Question 1: Data Query

Consider the following table schema:

Table: transactions	
Variable	**data_type**
Transaction_id	*Integer*
User_id	*Integer*
Recipient_ id	*Integer*
Timestamp	*datetime*
CashDirection	*In/Out*
Amount	*float*
CTR	*binary. A CTR is filed when Cash transaction is more than 10,000.*
Transaction_type	*Cash/ACH/Check/Wire*

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

a) List the top 5 user_id which had the highest number of CTR filed during in any 7 days window period. (5 points)

b) List top 5 user_id which had the largest amount of incoming amount over any 30 day period exceeding $1,000,000? (5 points)

Question 2: Modeling

In order to recognize the malpractice of using cryptocurrency exchange towards money laundering, a logistic regression model is trained to predict whether a certain user at a bank will engage in cryptocurrency transaction. The coefficients of this model are:

The variables are defined as:

- Male - coded as 1 if the user is male, 0 if not
- Account Balance- natural log of the user's account balance
- Age - User's age (years)
- Age_Sq - User's Age squared
- Investor - If a user is investor or not, based on transactions
- Works_at_Y: If user works at company "Y".

➢ Constant - The constant term

Variable	Coefficient	Standard_Error
Male	2.45	0.12
Account balance	-0.109	0.041
Age	-0.0135	0.00096
Age_Sq	0.0001	0.000029
Investor	3.21	0.67
Works_at_Y	-5.12	0.399
Constant	2.8	0.584

a) For a not 'Male' customer, what is the most important features in predicting likelihood of customer doing cryptocurrency transaction? (5 points)

b) How do we interpret the difference in probability using cryptocurrency exchange between users of different ages? (5 points)

Q3. Data Analysis

Refer to the data file "transaction_data.csv" to answer the following questions. Use of Python/R is highly recommended.

To detect money laundering, your analysis can use some signals based on pattern of transactions. Lets say that if you observe unusual amount of money coming to your account, then that is considered to be suspicious.

a) How you would determine if an amount is unusual for a bank's user? (3 points)

b) A relative of user is defined as someone who shares same phone number. Find the set of relatives for all user_id's. (3 points)

c) When a transaction happens among relatives, call it circulatory (among relatives) transaction. Create a binary feature if a transaction is circulatory transaction. (4 point)

d) Suppose that you would like to track weekly incoming (from counterparty to user's account) and outgoing (from user's account to counterparty's account) amount (i.e. sum of weekly amount). If this amount is more than $100,000, then an alert should be generated. Create features using these logics. (5 points)

e) Addition to features in d), you would like to use features which tracks weekly number of transaction by "trx_type" and "activity_type". Create features which uses this logic. (5 points)

f) Using the features as created in part (c), (d) and (e), build a statistical machine learning model to detect the suspicious money laundering cases. (To build a model, data (features, and Y) can be aggregated at ["user_id", "date"] level.) Report model performance using a relevant metric (and why you would choose this metric?) (8+2 points)