# MSAN 502 - Homework 3

*Andre Guimaraes Duarte*

*August 3rd, 2016*

## Python problem

In this problem, we will use `pandas` and `numpy` to find the best fit line in the sense of least squares to a set of data consisting of paired observations in the form $(x; y)$. The code for my implementation can be found in files `TVlife.py`, `population.py`, and `nba.py`. In addition, I found a data set online concerning XXX that I analyzed using the same system as these files. Here, I will explain how I proceeded, and show the results and graphs.

### population.txt

This file contains information concerning the national population $(y)$ as a function of the year $(x)$. Plotting $y$ against $x$, we get the graph seen in figure 1. We can see that a linear regression seems to be a likely candidate for regression.

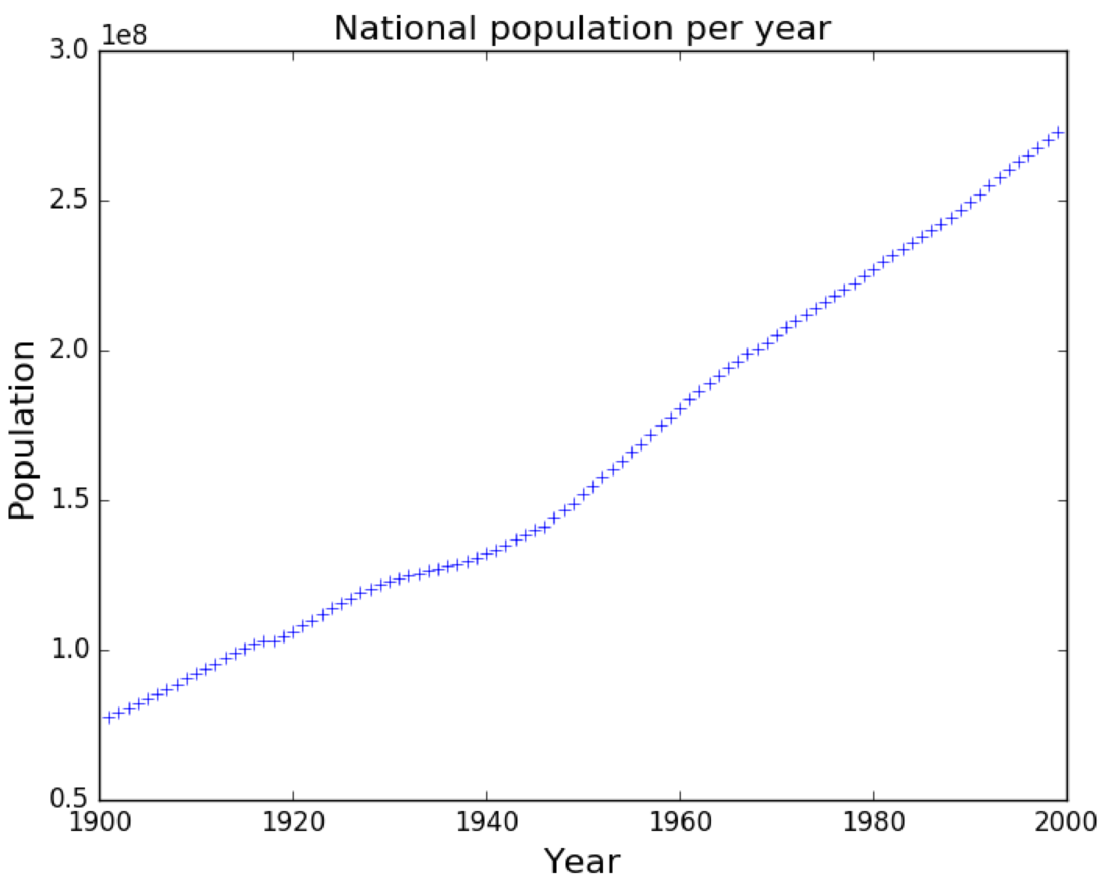


Figure 1: National US population from 1900 to 2000

With `numpy`, we can easily obtain $\widehat{\mathbf{x}}$ that minimizes the error. We just need to compute: $\widehat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$.

We get $\widehat{\mathbf{x}} = \begin{bmatrix} -3.741 \cdot 10^9 \\ 2.003 \cdot 10^6 \end{bmatrix}$.

If we draw this line on top of the data, we get the image in figure 2.

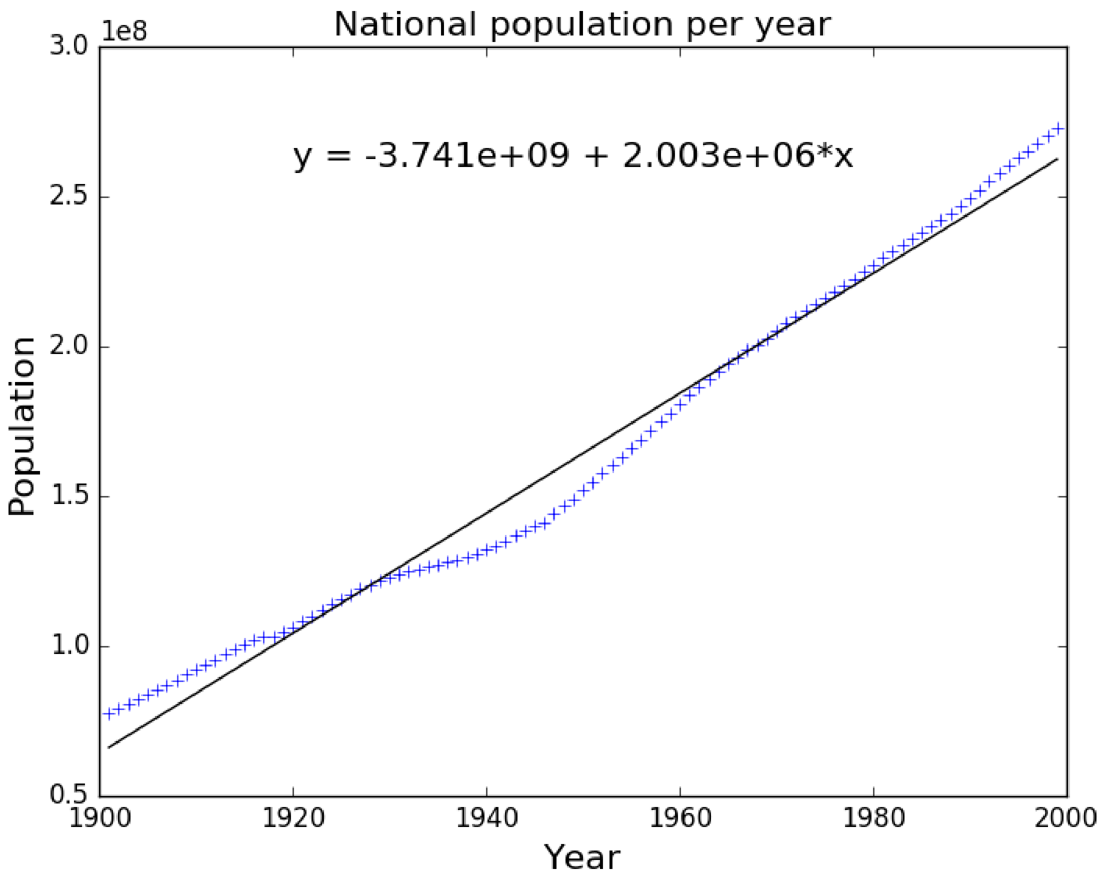


Figure 2: National US population from 1900 to 2000 and regression line

The fit that we computed seems to be what we would expect.

### nba.txt

This file contains information concerning team winning percentage in basketball games ($y$) as a function of PM (the average point difference over all that team's games) ($x$). Plotting $y$ against $x$, we get the graph seen in figure 3. We can see that a linear regression seems to be a likely candidate for regression.

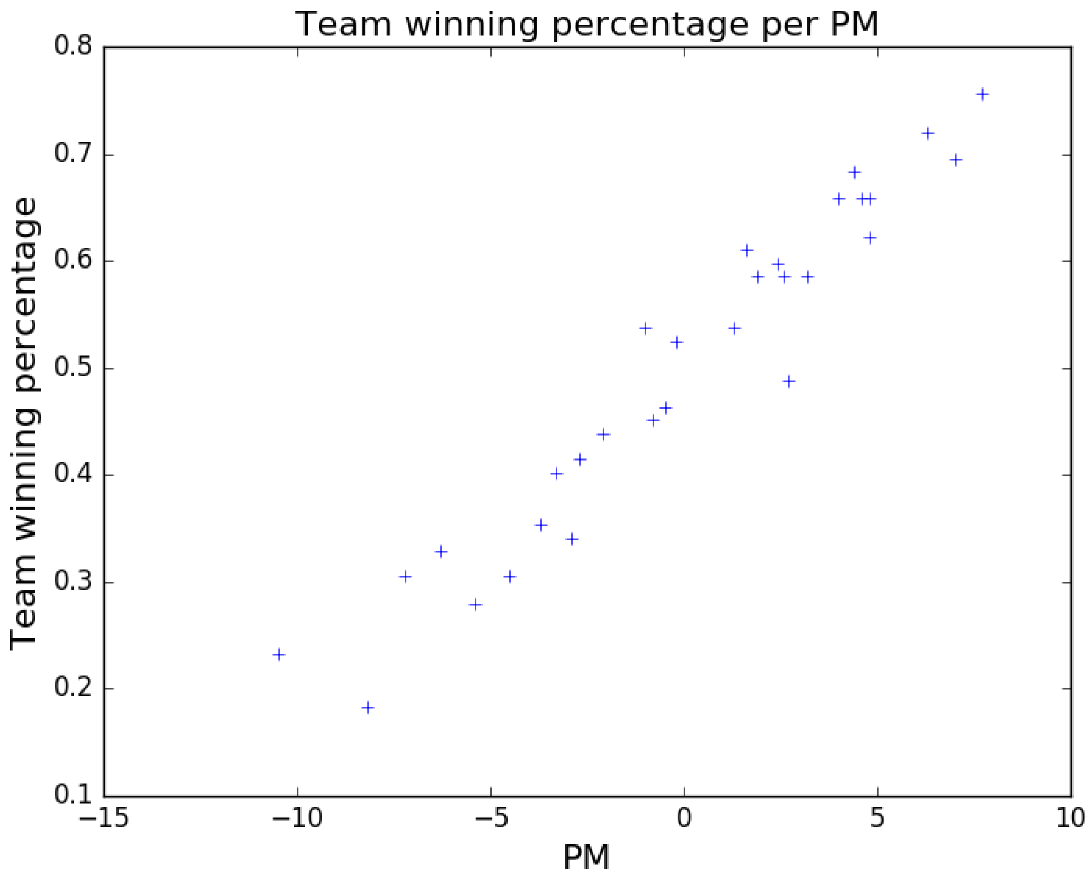


Figure 3: Team winning percentage as a function of PM

With `numpy`, we can easily obtain $\widehat{\mathbf{x}}$ that minimizes the error. We just need to compute:

$\widehat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$.

We get $\widehat{\mathbf{x}} = \left[\begin{smallmatrix} 0.500 \\ 0.032 \end{smallmatrix}\right]$.

If we draw this line on top of the data, we get the image in figure 4.

The fit that we computed seems to be what we would expect.

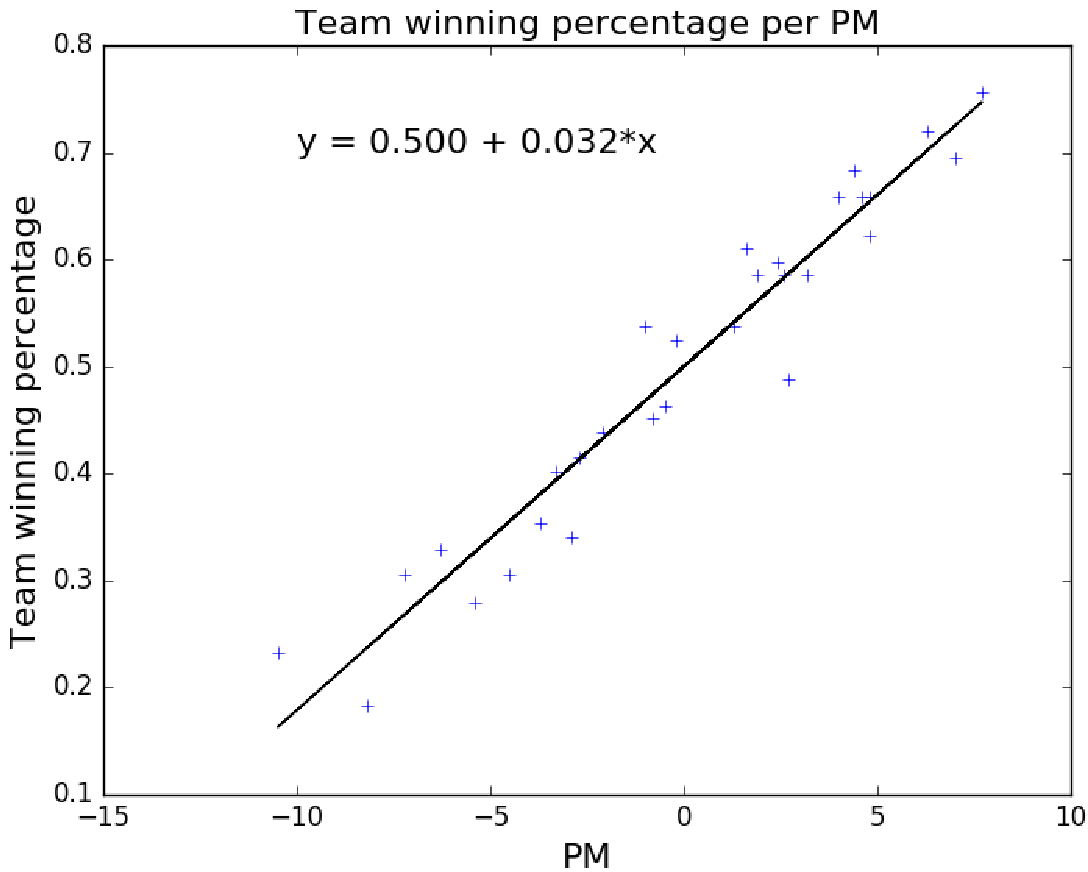


Figure 4: Team winning percentage as a function of PM and regression line

## TVlife.txt

This file contains information concerning life expectancy ($y$) as a function of televisions per thousand people ($x$). Plotting $y$ against $x$, we get the graph seen in figure 5. In this case, linear regression does not seem like the best candidate for regression. Maybe a polynomial regression would work better in this case. But we will proceed with linear regression for this problem.

With `numpy`, we can easily obtain $\widehat{\mathbf{x}}$ that minimizes the error. We just need to compute:

$\widehat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$.

We get $\widehat{\mathbf{x}} = \left[ \begin{smallmatrix} 57.337 \\ 0.032 \end{smallmatrix} \right]$.

If we draw this line on top of the data, we get the image in figure 6.

The linear fit seems adequate, but a polynomial regression would produce a better result in this particular case.

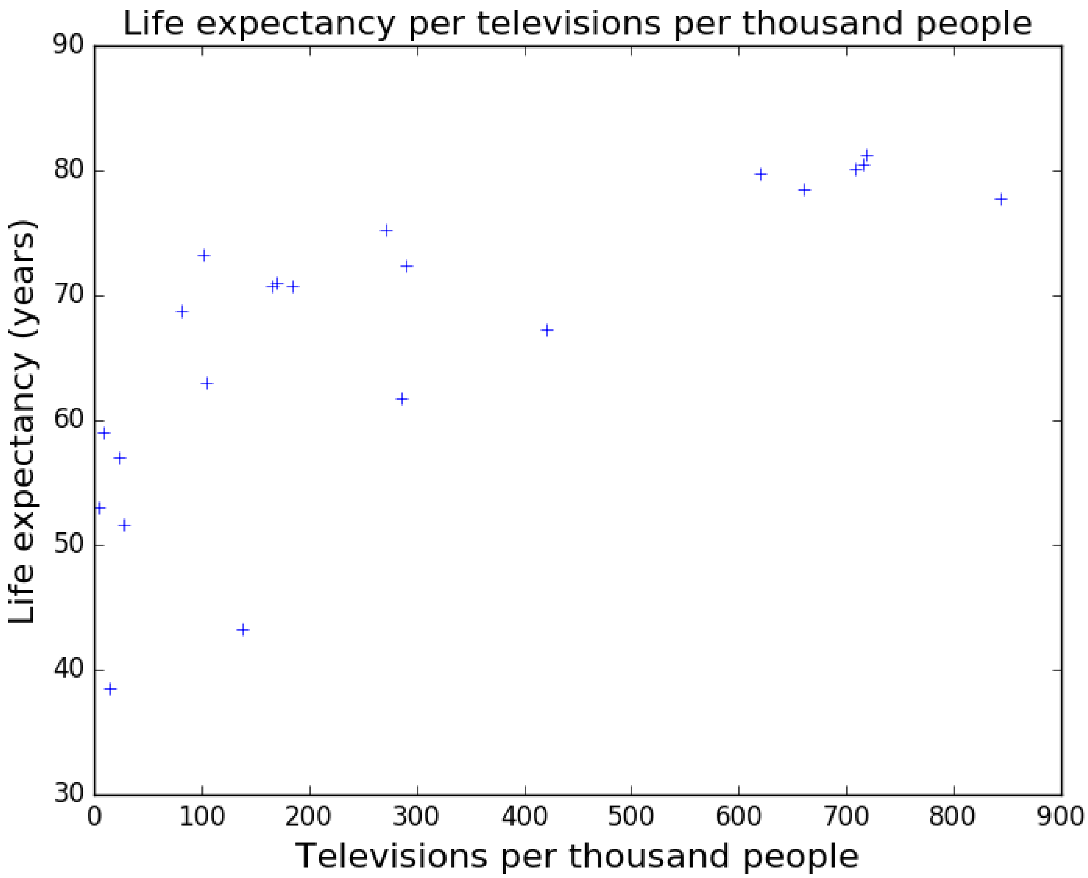


Figure 5: Life expectancy per televisions per thousand people

### 30oysters.txt

I got this file from the *Journal of Statistics Education*, via the website http://www.amstat.org/publications/jse/jse_data_archive.htm. The file consists of 30 observations of 5 variables concerning oysters that was collected in 2001. The direct link to the file is: http://www.amstat.org/publications/jse/datasets/30oysters.dat.txt.

This file contains information concerning 30 oysters' volume in cc ($y$) as a function of their weight in grams ($x$). Plotting $y$ against $x$, we get the graph seen in figure 7. We can see that a linear regression seems to be a likely candidate for regression. The data seems very linearly correlated.

With `numpy`, we can easily obtain $\widehat{\mathbf{x}}$ that minimizes the error. We just need to compute:

$\widehat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$.

We get $\widehat{\mathbf{x}} = \left[\begin{smallmatrix} 0.714 \\ 0.955 \end{smallmatrix}\right]$.

If we draw this line on top of the data, we get the image in figure 8.

The fit that we computed seems to be what we would expect.

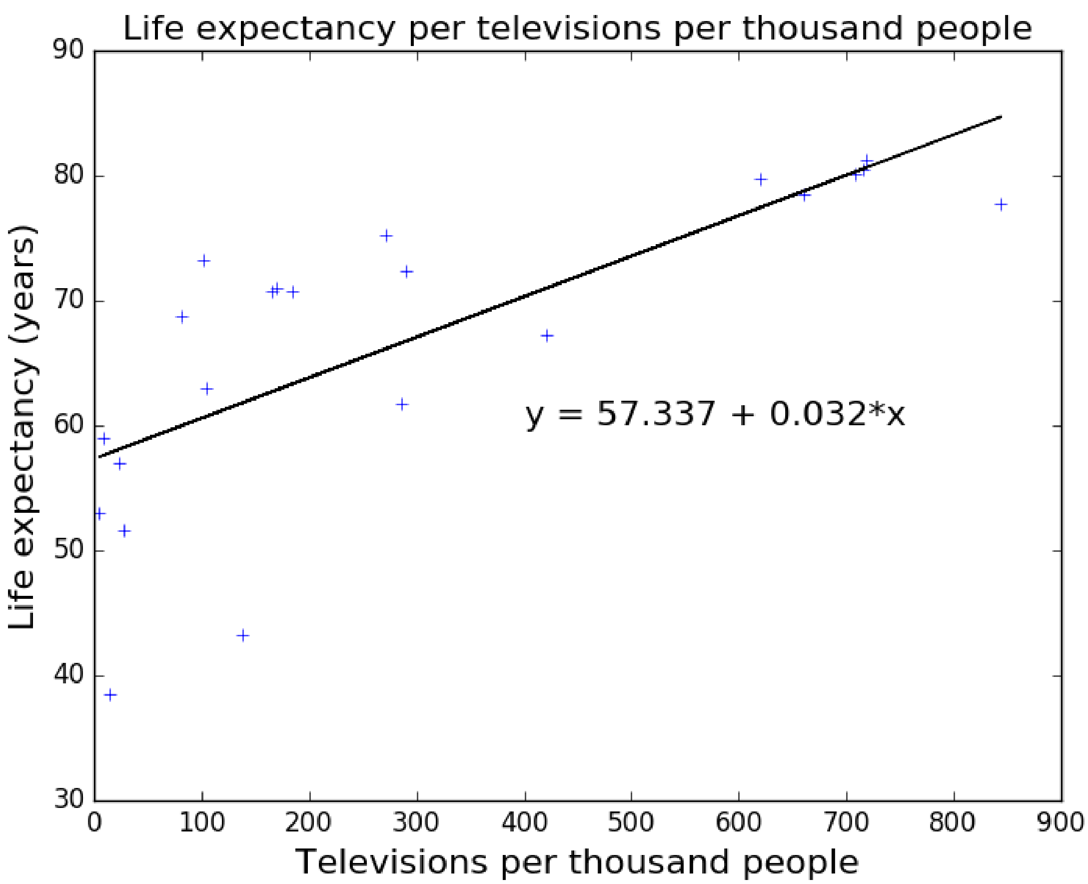


Figure 6: Life expectancy per televisions per thousand people and regression line

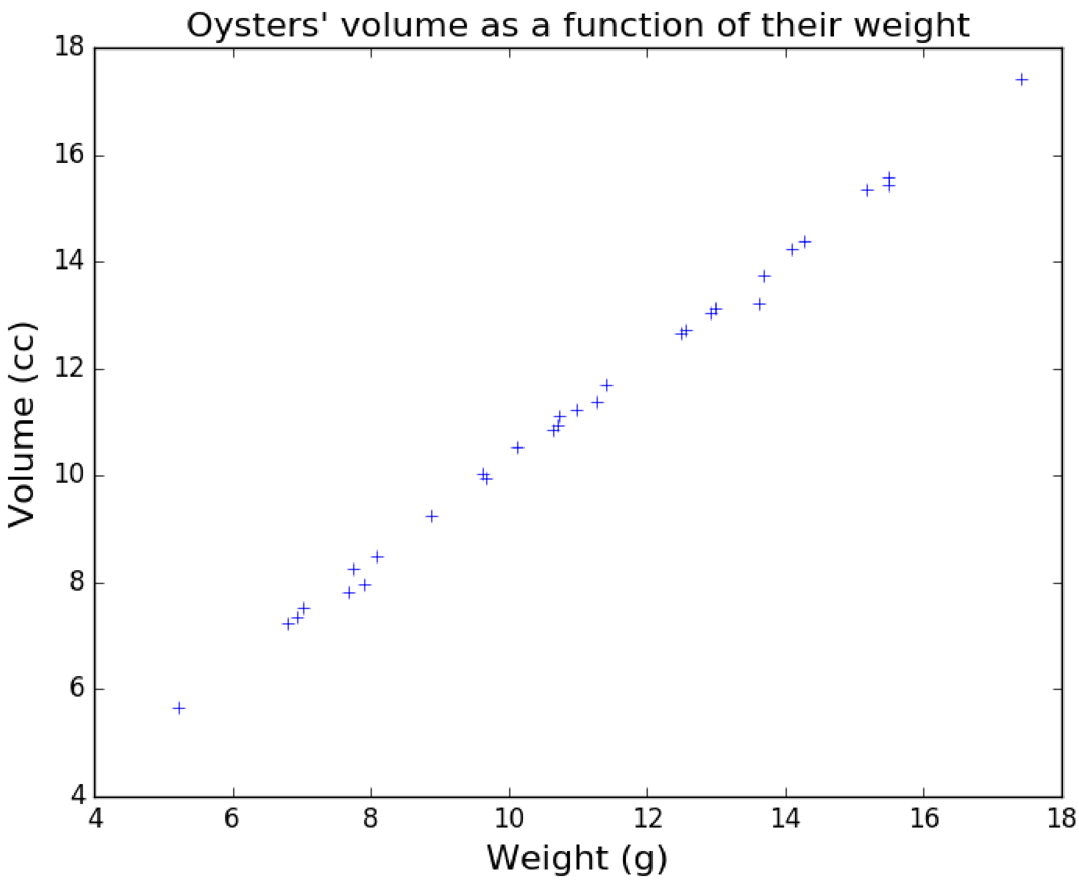


Figure 7: Oysters' volume as a function of their weight

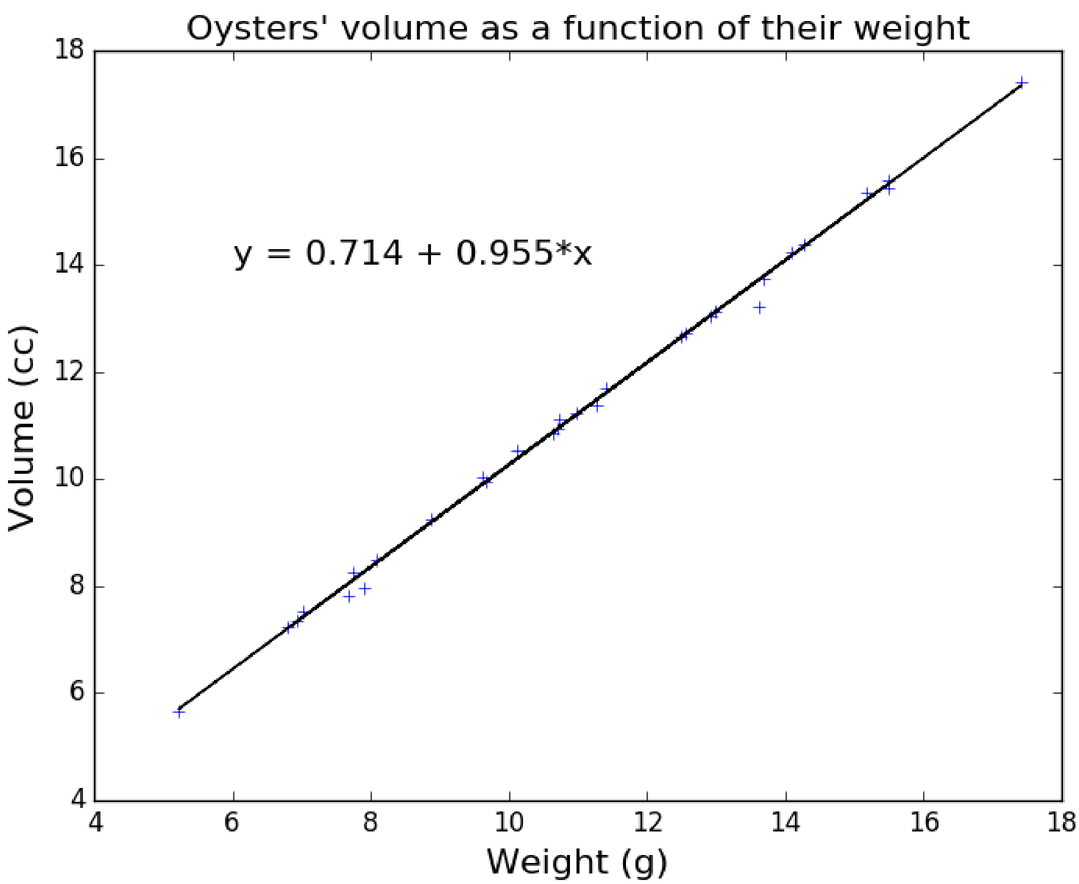


Figure 8: Oyster's volume as a function of their weight and regression line