Final Draft Submission

Mckenzie Hebert, Kassia Crouse, Jamie Yu

November 2025

Abstract

Influenza poses a significant annual risk to global populations, particularly to children aged two to five, among whom hospitalization rates can reach up to 60 percent. Given the recurrent burden of influenza outbreaks on healthcare systems, this project aims to explore how influenza forecasting can be used to predict and potentially reduce hospital strain. We developed and evaluated time series and machine learning models, including SARIMA, auto-ARIMA, and random forest methods, to forecast Influenza-Like Illness (ILI) percentages at national and state levels. Our initial SARIMA model on the state level data achieved a root mean squared error (RMSE) of 1.348 percent, indicating reasonable performance but not fully capturing the actual seasonal peaks. The SARIMA method on the national data performed worse, with an RMSE of 4.093 percent. In contrast, the optimized random forest model demonstrated strong predictive accuracy with an RMSE of 0.1599 percent, successfully capturing seasonal trends. We further incorporated the MAPIE library to generate 95 percent prediction intervals, providing confidence estimates for our forecasts. These models can inform hospitals to allocate resources and staff in advance of anticipated surges, reducing patient overflow and staff burnout. Future work includes adapting threshold values for state-specific forecasts and exploring dynamic parameters that account for seasonal and regional variability.

1 Introduction and Background

1.1 Introduction

The influenza virus is one of the most prevalent seasonal illnesses across the globe, with estimates by the WHO marking 3 to 5 million cases of severe illness and between 290,000 and 650,000 respiratory deaths per year[19]. "Flu season" is a period of the year, generally from October to May, where influenza cases and hospitalizations rapidly rise, increasing the global burden of the disease. With every year, this virus mutates creating new strains that are not protected from previous vaccines. This leads to more people being infected with the virus and putting pressure on hospitals and their workers both physically and mentally.

One of the qualities that makes them so difficult to fight is that there are generally no "cures" for viruses; the only way to recover is for the body to fight it off. In addition, viruses are highly transmissible and mutate frequently, resulting in new strains that bodies are not prepared to fight off. Particularly, the influenza virus has plagued the world for centuries, with more than five major influenza pandemics in the last 150 years. The most recent influenza pandemic was the 2009 H1N1 swine flu pandemic. This influenza pandemic was so severe that the NIH actually estimates there were about 41 million cases of H1N1 during this pandemic [18].

This "flu season" typically runs for about six months and can start in October and end in May, with every year this virus mutates creating new strains that previous vaccines don't protect against. This leads to more people being infected with the virus, increasing the number of cases and hospitalizations. This influx of cases puts pressure on hospitals and their workers both physically and mentally. This leads to one of the reasons our project is important: if we can forecast influenza outbreaks, we can warn hospitals so they can allocate equipment and staff to prepare for the influx of patients and be in a better position to effectively help them.

While influenza forecasting can be useful, it can also be very difficult due to numerous factors like location, age, lack of data, and the season. Influenza transmission also adds another layer of uncertainty for both influenza forecasting and for hospitals. Influenza as it stands is quite difficult to pin down, with many variables at play. Despite those variables being identifiable, such as droplet particles and physical contact, the exact efficiencies and specifics of these methods still remain unclear [10]. For instance, how do air currents or buildings play into how an aerosol droplet travels? Hand washing proves effective in countering contact spread, but what is the threshold for optimal halting of transmission? How much does this threshold change in different sized communities? Although we can't directly quantify and reduce transmission, with an effective model, we can help hospitals anticipate the potential volume of patients. So with our analysis, we intend to provide hospitals and other medical centers the information they need to reduce uncertainty about potential outbreaks and reduce the strain on hospitals.

1.2 Background

The idea of forecasting the next year's flu outbreak is not novel and has been done many times by numerous studies and government organizations. The data used for forecasting is called ILI or influenza-like illness which is defined as: a temperature equal to or greater to 100 degrees Fahrenheit with a cough or sore throat, and no other known cause. Approximately 3,000 providers contribute to the CDC's U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) with the number of patients seen and how many exhibited influenza-like illness [13]. This ILINet data is then used for one of the most comprehensive resources for this topic: the FluView interactive webpage from the United States Center for Disease Control and Prevention [7]. This website has a wealth of information on the subject, including a weekly influenza surveillance report, a report of cases per age group, a map displaying areas with the most new cases, and many more figures. It also includes data on hospitalization rates that can vary from strain to strain. This is a great starting point for many studies that are looking to predict the timing and magnitude of future flu outbreaks.

As with previous research efforts, we plan to implement this dataset as well. The dataset's varied tools and modules can prove valuable, as it separates by age range, which can assist with determining differences in age groups as we intend to do. The data can also be grouped by viral type, which will allow us to extrapolate some conclusion to help medical personnel fine-tune their response to the outbreak's particular strand.

While there are numerous flu forecasting models, there is a general focus on forecasting at the national and state level. It is uncommon to see a model that accounts for differences in preparedness at the rural and urban level, which is important as there are notable differences in their priorities. Looking at a study conducted in New York, in which surveys were sent out to hospitals asking them to evaluate six aspects of disaster preparedness: disaster plan development, on-site surge capacity, available materials and resources, disaster education and training, disaster preparedness funding levels, and perception of disaster preparedness [22]. The study found there to be differences in all but disaster preparedness funding levels, which is significant as it highlights the different priorities that rural and urban hospitals have. But these differences are often left unaccounted for, and this is likely due to the measured data. Many prevalent models utilize the CDC data which is not as effective for forecasting at a finer geographic scale [11]. For example, looking at the ILINet State Indicator Map, it states that it does not, "measure the extent of geographic spread of flu within a state" and "may not accurately depict the full picture of influenza activity for the whole state" [5].

While influenza forecasting models have been getting increasingly accurate, the question arises: how do these predictions turn into actionable recommendations for hospitals? This question is difficult to answer. There are many factors at play when attempting to predict influenza, many of which are extremely variable so it becomes more difficult to set strict cutoffs. There is not a singular measure that determines when hospitals should be alerted to potential outbreaks. However, we are considering using the CDC flu severity metrics, which range from low severity to very high severity, assessed across three age groups using intensity thresholds. The metrics used are: the percentage of visits due to ILI, rate of flu-related hospitalizations, and percentage of deaths due to flu [2].

Another challenge or unknown regarding influenza transmission is the differences in various age ranges. Children for instance are more susceptible to being affected by the virus, due to their immune systems naturally being weaker and less "experienced" than a typical adult's. One reason for this is due to the lack

of sufficient memory cells, which are crucial in defending against repeat cases of an infection [**?**]. This means that patterns that emerge among adult infections may not apply to pediatric cases of influenza, meaning that they have to be categorized separately. To provide an example of such differences, hospitals should ideally be aware of different critical symptoms for children and adults. In adults, a fever is of little worry unless that fever clears, then returns later. In children, any fever regardless of consistency above 104 degrees Fahrenheit should be treated as crucial. This is even more prevalent in children younger than 12 weeks, where any fever at all requires immediate attention [9].

Another challenge to influenza forecasting is the data itself and its limitations. Most models rely on the same ILINet influenza data collected by the CDC, and while it is extremely useful in tracking visits and influenza-related illness, there are still some potential limitations, particularly in an area we are focusing on. Because the CDC collects its data from healthcare centers across the country, it is likely that some areas, specifically more rural populations are underrepresented within the data set, which could lead to less accurate predictions in these regions. Additionally, ILINet lacks more specific information on factors such as vaccination rates, which could potentially be a significant variable for a model. To potentially address this, using a vaccination dataset like the FluVaxView could be beneficial in decreasing some uncertainty in predictions [4].

Taking into account these challenges, we plan to incorporate multiple data sources and modeling approaches in order to provide accurate forecasts while minimizing some of the limitations within the available data.

1.3 Methodologies in Literature

There are many ways of approaching disease forecasting. It is often difficult to choose one because all have distinct advantages and disadvantages. A recent strategy that has begun to emerge is using some combination of complementary methods in order to help account for model weaknesses. But focusing on individual methods that have been used in the past, one of the earliest and most common ways is to forecast similar to the methods in meteorology and weather. This method is based on historical outbreak and pandemic patterns and matching current numbers to predict what will happen [21]. This technique is often referred to as the "Method of Analogs" and is common in less recent studies. Some advantages of this method are that it can sometimes outperform traditional time series methods as long as it is nonparametric. Some limitations of this method are that it is fairly sensitive and it can sometimes be difficult to find patterns in history that repeat and are aligned [12].

After the method of analogs, another approach to predicting flu outbreaks that became more prevalent was using a time series model. A time series model is a statistical method that extracts trends and information from some data over time and aims to apply it to the future with little to no error. One of the most common time series analyses used in disease prediction is ARIMA (auto-regressive integrated moving average model) [15]. The advantages of using a time series model to predict disease is that ARIMA can capture relationships with lag as well as generally capture seasonality of the outbreak. Some disadvantages are that flu outbreaks can have very different timing from year to year, and that they aren't good at predicting any off-season pandemics [12]. Alongside ARIMA, more recent models have begun to pair it with different machine learning methods to analyze these time series. Some methods include random forest, a support vector machine learning process, and creation of an artificial neural network. In the past, it has been identified that random forest demonstrated the highest accuracy between RF, SVM, and ANN analysis [17]. However, as random forest and ARIMA operate on different principles, we may find that one method proves more effective than the other in our analysis.

In more recent years, a common statistical method for influenza forecasting is using Bayesian models, specifically hierarchical Bayesian models and/or dynamic Bayesian models. Focusing on the general method for Bayesian models: a prior distribution is specified by using factors such as historical influenza data and other knowledge about its transmission, impact, etc. Then as new data is observed, the likelihood of various model parameters are evaluated. Finally, we combine both of those using Bayes' Formula to find the posterior distribution, which can then be used to predict outbreaks [20]. Bayesian models can be particularly useful when data is more limited, as its use of prior knowledge can supplement the lack of observations, so this

could be potentially useful for rural areas where there may not be as many observed cases of influenza. But a potential drawback is that Bayesian models can be computationally expensive, especially with larger datasets.

1.4 Project Plan

This project aims to develop predictive models that forecast influenza outbreaks to help hospitals anticipate and reduce strain during flu season. Influenza causes millions of severe illnesses and hundreds of thousands of deaths globally each year, with young children and rural populations particularly vulnerable. By forecasting the timing and magnitude of influenza outbreaks, hospitals can better allocate beds, staff, and resources to manage patient surges and prevent healthcare worker burnout. The central research question guiding this project is: How can influenza forecasts be used to predict and potentially reduce overall hospital strain? Sub-questions focus on the predictive accuracy of epidemiological trends, the unique needs of pediatric populations, and the varying impacts of outbreaks on rural hospitals.

To address these aims, the team will analyze national and regional influenza data from the CDC's FluView, FluVaxView, and ILINet datasets, along with local surveillance data from California and its various different regions. Exploratory Data Analysis (EDA) will be performed using R, Python, and Tableau to assess data quality, visualize spatial and temporal patterns, and identify influential variables. For modeling, a combination of statistical and machine learning methods—such as ARIMA, Random Forest, and potentially Bayesian ensemble approaches—will be employed to capture both national and regional influenza trends. The goal is to produce actionable, high-accuracy forecasts with uncertainty estimates that health systems can use to make proactive operational decisions, especially for high-risk and resource-limited settings.

2 Methods

In this section, we will go over the analysis process mentioned above. We will go more in-depth into how R, Python, and Tableau were used to develop visualizations and achieve the results (displayed in the next section). Python was used for a majority of the project, with R and Tableau used to generate only a handful of visualizations that were more easily done in that respective program compared to Python.

2.1 Methods: Datasets

We use two main datasets for the modeling and data analysis: CDC FluView ILI Dataset [8] and California ILI Dataset [6]. ILI data is a measure of the percentage of cases that contributing health providers see that exhibit influenza-like illness symptoms. We downloaded the CDC FluView ILI dataset twice, once with the region set to National and another time with the region set to State. The CDC State Region dataset groups the data by state, making it very easy to compare data features by individual states, and how they compare to each other. This was the dataset used for spatial analyses. The National ILI region dataset groups all the data by week, regardless of the state it came from and also includes an age breakdown on cases. This makes it easy to compare week-by-week analysis, which is how we performed our temporal analyses. We used the State region dataset more for the initial data analysis, and the National region ILI dataset was the one used for the actual modeling.

The State ILI is from the California Department of Public Health, but is similar in structure and content to the National ILI dataset with State regions from the CDC. One difference is that the primary variable is total ILI, so it's a numerical count rather than a percentage. Our primary focus is on predicting National influenza rates, with a secondary focus on seeing how well our models perform on the state level.

While we are primarily using the National ILI dataset for modeling, it does not provide as much insight into the specific characteristics of those who contracted influenza beyond age groups. So for the purpose of the EDA, we're going to use the the FluSurv Characteristics Datasets [3]. This dataset comes from laboratory confirmed cases rather than ILI symptoms reported by providers. So while useful for data analysis and exploration, laboratory cases tend to lag behind influenza trends due to the time required for testing and

then confirmation, making them not as ideal for weekly forecasting. The dataset is organized such that each row is a unique combination of the season, characteristic, age category, a group (within the characteristic), and then the corresponding percentage of hospitalizations. Our primary focus will be on the characteristics which are: Sex, Race, ICU Admission, Mechanical Ventilation, Pneumonia Diagnosis, Deaths, Antiviral Treatments, and Age Group. The dataset was loaded in, skipping the first row as that was the title of the entire data table, and then typos in the column names were corrected. The dataset had quite a few missing percentages and NA values, so those rows were dropped. Since each row represents a unique week, manually inputting values could potentially be misleading because it would also use information from rows with different combinations of variables and influenza trends vary across seasons and years. The first five entries are shown below.

Looking at the first five entries of each dataset:

```
ILINet.head()
```

	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
0	National	X	2010	40	1.10939	1.13505	2627	1677	X	3142	627	400	8473	1838	746485
1	National	X	2010	41	1.24341	1.25256	2953	1779	X	3522	649	444	9347	1875	746230
2	National	X	2010	42	1.25726	1.24570	3044	1898	X	3641	690	411	9684	1907	777397
3	National	X	2010	43	1.25734	1.26774	3226	1754	X	3822	682	420	9904	1929	781234
4	National	X	2010	44	1.43414	1.43723	3451	1981	X	4397	736	455	11020	1947	766753

Figure 1: The head of the National ILI dataset (National Region).

```
ILINetByState.head()
```

	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
0	States	Alabama	2010	40	X	2.13477	X	X	X	X	X	X	249	35	11664
1	States	Alaska	2010	40	X	0.875146	X	X	X	X	X	X	15	7	1714
2	States	Arizona	2010	40	X	0.674721	X	X	X	X	X	X	172	49	25492
3	States	Arkansas	2010	40	X	0.696056	X	X	X	X	X	X	18	15	2586
4	States	California	2010	40	X	1.95412	X	X	X	X	X	X	632	112	32342

Figure 2: The head of the State ILI dataset.

	SEASON	CHARACTERTISTIC	AGE_CATEGORY	GROUP	PERCENT
0	2010-11	Sex	Ages 0-17	Male	55.1
1	2010-11	Sex	Ages 0-17	Female	44.9
2	2010-11	Race	Ages 0-17	White	39.3
3	2010-11	Race	Ages 0-17	Black	26.7
4	2010-11	Race	Ages 0-17	Hispanic/Latino	25.0

Figure 3: The head of the National FluSurv Characteristics dataset.

2.2 Exploratory Data Analysis

As typical with any data analysis, we start with exploratory data analysis. We looked at the aforementioned National ILI data, as well as national data from the CDC's FluSurv Characteristics dataset. We additionally analyzed the California ILI dataset. We constructed a variety of different visualizations in order to find any trends in ILI cases.

2.3 Methods: Correlation Analysis

Before going into the various analyses and forms of machine learning, it is a good idea to perform some manual data analysis prior, to get an idea of any patterns that may emerge. Looking at the National ILI data, we will be primarily examining the "%UNWEIGHTED ILI" variable, which measures the percentage of doctor visits that exhibited ILI (Influenza-like illness) symptoms, a heatmap was generated to study the size of the percentages week by week using R's ggplot2 package [23]. Similarly, we did that for the California ILI data as well, examining the total ILI measure.

2.4 Methods: Inferential Statistics

We performed some hypothesis testing on the National ILI and California ILI datasets. We wanted to determine whether or not the mean number of ILI cases was equivalent across age groups nationally and across the different regions in California. To examine this, we employed one-way ANOVA testing, as it determines whether there are any statistically significant differences between the means of two or more groups.

2.5 Methods: SARIMA

Time-series data is defined as a sequence of points collected over time either regularly or randomly. Disease tracking data and especially influenza outbreak data fall into this category. There are a multitude of time series models that have been specially developed to handle this datatype. The two we will focus on are ARIMA and SARIMA. ARIMA stands for Autoregressive Integrated Moving Average and SARIMA stands for Seasonal Autoregressive Integrated Moving Average. Breaking this down into parts starts with the "AR" section of the abbreviation. Autoregressive essentially means that the model uses past values to predict future values. The "I" part of the abbreviation stands for Integrated, which means that the model removes trends in the data and makes the data stationary. Lastly, the "MA" part of the abbreviation or moving average means that the models uses past forecasting errors to correct the model and improve future forecasts [**?**]. For a SARIMA model, the extra parameters tell the model not only to look at trends from previous weeks, but also to focus on trends that repeat or happen often, such as a yearly influenza outbreak.

2.6 Methods: Random Forest

A random forest regressor method involves the usage of several decision trees, which each data point is fed into. At the end of each decision tree, a determination is made as to what the appropriate output value for the given input is. The average of each tree's determinations is what the final prediction for a given input is. This process repeats for all data points, taking in all inputs associated with that data point.

2.7 Methods: Ensemble

Ensemble learning is a machine learning technique that, rather than using only one technique such as SARIMA, uses multiple at once. Multiple techniques are employed to achieve better results than any one technique could achieve alone. Ensemble methods can be achieved though methods of bagging, boosting, stacking, or (weighted) averages just to name a few. We plan to test an ensemble method of averages to see if we could create a more effective model leveraging data from two models.

3 Results

3.1 Results: Exploratory Data Analysis

3.1.1 National ILI Data

The data we extrapolated was grouped to only view one type of data, resulting in some columns being filled with Xs, which represented NA values. These columns had to be filtered out before any actual analysis could be done.

As discussed for one of the sub-questions we aim to answer, we wanted to determine the impact of influenza among different age groups, particularly pediatric patients. However, for completeness, a full age breakdown was constructed, which is displayed below. The ranges are unevenly split, with one being only 0-4, and another being much larger at 5-24.

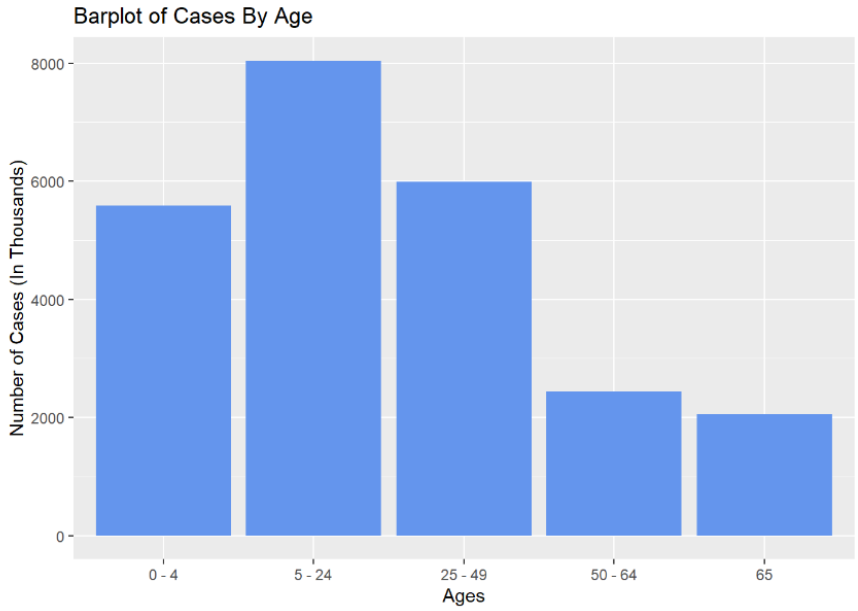


Figure 4: A barplot showing the number of ILI (Influenza-like illness) cases by age group.

We also split the data above into separate graphs by year to examine any trends that may have occurred.

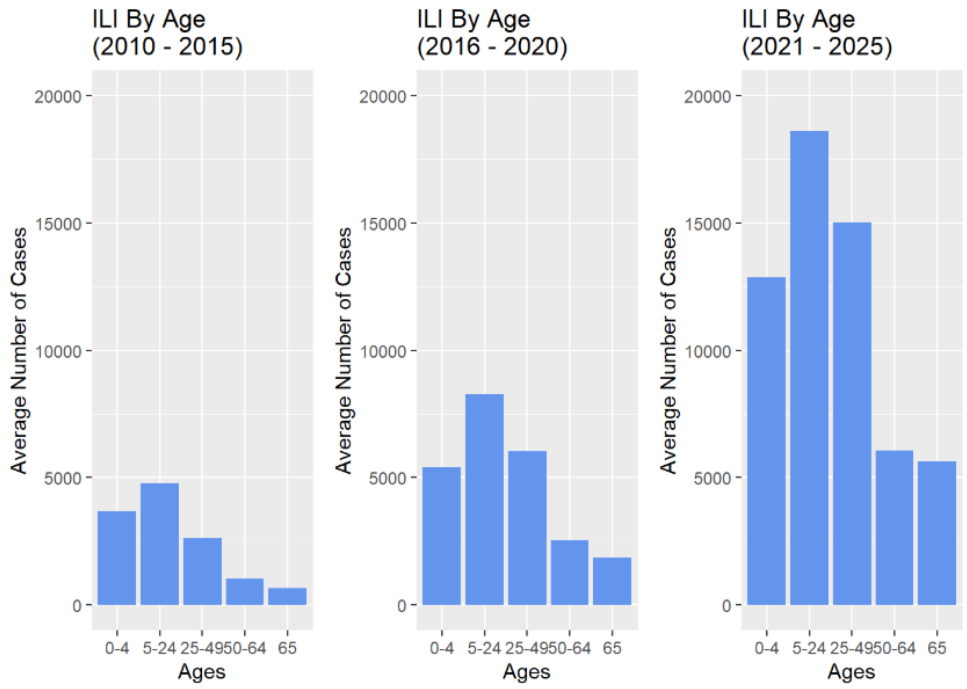


Figure 5: A barplot showing the number of ILI (Influenza-like illness) cases for 2021-2025, split between three 5-year sections.

To test a potential reason the number of influenza cases were increasing, we looked at the difference in the number of providers contributing over the years.

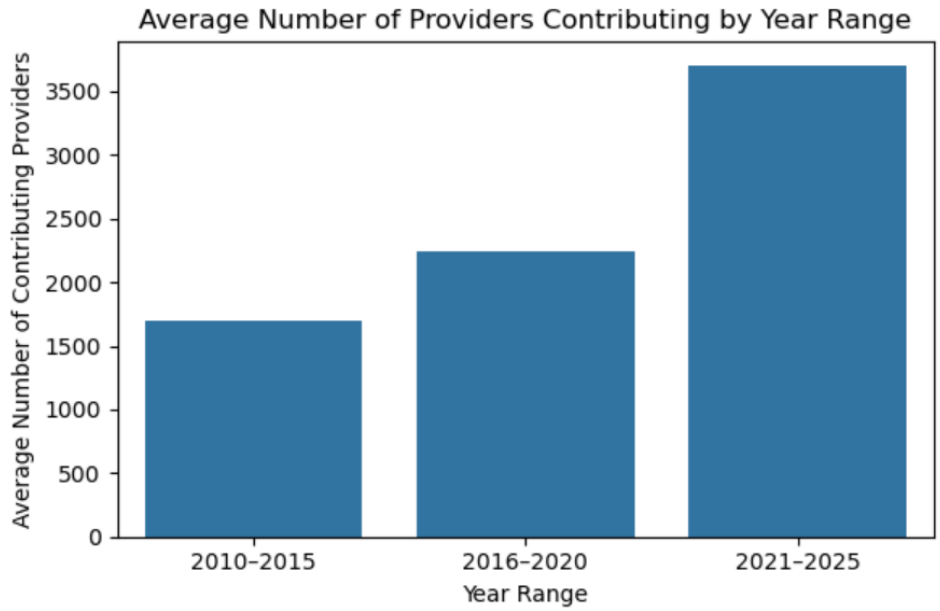


Figure 6: A barplot showing the average number of contributing providers by year.

3.1.2 National FluSurv Data

For this dataset, we primarily focused on the different characteristics and trends across age to see if there were certain characteristics that led to more hospitalizations. The figure below shows the sex distribution across different age groups from the most recent five seasons.

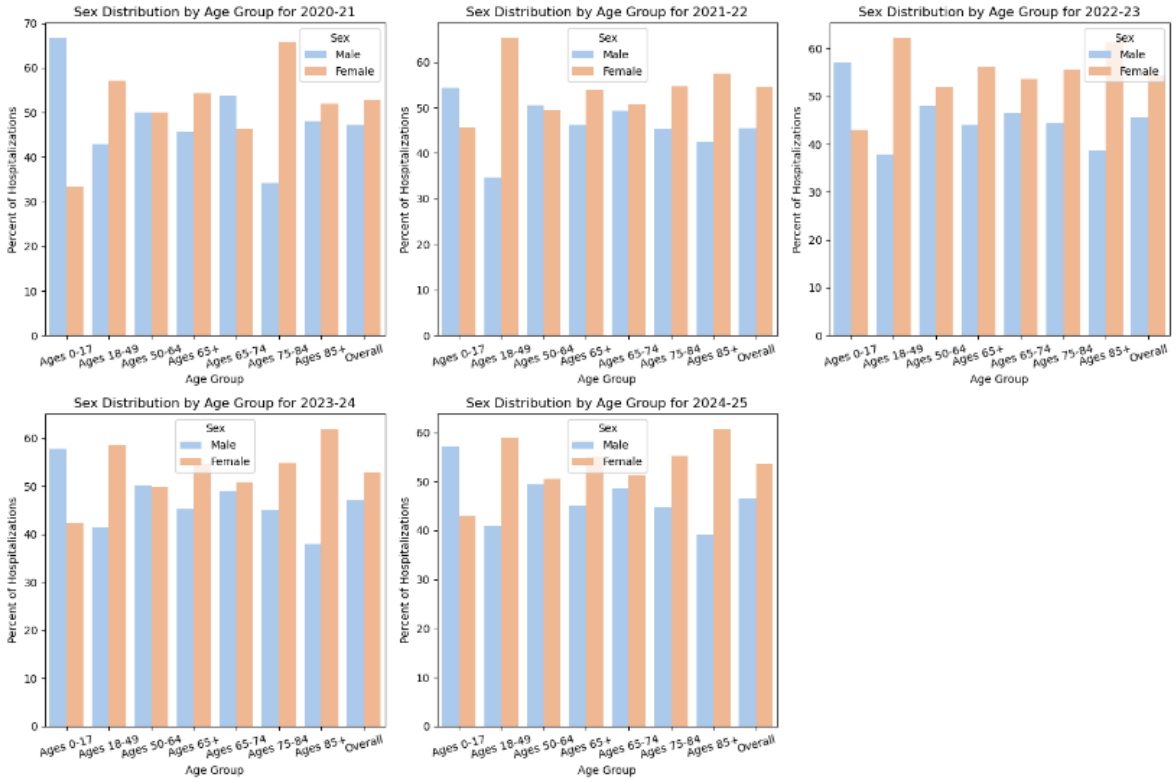


Figure 7: Bar graphs comparing the sex distribution across different age groups from 2020-2025.

Taking a look at some of the other characteristics with interesting trends, since distributions were similar across the past five seasons, only the graph from the most recent season (2024-2025) was included.

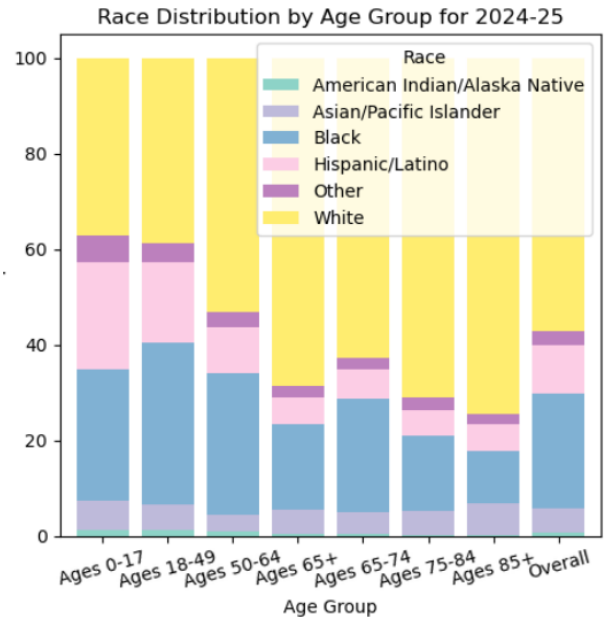
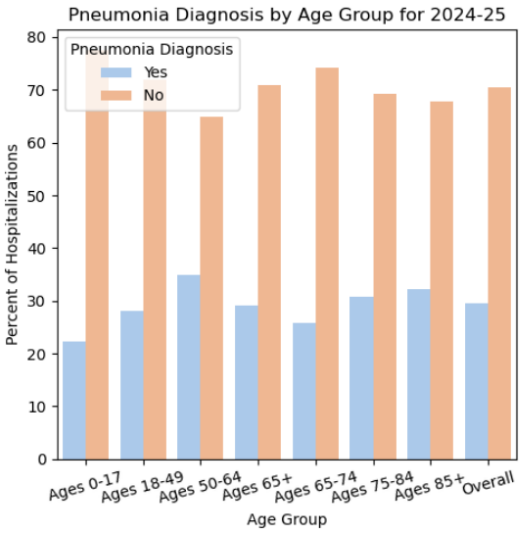
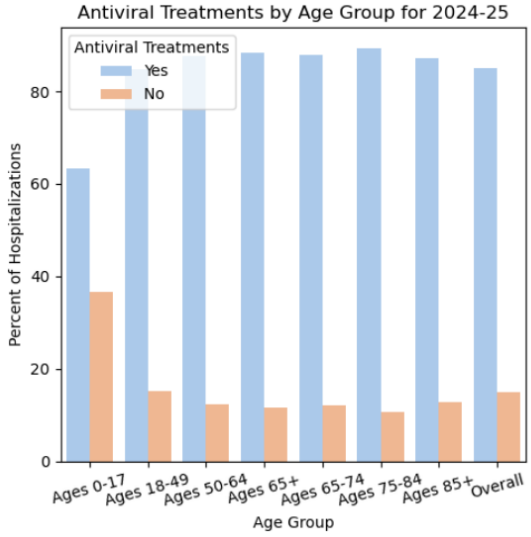


Figure 8: Stacked bar plot of the race distribution by age group for influenza hospitalizations in 2024-2025.



(a) Hospitalizations by pneumonia diagnosis across age groups for 2024-2025.



(b) Hospitalizations by antiviral treatments across age groups for 2024-2025.

Figure 9: Distribution of characteristics across age groups for 2024-2025.

3.1.3 California ILI Data

In this data set, the state of California is divided into 5 regions: Bay Area, Central, Lower Southern, Upper Southern, and Northern. Additionally, California is included as a region, containing the overall data. This regional data can help to provide a closer look at the influenza trends within the state.

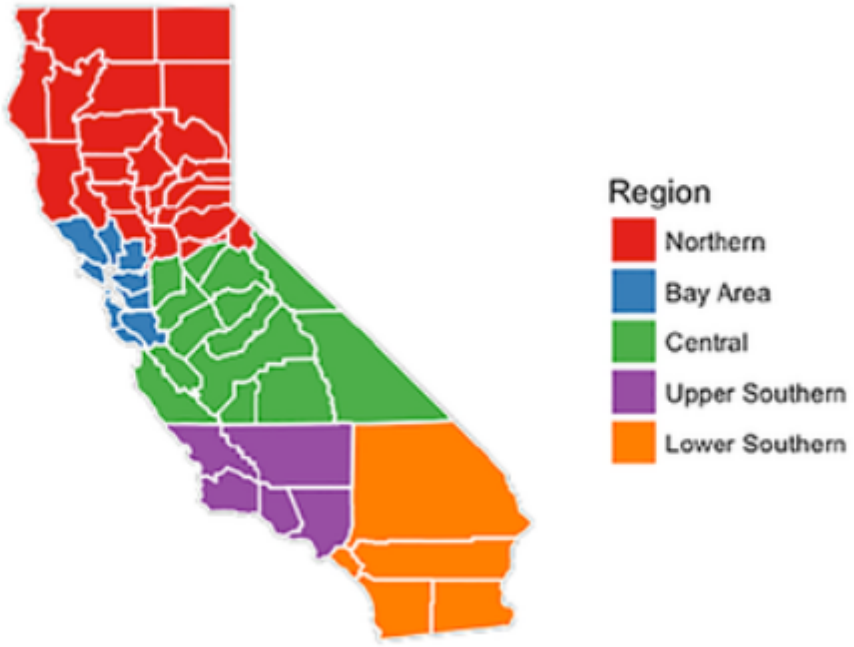


Figure 10: The regions of the state of California

Although there is no explicit information from the California Department of Public Health on how the regions are divided, this is a general map of the regions of California and what we believe to be the approximate regions referenced in the dataset [14]. Below is an image of the population density in regions across California based on the 2010 census [1].

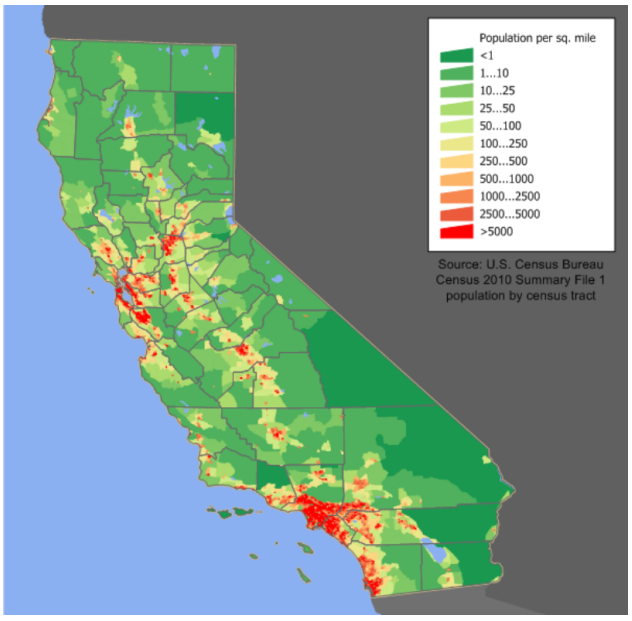


Figure 11: A map of California with population density across regions.

Next, we look at the percentage of ILI visits compared to total visits throughout the years.

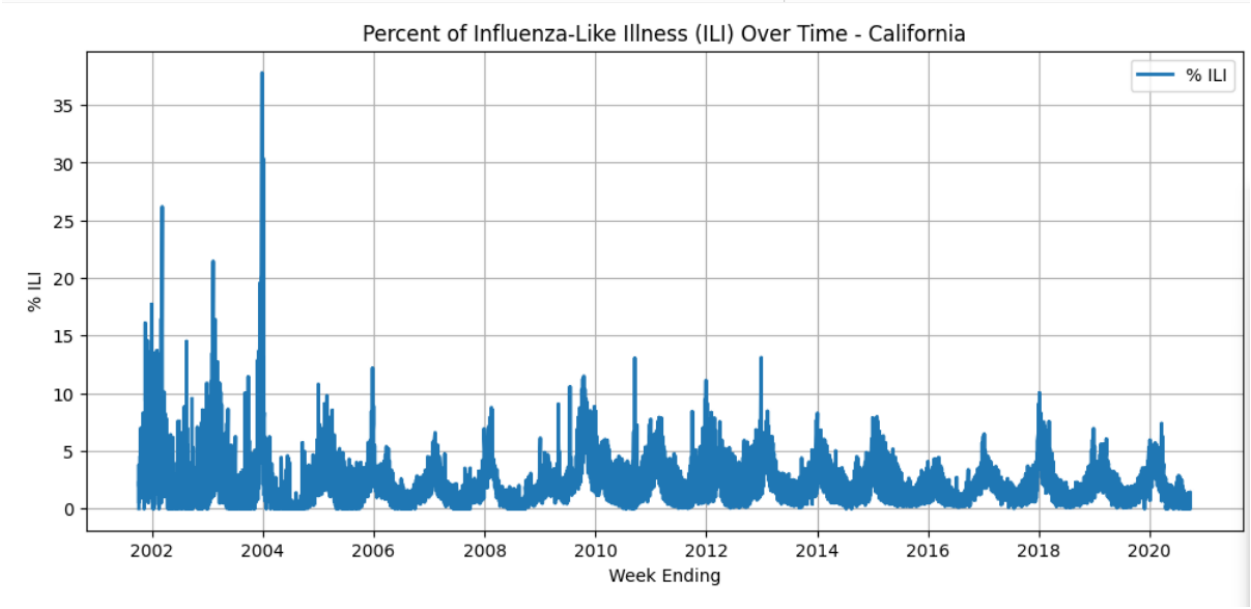


Figure 12: Trend of Percent ILI (the percent of how many Influenza-Like-Illnesses are seen per day) from 2001-2020.

Next, we want to know which months are the prominent months for flu season in California. The literature suggests that the flu season can start as early as October and run as late as May [19], so we want to see for California which months are the most prominent months for flu season.

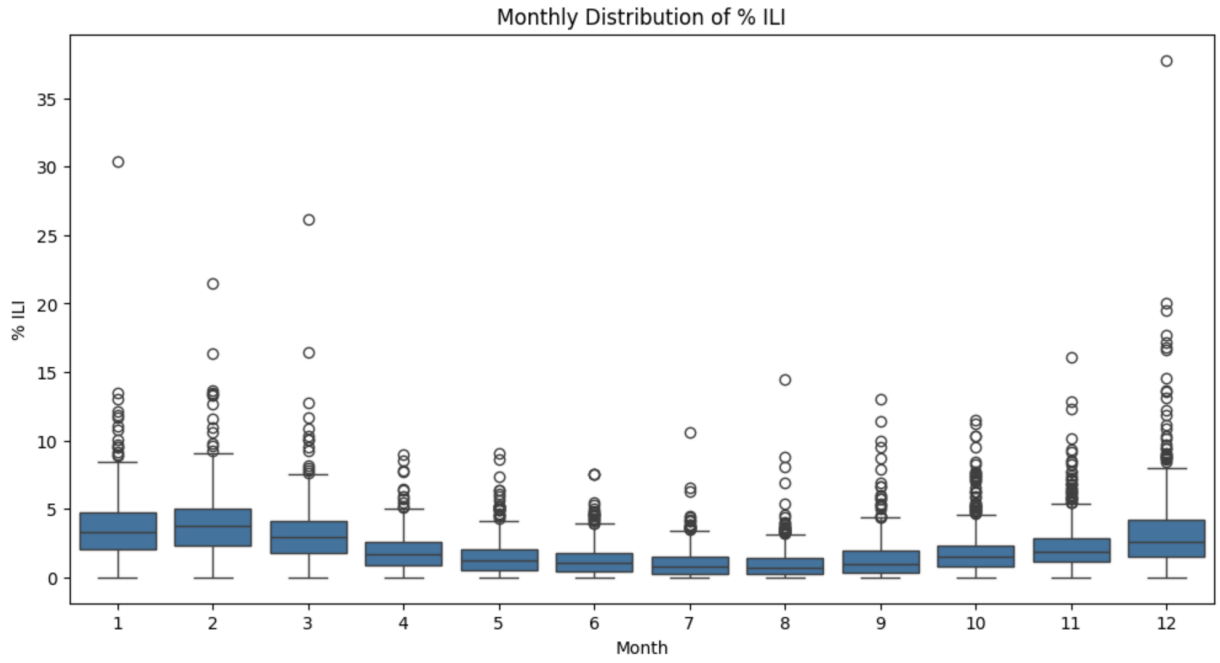


Figure 13: A box plot displaying the distribution of percent of ILI per month for years 2001-2020

The next thing we want to look at is which regions in California have the most cases of ILI and how we can adjust recommendations to different areas of the state to be more urban or rural. In our dataset California is split into 6 different regions so we want to determine where the outbreaks are the largest to better help the California hospital system prepare.

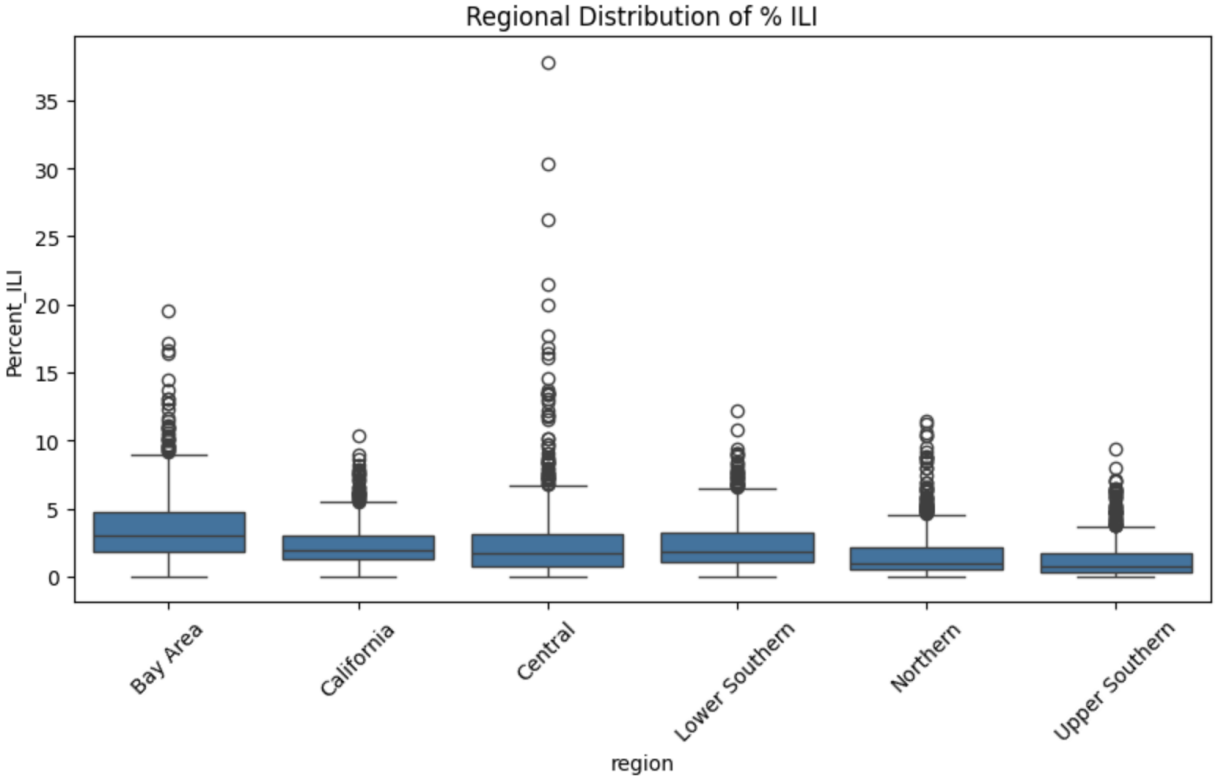


Figure 14: A box plot displaying the distribution of percent of ILI per region of California.

3.2 Results: Correlation Analysis

Below is a heatmap, constructed using R using the National ILI dataset mentioned earlier. Each small rectangular "block" represents the average "% UNWEIGHTED ILI" per week across 15 years.

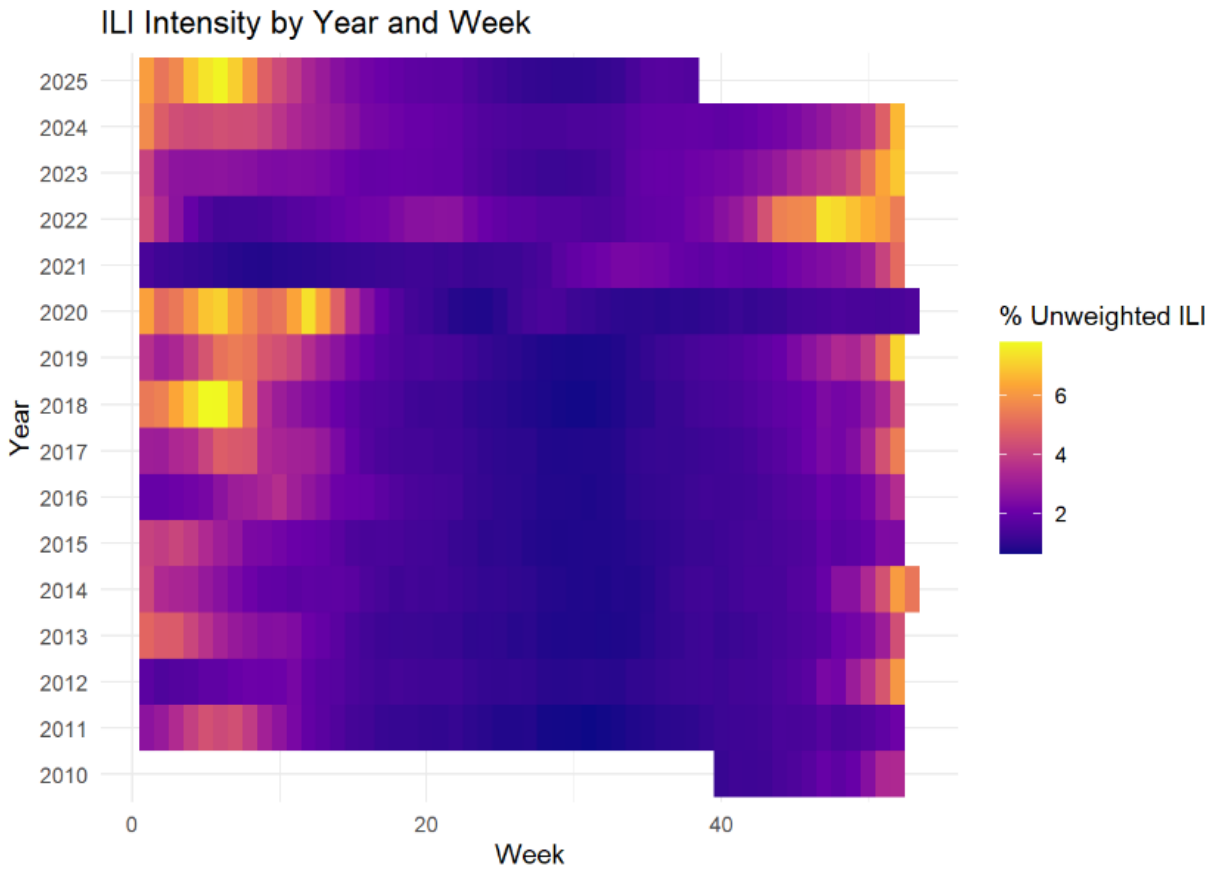


Figure 15: A heatmap depicting the value of the "%UNWEIGHTED ILI" variable, split by week.

The "% UNWEIGHTED ILI" variable measures the average percentage of medical visits are related to influenza-related illness. A hospital, doctor's office, or other facility receives several cases each day, with a fraction of them being related to influenza-like illnesses, which consists of illnesses that result in fevers, coughs, chills, and other symptoms commonly associated with influenza.

Below, we created another heat map for the California ILI data.

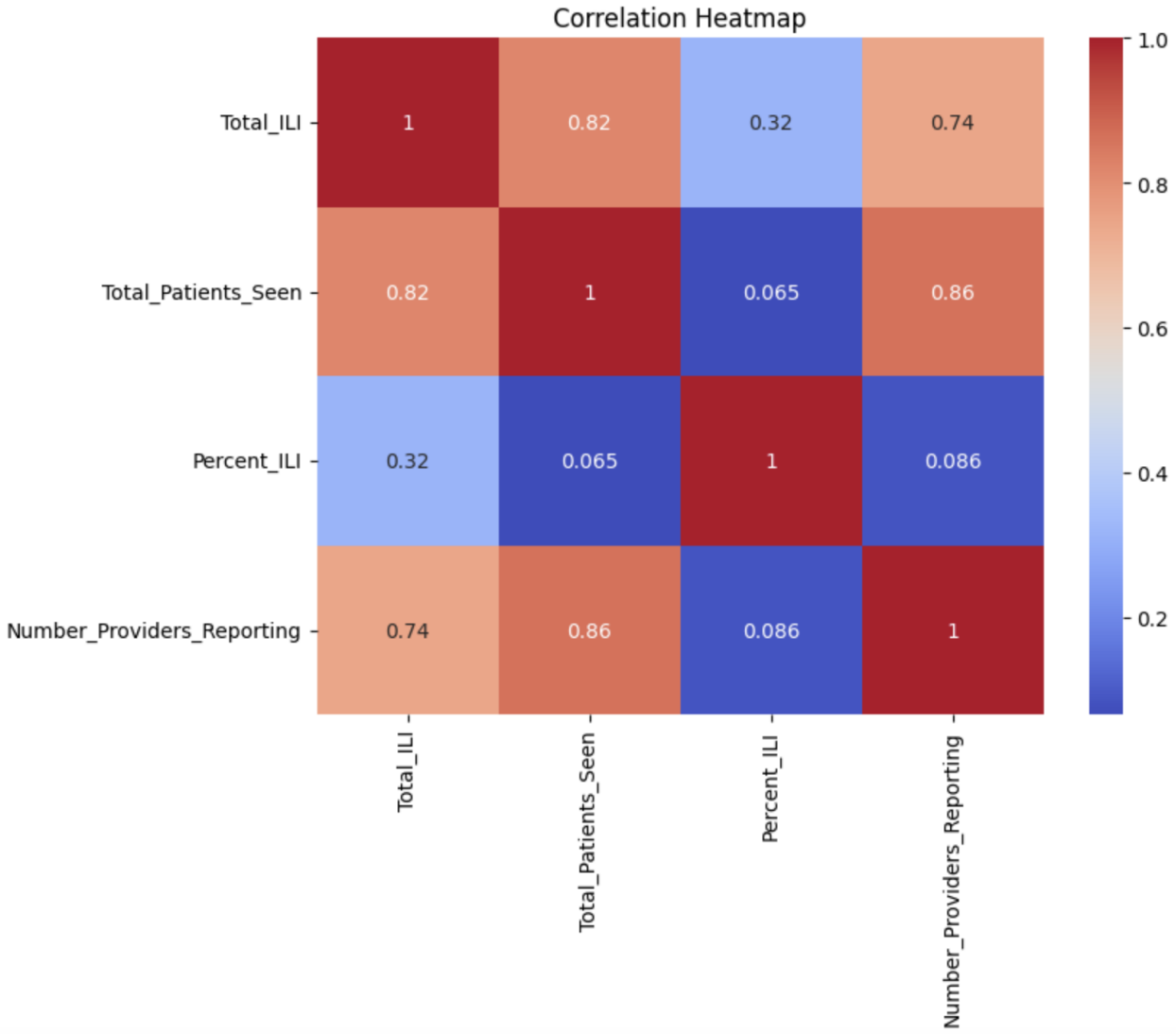


Figure 16: The heat map correlation matrix between the California ILI variables.

3.3 Results: Inferential Statistics

Since one of our sub-questions focuses on the differences between different age groups, pediatric patients in particular, we wanted to see if there was an actual significant differences in the mean case counts between the different age groups. We did an ANOVA test assuming the age group counts were independent, the variances were homogeneous, and since the number of cases was greater than 30 for each group, we assumed they were approximately normally distributed by the Central Limit Theorem.

We set alpha = 0.05 and conducted the test under these hypotheses:
H_0 = the mean number of ILI cases is equal for all of the age groups
H_a = at least one of the ILI mean cases is different
For the results: p-value = 0.0000 and F-value = 151.1172.

Since we also wanted to examine regional differences, we decided to look into the regions within the Cal-

ifornia ILI dataset. We conducted another ANOVA test under the same assumptions to see if the differences between regions were significant with alpha = 0.05 and hypotheses:

H_0 = the mean number of ILI cases is equal for regions in California
H_a = at least one of the regions has a different mean number of ILI cases
For the results: p-value = 0.0000 and F-value = 554.3224.

Again we see that the p-value is effectively zero, less than alpha, so we reject the null hypothesis and can conclude that at least one of the regions has a different mean number of ILI cases. While the regions do have different populations, it represents why there is need to consider whether the region, like whether it is more urban or rural. For example, one hundred cases would be more concerning for a region that is more rural and typically sees less ILI cases compared to an urban region with high population that sees a much greater number of ILI cases on average.

3.4 Results: SARIMA

Before we could start working on the model itself, the data had to be transformed into time series data. Before we log transform the data and train the model, here is our seasonality data.

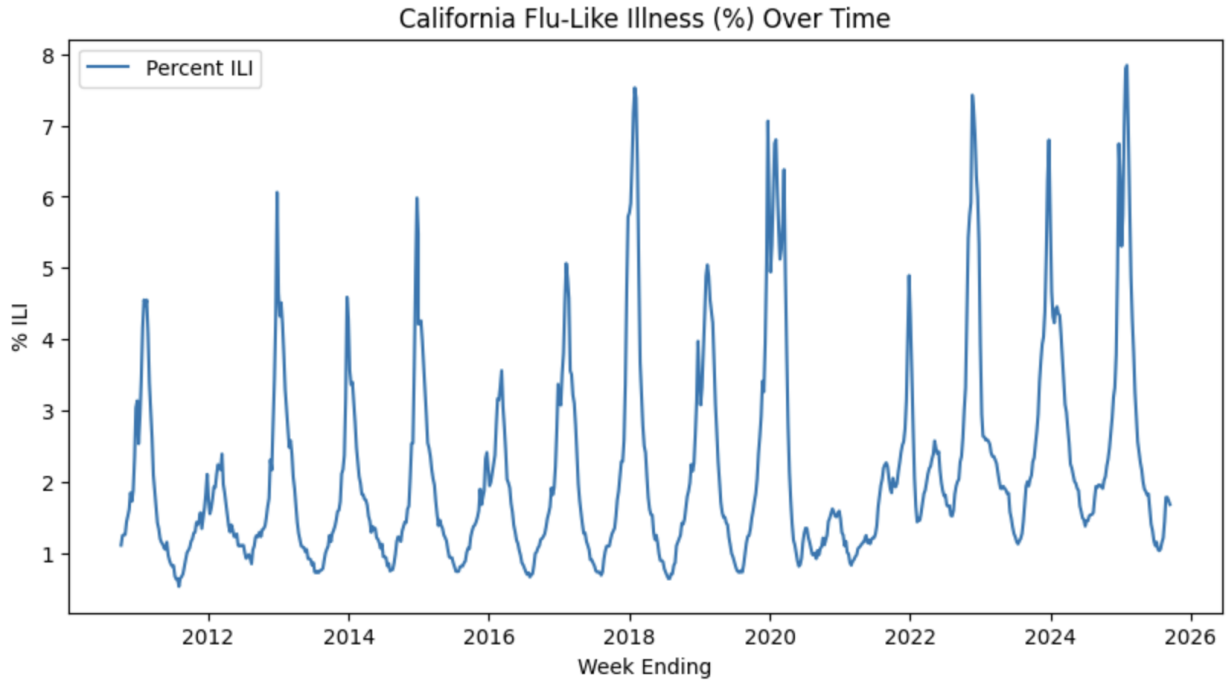


Figure 17: The seasonality of Weighted percent ILI cases for the national data.

As we can see, there are clearly yearly cycles of times when influenza rates rise, demonstrating what we refer to as the flu season. As we can see, the National ILI data runs from 2010 to present, giving us a good picture of the most recent 15 years of ILI cases. Next, it was time to prepare our model for training. We chose a 80/20 train test split with the data to reserve some of it for later testing the model. Next, we had to log transform the data to reduce the skew of the data and make them more normally distributed and ready for the model to be applied to it. After this, it was finally time to train the SARIMA model. Our data appeared to be stationary already, and so we chose 2 auto-regressive coefficients, 1 differencing, and 2 moving average, as well as a season period of 52 because there are 52 weeks in the year, and our data is

weekly. This is just a basic set of parameters and the SARIMAX results for our model are shown below.

SARIMAX Results

Dep. Variable:	% WEIGHTED ILI	**No. Observations:**	781
Model:	SARIMAX(2, 1, 2)x(1, 1, [1], 52)	**Log Likelihood**	1052.043
Date:	Sun, 09 Nov 2025	**AIC**	-2090.086
Time:	16:55:37	**BIC**	-2057.954
Sample:	0	**HQIC**	-2077.688
	-781		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8164	1.527	0.535	0.593	-2.176	3.809
ar.L2	-0.1637	0.896	-0.183	0.855	-1.919	1.592
ma.L1	-0.4015	1.524	-0.264	0.792	-3.388	2.585
ma.L2	0.0841	0.258	0.326	0.744	-0.421	0.589
ar.S.L52	0.1707	0.058	2.952	0.003	0.057	0.284
ma.S.L52	-0.6589	0.047	-14.021	0	-0.751	-0.567
sigma2	0.0032	0	26.306	0	0.003	0.003

Ljung-Box (L1) (Q):	0	**Jarque-Bera (JB):**	137.47
Prob(Q):	0.95	**Prob(JB):**	0
Heteroskedasticity (H):	0.83	**Skew:**	0.14
Prob(H) (two-sided):	0.16	**Kurtosis:**	5.11

Figure 18: The Sarimax output displaying the results of our SARIMA model.

After this, we viewed some of the residual plots; these are plots that can give us further insight into what the model missed. We view four standard residual plots for this [**?**]. One that checks the white noise, one that sees if the model is independent, one that checks the constant variance, and one that sees if it is normally distributed. Here are our 4 residual plots for our model below.

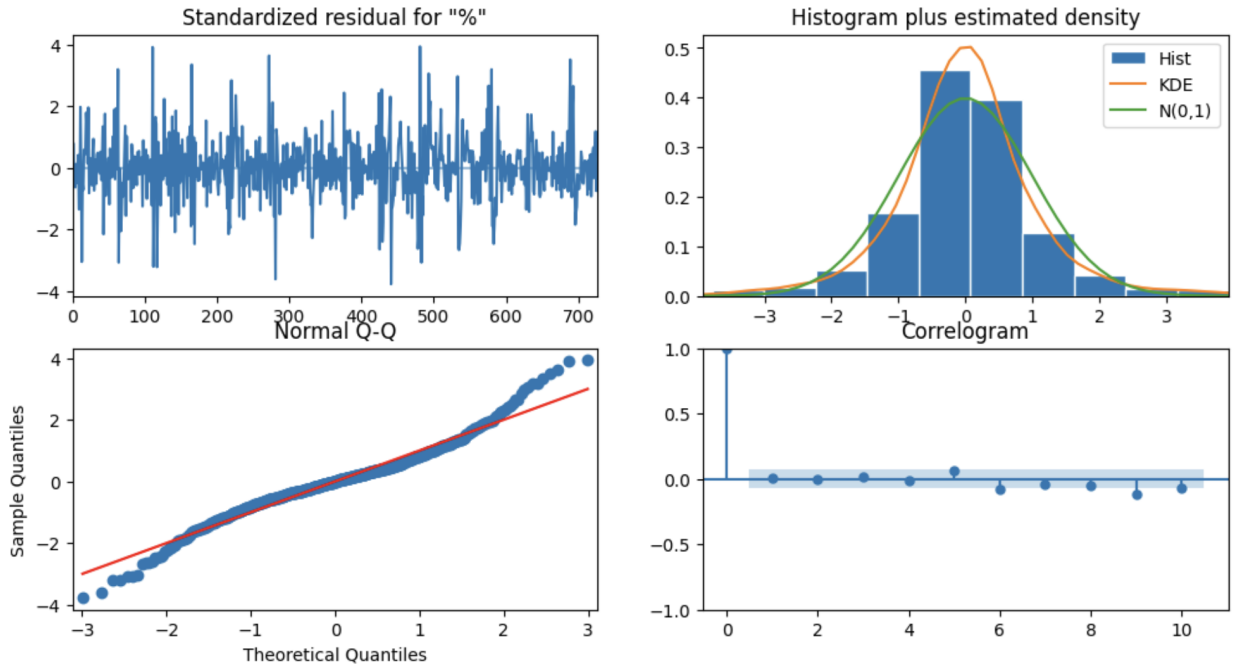


Figure 19: The four residual plots to give further insight into the SARIMA model.

Lastly, we used our model to forecast the percent ILI for the test data. In the plot below, our forecast is in orange, with the shaded region representing the 95 percent confidence interval. The plot is just displaying the most recent 5 years for better visibility of the trends and forecasting.

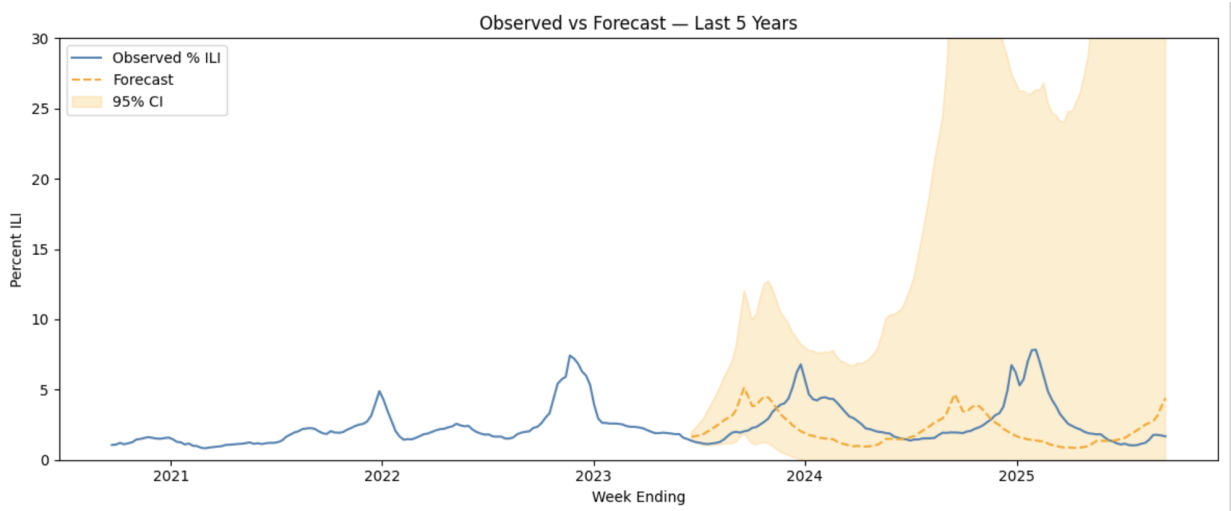


Figure 20: The SARIMA model forecasting overlaid on the test data.

As you can see in the plot above, our model actually does a fairly good job at capturing the seasonality of an annual flu season spike. However, it is clear that our model is predicting these spikes to be earlier than they are actually occurring.

A good measure of the model is the RMSE (root mean squared error). This is a measure of how far our

model's estimate is from the actual observed percentages [**?**]. Our model's RSME is approximately 1.348 meaning that our model predicts within 1.348 percent of the actual value of percent ILI cases. This is not as ideal as the percent of ILI cases generally hover around 1 to 10 percent, so being around 1.348 percent off is a fairly significant margin of error. We believe that this value is this high because the model predicts the seasonality at the wrong time. If the predictions were aligned with the correct seasonality the RMSE would be lower as the actual magnitude of our predictions is close to the actual values.

So we decided to try the auto-ARIMA method in order to find the optimal values for the SARIMA model. We ended up limiting the maximum values to 1 or 2 to see if a decent model could be found without having to run through all the combinations of all the possible values, which would take a very long time. But the auto-ARIMA actually ended up performing worse than our original SARIMA model, with a RMSE of 4.093%. Due to the long run time, we decided to keep the original SARIMA method rather than running auto-ARIMA with a wider range for the values.

3.5 Results: Random Forest

While the random forest method can simply be called without specifying the parameters, we were aware of optional parameters which can be set, which could improve our model's accuracy. To find the optimal values for those parameters, we conducted a GridSearchCV through Python's sklearn package. Through simply calling the function after feeding it our training inputs and outputs, we found the optimal parameters, as determined by the function to be:

max_depth = 20, min_samples_split = 5, n_estimators = 50

Each decision tree will have a maximum depth of 20 layers. If a node contains 5 or more samples, it has the potential to branch, being split into 2 nodes. The forest contains 50 decision trees.

Below is a figure with the above optimal parameters set:

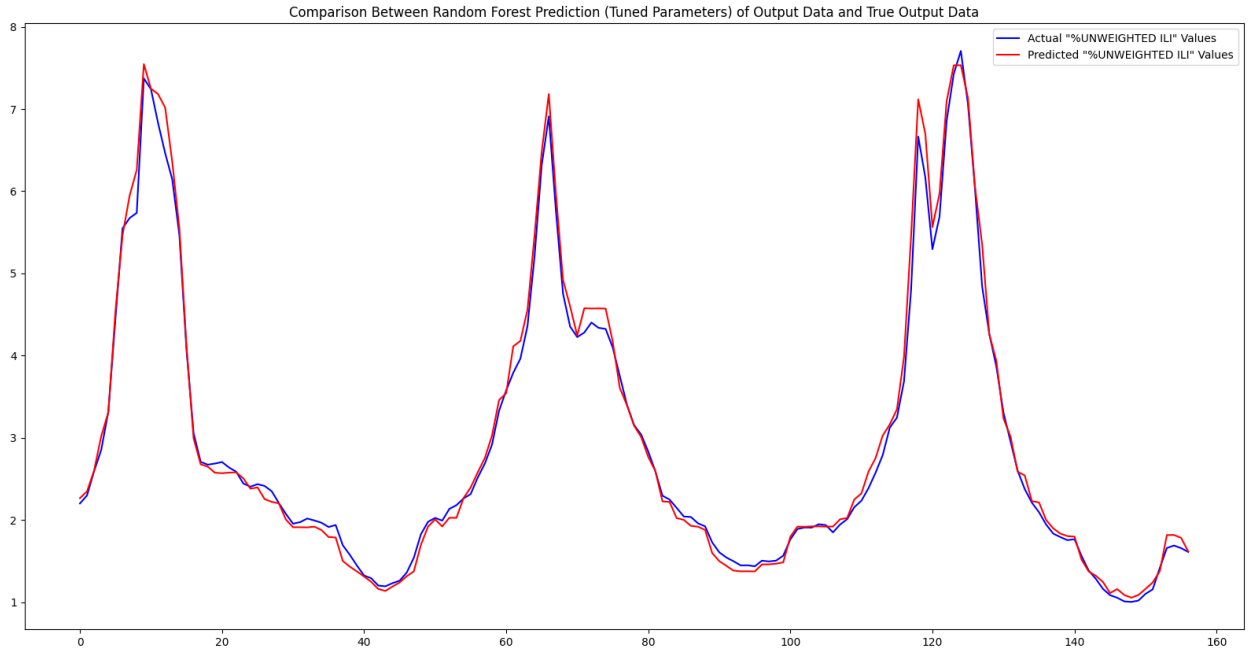


Figure 21: A line plot showing the predicted outputs of our tuned random forest method, and the actual outputs.

For comparison, the following figure has the following default parameters, which was used previously while describing methodology, and is less optimized:

max_depth = None (default), min_samples_split = 2 (default), n_estimators = 100 (default)

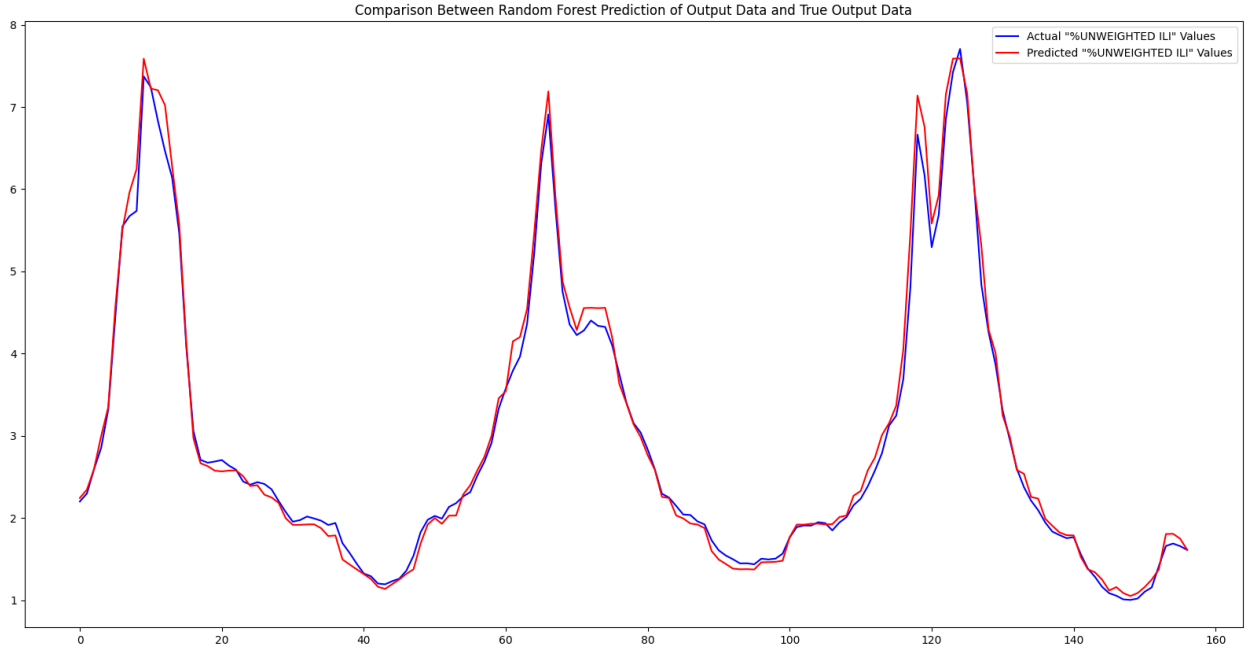


Figure 22: A line plot showing the predicted outputs of our untuned random forest method, and the actual outputs.

These two plots are largely identical, except for a few minor differences. In the tuned model, the upper spikes are slightly lowered. However, upward and downward slopes are slightly lengthened. The valleys remain largely the same, as well as the graph as a whole; any changes that did occur were very slight.

We can see that our random forest model is very accurate in capturing the trends in percentage of ILI cases, which the root mean squared error value of 0.1599 reflects. Even within the context of our data, 0.1599% reflects the model is very accurate, predicting within 0.1599% of the actual national ILI case percentage rate. This is actually more accurate compared to our non-optimized version of our random forest method, which had root mean squared error value of 0.1607%. In other words, our non-optimized version has the chance to be slightly more "off".

But while accurate, we still wanted to provide a prediction interval as point predictions can be more inaccurate. To do this, we used the python Model Agnostic Prediction Interval Estimator (MAPIE) library to wrap around our model and provide a 95% confidence interval. From the MAPIE library, we used the split conformal regressor, which allows us to calculate the prediction intervals for our data [**?**]. We then graphed the predictions with a 95% confidence interval. This means that we are 95 percent confident that the true mean percentage of ILI cases falls within our interval.

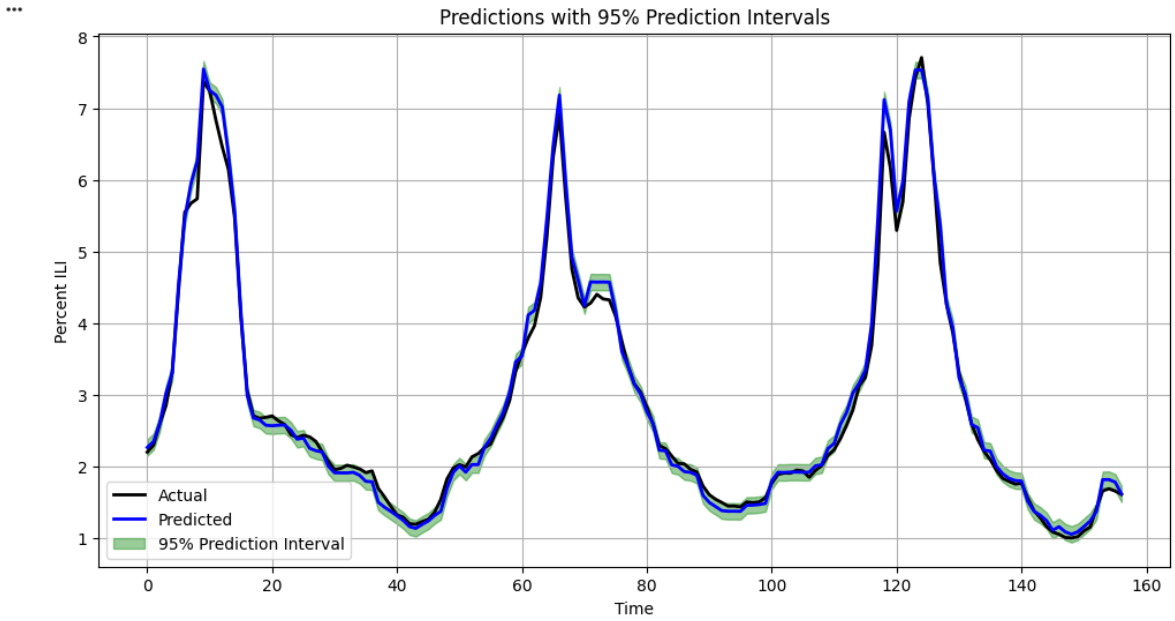


Figure 23: A graph showing the predicted values with the confidence interval against the actual values.

We hope that by providing a confidence interval, hospitals can get a more accurate estimation of the potential influenza case percentages for the following week and prepare adequately.

3.6 Ensemble Method

Previously we had tested an ensemble method using a simple average of the two methods and were planning to test different weighted averages, but since the random forest model accuracy ended up improving with hyperparameter tuning, we decided to just proceed with the random forest model and not pursue the ensemble method, as the SARIMA model has a much higher RMSE compared to the random forest.

3.7 Results: Outbreak Threshold

Since we're mainly predicting the National ILI case percentage, we wanted to find a threshold value that if the percentage of cases exceeds, could alert hospitals around the country of potential outbreaks. We decided to start with the 75th quartile, which we calculated to be 2.314%. Next we flagged prediction values that were over the 75th quartile value and then graphed them below.

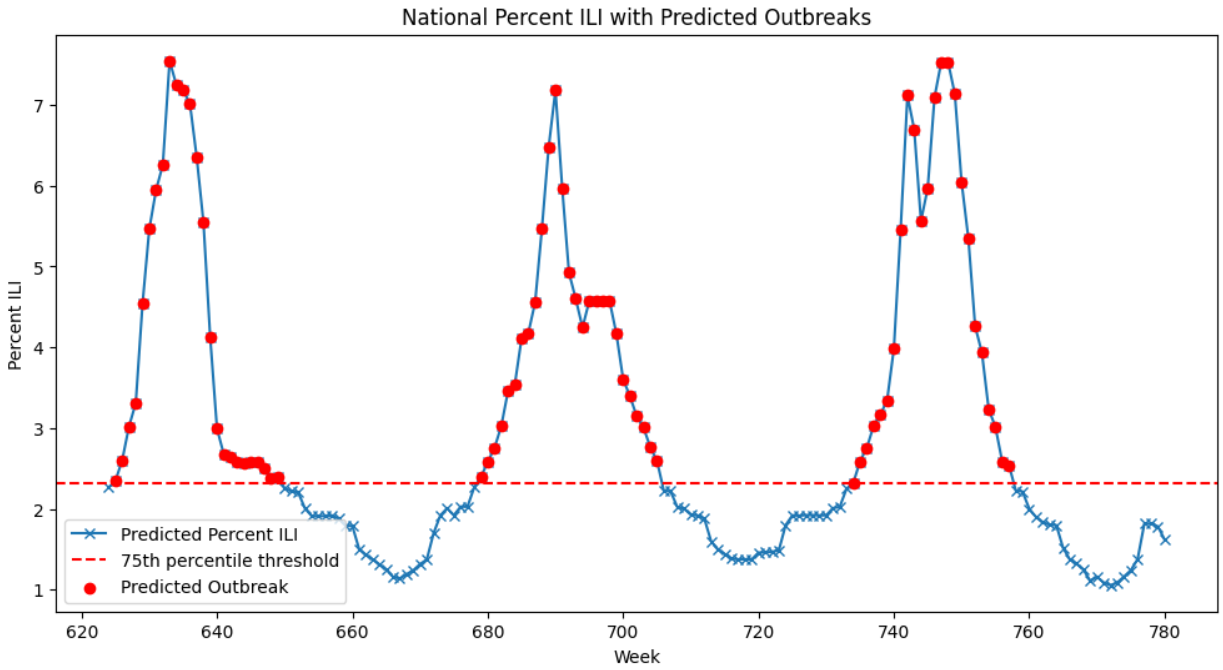


Figure 24: A graph showing the predicted ILI case percentages with outbreak values flagged.

3.8 Testing Model on State Data

Our model was fitted on the National ILI data, so we wanted to see how well our model performed with data on a different geographic scale. So using the California ILI dataset [6], we ran our random forest model. It also did fairly well, with an RMSE of 0.331%, which is slightly worse than the RMSE of the national data, but the random forest model was fitted on the national data so that makes sense. Below is a graph of the random forest model predictions versus the actual testing values.

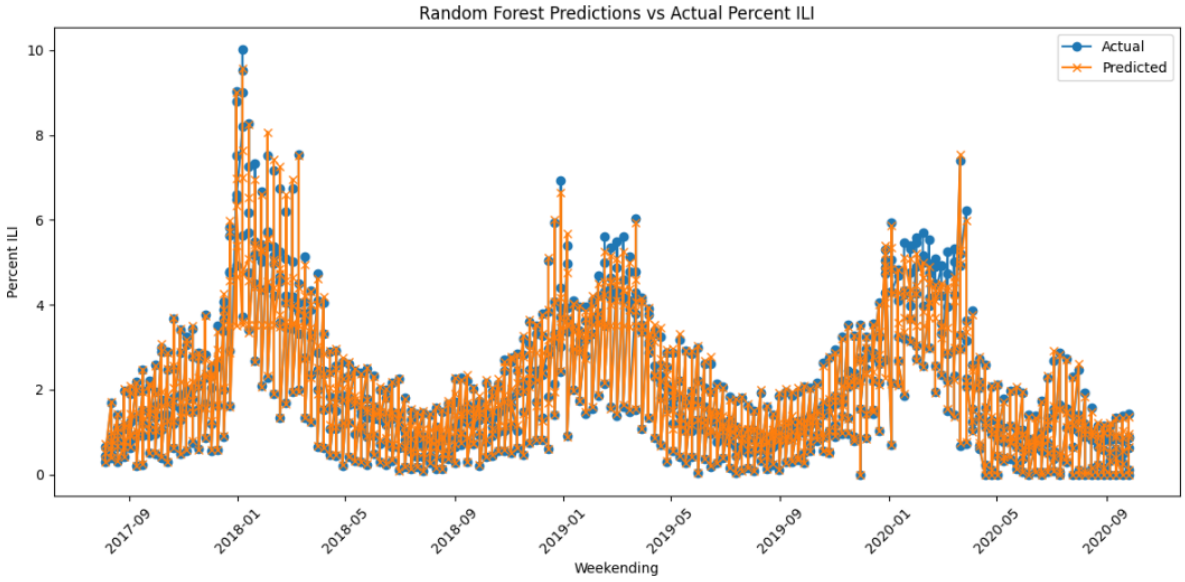


Figure 25: A graph with the random forest predictions compared to the testing values.

We see that the model is still quite accurate for the California state data, which makes sense as we already noticed the California data mirrors the national data trends fairly closely.

3.9 Challenges

One challenge we faced was running searches for optimal values, like the Auto ARIMA and GridSearchCV. We had to find a balance between including a wide enough range of values to try to find the best values and the computation time.

Another challenge was trying to set the threshold value, a singular value suffices for now, but if we wanted to do state by state predictions in the future, the threshold value would probably need to adjust based on the state and other factors.

3.10 Next Steps

Currently we have a model that's effective for predicting the National ILI Percentage and California ILI Percentage, the next step could be to see if the model is effective for predicting ILI case percentages in other states as well. But we have to consider influenza trends may be different so they may require different hyperparameters.

We could also explore different methods for setting the threshold value that takes more variable into account, such as time of the year or compares the predicted value to the previous week's values.

4 Discussion

4.1 Discussion: Exploratory Data Analysis

4.1.1 National ILI Data

As we can see in figure 4, the bins are highly variable. There is a bin that spans only the ages 0-4, whereas the others span a much larger range. However, even in the relatively narrow bin of 0-4, it has the third highest number of cases, close to the number of cases compared to the relatively wide bin of 25-49. As such, the ages of early childhood seem to see more cases relative to the age range compared to other age groups.

As expected, the bins of 5-24 and 25-49 have the greatest number of cases, likely due to the fact that they encompass such wide age ranges. Ideally there would be more evenly-split bins, which would provide a clearer view of the age distribution of ILI cases, but unfortunately the uneven age ranges are a limitation of this dataset.

In figure 5, we noticed a trend: throughout the years, the total number of cases were growing, reaching a very high amount at 2021-2025. This is likely is that the number of providers contributing to the ILI Network increased, so it could've led to an overall increase in cases across all age groups. This is supported by figure, as we see, the average number of providers increased over each range, which could potentially correlate to the increase in cases.

Looking at figure 6, California and Florida seem to have the highest amount of ILI across the 2010 - 2025 period, followed by Virginia, whereas the rest of the state seem to have significantly fewer cases. While it would make sense California and Florida also have large populations, some other high population states such as Texas and New York still don't have as many cases. Additional analysis is being considered to examine this further in the future, to see what could have led to such discrepancies in case averages across states.

4.1.2 National FluSurv Data

We can see that in figure 7 each year over the last five years, females consistently make up a larger percentage of influenza hospitalizations compared to males. By age groups, we see that within the 0-17 range, more males are hospitalized, while for the 18-49 range, it's females by a substantial margin. For the 50-64 group,

it's fairly evenly split. For those 65+, with the exception of the 75-84 age group from 2020-2021, females have higher hospitalization rates.

Moving on, in figure 8, as expected, white individuals made up the largest percentage of hospitalizations overall, reflecting their majority proportion in the US population. However, we do see they make up a smaller percentage of the hospitalizations among the younger age groups compared to the older ones, which may be due to increased racial diversity among the more recent generations.

In figure 9b, overall we see that around 20-30 percent of hospitalizations are also accompanied by a pneumonia diagnosis. So there could be a potential correlation there. Looking at antiviral treatments, we see they are administered majority of the time for hospitalizations, but for those 0-17, the percentage is significantly lower. This could be due to pediatric patients requiring different treatment, which is related to our sub-question on pediatric patients, further demonstrating the need to take into consideration the age of patients.

4.1.3 California ILI Data

As seen in 11, the population density varies greatly throughout the state, highlighting why it may be important to consider whether regions are more rural and urban. While population density isn't the only indication of rural versus urban areas, it can still be fairly informative. We see most regions have areas of lower population and higher population densities. But it appears the Northern region generally has the lowest population density. The regions with the most populous cities: Bay Area, Lower Southern, and Upper Southern all have large populations and a large area of high population density. The Central California region seems to contain both higher and lower population densities. This will be useful context to keep in mind as we look prepare to model the data.

In 12, there are clearly trends at certain times in the year where there are more cases, this represents the "flu season". We see that over time, the percentage of ILI visits has slightly decreased. Also seen above is that every flu season has a different severity. While most seasons seem to be more consistent, some can skyrocket to great magnitudes, which we see occurring in 2004, where the percentage of ILI visits jumps to over 35 percent. This just highlights one of the considerations of our primary question: how flu season puts extra strain on hospitals.

Moving on, in 13, we can see that the months of November to March have much higher averages as well as higher upper outliers than the other months. This confirms what the literature suggests about the month range, but also narrows our scope of months when specifically looking at California.

And lastly, in 14, the Bay Area has the highest average percent of ILI of the regions in California, this makes sense because, as seen in Figure 4 the population density of the Bay Area is high since it contains San Francisco. The Upper Southern region has the lowest, which is surprising given that the population density there is quite high due to Los Angeles. Something else that is interesting to note is that the Central region has the greatest number of outliers, this is surprising because this indicates that from year to year the severity of flu season in the Central region varies widely. This could indicate that the region has both significant rural and urban areas leading to more variation.

4.2 Discussion: Correlation Analysis

4.2.1 National ILI Data

In figure 15, immediately, we can visualize some regular patterns: near the tail end of each year, and the beginning end of a subsequent year, there is a regular spike of percentage of ILI cases in doctor's offices and hospitals. As such, it is not a stretch to assume that is when hospitals may want to strengthen their efforts. However, it is not ideal to rely solely on visual information and heatmaps to make any sort of actionable data or recommendations. As correlation heatmaps only display the strength of linear relationships.

We also notice that there is a noticeable spike of this variable's data within the early quarter of 2020. This is likely due to the COVID-19 pandemic which occurred around the same time. COVID-19, while it being its own separate virus, has many similar symptoms that were likely counted as ILI cases. As such, we

predict that there are COVID-19 cases that are being captured by searches for ILI symptoms, causing the larger continual spike.

4.2.2 California ILI Data

Seen in 16, there is a relatively high correlation between the "number of providers reporting" and the "total patients seen" variables, indicating a strong, positive, linear relationship between them. What this suggests is that as the number of providers reporting increases, the number of patients seen also increases. With a coefficient of 0.86, the R-squared value is 0.74. This means that about 74 percent of the variation in the total patients seen variable can be explained by the number of providers reporting variable. This would make sense because if we have more providers, there would be more patients overall reported in the data set. Something that is important to note is that correlation does not imply causation and just because these variables are correlated does not mean that one causes the other. There could be many confounding variables that have not been accounted for.

4.3 Inferential Statistics

4.3.1 National ILI Data

The p-value is the probability of observing this data assuming that the null hypothesis is true. This value then gets compared to the alpha value, and if it's less it gives us grounds to reject the null hypothesis, but if it's greater, we fail to reject the null hypothesis. In this case, the p-value is effectively zero, which means we reject the null hypothesis, showing that at least one of the mean number of cases for an age group was different than the others.

However it is important to keep in mind that since the age groups were not grouped with equal ranges, they have different populations, so age groups with a greater population would have a greater number of cases. It would be more accurate to compare the mean number of cases per 100,000 individuals in each age group. Unfortunately, since we don't have the age population data for each group during each year, this is not feasible for further analysis.

4.3.2 California ILI Data

Again we see that the p-value is effectively zero, less than alpha, so we reject the null hypothesis and can conclude that at least one of the regions has a different mean number of ILI cases. Although regions do have different populations, it represents why there is a need to consider whether the region is more urban or rural. For example, one hundred cases would be more concerning for a region that is more rural and typically sees less ILI cases compared to an urban region with high population that sees a much greater number of ILI cases on average.

4.4 Discussion: SARIMA

Here are some of the key things to point out about the SARIMAX output [16] in figure 17:

AIC/BIC: For this measure, the lower value is better. The SARIMA displays a value of about -2,000 for both. This is good because it means that the model is a relatively good fit in relation to the noise.

Ljung-Box Q: This displays whether or not the residuals are autocorrelated (we do not want them to be). A value of over 0.05 is better, our model is at 0.95 which is good.

Jarque-Bera: This measure checks to see if the residuals are normally distributed. Again, a measure of over 0.05 is better. Our model is currently sitting at 0 for this measure, which is not great and is indicative of a model that is skewed or heavily tailed.

Heteroskedasticity: This is a measure of whether the variance changes over time. A measure of more than 0.05 is best, and our model is at 0.16 for this measure. This is a good sign and indicates that our model is catching that seasonal variability in ILI cases.

Although some of these measurements are not ideal, the model performs fairly well.

Next, looking at the plots in figure 19 The plot in the upper left is the standard residual for p and shows the residuals as a time series. An ideal plot would have the residuals centered around zero with no patterns or trends. Our plot does have many residuals centered around zero, but there are also many outliers that are reaching to the -4/4 area. Additionally, there may be some pattern to the large spikes, which is not ideal. The plot in the upper right is the histogram and the kernel density estimate for the model. An ideal plot would display a normal distribution curve of the residuals. Again we can see that our plot displays ideal behaviors, with the kernel density curve in orange almost exactly following the green normal distribution curve. The bottom left plot is a plot of the residuals compared to a normal curve. Ideally, the plotted residuals (blue) would follow the red line going through the center of the plot (red) representing the normal distribution. In the middle of the line, our plot again follows the ideal plot with out residuals mostly falling right on the red line. However, during the tail ends of the plot the residuals fall off the line. Lastly, the bottom right plot displays a correlogram which is a plot of autocorrelation at different lags. Ideally, this plot would have all bars in the blue-shaded confidence interval. As you can see, our plot displays ideal behavior with many of the bars inside that 95 percent confidence interval.

4.5 Discussion: Random Forest

Looking at the random forest results in figure 21, we see that it matches the testing data extremely accurately. The RMSE, or Root Mean Squared Error of the random forest method is 0.1599%. This implies that our method is very accurate predicting within 0.1599% of the actual national ILI case percentage rate. This is actually more accurate compared to our non-parameter-optimized version of our random forest method, which had root mean squared error value of 0.1607%. This entails that our tuned parameters did have a positive effect on our model's accuracy, albeit slightly.

The model seems roughly in alignment with the actual output data. Common inaccuracies the model has are that it overestimates peaks and valleys, predicting them to be at a greater magnitude than they actually are. But that actually works in this context, as for influenza hospitalization rates, it may be more beneficial to err on the side of caution.

5 Conclusions

We now have all the pieces we need to answer our initial question: when should healthcare facilities begin to prepare countermeasures against ILI? How severe are those times of high ILI? How well can a predictive model work on time-series data such as this?

Looking at the first question, we can determine from the heatmap generated earlier in the project that there is a spike in ILI cases at the end of each year, and the beginning of the next. As this appears to happen regularly, hospitals should aim to prepare additional resources close to the end of the year, and to continue to be ready into the first approximately weeks into the next year.

From our modeling results, we can observe the intensity of each of the aforementioned spikes. Particularly from our ensemble results, we can tell that the peaks of these major ILI spikes can reach up to 8x the typical amount across the rest of the year. However, the spikes are sudden: they don't persist at this high level for very long, appearing and disappearing quickly.

Overall the random forest performed much better than the SARIMA, so we decided not to further pursue ensemble model. Interestingly enough, the SARIMA model performed much more effectively on the California state data compared to the National data.

A Code Appendix

Code Repository: https://github.com/kenzihebert/Data-Science-Capstone

References

[1] California population map. Wikimedia Commons. Accessed: 13 October 2025.

[2] How cdc classifies flu severity each season in the united states, Nov 2024.

[3] FluSurv-NET: Hospitalization Surveillance Data, 2025. Accessed: October 8, 2025.

[4] Weekly flu vaccination dashboard, May 2025.

[5] Outpatient Respiratory Illness Activity Map Determined by Data Reported to ILINet, 2025. Accessed: 13 September 2025.

[6] California Health and Human Services Open Data Portal. California Department of Public Health Influenza Surveillance, 2025. Accessed: October 12, 2025.

[7] Centers for Disease Control and Prevention. FluView: Weekly U.S. Influenza Surveillance Report, 2025. Accessed: 10 September 2025.

[8] Centers for Disease Control and Prevention (CDC). FluView Interactive: Influenza Hospitalizations Dashboard, 2025. Accessed: October 8, 2025.

[9] Center for Disease Control and Prevention.

[10] Ben Killingley and Jonathan Nguyen-Van-Tam. Routes of influenza transmission. *Influenza and Other Respiratory Viruses*, 7(s2):42–51, Aug 2013.

[11] Hidetaka Morita, Stephanie Kramer, Alex Heaney, Henry Gil, and Jeffrey Shaman. Influenza forecast optimization when using different surveillance data types and geographic scale. *Influenza and Other Respiratory Viruses*, 12(6):755–764, 2018.

[12] Elaine O. Nsoesie, John S. Brownstein, Naren Ramakrishnan, and Madhav V. Marathe. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and Other Respiratory Viruses*, 8(3):309–316, 2014.

[13] Dave Osthus. Fast and accurate influenza forecasting in the united states with inferno. *PLoS Computational Biology*, 18(1):e1008651, 2022.

[14] Vladimir S. Petrovic Sarah F. Ackley, Sarah Pilewski and Travis Porco. Assessing the utility of a smart thermometer and mobile application as a surveillance tool for influenza and influenza-like illness. *Health Informatics Journal*, 26(3):1899–1908, 2020.

[15] Aditya Saxena. Revisiting the equity premium puzzle: Time series analysis and forecasting from 1990 to 2012. In *2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, pages 1–5, 2025.

[16] statsmodels-developers. Sarimax: Introduction, 2024.

[17] Leili Tapak, Omid Hamidi, Mohsen Fathian, and Manoochehr Karami. Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare centers in iran. *BMC Research Notes*, 12(1), Jun 2019.

[18] Jeffery K. Taubenberger and David M. Morens. Influenza: the once and future pandemic. *Public Health Reports*, 125(Suppl 3):16–26, April 2010.

[19] the World Health Organization. The burden of Influenza, 2024. Accessed: 10 September 2025.

[20] Rens van de Schoot, David Kaplan, Jaap Denissen, Jens B Asendorpf, Franz J Neyer, and Marcel AG van Aken. A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, 85(3):842–860, 2014.

[21] Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, and Antoine Flahault. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*, 158(10):996–1006, 11 2003.

[22] Dan J. Vick, Asa B. Wilson, Michael Fisher, and Carrie Roseamelia. Comparison of disaster preparedness between urban and rural community hospitals in new york state. *Disaster Medicine and Public Health Preparedness*, 13(3):424–428, 2019.

[23] Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, Teun van den Brand, and PBC Posit. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2025. R package version 4.0.1.