DATA SCIENCE 102: CLUSTERING

AGENDA

- Unsupervised Learning
- Clustering
  - Use Cases
  - Types of Clustering Algorithms
- K-Means Clustering
  - K-Means Algorithm
  - Optimal K
  - Coded Example

UNSUPERVISED LEARNING

UNSUPERVISED LEARNING

- Unsupervised learning is used when there is **no outcome variable (*y*)** to predict or classify
- Attempts to learn patterns in the data other than predicting ***y***
- Unsupervised learning methods include:
  - Clustering Techniques
  - Association Rules
  - Dimension Reduction Methods

CLUSTERING TECHNIQUES

- Use Cases
- Types of Clustering Algorithm
- Distance Scoring

USAGE

- Segmentation of data into sets of homogenous clusters of records to generate insight
- Clustering can help improve the performance of supervised methods by modelling each cluster rather than the entire heterogeneous dataset
- Cluster analysis helps to form groups (clusters) of similar observations based on several measurements made on those observations

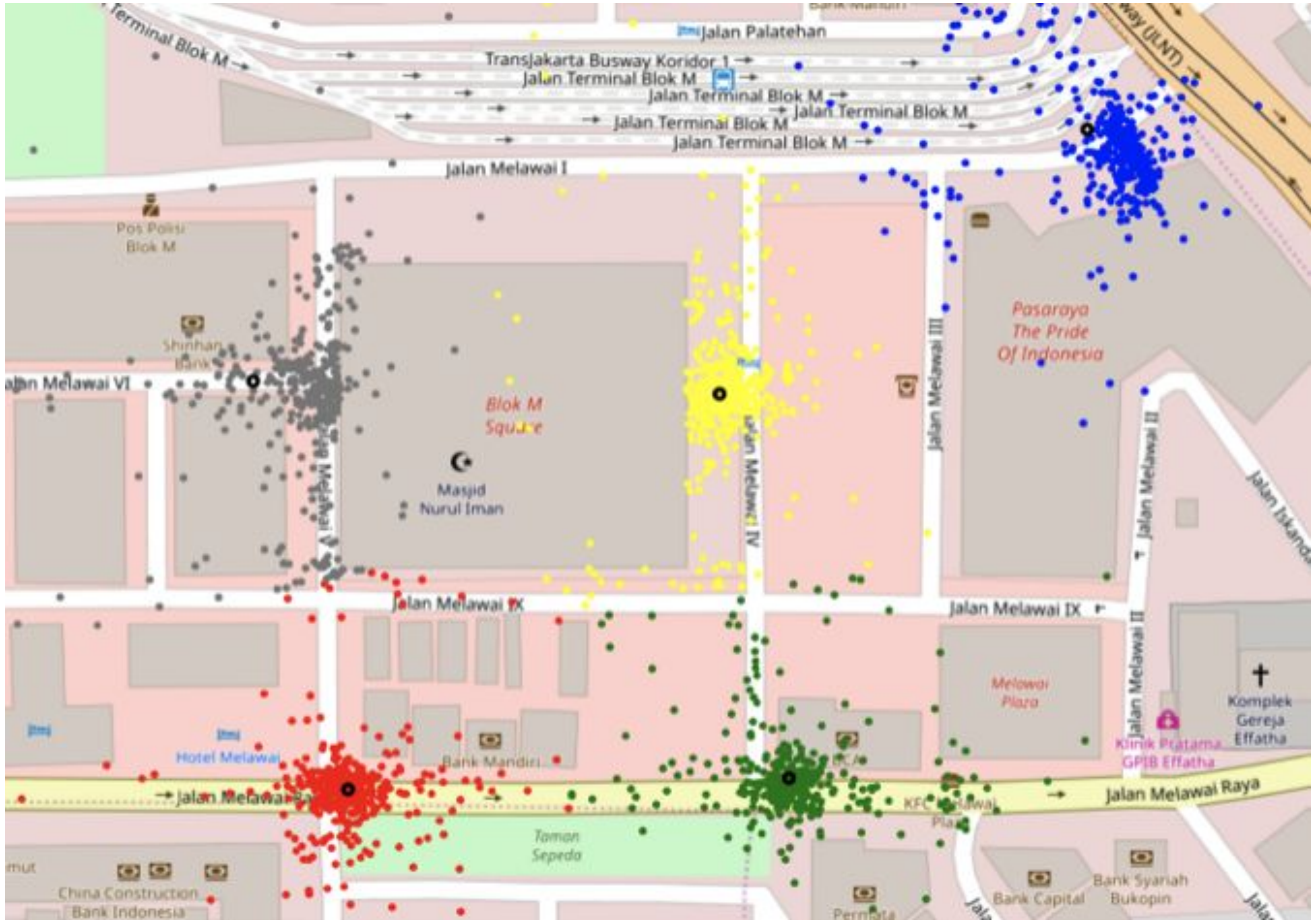
USE CASES

- Marketing
  - **Market segmentation** of customers based on demographic and transaction history and tailored marketing strategy
  - **Market structure analysis** identifying groups of similar products according to competitive measures of similarity
- Finance
  - **Balanced portfolios** by choosing stocks from different clusters
  - **Industry Analysis** finding similar firms through “market measures”
- Accounting
  - **Group transactions** by type
  - **Anomaly detection**

Data Mining for Business Analytics: Concepts, Techniques, and Applications in R by GalitShmueli, Peter C. Bruce, InbalYahav, Nitin R. Patel, Kenneth C. LichtendahlJr. (2018)

USE CASE - GO-JEK FANTASTIC DRIVERS

- Go-Jek used K-Means algorithm to identify their better drivers
- It also helped them "pin" pick up points at popular locations



Find out more here: https://blog.gojekengineering.com/fantastic-drivers-and-how-to-find-them-a88239ef3b29

TYPES OF CLUSTERING ALGORITHMS

- There are two general types of clustering algorithms:
  a. Hierarchical
     - Agglomerative - begins with *n* clusters and sequentially merge similar clusters until a single cluster is obtained
     - Divisive - starts with a single cluster including all records and does the opposite
  b. Non-hierarchical *(Focused for this class; k-means clustering)*
     - Using predetermined number of clusters to assign observations to each cluster
     - Less computationally intensive and preferred for larger datasets

Data Mining for Business Analytics: Concepts, Techniques, and Applications in R by GalitShmueli, Peter C. Bruce, InbalYahav, Nitin R. Patel, Kenneth C. LichtendahlJr. (2018)

K-MEANS CLUSTERING

- K-Means Algorithm
- Distance Scoring
- Optimal K
- Limitations of K-Means

K-MEANS ALGORITHM

- Start with *k* initial clusters (*k* needs to be pre-defined)
- At every step, each record is reassigned to the cluster with the "closest" centroid
- Recompute the centroids of clusters that lost or gained a record, and repeat Step 2
- Stop when moving any more records between clusters **increases cluster dispersion**

DISTANCE SCORING - BETWEEN TWO OBSERVATIONS

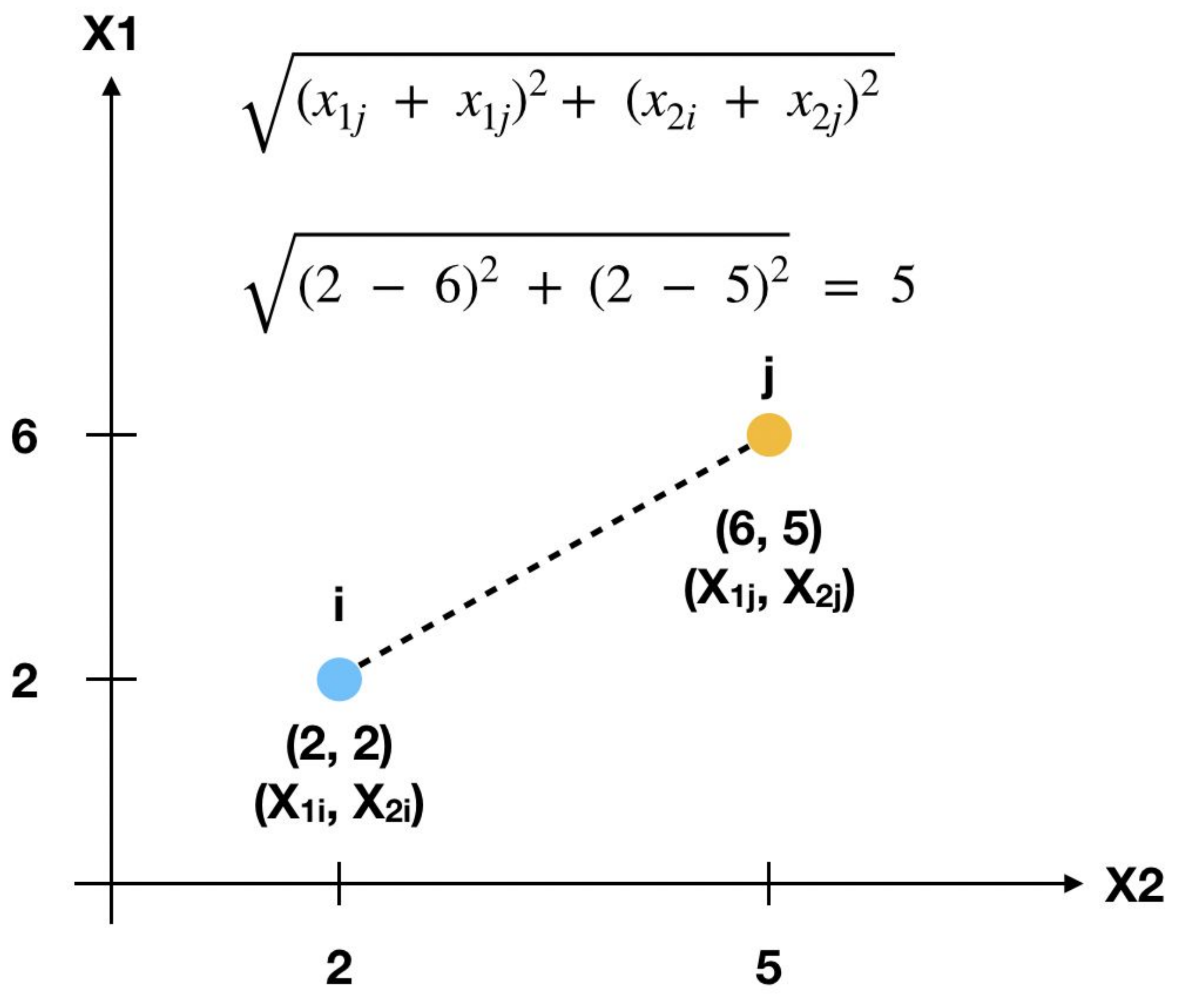
- Suppose two different observations are i and j, the distance metric for them is d_{ij}
- The formula to calculate the distance between two observed points is the **Euclidean Distance**:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

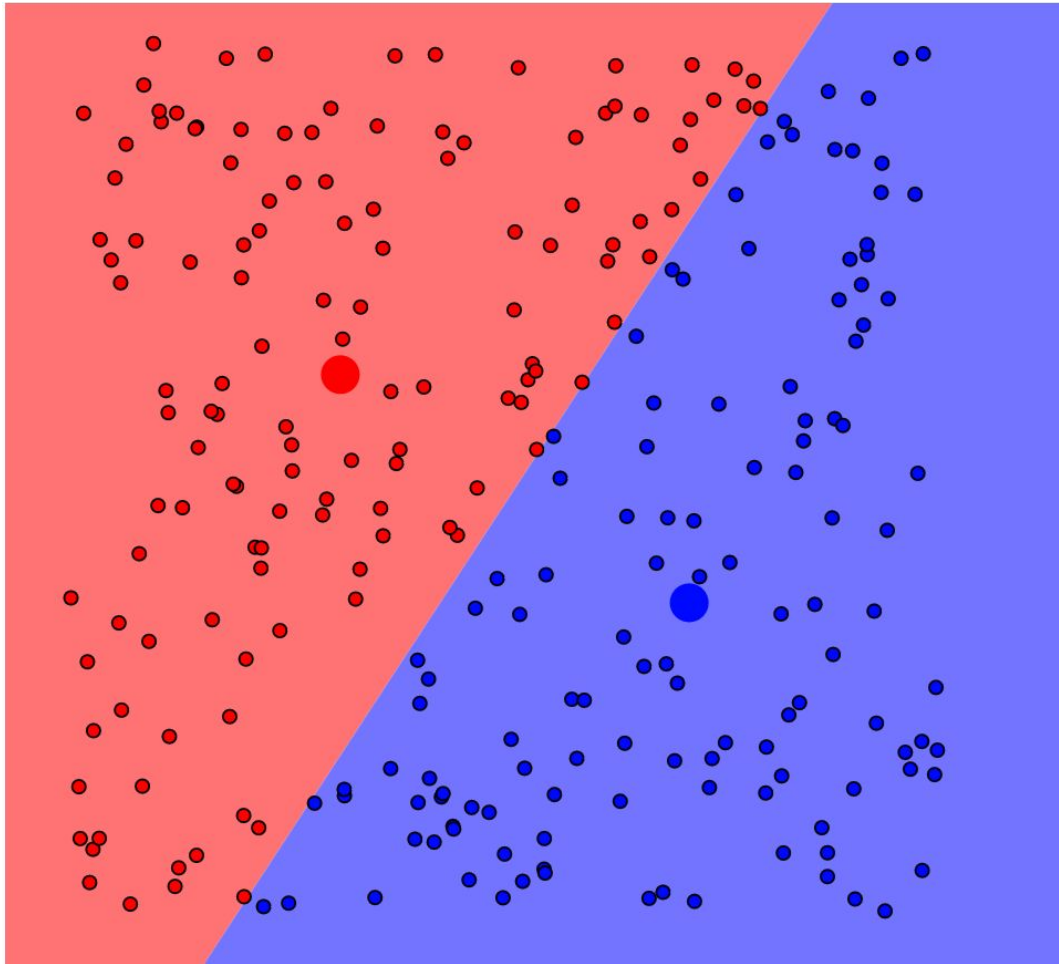
Data Mining for Business Analytics: Concepts, Techniques, and Applications in R by GalitShmueli, Peter C. Bruce, InbalYahav, Nitin R. Patel, Kenneth C. LichtendahlJr. (2018)

DISTANCE SCORING - BETWEEN TWO OBSERVATIONS

ID	X1	X2
i	2	5
j	2	6



K-MEANS ALGORITHM - VISUALIZED



Click here for an interactive k-means algorithm animation: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

OPTIMAL K

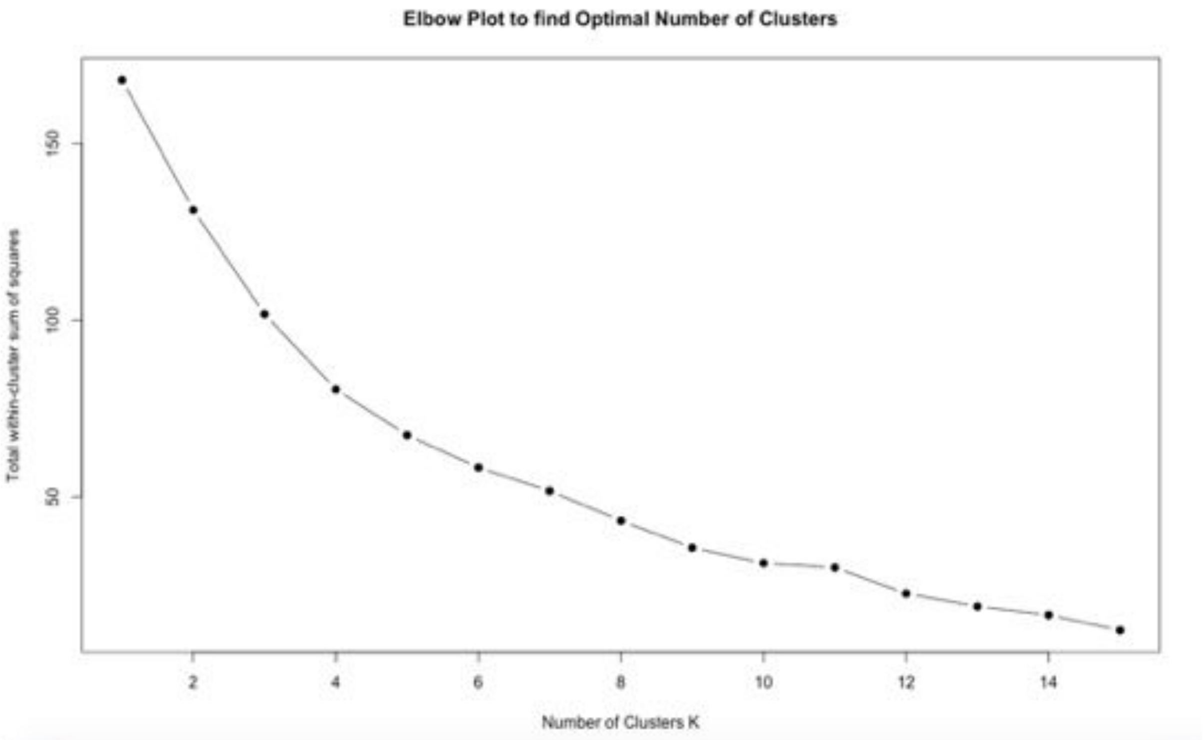
- In many cases, there is a lack of information to be used for the initial number of *k*.
- K-means algorithm, like other clustering algorithms, work towards compressing the *k* number of data-points by summarising them by their "means" (hence, k-means)
- There is no right number of *k*'s. However, information on within-cluster dispersion can assist in determining the optimal number of *k*
- With an "elbow chart", we can graphically evaluate whether there is a decline in cluster heterogeneity when more clusters are added (i.e *k* is increased)

Data Mining for Business Analytics: Concepts, Techniques, and Applications in R by GalitShmueli, Peter C. Bruce, InbalYahav, Nitin R. Patel, Kenneth C. LichtendahlJr. (2018)

OPTIMAL K - ELBOW PLOT

- Select a K where the next increase in K has **little to no improvement** in the total within cluster sum of squares (WSS)
- Higher the K, the more computation required

choose the k with the sharpest drop



Data Mining for Business Analytics: Concepts, Techniques, and Applications in R by
Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, Kenneth C. Lichtendahl Jr. (2018)

SPOTIFY DATASET - SELECTING K*

```python
X = df_spotify_cluster # <<< Numerical DataFrame here
distorsions = []
for k in range(2, 20):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(X)
    distorsions.append(kmeans.inertia_)

fig = plt.figure(figsize=(15, 5))
plt.plot(range(2, 20), distorsions)
plt.grid(True)
plt.title('Elbow curve')
```

Loops for clusters from 2 to 19

Initialises based on K clusters

Calculates within cluster sum of squared

Plots elbow plot

SPOTIFY DATASET - K-MEANS ALGORITHM*

```
k = 5
model = KMeans(n_clusters=k,   # < Initialise Number Of Clusters here
                   random_state=0)

spotify_kmeans = model.fit(df_spotify_cluster) # < DataFrame of All Variables
print(spotify_kmeans)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=5, n_init=10, n_jobs=1, precompute_distances='auto',
    random_state=0, tol=0.0001, verbose=0)
```

Initialises the KMeans model based on K clusters

Trains the model by fitting in all variables into the model and **returns a kmeans result set**

IN-CLASS PRACTICE - CLUSTERING*

- Try out the in-class practice with the credit card spending behaviour
- Do a summary statistics on the different clusters of credit card spending behaviours