GLOBAL POLLUTION

Lee Xuan Le, Chloe (U2230069E)
Valen Tang Wenwen (U2230862L)
Zann Ang (U2230204A)

W132 Team 6

TABLE OF CONTENTS

01

PROBLEM DEFINITION

Introduction of dataset used

APPROACHING THE PROBLEM

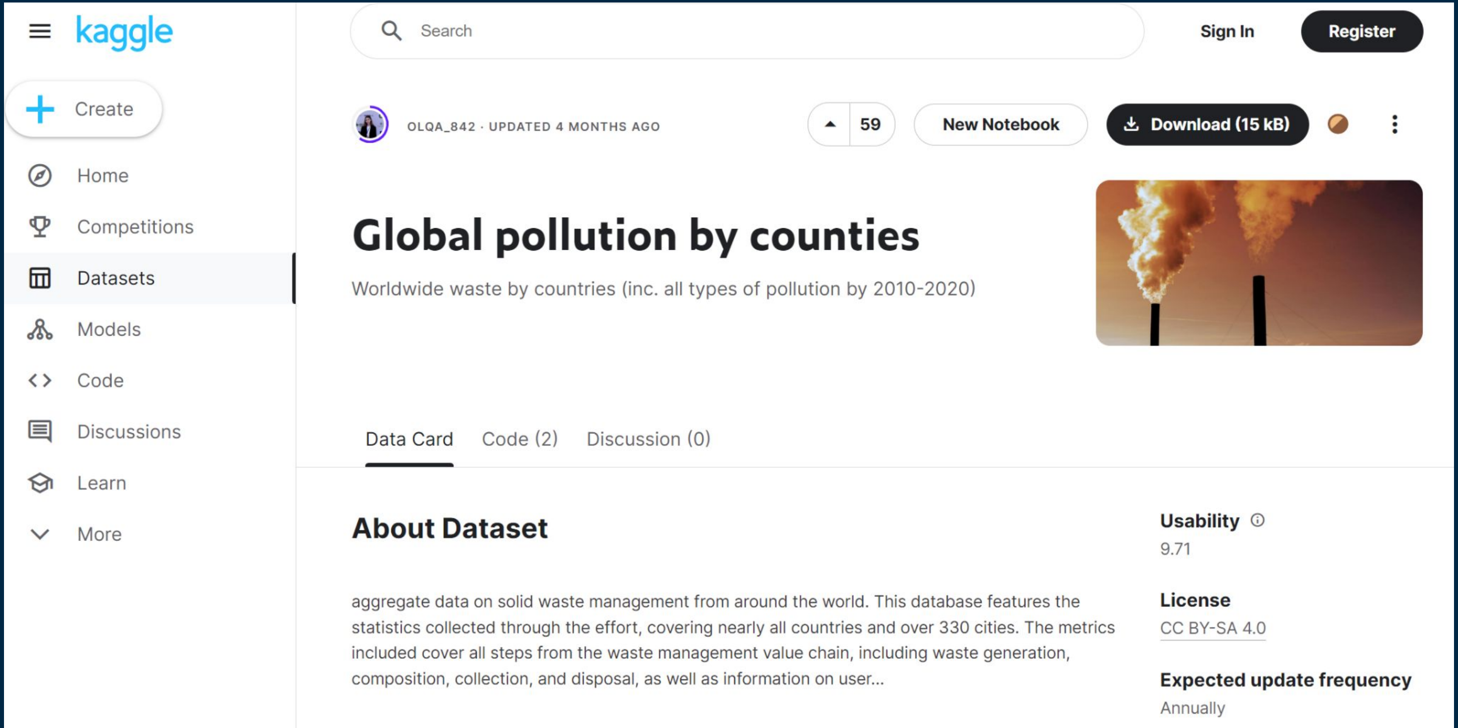
OBSERVATIONS LEADING TO PROBLEM FORMATION

- High food wastage in developed countries
- Lack of focus on organic waste
- Lack of awareness and solutions to tackle problem

POSSIBLE TARGETS FOR DISCOVERY

- Significance of problem
- Factors affecting organic waste and recycling efforts
- Significance of these factors
- Existing solutions

CHOSEN DATASET



49 columns x 218 rows

PROBLEM STATEMENT

We want to find out how the **amount of food organic waste** produced by **different countries varies** with many **factors**, and thus determine the **most significant factor** that contributes to high waste levels.

02

OUR PROCESS

Data cleaning

OUTLINE OF DATA CLEANING AND PROCESSING

DIFFERENTIATING

Identifying the columns that we needed

SORTING

Identifying which rows have to be removed

NEW TABLE

Creating a new table with the columns that we need

PREDICTION

Predicting NA values so that there are no more null values

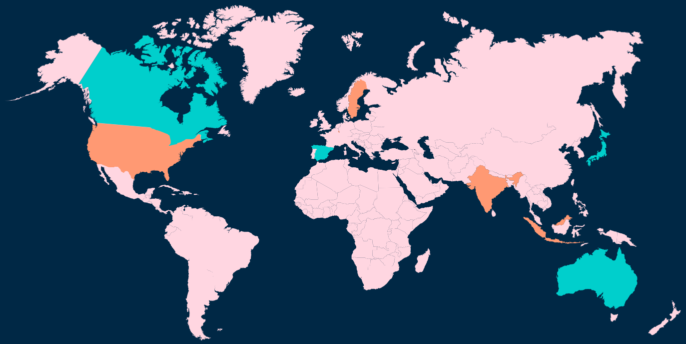
DIFFERENTIATING - SELECTING COLUMNS

Columns we have chosen:

- Country
- Population
- Total municipal solid waste (MSW)
- Percentage of food organic waste
- GDP
- Recycling percentage
- Presence of national law governing solid waste management
- Presence of national agency enforcing solid waste laws and regulations

Why we chose these columns:

- Factors affecting food organic waste percentage
- Population and total MSW can be used to predict NA values
- Provides a range of data for analysis



SORTING – DELETING ROWS

D225

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	region_id	country_name	gdp	compositic	compositic	compositic	compositic	compositic	compositic	compositic	compositic	compositic	other_info	other_info	other_info	other_info	other_info
200	EAS	NA	57362.93	23.34	4.11	3.07	14.11	30.01	20.23	0.6	NA	4.43	NA	NA	Yes	Yes	NA
201	SSF	Tanzania	2128.554	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Yes	NA	NA

Deleting row 200: Unable to predict country’s name

E193 4.64

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	region_id	country_name	gdp	compositic	compositic	compositic	compositic	compositic	compositic	compositic	compositic	compositic	other_info	other_info	other_info	other_info
185	LCN	Sint Maarten (Dutch part)	NA	46	7	7	12	15	13	NA	NA	NA	NA	Yes	Yes	NA
186	SSF	Seychelles	23302.95	48.5	5.2	4.6	25.8	5.9	9.9	0.2	0.5	NA	NA	Yes	Yes	Yes

region_i	country_	populati	special_	special_	special_	special_	special_	special_	total_m
LCN	Turks and (	34900	NA	NA	NA	NA	NA	NA	NA

Deleting rows 185 and 187: The dataset is sufficiently large to answer our question without these rows

PREDICTION – PREDICTING NA VALUES

CHECKPOINT 1

Identify the rows with null values

CHECKPOINT 2

Drop the rows with null values for future prediction

CHECKPOINT 3

Predicting rows with null values

CHECKPOINT 4

Merge tables of non-null and null values

CHECKPOINTS 1 & 2: Identify and Drop

```
# Drop the rows with null values to be used during prediction
food.dropna(inplace=True)
food
```

	composition_food_organic_waste_percent	total_msw_total_msw_generated_tons_year	population_population_number_of_people
2	51.800000	4.213644e+06	25096150
3	51.400000	1.087447e+06	2854191
4	31.200000	4.300000e+04	82431
5	39.000000	5.617682e+06	9770529
6	38.740000	1.791055e+07	42981516
...	...	...	...
208	42.600000	2.739909e+04	187665
209	42.000000	3.190000e+05	1801800
210	65.000000	4.836820e+06	27584212
211	16.381655	1.845723e+07	51729344
213	36.000000	1.449752e+06	12500525

CHECKPOINTS 3 & 4 (NUMERICAL)

Extension

```
# Predicting null values for food organic waste
y_pred = lr.predict(x_test)
y_pred
```

```
array([ 41.56654727,  43.68141007,  41.59748735,  41.74546417,
        41.54652823,  42.48651747,  45.88354354,  41.73557005,
        41.59798784,  41.59065586,  41.62829328,  41.85235493,
        41.56862933,  41.62895414,  41.68797062,  41.65525734,
        41.57515394, 132.33872991,  41.89941798,  41.7959195 ,
        41.77254584,  41.58005956,  41.57183639,  43.17043758,
        45.19951734,  42.99530528,  41.6747978 ,  50.19270821,
        41.7536694 ,  41.76621836,  41.58419395,  41.99113489,
        42.44871912,  42.00450979,  41.59445152,  41.49834991,
        41.66262505,  41.94197741,  44.17237147,  41.94683762,
        42.36009665])
```

```
# Merging the two tables together (previously null and non-null values) (with country names)
updated_food = pd.merge(food_country, food_null_country, how='outer')
updated_food
```

	country_name	composition_food_organic_waste_percent	total_msw_total_msw_generated_tons_year	population_population_number_of_people
0	Angola	51.800000	4.213644e+06	25096150
1	Albania	51.400000	1.087447e+06	2854191
2	Andorra	31.200000	4.300000e+04	82431
3	United Arab Emirates	39.000000	5.617682e+06	9770529
4	Argentina	38.740000	1.791055e+07	42981516
...	...	...	...	...
208	Eswatini	41.662625	2.181995e+05	1343098
209	Tajikistan	41.941977	1.787400e+06	8177809
210	Tanzania	44.172371	9.276995e+06	49082996
211	Venezuela, RB	41.946838	9.779093e+06	29893080
212	Zambia	42.360097	2.608268e+06	14264756

213 rows × 4 columns

CHECKPOINTS 3 & 4 (CATEGORICAL)

Extension

```
# Predicting null values for national law
y_pred = lr.predict(x_test)
y_pred
```

```
array([0.91791585, 0.9165048 , 0.91394896, 0.91513661, 0.91607257,
       0.91390421, 0.91599359, 0.91548814, 0.91580576, 0.91493718,
       0.90599427, 0.90436968, 0.91561371, 0.91387206, 0.88785023,
       0.91285784, 0.91652111, 0.90815062, 0.91151127, 0.91639827,
       0.91573468, 0.90856219, 0.89837645])
```

```
# Analysing the predicted values to determine which is yes and which is no
arr = np.array([0.91791585, 0.9165048 , 0.91394896, 0.91513661, 0.91607257,
       0.91390421, 0.91599359, 0.91548814, 0.91580576, 0.91493718,
       0.90599427, 0.90436968, 0.91561371, 0.91387206, 0.88785023,
       0.91285784, 0.91652111, 0.90815062, 0.91151127, 0.91639827,
       0.91573468, 0.90856219, 0.89837645])

# Create an empty list
filter_arr = []

# Go through each element in arr
for element in arr:
  # If the element (y/1-y) is higher than y+/y-, set the value to 1, otherwise 0
  if (element/(1-element)) > 10.1:
    filter_arr.append(1)
  else:
    filter_arr.append(0)

newarr = arr[filter_arr]

# Converting the predicted values into 1 and 0
print(filter_arr)
```

```
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0]
```

```
# Converting 1 and 0 back to yes and no respectively
final_national_law = updated_national_law[['country_name', 'other_information_national_law_governing_solid_waste_management_i
final_national_law.other_information_national_law_governing_solid_waste_management_in_the_country[final_national_law.other_in
final_national_law.other_information_national_law_governing_solid_waste_management_in_the_country[final_national_law.other_in
print(final_national_law)
```

```
              country_name  \
0                    Aruba
1              Afghanistan
2                   Angola
3                  Albania
4                  Andorra
..                     ...
207            South Sudan
208  São Tomé and Príncipe
209               Eswatini
210                   Chad
211                 Uganda

    other_information_national_law_governing_solid_waste_management_in_the_country
0                                                  Yes
1                                                  Yes
2                                                  Yes
3                                                  Yes
4                                                  Yes
..                                                 ...
207                                                Yes
208                                                Yes
209                                                Yes
210                                                 No
211                                                 No

[212 rows x 2 columns]
```

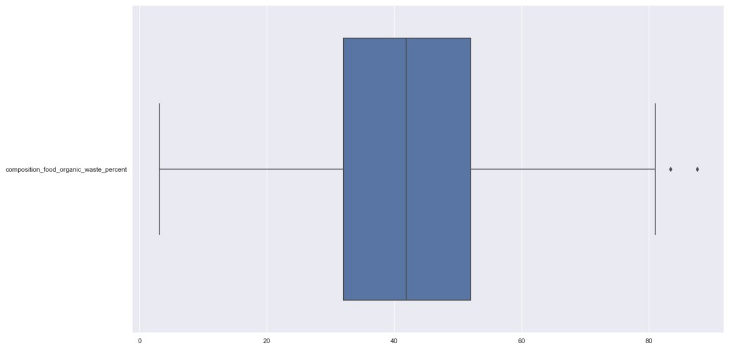
03

DATA ANALYSIS

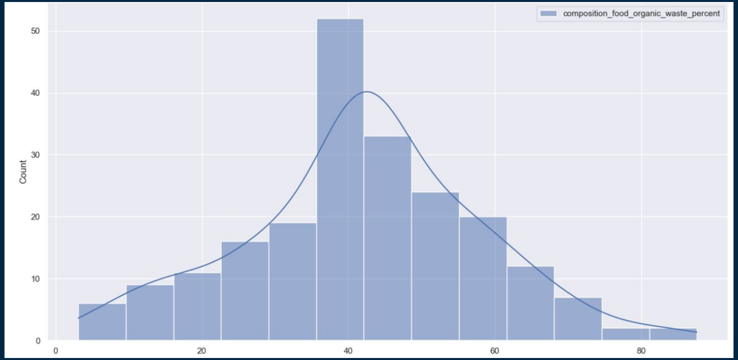
Techniques used, conclusion and insights

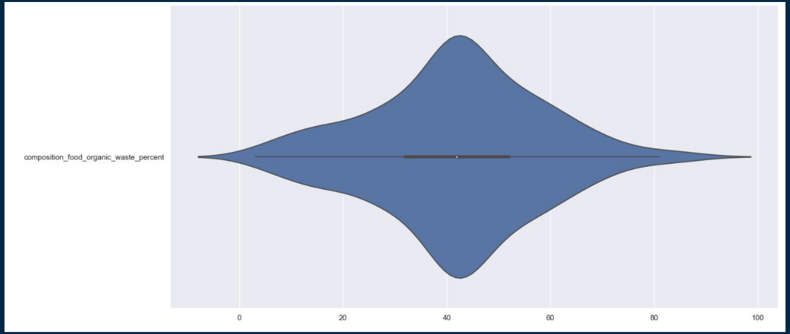
UNI-VARIATE ANALYSIS (NUMERICAL)

BOX-PLOTS



DENSITY PLOT





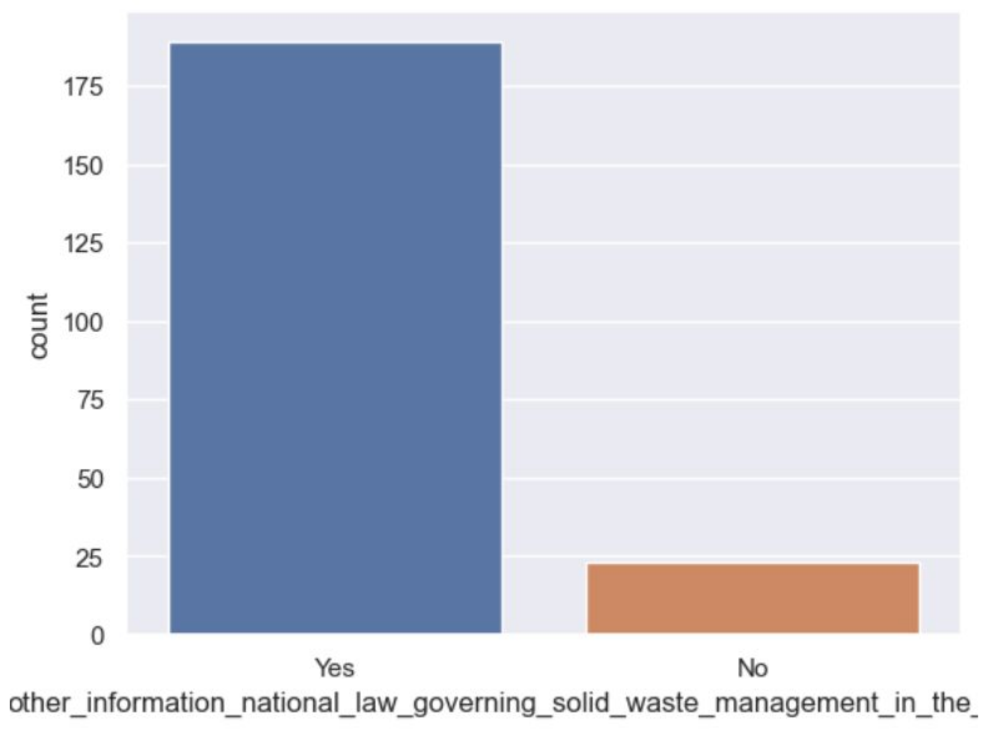
VIOLIN-PLOTS

UNI-VARIATE ANALYSIS (NUMERICAL)

What we obtained from this analysis:

1. Mean
2. Median
3. Mode
4. Distributions of the different variables
5. Outliers in the data

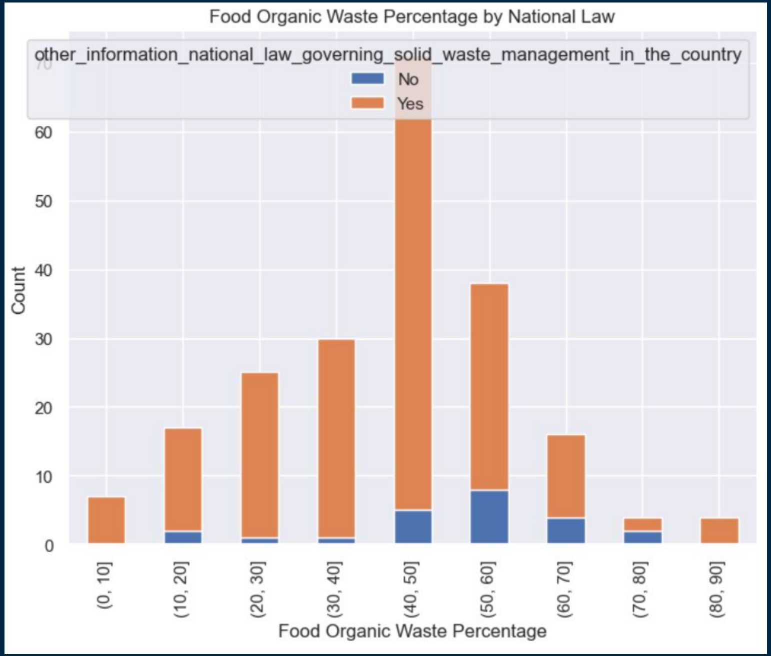
UNI-VARIATE ANALYSIS (CATEGORICAL)

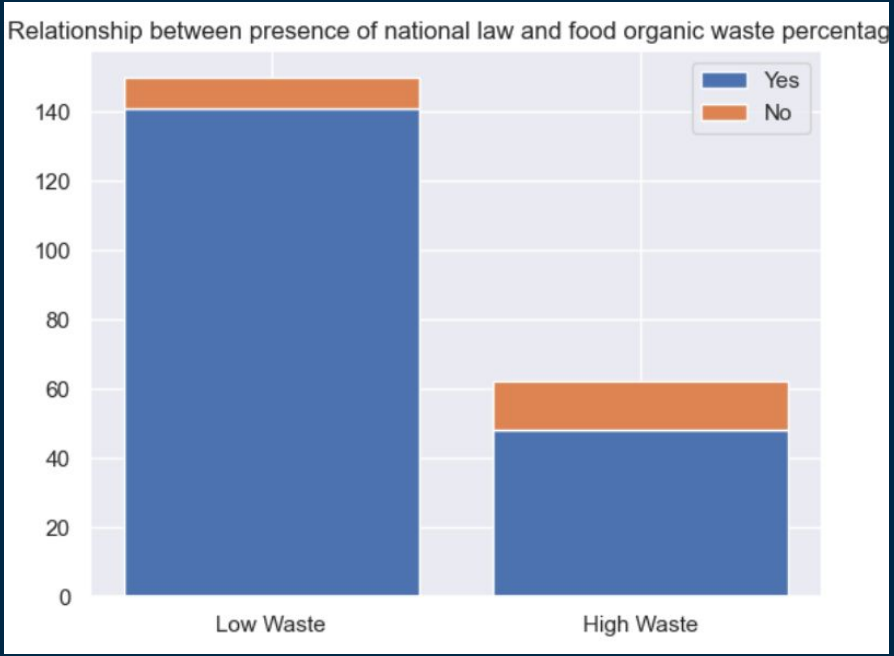


COUNT PLOT

CHECKING IF THE DATASET IS BALANCED

Extension





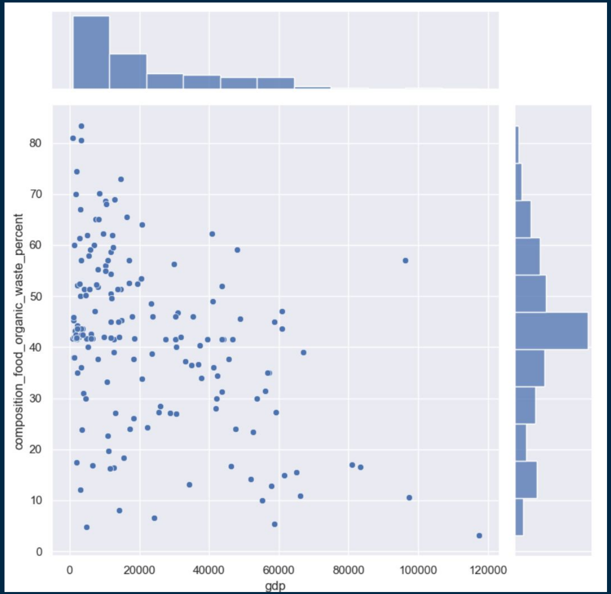
BI-VARIATE ANALYSIS

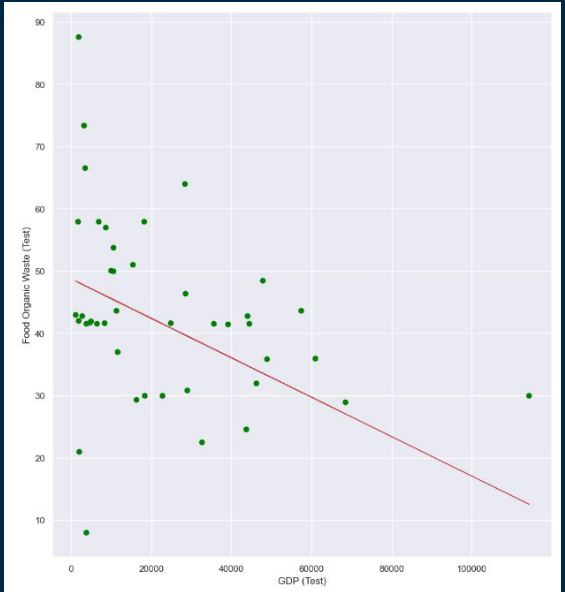
RELATIONSHIPS (y vs x):

1. Food organic waste percentage vs GDP
2. Food organic waste percentage vs Recycling percentage
3. Food organic waste percentage vs Presence of national law
4. Food organic waste percentage vs Presence of enforcement

BI-VARIATE ANALYSIS (NUMERICAL)

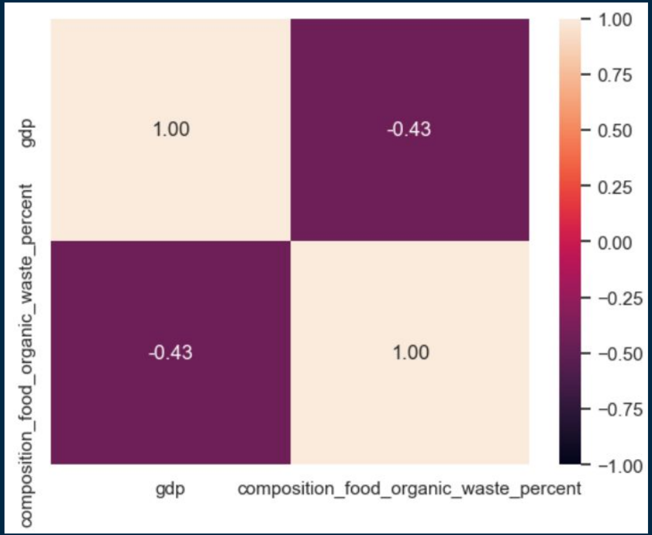
JOINTPLOT



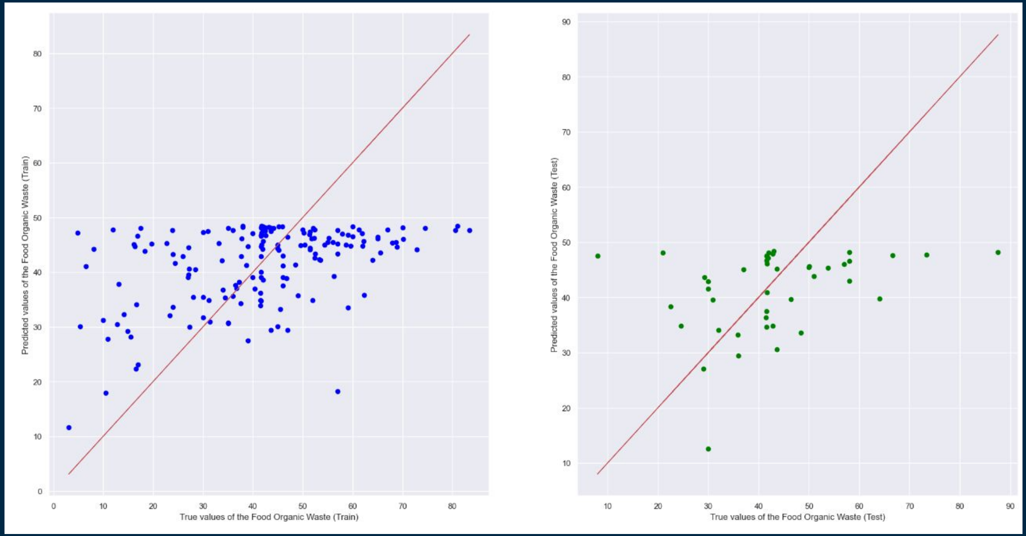


SCATTER PLOT- LINEAR REGRESSION

BI-VARIATE ANALYSIS (NUMERICAL)

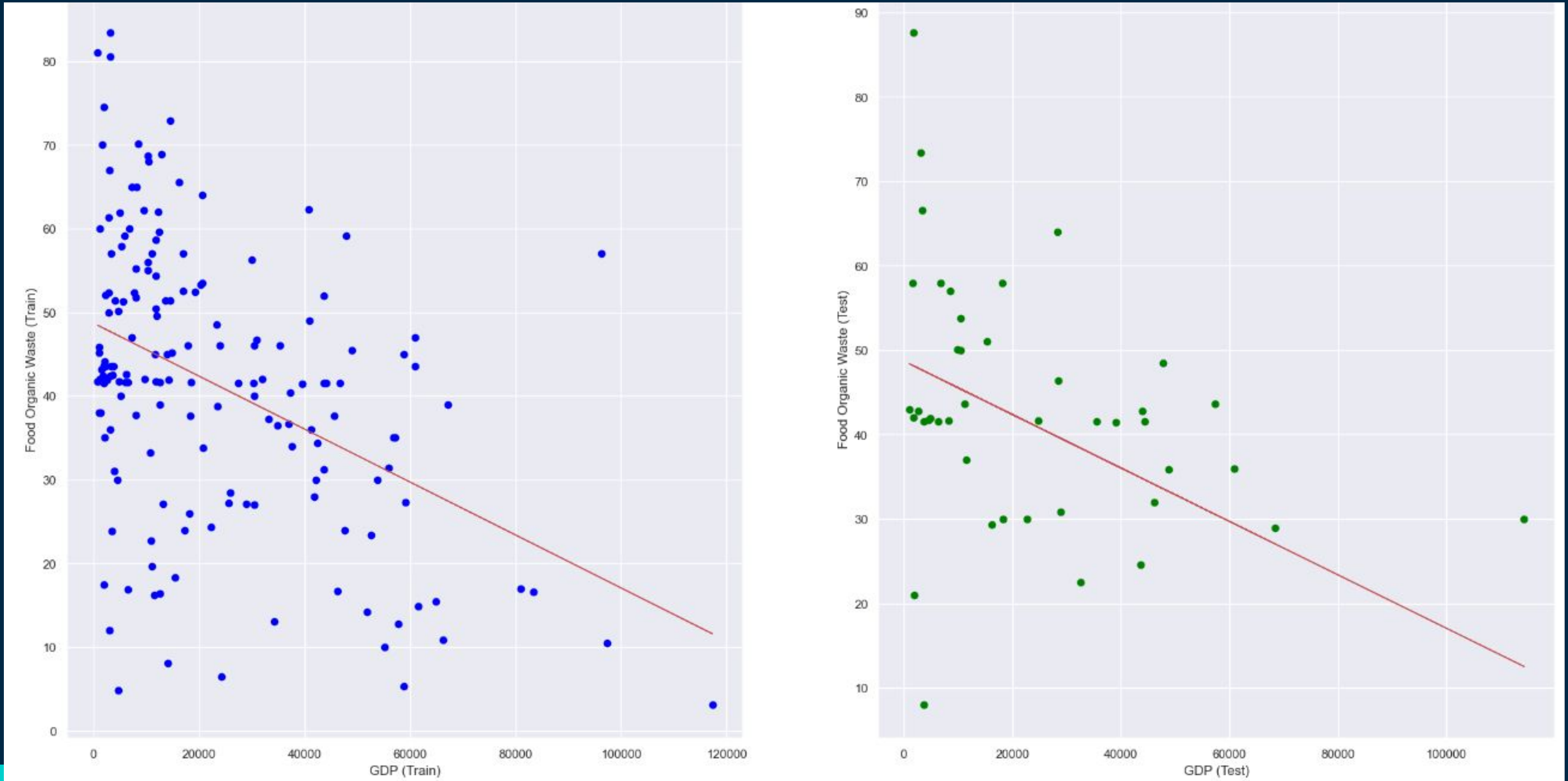


HEATMAP



LINEAR REGRESSION

BI-VARIATE ANALYSIS (NUMERICAL)



BI-VARIATE ANALYSIS: GOODNESS OF FIT

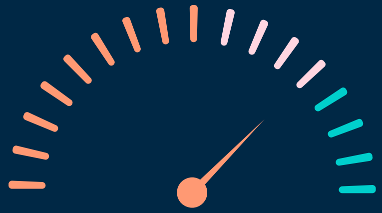
Explained Variance

R^2

Mean Squared Error

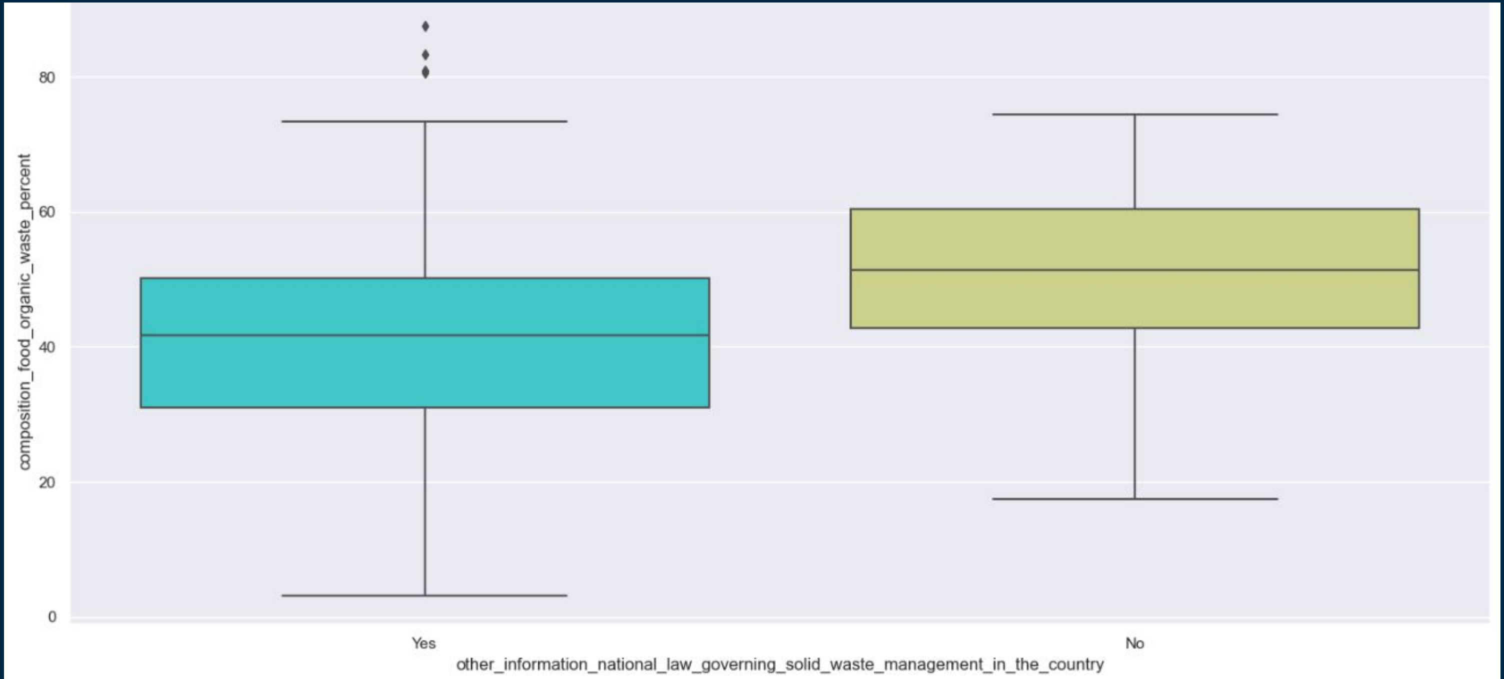
MSE

Root Mean Squared Error



RMSE

BI-VARIATE ANALYSIS (CATEGORICAL)



BOX-PLOT

MULTIVARIATE ANALYSIS

Extension

RANDOM FOREST CLASSIFIER

To compare all 4 variables, both numerical and categorical

F AND PROBABILITY VALUES

To predict the significance of impact of each variable on food organic waste percentage

RANDOM FOREST REGRESSOR

Extension

Comparison of 4 variables:

1. Numerical variables
   - GDP
   - Recycling percentage

2. Categorical variables
   - Presence of national law
   - Presence of enforcement

Feature importances:

The higher the value, the greater the significance of the variable in contributing to food organic waste.

1. Numerical feature importances
2. Categorical feature importances
   - Ignore the last two numbers (dummies)

```
Mean Squared Error: 320.2608265498277
Root Mean Squared Error: 17.895832658745658
R-squared: 0.0019115142812655694
Feature Importances: [0.51054631 0.44286293 0.01077362 0.01032748 0.01307536 0.0124143 ]
Numerical Feature Importances: [0.5105463134160089, 0.44286292540008676]
Categorical Feature Importances: [0.01077362028474791, 0.010327478844332057, 0.013075360549516338, 0.012414301505530811]
```

COMPARING F AND P-VALUES

Extension

```
                                  OLS Regression Results
==========================================================================================
Dep. Variable:     composition_food_organic_waste_percent   R-squared:                0.188
Model:                                                OLS   Adj. R-squared:           0.180
Method:                                     Least Squares   F-statistic:              24.21
Date:                                    Fri, 21 Apr 2023   Prob (F-statistic):    3.51e-10
Time:                                            12:56:36   Log-Likelihood:         -868.00
No. Observations:                                     212   AIC:                      1742.
Df Residuals:                                         209   BIC:                      1752.
Df Model:                                               2
Covariance Type:                                nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                51.2219      1.848     27.719      0.000      47.579      54.865
gdp                                  -0.0002   4.91e-05     -4.932      0.000      -0.000      -0.000
waste_treatment_recycling_percent    -0.2201      0.095     -2.313      0.022      -0.408      -0.033
==============================================================================
Omnibus:                        1.880   Durbin-Watson:                   2.200
Prob(Omnibus):                  0.391   Jarque-Bera (JB):                1.512
Skew:                          -0.165   Prob(JB):                        0.469
Kurtosis:                       3.250   Cond. No.                     5.89e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.89e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

NUMERICAL DATA

	sum_sq	df
other_information_national_agency_to_enforce_so...	597.655588	1.0
other_information_national_law_governing_solid_...	1246.859067	1.0
other_information_national_agency_to_enforce_so...	4.889152	1.0
Residual	52300.005901	208.0

	F	PR(>F)
other_information_national_agency_to_enforce_so...	2.376909	0.124661
other_information_national_law_governing_solid_...	4.958827	0.027030
other_information_national_agency_to_enforce_so...	0.019444	0.889235
Residual	NaN	NaN

CATEGORICAL DATA

OUR CONCLUSION

GDP

The most significant factor in affecting food organic waste

%

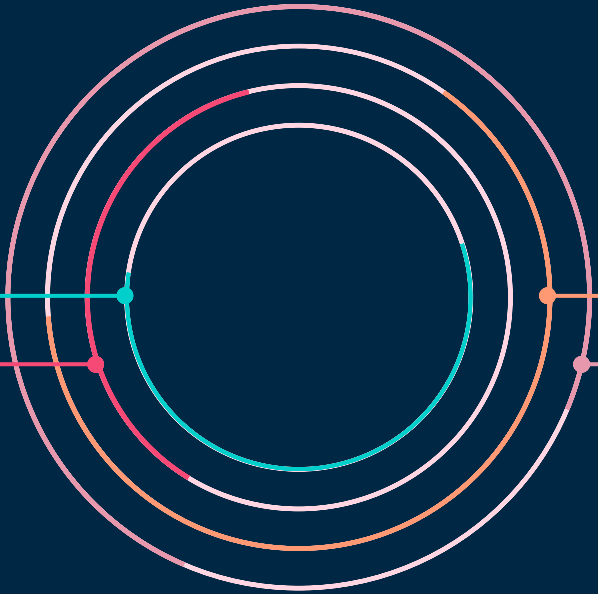
INSIGHTS AND POSSIBLE SOLUTIONS

Focus on increasing GDP

GOVERNMENT

COMMUNITY

Educational and awareness campaigns



More development of longer-lasting food

R&D

INDIVIDUAL

Recycling convenience

THANKS