**Homework – Questions 3.1, 4.1, and 4.2**

Note: Deeper analysis for code methodology is inside both code files I have.

**Question 3.1**  
Using the same data set (credit\_card\_data.txt or credit\_card\_data-headers.txt) as in Question 2.2, use the ksvm or kknn function to find a good classifier:  
(a) using cross-validation (do this for the k-nearest-neighbors model; SVM is optional); and  
(b) splitting the data into training, validation, and test data sets (pick either KNN or SVM; the other is optional).

**Methodology:**  
I applied the k-nearest-neighbors (KNN) algorithm to the credit card dataset. For part (a), I used leave-one-out cross-validation and tested multiple values of k through a validation loop. This allowed me to evaluate model performance across different neighborhood sizes. For part (b), I split the dataset into training (60%), validation (20%), and test (20%) subsets. The validation set was used to select the best k, and the final model was retrained on the combined training + validation data before being tested on the held-out test set.

**Results:**

* **Cross-validation:** Accuracy was maximized at **k = 11**, with mean classification accuracy of approximately **84%**.
* **Train/validation/test split:** Validation results also favored **k = 11** with accuracy ≈ **83.9%**. Retraining the model on train+validation data and evaluating on the test set gave a final test accuracy of **83.9%**.

**Discussion:**  
The results indicate that **k = 11** provides the best generalization on this dataset. Accuracy was consistent across validation and test splits, showing the model is stable. While support vector machines (SVMs) can also achieve strong performance, the KNN model was sufficient for this assignment and provided interpretable results.

**Question 4.1**  
Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

**Response:**

I work with the stock market daily, and major current events or differences in 10-Q and 10-k reports versus expectations often impact how I manage my portfolio. When analyzing stocks and ETFs, I approach them from several different angles. For this assignment on clustering, I would focus on grouping securities by their exposure to risk factors.

**Predictors I would use include:**

1. Historical volatility
2. Beta to a market index
3. Market capitalization or return correlation with major factors
4. Valuation ratios (e.g., P/E, P/B)
5. Momentum scores (12-month rolling return)

This clustering approach would reveal natural groups of securities with similar profiles, which could then be used to rebalance or hedge my portfolio more effectively.

**Question 4.2**  
The Iris dataset contains 150 data points, each with four predictor variables (sepal length, sepal width, petal length, petal width) and one categorical response (flower species). The response is only for evaluation. Using kmeans, cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your clustering predicts flower type.

**Methodology:**  
I first explored clustering with the elbow method by plotting total within-cluster sum of squares (WSS) against k. The elbow point appeared near **k = 3**, which matches the number of true flower species. I then tested clustering with different predictor subsets: sepal features only, petal features only, and all four features.

**Results:**

* **Petal-only predictors (Petal.Length, Petal.Width)** provided the best clustering performance.
* With **k = 3**, all **setosa** samples were perfectly separated. Most overlap occurred between **versicolor** and **virginica**.
* The overall clustering accuracy (using majority-vote alignment of clusters to species) was **96%**.
* In comparison, using all four predictors gave ≈ 89% accuracy, while sepal-only features were weaker at ≈ 82%.

**Discussion:**  
The results confirm that **petal features carry the strongest discriminative structure** in the Iris dataset. K-means with k = 3 effectively captures species separation, particularly isolating setosa perfectly. Versicolor and virginica show partial overlap, which is a known limitation of linear clustering on this dataset.

**References:**

R Core Team. (2025). *kmeans: K-Means Clustering* [R Documentation]. R Project. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html>

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). STHDA. <https://www.sthda.com/english/articles/29-cluster-validation/97-cluster-validation-statistics-must-know-methods/>

OpenAI. (2025). *ChatGPT (March 2025 version)* [Large language model].

**GeeksforGeeks**. *“Elbow Method for Optimal Value of k in K-Means.”* GeeksforGeeks, 11 July 2025, <https://www.geeksforgeeks.org/machine-learning/elbow-method-for-optimal-value-of-k-in-kmeans/>.[XSLiuLab.github.io](https://xsliulab.github.io/Workshop/2021/week10/r-cluster-book.pdf?utm_source=chatgpt.com)[GeeksforGeeks+1](https://www.geeksforgeeks.org/machine-learning/elbow-method-for-optimal-value-of-k-in-kmeans/?utm_source=chatgpt.com)