Capstone Project: You and Data Science

Due: Sunday, July 17th at 11:59 PM

Throughout this semester, you have grown into an amazing Data Scientist! You are analyzing datasets in RStudio, performing advanced statistical tests, and finding the answers to complex questions using data. You have seen dozens of datasets we have provided. For the final project, we want you to teach us something about your Chicago community - we want to learn about something you are passionate about!

For this final project in Discovering Data Science, you will use Data Science to explore something you are passionate about or interested in learning more about in your Chicago community. In the end, you will write a small paper telling us about what you found and teaching us something! We only have a few minimal requirements:

- You must use a non-trivial dataset. The dataset must have at least 200 data points (this could be 20 rows with 10 columns, 50 rows with 4 columns, etc).
- You must do some analysis using RStudio. You will turn in your code. You must do something, but it could be anything.
- You must submit a paper/report that provides a summary of what you found and teach us about your passion/interest. The paper must be at least 2 pages (and double-spaced), but up to half of that can be figures/graphs. Full details are below.

With students from so many different majors in Data Science Discovery, we recommend collaborating on ideas for your project. Keep in mind plagiarism and collaboration are two different ideas. Plagiarism will not be tolerated, be careful. We are excited about everything we are going to learn from you! :)

Setting Up Your Project Workspace

To complete this project, there is no starter code – you are building it from scratch! However, we do want to check out your work so make sure you keep your code, dataset, and other files together in one place to share or submit your project when you turn it in at the beginning of Week 5.

Dataset

We hope that you will use a dataset you are passionate about that relates to your Chicago community. It can be anything – it can be a dataset used from another class (eg: think if you had any data you get in Excel), it can be a dataset you found online, or it can be a dataset you gather yourself. Some ideas include:

- A dataset about a hobby you're interested in (eg: vacation destinations, best beaches, fashion trends, Instagram, etc.)
- A dataset about something you enjoy doing or watching (eg: swimming, volleyball, Rocket League, Chicago White Sox, etc.)
- A dataset about a topic related to your major (economics, communications, political science, etc)
- Any dataset that means something to you.

The dataset must relate to Chicago. A great place to find lots of Chicago data is the Chicago Data Portal. You should be able to find plenty of datasets in the Chicago Data Portal, but if you want to look in other places for Chicago-related data sets you can search through these other free resources that contain millions of datasets:

- Google Dataset Search
- Kaggle
- Data.Gov (U.S. Open Data)
- FBI Crime Data Explorer

Project Report

The major deliverables for this project are a small paper or report over what you found and your code. We want to learn something from you about your interest/passion in the Chicago community, so tell us a story about what you discovered!

The only requirements are:

1. Your report must be at least two pages. (It can be more, use enough space to tell us what amazing things you found.)
2. Your report must be double-spaced.
3. Your font size should not be greater than 12. Feel free to include images, diagrams, figures, etc. The only requirement is that we want at least a page of text in your report (you can have three pages of diagrams so long as there's at least a page of text somewhere in it all.)

4. Your audience is going to be the class. You do not need to explain R or Data Science to us, but you should not assume we know anything about your specific interest/passion.
5. Your report must include four sections:
    a. Introduction: Why is the dataset important to you? What do you want to discover about it?
        i. Must include a source of the dataset.
    b. Methods: Briefly summarize what steps you took to analyze the data. This would include processing, cleaning, modifying, grouping, using algorithms, etc. - write about what you did but keep it concise.
    c. Results: A summary of the exciting discovery you made! This will be when you write about your discovery - you can show any interesting plots here as well.
    d. Conclusion: What would you like to do next? Did this analysis inspire you to go and discover new things? Tell us about the next analysis you'd like to do! (It can be about the same topic or even a new one that this project got you thinking about!)

Project Presentation

You need to prepare a short 5-10 minute presentation (i.e. google slides, Prezi, or any other presentation tool you want to use) to share your exciting discovery with your class! You don't have to explain any of the R or Data Science to us, but you should not assume we know anything about your specific interest/passion.

Presentations will take place during the last two days of classes (i.e. Monday and Tuesday, July 18-19). You will sign up for a presentation date on the first day of class. The presentation order will be assigned at random for each day.

Submission

When you are ready to submit, there are three things you will submit.

- ☐ Dataset Source Link
- ☐ Code (Rmd file only; No knitted PDF required, unlike the labs.)
- ☐ Report (PDF file)

You must submit your project documents on Gradescope before the deadline. If the submission fails, then email your files or link to one of the instructors at paulch2@illinois.edu or athanvi2@illinois.edu.

We can't wait to read your project and see your presentation! :)

Rubric

- **Paper Total: 150 Points**
  - **Intro:** 30 points
    - *Make sure to answer the 2 questions and introduce your report formally.*
  - **Methods:** 45 points
    - *Should be statistical and/or logical steps.*
  - **Results:** 45 points
    - *A detailed summary of the results.*
  - **Conclusion:** 30 points
    - *Make sure to answer the 2 questions and wrap your report formally.*

- **Presentation Total: 50 points**
  - **Length (within 5-10 mins):** 10 points
  - **Well-Prepared and Clear:** 10 points
  - **Visuals:** 10 points
  - **Content of Presentation:** 20 points

Penalties to Keep in Mind

- You will lose 25 points if you do not use double space or your font size is greater than 12.
- You will lose 50 points if the report is not 2 pages with at least 1 page of text.
- You will lose 75 points if your data set does not meet the criteria listed in the syllabus.
- You will receive 0 points if your dataset is not related to the city of Chicago.
- You will also receive 0 points if you do not turn in your code for the project.

If your report is poorly written, it is up to the instructor's discretion to take points off from the report side of the rubric, as deemed appropriate.