Conclusion

Question 1

- Programmatically download and load into your favorite analytical tool the transactions data. This data, which is in line-delimited JSON format, can be found here
- Please describe the structure of the data. Number of records and fields in each record?
- Please provide some additional basic summary statistics for each field. Be sure to include a count of null, minimum, maximum, and unique values where appropriate.

There have total of 786363 records in the dataset and in all records,12417 of them appears to be fraud record. There are 5000 users in this dataset and their transaction records come from 2490 merchants in four counties: US(774709), MEX(3130), CAN(2424) and PR(1538). May be because of mobile payment is blooming, 352868 records are recorded with no card present. There are 19 categories for the merchants, and online_retail reached the top count of records with number of 202156. In order to analyze the data more conveniently, I deleted 6 columns of data (echoBuffer, merchantCity, merchantState, merchantZip, posOnPremises, recurringAuthInd) and tranformed 2 columns into data format (transactionTime, accountOpenDate). For the remaining data, 5 columns have null values (acqCountry:4562, merchantCountryCode:724, posConditionCode:409, posEntryMode:4054, transactionType:698).

Question 2

- Plot a histogram of the processed amounts of each transaction, the transactionAmount column.
- Report any structure you find and any hypotheses you have about that structure.

The majority of transactions are under 500$. There might have three reasons lead to this situation, 1)Few people would make large transactions using credit card. 2) Credit of credit card is limited, people would like to use cheque or debit card to make large transactions. 3) The society is becoming a cashless one, people would like to use credit card to make purchuse instead of using cash.

Also, I found that there are several transactions with 0. After reviewing the data, I found that all the cases of 0 are for address verification. In those transactions, there is still have fraud situations. So we cannot assume that it will not be a fraud if the transaction type is 'ADDRESS_VERIFICATION'.

Question 3

You will notice a number of what look like duplicated transactions in the data set. One type of duplicated transaction is a reversed transaction, where a purchase is followed by a reversal. Another example is a multi-swipe, where a vendor accidentally charges a customer's card multiple times within a short time span.

•Can you programmatically identify reversed and multi-swipe transactions?
•What total number of transactions and total dollar amount do you estimate for the reversed transactions? For the multi-swipe transactions? (please consider the first transaction to be "normal" and exclude it from the number of transaction and dollar amount counts)
•Did you find anything interesting about either kind of transaction?

If we consider 2 minutes as the time slot to define a Data Wrangling and Dupilcate Transactions, then the estimate numer of transactions for reversed transactions is 1901, the estimate amount is 270312.77 dollars. For the duplicate transactions, the estimate number of it is 2490 and the amount reached 370634.35 dollars.

Question 4

•Each of the transactions in the dataset has a field called isFraud. Please build a predictive model to determine whether a given transaction will be fraudulent or not. Use as much of the data as you like (or all of it).
•Provide an estimate of performance using an appropriate sample, and show your work.
•Please explain your methodology (modeling algorithm/method used and why, what features/data you found useful, what questions you have, and what you would do next with more time)

In Question 4, I used Random Forest as the basic model to make prediction. The accuracy using balanced data (number of fraud and not fraud records are the same) reached 0.684 while using original data is 0.724. In 6 features I seleced, 'transactionAmount' and 'posEntryMode' should be paid more attention because they are the key factors when predicting with Random Forest. It's not a 'good enough' model because the accuracy is not high enough, especially consider that most cases are not fraud in original data. Maybe I can try different ways to get a higher accuracy with more times.

Question I Have

1. Some columns have no data in it, like 'echoBuffer', 'merchantZip',etc. I want to know why all these columns of data are null? Because I think merchantZip might be an important factor to predict fraud as some locations may have higher chance to fraud than other area.

2. I'm not sure about the definition of reversed transaction. The definition is a purchase is followed by a reversal, but I'm not sure if there have other transactions between a purchase and a reversal can be defined as a reversed transaction.

If I Have More Time

If I have more time, I want to do the following things:

1. I want to change the ratio of records between non-fraud and fraud to see if it will affect the accuracy. For example, the ration is 2:1 or 3:1.
2. Use different models or optimize models I tried, like Logistic Regression and Naive Bayes.
3. I didn't dig deep in the column 'merchantName' because there are too many stores or sevices providers, I want to figure out if 'merchantName' have effect when predicting a fraud transaction.