(**Permutation Test: 10+20 points**) The same social scientist as in the previous task asks you to investigate a dataset on wages, since you've done such a good job on the previous task. The dataset wages.csv contains the annual salaries of male and female persons and a column which indicates the respective salary and gender.

(a) First of all, load the file and try to get an overall impression of the data. Since we aim to test for differences between female and male people, plot the respective empirical distribution functions of female and male people, as well as their empirical cumulative distribution function.

(b) Now, after having gotten an overall idea, you aim to find statistically significant differences in the distribution using the Kolmogorov-Smirnov test. Recall the Kolmogorov-Smirnov (KS) test-statistic which compares two distributions of random variables Y_1 and Y_2 via the following test-statistic:

$$D_n = \max_x |F_n^{Y_1}(x) - F_n^{Y_2}(x)|$$

Now you perform a permutation test which evaluates 10.000 permutations to investigate the interchangeability of data between the two groups. For each permutation you evaluate the test-statistic. In the end you obtain a vector of 10.000 numbers. Get hold of the density of the test-statistic by plotting the empirical density. Within this plot, mark the realization of the test-statistic on the original data where you have not yet interchanged anything. Estimate the p-value according to the permutation test and decide if you reject the null-hypothesis of the KS-test at a significance level of 3.5%. Could you already tell the outcome of the test by looking at the respective ecdf?

(c) Another test-statistic which measures the difference in the wages would consider the difference in mean of the two groups. So formulate a Null-hypothesis which assumes the smaller observed mean, besides being smaller, however is not statistically significantly smaller. Using a permutation test with 10.000 permutations check, if you would reject the Null-hypothesis at a level of 1%.

(d) Which conclusion about fairness do you draw? Base your argument on the parts of the exercise which you solved.

Hints: For plotting the ecdf and epdf you may want to use seaborn.kdeplot where you set the argument "cumulative" to your needs. The realization of test-statistics can neatly be marked via matplotlib.pyplot.axvline. You may use numpy.searchsorted to compute the ecdf for the KS-test. You only need to be able to evaluate the ecdf at the observations so the rank of the element in the data together with the size of the group is sufficient to do so. In order to perform a permutation of the observations, you permute only one column of the data. You may use pandas.Series().sample(frac=1) or np.random.permutation. A suitable data-structure which allows you to access rows according to labels would be a pandas.DataFrame which is recommended for this task.