Dimension Reduction in Practice

February 12, 2026

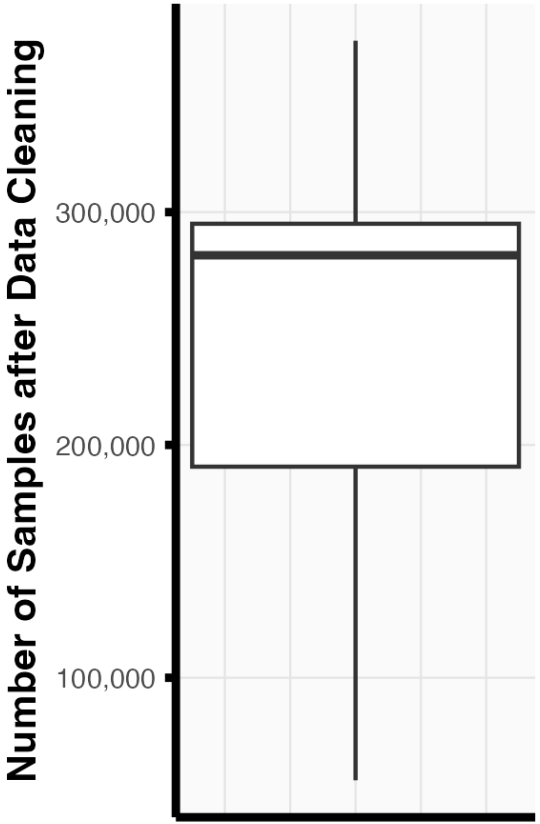
Lab 1 Recap

- Lab 1 has been graded
- Example "solution" is on GitHub/course website (qmd and html)
- Main Takeaways:
  - It is possible to do data cleaning without relying on statistical measures of "outliers".
  - When possible, the data collection process should inform our data cleaning.
  - Document and justify your choices.
- For future labs:
  - Grading for code reproducibility will be more strict
  - Do not push data/ folder to GitHub (check github.com if you aren't sure what you pushed

Lab 1 Data Cleaning Issues

- NAs
- Duplicates
  - Duplicate entries that appear on log and network
  - Duplicate entries for the same (epoch, nodeid) combination
- Erroneous temperature, humidity, and PAR measurements
- Battery failure is the primary cause of erroneous measurements
  - **Voltage units differ between log and network data**
  - To use voltage as a filter, need to convert voltage to same units
- **Two trees (interior and edge; see *readme*)**
- "result_time" variable in log data is a constant
  - Need to merge by epoch, not "result_time"
- Motes with missing location information
- Incident PAR > Reflected PAR

Lab 1 Recap



Max ~ 373,632 (90%)

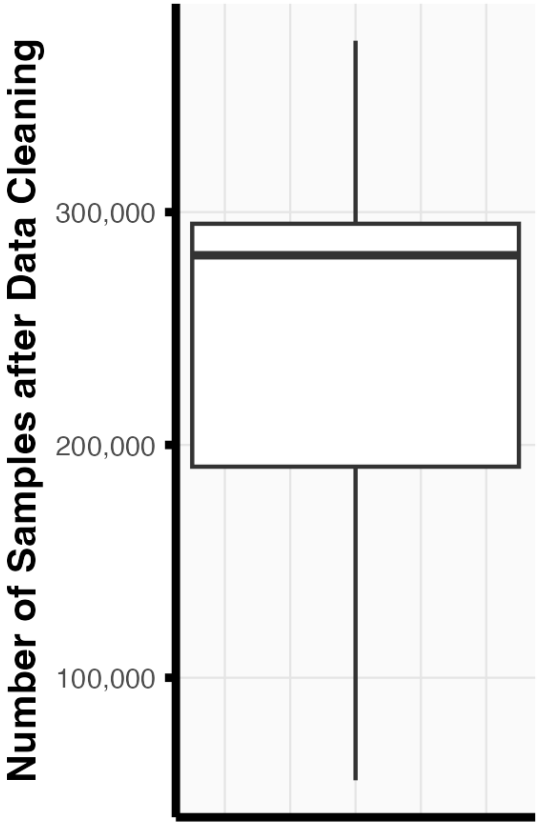
Q3 ~ 295,005 (71%)

Median ~ 281,469 (68%)

Q1 ~ 190,661 (46%)

Min ~ 55,953 (13%)

Lab 1 Recap



Max ~ 373,632 (90%)

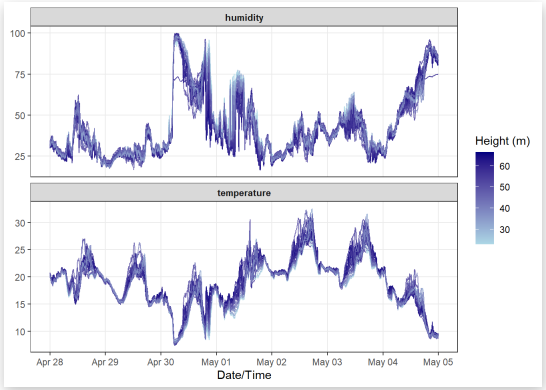
Q3 ~ 295,005 (71%)

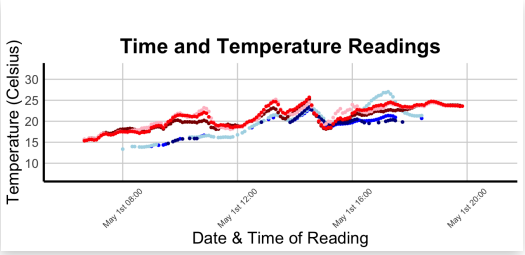
Median ~ 281,469 (68%)

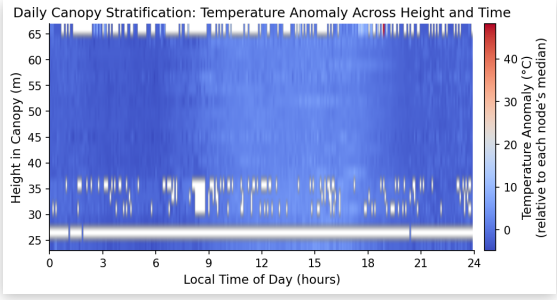
Q1 ~ 190,661 (46%)

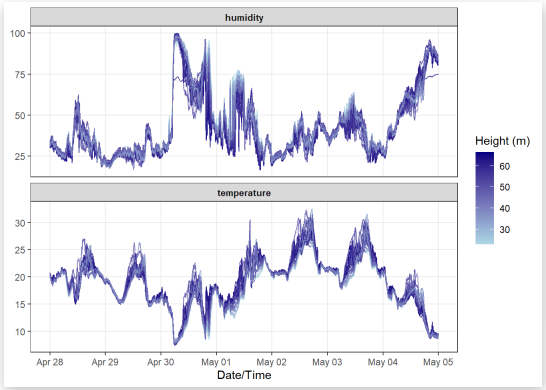
Min ~ 55,953 (13%)

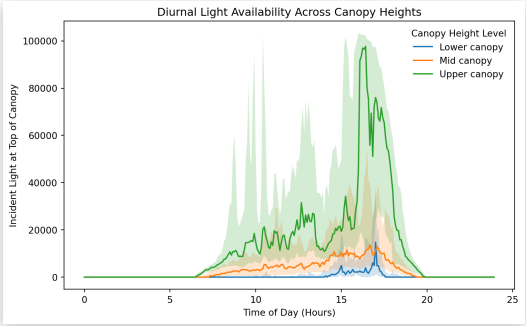
* Example solution ~ 225,000 (54%)

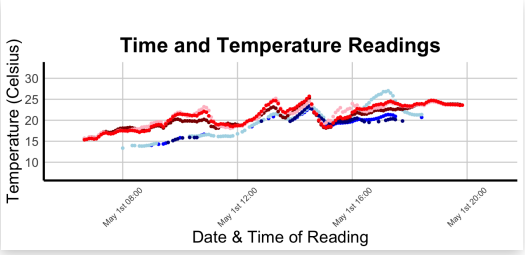

humidity
100
75
50
25
temperature
30
25
20
15
10
Apr 28
Apr 29
Apr 30
May 01
May 02
May 03
May 04
May 05
Date/Time
Height (m)
60
50
40
30

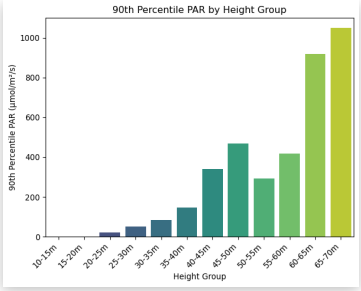

Time and Temperature Readings
Temperature (Celsius)
30
25
20
15
10
May 1st 08:00
May 1st 12:00
May 1st 16:00
May 1st 20:00
Date & Time of Reading

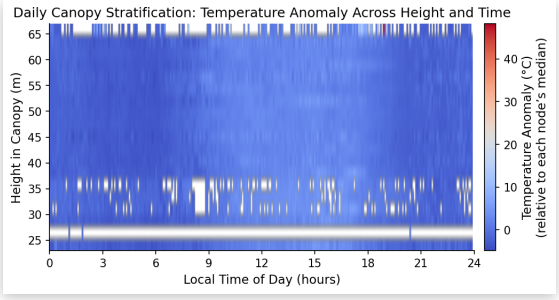

Daily Canopy Stratification: Temperature Anomaly Across Height and Time
Height in Canopy (m)
65
60
55
50
45
40
35
30
25
0
3
6
9
12
15
18
21
24
Local Time of Day (hours)
Temperature Anomaly (°C)
(relative to each node's median)
40
30
20
10
0

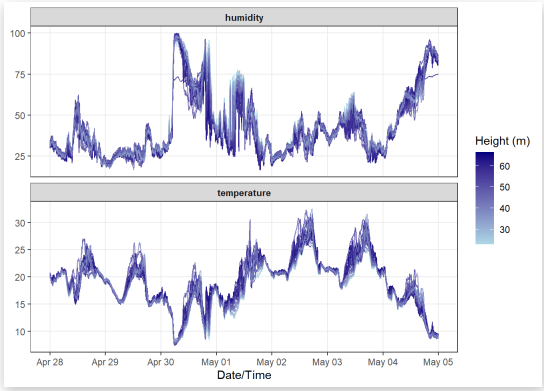

humidity
temperature
Height (m)
60
50
40
30
Apr 28
Apr 29
Apr 30
May 01
May 02
May 03
May 04
May 05
Date/Time

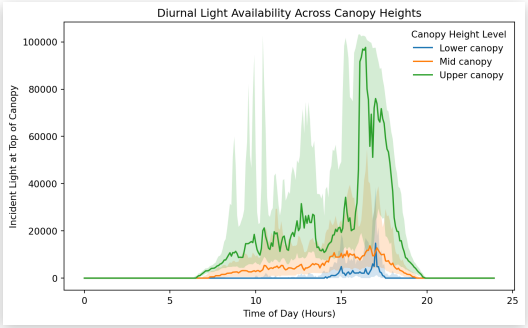

Diurnal Light Availability Across Canopy Heights
Canopy Height Level
Lower canopy
Mid canopy
Upper canopy
Incident Light at Top of Canopy
Time of Day (Hours)

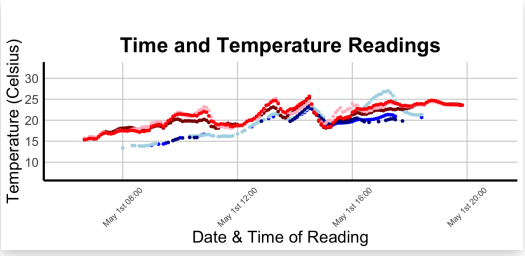

Time and Temperature Readings
Temperature (Celsius)
May 1st 08:00
May 1st 12:00
May 1st 16:00
May 1st 20:00
Date & Time of Reading

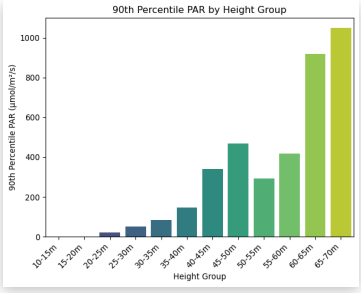

90th Percentile PAR by Height Group
90th Percentile PAR (µmol/m²/s)
10-15m
15-20m
20-25m
25-30m
30-35m
35-40m
40-45m
45-50m
50-55m
55-60m
60-65m
65-70m
Height Group

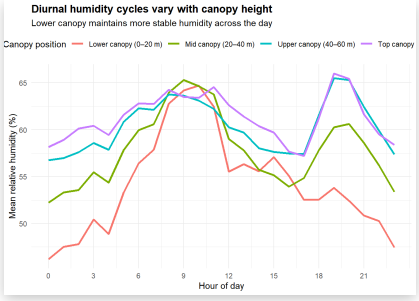

Daily Canopy Stratification: Temperature Anomaly Across Height and Time
Height in Canopy (m)
Local Time of Day (hours)
Temperature Anomaly (°C)
(relative to each node's median)

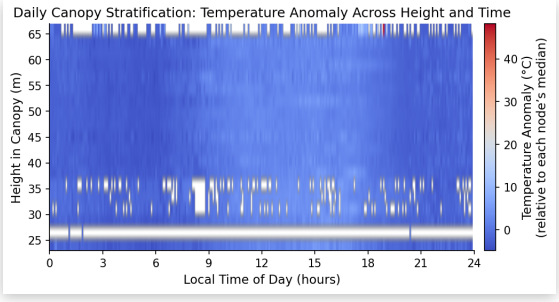

humidity
temperature
Height (m)
60
50
40
30
Apr 28
Apr 29
Apr 30
May 01
May 02
May 03
May 04
May 05
Date/Time

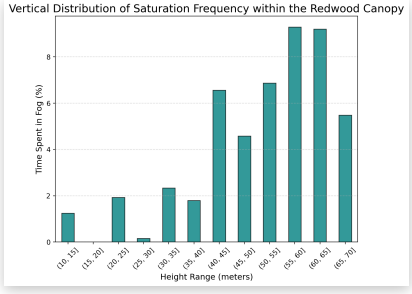

Diurnal Light Availability Across Canopy Heights
Canopy Height Level
Lower canopy
Mid canopy
Upper canopy
Incident Light at Top of Canopy
Time of Day (Hours)

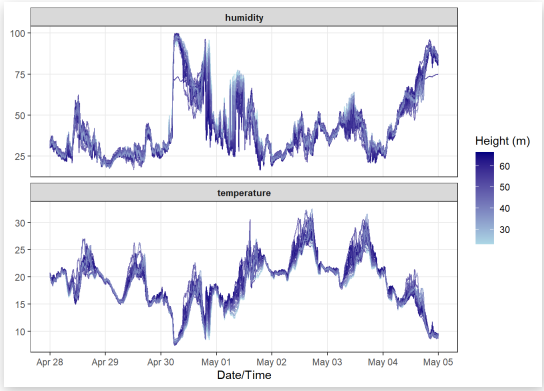

Time and Temperature Readings
Temperature (Celsius)
May 1st 08:00
May 1st 12:00
May 1st 16:00
May 1st 20:00
Date & Time of Reading

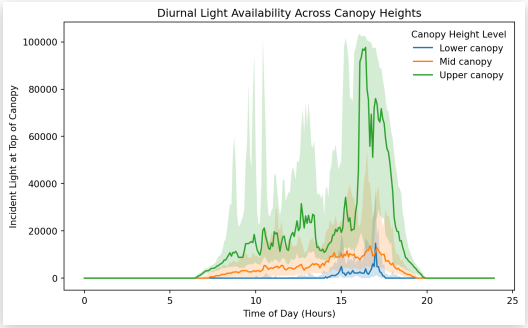

90th Percentile PAR by Height Group
90th Percentile PAR (µmol/m²/s)
10-15m
15-20m
20-25m
25-30m
30-35m
35-40m
40-45m
45-50m
50-55m
55-60m
60-65m
65-70m
Height Group

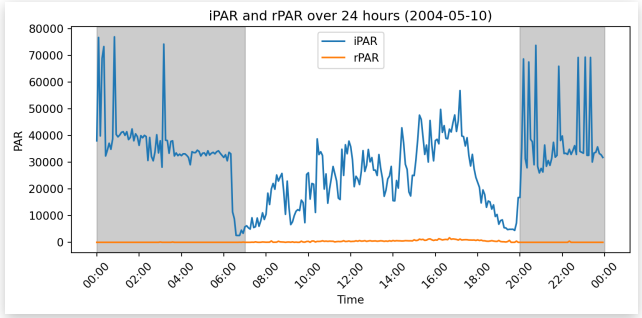

Diurnal humidity cycles vary with canopy height
Lower canopy maintains more stable humidity across the day
Canopy position
Lower canopy (0–20 m)
Mid canopy (20–40 m)
Upper canopy (40–60 m)
Top canopy
Mean relative humidity (%)
Hour of day

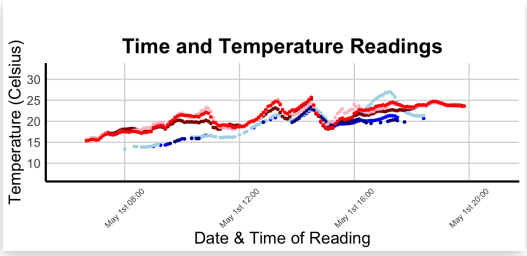

Daily Canopy Stratification: Temperature Anomaly Across Height and Time
Height in Canopy (m)
Local Time of Day (hours)
Temperature Anomaly (°C)
(relative to each node's median)

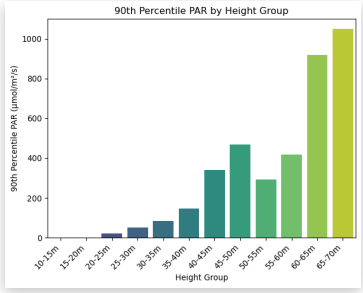

Vertical Distribution of Saturation Frequency within the Redwood Canopy
Time Spent in Fog (%)
Height Range (meters)

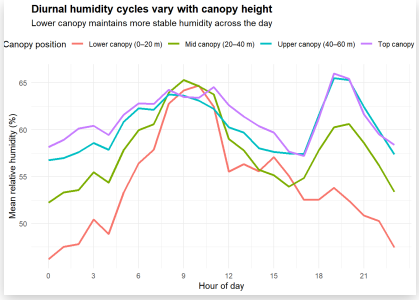

humidity
temperature
Height (m)
60
50
40
30
Apr 28
Apr 29
Apr 30
May 01
May 02
May 03
May 04
May 05
Date/Time

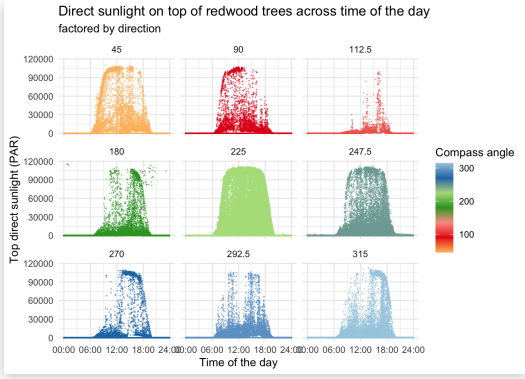

Diurnal Light Availability Across Canopy Heights
Canopy Height Level
Lower canopy
Mid canopy
Upper canopy
Incident Light at Top of Canopy
Time of Day (Hours)

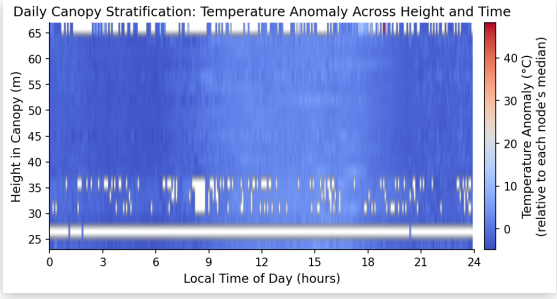

iPAR and rPAR over 24 hours (2004-05-10)
iPAR
rPAR
PAR
Time

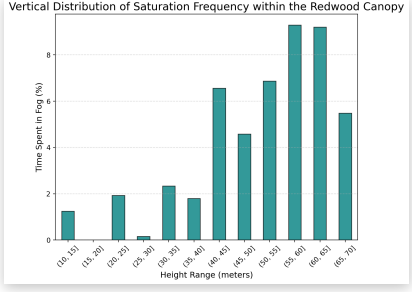

Time and Temperature Readings
Temperature (Celsius)
May 1st 08:00
May 1st 12:00
May 1st 16:00
May 1st 20:00
Date & Time of Reading

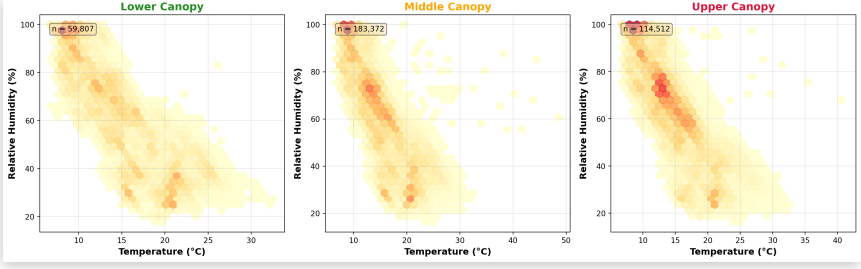

90th Percentile PAR by Height Group
90th Percentile PAR (µmol/m²/s)
10-15m
15-20m
20-25m
25-30m
30-35m
35-40m
40-45m
45-50m
50-55m
55-60m
60-65m
65-70m
Height Group

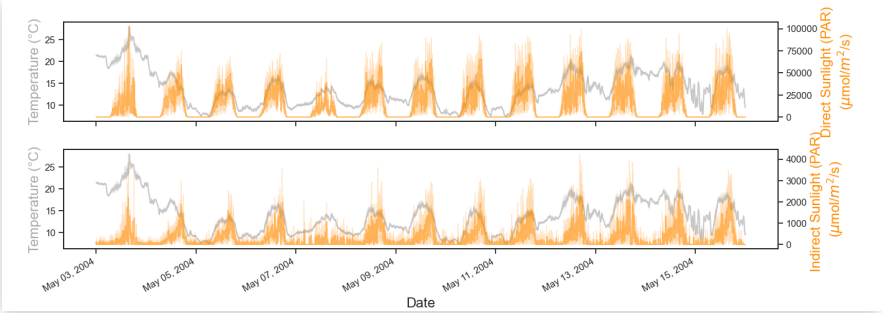

Diurnal humidity cycles vary with canopy height
Lower canopy maintains more stable humidity across the day
Canopy position
Lower canopy (0–20 m)
Mid canopy (20–40 m)
Upper canopy (40–60 m)
Top canopy
Mean relative humidity (%)
Hour of day

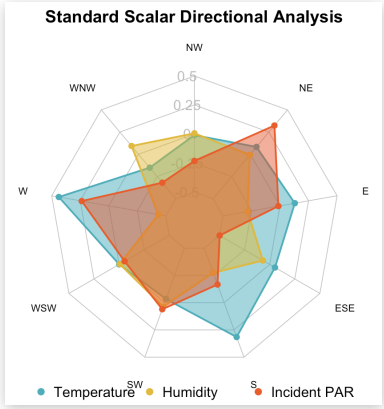

Direct sunlight on top of redwood trees across time of the day
factored by direction
45
90
112.5
180
225
247.5
270
292.5
315
Top direct sunlight (PAR)
Compass angle
Time of the day

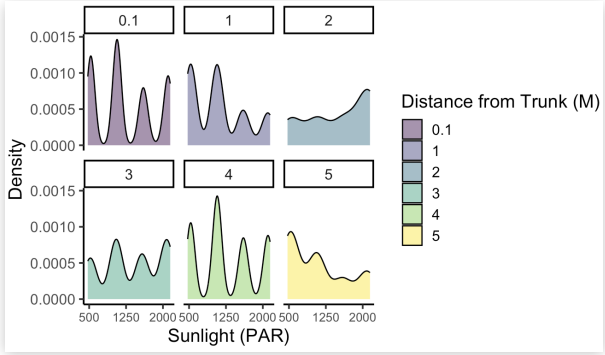

Daily Canopy Stratification: Temperature Anomaly Across Height and Time
Height in Canopy (m)
Local Time of Day (hours)
Temperature Anomaly (°C)
(relative to each node's median)

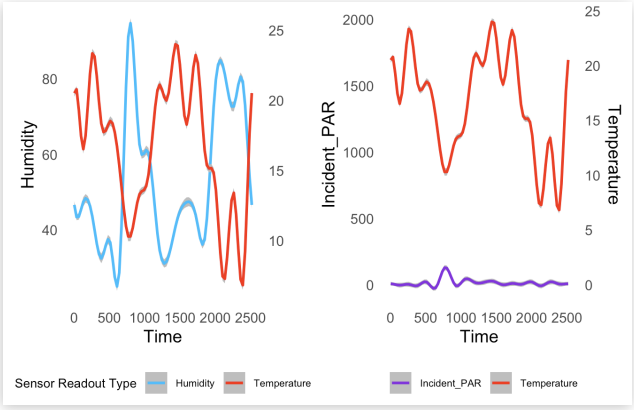

Vertical Distribution of Saturation Frequency within the Redwood Canopy
Time Spent in Fog (%)
Height Range (meters)

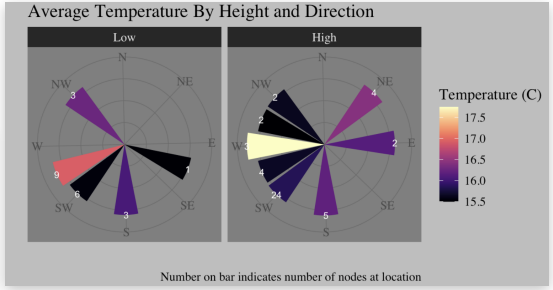

Lower Canopy
Middle Canopy
Upper Canopy
n = 59,807
n = 183,372
n = 114,512
Relative Humidity (%)
Temperature (°C)

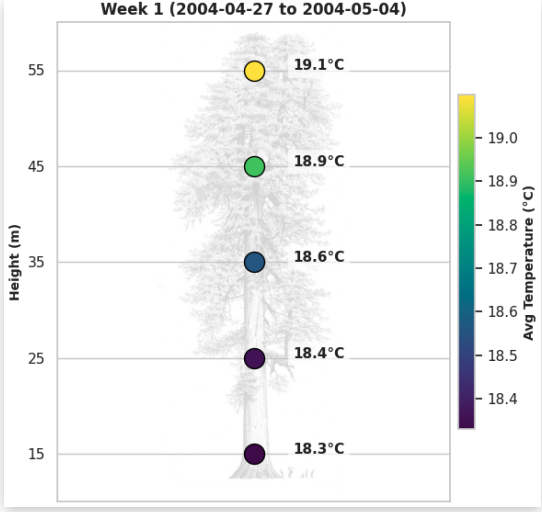

Temperature (°C)
Direct Sunlight (PAR) (μmol/m²/s)
Indirect Sunlight (PAR) (μmol/m²/s)
May 03, 2004
May 05, 2004
May 07, 2004
May 09, 2004
May 11, 2004
May 13, 2004
May 15, 2004
Date

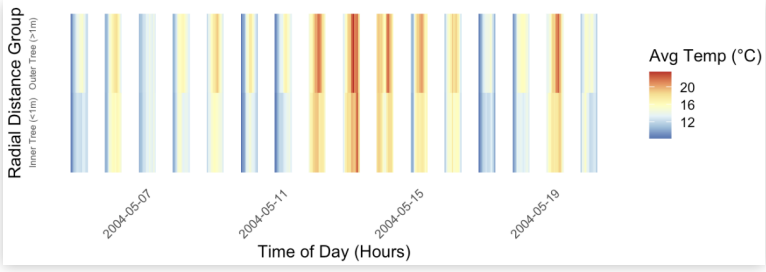

Standard Scalar Directional Analysis
NW
NE
WNW
W
E
WSW
ESE
SW
S
Temperature
Humidity
Incident PAR

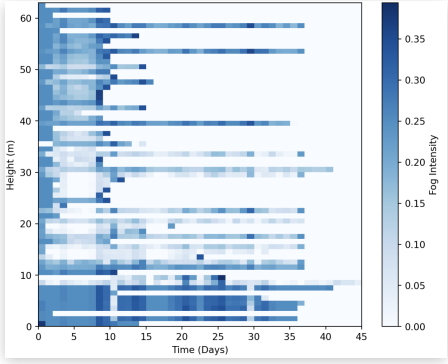

0.1
1
2
3
4
5
Density
Sunlight (PAR)
Distance from Trunk (M)


Humidity
Incident_PAR
Temperature
Time
Sensor Readout Type
Humidity
Temperature
Incident_PAR
Temperature


Average Temperature By Height and Direction
Low
High
Temperature (C)
17.5
17.0
16.5
16.0
15.5
Number on bar indicates number of nodes at location


Week 1 (2004-04-27 to 2004-05-04)
19.1°C
18.9°C
18.6°C
18.4°C
18.3°C
Height (m)
Avg Temperature (°C)


Radial Distance Group
Inner Tree (<1m)
Outer Tree (>1m)
Avg Temp (°C)
2004-05-07
2004-05-11
2004-05-15
2004-05-19
Time of Day (Hours)


Height (m)
Time (Days)
Fog Intensity

Quiz

Question 1

What word(s) do you use to address a group of two or more people?

Question 1

What word(s) do you use to address a group of two or more people?

A. You

B. You guys

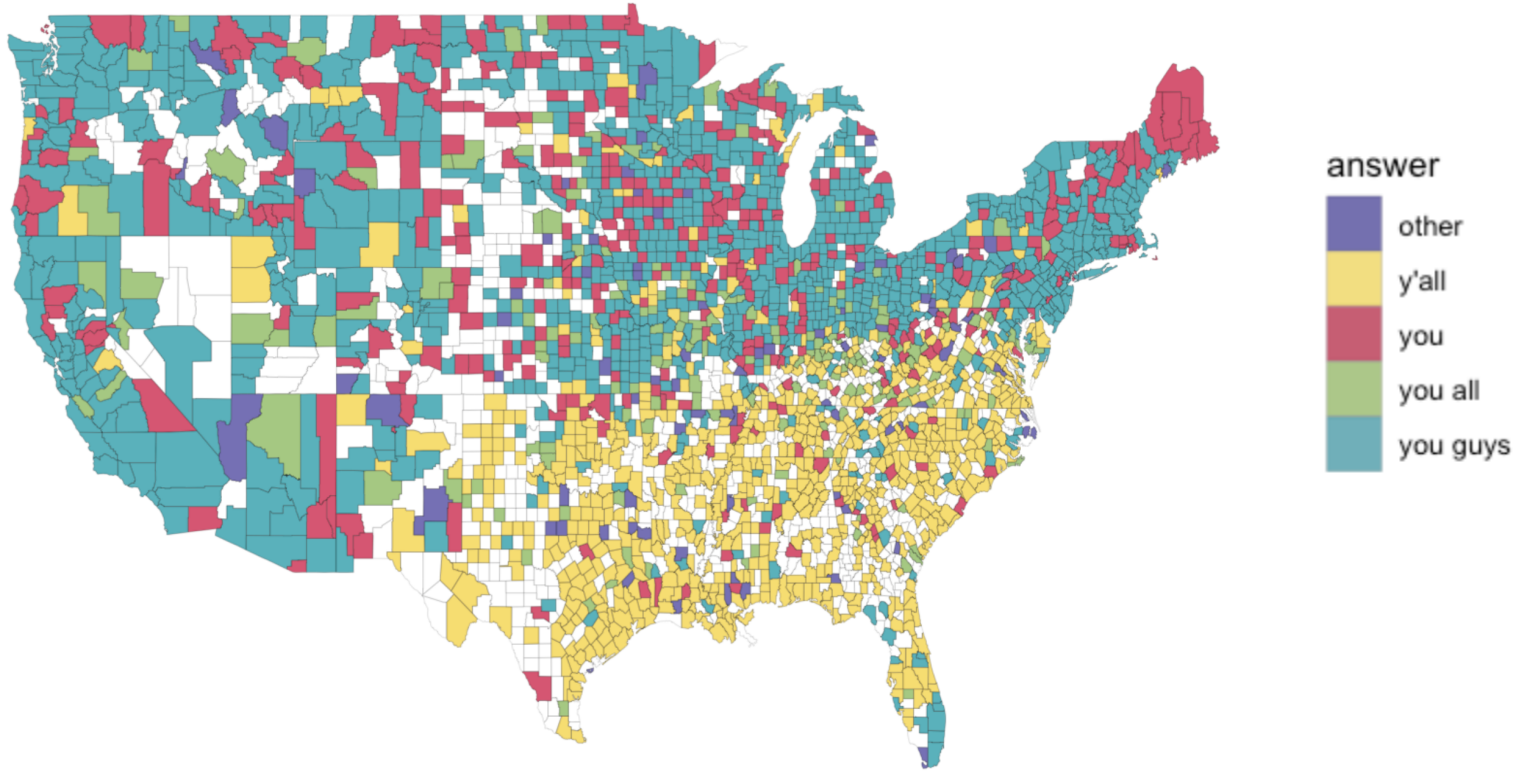
C. You all

D. Y'all

E. Other

Question 1

What word(s) do you use to address a group of two or more people?



Question 2

What do you call the insect that flies around in the summer and has a rear section that glows in the dark?

Question 2

What do you call the insect that flies around in the summer and has a rear section that glows in the dark?

A. Firefly

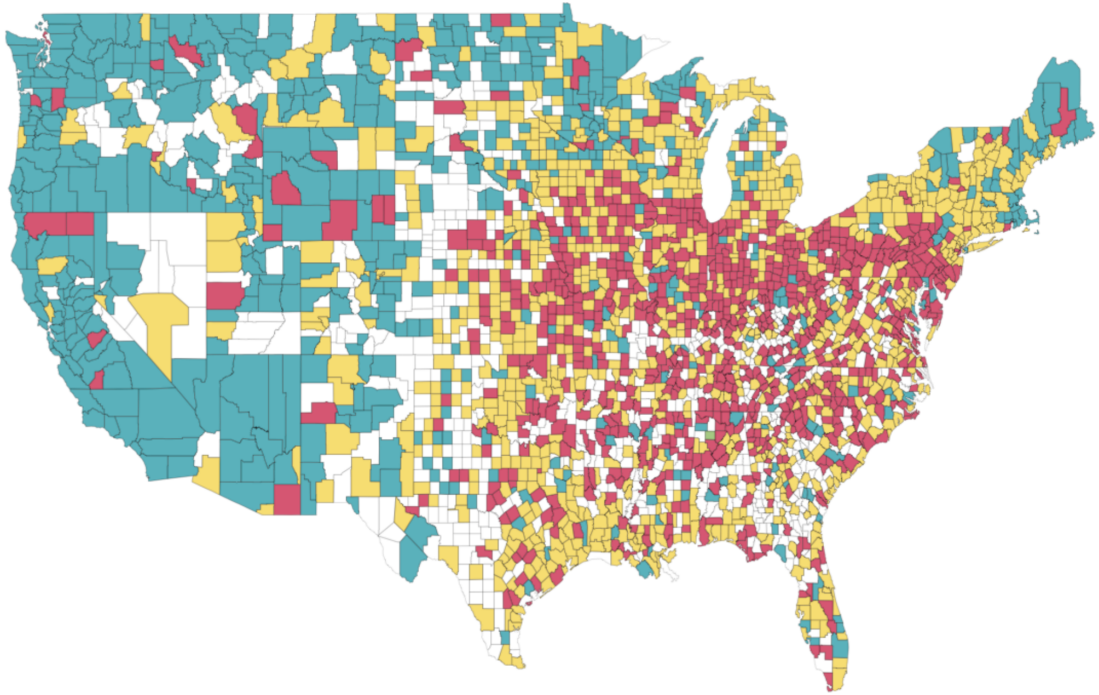
B. Lightning bug

C. I use lightning bug and firefly interchangeably

D. Other

Question 2

What do you call the insect that flies around in the summer and has a rear section that glows in the dark?



answer
- firefly
- I use lightning bug and firefly interchangeably
- lightning bug
- other

Question 3

What do you call the miniature lobster that one finds in lakes and streams for example (a crustacean of the family Astacidae)?

Question 3

What do you call the miniature lobster that one finds in lakes and streams for example (a crustacean of the family Astacidae)?

A. Crawdad

B. Crawfish

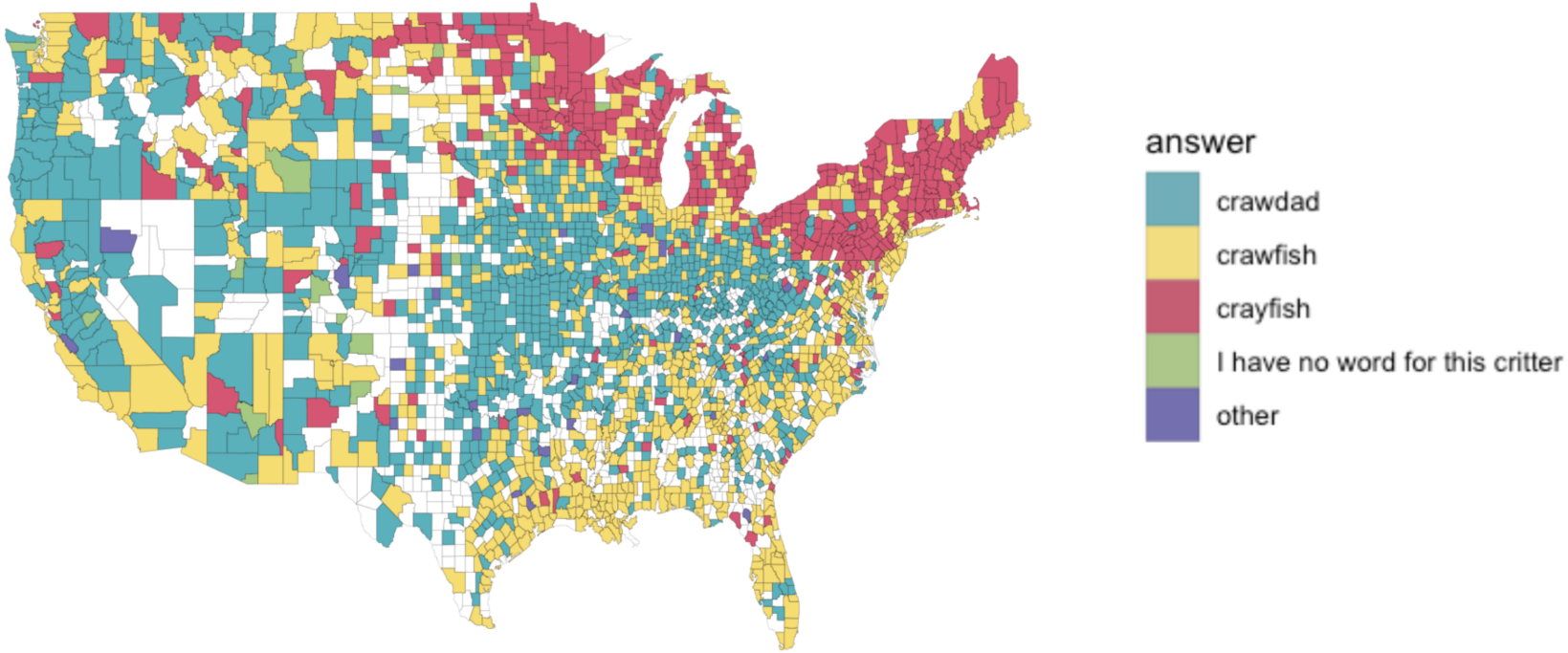
C. Crayfish

D. I have no word for this

E. Other

Question 3

What do you call the miniature lobster that one finds in lakes and streams for example (a crustacean of the family Astacidae)?



Question 4

What is your general term for the rubber-soled shoes worn in gym class, for athletic activities, etc.?

Question 4

What is your general term for the rubber-soled shoes worn in gym class, for athletic activities, etc.?

A. Sneakers

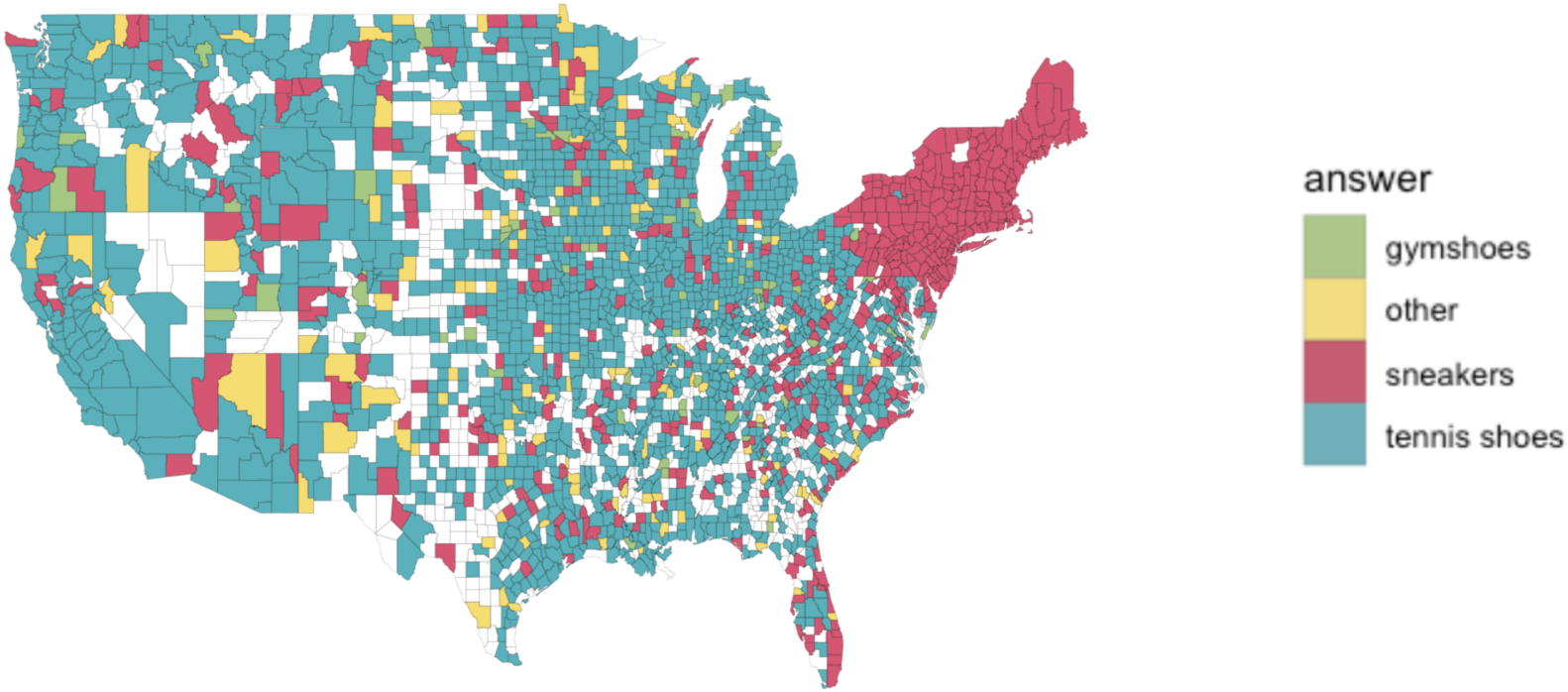
B. Tennis Shoes

C. Gym Shoes

D. Other (running shoes, shoes, ...)

Question 4

What is your general term for the rubber-soled shoes worn in gym class, for athletic activities, etc.?



Question 5

What do you call it when rain falls while the sun is shining?

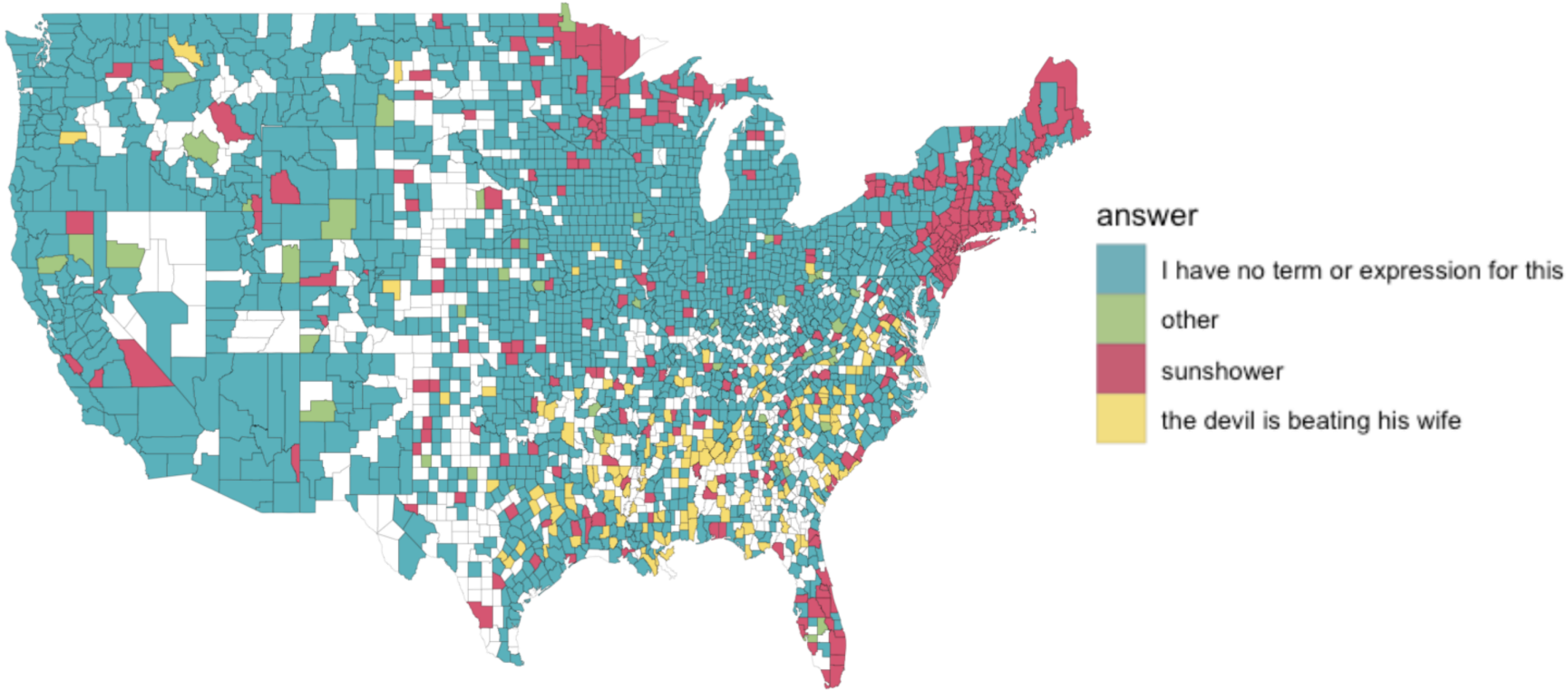
Question 5

What do you call it when rain falls while the sun is shining?

A. I have no term or expression for this
B. Sunshower
C. The devil is beating his wife
D. Other

Question 5

What do you call it when rain falls while the sun is shining?



Question 6

What do you call a traffic situation in which several roads meet in a circle and you have to get off at a certain point?

Question 6

What do you call a traffic situation in which several roads meet in a circle and you have to get off at a certain point?

A. Rotary

B. Roundabout

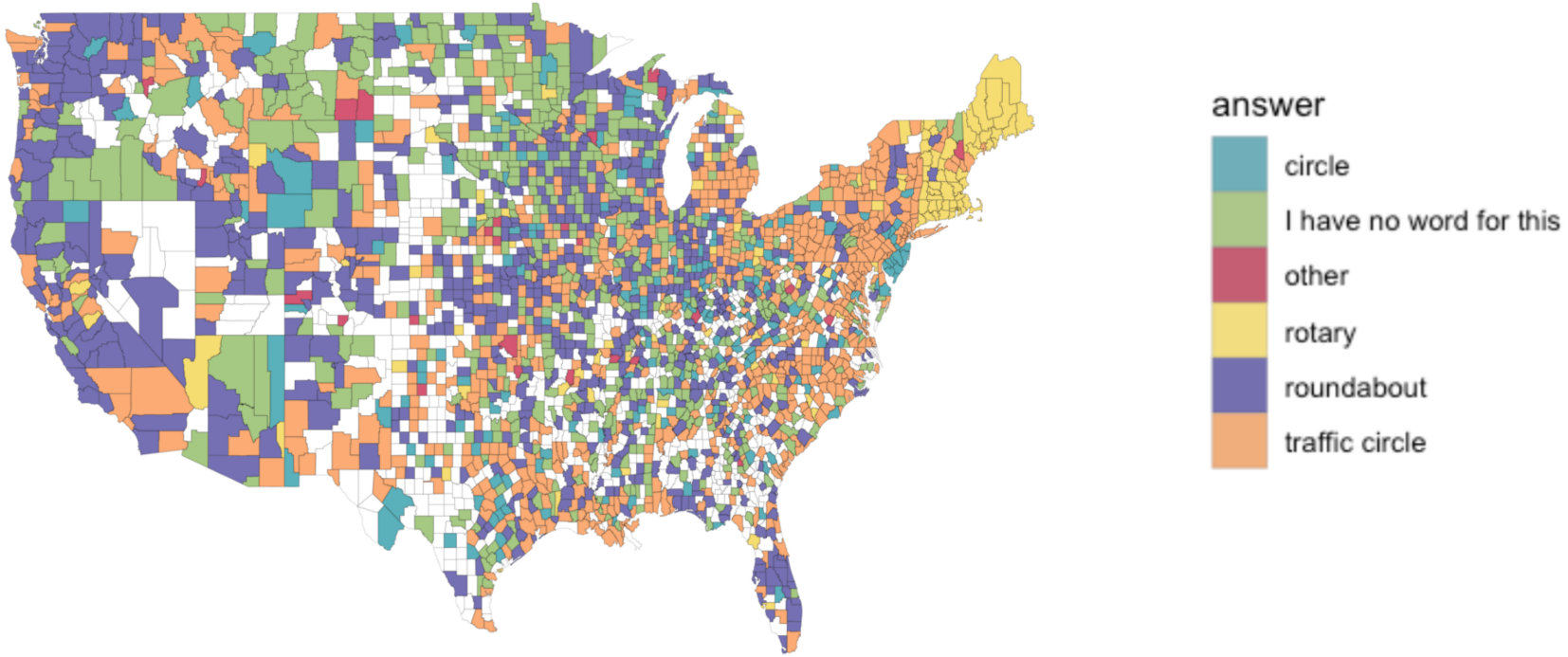
C. Traffic circle

D. Circle

E. Other/I have no word for this

Question 6

What do you call a traffic situation in which several roads meet in a circle and you have to get off at a certain point?



Question 7

Which of these terms do you prefer?

Question 7

Which of these terms do you prefer?

A. Trash can

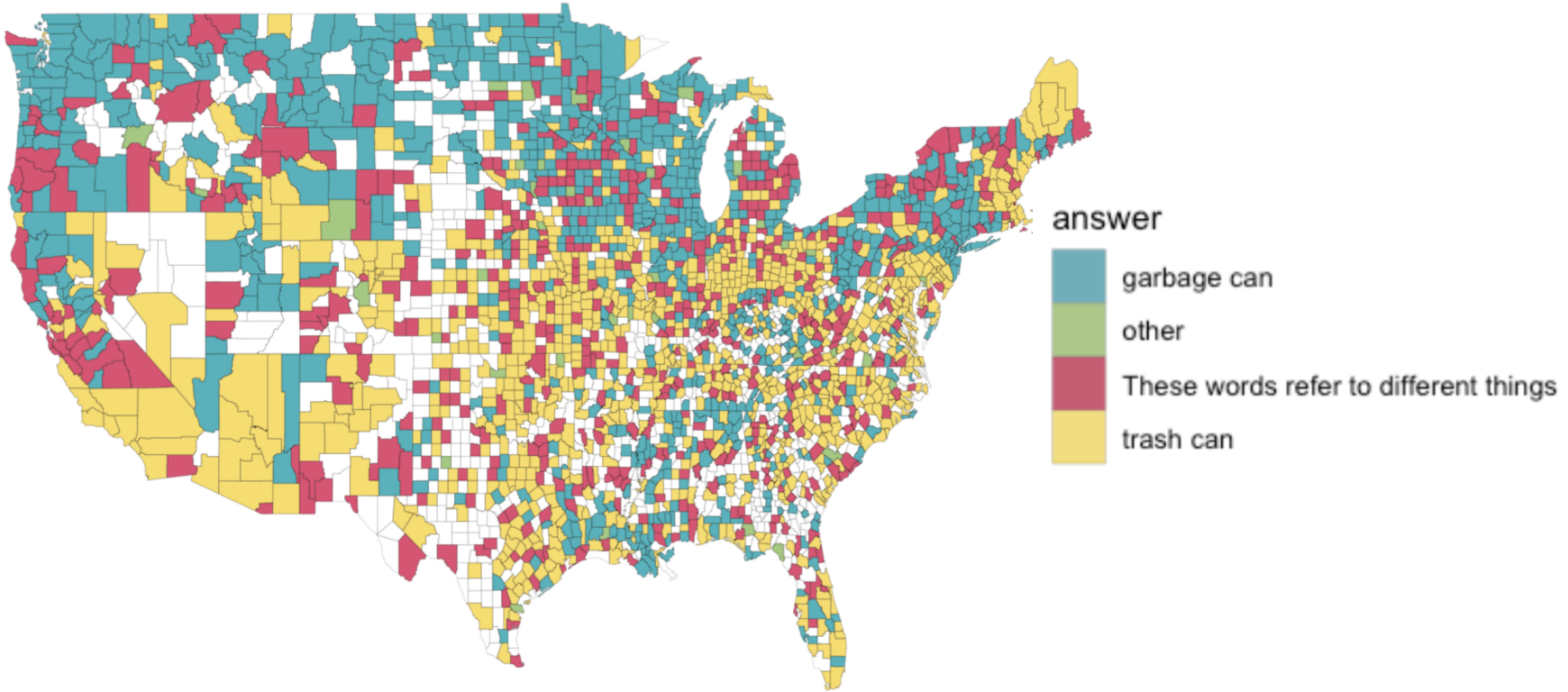
B. Garbage can

C. These words refer to different things

D. Other

Question 7

Which of these terms do you prefer?



Question 8

What do you call the thing from which you might drink water in a school?

Question 8

What do you call the thing from which you might drink water in a school?

A. Drinking fountain

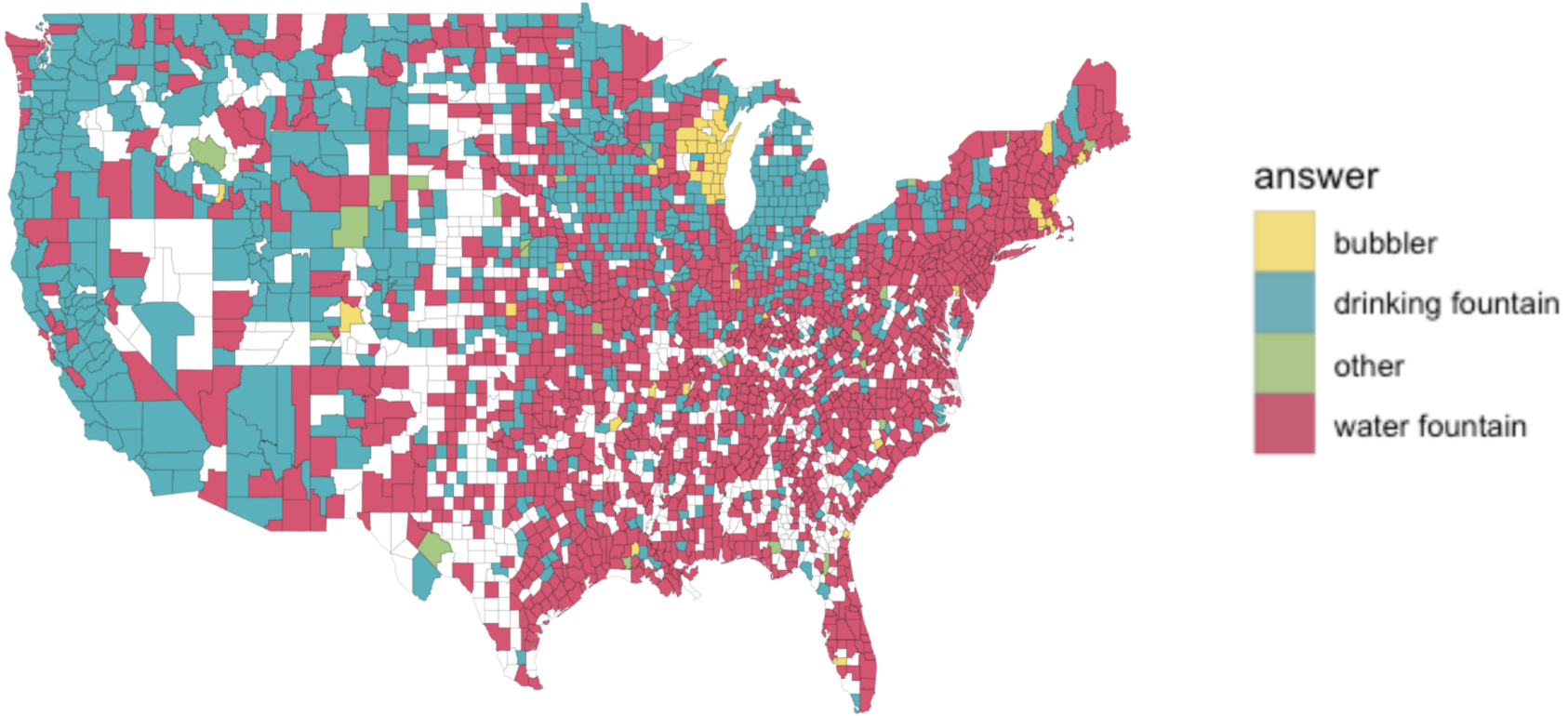
B. Water fountain

C. Bubbler

D. Other

Question 8

What do you call the thing from which you might drink water in a school?



Question 9

What is your generic term for a sweetened carbonated beverage?

Question 9

What is your generic term for a sweetened carbonated beverage?

A. Coke

B. Pop

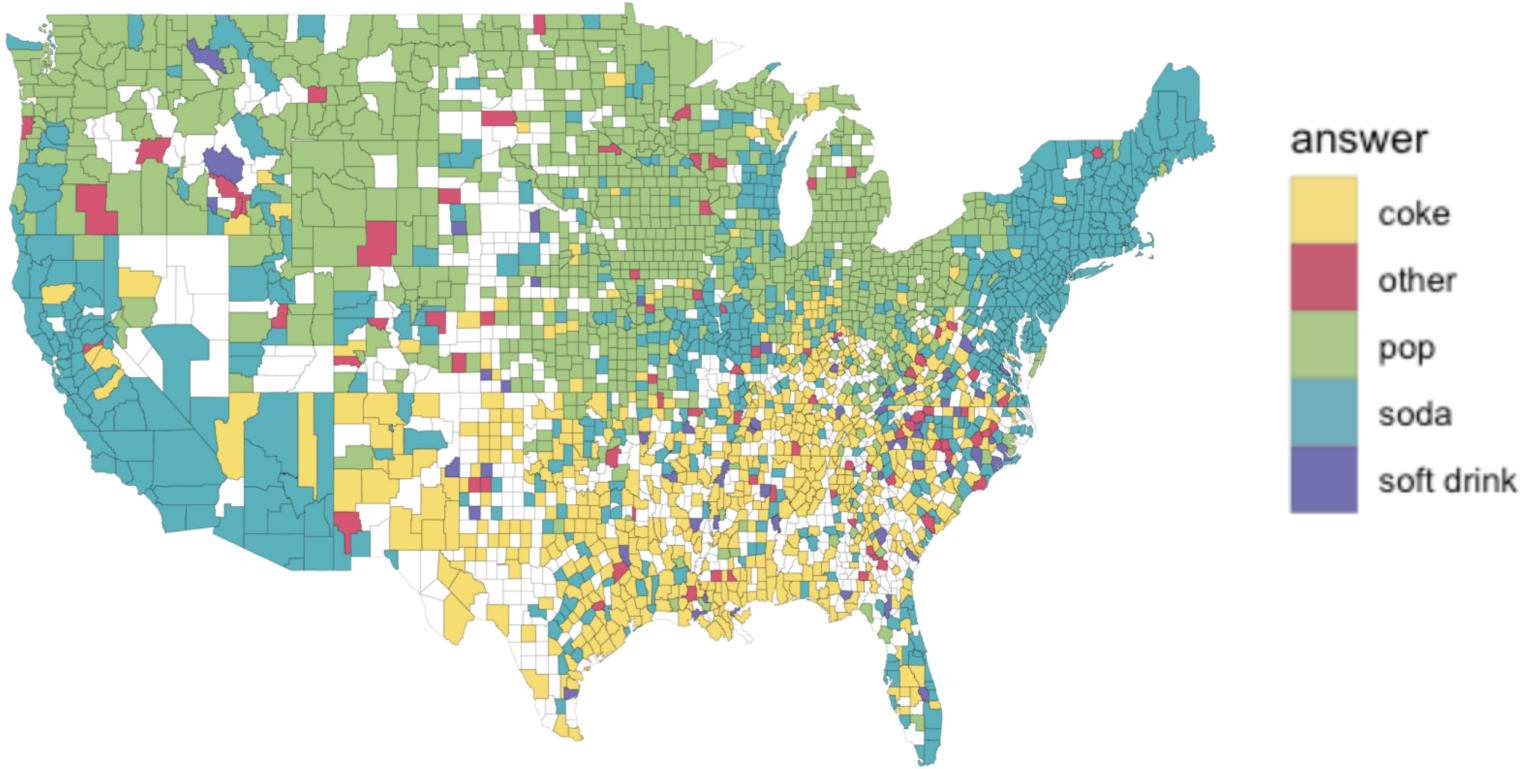
C. Soda

D. Soft drink

E. Other

Question 9

What is your generic term for a sweetened carbonated beverage?



Question 10

What do you call the night before Halloween?

Question 10

What do you call the night before Halloween?

A. I have no word for this

B. Cabbage night

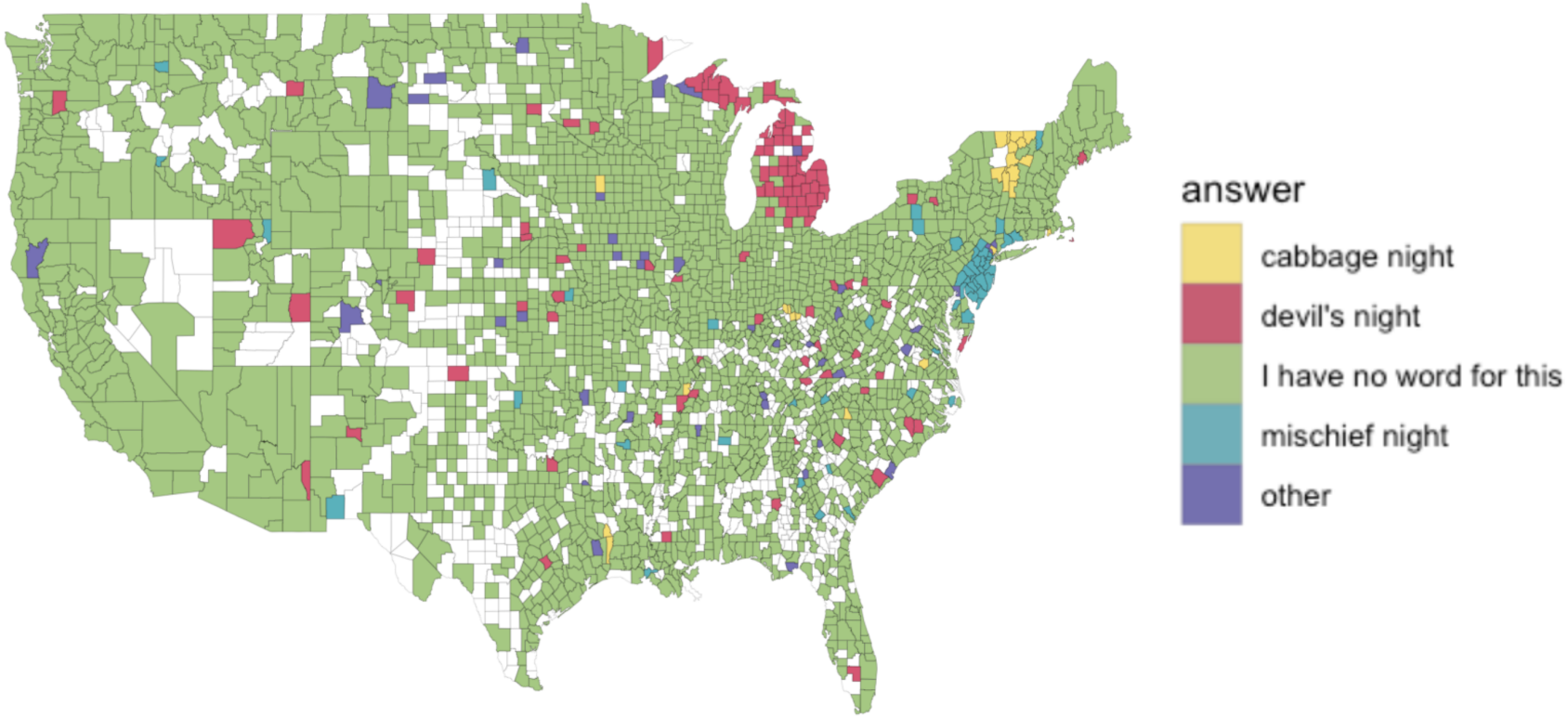
C. Devil's night

D. Mischief night

E. Other

Question 10

What do you call the night before Halloween?



End of quiz!

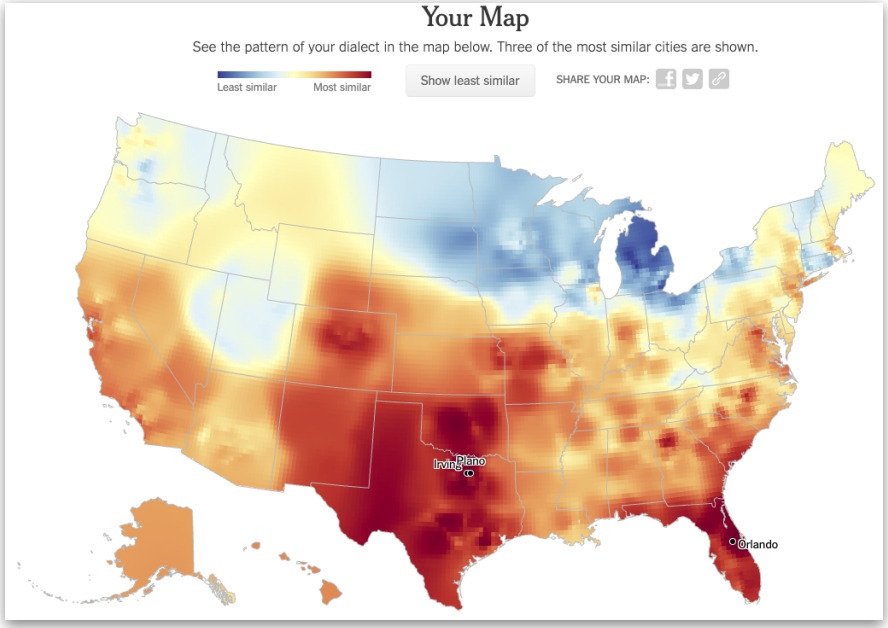
Linguistics quiz discussion

- Did your personal responses agree with the other respondents from where you are from?
- Was there another classmate who tended to answer similarly as you? Are you from similar regions?

New York Times Dialect Quiz* [Josh Katz and Wilson Andrews]

Goal: guess where you are from based upon 25 questions

https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html



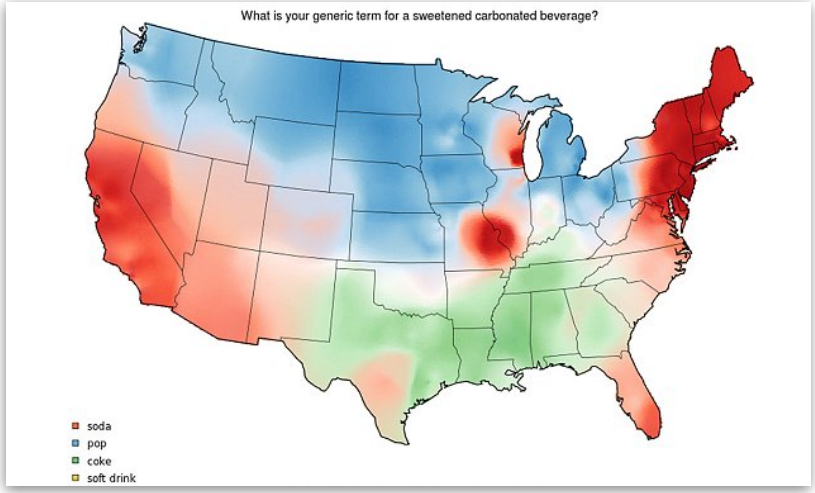
* won a peabody award

Linguistics Data Case Study

Introduction to Linguistics Case Study

Background: Variation in languages/dialects can provide insights into historical, social, and geographical factors in society

Domain question: Are there geographical regions in the US with distinctive dialects? Can we find clusters of people who speak similarly in the US?



Linguistics Data Collection

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

- Survey contains a series of questions: do you say *a*, *b*, or *c*?

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

- Survey contains a series of questions: do you say *a*, *b*, or *c*?

A subset of the original Harvard Dialect survey data can be found on Canvas

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

- Survey contains a series of questions: do you say *a*, *b*, or *c*?

A subset of the original Harvard Dialect survey data can be found on Canvas

- n = 47,471 respondents

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

- Survey contains a series of questions: do you say *a*, *b*, or *c*?

A subset of the original Harvard Dialect survey data can be found on Canvas

- n = 47,471 respondents
- p = 67 questions of interest

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

- Survey contains a series of questions: do you say *a*, *b*, or *c*?

A subset of the original Harvard Dialect survey data can be found on Canvas

- n = 47,471 respondents
- p = 67 questions of interest
- Also have respondents' geographical location, encoded by city, state, ZIP

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

- Survey contains a series of questions: do you say *a*, *b*, or *c*?

A subset of the original Harvard Dialect survey data can be found on Canvas

- n = 47,471 respondents
- p = 67 questions of interest
- Also have respondents' geographical location, encoded by city, state, ZIP

What are we going to do?

Linguistics Data Collection

Data: original Harvard Dialect survey conducted by Bert Vaux

- Survey contains a series of questions: do you say *a*, *b*, or *c*?

A subset of the original Harvard Dialect survey data can be found on Canvas

- n = 47,471 respondents
- p = 67 questions of interest
- Also have respondents' geographical location, encoded by city, state, ZIP

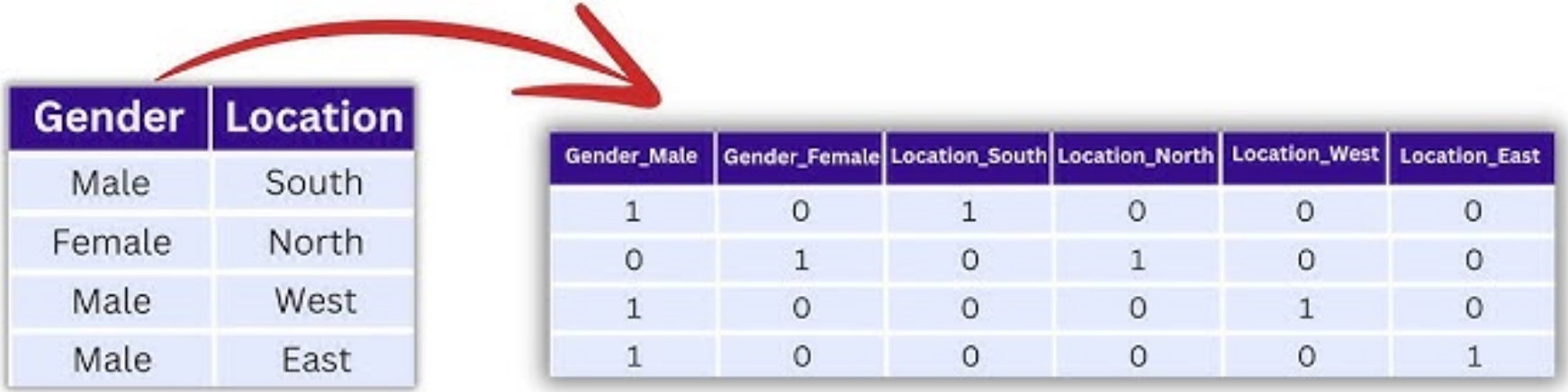
What are we going to do?

- Perform dimension reduction + clustering **to identify regional dialects in the US**

Linguistics Data Case Study

1. Go to our `dsip-s26/` course repository and `git pull`
    - This should have pulled a new `course_materials/linguistics/` folder
2. Restore renv/conda environment
    - **R:**
        1. Either open `linguistics.Rproj` in RStudio or open `linguistics/` folder in Positron
        2. In R console: run `renv::restore()`
        3. Open `linguistics_R.qmd (or linguistics_R.html)`
    - **Python:**
        1. Open `linguistics/` folder in Positron or VS Code
        2. In terminal: run conda-lock install —name=dsip_linguistics
            - Be sure that your current working directory is `dsip-s26/course_materials/linguistics/`
        3. Open `linguistics_python.qmd (or linguistics_python.html)`

One-hot Encoding

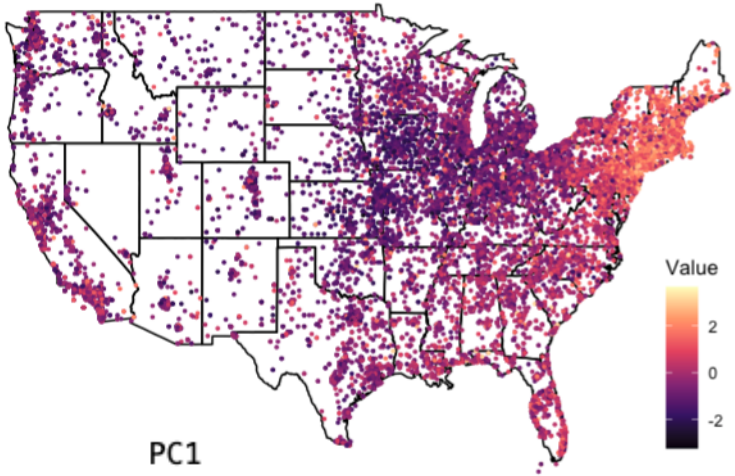


Gender	Location
Male	South
Female	North
Male	West
Male	East

Gender_Male	Gender_Female	Location_South	Location_North	Location_West	Location_East
1	0	1	0	0	0
0	1	0	1	0	0
1	0	0	0	1	0
1	0	0	0	0	1

Interpreting PCA

Component 1

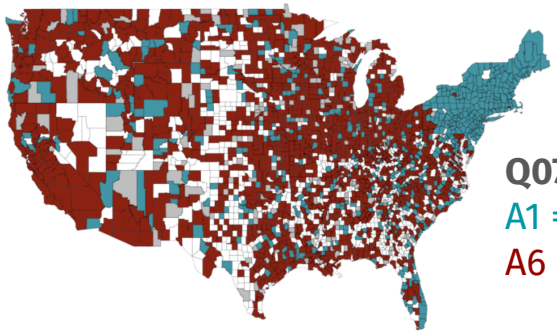


PC1

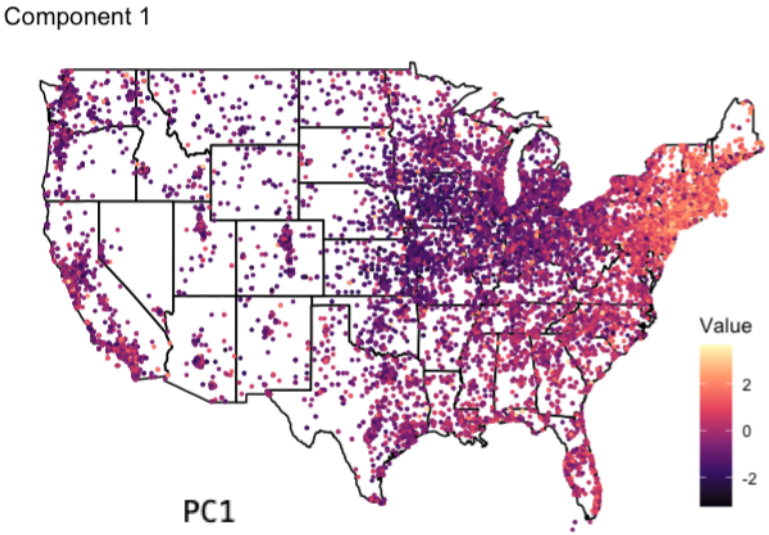
Q073.1 (0.27)
Q073.6 (−0.24)
Q105.1 (0.20)
Q080.1 (0.18)
Q080.8 (−0.18)
Q105.2 (−0.17)
⋮

PC1 = 0.27 × Q073.1 - 0.24 × Q073.6
+ 0.2 × Q105.1 - 0.17 × Q105.2
+ 0.18 × Q080.1 - 0.18 × Q080.8
+ …

Interpreting PCA



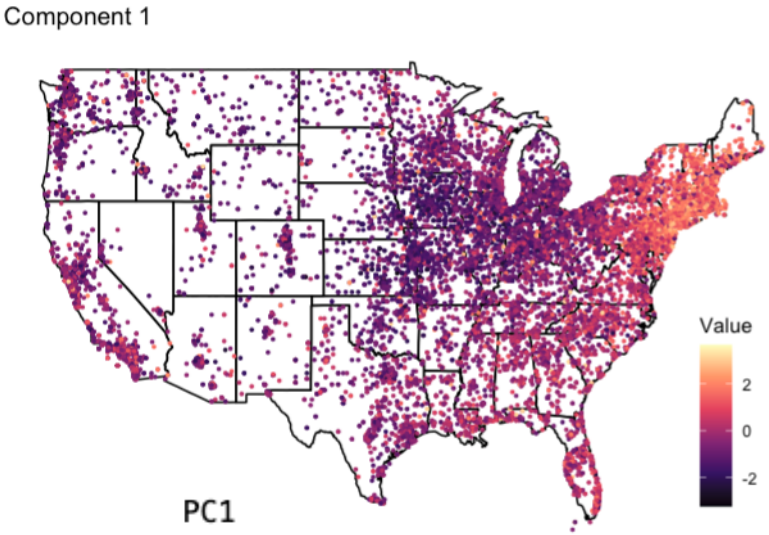
Q073: "Shoes" question
A1 = sneakers
A6 = tennis shoes



PC1
Q073.1 (0.27)
Q073.6 (−0.24)
Q105.1 (0.20)
Q080.1 (0.18)
Q080.8 (−0.18)
Q105.2 (−0.17)
⋮

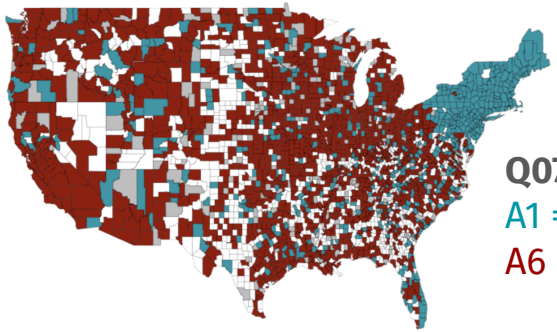
PC1 = 0.27 × Q073.1 - 0.24 × Q073.6
+ 0.2 × Q105.1 - 0.17 × Q105.2
+ 0.18 × Q080.1 - 0.18 × Q080.8
+ …

Interpreting PCA

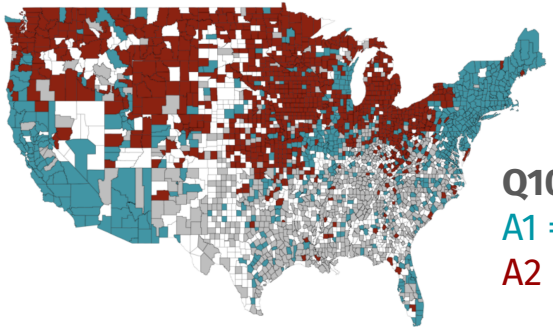


PC1
Q073.1 (0.27)
Q073.6 (−0.24)
Q105.1 (0.20)
Q080.1 (0.18)
Q080.8 (−0.18)
Q105.2 (−0.17)
⋮

PC1 = 0.27 × Q073.1 - 0.24 × Q073.6
+ 0.2 × Q105.1 - 0.17 × Q105.2
+ 0.18 × Q080.1 - 0.18 × Q080.8
+ …

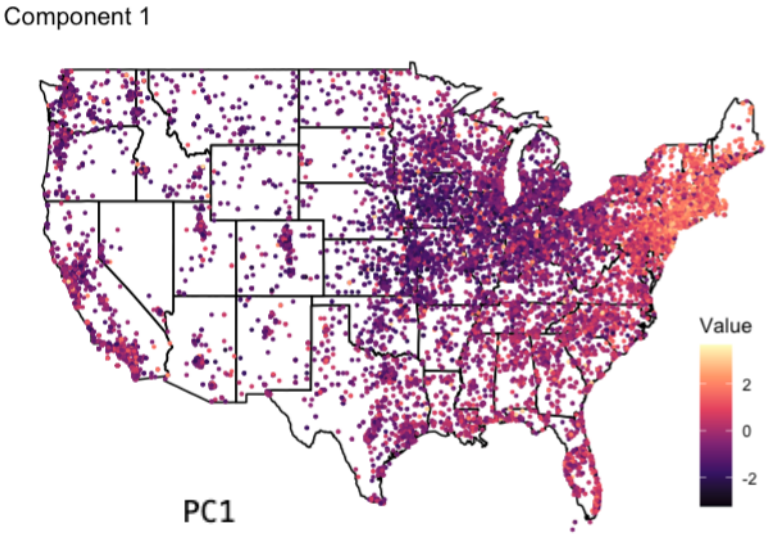


Q073: "Shoes" question
A1 = sneakers
A6 = tennis shoes



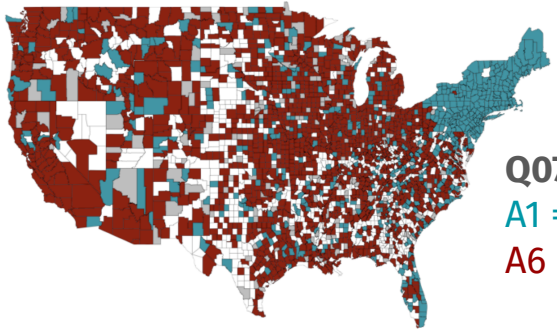
Q105: "Soda" question
A1 = soda
A2 = pop

Interpreting PCA

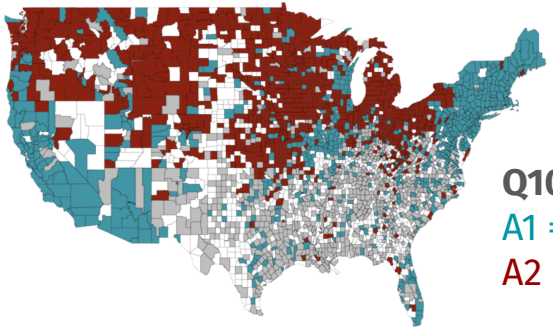


PC1
Q073.1 (0.27)
Q073.6 (−0.24)
Q105.1 (0.20)
Q080.1 (0.18)
Q080.8 (−0.18)
Q105.2 (−0.17)
⋮

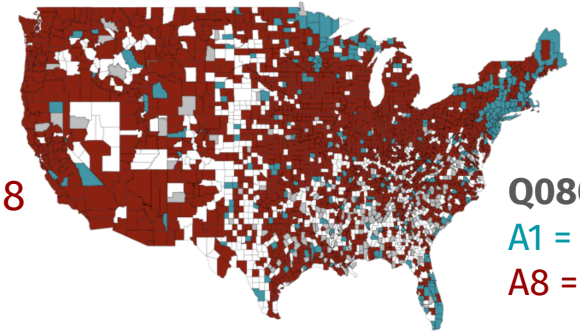
PC1 = 0.27 × Q073.1 - 0.24 × Q073.6
+ 0.2 × Q105.1 - 0.17 × Q105.2
+ 0.18 × Q080.1 - 0.18 × Q080.8
+ …



Q073: "Shoes" question
A1 = sneakers
A6 = tennis shoes

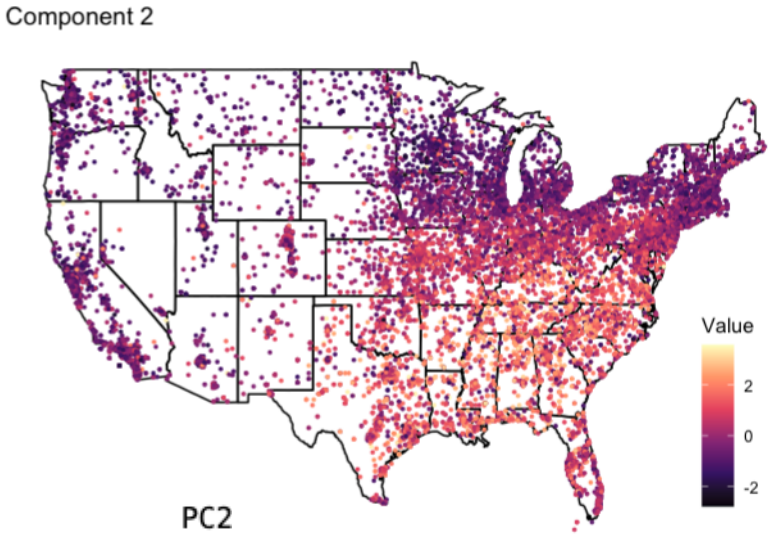


Q105: "Soda" question
A1 = soda
A2 = pop



Q080: "Sunshower" quest
A1 = sunshower
A8 = no term for this

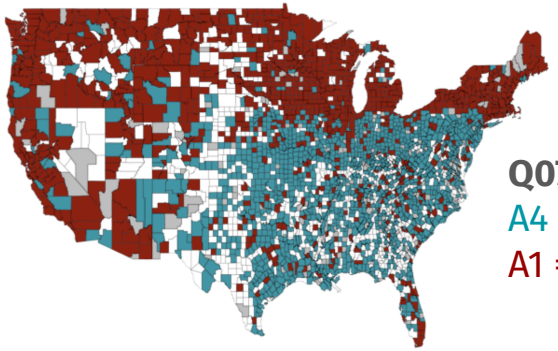
Interpreting PCA



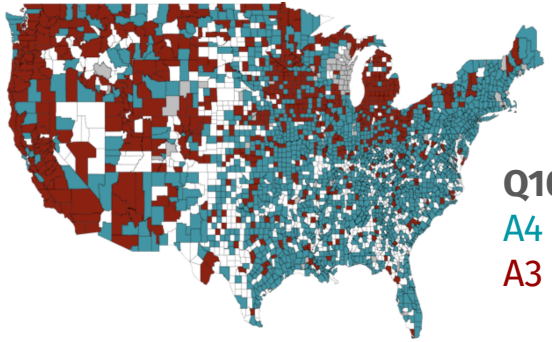
PC2

Q076.1 (−0.25)
Q076.4 (0.24)
Q103.4 (0.22)
Q050.9 (0.19)
Q103.3 (−0.19)
Q071.5 (0.16)
⋮

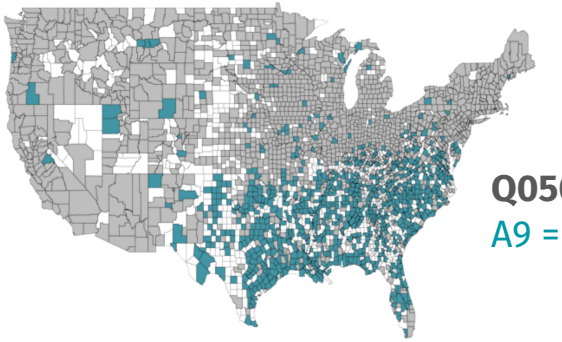
PC2 = 0.24 × Q076.4 - 0.25 × Q076.1
+ 0.22 × Q103.4 - 0.19 × Q103.3
+ 0.19 × Q050.9
+ 0.16 × Q071.5
+ …



Q076:
A4 = catty-corner
A1 = kitty-corner



Q103:
A4 = water fountain
A3 = drinking fountain



Q050:
A9 = y'all

Summary of PCA journey so far

Using PCA, we found weighted linear combinations of questions that correspond to the maximal variation patterns in our data

It happens to be that these directions of maximal variation are associated with geographical regions

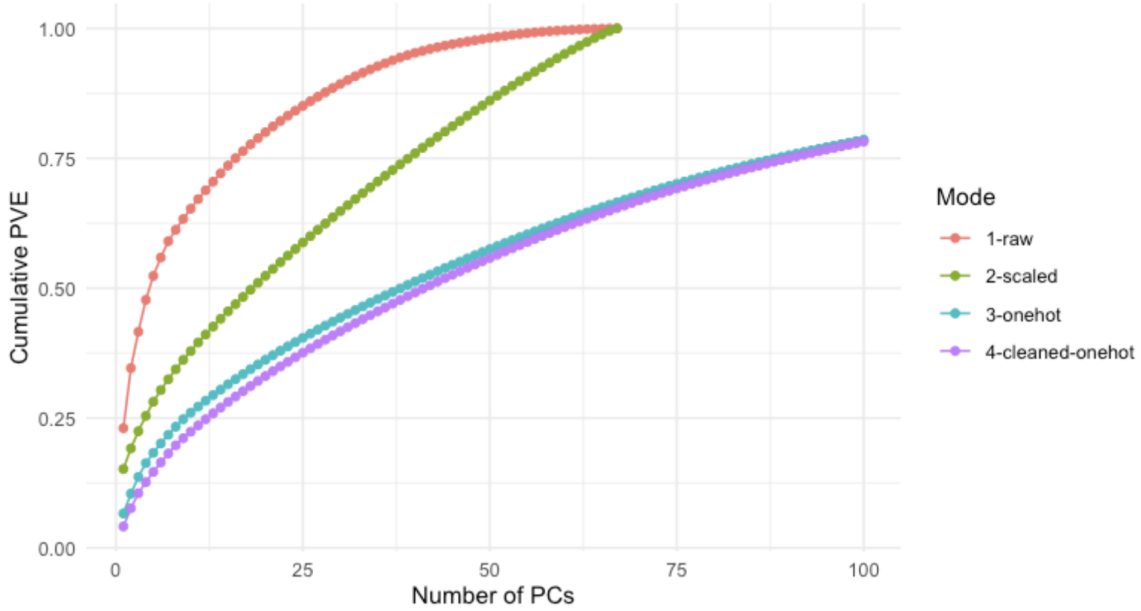
- Importantly, the geographical information was never used when fitting PCA

Moreover, these directions/components are interpretable:

- **PC1** (NE versus the rest): people who say sneakers, soda, sunshower, … versus people who say tennis shoes, pop, no term for sunshower, …
- **PC2** (South versus the rest): people who say catty-corner, water fountain, y'all, … versus people who say kitty-corner, drinking fountain, …

Interpreting PCA

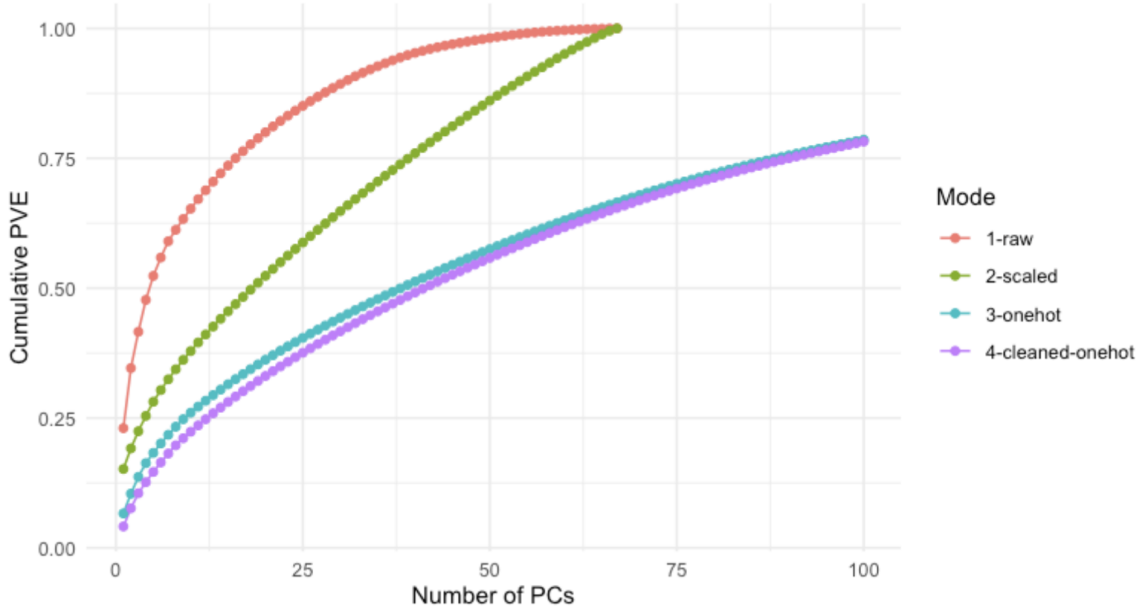
Does the proportion of variance explained (PVE) tell us anything about the "quality" of the PCA results?



Interpreting PCA

Does the proportion of variance explained (PVE) tell us anything about the "quality" of the PCA results? No!

- PC1 from attempt #1 explained the most variance, but this was variance in the data that we did not care about
- PVE is relative to the *total amount of variation* in the data, which differs across datasets (or even the same dataset, scaled differently)



Key Takeaways + Big Picture Lessons for Practice

Key Takeaways + Big Picture Lessons for Practice

- We've finally found a "meaningful" dimension reduction
  - The top PCs can be thought of as a weighted combination of questions that are indicative of the NE (PC1) and the South (PC2)
  - How do we assess "meaningful"? Need to inject some knowledge about the domain problem. Here, we inserted geographical data

Key Takeaways + Big Picture Lessons for Practice

- We've finally found a "meaningful" dimension reduction
  - The top PCs can be thought of as a weighted combination of questions that are indicative of the NE (PC1) and the South (PC2)
  - How do we assess "meaningful"? Need to inject some knowledge about the domain problem. Here, we inserted geographical data
- Our progress thus far had been made possible because:
  - Our understanding of PCA
  - Interpretability of PCA (look at the PC loadings!)

Key Takeaways + Big Picture Lessons for Practice

- We've finally found a "meaningful" dimension reduction
  - The top PCs can be thought of as a weighted combination of questions that are indicative of the NE (PC1) and the South (PC2)
  - How do we assess "meaningful"? Need to inject some knowledge about the domain problem. Here, we inserted geographical data
- Our progress thus far had been made possible because:
  - Our understanding of PCA
  - Interpretability of PCA (look at the PC loadings!)
- This would be more difficult if we had started with tSNE or UMAP

Key Takeaways + Big Picture Lessons for Practice

- We've finally found a "meaningful" dimension reduction
  - The top PCs can be thought of as a weighted combination of questions that are indicative of the NE (PC1) and the South (PC2)
  - How do we assess "meaningful"? Need to inject some knowledge about the domain problem. Here, we inserted geographical data
- Our progress thus far had been made possible because:
  - Our understanding of PCA
  - Interpretability of PCA (look at the PC loadings!)
- This would be more difficult if we had started with tSNE or UMAP
- EDA hinted at these geographical linguistics patterns that we can now summarize more effectively via dimension reduction

Key Takeaways + Big Picture Lessons for Practice

- We've finally found a "meaningful" dimension reduction
  - The top PCs can be thought of as a weighted combination of questions that are indicative of the NE (PC1) and the South (PC2)
  - How do we assess "meaningful"? Need to inject some knowledge about the domain problem. Here, we inserted geographical data
- Our progress thus far had been made possible because:
  - Our understanding of PCA
  - Interpretability of PCA (look at the PC loadings!)
- This would be more difficult if we had started with tSNE or UMAP
- EDA hinted at these geographical linguistics patterns that we can now summarize more effectively via dimension reduction

Start simple, and work our way up!