**Flu In the U.S.:A Closer Look**

**Team 3: Morgan, Qin, Ryan, So Jung.**

**Objective:**

Last year, the CDC estimated that 49 million individuals in the United States contracted influenza, 960,000 were hospitalized, and about 79,000 individuals died due to respiratory illness. This makes the 2017-2018 flu season the largest outbreak since 2009. The objective of this research was to use publicly available information to determine what socioeconomic, geographic, and other factors may have influenced the severity of the outbreak. The following report summarizes our findings.

**Data Used:**

* Flu Occurrence data: CSV file from the CDC website - specifically the flu occurrence data which is collected by both the World Health Organization labs and National Respiratory and Enteric Virus Surveillance System labs.
  + Disclaimer: CDC ILINet (Influenza Like Illness) data was used in determining a baseline of the respiratory illness. However, it is important to note that the ILINet Surveillance data only accounts for about 8.5% of the US population. Therefore, all information used was analyzed as a direct proportion of the United States to evaluate the country as a whole.
* US Census data: API from the Census Bureau as it relates to the United States population. We used the Census Wrapper available within Python to pull our data in after identifying the variables we needed from the technical documentation for the Census API. The data pulled was at the state level, and since we used the ACS5 census API, the data pulled is in the format of five-year estimates. The importance of this is that there is increased statistical reliability of the data for less populated areas and small population subgroups. The specific columns that we pulled for each state include:
  + Household Income
  + Population (Total)
  + Population (Under 18)
  + Population (Over 65)
  + Household Per Capita Income
  + Median Age
  + Poverty Count
  + Unemployment Count
  + Number of Bachelor Degrees
  + Healthcare Coverage (%) in each state

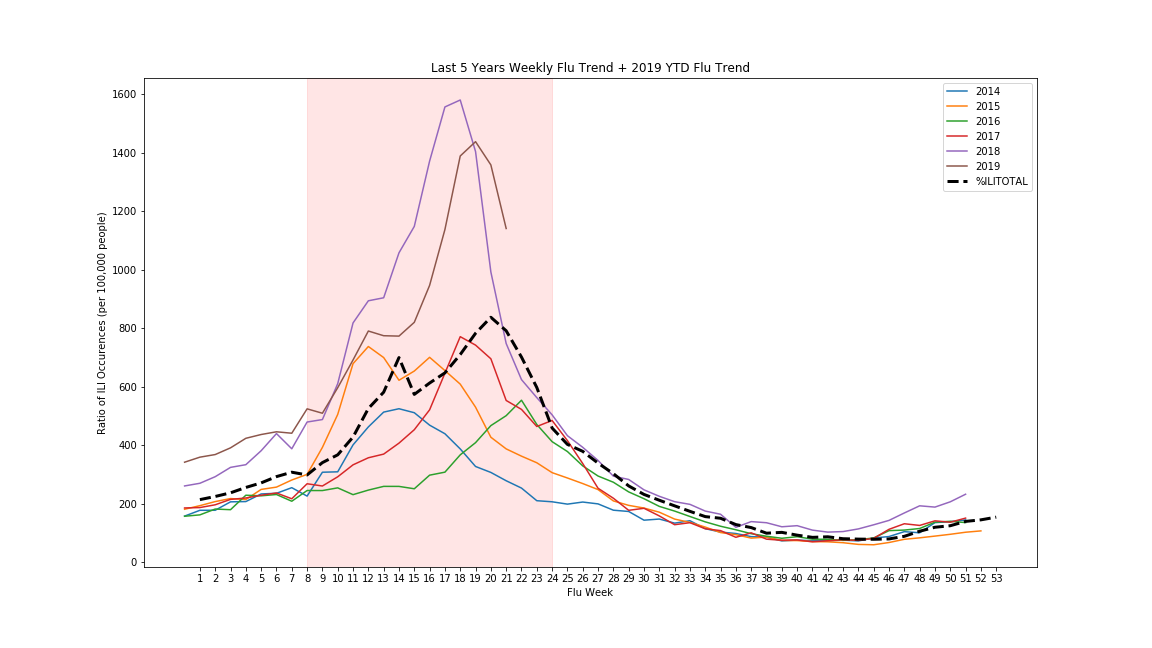
**Cleaning the Data:**

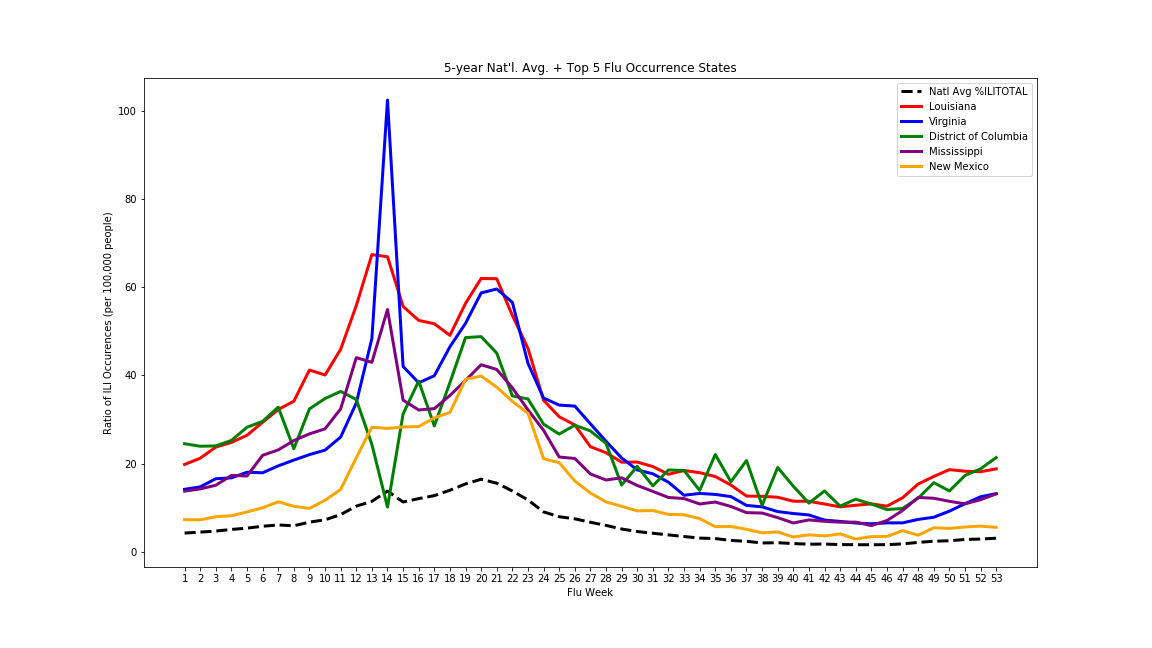
* Flu Data cleaning:
  + Each state’s 5-year total ILI reported numbers were divided by five to get an annual number. To reduce risks about more ILI numbers that are affected by highly populated states, we divided the number by state population. In order to normalize the number, we multiply it by 10,000. Thus, our ILI numbers represent 1/10,000 people.
* Census Data cleaning:
  + Format data types: Census data were the average numbers of 5 years. We took all people counting numbers and divide the numbers by each state population to proportionate them. Then, the numbers were multiplied by 10,000. However, unemployment and aggregate travel time(commuting time) were divided by (total population-population under 18). These two numbers are related to the job, we deduct population under 18 from the total population because they are not involved in these categories. Two income related numbers were not modified.
* Florida, City of New York, and Virgin Island were deleted from our data set since those areas’ one of flu data or census data were not available.

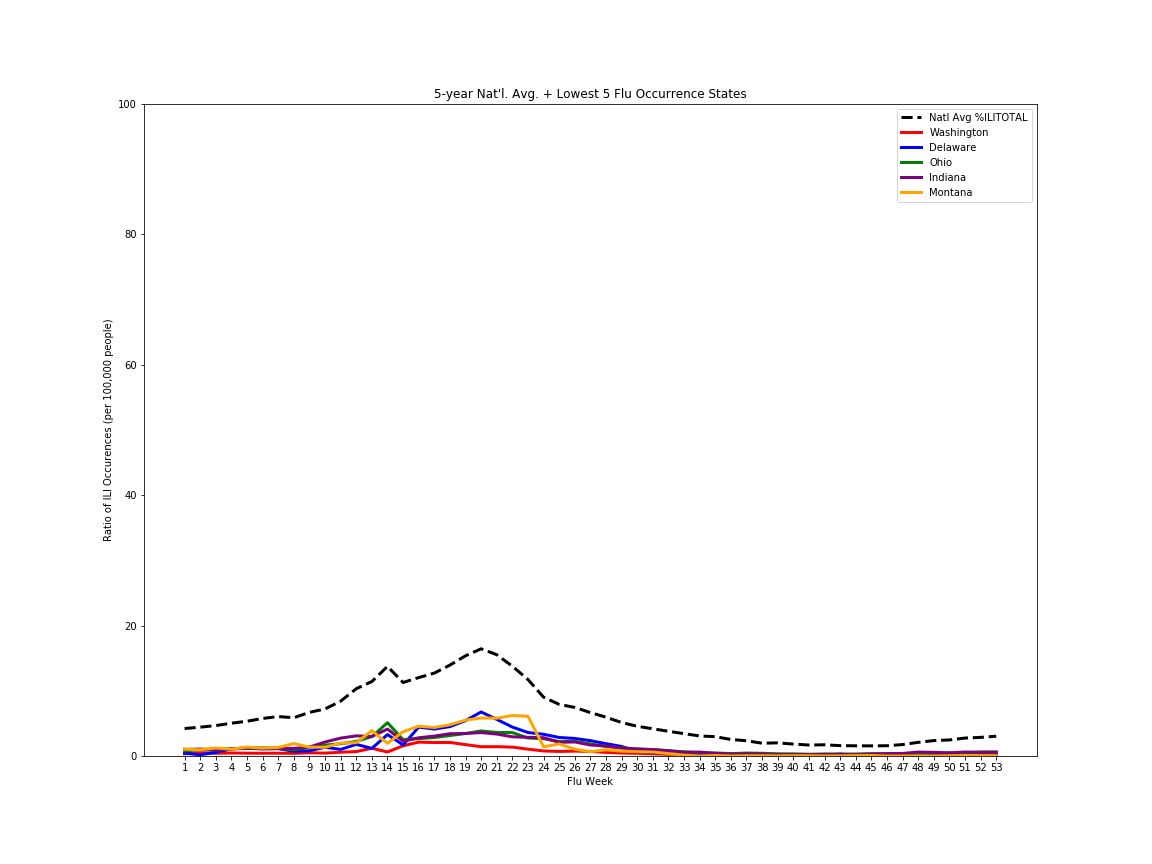
**Question: How do the last five flu seasons stack up against each other and the current 2019 flu season?**

**Data Used:**

* CDC ILINET data (2013-2019)







In comparing states and their average flu occurrence ratio of the last five years, we determined that Louisiana, Virginia, DC, Mississippi and New Mexico experienced the highest rates of flu. Contrarily, Washington state, Delaware, Ohio, Indiana, and Montana tend to have lower rates of the flu. The black dotted line represents the national average for all states.

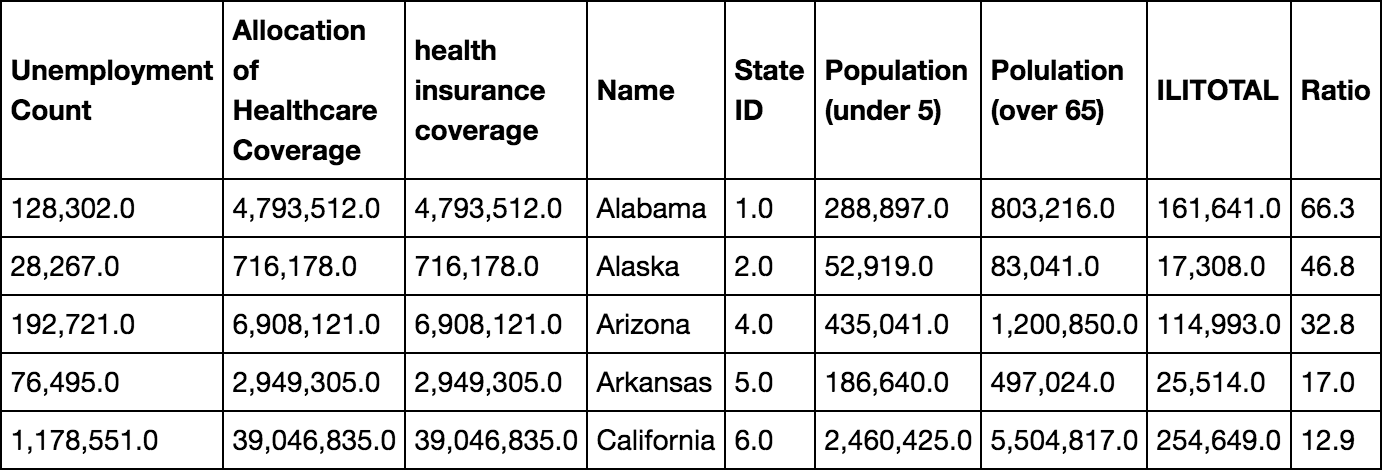
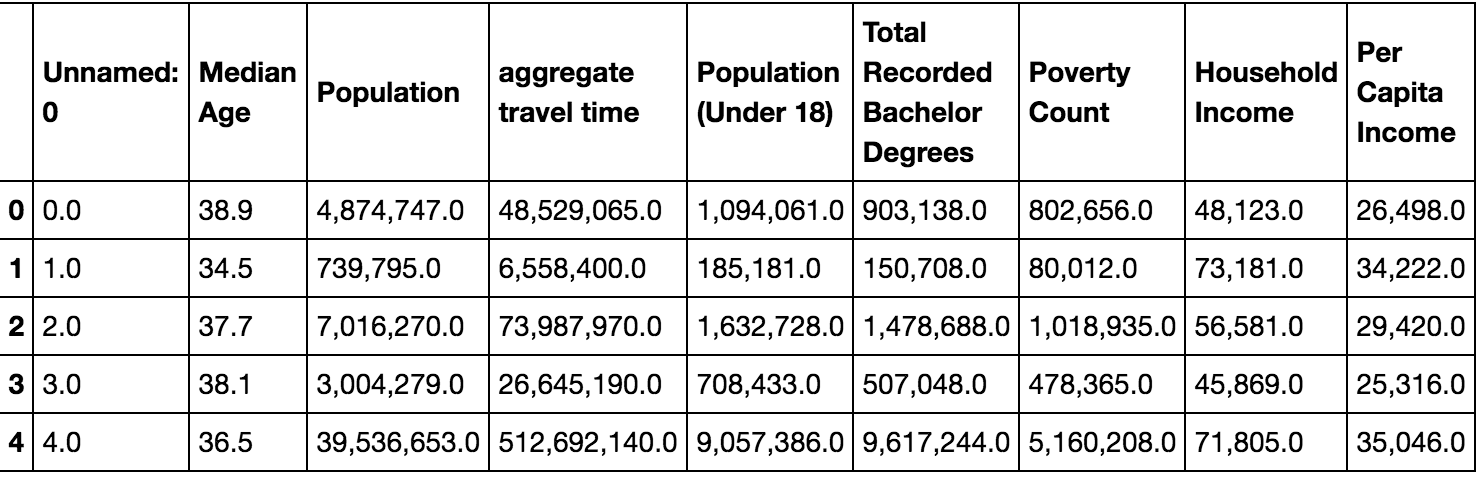
**Question: Which states or regions have a higher proportion of the flu?**

**Data Used:**

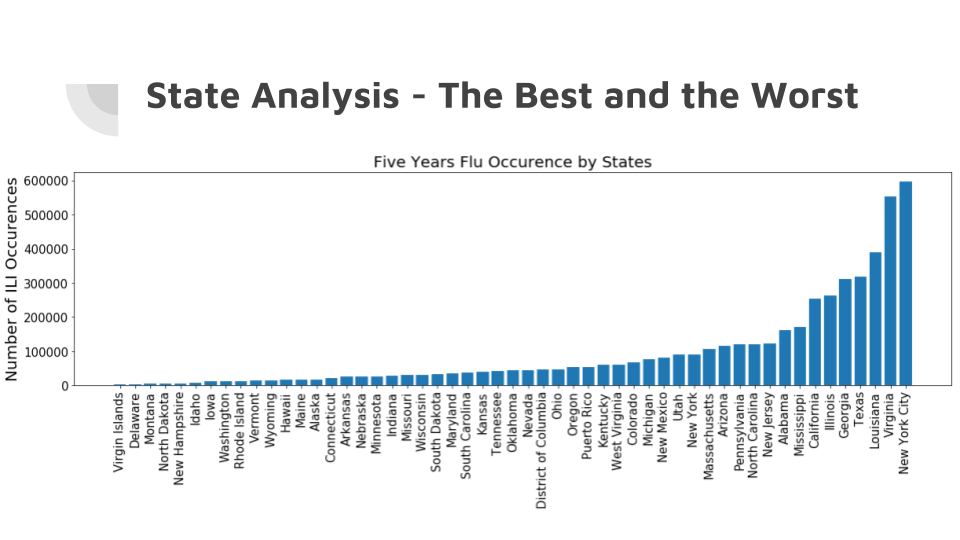
* Census.gov
* Census Wrapper
* CDC ILINET data (2013-2019)

**Cleaning the Data:**

For this section, we pulled data from the ACS5 API. The 5-year estimates from the ACS are "period" estimates that represents data collected over a period of time. The primary advantage of using multi-year estimates is the increased statistical reliability of the data for less populated areas and small population subgroups. From here, we cleaned flu data to remove all NANs and merged it with the census data to create the following table:

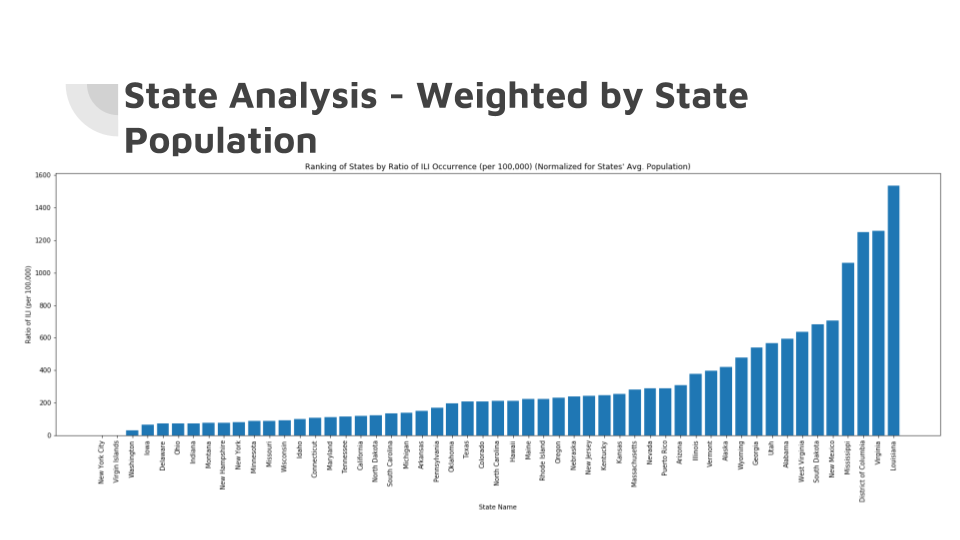


First, we looked at the total number of flu occurrences across all states and then looped through the latitudes and longitudes of each state and used a heat map to represent the flu dispersion. However, these charts are not proportionate to state population counts. Therefore, in order to get a better representation of the density of the flu per state population, we adjusted for the weight of state populations, which accounts for the difference between the first two and second two graphs.

****



The above two graphics show the number of total occurrences within each state. This was interesting to look at but not meaningful in terms of comparing states as the states with larger circles could be due to greater populations relative to their counterparts. In order to address this, we had to normalize the data by dividing the number of occurrences into each states estimated population. In doing so, we created the two visualizations below. States such as Wyoming and South Dakota, which have relatively low number of occurrences compared to other states, actually showed they had a higher percentage occurrence when normalized for population total.





In all, there wasn’t a huge change. The Southeastern and Eastern parts of the U.S. tend to have higher occurrences of the flu, with the exception of a few states in the Midwest. This normalized percentage occurrence of the flu is what we used for comparison in our statistical analysis, rather than using the raw number. Again, it’s important to note that the raw numbers we used only account for a percentage of the actual population (one of the limits of our dataset from the CDC).

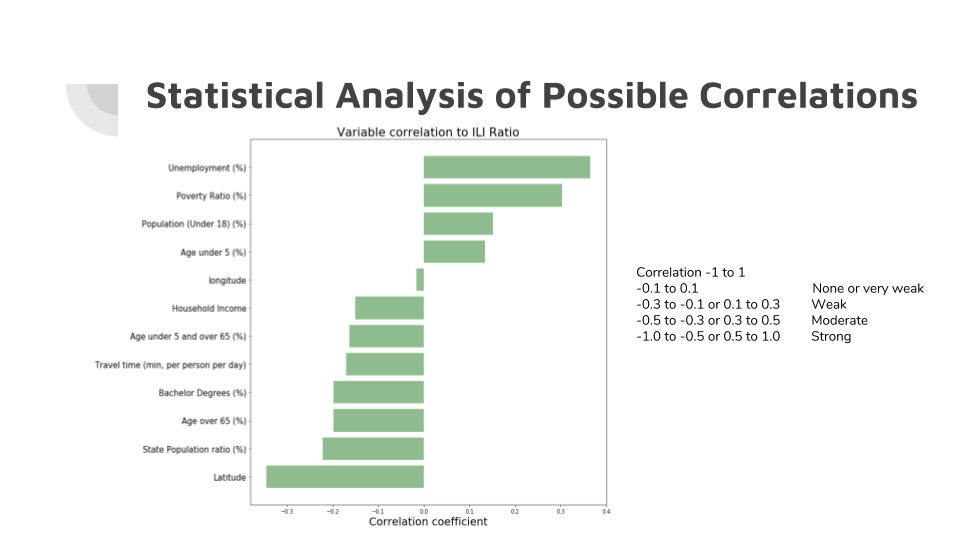
**Question: Are there any correlations with Flu Data and Census Data?**

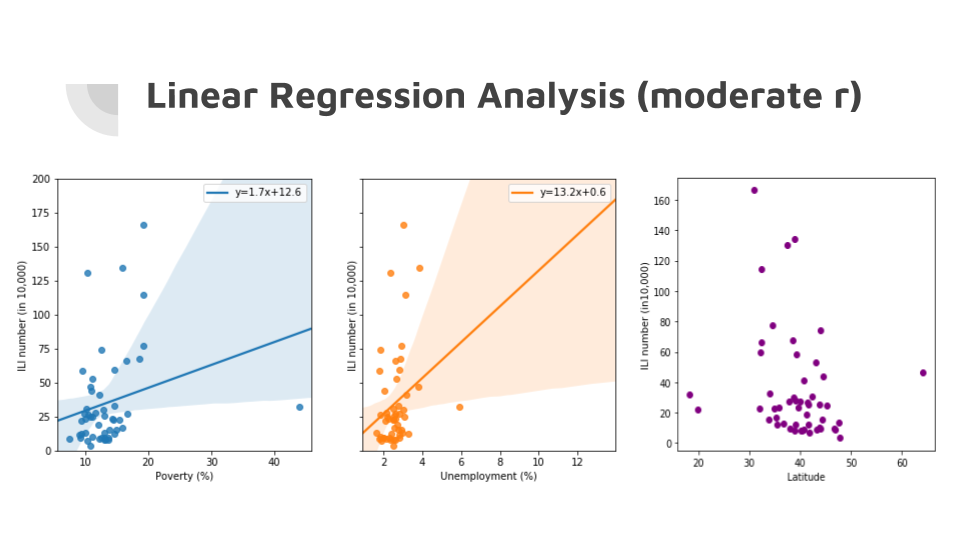
**Data Used:**

* Census.gov
* Census Wrapper
* CDC ILINET data (2013-2019)

**Cleaning the Data:**

This section used the same dataset as the previous Census sections. However, we also used “from scipy.stats import spearmanr” to uncover the Variable Correlation to ILI Ratio.





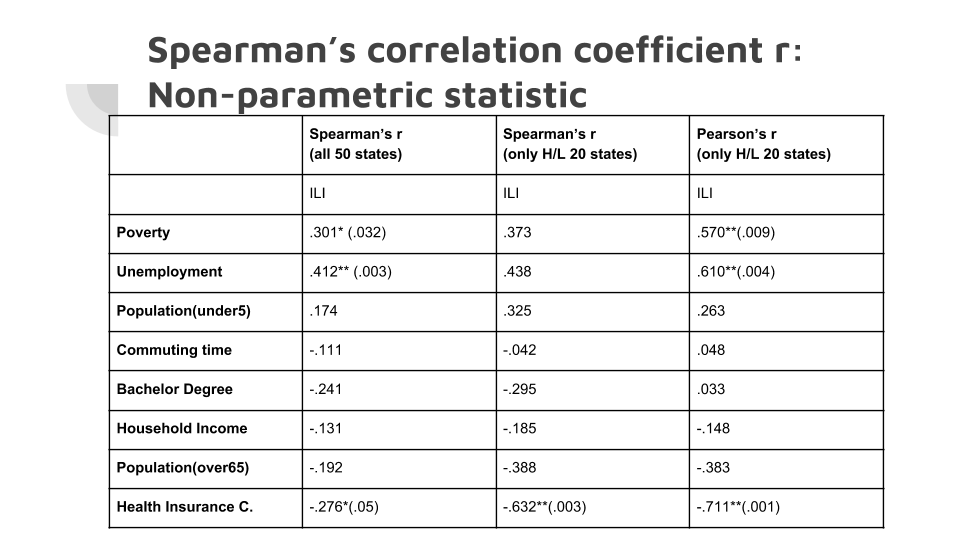
**Findings:**

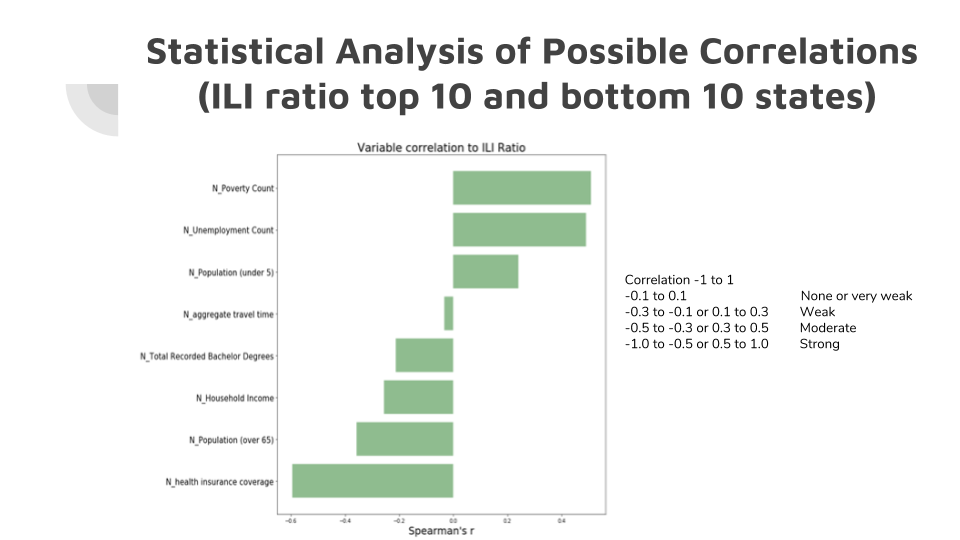
As seen in the graphs above, there is a slight positive correlation with flu occurrences in the poverty graph (left) and the unemployment graph (middle). The latitude graph (right) has no significant correlation considering the majority of the united states resides between the 30 and 50 latitude lines. Nevertheless, the first two graphs seem to imply that the higher state unemployment and state poverty levels suggests a higher risk for contracting the flu.

**Question: What deeper insight can we uncover with the correlation of multiple variables using SPSS?**

**Data Analysis:**

* First, look at the scatter plots of each socio-economic variable with ILI. After we found moderate relationships from the scatter plots, then, Spearman’s correlation coefficients were computed using all 50 states. (Scatter plots and Spearman’s r were obtained from Python, and Spearman’s r and Pearson’s r also were calculated using the SPSS statistical program.)
* There are big disparities among higher flu and lower flu occurring states, we need to look at if the differences are meaningful.
  + So, we sorted the state’s ILI flu number from top to bottom and only selected opt 10 and bottom 10.
  + We checked their correlation coefficients and tested independent sample t-tests using only the top 10 and bottom 10 states. (These were both conducted by Python and SPSS).





**Findings:**

* Even though scatter plots show moderate relationships among socio-economic variables with ILI flu number, significant correlation coefficients were found amongst only the ‘Poverty’, ‘Unemployment’ (positive correlation) and ‘Health Insurance Coverage’ (negative correlation) variables.
* When we tested correlations using only 20 states, all correlation coefficients were improved and those 3 variables’ correlations were still moderately significant. (Although Spearman’s r improved, two of the significances disappeared. Since Spearman’s r is a more constricted measurement when measuring r, we also checked Pearson’s r.)
* Independent sample T-Test was conducted to confirm whether the two groups’ mean differences were statistically different. Significant t-test result was obtained.
* The ‘Poverty’, ‘Unemployment’, and ‘Health insurance coverage’ variables are significantly correlated with the ILI flu numbers and two groups’ (High/Low flu occurring state) differences were significantly different, these numbers should not be used as predictors. We need further analysis.

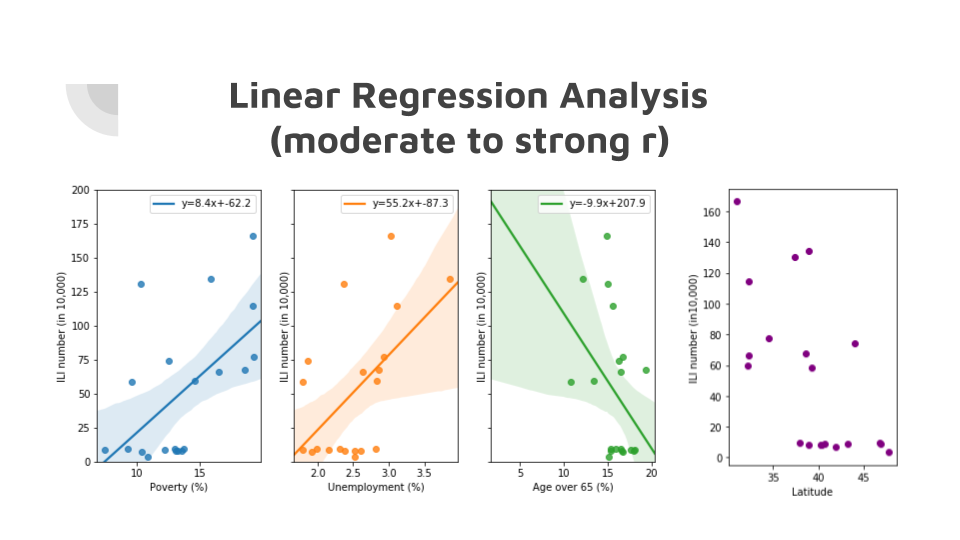
**Question: What other factors correlate with Flu Data?**

**Data Used:**

* Census.gov
* Census Wrapper
* CDC ILINET data (2013-2019)

**Cleaning the Data:**

Again, the same flu and Census data was used as before. However, this time we ran a regression on age over 65 and Latitude as well.



**Findings:**

Interestingly, there is a negative correlation between Age over 65 and flu occurrences (green). This means that the states that had a larger percent of their population made up of older individuals did not positively correlate like one may assume.

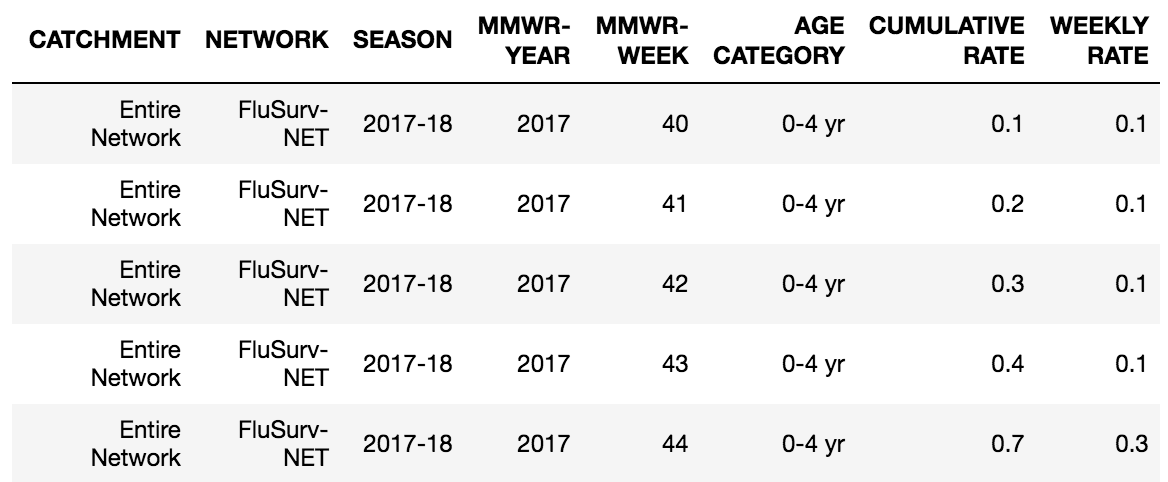
**Question: What can these correlations tell us about age groups affected/How does the flu affect hospitalization rates?**

**Data Used:**

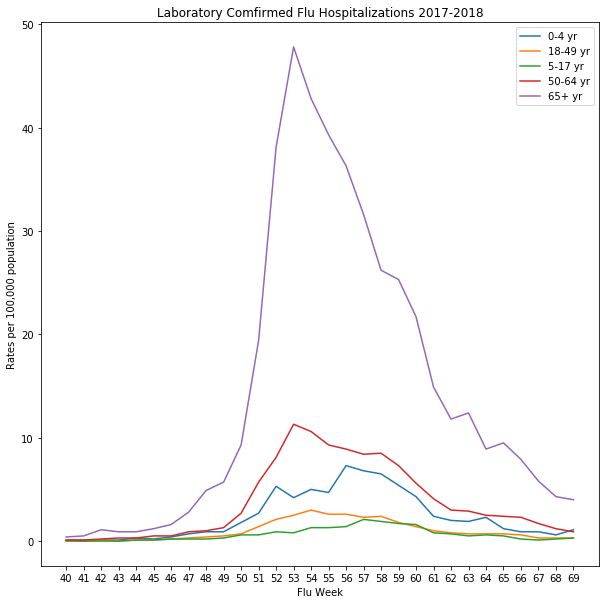
* CDC hospitalization rates 2017-2018 season
* CDC hospitalization rates 2013-2018 season

**Cleaning the Data:**

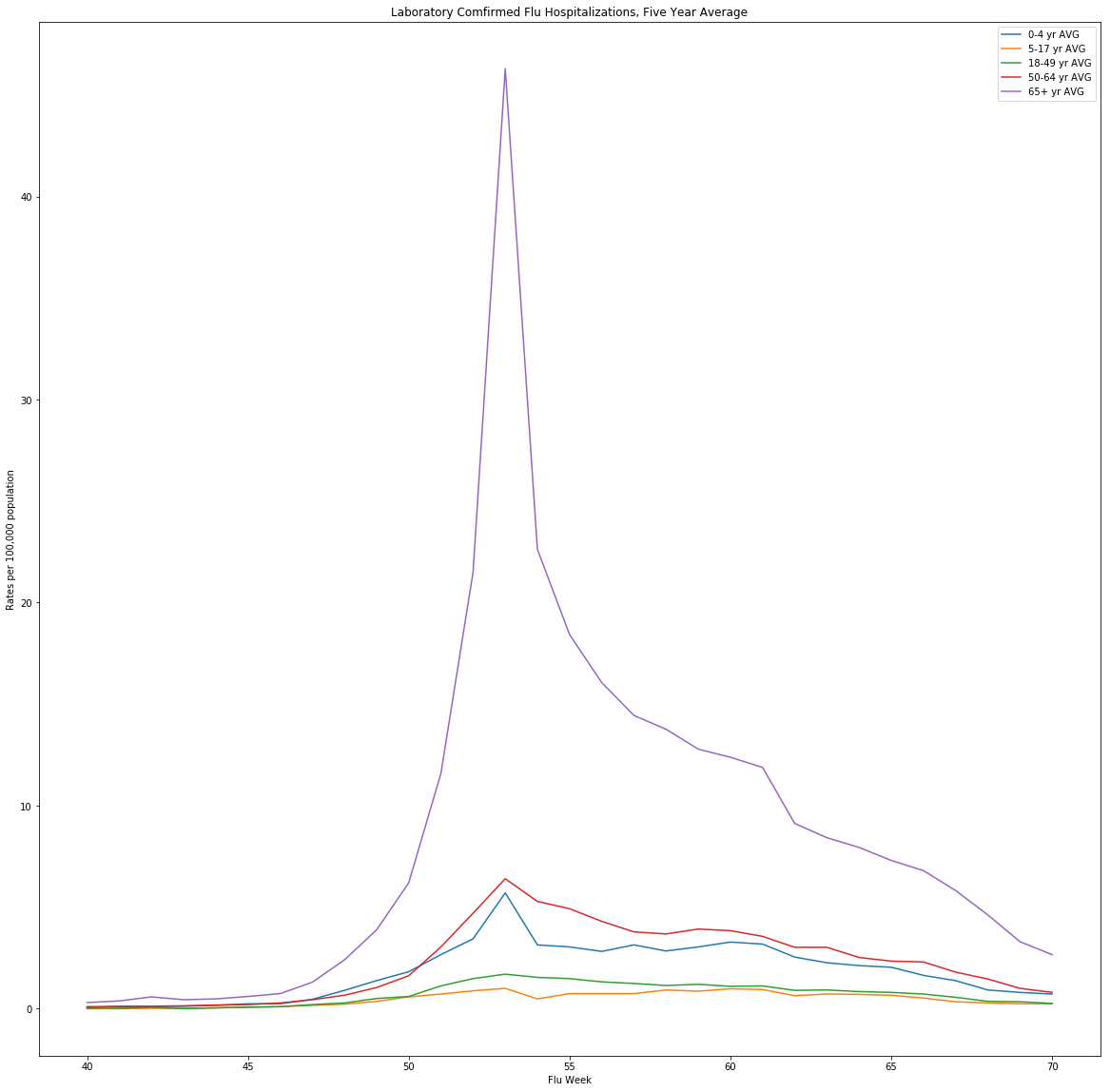
The raw hospitalization data from the CDC has the following columns:



The data was then cleaned by adjusting the “MMWR-WEEK” column. The natural flu season falls between October of the first year to May the following year. Therefore, the data had to be adjusted accordingly since week 1 falls in the middle of the season. From there, the data was grouped by age, week, and year and manipulated to generate the line chart below.



A similar method was used to generate a 5-year average. However, additional steps were taken to a average the “WEEKLY RATE” at each day of each season.



**Findings:**

Despite popular beliefs, our findings in addition to previous age correlations found suggest that individuals that fall in the 65+ are not more susceptible to contracting the illness. However, interestingly, they are significantly more likely to be hospitalized if contracted. In the 2017-2018 flu season, the CDC estimated that 3,984,513 individuals between 0-4 showed symptoms of the illness, 7,512,601 individuals between 5-17 years old showed symptoms of the illness, 14,428,065 individuals between 18-49 years old showed symptoms of the illness, 15,588,035 individuals between 50-64 years old showed symptoms of the illness, and 7,309,120 individuals 65+ contracted the illness. Therefore, the 65+ age bracket is not necessarily more susceptible to the illness. Yet, they do result with larger repercussions if contracted.

The large purple spike on each graph supports these finding, with the 65+ age bracket as the clear outlier.

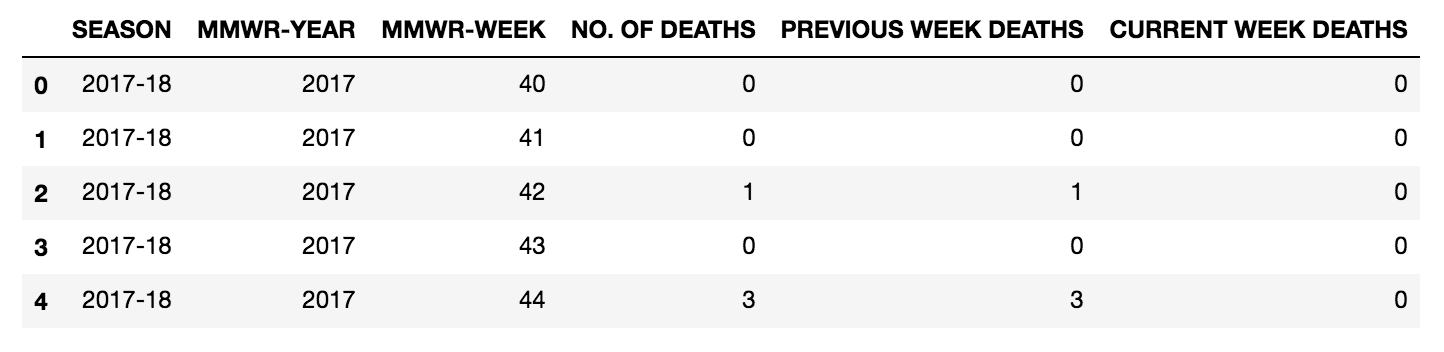
**Question: How many of the flu hospitalization incidences lead to mortality?**

**Data Used:**

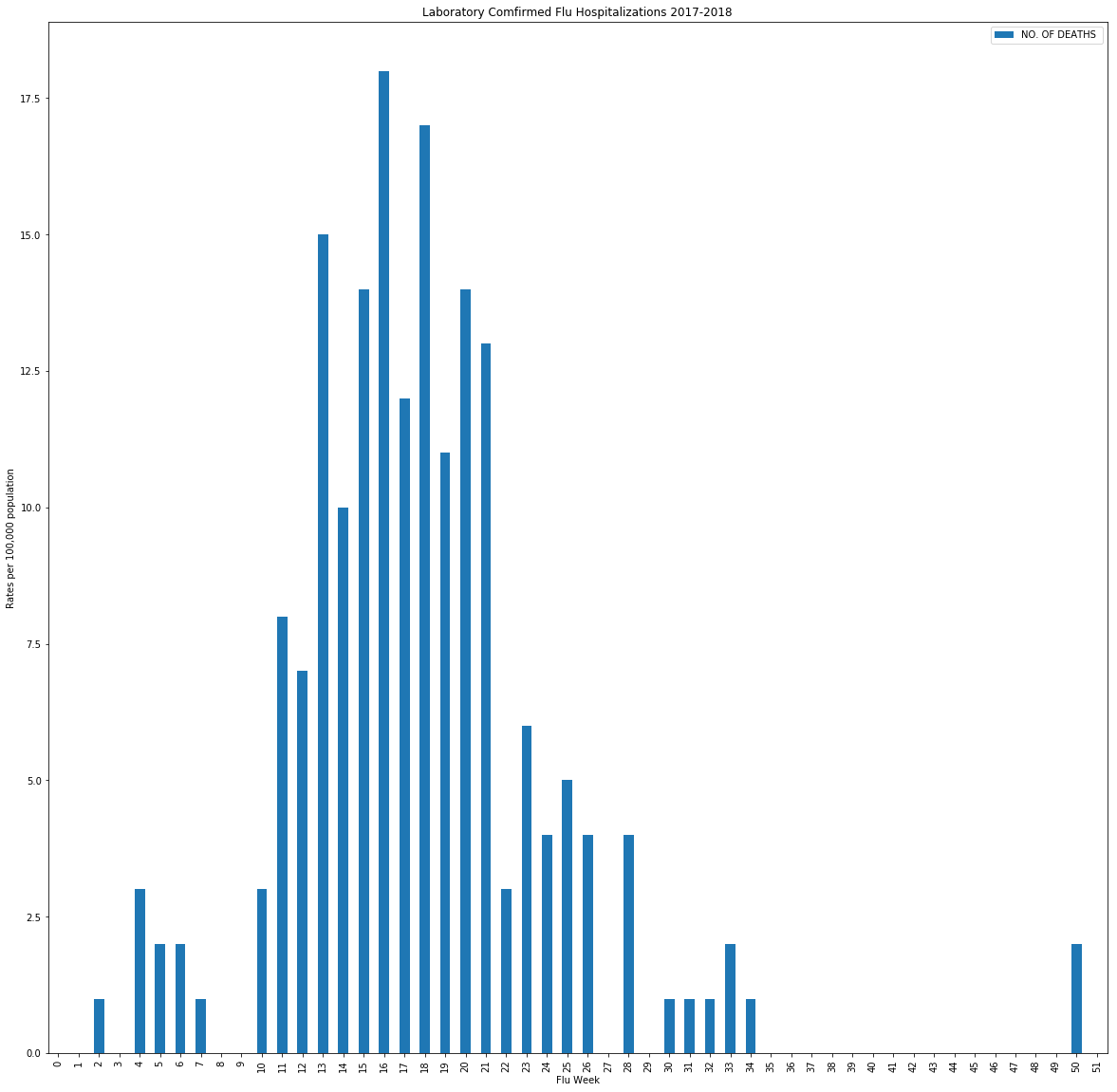
* CDC mortality counts for the 2017-2018 season

**Cleaning the Data:**

The raw mortality data from the CDC has the following columns:



The mortality CDC data was manipulated to only include “MMWR-WEEK” and “NO. OF DEATHS” to graph a bar chart of total count of deaths by week. Again, it is important to remember that the data provided only represents a subset of the total US population and was used to estimate the severity of illness across all 50 states.



**Findings:**

Again, this chart helps visualize the spikes of the severity of the illness broken down my week. The week where most individuals passed due to the illness was during week 1, which correlates well with the trend lines on page 3 of this report.

**Conclusion:**

Although correlation and causation should always be viewed separately, we were able to see correlations between Census data and past flu seasons. Gaps in the CDC’s information resulted in year-over-year estimates based on small sample size. Therefore it is clear to see why other companies like Google have attempted to make their own flu season predictions in the past.

Other factors such as vaccine and the yearly flu strains also play a large role in determining the severity of the season. For the purpose of this research however, we felt it was important to focus on and analyze the largest publicly available data sets. Nevertheless, if given more time we agree that it would also be interesting to further research additional factors in order to gain a complete picture of why the flu season varies from year to year.