The dataset is collected by Jia el al. [1]. Each data point is a county (indexed by FIPS code) with statistics as follow,

1. The number of civilian labor force in 2012 and 2016 (Continuous variable);
2. Unemployment rate in 2012 and 2016 (Continuous variable);
3. Bachelor Rate in 2012 and 2016 (Continuous variable);
4. Median Income in 2012 and 2016 (Continuous variable);
5. Population in 2012 and 2016 (Continuous variable);
6. 2013 Urban Influence Codes (Discrete variable);
7. The economic dependence types (Category);
8. Election Results of Democrats and Republicans in 2012 and 2016 (Continuous variable).

Meanwhile, the dataset has the record of the connection of bordering counties.

In the original paper [1], the author uses this dataset to test the performance of their novel Graph Neural Networks. They select one statistic as the outcome; the remaining are features. Prediction accuracy is the only measurement.

In our project, we want to use GLM and other more interpretable statistical methods (e.g., random forests) to analysis this dataset. We have two plans of choosing the response variable: first, the voting rate of Democrats in 2016 with the sigmoid function on the output of GLM to get a percentage result as the response variable; second, using the number of votes actually obtained rather than the percentage as response. Then, the remaining statistics are input features (except Election Results of Republicans in 2016). Comparing with only seeking accuracy, we hope, by doing these, we can get a balance of accuracy and interpretation with deeper understanding of the relationship between election results and demographic statistics and the influence of bordering counties.

[1] Jia, Junteng, and Austion R. Benson. "Residual correlation in graph neural network regression." *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020.