The first four methods below are all from AI fairness 360. More info can be read on them there, but more importantly for the development team there is code provided. I didn't look at the code very in depth, but it looks to be at least somewhat annotated and would be a good starting point.

https://aif360.mybluemix.net/

There was actually more unfairness in Compas' algorithm between sexes than races. It should be noted that the difference in race was just set up as Caucasian vs all other races, so there may still be heavier bias when looking closer. Sex was set up with Females being the privleged group. I mostly ignored the Theil Index for this because it measures inequality of benefit allocation between individuals, and was considered fair both before and after mitigation techniques were applied.

Also I wanted to note that all of these methods showed a great improvement in disparate impact, which is promising.

1) **Reweighing** was all-around shown to be very effective for reducing bias. There was no loss to accuracy and all of the statistical tests showed a sizeable reduction in bias, as well as putting the measures within "acceptable" ranges that would indicate fairness for both Sex and Race.

   Brief Description: Weights the examples in each (group, label) combination differently to ensure fairness before classification.

2) **Optimized Pre-processing** had mixed results. It did a worse job with sex, but better with race.

   Brief Description: Learns a probabilistic transformation that can modify the features and the labels in the training data.

   - Regarding Sex: Overall accuracy actually got slightly worse, (66% to 65%). Bias was still reduced across the board for all tests, but not to a measure that was considered fair.
   - Regarding Race: Overall accuracy was improved from 66% to 67%. Bias was reduced to acceptable levels for all tests.

3) **Adversarial Debiasing algorithm:** This was similar to the last method, except flipped

   Brief Description: Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

- Regarding Sex: Accuracy stayed the same, and all tests showed a reduction of bias to acceptable levels.
- Regarding Race: Accuracy was reduced to 65%. Bias was still reduced for all tests, however only 1 of them reached the threshold to be called “fair”.

4) **Reject Option Based Classification:** This method was effective for reducing bias across the board for both sex and race.

   Brief Description: Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

   - Regarding Sex: Accuracy was again reduced to 65%. Also, while this method did reduce bias for all tests, according to the equal opportunity difference test it actually favored Females over Males a little bit outside of the range considered to be fair.
   - Regarding Race: This was all around effective. One thing that was strange was that it did significantly worse on the Theil Index. It was still in the fair range, but went up from 0.21 to 0.29 (lower scores are better). I’m not very concerned about this.

5) IGD method: This method would be more work to implement and wrap our heads around, but it was shown to be significantly more effective than both EOP and ROC methods. This is the algorithm described in “Bias Mitigation Post Processing”. There is a separate document in the statistics folder breaking this method down:

https://docs.google.com/document/d/1LbmqGqN2f7G-IeQLteX3p68RKHIlbXj8fOyu3uyVgYE/edit

6) Prejudice Remover: This method adds something called a discrimination-aware regularization term to the learning objective. Meaning this also helps with regarding race and sex. The link below is the code that explains this method a little more in depth.

https://github.com/Trusted-AI/AIF360/blob/master/aif360/algorithms/inprocessing/prejudice_remover.py

The following methods are from "Using algorithms to address trade-offs inherent in predicting recidivism"

This article uses a sample of federal probationers assessed by Post Conviction Risk Assessment (PCRA). For their methods they only selected those in the sample that were identified as "black" or "non-hispanic white"

1. **Conservative matching approach:** Isolate effect of race without creating non-representative groups. "We randomly matched each Black participant to a White participant on age, sex, offense, and district using ccmatch in STATA (Cook, 2015)." This ccmatch function seems to be similar to the match() function in R.
2. **Race eliminated algorithm:** The "race eliminated" algorithm was generated by regressing each of the 15 PCRA items on race and retaining the residual values, and then using the residual PCRA item scores as predictors of violent reoffending (see Gottfredson & Jarjoura, 1996; Kleinberg et al., 2018). The "criminal history discount" algorithm was calculated by first reducing Black supervisees' scores on this domain by 23% (based on a meta-analytic estimate of the effect of race on police decisions to arrest; Kochel, Wil-son & Mastrofski, 2011), and then using this revised domain and the remaining PCRA items as predictors.
    a. Couldn't find a complete list of the 15 predictors they used, but will keep searching before Tuesday.
    b. Simple and Multiple Linear Regression in Python
3. **Describe Predictive Use of Algorithms and Association with Race:** Use AUC and correlation with black race. The "race eliminated" algorithm was the one with least correlation to black race, out of the 5 algorithms evaluated.
    a. Finding AUC in Python : Under section "ROC Curves and AUC in Python"