DataEng: Data Integration Activity

NEHA AGRAWAL

WEEK 8 – In class assignment
My responses are inlined below.

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: Loudon County Virginia, Washington County Oregon, Harlan County Kentucky, Malheur County Oregon

1. Loudoun County Virginia:

```
   state          county  Population     Men   Women   Hispanic  \
Virginia  Loudoun County      374558  185575  188983  13.898438

    White    Black    Native      Asian   Pacific  VotingAgeCitizen  \
59.409375  7.19375  0.232812  15.234375  0.071875            231130

       Income    IncomeErr  PerCapitaIncome  IncomePerCapErr   Poverty  \
129669.703125  15198.28125     50455.645745      5336.265625  3.689598

ChildPoverty  Professional    Service     Office  Construction  \
    4.434375      56.44375  13.384375  20.460937      4.398438

Production     Drive  Carpool   Transit      Walk  OtherTransp  \
  5.307812  77.05625  9.10625  3.446875  1.842188     1.089063

WorkAtHome  MeanCommute  Employed  PrivateWork  PublicWork  \
  7.459375    33.417187    201528    78.528125   16.539062

SelfEmployed  FamilyWork  Unemployment
    4.732813    0.198438         3.825
```

2. Washington County Oregon

```
 state             county  Population     Men   Women   Hispanic  \
Oregon  Washington County      572071  282381  289690  16.461538

    White     Black    Native  Asian   Pacific  VotingAgeCitizen  \
68.310577  1.795192  0.281731  8.675  0.360577            384659

      Income     IncomeErr  PerCapitaIncome  IncomePerCapErr    Poverty  \
76556.817308  10490.596154     35369.047499      4091.605769  10.321202

ChildPoverty  Professional    Service     Office  Construction  \
   13.657692     44.122115  15.674038  23.052885      7.226923

Production      Drive    Carpool   Transit      Walk  OtherTransp  \
  9.925962  73.274038  10.139423  6.106731  2.588462     1.805769

WorkAtHome  MeanCommute  Employed  PrivateWork  PublicWork  \
  6.081731       24.875    292979        84.525    9.597115

SelfEmployed  FamilyWork  Unemployment
    5.751923    0.125962        5.5125
```

3. Harlan County Kentucky

```
   state         county  Population    Men  Women  Hispanic      White  \
Kentucky  Harlan County       27548  13323  14225       0.7  95.209091

   Black  Native  Asian  Pacific  VotingAgeCitizen        Income  \
2.309091     0.1    0.8      0.0             21193  26472.181818

  IncomeErr  PerCapitaIncome  IncomePerCapErr    Poverty  ChildPoverty  \
6359.363636     15456.971032           2455.0  35.669482     42.018182

Professional    Service     Office  Construction  Production      Drive  \
   28.263636  20.145455  22.745455     14.745455        14.1  84.072727

  Carpool   Transit      Walk  OtherTransp  WorkAtHome  MeanCommute  \
11.481818  0.545455  2.354545     0.181818    1.363636    22.072727

Employed  PrivateWork  PublicWork  SelfEmployed  FamilyWork  \
    7555    71.945455   22.045455      5.990909    0.027273

Unemployment
    9.072727
```

4. Malheur County Oregon

```
 state          county  Population    Men  Women   Hispanic      White  \
Oregon  Malheur County       30421  16514  13907  32.185714  62.671429

Black    Native     Asian  Pacific  VotingAgeCitizen        Income  \
  0.7  1.014286  1.585714      0.1             20573  38880.285714

  IncomeErr  PerCapitaIncome  IncomePerCapErr    Poverty  ChildPoverty  \
6669.857143     17567.504323      2329.285714  24.298225     35.628571

Professional    Service     Office  Construction  Production      Drive  \
   27.614286  20.814286  18.614286     16.514286   16.414286  74.771429

  Carpool   Transit  Walk  OtherTransp  WorkAtHome  MeanCommute  \
10.828571  0.257143   4.8     2.542857    6.785714         19.5

Employed  PrivateWork  PublicWork  SelfEmployed  FamilyWork  \
   10824    72.414286   17.685714      9.171429         0.7

Unemployment
    8.057143
```

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.

1. Loudoun County Virginia:

```
 county     state  TotalCases  TotalDeaths  Dec2020Cases  Dec2020Deaths
Loudoun  Virginia     2496450        35820        376223           4729
```

2. Washington County Oregon

```
    county   state  TotalCases  TotalDeaths  Dec2020Cases  Dec2020Deaths
Washington  Oregon     2157339        22455        424620           3860
```

3. Harlan County Kentucky

county	state	TotalCases	TotalDeaths	Dec2020Cases	Dec2020Deaths
Harlan	Kentucky	205984	3994	38959	506

4. Malheur County Oregon

county	state	TotalCases	TotalDeaths	Dec2020Cases	Dec2020Deaths
Malheur	Oregon	453634	7770	82916	1465

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.

The below output contains the normalized columns as well -

county	state	TotalCases	TotalDeaths	Dec2020Cases	Dec2020Deaths	Population	Poverty	PerCapitaIncome	norm_TotalCases	norm_TotalDeaths	norm_Dec2020Cases	norm_Dec2020Deaths
Baker County	Oregon	55586	663	11688	133	15980	15.083854818523154	25820.273153942428	347847.30913642055	4148.936170212766	73141.42678347934	832.2903629536921
Benton County	Oregon	180225	2304	34260	278	88249	22.42115151446475	30872.824360615985	204223.27731759	2610.794456594409	38821.96965404707	315.017733911999
Clackamas County	Oregon	1284402	20040	261810	3125	399962	8.97611998139823	37550.849108165276	321131.00744570734	5010.475995219545	65458.71857826493	781.3242258014511
Clatsop County	Oregon	77666	287	14439	47	38021	12.190089687278082	28114.625522737435	204271.3237421425	754.8460061544936	37976.38147339628	123.61589647826202
Columbia County	Oregon	105324	1363	21459	266	50207	12.31532853984504	28459.68805146693	209779.51281693787	2714.760889915749	42741.05204453562	529.8066006732129
Coos County	Oregon	100097	969	18806	151	62921	17.896487659128113	26007.21299725052	159083.61278428507	1540.026382288902	29888.272595794726	239.98347133707347
Crook County	Oregon	55863	1134	11048	196	21717	15.320863839388496	24238.814477137726	257231.6618317447	5221.715706589308	50872.58829488419	902.5187641018557
Curry County	Oregon	30045	393	6741	72	22377	15.408656209500826	26925.536398981098	134267.32806006167	1756.267596192519	30124.6815927068	321.75894892076684
Deschutes County	Oregon	509974	4141	102490	563	175321	12.100897781783132	31574.934092322084	290880.1569692165	2361.953217241517	58458.48472230937	321.1252502552461
Douglas County	Oregon	174952	3983	37590	964	107576	17.02599464564587	25001.732923700452	162631.0701271659	3702.498698594482	34942.73815720979	896.1106566520414
Gilliam County	Oregon	4691	76	898	25	1910	9.9	24178.0	245602.0942408377	3979.057591623037	47015.706806282724	1308.9005235602094
Grant County	Oregon	18551	94	4895	31	7209	13.635802469135804	25154.16174226661	257331.1138854210 1	1303.925648494937	67901.23456790124	430.0180330142877
Harney County	Oregon	17024	291	3717	34	7195	17.52876997915219	24397.712578179293	236608.75608061152	4044.4753300903403	51660.87560806116	472.55038220986796
Hood River County	Oregon	107383	1444	19348	216	22938	12.123144999564044	29594.972796233324	468144.56360624294	6295.230621675822	84349.11500566745	941.6688464556631
Jackson County	Oregon	713288	7221	154535	1655	212070	16.85834960154666	27080.538534446172	336345.54628188803	3405.0077804498515	72869.80713915217	780.4026972226152
Jefferson County	Oregon	200346	2630	36278	409	22707	20.694856211740877	22956.83529308143	882309.4200026423	11582.331439644162	159765.71101422468	1801.206676355309
Josephine County	Oregon	153675	2638	27180	407	84514	18.646375748396714	24348.609449322004	181833.77901886078	3121.376340014672	32160.35213100788	481.5770168256147
Klamath County	Oregon	224256	2857	45118	373	66018	18.688624314580867	23793.066678784573	339689.17567936017	4327.607622163652	68341.96734224	564.9974249447121
Lake County	Oregon	25357	348	5358	76	7807	20.1393108748559	21004.5893428974	324798.2579736 1345	4457.538106827206	68630.71602408095	973.4853336749071
Lane County	Oregon	850956	10372	178816	2215	363471	19.23047148190639	27032.412178688257	234119.36578158918	2853.597673542043	49196.77223217258	609.4021256166242
Lincoln County	Oregon	153979	3117	24041	502	47307	18.376280465892997	25782.113704102987	325488.82829179615	6588.87691039381	50819.11767814488	1061.1537404612425
Linn County	Oregon	324636	5949	66702	891	121074	16.063928671721424	24448.46735880536	268130.2344021012	4913.5239605530505	55091.92725110263	735.9135735170227
Malheur County	Oregon	453634	7770	82916	1465	30421	24.298224910423723	17567.504322671837	1491187.008974064	25541.56668091121	272561.7172348049	4815.7522763880215
Marion County	Oregon	1974030	34089	365801	5720	330453	16.128516309429784	24791.074830611313	597370.8817895434	-2681.371620169888	110696.8313194312	1730.9572011753562
Morrow County	Oregon	139209	1447	23219	227	11153	14.699049583071819	21742.930153321977	1248175.3788218417	12974.08768941092	208186.13825876446	2035.32661789653
Multnomah County	Oregon	3374737	58787	680418	10244	788459	16.474667801369506	34848.16561165514	428016.8023955589	2008.643067045972	86297.19490804215	1299.2432073195944
Polk County	Oregon	268036	5480	50986	743	79666	15.639958074962973	25928.364057439812	336449.67740315816	6878.718650365275	63999.69874224889	932.6437878141239
Sherman County	Oregon	5807	0	855	0	1635	13.700000000000001	34226.0	355168.19571865443	0.0	52293.577981651375	0.0
Tillamook County	Oregon	34370	92	6850	0	25840	15.512716718266253	25458.191137770897	133010.83591331268	356.0371517027864	26509.287925696593	0.0
Umatilla County	Oregon	933975	10661	154995	1645	76736	17.825221538782323	22153.237007402	1217127.5542118433	13893.087989999166	201984.7268557131	2143.713511259383
Union County	Oregon	161223	1533	28227	338	25810	17.618597442851 61	26585.728709802403	624653.2351801627	5939.558310732275	109364.58736923673	1309.5699341340567
Wallowa County	Oregon	13017	449	2306	93	6864	13.748776223776222	26897.38986013986	189641.6083916084	6541.375291375291	33595.5710955711	1354.8951048951049
Wasco County	Oregon	121202	3039	22511	621	25687	13.670817923463233	24727.50613150621	471841.78767469927	11830.887219215945	87635.76906606455	2417.5653054074046
Washington County	Oregon	2157339	22455	424620	3860	572071	10.321201738944987	35369.04749934886	377110.35867925483	-3582.5400972956154	74225.05248474402	674.7414219563656
Wheeler County	Oregon	1454	53	359	2	1415	20.600000000000005	21268.0	102756.18374558304	3745.583038869258	25371.024734982333	141.34275618374556
Yamhill County	Oregon	356425	6010	69481	812	102366	13.8026581091378	28539.604790653146	348186.8979934744	5871.090010355	67875.07570873141	793.2321278549518

Show [50] per page

The above output is too blurry, so the below figure shows the output without the normalized columns-

county	state	TotalCases	TotalDeaths	Dec2020Cases	Dec2020Deaths	Population	Poverty	PerCapitaIncome
Baker County	Oregon	55586	663	11688	133	15980	15.083854818523154	25820.273153942428
Benton County	Oregon	180225	2304	34260	278	88249	22.42115151446475	30872.824360615985
Clackamas County	Oregon	1284402	20040	261810	3125	399962	8.97611998139823	37550.849108165276
Clatsop County	Oregon	77666	287	14439	47	38021	12.190089687278082	28114.625522737435
Columbia County	Oregon	105324	1363	21459	266	50207	12.31532853984504	28459.68805146693
Coos County	Oregon	100097	969	18806	151	62921	17.896487659128113	26007.21299725052
Crook County	Oregon	55863	1134	11048	196	21717	15.320863839388496	24238.814477137726
Curry County	Oregon	30045	393	6741	72	22377	15.408656209500826	26925.536398981098
Deschutes County	Oregon	509974	4141	102490	563	175321	12.100897781783132	31574.934092322084
Douglas County	Oregon	174952	3983	37590	964	107576	17.02599464564587	25001.732923700452
Gilliam County	Oregon	4691	76	898	25	1910	9.9	24178.0
Grant County	Oregon	18551	94	4895	31	7209	13.635802469135804	25154.16174226661
Harney County	Oregon	17024	291	3717	34	7195	17.52876997915219	24397.712578179293
Hood River County	Oregon	107383	1444	19348	216	22938	12.123144999564044	29594.972796233324
Jackson County	Oregon	713288	7221	154535	1655	212070	16.85834960154666	27080.538534446172
Jefferson County	Oregon	200346	2630	36278	409	22707	20.694856211740877	22956.83529308143
Josephine County	Oregon	153675	2638	27180	407	84514	18.646375748396714	24348.609449322004
Klamath County	Oregon	224256	2857	45118	373	66018	18.688624314580867	23793.066678784573
Lake County	Oregon	25357	348	5358	76	7807	20.1393108748559	21004.5893428974
Lane County	Oregon	850956	10372	178816	2215	363471	19.23047148190639	27032.412178688257
Lincoln County	Oregon	153979	3117	24041	502	47307	18.376280465892997	25782.113704102987
Linn County	Oregon	324636	5949	66702	891	121074	16.063928671721424	24448.46735880536
Malheur County	Oregon	453634	7770	82916	1465	30421	24.298224910423723	17567.504322671837
Marion County	Oregon	1974030	34089	365801	5720	330453	16.128516309429784	24791.074830611313
Morrow County	Oregon	139209	1447	23219	227	11153	14.699049583071819	21742.930153321977
Multnomah County	Oregon	3374737	58787	680418	10244	788459	16.474667801369506	34848.16561165514
Polk County	Oregon	268036	5480	50986	743	79666	15.639958074962973	25928.364057439812
Sherman County	Oregon	5807	0	855	0	1635	13.700000000000001	34226.0
Tillamook County	Oregon	34370	92	6850	0	25840	15.512716718266253	25458.191137770897
Umatilla County	Oregon	933975	10661	154995	1645	76736	17.825221538782323	22153.237007402
Union County	Oregon	161223	1533	28227	338	25810	17.61859744285161	26585.728709802403
Wallowa County	Oregon	13017	449	2306	93	6864	13.748776223776222	26897.38986013986
Wasco County	Oregon	121202	3039	22511	621	25687	13.670817923463233	24727.50613150621
Washington County	Oregon	2157339	22455	424620	3860	572071	10.321201738944987	35369.04749934886
Wheeler County	Oregon	1454	53	359	2	1415	20.600000000000005	21268.0
Yamhill County	Oregon	356425	6010	69481	812	102366	13.8026581091378	28539.604790653146

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

    ```
    0.28707860802137747
    ```

    b. COVID total deaths vs. % population in poverty

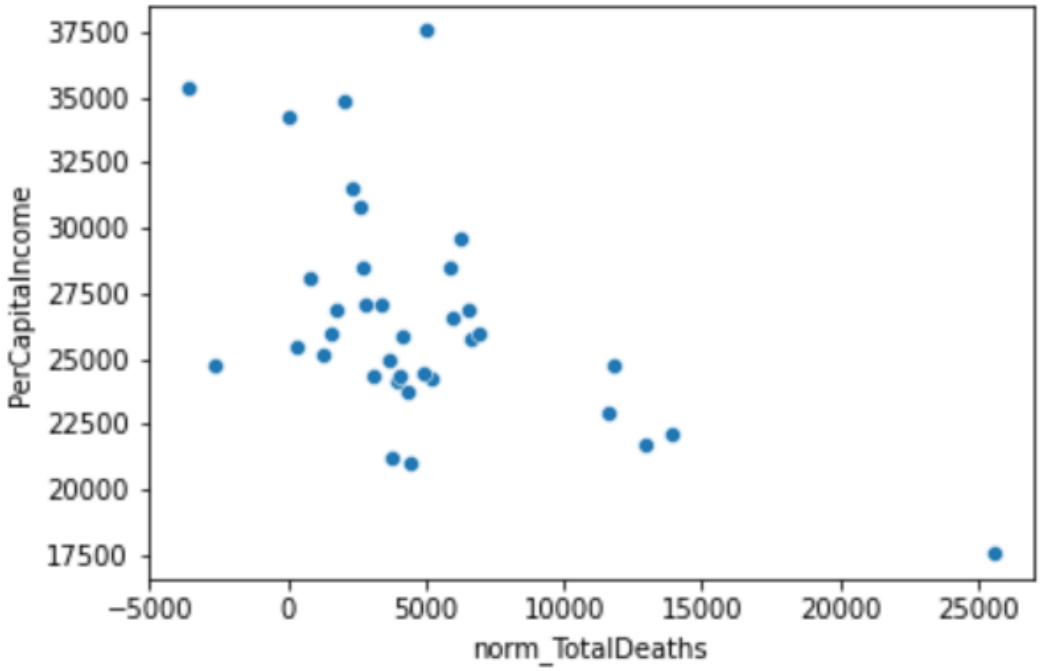
    ```
    0.3963339291874621
    ```

    c. COVID total cases vs. Per Capita Income level

    ```
    -0.37568502761471967
    ```

    d. COVID total deaths vs. Per Capita Income level

    ```
    -0.5477168607765278
    ```



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

0.29815203013315383

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

0.3027269512831473

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

-0.3853971943730501

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

-0.45595519506866566

2. Across all of the counties in the entire USA

a. COVID total cases vs. % population in poverty

0.1927585925100599

b. COVID total deaths vs. % population in poverty

0.25512753235836966

c. COVID total cases vs. Per Capita Income level

-0.20391612903601486

d. COVID total deaths vs. Per Capita Income level

-0.31183735299611814

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

0.06359486143114673

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

0.2120341447092468

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

-0.14580811249276734

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

```
-0.2483643330208416
```

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.