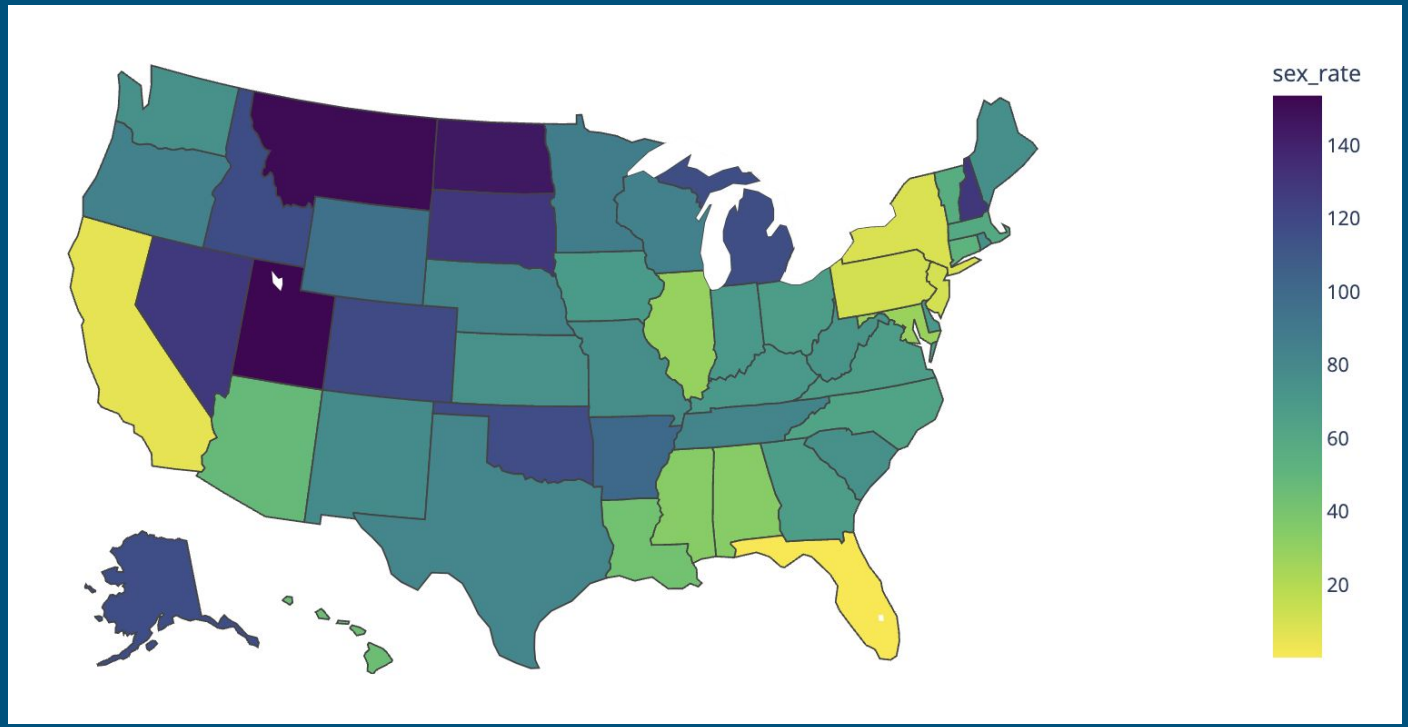
Sexual Offense Rate in U.S.

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

Heat map of sexual offense rate in US



Data set: 8 variables v.s. Sex Offense Rate

Police spending

Teacher salaries

Poverty rate

GDP per capita

Education spending per pupil

Drug overdose mortality rate

Binge drinking rate

Bachelor degree attainment rate

Research questions

1. **Which of the demographic features is most strongly correlated with sexual offense rates?**
    - Is there any relationship between 8 variables and sexual offense rate?

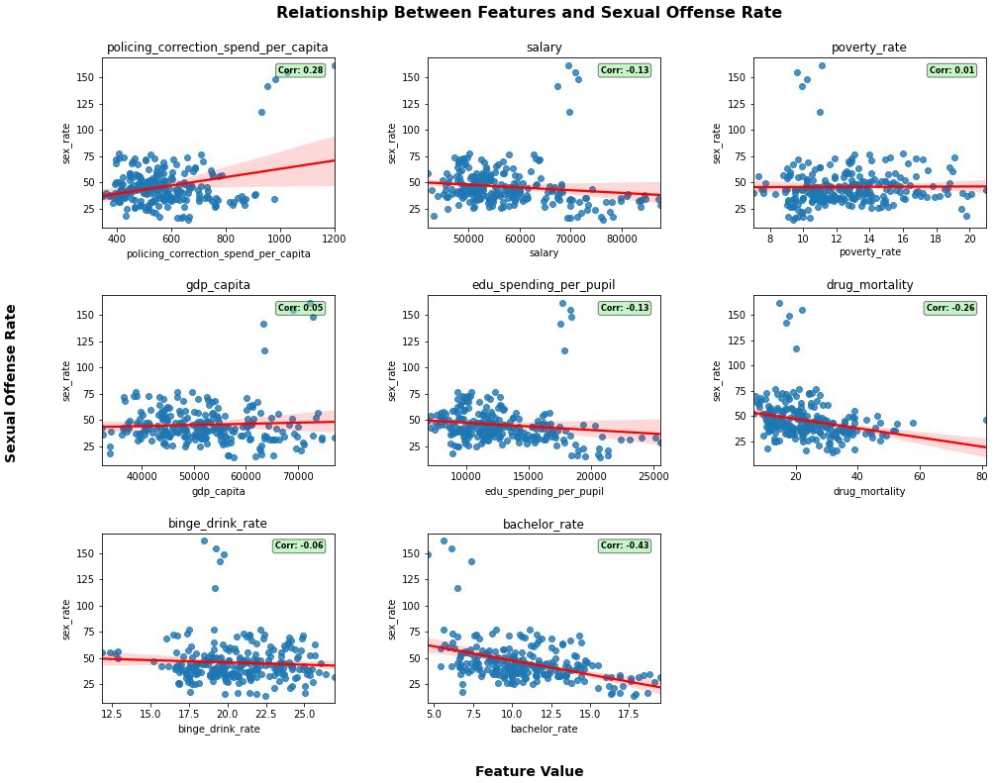
2. **Find the best machine-learning model.**

Data Processing

1. Use SimpleImputer() to fill in the missing data on binge_drinking_rate

2. Drop D.C. data due to multiple missing feature data

Linear Model




```
linear = LinearRegression()
linear.fit(X_train, y_train)
linear.score(X_test, y_test)
```

1. Correlation:
   - Policing spending
   - Bachelor rate

2. Linear model test:
   - Test score: -0.1609

KNN

```
k value = 1,scores = 0.840137846984906
```

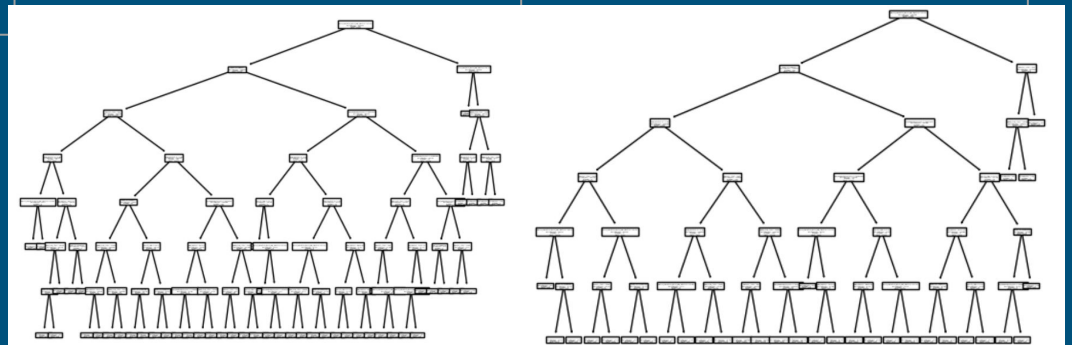
```
k value = 1,scores = 0.4822428637054106
```

High risk of overfitting with k=1

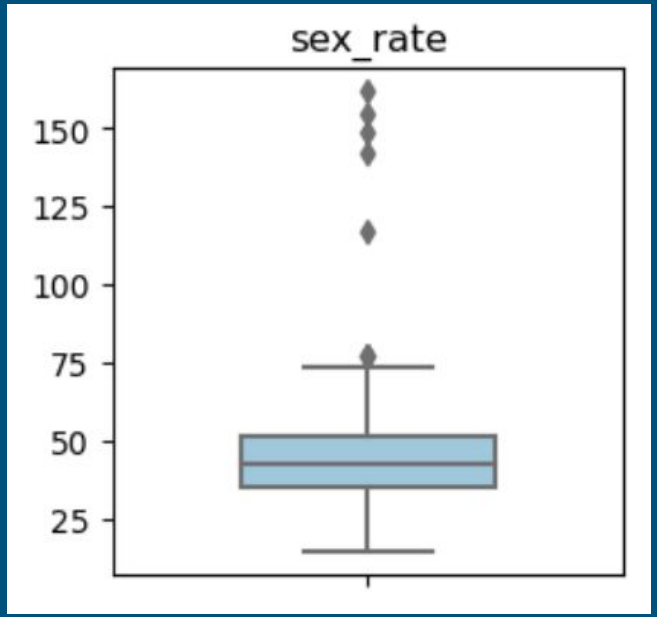
Assumption: **1. Feature Selection**

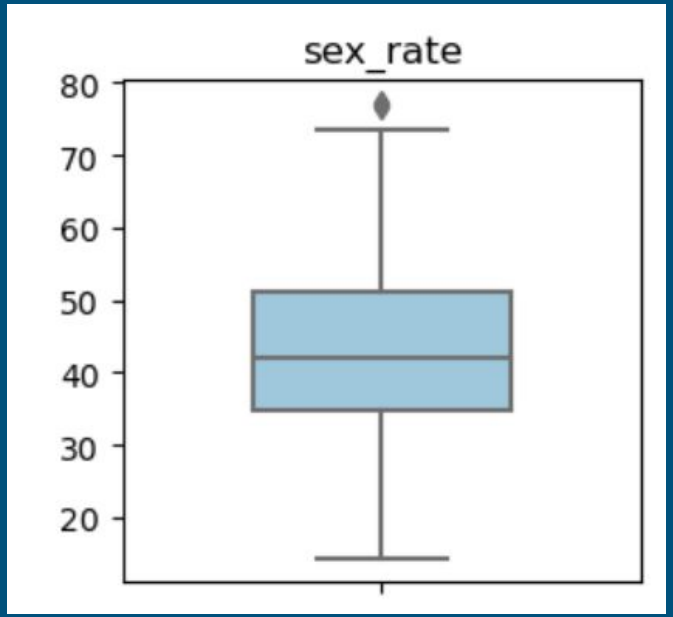
Decision Tree Regression

Splitting train, valid & test data:		
Random_state =	0	5
Grid Search & Cross-validation:		
Max_depth =	7	6
R^2 of Test Data =	(-) 5.0	0.814



Drop Outliers





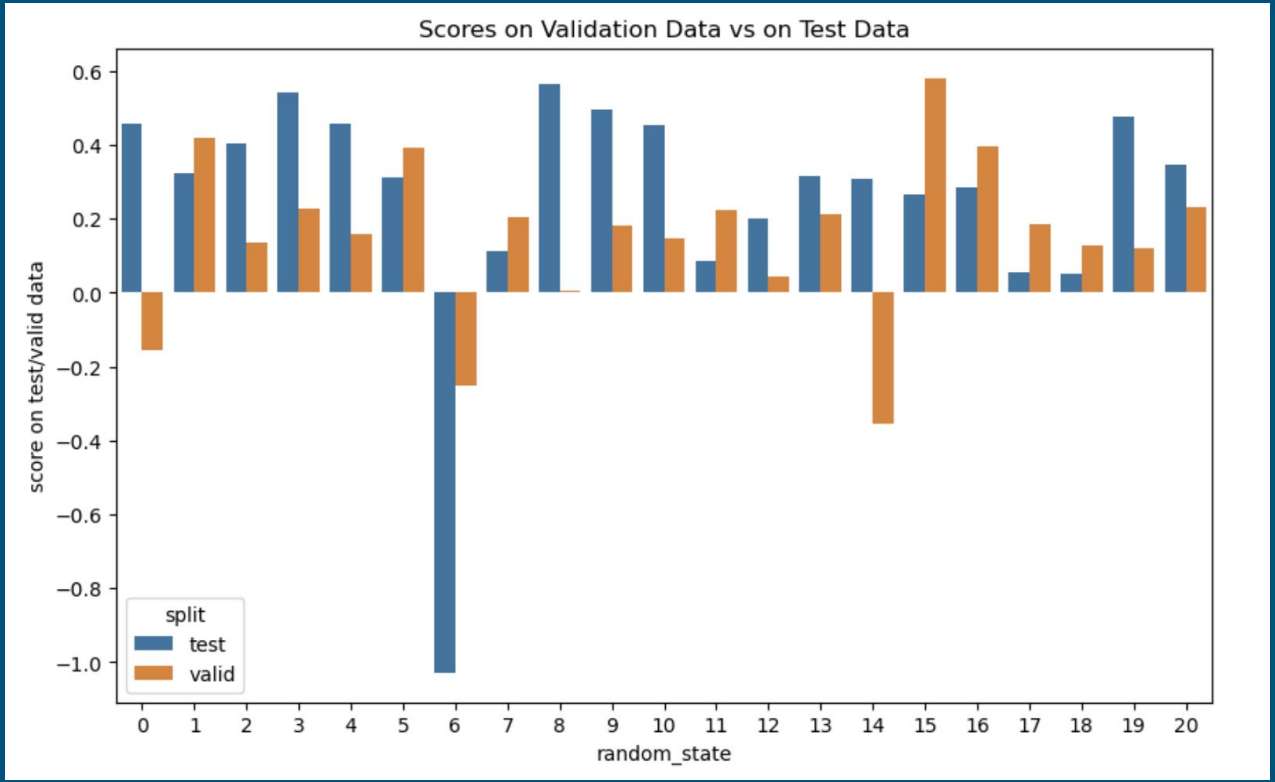
Linear and KNN

Accuracy score of linear model on data without outliers: 0.398 compare to -0.16

Accuracy score of KNN on data without outliers: 0.588 compare to 0.482

- k=1: new data point is based solely on the single nearest neighbor in the training set.

R^2 : Test different random states (Decision Tree)



Takeaways

1. Demographic features and predictive sexual offense rate are weakly correlated
2. Collinearity
3. Model is too simple
4. Data collection methodology
    a. Census data/FBI crime data underreporting.
5. Data splitting method

Thank you!
Question?