Predicting COVID-19 Vaccine Uptake in the United States

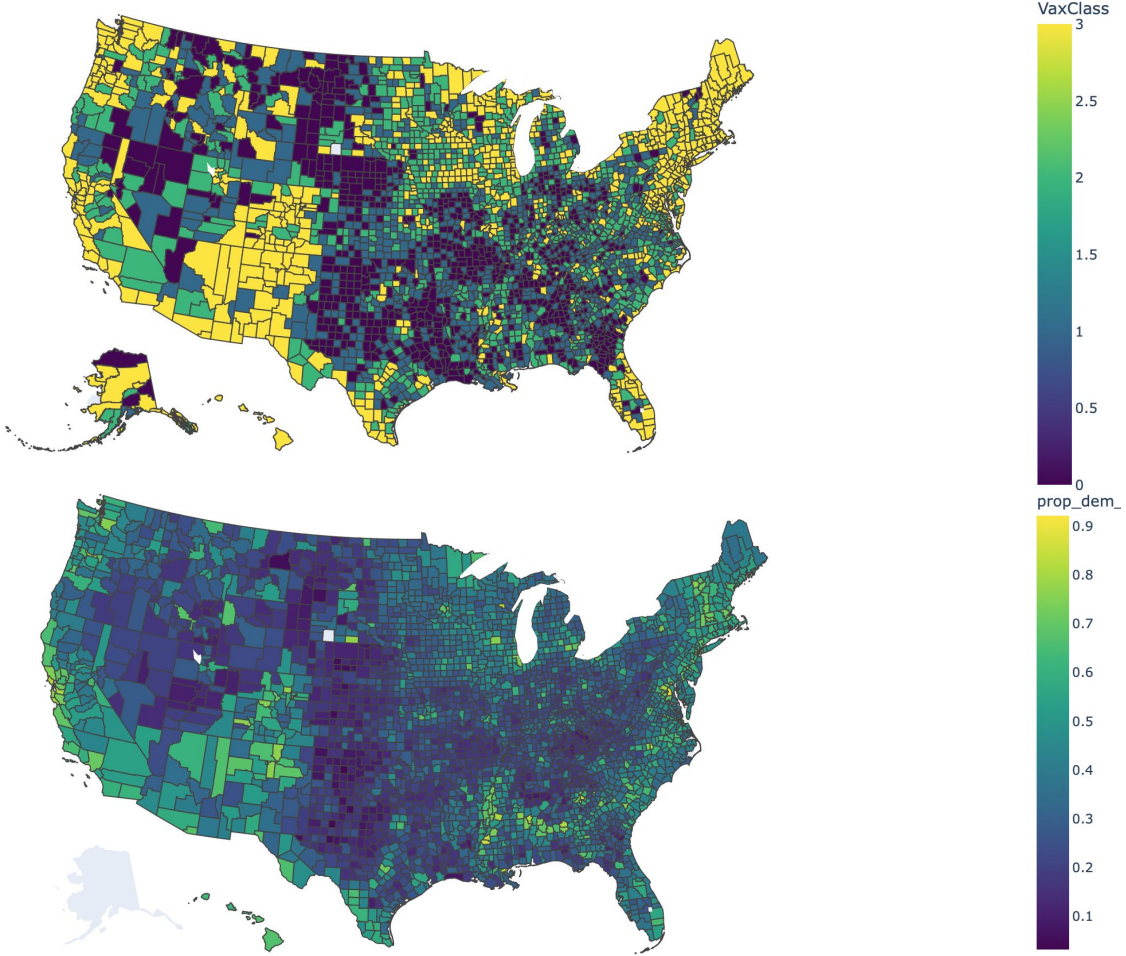
By:
Princess Allotey, *Carnegie Mellon University*
Robin Armstrong, *Cornell University*
Erik Bergland, *Brown University*
Eddie Mitchell, *University of Tennessee-Knoxville*
Nikki Wang, *Johns Hopkins University*

Research Questions

1. What are the most influential factors leading to individuals in the US choosing not to get vaccinated for COVID-19, beyond party vote?

2. What kind of policies/actions might encourage more individuals to choose to get the COVID-19 vaccine?

Vaccination Rate and Party Affiliation

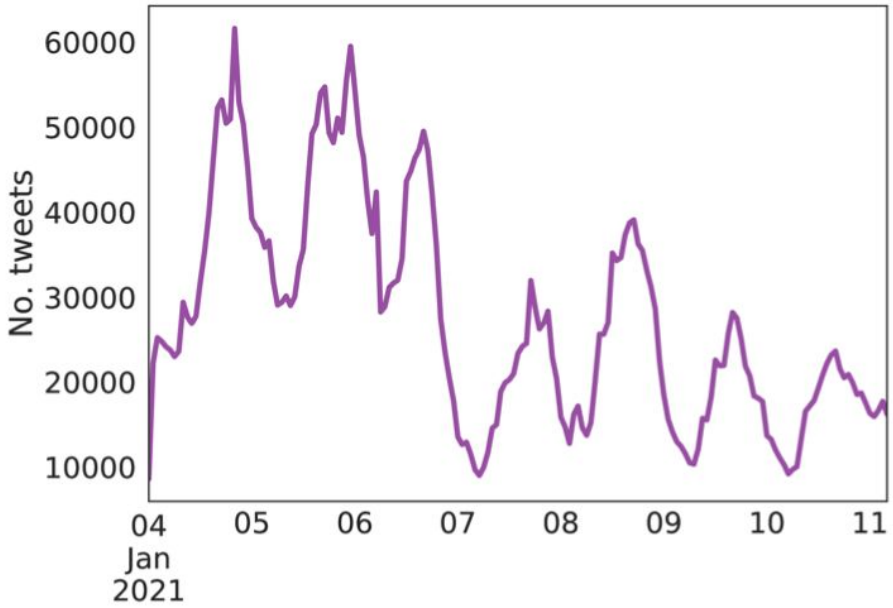


Summary

- Issues with Misinformation
  - CoVaxxy and Sparseness
- Exploratory Analysis
  - Correlation Matrix
  - PCA
- Regression Models
  - Logistic Regression + LASSO
  - Random Forest
- Classification Models
  - KNN
- Key Takeaways and Future Directions

Misinformation

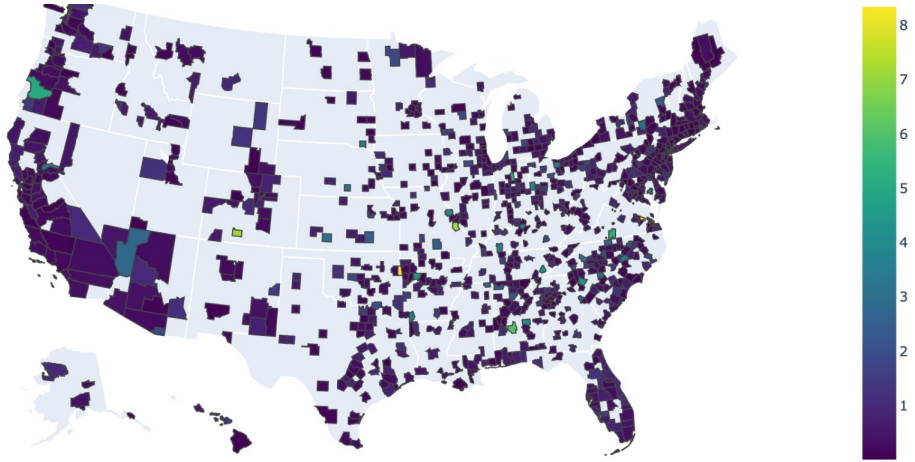
- Goal: use twitter data to identify regions featuring most misinformation
- Dataset: CoVaxxy[4]
  - Scrape tweets about covid and vaccines using specific search terms
  - Identify tweets with low-credibility
- Potential issue: collected over week of 1/4/2021-1/11/2021



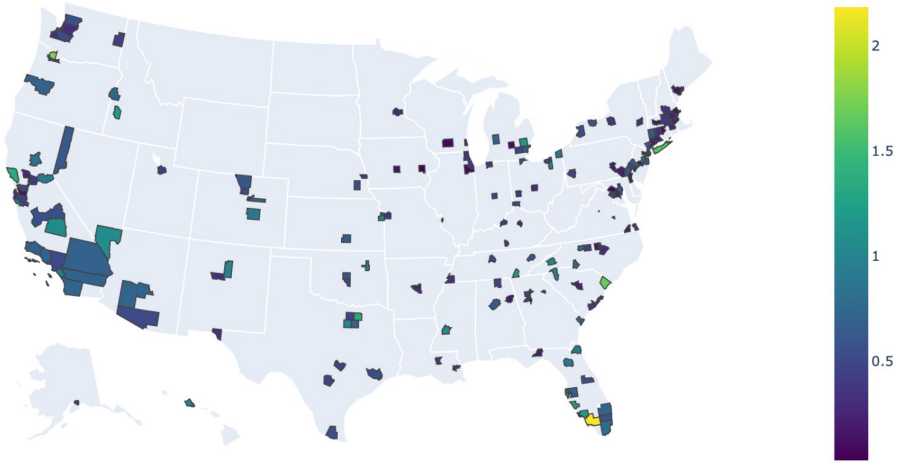
CoVaxxy and Misinformation Data

- Potential issue: most counties did not tweet about covid or vaccines

1 account 1 tweet Mean % Low-Credibility



100 accounts 10 tweets Mean % Low-Credibility

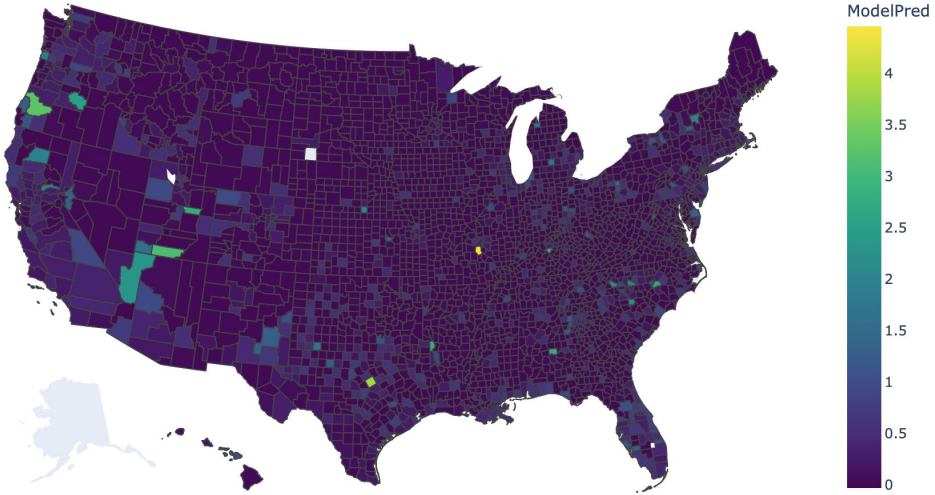


Extrapolating CoVaxxy

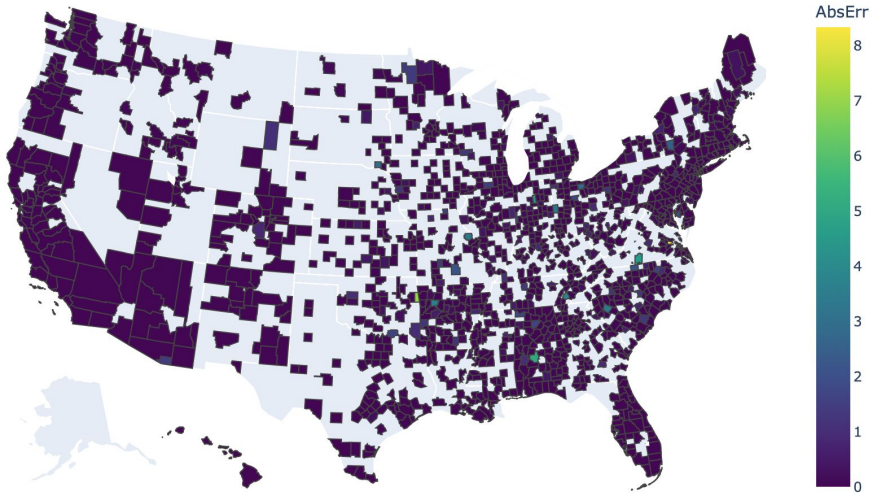
- Idea: use neural network to overcome lack of counties tweeting about COVID vaccines

1 account 1 tweet Mean % Low-Credibility

Prediction



Absolute Error



Summary

- Issues with Misinformation
  - CoVaxxy and Sparseness
- Exploratory Analysis
  - Correlation Matrix
  - PCA
- Regression Models
  - Logistic Regression + LASSO
  - Random Forest
- Classification Models
  - KNN
- Key Takeaways and Future Directions

Potential Factors

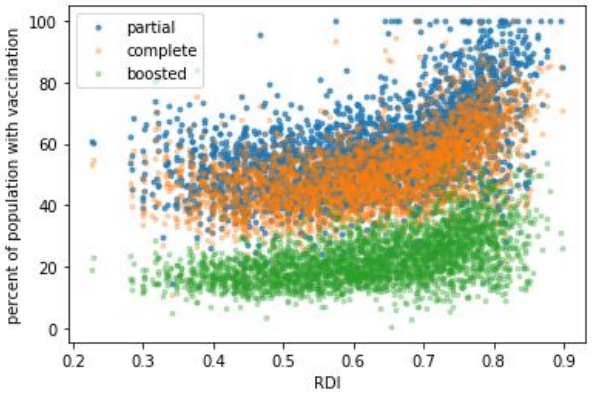
Dependent variables:

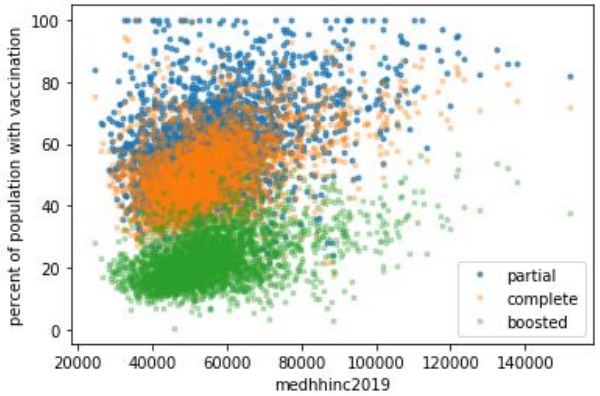
- **CVR_2022_03_03**: percent of population with complete vaccination on 3/3/2022
- **PVR_2022_03_03**: percent of population with partial vaccination on 3/3/2022
- **BR_2022_03_03**: percent of population with boosted vaccination on 3/3/2022

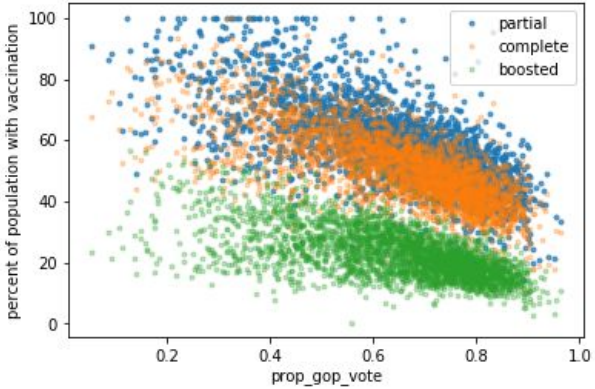
Potential factors of COVID-19 vaccine rate for each county:

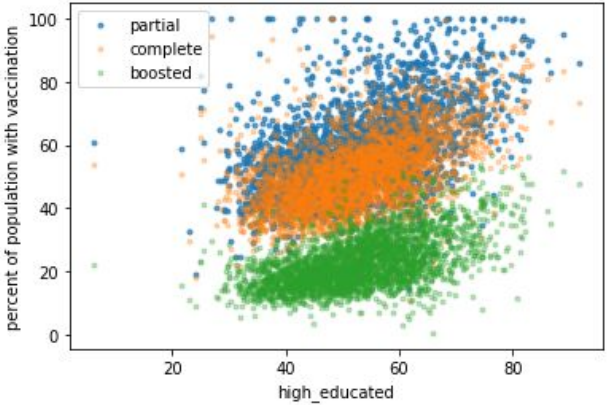
- **Prop_gop_vote**: proportion of people vote for gop (republican party)
- **RDI**: religion diversity index
- **high_educated**: percent of adults completing college or with a bachelor's degree or higher in 2015-19
- **medhhinc2019**: median household income in 2019
- **POPPCT_URBAN**: percentage of the total population of the county represented by the urban population
- **pop_density**: population Density
- **case_rate**: cumulative cases per 100,000 people
- **death_rate**: cumulative deaths per 100,000 people

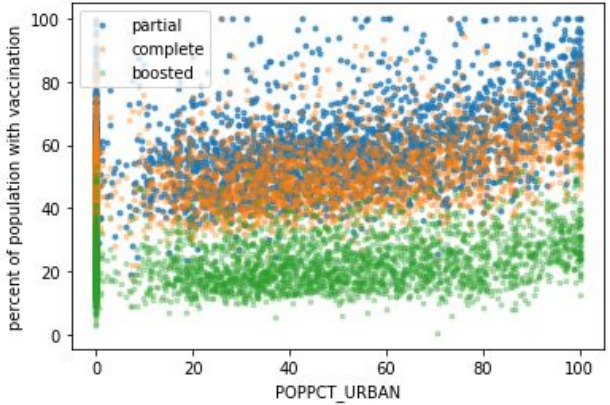
Partial / Complete / Boosted Vaccine Rates

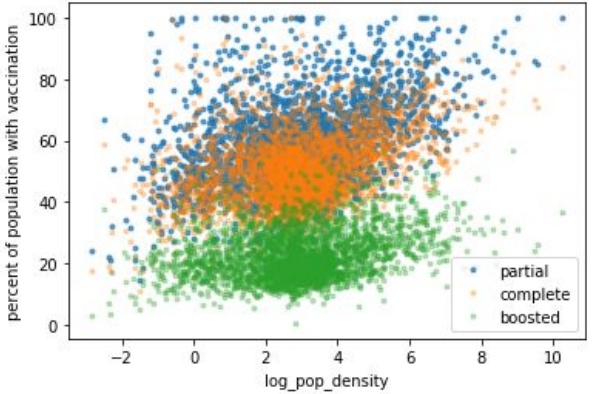


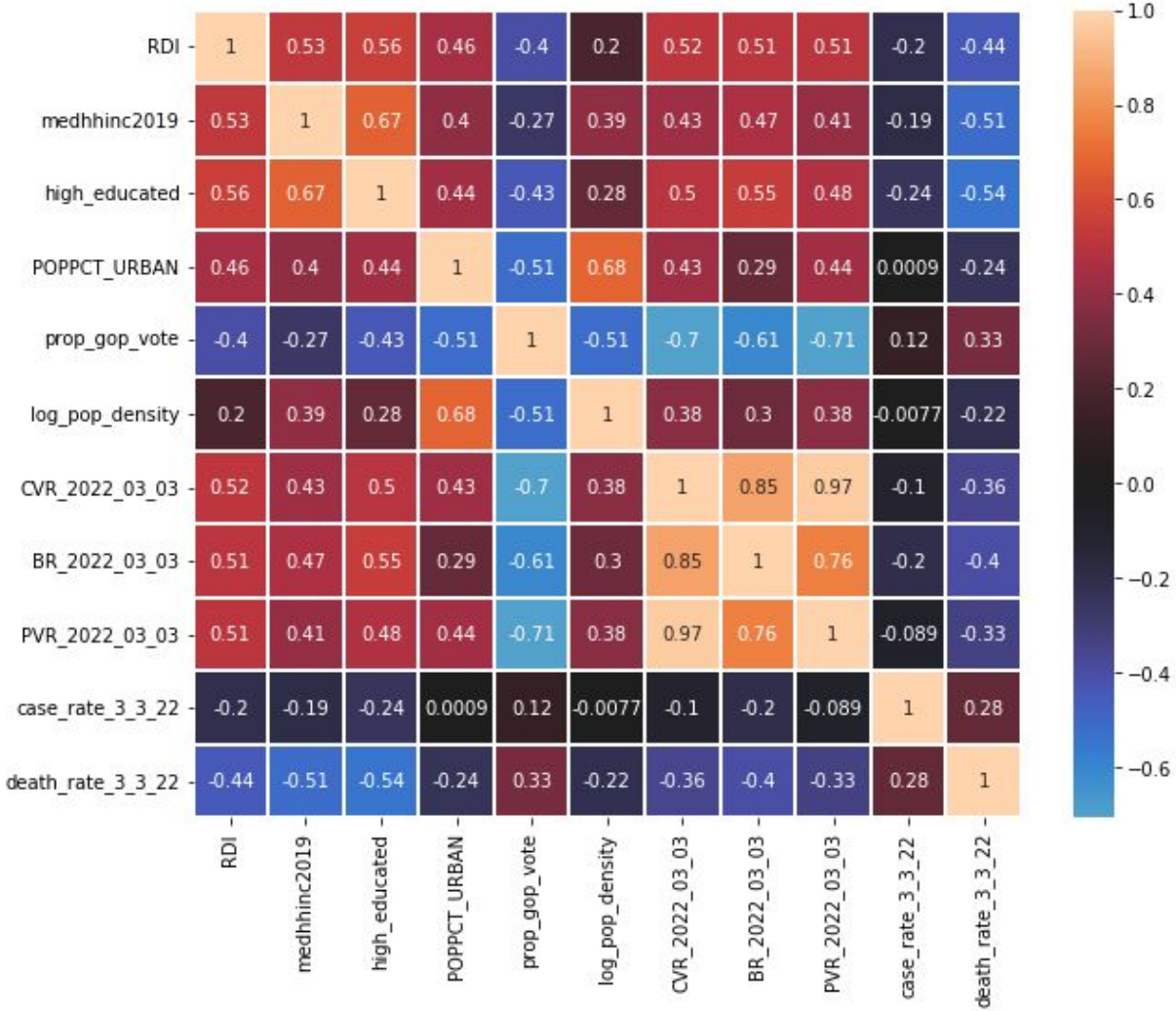












Correlation Matrix

What can we see from correlation matrix:

1. The higher RDI, income, education, and percentage of urban are, the higher vaccine rates are.
2. The higher proportion of people voting for the GOP, the lower vaccine rates are.
3. There are positive linear correlations between RDI, income, and education; percentage of urban and logarithm of population density.
4. Vaccine helps to reduce the death rate, but not the case rate.

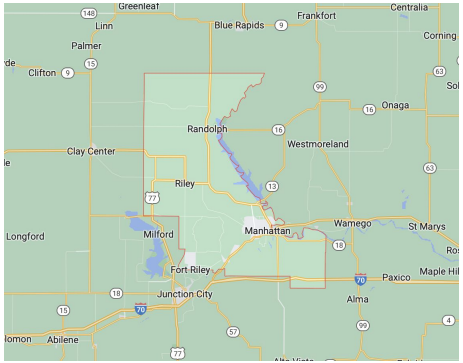
States	County	GOP_vote	RDI	Income	Education	Urban	vax_rate
Alabama	Russell County	1	1	1	2	0	1
Arkansas	Lee County	1	1	1	1	0	1
California	Alpine County	1	4	3	4	0	1
Georgia	Muscogee County	1	2	2	4	1	1
Georgia	Richmond County	1	2	1	2	1	1
Georgia	Stewart County	1	1	1	1	0	1
Kansas	Riley County	1	4	2	4	1	1
Louisiana	St Helena Parish	1	1	1	1	0	1
Mississippi	Issaquena County	1	1	1	1	0	1
Mississippi	Tallahatchie County	1	1	1	1	0	1
North Carolina	Hoke County	1	2	2	3	0	1
South Carolina	Jasper County	1	1	2	1	0	1
South Dakota	Ziebach County	1	4	1	1	0	1
Vermont	Caledonia County	1	4	2	3	0	1
Vermont	Franklin County	1	4	4	2	0	1

Low GOP Vote & Low Vaccination Rate

More on peculiarities

Alpine County, California:

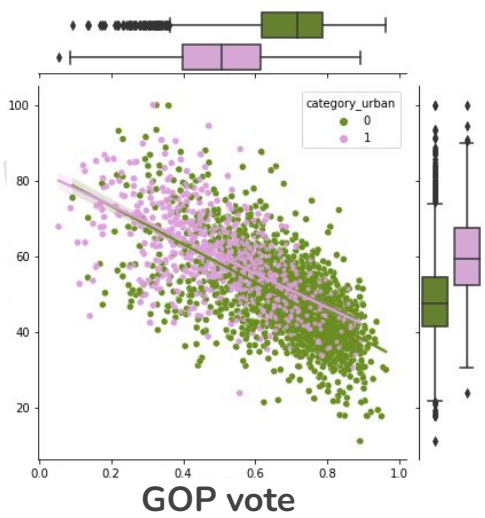
- 40.48% Vaccination rate as at 03/03/2022
- 32.9% Republican vote
- 100% Rural Population
- 0.599 people/square kilometre
- 29.93% are 65+ years old (*the max percentage across all counties)
- 8400 case rate as at 10/01/2022 (1th percentile)
- 0 death rate as at 10/01/2022



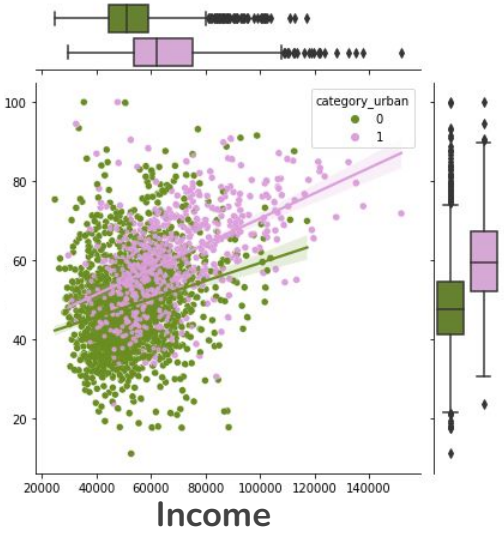
Riley County, Kansas:

- 38.8% Vaccination rate as at 03/03/2022
- 47% Republican vote
- High RDI and education level, 86% Urban
- 48 people/square kilometre (76th percentile)
- 18076 case rate as at 10/01/2022 (20th percentile)
- 80.6 death rate as at 10/01/2022(7th percentile)

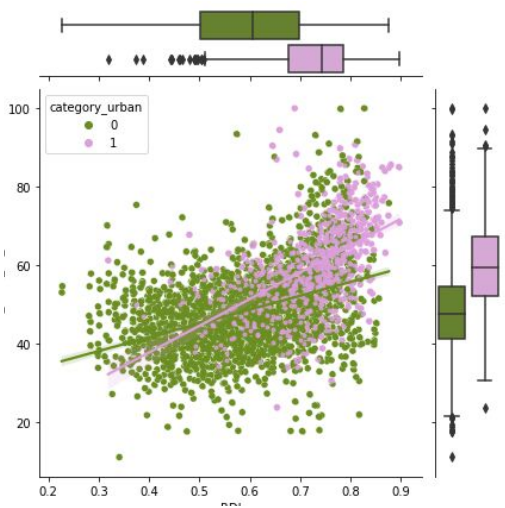
Vax Rate



GOP vote

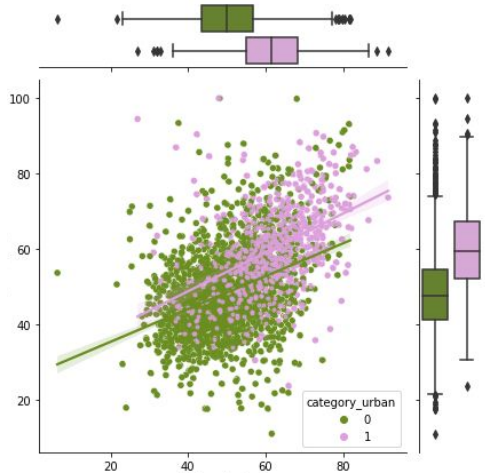


Income

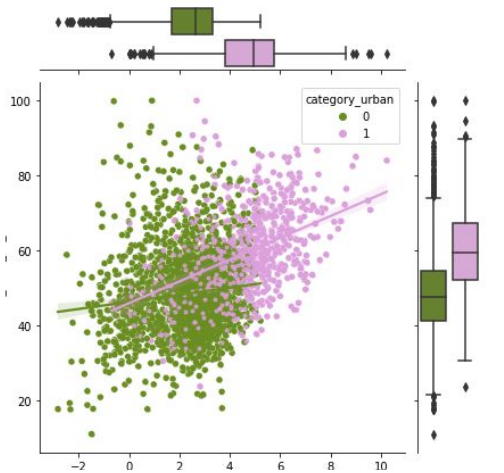


RDI

Vax Rate



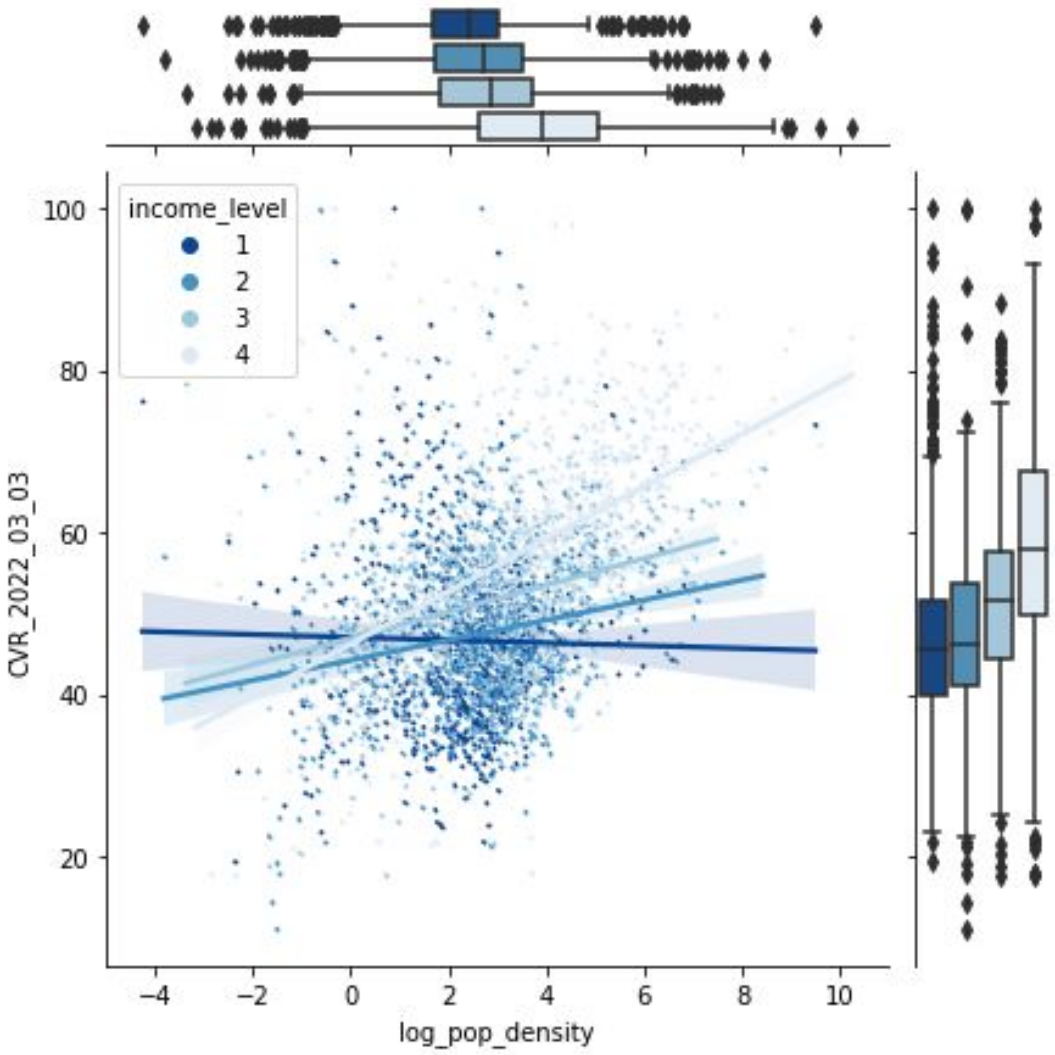
Education



log(population density)

Urban vs **Rural**

Vaccine Rate and the 5 variables



As income increases, the linear correlation between population density and vaccine rate becomes stronger.

Correlation Coefficient between log(pop_density) and Vaccine Rate:

- Total 0.36
- Low -0.022
- Lower Middle 0.2
- Upper Middle 0.28
- High 0.49

Lasso Regression:

Vax Rate = 46.45
+ 0.0 * log(pop_density) * category_income1
+ 0.53 * log(pop_density) * category_income2
+ 1.72 * log(pop_density) * category_income3
+ 3.19 * log(pop_density) * category_income4
Explained variance score = 0.2456

Vax Rate = 43.41 + 2.6 * log(pop_density)
Explained variance score = 0.0514

Principal Component Analysis (PCA) - 2D

Ideas behind PCA:

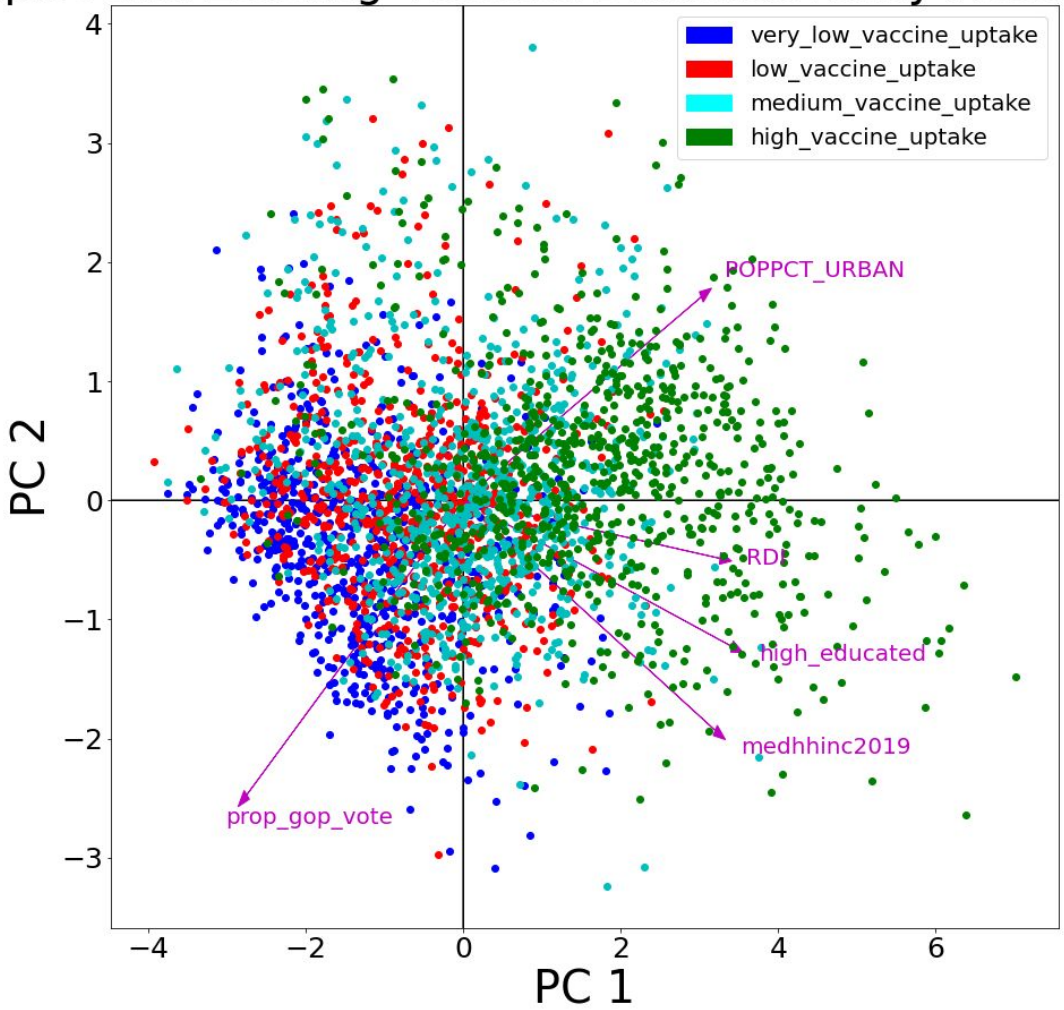
- How do we find the interrelation between variables?
- Reduce the dimensionality of a dataset with interrelated variables while retaining as much as possible of the variation present in the data set.[1]

Principal Component Analysis (PCA) - 2D

Variables Used:

- POPPCT_URBAN
- RDI (Religious Diversity Index)
- high_educated
- medhhinc2019
- prop_gop_vote
- Complete vaccination (03/03/2022)

Biplot and Loading Vectors: Dimensionality Reduction

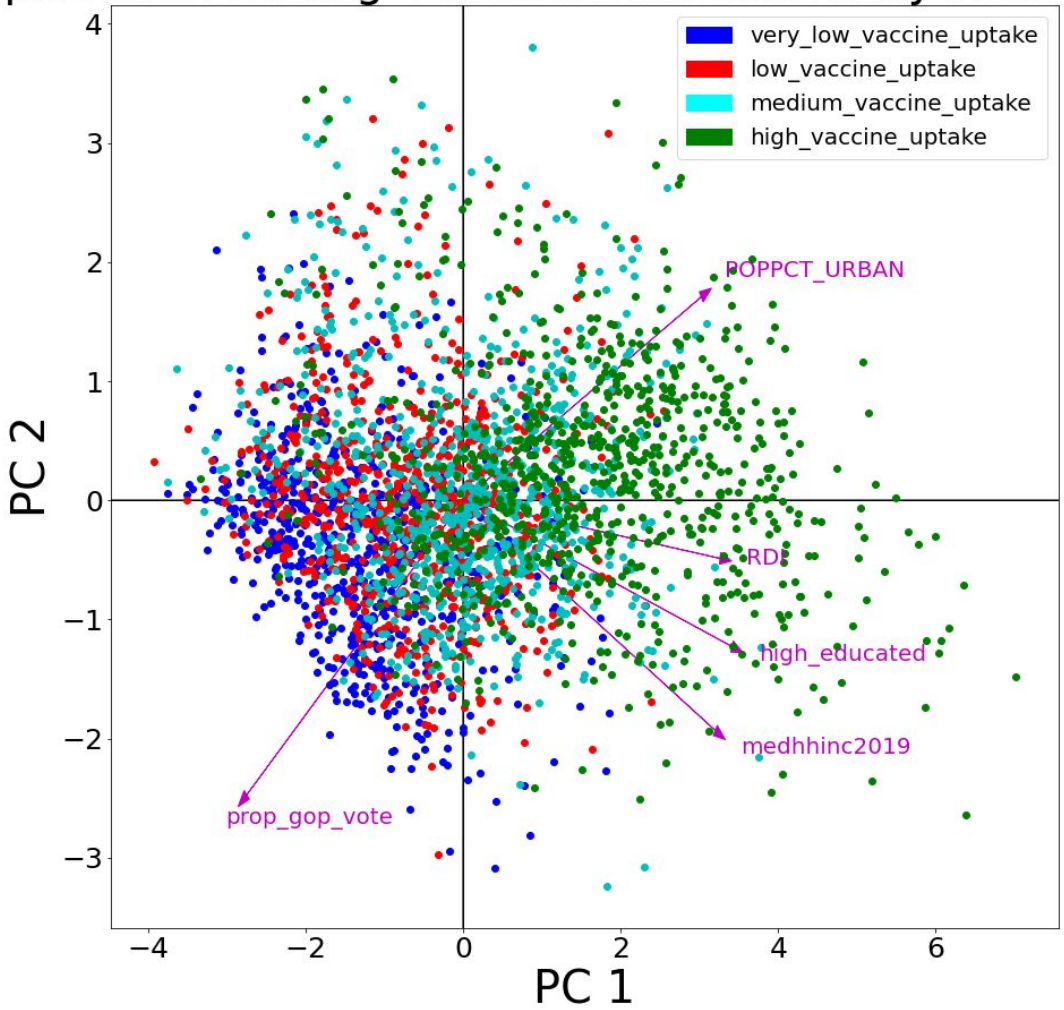


Interpretation[2]:

- **Length** of loading vectors
  - Similar size hence approx. equally represented by PC1 and PC2 (except RDI)
- **Proximity** to PC1 & PC2 axes
  - Approximately 45 degrees away from PC1 & PC2 so each variable contributes somewhat equally to PC1 & PC2
- **Angles** between loading vectors
  - ***Prop_gop_vote*** and ***high_educated*** vectors are 90 degrees apart hence have no correlation
  - ***RDI***, ***medhhinc2019*** and ***high_educated*** are close to each other hence highly correlated to each other

*Total Explained Variance: 74.52%

Biplot and Loading Vectors: Dimensionality Reduction

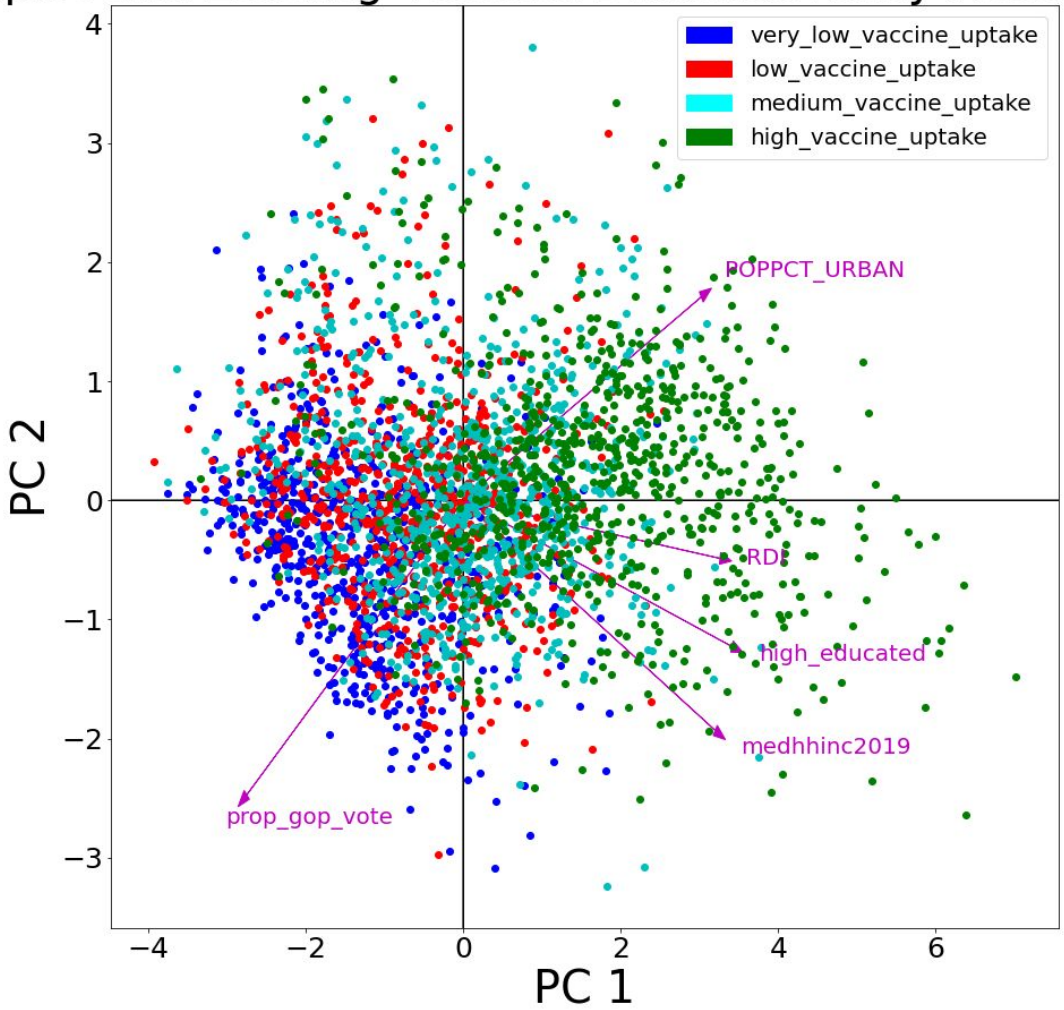


Interpretation[2]:

- **Length** of loading vectors
  - Similar size hence approx. equally represented by PC1 and PC2 (except RDI)
- **Proximity** to PC1 & PC2 axes
  - Approximately 45 degrees away from PC1 & PC2 so each variable contributes somewhat equally to PC1 & PC2
- **Angles** between loading vectors
  - ***Prop_gop_vote*** and ***high_educated*** vectors are 90 degrees apart hence have no correlation
  - ***RDI***, ***medhhinc2019*** and ***high_educated*** are close to each other hence highly correlated to each other

*Total Explained Variance: 74.52%

Biplot and Loading Vectors: Dimensionality Reduction



Interpretation[2]:

- **Length** of loading vectors
  - Similar size hence approx. equally represented by PC1 and PC2 (except RDI)
- **Proximity** to PC1 & PC2 axes
  - Approximately 45 degrees away from PC1 & PC2 so each variable contributes somewhat equally to PC1 & PC2
- **Angles** between loading vectors
  - ***Prop_gop_vote*** and ***high_educated*** vectors are 90 degrees apart hence have no correlation
  - ***RDI***, ***medhhinc2019*** and ***high_educated*** are close to each other hence highly correlated to each other

*Total Explained Variance: 74.52%

Correlation Matrix & Principal Component Analysis:

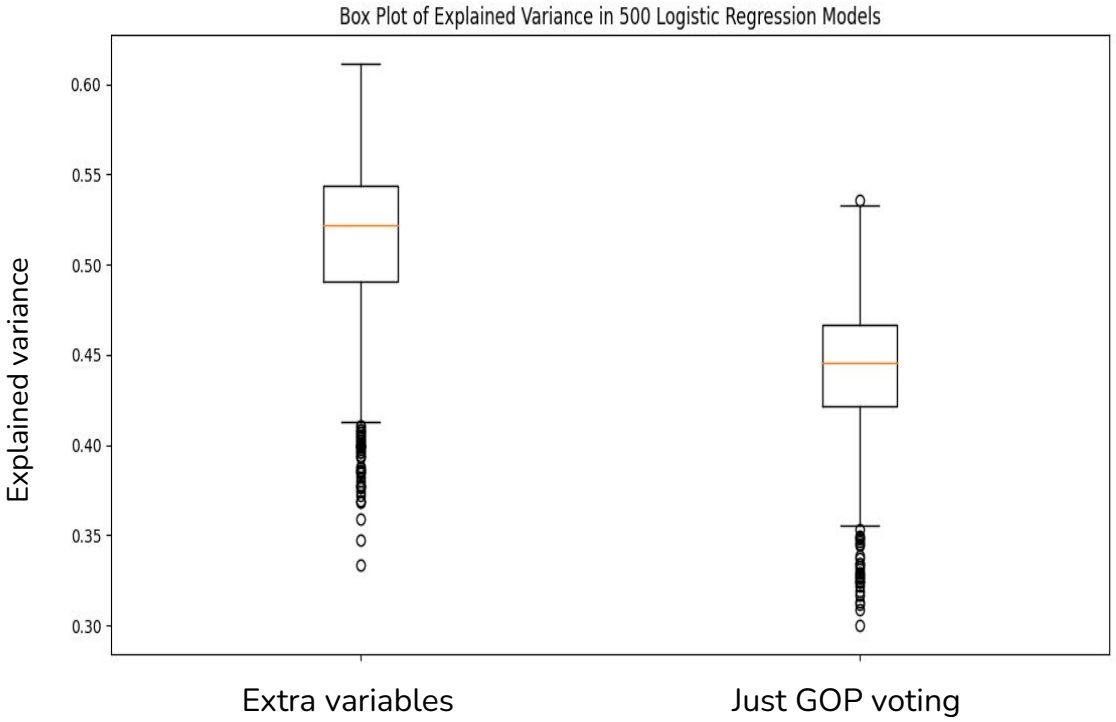
Connecting the dots

Variable	Variable	Correlation Coefficient	PCA Interpretation
Prop_gop_vote	high_educated	-0.43	No correlation
RDI	medhhinc2019	0.53	High positive correlation
RDI	high_educated	0.56	High positive correlation
medhhinc2019	high_educated	0.67 (*one of highest correlation coefficients on the correlation matrix)	High positive correlation
Prop_gop_vote	POPPCT_URBAN	-0.51	High negative correlation

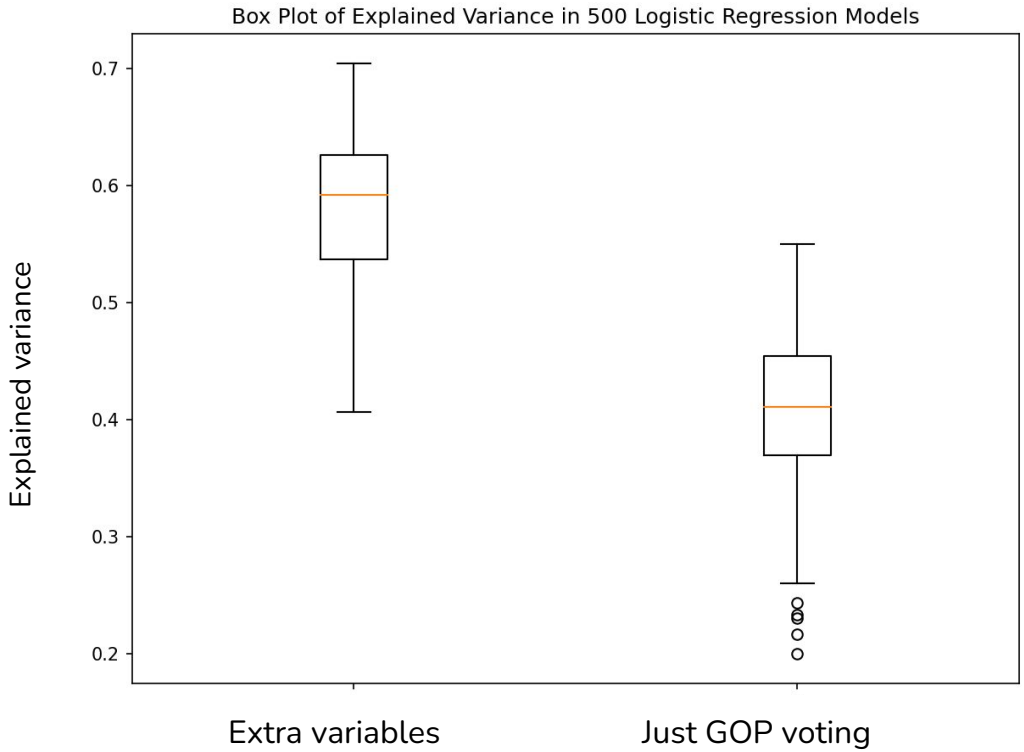
Summary

- Issues with Misinformation
  - CoVaxxy and Sparseness
- Exploratory Analysis
  - Correlation Matrix
  - PCA
- **Regression Models**
  - **Logistic Regression + LASSO**
  - **Random Forest**
- Classification Models
  - KNN
- Key Takeaways and Future Directions

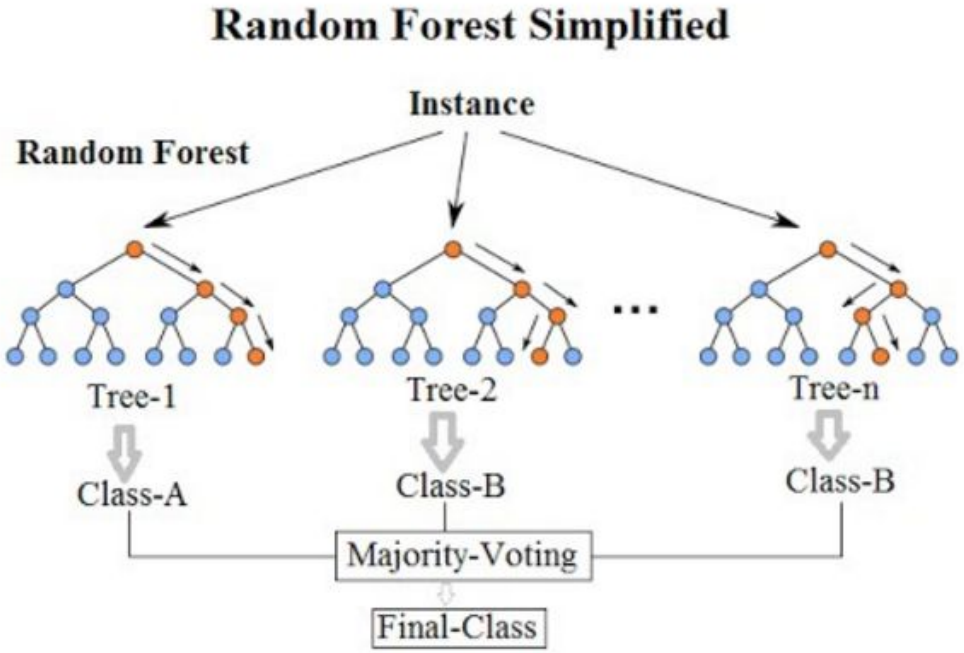
Logistic Regression Results



Controlling for CoVaxxy

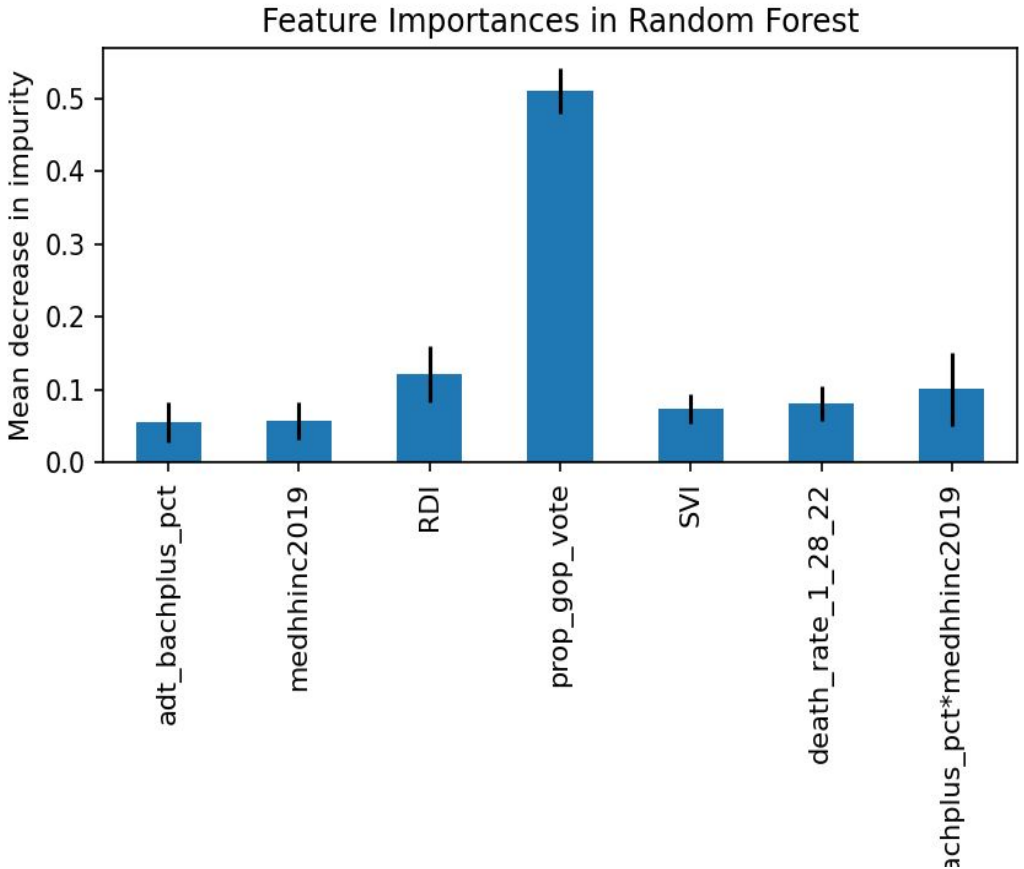


Random Forests

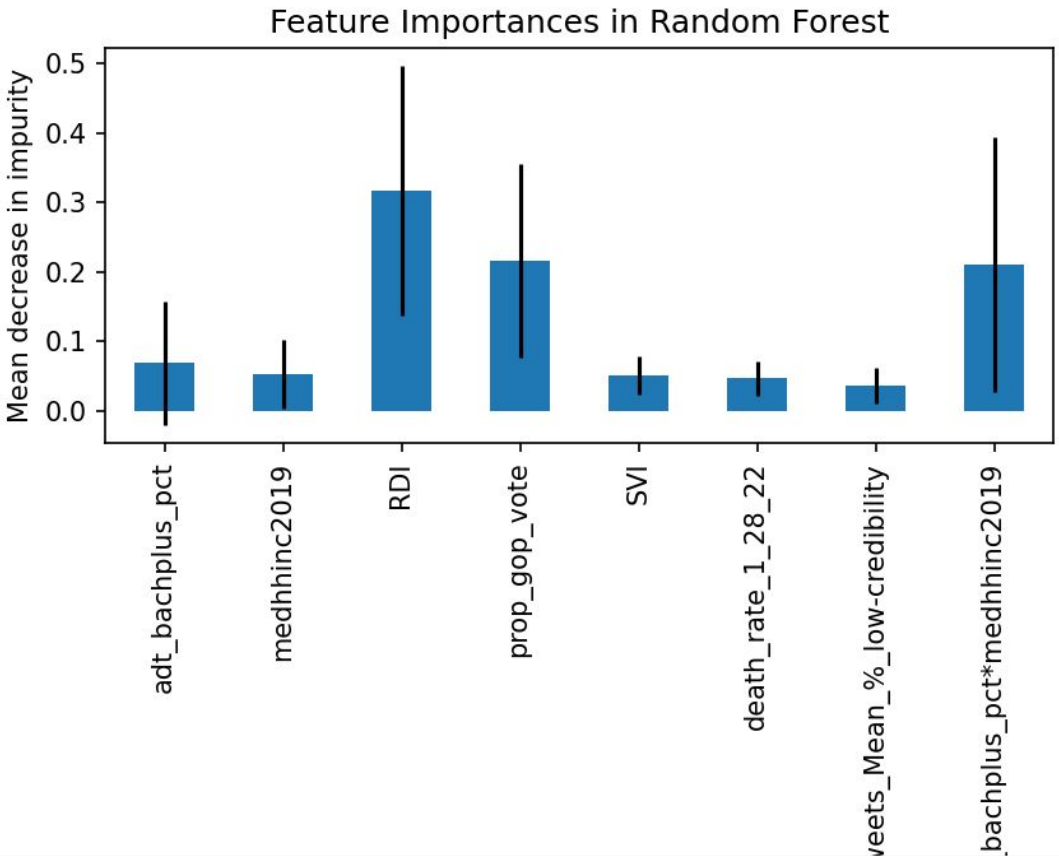


$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

Random Forest Results



Controlling for CoVaxxy



Summary

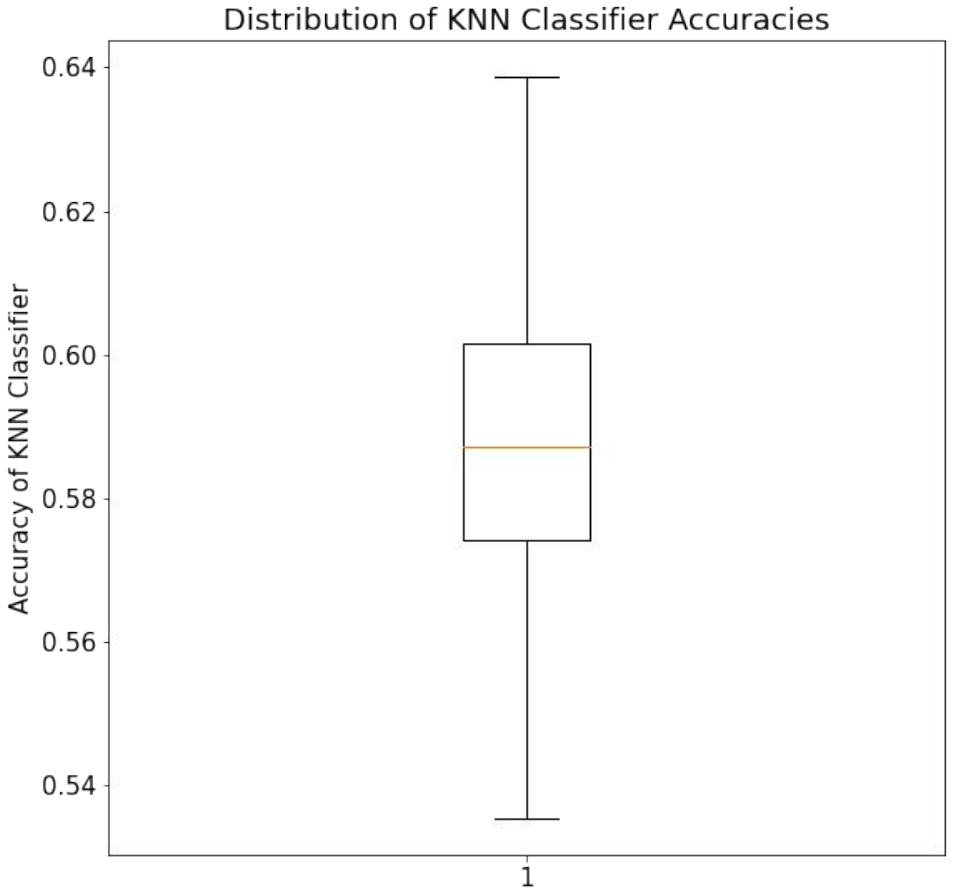
- Issues with Misinformation
  - CoVaxxy and Sparseness
- Exploratory Analysis
  - Correlation Matrix
  - PCA
- Regression Models
  - Logistic Regression + LASSO
  - Random Forest
- **Classification Models**
  - **KNN**
- Key Takeaways and Future Directions

k-Nearest Neighbors (KNN)

- The k-nearest-neighbors algorithm performs supervised classification of categorical data.
- New test data are compared to the test data which are “closest” in some metric.
- How do we treat vaccine update (continuous) as categorical?
  - We classify counties into four categories, according to which quartile of vaccine uptake they fall into.
- We classify according to:
  - Rural/urban distribution, religious diversity, education level, median household income, and proportion of GOP vote in the 2020 election.
- Estimator is trained using 20-fold cross validation on a randomly permuted dataset.

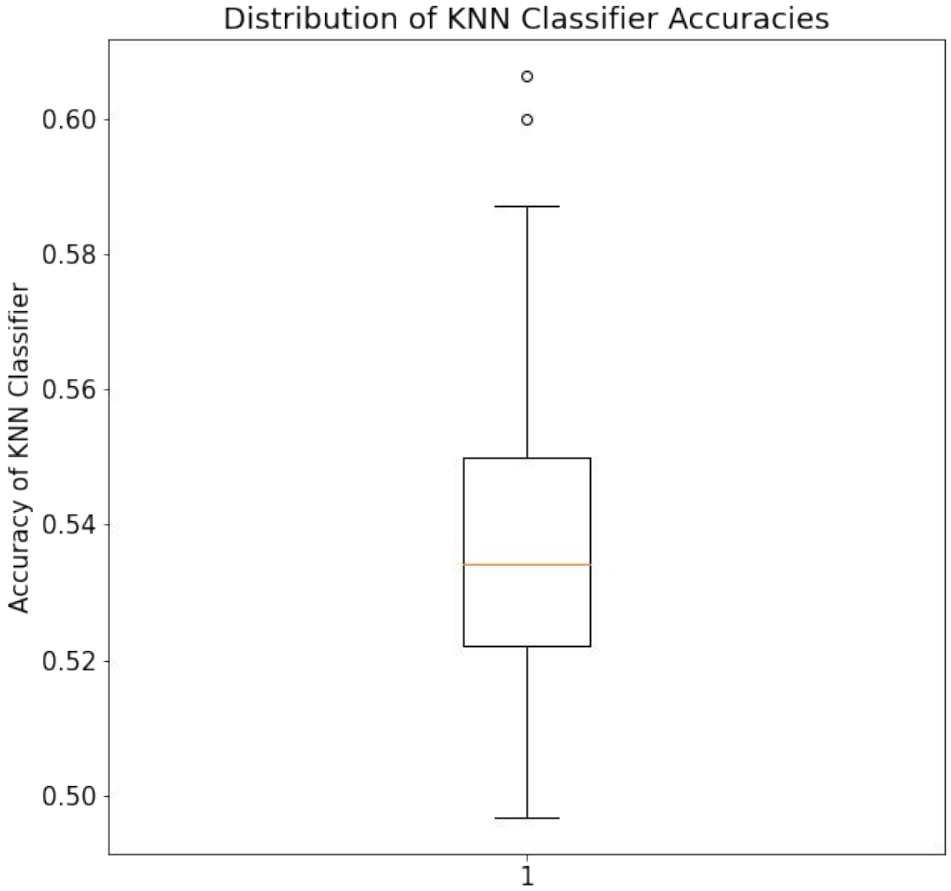
Results

- Accuracy has high variance over the randomly chosen permutation of the dataset.
- Mean accuracy is around 59% on the test data (taken over 100 independent training sessions).
- For comparison: considering only the proportion of the GOP vote, mean accuracy is around 54%.



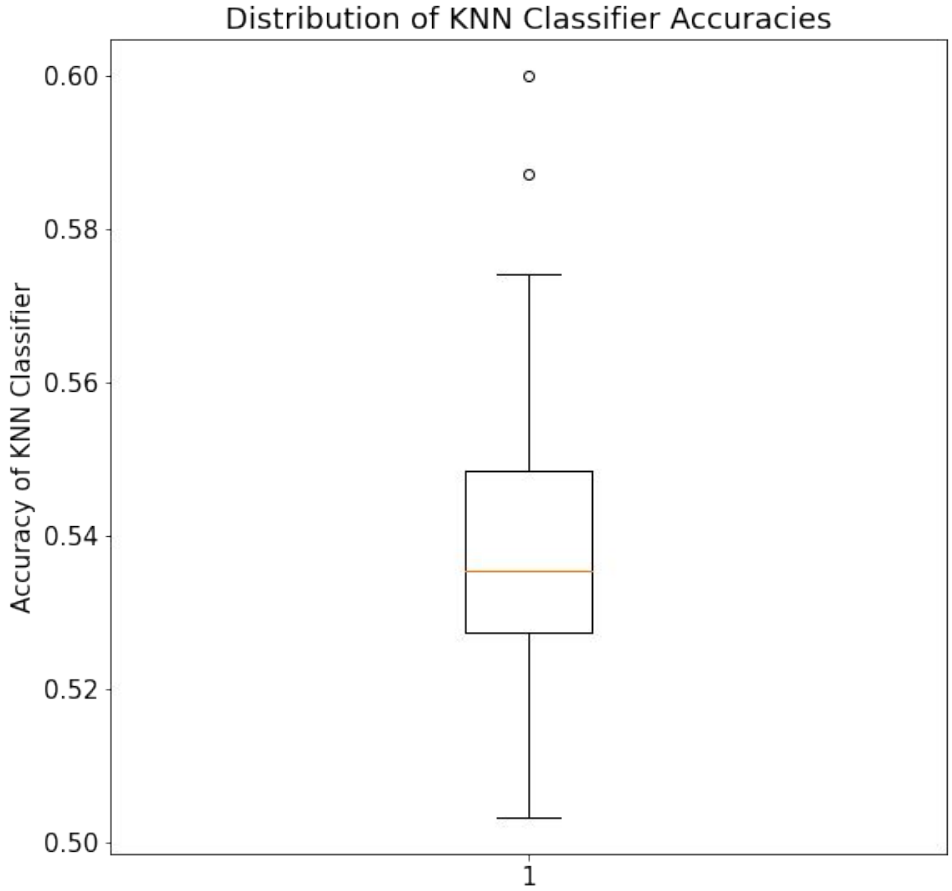
Results

- How to determine which variables determine the KNN classification?
- We omit each variable from the training data and record the loss of accuracy.
- Proportion of the GOP vote is most influential; if omitted, mean accuracy drops to about 50%.
- Omitting other variables causes no significant change in accuracy.

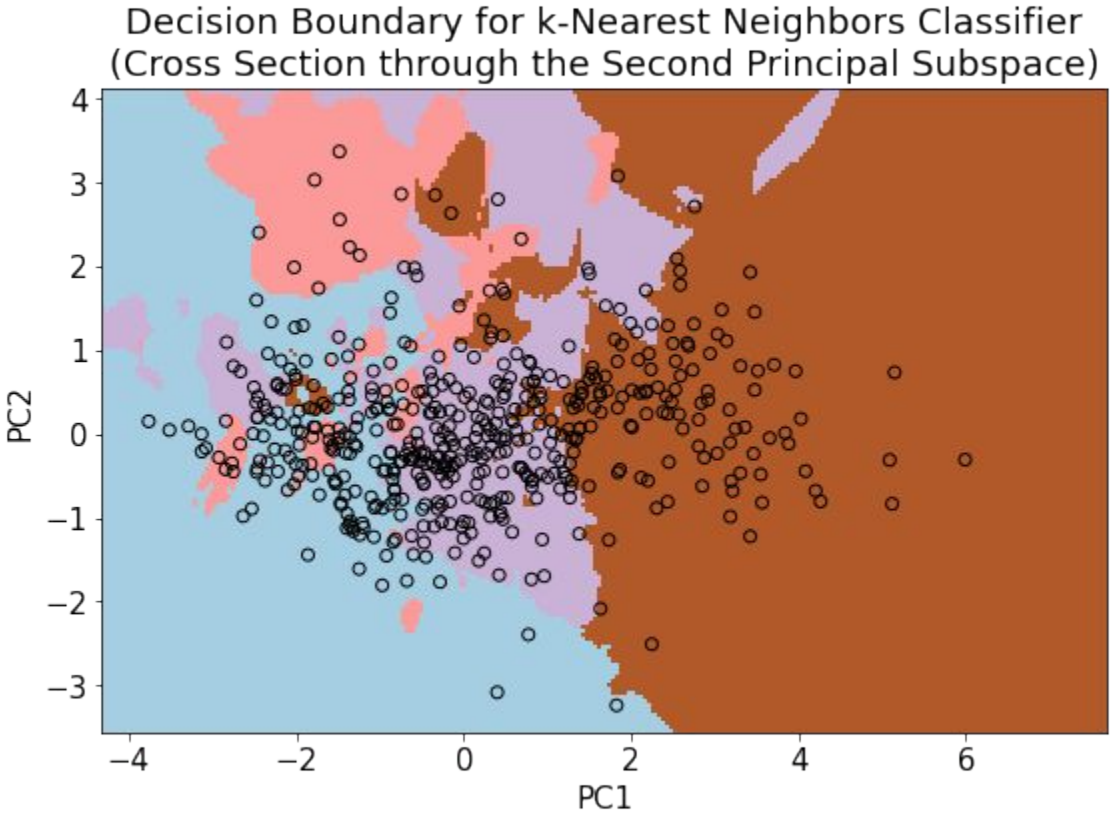


Results

- On the other hand, predicting using *only* the GOP vote has an accuracy of about 54%.



Visualization



Summary

- Issues with Misinformation
  - CoVaxxy and Sparseness
- Exploratory Analysis
  - Correlation Matrix
  - PCA
- Regression Models
  - Logistic Regression + LASSO
  - Random Forest
- Classification Models
  - KNN
- **Key Takeaways and Future Directions**

Conclusions

- **Voting patterns** account for the majority of the predictive power of our models
- There is evidence that one can improve on this state of affairs by considering religious diversity, education, and income as predictors
- “Secondary variables” are more predictive when considering **Democratic or purple counties** that nonetheless have low vaccination rates

Future Directions

- More localized analysis: careful consideration of outliers, applying algorithms to sub-samples with different vaccination rates
- Additional misinformation dataset: the Miller Center at Rutgers University has cultivated a database of anti-mask and anti-vaccination protests, and has found interesting predictors as well

Acknowledgement

- Funding:
  - Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University
  - Stavros Niarchos Foundation SNF Agora Institute, Krieger School of Arts and Sciences, Johns Hopkins University
- Mentors:
  - Professor Lauren Gardner, Civil and Systems Engineering, JHU
  - Ensheng (Frank) Dong, PhD Student, Civil and Systems Engineering, JHU
  - Haoji (Barry) Zeng, MSc. Student, Engineering Management, JHU
  - Dr. Sara Bertran De Lis Mas, Data Scientist at GovEx
  - Professor Fadil Santosa, Applied Mathematics and Statistics, JHU
  - Dr. Bryce Corrigan, Senior Statistician and Lecturer, SNF Agora Institute, JHU

References

1. The Mathematics Behind Principal Component Analysis, Dubey, A. 2018.
2. E-Learning Project SOGA: Statistics and Geospatial Data Analysis, Hartmann, K., Krois, J., Waske, B., 2018
3. The relationship between vaccination rates and COVID-19 cases and deaths in the USA, E. Dong & L. Gardner., JHU CSSE, 2021
4. CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines, DeVerna, M. R., Pierri, F., Truong, B. T., Bollenbacher, J., Axelrod, D., Loynes, N., Torres-Lugo, C., Yang, K.-C., Menczer, F., & Bryden, J., Proceedings of the International AAAI Conference on Web and Social Media, (2021)
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011
6. https://en.wikipedia.org/wiki/Alpine_County,_California
7. Random Forests. Wikipedia, collected 2022
8. Gini Impurity vs. Information Gain vs. Chi-Square. October 2021
9. https://i.imgflip.com/3g43lo.jpg


THANK
YOU