README.md

Interview-Practice

Interview practice project #1

1. Describe a data project you worked on recently.

The data project I completed most recently was called "Explore and Summarize Data". In this project I used R to explore and perform statistical analysis on a dataset containing financial contributions made by California residents to Presidential candidates in the 2016 Presidential election. After looking through the 19 variables I decided that only four of the existing variables would be useful to my analysis. Those four variables were the names of the candidates, the cities where the contributions came from, the the occupation of the people who donated, and the donation amounts. However, because there were far too many individual occupations to easily graph and examine, I ended up writing a function that used information from the United States Census Bureau that used the North American Industry Classification System to group the occupations into categories based on industry groupings. I also wrote a function to create a variable for each candidates political party. So in total, I used five variables for my analysis. I then began my analysis by plotting the counts of all five variables and moved on to get the total donation amounts for the other four variables. To finish up, I analyzed total donation amounts for combinations of variables (ex. donation amounts given to each political party by occupation). The main points of conclusion were that California residents overall, and especially those in bigger cities, lean Liberal and voted for and supported the Democratic candidates. In addition, Democratic candidates overall had the most people donate to them and the most money donated to them, and when broken down by occupation they won the support of every occupational category in both count and dollar amount.

I am also currently working on a project that is very relevant to the kind of work I would be doing here at Duolingo. It involves developing an A/B test on student data for an online education company called Udacity. Udacity courses currently have two options on the home page: "start free trial" and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback. For the experiment, a change will be tested on what happens if the student clicks "start free trial". They will be asked how much time they have available to devote to the course, and if the student indicates 5 or more hours per week, they will be taken through the checkout process as usual. If they indicate fewer than 5 hours per week, a message will appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggest that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead. The hypothesis is that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who leave the free trial because they don't have enough time (without significantly reducing the number of students to continue past the free trial and eventually complete the course). If this hypothesis holds true, Udacity could improve the overall student experience and improve their coaches' capacity to support students who are likely to complete the course.

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?

Total Pieces	Orange Cream	Coconut Filling
10	6	4

P(1st Orange Cream)	P(2nd Orange Cream)	P(1st Coconut Filling)	P(2nd Coconut Filling
6/10	5/9	4/8 or 1/2	3/7

In order to find the probability of this exact sequence, the probabilities above need to be multiplied together:

```
6/10 * 5/9 = 30/90 or 1/3

1/2  * 3/7 = 3/14

1/3 * 3/14 = 3/42 or 1/14
```

Answer
1/14 or ~0.07

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?

First, let's list all the possible combinations where there are more than exactly 2 chocolates with coconut filling:

1	2	3
CCCC	CCCO	OCCC

Next, let's list all of the possible combinations where there are exactly 2 chocolates with coconut filling:

1	2	3	4	5	6
CCOO	COCO	COOC	OCCO	OCOC	OOCC

Finally, let's list all of the possible combinations where there are less than exactly 2 chocolates with coconut filling:

1	2	3	4	5
COOO	OCOO	OOCO	OOOC	OOOO

So, to sum up:

Total Combinations of 4 Chocolates	Total Combinations with Exactly 2 Coconut Filling	P(Of Exactly 2 Coconut Filling)
14	6	6/14 or ~0.43

Answer
6/14 or ~0.43

3. Given the table "users":

Column	Type
id	integer
username	character
email	character
city	character

Column	Type
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

My Code:

```
SELECT state, COUNT(*) AS num_active_users  FROM users
WHERE active = TRUE
GROUP BY state
ORDER BY num_actve_users  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.

Example:

```
def first_unique(string):
    # Your code here
    return unique_char

> first_unique('aabbcdd123')
> c

> first_unique('a')
> a

> first_unique('112233')
> None
```

My Code:

```
def first_unique(string):
    """
    Iterates through the string and counts how many times each character occurs
    in the string.  If the character only occurs once, return that character.
    If the all the characters occur more than once, return "None".
    """
    for char in string:
        if string.count(char)  == 1:
            return char
    return None
```

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

With machine learning, the goal is to use an algorithm to try and classify data points into groups depending on the characteristics of the data points. In order to do this, I would first train the algorithm on some training data, and this is where underfitting and overfitting can start. If the algorithm is underfitting the training data, it would not be able to correctly categorize the training data nor any new data accurately. If it is overfitting the data, it is able to categorize the training data with great accuracy (say at 90% or higher), but is not able to generalize that knowledge and categorize any new data accurately.

For example, let's say you have some data for road conditions that is used to train self-driving cars, and the data represents the steepness and bumpiness of road terrain. The goal of the algorithm would be to categorize the data points into terrain that requires the car to drive at a certain speed depending on the steepness and bumpiness of the terrain. If the algorithm is underfitting the data, the car would not be able to drive on real roads without crashing a lot. The same goes for overfitting, but in a slightly different way. If the algorithm is doing too good of a job at categorizing the training data, it would do very well on the terrain that fit the training data well, but would do very poorly on terrain that is vastly different than the training data.

In order to balance them, it is best to try and tune the parameters so that the algorithm can generalize to new data. This can be done by choosing good data, limiting the number of variables you use to train your algorithm to the ones that have the most predictive power, and properly randomizing the data chosen for the training data. If all else fails, especially in the case of underfitting, the algorithm chosen might not be the best fit, so playing around with different algorithms to find the best one will help.

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

My goals a year from now are to have improved as a data analyst in general. I am still new to the world of data analytics and I know I still have a lot to learn. I want to continue to improve my statistical analysis skills, become a more efficient programmer in Python and R, and gain more experience performing A/B Tests and predictive analysis using machine learning. I also want to learn how to become more business oriented in my analysis so that I can bring value to the company and have helped the company as a whole grow.

Beyond a year from now, I would want to have become proficient in Java and "big data" systems like Hadoop so that I can better deliver the results that Duolingo needs. Also, coming from a background in teaching English as a foreign language, I hope to become more involved in the teaching side of Duolingo. I would love to help to improve Duolingo's learning experience and help expand Duolingo's service offerings to better meet the needs of the hundreds of thousands of users around the world.