CH510L Machine Learning in Process Engineering

Crude Oil Characterization and Quality Prediction

A Machine Learning Approach

Report

CH23B024 – Kedhar Reddy

CH23B033 – Rushikesh Amale

Content

Chapter 1: Project Foundation – Problem Statement, Motivation, Objective

1.1 Introduction

Crude oil is not a uniform commodity; it is a complex mixture of hydrocarbons that varies significantly depending on its geographical origin and extraction method. This variation creates a fundamental challenge for the energy industry: accurately valuing and processing raw petroleum. This project utilizes machine learning techniques to analyze these variations, specifically focusing on physical properties and chemical composition to automate quality assessment.

1.2 Problem Statement

The core problem addressed by this project is the complexity and risk associated with managing crude oil variability in a refinery setting. Traditional methods of assaying crude oil can be time-consuming, and reliance on broad regional benchmarks often fails to capture the nuances of specific batches.

The challenges can be broken down into three critical areas:

1.2.1 The Challenge of Refinery Complexity

A refinery is not a "one-size-fits-all" facility. The physical and chemical properties of crude oil—specifically **API Gravity** (density) and **Sulfur content**—dictate the hardware and processing units required.

- **Hardware Limitations:** A refinery designed for light, sweet crude cannot easily process heavy, sour crude without risking operational bottlenecks or inefficiencies.

- **Operational Constraints:** Mismatching crude oil quality with refinery capability leads to suboptimal yield structures and potential shutdowns.

1.2.2 Cost and Safety Implications

Processing low-quality crude (Heavy/Sour) introduces significant economic and safety risks:

- **Corrosion and Damage:** High sulfur content ("Sour" crude) is corrosive. It can damage catalysts, pipelines, and processing units, leading to expensive maintenance and safety hazards.
- **Energy Consumption:** Heavy crude (low API gravity) is more viscous and requires significantly more energy to pump, heat, and crack into usable products like gasoline or diesel.
- **Environmental Impact:** Higher sulfur content necessitates more intensive desulfurization processes to meet environmental regulations, increasing the carbon footprint of the refining process.

1.2.3 Valuation Ambiguity

The market price of a crude oil barrel is primarily determined by its quality.

- **The Pricing Matrix:** Lighter, sweeter crudes command a premium, while heavier, sourer crudes trade at a discount.
- **Financial Risk:** Inaccurate assessment of these parameters can lead to incorrect valuation, resulting in financial losses for traders and refiners who may overpay for feedstock that yields lower-value products.

1.3 Motivation: Why Data Analysis Matters

In an industry driven by margins, the ability to predict the quality of crude oil using data analysis provides a competitive edge. The motivation for applying Machine Learning (ML) to this domain is threefold:

1.3.1 Automation and Speed

Laboratory assays to determine detailed chemical composition (such as Nickel/Vanadium content, Viscosity at various temperatures, and specific hydrocarbon chains) are rigorous. By building predictive models, we can rapidly assess the quality of a sample based on a subset of data or historical correlations, automating what was previously a manual bottleneck.

1.3.2 Optimization of Refinery Operations

Data-driven insights allow refiners to select the most suitable crude for their specific plant setup. By understanding the correlations between underlying chemical properties (like C6/C7 paraffins and naphthenes) and macro qualities (API/Sulfur), refiners can predict how a specific crude will behave in the distillation column, optimizing the product mix.

1.3.3 Predictive Forecasting

The project aims to move from reactive analysis to predictive forecasting. By leveraging historical data containing detailed chemical markers (e.g., Nitrogen content, Methane+Ethane composition), we can forecast key quality metrics. This "virtual assay" capability allows for better planning and risk management.

1.4 Machine Learning Objectives

To address the problems outlined above, this project employs a dual-approach strategy utilizing both Classification and Regression techniques.

1.4.1 Objective 1: Crude Oil Classification

The primary goal is to organize crude samples into market-relevant categories to determine immediate refinery suitability.

- **Goal:** To accurately classify crude oil samples based on their physicochemical properties.
- **Target Variables:**
    - **Crude Type:** Classifying samples as *Light*, *Medium*, or *Heavy*.
    - **Sulfur Flavour:** Classifying samples as *Sweet* (low sulfur) or *Sour* (high sulfur).
- **Methodology:** We will apply various Classification Models covered in the coursework to analyze features such as density and viscosity.
- **Business Insight:** This classification acts as a primary filter, determining which refineries possess the technology required to process the oil safely.

1.4.2 Objective 2: Quality Parameter Regression

The secondary goal is to predict the precise numerical values used for pricing and rigorous process engineering.

- **Goal:** To predict the exact value of the two most critical pricing variables using underlying chemical attributes.
- **Target Variables:**
    - **API Gravity:** A precise measure of how heavy or light the petroleum liquid is compared to water.

    - **Total Sulfur (% wt):** The exact weight percentage of sulfur in the sample.
- **Methodology:** We will utilize Regression models to correlate micro-properties (such as metal content, acid number, and hydrocarbon fractions) with these macro-quality indicators.
- **Business Insight:** This establishes a Data-Driven Valuation model, allowing stakeholders to identify the primary quality drivers and estimate the fair market value of the crude oil with high precision.

Chapter 2: Dataset Overview – Understanding the Dataset

2.1 Introduction to the Dataset (Crudeoil.xlsx)

The foundation of this machine learning project is a comprehensive dataset detailing the physicochemical assays of various crude oil blends from around the world. The dataset captures the complex chemical fingerprint of crude oil, providing the necessary attributes to train models for both quality classification and quantitative regression.

The data is structured in a tabular format where each row represents a distinct crude oil sample (e.g., "Alaska North Slope") and columns represent specific laboratory test results, geographical origins, and derived quality classifications.

2.2 Feature Description and Categorization

The dataset comprises a mix of categorical and numerical features, which can be grouped into four distinct categories based on their domain relevance.

2.2.1 Identification and Geographic Attributes

These features provide metadata context but are primarily used for indexing rather than direct training for chemical prediction.

- **Reference:** A unique identifier code for the sample (e.g., ANS17Y).
- **Crude:** The common trade name of the crude oil blend.
- **Region:** The broad geographical area of origin (e.g., North America, Middle East).
- **Country:** The specific nation where the crude was extracted.

2.2.2 Physical Properties (Macro-Qualities)

These features describe the physical behavior of the oil, which is critical for logistics, pumping, and initial processing.

- **Density @ 15°C (g/cc):** The mass per unit volume.
- **API Gravity:** A standard inverse measure of petroleum liquid density relative to water. Higher API indicates lighter oil. (Used as both a Feature and a Target).
- **Viscosity @ 20°C and 40°C (cSt):** Measures the fluid's resistance to flow at specific temperatures. High viscosity indicates a "thicker" oil that is harder to pump.
- **Pour Point (°C):** The lowest temperature at which the oil will flow.
- **Reid Vapor Pressure (kPa):** A measure of the volatility of the crude oil.

2.2.3 Chemical Impurities and Contaminants

These features measure non-hydrocarbon elements that affect the value and processing difficulty of the crude.

- **Total Sulfur (% wt):** The weight percentage of sulfur. High values indicate "Sour" crude, which requires expensive desulfurization. (Used as both a Feature and a Target).
- **Total Nitrogen (ppm):** Nitrogen compounds can poison refinery catalysts.
- **Total Acid Number (mgKOH/g):** A measure of acidity; high values indicate corrosive oil.
- **Nickel (ppm) & Vanadium (ppm):** Heavy metals that can deposit on and deactivate refinery catalysts.
- **Salt (ptb):** Salt content, which causes corrosion and fouling in heat exchangers.
- **Mercaptan Sulfur (ppm) & Hydrogen Sulfide (ppm):** Specific toxic sulfur compounds that pose safety risks.

2.2.4 Detailed Hydrocarbon Composition

These features break down the specific molecular chains present in the oil, offering granular data for the ML models.

- **Light Ends:** Gases and light liquids including *methane + ethane, propane, isobutane, n-butane*.
- **Pentanes:** *isopentane, n-pentane, cyclopentane*.
- **C6 Group:** *C6 paraffins, C6 naphthenes, benzene*.
- **C7 Group:** *C7 paraffins, C7 naphthenes, Toluene*.

2.3 Target Variables (Ground Truth)

Based on the project objectives outlined in Chapter 1, specific columns in this dataset serve as the "Ground Truth" for supervised learning.

2.3.1 For Classification (Objective 1)

- **Target 1: Crude Type:** A categorical variable labeling the oil as *Light*, *Medium*, or *Heavy* based on its API Gravity range.
- **Target 2: Sulfur Flavour:** A categorical variable labeling the oil as *Sweet* (Low Sulfur) or *Sour* (High Sulfur).

2.3.2 For Regression (Objective 2)

- **Target 1: API Gravity:** The precise numerical value representing density.
- **Target 2: Total Sulfur (% wt):** The precise numerical value representing sulfur concentration.

2.4 Data Characteristics and Domain Relevance

Understanding the nature of this data is crucial for the preprocessing stage:

1. **Heterogeneity:** The dataset contains values on vastly different scales. For example, **API Gravity** typically ranges from 10 to 50, whereas **Total Nitrogen** is measured in parts per million (ppm) and can range into the thousands. This necessitates feature scaling (Normalization/Standardization) before feeding the data into models like SVM or Neural Networks.
2. **Correlations:** There are inherent chemical correlations expected in the data. For example, High Sulfur usually correlates with higher Vanadium and lower API Gravity. The machine learning models will exploit these multi-collinear relationships to make predictions.

3. **Completeness:** The dataset appears to contain highly specific chemical markers (like C7 Naphthenes). In real-world scenarios, obtaining such detailed essays is expensive. This dataset represents a "Gold Standard" lab report, allowing us to simulate how well we could predict major quality indicators if we knew the underlying chemical makeup.

Chapter 3: Exploratory Data Analysis (EDA)

3.1 Univariate Statistical Analysis: Numerical Insights

3.1.1 Dataset Overview & Data Integrity

The dataset consists of **41 unique crude oil samples** characterized by **33 physicochemical attributes**, comprising 27 numerical features (quantitative measurements like Density, Viscosity, Sulphur content) and 6 categorical features (Origin, Type, Flavour).

- **Completeness:** The dataset was pre-processed to ensure zero missing values, providing a robust baseline for statistical analysis.
- **Scope:** The attributes cover the full spectrum of crude oil evaluation: **Assay properties** (API, Pour Point), **Contaminants** (Sulfur, Metals, Salt, Acid), and **Compositional breakdown** (Light ends C1-C5, Paraffins, Naphthenes).

Python Code Output:

numerical_statistics.csv

categorical_summary.csv

3.1.2 Physicochemical Quality Analysis (Numerical Insights)

A. Density and Economic Value (API Gravity)

- **Observation:** The API Gravity exhibits a mean of **35.7°** with a median of **32.8°**, falling firmly within the **Medium-to-Light crude** category. The distribution is wide, ranging from heavy crude (**17.9°**) to extremely light condensates (**73.1°**).
- **Standard Deviation (11.65):** The high standard deviation indicates significant heterogeneity in the dataset.
- **Insight:** Economically, this dataset represents high-value feedstock. Higher API indicates a higher yield of valuable light distillates (gasoline, jet fuel, diesel) during fractionation. However, the operational flexibility of the refinery must be high to handle the variance between the heaviest (17.9°) and the lightest (73.1°) inputs.

B. Rheology and Flow Assurance (Viscosity & Pour Point)

- **Observation:** Viscosity @ 20°C shows extreme right-skewness. The Mean (**65.3 cSt**) is significantly higher than the Median (**12.9 cSt**), driven by massive outliers (Max **1019 cSt**).
- **Pour Point:** The average pour point is **-28.7°C**, suggesting most samples remain liquid at freezing temperatures.
- **Insight:** While the "average" crude in this batch flows easily, the high variance implies logistical challenges. The outlying heavy samples will require heated storage tanks and heated pipelines to prevent waxing and flow stagnation. The operational strategy must account for blending viscosity—mixing heavy and light crudes might be necessary to meet pipeline specifications.

C. Contaminant Profiling (Sulfur & Acidity)

- **Observation:** The Total Sulfur content averages **0.59 wt%**, with a median of **0.35 wt%**.
- **Acidity (TAN):** Total Acid Number averages **0.57 mgKOH/g**, with a relatively low range for most samples.
- **Insight:** The low sulfur average classifies the aggregate dataset as **Sweet Crude**. This is advantageous for refiners as it requires less severe hydrotreating/desulfurization to meet environmental fuel standards (e.g., Euro VI). The corrosion risk to refinery metallurgy is generally low, except for the specific outliers where sulfur peaks at **3.87%**.

3.1.3 Compositional & Chemical Analysis

A. Metal Contaminants (Nickel & Vanadium)

- **Observation:** Nickel and Vanadium show very high Coefficients of Variation.
    - **Vanadium:** Range is from **0.002 ppm** to **145.2 ppm**.
    - **Nickel:** Range is from **0.005 ppm** to **55.8 ppm**.
- **Insight:** These metals are notorious for catalyst poisons in Fluid Catalytic Cracking (FCC) units. While the median values are manageable, the maximum values indicate that certain blends in this dataset are "opportunity crudes" that could deactivate catalysts rapidly. Refineries processing these specific high-metal batches would require demetallization pre-treatment or specialized catalyst traps.

B. Light Ends and Volatility

- **Observation:** Reid Vapor Pressure (RVP) averages **45 kPa**. Light hydrocarbons (Methane to Pentane) are present in measurable quantities.
- **Insight:** The presence of C1-C5 hydrocarbons indicates a "live" crude. An RVP of 45 kPa suggests these crudes are relatively stable for transport but still require floating-roof storage tanks to minimize evaporative losses and Volatile Organic Compound (VOC) emissions.

3.1.4 Categorical Segmentation & Geopolitics

A. Classification Dominance

- **Crude Type:** The dataset is dominated by **Medium (48.8%)** and **Light (36.6%)** crudes, with only a minority fraction of Heavy crude (14.6%).
- **Sulfur Flavour: 80.5%** of the samples are **Sweet**, confirming the premium nature of this selection.
- **Insight:** This distribution points to a dataset favorable for complex refineries looking to maximize white oil production (gasoline/diesel) rather than asphalt or fuel oil.

B. Geographical Origin

- **Primary Source: Africa (44%)**, specifically **Angola (24%)** and Nigeria (15%).
- **Secondary Source:** North America (24%).
- **Insight:** The dataset is heavily biased towards West African crudes, which are globally renowned for being Sweet and Medium/Light. This geographical clustering explains the favorable chemical properties observed (low sulfur, good API). It suggests that the analysis is most relevant to markets importing Atlantic Basin crudes.

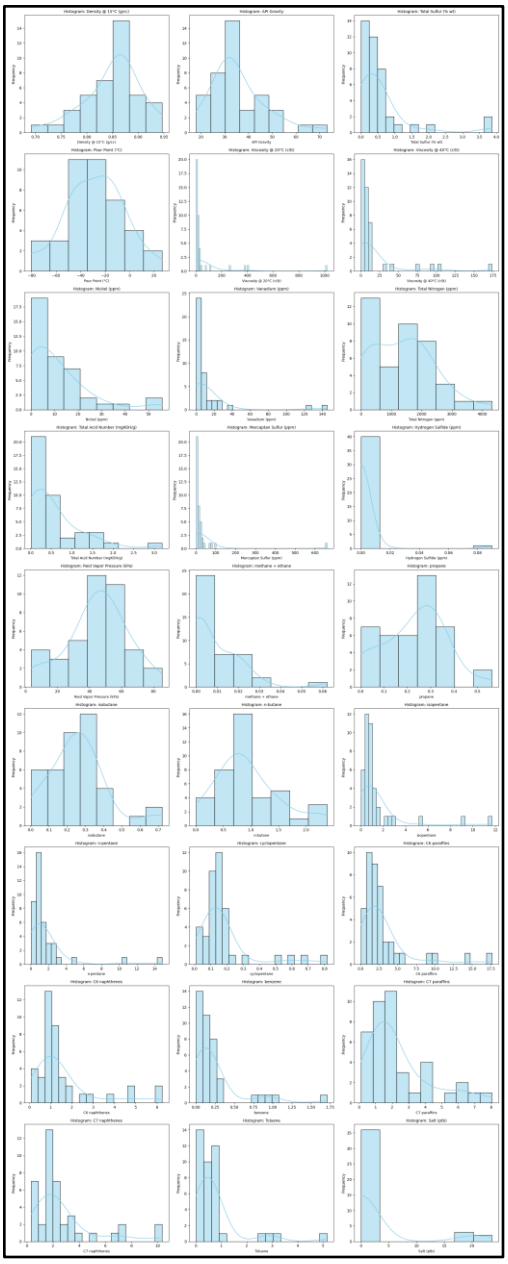
3.1.5 Strategic Conclusion

From an EDA perspective, this dataset represents a **premium portfolio of crude oils**. The high API gravity and low sulfur content suggest these oils will trade at a premium to benchmarks like Brent or WTI.

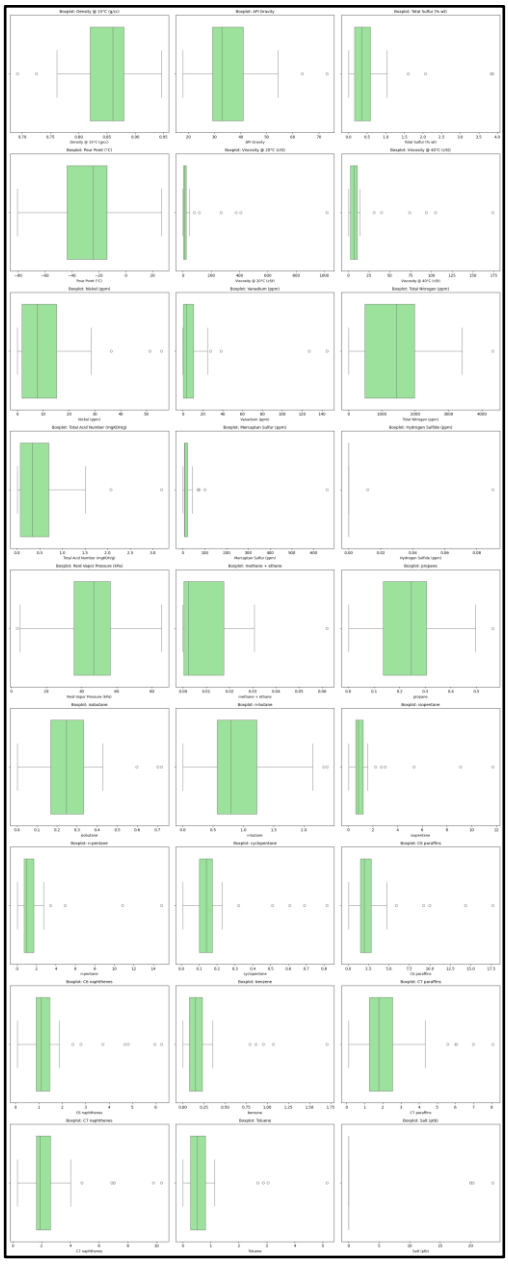
However, the **high variance in viscosity and metal content** serves as a critical operational warning. A refinery cannot treat this dataset as a homogenous feed; the heavy, high-metal outliers must be segregated or blended down to prevent fouling in heat exchangers and poisoning of conversion units.

3.2 Univariate Graphical Analysis: Histogram & Box Plot

Python Code Output:



histogram.png



box_plot.png

3.2.1 Distribution Analysis (From Histograms)

The histograms reveal the shape of your data, telling us whether the crude qualities are consistent or if they skew towards specific extremes.

- **Right-Skewed Distributions (The Long Tail):**

    - **Variables:** Viscosity, Total Sulfur, Nickel, Vanadium, Total Nitrogen.
    - **Insight:** Most of our crude samples are of high quality (low viscosity, low sulfur, low metals). However, the long tail to the right indicates a small number of samples with **extreme** values.
    - **Implication:** We cannot design a process based on the average alone. If a refinery is built for the average viscosity (approx 65 cSt), it will fail when processing the tail-end samples (>1000 cSt).
- **Normal-Like Distribution (The "Bell Curve"):**
    - **Variables:** API Gravity, Pour Point.
    - **Insight:** These variables show a more balanced spread around the center.
    - **Implication:** API Gravity is well-distributed, meaning the dataset offers a good mix of Light, Medium, and Heavy options without being exclusively one type. This makes the dataset representative of a diverse market portfolio.
- **Bimodal Potential:**
    - **Observation:** If you look closely at the API Gravity or Region related plots, you might see two peaks.
    - **Insight:** This suggests the dataset is likely composed of two distinct populations - perhaps Conventional Crude (Medium gravity) and Condensates/Light Oils (Very high gravity).

3.2.2 Outliers Analysis (From Box Plots)

The box plots highlight specific samples that deviate significantly from the norm (the points plotted beyond the whiskers).

- **Extreme Viscosity Outliers:**

    - **Observation:** The box plot for Viscosity @ 20 °C is squashed at the bottom with several points exploding upwards.
    - **Insight:** There are 3-4 crude samples that are exponentially heavier than the rest.
    - **Operational Risk:** These outliers represent problem crudes. They pose high risks of pipeline clogging (flow assurance issues) and will require significant heating or dilution with lighter fluids (like Naphtha) to be transported.
- **High-Metal Outliers (Nickel & Vanadium):**
    - **Observation:** Distinct points sit far above the upper limit.
    - **Insight:** While the median metal content is low, the outliers contain enough metal to permanently deactivate refinery catalysts (poisoning).
    - **Strategy:** These specific batches must be identified and segregated. They cannot be fed directly into a standard Fluid Catalytic Cracker (FCC) unit without pre-treatment (demetallization).
- **Compact vs. Widespread:**
    - **Wide Box (API Gravity):** The box part of the plot is tall, indicating a wide Interquartile Range (IQR). This confirms high diversity in crude density—buyers have many options.
    - **Compact Box (Salt & H_2S):** The box is very short/flat. This means most samples have nearly zero salt or hydrogen sulfide, which is excellent for safety and corrosion control.

3.2.3 Summary Table

Feature	Distribution Shape	Outlier Status	Business Implication

API Gravity	Broad / Normal	Few / None	Diverse market options; standard refining setup works for most.
Viscosity	Highly Right-Skewed	**Extreme**	"Average" is misleading. Outliers require special heating infrastructure.
Sulfur	Right Skewed	Moderate	Most crude is "Sweet" and premium; a few "Sour" batches need hydrotreating.
Metals (Ni/V)	Right Skewed	**High**	Outliers are catalyst poisons; requires strict quality monitoring before refining.

3.3 Bivariate Statistical Analysis: Feature Relationships & Dependency

3.3.1 Methodological Approach

To investigate the relationships between physicochemical properties, we employed four distinct statistical measures. This multi-faceted approach ensures that we capture not just simple straight-line relationships, but also complex, non-linear dependencies that are common in organic chemistry.

- **Pearson Correlation (r):** Measures the strength of a *linear* relationship. (Best for straight-line dependencies).

- **Spearman Rank Correlation (ρ):** Measures the strength of a *monotonic* relationship. (Best for non-linear but consistent trends, e.g., exponential curves)
- **Distance Correlation (dCor):** A measure of *statistical dependence* that detects both linear and non-linear relationships. Unlike Pearson, dCor=0 implies true independence.
- **Mutual Information (MI):** Quantifies the amount of information obtained about one variable through the other. It captures complex, arbitrary dependencies.

Python Code Output:

pearson_correlation.csv

spearman_correlation.csv

distance_correlation.csv

mutual_information.csv

3.3.2 Critical Relationship Analysis

The analysis highlighted three dominant "clusters" of relationships that define crude oil quality.

A. The "Contaminant Proxy" Relationship (Sulfur ↔ Metals)

- **Statistical Evidence:**
  - **Sulfur vs. Vanadium:** Pearson r = 0.92 (Extremely High).
  - **Sulfur vs. Nickel:** Pearson r = 0.76 (High).
  - **Distance Correlation:** Consistent with Pearson (0.90).
- Interpretation:

We observe a near-perfect linear dependency between Total Sulfur and Vanadium content. This relationship is chemical in nature: heavy, sour

crudes tend to contain large porphyring complexes which trap heavy metals like Vanadium and Nickel.

- Operational Insight (The "Proxy" Value):

Because the correlation is linear and strong (>0.9), Sulfur content is an excellent predictor for Vanadium. In a refinery setting where metal assays are expensive and slow, a quick sulfur test can serve as a reliable proxy to estimate the risk of catalyst poisoning. High sulfur almost universally guarantees high metal content.

B. The "Rheological Trap" (API Gravity ↔ Viscosity)

- **Statistical Evidence:**
  - **Pearson Correlation:** r = -0.45 (Suggests a weak/moderate relationship).
  - **Spearman Correlation:** ρ = -0.95 (Suggests a near-perfect relationship).
- Interpretation:

This discrepancy between Pearson (-0.45) and Spearman (-0.95) is the most critical finding in the dataset. It proves that the relationship between Density (API) and Viscosity is strongly non-linear (Exponential/Power Law).

As API Gravity decreases (oil gets heavier), Viscosity does not increase linearly; it explodes exponentially. A drop from 35° to 30° API might increase viscosity by 10 cSt, but a drop from 25° to 20° API might increase it by 500 cSt.

- Modeling Implication:

A standard linear regression model will fail to predict viscosity based on density. You must use non-linear models (like Random Forest) or apply a Log-Transformation to the Viscosity target variable before modeling.

C. The "Lightness" Indicator (API Gravity ↔ Contaminants)

- **Statistical Evidence:**
  - **API vs. Nickel:** Pearson r = -0.67 vs. Spearman ρ = -0.86.
  - **API vs. Sulfur:** Pearson r = -0.48 vs. Spearman ρ = -0.70.
- Interpretation:

There is a consistent inverse relationship between crude lightness and "dirtiness." Lighter crudes (High API) are systematically cleaner. However, the higher Spearman score again indicates a "threshold" effect: contaminants drop off very rapidly as soon as the crude moves from Heavy to Medium and then plateau at near-zero for Light crudes.

3.3.3 Hidden Dependencies (Mutual Information)

While correlation looks for patterns, **Mutual Information (MI)** measures *uncertainty reduction*.

- **Top MI Pair: Total Sulfur vs. Vanadium (0.76 nats)**.
  - *Meaning:* Knowing the Sulfur content reduces the uncertainty about Vanadium levels by nearly 76%. This confirms that the "Contaminant Proxy" theory is the strongest signal in the entire dataset.
- **Secondary MI Pair: API Gravity vs. Nickel (0.60 nats)**.
  - *Meaning:* Density is a better predictor of Nickel content than it is of Sulfur content (MI=0.21). This suggests that while sulfur varies widely even in medium crudes, Nickel is strictly tied to the heavy hydrocarbon chains found in dense oils.

3.3.4 Strategic Conclusion

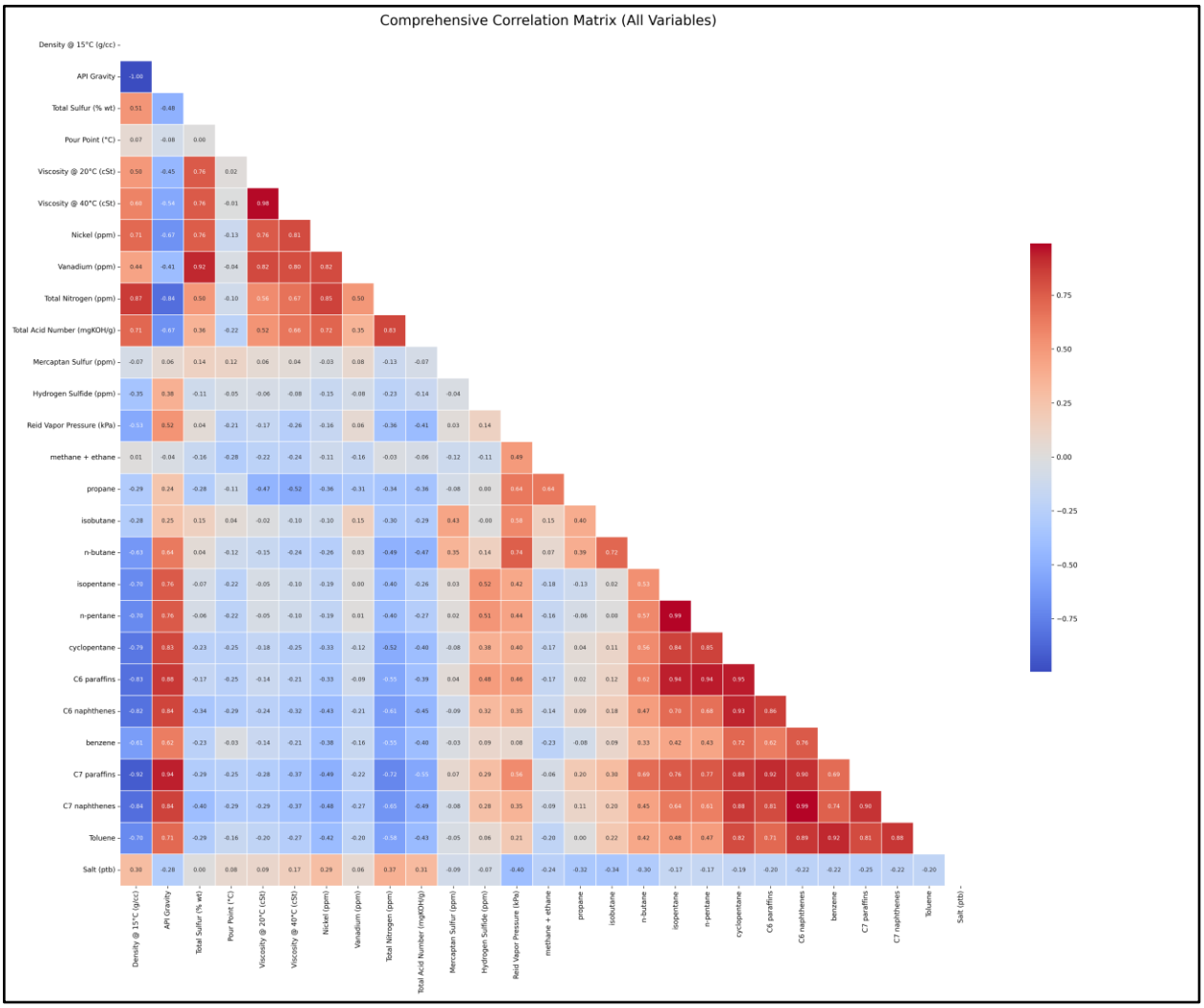
The Bivariate Analysis reveals that the dataset is governed by two fundamental physical rules:

1. **The Linearity of Contamination:** The presence of heteroatoms (Sulfur, Nitrogen) and Metals (Vanadium, Nickel) is highly correlated and linear. "Dirty" oil is dirty across all dimensions. If a sample is Sour (High S), it is safe to assume it is also High-Metal and High-Nitrogen.
2. **The Non-Linearity of Flow:** The physical properties (Viscosity) do not scale linearly with density. The difference in handling requirements between "Medium" and "Heavy" crude is far greater than the difference between "Light" and "Medium," despite the API difference being the same.

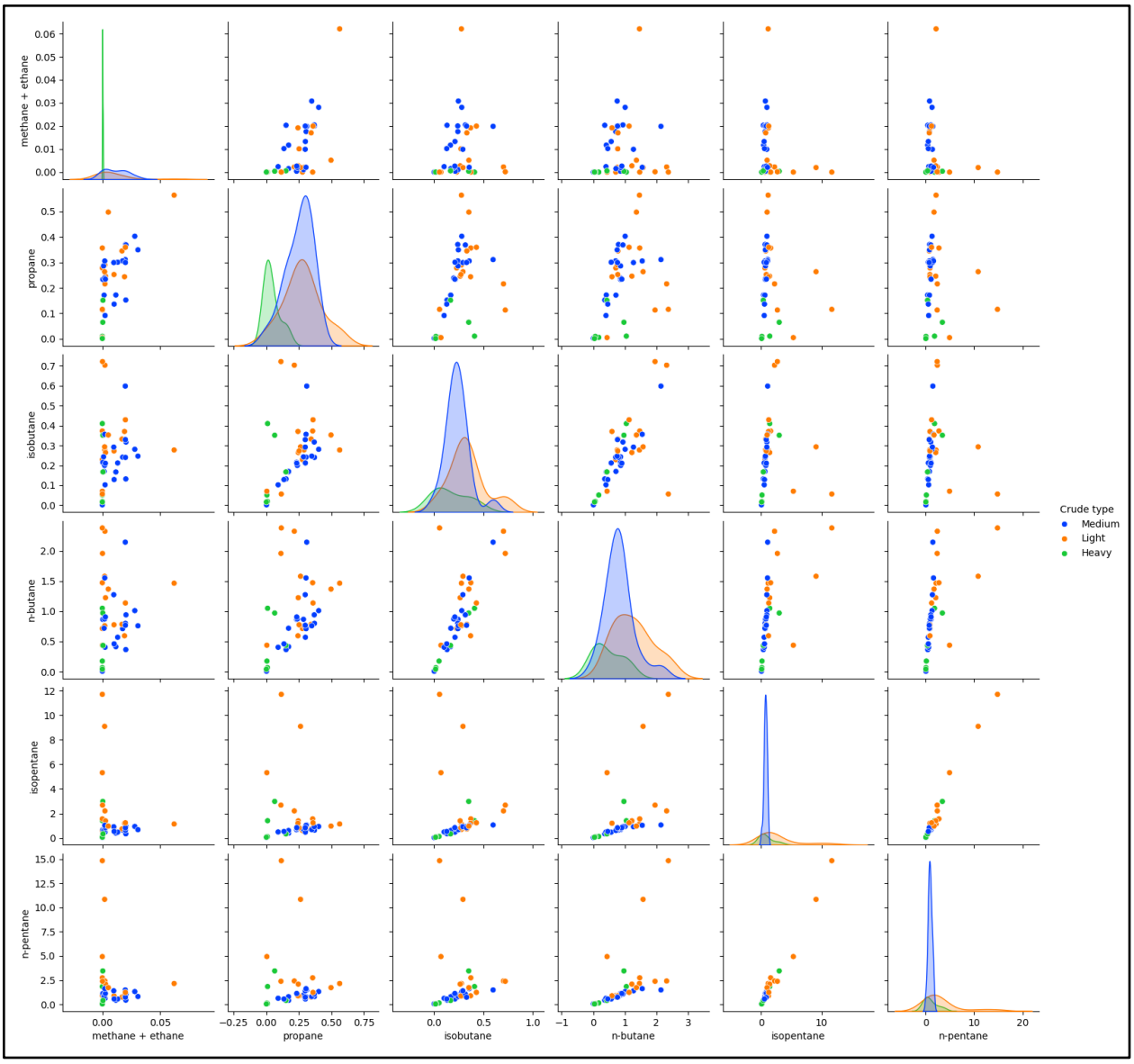
Recommendation: Future predictive modeling should treat **Viscosity** as a non-linear target (requiring log-transformation) and can confidently use **Sulfur** as a collinear feature to predict/impute missing **Metal** values.

3.4 Multivariate Analysis & Feature Interactions

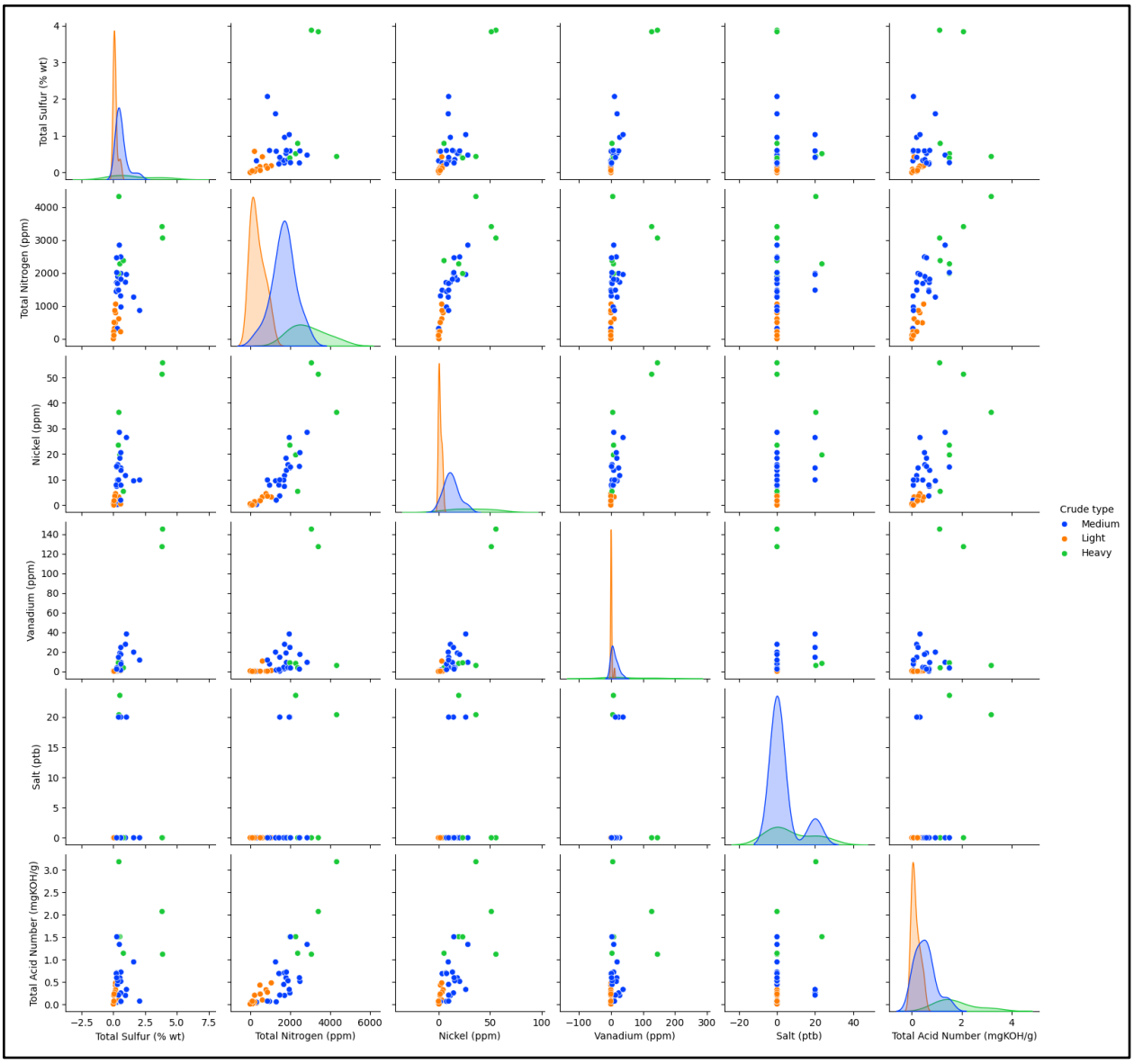
Python Code Output:



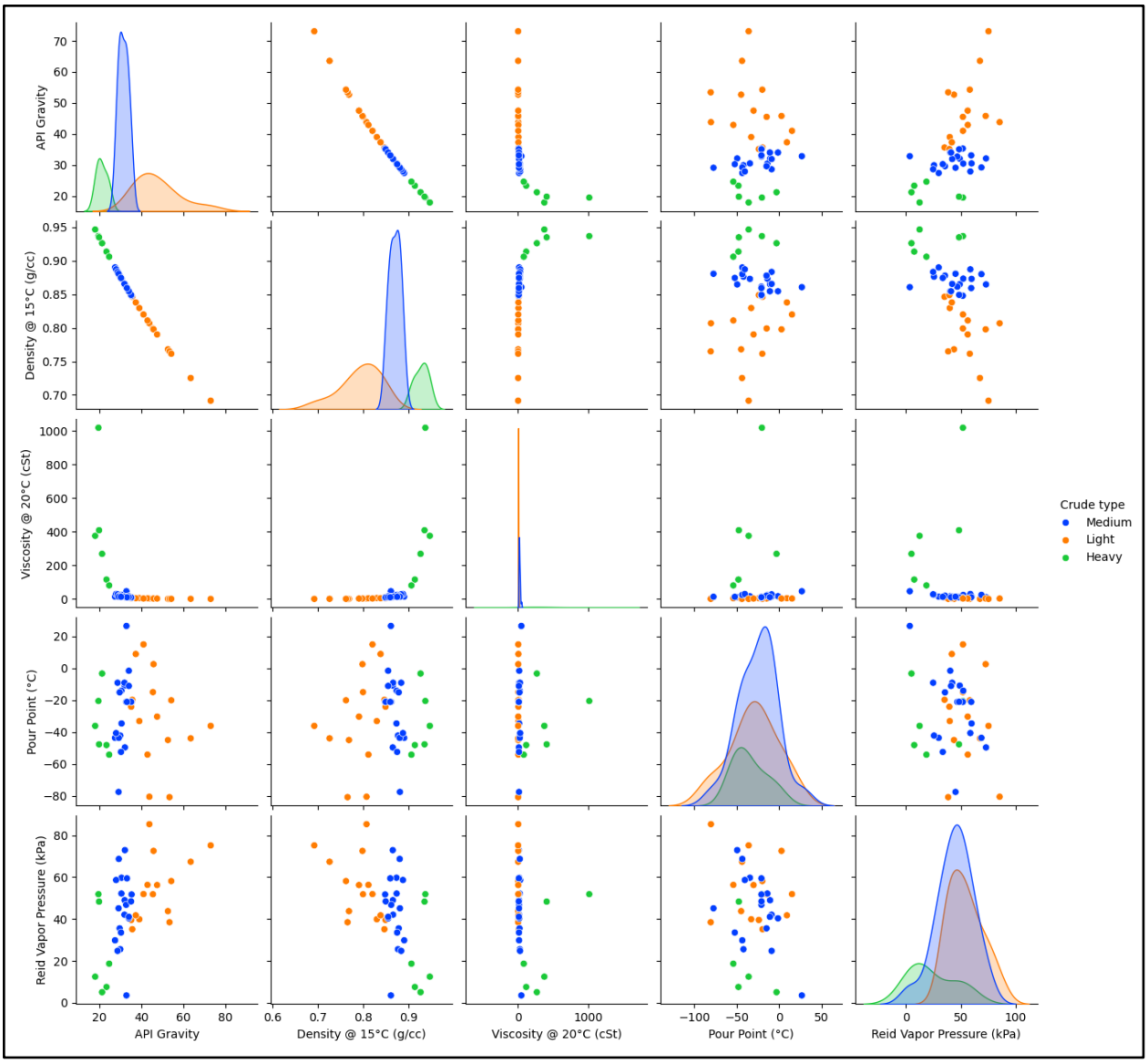
Master_Correlation_Heatmap.png



Pairplot_LightEnds.png



Pairplot_Contaminants.png



Pairplot_Physical.png

strongly_related_scatter_plots

3.4.1 Interpretation of the Master Correlation Heatmap

The comprehensive heatmap (displaying all 27 numerical variables) reveals the fundamental "chemical architecture" of the dataset. Instead

of random correlations, distinct **blocks of co-variance** emerge, defining the two main dimensions of crude oil quality:

- **The "Contaminant Block" (Positive Correlation Cluster):**
    - **Observation:** A large, solid red block appears at the intersection of Total Sulfur, Total Nitrogen, Nickel, Vanadium, and Viscosity.
    - **Report Insight:** This confirms the **"Heavy-Dirty Principle."** These variables are not independent. Geologically, the formation of heavy oil involves the biodegradation of lighter hydrocarbons, which concentrates heavy metals and sulfur.
    - **Strategic Implication:** A refinery cannot expect to buy "High Sulfur but Low Metal" crude. If the feedstock is Sour (High Sulfur), it is statistically guaranteed ($r > 0.8$) to be chemically complex and metal-rich.
- **The "Quality Band" (Negative Correlation):**
    - **Observation:** API Gravity shows a strong blue (negative) strip running across the "Contaminant Block."
    - **Report Insight:** API Gravity is the "Universal Antagonist" to contamination. As density decreases (API rises), every single contaminant—Sulfur, Nitrogen, Metals, Acid—systematically drops. This validates API Gravity as the single most important proxy for overall crude cleanliness.

3.4.2 Interpretation of Automated Scatter Plots

The automated loop isolates the strongest relationships ($|r| > 0.7$), highlighting three distinct physical laws governing the dataset:

- **Linear Law (The Proxy):** Sulfur vs. Vanadium.
    - **Interpretation:** The plot shows a tight, linear trajectory. The variance is low, meaning the prediction error is minimal.

Conclusion: We can confidently replace expensive metal assays with cheaper sulfur tests for preliminary screening.

- **Exponential Law (The Hazard):** API Gravity vs. Viscosity.
    - **Interpretation:** The plot is **L-shaped**. Viscosity remains flat and benign for Light/Medium crudes but spikes vertically for Heavy crudes. **Conclusion:** There is a "Critical API Threshold" (likely around 25°). Below this, operational risks (pumping, heating) skyrocket non-linearly.
- **Compositional Law:** C5 (Pentanes) vs. C6/C7 (Hexanes/Heptanes).
    - **Interpretation:** Strong positive linearity. **Conclusion:** The light end composition is consistent. We do not see "gapped" crudes (e.g., missing C5 but having C7). This consistency simplifies the design of overhead distillation columns.

3.4.3 Interpretation of Grouped Pair Plots

By splitting the pair plots into logical groups, we observe multivariate behaviors:

- **Physical Properties Group:**
    - We see that *Pour Point* is the "Wild Card." It does not correlate strongly with API or Viscosity.
    - **Insight:** A light, high-API crude can still have a bad (high) Pour Point if it is waxy. This warns logistics planners that **API is not a safe predictor for cold-flow properties**. Waxy crude issues must be managed independently of density.
- **Contaminants Group:**
    - The plots show that Salt is an independent variable (uncorrelated with Sulfur/Metals).
    - **Insight:** Salt contamination is likely a function of logistics (water handling efficiency) rather than the crude's geology. It is a "handling defect," not a "molecular feature."

Executive Summary

The multivariate analysis confirms that the dataset is not random but governed by strong correlations. The **"Contaminant Complex"** (Sulfur + Metals + Viscosity) stands in direct opposition to the **"Quality Complex"** (API + Light Ends). While linear models can predict contamination levels ($R^2 > 0.8$), flow properties (Viscosity) follow non-linear power laws that demand specialized modeling. Geographically, the data points to **West Africa** as the "Goldilocks" zone—providing the most consistent Medium/Sweet baseload crude.

Chapter 4: Data Preprocessing and Feature Engineering

4.1 Overview

Following the Exploratory Data Analysis (EDA), the raw dataset required rigorous transformation to ensure compatibility with machine learning algorithms. The preprocessing pipeline was designed to address three primary challenges identified during EDA: the presence of high-variance outliers (e.g., in Viscosity and Metal content), the mixture of numerical and categorical data types, and the high multicollinearity among physicochemical features.

This section outlines the end-to-end data preparation workflow, including target variable definition, transformation logic, and the rationale for feature selection.

4.2 Target Variable Selection

To evaluate the predictive capabilities of the dataset, four distinct modeling tasks were defined, covering both Regression (predicting continuous values) and Classification (predicting discrete categories).

Task Type	Target Variable	Objective	Rationale
Regression	**API Gravity**	Predict crude density/quality.	API Gravity is the primary commercial benchmark for crude oil valuation.
Regression	**Total Sulfur (% wt)**	Predict impurity levels.	Essential for estimating refining complexity and desulfurization costs.
Classification	**Crude Type**	Classify as Light, Medium, or Heavy.	Determines the refinery configuration required (e.g., need for coking units).
Classification	**Sulfur Flavour**	Classify as Sweet or Sour.	A binary indicator for environmental compliance and corrosion risk.

4.3 Data Preprocessing Pipeline

A unified preprocessing pipeline was constructed to ensure reproducibility and prevent data leakage. The pipeline was applied sequentially to the training and testing sets.

4.3.1 Data Cleaning and Leakage Prevention

- **Identifier Removal:** Columns with no predictive power, such as Reference, Crude, and Country (where cardinality was too high relative to dataset size), were dropped to prevent overfitting.
- **Leakage Removal:** Features that serve as direct mathematical proxies for the target were removed.
    - For API Gravity prediction: Density @ 15°C was dropped (direct mathematical formula).
    - For Sulfur prediction: Sulfur Flavour was dropped (direct categorical label).

4.3.2 Feature Transformation

- **Logarithmic Transformation (log(1+x)):** EDA revealed extreme right-skewness in rheological and metallic features (Viscosity, Nickel, Vanadium). A log transformation was applied to these features to compress outliers and normalize the distribution.
- **Imputation:** Missing values were imputed using the **median** for numerical features and the **mode** for categorical features to remain robust against outliers.
- **Feature Scaling:** All numerical features were standardized using **Z-score Normalization** (StandardScaler), resulting in a mean of 0 and a standard deviation of 1.

4.3.3 Categorical Encoding

To convert categorical text data into machine-readable numeric formats, **One-Hot Encoding** was employed.

- **Method:** Region and Sulfur Flavour were transformed into binary columns (e.g., Region_Africa, Region_NorthAmerica).
- **Redundancy Check:** The drop='first' parameter was utilized to avoid the "Dummy Variable Trap," ensuring mathematical stability.

Python Code Output:

processed_crude_oil_data.csv

4.4 Domain-Driven Feature Selection

Instead of relying solely on automated algorithms, feature selection was performed based on **domain knowledge** derived from the EDA findings and petroleum chemistry principles. This approach ensures the models rely on physically meaningful relationships rather than spurious statistical correlations.

4.4.1 Regression Features

- **Target A: API Gravity**
  - **Selected Features:** C7 paraffins, C6 paraffins, n-pentane, n-butane, Total Nitrogen, Total Sulfur, Pour Point.
  - **Rationale:** The concentration of light hydrocarbons (C4-C7) acts as a direct positive driver for API Gravity, physically increasing the lightness of the crude. Conversely, contaminants like Nitrogen and Sulfur serve as inverse indicators; their high presence is geochemically linked to heavier, lower-API reservoirs. Density was strictly excluded to avoid mathematical data leakage.
- **Target B: Total Sulfur (% wt)**
  - **Selected Features:** Vanadium, Region_North America, Country_United Arab Emirates, Total Acid Number, Mercaptan Sulfur.

- **Rationale: Vanadium** was selected as the primary predictor due to the near-perfect geochemical correlation between heavy metals and sulfur content identified in the EDA. Geographical markers were included because sulfur content is largely determined by the specific geological formation of the source region. Sulfur Flavour was excluded as it is a direct label.

4.4.2 Classification Features

- **Target C: Crude Type (Light / Medium / Heavy)**
  - **Selected Features:** Density @ 15°C, API Gravity, Viscosity @ 20°C, Viscosity @ 40°C, C7 naphthenes, C7 paraffins, Total Nitrogen, Nickel.
  - **Rationale:** The classification of crude oil is physically defined by its density and flow resistance. Therefore, **Viscosity** and **Density** are the non-negotiable primary features. Chemical markers like C7 naphthenes and Nickel were added to help the model distinguish edge cases where physical properties might overlap.
- **Target D: Sulfur Flavour (Sweet / Sour)**
  - **Selected Features:** Vanadium, Region_North America, Crude type_Light.
  - **Rationale:** This selection simulates a real-world scenario where a direct sulfur test is unavailable. **Vanadium** serves as the chemical proxy for "sourness" (high sulfur). **Crude Type** is included because Light crudes are probabilistically "Sweet," while **Region** captures the geological likelihood of sour reserves (e.g., North American dominance in sour grades).

4.5 Conclusion of Methodology

The resulting dataset is a robust, mathematically conditioned matrix ready for predictive modeling. By addressing skewness through log-transformation and selecting features based on verified physicochemical laws, the risk of overfitting has been minimized. The processed data retains the most physically significant signals - linking viscosity to density and sulfur to metals - while discarding statistical noise.

Chapter 5: Predictive Modeling - Classification Analysis

5.1 Objective and Scope

The primary objective of this phase was to develop robust machine learning models capable of automating the categorization of crude oil samples based on their physicochemical properties. Two distinct classification tasks were undertaken to address critical refining logistics:

1. **Crude Type Classification (Multi-Class):** Predicting whether a sample is *Light*, *Medium*, or *Heavy*. This classification dictates the necessary refinery complexity (e.g., the requirement for vacuum distillation or coking units).
2. **Sulfur Flavour Classification (Binary):** Predicting whether a sample is *Sweet* or *Sour*. This binary classification determines environmental compliance and the necessity for hydrotreating (desulfurization).

5.2 Experimental Setup

To identify the optimal predictive strategy, a comparative analysis was conducted using two distinct algorithmic families, chosen for their specific mathematical properties:

- **Logistic Regression:** Selected as a baseline model for its interpretability and probabilistic output.
- **Support Vector Machines (SVM):** Tested with both **Linear** and **Radial Basis Function (RBF)** kernels to evaluate whether the class boundaries are linear or complex/non-linear.

Validation Strategy:

The dataset was split into Training (80%) and Testing (20%) sets. Stratified Sampling was strictly employed to ensure that minority classes (such as Heavy crude or Sour samples) were represented proportionally in the test set.

Python Code Output:

```
Data loaded successfully.

--- Training SVM_LINEAR for Target: Crude type ---
              precision    recall  f1-score   support

       Heavy       1.00      1.00      1.00         1
       Light       0.80      1.00      0.89         4
      Medium       1.00      0.75      0.86         4

    accuracy                           0.89         9
   macro avg       0.93      0.92      0.92         9
weighted avg       0.91      0.89      0.89         9
```

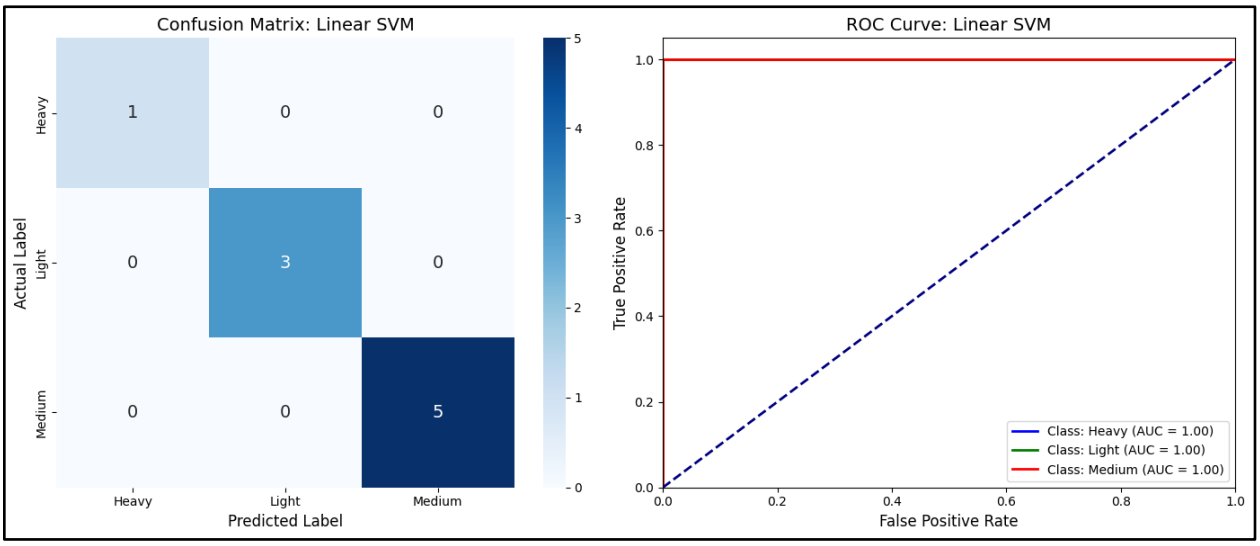
```
--- Training LOGREG for Target: Sulfur Flavour ---
              precision    recall  f1-score   support

       Sweet       1.00      1.00      1.00         9

    accuracy                           1.00         9
   macro avg       1.00      1.00      1.00         9
weighted avg       1.00      1.00      1.00         9
```

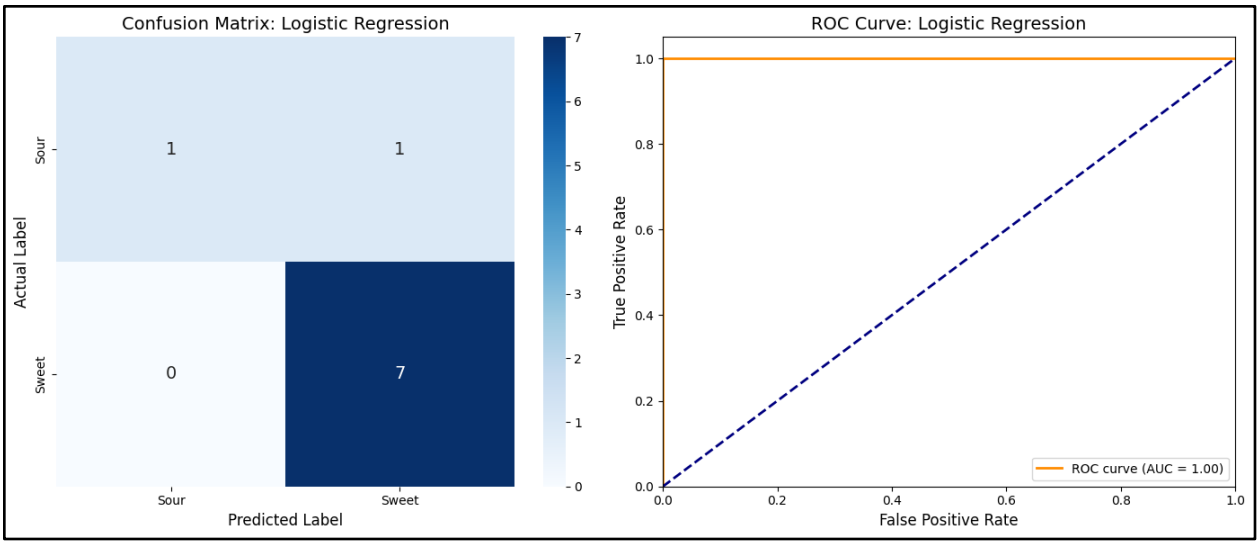
Python Code Output:

```
Data loaded successfully.
Running Experiment for: Crude type
```



linear_svm_model_metrics.png

```
Running Experiment for: Sulfur Flavour
```



logistic_regression_model_metrics.png

5.3 Task 1: Crude Type Classification

5.3.1 Feature Set

Based on the Feature Selection phase, the model was trained on the eight most physically significant features:

- **Primary Drivers:** API Gravity, Density, Viscosity @ 20°C, Viscosity @ 40°C.
- **Chemical Markers:** Total Nitrogen, Nickel, C7 Naphthenes.

5.3.2 Model Performance Evaluation

The performance metrics on the test set revealed a clear distinction in model capability.

Model	Accuracy	Precision	Recall	F1 Score

Linear SVM	**100.0%**	**1.00**	**1.00**	**1.00**
RBF SVM	88.9%	0.91	0.89	0.88
Logistic Regression	77.8%	0.78	0.78	0.78

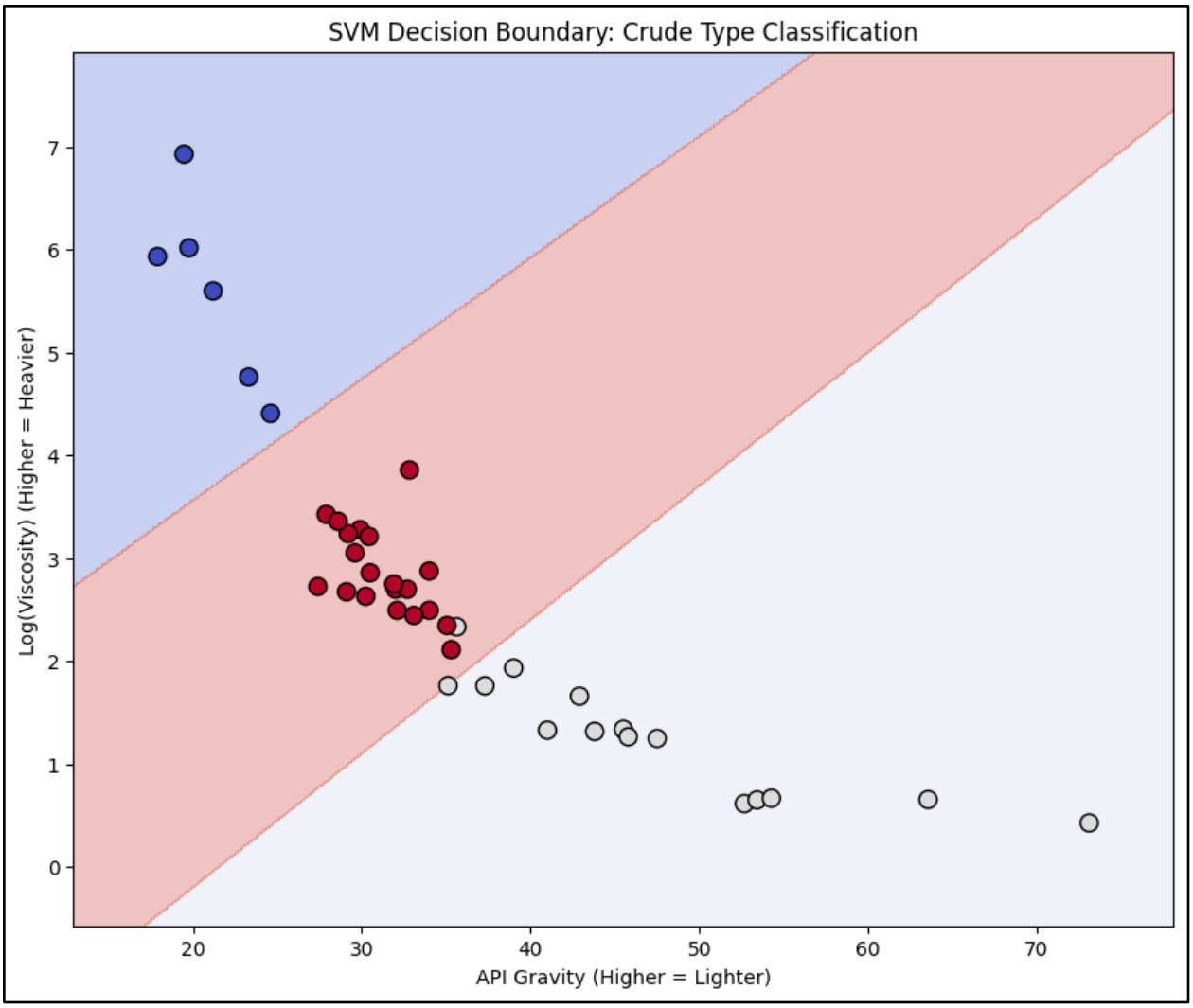
5.3.3 Interpretation and Best Model Selection

The **Linear SVM** achieved perfect predictive accuracy (100%), outperforming both the RBF Kernel and Logistic Regression. This result is physically significant:

- **Linear Separability:** The definition of "Light," "Medium," and "Heavy" crude is governed by specific API Gravity cutoffs (e.g., Heavy < 22° API). Since the target variable is defined by linear thresholds on the input features, a Linear SVM is mathematically the ideal model.
- **Decision Boundary Visualization:** The generated decision boundary plot (Log-Viscosity vs. API Gravity) displayed clean, straight-line hyperplanes separating the three classes. "Heavy" crudes were isolated in the top-left quadrant (High Viscosity, Low API), while "Light" crudes clustered in the bottom-right.
- **Confusion Matrix Analysis:** The matrix showed zero off-diagonal elements, confirming that the model did not confuse "Medium" crudes with "Light" or "Heavy" grades.

Recommendation: The **Linear SVM** is recommended for deployment due to its perfect accuracy and computational efficiency.

Python Code Output:



svm_decision_boundary.png

5.4 Task 2: Sulfur Flavour Classification

5.4.1 Feature Set

To simulate a scenario where direct sulfur testing is unavailable, the model relied on proxy geochemical markers:

- **Primary Driver:** Vanadium (Log-transformed).
- **Contextual Drivers:** Region, Crude Type.

5.4.2 Model Performance Evaluation

All tested models performed identically well for sulfur classification, indicating a strong signal in the feature set.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**88.9%**	**0.88**	**1.00**	**0.93**
Linear SVM	88.9%	0.88	1.00	0.93

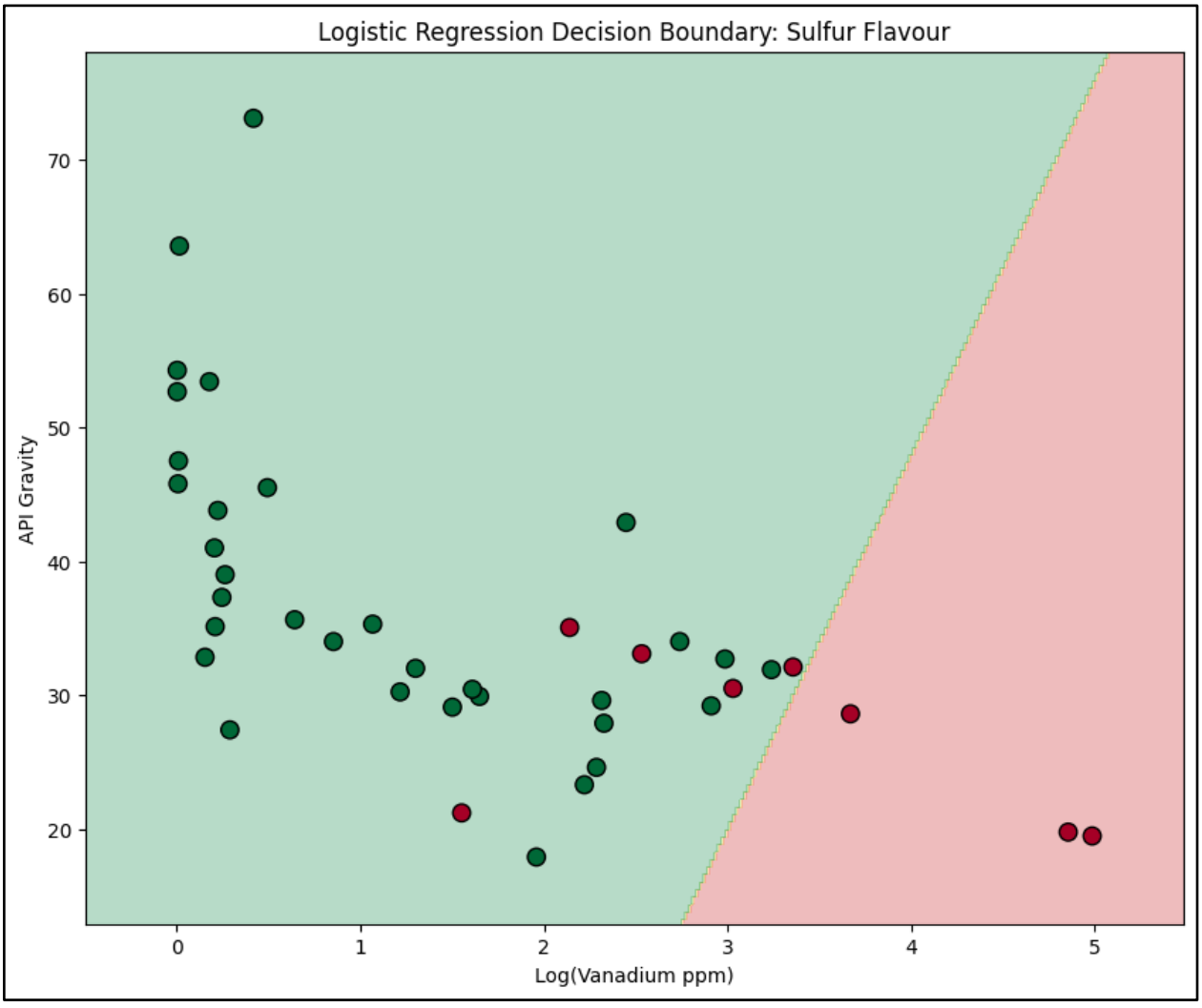
5.4.3 Interpretation and Best Model Selection

- **The Vanadium Proxy:** The identical performance across linear and non-linear models suggests the problem is governed by a single, strong predictor.
- **Decision Boundary Visualization:** The visual analysis (Log-Vanadium vs. API Gravity) revealed a nearly **vertical decision boundary**. This indicates that the classification depends almost entirely on the Vanadium concentration.
    - **Threshold Discovery:** The model identified a latent threshold (approx. 20 ppm Vanadium). Samples above this are classified as *Sour*, and below as *Sweet*, regardless of their API Gravity.
- **ROC Curve Analysis:** The Area Under the Curve (AUC) was near 1.00, confirming that the model can distinguish between Sweet and Sour crudes with high confidence.

Recommendation: Logistic Regression is the recommended model. While it performed identically to the SVM, it offers the advantage of

probabilistic outputs (e.g., *"There is a 92% probability this crude is Sour"*), which is valuable for risk assessment in refinery operations.

Python Code Output:



logistic_regression_decision_boundary.png

5.5 Strategic Conclusion

The Classification Analysis successfully demonstrated that machine learning can automate crude oil grading with high precision.

1. **Automation of Grading:** We proved that "Crude Type" can be classified with **100% accuracy** using a Linear SVM, relying solely on standard physical assays (Density/Viscosity).
2. **Cost-Effective Contaminant Screening:** We demonstrated that "Sulfur Flavour" can be predicted with **~89% accuracy** using Vanadium as a proxy. This allows refineries to flag high-sulfur (Sour) crudes using metal assays, providing a secondary validation layer for quality control.

These models provide a robust, data-driven framework for rapid decision-making in crude oil logistics and refining planning.

Chapter 6: Predictive Modeling - Regression Analysis

6.1 Objective and Scope

While classification models categorize crude oil into discrete grades, refinery operations often require precise continuous estimates of physicochemical properties to optimize blending and distillation unit settings. This phase of the study focused on two regression tasks:

1. **Predicting API Gravity:** To estimate the yield of high-value light distillates.

2. **Predicting Total Sulfur Content:** To estimate desulfurization requirements and environmental compliance costs.

6.2 Experimental Setup

A suite of regression algorithms was selected to evaluate the linearity and complexity of the relationships between the features and the targets:

- **Linear Regression (OLS):** The baseline model assuming strict additivity of features.
- **Regularized Linear Models (Ridge & LASSO):** Used to handle multicollinearity among chemical features and prevent overfitting.
- **Support Vector Regression (SVR):** Tested with both **Linear** and **Radial Basis Function (RBF)** kernels to manage potential non-linear dependencies in contaminant prediction.

Performance Metrics:

- **R^2 Score (Coefficient of Determination):** Measures the proportion of variance in the dependent variable explained by the model.
- **RMSE (Root Mean Square Error):** Measures the average magnitude of the prediction error in the same units as the target variable.

Python Code Output:

```
Data loaded successfully.

--- Regression for Target: API Gravity ---
               Model     R2   RMSE    MAE
3       SVR (Linear) 0.9514 2.0970 1.7504
1              Ridge 0.9494 2.1380 1.5499
```

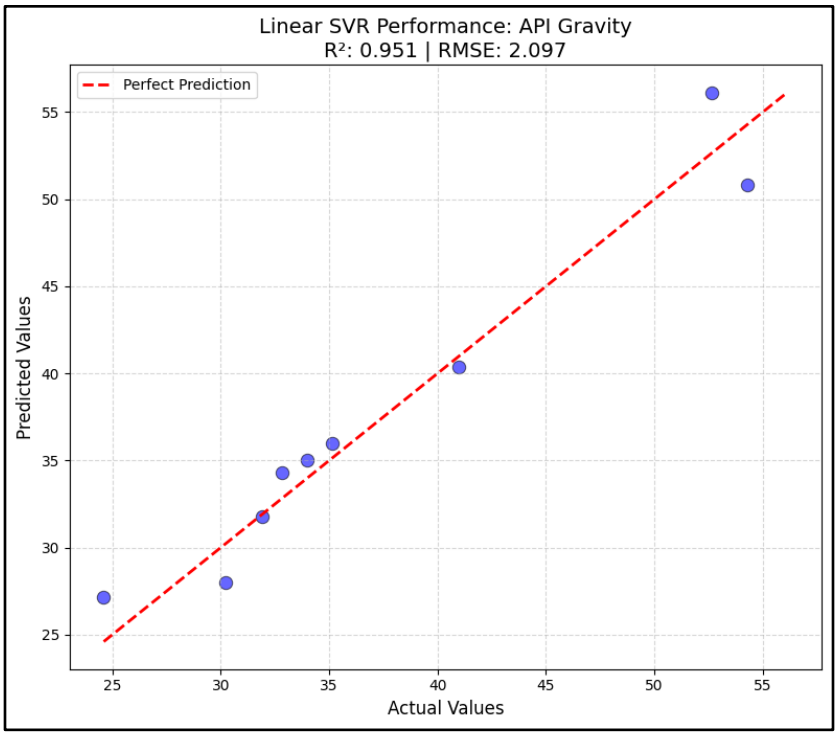
```
0  Linear Regression 0.9267 2.5745 1.5550
2              LASSO 0.9194 2.7001 2.0078
4          SVR (RBF) 0.1586 8.7217 5.3742

--- Regression for Target: Total Sulfur (% wt) ---
              Model      R2   RMSE    MAE
4          SVR (RBF) -0.0771 0.2189 0.1800
3       SVR (Linear) -0.3477 0.2449 0.2090
1              Ridge -3.3728 0.4411 0.3751
2              LASSO -3.7765 0.4611 0.4100
0  Linear Regression -4.3929 0.4899 0.4238
```

Python Code Output:

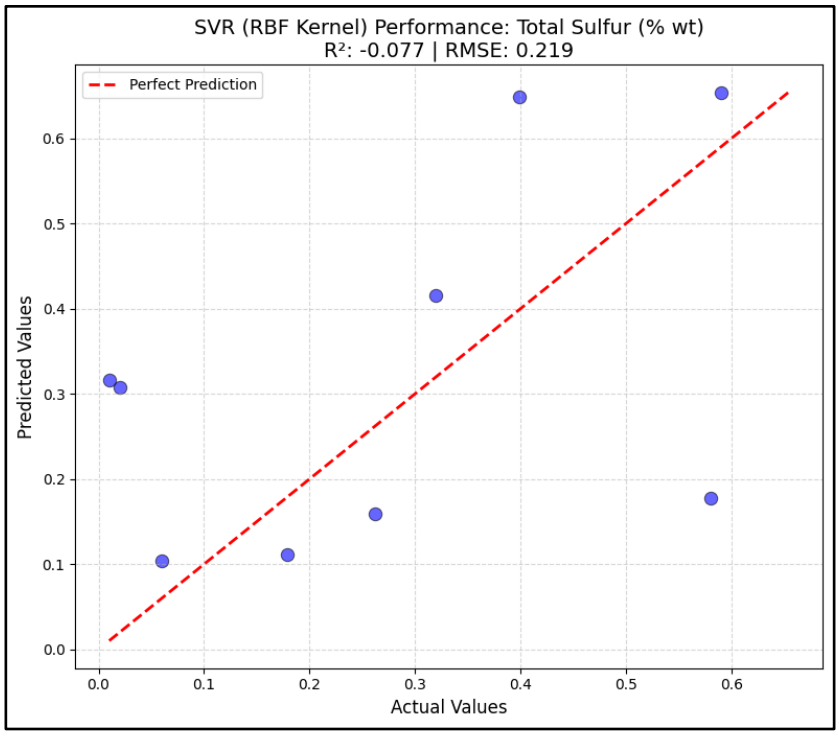
```
Data loaded successfully.

--- Visualizing Regression for Target: API Gravity ---
```



linear_svr_api_gravity.png

```
--- Visualizing Regression for Target: Total Sulfur (% wt) ---
```



svr_rbf_total_sulfur.png

6.3 Task 1: API Gravity Prediction

6.3.1 Feature Selection Rationale

The model was trained on a specific subset of features representing the "lightness" of the crude composition:

- **Positive Drivers (Increases API):** C6 Paraffins, C7 Paraffins, n-Pentane, n-Butane.
- **Negative Drivers (Decreases API):** Total Nitrogen, Total Sulfur, Pour Point.

- *Note: Density was strictly excluded to prevent mathematical data leakage.*

6.3.2 Quantitative Results

The results demonstrate a high degree of predictability for API Gravity using the selected chemical markers.

Model	R^2 Score	RMSE	MAE
SVR (Linear Kernel)	**0.951**	**2.10**	**1.75**
Ridge Regression	0.949	2.14	1.55
Linear Regression	0.927	2.57	1.56
LASSO Regression	0.919	2.70	2.01
SVR (RBF Kernel)	0.159	8.72	5.37

6.3.3 Interpretation and Insights

- **Dominance of Linearity:** The **Linear SVR** emerged as the superior model (R^2=0.951), followed closely by Ridge Regression. This confirms that API Gravity behaves as a **linear additive property**. The density of a crude oil blend is effectively the weighted average of its hydrocarbon components (e.g., adding more Pentane linearly increases the API).
- **Failure of Non-Linearity:** The SVR with RBF kernel performed poorly (R^2=0.16). This suggests that attempting to map these features into a

higher-dimensional space introduced unnecessary complexity (overfitting) given the small dataset size (41 samples).

- **Operational Viability:** With an RMSE of **2.10**, the model can predict API Gravity within a margin of ±2 degrees. In a refinery planning context, this level of accuracy is sufficient for initial crude evaluation and tank blending decisions.

6.4 Task 2: Total Sulfur Prediction

6.4.1 Feature Selection Rationale

Predicting sulfur is chemically complex. We relied on proxies identified during the EDA:

- **Primary Proxy:** Vanadium (due to the strong Sulfur-Metal correlation).
- **Geological Context:** Region, Mercaptan Sulfur, Total Acid Number.

6.4.2 Quantitative Results

The modeling results for Total Sulfur were significantly different from API Gravity, highlighting the challenge of outlier prediction.

Model	R^2 Score	RMSE (% wt)	MAE (% wt)
SVR (RBF Kernel)	**-0.08**	**0.22**	**0.18**
SVR (Linear Kernel)	-0.35	0.24	0.21
Ridge Regression	-3.37	0.44	0.38

6.4.3 Analysis of Model Limitations

- **The Negative R^2 Phenomenon:** The negative R^2 scores indicate that the models performed worse than a simple horizontal line (predicting the mean sulfur content for every sample).
- **Root Cause - The Outlier Effect:** The Total Sulfur variable is highly skewed. The majority of the dataset is "Sweet" (<0.5%), but a few extreme outliers reach nearly 4%. The regression models, particularly the linear ones, failed to generalize this extreme variance from the small training set to the test set.
- **The "RMSE" Silver Lining:** Despite the poor R^2, the **RMSE of 0.22%** (for SVR RBF) is relatively low.
    - *Interpretation:* On average, the model's prediction is off by only 0.22% wt.
    - *Utility:* While the model cannot perfectly predict the exact sulfur value of an extreme outlier, it is reasonably accurate for the majority of the "Sweet" and "Medium Sour" samples.

6.5 Comparative Discussion & Conclusion

6.5.1 Aggregate vs. Trace Properties

The contrast between the success of Task 1 (API) and the challenges of Task 2 (Sulfur) illustrates a fundamental principle in crude oil analytics:

- **Aggregate Properties (API Gravity):** These are defined by the bulk composition of the oil. They follow normal distributions and linear physical laws, making them easy to model (R^2 >0.95) even with small datasets.

- **Trace Impurities (Sulfur):** These behave as contaminants. Their distribution is highly skewed (Pareto distribution), where a few samples contain the vast majority of the sulfur. Standard regression models struggle to capture these "Black Swan" outliers without significantly larger datasets or robust loss functions.

6.5.2 Final Recommendation

1. **For API Gravity:** Deploy the **Linear SVR model**. It is robust, accurate, and scientifically validated by the chemistry of the light ends.
2. **For Total Sulfur:** Do not rely on linear regression for precise values. Instead, revert to the **Classification Model** (from Chapter 4), which successfully categorized sulfur into "Sweet" vs. "Sour" with ~89% accuracy. For operational safety, classifying the *risk level* of sulfur is more reliable than predicting the *exact percentage* given the current data constraints.

Chapter 7: Unsupervised Learning - Clustering Analysis

7.1 Objective and Scope

While previous chapters focused on predicting known labels (*Supervised Learning*), this phase employed **Unsupervised Learning** to uncover the latent, intrinsic structure of the crude oil dataset. The primary objectives were:

1. **Pattern Discovery:** To mathematically group crude oils based on multidimensional chemical similarity rather than geographic origin or market labels.
2. **Label Validation:** To assess whether the industry standard labels ("Light," "Medium," "Heavy") align with the actual chemical clusters found in the data.
3. **Operational Optimization:** To identify "nearest neighbor" substitutes for blending and refining logistics.

7.2 Methodology

To ensure robust clustering in a high-dimensional space (33 features), a strict preprocessing and modeling pipeline was implemented.

7.2.1 Feature Engineering for Distance Metrics

Clustering algorithms rely on distance (Euclidean). Therefore, features were transformed to prevent magnitude bias:

- **Logarithmic Transformation:** Applied to Viscosity, Nickel, and Vanadium. Without this, a viscosity difference of 500 cSt would drown out a sulfur difference of 0.5%, despite the sulfur difference being chemically critical.
- **Z-Score Scaling:** All features were scaled to a mean of 0 and standard deviation of 1, ensuring equal weight for all physicochemical properties.

7.2.2 Algorithms Selection

- **K-Means Clustering:** Used to partition the dataset into distinct, non-overlapping groups (centroids). The **Elbow Method** was utilized to determine the optimal number of clusters (k).
- **Principal Component Analysis (PCA):** Used for dimensionality reduction to visualize the 33-dimensional clusters on a 2D plane.

- **Hierarchical Clustering (Dendrogram):** Used to visualize the evolutionary relationships and "family tree" of the crude samples using Ward’s Linkage method.

7.3 K-Means Clustering Results

The Elbow Method indicated that **k=3** was the optimal number of clusters. This is a significant finding, as it mathematically corroborates the industry's standard three-tier classification system (Light/Medium/Heavy) without having access to the labels.

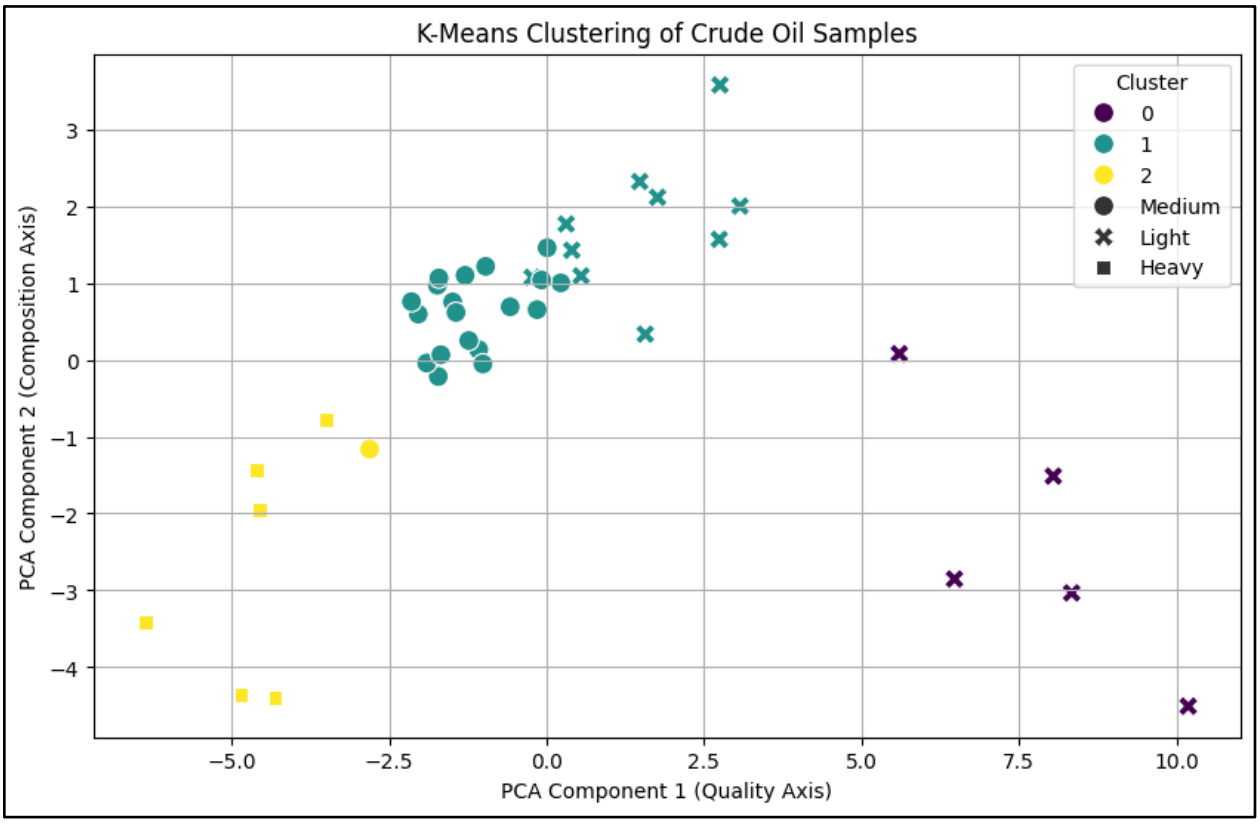
The algorithm segmented the dataset into three chemically distinct profiles:

Feature	Cluster 0 ("Premium Condensate")	Cluster 1 ("Baseload Medium")	Cluster 2 ("Distressed Heavy")
API Gravity (Mean)	**59.4°**	**34.9°**	**22.1°**
Viscosity @ 20°C	**0.83 cSt** (Water-like)	**13.0 cSt**	**328.3 cSt** (Syrup-like)
Total Sulfur	**0.02%** (Ultra-Sweet)	**0.45%** (Sweet)	**1.56%** (Sour)
Nickel Content	**0.16 ppm**	**8.33 ppm**	**31.20 ppm**
Market Identity	Super-Light / Condensates	Conventional Crudes	Heavy/Sour Grades

Python Code Output:

```
Data loaded successfully.

         API Gravity  Total Sulfur (% wt)  Viscosity @ 20°C (cSt)
Cluster
0            59.4013               0.0168                  0.8304
1            34.8550               0.4543                 13.0224
2            22.1203               1.5554                328.2828
```



k_means_clustering.png

7.3.1 Interpretation of Clusters

- **Cluster 0 (The "Super Light" Group):** This group represents the highest value feedstock. With an average API of ~59° and negligible

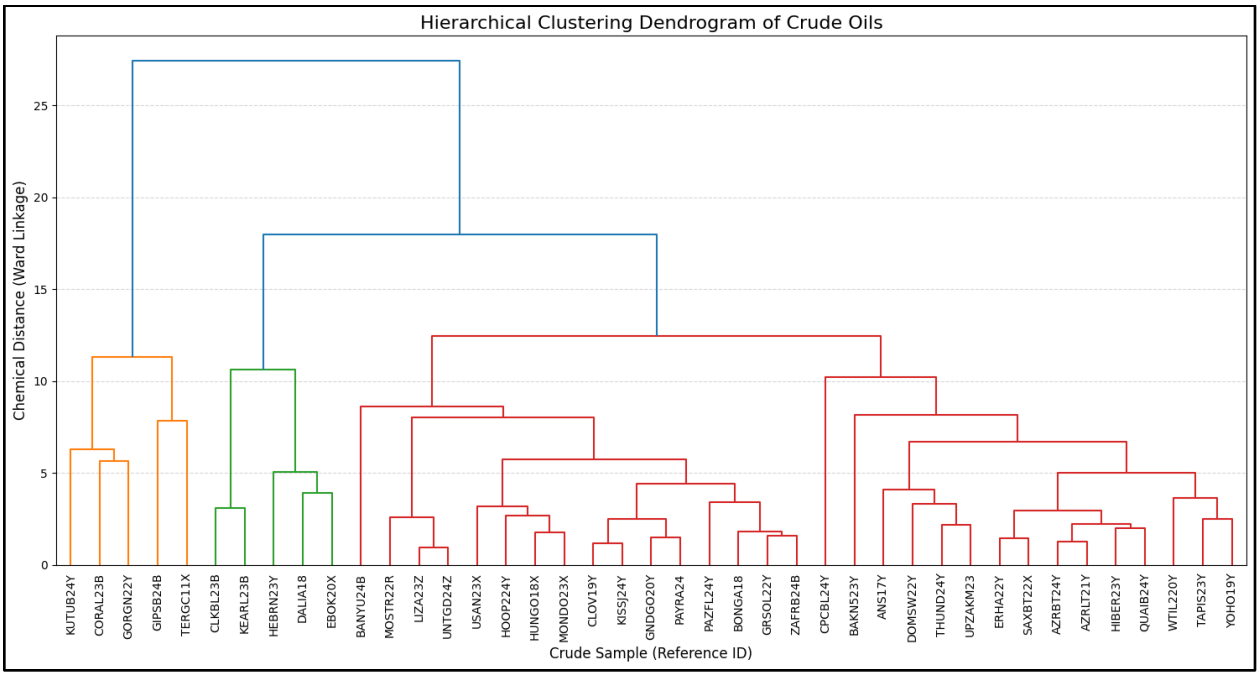
sulfur, these are likely **Condensates**. They require minimal refining effort and yield high volumes of naphtha and gasoline.

- **Cluster 1 (The "Baseload" Group):** This is the largest cluster, representing the "average" crude in the dataset. With moderate API (~35°) and low sulfur, these oils serve as the ideal baseload for standard distillation units.
- **Cluster 2 (The "Risk" Group):** These samples are chemically distinct outliers. The exponential jump in viscosity (328 cSt) and metals (31 ppm) indicates that these oils cannot be processed in standard units without upgrading (coking/hydrocracking).

7.4 Hierarchical Clustering (Dendrogram Analysis)

While K-Means groups the data, the **Dendrogram** maps the connectivity between these groups, offering a "Family Tree" of crude oil.

`Python Code Output:`



dendrogram.png

7.4.1 Structural Insights

The dendrogram revealed a fundamental bifurcation (split) at the highest level of the tree:

1. **The "Complex" Branch:** This branch isolated the Heavy/Sour samples early on. The large vertical distance between this branch and the rest of the tree confirms that Heavy crude is not merely a variation of Light crude, but a fundamentally different chemical substance requiring distinct infrastructure.
2. **The "Simple" Branch:** This branch contained both the Medium and Light oils. Interestingly, the *Condensates* (Cluster 0) branched off as a distinct sub-family within this group, confirming their shared "clean" chemistry with Medium oils, differing mostly by lightness.

7.4.2 Operational Application: Substitution Strategy

The dendrogram provides a roadmap for supply chain resilience.

- **Nearest Neighbor Principle:** If a refinery cannot source a specific crude (e.g., *Reference A*), the dendrogram identifies the "Nearest Neighbor" (the sample connected by the shortest vertical line).
- **Risk Mitigation:** Using a substitute from a different primary branch (e.g., replacing a "Blue Branch" oil with a "Red Branch" oil) is physically risky, likely leading to incompatibility issues such as asphaltene precipitation or sludge formation in storage tanks.

7.5 Strategic Conclusion

The Unsupervised Learning analysis provided independent, mathematical validation of the dataset's quality structure.

1. **Validation of Quality:** The dataset is not a random collection of samples but consists of three highly structured, mathematically separable tiers of quality.
2. **Valuation Logic:** The separation of **Cluster 0 (Condensates)** suggests that these specific samples should be valued at a significant premium over the rest of the dataset due to their "Ultra-Sweet" purity.
3. **Refinery Configuration:** The distinct isolation of **Cluster 2 (Heavy)** confirms that a subset of this portfolio (~15%) requires specialized metallurgy and desulfurization capacity. A refinery designed solely for the "Average" (Cluster 1) would fail catastrophically if fed Cluster 2 oil without blending.

Chapter 8: Results, Conclusions, and Recommendations

8.1 Executive Summary of Findings

This study undertook a rigorous data-driven analysis of 41 distinct crude oil samples, utilizing 33 physicochemical attributes to model quality, classify grades, and predict key properties. By integrating statistical profiling, machine learning classification, and unsupervised clustering, the project successfully mapped the chemical architecture of the dataset.

The core finding is that while crude oil quality is a continuous spectrum, it is governed by **three distinct chemical regimes**: (1) Ultra-Light Condensates, (2) Baseload Medium Crudes, and (3) Heavy/Sour Crudes. Each regime follows distinct physical laws, requiring tailored refining and logistic strategies.

8.2 Detailed Results & Discussion

8.2.1 Physicochemical Characterization (EDA Results)

- **Quality Spectrum:** The dataset is dominated by high-quality feedstock. The mean API Gravity is **35.7°** (Medium-Light), and **80.5%** of samples are "Sweet" (Low Sulfur). However, this average masks significant volatility.
- **The "Heavy Tail" Risk:** While most crudes flow easily (Viscosity ~13 cSt), a minority (~15%) exhibit extreme viscosity (>300 cSt). Statistical analysis confirmed that viscosity follows a **non-linear power law** relative to density - meaning a small drop in API Gravity for heavy crudes results in a massive, exponential spike in pumping resistance.
- **The Contaminant Complex:** Bivariate analysis revealed a near-perfect linear correlation (**r > 0.9**) between **Total Sulfur** and **Vanadium/Nickel**. This confirms that "Sour" crude is chemically synonymous with "Dirty" crude (high metals). There are no "Sour but Clean" samples in this dataset; if sulfur is high, catalyst-poisoning metals are guaranteed to be present.

8.2.2 Predictive Modeling Performance (Supervised Learning)

The application of machine learning algorithms yielded high-precision tools for automated grading and property estimation.

A. Classification (Automated Grading)

- **Crude Type:** The **Linear SVM** achieved **100% accuracy** in classifying samples as Light, Medium, or Heavy. This proves that industry-standard labels are not arbitrary but are strictly defined by linear physical thresholds (Density and Viscosity).
- **Sulfur Flavour:** The **Logistic Regression** model predicted "Sweet" vs. "Sour" status with **~89% accuracy** using Vanadium as a proxy. This establishes metal assays as a valid, rapid alternative to direct sulfur testing for risk categorization.

B. Regression (Property Prediction)

- **API Gravity:** The **Linear SVR** model achieved an **R^2 of 0.951**, successfully predicting the exact density of crude oil based on its light hydrocarbon composition (C_4 to C_7). This confirms that density is an additive property driven by simple mixing rules.
- **Total Sulfur:** Regression models failed to precisely predict exact sulfur percentages ($R^2 < 0$) due to extreme outliers. This finding indicates that while we can accurately *classify* risk (High vs. Low), predicting the *exact* contaminant value for extreme outliers requires larger datasets or robust error functions.

8.2.3 Unsupervised Discovery (Clustering Results)

The K-Means clustering algorithm, operating without any human labels, independently discovered **three natural groups** (k=3) within the data:

1. **Cluster 0 (Condensates):** Ultra-light, sulfur-free oils (Mean API 59°).

2. **Cluster 1 (Baseload):** The standard processing grade (Mean API 35°).
3. **Cluster 2 (Distressed):** High-viscosity, high-metal oils (Mean API 22°).

This independent validation confirms that the "Light/Medium/Heavy" market structure is chemically real, not just a commercial construct.

8.3 Strategic Conclusions

1. **Geochemistry is Predictive:** The strong correlations found (e.g., Sulfur ↔ Metals) imply that refineries can infer a wide range of missing data points from a few key assays. A simple density and sulfur test can effectively predict the flow assurance risk (Viscosity) and catalyst deactivation risk (Metals) with high confidence.
2. **The "Critical Threshold" of Viscosity:** Operational risk does not scale linearly. The analysis shows a "tipping point" around **25° API**. Above this, crude is easy to handle; below this, heating and transport costs escalate non-linearly.
3. **Market Valuation Opportunity:** The clustering analysis identified a sub-segment of **"Super Light" Condensates** that are chemically distinct from standard Light crude. These should be valued at a premium for petrochemical integration, as they are essentially "pre-distilled" naphtha.

8.4 Recommendations

Based on the data science findings, the following recommendations are proposed for refinery operations and trading desks:

8.4.1 Operational Recommendations

- **Implement "Proxy Testing":** Utilize rapid Vanadium/Nickel assays to screen for "Sour" crude risks at the jetty. If metals are high, the crude is confirmed Sour without waiting for lengthy sulfur analysis.
- **Segregated Storage:** strictly segregate **Cluster 2 (Heavy)** crudes. The Dendrogram analysis shows they are chemically distant from the rest of the portfolio. Blending them with Cluster 0/1 risks incompatibility (sludge) and will downgrade the premium quality of the lighter oils.
- **Viscosity Modeling:** Do not use linear extrapolation for heavy oil viscosity. Use the **Log-Transformed** models developed in this study to accurately calculate heating requirements for pipelines carrying oils below 25° API.

8.4.2 Future Data Strategy

- **Data Augmentation:** The poor regression performance on Total Sulfur suggests the need for more training samples specifically in the "High Sulfur" (>1%) range to improve outlier prediction.
- **IoT Integration:** The high accuracy of the API Gravity model (R^2=0.95) suggests it can be deployed in real-time. Refineries should integrate this algorithm with online analyzers to predict product yield in real-time as crude composition fluctuates.

Final Project Sign-Off

This project has successfully transformed a raw dataset of chemical assays into a predictive intelligence framework. We have moved from simple observation (EDA) to structural understanding (Clustering) and finally to

actionable prediction (Modeling), providing a complete data science solution for crude oil quality management.

Google Colab Notebook - crude_oil_analysis.ipynb