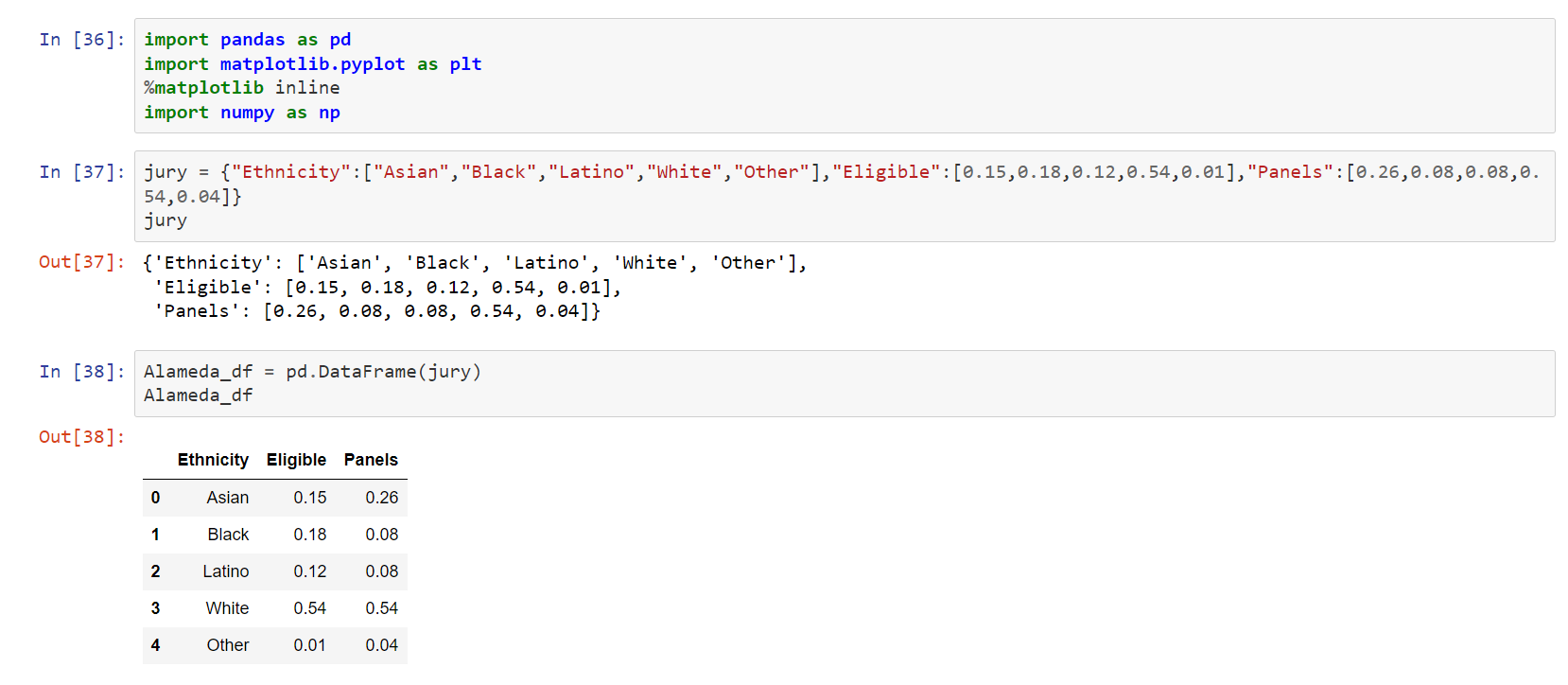
Pullela Meghana

CB.EN. U4CSE20239

**19CSE304 – Fundamentals of Data Science**

**Assignment-3**

Case Study: **Jury Selection in Alameda County**



There are 3 attributes in jury dataset Ethnicity, Eligible, Panels

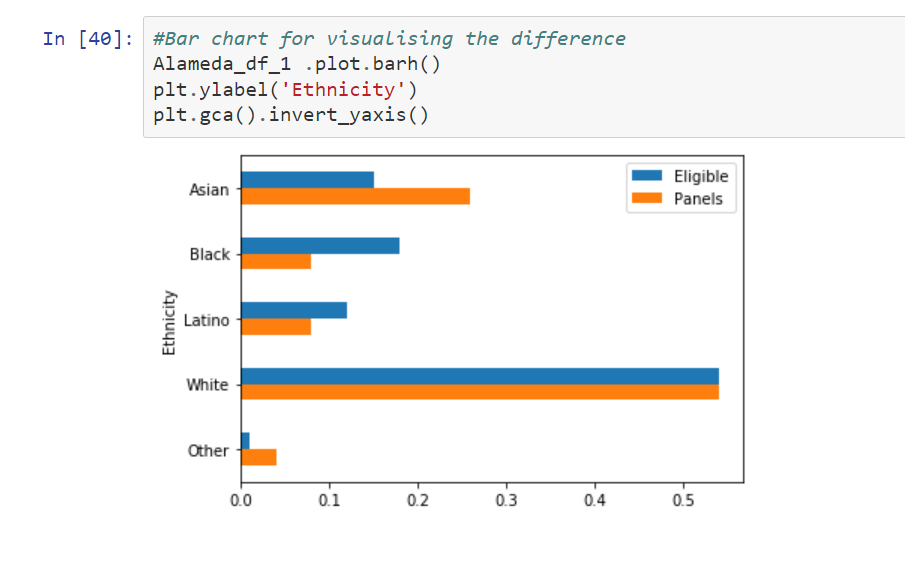
Ethnicity: Describes about the division of categories of people in Alameda.

Eligible: proportion of all eligible juror candidates of that ethnicity.

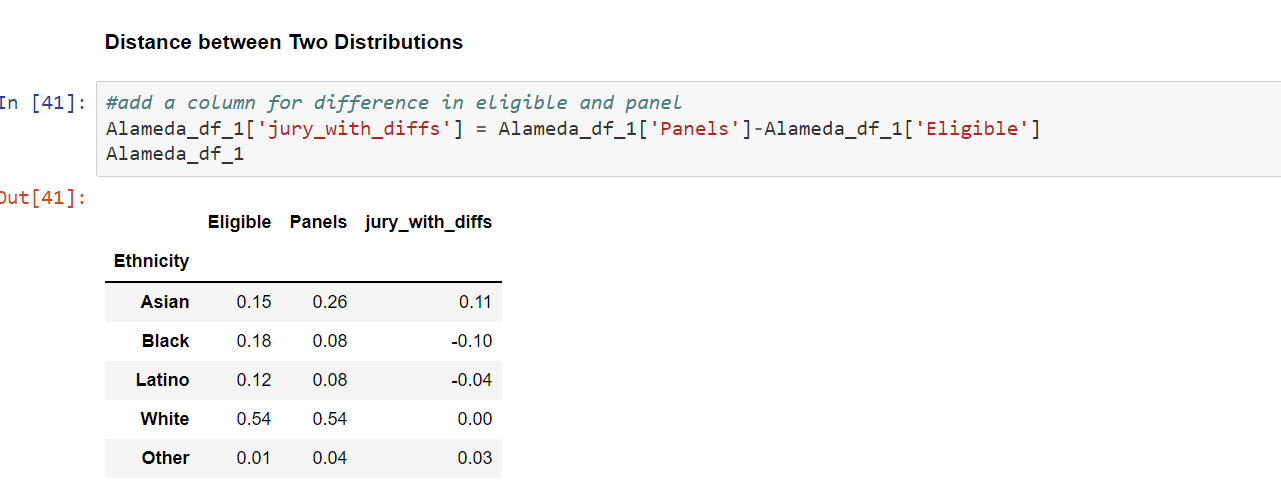
Panels: proportion of people of that ethnicity among those who appeared for the process of selection into the jury.

**Null Hypothesis**: panels were selected at random from the population of eligible jurors.

**Alternate Hypothesis**: panels were not selected at random.

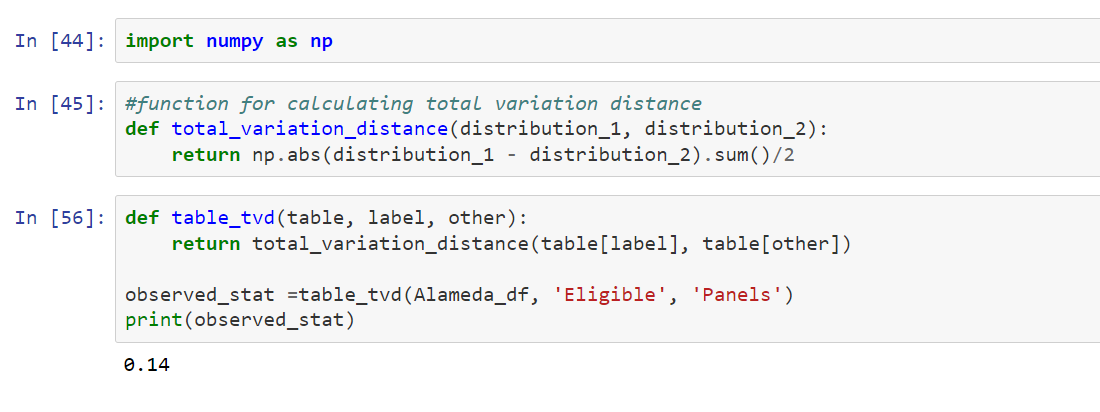


Plotting bar graph for two distributions of ethnicity.

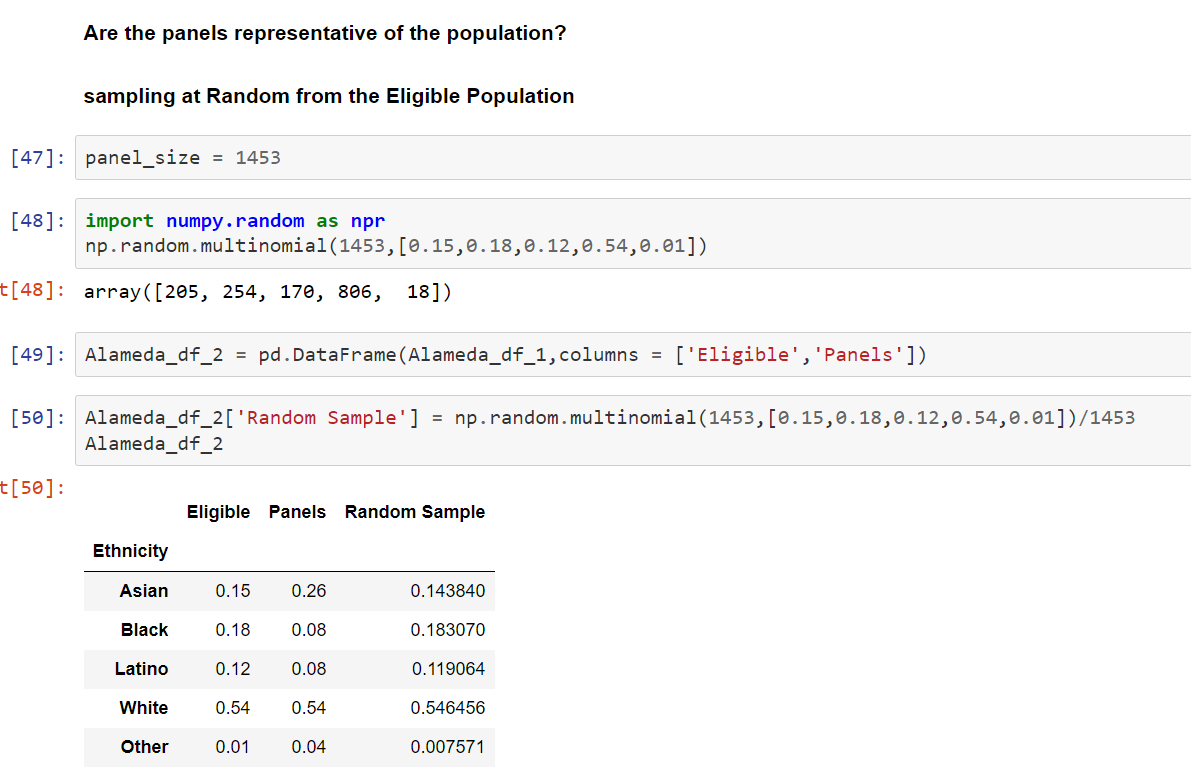


Compute the difference between the eligible and panel proportions and add that as a new column for jury dataset.



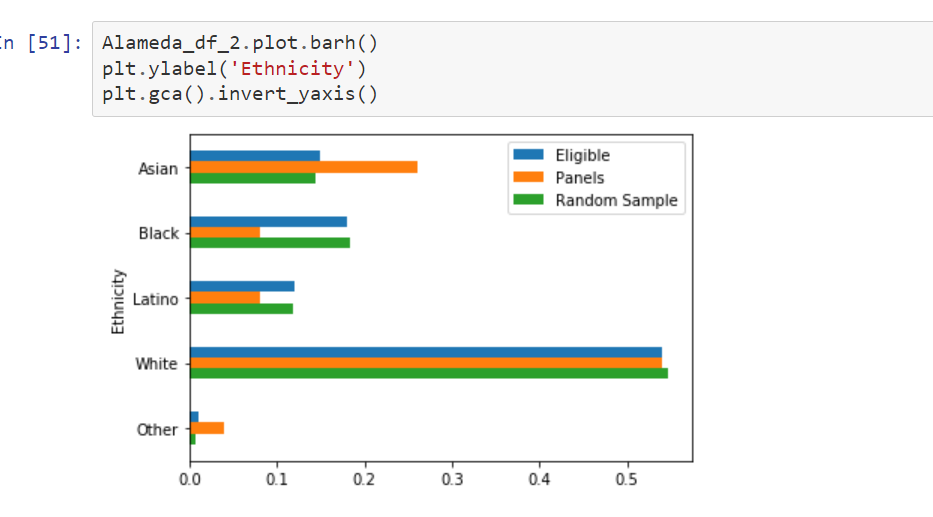


Here the total sum of differences is 0. So, to avoid cancellation find absolute positives values, as these doubles positive values divide it by 2. The quantity obtained 0.14 is the total variation distance (TVD) between the distribution of ethnicities in the eligible juror population and the distribution in the panels.

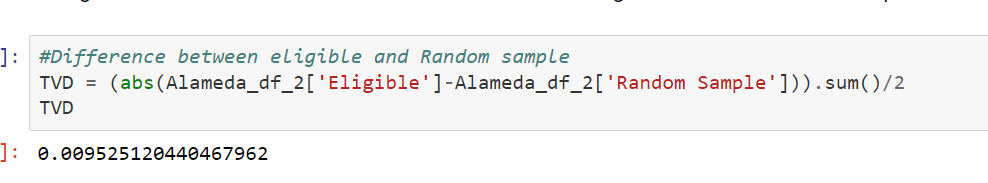


Taking sample panel size as 1453, we are finding proportion of all ethnicities based on the proportion of eligible juror population.

The calculated random sample is close to eligible jury population.

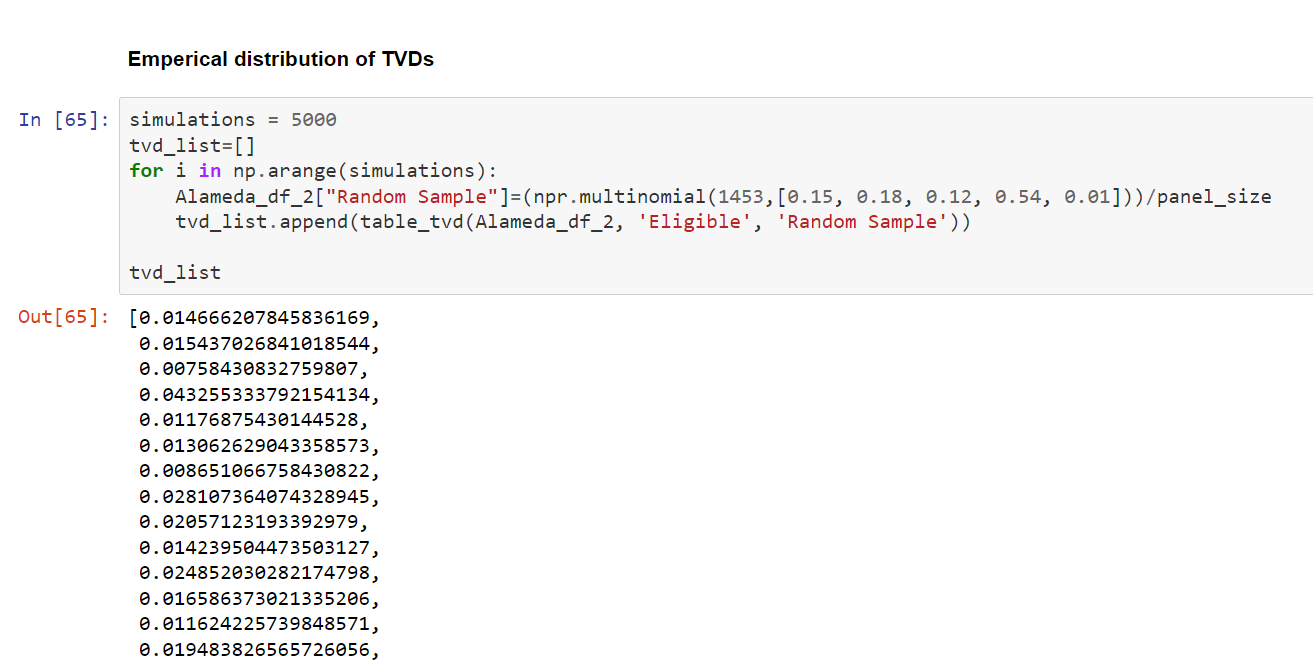


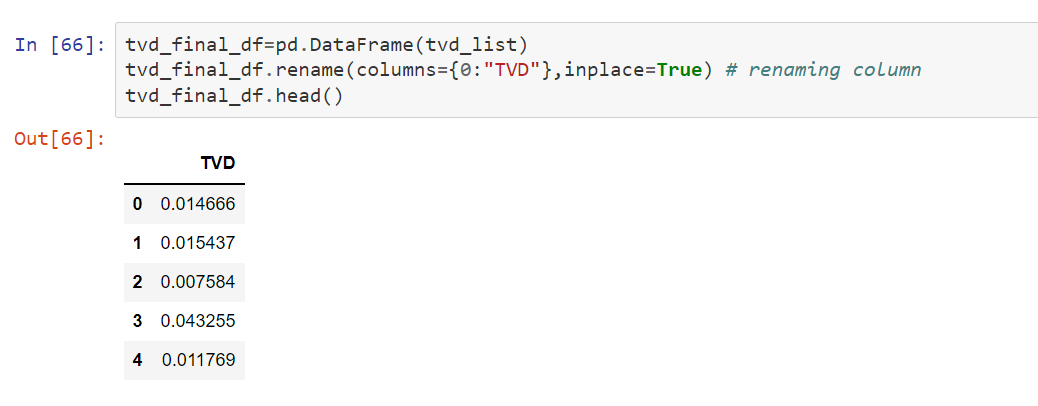
The green bar is closer in size to the blue bars than the orange bars are. The random sample resembles the eligible population, but the panels don't.



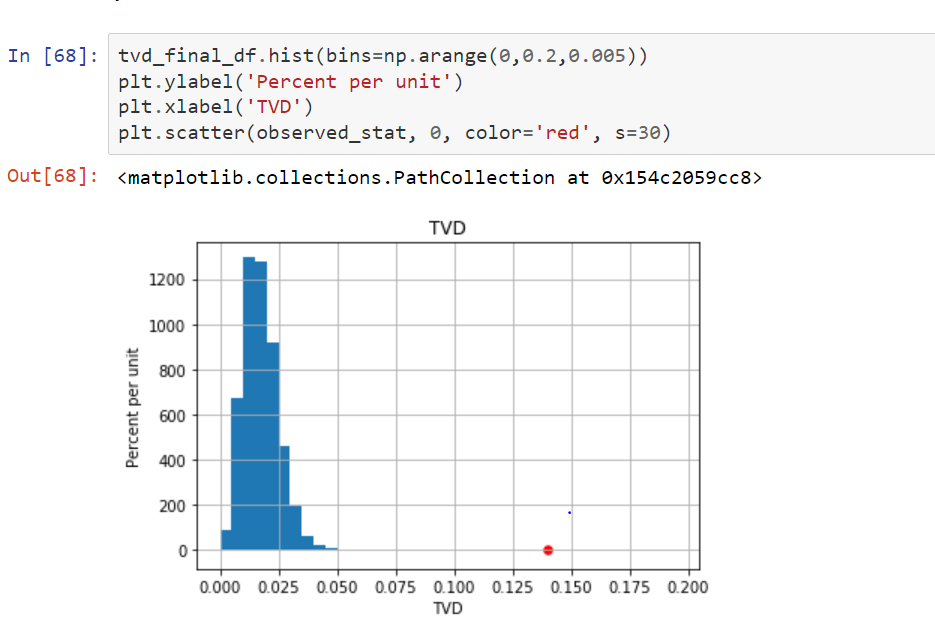
TVD between eligible proportion and random sample is much smaller compared to TVD between eligible and panel proportions.

This concludes that eligible and random samples almost resemble same.





Taking 5000 samples and calculating TVD between eligible and random sample of size 1453, for all samples and store it in a data frame.



Plot a histogram for TVD calculated for all samples and also use scatter plot to plot a point of distance between the panels and the population was 0.14. As the point is far away from the distribution. Panels were not representative of the distribution provided for the eligible jurors. As we have enough evidence to support alternate hypothesis, so we conclude that panels were not selected at random from eligible proportion. Hence Null hypothesis is rejected.