CIS 5270 Project - 2 Death Rate for Pneumonia and Influenza Using RStudio

# **Death Rate for Pneumonia and Influenza**

# **Using RStudio**



**SUBMITTED BY: -**

**DIVYA PAKHALE**

**NEHA GULI.**

**Table of Contents**

**Objective** ----------------------------------------------------------------------------------------------- 3

**Dataset URL** -------------------------------------------------------------------------------------------3

**Data Cleaning** -----------------------------------------------------------------------------------------4

**Data Visualization** ------------------------------------------------------------------------------------6

**R code for all Analysis & Visualizations**--------------------------------------------------------11

**Statistical Summary** --------------------------------------------------------------------------------14

**Functions** --------------------------------------------------------------------------------------------- 14

**References** -------------------------------------------------------------------------------------------- 17

**A) Objective of the Study:**

Pneumonia is a serious infection or inflammation of the lungs. The air sacks fill with pus and other liquid, blocking oxygen from reaching the bloodstream. If there is too little oxygen in the blood, the body’s cells cannot work properly, which can lead to death. Influenza(flu) is a highly contagious viral infection that is one of the most severe illnesses of the winter season. Influenza is spread easily from person to person, usually when an infected person coughs or sneezes. Influenza is a common cause of pneumonia, especially among young children, the elderly, pregnant women, or those with certain chronic health conditions or who live in a nursing home. Most cases of flu never lead to pneumonia, but those that do tend to be more severe and deadly. In fact, flu and pneumonia were the 8-leading cause of death in the United States in 2015.

**B) Dataset URL's:**

Dataset:<https://data.cdc.gov/NNDSS/Deaths-from-pnuemonia-and-influneza-and-all-deaths/pp7x-dyj2>

This dataset consists of 29.2K rows and 12 columns. The data present in the dataset is from year 2009 to 2018. Data was last Updated on April 13, 2018, the Metadata Last Updated was on January 9, 2018 and the dataset was Created on September 13, 2016. The dataset consists of following columns:

Deaths from influenza – It determines the number of deaths due to influenza.,

Deaths from pneumonia – It determines the number of deaths due to pneumonia.

Geoid – It describes the geographical area like national, region and state.

State – Based on geographical id, states were defined.

Age – This determined the different age group from infant to adults.

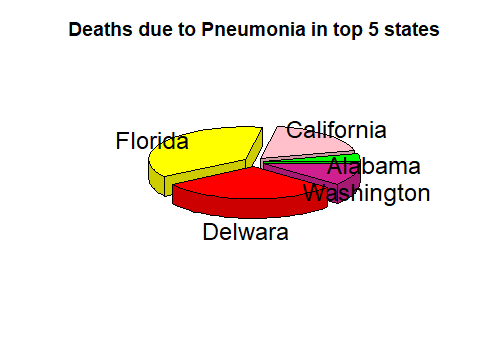
Deaths from pneumonia and influenza – It determined the deaths occurred due to both.

**C) Data Cleaning:**

|  |  |
| --- | --- |
| **Problem** | **Un-cleaned and Cleaned Data** |
| **Deleting**  **/Removing irrelevant columns** | The dataset contained region column which was not required in any of the visualizations. |
| **Splitting Columns** | The "MMWR\_Year\_Week" column had a united date, but we required separated MMWR by MMWR\_Year, MMWR\_Week. Hence, we splitted the column into two parts. |
| **Removing the NULL values** | There were Null values for States, we omitted it with the help of omit function. |

**D) Data Visualization:**

**1) Deaths Due to Pneumonia in top 5 States**

**(Application Used: plotrix, dplyr, pie chart)**

> install.packages("dplyr")

> library(dplyr)

> install.packages("plotrix")

> library('plotrix')

> top\_5<-a12 %>%

+ group\_by(State) %>%

+ summarise(State,Deaths\_from\_pneumonia,Region) %>%

+ head(5)

> colors1<-c("Green","Pink","Yellow","Red","VioletRed")

> pie3D(top\_5$Region, labels = top\_5$State, explode=0.1,main = "Deaths due to Pneumonia in top 5 states", col = colors1)

The above Pie-Chart, shows the death rate due to pneumonia in the top 5 states like Florida,

California, Delaware, Alabama and Washington. The package used in this analysis are dplyr

and plotrix. We have taken the top 5 states with the help of head () function and group\_by ()

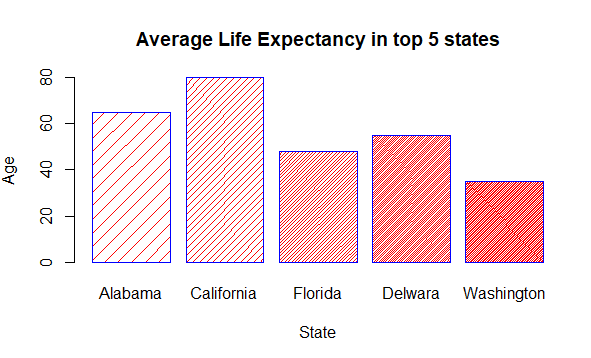
function. With the help of above analysis, we are able to determine which state has highest

number of deaths due to Pneumonia. From the visualization, we are able to depict that the

states of Florida and Delaware has almost the high number of deaths compared to other

states.

**2) Average Life Expectancy**



**(Application Used: dplyr, group\_by, barplot)**

> setwd("C:/Users/divya/Desktop")

> getwd()

[1] "C:/Users/divya/Desktop"

> load<-read.csv("P&I.csv",header=TRUE,sep=",")

> View(load)

> attach(load)

The following objects are masked from load (pos = 4):

Age, geoid, Region, State, Year

> library(dplyr)

> load %>%

+ group\_by(State)%>%

+ head(load,n=5)

> top\_5<-head(load,n=5)

> View(top\_5)

> barplot(top\_5$Age,names.arg = top\_5$State,xlab="State",ylab="Age",col="red",main="Average Life Expectancy in top 5 states",border="blue",density=c(10,20,40,60,80))

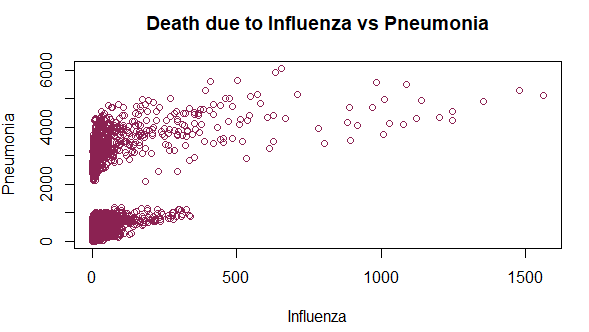
Above given barplot shows the Average life expectancy of the people according to their age

in the top 5 states. As there were lot of data for the states so we have extracted only the top 5

states with the help of group\_by function and by using head function. From this visualization, we are trying to analyze what is the average life expectancy for the people in different states

if they are suffering from any of the two diseases. From the bar chart we can analyze that, the highest average life expectancy in the state of California is about 80 years and the lowest is in the state of Washington between the age of 20-40 years.

**3) Death due to Influenza vs Pneumonia**



**(Application Used:plot)**

> setwd("C:/Users/divya/Desktop")

> getwd()

[1] "C:/Users/divya/Desktop"

> data1<-read.csv("P&I.csv",header=TRUE,sep=",")

> View(data1)

> keepcols<-c(3,4,5,8,9)

> data4<-data1[,keepcols]

> View(data4)

> plot(data4$Deaths\_from\_influenza ,data4$Deaths\_from\_pneumonia,xlab = "Influenza",

ylab="Pneumonia", main="Death due to Influenza vs Pneumonia",col= "VioletRed4")

Scatter plot are best to depict the relationship between the two attributes. The above scatter

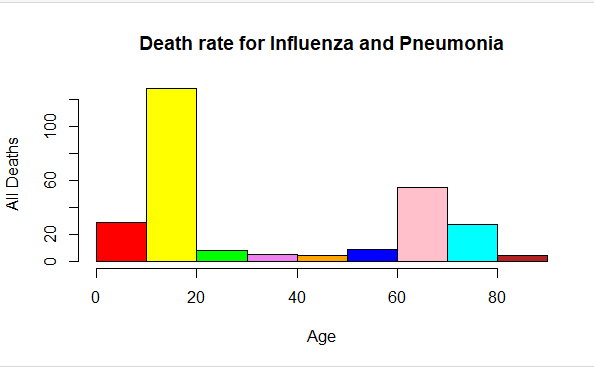
plot shows the comparison of deaths due to influenza versus death due to pneumonia. Plot

function is used to draw the visualization for the scatter plot. Analysis done by scatter plot

helps us to determine what were the number of deaths for adults or infants due to pneumonia

and Influenza.

**4) Mean of All Death Rate due to Influenza and Pneumonia**



**(Application Used: mean, histogram)**

> setwd("C:/Users/divya/Desktop")

> getwd()

[1] "C:/Users/divya/Desktop"

> data<-read.csv("P&I.csv",header=TRUE,sep=",")

> View(data)

> attach(data)

> mean(All\_Deaths)

[1] 5029.227

> hist(Age)

> colors<-c("red","yellow","green","violet","orange","blue","pink","cyan","firebrick","deeppink","lightpink4","mediumpurple4","yellow4")

> hist(Age, main = "Death rate for Influenza and Pneumonia", xlab = 'Age', ylab ='All Deaths', col = colors)

In the above histogram chart, we have tried to analyze all deaths of adults and infants with its age. We can analyze from this histogram that; maximum number of all deaths were above 100 at the age of 10-20 years. We can also determine that the number of deaths at the age between 40-60 years were very less. Mean function is used to determine the range of all deaths among different age groups.

**R Code for all the Analysis and Visualizations:**

**Data – Cleaning:1**

> setwd("F:/MIS/5270/R Project")

> getwd()

[1] "F:/MIS/5270/R Project"

> data1<-read.csv("PI.csv",header=T,sep=",")

> View(data1)

> keepcols<-c(1,3,4,5,6,7,8,9,10,11,12)

> data4<-data1[,keepcols]

> View(data4)

**Data – Cleaning:2**

> setwd("F:/MIS/5270/R Project")

> getwd()

[1] "F:/MIS/5270/R Project"

> b<-read.csv("Year-Week.csv",header=T,sep=",")

> View(b)

> b1<-separate(b,MMWR.Year.Week,c("MMWR Year","MMWR Week"),sep="-")

> View(b1)

**Data – Cleaning:3**

> setwd("F:/MIS/5270/R Project")

> getwd()

[1] "F:/MIS/5270/R Project"

> c<-read.csv("PI.csv",header=T,sep=",")

> View(c)

> cc<-na.omit(c)

> View(cc)

**Visualization – 1**

> install.packages("dplyr")

> library(dplyr)

> install.packages("plotrix")

> library('plotrix')

> top\_5<-a12 %>%

+ group\_by(State) %>%

+ summarise(State,Deaths\_from\_pneumonia,Region) %>%

+ head(5)

> colors1<-c("Green","Pink","Yellow","Red","VioletRed")

> pie3D(top\_5$Region, labels = top\_5$State, explode=0.1,main = "Deaths due to Pneumonia in top 5 states", col = colors1)

**Visualization-2**

> setwd("C:/Users/divya/Desktop")

> getwd()

[1] "C:/Users/divya/Desktop"

> load<-read.csv("P&I.csv",header=TRUE,sep=",")

> View(load)

> attach(load)

The following objects are masked from load (pos = 4):

Age, geoid, Region, State, Year

> library(dplyr)

> load %>%

+ group\_by(State)%>%

+ head(load,n=5)

> top\_5<-head(load,n=5)

> View(top\_5)

> barplot(top\_5$Age,names.arg = top\_5$State,xlab="State",ylab="Age",col="red",main="Average Life Expectancy in top 5 states",border="blue",density=c(10,20,40,60,80))

**Visualization-3**

> setwd("C:/Users/divya/Desktop")

> getwd()

[1] "C:/Users/divya/Desktop"

> data1<-read.csv("P&I.csv",header=TRUE,sep=",")

> View(data1)

> keepcols<-c(3,4,5,8,9)

> data4<-data1[,keepcols]

> View(data4)

> plot(data4$Deaths\_from\_influenza ,data4$Deaths\_from\_pneumonia,xlab = "Influenza", ylab="Pneumonia", main="Death due to Influenza vs Pneumonia",col= "VioletRed4")

**Visualization-4**

> setwd("C:/Users/divya/Desktop")

> getwd()

[1] "C:/Users/divya/Desktop"

> data<-read.csv("P&I.csv",header=TRUE,sep=",")

> View(data)

> attach(data)

> mean(All\_Deaths)

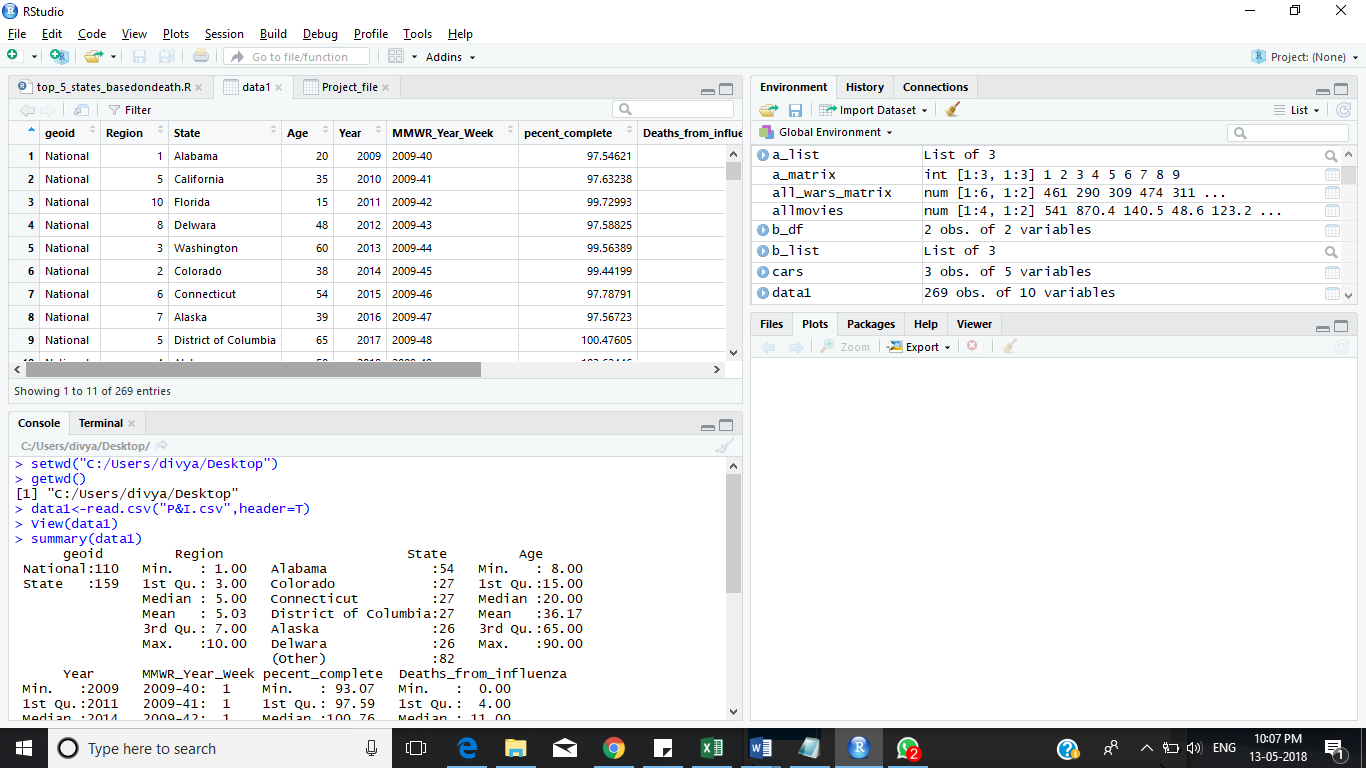
[1] 5029.227

> hist(Age)

> colors<-c("red","yellow","green","violet","orange","blue","pink","cyan","firebrick","deeppink","lightpink4","mediumpurple4","yellow4")

> hist(Age, main = "Death rate for Influenza and Pneumonia", xlab = 'Age', ylab ='All Deaths', col = colors)

**Statistical Summary**



> setwd("C:/Users/divya/Desktop")

> getwd()

[1] "C:/Users/divya/Desktop"

> data1<-read.csv("P&I.csv",header=T)

> View(data1)

> summary(data1)

geoid Region State Age

National:110 Min. : 1.00 Alabama :54 Min. : 8.00

State :159 1st Qu.: 3.00 Colorado :27 1st Qu.:15.00

Median : 5.00 Connecticut :27 Median :20.00

Mean : 5.03 District of Columbia:27 Mean :36.17

3rd Qu.: 7.00 Alaska :26 3rd Qu.:65.00

Max. :10.00 Delwara :26 Max. :90.00

(Other) :82

Year MMWR\_Year\_Week pecent\_complete Deaths\_from\_influenza

Min. :2009 2009-40: 1 Min. : 93.07 Min. : 0.00

1st Qu.:2011 2009-41: 1 1st Qu.: 97.59 1st Qu.: 4.00

Median :2014 2009-42: 1 Median :100.76 Median : 11.00

Mean :2014 2009-43: 1 Mean :102.19 Mean : 54.67

3rd Qu.:2016 2009-44: 1 3rd Qu.:106.10 3rd Qu.: 53.00

Max. :2018 2009-45: 1 Max. :127.78 Max. :653.00

(Other):263

Deaths\_from\_pneumonia All\_Deaths

Min. :2825 Min. :44281

1st Qu.:3142 1st Qu.:46432

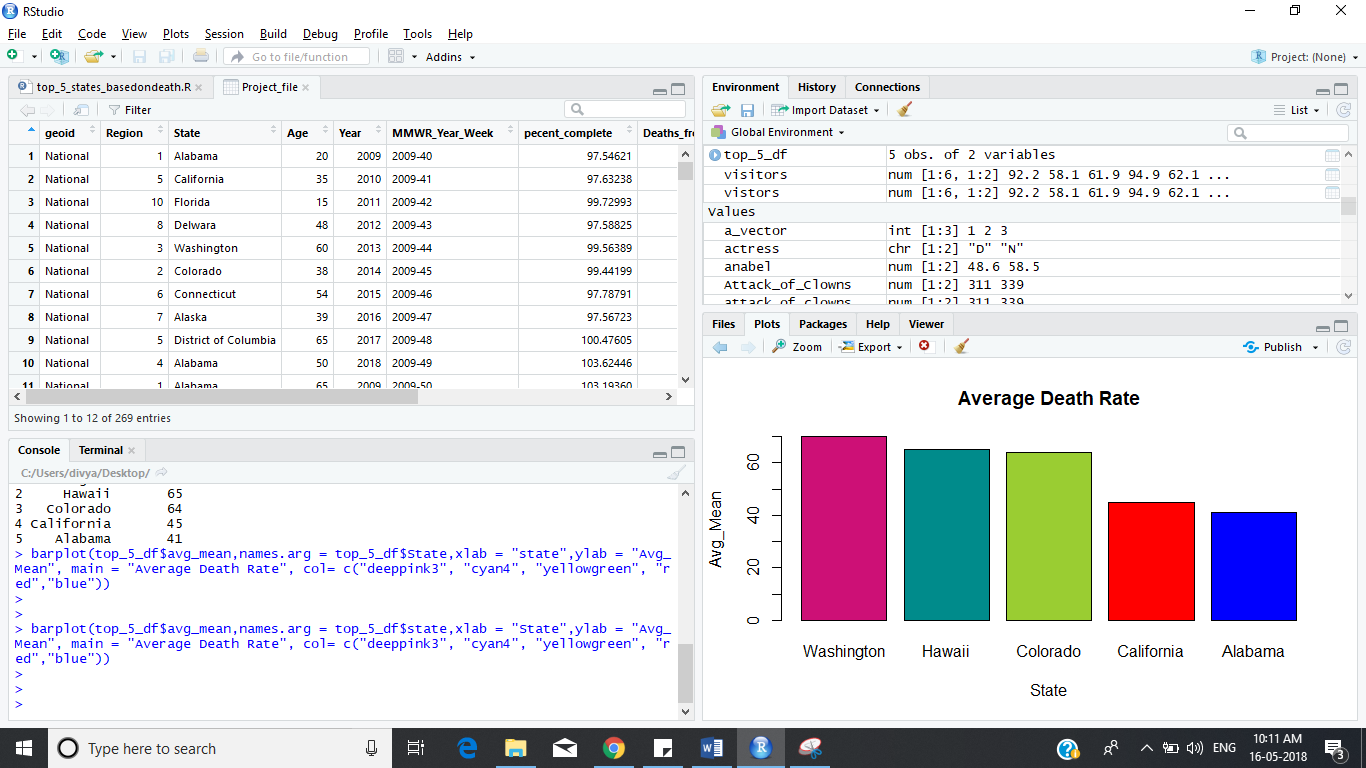
Median :3493 Median :47939

Mean :3634 Mean :48621

3rd Qu.:3986 3rd Qu.:50484

Max. :6065 Max. :60795

**Function**



**Code for script**

top\_5\_states<-function(Project\_file\_df,topn) {

library(dplyr)

top\_5\_states\_1<-Project\_file\_df %>%

group\_by(State) %>%

summarize(avg\_mean=mean(Age)) %>%

arrange(desc(avg\_mean)) %>% head(topn) %>%

transform(avg\_mean = as.integer(avg\_mean))

return(top\_5\_states\_1)

}

**Output**

> Project\_file<-read.csv("P&I.csv",header=T)

> Project\_file\_df<-as.data.frame(Project\_file)

> View(Project\_file)

> source("top\_5\_states\_basedondeath.R")

> top\_5\_states(Project\_file\_df,5)

> state<-c("Washington","Hawaii","Colorado","California","Alabama")

> avg\_mean<-c(70,65,64,45,41)

> top\_5\_df<-data.frame(state,avg\_mean)

> top\_5\_df

state avg\_mean

1 Washington 70

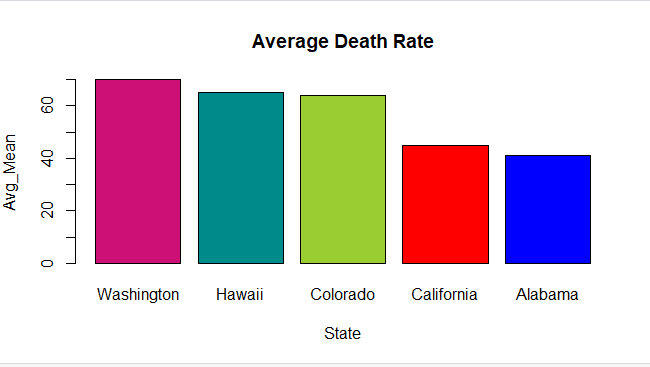
2 Hawaii 65

3 Colorado 64

4 California 45

5 Alabama 41

> barplot(top\_5\_df$avg\_mean,names.arg = top\_5\_df$state,xlab = "State",ylab = "Avg\_Mean", main = "Average Death Rate", col= c("deeppink3", "cyan4", "yellowgreen", "red","blue"))



Above code is for creating a function to find All Deaths in top 5 states by using barplot. With

the above analysis, we have tried to describe the death rate in 5 states by taking average mean of all Ages. From this, we can depict that the Washington state has the highest number of

average death rate from rest of the other states and Alabama has the lowest average death rate.

**Mean & Standard Deviation**

> attach(Project\_file)

> mean(Age)

[1] 36.16729

> sd(Age)

[1] 26.09808

**References**

1. “What Is the Connection Between Influenza and Pneumonia?” *American Lung Association*, [www.lung.org/lung-health-and-diseases/lung-disease-lookup/pneumonia/what-is-the-connection.html](http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/pneumonia/what-is-the-connection.html).

2. *Deaths from Pneumonia and Influenzas and All Deaths, by State and Region, National Center For Health Statistics Mortality Surveillance System | HealthData.gov*, [www.healthdata.gov/dataset/deaths-pnuemonia-and-influneza-and-all-deaths-state-and-region-national-center-health](http://www.healthdata.gov/dataset/deaths-pnuemonia-and-influneza-and-all-deaths-state-and-region-national-center-health).

3. HHS Office of the Secretary, Office of Intergovernmental and External Affairs. “Regional Offices.” *HHS.gov*, US Department of Health and Human Services, 15 Apr. 2014, [www.hhs.gov/about/agencies/iea/regional-offices/index.html](http://www.hhs.gov/about/agencies/iea/regional-offices/index.html).

4. “Deaths from Pnuemonia and Influneza and All Deaths, by State and Region, National Center For Health Statistics Mortality Surveillance System | Data.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, data.cdc.gov/NNDSS/Deaths-from-pnuemonia-and-influneza-and-all-deaths/pp7x-dyj2.

5. Shrestha, Sourya, et al. “The Role of Influenza in the Epidemiology of Pneumonia.” *Nature News*, Nature Publishing Group, 21 Oct. 2015, [www.nature.com/articles/srep15314](http://www.nature.com/articles/srep15314).