# Capstone Project 1 – Data Wrangling

Describe the data wrangling steps that you undertook to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them? This document will eventually become part of your milestone report.

**Step 1**: Imported CSV file as a Pandas DataFrame.

**Step 2**: Printing the DataFrame, I can see there are 2005 rows and 13 columns. That means there are 2005 firefighter fatalities recorded over 15 years.

**Step 3**: Converted the ‘Date Of Death’ column to a datetime object. Dateime is a combination of time and date. First, I imported the date time modules:

from datetime import date

from datetime import time

from datetime import datetime

Secondly, I converted the ‘Date of Death’ column to datetime using pd.to\_datetime:

df['Date of Death'] = pd.to\_datetime(df['Date of Death'])

Lastly, I extracted the year from the date column and stored it in another column, ‘Year’.

df['Year'] = df['Date of Death'].dt.year

The purpose for creating a new column for just the year is to compare causes of death by year. This will also come in handy when analyzing the number of heart attack fatalities each year.

**Step 4**: Used the column ‘Year’ to count number of deaths by year. The year 2001, had a large number compared to the rest and this is most likely due to the 9-11 attacks. When I plotted the number of fatalities each year, I noted that 2001 is a rare case.

**Step 5**: Checked for missing values by using df.info() to get a count of entries for each column. This gave me useful information about my dataset. I found out that there are 359 ages missing. 18% of the firefighters fatalities did not have an age recorded. This will not make a difference to the count of deaths per year, but will make a slight difference when comparing deaths by age. I decided not to drop the rows with missing age because this will ignore 359 deaths that occurred throughout the 15 years. Another interesting find from my dataset is that ten incidents do not have a cause of death. I replaced these empty values with ‘unknown’. This will be a category when I create a chart showing the percentage of each cause of death.

**Step 6**: The remainder of my data wrangling steps involve manipulating the DataFrame. For example, I filtered the DataFrame for only Career and Volunteer fatalities. I then take the Career and Volunteer DataFrame to filter by ‘Cause Of Death’ and ‘Year’. This will allow me to look at the causes of death for Career and Volunteer Firefighters between 2000 – 2015 and see if there is a consistent trend on a type of cause each year.