readme

File Directories

Data

1. legal_compendium_oct_2019_update_.xlsx
2. master_table1.csv
3. kentuckymod.csv
4. Northdmod.csv

Note: Data #1 (raw data obtained from Urban Institute) is cleaned using Programs #1-#3 into Section_codes_current.xls (Output/Formating Urban/ Section_codes_current.xlsx). Section codes in Section_codes_current.xls are then matched with section codes in Data #2 (Drexel data). However, to match section codes for Kentucky and North Dakota, Data #3 and #4 respectively are used instead of Data #2 due to some data inconsistencies.

Code

1. Step 1a Back(sources) transposed.ipynb
2. Step 1b Back(sources) transposed-Regulation Indicator.ipynb
3. Step 1c Extract_SectionCodes.ipynb

Note: Programs #1-#3 are run to clean Data #1 and extract Section Codes in workable format for further matching with Drexel data.

4. State-wise Notebooks (Folder)

Note: This folder contains code files for merging regulatory data for individual states. Program #5 runs all files in #4 State-wise Notebooks (Folder) together. User will not generally be required to run these files individually, but only refer to them to make any state-specific changes to the merging code. The output for this gets stored in Output #2 (Output/Matching Urban to Drexel) as individual state-wise files. To run Program #5, user will need to change "directory" in all individual state code files in folder Programs #4. This can be done using Program #9.

5. Master code.ipynb
6. Match rate.ipynb
7. Match rate – Type.ipynb
8. Unmatched Section Codes.ipynb

Note: Program #6 and #7 calculate two types of match rates, overall and by regulation type, for the merged results. Output is stored in Output #3. Program #8 identifies and list all the unmatched section codes for future work.

9. Change directory.ipynb
10. Final Formating.ipynb

Output Folders

1. Formatting Urban
   1.1. Step 1a Back(sources) transposed.csv
   1.2. Step 1b Back(sources) transposed - Regulation Indicator.csv
   1.3. Section_codes_current.xlsx

2. Matching Urban to Drexel Formatted
    2.1. Section_codes_merged_alabama.xlsx
        .
        .
        .
    2.51. Section_codes_merged_wyoming.xlsx
3. Match Rate
    3.1. match_percentages.csv
    3.2. match_percentage_by_type_pivot_new.csv
    3.3. Unmatched Section List.xlsx

Procedure – data/program/output

1. Use Data #1 and Programs #1 – 3 (Step 1a – 1c) to create Output #1.1-1.3 respectively.
2. Use Program #5 to run all state-wise merge on Data #2 and Output #1.3 through the master code method. Remember to change "directory" in all individual state code files in folder Programs #4 to run this. Program #9 does this for all state code files together.
3. Use Program #6 to get excel table Output #3.1 which contains match rates for each state.
4. Use Program #7 to get excel table Output #3.2 which contains match rates for each state broken by regulatory type.
5. Use Program #8 to get excel table Output #3.3 that contains unmatched section codes for every state.

Note: Programs #6, #7, #8 use Output #2. So user can choose to skip the time-taking state-wise matching mechanism in Program #5.

The project can be found in the folder Urban Final. Inside Urban Final are 3 folders organized as follows:

1. Code
    1.1. Step 1a Back(sources) transposed.ipynb
    1.2. Step 1b Back(sources) transposed-Regulation Indicator.ipynb
    1.3. Step 1c Extract_SectionCodes.ipynb
    1.4. State-wise Notebooks (Folder)
    1.5. Master code.ipynb
    1.6. Match rate.ipynb
    1.7. Match rate – Type.ipynb
    1.8. Unmatched Section Codes.ipynb
2. Data
    2.1. legal_compendium_oct_2019_update_.xlsx
    2.2. master_table1.csv
    2.3. kentuckymod.csv
    2.4. northdmod.csv
3. Output
    3.1. Formating Urban
        3.1.1.Step 1a Back(sources) transposed.csv
        3.1.2.Step 1b Back(sources) transposed - Regulation Indicator.csv
        3.1.3.Section_codes_current.xlsx

3.2. Matching Urban to Drexel

3.2.1. Section_codes_merged_alabama.xlsx

.

.

.

3.2.51. Section_codes_merged_wyoming.xlsx

3.3. Match Rate

3.3.1. match_percentages.csv

3.3.2. match_percentage_by_type_pivot_new.csv

3.3.3. Unmatched List.xlsx

Code: This folder contains 9 code files and 1 folder containing 51 code files.

1. **Step 1a Back(sources) transposed.ipynb:** This notebook uses Data #1 and transposes the file. The output file is "Step 1a Back(sources) transposed.csv".
2. **Step 1b Back(sources) transposed-Regulation Indicator.ipynb**: This notebook uses "Step 1a Back(sources) transposed.csv". It then creates a column "Regulation Indicator". Column "Regulation Indicator" takes the values Yes or No (Yes, if the particular state has some laws/regulations in place on the said regulatory aspect/No, if not). The output file is "Step 1b Back(sources) transposed - Regulation Indicator.csv".
3. **Step 1c Extract_SectionCodes.ipynb:** This notebook uses "Step 1b Back(sources) transposed - Regulation Indicator.csv" and extracts Section Codes from the "Notes" column. The output file is "Section_codes_current.xlsx" which is the final cleaned Urban Institute data used in the matching process.
4. **State-wise Notebooks (Folder)**: This folder contains code files for merging regulatory data for individual states. User will not generally require to run these files individually, but only refer to them to make any state-specific changes to the merging code.
    a. Usage:
        - Navigate to the specific state code file (e.g., Merging-ARIZONA.ipynb).
        - Execute the notebook to perform data merging for the respective state.
        - The merged data will be saved in Output #2.
    b. Notes:
        - The state-specific code files require two input files: Section_codes_current.xlsx and master_table1.csv.
        - Section_codes_current.xlsx is a cleaned version of the Legal Compendium data file put together by Urban Institute. master_table1.csv is the regulatory data (section codes and text) scraped by Drexel University.
5. **Master code.ipynb**: This notebook automates the execution of state-specific codes.
    a. Usage:
        - Open the Master Code.ipynb notebook located in the Code folder.
        - Execute the notebook to run all state-specific codes sequentially.
        - The output files (state-specific merged files) will be saved in folder Output #3.
6. **Match rate.ipynb:** This notebook calculates overall match rate between Section Codes in Urban Data and Section Codes with text in Drexel data.
    a. Usage:
        - Open the Match rate.ipynb notebook located in Code folder.

- Execute the notebook to calculate match rates from the merged data.
- The match rates will be saved in folder Output #3 as match_percentages.csv.

b. Notes:

- The overall match rate is calculated as a ratio of matched "section" (Section Codes from Drexel data) to "Section Code" (Section Codes from Urban Institute data).
- The formula is: match_percentage = (total_non_empty_section / total_non_empty_section_code) * 100

7. **Match rate – Type.ipynb:** This notebook calculates match rate by regulation type between Section Codes in Urban Data and Section Codes with text in Drexel data.
   a. Usage:
      - Open the Match rate - Type.ipynb notebook located in Code folder.
      - Execute the notebook to calculate match rates from the merged data.
      - The match rates will be saved in folder Output #3 as match_percentage_by_type_pivot_new.csv.
   b. Notes:
      - The by-type match rate is calculated as a ratio of matched "section" (Section Codes from Drexel data) to "Section Code" (Section Codes from Urban Institute data), but within regulation types.
8. **Unmatched Section Codes:** This notebook lists unmatched section codes for future work.
   a. Usage:
      - Open the Unmatched Section Codes.ipynb notebook located in Code folder.
      - Execute the notebook to unmatched section codes.
      - The file will be saved in folder Output #3 as Unmatched List.xlsx.
9. Final Change directory.ipynb: This notebook changes directory for all individual state matching files centrally.
10. Final Formating.ipynb: This notebook carries out final formating of merged files as suggested by Urban Institute,

Data Dictionary for Output #2 (which are the final matched excel files).

1. `Unnamed: 0`: Index. (Links to index in Section_codes_current.xlsx)

2. `Unnamed: 0.1`: Index. (Links to within state index in Section_codes_current.xlsx)

3. `STATE`: The US state name. (Extracted from Urban data)

4. `Regulatory Type`: The type or category of regulation. (Created in Step 1c Extract_SectionCodes.ipynb)

5. `Regulatory Type Full`: The regulatory type along with regulator type. (Extracted from Urban data)

6. `Regulatory Body`: The governing body responsible for the regulation. (Created in Step 1c Extract_SectionCodes.ipynb)

7. `Regulation Indicator`: An indicator of whether the regulation exists or not. (Created in Step 1b Back(sources) transposed-Regulation Indicator.ipynb)

8. `Notes`: Additional notes or comments related to the regulation. (Extracted from Urban data)

9. `Year legislation originally enacted`: The year when the legislation associated with the regulation was first enacted. (Created in Step 1c Extract_SectionCodes.ipynb – Empty)

10. `Year legislation changed`: The year when the legislation associated with the regulation was changed. (Created in Step 1c Extract_SectionCodes.ipynb – Empty)

11. `Type of Change`: The type of change made to the legislation. (Created in Step 1c Extract_SectionCodes.ipynb – Empty)

12. `Section Code`: The Section Code for the regulation. (Extracted from Urban data)

13. `State`: The US state name (Extracted from Drexel data)

14. `section`: The Section Code for the regulation. (Extracted from Drexel data)

15. `text`: The text under the Section Code. (Extracted from Drexel data)

16. `ID-State`: Identifier for row. (Created in Final Formating.ipynb)

17. `ID`: Identifier from Section_codes_current.xlsx.

18. `Regulatory Type Abbr`: Abbreviation of the regulatory type. (Created in Final Formating.ipynb)