TABLE OF CONTENTS

1. INTRODUCTION

1.1. Overview of Problem Statement and Weather Data Analysis

Weather impacts our everyday lives. Of the 8 billion people who reside on our planet, virtually everyone you could come across would agree with this statement. While practically all would consent to this, the manner in which weather impacts us can drastically vary from person to person and region to region across nations themselves and the world as a whole.The magnitude in which weather impacts our species cannot be understated and as a result, an improved understanding of weather can serve to benefit everyone. Unfortunately, these climatic phenomena have a reputation of being highly unpredictable and for most of time have left humans unable to consistently comprehend and accurately forecast weather. Modern advances in technology have dramatically enhanced a process of being able to view, interpret, and store data pertaining to weather developments. Having access to such a massive amount of accurate data coupled with the means to analyze this data have presented the ability to come to an understanding of weather in an unprecedented manner. As we all know, weather varies drastically from region to region which renders data and its subsequent interpretations only relevant to a specific area. Due to this, in our project we tasked ourselves with collecting, processing, exploring, and interpreting data unique to a specific region. After beginning this project and sifting through many regions, we elected a specific region which we determined holds immense relevance to a vast amount of people. Our objective for this project is to utilize a dataset that shows various weather features over time in London in an effort to draw an understanding and attempt to generate relevant forecasts. This data provides information on a diverse range of climatic factors from which we seek to identify patterns by analyzing this data. This process is important as London is one of the largest cities in the world so an understanding of the nature of their weather coupled with forecasts into the future would be useful for many people.

1.2. Specific Importance of Weather Forecasting

In many different industries, weather forecasting is essential to optimize preparation for a seemingly limitless array of tasks which are imperative to our species. Some specific tasks where precise forecasts are critical are in preventing natural calamities, directing farming

methods, supply chain, construction, and much more. In a broader sense, accurate weather forecasts impact human ability to maintain and sustain our natural resources.

1.3. Objectives of the Study

Our project aims to garner an understanding of the weather patterns and variations in London and subsequently construct models for predicting future weather conditions with a goal to contribute to more accurate and dependable weather forecasting methods by utilizing statistical and machine learning techniques.

1.4. 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.

2. Data Preparation and Exploration

2.1. Data Collecting, Loading and Cleaning

This process starts with an intricate evaluation of potential locations and their data. After much consideration, we elected to utilize a dataset for London weather features from 1979 to 2021. This was chosen due to the importance of London coupled with a large volume of data over decades for many features. This data is from Kaggle and was sourced from this link: https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

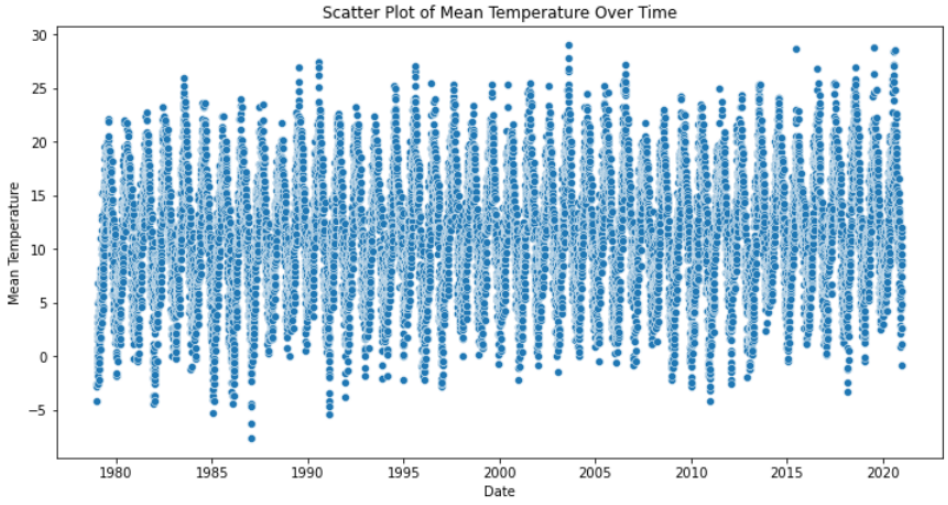
The analysis begins by importing necessary libraries (such as pandas, numpy, matplotlib, seaborn, etc.) and loading the London weather dataset from a CSV file using pandas (df = pd.read_csv(path)). The initial data cleaning involves formatting date columns for consistency and handling any missing or null values in the dataset. We note the total presence of null values, 1533 along with their unique distribution about specific rows and columns. Following this, we note that the column pertaining to snow depth has 1441 of these null values and therefore drop the feature. For the other columns, these null values were dispersed and minimal relative to the whole dataset. Given the temporal nature of the data, we elected to fill these values with the previous value as data closest in time would be the most similar. In addition, we utilize the describe() function for each feature to get summary statistics. We also construct boxplots for each feature to note the data distribution as well as outlier presence. From this, we note a very slight outlier presence for maximum temperature and minimum temperature along with a substantial outlier presence for precipitation and pressure. In the context of precipitation, the reasoning for the substantial outliers is that most days there is no rain, meaning the data is declared an outlier when there is a day with more than minimal rain. Upon further investigation

later in the project, we came to the realization that we would not have to deal with these outliers because they would have no impact on the temperature forecast models. This also applies to the pressure outliers.

2.2. Exploratory Data Analysis (EDA)

2.2.1. Visualizing Temperature Trends

During initial exploration, we graphed a mean temperature vs. date to see if we could draw any early conclusions based on the visual.

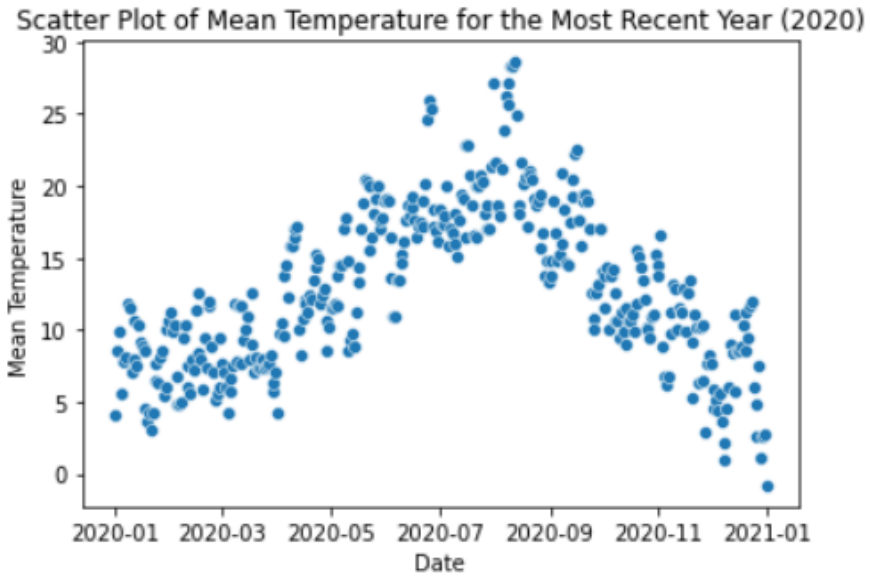


Here we can see that it is difficult to draw many conclusions from this graph because of the significant number of data points that there are. However, if you inspect closely you can see that there is a degree of seasonality present with little trend. This makes sense because the temperature should have very little difference year in and year out even with global warming

potentially being a factor. We can also see a degree of randomness (error). Lets see what it looks like over the course of just the most recent year in the dataset.

2.2.2. Seasonal Patterns in Weather Data

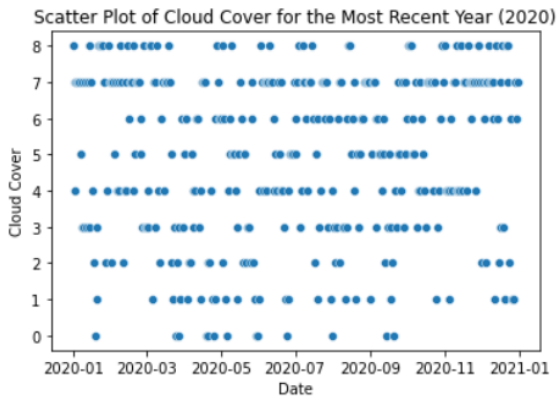
We now look at a graph of the mean temperature over the course of a year to see the seasonality more closely.

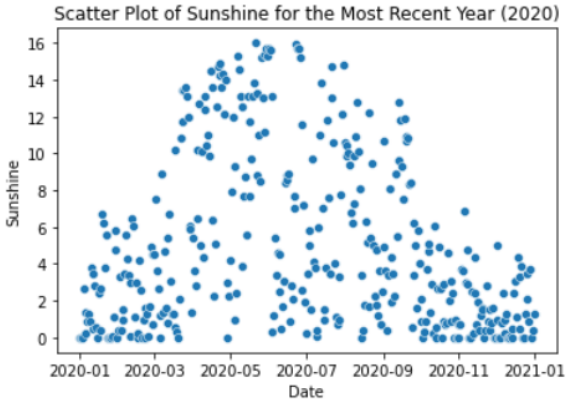


Here we can discuss some clear insights. Over the course of the year, it is clear that seasonality plays a substantial role which makes sense because the temperature changes in London depending on the season. The winter months here, December-March, are much colder than the other months.

2.2.3. Cloud Cover and Sunshine Analysis

We decided to create a few plots of other features in the dataset to see if they had any patterns as well.





Some of the other visualizations included in the notebook include global radiation, mean temperature, precipitation, pressure and more. These visualizations give us key insights on certain weather variables which we will use in the future of our analysis.

2.2.4. Correlation Analysis

Let's now look at some correlations between different features in this dataset.

```
correlation_matrix = df[['mean_temp', 'cloud_cover', 'sunshine', 'global_radiation','precipitation', 'pressure']].corr()
correlation_with_mean_temp = correlation_matrix['mean_temp']
print("Correlation of each column with 'mean_temp':")
print(correlation_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
```

Here, we see the correlation coefficient between mean temperature and all of the variables except max and min temperature. This tells us how the magnitude and direction which each variable is correlated with mean temperature. We see cloud cover to be negatively correlated with temperature. While sunshine and global radiation are positively correlated. Remember correlation does not necessarily mean causation but it does make sense that a sunny day would typically raise the temperature.

This time we see every correlation with the features.

	cloud_cover	sunshine	global_radiation	max_temp	mean_temp	min_temp	precipitation	pressure
cloud_cover	1.000000	-0.738229	-0.487112	-0.213226	-0.112260	0.047633	0.234323	-0.240417
sunshine	-0.738229	1.000000	0.852159	0.472089	0.396763	0.219088	-0.231565	0.226804
global_radiation	-0.487112	0.852159	1.000000	0.691314	0.635904	0.478476	-0.163145	0.150449
max_temp	-0.213226	0.472089	0.691314	1.000000	0.912314	0.810526	-0.071918	0.100396
mean_temp	-0.112260	0.396763	0.635904	0.912314	1.000000	0.955520	-0.010809	0.004805

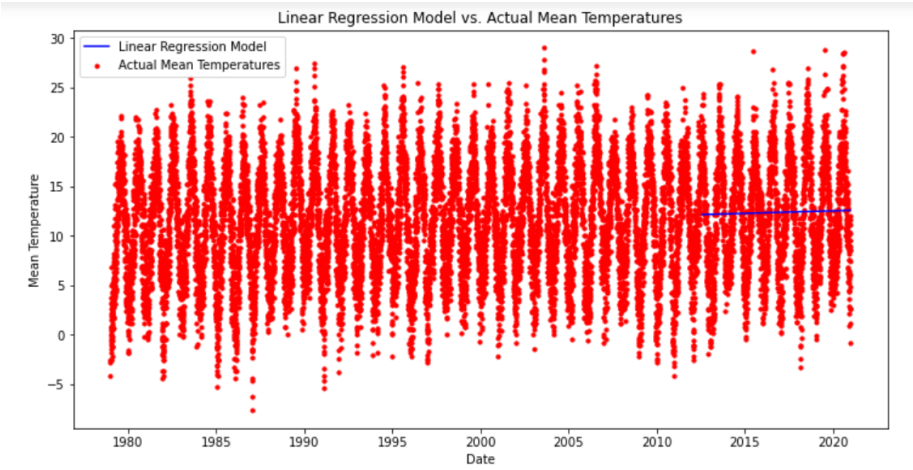
min_temp	0.047633	0.219088	0.478476	0.810526	0.955520	1.000000	0.037208	-0.074227
precipitation	0.234323	-0.231565	-0.163145	-0.071918	-0.010809	0.037208	1.000000	-0.349418
pressure	-0.240417	0.226804	0.150449	0.100396	0.004805	-0.074227	-0.349418	1.000000

After careful analysis, we determined that most of the features would not have significant use in our model due to the fact that our model was time series, meaning that much of the impact of the other features is captured in the historical temperature data.

3. Model and Forecasting:

3.1 Linear Regression Model:

We will start by training a linear model on the first 80% of the data and then we will test it on the next 20%. We can assume this will perform poorly based on the fact that there is substantial seasonality in the dataset.



Mean Squared Error on Testing Set: 32.155905666823166

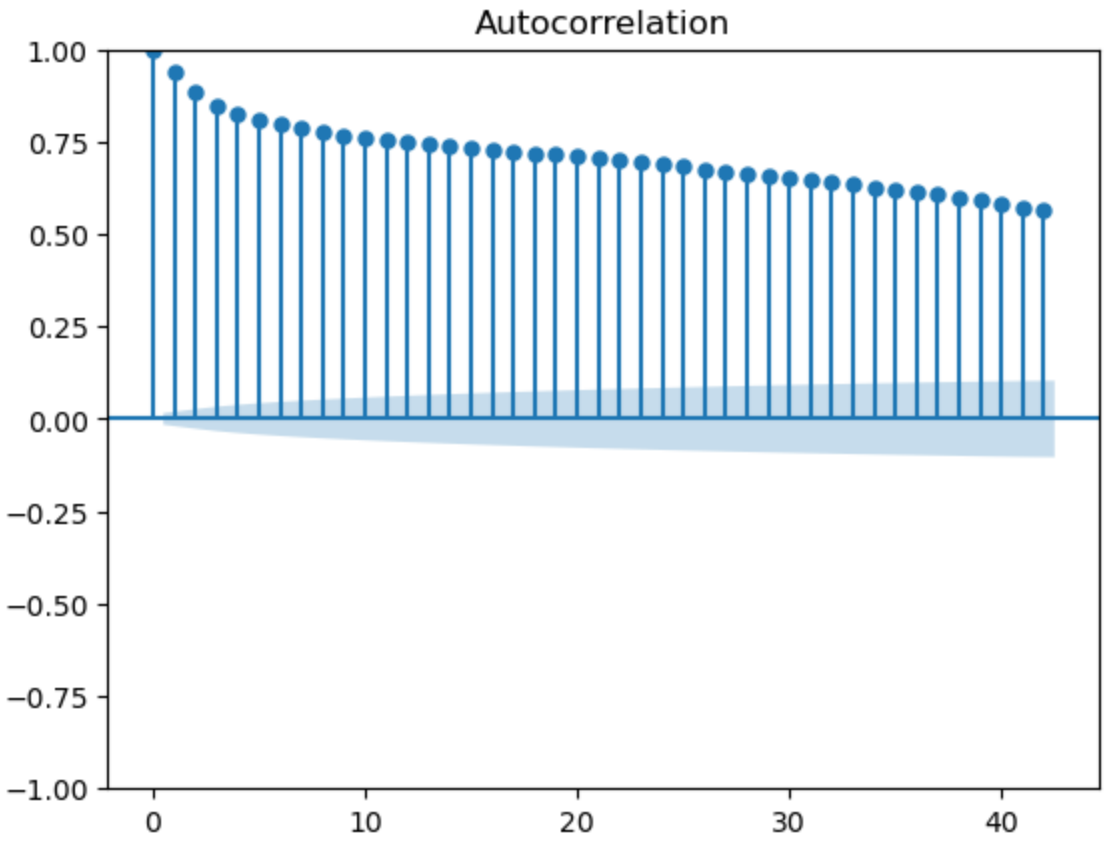
Slope: 1.6085806244318141e-18

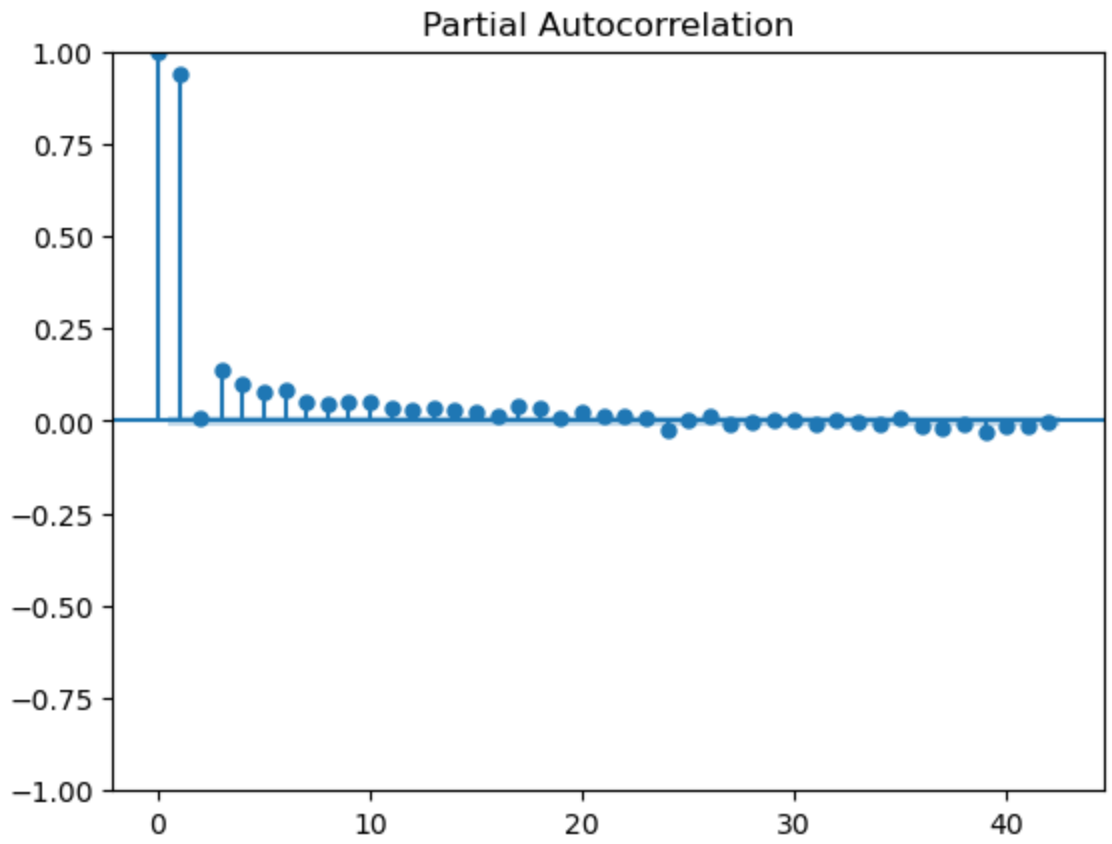
It is difficult to draw conclusions from the MSE without something to compare it to so we will touch on this later, however, we can visually see that the model performs poorly. Moreover, we can conclude that there is no significant trend in this dataset because the slope in the linear model is nearly 0.

Can be seen that a linear model is a very poor representation of this dataset. We see that the actual data exhibits strong seasonality and the linear model is not representative of the data. Another interesting thing is that the slope is nearly 0, meaning that the trend is negligible. We should focus more on seasonality.

3.2 Analyzing the Autocorrelation and Partial Autocorrelation while Experimenting with ARIMA and ETS

We start here with getting the ACF and PACF for the mean temperature data. This is visualized below

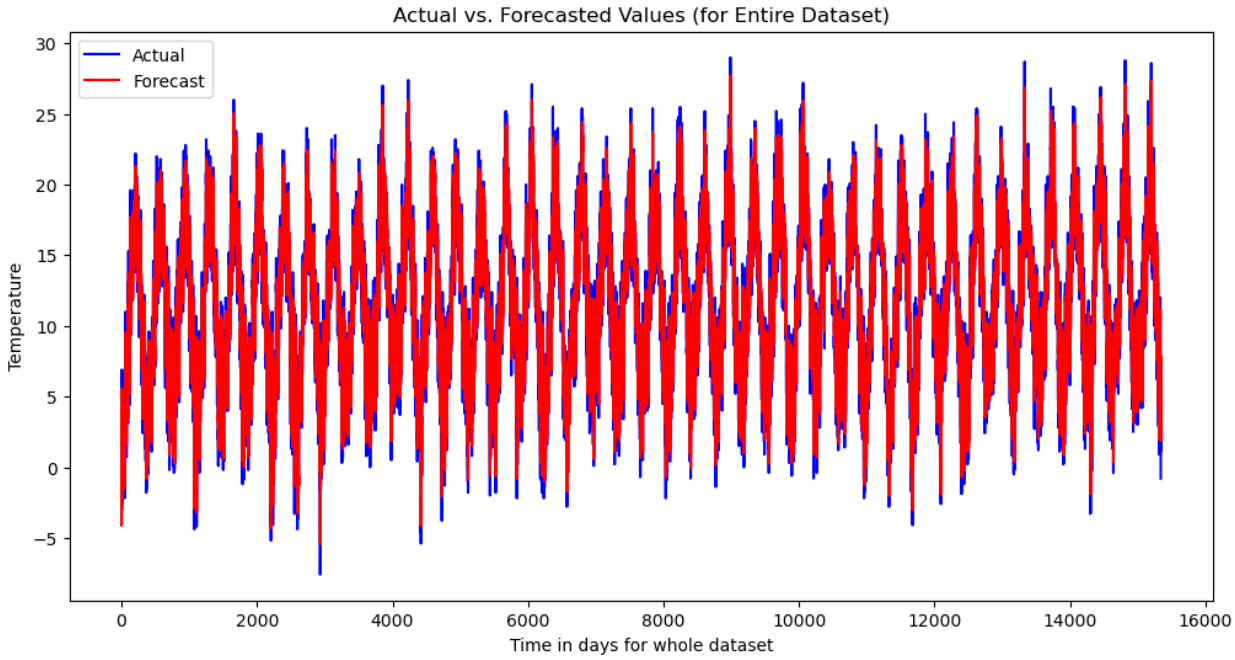

Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40



From this we can see that the data exhibits notable ACF especially at the initial lags which decreases over time at a slower rate as it approaches 50 lags. This makes sense in the context of temporal temperature data since the most recent datapoint bears heavy influence on some hypothetical current datapoint. Essentially, the temperature at any point in time is quite similar to that of the previous day and this decreases as the one gets farther from the current day. For PACF fixates on specific relationships between a current point and a lagged point without being influenced by the intermediate lags. From the PACF we can note two initial spikes followed by a drastic drop and subsequent fluctuations about 0 to lag 50. This is reasonable given the context since temporal temperature data bears significant direct similarity to the closest data points.

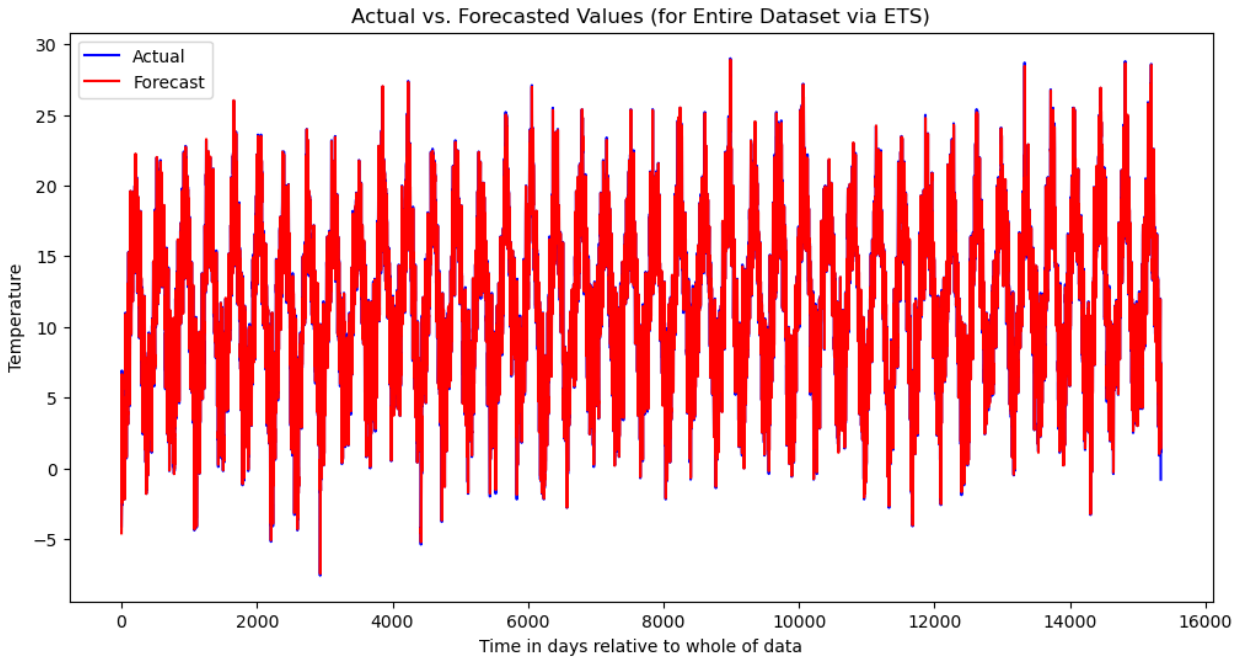
From this insight, we designed an ARIMA model utilizing an ARIMA function with the parameters for Autoregression, Integrated (which is for differencing) and moving average

3 is used for AR since in the ACF plot we have the first three lags which are drastic from which it tapers off, 1 for I since there is not notable trend aside from seasonality, and 2 for MA since there are two notable spikes in PACF. This is plotted below with the red as the model and the blue as the actual values.



This is done on the full dataset as we first wanted to see how the model would resemble it. From this point we noticed it is highly fitted which raised concerns of overfitting. As a result, this incited thinking in our group that a model of this nature may be overly complex for a consistent seasonal component and a simpler approach could be used to forecast this can view the model below.

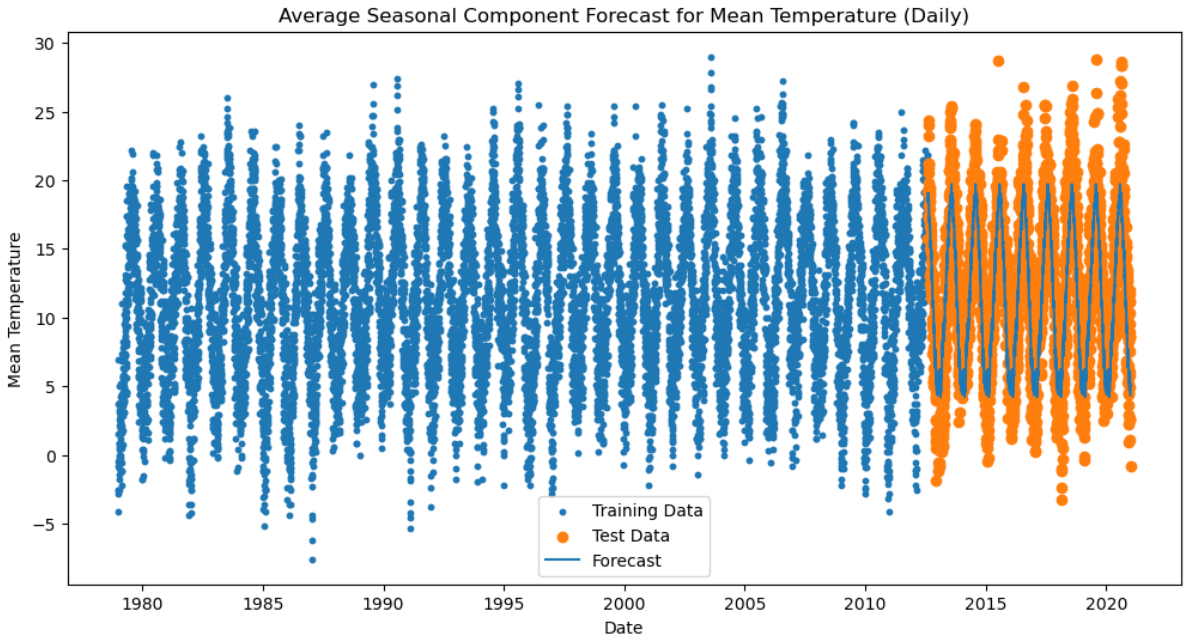
Following the ARIMA experimentation, we attempted to employ the ETS modeling technique which specializes in error, trend, and seasonality. We utilized a similar technique as with the ARIMA model on the dataset to see how it would capture the data. This can be seen below with the red as the model and the blue as the actual data.



From ARIMA we pivoted to ETS and used the full dataset to see how fitted the model would be on the data. Once again, this model was highly fitted (can barely see the blue) onto the data and led us to consider another approach that captures the seasonality of the data without being a risk of being overfitted.

3.3 Seasonal Analysis Using GroupBy and Mean Calculations

Now, we will adjust for seasonality.

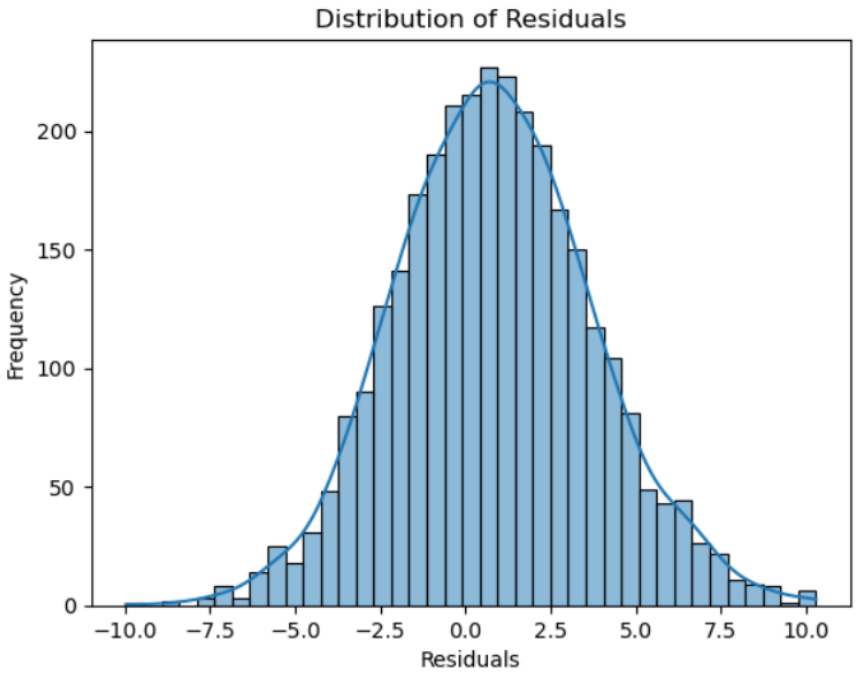


Mean Squared Error: 8.938186477935957

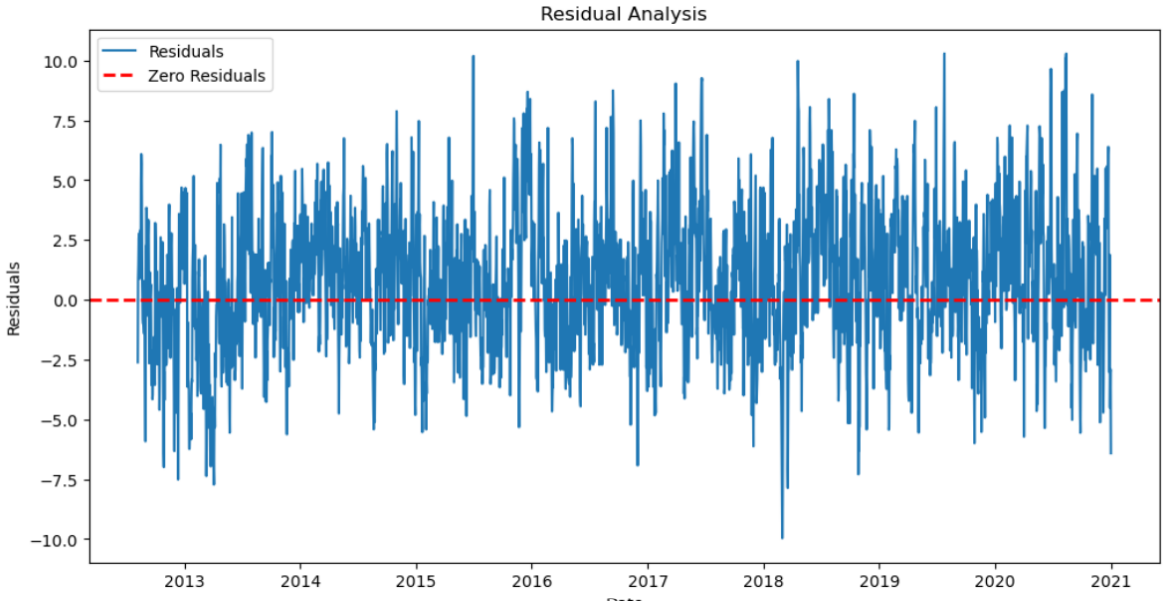
Based on the MSE, we can see that this is a much better forecast. As expected, adjusting for seasonality gives much better predictions – which makes sense considering the nature of weather patterns. This plot helps in understanding the detailed fluctuation of temperature over this period.

3.4 Residuals Distribution Visualization:

A histogram of residuals shows many things about a mode. It should be a bell shaped curve with a mean around 0. This is because in a good model, the residuals would be randomly spread out with little pattern and most should be around 0, meaning as small as possible. The variance should be consistent rather than more negative than positive or vice versa.



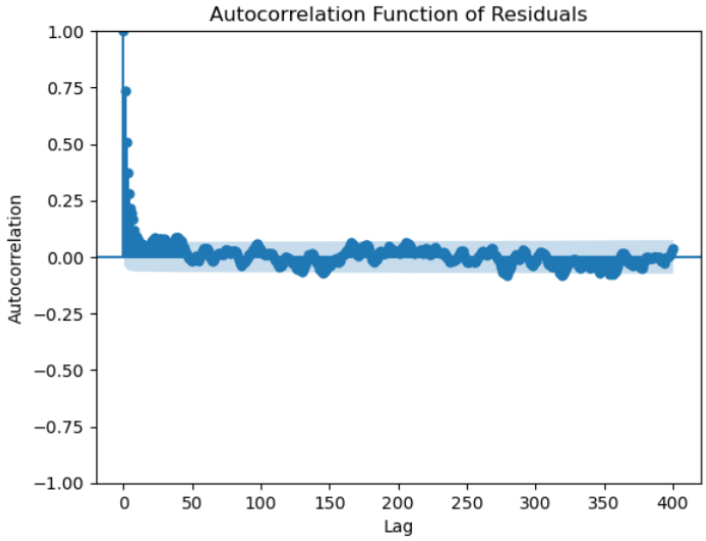
This graph shows what we want in a good model because it is a bell shaped curve with a mean around 0.



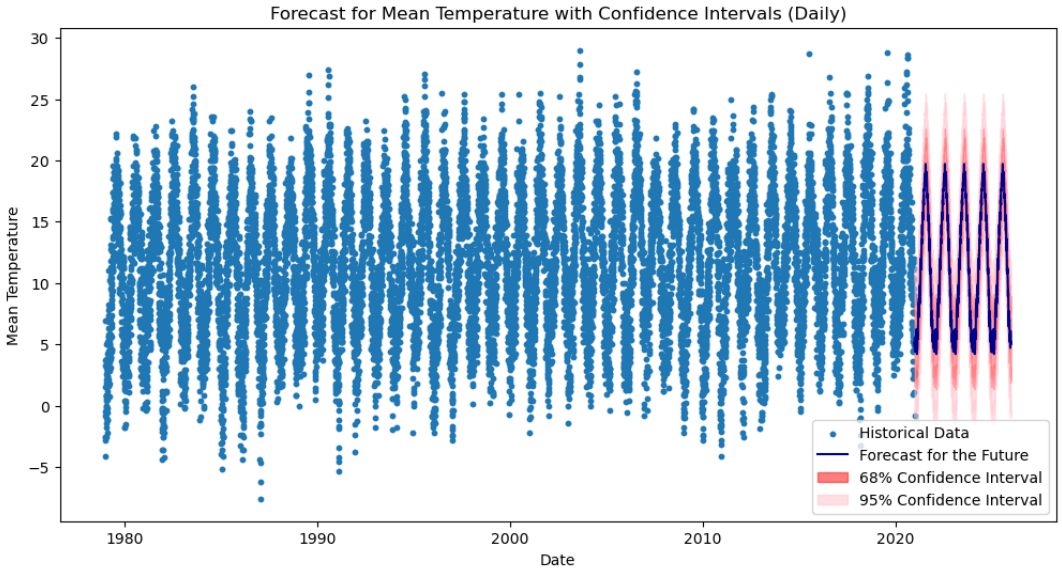
The residuals should be random with a mean of 0, with little fanning out. This model seems to satisfy these requirements because the distribution seems to be random.

3.5 Residual ACF Analysis:

The Autocorrelation Function (ACF) graph shows the autocorrelation of the residuals from the model with 400 lags.



The autocorrelation graph shows how correlated each of the residuals are to one another. The first lag is one because the residual has perfect correlation to itself but then decreases significantly as we move forward in the lags. It makes sense that the first few lags have some autocorrelation since temperature changes usually come in waves rather than drastic changes at once.



This graph shows what future visualizations would look like based on the model provided. It also includes standard deviation confidence intervals that attempt to capture an area where the real data point is likely to land in. In many models, this interval would spread out more and more over time since it is often difficult to forecast values in the distant future when trends and other factors have a substantial role in the model. However, this is not the case with our model because temperature remains consistent year after year so only seasonality had a substantial impact on the model. This is why our interval looks the way it does.

The project ends with a piece of code that takes any future date as an input and gives a predicted temperature for that date.

4. Conclusions and Future Work:

4.1 Challenges

We had a variety of problems that we had to overcome and we believe that we were successful for the most part in doing so. Some of the issues we had are discussed more thoroughly below.

- **Data Preprocessing:** Some challenges we had were finding an adequate dataset, handling the null values, and deciding on how to deal with the outliers present in some of the features.
- **Feature Selection:** When we first thought of this project, the initial intention was to utilize various weather features to collectively come together to construct a forecast for temperature. This would have been heavily based on the historical temperature data while factoring in various other features at weights which would have been determined to optimize the forecast. After initially embarking on the project, we realized that this presented an issue not initially noted. As a time series analysis for temperature, the impact which these other factors had on the temperature would have already been in essence captured by the temperature data itself. Each of these features are notable when discussing weather, however the collective impact they have upon temperature is what is useful given the goal of temperature forecasting. In turn, we realized that using these additional features in conjunction with the historical data would be pointless since their impacts are already captured in the past temperature data. Due to this, we pivoted the project idea to utilizing some of these features to generate models, and forecast each feature. As we progressed in the project and were in the exploratory data analysis stage. We discovered the unpredictability, inconsistency, and outlier presence of some of these

other features. As a result, the patterns and relationships in these features were quite loose and attempted models would encounter inconsistent accuracy. In turn, we opted to concentrate our focus on temperature data as this had a clear seasonal element which could be accounted for in an accurate forecast.

- **Seasonality and Trend:** Accounting for seasonality and trends when working with time series is difficult because putting too much weight in either can lead to a model which is too flexible. We determined that trend was not necessary because of how little positive slope there was in the linear model.
- **Interpreting the Model:** Some minor issues were with model interpretation, specifically with getting conclusions from the residual graphs.
- **Model Selection:** We tested several models to see which would best fit our data. We attempted to implement models such as ETS and ARIMA but they were not ideal for this project. They seemed to follow the data too closely so they had a high chance of overfitting as we forecast deeper into the future.

4.2 Lessons Learned

Overcoming the challenges that we encountered proved to be beneficial in improving our data analysis skills. Some of the lessons learned include:

- **Data Preprocessing:** Dealing with our dataset and cleaning our data proved to be quite a challenge. We used a variety of different methods from class as well as external resources to clean our data which is valuable experience. Considering the data was daily records

from 1979 to 2021 for various features, it was quite large and this presented new challenges as pinpointing where issues where with the data as well as figuring out the techniques to be able to properly uncover and handle these was a good learning experience. Handling null values, outliers, and more was carried out by the techniques learned in class alongside some from external sources more catered to data of a temporal nature.

- **Exploratory Data Analysis (EDA):** Garnering a proper understanding of the data through visualization and statistics is crucial before any modeling. Our EDA revealed underlying patterns, trends, and anomalies that proved critical in guiding the modeling process. We employed many techniques from class such as summary statistics for features, box plots to understand data distribution, scatter plots to visualize this distribution over time and more. This collectively facilitated our comprehension of the data and the underlying patterns and relationships and properties. Overall, this project has given us more valuable experience performing EDA.
- **Modeling Techniques:** From this project, we gained hands-on experience in applying linear regression and learned about its assumptions, limitations, and when it may or may not be appropriate. In addition, we experimented with more complex modeling methods such as ARIMA and ETS before ultimately settling on a more straightforward seasonal model primarily focused on the seasonal fluctuations of the temporal temperature data.
- **Seasonality in Time Series Data:** We learned how to handle data that exhibits seasonality as a result of executing this project. Going into a project based on weather over time, seasonality was not surprising however the manner and methods to handle it

was a valuable experience. Using a variety of models in an effort to optimally handle this was fantastic insight and doing so with such a large volume of data was fascinating.

- **General Programming Skills:** We enhanced our proficiency in Python and in using libraries like pandas, NumPy, scikit-learn, matplotlib, seaborn, statsmodels, and more.
- **Machine Learning Workflow:** We gained experience with time series machine learning workflow, from understanding the problem to deploying the model and making predictions.
- **Critical Analysis of Models:** Oftentimes it is helpful to discuss the analysis with team members so everyone can understand and give their own interpretation.

4.3 Scope for Improvements and Extension

While the current project has provided valuable insights into weather patterns and their relationships coupled with a handful of techniques for time series analysis, there are several potential improvements and extensions that could further enhance our understanding and forecasting capabilities. For example, a few ideas for future additions could be:

- **Incorporating Additional Data Science Techniques:** In an effort to improve the model's predictive power, additional data sources could be integrated. This might include more macro-environmental focused variables such as more intricate atmospheric pressure changes, oceanic cycles, and even socio-economic factors that could influence weather patterns. By enriching the dataset, we can provide the model with a more holistic view of the factors which could impact weather.

- **Exploring Advanced Time-Series Models:** Another extension would be to experiment with additional models in our time series analysis which were not employed in the project. Some models may include SARIMA (ARIMA with a seasonal emphasis), or a Holt Winters approach. While we worked with ARIMA and ETS models, further experimentation and usage of these and other models in tandem with other methods could potentially enhance the process and end result of the project.
- **Additional Machine Learning Approaches:** Some further machine learning models which could be implemented could be, random forests and gradient boosting machines, may capture more complex non-linear trends in the data. A potential application of deep learning techniques would be an interesting addition to the overall process in an effort to develop accurate forecasts.
- **Real-Time Data Integration:** Another extension that we could potentially add in the future would include the feature to update with real time data. This could transform the model to be used in real time. This however would be rather complicated to implement and would likely require an incredibly detailed and reliable data source up to present day that consistently updates in conjunction with the model.
- **Climate Change Analysis:** Another section we could add to this project would be an analysis of climate change; specifically how it is impacting temperature over time. This could provide valuable insights on how climate change has been influencing weather patterns as well as provide insights on what we can expect for climate change going forward into the future.
- **Location Customization:** Another key feature that we could add is the ability for the user to select and view data based on location. This would require access to an expansive

array of data for regions across the world to apply a modeling process which could be adapted for different data in various regions across the globe. Our project only utilizes data from London and therefore would have to be enhanced to be applicable to other data. Overall, to add this feature, we would have to find a variety of data sources likely from multiple databases to gather weather data from any location.

https://github.com/adacersos/INFO-501/blob/c7155fe4053563da486a6b1f4c66637685f888c6/Data%20mining%20project%20(1)%20(2)%20(1)%20(2).ipynb