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

```
In [76]: result = acs_grp_df.loc[[('Oregon', 'Washington County'),('Virg
         result
```

Out[76]:

State	County	TotalPop	Poverty	IncomePerCap
Oregon	Washington County	572071	10.321202	35369.047499
Virginia	Loudoun County	374558	3.689598	50455.645745
Kentucky	Harlan County	27548	35.669482	15456.971032
Oregon	Malheur County	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.

Out[181]:

State	County	date_x	dec_cases	dec_deaths	date_y	tot_cases	tot_deaths
Oregon	Washington County	2020-12-31	16070	142.0	2021-02-20	20866	209.0
Virginia	Loudoun County	2020-12-31	14169	159.0	2021-02-20	22557	199.0
Kentucky	Harlan County	2020-12-31	1538	18.0	2021-02-20	2352	68.0
Oregon	Malheur County	2020-12-31	2914	50.0	2021-02-20	3331	58.0

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	date_x	dec_cases	dec_deaths	date_y	tot_cases	tot_deaths	TotalPop	Poverty	IncomePerCap	casesPer100k
Oregon	Baker County	2020-12-31	472	5.0	2021-02-20	629	7.0	15980	15.083855	25820.273154	3936.170213
	Benton County	2020-12-31	1347	11.0	2021-02-20	2248	16.0	88249	22.421152	30872.824361	2547.337647
	Clackamas County	2020-12-31	10058	114.0	2021-02-20	13196	172.0	399962	8.976120	37550.849108	3299.313435
	Clatsop County	2020-12-31	553	3.0	2021-02-20	766	6.0	38021	12.190090	28114.625523	2014.676100
	Columbia County	2020-12-31	837	14.0	2021-02-20	1208	21.0	50207	12.315329	28459.688051	2406.038999
	Coos County	2020-12-31	756	9.0	2021-02-20	1347	18.0	62921	17.896488	26007.212997	2140.779708
	Crook County	2020-12-31	448	7.0	2021-02-20	765	18.0	21717	15.320864	24238.814477	3522.585993
	Curry County	2020-12-31	278	3.0	2021-02-20	394	6.0	22377	15.408656	26925.536399	1760.736470
	Deschutes County	2020-12-31	3976	22.0	2021-02-20	5839	58.0	175321	12.100898	31574.934092	3330.462409
	Douglas County	2020-12-31	1387	39.0	2021-02-20	2312	51.0	107576	17.025995	25001.732924	2149.178255
	Gilliam County	2020-12-31	37	1.0	2021-02-20	53	1.0	1910	9.900000	24178.000000	2774.869110
	Grant County	2020-12-31	170	1.0	2021-02-20	221	1.0	7209	13.635802	25154.161742	3065.612429
	Harney County	2020-12-31	134	2.0	2021-02-20	266	6.0	7195	17.528770	24397.712578	3697.011814
	Hood River County	2020-12-31	816	14.0	2021-02-20	1057	29.0	22938	12.123145	29594.972796	4608.073938
	Jackson County	2020-12-31	5884	72.0	2021-02-20	8115	108.0	212070	16.858350	27080.538534	3826.566700
	Jefferson County	2020-12-31	1425	17.0	2021-02-20	1918	27.0	22707	20.694856	22956.835293	8446.734487
	Josephine County	2020-12-31	1193	22.0	2021-02-20	2266	48.0	84514	18.646376	24348.609449	2681.212580
	Klamath County	2020-12-31	1910	18.0	2021-02-20	2752	54.0	66018	18.688624	23793.066679	4168.560090
	Lake County	2020-12-31	197	4.0	2021-02-20	373	6.0	7807	20.139311	21004.589343	4777.763546
	Lane County	2020-12-31	6929	92.0	2021-02-20	10033	121.0	363471	19.230471	27032.412179	2760.330260
	Lincoln County	2020-12-31	880	17.0	2021-02-20	1120	19.0	47307	18.376280	25782.113704	2367.514321
	Linn County	2020-12-31	2650	32.0	2021-02-20	3533	55.0	121074	16.063929	24448.467359	2918.050118
	Malheur County	2020-12-31	2914	50.0	2021-02-20	3331	58.0	30421	24.298225	17567.504323	10949.672923
	Marion County	2020-12-31	13928	210.0	2021-02-20	18171	280.0	330453	16.128516	24791.074831	5498.815263

Marion County	2020-12-31	13928	210.0	2021-02-20	18171	280.0	330453	16.128516	24791.074831	5498.815263
Morrow County	2020-12-31	815	8.0	2021-02-20	1031	13.0	11153	14.699050	21742.930153	9244.149556
Multnomah County	2020-12-31	25290	394.0	2021-02-20	31526	516.0	788459	16.474668	34848.165612	3998.432385
Polk County	2020-12-31	1977	30.0	2021-02-20	2978	42.0	79666	15.639958	25928.364057	3738.106595
Sherman County	2020-12-31	31	0.0	2021-02-20	52	0.0	1635	13.700000	34226.000000	3180.428135
Tillamook County	2020-12-31	308	0.0	2021-02-20	403	2.0	25840	15.512717	25458.191138	1559.597523
Umatilla County	2020-12-31	5640	57.0	2021-02-20	7580	80.0	76736	17.825222	22153.237007	9878.023353
Union County	2020-12-31	980	14.0	2021-02-20	1264	19.0	25810	17.618597	26585.728710	4897.326618
Wallowa County	2020-12-31	76	3.0	2021-02-20	142	4.0	6864	13.748776	26897.389860	2068.764569
Wasco County	2020-12-31	905	22.0	2021-02-20	1218	25.0	25687	13.670818	24727.506132	4741.698135
Washington County	2020-12-31	16070	142.0	2021-02-20	20866	209.0	572071	10.321202	35369.047499	3647.449355
Wheeler County	2020-12-31	17	1.0	2021-02-20	22	1.0	1415	20.600000	21268.000000	1554.770318
Yamhill County	2020-12-31	2641	35.0	2021-02-20	3716	62.0	102366	13.802658	28539.604791	3630.111560

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'])
```

```
In [166]: R = oregon_result['tot_cases'].corr(oregon_result['Poverty'])
          R

Out[166]: -0.11895289856840817
```

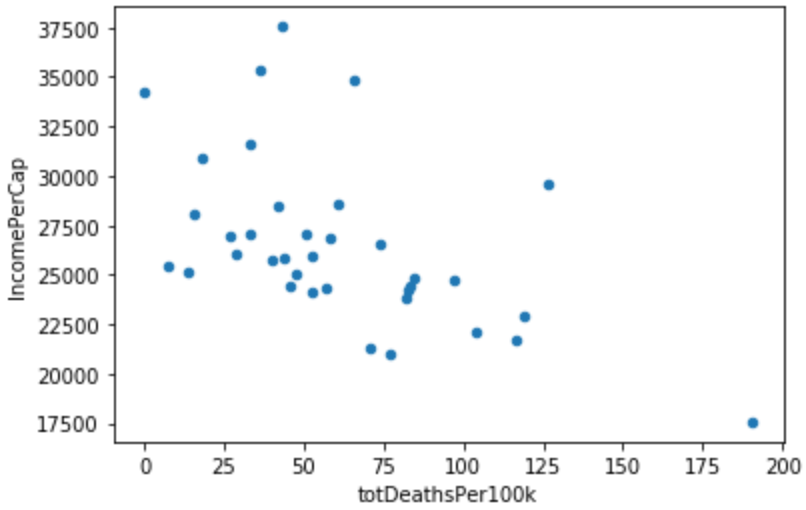
For any R that is > 0.5 or < -0.5 also display a scatter plot (see pandas scatterplot and seaborn documentation for information about how to display scatter plots from DataFrame data).

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
    0.31806336246419203
    b. COVID total deaths vs. % population in poverty
    0.36387526180001856

    c. COVID total cases vs. Per Capita Income level
    -0.41300604347954034

    d. COVID total deaths vs. Per Capita Income level
    -0.5367952720335063



e. COVID cases during December 2020 vs. % population in poverty
0.29487677050379707

f. COVID deaths during December 2020 vs. % population in poverty
0.38604737818347395

g. COVID cases during December 2020 vs. Per Capita Income level
-0.38717063280245867

h. COVID cases during December 2020 vs. Per Capita Income level

-0.38717063280245867

2. Across all of the counties in the entire USA
    a. COVID total cases vs. % population in poverty
    0.1280074916825624

    b. COVID total deaths vs. % population in poverty
    0.22670239808714923

    c. COVID total cases vs. Per Capita Income level
    -0.21209843702504072

    d. COVID total deaths vs. Per Capita Income level
    -0.25186712152583435

    e. COVID cases during December 2020 vs. % population in poverty
    0.0724852534999747

f. COVID deaths during December 2020 vs. % population in poverty
0.17970146156874178

g. COVID cases during December 2020 vs. Per Capita Income level
-0.16790759951927262

h. COVID cases during December 2020 vs. Per Capita Income level
-0.16790759951927262

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.