|  |
| --- |
| **## Portfolio Assignment week 01** |
|  |  |
|  | **Study the Tutorial tutorial\_cluster\_scanpy\_object and the tutorial\_Clustering\_Methods** |
|  |  |
|  | **Write a brief summary about the following:** |
|  |  |
|  | **- What are common preprocessing steps? Explain for each step why and when you should execute this step and when not.**  **Answer:**   |  |  | | --- | --- | | Preprocessing | Why? | | Loading Data+ shape | **To know the size of data** | | .dtype | **To know what the type of columns are (continuous, categorical, and boolean)** | | If in .dtype, there is object type, .groupby, or .value\_counts() | **To know how many and what type we are** | | Check null data | **to know if the number of them is low, eliminate them. If not, use interpretation methods to replace them.** | | Check if data are normal or not | **To decide normalize or not** | | Normalization | **Decide when normalize data and when not.  we should normalize data when the range of values for different fratures is very different.(one is 0-1, other is 100-1000)** | |
|  | **- What visualization methods are used in the cluster methods tutorial? Explain why the selected method is the most appropriate method for the visualization. Bonus points: do this as well for the scanpy tutorial.**   |  |  | | --- | --- | | Visualization methods | Why? | | Histogram plot | **To have a overall view of the range of categorical data and skewness of them.** | | pairplots | **to explore the relationship between different pairs of variables in a dataset.** | | Plot a scatter plot based on different clusters and inertia | **To find the number of best cluster** | | Dendrogram | **to obtain the linkage and dendrogram functions** | |
|  | - **What performance/evaluation metrics are in the cluster methods tutorial? Explain why the used methods are the most appropriate method for the evaluation.**  1- correlation matrix:This matrix Shows relationships between different variables and varies between -1 , 1. If the value is -1 or near this number, the relationship between those variables are negative, while +1 shows the strong positive relation, and zero is for no relations.  2- ROC-AUC: (Receiver Operating Characteristic - Area Under the Curve) is a performance metric to evaluate the quality of a classification model. It assesses the model's ability to discriminate between the positive and negative classes by measuring the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity). If the number is 1 or near to that, it is a perfect classifier |
|  |  |
|  |  |
|  | **Bonus:** |
|  | **You practice the steps yourself with the breast\_cancer dataset (clustering\_data.csv)** |