Distance Measures

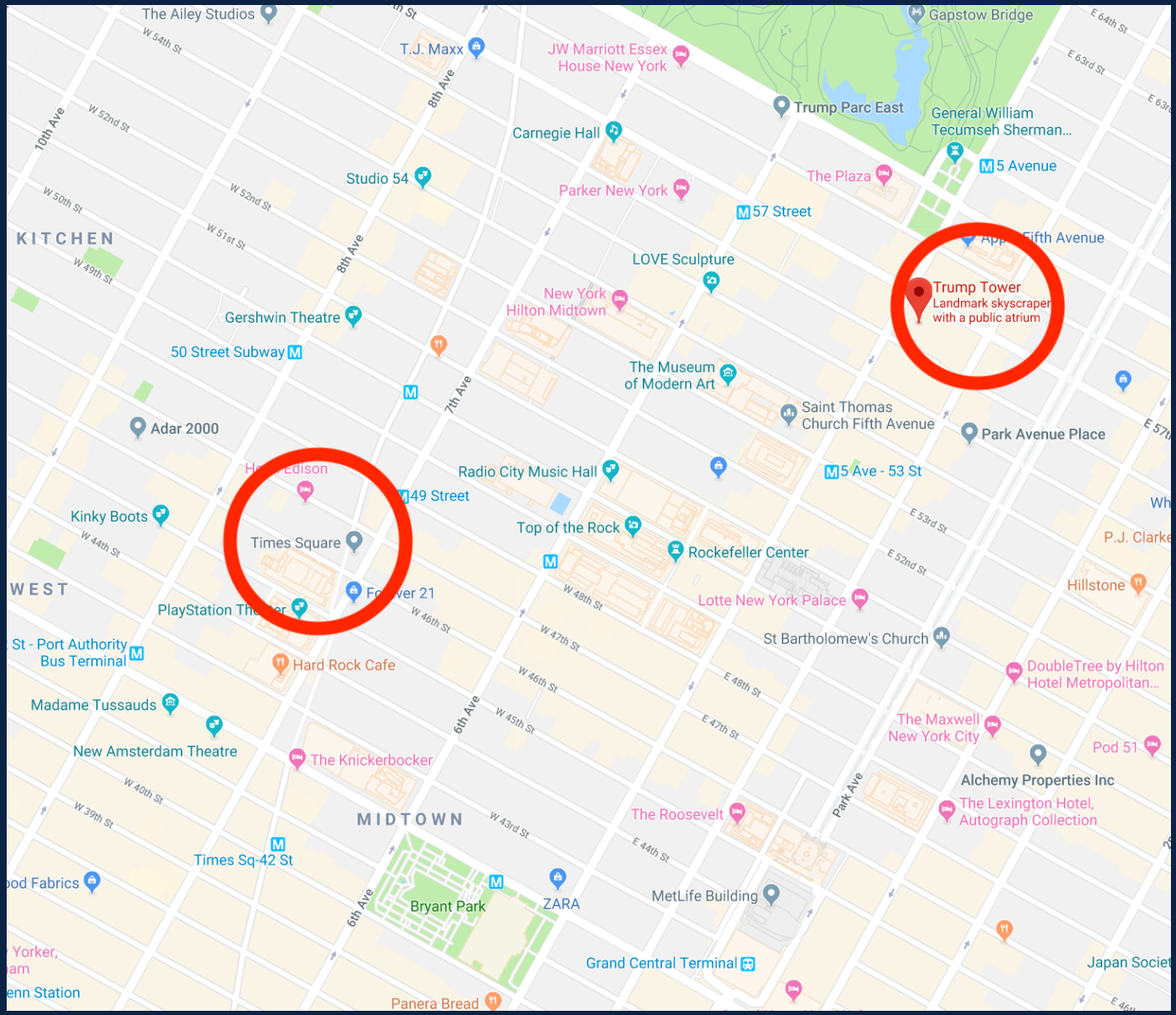
UBCO MDS — DATA 573

Distances

- Prior to diving deep on unsupervised methods, it's important to note that the vast majority are based on distance calculations (of some form)
- Distance seems like a straightforward idea, but it is actually quite complex....

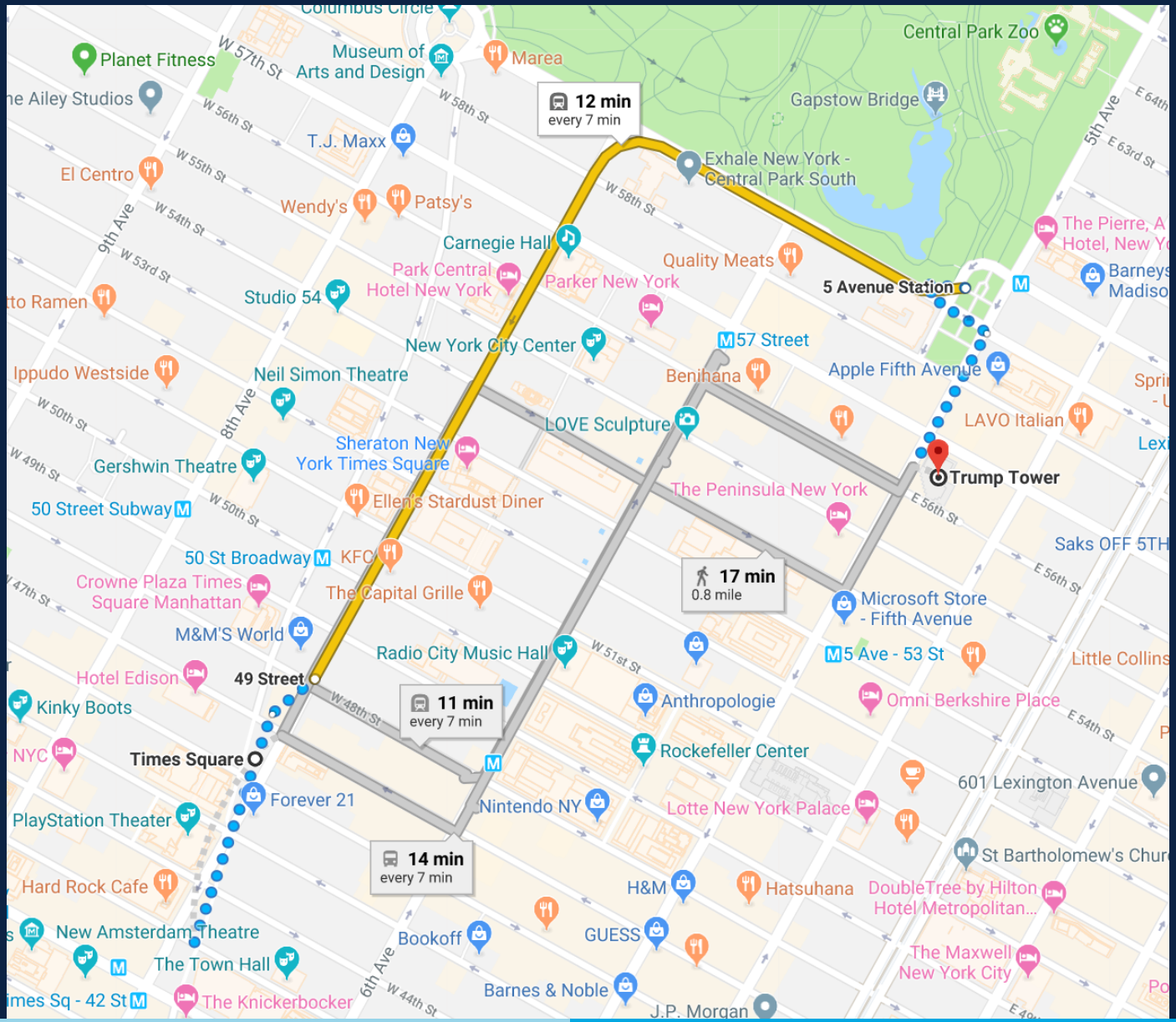
Distance Motivation

- Distance from Trump Tower to Times Square?



Distance Motivation

- Distance from Trump Tower to Times Square?



Distance Motivation

- Distance from Trump Tower to us?

Euclidean Distance

- To begin, we'll consider that all predictors are numeric.
- First, the main distance that you know of...
- Euclidean distance between observations i and j simply as

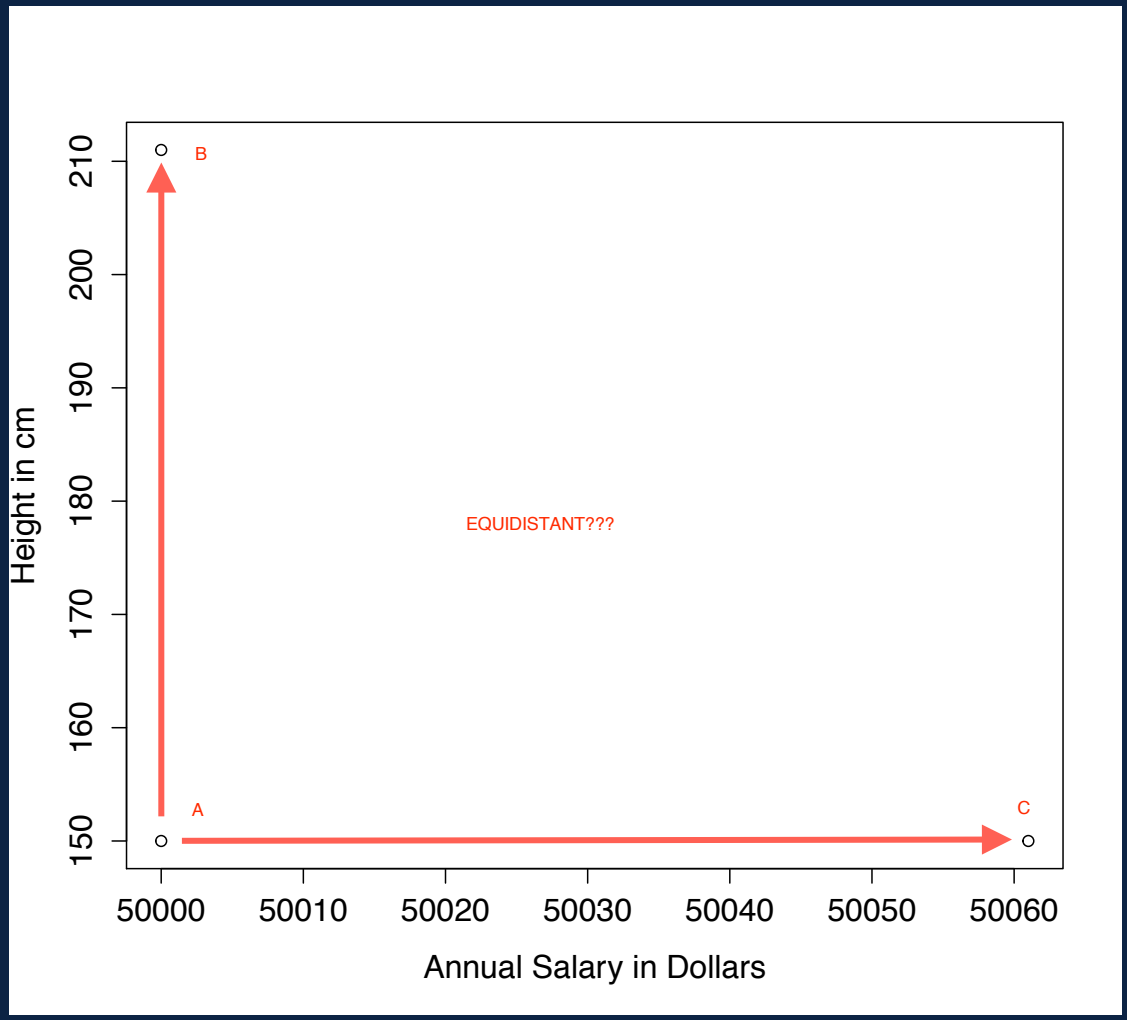
$$d_{ij}^{\text{EUC}} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

is inappropriate in many settings.

Defining Distances

- And now, why the main distance that you know of is immediately problematic...
- Consider the following measurements on people:
  - Height (in cm)
  - Annual salary (in $)
- A $61 difference in annual salary would be considered a miniscule difference, whereas a 61 cm difference in height (approx 2 feet) would be substantial!

Defining Distances



Standardized Euclidean Distance

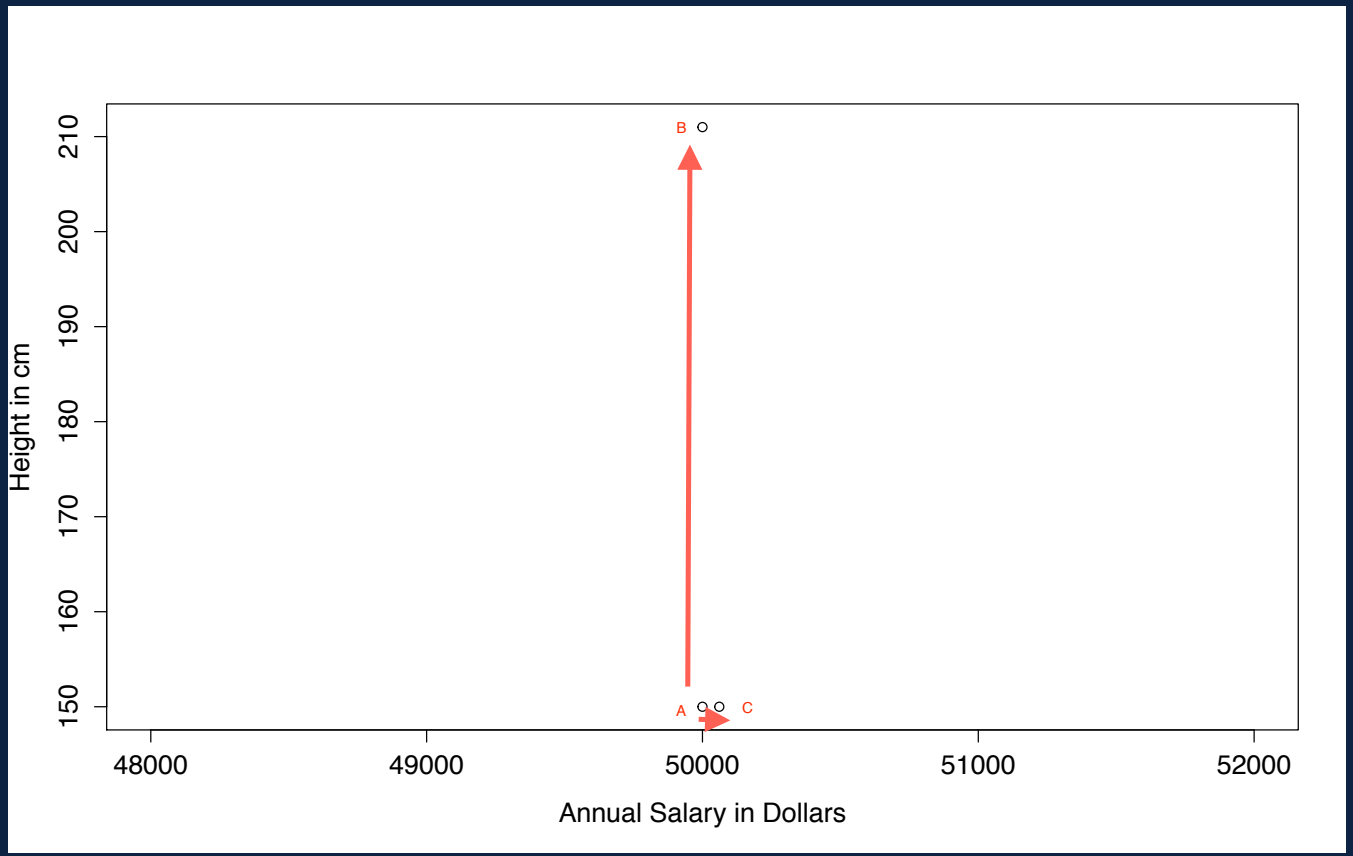
- Scale the data to have mean 0, variance 1 via

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}$$

- Then we can define pairwise distances as

$$d_{ij}^{\text{STA}} = \sqrt{\sum_{k=1}^{p} (z_{ik} - z_{jk})^2}$$

Defining Distances



Manhattan (or City-Block) Distance

- Manhattan or city-block distance merely measures distance as though one could only travel along the axes.
- So pairwise distance are calculated as

$$d_{ij}^{\text{MAN}} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

- Similarly, one might want to consider the standardized form

$$d_{ij}^{\text{MANs}} = \sum_{k=1}^{p} |z_{ik} - z_{jk}|$$

Mahalanobis Distance

- Mahalanobis distance takes into account the covariance structure of the data.
- Recall, we saw this distance crop-up in our discussion of LDA/QDA.
- It is easiest defined in matrix form

$$d_{ij}^{\text{MAH}} = (\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

- Note that Mahalanobis distance is a generalization of standardized euclidean distance (you can thing of it as standardizing by both variance and covariances).

When to Standardize?

- This brings us back to a question we've now asked at least once before...when should we use a standardized measure, and when should we not?
- Same general rule as with PCA — that is, it is often a good idea, and generally necessary if measurements are on vastly different units.
- Another way to put it: unless you have a good reason to believe that higher variance measures should be heavier weighted in your unsupervised method, then you probably want to standardize in some form.
- FYI: some unsupervised methods are scale invariant, in which case you don't have to worry about it.

From Numeric to Mixed

- What if some/all the predictors are not numeric?
- We'll start with the binary case through the following example

President	Democrat	Governor	VP	2nd Term	From Iowa
GW Bush	0	1	0	1	0
Obama	1	0	0	1	0
Trump	0	0	0	0	0
Biden	1	0	1	?	0

Matching Binary Distance

President	Democrat	Governor	VP	2nd Term	From Iowa
GW Bush	0	1	0	1	0
Obama	1	0	0	1	0
Trump	0	0	0	0	0
Biden	1	0	1	?	0

- Equivalent to Manhattan/City-block,

$$d_{ij}^{\text{MAT}} = \#\text{ of variables with opposite value}$$

- Example GWBush vs Obama:

all these match

$$d_{12}^{\text{MAT}} = \underbrace{|0-1|}_{1} + \underbrace{|1-0|}_{1} + |0-0| + |1-1| + |0-0| = 2$$

- Often divided by p for simple interpretation

$$d_{12}^{\text{MATp}} = \frac{2}{5} = 0.4$$

Asymmetric Binary Distance

President	Democrat	Governor	VP	2nd Term	From Iowa
GW Bush	0	1	0	1	0
Obama	1	0	0	1	0
Trump	0	0	0	0	0
Biden	1	0	1	?	0

- But are the 0's in "From Iowa" really matches?
- Asymmetric binary:

$$d_{ij}^{\text{ASY}} = \frac{\#\text{ of 0-1 or 1-0 pairs}}{\#\text{ of 0-1, 1-0, or 1-1 pairs}}$$

- Example GWBush vs Obama:

$$d_{12}^{\text{ASY}} = \frac{2}{3} = 0.667$$

Distance for Qualitative Variables

- For nominal variables, a common measure is essentially standardized matching once again
- If $u = \#$ of variables that match between observations i and j and $p = \#$ of variables then

$$d_{ij}^{\text{NOM}} = 1 - \frac{u}{p}$$

Gower's Distance

- Often data has a mix of variable types. Gower's distance is a common choice for computing pairwise distances in this case.
- Basic idea is to ensure that each variable's contribution to total distance is standardized between 0 and 1, and then sum them up
- Gower's distance:

$$d_{ij}^{\text{GOW}} = \frac{\sum_{k=1}^{p} \delta_{ijk} d_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

where

- $\delta_{ijk} = 1$ if both x_{ik} and x_{jk} are non-missing (and note that a 0-0 binary match is often counted as "missing" to match asy-binary)
- $\delta_{ijk} = 0$ otherwise.
- and d_{ijk} depends on variable type...

Gower's Distance

- Quantitative numeric

$$d_{ijk} = \frac{|x_{ik} - x_{jk}|}{\text{range of variable k}}$$

- Qualitative (nominal, categorical)

$d_{ijk} = 0$ if obs i and j agree on variable k, 1 otherwise

- Binary (asymmetric)

$d_{ijk} = 0$ if obs i and j are a 1-1 match, 1 otherwise

- Example on board.


THE UNIVERSITY OF BRITISH COLUMBIA