Matt Burns

DSC 540

Bellevue University

**Preparation**

There were two main steps to prepare for the data cleansing exercise. First, I pulled in many of the packages that were helpful in 530. This included NumPy, random, pandas and MATLAB. The second step was to pull in the data which was a census estimate from 2018 containing population data cut by age, race and cohabitation status. During this process I knocked out column called “SUMLEV” that contained the number 50 for every row.

**Header Replacement**

For this I replaced the names given by the census bureau with names that were user friendly. To do this I used the *rename* function for each column. I used *describe* and *head* to check my work.

**Data Format**

There were a few columns that need to be reformatted as well as a column that needed its bin range to be explicitly stated. When I used *dtypes*, I could see that many rows were properly stored as integers. The State and County numbers needed to be changed to strings via the *astype* function. Year needed to be changed from 11 to 7/1/2018 and put into the date format via a pandas *to\_datetime* function. Finally, the age groups needed to be decoded. I used supplemental information from the census bureau and the *replace* function to accomplish this.

**Outlier Identification**

First, I confirmed that they were no blanks. I did this for each column using *isnull* as well as the dataframe in aggregate. Both reported no errors. Next I made box plots cut by age band to see if there were any outliers. I didn’t graph totals or the largest county intentionally. The outliers that were identified across gender or race were from huge counties (e.g. Maricopa, Cook) and were intentionally left in.

**Duplicates Identification**

The data is aggregated at the county level. I checked to make sure that duplicate county names were in different states. Two Lake Counties in Illinois would be a problem. Lake County, Illinois and Lake County, Indiana are not a problem. I used *counter* to identify duplicate names. Then I checked out two instances to make sure they were different states which is acceptable. There are two Crook Counties. One in Wyoming and one in Oregon. There are three Cook counties across Illinois, Minnesota and Georgia. All is well.

**Fuzzy Match Execution**

I used fuzzywuzzy to look for counties that have a similar name Ellis County. The four out of the five selected had “ell” in their name.

Note:

I am running Python 3 so the textbook was of little help to me. That is why submission doesn’t align with the text’s methodology.