E194 - Institute of Information Systems Engineering

188.475 Digital Preservation - Exercises

FAIR Data Science – Part 1

Presented to: Dr. Tomasz Miksa

Student name : Obeida Rimawi 11742797

Aysar Ishtaiwi 11742796

Tue, Mar 28th 2017

Part 1.1 – data science use case:

Experiment overview:

Our experiment is about the relation between Number of Unemployment people in New York and Number of DNA Structures Released by protein data bank, we took a data sets from this two different open data repository:

- Unemployment data from the U.S. Government's repository (https://www.data.gov)

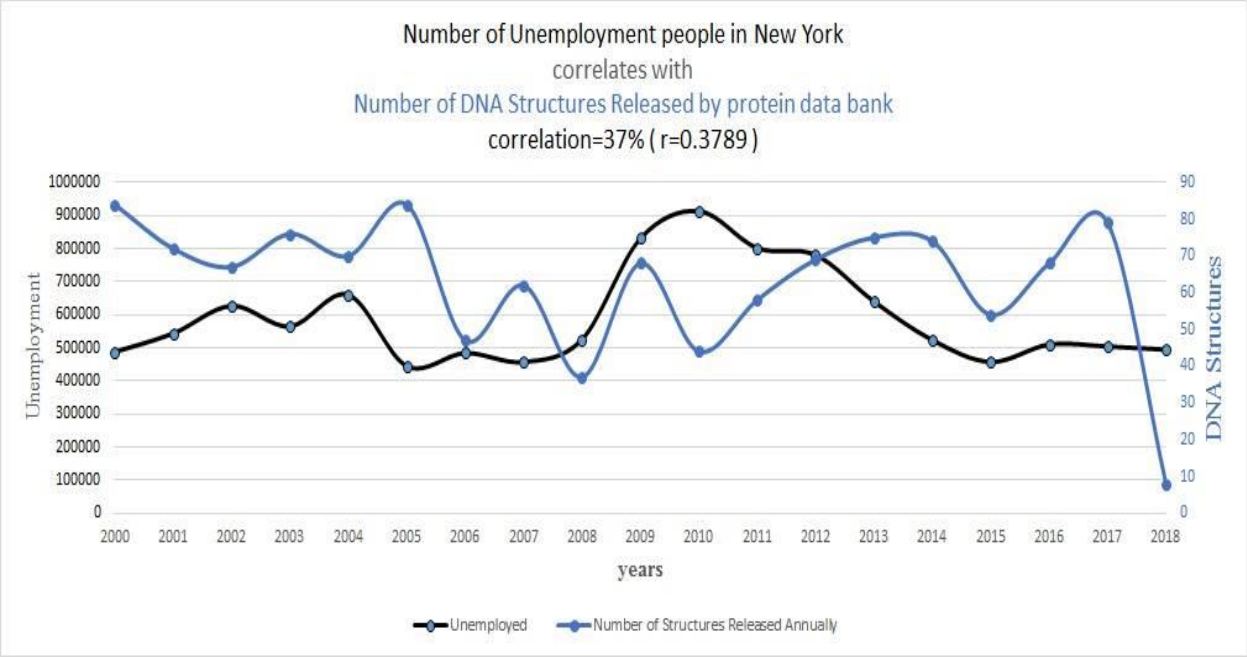
- DNA structures from Protein Data Bank (https://www.rcsb.org)

we observed from this experiment there is no similarity between the results in the same year because as we know that the Number of Unemployment people is quite large in most countries, and DNA Structures isn't the easy thing and it costs a lot of money.

Input:

YEAR	NUMBER OF STRUCTURES RELEASED ANNUALLY	UNEMPLOYED
2000	84	487700
2001	72	543400
2002	67	626800
2003	76	564800
2004	70	660000
2005	84	444900
2006	47	484300
2007	62	457200
2008	37	525400
2009	68	830500
2010	44	913400
2011	58	801700
2012	69	780500
2013	75	640200
2014	74	523100
2015	54	457400
2016	68	510100
2017	79	504100
2018	8	494500

Output:



UML diagram: