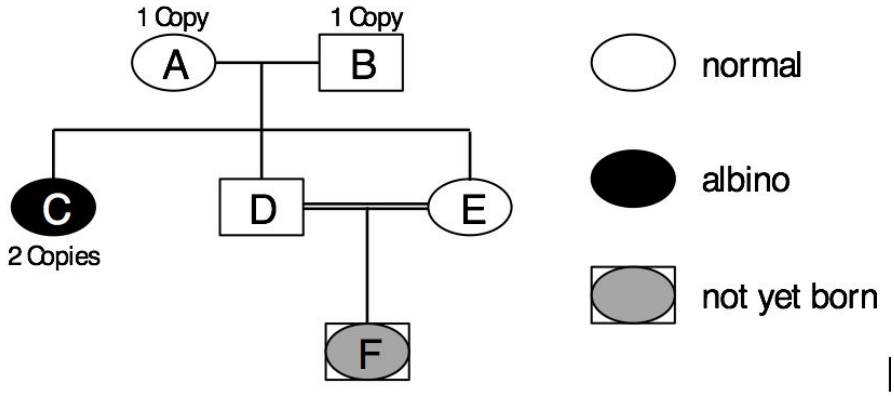
Question 1: Probability

Below is the family tree for a small parade of elephants, one of which ("C") exhibits albinism caused by a mutation in the gene for skin pigmentation. Each elephant carries two copies of this gene and in order to be an albino, both copies must have the mutation. When the elephants mate, each parent passes one of their copies onto the child with equal probability (regardless of mutation).

Elephants A and B each carry only 1 copy of the mutation (and thus don't exhibit albinism) and elephant C carries 2 copies (and is affected), as shown in the figure. Elephants C, D, and E are all descendants of A & B, and Elephant F is a descendant of D & E.



Please show your work or reasoning for each of the following questions:

a) Given that elephant D is not an albino, what is the probability it has exactly one copy of the mutation? *[1 point]*
b) Given that elephant E is not an albino, what is the probability it has exactly one copy of the mutation? *[1 point]*
c) What is the probability that elephant F, when born, will be an albino? *[2 points]*
d) What is the probability that elephant F, when born, will carry exactly one copy of the mutation? *[2 points]*
e) Suppose that elephant F is born and is not an albino. Given this, what is the probability that elephant F carries exactly one copy of the mutation? *[2 points]*
f) Suppose that elephant F is born and is not an albino. Given this, what is the probability that elephant D carries exactly one copy of the mutation? *[2 points]*

Question 2: Model interpretation

In order to understand and predict whether a certain user will buy a stock of company X, a simplistic logistic regression that predicts this probability is trained. The coefficients of this model are:

Variable	Coefficient	Standard Error	z-score
Male	1.45	0.09	15.81
Female	-2.11	0.10	-21.95
Ln_Equity value	-0.109	0.041	-2.63
Age	-0.013	0.0096	-1.4
Age_Sq	0.0001	0.00009	1.52
Sector trades	1.77	0.33	5.37
Company Y	2.07	0.399	5.18
Company Z	-0.872	0.437	-2.01
Constant	1.2	0.484	2.48

The variables are defined as:

- Male - coded as 1 if the user is male, 0 if not
- Female - coded as 1 if the user in female, 0 if not
- Ln_ Equity value - natural log of the user's equity value (total dollars in Robinhood account)
- Age - User's age (years)
- Age_Sq - User's Age squared
- Sector trades - percentage of trades made in the same vertical or sector as company X with respect to the total trades made by the user
- Company Y - 1 if the user currently holds Company Y
- Company Z - 1 if the user currently holds Company Z
- Constant - The constant term

a) Consider 4 users, Adam, Bob, Chris and David. Adam and Chris share identical characteristics except for their equity values. Bob and David also share identical characteristics (with each other, not necessarily Adam and Chris), except for their equity values.

Name	Equity value	Modeled Probability
Adam	$50,000	50%
Bob	$200,000	50%
Chris	$40,000	?
David	$190,000	?

Based on the coefficients above, who would you think has a higher probability of buying a stock of Company X?

- Chris
- David
- They have the same probability
- Cannot tell based on the information provided

What is your reasoning? (you need not calculate an exact probability to answer this question. Just explain your reasoning in general terms.) *[2 points]*

b) The coefficient for Company Z is negative. How do you interpret this? *[2 points]*

c) How do we interpret the difference in buying probability between users of different ages? How do the variables in the model estimate such support? *[2 points]*

d) Are there any variables in this model that you would choose to drop? Why or why not? Would you need more information in order to make this decision? *[2 points]*

Question 3: Data Query

Consider the following table schema:

Table: trades
Order_id - integer
User_id - integer
Timestamp - datetime
Symbol - string
Price - float
Quantity - int
Side - (buy,sell)
Status - (completed, cancelled)

Table: users
User_id - integer
City - string
Email - string
Age - int
Address - string
Created_at - timestamp

Using SQL or a similar query language and the table schema presented above, write queries that answer the following:

a) List the top 5 cities which had the highest number of users cancel orders on 1st Dec 2016. *[2 points]*

b) Sometimes when one user is selling and the other buying the same symbol for the same price and same quantity at the same time, they trade against one another and not with the exchange. List the top 3 symbols which have been traded amongst users. *[3 points]*

Question 4: Modeling

Please use python, R or a similar tool for answering this question.

In order to improve user retention and lower churn, the growth team at Robinhood is interested in understanding why and which users withdraw money from their Robinhood account. A user is considered *churned* when their equity value (amount of money in Robinhood account) falls below $10 for a period of 28 consecutive calendar days or longer.

Using the datasets given below answer the following questions:

a) What percentage of users have churned in the data provided? *[4 points]*
b) Build a classifier that given a user with their features assigns a churn probability for every user and predicts which users will churn. How well does your classifier perform? State any metrics you deem important here. Based on the classifier output classify each user in the dataset as churned or not churned. *[5 points]*
c) List the most important features that correlate to user churn. *[3 points]*
Please provide the code and any explanation of your assumptions and methodology.

Datasets:

- *features_data.csv* - contains user level data such as:

  user_id - unique id for every user
  risk_tolerance - self-reported risk tolerance of the user
  investment_experience - self reported investment experience of the user
  liquidity_needs - self reported liquidity needs of the user
  time_horizon - self reported investment time horizon of the user
  platform - which platform (iOS or Android) the user is on
  time_spent - amount of time spent on the app
  first_deposit_amount - $ value of the amount first deposited
  instrument_type_first_traded - type of instrument first traded

- *equity_value_data.csv* - contains *user_id* and *equity_value* for user along with timestamps for days when the user's equity value is greater than or equal to $10.