Readme for Reproducibility submission of paper

“DataPrism: Exposing Disconnect between Data and Systems”

Sainyam Galhotra, Anna Fariha, Raoni Lourenço, Juliana Freire, Alexandra Meliou and Divesh Srivastava

This document provides details of the different experiments, the script to generate their plot and the description of the conclusion from the plot.

A. **Link to paper:** https://dl.acm.org/doi/abs/10.1145/3514221.3517864
B. **Link to Code and Scripts:** https://github.com/sainyam/DataPrism.git
C. **Readme to run and interpret the plots:** https://github.com/sainyam/DataPrism/tree/main/reproducibility/readme.pdf

Note: The runtime results for the experiments may not match, but the plot trends should reproduce. Please refer to the README file (https://github.com/sainyam/DataPrism/tree/main/reproducibility/readme.pdf) to verify the key conclusions.

Figure 6

- Script : bash Figure6.sh
- Time taken: 12 hrs
- Conclusion:
  - DataPrism requires the least number of interventions
  - Anchors requires the maximum number of interventions
  - GrpTest is 2nd best, but it does not run for 3 cases.

Figure 7

- Script : bash Figure7.sh
- Time taken: 4 hrs
- Conclusion:
  - DataPrism requires the least number of interventions as compared to the other two variations.

Figure 8 and 9

- Script : bash Figures_8_and_9.sh
- Time taken: 6 hrs
- Conclusion:
  - Figure 8: DataPrism requires the least number of interventions
  - Figure 8: Anchors requires the most number of interventions

- Figure 9: Median number of interventions reduce with increasing threshold.