DataEng: Data Integration Activity

Priyanka Pandey

This week you will gain hands-on experience with Data Integration by combining data from two distinct sources into a unified DataFrame for analysis.

Submit: Make a copy of this document and use it to record your results. Store a PDF copy of the document in your git repository along with any needed code before submitting for this week.

Your job is to integrate **county-level COVID-19 data** with the **ACS Census Tract data for 2017** to build a model that allows you to relate COVID numbers with economic data such as population, per capita income and poverty level. To do this you should build a pandas DataFrame that has a row per USA county (there are more than 3000 counties in the USA) and includes the following columns:

County - name of the county
State - name of the state in which the county resides
TotalCases - total number of COVID cases for this county as of February 20, 2021
Dec2020Cases - number of COVID cases recorded in this county in December of 2020
TotalDeaths - total number of COVID deaths for this county as of February 20, 2021
Dec2020Deaths - number of COVID deaths recorded in this county in December of 2020
Population - population of this county
Poverty - % of people in poverty in this county
PerCapitaIncome - per capita personal income for this county

We hope that you make it all the way through to the end. Regardless, use your time wisely to gain python programming experience and learn as much as you can about building integrated multi-source data models using python and pandas.

For this activity you should use whichever environment is convenient for you to develop with python 3 and pandas. You are not required to use GCP, but you can use it if you prefer.

Submit: In-class Activity Submission Form

A. Aggregate Census Data to County Level

Your integration will use two different dimensions: location (as indicated by state and county) and time. You should greatly simplify your processing and reduce your time by pre-processing your data along each of these dimensions.

The ACS data is separated into “Census Tracts” which are regions within counties that correspond to groups of approximately 4000 people. The Census Bureau defines these to help organize the actual job of collecting census data, but this grouping can make your Data Engineering job more more challenging. This level of detail is not needed for your county-level analysis, and you can greatly decrease your efforts by aggregating per-tract data to the county level.

Create a python program that produces a one-row-per-county version of the ACS data set. To do this you will need to think about how to properly aggregate Census Tract-level data into County-level summaries.

In this step you can also eliminate unneeded columns from the ACS data.

Question: Show your aggregated county-level data rows for the following counties: Loudoun County Virginia, Washington County Oregon, Harlan County Kentucky, Malheur County Oregon

State	County	TotalPop	Poverty	IncomePerCap	Poverty(In Num)	TotalIncome
Oregon	Washington County	572071	10.321202	35369.047499	59044.602	2.023361e+10
Virginia	Loudoun County	374558	3.689598	50455.645745	13819.683	1.889857e+10
Kentucky	Harlan County	27548	35.669482	15456.971032	9826.229	4.258086e+08
Oregon	Malheur County	30421	24.298225	17567.504323	7391.763	5.344210e+08

B. Simplify the COVID Data

You can simplify the COVID data along the time dimension. The COVID data set contains day-level resolution data from (approximately) March of 2020 through February of 2021. However, you will only need four data points per county: total cases, total deaths, cases reported during December of 2020 and deaths reported during December 2020.

Create a python program that reduces the COVID data to one line per county.

Question: Show your simplified COVID data for the counties listed above.

a)
```
#TotalCases - total number of COVID cases for this county as of February 20, 2021
#Dec2020Cases - number of COVID cases recorded in this county in December of 2020
```

State	County	date	total cases	total deaths
Oregon	Washington County	2021-02-20	20866	209.0
Virginia	Loudoun County	2021-02-20	22557	199.0
Kentucky	Harlan County	2021-02-20	2352	68.0
Oregon	Malheur County	2021-02-20	3331	58.0

b) `#TotalDeaths - total number of COVID deaths for this county as of February 20, 2021`
`#Dec2020Deaths - number of COVID deaths recorded in this county in December of 2020`

State	County	date	dec_cases	dec_deaths
Oregon	Washington County	2020-12-31	16070	142.0
Virginia	Loudoun County	2020-12-31	14169	159.0
Kentucky	Harlan County	2020-12-31	1538	18.0
Oregon	Malheur County	2020-12-31	2914	50.0

C. Integrate COVID Data with ACS Data

Create a single pandas DataFrame containing one row per county and using the columns described above. You are free to add additional columns if needed. For example, you might want to normalize all of the COVID data by the population of each county so that you have a consistent “number of cases/deaths per 100000 residents” value for each county.

Question: List your integrated data for all counties in the State of Oregon.

State	County	TotalPop	Poverty	IncomePerCap	Poverty(In Num)	date	dec_cases	dec_deaths	date	total cases	total deaths
Oregon	Baker County	3344.0	17.769557	29594.000000	594.214	2020-12-31	472.0	5.0	2021-02-20	629.0	7.0
	Benton County	9020.0	31.088814	38541.960532	2804.211	2020-12-31	1347.0	11.0	2021-02-20	2248.0	16.0
	Clackamas County	8815.0	18.001872	58235.322292	1586.865	2020-12-31	10058.0	114.0	2021-02-20	13196.0	172.0
	Clatsop County	4866.0	12.900000	28006.476367	627.714	2020-12-31	553.0	3.0	2021-02-20	766.0	6.0
	Columbia County	6964.0	20.100000	29617.062321	1399.764	2020-12-31	837.0	14.0	2021-02-20	1208.0	21.0
	Coos County	7505.0	22.558561	24989.045703	1693.020	2020-12-31	756.0	9.0	2021-02-20	1347.0	18.0
	Crook County	7132.0	17.752370	24280.000000	1266.099	2020-12-31	448.0	7.0	2021-02-20	765.0	18.0
	Curry County	5440.0	20.352849	29275.000000	1107.195	2020-12-31	278.0	3.0	2021-02-20	394.0	6.0
	Deschutes County	14329.0	14.999079	45883.227162	2149.218	2020-12-31	3976.0	22.0	2021-02-20	5839.0	58.0
	Douglas County	8008.0	20.200000	33676.885365	1617.616	2020-12-31	1387.0	39.0	2021-02-20	2312.0	51.0
	Gilliam County	1910.0	9.900000	24178.000000	189.090	2020-12-31	37.0	1.0	2021-02-20	53.0	1.0
	Grant County	5398.0	11.400000	26466.000000	615.372	2020-12-31	170.0	1.0	2021-02-20	221.0	1.0
	Harney County	5071.0	19.300000	23278.000000	978.703	2020-12-31	134.0	2.0	2021-02-20	266.0	6.0
	Hood River County	6862.0	18.301006	34140.000000	1255.815	2020-12-31	816.0	14.0	2021-02-20	1057.0	29.0
	Jackson County	10929.0	18.146253	31145.000000	1983.204	2020-12-31	5884.0	72.0	2021-02-20	8115.0	108.0
	Jefferson County	5092.0	30.552082	27894.000000	1555.712	2020-12-31	1425.0	17.0	2021-02-20	1918.0	27.0
	Josephine County	9210.0	25.770489	24814.564169	2373.462	2020-12-31	1193.0	22.0	2021-02-20	2266.0	48.0
	Klamath County	5465.0	22.661354	25960.638243	1238.443	2020-12-31	1910.0	18.0	2021-02-20	2752.0	54.0
	Lake County	5213.0	22.000000	20773.000000	1146.860	2020-12-31	197.0	4.0	2021-02-20	373.0	6.0
	Lane County	8655.0	66.362103	31604.000000	5743.640	2020-12-31	6929.0	92.0	2021-02-20	10033.0	121.0
	Lincoln County	4085.0	22.134051	22244.418360	904.176	2020-12-31	880.0	17.0	2021-02-20	1120.0	19.0
	Linn County	10045.0	14.835839	26674.000000	1490.260	2020-12-31	2650.0	32.0	2021-02-20	3533.0	55.0
	Malheur County	5822.0	30.043988	15641.183786	1749.161	2020-12-31	2914.0	50.0	2021-02-20	3331.0	58.0
	Marion County	10676.0	29.898839	30128.718153	3192.000	2020-12-31	13928.0	210.0	2021-02-20	18171.0	280.0
	Morrow County	8308.0	16.000000	20255.000000	1329.280	2020-12-31	815.0	8.0	2021-02-20	1031.0	13.0
	Multnomah County	11030.0	30.800000	56678.037715	3397.240	2020-12-31	25290.0	394.0	2021-02-20	31526.0	516.0
	Polk County	10051.0	18.561287	33400.516566	1865.595	2020-12-31	1977.0	30.0	2021-02-20	2978.0	42.0
	Sherman County	1635.0	13.700000	34226.000000	223.995	2020-12-31	31.0	0.0	2021-02-20	52.0	0.0
	Tillamook County	8080.0	17.900000	23415.000000	1446.320	2020-12-31	308.0	0.0	2021-02-20	403.0	2.0
	Umatilla County	9236.0	24.954060	21463.000000	2304.757	2020-12-31	5640.0	57.0	2021-02-20	7580.0	80.0
	Union County	3825.0	25.336000	35782.000000	969.102	2020-12-31	980.0	14.0	2021-02-20	1264.0	19.0
	Unknown County	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2020-11-29	1.0	0.0
	Wallowa County	3173.0	11.200000	26189.000000	355.376	2020-12-31	76.0	3.0	2021-02-20	142.0	4.0
	Wasco County	4439.0	19.378126	25911.302546	860.195	2020-12-31	905.0	22.0	2021-02-20	1218.0	25.0
	Washington County	12595.0	17.998253	53974.000000	2266.880	2020-12-31	16070.0	142.0	2021-02-20	20866.0	209.0
	Wheeler County	1415.0	20.600000	21268.000000	291.490	2020-12-31	17.0	1.0	2021-02-20	22.0	1.0
	Yamhill County	9940.0	19.011911	32092.069416	1889.784	2020-12-31	2641.0	35.0	2021-02-20	3716.0	62.0

D. Analysis

For each of the following, determine the strength of the correlation between each pair of variables. Compute the correlation strength by calculating the Pearson correlation coefficient R for pairs of columns in your DataFrame. For example, if you have a DataFrame df with each row

representing a distinct county, and columns named ‘TotalCases’ and ‘Poverty’, then you can compute R like this:

```
R = df['TotalCases'].corr(df['Poverty'])

0.3318299700396194
```

For any R that is > 0.5 or < -0.5 also display a scatter plot (see pandas scatterplot and seaborn documentation for information about how to display scatter plots from DataFrame data).

The COVID numbers should be normalized to population (# of cases per 100,000 residents) so that different sized counties are comparable. So for example, “COVID total cases” below really means “((COVID total cases in county * 100000) / population of county)”.

1. Across all of the counties in the State of Oregon
    a. COVID total cases vs. % population in poverty
    b. COVID total deaths vs. % population in poverty
    c. COVID total cases vs. Per Capita Income level
    d. COVID total deaths vs. Per Capita Income level
    e. COVID cases during December 2020 vs. % population in poverty
    f. COVID deaths during December 2020 vs. % population in poverty
    g. COVID cases during December 2020 vs. Per Capita Income level
    h. COVID cases during December 2020 vs. Per Capita Income level

```
#COVID total cases vs. % population in poverty
R1 = final_result['CovidTotalCases'].corr(final_result['Poverty'])
print(R1)
#COVID total deaths vs. % population in poverty
R2 = final_result['CovidDeathCases'].corr(final_result['Poverty'])
print(R2)
#COVID total cases vs. Per Capita Income level
R3 = final_result['CovidTotalCases'].corr(final_result['IncomePerCap'])
print(R3)
#COVID total deaths vs. Per Capita Income level
R4 = final_result['CovidDeathCases'].corr(final_result['IncomePerCap'])
print(R4)
#COVID cases during December 2020 vs. % population in poverty
R5 = final_result['CovidDecTotalCases'].corr(final_result['Poverty'])
print(R5)
#COVID deaths during December 2020 vs. % population in poverty
R6 = final_result['CovidDecDeathCases'].corr(final_result['Poverty'])
print(R6)
#COVID cases during December 2020 vs. Per Capita Income level
R7 = final_result['CovidDecTotalCases'].corr(final_result['IncomePerCap'])
print(R7)
#COVID cases during December 2020 vs. Per Capita Income level
R8 = final_result['CovidDecDeathCases'].corr(final_result['IncomePerCap'])
print(R8)
```

```
0.4107959010799331
0.39261885108213307
0.67217844804829
0.6045701150884264
0.3846509420347484
0.4074096277848708
0.664637225455696
0.5701599294577229
```

2. Across all of the counties in the entire USA
    a. COVID total cases vs. % population in poverty
    b. COVID total deaths vs. % population in poverty
    c. COVID total cases vs. Per Capita Income level
    d. COVID total deaths vs. Per Capita Income level
    e. COVID cases during December 2020 vs. % population in poverty
    f. COVID deaths during December 2020 vs. % population in poverty
    g. COVID cases during December 2020 vs. Per Capita Income level
    h. COVID cases during December 2020 vs. Per Capita Income level

```
#COVID total cases vs. % population in poverty
R1 = final_result['CovidTotalCases'].corr(final_result['Poverty'])
print(R1)
#COVID total deaths vs. % population in poverty
R2 = final_result['CovidDeathCases'].corr(final_result['Poverty'])
print(R2)
#COVID total cases vs. Per Capita Income level
R3 = final_result['CovidTotalCases'].corr(final_result['IncomePerCap'])
print(R3)
#COVID total deaths vs. Per Capita Income level
R4 = final_result['CovidDeathCases'].corr(final_result['IncomePerCap'])
print(R4)
#COVID cases during December 2020 vs. % population in poverty
R5 = final_result['CovidDecTotalCases'].corr(final_result['Poverty'])
print(R5)
#COVID deaths during December 2020 vs. % population in poverty
R6 = final_result['CovidDecDeathCases'].corr(final_result['Poverty'])
print(R6)
#COVID cases during December 2020 vs. Per Capita Income level
R7 = final_result['CovidDecTotalCases'].corr(final_result['IncomePerCap'])
print(R7)
#COVID cases during December 2020 vs. Per Capita Income level
R8 = final_result['CovidDecDeathCases'].corr(final_result['IncomePerCap'])
print(R8)
```

```
0.4107959010799331
0.39261885108213307
0.67217844804829
0.6045701150884264
0.3846509420347484
0.4074096277848708
0.664637225455696
0.5701599294577229
```

Note that this exercise does not constitute a competent, thorough statistical analysis of the relationships between immunological data and demographic data. It is just an illustration of the types of computations that might be accomplished with an integrated data set.