. *Interim Report 2018* .

Medical staffing for the upcoming influenza season

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 patients are over 65 or under 5 years of age, they are more likely to develop severe symptoms or die from influenza, since they are more vulnerable. As a result, if a state has a large number of vulnerable people, there is a greater likelihood of severe cases of flu, hospitalisation and high mortality rates.

Data overview

US Census dataset :

This is an external source. The source is provided by the *US Census Bureau*, which means that it is **government data.** The data contains information on each US individual per county, sex and age from 2009 to 2017. Counts are broken into three categories : county, sex and age. There are different age groups, which start from under 5 years old and end at 85 years and over old. Each age group spans 5 years (eg. 5 to 9 years,10 to 14 years,15 to 19 years, 20 to 24 years, etc.).

Influenza Deaths dataset :

This is an external data source. The data is provided by the *Centers for Disease Control and Prevention (CDC)* through their National Center for Health Statistics, which means that it is **government data**. The data contains influenza-related death counts per month and per year in the United States from 2009 to 2017. Counts are categorised into different age groups (From under <1 year years old to 85+ years old) and per state.

Data limitations

US Census dataset :

The data shows the geographic counts of people living in the US over multiple years, regarding their age and sex. As it is a government data, we can assume that most of the information is accurate, though we should take into account the limitation in the database regarding **non-participants rates**, **errors** or **missing/incomplete data**.

Influenza Deaths dataset :

- **Human errors limitation :** The data might contain some **errors or missing/incomplete data** in the reporting. Indeed, the deaths could be underreported or misclassified. For instance, deaths on a death certificate only list one cause of death, which means that an influenza death could be reported instead of another disease which might have played a significant part in the death of the patient, or some death might not be reported as caused by influenza because, at that moment, the patients might have another disease that would have been reported as the cause of death instead of influenza. This could, then, lead to misrepresentation and bias.
- **Death counts limitation** : We are not able to determine the exact number of deaths for the "Suppressed" death. Indeed, the value "Suppressed" indicates that the number of deaths was between 0 to 9. Not having the exact number limitate ourselves in our counting and changing those values could lead to bias in our analysis.
- **Regional limitation** : We might be able to identify the death rate per state, but lack in-depth information about the geolocalisation, such as counties, clinics or hospitals of patients, etc. which could have helped to have an in-depth insight of the influenza-patients geo-spatial representation.

Descriptive analysis

	Mean	Standard deviation
Var.1 - Census dataset (65+ years old)	806988,94	887017,19
Var. 2 - Influenza deaths dataset (65+ years old)	826,29	1014,14

Table 1. *Analytics study for the 65+ years old people*

- **Correlation** : The analyses demonstrated a strong positive linear relationship between the two variables (Census data 65+ years old and Influenza death 65+ years old) with a coefficient of **0.942647681**. It shows, then, that as one variable increases, the other variable tends to increase as well.
- **Correlation on normalised data :** For in depth study and accuracy in the statistical analysis, the data were normalised (%= Var.2/Var.1) and the correlation was calculated on the normalised data. The analyses demonstrated, this time, a moderated positive linear relationship between the two variables with a coefficient of **0,26**, rejecting the perfection of our 1st correlation study, still, it shows that as one variable increases, the other variable tends to increase as well.

Under 5 years old study ? The analysis for the people under 5 years old could not be led further since all data in the influenza death, for the age group, are "Suppressed". Since we do not have an exact number of the deaths and these numbers would be low in any case, **no further study should be conducted.**

- **Results :** We can say that the statistical analysis, at this stage, is proving a part of our hypotheses : *if a state has a large number of vulnerable people, there is a greater likelihood of severe cases of flu, hospitalisation rates and mortality.* Indeed, the data are clearly showing that the higher the population of people aged 65+ in a state, the higher the death rate of people aged 65+.

Results and insight

- **Null hypothesis :** People over 65 years old are not vulnerable and do not have more risk of dying from influenza.
- **Alternative hypothesis :** People over 65 years old are vulnerable and have more risk of dying from influenza.

t-Test: Two-Sample Assuming Unequal Variances

	65 - 85+ years (Influenza deaths)	*Under 65 years old (Influenza deaths)*
Mean	826,2875817	78,76470588
Variance	1028483,747	22903,91395
Observations	459	459
Hypothesized Mean Difference	0	
df	478	
t Stat	15,61886217	
P(T<=t) one-tail	**4,95504E-45**	
t Critical one-tail	1,648047653	
P(T<=t) two-tail	9,91009E-45	
t Critical two-tail	1,964939272	

Table 2. *t-Test on the 65+ years vs -65 years old influenza deaths*

The P(T<=t) is **4,95504E-45**. Since the value is under the significance level 0.05, then we have full confidence to reject the null hypothesis in favour of the alternative hypotheses. Indeed, the two groups are significantly different. We can then conclude that people over 65 years old are vulnerable and have a higher risk of dying from influenza than people under 65 years old.

➢ **Results :** We can say that, at this stage of the analysis, the data are proving the initial hypotheses. Indeed, not only can we confirm that the number of influenza deaths is higher in the age group 65+ than the age group under 65+ years old, but also that the states with the highest number of people from this age group 65+ have a high influenza death rate compared to people under 65 years old.

Remaining analysis and next steps

- Determine which states to prioritise for medical staff deployment, now that we could prove our hypothesis. If possible in the study, determine which counties per state to prioritise.

- Determine how many medical professionals to send per state : all states have different death rates numbers for 65+ year old people. In fact, some states have very high death rates compared to others. On which baseline figures should we define the number of medical staff to send in those states ?

- Determine the exact timeframe when the influenza is the most active per state and per season (Some states may have high flu activity in November and others in December, ect. - Does climate and temperature play a role in this ?).

- Conduct temporal, statistical, spatial and textual visualisations / Prepare and present final presentation to stakeholders.

Appendix

- **Business requirement :**
https://docs.google.com/document/d/1AhnxZ0FNvL2AZ__ea3vdojt4iA_HSFURBi0o2ezD_So/edit?usp=sharing

- **Influenza dataset :**
https://coach-courses-us.s3.amazonaws.com/public/courses/da_program/CDC_Influenza_Deaths_edited.xlsx

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

- **Data sourcing :**
https://docs.google.com/document/d/1epKcDXb5zhEhJqzWgWxfwv2X6etBp8RcemX5yu_BD6s/edit?usp=sharing

- **Statistical analysis :**
https://www.icloud.com/iclouddrive/007Z0brDOh_FnpQstdfmojOUQ#Task_1.8._Melissa

- **Statistical hypothesis testing :**
https://www.icloud.com/iclouddrive/044kVQQ8wZoAzNADubygdN8Ew#Task-1.9.Me-lissa_(1)