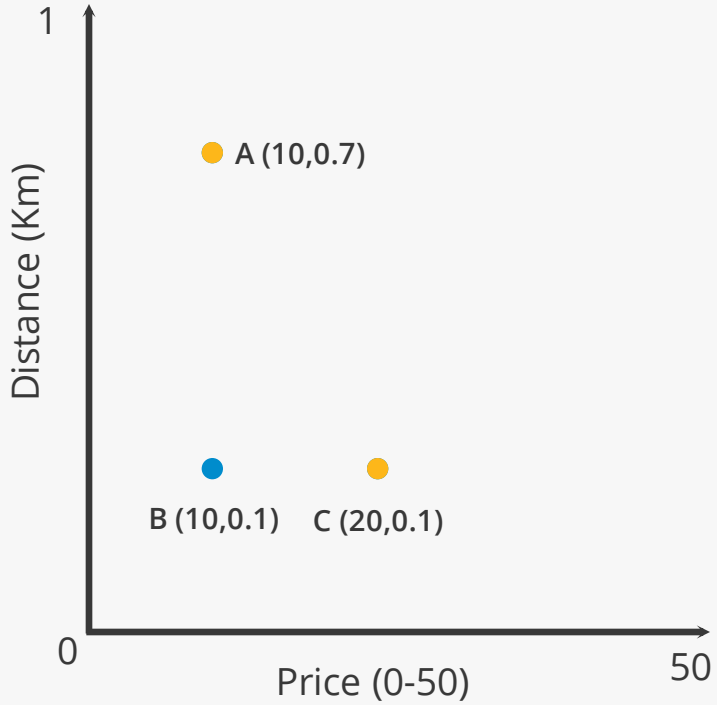
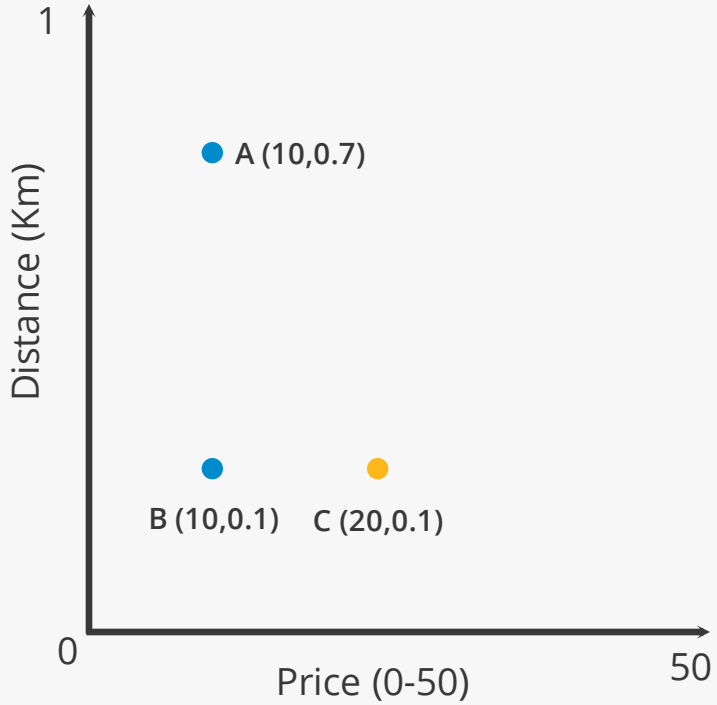
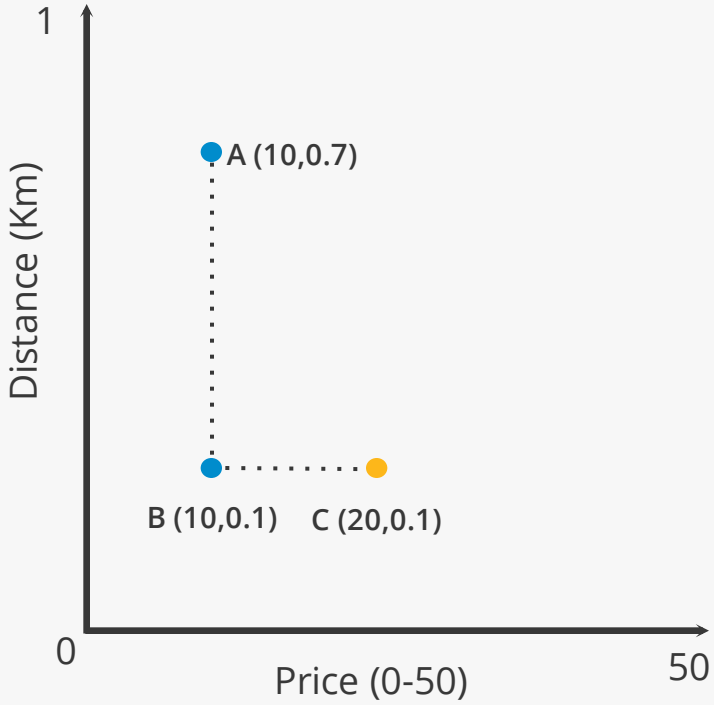
# Problems with distance based algorithms



# Problems with distance based algorithms



# Problems with distance based algorithms

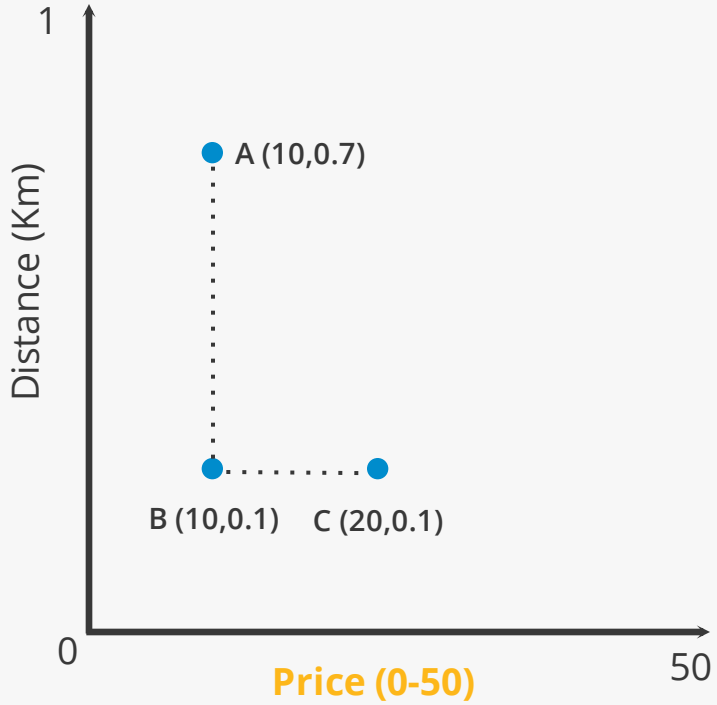



**Manhattan Distance**

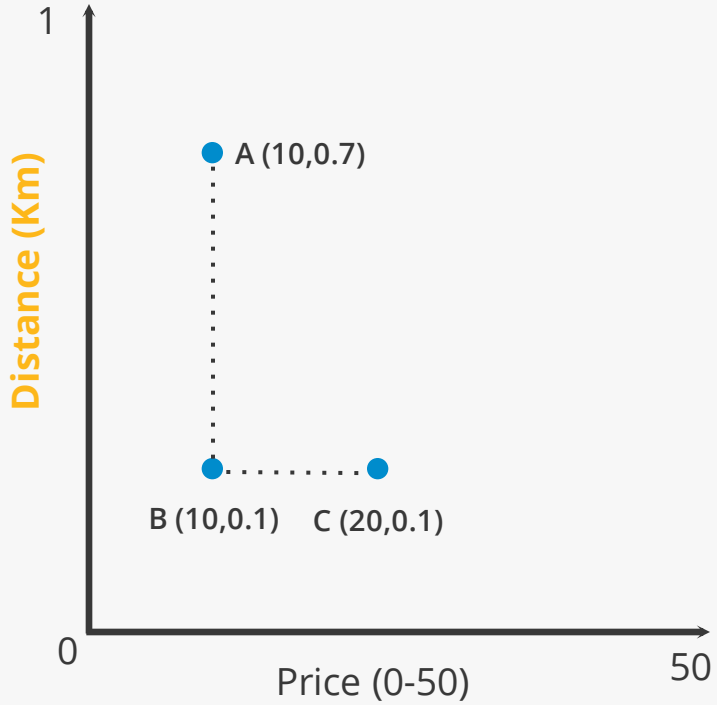
**BA** = 0.6 Units

**BC** = 10 Units

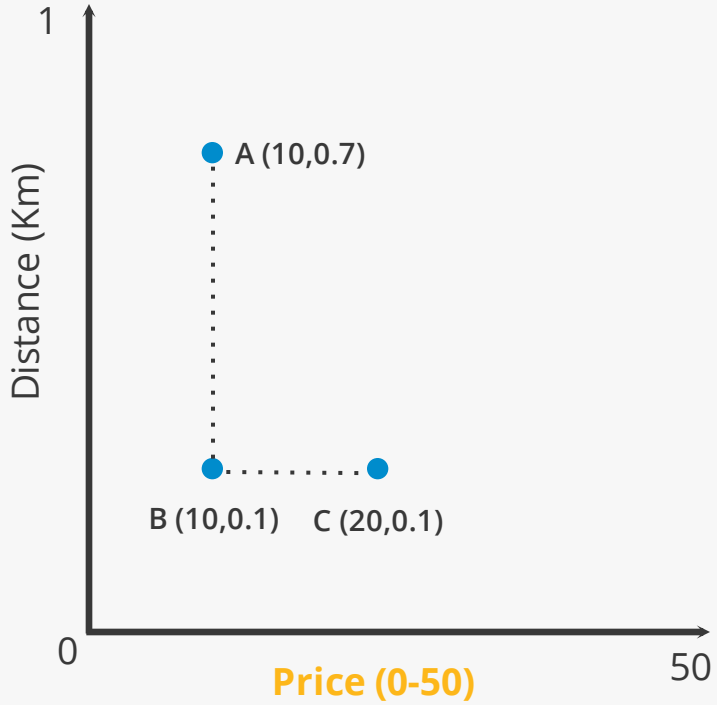
# Problems with distance based algorithms



# Problems with distance based algorithms



# Problems with distance based algorithms



# What is the appropriate solution?

# Scaling the Variables

# Scaling the Variables

- **Scaling** changes the range and distribution of data values to make them consistent across variables.
- Scaling ensures **different features contribute equally** to the performance of the ML model.
- **Scaled variables become unitless** but withhold all the information it is associated with.

# Methods to Scale Variables

# Min-Max Scaler or Normalization

Min-Max Scaling Transforms features by scaling each feature to a range between **0 and 1**.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

# Standard Scaler or Z-Score Normalization

Standard Scaling centers the features **around zero with a standard deviation of one**.

$$X_{scaled} = \frac{X - mean(X)}{std(X)}$$

# Robust Scaler

Standard Scaling centers the features **based on percentiles** and are not influenced by outliers.

$$X_{scaled} = \frac{X - median(X)}{IQR(X)}$$

# Robust Scaler

Standard Scaling centers the features **based on percentiles** and are not influenced by outliers.

$$X_{scaled} = \frac{X - median(X)}{IQR(X)}$$

# Choice of Scaling

- Nature of the data
- Algorithm being used

***Note:** When scaling, we lose the context of the value with reference to our problem statement