Great — you already have a strong dataset portfolio for MRT station and line planning. Here’s a structured plan on what to **do next** using your data:

**🔍 Step 1: Identify Underserved Areas**

Use:

* ✅ **Subzone → Nearest MRT Station Distance**
* ✅ **Population Density by Subzone**
* ✅ **MRT User Ratio**

**Goal:** Find high-density subzones that are far (>800m or >1km) from any MRT station and have high MRT usage (suggesting suppressed demand).

📌 Output: A list/map of high-potential subzones underserved by current MRT.

**🧮 Step 2: Score & Rank Subzones**

Create a **composite index or score** for each subzone:

| **Metric** | **Rationale** | **Weight** |
| --- | --- | --- |
| Population Density | Captures demand | 0.25 |
| Employment Density | Captures work-trip volume | 0.20 |
| Distance to MRT | Underserved status | 0.25 |
| MRT User Ratio | Existing reliance | 0.15 |
| Morning Tap-out/Tap-in Skew | Station pressure indicator | 0.10 |
| Income | Can invert to prioritize equity | 0.05 |

Normalize each metric and calculate:

# Example score formula

score = 0.25 \* pop\_density\_norm + 0.20 \* emp\_density\_norm + ...

📌 Output: A ranked list of subzones by suitability for new station(s)

**🗺️ Step 3: Map High-Scoring Zones**

Use **Plotly or Folium** to visualize:

* High scoring subzones (heatmap or top 10 as markers)
* Existing MRT lines
* Nearest MRT stations (to show service gaps)

📌 Output: Interactive map highlighting strategic candidate locations

**🚉 Step 4: Simulate Station Placement**

Select 5–10 top-ranked subzones as **proposed station nodes**.

Check:

* Distance to nearby existing lines
* Whether several nodes lie in a corridor (→ new line potential)
* Whether they serve different residential or employment hubs

📌 Output: Proposed new station candidates (point layer or lat/lon list)

**🔄 Step 5: Form MRT Line(s)**

Now, connect high-priority station candidates:

* Manually draw plausible MRT line paths
* OR, use straight-line proximity or routing to cluster candidates into lines
* Ensure each proposed line:
  + Has a logical route
  + Connects with existing network
  + Covers multiple key nodes (residential, work)

📌 Output: Line geometry (Polyline), estimated ridership corridor

**📊 Step 6: Justify and Evaluate**

For each proposed station:

* Area population served (within 800m buffer)
* Employment density served
* Feeder bus availability or gaps
* Equity (if it serves low-income estates)

Optional:

* Estimate time savings (e.g. average travel time to CBD drops X mins)
* Estimate decongestion effect (if near busy stations)

📌 Output: Table of benefits per proposed station & line

**✅ Summary of What You Can Build**

You can now create:

1. **An interactive map** of MRT service gaps and proposed expansions
2. **A report** justifying each new station and proposed line
3. **Visuals**: heatmaps, station scoring, line overlays, tap-in/out imbalance
4. Optional: **Dashboard** in Dash or Streamlit to present this interactively

If you tell me your preferred tool (e.g. Python + Plotly, or QGIS), I can help you **generate visuals or scoring formulas** for the next step.

Option 1:

SUBZONE\_N MRT\_Access\_Gap

48 CLEMENTI WEST True

51 CLEMENTI WOODS True

107 FRANKEL True

124 WEST COAST True

174 FLORA DRIVE True

194 TAMAN JURONG True

199 BUKIT BATOK SOUTH True

218 PASIR RIS WEST True

228 TRAFALGAR True

234 SELETAR HILLS True

275 YISHUN EAST True

278 SEMBAWANG NORTH True

**MRT Accessibility Gap**

45 DOVER

107 FRANKEL

163 BALESTIER

174 FLORA DRIVE

189 TAI SENG

218 PASIR RIS WEST

228 TRAFALGAR

275 YISHUN EAST

278 SEMBAWANG NORTH

281 NORTHLAND

Great question — let’s break this down clearly.

Both **Option 1** and **MRT Accessibility Gap** are methods to identify where new MRT stations might be needed, but they differ in **scope**, **flexibility**, and **how you use MRT user ratio**.

**🔹 Option 1: Use MRT User Ratio as a Filter (High or Low)**

**🔧 What It Is:**

You use the **Ratio\_of\_MRT\_Users** metric *directly* to filter areas:

* **High usage → likely overloaded**
* **Low usage → possibly suppressed demand**

**🧠 Logic:**

You're assuming:

* High MRT user ratio = high demand → maybe **more access needed**
* Low MRT user ratio = possibly **suppressed demand** if also far from MRT

**🧪 How It Looks:**

# High usage areas

df\_high\_usage = df[df['Ratio\_of\_MRT\_Users'] > df['Ratio\_of\_MRT\_Users'].quantile(0.7)]

# Low usage + far → possible suppressed demand

df\_low\_usage = df[

(df['Ratio\_of\_MRT\_Users'] < df['Ratio\_of\_MRT\_Users'].quantile(0.3)) &

(df['dist\_to\_nearest\_mrt\_km'] > 0.8)

]

**✔️ Pros:**

* Simple and fast
* Easy to explain

**❌ Cons:**

* Doesn’t consider **population density** or other factors
* Might give false positives in areas with genuinely low demand

**🟢 MRT Accessibility Gap**

**🔧 What It Is:**

A **composite flag** that combines:

1. **Distance to MRT**
2. **Population density**
3. **MRT user ratio** (either high or low)

You define this as an area being underserved *either because* MRT demand is **suppressed** (low usage + far) or **overloaded** (high usage + far), **and** enough people live/work there to justify it.

**🧪 How It Looks:**

df['MRT\_Access\_Gap'] = (

((df['Ratio\_of\_MRT\_Users'] < 30th percentile) |

(df['Ratio\_of\_MRT\_Users'] > 70th percentile)) &

(df['dist\_to\_nearest\_mrt\_km'] > 0.8) &

(df['population\_density'] > median)

)

**✔️ Pros:**

* More complete and accurate
* Balances **access**, **demand**, and **density**
* Avoids false positives (e.g. low MRT usage in low-density areas)

**❌ Cons:**

* Slightly more complex
* Harder to explain in one sentence

**✅ Summary Table**

| **Feature** | **Option 1 (Filter by MRT Ratio)** | **MRT Accessibility Gap** |
| --- | --- | --- |
| Simplicity | ✅ Very simple | ❌ Needs 3 conditions |
| Customizability | ❌ Rigid (just usage) | ✅ Flexible (density + distance + usage) |
| Captures suppressed demand | ✅ Yes | ✅ Yes |
| Captures overloaded demand | ❌ Not directly | ✅ Yes |
| Best for? | Quick analysis, single factor | Detailed planning, better targeting |

**🧭 Final Verdict**

* Use **Option 1** for quick checks or basic filtering.
* Use **MRT\_Access\_Gap** when you're ready for **serious analysis and planning** — it's the more powerful and realistic metric.

Would you like to now visualize both on the same map or do a side-by-side comparison?

Absolutely! Let me walk you through the **entire clustering process** in the context of your Singapore MRT planning project — using subzone-level data like population density, MRT access, and employment.

**🎯 Goal**

You want to **group subzones** with similar characteristics to identify:

* Underserved areas (e.g., far from MRT, high density, low MRT use)
* Priority candidates for future MRT station planning

**🧩 What is Clustering?**

Clustering is **unsupervised machine learning**.  
We group data points (subzones) based on **similar patterns** in selected features.

We’re using **K-Means Clustering**:

* It groups subzones into *k clusters*.
* Each subzone goes to the cluster with the **closest centroid** (average of all points in that cluster).

**🧮 Features You Use for Clustering**

These come from your filtered GeoDataFrame:

| **Feature** | **Description** | **Why it's important** |
| --- | --- | --- |
| distance\_km | Distance to nearest MRT station | Subzones farther from MRT might be underserved |
| population\_density | Residents per km² | Denser areas = more demand |
| employment\_density | Workers per km² | Places with many jobs need MRT access |
| mrt\_user\_ratio | Proportion of MRT users | Helps find areas where MRT isn't used much |

✅ You *inverted* mrt\_user\_ratio so higher means **worse access**.

gdf\_cluster['mrt\_user\_inverse'] = 1 - gdf\_cluster['mrt\_user\_ratio']

**🔢 Why StandardScaler?**

Clustering needs features on the **same scale**:

* distance\_km might range from 0 to 10+
* population\_density might go to thousands
* mrt\_user\_inverse is between 0 and 1

So, we use **StandardScaler** to normalize them:

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

scaled\_features = scaler.fit\_transform(gdf\_cluster[['distance\_km', 'population\_density', 'employment\_density', 'mrt\_user\_inverse']])

Now they all have:

* Mean = 0
* Std Dev = 1

**🔗 KMeans Clustering**

You pick how many groups (n\_clusters=4, for example):

from sklearn.cluster import KMeans

kmeans = KMeans(n\_clusters=4, random\_state=42)

gdf\_cluster['cluster'] = kmeans.fit\_predict(scaled\_features)

Each subzone gets a **cluster label** (0 to 3) based on the grouping.

**🗺️ Visualizing the Clusters**

If you're working with a GeoDataFrame, you can **plot the results on a map**:

import matplotlib.pyplot as plt

fig, ax = plt.subplots(1, 1, figsize=(10, 8))

gdf\_cluster.plot(column='cluster', categorical=True, legend=True, cmap='Set2', ax=ax)

ax.set\_title("Subzone Clusters Based on MRT Access & Demographics")

plt.axis('off')

plt.show()

This helps you **see spatial patterns**:

* Are there clusters of underserved areas in certain parts of Singapore?
* Which clusters have high development potential?

**🏁 Summary of Flow**

1. **Prepare** the key features (drop income if not useful)
2. **Scale** the features to normalize them
3. **Cluster** using KMeans
4. **Assign** cluster labels to your dataset
5. **Plot** the clusters on a map (optional)
6. **Interpret** the clusters (which group represents underserved?)

If you'd like:

* A *ranking score* instead of clusters → we can create a formula.
* To export clusters as GeoJSON or visualize with Plotly → I can help with that.

Would you like to know how to interpret the cluster results or choose the best number of clusters (n\_clusters)?

A map of the united states

AI-generated content may be incorrect.

A diagram of a graph

AI-generated content may be incorrect.

Great! The 3D scatter plot you've provided visualizes **subzones in Singapore**, clustered into **4 groups** using K-Means, based on:

* **X-axis:** Distance to Nearest MRT (in km)
* **Y-axis:** Population Density (residents per km²)
* **Z-axis:** Working Population Density (workers per km²)
* **Colors:** K-Means Cluster Labels
  + Red, Blue, Green, Yellow → 4 different cluster groupings

Thanks — that’s an important correction. Based on your updated observation, let’s **reframe your cluster interpretations** for clarity and accuracy.

**✅ Revised Cluster Interpretations**

**🟨 Yellow — Central Business District (CBD)**

* **Very High Working Population Density**
* **Very Low Residential Population**
* **MRT Distance: Low to Moderate**
* 📍 **Likely Areas**: Raffles Place, Downtown Core, Marina Bay, Shenton Way
* **Insight**: These are the **economic hubs** of Singapore. Very intense daytime influx; likely already well served by MRT. Focus could shift to **last-mile and crowd handling**, not new lines.

**🟥 Red — Regional & Mixed-Use Centres**

* **High Working & Residential Density**
* **MRT Distance: Low to Moderate**
* 📍 **Likely Areas**: Woodlands, Tampines, Jurong East, Kallang, Novena, Buona Vista, Newton
* **Insight**: These areas are planning area centres or **decentralised nodes**. They act as smaller CBDs. MRT access is decent, but may benefit from **station upgrades** or **cross-line connectors**.

**🟦 Blue — Mature HDB Heartlands**

* **Very High Residential Density**
* **Low Distance to MRT**
* **Moderate Working Density**
* 📍 **Likely Areas**: Bishan, Toa Payoh, Ang Mo Kio, Bedok, Queenstown
* **Insight**: These are **well-connected residential hubs**. Likely optimal MRT coverage. Not priorities for expansion — may instead benefit from service frequency or interchange upgrades.

**🟩 Green — Emerging / Underserved Zones**

* **Moderate Residential + Working Density**
* **Longer Distance to MRT**
* 📍 **Likely Areas**: Seletar, Sengkang West, Tengah, parts of Punggol, outer Woodlands, etc.
* **Insight**: **Top candidates** for **new MRT lines or infill stations**. Especially if population is expected to grow.

**✅ Now What?**

With this improved interpretation, your **MRT planning narrative** becomes stronger:

**1. Use 🟩 Green + Distance > 1.5 km → Propose new MRT coverage**

* Look for **high population/employment density** among these

**2. Use 🟥 Red → Check for potential congestion or interchange bottlenecks**

* These areas may be **next-tier investment zones**

**3. 🟦 Blue & 🟨 Yellow → Already mature → Maintain, enhance crowd flow**

Would you like me to:

* Help produce a **summary table** of all cluster types + strategy?
* Filter and export a list of 🟩 underserved subzones?
* Suggest **new MRT lines** between selected zones?

Let me know what you're aiming to present next.

**✅ What This Plot Helps You Do**

* Visually **separate zones** by their **transit access vs. population pressure**.
* Identify areas underserved by MRT (e.g. green zones far from MRT with dense residents).
* Distinguish **commuting-focused subzones** (yellow) vs **residential hubs** (blue).
* Prioritize zones for **transit upgrades, housing plans, or business development**.

Would you like a 2D PCA version of this or a map to match each cluster to planning areas?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Colour | Type of Zones | Description | Insight | Areas |
| 0 | Red | **Regional & Mixed-Use Centres** | * **High Residential Density** * **High Working Population Density** * **MRT Distance: Low to Moderate** | These areas are planning area centres or **decentralised nodes**. They act as smaller CBDs. MRT access is decent, but may benefit from **station upgrades** or **cross-line connectors**. | Woodlands, Tampines, Jurong East, Kallang, Novena, Buona Vista, Newton |
| 1 | Blue | **Mature HDB Heartlands** | * **Very High Residential Density** * **Moderate Working Density** * **Low Distance to MRT** | These are **well-connected residential hubs**. Likely optimal MRT coverage. Not priorities for expansion — may instead benefit from service frequency or interchange upgrades.  Established HDB heartlands with strong MRT access. MRT serves them well → efficient transport catchments | Bishan,  Toa Payoh, Ang Mo Kio, Bedok, Choa Chu Kang |
| 2 | Yellow | **Central Business District (CBD)** | * **Very High Working Population Density** * **Very Low Residential Population** * **MRT Distance: Low to Moderate** | These are the **economic hubs** of Singapore. Very intense daytime influx; likely already well served by MRT. Focus could shift to **last-mile and crowd handling**, not new lines. | Raffles Place,  Downtown Core, Marina Bay, Shenton Way,  Orchard |
| 3 | Green | **Emerging / Underserved Zones** | * **Moderate population density (residential + working)** * **Farther distance to MRT** | **Newer or fringe residential areas** still developing, where MRT hasn't reached fully. These could be **targets for MRT expansion**. | Seletar, Sengkang West, Tengah, parts of Punggol, outer Woodlands |

Great question — now that you've successfully clustered Singapore’s subzones using features like distance to MRT, population density, and working population density, the next steps depend on your **project goal**. Based on your previous MRT planning intent, here’s a structured guide on **what to do next**:

**🎯 Main Goal: Propose where to improve or expand MRT coverage.**

**✅ Step 1: Interpret and Label Each Cluster**

You've already done a strong visual analysis. Now:

* Assign **semantic meanings** to clusters (e.g. "well-served residential", "transit deserts", "commuter-only zones").
* For example:
  + **Blue** = Well-connected HDB areas
  + **Green** = Underserved, possibly MRT-sparse regions
  + **Red** = CBD/mixed-use zones
  + **Yellow** = High-working-pop industrial/business zones

**📍 Step 2: Map Clusters Spatially**

You already have a GeoDataFrame KMeansDF1 (or gdf\_cluster). Now:

import plotly.express as px

fig = px.choropleth\_mapbox(

gdf\_cluster,

geojson=gdf\_cluster.geometry,

locations=gdf\_cluster.index,

color='cluster',

mapbox\_style="carto-positron",

center={"lat": 1.35, "lon": 103.82},

zoom=10,

opacity=0.6,

color\_discrete\_map={

'0': 'red',

'1': 'blue',

'2': 'yellow',

'3': 'green'

}

)

fig.show()

🧭 This will let you **see which geographic areas** belong to each cluster — key for identifying where interventions are needed.

**📊 Step 3: Generate Cluster-Level Summary Stats**

Get insight into each cluster's profile:

gdf\_cluster.groupby('cluster')[['dist\_to\_nearest\_mrt\_km', 'population\_density', 'Working Population Density', 'Ratio\_of\_MRT\_Users']].mean()

This helps you **rank clusters** for MRT need — e.g. high population + long distance = MRT underserved.

**📌 Step 4: Identify MRT Underserved Areas**

Focus on clusters with:

* 🔴 High population density
* 🔴 High MRT user ratio
* 🔴 Long distance to MRT

These areas are your **high-priority zones for MRT expansion**.

You could:

* Flag subzones in the “green” cluster with highest dist\_to\_nearest\_mrt\_km
* Calculate working\_population + resident\_population to find total activity pressure

**🛤️ Step 5: Suggest New MRT Lines or Stations**

Now that you know where the needs are:

* **Visualize existing MRT lines** (you already have the shapefile)
* **Plot centroids** of target clusters (e.g. underserved green cluster)
* Connect them using:
  + KMeans cluster centroids
  + Network design (e.g. MST or shortest path)
  + Visual/heuristic planning

**🧩 Optional Next Steps:**

1. **Add more features**: land cost, bus access, future plans, housing types
2. **Multi-criteria scoring**: combine standardized distance/pop/work pop into a priority score
3. **Run simulation or optimization model**: e.g. where to build new MRT stops to minimize average distance

Would you like help visualizing suggested new MRT stops or computing the best subzones to prioritize next?

Thanks for sharing the map! Based on what you’ve described and shown:

* 🟦 **Blue** = Well-connected HDB areas
* 🟩 **Green** = Underserved, MRT-sparse regions
* 🟥 **Red** = CBD/mixed-use hubs (often well-served, but could still face peak congestion)

Here are **strategic suggestions** to better serve the **green underserved areas**, either through **new stations on existing lines** or by **proposing new MRT lines**:

**🚉 1. Add Infill Stations on Existing Lines**

Infill stations are added **between existing MRT stations**, ideal when lines pass near but don't serve a high-density area.

**🔹 Example A: Circle Line (Orange) or Downtown Line (Blue)**

* Near **Geylang area or East Coast**, you could add a station between existing Downtown Line or EWL stops (e.g., between Kembangan and Bedok or around Marine Parade).
* Could better serve underserved areas in the southeast green zones.

**🔹 Example B: North-East Line (Purple) – Upper Serangoon**

* Near **Hougang/Punggol fringe** or **Seletar**, where green zones exist.
* Add infill stations between Buangkok → Sengkang or NE17 → NE18.

**🆕 2. Propose Branch Lines / Spurs**

Branch lines extend existing MRT lines into green zones without building an entirely new line.

**🔸 Red Line (NSL) Extension – North Woodlands / Mandai**

* Add a spur heading west into the green zones near Mandai or northern Woodlands.
* These areas are **low MRT access but have rising residential/working demand**.

**🔸 East-West Line (EWL) – Between Chinese Garden and Lakeside**

* A spur can serve the **Jurong West Industrial / residential pockets** that aren’t close to current stations.

**🟢 3. New MRT Lines – Connect Peripheral Green Zones**

**🟢 East Coast – North-East Link**

* Draw a **new diagonal line** from the **East Coast green zone** (e.g., Siglap/Bedok South) → North-East zone (e.g., Seletar).
* Could connect to Circle Line (East) and North-East Line (Hougang/Punggol).

**🟢 Northern Arc Line**

* Serve green zones in **northeast → northwest arc** (e.g., Seletar, Lentor, Mandai).
* Connect from **Sengkang → Seletar → Mandai → Bukit Panjang**, potentially integrating with LRT lines.

**⚠️ 4. Prioritize Red Zone Management – CBD Congestion**

Red zones (e.g., Downtown Core) are **already well served**, but:

* Consider **express shuttles**, **short loop lines**, or **peak-hour crowd management**.
* Alternatively, **decentralize employment**: improve MRT access to green (fringe) zones to **encourage offices/business parks** outside of CBD.

**🔧 Tools to Help Further**

If you want to **simulate MRT expansion scenarios**, consider:

* **Isochrone maps** for travel time
* **AccessScore** by distance + ridership
* **Multi-criteria site selection**: Score each subzone by distance, population, jobs

Would you like me to help you build a simulation/dashboard for this?

A map of a city

AI-generated content may be incorrect.

**East Coast – North-East Link**

**EN1: Marine Terrace (TEL)**

**EN2: Frankel (midpoint)**

**EN3: Kembangan (EWL)**

**EN4: Kaki Bukit (DTL)**

**EN5: Paya Lebar Air Base (PLAB) (midpoint)**

**EN6: Defu (CRL)**

**EN7: Hougang (NEL, CRL)**

**EN8: Trafalgar (midpoint)**

**EN9: Seletar Hills (midpoint)**

**EN10: Sengkang West (midpoint)**

**EN11: Seletar Aerospace Park (midpoint)**

**EN12: Yishun East (midpoint)**

**EN13: Yishun (NEL)**

**A map of a city

AI-generated content may be incorrect.**

**DOVER**

**A map of a city

AI-generated content may be incorrect.**

**BALASTIER**

**A map of a city

AI-generated content may be incorrect.**

**SEMBAWANG NORTH**

**A map of a city

AI-generated content may be incorrect.**

**PASIR RIS WEST & FLORA DRIVE**

**A map of a city

AI-generated content may be incorrect.**

**TAI SENG**

**A map of a city

AI-generated content may be incorrect.**