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

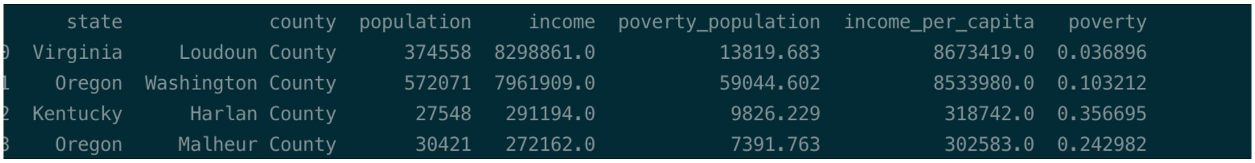
Your integration will use two different dimensions: location (as indicated by state and county) and time. You should greatly simplify your processing and reduce your time by pre-processing your data along each of these dimensions.

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



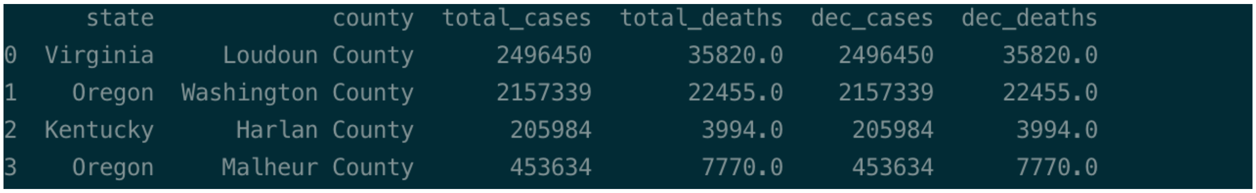
```
      state            county  population    income  poverty_population  income_per_capita   poverty
0  Virginia    Loudoun County      374558  8298861.0           13819.683          8673419.0  0.036896
1    Oregon  Washington County     572071  7961909.0           59044.602          8533980.0  0.103212
2  Kentucky     Harlan County       27548   291194.0            9826.229           318742.0  0.356695
3    Oregon    Malheur County       30421   272162.0            7391.763           302583.0  0.242982
```

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.

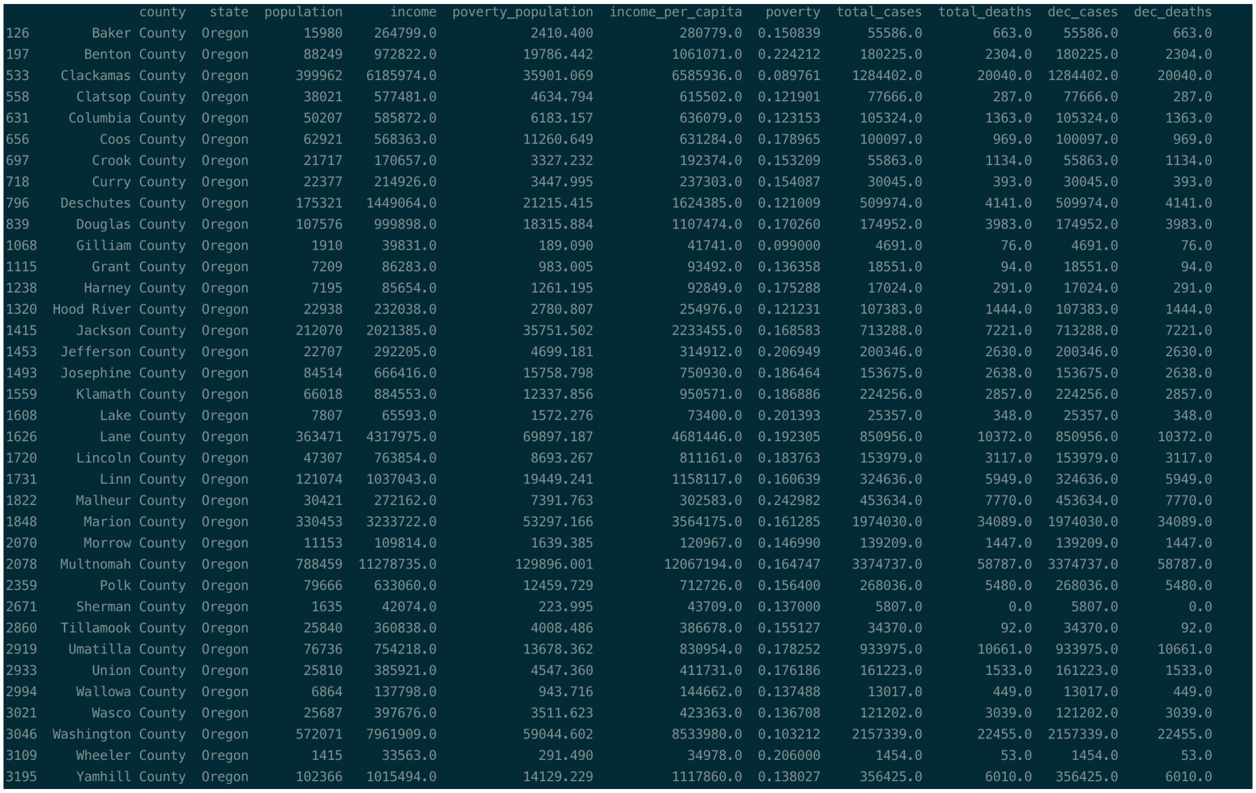


```
      state             county  total_cases  total_deaths  dec_cases  dec_deaths
0  Virginia     Loudoun County      2496450       35820.0    2496450     35820.0
1    Oregon  Washington County      2157339       22455.0    2157339     22455.0
2  Kentucky      Harlan County       205984        3994.0     205984      3994.0
3    Oregon     Malheur County       453634        7770.0     453634      7770.0
```

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      income  poverty_population  income_per_capita   poverty  total_cases  total_deaths  dec_cases  dec_deaths
126       Baker County  Oregon       15980    264799.0            2410.400           280779.0  0.150839      55586.0         663.0    55586.0       663.0
197      Benton County  Oregon       88249    972822.0           19786.442          1061071.0  0.224212     180225.0        2304.0   180225.0      2304.0
533    Clackamas County  Oregon      399962   6185974.0           35901.069          6585936.0  0.089761    1284402.0       20040.0  1284402.0     20040.0
558      Clatsop County  Oregon       38021    577481.0            4634.794           615502.0  0.121901      77666.0         287.0    77666.0       287.0
631     Columbia County  Oregon       50207    585872.0            6183.157           636079.0  0.123153     105324.0        1363.0   105324.0      1363.0
656         Coos County  Oregon       62921    568363.0           11260.649           631284.0  0.178965     100097.0         969.0   100097.0       969.0
697        Crook County  Oregon       21717    170657.0            3327.232           192374.0  0.153209      55863.0        1134.0    55863.0      1134.0
718        Curry County  Oregon       22377    214926.0            3447.995           237303.0  0.154087      30045.0         393.0    30045.0       393.0
796    Deschutes County  Oregon      175321   1449064.0           21215.415          1624385.0  0.121009     509974.0        4141.0   509974.0      4141.0
839      Douglas County  Oregon      107576    999898.0           18315.884          1107474.0  0.170260     174952.0        3983.0   174952.0      3983.0
1068    Gilliam County  Oregon        1910     39831.0             189.090            41741.0  0.099000       4691.0          76.0     4691.0        76.0
1115      Grant County  Oregon        7209     86283.0             983.005            93492.0  0.136358      18551.0          94.0    18551.0        94.0
1238     Harney County  Oregon        7195     85654.0            1261.195            92849.0  0.175288      17024.0         291.0    17024.0       291.0
1320 Hood River County  Oregon       22938    232038.0            2780.807           254976.0  0.121231     107383.0        1444.0   107383.0      1444.0
1415    Jackson County  Oregon      212070   2021385.0           35751.502          2233455.0  0.168583     713288.0        7221.0   713288.0      7221.0
1453  Jefferson County  Oregon       22707    292205.0            4699.181           314912.0  0.206949     200346.0        2630.0   200346.0      2630.0
1493  Josephine County  Oregon       84514    666416.0           15758.798           750930.0  0.186464     153675.0        2638.0   153675.0      2638.0
1559    Klamath County  Oregon       66018    884553.0           12337.856           950571.0  0.186886     224256.0        2857.0   224256.0      2857.0
1608       Lake County  Oregon        7807     65593.0            1572.276            73400.0  0.201393      25357.0         348.0    25357.0       348.0
1626       Lane County  Oregon      363471   4317975.0           69897.187          4681446.0  0.192305     850956.0       10372.0   850956.0     10372.0
1720    Lincoln County  Oregon       47307    763854.0            8693.267           811161.0  0.183763     153979.0        3117.0   153979.0      3117.0
1731       Linn County  Oregon      121074   1037043.0           19449.241          1158117.0  0.160639     324636.0        5949.0   324636.0      5949.0
1822    Malheur County  Oregon       30421    272162.0            7391.763           302583.0  0.242982     453634.0        7770.0   453634.0      7770.0
1848     Marion County  Oregon      330453   3233722.0           53297.166          3564175.0  0.161285    1974030.0       34089.0  1974030.0     34089.0
2070     Morrow County  Oregon       11153    109814.0            1639.385           120967.0  0.146990     139209.0        1447.0   139209.0      1447.0
2078  Multnomah County  Oregon      788459  11278735.0          129896.001         12067194.0  0.164747    3374737.0       58787.0  3374737.0     58787.0
2359       Polk County  Oregon       79666    633060.0           12459.729           712726.0  0.156400     268036.0        5480.0   268036.0      5480.0
2671    Sherman County  Oregon        1635     42074.0             223.995            43709.0  0.137000       5807.0           0.0     5807.0         0.0
2860  Tillamook County  Oregon       25840    360838.0            4008.486           386678.0  0.155127      34370.0          92.0    34370.0        92.0
2919   Umatilla County  Oregon       76736    754218.0           13678.362           830954.0  0.178252     933975.0       10661.0   933975.0     10661.0
2933      Union County  Oregon       25810    385921.0            4547.360           411731.0  0.176186     161223.0        1533.0   161223.0      1533.0
2994    Wallowa County  Oregon        6864    137798.0             943.716           144662.0  0.137488      13017.0         449.0    13017.0       449.0
3021      Wasco County  Oregon       25687    397676.0            3511.623           423363.0  0.136708     121202.0        3039.0   121202.0      3039.0
3046 Washington County  Oregon      572071   7961909.0           59044.602          8533980.0  0.103212    2157339.0       22455.0  2157339.0     22455.0
3109    Wheeler County  Oregon        1415     33563.0             291.490            34978.0  0.206000       1454.0          53.0     1454.0        53.0
3195    Yamhill County  Oregon      102366   1015494.0           14129.229          1117860.0  0.138027     356425.0        6010.0   356425.0      6010.0
```

D. Analysis

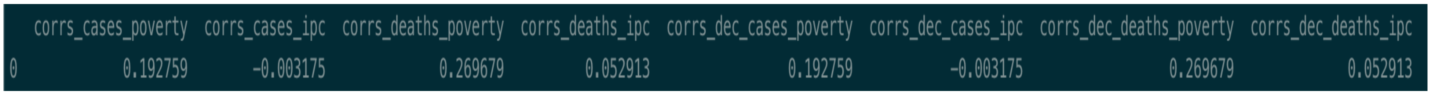
For each of the following, determine the strength of the correlation between each pair of variables. Compute the correlation strength by calculating the Pearson correlation coefficient R for pairs of columns in your DataFrame. For example, if you have a DataFrame df with each row representing a distinct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

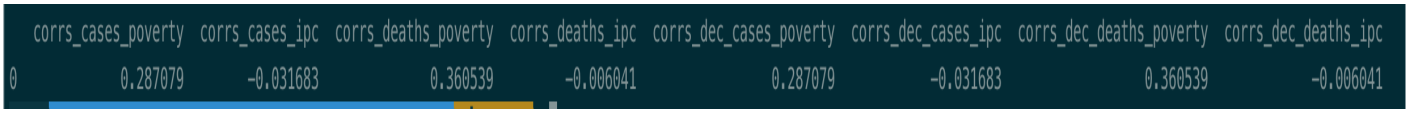
The COVID numbers should be normalized to population (# of cases per 100,000 residents) so that different sized counties are comparable. So for example, “COVID total cases” below really means “((COVID total cases in county * 100000) / population of county)”.

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



```
   corrs_cases_poverty  corrs_cases_ipc  corrs_deaths_poverty  corrs_deaths_ipc  corrs_dec_cases_poverty  corrs_dec_cases_ipc  corrs_dec_deaths_poverty  corrs_dec_deaths_ipc
0             0.192759        -0.003175              0.269679          0.052913                 0.192759            -0.003175                  0.269679              0.052913
```

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



```
   corrs_cases_poverty  corrs_cases_ipc  corrs_deaths_poverty  corrs_deaths_ipc  corrs_dec_cases_poverty  corrs_dec_cases_ipc  corrs_dec_deaths_poverty  corrs_dec_deaths_ipc
0             0.287079        -0.031683              0.360539         -0.006041                 0.287079            -0.031683                  0.360539             -0.006041
```

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.