CT Academy | Exercises

Module 1 (Blue) – Data Foundations

1. Consider the following situation: you are away on business and you forgot to hand in a very important (and urgently required) architectural drawing to your supervisor before leaving. Your office will send an intern to pick it up in your living space. How would you explain to them, by phone, how to find the document? If the intern has previously been in your living space, if their living space is comparable to yours, or if your spouse is at home, the process may be sped up considerably, but with somebody for whom the space is new (or someone with a visual impairment, say), it is easy to see how things could get complicated. Time is of the essence – you and the intern need to get the job done correctly as quickly as possible. What is your strategy?
2. Write a data ethics statement for yourself as a data or a data adjacent professional.
3. Discuss how the data ethics best practices are applied in your organization.
4. Answer the University of Virginia's Centre for Data Ethics and Justice questions, as they apply to data used by your organization.
5. Translate the cognitive biases to analytical contexts. What cognitive biases are you, your team, and your organization most susceptible to? Least?
6. In pairs, identify times where you have had issues because of data availability, usability, consistency, integrity, quality, security, and/or trustworthiness.
7. In pairs, answer the following questions:
   i) does your group create or generate data? if so, what data?
   ii) do you use data from external sources? if so, which ones?
   iii) how many sources of data (e.g., databases) does your group use, roughly speaking?
   iv) do you publish analysis of data internally to your group? externally? both?
8. Where does your organization's data come from? Is it generated locally? Is it a sample or the full population? Was any of it collected from or provided by 3rd parties? Was any of it scraped online?
9. You must estimate the yearly salary of data scientists in Canada. Identify potential:
   i) populations (target, study, respondent, sampling frames);
   ii) samples (intended, achieved);
   iii) unit information (unit, response variate, population attribute);
   iv) sources of bias (coverage, nonresponse, sampling, measurement) and variability (sampling, measurement).
10. Write down examples of when you are acting in each of the following roles, if applicable: data consumer, data steward, data trustee, data contributor, and data custodian.
11. Review the provided example datasets and identify parts of the data that are associated with ethical risk (e.g., sharing of personal information). Identify potential areas of bias in the data (if any).
12. Identify 3 ways in which your organization collects data. For each way, identify potential issues with how the data is collected and how you manage those issues (if any issues exist).
13. Make a list of the top 3 data governance activities that you think have the biggest impact on your organization; for each, detail the reason(s) why it is important and the barriers to implementation.
14. In your own words, write out a list of guiding principles that could apply in the world of Government finance. Be specific and avoid generic words. For example, "Cost center managers are accountable for providing an accurate representation of their finances at the end of each quarter".

Module 2 (Green) – Data Analysis, Data Science, and Business Intelligence

1. Sketch a possible data quality process for a data quality problem that has been identified in your workplace.
2. What data role do you hold in your organization? Which role do you think you are currently best suited for? Which role do you aspire to?
3. Have you encountered the Analysis Cheat Sheet lessons in your work? Have you encountered others?
4. Find examples of recent "Data in the News" stories. Were they successes or failures? What social consequences could emerge from the technologies described in the stories?
5. Which of the quantitative skills presented in this section do you possess? Which interest you? Which do you plan on learning about?
6. Which of the software skills presented in this section do you possess? Which interest you? Which do you plan on learning about?
7. In what format is your organization's data available? Are you able to access it easily? Is it updated regularly? Are there data dictionaries? Have you read them?
8. Are the following examples of good questions? Are they vague or specific? What are the ranges of answers we could expect? How would you improve them?
   i) How does rain affect goal percentage at a soccer match?
   ii) Did the Toronto Maple Leafs beat the Edmonton Oilers?
   iii) Did you like watching the Tokyo Olympics?
   iv) What types of recovery drinks do hockey players drink?
   v) How many medals will Canada win at the Paris 2024 Olympics?
   vi) Should we fund the Canadian Basketball team more than the Canadian Hockey team?
9. What questions could you ask about the provided example datasets? Write the questions out as per the guidelines. Try to build "bad" questions as well, to get a handle on the difference.
10. Review the provided example datasets and identify any data quality issues you can. For each of the issues categorize them into one of the data quality dimensions.
11. For the data quality issues identified in the previous question, identify or create:
    i) the root cause(s);
    ii) short term corrective action(s), and
    iii) long term corrective action(s).

Module 3 (Red) – Data Analysis and Visual Storytelling

1. The file cities.txt contains population information about a country's cities. A city is classified as "small" if its population is below 75K, as "medium" if it falls between 75K and 1M, and as "large" otherwise. Locate and load the file into the workspace of your choice. How many cities are there? How many are there in each group? Display summary population statistics for the cities, both overall and by group.
2. Are there opportunities for computations like correlation, linear regression, and times series analysis in this dataset?
3. Are there opportunities for computations like correlation, linear regression, and times series analysis in your workplace datasets?
4. Are there opportunities for machine learning tasks/anomaly detection in your workplace datasets?
5. Turn the dataset found in the file cities.txt into a tidy dataset.
6. Does the dataset found in the file cities.txt appear to be of good quality (is it sound? does it have invalid entries?)
7. Create a list of items that could be used in a methodical data cleaning checklist. Use data that you have encountered in the past as inspiration (numerical, categorical, text data).
8. Find anomalous observations in the cities.txt and HR_2016_Census_simple.xlsx datasets (if applicable).
9. Find anomalous observations in a dataset of your choice.
10. Scale, discretize, and create new variables out of the cities.txt and HR_2016_Census_simple.xlsx datasets.
11. Scale, discretize, and create new variables out of a dataset of your choice.
12. Find examples of data presentations that you consider to be particularly insightful and/or powerful. Discuss their strengths/weaknesses.
13. Find examples of data presentations that you consider to be particularly misleading and/or useless. Discuss their strengths/weaknesses.
14. How do you think new technologies (e.g. virtual or augmented reality, 3D-printing, wearable computing) will influence data presentations?
15. In teams or individually, identify a few data visualizations that appeal to you. What is the story being told by the visualization? What kind of data is needed to build these visualizations?
16. In teams or individually, identify work scenarios for which data visualization could prove useful. What insight could be drawn from such visualizations? Would such visualizations get a buy-in from your supervisors/employers? How much work would be required to get from design to completion? Are the obstacles mostly of a technical nature? Related to data procurement?
17. Consider the dashboards on pp. 130-132. Can you figure out at a glance who their audience is? What are their types? Their strengths and limitations? How could you improve them?
18. Consider a data question of interest to you. Identify the target audience and the goals for your storytelling dashboard.
19. Identify the presentation requirements for your dashboard.
20. Create a storyboard for your dashboard.
21. What type of narrative and logic do you think would best serve your needs?

Module 4 (Yellow) – Decision-Making and Evaluation

1. With 26 seconds left in the Super Bowl, the Seattle Seahawks were trailing the New England Patriots by 4 points. At 2nd & Goal, the Seahawks had the ball at the Pats' 1 yard line. The common wisdom in this situation is to hand the ball to the running back and let them try to punch through the defensive line. The Seahawks had two options:
    a) Run the ball (1 play). Risk: Fails to score and time runs out.
    b) Throw the ball instead, then run if necessary (2 plays). Risk: 2% chance of interception.

    What play should the coach call? Why?
2. In the Vanity Fair article "You Could Fit All the Voters Who Cost Clinton the Election in a Mid-Size Football Stadium", Tina Nguyen writes:

    "While nearly 138 million Americans voted in the presidential election, the stunning electoral victory of Donald Trump came down to upsets in just a handful of states that Hillary Clinton was expected to win. It has been cold comfort for Democrats that Clinton won the popular vote—at the last count, she was up by about 2.5 million votes, and climbing, as ballots continue to be counted. Even more distressing is the tiny margin by which Clinton lost Wisconsin, Michigan, and Pennsylvania—three states that were supposed to be her firewall in the Rust Belt, but that ultimately tipped the electoral college map decisively in Trump's favor.

    Trump's margin of victory in those three states? Just 79,316 votes.

    This latest number comes from Decision Desk's final tally of Pennsylvania's votes, where Trump won 2,961,875 votes to Clinton's 2,915,440, a difference of 46,435 votes. Add that to the official results out of Wisconsin, where Clinton lost by 22,177 votes, and Michigan, which she lost by 10,704 votes, and there you have it: 0.057 percent of total voters cost Clinton the presidency.

    It is not entirely unusual for the electoral college to be lost by such a slim margin. In 2000, Al Gore lost Florida (and therefore the election) by 1,754 votes, triggering a painfully drawn out recount drama that only ended with a Supreme Court ruling. And in 2004, John Kerry lost to George W. Bush by losing Ohio by a little over 118,000 votes. But it is worth considering just how few voters ultimately set the country on its current, arguably terrifying course. The 79,316 people who voted for Trump in Wisconsin, Michigan, and Pennsylvania—all states that Democrats carried since 1992—is less than the entire student body of Penn State (97,494 students), or only slightly more than the number of people who attended Desert Trip, the Baby Boomer-friendly music festival colloquially known as "Oldchella." If you put all these voters in the Rose Bowl, there would be slightly over 13,000 seats left over.

    There are more people living in Nampa, Idaho, a city you have never heard of.

    To put things in even more painful perspective, Green Party candidate Jill Stein won about 130,000 votes in those three states. Libertarian candidate Gary Johnson won about 422,000.

    But perhaps the most painful data point for Clinton is this: the Democratic nominee for president never made a single campaign stop during the general election, and largely neglected Pennsylvania and Michigan, too, while Trump canvassed all three states relentlessly. His furious, last-minute blitz throughout the Rust Belt to win white, working-class voters, combined with the lack of resources Clinton invested, essentially handed their combined 46 electoral votes to Trump. Instead, Clinton spent the last few weeks of her campaign expending resources in places like Arizona and Texas—states which went for Trump by huge margins."

    So was it bad luck, or a mistake? Why?
3. Revisit the last two questions in light of the Luck and Information slide.

4. A port-mortem is good for learning the causes of a bad outcome, with one tiny limitation: the patient is already dead. For a **pre-mortem**, we imagine ourselves at some time in the future, having **failed to achieve a goal**, and looking back at how we arrived at that destination – it is an autopsy **before** the patient dies. With **backcasting**, we instead imagine that things worked out. In general, we conduct a pre-mortem/backcasting exercise by first identifying the goal to achieve, or the decision being considered, then picking a timeline for achieving that goal, and finally imagining that it is the day after the deadline, at which point we are looking back at the process. We then try to give 5 reasons "within our control" and 5 reasons "outside of our control" for why things failed (pre-mortem) or for why they succeeded (backcasting). After these exercises, we might modify our decision based on the new insights, increasing the chance of good things happening and reducing the chance of bad things happening; in effect, we are looking for ways to mitigate the impact of bad luck (flood insurance works along those lines). Conduct premortem/backcasting for a new youth mental health initiative. Assume that your department has created an app which aims to improve the mental health of Canadian teenagers. It is now two years from today and you are looking back on the app's launch.
5. Are the following arguments strong? If they are weak, what are their flaws? Could they be improved?
    i) COVID vaccinations lead to increased hospitalizations as half of the hospitalizations were vaccinated.
    ii) Turning the Large Hadron Collider on was a mistake because either it destroys the Earth or it does not; a 50% chance is way too risky.
    iii) We know that the Earth is not flat because none of the other planets we know are flat.
    iv) You should not vote in the next election because one vote never makes a difference.
    v) The solution to reduce congestion is to reduce the number of lanes because with fewer lanes, people will seek alternative modes of transportation.
    vi) Airport security measures are proportionate to the risk because it's ok to wait a few hours if it means that my plane won't be hijacked.
6. Consider the items found in a briefing note relating to building a pipeline through caribou territory:
    i) The last 7 times pipelines were constructed in caribou territories, populations decreased in the territory.
    ii) Biologists created a map showing the caribou migration paths. Based on this map, we conclude that placing the pipeline over the territory will not interfere with caribou migration.
    iii) Pipelines have not affected geese populations; as they and caribous are both social animals, the pipeline will not affect the caribou population.
    iv) Biologists have shown that caribous are not scared of large objects. If caribous are not scared, their breeding habits will not be affected. As pipelines are large objects, constructing this pipeline will not affect the breeding habits of the caribous on the territory.

    Identify the reasoning strategies being used in each of these arguments. Applying a plausible reasoning lens to this, what would you conclude? What additional information would you need/want, before drawing a conclusion?
7. For one of your programs, identify outputs, outcomes, and impacts (as in the logic model). You may consider the following: as a finance professional, what things are you able to measure and what data would you need to use to measure them? Is financial data part of an outcome or is it always an input into the program itself?
8. Define evaluations for each step in a logic model for one of the initiatives on which you are working: define a methodology (e.g., it will take X amount of time to complete an activity), measurements (e.g., compare actual $ spent to budget), and an evaluation criteria (e.g., complete in schedule is "effective", up to 1 month over is "partially effective" and more than 1 month over is "not effective").