Yuri Almeida Cunha
Capstone Project Summary
14/06/2022

Webscraping and deploying financial dashboards on Heroku

Introduction

As technology evolves, the amounts of data generated exponentially increase. Research of the Statista institute forecasts that the amount of data created, captured, and copied will reach the total of 181 zettabytes by 2025. So, not only is data already a substantial part of our days, but will become even more important in the upcoming years.

As a final project for the Masters in Business Analytics at Central European University (CEU), I had the job to be a bridge, a point of contact, between a small corporation and this new and constantly updating data-world.

Using Python, Heroku, Git and some HTML, I have developed a small application that can provide this small company a better understanding of their costs based on data that already existed, was available; however, due to the lack of some technical aspects, difficult to be comprehensive enough to facilitate their decision-process.

The goal

After my initial contact with Rosen from PhaseGrowth, I have realized that although the company lacked automated and structured basic operational KPI's collections. Some important decisions related to the company's operations regarding location, cost margins, and pricing weren't easily done due the difficulty of processing the data associated to basic KPI's.

When I started arguing about what the main data problems that could be prioritized were, it was quite difficult even to define a basic strategy. This was because it took an entire week for him to collect / finalize a balance sheet for the entire period of the company exercise. Having that in mind, we had to decide that, in terms of time consumed, the most difficult aspect of that data collection were the invoices from ENVOICE:EU. Although the system helps PhaseGrowth on its accounting perspective, it doesn't offer any dashboarding or visual capabilities for the plan that was hired. With that in mind and knowing how tough and painful the process was to collect all the operational costs related to the company's operation, the final scope of the project could be defined: automate the cost operation collection and display that information in useful and interactive dashboards.

The process

There are two main parts to this project: the web scraping of the data and the washboarding of the results.

The two main important parts of content that should be scrapped are the invoice and the report sections. All the invoices and reports are registered under the Purchase Tab, and they are displayed in a table formatted way which details the most generic information related to each one of them. In addition, there is also a specific page for each one of those invoices / reports where more detailed information can be obtained. So, my final data frame would be a combination of the information displayed in the initial section, with the detailed information displayed in each page of a single invoice. For that, I have created two different functions for each part (reports and invoices).

With the invoice and report data frames ready, it is time for the data preparation. The invoice data frame is the one requiring the most attention. And although the information is quite clean, some small modifications should be applied to it. An important factor that needs to be analyzed is that the currency of the invoices is not all the same. For that, a function that uses an API connection that allows me to convert different currencies into EUR (defined by the client) was used. The end of the data preparation process consisted of merging both datasets web scrapped, the report and the invoice data frames. This allowed me to distinguish the invoices that belong to specific reports and those that do not. At the end, the final and ready to analysis dataset was converted into a .csv format to be used as the data source / database for the final application.

To be able to deploy the app, it's necessary to create a local virtual environment. It will activate the environment after that so the application can work. The libraries, pandas and plotly, helped me out to manipulate the data inside my application and turn it into interactive and publication-quality graphs. The dash library allowed me to build those analysis into production-ready data dashboards. The drawing, the borders, and patterns of each single element of my dashboard are defined inside this css file. This is the location where I can change the size, the font-color, and the positioning of all the elements presented like titles, graphs, and images. Located within app.py are the building blocks that compound the structure of the application.

I have used Heroku to deploy my web application. This was a great solution since the client didn't have a web service and there was no budget to get one. It allowed me to deploy the project with relative easy. While this means the information is public, this is an accepted risk for the first version. It is also a small risk currently with almost nobody aware of its existence and the name of the application.

Outcome

The main part of the of this project was to provide a simply descriptive analytical solution that at one point would improve the relationship of the organization with their own data - not only by having an easy and initial understanding of its own costs but doing so in an automated way that wouldn't deviate them to solve the constant number of new problems popping up the whole time.

The focus here was to develop something applicable to any client's reality focusing on simple but very useful development tools that every data analyst or data scientist should be able to apply on their daily work life.

Data preparation and simple EDA skills is not only the core but the basic fundaments of professions that have their core based on retrieving insights from data. The decision-making process should always be the main target of every single analysis despite of any fancy technique or algorithms applied.

Takeaways

Although I have faced quite unexpected risks over the course of the project, I honestly enjoyed having deployed a project like this. It was a great chance to put the skills learned over the course of this masters. I got into touch with new HTML applications and structures such as Dash, Heroku and Flask. Furthermore, I have created a fantastic relationship with my client, and I am very grateful to CEU for helping me out to boost my career like that.