DATA INGEST INSTRUCTIONS

AMERICAN COMMUNITY SURVEY (ACS) DATA TABLES

1. Go to https://www1.nyc.gov/site/planning/planning-level/nyc-population/american-community-survey.page.
2. Under the "Profiles" Tab, select 2017. Download the social, economic, and demographic profiles by Neighbourhood Tabulation Areas for 2013-14-15-16-17.
3. Open each of the Excel workbooks. To avoid formatting issues, save the sheets containing the data (not the dictionary or about sheets) as tab-delimited text files rather than csv files. You should end up with three text files : soc_2017_acs5yr_nta.txt, soc_2017_acs5yr_nta.txt, and soc_2017_acs5yr_nta.txt.
4. Copy the files to your Dumbo home directory, e.g.:
   $ scp ~/Downloads/demo_2017_acs5yr_nta.txt ~/Downloads/econ_2017_acs5yr_nta.txt ~/Downloads/soc_2017_acs5yr_nta.txt <YOUR_NETID>@dumbo.es.its.nyu.edu:~
5. Log into your Dumbo machine, create an HDFS project directory to hold the data, and then put the files in HDFS:
   $ hdfs dfs -mkdir /project
   $ hdfs dfs -put demo_2017_acs5yr_nta.txt project/
   $ hdfs dfs -put econ_2017_acs5yr_nta.txt project/
   $ hdfs dfs -put soc_2017_acs5yr_nta.txt project/

RETAIL FOOD STORES

1. Go to https://data.ny.gov/Economic-Development/Retail-Food-Stores/9a8c-vfzj.
2. Export the data as a csv file.
3. On opening the csv file with a text editor, you will find that the data is poorly formatted. Open the csv file in Excel. Use the CLEAN function on the "Location" column to fix cells that contain line breaks. Save the file as a tab-delimited text file, and as a csv file.
4. Copy the files to your Dumbo home directory:
   $ scp ~/Downloads/Retail_Food_Stores.txt ~/Downloads/Retail_Food_Stores.csv <YOUR_NETID>@dumbo.es.its.nyu.edu:~
5. Log into your Dumbo machine and put your files into the HDFS project directory:
   $ hdfs dfs -put Retail_Food_Stores.txt project/
   $ hdfs dfs -put Retail_Food_Stores.csv project/

NEIGHBOURHOOD TABULATION AREAS

1. Go to https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq.
2. Export the boundaries as a shapefile, and unzip the folder. Remove whitespaces from the folder name.
3. Copy the folder to your Dumbo home directory:
   $ scp -r ~/Downloads/NeighborhoodTabulationAreas <YOUR_NETID>@dumbo.es.its.nyu.edu:~
4. Put your folder into the HDFS project directory:
   $ hdfs dfs -put NeighborhoodTabulationAreas project/
5. You will also need the area of each neighbourhood tabulation area. On your local machine, or in NYU's Virtual Lab, open the shapefile (.shp) in ArcGIS Map and then save the attribute table as a csv file.
6. Copy the file to your Dumbo home directory:
   $ scp ~/Downloads/nta_area.csv <YOUR_NETID>@dumbo.es.its.nyu.edu:~
7. Put the file into the HDFS project directory:
   $ hdfs dfs -put nta_area.csv project/

FOOD AVAILABILITY PER CAPITA

1. Go to https://www.ers.usda.gov/data-products/food-availability-per-capita-data-system/.
2. Under the Food Availability section, download the Vegetables (frozen) excel file. This file has more precise values than the csv version.
3. Open the file in Excel. Each vegetable has its own sheet. Locate the sheets for the following leafy greens: collard greens, escarole, kale, head lettuce, romaine and leaf lettuce, mustard greens, spinach and turnip greens. For each of them, note the per capita availability in pounds (farm weight, not retail weight) in 2018.
4. Using the information you have noted, create a new csv file, where the first column is the name of the leafy green and the second is the per capita availability in pounds.
5. Copy the file to your Dumbo home directory:
   $ scp ~/Documents/leafygreens.csv <YOUR_NETID>@dumbo.es.its.nyu.edu:~
6. Create a new folder in your HDFS project directory, and put the file in that folder:

$ hdfs dfs -mkdir project/leafygreens/
$ hdfs dfs -put leafygreens.csv project/leafygreens/

INSTACART GROCERY ORDERS

1. *Go to https://www.instacart.com/datasets/grocery-shopping-2017*
2. Download the dataset, unzip the folder and copy the files to your local machine. There are 6 files in the dataset. Only 4 of the 6 are used for the analytics analysis.
3. Log into to Dumbo, create a folder in your home directory and copy the files from your local machine to your Dumbo directory you just created
   $ mkdir instacart_data
   $ scp departments.csv orders.csv order_products__prior.csv products.csv dumbo:/home/<YOUR_NETID>/instacart_data
4. Create a folder in HDFS for the files and copy the files from your local directory to the folder you just created in HDFS
   $ hdfs dfs -mkdir /user/<YOUR_NETID>/instacart_data
   $ hdfs dfs -put instacart_data/*.csv /user/<YOUR_NETID>/instacart_data