Data Analytics

Economic Activities – Portugal

RNCP

Edgar Tomé

June, 2022

Table of Contents

Introduction

Business study economic activities in Portugal

The goal of this project was to create a data set to future analyse of the evolution of the economy in factors of production activities, efficiency and return, internationalization of the economy, and they sub factors.

Table 1 - Table of the factors and sub factors

Production Activities	**Efficiency and Return**	**Internationalization**
Export	Apparent labour productivity	Degree of exposure to international trade
Import	Investment rate	Export intensity
Production	Degree of production	Import penetration rate

The plan of the project was to chouse a data source that was able to provide the different factors of the economy in Portugal, by industry and the evolution along the years, and retrieve data from the choice source.

In this case the data source dint had API access, so for retrieving the data was by download of Excel files. Import the data to python script for clean, the none necessary columns and rows and nan values, and the years with complete data and the same industries in all tables.

After export of the clean data in to CSV files, imported the files in to a new script of Python, to produce visualizations, to perform a primary data analysis and verify that there were no outlier values, our still missing values and data it no relevance for the study. If necessary, return to the script in python and perform more data clean, to have the files exported in csv, to be imported in to database and future analysis

Create a database, create tables for import of the data information in all the clean files retrieved in csv for all the factors for analysis, and produce some queries that can give some evolution of the economy production.

Data and data sources

Data source is PORDATA (https://www.pordata.pt/), the Database of Contemporary Portugal, organized and developed by the Francisco Manuel dos Santos Foundation, was created in 2009.

The collection, compilation, systematization and dissemination of data on multiple areas of society, for Portugal and its municipalities, and for the European countries. The reported statistics derive from official and certified sources, with data production skills in the respective areas.

Consists in collecting and organizing the data available, making it as clear and accessible as possible. Also, important work of contextualized information, the so-called "metadata", as an inextricable part of the data, enabling its adequate interpretation, to be providing a public service to the Portuguese society, free of charge and without any cost to the user.

PORDATA retrieves information from over sixty official agencies, with particular emphasis to Statistics Portugal, cooperate with PORDATA.

Data collection

The data collection was retrieved in the format of excel files from PRODATA, for the three subfactors from each group of factors of the economy in Portugal, that are more important for the development and growth of the economy in the country.

The files were retrieved by download, as the site don't have API access to retrieve the data.

Table 2 - Files collected

Production Activities	**Efficiency and Return**	**Internationalization**
Export of goods and services total and by product (2016).xlsx	Apparent labour productivity total and by industry.xlsx	Degree of exposure to international trade total and by product.xlsx
Import of goods and services total and by product (2016).xlsx	Investment rate total and by industry.xlsx	Export intensity total and by product.xlsx
Gross value of production total and by industry (2016).xlsx	Degree of production processing total and by industry.xlsx	Import penetration rate in the domestic market total and by product.xlsx

Metadata

Information about the data collected.

Table 3 - Metadata production activities

Production Activities	Definition	Economic Activity	Geographic	Responsible entity
Export of goods and services total and by product (2016).xlsx	Exports of goods and services consist of transactions in goods and services	Activity is characterized by an input of products, production process and an output of products	Portugal	INE – National institute of statists
Import of goods and services total and by product (2016).xlsx	Imports of goods and services consist of transactions in goods and services			
Gross value of production total and by industry (2016).xlsx	Output is the total of products created during the accounting period			

Table 4 - Metadata efficiency and return

Efficiency and Return	Definition	Economic Activity	Geographic	Responsible entity
Apparent labour productivity total and by industry.xlsx	Measures the ratio between value added and the number of workers	Activity is characterized by an input of products, production process and an output of products	Portugal	INE – National institute of statists
Investment rate total and by industry.xlsx	Measures the relationship between gross fixed capital formation and gross value added			
Degree of production processing total and by industry.xlsx	Measures the ratio between gross value added and output, i.e., the relative proportion of value added per unit produced			

Table 5 - Metadata internationalization

Internationalization	Definition	Economic Activity	Geographic	Responsible entity
Degree of exposure to international trade total and by product.xlsx	Indicator that assesses the extent to which the production of each product is exposed to international competition both via export of domestic production and by competition with imports in the domestic market	Activity is characterized by an input of products, production process and an output of products	Portugal	INE – National institute of statists
Export intensity total and by product.xlsx	Measures the ratio between the value of exports and output			
Import penetration rate in the domestic market total and by product.xlsx	Import penetration assesses the growth of imports of goods and services in view of the growing global demand			

Data cleaning and Exploratory data analysis

Data cleaning

The data collection and exploratory data analysis process, had different process, because the excel files where not with the same formation, had different values, different types of values, missing values, different quantity of rows and columns. The process was in a loop, as after the first cleaning, in visualization verified that was necessary to remove more irrelevant information, and in the first creation of the database also realized that had to transpose the data frame to be stubble to the entity relationship model.

The process for the files export of goods, and import of goods, of the group production activities, the process of cleaning and preparation of CSV files to export to visualization and database, was the follow.

- Import of the libraries Pandas and Numpy;
- Import files with pandas to read Excel files;
- Listed the name of the columns to verifier;
- Place the name Years, in the same row as was the correct names of the columns;
- Delete rows with not relevant data, and non-values, and delete columns with not relevant data, and non-values;
- Rename the columns with the information on the row “6”, and print the data frame to verified the status of the data;
- As the names of the industries are long length, placed encoding to the columns;
- Placed new index to have order, as with the remove of rows the index add no order, and removed the old index, and the old columns name;
- Exported the files to csv, to perform visualizations, noticing that was some values with category of object, had to convert to integer and floats.
- Exported the files to csv, to perform visualizations, noticing that was some columns of industries that were no common to every file and category of study. Had to delete the row Col_39, as the industries was no common to the different groups, and for that reason, not comparable;
- Exported the files to csv, and uploaded on database with the entity relationship model as base, verified, that a new clean had to be performed to the files, to have them suitable for the creation of data base;

- Had to transform the data frame, so performed a transpose function, to convert e columns names in to index. Rename the columns names with the values of the row of Years, and remove the rows of Years and Total, then replace the columns names with range from 1995 to 2020, as the years were in float, and had to convert in integer from the year 1995 to 1999;

- Exported the file into csv with pandas, to be used on database;

1.2. For the file (Import of goods and services total and by product (2016).xlsx)

```
In [17]: #Import data using pandas
         pa_imports=pd.read_excel(r"/Users/edgartome_1/IronHack/IronProjects/Project4/Data_Economic_Activities/Production_Act

In [18]: #Place the name "Years" on the row that i will have the columns name, then drop the rest of columns and rows
         pa_imports.loc[pa_imports.index[6], 'Unnamed: 0'] = "Years"

In [19]: #Delete all rows with data not relevant
         pa_imports = pa_imports.drop(labels=range(0, 6), axis=0)

In [20]: #Delete all rows with data not relevant
         pa_imports = pa_imports.drop(labels=range(32, 54), axis=0)

In [21]: #Delete all columns with NAN values
         pa_imports = pa_imports.dropna(thresh=10, axis=1)

In [22]: #Rename the columns name with the values of the row index 6
         pa_imports.rename(columns=pa_imports.iloc[0], inplace = True)

In [23]: #Rename the columns by code "Col_" plus position
         pa_imports.columns = ['Col_' + str(i) if 3 <= i <= 39 else x for i, x in enumerate(pa_imports.columns, 1)]

In [24]: #Reset index
         pa_imports = pa_imports.reset_index()

In [25]: #Delete old index
         pa_imports = pa_imports.drop(labels=['index'], axis=1)

In [26]: #Delete old columns name
         pa_imports = pa_imports.drop(labels=[0], axis=0)

In [27]: #Convert to interger and floats, because on data visualiztion verifation that some tables had the data in objects
         pa_imports.apply(pd.to_numeric)
```

Figure 1 - Code process 1 part 1

```
In [28]: #Delete columns Col_39
         pa_imports = pa_imports.drop(labels=['Col_39'], axis=1)

In [29]: #Transpose
         pa_imports = pa_imports.T

         #Rename the columns name with the values of the row index 0
         pa_imports.rename(columns=pa_imports.iloc[0], inplace = True)

         pa_imports = pa_imports.drop(labels=['Years'], axis=0)
         pa_imports = pa_imports.drop(labels=['Total'], axis=0)


         cols=range(1995, 2020)
         pa_imports.columns=cols

In [30]: #Export to CSV file
         pa_imports.to_csv(r'/Users/edgartome_1/IronHack/IronProjects/Project4/Data_Economic_Activities/Production_Activities
         pa_imports
```

Figure 2 - Code process 1 part 2

For the rest of the files retrieved, the process of cleaning and preparation of CSV files to export to visualization and database, was the follow.

- On the excel files, delete rows and columns with nan values, or with legend information on the excel formation, not relevant for the analysis that was to be performed, and moved

the columns name to the first row. This was performed different from the first two files, because, during the first process of data cleaning, the python was unable to convert all the object types in to numeric values, like float our integer.

- Import files with pandas to read Excel files;
- Listed the name of the columns to verifier;
- Delete rows with not relevant data, and non-values, and delete columns with not relevant data, and non-values;
- To perform visualizations, noticing that was some values with category of object, had to convert to integer and floats, and verified that had changed;
- As the names of the industries are long length, placed encoding to the columns;
- Exported the files to csv, to perform visualizations, noticing that was some columns of industries that were no common to every file and category of study. Had to delete the row Col_39, as the industries was no common to the different groups, and for that reason, not comparable;
- Exported the files to csv, and uploaded on database with the entity relationship model as base, verified, that a new clean had to be performed to the files, to have them suitable for the creation of data base;
- Had to transform the data frame, so performed a transpose function, to convert e columns names in to index. Rename the columns names with the values of the row of Years, and remove the rows of Years and Total, then replace the columns names with range from 1995 to 2020, as the years were in float, and had to convert in integer from the year 1995 to 1999;
- Exported the file into csv with pandas, to be used on database;

1.3. For the file (Gross value of production total and by industry (2016).xlsx)

```
In [31]: #Import data using pandas
         pa_production=pd.read_excel(r"/Users/edgartome_1/IronHack/IronProjects/Project4/Data_Economic_Activities/Production_
         pa_production
```

Figure 3 - Code process 2 part 1

```
In [32]: #Liste the name of columns
         pa_production.columns

Out[32]: Index(['Years', 'Total', 'Agriculture, forestry and fishing',
                'Mining and quarrying',
                'Manufacture of food products, beverages and tobacco products',
                'Manufacture of textiles, wearing apparel and leather products',
                'Manufacture of wood and paper products, and printing',
                'Manufacture of coke and refined petroleum products',
                'Manufacture of chemicals and chemical products',
                'Manufacture of basic pharmaceutical products and pharmaceutical preparations',
                ...
                'Unnamed: 246', 'Unnamed: 247', 'Unnamed: 248', 'Unnamed: 249',
                'Unnamed: 250', 'Unnamed: 251', 'Unnamed: 252', 'Unnamed: 253',
                'Unnamed: 254', 'Unnamed: 255'],
               dtype='object', length=256)

In [33]: #Delete all rows with data not relevant
         pa_production = pa_production.drop(labels=range(25, 48), axis=0)

In [34]: #Delete all columns with NAN values
         pa_production = pa_production.dropna(thresh=10, axis=1)

In [35]: #Change the types of data have  to be used in visulizations and database
         pa_production = pa_production.astype(float)
         pa_production['Years'] = pa_production['Years'].astype(int)

In [36]: #Verify that the types of data have changed to be used in visulizations and database
         pa_production.dtypes
```

Figure 4 - Code process 2 part 2

```
In [37]: #Rename the columns by code "Col_" plus position
         pa_production.columns = ['Col_' + str(i) if 3 <= i <= 39 else x for i, x in enumerate(pa_production.columns, 1)]

In [38]: #Delete columns Col_39
         pa_production = pa_production.drop(labels=['Col_39'], axis=1)

In [39]: #Transpose
         pa_production = pa_production.T


         #Rename the columns name with the values of the row index 0
         pa_production.rename(columns=pa_production.iloc[0], inplace = True)

         pa_production = pa_production.drop(labels=['Years'], axis=0)
         pa_production = pa_production.drop(labels=['Total'], axis=0)


         cols=range(1995, 2020)
         pa_production.columns=cols

In [40]: #Export to CSV file
         pa_production.to_csv(r'/Users/edgartome_1/IronHack/IronProjects/Project4/Data_Economic_Activities/Production_Activit
         pa_production
```

Figure 5 - Code process 2 part 3

Table 6 - Codification of industries

Col_3	Products of agriculture, forestry and fishing
Col_4	Mining and quarrying
Col_5	Food products, beverages and tobacco products
Col_6	Textiles, wearing apparel and leather products
Col_7	Wood and paper products, and printing services
Col_8	Coke and refined petroleum products
Col_9	Chemicals and chemical products
Col_10	Basic pharmaceutical products and pharmaceutical preparations
Col_11	Rubber and plastics products, and other non-metallic mineral products
Col_12	Basic metals and fabricated metal products, except machinery and equipment
Col_13	Computer, electronic and optical products
Col_14	Electrical equipment
Col_15	Machinery and equipment n.e.c.
Col_16	Transport equipment
Col_17	Furniture, other manufactured goods, repair and installation services of machinery and equipment
Col_18	Electricity, gas, steam and air-conditioning
Col_19	Water supply, sewerage, waste management and remediation services
Col_20	Constructions and construction works
Col_21	Wholesale and retail trade services, repair services of motor vehicles and motorcycles
Col_22	Transportation and storage services
Col_23	Accommodation and food services
Col_24	Publishing, audiovisual and broadcasting services
Col_25	Telecommunications services
Col_26	Computer programming, consultancy and related services, information services
Col_27	Financial and insurance services
Col_28	Real estate services
Col_29	Legal and accounting services, services of head offices, management consulting services, architectural and engineering services, technical testing and analysis services
Col_30	Scientific research and development services
Col_31	Advertising and market research services, other professional, scientific and technical services, veterinary services
Col_32	Administrative and support services
Col_33	Public administration and defence services, compulsory social security services
Col_34	Education services
Col_35	Human health services
Col_36	Social work services
Col_37	Arts, entertainment and recreation services
Col_38	Other services
Col_39	Services of households as employers, undifferentiated goods and services produced by households for own use (DROP because missed on Investment rate total and by industry.xlsx)

Data visualization

During the process of data cleaning, was performed at the same time data visualization, in python with the libraries matplotlib and seaborn, to analyse if the data was clean, and the relations between the values were corrected, and if was need to removed outliers our columns and rows with no relevance to the study. This process was performed to all files, extracted during the loop process in data cleaning, with same visualizations and plots for each file.

In this example was applied the function to obtain bars plot for the evolution true the years for production activities exports, because of the relevance of exporting as to the economy in Portugal.

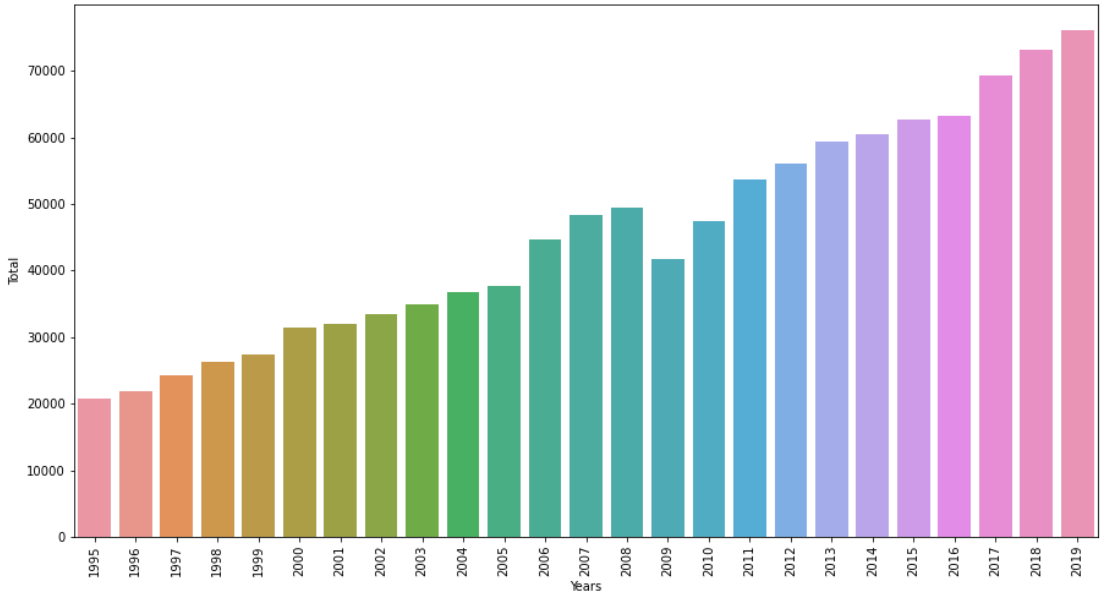

Figure 6 - Bars plot production activities exports Portugal

To be verified if existed outliers in the data, was used a function of box plot, for visualization of the distribution of the values for all the industries. Was identified outliers in the industries Col_16 (Transport equipment), that was verified has a growth on that industry in the past years, doubling the value in five years.

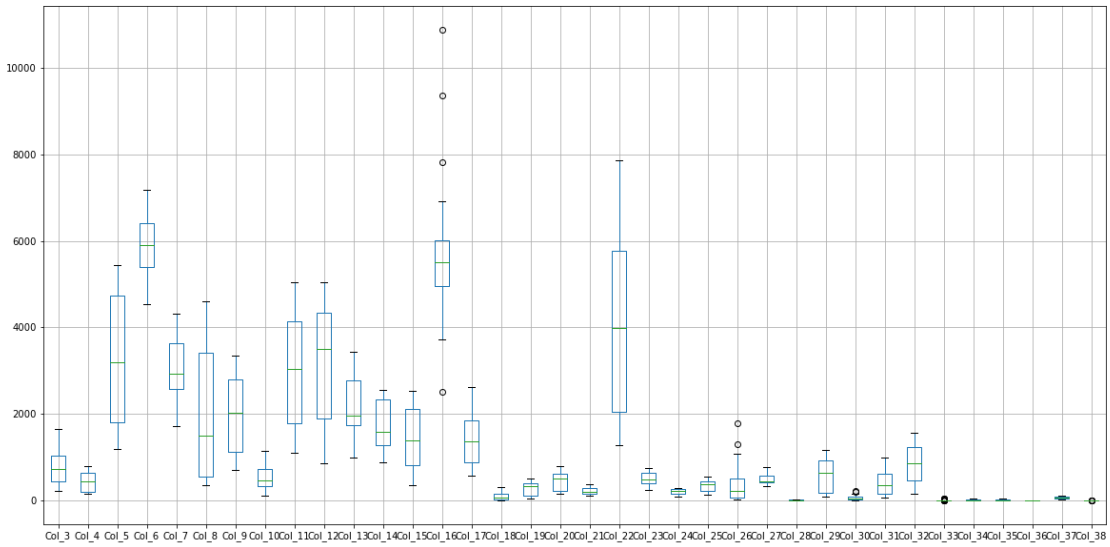

Figure 7 - Boxplot production activities exports

To be verified the relation between the industries in production activities exports, was applied the function correlation of numeric values and plotted using the function of heatmap.

Verifying that the industries Col_23 (Accommodation and food services), have higher correlation to the industries Col_3 (Products of agriculture, forestry and fishing) and Col_5(Food products, beverages and tobacco products), the increased industries of turism makes that the industries that provides food, beverage increases with the consumption of the tourists in the accommodations and in bars and restaurants.

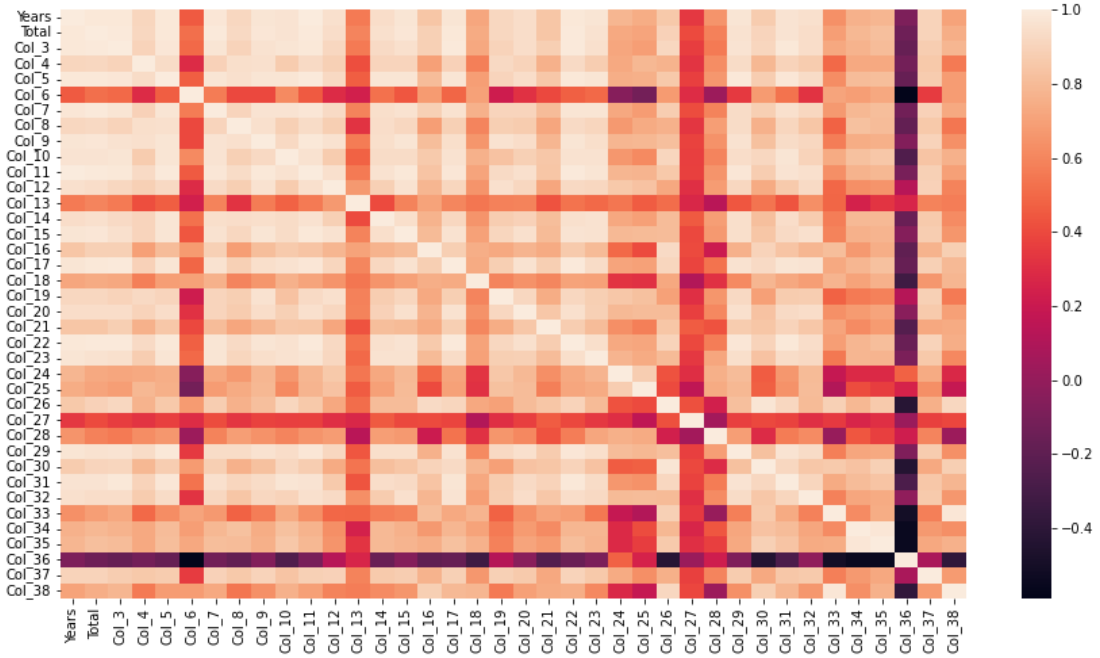

Figure 8 - Correlation heatmap production activities exports

Database

Types of databases

In order to store the data collected in a database, that can produce queries and retrieve information with relations our not, is necessary to choose the database that is more suitable for the project. Choosing the wrong database, can came in the long run as sometimes painful to fix, and is important to know the limitations in choosing one database for the project. There are two main types, relational databases (SQL based), and NoSQL databases.

The relational databases (SQL based), collects data in tables like csv files, each row in a table represents a record. It this structed database can be proceeded relations between rows, production of queries that relates information in different tables, but that have a relation key that are common in the tables for correlation. For that reason, before insertion of data, is need to produce the entity relationship model.

The NoSQL databases there is no common structured schema for all records, most of the NoSQL database contain JSON records, and different records can include different fields. The main types are document-oriented the schema can vary between different documents and contain different fields, as the records are not depended it supports parallel computations. Columnar database, the data is store column by column, that makes column-based queries very efficient. Key-value database, is based on key only, requesting for a key and getting its value, not supporting queries across different record values.

Comparing the relation database and document database, the relational as advantages in simple structure that matches most kind of data, supports join operations, allows fast data updating with relations between records, and performs atomic transitions. The disadvantage is the query execution time that depends of the size of the table. The advantages in document database allows to keep object with different structures, can represent almost all data structures, supports schema validation making collection schematized, the querying is very fast as the independent and therefore the query time is independent of database size, the disadvantages are that the process of updating the database is slow, and not atomic transactions are possible.

Entities. ER Model

For the project considering the data collected and relations between industries that are of great value for analyses, it was chooses the relation data base in SQL. For the use of relations database is necessary, before upload the data into SQL, create entity relationship model, with the different entities, and their relations, primary keys and foreign keys.

In this project for each table was create a primary key with the code of industries, and created a dictionary table with primary key being the name of the industries and the foreign keys the code of industries, to be related with each table of information. The tables can be related between them with the primary key that is the code of the industries and is common to all tables, producing queries of relations between exports and imports and production values

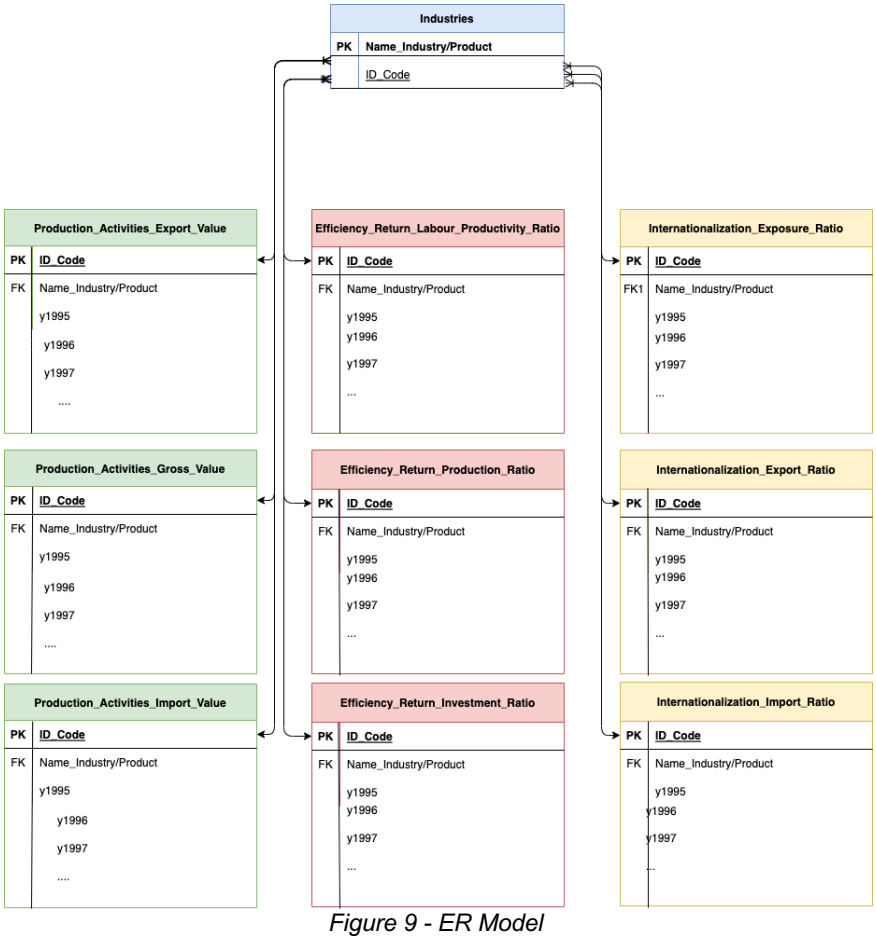

Figure 9 - ER Model

SQL database

For the project considering the data collected and relations between industries that are of great value for analyses, it was chooses the relation data base in SQL. The import of the data was performed true the wizard process as the data was big with many files and columns and rows and many numeric values.

After the imported data was placed in the tables created in SQL script, was performed some queries, with join tables and left join of tables, using has primary key the code of the industries, to relate the information on the different tables, and analyze the relations between the information collected.

The five queries created are the follow:

1. Production value comparison for import and export for the year 2019, for the industry with more value of export and industries with more value of import;

2. Production value comparison for import and export for the year 2019;

3. Internationalization ratio comparison for import and export for the year 2019;

4. Efficiency ratio comparison for investment and production for the year 2019;

5. Total of production export and production import for last 5 years all years

Conclusion

The process of collection data, cleaning data, visualization data, creation of database and retrieving queries from all the data inputted. It serves as the principal base for analyzed the data of the object of study required, giving information, that have to contextualized with object in study and whit is required of the project.

This process is a loop, that for each iteration, some conclusions can be made and a new perspective can be observed, being necessary to perform all the process again to refine the data, so can be obtain the conclusion with data support to the question placed of the object in study.