Thank you for taking the time to complete the Capital One Data Science Challenge! Please confirm receipt of this email and follow the instructions below.   
  
**Instructions for Submission:**  
**Due Date:** The DSC is due no later than 9 AM EST on Monday, December 26th EST.  
*Challenges must be received within 7 calendar days. It is imperative that you return your submission by that date or your challenge will be marked as failed, which will prohibit you from applying to Data Science roles for 6 months from the deadline date.*  
**Zip File Name:**DSC\_C2561641  
**Send to:**[**DataScienceRecruitingMailbox@CapitalOne.com**](https://mailto:DataScienceRecruitingMailbox@capitalone.com/)**and CC me (**[**kristina.skordas@capitalone.com**](https://mailto:kristina.skordas@capitalone.com/)**).**  
**Please copy and paste the following into the body of your email:**  
**CID: C2561641**  
**Recruiter:** Kristina Tichner   
**Version:**v20.01  
  
We estimate that this Data Challenge will take approximately five to seven hours to complete. When you are finished, please create a zip file (**see required file name above**) containing ONLY your work/submission and email the file to the above addresses.   
In an attempt to keep your submission anonymous to our reviewers, **you must refrain from including your name, initials or any other Personal Identifiable Information within your zip file.**Including such information may result in delays to the processing of your submission.

Please do not submit the downloaded data set. If you reference materials/libraries/code found online, please source appropriately. If you have any questions or concerns, feel free to email the mailbox.  
To maintain the integrity of the assessment process please complete the assessments independently and keep the assessment content confidential. All content in this assessment is proprietary and strictly confidential. Failure to maintain the confidentiality of the assessment process and materials will disqualify you from further consideration and other consequences, up to and including termination of employment.  
If you require a reasonable accommodation to take this assessment, or for any step in the application or interview process, please contact Capital One Recruiting at 1-800-304-9102 or [RecruitingAccommodation@capitalone.com](https://mailto:RecruitingAccommodation@capitalone.com/).  
All information you provide will be kept confidential and will be used only to the extent required to provide a reasonable accommodation.

**Your submission**  
At Capital One, you'll almost always be working with a diverse team, often including business analysts, engineers, product managers, and senior leaders. The ability to not just complete data science work, but also to explain and present your work is important.  
With that in mind, please deliver your answers as if you were handing off work to another data scientist joining your team. Tying together your intent, code, analysis and results is the goal.  
The following artifacts are valued:

* explanations of your intent, methods, conclusions and any assumptions
* clear, documented, and well-structured code
* instructions for running your code
* methods you attempted that didn't work
* ideas you didn't have time to complete but would have done with more time
* a thorough write up with any pertinent visualizations

We ask that any code written by others is sourced/cited appropriately. We are excited to review your work and your interpretation of the data.  
**Data Science Challenge: Card Transactions!**  
This coding and analysis challenge is designed to test your skill and intuition analyzing real[-ish] world data. For the challenge, we will use credit card transactions data. Note that this dataset loosely resembles real transactional data from Capital One credit card customers, but the entities and relations within are purely fictional. No persons, places, or things lost their identity in the making of this dataset.  
**Required Questions: Please answer all four  required questions completely.**  
**Question 1: Load**

* 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**](https://urldefense.com/v3/__https:/github.com/CapitalOneRecruiting/DS__;!!EFVe01R3CjU!OqjAT7sK8gyc70Vt0xU4WDiHTmpZTHNfTCyUVAIPTd3BAzZfW9xxwUi-euxuqQaPlyANJz8$)
* 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.

**Question 2: Plot**

* 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.

**Question 3: Data Wrangling - Duplicate Transactions**  
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?

**Question 4: Model**  
Fraud is a problem for any bank. Fraud can take many forms, whether it is someone stealing a single credit card, to large batches of stolen credit card numbers being used on the web, or even a mass compromise of credit card numbers stolen from a merchant via tools like credit card skimming devices.

* 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)

Thank you very much for your efforts! We look forward to reviewing your insights!