MDPI

Article

Full title of the paper (Capitalized)

[1] Statistical and Data Sciences Smith College Northampton, MA 01063;
[2] ; aacharya@smith.edu
[3] ; dcaravela@smith.edu
[4] ; ekim89@smith.edu
[5] ; ekornberg@smith.edu
[6] ; enesmith@smith.edu
* Correspondence: leutnant@fh-muenster.de; Tel.: +XX-000-00-0000.
† Current address: Updated affiliation
‡ These authors contributed equally to this work.

Version March 6, 2022 submitted to Water

Abstract: A variety of disciplines use risk assessment instruments to help humans make data-driven decisions. Northpointe, a software company, created a risk assessment instrument known as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). COMPAS uses various behavioral and psychological metrics related to recidivism to assist justice systems in assessing a defendant's potential recidivism risk. ProPublica published an article which concludes that the biases in the criminal justice system are reflected in the COMPAS software. Using a human rights framework adopted from the organization Women at the Table, we use various debiasing algorithms to analyze both sides of the argument between Northpointe and ProPublica and determine the level and extent of racial bias in the COMPAS algorithm.

1. Version

This Rmd-skeleton uses the mdpi Latex template published 2019/02. However, the official template gets more frequently updated than the 'rticles' package. Therefore, please make sure prior to paper submission, that you're using the most recent .cls, .tex and .bst files (available here).

2. Introduction

Women @ the Table, the sponsor organization for this project, is "a growing, global gender equality & democracy CSO based in Geneva, Switzerland focused on advancing feminist systems change by using the prism of technology, innovation & AI exercising leverage points in technology, the economy, sustainability & democratic governance" [1]. We have been asked to collaborate on their AI & Equality [2] initiative, tasked with debiasing the COMPAS algorithm [3] and producing a corresponding data story that will be added to their library.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm was created by the private for-profit company Northpointe, also known by its parent company Equivant [**?**], to predict defendants' risk of recidivism. It generates a score that classifies defendants' risk of recidivism as either low, medium, or high [4]. Jurisdictions across the United States use the COMPAS risk assessment instrument. In 2016, ProPublica published a piece that analyzed the methods and algorithms used by Northpointe to uncover racial biases in defendants' scores [4]. Their analysis looks at the distribution of decile COMPAS scores among Black and white defendants. ProPublica concludes, after using a statistical parity metric on false positives or false negatives, that the algorithm is racially biased [5]. Northpointe denies the allegations of racial bias, conducting their own analyses

based on different statistical parity metrics [6]. Since then, the two parties have had several exchanges, maintaining their original arguments. Currently, the two parties' disagreement focuses on pretrial Risk Assessment Instruments (RAIs). ProPublica maintains that there are biases in the outcome values, protected attributes, and covariates during Northpointe's data processing phase. ProPublica accounts for these biases in their analyses, and Northepointe's response, highlights how ProPublica did not account for base rates of recidivism in their analysis, which are important to understand initial percentages without the presence of other information.

Our project builds on Women at the Table's various debiasing algorithms to conduct our own analyses on the COMPAS dataset. Based on this analysis, we employ a human rights framework to contribute to the ongoing ProPublica and Northpointe debate and investigate whether or to what extent there is racial bias in the COMPAS algorithm. With a solid understanding of the two sides, we aim to pinpoint the shortcomings of both arguments and correct them in our analyses. We will use various machine learning algorithms including logistic regression, cross validation, and lasso and ridge techniques to choose models. We will summarize our results using the JupyterNotebook framework from Women at the Table, to be used by members of the organization to teach in a workshop setting. We hope that our findings will highlight the importance of checking statistical analyses using varied methods and contribute to the ongoing discussion of the effects of machine biases in the justice system.

3. Data

The data we are using for this project is the COMPAS General Recidivism Risk Scores dataset from ProPublica used in their original analysis [**?**]. The COMPAS dataset we are using is from AI Fairness 360 (AIF360) toolkit [7] which does the same initial preprocessing as ProPublica. The raw data has 6,167 rows and each row represents an arrest charge for a defendant. ProPublica's COMPAS data includes the defendant's age, race, sex, what they were charged with, and whether or not the defendant ultimately recidivated within a two-year period after their arrest. For the purposes of our project, which endeavors to evaluate the differing effects of the COMPAS algorithm on white defendants and Black defendants, we have filtered the data to only include individuals whose race is listed as Caucasian or African-American. Our data therefore has 5,723 rows (Figure 1), with the below distributions of race (Figure 2), age (Figure 3), and two year recidivism rate (Figures 4 & 5).

```
knitr::include_graphics("/Users/elisabethnesmith/womenatthetable/images/table_snippet.png")
```

```
knitr::include_graphics("/Users/elisabethnesmith/womenatthetable/images/race_bar_plot.png")
```

```
knitr::include_graphics("/Users/elisabethnesmith/womenatthetable/images/race_age_plot.png")
```

```
knitr::include_graphics("/Users/elisabethnesmith/womenatthetable/images/recid_bar_plot.png")
```

```
knitr::include_graphics("/Users/elisabethnesmith/womenatthetable/images/race_recid_plot.png")
```

4. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

4.1. Subsection Heading Here

Subsection text here.

		sex	age	age_cat	
id	**sex**				
1	**0**	Male	34	25 - 45	Afr Ame
2	**0**	Male	24	Less than 25	Afr Ame
4	**0**	Male	41	25 - 45	Cauc
6	**1**	Female	39	25 - 45	Cauc
7	**0**	Male	27	25 - 45	Cauc

Figure 1. This table is a snippet the data set we will be using.

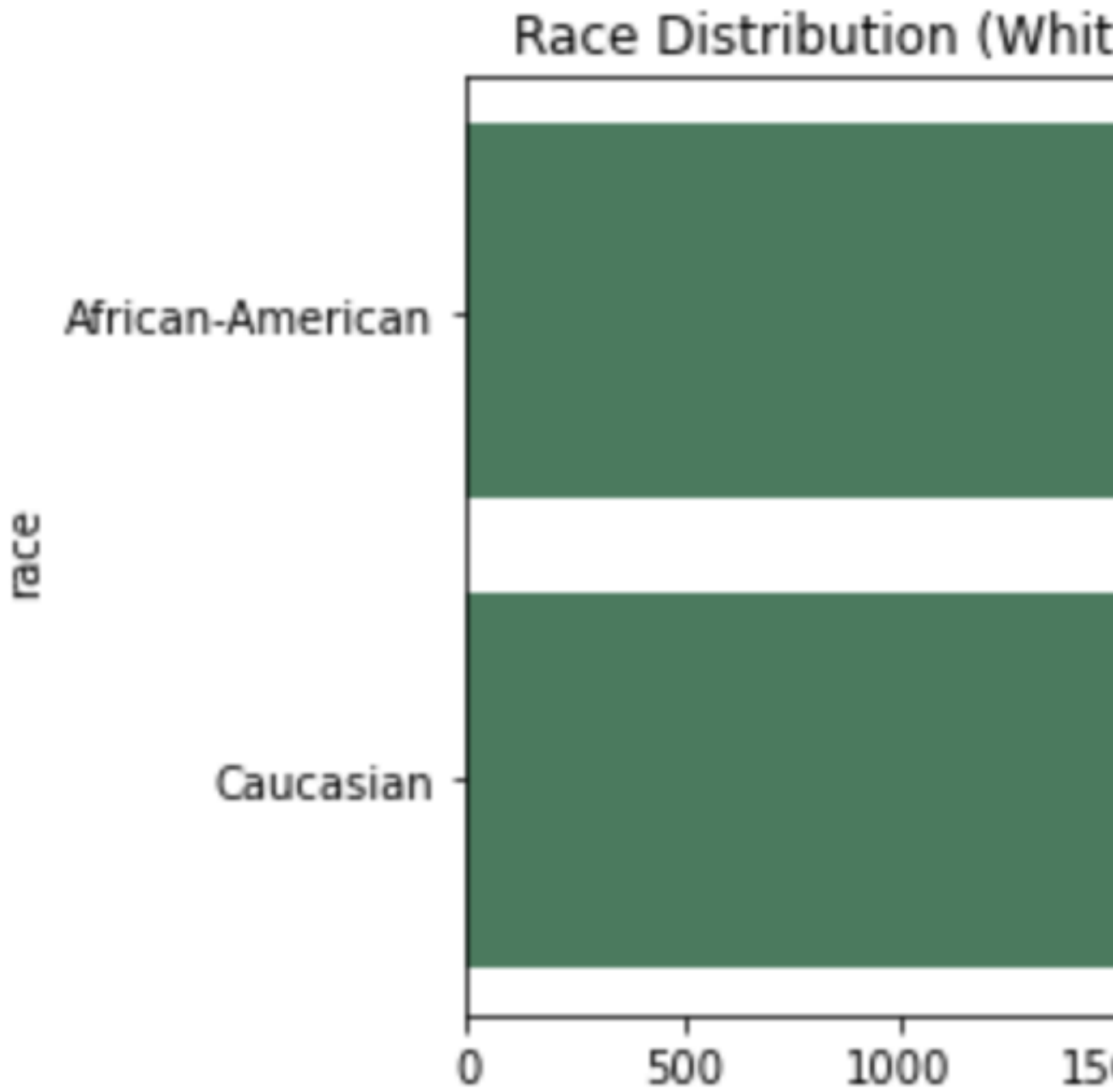

Figure 2. This plot shows the distribution of defendant races.

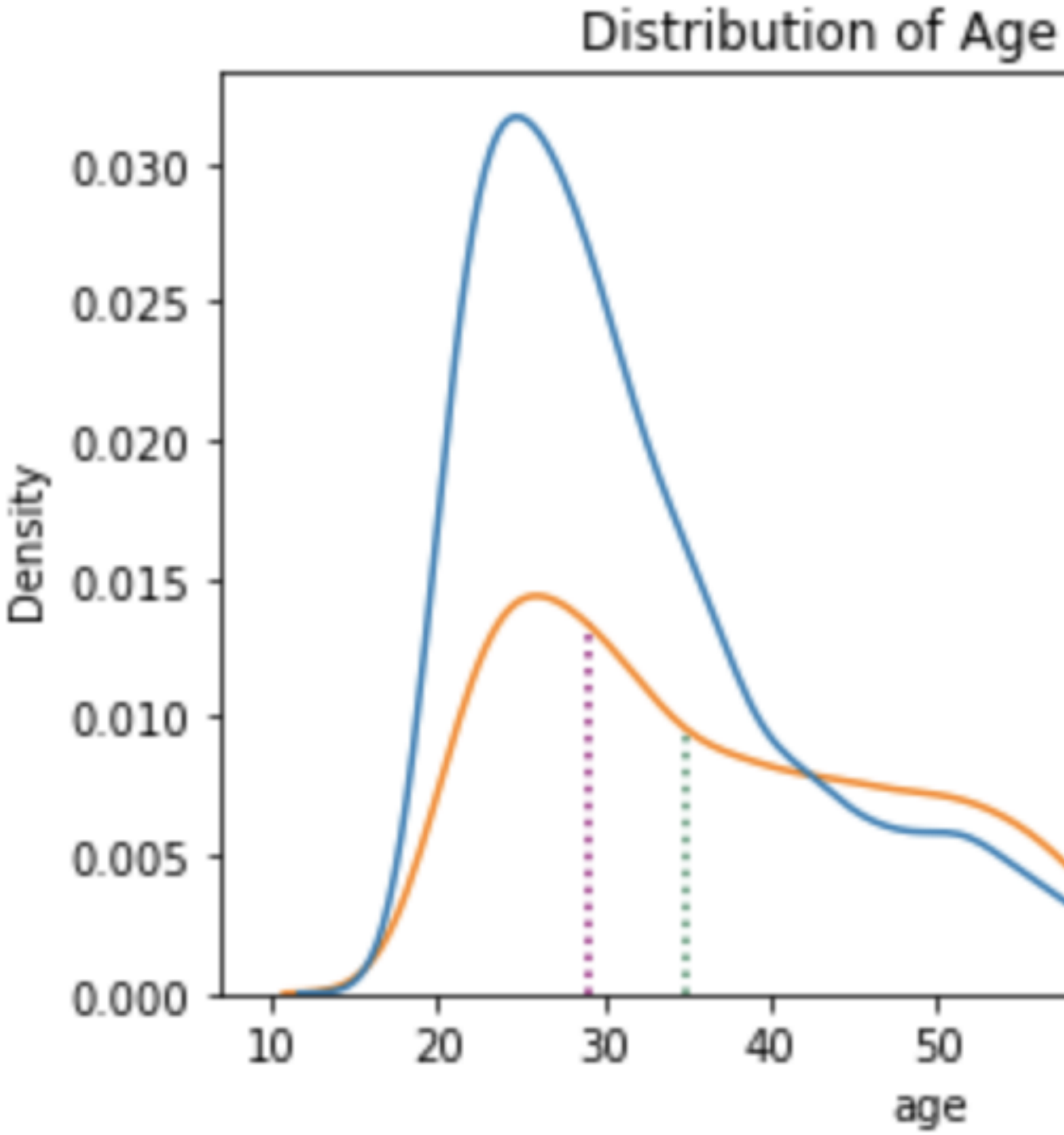

Figure 3. This plot shows the distribution of defendant ages by race.

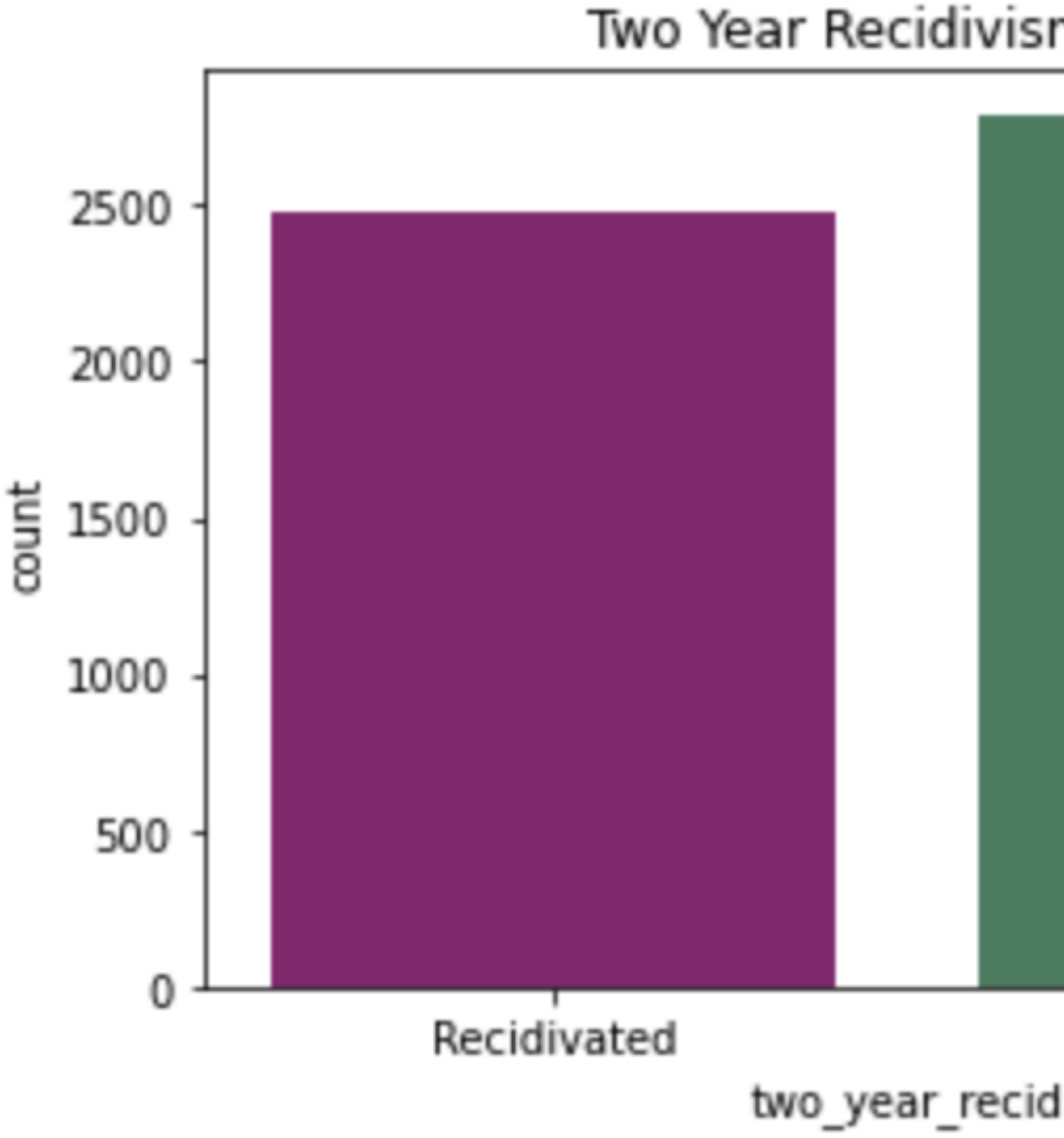

Figure 4. This plot shows the distribution of two year recidivism outcomes.

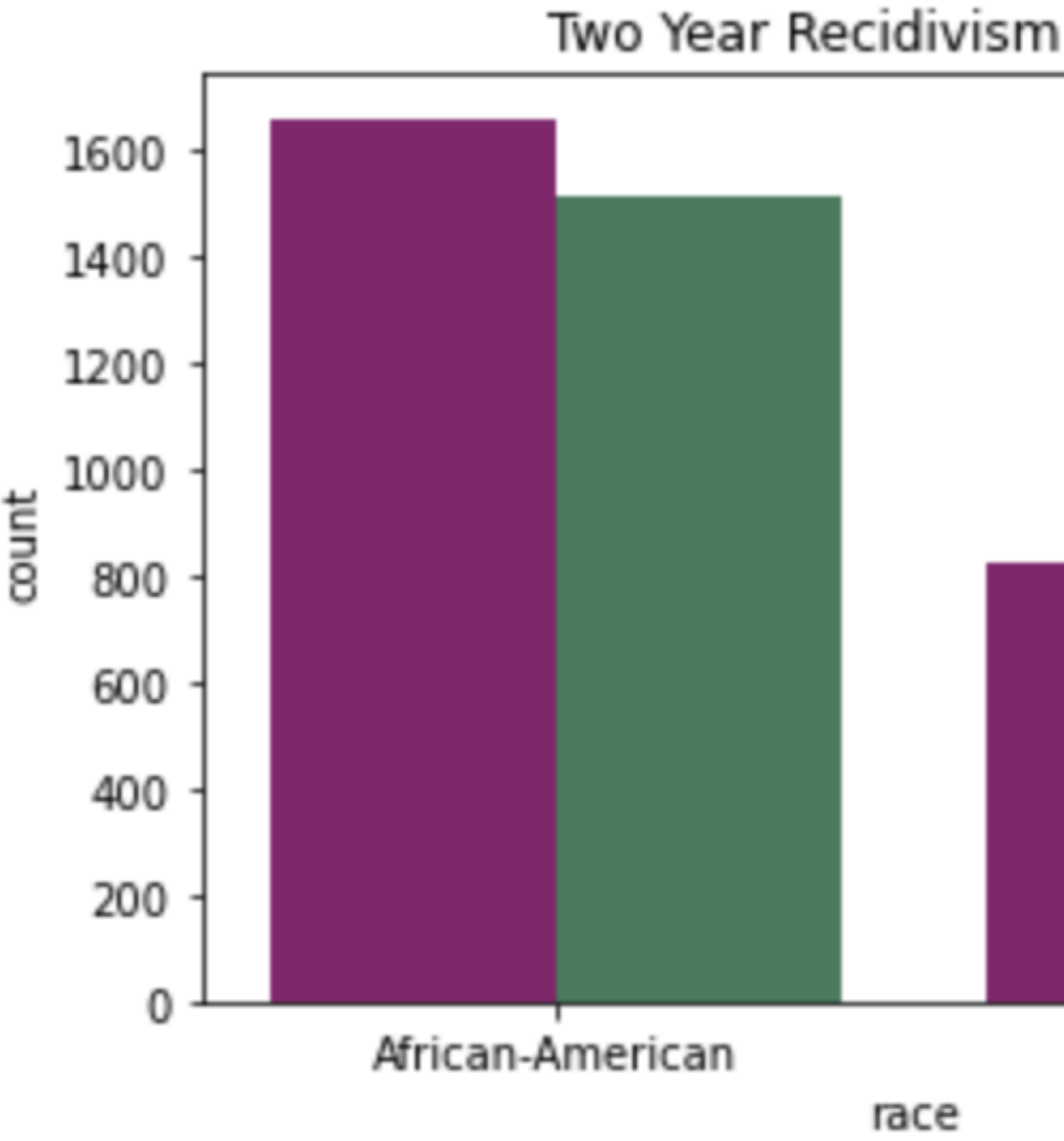

Figure 5. This plot shows the distribution of two year recidivism outcomes by race.

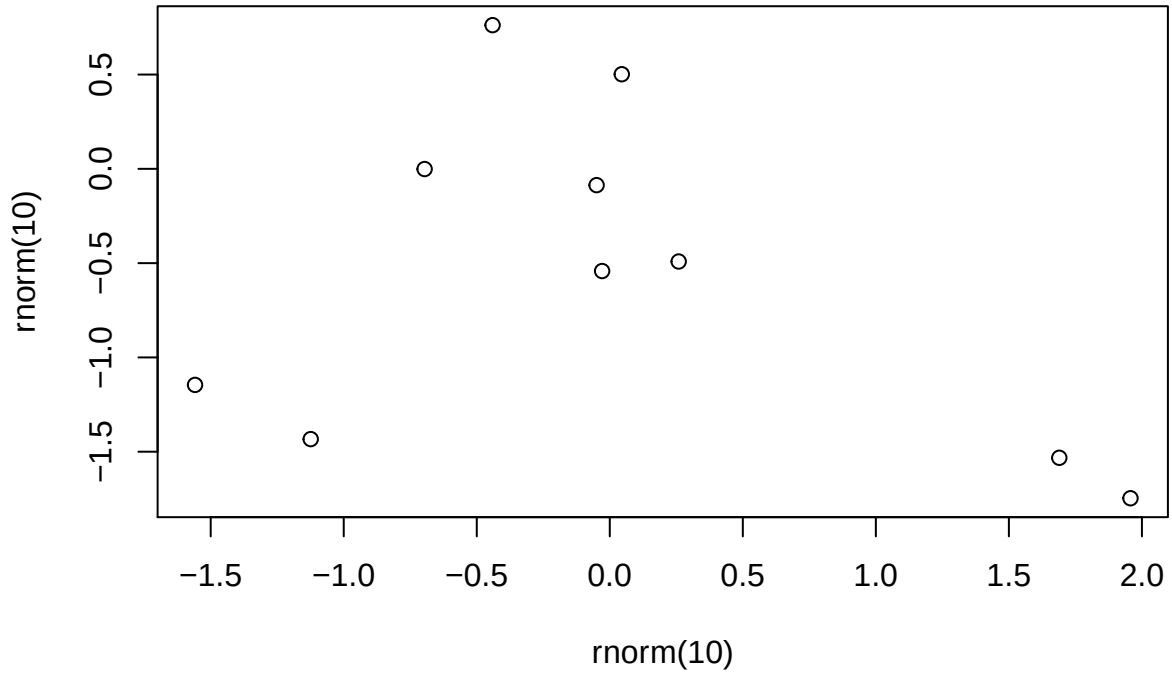

Figure 6. This is a plot.

Figure 7. This is a figure, Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

4.1.1. Subsubsection Heading Here

Bulleted lists look like this:

- First bullet
- Second bullet
- Third bullet

Numbered lists can be added as follows:

1. First item
2. Second item
3. Third item

See Figure 6 below.

```
plot(rnorm(10), rnorm(10))
```

The text continues here.
All figures and tables should be cited in the main text as Figure 7, Table 1, etc.
Please see Table 1.

```
x <- tibble::tribble(~‘Title 1‘, ~‘Title 2‘, ~‘Title 3‘,
"entry 1", "data", "data",
"entry 2", "data", "data"
)

knitr::kable(x, caption = "This is a table caption. Tables should be placed in the main text near to
```

Table 1. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Title 1	Title 2	Title 3
entry 1	data	data
entry 2	data	data

This is an example of an equation:

$$\mathbb{S} \tag{1}$$

Example of a theorem:

Theorem 1. *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:
Example of a proof:

Proof of Theorem 1. Text of the proof. Note that the phrase ‘of Theorem 1’ is optional if it is clear which theorem is being referred to. □

The text continues here.

5. Discussion

Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

6. Conclusion

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

7. Bibliography

Angwin *et al.* [4] Bao *et al.* [8] Barocas *et al.* [9] Baumer *et al.* [10] equivant [6] Gebru *et al.* [11] Hardin *et al.* [12] James *et al.* [13] Knight [14] Kypraiou [15] Larson and Angwin [16] Larson *et al.* [5] noa [2] noa [3] noa [7] noa [17] noa [1] Vartan [18]

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text. For example, explanations of experimental details that would disrupt the flow of the main text,

but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix B

All appendix sections must be cited in the main text. In the appendixes, Figures, Tables, etc. should be labeled starting with 'A', e.g., Figure A1, Figure A2, etc.

References

1. women@thetable.
2. <AI & Equality>: A Human Rights toolbox.
3. aif360.datasets.CompasDataset — aif360 0.4.0 documentation, 2018.
4. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. ProPublica, May 23, 2016, 2016.
5. Larson, J.; Mattu, S.; Kirchner, L.; Angwin, J. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* **2016**, *9*, 3–3.
6. equivant. Response to ProPublica: Demonstrating accuracy equity and predictive parity, 2018.
7. Trusted-AI / AIF360, 360.
8. Bao, M.; Zhou, A.; Zottola, S.; Brubach, B.; Desmarais, S.; Horowitz, A.; Lum, K.; Venkatasubramanian, S. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *arXiv preprint arXiv:2106.05498* **2021**.
9. Barocas, S.; Hardt, M.; Narayanan, A. Fairness in machine learning. *Nips tutorial* **2017**, *1*, 2.
10. Baumer, B.S.; Kaplan, D.T.; Horton, N.J. *Texts in Statistical Science: Modern Data Science with R*; Chapman and Hall/CRC, 2017.
11. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Iii, H.D.; Crawford, K. Datasheets for datasets. *Communications of the ACM* **2021**, *64*, 86–92.
12. Hardin, J.; Hoerl, R.; Horton, N.J.; Nolan, D.; Baumer, B.; Hall-Holt, O.; Murrell, P.; Peng, R.; Roback, P.; Temple Lang, D.; others. Data science in statistics curricula: Preparing students to "think with data". *The American Statistician* **2015**, *69*, 343–353. doi:10.1080/00031305.2015.1077729.
13. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Vol. 112, Springer, 2013.
14. Knight, C. Automated Decision-making and Judicial Review. *Judicial Review* **2020**, *25*, 21–27.
15. Kypraiou, S. What is Fairness?, 2021.
16. Larson, J.; Angwin, J. Technical response to Northpointe. *ProPublica, July* **2016**, *29*.
17. propublica/compas-analysis, 2022. original-date: 2016-05-21T03:36:35Z.
18. Vartan, S. 1st Greek ACM-W Chapter Winter School | Fairness in AI | Day 1, 2022.