# Data Engineering Test

The following exercise aims at testing the candidates’ ability to write clear and concise coding, scraping, cleaning, and manipulating data. At the end, please send a zip file with your code and this word file to [ol-ir-liquid@btgpactual.com](mailto:ol-ir-liquid@btgpactual.com). Our servers may block the zip file, in that case, you can replace the .zip extension for a .txt file.

## Scraping price data

Ibovespa is the main benchmark of the Brazilian stock market. This benchmark is crucial for asset managers to generate insight, monitor activities and act. In this section we will scrape yahoo finance for daily price and volume data for the Ibovespa benchmark. You can find Ibovespa’s quotes here: [https://finance.yahoo.com/quote/%5EBVSP?p=^BVSP&.tsrc=fin-srch](https://finance.yahoo.com/quote/%5EBVSP?p=%5eBVSP&.tsrc=fin-srch).

Write a code that scrapes Yahoo Finance for the last 5 years of daily price and volume data.

## Feature creation

Now that we have our table with the last 5 years of price ticks and daily volume, write a code that calculates the monthly log return of the benchmark and the total monthly volume.

Monthly log returns can be calculated by taking the difference between the log of the price at t+1 and the log of the price at t. Monthly volume can be calculated by simply adding all volumes of the daily series in that given month.

Once you have monthly log returns and monthly value, plot both curves in a single graph and paste it below.

**ANSWER HERE:**

## Analyzing results

We now have a broad picture of the monthly behavior of our main Brazilian stock market benchmark.

Now, for each question below, write a function that returns just a single decimal number containing the answer.

For this question, please refrain from using any ready-made library that could aid in the calculations (eg. you can use pandas to manipulate the DataFrame, but not the corr() method to calculate the correlation between the two series). If you absolutely need to use a ready-made library, please explain why it wasn’t possible to code the method by hand.

1. Which month of the year shows, on average, the highest log returns?
2. What is the correlation between monthly log return absolute value and volume?
3. Which day of the week shows, on average, the lowest log returns?
4. Which is the highest difference between average daily log return (within a month) and that same month’s log return

## Defining an architecture

We now can scrape and analyze Ibovespa’s data. The next step would be to design a production-level flow to automate the extraction, transformation, loading process, while also making this information readily available to all our analysts. Please give a brief explanation of how you would design this process, what technologies you would use and why. If possible, draw a simple diagram explaining your architecture (you may use <https://app.diagrams.net/>).

**ANSWER HERE:**