# Waste Management EDA proyect

## ***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:

## 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/>

## Imports the required libraries for EDA

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

## 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.  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

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.

*5.2 WASTE COLLECTION & DISPOSAL*

*It has been observed that both the collection and the disposal base have a similar format, therefore a similar cleaning function is applied for both.*

*5.3. WASTE EMISSIONS DATABASE CLEANED*

*For this database, no disaggregated information on pollutants by Autonomous Community has been found. However, to obtain the amount of pollution that has been produced by waste management, this information has been retained, which, when joined to the general database, will be NaN.*

*5.4 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.*

*5.5 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, and E for emissions.*

*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.*

*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*

*5.1 EXAMINE THE DISTRIBUTION OF THE VARIABLES*

*The objective of this project is to analyse solid waste management in the Spanish Autonomous Communities and to verify the relationship with economic and population growth.*

*Distribution of key variables*

*The following have been selected as main or explanatory variables: Total Waste-C,' Per Capita(kg/capita/day)', 'Processed\_waste', 'Not Processed\_waste'*

*CORRELATION ANALYSIS*

*Correlation is a type of association between two numerical variables, specifically, it assesses the trend (increasing or decreasing) in the data. For this project, it has been hypothesised that the growth of waste generation may be related to population growth and economic growth, which will be proved or disproved at the end of this project.*

Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media

SPENT TIME