Weather Data Analysis and Temperature Forecasting

Group 15
Adam Cersosimo, Daniel Cersosimo,
Jibin Solomon, Owen Hovey

Contextualization

CNN Weather
RAIN ACCUMULATION
1" 2" 4" 6" 10" 20"+
NEXT 5 DAYS
Boston
Chicago
Cleveland
Omaha
City
Denver
New York
Philadelphia
Washington
Cincinnati
St. Louis
Nashville
Raleigh
Charlotte
querque
Oklahoma City
Little Rock
Atlanta
Charleston
Dallas
Jacksonville
Houston
New Orleans

Overview of Project

- Objective for the project: Analyze weather data from London to capture long term temperature patterns and generate relevant forecasts

- Weather has a significant impact on our lives

- Modern advances in technology enhanced weather data analysis capabilities

Importance of Weather Forecasting

- Weather forecasting is crucial across many industries.
- Aids in preventing natural calamities, optimizing farming methods, managing supply chains, and guiding construction projects.

Problem Statement

- London is one of the largest cities in the world so an understanding of the nature of their weather coupled with long term temperature forecasts into the future would be useful for many people.

- This forecast would help people gain an insight into how the temperature tends to fluctuate from year to year based on decades of data

Scope

- The scope of the temperature time series analysis is to create models that predict temperature in the future given historical dataset.
- The goal is to compare several models to see what would perform well and not be prone to overfitting in the future.
- To do this we have to gather, clean, and implement weather data from a specific location.Then we have to perform exploratory data analysis and build models to predict the future weather data.

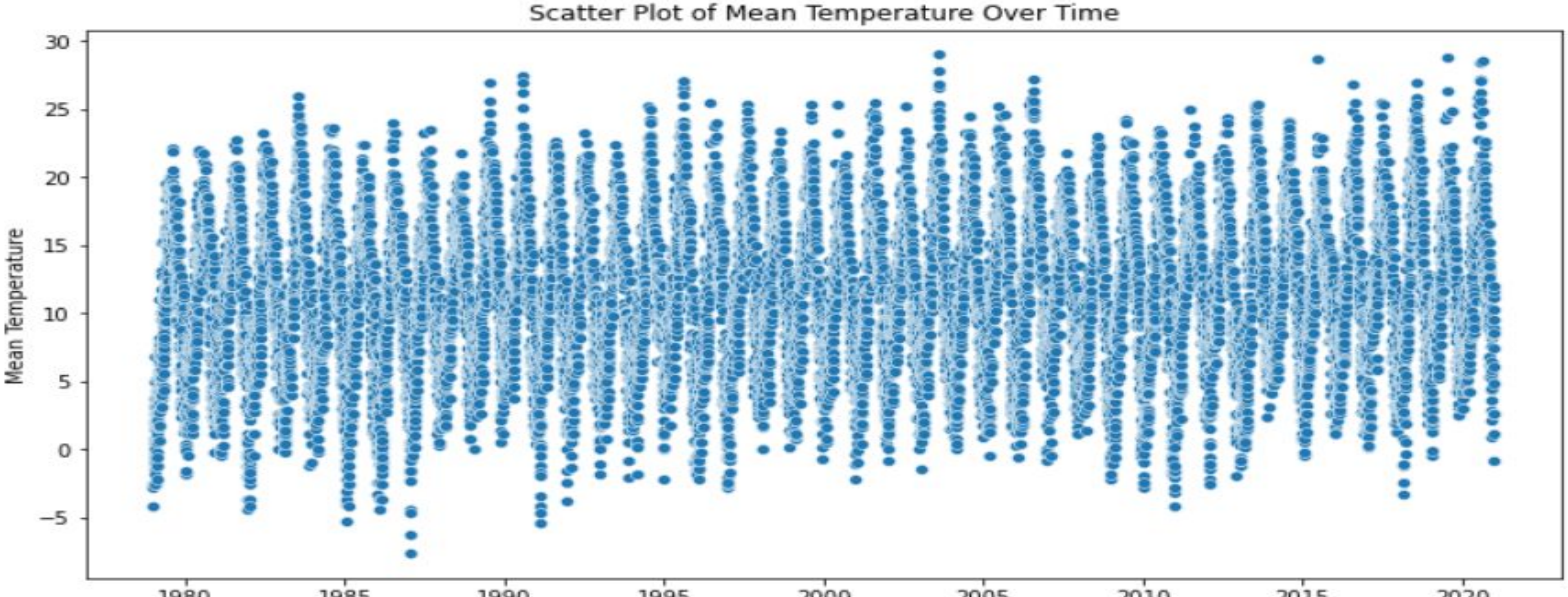
Data Preparation and Data Exploration

Data Acquisition, Loading, and Cleaning

- Selection of the London weather dataset (1979-2021) from Kaggle.
- Data cleaning: Formatting dates, handling missing/null values.
- Specific actions: Dropping the snow depth column, handling null values, summarizing statistics with describe(), and constructing box plots to view the data distribution of features

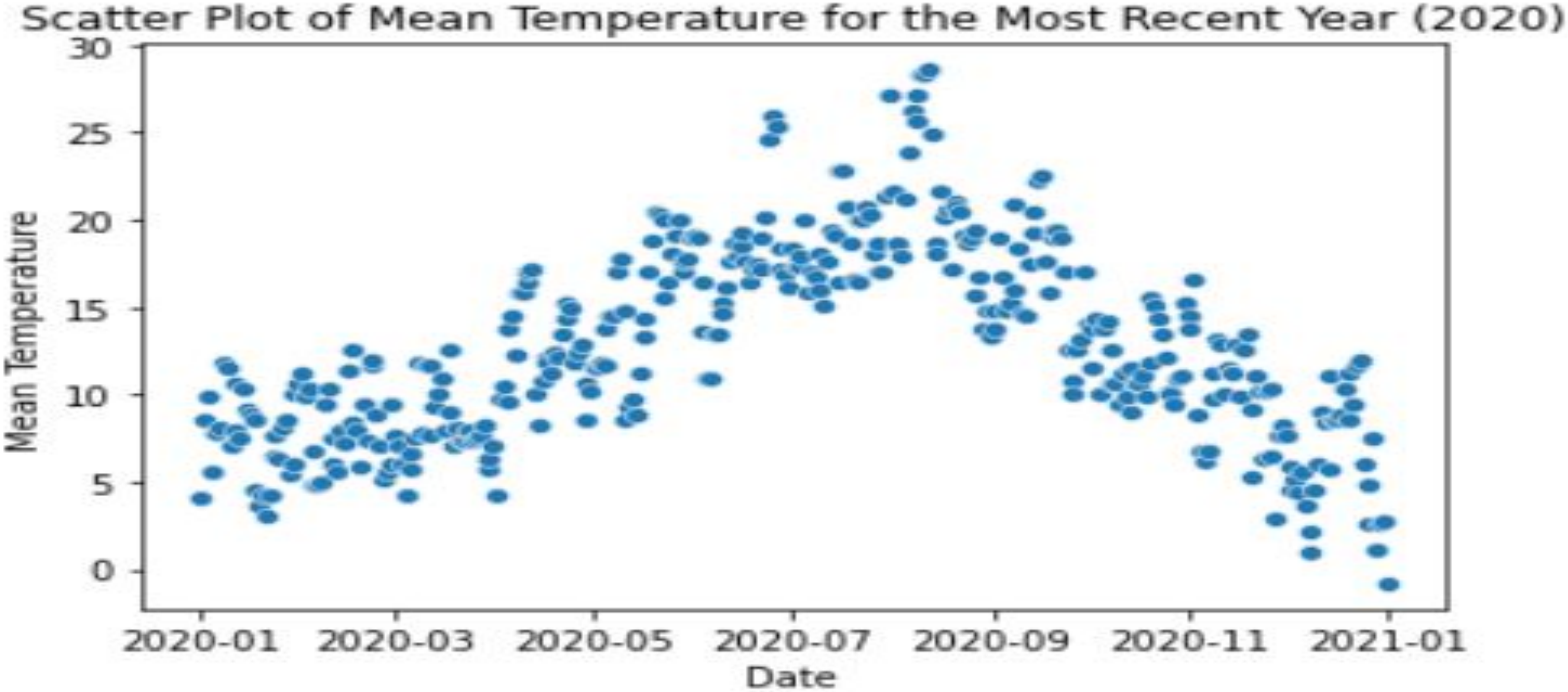
Exploratory Data Analysis (EDA)

- Initially found difficulty drawing conclusions due to condensed seasonality over decades in the visualization for mean temperature



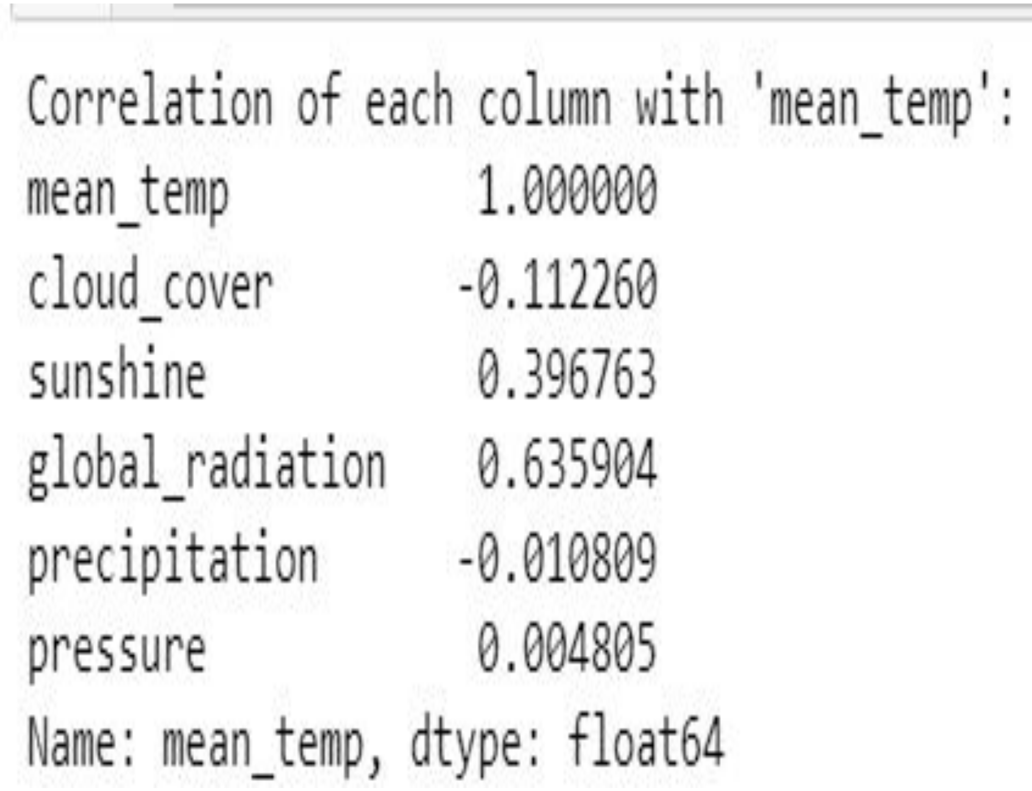
EDA (cont.)

- Closer view of seasonality over a year



EDA – Correlation Analysis

- Analyzed correlation between many variables, for example relative to mean temperature
- Tells us the magnitude and direction to which each variable is correlated with mean temp.



```
Correlation of each column with 'mean_temp':
mean_temp           1.000000
cloud_cover        -0.112260
sunshine            0.396763
global_radiation    0.635904
precipitation      -0.010809
pressure            0.004805
Name: mean_temp, dtype: float64
```

Model and Forecasting

Linear Regression Model

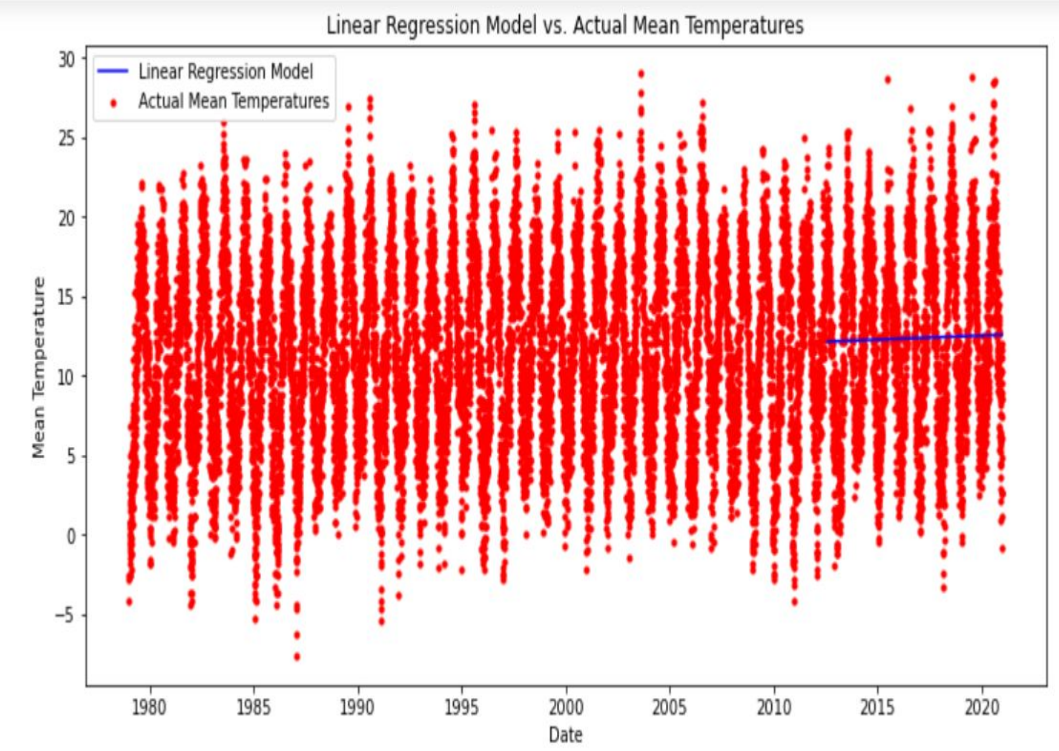
- Trained a linear model on the first 80% of the dataset.
- Tested on the remaining 20%.
- Initial observation: Mean Squared Error or Testing Set is 32.16.

Model Performance:

- Linear model exhibits poor performance.
- Visual inspection indicates a significant mismatch with the actual data.

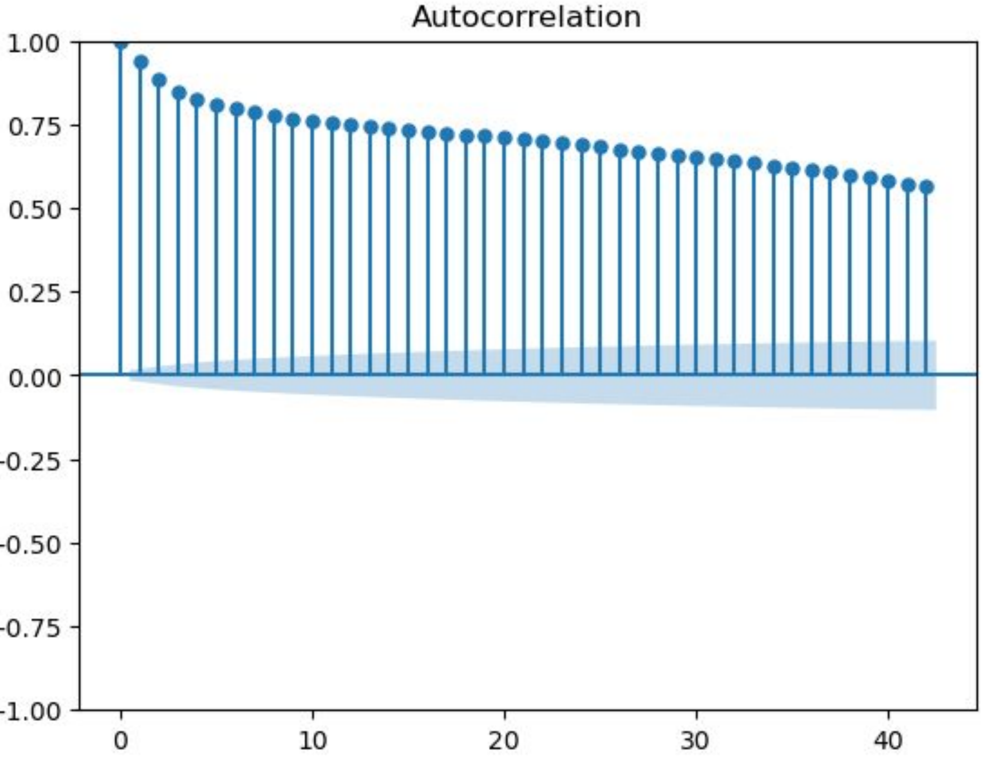
Conclusion:

- Linear regression is not a suitable representation.



Analyzing the Autocorrelation for the Mean Temperature in Preparation for more Advanced Modeling Techniques

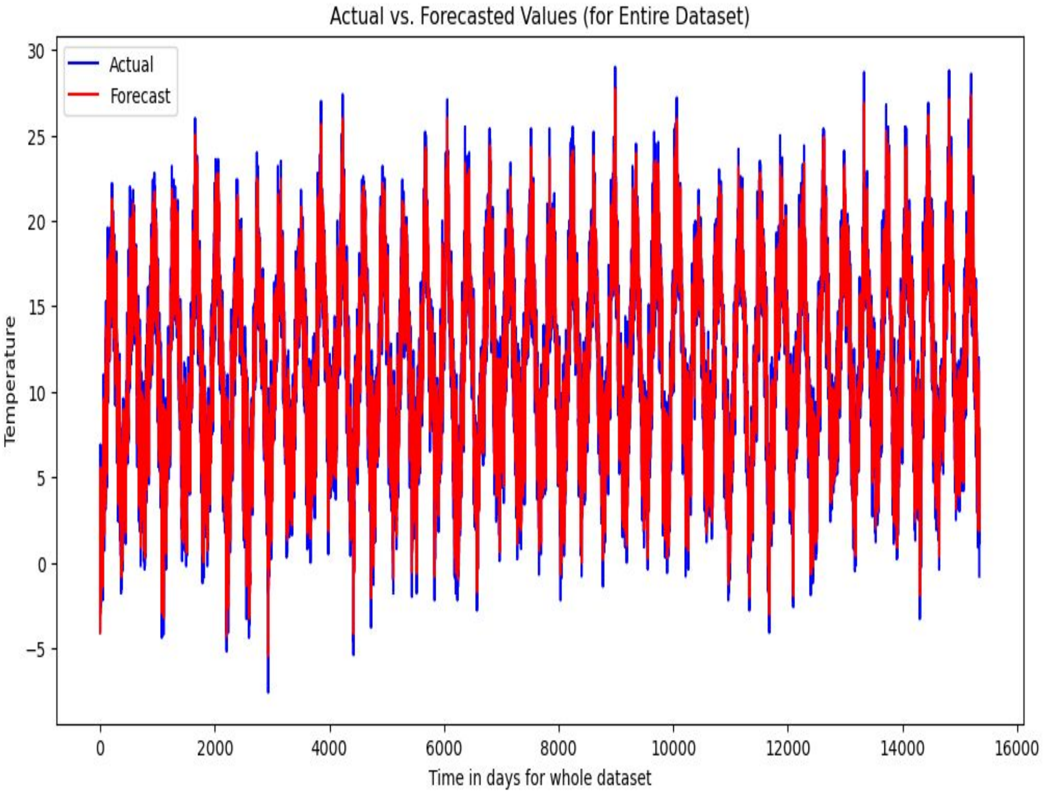
Start by getting autocorrelations and partial autocorrelations:



- From this, we can see that the data exhibits notable ACF at the initial few lags, decreasing over time at a slower rate as it approaches 50 lags
- Essentially, the temperature at any point is quite similar to that of the previous day, decreasing as we get farther from the current day

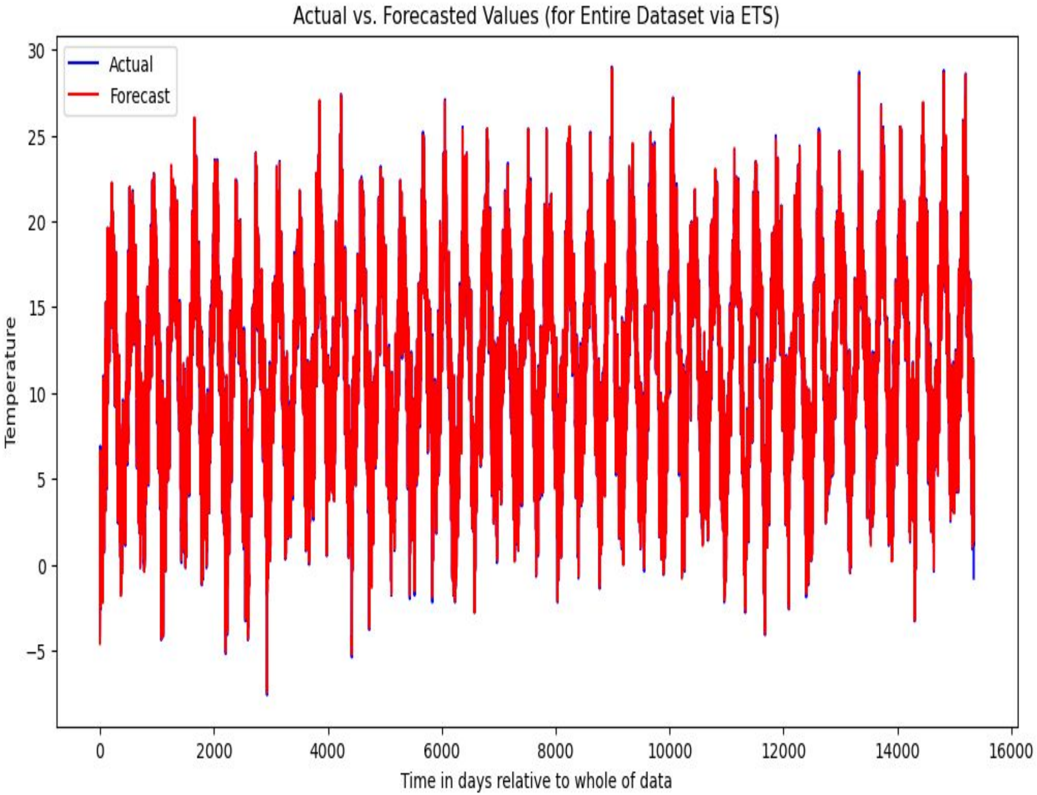
Experimenting with ARIMA

- Based on ACF evaluation, we chose to look at the ARIMA model on the entire dataset
- See signs of overfitting, indicating this model may be too complex for a singular seasonal component
- Model had AR of 3, I of 1, and MA of 2



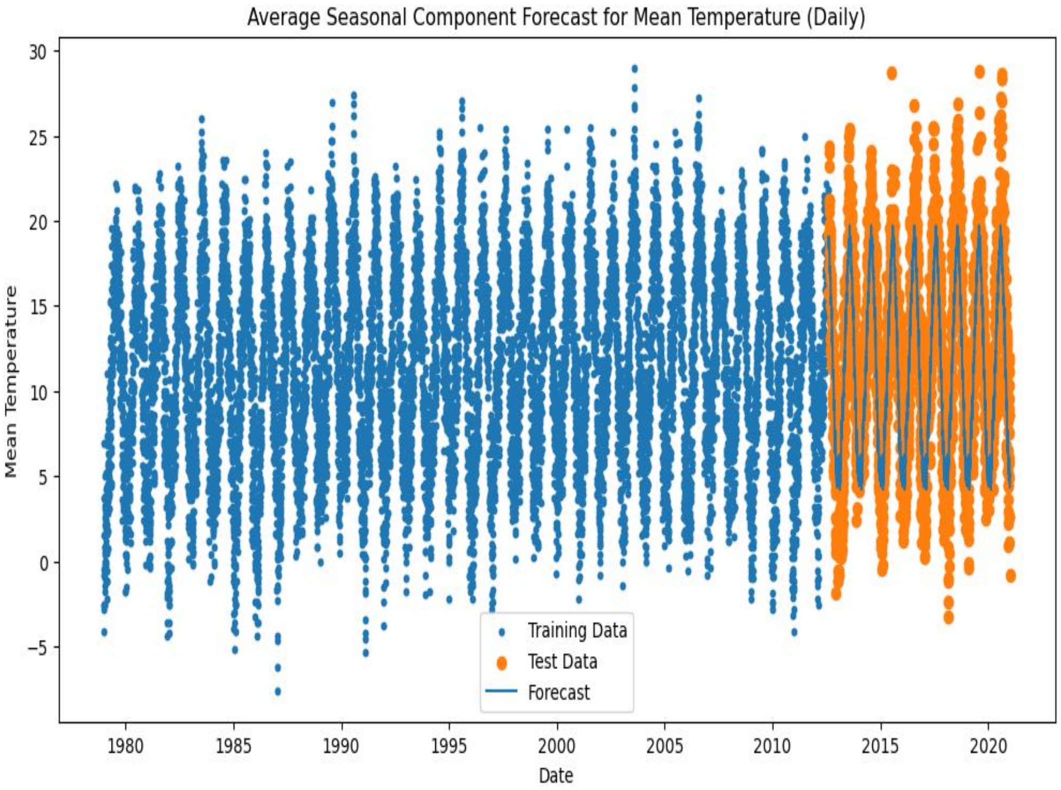
Experimenting with ETS

- When analyzing ETS, we noticed similar results as ARIMA – potential overfitting
- Can barely see the blue line, need to find another approach to fully capture seasonality and avoid risk of overfitting
- Model had 12 seasonal periods with additive seasonality and trend



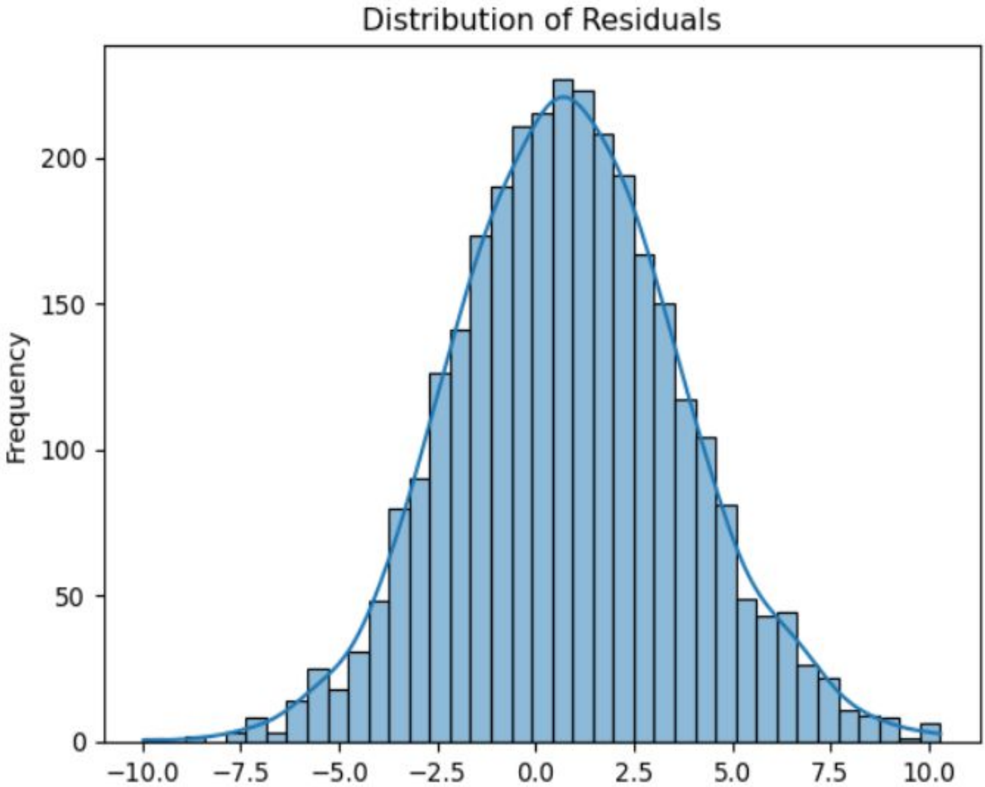
Seasonal Analysis Using GroupBy and Mean Calculations

- Mean Squared Error based on training set is 8.9382
- Based on MSE, this is a better fit for our data as it captures yearly patterns without overfitting the data
- As expected, adjusting for seasonality gives much better prediction



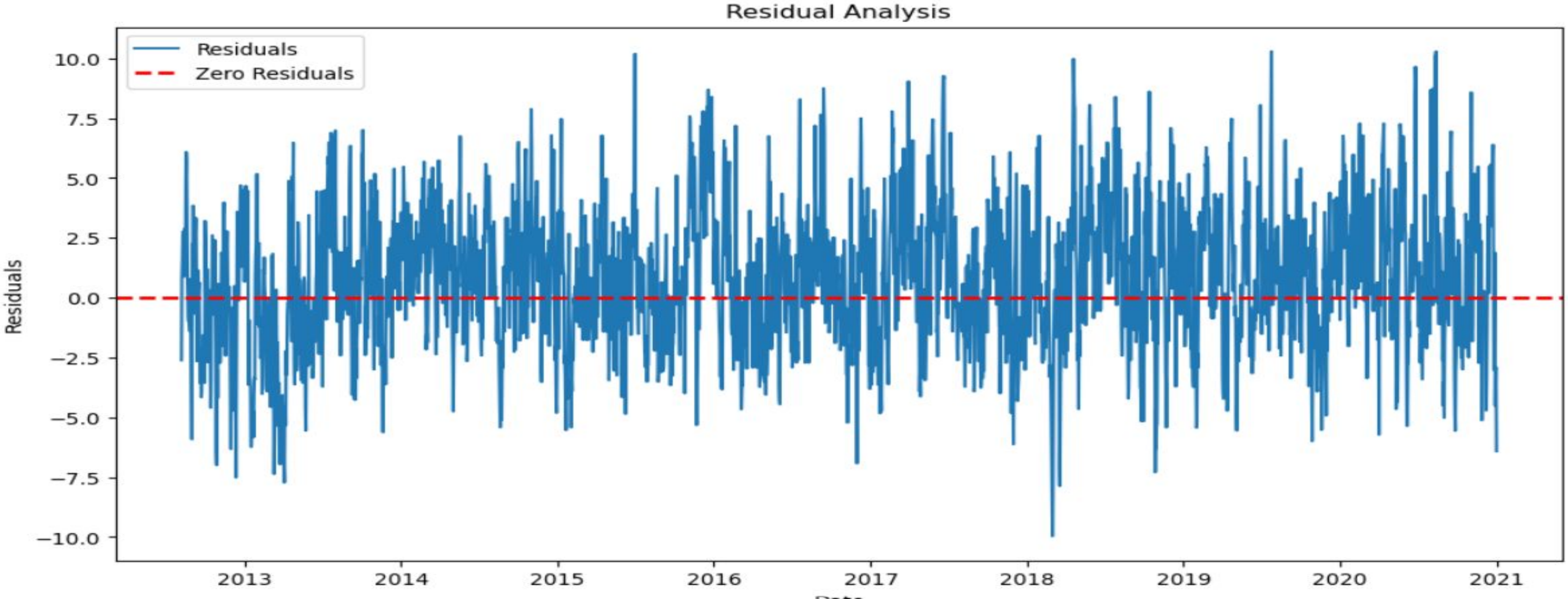
Residuals Distribution Visualization

- Looked at residuals to evaluate model performance
- As expected, we see normally distributed residuals with mean around 0 – an indication of a good model



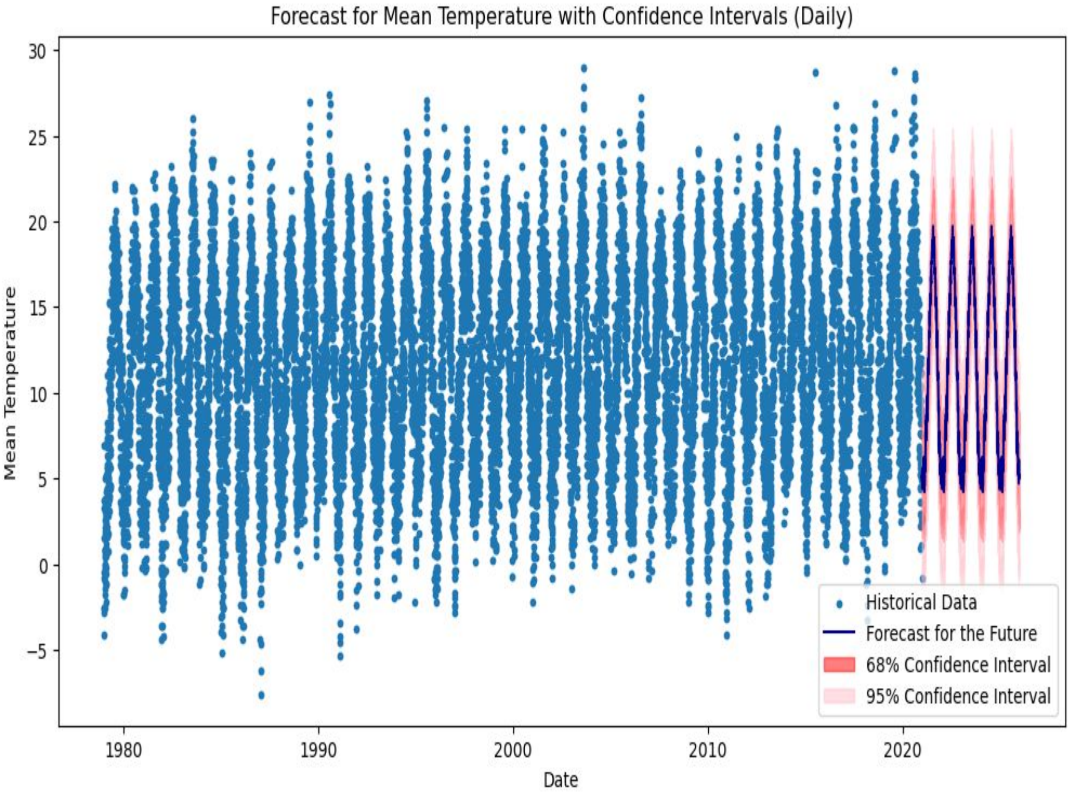
More Residual Analysis

- The residuals should be random with a mean of 0, with little fanning out.
- This model seems to satisfy these requirements because the distribution seems to be random.



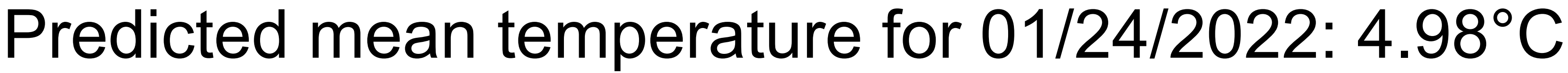
Future Forecasts

- This graph shows what future visualizations would look like based on our model
- Because temperature stays mostly constant over years, we can forecast pretty far into the future
- Only seasonality makes significant impact



Final Code Snippet

- The project ends with a piece of code that takes any future date as an input and gives a predicted mean temperature for that date.
- Below is an output from a user entering in January 24, 2022



```
Predicted mean temperature for 01/24/2022: 4.98°C
```

Conclusions

Real World Applications

- This forecast would be relevant to people who live in London and would like to gain insight into the year to year fluctuations of mean temperature
- This could help for those traveling to london as well since these seasonal temperature fluctuations can aid in planning when in the year someone may want to travel
- The mean temperature predictor could be useful for people who are interested in getting an estimation of where a future mean temperature would be around for a certain day

Challenges

- Data Acquisition
- Data Preprocessing
- Feature Selection
- Capturing seasonality, trends, and overall pattern and data relationships
- Model Selection (ETS and ARIMA experimentation)

Improvements/Extensions

- Additional Data Science Techniques:
- Advanced Time-Series Models: SARIMA
- Additional Machine Learning Approaches
- Real Time Data Integration
- Climate Change Analysis

Questions

Thank you!