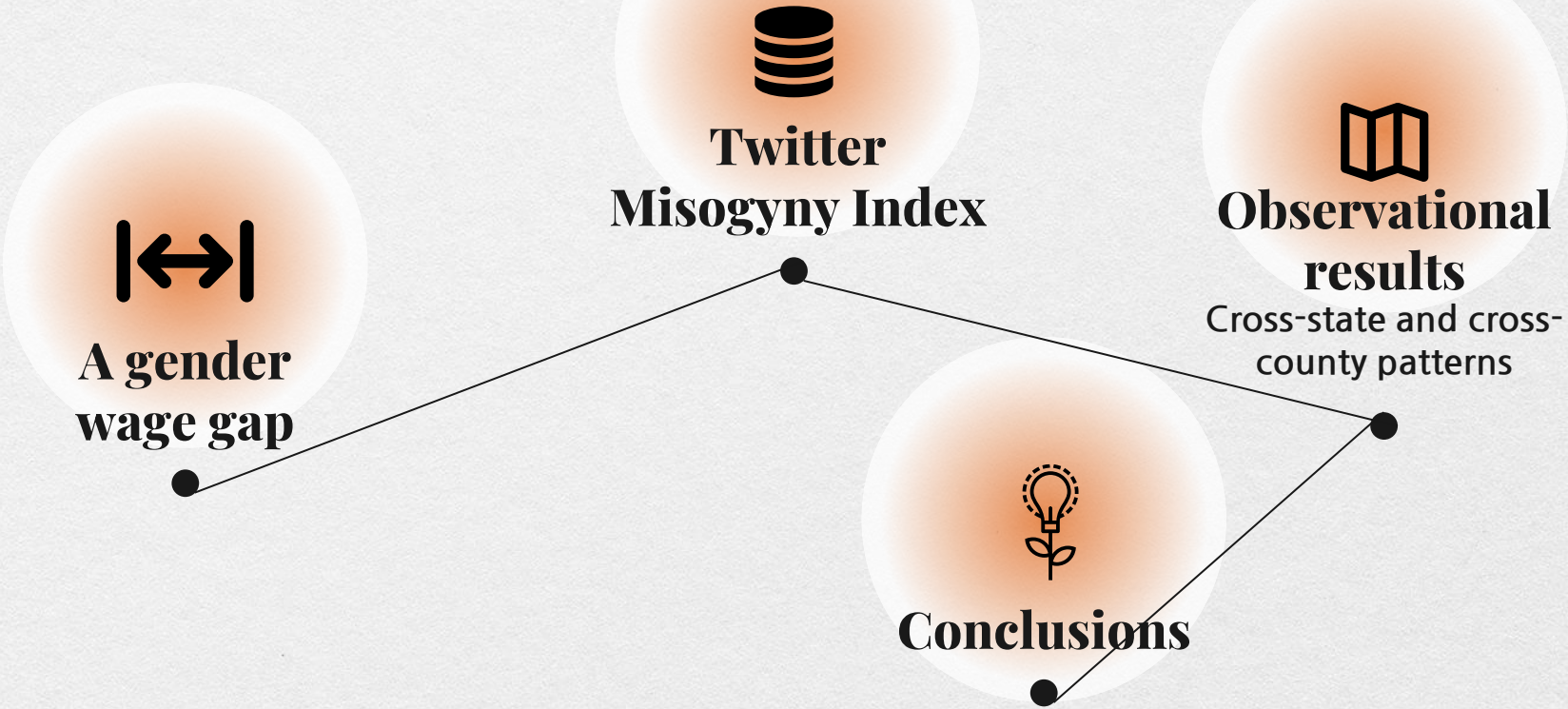
Final Presentation

Social Media, *Misogyny* and Labor Outcomes

Qiuyu (Jora) Li

Misogyny & Gender Wage Gap (Becker, 1957)

Table of *Content*



The Measurement of *Misogyny*

- **STEP 3: Construct the measurement – a Twitter Misogyny Index**
  - $TMI_{c,k} = \frac{NumMisoTweets_{c,k}}{TwitterPopulation_c}$
  - **c: county; k: term**
  - **Denominator: the number of tweets in county c I got by scrapping tweets with the searching keyword "the".**
- STEP 2: Construct a lexicon of misogynistic terms
  - From a lexicon of 1,300 potentially misogynistic terms built by Farrell et al. (2019)
  - Only keep those words that return misogynistic tweets in a sample of limited size
- STEP 1: Train a BERT classifier

The Measurement of *Misogyny*

- **STEP 3: Construct the measurement – a Twitter Misogyny Index**
  - $TMI_{c,k} = \frac{NumMisoTweets_{c,k}}{TwitterPopulation_c}$
  - **c: county; k: term**
  - **Denominator: the number of tweets in county c I got by scrapping tweets with the searching keyword "the".**
- **STEP 2: Construct a lexicon of misogynistic terms**
  - **From a lexicon of 1,300 potentially misogynistic terms built by Farrell et al. (2019)**
  - **Only keep those words that return misogynistic tweets in a sample of limited size**
- **STEP 1: Train a BERT classifier**

The Measurement of *Misogyny*

- STEP 3: Construct the measurement – a Twitter Misogyny Index
  - $TMI_{c,k} = \frac{NumMisoTweets_{c,k}}{TwitterPopulation_c}$
  - c: county; k: term
  - Denominator: the number of tweets in county c I got by scrapping tweets with the searching keyword “the”.
- **STEP 2: Construct a lexicon of misogynistic terms, and calculate the “efficiency” of each of them**
  - **From a lexicon of 1,300 potentially misogynistic terms built by Farrell et al. (2019)**
  - **Only keep those words that return misogynistic tweets in a sample of limited size**
- **STEP 1: Train a BERT classifier**
  - **Find an annotated dataset**
  - **Feed it to BERT**

The *process*: Train the classifer

01

Guest et al. (2021): the annotated dataset for training

From Reddit.
5,868 nonmisogynistic posts
v.s. 699 misogynistic posts

02

“bert-base-uncased”

Accuracy score on the test set: 0.9

The *process*

01

Guest et al. (2021): the annotated dataset for training

From Reddit.
5,868 nonmisogynistic posts v.s. 699 misogynistic posts

02

“bert-base-uncased”

Accuracy score on the test set: 0.9

03

Primary tweets collection

Five random dates in 2021: 02/26, 05/10, 07/22, 10/03, and 12/15.
In total: 3,894,780

04

Misoginistic terms

16 terms, “bitch” most frequently

05

Second round of tweets collection

10 random dates in 2021.
In total: 1,516,565

06

Identify user location

Number of tweets with state tags: 315,151
Number of tweets with county tags: 90,598

The *process*

01

Zhou et al. (2021): the annotated dataset for training

From Reddit.
5,868 nonmisogynistic posts v.s. 699 misogynistic posts

02

bert-base-uncased”

Accuracy score on the test set: 0.9

03

Primary tweets collection

Five random dates in 2021: 02/26, 05/10, 07/22, 10/03, and 12/15.
In total: 3,894,780

04

Misoginistic terms

16 terms, “bitch” most frequently

05

Second round of tweets collection

10 random dates in 2021.
In total: 1,516,565

06

Identify user location

Number of tweets with state tags: 315,151
Number of tweets with county tags: 90,598

The *process*

01

Guest et al. (2021): the annotated dataset for training

From Reddit.
5,868 nonmisogynistic posts v.s. 699 misogynistic posts

02

bert-base-uncased”

Accuracy score on the test set: 0.9

03

Primary tweets collection

Five random dates in 2021: 02/26, 05/10, 07/22, 10/03, and 12/15.
In total: 3,894,780

04

Misoginistic terms

16 terms, “bitch” most frequently

05

Second round of tweets collection

10 random dates in 2021.
In total: 1,516,565

06

Identify user location

Number of tweets with state tags: 309,083
Number of tweets with county tags: 59,121

The *process*

01

Guest et al. (2021): the annotated dataset for training

From Reddit.
5,868 nonmisogynistic posts v.s. 699 misogynistic posts

02

bert-base-uncased”

Accuracy score on the test set: 0.9

03

Primary tweets collection

Five random dates in 2021: 02/26, 05/10, 07/22, 10/03, and 12/15.
In total: 3,894,780

04

Misoginistic terms

16 terms, “bitch” most frequently

05

Second round of tweets collection

10 random dates in 2021.
In total: 1,516,565

06

Identify user location

Number of tweets with county tags: 59,121
Number of tweets with state tags: 309,083

The *process*

06

Identify user location

Number of tweets with county tags: 59,121
Number of tweets with state tags: 309,083

07

Third round of tweets scrapping by searching for the word “the”

On March 10
1,702,247 tweets were collected

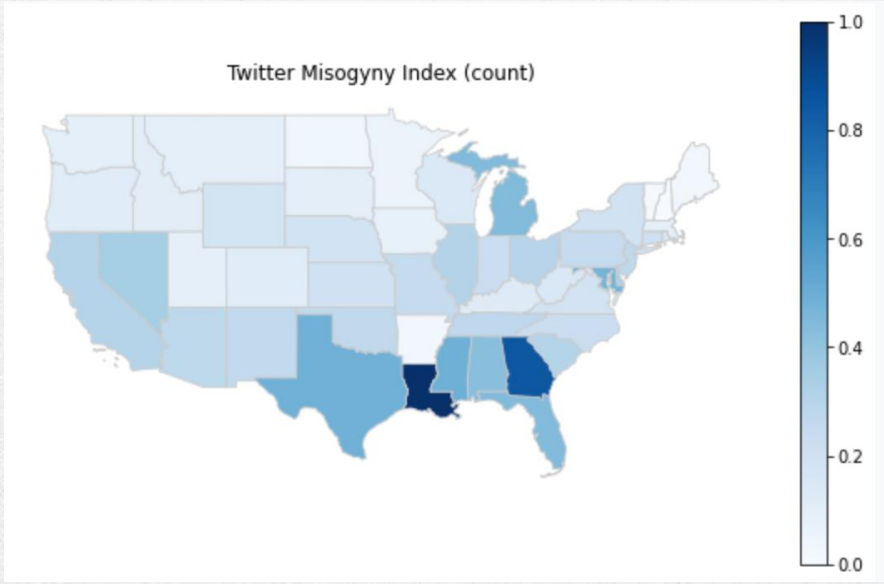
08

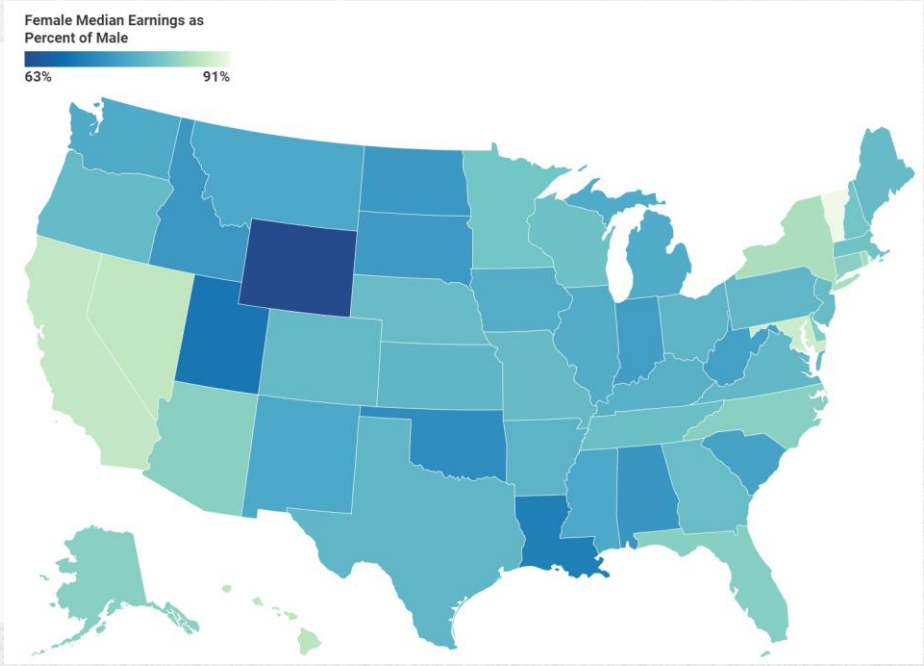
Calculating the Twitter Misogyny Index

$$TMI_{c,k} = \frac{NumMisoTweets_{c,k}}{TwitterPopulation_c}$$

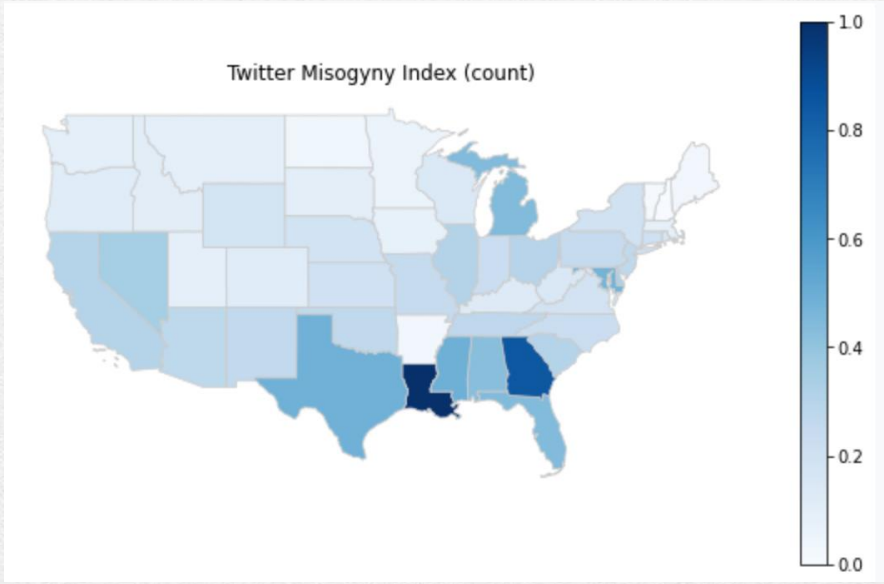
c means county and k means term

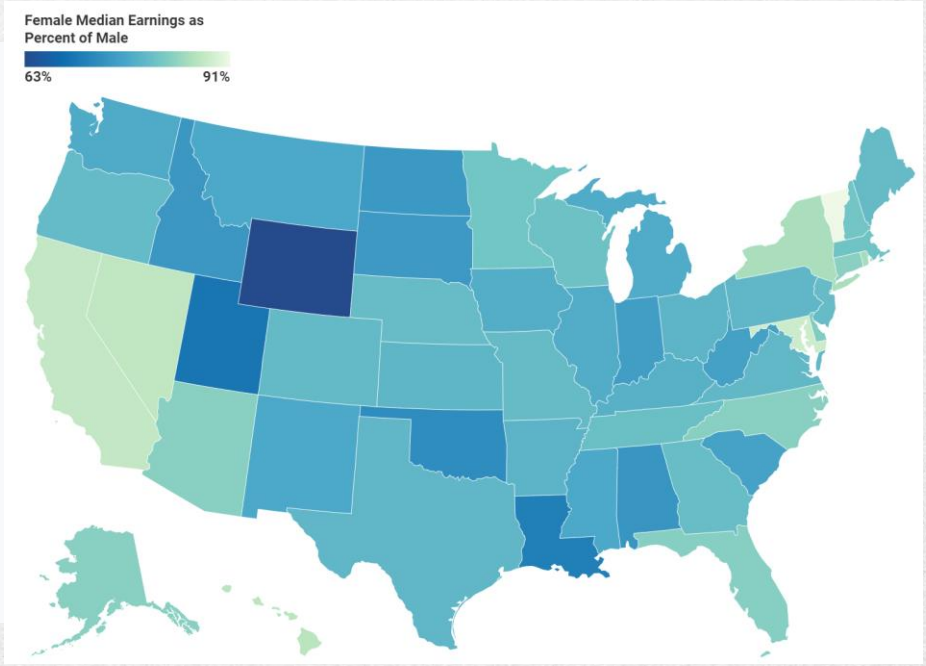
Primary results: state-level



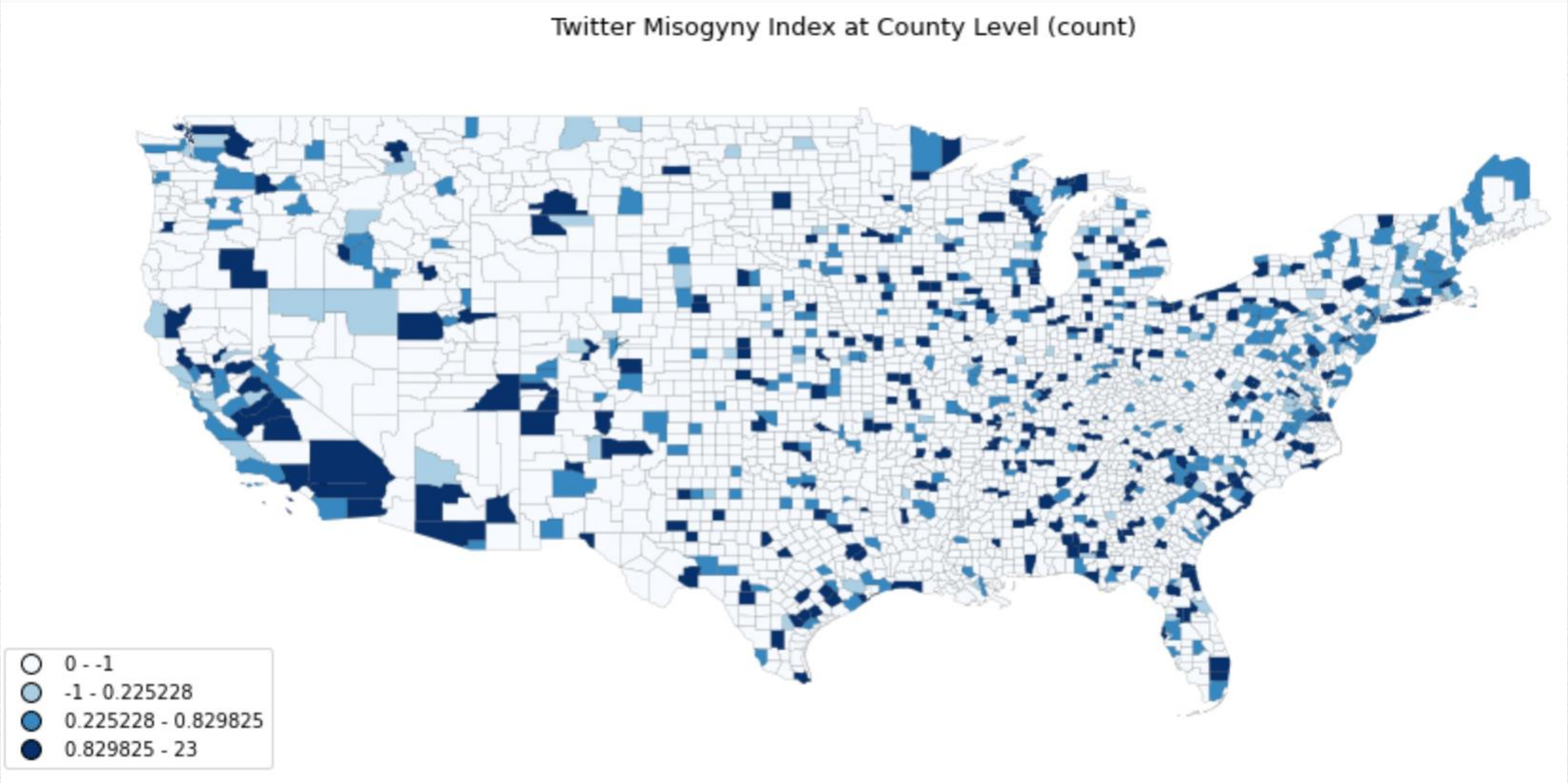


Primary results: state-level -- INCONSISTENT!





Primary results: county-level



Another Problem

Different terms in my lexicon are unequally efficient in returning misogynistic tweets

Solution

Calculate a coefficient matrix between the frequency of each of the terms and the labor outcomes

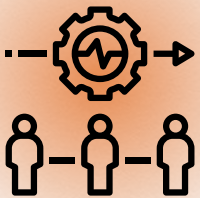
American Community Survey

Earnings

Created by supalerk laipa

1. Median wage of female workers / male workers
2. Median wage of female full-time workers / male workers
3. Average wage of female workers / male workers

Larbor market participation



Created by Nithinan Tatai

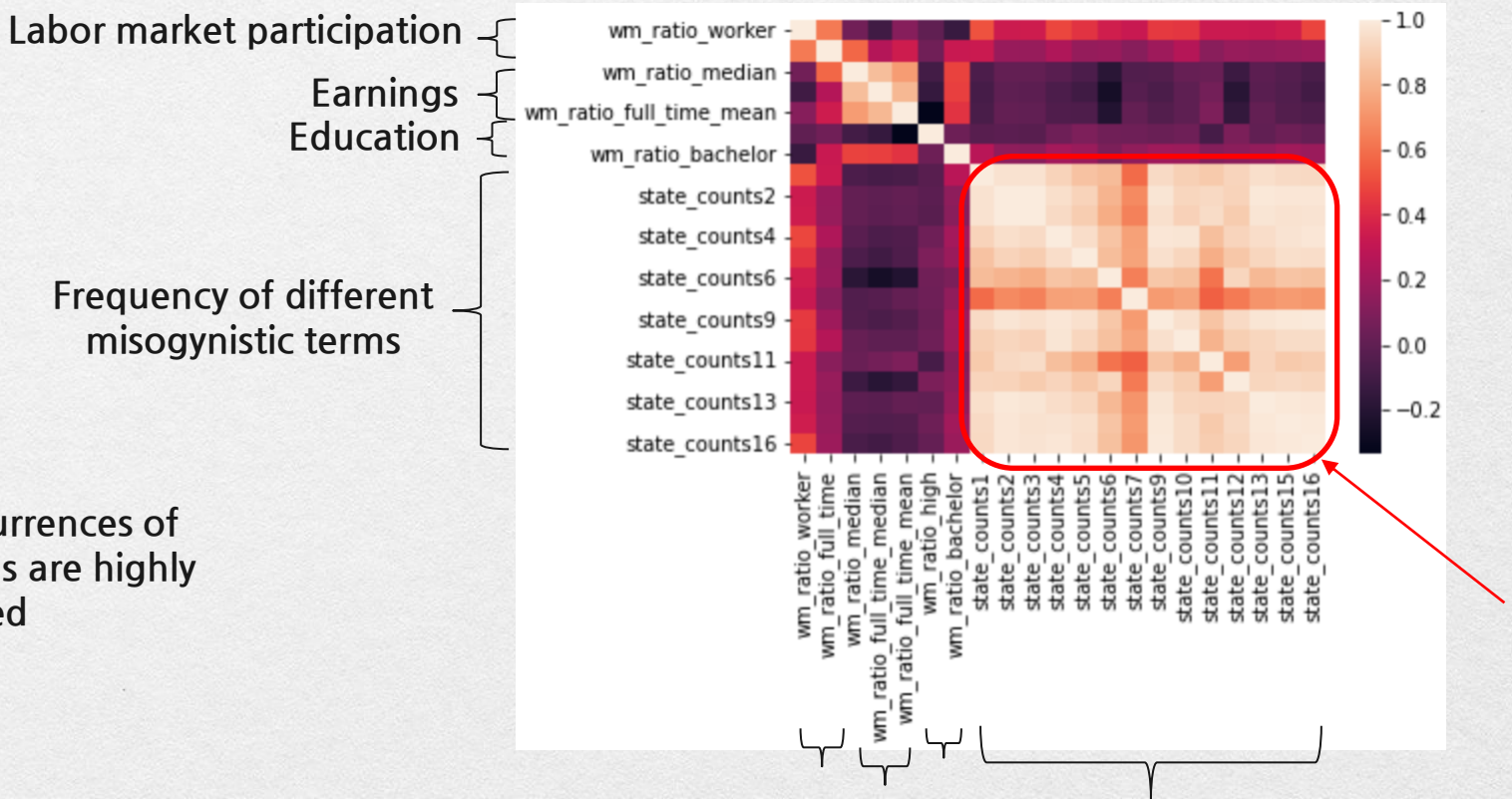
1. Proportion of workers with earnings among female population / male population
2. Proportion of full-time workers among female population / male population

Education

Created by Vectorstall

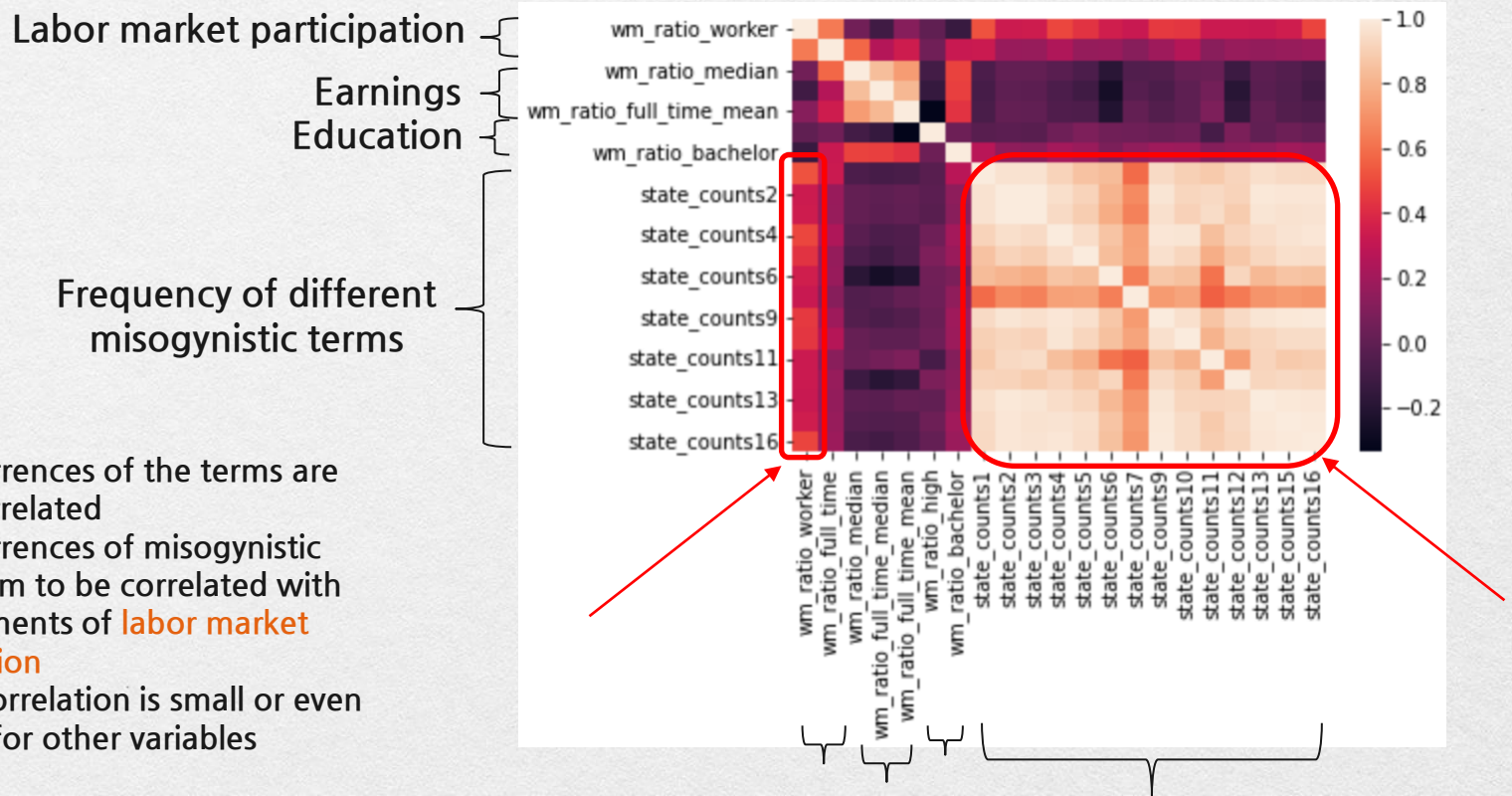
1. Proportion of people high school degree or above among 25-or-above female / male
2. Proportion of people with Bachelor's degree or above among 25-or-above female / male

Heatmap of correlation matrix – state-level

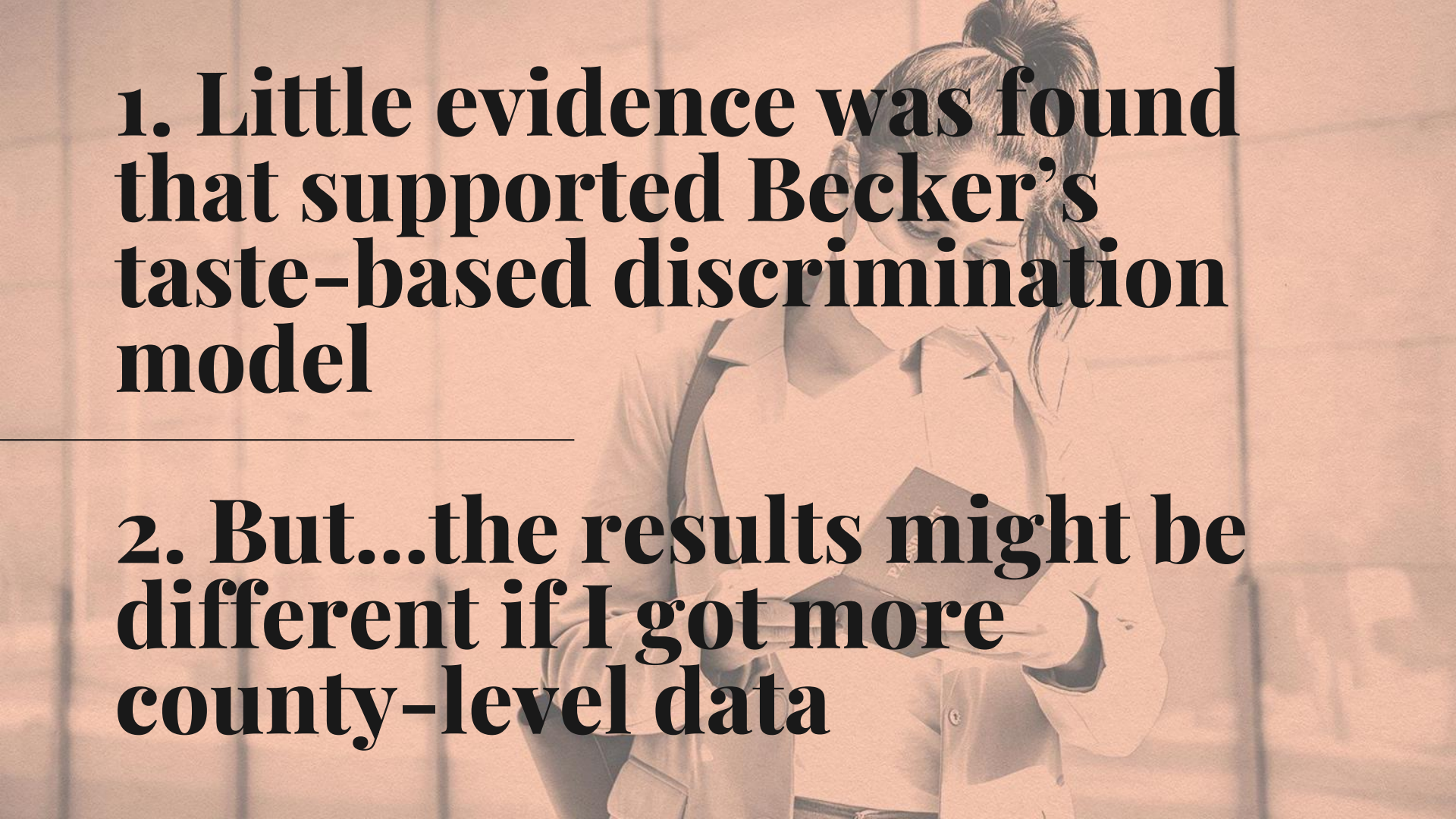


1. The occurrences of the terms are highly correlated

Heatmap of correlation matrix – state-level



1. The occurrences of the terms are highly correlated
2. The occurrences of misogynistic terms seem to be correlated with measurements of labor market participation
3. But the correlation is small or even negative for other variables



1. Little evidence was found that supported Becker's taste-based discrimination model

---

2. But...the results might be different if I got more county-level data