# **ETL Project**

## Team 4

**Bianca Moreira**

**Grecia Villarreal**

**Heidy Guzman**

## **Introduction**

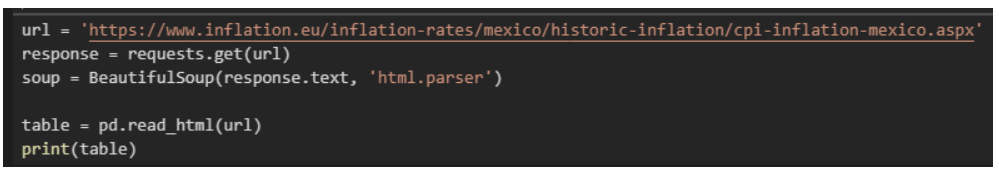
These data sources were chosen to analyze the correlation between financial indexes in Mexico by different periods of time. We got data from 3 different sources and we also used multiple methods to read and extract data.

## **Extract**

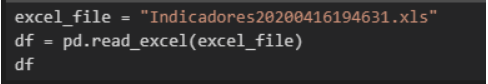
**Data Sources**

* **Inflation –** Inflation.eu **- HTML Web Scraping**

<https://www.inflation.eu/inflation-rates/mexico/historic-inflation/cpi-inflation-mexico.aspx>

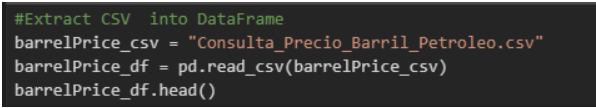
****

* **INPC –** Banxico - **XLS** <https://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?accion=consultarCuadro&idCuadro=CP154&locale=es>

****

* **Barrel Price -** Banxico **- XLS saved as CSV**

<https://www.banxico.org.mx/apps/gc/precios-spot-del-petroleo-gra.html>

****

* **GDP –** INEGI – **XLS saved as CSV**

<https://www.inegi.org.mx/temas/pib/default.html#Tabulados>

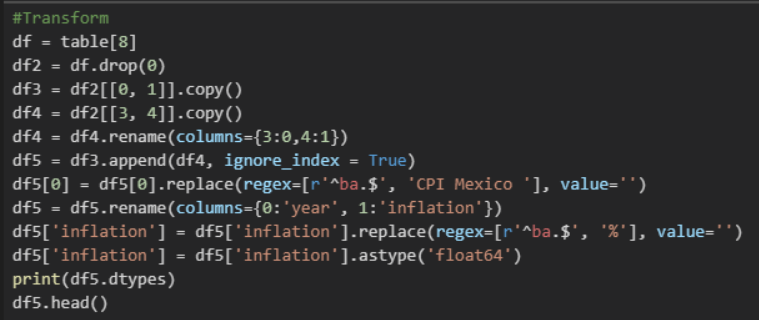
****

## **Transform**

* **Inflation**

1. Read HTLM data using BeautifulSoup and pandas.
2. Merge information from different columns to create a single one using append method (Inflation value and Year).
3. Drop first row header.
4. Rename Columns using regular expression method.
5. Remove special characters (%) and text.
6. Update data type

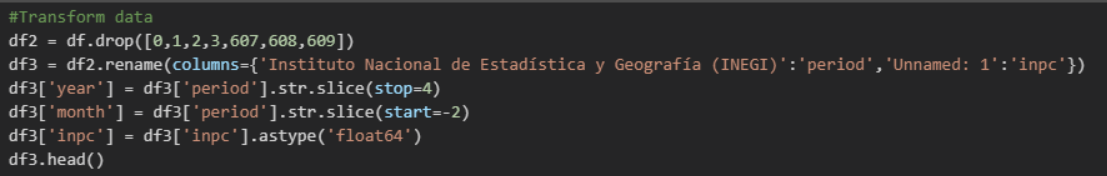
File Name: Inflation.ipynb



* **INPC**

1. Read xls file
2. Drop columns and rows that will not be used
3. Rename Columns
4. Calculate Year and Month using str.slice method
5. Update data type

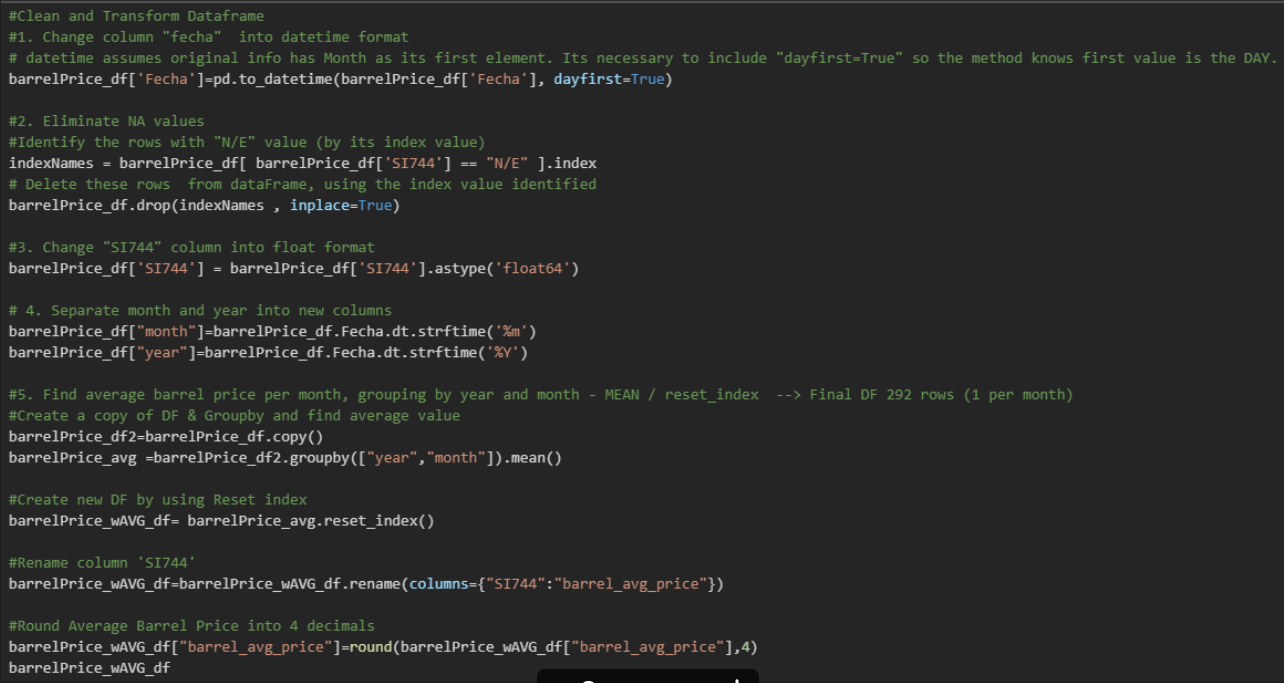
File Name: INPC.ipynb

****

* **Barrel Price**

1. Extract CSV into DataFrame
2. Clean and Transform Dataframe
   1. Change column "fecha" into datetime format
   2. datetime assumes original info has Month as its first element. It’s necessary to include "dayfirst=True" so the method knows first value is the DAY.
3. Eliminate NA values. Identify the rows with "N/E" value (by its index value)
4. Delete rows from dataFrame, using the index value identified
5. Update datatype
6. Separate month and year into new columns using time series method
7. Groupby and find average value
8. Create new DF by using Reset index
9. Rename columns

File Name: barrelPrice\_etl.ipynb



* **GDP**

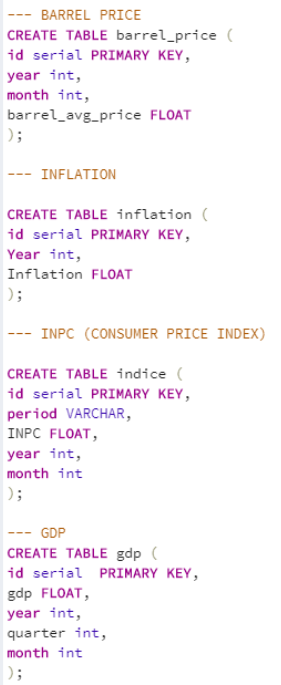
1. Read CSV File
2. Transpose data – Period Columns to rows
3. Exclude column that will not be used
4. Rename Columns
5. Create new month column using dictionaries, list comprehension and explode method
6. List values displays into multiple rows

File Name: GDP.ipynb

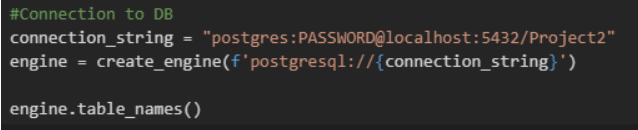


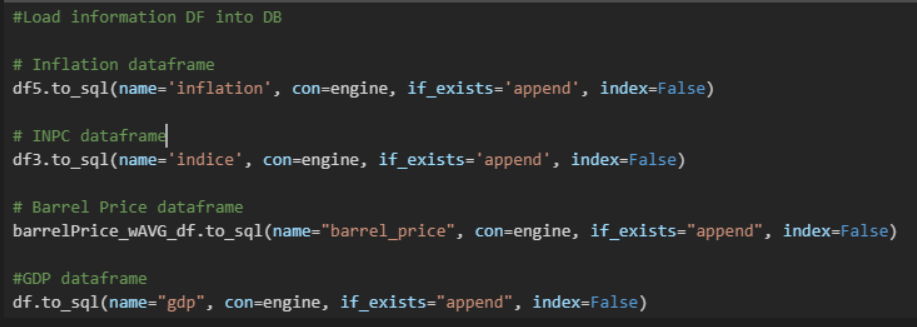
## **Load**

1. Data was loaded in SQL Server (PgAdmin). We are using SQL server because we are using relational data and will be easier to analyze and create different views.
2. We created 4 different Tables for each data source



1. Connection into database (Project2) and Data Load

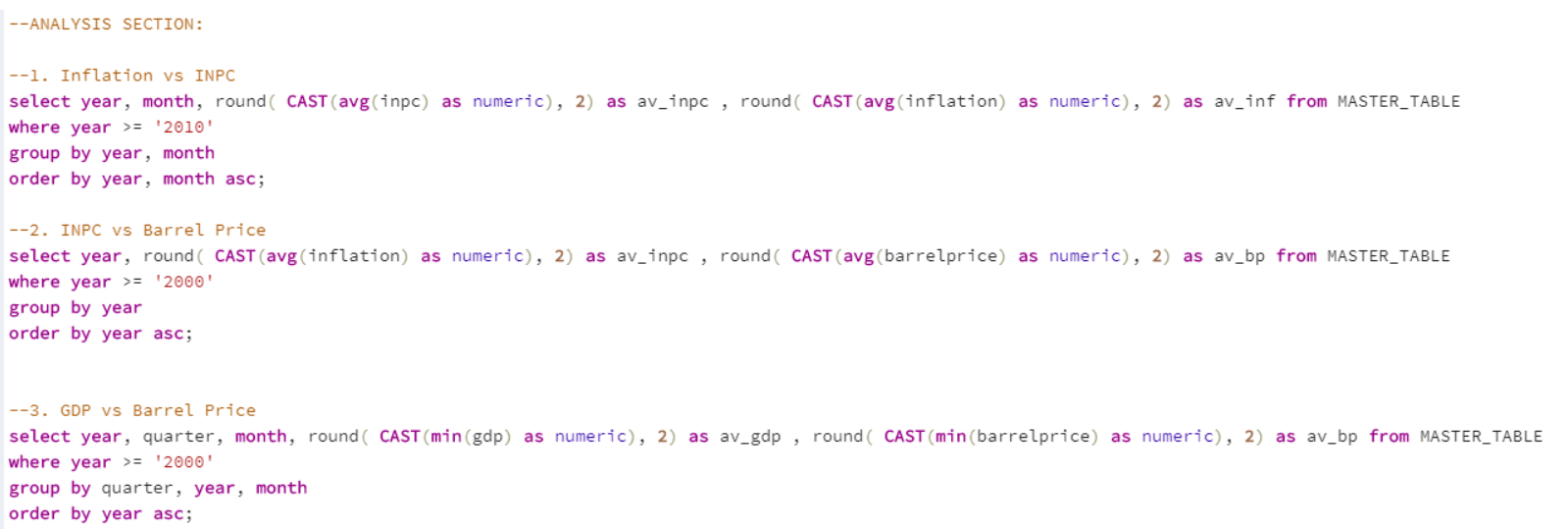




1. View was created into a Master Table by joining all data sources to facilitate the analysis.



1. Some Queries were created to analyze data



## **Conclusion & Learnings**

Using the queries, we could observe that apparently exists a positive relationship between the variables compared.

While analyzing the final data load we noticed several issues that we needed to get fixed during the transformation process such as the correct use of time series, datetime assumes original info has Month as its first element. It’s necessary to include "dayfirst=True" so the method knows first value is the DAY.

We also had to restructure the gpd dataframe to correctly merge our final master table by creating a new month column to stablish a relationship with quarter columns. We had to create a dictionary and also use the explode method.