Emily’s Project 4 Data Cleaning Notes

1. May 8, 2024
   1. daycare slots
      1. I went through all the files in all the subfolders of the download and found 3 files with data in them, and many copies of a data dictionary. I read all 3 data files into python and ran tests on them to determine whether they were the same. Files 1 and 2, which were both originally saved as CSVs in the download, passed easily. File 3 presented difficulties. After accounting for datetime formatting, null values, and ASCII characters, I concluded there were at most 18 discrepant cells. Out of a 20100x28 dataframe, this discrepancy did not warrant further examination. Files 2 and 3 were discarded, and file 1 used for analysis.
   2. poverty rate
      1. After doing extensive research on the daycare data, I decided that, for datasets from the same source, I could make decisions based on similar patterns without repeating the extensive validation work. Therefore, I used git bash to check that the two CSV data files in the “data” folder were the same, which they were, and then just visually assessed “poverty-rate-200-federal-poverty-level.xlsx.” I could not immediately discern what was going on with that file, but given that none of the data files were actually distinct in the last batch, I concluded it was unlikely that this one was either. The file “poverty-rate-200-fpl-and-child-under-18-poverty-rate-by-california-regions-3.xlsx” is a duplicate, verified in git bash.
   3. abortion services funded
      1. I went through the files in the download and found two distinct spreadsheets, one containing counts of services provided per county, and the other containing the amounts spent on services per county. I will include both. All other files were either metadata or state-level.
   4. educational attainment
      1. This download followed the same pattern of duplicates, but thankfully only one copy per file type. I reserved the CSV and discarded the XLSX.
   5. ecig availability
      1. This download had multiple duplicates per file type, as verified in git bash, and duplicates within and across the “data” and “original” folders. After briefly comparing each unique file, and assuming that the same patterns would hold as for the other downloads from this source, I identified one data file and one data dictionary to keep, and discarded the rest.
   6. income, poverty, and health insurance
      1. This download only contains state-level data. I did not wade through it for anything, but there are a few interesting statistics that we may want to pull out for context.