MTHM501 Assessment Report

Kevin Donkers (EI CDT) - 700063874

16/11/2020

1 Introduction

The United Nations (UN) have Sustainable Development Goals, which are used to guide human development in a sustainable way. One of thse goals is to reduce waste[1], while another is to increase human prosperity and economic grown[2]. A measure of economic growth is Gross Domestic Product (GDP).

Might these two goals work against each other? We can explore this by asking the question: What is the relationship between gross domestic product (GDP) and household waste?

One theory would be that increased GDP is linked to an increase in consumerism, and so a positive correlation would be expected. On the other hand, high GDP usually means a country has more money to spend on infrastructure, hence a higher recycling rate would be expected and thus a negative correlation would be expected.

In addition, how might a country's population factor into this? Do countries with larger populations produce more waste per person, or less?

United Nations Statistics Division (UNSD)[3] collects data from member states. This includes data on total municipal waste generation, GDP per capita and total population. We will use these datasets to explore the questions posed above.

2 Objectives

The aim of this investigation is to determine whether there is a relationship between national GDP and municipal waste generation. After obtaining data from the UNSD, we must tidy the data. This involves combining multiple datasets and processing variable so they are comparable. Next we should visualise each variable to determine how values are distributed, both statistically and geospatially. The datasets being merged do not all cover the same set of countries. Therefore the merged dataset will contain some missing values. This is addressed by two methods: removal and multiple imputation. Linear models are fitted to both "corrected" datasets to determine whether GDP and population have an effect on municipal waste generation, and as an evaluation of the missingness methods. Limitations of the approach are discussed and final analyses presented in the conclusion.

[1] Sustainable Development Goal 12 - https://sdgs.un.org/goals/goal12
[2] Sustainable Development Goal 8 - https://sdgs.un.org/goals/goal8
[3] UNSD data portal - https://data.un.org/

3 Data

3.1 Data sources

UNSD provides datasets open to the public on various topics related to international development, economic metrics and trade (see 7 below). For this investigation data on total municipal waste generation and GDP were needed.

Total municipal waste generation was available for the years 1990-2016, aggregated by year and country. GPD per capita data was available for the years 1970-2018, aggregated by year and country. Data for the years 1989-2018 were downloaded for this investigation.

To allow fair comparison between waste generation and GDP, the waste data was processed to represent municipal waste generated per capita (in tonnes per capita). To calculate this we needed population data for each of the corresponding countries and years.

Population data was available from UNSD for the years 1948-2020, aggregated by year and country. Data for the years 1990-2020 were downloaded for this investigation.

All data was downloaded as Comma Separated Values (CSV) format. Municipal waste per capita was calculate by dividing the total municipal waste by the total population, for each year and country in the dataset.

3.2 Data variability

We can see from the table below that the number of countries covered by each dataset varies, as does the number of data point each one contains.

Table 1: Summary of datasets

Dataset	Number of countries	Years	Number of values	Max	Min	Unit
Municipal waste	128	1990-2016	1,587	4.64	234,471	1000s tonnes
GDP per capita	220	1989-2018	6,317	79.95	189,162	$USD
Population	239	1990-2020	8,022	44	1,398,000,000	

This reflects the variability in data collection when working at an internation level, since data is volunteered by governments to the UN, rather than it being collected by a central system.

Given that municipal waste generation is our response variable, it makes sense to retain the number of entries in this dataset and match the corresponding entries for countries and years from the GDP and population datasets. The consequences of this are discussed in Sections 3.3 and 3.4.

To understand how each dataset is distributed we can plot the distributions of the values and their geographical spreads.

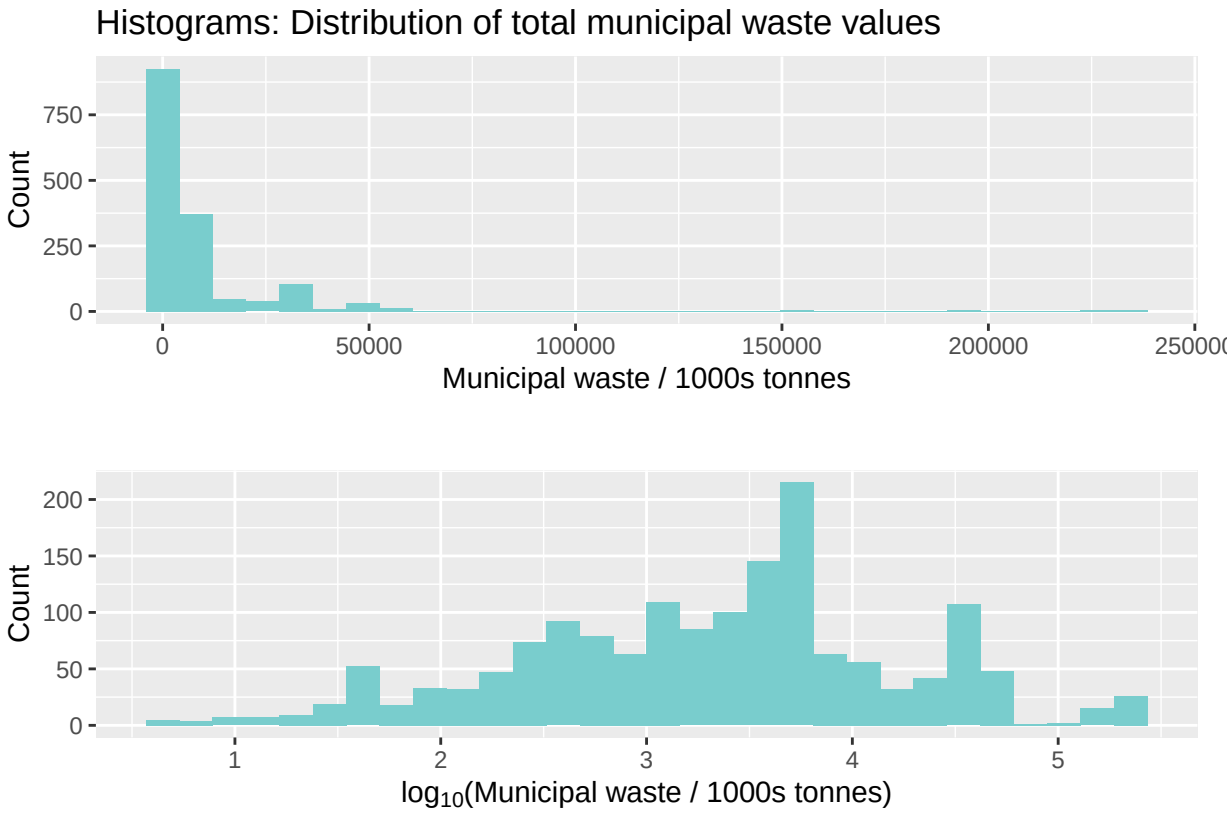

Figure 1: Distribution of municipal waste values. A logarithm of the data appears to have a more normal distribution.

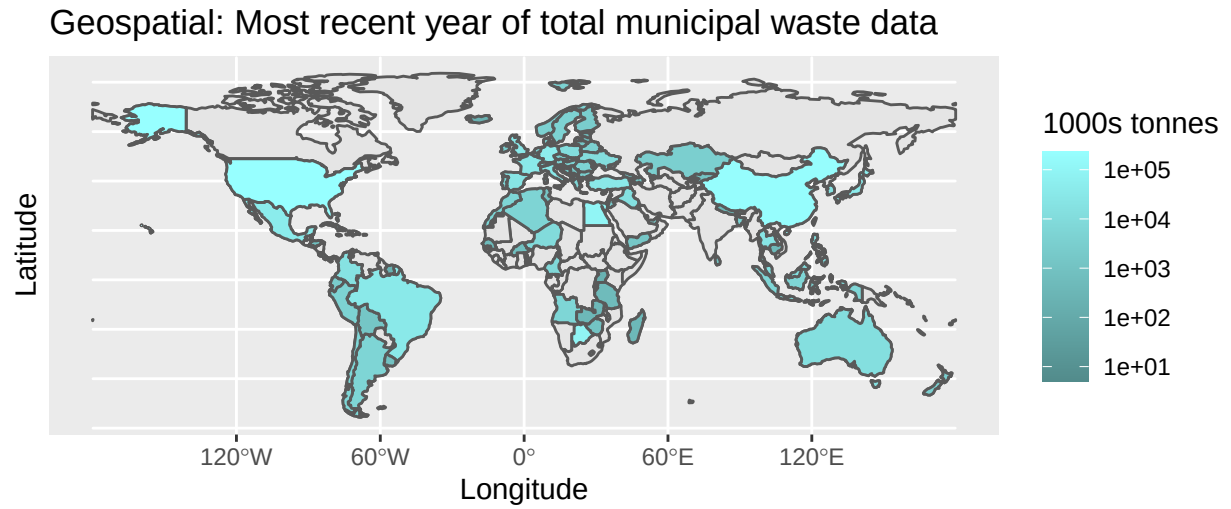

Figure 2: Geographical distribution of municipal waste values, using most recent year of data for each country in the dataset.

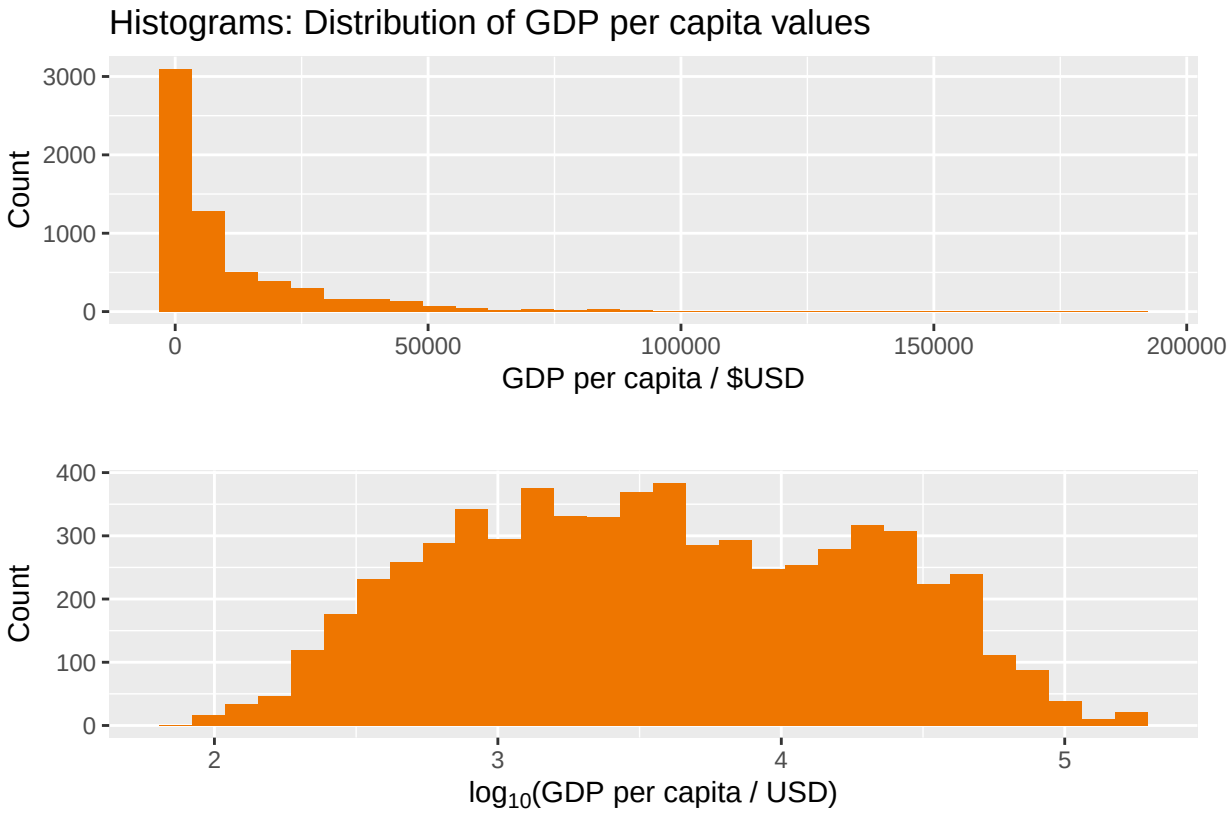

Figure 3: Distribution of GDP per capita values. A logarithm of the data appears to have a more normal distribution.

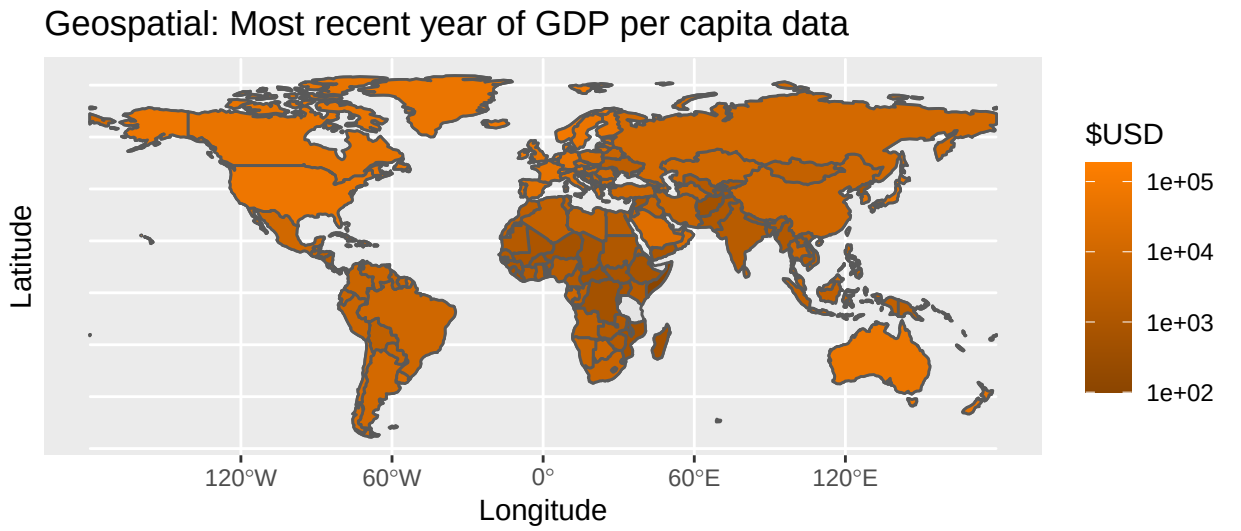

Figure 4: Geographical distribution of GDP per capita values, using most recent year of data for each country in the dataset.

We can see from Figures 1, 3 and 5 that the data are more normally distributed when done so logarithmically. Therefore we should use log transforms of the data when fitting linear models.

We can see from Figures 2, 4 and 6 that each of the datasets don't cover exactly the same set of countries. It is also the case that each of the datasets also don't cover the same number of years for each country. We will explore this in Section 3.4 below.

3.3 Data merging

This investigation is trying to understand the variation of municipal waste generation with GDP and possibly population. Therefore we are treating municipal waste generation as a response variable, and GDP and population as explanatory variables. We can merge our datasets together, but as we saw in above 3.2 each dataset does not cover the same number of countries. Given municipal waste is our response variable, it was chosen to keep all the data from this dataset and merge on only matching countries and years from the other two datasets.

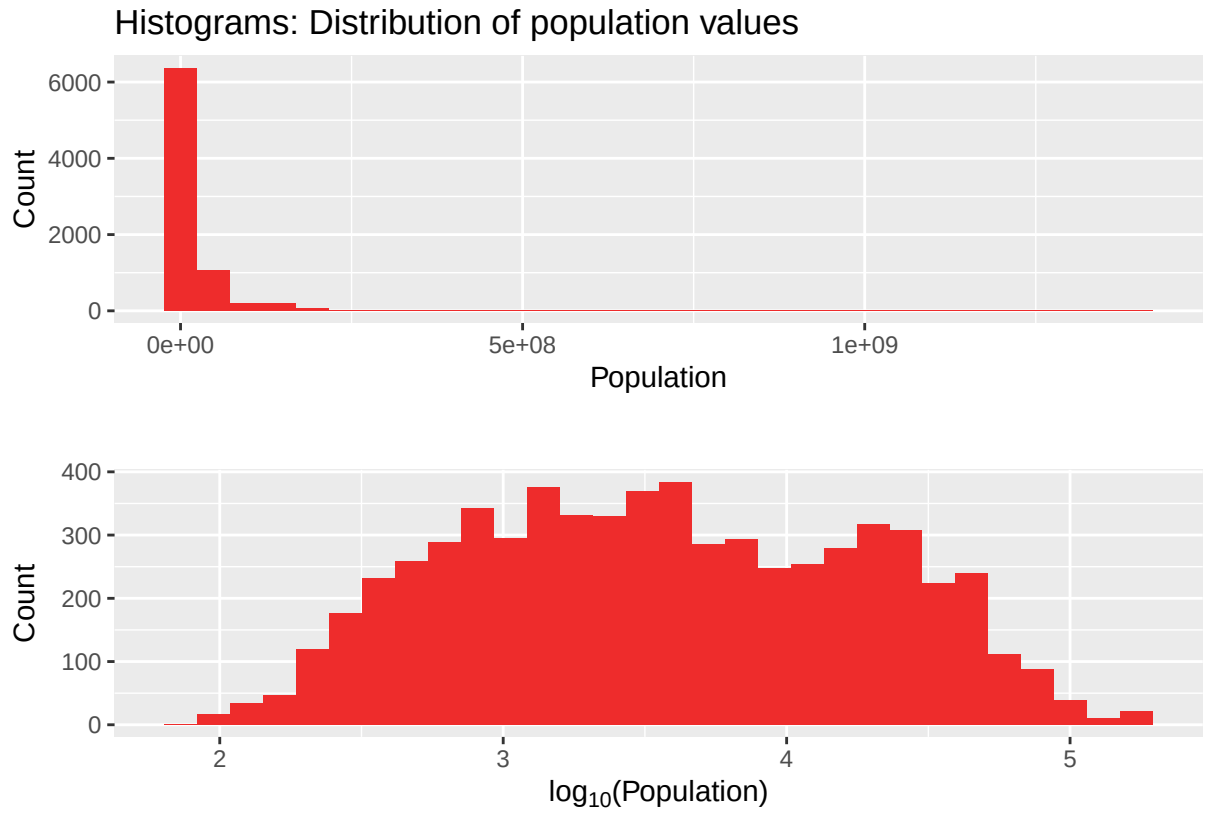

Figure 5: Distribution of population values. A logarithm of the data appears to have a more normal distribution.

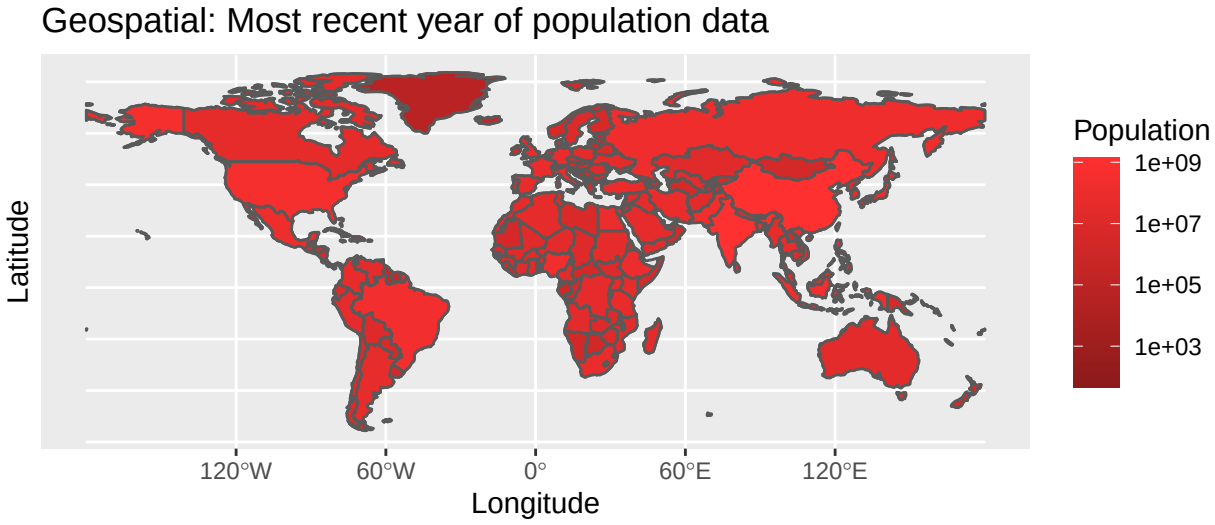

Figure 6: Geographical distribution of population values, using most recent year of data for each country in the dataset.

3.4 Missing data

A consequence of merging datasets, as outlined above, was that some of the values for GDP and/or population for some combinations of country and years were missing.

The reasons for these missing entries is that, despite the municipal waste data containing data for fewer countries than GDP or population, not every country in the UN was included in those datasets for every year covered. It is unclear why this is the case. Given that the data is voluntarily submitted to the UN, some of it could be explained as Missing Completely At Random (MCAR) given that some countries could simply have missed the submission deadline for particular years or a country wasn't aware of it's obligations to report for a number of years e.g. an new/young UN member state.

One proposed strategy for dealing with these missing data values is to simply remove any entries which have a missing value for any of the variables. This leads to a ~2% smaller dataset, as seen in Figure 7.

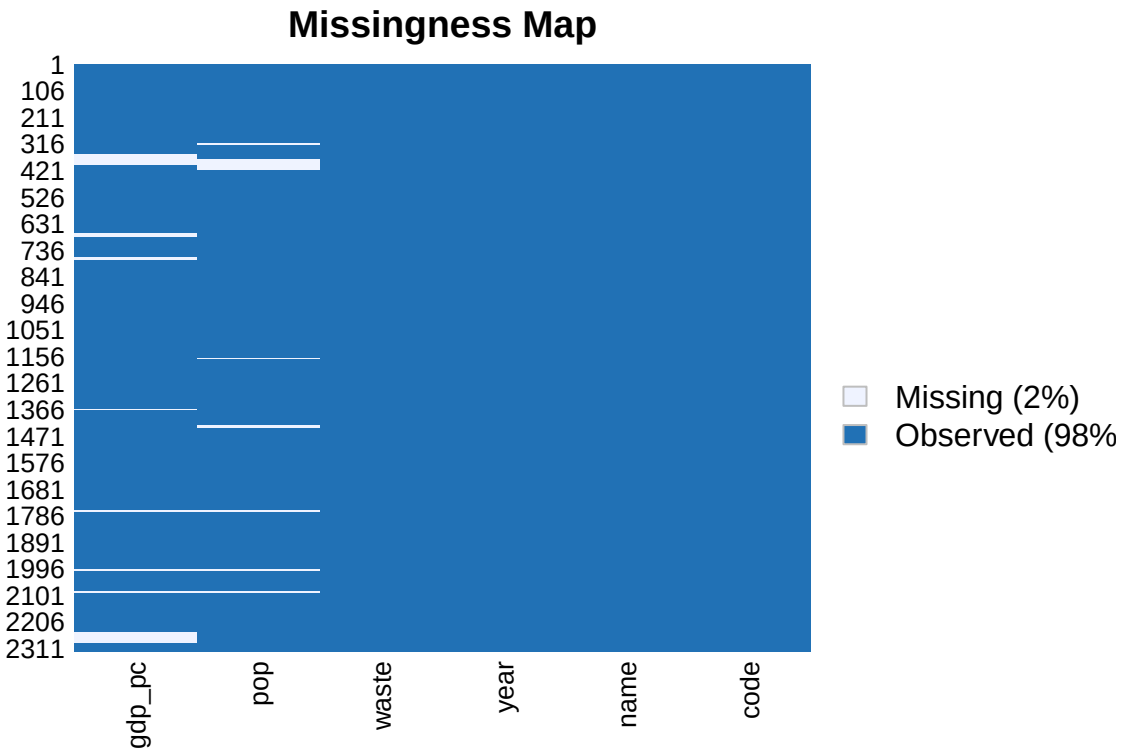

Figure 7: This graphic highlights the missing values in the merged datset. There are 142 missing values in the GDP per capita variable (`gdp_pc`) and 89 missing values in the total population variable (`pop`).

Another method would be to use multiple imputation to generate multiple possible complete dataset by sampling (the logarithm of) the non-missing values for each variable. This was conducted using the `Amelia` package to create ten imputed datasets. The benefit of this was that all values of the response variable, total waste, could be retained and converted to waste per capita given the now complete number of population values. The multiple imputed datasets allowed a level of uncertainty to be calculated for the resulting linear models.

For this analysis, 150 imputed datasets were generated in order to have a number of data sets simliar to the number of missing values in any one variable (GDP per capita has 142 missing values).

4 Analysis and Results

Linear models were generated using the `lm` function. A model was generated for the dataset with missing entries removed, and another for the dataset which imputed missing values using multiple imputation.

Both linear models try to fit the relatonship `log(waste_pc) ~ log(gdp_pc) + log(pop) + log(gdp_pc):log(pop)`.

4.1 Linear model with missing entries removed

Below is a table outlining the estimates for the linear model intercept, coefficients and their corresponding standard errors.

Table 2: Linear model coefficient and standard errors, for dataset with missing entries removed.

	Intercept	`log(gdp_pc)`	`log(pop)`	`log(gdp_pc):log(pop)`
Estimate	5.934065	-0.543428	-0.686317	0.060929
Standard error	0.802349	0.080042	0.050511	0.005061

We can see that waste per capita has a negative correlation with both GDP per capita and population, the latter having the stronger negative correlation. The interaction with GDP per capita and population is an order of magnitude smaller than the main correlation, and it is positive rather than negative.

The standard errors for these estimates are ok but could definitely be improved.

The adjusted R^2 value was 0.5227, which is not a particularly good fit to the data points.

4.2 Linear model with missing entries removed

Below is a table outlining the estimates for the linear model intercept, coefficients and their corresponding standard errors, aggregated from fitting a linear model to the 150 imputed datasets.

Table 3: Linear model coefficient and standard errors, for multiple imputation dataset.

	Intercept	`log(gdp_pc)`	`log(pop)`	`log(gdp_pc):log(pop)`
Estimate	4.898117	-0.4513177	-0.634285	0.0564811
Standard error	0.7941083	0.0792634	0.0500456	0.0050143

The aggregation was conducted using `Amelia::mi.meld`.

Multiple imputation generated similar results to the non-imputed linear model (4.1), with reduced magnitude in estimates and marginally improved standard errors.

No R^2 was calculated for the multple imputations.

5 Limitations

5.1 Data

We were able to calculate the waste per capita for each country from the total waste generation and total population, but this calculate assumes that 100% of the population have access to municipal waste management infrastructure and that 100% of the population is surveyed in the initial dataset. To correct for this we can factor in the percentage of the population served by waste management infrastructure, which the UNSD provides[4].

5.2 Missing data

Multiple imputation was conducted to fill in missing values from the whole dataset. However, this may not be the most sensible way to conduct this multiple imputation. Missing values tended to be clustered around particular countries (few countries with missing values for many years). Therefore, sampling for the imputation should be clustered by country, since the intra-country variability is more likely to be normally distributed than inter-country variability.

The UNSD provides it's own guidelines[5] on missing values. It boils down to imputation and extrapolation of time series data, based on how many missing values are in each time series. This could be implemented and compared with multiple imputation.

5.3 Linear models

The low R^2 value for the fitted linear model might have been improved by removing outliers from the data.

5.4 Plotting

Latitude values are missing from the geographical plots. I could not find a way to add them.

6 Conclusion

From the linear models fitted to dataset with varying levels of missingness correction, we can conclude that GDP per capita and total population of a country are negatively correlated with municipal waste generation per capita.

With an R^2 of these linear models could greatly be improved, so drawing strong conclusions from these relations are difficult.

However, if we can rely on these coefficients then they would support the theories that: 1. Higher GDP per capita lead to less waste per capita due to improved waste management infrastructure, as there is more money to support such infrastructure. 2. Larger populations allow more efficient utilisation of waste management infrastructure.

However, it is likely that the relationship is much better explained by incporporating other development indicators or social factors.

[4]UNSD percentage of population served by municipal waste collection - https://data.un.org/Data.aspx?q=waste+population&d=ENV&f=variableID:1878&c=2,3,4,5&s=countryName:asc,yr:desc&v=1

[5]UNSD data aggregation and missing value guidelines - https://uneplive.unep.org/media/docs/graphs/aggregation_methods.pdf

7 Data Sources

UNSD total municipal waste data:

https://data.un.org/Data.aspx?q=waste+datamart%5BENV%5D&d=ENV&f=variableID%3a1814&c=1,2,3,4,5&s=countryName:asc,yr:desc&v=1

UNSD GDP per capita data:

https://data.un.org/Data.aspx?q=gdp&d=SNAAMA&f=grID%3a101%3bcurrID%3aUSD%3bpcFlag%3atrue%3byr%3a1989%2c1990%2c1991%2c1992%2c1993%2c1994%2c1995%2c1996%2c1997%2c1998%2c1999%2c2000%2c2001%2c2002%2c2003%2c2004%2c2005%2c2006%2c2007%2c2008%2c2009%2c2010%2c2011%2c2012%2c2013%2c2014%2c2015%2c2016%2c2017%2c2018&c=1,2,3,5,6&s=_crEngNameOrderBy:asc,yr:desc&v=1

UNSD population data:

https://data.un.org/Data.aspx?q=population+datamart%5BENV%2CPOP%5D&d=POP&f=tableCode%3a1%3brefYear%3a1990%2c1991%2c1992%2c1993%2c1994%2c1995%2c1996%2c1997%2c1998%2c1999%2c2000%2c2001%2c2002%2c2003%2c2004%2c2005%2c2006%2c2007%2c2008%2c2009%2c2010%2c2011%2c2012%2c2013%2c2014%2c2015%2c2016%2c2017%2c2018%2c2019%2c2020&c=1,2,3,6,8,10,12,13,14&s=_countryEnglishNameOrderBy:asc,refYear:desc,areaCode:asc&v=1