José Mariano Portilla Landa 04/04/2022

HO-2: Data processing and analytics

Propose a workflow (i.e., pipeline) that shows the steps implemented in the notebook. For every step, clearly show the I/O, objective and partial conclusions. Out of conclusions explain what are the elements used to decide to propose the following steps.

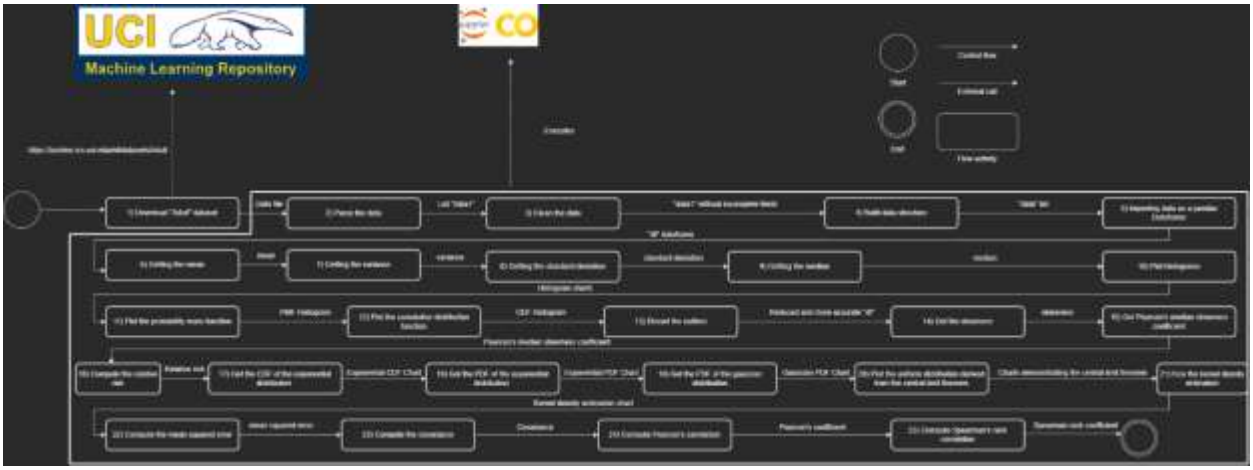

Figure 1: workflow of the steps implemented in the notebook.

Objectives and partial conclusions of each step:

1) Get the dataset to work with. The dataset contains around 32,000 observations about different financial parameters of US population, which will come handy for the statistic measures that we will be doing.
2) Parse the data. Now that we have the dataset, we can parse the data by opening the file and splitting each line of the file by “,”. After doing that we generate a list that will input to the next step.
3) Clean the data. To clean the data we are using a simple approach by just ignoring the records (lines) that are not complete from source (not having 15 elements), this will avoid having troubles by the size of each record.
4) Build the data structure. To build the general data structure we are creating a new list and appending each element of each line of the dataset. After doing this we are ready to work with our data, however we are doing an extra step to be able to manage our data more easily.

5) Create a dataframe from our data. To make it easy to manage our data we are creating a dataframe structure from the library pandas. This structure is a two-dimensional size-mutable tabular data structure.

6) Compute the mean. Now that we have the dataframe, we can calculate the mean of the age of men and women, and the average age of high-income men and women. The dataframe structure helps us to do this very easily by just invoking mean() and we can see that women reach high income first than men.

7) Compute the variance. We can also get the variance, which describes the spread of data. Because variance is hard to interpret, we prefer the standard deviation.

8) Compute the standard deviation. The standard deviation can also be obtained, which is the square root of variance. From the exercise we get that the mean of hours per week for men is 42, however the standard deviation is of 12 hours so the spread of data is kind of high, for women we have that the mean of hours per week is less, but the deviation is almost the same.

9) Compute the median. We can also compute the median, which is the middle value of a sample, and from the exercise we get that the median is fairly similar to the mean, compared to the previous results. The highest value of disagreement is hours per week for women with only 4 hours of difference.

10) Plot the data distribution histogram. Plotting the histogram shows us how often each value appears in the data distribution. We are plotting the age of working men and women separately. From the notebook, we can see that the predominant ages for men in the dataset are from 30 to 45, with a drop in 35. For women we can see that that the predominant ages are from 20 to 30, and is almost like a ladder.

11) Compute and plot the probability mass function. If we divide the number of samples by n, then we can normalize the frequencies of the histogram, which gives us the Probability Mass Function. From the notebook we can see that the light color indicates the male population and the dark color the female. In the female population we can see that is more detailed because we configure the bins (bars) to be 20.

12) Compute and plot the cumulative distribution function. The cumulative distribution function describes the probability that a real-valued random variable X with a given

probability distribution will be found to have a value less than or equal to x. From the notebook we can see for males that around age 30 is where we have a larger step, so the sample is larger in there. And for women, we can see that around age 20 is where we have a larger step, so the sample is larger in there.

13) Find and remove outliers. If we eliminate outliers by computing samples whose value exceeds the mean by 2 or 3 standard deviations, then we are getting a much more representative sample of what we want to accomplish.

14) Compute the asymmetry of the dataset. Skewness can be affected by outliers, however we already dealt with that problem by removing them in the previous step. In the notebook we got a skewness of 0.266 for the male population and 0.386 for the female, this means that the data is asymmetric, extending further to the right than to the left.

15) Compute the Pearson's coefficient. Pearson's coefficient is more robust that the skewness coefficient. From the notebook we get that the male population Pearson's coefficient is of 0.095 and for the female 0.26, we can see that the difference is greater in the Pearson's coefficient.

16) Compute the relative risk. The relative risk is the ratio of two probabilities, for the example we are considering the probabilities of being promoted early for men and women.

17) Plot the CDF of the exponential function. In this step we are only plotting the continuous distribution function to denote that many real problems are well approximated by fitting them into CDF's.

18) Plot the PDF of the exponential function. This step works as a comparison between the CDF and the probability density function for the exponential.

19) Plot the PDF of the gaussian distribution. The PDF representation of the gaussian distribution is the most common one, it describes a lot of phenomena and it is amenable for analysis.

20) Plot the uniform distribution derived from the central limit theorem. In this step we are demonstrating the central limit theorem by drawing n samples (n=2,3,4…) at random from the parent distribution, computing the average and repeating the process. The end result is a normal probability density.

21) Plot the kernel density estimation. We are estimating the kernel density to get a continuous representation of data that we do not know their distribution.
22) Compute the mean squared error. From the notebook we get a really small error by estimating 200 times. This is due to the removal of outliers.
23) Compute the covariance. Covariance is a measure of the tendency of two variables to vary together. From the notebook we get that the covariance(x, y) is of 0.18, which means that they are positive related but weakly.
24) Compute the Pearson's correlation. Pearson's correlation only measures linear correlations, it is a normalized measurement of the covariance, meaning that the result can only be between -1 and 1. In the example we are getting 0.28 as the coefficient.
25) Compute Spearman's rank correlation. This correlation is helpful when we have outliers and skewed distributions, to compute it we have to use the rank of each value, which is its index in the sorted sample. In the same example as the Pearson's correlation we got a 0.8, which means that there is a strong correlation and that there are probably outliers and the data is skewded.

What is the purpose of introducing the Limits Theorem and how is it use in the analysis?

Its purpose is to measure how much the means of various samples will vary without having to take any other sample means to compare it with. It is used by taking the uniform distribution as the parent distribution and derive a normal distribution from it.

What about kernel density? how does the use of these strategies modify and impact the conclusions of the analysis?

Kernel density helps to estimate the distribution of our data without making assumptions about the form of the underlying distribution, if we consider a Gaussian kernel around the data, the sum of them can give us a function that approximates the density of the distribution, thus estimating the distribution non-parametrically.