Waste Management EDA project- Technical Report

Autor: Mary Cruz Meza

Brief

Solid waste management is the one thing just about every city government provides for its residents. Although levels of service, impact, environmental impact and costs vary from one territory to another, solid waste management is undoubtedly one of the most important municipal services. It is for this reason that this project analyses the management in the autonomous communities of Spain.

Hypothesis

"The development of societies and population growth, among other things, influence the increase in municipal solid waste generation. The world's population is growing by leaps and bounds, projected to reach 8.5 billion by 2030, 9.7 billion by 2050 and 11.2 billion by 2100 (United Nations, 2019). Similarly, global environmental indicators show that more waste is being generated over time, creating environmental problems in urban ecosystems, especially because of the challenge faced by governments in the management and disposal of such waste. According to a report presented by the World Bank in 2017 (2018), 2010 million tonnes of MSW are generated annually in the world, and an important fact to highlight is that at least 33% of this waste is managed at risk to the environment. Therefore, the hypothesis of this project is: Population growth and economic development generate an increase in the production of solid waste."

For the execution of this project, I have followed the following steps:

1. Data Sourcing

Data Sourcing is the process of finding and loading the data into our system. For this project, the data submitted by the INE have been used, which are considered public data.

https://www.ine.es/

At this point in the project, my initial focus and hypothesis were different when I chose the subject of the project. This had to be changed due to a lack of information, which meant that I was delayed in starting the project.

2. Imports the required libraries for EDA

First of all, import the needed libraries. Common data science functions have been used: NumPy, pandas, matplotlib.

3. Load the data into the data frame.

Loading the data into the pandas data frame is certainly one of the most important steps in EDA. First, we can list all the available data files. There are a total of 5 files. 3 files representing waste management: 1 file with information on waste generation, 1 file on waste disposal, 1 file on emissions produced by waste management. And 2 files to contrast the previous variables: 1 on the Spanish economy (GDP) and 1 on the population.

4. Basic Data Exploration

In this step, it will be checked what the datasets are composed of. As they are different databases, they have different formats and types of objects.

5. Data Cleaning

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into the system.

For the development of this project, a database cleaning is carried out independently or following patterns in databases with a similar format. To subsequently join the 5 databases.

* Missing Values
* Incorrect Format
* Incorrect Headers
* Anomalies

The most complicated part of the project was undoubtedly the cleaning of the database. Having 5 independent datasets, each with a different format, required care and dedication when standardising them.

5.1. Melt row into column

For this database, it has been decided to place the years and the autonomous communities as rows, and the waste management variables will be placed as columns. The first difference is observed in the GDP and population databases. We proceed to eliminate the years that are not going to be considered in the analysis and to invert the shapes of the bases, to bring them closer to the desired format.

This section, together with the previous one, required a greater amount of time due to the complexity of linking the various databases. This step generated a lot of frustrations and fights in the process of my project. And above all, it took up more time than I thought it would.

5.2 Databases Merge

After the different databases have been adapted to a similar format, they are merged into a single database that will be used for the exploratory analysis. In this step of the project, the database is clean, without any NAN, However, in the next steps, the test for nulls and outliers will be performed.

5.3 Renaming the columns

In this instance, most of the column names are very confusing to read, so I just tweaked their column names. This is a good approach it improves the readability of the data set. The language of the variables in this database has been changed to standardise it according to this report. Identifying with C for collection, D for disposal variables.

6. Exploratory Data Analysis

In this part of the process, statistics are computed and calculations are made to find trends, anomalies, patterns or relationships within the data. Will serve to see what the data can tell us beyond the formal task of modelling or hypothesis testing.

After merging the databases in a standard format, the database is cleaned, important columns are created to analyse the indicators and, above all, graphs are plotted and conclusions are drawn.

6.1. Dropping the missing or null values

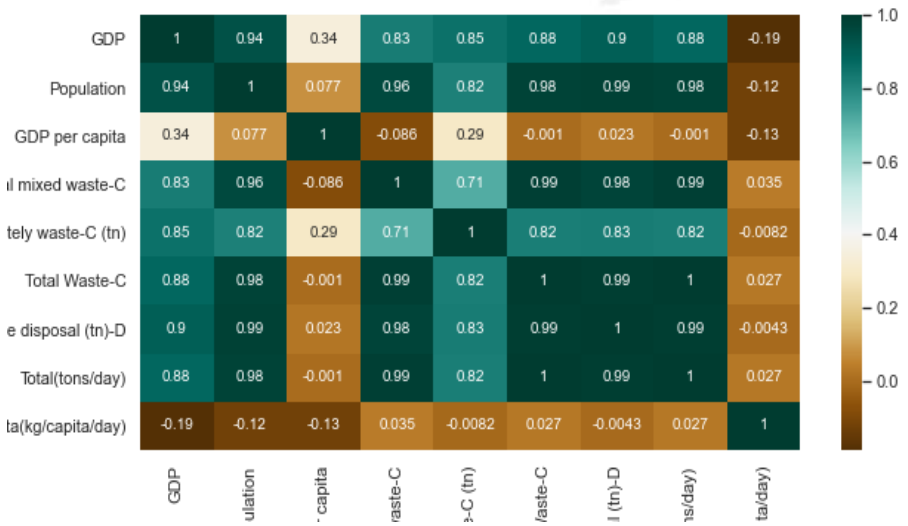
In this step the missing values are detected, however, in this case, no null value found. To answer the hypothesis, and to continue the exploratory analysis, the following variables are calculated: COLLECTION PER DAY, KG PER CAPITA, PROCESSED WASTE & NOT PROCESSED WASTE.

6.2 Examine the distribution of the variables

The objective of this project is to analyse solid waste management in Spain Autonomous Communities and to verify the relationship with economic and population growth. For this reason, some target variables have been selected to obtain the result of the hypothesis.

Distribution of key variables: 'Total Waste-C', 'Per Capita (kg/capita/day)', 'Processed_waste', 'Not Processed_waste'.

Distribution of all columns, bivariate analysis. To plot the database in an orderly manner, it has been separated into graphs that analyse the correlation, and time series analysis. The final result, and above all the image that best defines the hypothesis, would be the following:



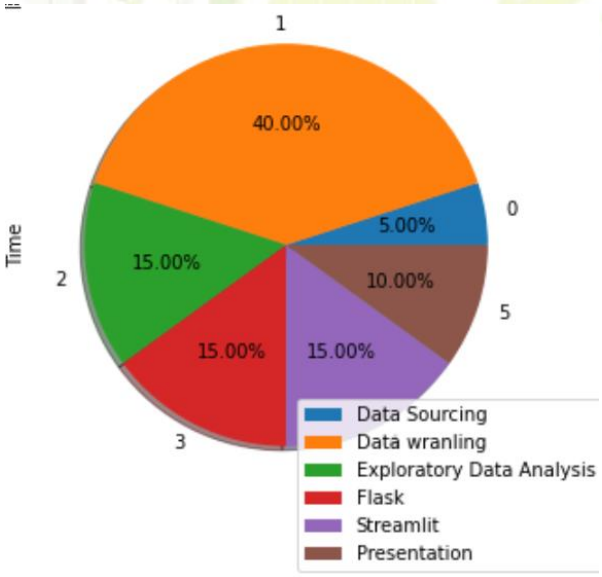
It is observed that all the variables that reflect both waste generation and final waste disposal show a positive and significant relationship with population and GDP, i.e. the hypothesis can be affirmed.

It can therefore be concluded that as a population has a higher income, or has a higher population development, it tends to consume more and therefore will generate more solid waste.

To do another EDA project, I would change the organisation at the beginning of the project, specifically with the database, in the choice of data and data cleansing. And especially to set the tasks.

By doing this project, I have reaffirmed above all, the knowledge acquired in the Bride, because previously I had done data research projects, but using different tools and from another approach, however, this project has allowed me to feel more confident in the knowledge of the Bootcamp.

Most of the time spent on this project has been devoted to the cleaning and merging of the database as mentioned above, which is reflected in the graph below.



The process of working with flask and streamlit also turned out to be a bit complex for me, especially because at the end of this process, I noticed that I had some errors in the previous steps, which I had to correct and fix so that the graphics could be presented without any problems in the dashboard. But it has certainly all been worth learning from.