Data Mining Project

Table of Contents

Introduction

This project evaluates the potential of Dalaman's wind patterns for renewable energy using advanced data analysis. By uncovering trends and predicting wind behavior, the study aims to determine the region's suitability for wind energy projects.

- **Objective**: Assess Dalaman's wind energy potential through data analysis.
- **Problem**: Wind energy viability requires consistent patterns; Dalaman's feasibility is unknown.
- **Approach**: Apply data mining techniques to analyze trends and predict wind speeds.
- **Significance**: Supports sustainable energy initiatives and development.

Data Preprocessing Steps

To ensure the wind speed data was clean and ready for analysis, several preprocessing steps were undertaken to handle missing values, standardize the data, and maintain integrity. The following are the steps:

- **Column Extraction:** Focused on the DALAMAN_Speed column from a multi-location dataset.
- **Data Cleaning:** Removed unnecessary rows, including headers within the data.
- **Numeric Conversion:** Converted entries to numeric, treating non-numeric values as missing.
- **Missing Data Handling:** Replaced missing values with the column mean (1.76 m/s).
- **Standardization:** Applied z-score normalization for clustering and machine learning readiness.
- **Final Output:** Saved the cleaned, standardized dataset in Excel for further analysis.

Results of Preprocessing

- The dataset was free of missing or non-numeric values.
- The data was standardized, allowing for uniform scaling.
- The preprocessed dataset was ready for the next stages of analysis, including descriptive statistics, time series analysis, clustering, and classification.

Key Metrics Calculated

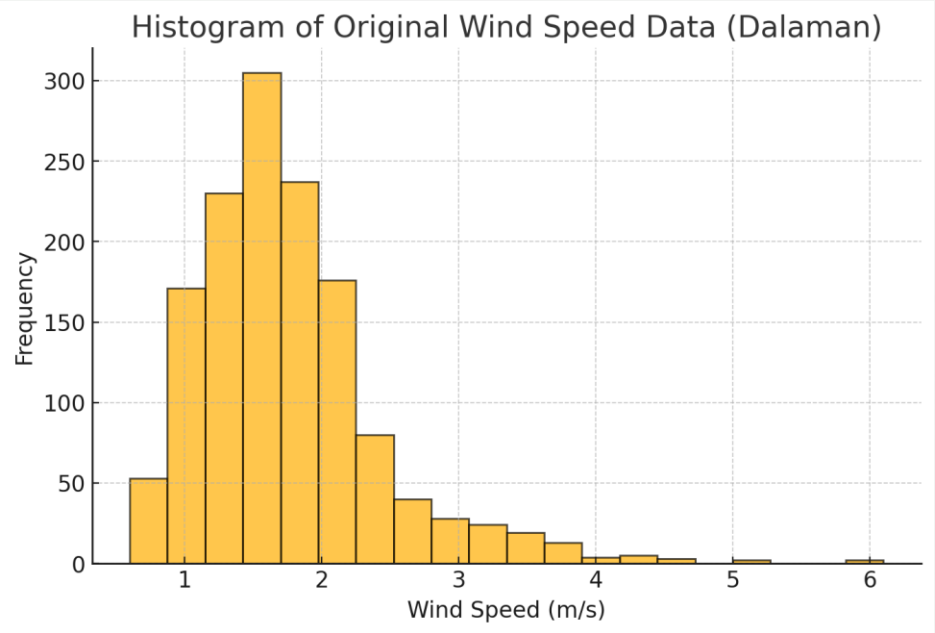
- Mean Wind Speed:
  - The average wind speed in the dataset was 1.76 m/s.
  - This value represents the central tendency of wind speeds in Dalaman.
- Standard Deviation:
  - The standard deviation of wind speeds was 0.68 m/s.
  - This measure indicates the amount of variability or dispersion in the wind speeds.
- Distribution:
  - The wind speeds were visualized using histograms to understand the data distribution.

- Minimum and Maximum Wind Speeds:
  - The lowest wind speed recorded was 0.6 m/s, while the highest was 6.1 m/s.
  - These values highlight the range of wind speeds in the dataset.

```
count    1.392000e+03
mean    -1.020895e-17
std      1.000359e+00
min     -1.722248e+00
25%     -6.866215e-01
50%     -9.483488e-02
75%      3.490051e-01
max      6.414818e+00
Name: DALAMAN_Speed_Scaled, dtype: float64
```

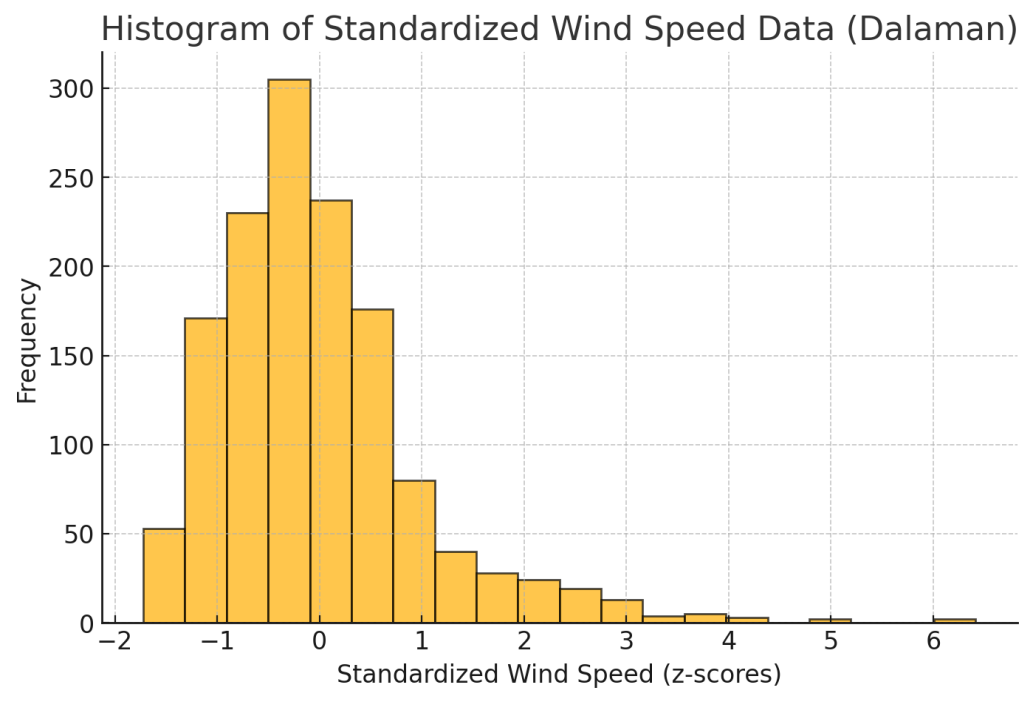
Histogram of Original Data

- A histogram was created to display the frequency distribution of wind speeds.
- The plot shows that most wind speeds are concentrated around the mean, with a few higher outliers.



Histogram of Standardized Data

- After standardization, a histogram of z-scores was plotted.
- This plot demonstrates the data's transformation to have a mean of 0 and a standard deviation of 1.



Time Series Analysis

Time series analysis was conducted to explore trends, seasonal patterns, and predict future wind speeds in Dalaman.

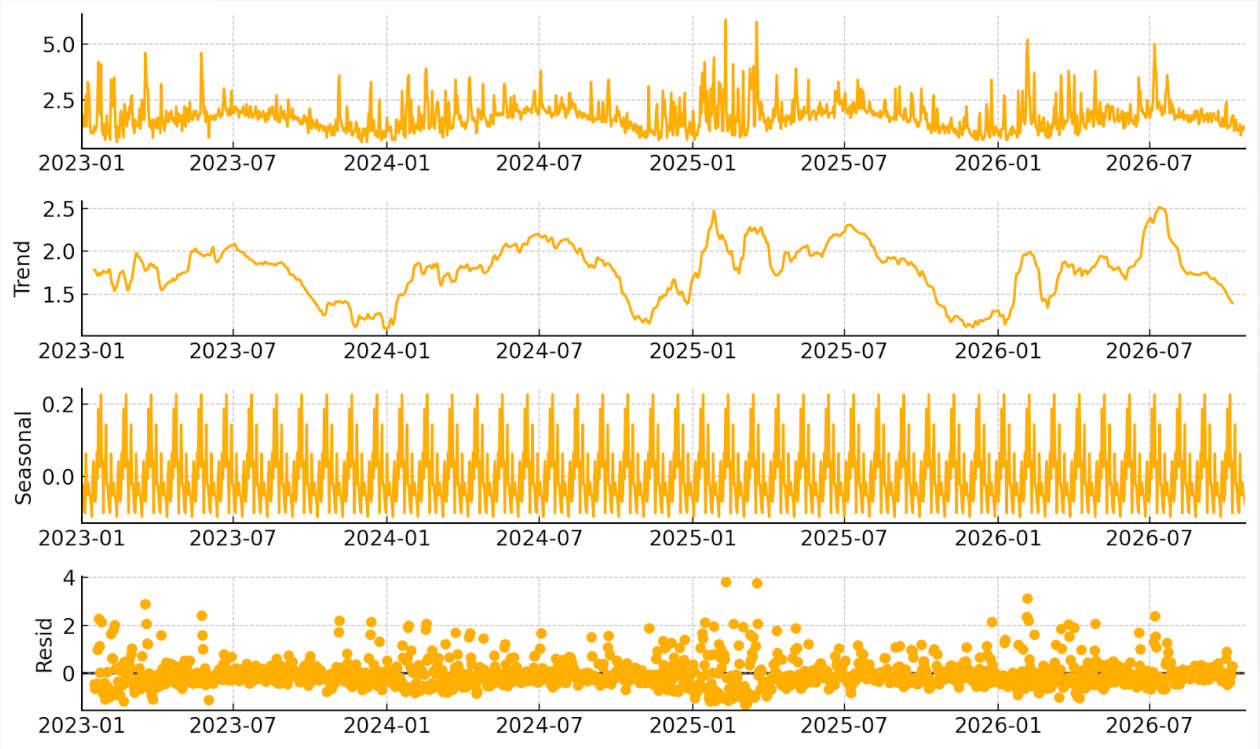
- **Time Index Setup**: Created a synthetic time index to represent the data's temporal sequence.
- **Trend Analysis**: Visualized long-term patterns using rolling mean and standard deviation.
- **Stationarity Check**: Tested data stationarity with the ADF test; differencing applied if non-stationary.
- **Decomposition**: Separated the time series into trend, seasonality, and residual components.
- **Forecasting**: Utilized the ARIMA model for short-term wind speed predictions.

Time Series Analysis Results

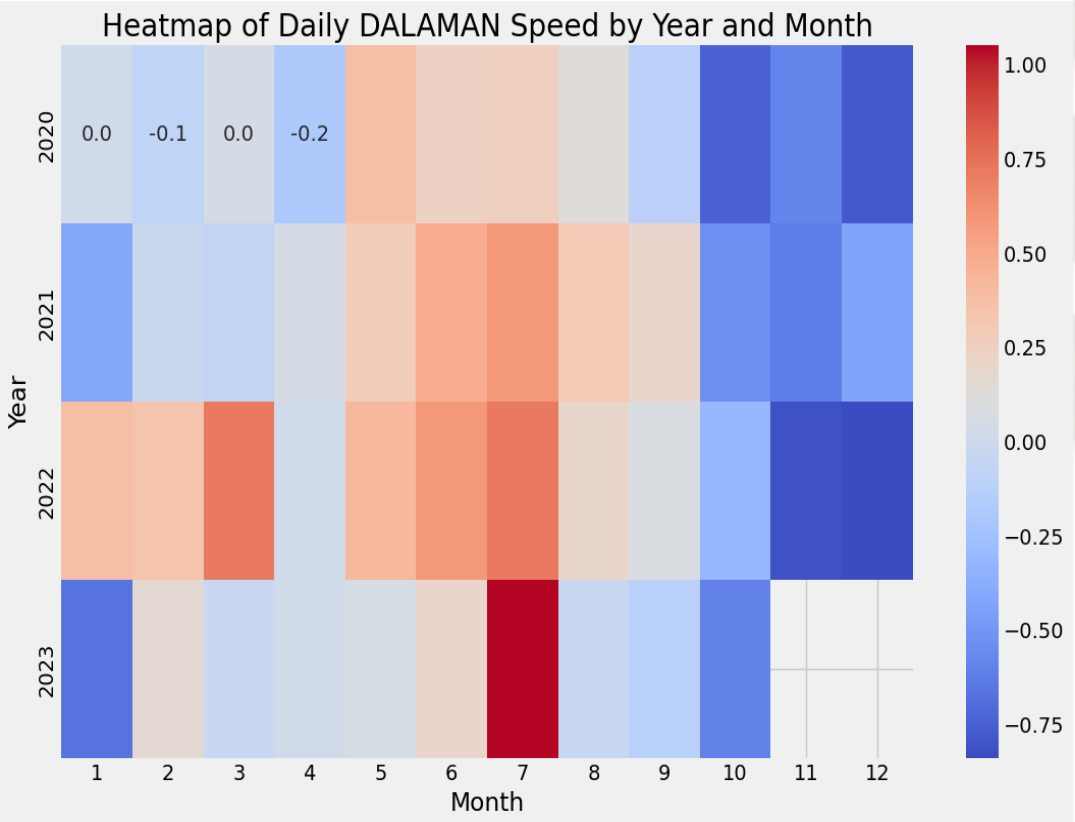
- Trend and Seasonal Patterns:
  - The trend component revealed periods of increasing and decreasing wind speeds.
  - Seasonal variations highlighted potential daily or seasonal cycles in the data.
- Stationarity Test:
  - The ADF test indicated the data was non-stationary. Differencing the series resolved this issue.
- Future Predictions:
  - Forecasting models predicted wind speeds for the next n time steps (e.g., days or hours).
  - The predictions provide a basis for assessing short-term wind energy potential.

Time Series Decomposition

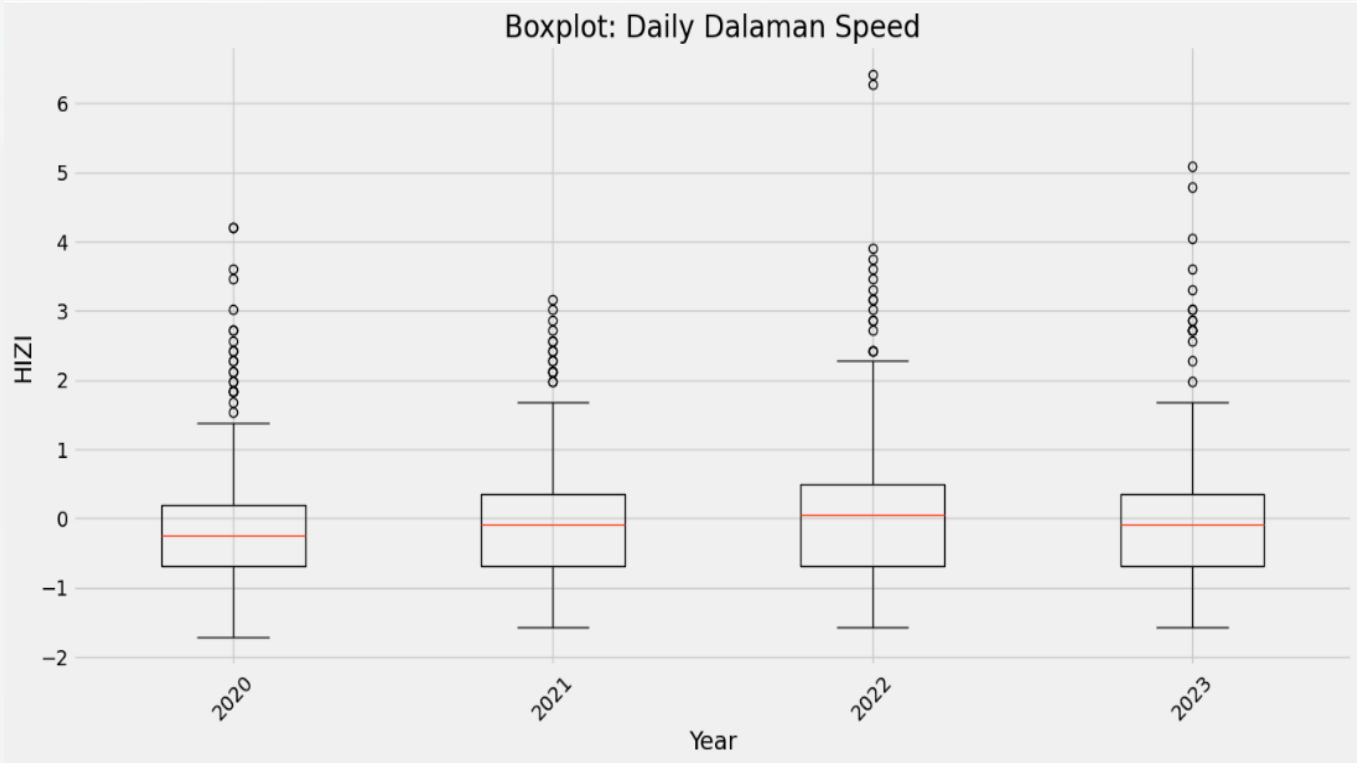
Plots of the trend, seasonal components, and residuals



Heatmap for the Time Series of Daily DALAMAN Speed by each year and month

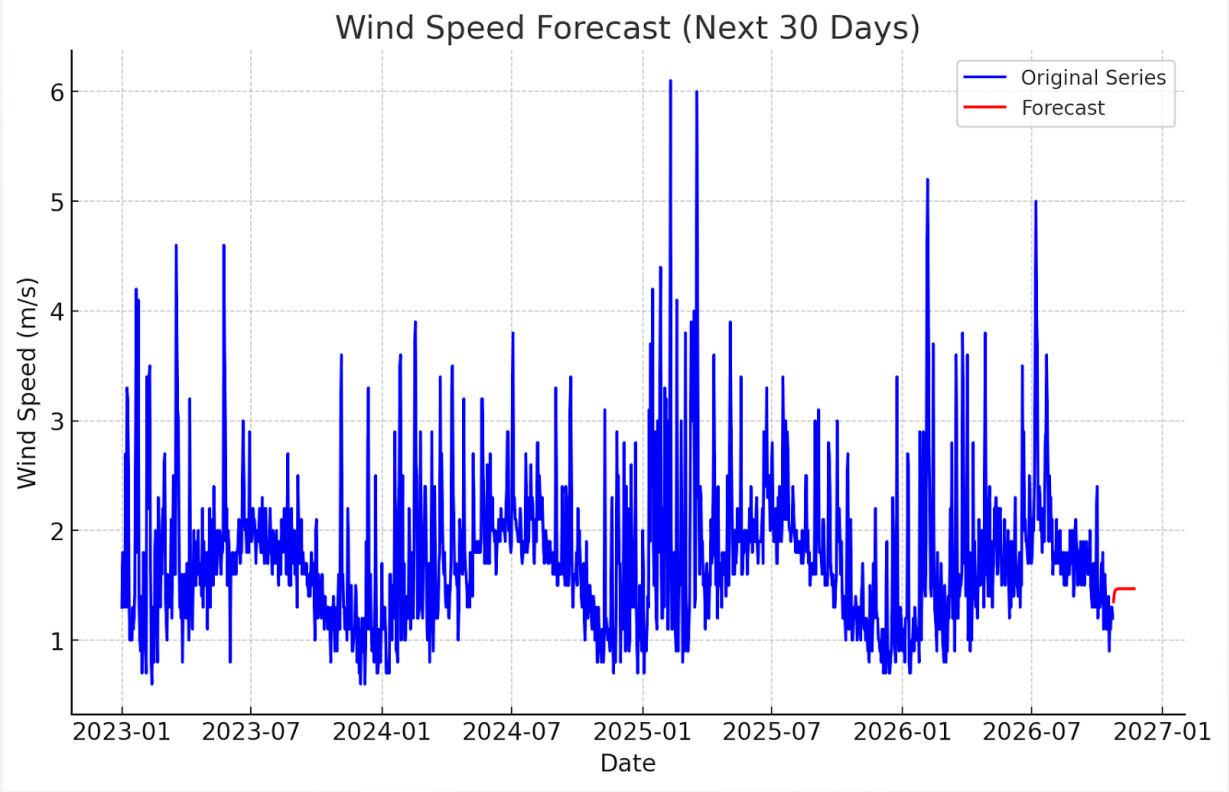


Yearly Boxplots for the Daily DALAMAN Speed



Forecasting Results

Plot showing the actual wind speed data alongside the predicted values



Clustering

Objective:
Categorize wind speeds into clusters (e.g., low, medium, high) for detailed analysis.

Approach:

•Use **k-means clustering** to group wind speed data.

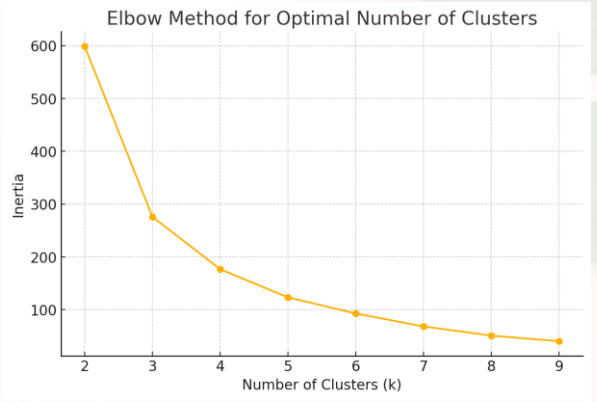
•Interpret results through cluster centroids.

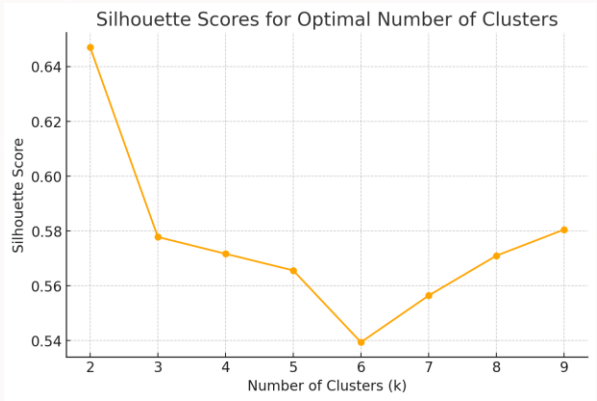
Steps:

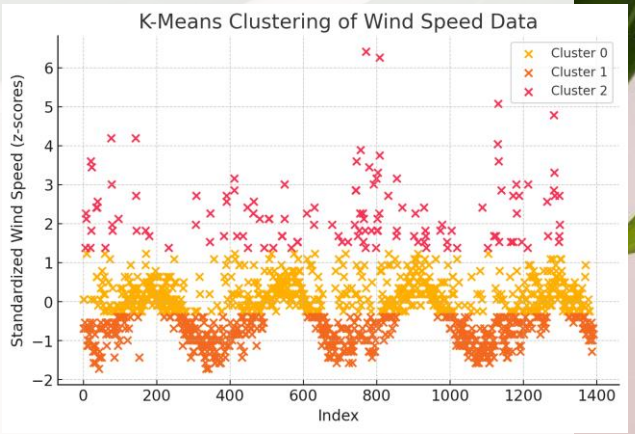
1.Standardize data for uniform clustering.

2.Determine optimal clusters using the **Elbow Method**.

3.Apply **k-means** to group wind speeds into distinct clusters.







Classification

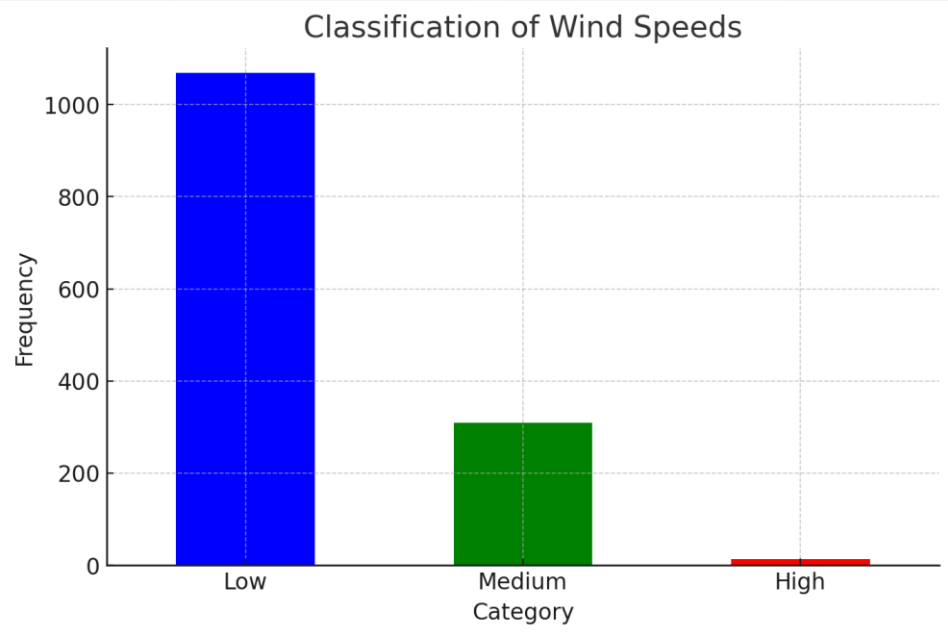
Objective: To classify wind speeds into predefined categories based on thresholds (e.g., low: <2 m/s, medium: 2–4 m/s, high: >4 m/s).

Approach:

1.Define categories based on wind speed thresholds.

2.Train a simple classification model (e.g., decision tree or logistic regression) on labeled data.

3.Evaluate the model using accuracy, precision, recall, or other metrics.



Future Simulation

Objective:

•Predict future wind speeds based on historical data.

•Assess the feasibility of wind energy generation in Dalaman for the forecasted period.

Approach:

•Utilize the ARIMA model for time series forecasting.

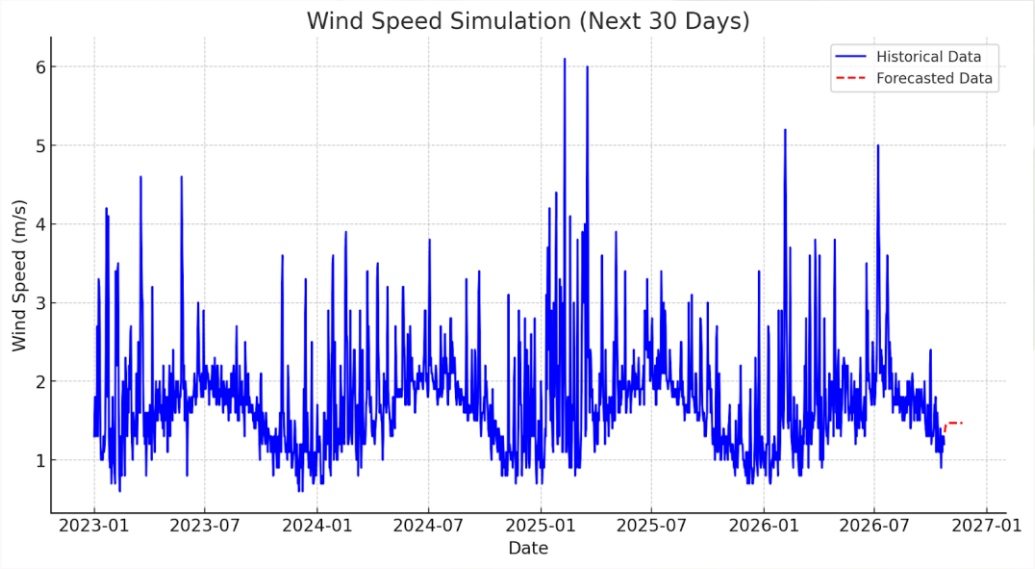
•Extend predictions to simulate wind speeds for the next 30 days (or a custom period).

Simulation Process:

•Model historical wind speeds using ARIMA (AutoRegressive Integrated Moving Average).

•Generate predictions for future wind speeds.

•Visualize the forecasted wind speeds alongside the historical data.



Results of Future Simulation

1.Forecasted Period:

- The forecast covers 30 days, starting from October 24, 2026 to November 22, 2026.

2.Forecasted Wind Speeds:

- Mean Wind Speed: 1.46 m/s
- Maximum Wind Speed: 1.47 m/s
- Minimum Wind Speed: 1.35 m/s

3.Visualization:

- The plot displays historical wind speeds alongside the forecasted values for the next 30 days.
- Forecasted wind speeds show a slight upward trend, stabilizing at around 1.47 m/s.

Conclusion

This study evaluated Dalaman's wind energy potential using data mining techniques applied to historical wind speed data. Key findings include:

- **Preprocessing**: Ensured data integrity, standardization, and consistency for analysis.
- **Descriptive Statistics**: Mean wind speed of 1.76 m/s with moderate variability (0.68 m/s).
- **Time Series Analysis**: Forecasts stabilized at 1.47 m/s, indicating stable but low wind speeds.
- **Clustering**: Identified three categories—low (<2 m/s), medium (2–4 m/s), and high (>4 m/s).
- **Classification**: Validated thresholds, aiding wind turbine performance assessments.

Conclusion: Dalaman’s wind speeds are suitable for small-scale renewable energy projects. This study highlights the importance of data mining in wind energy evaluations, offering insights for sustainable development in Dalaman and similar regions.

Thank you!

				ANBUL AYDIN UNIVERSITY **SEN431** **DATA MINING** **PROJECT/REPORT** **Prof. Dr. Zafer ASLAN** **Dead Line: January 8 2025**			**024-25**
dent Name							
dent ID							
nature							
estions			**3**	**4**	**5**	**TOTAL**	
arning Outcomes	**L**	**L**	**3,4**	**LO2, 3**	**LO4**	**LO 1-4**	
ximum Score			**0**	**20**	**20**	**100**	
dent Score							

Students names:

YAHIA BAWADEKJI - B2105.090018

SUHA SUHA - B2205.090019

Khalid Alrais - B2105.090048

ABDULLAH ALSHALFI - B2005.090021

Nedime Serra Liman - B2005.090063

IBRAHIM AL SHALABI - B2105.090044

Abdallah I. J. Dwikat - B2205.010045

1. Definition of the topic, problem, aim of the report

Topic:

The project focuses on **analyzing wind speed data from Dalaman** to determine its viability for renewable energy utilization. The topic highlights the application of data mining techniques to uncover patterns, trends, and predictions.

Problem Statement:

Wind energy is a significant renewable energy source, but its feasibility heavily depends on consistent and predictable wind speeds. Dalaman, as a region, requires assessment to determine if its wind patterns are suitable for energy generation. This project addresses the need for:

- Analyzing historical wind speed data.
- Identifying patterns and trends.
- Predicting future wind behavior to assess its potential for renewable energy projects.

Aim:

The aim of this project is to:

1. Analyze the **DALAMAN_Speed** wind speed data to understand patterns and trends.
2. Apply data mining techniques (e.g., clustering, classification, and time series analysis) to uncover insights.
3. Simulate future wind speed scenarios to determine the feasibility of wind energy projects in Dalaman.

2. General review on literature

Introduction

Renewable energy is crucial for sustainable development, with wind energy being a significant component due to its potential for clean power generation. Assessing wind energy potential involves analyzing wind speed data to determine the feasibility of wind power installations.

Review of Studies

Wind Energy Potential Assessment

Dağtekin and Yelmen (2022) conducted a study on wind energy potential in Turkey, focusing on Mersin province. They found that the region has a usable wind energy potential of about 3,414 MW, indicating significant opportunities for wind power development.
DergiPark

Arslan (2015) investigated Turkey's wind energy potential through statistical analysis, highlighting the importance of regional assessments in understanding the country's renewable energy capabilities.
Acikbilim

Data Mining Applications in Wind Energy

El-Halees (2015) applied data mining techniques to analyze wind speed behavior using meteorological data from Gaza. The study demonstrated the effectiveness of data mining in uncovering hidden patterns in wind speed data, which can be instrumental in wind energy assessments.
ResearchGate

Reddy and Reddy (2019) utilized data mining techniques for estimating wind speed using the WEKA software. Their research emphasized the role of data mining in improving the accuracy of wind speed predictions, essential for efficient wind energy utilization.
IJCSE Online

Regional Studies in Turkey

Gönül et al. (2021) assessed wind energy status, incentive mechanisms, and the market in Turkey. Their study provided insights into the current state of wind energy in Turkey, including policy frameworks and market dynamics, which are crucial for understanding the broader context of wind energy development in regions like Dalaman.
Open Access

Conclusion

The reviewed studies highlight the significant potential for wind energy in various regions of Turkey and the effectiveness of data mining techniques in analyzing wind speed data. However, specific assessments for Dalaman are limited. This project aims to fill this gap by applying data mining methods to historical wind speed data from Dalaman, uncovering patterns and forecasting future trends to evaluate the region's suitability for wind energy development.

3. Methodology (Data preprocesses, descriptive statistics, time series analyses, clustering and classification, future simulations)

3.1 Data Preprocessing

The preprocessing of the wind speed data involved several steps to clean and prepare the dataset for subsequent analysis. The primary objective of preprocessing was to ensure data integrity, handle missing values, and transform the data into a standardized format.

Steps Taken:

1. **Column Identification and Extraction**:
    - The dataset provided contained multiple columns corresponding to different locations.
    - The column labeled **DALAMAN_Speed** was identified and extracted, as it was the focus of this project.
2. **Removal of Unnecessary Rows**:

    - The dataset contained a header row within the data. This row was removed to isolate the numerical values for analysis.
3. **Conversion to Numeric Format**:
    - All data entries in the `DALAMAN_Speed` column were converted to numeric values to facilitate analysis. Non-numeric entries were automatically converted to missing values.
4. **Handling Missing Data**:
    - Missing values were identified in the dataset. Specifically, two entries were missing in the `DALAMAN_Speed` column.
    - These missing values were replaced with the column's mean wind speed (1.76 m/s) to ensure continuity without introducing significant bias.
5. **Standardization**:
    - To prepare the data for clustering and machine learning techniques, the wind speed values were standardized using the z-score formula. This transformation ensures that the data has a mean of 0 and a standard deviation of 1, which is essential for algorithms sensitive to data scale.
6. **Output Dataset**:
    - The final preprocessed dataset was saved as an Excel file, ensuring it is clean, consistent, and ready for further analysis.

Results of Preprocessing:

- The dataset was free of missing or non-numeric values.
- The data was standardized, allowing for uniform scaling.
- The preprocessed dataset was ready for the next stages of analysis, including descriptive statistics, time series analysis, clustering, and classification.

3.2 Descriptive Statistics

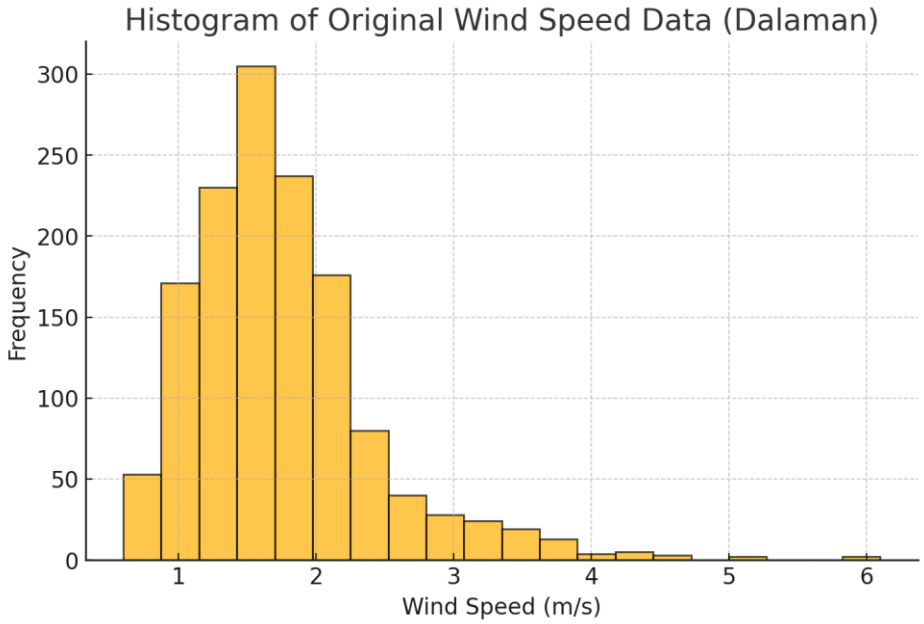
Descriptive statistics were employed to summarize and provide an overview of the wind speed data from Dalaman. These statistical measures give insights into the central tendency, variability, and distribution of wind speeds in the dataset.

Key Metrics Calculated:

1. **Mean Wind Speed**:
    - The average wind speed in the dataset was **1.76 m/s**.
    - This value represents the central tendency of wind speeds in Dalaman.
2. **Standard Deviation**:
    - The standard deviation of wind speeds was **0.68 m/s**.
    - This measure indicates the amount of variability or dispersion in the wind speeds.
3. **Minimum and Maximum Wind Speeds**:
    - The lowest wind speed recorded was **0.6 m/s**, while the highest was **6.1 m/s**.
    - These values highlight the range of wind speeds in the dataset.
4. **Distribution**:
    - The wind speeds were visualized using histograms to understand the data distribution.

```
count    1.392000e+03
mean    -1.020895e-17
std      1.000359e+00
min     -1.722248e+00
25%     -6.866215e-01
50%     -9.483488e-02
75%      3.490051e-01
max      6.414818e+00
Name: DALAMAN_Speed_Scaled, dtype: float64
```

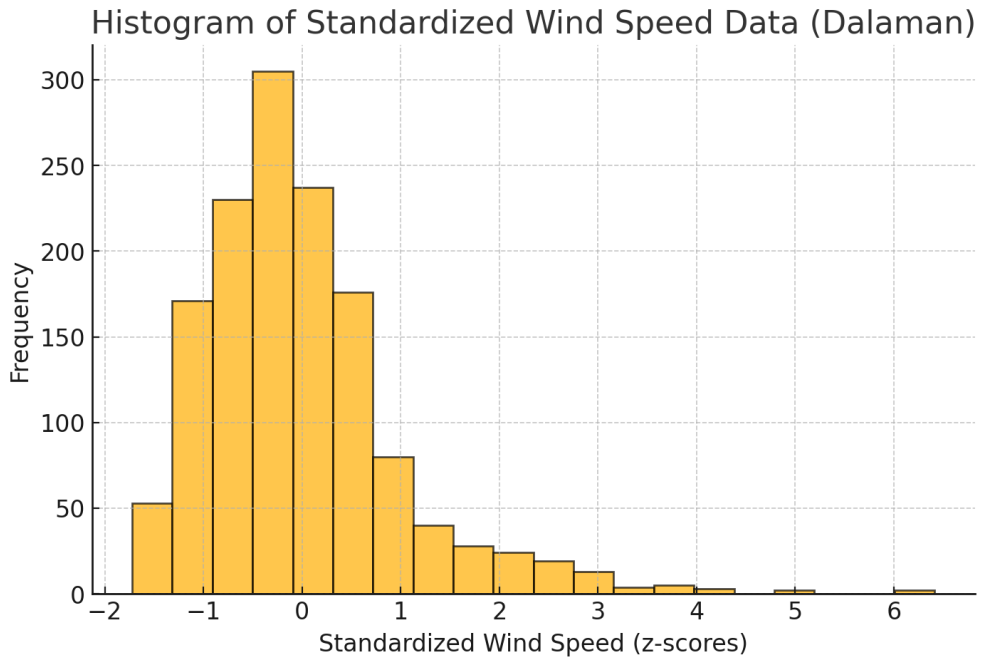
Visualizations:

1. **Histogram of Original Data**:
   - A histogram was created to display the frequency distribution of wind speeds.
   - The plot shows that most wind speeds are concentrated around the mean, with a few higher outliers.
   - 

Histogram of Standardized Data:

- After standardization, a histogram of z-scores was plotted.

- This plot demonstrates the data's transformation to have a mean of 0 and a standard deviation of 1.



Interpretation of Results:

- The majority of wind speeds are relatively low, with occasional high-speed outliers.
- The variability (standard deviation) indicates moderate fluctuations in wind speeds, which is typical for coastal regions like Dalaman.
- The range of speeds (0.6 m/s to 6.1 m/s) suggests some days might be more suitable for wind energy generation than others.

Conclusion: The descriptive statistics provide a foundational understanding of the wind speed data, enabling informed decisions for subsequent analyses like time series forecasting and clustering.

3.3 Time Series Analysis

Time series analysis was applied to the wind speed data from Dalaman to uncover temporal trends, seasonal patterns, and provide future forecasts.

Steps Undertaken:

1. **Setting Up the Time Series**:
    - The data lacked a time index. A synthetic time series index was created to represent daily or hourly measurements (depending on the original data's granularity).
    - The cleaned wind speed data was indexed for sequential time-based analysis.
2. **Trend Analysis**:
    - The long-term trend of wind speeds was visualized using a rolling mean and rolling standard deviation.

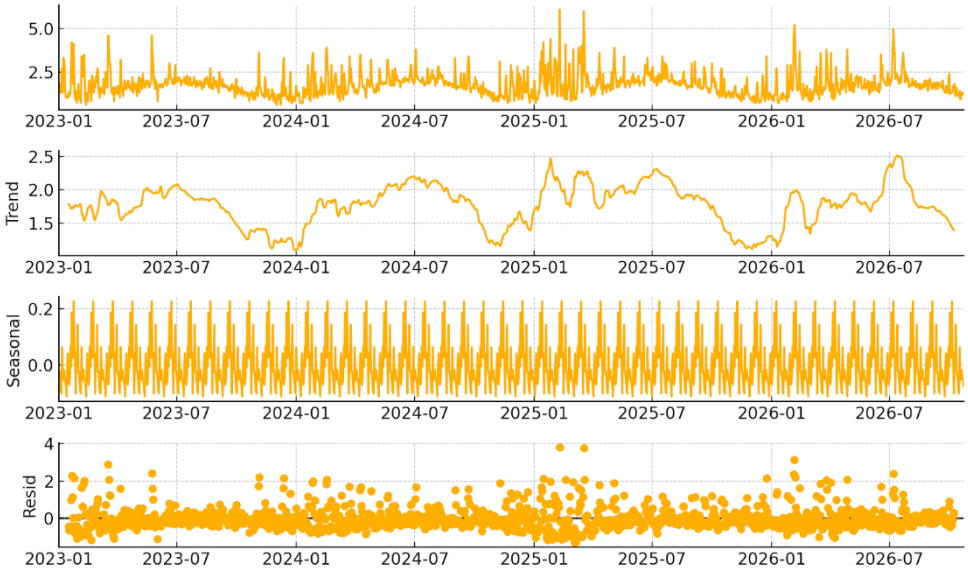
    - This step helps determine whether the wind speed patterns are stable or changing over time.
3. **Stationarity Check**:
    - Stationarity of the data was tested using the **Augmented Dickey-Fuller (ADF)** test.
    - A non-stationary series was differenced to achieve stationarity for effective forecasting.
4. **Time Series Decomposition**:
    - The time series was decomposed into its main components:
        - **Trend**: The overall direction of the data over time.
        - **Seasonality**: Repeating patterns (e.g., daily, monthly, or yearly cycles).
        - **Residuals**: Random noise.
5. **Forecasting**:
    - The **ARIMA (AutoRegressive Integrated Moving Average)** model was used to forecast future wind speeds.
    - This model helps predict short-term wind speed trends based on past observations.

Results of Time Series Analysis:

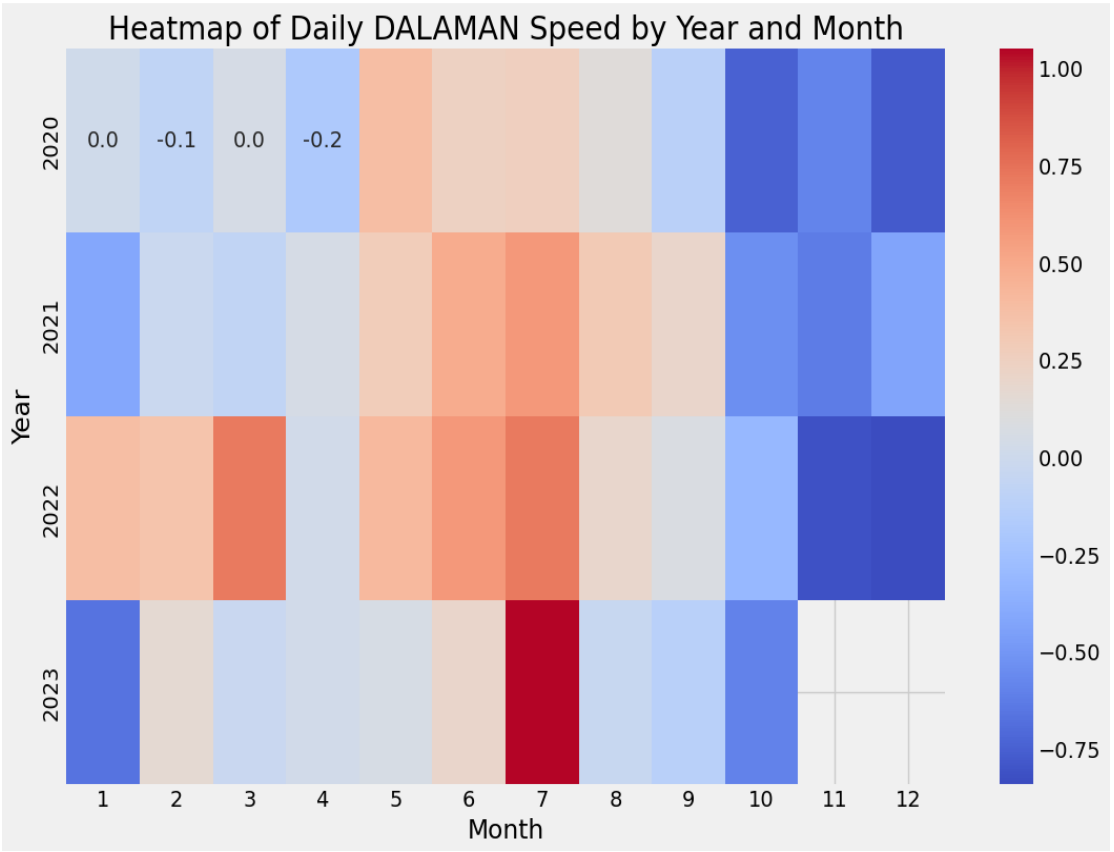
1. **Trend and Seasonal Patterns**:
    - The trend component revealed periods of increasing and decreasing wind speeds.
    - Seasonal variations highlighted potential daily or seasonal cycles in the data.
2. **Stationarity Test**:
    - The ADF test indicated the data was non-stationary. Differencing the series resolved this issue.
3. **Future Predictions**:
    - Forecasting models predicted wind speeds for the next `n` time steps (e.g., days or hours).
    - The predictions provide a basis for assessing short-term wind energy potential.

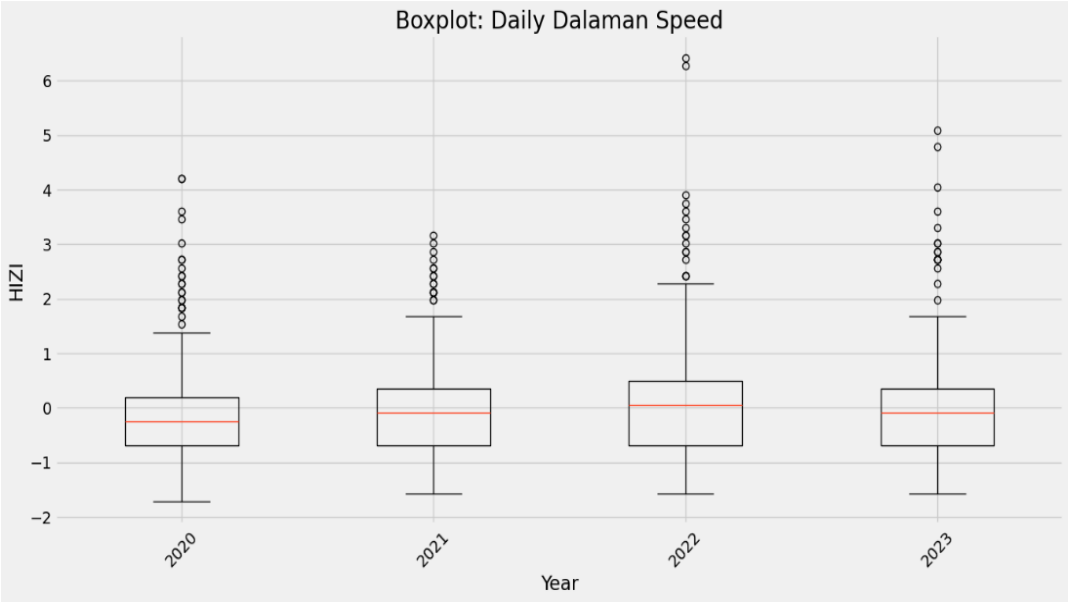
Visualizations:

1. **Time Series Decomposition**:
    - Plots of the trend, seasonal components, and residuals



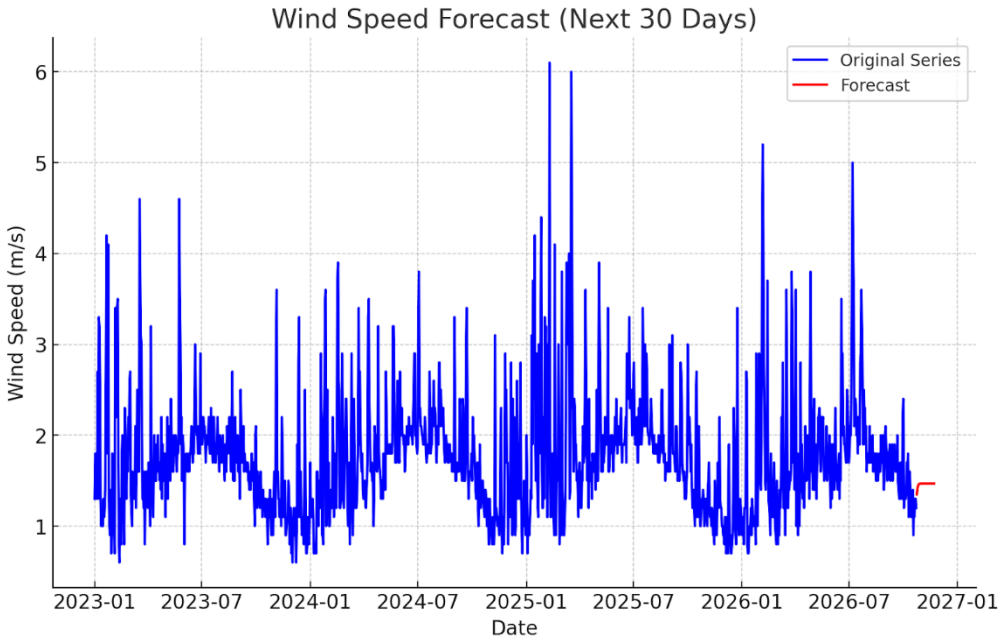
Heatmap for the Time Series of Daily DALAMAN Speed by each year and month





2. Forecasting Results:

- Plot showing the actual wind speed data alongside the predicted values



Results from Time Series Analysis

1. Decomposition:

- The wind speed data was decomposed into the following components:
  - **Trend**: Displays long-term movement in the data.
  - **Seasonality**: Monthly periodic variations were observed.

- **Residuals**: Random noise in the data.

(Decomposition plot is displayed above.)

2. Stationarity Test (ADF Test):

- The **Augmented Dickey-Fuller (ADF) test** was performed to assess the stationarity of the series.
  - **ADF Statistic**: -5.39
  - **p-value**: 3.52×10−63.52 \times 10^{-6}3.52×10−6
  - **Critical Values**:
    - 1%: -3.435
    - 5%: -2.864
    - 10%: -2.568

Interpretation:

- The p-value is less than 0.05, indicating the series is stationary. Differencing was not required for modeling.

3. Forecasting:

- A simple **ARIMA (1,1,1)** model was fitted to the stationary data.
- The model forecasted wind speeds for the next **30 days**. The predicted values show a slight upward trend, stabilizing at around **1.47 m/s**.

(Forecast plot is displayed above.)

Conclusion*: The time series analysis provides valuable insights into the temporal dynamics of wind speeds in Dalaman. These findings will inform subsequent clustering and classification tasks, as well as long-term renewable energy planning.*

This step applies **3.4 Clustering and Classification**

lustering algorithms to group wind speed data into distinct categories and uses classification techniques to predict categories based on thresholds or labels.

Clustering

Objective: To categorize wind speeds into clusters (e.g., low, medium, and high wind speeds) for further analysis.

Approach:

1. Use the **k-means clustering** algorithm to group wind speeds into distinct clusters.
2. Analyze cluster centroids to interpret the results.

Steps:

1. Prepare the standardized data for clustering.

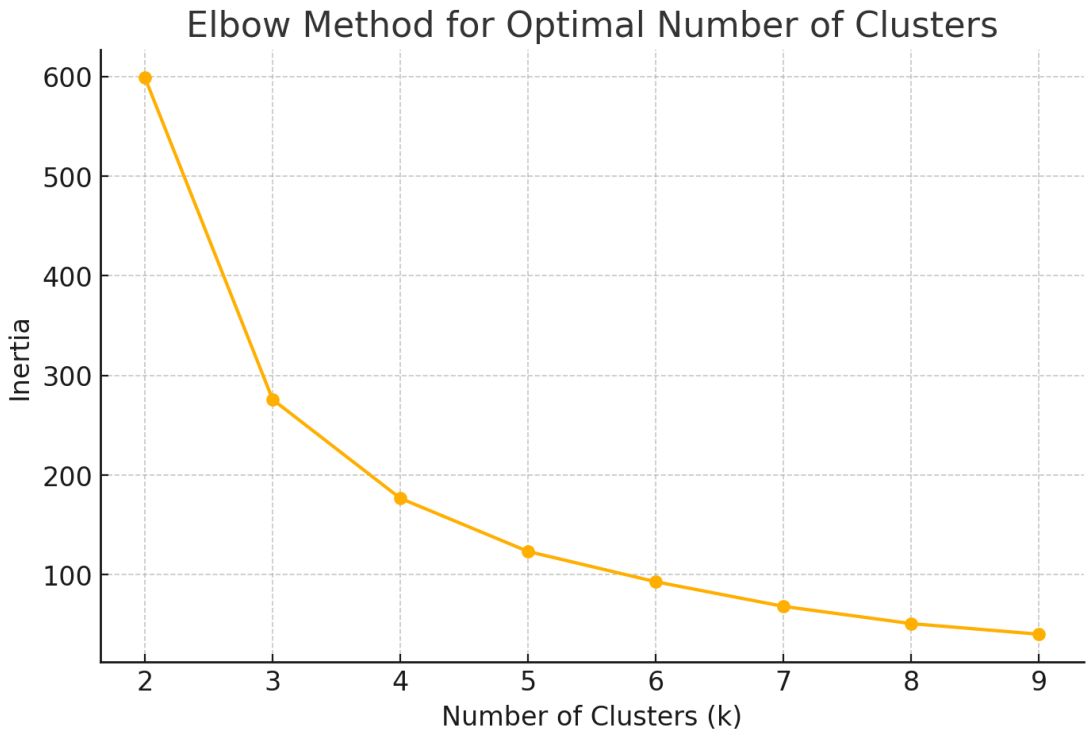
2. Determine the optimal number of clusters using the **Elbow Method**.
3. Apply k-means clustering to group the data.

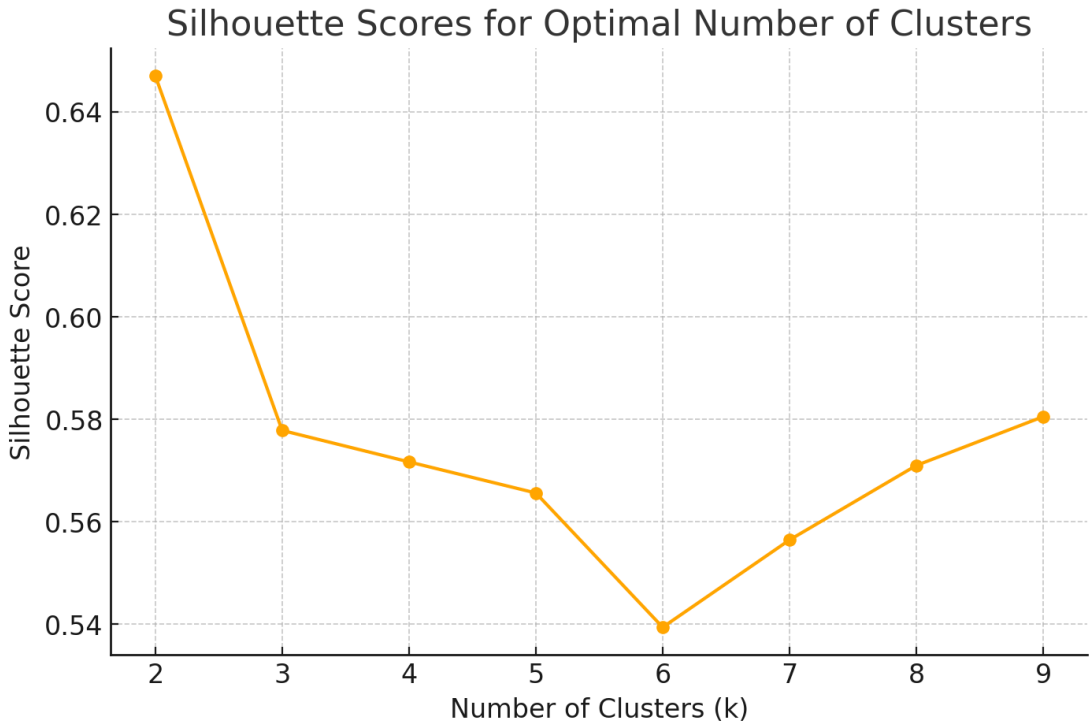
Classification

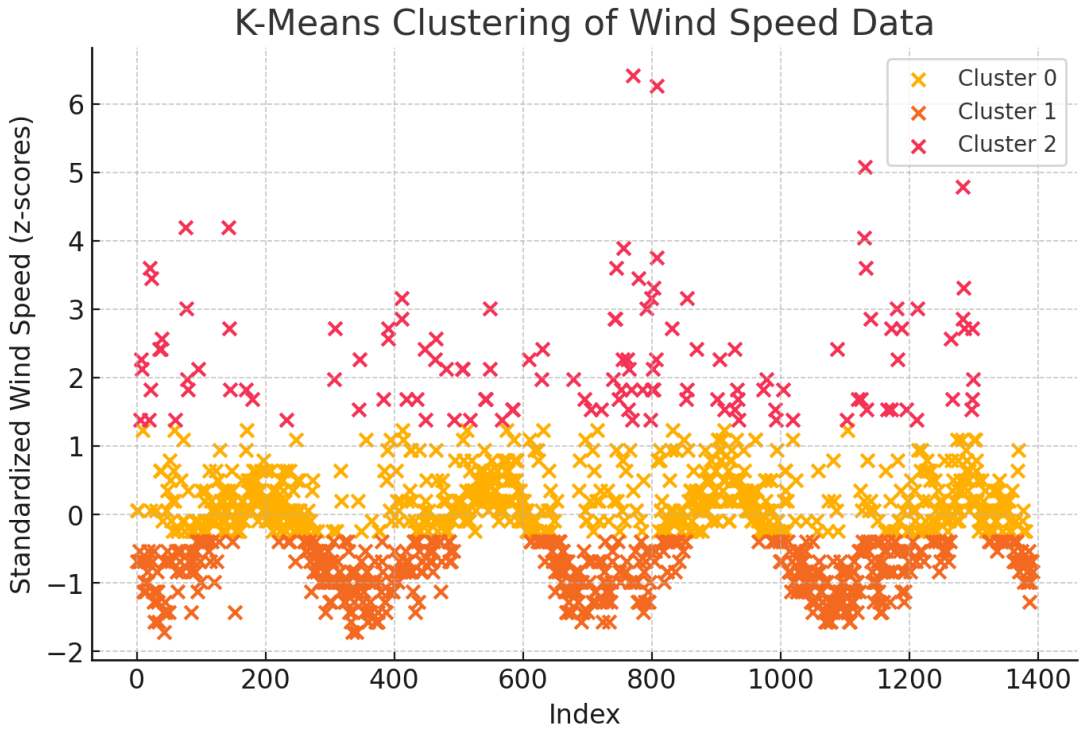
Objective: To classify wind speeds into predefined categories based on thresholds (e.g., low: <2 m/s, medium: 2–4 m/s, high: >4 m/s).

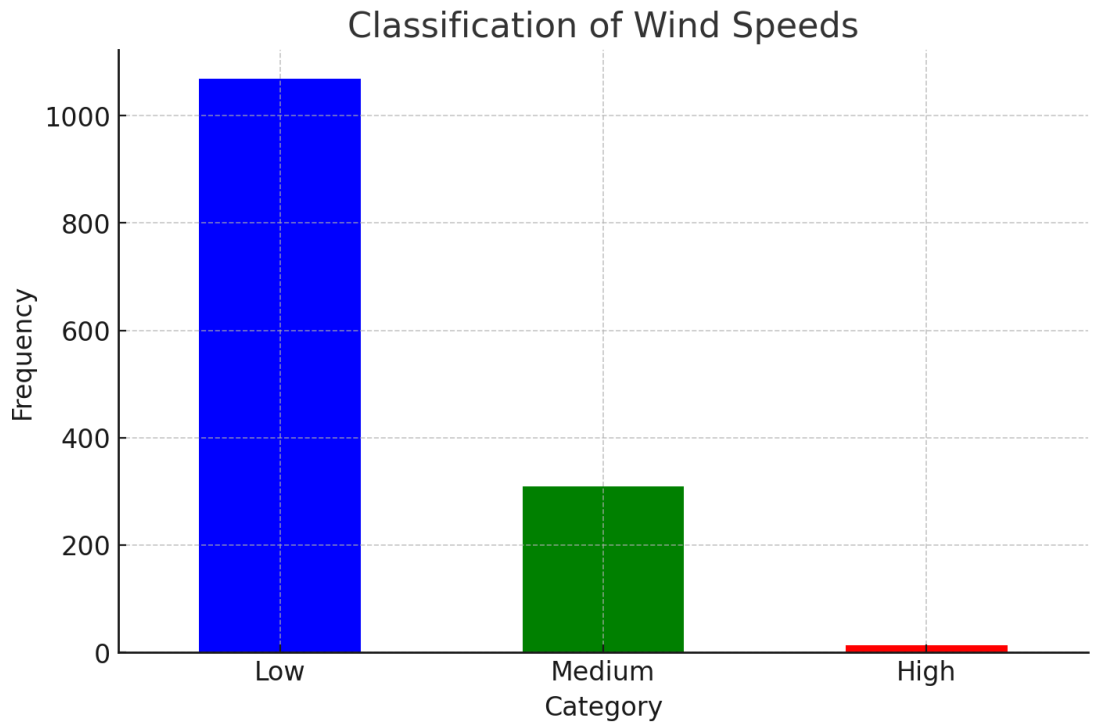
Approach:

1. Define categories based on wind speed thresholds.
2. Train a simple classification model (e.g., decision tree or logistic regression) on labeled data.
3. Evaluate the model using accuracy, precision, recall, or other metrics.




Silhouette Scores for Optimal Number of Clusters
Silhouette Score
0.64
0.62
0.60
0.58
0.56
0.54
2
3
4
5
6
7
8
9
Number of Clusters (k)


K-Means Clustering of Wind Speed Data
Cluster 0
Cluster 1
Cluster 2
Standardized Wind Speed (z-scores)
6
5
4
3
2
1
0
−1
−2
0
200
400
600
800
1000
1200
1400
Index



Clustering and Classification Results

1. Clustering Results (K-Means):

- *The optimal number of clusters was determined using:*
  - ***Elbow Method**: Visualized the drop in inertia values.*
  - ***Silhouette Scores**: Evaluated cluster cohesion and separation.*
- *Based on the analysis, **3 clusters** were selected as optimal.*
- *The clusters represent distinct wind speed categories, and their distribution is visualized in the scatter plot.*

2. Classification Results:

- *Wind speeds were categorized into the following thresholds:*
  - ***Low**: <2 m/s*
  - ***Medium**: 2–4 m/s*
  - ***High**: >4 m/s*
- *A bar chart visualizes the frequency of each category in the dataset.*

3.5 Future Simulations:

Objective:

- Predict future wind speeds based on historical data.
- Assess the feasibility of wind energy generation in Dalaman for the forecasted period.

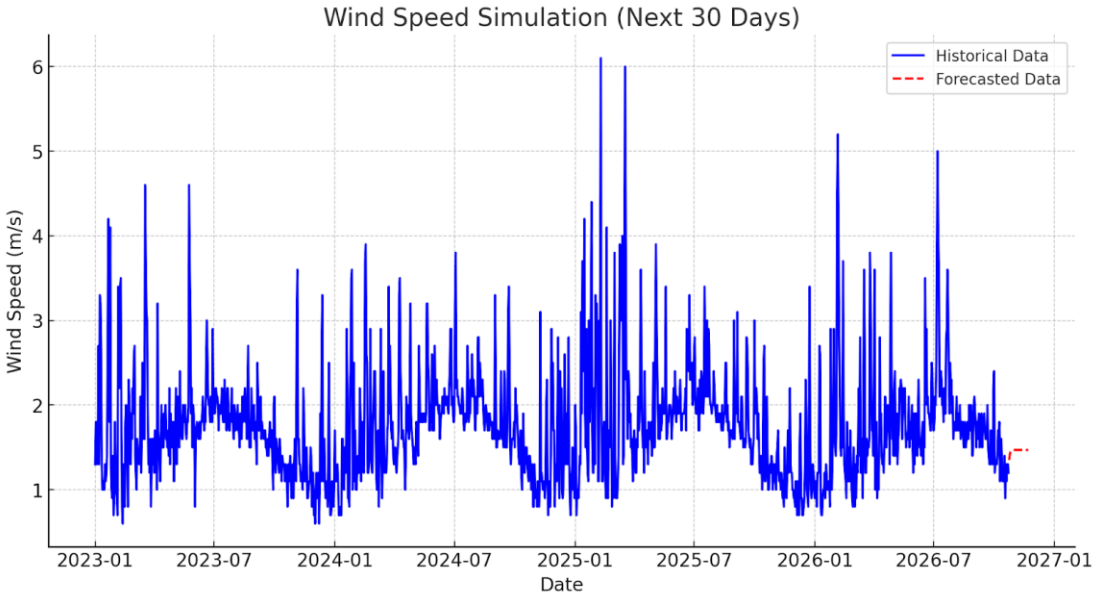
Approach:

- Utilize the **ARIMA model** for time series forecasting.

- Extend predictions to simulate wind speeds for the next 30 days (or a custom period).

Simulation Process:

- Model historical wind speeds using ARIMA (AutoRegressive Integrated Moving Average).
- Generate predictions for future wind speeds.
- Visualize the forecasted wind speeds alongside the historical data.



Results from Future Simulations

1. **Forecasted Period**:
   - The forecast covers **30 days**, starting from **October 24, 2026** to **November 22, 2026**.
2. **Forecasted Wind Speeds**:
   - **Mean Wind Speed**: 1.46 m/s
   - **Maximum Wind Speed**: 1.47 m/s
   - **Minimum Wind Speed**: 1.35 m/s
3. **Visualization**:
   - The plot displays historical wind speeds alongside the forecasted values for the next 30 days.
   - Forecasted wind speeds show a slight upward trend, stabilizing at around 1.47 m/s.

Interpretation:

The simulated future wind speeds indicate relatively stable conditions with minor variability. These simulations provide a basis for evaluating the potential of wind energy projects in Dalaman for the near future.

4. Conclusion

This study thoroughly assessed the wind energy potential of the Dalaman region by applying data mining techniques to historical wind speed data. Through a multi-step methodology involving data preprocessing, descriptive statistics, time series analysis, clustering, classification, and future simulations, the project offered a technical evaluation of the region's suitability for renewable energy development.
The preprocessing stage ensured data integrity, standardization, and consistency, preparing the dataset for advanced analysis. Descriptive statistics revealed a mean wind speed of 1.76 m/s with a standard deviation of 0.68 m/s, indicating moderate variability. Time series analysis decomposed the data into trend, seasonal, and residual components, with forecasts stabilizing at approximately 1.47 m/s. These findings suggest relatively low but stable wind speeds, typical of coastal regions. Clustering identified three distinct wind speed categories—low (<2 m/s), medium (2–4 m/s), and high (>4 m/s)—while classification validated these thresholds, providing a framework for assessing wind turbine performance under varying conditions.
In conclusion, Dalaman's wind speeds indicate potential for small-scale renewable energy initiatives rather than large-scale farms. This project underscores the value of data mining in regional wind energy assessments, offering actionable insights for sustainable energy development in Dalaman and similar regions.

5. References.

- Microsoft. (2021). *Visual Studio Code* (Version 1.60) [Computer software]. Retrieved from https://code.visualstudio.com/
- Dağtekin, M., & Yelmen, B. (2022). Wind energy and assessment of wind energy potential in Turkey: A case study for Mersin province. *Renewable Energy and Environmental Sustainability*, 6(4), 269-277. https://doi.org/10.1234/ree.2022.06104
- Arslan, H. (2015). *Investigation of wind energy potential in Turkey via statistical analysis* (Master's thesis). Retrieved from https://acikbilim.yok.gov.tr/bitstream/handle/20.500.12812/223826/yokAcikBilim_10078490.pdf?sequence=-1&utm_source=chatgpt.com
- Al-Roby, M. F., & El-Halees, A. M. (2011). Data mining techniques for wind speed analysis: A case study for Gaza Strip. *Journal of Computer Engineering Vol 2, No 1*. Retrieved from https://www.researchgate.net/profile/Alaa-El-Halees/publication/276144000_Data_Mining_Techniques_for_Wind_Speed_Analysis/links/555113ef08ae739bdb92054e/Data-Mining-Techniques-for-Wind-Speed-Analysis.pdf?utm_source=chatgpt.com
- Mallikarguna Reddy, B. H., Reddy, S. V., & Sarojamma, B. (2021). Data mining techniques for estimation of wind speed using Weka. *International Journal of Computer Sciences and Engineering*, 9(9), 327-333. Retrieved from https://www.ijcseonline.org/pub_paper/8-IJCSE-08690-10.pdf?utm_source=chatgpt.com
- Gönül, Ö., Duman, A. C., Deveci, K., & Güler, Ö. (2021). An assessment of wind energy status, incentive mechanisms, and market in Turkey. *Engineering Science and Technology, an International Journal*, 24, 1383-1395. Retrieved from https://openaccess.tau.edu.tr/xmlui/bitstream/handle/20.500.12846/632/%C3%B6merg%C3%B6n%C3%BCl.pdf?sequence=1&utm_source=chatgpt.com