Group 14: Yahan Chen, Yani Sun, Yuxin Liang, Jingyi Bai, Xilin Chen

Demographic Features and Sexual Offense Rate in United States

Research questions:

The purpose of the project is to find out whether the demographic features will affect sexual offense rate.

Subquestion:

1. Which of the eight features - police spending, teacher salaries, poverty rate, GDP per capita, education spending per pupil, drug overdose mortality rate, binge drinking rate, and bachelor degree attainment rate - is most strongly correlated with sexual offense rates?
2. How does the relationship between education spending and sexual offense rates vary across regions with different GDP per capita?
3. Is there a significant interaction effect between police spending and poverty rates on the sexual offense rates?
4. Find the best machine-learning model.

Variables:

Link to Data: https://uwmadison.box.com/s/98rfdhyw5yh2kj312523atlvuah8lgxk

Data is collected from: URC Crime Database, Bureau of Justice Statistics, US Census Bureau, National Center for Education Statistics, and Kaggle https://www.kaggle.com/code/marshuu/poverty-rate-in-the-us-animation/input

8 Features: police spending, primary and secondary school teacher average salary, poverty rate, GDP per capita, education spending per pupil, drug overdose mortality rate, binge drinking rate, and bachelor degree attainment rate. Data Size: 255 rows and 9 columns.

Our study's dataset is limited by a shorter observational timeframe and lacks the most recent data, with all data and variables ranging only from 2016 to 2020. Due to potential gaps in the initial data collection process, some variables might have missing data. However, we plan to address this issue at a later stage, implementing strategies to mitigate the impact of these data omissions and reduce any resulting distortion in our analysis.

Methods:

We would like to try Linear regression, decision tree, and logistic regression method, and compare which one has the most accuracy.

```
df = pd.read_csv('sexual_offense_data.csv', index_col=0)
df.head()
```

state	policing_correction_spend_per_capita	year	salary	poverty_rate	sex_rate	gdp_capita	edu_spending_per_pupil	drug_mortality	binge_drink_rate	b
Alabama	405.28	2016	49781.0	17.2	39.4	37158.0	9242.677695	16.2	17.69	
Alaska	954.48	2016	67443.0	9.9	141.9	63304.0	17509.975316	16.8	19.53	
Arizona	569.62	2016	45477.0	16.4	47.5	38940.0	7613.006435	20.3	18.50	
Arkansas	396.26	2016	48220.0	17.2	71.7	36502.0	9845.568548	14.0	16.71	
California	843.21	2016	72842.0	14.4	34.9	58974.0	11495.363449	11.2	19.98	

Notes: Since there are many sources and sharing them individually is not ideal and reflective of what we want to do. We have linked the organized data that can be cleaned further for modeling.