Basics of k-means clustering

CLUSTER ANALYSIS IN PYTHON

Shaumik Daityari
Business Analyst

datacamp

Why k-means clustering?

- A critical drawback of hierarchical clustering: runtime
- K means runs significantly faster on large datasets

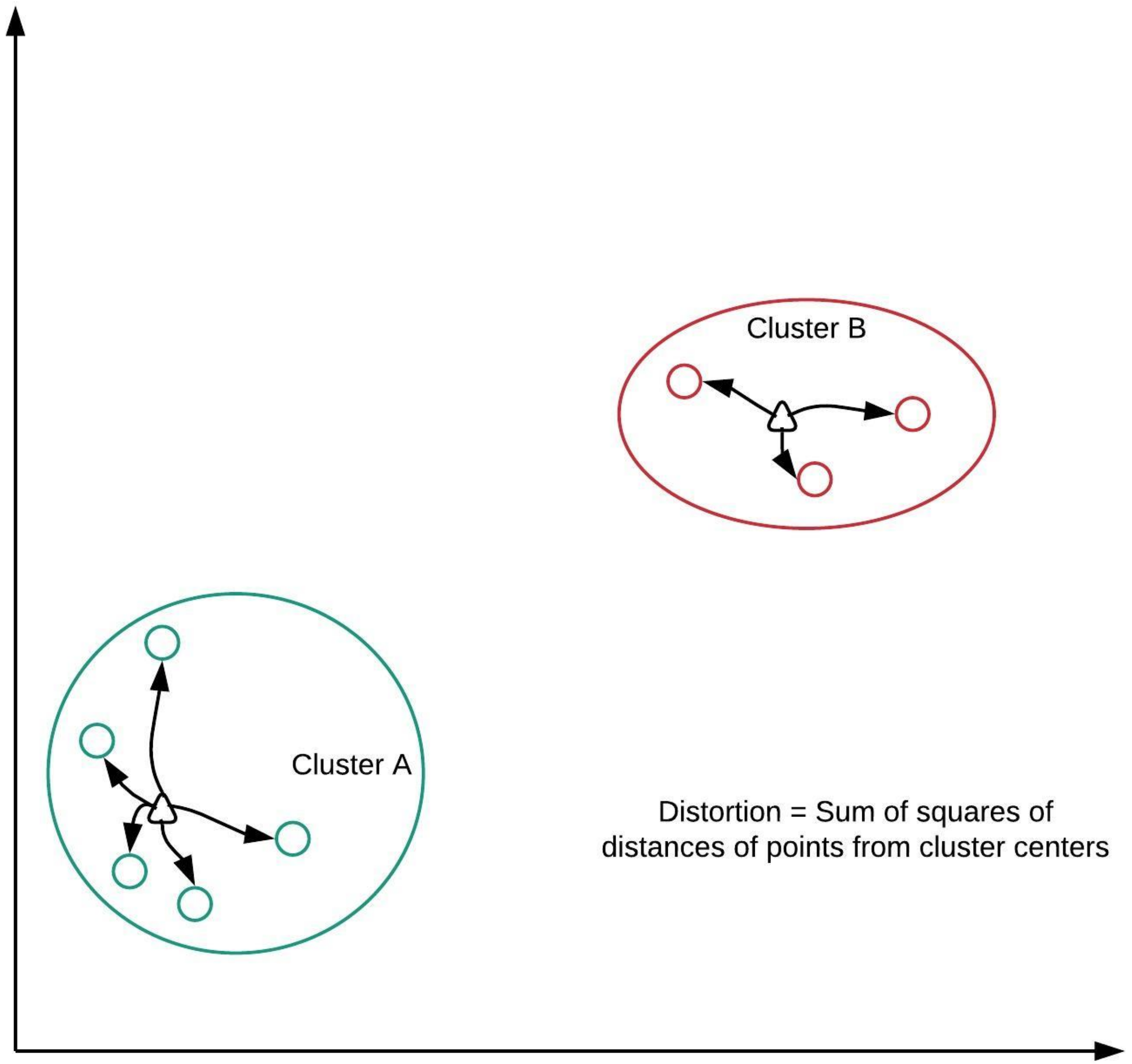
Step 1: Generate cluster centers

```
kmeans(obs, k_or_guess, iter, thresh, check_finite)
```

- `obs` : standardized observations
- `k_or_guess` : number of clusters
- `iter` : number of iterations (default: 20)
- `thres` : threshold (default: 1e-05)
- `check_finite` : whether to check if observations contain only finite numbers (default: True)

Returns two objects: cluster centers, distortion

How is distortion calculated?



Distortion = Sum of squares of distances of points from cluster centers

Step 2: Generate cluster labels

```
vq(obs, code_book, check_finite=True)
```

- `obs` : standardized observations
- `code_book` : cluster centers
- `check_finite` : whether to check if observations contain only finite numbers (default: True)

Returns two objects: a list of cluster labels, a list of distortions

A note on distortions

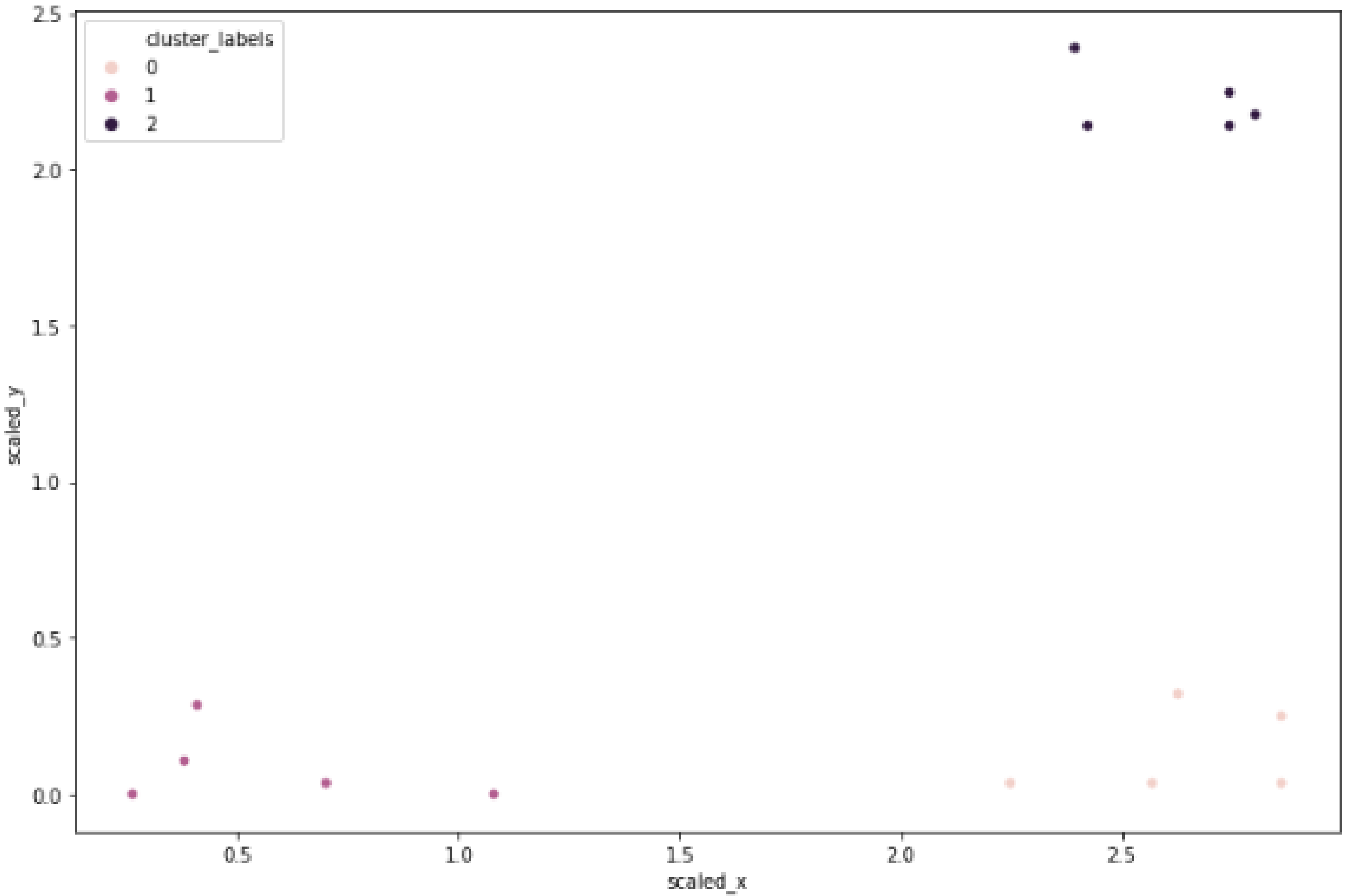
- `kmeans` returns a single value of distortions
- `vq` returns a list of distortions.

Running k-means

```
# Import kmeans and vq functions
from scipy.cluster.vq import kmeans, vq
```

```
# Generate cluster centers and labels
cluster_centers, _ = kmeans(df[['scaled_x', 'scaled_y']], 3)
df['cluster_labels'], _ = vq(df[['scaled_x', 'scaled_y']], cluster_centers)
```

```
# Plot clusters
sns.scatterplot(x='scaled_x', y='scaled_y', hue='cluster_labels', data=df)
plt.show()
```


cluster_labels
0
1
2
2.5
2.0
1.5
1.0
0.5
0.0
scaled_y
0.5
1.0
1.5
2.0
2.5
scaled_x

Next up: exercises!

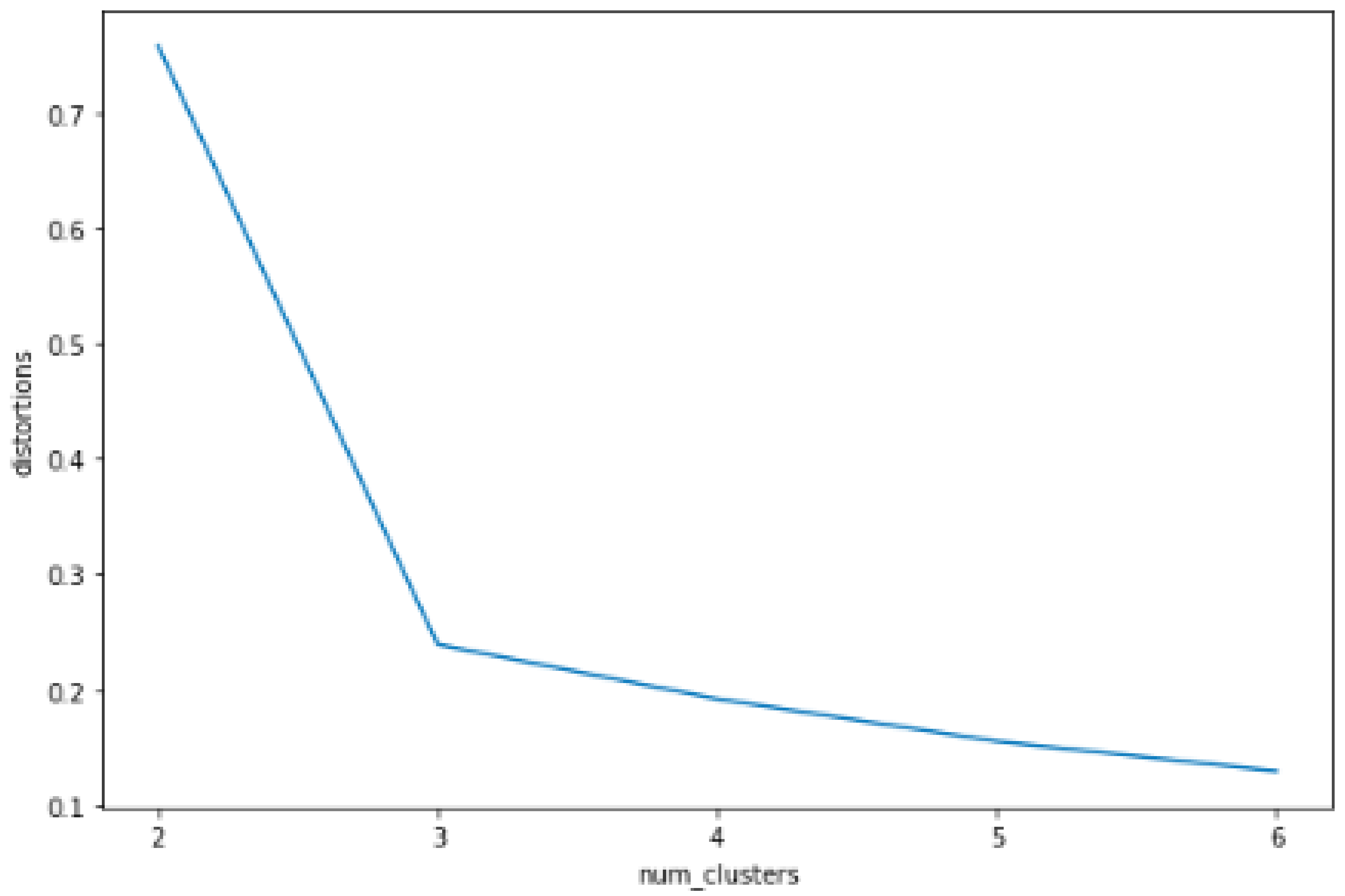
How many clusters?

CLUSTER ANALYSIS IN PYTHON

Shaumik Daityari
Business Analyst

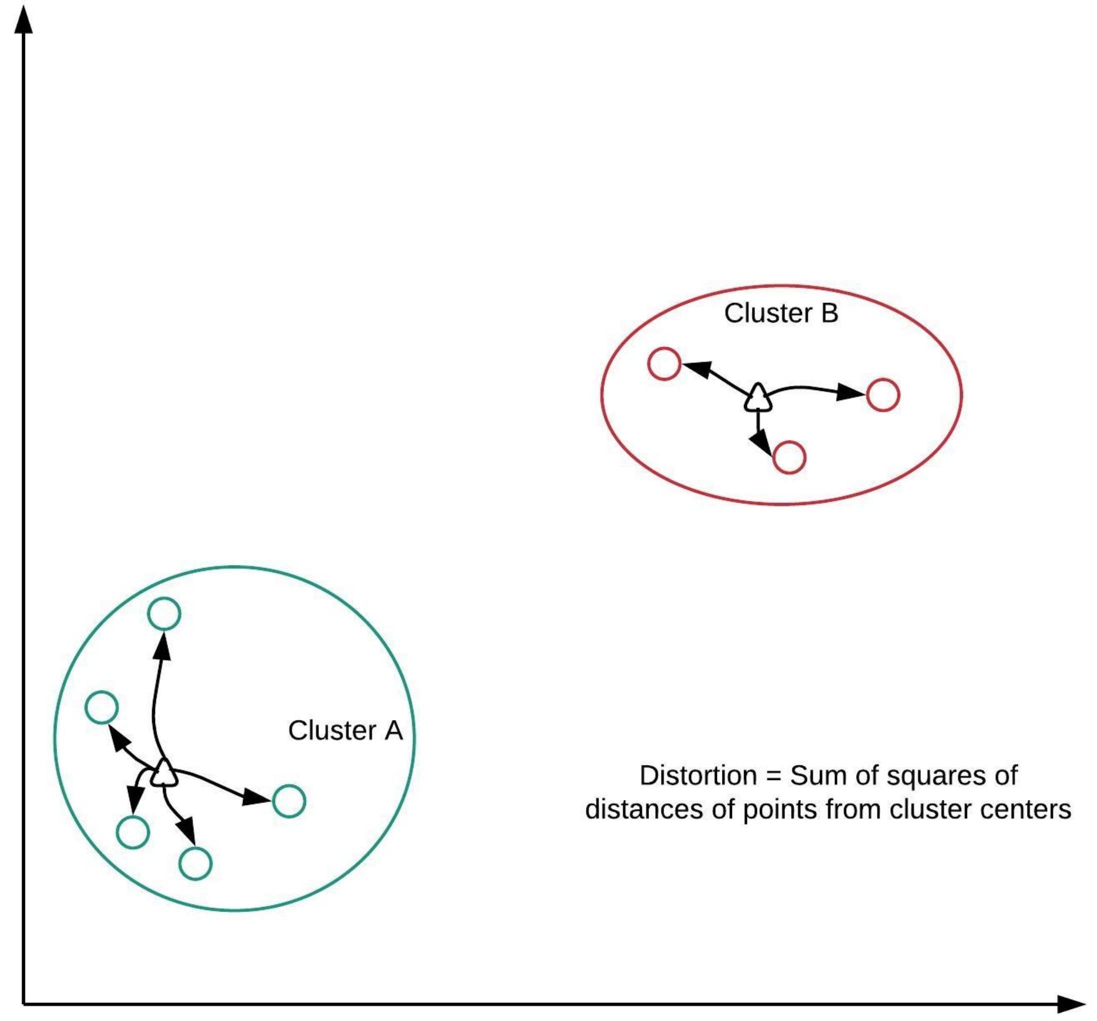
How to find the right k?

- No *absolute* method to find right number of clusters (k) in k-means clustering
- Elbow method



Distortions revisited

- Distortion: sum of squared distances of points from cluster centers
- Decreases with an increasing number of clusters
- Becomes zero when the number of clusters equals the number of points
- Elbow plot: line plot between cluster centers and distortion


Cluster B
Cluster A
Distortion = Sum of squares of distances of points from cluster centers

Elbow method

- Elbow plot: plot of the number of clusters and distortion
- Elbow plot helps indicate number of clusters present in data

Elbow method in Python

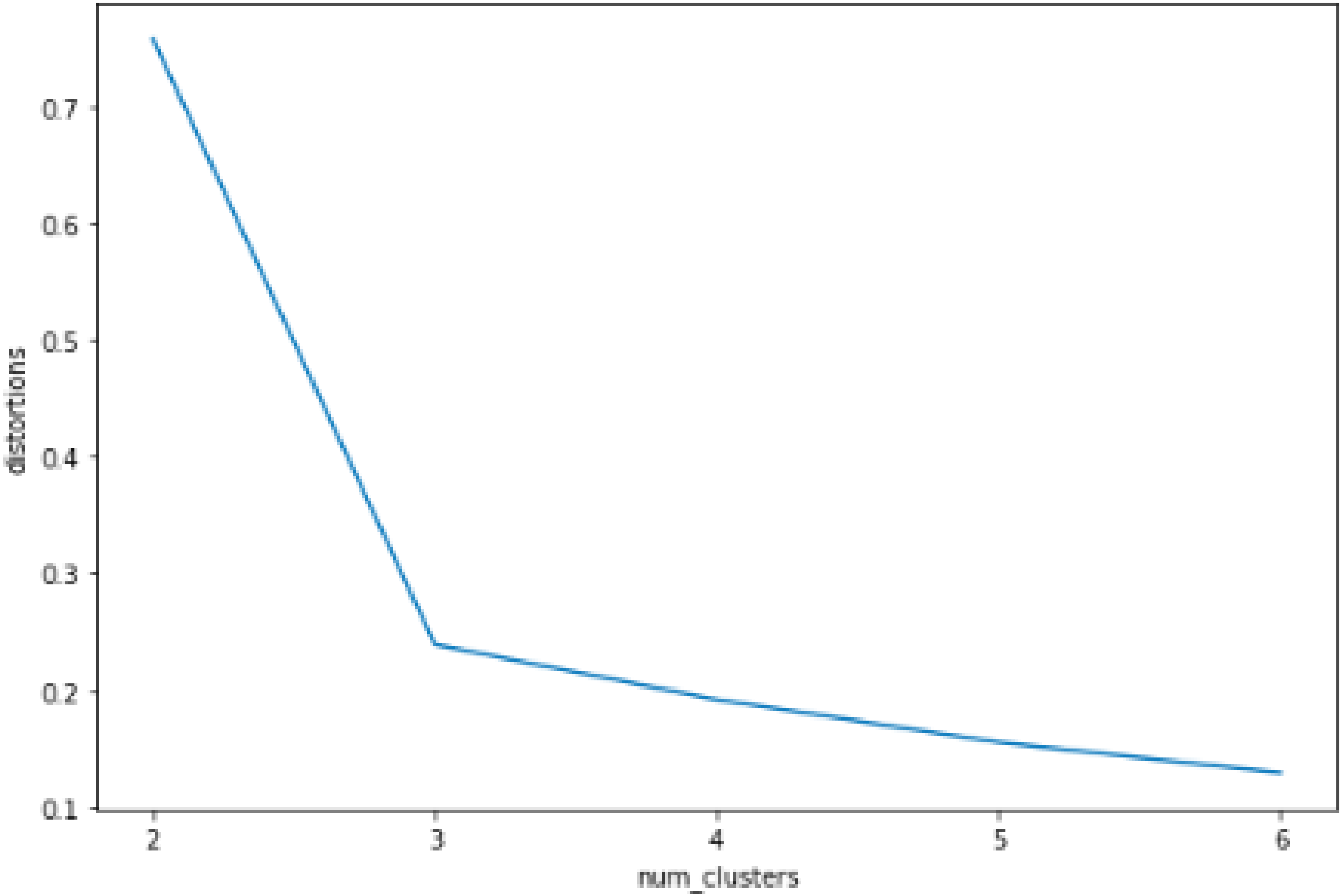
```
# Declaring variables for use
distortions = []

num_clusters = range(2, 7)
```

```
# Populating distortions for various clusters
for i in num_clusters:
    centroids, distortion = kmeans(df[['scaled_x', 'scaled_y']], i)
    distortions.append(distortion)
```

```
# Plotting elbow plot data
elbow_plot_data = pd.DataFrame({'num_clusters': num_clusters,
                                'distortions': distortions})

sns.lineplot(x='num_clusters', y='distortions',
             data = elbow_plot_data)
plt.show()
```


0.7
0.6
0.5
0.4
0.3
0.2
0.1
distortions
2
3
4
5
6
num_clusters

Final thoughts on using the elbow method

- Only gives an indication of optimal *k* (numbers of clusters)
- Does not always pinpoint how many *k* (numbers of clusters)
- Other methods: average silhouette and gap statistic

Next up: exercises

CLUSTER ANALYSIS IN PYTHON

Limitations of k-means clustering

CLUSTER ANALYSIS IN PYTHON

Shaumik Daityari
Business Analyst

datacamp

Limitations of k-means clustering

- How to find the right _K_ (number of clusters)?
- Impact of seeds
- Biased towards equal sized clusters

Impact of seeds

Initialize a random seed

```
from numpy import random
random.seed(12)
```

Seed: `np.array(1000, 2000)`

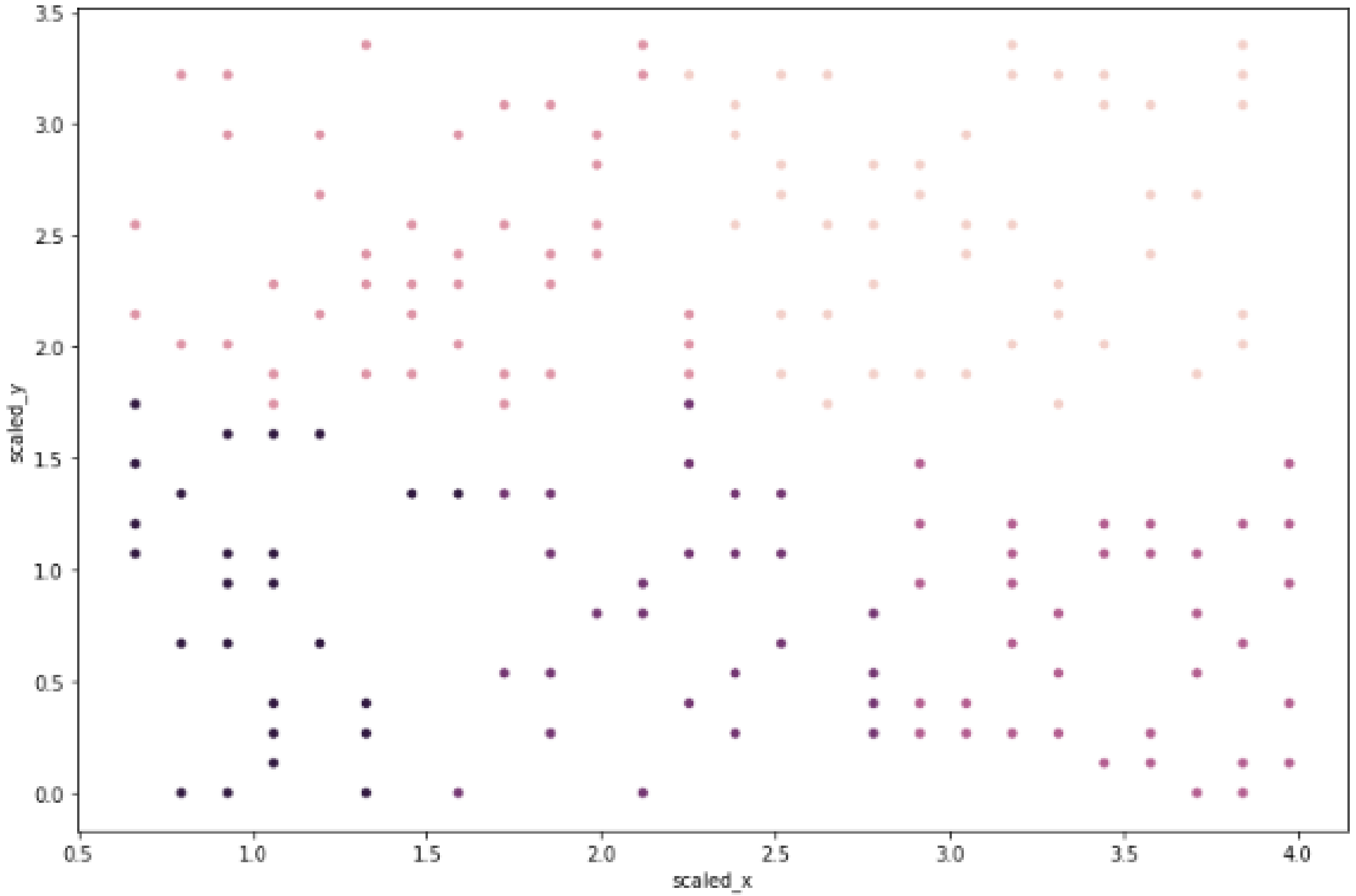
Cluster sizes: 29, 29, 43, 47, 52

Seed: `np.array(1,2,3)`

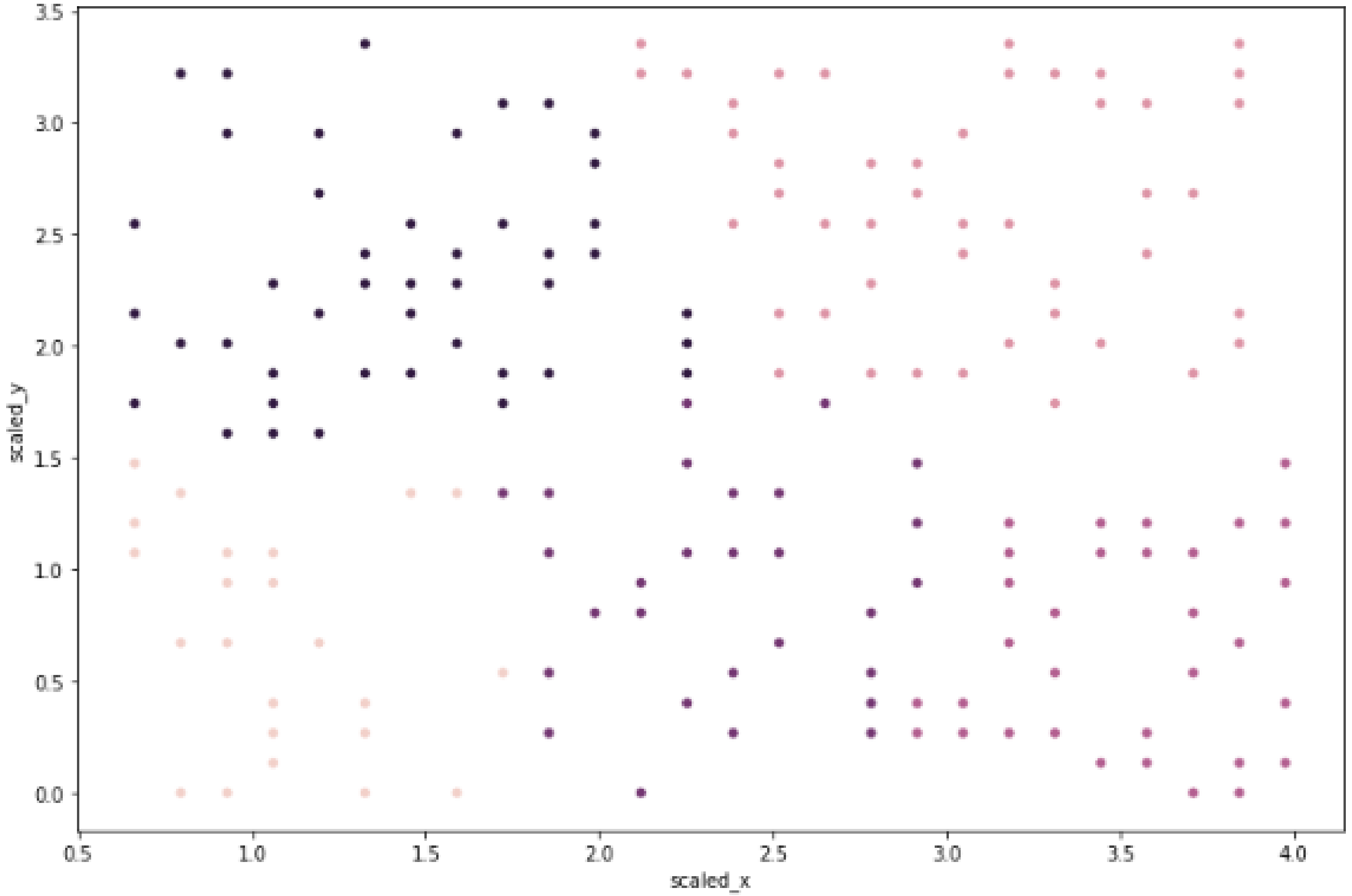
Cluster sizes: 26, 31, 40, 50, 53

Impact of seeds: plots

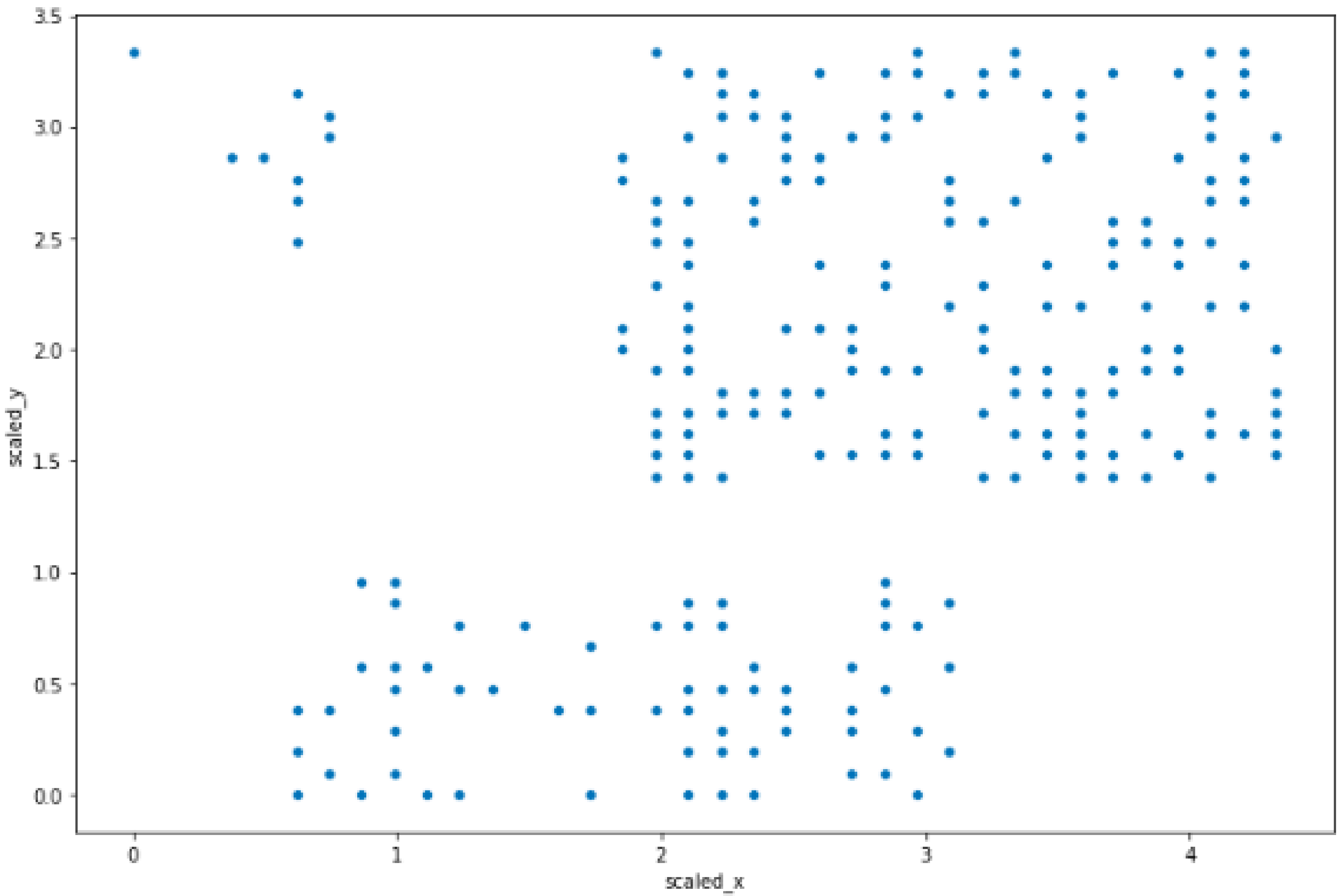
Seed: `np.array(1000, 2000)`



Seed: `np.array(1,2,3)`

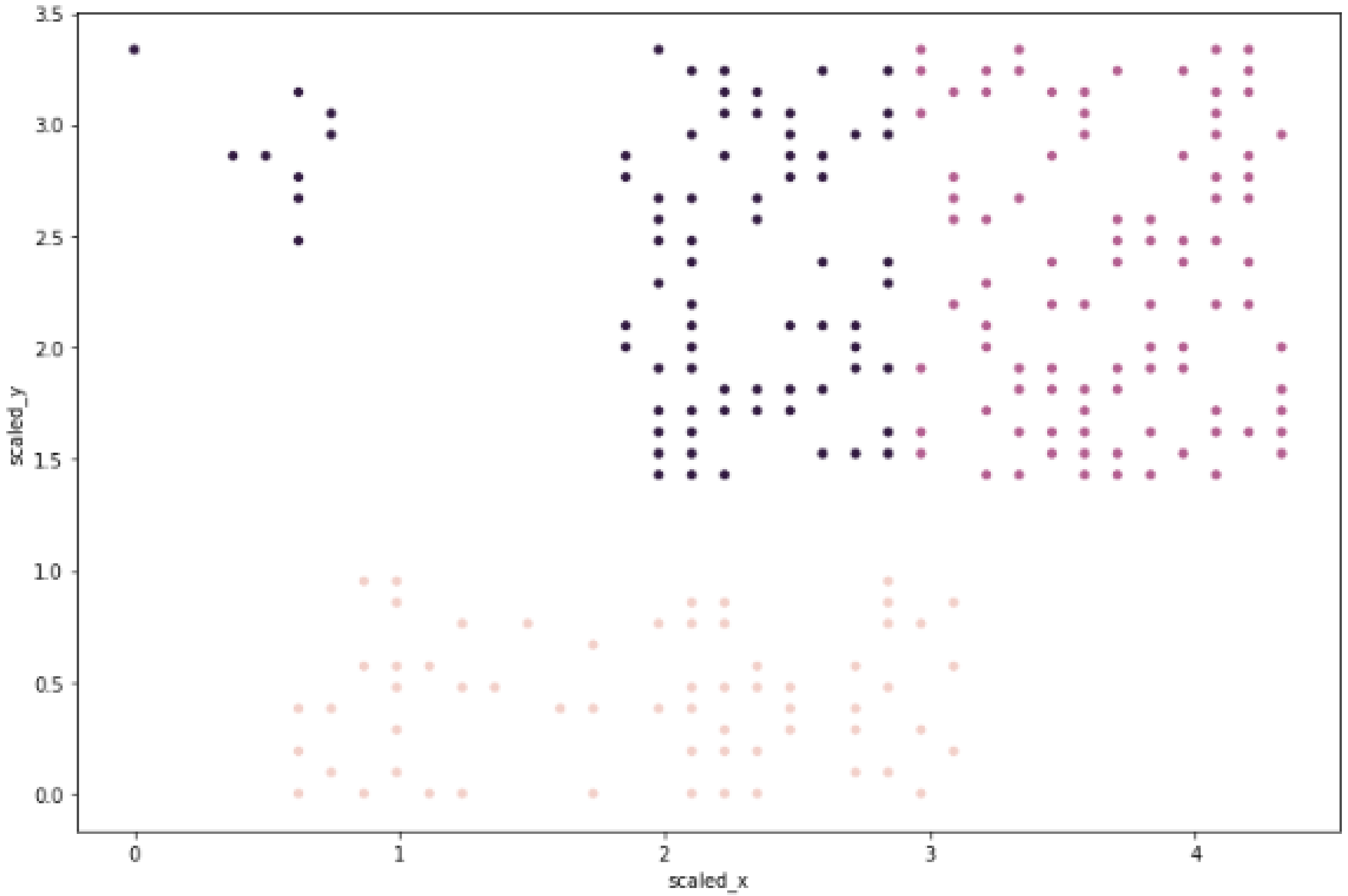


Uniform clusters in k means

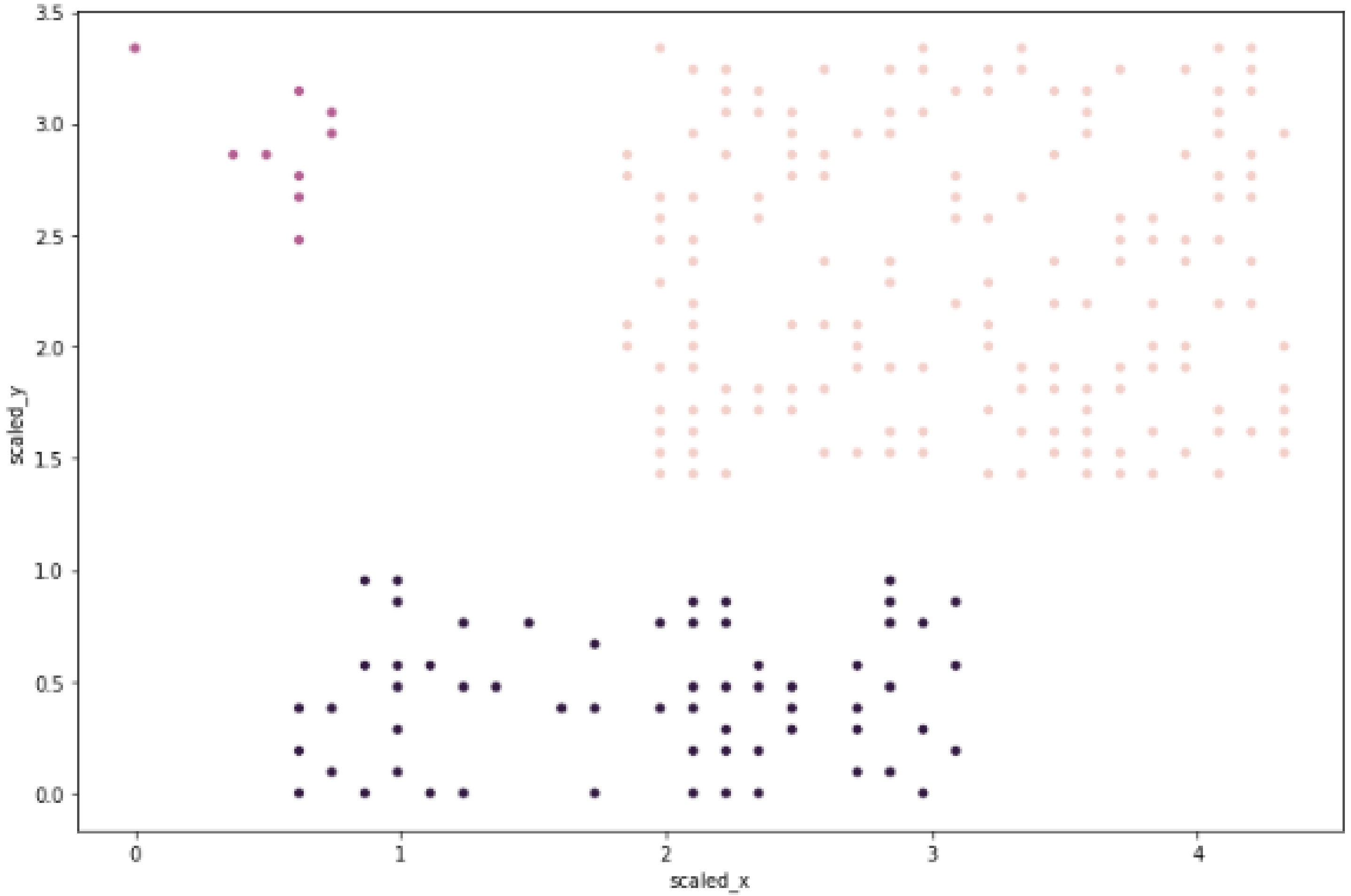


Uniform clusters in k-means: a comparison

K-means clustering with 3 clusters



Hierarchical clustering with 3 clusters



Final thoughts

- Each technique has its pros and cons
- Consider your data size and patterns before deciding on algorithm
- Clustering is exploratory phase of analysis

Next up: exercises

CLUSTER ANALYSIS IN PYTHON