1. Summarise main points from Week 3 and 4

- **Week 3**: We jumped into unsupervised learning, where the computer finds patterns on its own without being told what to look for. A big part of this is clustering, which groups similar data points together. We learned how to measure "similarity" or "distance" between these points using different methods. The main clustering technique we looked at was K-means. We saw how it works by finding cluster centers and how we can tell if the clustering is good, though K-means does have some limitations. We also touched on K-means++ as an improvement and briefly mentioned other ways to cluster data, like DBSCAN and Hierarchical clustering, finishing up with how to do K-means using Python.

- **Week 4:** We tackled the problem of having data with way too many features or dimensions, which makes things complicated. This is called the "curse of dimensionality." The main way we looked at solving this was using something called PCA (Principal Component Analysis). PCA helps by finding the most important directions or patterns in the data to shrink it down while keeping as much key information as possible. We touched on the math behind it and how to actually do it using code, even seeing an example with images. We also briefly checked out other ways to reduce dimensions, like t-SNE, which is good for visualizing complex data.

2. Summary of reading list – external resources, websites, book chapters, code libraries, etc.

External Resources/Websites:

- StatQuest: K-means clustering: https://www.youtube.com/watch?v=4b5d3muPQmA
- K-Means Clustering Explained: An Easy Guide to Cluster Analysis: https://www.youtube.com/watch?v=YEwt6BJROug
- DBSCAN Clustering Algorithm Explained Simply: https://m.youtube.com/watch?v=Lh2pAkNNX1g&pp=ygUUI2NsdXN0ZXJpbmdleHBsYWluZWQ%3D
- StatQuest: Principal Component Analysis (PCA), Step-by-Step: https://www.youtube.com/watch?v=FgakZw6K1QQ
- Principal Component Analysis (PCA) - YouTube: https://www.youtube.com/watch?v=fkf4IBRSeEc&pp=0gcJCdgAo7VqN5tD
- Machine Learning | t-SNE -: https://m.youtube.com/watch?v=xLwEo_lGVrk&pp=ygUQIzJiaGtmbGF0c25lYXJtZQ%3D%3D
- Silhouette Score for clustering Explained : https://m.youtube.com/watch?v=_jg1UFoef1c&pp=ygUVI3NpbGhvdWV0dGVkZWZpbml0aW9u
- Machine Learning | Purity : https://www.youtube.com/watch?v=FKV_SMMnCTk
- K Means ++ for Initialization : https://www.youtube.com/watch?v=5k-ngBquGBI

Code Libraries:

- **Pandas**
  - **Functions:**
    - DataFrame.sum(): Documentation (used with isnull())
    - DataFrame.drop(): Documentation
    - DataFrame.boxplot(): Documentation
- **NumPy**
  - **Functions:**
    - numpy.nanmean(): Documentation
    - numpy.cumsum(): Documentation

        - numpy.argmax(): Documentation
        - numpy.cov(): Documentation
        - numpy.sum(): Documentation
        - numpy.amax(): Documentation
        - numpy.nan: Documentation
- **Seaborn**
    - **Functions:**
        - seaborn.histplot(): Documentation
        - seaborn.kdeplot(): Documentation
- **Scikit-learn**
    - **Functions:**
        - sklearn.preprocessing.StandardScaler: Documentation
        - sklearn.cluster.KMeans: Documentation
        - sklearn.cluster.DBSCAN: Documentation
        - sklearn.decomposition.PCA: Documentation
        - sklearn.manifold.TSNE: Documentation
        - sklearn.metrics.silhouette_score: Documentation
        - sklearn.metrics.adjusted_mutual_info_score: Documentation
        - sklearn.metrics.cluster.contingency_matrix: Documentation
        - sklearn.neighbors.NearestNeighbors: Documentation
        - sklearn.datasets.make_blobs: Documentation
- **SciPy**
    - **Functions:**
        - scipy.linalg.inv: Documentation

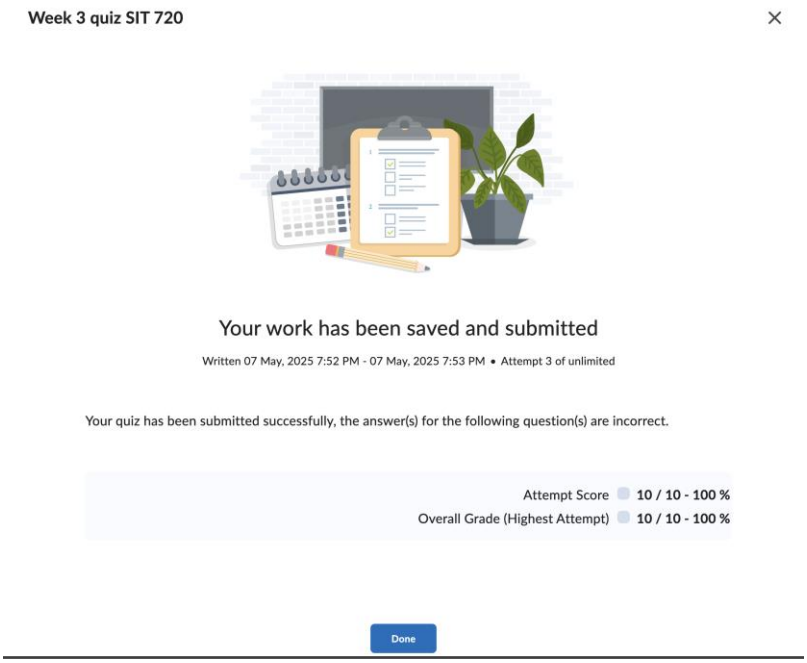
3. Reflection on knowledge gained by reading contents of the week 3 and 4 with respect to machine learning.

- The core idea of clustering is intuitive but learning how to measure that distance properly is key. K-means feels like the foundational tool for this, straightforward in principle but with nuances like needing to pick the right number of groups and where to start. Seeing there are other methods, and ways to check if the grouping

actually makes sense even without the right answers, shows it's a whole field of figuring out hidden structures in data.

- Learning about dimensionality reduction, especially PCA, feels like getting a powerful tool to tackle this. It's like being able to find the main underlying patterns and simplify things without losing the most important information. It makes complex data sets feel a lot more manageable, like finding the most essential summary instead of reading every single detail. And seeing how you can actually apply it makes it feel really practical.

4. Attempted Week 3 and 4 Quiz results



Week 4 Quiz for SIT720

Your work has been saved and submitted

Written 07 May, 2025 8:02 PM - 07 May, 2025 8:03 PM • Attempt 5 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score	**10 / 10 - 100 %**
Overall Grade (Highest Attempt)	**10 / 10 - 100 %**

Done