Project for Data Science Candidates

This project is about extracting actionable insights from enterprise transaction data. Avoid non-specific statements or general observations on the data. We are looking for a detailed result providing information that an enterprise decision maker could act upon to save money, reduce risk or otherwise improve the efficiency of operating their business.

Some examples include:

* Duplicate transactions (can request reimbursement)
* Fragmented categories: categories of spend for which there is no dominant supplier or handful of suppliers (can consolidate many suppliers into few)
* Outlier transactions: transactions that are much larger than the typical transaction amount for a given supplier in a given category (can investigate further)

Included tables:

* Transactions (described below)
* Normalized vendor name to vendor URL mapping

Transaction table columns:

* transaction\_date
* transaction\_id: unique transaction identifier.
* transaction\_type: **PCard** (i.e., purchase card or corporate card), made by specific employees on a credit card, or **AP** (i.e., accounts payable), payment on an invoice directly by the company.
* employee\_name: NULL for AP transactions.
* business\_unit\_name: the name of the company’s business unit (independent division)
* Different ways to group transactions:
  + account\_name
  + cost\_center\_name
  + expense\_type
  + category
* amount: the dollar amount of the transaction (negative for refunds).
* vendor: the original vendor label that came, e.g., from the credit card company.
* normalized\_vendor: somewhat normalized version of the vendor label.

Some actionable insights that can be potentially extracted from data:

* The data is mostly but not perfectly clean. What are some ways to clean it up?
  + The normalized vendor name to vendor URL mapping table can be handy for cleaning up the normalized vendor column.
  + Other columns might contain dirty data as well.
* Are there time periods for which data is obviously missing?
* Based on vendor usage, can you group some of them, with an eye towards reducing the number of unique vendors per transaction category?
* Can you predict monthly/weekly spend for the next 6 periods? Your predictions can be by vendor, category, employee name, etc. How precise do you expect your estimates to be?
* Can you identify “suspicious” transactions, e.g., those where a specific employee spends much more than predicted on a single transaction or over the course of a time period?
* Anything else you can think of.

We don’t expect you to spend more than 4 hours on this project. In case you do, please include the number of hours worked.

Please include all your artifacts: code, output, findings. A Jupyter notebook would work. We are looking for documentation of your thought process with detailed description of the operations/transformation you executed and a summary statement of the findings.

Questions? Please ask. Contact: Yakov Keselman, [yakov@suplari.com](mailto:yakov@suplari.com)

**Project deadline: Sunday, April 8, 8 pm (if you need more time, please email Yakov).**