Final Project – Choose Your Own Adventure

Report, Notebook and Slide Due 4/14 ***at noon***

Presentations 4/14 2-5pm

Purpose: Utilize your developing statistical and data visualization skills to design and complete a data-driven investigation of your own design. You will choose the dataset and design the "question" that is the focus of your investigation yourself, in consultation with Prof. Follette and your TAs.

A good dataset will be:

1. Reasonably sized for statistical investigation, but not so large that it is unwieldy
2. Sufficiently documented such that you can research the nature of the measurements (columns) and objects (rows) and form a well-informed question

As with your previous project, a good question should be:

1) Tractable (the data are capable of answering it)
2) Specific
3) Substantive (the answer is not intuitively obvious)

Report Formatting Constraints/Requirements:
(Please read ALL of these VERY carefully)

1. Graphics that you generate should adhere to the following basic constraints:
    a. All box plots should be notched box plots
    b. All histograms should have explicit user-defined bins and an explanation in the text of how these bins were selected
    c. All scatter plots should have error bars in x and y wherever possible.
    d. Axis scales (log vs. linear) for all plots should be carefully considered and decisions justified in the text of the report
    e. Fits should include a label specifying the slope and intercept, or relevant alternative parameters if fitting something non-linear
    f. If more than one type of point or line is used, the plot should include a legend

2. Formatting:
    a. Your report should have a title that describes the nature of your investigation at the top, together with your name
    b. All graphics should be numbered and should have a descriptive caption written by you (including graphics that you obtain elsewhere)
    c. All graphics should be high-resolution and a full page width, with legible axis and legend labels
    d. Sections should be numbered and labeled
    e. The report should be single spaced with 1 inch margins and a standard 12pt font
    f. The report should be submitted as a .pdf

Presentation Constraints/Requirements/Tips:
Please read ALL of these VERY carefully

- You will be given 7 minutes to present and 2 minutes to answer questions at the conclusion of your talk
- When there is 1 minute remaining in your talk, Professor Follette will hold up a red card. With 10 seconds left she will stand up. At the 7 minute mark she will stop you in order to leave time for questions. Please practice your talk and make sure that it fits within the allotted time so that you aren't interrupted mid-idea
- Each of you will be expected to ask 2 questions of your peers and to pay full attention to all presentations without using devices. Those who have already asked two questions will be asked to wait to ask further questions until all classmates have asked their 2.
- Remember to explicitly introduce every graphic in your presentation by describing the axes, types of points/lines, and "big picture" of what it is showing
- Remember to speak loudly and slowly and to speak toward the audience rather than the board
- Restate questions to check your understanding and ask for clarification where necessary

Products:

1. **An 8-10 page (no more than 10 pages!) written report that includes:**
   a. ***Background Section (~2 pages)***
      i. A description of how you found your database, what drew you to it, and what (if anything) you did to narrow the available data down to the specific data that you chose to investigate.
      ii. A description of what is contained in the dataset in both the rows and columns. This should include an explanation of how the data were collected as well as what the variables mean or measure. The nature of the objects being measured should also be described.
      iii. A description of why this dataset is interesting and important in a broader astronomical context. What types of questions were the data taken in order to inform?
      iv. This section should be supported by at least two graphics that were not designed by you, with descriptive captions written by you.
      v. Define any new terms introduced in this section (e.g. "radial velocity") in language that a student who has not taken this class and does not know astronomy jargon would understand.

   b. ***Procedure/ Data Exploration Section (~2-3 pages)***
      i. A description of how you found and imported the data that is sufficiently detailed for a peer to obtain, manipulate (where necessary), and read it into a Jupyter notebook.
      ii. A description of your initial exploration of the dataset. Describe what you did to understand the distributions of values and relationships between

variables. This description should include at least one of each of the following: a histogram, a box plot, and a scatter plot.

iii. A detailed description of any selections or manipulations to the "raw" data from the database that were necessary to establish your sample and question (e.g. turning a continuous variable (e.g. mass) into a categorical one (e.g. "high" vs "low" mass), elimination of outliers, log scaling)

iv. You should end this section with a description of the "question(s)" that you arrived at for investigation and describe the thought process, sanity checks, and background research that you did in order to ensure that it was interesting and tractable.

c. *Discussion and Analysis Section (~3-4 pages)*

i. This section should be focused on how you went about answering the question that you presented at the end of the last section and should be arranged in a narrative that tells the story of how you went about your investigation, including any dead ends or uninformative results (these don't necessarily need to be shown graphically, though you are welcome to include them, but they should be described in the text).

ii. The story of your investigation should be interspersed with data visualizations (histograms, box plots, scatter plots, etc.), computed statistics, hypothesis tests, model fits, etc. and interpretations of what they mean arranged in a narrative with data-driven arguments for how they informed or answered your question(s) or investigation

iii. Your investigation must involve at least THREE of the following

1. A model fit with a reported "quality of fit" metric. Note that this can be a fit to means/medians and does not necessarily need to be a fit to raw data. In the case of fits to means/medians you should justify your choice of statistic and use either confidence intervals or standard deviations as error bars, with justification of your choices. Do not fit anything without at least 5 data points.
2. A classical hypothesis test, including justification of why the data are amenable to it, discussion of any decisions that you made regarding p-values, confidence intervals, etc.
3. A Monte Carlo simulation and an explanation of how you decided on input parameters for the simulation and why it was necessary in the context of your investigation.
4. A Bayesian hypothesis test, including computation of a Bayes factor. For the most part, these should not be simple point estimate likelihood ratios, but should incorporate a prior and integration or summing over the likelihood to create a more informative hypothesis.

iv. You should end this section with a brief conclusion describing what you learned from this project and what you envision the next steps being were you to continue your investigation.

d. ***Tufte appendix (~1 page)***
    i. At the end of your report, you should include an appendix with a table. The columns of this table should be (a) the figure number (b) aspects of the graphic that adhere to the principles of graphical excellence, integrity, and aesthetics that we have been discussing throughout the semester and (c) aspects of the graphic that violate those principles or had to be compromised in order to make it. Columns (b) and (c) can be in bullet point form, but should be clear and specific

2. **A jupyter notebook** containing the code needed to read in the data, manipulate it, and create your plots. The notebook must:
    a. Be zipped together with any supplemental files (e.g. data) that are necessary to run it
    b. Execute linearly and without errors
    c. Contain only the code necessary to read in the data and generate the visualizations and calculations in your report. You may wish to keep a separate notebook with all of your exploratory code and visualizations that don't make it into your final report for your own reference, but please don't include anything that isn't necessary to generate the visualizations and statistics in your report in your submitted notebook.
    d. Employ markdown headings and annotations that make it easy to follow and find what is being done where
    e. Functions should have docstrings, and all non-trivial code lines should have descriptive comments.

3. **A 5 slide presentation** highlighting the results of your investigation. It should:
    a. Include 1 slide for each of the sections of your report (1 a-d above) except results (1c), which should be 2 slides. You should end with your results and mix your Tufte slide in somewhere beforehand.
    b. Be submitted as a .pdf
    c. Include minimal text and no complete sentences

A rubric for evaluation of the project is below. ***Please carefully study the criteria against which you will be evaluated before beginning the project***

Final Project Rubric

	4 points	7 points	10 points
Background/ Motivation	Student explanation of background had many and/or severe deficiencies in completeness or clarity	Student explanation of background fell a little short in completeness or clarity	Student clearly and thoroughly explained the background necessary to understand their investigation such that a peer with no astronomical knowledge could understand the context and significance of the investigation and dataset
Procedure/ Data Exploration	There were many and/or severe deficiencies in the clarity or thoroughness of the procedure explanation or the procedure itself	There were some deficiencies in the clarity or thoroughness of the procedure explanation or the procedure itself	The procedure used to analyze the data and generate the plots was clearly and thoroughly explained and justified such that a peer could reproduce results, and was appropriate to the investigation
Question	The question(s) developed for investigation were not clearly outlined or were inappropriately narrow, broad, uninformative, or unambitious	The question(s) arrived at for investigation was motivated, but could have been more clear or data-driven or didn't strike quite the right balance of feasibility and ambition	The question(s) arrived at for investigation was motivated clearly within a data-driven narrative and struck the right balance between feasibility and ambition
Final Graphics	Student-designed graphics had several and/or severe deficiencies in design or appropriateness to the investigation	Student-designed graphics had minor deficiencies in design or appropriateness, were occasionally missing where needed, or unnecessary ones were included	Student-designed graphics were well-designed, clearly and legibly labeled, informative to the investigation, and supportive of the narrative.
	7 points	**14 points**	**20 points**
Results/Analysis (worth double)	Analysis showed many and/or severe deficiencies in clarity, specificity, design, or accuracy	Analysis fell somewhat short in clarity, specificity, design, or accuracy	Plots and statistical analyses were thoroughly explained, appropriate to the investigation, and accurately interpreted

	4 points	**7 points**	**10 points**
Code	Written code applies few or none of the coding and statistical concepts from the course, is impossible to follow, or is largely incorrect	Written code applies some of the coding and statistical concepts from the course but is moderately incorrect, unambitious, or difficult to follow in places	Written code represents a clear, ambitious and correct attempt by the student to apply both computational and statistical concepts from the course
	1 points	**3 points**	**5 points**
Notebook Formatting (worth half)	Code was poorly commented or did not meet several of the requirements	Code comments could have been improved or one of the requirements for submission was not met	Notebook has headings, is easy to follow and contains only necessary code. Code is commented, runs linearly, and was packaged and submitted appropriately
Slide Design (worth half)	Slides were unreadable, extremely text heavy, or submitted in the wrong format	Slides were somewhat deficient in design or readability or were somewhat text heavy	Slides were well designed and submitted with an appropriately clear and readable format (including axis/plot labels!) with minimal text
	4 points	**7 points**	**10 points**
General Presentation Skills	The presentation as a whole had many and/or severe deficiencies in volume, appropriateness, pace, thoroughness, etc. or questions were answered rudely.	The presentation as a whole had some deficiencies in volume, appropriateness, pace, thoroughness, fielding of questions, etc.	The presentation as a whole was practiced, thorough, and given at an appropriate pace and volume. Questions were fielded thoughtfully and answered thoroughly.
Presentation Content	Description of the context and results of the investigation was unclear or insufficient or was not accessible to peers	Description of the context and results of the investigation had some deficiencies in clarity or thoroughness or was not always accessible to peers	Student clearly and thoroughly explained the context and nature of their investigation and its results, including all necessary statistical, astronomical, and computational background necessary for their peers to interpret them

	4 points	**7 points**	**10 points**
Presentation of Graphics	Graphics were glossed over or interpreted incorrectly	Student made some attempt to introduce their graphics systematically, but missed one or more key aspects	Student introduced all graphics in their presentation clearly and systematically so that they were easily interpreted by their peers

Total: ____________/100