DataEng: Data Integration Activity

Student: Kaushik Chaudhary

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	PerCapitaIncome
0	Loudoun	Virginia	374558	3.884375	50391.015625
1	Washington	Oregon	572071	10.446154	34970.817308
2	Harlan	Kentucky	27548	33.318182	16010.363636
3	Malheur	Oregon	30421	24.414286	17966.428571

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	Dec2020Cases	TotalDeaths	Dec2020Deaths
0	Loudoun	Virginia	2496450	376223	35820.0	4729.0
1	Washington	Oregon	2157339	424620	22455.0	3860.0
2	Harlan	Kentucky	205984	38959	3994.0	506.0
3	Malheur	Oregon	453634	82916	7770.0	1465.0

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	PerCapitaIncome	TotalCases	Dec2020Cases	TotalDeaths	Dec2020Deaths
1	Washington	Oregon	572071	10.446154	34970.817308	3.771104e+05	424620	3925.212080	3860.0
3	Malheur	Oregon	30421	24.414286	17966.428571	1.491187e+06	82916	25541.566681	1465.0

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

	State	County	TotalCases	...	Population	Poverty	PerCapitaIncome
0	Oregon	Washington	7.931876e+05	...	2564646	13.422243	30729.954380
1	Oregon	Jackson	7.416558e+05	...	1722755	18.493882	26107.727059
2	Oregon	Klamath	3.396892e+05	...	66018	18.930000	23712.400000
3	Oregon	Douglas	9.380390e+05	...	1440952	13.097015	33581.874627
4	Oregon	Marion	9.534415e+05	...	2018526	19.951429	24578.778022
5	Oregon	Multnomah	4.280168e+05	...	788459	15.730588	36739.558824
6	Oregon	Deschutes	2.908802e+05	...	175321	12.208333	31834.375000
7	Oregon	Linn	7.293391e+05	...	362932	13.113699	28612.260274
8	Oregon	Polk	1.040834e+06	...	1456426	16.184211	26045.102167
9	Oregon	Umatilla	1.217128e+06	...	76736	16.520000	23200.466667
10	Oregon	Clackamas	3.211310e+05	...	399962	9.320000	37502.712500
11	Oregon	Yamhill	3.481869e+05	...	102366	13.935294	28578.882353
12	Oregon	Benton	9.071618e+05	...	647670	14.231111	29409.385185
13	Oregon	Lane	2.367902e+05	...	365173	18.439080	27571.758621
14	Oregon	Grant	9.973480e+05	...	331216	16.445333	23628.626667
15	Oregon	Union	1.165808e+06	...	1128752	12.091453	31687.132479
16	Oregon	Josephine	1.818338e+05	...	84514	19.131250	24179.062500
17	Oregon	Hood River	4.681446e+05	...	22938	12.150000	29178.000000
18	Oregon	Clatsop	2.042713e+05	...	38021	12.481818	28357.363636
19	Oregon	Lincoln	9.994460e+05	...	596238	14.535172	26656.834483
20	Oregon	Tillamook	1.330108e+05	...	25840	15.437500	25805.750000
21	Oregon	Wasco	4.718418e+05	...	25687	13.037500	25089.750000
22	Oregon	Columbia	8.618110e+05	...	475291	13.003125	29630.302083
23	Oregon	Morrow	7.754947e+05	...	46088	10.825000	24286.875000
24	Oregon	Malheur	1.491187e+06	...	30421	24.414286	17966.428571
25	Oregon	Wallowa	1.896416e+05	...	6864	14.400000	26943.000000
26	Oregon	Crook	3.393924e+05	...	29064	11.683333	27464.666667
27	Oregon	Sherman	7.735617e+05	...	13749	13.220000	28958.000000
28	Oregon	Curry	7.464575e+05	...	72660	20.747059	23559.470588
29	Oregon	Coos	1.858439e+05	...	95040	15.391667	26641.500000
30	Oregon	Jefferson	1.081458e+06	...	3208539	16.088481	29289.780856
31	Oregon	Harney	2.366088e+05	...	7195	16.300000	25174.500000
32	Oregon	Baker	1.050432e+06	...	46768	16.191667	24281.750000
33	Oregon	Lake	9.394612e+05	...	1900172	14.451781	32451.424171
34	Oregon	Gilliam	2.456021e+05	...	1910	9.900000	24178.000000
35	Oregon	Wheeler	8.175753e+05	...	15812	19.850000	20786.833333

```
[0.297471134394528, 0.20946190596896594, -0.18950461546988256, -0.07405908604882054, 0.057059538757632346, 0.05571798839796383, 0.2851092192782961, 0.2664467664558328]
[]
```

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.

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:49: RuntimeWarning: in
                    County    TotalCases  ...    Poverty  PerCapitaIncome
0                Snohomish  4.587665e+05  ...   8.934899     36148.134228
1                     Cook  1.180935e+06  ...  17.988948     33006.301287
2                   Orange  8.418201e+05  ...  13.075722     35621.433298
3                 Maricopa  1.267895e+06  ...  16.130955     30761.265642
4              Los Angeles  1.122602e+06  ...  17.323803     31389.413867
...                    ...           ...  ...        ...              ...
1925             Petroleum  2.445916e+05  ...  10.100000     31549.000000
1926  Skagway Municipality           inf  ...        NaN              NaN
1927             Esmeralda  2.421053e+05  ...   6.800000     23755.000000
1928                Loving  1.297297e+05  ...  17.100000     35530.000000
1929               Kalawao  8.488372e+04  ...  12.700000     46024.000000
```

```
[0.16690037795100285, 0.199728642737955, -0.1870637430384339, -0.1413631086967846, -0.027528196894307654, -0.02187824035470475, 0.2021547054177172, 0.21840972654694812]
[]
```