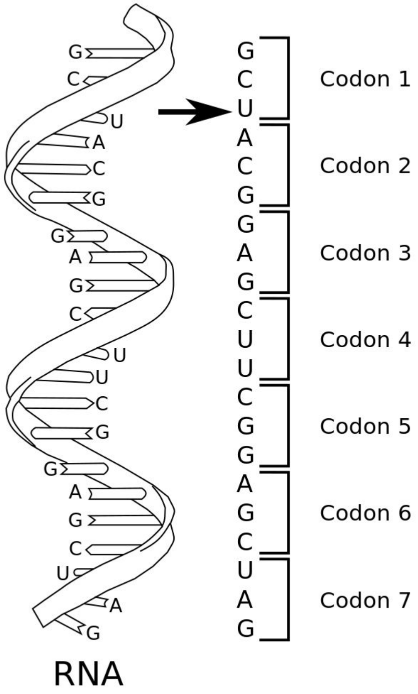
Codon Frequency Classification Project

By Pier Bruno Pompilii

Problem Overview



Ribonucleic acid

This project aims to classify species into different kingdoms based on the frequencies of codons in their genomic coding DNA sequences that are transcribed from the RNAm. The dataset contains codon usage frequencies for various species, and the goal is to build a classification model to predict the kingdom to which a species belongs.

“Solution”:

- Machine Learning
  - KNN
  - Logistic Regression
  - Unsupervised: Clustering

Dataset Overview

The dataset is a CSV file with 13,026 entries and 69 columns. Each row represents a species, and the columns include:

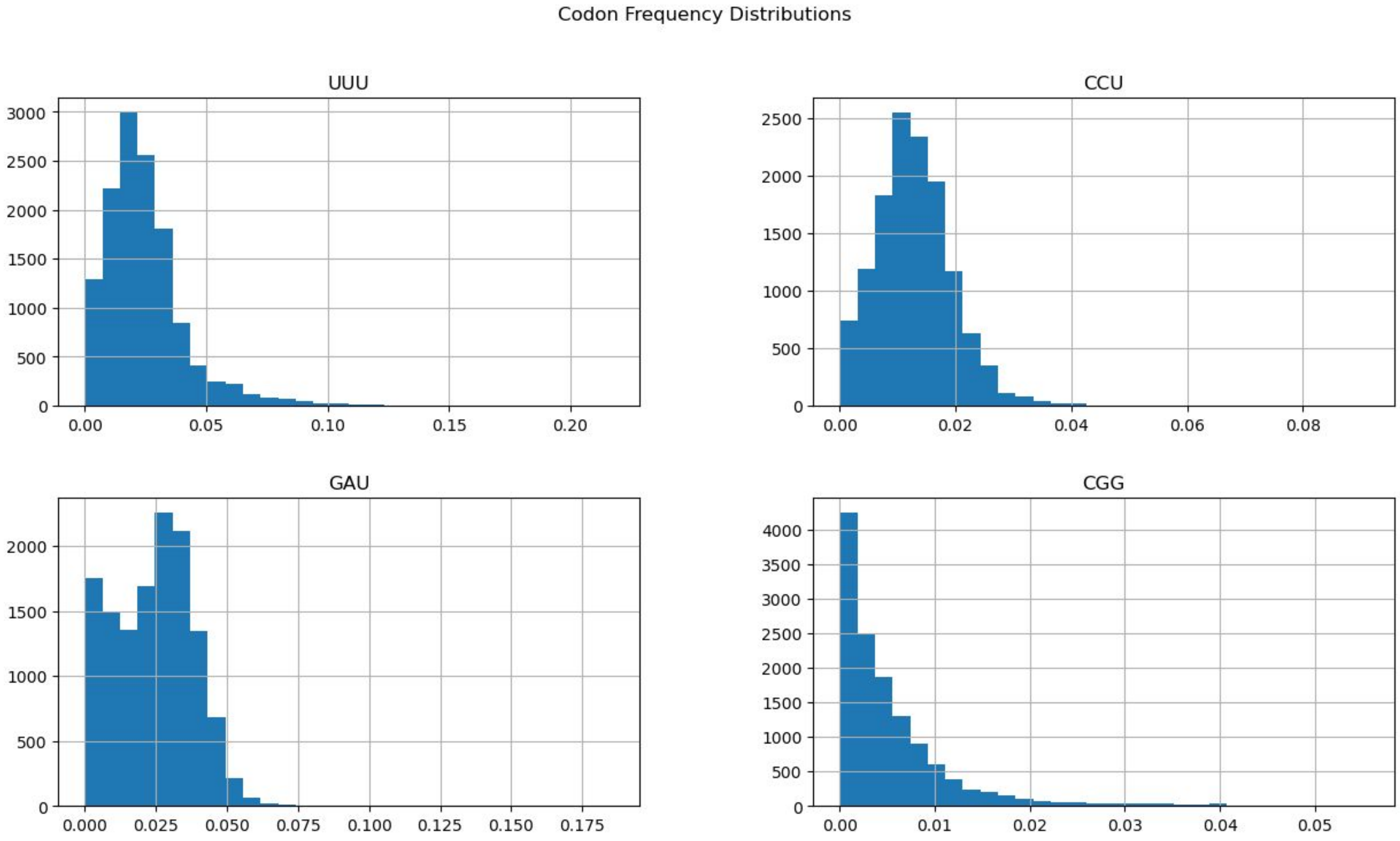
- **Kingdom**: Biological classification (e.g., Animalia, Plantae).
- **DNAtype**: Type of DNA (e.g., mitochondrial, nuclear).
- **SpeciesID**: Numeric identifier for each species.
- **Ncodons**: The algebraic sum of the numbers listed for the different codons. Codon frequencies are normalized to the total codon count, hence the number of occurrences divided by 'Ncodons' is the codon frequencies listed in the data file.
- **SpeciesName**: Name of the species.
- **UUU to UGA:** Frequencies of each codon

Hallee, L., Khomtchouk, B.B. Machine learning classifiers predict key genomic and evolutionary traits across the kingdoms of life. Sci Rep 13, 2088 (2023)

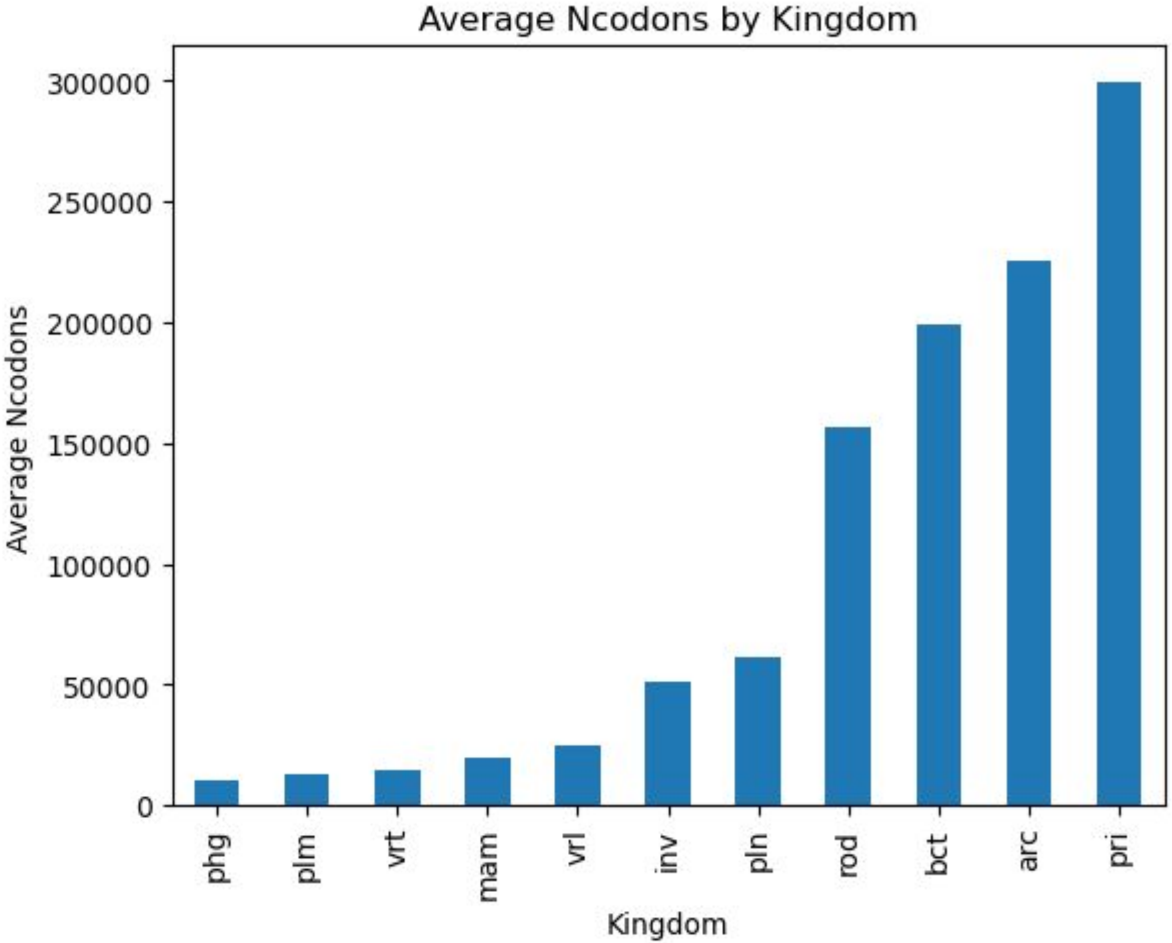
After exploration of the data set :

- After exploring the dataset, found no duplicate values.
- There was an issue with two rows containing strings in a numerical column. This problem was resolved by deleting the affected rows, which only constituted 2 out of a total of 13,026.
- The dataset is now clean and ready for analysis.

Insights from EDA

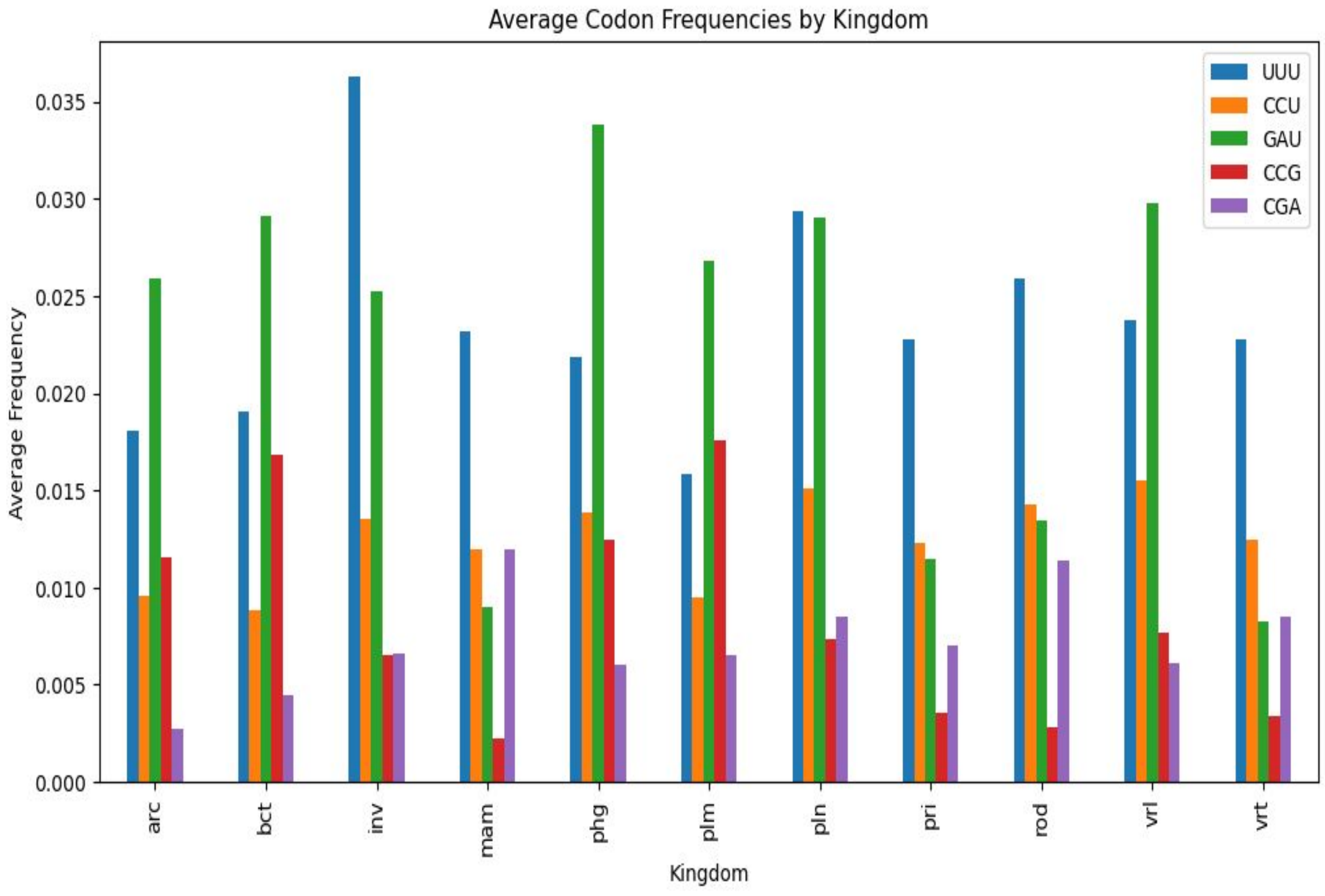


The distribution of codons is not normal.

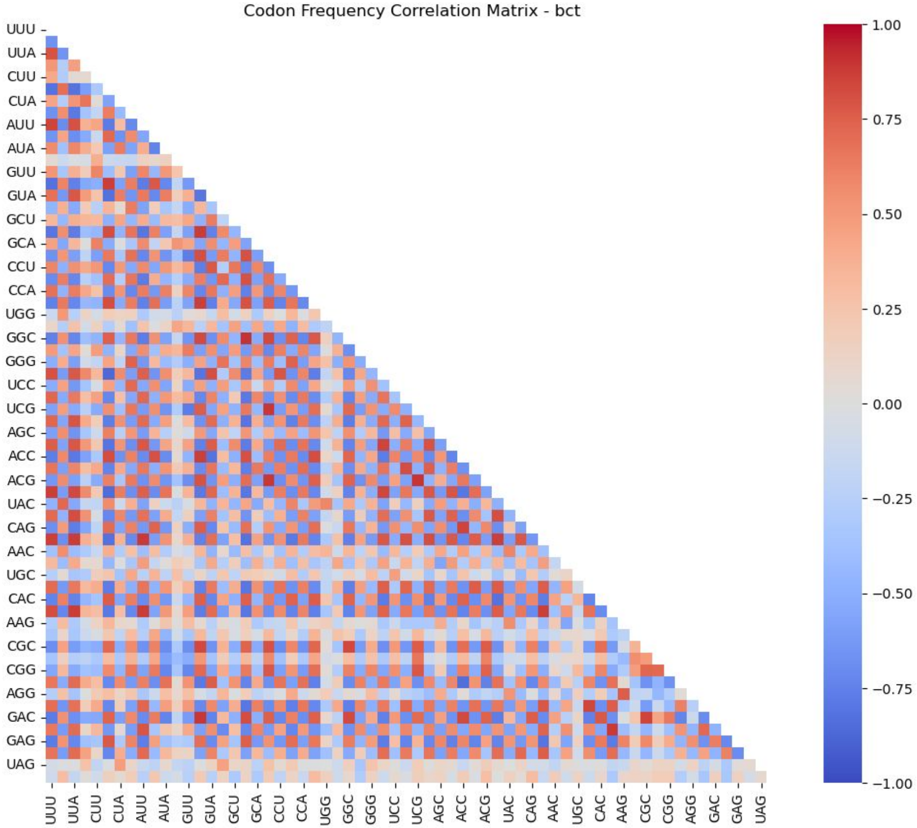


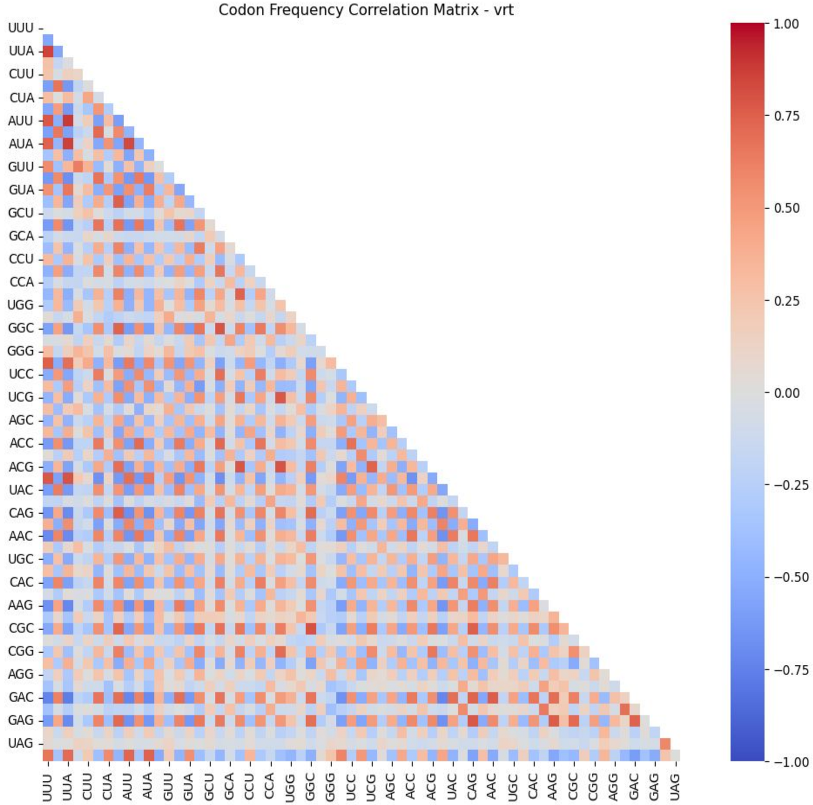
Organisms with higher codon counts might have more genes or longer genes, reflecting more complex

Further analysis needed



The distribution of codons varies across different kingdoms.





There is colinearity among codons, and it varies across kingdoms.

What is next

- An exploratory logistic regression has been conducted, necessitating further exploration due to the complexity of analyzing 64 codons.

- The kingdoms will be separated into 6 distinct categories, along with an additional grouping (virus).

- Models such as KNN and unsupervised clustering will be applied for further analysis.