CCT College Dublin

Assessment Cover Page

Module Title:	**Strategic Thinking**
Assessment Title:	Ca2: Global plastic usage: future impact awareness
Lecturer Name:	**James Garza**
Student Full Name:	**Cristina Priolo**
Student Number:	Sba23037
Assessment Due Date:	**17th December 2023**
Date of Submission:	**16th December 2023**

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Global plastic usage: future impact awareness

Table of content:

Introduction:

The main purpose of this report is to examine the usage of plastic and the global impact that it has to bring awareness to the readers.

Plastic pollution has emerged as one of the most pressing global environment challenges of our time. The report will delve into the complexities of the situation and its far-reaching significance, aiming to shed light on the myriad issues that underlie this crisis.

By examining the contributing factors of the increasing plastic usage, the economics of plastic production this report seeks to highlight a future forecast about the situation.

Objectives:

This assignment aims to explore the following objectives:

1- Global usage of plastic: Investigate and examine the current usage of plastic worldwide. Highlight main factors of consumption based on previous and actual data with mass production trends.

2- Impact of plastic utilisation: Forecast will indicate: What are the consequences? How will pollution and the ecosystem be?

3- Waste management: Exhibit the recycling data in order to respond to possible future risks.

By addressing these objectives, the goal is to raise awareness to have hypothetical sustainable solutions.

Problem Definition:

This report will identify the key issues of global plastic usage. The global problem of plastic usage presents a multifaceted crisis that demands immediate attention. With each day passing, the excessive reliance on plastic deepend, leading to dire consequences for the environment and future generations.

Awareness is crucial for this report will delve into the gravity of the situation, emphasising the extreme need for immediate action. By addressing this crisis as early as possible this is essential to safe-guarding and preserving the planet.

Although, By addressing it holistically, this include:

1. Reducing single-use plastics
2. Recycling resources
3. Fostering global cooperation

Challenge of the project are the following:

- Appropriate data: finding dataset that have enough rows to support the objective of the analysis
- Avoid Bias in the analysis: staying neutral and focusing on the facts.
- Finding sustainable solutions for the problem supported with reliable dataset: the lack of insufficient data has led to continuous research. The project aims to raise awareness which gives the reader an opportunity to research sustainable solutions.

The context of the problem and the important of this to be addressed are:

> *"Plastic pollution is a planetary threat, affecting nearly every marine and freshwater ecosystem globally. In response, multilevel mitigation strategies are being adopted but with a lack of quantitative assessment of how such strategies reduce plastic emissions The global threat from plastic pollution"* (Borrelle et al. 2020, p.1).

Scope:

Over the two semester the scope of the project is to analyse the following topics and try to answer the following questions:

- Usage of plastic worldwide: which factors are impacting, mass production, forecast for future production of the plastic if nothing will change.

- Pollution and Ecosystem: How much the pollution increased and the ecosystem degraded? What are the causes?

- Analyse the recycling waste: what do we recycle? What is the capacity of the recycling facilities? Based on the analysis will we be able to plan and respond to the demand?

The project aims to bring at the end of the two semester attention on the topic to be more conscious about the long term effect on this concept.

Inclusions of the project:

- Definition of the problem and objective
- Analysis of the dataset of worldwide plastic usage
- Forecast on plastic usage
- Analysis of global pollution dataset
- Forecast about pollution
- Analysis of general waste
- Forecast about capacity of recycling facilities
- Conclusion of the analysis to bring awareness on the topic

Exclusions:

- Bias
- Avoid using personal data
- Not providing sustainable solution

Role and responsibilities:

Role	**Who:**
Project Manager	Cristina/Hodan
Researcher	Cristina/Hodan
Data collection and cleaning	Cristina/Hodan
Data Analysis	Cristina/Hodan
Data visualisation	Cristina/Hodan
Writer	Cristina/Hodan

Table 1: Role and Responsibilities.

Boundaries:

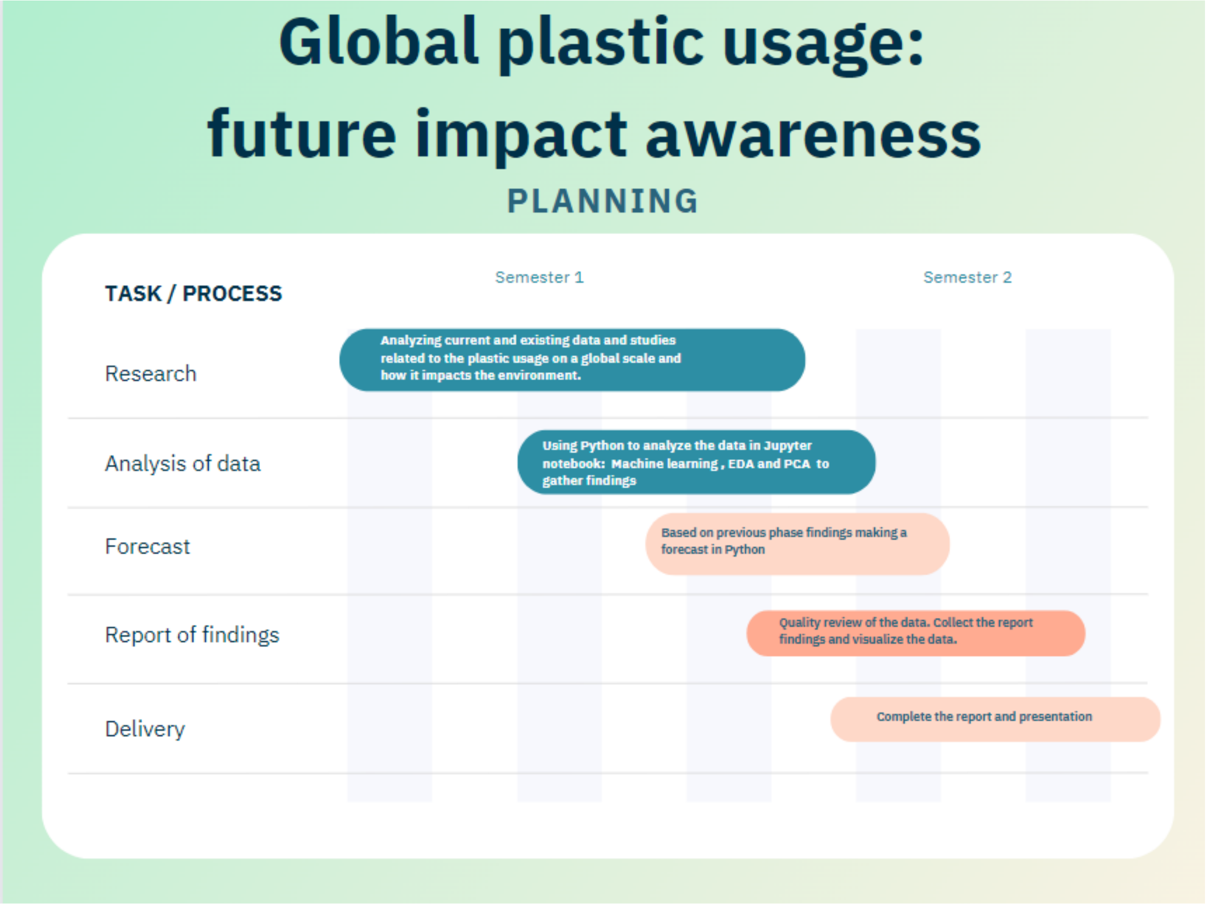
- Dataset limitation
- To avoid the limitation of the geographic area the project aims to analyse the situation on a global scale as to why the project title evolved during the time
- Research of solution to the argument
- Time frame of the project: Two semesters

In-depth analysis:

	First semester	Second semester
Focus	Foundational research: extensive literature review and data set collection	Analysis of garthing, the synthesis of findings and the formulation of rememendations
Aims	Prepare the material to move to the second semester phase.	Allow for a deeper exploration of global policies, case studies as well best practices and will also provide ample time for additional review if required which will ensure that the report quality.

Table 2: Analysis of tasks

Planning:



Global plastic usage:
future impact awareness

PLANNING

TASK / PROCESS

Semester 1

Semester 2

Research: Analyzing current and existing data and studies related to the plastic usage on a global scale and how it impacts the environment.

Analysis of data: Using Python to analyze the data in Jupyter notebook: Machine learning , EDA and PCA to gather findings

Forecast: Based on previous phase findings making a forecast in Python

Report of findings: Quality review of the data. Collect the report findings and visualize the data.

Delivery: Complete the report and presentation

Table 3: Project planning

Project accomplishment:

The main recommendations and strategies for raising awareness, driving change as well as promoting recycling best practices. This approach will ensure a well rounded, evidence-based exploration of the global plastic usage issue and future awareness.

Final consideration on the project scope:

By the end of semester two we aim to deliver a comprehensive academic report that includes the following:

1. Extensive research: a well-researched report featuring a thorough literature review, primary and secondary research, data analysis and an in depth exploration of global plastic usage. This will include:
    - Daily plastic consumption
    - Global pollution
    - Recycling waste facilities
    - Forecast based on the data
2. Data-driven recommendations: This will include evidence based recommendations for raising awareness to mitigate the future impact of plastic usage

Potential data for the project:

Data Source	Data amount	Permission
www.kaggle.com. (n.d.). Global Plastic Pollution. [online] Available at: https://www.kaggle.com/datasets/sohamgade/plastic-datasets	Full	open resources allowed by their terms and conditions
Datopian (n.d.). Daily_csv plastic monkey 78. [online] DataHub. Available at: https://www.datahub.io/gitchenze/daily_csv-plastic-monkey-78 [Accessed 15 Oct. 2023]	Full	open resources allowed by their terms and conditions
Our World in Data. (n.d.). Extrapolated change in plastic fate. [online] Available at: https://ourworldindata.org/grapher/plastic-fate-to-2050.	Full	open resources allowed by their terms and conditions

Table 4: Data sources

The data that has been found as potential for the project are from open resources so they are allowed to be used by their terms and conditions.

Ethical considerations:

While the report global plastic usage: future impact awareness does not directly involve sensitive data, user privacy or potential societal impact, however we believe ethical considerations remain essential to the report.

The report will prioritise transparency and accuracy in the use of data, adhering to proper dataset usage.

All the sources referenced will be appropriately cited under the guidelines of Harvard Reference to acknowledge the contribution of others to avoid plagiarism.

Additionally the report will emphasise the importance of responsible data handling and the ethical use of information for academic and educational purposes.

Report Paper of Artefact:

This section aims to explore the dataset “per-capita-plastic-waste-vs-gdp-per-capita” utilising data analysis techniques.

The section will cover:

- **Dataset analysis and exploration**:
  - Explanation of the dataset decision
  - Preliminary analysis of the dataset
- **Visualisation techniques:** visual representations to show:
  - Trends
  - Patterns
  - Relationships within the dataset.
- **EDA and Visualization**:
  - Data cleaning methods
  - Encoding categorical variables
  - Handling missing values
  - Methodology explanation
- **Model discussion**:
  - Explanation of machine learning models
- **Conclusion**:
  - Considerations
  - Challenges.

Dataset analysis and exploration:

The project utilises the public dataset from Keggle “Global Plastic Pollution”. A preliminary analysis was performed to determine which dataset was suitable for the project and the dataset “per-capita-plastic-waste-vs-gdp-per-capita.csv” has been selected for its number of observations and features.

Analysis and exploration of the dataset are necessary steps to start understanding the raw data and the pattern in it to extract vital information and details from it.

Utilising various libraries the dataset is imported and analysed:

- Pandas: for data manipulation
- Matplotlib:for visualisation
- Numpy: for numerical computation
- Seaborn: for statistical visualisation.

The dataset has 48168 rows and 7 columns, which imply: 48168 observations and 7 features.

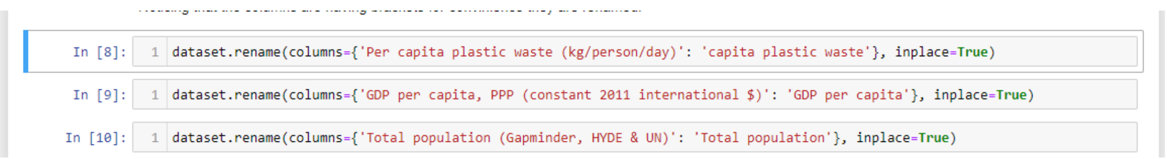
An essential step is to understand what type of variables are in the dataset. This preliminary check helps to determine how to proceed in the next steps. The function ‘dataset.dtype’ gives the following information:

Feature	Type
Entity	object
Code	object
Year	int64
Per capita plastic waste (kg/person/day)	float64
GDP per capita, PPP (constant 2011 international $)	float64
Total population (Gapminder, HYDE & UN)	float64
Continent	object

Table of content of variable types

This is vital to understand how to proceed when it comes to handling the missing values.

Some features' name contains brackets, for convenience they have been renamed:

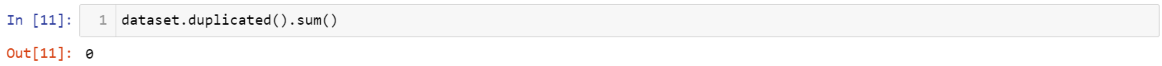


```
In [8]:  dataset.rename(columns={'Per capita plastic waste (kg/person/day)': 'capita plastic waste'}, inplace=True)

In [9]:  dataset.rename(columns={'GDP per capita, PPP (constant 2011 international $)': 'GDP per capita'}, inplace=True)

In [10]: dataset.rename(columns={'Total population (Gapminder, HYDE & UN)': 'Total population'}, inplace=True)
```

Image of the code of renaming

The next step involves starting checking the dataset for duplicates and missing values in order to proceed with the visualisation and the data processing.

The dataset has zero duplicated values:



```
In [11]: dataset.duplicated().sum()
Out[11]: 0
```

Image of the code of duplicated values

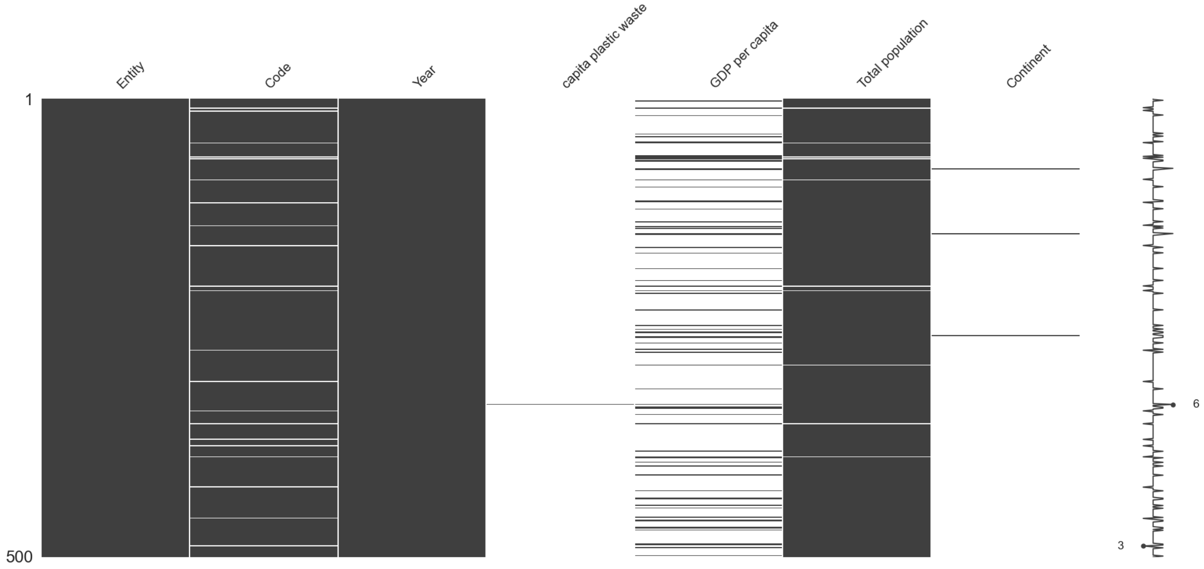
However it has the following missing values for a total of 140925:

Feature	Amount of missing Values
Entity	0
Code	2014
Year	0
capita plastic waste	47982
GDP per capita	41761
Total population	1285
Continent	47883

Table of missing values

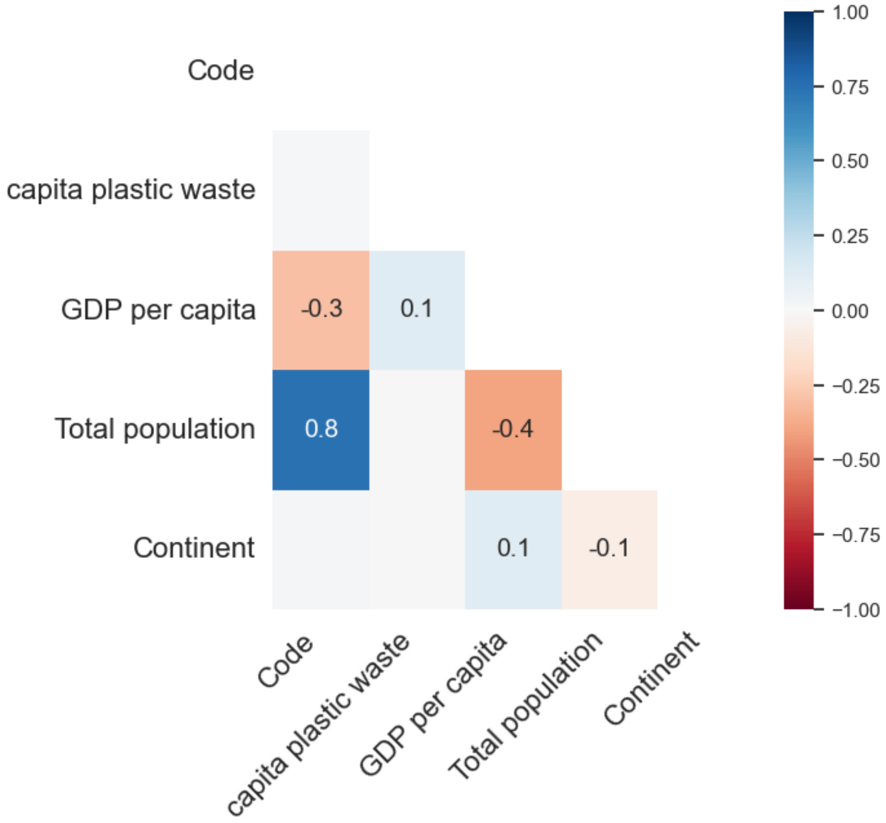
The library Missingno comes to help with the missing value visualisation.

- The matrix identify the location of the missing data within the dataset:



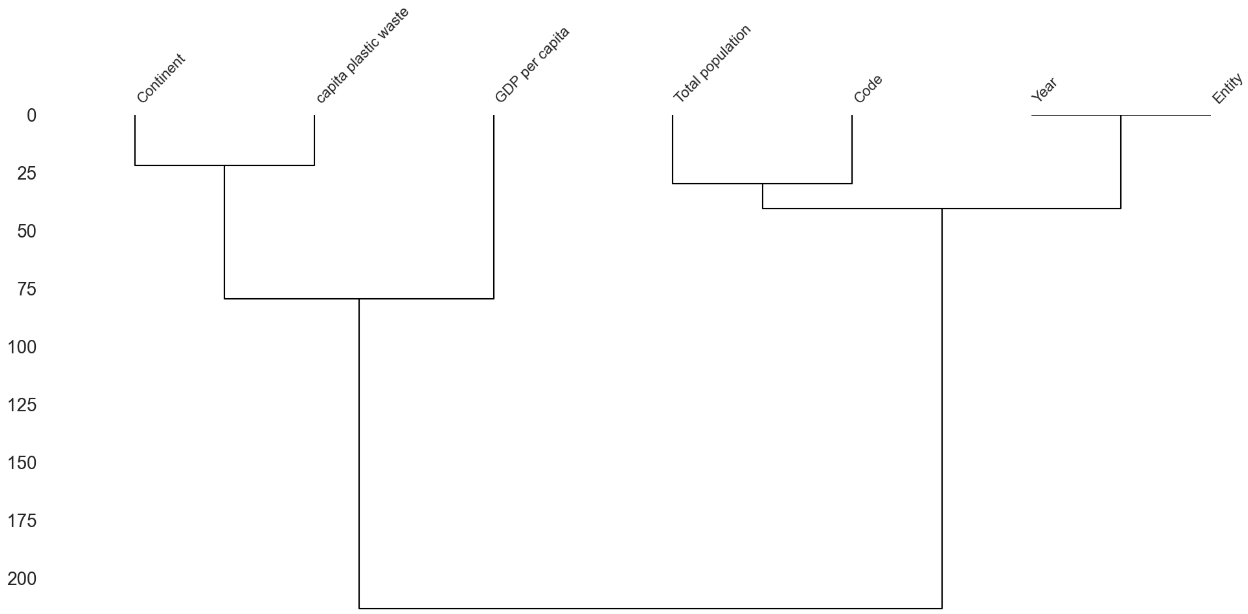
Graphic: Matplot of missing values

- The heatmap shows the correlation between missing values across different columns. This help understand the relationship inside the missing values inside different columns, if there is any:



Graphiac: HeatMap

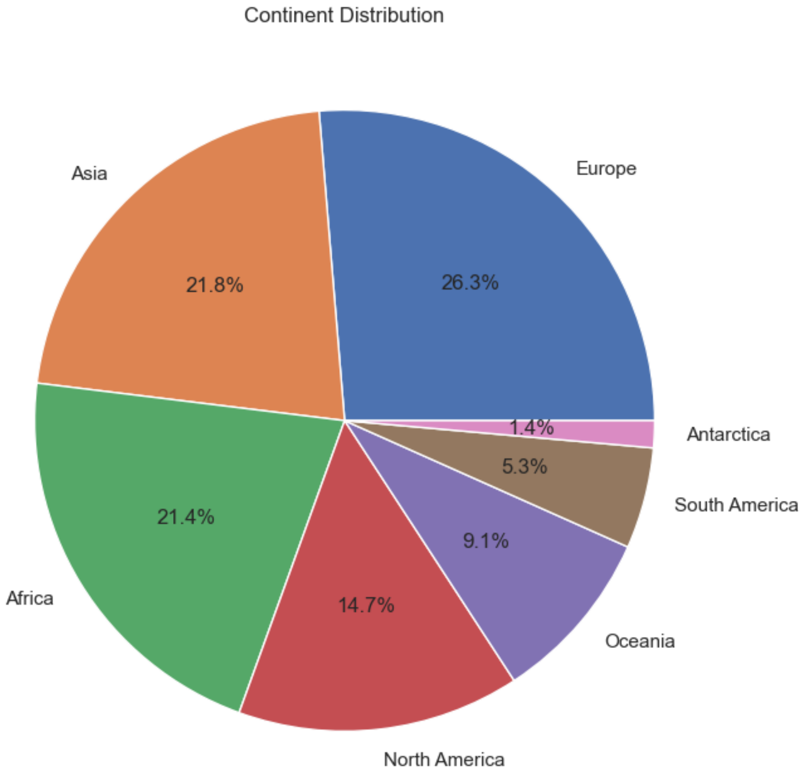
- The dendrogram identifies the group of columns that have similar patterns of the missing values:



Graphic: Dendrogram

Visualisation:

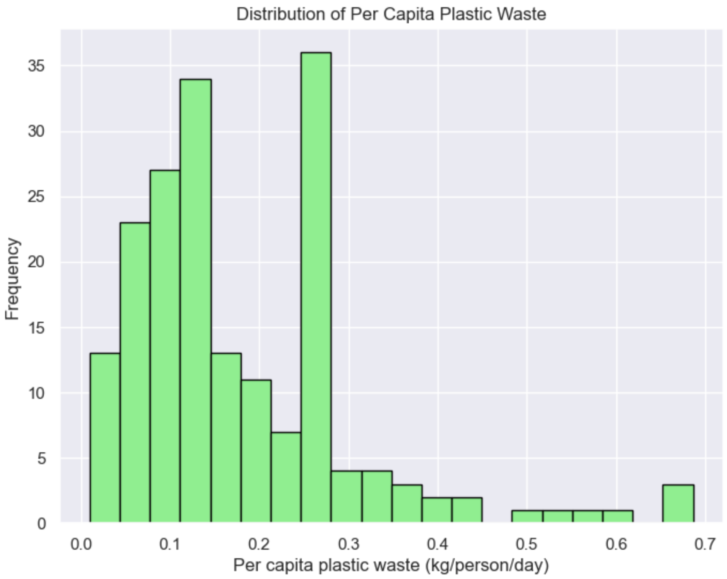
Before addressing the missing values to see the pattern and understand the dataset it is performed some visualisation. Visualise the distribution of the categorical variable is performed at this stage before transforming them into numerical ones to understand the pattern:



Pie chart of distribution of continent

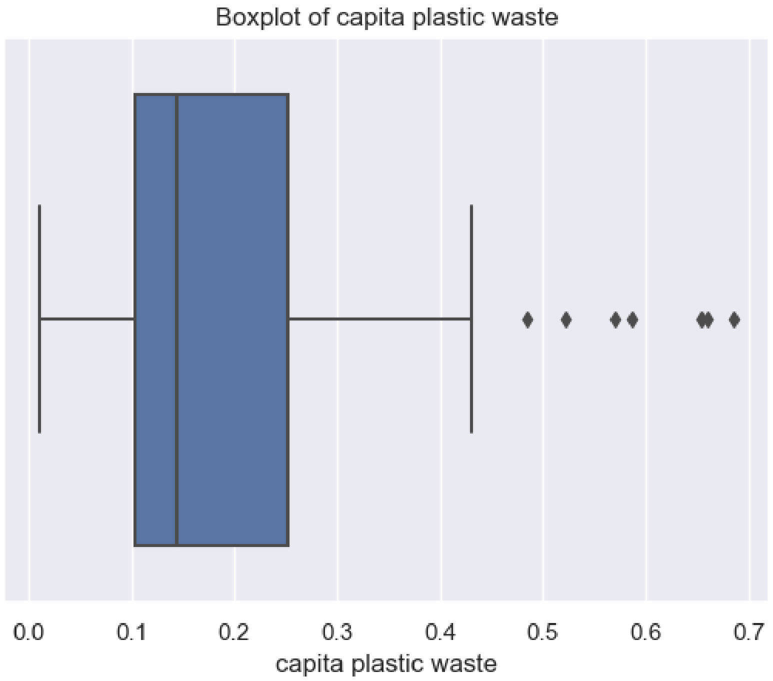
Africa,Asia and Europe are the majority in the distribution of the Continent feature.

After it has been visualised with the Histogram the frequency for “Per capita plastic waste” which shows the fluctuations of it:

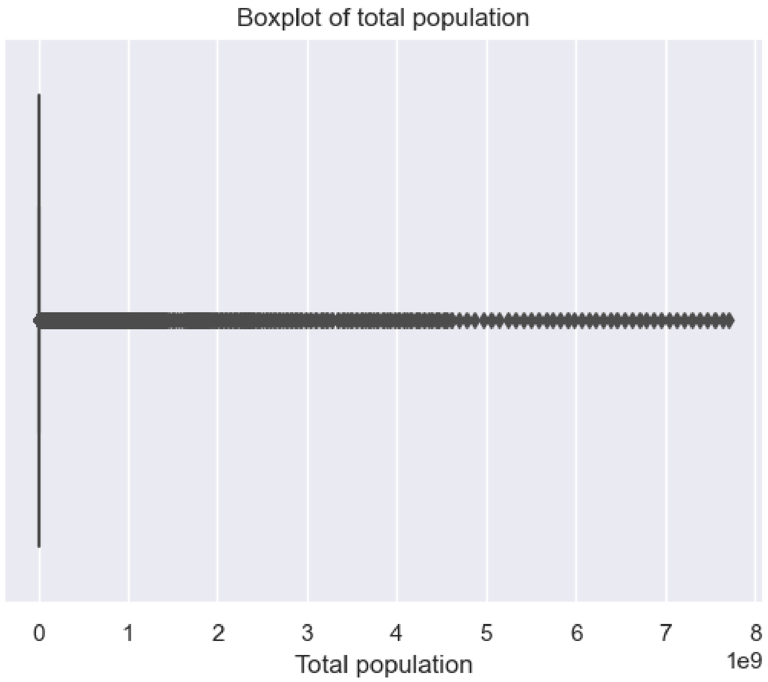


Histogram of Per capita plastic waste

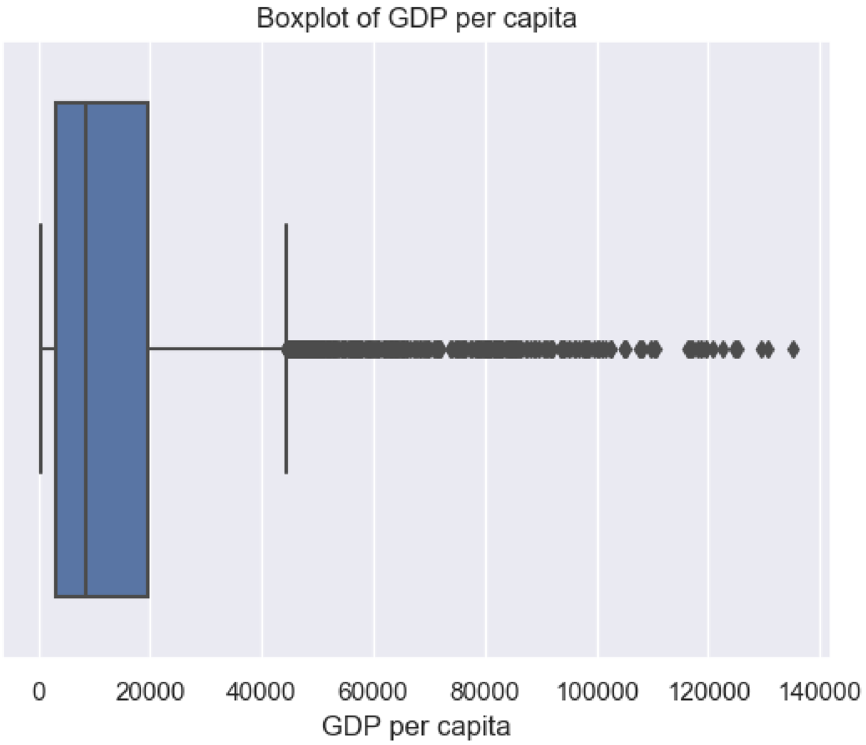
Thanks to boxplot visualisation it has been analysed if the dataset has outliers. Outliers have an impact on the results also if the missing values will be filled with the median, the mean and the mode.



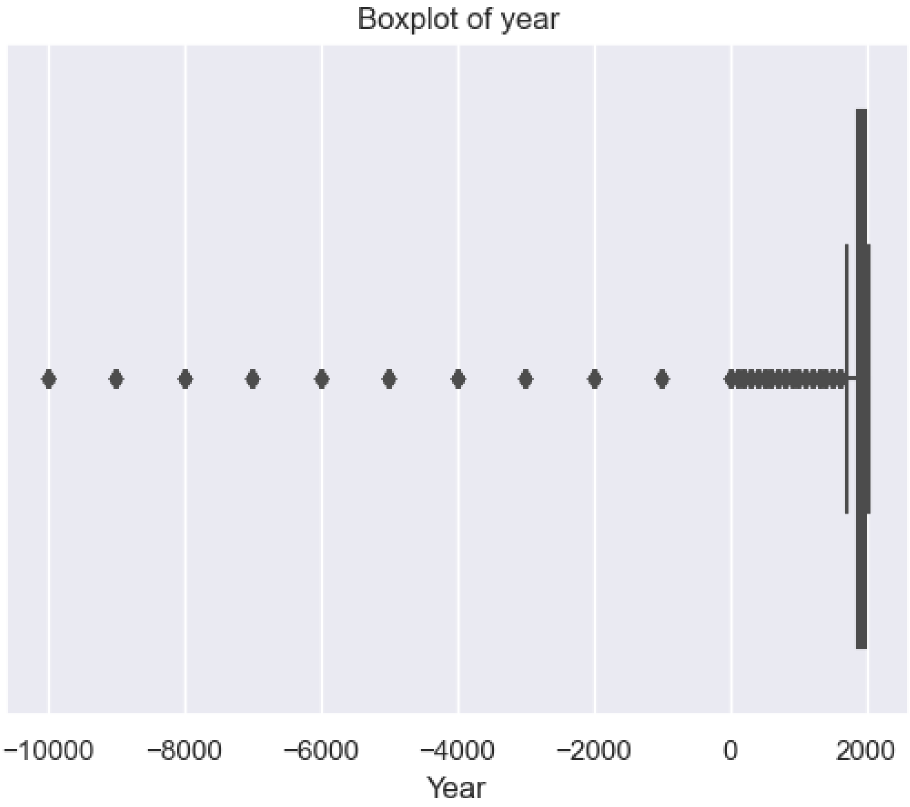
Graphic: Box plot for Outliers of “Capita plastic waste”



Graphic: Box plot for Outliers of “Total Population”



Graphic: Box plot for Outliers of “GDP per capita”

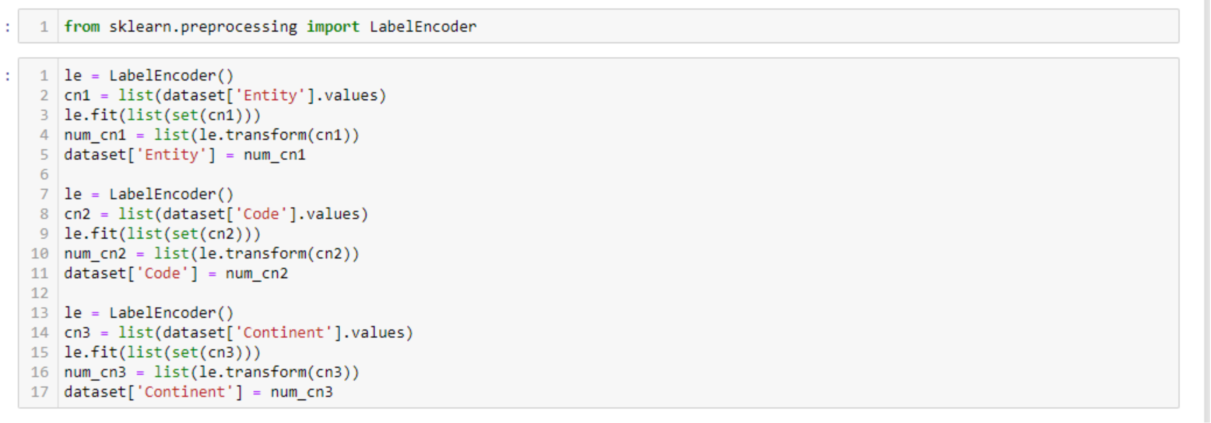


Graphic: Box plot for Outliers of “Year”

EDA and Visualization

Missing values in the dataset can affect the application and accuracy of machine learning models therefore it is crucial to address them before moving on to the next stage.

“Entity”, “Code”, “Continent” are categorical variables in the dataset therefore they need to be transformed into numerical variables with the LabelEncoder function:

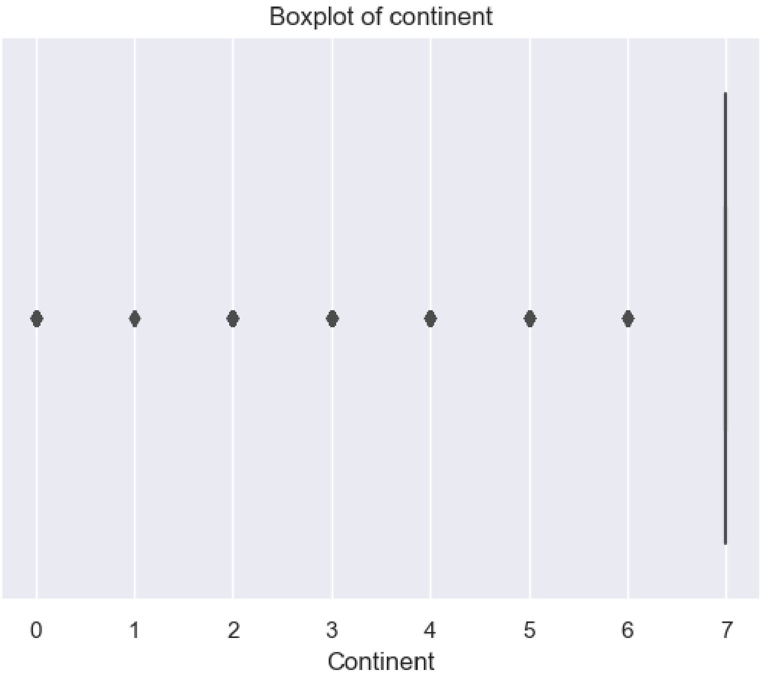


```
from sklearn.preprocessing import LabelEncoder
```

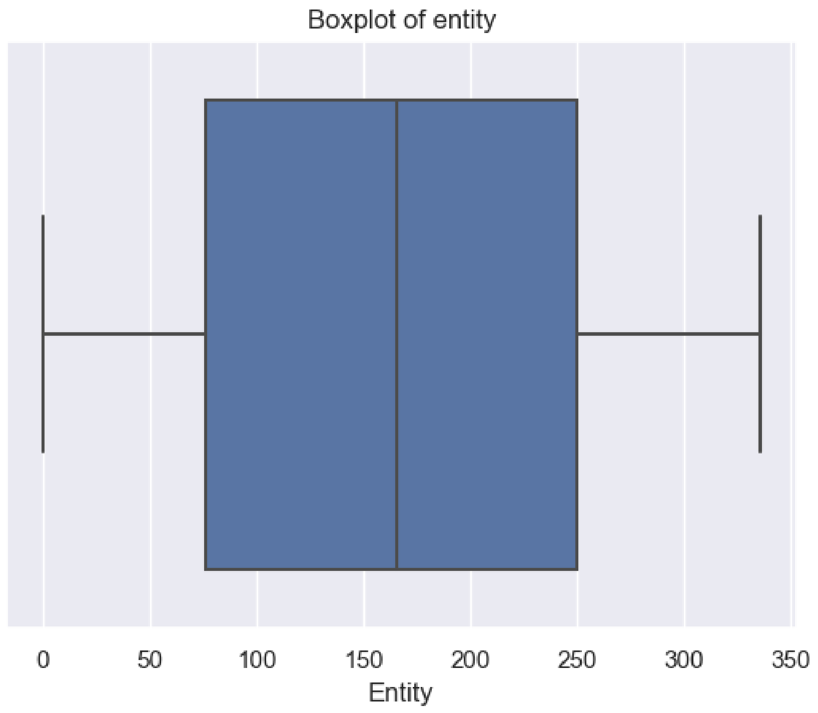
```
le = LabelEncoder()
cn1 = list(dataset['Entity'].values)
le.fit(list(set(cn1)))
num_cn1 = list(le.transform(cn1))
dataset['Entity'] = num_cn1

le = LabelEncoder()
cn2 = list(dataset['Code'].values)
le.fit(list(set(cn2)))
num_cn2 = list(le.transform(cn2))
dataset['Code'] = num_cn2

le = LabelEncoder()
cn3 = list(dataset['Continent'].values)
le.fit(list(set(cn3)))
num_cn3 = list(le.transform(cn3))
dataset['Continent'] = num_cn3
```

Image of the code of encoding

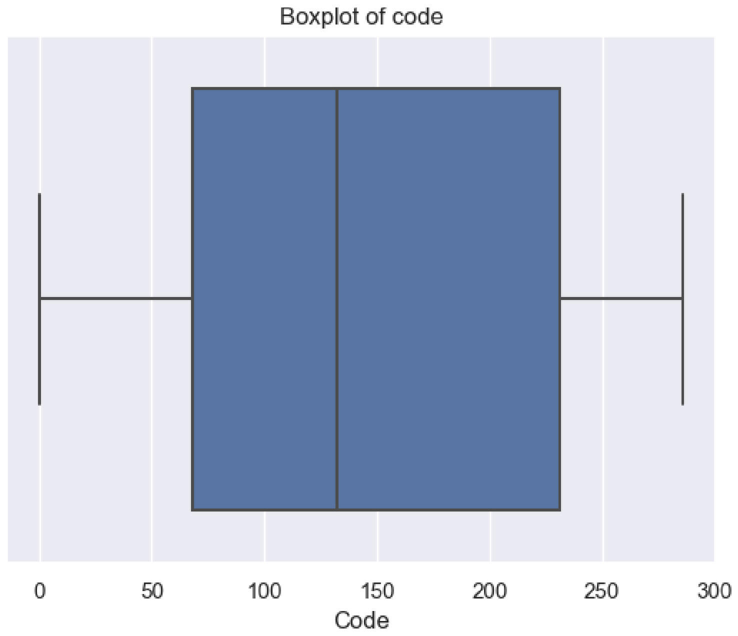
Once they are numerical they can be checked for outliers with the boxplot:



Graphic: Box plot for Outliers of "Continent"

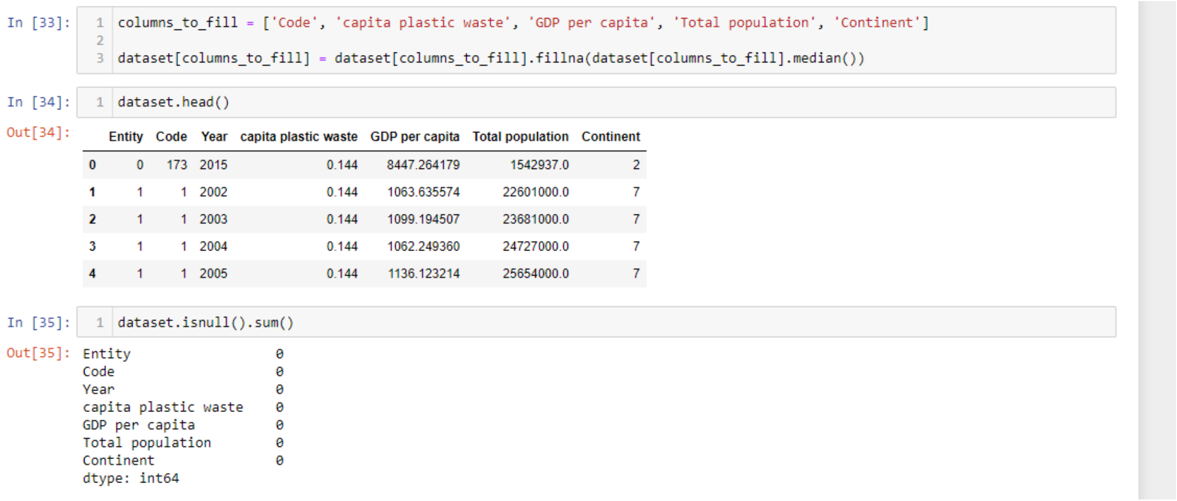


Graphic: Box plot for Outliers of "Entity"



Graphic: Box plot for Outliers of "Code"

Once the outliers have been checked it is time to fill the missing values. The median is less influenced by outliers in comparison to the mode and the mean.The median is a good fit for the binary values within the dataset post-transformation of the categorical value using LabelEncoder. Only the features containing missing variables have been filled. "Year" and "Entity" have zero missing variables so they can remain unchanged.



```
In [33]: columns_to_fill = ['Code', 'capita plastic waste', 'GDP per capita', 'Total population', 'Continent']

         dataset[columns_to_fill] = dataset[columns_to_fill].fillna(dataset[columns_to_fill].median())
```

```
In [34]: dataset.head()
```

Out[34]:

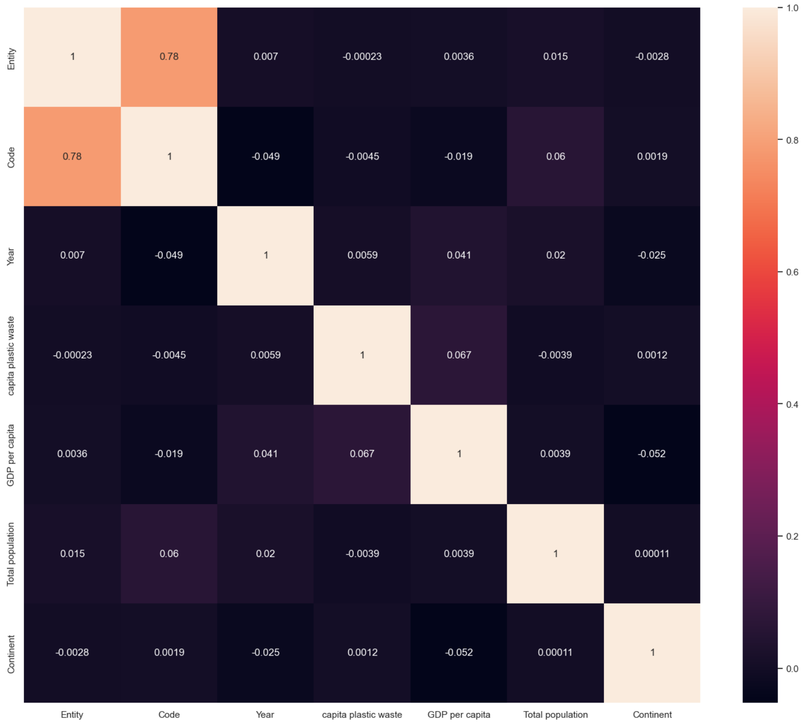
	Entity	Code	Year	capita plastic waste	GDP per capita	Total population	Continent
0	0	173	2015	0.144	8447.264179	1542937.0	2
1	1	1	2002	0.144	1063.635574	22601000.0	7
2	1	1	2003	0.144	1099.194507	23681000.0	7
3	1	1	2004	0.144	1062.249360	24727000.0	7
4	1	1	2005	0.144	1136.123214	25654000.0	7

```
In [35]: dataset.isnull().sum()

Out[35]: Entity                  0
         Code                    0
         Year                    0
         capita plastic waste    0
         GDP per capita          0
         Total population        0
         Continent               0
         dtype: int64
```

Image of the code to fill the missing values

Model discussion:

The clean dataset is now ready to be tested with Machine learning models. The target variable is “Capita plastic Waste” due to the domain nature. To see which input variables align well for the machine learning accuracy it has been utilised the correlation matrix to examine their relationship. Positive values closer to 1 indicate a strong positive linear relationship, negative values closer to -1 indicate a strong negative linear relationship, and values closer to 0 indicate a weaker or no linear relationship. This helps understand how strong the correlation is and this is crucial for the modelling, in particular to identify factors or features inside the dataset.



Graphic: Correlation Matrix

It has been considered as target variable “Capita Plastic Waste” and as input variable Continent:

X= Capita plastic waste
y= Continent.

The machine learning models considered based on the nature of the problem are:

- Decision tree
- Random forest

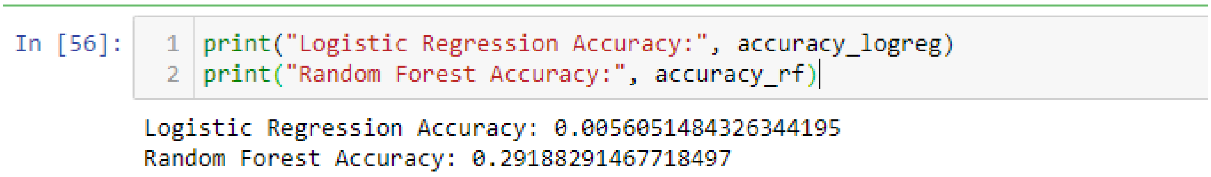
The reason why is because this is a classification problem and the simplicity of the interpretation and the accurate prediction are suitable with the project goals.

To avoid overfitting and underfitting the dataset has been splitted in 3 test: 10%, 20 %, 30%:

	Random Forest		**Decision Tree**	
	Accuracy	**Accuracy**	**Accuracy**	**Accuracy**
10% split test	1.0	1.00	0.9929416649	0.99
20% split test	1.0	1.00	0.99387585634	0.99
30% split test	1.00	1.00	0.99397965538	0.99

Table of Machine learning models results for Plastic capita waste and continent

Also due to the ultimate goal of the project that is to bring awareness it has been tested also the input variable Year. It has been evaluated the Logistic regression and the random forest classifier to see the accuracy score that are:



```
In [56]:  print("Logistic Regression Accuracy:", accuracy_logreg)
           print("Random Forest Accuracy:", accuracy_rf)

           Logistic Regression Accuracy: 0.0056051484326344195
           Random Forest Accuracy: 0.29188291467718497
```

Image of the code logistic regression and random forest accuracy

	Accuracy score in percentage
Logistic Regression Accuracy	0.56%
Random Forest Accuracy	29.28%

Table of content for machine learning models results for Capita plastic waste and Year

Conclusion:

The dataset aims to predict capita plastic waste to bring awareness. The target variable has been tested with two input variables: Continent and Year.

The classification problem of predicting the plastic waste based on the continent was explored with three different splits. Decision tree and random forest models were applied and give 100% accuracy; the model can predict the plastic waste for the continent.This outcome holds different possible solutions like:

- Tailor sustainable techniques based on the areas
- Areas recycle solutions.

Differents question can be address:

- Why is there more plastic waste in one continent than others?
- What is the reason why some continents have more plastic?
- Could it be a behavioural problem?

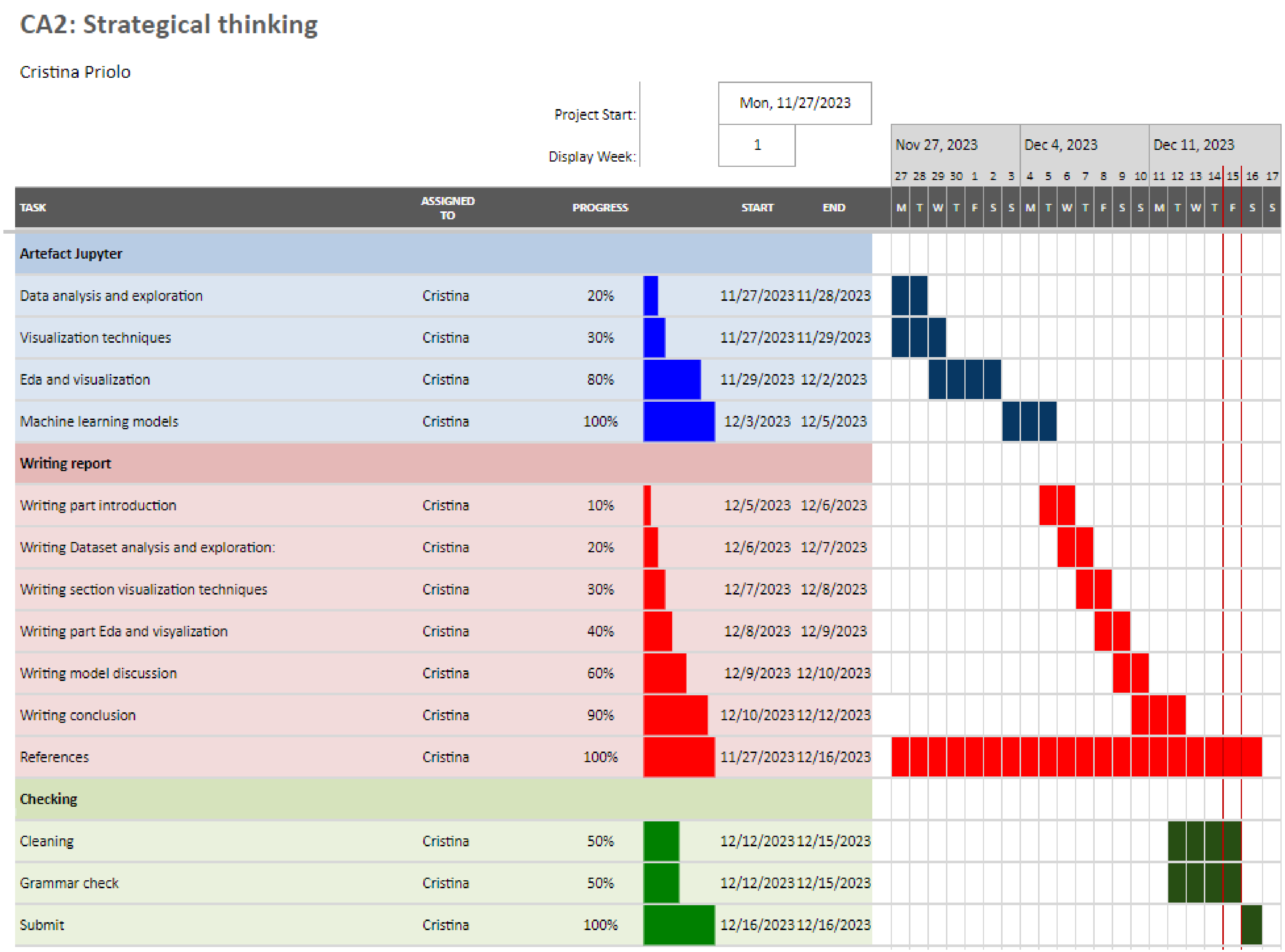
However it is crucial to note that biases might influence this analysis like:

- The given dataset is based on restricted selected data
- Personal opinion
- Local education and laws

Bringing awareness for the future consequences is the ultimate goal of the project so that's why also the Year has been tested with logistic regression and random forest. The accuracy results are low which is a challenge. Future improvements may include expanding the analysis across multiple datasets, incorporating additional variables.

This would give an overview to the readers and possible stakeholders - such as recyclable companies and sustainable advocates - to mitigate the potential future repercussions.

Gantt Chart:



CA2: Strategical thinking

Cristina Priolo

Project Start: Mon, 11/27/2023

Display Week: 1

TASK	ASSIGNED TO	PROGRESS	START	END	Nov 27, 2023							Dec 4, 2023							Dec 11, 2023						
					27	28	29	30	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
					M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S
Artefact Jupyter																									
Data analysis and exploration	Cristina	20%	11/27/2023	11/28/2023	🟦	🟦																			
Visualization techniques	Cristina	30%	11/27/2023	11/29/2023	🟦	🟦	🟦																		
Eda and visualization	Cristina	80%	11/29/2023	12/2/2023			🟦	🟦	🟦	🟦															
Machine learning models	Cristina	100%	12/3/2023	12/5/2023							🟦	🟦	🟦												
Writing report																									
Writing part introduction	Cristina	10%	12/5/2023	12/6/2023									🟥	🟥											
Writing Dataset analysis and exploration:	Cristina	20%	12/6/2023	12/7/2023										🟥	🟥										
Writing section visualization techniques	Cristina	30%	12/7/2023	12/8/2023											🟥	🟥									
Writing part Eda and visyalization	Cristina	40%	12/8/2023	12/9/2023												🟥	🟥								
Writing model discussion	Cristina	60%	12/9/2023	12/10/2023													🟥	🟥							
Writing conclusion	Cristina	90%	12/10/2023	12/12/2023														🟥	🟥	🟥					
References	Cristina	100%	11/27/2023	12/16/2023	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
Checking																									
Cleaning	Cristina	50%	12/12/2023	12/15/2023																🟩	🟩	🟩	🟩		
Grammar check	Cristina	50%	12/12/2023	12/15/2023																🟩	🟩	🟩	🟩		
Submit	Cristina	100%	12/16/2023	12/16/2023																				🟩	

Project plan gantt chart

GitHub Link:

https://github.com/Sba23037/Hdip_StrategicThinking_CA2_CristinaPrioloSBA23037.git

References:

1. www.kaggle.com. (n.d.). Global Plastic Pollution. [online] Available at: https://www.kaggle.com/datasets/sohamgade/plastic-datasets.

2. Datopian (n.d.). Daily_csv plastic monkey 78. [online] DataHub. Available at: https://www.datahub.io/gitchenze/daily_csv-plastic-monkey-78 [Accessed 15 Oct. 2023].

3. Our World in Data. (n.d.). Extrapolated change in plastic fate. [online] Available at: https://ourworldindata.org/grapher/plastic-fate-to-2050.

4. Oa, A. (2019). Public and Environmental Health Effects of Plastic Wastes Disposal: A Review. clinmedjournals.org, [online] 5(1). doi:https://doi.org/10.23937/2572-4061.1510021.

5. Borrelle, Stephanie B., et al. “Predicted Growth in Plastic Waste Exceeds Efforts to Mitigate Plastic Pollution.” Science, vol. 369, no. 6510, 18 Sept. 2020, pp. 1515–1518, https://doi.org/10.1126/science.aba3656

6. Avinash Navlani. “Random Forests Classifiers in Python.” DataCamp Community, 2018, www.datacamp.com/community/tutorials/random-forests-classifier-python.

7. Avinash Navlani. “Decision Tree Classification in Python.” DataCamp Community, 2018, www.datacamp.com/community/tutorials/decision-tree-classification-python.
8. javaTpoint. “Machine Learning Decision Tree Classification Algorithm - Javatpoint.” Www.javatpoint.com, 2021, www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.
9. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
10. Jaiswal, Abhishek Sheshnath . Importance of Data Exploration in Data Analysis a Review Paper. 2 Apr. 2022, p. 1.
11. “Pyplot Tutorial — Matplotlib 2.0.2 Documentation.” Matplotlib.org, matplotlib.org/2.0.2/users/pyplot_tutorial.html.
12. “What Are Data Types and Why Are They Important?” Amplitude, amplitude.com/blog/data-types#integer-int. Accessed 11 Dec. 2023.
13. Yennhi95zz. “The Importance of Outlier Detection in Machine Learning: Methods and Implementation in Python.” Medium, 21 Apr. 2023, medium.com/@yennhi95zz/the-importance-of-outlier-detection-in-machine-learning-methods-and-implementation-in-python-125e3d5ada7d.
14. “Simple Gantt Chart.” Vertex42.com, www.vertex42.com/ExcelTemplates/simple-gantt-chart.html?utm_source=ms&utm_medium=file&utm_campaign=office&utm_content=url.
15. Nautiyal, Dewang. “Underfitting and Overfitting in Machine Learning.” GeeksforGeeks, 23 Nov. 2017, www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/.

16. javaTpoint. “Machine Learning Decision Tree Classification Algorithm - Javatpoint.” Www.javatpoint.com, 2021, www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.
17. “Random Forest Classifier Giving 100% Accuracy on the Test Split.” Stack Overflow, stackoverflow.com/questions/76643983/random-forest-classifier-giving-100-accuracy-on-the-test-split. Accessed 16 Dec. 2023.