linear regression model. Include all necessary figures. (12 pts)

9. Conclusion (10 pts)

- Summarize the major findings of your data analysis and discuss any interesting observations. (6 pts)
- Discuss possible explanations for the observed relationship (or lack of) between your data sets. (4 pts)

10. Presentation (15 pts)

In addition to completeness and correctness, you will be graded on the presentation of your notebook. Refer to 'Markdown_Guide' in bCourses. Here are some guidelines that **should** be followed:

- Keep things looking nice! Use different header levels, colors, tables, text emphasis etc.
- Think of your notebook as a polished report you are giving to your PI or boss to summarize your work researching a topic.
- Any code cell should be preceded by a markdown cell that describes what is it you are trying to calculate/plot in the code.
- All plots should have a title, axes labels, and units (when appropriate).
- You will lose points if you print out your entire data set(s), have terrible formatting, etc.
- All required output should be printed using a combination of text and numeric values. Units should be included and the output should be rounded to show no more than 3 digits after the decimal point (or less for large numbers). It is **not sufficient** to calculate the required outputs in your code without printing them. For every task on the previous pages that requires you to calculate and output a value, your code should include something similar to this:

```
mean_temp = np.mean(temperature)
print(f'The mean value of the daily temperature is {round(mean_temp, 3)} degrees F')
```

11. Group Work Assessment Survey (5 pts)

Every member is required to complete the following survey separately to assess the group work performance on this project. You have to complete the survey to receive the 5 points.

Submission Instructions

Only one member should submit the Jupyter notebook and the exported PDF file on behalf of the group. When you are ready to submit, follow the instructions on the first page.

Read the instructions here on how to submit the exported PDF to bCourses as a group submission.

Watch this tutorial on how to submit the Jupyter notebook file to Gradescope as a group submission. Note that the groups in bCourses are not synced in Gradescope. So, you have to follow the instructions in this tutorial.

Failure to submit both files per the instructions will result in a 0.