

Final project

The final project is a group assignment. Each student will work with their assigned discussion group to complete the assignment.

Overview: The final project is a group data analysis based on a data set of your group's choosing (see below for details on the data selection). There are several steps to the project, but it will culminate in a data analysis report and a short group presentation to your discussion section during the final week of the semester. The analysis will be based on a question your group wants to investigate with the selected data set. The main part of your analysis should involve inference (e.g., confidence interval estimation, hypothesis testing, regression), and interpretation of the result.

The project must include:

- the acquisition of a set of data;
- one or more questions to address with the selected data set;
- the creation of an R Markdown document that:
 - uses appropriate methods from the course;
 - is knitted into an HTML document;
 - contains:
 - a reproducible process of data manipulation;
 - graphical exploration;
 - modeling and analysis;
 - interpretation;

Due dates on home page.

Grading

The entire final project will be worth 100 points. Note that 10 of the 100 points are based on the quality of *your individual assessments of the work of other groups* with the Peer Review. (The data description and background will be graded as a group discussion assignment rather than part of the project grade.)

part	points
Proposal	20
Draft	20

part	points
Peer Review	10
Final Report	50

Data

You are **strongly encouraged** to find a novel data set for the project. You must have permission to use any data set selected. The data should contain multiple variables where there are interesting questions to address about relationships among these variables. You may choose to use data *related* to some of the case studies in class. For example, you could gather:

- weather data from a city other than Madison;
- variables from the exoplanet data set we have not examined;
- freeze dates from a different lake;

The best choice will be finding data on your own that is related to a topic of common interest in your group.

Here are some places to find data sources:

<https://github.com/awesomedata/awesome-public-datasets> ➞

(<https://github.com/awesomedata/awesome-public-datasets>)

<https://www.data-is-plural.com/> ➞ (<https://www.data-is-plural.com/>)

<https://archive.ics.uci.edu/ml/datasets.php> ➞ (<https://archive.ics.uci.edu/ml/datasets.php>)

<https://msropendata.com/categories> ➞ (<https://msropendata.com/categories>)

<https://methods.sagepub.com/Datasets> ➞ (<https://methods.sagepub.com/Datasets>)

<https://www.pewresearch.org/download-datasets/> ➞ (<https://www.pewresearch.org/download-datasets/>)

<https://opendata.cityofnewyork.us/> ➞ (<https://opendata.cityofnewyork.us/>)

<https://ieee-dataport.org/datasets> ➞ (<https://ieee-dataport.org/datasets>)

<https://catalog.data.gov/dataset> ➞ (<https://catalog.data.gov/dataset>)

<https://datahub.io/collections> ➞ (<https://datahub.io/collections>)

Note: You may get a data set from a repository such as Kaggle, but be warned if you do so that you need to answer many questions about the primary collectors of the data, not just this secondary source. In other words, you may not only cite the repository as the data source, but you will need to find the original source of the data.

Question of interest:

Your project must be centered around **a central question of interest (usually just one, sometimes two if you want)** that relates to your dataset and should inform and direct the flow of your entire project. This question should be some kind of **inference-style question**, i.e. NOT just saying what is the highest number in a column or making a plot, but rather some deeper question about the population distribution. Examples of kinds of inference questions that might be appropriate that we will cover in class include:

- Inference on a single population proportion (e.g. what is the true proportion of X in some population, or is there evidence that the proportion is/is not X in the population?)
- Comparing two population proportions (e.g. are the proportions of X and Y in the population (or the proportions of X in two different populations) equal or different?)
- Inference on a single population mean (e.g. what is the true mean of X in some population, or is there evidence the mean is/is not X in the population?)
- Comparing two population means (e.g. are the means of X and Y in the population (or the means of X in two different populations) equal or different?)
- Simple regression of two numeric variables (e.g. using continuous X to predict continuous Y).
 - This topic is covered fairly late in the course, so you may need to study up on this yourself more if you are using this as a question.

Project Proposal

The project proposal should be a knitted R Markdown document (.html) which contains the following elements.

- The names of the students in the group.
- A brief description of how the group will communicate and work together.
 - When during the week can the group meet? Consider needs of all group members.
 - What platforms will the group use to communicate and share documents.
- The question or questions of interest that motivate the planned study.
 - The topic selected must be unanimous among group members.
 - Each question should be in the form of a research question which may be supported or refuted by the data analysis and typically proposes or predicts a relationship between some variables in the study.
 - Note that your project should aim to answer a limited number of questions (preferably one) using data from one or more source.
 - Do not try to answer too many questions. *Better projects will have a focused set of one or two questions.*
- The source(s) of data.
- A description of the data. Include in this description answers to these questions:
 - Who collected the data?

- How were the data collected?
- What are key variables in the data set, what do they measure, and how are they related to your key questions?
- **What does a single row of the data set represent in real terms?**
- If the data may be thought of as a sample from some larger population, what is that larger population?
- Note: if you get the data from a data repository such as Kaggle, you will need to dig into supplementary information to answer these questions.
 - It is insufficient to simply report a link to a Kaggle repository.
- Include a description of a graph which will show how the data informs the primary question of interest.
 - Describe what variables you will include and what type of graph you will create.
- Finally, please do your best to identify which "type" of question from the "Question of Interest" section yours maps to.
 - In particular, please identify if your variable(s) of interest are continuous or categorical, and which (if you have two variables) you are treating as the cause and which you are treating as an effect.
 - If you have a question that is too vague like "What predicts COVID" or "What is the best basketball team of all time", please consider refining it to line up with one of those forms above!
 - If you are having trouble; please email us and we are happy to help you!

Project Draft

You should turn in three (or more) documents:

- One (or more) files with your data.
- An R Markdown file with your analysis.
- A knitted HTML file which contains your *full* report.
 - The format of the report is described in detail below.

Notes

- The R Markdown file should include a section in which you read in the raw data files and transform the data for analysis. Exclude this R code from the knitted HTML report by using `include=FALSE` in the corresponding R chunks.
- Your data analysis should include data exploration including graphical and numerical summaries which do not appear in the final report. You may exclude such analysis by using `include=FALSE` in the corresponding R chunks.
- The methods of analysis should primarily be methods discussed from the course.

Data

[5Ka1BURmNkVEF3TWpabVIXeHNZbUZqYTE5MGN6MHhOalEzT0RnM056QTFJaXdpWIhod0lqb3hOalEzT0RnNE1EQTFmUS4zU2JtbmdsOGhOdXJuUG1YZTJtdTBGVWZQVEk0cIRhV3NGN3lwZHBYa2JMcTB3d3kzS1RPanluRm80dDIwVjhKMIBRdl90V2JCUVBMQk5ndkdZeUNRUSlslmV4cC16MTY0Nzk1NDEyNH0.975Y3C5zMmiO1xBDxjzdXsFEHHJFhbde8ACH4IlbZYzY4HMMjQBo_u_sqbvPGH3B1xTE72o13-2MZ20K-nZ2Fjg#fn1\)](#)

- Background
 - Describe the data set, how it was collected, what the variables mean.
 - Cite the source of your data.
 - Describe any background information needed to better comprehend the question or questions you are posing and how the data relate to the question.
 - Describe any unusual factors which may affect interpretation of results.
 - Describe what you intend to do in the rest of the report.
- Analysis
 - Include numerical and graphical summaries of the data. Be sure to introduce and describe your numerical and graphical summaries. Do not display a series of summaries without text explaining them.
 - Make sure to include at least one graphical display of the data that best supports your main conclusions for your primary question, and explain how it supports your conclusions.
 - For each inference you make, briefly describe the method you use and what the results are.
 - Provide evidence to support each of your claims.
 - This section should not include substantial interpretation of results. Leave that for the discussion.
- Discussion
 - Provide broader interpretations of your analysis and describe how to interpret your results with respect to your question of interest.
 - Discuss any potential short-comings of the analysis.
 - Discuss potential future directions for additional work
 - New questions
 - Different methods to address the same questions
 - New data you might collect to refine your understanding
 - Summarize your primary conclusions and the primary evidence that supports these conclusions.
- References
 - In the R Markdown file, you may automatically cite other resources by using a caret followed by the reference between square brackets: `^[your reference text]` which will place a marker at this location and a footnote at the end of the document.

Peer Review

- Each individual student will be asked to read and comment on the draft report HTML of another group.
 - In this way, each group should receive feedback from as many students as there are in the group.
- You will be graded on the quality and usefulness of your review.
 - This is the only part of the assignment that is graded individually.
- Do not use abusive language when providing feedback.
- Full credit requires specificity.
- Your peer review will be evaluated in each of five areas, and you will receive up to 2 points for your comments on each area.
- Areas: Background, Data, Graphs, Analysis, Interpretation
- The point distribution below is how we will evaluate *your* peer review.
- Note: you do not assign grades/points to the report you review, you only provide comments/feedback.

points	description
2.0	detailed responses with helpful ways to improve
1.5	responses, but either too general or too few
1.0	minimal unhelpful response or responses
0.0	not addressed or blank

Final Project Report

- The format of the final project is the same as the draft report.

Oral Presentation

- During the last discussion meeting of the year, your group will share an oral presentation of your work.
- This will be graded as a group discussion assignment and not part of the project grade.
- There will be a time limit for your presentation based on the number of groups in your discussion section.
- All group members should be involved with the creation of the presentation, but not everyone needs to speak during it.
- Follow your TA's instructions on how to share your work with him prior to your discussion meeting so that all students may see the presentation.
 - Some TAs will let you hook up your computer, some will want the presentation shared or a link provided

- Each presentation should include no more than seven slides with the following:
 1. A title slide with a meaningful project title and the names of the group members
 2. A single slide to describe background material
 - Use bullet points and/or images
 - Be prepared to say more than what the slide says with words
 3. A single slide on your main question of interest
 - Pick the most interesting results and conclusions from your analysis
 4. One to three slides with graphs
 - The key question that you discuss in the presentation should have an accompanying graph
 - A single slide may include an arrangement of a few graphs
 5. A summary slide with your key conclusions and the final take-home message for the audience

Group Experience Feedback

- The expectation is that each member of the group will share in the work of completing the project.
- Different aspects of the group work may be partitioned among different group members.
- The group should come to an agreement on how to decide who is responsible for different parts of the project.
- Each member of the group should contribute substantially to the project, but it is normal that some individuals will do more work on some aspects while others do more work on other aspects.
- However the work is divided, every individual group member is responsible for knowing details about all of the work that is turned in by the group.
- When the project is complete, each individual will have a chance to answer survey questions about the group dynamics and if it is fair for each group member to receive full credit for the project.

Example Reports from Previous Semesters

Here are a couple of anonymized student reports from previous semesters of 240 that received higher scores. Note a few items before you look at these examples.

- The purpose of providing these examples is to give you a general idea of what a data analysis report is like.
- *The example reports are not perfect.* Do not follow all of their features. For examples, ...
 - Do *not* include R code or raw R output (such as from `t.test()` or `summary(lm())`) in the report.
 - Do suppress warning messages.
 - Do *not* include a sequence of graphs without explanation between them.
- The directions for the example reports were different from the directions you are given above. Follow the directions for this semester.


- The description of a model thesis statement these students follow differs from what you were given.

Follow the description you were given.

- Students do not always describe or interpret hypothesis tests, p-values, and confidence intervals correctly.
- The choice of analysis methods are fine, but not always the most appropriate.

[nba-mvp-prediction.html \(https://canvas.wisc.edu/courses/397597/files/37880151?wrap=1\)](https://canvas.wisc.edu/courses/397597/files/37880151?wrap=1) 




(https://canvas.wisc.edu/courses/397597/files/37880151/download?download_frd=1)

[nyc-graffiti.html \(https://canvas.wisc.edu/courses/397597/files/37880153?wrap=1\)](https://canvas.wisc.edu/courses/397597/files/37880153?wrap=1) 
 (https://canvas.wisc.edu/courses/397597/files/37880153/download?download_frd=1)

[tumors.html \(https://canvas.wisc.edu/courses/397597/files/37880152?wrap=1\)](https://canvas.wisc.edu/courses/397597/files/37880152?wrap=1) 
 (https://canvas.wisc.edu/courses/397597/files/37880152/download?download_frd=1)

References

- Your written report should include references to data sources and also to other references with background information.
- You may use R Markdown to cite references in place and automatically generate a reference section at the end of the report

1. <http://www.guide2research.com/research/how-to-write-a-thesis-statement> 
(<http://www.guide2research.com/research/how-to-write-a-thesis-statement>)   (

[B3d3kzS1RPanluRm80dDIwVjhKMIBRdl90V2JCUVBMQk5ndkdZeUNRUSIsImV4cCI6MTY0Nzk1NDEyNH0.975Y3C5zMmiO1xBDxjzdXsFEHHJFhbde8ACH4IlbZYzY4HMMjQBou_sqbvPGH3B1xTE72o13-2MZ20K-nZ2Fjg#fnref1](#)