SIADS 591/592

Milestone I Project Guidelines

version 2022.11.17.1.CT

## 

[1.0 Introduction](#_hi1g9fls4kkt)

[2.0 Project proposal and team formation](#_9kg3gk4iiqe1)

[3.0 Draft proposal (10 points)](#_ane1z18fivlb)

[4.0 Draft proposal peer review (10 points)](#_6v3iokgrokfi)

[5.0 Project report (50 points)](#_ckyd84w4synq)

[5.1 Report length](#_lek9l522rwab)

[5.2 Report format](#_7fdq45ewskml)

[5.3 Report Submission](#_565cez4pei8x)

[6.0 Project Report Feedback (10 points)](#_b0zcuz8ozbhr)

[7.0 Technology Choices](#_p8s3inb9qb5p)

[Changelog](#_hg9yk5wxcr1)

## 

## 1.0 Introduction

The team project component of SIADS 591/592 Milestone I is designed to give you the opportunity to apply the skills, tools, techniques, and knowledge that you've learned in your prerequisite MADS courses to a new problem.

Overall, the goal of this project is to provide you with an opportunity to bring your unique interests, creativity and ingenuity to our data world into alignment with one or two team members and demonstrate your team’s ability to apply data science skills and concepts on public datasets of your choice.

In particular, the project should show off your team’s ability to take a minimum of two datasets that have different features and/or access methods and clean and manipulate them in order to extract a useful byproduct: something more than you could have gotten from either dataset by itself. The cleaning and manipulation could involve filtering, format conversion, handling missing or noisy data, matching records from one data source with corresponding records in the other, and so on, and must make significant use of the programming tools we've covered in prior MADS courses.

**You are not expected to use tools and techniques from courses that are not prerequisites of Milestone I, and in fact you are discouraged from doing so as neither your peers nor your instructors may be in a position to offer helpful advice.**

## 2.0 Project proposal and team formation

There are several steps involved in the project:

1. Problem formulation
2. Team formation
3. Selection of data sources
4. Project proposal creation
5. Peer feedback on project proposals
6. Final project report
7. Peer feedback on project reports

Team formation and the selection of project topics and datasets is a student responsibility. Advertise interests via Slack and coordinate via the Project proposal and team tracker Google Sheet (see Coursera week 1 readings for the link to the document).

We expect that **most teams will consist of two to three (2-3) students**. The maximum team size is three (3) students. Working alone or working in groups of more than three is ***not*** permitted. It has been our experience that working in pairs or groups of three balances high-quality work with the amount of work each person needs to do to complete the project. Each of the steps is described below.

Before starting work on the actual project, you'll write a short project proposal that is meant to be a high-level summary and does not need to contain technical details or code. You will then review your peers' proposals, and they will review yours. You can then fine-tune your project based on the feedback you receive. Requiring the proposal to be submitted early is intended to get you thinking about questions and datasets you're interested in, and how those might be answered with the tools you've been exposed to in previous classes.

## 3.0 Draft proposal (10 points)

You *must* use a copy of the project proposal template to create your project proposal (see Coursera week 1 readings for the link to the document, which you *must* copy).

Your proposal will be reviewed by a minimum of two peers from the class and your team should take their feedback into consideration when you work on your project. The proposal draft is worth up to 10 points.

**Please note:** when your project proposal is ready for review please provide a link to it via the project tracker and coordination spreadsheet. Be sure to enable commenting on your Google Doc in order to permit peer reviews.

## 4.0 Draft proposal peer review (10 points)

You will receive credit for completing reviews of your peers' proposals. Each review is worth up to 5 points toward your final grade. Note that your proposal grade will not be based on the content of your peers' reviews.

The purpose of the peer reviews is threefold:

* gain experience reviewing proposals,
* learn about other work going on in class,
* get feedback on how to make your project better.

Your reviews should be several sentences long and should take into consideration the following points:

* **professional:** what would a co-worker think about your review?
* **pleasant:** courtesy goes a long way
* **helpful:** what sort of advice would you want?
* **scientific:** focus on facts, not opinions
* **realistic**: keep scope in mind
* **empathetic**: how would you feel if you received the review you wrote?
* **organized:** make it easy for the recipient to follow your train of thought

A useful approach when writing peer reviews is the "two stars and a dog" approach. In other words, highlight two things that the authors did well and identify one area where they might spend some time improving their work (and make constructive suggestions about how to do so).

Peer reviews will take the form of comments on the Google Doc containing the proposal and will be coordinated via the Google sheet.

## 5.0 Project report (50 points)

First and foremost, following the individual original work policy stated at the start of the course, the topic and questions you ask in your project must be of your own invention.If you used ideas from a particular web site or previous project, or did your project as part of an existing research collaboration, you must identify your sources and/or collaborators and provide links and citation(s) where appropriate.

## 5.1 Report length

**The report MUST be no more than 10 pages** including all material (code snippets, visualizations, references, etc.). The final report must be delivered as a PDF. Font size must be at least 10 point and margins must be at least 1" all around. Avoid the use of a cover page. Instead, include the title and authorship at the top of the first page. The page limit will be strictly enforced: we will only consider the first 10 pages when assessing your report.

## 5.2 Report format

The format of the report is semi-flexible - you can include additional information (keeping in mind the 10-page limit), but at a minimum the report should include the following sections:

1. **Motivation** (5 points): Briefly state the nature of your project and why you chose it. What specific question or goal did you try to address?

2. **Data Sources** (5 points): Describe the properties of the two dataset(s) or API services you used. Be specific. Your information at a minimum should include but not be limited to:

* where the datasets or API resources are located,
* what formats they returned/used,
* what were the important variables contained in them,
* how many records you used or retrieved (if using an API), and
* what time periods they covered (if there is a time element)

For example, if you downloaded data or used API services, you should state the specific URLs to those files or resources. It should require zero effort on the teaching team’s part to find and access the exact resources you used if we need to do so.

3. **Data Manipulation Methods** (20 points): For each of your two sources, describe how you manipulated the data. For example:

* How specifically did you need to manipulate the data?
* How did you handle missing, incomplete, or incorrect data?
* How did you perform conversion or processing steps?
* What variables and steps did you use to join the two data resources to perform your data analysis?
* Briefly describe the workflow of your source code and what the main parts do.
* What challenges did you encounter and how did you solve them?

4. **Analysis and Visualization** (20 points):

* A key goal of this project was bringing together two different data resources to answer an interesting question or find a new insight that could not have been answered with either data resource alone (which you summarized in part 1). Now describe the analysis steps you performed on your combined dataset to address that goal/question. Be specific, and include references to key functions or parts of your code.
* What interesting relationships or insights did you get from your analysis?
* What didn't work, and why?
* To summarize your findings, include at least one visualization (chart, plot, tag cloud, map or other graphic) that summarizes your analysis.

5. **Statement of Work** (0 points)

* You must include a statement that describes the contribution that each team member made to the project.

## 5.3 Report Submission

Please submit a zip file containing following files:

* Your project report, as a single PDF document. Remember, the project report should be no more than 10 pages in length. Keep this in mind if you are planning to generate the PDF from a Jupyter notebook. We strongly recommend that you use Google Docs, Word, or some other word processing package to generate your final PDF.
* All source code files/scripts (Python, or any other code) used for your project in a source/ folder in your zip file.
* Working URLs that point to either (a) the actual data/API resources you used or (b) if the datafile is over 10 Mb or not available in file form, create a sample file containing the first 100 records.

Your project report MUST be named the same as your team name in Coursera and MUST be in PDF format (e.g. 27-arwhyte-cteplovs.pdf). That PDF file should be contained, along with any code or ancillary material, in a zip file that Your zip file MUST be named the same as your team's name in Coursera (e.g. 27-arwhyte-cteplovs.zip). Do not submit large data files as part of your zip file, but be sure that the data is accessible by all class members, including the teaching team.

You will also be required to submit your PDF to a platform that will facilitate the sharing of your report with the class. The details of this will be announced during the course.

## 6.0 Project Report Feedback (10 points)

You will have an opportunity to provide feedback on two project reports. We recommend that you choose the same projects that you provided feedback for at the proposal stage. We will likely use a gallery tool built by the Center for Academic Innovation (CAI). More information will be provided regarding the review mechanism later in the course.

## 7.0 Technology Choices

This course differs from other MADS courses in many ways including technology. We have created a Jupyter environment for you that is functionally equivalent to SIADS 516, which is a superset of the base MADS environment, and you can access that environment via the "ungraded lab assignment" in Coursera. You can use that environment or choose to use any of the environments from courses you have already completed to build and test data manipulations and visualizations for your project. Alternatively, you can use your own locally installed environment. Another possibility is to use [Google Colaboratory](https://colab.research.google.com/notebooks/intro.ipynb), which may facilitate collaboration.

***As part of the grading the teaching team may attempt to reproduce your results using your code and data, and you are expected to assist with this if we request it.***

## Changelog

(2022.11.17.1.CT) Remove bonus points

(2021.07.24.1.AW) Adjust title, number sections, adjust bonus points, edit text, remove links to companion Google docs (refer to Coursera for links)