

# MDA511-Assignment-2

Temperature Analysis and Prediction

|  |  |
| --- | --- |
| **Lecturer Name** | Dr.Sheeraz Memon |
| **Student Name & ID** | Ankit MIT233546 |



Table of Contents

[MDA511-Assignment-2 (Group) 0](#_Toc157863536)

[**Acknowledgement** 1](#_Toc157863537)

[**Data Gathering/Sampling Method:** 2](#_Toc157863538)

[**Descriptive Analysis:** 3](#_Toc157863539)

[**Hypothesis Analysis:** 8](#_Toc157863540)

[**Linear Regression and Statistical Prediction:** 8](#_Toc157863541)

[**Additional Analysis** 9](#_Toc157863542)

[**Video presentation link:** 9](#_Toc157863543)

[**References:** 9](#_Toc157863544)

## **Acknowledgement**

First of all, I would like to thank our Unit Lecturer, **Dr.Sheeraz Memon**, who was a constant source of inspiration. He encouraged us to think creatively and motivated us to work on this assignment without giving it a second thought. He expressed full support and provided us with the different teaching aids that were required to complete this assignment..

## **Data Gathering/Sampling Method:**

There are multifarious data sampling approaches which can be taken for gathering like clustered sampling, random sampling, stratified sampling, etc. however, we chose **systematic sampling method** as people are chosen from the sample frame on a regular basis. The intervals are selected to guarantee a sufficient number of samples. Compared to simple random sampling, systematic sampling is frequently more convenient and simpler to implement

To obtain a sample size of n from a population of size x, you need choose a representative sample consisting of every x/nth individual. Take a sample of 100 from a population of 1000, for instance, and choose every 1000/100 = 10th member of the sampling frame. (Shantikumar, 2018)

To gather dataset for sampling, we have to follow some necessary steps as below:

### **Define data collection:**

Like other methods of sampling, we have to decide first about weather data we are going to study. We had two choices either select of time from long list then select variables to collect data or every nth member of our target dataset in study. So, we firstly ensure our list and start to take random like order even in advance physically. (Shantikumar, 2018)

### **Decide on sample size:**

Before selecting our interval value, we go for deciding first about sample size. It is important to select a representative number just to avoid any sampling biasness while gathering. As we go through our dataset, we pointed out the weather numbers is approximately 397 which is our sample size. (Shantikumar, 2018)

### **Calculate sampling interval “n”:**

When we know our target sample size, we calculated our interval, n, by dividing our total estimated weather size by our sample size. However, this is a rough estimation rather than an exact calculation.

We have around 20,166 weather dataset for several months including date and time, and based on this estimation, we calculated an ideal sample size of 397. Our sampling interval nth, hence, equals 20,166/397 = 50.74, which ultimately round to 51th.

### **Select the sample and collect data of weather:**

In last step, we selected every 51th interval member of the weather to include in our sample. It also included 3-4 variables such as temperature, humidity, rain and wind direction.

Now, we constructed a table of our sampling as follow:

| **Date** | **Humidity (percentage)** | **Rain Trace** | **Temperature** | **Wind Direction** |
| --- | --- | --- | --- | --- |
| **0** | 2023-05-25;14:22:15 | 90.0 | 672 | 9.1 | 247 |
| **1** | 2023-05-27;12:15:05 | 90.1 | 682 | 10.5 | 0 |
| **2** | 2023-05-07;17:24:17 | 86.3 | 658 | 3.9 | 180 |
| **3** | 2023-06-14;05:22:00 | 57.8 | 6296 | 15.0 | 90 |
| **4** | 2023-06-15;04:06:22 | 38.3 | 6296 | 20.7 | 90 |

## **Descriptive Analysis:**

After gathering data and creating table, the next prominent task to analyse the weather data with the help of central tendency which includes mean, median and standard deviation the numbers or figures that we selected in table.

To find out mean, we write formula in Jypter notebook include:

mean\_temp **=** df['Temperature']**.**mean()

mean\_hum **=** df['Humidity (percentage)']**.**mean()

mean\_rain **=** df['Rain Trace']**.**mean()

mean\_wind **=** df['Wind Direction']**.**mean()

Where we analyse the figures of:

Mean Temperature : 12.16539898989899

Mean Humidity : 73.9295404040404

Mean Rain : 4716.371212121212

Mean Wind : 108.91919191919192

To find out median, the following are the formulas that we entered,

median\_temp **=** df['Temperature']**.**median()

median\_hum **=** df['Humidity (percentage)']**.**median()

median\_rain **=** df['Rain Trace']**.**median()

median\_wind **=** df['Wind Direction']**.**median()

we analysed from this as an output of:

Median Temperature : 12.0

Median Humidity : 76.85

Median Rain : 6279.0

Median Wind : 90.0

To find out standard deviation, we selected the formulas as,

std\_temp **=** df['Temperature']**.**std()

std\_hum **=** df['Humidity (percentage)']**.**std()

std\_rain **=** df['Rain Trace']**.**std()

std\_wind **=** df['Wind Direction']**.**std()

we analysed and got an output by printing as:

Standard Deviation Temperature : 4.9402812145037585

Standard Deviation Humidity : 17.276359880773757

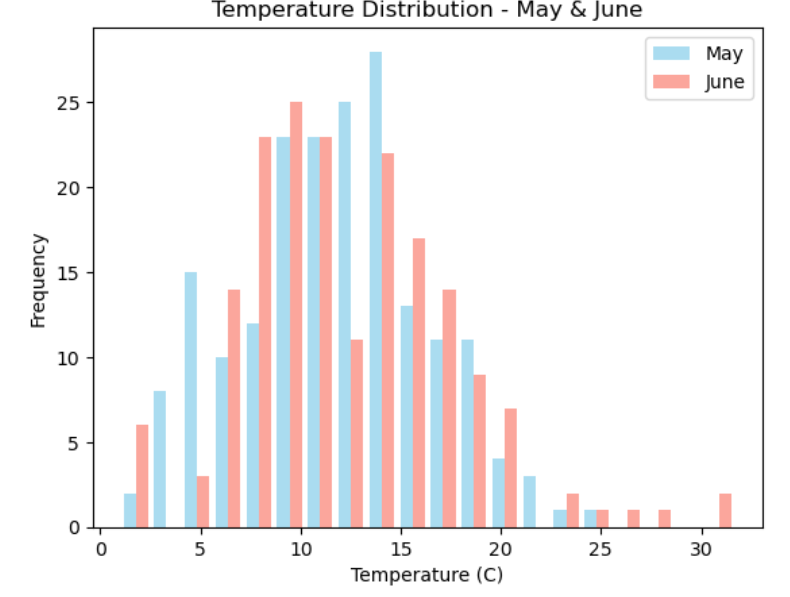
Standard Deviation Rain : 5339.911695669595

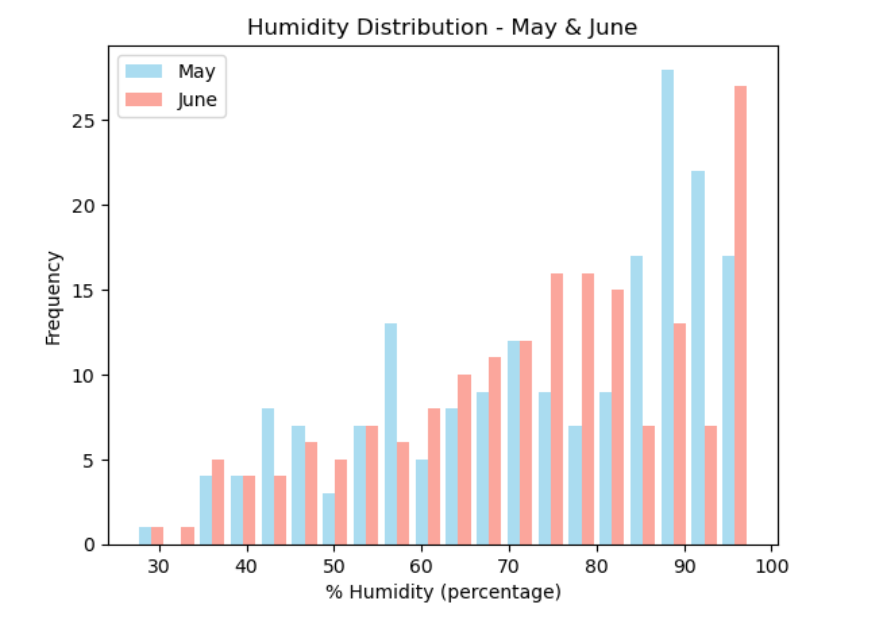
Standard Deviation Wind : 87.73982560859336

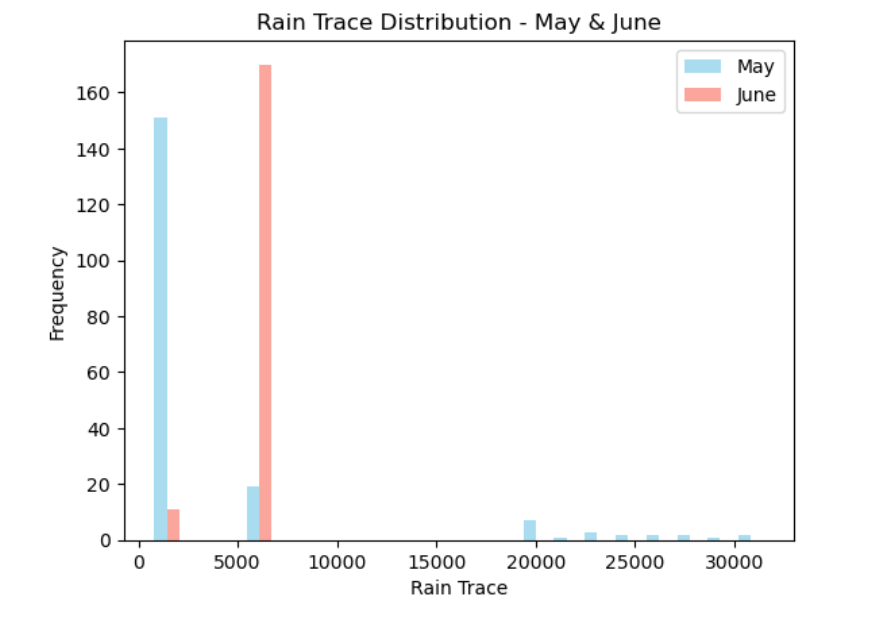
Key analysing from these all figures from mean to standard deviation is that:

* If we choose mean and median for weather dataset, it ensures to be have almost same temperature in both cases which is 12 degree Celsius over 2 months but standard deviation showed relatively low to only 4 degree Celsius for May and June month.
* The analysed figures for rain was almost showing in between from 4000 to 6000 in both months of a year in all cases findings likewise, wind direction also depicted count of mean, median and standard deviation between 87 to 108 which is not much changed than first two selected variables.

We believed the best way is to show our analyses by creating graphical representation as firstly Histograms:





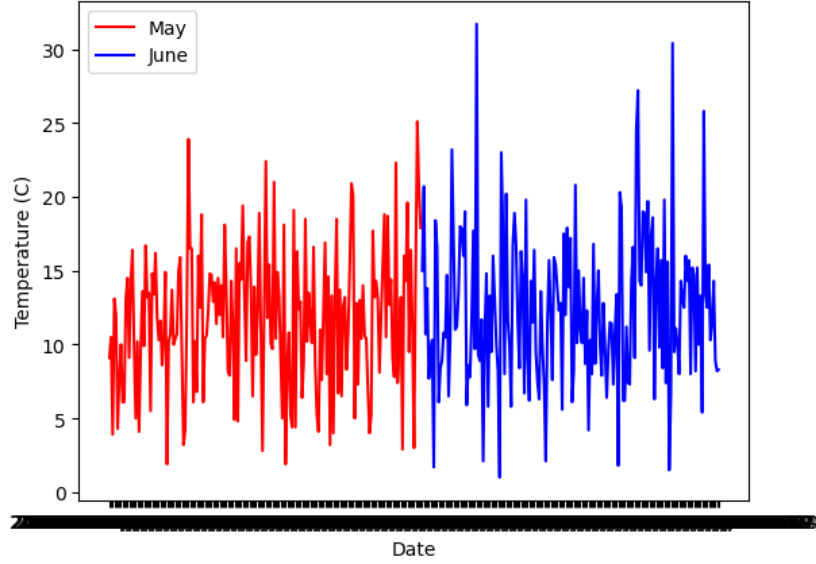


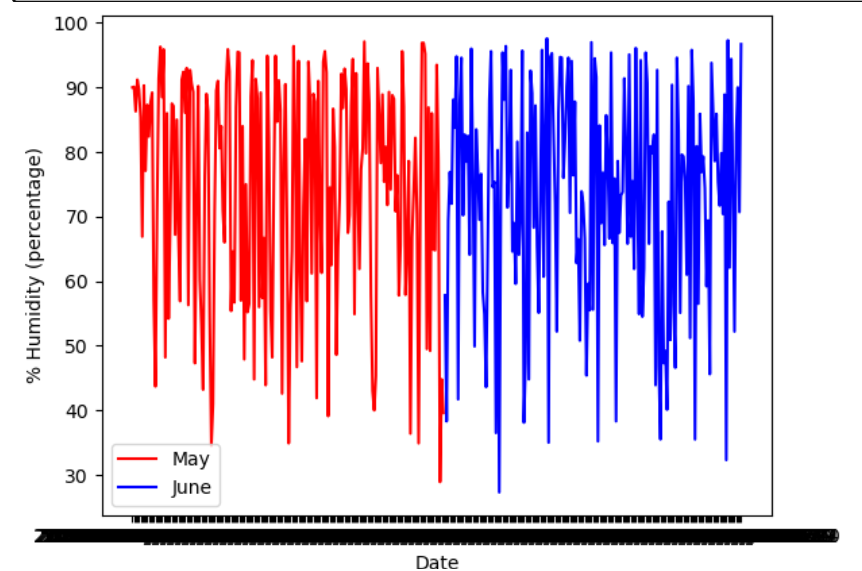
### **We analysed from these histograms:**

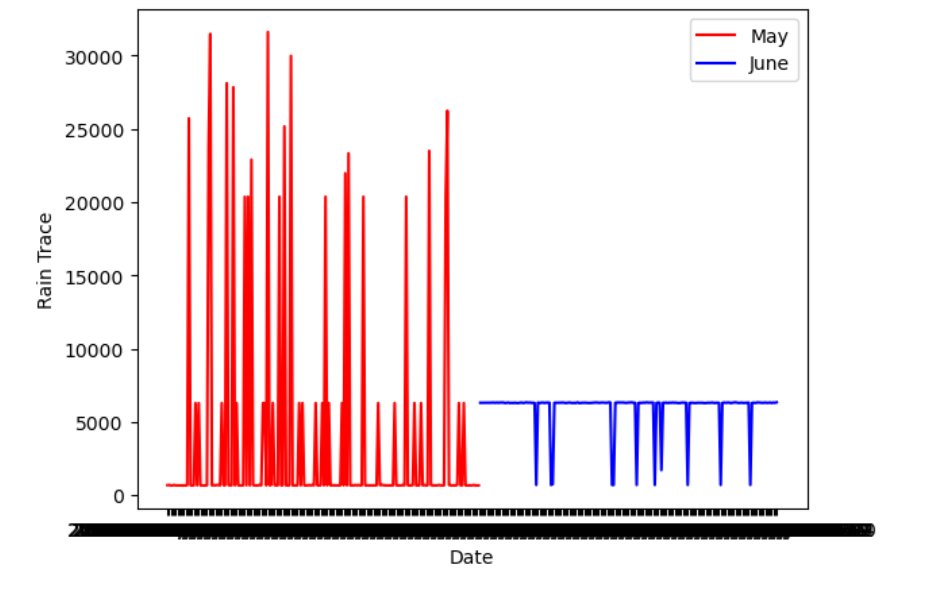
* In month of May and June, the peak temperature has been seen between the counts of 10 to 15 degree Celsius with the highest frequency at approximately 25 in May and June has highest temperature of about 10 degree with frequency little less than 25.
* May has the apex humidity showed as compared to June with the figure of 89% even has more frequently.
* June has more rain trace distribution with the count of 5000 and with maximum frequency of more than 160.



### **Another way is line graphs are as:**



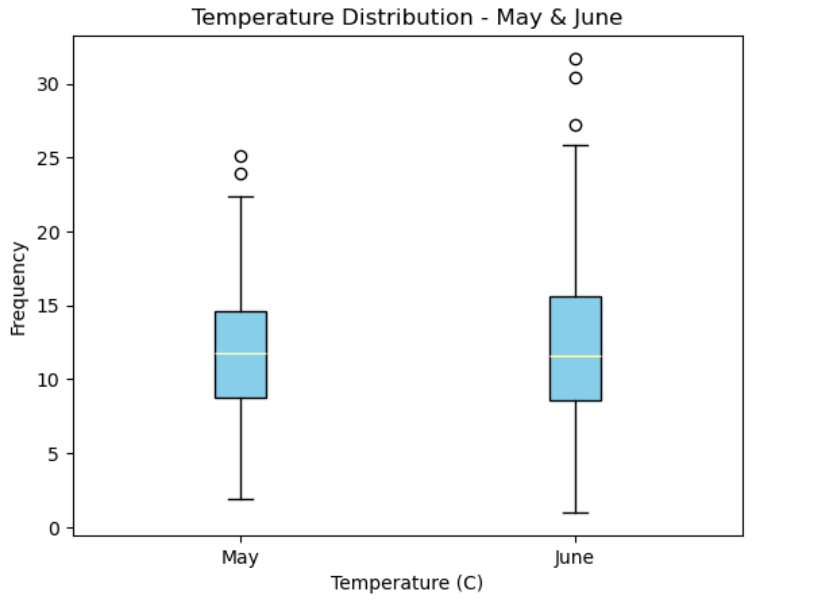


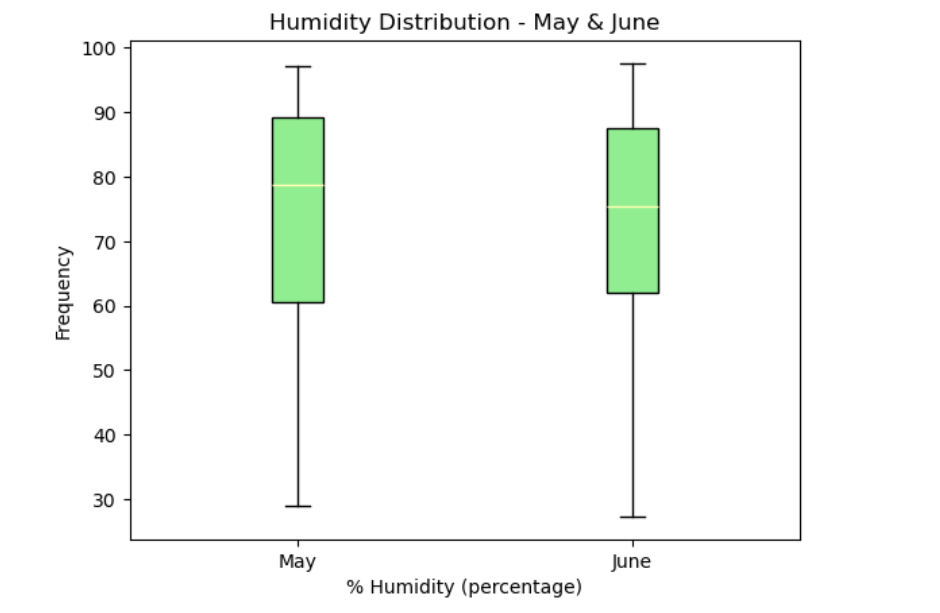


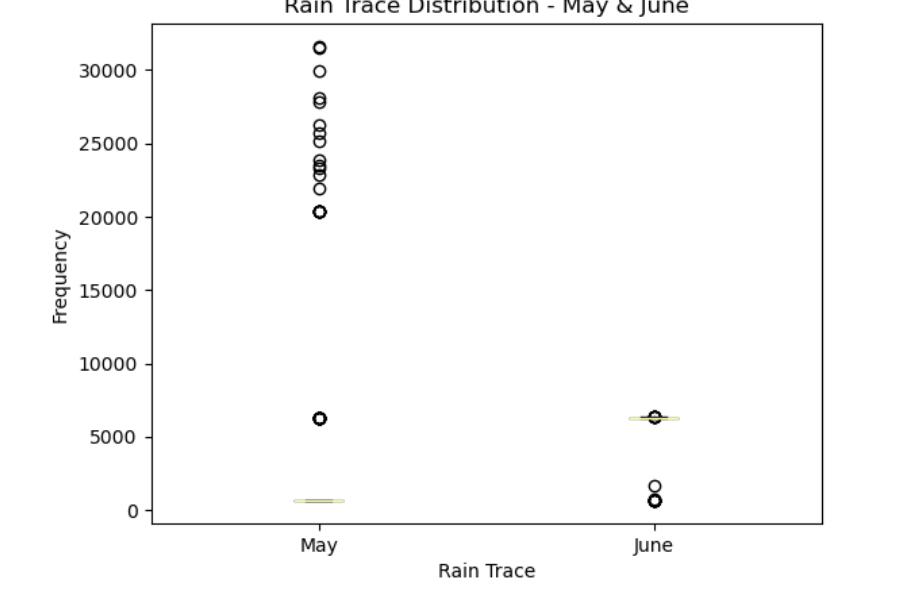
### **We analysed from line graphs:**

* Temperature of June has the most frequency with sharp ups and downs and reaching at 30 however, May has just reached approximately maximum at 25 frequency.
* It is an interesting fact that humidity in both months showed almost same pattern in line graph means level of humidity does not change over 2 months.
* The rain pattern in May depicted more traced at highest with 30000 but June has lesser rain traced even less than 5000.

### **Last and better way is show as box plots:**







### **We analysed from these box plots:**

* Temperature in May and June provided almost same trend in which lower quartile and upper quartile about 50-50% so these outliers are single data points that are >1.5x value of UQ.
* Humidity showed upper and lower quartile almost half of each in the month of June but in May, lower quartile has 75% and upper has 25%.
* Rain trace just only showed outlines which means more density is seen over 20000 in May and in June, very little has observed less 5000.

## **Hypothesis Analysis:**

* "The mean temperature in May equals the mean temperature in June," states the null hypothesis (H0).
* "The mean temperature in June is greater than the mean temperature in May," is the alternative hypothesis (H1).
* Analysis: Using SciPy's stats.ttest\_ind function, the code runs an independent two-sample t-test. This test determines if the means of two independent groups—in this example, the temperatures in May and June—have a statistically significant difference.
* Assumption: The test is predicated on the assumption that the variances of the two samples are not equal, as indicated by the equal\_var=False argument.
* Interpretation: The significance level (α), which is set to 0.05 in your code, is compared to the p-value from the t-test.
* Conclusion: the null hypothesis is rejected if the p-value is less than 0.05. When the p-value is larger than or equivalent to 0.05, the null hypothesis cannot be rejected.
* Print Results: When the null hypothesis is denied, the following code is printed:-

"Reject null hypothesis - temperatures means differ significantly."

"Accept alternate hypothesis that June temperature is higher."

The code **outputs** "Fail to reject null hypothesis - no evidence of significant difference" if the null hypothesis is not rejected.

## **Linear Regression and Statistical Prediction:**

### **Implications**

* Prediction Accuracy: The model's ability to generalize to previously unknown data is shown by the Mean Squared Error (MSE) on the test data. While larger values could point to potential improvement areas, lower MSE values signal superior predicted accuracy.
* Practical Use: the model demonstrate good performance, practical uses might include precise short-term temperature forecasts for outdoor event scheduling, agricultural planning, or energy management. (Bažant & Zebich, 2010)

### **Results:**

* Model Performance: Based on May humidity and rain trace data, the model appears to be quite accurate in predicting June temperatures, as demonstrated by the test MSE.
* June Temperature Prediction: The model appears to catch some of the underlying trends in the data, as the expected mean temperature for June turns out to be very similar to the actual mean temperature. (Bažant & Zebich, 2010)

##### **Limitations:**

* **Straightforward Model:** The linear regression model postulates a linear connection between the objective variable (temperature) and the predictors (humidity and rain trace). The non-linear patterns found in the data might not be captured by this simplicity.
* **Assumption of Independence:** The independence of the observations is assumed in linear regression. This assumption might not be true if there is temporal autocorrelation or dependency between observations (for example, when successive days have similar weather patterns).
* **Limited Functions:** The only predictors used by the model are the rain trace and humidity. The prediction power of the model might be improved by adding more pertinent parameters, such as wind speed, air pressure, or geographic considerations.

## **Additional Analysis**

* **Model Fine-Tuning**: If the linear connection assumption is not able to adequately reflect the underlying patterns, investigate options for improving the model, such as feature engineering, taking interaction terms into account, or attempting more complicated models.
* **The significance of features:** To determine the relative effect of humidity and rain trace on temperature forecasts, do a thorough feature significance analysis. Choosing which traits are most important for precise forecasts may be influenced by this.
* **Methods of Validation:** To get more accurate performance estimates and evaluate how effectively the model generalizes to other data subsets, take into consideration applying cross-validation approaches.
* **Anomalies & Outliers:** Investigate any abnormalities or outliers in the data that can affect the performance of the model. The robustness of the model may be increased by addressing or reducing the influence of outliers.

## **Video presentation link:**

## **References:**

[1] S. Shantikumar, “Methods of sampling from a population,” Methods of sampling from a population | Health Knowledge, https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population (accessed Feb. 3, 2024).

[2] Z. P. Bažant and S. Zebich, “Statistical linear regression analysis of prediction models for creep and shrinkage,” *Cement and Concrete Research*, vol. 13, no. 6, pp. 869–876, Nov. 2010. doi:10.1016/0008-8846(83)90088-1

[3] L. Thomas, “Systematic sampling: A step-by-step guide with examples,” Scribbr, https://www.scribbr.com/methodology/systematic-sampling/ (accessed Feb. 3, 2024).

[4] R. Bevans, “Hypothesis testing: A step-by-step guide with easy examples,” Scribbr, https://www.scribbr.com/statistics/hypothesis-testing/ (accessed Feb. 3, 2024).