**Unit 7 Homework Assignment: Looking for Suspicious Transactions**



[*Credit Card Fraudster by Richard Patterson*](https://www.flickr.com/photos/136770128@N07/42252105582/) *|* [*Creative Commons Licensed*](https://creativecommons.org/licenses/by/2.0/)

**Background**

Fraud is everywhere these days—whether you are a small taco shop or a large international business. While there are emerging technologies that employ machine learning and artificial intelligence to detect fraud, many instances of fraud detection still require strong data analytics to find abnormal charges.

In this homework assignment, you will apply your new SQL skills to analyze historical credit card transactions and consumption patterns in order to identify possible fraudulent transactions.

You are asked to accomplish three main tasks:

1. [Data Modeling](#Data-Modeling): Define a database model to store the credit card transactions data and create a new PostgreSQL database using your model.
2. [Data Engineering](#Data-Engineering): Create a database schema on PostgreSQL and populate your database from the CSV files provided.
3. [Data Analysis](#Data-Analysis): Analyze the data to identify possible fraudulent transactions.

**Files**

* [card\_holder.csv](/asu-coding-bootcamp/asu-tem-fin-pt-09-2020-u-c/-/blob/master/Homework/Homework7/Instructions/Data/card_holder.csv)
* [credit\_card.csv](/asu-coding-bootcamp/asu-tem-fin-pt-09-2020-u-c/-/blob/master/Homework/Homework7/Instructions/Data/credit_card.csv)
* [merchant\_category.csv](/asu-coding-bootcamp/asu-tem-fin-pt-09-2020-u-c/-/blob/master/Homework/Homework7/Instructions/Data/merchant_category.csv)
* [merchant.csv](/asu-coding-bootcamp/asu-tem-fin-pt-09-2020-u-c/-/blob/master/Homework/Homework7/Instructions/Data/merchant.csv)
* [transaction.csv](/asu-coding-bootcamp/asu-tem-fin-pt-09-2020-u-c/-/blob/master/Homework/Homework7/Instructions/Data/transaction.csv)

**Instructions**

**Data Modeling**

Create an entity relationship diagram (ERD) by inspecting the provided CSV files.

**Note:** For the credit\_card table, the card column should be a VARCHAR(20) datatype rather than an INT.

Part of the challenge here is to figure out how many tables you should create, as well as what kind of relationships you need to define among the tables.

Feel free to discuss your database model design ideas with your classmates. You can use a tool like [Quick Database Diagrams](https://app.quickdatabasediagrams.com/#/) to create your model.

**Data Engineering**

Using your database model as a blueprint, create a database schema for each of your tables and relationships. Remember to specify data types, primary keys, foreign keys, and any other constraints you defined.

After creating the database schema, import the data from the corresponding CSV files.

**Data Analysis**

Now that your data is prepared within the database, it's finally time to identify fraudulent transactions using SQL and Pandas DataFrames. In this part of the homework assignment, you will analyze the data and then create a report within a Jupyter Notebook file to present your findings. Your report should answer the following questions:

* How can you isolate (or group) the transactions of each cardholder?
* Consider the time period 7:00 a.m. to 9:00 a.m.
  + What are the 100 highest transactions during this time period?
  + Do you see any fraudulent or anomalous transactions?
  + If you answered yes to the previous question, explain why you think there might be fraudulent transactions during this time frame.
* Some fraudsters hack a credit card by making several small payments (generally less than $2.00), which are typically ignored by cardholders. Count the transactions that are less than $2.00 per cardholder. Is there any evidence to suggest that a credit card has been hacked? Explain your rationale.
* What are the top five merchants prone to being hacked using small transactions?
* Once you have a query that can be reused, create a view for each of the previous queries.

Create a report for fraudulent transactions of some top customers of the firm. To achieve this task, perform a visual data analysis of fraudulent transactions using Pandas, Plotly Express, hvPlot, and SQLAlchemy to create the visualizations.

* Verify if there are any fraudulent transactions in the history of two of the most important customers of the firm. For privacy reasons, you only know that their cardholders' IDs are 18 and 2.
  + Using hvPlot, create a line plot representing the time series of transactions over the course of the year for each cardholder. In order to compare the patterns of both cardholders, create a line plot containing both lines.
  + What difference do you observe between the consumption patterns? Does the difference suggest a fraudulent transaction? Explain your rationale.
* The CEO of the firm's biggest customer suspects that someone has used her corporate credit card without authorization in the first quarter of 2018 to pay for several expensive restaurant bills. You are asked to find any anomalous transactions during that period.
  + Using Plotly Express, create a series of six box plots, one for each month, in order to identify how many outliers there are per month for cardholder ID 25.
  + Do you notice any anomalies? Describe your observations and conclusions.

**Challenge**

Another approach to identify fraudulent transactions is to look for outliers in the data. Standard deviation or quartiles are often used to detect outliers.

Read the following articles on outliers detection, and then code a function using Python to identify anomalies for any cardholder.

* [How to Calculate Outliers](https://www.wikihow.com/Calculate-Outliers)
* [Removing Outliers Using Standard Deviation in Python](https://www.kdnuggets.com/2017/02/removing-outliers-standard-deviation-python.html)
* [How to Use Statistics to Identify Outliers in Data](https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/)

**Submission**

* Create an image file of your ERD.
* Create a .sql file of your table schemata.
* Create a .sql file of your queries.
* Create a Jupyter Notebook for the visual data analysis and the challenge.
* Create and upload a repository with the above files to GitHub and post a link in BootCamp Spot.

**Hint**

For comparing time and dates, take a look at the [date/time functions and operators](https://www.postgresql.org/docs/8.0/functions-datetime.html) in the PostgreSQL documentation.

© 2020 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.