Note: This is meant to be a brief examination of your skills. You can select any number of the 6 questions below as long as the point total adds to at least 40.

1 The "Sudo" Code Problem: 20

The joke is intentional...

1.1 Iterator Problem

Write a program that scrolls through the numbers 0 to 1000 in increments of two. If the number is evenly divisible by 8 your program should let the user know.

On the other hand, we don't like things that are divisible by 12 (12 is an unlucky number in advertising) so your program should **never** pass these indicies. In other words any $i \ \ s.t. \ \ i \equiv 0 \, (mod \, 12)$ should never be explicitly checked.

1.2 Search Function

Consider another set of data, a matrix D of size 1000 by 30. The previous bit of code has given you important row indicies from this matrix. Column 30 in this matrix is a vector of values $v \in (0, 1)$. The goal now is the write a program that allows us to predict the value of v for a row (r_i) identified by our program in section 1.

We first want to prototype some code and strategies:

1. Write a function to subset the matrix D to the required rows.
2. Generate a random value for column 30 for each row you slice.
3. Write a function to sort the resulting rows by this random value.

1.3 Complexity

Estimate the complexity of your program in sections 1.1 and 1.2

2 Quadtrees: 40

Quadtrees are a form of simplified clustering algorithm. Given a set of features $S \in \mathbb{R}^2$ we seek to partition the space into segments such that no more than k points are members of any single segment. In quad trees this is done by recursively splitting the space into quadrants until this condition is met.

Write a program that allows a user to specify the number of points to be separated n and the threshold k. Upon completion, your program should plot the space, segments, and points, it should report the total number of segments as well as the area of the space with the most densely packed sub-segments (this definition is entirely up to you).

The two features used in your program ($s_1, s_2 \in S$) can be randomly generated within. You may use the programming language if your choice to solve this problem. Please provide your code so we can test your function. Also please provide an estimate of your function's complexity.

3 Hyperspace: 40

We are given a massive set of measurement data from our client. Specifically, it consists of n instances of $m \times m \times 3$ tensors. We need to effectively process this data and, given a single observation, o_i from the set of observations $\{o_1, ..., o_n\}$ we need to estimate the probability of an event $e \in \{1, 2, 3\}$

3.1 Space $\rightarrow$ Space

We first decide to transform the input data from $m \times m \times 3$ to $m \times m \times 1$.

Describe at least two methods for accomplishing this transformation. Please be specific, write (to the best of your ability) any relevant mathematical formulations and highlight the pros and cons of each method.

3.1.1 Computer time?

What are the computation constraints, if any, of the methods you described above? How might we implement them in practice?

3.2 Space $\rightarrow \rightarrow$ Space

Suppose that we discover a low variation in point values for localized sets within the last layer dimension of the input tensor. Because of this we want to change our transformation

strategy *s.t.* we take $m \times m \times 3$ to $p \times p \times 1$, where $p < m$.

Describe a method for accomplishing this transformation. Please be specific and write (to the best of your ability) any relevant mathematical formulations. If you describe multiple methods (not required) please highlight the pros and cons of each.

3.2.1 Computer time?

What are the computation constraints, if any, of the method you described above? How might we implement it in practice?

3.3 $\mathbb{P}$

How might we adjust our transformation approaches when it comes time to estimate, $\mathbb{P}(e|o_i)$? What are some ways we can estimate this probability?

4 Business Intelligence Output: 20

One of our financial services clients is looking for help in creating an analytical dashboard to identify demand for certain financial products throughout the year. The client has sent us a few different data schemas and would like our help designing the process to merge the information to come up with metrics to fill the dashboard.

Please use the schemas provided in the document FinancialDashboard.xlsx and document your logic for merging these data sources to complete the following tasks:

1. Document the data ingestion process. Please include any code or logic that would be used to define the included data sources
2. Produce the code or logic that uses the provided tables to create the output table. Please include documentation of your process and assumptions

Field Name	Type	Description
Month	string	"01" to "12"
ZipCode	string	5 digit ZipCode
State	string	Full State Name
CreditCardIndex	integer	Index value (Mean at 100)
SavingsAccountIndex	integer	Index value (Mean at 100)
MortgageIndex	integer	Index value (Mean at 100)
EmployeeHours	integer	Total Number of Employee Hours

5 Predicting Flight Delays 20

Our client is a large travel agency that has asked us to build a model to predict whether a flight will be canceled based on several factors. Additionally, our client can only sell tickets for three airlines (AA,UA, and DL) and would like to be able to advise their customers on which airline has the least risk of cancellation.

Use the provided data to build a model that predicts whether a flight will be canceled and complete the tasks below.

1. Provide fully commented code and model output for your analysis
2. Include your own function that uses the model output to predict whether a flight will be canceled. Make sure to document your approach

Field Name	Type	Description
Canceled	Binary	Canceled=1
Month	Integer	1=Jan
DepartureTime	Integer	Military Time (1:00 PM = 1300)
UniqueCarrier	String	Airline Carrier Code
SchedElapsedTime	Integer	Scheduled Flight time in minutes
ArrDelay	Integer	Arrival delay in minutes
DepDelay	Integer	Departure delay in minutes
Distance	Integer	Distance in miles

6 Retention Games: 20

A daily deals eCommerce company recently asked our team to predict which customers are likely to cancel their subscription. By identifying these customers in advance, the client would be better positioned to intercept potential defectors with retention offers and to forecast financial performance more accurately.

The client provided three years of customer-level data, such as purchase transactions and email offer responses. After conducting extensive exploratory analyses, our team built a logistic regression model and performed a decile-based lift analysis to assess how efficient our model is at identifying the most likely defectors. The final task is to share the results with our client.

Note: The model output has been provided in the appendix materials to this case document. You'll find the relevant tables in the PDF titled *Case_question_reporting.pdf*

6.1 For the CMO

Using the model output and results from the lift analysis below, please develop a brief presentation that summarizes the business impact of these results in a clear, digestible and visually appealing way. Be sure to draw relevant conclusions and to recommend how the our client might implement these results via a retention program.

6.2 For the Stats Guys

We found out at the last minute that the client's statistical modeling team will also be in the presentation. They requested that we provide a brief technical summary as part of our presentation. Add one slide that summaries our statistical results (for a technical audience) and another that proposes how we might implement a scoring solution in production.