Back to top

TABLE OF CONTENTS

1. Overview of the hackathon.
2. Overview of rules
3. Grading Rubrics
4. Examples of good reasoning
    a. Data Cleaning
    b. Cat2: Descriptive Stats
    c. Cat3: Prescriptive Stats
    d. Cat4: Predictive Analysis

Back to top

THE PYTHON HACKATHON
WHAT IT IS AND HOW YOU CAN PARTICIPATE

Like all hackathons, the purpose is to help you get better at a specific skill needed to be a DA. In this case Python.
However, being good at just writing Python is no longer a standalone skill that is in demand.
With AI taking over, writing code is no longer a barrier.
Instead, reading code and knowing what code to write, is. Which is the goal for this hackathon.
We will be testing your ability to:

1. Ask good questions.
2. Understand what questions need to be asked to provide value.
3. Present the value and meaning of the data, not just the data itself.

You will no longer receive pre-written questions and just be asked to answer them. **Instead, you will be writing questions and answers that help with insights. The analysis and insights you decide to show are completely up to you. Organizers will not give you a goal/insight to work towards.** You will be the analyst that decides what insights have value and how to show them.
With that in mind, here are the rules and grading metrics

ASKING GOOD QUESTIONS: WHAT THAT MEANS AND HOW WE CAN JUDGE THE VALUE OF THOSE QUESTIONS.

To help us evaluate your abilities fairly, we have identified 5 types of questions that qualify for the hackathon. A look at each of these, along with rules for evaluation, is given below. (Jump to examples)

Category	Requirements	Rules
Category 1: Data Preprocessing / Cleaning	• Clearly explain cleaning steps with reasoning Explain why a certain process/method was needed Explain how it helps your analysis Explain how analysis is affected if you do not do this step	• No limit on steps • 100 points available (maximum)
Category 2: Descriptive Analysis	• Each question should include: - Choose a good column to base the analysis on - Explain why the marker was chosen and why it's important to the medical condition at hand - Write code that achieves this objective	• 10 questions allowed(maximum) • 50 points available (maximum) • 5 pts per question

Back to top

Category 3: Prescriptive Analysis	• For each question: - Explain why a certain group of markers is chosen and how it influences the medical condition - Show strong correlation within the data given as supporting evidence (results or charts) - Write effective code that achieves the above objective	• 30 questions allowed (maximum) • 600 points available (maximum)
Category 4: Predictive Analysis	• For each hypothesis: - Set up an accurate hypothesis that has meaning and shows strong understanding of the subject and statistics - Choose the correct markers - Write code	• Unlimited questions allowed. • 90 points per well-explained hypothesis +model
Category 5: Demo of insights using Python	• Must be in the form of a dashboard or predictive learning tool Only Python or python-based tools can be used (plotly, streamlit, flask etc) Must have a mix of all the above components and must solve a problem	• 100 points available. If the following conditions are met - Explain what problems the dashboard solves and how it can be used - Explain the technical aspects, especially how you decided on the idea and how you overcame any technical obstacles - Innovative design and layout

GRADING RUBRICS: HOW TO SCORE MAXIMUM POINTS IN EACH CATEGORY

Category	Grading Rubrics
Data Preprocessing / Cleaning	**100 Points:** Demonstrates a clear understanding of what data preprocessing is, specifically w.r.t to analysis. All steps outlined have purpose. No steps have been added without understanding (for example switching blanks to nulls without specifically explaining what outcomes are expected in this dataset from making that change). **90-99:** Demonstrates fairly clear understanding of data cleaning with some exceptions. Missed or added no more than 1 cleaning step that was unnecessary/erroneous.

Back to top

	70-89: Good understanding of cleaning but needs improvement. Lots of steps added without understanding. **50-69:** Passable understanding of data cleaning. **30-49:** Below average understanding of data cleaning. 1-2 steps are correct, but more steps are incorrect than correct. **11-29:** Code and some steps show technical expertise, understanding of data cleaning is poor. **0-10:** No understanding of the process w.r.t analysis. Most of the processing recommended is inaccurate.
Descriptive Analysis	**5 points:** Column chosen was a valuable addition that helped improve understanding of the data or set the stage for telling the story of the data. (For example mean and stdev of glucose values in a diabetes dataset can help with an understanding of what the baseline population is) **4 points:** Code is accurate and gets all points. But either the column chosen does not improve the analysis in the ways stated above (or) the explanation for why the marker was chosen is incomplete(for example it's from google) **1-3 points:** code is accurate but the other 2 explanations do not clearly tell us why a marker was chosen and how it is important. **0 points:** Code exists, but explanation is incomplete (points for code will only be given if it accompanies a good analysis)
Prescriptive Analysys	**4 points for code to be awarded in full if the other 2 items on the list are complete. If not, 0 points for code without an explanation.** **16 points:** Demonstrates a clear understanding of the subject and how the marker is related to the subject. Provided excellent evidence of either correlation or causation to the subject from within the dataset. Evidence is in the form of graphs/tables/a query result or a clear correlation matrix. **12-15 points:** Demonstrates some understanding of the subject and provides some evidence of correlation. But either correlation is weak (or) its from google and does not clearly show a connection to the dataset (or) the explanation of why they are correlated is not clear. 8-11 points: Below average explanation of marker. Some, incomplete evidence provided. **1-7 points:** Either the explanation is completely missing but the evidence is partially available (or) the evidence is missing but there is a partial explanation and a partial understanding of the subject. 0 points: Insufficient evidence and explanations given. Selection of variables seem random and only help get the team closer to 30 questions

Back to top

Predictive Analysis	**20 points for code to be awarded in full if the other 2 items on the list are complete. If not, code gets the same 'grade' as the model. For example if the model was evaluated at 50/70 points, thats 71%, code will also receive 71% which is 14 points.** **70 points:** Clearly demonstrated the need for a predictive model. Chose parameters that help with prediction and set up the hypothesis accurately. The predictive model clearly improves the analysis/solves a problem and shows a strong knowledge of the subject. **60-69 points:** The model does make a prediction, but the need for the prediction is not clearly demonstrated (or) the parameters chosen are good, but there are flaws in the understanding of the subject. **40-59 points:** The model makes a prediction that somewhat improves the analysis/overall solution. Parameters are accurate, prediction accuracy is high, but the model was mainly created for points and has little to no real purpose. **20-39 points:** Low understanding of the subject, some parameters were chosen accurately, most were not. **1-19 points**: Inaccurate subject knowledge, mostly inaccurate parameters. Either by luck or partial understanding, some parameters were correctly identified. **0 points:** No understanding of the subject. Model serves no logical purpose and was built for the sake of points.
Demo of final product in Python	**30 points for design and layout at the discretion of the evaluator.** **70 points:** Dashboard is easy to read and intelligent. I.e: it goes beyond just summarizing data and actually helps provide valuable insight. It answers the question of why business should care about the data and how the data affects healthcare. It clearly shows what problems it solves and is useful to read and make decisions from. **40-69 points:** Some intelligence and some good insights are available, but the dashboard is mostly just a summary of the data with no specific goal and/or very few problem-solving abilities. **30-39 points:** The dashboard mainly serves as a descriptive tool with no specific purpose. It has 1-2 insights but no specific explanation of why it matters or what problems are solved. **15-29- points:** This is a dashboard mainly for the sake of points. It may have 1 insight but most statistics on it are purely technical or descriptive. No specific problems or solutions to problems were mentioned. **0-14 points:** This dashboard was created just for points, team demonstrates little to no understanding of the subject.

Back to top

Presentation	**45–50 points:** exceptional presentation. team strategically selects its strongest insights and communicates them clearly within the time limit. every team member contributes. the findings are compelling and strongly tied to the problem. professional delivery. questions during the Q&A are answered confidently and accurately. **40–44 points:** strong presentation. good insight selection, mostly clear communication, and within the time limit. most team members participate effectively. good delivery with minor timing issues. Q&A responses are mostly accurate and complete. **30–39 points:** adequate presentation. includes basic insights but may come across as rushed or lacking depth due to inconsistent delivery. while all team members are present, individual contributions may vary. key findings might be loosely tied to the problem. minor timing issues or unclear communication may affect flow. Q&A responses tend to be surface-level or partially addressed. **20–29 points:** below average. weak insight selection or poor delivery. team may exceed time limit or struggle to present clearly. uneven or minimal team participation. significance of findings is unclear. time not managed well. Q&A responses are uncertain or off-topic. **0–19 points:** poor presentation. major time overruns, minimal team contribution, and unclear or irrelevant insights. delivery lacks structure or professionalism. Q&A reveals limited understanding of content.
GIT	**135–150 Points: Excellent.** Regular commits with clear messages. All team members contribute notebooks. Clean, well-documented Jupyter notebooks. Well-organized repository with detailed README. **120–134 Points: Good.** Consistent commits tracking hackathon progress. Most team members contribute regularly. Notebooks are mostly clean with some explanations. Basic repository documentation. **90–119 Points: Adequate.** Basic commits but inconsistent. All team members have some contributions. Notebooks may be messy or lack explanations. Repository is functional but lacks documentation. **60–89 Points: Below Average.** Few commits or unclear messages. Uneven team participation. Poor notebook organization. Missing documentation. **0–59 Points: Poor.** Minimal commits. Missing team contributions. Disorganized notebooks and repository. Major collaboration issues with notebook conflicts.
Submission	**50 points:** On time, no subfolders **35 points:** Late submission (-15 points) **40 points:** Has subfolders (-10 points) **25 points:** Late + subfolders (-25 points) **0 points:** No usage instructions OR missing files

Back to top

EXAMPLES OF GOOD REASONING

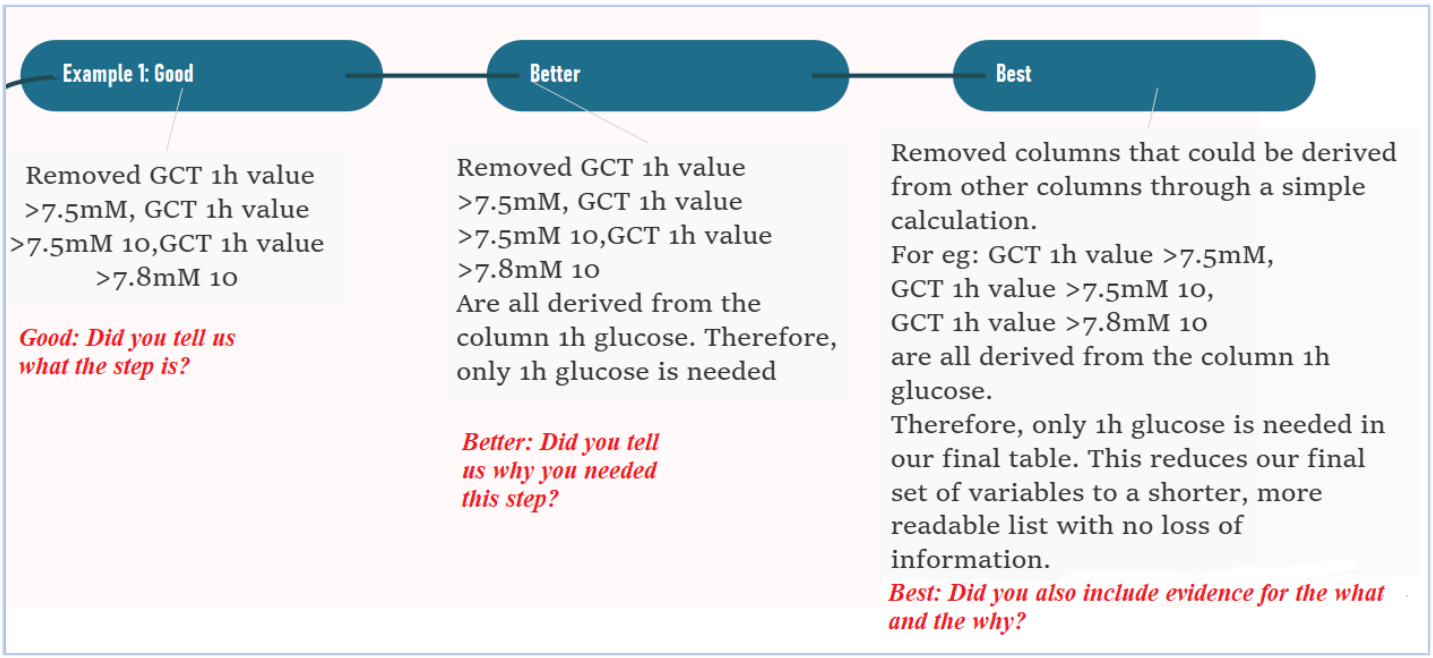
CATEGORY 1: DATA CLEANING

Rules and Requirements

- 100 points(max)
- Clearly explain cleaning steps with reasoning
- Explain why a certain process/method was needed
- Explain how it helps your analysis
- Explain how analysis is affected if you do not do this step

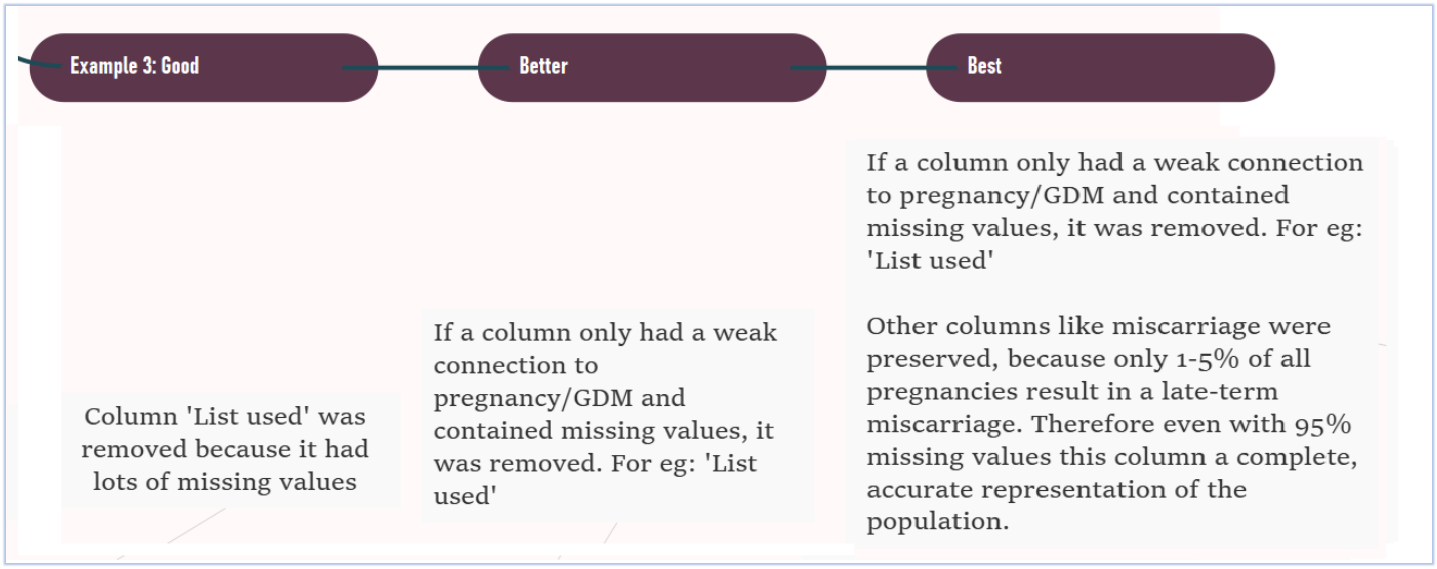
You get the most points for the **explanation behind why a cleaning step was created** along with evidence (where possible). Here are some examples of reasoning, ranging from what will get you minimum points to what will get you the maximum.

Example 1

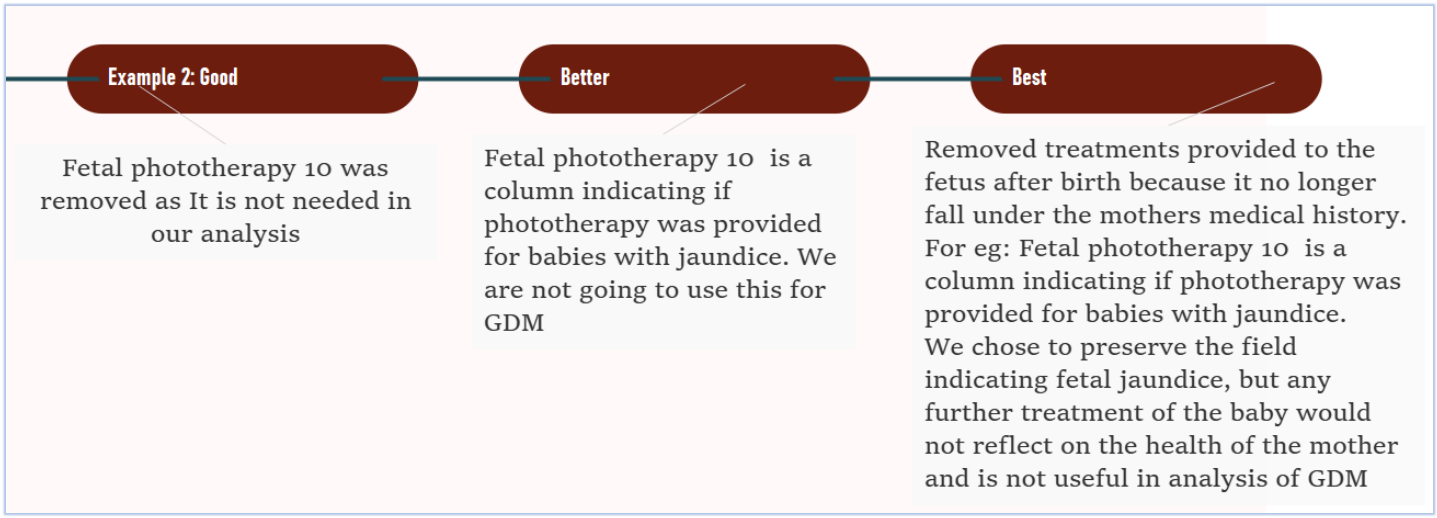


Back to top

Example 2



Example 3



Back to top

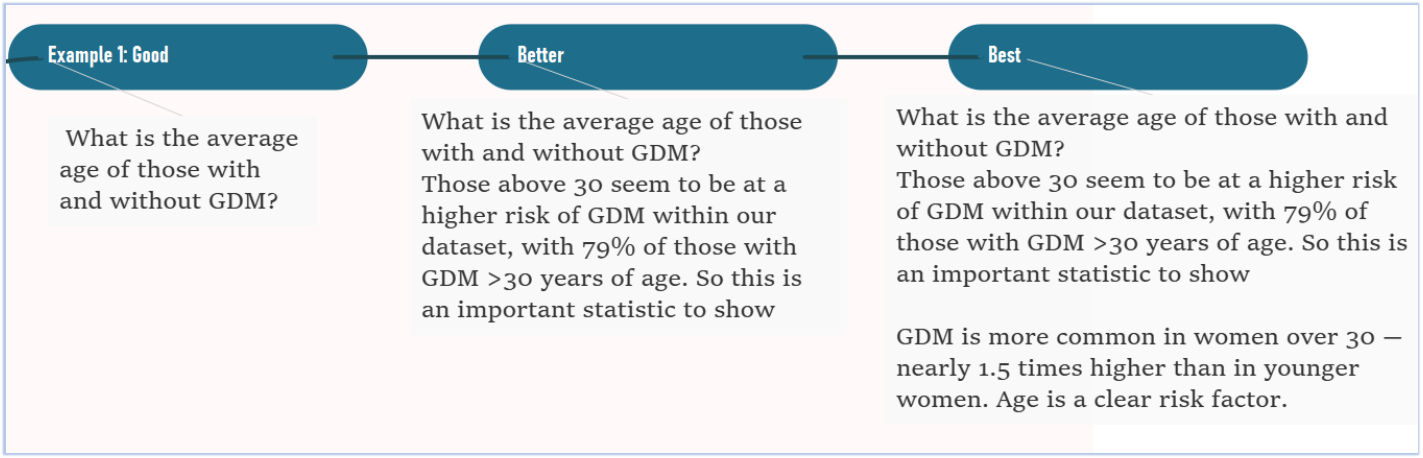
CATEGORY 2: DESCRIPTIVE STATISTICS

Rules and Requirements

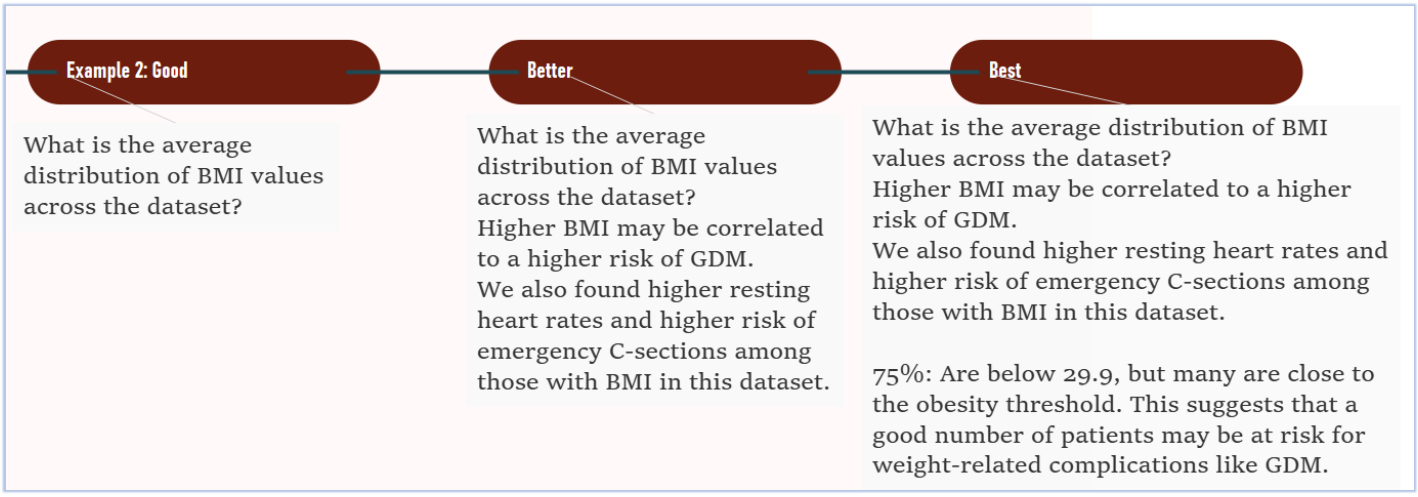
- Maximum of 10(50 total points) questions allowed.
- Typically includes counts, distribution charts, frequency/outlier related charts, averages and variance.
- Each question should include:
  - Choose a good column to base the analysis on
  - Explain why the marker was chosen and why it's important to the medical condition at hand
  - Write code that achieves this objective

You get the most points for the **choosing good columns that describe the data well,** along with evidence (where possible). Here are some examples of reasoning, ranging from what will get you minimum points to what will get you the maximum.

Example 1:

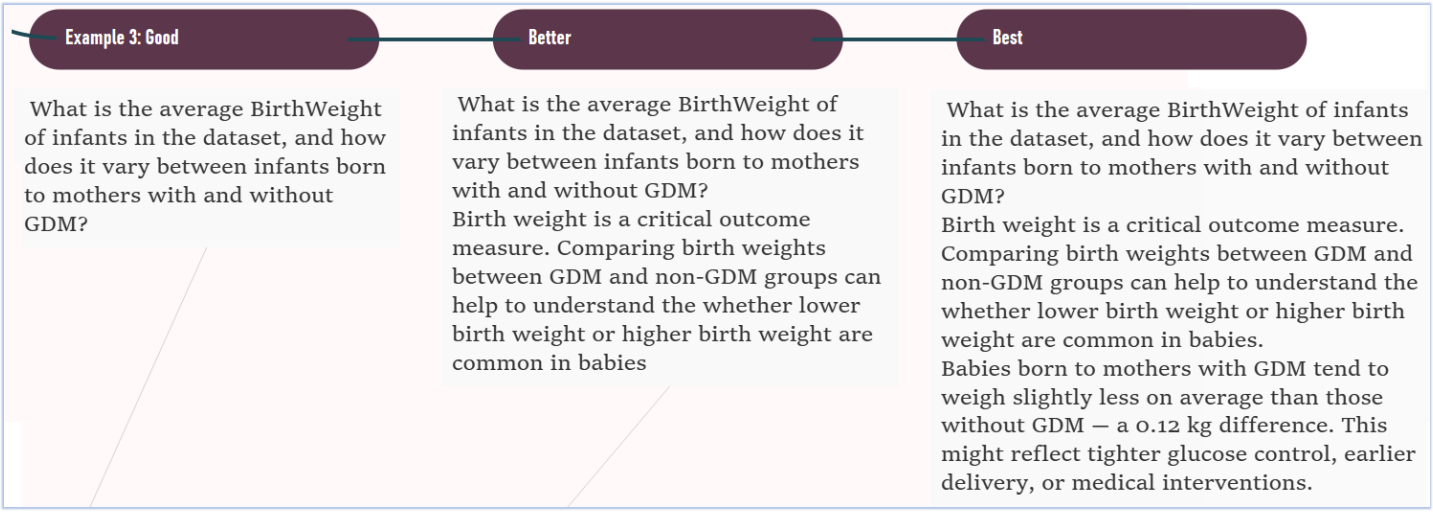


Example 2:

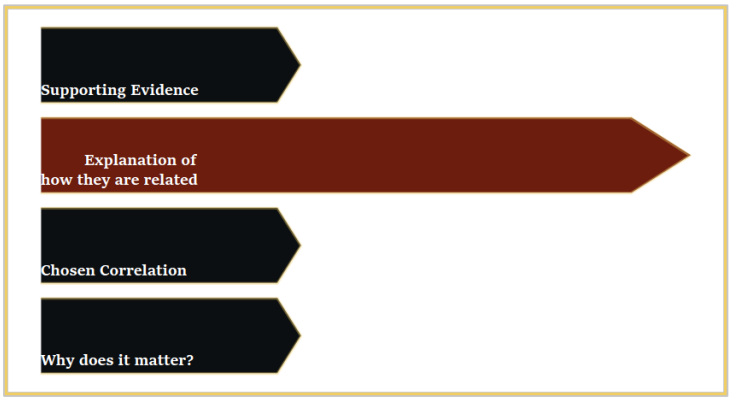


Back to top

Example 3:



CATEGORY 3: PRESCRIPTIVE/MUTIVARIATE ANALYSIS



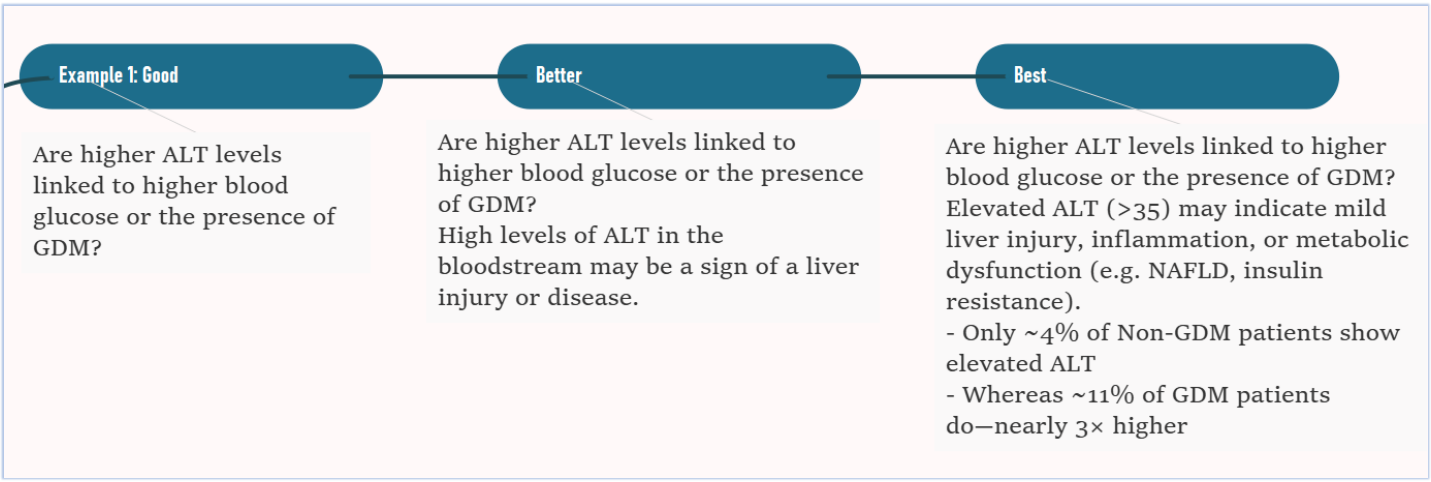
Rules and Requirements

- Maximum 30 (600 points total) questions allowed
- Typically involves correlating 2 markers and providing insights about how they are correlated and why.
- For each question:
  - Explain why a certain group of markers is chosen and how it influences the medical condition
  - Show strong correlation within the data given as supporting evidence (results or charts)
  - Write effective code that achieves the above objective

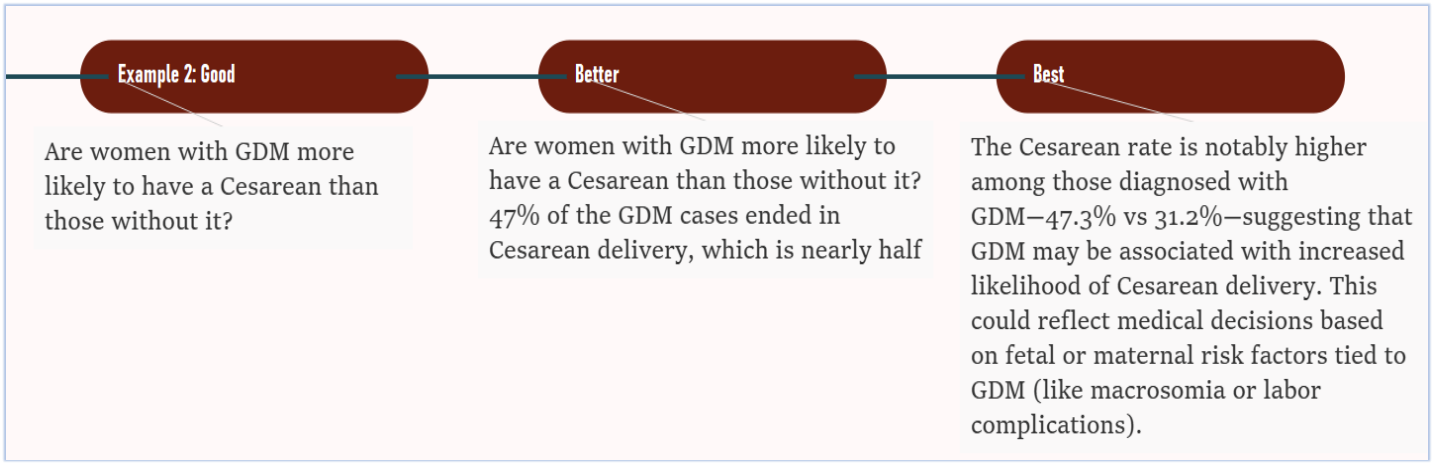
You get the most points for choosing a good **question**, providing a good **reason for the question** and following up with excellent **evidence** through your **python code**.

Back to top

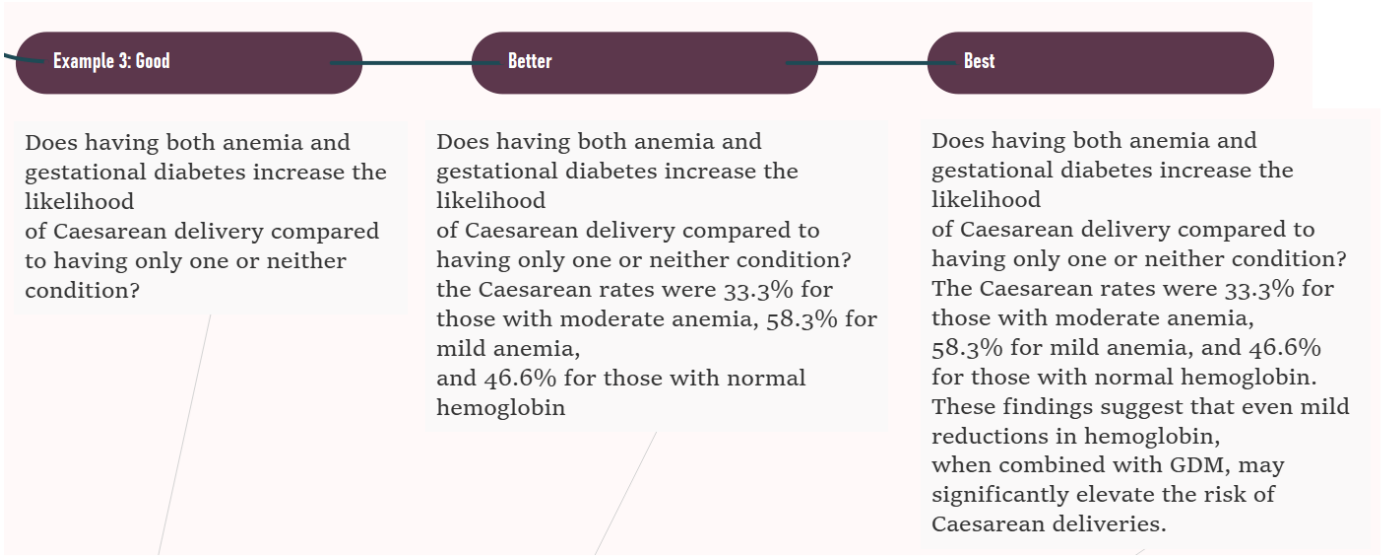
Example 1



Example 2

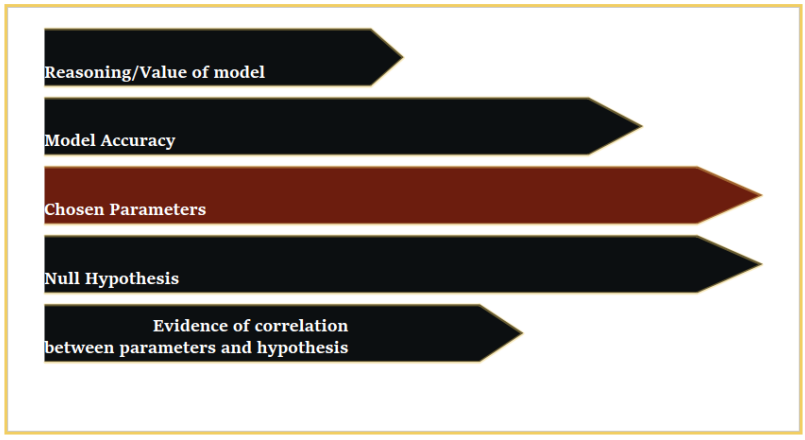


Example 3



Back to top

CATEGORY 4: PREDICTIVE ANALYSIS



Rules and Requirements

- Unlimited questions allowed (90 points each)
- For each hypothesis:
  - Set up an accurate hypothesis that has meaning and shows strong understanding of the subject and statistics
  - Choose the correct markers
  - Write code

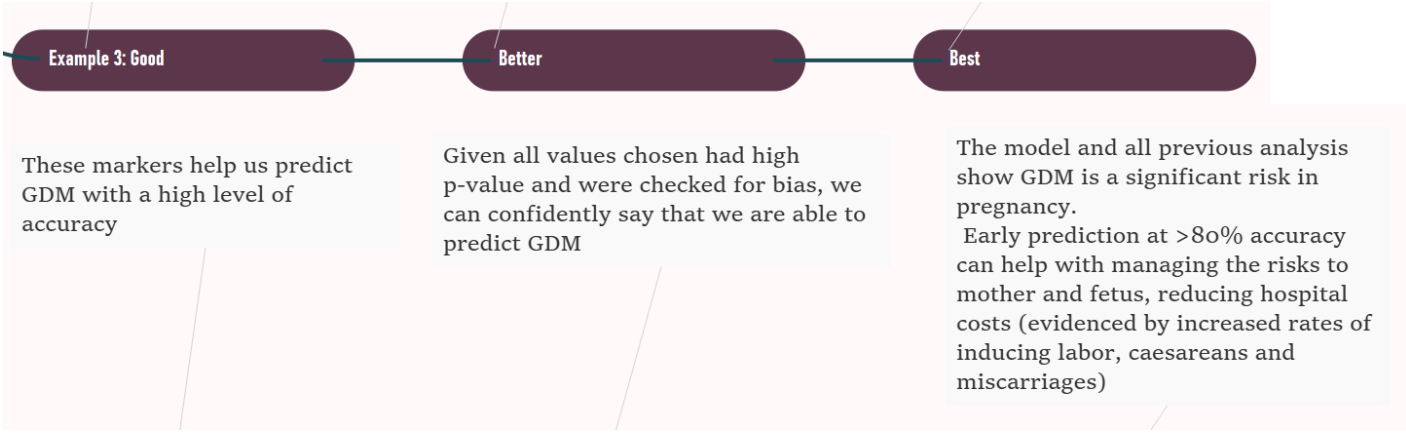
You get the most points for explaining why this dataset needed a model. What problems the model solves and following up with a solid understanding of the dataset, statistics and model choice.

Example 1

Hypothesis: Good	Better	Best
Early prediction of risk of GDM	Null hypothesis- GDM is not a significant risk factor in pregnancy or pregnancy related complications	Null hypothesis- GDM is not a significant risk factor in pregnancy or pregnancy related complications In testing out this hypothesis we hope to find markers of health that act as early indicators of risk of GDM. For example, hypercalcemia seems to show a strong correlation to the occurance of GDM, but in combination with other markers, it may help doctors order lab reports and intervene early.

Example 2

Variables Chosen: Good	Better	Best
Pulse rate, CRP,White Cell Count, BMI, and Platelet, white cell count show higher correlation.	-Pulse rate, CRP,White Cell Count, BMI, and Platelet, White Cell Count show higher correlation and low p-values, which means they are consistently elevated in women diagnosed with GDM. These could be strong early indicators.	- Based on the mann whitney u test,the top 5 p-values with the highest rate of confidence was chosen, they are: -Pulse rate, CRP,White Cell Count, BMI, and Platelet, White Cell Count show higher correlation and low p-values, which means they are consistently elevated in women diagnosed with GDM. These could be strong early indicators.

Back to top

Example 3