VISUALIZING INSIGHTS: EXPLORING TRENDS IN THE DISCOMFORT INDEX

A Project Report submitted to the Department of Computer Applications, SRMIST, Kattankulathur for partial fulfilment of the requirement for the award of degree of

MASTER OF SCIENCE
IN
APPLIED DATA SCIENCE

Submitted by

JOHN VARSHAN J

Registration number: RA2332014010040

M.Sc. Applied Data Science (2nd year)
Department of Computer Applications
SRM INSTITUTE OF TECHNOLOGY

Supervisors

Bibraj.R,
Scientist-D,
IT Cell/X-Doppler Weather Radar,
Regional Meteorological Centre, Chennai

Dr.Mothilal Nehru.G,
Assistant Professor,
Department of Computer Applications,
SRMIST , Kattankulathur

Regional Meteorological Centre,
Indian Meteorological Department,
College Rd, Nungambakkam, Chennai,
Tamil Nadu – 600006

Department of Computer Applications,
Faculty of science and humanities,
SRMIST , Kattankulathur,
Tamil Nadu - *603203*

ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to our mentor, **Dr. Bibraj**, for their invaluable guidance and unwavering support throughout our internship at the **Regional Meteorological Centre, Nungambakkam**. Their insights and encouragement have been instrumental in our professional and personal development.

I am also deeply appreciative of the **Regional Meteorological Centre, Nungambakkam**, for providing us with such a remarkable opportunity. This experience has significantly broadened our practical knowledge and industry skills.

I am grateful to everyone who has played a part in this journey, and I look forward to applying these learnings in my future endeavors.

Thank you all for your support and encouragement.

John Varshan

INDEX

INTRODUCTION

This report documents my internship experience at **Regional Meteorological Centre, Nungambakkam**, where I worked as a Data Visualization Intern from **19-05-2024** to **19-06-2024.** As an essential component of my academic program at **SRM IST, Kattanakulathur**, this internship provided a valuable opportunity to bridge the gap between theoretical knowledge and practical application in the field of **Data Visualization.** designed to enhance my practical skills and understanding of data visualization, a key area in the modern data-driven decision-making process.

Data visualization is an essential tool in the current landscape of big data, enabling organizations to interpret complex data sets and derive meaningful insights. My role involved transforming raw data into clear, actionable visual representations that facilitate better understanding and communication across various departments.

During my tenure, I was tasked with a variety of responsibilities that not only enhanced my technical skills but also honed my professional capabilities, such as communication, teamwork, and problem-solving. The primary aim of this internship was to gain hands-on experience with **Visualization of time series data,** develop proficiency in industry-standard tools and techniques, and contribute meaningfully to the organization's objectives

This report encompasses an overview of the organization, a detailed account of my role and responsibilities, the projects I undertook, and the skills I developed. It also reflects on the challenges encountered and the strategies used to overcome them. This internship not only allowed me to apply my academic knowledge in a real-world setting but also provided a platform to build a professional network and gain insights into the industry's dynamics. I am immensely grateful to **Dr. BIBRAJ** for their guidance and support, and to the **RMC** for their cooperation, which made this experience profoundly enriching and impactful.

PROFILE

Regional Meteorological Centre, Chennai

About RMC

The Regional Meteorological Centre (RMC) in Chennai is one of the key meteorological offices under the India Meteorological Department (IMD), which is responsible for weather forecasting, monitoring, and climate research in the southern region of India. Located in Chennai, Tamil Nadu, the RMC plays a crucial role in providing accurate and timely weather information, which is essential for various sectors including agriculture, aviation, shipping, and disaster management.

Functions and Responsibilities

Weather Forecasting: The RMC Chennai provides daily weather forecasts, including short, medium, and long-term predictions. These forecasts are vital for preparing for adverse weather conditions and planning activities accordingly.

Climate Monitoring: The centre monitors and records climatic conditions, analyzing long-term weather patterns and contributing to climate research. This data is crucial for understanding climate change impacts and variability.

Severe Weather Warnings: The RMC is responsible for issuing warnings about severe weather events such as cyclones, heavy rainfall, thunderstorms, and heatwaves. These warnings are disseminated to the public and relevant authorities to take preventive measures.

Data Collection and Analysis: The centre collects meteorological data from various observation stations across the region. This data includes temperature, humidity, rainfall, wind speed, and pressure, which are analysed to produce accurate weather reports and forecasts.

Research and Development: RMC Chennai engages in meteorological research to improve forecasting techniques and understand weather phenomena better. This research contributes to the overall advancement of meteorological science.

Public Awareness: The centre also focuses on educating the public about weather-related issues and the importance of preparedness. It conducts outreach programs and provides information through various media channels.

Infrastructure and Technology:

The RMC Chennai is equipped with advanced meteorological instruments and technology, including weather radars, satellite imagery, automatic weather stations, and computer models for weather prediction. These tools enable the centre to monitor weather conditions in real-time and produce accurate forecasts.

Importance to the Region:

RMC Chennai's services are critical for the southern region of India, which frequently experiences extreme weather events like cyclones and monsoons. Accurate weather forecasts and timely warnings help mitigate the impact of such events on lives and property, making the centre an essential component of the region's disaster management and planning infrastructure.

Overall, the Regional Meteorological Centre in Chennai plays a pivotal role in safeguarding the southern region of India through its comprehensive weather monitoring, forecasting, and research activities.

OBJECTIVE

Communicate Insights

The primary objective of data visualization is to effectively communicate complex data and insights in a visual format that is easy to understand and interpret. It involves selecting the most relevant data points, choosing the appropriate visual representation (such as charts, graphs, or maps), and designing the visualization in a way that highlights key findings.

Identify Trends and Patterns

Data visualization helps in identifying trends, patterns, and relationships within datasets that may not be apparent from raw data alone. Identifying trends and patterns through data visualization is about uncovering meaningful insights hidden within the data. By visualizing data over time or across different variables, trends and patterns that might not be apparent in raw data can be revealed. This process often involves using techniques such as line charts for trend analysis, scatter plots for identifying correlations, and heatmaps for spotting patterns in large datasets.

Develop Proficiency in Data Visualization Tools

Gain hands-on experience with industry-standard tools such as Tableau, Power BI, or Python libraries like Matplotlib and Seaborn. It also involves learning best practices for data visualization, such as choosing the right type of chart or graph for the data, effect

Create Effective Visualizations

Learn to design and implement visualizations that effectively communicate complex data insights to both technical and non-technical stakeholders. Key principles for creating effective visualizations include choosing the right type of chart or graph for the data, using appropriate colours and visual elements to highlight important information, and ensuring that the visualization is well-organized and easy to navigate.

Handle and analyse large datasets

Acquire skills in managing, cleaning, and analyzing large datasets to ensure accuracy and reliability of visualized information. By effectively handling and analyzing large datasets, analysts can uncover trends, patterns, and correlations that can inform decision-making and drive business strategies.

Understand User Experience (UX) Principles

Apply UX principles to create visualizations that are intuitive, user-friendly, and enhance the overall user experience. This involves considering factors such as the target audience, their goals and needs, and the context in which they will interact with the visualization. Key UX principles include simplicity, consistency, clarity, and interactivity

Explore Innovative Visualization Technique

Experiment with new and innovative visualization techniques to present data in engaging and insightful ways. By exploring innovative visualization techniques, analysts can uncover new insights and patterns in the data that may not be apparent with conventional visualization methods. This can lead to a deeper understanding of the data and help stakeholders make more informed decisions based on these insights.

Enhance Data Quality

This process includes cleaning and validating the data to remove errors, duplicates, and inconsistencies. It also involves verifying the source of the data and ensuring that it is up to date and relevant for the analysis. By enhancing data quality, analysts can improve the accuracy and reliability of their visualizations, leading to more meaningful insights and informed decision-making.

Gain Industry Experience

Gain insights into how data visualization is used in real-world scenarios within your specific industry or field of interest. This experience helps us to understand the unique challenges and requirements of the industry, allowing them to create visualizations that are tailored to the industry's needs.

DATA PREPARATION

Dataset Overview

The dataset contains 94,704 entries and 7 columns, with the following structure:

YEAR: The year the data was recorded

MONTH: The month the data was recorded

DATE: The date of the month the data was recorded

HR: The hour of the day the data was recorded

T: Temperature in degrees Celsius

RH: Relative humidity in percentage

DI: Discomfort index

VARIABLE EXPLAINATION

TEMPERATURE

Temperature is a crucial parameter used to describe the state of the atmosphere and predict weather patterns. Temperature is a fundamental aspect of meteorology, influencing weather patterns, forecasting, and climate studies. Accurate temperature measurement and analysis are essential for understanding and predicting atmospheric phenomena.

In the received data we got temperature data in Celsius scale.

RELATIVE HUMIDITY

Relative humidity (RH) is a key concept in meteorology, representing the amount of moisture in the air relative to the maximum amount of moisture the air can hold at a given temperature. The ratio of the current absolute humidity to the highest possible absolute humidity (which depends on the current air temperature), expressed as a percentage. It indicates how close the air is to being saturated with water vapour.

DISCOMFORT INDEX

The Discomfort Index (DI), also known as the Temperature-Humidity Index (THI) or Humidex, is a metric used in meteorology to assess the combined effects of temperature and humidity on human comfort. It provides an indication of how hot or uncomfortable it feels to the average person, taking into account both the air temperature and the moisture content in the air.

DI Values and Comfort Levels:

Below 21°C: Comfortable, no discomfort.

21°C - 24°C: Slight discomfort.

24°C - 27°C: Noticeable discomfort.

27°C - 29°C: Increasing discomfort.

29°C - 32°C: High discomfort, avoid physical activity.

Above 32°C: Very high discomfort, potential for heat-related health issues.

Brief Explanation of Data Pre-processing

Data pre-processing is a crucial step in the data analysis pipeline. It involves transforming raw data into a clean and usable format, which improves the quality of the data and ensures accurate results during analysis and modeling. Key steps in data pre-processing include:

Data Cleaning:

Handling Missing Values: Identifying and managing missing data, which may involve filling, interpolation, or removal of rows/columns with missing values.

Removing Duplicates: Ensuring each data entry is unique to prevent bias and redundancy.

Data Transformation:

Normalization/Standardization: Adjusting the scale of the data for better performance in algorithms that are sensitive to data magnitude.

Encoding Categorical Variables: Converting categorical data into numerical formats using techniques like one-hot encoding or label encoding.

Implementation in python

The data csv file has been imported in python

Initially there were 94704 rows and 7 columns

```
In [1]: import pandas as pd

In [2]: df = pd.read_csv('DI_FINAL.csv')

In [3]: df
Out[3]:
       YEAR  MONTH  DATE     HR     T    RH        DI
0      2006      1     1   1:00  22.1  85.0  21.47300
1      2006      1     1   2:00  21.6  88.0  21.13140
2      2006      1     1   3:00  21.2  91.0  20.86835
3      2006      1     1   4:00  20.8  93.0  20.55745
4      2006      1     1   5:00  20.8  93.0  20.55745
...     ...    ...   ...    ...   ...   ...       ...
94699  2023     12    31  20:00  25.9  70.0  24.01900
94700  2023     12    31  21:00  26.1  68.0  24.05840
94701  2023     12    31  22:00  25.6  71.0  23.82955
94702  2023     12    31  23:00  24.6  74.0  23.15570
94703  2023     12    31   0:00  24.1  77.0  22.88560

[94704 rows x 7 columns]
```

```
In [4]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 94704 entries, 0 to 94703
Data columns (total 7 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   YEAR    94704 non-null  int64
 1   MONTH   94704 non-null  int64
 2   DATE    94704 non-null  int64
 3   HR      94704 non-null  object
 4   T       94318 non-null  float64
 5   RH      94174 non-null  float64
 6   DI      93852 non-null  float64
dtypes: float64(3), int64(3), object(1)
memory usage: 5.1+ MB
```

DATA PREPROCESSING

DATA CLEANING

There were 1768 missing data present in the data which was removed by dropna() function

```
In [5]: missing_values_count = df.isnull().sum()
```

```
In [7]:
   ...: total_missing_values = missing_values_count.sum()
   ...:
   ...: print("Number of missing values in each column:")
   ...: print(missing_values_count)
   ...:
   ...: print("\nTotal number of missing values in the DataFrame:", total_missing_values)
Number of missing values in each column:
YEAR       0
MONTH      0
DATE       0
HR         0
T        386
RH       530
DI       852
dtype: int64

Total number of missing values in the DataFrame: 1768
```

```
In [8]: df1 = df.dropna()

In [9]: missing_values_count = df1.isnull().sum()
   ...:
   ...: # Total number of missing values in the DataFrame
   ...: total_missing_values = missing_values_count.sum()
   ...:
   ...: print("Number of missing values in each column:")
   ...: print(missing_values_count)
   ...:
   ...: print("\nTotal number of missing values in the DataFrame:", total_missing_values)
Number of missing values in each column:
YEAR     0
MONTH    0
DATE     0
HR       0
T        0
RH       0
DI       0
dtype: int64

Total number of missing values in the DataFrame: 0
```

DATA TRANSFORMATION

The variables DATE, YEAR, MONTH, HR was transformed into datetime for better indexing and there is no data recorded from the year 2009 to 2012 and 2019. We were instructed to do decadal trend analysis from the year 2013 to 2023. So we filtered the data from 2013 to 2023.

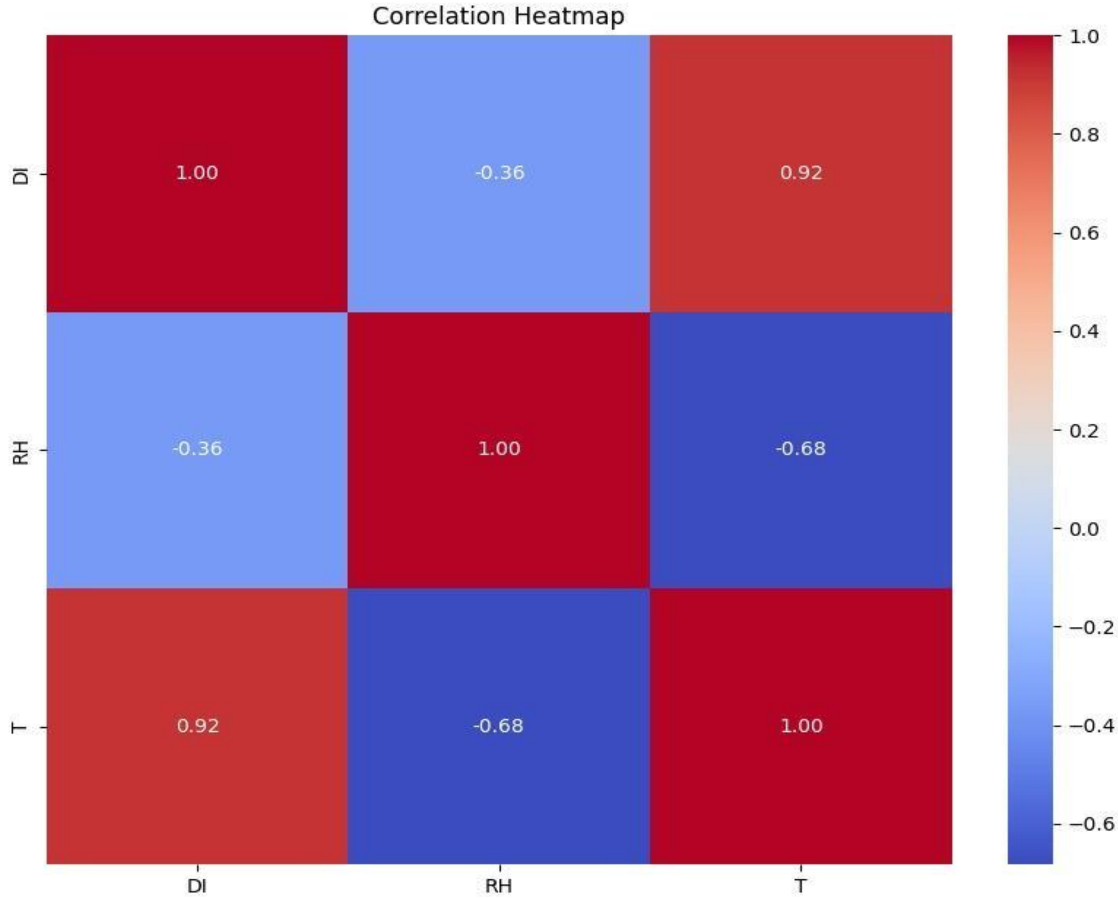
```
In [13]:
    ...: df1 = df1.rename(columns={'DATE': 'DAY'})
    ...: df1['HOUR'] = df1['HR'].str[:-3]
    ...: df1['Datetime'] = pd.to_datetime(df1[['YEAR', 'MONTH', 'DAY', 'HOUR']])
    ...: df1.set_index('Datetime', inplace=True)
    ...:
    ...: df1.drop(columns=['YEAR', 'MONTH', 'DAY', 'HR',"HOUR"], inplace=True)
    ...: print(df1.head())
                        T     RH        DI
Datetime
2006-01-01 01:00:00  22.1  85.0  21.47300
2006-01-01 02:00:00  21.6  88.0  21.13140
2006-01-01 03:00:00  21.2  91.0  20.86835
2006-01-01 04:00:00  20.8  93.0  20.55745
2006-01-01 05:00:00  20.8  93.0  20.55745
```

```
In [19]:
    ...: df_filtered = df1[~((df1.index.year >= 2006) & (df1.index.year <= 2008))]
    ...:
    ...: print("\nDataFrame after removing data from 2006 to 2008:")
    ...: print(df_filtered)

DataFrame after removing data from 2006 to 2008:
                        T     RH        DI
Datetime
2013-02-01 01:00:00  24.7  72.0  23.12920
2013-02-01 02:00:00  24.4  74.0  22.98430
2013-02-01 03:00:00  23.6  80.0  22.59900
2013-02-01 04:00:00  22.9  85.0  22.20700
2013-02-01 05:00:00  22.8  85.0  22.11525
...                   ...   ...       ...
2023-12-31 20:00:00  25.9  70.0  24.01900
2023-12-31 21:00:00  26.1  68.0  24.05840
2023-12-31 22:00:00  25.6  71.0  23.82955
2023-12-31 23:00:00  24.6  74.0  23.15570
2023-12-31 00:00:00  24.1  77.0  22.88560

[72022 rows x 3 columns]
```

Week 1

CORRELATION HEATMAP

After processing the data, we tried to find the relation between the variables. For that we have used the seaborn library in python to plot the correlation heatmap.



Observations from the heatmap

- Humidity and discomfort index shows negative correlation
- Temperature and discomfort index shows positive correlation
- Temperature and humidity negative correlation

CALENDAR HEATMAPS

Heatmap of Discomfort index

We have plotted average of Discomfort Index in month against year using the tableau desktop.

HEAT MAP of Discomfort Index (Month vs Year)

	Year calc field									
month calc ..	2013	2014	2015	2016	2017	2018	2020	2021	2022	2023
1			24.195	23.937	24.290	23.670	24.734	24.848	24.841	23.842
2	25.270		24.706	25.782	24.163	24.632	25.737	24.214	24.912	24.236
3	26.444	25.920	26.566	27.279	26.659	25.985	27.539	26.352	26.853	26.306
4	28.926	28.028		28.824	28.422	28.210	28.964	28.472	29.322	28.083
5	29.977	28.723		29.588	28.546	28.816	30.085	28.935	29.648	28.730
6	28.305		28.585	28.420	27.894	27.919	28.500	28.518	29.100	28.444
7	28.060	28.462		28.465	27.535	27.531	28.161	27.879	27.498	27.603
8	27.934	27.954		27.827	27.426	27.239	28.179	28.055	27.626	27.588
9	27.678	27.476	28.248	27.389	27.747	27.732	27.656	28.244	27.461	27.244
10	27.372	26.947		27.964		26.757	26.944	27.583	26.326	26.983
11		25.384	25.528	25.313	25.786	25.669	25.590	25.944	24.721	
12		24.774	24.979	24.196	24.906	24.558	24.380	24.248	24.833	24.789

Avg. DI

23.670 30.085

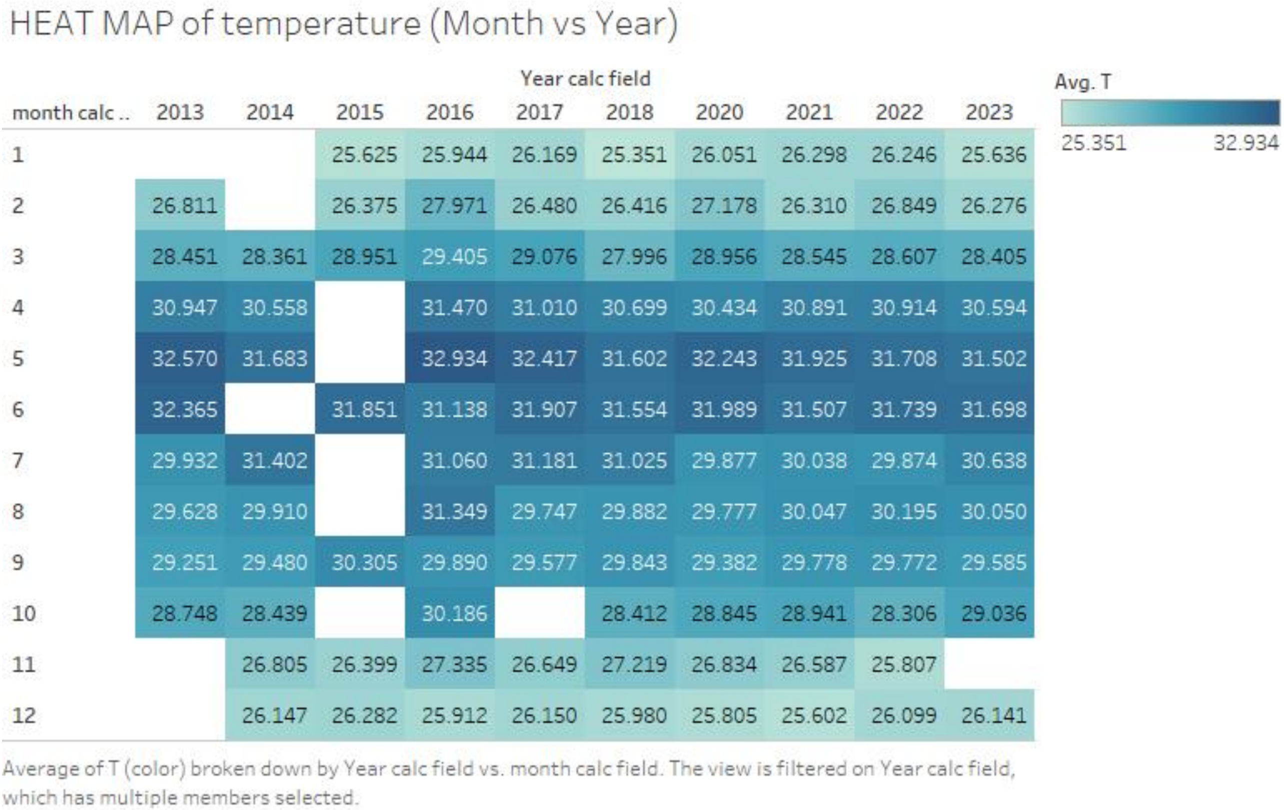
Average of DI (color) broken down by Year calc field vs. month calc field. The view is filtered on Year calc field, which has multiple members selected.

Observations from the heatmap

- January and February: These months generally show lower discomfort levels, with a few exceptions such as February 2023.
- May to August: These months tend to have higher discomfort levels,
- The discomfort index is relatively lower in December compared to summer months but shows some variability over the years.

HEATMAP OF TEMPERATURE

We have plotted average of temperature in month against year.



HEAT MAP of temperature (Month vs Year)

	Year calc field									
month calc ..	2013	2014	2015	2016	2017	2018	2020	2021	2022	2023
1			25.625	25.944	26.169	25.351	26.051	26.298	26.246	25.636
2	26.811		26.375	27.971	26.480	26.416	27.178	26.310	26.849	26.276
3	28.451	28.361	28.951	29.405	29.076	27.996	28.956	28.545	28.607	28.405
4	30.947	30.558		31.470	31.010	30.699	30.434	30.891	30.914	30.594
5	32.570	31.683		32.934	32.417	31.602	32.243	31.925	31.708	31.502
6	32.365		31.851	31.138	31.907	31.554	31.989	31.507	31.739	31.698
7	29.932	31.402		31.060	31.181	31.025	29.877	30.038	29.874	30.638
8	29.628	29.910		31.349	29.747	29.882	29.777	30.047	30.195	30.050
9	29.251	29.480	30.305	29.890	29.577	29.843	29.382	29.778	29.772	29.585
10	28.748	28.439		30.186		28.412	28.845	28.941	28.306	29.036
11		26.805	26.399	27.335	26.649	27.219	26.834	26.587	25.807	
12		26.147	26.282	25.912	26.150	25.980	25.805	25.602	26.099	26.141

Average of T (color) broken down by Year calc field vs. month calc field. The view is filtered on Year calc field, which has multiple members selected.

Observations from the heatmap

- The heat map illustrates clear seasonal temperature patterns, with peak temperatures occurring between April and August and cooler temperatures from September to March.
- There is a general trend of rising temperatures over the years, particularly evident in the hottest months.
- Significant year-to-year variability highlights the dynamic nature of climate patterns, influenced by various factors including global climatic changes.

HEATMAP OF HUMIDITY

We have plotted average of humidity in month against year.

HEAT MAP of Relative humidity (Month vs Year)

	Year calc field									
month calc ..	2013	2014	2015	2016	2017	2018	2020	2021	2022	2023
1			77.93	69.89	71.70	74.12	80.77	79.27	79.99	72.60
2	78.76		76.32	71.74	66.49	74.95	81.40	69.72	73.44	71.90
3	75.72	69.81	71.14	74.60	70.75	74.97	83.20	73.80	79.46	74.09
4	78.72	72.55		72.26	72.34	73.06	84.30	74.53	83.56	72.85
5	75.33	70.09		68.08	61.90	71.23	79.13	69.97	81.07	71.92
6	60.11		67.66	71.53	59.60	62.70	65.28	69.67	75.10	68.49
7	79.78	69.89		72.20	61.69	63.10	81.57	76.70	73.67	67.10
8	81.50	78.60		63.57	73.60	70.86	82.25	78.42	71.94	72.97
9	82.17	77.08	77.27	71.79	79.23	76.43	80.60	83.30	74.24	73.58
10	84.00	82.28		75.23		79.98	78.15	84.26	75.63	75.61
11		80.23	87.55	72.49	88.62	79.17	83.34	91.20	83.68	
12		79.55	81.35	73.98	82.14	78.46	78.39	80.31	81.22	79.81

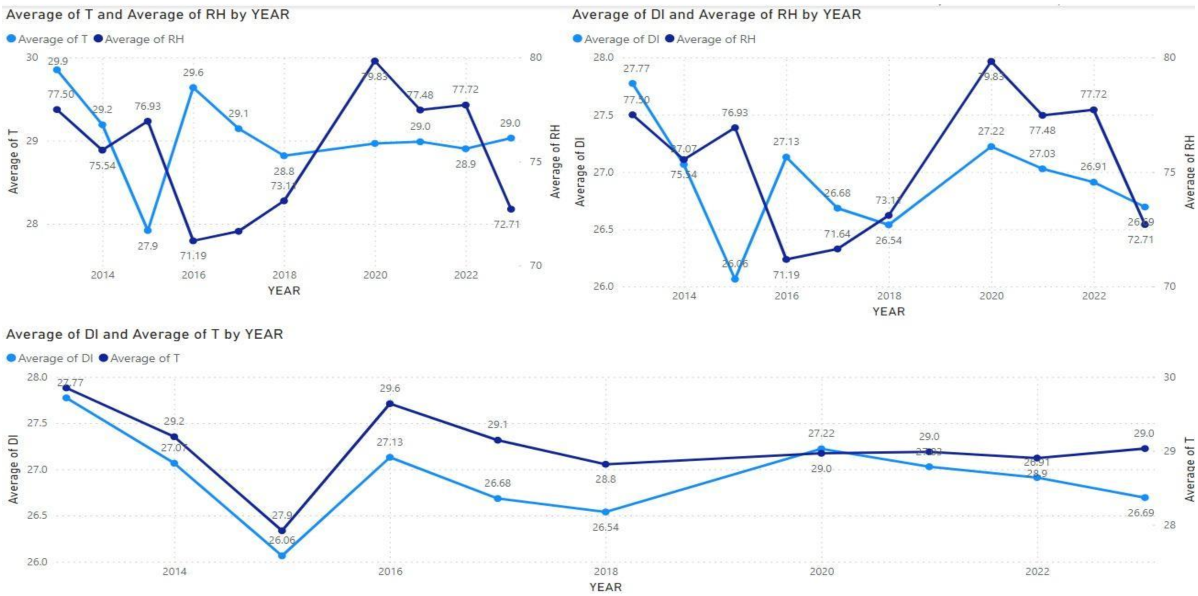
Avg. RH
59.60 91.20

Average of RH (color) broken down by Year calc field vs. month calc field. The view is filtered on Year calc field, which has multiple members selected.

Observations from the heatmap

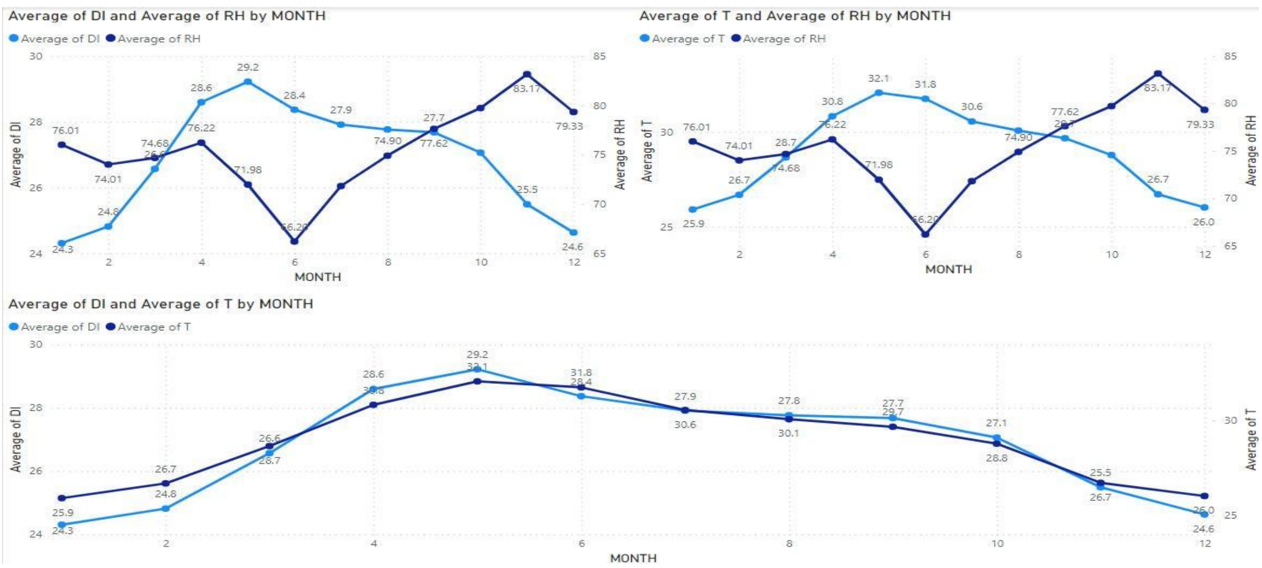
- The colour gradient from dark to light blue indicates changes in average RH, where darker shades represent higher humidity levels and lighter shades indicate lower levels.
- August 2022: This month shows the highest RH value (91.20), indicating an unusually high humidity level.
- May 2013 and 2016: These months show the lowest RH values (around 60), indicating a significant drop in humidity.
- January and December: These months generally show higher humidity levels, with some exceptions like January 2016 and December 2017.
- May and July: Variability is more pronounced in these months. For example, May month shows lower humidity in 2013 and 2016.

Line chart by year-wise average



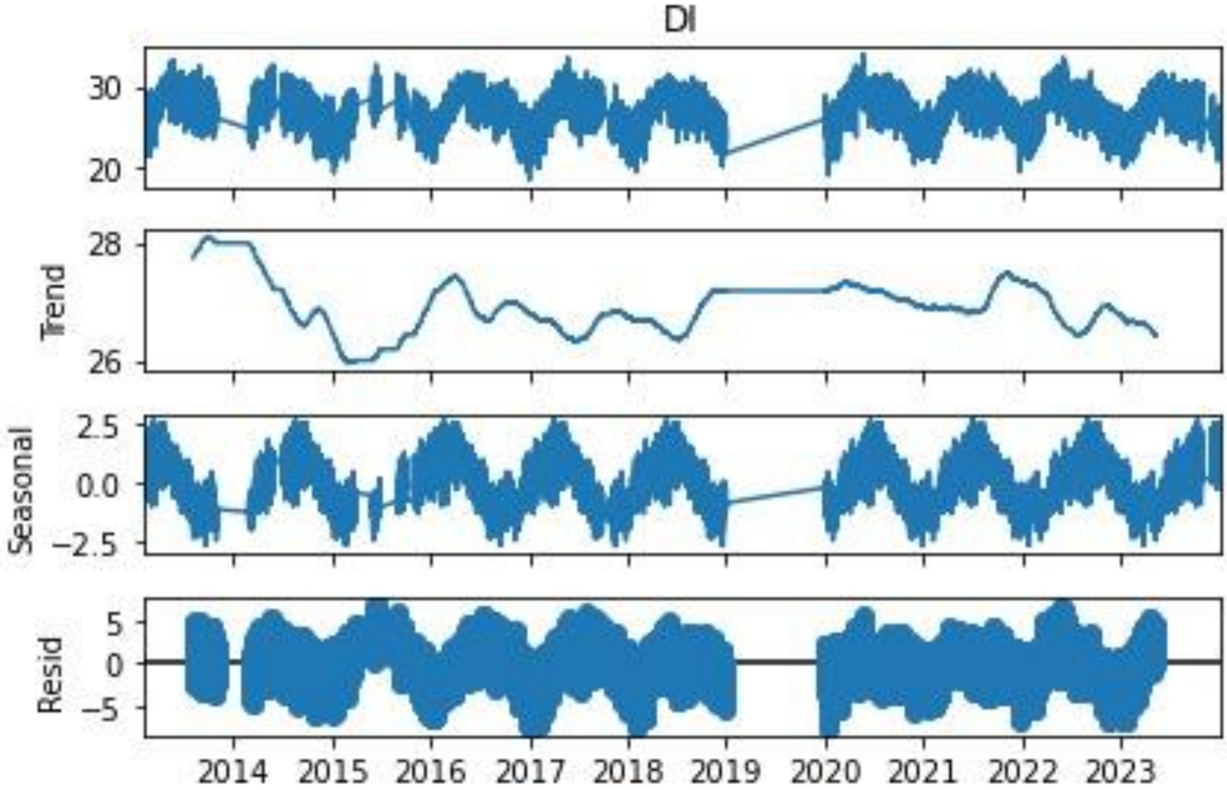
- We can see that when temperature increases DI also increases, When Temperature Decreases DI is also decreases
- 2016 stands out as a year with the highest average temperature (29.6°C) and DI (29.6), indicating the most uncomfortable year.

Line chart by month-wise average

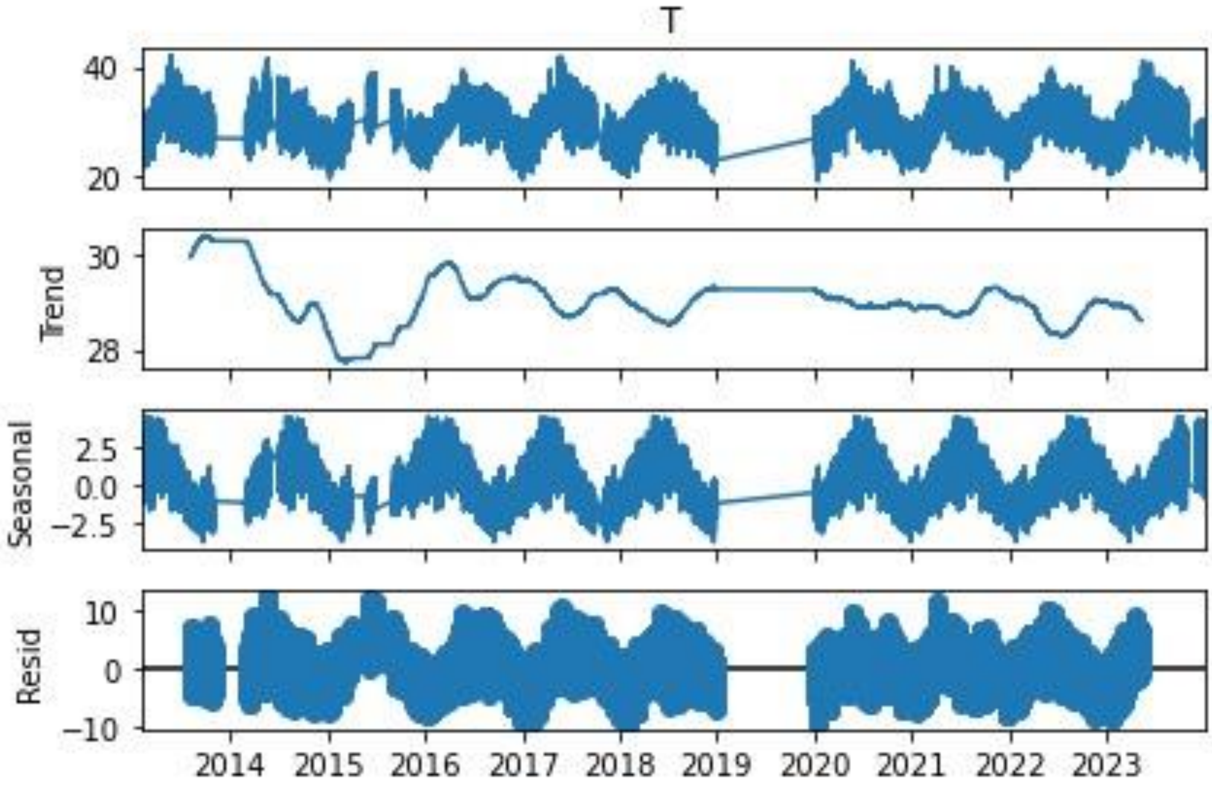


- There is a strong correlation between temperature and the Discomfort Index. Higher temperatures typically result in higher DI values.
- Most comfortable months: January and December, with the lowest DI values and moderate temperatures.
- Least comfortable months: May through August, with the highest DI values due to peak temperatures and varying humidity levels.

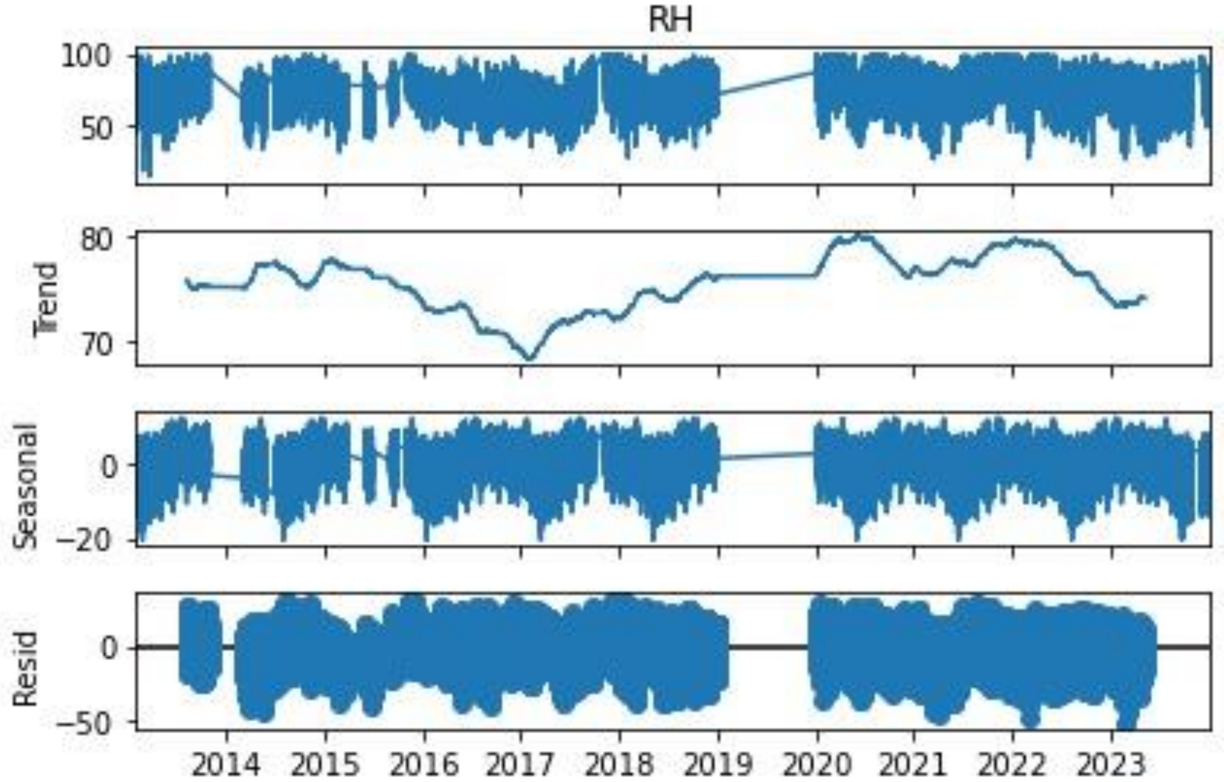
Seasonal decomposition of Discomfort Index over years



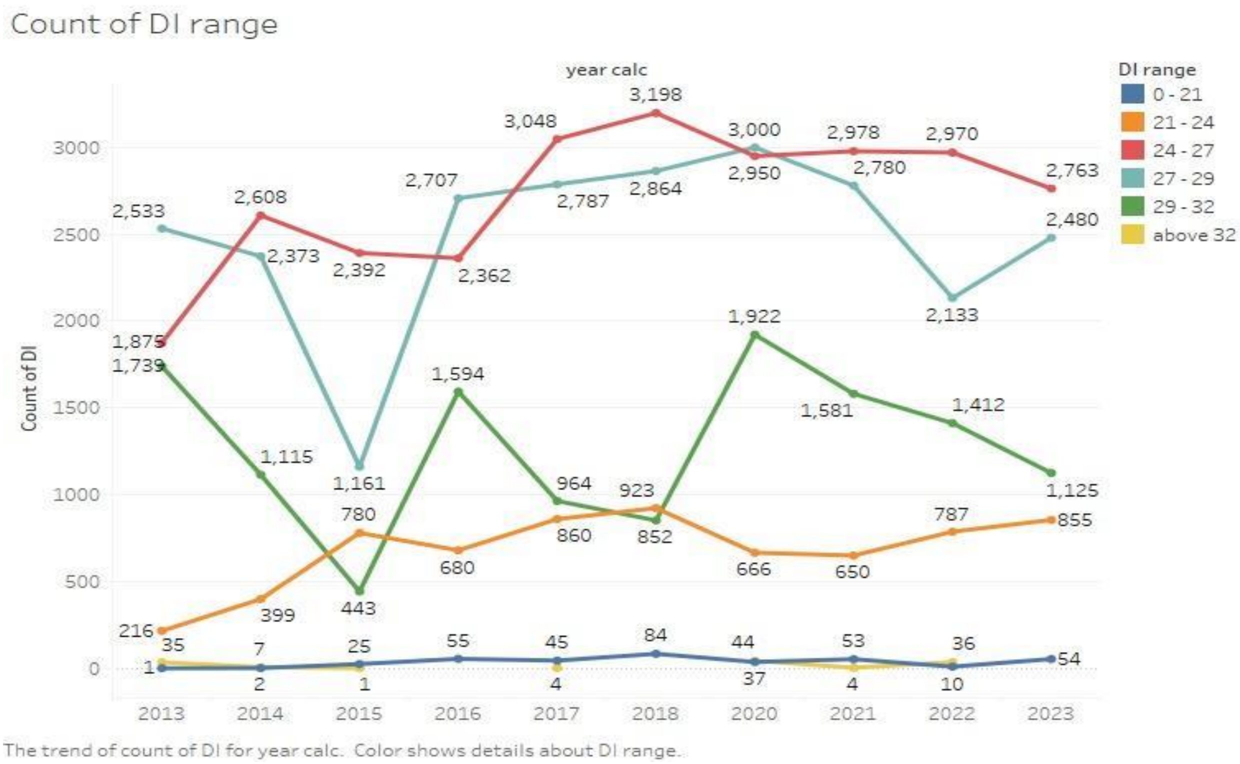
Seasonal decomposition of Temperature over years



Seasonal decomposition of Humidity over years



Trend count of Discomfort Index observed in all ranges



SARIMAX

We built a sarimax model in python using autoarima function. In that we found that the time series data is non stationary by using the Dickey Fuller test, after that we did convert the data to stationary. We trained the model with 80% train data.

We found that the best model has

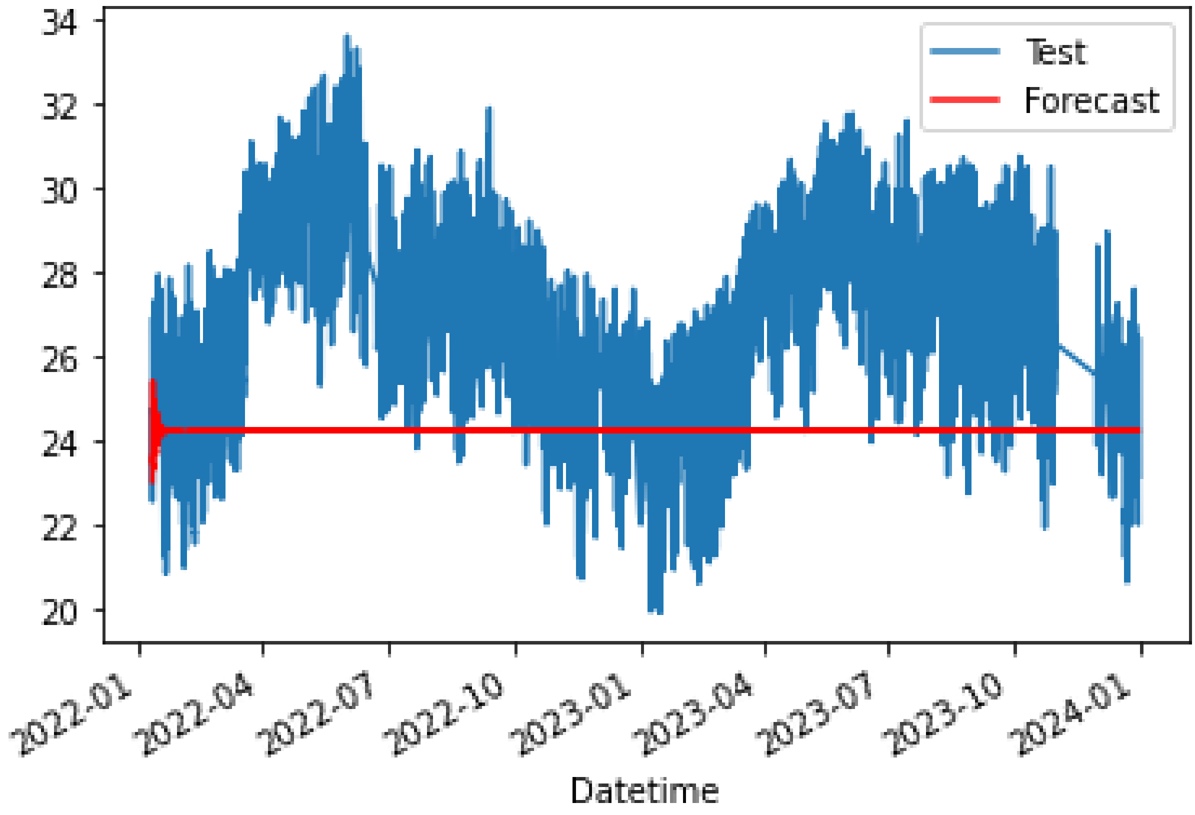
- 4 auto regressive parts
- 1 integration part
- 4 moving average parts

MODEL SUMMARY

```
Best model:  ARIMA(4,1,4)(0,0,0)[0]
Total fit time: 1652.622 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                57617
Model:               SARIMAX(4, 1, 4)   Log Likelihood              -35495.964
Date:                Fri, 24 May 2024   AIC                          71009.928
Time:                        13:00:21   BIC                          71090.582
Sample:                             0   HQIC                         71035.027
                              - 57617
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.4370      0.070     20.546      0.000       1.300       1.574
ar.L2         -0.2709      0.153     -1.769      0.077      -0.571       0.029


ar.L3         -0.1032      0.133     -0.775      0.439      -0.364       0.158
ar.L4         -0.1694      0.052     -3.282      0.001      -0.271      -0.068
ma.L1         -1.3279      0.070    -19.013      0.000      -1.465      -1.191
ma.L2          0.1361      0.146      0.933      0.351      -0.150       0.422
ma.L3          0.0315      0.120      0.263      0.793      -0.203       0.266
ma.L4          0.2193      0.044      4.937      0.000       0.132       0.306
sigma2         0.1995      0.000    436.038      0.000       0.199       0.200
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):            606502.59
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.88   Skew:                            -0.37
Prob(H) (two-sided):                  0.00   Kurtosis:                        18.88
===================================================================================
```

FORECASTING

After successfully building the model, we tested the model with test dataset and plotted the forecast. In that, we found that trendline of discomfort index will lie near 24



Learnings from 1st week

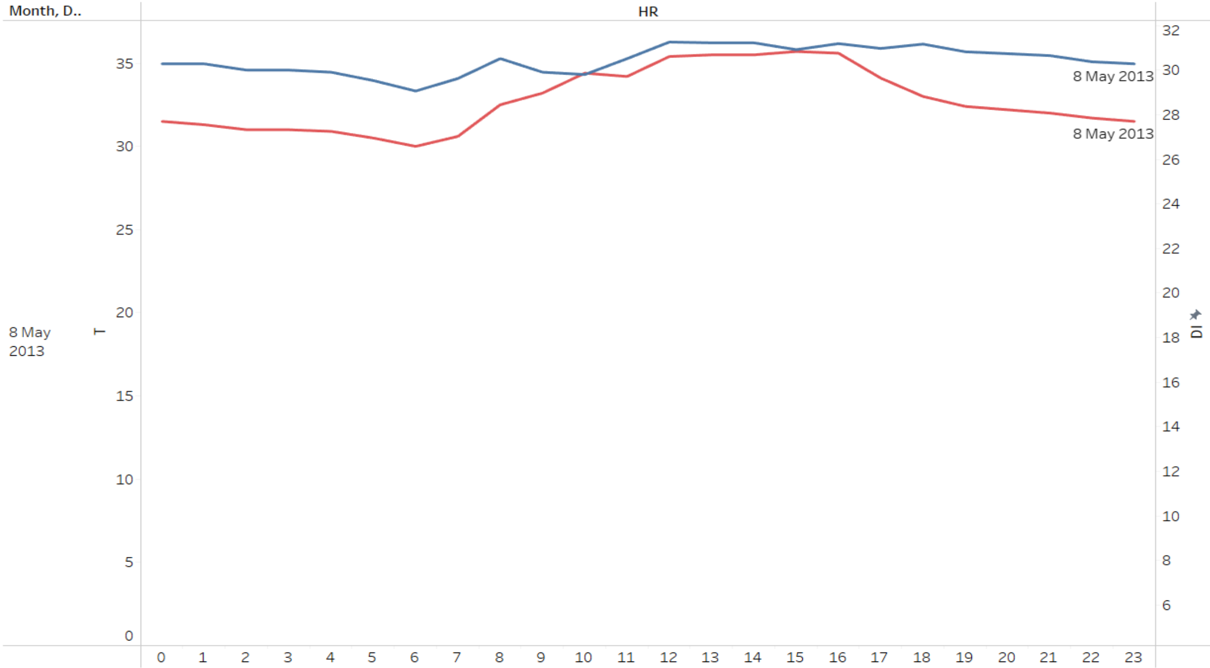
While plotting the heatmaps and trendlines we have used the average values of the variables but in time series data which contains hourly data by using the mean values we may not generate appropriate insights because the variations of each data point play an important role and the by averaging the values we loss the uniqueness of the data.

****

Week 2

Temperature and Discomfort (Day-wise)

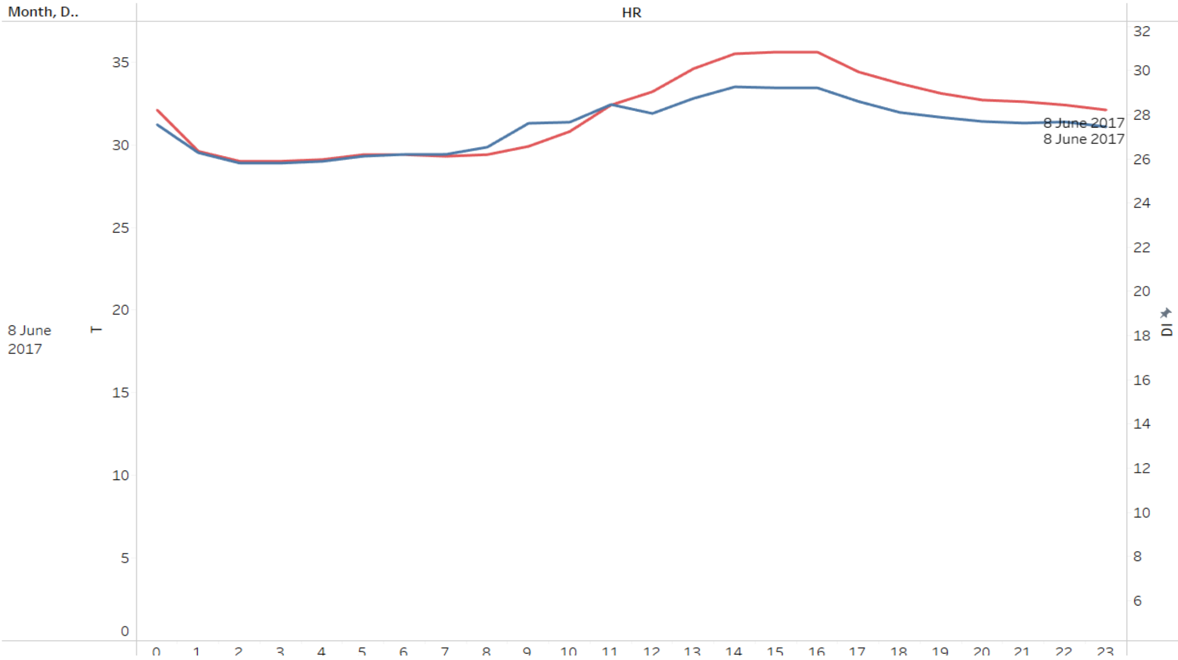
Correlation value is **0.7639**



Here, the blue line indicates Discomfort (DI) and red line indicates Temperature (T). In the beginning we compare the correlation of T and DI for the whole day, which is always positive.

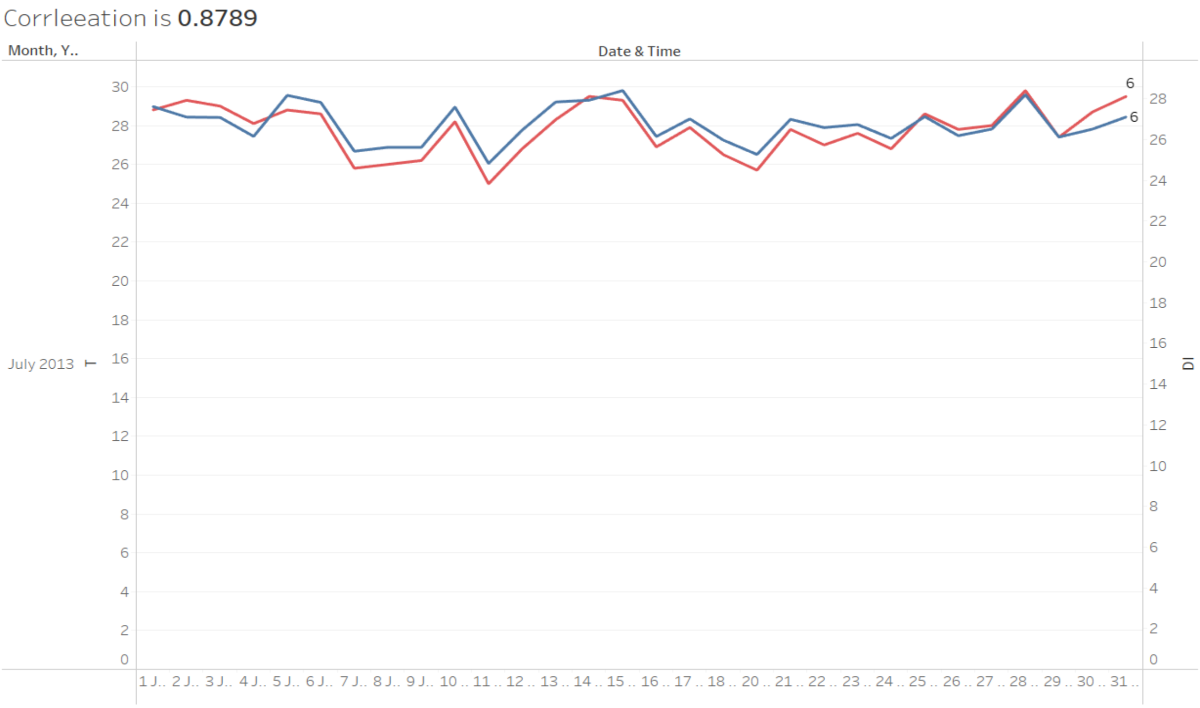
Here are some other dates of correlation for T and DI,

Correlation value is **0.9528**



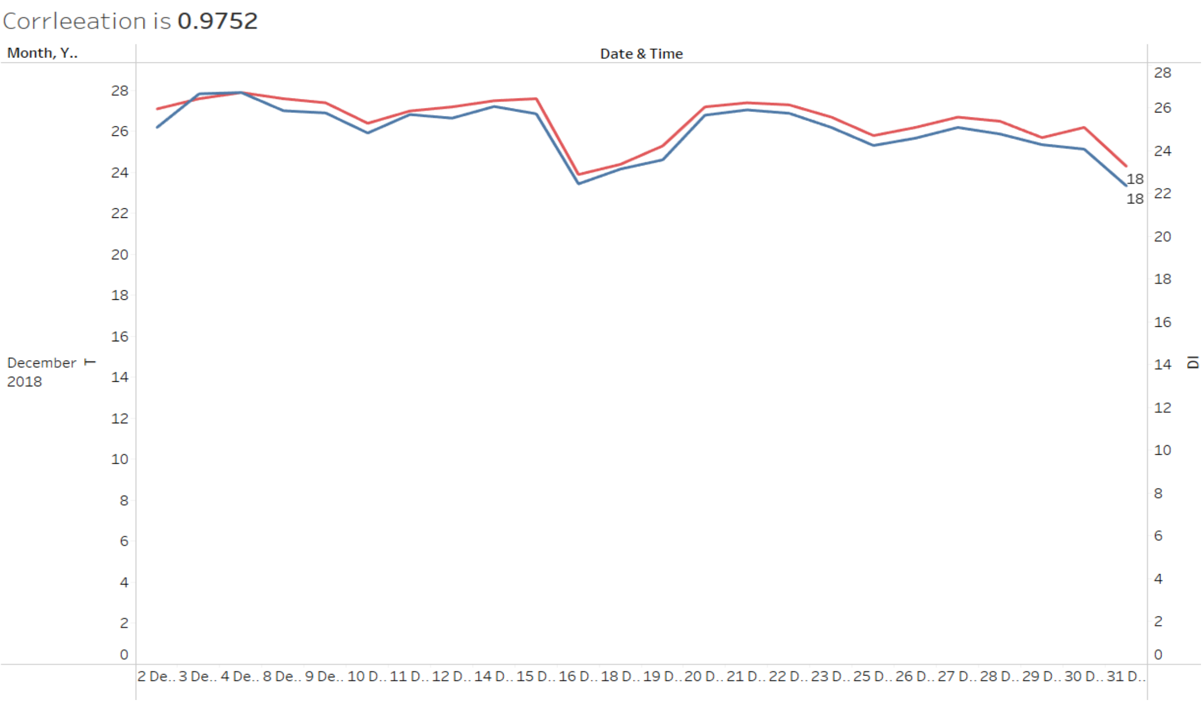
As we can see its another random date which is 8th June 2017, and the correlation of T and DI were positive which is 0.9528

Temperature and Discomfort (Hour-wise)



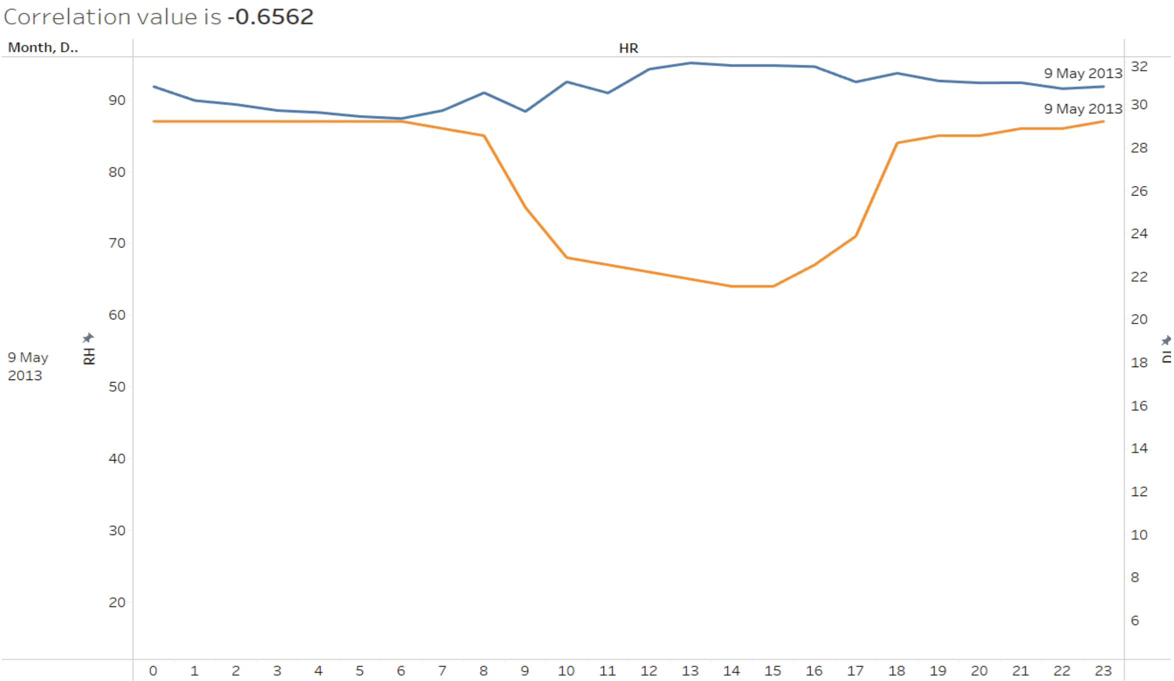
It's the correlation between T and DI in hour-wise, which is also positive for a random date July 2013. T and DI were always positive for day-wise as well as hour-wise.

Here is another date for hour-wise to check the correlation for T and DI,



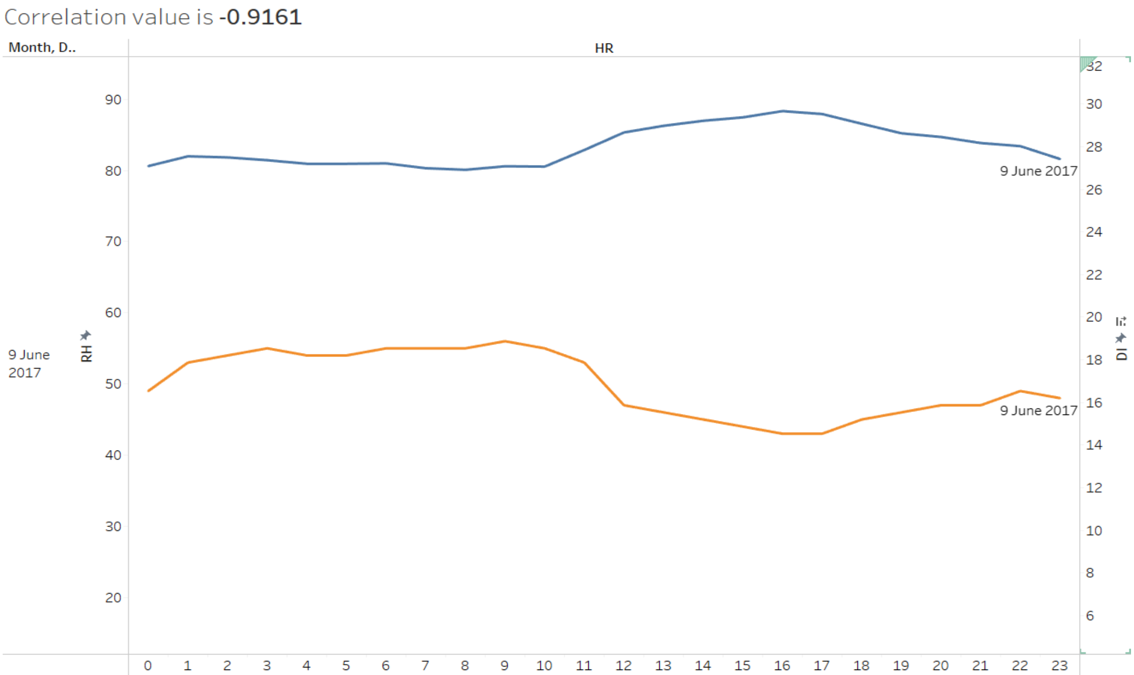
So, it is concluded that the T increases DI also increases, there is always positive correlation for T and DI.

Humidity and Discomfort (Day-wise**)**



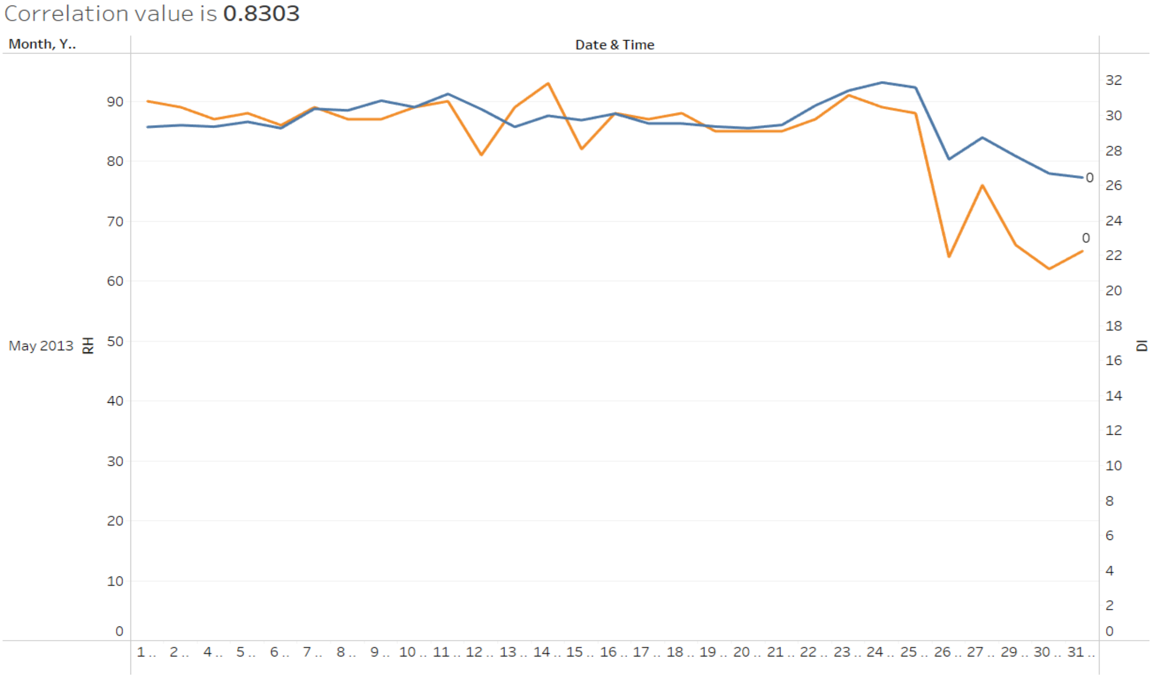
Here, the blue line indicates Discomfort (DI) and orange line indicates Humidity (RH). In the beginning we compare the correlation of RH and DI for the whole day, unlike T and DI, RH and DI is negatively correlated for day-wise.

Here are some other dates of correlation for RH and DI,



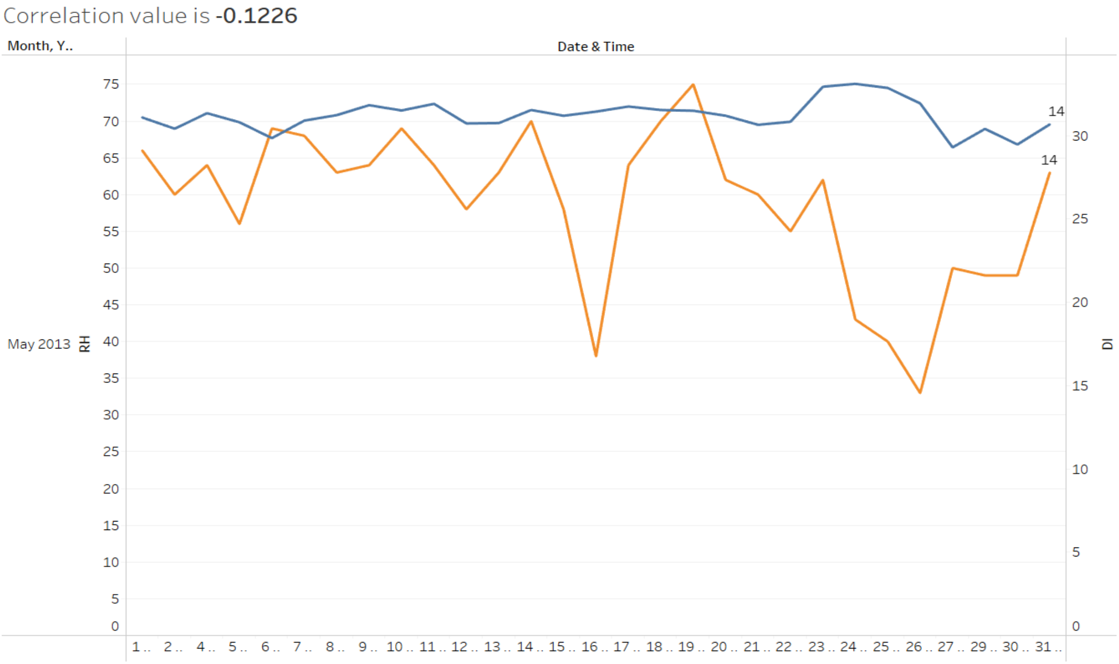
So, it is concluded that RH and DI were negatively correlated for day-wise. Now we are going to calculate for hour-wise to see any change.

Humidity and Discomfort (Hour -wise)



Correlation of RH and DI in hour-wise, which is positive for few hours and negative for few hours. Here in May 2013 for first few hours the correlation positive and in the middle hours which is in between 9 to 16 hours the correlation value is negative.

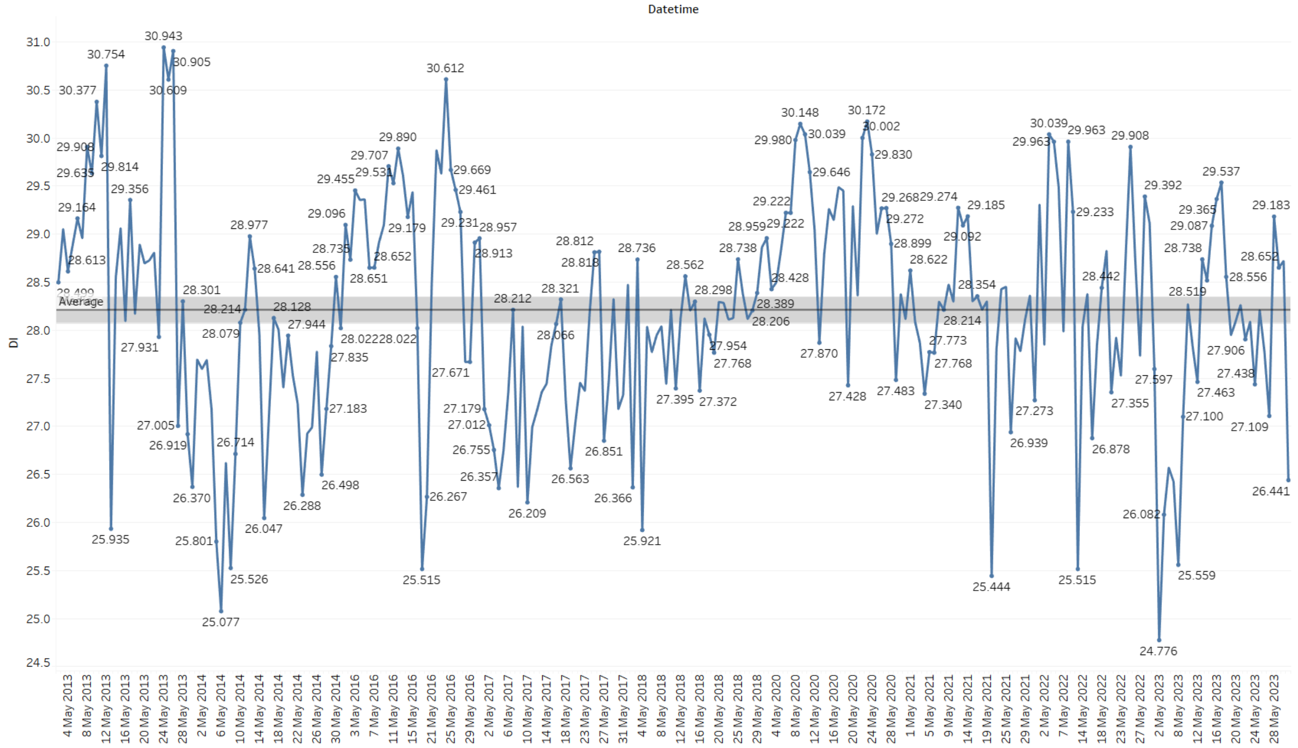
In hour-wise, RH and DI do not have always positive or always negative correlation it varies hour to hour.



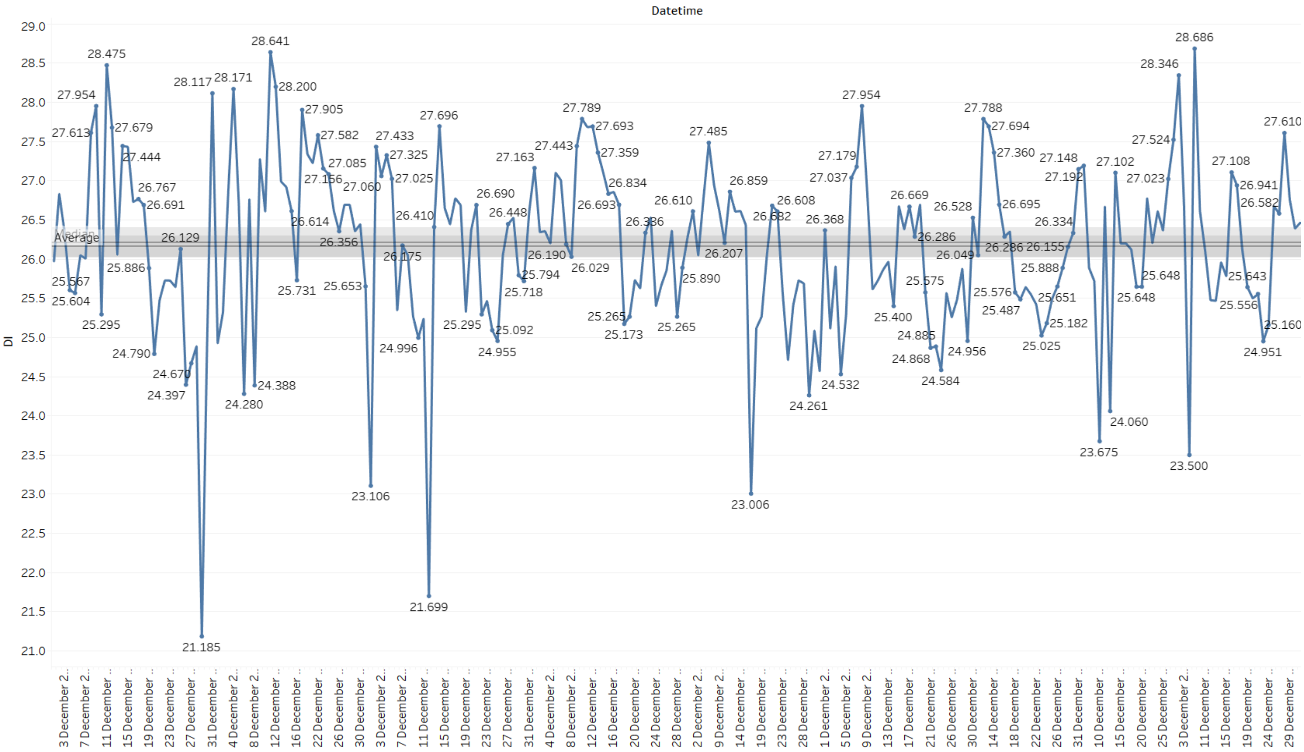
Now we can see, there is negative correlation for 14th hour for the same date. It is concluded that the correlation of RH and DI were positive for some hour of the day and negative for some hour of the same day. It varies from hour to hour.

Trend line of Discomfort

The trend line of discomfort will show how it changes for past 10 years. Each hour the discomfort index trend line was visualized, some of them were given below,



The above chart shows the 4th hour of May month for past 10 years, which is from 2013 to 2023.



This is the 13th hour of December for past 10 years. Likewise, this can be visualized for every hour and every month

Learnings from 2nd week

- There is always positive correlation between Temperature and Discomfort index, whether it is day-wise or hour-wise analyzation it is always correlated positively
- But for Humidity and Discomfort index, the correlation is negative for day-wise analyzation and for hour-wise it changes.
- In hour-wise analyzation we’ve found that the correlation is positive for few and negative for few hours.
- In a single day correlation varies from positive to negative for each hour.
- We’ve visualized the trend of Discomfort index for each hour in past 10 years, which is from 2013 to 2023.

****

Week 3

Data Visualization

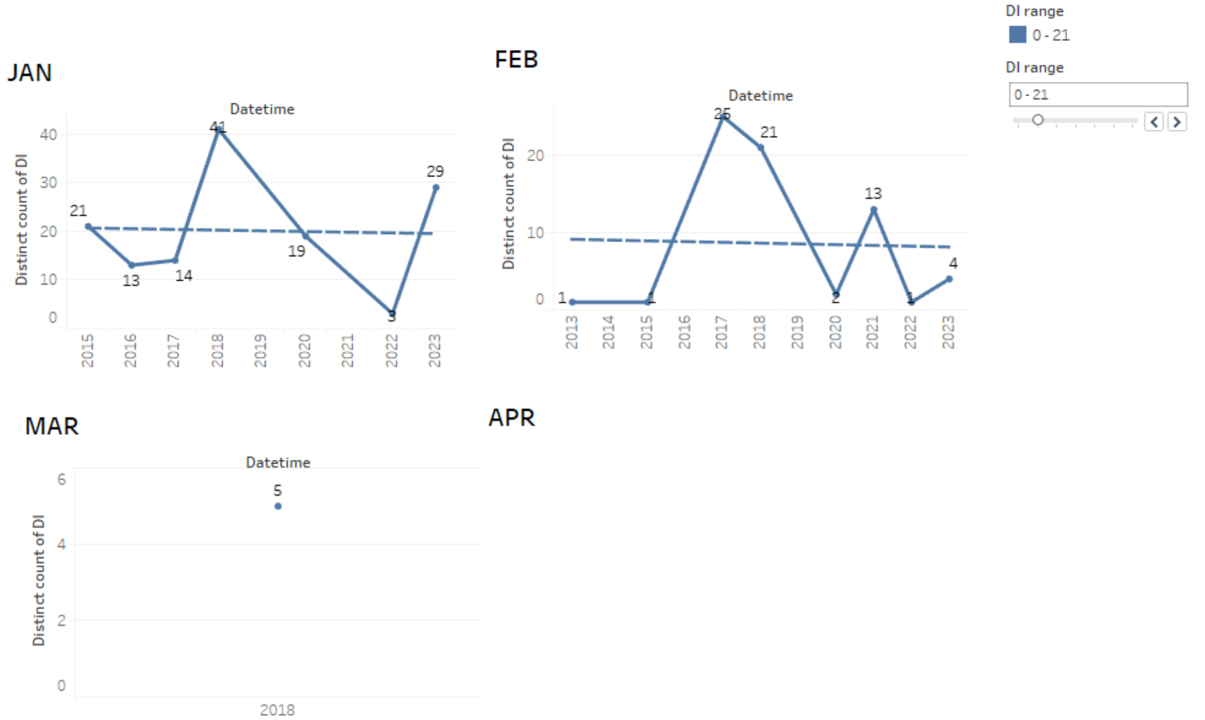
DI range and DI conditions

DI range (˚C)	Discomfort conditions
DI < 21	No discomfort
21 ≤ DI < 24	Less than 50% of people feel discomfort
24 ≤ DI < 27	More than 50% of people feel discomfort
27 ≤ DI < 29	Most of the population feel discomfort
29 ≤ DI < 32	Everyone feels severe stress
DI ≥ 32	Medical emergency

DI for past 10 years of data from 2013-2023

DI range (0-21) No discomfort

January to April

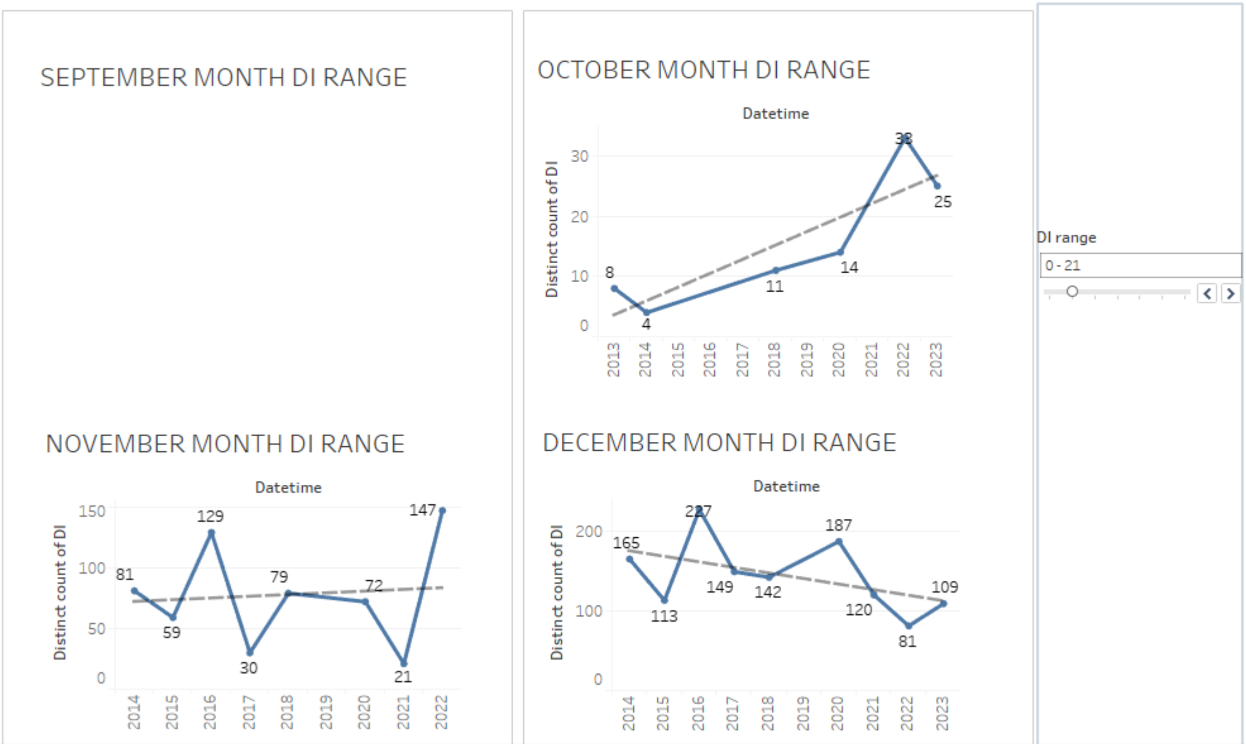


- The discomfort index range of 0-21, in January the trend lines sems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 0-21 occur in 2018 (42 times).
- The discomfort index range of 0-21, in February, the trend lines sems to be decrease in decadal analysis from 2013 to 2023. • The highest count of DI of range 0-21 occur in 2017 (36 times).
- The discomfort index range of 0-21, in march we observe only 5 observation observations in 2018.
- No observation recorded.

May to August

No observation recorded

September to December

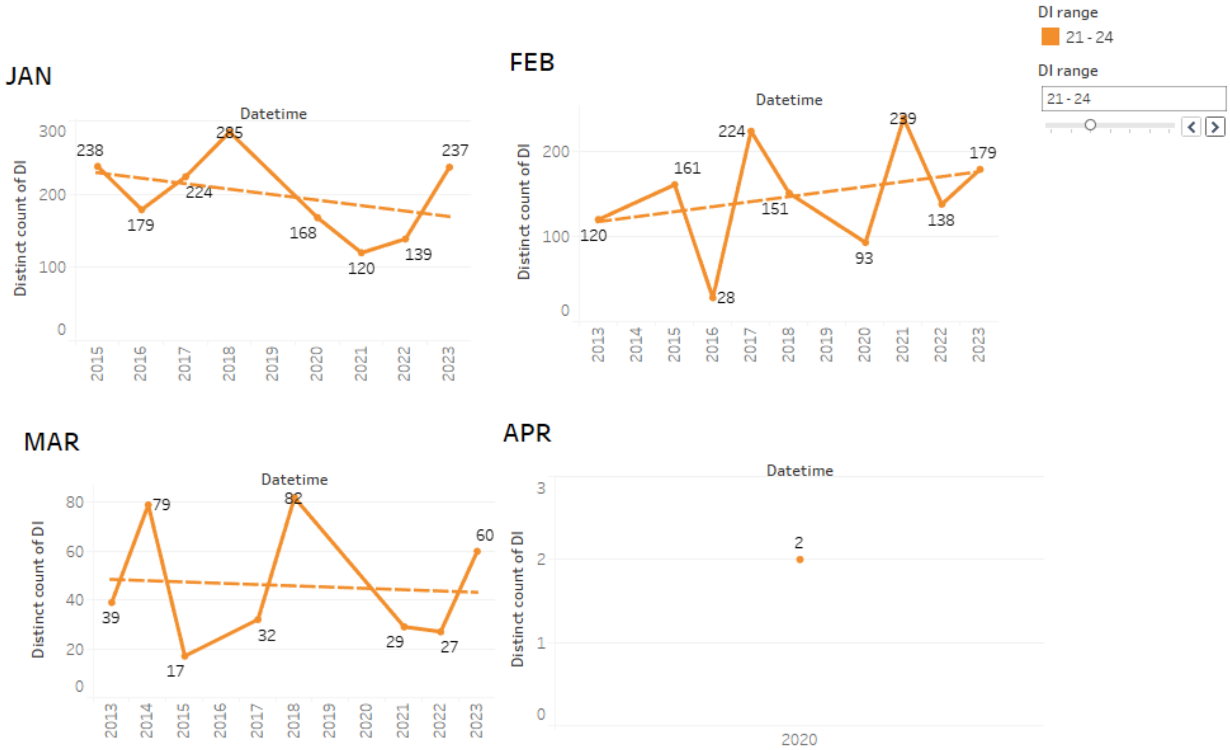


- The discomfort index range of 0-21, in October, the trend lines sems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 0-21 occur in 2022 (36 times).
- The discomfort index range of 0-21, in November, the trend lines sems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 0-21 occur in 2022 (147 times).

➢ The discomfort index range of 0-21, in December, the trend lines sems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 0-21 occur in 2016 (287 times).

DI range (21-24) less than 50% people feel Discomfort

January to April



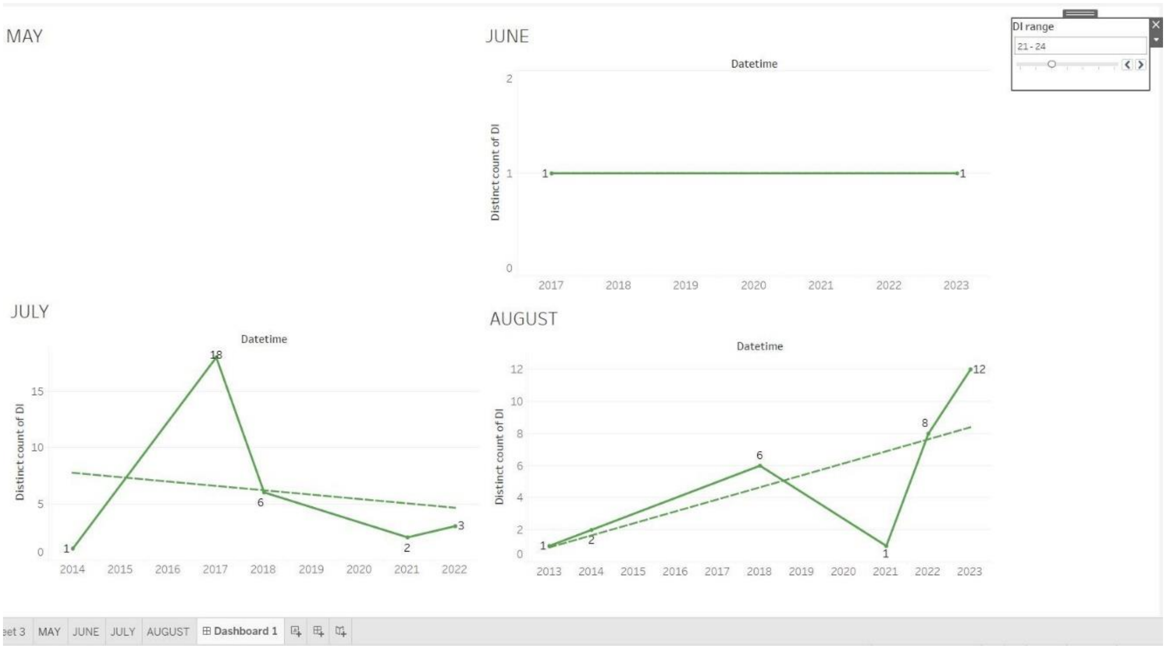
➢ The discomfort index range of 21-24, in January, the trend lines sems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 21-24 occur in 2018 (285 times).

➢ The discomfort index range of 21-24, in February, the trend lines sems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 21-24 occur in 2018 (82 times).

➢ The discomfort index range of 21-24, in march the trend lines sems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 21-24 occur in 2018 (82 times).

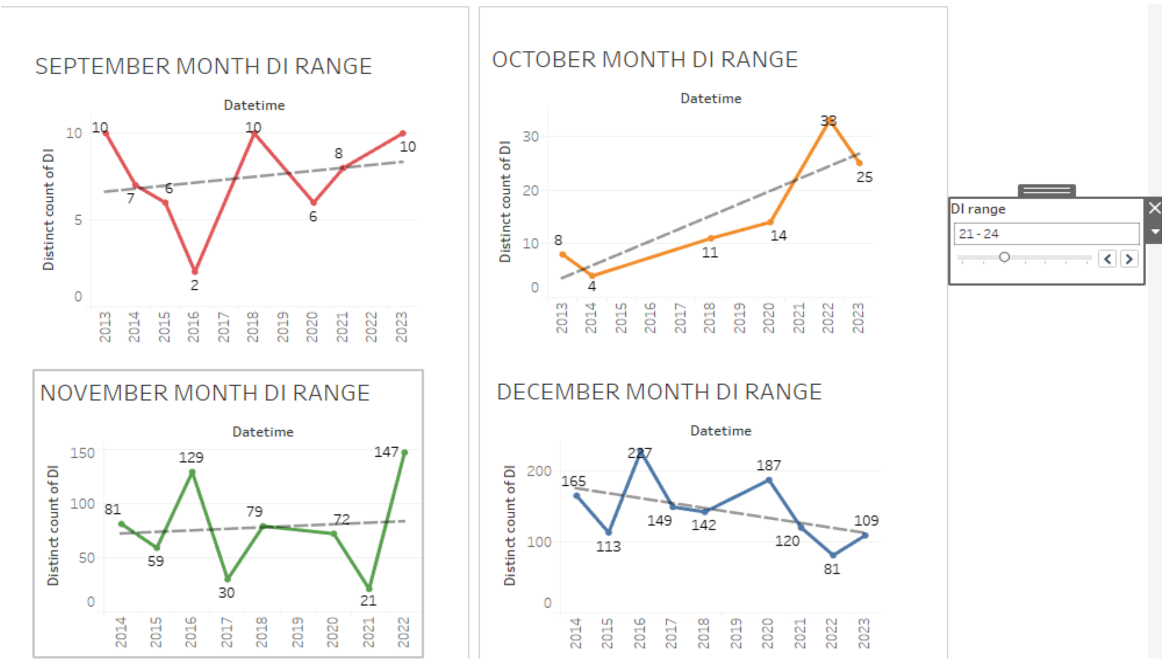
➢ The discomfort index range of 21-24, in April, only two observations has been recorded in 2020.

May to August



- ➢ The discomfort index range of 21-24 in May, there is no observation recorded.
- ➢ The discomfort index range of 21-24 in June ,there is only two observation has been recorded,one in 2017 and one in 2023.
- ➢ The discomfort index range of 21-24 in July ,there is decrease in decadal trend the highest count of DI occur in 2017(18 times).
- ➢ The discomfort index range of 21-24 in August , there is an increase in decadal trend ,the highest count of DI in range 21-24 is in 2023 (12 times)

September to December



➢ The discomfort index range of 21-24,in September the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 21-24 occur in three year 2014,2020,2023 (10 times).

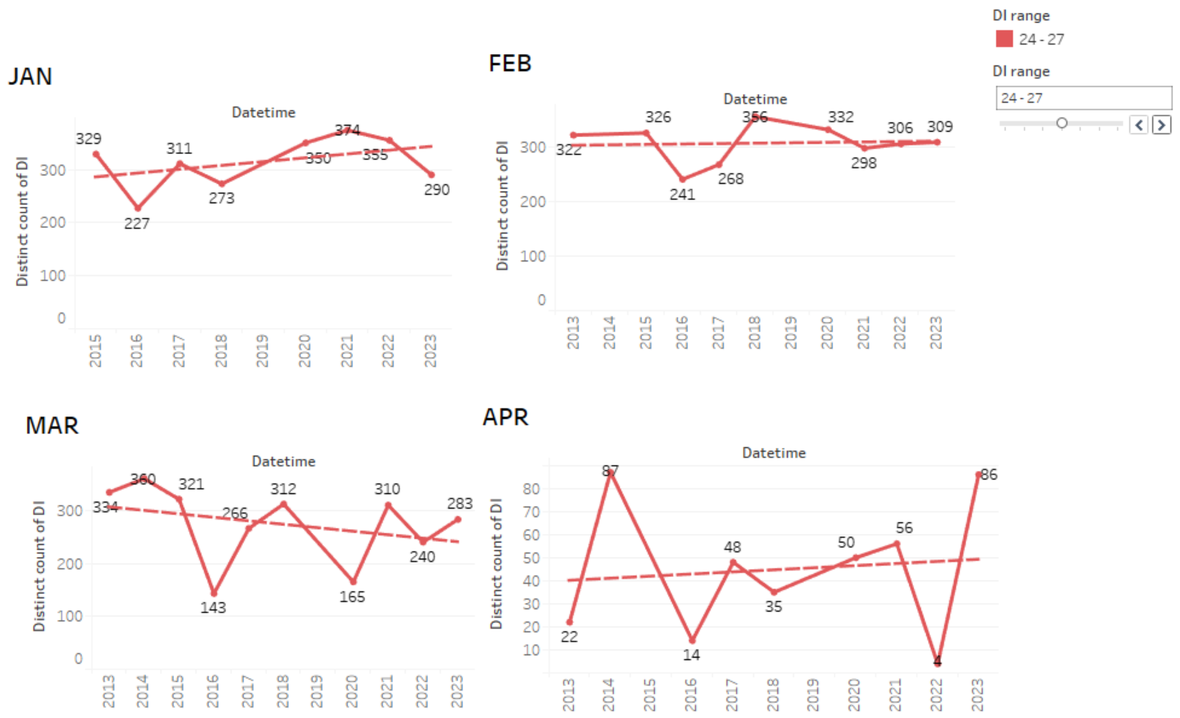
➢ The discomfort index range of 21-24 in October the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 21-24 occur in 2022 (38 times).

➢ The discomfort index range of 21-24,in November the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 21-24 occur in 2022 (147 times).

➢ The discomfort index range of 21-24,in December the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 21-24occur in 2016 (287 times).

DI range (24-27) more than 50% people feel Discomfort

January to April



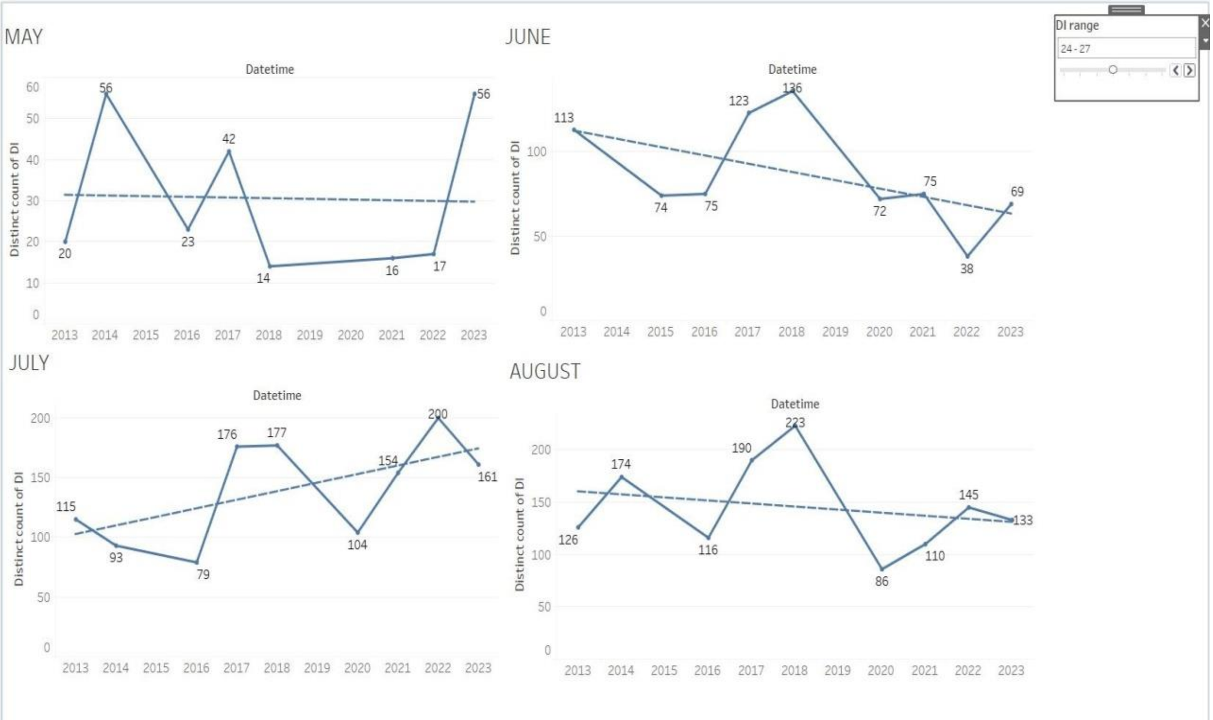
➢ The discomfort index range of 24-27,in January the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 24-27 occur in 2022 (374 times).

➢ The discomfort index range of 24-27,in February the trend lines seems to be increase in decadal analysis from 2013 to 2023. •   The highest count of DI of range 24-27 occur in 2018 (356 times).

➢ The discomfort index range of 24-27,in march the trend lines seems to be decrease in decadal analysis from 2013 to 2023.The highest count of DI of range 24-27 occur in 2015(321 times).

➢ The discomfort index range of 24-27,in April the trend lines seems to be increase in decadal analysis from 2013 to 2023.The highest count of DI of range 24-27 occur in 2023(86 times).

May to August



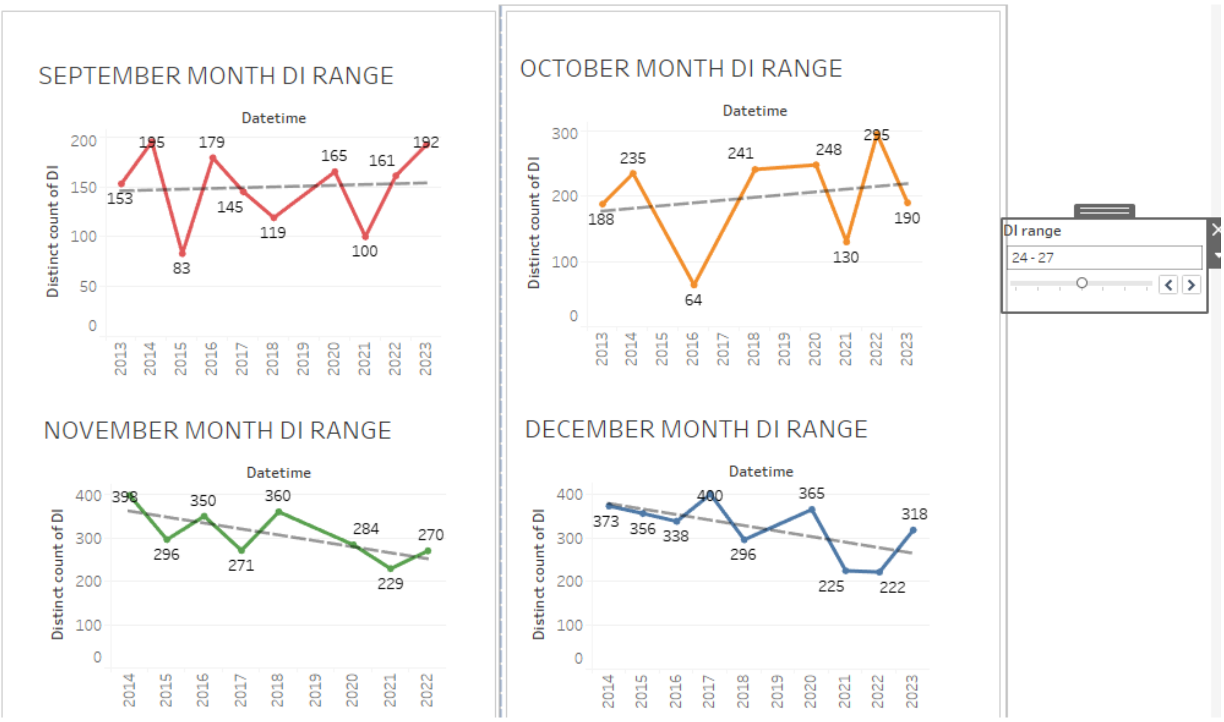
➢ The discomfort index range of 24-27 in may, there is an decrease in decadal trend ,the highest count of DI in range 24-21is in 2014,2023 (56 times).

➢ The discomfort index range of 24-27 in June, there is an decrease in decadal trend ,the highest count of DI in range 24-21is in 2018 (136times).

➢ The discomfort index range of 24-27 in July , there is an increase in decadal trend ,the highest count of DI in range 24-21is in 2022 (200times).

➢ The discomfort index range of 24-27 in August , there is an decrease in decadal trend ,the highest count of DI in range 24-21is in 2018 (223times).

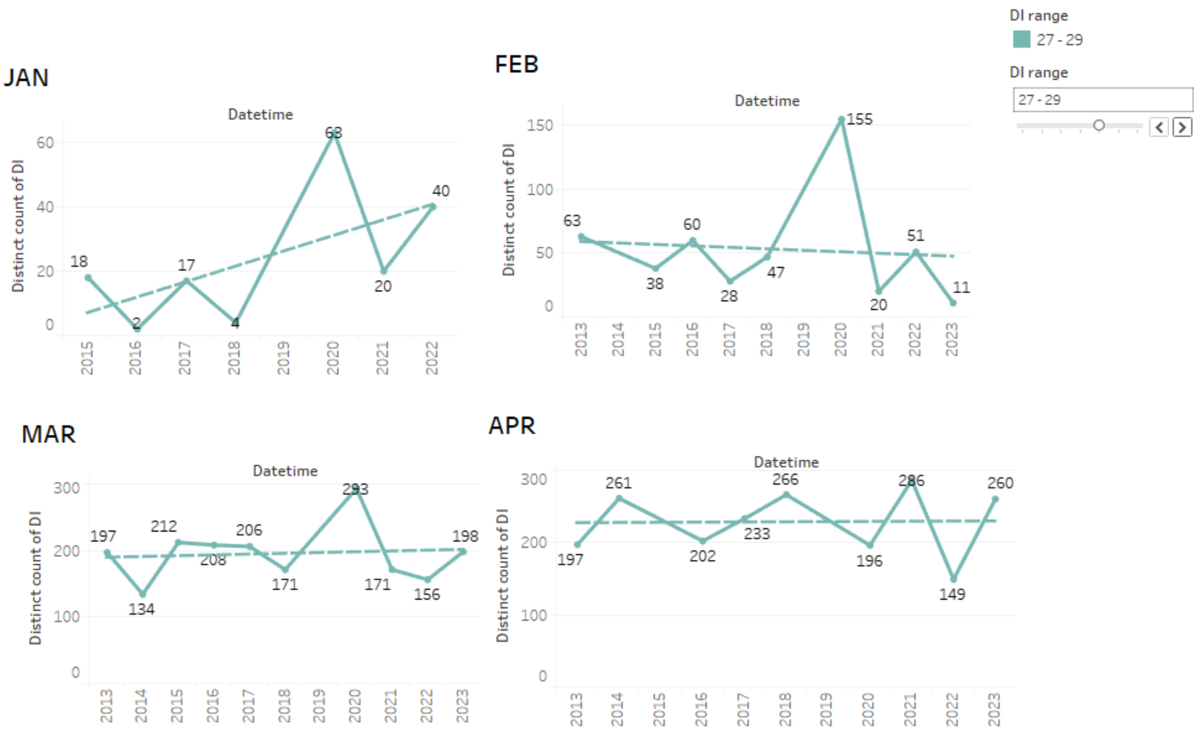
September to December



- The discomfort index range of24-27,in September the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 24-27occur in three year 2023 (192 times).
- The discomfort index range of 24-27 in October the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 24-27occur in 2022 (295 times).
- The discomfort index range of24-27,in November the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 24-27occur in 2014 (390 times).
- The discomfort index range of24-27,in December the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 24-27 occur in 2020 (365 times).

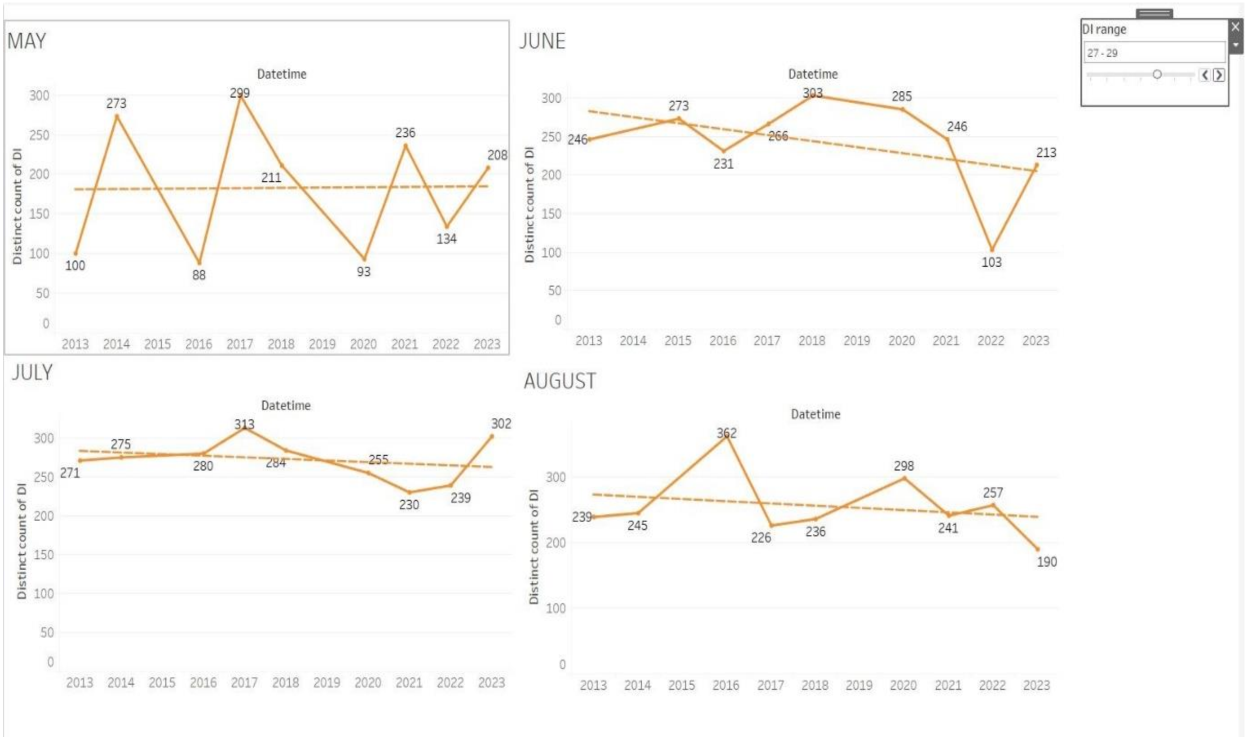
DI range (27-29) most of the people feel discomfort

January to April



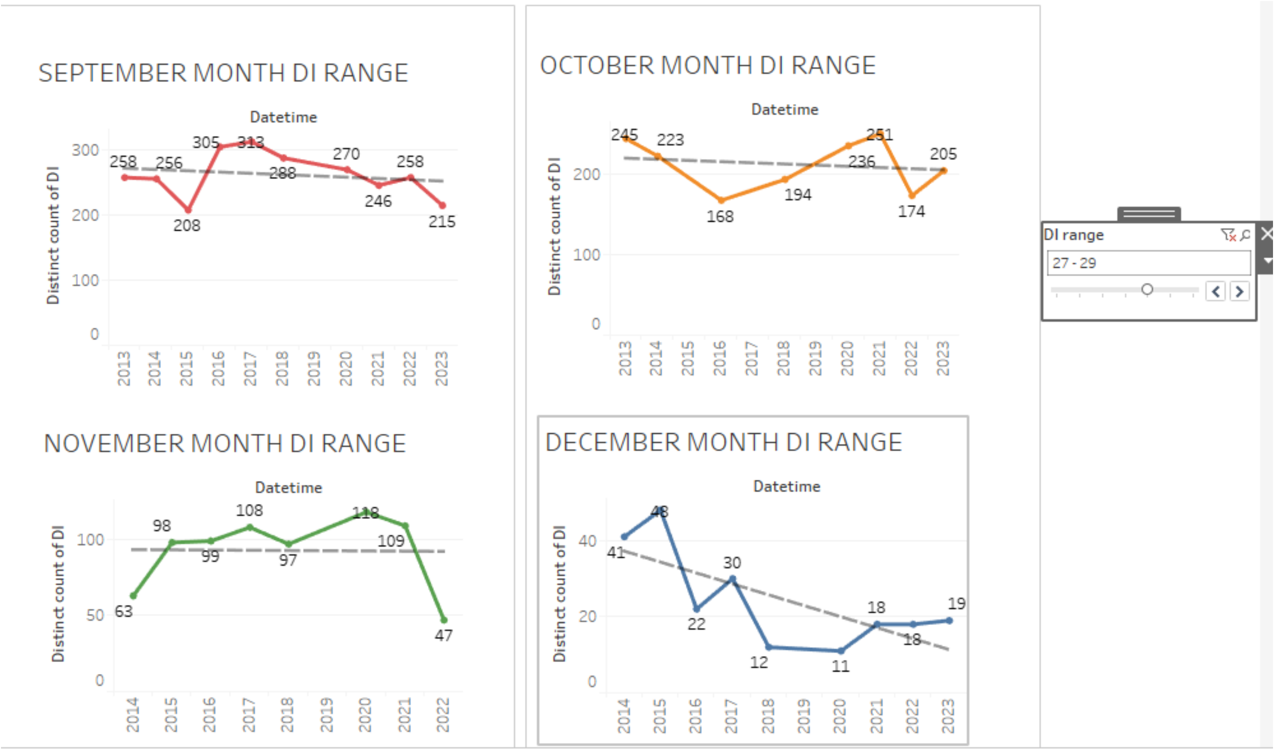
- ➢ The discomfort index range of 27-29,in January the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 27-29 occur in 2020 (68 times).
- ➢ The discomfort index range of 27-29,in February the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 27-29occur in 2020 (155 times).
- ➢ The discomfort index range of27-29,in march the trend lines seems to be increase in decadal analysis from 2013 to 2023.The highest count of DI of range 27-29 occur in 2020 (293 times).
- ➢ The discomfort index range of 27-29,in April the trend lines seems to be increase in decadal analysis from 2013 to 2023.The highest count of DI of range 27-29 occur in 2021(286 times).

May to August



- The discomfort index range of 27-29,in may the trend lines seems to be increase in decadal analysis from 2013 to 2023. The highest count of DI of range 27-29 occur in 2017 (299 times).
- The discomfort index range of 27-29,in June the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 27-29occur in 2018 (303 times).
- The discomfort index range of27-29,in July the trend lines seems to be decease in decadal analysis from 2013 to 2023.The highest count of DI of range 27-29 occur in 2020 (313 times).
- The discomfort index range of 27-29,in August the trend lines seems to be decrease in decadal analysis from 2013 to 2023.The highest count of DI of range 27-29 occur in 2016(362 times).

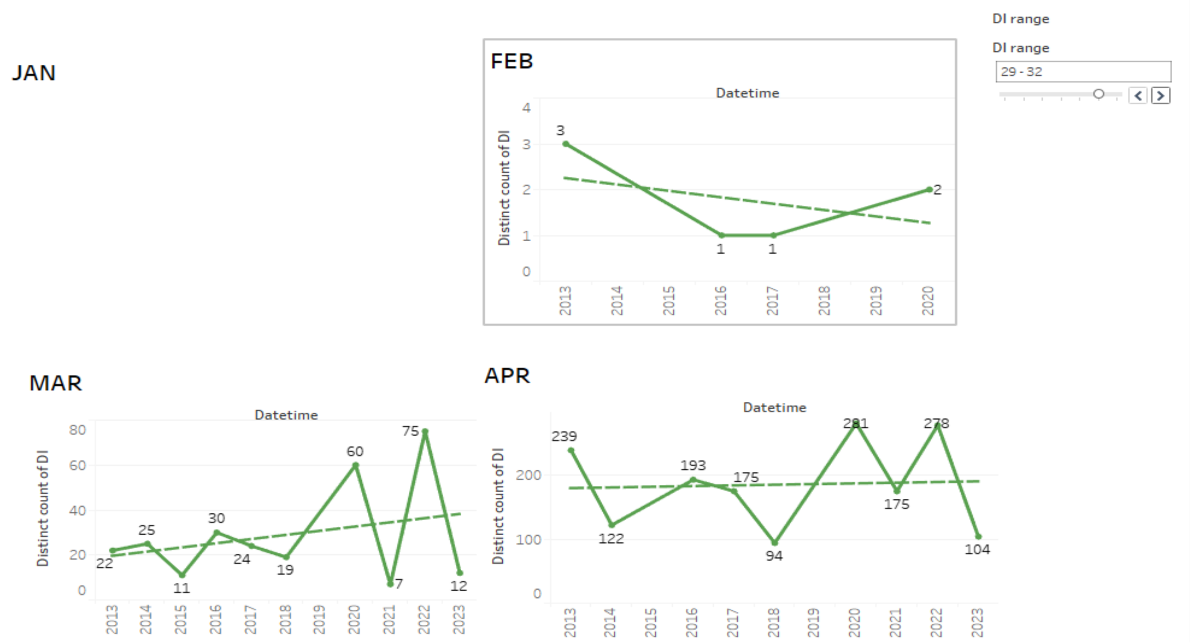
September to December



- ➢ The discomfort index range of 27-29,in September the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 27-29 occur in 2017 (313 times).
- ➢ The discomfort index range of 27-29,in October the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 27-29occur in 2021 (251 times).
- ➢ The discomfort index range of 27-29,in November the trend lines seems to be decease in decadal analysis from 2013 to 2023.The highest count of DI of range 27-29 occur in 2020 (116 times).
- ➢ The discomfort index range of 27-29,in December the trend lines seems to be decrease in decadal analysis from 2013 to 2023.The highest count of DI of range 27-29 occur in 2015(46times).

DI range (29-32) Everyone feels discomfort and stress

January to April

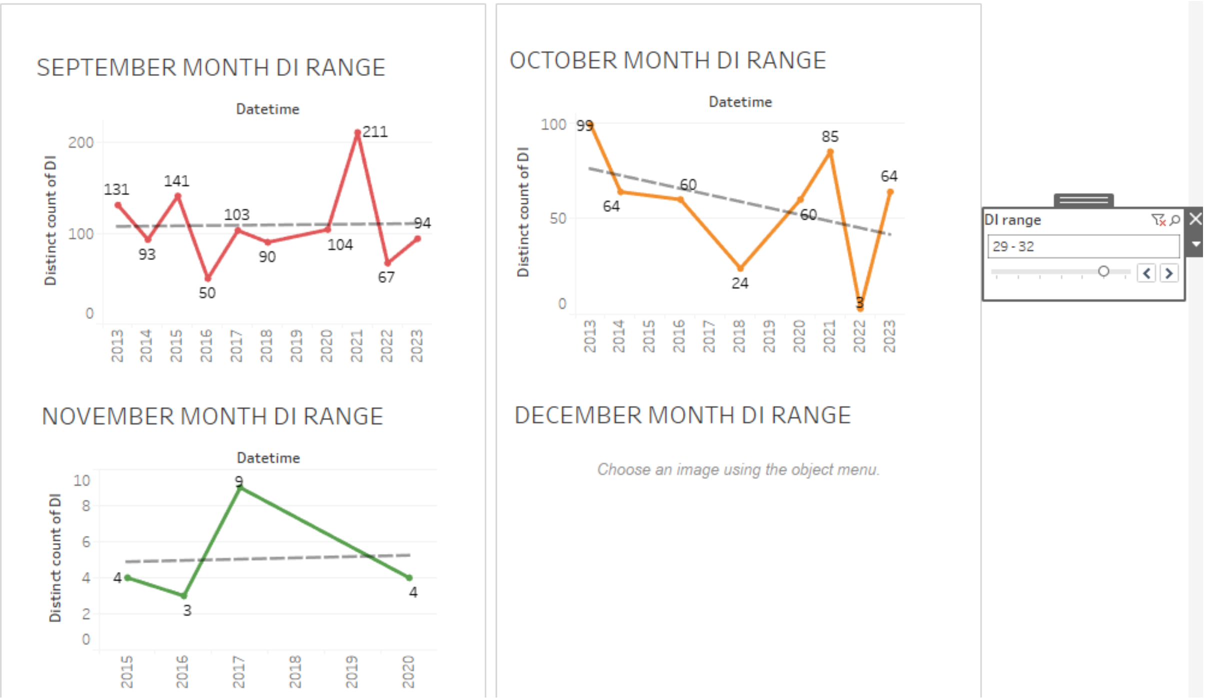


- ➢ The discomfort index range of 29-32,in January no observation is recorded.
- ➢ The discomfort index range of 29-32,in February the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2013 (3times).
- ➢ The discomfort index range of 29-32,in march the trend lines seems to be increase in decadal analysis from 2013 to 2023.The highest count of DI of range 29-32 occur in 2022 (75 times).
- ➢ The discomfort index range of 29-32,in April the trend lines seems to be increase in decadal analysis from 2013 to 2023.The highest count of DI of range 29-32 occur in 2020(281 times).

May to August

- ➢ The discomfort index range of 29-32,in may the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2020 (423times).
- ➢ The discomfort index range of 29-32,in June the trend lines seems to be icrease in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2021 (235 times).
- ➢ The discomfort index range of 29-32,in July the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2014 (219times).
- ➢ The discomfort index range of 29-32,in August the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2020 (173times).

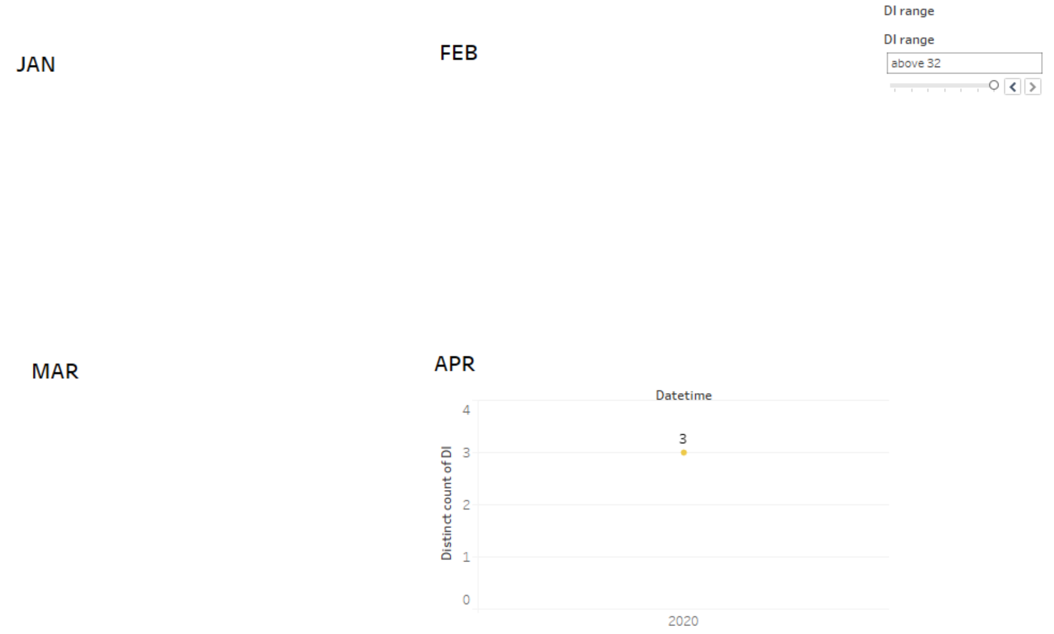
September to December



- ➢ The discomfort index range of 29-32,in September the trend lines seems to be slight increase in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2021 (221times).
- ➢ The discomfort index range of 29-32,in October the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2013 (99 times).
- ➢ The discomfort index range of 29-32,in November the trend lines seems to be decrease in decadal analysis from 2013 to 2023. The highest count of DI of range 29-32 occur in 2017 (9 times).
- ➢ The discomfort index range of 29-32 in December no observation is recorded.

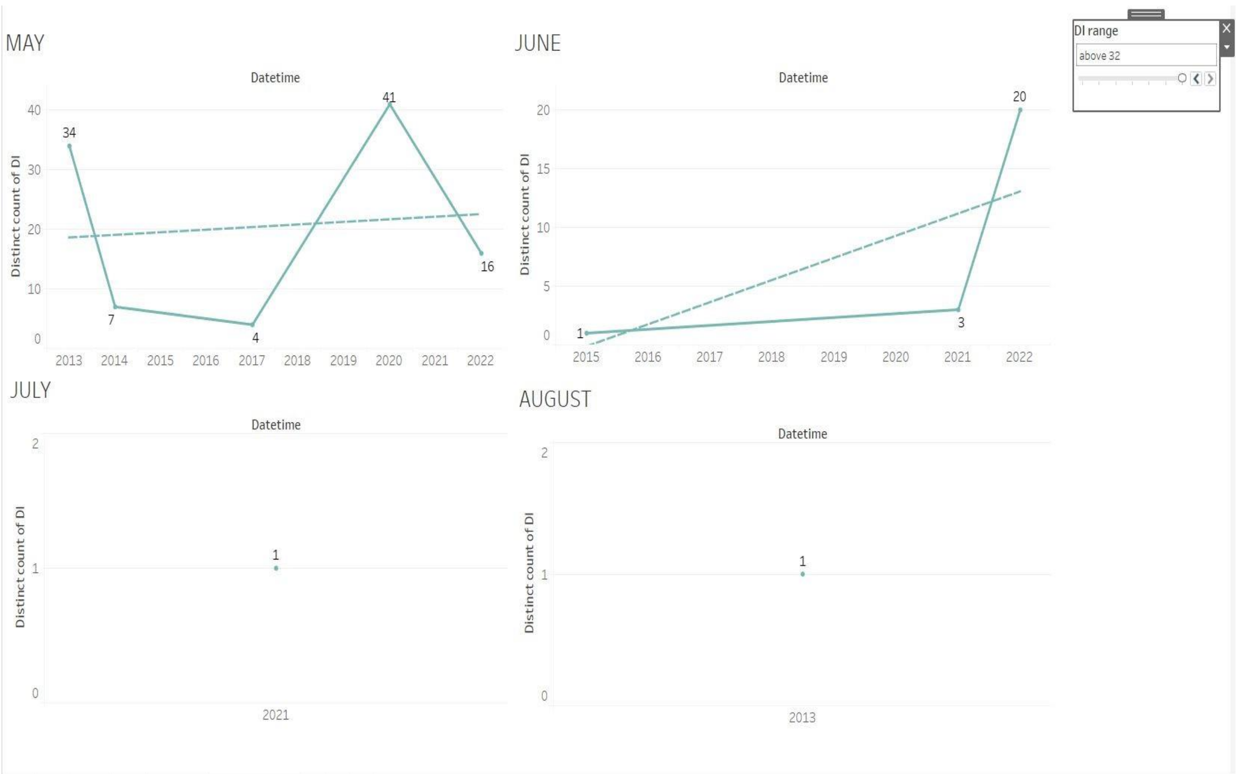
DI range (above32) Medical emergency

January to April



- The discomfort index range above 32 is observed only 3 times in April 2020

May to August



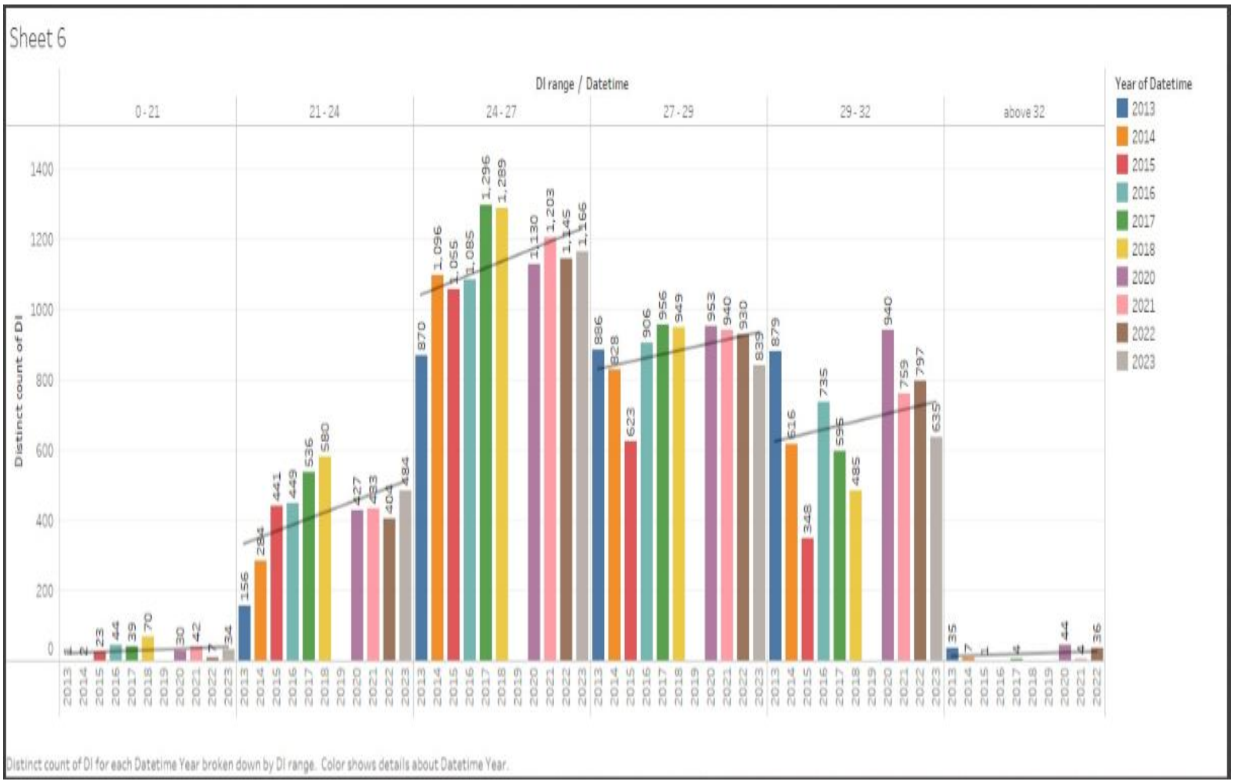
- The discomfort index range above 32 in may the trend line seems to be increasing the highest count of DI range of 32 above is found in 2020(41times).

- ➢ The discomfort index range above 32 in June the trend line seems to be increasing the highest count of DI range of 32 above is found in 2022(20 times).
- ➢ The discomfort index range above 32 in July only one observation is recorded in 2020.
- ➢ The discomfort index range above 32 in August only one observation is recorded in 2013

September to December

No observation recorded

DI range year-wise



General Observations:

1. **DI Ranges**: The chart is divided into five distinct ranges: 0-21, 21-24, 24-27, 27-29, and above 32.
2. **Yearly Trend**: Each year from 2013 to 2023 is represented with a distinct color, showing how the counts have changed over time within each range.

Discomfort Index Insights

DI Ranges

- **0-21**: This range indicates minimal discomfort. The counts in this range are very low across all years, suggesting that conditions causing minimal discomfort are rare.
- **21-24**: Represents low discomfort. There's a noticeable increase over the years, particularly in 2023, indicating that low discomfort situations have become more common recently.
- **24-27**: Represents moderate discomfort. This range has the highest counts overall, with a steady increase from 2013 to 2023, peaking in 2019. This suggests that moderate discomfort has been the most common condition.
- **27-29**: Represents high discomfort. Counts in this range have shown a steady rise, particularly peaking in 2021, indicating an increase in high discomfort conditions.
- **Above 32**: Represents extreme discomfort. The counts are minimal across all years, with slight upticks in certain years (2014, 2022, 2023), suggesting that extreme discomfort conditions are rare but have been slightly increasing.

Yearly Trends

- **2013**: Low discomfort index values across all ranges.
- **2014-2015**: Slight increases in low to moderate discomfort ranges.
- **2016-2019**: Significant increases, especially in the moderate discomfort range (24-27). This period shows a marked rise in discomfort levels.
- **2020-2021**: High discomfort levels, with 2021 peaking in the 27-29 range.
- **2022-2023**: Mixed trends, with 2023 showing an increase in the low discomfort range (21-24) and slight increases in higher ranges.

Overall Trends

- **Increasing Trend**: There is a clear upward trend in the discomfort index over the years, with more occurrences in the moderate and high discomfort ranges (24-27, 27-29).
- **Extreme Discomfort**: While rare, the counts for extreme discomfort (above 32) have shown slight increases, indicating a growing occurrence of extreme discomfort conditions.

Learnings from 3rd week

The discomfort index has shown a significant upward trend from 2013 to 2023. The increase is most pronounced in the moderate (24-27) and high (27-29) discomfort ranges, suggesting that overall discomfort levels are rising. The slight increases in extreme discomfort (above 32) are concerning and warrant further investigation into the underlying causes.

OVERVIEW

During Our internship, we engaged in a comprehensive analysis of the discomfort index (DI) spanning a decade from 2013 to 2023. This involved various data visualization techniques to uncover patterns and trends. Initially, we used Seaborn in Python to create correlation heatmaps, revealing that temperature positively correlates with DI while humidity shows a negative correlation. Subsequently, we utilized Tableau Desktop to generate calendar heatmaps for DI, temperature, and humidity, highlighting seasonal variations and long-term trends.

In the first week, we observed that discomfort levels peaked between May and August and were lowest in January and February. The analysis also included seasonal decomposition and trend analysis using a SARIMAX model for forecasting. The model predicted that the DI trend line would hover around 24 in the future. In the second week, we focused on hourly and day-wise correlations between temperature, humidity, and DI. Temperature consistently showed a positive correlation with DI, while humidity's correlation varied hourly.

During the second week of my internship, my focus was on delving deeper into the correlations between temperature, humidity, and the discomfort index (DI) on an hourly and daily basis. we used data visualization tools to create detailed charts and graphs that revealed how these variables interacted over different times of the day and across various days of the week.

In the third week, we categorized DI into different ranges and analyzed these across months and years. For instance, we found that the DI range of 21-24, where less than 50% of people feel discomfort, had varying trends throughout the year. Our analysis provided insights into how discomfort levels fluctuate, highlighting specific months and years where extreme discomfort was most prevalent. This comprehensive analysis of DI trends over the past decade has equipped me with valuable skills in data processing, visualization, and trend forecasting.

INTERPRETATION

The internship report focuses on data visualization techniques applied to a dataset spanning 10 years, aiming to uncover correlations, trends, and patterns. The report begins with an introduction highlighting the importance of data visualization in uncovering insights from complex datasets. The data collection and preparation section outline the dataset used and any preprocessing steps undertaken.

The main analysis includes correlation and trend analysis, utilizing various visualization methods such as scatter plots and line charts to illustrate relationships between variables over the 10-year period. Hourly and monthly trends are also explored, showcasing how trends vary throughout the day and across different months. Additionally, a discomfort analysis is conducted, using a relevant metric to visualize discomfort levels over the 10-year period.

The skills and knowledge acquired in data processing, visualization, and trend analysis have been invaluable. This experience not only deepened my understanding of environmental data analysis but also equipped me with practical skills in using tools like Python and Tableau.

The report concludes with a summary of key findings, emphasizing the significance of the identified correlations, trends, and discomfort patterns. Recommendations for future research or applications are also provided, based on the insights derived from the analysis. The report is supported by references and an appendix containing additional charts, tables, and data used in the analysis.