DS4A Colombia AI Training:
Projects Overview

Overview

As discussed in the Welcome Packet, a meaningful part of the Data Science For All Colombia program is the completion of applied, multi-week projects. These projects come in two types: a *Datathon project* and a *Final project*.

Each of these two projects is described in detail below. The Datathon project will be introduced in class on Day 1, and the final project will be introduced in class on Day 2.

You will work with your team and your TA to scope these projects. There is significant time allocated during each class session for you to make progress on your Datathon and Final projects.

Datathon Project

Basic Information: The Datathon Project is based on Correlation One's Datathon competition series, and is designed for you and your team to apply and showcase the data science and analysis skills you have accumulated throughout this program. Your team will be expected to write a report that details your execution of the entire data science process, from start to finish.

You will be given a particular topic of focus - e.g. US commercial airline travel, West Coast Airbnb rental activity - and several datasets related to this topic. You are asked to explore the provided data and based on this, pose a question which you believe would be interesting to answer. You and your team will then write a report detailing why you believe this question is important and how you went about answering it.

We have included the particular Datathon problem statement we will be using for this program separately.

Report Details: The report should incorporate the entire data science process in order to answer the proposed question. Specifically, this means it should contain the following:

- *Topic Question.* What is the question that your team set out to answer? How did you come about to choosing this question, and why is it an important question?
- *Data Wrangling & Cleaning Process.* Do you conduct proper quality control and handle common error types? How do you transform the datasets to better use them together? What sorts of feature engineering do you perform?

- *Exploratory Data Analysis.* What hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
- *Statistical Analysis & Modeling.* What assumptions and choices did your team make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built?
- *Results Interpretation & Conclusions.* What were your team's key findings, and what is their significance? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations? You should use summary statistics and visualizations to help explain your thoughts.

For each section and sub-section, your team must clearly indicate what the percentage split in effort was among the team members.

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **they MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if a report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

Additionally, your team should submit the following with the final report:

- **Code** that was used to generate your results
- **Source File** used to generate your report (e.g. for a PDF with math-type, equations, or symbols, you should include your raw LaTeX source file)
- **PowerPoint Presentation** which gives a high-level overview of your report. It should include the topic question, key insights, results, and any data visualizations that your team thinks is helpful to communicate and support your points

Your team will be expected to present your findings to your peers for 10 - 15 minutes on the second-to-last day of the program, with your PowerPoint presentation as support, so this component should be optimized for that purpose.

Other Responsibilities: Your team will be assigned to a TA. You will have ~15 minutes on select afternoons to meet with said TA to discuss your progress and any issues you have run into. The TA will help guide your team in the right direction based on this.

Your team is also expected to keep a Google Doc journal of your progress on the Datathon project, and write 1 - 2 paragraphs at the end of each day in this journal. Each journal entry should cover:

- What your team accomplished that day
- Any interesting discoveries you made that day and how it will impact what you do next
- Any blockers you currently have

TAs will review these at the end of each day and determine any obvious inefficiencies in your team's progress (e.g. if you are going down a rabbit hole, if you have a simple technical blocker the TA can clear up, etc.) that they can help with during office hours.

At the end of week 3, your team is expected to submit a project scoping plan to your assigned TA. The TA will review this proposal and request that your team make changes based on their judgment of feasibility and informativeness. This proposal should include:

- Your team's question you set out to answer
- The datasets your team is planning on using (including any externals)
- How you plan on using each of these datasets/how they contribute to answering the question

Milestones: There are no official milestones for the Datathon project, but the following is our recommended timeline. Your TA will keep your team roughly on this schedule (some deviation is fine) and assist you as necessary, particularly with the first step (choosing an appropriate question). Each week number corresponds to the end of the indicated week (with the exception of week 10 as teams need to present on the second-to-last day of week 10):

- Week 3
    - Appropriate topic question chosen
        - It should not be TOO ambitious, as it may be impossible to demonstrate in the span of the report
        - TAs will assist you here
    - Project scoping plan/proposal written
- Week 5
    - Datasets wrangled & cleaned, associated section written up
- Week 7:
    - EDA completed, associated section written up
- Week 9:
    - Analysis & modeling step completed, associated section written up
- Week 10:
    - Conclusions completed, report touch-ups completed

Final Project

Basic Information: The Final Project is meant for you to apply and showcase the data engineering and visualization skills you have developed throughout the program. Teams will be asked to scour available datasets from a number of pre-provided data repositories (the full list will be provided separately) and build a production-level application centered around a subset of those datasets. This application should either solve an existing problem or improve upon an existing process. It will be up to each team to identify a problem or area of improvement to focus on.

Your team will also be expected to write a report which details what your application does, how you chose to design it and why, and how you implemented each major piece.

Teams will be presenting their final projects to government officials and private sector representatives on the final day of the program. Therefore, at least some of the time will be dedicated to TAs mentoring the teams on how to present to this type of audience, which may well be non-technical and need to be convinced of the productive possibilities of data science.

Application Details: At the very minimum, each application MUST have the following three technical components (they are allowed to have more):

- **Interactive Front-End:** A non-technical user should be able to use this to get meaningful outputs and visualizations. The target audience is a government official or someone in the private sector - it should be clear to them how to get value out of this and what that value is.
- **AWS-hosted Database:** This should persist all relevant data, and supply it to the front-end.
- **AWS-hosted Data Analysis & Computation:** These should perform all computation on the data itself that is relevant to the application. They can be structured in a manner of the team's choice - microservices, chron jobs, scripts, etc. However, they must live in and run off of the AWS compute engine (NOT a local machine)

Report Details: Each report should have three major sections:

- **Introduction:** This should clearly state the context of your team's application and the problem you set out to solve, as well as your justification for why it is an important problem.

- **Application Overview:** This should cover what the application does, what the primary use cases are, and how a user would interact with it. Your team should also discuss which features you chose to include in the front-end, and why they are important.
- **Technical Exposition:** At minimum, this should contain at least three sub-sections, one for each of the three required components of the application (interactive front-end, AWS-hosted database, AWS-hosted data analysis & computation). Below are examples of questions that ought to be addressed in each of those three sub-sections:
    - Interactive Front-end:
        - What technologies did you use to set up the front-end?
        - What types of visualizations do you display and why?
        - How does this pass and receive information for use from the AWS-hosted components?
    - AWS-hosted Database:
        - What type of database did you use?
        - What are the main datasets you chose to include?
        - What are the main data tables that you set up?
        - How did you choose to design those tables, and why?
    - AWS-hosted Data Analysis & Computation:
        - What computation tool did you use?
        - What does each part of your data computation package do (e.g. run statistical tests, build a model, or something else)?
        - What was your data wrangling & cleaning process? Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform?
        - How did you incorporate models and/or statistical tests into the application? What was the motivation behind the particular ones used?
    - Additional topics that ought to be covered if relevant for your project:
        - Code/program design paradigms used
        - Flow charts/diagrams indicating how the different parts interact with each other

As with the Datathon project, **for each section AND sub-section, your team must clearly indicate what the percentage split in effort was among the team members.** Additionally, although not directly part of their writeup, **all of your code MUST be commented and submitted.**

Additional Requirements: Your team's application MUST work in real-time as it will be presented to government officials as well as private sector representatives. Your team will need to be polished when showing its work to these individuals. As such, you are expected to dedicate some time to putting together and rehearsing a presentation script which covers the following:

- What problem does your application try to solve? Why is this an important problem?
- On a high level, how does your application work? How do you use data science & engineering technologies and methods to do what it does?
- How does a non-technical user interact with and get value out of your application? Where do they put in their inputs and how do they receive the desired outputs? How should that user interpret and use those outputs?

You are expected to do dry-runs of this with the TAs, whom will be expected to give feedback and ensure that the presentations are at a level suitable to the intended audience.

Other Responsibilities: As with the Datathon project, team are expected to keep a Google Doc journal of their progress on the final project, and write 1 - 2 paragraphs at the end of each day in this journal (a single journal document should contain both the Datathon project and Final project notes). You will also be expected to submit a project scoping document at the end of week 3. Unlike the Datathon project, this scoping document will be more involved, and should include:

- Your team's problem you set out to solve (multiple versions, starting from V1 (minimal viable products), and work your way up to more ambitious versions)
- High-level overview of how your application plans on solving this problem
- Outline of who the primary users would be and how they would interact with the app
- Descriptions of app features
- The datasets your team is planning on using
- How you plan on using each of these datasets/how they contribute to your app

You can expect your TA to actively work with you on this as well as any issues you have identified in office hours and in your daily Google Doc journal.

Milestones: There are official milestones for this project because of the requirement that this be presented to government officials and private sector representatives on the last day of the program. As such, TAs are responsible for ensuring that teams are managing their time appropriately and hitting the below deadlines throughout the program (each week number corresponds to the end of the indicated week). TAs are STRONGLY encouraged to keep teams to these except in special circumstances:

- Week 3
    - Multiple problem versions chosen
        - All versions should naturally build on top of each other

        - V1 should be pretty easy and reasonable
        - The last version can be moonshot - the idea is that you must implement for V1 first, then V2, etc. so as to guarantee some sort of finished build by the end of Week 10. If you can get to V2, V3, etc. by the end, great, if not, at the very least you have something done that you can present)
    - Datasets sourced
    - Project scoping plan/proposal written
    - Introduction section of final report written
- Week 5
    - Front-end design finalized
    - Information channel to and from AWS components designed
- Week 6:
    - Datasets wrangled & cleaned
- Week 7:
    - Database tables designed
    - Finalized datasets loaded into AWS-hosted database
- Week 8:
    - Front-end build finalized
    - Link from front-end to AWS-hosted database established
- Week 9:
    - Analysis & modeling of datasets finished
        - Relevant code from this embedded into AWS compute engine
        - AWS compute engine integrated with database and front-end
- Week 10:
    - App final touches and quality assurance checks
    - Report finished
    - Presentation prepared