Part II

The summary of the email findings indicates that the ages of death of World Bank retirees within the past 6 years has a wide distribution, with 50 as the minimum age and a maximum age of 104. The mean age at death is 81 years old, which is lower than the life expectancy at age 65 for both men and women in the United States. And the analysis was carried out with 919 data points.

After reviewing the analysis, I have a few suggestions to further improve the analysis and findings.

It is important to remove outliers before plotting the distribution and calculating the mean value. Outliers are the data point that lies far outside the range of the other data points. These data points are either much larger or smaller than the other values in the data set, and they can affect the interpretation of the data by skewing the mean and range of the data. By visual inspection, it looks like data point with death age 104 could be an outlier.

There are different methods to calculate outliers, such as IQR (Interquartile range) method, Zscore method etc.

By visualizing the plot it looks the distribution of the data follows approximately normal or Gaussian distribution. In this case, Zscore could be a more relevant method; since the Z-score method is useful for datasets that are normally distributed or approximately normally distributed and can easily identify outliers that are significantly different from the majority of the data. The formula for calculating Z score is as follows,

$$Z - score = \frac{x - \overline{x}}{\sigma} \quad (1)$$

Where, x= raw data point, $\overline{x}$ is the mean and σ is the standard deviation.

The Z-score method can identify outliers that are more than 3 standard deviations away from the mean. Standard deviation is a measure of how much the data values vary from the average value.

So, If a data point has a Z-score that is greater than 3 or less than -3, it is generally considered to be an outlier. There is also a Python inbuilt library to calculate the Zscore, it can be imported from scipy.stats module.

I have also included an example notebook (PartII_example.pynb) to calculate Z score and to remove the outlier points using Zscore.

However, it's important to consider that if data don't follow a normal distribution, this approach[1] might not be correct.

Another suggestion would be to derive the conclusion using more data points and examining the results separately for both men and women. As this would provide more detailed insights.