Preparing for Influenza Season

Tools: Excel (Pivot, VLOOKUP, integration)

Tableau Public

Posted by Graham Field on February 10, 2024 – 10 mins read

Duration of the project: 1 week

Role: Data Analyst

Tools:

- Excel (Pivot, VLOOKUP, integration)
- Tableau Public

Data:

- Open Source Data
- 4 Possible Data Sets

Key Research Questions

● Provide information to support a staffing plan, detailing what data can help inform the timing and spatial distribution of medical personnel throughout the United States.

● Determine whether influenza occurs seasonally or throughout the entire year. If seasonal, does it start and end at the same time (month) in every state?

● Prioritize states with large vulnerable populations. Consider categorizing each state as low-, medium-, or high-need based on its vulnerable population count.

● Assess data limitations that may prevent you from conducting your desired analyses.

Introduction

Introduction: The annual influenza season in the United States poses significant challenges for healthcare providers. Vulnerable populations, such as young children and the elderly, are particularly at risk of developing severe complications from the flu, which often leads to increased hospitalizations. Medical staffing agencies play a crucial role in supporting hospitals and clinics by providing temporary staff to manage the surge in patients.

Objective: The primary objective of this analysis is to determine the optimal timing and allocation of staff to hospitals across the United States during the influenza season.

Scope: The project encompasses all hospitals within the 50 states of the United States, focusing on the upcoming influenza season to ensure adequate staffing and resource distribution. By examining historical data and identifying trends, we aim to enhance preparedness and response strategies for healthcare facilities.

Step 1: Initial Data Exploration and Cleaning

After drawing requirements from the stakeholders and the business requirements document I evaluated each data set. There were several considerations with the data namely: its reliability, ownership and ethical concerns. I don't want to present any data that could lead to the identification of individuals, given the scope and scale of this medical data.

The focus of this was to send staff and allocate resources to areas which experience the greatest need. I focussed on areas with the highest mortality rates.

From the 4 posible data sets i selected :

1. US Geographical Census Data
2. US Influenza Deaths

Step 2: Cleaning and Consistency Checks

During the exploration of the data i found several inconsistancies namely: the naming of states and areas which were recored but not part of the US such as Puerto Rico.

There was some missing data due to privacy requirements where data that was less than 10 was removed. I imputed using both random and averaged methods. But ultimately chose the random seed.

Step 3: Integration

In orderto use the data sets for analysis i had to integrate or merge them. This is due to the data sets often complete and ready for use.

I bean by mapping the data and trying to find commonalities between the data sets. I identified “State” and “Year” as common columns. There was an issue with granularity as The US census data had state and county with year while the Influenza Deaths had State with Year and Month.

Now that I have a common key to work with, I created Pivot tables which summarised the data on both data sets and allowed me to Integrate the two data sets. I used VLOOKUP to integrate the data sets into one.

County	State	Year	Total population	N
Autauga County	Alabama	2009	49584	
Baldwin County	Alabama	2009	171997	
Barbour County	Alabama	2009	29663	
Bibb County	Alabama	2009	21464	
Blount County	Alabama	2009	56804	
Bullock County	Alabama	2009	10917	
Butler County	Alabama	2009	20189	
Calhoun County	Alabama	2009	112969	
Chambers County	Alabama	2009	34704	
Cherokee County	Alabama	2009	24427	

US Census Data

State	State Cleaned	State Code	Year	Year Cleane	Month	Month Code
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04
Alabama	Alabama	1	2009	2009	Apr., 2009	2009/04

US Influenza Deaths Data

Now that I have the data sets integrated I began to segment them into age groups of 10 years and then created a secondary segment for retired vs non-retired or younger than 65 and older than 65.

Combined Key State and Years	Sum of <5 Years	Sum of 5-14 years	Sum of 15-24 years	Sum of 25-34 years
Alabama, 2009	307929	619584	656445	601455
Alabama, 2010	195281	453124	455090	432564
Alabama, 2011	170158	348517	394545	361849
Alabama, 2012	529615	995149	984038	1000440
Alabama, 2013	179409	354299	373793	379386
Alabama, 2014	299508	636920	673430	592431
Alabama, 2015	122440	248372	270865	270636
Alabama, 2016	161581	331488	348691	340797
Alabama, 2017	223096	450427	447310	463235
Alaska, 2009	52103	98092	113847	97175
Alaska, 2010	72284	151703	148286	137525

State and Year	<5 Years	5-14 years	15-24 years	25-34 years	35-44 years	45-54 years
Alabama,2009	103	54	49	37	39	77
Alabama,2010	107	60	56	54	56	60
Alabama,2011	68	62	52	58	60	74
Alabama,2012	98	45	56	52	46	56
Alabama,2013	113	48	47	53	57	47
Alabama,2014	110	37	55	67	62	79
Alabama,2015	100	52	42	50	55	55
Alabama,2016	119	48	57	43	60	47
Alabama,2017	101	72	44	48	58	56
Alaska,2009	109	53	59	62	46	39
Alaska,2010	111	48	48	40	53	54

Now that I had the two tables with matching data I could create a new normalised data table that would give me death rates by dividing the census data to the matching Influenza deaths.

Deaths 65-	Deaths65+	Total Deaths	Census 65-	census65+	Total Census	Death rate 65-	death rate 65+
420	726	1146	4639060	626542	4633360	0.009%	0.116%
478	765	1243	3124602	412999	3119042	0.015%	0.185%
435	772	1207	2851565	416946	2882272	0.015%	0.185%
417	744	1161	6804458	634149	6532201	0.006%	0.117%
472	783	1255	2794271	414784	2837469	0.017%	0.189%
505	783	1288	4741737	708257	4851671	0.011%	0.111%
483	883	1366	1926478	278403	1964357	0.025%	0.317%
499	757	1256	2555218	375229	2596535	0.020%	0.202%
506	940	1446	3302156	519888	3400030	0.015%	0.181%
415	170	585	731898	47809	683142	0.057%	0.356%
400	146	546	1120682	169413	1137974	0.036%	0.086%
447	159	606	2000753	269422	2009724	0.022%	0.059%

Step 4: Statistical Analysis and Testing

I began by stating 2 hypothesis.

Null Hypothesis (H0): The mean influenza mortality rate for individuals aged 65 and older is equal to the mean influenza mortality rate for individuals under 65.

H0: μInfluenza mortality 65+ years = μInfluenza mortality <65 years

Alternative Hypothesis (HA): The mean influenza mortality rate for individuals aged 65 and older is not equal to the mean influenza mortality rate for individuals under 65.

HA: μInfluenza mortality 65+ years <> μInfluenza mortality <65 years

The goal was to test whether the two groups had the same risk of death from influenza the first test I ran was a statistical test to determine: Distribution, Standard Deviation, Number of outliers, and correlation. I wanted to determine if the average mortality rate was higher or lower for the two age groups.

Data Spread Combi Data			
Dataset Name	65-Years CDC	65+Years CDC	Grand Total cdc
Sample or Population ?	Sample	Sample	Population
Normal Distribution ?	Normal	Normal	Normal
Variance	15299	952080	1179549
Standard Deviation 1	124	976	1086
Standard Deviation 2	247	1951	2172
Standard Deviation 3	371	2927	3258
Mean	490	890	1380
Number of outliers STD1	0	0	0
Number of outliers STD2	349	29	19
Number of outliers STD3	861	3817	4639
Outlier Percentage STD1	0.126361656	0.093681917	0.124183007
Outlier Percentage STD2	0.039215686	0.063318777	0.041394336
Outlier Percentage STD3	0	0	0
+1 Standard deviation upper	614	1866	2466
+2 Standard deviation upper	738	2842	3552
+3 Standard deviation upper	861	3817	4639
-1 Standard deviation lower	367	-86	294
-2 Standard deviation lower	243	-1061	-792
-3 Standard deviation lower	119	-2037	-1878
Correlation Combi Data			
Variables	65-Years CDC	65+Years CDC	Grand Total cdc
Proposed Relationship	People over retirement age are likely to die from influenza		
Correlation Coefficient	0.173	0.174	0.174
Strength of Correlation	Weak	weak	weak
Usefulness / Interpretation	From this data there is a low correlation and low chance of death from influenza		
Notes: •The population in every case is over 30 and so the distribution will be considered normal. •For the lower standard deviations the Numbers are negative, this is impossible for this type of data so we will only use the upper when counting the outliers. •The Census and CDC data for over 65 and in total would be with in the 2nd standard deviation as the data does not exceed 95%.			

The next test I ran was the two-tail test because I was interested in confirming the results from the previous test and confirming or rejecting the null hypothesis with certainty. I decide to include both Equal and unequal variance tests for the following reasons.

The first test shows the difference between the deaths in the 65+ age group and the overall census data for the 65+ population, highlighting the impact of influenza within this age group relative to its population size.

The second test directly compares the mortality rates between the two age groups, showing the disparity in influenza mortality between younger and older individuals.

t-Test: Two-Sample Assuming Unequal Variances		
	Deaths65+	*census65+*
Mean	890.0827887	730435.789
Variance	952080.3818	6.61922E+11
Observations	459	459
Hypothesized Mean Difference	0	
df	458	
t Stat	-19.2112299	
P(T<=t) one-tail	4.62535E-61	
t Critical one-tail	1.648187415	
P(T<=t) two-tail	9.25071E-61	
t Critical two-tail	1.965157098	

t-Test: Two-Sample Assuming Unequal Variances		
	Deaths 65-	*Deaths65+*
Mean	490.2745098	890.0827887
Variance	15298.88518	952080.3818
Observations	459	459
Hypothesized Mean Difference	0	
df	473	
t Stat	-8.7088285	
P(T<=t) one-tail	2.58544E-17	
t Critical one-tail	1.648081483	
P(T<=t) two-tail	5.17089E-17	
t Critical two-tail	1.964991997	

Analysis of the t-Tests

Test 1: Deaths65+ vs. Census65+

- **t Stat**: -19.2112299

- **P(T<=t) two-tail**: 9.25071E-61
- **t Critical two-tail**: 1.965157098

This test shows a highly significant difference between the mortality data for the 65+ age group and the census data for the same age group. However, this is not directly related to the null hypothesis about comparing mortality rates between age groups.

Test 2: Deaths 65- vs. Deaths65+

- **t Stat**: -8.7088285
- **P(T<=t) two-tail**: 5.17089E-17
- **t Critical two-tail**: 1.964991997

This test directly compares the mean influenza mortality rates between the two age groups:

- **t Stat**: -8.7088285
- **t Critical two-tail**: 1.964991997
- **P(T<=t) two-tail**: 5.17089E-17

Conclusion

The p-value for the second test (5.17089E-17) is extremely low, much lower than the typical significance level (α) of 0.05. This indicates that the difference in mean mortality rates between individuals aged 65 and older and those under 65 is statistically significant.

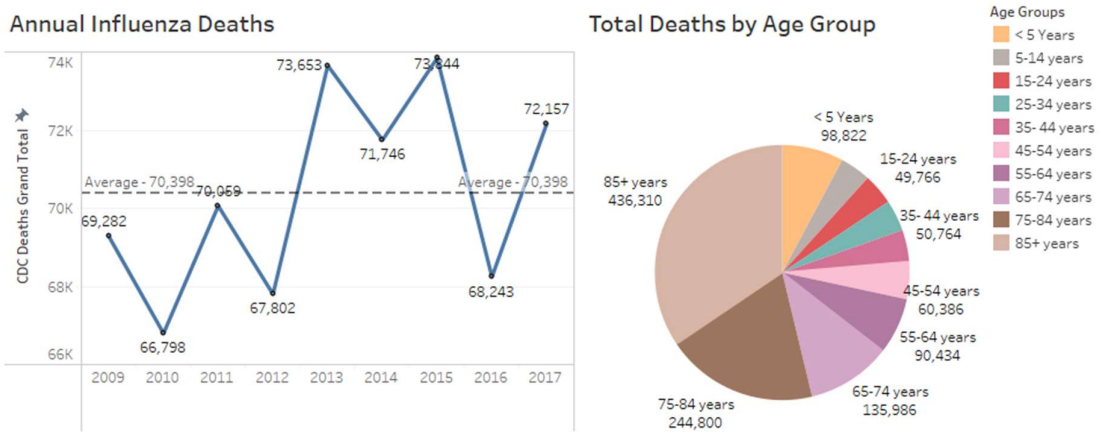
Decision:

- Since the p-value is much less than 0.05, we reject the null hypothesis (H0).

Summary:

Based on the results of the t-test comparing the mortality rates between the two age groups, we have sufficient evidence to reject the null hypothesis. This indicates that there is a statistically significant difference in the mean influenza mortality rates between individuals aged 65 and older and those under 65.

Step 5: Visual Analysis



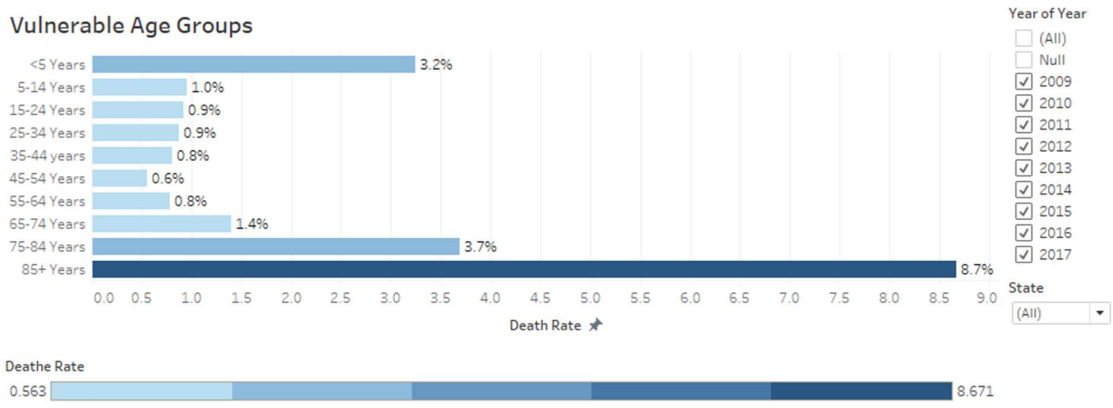
Defining Influenza and Who Is at Risk?

What is Influenza? Influenza, commonly known as the flu, is a contagious respiratory illness caused by influenza viruses. It can cause mild to severe illness and at times can lead to death. The flu spreads through respiratory droplets when people with the flu cough, sneeze, or talk.

Who is at Risk? Influenza significantly impacts certain age groups more than others. The most vulnerable age groups include:

- **Children younger than 5 years old**, with a mortality rate of 3.2%.
- **People over the age of 65 years**, with an average mortality rate of 4.6%. Within this group:
    - 65-74 years: 1.4% mortality rate
    - 75-84 years: 3.7% mortality rate
    - 85+ years: 8.7% mortality rate

Vulnerable Groups: The chart shows the death rates across various age groups, highlighting the higher risk among the very young and the elderly. These populations should be the focus of targeted interventions to reduce mortality rates during the influenza season.

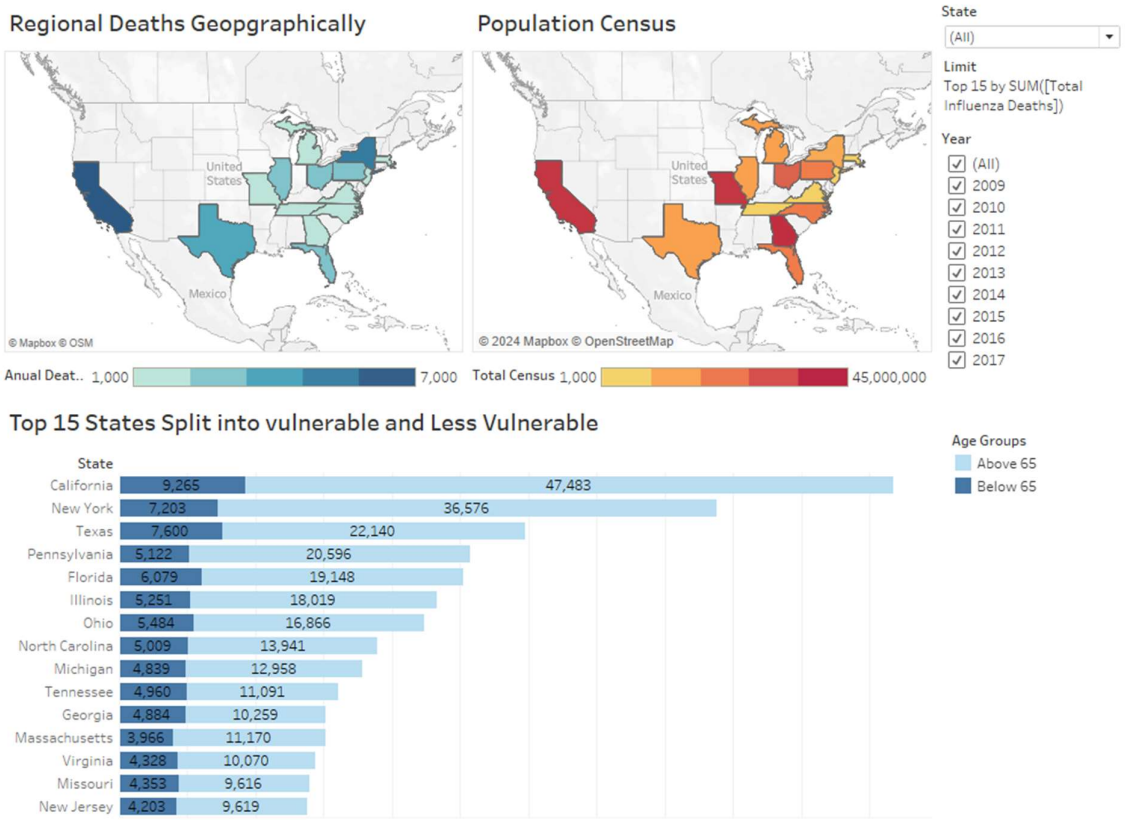


Where and Who is at Risk?

Geographical Analysis: The geographical distribution of influenza deaths shows significant variations. The top 15 states with the highest number of deaths include California, New York, Texas, Pennsylvania, and Florida, which are also among the most densely populated states.

Regional Deaths Geographically: Mapping the deaths geographically allows us to identify high-risk areas. This information is crucial for directing resources and planning healthcare strategies.

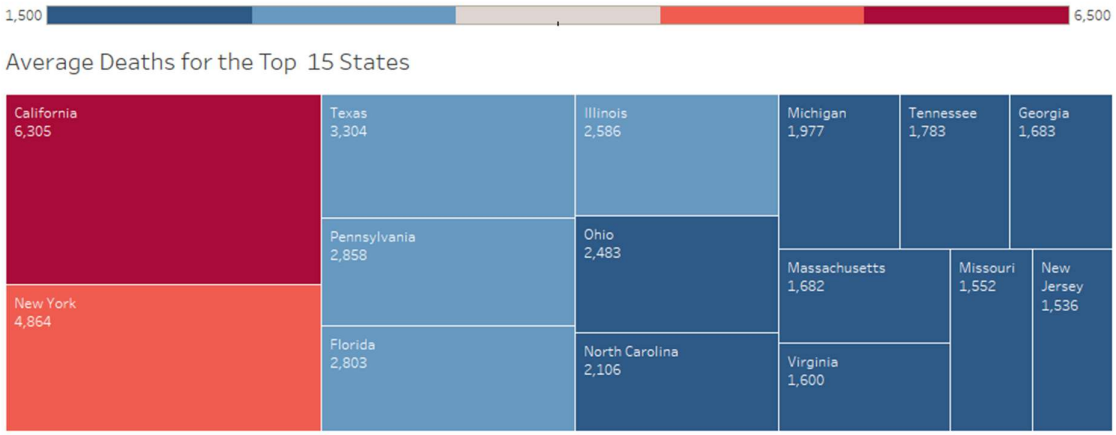
Total Deaths by Age Group: Analyzing the total deaths by age group reveals that older adults, especially those above 65, are at the highest risk. This demographic data is essential for preparing healthcare systems to meet the needs of these vulnerable populations.



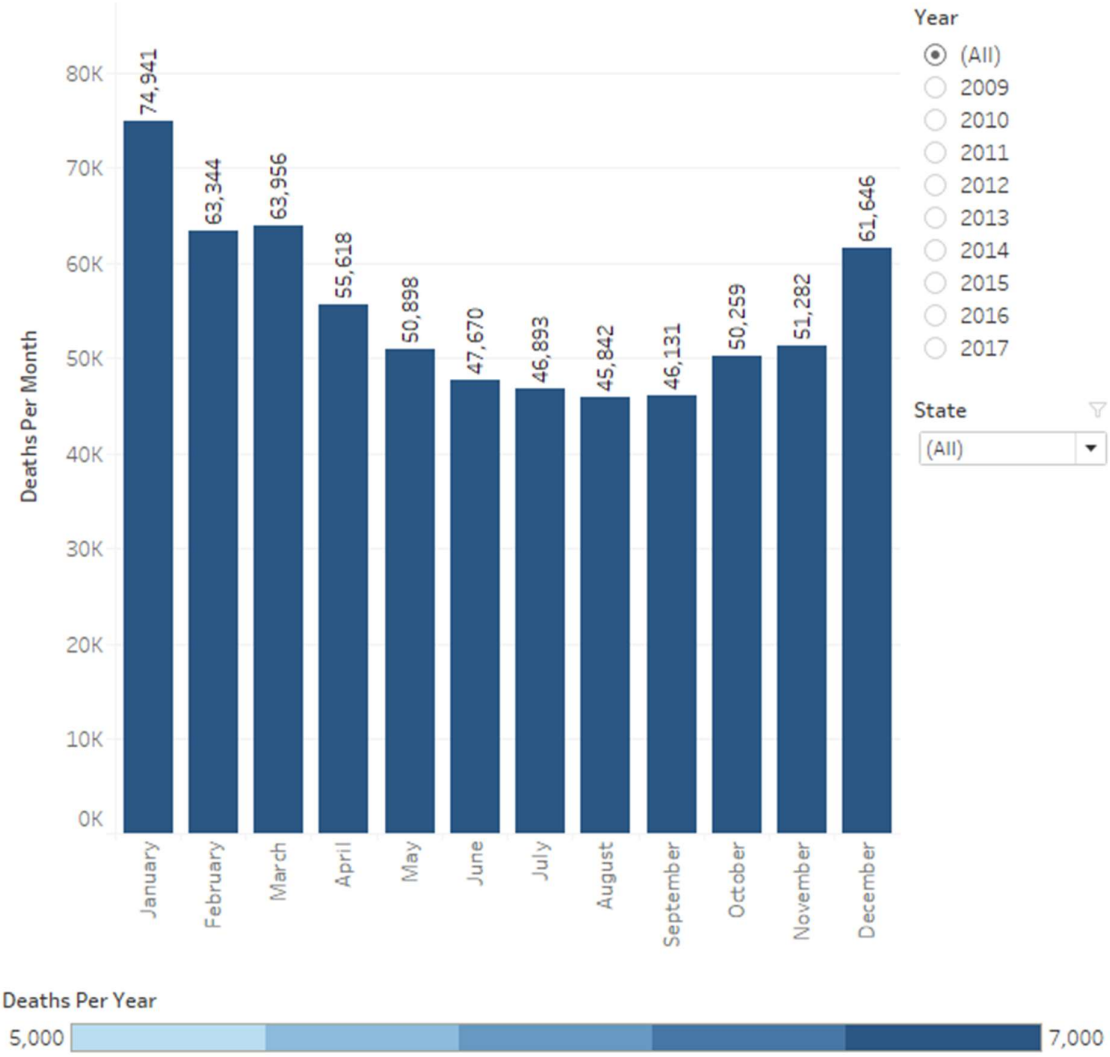
When is Influenza Season?

Influenza Seasonality: Influenza season typically starts in December, peaks in January, and declines by March. These months coincide with colder weather, which contributes to the spread of the virus.

State-specific Analysis: California, New York, and Texas have the highest number of deaths, correlating with their large populations. This trend underscores the need for targeted resource allocation during peak months.



Monthly Distribution of Deaths: Understanding the monthly distribution of deaths can help in planning the deployment of medical staff and resources. Ensuring that healthcare facilities are adequately staffed during peak months can save lives and improve patient outcomes.



Conclusion and Recommendations

Conclusion: The analysis shows that certain age groups and geographical regions require more resources during the influenza season. Specifically, children under 5 and adults over 65 years need prioritized care due to higher mortality rates.

Recommendations:

1. **Prioritize Care for Vulnerable Age Groups:** Ensure that resources are directed towards children under 5 and adults over 65.
2. **Prepare for Seasonal Peaks:** Staffing and resources should be heightened from December to March.

3. **Allocate Resources to High-Risk States:** States like California, New York, Texas, Pennsylvania, and Florida should receive the majority of resources. The next tier of states should receive moderate resources.
4. **Enhance Public Health Campaigns:** Increase awareness and vaccination efforts, particularly among high-risk groups and in high-risk states.

Further Analysis: Future research should include additional data on vaccination rates, living standards, nutrition, and substance use to better understand the spread and survival rates of influenza. Collaboration with public health agencies can provide a more comprehensive approach to managing influenza seasons.

Ongoing Monitoring and Data Limitations

Ongoing Monitoring:

1. Continue to track death rates over time to identify trends and emerging patterns.
2. Include comparisons between vaccinated and unvaccinated populations to assess the effectiveness of vaccination campaigns.
3. Assess how influenza deaths compare to other leading causes of death to allocate healthcare resources effectively.

Data Limitations:

1. Limited information on general health, fitness, diet, and hereditary factors can affect the analysis. More detailed health data could improve predictive models.
2. High-population density states with low death rates should be studied to identify effective health practices that could be implemented elsewhere.
3. Data quality and completeness are crucial; ensuring accurate reporting from healthcare facilities will enhance the reliability of future analyses.

Assessment Recommendations:

- Regularly update and review data to reflect current trends.
- Engage with local health departments to validate and supplement data.
- Implement technological solutions to improve data collection and analysis.

By addressing these limitations and continuing to monitor relevant data, we can improve our understanding of influenza patterns and enhance our preparedness for future seasons.