## ***Author: Paulina Grunwald***

***Date: 07/04/2017***

## **Project: Answer the Following Interview Questions (6 total)**

For each of the questions below, answer as if you were in an interview, explaining and justifying your answer with two to three paragraphs as you see fit. For coding answers, explain the relevant choices you made writing the code.

1. ***Describe a data project you worked on recently.***

Recently, I have been working on various projects which were assignments In my Udacity Data Analyst Nanodegree. One of them was performing data analytic on white wine dataset for which I have used R programming language. The goal was to asses which wine properties have influence on it’s quality. The dataset contained 13 variables which were representing white wine properties e.g alcohol content, acidity, quality, sweetness. I have used various R packages for my project. Most of the plots were done using ggplot. I have created bar plots, scatter plots, boxplots and histograms. I applied different techniques to get bettet insights from my plots: adding jitter, adjusting axes (limits, log10 etc.), remove top and bottom 1% of the outliers. As a first step of I have done univariate analysis. i have have looked at every white wine property separately by creating histogram to asses if the variable has normal distribution or not and if the outliers are present. In bivariate section of my project I looked at correlation between two different white wine properties. Using ggpairs and corrplot I have create correlation plots. The last part of my analysis consisted of multivariate plot section. Using ggplot I plotted 3 different wine properties on one scatter plot. To adjust colours of the plots I used RColorBrewer package. I have also created a model with lm function but unfortunate it did not result in a good prediction for white wine quality

Analyzing the data I have observed that most frequent quality value of white wine are 5 and 6. Alcohol seem to be a main factor correlated to the wine quality. The data strongly suggest that the higher the alcohol content the better wine quality. When alcohol percentage increases, density decreases.

- The residual suger content increases with increase of density and decrease of alcohol content.

.

- High quality wines have high content of alcohol and are either dry or medium dry (wine sweetness).

I believe the quality score given by the jugdges is linked with their personnal perference or i could depend on other variables which were not provided in white wine datas

I also found the relationship between sugar, density, and alcohol to be interesting and I explored that in detail. The suprising thing for me was that the high quality wines can be characterized by two types of sweetness: medium dry and dry. Actually it's also suprizing in the end that alcohol is very well correlated with wine quality. I would expect that not all the wines with high content of alcohole would be high quality wines. Since the wine quality jugment can be very personal it is possible that juges that gave the scores for the wine in this dataset enjoyed white wines with higher content of alcohol.

This project made me quite interested in the wine quality. I would be very interested to do similar experiment either with red wine dataset or beer dataset.

1. ***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 number of truffles: 10

6 - cream orange filling

4 - coconut filling

Instruction: eat 4 pieces in a row choosing random truffles

At the start we are given 10 truffles. 6 of them have orange filling and 4 coconut filling. The probability of selecting cream orange truffle first is 6/10. Now, we have remaining 9 truffles. The probability of selecting 2nd cream orange truffle in a row is 5/9 (there are 5 cream orange truffles and 4 coconut filling truffles left. So, we are left with 8 truffles: 4 that have cream orange filling and 4 that have coconut filling. The probability of selecting truffle with coconut filling is now 4/8. Now we remain only with 7 truffles and probability of selecting another coconut filling truffle is 3/7.

(6/10)\*(5/9)\*(4/8)\*(3/7) = 0.07*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?

There are 14 various combinations of coconut and orange truffes. I will have a lookk now how many variations of set with 2 coconut truffles are possible

O - orange

C - coconut

1: C C O O

2: C O C O

3: C O O C

4: O C C O

5: O C O C

6. O O C C

There are 6 possible combinations of drafting set of 4 truffles with exacly two coconut truffles. The probability of drafting 4 truffles from which exacly two have coconut filling would be 6/14 = 0.42857

***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. 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 |*

*+------------+------------------+*

*Solution for the problem:*

SELECT state, sum(active) as num\_active\_users

FROM users

GROUP BY state

ORDER BY sum(active) DESC

LIMIT 5

***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.***

def first\_unique(string):

count = []

for i in string:

count[c] += 1

return unique\_char

> first\_unique('aabbcdd123')

> c

> first\_unique('a')

> a

> first\_unique('112233')

> None

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

Overfitting and underfitting are the two biggest causes for poor performance of machine learning algorithms. Overfitting occurs when a statistical model or machine learning algorithm captures the noise of the data and the algorithm fits the data too well. It often happens when the model or algorithm shows low bias but high variance.

Overfitting can be prevented by fitting multiple models and using validation or cross-validation to compare their predictive accuracies on test data.

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

My goal would be to develop my data analytic skills and