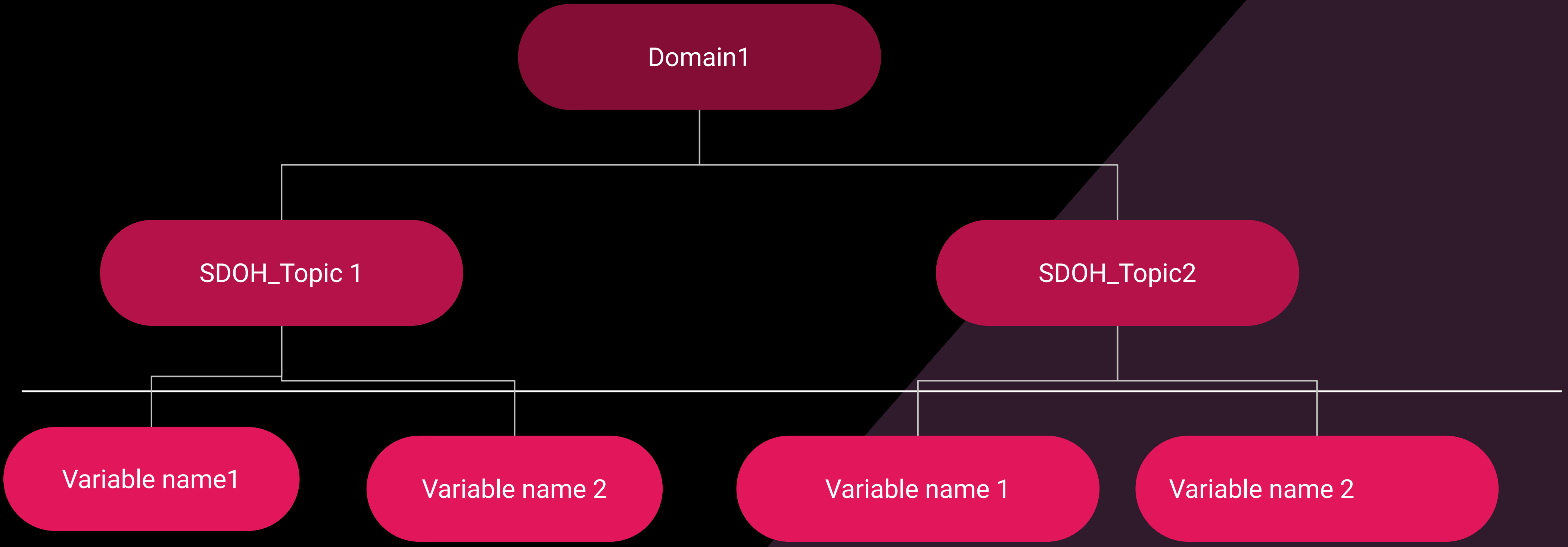
EDA AND PATTERNS OF MISSINGNESS IN SDOH for the EMORY CXR

Team 5 - CXR Dataset

Ali Aslam
Anudeep Errabelly
Enamul Hoq
Zeph Kaffey
David Nyarko
Chiratidzo Sanyika
Veera Venkata Satyavathi
Enzo Ferrante

SOCIAL DETERMINANTS OF HEALTH (SDOH)

- Social factors allocated according to zip code
- 44 sources
- 8 categories (domains)



SO MANY IDEAS

Merge SDOH, metadata, findings

+

Emeddings

But….

- We discovered there are a lot of missing data in the SDOH dataset

We decided to redirect our plan and focused on

Characterizing the fingerprint of “missingness”
in the SDOH Dataset

+

Creating embeddings to allow for easier analysis

Why focus on EDA?

- You cannot start running analysis on a dataset that is not well understood
- Patterns of missingness could lead to underrepresenting some groups and therefore sample bias
- Embeddings allow for easier analysis such as model training, and clustering identifies potential shortcuts for classification models

Hypothesis

- Observe patterns of missingness in the SDOH table for different demographic group (sex and race)
- SDOH (like Gini index) vary strongly for different diseases, races, and sexes
- The embeddings will clusterize by demographics (sex and race) and SDOH

Methods - SDOH

Merged SDOH, metadata and findings and split by demographic

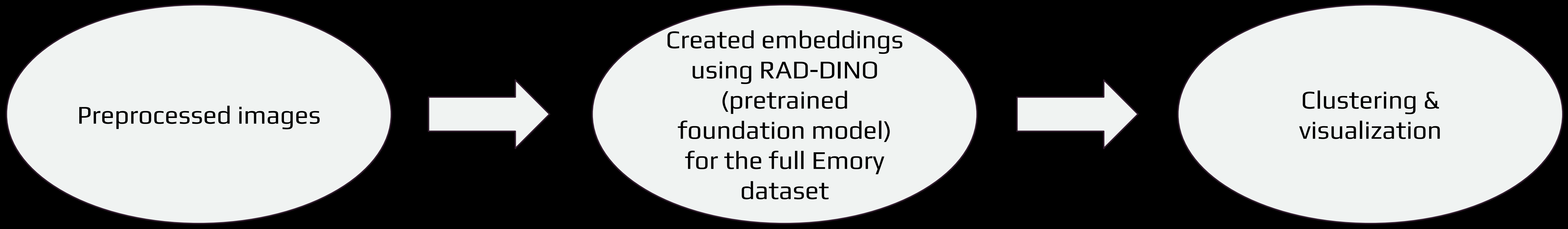
→

% missingness by demographic by variable name

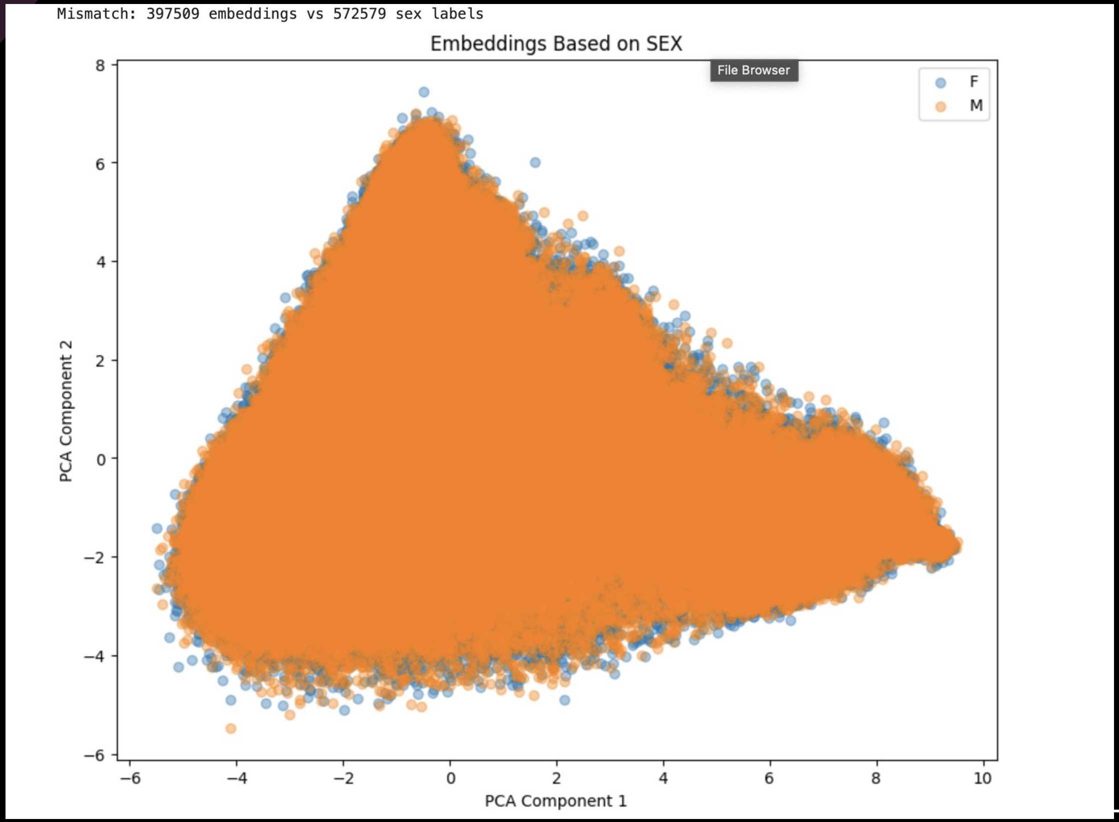
→

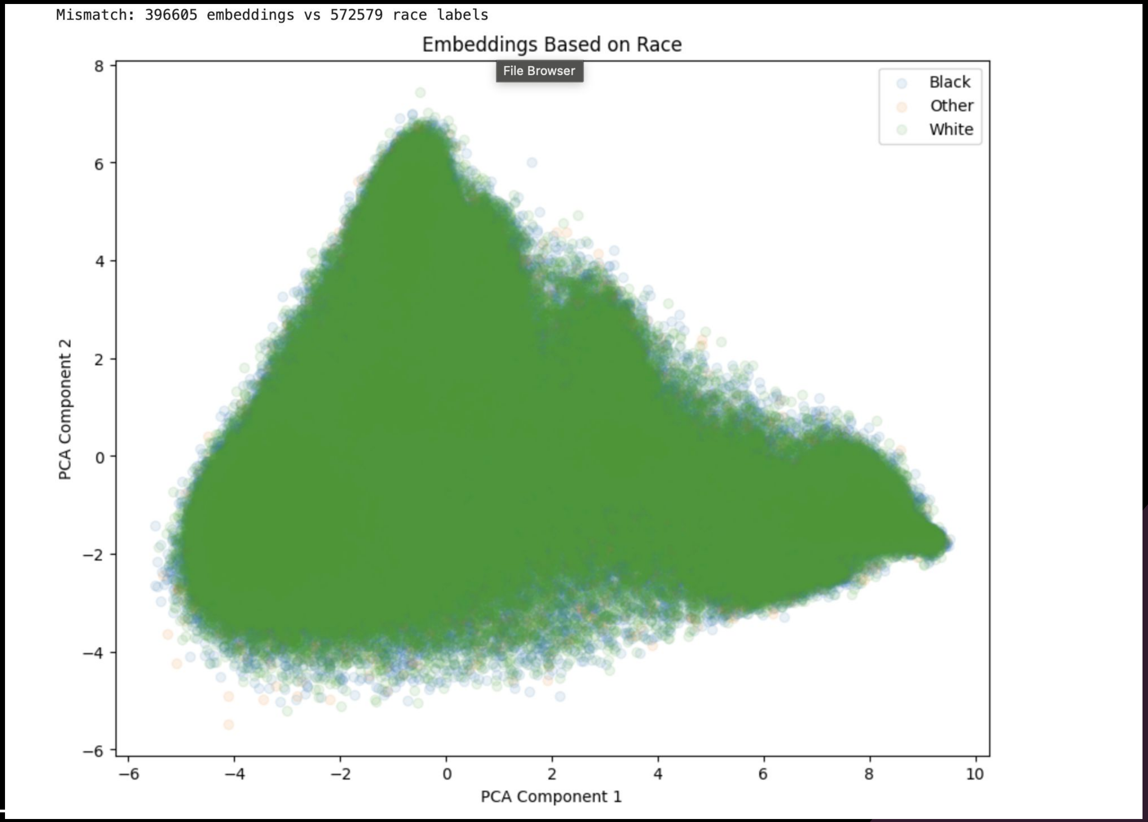
Plotted missingness by domain by demographic (sex, race)

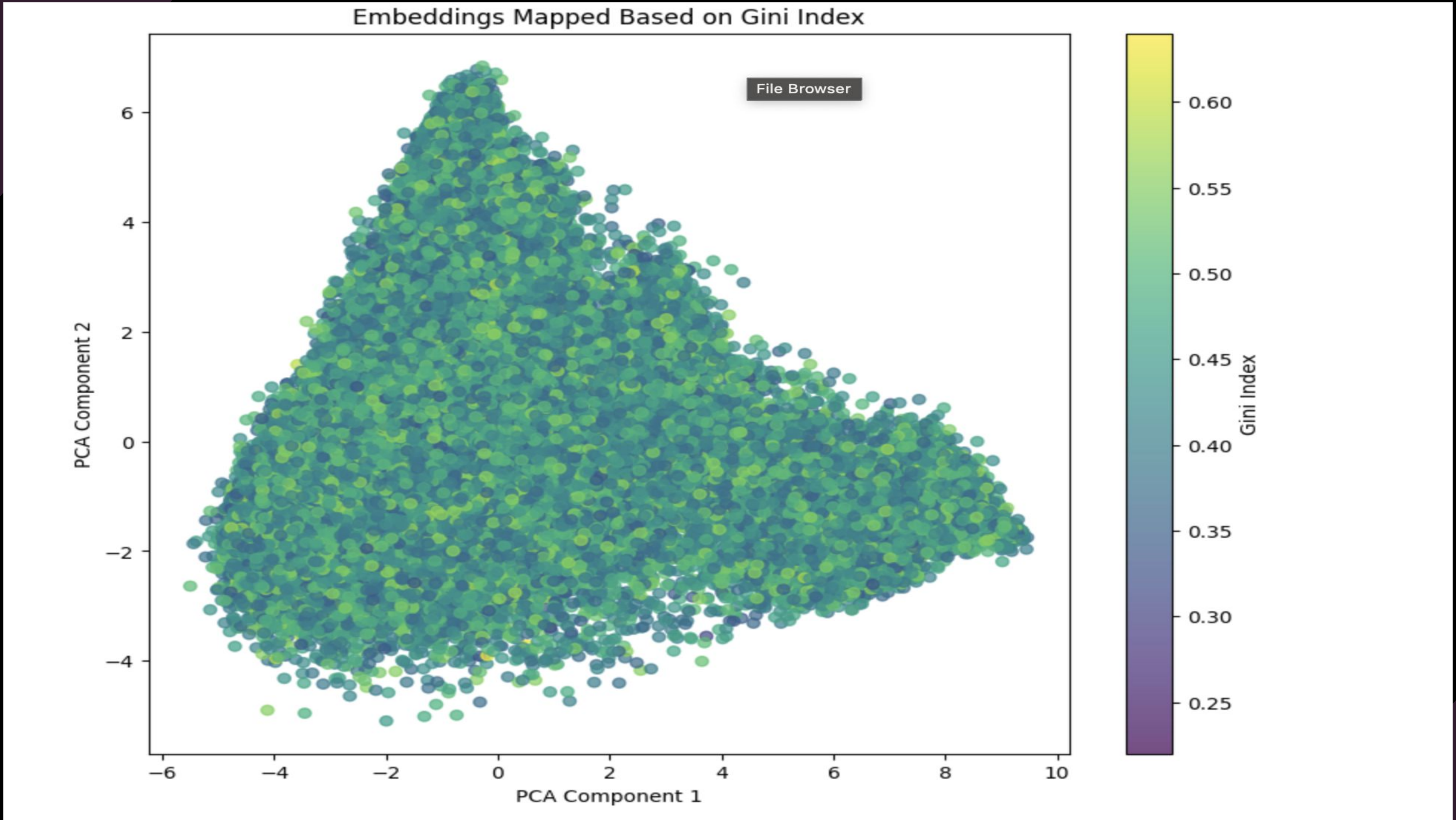
Methods - Embeddings

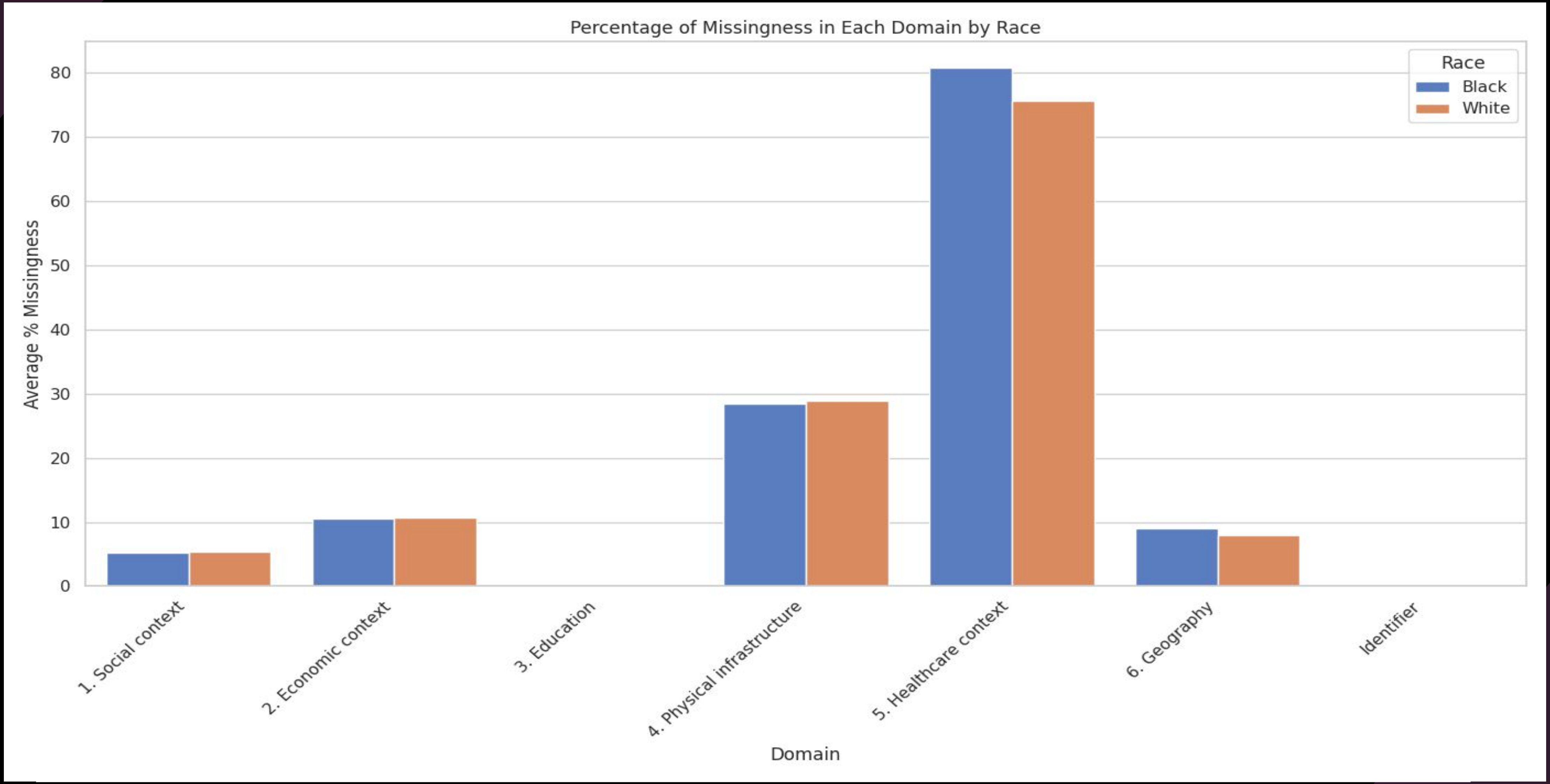


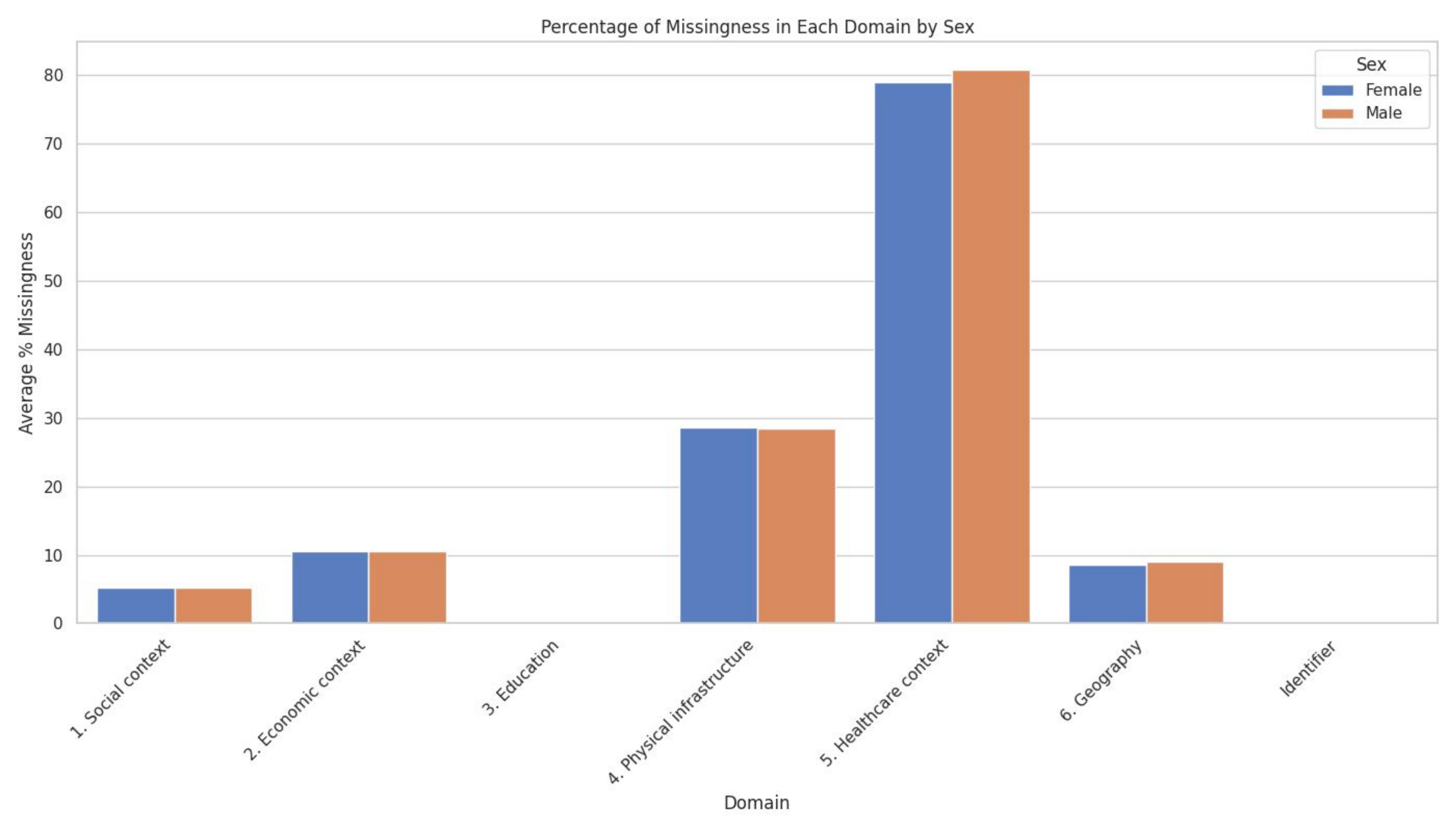
RESULTS: Embeddings



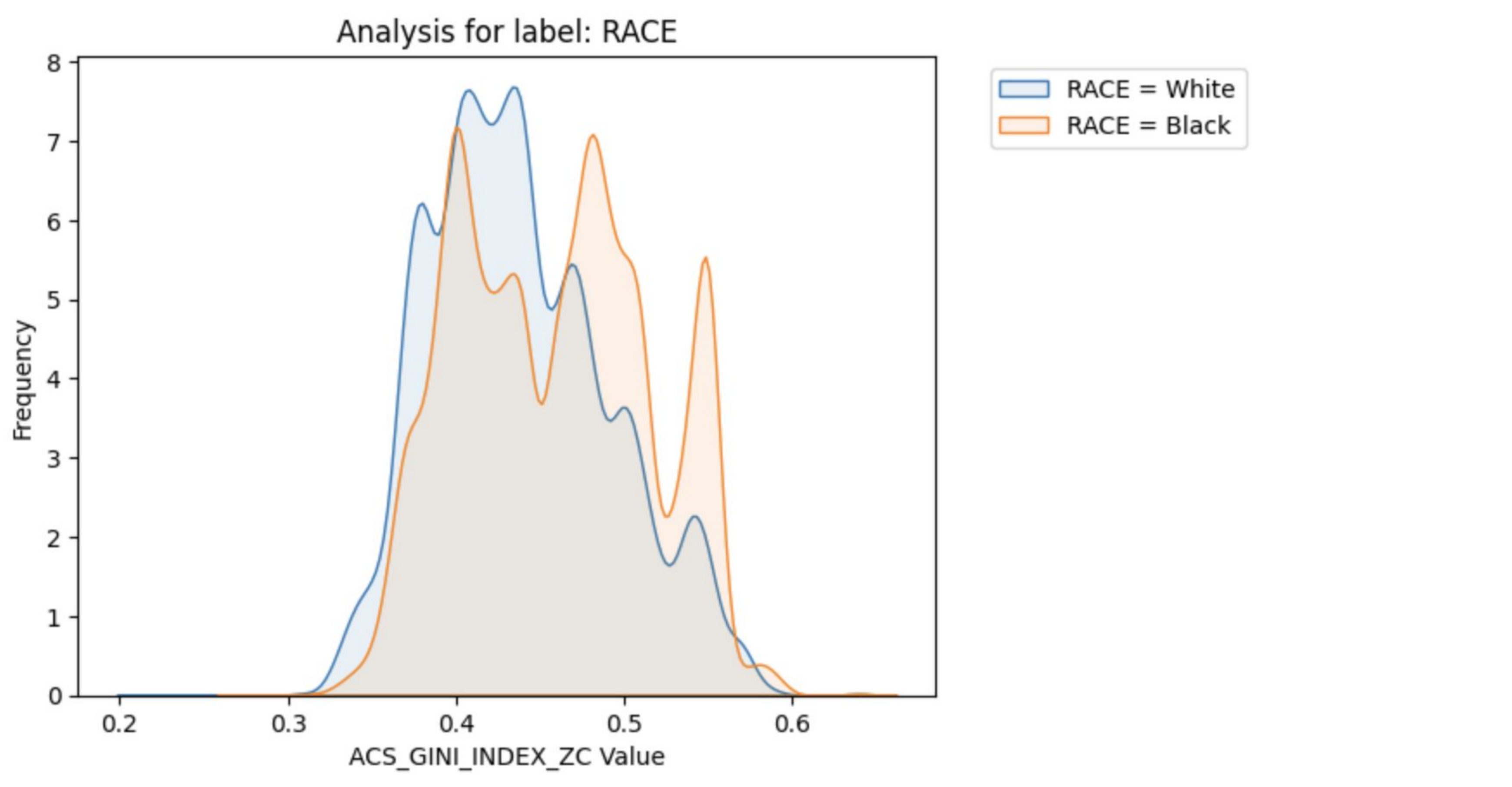



Embeddings Mapped Based on Gini Index
File Browser
PCA Component 2
PCA Component 1
Gini Index
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8
10
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25


Percentage of Missingness in Each Domain by Race
Race
Black
White
Average % Missingness
0
10
20
30
40
50
60
70
80
1. Social context
2. Economic context
3. Education
4. Physical infrastructure
5. Healthcare context
6. Geography
Identifier
Domain


Percentage of Missingness in Each Domain by Sex
Sex
Female
Male
Average % Missingness
0
10
20
30
40
50
60
70
80
1. Social context
2. Economic context
3. Education
4. Physical infrastructure
5. Healthcare context
6. Geography
Identifier
Domain

RESULTS: Example of subpopulation analysis for the SDOH per Race



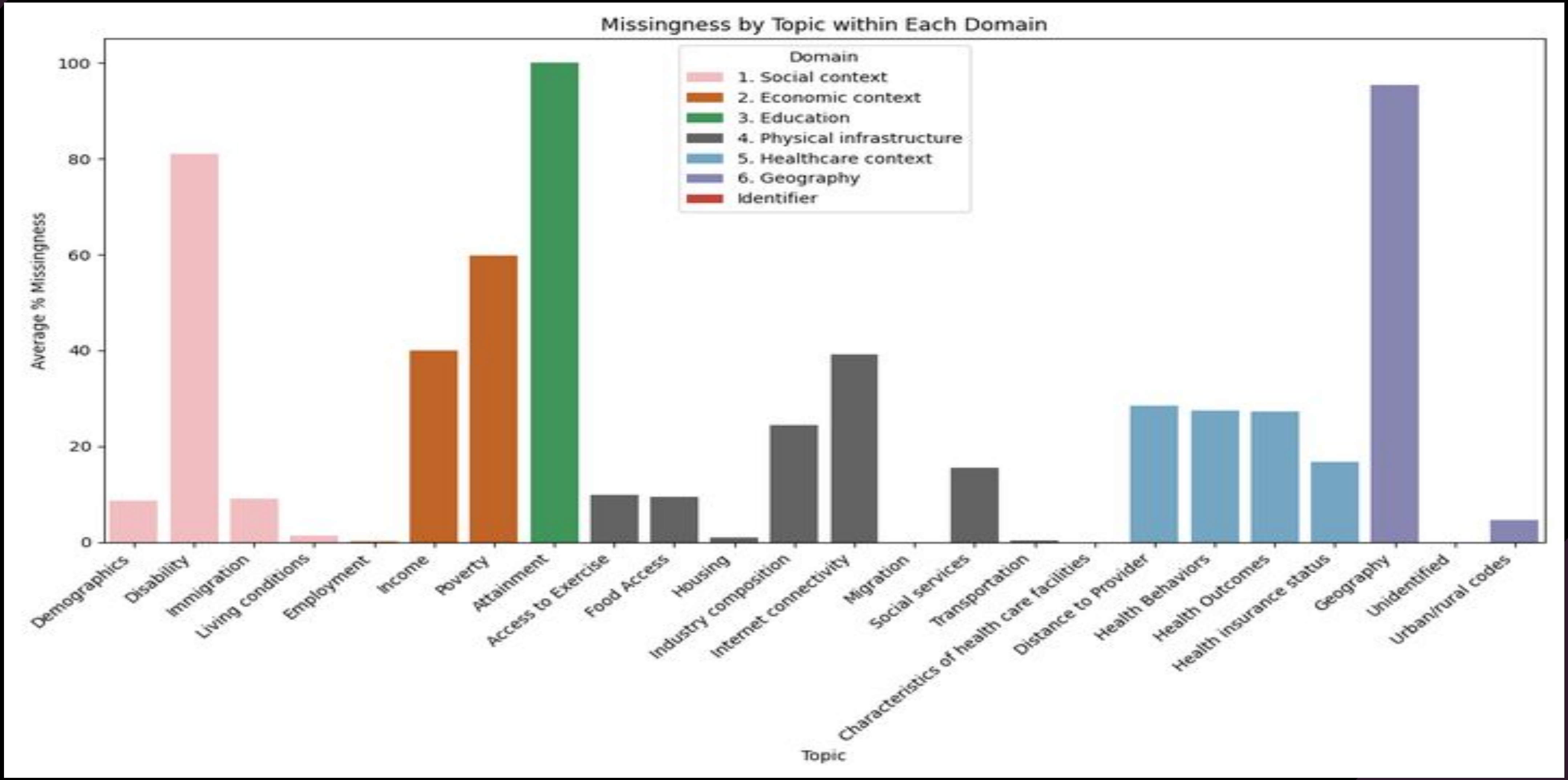
Conclusion/Relevance

- Lay the framework for preparing & clustering embeddings, and interactive graphs to see who is and is not represented in the SDOH dataset

MEET THE TEAM

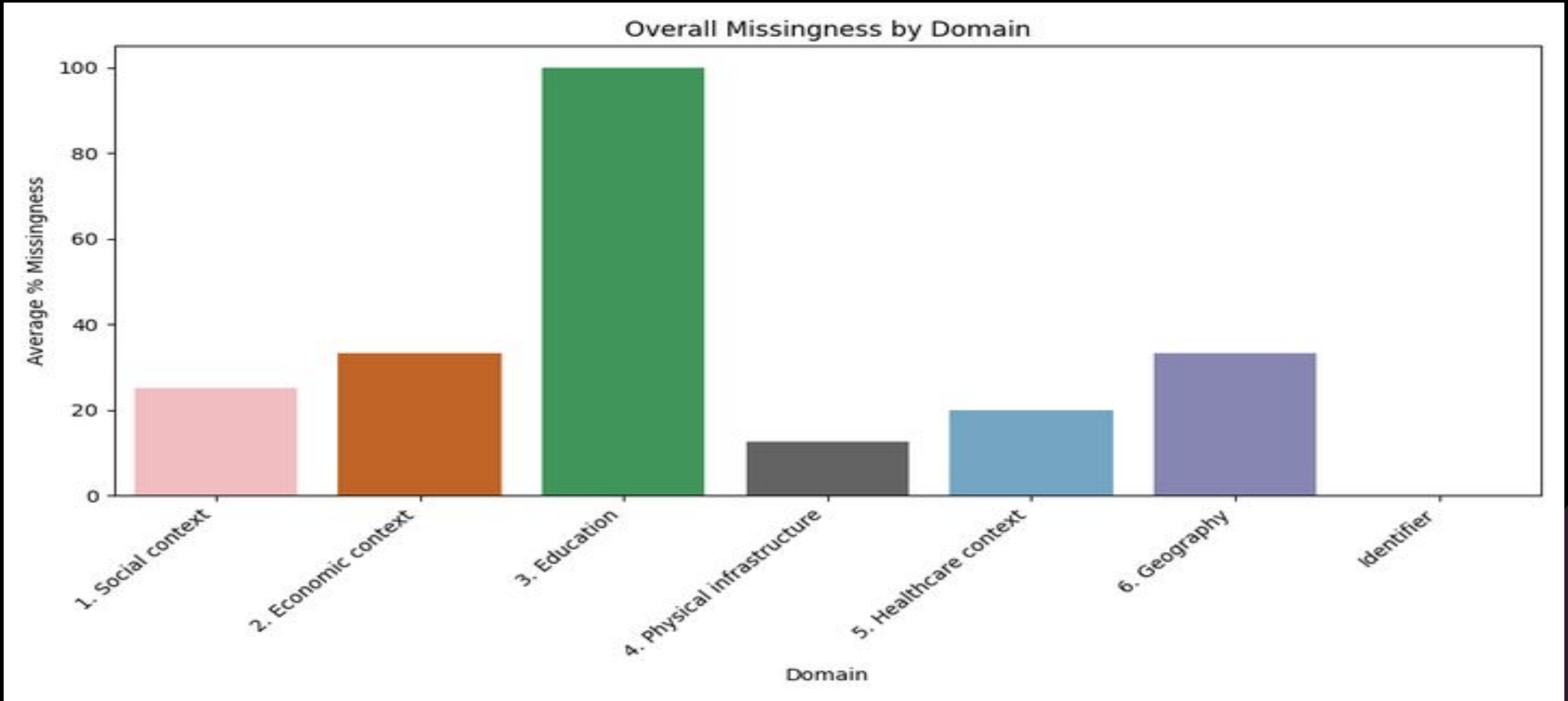
Extra Slides

RESULTS



Missingness by Topic within Each Domain
Domain
1. Social context
2. Economic context
3. Education
4. Physical infrastructure
5. Healthcare context
6. Geography
Identifier
Average % Missingness
0
20
40
60
80
100
Demographics
Disability
Immigration
Living conditions
Employment
Income
Poverty
Attainment
Access to Exercise
Food Access
Housing
Industry composition
Internet connectivity
Migration
Social services
Transportation
Characteristics of health care facilities
Distance to Provider
Health Behaviors
Health Outcomes
Health insurance status
Geography
Unidentified
Urban/rural codes
Topic

RESULTS


Overall Missingness by Domain
Average % Missingness
100
80
60
40
20
0
1. Social context
2. Economic context
3. Education
4. Physical infrastructure
5. Healthcare context
6. Geography
Identifier
Domain