Air Quality Trends

TEAM 1

FINAL PROJECT

TIME SERIES ANALYSIS AND FORECASTING

Table of Contents

Roadmap for Key Sections and Topics of Discussion

Introduction

PROBLEM STATEMENT,
PROJECT OBJECTIVE,
CONTEXT SETTING,
AND ASSUMPTIONS

Problem Statement

"Rising benzene levels in the atmosphere present a growing public health concern!"

Group 1 Carcinogen:
- Highest Level of Carcinogenicity

Blood Disorders:
- Chronic Exposure to Benzene Leads to Blood Disorders, Including but Not Limited to Leukemia and Aplastic Anemia

Reproductive and Developmental Toxicity:
- Exposure Is Linked to Increased Risk of Miscarriage and Fetal Development Disorders

Liver and Kidney Damage:
- Prolonged Exposure Leads to Impaired Liver and Kidney Functions

Project Objective

- Develop a Time Series Model to Accurately Forecast Benzene Concentrations
- Provide Insights That Alert Authorities to Take Timely Measures to Reduce Exposure and Protect Public Health

Regulatory Guidelines

- World Health Organization (WHO) Guideline for Benzene in Ambient Air:

  10 $\mu g/m^3$ as an annual average

- European Union (EU) Directive:

  5 $\mu g/m^3$ as an annual average

Project Context

13 Months

03/10/2004 — 4/4/2005

- The Dataset Originates From an Urban Area in Italy
- 13 Months of Hourly Data From a Gas Multi-Sensor Device

Forecast Analysis of Benzene Levels Allows For:

1. Identification of Pollution Sources
2. Implementation of Mitigation Strategies

Results Can Guide Policy Decisions, Public Health Strategies, and Timely Warnings to the Public

Assumptions

DATA

- Stationarity
  - Differencing on Seasonal Lag + Trend Has Made the Data Stationary
- Sensor Accuracy
  - The Device Used for Data Collection Is Accurate and Reliable

MODEL

- Train-Test Data
  - 12 Months Training Data
  - 1 Month Test Data
- External Factors
  - No External Effects

REAL-WORLD

- Regulation & Standards
  - WHO
  - European Union

Data

DATA OVERVIEW, PRE-PROCESSING, DATASET CHARACTERISTICS AND DATA DECOMPOSITION

Data Overview

Source: UCI Machine Learning Repository, "Air Quality" Dataset

Timeframe: March 2004 to February 2005

Observations: 9,357 Hourly Records From an Italian Urban Area

Features: 15 Variables, Including Gas Concentrations (CO, NMHC, C6H6, NOx, NO2, O3) and Environmental Conditions (Temperature, Relative Humidity, Absolute Humidity)

Objective: Forecast Daily Average Benzene (C6H6) Concentration Using Various Time Series Models to Identify the Most Accurate Approach

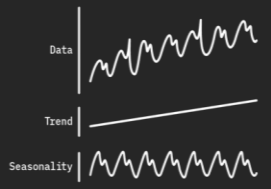
Pre-Processing Methodology

Missing Values

- Linear Interpolation for Imputation

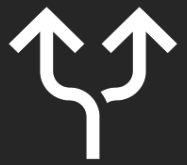
Data Aggregation

- Daily Averages (Downsampling)
- De-Noising



Stationarity

- ADF Test
- KPSS Test
- Differencing
- Seasonal Differencing

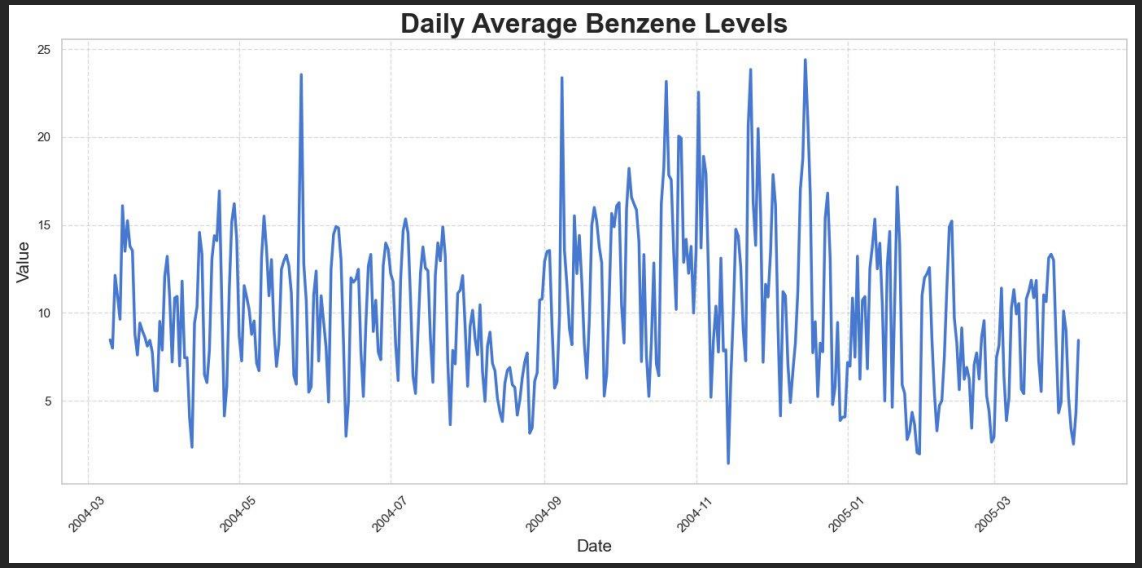


Data Split

- 12 Months Train Data
- 1 Month Test Data

Dataset Characteristics

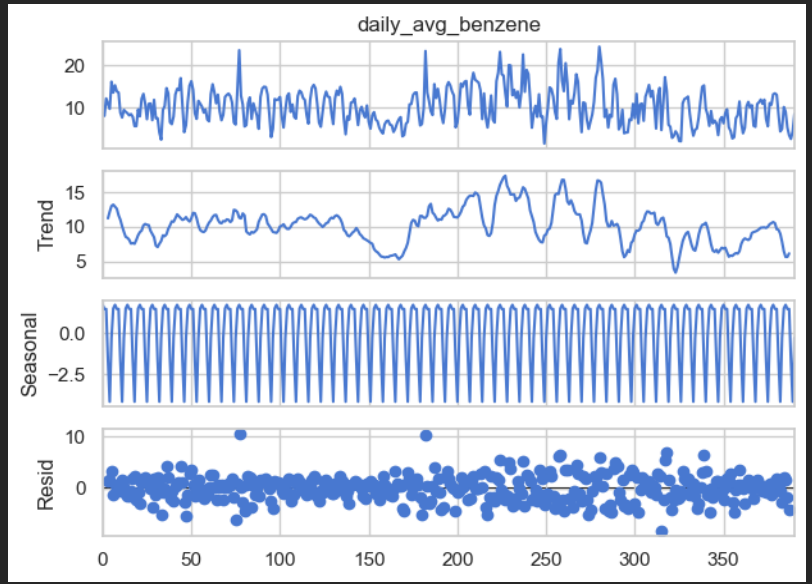
- Non-Stationary Data
- Inconsistent Variance
- Fluctuation Magnitude: Additive
- Systematic Components: Imperfect Recurrence + Consistency



Dataset Decomposition

Decomposition

- Type: Additive
- Period: 7
- Results: The Data Is Effectively Separated Into Its Trend, Seasonal, and Residual Components



Residuals

- Stationary Residuals
- Decomposition Was Able to Remove Trend and Seasonality From the Data

Residuals Analysis	
Metric	Result
ADF	Stationary
KPSS	Stationary

Experimental Results and Analysis

ALGORITHMS
AND RESULTS

Modeling Strategy: Overview

Objective

Develop and Compare 3 Models to Forecast Benzene Concentrations, Identifying and Selecting the Most Accurate and Reliable Method

Benchmark Model

The Seasonal Naïve Model Serves as the Baseline for Comparison to Make Comparison Against Higher Complexity Models

Model Evaluation

Use Information Criteria to Select Best Model of Each Model Type and Evaluate Forecasts With Mean Relative Error (MRE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) As Performance Metrics

Modeling Algorithms Applied

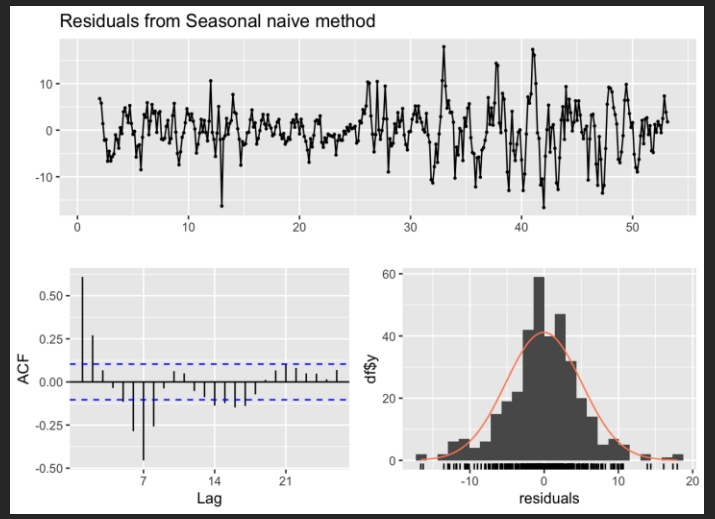
- Seasonal Naïve Model (Benchmark)
- Exponential Smoothing (ETS)
- Autoregressive Fractional Integrated Moving Average (ARFIMA)
- Neural Network Autoregression (NNAR)

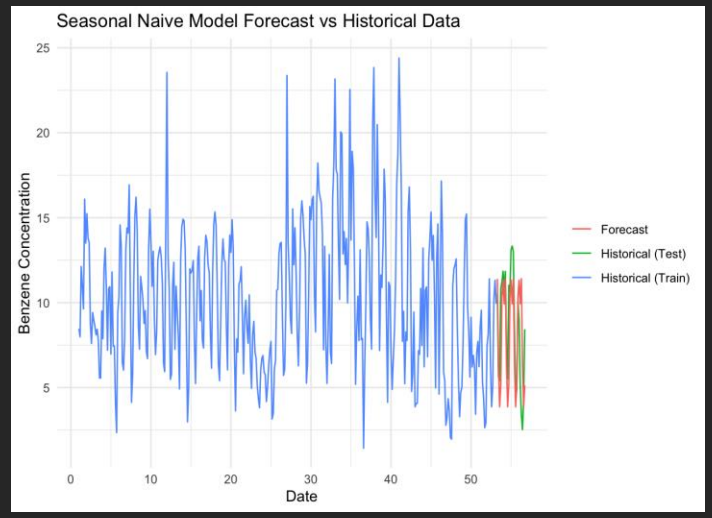
Seasonal Naïve Model

Benchmark

- **Model Overview:** The Seasonal Naïve Model Assumes That the Future Value Will Be the Same as the Value From the Previous Season (e.g., Last Year's Value for the Same Month/Day)
- **Purpose:** Used as a Simple Baseline to Compare Against More Complex Models

Residuals Analysis	
Metric	Result
Ljung-Box (Autocorrelation – Y/N)	Y
Shapiro-Wilk (Normality – Y/N)	N





- **Residuals Do NOT Resemble White Noise**
- **Significant Lags in ACF**
- Distribution Is NOT Normal
- Forecast Somewhat Resembles Test Data

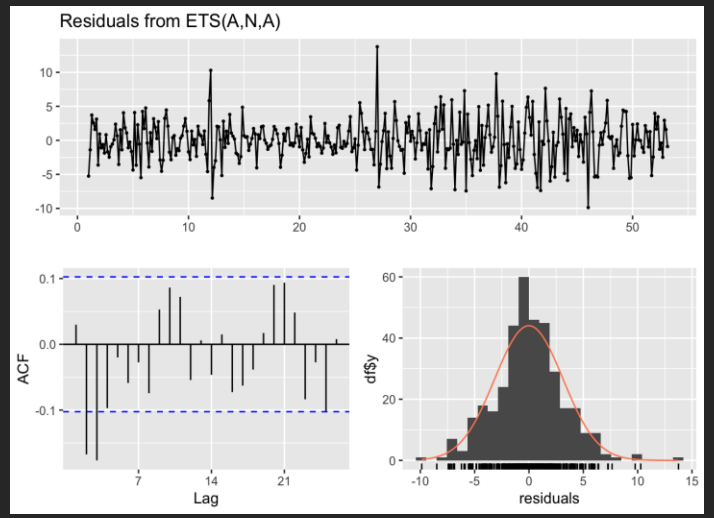
1. ETS

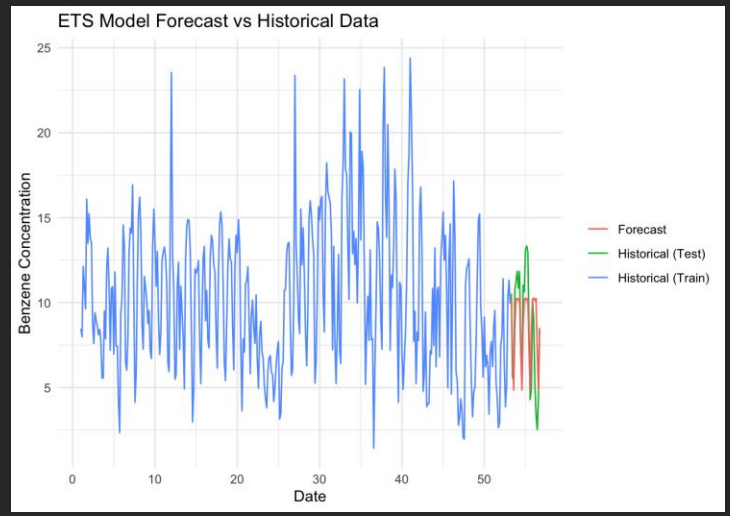
Exponential Smoothing

- **Model Overview**: A Family of Models That Forecast by Smoothing Past Observations With Exponentially Decreasing Weights, Handling Various Trends and Seasonal Patterns
- **Purpose**: Adaptively Forecasts Time Series With Trends and Seasonality

Residuals Analysis	
Metric	Result
Ljung-Box (Autocorrelation – Y/N)	Y
Shapiro-Wilk (Normality – Y/N)	N

ETS(A, N, A)





- Residuals Do NOT Resemble White Noise
- Significant Lags in ACF
- Distribution Is NOT Normal
- Forecast Somewhat Resembles Test Data

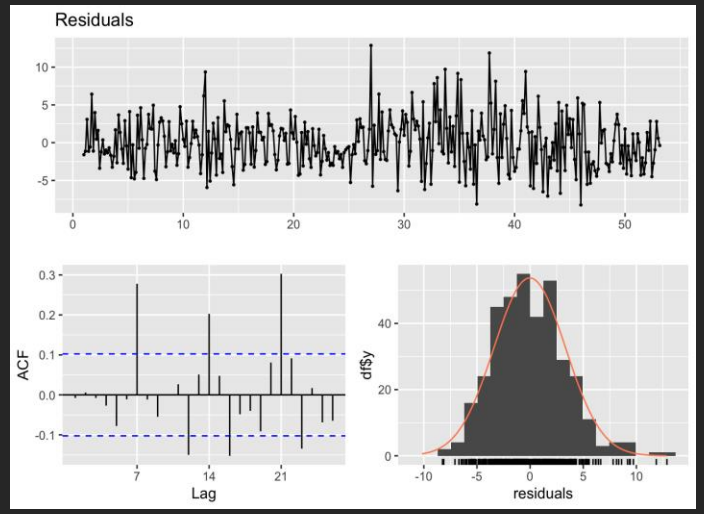
2. ARFIMA

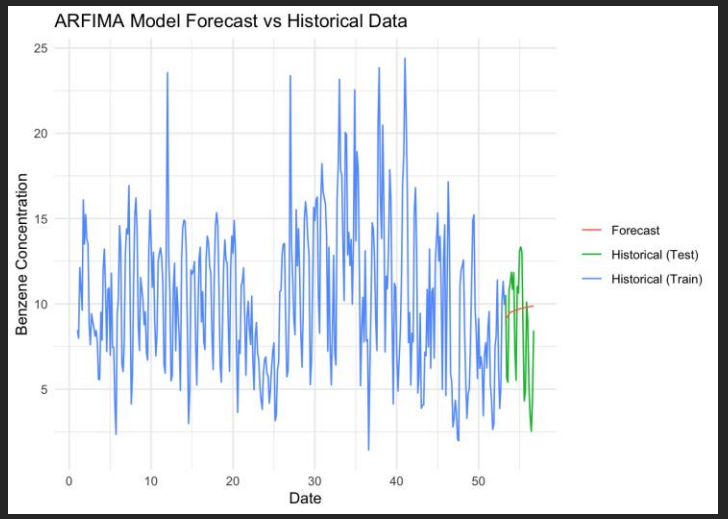
Autoregressive Fractional Integrated Moving Average

- **Model Overview**: An Extension of Arima That Allows for Fractional Differencing, Making It Suitable for Modeling Long-Memory Processes With Slow-Decaying Autocorrelations
- **Purpose**: Models Time Series With Long-Term Dependencies

Residuals Analysis	
Metric	Result
Ljung-Box (Autocorrelation – Y/N)	Y
Shapiro-Wilk (Normality – Y/N)	N

ARFIMA(2, 0.17, 0)





- **Residuals Do NOT Resemble White Noise**
- **Significant Lags in ACF**
- Distribution Is NOT Normal
- Forecast Does NOT Resemble Test Data

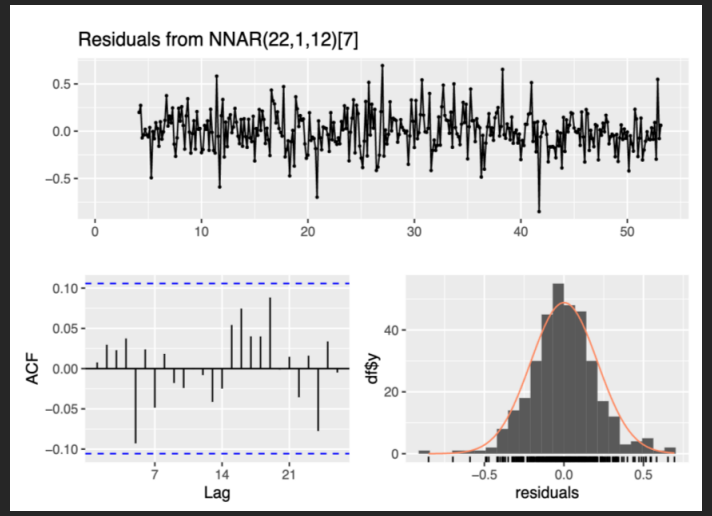
3. NNAR

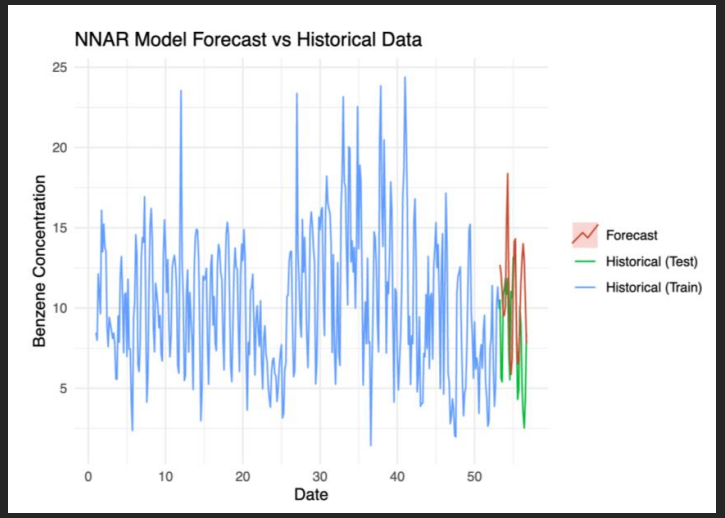
Neural Network Auto Regression

- **Model Overview**: A Neural Network Model That Combines Autoregressive Lags With a Neural Network Structure to Capture Non-linear Relationships in Time Series Data
- **Purpose**: Captures Complex, Non-Linear Patterns in Time Series

Residuals Analysis	
Metric	Result
Ljung-Box (Autocorrelation – Y/N)	N
Shapiro-Wilk (Normality – Y/N)	N

NNAR(22, 1, 12)[7]





- **Residuals Mostly Averaging Around Mean of Zero**
- **Residuals Resemble White Noise**
- Distribution Is NOT Normal
- Forecast Resembles Test Data in Some Areas

Model Selection and Derived Solution

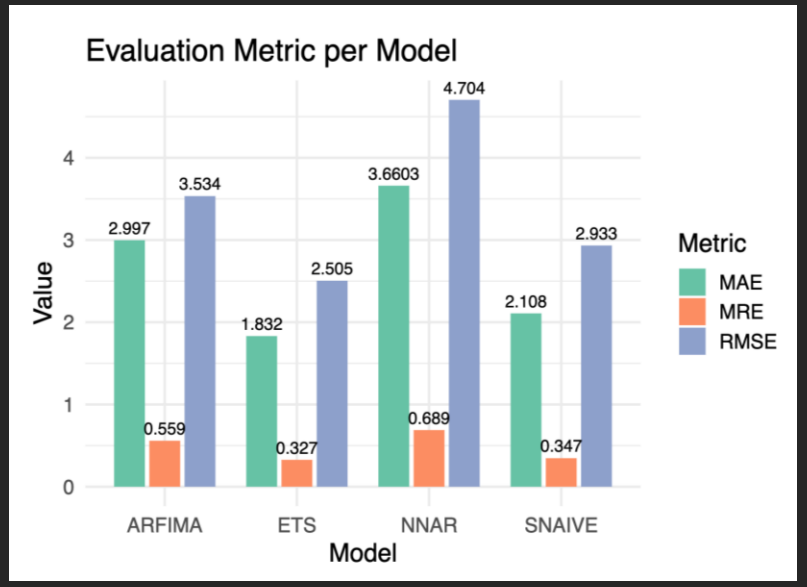
CHOSEN MODEL AND FORECASTING EVALUATION

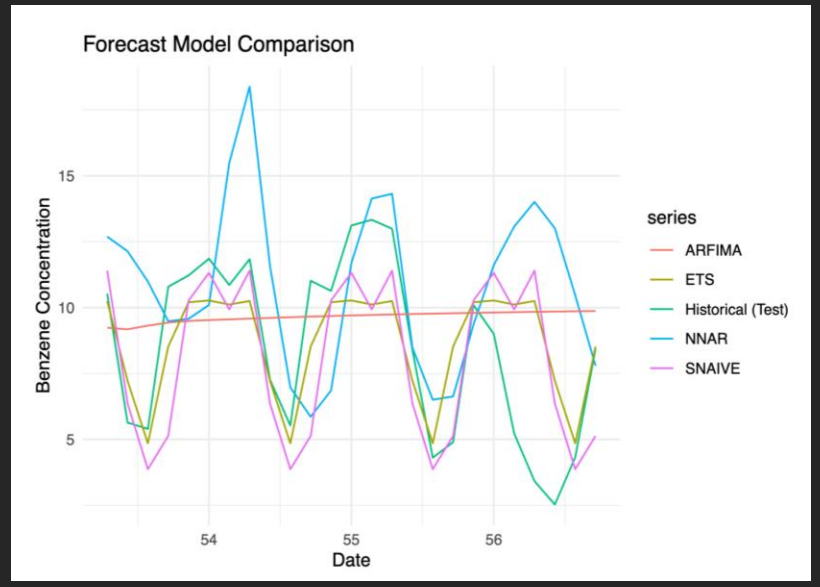
Results: Algorithm Evaluations

	Forecast Evaluation			Residuals Analysis	
Model	MAE	MRE	RMSE	Ljung-Box (Presence of Autocorrelation – Y/N)	Shapiro-Wilk (Normality – Y/N)
Seasonal Naïve	2.108	0.347	2.933	Y	N
ETS	**1.832**	**0.327**	**2.505**	Y	N
ARFIMA	2.997	0.559	3.534	Y	N
NNAR	3.6603	0.689	4.704	N	N

- Best Forecast: Exponential Smoothing (ETS)
- Best Model per Residuals: Neural Network Auto Regression (NNAR)

Model Comparison





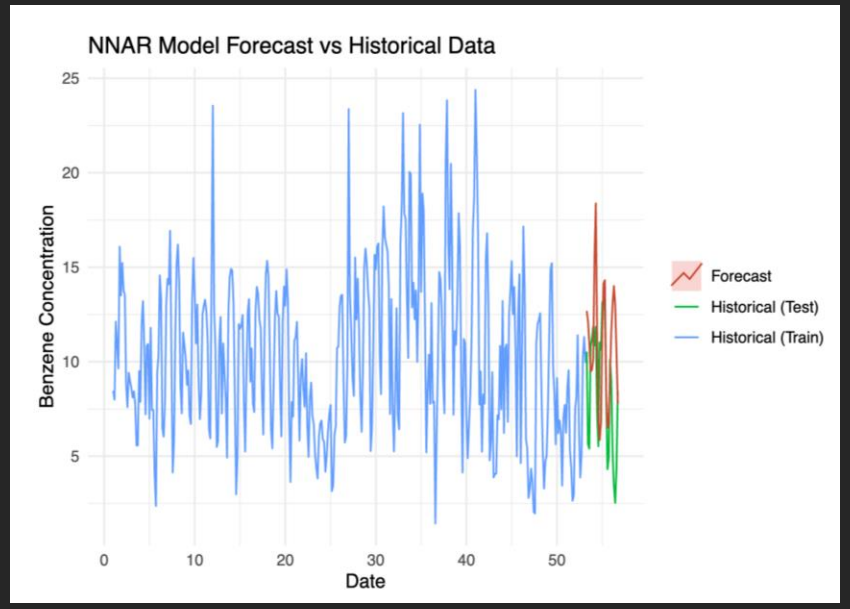
- ETS Has the Lowest Errors, but Is Insufficient Because It Does Not Capture All the Autocorrelations of the Underlying Data
- NNAR's Forecasts Are in Line in Some Areas of the Forecast, but Struggles When the Actual Benzene Concentrations Are Low

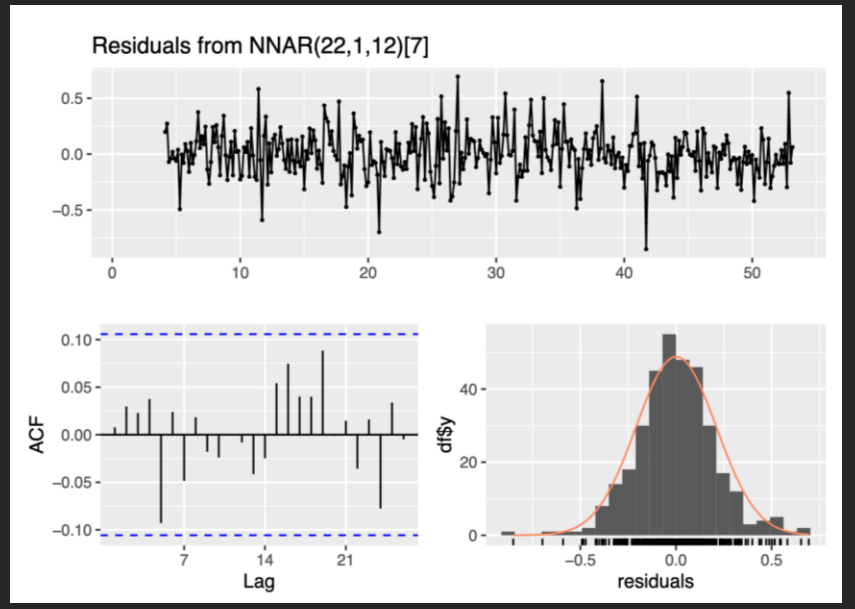
Final Model Selection

Chosen Model:
NNAR (Neural Network Autoregressive)

- Reasoning:
  - Modeling Capabilities: Ability to Capture Non-Linear Relationships in the Data
  - Seasonality: Captures Seasonal Patterns in the Benzene Concentration, Leveraging the Ability to Model Complex Dependencies
  - Model Complexity: Captured All Autocorrelations That Other Models Were Unable to Capture
  - Residual Analysis: Residuals Indicate Model Captured All Autocorrelations

Final Model Selection, Continued



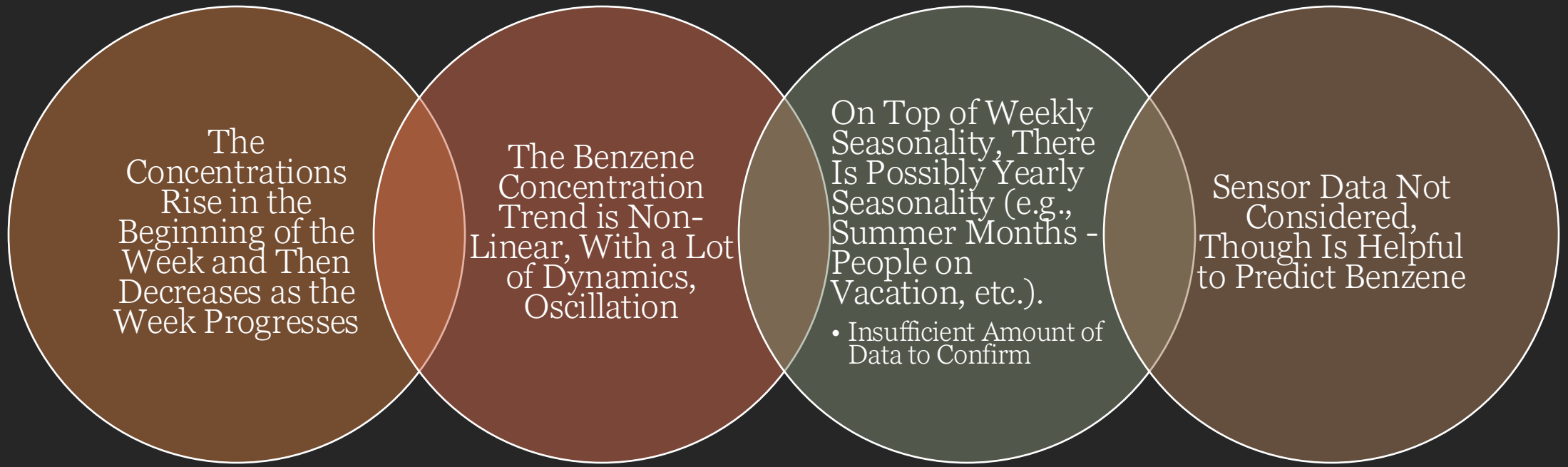


- **NNAR Model Forecast:** The NNAR Model Effectively Captures Certain Areas of Benzene Concentration Over Time
- **Residual Analysis:** The Residuals of the NNAR Model Fluctuates Around Zero With No Obvious Patterns, Indicting the Model Has Captured Most of the Data's Underlying Structure Successfully

Results, Conclusion and Future Work

FINDINGS, PROJECT CONCLUSION(S), AND NEXT STEPS

Key Findings



Conclusion: Learnings and Findings

! Stationarity is Essential

- Differencing and Appropriate Transformations for Stationarity of Data Are Critical for Effective Modeling

Model Effectiveness

- NNAR Captures Trend and Season Patterns, Reflecting a Good Fit of the Underlying Data

Implications for Public Health

- The Resulting Accurate Forecasting of Benzene Levels Is Helping the Public by Enabling Informed Decisions on Exposure and Health Risks
- The Model's Ability to Support Real-Time Monitoring and Policy-Making Decisions Allows for Timely Interventions to Mitigate Pollution and Protect Public Health

Future Works

- Future Work:
  - Predict Other Pollutants: CO (Carbon Monoxide), NOx (Nitrogen Oxides), NO2 (Nitrogen Dioxide), O3 (Ozone), SO2 (Sulfur Dioxide), NMHC (Non-Methane Hydrocarbons)
  - Model Improvement: Train on More Data, Incorporate the Sensor Data to Improve the Accuracy, Use Models that Can Predict Other Highly Correlated Pollutants (e.g., C6H6, CO, NOX, NO2), Change the Objective Function to Penalize More on Predicting Low Values When Actual is High

- Final Thoughts:
  - Project Impact:
    - Evaluate the Broader Impact of This Forecasting Model on Urban Planning and Environmental Policy
    - Potential Impacts Include Enabling Real-Time Monitoring of Air Quality, Improving Public Health Outcomes by Issuing Timely Warnings for High Pollution Levels, and Guiding Policy Decisions for Pollution Control Measures

References

- Vito,Saverio. (2016). Air quality. UCI Machine Learning Repository. https://doi.org/10.24432/C5060Z.
- Benzene - annual limit value for the protection of human health. European Environment Agency. (2014, September 4). https://www.eea.europa.eu/data-and-maps/figures/benzene-annual-limit-value-for-the-protection-of-human-health#:~:text=In%20the%20air%20quality%20directive,measured%20from%201%20January%202010.
- Harrison R, Delgado Saborit JM, Dor F, et al. Benzene. In: WHO Guidelines for Indoor Air Quality: Selected Pollutants. Geneva: World Health Organization; 2010. 1. Available from: https://www.ncbi.nlm.nih.gov/books/NBK138708/