1. Describe a data project you worked on recently.

While working on my machine learning project, I didn't fit SelectKBest to my features correctly. This created all kinds of problems for me with my machine learning algorithm. Without a proper fit, I wasn't able to find a minimal solution for the problem as required by the grader. Once I fitted it correctly, I was able to find a solution immediately, and optimize it. This turned out to be an important lesson for me. The lesson being, when things aren't working, take a step back and try to understand what has gone wrong with my algorithm.

This brings me to this job interview. I'm interested in working for Convoy for the challenges that you would offer me. While working for you, I would be able to hone the various data mining algorithms to match the varying business analytics such as classification and probability estimation, regression, clustering, and profiling. I would begin honing my skills for them by framing the exact problem that they would liked addressed the most – the single issue that is most important to them. I would try to solve this immediately. To solve it, I would collect their data, discover and visualize it to gain insights, prepare the data for machine learning algorithms, select a model & train it. And finally, I would fine-tune the model and present the solution to Convoy.

Truly understanding all of these tools would allow me to do more for Convoy as a data analyst, and with their data science team to find meaningful insights which would hopefully drive their bottom line, and perhaps, to provide enough experience not to make simple mistakes as I described above.

2a. 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?

(6/10)(5/9)(4/8)(3/7) = 0.0714

2b. 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?

Calculate the possible chocolate combinations:

C(n,r) = n!/(r!(n-r)! = 6!/4!(4-2)!= 24/4 = 6 possible combinations for the chocolates.

Here are the 6 possibilities for the chocolates:

Now, calculate their probabilities and then sum their total.

OOCC = (6/10)(5/9)(4/8)(3/7) = 0.0714
OCCO = (6/10)(4/9)(3/8)(5/7) = 0.0714
CCOO = (4/10)(3/9)(6/8)(5/7) = 0.0714
COOC = (4/10)(6/9)(5/8)(3/7) = 0.0714

COCO = (4/10)(6/9)(3/8)(5/7) = 0.0714
OCOC = (6/10)(4/9)(5/8)(3/8) = 0.0714

0.0714 X 6 = **0.4285**

3. 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?

```
SELECT state, sum(active) as number_active_users
FROM users
GROUP BY state
ORDER BY sum(active) DESC
LIMIT 5;
```

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.

```
### slice through string for each character
### check before and after for more charcters

def first_unique(string):
        for i in range(len(string)):
                if string[i] not in string[0:i] and string[i] not in string[i+1:]:
                        return string[i]
        return None
```

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

Overfitting is when a model performs well on the training data, but it doesn’t generalize well to new data, and underfitting is when your model doesn’t perform well even on training data. Overfitting can general be fixed by simplifying the model by selecting one with fewer parameters (perhaps, a linear

model rather than a high-degree polynomial model), and by reducing the number of attributes in the training data or by constraining the model. You can also reduce overfitting by gathering more training data, and by reducing the noise in the training set by fixing data errors & removing outliers. Constraining the model to make it simpler and to reduce the risk of overfitting is called regularization. The amount of regularization to apply during learning can be controlled by your hyperparameter values. These values be set with GridSearch. The best way to learn how your model will generalize is to split your dataset so that you train on a portion of your data while testing or measuring the model's performance on the remaining data. The best way to split your model is through cross-valiation. Cross-validation splits the training set into multiple complementary subsets, and each model is trained against a different combination of these subsets and validated against the remaining part. With cross-validation the data is used more effectively. A model that is underfit is said to have high bias and low variance. High bias is where a model makes many mistakes on the training set, or that it errors from erroneous assumptions in the learning algorithm. Variance is an error from sensitivity to small fluctuations in the training set. The goal for a machine learning algorithm is to find a balance between these two errors in your algorithm's ability to generalize beyond the training set. One way to balance these two is with the use of feature selection. If you model has high bias, then you would add features to correct it. If you model has high variance, then you would similarly remove features. Finding the right balance will help your model make better generalizations to new data.

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

My goals for a year from now would be to hone my all of my data analyst skills. While I have learned a lot in the past 9 months regarding my Udacity Nanodegree, I still have a long way to go. Every time I learn something new, I realize that there is so much that I don't know. I would work on improving the tools I would need to be a valued team member at Conoy. I would work towards honing that data analyst process of importing the data, cleaning it, transforming it, visualizing it, modeling it, and finally communicating it. And in addition to this, I would like to improve my Python skills. I realize that so much more can be done as a data analyst with better Python skills. So while I'm honing my skills as a data analyst, I would be simultaneously improving my Python skills. And finally, I know that if I put as much effort into my first year while working at Convoy as I did for my Data Analyst Nanodegree that I would be successful in my endeavors to become a better data analyst, and a better team member for Convoy. And this would go far in supporting Convoys mission of reinventing the world's supply train needs through actionable business analytics to improve their supply chain.