Interview Practice

1. Project Experience

> *Describe a data project you worked on recently*

I did Exploratory Data Analysis Using R recently while I was in the Nanodegree program with Udacity.
The dataset I used was downloaded from Google which was an already curated .CSV file consisting data of about 1600 records and had 11 variables that included the chemical properties of Red wine samples and the quality rating from 1 -10 as ranked by 3 wine experts was also a part of the data.

In this project I have explored the data through various Staistical visualizations using the R language. The goal was to analyse different chemical properties of the red wine and figure out how these influenced the quality of the red wine.

To begin with, I used the basic functions of R language, like Summary(), str(), table() to get an idea of the data and its properties.
Then I started plotting Univariate plots against all the checmical properties in the data versus the count of all the samples. This gave me an idea of how they are distributed.
Using R's built-in functions like grid.arrange(), ggcorr(), it was very easy to plot the visualizations and checking on the trends.
As expected, Alcohol content, sulphates and citric acid affcted the quality

of wine, this was further proven through visualizations in bi-variate and multivariate analysis.

I am always very interested in speculating the new trends or old practices for the betterment of self and society in various fields like Health care and fitness, fashion, consumer market etc. Such analysis using various and randomly available data everywhere to make informed decissions using satistical concepts and using visualization tools to communicate these facts to make a difference is very exciting.

Robert Half Technology is a company that can provide various such projects to occupy consultants with different skills and thus maintaining employment within the company between engagements. There is always a fantastic potential for growth and learning in such jobs that require the employees to work on different platforms and projects for full time without having to worry about finding the projects on oneself.

2. Probability

> **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?*

I am not sure if my thinking is right, but this is how I am calculating the total probability of me eating the first two orange and last two coconut

filling is:

Prob of getting the orange filling in first pick = 6/10

Prob of getting the orange filling again in second pick = 5/9

Prob of getting the coconut filling in third pick = 4/8

Prob of getting the coconut filling in fourth pick = 3/7

so total probability = 6/10 * 5/9 * 4/8 * 3/7 = .071 0r 7.135%

> **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?*

This is a combination probability question.

I am allowed to eat 4 pieces. The number of coconut filling in that should be 2

So considering nCr, C(4,2) = 4! / (2! * (4-2)!) = 6

Therefore there are 6 combinations of chocolates with 2 coconut fillings I can get when I choose 4 pieces.

they are: CCOO, COOC, OOCC, OCCO, OCOC, COCO

We can have a total of 14 combinations like this with Orange and Coconut being any number of times or not at all in the given pick.

So the probability of picking 4 pieces in a row with 2 coconut fillings = 6/14 = 0.4285 or 42.85%

> **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 |*
>
> *±----------±----------------+*

SELECT state, SUM(active) as num_active_users

```
FROM users WHERE active == 1
GROUP BY state
ORDER BY SUM(active) DESC
LIMIT 5;
```

> **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.*

```
from collections import defaultdict
def first_unique(word):
‘’‘initiate defaultdict for count’’’
counts = defaultdict(int)
‘’’ create empty list’’’
l = []
‘’‘loop through each character in a string’’’
for c in word:
counts[c] += 1
‘’‘if there’s first unique character, append them to list’’’
if counts[c] == 1:
l.append©
‘’’ if list only contains 1 character, return the result’’’
for c in l:
if counts[c] == 1:
return c
‘’’ otherwise, return “None”" ‘’’
return “None”
```

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

In Machine Learning, a over fitting or underfitting the data can lead to poor performance of the algorithm.
Overfitting happens when an algorithm tunes the data too well that it learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. So, when a training data is tuned too well to fit the result, the model then does not apply those concepts to the new data and thus negatively impacts the models ability to genaralise.
Underfitting refers to a model that can neither model the training data nor generalize to new data.
Using k-fold cross validation method that uses the different subset of training data k times on unseen data can build up an estimate of the performance of a Machine Learning algorithm.
Cross-validation is a powerful preventative measure against overfitting.
In standard k-fold cross-validation, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set.
Cross-validation allows us to tune hyperparameters with only our original training set. This allows us to keep our test set as a truly unseen dataset for selecting our final model.

Regularization refers to a broad range of techniques for artificially forcing the model to be simpler.The method will depend on the type of learner that is being used. For example, we could prune a decision tree, use dropout on

a neural network, or add a penalty parameter to the cost function in regression.
Oftentimes, the regularization method is a hyperparameter as well, which means it can be tuned through cross-validation.

Another way of overcoming overfit model is by removing features.
Some algorithms have built in feature selection like ‘Regularized Regression’ and ‘Random Forests’. In others we can manually improve their generalizability by removing irrelevant input features. However, feature selection keeps a subset of the original features.

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

I would see myself as a valuable team member. My goals would be to be a proficient Python and SQL coder and an excellent Tableau presenter. Perform complex Data Analysis for making productive business decissions. Robert Half Technology is stategic partners with top IT companies and are committed to serve their IT hiring needs. With the comapany’s learning environment I would like to aquire best knowledge for further improvement of my skill set and seek top-rated opportunities to use and work on various platforms as a consultant to projects through Robert Half Technology’s conncections with other great companies.