# The effect of flu jabs in children on deaths by influenza

Case study by Julia Fortuny Wollny

2021

# Context

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.

A medical staffing agency provides this temporary staff.

# Objective

To help this medical staffing agency on an as-needed basis by examining trends in influenza and how they can proactively be used to plan for flu season.

In order to achieve this, I asked myself:
- How is the ratio of casualties vs vaccinated children in each state?
- Will increasing vaccinations in children help decrease the number of deaths?

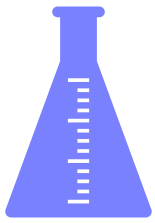
This project was created as part of the CareerFoundry Data Analytics program.

| **Role:** | **Project duration:** |
|---|---|
| Data Analyst | 2 months, delivered on time |
| **Data used:** | **Tools used:** |
| Data on the demographics of flu deaths, US population and flu shots in children provided by the US Census and Centre for Disease Control (CDC). | • Excel<br>• Tableau |

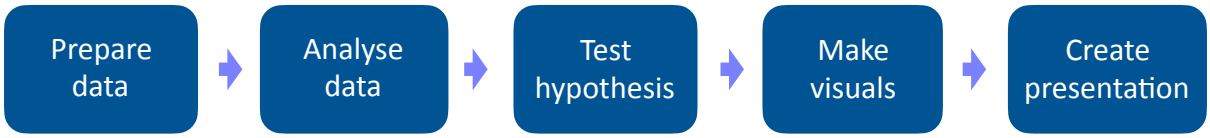
**Download project brief here:**
Preparing for Influenza Season Project Brief

# Skills

| Excel | Tableau |
|---|---|
| ✦ Sourcing data | ✦ Composition & comparison charts |
| ✦ Data profiling & integrity | ✦ Temporal visualisations & forecasting |
| ✦ Data quality measures | ✦ Histograms & box plots |
| ✦ Data transformation & integration | ✦ Scatterplots & bubble charts |
| ✦ Conducting statistical analyses | ✦ Spatial & textual analysis |
| ✦ Statistical hypothesis testing | ✦ Storytelling & presenting results to stakeholders |



# Approach & Methodology




I started by translating business requirements to data questions. For example:

- Which states have the most severe cases and deaths?
- Can we define risk groups or areas?

I defined a research hypothesis:

**If we increase vaccines in children, then the number of casualties will decrease.**

I went on to **draw up a project management plan** including the following sections:

- Stakeholder communication
- Schedule and milestones
- Project deliverables
- Audience definition

Following that I examined the provided data sources for:

- Collection methods
- Relevance

and wrote up a description of them.

Next, I started to **prepare the data**. I examined it for **terminology and structure.**

I resolved integrity issues through examining summary statistics, amongst others, and created a data profile.

| | | | | | |
|---|---|---|---|---|---|
| State code (B): structured, qualitative nominal | | | | | |
| Year (C): time invariant, structured, quantitative discrete | | | | | |
| Month (D): time invariant, structured, quantitative discrete | | | | | |
| Month code (E): time invariant, structured, quantitative discrete | | | | | |
| Ten-Year Age Groups (F): time invariant, structured, qualitative ordinal | | | | | |
| Ten-Year Age Groups Code (G): time invariant, structured, quantitative discrete | | | | | |
| Deaths (H): time variant, structured, quantitative discrete | | | | | |
| | | | | | |
| Year | | | | | |
| | min | 2009 | | | |
| | max | 20133 | | | |
| | mean | 2017,66 | | | |
| | | | | | |
| | | | | | |
| Deaths | | | | | |
| | min | 10 | | | |

Here for example there was an error, which I corrected by inputing an educated guess (2013).

Excerpt of my data profile

Following that I determined and improved the **quality of the data sets**, including addressing:

- Uniqueness (duplicates)
- Completeness (missing values)
- Timeliness (is the data in this set timely enough for this project?)

I used pivot tables to create **frequency tables**.

| | | | |
|---|---|---|---|
| 2010 | 1 | | 1 |
| 2011 | 1 | | 1 |
| 2012 | 1 | | 1 |
| 2013 | 1 | | 1 |
| 2014 | 1 | | 1 |
| 2015 | 1 | | 1 |
| 2016 | 1 | | 1 |
| 2017 | 1 | | 1 |
| **Accomack County, Virginia** | **7** | | 1 |
| 2009 | 1 | | 1 |
| 2010 | 1 | | 1 |
| 2011 | 1 | | 1 |
| 2012 | 1 | | 1 |
| 2013 | 1 | | 1 |
| 2015 | 1 | | 1 |
| 2016 | 1 | | 1 |
| **Ada County, Idaho** | **9** | | 1 |
| 2009 | 1 | | 1 |
| 2010 | 1 | | 1 |

Excerpt of my frequency tables

In order to be able to conduct the analysis I had to **unify 2 data sets** in Excel.

For that, it was necessary to:

- Map the data
- Transform the data
- Integrate the data

The next step was to **analyse** the data. I used statistical methods like:

- Standard deviation
- Variance
- Correlation

Amongst others, I checked if there was a correlation between the percentage of males in a population and the death rate.

The result was a coefficient of 0,4 (moderate relationship).

**I ruled out that this information would be relevant to the results.**

I also conducted an **inferential analysis** via **hypothesis testing**.

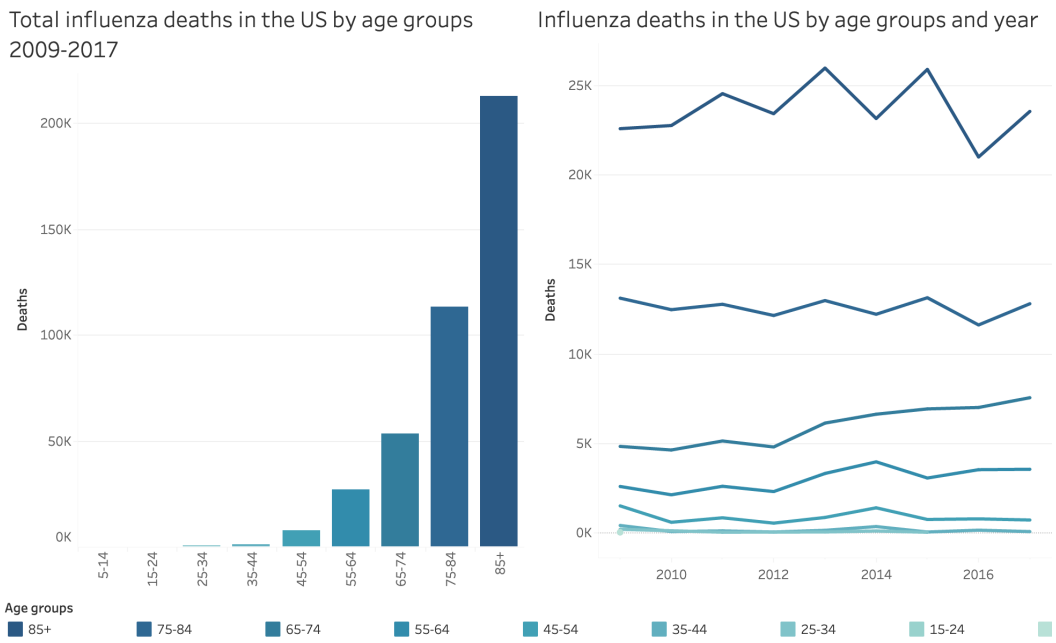
I decided to test if the female population is larger than the male population.

**Null hypothesis: female population ≤ male population**
**Alternative hypothesis: female population > male population**

I used the **T-test** and computed the **P value**.

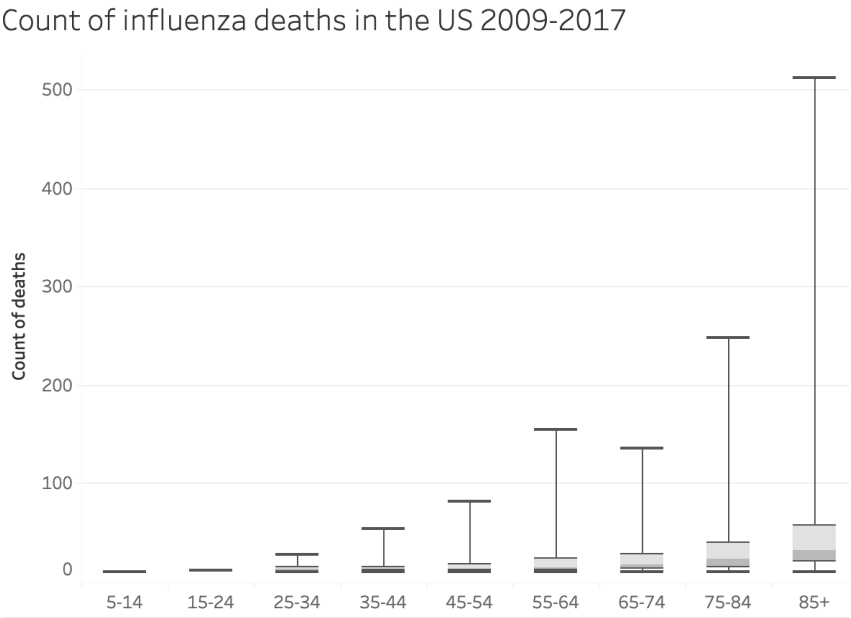
**At an alpha of 5% I couldn't find any significant difference between the female and male populations.**

In preparation for the presentation, I turned some descriptive statistics into bar and line plots.



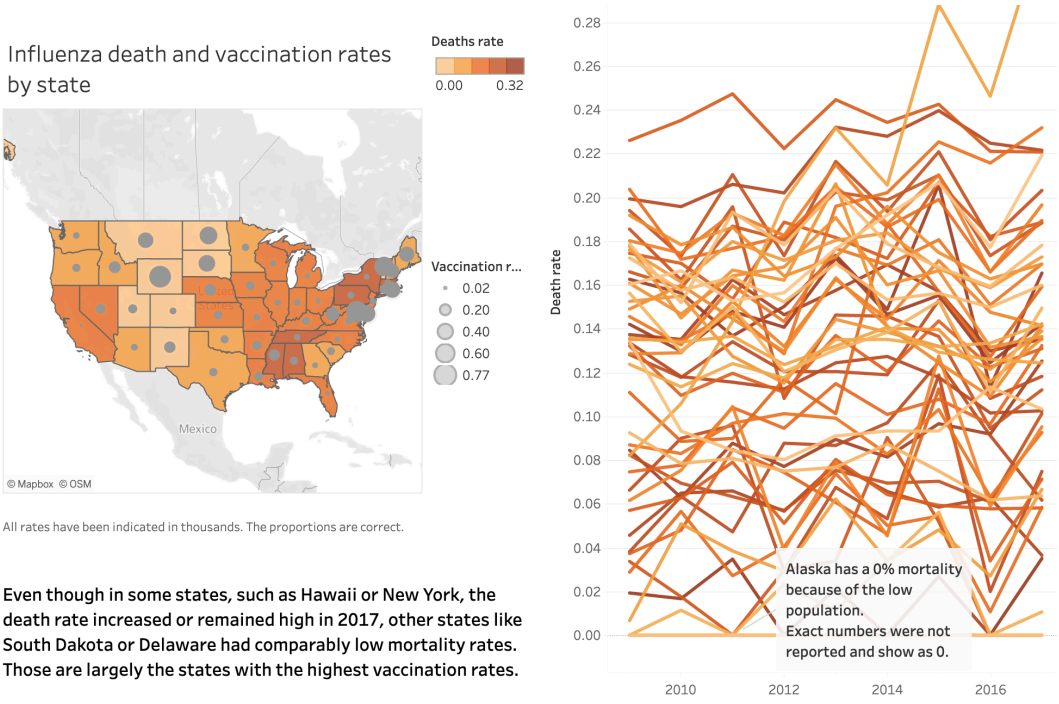

Here we can see that the 85+ group is the most affected in terms of number of deaths and through time.

Using **statistical visualisations** such as a box and whisker plot, I could further check the distribution of casualties in the population.




The box and whisker plot shows the distribution and spread of deaths in each age group.

I also created an interactive spatial and temporal chart.




By clicking on one state on the left it will highlight the corresponding line on the right.

Finally, I checked the relationship between the casualty rate and the percentage of vaccinations in children in each state and I found a **strong correlation**.

For the presentation I turned all this into a **compelling story** with a beginning, middle and end.

You can see the full Tableau presentation here and listen to my video presentation of it here.

# Recommendation

At the end of the presentation I recommended to test the correlation for causation, to find out if the link between the two variables can be confirmed as cause and effect.

Through **finding out if there is a causation** we can then **rule out** the the null hypothesis: *if vaccinations in children increase, then the number of deaths increase or remain the same.*

# Future steps

If the hypothesis can be rejected, this would prove the effectiveness of vaccinations in children. As a result, it would be recommendable to **launch a nationwide vaccination campaign**.

If the hypothesis cannot be rejected then it would be necessary to **summon the stakeholders, brainstorm for another research hypothesis** and **conduct another analysis**.

# Deliverables

- Interim report, download here,
- Document for stakeholders with reflections on data limitations & metrics, download here,
- Tableau presentation,
- Video presentation.

**Full link to the presentation:**

https://public.tableau.com/app/profile/julia.fortuny/viz/Theeffectofflujabsinchildrenondeathsbyinfluenza/Presentation

**Find me on:**

Tableau: https://public.tableau.com/app/profile/julia.fortuny

LinkedIn: https://www.linkedin.com/in/julia-f-18144718/

Github: https://github.com/juliafor/