CREDIT TASK

About this task

Step-1

This task is designed to assess the Credit level expectations. There are two sets of evidence requested in this task: one is for SIT307 students and the other one for SIT720 students. Please select the set based on your unit code. DO NOT SUBMIT BOTH SETS OF EVIDENCE.

Step-2

Your tutor will then review your submission and will give you feedback. If your submission is incomplete the tutor will ask you to include missing parts. Tutor can also ask follow-up questions, either to clarify something that you have submitted or to assess your understanding of certain topics.

Feedback and submission deadlines

Feedback deadline: Friday 14 April (No submission before this date means no feedback!)

Submission deadline: Before creating and submitting portfolio.

Background

Building development activity monitor is important for city construction and planning, in this task a real world data set from City of Melbourne is provided, it monitors major new commercial and residential property development in the City of Melbourne municipality. Analysis of the data can inform short to medium-term supply forecasts in the commercial and residential markets. This data provides our current understanding of major development sites that are recently completed, under construction, planned or mooted in all of 13 small areas of the city.

Dataset

Dataset file name: development-activity-monitor.csv

Dataset description: Dataset contains different features along with the development activity. It contains total 42 features. It contains different types of data including int, float and string. Feature names, data type and values are described in the following link. Each observation is a datapoint along the row of the dataset. The data set can be found in :

https://data.melbourne.vic.gov.au/explore/dataset/development-activity-monitor/information/ .

Evidence of Learning – SIT307

Answer the following questions in a .jpynb file, execute your code and keep the output, submit the .jpynb file to **Ontrack** (**https://ontrack.deakin.edu.au**)

1. Load the data from supplied data file. Print the data dimension.
2. Continue from question 1. Display the data type of all features. If the data type is integer, print the median values of the features.
3. Continue from question 2. Print all the possible values of the feature "status" and calculate the ratio of each "status" value.
4. Is there any association between clue_small_area and floors_above? Explain your results from given dataset.
5. Print the number of properties for different suburbs. Please report the pattern found in the result, if any.
6. Continue from question 5, which suburb has the smallest number of properties which are under construction?
7. Continue from question 6, which suburb has the biggest number of hotel rooms?
8. Create and print a data frame of the number of different status values for different year groups (based on 5 years interval).
9. Continue from question 8. Draw a histogram of number of status values against different year groups. Explain the result.
10. Based on the original dataset, exclude the clue_small_area feature, use the rest available features and perform clustering on all the properties and determine the number of clusters. Is this the same as the number of suburbs in the data set?
11. Continue from question 10, perform K-Means on the data set, report the purity score.
12. Continue from question 11, try at least three different distance metrics for K-Means, select the best distance metric for each corresponding clustering algorithm, explain why the chosen distance metric is the best for the given data set.
13. Apart from K-Means, try another clustering method, and compare the results.

Evidence of Learning – SIT720

Answer the following questions in a .jpynb file, execute your code and keep the output, then submit the .jpynb file to **Ontrack** (**https://ontrack.deakin.edu.au**)

1. Load the data from supplied data file. Print the data dimension.

2. Continue from question 1. Display the data type of all features. If the data type is integer, print the median values of the features.

3. Continue from question 2. Print all the possible values of the feature “status” and calculate the ratio of each “status” value.

4. Is there any association between status and clue_small_area? Explain your results from given dataset.

5. Print the number of properties for different suburbs. Please report the pattern found in the result, if any.

6. Continue from question 5, which suburb has the biggest number of properties which are under construction?

7. Continue from question 6, which suburb has the biggest number of student apartments?

8. Create and print a data frame of the number of different status values for different year groups (based on 5 years interval).

9. Continue from question 8. Draw a histogram of number of status values against different year groups. Explain the result.

10. Based on the original dataset, exclude the clue_small_area feature, use the rest available features and perform clustering on all the properties and determine the number of clusters. Is this the same as the number of suburbs in the data set?

11. Continue from question 10, perform K-Means and Hierarchical clustering on the data set, report the purity score.

12. Continue from question 11, try at least three different distance metrics for K-Means and Hierarchical clustering, select the best distance metric for each corresponding clustering algorithm, explain why the chosen distance metric is the best for the given data set.

13. Apart from K-Means and Hierarchical clustering, try another clustering method, and compare the results.