

Machine Learning - Lab0

Santiago Álvarez Sepúlveda

October 1 2018

1 Introduction to the Research Environment

This lecture introduces the research environment, which is based on iPython Jupyter Notebooks, that allows to perform data analysis and statistical validation. The first objects described in this lecture belong to iPython Jupyter Notebook's main elements, commands and shortcuts, i.e. code cells, text cells, command executions, how to understand the GUI, using Python language, Python tools for coding and documentation, usage of libraries as Numpy, Pandas and Matplotlib as a formula to work simultaneously with database handling, data processing and data visualization. Of course some Quantopian aspects take part in this lecture, the examples that involve the usage of Python programming language use data of the prices and the returns of the stock "MSFT" (Microsoft) as a function of time, then with this information we estimated some statistical data of the prices like mean and standard deviation and moving averages, which is an average of a window with a given size that is applied to a signal as a low pass filter.

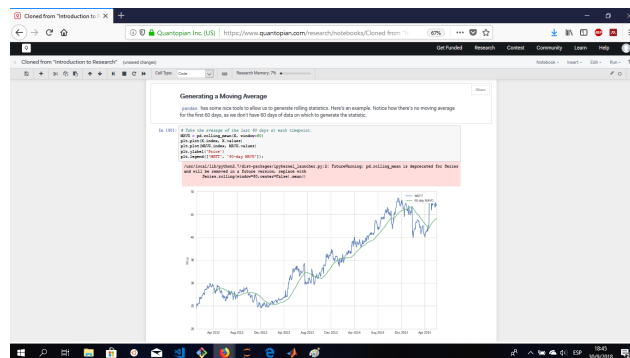
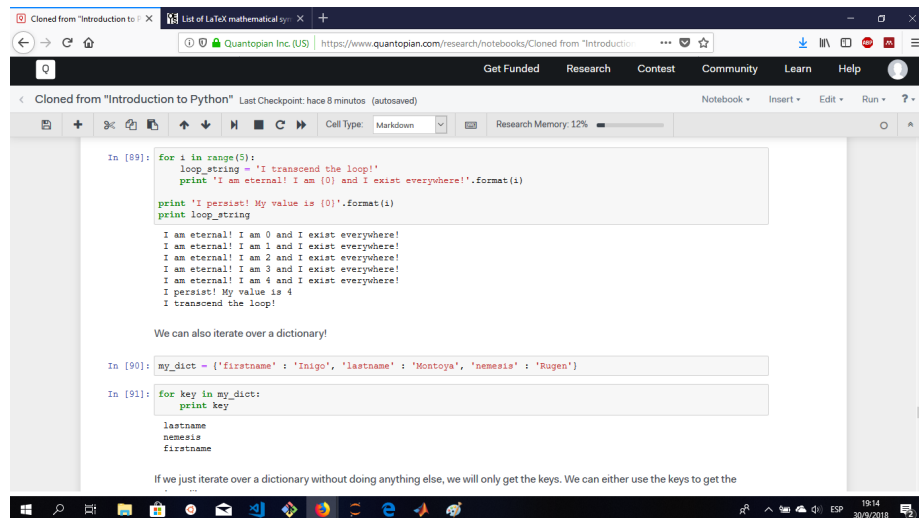


Figure 1: Screenshot of Lecture 1 Cloned to Quantopian.

2 Introduction to Python

Since everything in Quantopian is programmed using Python, this is a very important lecture to pay attention. As many other introductions to Python programming language, this shows the main syntax and the semantics behind every basic command that Python understands. Some of this basic commands are the single-line and multi-line comments, how to define and print a variable, which in Python makes reference to an object (Python is an object oriented language), the usage of different native data types such as integers, floats, strings, Boolean, some mathematical operations like addition, subtractions, multiplication, division, modulo, floating point operations, and the most notable of the built-in math functions that Python has. Also here is presented the math library that defines multiple irrational numbers and functions to make more extensive mathematical processing of data with Python. This lecture also shows some examples involved with initializing and handling lists, tuples, sets and dictionaries, which are data structures that allows us to organize information very intuitively and which contain a lot of methods in order to process and access to this information. Among many other things, it is important to mention the correct syntax of If-Else statements and loop cycles, also the definition of functions and the way they are called and how to handle the values this return.



The screenshot shows a web browser window with a Quantopian notebook. The browser address bar shows the URL <https://www.quantopian.com/research/notebooks/Cloned from 'Introduction to Python'>. The notebook interface includes a toolbar with icons for file operations and a 'Cell Type' dropdown set to 'Markdown'. The main content area displays Python code in a light blue box, followed by its output in a white box. The code includes a for loop that prints a message 5 times, a dictionary iteration, and a comment about iterating over a dictionary without doing anything else.

```
In [89]: for i in range(5):
         loop_string = 'I transcend the loop!'
         print 'I am eternal! I am {} and I exist everywhere!'.format(i)

         print 'I persist! My value is {}'.format(i)
         print loop_string

I am eternal! I am 0 and I exist everywhere!
I am eternal! I am 1 and I exist everywhere!
I am eternal! I am 2 and I exist everywhere!
I am eternal! I am 3 and I exist everywhere!
I am eternal! I am 4 and I exist everywhere!
I persist! My value is 4
I transcend the loop!

We can also iterate over a dictionary!

In [90]: my_dict = {'firstname': 'Inigo', 'lastname': 'Montoya', 'nemesis': 'Bagen'}

In [91]: for key in my_dict:
         print key

lastname
nemesis
firstname

If we just iterate over a dictionary without doing anything else, we will only get the keys. We can either use the keys to get the
```

Figure 2: Screen shot of Lecture 2 Cloned to Quantopian.

3 Introduction to Numpy

Numpy adds a support to multidimensional arrays and mathematical functions that allows us to easily perform linear algebra calculations, for this reason this lecture is a collection of linear algebra examples computed using Numpy.

To start it is necessary to define a Numpy array, it's main methods and attributes associated (like shape and splitting), also the mathematical functions that can handle this multidimensional arrays, this way it is possible to work with the information of a portfolio, in an ordered way, and applying the mathematical axioms of the vector spaces and linear algebra.

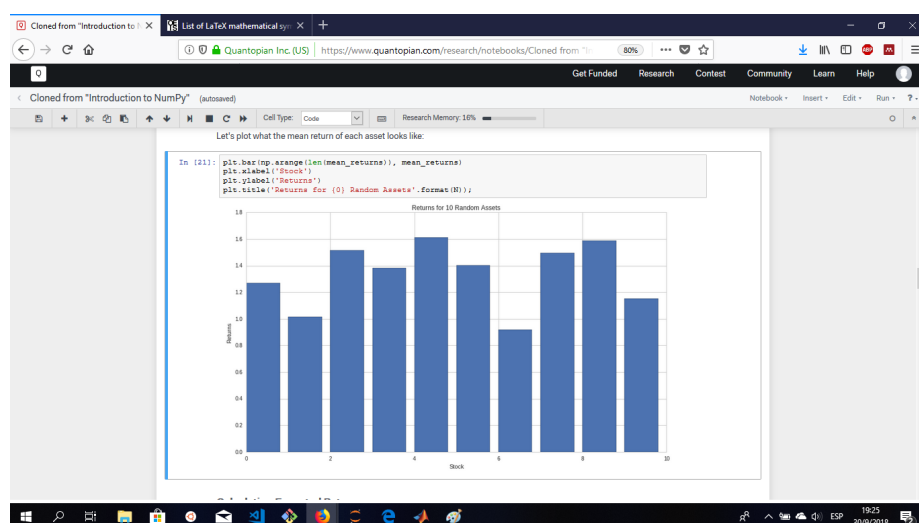


Figure 3: Screen shot of Lecture 3 Cloned to Quantopian.

4 Introduction to Pandas

This lecture introduces the Pandas library, which is a Python collection of powerful data structures to make easy the handling of data, mainly the Series (1-Dimensional arrays with labels that can contain any data type) and DataFrame objects where used through this lecture. Another important fact about this library is that it has a very close relation with Numpy, so it is possible to do mathematical and logical processing of this data structures, very helpful for finances.

For each of the data structures used from Pandas (Series and DataFrames), we see some examples of how to define variables of this types, apply mathematical operations and functions to this objects, filter them with logical operations, indexing the elements, how to fill the missing values so we avoid numerical errors and finally how to generate plots from the values stored in this structures.

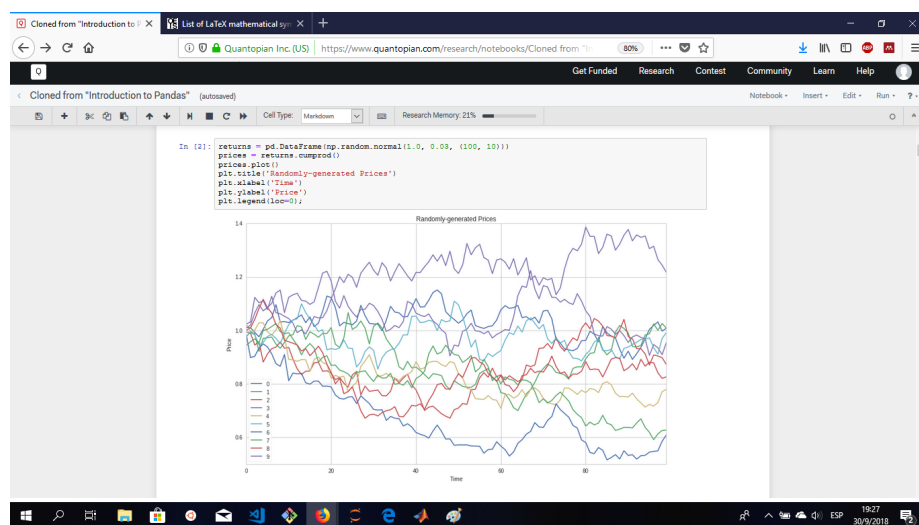


Figure 4: Screen shot of Lecture 4 Cloned to Quantopian.

5 Plotting Data

This lecture contains a lot of examples of the usage of the library Matplotlib and the way it displays graphical information of data stored in a certain multidimensional array. We can compute statistical distributions as histograms, cumulative histograms scatter plots and line graphs.

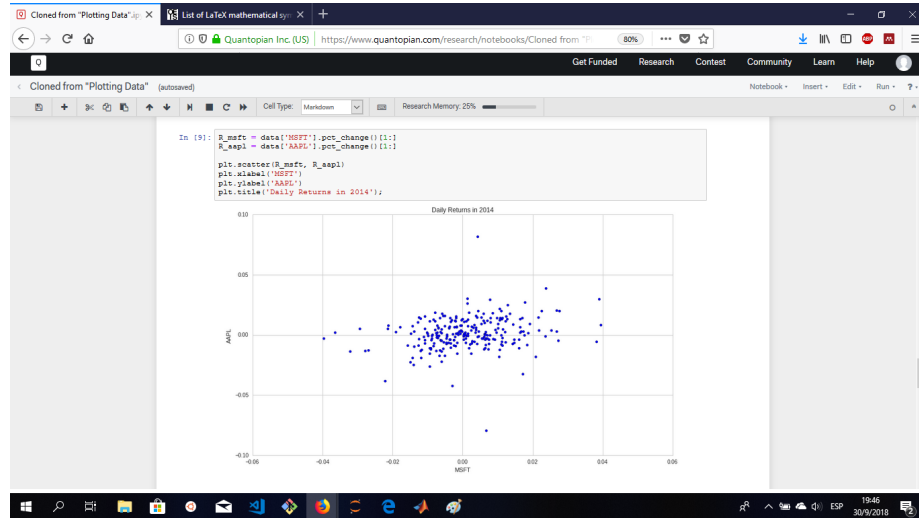


Figure 5: Screen shot of Lecture 5 Cloned to Quantopian.

6 Means

This lecture contains a way to summarize a set of data using one single parameter, this parameter captures information about the distribution of data and it's central tendency. The first parameter to do this is called the Arithmetic Mean, sometimes the arithmetical mean is also called the average of a series of numbers. Another different concept is the one of Median, which is the number that appears in the middle of a sorted array, the mode, is the most frequently repeated value in the data set, and finally the geometric mean, which is similar to the arithmetic mean, but this one is calculated from the product of the values and a square root, while the arithmetic is calculated from the sum of the values and a division.

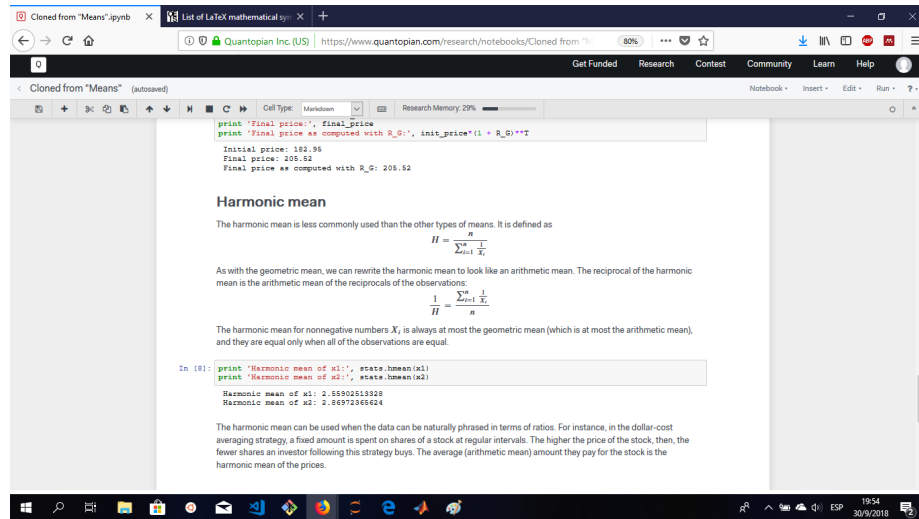


Figure 6: Screen shot of Lecture 6 Cloned to Quantopian.

7 Variance

Dispersion measures how spread out a set of data is, somehow related to a measure of the precision of a set or how close each element can be from another. Data with low dispersion is heavily clustered around the mean, while data with high dispersion indicates many very large and very small values. This is especially important in finance because one of the main ways risk is measured is in how spread out returns have been historically.

Some of the usual measures of dispersion are the Mean Absolute Deviation (MAD), Variance and Standard Deviation, Semivariance and Semideviation, among others.

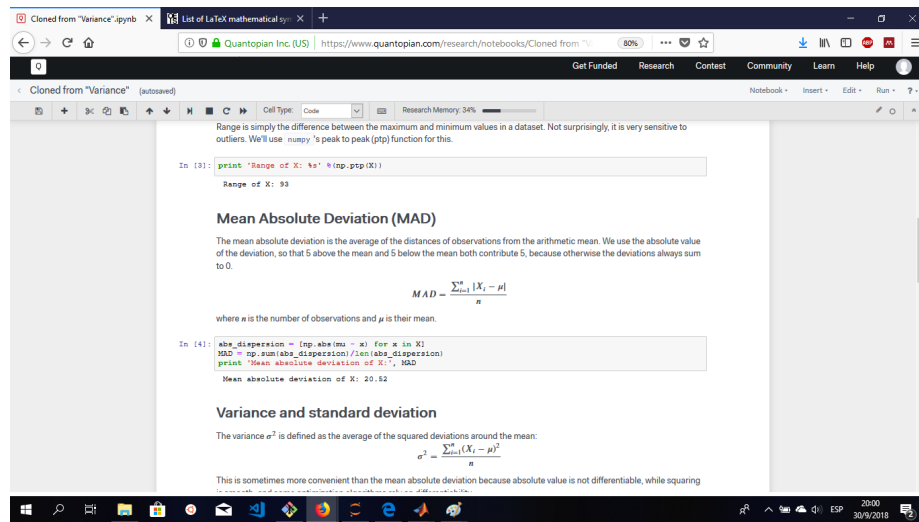


Figure 7: Screenshot of Lecture 7 Cloned to Quantopian.

8 Statistical Moments

This lecture explains the concepts of statistical moments, skewness and kurtosis. Sometimes when a distribution is not symmetric, it is called skewed, the kurtosis explains how peaked a distribution function is, or how sharp is the top of the distribution, from it's mean. This is very important in cases when distributions are not only of a normal behavior, but they tend to be very sharp or to be skewed.

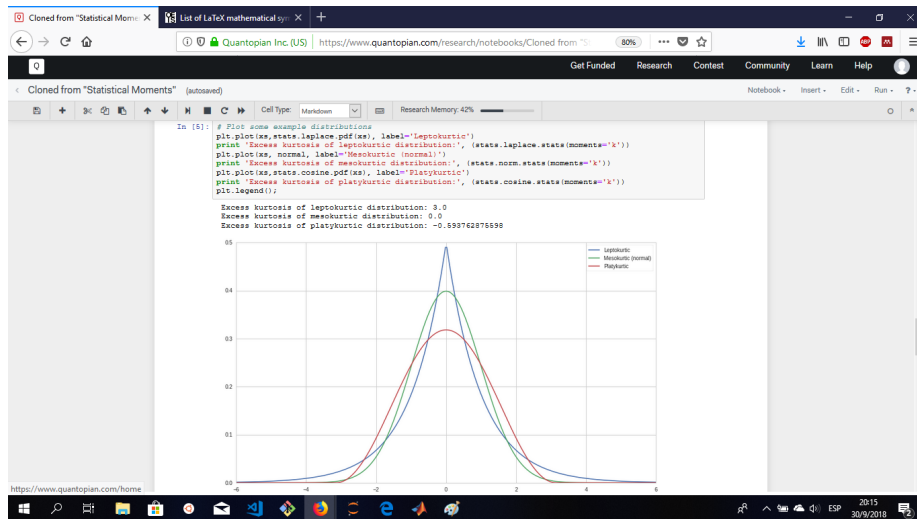


Figure 8: Screenshot of Lecture 8 Cloned to Quantopian.

9 Linear Correlation Analysis

The correlation coefficient measures the extent to which the relationship between two variables is linear or not. Its value is always between -1 and 1. A positive coefficient indicates that the variables are directly related and when one increases the other one also increases, and a negative coefficient indicates that the variables are inversely related, so that when one increases the other decreases. The closer to 0 the correlation coefficient is, the weaker the relationship between the variables, this means, they don't present variations related linearly.

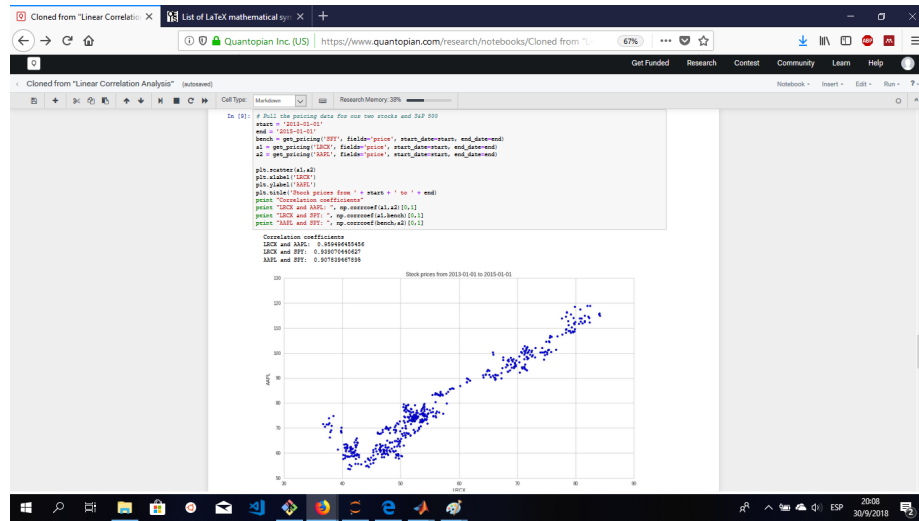


Figure 9: Screenshot of Lecture 9 Cloned to Quantopian.

10 Instability of Estimates

This lecture introduces the concept of volatility of a parameter, which has to be calculated or otherwise, we do not know whether or not we should expect new data points to be aligned with this parameter. A good way of computing volatility is dividing the data into subsets and estimating the parameter from each one, then finding the variability among the results, something similar to a divide and conquer approach. The instability analysis and testing for standard error is still very useful for telling us how much we should distrust our estimates.

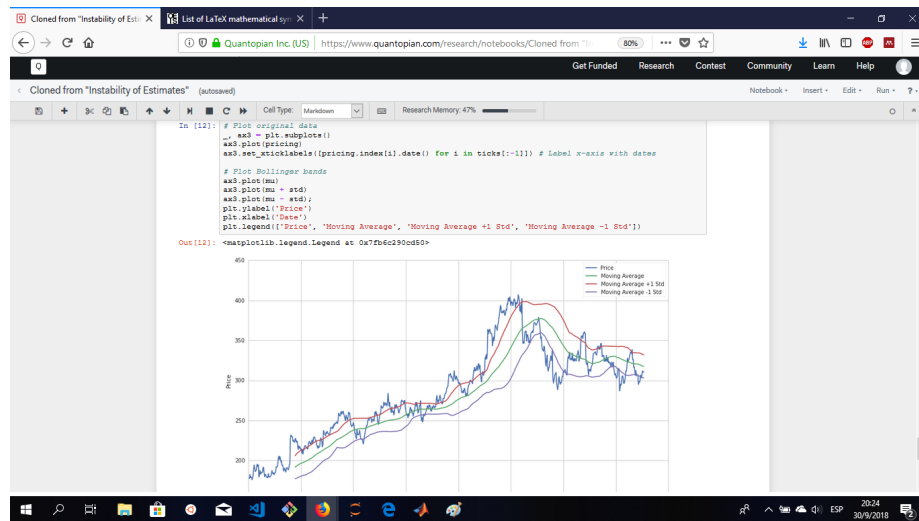


Figure 10: Screenshot of Lecture 10 Cloned to Quantopian.

11 Random Variables

This lecture introduces the concept of random variable, which is a variable that takes values according to a chance. This random variables are commonly separated as discrete random variables and continuous random variables, the main difference comes from the space of the possibilities, whether if it is countable or not. Some very famous discrete distributions are the Binomial, the uniform distribution. On the other hand a continuous random variable is one that has a non countable set of outputs, some very known examples are the normal distribution. Another important aspect is to estimate how fit is a distribution with the one of a data set.

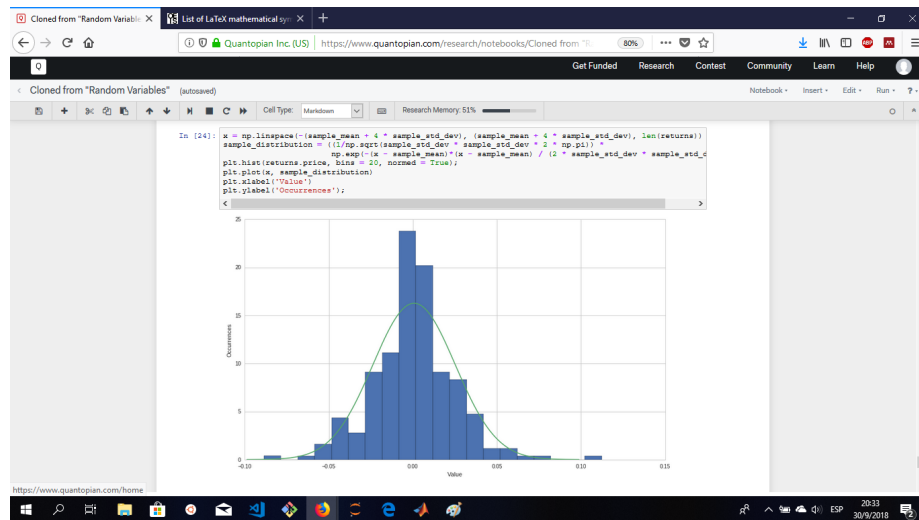


Figure 11: Screenshot of Lecture 11 Cloned to Quantopian.

12 Linear Regression

This lecture talks about linear regression, which is a technique that measures the relationship between two variables. If we have an independent variable X and a dependent outcome variable Y linear regression allows us to determine which linear model $Y = \alpha + \beta X$ best explains the data, note that this will give a line of best fit, whether or not the relationship it shows is significant is for you to determine. Python's statsmodels library has a built-in linear fit function. The output will also have some statistics about the model, such as R-squared and the F value, which may help you quantify how good the fit actually is.

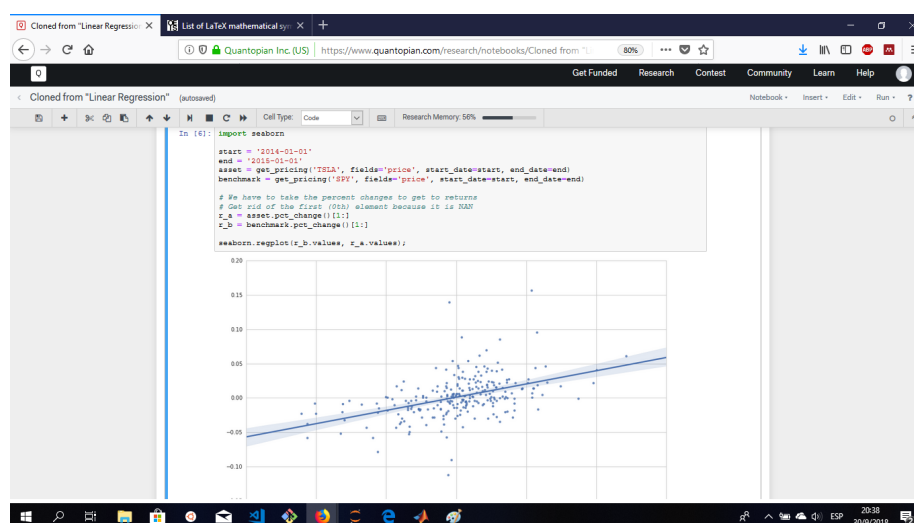


Figure 12: Screenshot of Lecture 12 Cloned to Quantopian.