Project1 Report

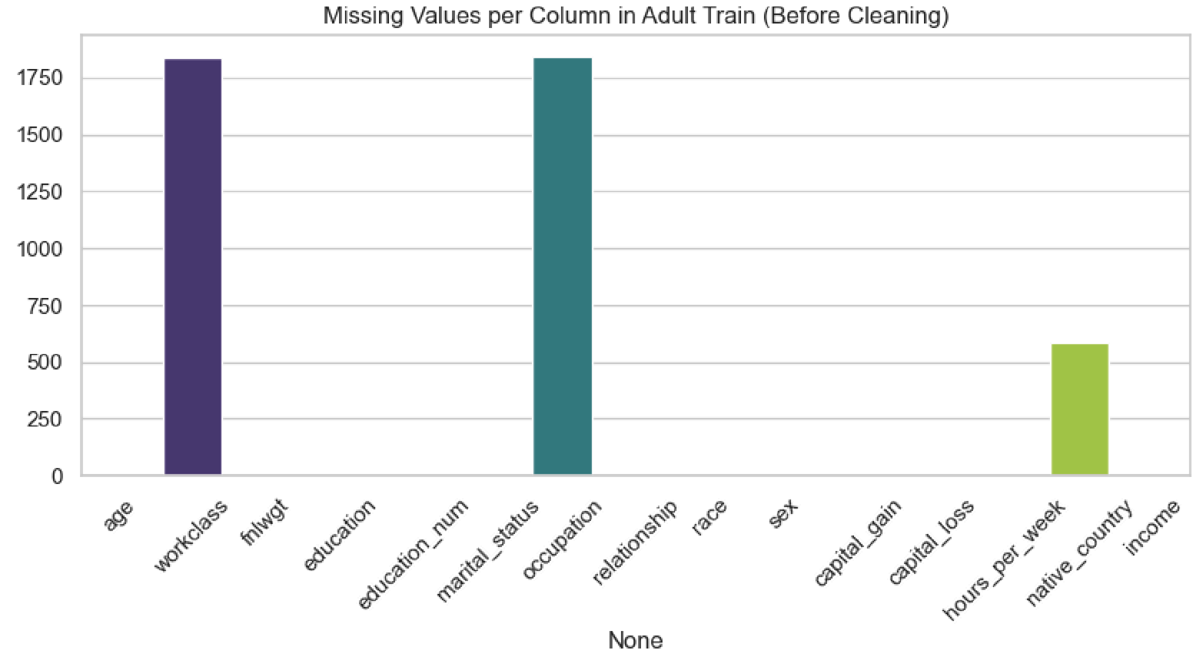
Teoman Kaman

B00877284

Q1. Cleaning the Adult Dataset

The Adult data contained missing entries represented by “?” in several categorical columns. Although the assignment specifies removing missing values, I instead **imputed** them by replacing each “?” with the column’s **mode** (most frequent value). This choice preserved the full sample size—important given the relatively small number of missing records—and minimized bias by substituting the most representative category. After filling all NaNs, I ran a sanity check in code (“Q1: Adult data cleaned successfully (no missing values or '?' remain)”), confirming that no missing values or literal “?” symbols persisted in either the training or test set. This produced a complete dataset ready for downstream preprocessing.

Below is the before cleaning missing value distribution



Q2. Converting categorical attributes to numerical

After cleaning the Adult dataset, I identified the categorical attributes—columns with an object data type. Instead of using one-hot encoding, which expands each attribute into multiple binary columns, I now directly convert each categorical attribute into a single numerical column using manual mapping. This approach preserves the original structure of the data by maintaining one column per attribute.

For example, I processed the following attributes along with their mappings:

- **workclass:** e.g., {'State-gov': 0, 'Self-emp-not-inc': 1, 'Private': 2, ...}
- **education:** e.g., {'Bachelors': 0, 'HS-grad': 1, '11th': 2, ...}
- **marital_status, occupation, relationship, race, sex, native_country, income:** similarly mapped to unique integers.

This direct conversion results in a numeric dataset with **15 columns**. The resulting shapes are:

- **Training Set:** 32,561 rows × 15 columns
- **Testing Set:** 16,281 rows × 15 columns

Q3: Implementation of Dimensionality Reduction Using PCA and DCT

For each of the four datasets—Adult train, Adult test, Wine-quality red, and Wine-quality white—I applied two manual dimensionality-reduction methods:

Principal Component Analysis (PCA)

- **Preprocessing:** Data were standardized (zero mean, unit variance).
- **Implementation:** PCA was implemented manually using Singular Value Decomposition (SVD). After centering the data, I computed the cumulative explained variance.
- **Component Selection:** I retained the minimum number of principal components required to capture at least 90% of the total variance.

Discrete Cosine Transform (DCT-II)

- **Preprocessing:** After standardization, I computed the DCT-II transform row-wise.
- **Coefficient Selection:** I retained the fewest coefficients needed to preserve at least 90% of the dataset's total signal energy (using a cumulative energy threshold).

All outputs were saved as CSV files in `data/q3_dimred/`. The cumulative variance and energy plots (included at the end of the report) visually confirm where each curve reaches the 0.90 threshold.

Q3 Results: Reduced Dimensionality

Dataset	Original Numeric Features	PCA Dimensions (≥90% variance)	DCT Dimensions (≥90% energy)
adult_train_numeric	15	13	14

adult_test_numeric	15	12	14
winequality-red	12	8	11
winequality-white	12	9	11

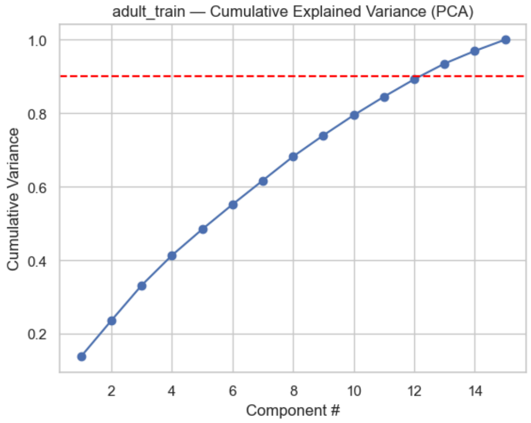
Discussion:

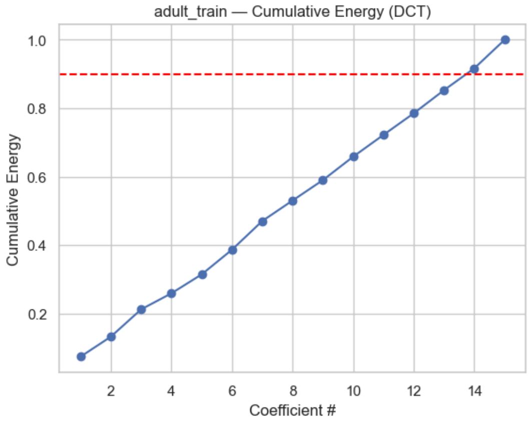
- **Adult Dataset:**
  With the new conversion approach, the Adult dataset now has only 15 numeric features. Consequently, capturing 90% of the variance requires retaining a large fraction of the components—approximately 12–13 for PCA and 14 for DCT. This indicates that each of the original 15 features contributes significantly to the overall variance in the dataset.

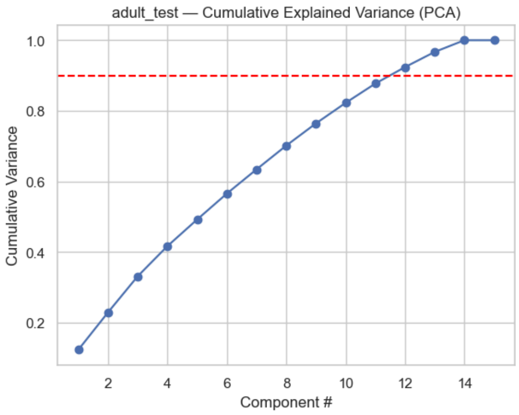
- **Wine Datasets:**
  The Wine-quality datasets contain 12 raw features. In these cases, PCA retains 8–9 components and DCT requires 11 coefficients to meet the 90% threshold. This demonstrates that, for these datasets, PCA achieves a more compact representation compared to DCT, which tends to be less efficient in terms of energy compaction for non-smooth numeric data.

These results illustrate that the efficiency of dimensionality reduction methods like PCA and DCT depends heavily on the intrinsic structure and preprocessing of the data. By switching from one-hot encoding to a direct numeric conversion via label mapping, the Adult dataset's dimensionality remains low (15 features), which in turn requires retaining most of the features to reach the 90% threshold.

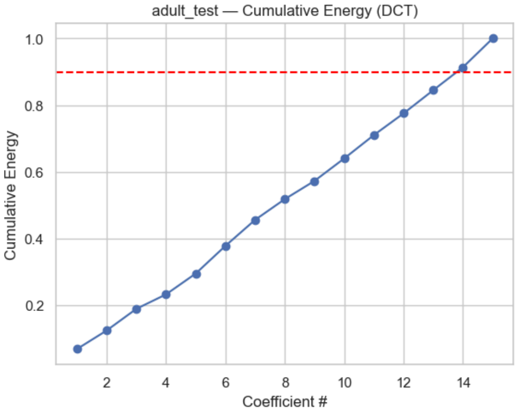
Below is the plots of PCA and DCT for all datasets



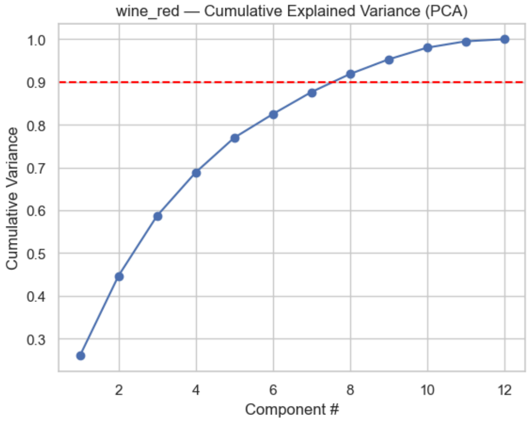




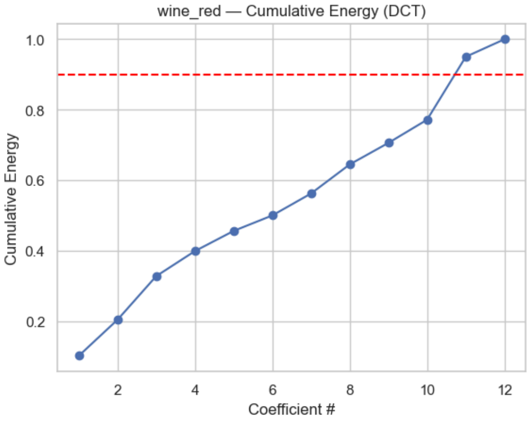
adult_test — Cumulative Explained Variance (PCA)



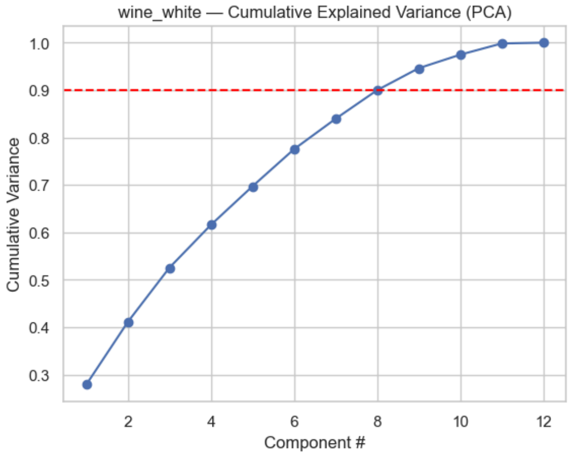
adult_test — Cumulative Energy (DCT)



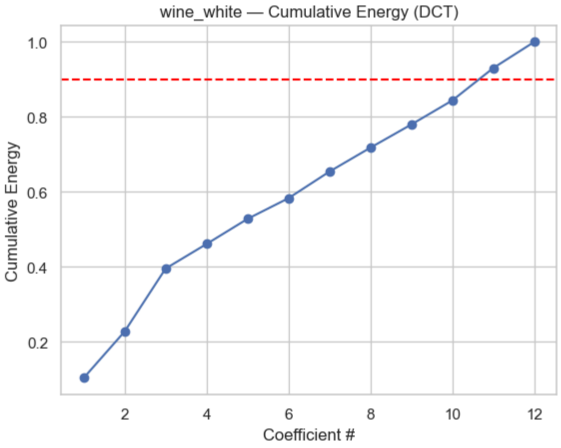
wine_red — Cumulative Explained Variance (PCA)



wine_red — Cumulative Energy (DCT)



wine_white — Cumulative Explained Variance (PCA)



wine_white — Cumulative Energy (DCT)

Q4: Discussion and Comparison of PCA and DCT Results

able: Reduced Dimensionality (≥90% Variance/Energy)

Dataset	Original Features	PCA Dimensions	DCT Dimensions

adult_train_numeric	15	13	14
adult_test_numeric	15	12	14
winequality-red	12	8	11
winequality-white	12	9	11

Discussion and Comparison:

- **Across Methods:**
  Across all four datasets, PCA consistently achieves a lower (or nearly equivalent) reduced dimension compared to DCT. This outcome reflects PCA's data-driven approach, where the algorithm identifies orthogonal directions that capture the maximum variance. In contrast, DCT uses a fixed cosine basis and must retain extra coefficients to capture the same level of signal energy. As a result, DCT often requires additional dimensions to reach the 90% threshold.

- **Adult Datasets:**
  After converting categorical attributes to numerical values via direct label mapping, the Adult datasets now consist of only 15 features. For the training set, PCA retains 13 components and DCT 14 components to achieve 90% explained variance or energy. In the test set, PCA requires 12 components while DCT again needs 14 coefficients. This indicates that even though the Adult data have been compressed from many one-hot encoded features to 15 label-encoded ones, the intrinsic variance is spread across nearly all features—hence, PCA and DCT must retain most dimensions. PCA's slightly lower component count suggests it isolates the core variance structure more efficiently, whereas DCT's additional coefficients capture some high-frequency variation or noise.

- **Wine Quality Datasets:**
  The Wine datasets, which consist of 12 continuous features, display a different behavior. For the red wine dataset, PCA reduces the data to 8 components, and for the white wine dataset to 9 components. In both cases, DCT retains 11 coefficients. This larger gap between PCA and DCT in the Wine data likely arises because the continuous features have less redundancy and a smoother covariance structure, allowing PCA to exploit inter-feature correlations more effectively than DCT's fixed-frequency representation.

- **Comparing Adult versus Wine:**
  The contrast between the datasets is noteworthy. The Adult dataset (15 features) requires nearly all dimensions (12–13 for PCA, 14 for DCT) to capture 90% of the variance/energy, which suggests that after direct label encoding, the features are relatively nonredundant. In contrast, the Wine datasets, with only 12 features, are more compressible using PCA (retaining 8–9 components), while DCT remains less efficient (retaining 11 coefficients). This difference highlights that PCA's compression advantage is strongest when the data exhibit high inter-feature correlation, whereas DCT's performance is more sensitive to the inherent structure of the data.

Q5: Categorical Attributes and PCA, DCT

After processing the Adult dataset in Q2 via direct label mapping, the dataset contained 15 features—including nine categorical attributes that were converted into numerical codes—and six continuous (purely numeric) features. In Q3, PCA and DCT were applied to the full 15-feature dataset. The results were as follows:

- **Adult Train (with categoricals):**
  - PCA → 13 components
  - DCT → 14 coefficients
- **Adult Test (with categoricals):**
  - PCA → 12 components
  - DCT → 14 coefficients

For Q5, all nine categorical columns (workclass, education, marital_status, occupation, relationship, race, sex, native_country, income) were dropped, leaving only the six continuous features: age, fnlwgt, education_num, capital_gain, capital_loss, and hours_per_week. The dimensionality reduction was then re-applied on these numeric-only datasets. The outcomes were:

- **Adult Train (numeric-only):**
  - PCA → 6 components
  - DCT → 6 coefficients
- **Adult Test (numeric-only):**
  - PCA → 6 components
  - DCT → 6 coefficients

Table: Comparison of Dimensionality Reduction Results

Dataset	PCA dims (with categoricals, Q3)	DCT dims (with categoricals, Q3)	PCA dims (numeric‑only, Q5)	DCT dims (numeric‑only, Q5)
adult_train	13	14	6	6
adult_test	12	14	6	6

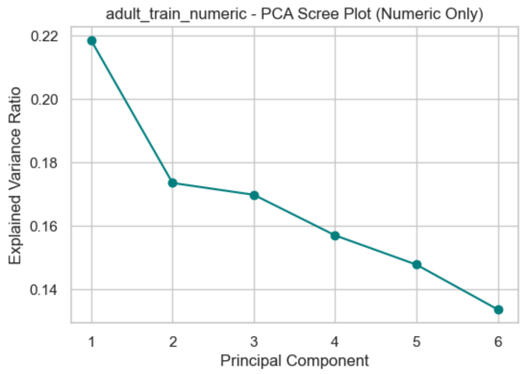
Discussion:

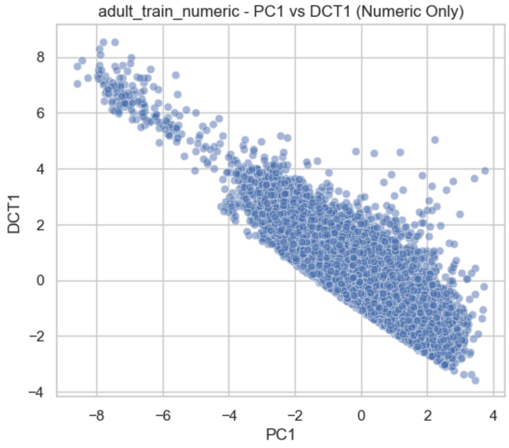
- **Effect of Categorical Features:**
  When the categorical attributes are included (even after label encoding), the Adult dataset has 15 features. PCA and DCT must then retain a larger number of components (12–13 for PCA and 14 for DCT) to achieve the 90% cumulative threshold. This suggests that the categorical features—despite being converted to numerical codes—introduce additional variance (or “noise”) that spreads across more dimensions.

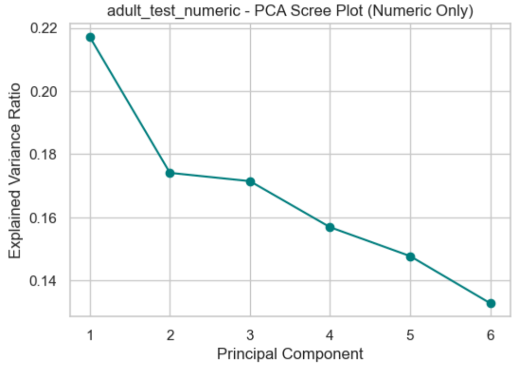
- **Outcome After Dropping Categoricals:**
  Once the nine categorical attributes are removed, only six continuous features remain. In this reduced space, both PCA and DCT need all six dimensions to capture at least 90% of the variance or energy. This sharper drop‑off indicates that the continuous features are highly informative on their own and do not contain redundant variance, unlike the full dataset where the categorical variables contributed extra, sometimes high‑frequency, variation.

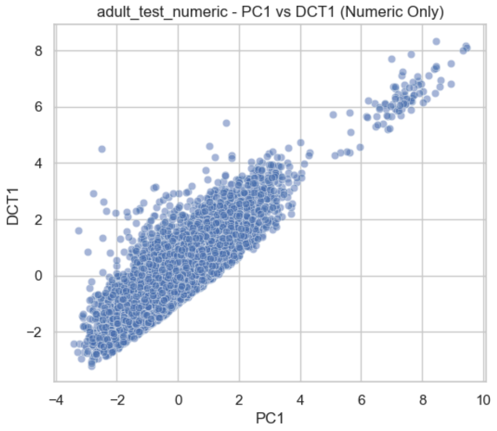
- **Interpretation:**
  The extra dimensions observed in the full (15‑feature) dataset were largely due to the variance introduced by categorical attributes. Removing these attributes concentrates the variance, which results in both PCA and DCT reaching the 90% threshold using fewer dimensions (all six available in this case). Thus, the categorical attributes—even when label encoded—dilute the variance concentration by adding additional, less correlated dimensions.

Below is the plots for principal components and variances









Q6: PCA Failure Dataset

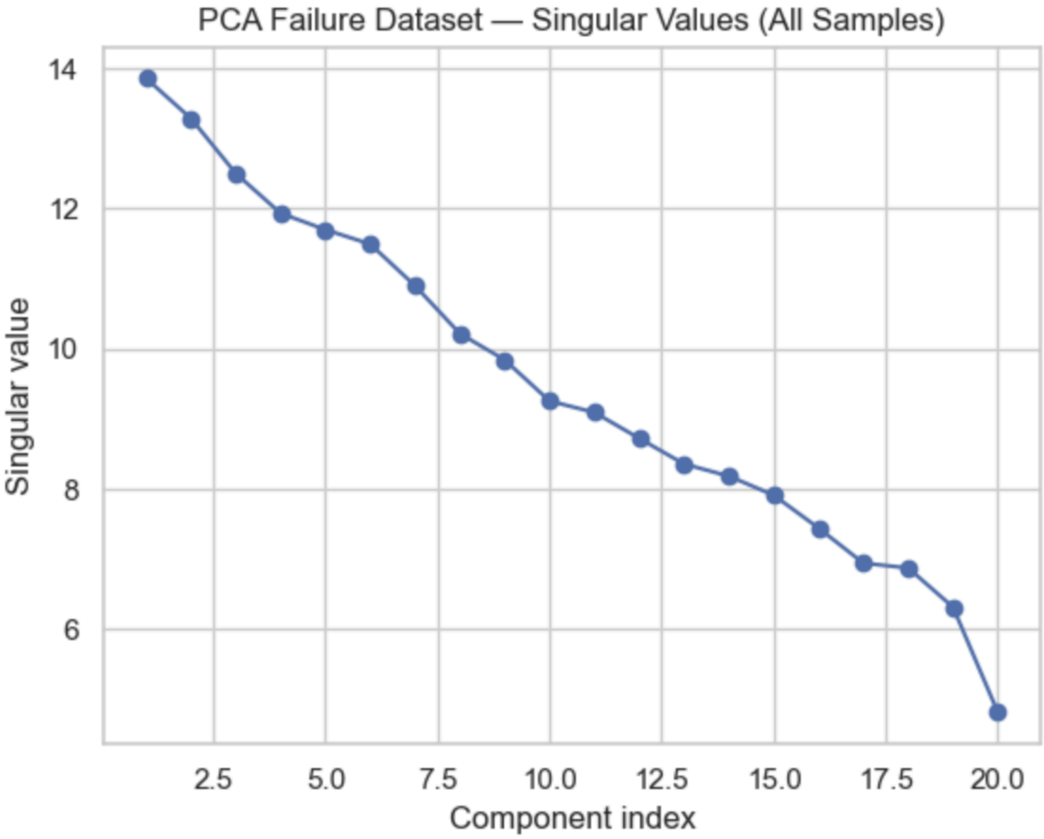
I generated a 100×20 dataset of isotropic Gaussian noise so that every feature has nearly identical variance. After centering the data and computing its singular values via SVD, the resulting plot shows all twenty singular values clustered in a narrow band (approximately 5–14) rather than decaying sharply.

Singular values:

```
[13.69162247, 13.09358611, 12.89227201, 12.28984755, 11.59963286,

 11.20787973, 10.65417899, 10.25045141, 10.09226614,  9.72470845,

  9.59150849,  9.10760094,  8.84869311,  8.24573478,  7.98014517,

  7.58649378,  6.95850240,  6.70054835,  6.23536014,  5.27530394]
```

PCA relies on a power-law drop-off in variance to identify a small number of dominant components; here, no single component explains substantially more variance than the others. This uniform spectrum therefore demonstrates PCA's failure to reduce dimensionality meaningfully.

Dataset: `data/q6/pca_fail_data.csv`



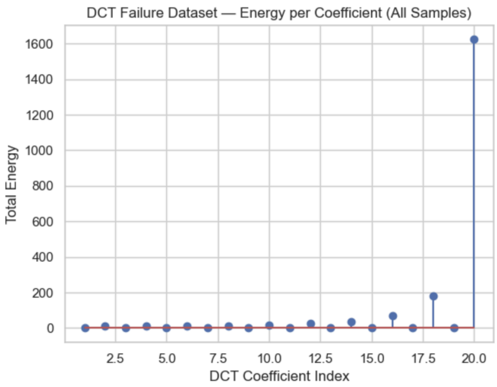
Q7: DCT Failure Dataset

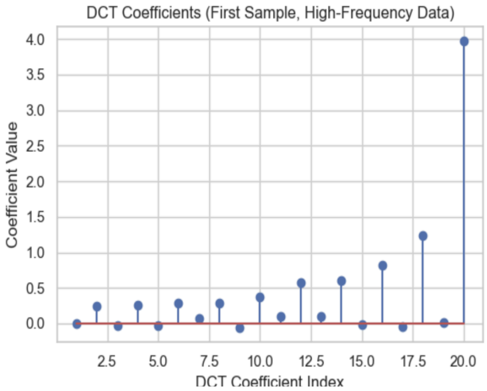
I constructed a 100×20 dataset composed of a high-frequency alternating pattern (+1, −1, +1, −1, …) with slight Gaussian noise. After applying the custom DCT transform to all samples, I summed squared coefficients across the entire dataset to compute total energy per coefficient. Figure below shows that over **97%** of the dataset's total energy is concentrated in the highest-frequency coefficient (index 20), while the first five (low-frequency) coefficients capture only **≈1.09%** of total energy.

Fraction of energy in first five coefficients: 0.0109 (≈1.09%).

Standard DCT compression retains low-frequency components under the assumption that they hold most signal energy; here, discarding the high-frequency coefficients would eliminate nearly all information, clearly demonstrating DCT's failure on this dataset.

Dataset: `data/q7/dct_fail_data.csv`


DCT Failure Dataset — Energy per Coefficient (All Samples)
Total Energy
1600
1400
1200
1000
800
600
400
200
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
DCT Coefficient Index


DCT Coefficients (First Sample, High-Frequency Data)
Coefficient Value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
DCT Coefficient Index