2025

Exploratory Data Analysis (EDA) Summary Report

ALVIN PHAN

1. Dataset Overview

- **Dataset Name:** Combined Cycle Power Plant
- **Total Rows:** 9568 (before cleaning) → **9527 (after cleaning)**
- **Total Columns:** 5
- **Target Variable (Output to Predict):** PE (Power Output)
- **Predictor Variables (Features):**
  - AT (Ambient Temperature)
  - V (Exhaust Vacuum)
  - AP (Ambient Pressure)
  - RH (Relative Humidity)

---

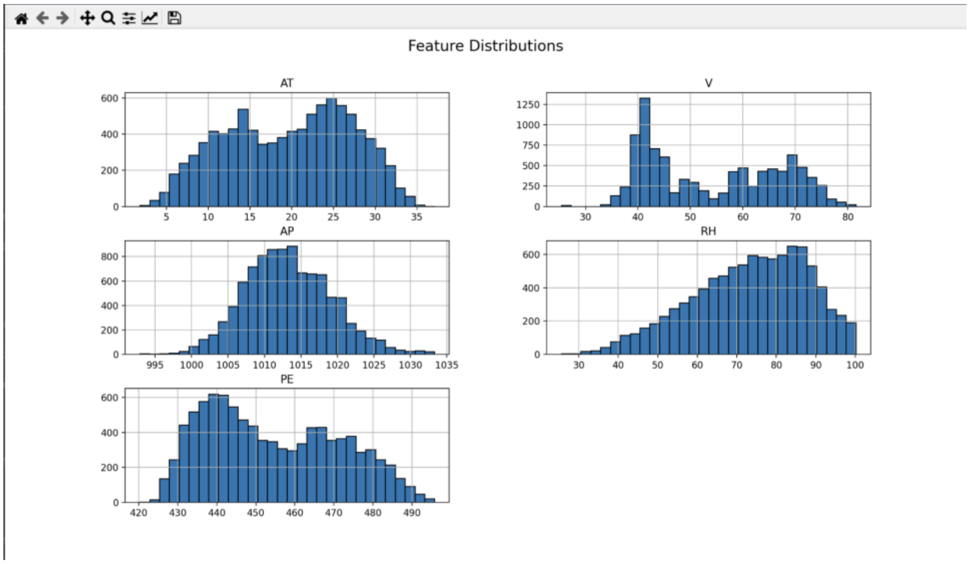
2. Data Cleaning Summary

- **Duplicates Removed:** 41 duplicate rows were dropped.
- **Missing Values:** No missing values detected in any column.
- **Data Types:** All features are numerical (float64).
- **Column Name Formatting:** Stripped spaces to ensure consistency.

---

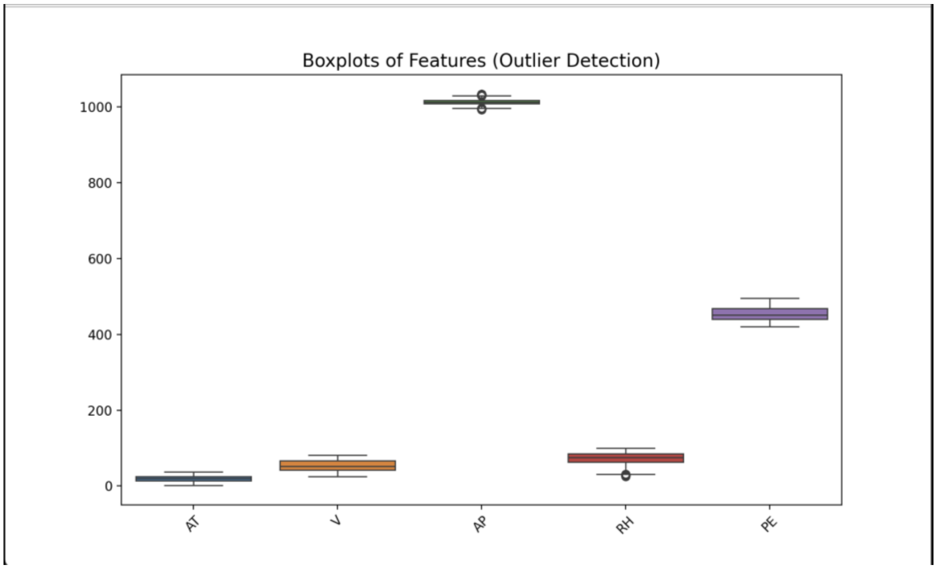
3. Univariate Analysis (Feature Distributions)

- **Histograms:**



  - AT and RH show a normal-like distribution.
  - V is left-skewed, meaning lower values are more frequent.
  - AP has a slight right skew.
  - PE (Power Output) is approximately normally distributed.

- **Boxplots:**

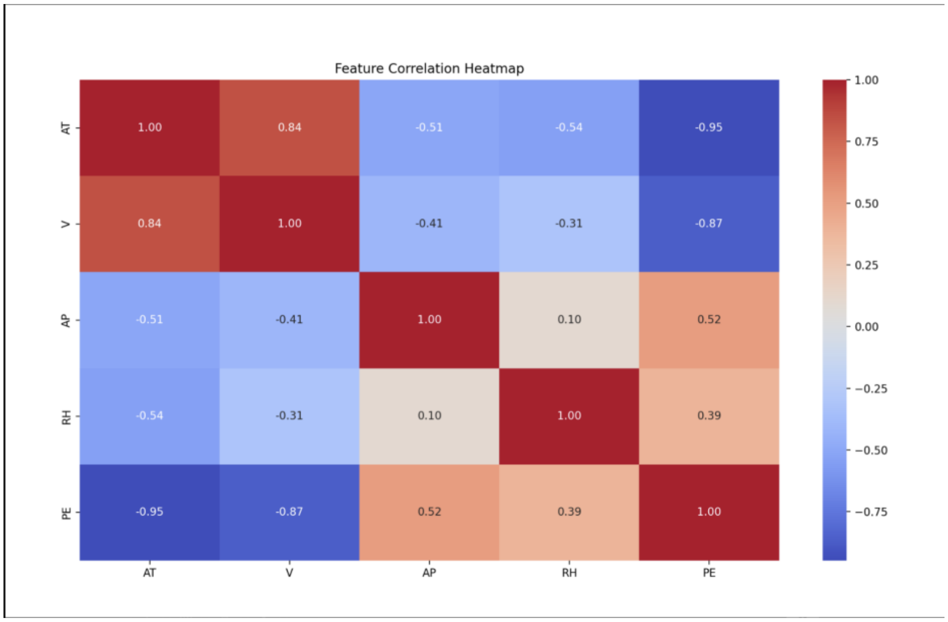


  - Possible outliers detected in V (Vacuum) and RH (Humidity).
  - Other features appear within normal ranges.

---

4. Multivariate Analysis (Feature Relationships)

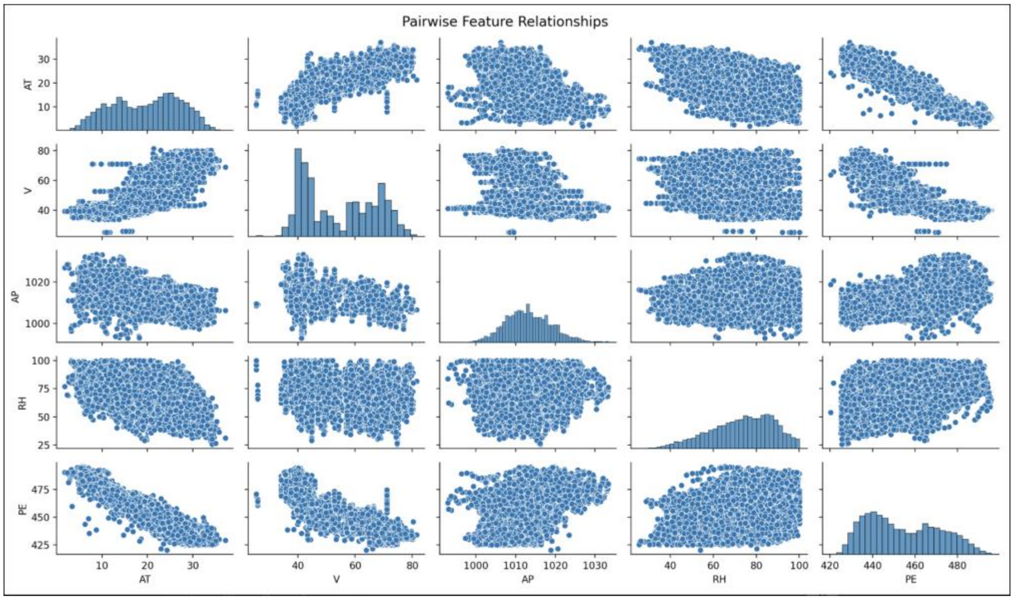
- **Correlation Heatmap Findings:**



  - PE (Power Output) has a **strong negative correlation with AT (-0.95)**.
  - PE also negatively correlates with V (-0.87) but positively correlates with AP (0.52) and RH (0.39).

    - AT and V are strongly correlated (0.84), meaning one might be redundant in a model.

- **Pairplot Analysis:**



    - Linear relationships are visible between PE and features like AT and V.
    - AP and RH show weaker correlations with PE.

Conclusion

The Exploratory Data Analysis (EDA) of the Combined Cycle Power Plant dataset has provided valuable insights into the relationships between variables. We identified that **Ambient Temperature (AT) has the strongest negative correlation with Power Output (PE), making it the most significant predictor.** Exhaust Vacuum (V) also negatively correlates with PE, while Ambient Pressure (AP) and Relative Humidity (RH) show positive but weaker correlations.

Additionally, we found potential outliers in the V and RH features, which may require further investigation in modeling. The dataset is **clean, well-structured, and ready for machine learning applications.**