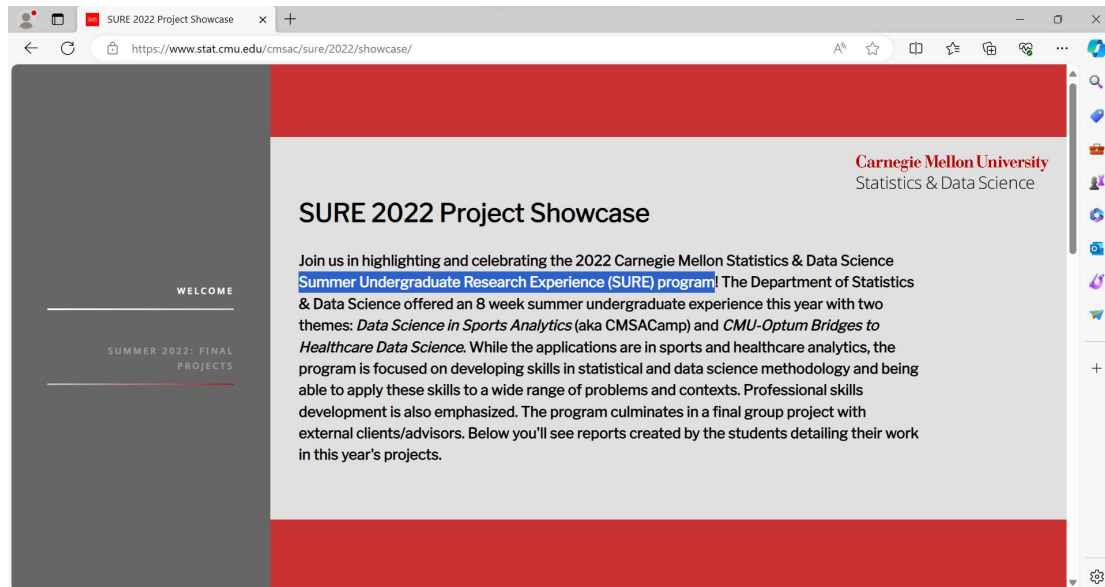


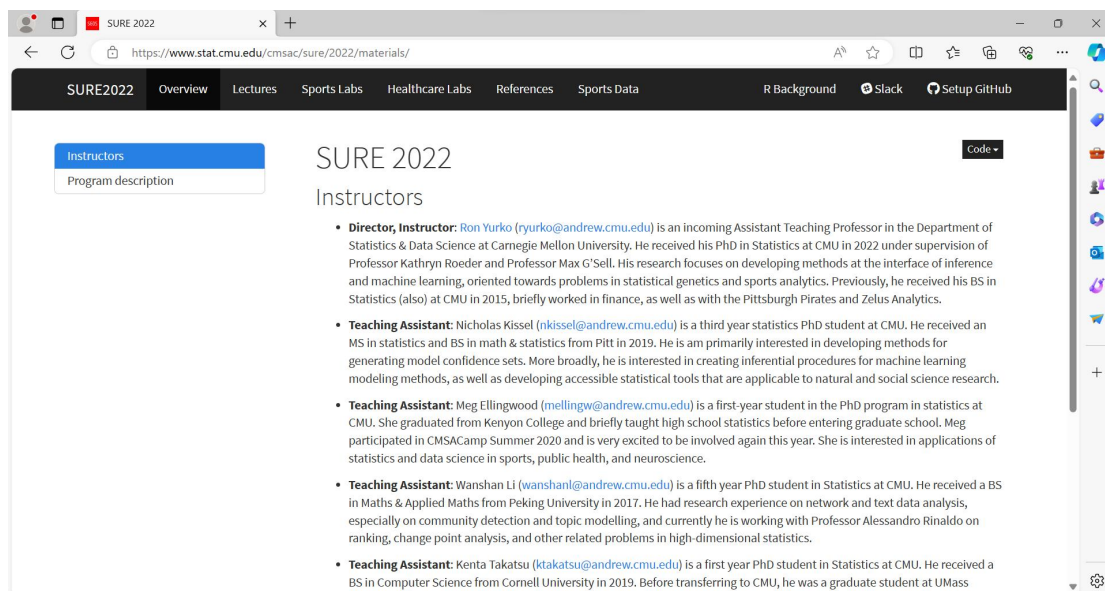
Carnegie Mellon University

University Undergraduate Course: The 2022 Carnegie Mellon Statistics & Data Science Summer Undergraduate Research Experience (SURE) program

<https://www.stat.cmu.edu/cmsac/sure/2022/showcase/>
(retrieved 11 March 2024)



<https://www.stat.cmu.edu/cmsac/sure/2022/materials/>



<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/>

Lectures

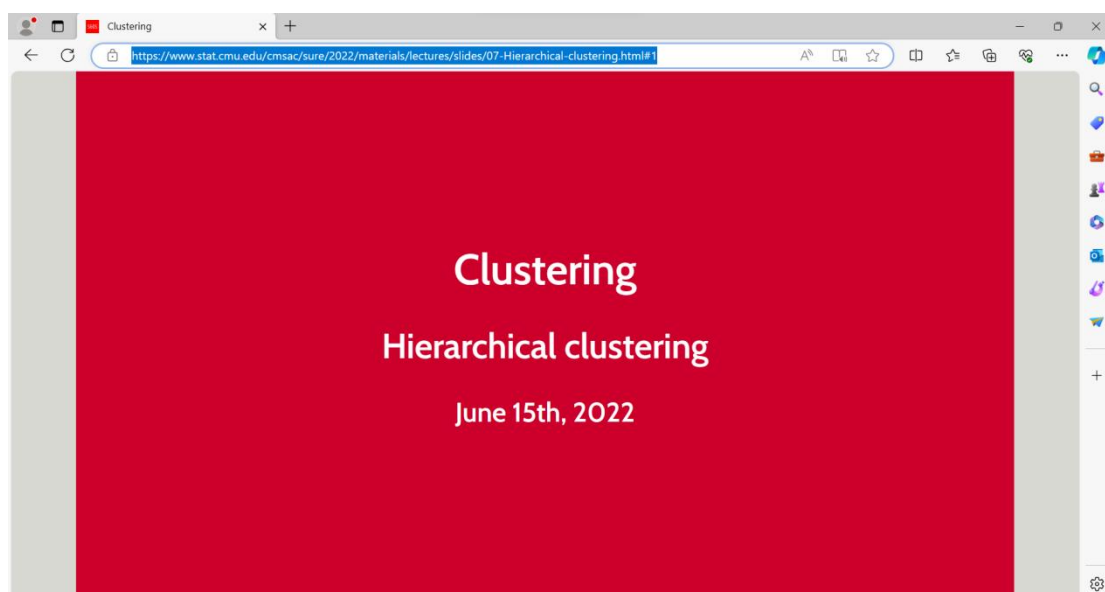
Ron Yurko

Contents

Lecture	Title	HTML	Rmd
Lecture 0	Welcome to SURE: Background and overview	HTML	Rmd
Lecture 1	Exploring data: Into the tidyverse	HTML	Rmd
Lecture 2	Data Visualization: The grammar of graphics and ggplot2	HTML	Rmd
Lecture 3	Data Visualization: Visualizing 1D categorical and continuous variables	HTML	Rmd
Lecture 4	Data Visualization: Visualizing 2D categorical and continuous by categorical	HTML	Rmd
Lecture 5	Data Visualization: Density estimation	HTML	Rmd
Lecture 6	Clustering: K-means	HTML	Rmd
Lecture 7	Clustering: Hierarchical clustering	HTML	Rmd
Lecture 8	Presentations: And working with xaringan and xaringanthemer	HTML	Rmd
Lecture 9	Model-based clustering: Gaussian mixture models	HTML	Rmd

<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html>

<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#1>
(retrieved 11 March 2024)



<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#32>

Clustering

https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#32

Minimax linkage

- Each cluster is defined **by a prototype** observation (most representative)
- Identify the point whose farthest point is closest** (hence the minimax)

Complete

Centroid

Minimax

- Use this minimum-maximum distance as the measure of cluster dissimilarity
- Dendrogram interpretation: each point is $\leq h$ in dissimilarity to the **prototype** of cluster
- Cluster centers are chosen among the observations themselves - hence prototype**

19

<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#33>

Clustering

https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#33

Minimax linkage example

- Easily done in R via the **protoclust** package
- Use the **protoclust()** function to apply the clustering to the **dist()** object

```
library(protoclust)
nba_minimax <- protoclust(player_dist)
ggdendrogram(nba_minimax,
  theme_dendro = FALSE,
  labels = FALSE,
  leaf_labels = FALSE) +
  labs(y = "Maximum dissimilarity from prototype")
theme_bw() +
  theme(axis.text.x = element_blank(),
    axis.title.x = element_blank(),
    axis.ticks.x = element_blank(),
    panel.grid = element_blank())
```

20

<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#34>

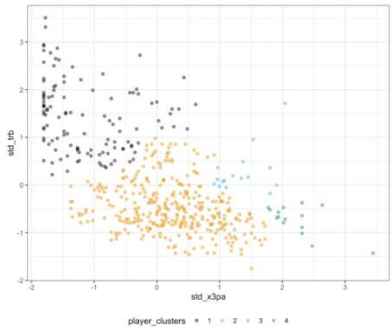
Clustering

https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#34

Minimax linkage example

- Use the `protocut()` function to make the cut
- But then access the cluster labels `c1`

```
minimax_player_clusters <-
  protocut(nba_minimax, k = 4)
nba_filtered_stats %>%
  mutate(player_clusters =
    as.factor(minimax_player_clusters$c1))
ggplot(aes(x = std_x3pa, y = std_trb,
  color = player_clusters)) +
  geom_point(alpha = 0.5) +
  ggthemes::scale_color_colorblind() +
  theme_bw() +
  theme(legend.position = "bottom")
```



21

<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#35>

Clustering

https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#35

Minimax linkage example

- Want to check out the prototypes for the three clusters
- `protocut` returns the indices of the prototypes (in order of the cluster labels)

```
minimax_player_clusters$protos
```

```
## [1] 468 347 103 251
```

- View these player rows using `slice`:

```
nba_filtered_stats %>%
  dplyr::select(player, pos, age, std_x3pa, std_trb) %>%
  slice(minimax_player_clusters$protos)
```

```
## # A tibble: 4 x 5
##   player      pos    age std_x3pa std_trb
##   <chr>    <chr> <dbl>   <dbl>   <dbl>
## 1 Domantas Sabonis C-PF    25   -1.02    1.99
## 2 Jalen Suggs      PG     20    0.161  -0.691
## 3 Luka Dončić      PG     22    1.53    0.955
## 4 Ben McLemore     SG     28    2.47   -1.28
```

22

<https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#39>

Clustering

https://www.stat.cmu.edu/cmsac/sure/2022/materials/lectures/slides/07-Hierarchical-clustering.html#39

Wrapping up...

- For context, how does player position (**pos**) relate to our clustering results?

```
table("Clusters" = minimax_player_clusters$cl, "Positions" = nba_filtered_stats$pos)
```

```
##           Positions
## Clusters  C  C-PF PF  PF-SF PG  PG-SG SF  SF-SG SG  SG-PG SG-SF
## 1  71    2  34    0  0    0  8    0  3    0  0
## 2  13    0  54    1  88    1  76    5  90    2  4
## 3  1     0  4     0  2     0  6     0  3     0  0
## 4  0     0  1     0  2     0  2     0  9     1  0
```

- Can see positions tend to fall within particular clusters...
- What's the way to visually compare the two labels?
- We can easily include more variables** - just changes our distance matrix
- But we might want to explore **soft** assignments instead...

23