DSCI 551 - Data Science Challenge

Yuchen Zhang, Zian Fan

DIRECTIONS:
Choose one of the three Data Science Challenges below. For your chosen Challenge, complete the prescribed data upload and cleaning steps, and then choose and complete one of the Options provided.

Challenge: Walmart Affiliate API

The product catalog data exists as json files where each file contains a page of 200 items. Also the taxonomy data exposes the category taxonomy used by Walmart.com to categorize items. The dataset can be found here: https://drive.google.com/drive/folders/1WHRy07hECWWJGjJB1wPCvVXJxi-AQf3K?usp=sharing

1. Read in the product catalog json files and taxonomy json file to create a pandas Dataframe. Be sure to include at least the following columns:
    a. itemId
    b. name
    c. msrp
    d. salePrice
    e. categoryPath
    f. shortDescription/longDescription
    g. brandName
    h. standardShipRate/twoThreeDayShipRate
    i. bestMarketplacePrice
    j. stock
    k. numReviews
2. Data Cleaning
    a. For sellerInfo, you can read the json objects into separate columns;
    b. Fix missing data with nulls. Also remove duplicate and invalid data;
    c. Change the columns to correct dtypes.

3. **(OPTION 1)** Exploratory Data Analysis: Use python, pandas, and any other libraries to answer the following questions:
    a. How many different main categories are in the datasets? (summarize from categoryPath)
    b. How many percent of products that are in stock, not available, limited supply?

c. Which 20 in-stock products(itemId) have the largest discount?
d. Which 10 products(itemId) have most reviews?
e. Which 5 products have the longest names?
f. Which product's description has the highest numerical frequency? I.e. which product has the highest number to words frequency?

4. **(OPTION 2)** Data Integration/Entity Resolution:
    a. Read in the taxonomy data into a pandas dataframe.
    b. Try to create your own way to link the taxonomy data with the categoryPath value in each product's json file.