Data Set

**Meetup\_Washington\_DC\_samples.csv**:

Original Data Set consisted of: 200 rows of data, plus 1 header row

**Preprocessing Actions:**

***Changed the Data Element “City” – Data Cleaning:***

|  |  |
| --- | --- |
| Data Element(s) City | Action |
| “20010” | Changed this to “Washington” (1 row) |
| “Washington 20004” | Changed this to “Washington” (1 row) |
| “Washington DC” | Changed this to “Washington” (2 rows) |
| “ Washington DC” | Changed this to “Washington” (1 row) |
| “Washington, DC 20036” | Changed this to “Washington” (1 row) |
| “Washngton DC” | Changed this to “Washington” (1 row) |
| “Washington “ | Changed this to “Washington” (6 rows) |
| “washington” | Changed this to “Washington” (1 row) |
| “Alexandria “ | Changed this to “Alexandria” (1 row) |
| “FAIRFAX” | Changed this to “Fairfax” (1 row) |
| “Herndon “ | Changed this to “Herndon” (1 row) |
| “manassas” | Changed this to “Manassas” (1 row) |
| “MD” | Changed to “Timonium” (1 row) |
| “Silver spring” | Changed this to “Silver Spring” (1 row) |
| “Sterling “ | Changed this to “Sterling” (1 row) |
| “NA” | **Deleted** per instructions, though we note these are all in Washington, DC based on the latitude and longitude. (37 rows) |

* 37 rows were deleted
* Rows changed city name, as described above
* Rows corrected state entries based on city information, removing “NA” and data entry errors – e.g., “Di” meaning “DC” and “VI” meaning “VA”
* New data set consists of 163 rows of data, plus 1 header row

**Categorized the Meetup Activities into 5 Categories:**

Identify 5 Topics (my choice). Add a new column and categorize the rows.

My Topics are:

* Arts (32)– Books, Culture, Film (var. arts), Language
* Food (30) – Food and Drink
* Hobbies (14) – Games and Hobbies, incl. pets
* Local (28) – Local Community (incl politics), Business and Tech (incl speaking, e.g., Toastmasters)
* Wellness (59) – Fitness, Mindfulness, Religious

= 163 records + 1 header (same as before, as it should be)

Data Analysis

Next:

*“Bin” the Data*

Create a 5-bin vector of topics

City -> [ topic0, topic1, topic2, topic3, topic4 ]

Create a Python Dictionary:

Each city has an entry with [CityName, [ topic0, topic1, topic2, topic3, topic4 ] ]

Calculate “reversed HHI” for each city 🡪 this is a quantification of the cultural diversity for each city.

Then: Write a 1-page report that includes a table that compares the reversed HHI between the cities and discuss the results (e.g., why certain cities would have more diverse Meetup activities while others are not. What're the implications of such results from a S&CC perspective?).

*{Comments: include caveats… the sample size is small, a more granular categorization might provide more insight in a larger data set, more care in categorization (or different categorization) might provide a useful lens…}*

Rough Program Flow

Import CSV to a Pandas data frame in Python

import pandas as pd

df = pd.read\_csv (r'~/Assignment 2 – AIT722/Meetup\_File\_Data.csv')

print (df)

Is there a way to go directly to the dictionary view or do I have to first pull the cities, then parse out the number per line? Seems like there might be a function to do it in one step…??

OR maybe import, then export a new CSV file with only city + category?