**Interview Practice (Data Analyst)**

Question 1: Describe a data project you worked on recently

One of my favorite project that I did as part of my Data Analyst Nanodegree on Email Fraud detection using Machine Learning. We had some data about the Enron company emails and financial information of the employees, the goal was to analyze the data and identify employees who committed fraud. I used Python and Pandas to analyze imbalance in the distinctive features, imputed the null values and detected anomalies. Then I performed exploratory data analysis and created a lot of visualizations to see correlation among the features, this process helped in feature selection as I wanted to capture trends in the data, I did not use features that do not give any information. So to benchmark, I implemented few classifiers like Naïve Bayes, KNearest Neighbors, Random Forest and Support Vector Machines with default settings. Naïve Bayes, KNN and RandomForest performed poorly so I rejected those and chose SVM as my baseline model. Then I did feature engineering and created two new features which described the number of emails sent and received by each employee, and salary vs other means of income ratio. I also perfomed hyper parameter optimization using sklearn’s grid search cv on 1000 epoch cycles and tuned by SVM to rbf kernel and optimal C value which gave me accuracy over 90% on the test data, and 0.88 and 0.92 recall rates. In conclusion, I found out that people who committed fraud had way more income than their salary, the people from they received and sent email the most were also involved in the fraud. Some of the challenges I faced was in tuning SVM so I had to look up online and update by knowledge. Overall, I enjoyed working on a real-world dataset and I feel someone can take this project further by engineering more features of importance or use deep learning algorithms to identify the email fraud.

Question 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?

*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?

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 |

+------------+------------------+