Fundamental Data Science

Student ID: 22097222 Name: Kavitha Subramaniyam
Coding project:

Data

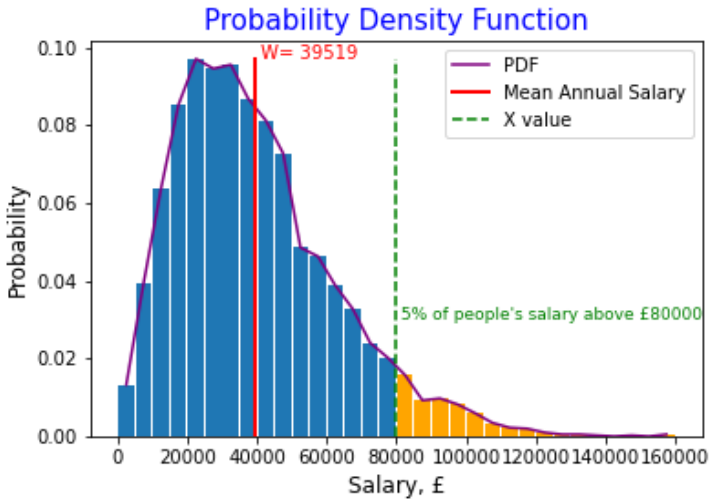
Datafile provided as csv file which contains 4000 entries for people's salary. Each entry represents the salary value of a person in euros.

Distribution

Created probability density function using 32 numbers of equal-width bins in the range zero to a value which is greater than max value of data(max value of given data is 158335, hence range selected 0-160000.00). Retrieved numbers of entries in each bin(hist) and bin boundaries(edge). Calculated bin centre location, bin width and normalise distribution to get discrete PDF.

This distribution is positively skewed and can observe a longer tail to the right hence the weight of the distribution is on the left which indicates more people's salary is less than average and therefore the extreme values will affect the mean salary value.

Histogram



Calculate Mean

Calculated mean salary value(W) using sum function with PDF(ydist), bin centre location(xdist) and bin width(wdist) values as input.

$W = {}_{k=-\infty}\Sigma^{+\infty}\ X_k\ f(X_k)$

W = 39519.25£

As distribution is positive skew the mean salary is greater than the median salary(35716£).

Calculate X

To find X such that 0.05 of the distribution corresponds to values >X value, calculate the k^{th} index using cumulative distribution and find X value as k^{th} index of bin boundaries.

$f_X\ (x_k) = P(X=x_k)$

X = 80000.00£

https://github.com/kavithasub/uh-study-modules/tree/main/fundamental-data-science/assignment