FOUNDATIONS OF DATA SCIENCE

ASSIGNMENT -3

Aravind S

CB.EN.U4CSE20208

Problem Statement

In 2010, the American Civil Liberties Union (ACLU) of Northern California has presented a report on jury selection in Alameda County, California. It concluded that certain ethnic groups are underrepresented among jury panelists in Alameda County. A jury panel is a group of people chosen to be prospective jurors. It can consist of a few dozen people or several thousand, depending on the trial. By law, a jury panel is supposed to be representative of the community in which the trial is taking place

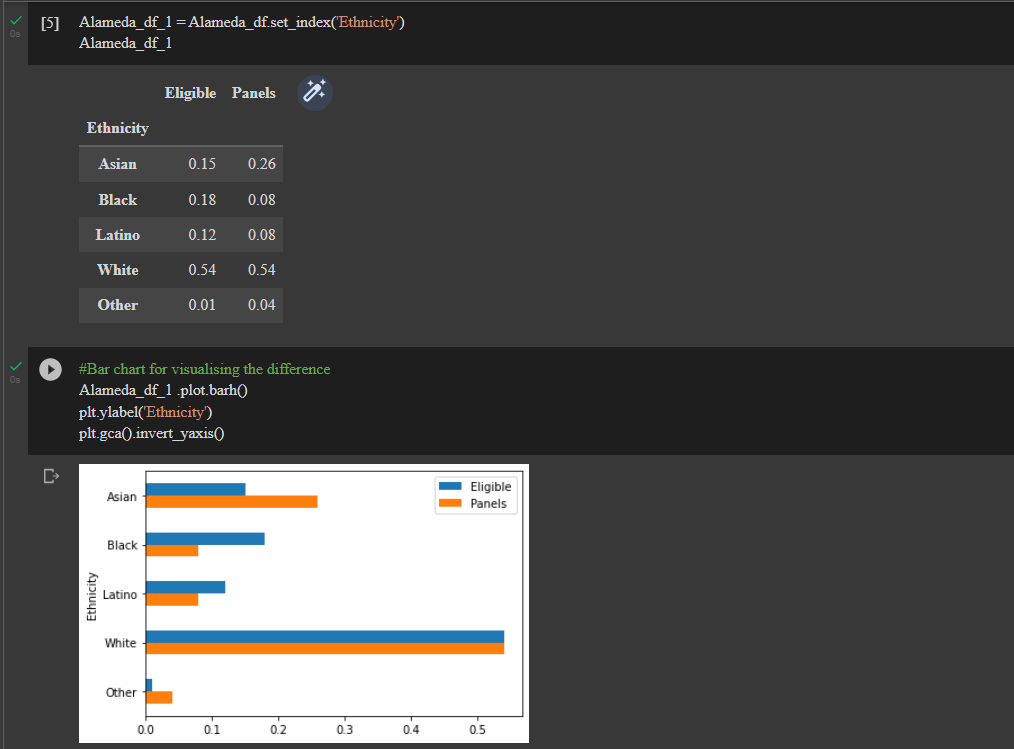
Attributes given in the dataset:

* “Ethnicity” = ethnic categories of people in Alameda.
* “Eligible” = fraction of people eligible grouped by ethnicity
* “Panel” = proportion of people currently chosen for the panel.

Hypothesis :

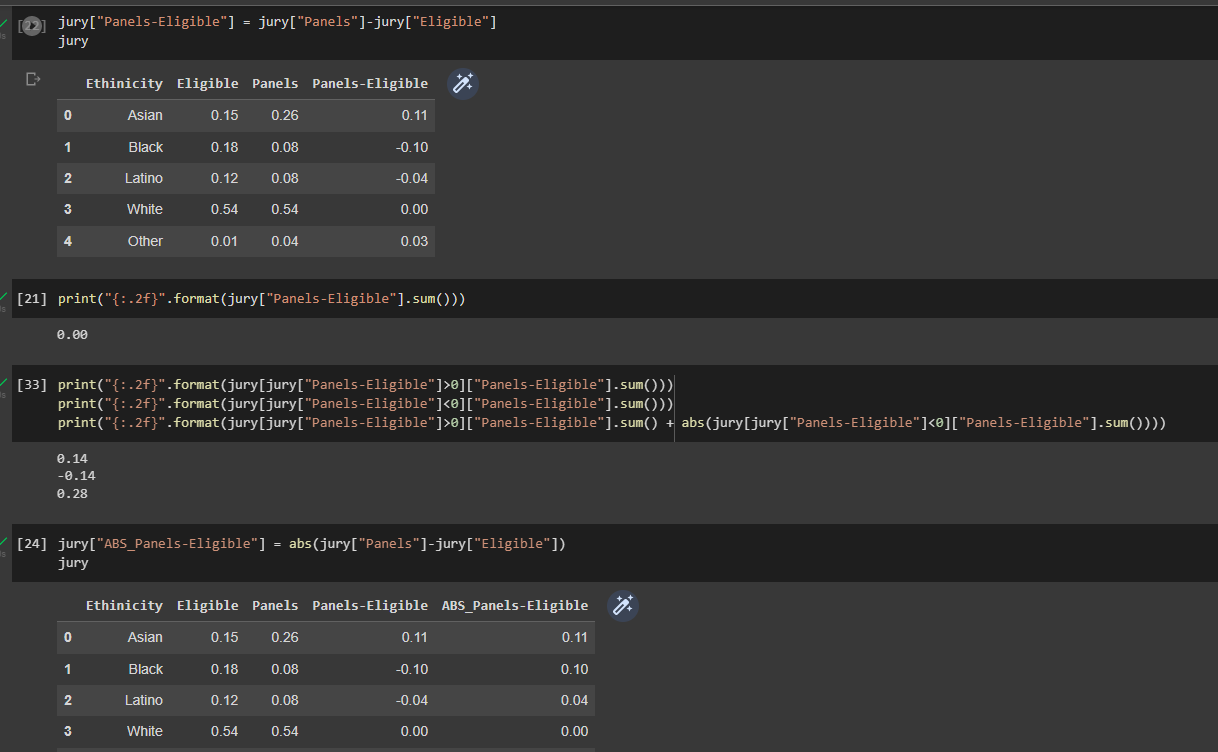
* H0 - Null Hypothesis =Panels were selected at random from the population
* H1 - Alternative Hypothesis = Panels were not selected at random

After setting column ‘Ethnicity’ as index and visualizing the horizontal bar graph.



Now that we have both a null and an alternate hypothesis, let’s find evidence reject null hypothesis

So, now to find the Bias that was done in selection, we Subtract Panel from eligible and thus arrive at either +ve, -ve, 0



So, it would only logically make sense if the number of members chosen in random that are in excess are same as the one in deficit. 14%,-14% in our case.

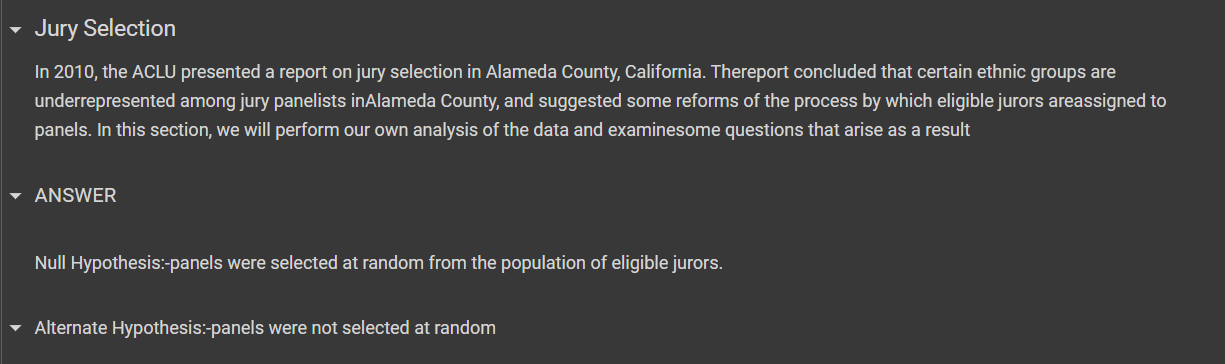
So, the positive and negative sign only imply excess or deficit. So, to get the total amount of tempered values, we can use abs() to get the Absolute value, in our case 28%.

So, now let’s try proving the hypothesis

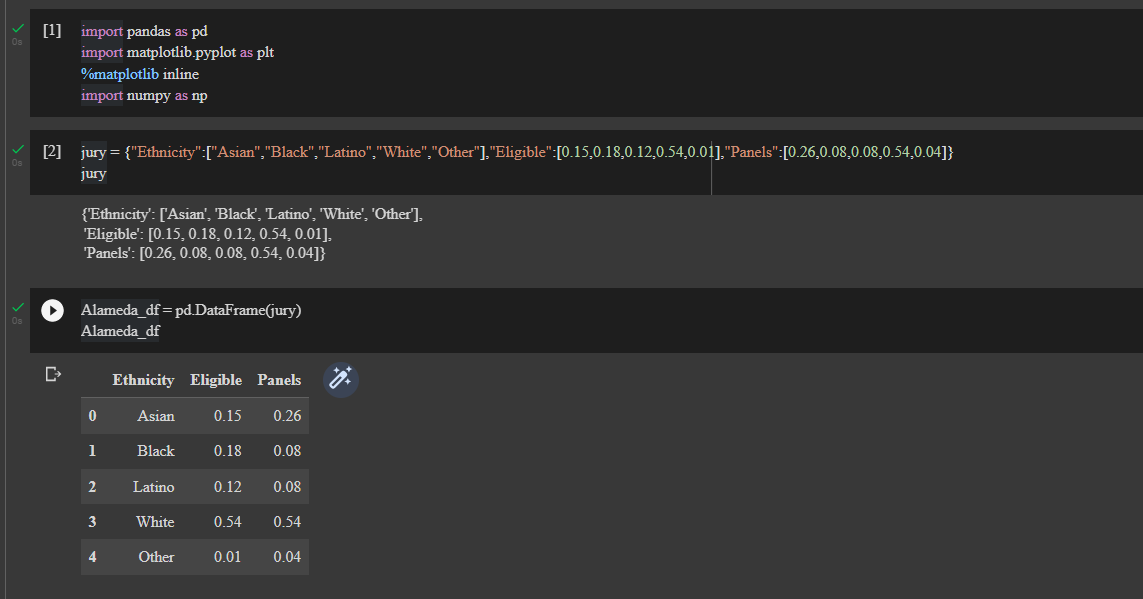
Alameda Jury Selection:

H0 – Panels were selected at random from the population(Null Hypothesis)

H1 – Panels were not selected at random (Alternative Hypothesis)

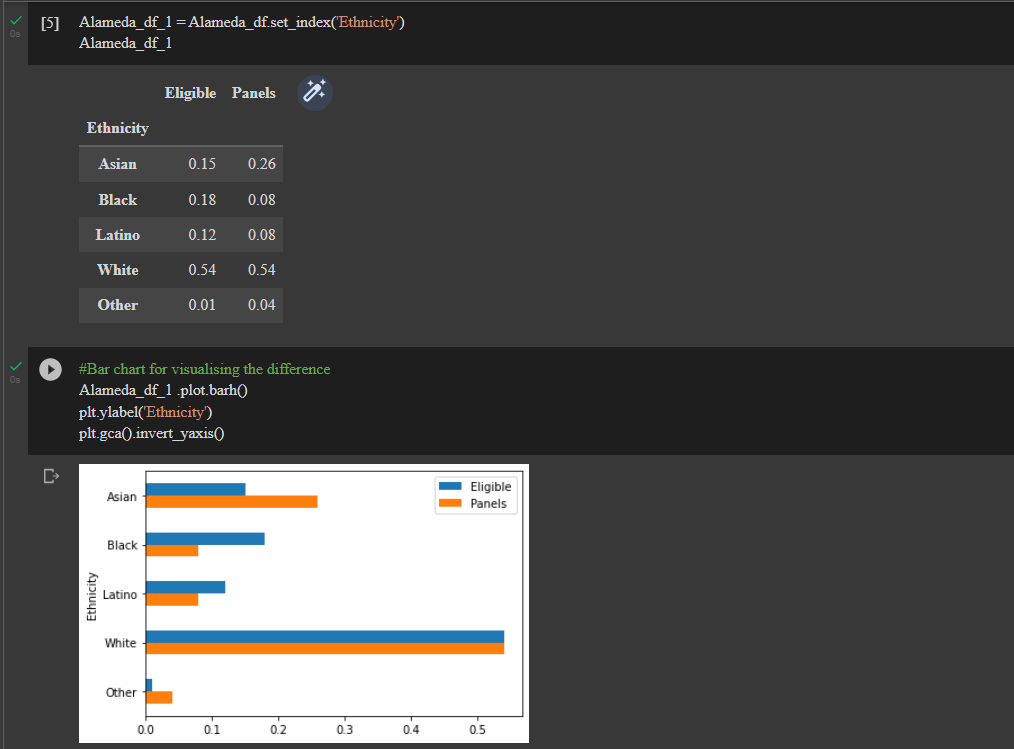


Making the data into a DataFrame



Setting column ‘Ethnicity’ as index

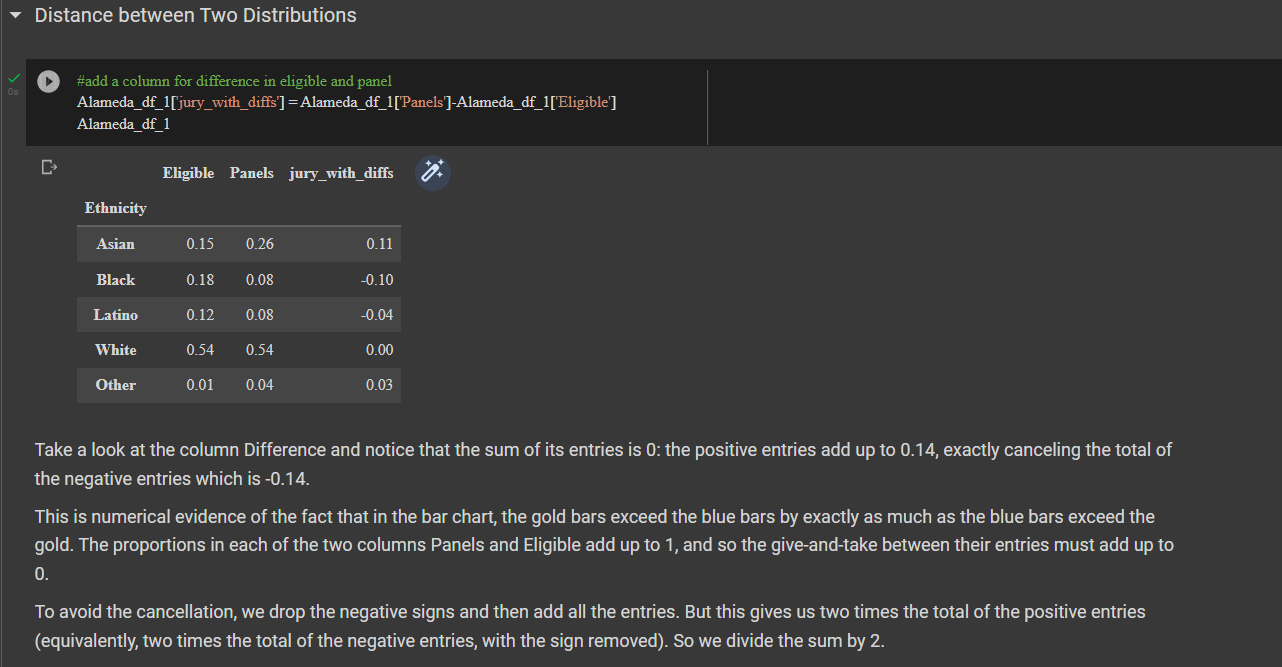
Visualizing the horizontal bar graph for ‘Ethnicity’

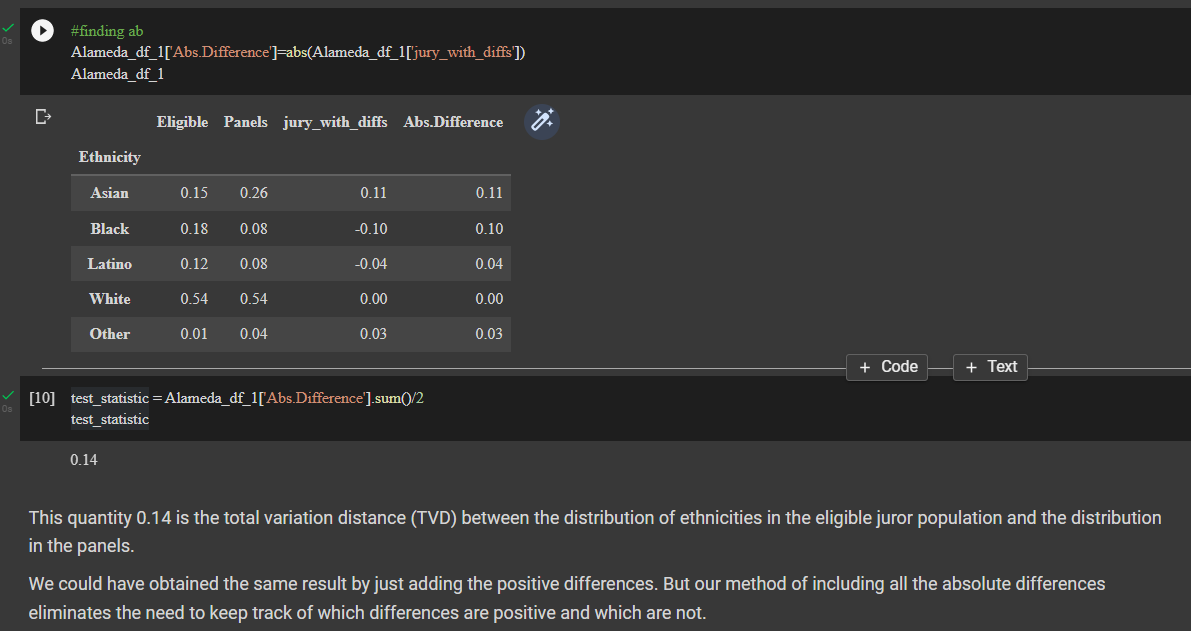


Now we have a null hypothesis and a alternative hypothesis, We should find a evidence in order to reject null hypothesis.

Here we take a column difference between eligible and panels, from the results, we can find that there is a evidence that give and take between the categories add up to 0.

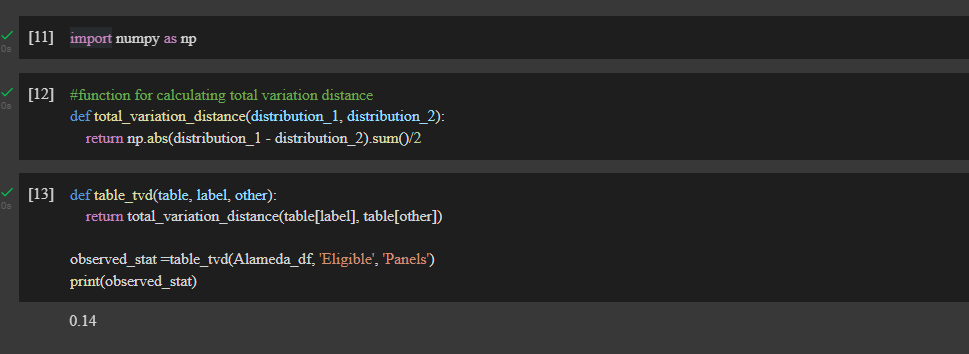
We take the absolute value of the difference column in order to avoid cancellation, so we add the magnitude and divide that by 2.



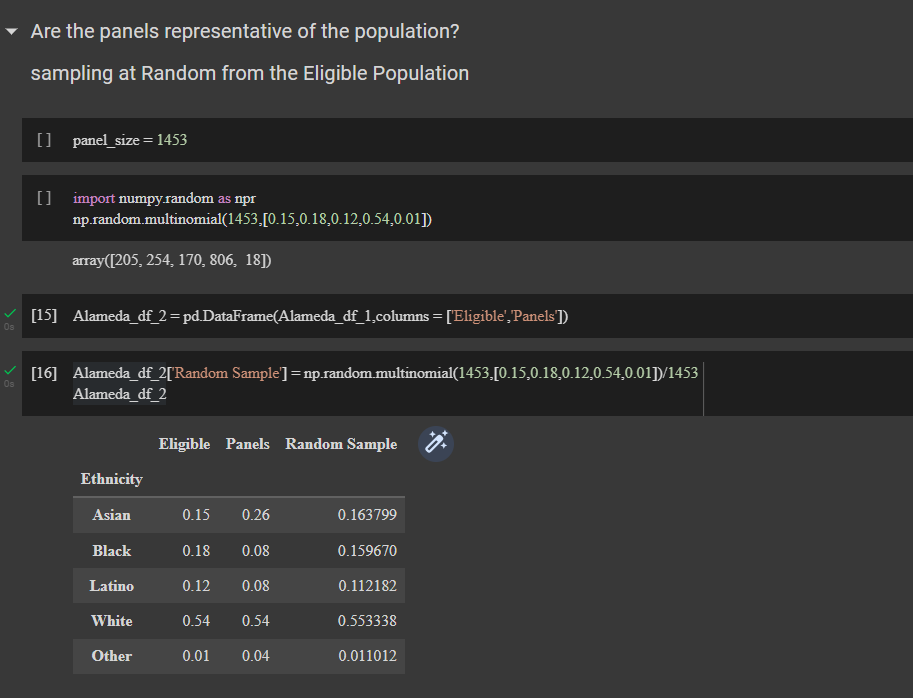


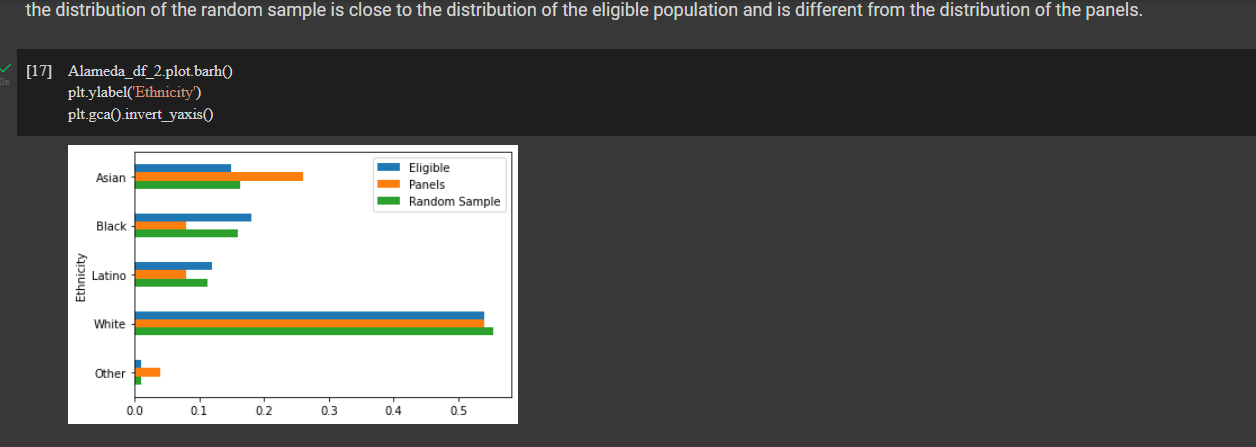
Total Variation Distance is the difference between probability distribution. We have found that value in above cells.

And below is the function that is interpreted for calculating TVD and one more function for calculating TVD table for various columns over the data



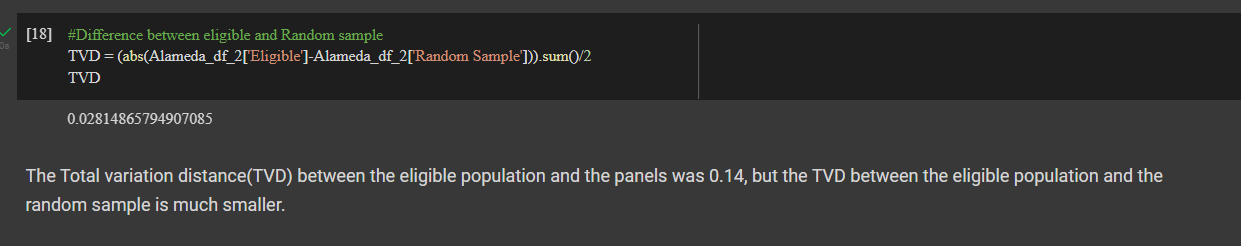
Here we sample the distribution at random and calculate the statistic for that. We visualize the sample distribution and calculated TVD for this case also between the eligible jurors and random samples.



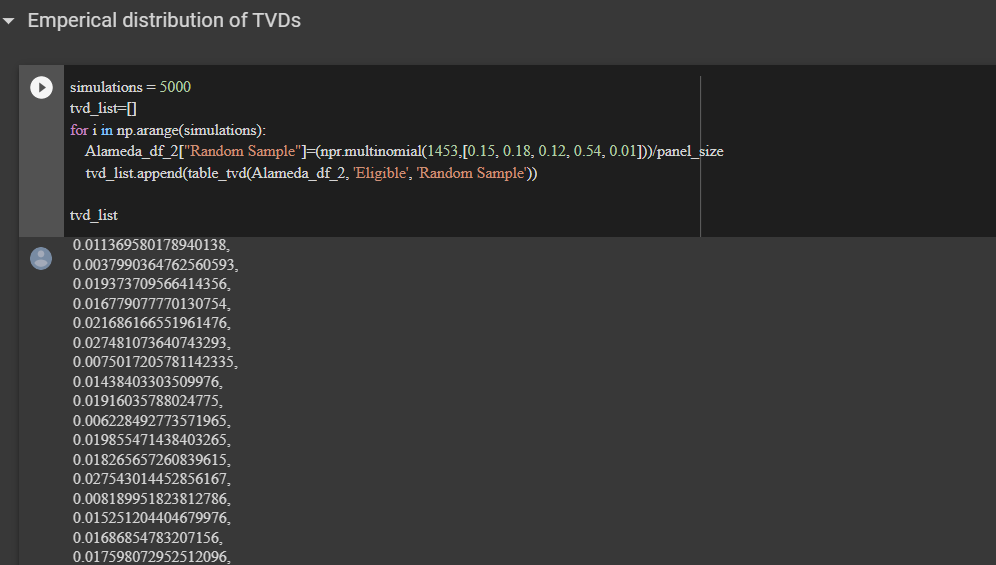


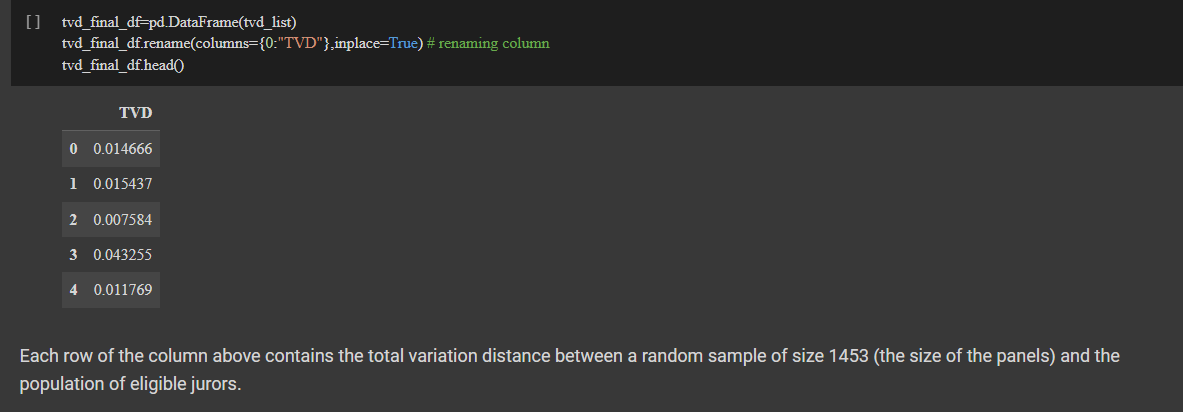
TVD – Total Variation Distance

It is the difference between probability distributions. Here the we are taking between eligible population and random sample



Using the table for TVD, we do the empirical distribution for the total variation distance for the case of random sample as we did before





Here we visualized the distribution of the tvd for the case of random sample.

So on a conclusion, distance that we calculated before is far from the distribution that we calculated. Hence as there is enough evidence to reject the null Hypothesis, we reject null hypothesis

