Recycling in Porto, PT

a study presented by Karen Pereira

June 2020

1. INTRODUCTION

Sustainability has became a big word in the past 30 years. Neglected by years not to sacrifice the so-called development by most companies and ignored by the average citizen, being green is finally becoming natural and, why not?, cool. Social responsibility leads the game for outstanding in competitive markets as new crowds look for reducing its environmental impact.

But how are portuguese people dealing with their own waste at home?

To understand what happens behind closed doors, this project will display how homemade waste is treated across the District of Porto.

I believe this can be the key to understanding good policies to expand across communities and where, and maybe what, is missing in the areas where waste is not treated well by its families.

The goal of this report is to understand the differences in waste treatment in Porto District by families and then propose new action plans to expand awareness.

2. DATA DESCRIPTION

2.A) Data source

All data used for this study comes mainly public source, namely https://dados.gov.pt/ and https://www.pordata.pt/

For this project, it was required to retrieve information on some demographics, such as population and its details, and also to state their coordinates to make possible to connect results to the Foursquare API. Some datasets could be just unpacked from the original location to be used on this study. Still, some cases required to consolidate information into one unique csv file as no database was available for direct extraction.

We chose mainly to use information collected from 2018, so that it would have no mismatch caused by time effects.

2.B) Data cleaning

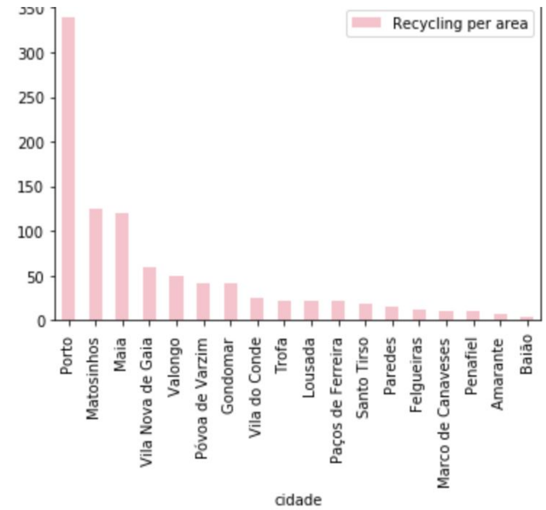
As described in the introduction section, our main focus was Porto District. For that reason, as most databases found relate to National or European information, it was necessary to perform some cleaning and filtering prior to working on the relevant data.

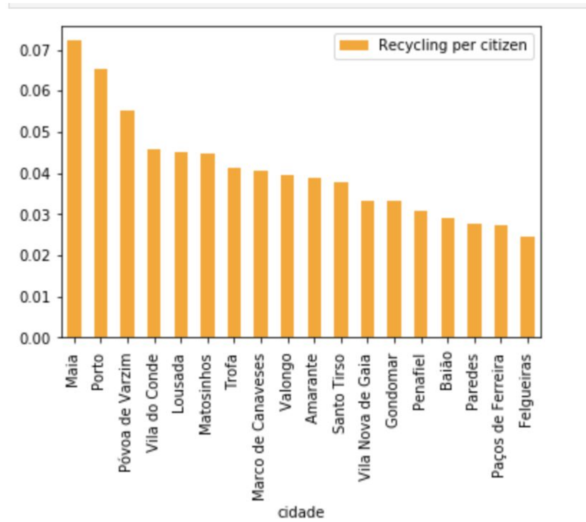
It was necessary to replace values due to some misspelling or absence of data, to modify their types and to merge a few datasets into a consistent base to be used.

It was also important for comparison to exclude data when it wasn't available for the whole set of cities on this analysis. This is mostly relevant for the amount of waste to be considered during this study, meaning that only the main 4 recyclables were considered: glass, plastic, paper and metal.

3.C) Data Exploration

During data exploration, it was used some data visualization to descendingly display the amounts of waste to recycle by its type and its origin. Also in this step, it became visible that there's no pattern related to area or inhabitants when it comes to type of recycling. Those primary visualizations led to the first assumption to be taken during this study: as some cities have very low density, it was more relevant to stick to the number of citizens when comparing data in between cities.





During this step, it was ranked for each city the amount of waste by type, and type-related for each inhabitant of each city.

3. METHODOLOGY & RESULTS

It was chosen to use multiple linear regression to figure out if it was possible to somehow relate the main demographic KPI's to family waste. As exposed during data exploration, KPI's were split into categories of demographics to be analyzed together with data waste. From this exercise, it was expected to figure out which category of KPI can somehow 'predict' an approximate value of family waste in the city. For this, it was used a split into city dataset by 75%.

3.A) Relationship between total waste and population age

Variables used: 'Jovens (%) <15', 'População em idade activa (%)', 'Idosos (%)'
Findings: Variance = 0.27 for R2 = 0.62.
The age split into one community is not statistically relevant for predicting individual recycling amount.

A) Relationship between total waste and population age

```
population_analysis = kpi[['geodsg', 'Jovens (%) <15', 'População em idade activa (%)', 'Idosos (%)']]
population_analysis = weighted_waste.merge(population_analysis, how = 'inner', left_on = 'cidade', right_on ='geodsg')
population_analysis.corr()
```

	total per citizen	Jovens (%) <15	População em idade activa (%)	Idosos (%)
total per citizen	1.000000	0.159421	-0.562734	0.430471
Jovens (%) <15	0.159421	1.000000	0.381354	-0.623723
População em idade activa (%)	-0.562734	0.381354	1.000000	-0.960291
Idosos (%)	0.430471	-0.623723	-0.960291	1.000000

```
# Split set for training
msk = np.random.rand(len(weighted_waste)) < 0.75
train = population_analysis[msk]
test = population_analysis[~msk]
```

```
from sklearn import linear_model
from sklearn.metrics import r2_score
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['Jovens (%) <15', 'População em idade activa (%)', 'Idosos (%)']])
y = np.asanyarray(train[['total per citizen']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
```

```
Coefficients:  [[0.0628092  0.05407142 0.05723514]]
```

```
y_hat= regr.predict(test[['Jovens (%) <15', 'População em idade activa (%)', 'Idosos (%)']])
x = np.asanyarray(test[['Jovens (%) <15', 'População em idade activa (%)', 'Idosos (%)']])
y = np.asanyarray(test[['total per citizen']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
print("R2-score: %.2f" % r2_score(y_hat , y) )
```

```
Residual sum of squares: 0.00
Variance score: 0.27
R2-score: 0.62
```

3.B) Relationship between waste and house pricing

Variables used: 'Valores médios de avaliação bancária dos alojamentos (€/m2)','Alojamentos familiares clássicos'

Findings: Variance = 0.92 for R2 = 0.91
Those results show that house pricing can be relevant to understand how engaging community can be over recycling, as it shows a high accuracy of the model with low error.

B) Relationship between waste and house pricing

```
house_analysis = kpi[['geodsg', 'Valores médios de avaliação bancária dos alojamentos (€/m2)','Alojamentos familiares clássicos']]
house_analysis = weighted_waste.merge(house_analysis, how = 'inner', left_on = 'cidade', right_on ='geodsg')
house_analysis.corr()
```

	total per citizen	Valores médios de avaliação bancária dos alojamentos (€/m2)	Alojamentos familiares clássicos
total per citizen	1.000000	0.682526	0.357940
Valores médios de avaliação bancária dos alojamentos (€/m2)	0.682526	1.000000	0.764249
Alojamentos familiares clássicos	0.357940	0.764249	1.000000

```
# Split set for training
msk = np.random.rand(len(weighted_waste)) < 0.75
train = house_analysis[msk]
test = house_analysis[~msk]
```

```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['Valores médios de avaliação bancária dos alojamentos (€/m2)','Alojamentos familiares clássicos']])
y = np.asanyarray(train[['total per citizen']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
```

```
Coefficients:  [[ 4.63081863e-05 -8.30347573e-05]]
```

```
y_hat= regr.predict(test[['Valores médios de avaliação bancária dos alojamentos (€/m2)','Alojamentos familiares clássicos']])
x = np.asanyarray(test[['Valores médios de avaliação bancária dos alojamentos (€/m2)','Alojamentos familiares clássicos']])
y = np.asanyarray(test[['total per citizen']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
print("R2-score: %.2f" % r2_score(y_hat , y) )
```

```
Residual sum of squares: 0.00
Variance score: 0.92
R2-score: 0.91
```

3.C) Relationship between waste and students

Variables used: 'Estabelecimentos do ensino préescolar', 'Estabelecimentos do 1.º ciclo do ensino básico', 'Estabelecimentos do 2.º ciclo do ensino básico', 'Estabelecimentos do 3.º ciclo do ensino básico', 'Estabelecimentos do ensino secundário', 'Alunos do ensino não superior (5)', 'Estabelecimentos do ensino superior', 'Alunos do ensino superior (5)'

Findings: Variance = -1.75 for R2 = -4.75

The students data weren't relevant for this study as it could not validate the model proposed, as shown below.

C) Relationship between waste and students

```
students_analysis = kpi[['geodsg', 'Estabelecimentos do ensino préescolar',
       'Estabelecimentos do 1.º ciclo do ensino básico',
       'Estabelecimentos do 2.º ciclo do ensino básico',
       'Estabelecimentos do 3.º ciclo do ensino básico',
       'Estabelecimentos do ensino secundário',
       'Alunos do ensino não superior (5)',
       'Estabelecimentos do ensino superior', 'Alunos do ensino superior (5)']]
students_analysis = weighted_waste.merge(students_analysis, how = 'inner', left_on = 'cidade', right_on ='geodsg')
students_analysis.corr()
```

	total per citizen	Estabelecimentos do ensino préescolar	Estabelecimentos do 1.º ciclo do ensino básico	Estabelecimentos do 2.º ciclo do ensino básico	Estabelecimentos do 3.º ciclo do ensino básico	Estabelecimentos do ensino secundário	Alunos do ensino não superior (5)	Estabelecimentos do ensino superior	Alunos do ensino superior (5)
total per citizen	1.000000	0.364232	0.277486	0.423620	0.377276	0.469376	0.409907	0.475341	0.505609
Estabelecimentos do ensino préescolar	0.364232	1.000000	0.973377	0.882203	0.891396	0.768928	0.965079	0.746970	0.656021
Estabelecimentos do 1.º ciclo do ensino básico	0.277486	0.973377	1.000000	0.812765	0.837416	0.693796	0.907127	0.664314	0.561898
Estabelecimentos do 2.º ciclo do ensino básico	0.423620	0.882203	0.812765	1.000000	0.991486	0.960568	0.966840	0.927230	0.892454
Estabelecimentos do 3.º ciclo do ensino básico	0.377276	0.891396	0.837416	0.991486	1.000000	0.957152	0.966228	0.922340	0.883627
Estabelecimentos do ensino secundário	0.469376	0.768928	0.693796	0.960568	0.957152	1.000000	0.889008	0.984092	0.976484
Alunos do ensino não superior (5)	0.409907	0.965079	0.907127	0.966840	0.966228	0.889008	1.000000	0.866236	0.800257
Estabelecimentos do ensino superior	0.475341	0.746970	0.664314	0.927230	0.922340	0.984092	0.866236	1.000000	0.986622
Alunos do ensino superior (5)	0.505609	0.656021	0.561898	0.892454	0.883627	0.976484	0.800257	0.986622	1.000000

```
train = students_analysis[msk]
test = students_analysis[~msk]
```

```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['Estabelecimentos do ensino préescolar',
       'Estabelecimentos do 1.º ciclo do ensino básico',
       'Estabelecimentos do 2.º ciclo do ensino básico',
       'Estabelecimentos do 3.º ciclo do ensino básico',
       'Estabelecimentos do ensino secundário',
       'Alunos do ensino não superior (5)',
       'Estabelecimentos do ensino superior', 'Alunos do ensino superior (5)']])
y = np.asanyarray(train[['total per citizen']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
```

```
Coefficients:  [[ 1.57574964e-04  5.11802029e-04  9.31336459e-04 -6.35880713e-03
   3.97356655e-03  1.50539130e-06 -4.58093330e-03  2.61239566e-06]]
```

```
y_hat= regr.predict(test[['Estabelecimentos do ensino préescolar',
       'Estabelecimentos do 1.º ciclo do ensino básico',
       'Estabelecimentos do 2.º ciclo do ensino básico',
       'Estabelecimentos do 3.º ciclo do ensino básico',
       'Estabelecimentos do ensino secundário',
       'Alunos do ensino não superior (5)',
       'Estabelecimentos do ensino superior', 'Alunos do ensino superior (5)']])
x = np.asanyarray(test[['Estabelecimentos do ensino préescolar',
       'Estabelecimentos do 1.º ciclo do ensino básico',
       'Estabelecimentos do 2.º ciclo do ensino básico',
       'Estabelecimentos do 3.º ciclo do ensino básico',
       'Estabelecimentos do ensino secundário',
       'Alunos do ensino não superior (5)',
       'Estabelecimentos do ensino superior', 'Alunos do ensino superior (5)']])
y = np.asanyarray(test[['total per citizen']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
print("R2-score: %.2f" % r2_score(y_hat , y) )
```

```
Residual sum of squares: 0.00
Variance score: -1.75
R2-score: -4.75
```

3.D) Relationship between waste and spent on sports & culture

Variables used: 'Museus', 'Sessões de espectáculos ao vivo', 'Ecrãs de cinema', 'Despesas da Câmara Municipal em cultura e desporto (%)'

Findings: Variance = 0.95 for R2 = 0.92

That means that one city's spend on sports & culture can be statistically relevant for predicting individual recycling amount.

D) Relationship between waste and spent on sports & culture

```
sports_analysis = kpi[['geodsg', 'Museus', 'Sessões de espectáculos ao vivo', 'Ecrãs de cinema']]
sports_analysis = weighted_waste.merge(sports_analysis, how = 'inner', left_on = 'cidade', right_on ='geodsg')
sports_analysis.corr()
```

	total per citizen	Museus	Sessões de espectáculos ao vivo	Ecrãs de cinema
total per citizen	1.000000	0.445524	0.500396	0.176939
Museus	0.445524	1.000000	0.977023	0.415655
Sessões de espectáculos ao vivo	0.500396	0.977023	1.000000	0.309360
Ecrãs de cinema	0.176939	0.415655	0.309360	1.000000

```
# Split set for training
msk = np.random.rand(len(weighted_waste)) < 0.75
train = sports_analysis[msk]
test = sports_analysis[~msk]
```

```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['Museus', 'Sessões de espectáculos ao vivo', 'Ecrãs de cinema']])
y = np.asanyarray(train[['total per citizen']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
```

```
Coefficients:  [[-4.15936675e-03  3.01764848e-05  2.72192916e-04]]
```

```
y_hat= regr.predict(test[['Museus', 'Sessões de espectáculos ao vivo', 'Ecrãs de cinema']])
x = np.asanyarray(test[['Museus', 'Sessões de espectáculos ao vivo', 'Ecrãs de cinema']])
y = np.asanyarray(test[['total per citizen']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
print("R2-score: %.2f" % r2_score(y_hat , y) )
```

```
Residual sum of squares: 0.00
Variance score: 0.95
R2-score: 0.92
```

3.E) Relationship between waste and wages

Variables used: 'Ganho médio mensal dos trabalhadores por conta de outrem. €','Desempregados inscritos nos centros de emprego'
Findings: Variance = -1.54 for R2 = 0.01
The wages received by one's living in a city is not statistically relevant for predicting individual recycling amount.

```
wages_analysis = kpi[['geodsg', 'Ganho médio mensal dos trabalhadores por conta de outrem. €','Desempregados inscritos nos centros de emprego']
wages_analysis = weighted_waste.merge(wages_analysis, how = 'inner', left_on = 'cidade', right_on ='geodsg')
wages_analysis.corr()
```

	total per citizen	Ganho médio mensal dos trabalhadores por conta de outrem. €	Desempregados inscritos nos centros de emprego
total per citizen	1.000000	0.679653	0.196591
Ganho médio mensal dos trabalhadores por conta de outrem. €	0.679653	1.000000	0.633541
Desempregados inscritos nos centros de emprego	0.196591	0.633541	1.000000

```
# Split set for training
msk = np.random.rand(len(weighted_waste)) < 0.75
train = wages_analysis[msk]
test = wages_analysis[~msk]
```

```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['Ganho médio mensal dos trabalhadores por conta de outrem. €','Desempregados inscritos nos centros de emprego']])
y = np.asanyarray(train[['total per citizen']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
```

```
Coefficients:  [[ 9.89505286e-05 -1.59086861e-06]]
```

```
y_hat= regr.predict(test[['Ganho médio mensal dos trabalhadores por conta de outrem. €','Desempregados inscritos nos centros de emprego']])
x = np.asanyarray(test[['Ganho médio mensal dos trabalhadores por conta de outrem. €','Desempregados inscritos nos centros de emprego']])
y = np.asanyarray(test[['total per citizen']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
print("R2-score: %.2f" % r2_score(y_hat , y) )
```

```
Residual sum of squares: 0.00
Variance score: -1.54
R2-score: 0.01
```

3.F) Relationship between waste and spent on environmental matters

Variables used: 'Despesas do município em ambiente (%)', 'Despesas da Câmara Municipal (7)'
Findings: Variance = 0.51 for R2 = -2.14
The city spent on environmental matters surprisingly has no statistical relation to the amount recycled by citizens.

F) Relationship between waste and spent on environmental matters

```
environment_analysis = kpi[['geodsg', 'Despesas do município em ambiente (%)', 'Despesas da Câmara Municipal (7)']]
environment_analysis['Despesas em MA']= environment_analysis['Despesas do município em ambiente (%)'] * environment_analysis['Despesas da Câmara
environment_analysis = environment_analysis[['geodsg', 'Despesas em MA', 'Despesas da Câmara Municipal (7)']]
environment_analysis = weighted_waste.merge(environment_analysis, how = 'inner', left_on = 'cidade', right_on ='geodsg')
environment_analysis.corr()
```

```
# Split set for training
msk = np.random.rand(len(weighted_waste)) < 0.75
train = environment_analysis[msk]
test = environment_analysis[~msk]
```

```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['Despesas da Câmara Municipal (7)','Despesas em MA']])
y = np.asanyarray(train[['total per citizen']])
regr.fit (x, y)
# The coefficients
print ('Coefficients: ', regr.coef_)
```

```
Coefficients:  [[6.36763233e-08 1.64967897e-09]]
```

```
y_hat= regr.predict(test[['Despesas da Câmara Municipal (7)','Despesas em MA']])
x = np.asanyarray(test[['Despesas da Câmara Municipal (7)','Despesas em MA']])
y = np.asanyarray(test[['total per citizen']])
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
print("R2-score: %.2f" % r2_score(y_hat , y) )
```

```
Residual sum of squares: 0.00
Variance score: 0.51
R2-score: -2.14
```

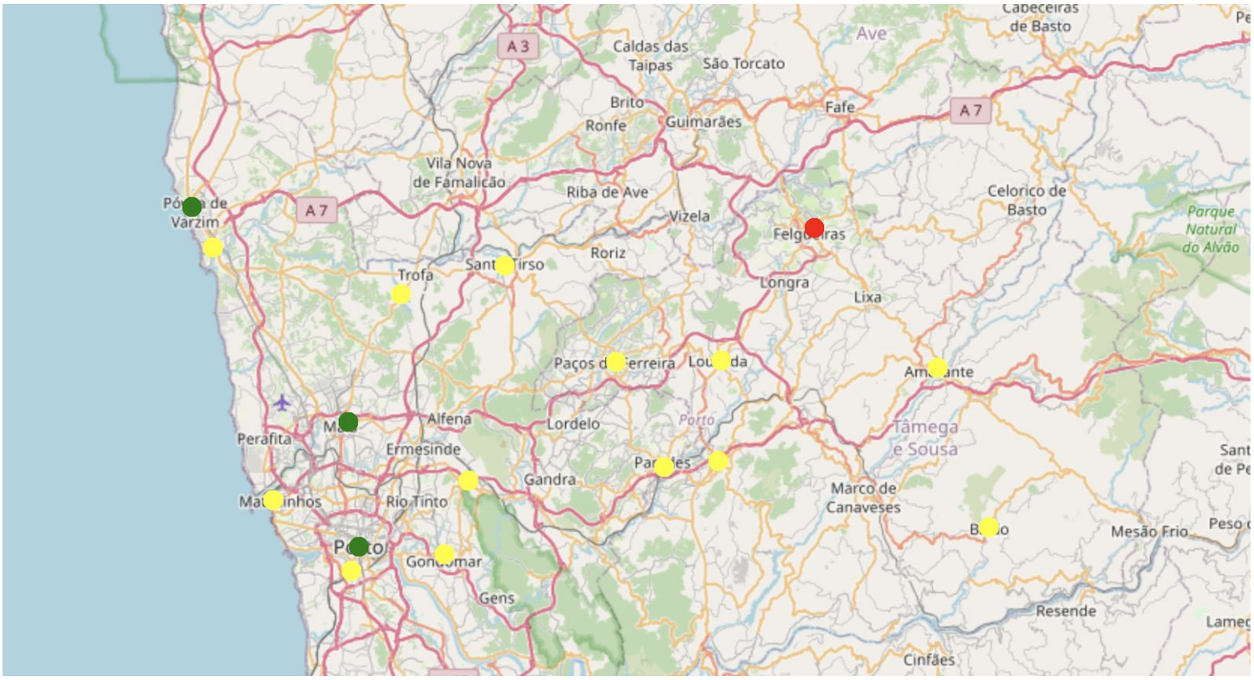
4. NOTES ON RESEARCH

First challenge faced in this study was to find data that could be used. Although multiple datasets can be found on the internet, many are outdated or only apply to national numbers. At first, the idea was to go on neighborhood detail but as information for North Portugal was not available that deep, we took an approach on a city basis.

It was expected to find some correlation in between some KPI sets and recycling in the cities. But in the level of detail that the dataset was exposed to, the strongest relation found for possible prediction was the average price of houses in the cities and city spent on sports & culture. Population aging also played a part in this study and should be reconsidered for future steps.

To complete this study, cities were classified as a green, yellow or red city, depending on the way families waste are treated within the community, accordingly to this structure:
Over 50kg of recycling per year, per inhabitant = Green City
Between 25kg and 50kg = Yellow City
Below 25kg per year = Red City

This classification is displayed in a map with a pin to the city and its correspondent classification, as it can be seem below:.



5. CONCLUSION

In this study, relationships between several demographic KPI's and home waste were tested and analyzed. After understanding their behavior, some KPIs were chosen to be tested upon a geographical approach. Also cities were classified as a green, yellow or red city, depending on the way families waste is treated within the community.

Also it was learned that using the same standards across datasets may not result in a strong and reliable model.