**Innovation Lab Interview Project**

**The Motivation**

In the Innovation Lab, our data scientists interact with many types of data sets and technologies. They also present their findings and contribute to business strategy due to the start-up nature of the lab. As a result, this project will allow us your skill in these different crucial areas of being a data scientist:

1. Develop a goal or problem statement (can come after step 4 when you explore data more)
2. Ingesting data from a data source
3. Validating and cleaning the data
4. Explore the data set
5. Extract and transform data that you need
6. Model the data/Gather some insight
7. Interpret the model/data findings and present in a meaningful way

**What we are looking for**

Depending on the data set and your objectives, certain steps will be more important and take more time than others. We want to see a **breadth** and **depth** of skill.

For example, different steps have certain technologies well suited for them (Step 1: Ingestion pipeline or API, Step 3: SQL or Pandas, Step 7: Website or Dashboard). In these steps, picking the right technology based on your experience and research you do online will show us you can pick up or use technologies appropriately. Therefore, we will reward a variety of technologies used (so long as used appropriately). This is because day-to-day our data scientists can interact with many different technologies.

We will also reward depth of expertise and solution. For example, if certain steps are more difficult technology-wise or are unfamiliar, do not worry. Show us your skill and specialty in one area, e.g. modelling and interpreting that model clearly. This will allow us to know your strengths which contribute to the strengths of the lab because we can rely on you to use or teach your specialty to others.

Your goal is to **recommend a product or list of products to customers.** We do not expect you to be able to hit all of these in this time period, but the more breadth and depth you can show, the better we can judge your skills. Be able to explain your assumptions and your strategy behind giving recommendations. Please also tell us, given more time (say 1 week or 1 month), what else you would do/prioritize.

Tips: For customers with too few purchases, feel free to ignore or change your strategy, just explain your choice. Also, it would be helpful to understand your objective when giving recommendations. As an example, you can imagine the recommendations you provide could be used on an e-commerce website for customers.

**Data**

The data set is in the zipped folder. Accounts.csv has the transaction ids and the account numbers. Along with transactions.csv, this will allow you to link customers to all of their transactions. The items\_transactions.csv will allow you to see for each transaction, what are their associated purchased items. Finally, the items\_descriptions.csv will allow you to link the item id with the item’s description.

Here are some ideas that can help you with the recommendation, but they are not at all exhaustive, and you can be creative with what you do. You can pick multiple, or one and they can be from this list or whatever you pick.

* Data Exploration (correlations, regressions, interesting phenomena, etc)
* Natural Language Processing (NLP models both simple or complex)
* Machine Learning Models (Classification or Regression)
* Clustering Analysis of transactions or items
  + Classify product catalog entries into consistent groupings, for example ketchup is closer to tomatoes than pens
* Create a dashboard
  + Make it easy for a business analyst or product manager to investigate the dataset or your recommendations

The schema for the data set is at the bottom of the document.

**Technologies**

For each of the important steps for a data scientist above, here are some technologies/strategies you could use to start your research. Feel free to pull something outside of this list.

1. Develop a goal or problem statement (can come after step 4 when you explore data more)
   1. Feel free to pull from the data section’s ideas on what to do
2. Ingesting data from a data source
   1. Set up a pipeline, so that you can ingest not only our data set, but other data sets in the future
   2. Some API request that can query an endpoint for the data set
      1. An API endpoint to pass the data to your own request (to simulate when others will send API requests to you)
   3. Pandas, R, or whatever language to read from CSV to read in from a CSV
3. Validating and cleaning the data
   1. Python packages like
      1. Pandas
      2. Pydantic
   2. SQL
   3. R packages from tidyverse
      1. Dplyr
      2. Tidyr
4. Explore the data set
   1. Python packages like
      1. Pandas (profiling, etc)
      2. Visualization platform
         1. Matplotlib
         2. Seaborn
         3. Altair, or other platform
   2. R packages for visualization like ggplot2
5. Extract and transform data that you need
   1. Pandas (Python)
   2. Tidyverse
      1. Dplyr
      2. Tidyr
6. Model the data/Gather some insight
   1. Python Packages
      1. Machine Learning
         1. StatsModels
         2. ScipPy
         3. Scikit-learn
      2. Neural Network
         1. TensorFlow
         2. PyTorch
      3. Time Series
         1. Tsfresh
         2. autots
   2. R Packages
      1. Classification and Regression Training (Caret)
      2. CRAN Time Series
7. Interpret the model/data findings and present in a meaningful way
   1. Website
      1. Django
      2. Dash
      3. Flask
      4. Streamlit
   2. API
      1. FastAPI
      2. Flask
   3. Dashboard
      1. Dash
      2. Jupyter Dashboards
   4. Presentation
      1. Powerpoint Presentation

**Schema**

accounts.csv (75,252 rows):

* **global\_transaction\_id:** Global transaction identifier for this dataset
* **ticket\_num:** Original ticket number for the transaction
* **account\_num\_hash:** Hashed account number to link transactions

transactions.csv (124,934 rows):

* **global\_transaction\_id:** Global transaction identifier for this dataset
* **store\_num:** Physical store number
* **ticket\_num:** Original ticket number for the transaction
* **date:** Date of the transaction (YYYY-MM-DD)
* **transaction\_start\_time:** Time the transaction began (HH:MM:SS)
* **transaction\_end\_time:** Time the transaction concluded (HH:MM:SS)
* **num\_items:** The sum of ‘qty\_sold’ of items that make up this transaction. Note: if qty\_is\_weight = 1, the total is incremented by 1 and not the weight value.
* **ticket\_total\_value:** The total value in cents of the transaction/ticket

items\_transactions.csv (1,293,520 rows):

* **global\_transaction\_id:** Global transaction identifier for this dataset
* **item\_id:** Unique item identifier
* **dept\_num:** Unique department identifier
* **qty\_sold:** The total number of this item being sold, can be weight if qty\_is\_weight = 1, else it’s a count
* **item\_price:** Price of the item in cents
* **qty\_is\_weight:** Boolean signifying if the qty\_sold field is weight (1) or count (0)
* **ticket\_num:** Original ticket number for the transaction
* **date:** Date of the transaction
* **time\_scanned:** The time that the particular item was scanned/rung up

items\_descriptions.csv (48,551 rows):

* **item\_id**: Unique item identifier
* **description**: Item description
* **ecomm\_description**: E-commerce item description, not available for all items
* **category**: A category id number - we don’t have the mapping for what the id stands for however. Could help with categorization.
* **item\_type**: Another id number for which we don’t have the mapping for what it actually means. Could help with categorization.
* **upc**: The given Universal Product Code. However - does not seems to be globally registered UPCs. Item\_id is derived from this column.

**Dataset Notes:**  
• The dataset containing item names (items\_descriptions.csv) is large! If you want to work with a smaller set to start, focus on the items that appear in the transactions data first.  
• If using pandas, it is recommended to read in the CSVs with dtype=str to avoid auto-formatting to ints. You will notice that Excel will do this as well. Use a text editor to view the raw data.  
• Not all items in items\_transactions.csv have a description in items\_descriptions.csv (but about 97% of them do).