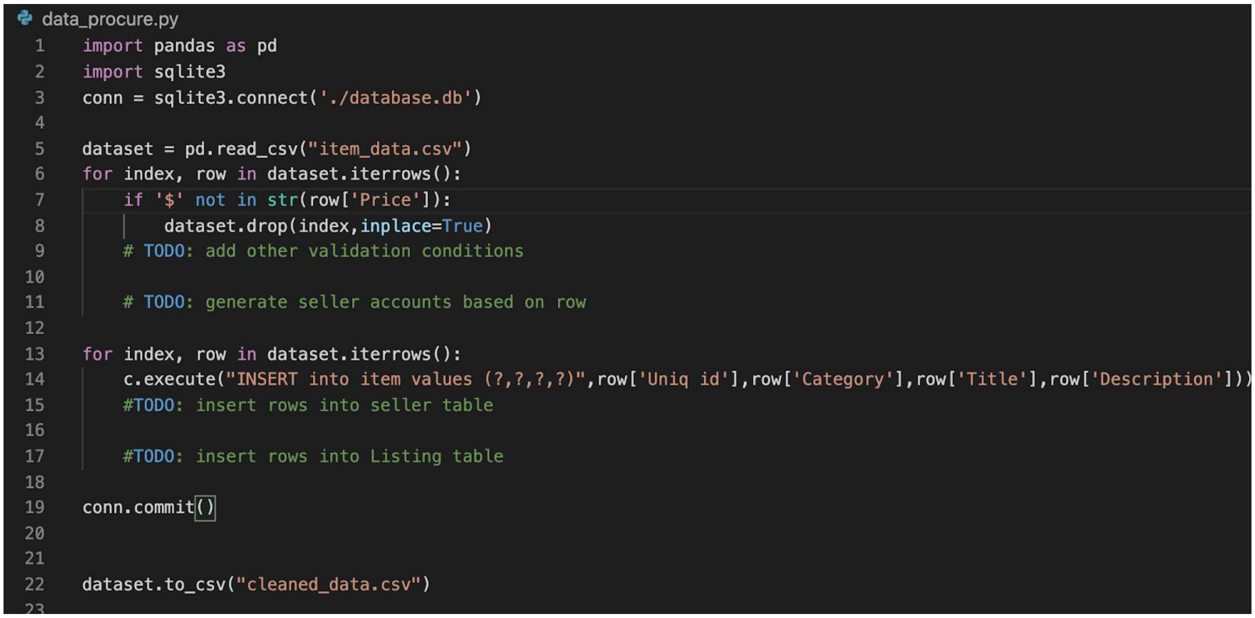
MS2 Submission

Mariusz Derezinski-Choo, Nicholas Aldana, Robert Chen, Justin Du, Karoline Xiong

Data Procurement Plan:

We have found a dataset of 30,000 amazon items from kaggle, available at https://www.kaggle.com/promptcloud/amazon-product-details. To procure the data, we are writing a python script that will load the dataset using the pandas package, and validate each row of the dataset to make sure it is complete. For example, by visually inspecting the dataset we have noticed that some rows do not have a properly formatted price, or may be missing an image url. We will validate each row against a set of conditions to ensure that the data entry is complete, and delete all the rows that are incomplete. We will also have to dynamically a table of storefront rows based on the sellers listed in the dataset. Each item in our dataset has anywhere between 1-3 sellers listed, so we will keep track of all the sellers we encounter and generate emails, passwords, and balances for each of them. Then, we can use sqlalchemy to insert all of these rows into the database. A rough outline of the script is shown below.



data_procure.py

```python
import pandas as pd
import sqlite3
conn = sqlite3.connect('./database.db')

dataset = pd.read_csv("item_data.csv")
for index, row in dataset.iterrows():
    if '$' not in str(row['Price']):
        dataset.drop(index,inplace=True)
    # TODO: add other validation conditions

    # TODO: generate seller accounts based on row

for index, row in dataset.iterrows():
    c.execute("INSERT into item values (?,?,?,?)",row['Uniq id'],row['Category'],row['Title'],row['Description']))
    #TODO: insert rows into seller table

    #TODO: insert rows into Listing table

conn.commit()


dataset.to_csv("cleaned_data.csv")
```

Changes from MS2:

- User table was split into Buyer and Seller tables. The schema for each remained largely the same except sellers do not have a photo.
- The selling table was renamed to Listing to provide more clarity.