Machine Learning Project

Approach for Investigating the Impact of Circular Economy on Taiwan's Economy and the Energy efficiency

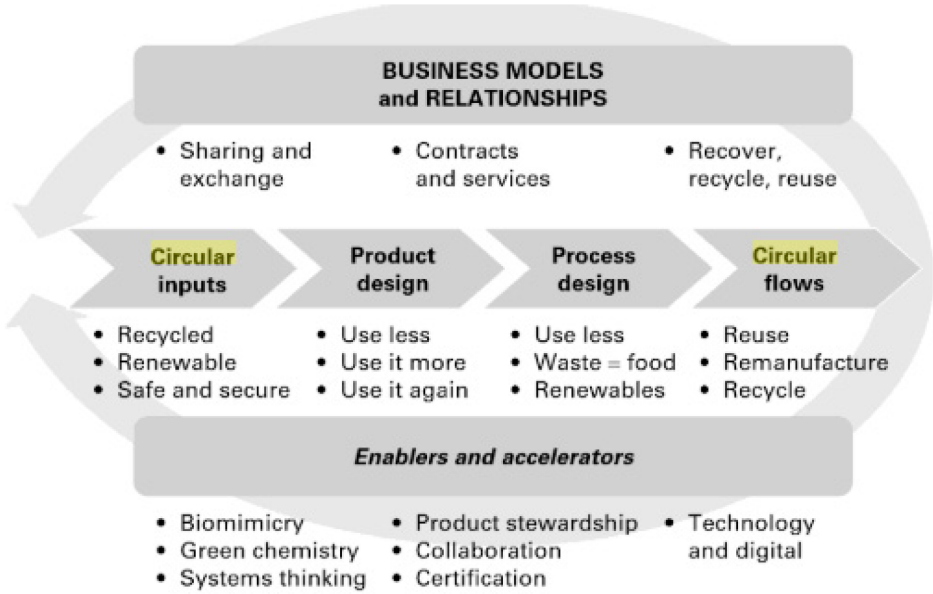
Yuze Tsai

May 17, 2025

Table of Contents

Introduction

- What is Circular Economy?
- Why this topic important?



SOURCE: © Catherine Weetman

References

Variables of interest

- Tanatu, A., et al.(2018): Resource productivity and domestic material consumption has a strong relationship with municipal waste recycling,
- The Ex'TAX project (2022) study: Taxshift can accelerate the transition to circular in construction industry
- Radivojevi´c, V, et al.(2024): A strong positive correlation between circular economy indicators and GDP per capita.

What we have right now...

- **Dependent Variables**: Total city/county income, Per capita electricity consumption, Vacancy Rates of House
- **Independent and Policy-Related Variables** Water pollution environmental tax, Air pollution environmental tax, General waste recycling rate...(total = 11)
- **Year of Interest**: 2014 - 2023 (10 years)

Data Manipualtion

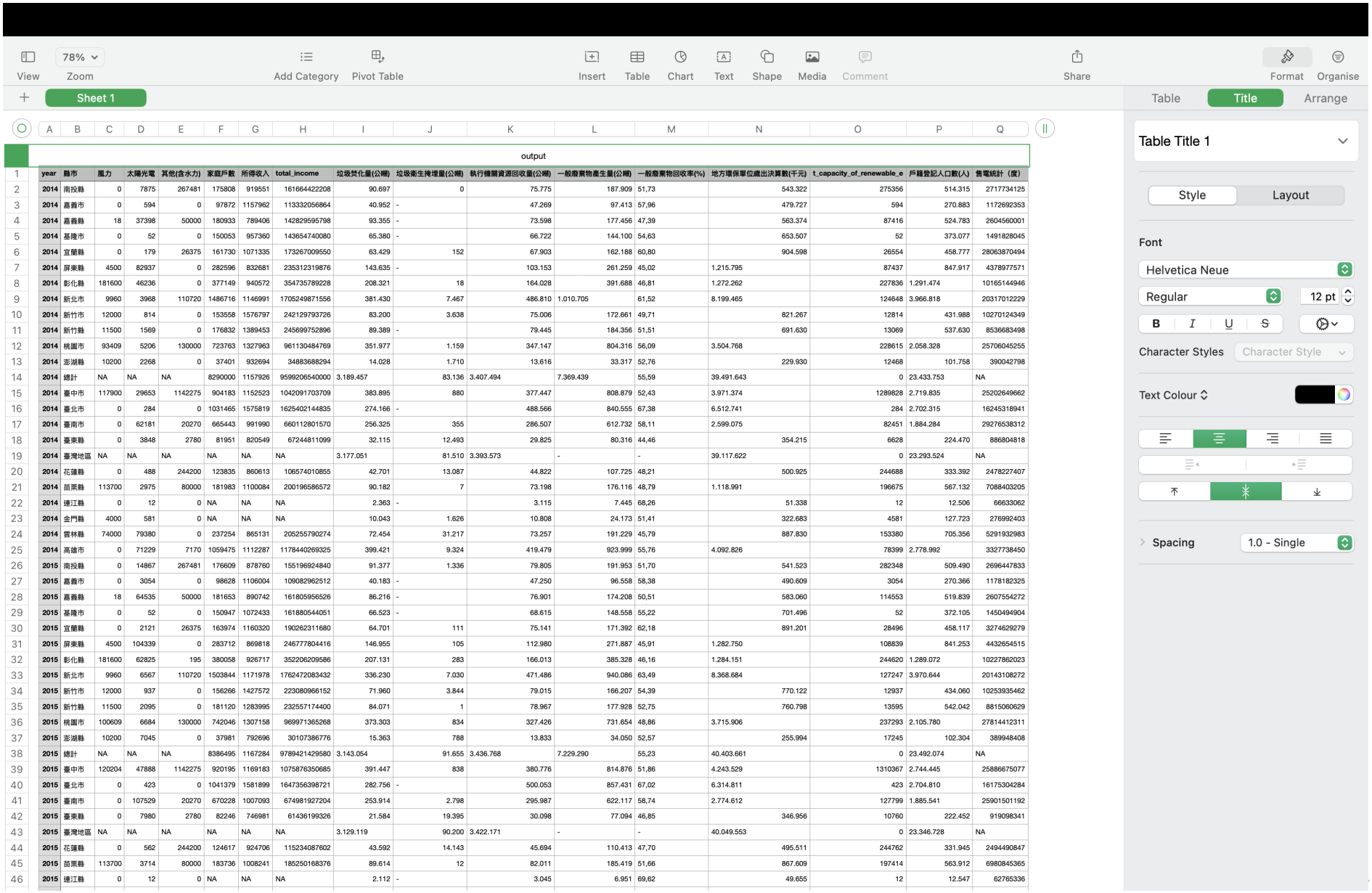
ML_Project.R*

Source on Save | Run | Source

```r
# Data manipulation

library(ggplot2)
library(tidyverse)
library(data.table)
library(readr)
install.packages("arrow")
library(arrow)
gc()

df_elec <- fread("~/Desktop/untitled folder/歷年各縣市再生能源裝置容量(101-11401).csv")
rm(df, data1)
df <- fread("~/Desktop/untitled folder/q15019-2140091076.csv")
df2 <- fread("~/Desktop/Result.csv")

df_elec[, year:= 年別+1911]
df_elec[, 1] = NULL
df[, year:= as.numeric(gsub("[^0-9]", "", 統計期)) + 1911]
df[, 1]= NULL
df[, total_income := as.numeric(家庭戶數)*as.numeric(所得收入)]
df_elec <- df_elec[-(287:308), ]
df_elec[, 縣市 := gsub("台北市", "臺北市", 縣市)]
df_elec[, 縣市 := gsub("台中市", "臺中市", 縣市)]
df_elec[, 縣市 := gsub("台南市", "臺南市", 縣市)]
df_elec[, 縣市 := gsub("台東縣", "臺東縣", 縣市)]

result <- merge(df_elec, df, by.x = c("year", "縣市"), by.y = c("year", "地區別"), all = TRUE)
colnames(df2) <- c('year','縣市', '垃圾焚化量(公噸)',
                   '垃圾衛生掩埋量(公噸)',
                   '執行機關資源回收量(公噸)',
                   '一般廢棄物產生量(公噸)',
                   '一般廢棄物回收率(%)',
                   '地方環保單位歲出決算數(千元)')
df2 <- df2[-(1), ]
df_final <- merge(result, df2, by = c("year", "縣市"), all = T)
rm(df, df_elec, df2)
gc()
df_final[, t_capacity_of_renewable_e := rowSums(.SD, na.rm = TRUE), .SDcols = c("風力", "太陽光電", "其他(含水力)")]


write_csv(df_final, "~/Desktop/untitled folder/output.csv")


people <- fread("~/Desktop/people.csv")
elec <- fread("~/Desktop/untitled folder/elec.csv")
df_transit <- merge(people, elec, by = c("year", "縣市"), all = T)
df_transit <- df_transit[, 1:4]
rm(people, elec)
df_final <- merge(df_final, df_transit, by = c("year", "縣市"), all = T)
df_final <- df_final[-(1:44), ]
df_final <- df_final[-(241:262), ]
write_csv(df_final, "~/Desktop/untitled folder/output.csv")
```

27:59 (Top Level) R Script

Data Manipualtion



78% View Zoom Add Category Pivot Table Insert Table Chart Text Shape Media Comment Share Format Organise

Sheet 1

Table Title Arrange

Table Title 1

Style Layout

Font

Helvetica Neue

Regular 12 pt

Character Styles Character Style

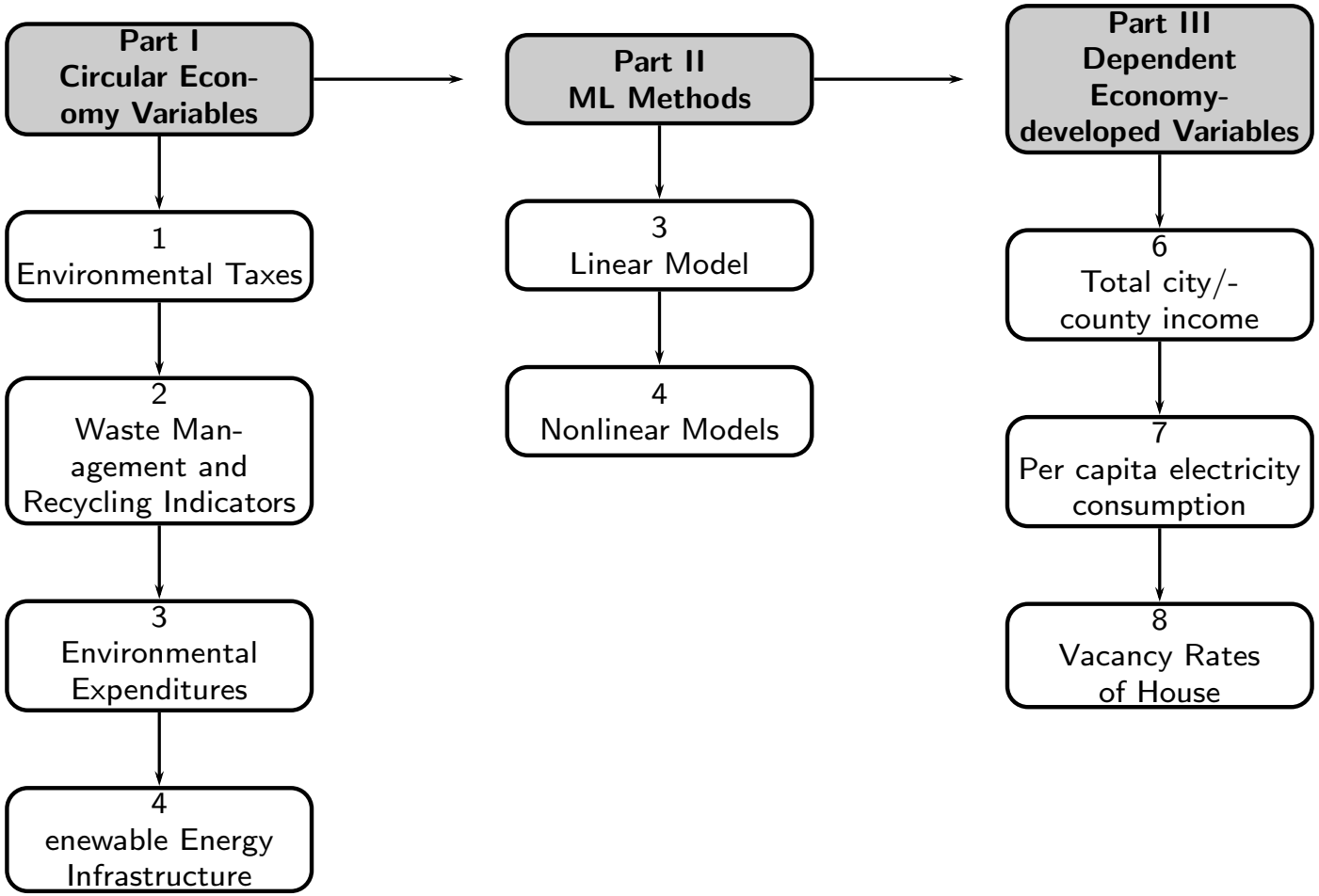
Text Colour

Spacing 1.0 - Single

output

year	縣市	風力	太陽光電	其他(含水力)	家庭戶數	所得收入	total_income	垃圾焚化量(公噸)	垃圾衛生掩埋量(公噸)	執行機關資源回收量(公噸)	一般廢棄物產生量(公噸)	一般廢棄物回收率(%)	地方環保單位歲出決算數(千元)	t_capacity_of_renewable_e	戶籍登記人口數(人)	售電統計（度）
2014	南投縣	0	7875	267481	175808	919551	161664422208	90.697	0	75.775	187.909	51,73	543.322	275356	514.315	2717734125
2014	嘉義市	0	594	0	97872	1157962	113332056864	40.952	-	47.269	97.413	57,96	479.727	594	270.883	1172692353
2014	嘉義縣	18	37398	50000	180933	789406	142829595798	93.355	-	73.598	177.456	47,39	563.374	87416	524.783	2604560001
2014	基隆市	0	52	0	150053	957360	143654740080	65.380	-	66.722	144.100	54,63	653.507	52	373.077	1491828045
2014	宜蘭縣	0	179	26375	161730	1071335	173267009550	63.429	152	67.903	162.188	60,80	904.598	26554	458.777	28063870494
2014	屏東縣	4500	82937	0	282596	832681	235312319876	143.635	-	103.153	261.259	45,02	1.215.795	87437	847.917	4378977571
2014	彰化縣	181600	46236	0	377149	940572	354735789228	208.321	18	164.028	391.688	46,81	1.272.262	227836	1.291.474	10165144946
2014	新北市	9960	3968	110720	1486716	1146991	1705249871556	381.430	7.467	486.810	1.010.705	61,52	8.199.465	124648	3.966.818	20317012229
2014	新竹市	12000	814	0	153558	1576797	242129793726	83.200	3.638	75.006	172.661	49,71	821.267	12814	431.988	10270124349
2014	新竹縣	11500	1569	0	176832	1389453	245699752896	89.389	-	79.445	184.356	51,51	691.630	13069	537.630	8536683498
2014	桃園市	93409	5206	130000	723763	1327963	961130484769	351.977	1.159	347.147	804.316	56,09	3.504.768	228615	2.058.328	25706045255
2014	澎湖縣	10200	2268	0	37401	932694	34883688294	14.028	1.710	13.616	33.317	52,76	229.930	12468	101.758	390042798
2014	總計	NA	NA	NA	8290000	1157926	9599206540000	3.189.457	83.136	3.407.494	7.369.439	55,59	39.491.643	0	23.433.753	NA
2014	臺中市	117900	29653	1142275	904183	1152523	1042091703709	383.895	880	377.447	808.879	52,43	3.971.374	1289828	2.719.835	25202649662
2014	臺北市	0	284	0	1031465	1575819	1625402144835	274.166	-	488.566	840.555	67,38	6.512.741	284	2.702.315	16245318941
2014	臺南市	0	62181	20270	665443	991990	660112801570	256.325	355	286.507	612.732	58,11	2.599.075	82451	1.884.284	29276538312
2014	臺東縣	0	3848	2780	81951	820549	67244811099	32.115	12.493	29.825	80.316	44,46	354.215	6628	224.470	886804818
2014	臺灣地區	NA	NA	NA	NA	NA	NA	3.177.051	81.510	3.393.573	-	-	39.117.622	0	23.293.524	NA
2014	花蓮縣	0	488	244200	123835	860613	106574010855	42.701	13.087	44.822	107.725	48,21	500.925	244688	333.392	2478227407
2014	苗栗縣	113700	2975	80000	181983	1100084	200196586572	90.182	7	73.198	176.116	48,79	1.118.991	196675	567.132	7088403205
2014	連江縣	0	12	0	NA	NA	NA	2.363	-	3.115	7.445	68,26	51.338	12	12.506	66633062
2014	金門縣	4000	581	0	NA	NA	NA	10.043	1.626	10.808	24.173	51,41	322.683	4581	127.723	276992403
2014	雲林縣	74000	79380	0	237254	865131	205255790274	72.454	31.217	73.257	191.229	45,79	887.830	153380	705.356	5291932983
2014	高雄市	0	71229	7170	1059475	1112287	1178440269325	399.421	9.324	419.479	923.999	55,76	4.092.826	78399	2.778.992	3327738450
2015	南投縣	0	14867	267481	176609	878760	155196924840	91.377	1.336	79.805	191.953	51,70	541.523	282348	509.490	2696447833
2015	嘉義市	0	3054	0	98628	1106004	109082962512	40.183	-	47.250	96.558	58,38	490.609	3054	270.366	1178182325
2015	嘉義縣	18	64535	50000	181653	890742	161805956526	86.216	-	76.901	174.208	50,51	583.060	114553	519.839	2607554272
2015	基隆市	0	52	0	150947	1072433	161880544051	66.523	-	68.615	148.558	55,22	701.496	52	372.105	1450494904
2015	宜蘭縣	0	2121	26375	163974	1160320	190262311680	64.701	111	75.141	171.392	62,18	891.201	28496	458.117	3274629279
2015	屏東縣	4500	104339	0	283712	869818	246777804416	146.955	105	112.980	271.887	45,91	1.282.750	108839	841.253	4432654515
2015	彰化縣	181600	62825	195	380058	926717	352206209586	207.131	283	166.013	385.328	46,16	1.284.151	244620	1.289.072	10227862023
2015	新北市	9960	6567	110720	1503844	1171978	1762472083432	336.230	7.030	471.486	940.086	63,49	8.368.684	127247	3.970.644	20143108272
2015	新竹市	12000	937	0	156266	1427572	223080966152	71.960	3.844	79.015	166.207	54,39	770.122	12937	434.060	10253935462
2015	新竹縣	11500	2095	0	181120	1283995	232557174400	84.071	1	78.967	177.928	52,75	760.798	13595	542.042	8815060629
2015	桃園市	100609	6684	130000	742046	1307158	969971365268	373.303	834	327.426	731.654	48,86	3.715.906	237293	2.105.780	27814412311
2015	澎湖縣	10200	7045	0	37981	792696	30107386776	15.363	788	13.833	34.050	52,57	255.994	17245	102.304	389948408
2015	總計	NA	NA	NA	8386495	1167284	9789421429580	3.143.054	91.655	3.436.768	7.229.290	55,23	40.403.661	0	23.492.074	NA
2015	臺中市	120204	47888	1142275	920195	1169183	1075876350685	391.447	838	380.776	814.876	51,86	4.243.529	1310367	2.744.445	25886675077
2015	臺北市	0	423	0	1041379	1581899	1647356398721	282.756	-	500.053	857.431	67,02	6.314.811	423	2.704.810	16175304284
2015	臺南市	0	107529	20270	670228	1007093	674981927204	253.914	2.798	295.987	622.117	58,74	2.774.612	127799	1.885.541	25901501192
2015	臺東縣	0	7980	2780	82246	746981	61436199326	21.584	19.395	30.098	77.094	46,85	346.956	10760	222.452	919098341
2015	臺灣地區	NA	NA	NA	NA	NA	NA	3.129.119	90.200	3.422.171	-	-	40.049.553	0	23.346.728	NA
2015	花蓮縣	0	562	244200	124617	924706	115234087602	43.592	14.143	45.694	110.413	47,70	495.511	244762	331.945	2494490847
2015	苗栗縣	113700	3714	80000	183736	1008241	185250168376	89.614	12	82.011	185.419	51,66	867.609	197414	563.912	6980845365
2015	連江縣	0	12	0	NA	NA	NA	2.112	-	3.045	6.951	69,62	49.655	12	12.547	62765336

Flow Chart of Study Structure



Linear Regression Method – Panel Data Regression

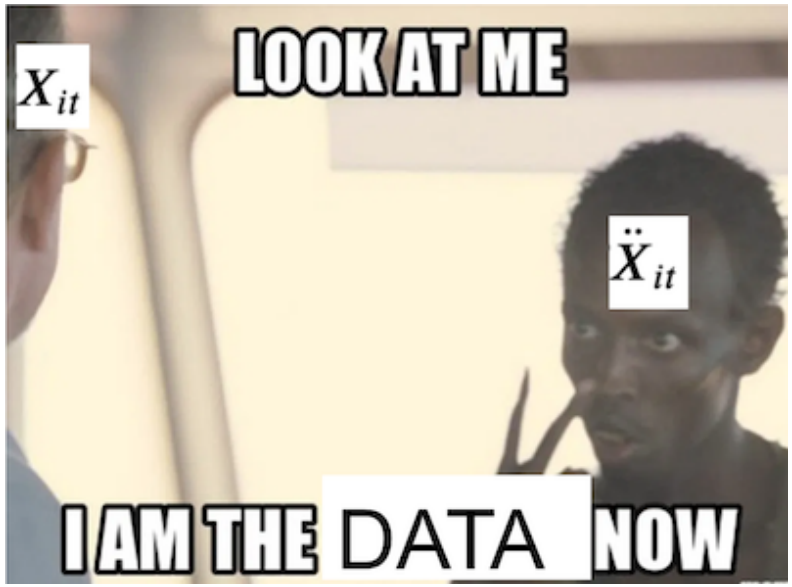
$$y_{it} = a + X'_{it}\beta + c_i + e_{it} \tag{1}$$

where we define:

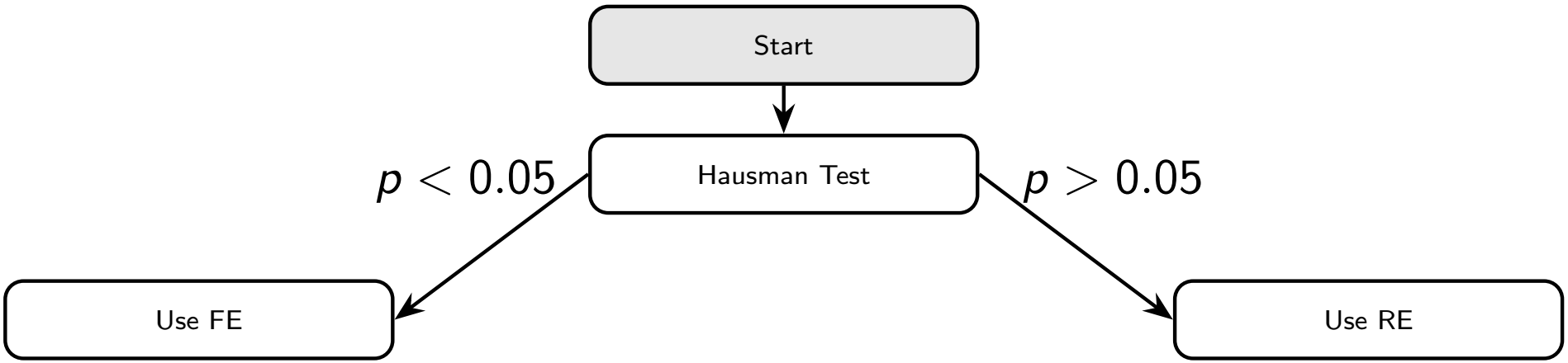
- $i = 1, \ldots, N$, where i represents a county/city, and we have N cross-sectional data ($N = 22$).
- $t = 2013, \ldots, 2023$.
- y_{it} is the dependent variable.
- X'_{it} are the independent variables of K-dimension.
- a is the intercept.
- β is the coefficient vector of K-dimension.
- c_i is the individual effect.
- e_{it} is the error term.

Fixed Effects vs. Random Effects

Feature	**Fixed Effects (FE)**	**Random Effects (RE)**
Individual Effect c_i	Fixed (correlated with X_{it})	Random (uncorrelated)
Main Assumption	Controls for time-invariant differences	Individual effects are uncorrelated
Interpretation	Within-group variation	Between and within variation
When to Use	If c_i is correlated with X_{it}	If c_i is random and uncorrelated



Panel Data Model Selection



Nonlinear regression method – XGBoost

XGBoost (Extreme Gradient Boosting) is an ameliorated gradient boosting method that efficiently handles large datasets and has performed exceptionally well in many competitions.

- Uses **gradient boosting** to iteratively minimize errors.

Mathematical Model:

$$\text{Loss} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where:

- The first term represents prediction error.
- The second term is a regularization term controlling model complexity.

Pros and Cons:

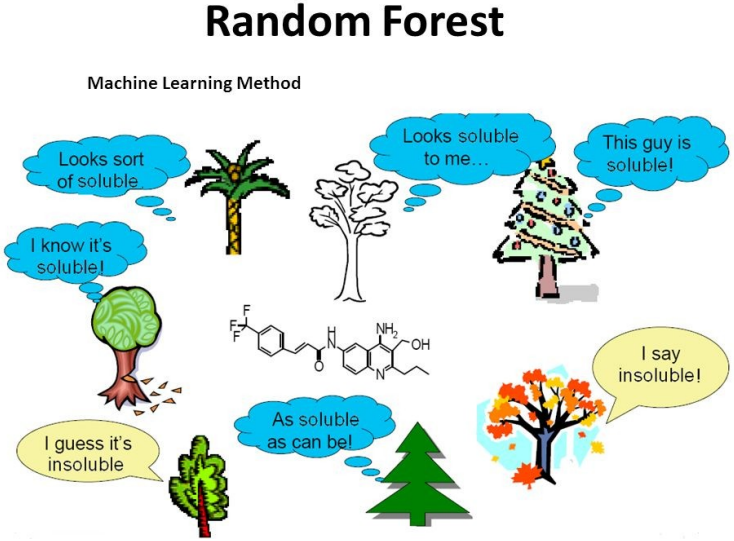
- Pros: High accuracy, fast training.
- Cons: Complex parameter tuning.

Nonlinear regression method – Random Forest

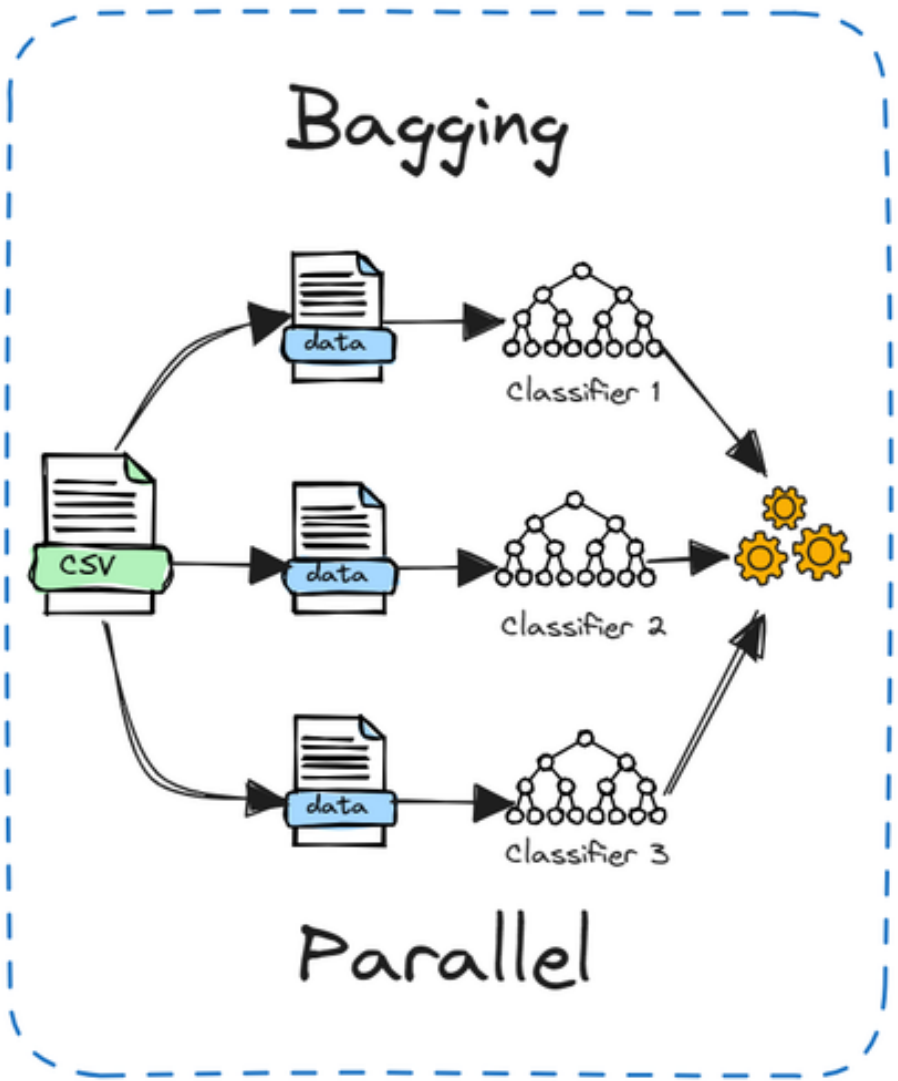
Random Forest is an **ensemble learning** method consisting of multiple decision trees. The final prediction is based on **voting (classification)** or **averaging (regression)**.

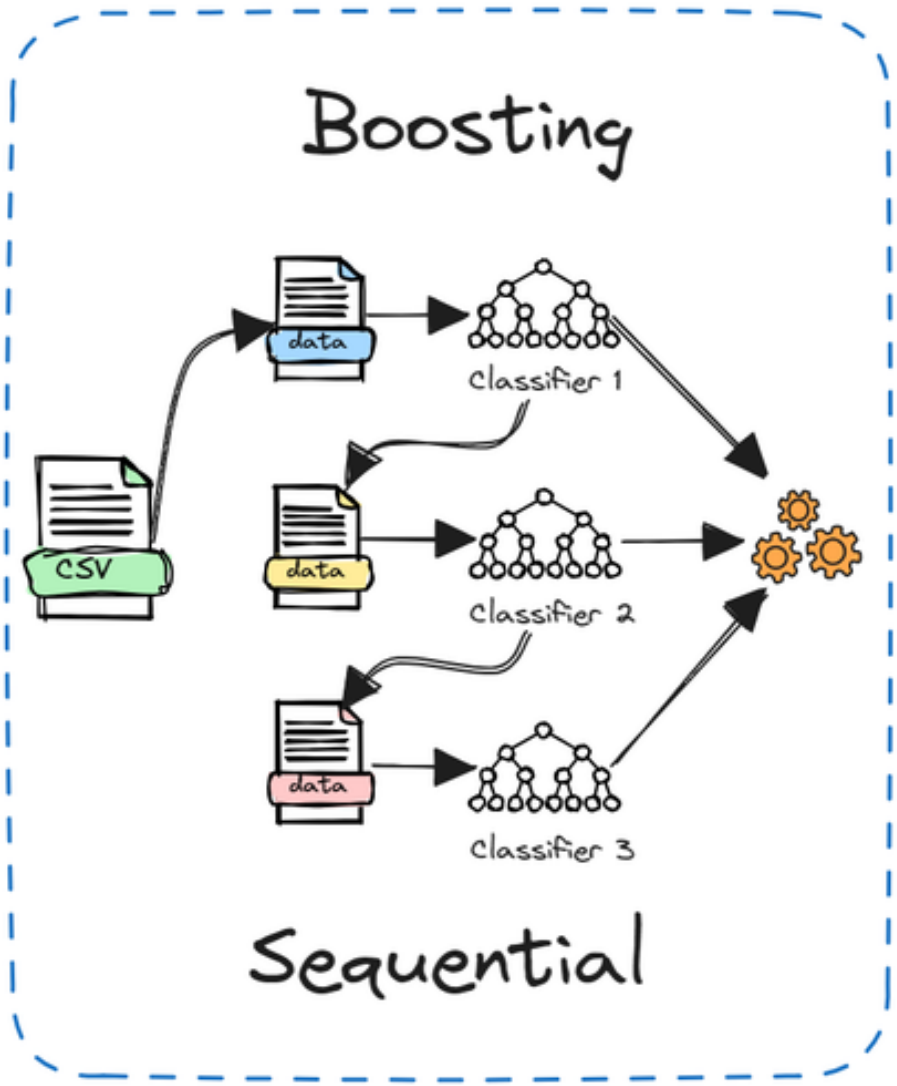
Pros and Cons:

- Pros: Handles high-dimensional data, reduces overfitting.
- Cons: Long training time, less interpretable.



Boosting vs Bagging





Nonlinear regression method – Support Vector Machine (SVM)

SVM is a supervised learning algorithm primarily used for classification.

- It finds the **optimal hyperplane** that best separates different classes.
- Uses **kernel functions** to handle non-linear classification.

Mathematical Formulation:

$$\max_{\mathbf{w},b} \quad \frac{1}{\|\mathbf{w}\|}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \forall i$$

Pros and Cons:

- Pros: Effective in small sample, high-dimensional data.
- Cons: Computationally expensive, slow for large datasets.

Comparison of ML Methods

Feature	SVM	Random Forest	XGBoost
Speed	Slow	Medium	Fast
Accuracy	High	High	Very High
Interpretability	Low	Medium	Low
Handles Large Data	No	Yes	Yes

Intuitive Clairvoyance...

- Environmental budgets and recycling efforts have a positive relationship with total income of a city.
- Cities that invested more in renewable energy infrastructure will generate some critical improvements in energy efficiency.
- Higher levels of incineration and landfill usage are associated with lower energy efficiency, possibly due to poor recycling practices.

Hausman Test for Model Selection

- Null Hypothesis: $H_0 : \mathbb{E}(c_i|X_{it}) = 0 \Rightarrow$ Random Effects (RE)
- Alternative Hypothesis: $H_1 : \mathbb{E}(c_i|X_{it}) \neq 0 \Rightarrow$ Fixed Effects (FE)
- Based on the Hausman test, FE model is chosen if $p < 0.05$, suggesting correlation between individual effects and regressors.
- In this study, FE models are generally preferred due to practical data nature.

Model Evaluation and Comparison

- **Performance Metrics:**
  - Mean Squared Error (MSE)
  - R-Squared (R^2)
  - Cross-validation (10-fold)
- **Findings:**
  - Nonlinear models generally outperform linear ones.
  - CatBoost achieved the highest R^2 for household income and vacancy rate.
  - SVM showed poor predictive performance; excluded from SHAP analysis.

SHAP (SHapley Additive Explanations)

- SHAP assigns each feature an importance value for prediction.
- Based on cooperative game theory—accounts for all possible feature combinations.
- **Advantages:**
  - Global and local interpretability
  - Indicates both direction (+/) and magnitude of effect
  - Helps identify nonlinear thresholds and interactions
- **Applied to:** XGBoost, CatBoost, LightGBM, Random Forest