**Project: Answer the Following Interview Questions (6 total)**

1. Describe a data project you worked on recently.

A. As part of data analyst Nano degree at Udacity, I worked on OpenStreetMap data. OpenStreetMap is open map data of the world. It maintains data about roads, railways, cafes, trails, and much more. This case study focuses on the wrangling map area **San Jose, CA, United States**. The steps followed are 1) Auditing the map data, 2) Cleaning data, 3) Loading data into database, 4) Querying the database.

Initially, raw data is transformed into a format that can be analyzed. During auditing phase, I have encountered some problems and cleaned them programmatically using python. Later, SQL is used to load the data into database and analyzed it to see whether data is complete. The experience I had in this project is very relevant to your requirements. I can apply SQL and python skills to maintain your electronic health record system.

I have experience with data visualizations tools like Matplotlib, Ggplot and Tableau. With your health related database, I can create dashboards and other reports. These visualizations are very useful to get complete picture of your data. At the same time, it will be easy to get an idea of what further changes can be made to improve data quality. As part of my work in map data analysis, I created a report using these tools to showcase my finding and the primary finding is that the data is incomplete and much more data has to be collected in San Jose region and also included some ideas of what further steps can be taken to improve data quality.

1. You are given a **ten piece** box of chocolate truffles. You know based on the label that six of the pieces have an orange cream filling and four of the pieces have a coconut filling. If you were to eat four pieces in a row, what is the probability that the **first two** pieces you eat have an orange cream filling and the **last two** have a coconut filling?
2. P (1st orange) = P(1) = 6/10

P (2nd orange) = P(2) = 5/9

P (3rd coconut) = P(3) = 4/8

P (4th coconut) = P(4) = 3/7

Total Probability = P(1)\*P(2)\*P(3)\*P(4)=7.14%

*Follow-up question:* If you were given an identical box of chocolates and again eat four pieces in a row, what is the probability that exactly **two** contain coconut filling?

* + 1. There are total of 6 possible combinations that can have exactly two coconut fillings

1. Orange, Orange, Coconut, Coconut
2. Coconut, Coconut, Orange, Orange
3. Coconut, Orange, Coconut, Orange
4. Orange, Coconut, Orange, Coconut
5. Orange, Coconut, Coconut, Orange
6. Coconut, Orange, Orange, Coconut

In the above question, we found probability for one such combination. Therefore the probability for all 6 = 6\*(6/10)\*(5/9)\*(4/8)\*(3/7) = 6/14= 42.86%

1. Given the table users:

Table "users" +-------------+-----------+ | Column | Type | +-------------+-----------+ | id | integer | | username | character | | email | character | | city | character | | state | character | | zip | integer | | active | boolean | +-------------+-----------+

construct a query to find the top 5 states with the highest number of active users. Include the number for each state in the query result. Example result:

+------------+------------------+ | state | num\_active\_users | +------------+------------------+ | New Mexico | 502 | | Alabama | 495 | | California | 300 | | Maine | 201 | | Texas | 189 | +------------+------------------+

* + 1. SELECT state, SUM(active) as num\_active\_users

FROM users

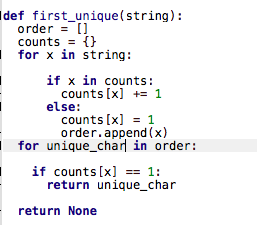
GROUP BY state

ORDER BY SUM(active) DESC

LIMIT 5

Here GROUP BY is used to group the users data by state and for each state SUM(active) will return the total sum. ORDER BY helps to arrange the column num\_active\_users in descending order. Limit 5 will make the query print top 5 states with the highest number of active users

1. Define a function first\_unique that takes a string as input and returns the first non-repeated (unique) character in the input string. If there are no unique characters return None. *Note: Your code should be in Python.*



In this program, order is a list and counts is a dictionary. In the first for loop, we loop through the string once. If x is a new character, we will store in counts. Else, we will increment the value of that character in counts. Order has list of all characters that are present in string. In the second for loop, we will loop through order until we find a character with value 1 in counts.

1. What are underfitting and overfitting in the context of Machine Learning? How might you balance them?

A. Overfitting occurs when the model learns noise and outliers in the data along with underlying pattern. These models have high variance and low bias. Variance is the measure of how spread-out values are and bias is the difference between estimated value and true value. In overfitting performance is great on trained data but fails to predict new data. These models are usually complex like decision trees, SVM, or neural networks that are prone to overfitting. This can be avoided by tuning the parameters like pruning the tree in decision tree after it has learned the details.

Underfitting occurs when the model is excessively simple and cannot capture the underlying pattern. These models have low variance and high bias. These models cannot capture complex patterns in data like linear and logistic regression. Underfitting can be avoided by using more data and reducing features by feature selection.

These steps helps to balance underfitting and overfitting:

* 1. Ensuring that the data is clear and relevant.
  2. Training on more data helps to capture signal better
  3. Using cross validation technique
  4. Removing unnecessary features
  5. Balancing bias and variance

1. If you were to start your data analyst position today, what would be your goals a year from now?
   * 1. My goal is to become technically sufficient in SQL and Python. Matching your needs as a Data Analyst. Within first few months, I want to get complete understanding of your database and able to maintain it according to your policies and procedures. As I get hang of the system, I want to support, implement, and optimize healthcare related software and applications. E.g good choice and efficient implementation of algorithms and data structures effect the system performance. I am interested to see whether any changes can be made in this section to improve system performance.

As I get more and more experience, I want to get in-depth knowledge of business process analysis and data modeling. Having this knowledge is very helpful to understand how different items are related to each other. This will ensure smooth flow of information between different health care agencies. As I get better understanding of how all this working, I am interested to prepare some educational materials like guidelines to store health data that helps to improve data quality in the system.