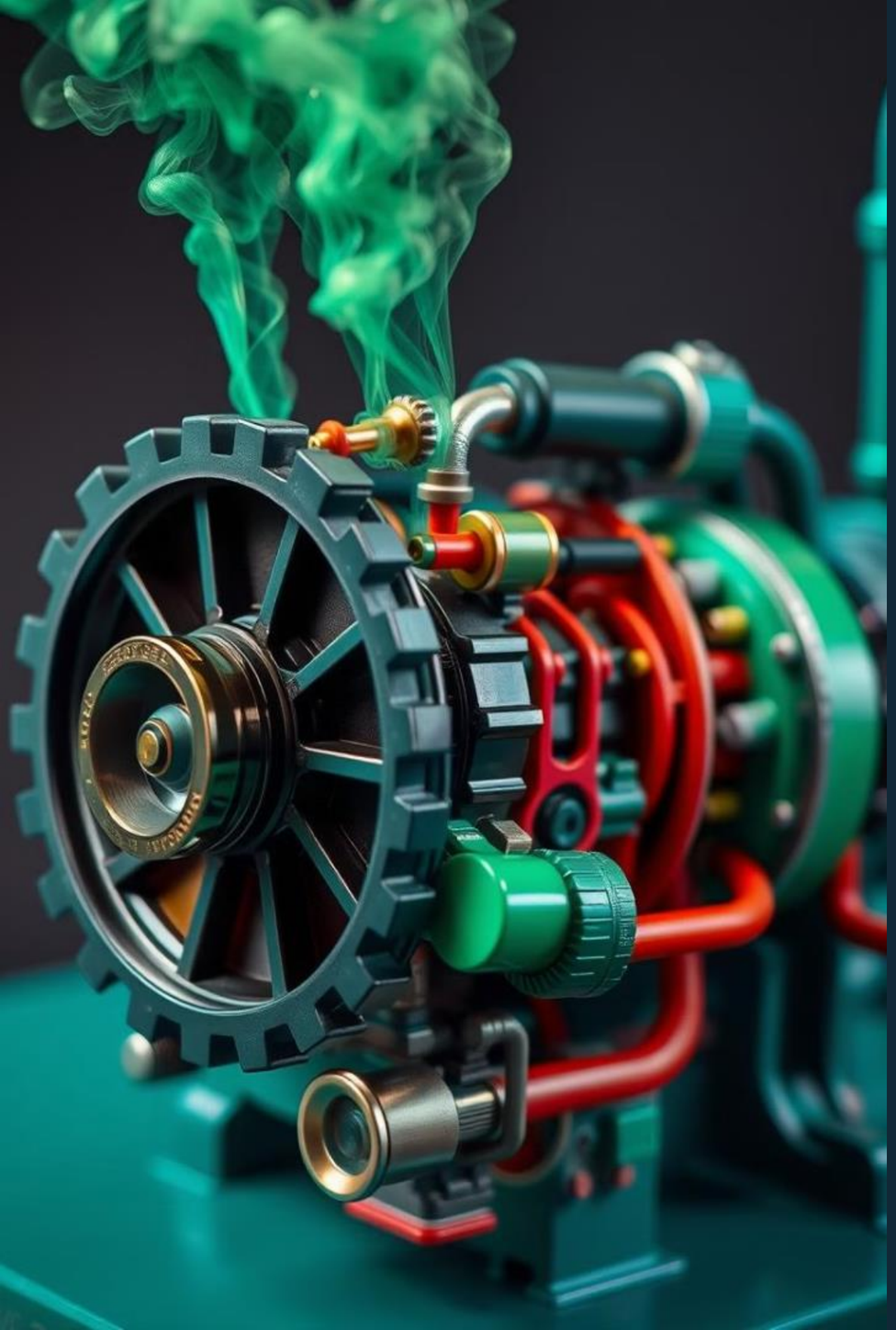


Gas Turbine Emissions Analysis: Uncovering Insights and Optimizing Performance

This presentation delves into a comprehensive statistical analysis of gas turbine emissions data, aiming to uncover key relationships and patterns that can inform operational improvements and environmental compliance. Through techniques like ANOVA, regression, and correlation, we explore the factors influencing emissions and provide insights for optimizing turbine performance.

Presented by

Yosr Charrada

Mariem Mojaat

Tessnim Etteib

Nour Amouri

Yasmine Bahri

Maram Sliti

Business Understanding: The Quest for Efficiency

Objective

Identify significant factors influencing gas turbine emissions to optimize processes and predict potential failures. This analysis leverages data from a turbine in Turkey, focusing on CO and NOx emissions.

Dataset

The dataset investigates CO and NOx emissions under varying operational conditions, providing a detailed picture of the turbine's behavior. Understanding these emissions is crucial for improving energy efficiency and complying with environmental regulations.

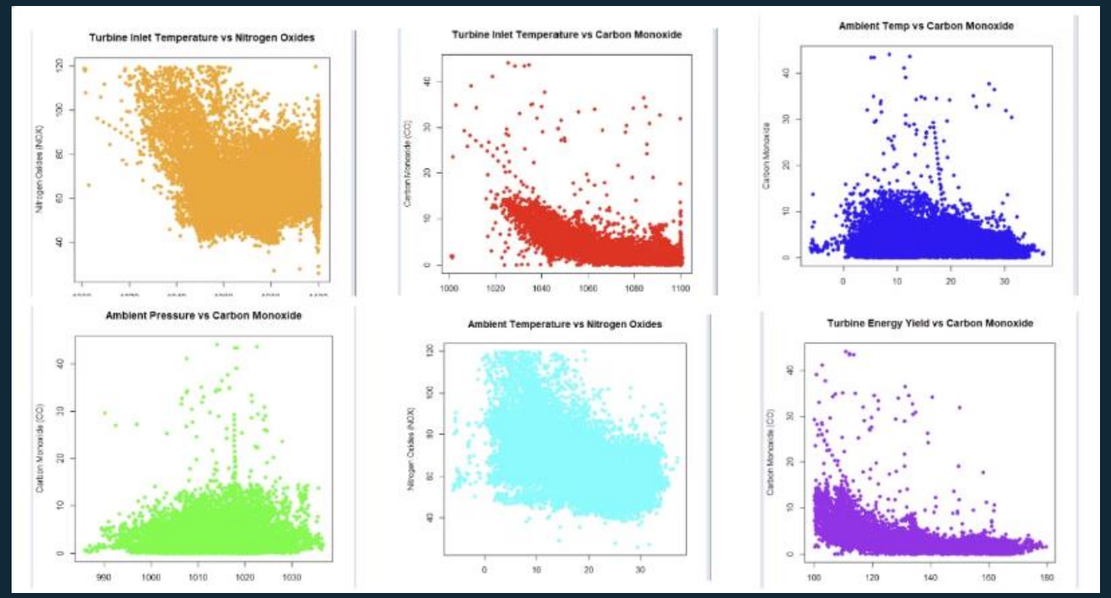
Data Understanding: Unveiling the Dataset's Secrets

Dataset Overview

Sourced from the UCI Machine Learning Repository, the dataset spans from 2011 to 2015. It includes 11 features representing various operational parameters and emission levels. This dataset is invaluable for predicting CO and NOx emissions, crucial for evaluating environmental impact and operational efficiency.

Data Merged

The merged dataset consolidates data from all years into a single comprehensive file, enabling a holistic analysis of trends and patterns. This integration simplifies comparisons across years, enhancing the accuracy and reliability of insights derived from the analysis.



Data Exploration: Unmasking Hidden Relations

- **Negative Correlations**: Higher turbine inlet temperatures and energy yields are associated with lower CO and NOx emissions.
- **Positive Correlations**: Higher ambient temperatures and pressures are linked to higher NOx and CO emissions, respectively.
- **No Strong Correlation**: Ambient temperature does not seem to significantly affect CO emissions.

Data Exploration: Unmasking Hidden Patterns

Data Types

All variables in the dataset are quantitative, indicating measurable values. This is ideal for applying statistical methods like regression, correlation, and ANOVA to uncover meaningful patterns and relationships.

Missing Values

The analysis revealed no missing data points in the dataset. This ensures that the analysis proceeds without needing to handle incomplete records, maintaining the integrity and completeness of the data throughout the statistical processes.

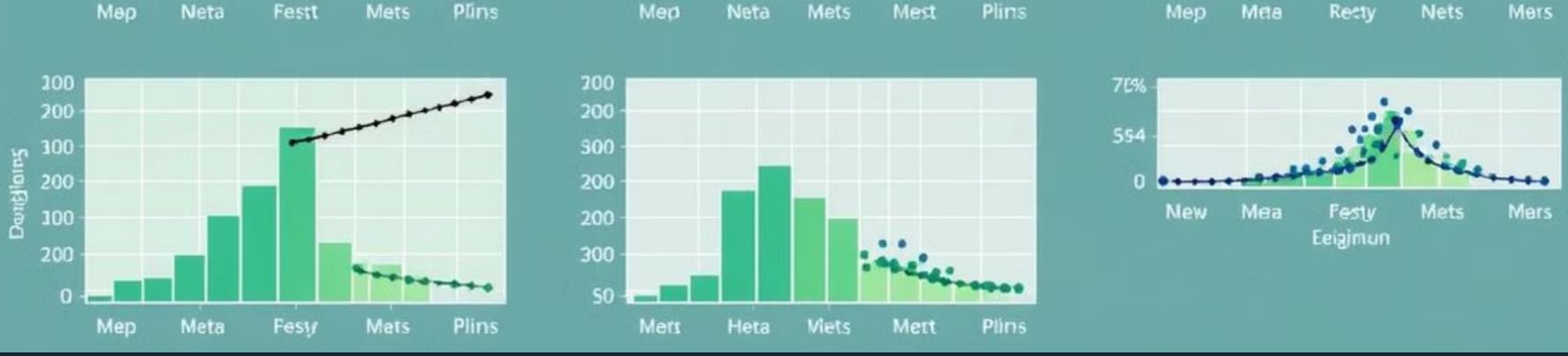
Data Exploration: Diving Deeper

Count of Unique Values

Variables like AT, CO, and NOx have the highest counts of unique values, suggesting significant variability in these measurements across the dataset. Columns like AP and TIT have fewer unique values, indicating more consistent or clustered measurements.

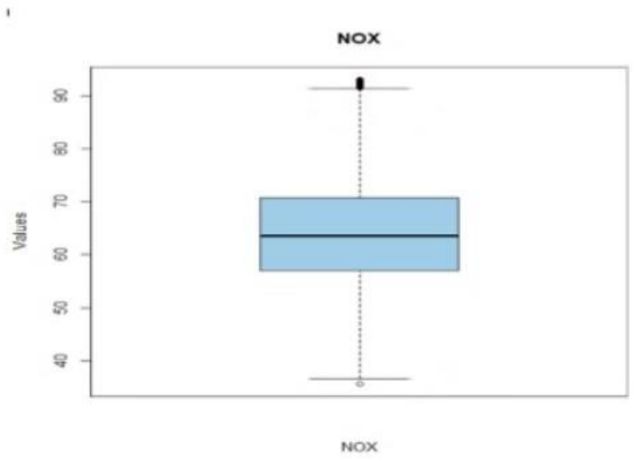
Duplicate Rows

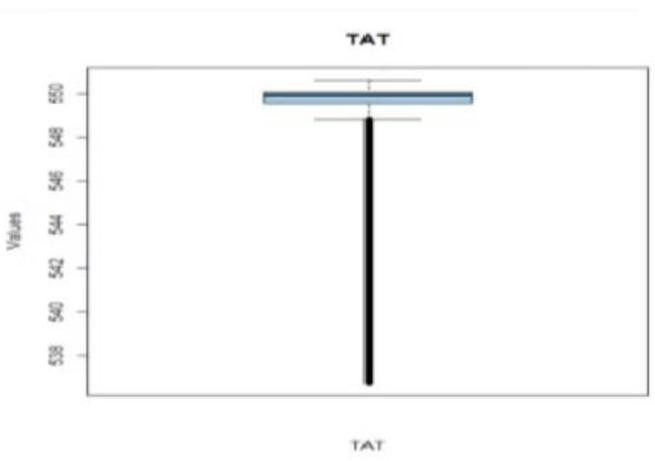
The dataset contains 7 duplicate rows out of a total of 36,733 instances. This represents a small fraction of the data (less than 0.02%).

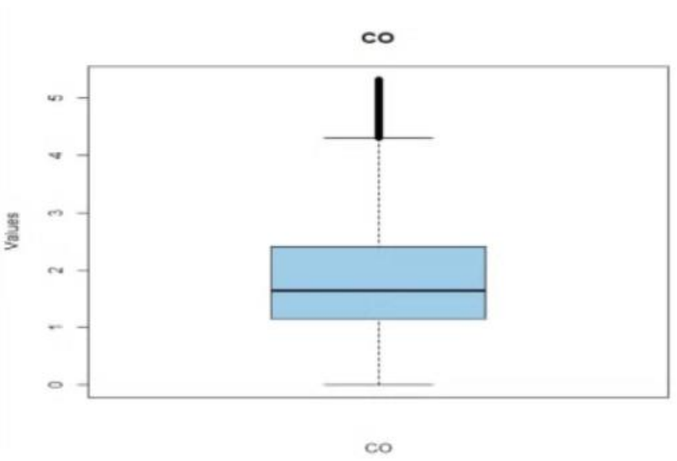


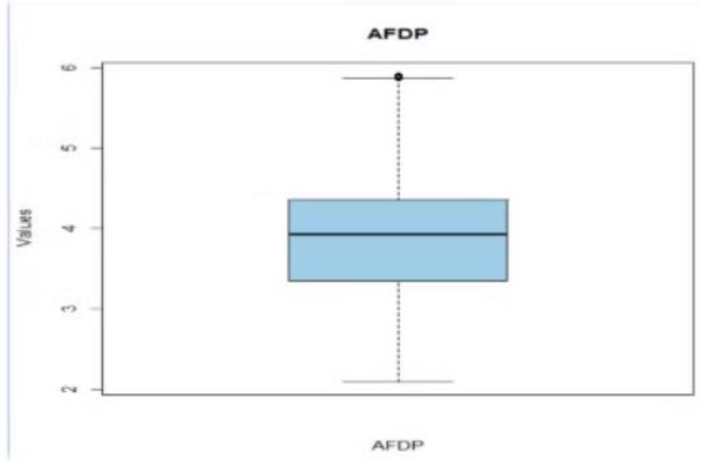
Outlier Detection: Identifying Anomalies

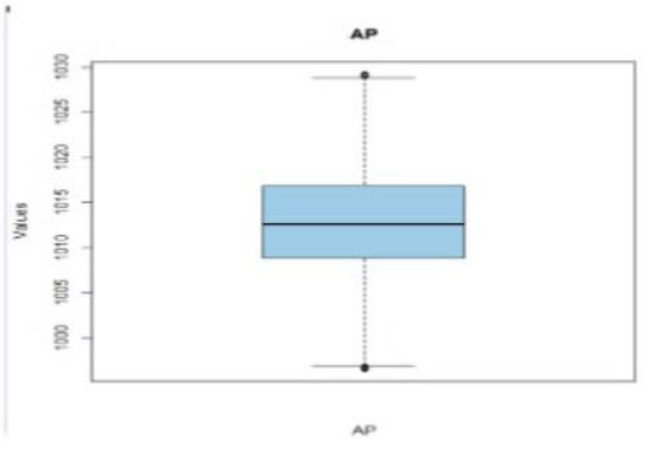
Boxplots reveal a mix of symmetric distributions and notable outliers across several variables. While AT and GTEP show stable distributions, variables like AP, AH, AFDP, TIT, TAT, TEY, CO, and NOX have outliers. These outliers may indicate rare events or errors and should be further investigated.

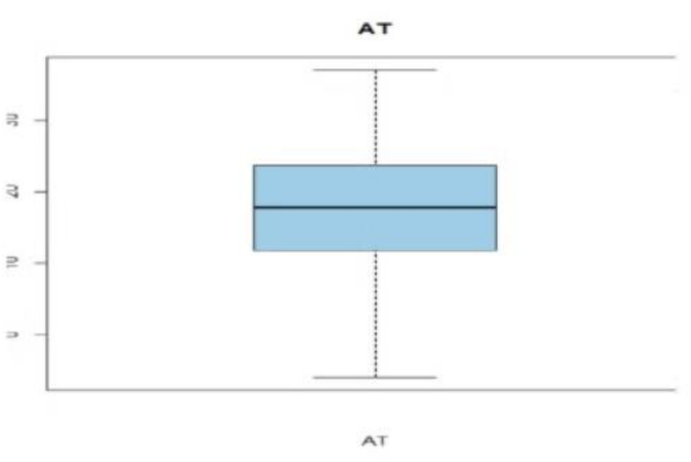


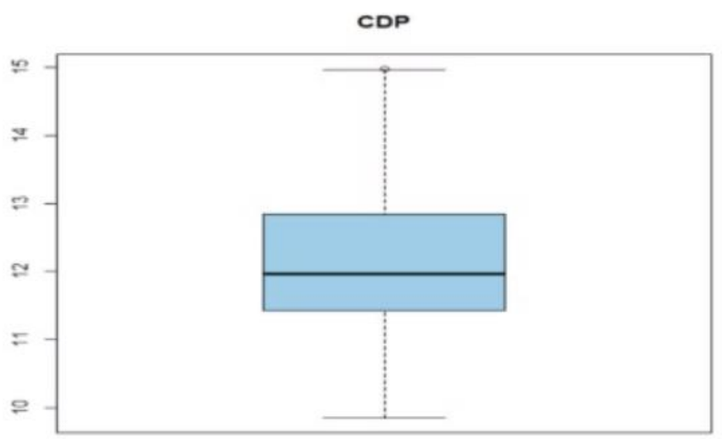


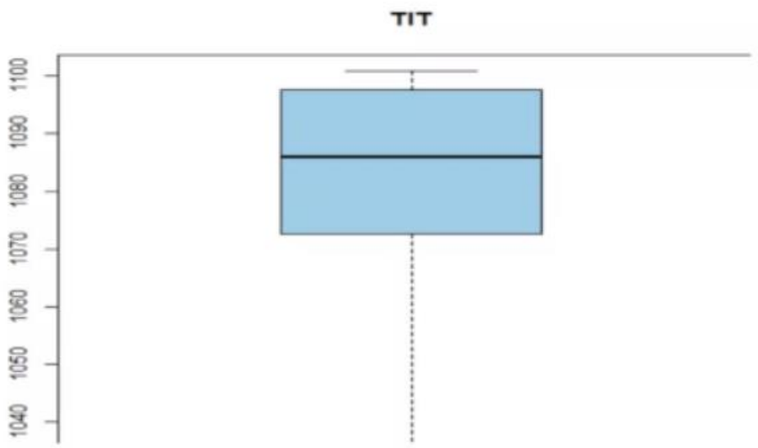


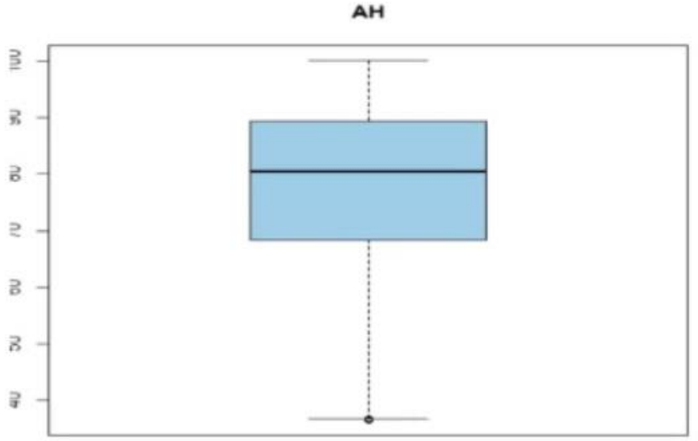


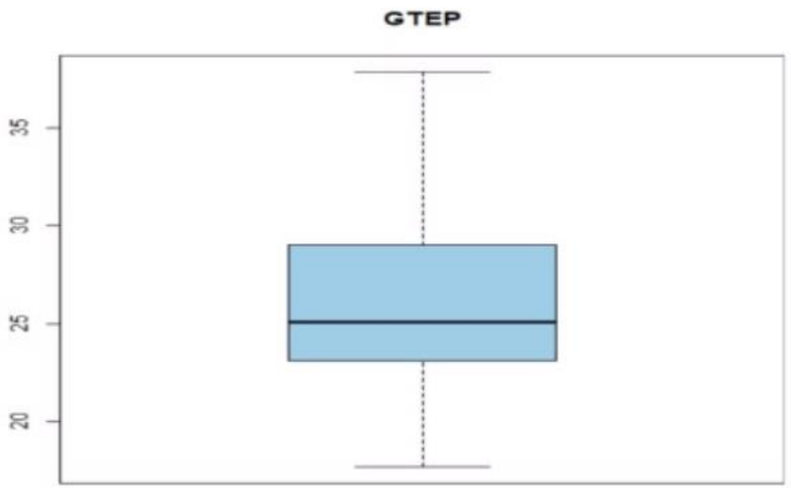


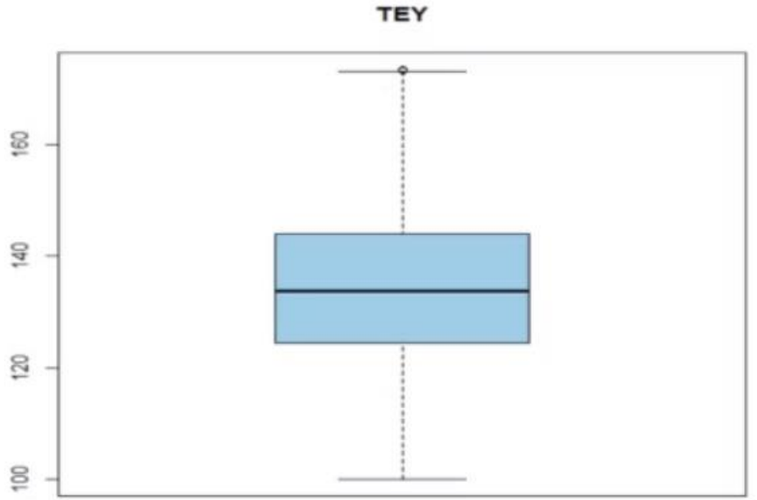


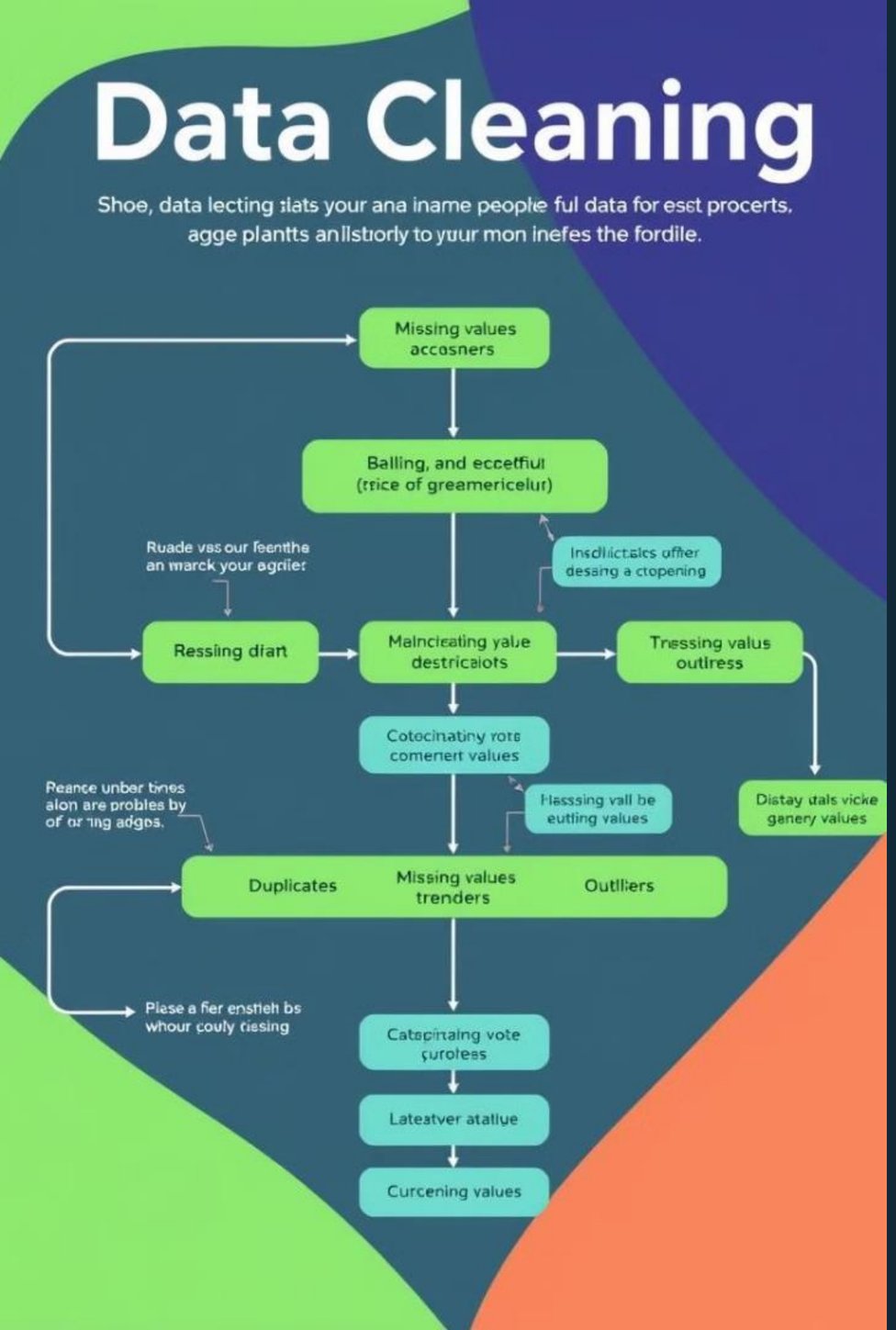












Data Cleaning: Ensuring Data Quality

1 Treating Outliers

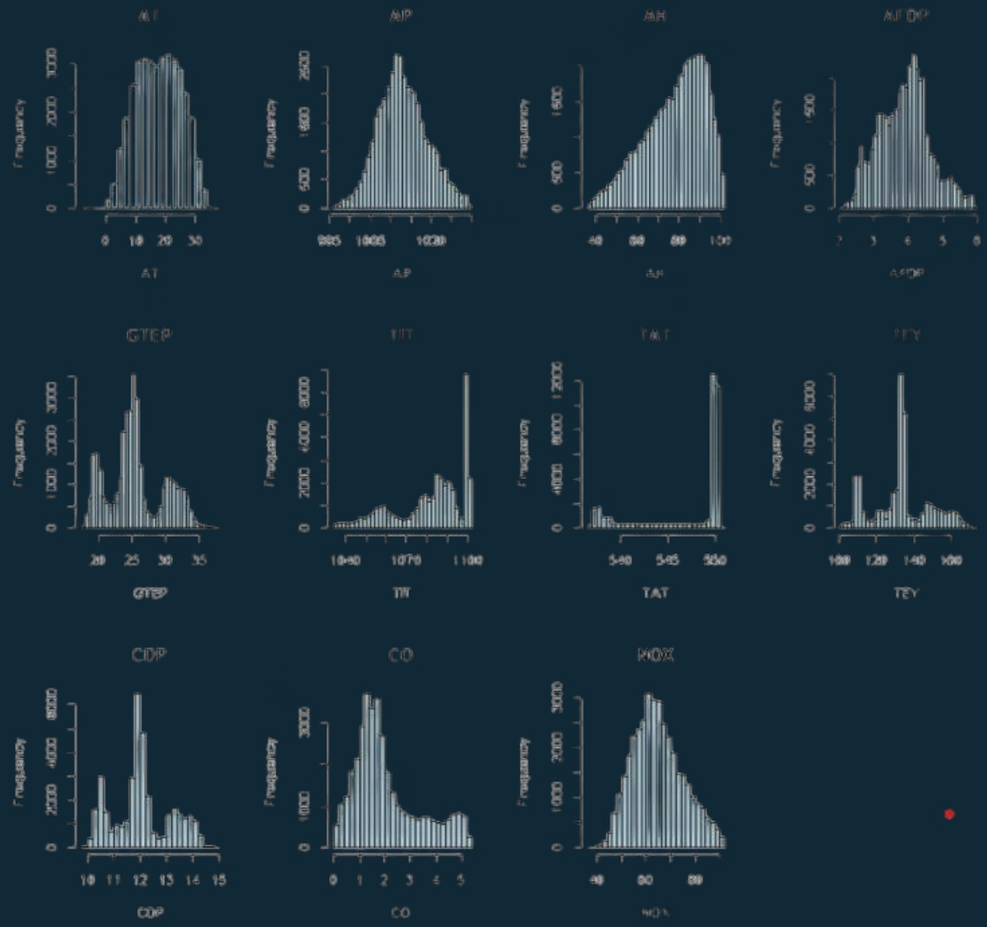
The data cleaning phase focused on addressing potential issues like missing values, duplicates, and outliers. We employed K-Nearest Neighbors (KNN) and median imputation to treat the outliers, ultimately choosing KNN as the more effective method for this dataset.

2 Scaling

Normalization was applied to the data, scaling the quantitative values to a consistent range between 0 and 1. This step prevents variables with larger ranges from disproportionately influencing the analysis, ensuring more accurate and reliable results.

Testing Normality: Assessing Data Distribution

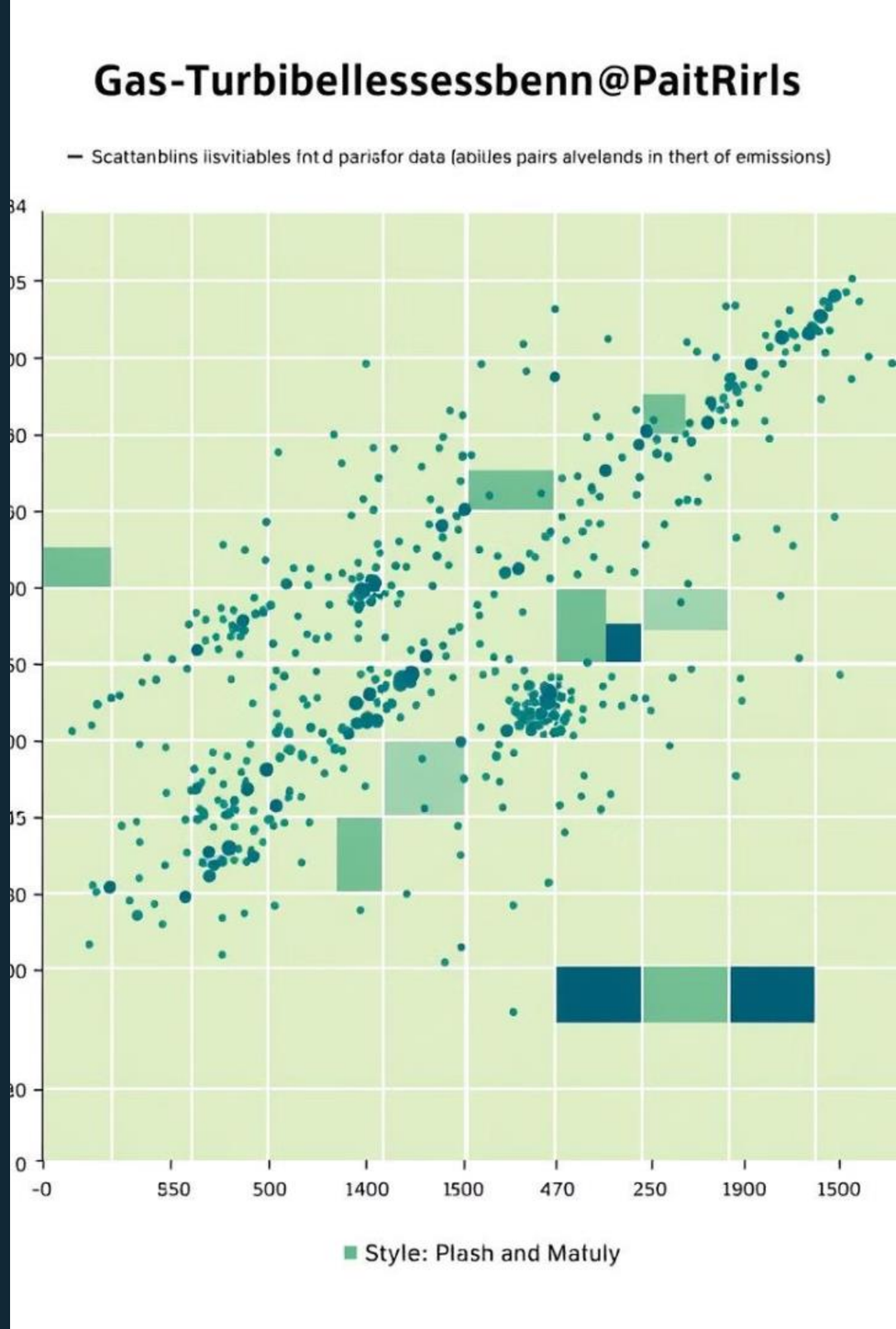
We used kurtosis, skewness, and histograms to evaluate the distribution of the variables. AT, AP, and NOX followed approximately normal distributions. However, variables like AH, GTEP, TIT, TAT, AFDP, TEY, CDP, and CO exhibited non-normal distributions. While some variables can be analyzed with standard parametric methods, others may require transformation or special handling.



Bivariate Analysis: Exploring Relationships

Bivariate analysis involved examining the relationship between two variables. We used t-tests to compare the means of two groups and the variance test (var-test) to assess if the variances between two groups are significantly different. These tests identified significant differences between groups and provided insights into the strength and nature of their relationshKey Takeaways & Next Steps

The analysis revealed significant relationships between various operational parameters and gas turbine emissions. Key takeaways include the significant influence of AT on NOX, the effectiveness of the KNN method for outlier treatment, and the need for further investigation into the non-normal distribution of some variables. Future steps involve developing predictive models based on the insights gained, further exploring the impact of outliers, and refining the analysis to incorporate additional data sources and variables.



Variance Test Results (F-tests)

Comparison	Variance Difference	p-value	Conclusion
NOX vs. AT	Significant (NOX variance slightly higher)	<2.2e-16	H1 chosen (p<0.5)
NOX vs. AP	No significant difference	0.1909	H1 chosen (p<0.5)
AT vs. AP	Significant (AT variance smaller)	<2.2e-16	H1 chosen (p<0.5)

Overall Conclusion: Significant variance differences found between NOX/AT and AT/AP, but not NOX/AP.

t-test Results

NOX and AT

There is a significant difference in means. AT has a higher mean (0.5526) than NOX (0.5034). p-value < 2.2e-16, indicating strong evidence against the null hypothesis.

NOX and AP

There is no significant difference in means. The means of NOX (0.5034) and AP (0.5036) are nearly identical. p-value = 0.9081, failing to reject the null hypothesis.

AT and AP

There is a significant difference in means. AT has a higher mean (0.5526) than AP (0.5036). p-value < 2.2e-16, indicating strong evidence against the null hypothesis.

Kruskal-Wallis Test Results

TIT Groups

Significant differences in GTEP, TAT, TEY, and CDP ($p < 0.05$). Reject H0.

AFDP Groups

Significant differences in GTEP, TAT, TEY, and CDP ($p < 0.05$). Reject H0.

TAT Groups

Significant differences in TEY and CDP ($p < 0.05$). Reject H0.

GTEP Groups

Significant differences in TAT, TEY, and CDP ($p < 0.05$). Reject H0.

AH Groups

Significant differences in GTEP and TAT ($p < 0.05$). No differences in AFDP, TEY, or CDP ($p > 0.05$). Reject H0 for GTEP and TAT.

TEY Groups

Significant difference in CDP ($p < 0.05$). Reject H0.

Linear Regression

Key Coefficients

Intercept: 0.8835 (baseline NOX when AP and AT are 0).

AP Coefficient (0.0414): A one-unit increase in AP increases NOX by 0.0414 ($p<2e-16$).

AT Coefficient (-0.8363): A one-unit increase in AT decreases NOX by 0.8363 ($p<2e-16$).

Model Fit & Conclusion

Model Fit (R-squared = 0.324): AP and AT explain 32.4% of NOX variability. The model is highly significant (F-statistic = 8808, $p<2.2e-16$).

Conclusion: While the model explains a moderate amount of NOX variability, AP and AT are highly significant predictors. AT's negative impact on NOX is considerably stronger than AP's. Further improvements, such as adding variables or testing interactions, could enhance the model.

Strategy to Improve the Model:

The current model's limited explanatory power (R-squared = 0.324) suggests that additional variables could significantly improve its predictive accuracy. Consider incorporating qualitative variables, to assess their impact through ANOVA. But since we don't have such we'll stop here .

Dimensional Reduction

NOX & TEY

Weak correlation (ρ = 0.0198), statistically significant (p < 0.0001). Action: Drop TEY due to weak correlation.

NOX & CDP

Weak negative correlation (ρ = -0.087), statistically significant (p < 2.2e-16). Action: Drop CDP due to weak correlation.

NOX & AH

Weak but stronger correlation (ρ = 0.1438), statistically significant (p < 2.2e-16). Action: Retain AH.

After reduction, the dataset comprises 36,733 rows and 9 columns, streamlining features while retaining key information.

Conclusion: Variables with weak correlations were dropped.

Conclusion

This analysis revealed key insights into the dataset's characteristics and relationships. Data cleaning ensured quality, and normalization provided consistent scaling. Bivariate analyses (t-tests, variance tests, Kruskal-Wallis) uncovered significant relationships. Dimensionality reduction simplified the dataset while preserving crucial information. Finally, linear regression highlighted the significant impact of **AP** and **AT** on **NOX**, explaining a substantial portion of its variability.

This report demonstrates the effective use of statistical methods for data understanding, cleaning, and analysis, providing valuable results for informed decision-making.