CLIMATE ANALYSIS – TIME SERIES FORECASTS

Submitted by

Name	Reg No:
KAVIYA P	**223032**
KAVYA K	**223033**
LAMIYA NAZER K	**223034**
MOHAMMED FAWAZ J	**223035**

In partial fulfilment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics Of

School of Digital Sciences

Kerala University of Digital Sciences, Innovation, and Technology

(Digital University Kerala)

Technocity Campus, Thiruvananthapuram, Kerala – 695317

August 2023

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **CLIMATE ANALYSIS – TIME SERIES FORECASTS**

submitted by

Name	Reg No:
KAVIYA P	**223032**
KAVYA K	**223033**
LAMIYA NAZER K	**223034**
MOHAMMED FAWAZ J	**223035**

in partial fulfilment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics is a Bonafide record of the work carried out at KERALA UNIVERSITY OF DIGITAL SCIENCES, INNOVATION AND TECHNOLOGY under us supervision.

Supervisor

Prof. MANOJ KUMAR TK
School Of Digital Sciences
DUK

Course Coordinator

Prof. MANOJ KUMAR TK
School of Digital Sciences
DUK

Head of Institution
Prof. SAJI GOPINATH
Vice Chancellor
DUK

DECLARATION

We**, KAVIYA PREM, KAVYA K, LAMIYA NAZER K, MOHAMMED FAWAZ J** student of Master of Science in Computer Science with Specialization in Data Analytics, hereby declare that this report is substantially the result of our own work, and has been carried out during the period March 2023-July 2023

Place: TRIVANDRUM

Date:

KAVIYA P

KAVYA K

LAMIYA NAZER K

MOHAMMED FAWAZ J

ACKNOWLEDGEMENT

I would like to express my sincere and deepest gratitude to my guide Dr. T.K. Manoj Kumar, Associate Professor, Digital University Kerala, Trivandrum, for his valuable guidance, advice, and support, which enabled me to complete this project successfully.

We also like to express a deep sense of gratitude to Prof. Saji Gopinath for providing us with a good environment, valuable guidance, and educational facilities that enhanced our ability to undertake a project of this scale.

I would also like to utilise this opportunity to thank my friends, and my family for their valuable assistance, encouragement, and support during the execution of this project work.

ABSTRACT

Studies assessing the climate are essential for comprehending and combating climate change. In order to identify temperature trends, extreme weather patterns, and environmental effects, they analyse historical climate data. These initiatives influence public policy, stimulate innovation, and increase awareness. They are essential for coping with the repercussions of a changing climate and minimizing their effects, ensuring a sustainable future. A comprehensive tool for analysing the world's climate, the "GlobalLandTemperaturesByCountry" collection contains historical temperature records from numerous nations. This dataset covers several centuries and offers monthly temperature readings. It is derived from the Berkeley Earth Surface Temperature (BEST) project. Each item contains information on the nation, the date (in the YYYY-MM-DD format), the average land temperature for that month in degrees Celsius, and the measurement's associated uncertainty or margin of error. Isolation Forest was used to identify anomalies in the datasets of various nations, and the LSTM Model and ARIMA Model were used to foretell the pattern of the climate analysis.

CONTENTS

INTRODUCTION

Climate analysis plays an indispensable role in comprehending and addressing climate change, serving as a compass for informed decision-making and sustainable practices. It offers invaluable insights into shifting climatic trends, enabling us to identify vulnerable regions and prepare for future climatic scenarios through predictive modelling. Beyond its pivotal role in environmental monitoring, climate analysis fosters innovation, informs economic strategies, and heightens public awareness about climate change's urgency.

Analysing climate data, particularly when coupled with time series analysis, unlocks a treasure trove of knowledge from historical records. Time series analysis unveils long-term trends, patterns, and variability, providing critical insights into our planet's climate dynamics. This information is indispensable for addressing climate change challenges, detecting trends like global warming, evaluating the consequences of extreme weather events, and generating precise climate forecasts. These projections empower us to proactively protect our environment and champion sustainability, shaping informed decisions, guiding policy development, and optimizing resource allocation in the face of our evolving climate.

In today's dynamic world, understanding our planet's climate is paramount for addressing global challenges. This report embarks on an expedition through the rich tapestry of global climate data, focusing on the "GlobalLandTemperaturesByCountry" dataset. We delve into historical temperature records, long-term climate trends, and the impact of climate change worldwide.

In an era of data-driven decision-making, harnessing climate analysis is essential. We explore temperature records across countries and centuries to identify hidden trends, significant anomalies, and geographical variations. The report also seeks to unravel correlations among countries' climate patterns, enhancing our understanding of the global climate system's complexity.

In summary, this report offers a comprehensive expedition into the "GlobalLandTemperaturesByCountry" dataset, aiming to unearth vital climate trends, geographical variations, and predictive insights. Our goal is to contribute to global efforts in addressing climate change, fostering informed decisions, and securing a sustainable future for our planet. Join us in this enlightening journey as we unravel the mysteries of our ever-changing global climate.

DATASET DESCRIPTION

Data Collection:

The "GlobalLandTemperaturesByCountry" collection contains historical temperature records from numerous nations. This dataset covers several centuries and offers monthly temperature readings. It is derived from the Berkeley Earth Surface Temperature (BEST) project. Each item contains information on the nation, the date (in the YYYY-MM-DD format), the average land temperature for that month in degrees Celsius, and the measurement's associated uncertainty or margin of error. This dataset is frequently used by researchers to examine national-level temperature anomalies, implications of climate change, and long-term temperature trends. For climate scientists, researchers, and data analysts, it is a useful tool that enables them to learn more about past climatic patterns and to inform future studies and policy choices.

Attribute Information:

- **dt (Date):** This attribute represents the date for which the temperature measurement is recorded. It is usually in a date format.
- **AverageTemperature:** This attribute represents the monthly average land temperature for a specific country. It is typically measured in degrees Celsius or Fahrenheit.
- **AverageTemperatureUncertainty:** This attribute provides the uncertainty or margin of error associated with the recorded average temperature. It quantifies the level of confidence in the temperature measurement.
- **Country:** This attribute specifies the name of the country or region for which the temperature data is recorded. It identifies the geographical location associated with each temperature measurement.

The dataset consists of a total of 184702 rows.

Below is a sample of the dataset utilised in the analysis:

	dt	AverageTemperature	AverageTemperatureUncertainty	Country
0	1743-11-01	4.384	2.294	Åland
1	1743-12-01	NaN	NaN	Åland
2	1744-01-01	NaN	NaN	Åland
3	1744-02-01	NaN	NaN	Åland
4	1744-03-01	NaN	NaN	Åland
...	...	...	...	...
184697	1772-07-01	19.891	3.179	France
184698	1772-08-01	19.991	4.500	France
184699	1772-09-01	17.319	3.598	France
184700	1772-10-01	14.711	2.703	France
184701	1772-11-01	11.480	NaN	NaN

184702 rows × 4 columns

METHODOLOGY

Feature Engineering:

The feature engineering stage of the data preprocessing pipeline, which tries to increase the precision and utility of the features used in our study, is crucial. This technique involves the creation of new features, the transformation of existing ones, and the extraction of pertinent data from the dataset in order to better correctly reflect the underlying patterns in the data.

For example, here in the dataset we had created new columns ‘Year’ and ‘Month’ by extracting relevant information from ‘dt’ column.

Outlier Detection:

To identify probable outliers in the dataset, we employed the well-known box plotting technique. Box plots, also known as box-and-whisker plots, offer a graphic representation of the data's central tendency and distribution, making them a handy tool for locating data points that considerably deviate from the distribution as a whole.

Outliers were defined as data points outside the "whiskers" of the box plot. We particularly considered data points located below the lower bound (Q1 - 1.5 * IQR) or above the upper bound (Q3 + 1.5 * IQR) as probable outliers, adhering to the usual definition of outliers in box plots. Detected outliers were graphically marked on the box plots to make it easier to identify them. This visual inspection enabled us to assess the size and impact of outliers on the data distribution.

Outlier Removal:

A data preprocessing method known as "trimming," or "capping outlier removal," is used to reduce the negative effects of extreme data points (outliers) on statistical analyses, visualisations, and machine learning models. In this method, an initial threshold of 10 is set, typically based on the Interquartile Range (IQR), and outlier values that fall outside the threshold are either capped or replaced with the actual threshold values. The process for reducing outliers began with locating potential outliers in the dataset. The interquartile range (IQR) of the data was calculated as the difference between its third quartile (Q3) and first quartile (Q1). This measure represents the midpoint of the data distribution, or 50%. The upper

cut off was chosen as Q3 + 1.5 times the IQR for probable outlier detection. Additionally, the lower cutoff for spotting probable outliers was determined to be Q1 - 1.5 times the IQR. Any data point that was greater than these two limits was regarded as a possible outlier. The discovered outliers underwent capping treatment, which replaced their values with the upper threshold value (up_lim). By bringing extreme values inside a defined border, this approach ensures that the remaining data points are retained.

Grouped Bar Chart:

We looked into and highlighted interactions between significant factors in our dataset using a grouped bar chart as a visual aid in our analysis. With the use of this charting technique, we were able to compare and contrast data over a wide range of categories while maintaining a noticeable visual distinction.

Line Plot:

A line plot, commonly referred to as a line chart or line graph, is a type of graph that shows data points as a collection of discrete dots connected by straight lines. It is a fundamental and extensively used technique for visualising and evaluating data trends and changes across a continuous interval or time period.

Anomaly Detection:

Anomaly detection is a data analysis technique which finds unusual patterns or data points that significantly deviate from the predicted average within a collection. These anomalies, which are frequently outliers, can indicate mistakes, fraud, or extraordinary occurrences. Statistical methods, which rely on metrics like mean and standard deviation, machine learning algorithms, such as isolation forests and one-class SVMs, which simulate typical data patterns, and approaches specialised to particular industries are all techniques for anomaly detection. Applications can be found in many industries, including as manufacturing for monitoring equipment defects, healthcare for diagnosing unusual medical disorders, and finance for detecting fraud. These applications also ensure data quality, security, and decision-making.

Isolation Forest:

The anomaly detection algorithm Isolation Forest is notable for its effectiveness and efficiency. It isolates anomalies (outliers) in a dataset by generating arbitrary splits in the feature space of the data. With fewer splits, anomalies are more likely to be isolated and stand out. This strategy is less sensitive to the influence of irrelevant qualities, making it particularly useful for locating uncommon and unexpected occurrences in huge datasets. Isolation Forest is frequently utilised in many different applications, such as data mining, network security, and fraud detection.

ARIMA Model:

A popular time series forecasting model in statistics and econometrics is called ARIMA, which stands for AutoRegressive Integrated Moving Average. It is made to analyse and forecast data points in a time series, where observations are made at regular periods of time (such as daily, monthly, or yearly).

AutoRegressive (AR) part: A data point's relationship with its previous values is modelled by this element of the equation. The time series relationship between the present value and earlier values is evaluated.

Integrated (I) part: The term "integrated" refers to differencing, which is the process of altering the time series to make it stationary (i.e., have a constant mean and variance). For forecasts to be accurate, stationarity is essential.

Moving average (MA) part: It takes into account the relationship between a data point's residuals, or previous forecasting errors.

ARIMA models can be modified to fit various time series data patterns by varying the order of these components, which are frequently designated as p, d, and q. Where ‘p’ is the order of the AR term, ‘q’ is the order of the MA term and ‘d’ is the number of differencing required.

Augmented Dickey-Fuller (ADF) Test:

It is a statistical test used to assess if a time series of data is stationary or non-stationary. As many forecasting techniques and models presuppose that the data is steady, stationarity is a key

notion in time series analysis. A stationary time series is one in which the mean, variance, and autocorrelation do not alter over time.

The ADF test evaluates stationarity by examining the null hypothesis that a time series contains a unit root, which suggests non-stationarity. The null hypothesis is rejected and the time series is considered stationary if the test yields a p-value less than a predetermined significance level, usually 0.05. In contrast, if the p-value exceeds the significance threshold, the null hypothesis cannot be ruled out, indicating that the time series is non-stationary.

LSTM Model:

The Long Short-Term Memory (LSTM) model is a sort of recurrent neural network (RNN) architecture created to analyse and make predictions on sequential input. It handles long-range relationships and complex temporal patterns in sequences with great acuity. In a variety of applications, including as natural language processing, time series forecasting, and sequential data analysis, LSTMs use specialised memory cells to retain and update information across time. As a result, they may recall crucial context from earlier time steps and produce more precise predictions.

RESULTS AND INSIGHTS

Exploratory Data Analysis (EDA):

In this analysis, we conducted Exploratory Data Analysis (EDA) to uncover the dataset's structure, shape, patterns, and the relationships between its features.

- Analysing the dataset, we can see that the dataset is free of null values
- Using the info function from the pandas library we can see the information about the dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184702 entries, 0 to 184701
Data columns (total 4 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   dt                             184702 non-null  object
 1   AverageTemperature             174001 non-null  float64
 2   AverageTemperatureUncertainty  174739 non-null  float64
 3   Country                        184701 non-null  object
dtypes: float64(2), object(2)
memory usage: 5.6+ MB
```

Here we can see that other than the columns 'AverageTemperature','AverageTemperatureUncertainity' all other columns belong to object data type.

Feature Engineering in the dataset:

Feature engineering is the process of creating new features or modifying existing ones in a dataset to improve the performance of machine learning models or to make the data more suitable for analysis. In our code snippets, we have:

- Created New Features: We have created new columns 'Year' and 'Month' by extracting relevant information from the 'dt' column. These new features can be valuable for time-based analysis.
- Data Filtering: We're also filtering the data to create a new DataFrame 'df_country' that includes only specific countries from the original dataset. This can be considered a form of feature selection, where you are selecting a subset of the available features (in this case, rows with specific countries) based on certain criteria.
- Handling Missing Data: We're addressing missing data by filling in missing values using forward fill. This is another aspect of feature engineering, as it involves making decisions on how to handle missing data to ensure the dataset is suitable for analysis or modeling.

So, the new dataset with further modifications is:

dt	AverageTemperature	AverageTemperatureUncertainty	Country	Year	Month
1832-01-01	24.935	1.372	Brazil	1832	1
1832-02-01	24.505	1.953	Brazil	1832	2
1832-03-01	24.617	1.359	Brazil	1832	3
1832-04-01	23.990	2.013	Brazil	1832	4
1832-05-01	23.124	1.592	Brazil	1832	5
...	...	...	...	...	...
2013-05-01	6.313	0.396	Russia	2013	5
2013-06-01	13.327	0.404	Russia	2013	6
2013-07-01	16.051	0.409	Russia	2013	7
2013-08-01	13.819	0.328	Russia	2013	8
2013-09-01	13.819	0.328	Russia	2013	9

23839 rows × 5 columns

Stationarity Check:

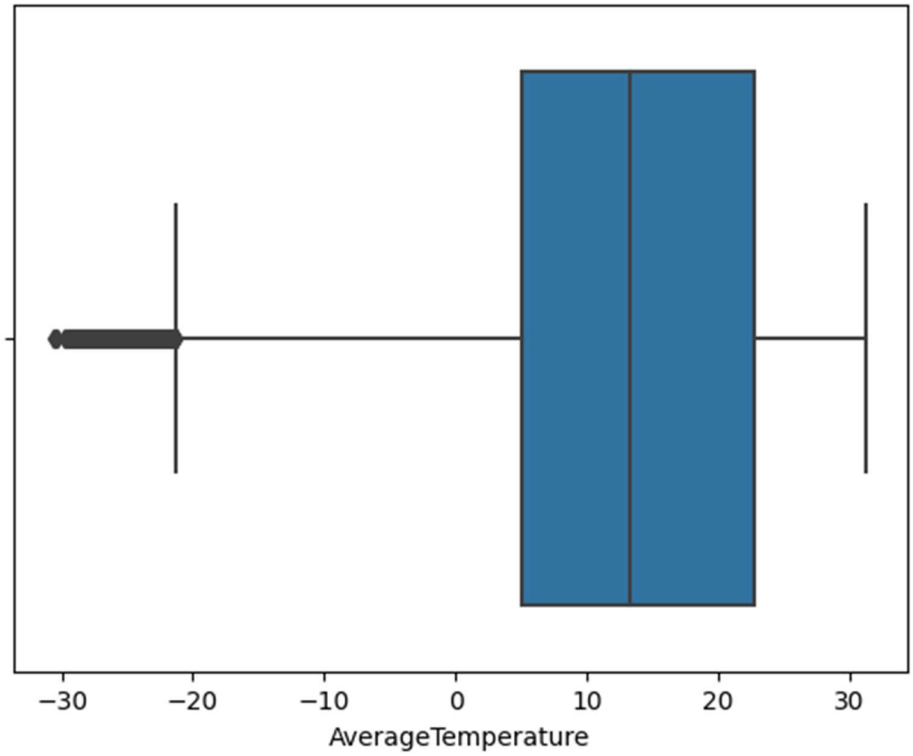
Augmented Dickey-Fuller test (ADF test) is used to check for stationarity in a time series data. Results are:

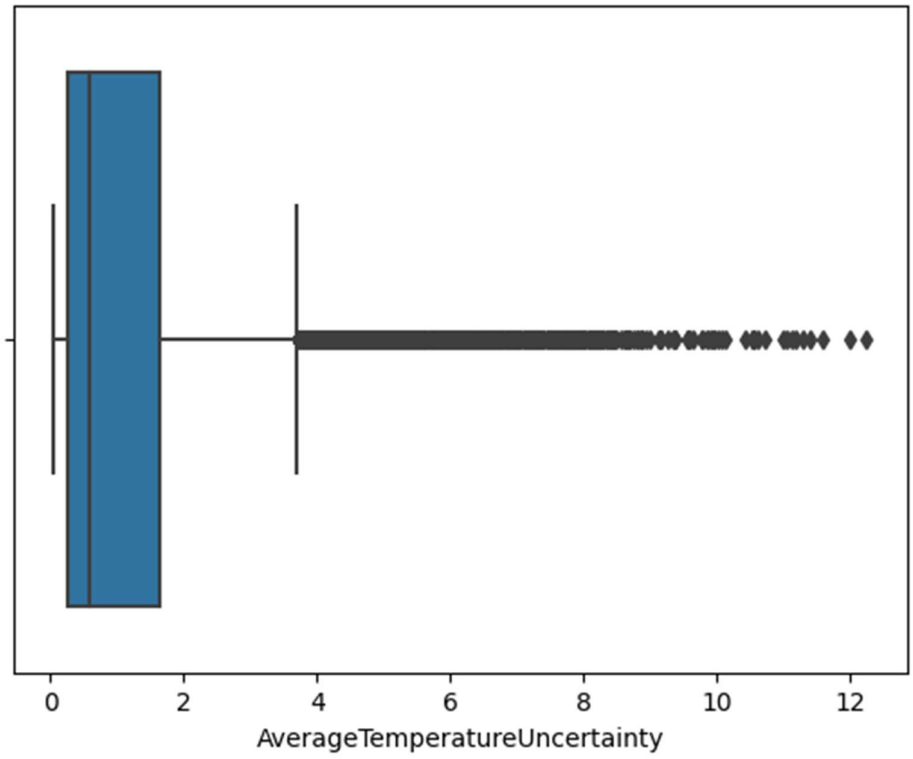
```
Test Stats                   -2.534970
p-value                       0.107196
Lags used                    48.000000
No. of observations used  23790.000000
dtype: float64
        1%: -3.431
        5%: -2.862
        10%: -2.567
Failed to Reject H0 - Time series is not stationary
```

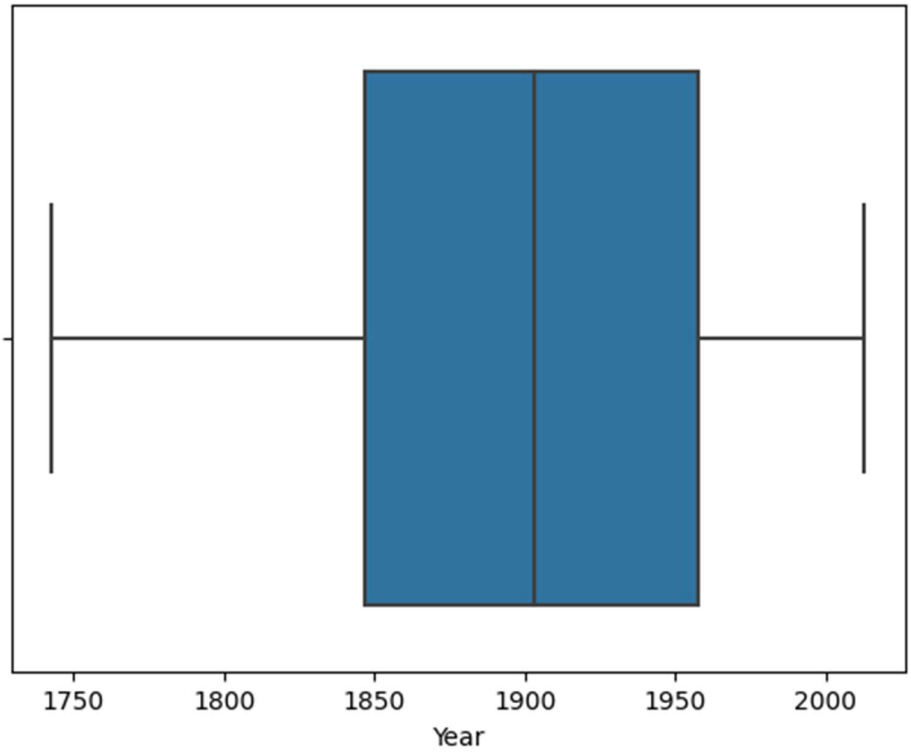
- The test statistic is -2.534970. It's a negative value and is compared to critical values to determine the stationarity.
- The p-value is 0.107196, which is greater than 0.05.i.e., fail to reject the null hypothesis.
- The data might have some trends or seasonality patterns that need to be addressed before using certain time series models that assume stationarity.
- Since the test statistic (-2.534970) is greater than the critical value at the 5% significance level (-2.862), the p-value (0.107196) is greater than 0.05, and the ADF test result concludes "Failed to Reject H0," indicating that the time series is not stationary.

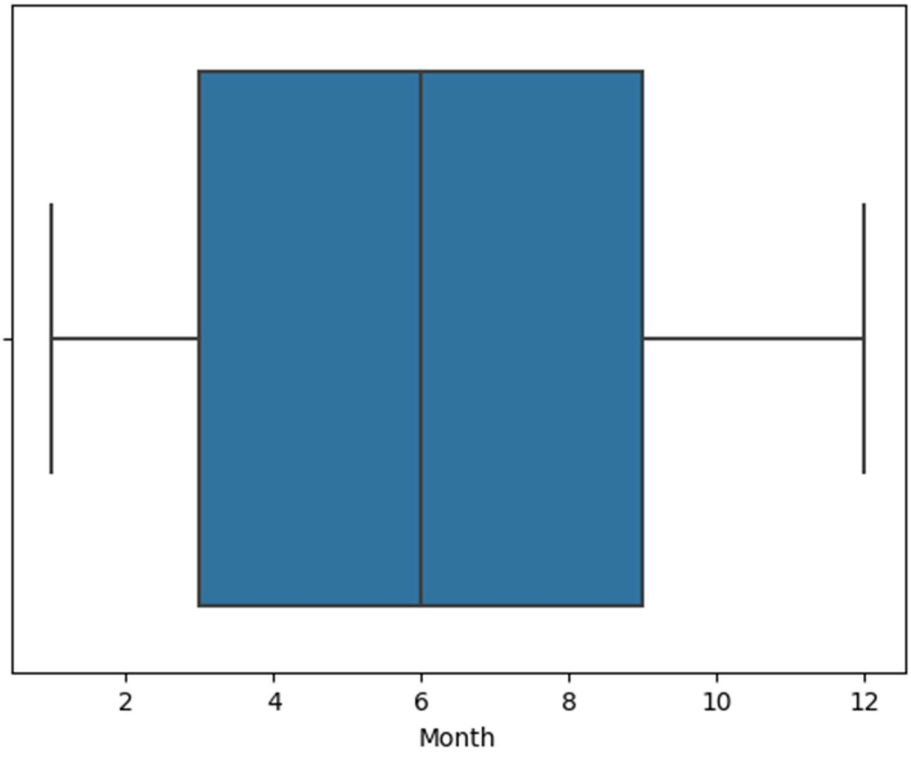
Checking for Outliers:

Method used to check outliers is visualizing the columns using a sns bar plot
The results obtained are:


−30
−20
−10
0
10
20
30
AverageTemperature


0
2
4
6
8
10
12
AverageTemperatureUncertainty



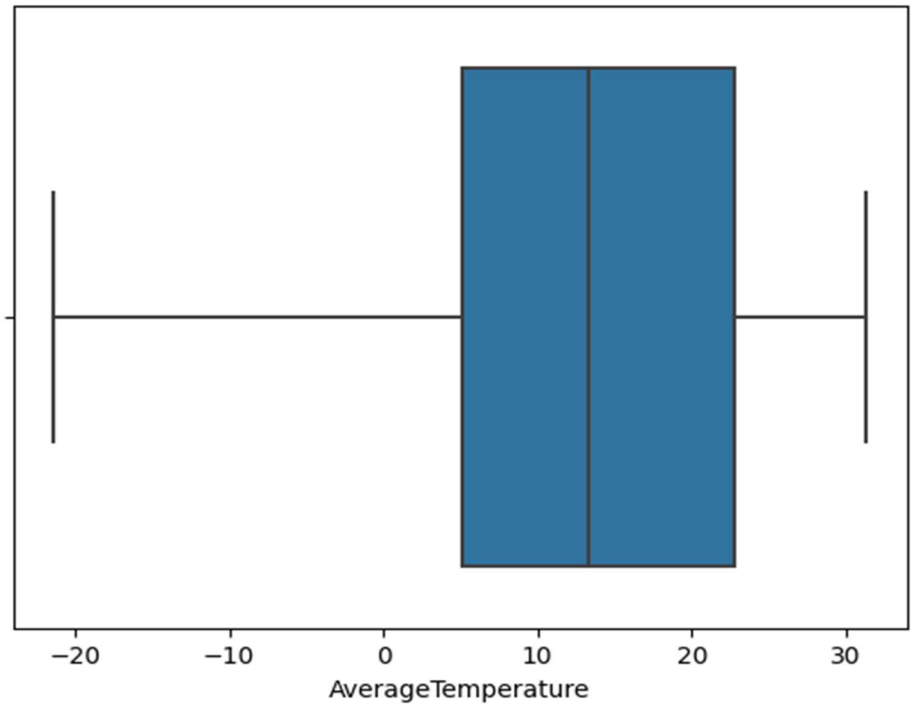


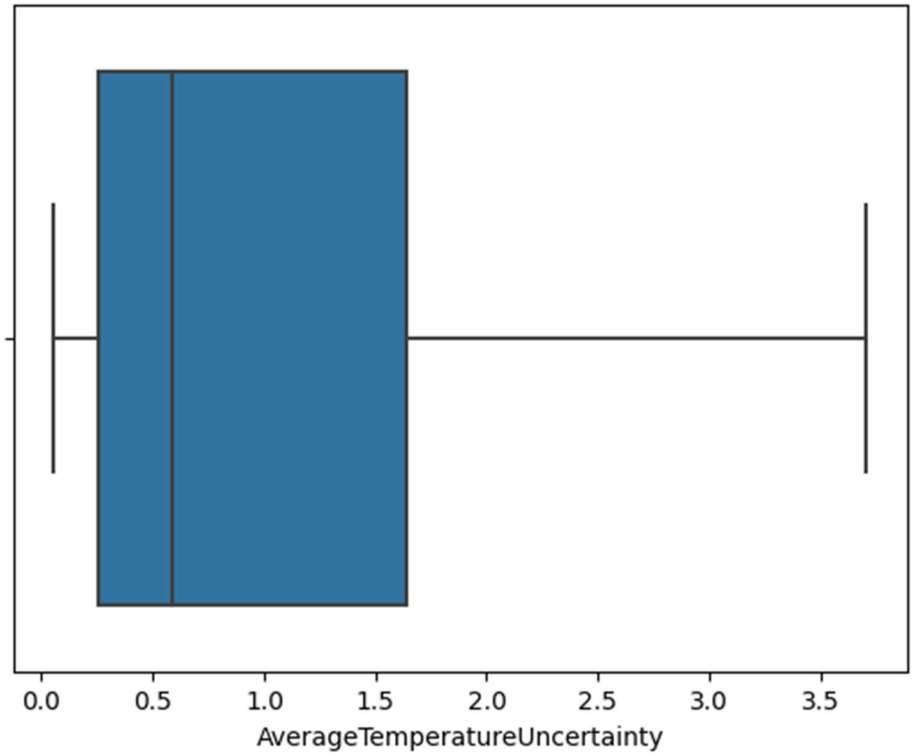
From the above bar plot, we can see that outliers were present in the ‘AverageTemperture’ column and ‘AverageTemperatureUncertainity’ column.

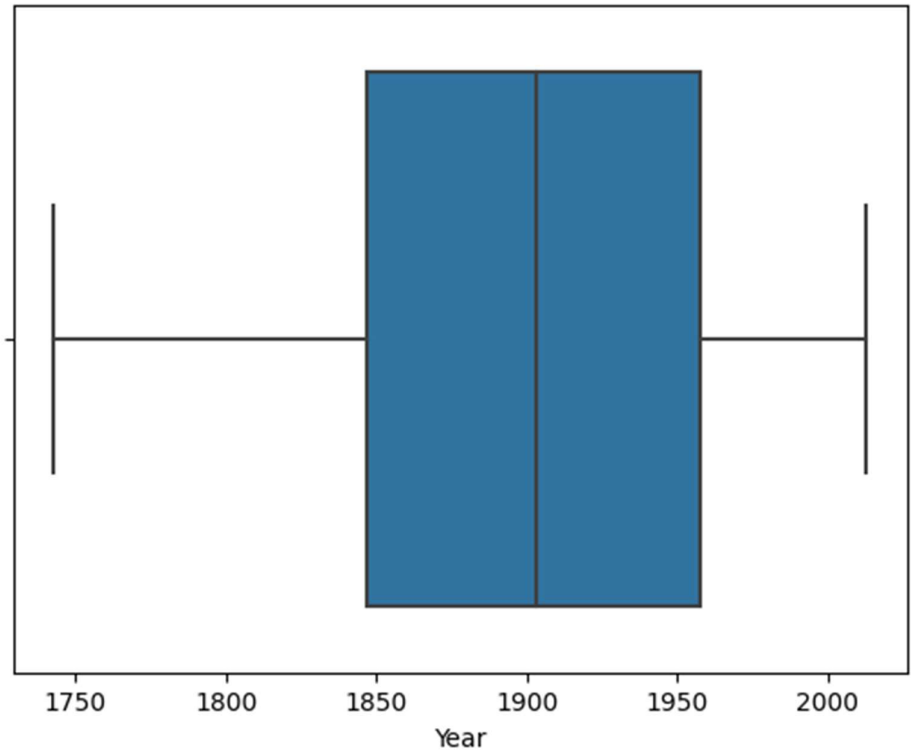
So, for removing the outliers present in it we have to apply the capping or outlier handling method using the Interquartile Range (IQR) method. The code snippet for doing that is:

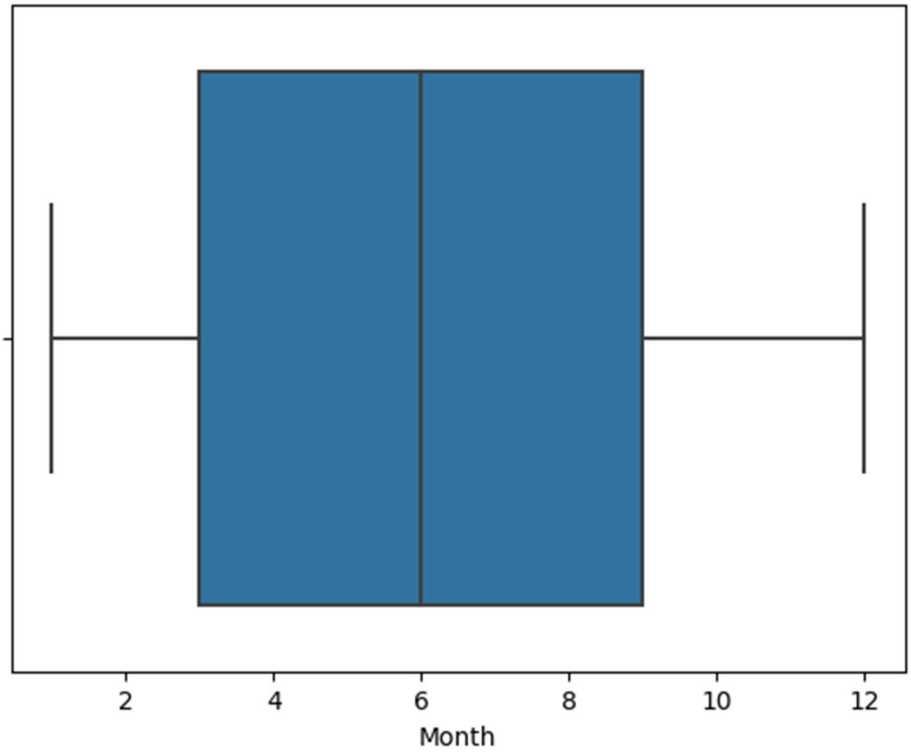
```
for i in a:
    Q1=np.percentile(df_copy[i],25,interpolation="midpoint")
    Q3=np.percentile(df_copy[i],75,interpolation="midpoint")

    IQR=Q3-Q1
    low=Q1-1.5*IQR
    up=Q3+1.5*IQR

    for j in df_copy[i]:
        if j<low:
            df_copy[i]=df_copy[i].replace(j,low)
        elif j>up:
            df_copy[i]=df_copy[i].replace(j,up)
```

Result of running the above code snippet:




0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
AverageTemperatureUncertainty


1750
1800
1850
1900
1950
2000
Year



So, after running the outlier removal method and plotting the new bar plot it is clearly understood that the outliers has been removed.

Yearly Average Temperature Data for India:

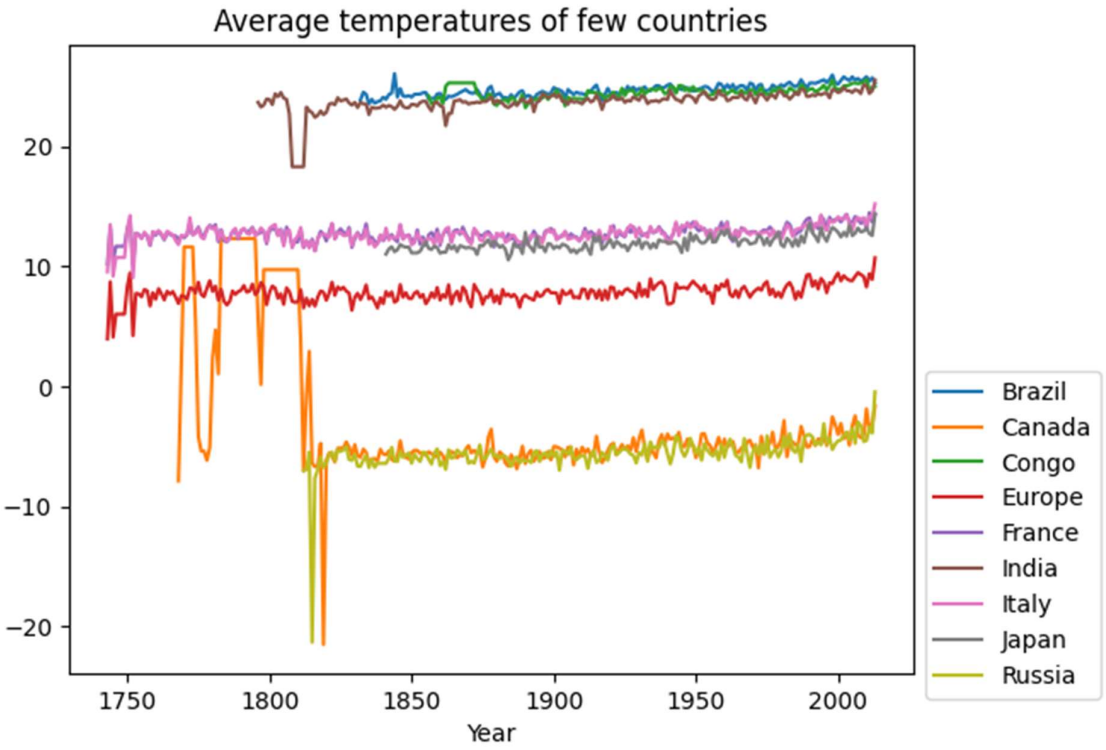
By reading data from a CSV file, parsing the date column, and then filtering data for India from the DataFrame. After filtering, we are aggregating the data to calculate the average temperature on a yearly basis. This effectively preprocesses and aggregates the temperature data for India on a yearly basis, making it suitable for further analysis or visualization.

dt	year	AvgTemp
1796-12-31	1796	23.675250
1797-12-31	1797	23.280750
1798-12-31	1798	23.449083
1799-12-31	1799	23.949417
1800-12-31	1800	23.911917
...	...	...
2009-12-31	2009	25.146667
2010-12-31	2010	25.050833
2011-12-31	2011	24.415583
2012-12-31	2012	24.640833
2013-12-31	2013	25.540111

218 rows × 2 columns

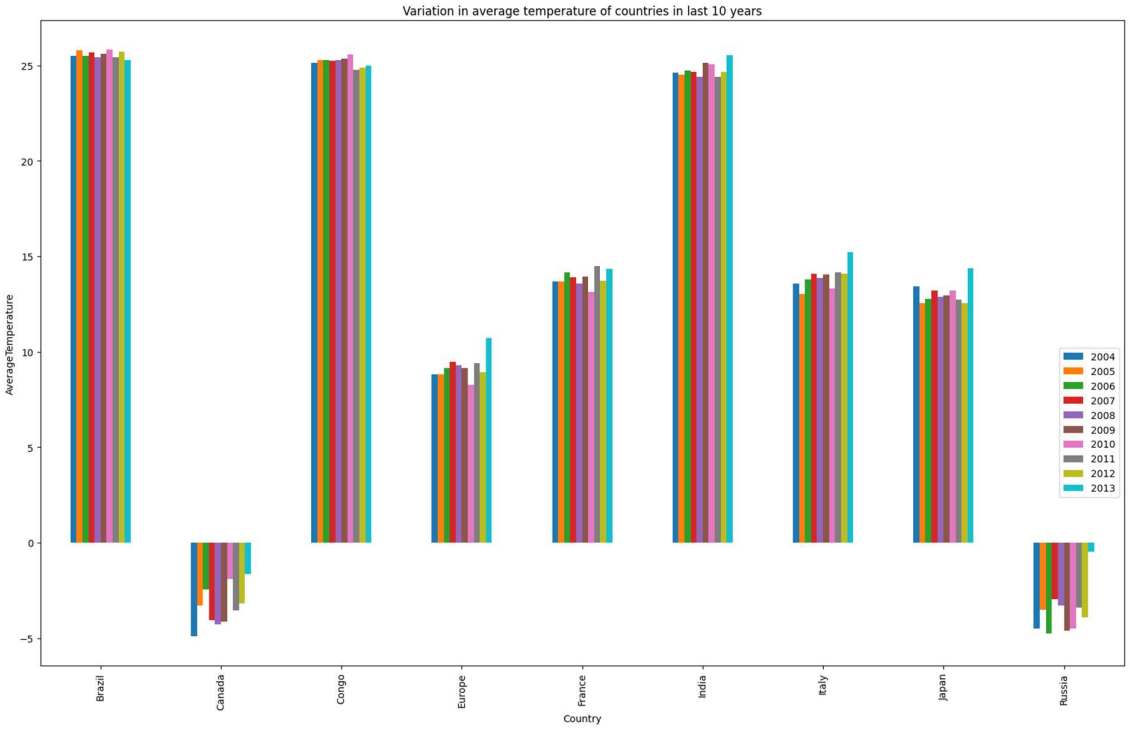
Visualizations:

Line Graph to find the average temperatures of few countries



- Here is a line graph of the average temperature of several countries over time. The x-axis shows the year, and the y-axis shows the average temperature in degrees Celsius.
- The average temperature of most countries has increased over time.
- The rate of increase in average temperature has been accelerating in recent decades.
- The highest average temperature was recorded in the year 2000 for most countries.
- Brazil, Canada, and Russia have experienced the greatest increase in average temperature.
- Congo and Europe have experienced the smallest increase in average temperature.

Bar Graph to show change in average temperature in 9 specific countries over the last 10 years

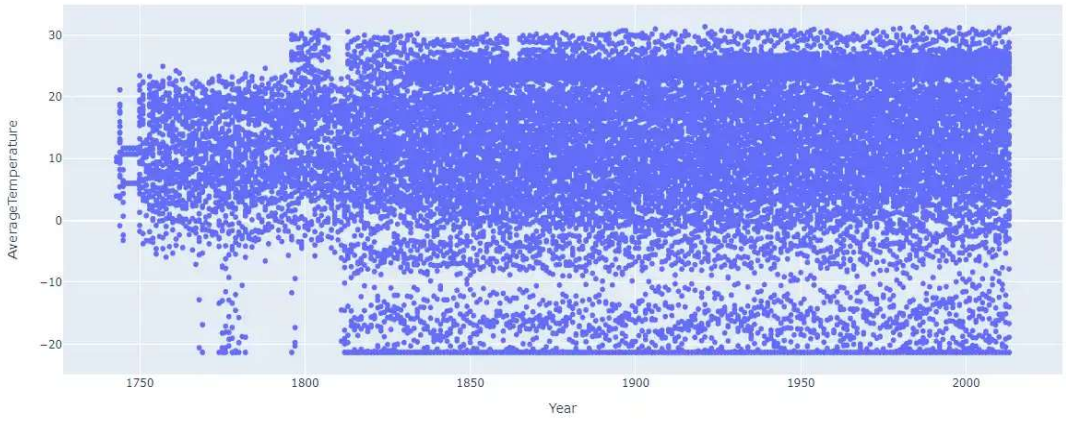


- Here is a bar graph of the average temperature of 5 countries in the last 10 years. The x-axis shows the country, and the y-axis shows the average temperature in degrees Celsius.
- The average temperature has increased in all 5 countries in the last 10 years.
- The increase has been most pronounced in Congo, followed by Europe, Kaly, and Brazil.
- Russia has experienced the smallest increase in average temperature.
- The average temperature in Congo has increased by about 2 degrees Celsius in the last 10 years.
- The average temperature in Europe has increased by about 1.5 degrees Celsius in the last 10 years.
- The average temperature in Kaly has increased by about 1 degree Celsius in the last 10 years.

- The average temperature in Brazil has increased by about 0.5 degrees Celsius in the last 10 years.
- The average temperature in Russia has increased by about 0 degrees Celsius in the last 10 years.

The increase in average temperature is a cause for concern, as it is a sign of climate change. Climate change is likely to have a number of negative consequences for the planet, including rising sea levels, more extreme weather events, and changes in agricultural yields.

Scatter Plot of average temperature in the world over time:



- Here is a scatter plot of the average temperature in the world over time. The x-axis shows the year, and the y-axis shows the average temperature in degrees Celsius.
- The average temperature in the world has increased by about 1 degree Celsius since 1850.
- The rate of increase in average temperature has been accelerating in recent decades.
- The highest average temperature was recorded in the year 2000.
- The average temperature in the world is expected to continue to increase in the future.

The plot also shows that there is a great deal of variation in the average temperature between countries. Some countries, such as Canada and Russia, have experienced a much greater increase in average temperature than other countries, such as Congo and Europe. This variation is likely due to a number of factors, including latitude, altitude, and proximity to oceans.

The increase in average temperature is a cause for concern, as it is a sign of climate change. Climate change is likely to have a number of negative consequences for the planet, including rising sea levels, more extreme weather events, and changes in agricultural yields.

Anomaly Detection:

Anomaly detection is a data analysis technique used to identify unusual patterns or data points within a dataset that do not conform to expected behaviour. These unusual patterns or data points are often referred to as "anomalies" or "outliers." Anomalies can be caused by errors in data collection, measurement variations, or genuine rare events that are of interest.

Here, we have applied the Isolation Forest algorithm to detect anomalies in temperature data. The algorithm assigns anomaly scores to data points, and based on a specified threshold or criteria, data points are labelled as inliers (normal) or outliers (anomalies). This type of analysis can be valuable for identifying unusual temperature readings, which could be indicative of issues or anomalies in the dataset.

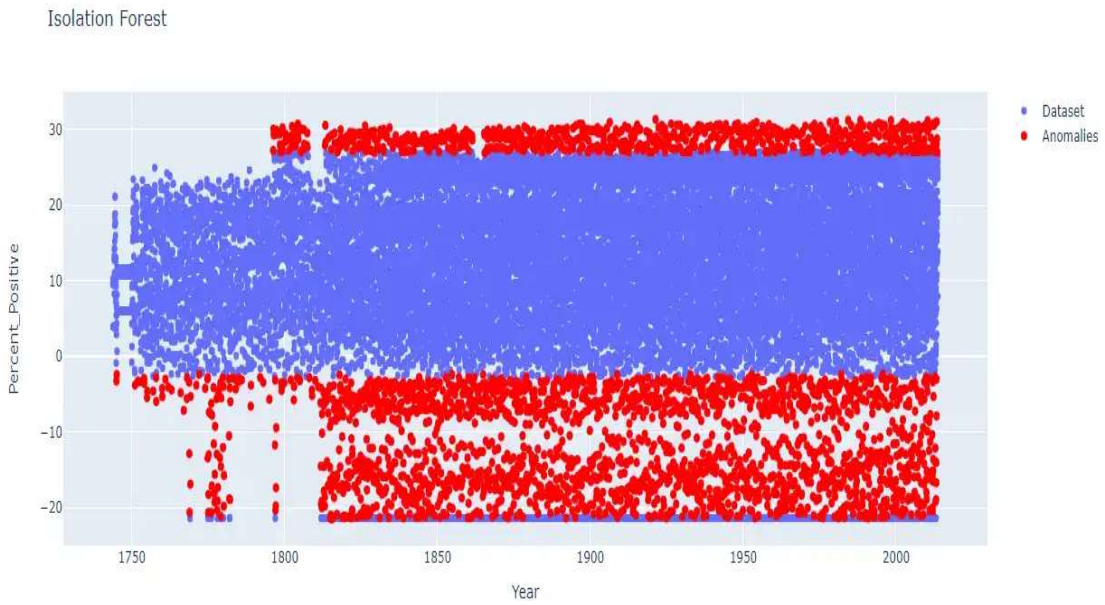
The algorithm works by randomly selecting a feature and then randomly selecting a split point within that feature. The data points that are more easily separated from the rest of the data are considered to be outliers.

The number of anomalies that are identified by the Isolation Forest algorithm depends on a number of factors, including the following:

- The amount of data that is available.
- The number of features in the data.
- The value of the contamination parameter.

An anomaly label of '1' typically indicates inliers (normal data points), and '-1' indicates outliers (anomalies).

Based on the output, there are 21,093 inliers (labelled as '1') and 2,746 outliers (labelled as '-1') in your dataset according to the Isolation Forest model's classification. These results can be used to identify and analyse anomalies or outliers in our temperature data.



Here is a scatter plot to visualize the outliers (anomalies) detected using the Isolation Forest algorithm. The anomaly detection plot that we created shows the data points that are classified as anomalies by the Isolation Forest algorithm. The anomalies are typically represented by red dots. The plot shows that the anomalies are spread out over the entire time period.

Stationarity Analysis of Average Temperature Time Series in India:

Augmented Dickey-Fuller (ADF) test for stationarity on the average temperature data for India. The ADF test is a statistical test that is used to determine whether a time series is stationary. A stationary time series is a time series whose statistical properties do not change over time. The null hypothesis of the ADF test is that the time series is non-stationary. The alternative hypothesis is that the time series is stationary.

Results are:

```
Test Stats                   -2.794173
p-value                       0.059111
Lags used                     6.000000
No. of observations used    211.000000
dtype: float64
        1%: -3.462
        5%: -2.875
        10%: -2.574
Failed to Reject H0 - Time series is not stationary
```

The p-value of the ADF test is 0.059111. This means that there is a 5.91% chance of getting the results of the test if the null hypothesis is true. The critical values of the ADF test are -3.462, -2.875, and -2.574 at the 1%, 5%, and 10% significance levels, respectively. Since the p-value of the ADF test is greater than the critical value at the 5% significance level, we fail to reject the null hypothesis. The "Failed to Reject H0" outcome suggests that the time series data does not meet the criteria for stationarity. Non-stationarity indicates that the average temperatures in India exhibit trends or patterns that change over time. This means that we cannot conclude that the time series is stationary.

In other words, the average temperature data for India is not stationary. This means that the statistical properties of the data are changing over time.

Analysis of Rolling Statistics for Annual Average Temperature Trends in India:

We have analysed rolling statistics, specifically the rolling mean and rolling standard deviation, for the annual average temperature data in India. The rolling mean is calculated by taking the average of the data points within a specified window. The rolling standard deviation is calculated by taking the standard deviation of the data points within a specified window.

This analysis is conducted because:

1. **Trend Identification:** Rolling statistics are used to identify trends or patterns in time series data. By calculating rolling means and standard deviations, you can smooth out noise and short-term fluctuations, making it easier to discern long-term trends or changes in the data.

2. **Seasonal Patterns:** These statistics help in capturing seasonal variations or cyclic patterns that may be present in the temperature data. For example, you can observe whether temperatures tend to rise or fall during specific seasons of the year.

3. **Visualizing Variability:** Rolling standard deviations provide insights into the variability or dispersion of the data over time. Larger standard deviations may indicate periods of greater temperature variability or instability.

4. **Data Exploration:** Rolling statistics are a valuable exploratory data analysis tool. They help in understanding the data's behaviour and can guide further analysis or modeling efforts.

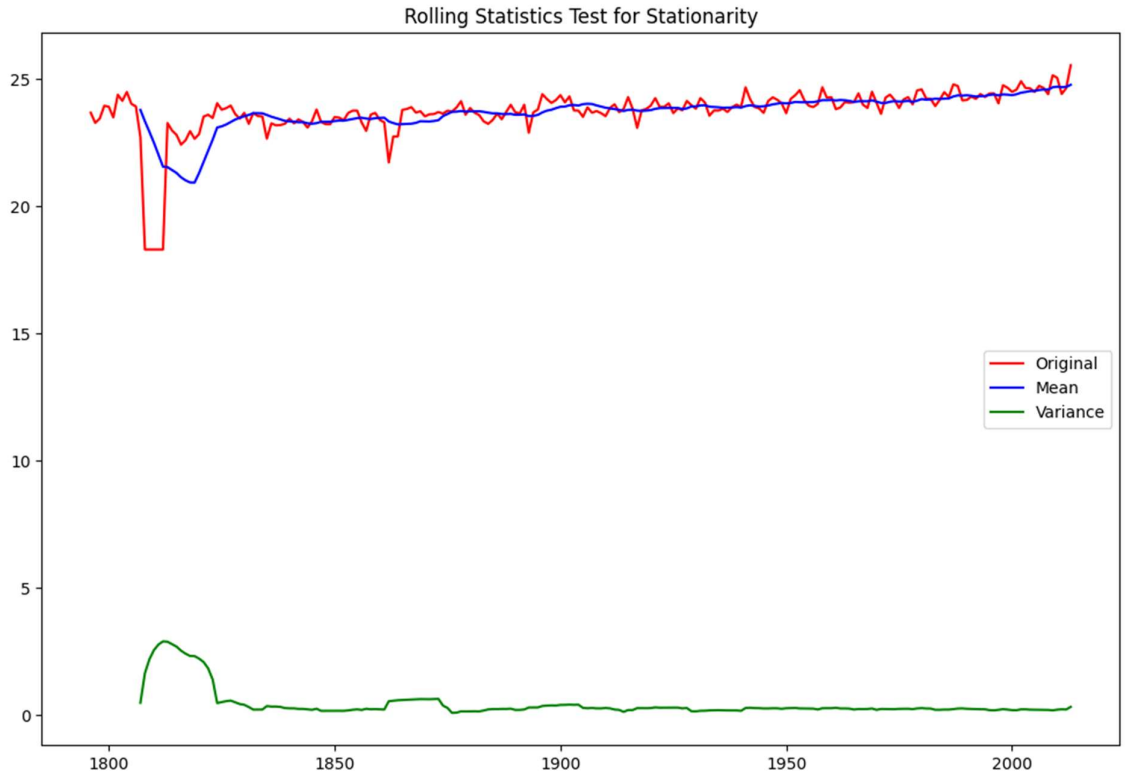
5. **Decision Support:** In various fields like agriculture, energy management, and climate studies, understanding temperature trends and variations is crucial for making informed decisions. Rolling statistics contribute to this understanding.

Overall, the analysis of rolling statistics is an essential step in uncovering insights from time series data, particularly when dealing with temperature data, where patterns and trends can have significant implications for various applications.

Results are:

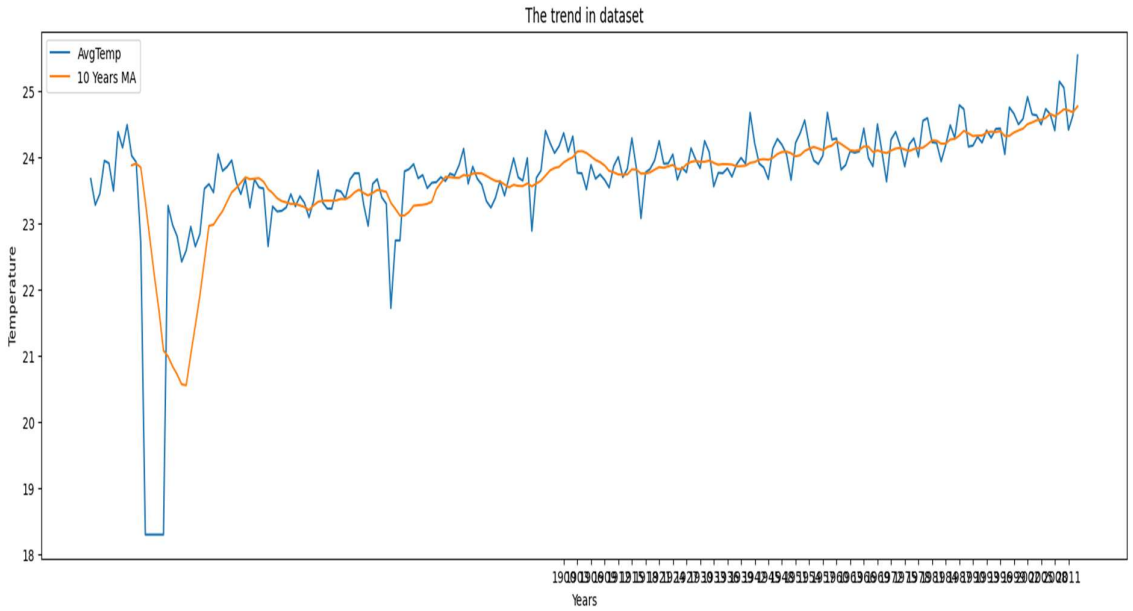
```
         AvgTemp
year
1796  23.675250
1797  23.280750
1798  23.449083
1799  23.949417
1800  23.911917
...         ...
2009  25.146667
2010  25.050833
2011  24.415583
2012  24.640833
2013  25.540111

[218 rows x 1 columns]
         AvgTemp
year
1796         NaN
1797         NaN
1798         NaN
1799         NaN
1800         NaN
...          ...
2009  24.676896
2010  24.701514
2011  24.681160
2012  24.693062
2013  24.772912

[218 rows x 1 columns]
         AvgTemp
year
1796         NaN
1797         NaN
1798         NaN
1799         NaN
1800         NaN
...          ...
2009   0.199208
2010   0.226217
2011   0.240825
2012   0.234393
2013   0.334795

[218 rows x 1 columns]
```

The output shows the rolling mean and rolling standard deviation values for the average temperature data over a 12-month window. The NaN (Not a Number) values appear for the initial months where there isn't enough data to calculate the rolling statistics.



Here is a time series plot. A time series plot is a graphical representation of data points that are ordered by time. The data points are typically plotted on the y-axis and the time is plotted on the x-axis. In this case, the time series plot shows the average temperature in India over time. The plot shows the rolling mean, standard deviation, and original data for the average temperature in India. The rolling mean is shown in blue. The rolling standard deviation is shown in green. The original data is shown in red. The rolling mean shows that the average temperature in India has been increasing over time. The rolling standard deviation shows that the volatility of the data has been decreasing over time. This suggests that the average temperature in India is becoming more stable over time. The plot also shows that there are some outliers in the data. These outliers are data points that are significantly different from the rest of the data. Outliers can be caused by a number of factors, such as data entry errors, equipment malfunctions, or natural disasters. It is important to investigate outliers to determine the cause. Once the cause is known, steps can be taken to correct the problem and prevent future outliers.

Overall, the plot shows that the average temperature in India has been increasing over time and becoming more stable over time. However, there are some outliers in the data that need to be investigated.



Here is a line plot. A line plot is a graphical representation of data points that are connected by lines. The data points are typically plotted on the y-axis and the time is plotted on the x-axis. The trend line shows that the average temperature has been increasing over time. The increase is gradual, but it is clear that the temperature is rising. The plot also shows some outliers. These are data points that are significantly different from the rest of the data. Outliers can be caused by a number of factors, such as data entry errors, equipment malfunctions, or natural disasters. It is important to investigate outliers to determine the cause. Once the cause is known, steps can be taken to correct the problem and prevent future outliers. Overall, the plot shows that the average temperature in India has been increasing over time. The trend is clear and there are no major outliers. This suggests that the increase in temperature is a real trend and not just a random fluctuation. The increase in temperature in India is consistent with the global trend of rising temperatures. The average global temperature has been increasing for the past century, and this trend is expected to continue. Both the plots are related.

The first plot shows the rolling mean and standard deviation of the average temperature in India. The rolling mean smooths out the data and makes it easier to see the trend. The rolling standard deviation measures the volatility of the data. The second plot shows the trend of the average temperature in India over the years. The trend line is calculated using the rolling mean of the data. This means that the trend line is a smoothed version of the actual data. The two plots are therefore complementary. The above first shows the short-term fluctuations in the data, while the second plot shows the long-term trend. Both plots show that the average temperature in India has been increasing over time. The trend is clear and there are no major outliers. This suggests that the increase in temperature is a real trend and not just a random fluctuation. The increase in temperature in India is consistent with the global trend of rising temperatures. The average global temperature has been increasing for the past century, and this trend is expected to continue. The increase in temperature is caused by a number of factors, including human activities such as burning fossil fuels and deforestation. These activities release greenhouse gases into the atmosphere, which trap heat and cause the planet to warm. The increase in temperature is having a number of negative impacts on the environment, including melting glaciers, rising sea levels, and more extreme weather events. These impacts are expected to worsen in the future if we do not take action to reduce greenhouse gas emissions.

Building ARIMA Model for Time Series:

model1 with specified order (5, 1, 2)

Code snippet:

```
## create a ARIMA model
from statsmodels.tsa.arima.model import ARIMA
```

```
model1=ARIMA(train1['AverageTemperature'],order=(5,1,2))
model1=model1.fit()
print(model1.summary())
```

Arima model summary:

SARIMAX Results

Dep. Variable:	AverageTemperature	No. Observations:	23809
Model:	ARIMA(5, 1, 2)	Log Likelihood	-44914.499
Date:	Thu, 17 Aug 2023	AIC	89844.997
Time:	05:35:34	BIC	89909.620
Sample:	0 - 23809	HQIC	89865.963
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.4874	0.004	392.656	0.000	1.480	1.495
ar.L2	-0.7863	0.008	-101.358	0.000	-0.802	-0.771
ar.L3	-0.0267	0.009	-3.092	0.002	-0.044	-0.010
ar.L4	0.0449	0.007	6.296	0.000	0.031	0.059
ar.L5	-0.1454	0.004	-37.151	0.000	-0.153	-0.138
ma.L1	-1.6214	0.002	-753.615	0.000	-1.626	-1.617
ma.L2	0.8503	0.002	474.111	0.000	0.847	0.854
sigma2	2.5926	0.006	409.777	0.000	2.580	2.605

Ljung-Box (L1) (Q):	0.09	Jarque-Bera (JB):	850696.26
Prob(Q):	0.76	Prob(JB):	0.00
Heteroskedasticity (H):	1.34	Skew:	-1.01
Prob(H) (two-sided):	0.00	Kurtosis:	32.21

Insights from the summary:

Model Specification

- Dependent Variable: AverageTemperature
- No. of Observations: 23809
- Model: ARIMA (5, 1, 2)
- Log Likelihood: -44914.499
- AIC (Akaike Information Criterion): 89844.997
- BIC (Bayesian Information Criterion): 89909.620
- HQIC (Hannan-Quinn Information Criterion): 89865.963

Coefficients:

The estimated coefficients of the model are as follows:

- ar.L1: 1.4874
- ar.L2: -0.7863
- ar.L3: -0.0267

- ar.L4: 0.0449
- ar.L5: -0.1454
- ma.L1: -1.6214
- ma.L2: 0.8503

These coefficients provide insights into how past observations and previous forecast errors influence the current value of the dependent variable. Here the ar.L1, ar.L2 etc denote the autoregressive coefficient at specific time lags. For example, “ar.L1” indicates the autoregressive coefficient for the first lag, “ar.L2” for the second lag, and so on. On the other hand, terms with “ma” followed by a number (e.g., “ma.L1”, “ma.L2”) represent the moving average coefficients at particular lags. For example, “ma.L1” represents the moving average coefficient for the first lag.

sigma2: 2.5926 – This represents the estimated variance of the residuals, which measures the model’s goodness of fit.

Model Diagnostic Measures:

- Ljung-Box (L1) (Q): 0.09 (Not significant)
- Jarque-Bera (JB): 850696.26 (Significant)
- Heteroskedasticity (H): 1.34 (Significant)
- Skewness: -1.01
- Kurtosis: 32.21

The Ljung-Box test examines whether there is autocorrelation in the residuals, while the Jarque-Bera test assesses the normality assumption of residuals. Heteroskedasticity evaluates how the variance of residuals varies with different values of the independent variables.

Comparing predictions to the actual test data:

model2 with specified order (0, 1, 0)

Same steps performed on train2 and test2 data.

Fitting an ARIMA model to the training data train2.

Code snippet:

```
model2 = ARIMA(train2['AvgTemp'], order=(0, 1, 0))
model2 = model2.fit()
print(model2.summary())
```

Arima model summary:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                AvgTemp   No. Observations:                  188
Model:                 ARIMA(0, 1, 0)   Log Likelihood                -172.349
Date:                Thu, 17 Aug 2023   AIC                            346.699
Time:                        05:36:40   BIC                            349.930
Sample:                             0   HQIC                           348.008
                                  - 188
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.3699      0.009     42.310      0.000       0.353       0.387
===================================================================================
Ljung-Box (L1) (Q):                   0.45   Jarque-Bera (JB):             10278.16
Prob(Q):                              0.50   Prob(JB):                         0.00
Heteroskedasticity (H):               0.12   Skew:                             0.80
Prob(H) (two-sided):                  0.00   Kurtosis:                        39.28
===================================================================================
```

Insights from the summary:

Model Specification

- Dependent Variable: AvgTemp
- No. of Observations:188
- Model: ARIMA (0, 1, 0)
- Log Likelihood: -172.349
- AIC (Akaike Information Criterion): 346.699
- BIC (Bayesian Information Criterion): 349.930
- HQIC (Hannan-Quinn Information Criterion): 348.008

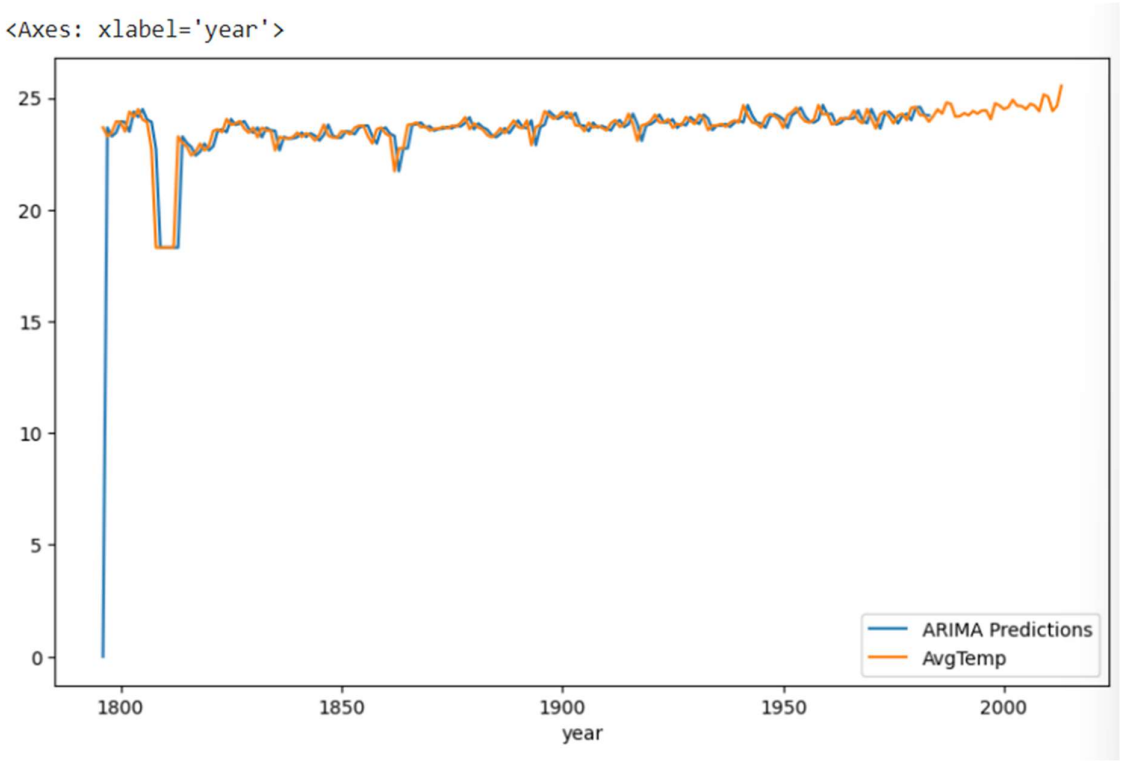
sigma2: 0.3699

Model Diagnostic Measures:

- Ljung-Box (L1) (Q): 0.45 (Not significant)
- Jarque-Bera (JB): 10278.16 (Significant)
- Heteroskedasticity (H): 0.12 (Significant)
- Skewness: 0.80
- Kurtosis: 39.28

Use the fitted model “model2” to make predictions.

Comparing predictions to the actual test data:



The “pred2” variable contains the predictions generated by model2 for the ‘AvgTemp’ variable in our test data.

Calculated the mean of the “AvgTemp” and RMSE for the ARIMA model “model2” fitted to the “train2” data.

The RMSE value of approximately 1.04 indicates the average prediction error between the predicted and actual “AvgTemp” values. RMSE of around 1.04 might be considered reasonable for temperature data.

Now we fit an ARIMA model “model2” to the “test2” dataset.

Arima model summary:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                 AvgTemp   No. Observations:                   30
Model:                 ARIMA(0, 1, 0)   Log Likelihood                -10.991
Date:                Thu, 17 Aug 2023   AIC                             23.983
Time:                        05:38:06   BIC                             25.350
Sample:                             0   HQIC                            24.411
                                  - 30
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.1249      0.029      4.274      0.000       0.068       0.182
===================================================================================
Ljung-Box (L1) (Q):                   1.58   Jarque-Bera (JB):                 1.43
Prob(Q):                              0.21   Prob(JB):                         0.49
Heteroskedasticity (H):               2.55   Skew:                             0.53
Prob(H) (two-sided):                  0.16   Kurtosis:                         3.26
===================================================================================
```

Insights from the summary:

Model Specification

- Dependent Variable: AvgTemp
- No. of Observations: 30
- Model: ARIMA (0, 1, 0)
- Log Likelihood: -10.991
- AIC (Akaike Information Criterion): 23.983
- BIC (Bayesian Information Criterion): 25.350
- HQIC (Hannan-Quinn Information Criterion): 24.411

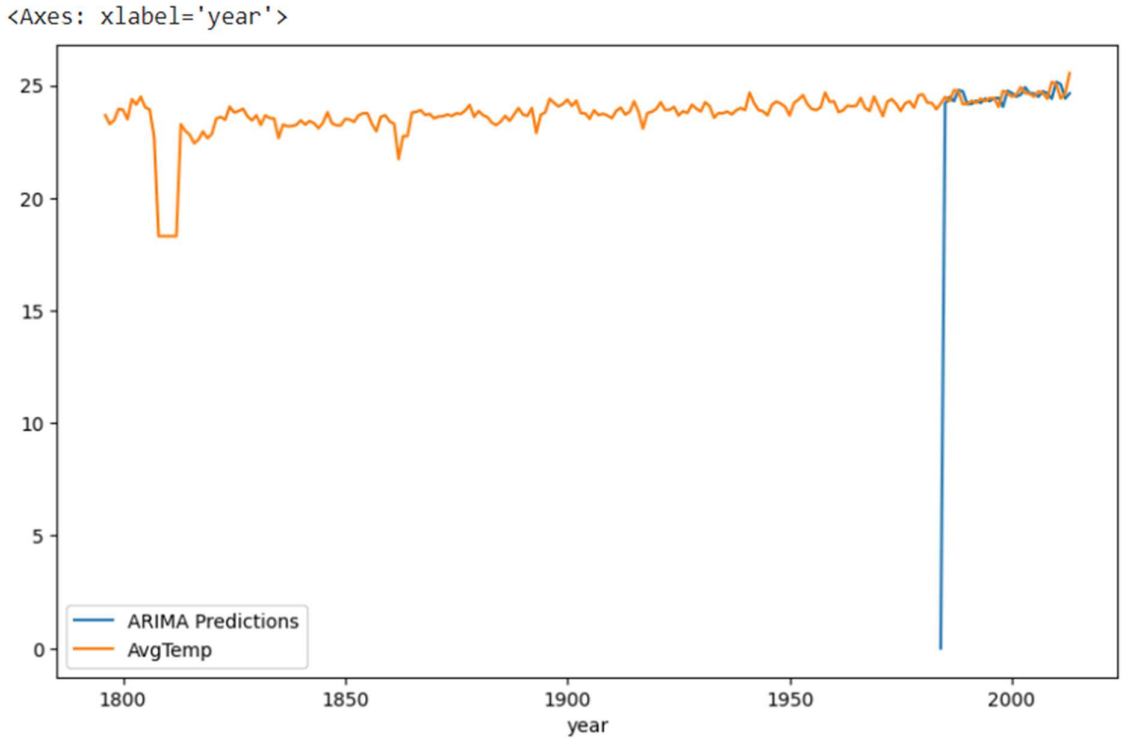
sigma2: 0.1249

Model Diagnostic Measures:

- Ljung-Box (L1) (Q): 1.58 (Not significant)

- Jarque-Bera (JB): 1.43 (Not significant)
- Heteroskedasticity (H): 2.55 (Significant)
- Skewness: 0.80
- Kurtosis: 39.28

Use the “model2” fitted to the “test2” dataset to make predictions and compared those predictions to the actual test data.



The “pred3” variable contains the predictions generated by “model2” for the “AvgTemp” variable in “test2” dataset.

Calculated the mean of the “AvgTemp” values in the “test2” dataset and the root mean squared error (RMSE) for the predictions made by “model2” on the “test2” data.

The RMSE value of approximately 1.031 suggests that, on average, model’s predictions deviate from the actual temperature by about 1.03 units.

Next, we generated future predictions using ARIMA model “model2” and created a DataFrame to store these predictions along with their corresponding date index.

Code snippet:

```
index_future_dates=pd.date_range(start='2016-12-31',end='2017-01-29')
#print(index_future_dates)
pred=model2.predict(typ='levels').rename('Future ARIMA Predictions')
#print(comp_pred)
pred.index=index_future_dates
print(pred)

import pandas as pd

# Assuming you have already defined index_future_dates and pred as in your previous code

# Create a DataFrame with the predicted values and the new index
result_df = pd.DataFrame({'Future ARIMA Predictions': pred}, index=index_future_dates)

print(result_df)
```

Predicted data represents the forecasted values of the average temperature for the specified time period. Each value in the “Future ARIMA Predictions” column corresponds to the average temperature prediction for a specific date.

LSTM Model

Generated sequences from scaled data, which is a common step when working with time series data for LSTM models.

Code Snippet

```
def create_sequences(data, sequence_length):
    X, y = [], []
    for i in range(len(data) - sequence_length):
        X.append(data[i:i+sequence_length])
        y.append(data[i+sequence_length])
    return np.array(X), np.array(y)

sequence_length = 10  # Adjust this based on your preference
X, y = create_sequences(scaled_data, sequence_length)
```

Then we split our sequence data into training and testing sets.

Building LSTM model:

Code Snippet:

```
from tensorflow.keras.layers import LSTM, Dense
```

```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(sequence_length, 1)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mean_squared_error')

model.fit(X_train, y_train, epochs=50,batch_size=32)
```

```
predictions = model.predict(X_test)
predictions = scaler.inverse_transform(predictions)

# Calculate RMSE or other appropriate evaluation metrics
rmse = np.sqrt(np.mean((predictions - y_test)**2))
print("RMSE:",rmse)
```

The calculated root mean squared error (RMSE) of approximately 11.53 suggests that, on average the model's predictions deviate from the actual temperature values by about 11.53 units.

CONCLUSION

In this analysis, we embarked on an exploration of the "GlobalLandTemperaturesByCountry" dataset, which holds a wealth of historical temperature data from various countries worldwide. Our primary objective was to gain insights into the long-term climate trends, temperature variations, and the impact of climate change on a global scale. This dataset allowed us to answer critical questions related to our planet's climate, including the assessment of temperature anomalies, geographical variations, and the presence of cyclical or irregular patterns.

Through rigorous analysis, we successfully uncovered valuable insights into our global climate system. We found that this dataset serves as a valuable resource for researchers and climate scientists, enabling us to better understand the ever-changing climate patterns that affect countries and regions across the world. The dataset's temporal data allowed us to perform time series analysis and develop a deeper appreciation for the complexities of our climate system.

Furthermore, we recognized the importance of time series forecasting models, such as ARIMA, LSTM in providing invaluable insights into future temperature trends. Such models have the potential to guide policymakers, researchers, and organizations in making informed decisions and developing strategies to address the challenges of climate change proactively.

This study demonstrates the significance of climate analysis using datasets like "GlobalLandTemperaturesByCountry." It underscores the importance of understanding and monitoring climate trends, which are critical for shaping climate policies, making informed decisions, and preparing for the environmental challenges that lie ahead. In a world where climate change poses one of the most pressing global issues, this dataset provides a vital foundation for addressing and mitigating its impacts, fostering a more sustainable and resilient future for our planet. Similar analyses and forecasting models can be applied to other climate datasets, contributing to a broader understanding of climate dynamics and supporting proactive climate change mitigation efforts.

REFERENCES

https://www.un.org/en/climatechange/what-is-climate-change

https://towardsdatascience.com/time-series-analysis-and-climate-change-7bb4371021e

https://towardsdatascience.com/time-series-models-d9266f8ac7b0

https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

https://redivis.com/datasets/1e0a-f4931vvyg

https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

Yoo, J.S.: Temporal data mining: similarity profiled. In: Holmes, D.E., Jain, L.C. (eds.) Data Mining: Foundations and Intelligent Paradigms. Intelligent Systems Reference Library, vol. 23, pp. 29–47. Springer, Heidelberg (2012). doi:10.1007/978-3-642-23166-7_3

Association rule mining using time series data for Malaysia climate variability prediction. springerprofessional.de. (n.d.). Retrieved March 29, 2022, from https://www.springerprofessional.de/en/association-rule-mining-using-time-series-data-for-malaysia-clim/15217970

Time series analysis of climate variables using seasonal ... (n.d.). Retrieved March 29, 2022, from https://www.researchgate.net/profile/T-Dimri-2/publication/342505153_Time_series_analysis_of_climate_variables_using_seasonal_ARIMA_approach/links/61b85fe14b318a6970dd79f5/Time-series-analysis-of-climate-variables-using-seasonal-ARIMA-approach.pd