Data-Driven Classification of Canadian Crude Oils

Ryan Bulger

University of Colorado Boulder

DTSA 5506

ABSTRACT

This project applies unsupervised machine learning to evaluate whether the current classification of Canadian crude oil grades accurately represents their chemical and physical relationships. Using clustering algorithms—K-Means, Agglomerative Clustering, and Gaussian Mixture Models—applied to CrudeMonitor.ca assay data, the analysis identifies natural groupings based on compositional similarity. Across all models, cluster sizes of two to three produced the highest performance scores and most distinct separations, while larger cluster counts resulted in overlapping boundaries. These results, supported by supervised learning comparisons, indicate that Canadian crude oils form a continuous spectrum better represented by a few broad categories rather than eleven discrete grades. The findings suggest that a simplified classification system could more accurately reflect the natural structure of Canadian crudes and improve future modeling, transportation, and market analysis.

1. INTRODUCTION

This study addresses the problem of whether hidden patterns within Canadian crude oil quality data can reveal more natural groupings of crude oils than those defined by current classification systems. Accurate grouping is important because the existing categories, often based only on density and sulfur content, may not fully capture the chemical and physical complexity of each crude. As a result, misclassification can distort pricing, complicate blending and transport decisions, and reduce overall efficiency in the supply chain.

To address these limitations, this project applies unsupervised machine learning techniques to cluster crude oils based on a comprehensive range of laboratory-measured properties. By uncovering data-driven groupings that reflect intrinsic relationships between crudes, the analysis provides a more objective framework for understanding quality variation. The findings can support more accurate market valuation, optimize pipeline and blending operations, and improve refinery yield predictability within the Canadian energy sector.

The analysis proceeds through several structured stages: data summary to understand the dataset's structure, exploratory data analysis (EDA) to identify patterns and correlations, and data cleaning to handle missing or inconsistent values. Three clustering algorithms, K-Means, Agglomerative Clustering, and Gaussian Mixture Models (GMM), are then applied and evaluated using multiple performance metrics and visualized through PCA and t-SNE. The unsupervised results are later compared with supervised learning models, Multinomial Logistic Regression, Support Vector Classifier, and Gradient Boosting, to assess alignment with true class labels. The project concludes with final discussion and takeaways, highlighting key insights and implications for applying machine learning to crude quality analysis in the energy industry.

2. RELATED WORK

Over the past two decades, several approaches have been developed to classify crude oil grades based on their physical and chemical properties. Traditional methods rely on API gravity and sulfur content to distinguish between light, medium, and heavy crudes, as well as sweet and sour categories [1]. While effective for general use, these classifications overlook finer compositional differences. Chemical engineering studies have expanded on this by incorporating indices such as the Watson characterization factor and correlation index to better capture variations in paraffinic and aromatic content [2].

More recently, researchers have introduced spectroscopic and chemometric techniques to automate and improve accuracy in crude classification. Studies using Fourier Transform Infrared (FTIR) spectroscopy, Near-Infrared (NIR) spectroscopy, and Nuclear Magnetic Resonance (NMR) have demonstrated that supervised machine learning models such as Support Vector Machines and Partial Least Squares Discriminant Analysis (PLS-DA) can predict crude oil grades and key properties with high accuracy [3][4]. These methods provide faster, non-destructive analysis compared to conventional lab assays.

Unsupervised statistical methods have also been used to discover natural groupings within crude oil datasets. Techniques like K-Means, hierarchical clustering, and self-organizing maps have revealed that many crude oils can be clustered into fewer groups than their assigned categories, often aligning with geographic origin or refining characteristics [5]. This supports the idea that data-driven clustering can yield more meaningful categorizations than fixed classification systems.

The proposed study builds upon this prior work by combining multiple clustering techniques, K-Means, Agglomerative Clustering, and Gaussian Mixture Models, and comparing them with supervised learning models. Unlike previous studies that focused on narrow datasets or specific analytical tools, this project integrates a broad range of crude quality features to evaluate whether unsupervised methods can uncover more representative groupings, offering both technical and commercial insights into crude oil classification.

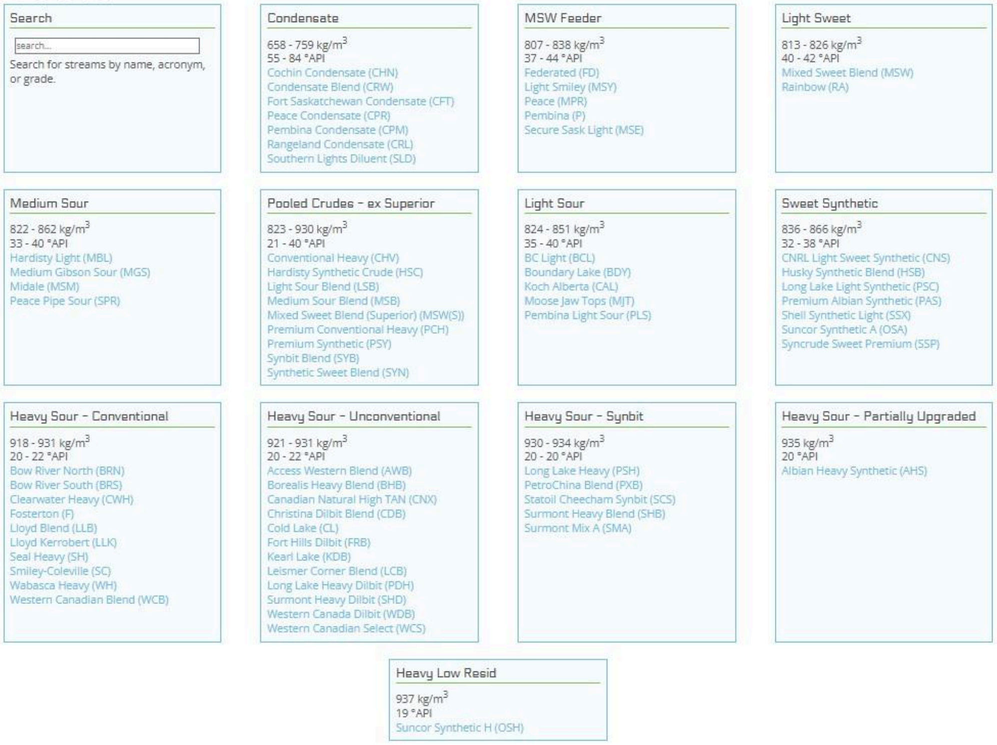
3. METHODOLOGY

3.1 Data Source

The dataset for this project is derived from CrudeMonitor.ca, an established platform managed by Crude Quality Inc., which collects and publishes detailed crude oil assay data from pipelines and refineries across Western Canada [6]. Each assay provides extensive information on the chemical and physical characteristics of crude oils, including density, sulfur content, metals (nickel, vanadium), distillation yields, and hydrocarbon compositions. As shown in Figure 1, CrudeMonitor categorizes these crude oils into 12 defined groups based on quality and refining characteristics.

Figure 1

CrudeMonitor crude oil grades



Search

search...

Search for streams by name, acronym, or grade.

Condensate

658 - 759 kg/m^3
55 - 84 °API
Cochin Condensate (CHN)
Condensate Blend (CRW)
Fort Saskatchewan Condensate (CFT)
Peace Condensate (CPR)
Pembina Condensate (CPM)
Rangeland Condensate (CRL)
Southern Lights Diluent (SLD)

MSW Feeder

807 - 838 kg/m^3
37 - 44 °API
Federated (FD)
Light Smiley (MSY)
Peace (MPR)
Pembina (P)
Secure Sask Light (MSE)

Light Sweet

813 - 826 kg/m^3
40 - 42 °API
Mixed Sweet Blend (MSW)
Rainbow (RA)

Medium Sour

822 - 862 kg/m^3
33 - 40 °API
Hardisty Light (MBL)
Medium Gibson Sour (MGS)
Midale (MSM)
Peace Pipe Sour (SPR)

Pooled Crudes – ex Superior

823 - 930 kg/m^3
21 - 40 °API
Conventional Heavy (CHV)
Hardisty Synthetic Crude (HSC)
Light Sour Blend (LSB)
Medium Sour Blend (MSB)
Mixed Sweet Blend (Superior) (MSW(S))
Premium Conventional Heavy (PCH)
Premium Synthetic (PSY)
Synbit Blend (SYB)
Synthetic Sweet Blend (SYN)

Light Sour

824 - 851 kg/m^3
35 - 40 °API
BC Light (BCL)
Boundary Lake (BDY)
Koch Alberta (CAL)
Moose Jaw Tops (MJT)
Pembina Light Sour (PLS)

Sweet Synthetic

836 - 866 kg/m^3
32 - 38 °API
CNRL Light Sweet Synthetic (CNS)
Husky Synthetic Blend (HSB)
Long Lake Light Synthetic (PSC)
Premium Albian Synthetic (PAS)
Shell Synthetic Light (SSX)
Suncor Synthetic A (OSA)
Syncrude Sweet Premium (SSP)

Heavy Sour – Conventional

918 - 931 kg/m^3
20 - 22 °API
Bow River North (BRN)
Bow River South (BRS)
Clearwater Heavy (CWH)
Fosterton (F)
Lloyd Blend (LLB)
Lloyd Kerrobert (LLK)
Seal Heavy (SH)
Smiley-Coleville (SC)
Wabasca Heavy (WH)
Western Canadian Blend (WCB)

Heavy Sour – Unconventional

921 - 931 kg/m^3
20 - 22 °API
Access Western Blend (AWB)
Borealis Heavy Blend (BHB)
Canadian Natural High TAN (CNX)
Christina Dilbit Blend (CDB)
Cold Lake (CL)
Fort Hills Dilbit (FRB)
Kearl Lake (KDB)
Leismer Corner Blend (LCB)
Long Lake Heavy Dilbit (PDH)
Surmont Heavy Dilbit (SHD)
Western Canada Dilbit (WDB)
Western Canadian Select (WCS)

Heavy Sour – Synbit

930 - 934 kg/m^3
20 - 20 °API
Long Lake Heavy (PSH)
PetroChina Blend (PXB)
Statoil Cheecham Synbit (SCS)
Surmont Heavy Blend (SHB)
Surmont Mix A (SMA)

Heavy Sour – Partially Upgraded

935 kg/m^3
20 °API
Albian Heavy Synthetic (AHS)

Heavy Low Resid

937 kg/m^3
19 °API
Suncor Synthetic H (OSH)

The platform serves as the industry-standard source for Canadian crude oil quality data, used extensively by producers, refiners, marketers, research institutions, and government agencies. Because the data is both publicly accessible and quality-assured, CrudeMonitor is an ideal and trusted source for developing a transparent and reproducible dataset for this project.The project is being developed in Python using Jupyter Notebooks within Visual Studio Code (VS Code). All scripts and supporting materials are stored and version-controlled through a public GitHub repository. The dataset itself is relatively small so it is stored in CSV format rather than a database, allowing for efficient access and portability within the repository.

Initially, CrudeMonitor defined 12 groups of Canadian crude oils; however, upon further inspection, the Condensate group was found to contain a unique set of features that are not directly comparable to the remaining crude grades. To maintain analytical consistency, the Condensate group was excluded from the study, resulting in a final dataset containing 11 crude oil grades for clustering analysis.

To construct the working dataset, a Python function was developed to interact directly with the CrudeMonitor API. This function loops through each of the 11 defined crude groups, extracts the associated crude assay data, and converts it into structured DataFrames. Once all groups are collected, the DataFrames are concatenated into a single unified dataset for analysis. This automated process ensures consistency across all groups and simplifies future data updates if new assays or crude types are introduced.

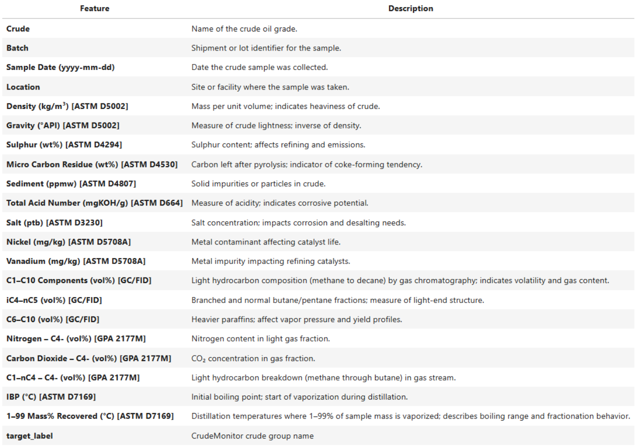
3.2 Data Summary

The first stage of analysis involved a systematic review of the dataset's structure and composition. After loading the dataset, it was found to contain 8,561 rows and 133 columns. The dataset consists primarily of numeric features, with 128 columns of type float32, one column of type datetime64[ns], and four columns of type object, which represent categorical or identifier fields.

A detailed table summarizing each column including its description, physical property, unit of measurement, and expected value range is presented in Figure 2. This summary establishes a clear understanding of the dataset's structure and provides the foundation for determining appropriate preprocessing, encoding, and modeling strategies in the subsequent stages of analysis.

Figure 2

Feature descriptions



Feature	Description
Crude	Name of the crude oil grade.
Batch	Shipment or lot identifier for the sample.
Sample Date (yyyy-mm-dd)	Date the crude sample was collected.
Location	Site or facility where the sample was taken.
Density (kg/m^3) [ASTM D5002]	Mass per unit volume; indicates heaviness of crude.
Gravity (°API) [ASTM D5002]	Measure of crude lightness; inverse of density.
Sulphur (wt%) [ASTM D4294]	Sulphur content; affects refining and emissions.
Micro Carbon Residue (wt%) [ASTM D4530]	Carbon left after pyrolysis; indicator of coke-forming tendency.
Sediment (ppmw) [ASTM D4807]	Solid impurities or particles in crude.
Total Acid Number (mgKOH/g) [ASTM D664]	Measure of acidity; indicates corrosive potential.
Salt (ptb) [ASTM D3230]	Salt concentration; impacts corrosion and desalting needs.
Nickel (mg/kg) [ASTM D5708A]	Metal contaminant affecting catalyst life.
Vanadium (mg/kg) [ASTM D5708A]	Metal impurity impacting refining catalysts.
C1–C10 Components (vol%) [GC/FID]	Light hydrocarbon composition (methane to decane) by gas chromatography; indicates volatility and gas content.
iC4–nC5 (vol%) [GC/FID]	Branched and normal butane/pentane fractions; measure of light-end structure.
C6–C10 (vol%) [GC/FID]	Heavier paraffins; affect vapor pressure and yield profiles.
Nitrogen – C4- (vol%) [GPA 2177M]	Nitrogen content in light gas fraction.
Carbon Dioxide – C4- (vol%) [GPA 2177M]	CO_2 concentration in gas fraction.
C1–nC4 – C4- (vol%) [GPA 2177M]	Light hydrocarbon breakdown (methane through butane) in gas stream.
IBP (°C) [ASTM D7169]	Initial boiling point; start of vaporization during distillation.
1–99 Mass% Recovered (°C) [ASTM D7169]	Distillation temperatures where 1–99% of sample mass is vaporized; describes boiling range and fractionation behavior.
target_label	CrudeMonitor crude group name

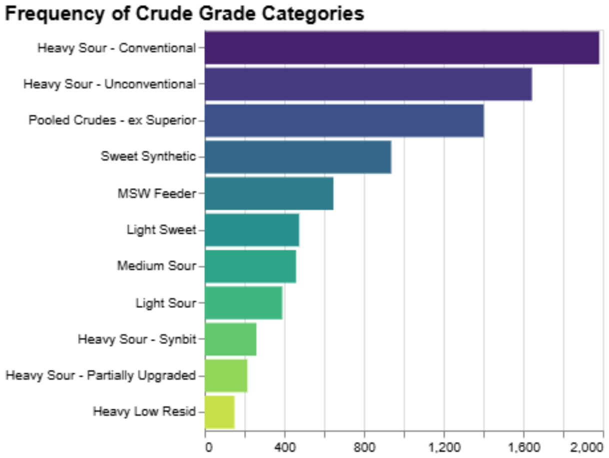
3.3 Exploratory Data Analysis

The exploratory analysis phase began with organizing the dataset into logical column groups to ensure consistency throughout the analysis process. A categorical group was created containing columns such as Sample Date, Location, and Crude. A second group of numeric columns was defined to capture features describing crude quality (e.g., density, sulfur, metals, and hydrocarbon composition). Finally, a third group of numeric columns was established for the distillation temperature measurements.

A frequency chart of Crude Grade Categories was created (Figure 3). The dataset is dominated by Heavy Sour – Conventional, Heavy Sour – Unconventional, Pooled Crudes – ex Superior, and Sweet Synthetic grades. This is expected since Western Canada primarily produces heavy crude oils derived from oil sands operations.

Figure 3

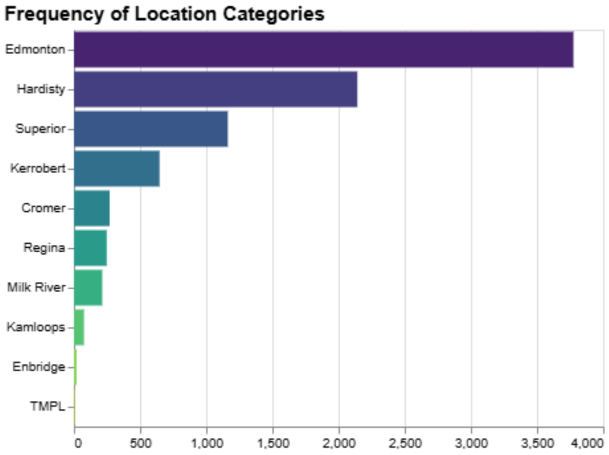
Crude grade categories



A frequency chart of Location Categories (Figure 4) shows that most samples originate from Edmonton, Hardisty, and Superior. This distribution aligns with the structure of Western Canada's pipeline system, where heavy crude and oil-sands by-products typically flow south from northern Alberta through Edmonton and then on to Hardisty, a major crude-oil hub.

Figure 4

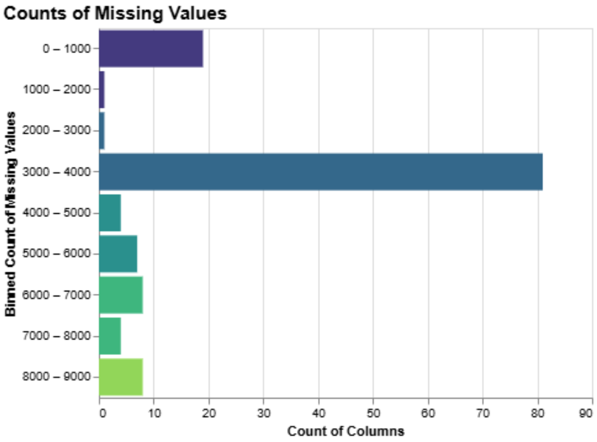
Location categories



A histogram comparing the number of columns to the binned count of missing values revealed that most columns have between 3,000 and 4,000 missing entries (Figure 5).

Figure 5

Missing value counts

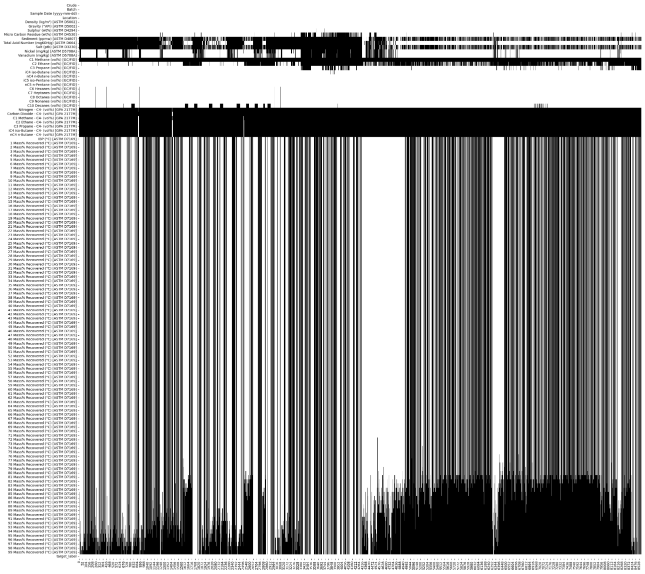


To visualize missing data more comprehensively, a missing-values heatmap was generated (Figure 6). The heatmap showed several columns that are almost entirely empty. These columns will be dropped during data cleaning, as imputing values where data are scarce would add bias. The proposed columns to drop are:

- Sediment (ppmw) [ASTM D4807],
- Total Acid Number (mgKOH/g) [ASTM D664],
- Salt (ptb) [ASTM D3230],
- C1 Methane (vol%) [GC/FID],
- Nitrogen – C4- (vol%) [GPA 2177M],
- Carbon Dioxide – C4- (vol%) [GPA 2177M],
- C1 Methane – C4- (vol%) [GPA 2177M],
- C2 Ethane – C4- (vol%) [GPA 2177M],
- C3 Propane – C4- (vol%) [GPA 2177M],
- iC4 iso-Butane – C4- (vol%) [GPA 2177M], and
- nC4 n-Butane – C4- (vol%) [GPA 2177M].

Figure 6

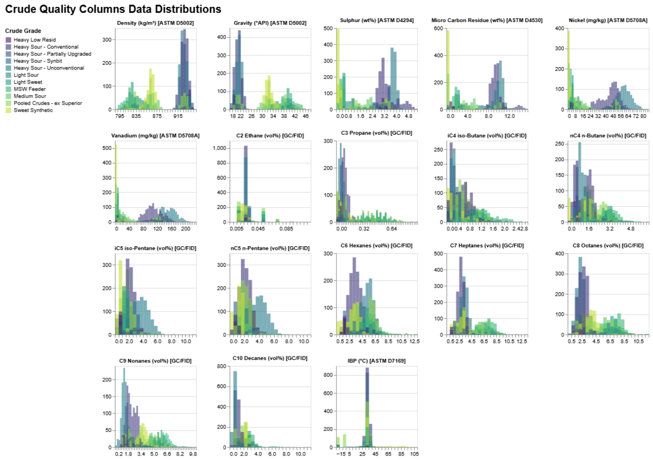
Missing values heatmap



Histograms of crude quality columns were then plotted, color-coded by target group (Figure 7). These plots reveal distinct groupings among certain crude grades, suggesting that several features (e.g., density, sulfur, metals, and hydrocarbon composition) are strong indicators of grade similarity and could enable the dataset to be clustered into fewer groups.

Figure 7

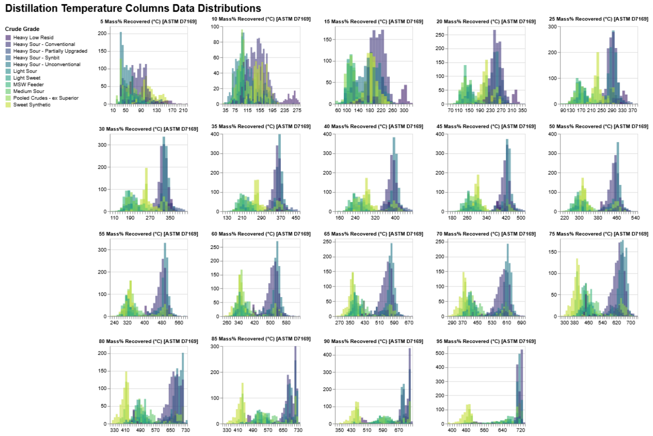
Crude quality distributions



Next, histograms of the distillation temperature columns were created (Figure 8). To avoid excessive visual complexity, only every fifth column was included out of the total 99. The resulting patterns show clear separation among crude grades, reinforcing that the dataset contains identifiable chemical signatures suitable for clustering.

Figure 8

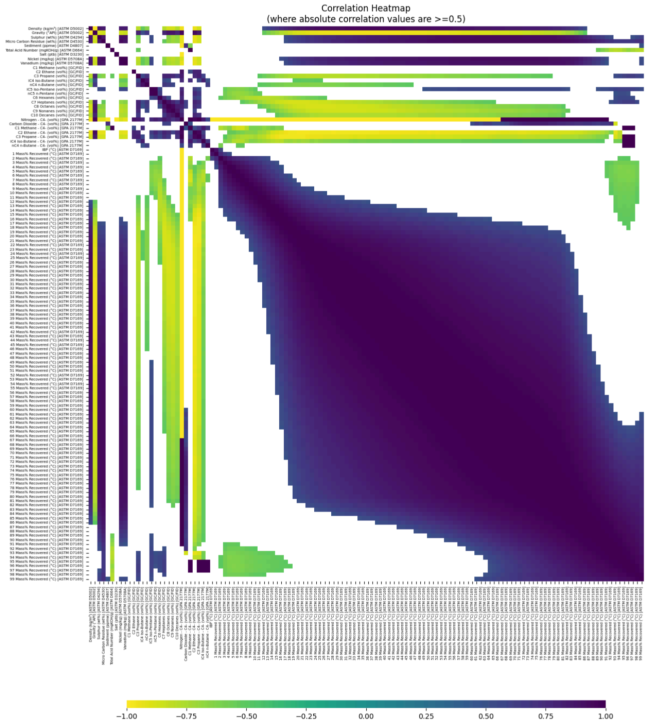
Distillation temperature distributions



Finally, a correlation heatmap was produced (Figure 9), limited to variables with correlation coefficients less than −0.5 or greater than 0.5 to focus on the strongest relationships. Several notable positive and negative correlations were observed among the crude-quality variables, and strong positive correlations were found among the distillation temperature columns which is an expected result since these measurements originate from the same experimental process.

Figure 9

Correlation heatmap



3.4 Pre-Processing

The preprocessing stage prepared the dataset for modeling. After importing the data from CrudeMonitor.ca, the column types were first converted into categorical, numeric, and datetime groups to ensure consistent handling throughout analysis. The data was then split into features (X) and the target (y) variable. The feature matrix X combines the three column groups defined during the EDA stage: a categorical group containing the Sample Date, Location, and Crude columns; a set of numeric columns describing the crude quality characteristics; and a second set of numeric columns representing the distillation temperature measurements. The target labels representing the eleven crude grades were then encoded into integer values using Scikit-Learn’s LabelEncoder to prepare them for modeling and external validation metrics.

A Scikit-Learn ColumnTransformer pipeline was created to standardize and automate the preprocessing workflow. The pipeline includes a FunctionTransformer that removes the columns with excessive missing values identified during EDA. A second FunctionTransformer adds two datetime-derived columns, Month and Year, where the Month value represents the count of months since the earliest sample date in the dataset. This approach ensures that the model treats time as a continuous sequence rather than resetting every calendar year, since machine learning algorithms do not inherently understand that month 1 follows month 12. A KNNImputer using ten nearest neighbors is then applied to estimate missing values in the numeric crude quality and distillation temperature columns. After imputation, a StandardScaler is applied to all numeric columns to normalize their magnitudes and prevent features with large ranges from dominating the clustering process. Finally, a OneHotEncoder is applied to the Location category to convert it into binary variables suitable for modeling.

Once the pipeline was configured, the feature matrix X was processed through the ColumnTransformer to produce a fully preprocessed dataset ready for clustering and dimensionality reduction.

3.5 K-means Clustering

The K-Means clustering algorithm was the first model applied in this project and is one of the most widely used unsupervised learning techniques for identifying patterns within multivariate data. It partitions a dataset into a predefined number of clusters by minimizing the sum of squared distances between data points and their assigned cluster centroids. The algorithm begins by initializing random centroids, then iteratively reassigns data points and recalculates centroids until convergence is achieved. The outcome is a set of compact, well-separated clusters where each data point is closer to its own centroid than to any other. K-Means is computationally efficient and works effectively on large datasets with continuous numeric features, making it suitable for high-dimensional crude assay data. However, it assumes spherical clusters of similar variance, is sensitive to feature scaling and centroid initialization, and can struggle when the true cluster boundaries are not clearly defined. Despite these limitations, its speed, simplicity, and interpretability make it an ideal starting point for exploring structural relationships in crude oil quality data.

For this analysis, K-Means models were fitted for ten cluster sizes ranging from two to eleven. For each configuration, a set of evaluation metrics was calculated, including the Silhouette Score, Davies–Bouldin Index (DBI), Calinski–Harabasz Index (CHI), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes–Mallows Index (FMI). These results are summarized later in the Model Evaluation section. In addition to the quantitative metrics, a multi-panel visualization was created for each cluster size to support visual interpretation. Each visualization includes a silhouette plot to assess intra-cluster cohesion and inter-cluster separation, a PCA-based scatter plot to show linear separability in two dimensions, and a t-SNE projection to reveal nonlinear relationships and confirm how well-separated clusters appear in reduced space.

The strongest performance was observed for cluster sizes of two and three, with silhouette scores of 0.551 and 0.554, respectively. Beyond three clusters, the models showed weaker definition, as both the silhouette values and visual separation declined. The PCA and t-SNE plots for higher cluster counts revealed overlapping groups and fuzzier boundaries, indicating that the data likely forms two or three natural clusters rather than many small ones. This observation supports the project's hypothesis that the existing classification of Canadian crude grades may be overly segmented and that fewer, broader groups could more accurately represent their chemical and physical relationships.

A key learning from this process was the importance of combining quantitative scoring metrics with qualitative visual inspection. While internal validation measures such as the silhouette score provided a numerical assessment of cluster cohesion and separation, the PCA and t-SNE visualizations offered crucial confirmation of which cluster configurations best captured the dataset's structure. Together, these methods provided a balanced understanding of the clustering results and strengthened the overall interpretation of the data.

3.6 Agglomerative Clustering

Agglomerative Clustering was the second model applied in this project and serves as a hierarchical alternative to the K-Means algorithm discussed previously. Unlike K-Means, which partitions data based on centroid distances, Agglomerative Clustering builds nested groupings by progressively merging the most similar data points or clusters. The process begins with each observation as its own cluster and iteratively combines them based on a linkage criterion such as Ward's method, average linkage, or complete linkage, until all points form a single hierarchy. This hierarchical structure can be represented as a dendrogram, which reveals how clusters merge at varying levels of similarity and provides a deeper view of the dataset's internal relationships. Because it does not require a predefined number of clusters, Agglomerative Clustering is particularly useful for exploring hierarchical relationships that may not be captured by fixed partitioning methods.

The same evaluation approach used for the K-Means model was applied here, including testing cluster sizes ranging from two to eleven and generating the same set of internal and external scoring metrics. A multi-panel visualization was also created for each configuration, consisting of a silhouette plot, a PCA projection, and a t-SNE projection. These visualizations allowed for direct comparison with the K-Means results and helped assess both numerical and spatial patterns in cluster separation.

Consistent with the earlier model, the best-performing configurations occurred at cluster sizes of two and three, with silhouette scores of 0.546 and 0.552, respectively. Beyond three clusters, both the silhouette scores and visual distinctions weakened, and the PCA and t-SNE plots showed increasing overlap between groups. This again suggests that the data's natural organization is better captured by a smaller number of broader clusters rather than multiple narrowly defined ones. The similarity in results between the two algorithms reinforces confidence in the presence of two or three meaningful groupings within the crude dataset.

A key learning from implementing this model was recognizing the complementary strengths of hierarchical clustering. While K-Means provides efficient partitioning and clear boundaries, Agglomerative Clustering offers interpretive depth through its hierarchical view of relationships among crude grades. The ability to observe how smaller subgroups merge into larger, more cohesive structures provided a richer understanding of the hierarchical nature of the data. This confirmed that while both methods differ in implementation, they converge on the same structural insight supporting the hypothesis that Canadian crude oils may be more effectively categorized into a reduced set of core groups.

3.7 Gaussian Mixture Model Clustering

The third model applied in this project was the Gaussian Mixture Model (GMM), a probabilistic clustering method that assumes the data are generated from a mixture of multiple Gaussian distributions, each representing a cluster. Unlike K-Means, which assigns each observation to a single cluster, GMM performs soft clustering by estimating the probability that each data point belongs to every cluster. The algorithm uses the Expectation–Maximization (EM) process to iteratively optimize the parameters of the Gaussian components namely the means, covariances, and mixing weights until the likelihood of the observed data under the model is maximized. This probabilistic approach allows GMM to capture more complex cluster geometries, including ellipsoidal and overlapping groups, that simpler models such as K-Means cannot model effectively.

The same modeling and evaluation process described in earlier sections was followed here. GMM was trained for cluster sizes ranging from two to eleven, and the same set of internal and external evaluation metrics was computed for each configuration. A multi-panel visualization consisting of a silhouette plot, PCA projection, and t-SNE projection was also produced for every cluster size, maintaining consistency with the earlier models for comparative analysis. This parallel setup allowed for a direct comparison of the probabilistic model's performance relative to the partition-based and hierarchical methods already discussed.

The two- and three-cluster configurations again produced the strongest performance, with silhouette scores of 0.550 and 0.515, respectively. Visual inspection of the PCA and t-SNE plots confirmed that these cluster sizes generated distinct and interpretable groupings, while larger numbers of clusters resulted in diminishing separation and more overlap. At higher cluster counts, the probabilistic nature of GMM led to visible blending between clusters which is an indication that many samples share transitional characteristics rather than forming sharply bounded groups. This outcome reinforces the conclusions drawn from the K-Means and Agglomerative Clustering results, suggesting that the crude oil dataset is most naturally represented by two or three principal clusters.

The inclusion of GMM in this project provided additional insight into the uncertainty and overlap inherent in crude oil quality data. While the earlier models identified the same optimal range of clusters, GMM added interpretive depth by estimating the degree of membership each sample holds across groups. This proved useful for understanding borderline cases since crude samples share properties of multiple grades.

A key learning from this stage was that probabilistic models like GMM can complement deterministic clustering methods by quantifying ambiguity rather than forcing discrete assignments. Together, these three approaches provide a coherent picture of the dataset's structure and collectively support the hypothesis that a reduced, probabilistically defined classification system may better reflect the true relationships among Canadian crude grades.

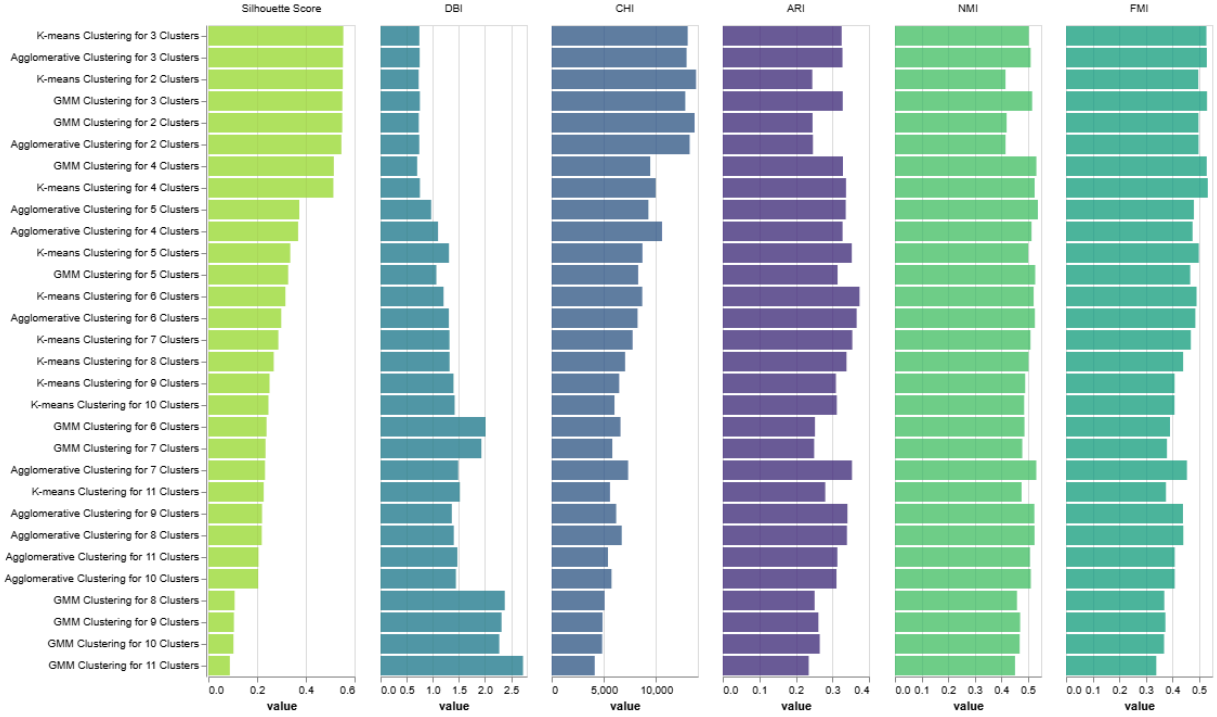
4. MODEL EVALUATION

4.1 Clustering Scoring Metrics

The performance of the clustering models was evaluated using both internal and external validation metrics to provide a comprehensive assessment of clustering quality. Internal metrics assess how well the model organizes data based solely on similarity within the feature space, while external metrics compare those cluster assignments to the known target labels representing the predefined crude oil groups. This dual evaluation approach captures both the intrinsic structure uncovered by the models and the degree to which that structure aligns with the existing classification system.

Three internal metrics were applied to each clustering result. The Silhouette Score measures how well each sample fits within its assigned cluster relative to its separation from other clusters, where higher values indicate stronger cohesion and clearer boundaries. The Davies–Bouldin Index (DBI) represents the average ratio of within-cluster similarity to between-cluster separation, with lower values corresponding to better-defined clusters. The Calinski–Harabasz Index (CHI) measures the ratio of between-cluster dispersion to within-cluster dispersion, where higher values signal more distinct partitioning. These internal metrics collectively describe how compact and well-separated the clusters are within the feature space.

Figure 10
Distillation temperature distributions



Three external metrics were also computed to evaluate how closely the model's clusters aligned with the known crude oil categories. The Adjusted Rand Index (ARI) measures the agreement between predicted clusters and true labels while adjusting for random chance, where higher scores indicate stronger correspondence. The Normalized Mutual Information (NMI) quantifies the shared information between cluster assignments and the ground truth, with values closer to one indicating greater consistency. Finally, the Fowlkes–Mallows Index (FMI) computes the geometric mean of precision and recall, balancing correct assignments against misclassifications. These metrics together gauge how effectively each unsupervised model replicated the underlying patterns reflected in the existing classification system.

During model training, all scoring metrics were stored in a Python dictionary at each iteration and later compiled into the performance visualization shown above (Figure 10). The chart ranks the models by Silhouette Score, enabling quick comparison of performance across algorithms and cluster sizes. Across all three clustering methods of K-Means, Agglomerative, and Gaussian Mixture Models the configurations with two or three clusters consistently achieved the highest scores across nearly every metric. The three-cluster models in particular stood out as the strongest performers, showing the best balance between internal cohesion and external alignment with known labels.

These results provide strong quantitative support for the project's central hypothesis that Canadian crude oils naturally group into two to three major categories rather than the larger number currently used. The agreement between all three clustering algorithms further reinforces this conclusion, as each model type of partition-based, hierarchical, and probabilistic is able to identify a similar underlying structure. Together, the scoring metrics demonstrate that the dataset's intrinsic patterns converge toward a smaller, more representative set of groupings, suggesting that a simplified classification scheme may better reflect the true relationships among crude oil grades.

4.2 Clustering Plots

To visually evaluate clustering performance, a consistent set of three plots was generated for every model and cluster size (Figure 11). Each iteration included a silhouette plot, a PCA projection, and a t-SNE projection, allowing both quantitative and qualitative interpretation of clustering behavior. The silhouette plots show the cohesion and separation of the clusters, while the PCA and t-SNE plots illustrate how these clusters appear in reduced two-dimensional space. When clusters are well-defined, the silhouette coefficients remain positive and consistent in width across groups, while the PCA and t-SNE plots display distinct and clearly separated clusters. Conversely, uneven or negative silhouettes and overlapping projections suggest weaker cluster formation and higher levels of misclassification.

Across all three clustering algorithms of K-Means, Agglomerative, and Gaussian Mixture Models the visual patterns consistently supported the numerical findings discussed earlier. The cluster sizes of two and three produced the most distinct groupings, with minimal overlap between clusters and strong alignment between the silhouette, PCA, and t-SNE visualizations. In contrast, models with more than three clusters exhibited greater overlap in the dimensionality reduction plots, indicating that additional clusters did not correspond to meaningful structural divisions in the data. This trend became especially apparent at cluster size eleven, where the silhouette plots displayed many negative values, signaling a significant number of misclassified or ambiguously assigned data points. These results reinforce the conclusion that increasing the number of clusters beyond three does not improve model performance and, in fact, leads to weaker and less interpretable structures.

To provide a visual reference for these findings, representative plots for cluster sizes two, three, and eleven are included in Appendix A, where Section A.1 presents the K-Means results, Section A.2 presents the Agglomerative Clustering results, and Section A.3 presents the Gaussian Mixture Model results. These plots collectively illustrate the progression from clearly separated to increasingly overlapping clusters as the number of groups increases. Together with the scoring metrics, they offer strong visual evidence in support of the project's hypothesis that the crude oil dataset naturally organizes into two to three major clusters, aligning closely with the structural insights observed throughout the modeling process.

4.3 Classification Confusion Matrix

A second visualization, presented in Figure 12, was created to complement the clustering evaluation by displaying 11 × 11 confusion matrices for each of the three clustering models: K-Means, Agglomerative, and Gaussian Mixture Model. Each heatmap compares the predicted cluster assignments to the known crude oil grade labels, providing a direct visual indication of how each algorithm grouped the data relative to the existing classification system.

Figure 11

Example of the silhouette plot and dimensionally reduced scatter plots

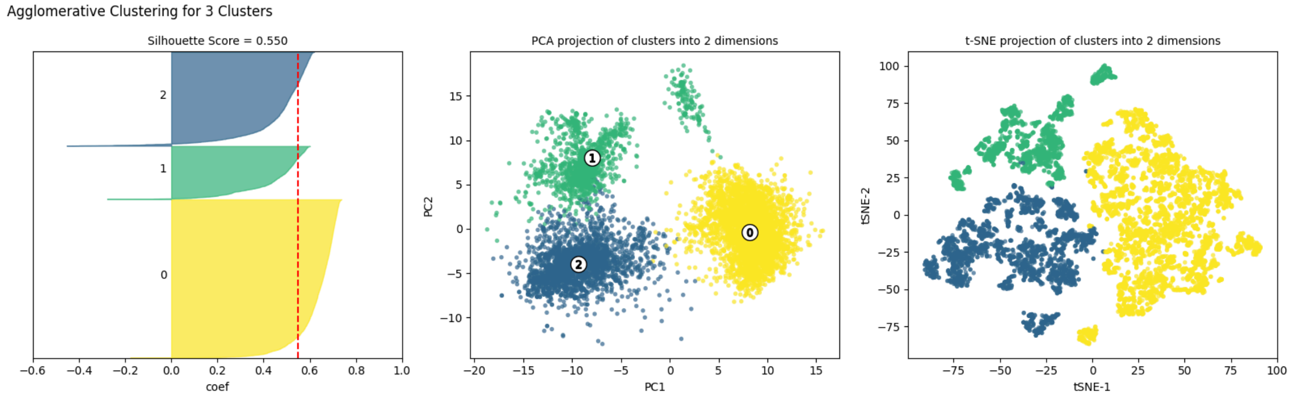
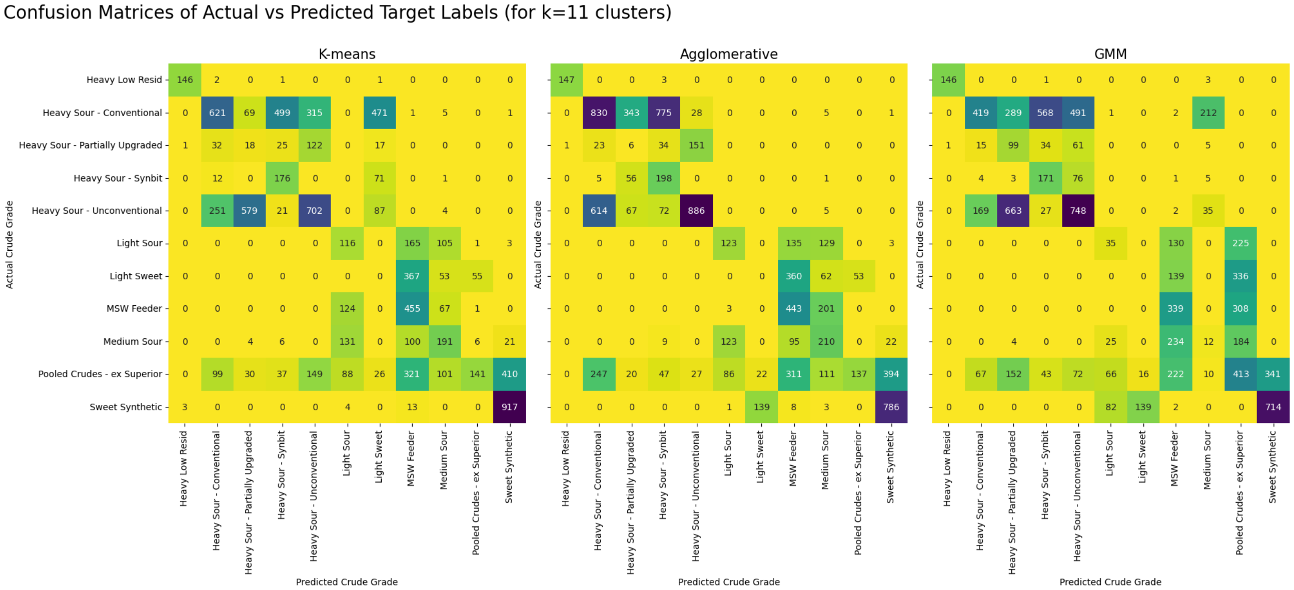


Figure 12

Classification confusion matrices - unsupervised models



Across all three matrices, there is a clear structural pattern suggesting at least two distinct groupings. The upper-left region of each matrix shows dense concentrations corresponding to the heavy oil grades, while the bottom-right region aligns with the lighter and synthetic grades. This visual separation reinforces the earlier quantitative and visual findings from the silhouette, PCA, and t-SNE analyses, further supporting the hypothesis that Canadian crude oils naturally organize into two to three major families.

A particularly interesting observation is the behavior of the Pooled Crudes – ex Superior category. This class is known to represent a blended stream containing multiple crude grades that together comprise the majority of western Canadian production. The confusion matrices confirm this, as every model showed that samples labeled as "Pooled Crudes" were distributed across nearly all other classes. This result validates the model outputs rather than exposing an error and it confirms that the blended nature of that category is being detected through the clustering process.

Although there is considerable misclassification across all models, this outcome is not a weakness but rather an important signal about the underlying structure of the dataset. The overlap and cross-classification observed in the confusion matrices demonstrate that many crude grades share similar chemical and physical characteristics, blurring the traditional boundaries used for classification. Instead of showing failure, these misclassifications reveal continuity within the crude quality spectrum, supporting the interpretation that the current eleven-grade system may over-segment a naturally continuous set of properties. In combination with the scoring metrics and dimensionality reduction plots, these findings strongly suggest that the crude oil dataset is best characterized by a smaller number of broader, compositionally related clusters.

4.4 Comparison to Supervised Learning

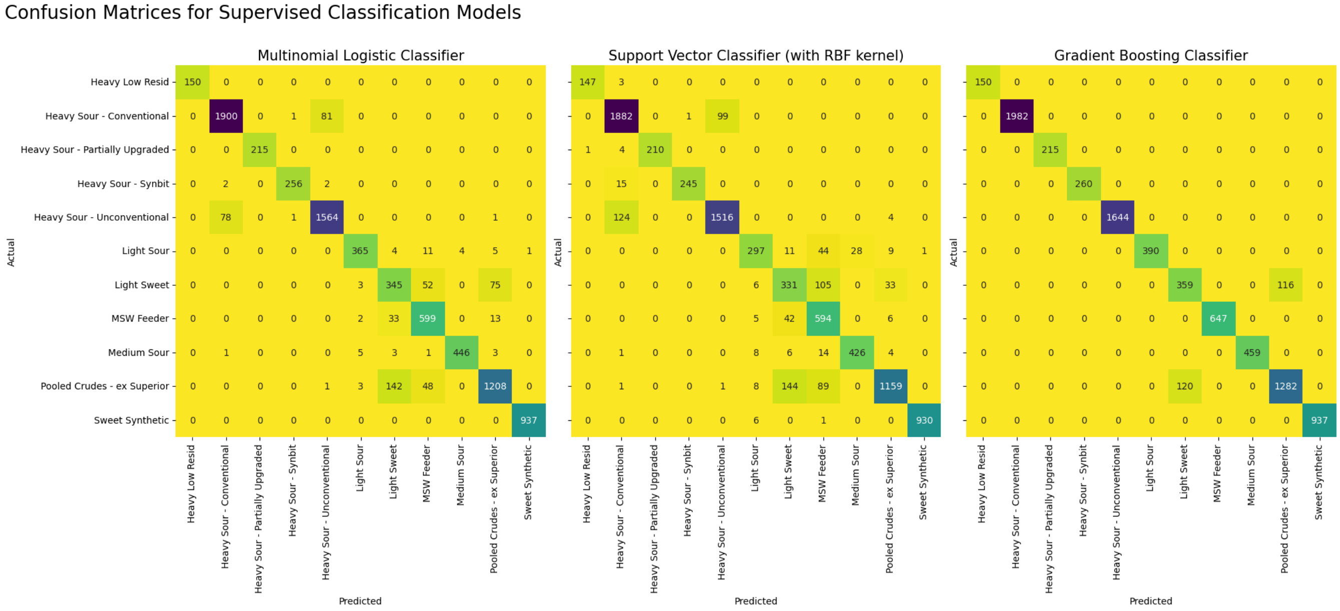
To assess how well the unsupervised clustering models align with label-driven classification methods, their results were compared against three supervised learning models trained on the same feature matrix with true target labels: a Multinomial Logistic Classifier, a Support Vector Classifier (SVC), and a Gradient Boosting Classifier (Figure 13). These algorithms were chosen to represent different modeling paradigms—logistic regression for linear separability, SVC for margin-based decision boundaries, and gradient boosting for nonlinear, ensemble-based learning. After training, each model generated predictions for the crude oil grades, which were then visualized as 11×11 confusion matrices to mirror the unsupervised evaluation and allow for direct visual comparison.

All three supervised models demonstrated strong overall predictive performance, confirming that the dataset contains enough feature differentiation to reliably separate most crude grades when labels are available. However, certain patterns of systematic misclassification appeared across multiple models, revealing meaningful overlaps between specific categories. Both the Logistic Classifier and SVC commonly confused Heavy Sour – Conventional and Heavy Sour – Unconventional, which aligns with expectations since these grades share highly similar physical and chemical characteristics. Likewise, all three models exhibited notable confusion between Light Sweet and Pooled Crudes – ex Superior, further reflecting the compositional overlap of these streams in real-world production.

These recurring misclassification patterns directly mirror the relationships observed in the unsupervised clustering results. The same grade pairings that proved difficult to separate in supervised learning were also found to overlap in the PCA and t-SNE visualizations from the clustering models. This consistency across both learning paradigms strengthens the interpretation that Canadian crude oils naturally fall into two to three broader categories, rather than the eleven discrete grades currently defined. The supervised models, while more precise due to label guidance, ultimately confirm the same structural insight uncovered by the unsupervised methods and that the boundaries between certain crude groups are fluid, reflecting a continuum of properties rather than rigid divisions.

Figure 13

Classification confusion matrices - supervised models



Confusion Matrices for Supervised Classification Models

Multinomial Logistic Classifier

Actual \ Predicted	Heavy Low Resid	Heavy Sour - Conventional	Heavy Sour - Partially Upgraded	Heavy Sour - Synbit	Heavy Sour - Unconventional	Light Sour	Light Sweet	MSW Feeder	Medium Sour	Pooled Crudes - ex Superior	Sweet Synthetic
Heavy Low Resid	150	0	0	0	0	0	0	0	0	0	0
Heavy Sour - Conventional	0	1900	0	1	81	0	0	0	0	0	0
Heavy Sour - Partially Upgraded	0	0	215	0	0	0	0	0	0	0	0
Heavy Sour - Synbit	0	2	0	256	2	0	0	0	0	0	0
Heavy Sour - Unconventional	0	78	0	1	1564	0	0	0	0	1	0
Light Sour	0	0	0	0	0	365	4	11	4	5	1
Light Sweet	0	0	0	0	0	3	345	52	0	75	0
MSW Feeder	0	0	0	0	0	2	33	599	0	13	0
Medium Sour	0	1	0	0	0	5	3	1	446	3	0
Pooled Crudes - ex Superior	0	0	0	0	1	3	142	48	0	1208	0
Sweet Synthetic	0	0	0	0	0	0	0	0	0	0	937

Support Vector Classifier (with RBF kernel)

Actual \ Predicted	Heavy Low Resid	Heavy Sour - Conventional	Heavy Sour - Partially Upgraded	Heavy Sour - Synbit	Heavy Sour - Unconventional	Light Sour	Light Sweet	MSW Feeder	Medium Sour	Pooled Crudes - ex Superior	Sweet Synthetic
Heavy Low Resid	147	3	0	0	0	0	0	0	0	0	0
Heavy Sour - Conventional	0	1882	0	1	99	0	0	0	0	0	0
Heavy Sour - Partially Upgraded	1	4	210	0	0	0	0	0	0	0	0
Heavy Sour - Synbit	0	15	0	245	0	0	0	0	0	0	0
Heavy Sour - Unconventional	0	124	0	0	1516	0	0	0	0	4	0
Light Sour	0	0	0	0	0	297	11	44	28	9	1
Light Sweet	0	0	0	0	0	6	331	105	0	33	0
MSW Feeder	0	0	0	0	0	5	42	594	0	6	0
Medium Sour	0	1	0	0	0	8	6	14	426	4	0
Pooled Crudes - ex Superior	0	1	0	0	1	8	144	89	0	1159	0
Sweet Synthetic	0	0	0	0	0	6	0	1	0	0	930

Gradient Boosting Classifier

Actual \ Predicted	Heavy Low Resid	Heavy Sour - Conventional	Heavy Sour - Partially Upgraded	Heavy Sour - Synbit	Heavy Sour - Unconventional	Light Sour	Light Sweet	MSW Feeder	Medium Sour	Pooled Crudes - ex Superior	Sweet Synthetic
Heavy Low Resid	150	0	0	0	0	0	0	0	0	0	0
Heavy Sour - Conventional	0	1982	0	0	0	0	0	0	0	0	0
Heavy Sour - Partially Upgraded	0	0	215	0	0	0	0	0	0	0	0
Heavy Sour - Synbit	0	0	0	260	0	0	0	0	0	0	0
Heavy Sour - Unconventional	0	0	0	0	1644	0	0	0	0	0	0
Light Sour	0	0	0	0	0	390	0	0	0	0	0
Light Sweet	0	0	0	0	0	0	359	0	0	116	0
MSW Feeder	0	0	0	0	0	0	0	647	0	0	0
Medium Sour	0	0	0	0	0	0	0	0	459	0	0
Pooled Crudes - ex Superior	0	0	0	0	0	0	120	0	0	1282	0
Sweet Synthetic	0	0	0	0	0	0	0	0	0	0	937

5. DISCUSSION

This project set out to determine whether the existing classification of Canadian crude oil grades, which are currently divided into eleven distinct categories, accurately reflects the natural structure of their physical and chemical characteristics. Using a combination of unsupervised and supervised machine learning techniques, the analysis systematically explored whether these grades could instead be represented by a smaller number of meaningful groups. The work encompassed the full analytical pipeline, beginning with data acquisition and preprocessing, continuing through clustering model development and validation, and concluding with comparative evaluation and interpretation.

The dataset, sourced from CrudeMonitor.ca, contained detailed assay measurements including crude quality indicators and distillation curve data. After extensive cleaning and feature engineering, non-representative or incomplete variables were removed to preserve data integrity. A fully automated preprocessing pipeline was implemented using Scikit-Learn's ColumnTransformer, integrating time-based feature generation, mean and KNN imputation, feature scaling, and categorical encoding. This ensured that each clustering model received a consistent, standardized input matrix. Exploratory analysis confirmed that the dataset primarily represented production from Alberta's major hubs of Edmonton, Hardisty, and Superior providing both geographic consistency and realistic coverage of western Canadian crude streams.

Three clustering models were applied: K-Means, Agglomerative (Hierarchical) Clustering, and the Gaussian Mixture Model (GMM). Each algorithm was evaluated across cluster sizes from 2 to 11, and for each iteration, a comprehensive suite of internal (Silhouette, Davies–Bouldin, Calinski–Harabasz) and external (ARI, NMI, FMI) validation metrics was computed. Multi-panel plots combining silhouette visualizations, PCA projections, and t-SNE projections were generated to interpret the structure of each model visually. Across all methods, the results consistently showed that cluster sizes of two and three produced the highest silhouette scores and the most distinct separations, while higher cluster counts led to overlapping boundaries and degraded cohesion. The GMM and Agglomerative models particularly emphasized this trend, with strong separation at cluster sizes of two to three, and rapidly diminishing clarity beyond that range. These findings provide robust quantitative and visual evidence that the natural structure of the crude oil dataset aligns more closely with two to three major groups rather than the eleven predefined grades.

To complement the clustering analysis, the supervised models of Multinomial Logistic Regression, Support Vector Classifier, and Gradient Boosting were trained on the same dataset to establish upper-bound classification performance. While these models achieved high accuracy, the confusion matrices revealed systematic misclassifications between categories that share similar chemical profiles, such as Heavy Sour – Conventional vs. Heavy Sour – Unconventional and Light Sweet vs. Pooled Crudes. These same patterns were reflected in the unsupervised results, confirming that even with label guidance, the boundaries between certain crude types remain ambiguous. Collectively, both learning paradigms support the conclusion that crude oil grades exist along a continuous compositional spectrum rather than as sharply separated classes.

The findings have important implications for both industrial practice and analytical modeling. From a technical perspective, this work demonstrates that unsupervised learning can replicate much of the structure captured by supervised models, even without label information which suggests that the current classification system may impose artificial divisions. From an industry standpoint, these results could motivate further exploration into whether a simplified grading scheme based on physical similarity (e.g., Heavy, Medium, and Light/Synthetic categories) might better represent real-world production and refining behavior.

There are several promising directions for future work. First, incorporating additional datasets (such as crude price differentials, carbon intensity data, or refinery yield information) could reveal whether economic and environmental factors follow the same clustering structure observed in the quality data. Second, applying advanced dimensionality reduction techniques like UMAP or autoencoders could improve the visualization and separation of nonlinear relationships. Third, expanding the analysis to dynamic or temporal clustering might identify how crude quality distributions shift over time, particularly as new upgrading and blending processes are introduced. Finally, integrating deep learning or graph-based clustering methods could capture more complex interdependencies between assay attributes, offering even deeper insight into how Canadian crude oils are related.

In summary, this project successfully demonstrated that unsupervised clustering can reveal a more parsimonious and physically meaningful grouping of Canadian crude oil grades. By combining technical rigor with domain knowledge, the analysis confirms that the natural relationships among these crudes likely fall into two to three primary categories, setting a foundation for ongoing research into data-driven crude classification and broader applications in the energy sector.

6. CONCLUSION

This project used unsupervised machine learning to test whether Canada's eleven crude oil grades accurately reflect their chemical and physical similarities. The three clustering models K-Means, Agglomerative Clustering, and Gaussian Mixture Models (GMM) were applied to refined data from CrudeMonitor.ca, processed through a fully automated Scikit-Learn pipeline for cleaning, scaling, and encoding.

Across all models, two to three clusters consistently achieved the highest silhouette scores and clearest visual separations, while higher cluster counts produced increasing overlap. The Agglomerative and GMM models in particular revealed strong grouping between heavy sour and light or synthetic crudes, confirming two dominant families within the dataset. The Pooled Crudes category, distributed across nearly all clusters, correctly reflected its blended nature.

Results from supervised models of Logistic Regression, Support Vector Classifier, and Gradient Boosting all showed the same recurring misclassifications between similar grades, reinforcing the conclusion that crude oils form a continuous spectrum rather than distinct categories.

Overall, the analysis indicates that Canadian crude oils naturally organize into two to three primary groups, suggesting that the existing eleven-grade classification may be simplified without loss of accuracy. Future work could integrate economic, carbon, or yield data and explore advanced nonlinear or graph-based methods to capture deeper structural relationships within Canada's crude oil system.

7. REFERENCES

[1] Sancho, C., López, M., & Acha, E. (2021). *Clustering of crude oils using key quality parameters*. Energy & Fuels, 35(7), 5671–5678.

[2] Al-Dahhan, M., & Mahmood, A. (2019). *Characterization of Basra crude oils using Watson and correlation indices*. Journal of Petroleum Research and Studies, 28(1), 1–10.

[3] Mohammadi, A., Kamal, M. R., & Shahraki, F. (2020). *Rapid classification of crude oils using FTIR and chemometric methods*. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 239, 118488.

[4] Moro, M., dos Santos, L. M., & Sousa, J. (2021). *Prediction of crude oil properties using NIR and machine learning models*. Fuel, 289, 119918.

[5] Olea, R. A., Houseknecht, D. W., & Garrity, C. P. (2021). *Cluster analysis of crude oil geochemical data for source characterization*. Marine and Petroleum Geology, 133, 105292.

[6] Crude Quality Inc. (n.d.). CrudeMonitor.ca. Retrieved October 19, 2025, from https://www.crudemonitor.ca/

[7] OpenAI. (2025). ChatGPT (Oct 15, 2025 version)

Note: Portions of this report were edited with ChatGPT to improve clarity and grammar.

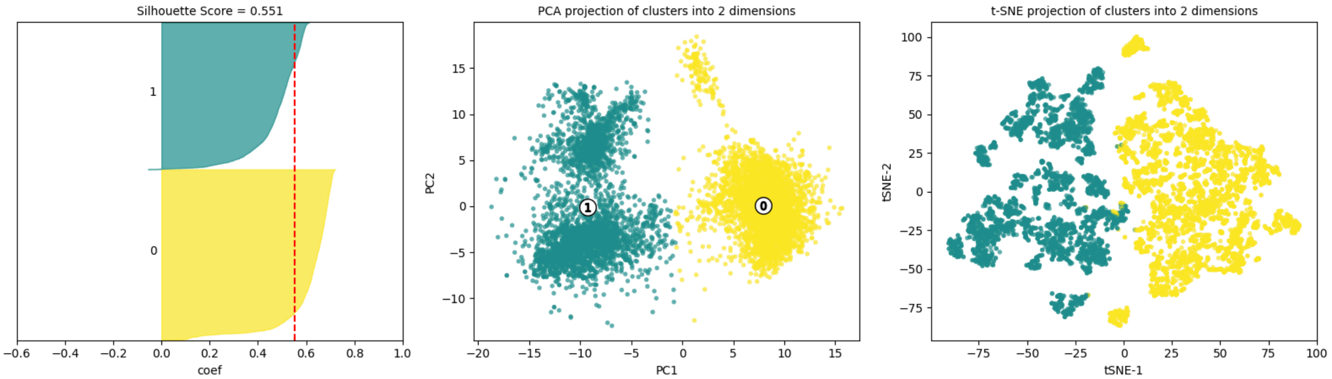
APPENDIX

A. CLUSTERING PLOTS

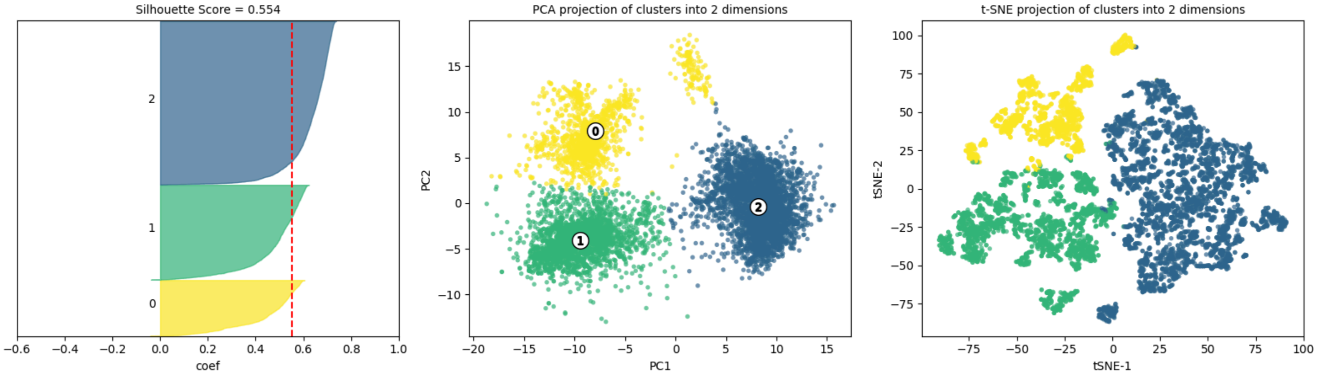
Shown below are the silhouette plots alongside the dimensionally reduced PCA and t-SNE visualizations for each clustering model at cluster sizes of 2, 3, and 11. These sizes were selected because all models achieved their highest internal validation scores at 2 and 3 clusters, while 11 corresponds to the number of original crude oil groups used for comparison.

A.1 K-means Clustering

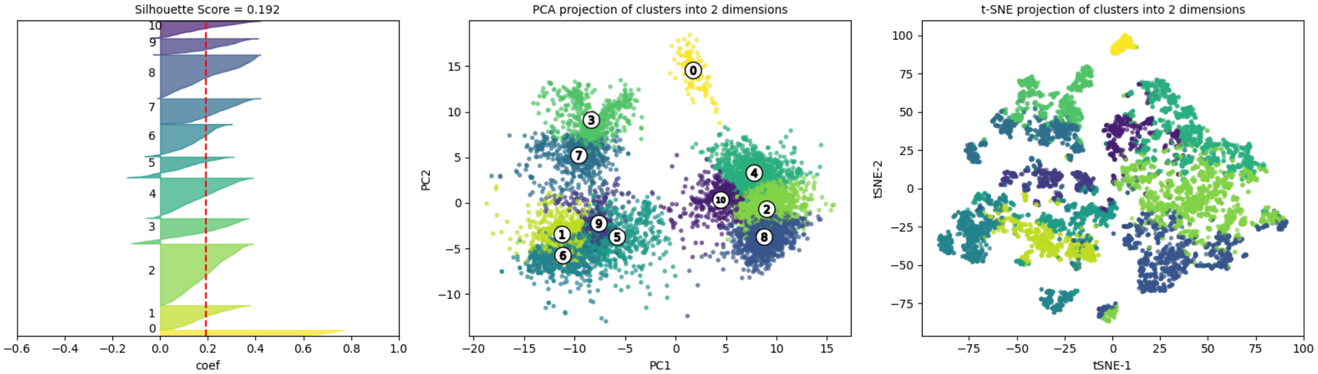
K-means Clustering for 2 Clusters



K-means Clustering for 3 Clusters

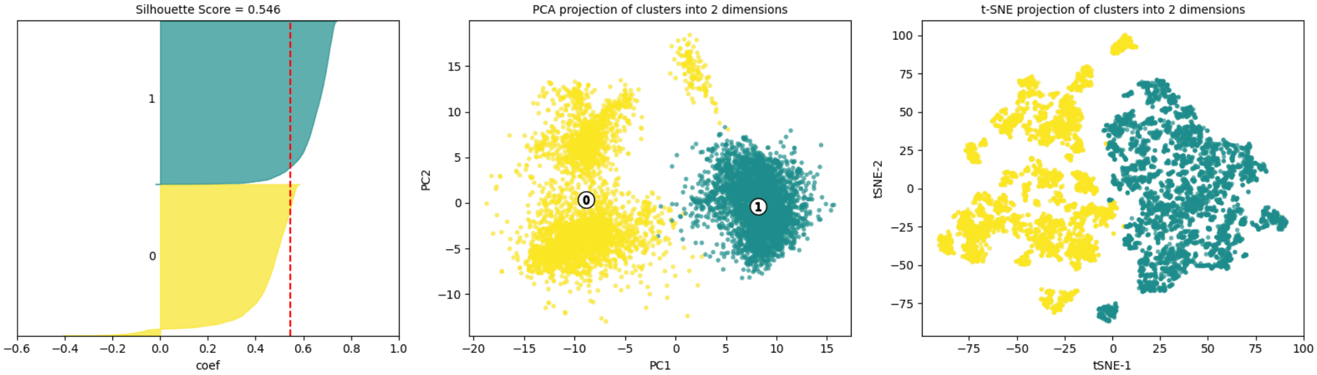


K-means Clustering for 11 Clusters

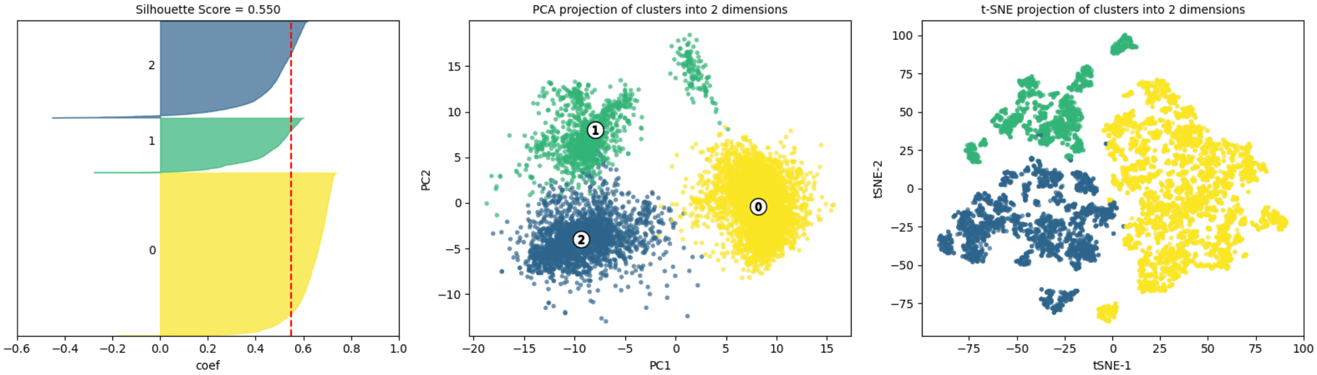


A.2 Agglomerative Clustering

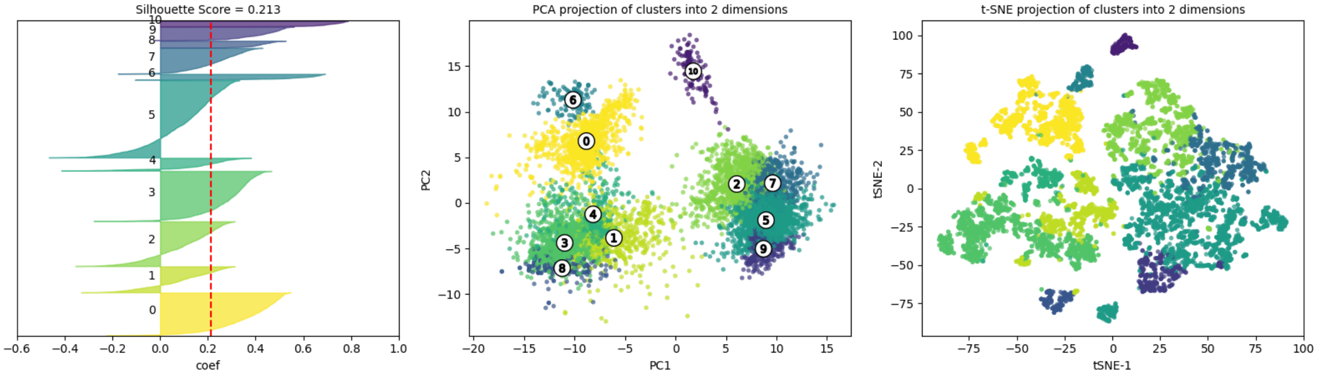
Agglomerative Clustering for 2 Clusters



Agglomerative Clustering for 3 Clusters

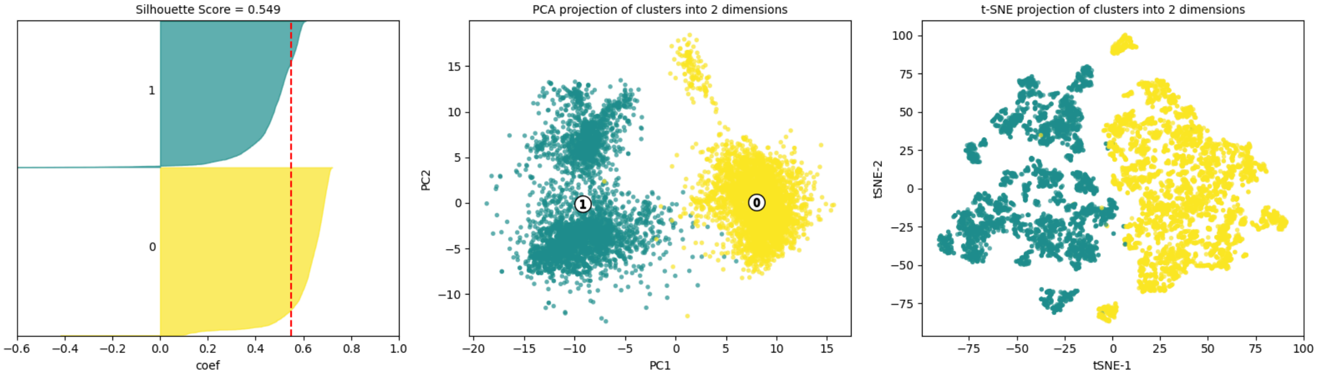


Agglomerative Clustering for 11 Clusters

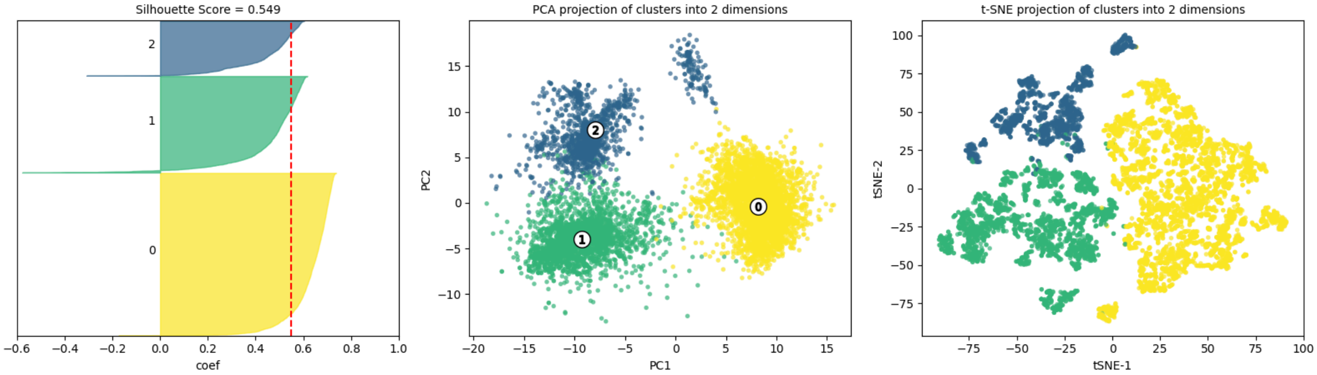


A.3 Gaussian Mixture Model (GMM) Clustering

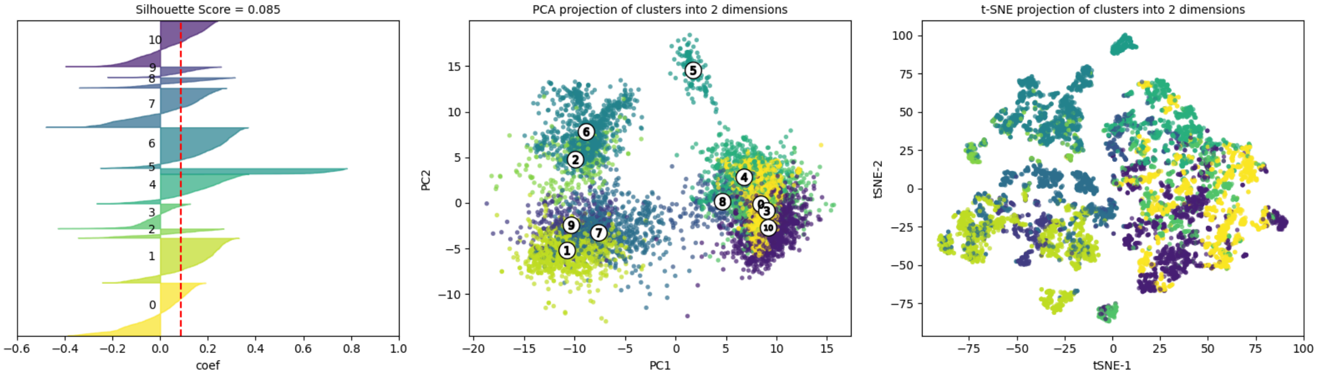
GMM Clustering for 2 Clusters



GMM Clustering for 3 Clusters



GMM Clustering for 11 Clusters


Silhouette Score = 0.085
coef
PCA projection of clusters into 2 dimensions
PC2
PC1
t-SNE projection of clusters into 2 dimensions
tSNE-2
tSNE-1