CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

Module Title:	Analytical Methods for Big Data
Assessment Title:	Repeat CA summer 2015
Lecturer Name:	Kislay Raj
Student Full Name:	Angel Thompson
Student Number:	SBA23353
Assessment Due Date:	15 / AUG / 2025
Date of Submission:	15 / AUG / 2025

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.

Table of Contents

The importance of analysing pollution and waste from certain industry sectors:

Industries are an important part of Europe, they supply essential goods while supporting employment. But they also impact the environment heavily with their release of pollutants and waste.

Pollutants can harm human health as well as natural habitats. In major industrial regions, air pollution affects many people through inhalation and heavy metals can contaminate food and water. These industrial emissions are serious threats to ecosystems for animal and plant life, by disrupting natural breeding cycles for animals, and reducing biodiversity, the variety of living organisms like plants. (Anon., 2025)

The above is why sustainability in industries is important. Many sectors are using sustainable products and non-hazardous materials. The EU's industrial sector has changed their strategy for the European green deal, their main focus is having a climate-neutral and cleaner economy. By having ambitions like low pollution and toxic-free waste transfers, future generations won't have to suffer from previous generations' carelessness. (Anon., 2025)

The EEV, the European Environment Agency, records all these releases to inform the public and policy makers, so that we can monitor the top polluting sectors and track progress towards sustainability.

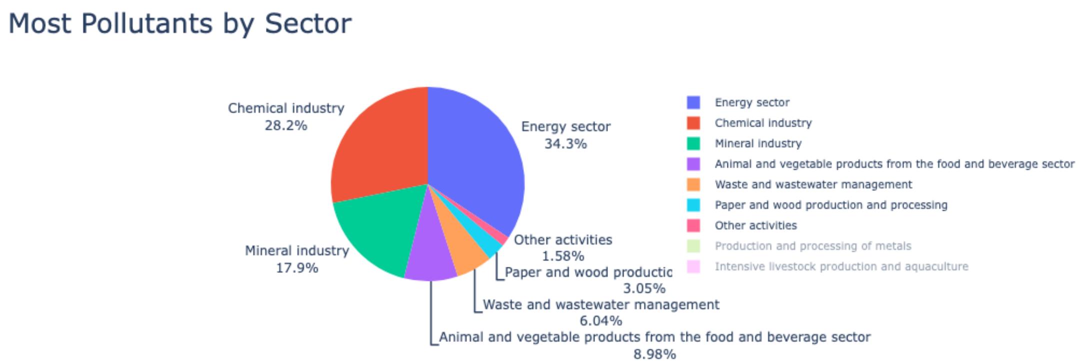
Data Exploration:

Visualising the data was not only incredibly informative and interesting but also showed inconsistensies within the data. There were missing years of data available, like austria seems to not have any data available after 2017 in both pollution and waste, this may have been an issue with the joins; some data may have been lost. There is also a very sudden spike for Italy in 2018 for pollutant transfers, for this reason i have decided to limit my dataset from 2007-2017.

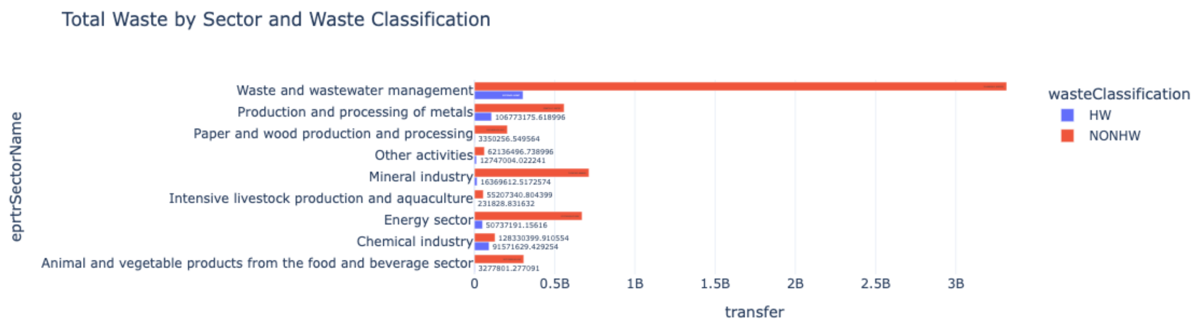
Sudden spikes in data could show things like outliers in normal datasets, but we are dealing with datasets that show very independent countries, which may not have stable waste and pollution releases each year.

Most contaminating industries:

To study the most contaminating industries I used Pie charts for most waste and pollutants, bar charts for highest pollution and grouped bar charts to show the highest waste grouped by hazardous to non-hazardous waste.



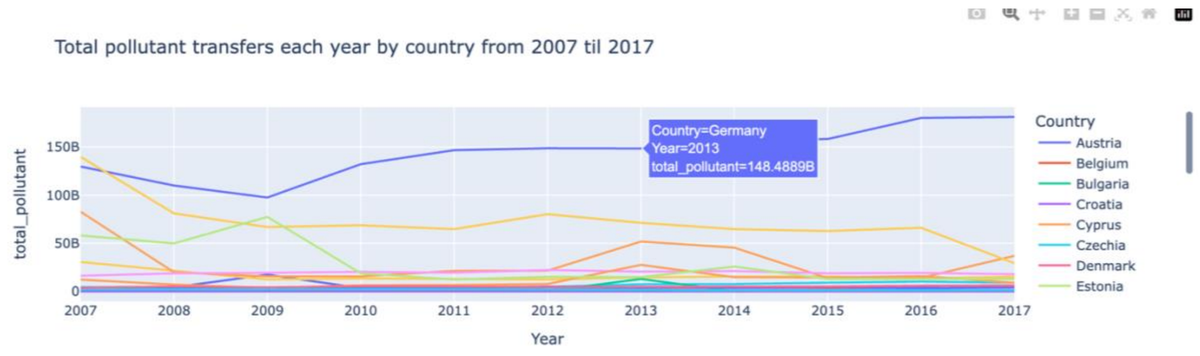
From the above pie chart of most pollutants by sectors, the top three sectors contributing to pollutants were the energy sector with 34.3%, the Chemical industry with 28.2% and the Mineral industry with 17.9%.

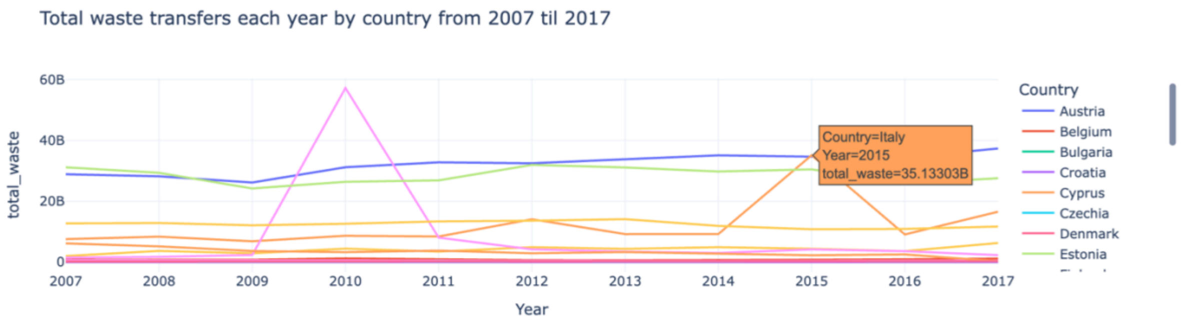


The above grouped horizontal bar chart shows waste water management had the largest waste, although most were non-hazardous, it was still the largest contributer to hazordous waste.

Most Contaminating Countries:

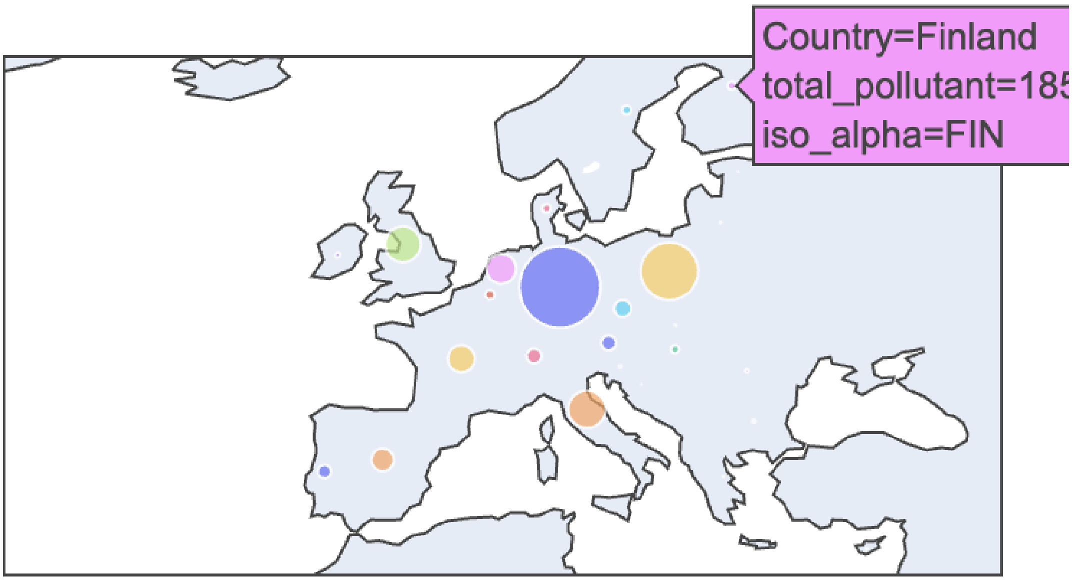
For easier readability, pollution visuals have a grey background, and waste transfer visuals have a white background.





The above graph shows the total waste transfers each year by country, from 2007 til 2017, with closer inspection there are sudden spikes for the Netherlands in 2010 and Italy in 2015.

I decided to combine the years 2007-2017, so that I could visualise each country's waste and pollution over a ten-year period. I named that dataframe wt_10, which stood for waste transfers 10 years. I wanted to use Plotly's Geo plot, which is a map and has data visuals displayed over it. I used the 2007-2017 dataset to apply the iso codes, assigning codes to countries such as Ireland as IRL. I then made that into another dataframe combining the data into a ten-year period, so that I could still visualise Geo plot on a yearly plot.



Looking at data integers versus looking at a visualisation really incapsulates the difference between the countries pollutants. I thought some countries weren’t showing up, but their pollutants were just so small compared to the rest of the countries, example Sweden and finland, referenced in the geo-plotly to the left.

Looking back I wish I created another dataset showing the avg waste and pollution for each country over the ten years, but I wanted to focus on the rest of the report instead of losing myself in the data.

Linear regression

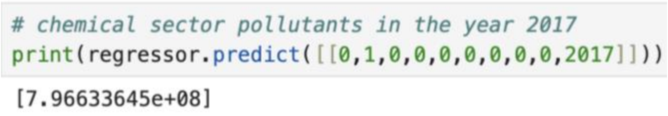
I used a machine learning model called linear regression to study and predict the amount of pollution from each sector by the year. Do certain industries produce more pollution as the years go by, or do they deplete. Which sectors contribute to higher pollution levels.

Routes taken with linear regression.

Conflicting routes were what to do with missing values and whether to use the sector code or to use my own encoder.

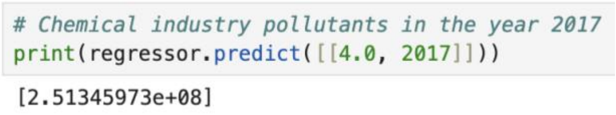
While studying my dataframe, I noticed there were null values after 2016 for waste, after 2017 for pollution. It would've been best to perform my linear regression model on the whole dataset if I wanted to estimate future transfer levels, but I did not want to delete rows that might lead to inconsistencies in some years or sectors. My goal was to analyse the historical patterns, like which sectors were polluting the most now. So I decided to only include the years 2007 to 2016 in my linear regression analysis, even though the prediction model seemed pointless because its prediction of the future was less likely due to the older data. We could use the prediction model to fill in the NaN values, but my report seemed more analytical.

The Linear regression model required only integer values, so it is common practice to encode any category VARCHAR values. I encoded the sector column, with year as a feature and pollution transfers as the column to be predicted. When the model was complete it was time to predict the values, This time consuming task was done by knowing the order of the one hot encoded features for sector. This method was taught to me, but with a smaller dataset, not one with 9 sectors like this one.
Below on the left, is the result of using the built-in one-hot encoder.



```
# chemical sector pollutants in the year 2017
print(regressor.predict([[0,1,0,0,0,0,0,0,0,2017]]))
```

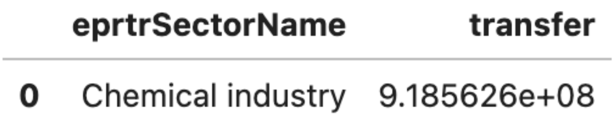
```
[7.96633645e+08]
```



```
# Chemical industry pollutants in the year 2017
print(regressor.predict([[4.0, 2017]]))
```

```
[2.51345973e+08]
```

The use of the column EPRTRSectorCode, shown on the right picture. Instead of using an encoder, I could also use the datasets built in sector code column, which is a numeric column that corresponds to the sector type, its easy to interpret and would work with the model. However with research, I had to consider that the model might treat the sector codes as an ordering, which would not be good because they are category codes. I decided to use both models to see if there was a difference in predicting.
The picture below shows the actual total transfer for the chemical industry in 2007, taken from the original data. So by comparing the two codes methods used to convert categorical data to integer data for linear regression predictions, the built-in one-hot encoder proved to work better with the machine.



	eprtrSectorName	transfer
0	Chemical industry	9.185626e+08

Conclusion and Reflections:

The context and importance of analysing waste and pollutants from sectors and countries.

Industries in Europe are crucial for goods and employement but they can contribute a lot to environmental changes.
Pollutants can affect human health, ecosystems and biodivsit. With new EU Green deals, sustainable practices can result in a cleaner and climate neautral world.

Data exploration and challenges:

The visualisations revealed inconsistences and missing data, most had lacked data post 2017, so I decided to focus on 2007-2017.
The visualisations I showed were pie charts, bar charts, grouped bar charts and geographical plots. I used them to determine the Main polluting industries and the main pollutant and waste transfers by country.

Findings:

The top polluting sectors were Energy with 34%, chemical with 28% and mineral with 18%.
The top polluting and wastful countries were Germany, Poland and The united kingdom