Congratulations! You passed!
TO PASS 80% or higher

Keep Learning

GRADE
100%

Graded Quiz: Test your Project understanding

LATEST SUBMISSION GRADE
100%

1. Why was Exploratory Data Analysis useful for our project? — 1 / 1 point

- [x] NaNs (nulls) will break the model if they are not removed.

Correct
Correct. Machine learning models need numbers explicitly.

- [x] Removing duplicates will prevent faulty clusters from forming.

Correct
Correct! False duplicate locations could bias our models, especially our density-based ones.

- [ ] It made our scatter plot much easier to visualize.

2. The K-Means algorithm will find the optimal number of clusters, *k*. — 1 / 1 point

- ( ) True. K-Means uses Silhouette Scores to find optimal clusters.
- ( ) True. K-Means iteratively tries different cluster values *k* in order to find the best performing one.
- (•) False. K-Means clusters iteratively but needs to be told how many clusters, *k*, to find.

Correct
Correct. This is one of the limitations of the K-Means algorithm and is what prompts us to use more advanced, density-based approaches.

3. K-Means will always result in the same answer, given enough iterations to run. — 1 / 1 point

- ( ) No. K-Means considers the density of your data, which never changes during each iteration.
- ( ) Yes, since the data is not changing, K-Means will always converge to the same result.
- (•) No. The K-Means algorithm starts with random values, and these can result in different clustering results even after many iterations.

Correct
Correct. K-Means is highly dependent on this random first iteration, and while most times results might be highly similar, there is no guaranteee that they will be the same.

4. DBSCAN can take into account varying densities in your data, and clusters them effectively. — 1 / 1 point

- ( ) Correct. By its definition DBSCAN find dense micro-regions and clusters them in your data.
- (•) Incorrect. DBSCAN has a global understanding of density.

Correct
Correct. This is in fact a limitation of DBSCAN, and is what HDBSCAN addresses.

5. HDBSCAN is an improvement over DBSCAN by finding the optimal hyperparameters for DBSCAN to run — 1 / 1 point

most effectively.

- (●) False. HDBSCAN basically uses DBSCAN with different hyperparameters to find different levels of density in a given dataset.
- ( ) True. This is why HDBSCAN will always outperform DBSCAN.
- ( ) False. HDBSCAN uses DBSCAN in conjunction with K-Means to find optimal clustering and minimize outliers.

✓ **Correct**

Correct. HDBSCAN varies the epsilon parameter in particular of DBCSAN to find clusters of different densities that DBSCAN may miss or skip over.

6. Outliers have to be dealt with in all clustering algorithms. **1 / 1 point**

- (●) Not necessarily. We had to define a way of dealing with them for our specific use case.
- ( ) Yes. Outliers do not tell us any useful information.

✓ **Correct**

Correct. We had a situation whereby we wanted all samples in our dataset to exist in some cluster.

7. Treating outliers as singletons was necessary to get a valid value for our Silhouette Scores. **1 / 1 point**

- (●) True. We treated outliers as single-sample clusters so as to not treat all outliers as one big cluster.
- ( ) Ignoring outliers was the correct implementation of the Silhouette Score, as this data does not tell us anything useful.

✓ **Correct**

Correct. This will allow for the Silhouette Score to give a score to the whole clustering result, not just the samples that were indeed clustered.

8. Our hybrid approach had, on average, larger clusters than K-Means. Would you say this is good for our task? **1 / 1 point**

Yes, It is.

✓ **Correct**

If we are to be building service centres for each cluster, we'd like these service centres to actually be reaching a good number of taxi ranks. Having some servicing many taxi ranks and some servicing 2-3 taxi ranks is not an ideal situation. On average, having a slightly higher cluster size across your clusters suggests a more even clustering that is logical for our usecase.