Final Project Proposal

Start Assignment

* **Due** Apr 22 by 11:59pm

* **Points** 10

* **Submitting** a text entry box, a website url, or a file upload

* **File Types** docx and pdf

**Purpose**

The final project is your opportunity to apply what you have learned in this course to answer a question that interests you, by analyzing a real-world data set of your choice. The goal of this proposal is to get you started thinking about the project, and to allow your instructors to give you feedback if needed.

**Task**

1. Read the [Final Project instructions](https://uweau.instructure.com/courses/566248/assignments/5533280) to understand your goals for the final project.
2. Choose a data set.
   1. You may choose a data set from [Kaggle.com Links to an external site.](https://www.kaggle.com/), from another public data source, or from your work.
      * Choose a data set that we have not worked with before in this course. If you choose a data set that you have worked with in another DS course, you will be held responsible for doing an analysis that goes substantially beyond what you did before--enough work to be worthy of a final project in a master’s level course.
      * If you choose a data set from Kaggle, you may choose data from the “Datasets” section or the training data from a competition. You may **not** choose data for a competition with a prize of “Knowledge.” (These data sets are typically too well-known and thoroughly analyzed; it’s difficult to develop your own original analysis for them.)
      * To ensure that your analysis is original, we recommend that you do **not** read any code posted by other data scientists about your data set on Kaggle until after you submit your final project.
      * If you choose a data set from your work, you must have permission to share your final report with your instructors. You must also have permission to share with your instructors the general information about the data set described in step 6b, below.
   2. To ensure that your final project analysis is sufficiently deep and complex, your data set should have at least 1000 rows and at least 10 variables (after removing any columns which are not meaningful for the analysis, such as Participant ID). Smaller data sets may be allowed if you can justify what other factors will make your analysis complex (such as a large amount of data cleaning or missing data imputation).
      * The added complexity / depth does not need to be extreme (don't over-extend the scope of your project). Simply be able to justify why your data and/or analysis is a step-up from your midterm work.
3. Identify what question(s) you will investigate using this data set.
   1. Is there a response variable you will try to predict or estimate? OR
   2. Is there a structural relationship among the variables you will try to model?
4. Identify a specific company, non-profit organization, or government agency who would be interested in the answer to your question, and explain why. This group will be your target audience for the executive summary you will submit as part of your final project.
   1. **EXAMPLE 1** If you chose the Game of Thrones data set, do **not**say that your audience is “myself and my friends who like Game of Thrones.” You could say that your audience is “Bantam Spectra, the publisher of the Game of Thrones series. By having predictions of which characters will die, they will be able to make informed decisions about which characters to promote in advertisements for the series.”
   2. **EXAMPLE 2** Alternatively, the group could be more interested in your method than your results: “My audience is Tor Books, a competitor of Bantam Spectra. By having a model for which characters are likely to die in a particular series of books, they will be able to identify other series that follow a similar model, which may appeal to the same readers.”
5. Identify **2 data mining approaches** from **2 different lessons** in this course that you can use to address your question. At least one of the approaches must be different than the approaches you used on the midterm project.
6. Write the project proposal. The proposal can be fairly short (~1 paragraph; certainly not more than 1 page). It just needs to answer the following questions:
   1. What data set will you use? If it is available online, include a hyperlink to it.
   2. Include the following information:
      * Number of rows in the data set
      * Total number of variables you will include in the analysis
        + Exclude any variables you plan to ignore, such as “Participant ID”, or if you plan to ignore complicated character variables (such as “Text of Review” for a data set about AirBnB)
      * Number of categorical variables
      * What data cleaning do you anticipate doing?
        + Imputing missing data
        + Transforming variables to reduce skew
        + Creating new variables or extracting information from existing columns (such as extracting the month from a column containing the date)
        + Other?
      * Do you feel confident about how you will accomplish the data cleaning you anticipate doing? If not, your instructors can give suggestions in the feedback on your project proposal. For example, maybe you want to transform variables to reduce skew, but some of your variables are left-skewed, or negative, so a log transformation won’t work.
      * If the data set is **not** from Kaggle:
        + If the data set has 15 or fewer columns, list all the column names.
        + If the data set has more than 15 columns, list the names of 10 columns that give a good example of the sorts of information available in the data.
   3. What question(s) will you investigate using this data set?
   4. What specific group would be interested in the answer to your question? Why?
   5. Which 2 data mining approaches will you use?  Describe how these will allow you to answer the question(s) you chose in part c.

**After submission**

* You may begin working on the final project as soon as you have submitted the proposal. You do **not** need to wait for feedback.
* Your final project may include methods different from the ones you proposed, but you will be given feedback on the ones in the proposal.
* We will provide feedback on your final project proposal in Canvas. If you have follow-up questions, please ask on Piazza.  (We do not regularly check the comments on Canvas assignments that are already graded.)

**Grading**

If your project proposal meets the following criteria, you will earn full credit for the proposal:

* Answers all the questions listed in step 6, above
* Data set has at least 1000 rows and at least 10 variables (after removing any columns which are not meaningful for the analysis, such as Participant ID), or proposal gives a credible explanation of why the analysis will still be sufficiently complex
* Data set is one we have not worked with before in this course, and is not from a Kaggle competition with a prize of “Knowledge”
* Question of interest is clearly stated and answerable based on the data set
* Credible explanation of why the target audience would be interested in the answer to the question
* Proposes two data mining techniques:
  + From different lessons
  + At least one of which is different from what you did on the midterm project
  + Which are relevant to the question of interest (e.g., don’t propose logistic regression if your response variable is quantitative)

If your proposal does not meet all of the criteria above, you will lose points, and you may be asked to re-submit the proposal.