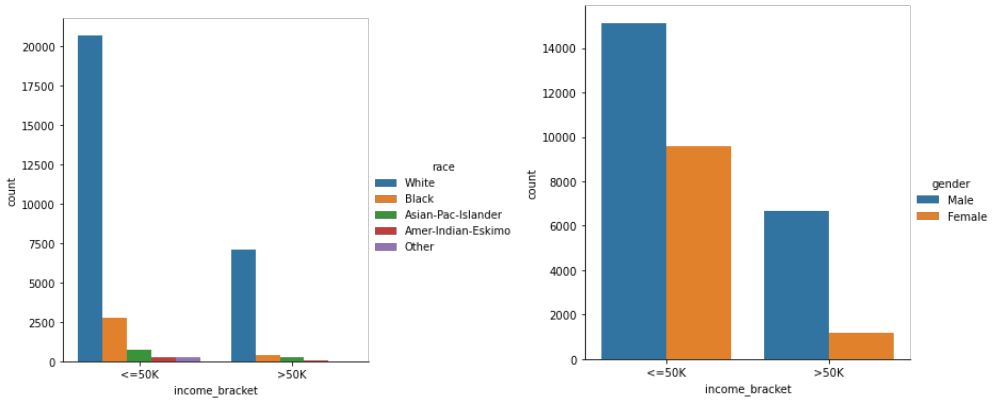
School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2023)

Week 12

1. We have a dataset containing demographic and income data from United States in 1994. We want to judge the fairness of a classifier we have trained on it. The data set consists of about 48,000 individuals, where each instance X is characterized through a range of 14 demographic attributes (gender, origin, education, race, occupation, etc.). The target variable Y is the income of the person (>50K or <=50K). To give you a snapshot of the training data, we have plotted the distribution of income across different genders and races.



(i). Discuss the following concepts in the context of this data set.

a) Historical Bias

b) Demographic disparity

c) Using the system in the context of (1) a bank which wants to use a model trained on this data for predicting credit ratings; and (2) a government institution in Australia which has access to the features of the Adult for a small population of Australians and wants to predict their income based on it.

(ii). You are asked to develop an income classifier that is fair with respect to the protected attribute *gender*. Your boss is a big believer in logistic regression classifiers and asks you to apply this particular classifier architecture with no modification. What approach(es) could you take to still test/improve the performance of your classifier?

2. Using the dataset in question 1, assume we selected *gender* as our protected attribute. We trained our classifier and observed the following outcomes. The label y=1 means "income >50K", and y=0 means "income <=50K.

P(ŷ=1|A=f)	P(ŷ=1|A=m)	P(ŷ=1|Y=1, A=f)	P(ŷ=1|Y=1, A=m)	P(Y=1|ŷ=1, A=f)	P(Y=1|ŷ=1, A=m)	P(Y=1|ŷ=1)	P(ŷ=1| Y=1)
0.81	0.75	0.80	0.86	0.73	0.74	0.74	0.85

(i). Name each of the statistics and provide a formula for its measurement. Be sure you understand the intuition / connection behind the statistical notion and its metric.

(ii). For each of the following criteria, decide whether the classifier meets this criterion.

a) Group Fairness (Demographic parity)

b) Equal opportunity

c) Predictive parity