Homework 5 : Analyze Celebrity Deaths

(Deadline as per Coursera)

This homework deals with the following topics:

- The *pandas* module
- Loading data
- Joining data
- Querying data
- Summarizing data
- Aggregate functions
- The *numpy* library
- The *matplotlib* library
- Data visualization

General Idea of the Assignment

In this assignment, you will analyze data from the file "celebrity_deaths_2016.xlsx" which contains records of deaths of famous humans and non-humans in 2016. You'll use functions from the *pandas* module for loading, inspecting and querying data. You are expected to summarize data, create pivot tables and apply aggregate functions, and to visualize data using histograms and other kinds of plots.

For each question, there are clear instructions in each cell. Follow those instructions and write the code after each block of:

```
# YOUR CODE HERE
raise NotImplementedError()
```

Make sure to delete the line raising an error! Please use the exact variable name if it is specified in the comment.

We'll run a Python test script against your program to test whether each function implementation is correct.

About the Data

All of the data is contained within the “celebrity_deaths_2016.xlsx” file which contains 2 sheets:

- "celeb_death": contains records of deaths of famous humans and non-humans
    - There are 5 columns: date_of_death, name, age, bio, cause_id

	date of death	name	age	bio	cause_id
2	2016-01-01	Tony Lane	71	American art director (Rolling Stone)	8915
3	2016-01-01	Gilbert Kaplan	74	American conductor and businessman	2860
4	2016-01-01	Brian Johns	79	Australian company director, managing director of the Australian Broadcasting C	2860
5	2016-01-01	Natasha Aguilar	45	Costa Rican swimmer, silver and bronze medalist at the 1987 Pan American Gan	33897
6	2016-01-01	Fazu Aliyeva	83	Russian Avar poet and journalist	10648
7	2016-01-01	Mike Oxley	71	American politician, member of the United States House of Representatives fror	7674

- "cause_of_death": contains the causes of the deaths
    - There are 2 columns: cause_id, cause_of_death

	cause_id	cause_of_death
2	753	ALS
3	1039	bomb
4	1120	shot
5	1499	fall
6	1629	shot
7	2132	gored
8	2151	tased

During this exercise, you’ll need to merge the “celeb_death” data with the “cause_of_death” data using the “cause_id” column. This will give you the cause for each death.

Other information about the dataset:

- The cause of death was not reported for all individuals
- The dataset might include deaths that took place in other years (you'll need to ignore these records)
- The dataset might contain duplicate records (you'll need to remove them)

Submission

To complete the assignment, download *celebrity_deaths_2016.ipynb* and *celebrity_deaths_2016.xlsx*.

Evaluation

Two points for each question.