

# **CASE STUDY 025**

## **[Python]**

### **Influenza Spread Analysis**

---

## Difficulty Level: 2 of 3

Influenza (flu) viruses can be detected year-round in the United States. But historical data from the CDC has evidence that there is a seasonal component in the spread of the disease, being much more common during the fall and winter.

CDC stands for Center for Disease Control and Prevention. It is one of the major operating components of the Department of Health and Human Services of the United States.

At the first sight, a question pops up: is the cold responsible for you to get sick?

You are a data scientist helping the CDC to answer this question. There are two datasets available for this analysis.

1. The number of deaths and the death rate associated with the Influenza virus by State in 2015;
2. The average temperature during the winter season from 2015/2016 by State;

### *Disclaimer:*

- *The number of deaths and death rate associated with the influenza was downloaded from the CDC;*
- *Death rate is the number of deaths by 100,000 of population;*
- *The average temperature is an estimate roughly calculated to this case study;*

Your analysis should follow these steps:

1. Load both datasets and merge them;
2. Create a function that receives a temperature in Celsius and converts it to Fahrenheit.
3. Create a function that receives a temperature in Fahrenheit and converts it to Celsius. Make both transformations in the same function, receiving as additional argument the type of conversion you want to perform.
4. Apply the conversion to your data frame, representing the temperature in Fahrenheit.
5. Calculate the Pearson's correlation coefficient to determine if the number of deaths associated with influenza and the average temperature are correlated. Do the same thing to the death rate, instead of absolute numbers.

6. Create a function that receives the correlation coefficient calculated earlier and interprets its results.
7. Graph the relationship between number deaths associated with influenza and average temperature using a scatter plot. Do the same thing to the death rate, instead of absolute numbers.

Good luck!

*Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file.*