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. As part of analysis, I created a report 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 details of what further steps can be taken to improve data quality.

2. 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?

A. 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?

A. 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%

3. Given the table users:

```
   Table "users" +-------------+-----------+ | Column      | Type      | +-------------+-----------+ | i
d         | integer    | | username    | character | | email       | character | | city        | chara
cter | | state       | character | | zip         | integer    | | active      | boolean    | +-------------+
-----------+
```

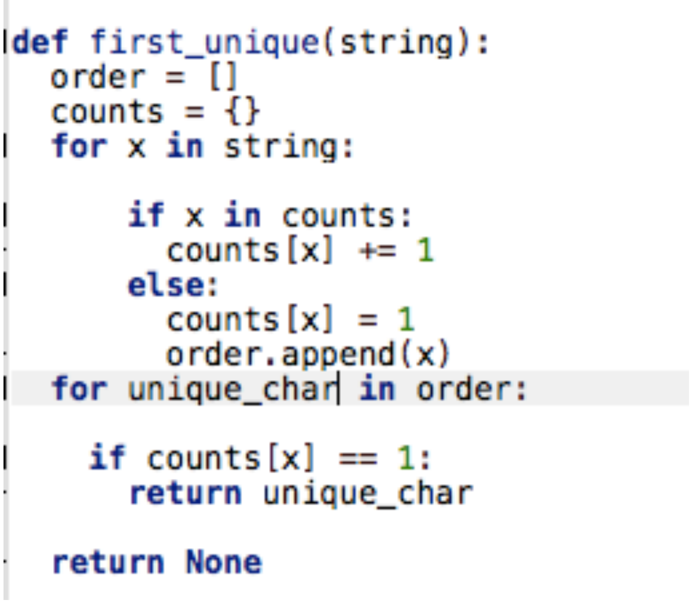
construct a query to find the top 5 states with the highest number of active users. Include the number for each state in the query result. Example result:

```
+------------+------------------+ | state      | num_active_users | +------------+------------------+ |
New Mexico | 502              | | Alabama    | 495              | | California | 300              | | Mai
ne       | 201              | | Texas      | 189              | +------------+------------------+
```

A. 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

4. 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.*



```
def first_unique(string):
    order = []
    counts = {}
    for x in string:

        if x in counts:
          counts[x] += 1
        else:
          counts[x] = 1
          order.append(x)
    for unique_char in order:

      if counts[x] == 1:
        return unique_char

    return None
```

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.

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

A. Underfitting occurs when the model is excessively simple where there is low variance and high bias. Overfitting occurs when a model learns noise instead of signal. In this case, its performance is great on trained data but fails to predict new data.

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

6. If you were to start your data analyst position today, what would be your goals a year from now?

A. I want to become proficient in data analysis by reading, applying, communicating with others, and receiving feedback. I am interested in SQL, python and R. Within a year I what to use database skills to maintain, develop and organize the data in databases and use programming skill to improve and develop health care software.
And at the same time, I want to develop necessary leadership qualities to reach to next level in my career ladder where I can help others.