DataEng: Data Integration Activity

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

	County	State	Population	Poverty	IncomePerCapita
	Loudoun County	Virginia	374558	3.689598	50455.645745
	Washington County	Oregon	572071	10.321202	35369.047499
	Harlan County	Kentucky	27548	35.669482	15456.971032
	Malheur County	Oregon	30421	24.298225	17567.504323

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.

	County	State	TotalCases	TotalDeaths	Dec2020Cases	Dec2020Deaths
	Loudoun County	Virginia	2496450	35820.0	24401.0	303.0
	Washington County	Oregon	2157339	22455.0	26784.0	245.0
	Harlan County	Kentucky	205984	3994.0	2470.0	33.0
	Malheur County	Oregon	453634	7770.0	5333.0	94.0

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.

	County	State	Population	Poverty	IncomePerCapita	TotalCases	TotalDeaths	Dec2020Cases	Dec2020Deaths
	Baker County	Oregon	15980	15.083855	25820.273154	55586.0	663.0	758.0	9.0
	Benton County	Oregon	88249	22.421152	30872.824361	180225.0	2304.0	2165.0	18.0
	Clackamas County	Oregon	399962	8.976120	37550.849108	1284402.0	20040.0	16606.0	196.0
	Clatsop County	Oregon	38021	12.190090	28114.625523	77666.0	287.0	921.0	3.0
	Columbia County	Oregon	50207	12.315329	28459.688051	105324.0	1363.0	1363.0	17.0
	Coos County	Oregon	62921	17.896488	26007.212997	100097.0	969.0	1199.0	12.0
	Crook County	Oregon	21717	15.320864	24238.814477	55863.0	1134.0	685.0	13.0
	Curry County	Oregon	22377	15.408656	26925.536399	30045.0	393.0	421.0	5.0
	Deschutes County	Oregon	175321	12.100898	31574.934092	509974.0	4141.0	6456.0	37.0
	Douglas County	Oregon	107576	17.025995	25001.732924	174952.0	3983.0	2354.0	57.0
	Gilliam County	Oregon	1910	9.900000	24178.000000	4691.0	76.0	60.0	1.0
	Grant County	Oregon	7209	13.635802	25154.161742	18551.0	94.0	307.0	2.0
	Harney County	Oregon	7195	17.528770	24397.712578	17024.0	291.0	236.0	3.0
	Hood River County	Oregon	22938	12.123145	29594.972796	107383.0	1444.0	1230.0	15.0
	Jackson County	Oregon	212070	16.858350	27080.538534	713288.0	7221.0	9835.0	107.0
	Jefferson County	Oregon	22707	20.694856	22956.835293	200346.0	2630.0	2287.0	28.0
	Josephine County	Oregon	84514	18.646376	24348.609449	153675.0	2638.0	1760.0	26.0
	Klamath County	Oregon	66018	18.688624	23793.066679	224256.0	2857.0	2820.0	22.0
	Lake County	Oregon	7807	20.139311	21004.589343	25357.0	348.0	335.0	5.0
	Lane County	Oregon	363471	19.230471	27032.412179	850956.0	10372.0	11370.0	135.0
	Lincoln County	Oregon	47307	18.376280	25782.113704	153979.0	3117.0	1526.0	32.0
	Linn County	Oregon	121074	16.063929	24448.467359	324636.0	5949.0	4119.0	55.0
	Malheur County	Oregon	30421	24.298225	17567.504323	453634.0	7770.0	5333.0	94.0
	Marion County	Oregon	330453	16.128516	24791.074831	1974030.0	34089.0	23442.0	356.0
	Morrow County	Oregon	11153	14.699050	21742.930153	139209.0	1447.0	1479.0	15.0
	Multnomah County	Oregon	788459	16.474668	34848.165612	3374737.0	58787.0	43115.0	638.0
	Polk County	Oregon	79666	15.639958	25928.364057	268036.0	5480.0	3277.0	49.0
	Sherman County	Oregon	1635	13.700000	34226.000000	5807.0	0.0	54.0	0.0
	Tillamook County	Oregon	25840	15.512717	25458.191138	34370.0	92.0	447.0	0.0
	Umatilla County	Oregon	76736	17.825222	22153.237007	933975.0	10661.0	10042.0	106.0
	Union County	Oregon	25810	17.618597	26585.728710	161223.0	1533.0	1777.0	22.0
	Wallowa County	Oregon	6864	13.748776	26897.389860	13017.0	449.0	147.0	6.0
	Wasco County	Oregon	25687	13.670818	24727.506132	121202.0	3039.0	1443.0	41.0
	Washington County	Oregon	572071	10.321202	35369.047499	2157339.0	22455.0	26784.0	245.0
	Wheeler County	Oregon	1415	20.600000	21268.000000	1454.0	53.0	22.0	1.0
	Yamhill County	Oregon	102366	13.802658	28539.604791	356425.0	6010.0	4409.0	52.0

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'])
```

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

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

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.