Project Update

Data-Driven Classification of Canadian Crude Oils

Ryan Bulger
DTSA-5506
University of Colorado Boulder

Problem Definition

Assessing Current Crude Oil Classifications

Grouping Limitations

Existing classifications of crude oils may not accurately reflect the **true similarities** among different types, leading to potential inefficiencies in management and utilization of these resources.

Importance of Reclassifying Crude Oils

Marketing

Improved classifications can enhance **targeted marketing strategies** effectively.

Pricing

Accurate groupings lead to **fair pricing structures** in the market.

Efficiency

Optimized classifications promote **efficient pipeline utilization** and resource management.

Related Work

Overview of traditional and modern methods

Traditional Approaches

Traditional methods like API gravity and sulfur content often **fail to capture the complexities** of crude oil. These simplistic classifications may overlook significant variations in crude oil characteristics.

Modern ML Studies

Recent machine learning approaches, such as **FTIR spectroscopy combined with clustering techniques**, have shown promise. However, they often rely on limited datasets, leading to potential inaccuracies in classifications.

Project Steps

Data Collection

Collect data from CrudeMonitor for robust analysis.

EDA

Perform exploratory data analysis to understand trends.

Data Cleaning

Prepare the dataset for modeling.

Clustering

Apply clustering algorithms to identify data-driven groups.

Supervised Comparison

Compare clustering results with supervised methods for validation.

COMPLETED

Data Collection

CrudeMonitor API

Data was captured using CrudeMonitor's API endpoints.

Dataset Shape

The dataset contains over **8,500 rows and 133 columns**

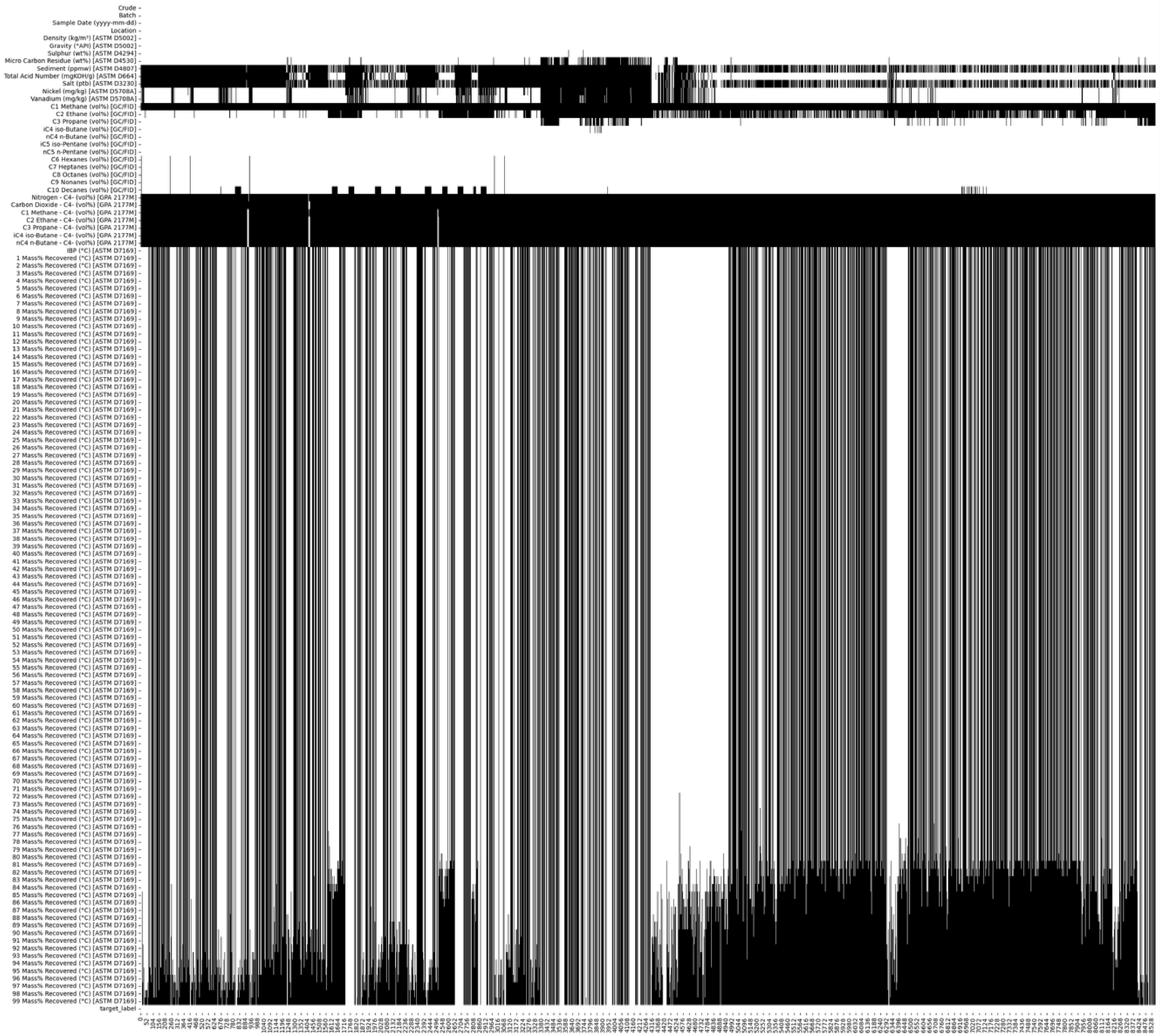
Data Types

The datatypes are 1 datetime, 4 categorical, and 128 numeric columns.

Missing Values

Heatmap

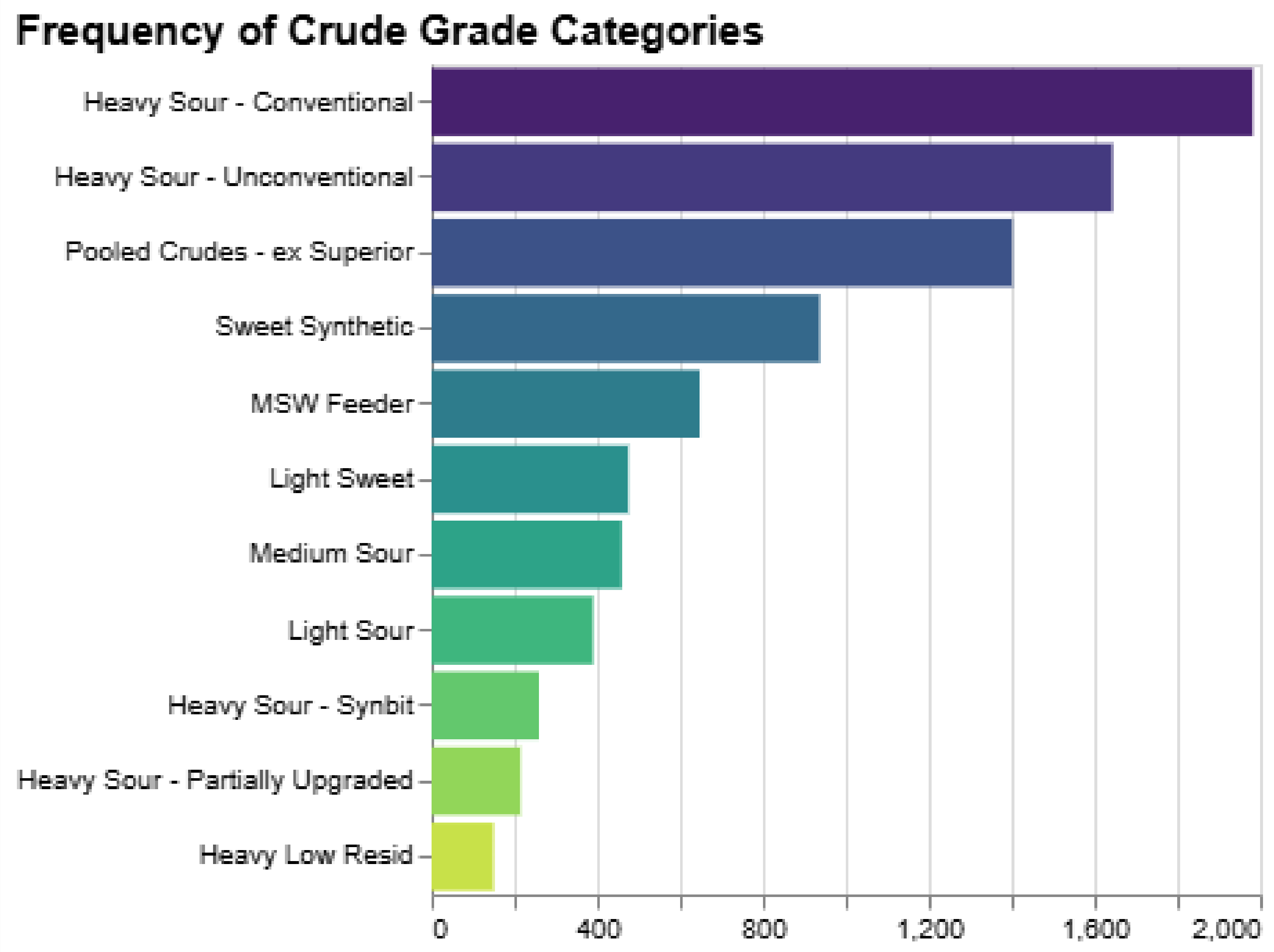
The heatmap color reveals **large swaths of missing data** throughout the dataset



Categorical Frequencies

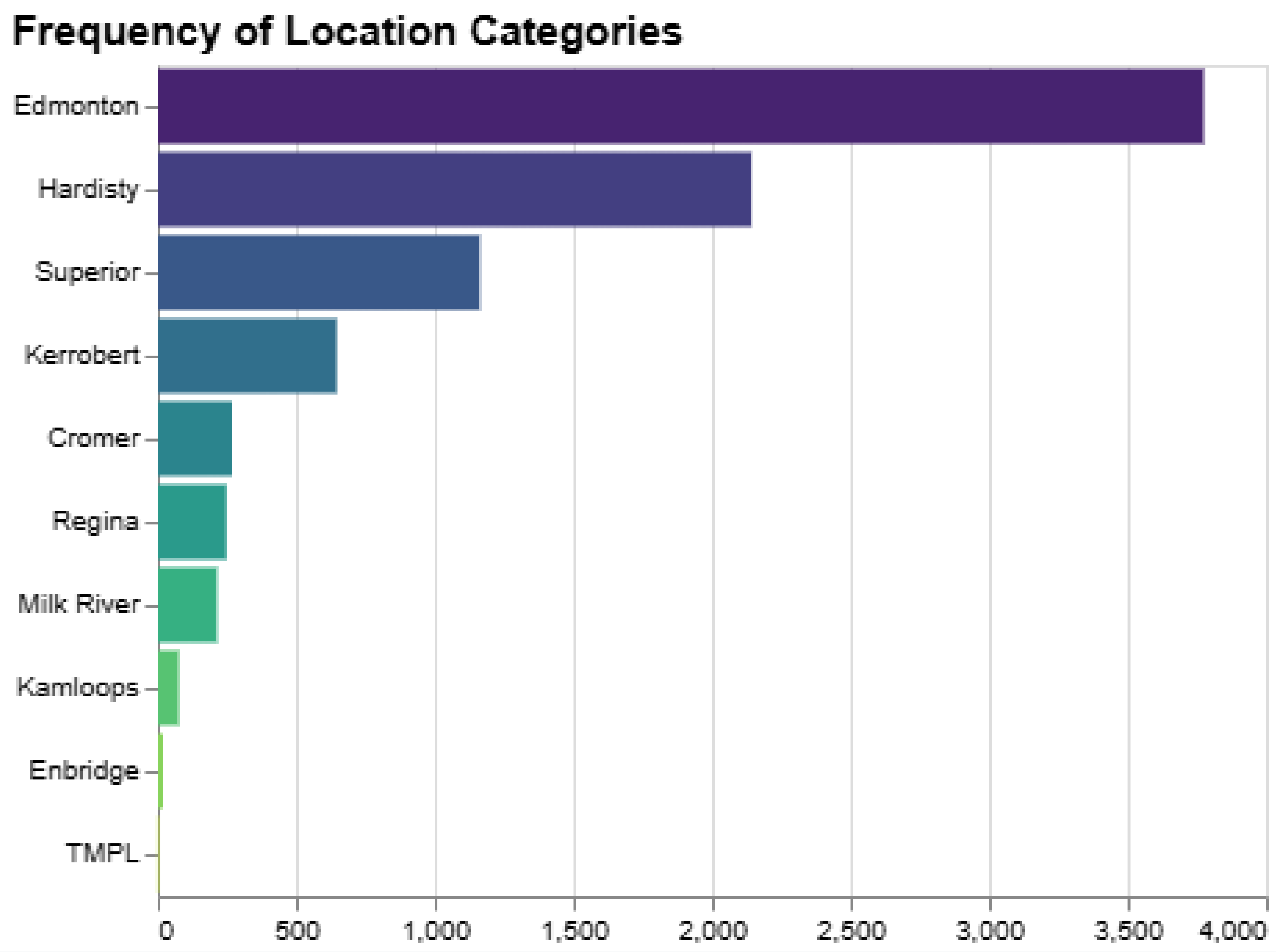
Crude Grades

Majority of the dataset's rows are related to the production of **heavy oils and their by-products**



Sample Locations

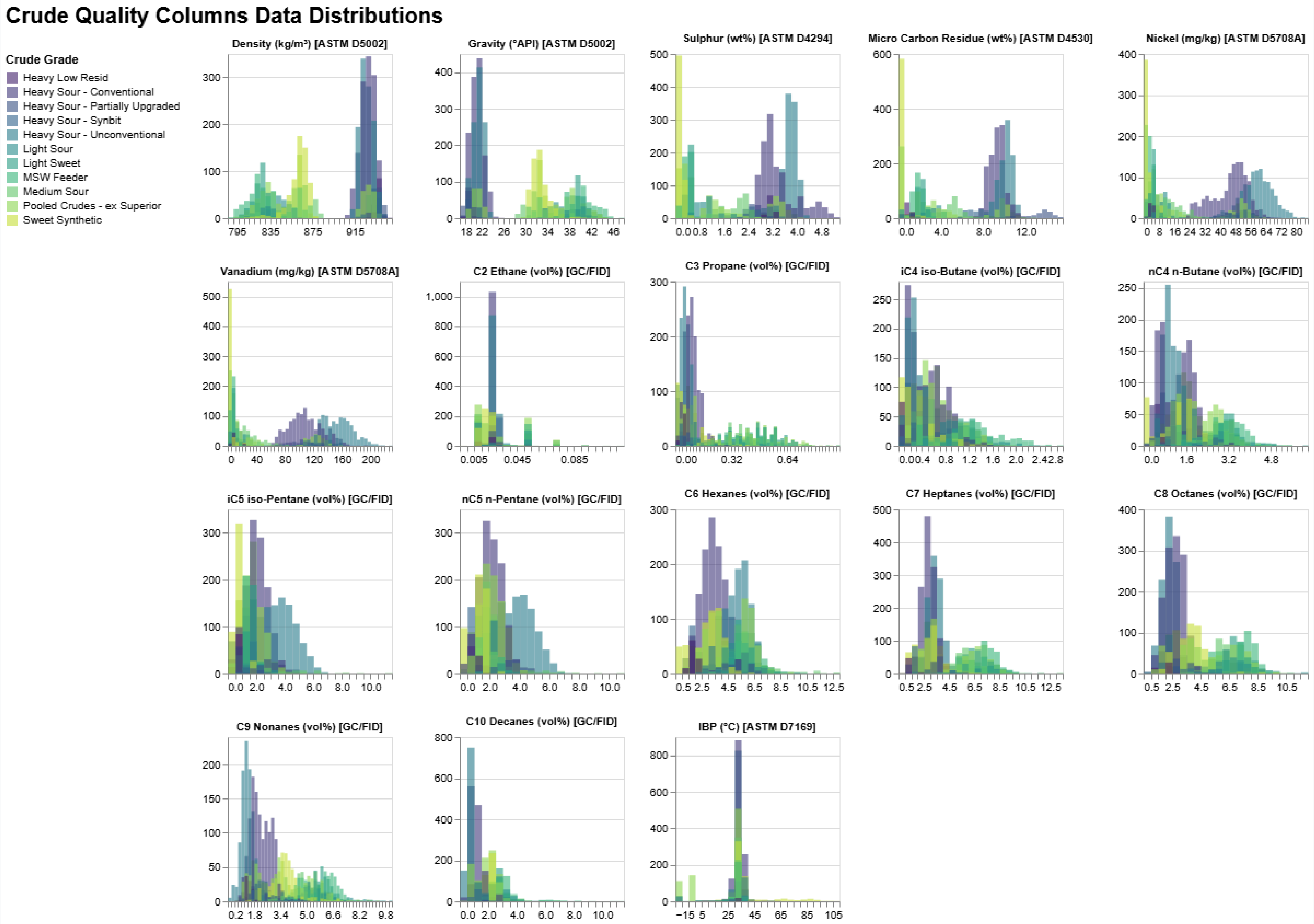
Oil production typically **flows from northern Alberta**, through Edmonton, and onto Hardisty



Numeric Distributions

Crude Quality Columns

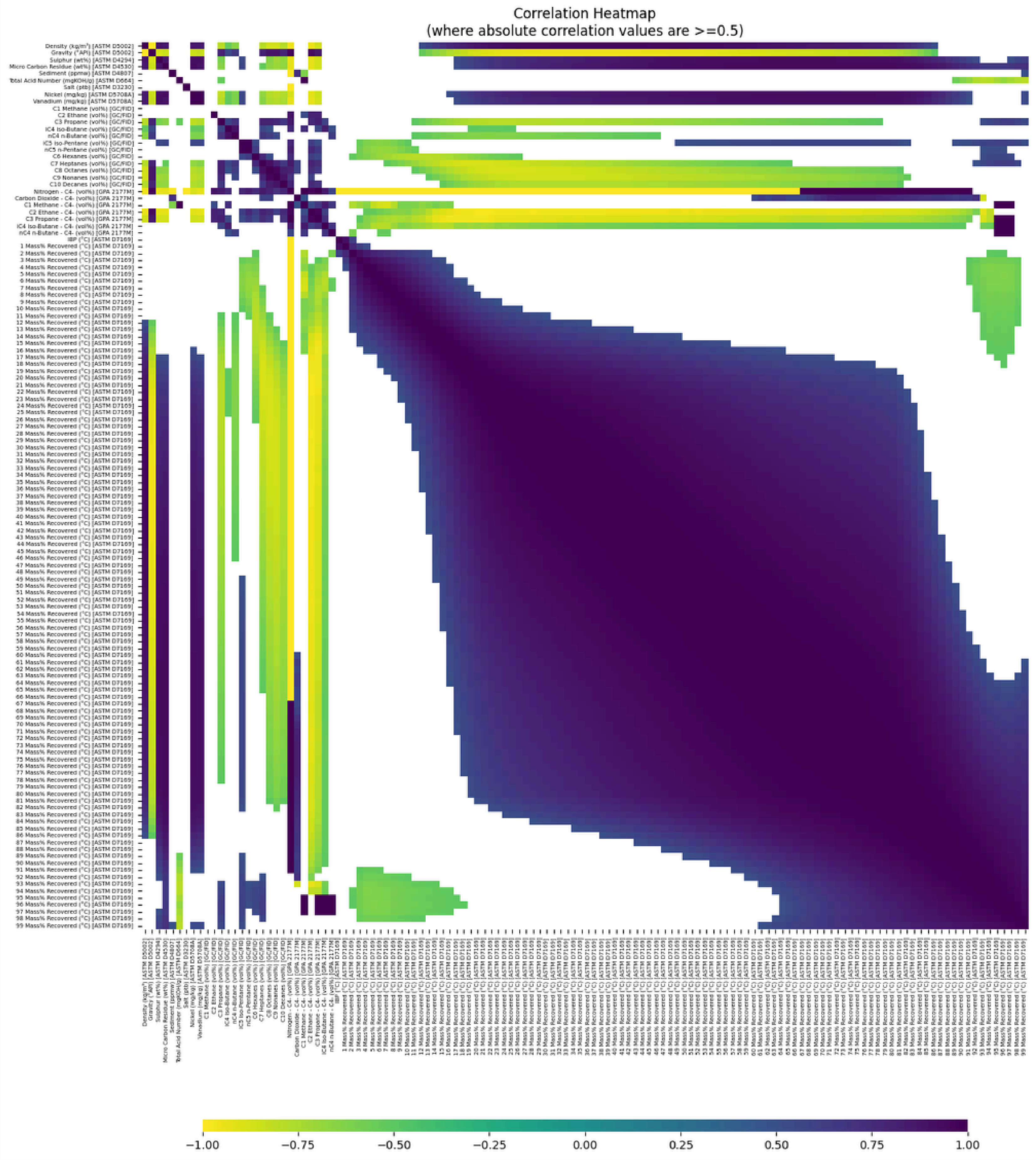
Noticeable patterns in the distributions suggest that **many groups of crudes share similar properties**



Numeric Correlations

Correlation Heatmap

Strong correlations are found throughout the crude quality and distillation temperature columns



Clustering Algorithms

K-Means

K-Means is a **baseline clustering** method that partitions data into K distinct clusters by minimizing intra-cluster variance, allowing for efficient and straightforward classification of crude oil samples.

Agglomerative

Agglomerative clustering is a **hierarchical approach** that builds clusters by merging similar data points iteratively, resulting in a tree-like structure that captures the relationships among crude oils.

GMM

Gaussian Mixture Models (GMM) utilize a **probabilistic framework** to model data as a mixture of multiple Gaussian distributions, providing flexibility and capturing complex data distributions in crude oil classification.

Model Evaluation

Internal Metrics

Internal metrics of **Silhouette Score, DBI, and CHI** are used to assess clustering quality using statistical measures

External Metrics

External metrics of **ARI, NMI, and FMI** are used to compare cluster assignments to known labels

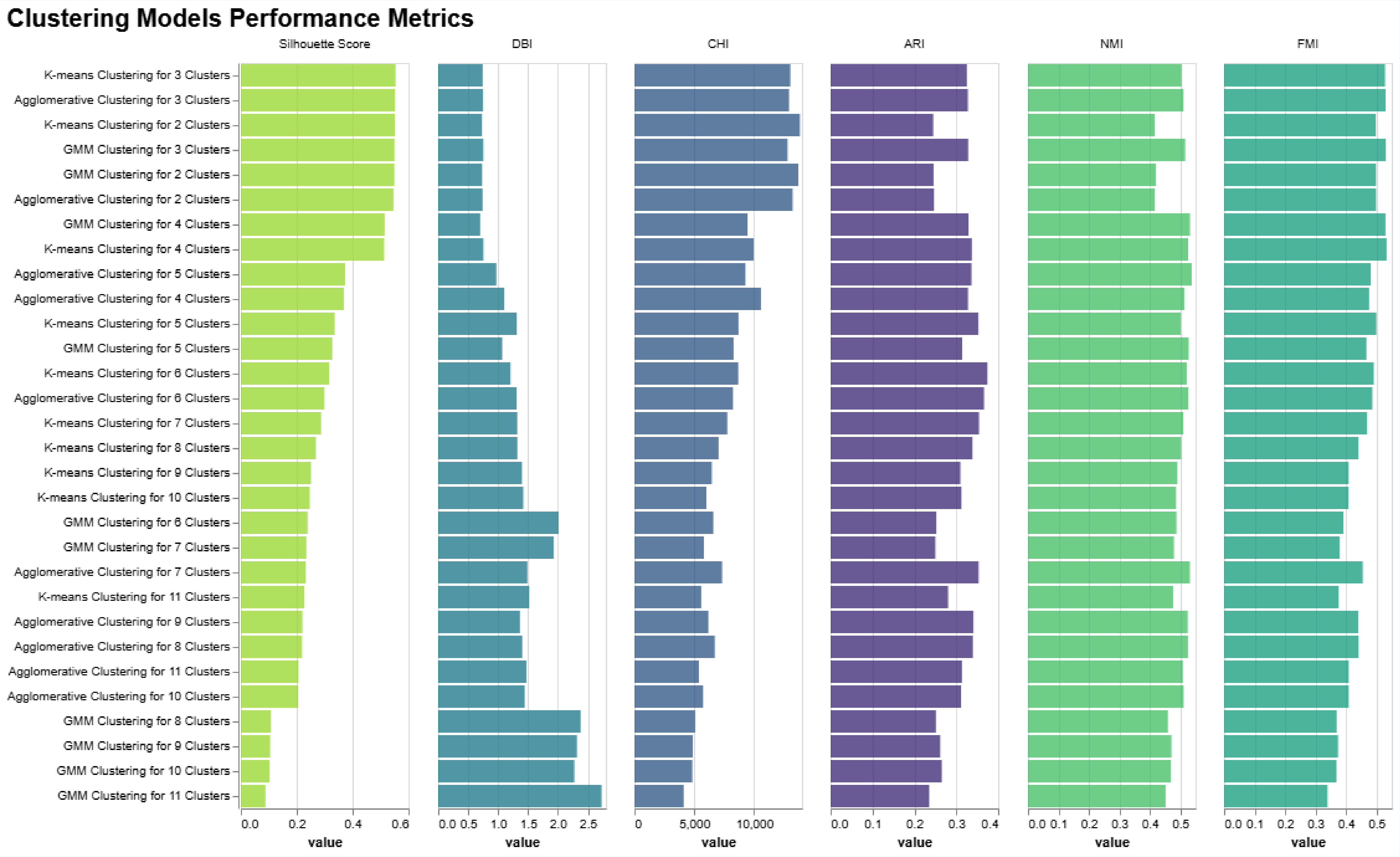
Visualization Plots

Visualization plots are used to illustrate clustering results and distribution patterns

Evaluation Metrics

Best Performer

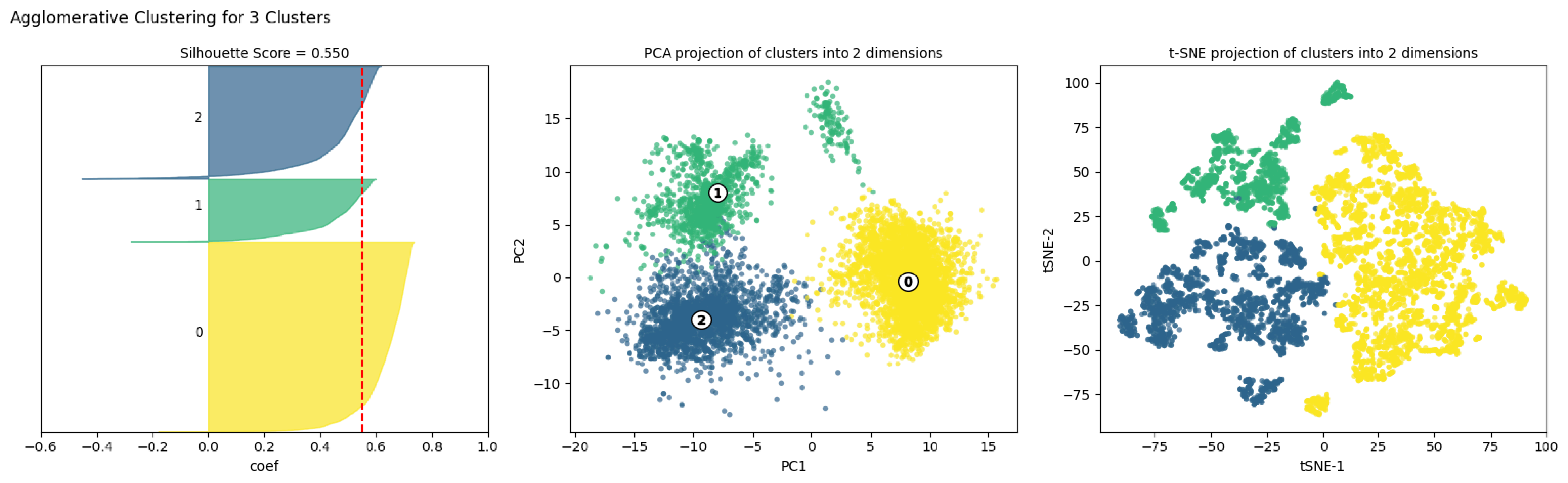
Models with **two to three clusters** showed the strongest performance, with Agglomerative Clustering and GMM performing best at three clusters


Clustering Models Performance Metrics

Evaluation Plots

Dimensionally Reduced Clustering

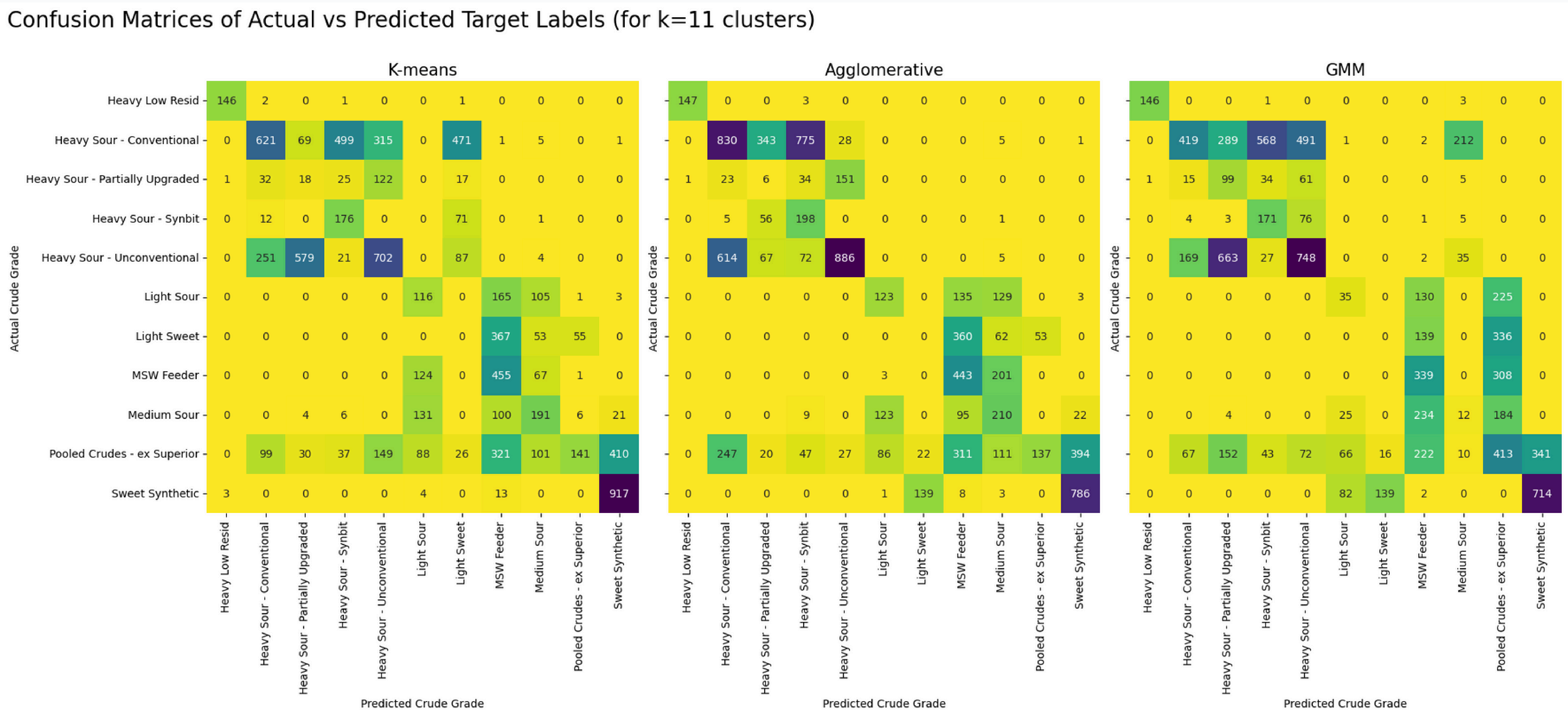
The silhouette and scatter plots clearly show that **three clusters** provide the most distinct and well-separated groupings



Unsupervised Confusion Matrix

Significant Misclassifications

Distinct patterns suggest **natural groupings**



Confusion Matrices of Actual vs Predicted Target Labels (for k=11 clusters)

K-means

	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	146	2	0	1	0	0	1	0	0	0	0
Heavy Sour - Conventional	0	621	69	499	315	0	471	1	5	0	1
Heavy Sour - Partially Upgraded	1	32	18	25	122	0	17	0	0	0	0
Heavy Sour - Synbit	0	12	0	176	0	0	71	0	1	0	0
Heavy Sour - Unconventional	0	251	579	21	702	0	87	0	4	0	0
Light Sour	0	0	0	0	0	116	0	165	105	1	3
Light Sweet	0	0	0	0	0	0	0	367	53	55	0
MSW Feeder	0	0	0	0	0	124	0	455	67	1	0
Medium Sour	0	0	4	6	0	131	0	100	191	6	21
Pooled Crudes - ex Superior	0	99	30	37	149	88	26	321	101	141	410
Sweet Synthetic	3	0	0	0	0	4	0	13	0	0	917

Agglomerative

	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	0	0	3	0	0	0	0	0	0	0
Heavy Sour - Conventional	0	830	343	775	28	0	0	0	5	0	1
Heavy Sour - Partially Upgraded	1	23	6	34	151	0	0	0	0	0	0
Heavy Sour - Synbit	0	5	56	198	0	0	0	0	1	0	0
Heavy Sour - Unconventional	0	614	67	72	886	0	0	0	5	0	0
Light Sour	0	0	0	0	0	123	0	135	129	0	3
Light Sweet	0	0	0	0	0	0	0	360	62	53	0
MSW Feeder	0	0	0	0	0	3	0	443	201	0	0
Medium Sour	0	0	0	9	0	123	0	95	210	0	22
Pooled Crudes - ex Superior	0	247	20	47	27	86	22	311	111	137	394
Sweet Synthetic	0	0	0	0	0	1	139	8	3	0	786

GMM

	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	146	0	0	1	0	0	0	0	3	0	0
Heavy Sour - Conventional	0	419	289	568	491	1	0	2	212	0	0
Heavy Sour - Partially Upgraded	1	15	99	34	61	0	0	0	5	0	0
Heavy Sour - Synbit	0	4	3	171	76	0	0	1	5	0	0
Heavy Sour - Unconventional	0	169	663	27	748	0	0	2	35	0	0
Light Sour	0	0	0	0	0	35	0	130	0	225	0
Light Sweet	0	0	0	0	0	0	0	139	0	336	0
MSW Feeder	0	0	0	0	0	0	0	339	0	308	0
Medium Sour	0	0	4	0	0	25	0	234	12	184	0
Pooled Crudes - ex Superior	0	67	152	43	72	66	16	222	10	413	341
Sweet Synthetic	0	0	0	0	0	82	139	2	0	0	714

Supervised Benchmarking

Comparing key machine learning methods

Logistic Regression

Logistic Regression serves as a fundamental method, modeling the **probability of class membership** using a linear approach, effective for binary classification problems in crude oil analysis.

Support Vector Classifier

SVC utilizes hyperplane separation to classify data points, **excelling in high-dimensional spaces**, providing robust classification through kernel trick adaptations tailored for complex crude oil datasets.

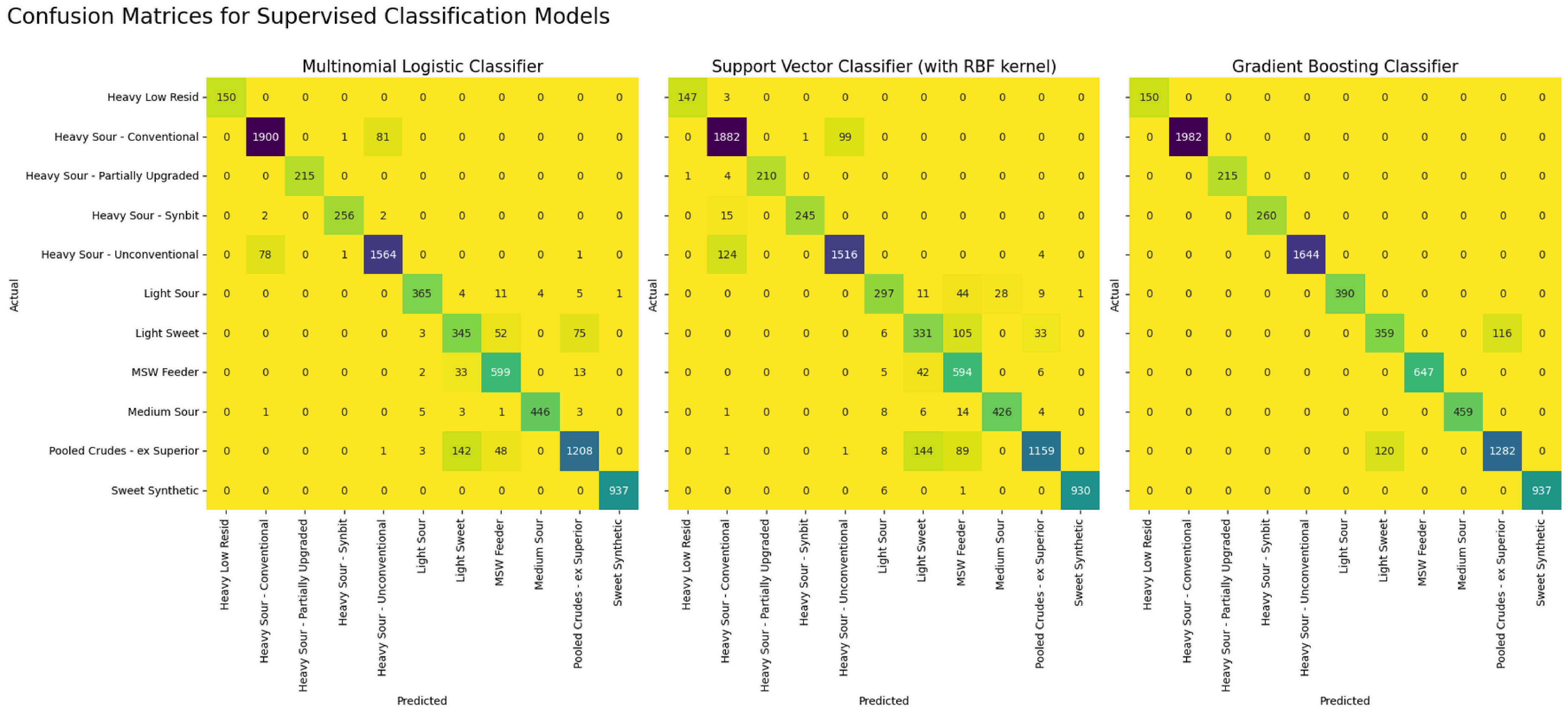
Gradient Boosting Classifier

Gradient Boosting Classifier builds an ensemble of decision trees where each model corrects the errors of the previous one, producing an **accurate and flexible classifier** that captures nonlinear relationships in complex crude oil data.

Supervised Confusion Matrix

Better Performance then Unsupervised

Persistent misclassification further suggest **common groupings**



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

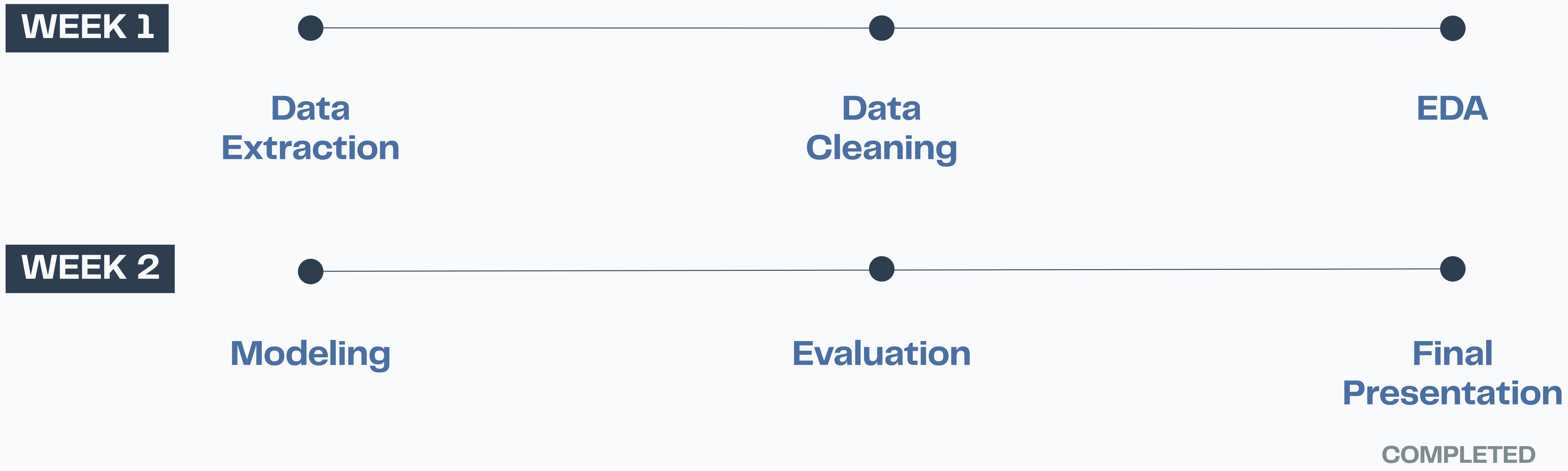
Support Vector Classifier (with RBF kernel)

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

Project Timeline



WEEK 1

Data Extraction

Data Cleaning

EDA

WEEK 2

Modeling

Evaluation

Final Presentation

COMPLETED

Conclusion

Key Findings

Unsupervised models consistently identified **two to three natural clusters**, revealing a simpler structure than the current eleven-grade classification

Model Agreement

K-Means, Agglomerative, and GMM **all produced similar groupings**, confirmed by supervised models showing the same overlapping crude categories

Industry Insight

Results suggest Canadian crudes exist on a **continuous quality spectrum**, better represented by broader heavy, medium, and light/synthetic families

Future Direction

Expanding the analysis with **economic, carbon, and yield data** could strengthen classification accuracy and industry applicability