FH | JOANNEUM
Electronic Engineering

Data Analysis Project Report

STM-WS2025

Project Assignment

Supervised By: Dr. Manu Gupta

Prepared by: Adeel Akhtar, Rabeet Tahir

Table of Contents

Contents

1. Contributions:

Adeel Akhtar:

- Data Preprocessing and Data Quality
- Visualization and Exploratory Analysis
- Probability and Event Analysis
- Slides Preparation

Rabeet Tahir:

- Statistical Theory Applications
- Regression and Predictive Modeling
- Dimensionality Reduction
- Report Preparation

2. Dataset:

- Electric Power Consumption.

Link To Dataset:

- Kaggle (**power consumption**)

Content Of Dataset:

The data consists of 52,416 observations of energy consumption on a 10-minute window. Every observation is described by 9 feature columns.

Data Overview:

Time range: 2017-01-01 00:00:00 to 2017-12-30 23:50:00
Average Time Period: 0 days 00:10:00
Sampling Frequency: 0.00167
Number of records: 52,416
Number of zones: 3
Missingness: 0
datatype: float

Statistical Overview:

- The key statistics of Carbon Diffuse Flows, humidity and Wind Speed

```
Diffuse Flows Statistics:
--------------------
  mean: 75.0280
median: 4.4560
   std: 124.2109
   min: 0.0110
   max: 936.0000
    q1: 0.1220
    q3: 101.0000

Humidity Statistics:
--------------------
  mean: 68.2595
median: 69.8600
   std: 15.5512
   min: 11.3400
   max: 94.8000
    q1: 58.3100
    q3: 81.4000
```

```
Wind Speed Statistics:
--------------------
  mean: 1.9595
median: 0.0860
   std: 2.3489
...
   min: 3.2470
   max: 40.0100
    q1: 14.4100
    q3: 22.8900
```

Figure 1: Statistical Values of datatset

3. Methods and Analysis:

3.1 Data Preprocessing and Data Quality:

- Data quality was ensured by filtering out readings exceeding valid ranges for each sensor.
- Outliers were identified and removed using the IQR method.

3.2 Exploratory Data Analysis

3.2.1 Distribution Analysis Using Histograms

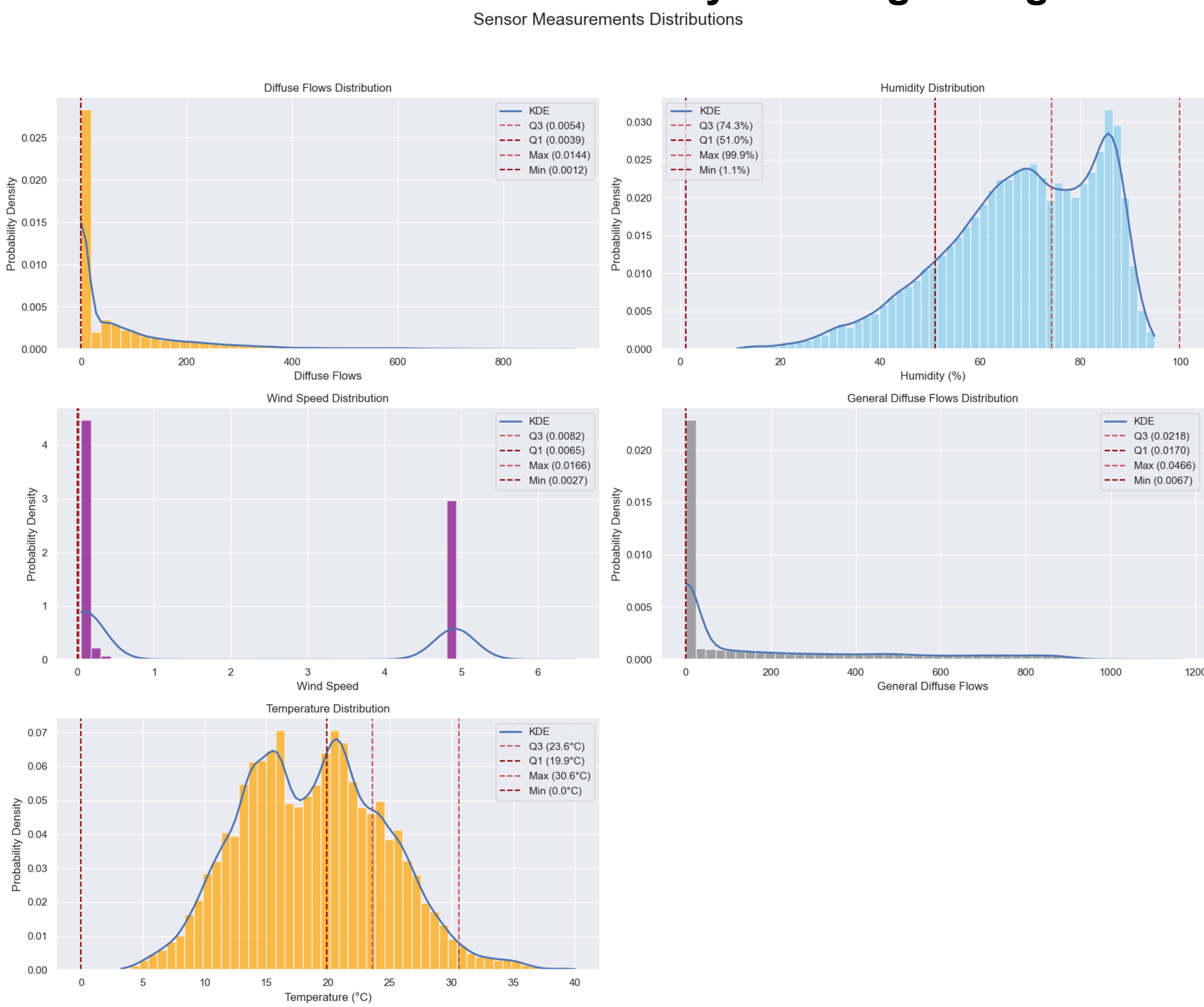


Figure 2: Sensor Measurements Distribution using matplot.lib

3.2.2 Original VS Clean Data:

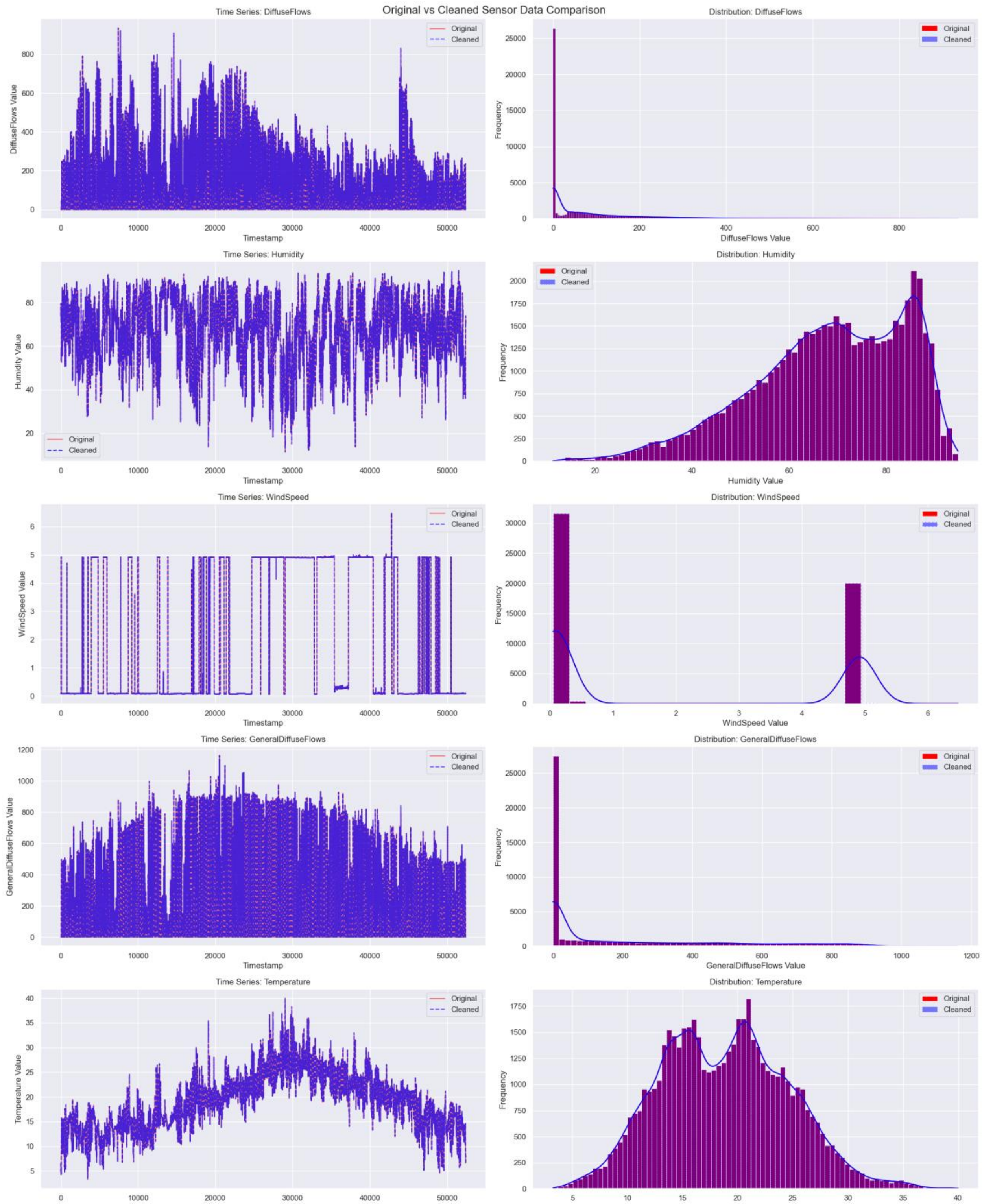


Figure 3: Original Vs Clean Data

3.2.3 Time Series Patterns:

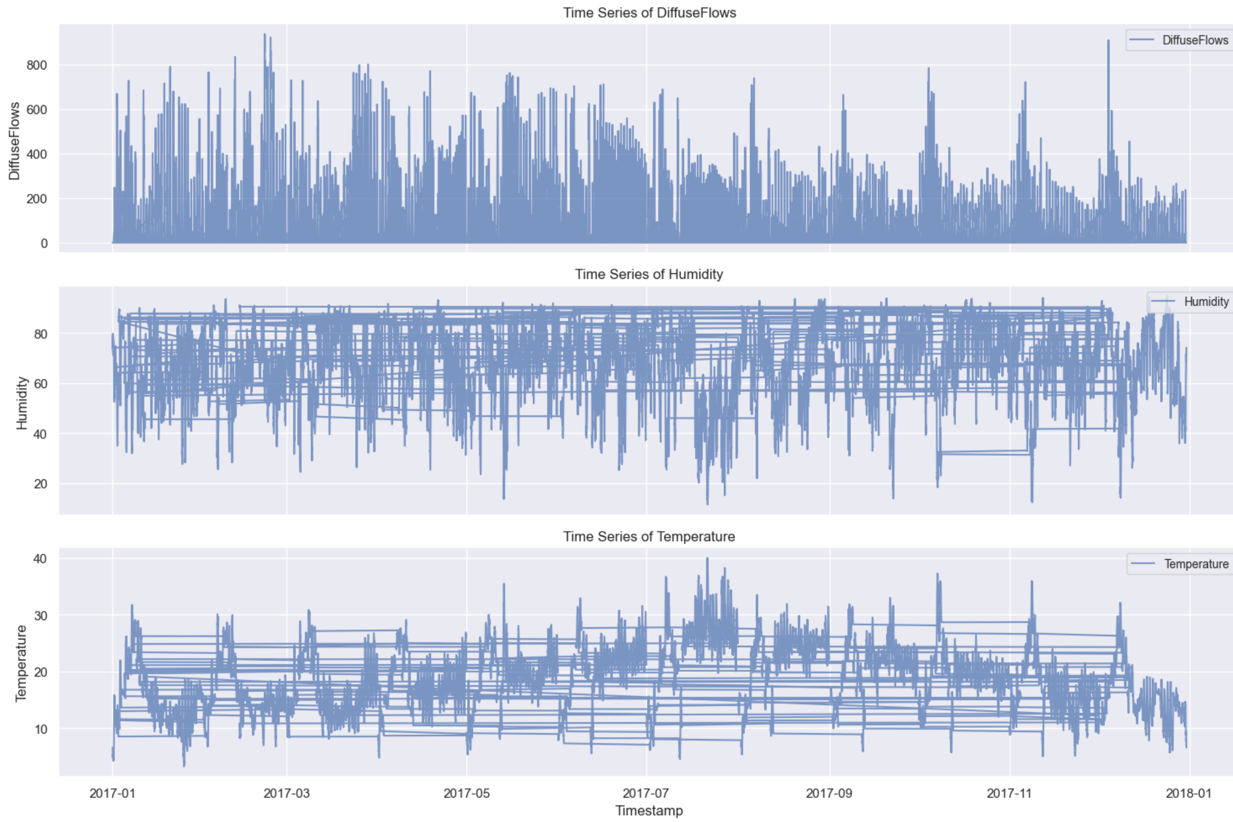


Figure 4: Time Series Patterns

3.2.4 Correlation Analysis:

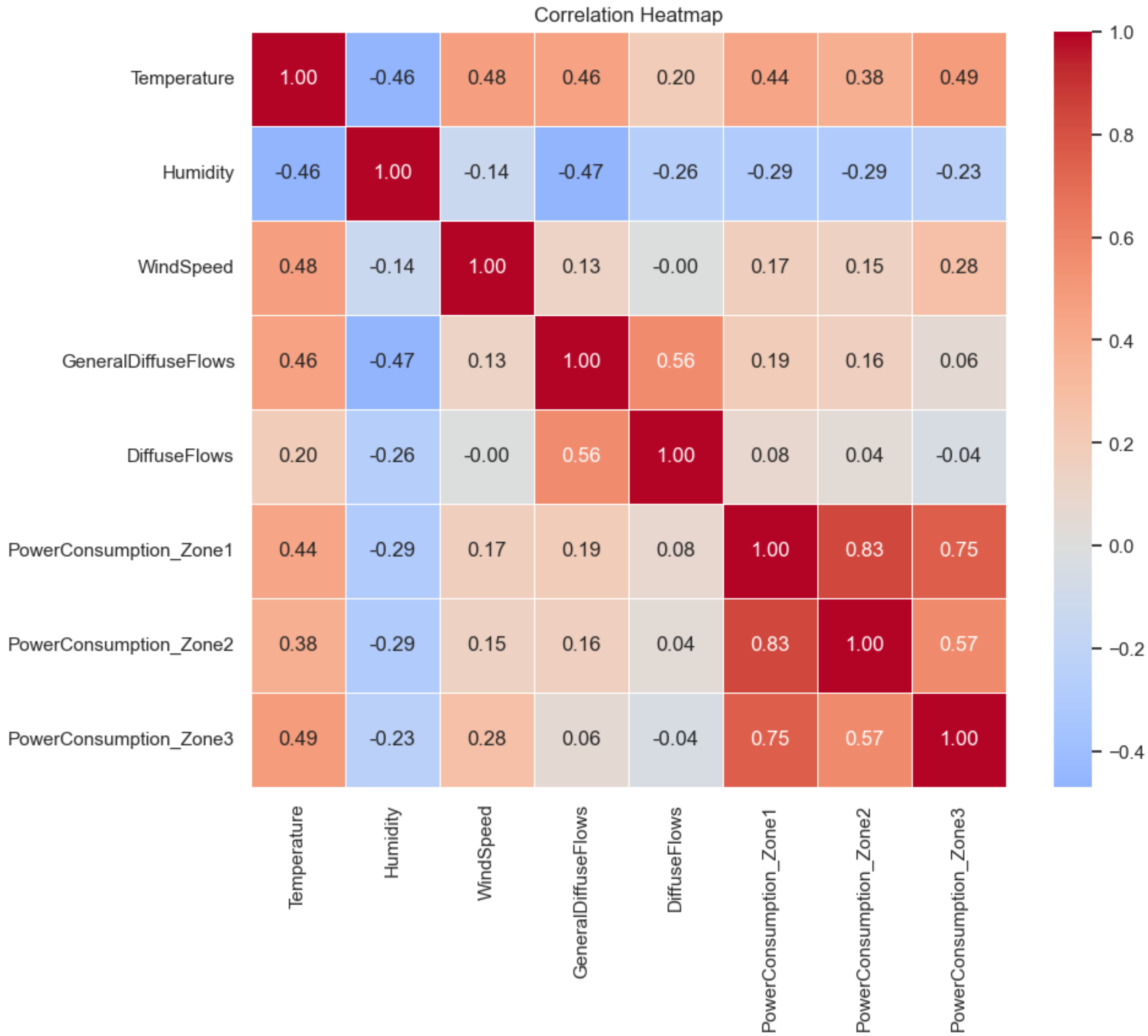


Figure 5: Correlation Heatmap

3.2.5 Daily Patterns:

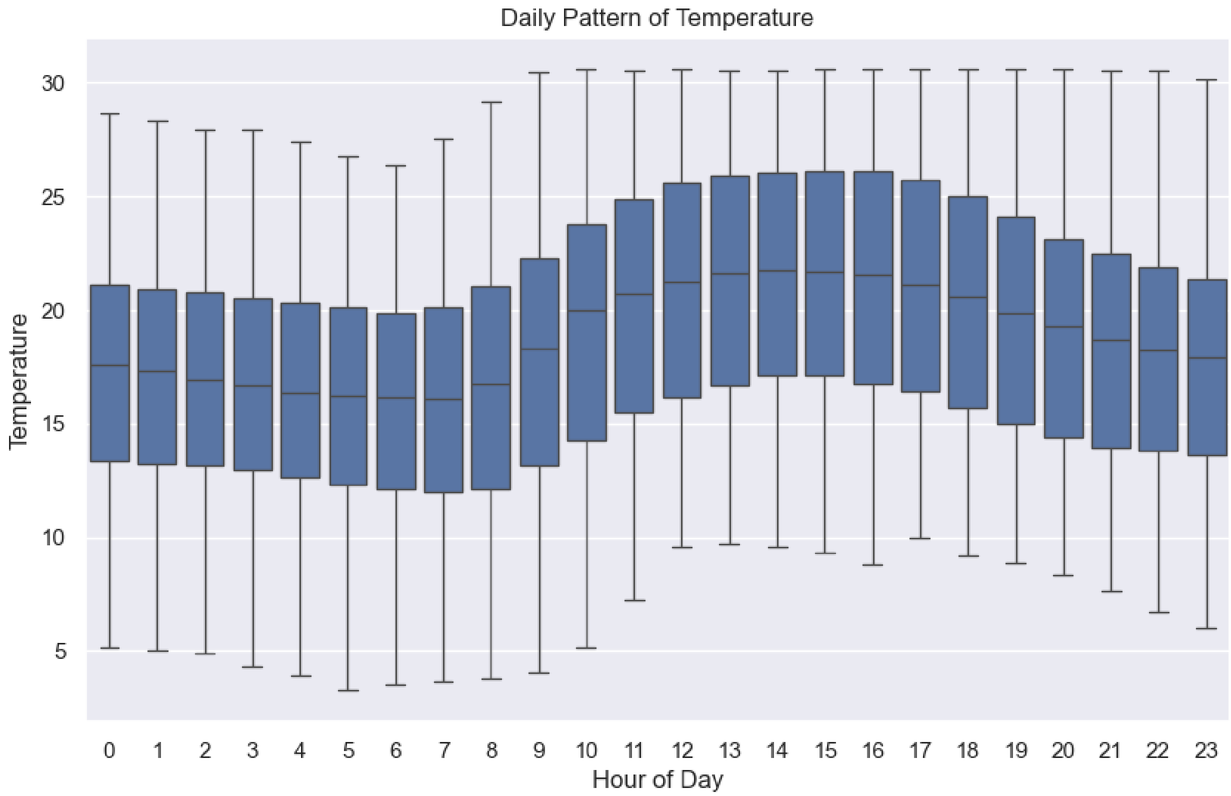


Figure 6: Daily Pattern Of Temperature

3.3 Statistical Data Analysis

3.3.1 Probability Analysis

```
Diffuse Flows Statistics from Dataset:
Mean: 0.0110
Standard Deviation: 0.0000
Threshold: 0.005
Standard deviation is zero or near-zero. Probability estimation using a normal distribution is not meaningful.
```

```
Humidity Statistics:
Mean: 68.2897
Standard Deviation: 15.4585
Threshold: 70
Z-Score: 0.1106
Probability of Humidity exceeding 70 is approximately 0.4560

Threshold-Based Probability Estimations:
Threshold: 50.0 -> Probability of exceeding: 0.8816
Threshold: 60.0 -> Probability of exceeding: 0.7041
Threshold: 70.0 -> Probability of exceeding: 0.4560
Threshold: 80.0 -> Probability of exceeding: 0.2244
```

```
Temperature Statistics:
Mean: 18.7473
Standard Deviation: 5.6653
Threshold: 25
Z-Score: 1.1037
Probability of Temperature exceeding 25 is approximately 0.1349

Threshold-Based Probability Estimations:
Threshold: 20.0 -> Probability of exceeding: 0.4125
Threshold: 22.0 -> Probability of exceeding: 0.2829
Threshold: 25.0 -> Probability of exceeding: 0.1349
Threshold: 28.0 -> Probability of exceeding: 0.0512
```

Figure 7 Threshold Probabilities

```
Cross Tabulation of Temperature and Humidity Categories:
Humidity       Low  Medium  High  Very High
Temperature
Low           1693    3742  9454      14668
Medium         688    1242  2928       3127
High           594     657  1446        793
Very High     3256    1793  2368       2325
Conditional Probability P(High Humidity | Medium Temperature):
0.366688
Conditional Probability P(Medium Humidity | Low Temperature):
0.126603

Conditional Probability P(Very High Temperature | High Humidity):
0.146209
```

Figure 8 Cross Tabulation

3.3.3.1 Summary of Observations

- High Thresholds have lower probability of being exceeded because they are further than the mean.

- Low Thresholds have higher probabilities of being exceeded as they are closer to or below the mean.

- The cross-tabulation helps identify patterns or associations between temperature and humidity levels..

- This can reveal whether certain ranges of temperature are more likely to co-occur with specific humidity levels.

- Conditional probability analysis, these probabilities reveal how certain ranges of temperature and humidity are likely to co-occur.

- The analysis can be used to infer relationships between temperature and humidity categories and predict one variable based on the other.

3.3.2 Law of Large Numbers

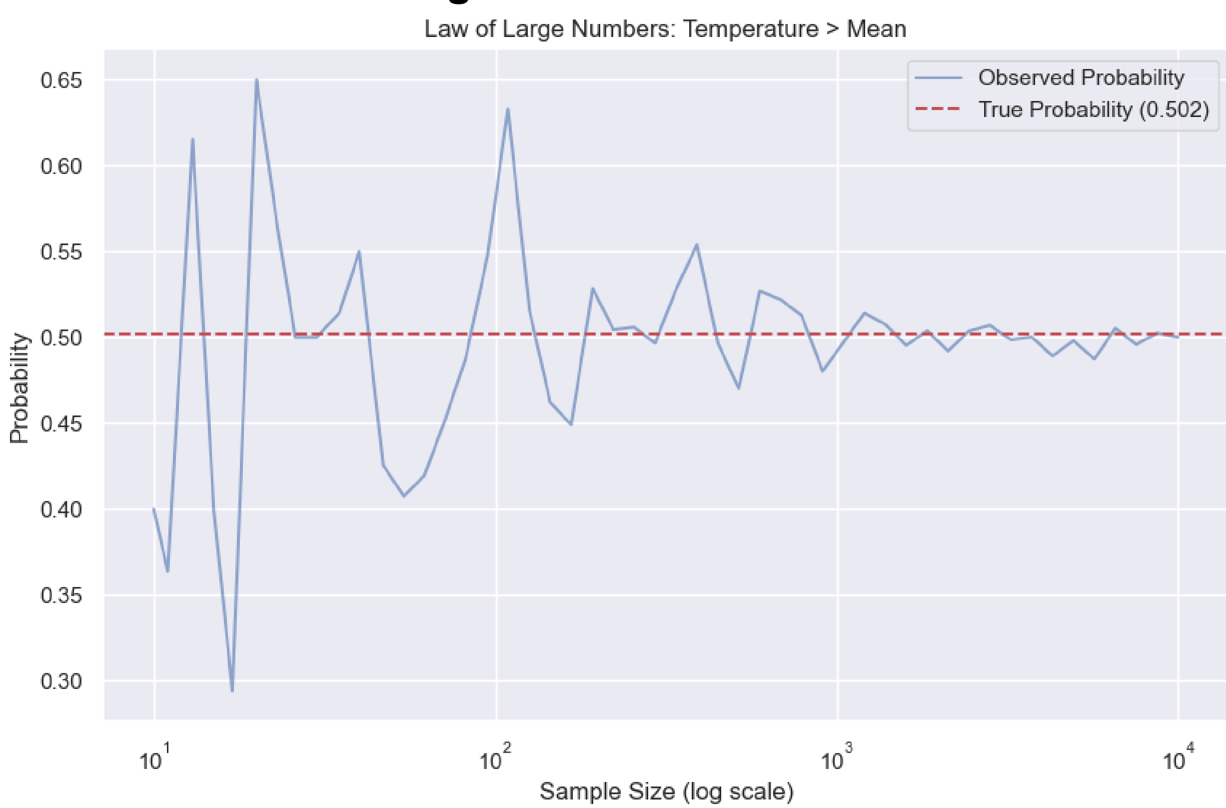
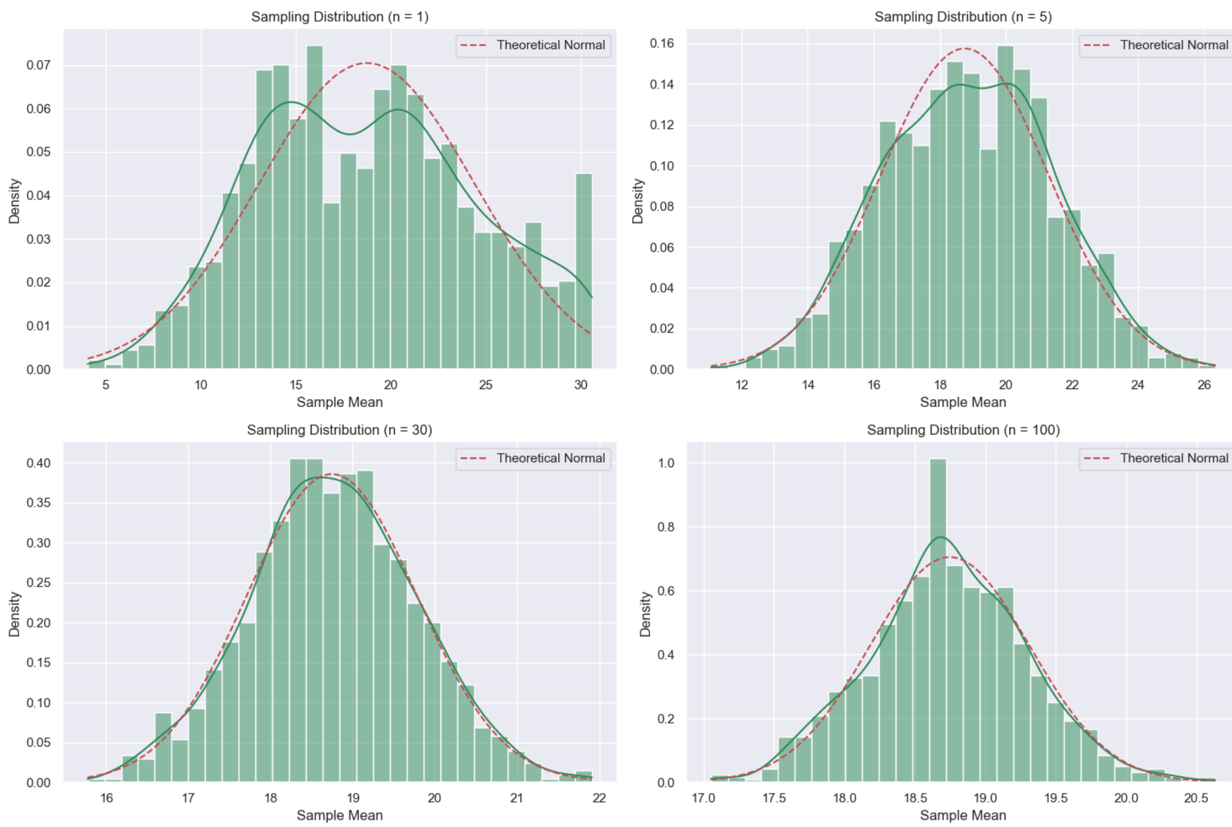


Figure 8: Law of Large Numbers

3.3.2.1 Result Interpretation

- As the sample size increases (logarithmic scale on the x-axis), the observed probability becomes closer to the true probability. This demonstrates the Law of Large Numbers, which states that as the size of a sample increases, the sample mean (or probability) will converge to the population mean (or true probability).

3.3.3 Central Limit Theorem



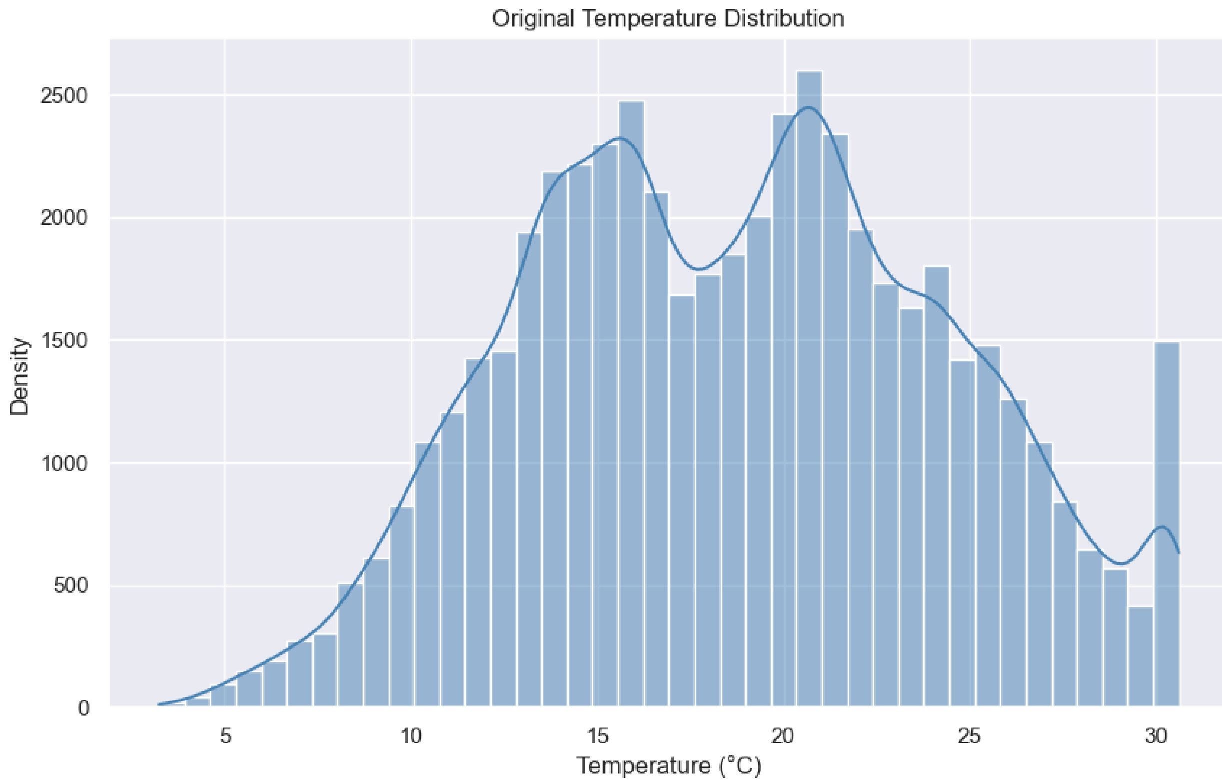


Figure 9: Central Limit Theorem

3.3.3.1 Result Interpretation

```
Sampling Distribution for n = 1
Mean of sample means: 18.8183
Population mean: 18.7473
Std of sample means: 5.7742
Expected std (σ/√n): 5.6652
--------------------------------------------------
Sampling Distribution for n = 5
Mean of sample means: 18.8459
Population mean: 18.7473
Std of sample means: 2.5206
Expected std (σ/√n): 2.5336
--------------------------------------------------
Sampling Distribution for n = 30
Mean of sample means: 18.7652
Population mean: 18.7473
Std of sample means: 1.0134
Expected std (σ/√n): 1.0343
--------------------------------------------------
Sampling Distribution for n = 100
Mean of sample means: 18.7357
Population mean: 18.7473
Std of sample means: 0.5591
Expected std (σ/√n): 0.5665
--------------------------------------------------
```

3.3.4 Linear Regression Analysis

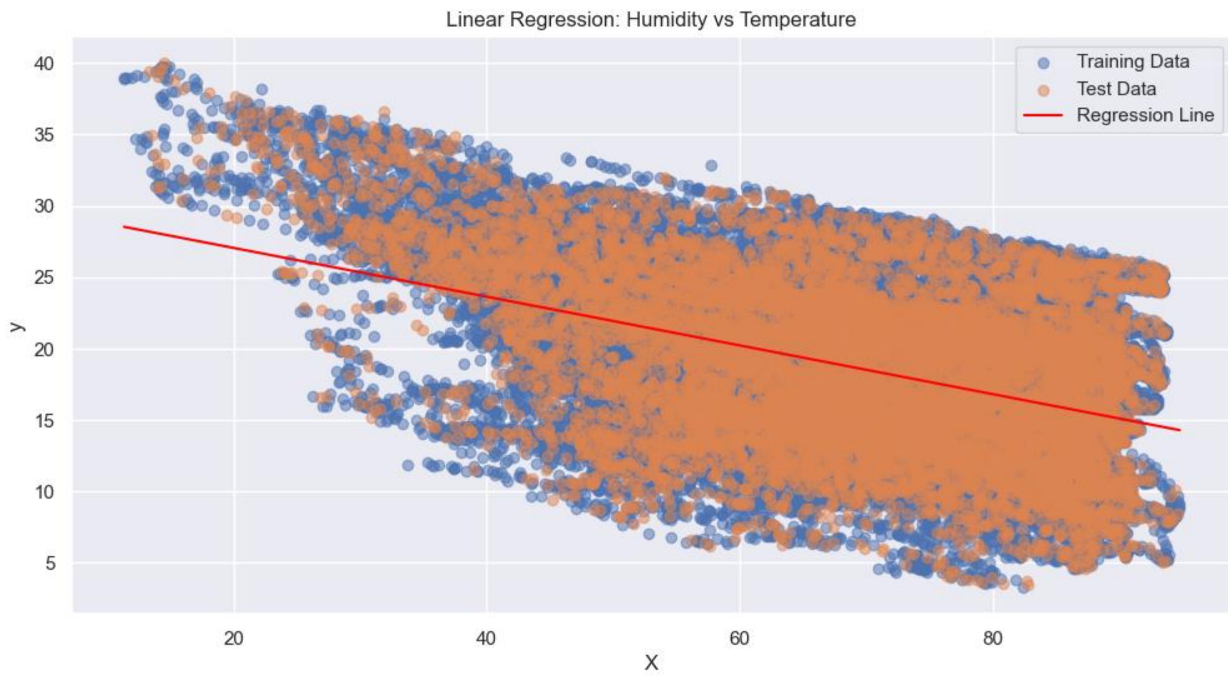
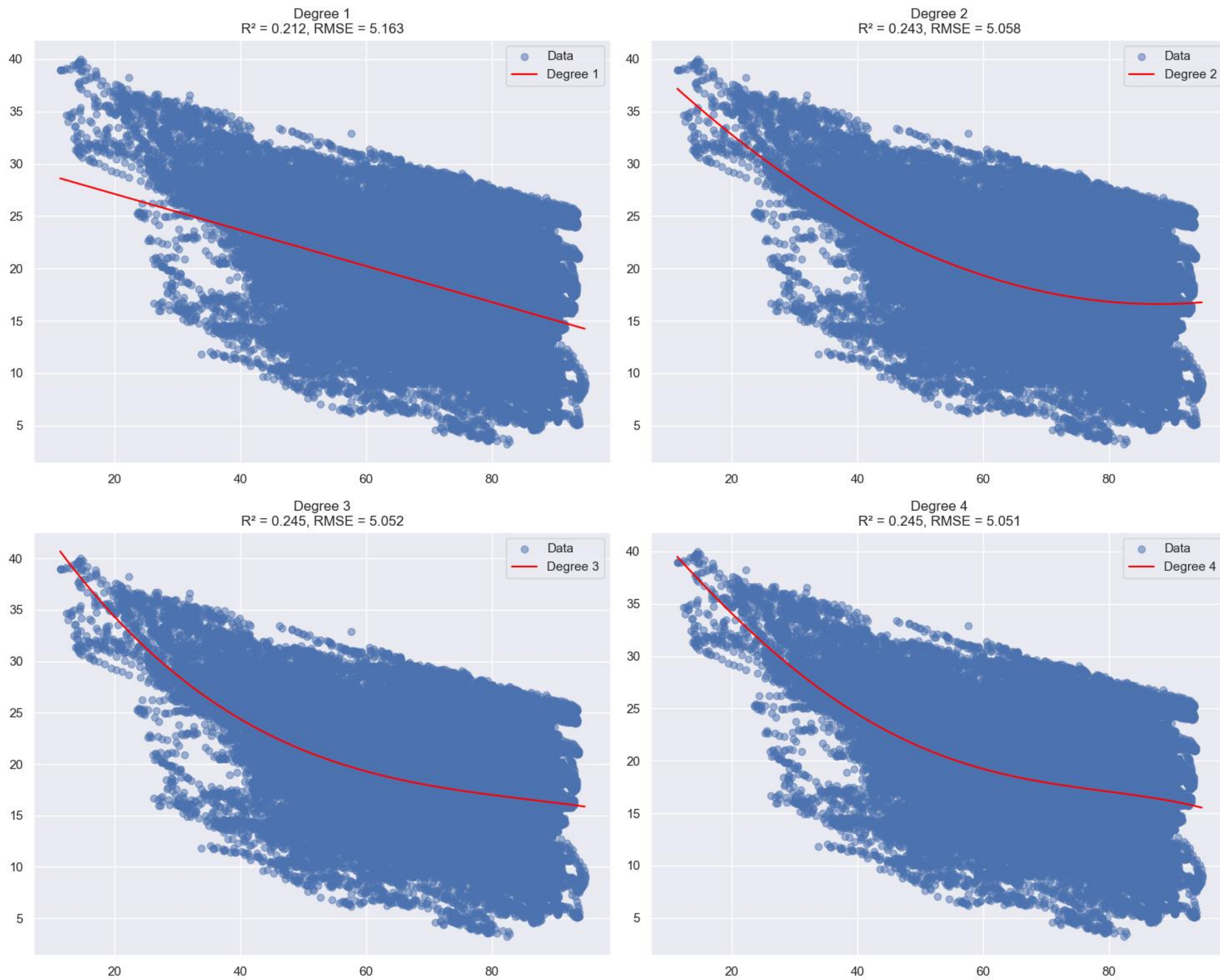


Figure 10: Linear Regression Analysis

3.3.4.1 Result Interpretation

- The poor fit of the regression line to the data (as indicated by the low R^2 values in your earlier output) confirms that the linear regression model is not a good fit.

3.3.5 Polynomial Regression Analysis



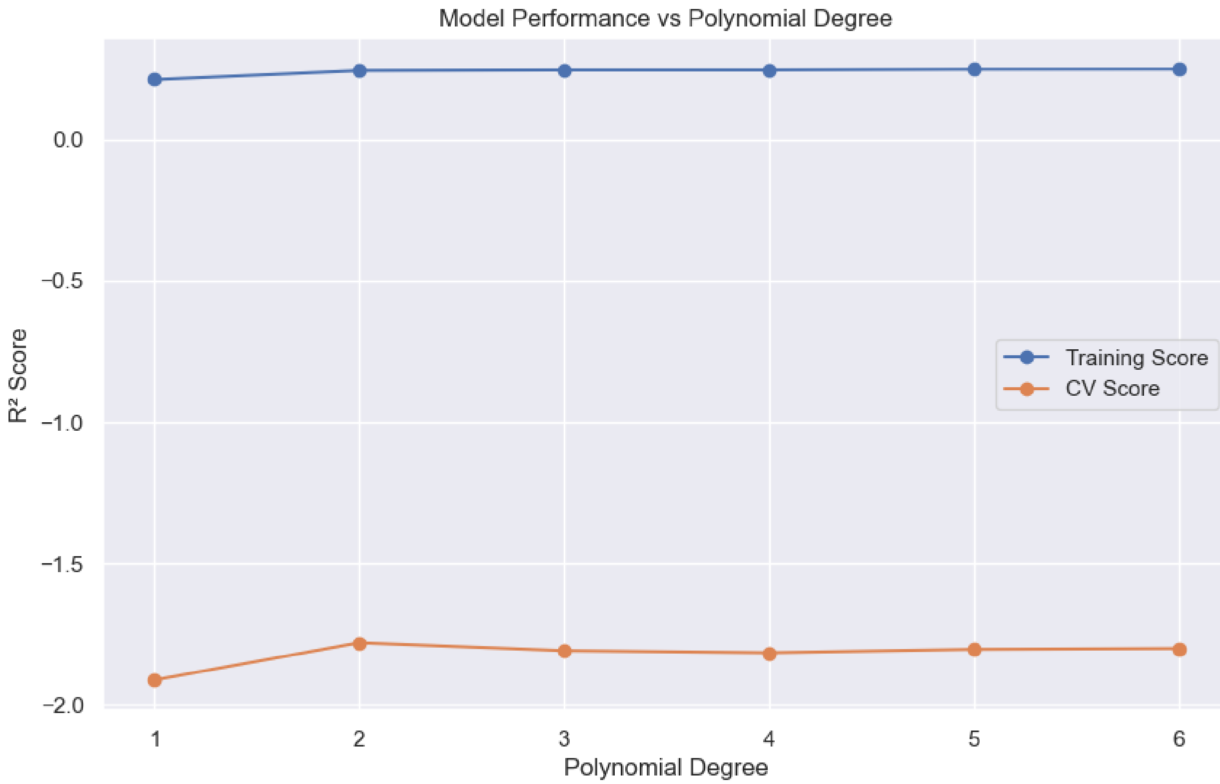


Figure 11: Polynomial Regression Analysis

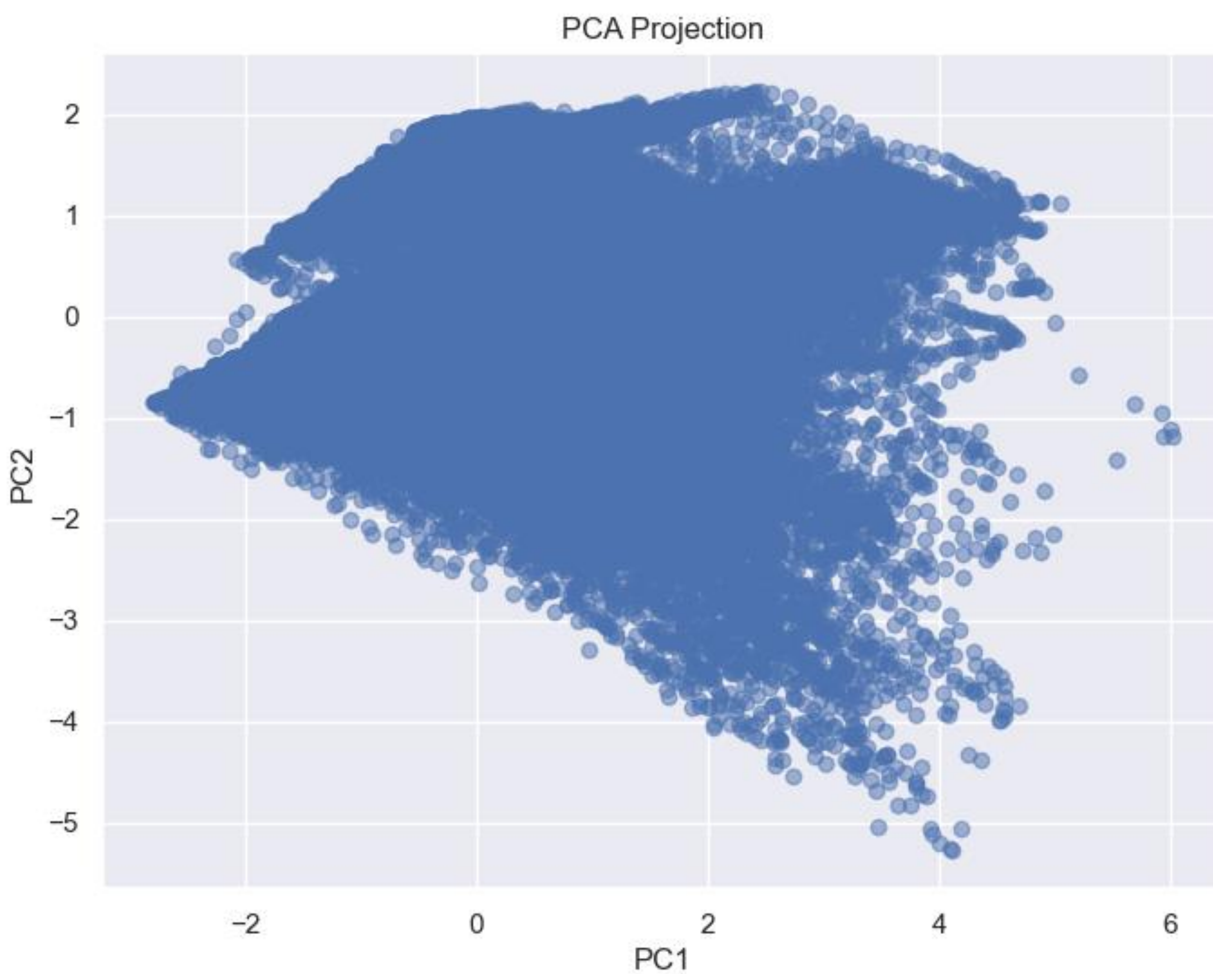
3.3.4.1 Result Interpretation

- Scatter plot: Visualizing the relationship between the independent variable (X-axis) and the dependent variable (Y-axis) for different polynomial degrees.

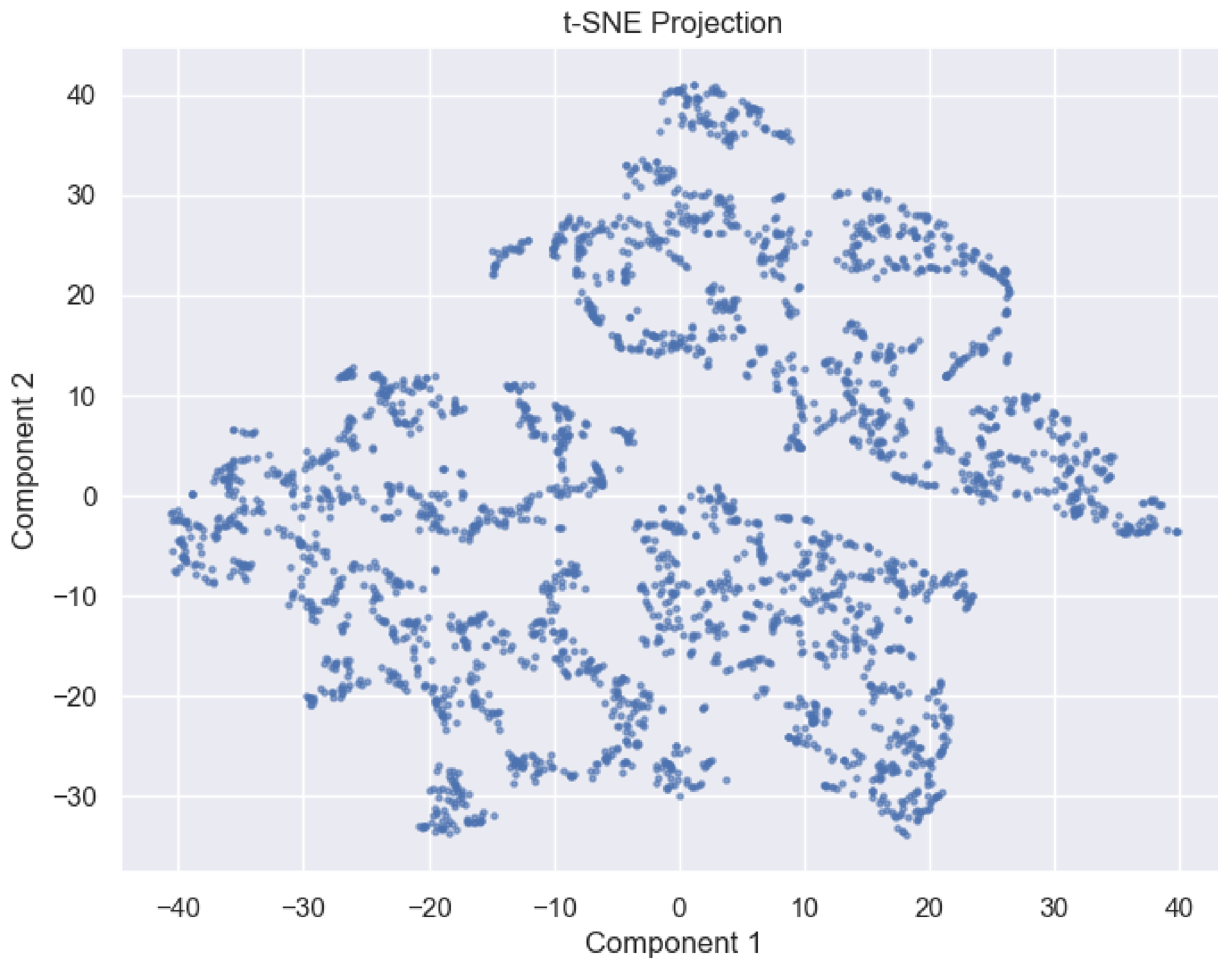
- Model performance plot: Showing how the model's performance (likely measured by metrics like R-squared or Mean Squared Error) changes with increasing polynomial degrees.

- Residual plots: Examining the distribution of residuals (the difference between predicted and actual values) for different models.

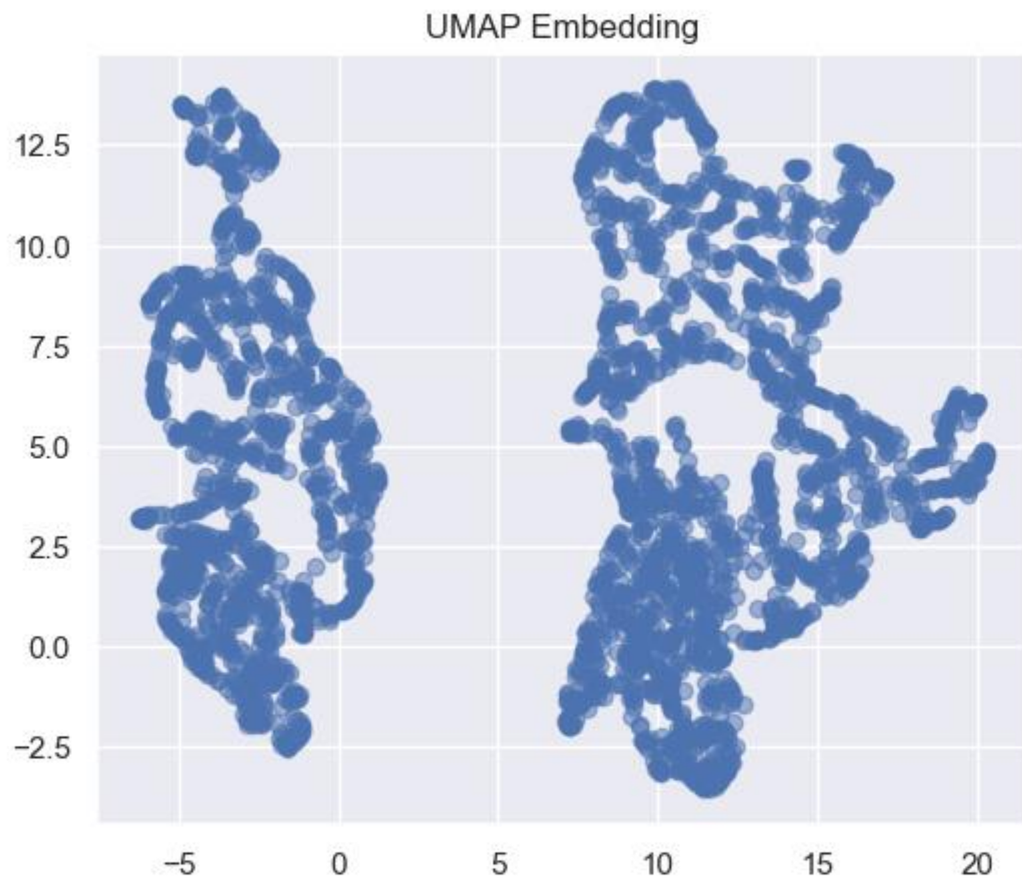
3.4.1 PCA Correction



3.4.1.1 T-SNE Projection



3.4.1.2 UMAP Embedding



UMAP Embedding
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5
0
5
10
15
20