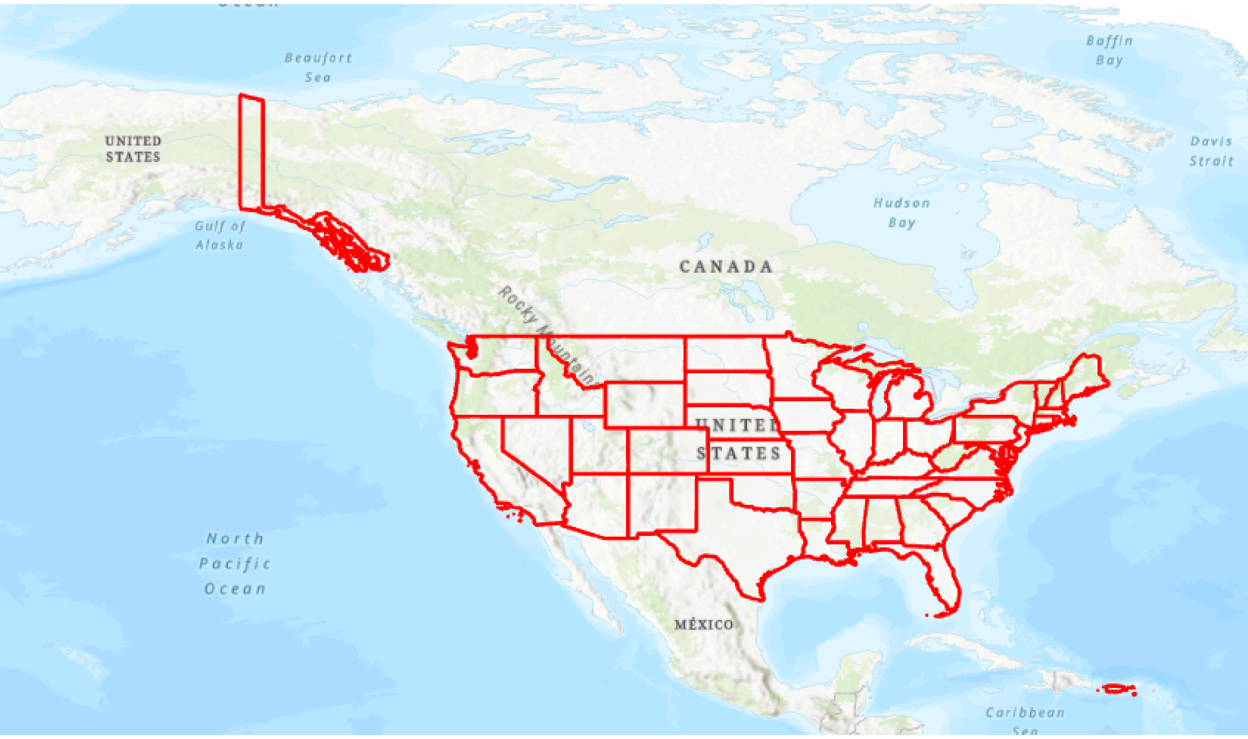
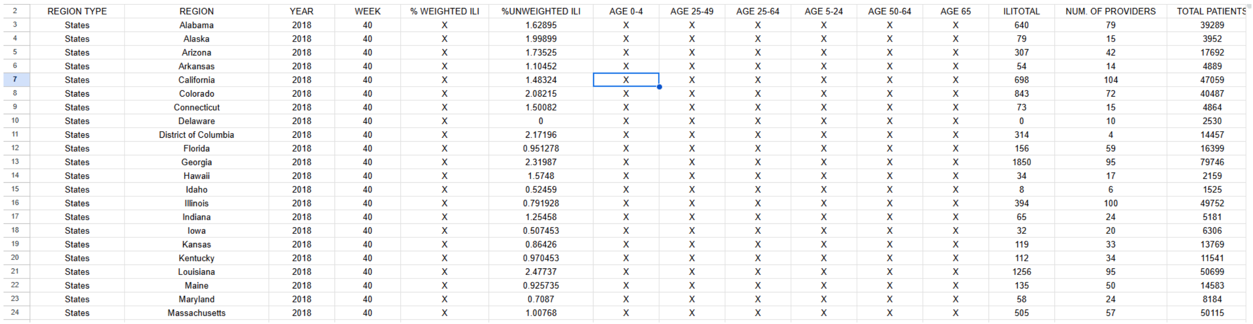
Let's start from collecting data:
We use ArcGISPro3 for our visualization.
So we download the state boundary shapefile for the US from this website:
https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html



Then we go for the Flu data from FLUVIEW CDS.
We are downloading season 2018-2019, Viral Surveillance by public health labs and Surveillance Area is State wide.
https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html
Climate Data Optionally:

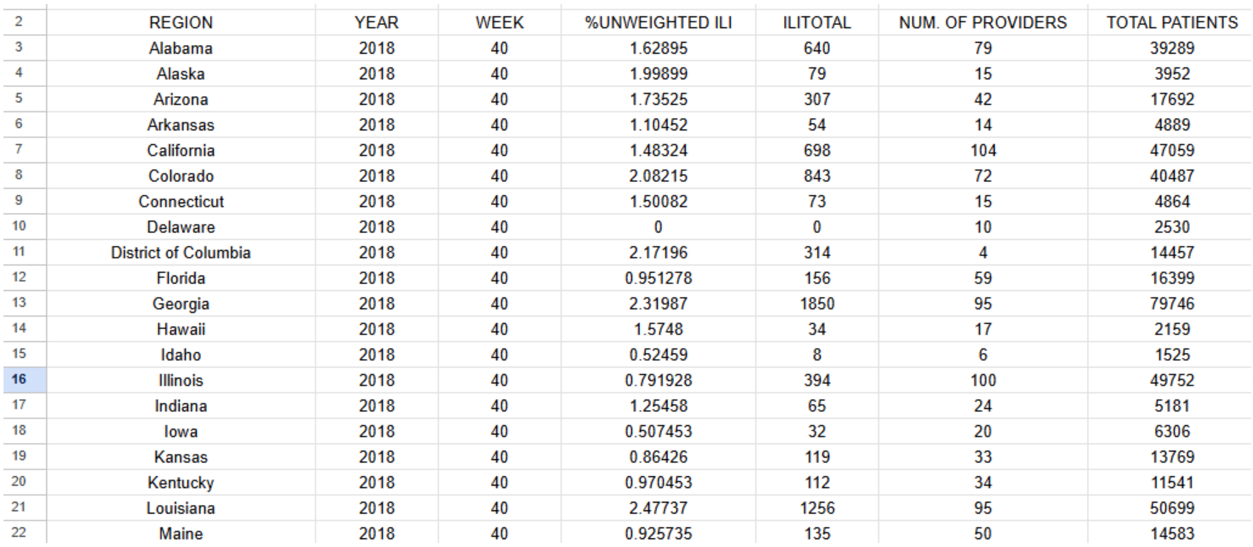
Process of the project:
ILINet (Influenza like illness) CSV



2	REGION TYPE	REGION	YEAR	WEEK	% WEIGHTED ILI	%UNWEIGHTED ILI	AGE 0-4	AGE 25-49	AGE 25-64	AGE 5-24	AGE 50-64	AGE 65	ILITOTAL	NUM. OF PROVIDERS	TOTAL PATIENTS
3	States	Alabama	2018	40	X	1.62895	X	X	X	X	X	X	640	79	39289
4	States	Alaska	2018	40	X	1.99899	X	X	X	X	X	X	79	15	3952
5	States	Arizona	2018	40	X	1.73525	X	X	X	X	X	X	307	42	17692
6	States	Arkansas	2018	40	X	1.10452	X	X	X	X	X	X	54	14	4889
7	States	California	2018	40	X	1.48324	X	X	X	X	X	X	698	104	47059
8	States	Colorado	2018	40	X	2.08215	X	X	X	X	X	X	843	72	40487
9	States	Connecticut	2018	40	X	1.50082	X	X	X	X	X	X	73	15	4864
10	States	Delaware	2018	40	X	0	X	X	X	X	X	X	0	10	2530
11	States	District of Columbia	2018	40	X	2.17196	X	X	X	X	X	X	314	4	14457
12	States	Florida	2018	40	X	0.951278	X	X	X	X	X	X	156	59	16399
13	States	Georgia	2018	40	X	2.31987	X	X	X	X	X	X	1850	95	79746
14	States	Hawaii	2018	40	X	1.5748	X	X	X	X	X	X	34	17	2159
15	States	Idaho	2018	40	X	0.52459	X	X	X	X	X	X	8	6	1525
16	States	Illinois	2018	40	X	0.791928	X	X	X	X	X	X	394	100	49752
17	States	Indiana	2018	40	X	1.25458	X	X	X	X	X	X	65	24	5181
18	States	Iowa	2018	40	X	0.507453	X	X	X	X	X	X	32	20	6306
19	States	Kansas	2018	40	X	0.86426	X	X	X	X	X	X	119	33	13769
20	States	Kentucky	2018	40	X	0.970453	X	X	X	X	X	X	112	34	11541
21	States	Louisiana	2018	40	X	2.47737	X	X	X	X	X	X	1256	95	50699
22	States	Maine	2018	40	X	0.925735	X	X	X	X	X	X	135	50	14583
23	States	Maryland	2018	40	X	0.7087	X	X	X	X	X	X	58	24	8184
24	States	Massachusetts	2018	40	X	1.00768	X	X	X	X	X	X	505	57	50115

So we open the ILINet (Influenza like illness) CSV file and start its cleaning. There are 2808 rows.

The region type column is useless so we remove it because it has only one constant value(zero variance). Now we see that the percentage of weighted ILI has missing values and by applying the filter function from Google Sheets, we come to know that the column is full of missing values and so we also drop that. In the same way, the Age Columns consist of ‘X’(non-numeric placeholders) values and we drop all the columns.

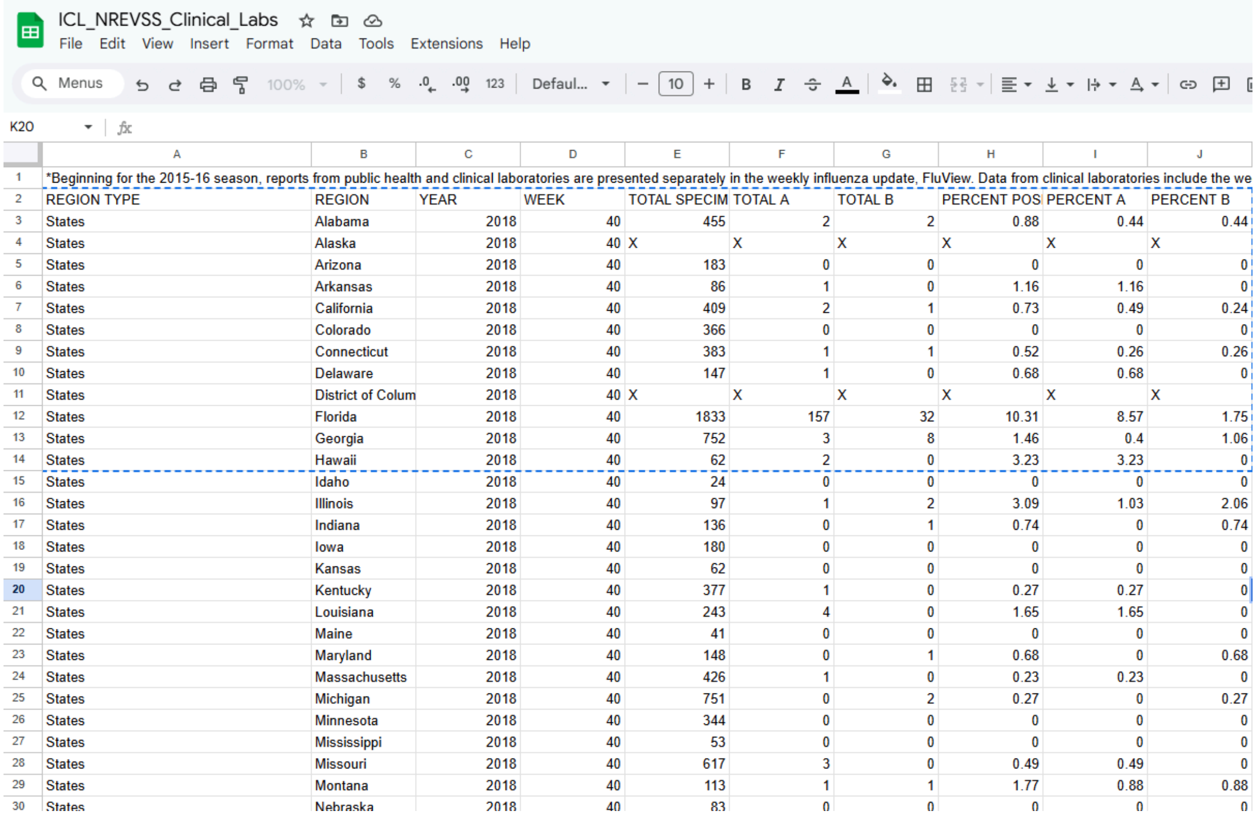


2	REGION	YEAR	WEEK	%UNWEIGHTED ILI	ILITOTAL	NUM. OF PROVIDERS	TOTAL PATIENTS
3	Alabama	2018	40	1.62895	640	79	39289
4	Alaska	2018	40	1.99899	79	15	3952
5	Arizona	2018	40	1.73525	307	42	17692
6	Arkansas	2018	40	1.10452	54	14	4889
7	California	2018	40	1.48324	698	104	47059
8	Colorado	2018	40	2.08215	843	72	40487
9	Connecticut	2018	40	1.50082	73	15	4864
10	Delaware	2018	40	0	0	10	2530
11	District of Columbia	2018	40	2.17196	314	4	14457
12	Florida	2018	40	0.951278	156	59	16399
13	Georgia	2018	40	2.31987	1850	95	79746
14	Hawaii	2018	40	1.5748	34	17	2159
15	Idaho	2018	40	0.52459	8	6	1525
16	Illinois	2018	40	0.791928	394	100	49752
17	Indiana	2018	40	1.25458	65	24	5181
18	Iowa	2018	40	0.507453	32	20	6306
19	Kansas	2018	40	0.86426	119	33	13769
20	Kentucky	2018	40	0.970453	112	34	11541
21	Louisiana	2018	40	2.47737	1256	95	50699
22	Maine	2018	40	0.925735	135	50	14583

So we have US states in Region, 2018 and 2019 in the Year Column, The week column consists of values from 1 to 42. Other than that we have checked there is no extra cleaning to do. We have 7 columns

ICL NREVSS Clinical Labs CSV:

This contains state level flu lab testing results reported by clinical labs like hospitals, healthcare providers and everything. There are 2808 rows.

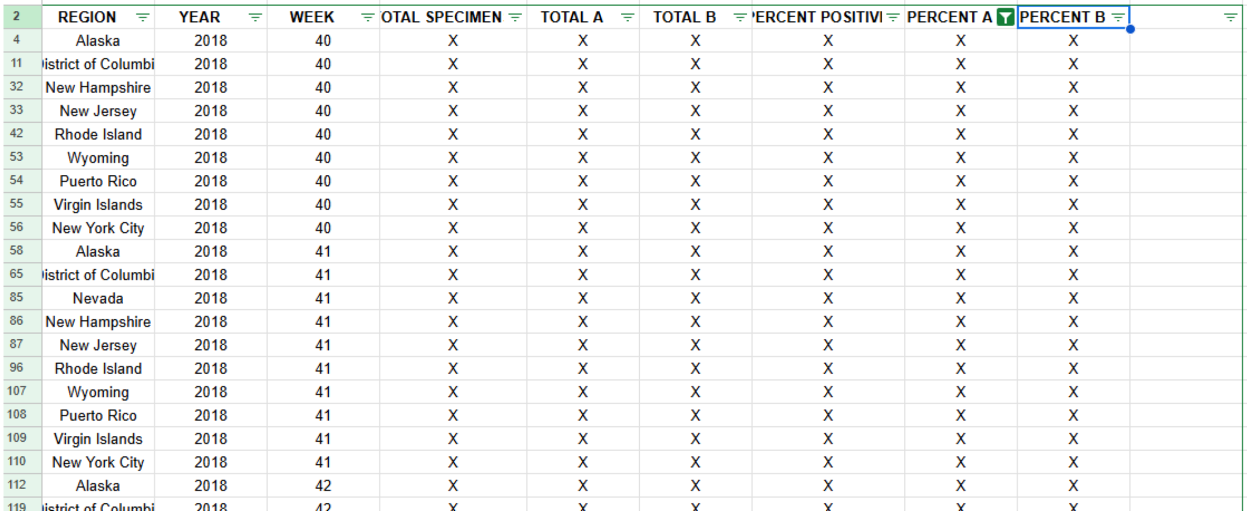


	A	B	C	D	E	F	G	H	I	J
1	*Beginning for the 2015-16 season, reports from public health and clinical laboratories are presented separately in the weekly influenza update, FluView. Data from clinical laboratories include the we									
2	REGION TYPE	REGION	YEAR	WEEK	TOTAL SPECIM	TOTAL A	TOTAL B	PERCENT POS	PERCENT A	PERCENT B
3	States	Alabama	2018	40	455	2	2	0.88	0.44	0.44
4	States	Alaska	2018	40	X	X	X	X	X	X
5	States	Arizona	2018	40	183	0	0	0	0	0
6	States	Arkansas	2018	40	86	1	0	1.16	1.16	0
7	States	California	2018	40	409	2	1	0.73	0.49	0.24
8	States	Colorado	2018	40	366	0	0	0	0	0
9	States	Connecticut	2018	40	383	1	1	0.52	0.26	0.26
10	States	Delaware	2018	40	147	1	0	0.68	0.68	0
11	States	District of Colum	2018	40	X	X	X	X	X	X
12	States	Florida	2018	40	1833	157	32	10.31	8.57	1.75
13	States	Georgia	2018	40	752	3	8	1.46	0.4	1.06
14	States	Hawaii	2018	40	62	2	0	3.23	3.23	0
15	States	Idaho	2018	40	24	0	0	0	0	0
16	States	Illinois	2018	40	97	1	2	3.09	1.03	2.06
17	States	Indiana	2018	40	136	0	1	0.74	0	0.74
18	States	Iowa	2018	40	180	0	0	0	0	0
19	States	Kansas	2018	40	62	0	0	0	0	0
20	States	Kentucky	2018	40	377	1	0	0.27	0.27	0
21	States	Louisiana	2018	40	243	4	0	1.65	1.65	0
22	States	Maine	2018	40	41	0	0	0	0	0
23	States	Maryland	2018	40	148	0	1	0.68	0	0.68
24	States	Massachusetts	2018	40	426	1	0	0.23	0.23	0
25	States	Michigan	2018	40	751	0	2	0.27	0	0.27
26	States	Minnesota	2018	40	344	0	0	0	0	0
27	States	Mississippi	2018	40	53	0	0	0	0	0
28	States	Missouri	2018	40	617	3	0	0.49	0.49	0
29	States	Montana	2018	40	113	1	1	1.77	0.88	0.88
30	States	Nebraska	2018	40	83	0	0	0	0	0

Like the ILINet we remove the Region Type Column. The total specimens indicate the number of flu tests performed. So there is also type A and B and there is the number of positive results of A and B. The percent positive is the percentage of specimens that tested positive for any flu type. Then the Percent A and B are the percentage of positive for Influenza for A and B respectively.

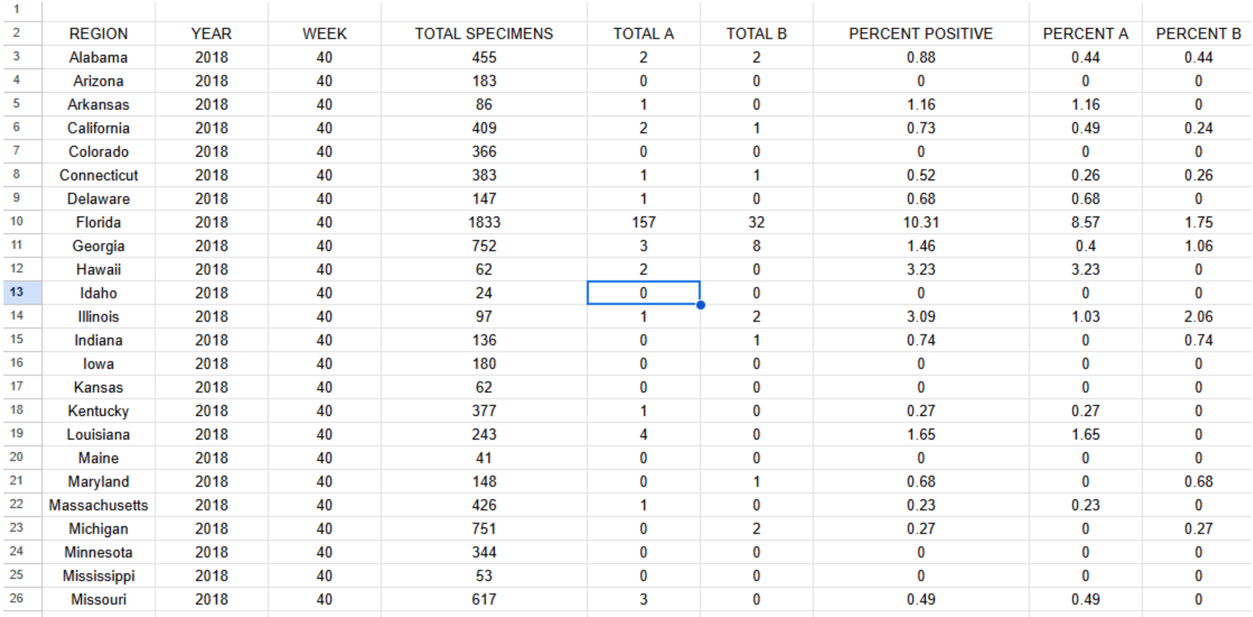
Data Cleaning:

There are like 531 rows with ‘X’ values in the total specimen out of 2808 rows.There are like 531 rows with ‘X’ values in the total A out of 2808 rows. There are 531 rows with ‘X’ values in the total B out of 2808 rows. The same number of rows are missing in PP and Percent A and B columns respectively. So, wherever there were the X values in Total specimens the whole row is missing. So we think it is better to delete the whole rows. So 531 rows out of 2808 will not be there.



2	REGION	YEAR	WEEK	OTAL SPECIMEN	TOTAL A	TOTAL B	ERCENT POSITIVI	PERCENT A	PERCENT B	
4	Alaska	2018	40	X	X	X	X	X	X	
11	istrict of Columbi	2018	40	X	X	X	X	X	X	
32	New Hampshire	2018	40	X	X	X	X	X	X	
33	New Jersey	2018	40	X	X	X	X	X	X	
42	Rhode Island	2018	40	X	X	X	X	X	X	
53	Wyoming	2018	40	X	X	X	X	X	X	
54	Puerto Rico	2018	40	X	X	X	X	X	X	
55	Virgin Islands	2018	40	X	X	X	X	X	X	
56	New York City	2018	40	X	X	X	X	X	X	
58	Alaska	2018	41	X	X	X	X	X	X	
65	istrict of Columbi	2018	41	X	X	X	X	X	X	
85	Nevada	2018	41	X	X	X	X	X	X	
86	New Hampshire	2018	41	X	X	X	X	X	X	
87	New Jersey	2018	41	X	X	X	X	X	X	
96	Rhode Island	2018	41	X	X	X	X	X	X	
107	Wyoming	2018	41	X	X	X	X	X	X	
108	Puerto Rico	2018	41	X	X	X	X	X	X	
109	Virgin Islands	2018	41	X	X	X	X	X	X	
110	New York City	2018	41	X	X	X	X	X	X	
112	Alaska	2018	42	X	X	X	X	X	X	
119	istrict of Columbi	2018	42	X	X	X	X	X	X	

So there are 2278 rows and there are 9 rows now.



1									
2	REGION	YEAR	WEEK	TOTAL SPECIMENS	TOTAL A	TOTAL B	PERCENT POSITIVE	PERCENT A	PERCENT B
3	Alabama	2018	40	455	2	2	0.88	0.44	0.44
4	Arizona	2018	40	183	0	0	0	0	0
5	Arkansas	2018	40	86	1	0	1.16	1.16	0
6	California	2018	40	409	2	1	0.73	0.49	0.24
7	Colorado	2018	40	366	0	0	0	0	0
8	Connecticut	2018	40	383	1	1	0.52	0.26	0.26
9	Delaware	2018	40	147	1	0	0.68	0.68	0
10	Florida	2018	40	1833	157	32	10.31	8.57	1.75
11	Georgia	2018	40	752	3	8	1.46	0.4	1.06
12	Hawaii	2018	40	62	2	0	3.23	3.23	0
13	Idaho	2018	40	24	0	0	0	0	0
14	Illinois	2018	40	97	1	2	3.09	1.03	2.06
15	Indiana	2018	40	136	0	1	0.74	0	0.74
16	Iowa	2018	40	180	0	0	0	0	0
17	Kansas	2018	40	62	0	0	0	0	0
18	Kentucky	2018	40	377	1	0	0.27	0.27	0
19	Louisiana	2018	40	243	4	0	1.65	1.65	0
20	Maine	2018	40	41	0	0	0	0	0
21	Maryland	2018	40	148	0	1	0.68	0	0.68
22	Massachusetts	2018	40	426	1	0	0.23	0.23	0
23	Michigan	2018	40	751	0	2	0.27	0	0.27
24	Minnesota	2018	40	344	0	0	0	0	0
25	Mississippi	2018	40	53	0	0	0	0	0
26	Missouri	2018	40	617	3	0	0.49	0.49	0

So now we have two datasets ILINet (2809x7) and ICL NREVSS Clinical Labs (2278x9).

Note:

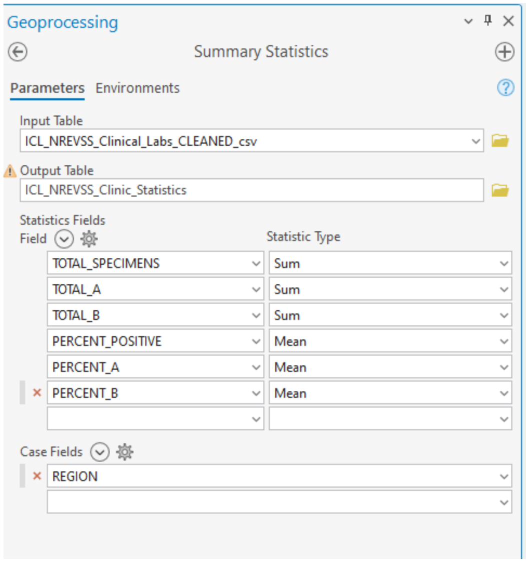
The Flu season timeline in CDC Format is from 2018 to 2019 which would mean Week 40 of 2018 to Week 52 of 2018 : October to December 2018
and Week 1 to Week 39 of 2019: jan to september 2019. So in order to avoid confusions we do feature engineering using the following formula in google sheets:

= IF(YEAR=2018, WEEK-39, WEEK+13). So there is a new feature called Season_Week which has values from 1 to 52.
We are still not deleting the YEAR and WEEK columns because we think that might be of some use in ArcGISPro3.

Process:
We projected the boundary file from NAD 83 (Geographic) to WGS 84(projected coordinate system).
Now we have uploaded the state boundary file and the two cleaned csv files to arcgispro3. Now we use the Joins and Relates and Add Joins from the csv files(region) and to the US State Boundaries shapefile (Name) and join it to the analysis.
Since the ILINet_CLEANED.csv and Clinical_Labs_CLEANED.csv contain multiple records per state (because of different SEASON_WEEKs), we are dealing with a one-to-many join, not a simple one-to-one join. Can't use the standard "Add Join" tool, which is only for one-to-one or many-to-one joins.
So we go ahead and import the CSV files as a table in the existing Geodatabase to then use the data management geoprocessing tools for that.

So first let us focus on creating a state level summary for the full flu season. So we will aggregate the columns
ILINET: ILITOTAL, TOTAL_PATIENTS, (SUM) PERC_UNWEIGHTED_ILI (MEAN)
ICL_NRVESS: TOTAL_SPECIMENS, TOTAL_A, TOTAL_B, (SUM) PERCENT_POSITIVE, PERCENT_A and PERCENT_B (MEAN)

So, then:



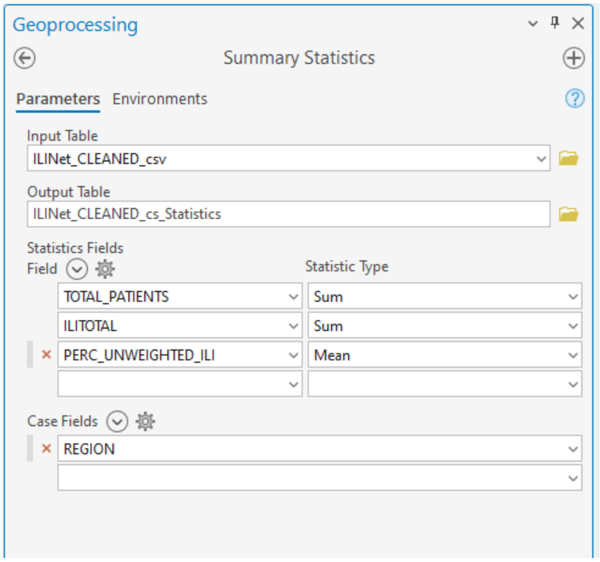
Input Table
ICL_NREVSS_Clinical_Labs_CLEANED_csv

Output Table
ICL_NREVSS_Clinic_Statistics

Statistics Fields

Field	Statistic Type
TOTAL_SPECIMENS	Sum
TOTAL_A	Sum
TOTAL_B	Sum
PERCENT_POSITIVE	Mean
PERCENT_A	Mean
PERCENT_B	Mean

Case Fields
REGION



Input Table
ILINet_CLEANED_csv

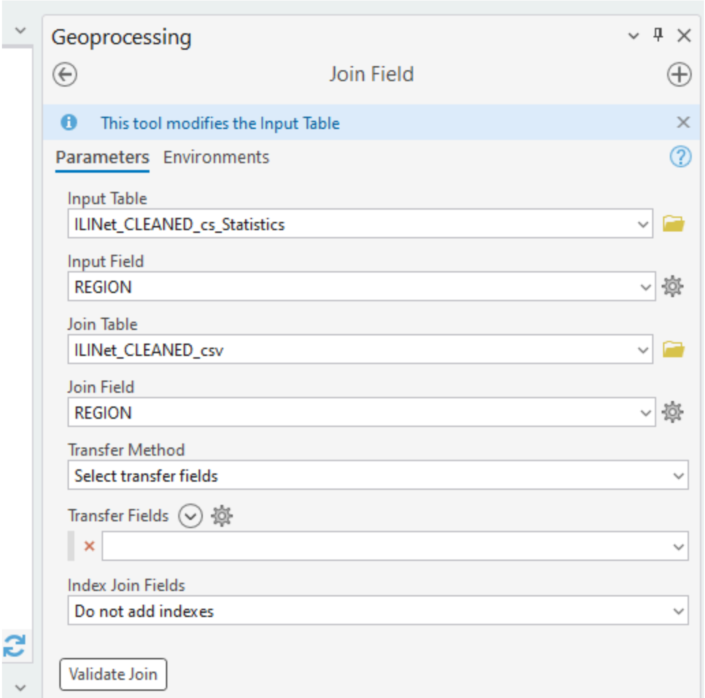
Output Table
ILINet_CLEANED_cs_Statistics

Statistics Fields

Field	Statistic Type
TOTAL_PATIENTS	Sum
ILITOTAL	Sum
PERC_UNWEIGHTED_ILI	Mean

Case Fields
REGION

Then, using the ‘Join Field’ tool from the geoprocessing toolbox, we join the ILI_summary table to the US State Boundaries using Region and Name Columns. Then we can merge the ICL_summary to the existing table.



This tool modifies the Input Table

Input Table
ILINet_CLEANED_cs_Statistics

Input Field
REGION

Join Table
ILINet_CLEANED_csv

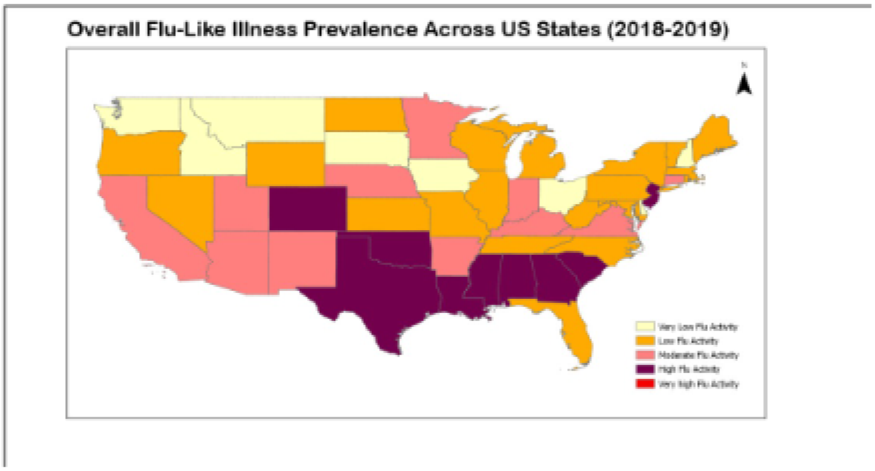
Join Field
REGION

Transfer Method
Select transfer fields

Transfer Fields

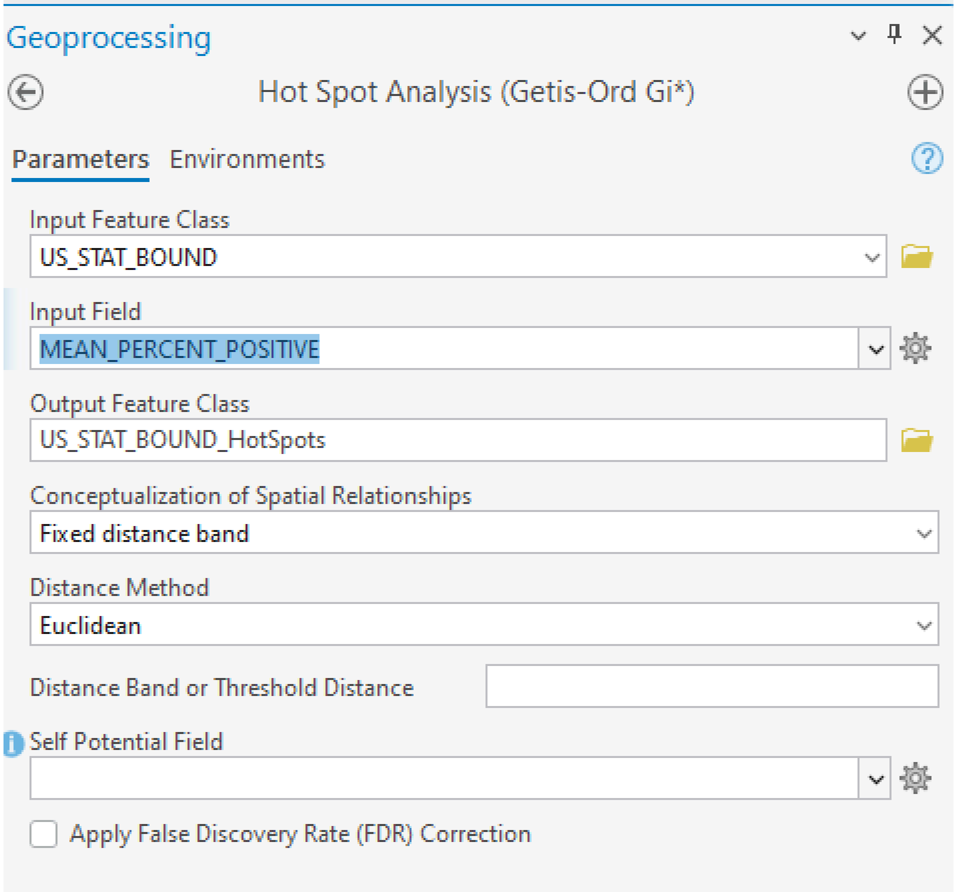
Index Join Fields
Do not add indexes

So, first we do some basic mapping of the overall flu prevalence. Symbology is Graduate Colors (which we eventually change into colors which we want). We use the mean percentage of unweighted ILI. The natural breaks are used for classification and we classify it into five classes.



Now we try to create a Strain dominance map which shows which type was more dominant during the season in each state. But, it seems that Type A was dominant in all of the states so we don’t create a map for that.

Next we try to do Hotspot Analysis.



Geoprocessing

Hot Spot Analysis (Getis-Ord Gi*)

Parameters Environments

Input Feature Class
US_STAT_BOUND

Input Field
MEAN_PERCENT_POSITIVE

Output Feature Class
US_STAT_BOUND_HotSpots

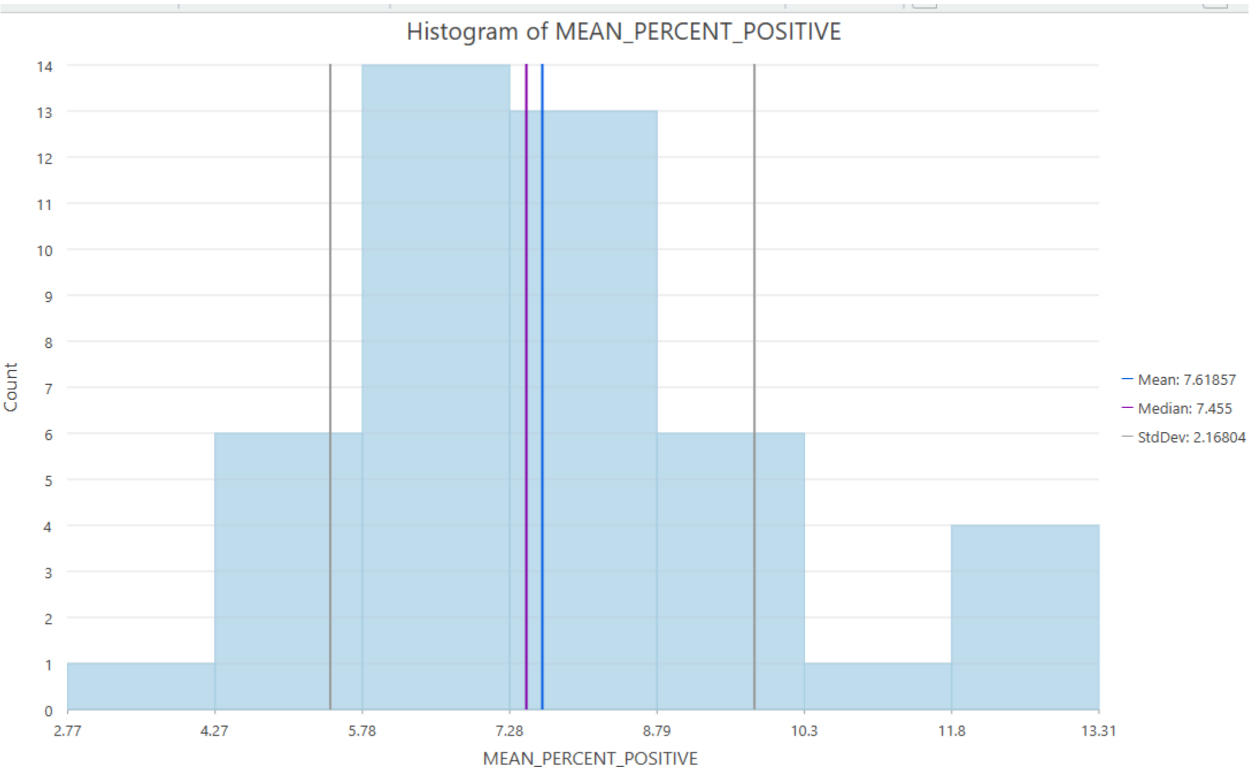
Conceptualization of Spatial Relationships
Fixed distance band

Distance Method
Euclidean

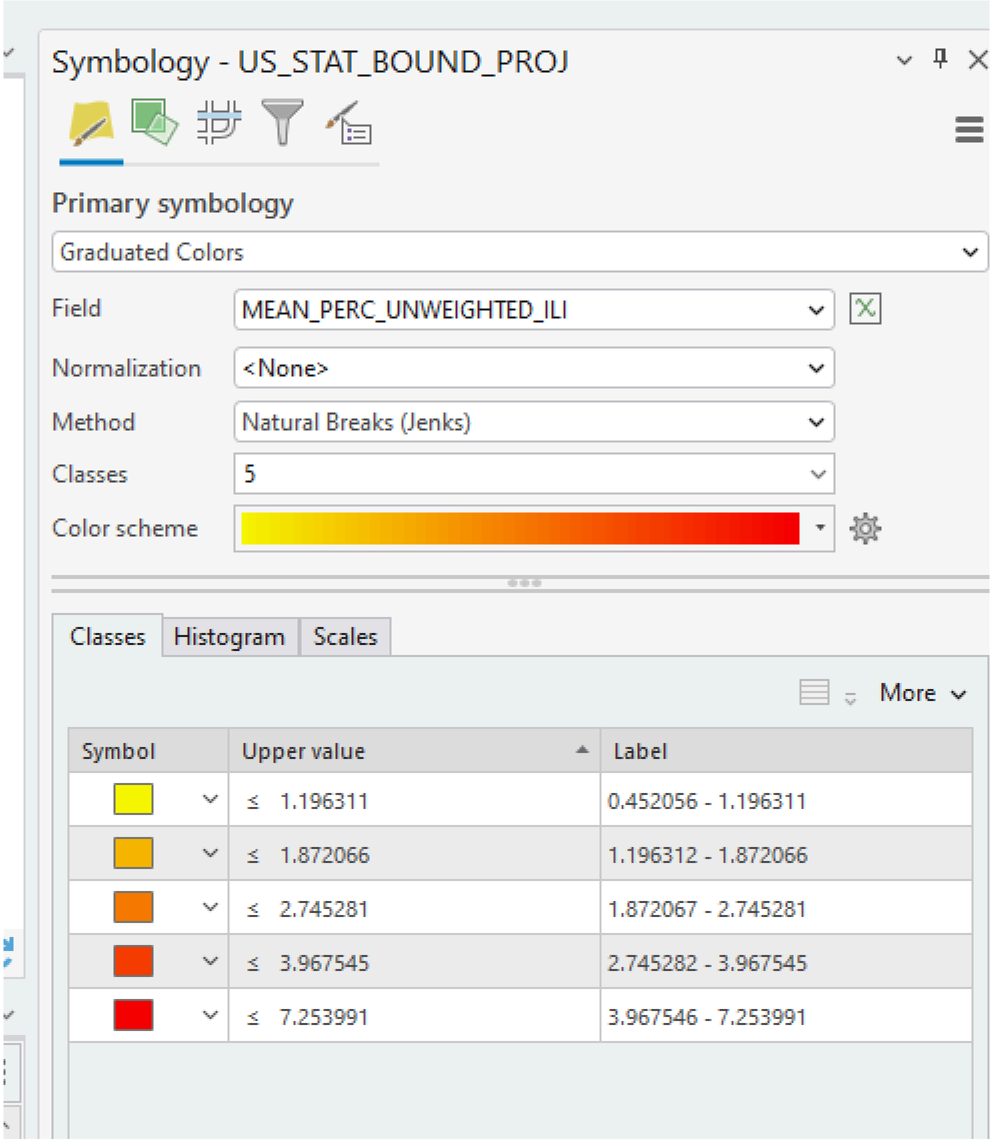
Distance Band or Threshold Distance

Self Potential Field

□ Apply False Discovery Rate (FDR) Correction



Now we start with visualisations:



Symbology - US_STAT_BOUND_PROJ

Primary symbology

Graduated Colors

Field: MEAN_PERC_UNWEIGHTED_ILI

Normalization: <None>

Method: Natural Breaks (Jenks)

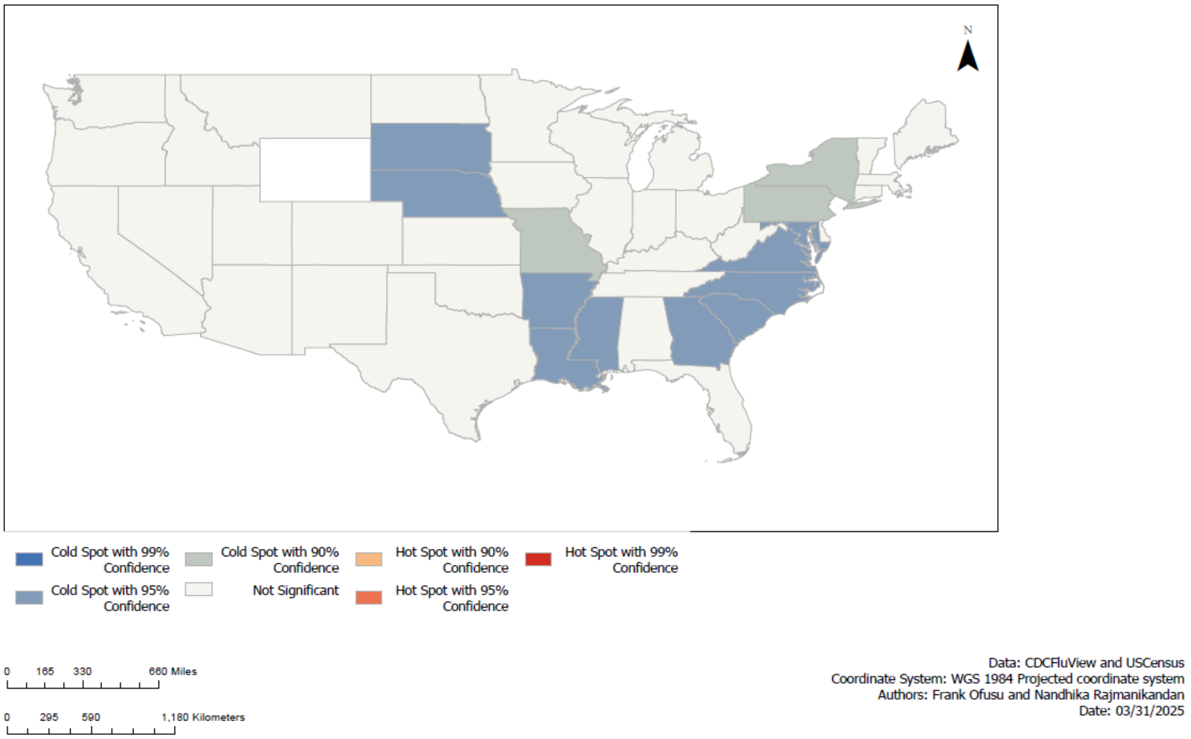
Classes: 5

Color scheme

Symbol	Upper value	Label
🟨	≤ 1.196311	0.452056 - 1.196311
🟧	≤ 1.872066	1.196312 - 1.872066
🟧	≤ 2.745281	1.872067 - 2.745281
🟥	≤ 3.967545	2.745282 - 3.967545
🟥	≤ 7.253991	3.967546 - 7.253991

Hotspot analysis is a spatial statistical method used to identify areas that show statistically significant clustering of high or low values. In the context of public health, hotspot analysis helps detect regions where disease prevalence, healthcare resource usage, or risk factors are unusually concentrated. By applying tools such as the Getis-Ord Gi* statistics, one can determine whether high flu positivity rates or high concentrations of risk factors (like multigenerational households or ILI cases) are spatially clustered beyond what would be expected by chance. In our study, hotspot analysis could be used to uncover geographic clusters of elevated lab-confirmed flu rates and explore whether these hotspots align with areas of increased household transmission risk or overburdened healthcare systems. This would provide valuable insight into localized vulnerabilities, enabling more targeted interventions in future flu seasons.

HOTSPOT ANALYSIS OF FLU SPOTS Across US States (2018-2019)


N
Cold Spot with 99% Confidence
Cold Spot with 95% Confidence
Cold Spot with 90% Confidence
Not Significant
Hot Spot with 90% Confidence
Hot Spot with 95% Confidence
Hot Spot with 99% Confidence
0 165 330 660 Miles
0 295 590 1,180 Kilometers

Data: CDCFluView and USCensus
Coordinate System: WGS 1984 Projected coordinate system
Authors: Frank Ofusu and Nandhika Rajmanikandan
Date: 03/31/2025

Interpretation:

The hotspot analysis map illustrates the spatial clustering of flu positivity rates across U.S. states during the 2018–2019 season, highlighting statistically significant hot and cold spots using the Getis-Ord Gi* statistic. States in the southeastern region—such as Georgia, South Carolina, Alabama, and Mississippi—emerge as hot spots with 95% confidence, indicating that these areas experienced higher-than-expected flu positivity rates that are not due to random chance. Conversely, states like North Dakota, Nebraska, and Montana appear as cold spots, showing significantly lower flu activity at 90–95% confidence. Several other states are marked as not significant, meaning their flu rates did not show a strong spatial pattern. This spatial distribution suggests that flu outbreaks during this season were not uniform but geographically clustered, reinforcing the importance of regional surveillance and localized interventions. The clustering of hot spots in the Southeast may point to socio-environmental factors, healthcare access disparities, or demographic patterns that warrant further investigation.

Next we will take multigenerational households data from ACS Website (data.census.gov)

Link:https://data.census.gov/table/ACSDT1Y2019.B11017?q=B11017:+MULTIGENERATIONAL+HOUSEHOLDS&g=010XX00US$0400000_040XX00US01,02,04,05,06,08,09,10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42&moe=false&tp=true

Data is Cleaned through Python: (Google Collab file or link)

The multigenerational household dataset represents the number and proportion of households in each U.S. state where three or more generations live together, derived from the American Community Survey (ACS) Table B11017 2018. The key variable, ‘pct_multigen’, reflects the percentage of multigenerational households out of all households in a state, offering important insights into potential influenza transmission dynamics. Multigenerational living arrangements are particularly relevant in public health contexts, as they can increase the risk of within-household flu spread especially from school-aged children to elderly family members, who are more vulnerable to severe illness. States with high percentages of such households, like California (5.8%), may experience elevated transmission risk due to housing density, cultural factors, and challenges with home isolation. Integrating this variable with flu surveillance data enables spatial comparisons to identify whether higher household mixing correlates with greater flu burden. This dataset, when used alongside socioeconomic indicators such as income or insurance coverage, supports a more holistic understanding of community-level vulnerability and can guide targeted public health interventions.

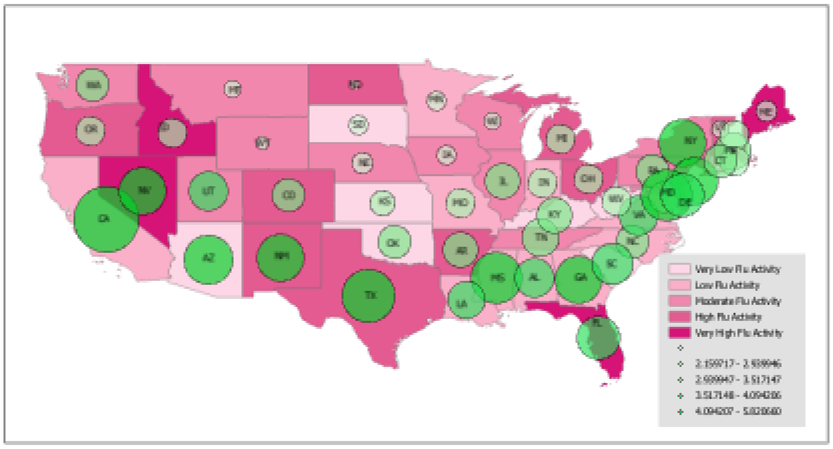
Process:

We have now included the Multigen_HH csv geotable also into the State Boundaries table using the ‘Join Field’.
Now we try to visualize the trends

First for the base map, we use the mean positive percentage and map flu activity across states. Then we use the percentage of multigenerational households to show the vulnerability and number of cases.

We feel that in a multigenerational household both grandparents and children would be vulnerable and so there might exist a correlation between those factors.

The correlation between that is still yet to be determined using python.



Machine Learning: Random Forest Model

```
R² Score: -0.45973398968773393
RMSE: 2.460987586347182
```

```
A worker process managed by the executor was unexpectedly terminated. This
could be caused by a segmentation fault while calling the function or by an
excessive memory usage causing the Operating System to kill the worker.
```

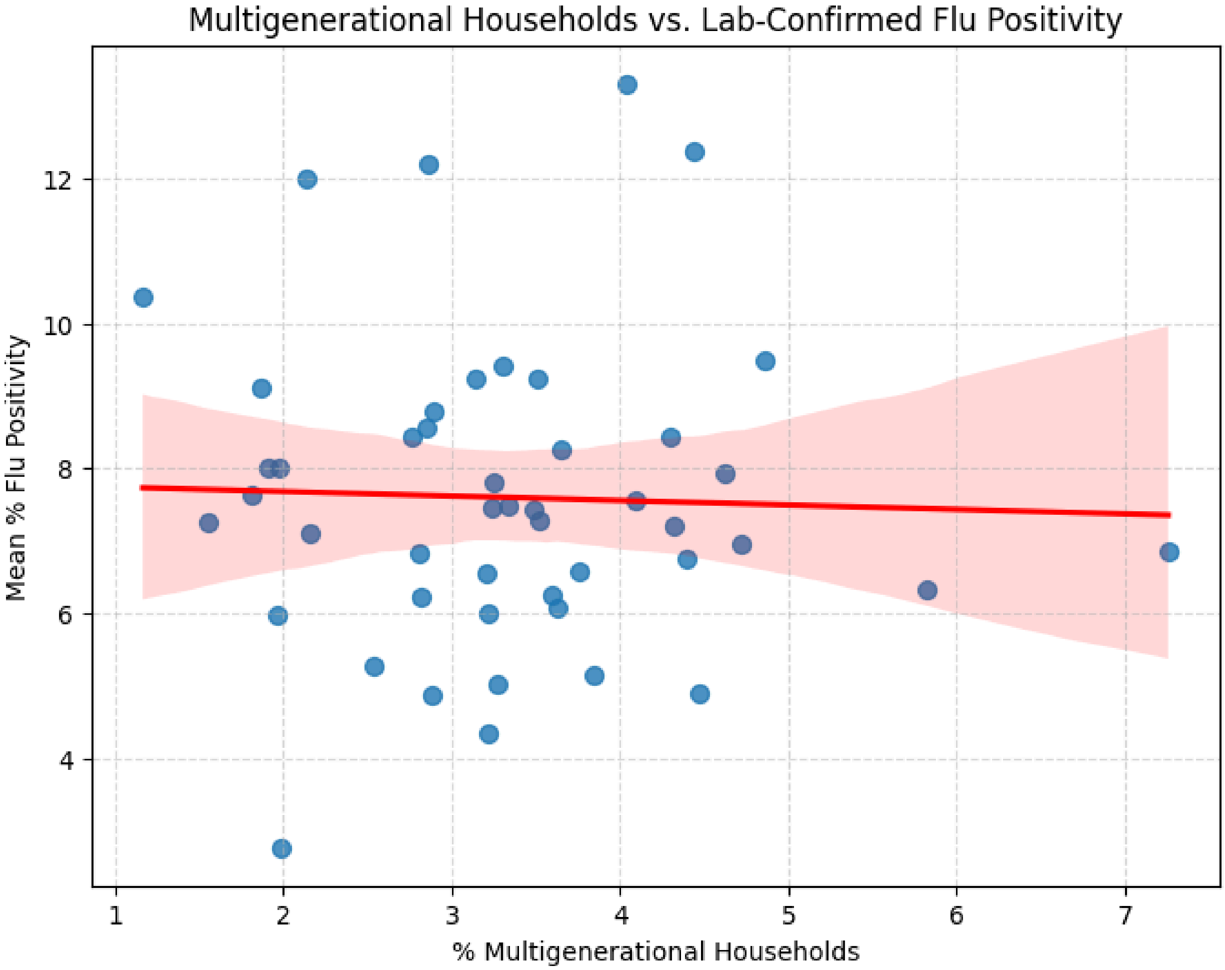
```
[20]:
```

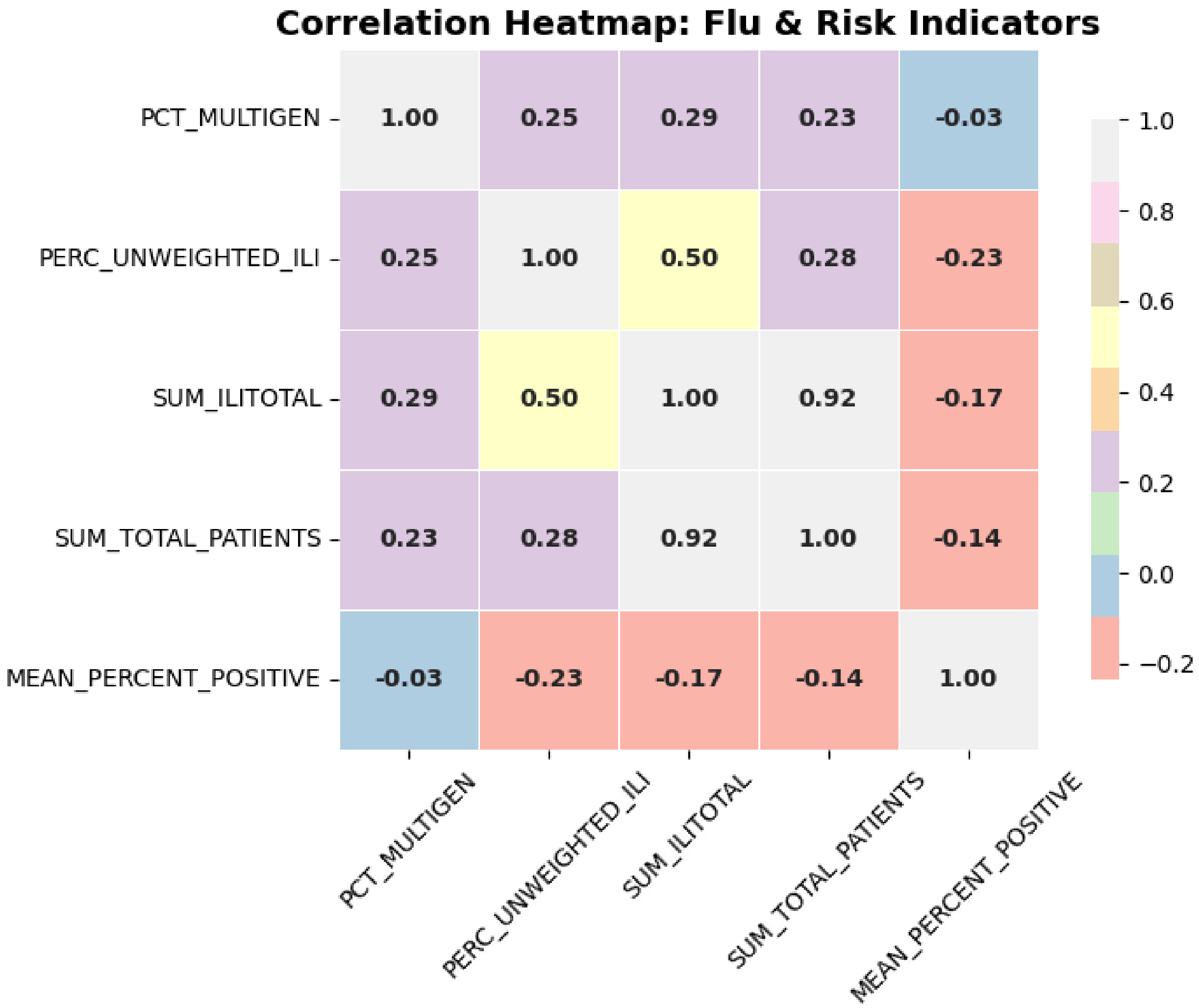
```
Selection deleted
```

We got the R square value and it does not seem good. Just 45% variance is very bad. So, in order to try different models, we are moving I think it outside of ArcGISPro3

We have used GridSearchCV as well, but it tends to give a lower R2 value compared to what we previously got. So, I think there is no point in visualizing the map for it

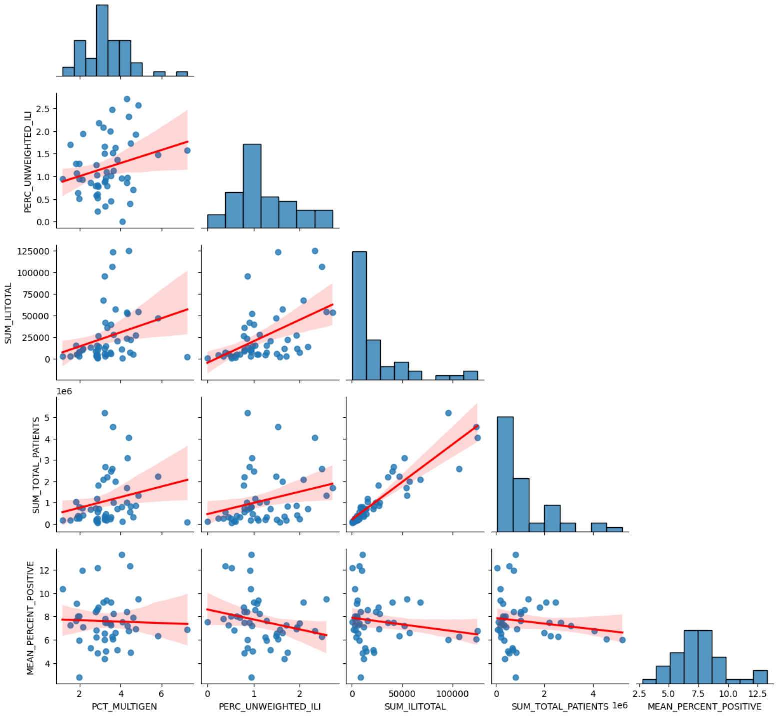
On the other hand, the correlation between the Percentage of MGH and LCDP also is not that great. Hence, there is no point in going in further for a machine learning model here.


Multigenerational Households vs. Lab-Confirmed Flu Positivity
Mean % Flu Positivity
% Multigenerational Households
4
6
8
10
12
1
2
3
4
5
6
7



Correlation Heatmap: Flu & Risk Indicators

Exploratory Data Analysis:


Pairwise Relationships Between Flu & Risk Features

The pairwise scatter plots and distribution histograms provide a comprehensive visual overview of the relationships between multigenerational household prevalence, influenza-like illness (ILI) reporting, patient volume, and lab-confirmed flu positivity across U.S. states. Among all variable pairs, the strongest linear relationship is observed between `SUM_ILITOTAL` and `SUM_TOTAL_PATIENTS`, which is expected as both represent patient volume metrics and are naturally correlated. Similarly, `PERC_UNWEIGHTED_ILI` (symptom-reported flu) and `SUM_ILITOTAL` show a mild positive trend, suggesting that as more individuals report flu-like symptoms, the number of ILI cases also increases.

Interestingly, `PCT_MULTIGEN` (percentage of multigenerational households) exhibits little to no visible linear association with the target variable `MEAN_PERCENT_POSITIVE` (lab-confirmed flu positivity), as shown by the nearly flat regression line and widely scattered data points. This suggests that multigenerational household structure, at the state level, may not be a strong standalone predictor of lab-confirmed flu prevalence. Additionally, the histogram for `MEAN_PERCENT_POSITIVE` indicates a slightly right-skewed distribution, while features like `SUM_TOTAL_PATIENTS` and `SUM_ILITOTAL` show long tails, implying skewed population distributions across states.

Overall, these visualizations reinforce the findings from your modeling: while volume-based variables (like ILI counts and patient totals) relate to one another, their direct predictive power on lab-confirmed flu positivity is limited without additional contextual or demographic variables.