ESTIMATING A CASH FLOW: THE CASE OF TERNA

Lorenzo Baglini

Edoardo Borriello

Filippo Piccaro

INTRODUCTION

For this Data Science in Action project, we chose to collaborate with Terna.

Terna - Rete Elettrica Nazionale S.p.A. is an Italian electricity transmission grid operator, listed on the FTSE MIB index of the Italian Stock Exchange. Through Terna Rete Italia, it operates Italy's national transmission grid with 74723 km of high-voltage power lines. The reason we wanted to undertake this project with Terna is to be found in the blazon that is recognized to the company. In addition, this opportunity was seen by all of us as a great challenge that was very captivating and stimulating not to be missed, also considering the relevance of the sector in which it currently operates and the position it has achieved within the sector itself.

We were asked to analyze various datasets containing data related to their economic activities, such as sales invoices, goods receipts, orders, deliveries, contracts, and others and, through them, be able to predict the cash flow, including income and expenses, relative to the first two quarters of 2022.

In this first introductory paragraph, we will briefly discuss the tasks we were assigned and completed, explaining how we operated and illustrating the results we arrived at. Next, we will go on to explain in detail all the various steps, following precisely each step that allowed us to obtain a good estimation of the result.

In order to be able to estimate as accurately as possible the cash flow for the quarters in question, we began by analyzing, understanding, cleaning, and rearranging all the various datasets, and then we collected the data that we found to be most useful and then entered them into two different regression models, linear and polynomial.

We chose to make use of these two regression models because they are among the most common, understandable, and tested for estimating numerical values, but also because we believe the inferred outputs to be truthful and to be able to reflect a potentially real and feasible situation. For the final outputs, we decided to consider only the results of the polynomial regression, which, in this case, was more accurate than the linear regression because, due also to the conformation of the data in our possession, it was better able to follow the trend and thus make a more accurate prediction that was more in line with the reality of the data provided by the company.

DATA UNDERSTANDING

The first thing we thought important to do was, of course, the part about understanding the data available to us. There were three datasets in question: one related to incoming invoices, another related to invoices for goods ordered and purchased by the company, thus concerning Terna's expenditures, and a dataset dedicated to the forecast of expenditures, relative to the various months of 2022.

To begin with, we went to check what other data, in addition to the amounts of the various transactions, had been provided, such as the date of issuance and payment of an invoice, the ID of the contract, of the goods purchased or of the company with which Terna entered business, etc... Next, we wondered if, within the various datasets, there were common data through which we could more easily conduct the process of understanding and analysis but, unfortunately, they were not present.

Once we understood this, we focused on understanding and studying the various datasets individually and discovered, especially in the one related to sales invoices, numerous problems or features that would have made our analysis inaccurate.

First regarding the dates of the invoices, themselves, we considered the two dates as the time of effected payment since we did not have a column containing the actual date of payment.

Immediately after that we noticed the presence of very numerous blocked invoices, with a ratio to free ones of 3 to 1 (74.7% blocked, 25.3% free) on which we particularly focused, going on to discover further problems.

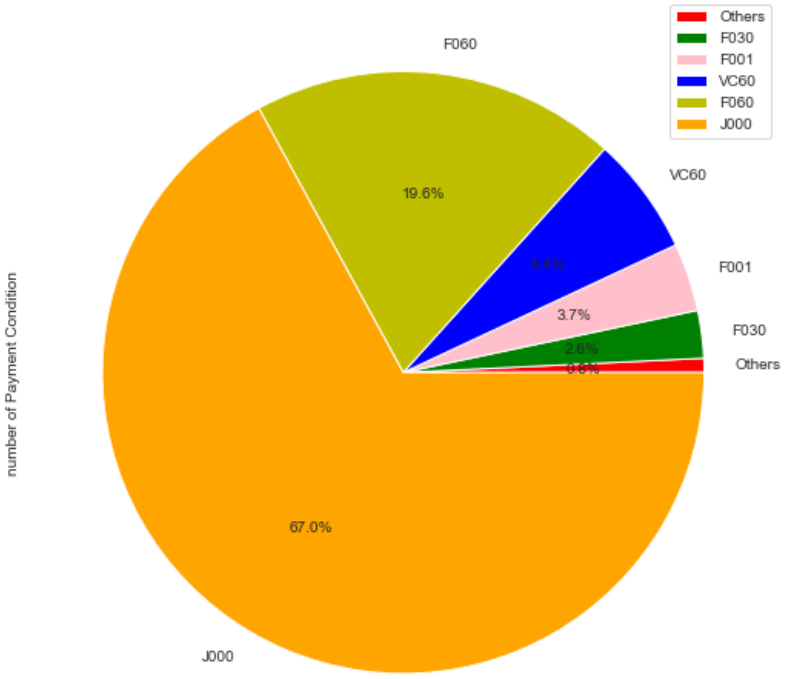
Among the numerous invoices, in fact, we noticed how many of them, mainly within the same order, were going to cancel thus finding the presence of numerous cancellations that, in most cases, completely canceled an invoice but, in others, simply went to touch up the amount of one or more invoices issued.

Also, within the dataset, we noticed that data from 2015 to 2022 were present but, in addition to the complete absence of 2016, 2015 and 2017 were characterized by the presence of very few invoices compared to later years and very often with discordant dates between issuance and payment. For 2015, for example, there is only one invoice, with an amount of -22692€, issued in 2021 with a due date for payment in 2015 while, for 2017, there are 14 invoices, with an amount of only 94,358.74€ but issued in 2018 and due in 2017. For these reasons, we thought that for the purposes of our analysis they might be unnecessary, if not even harmful.

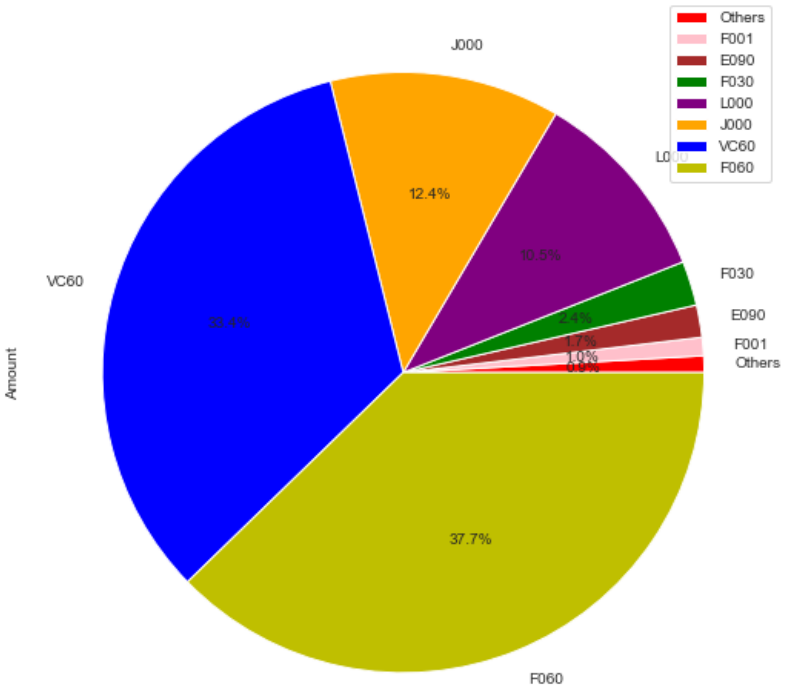
In addition, we noticed how there were a substantial number of suppliers for whom only negative invoices are associated, and so we began to wonder how to deal with this as well.

Finally, through another small dataset that served as a legend, we understood the meaning of the various payment terms by going back, even more in detail, to also understand how, and if, they have evolved over time.

Giving a couple of examples for explaining what we are talking about, we estimate that two-thirds (67%) of total payment invoices are of cautionary deduction.



From an amount-based perspective, most total invoices are collected within 60 days of recording the goods receipts (33.4%) and 60 days of the invoice date (37.7). Of course, all this attention we put on all the datasets provided is going to find, however, most of the problems, as mentioned earlier, in the one related to sales invoices.



As for the outgoing invoices for the purchase of goods, the work was much easier having simply analyzed the various columns, understood the link between them, and figured out what data we could use for further analysis. In this case we did not encounter all the problems observed for the incoming invoices dataset, but again we encountered the presence of reversals that went to cancel or correct invoices within the same order.

DATA CLEANING

Once we had made clear the overall picture of the situation in which we were operating, we began to clean and modify the data in such a way as to make them fit for the final analysis. Obviously, the data cleaning process was carried out by minimizing the number of values to be modified in order to avoid bias in the data. Having learned the significance of all the material at our disposal, we tried to figure out how many, and in which variables, there were missing values. As a matter of practicality, and because it would not bring negative consequences to our work, we opted for a cleanup of the dataset by acting directly on the file containing the data itself. For each variable in question, we devised a specific way to treat these data, considering that some could simply be kept as they were. In fact, we did not consider all the data to be compromising with our investigations precisely because they insisted on minor variables. Fortunately, there were not many values that had to be removed because they were particularly impactful, especially given the large amount of total data provided. This is true for the values that were null within the various datasets, but also for the problems highlighted earlier, such as those related to 2015 and 2017 invoices and those related to some suppliers, we had specific handling. For both years there were few but very confusing datasets, and therefore believing that they would simply cause a bias in the results, we decided to remove them. In addition, we reserved the same treatment for those invoices that came from suppliers with associated orders with a negative total value.

Another issue that we have already shed light on is that of the reversals that we noticed. As we said, we were able to divide the reversals into two types: those that went completely to offset an invoice that was part of the same order and those that simply went to touch up the total amount. As a result, we thought of lightening the dataset by eliminating the values of invoices having the same order that canceled each other out, while keeping, instead, the amounts, and the corresponding sales invoices, for the reversals that modified the first amount expressed.

DATA PREPARATION

After cleaning the datasets, we planned to select only a few variables in order to obtain the most useful and sensible results possible. In doing so, we relied on the variables that we ideally believe are most inherent to the actual formation of income and expenses, having as our goal the estimation of cash flow. Thus, in addition to the inevitable data regarding the amount and the various dates in the dataset, i.e., order date, invoice due date, and payment issue date, we selected the variables that lead back to the key elements of the transactions under consideration: invoice number, purchase order number, and supplier identification number. Finally, we also wanted to preserve conditions related to the contract, such as the respective ID number and invoice status. In fact, as we saw earlier some invoices are labeled as free and others as blocked, and we considered this distinction useful in the development of our work.

At this point, we were able to create, for sheer ease of understanding and to avoid errors due to confusion of various kinds, a new dataset containing all the variables listed above.

Finally, in this newly assembled dataset, we also divided the data by quarters in order to be able to stipulate as detailed an analysis as possible, even considering that the final output itself was to be expressed by quarters.

EXPERIMENTAL DESIGN

Having to arrive at an estimate of Terna's cash flow for the first two quarters of 2022, we asked ourselves what methods would be best. Strictly speaking, we assumed that the tactic that could most simply and correctly get to the point of the question might be one whereby we would conduct two separate analyses: one for revenue and the other for expenditures. And so we did. Step by step, we always conducted the two analyses in parallel, being able to take advantage of the fact that the process was common for both the study of sales invoices and that concerning invoices for the purchase of goods.

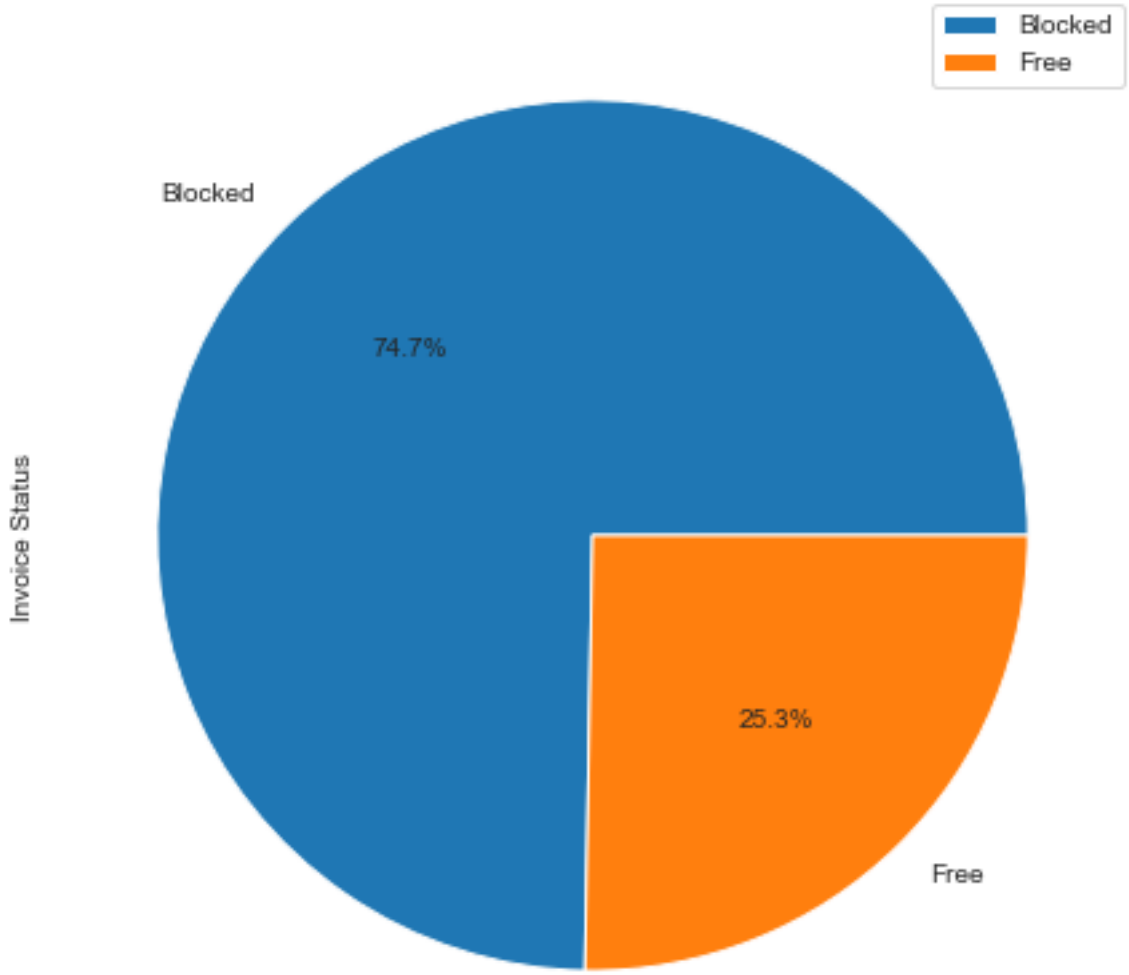
The goal of the analyses was to arrive at two numbers, which stood for the actual total of receipts and expenditures for the quarters in question. Thus, since we had to estimate a numerical value, the obvious choice seemed to be to select regression models as the algorithmic method to proceed.

In order to provide the timeliest results possible, we conducted our investigations through both linear and polynomial regression. From the beginning, we had assumed that a polynomial regression might provide more accurate results, and indeed it did. To highlight this, however, we have reported the results obtained with the linear regression method and compared them with the much better results we arrived at with polynomial regression.

In conclusion, we can say that most of our work was aimed at obtaining the input and output totals for the first two quarters of 2022, starting from the understanding of the data to their preparation and the mathematical operations performed then.

CODE DESCRIPTION

In the first part of the code, after importing the necessary libraries (numpy, pandas, seaborn, matplotlib.pyplot), we started by visualizing some of the raw data. We imported the necessary data from the excel using “pd.read_excel” first and replaced the missing data with a ‘0’ value only for visualization purposes (we imported “Invoices” sheet from "Terna Invoices.xlsx" as “dfvis1” dataframe and “Pivot Pc” from the same excel file as “dfvis2” dataframe. The first visualization is a pie chart made with matplotlib where we visualize the percentages of “Blocked” and “Free” Invoices: we extracted the number for each of the values in another dataframe (“df_pie”) and plotted. This is the result.

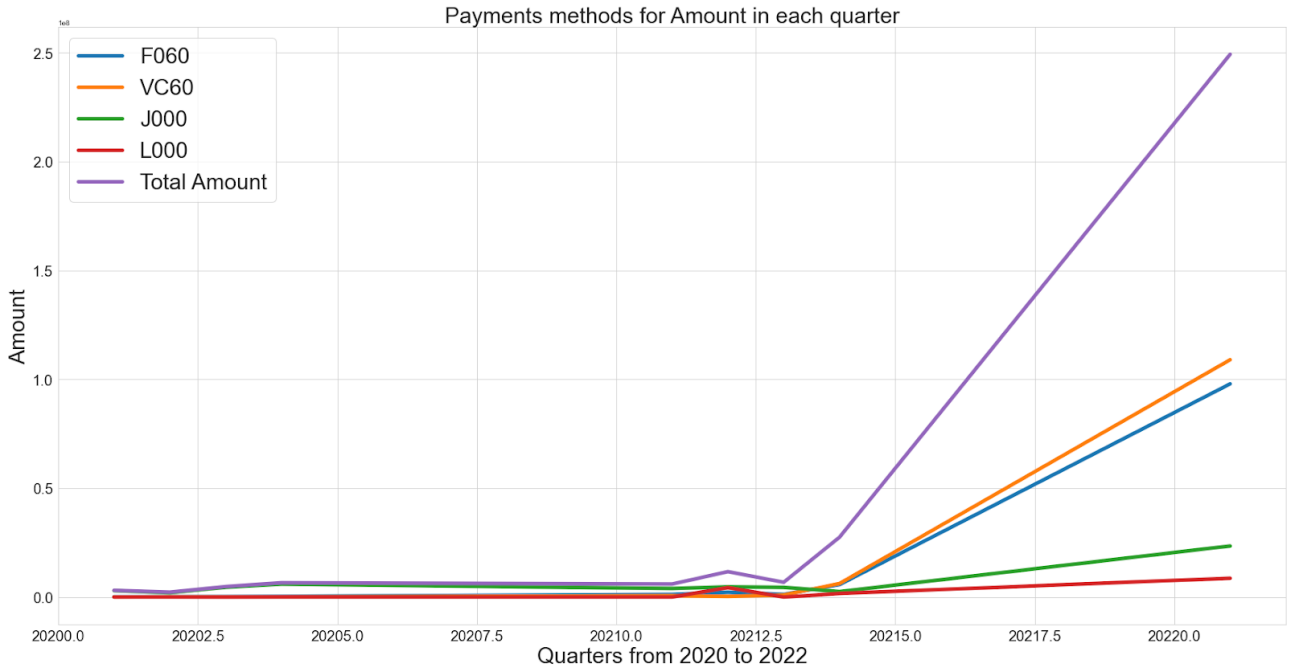


Then, we decided to analyze the “payment conditions” value, and, in order to do so, we created another dataframe called “df_pie1” (from “dfvis1”) where we put and sorted the number of different types of payment conditions. In order to plot them, we selected the first five values for the number of appearances and grouped the remaining values with a “for loop”, calling this new value “Others”. We associated each value with a color and plotted all those data in a pie chart (the first one in this report).

After, we did a similar thing, always with payment conditions, where we created a dataframe (data_pie2) in which we plotted the total amount of money for each payment condition. In order to plot it we used the same script of the first one, changing the variables, and the result is the second pie chart appeared in this report.

The last visualization is a line-graph in which we plotted the data extracted in “dfvis2”. We plotted the change in amount of the different payment conditions throughout the last two years grouped for quarters. We plotted the first four payment conditions, ordered for total amount, and we added the “Total Amount” (which represent the total value of the payment conditions) to give a sense of perspective. For this plot we used matplotlib. So, we create a list with labels and the base plot, setted the dimension and, with a for loop, we plotted the quarters with the corresponding amount: in each loop a different label for different payment methods is called from the list “c”.

Then we setted some visualization parameters such as the legend, the labels and the title. Finally, we have this as a result.



After the first visualization phase, we started with the real analysis.

We started by setting up different scripts for income and expenditure but still with the same purpose. Two regression models: linear and polynomial.

First, we imported the dataset we previously divided by quarterly values ("Terna Invoices.xlsx", sheet_name="Quarters Invoices").

Then, we selected the various quarters as the predictor variables and the total amount as the output variable to be searched and, subsequently, reformed them so as not to get errors of nonconforming magnitudes.

Next, we set up the division into training and test sets for possible subsequent analysis, but we did not find it necessary by deciding to rely on the results provided by the two computed regressions.

In this case, we did not choose to perform feature scaling since the data in our possession did not require it and, indeed, we felt it might be detrimental to the understanding of the results.

After that, we set up the script for the linear regression by importing from sklearn.linear_model.

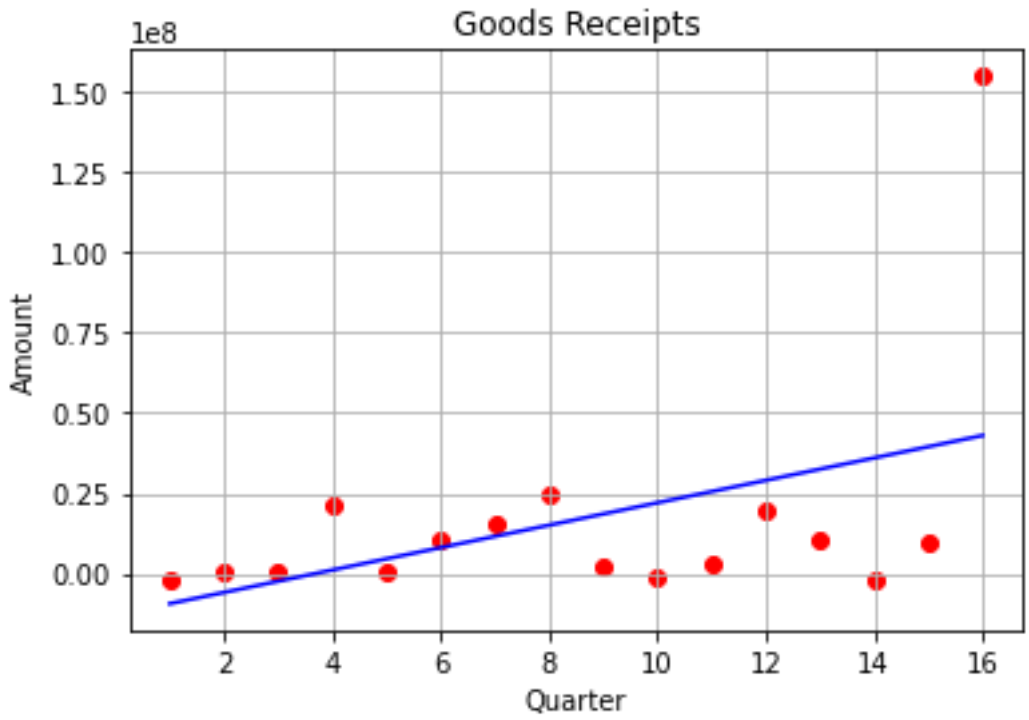
```
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression()
lin_reg.fit(X, y)
```

```
LinearRegression()
```

Next, we defined a function to display the results. Within it, we established colors: blue for the trend line and red for the various quarters in the plane, title, axis name, and the grid in the background to make the visualization clearer.

```
def viz_linear():
    plt.scatter(X, y, color='red')
    plt.plot(X, lin_reg.predict(X), color='blue')
    plt.title('Goods Receipts')
    plt.xlabel('Quarter')
    plt.ylabel('Amount')
    plt.grid()
    plt.show()
    return
viz_linear()
```

And this is the outcome:



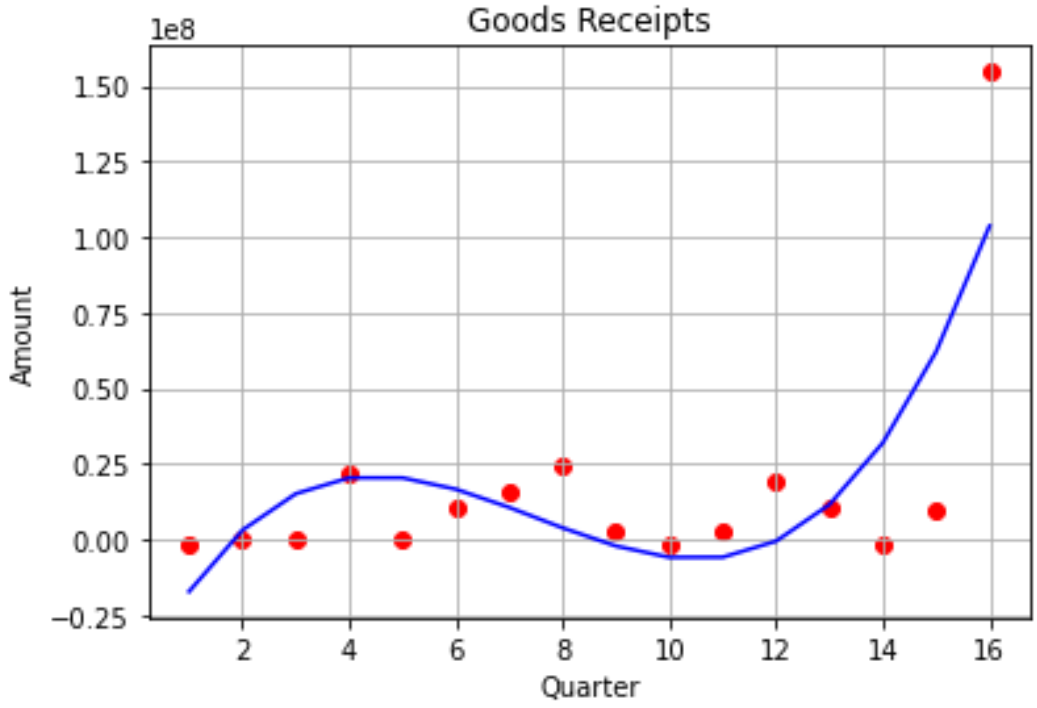
After that, we imported PolynomialFeatures from sklearn.preprocessing to develop our linear regression.

```
from sklearn.preprocessing import PolynomialFeatures
poly_reg = PolynomialFeatures(degree=3)
X_poly = poly_reg.fit_transform(X)
pol_reg = LinearRegression()
pol_reg.fit(X_poly, y)
```

Again, to plot, we defined a function in which we defined the same parameters used previously for linear regression.

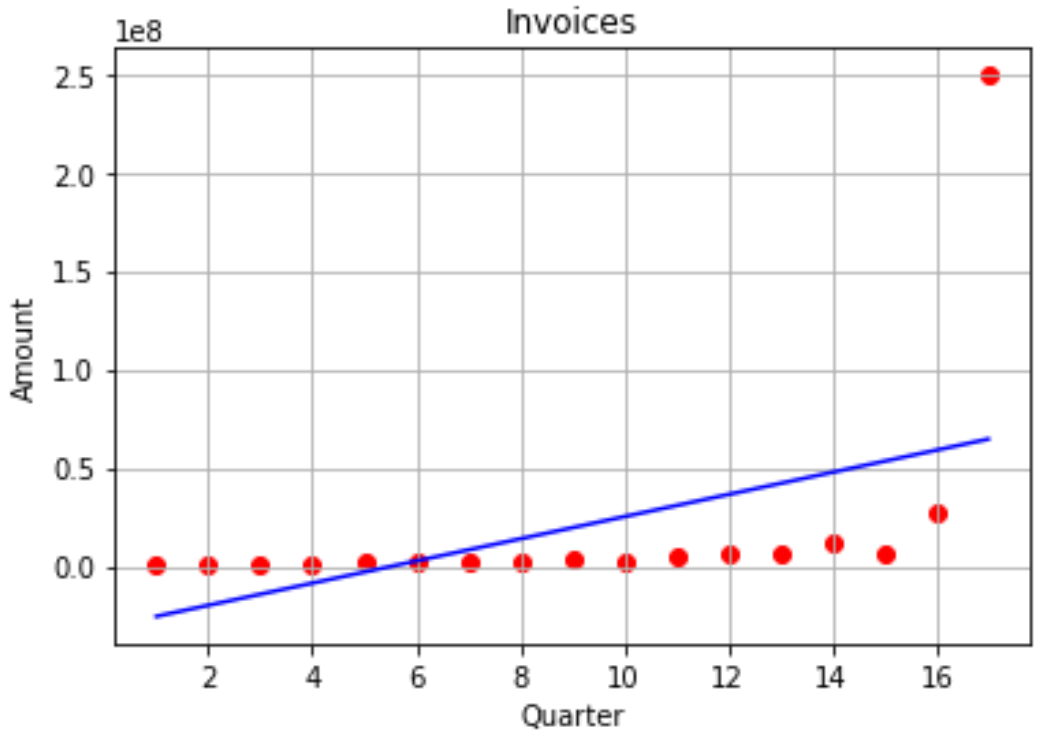
```
def viz_polymonial():
    plt.scatter(X, y, color='red')
    plt.plot(X, pol_reg.predict(poly_reg.fit_transform(X)), color='blue')
    plt.title('Goods Receipts')
    plt.xlabel('Quarter')
    plt.ylabel('Amount')
    plt.grid()
    plt.show()
    return
viz_polymonial()
```

The result obtained is as follows.

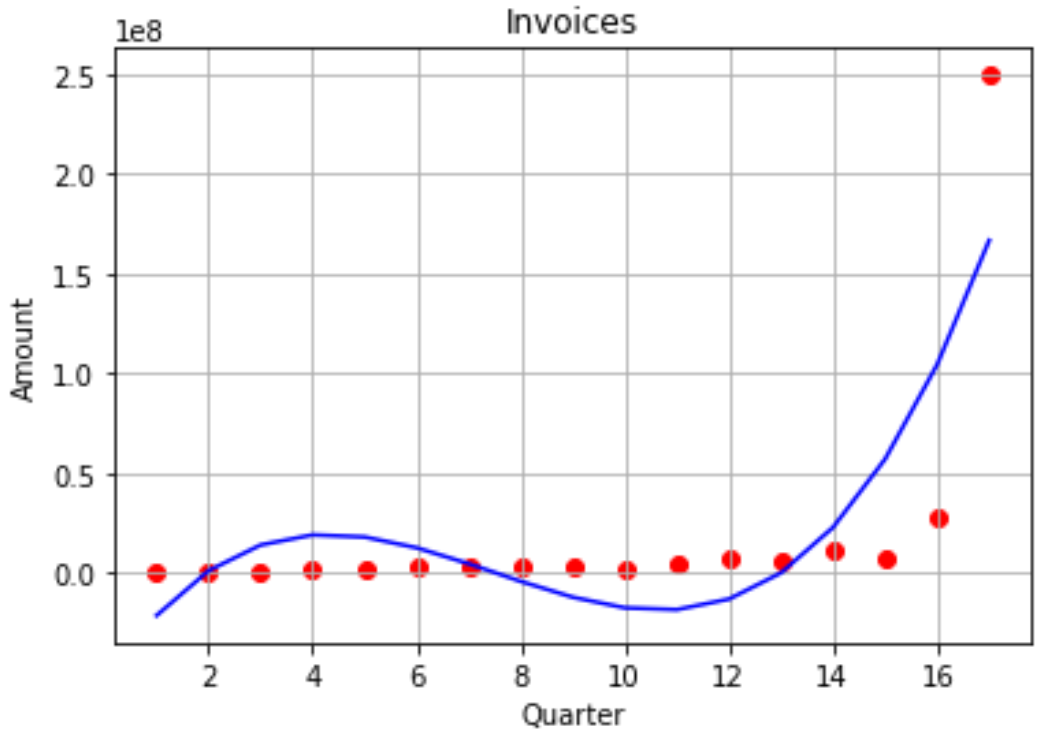


As for the dataset related to sales invoices, the procedures were the same. Obviously, the data were different, but the procedure was the same.

Below, we attach the results of the two regressions, the first concerning the linear regression.



Instead, these are those for polynomial regression for sales invoices.



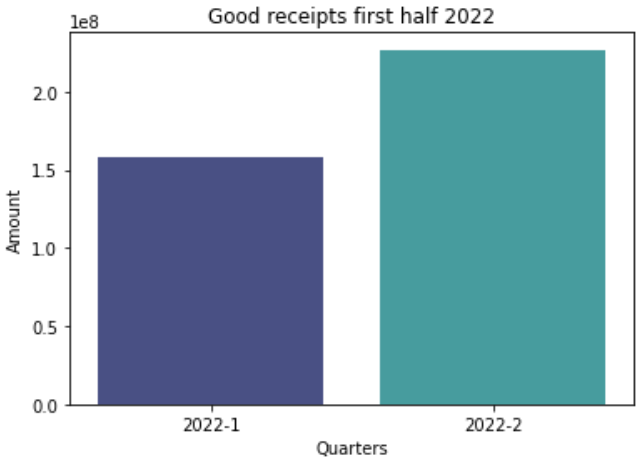
Finally, in the “result description” we plotted four different bar charts.

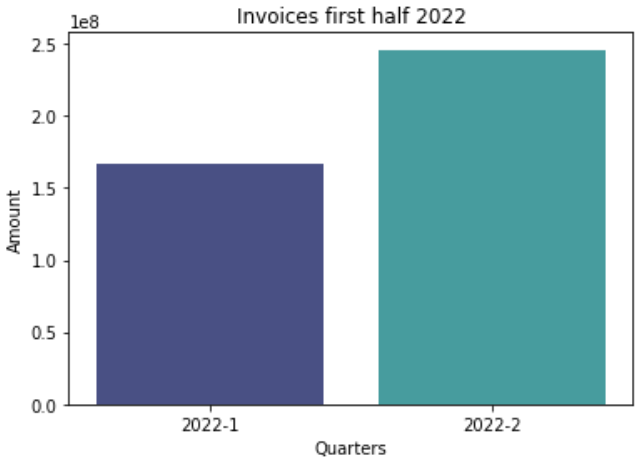
The first two (the first of the next part of the report) are a graphical representation of the values obtained from our analysis. We created four lists, two for the invoices and two for the goods receipts, with the obtained values and the corresponding quarter. Then we wrote the script in order to plot them using matplotlib.

The last two graphs are instead barplots, one for good receipts and one for Invoices, with all the amounts divided for quarters where we added our results. In order to create them we imported the necessary data from the “Quarter Pivot” sheet of “Terna Invoices.xlsx" and “Pivot DB Python” of "Terna Goods Receipts.xlsx" file. We respectively associated them to two dataframe, dfinvoices and dfgoodrecipts. After that we modified them by adding our predicted values with their respective quarter labels. Finally, we plotted these two dataframes using matplotlib. The resulting image of this chart is at the end of the next part of the report.

RESULTS

Here we have represented the values for the first and second quarters of 2022, relative to invoices and goods receipts, with a barplot that we believe can clearly indicate the growth we expect to occur in the two quarters.

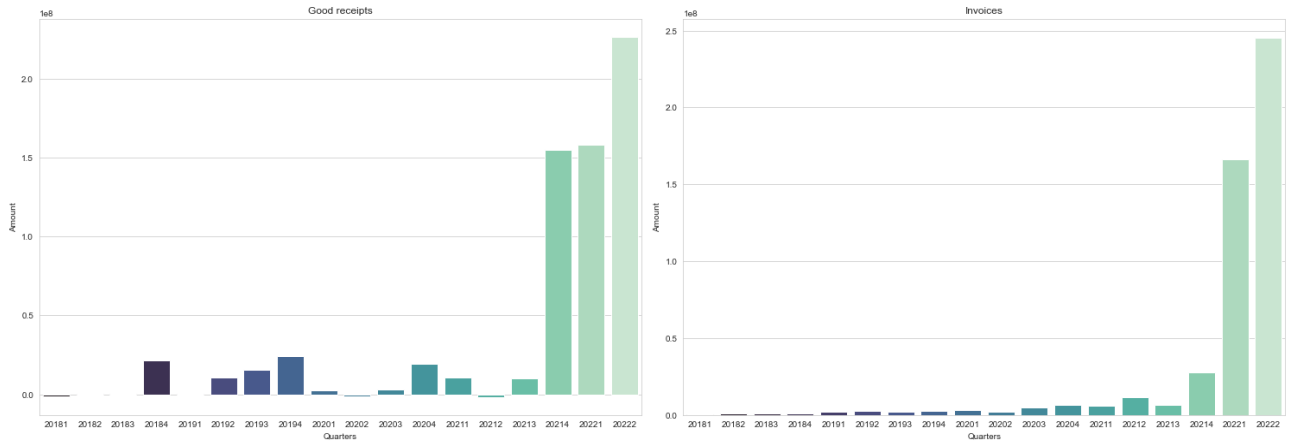




This below, on the other hand, is the table depicting the results of revenue and output predictions, again for the first half of 2022.

2022 Quarters	Payment invoices	Good receipts	Cash Flow
1° Quarter	166.491.388€	157.892.018€	+8.599.370€
2° Quarter	245.383.437€	226.468.893€	+18.914.544€
1° Half	411.874.825€	384.360.911€	+27.513.914€

These instead are two bar plots, one for the goods receipts and the second for the invoices, with the amounts for quarters, where we added our results in the last two columns.



Thus, our results show high revenues for Terna for each of the quarters subject to analysis and, overall, for the first half of 2022.

CONCLUSIONS

In summary, we can say that we arrived at the results we had hoped for with our analysis. In fact, we set ourselves the goal of obtaining estimates for revenues and expenditures for the first two quarters of 2022, and we did. Once these values were estimated, all we had to do to get the defined cash flow was to calculate the difference between revenues, which fortunately turned out to be the higher number, and expenditures.

However, as a group, we had already assumed that, through regression methods, correct estimates could come. Another prediction of ours that came true, turned out to be the one related to the concept that polynomial regression could provide more accurate results than linear regression.

In fact, the results are, in our opinion, satisfactory and make sense in logical and economic terms. From a logical point of view, they followed, step by step, our expectations and, from an economic point of view, they matched well, yielding a strongly positive cash flow for the first half of 2022.

From our conducted analysis, we can draw that even a simple regression model can be valid and that most of the tasks to be performed, and difficulties, are part of the process that culminates with the preparation of the final data.

SOURCES

The data available to us were provided directly by Terna.

SITOGRAPHY

https://en.wikipedia.org/wiki/Terna_Group

APPENDIX

In the course of our work, we worked mostly together because we realized that helping each other was the most fruitful way to complete the various steps.

In addition, we also found this kind of collaboration to be useful in order to maintain a common and unique style of writing and at the procedural level.

Overall, the parts that we preferred to divide among ourselves were minimized but this did not negatively affect the successful completion and planning of the final work.