PREPARING FOR INFLUENZA SEASON: INTERIM REPORT

Project Overview:

- **Motivation:** The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.
- **Objective:** Determine when to send staff, and how many, to each state.
- **Scope:** The agency covers all hospitals in each of the 50 states of the United States, and the project will plan for the upcoming influenza season.

Hypothesis:

"If any state has more vulnerable population (including people over 65 and under 5), then there are high chances of influenza outbreak."

Data Overview:

- **Population Data by geography:**
  The data is an external source that comes from the U.S Census Bureau. The data contains total population, female/male population, age group by state and county from 2009 to 2017. Snapshot of dataset attached for quick reference:

County	Year	Total p	Male To	Female	Under 5	5 to 9 y	10 to 14	15 to 19	20 to 24	25 to 29	30 to 34	35 to 39	40 to 44	45 to 49	50 to 54	55 to 59	60 to 64	65 to 69	70 to 74	75 to 79	80 to 84	85 years
Autauga County, Al	2009	49584	24057	25527	3421.296	3570.048	4412.976	4016.304	2727.12	2875.872	3024.624	3520.464	4363.392	3966.72	3222.96	2627.952	2429.616	1983.36	1437.936	892.512	644.592	495.84
Baldwin County, Al	2009	171997	84263	87734	10663.81	10663.81	11695.8	11007.81	9459.835	9803.829	9631.832	10835.81	12383.78	13071.77	12211.79	11179.81	10663.81	8599.85	7223.874	5847.898	3955.931	2923.949
Barbour County, Al	2009	29663	15687	13976	1839.106	1631.465	2017.084	2165.399	2224.725	1957.758	1987.421	2135.736	2195.062	2076.41	1987.421	2017.084	1542.476	1186.52	860.227	919.553	415.282	533.934
Bibb County, Alaba	2009	21464	11164	10300	1287.84	1416.624	1502.48	1330.768	1459.552	1201.984	1395.16	1738.584	1867.368	1695.656	1438.088	1223.448	1159.056	1116.128	450.744	536.6	536.6	128.784
Blount County, Ala	2009	56804	28216	28588	3749.064	3521.848	4317.104	3635.456	3181.024	3578.652	3521.848	4260.3	4203.496	4146.692	3976.28	3465.044	3408.24	2385.768	2272.16	1704.12	965.668	624.844
Bullock County, Ala	2009	10917	6165	4752	786.024	480.348	895.194	906.111	731.439	818.775	458.514	807.858	927.945	840.609	862.443	633.186	513.099	349.344	262.008	163.755	152.838	316.593
Butler County, Alab	2009	20189	9517	10672	1393.041	1393.041	1332.474	1393.041	1312.285	1231.529	1049.828	1251.718	1130.584	1453.608	1514.175	1332.474	1170.962	827.749	767.182	787.371	464.347	423.969
Calhoun County, Al	2009	112969	53981	58988	7455.954	7681.892	6778.14	7455.954	7568.923	8359.706	6778.14	6665.171	7907.83	8133.768	8133.768	7568.923	5874.388	4631.729	4179.853	3389.07	2485.318	1920.473
Chambers County,	2009	34704	16425	18279	2186.352	2429.28	2116.944	2151.648	1908.72	2221.056	2151.648	2394.576	2151.648	2463.984	2498.688	2394.576	2012.832	1631.088	1249.344	1214.64	902.304	659.376

- **Influenza deaths by geography, time, age, and gender:**
  The data is an external source that comes from The Center for Disease Control (CDC). The data for influenza visits tracks patient visits to a medical provider for influenza. It tracks the number of visits, number of providers, and total patients seen from 2010 to 2019 from 3,500 outpatient healthcare providers. Snapshot of dataset attached for quick reference:

State	State Cod	Year	Month	Month Code	Ten-Year Age Groups	Ten-Year Age Groups Code	Deaths
Alabama	1	2009	Apr., 2009	2009/04	< 1 year	1	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	1-4 years	1-4	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	15-24 years	15-24	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	25-34 years	25-34	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	35-44 years	35-44	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	45-54 years	45-54	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	5-14 years	5-14	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	55-64 years	55-64	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	65-74 years	65-74	Suppressed
Alabama	1	2009	Apr., 2009	2009/04	75-84 years	75-84	18
Alabama	1	2009	Apr., 2009	2009/04	85+ years	85+	28
Alabama	1	2009	Apr., 2009	2009/04	Not Stated	NS	Suppressed
Alabama	1	2009	Aug., 2009	2009/08	< 1 year	1	Suppressed
Alabama	1	2009	Aug., 2009	2009/08	1-4 years	1-4	Suppressed
Alabama	1	2009	Aug., 2009	2009/08	15-24 years	15-24	Suppressed

Data Limitations:

- **Population Data by geography:**
  The survey data is entered manually and is likely to contain unnecessary noise due to errors and omission.
- **Influenza deaths by geography, time, age, and gender:**
  In Influenza Deaths data is that the death certificate of patients lists only one cause of death. This could create some discrepancies within vulnerable populations, such as those with AIDs—while the cause of death may be related to AIDs, their decline in health may have been initiated by influenza. However, the data is government administrative data and can be considered trustworthy. Another key point that makes data bias is that count of deaths less than 10 deaths are marked as suppressed.

Descriptive Analysis:

In order to test our project hypothesis relevant datasets were collected, cleaned and integrated together to perform statistical analysis and determine whether those relationships are meaningful—or if they occur by chance. The snapshot of statistical analysis is as under:

Data Spread				
Data Set Name	**Deaths over and above 65 years**	**Population over and above 65 years**	**Deaths less than 5 Years**	**Population less than 5 Years**
Mean*	565	268417	55	385316
Sample or Population	Sample	Sample	Sample	Sample
Normal Distribution	Normal	Normal	Normal	Normal
Variance	149113.0793	1.25354E+11	107.3600887	2.03141E+11
Standard Deviation	386.1516273	354053.6567	10.36147136	450712.0894
2*Standard deviation	772	708107	21	901424
Upper	1338	976525	76	1286740
Lower**	-207	-439690	34	-516108
Total Records	459	459	459	459
Outliers	34	61	8	19
Outlier Percentage	0.07	0.13	0.02	0.04

Note:

a) **Mean is calculated on total of deaths (i.e., for the period 2009- 2017) and total population (i.e., for the period 2010- 2019).*
b) ***Lower range values are shown in negative since Mean is lower than SQRT of Standard Deviation.*

Correlation Coefficient:

Variables	Deaths and Population >65 years	Deaths and Population <5 years
Proposed Relationship	Since population >65 is vulnerable to death from influenza, the correlation between these two variables should be high	Since population <5 is vulnerable to death from influenza, the correlation between these two variables should be high
Correlation Coefficient	0.93	0.01
Strength of Correlation	Strong	Very Low
Usefulness/Interpretation	This is a helpful statistic as it supports the hypothesis. It shows a very strong correlation between a state's population over 65 and its number of influenza deaths for people over 65. In other words, it supports the high rate of influenza mortality for this age group	This may appear to be a surprising result, as children under 5 are considered a vulnerable population. However, in the original dataset, random values were inputted for this population, so this statistic isn't truly representative.

Result and Insights:

At an alpha of 0.05, or confidence level of 95 percent, we can reject our null hypothesis and state that the influenza death rate for individuals in age group of 65 years and above is greater than the death rate for individuals in age group of less than 65 years. The snapshot of Statistical testing is attached in the table below:

Statistical Testing	
Hypothesis to test:	If any state has more vulnerable population (including people over 65 and under 5), then there are high chances of influenza outbreak.
Independent Variable:	Proportion of Vulnerable individuals (above 65 and under 5)
Dependent Variable:	Influenza Death Rate
Null Hypothesis:	The death rate for individuals 65+ years of age is less than or equal to the death rate for individuals less than 65 years of age.
Alternative Hypothesis:	The death rate for individuals 65+ years of age is greater than the death rate for individuals less than 65 years of age.
T-test type	One-tailed test because we are only interested if the sample mean is higher or lower than the population mean - only interested in one direction not both simultaneously.
Significance Level	Alpha = 0.05
P-Value	0.0000000000000096
Significance Level Assessment:	This p-value is significantly less than 0.05. Therefore, we can reject our null hypothesis.

t-Test: Two-Sample Assuming Unequal Variances		
	Total Deaths >=65	**Total Deaths <65**
Mean	849.8779956	528.8082789
Variance	690632.6095	74047.36054
Observations	459	459
Hypothesized Mean Difference	0	-
df	555	-
t Stat	7.866215327	-
P(T<=t) one-tail	**9.60061E-15**	-
t Critical one-tail	1.647603773	-
P(T<=t) two-tail	1.92012E-14	-
t Critical two-tail	1.964247525	-

Remaining Analysis and Next Steps:

Based on statistical analysis insights, the next step of our analysis would be to dig in deeper and extract valuable insights from our dataset and help stakeholders prepare their staffing plan for the upcoming Flu season. The spatial aspect plays most important role in Project's primary goal i.e., to ascertain which regions are more likely to have vulnerable population (including people over 65 and under 5) and in our previous exercise we focused mainly on formulating and statistically analysis our hypothesis that vulnerable patients are at high risk of Flu.

However, we have not addressed the following geographically yet. Now, we should look at the geographical representation of vulnerable populations and analyze medical facilities available in these regions to plan for medical staff and prepare accordingly for the upcoming Flu Season.

With the help of Data Visualizations, we can effectively compare density of Vulnerable patients across each state with total number of deaths due to Flu and draw insights from it. To do so we will be working on powerful data visualization and analytics tool, **Tableau.** Adding a visualization would make the comparison much more readable, and it would be easier to see the states with a large contrast between vulnerable populations and available staff.

As a final step, we will be presenting final results extracted through combination of statistical analysis and visualization outcome in Powerpoint Presentation to our stakeholders to enable them to make informed decisions and combat upcoming pandemic.

Appendix:

CDC Influenza Deaths Data
https://coach-courses-us.s3.amazonaws.com/public/courses/da_program/CDC_Influenza_Deaths_edited.xlsx

Census Population Data
https://coach-courses-us.s3.amazonaws.com/public/courses/data-immersion/A1-A2_Influenza_Project/Census_Population_transformed_202101.csv