

1. Analysis, Domain, Questions, Plan

1.1 Overview of the Domain

UK statistics on waste show waste generation across the UK is growing rapidly with 177 million tonnes being generated each year in England alone [1]. This project attempts to gain an insight into how that waste is currently being handled. The government is setting **targets to improve efficiencies** in the industry – for example by enforcing a total ban on all food waste being sent to landfill by 2030 [2]. Targets to recycle or reuse 65% of packaging by 2025 have also been set, rising to 75% by 2030. Measures are being implemented to improve the quality of recycling waste being collected from the public.

1.2 Motivation for the Analysis

Local authorities are responsible for household waste collection and disposal services, as well as enforcing legislation and encouraging good waste management practices in their areas. There is reported disparity in waste management systems **across local authorities in England**. There are also varying practices across the country in terms of recycling within households which impacts on recycling. Barr et al. [2] found population density to have a negative and statistically significant effect on the recycling rate and suggests that a possible explanation could be that in densely populated areas the space to store recyclables separately from residual waste is limited.

1.3 Objectives for the Analysis

This project is an analysis of the 'WasteDataFlow' dataset from the data.gov.uk website. The primary objective is to discover if there are trends in terms of the recycling of different groups of waste recycling materials in the UK. The analysis aims to ascertain the **recycling patterns** in the data reported by local authorities between April 2015 and March 2016. With insight into these trends, waste management companies can target areas which are in further need of their services. For example, at local level we can discover strategic locations for a materials recovery facility. The benefit at government level is that waste management performance is being evaluated. The advantage is that the industry can ensure full potential is being met and plans to meet recycling targets are on track.

1.4 Analytical Questions

Some analytical questions this project aims to answer include:

- The project aims to identify trends in the quantities of different material groups and how this varies across the local authorities. The aim is to analyse the quality of recycling per waste type by investigating how much of each material group is **comingled waste** or **segregated waste**.
- Using census information, the project aims to provide an insight into quantities of each waste material being recycled per head of population. An objective is to deduce from the data if there are any local authorities which do not correlate with the general results?
- The project aims to look at **temporal patterns**. Is more of a certain material group being recycled at certain times of the year for instance? Garden waste is an interesting material group to analyse in this respect. For example, are there higher levels of garden waste recycled over the summer period than the winter period?
- Lastly, an interesting **visualisation** could identify where waste is being sent for final processing. A visualisation will be used to gain insight into the recycling destinations for the specific material groups.

1.5 Methodology

The analysis was very much an iterative process and evolved as the project developed. The project can be summarised by two separate methodologies which I will refer to as **Methodology A** and **Methodology B** within this report.

Methodology A:

This was the original strategy and the analysis focuses on a smaller number of authorities. Merging the waste dataset with known authority characteristics enables an ability to gain **richer insights** into the waste management practices. Five authorities were used for the purposes of the analyses. These include two rural authorities, Fenland and Mendips, Luton which is a mix of rural and urban and the Greater London boroughs of Islington and Hammersmith & Fulham. Further explanation into the choice of these authorities is explained in Section 2.1.

Methodology B:

During the development of the project, it became clear that an alternative analysis could be carried out using the entire range of authorities in England. This decision was taken to proceed with this as it was evident that interesting insights could be obtained from this analysis. A factor in changing strategy at this point was due to the time constraints involved in the task of sourcing and merging other local authority characteristics.

2. Analytical Process

2.1 Analysis Methodology

The initial stages of the project involved selecting which features to use in the analysis. The original scope as outlined in Section 1 was to carry out an analysis on **a select number** of authorities. Five authorities were selected as part of this plan and so I could have an appropriate comparison I picked authorities with varying population densities. Hammersmith and Islington with high population densities, Luton with medium, and Fenland and Mendip with low. The analysis would gain in richness by having the opportunity to expand further on the known characteristics of each of these authorities. Details of how the analysis methodologies developed are expanded on in Section 2.3.

2.2 Data Gathering, Wrangling and Transforming

The initial steps involved cleaning and transforming the dataset. Initially the challenges to complete this step arose from the inconsistency in reporting between the local authorities. There were rows of duplicate observations to be filtered. There were wrangling steps to achieve consistency in the observations and to make the dataset compatible for further analysis. I created five separate csv files for my selected authorities and began the wrangling process. This was done using **Trifacta**, a wrangling software. Within this I could upload my dataset and apply a 'wrangling recipe'. The recipe can be created for one csv and subsequently copied to another csv provided the original format is the same.

The Trifacta software had a file size upload limit of 10MB. This led to the manual separation of a small number of authorities into multiple csv's and wrangling them separately. At this stage of the process there was 31 variable columns. The technique involved transforming datasets by applying a Trifacta recipe to multiple csv's. The steps included deriving new separated columns and deleting duplicate rows. Trifacta provided a platform to explore many possible transformations and visualise how these affected the data.

2.3 Transformation and Derivation

Methodology A:

The first step of this stage involved saving the wrangled csv's to a **directory** of files. From this directory, I called each csv (local authority) iteratively into the python algorithm. The algorithm contained a 'for' loop which said that *for* each file in the directory – to create a dataframe to group the values for each of the comingled waste materials (which are mixed dry recyclables). The next part of the algorithm did the same process but for the source segregated materials (segregated in the household or civic amenity centre). The values in the 'comingled waste' and 'segregated waste' **dataframes** were then divided by the population of that authority. The population figure was called from an index using a new 'population' column which I had manually saved in each authority's csv file. Figure 1 below which illustrates the output of the program for two of the five authorities.

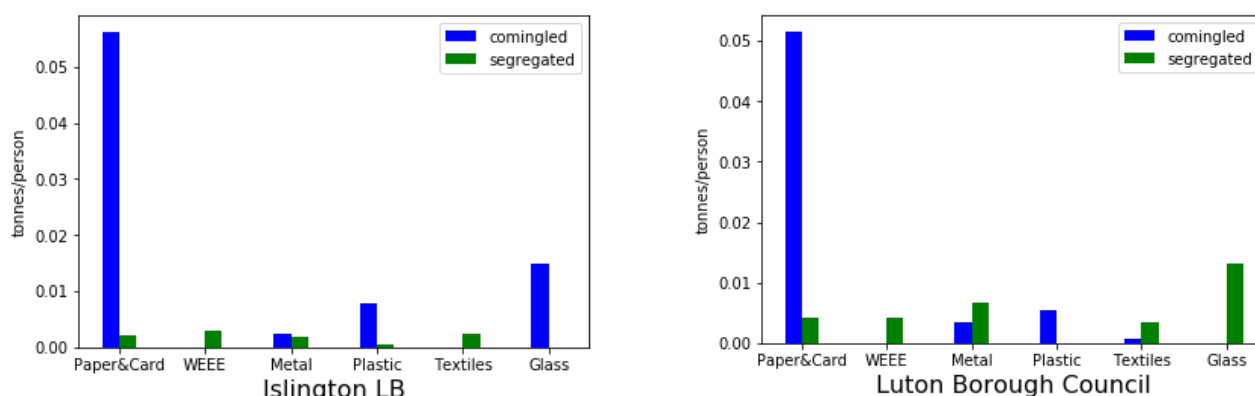


Figure 1. Identifying 'comingled' and 'segregated' rates of waste recycling

Methodology B:

At this point the methodology changed from looking at selected authorities to the entire range in the dataset (i.e. all authorities in England). I reduced the dataset size of 44MB by removing several columns as I was now more aware of how to utilise codes for values within variables. Removing 3 more columns which were redundant for the analysis, which included facility addresses, greatly reduced the file size. Additionally, there were rows which were not relevant to the material groups in the analysis and could be filtered out manually in excel. As these were from other material groups not pertaining to the analysis. This reduced the rows from ~160k to ~100k.

The total size of the file was now below the 10MB limit and could be wrangled in Trifacta. In Trifacta I then created new variables for the analysis. I derived new columns by **transforming** a single variable of categorical values (the individual material groups) into 12 new variables (comingled paper, segregated paper, comingled plastic etc.). Using Pandas, I created 2 dataframes, one with the original dataset and another which was a UK local authority lookup file. Having created this dataframe I could subsequently concatenate data for material recycling that was grouped by authority. I then merged the population data from the look up table and produced the dataframe in Figure 2. Steps 1 to 9 in Appendix A show the sequence of this process in Python. Step 10 shows where I used Matplotlib to illustrate material recycling per person to produce a graph output which is shown in Appendix B.

Index	Authority	ONSCode	POPULATION	Paper	Plastic	Metal	Textiles	Glass	WEEE
0	Adur District Council	E07000223	61182	0.0494518	0.00704047	0.00386633	0.000823608	0.0244526	0.000242718
1	Allerdale Borough Coun...	E07000026	96422	0.0470125	0.00806289	0.00588901	0	0.0258088	0.000195806
2	Amber Valley Borough Coun...	E07000032	122309	0.0436366	0.0080761	0.00576376	0	0.0318813	0.000369883
3	Arun District Council	E07000224	149518	0.0515783	0.00734734	0.00434897	0.000928249	0.0274081	0.000330663
4	Ashfield District Cou...	E07000170	119497	0.0469446	0.00879277	0.00404496	0.000288794	0.0213331	0.000518841
5	Ashford Borough Coun...	E07000105	117956	0.0535013	0.0151271	0.00588677	2.61962e-05	0.0215925	3.44196e-05
6	Aylesbury Vale Distric...	E07000004	174137	0.0584919	0.0116778	0.00442186	0.00133984	0.0258215	0
7	Barking and Dagenham LB	E09000002	185911	0.0158055	0.00475125	0.00545611	0.000649881	0.00572516	0.00516699
8	Barnet LB	E09000003	356386	0.0479763	0.00734891	0.004903	0.00211975	0.016518	0.00265748
9	Barnsley MBC	E08000016	231221	0.0336166	0.00737796	0.00920542	0.00173721	0.0277704	0.00521564
10	Barrow-in-Furness Boro...	E07000027	69087	0.0251195	0.014127	0.00593657	0	0.0254682	0
11	Basildon District Cou...	E07000066	174497	0.0535593	0.0160715	0.00536445	0.00128039	0.02386	0.000746563
12	Basingstoke and Deane Bo...	E07000084	167799	0.0417284	0.00364359	0.00274976	0.00290728	0.0258451	0.00121384
13	Bassetlaw District Cou...	E07000171	112863	0.0442237	0.00654324	0.00381675	0.000958419	0.00753834	0
14	Bath and North East S...	E06000022	176016	0.0530845	0.00759567	0.0115228	0.0020351	0.028148	0.00655974
15	Bedford	E06000055	157479	0.0520078	0.0105529	0.00854087	0.00230076	0.0141585	0.00569524
16	Bexley LB	E09000004	231997	0.0505912	0.015345	0.0121683	0.00590525	0.0193106	0.00397419

Figure 2. Dataframe with feature derivation of material quantities per person

2.4 Analytical Tools and Methods

To summarise these steps in Section 2.3, in Methodology A and B I used Trifacta to initially wrangle data and Pandas to derive new features (which were comingled and segregated tonnes for each material grouped by authority). In A I then used Numpy arrays to organise the quantities before plotting the bar chart with Matplotlib. Whilst in the case of Methodology B I used pandas dataframes to derive new variables and merge between data from my dataset and the local authority lookup table. The code used in **Methodology B** (along with comments) to produce the output is detailed fully in **Appendix A**.

To avoid the size limit there was an option of converting the Trifacta code so it could be used in Python but this was not explored. There are also steps in the code for Methodology B which could have been written more elegantly by using a different type of function in python, such as a 'for' loop for instance. The lines of code in question can be seen in Steps 4, 5 & 6 in Appendix A.

2.5 Incorporating Linear Dimensionality Reduction

Multi Dimensional Scaling (MDS) was implemented to see if there are outliers in the local authorities' recycling rates which could be detected. Figure 3 is a visualisation of the MDS for the data from Figure 2 in Section 2.3 above. 'Sklern' was used to compute the Euclidean Distance which is the **dissimilarity measurement** between authorities. The scales shown in Figure 3 are non-metric representations of the pairwise distance matrix. Galbraith, J. I., et al. [3] state that "the interpretation of any MDS solution must be invariant under reflection, translation and rotation". Step 11 in Appendix A shows the preparation step for this operation while the comments beside Steps 12 to 19 give further details on the implementation of MDS for this data.

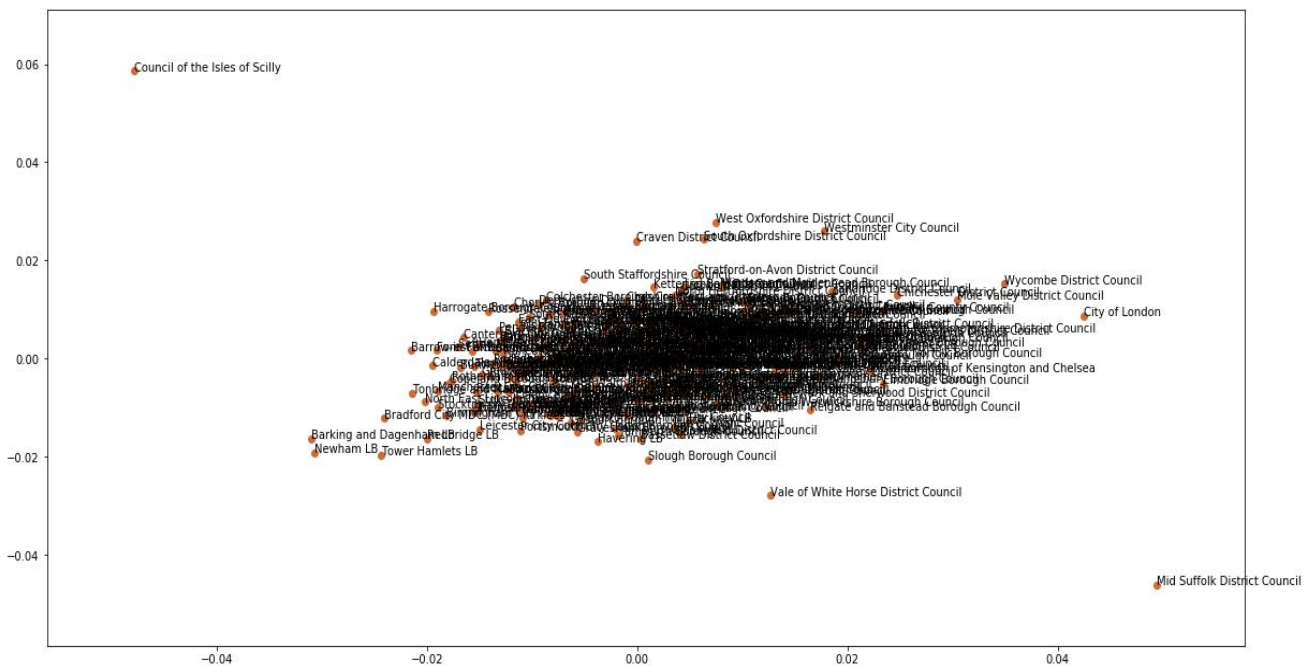


Figure 3. Configuration of authorities from a Two-Dimensional Scaling ratio MDS of the continuous numeric data from Figure 2 in Section 2.3.

2.6 Evaluation on the Approach and Findings

The structure of the dataset was a challenge and required a careful wrangling process. A strategy to speed up the process was to use Trifacta for the removal of duplicate data and aggregation of columns.

Methodology A served the dual purpose of comparing the tonnes of waste recycled per authority whilst also gaining insight into the most utilised waste collection system (commingled or segregated). Figure 1 from Section 2.3 highlights the output that showed authorities were overall **similar** in terms of quantities of **waste recycled per person**. It could also be deduced that Islington had roughly 0.015 tonnes of glass per person in its comingled recycling collection. Luton has a similar rate but in contrast this glass has all been segregated at source.

In the case of Methodology B a bar graph, shown in Appendix B, gave an insight into the patterns of waste recycled throughout England. The findings from this are that aside from a couple of outliers in terms of authorities, recycling rates are quite consistent throughout the country. To further analyse if there were outliers in the local authorities I implemented an MDS algorithm which showed the Council of the Isle of Scully and Mid Suffolk District Council were **dissimilar** to most of the dataset with respect to their quantities of recycling per person per annum.

2.7 Complementary Investigations

As part of the project I engaged in a complementary investigation to build a **holistic understanding** of the dataset. Investigations were made to determine whether patterns varied temporally. So, to complete this temporal analysis the waste material 'garden waste' was judged to be an appropriate material to investigate. Naturally more garden waste will be produced in the Spring and Summer months than the Winter period. And the quarterly data as illustrated in Figure 4 verified this. There is ~3 times as much garden waste recycled between April - June than the quarter of January – March.

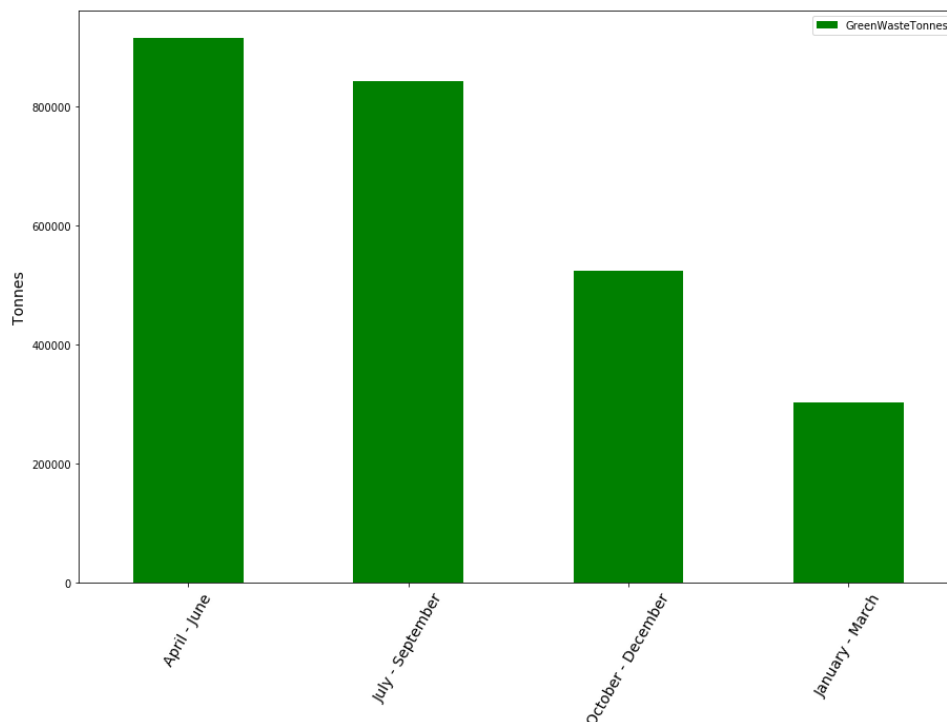


Figure 4. Temporal Trends in the recycling of garden waste.

3. Findings and Reflections

3.1 Critical Evaluation

Methodology A looked at recycling rates and determined that there were disparities in how the recycling was collected. For a small group of authorities (Fenland, Mendip, Islington, Luton and Hammersmith & Fulham) the range of recycling of paper and card for instance varied between 0.04 - 0.06 tonnes per person per annum (between 40 and 60kgs per person). An insight, illustrated in Figure 5, was that the Mendip authority had a fully source segregated waste management system, but also had the highest rates in the group when it came to recycling of paper & card and glass per person. However, there was no reporting of WEEE (Waste Electrical and Electronic Equipment) waste, a fact that highlights a **limitation** of the analysis in Methodology A. One explanation may be that this type of material is recycled within nearby authorities.

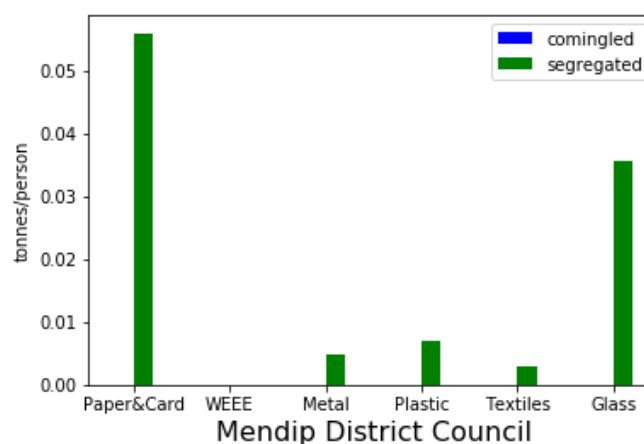


Figure 5. Mendip District Council implement a fully source segregated waste management system

The limitations of Methodology A lead to the wider comparative analysis being carried out in Methodology B. The wrangling process and algorithm developed in Methodology B enabled me to identify patterns on a nationwide scale. The graph contained in Appendix B and the MDS analysis in Figure 3 shows that there are few cases of outliers but that the recycling patterns are consistent. A limitation of Methodology B is that consideration was not given to the fact that energy recovery and incineration play a role in determining recycling rates. This is a **bias** which was outside the scope of the project however.

3.2 Impacts and Uses of the Findings

To assess the impact of my observations in recycling rates I undertook a visual investigation which involved mapping locations of the destinations for each waste stream. To carry this visualisation out I used **Tableau**. Within this software I used a look-up table for postcodes and combined this with my waste dataset postcodes. From there I could visualise the data on a map and filter the tonnes of each material separately. To check that the visualisation is correct I used an opensource platform called 'Mapbox' to underlay a satellite map. By zooming in I can closely analyse a couple of instances and decipher if the location looks like an industrial waste processing facility. The visualisation in Figure 6 illustrates the final destinations for waste in the material group of paper & card.

Although this visualisation does not detail the waste source, we can find out where much of the recycling waste is ending up. A **potential use** of the findings from my project would involve the operators at these facilities analysing the graph as shown in Appendix B and determining predictions for recycling tonnages to be processed by them. It could also be used to optimise locations of these processing facilities. A further study could analyse if recycling materials are travelling long distances and whether this long-distance transport outweighs any environmental benefit.

Tableau Visualisation of Paper Recycling Processing Locations

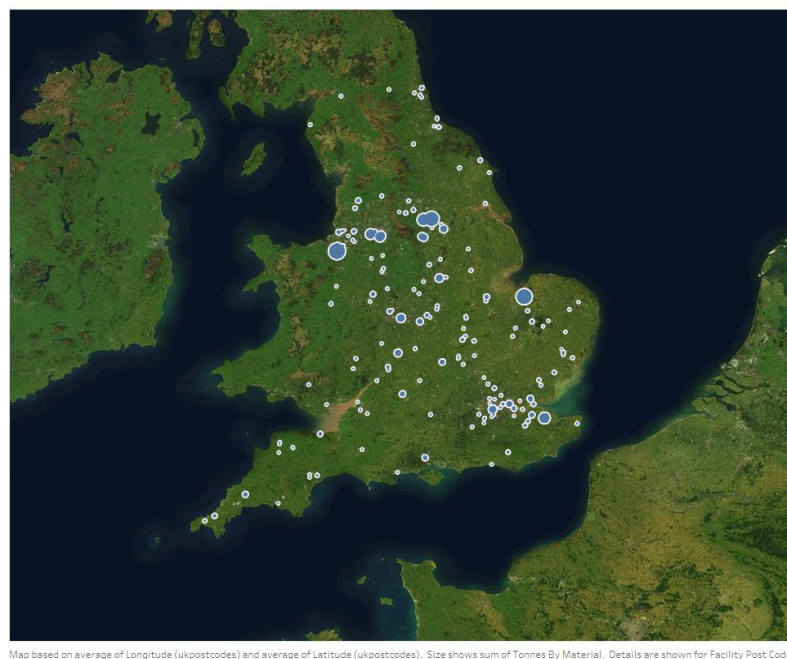


Figure 6. Visualisation of Final Destination for Paper & Card recycling

3.3 Conclusion

Section 1.2 outlined the **hypothesis** that increased population density has a negative effect on the recycling rate. This motivated the initial Methodology A which showed that rates in the less dense authorities of Fenland and Mendip had higher recycling rates than Islington and Luton. The primary motivation was to investigate patterns in the recycling of waste between local authorities and the MDS analysis and the bar chart in Appendix B illustrate that patterns of recycling did not vary widely across the authorities in England.

References

- [1] UK Statistics on waste (2016) Accessed from https://data.gov.uk/dataset/uk_statistics_on_waste
- [2] Barr, S. Factors Influencing Environmental Attitudes and Behaviors: A U.K. Case Study of Household Waste Management. Sage Publications Inc, 2007.
- [3] Galbraith, J. I., et al. The analysis and interpretation of multivariate data for social scientists. CRC Press, 2002.

Appendix A

THE FOLLOWING IS THE CODE USED FOR METHODOLOGY B (Section 2.3) & MDS (Section 2.5)

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import manifold
from sklearn.metrics import euclidean_distances

# Step 1: create dataframes for each file
df = pd.read_csv('recycling.csv', encoding='cp1252')
df1 = pd.read_csv('oa_la_lookup.csv', encoding='cp1252')

# Step 2: create a list of each individual authority
g = df.groupby("Authority", as_index=False)
authority = dict(iter(g))

# Step 3: purpose was to group the dataframes by authorities while only taking the first-row value of the ONS Code
code = ((df['ONSCode']).groupby(df['Authority']).agg(lambda
x:x.value_counts().index[0]))

# Step 4: create a list of each individual authority
paperc = ((df['Paper_Com']).groupby(df['Authority']).sum())
papers = ((df['Paper_Seg']).groupby(df['Authority']).sum())
plasticc = ((df['Plastic_Com']).groupby(df['Authority']).sum())
plastics = ((df['Plastic_Seg']).groupby(df['Authority']).sum())
metalc = ((df['Metal_Com']).groupby(df['Authority']).sum())
metals = ((df['Metal_Seg']).groupby(df['Authority']).sum())
textilesc = ((df['Textiles_Com']).groupby(df['Authority']).sum())
textiless = ((df['Textiles_Seg']).groupby(df['Authority']).sum())
glassc = ((df['Glass_Com']).groupby(df['Authority']).sum())
glasss = ((df['Glass_Seg']).groupby(df['Authority']).sum())
WEEEc = ((df['WEEE_Com']).groupby(df['Authority']).sum())
WEEEs = ((df['WEEE_Seg']).groupby(df['Authority']).sum())
#paperc = df['Paper_Com'].fillna(0)
#papers = df['Paper_Seg'].fillna(0)

# Step 5: reset the indices
code = code.to_frame().reset_index()
paperc = paperc.to_frame().reset_index()
papers = papers.to_frame().reset_index()
plasticc = plasticc.to_frame().reset_index()
plastics = plastics.to_frame().reset_index()
metalc = metalc.to_frame().reset_index()
metals = metals.to_frame().reset_index()
textilesc = textilesc.to_frame().reset_index()
textiless = textiless.to_frame().reset_index()
glassc = glassc.to_frame().reset_index()
glasss = glasss.to_frame().reset_index()
WEEEc = WEEEc.to_frame().reset_index()
WEEEs = WEEEs.to_frame().reset_index()

# Step 6: Merge together material groups as one dataframe using the Authority as the key
paperdf = pd.merge(paperc, papers, on='Authority')
plasticdf = pd.merge(plasticc, plastics, on='Authority')
metaldf = pd.merge(metalc, metals, on='Authority')
textilesdf = pd.merge(textilesc, textiless, on='Authority')
glassdf = pd.merge(glassc, glasss, on='Authority')
WEEEdf = pd.merge(WEEEc, WEEEs, on='Authority')

# Step 7: Merge all material groups
allwastedf =
paperdf.merge(plasticdf,on='Authority').merge(metaldf,on='Authority').merge(texti
lesdf,on='Authority').merge(glassdf,on='Authority').merge(WEEEdf,on='Authority')
```

```

df_merged = code.merge(allwastedf, how='outer', left_index=True,
right_index=True, on='Authority')
# Step 8: Merge with the look-up table dataframe. Drop duplicate rows and sum the populations simultaneously
a = df1['LOCAL_AUTHORITY_NAME'].drop_duplicates()
b = df1['LOCAL_AUTHORITY_CODE'].drop_duplicates()
d = pd.concat([a, b], axis=1)
c = df1['POPULATION'].groupby(df1['LOCAL_AUTHORITY_NAME']).sum().reset_index()

abcd = pd.merge(left=d, right=c, how='left', left_on='LOCAL_AUTHORITY_NAME',
right_on='LOCAL_AUTHORITY_CODE')

merged = pd.merge(left=df_merged, right=abcd, how='left', left_on='ONSCode',
right_on='LOCAL_AUTHORITY_CODE')
# Step 9: Drop any instances where the local authority code is equal to zero. And divide quantities by population
df4 = merged.dropna(subset=['LOCAL_AUTHORITY_CODE']).fillna(0)
df4['Paper'] = (df4.Paper_Com + df4.Paper_Seg) / df4.POPULATION
df4['Plastic'] = (df4.Plastic_Com + df4.Plastic_Seg) / df4.POPULATION
df4['Metal'] = (df4.Metal_Com + df4.Metal_Seg) / df4.POPULATION
df4['Textiles'] = (df4.Textiles_Com + df4.Textiles_Seg) / df4.POPULATION
df4['Glass'] = (df4.Glass_Com + df4.Glass_Seg) / df4.POPULATION
df4['WEEE'] = (df4.WEEE_Com + df4.WEEE_Seg) / df4.POPULATION
# Step 10: Plot a horizontal bar plot of the material waste per population as shown in Appendix B
df4.plot(x="Authority", y=["Paper", "Plastic", "Metal", "Textiles", "Glass",
"WEEE"], kind="barh", figsize=(10, 85), stacked=True)
plt.show()
# Step 11: Tidy up dataframe for MDS
del df4['LOCAL_AUTHORITY_NAME']
del df4['LOCAL_AUTHORITY_CODE']

df5 =
df4.drop(['POPULATION', 'Paper_Com', 'Paper_Seg', 'Plastic_Com', 'Plastic_Seg', 'Metal
_Com', 'Metal_Seg', 'Textiles_Com', 'Textiles_Seg', 'Glass_Com', 'Glass_Seg', 'WEEE_Com
', 'WEEE_Seg'], axis=1)

# Step 12: Only take the numeric columns
numericColumns = df5._get_numeric_data()
# Step 13: Centralise the data
numericColumns -= numericColumns.mean()
# Step 14: Store the authority names in a new variable
placeNames = df5["Authority"]
# Step 15: Compute the Euclidean Distance (dissimilarity measurement) and store the pair-wise distances between the
matrices in a distance matrix
distMatrix = euclidean_distances(numericColumns, numericColumns)
# Step 16: The input is the dissimilarity matrix, therefore set dissimilarity to 'precomputed'. n_components = 2 is the
number of dimensions in which to immerse the dissimilarities. max_iter = 300 is the maximum number of iterations.
n_init=2 is the number of iterations the algorithm will run with different initialisations.
mds = manifold.MDS(n_components = 2, max_iter=3000, n_init=2,
dissimilarity="precomputed")
# Step 17: Fits the data from the distance matrix and returns the embedded co-ordinates to Y
Y = mds.fit_transform(distMatrix)
# Step 18: Create a scatter plot
fig, ax = plt.subplots(figsize=(20, 10))
ax.scatter(Y[:, 0], Y[:, 1], c="#D06B36", s = 10, alpha = 0.8, linewidth='0')
# Step 19: Label the plots with authority names
for i, txt in enumerate(placeNames):
    ax.annotate(txt, (Y[:, 0][i], Y[:, 1][i]))

```

Appendix B

