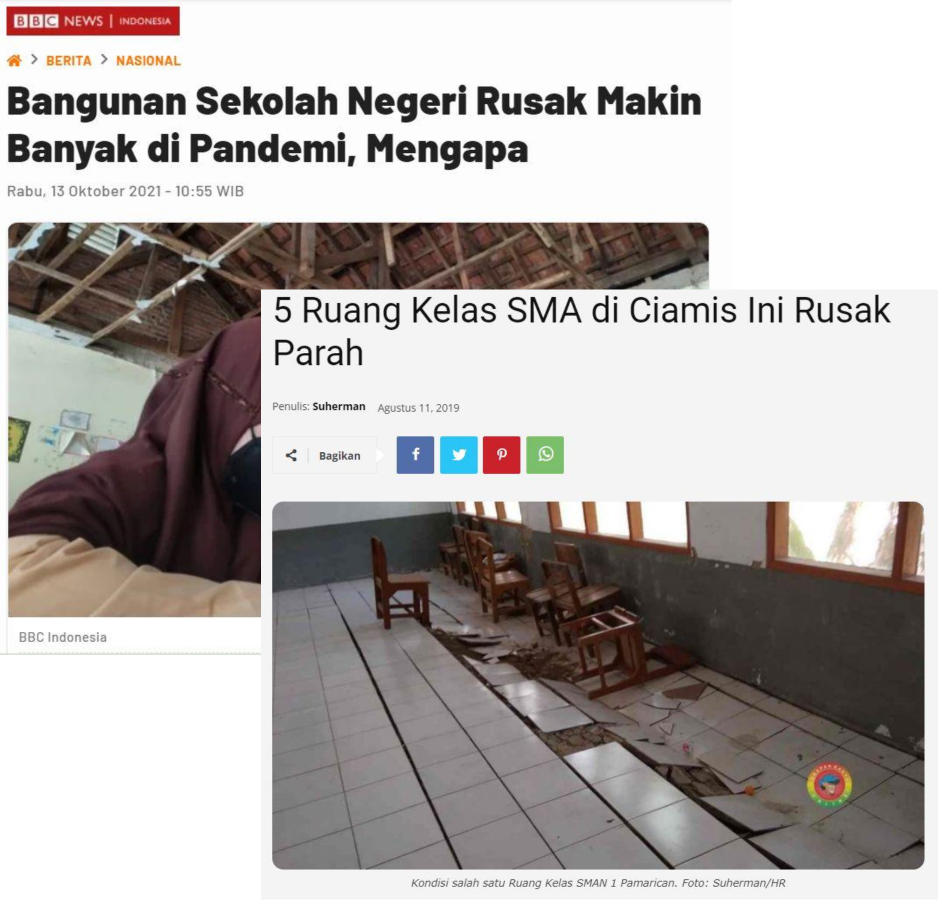
Clustering and Segmentation for High School Facilities in Indonesia

Based on facilities condition in 2020/2021 released by Kemendikbud

prepared by Mikael Dewabrata

Problem Definition

School facilities in Indonesia is a big issues that need to be reinstated. Even we have the data, but with big school numbers across all country, it need a better solution to segment the severity of the condition. With data science, we hope that we can cluster the facilities condition for each district or Kecamatan.



BBC NEWS | INDONESIA

> BERITA > NASIONAL

Bangunan Sekolah Negeri Rusak Makin Banyak di Pandemi, Mengapa

Rabu, 13 Oktober 2021 - 10:55 WIB

BBC Indonesia

5 Ruang Kelas SMA di Ciamis Ini Rusak Parah

Penulis: **Suherman** Agustus 11, 2019

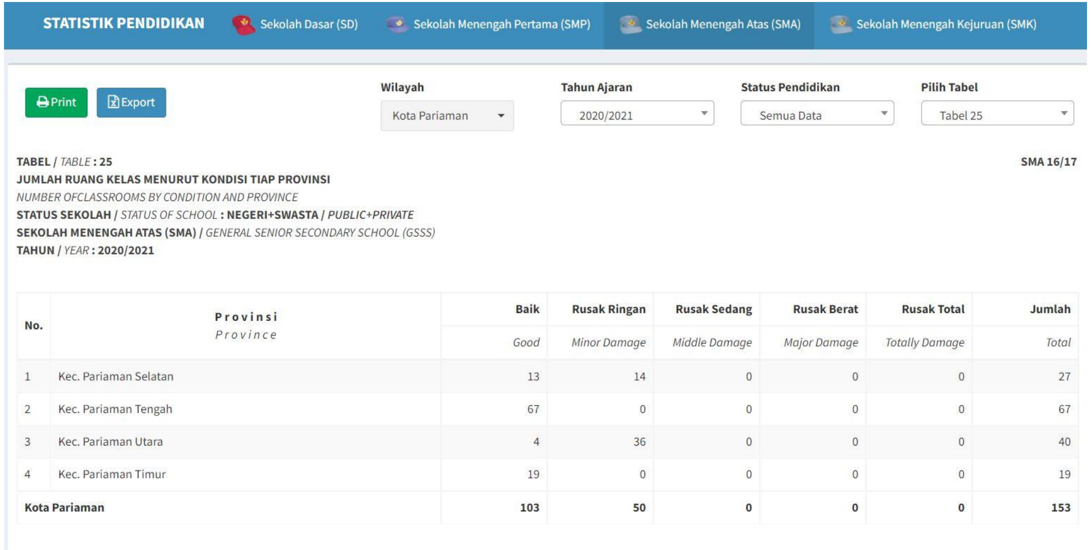
Kondisi salah satu Ruang Kelas SMAN 1 Pamarican. Foto: Suherman/HR

Business Question

- How can data science help to manage the situation about the school facilities
- What methodology that data science can be used and on what problem it can be solved
- How data science can make help to solve the problem more efficient and faster

Data Selection

For this project, the data is taken from Kemendikbud for table across facilities available (UKS, Toilet, Library, Classroom) for High School all over Indonesia


STATISTIK PENDIDIKAN
Sekolah Dasar (SD)
Sekolah Menengah Pertama (SMP)
Sekolah Menengah Atas (SMA)
Sekolah Menengah Kejuruan (SMK)
Print
Export

Wilayah: Kota Pariaman

Tahun Ajaran: 2020/2021

Status Pendidikan: Semua Data

Pilih Tabel: Tabel 25

TABEL / *TABLE* **: 25**

SMA 16/17

JUMLAH RUANG KELAS MENURUT KONDISI TIAP PROVINSI

NUMBER OFCLASSROOMS BY CONDITION AND PROVINCE

STATUS SEKOLAH / *STATUS OF SCHOOL* **: NEGERI+SWASTA** / *PUBLIC+PRIVATE*

SEKOLAH MENENGAH ATAS (SMA) / *GENERAL SENIOR SECONDARY SCHOOL (GSSS)*

TAHUN / *YEAR* **: 2020/2021**

No.	Provinsi	Baik	Rusak Ringan	Rusak Sedang	Rusak Berat	Rusak Total	Jumlah
	Province	*Good*	*Minor Damage*	*Middle Damage*	*Major Damage*	*Totally Damage*	*Total*
1	Kec. Pariaman Selatan	13	14	0	0	0	27
2	Kec. Pariaman Tengah	67	0	0	0	0	67
3	Kec. Pariaman Utara	4	36	0	0	0	40
4	Kec. Pariaman Timur	19	0	0	0	0	19
Kota Pariaman		**103**	**50**	**0**	**0**	**0**	**153**

* taken from statistik.data.kemendikbud.go.id

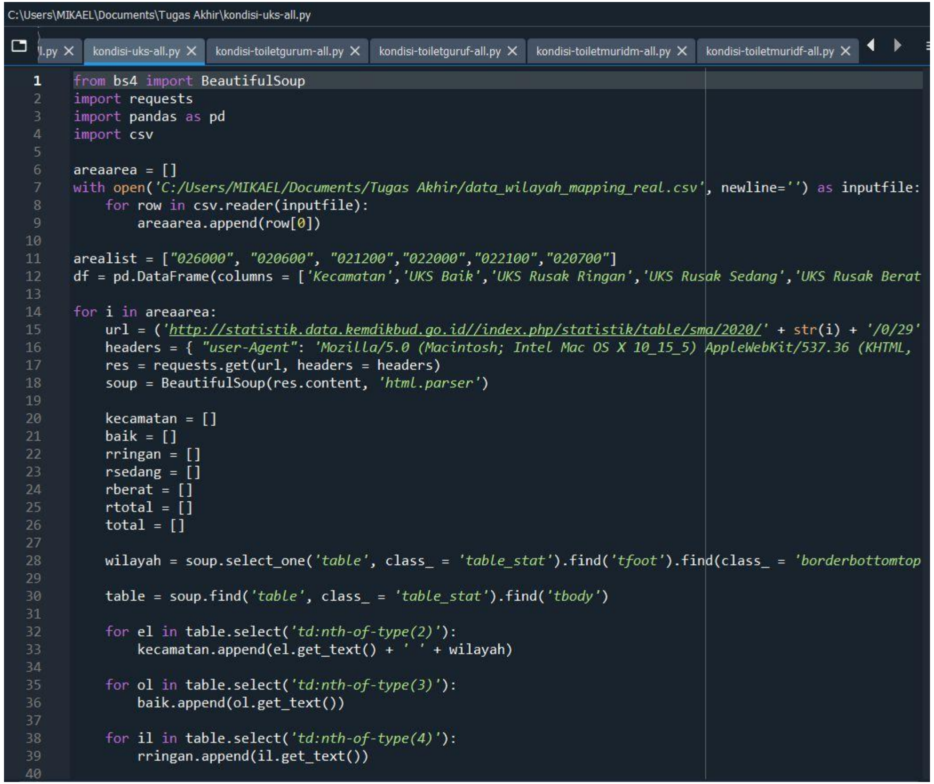
Data Problem

The data from Kemendikbud website is unstructured and separated to different table. E.g for toilet condition they have their own table. The website also doesn’t provide the data in csv or xlsx format.

Data Collection

To get the data, scraping is one method needed. Only relevant tables to the analysis subject are selected. And, since the data separated not only per category table, it also separated per region. So, the data scraped one by one through different category and different region.

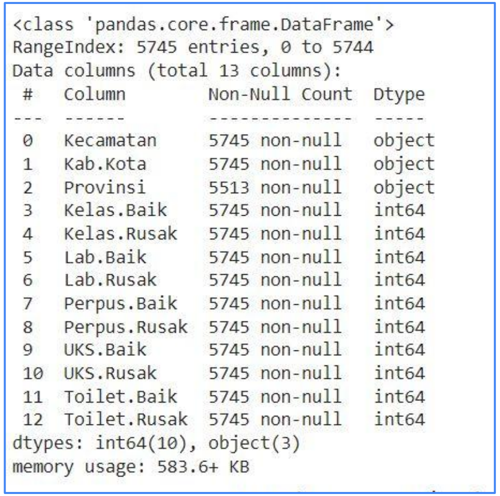
Scraping method done using Beautiful Soup through Python, stored to raw data in CSV. Total rows extracted **5745** from **5745** kecamatan for **9 categories**.


C:\Users\MIKAEL\Documents\Tugas Akhir\kondisi-uks-all.py
l.py
kondisi-uks-all.py
kondisi-toiletgurum-all.py
kondisi-toiletguruf-all.py
kondisi-toiletmuridm-all.py
kondisi-toiletmuridf-all.py

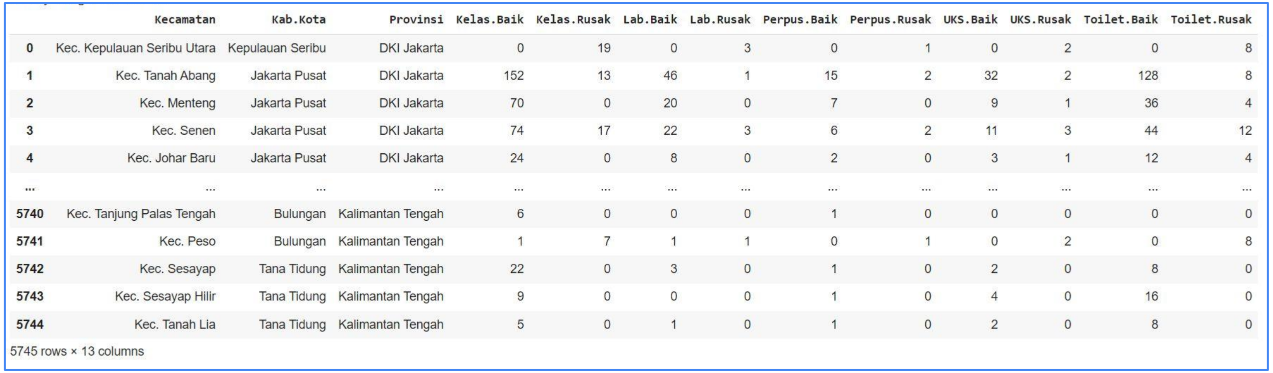
```python
from bs4 import BeautifulSoup
import requests
import pandas as pd
import csv

areaarea = []
with open('C:/Users/MIKAEL/Documents/Tugas Akhir/data_wilayah_mapping_real.csv', newline='') as inputfile:
    for row in csv.reader(inputfile):
        areaarea.append(row[0])

arealist = ["026000", "020600", "021200","022000","022100","020700"]
df = pd.DataFrame(columns = ['Kecamatan','UKS Baik','UKS Rusak Ringan','UKS Rusak Sedang','UKS Rusak Berat

for i in areaarea:
    url = ('http://statistik.data.kemdikbud.go.id//index.php/statistik/table/sma/2020/' + str(i) + '/0/29'
    headers = { "user-Agent": 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_5) AppleWebKit/537.36 (KHTML,
    res = requests.get(url, headers = headers)
    soup = BeautifulSoup(res.content, 'html.parser')

    kecamatan = []
    baik = []
    rringan = []
    rsedang = []
    rberat = []
    rtotal = []
    total = []

    wilayah = soup.select_one('table', class_ = 'table_stat').find('tfoot').find(class_ = 'borderbottomtop

    table = soup.find('table', class_ = 'table_stat').find('tbody')

    for el in table.select('td:nth-of-type(2)'):
        kecamatan.append(el.get_text() + ' ' + wilayah)

    for ol in table.select('td:nth-of-type(3)'):
        baik.append(ol.get_text())

    for il in table.select('td:nth-of-type(4)'):
        rringan.append(il.get_text())

```

Data Understanding & Cleaning

Once the data already collected, cleaning and wrangling process done in R to combine all tables so it would be relevant for analysis.



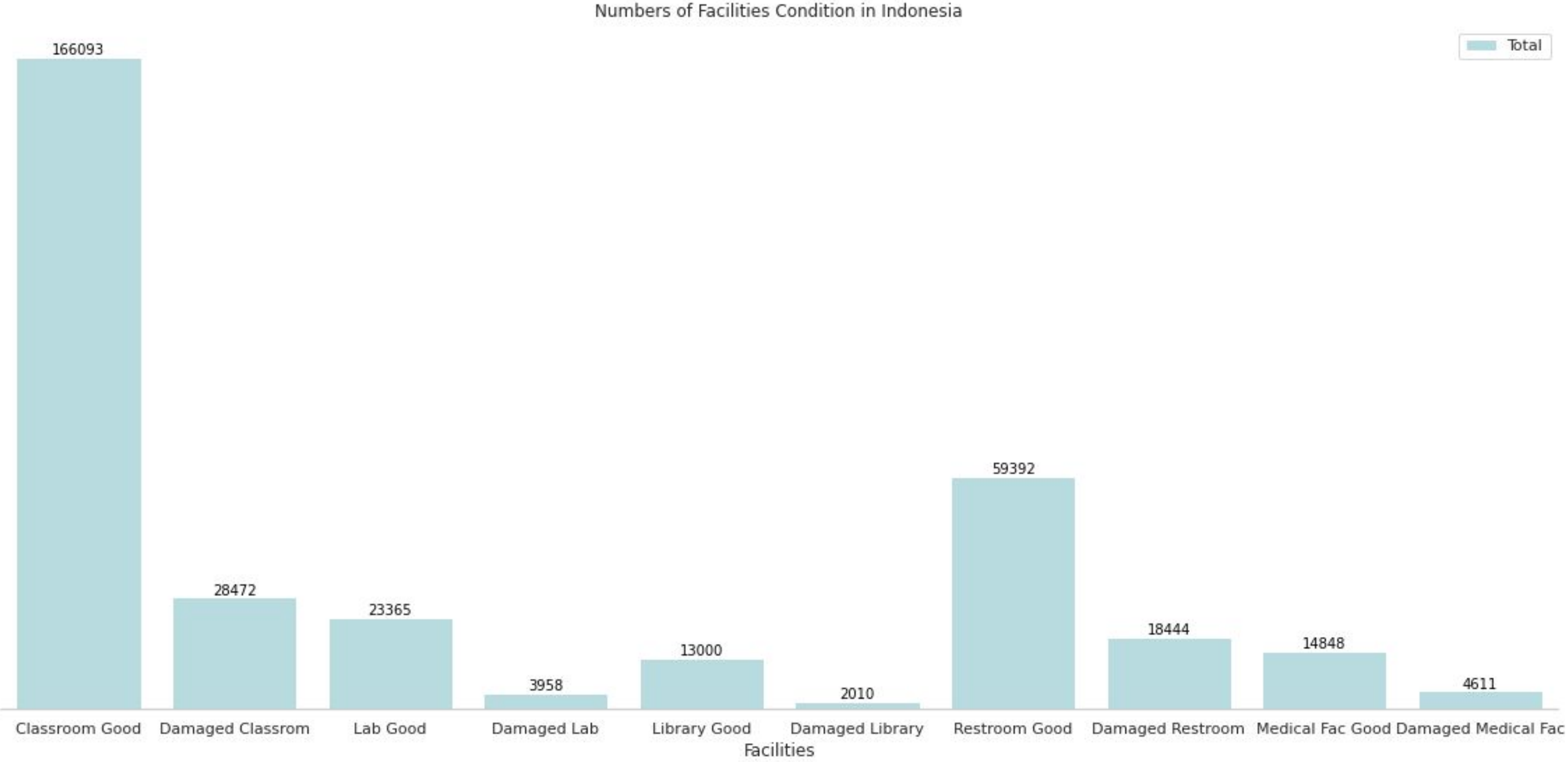
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5745 entries, 0 to 5744
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Kecamatan     5745 non-null   object
 1   Kab.Kota      5745 non-null   object
 2   Provinsi      5513 non-null   object
 3   Kelas.Baik    5745 non-null   int64
 4   Kelas.Rusak   5745 non-null   int64
 5   Lab.Baik      5745 non-null   int64
 6   Lab.Rusak     5745 non-null   int64
 7   Perpus.Baik   5745 non-null   int64
 8   Perpus.Rusak  5745 non-null   int64
 9   UKS.Baik      5745 non-null   int64
 10  UKS.Rusak     5745 non-null   int64
 11  Toilet.Baik   5745 non-null   int64
 12  Toilet.Rusak  5745 non-null   int64
dtypes: int64(10), object(3)
memory usage: 583.6+ KB
```



	Kecamatan	Kab.Kota	Provinsi	Kelas.Baik	Kelas.Rusak	Lab.Baik	Lab.Rusak	Perpus.Baik	Perpus.Rusak	UKS.Baik	UKS.Rusak	Toilet.Baik	Toilet.Rusak
0	Kec. Kepulauan Seribu Utara	Kepulauan Seribu	DKI Jakarta	0	19	0	3	0	1	0	2	0	8
1	Kec. Tanah Abang	Jakarta Pusat	DKI Jakarta	152	13	46	1	15	2	32	2	128	8
2	Kec. Menteng	Jakarta Pusat	DKI Jakarta	70	0	20	0	7	0	9	1	36	4
3	Kec. Senen	Jakarta Pusat	DKI Jakarta	74	17	22	3	6	2	11	3	44	12
4	Kec. Johar Baru	Jakarta Pusat	DKI Jakarta	24	0	8	0	2	0	3	1	12	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5740	Kec. Tanjung Palas Tengah	Bulungan	Kalimantan Tengah	6	0	0	0	1	0	0	0	0	0
5741	Kec. Peso	Bulungan	Kalimantan Tengah	1	7	1	1	0	1	0	2	0	8
5742	Kec. Sesayap	Tana Tidung	Kalimantan Tengah	22	0	3	0	1	0	2	0	8	0
5743	Kec. Sesayap Hilir	Tana Tidung	Kalimantan Tengah	9	0	0	0	1	0	4	0	16	0
5744	Kec. Tanah Lia	Tana Tidung	Kalimantan Tengah	5	0	1	0	1	0	2	0	8	0

5745 rows × 13 columns

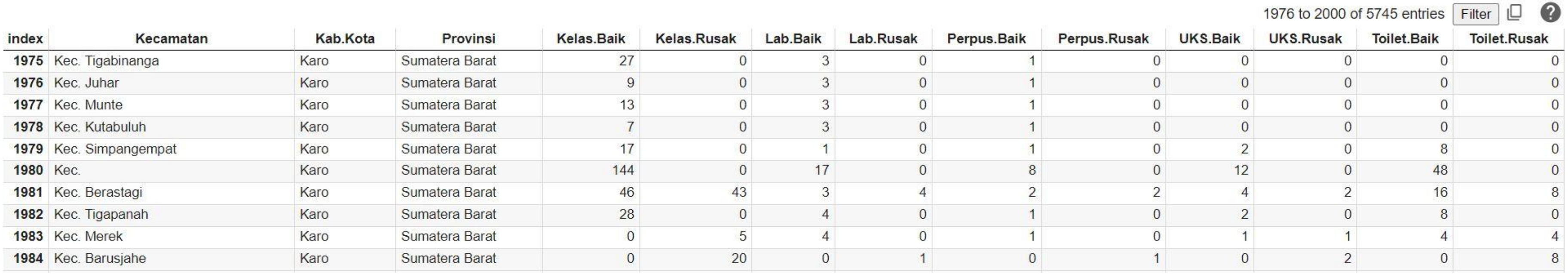
EDA - Data Summary



EDA - Data Summary

- Based the data, **28472** classes are damaged, it’s 14% of total classroom
- **18444** restrooms (for male or female) are damaged, 23% of total restroom
- Jawa Barat is the province with most damaged classroom **(3829)**, while Aceh is the most damaged outside Java **(2960)** - 20% from total classroom
- Papua is the province with most damage restroom outside Java **(804)** - 22% of total restroom
- Percentage wise, Bengkulu has the highest damaged facilities across all categories, with all categories have above 35% damaged facilities

Give the data the same weight



1976 to 2000 of 5745 entries

index	Kecamatan	Kab.Kota	Provinsi	Kelas.Baik	Kelas.Rusak	Lab.Baik	Lab.Rusak	Perpus.Baik	Perpus.Rusak	UKS.Baik	UKS.Rusak	Toilet.Baik	Toilet.Rusak
1975	Kec. Tigabinanga	Karo	Sumatera Barat	27	0	3	0	1	0	0	0	0	0
1976	Kec. Juhar	Karo	Sumatera Barat	9	0	3	0	1	0	0	0	0	0
1977	Kec. Munte	Karo	Sumatera Barat	13	0	3	0	1	0	0	0	0	0
1978	Kec. Kutabuluh	Karo	Sumatera Barat	7	0	3	0	1	0	0	0	0	0
1979	Kec. Simpangempat	Karo	Sumatera Barat	17	0	1	0	1	0	2	0	8	0
1980	Kec.	Karo	Sumatera Barat	144	0	17	0	8	0	12	0	48	0
1981	Kec. Berastagi	Karo	Sumatera Barat	46	43	3	4	2	2	4	2	16	8
1982	Kec. Tigapanah	Karo	Sumatera Barat	28	0	4	0	1	0	2	0	8	0
1983	Kec. Merek	Karo	Sumatera Barat	0	5	4	0	1	0	1	1	4	4
1984	Kec. Barusjahe	Karo	Sumatera Barat	0	20	0	1	0	1	0	2	0	8

Since numbers of damaged facilities cannot be same for each Kecamatan, we need to put the same weight across all values.

E.g despite the numbers of damaged class room in Kecamatan Berastagi (43) is bigger than in Kecamatan Barusjahe (20), but in Barusjahe is more urgent to be fixed since it has bigger damaged class than class in good condition (0). It means all the class are broken in Kecamatan Barusjahe!

That’s why the data need to be processed in order to find a more balanced number.

Damage Rate

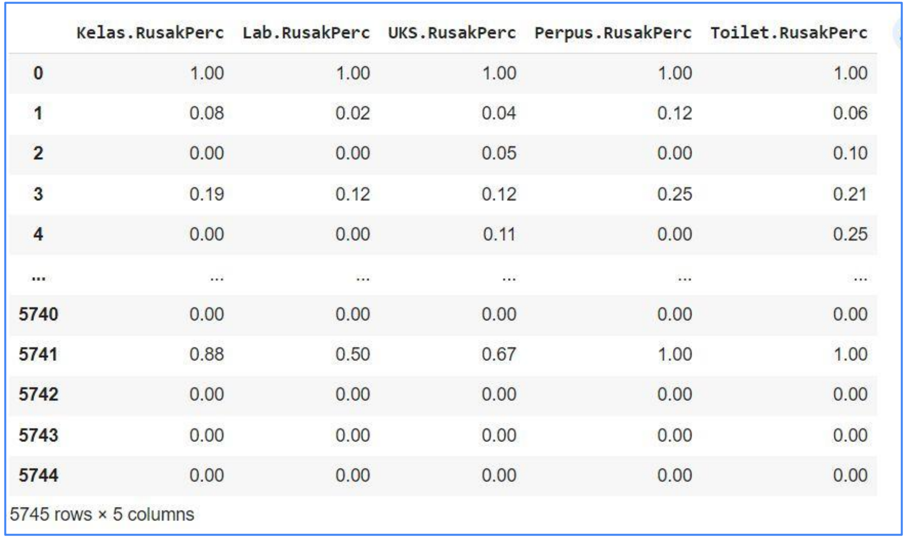
To make the data more balanced, the number then converted to damage rate.

Damage Rate formula:

Total Damage / (Total damage + Total Good Condition)

R formula:

round(df_clean$Lab.Rusak / (df_clean$Lab.Rusak + df_trim$Lab.Baik), digits = 2)



	Kelas.RusakPerc	Lab.RusakPerc	UKS.RusakPerc	Perpus.RusakPerc	Toilet.RusakPerc
0	1.00	1.00	1.00	1.00	1.00
1	0.08	0.02	0.04	0.12	0.06
2	0.00	0.00	0.05	0.00	0.10
3	0.19	0.12	0.12	0.25	0.21
4	0.00	0.00	0.11	0.00	0.25
...	...	...	...	...	...
5740	0.00	0.00	0.00	0.00	0.00
5741	0.88	0.50	0.67	1.00	1.00
5742	0.00	0.00	0.00	0.00	0.00
5743	0.00	0.00	0.00	0.00	0.00
5744	0.00	0.00	0.00	0.00	0.00

5745 rows × 5 columns

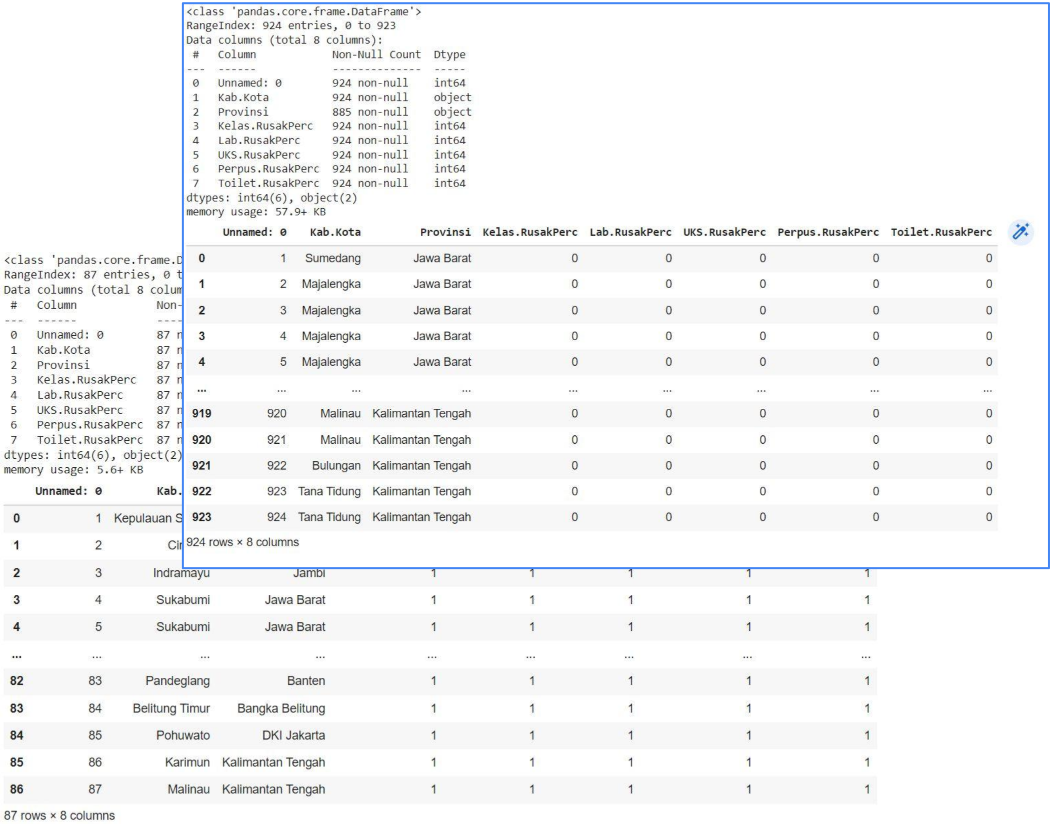
** Based on table above, 1 is having all damaged, while 0 is all in good condition for the total of all facilities*

Removing All High and Low Damage Rate

There are Kecamatan that have all facilities are all broken and all good. For these rows, we are going to remove from clustering since these rows are obvious.

We just put them as lowest and highest priorities. No need to put into the clustering calculation.

The summary as follow:
Good Condition: **2209 kecamatan**
Damaged: **87 kecamatan**



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 924 entries, 0 to 923
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Unnamed: 0          924 non-null    int64 
 1   Kab.Kota            924 non-null    object
 2   Provinsi            885 non-null    object
 3   Kelas.RusakPerc     924 non-null    int64 
 4   Lab.RusakPerc       924 non-null    int64 
 5   UKS.RusakPerc       924 non-null    int64 
 6   Perpus.RusakPerc    924 non-null    int64 
 7   Toilet.RusakPerc    924 non-null    int64 
dtypes: int64(6), object(2)
memory usage: 57.9+ KB
```

	Unnamed: 0	Kab.Kota	Provinsi	Kelas.RusakPerc	Lab.RusakPerc	UKS.RusakPerc	Perpus.RusakPerc	Toilet.RusakPerc
0	1	Sumedang	Jawa Barat	0	0	0	0	0
1	2	Majalengka	Jawa Barat	0	0	0	0	0
2	3	Majalengka	Jawa Barat	0	0	0	0	0
3	4	Majalengka	Jawa Barat	0	0	0	0	0
4	5	Majalengka	Jawa Barat	0	0	0	0	0
...	...	...	...	...	...	...	...	...
919	920	Malinau	Kalimantan Tengah	0	0	0	0	0
920	921	Malinau	Kalimantan Tengah	0	0	0	0	0
921	922	Bulungan	Kalimantan Tengah	0	0	0	0	0
922	923	Tana Tidung	Kalimantan Tengah	0	0	0	0	0
923	924	Tana Tidung	Kalimantan Tengah	0	0	0	0	0

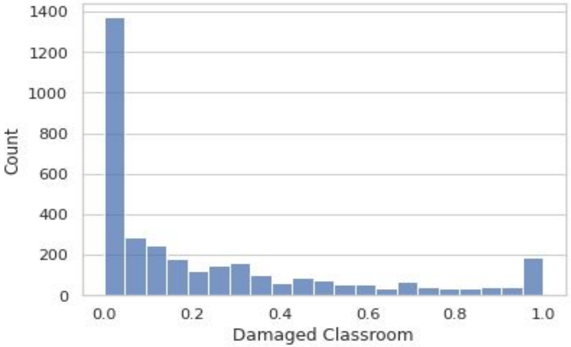
924 rows × 8 columns

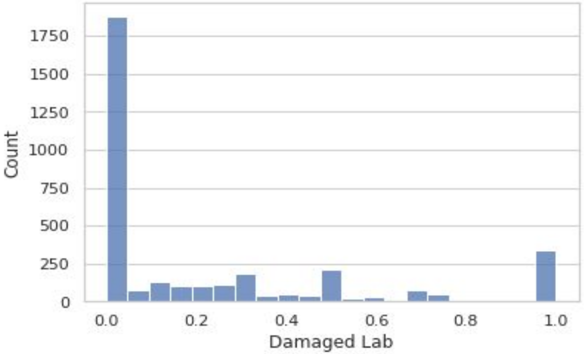
```
<class 'pandas.core.frame.D
RangeIndex: 87 entries, 0 t
Data columns (total 8 colum
 #   Column              Non-
---  ------              ----
 0   Unnamed: 0          87 n
 1   Kab.Kota            87 n
 2   Provinsi            87 n
 3   Kelas.RusakPerc     87 n
 4   Lab.RusakPerc       87 n
 5   UKS.RusakPerc       87 n
 6   Perpus.RusakPerc    87 n
 7   Toilet.RusakPerc    87 n
dtypes: int64(6), object(2)
memory usage: 5.6+ KB
```

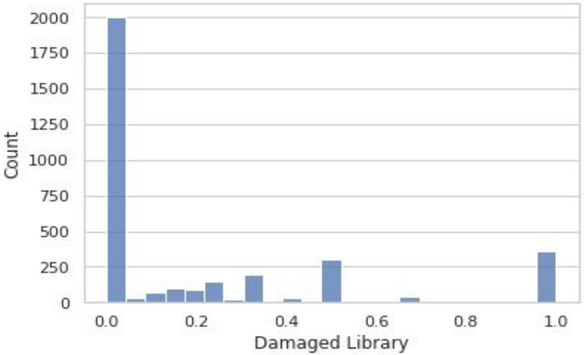
	Unnamed: 0	Kab.						
0	1	Kepulauan S						
1	2	Cir						
2	3	Indramayu	Jambi	1	1	1	1	1
3	4	Sukabumi	Jawa Barat	1	1	1	1	1
4	5	Sukabumi	Jawa Barat	1	1	1	1	1
...	...	...	...	...	...	...	...	...
82	83	Pandeglang	Banten	1	1	1	1	1
83	84	Belitung Timur	Bangka Belitung	1	1	1	1	1
84	85	Pohuwato	DKI Jakarta	1	1	1	1	1
85	86	Karimun	Kalimantan Tengah	1	1	1	1	1
86	87	Malinau	Kalimantan Tengah	1	1	1	1	1

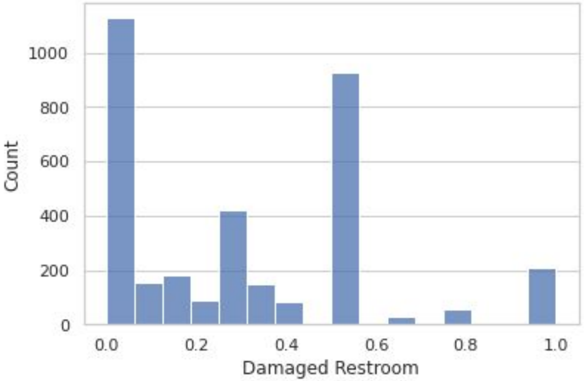
87 rows × 8 columns

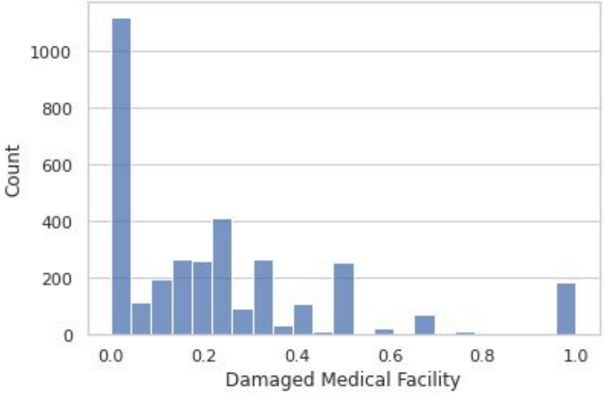
Distribution







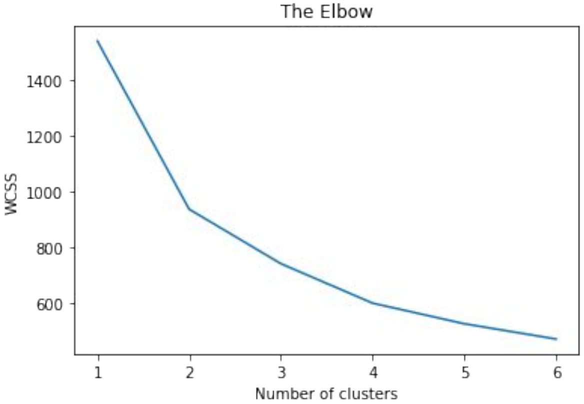


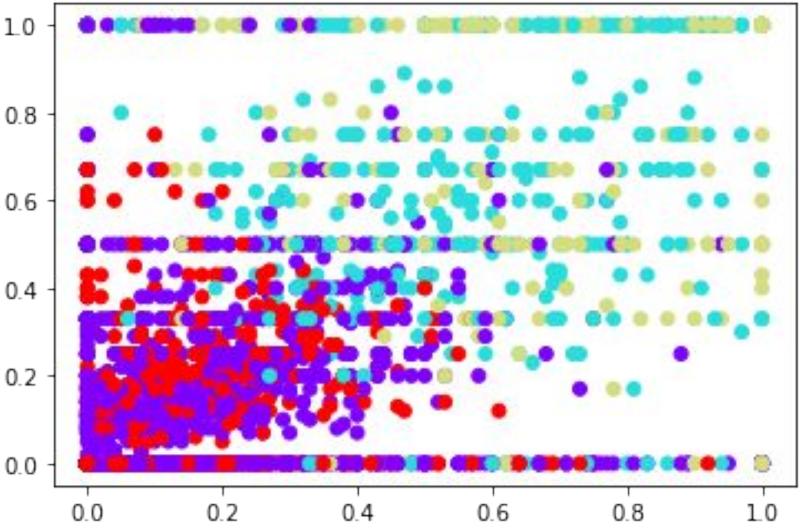


Restroom condition is urgently need to be investigated based on current bar.

Kmeans

In order to find the clusters, Kmeans clustering is applied to the data. The segmentation clusters decided to be 4 clusters.





Clustering Character

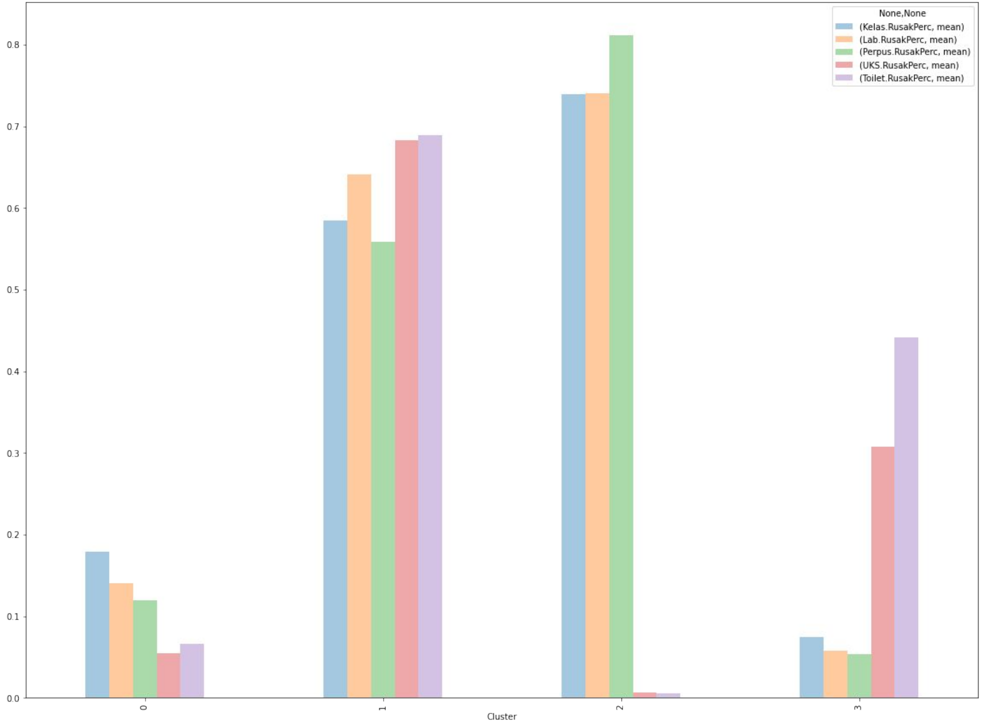
After clustering, we can now see that the data can be separated to several segments. The summary as follow:

Cluster 0: This cluster has low damage rate on all facilities. In average below 20%

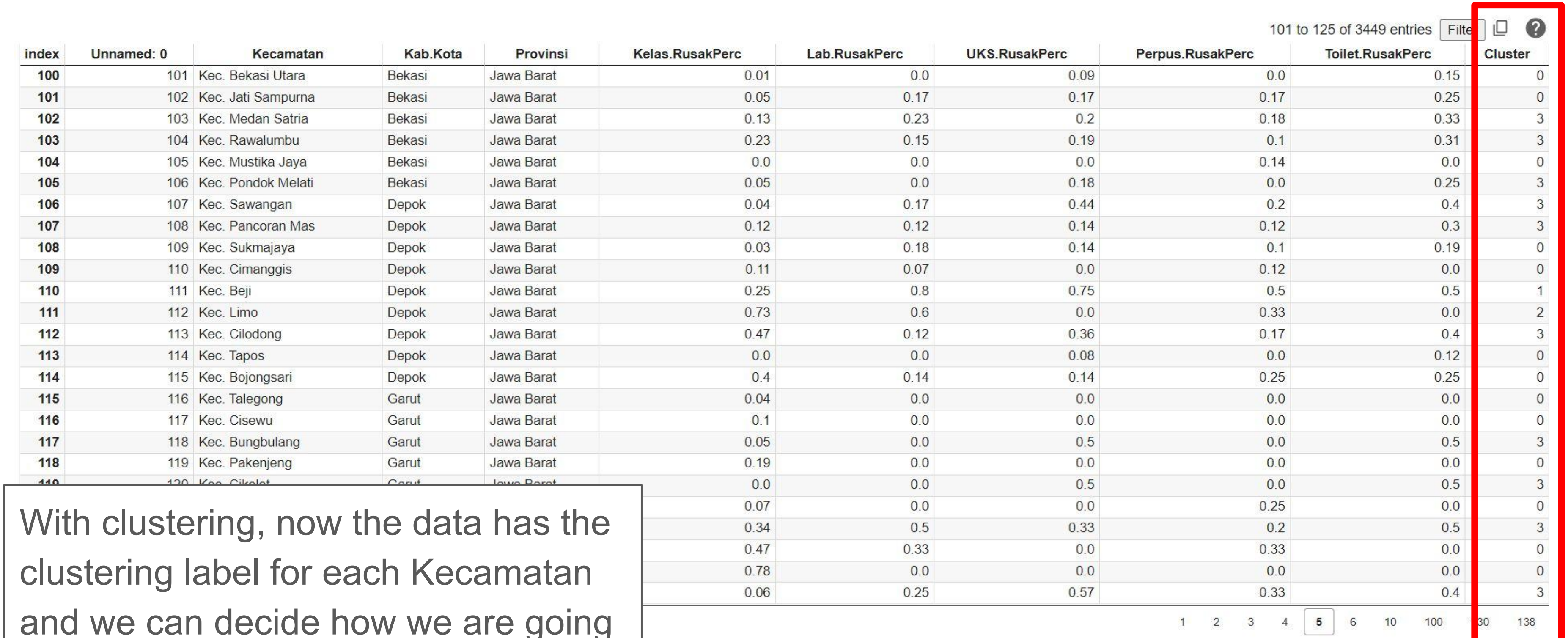
Cluster 1: This cluster has high damage rate on all facilities. All of the in average have above 50% of damage

Cluster 2: This cluster has high damage rate on Kelas, Perpus, Lab. All of them in average have above 70% of damage

Cluster 3: This cluster has high damage rate on Toilet and UKS



Labeling the Cluster



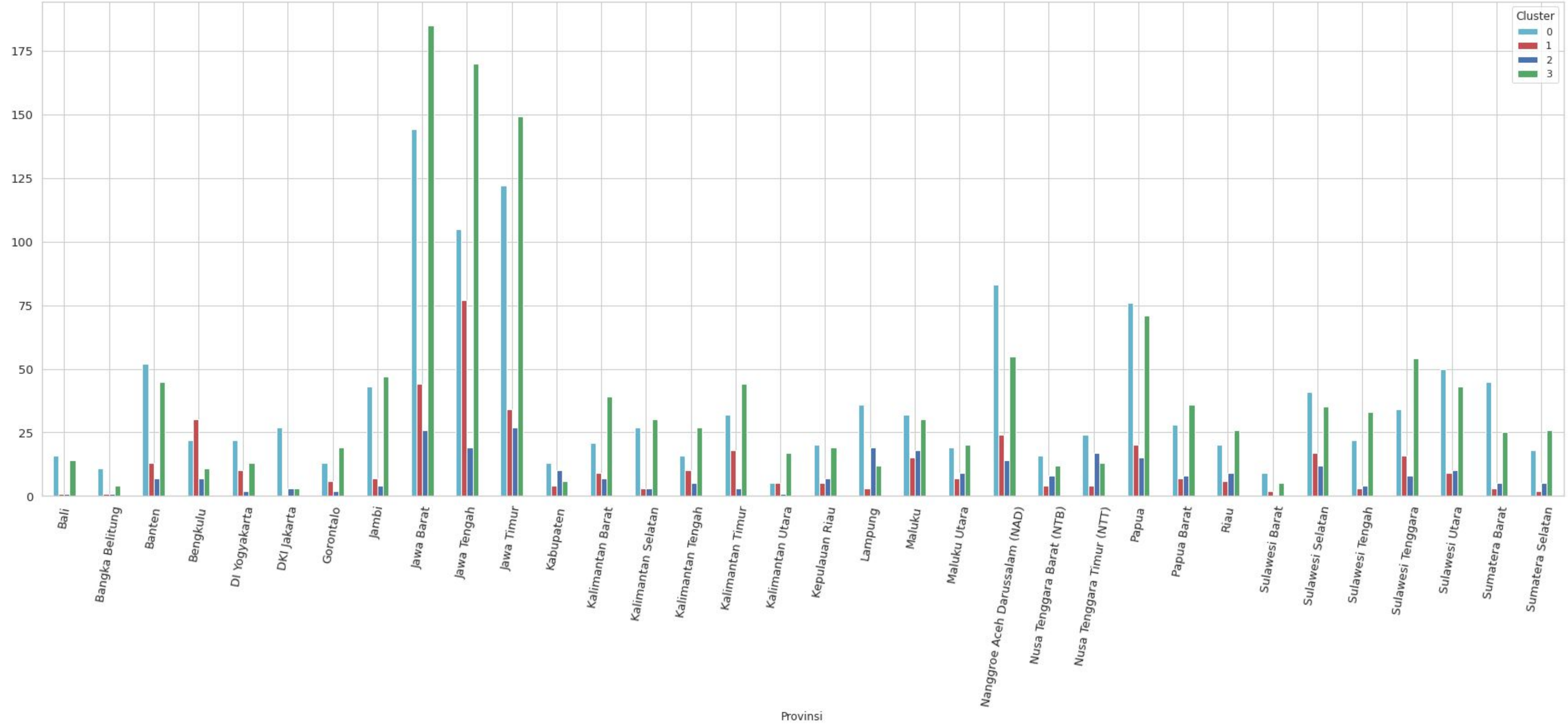
101 to 125 of 3449 entries Filte

index	Unnamed: 0	Kecamatan	Kab.Kota	Provinsi	Kelas.RusakPerc	Lab.RusakPerc	UKS.RusakPerc	Perpus.RusakPerc	Toilet.RusakPerc	Cluster
100	101	Kec. Bekasi Utara	Bekasi	Jawa Barat	0.01	0.0	0.09	0.0	0.15	0
101	102	Kec. Jati Sampurna	Bekasi	Jawa Barat	0.05	0.17	0.17	0.17	0.25	0
102	103	Kec. Medan Satria	Bekasi	Jawa Barat	0.13	0.23	0.2	0.18	0.33	3
103	104	Kec. Rawalumbu	Bekasi	Jawa Barat	0.23	0.15	0.19	0.1	0.31	3
104	105	Kec. Mustika Jaya	Bekasi	Jawa Barat	0.0	0.0	0.0	0.14	0.0	0
105	106	Kec. Pondok Melati	Bekasi	Jawa Barat	0.05	0.0	0.18	0.0	0.25	3
106	107	Kec. Sawangan	Depok	Jawa Barat	0.04	0.17	0.44	0.2	0.4	3
107	108	Kec. Pancoran Mas	Depok	Jawa Barat	0.12	0.12	0.14	0.12	0.3	3
108	109	Kec. Sukmajaya	Depok	Jawa Barat	0.03	0.18	0.14	0.1	0.19	0
109	110	Kec. Cimanggis	Depok	Jawa Barat	0.11	0.07	0.0	0.12	0.0	0
110	111	Kec. Beji	Depok	Jawa Barat	0.25	0.8	0.75	0.5	0.5	1
111	112	Kec. Limo	Depok	Jawa Barat	0.73	0.6	0.0	0.33	0.0	2
112	113	Kec. Cilodong	Depok	Jawa Barat	0.47	0.12	0.36	0.17	0.4	3
113	114	Kec. Tapos	Depok	Jawa Barat	0.0	0.0	0.08	0.0	0.12	0
114	115	Kec. Bojongsari	Depok	Jawa Barat	0.4	0.14	0.14	0.25	0.25	0
115	116	Kec. Talegong	Garut	Jawa Barat	0.04	0.0	0.0	0.0	0.0	0
116	117	Kec. Cisewu	Garut	Jawa Barat	0.1	0.0	0.0	0.0	0.0	0
117	118	Kec. Bungbulang	Garut	Jawa Barat	0.05	0.0	0.5	0.0	0.5	3
118	119	Kec. Pakenjeng	Garut	Jawa Barat	0.19	0.0	0.0	0.0	0.0	0
119	120	Kec. Cikelet	Garut	Jawa Barat	0.0	0.0	0.5	0.0	0.5	3
					0.07	0.0	0.0	0.25	0.0	0
					0.34	0.5	0.33	0.2	0.5	3
					0.47	0.33	0.0	0.33	0.0	0
					0.78	0.0	0.0	0.0	0.0	0
					0.06	0.25	0.57	0.33	0.4	3

1 2 3 4 5 6 10 100 30 138

With clustering, now the data has the clustering label for each Kecamatan and we can decide how we are going to take action for each segments.

Cluster Distribution per Province



Summary

- Based on clustering result, segment 0 labeled to **1306 kecamatan**, segment 1 to **435 kecamatan**, segment 2 to **315 kecamatan**, and segment 3 to **1393 kecamatan**
- Jawa Tengah is the province with most of segment 1 with **77 kecamatan**
- Jawa Timur has the most of segment 2 with **27 kecamatan**
- Jawa Barat has the most of segment 3 with **185 kecamatan**
- Jakarta doesn’t have any kecamatan in segment 1, and very low on segment 2 and 3
- Outside Java, Papua is quite concerning related to Medical Facility and Restroom, with **71 kecamatan** are part of segment 3
- Aceh also have problem with Medical Facility and Restroom
- Outside Java, Lampung and NTT are province with high number of segment 2, with **19 and 17 kecamatan** respectfully

Key Points & Action

- With data science, we can organize all data across all table consist of different facilities condition
- A more organized data can make data reading easier and more efficient
- Organized data can make a deeper data exploration so we get a lot of insight from there
- Applying clustering method can help segment each area so we can have a different character of each cluster
- Hopefully with clustering, when taken into action (e.g starting to fix the damaged facilities) decision maker can make a big project with specialization of the 3rd party fixing the facilities more close to cluster character
- After High School facilities data can be identified, it’s very possible with same methodology, Junior High School, Elementary School, and Vocational School data can also be processed

Thank You