作業3-2 Report Discussion

姓名:鄭至善 系級:資訊109 學號: F74051043

* Choose a dataset?

Ans:

Travel\_Review\_Ratings

* Analyze the data?

Ans:

1. Preprocessing

Remove unused column, unaligned rows

Remove 0 rating record, regarding as missing data.

Replacing 5 rating with 5 – 0.1\*(single user’s average rating in 24 categories) because 5 rating seems doubtful(Few (One) sample maybe).

1. Plotting statistical figure: Correlation coefficient, Histogram,

* Define a reasonable problem?

Ans:

Clustering the people into groups.

Expecting people in the same groups owning the same sense of values(How to value different travel spots). With accurate clustering method, people are more likely to find proper travelling partners, which consequently forms a fabulous trip.

* Explain how you improved your results step-by-step?

Ans:

First Try: K-Means clustering

Somehow is an unsupervised method with no labeled answer.

Moreover, visualizing with multiple features of data is a difficult task.

I tried to modify the number of clusters and evaluated with the performance of silhouette value(Silhouette Method) and inertia(Elbow Method) which gave some aspects to choose proper K. Finally pick K = 3. In other hand, I tried PCA for Dimension reduction. After reducing to 2 features left, I performed another scatter plot. And the result seemed quite similar with 3 clusters by intuition.

Second Try: DBSCAN clustering

As same by the evaluation of silhouette value, tuning the hyperparameters became more trustable. Compared with K-means, the result grouped people in a different way providing another point of view in the consideration of density distribution.