WEATHER DATA ANALYSIS

A.Harsha , A.Sairam , K.Saiteja , K.Sriram , M.Veda Sampreetha

Department of Artificial Intelligence Engineering, Amrita school of engeneering

***Abstract*— Weather plays an important role in many aspects like business, tourism, livelihood , agriculture , etc. This weather data affects the results of many processes.For an example if someone plan for a trip of 7 days because of sudden change in weather may disturb the trip not only in the case of trousim but even in many aspects this may happen. So, analysing of weather data is most important in our day to day life.We are having a huge data of weather to analyse better and accurate using python modules.**

Keywords*— *Weather analysis , Big data , correlation , data visualisation

I. INTROUDUCTION

Weather data analysis involves processing large data obtained from various sources, such as Satillites, weather stations and other sensors, to analyse about weather patterns and trends.Weather data analysis is becoming increasingly important as weather change will have continues effects on particular region. However, analysing huge amount of weather data can be a tough task and advanced techniques should be implemented to get more accurate result.

There is where we use python, a programming platform can be valuable, Python allows for processing and analysing of large volumes of data in a scalable and efficient manner, making it well good for weather data analysis. Using different modules in python can analyze large data, identify significant trends and patterns and gain insights. Understanding weather patterns and trends is important for many industries, including agriculture, energy and transportation as it can help in decision making and process of planning.However to analyze large data manually is impossible so we use python to analysze even using graphical representations such that getting insights is easier.

We are going to use data preprocessing, data cleaning , data analysis , data visulisation .Attributes that we are going to take to analyse the data are :

1.Instant Air Temperature (Celsius degrees)

2.Maximim Air Temperature (Celsius degrees)

3.Minimum Air Temperature (Celsius degrees)

4.Relative Humididty

5.Maximum and minimum Relative Air Humidity

6.Instant , max and min Dew point

7.Instant , max and min Air atmospheric pressure

8.Wind direction

9.Wind gust intensity

10.Solar radiation

11.Precipitation

12.Elevation

13.Observation Datetime

14.Station number (location)

Modules of pyhton are numpy ,pamdas ,matplotlib ,os ,seaborn ,plotly ,cufflinks.

Pandas: pandas is a python library which is for data manipulation and analysis. It provides operations and data structures. Pands is more flexible, fast and anyone can use it in ease.It is an open souce with high performance

Numpy: Numpy is a python module provides numerous operations to handle multidimentional arrays and matrices and having high level mathematical functions like comprehensive mathematics to work with arrays. It provides well optimized code and have a good speed.Numpy module is also an open source.Here we use this to deal with dataset.

Matplotlib: To visualize the data and results in python for better understanding matplotlib is used it is built to

integrate with nympy and design to do the things with scipy and having several types of plots like bar, line, pie, boxplot etc…Using it is so ease and being open source can be accesable to anyone.

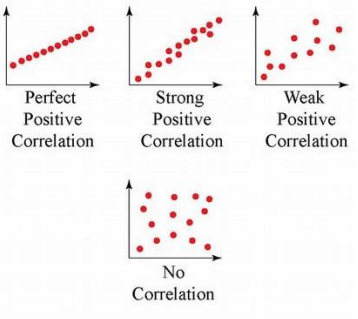
Os: os is a python library used to work and change the directory where you are going to work on.Os.chdir() is the inbuilt function to change the working directory.

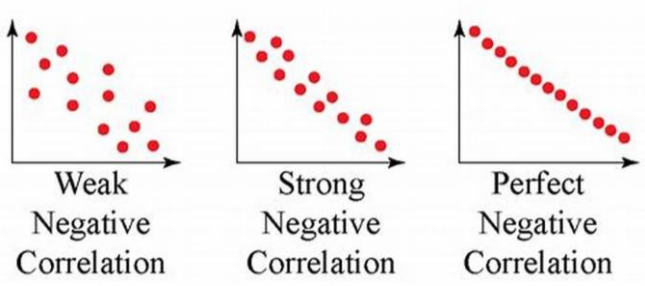
Seaborn: Seaborn is the extension library for matplotlib,it is used mostly for visualize the data in stastical plotting it provides varity number of default styles and patterns of colours to make the stastical plots to visualize in a clear way.It helps to mapp the data semantically and stastical aggregation to generate plots which are informative.

Plotly: Plotly has a nature of dynamic charts it allows additional operations to visualize like zoom in and out graphs, otlines identification in dataset and we can directly plot dynamic chats from web.it prodives an additional unique feature animations and interactive graphs which are highly helpful to visualize the data in much more detail and deep to have accurate insights.

Clufflinks: Clufflinks is also a python library that connects pandas and plotly that helps to create dataframe charts directly.It behaves lika a plugin.

And we use correlation matrix to find the relation between the attributes used in our datasets.
Correlation:Correlation represents a cofficent named correlation coefficient which lies in range of [-1 , 1] and tells the direction of a relation among variables





II. LITERATURE SURVEY

Hadoop has been proven to be a useful technology for analysing climate data in previous studies. For instance, Zhang et al. (2016) used Hadoop to examine climate data from a network of Chinese meteorological stations. According to the study, Hadoop's distributed processing capabilities sped up data processing and helped researchers to spot important climatic trends and patterns.
Hadoop was utilised in a different study by Duro et al. (2014) to examine climate information gleaned from satellite photos. The study concentrated on the examination of cloud cover patterns, which can significantly affect climate patterns and projections of the weather. Large amounts of satellite data may be processed effectively using Hadoop's MapReduce programming approach, according to the researchers.

Hadoop has also been utilised in other projects to analyse climate data for certain regions or nations. Hadoop was utilised, for instance, in a study by Patel et al. (2017) to examine patterns in temperature and precipitation across time using climate data from India. According to the study, Hadoop's distributed processing capabilities made it possible for the researchers to efficiently analyse the huge amounts of data and pinpoint important climatic trends.
Hadoop has also been used to analyse climate data in conjunction with other methods of data analysis, such as machine learning. Hadoop and machine learning methods were utilised in a study by Palacios-Callender et al. (2017) to examine climate data fromcolombia. According to the study, Hadoop and machine learning techniques allowed for more precise precipitation pattern forecasts than conventional approaches.
Overall, these studies show how Hadoop may be used to analyse climate data and highlight its advantages for handling vast amounts of climate data in a scalable and effective way.

III. METHADOLOGY

A)Collecting Datset:

We gathered dataset having weather data with the attributes listed above in introduction to analyse weather patterns .

B)Data preprocesing and cleaning

Before data analysis we make sure the data should be preprocessed and cleaned.Data preprocessing and cleaning is a process of

- Identifying the incorrect
- Incomplete
- Irrelevent
- And missing part in the data
- Modify ,replace and delete based on requirments

Data preprocessing steps that we are following to clean oyr data are:

- Tyding of data
- Dealing with Time related Columns
- Missing Data correlation
- Missing Data visualization

Tyding of data: The collected dataset is a mess with different crossovers and some value names will be weird as compared with remaing data this data is termed as untidy data.analysing of untidy data is so hard .We need to make the data tidy i.e each column should represents separate variables and rows should represent individual observations.

Dealing with time related columns: In the collected dataset we have the columns mdct, date, yr, month and hour which represents similar data but in different sections.With in those in analysis date and time columns are enough so we need to drop all other colums such that computational time reduces and data should be clear of unwanted stuff for clear analysis.

Missing Values:

a)Missing Data correlation:In the dataset of huge volume there may be some values that are not filled or having null value we need to remove all the null values or replace the null or missing values using statestical methods. We plotted customized heatmap that clear the repitated values by plotting only lower diagonal elements.The value of correlation across columns is in the range of >-0.25 and <0.25 which are classified as high correlation.

b)Missing data percentage caluculation:we calculate the percentage of missing value as a record of getting the original data. We caluculated using mathematical equation: (null_counts / sample_df.shape[0])*100 . null_counts represents the number of null vlaues in the data and .shape[] function is used to have the dimensions of dataset that we are taken and shape[0] represents number of columns.

c)Missing data visualization using Heatmap:

After the colums are getting sorted.analysing the heatmap of correlation to make the analysis precise divide that range in 3 interval.now we have to create a functions which fills the column values with different intervels of time and within the intervel the mean should be present.in our model we have created five intervels.

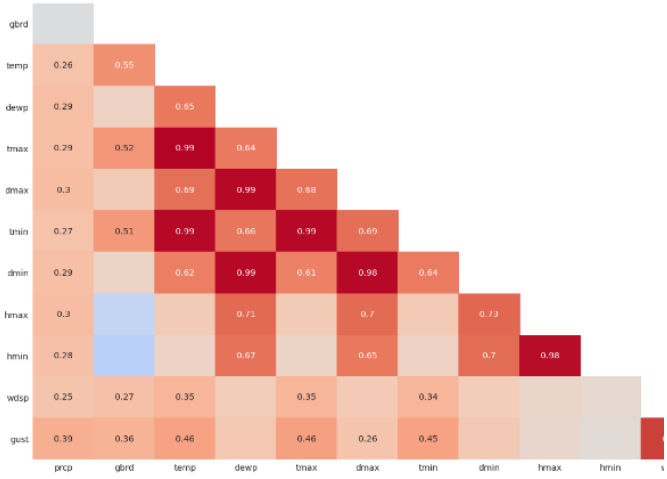
C)Data Analysis:

After completion of data preprocessing and cleaning next we moved to analize the data .

- Created Weather Data by province pie chat that contains the amont of data that is present of that particular place in the dataset. We used sklearn to plot the piechat using MinMaxScaler function
- Plotted a bar graph having data of average temperature of each city and the value is frequency normalized using iplot imported from plotly library
- Plotted a bar graph representing weather factors related to provinces. Factors:
  - Air pressure
  - Humididty
  - Solar Rdiation
  - Gust
  - Temperature
  - Dew Point
  - Wind speed
- Created a box-plot Yearly Average temperature from 2000 to 2015 with a scale of 5 years.

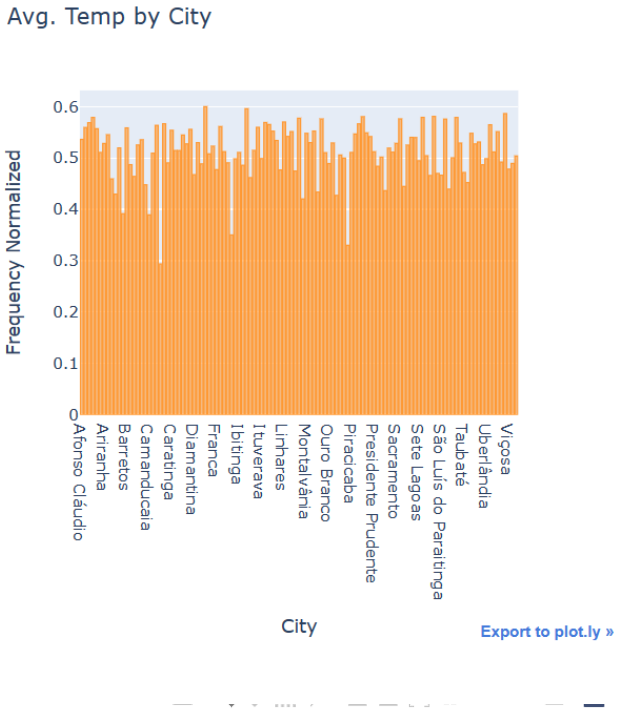
- Have a Temperature dataframe extract temperature records of our choice.and sketching chats between the attributes

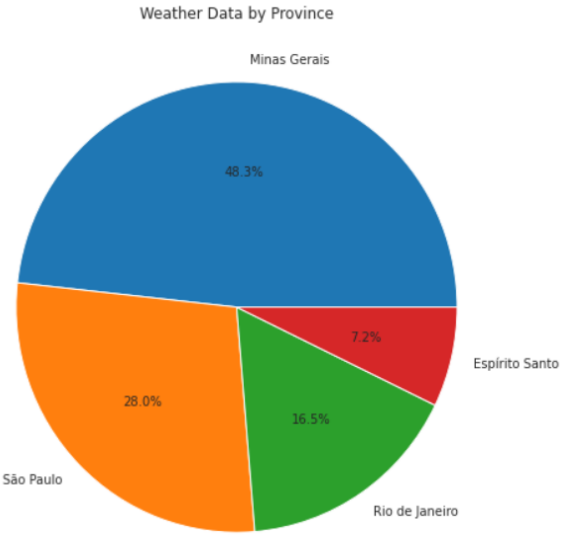
IV. RESULT:



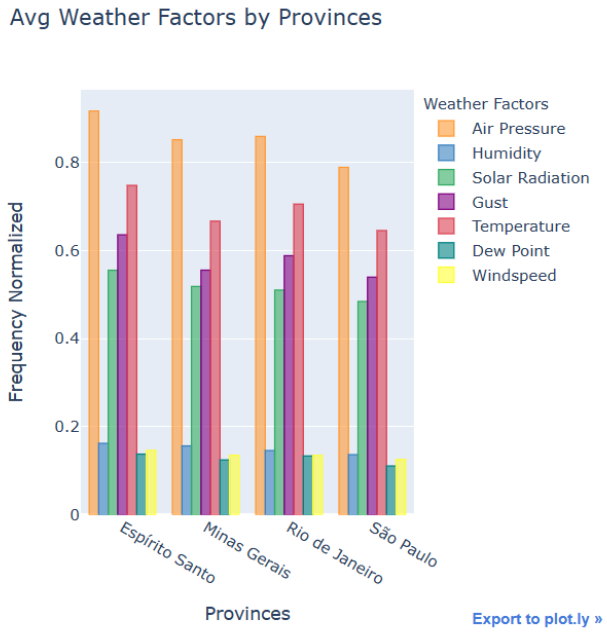
Correlatoin of Null Columns in Heatmap

Columns gust and wdsp have high correlation and both are positively affected by temperatures

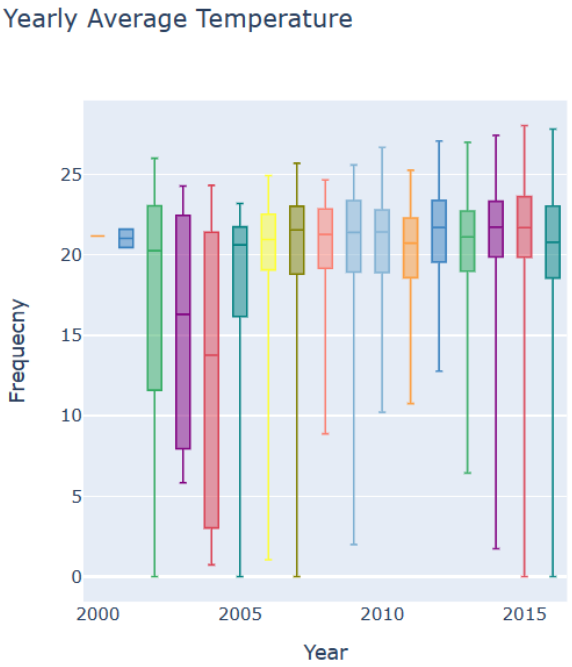




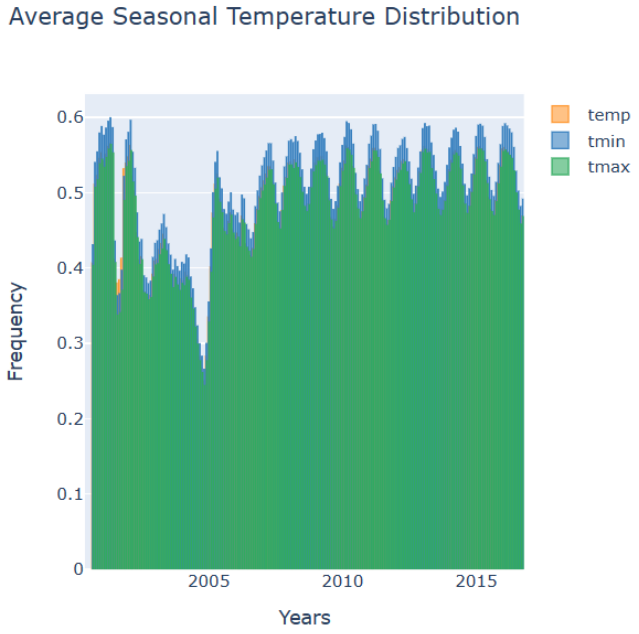
50% of data is mostly collected from Minas gerlas



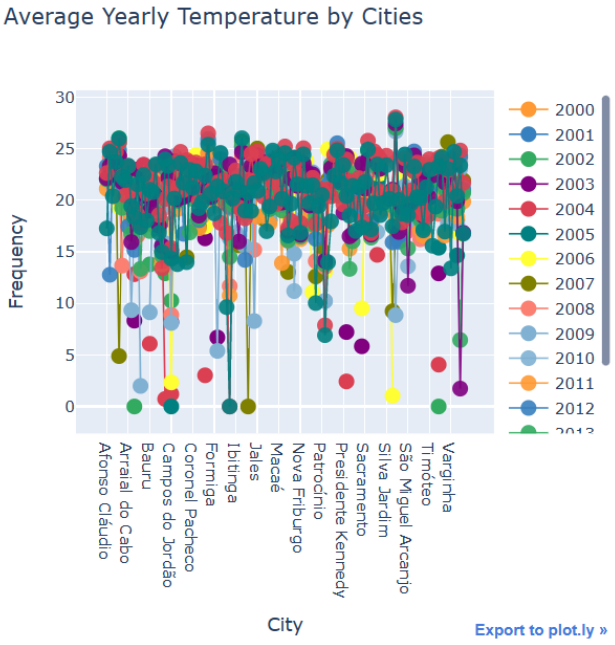
Province Espiritio is in the top spot on average weathwe factors



In 2011 the fluctuations recorded is less and the data is incomplete in early 2000 so it is not compared properly

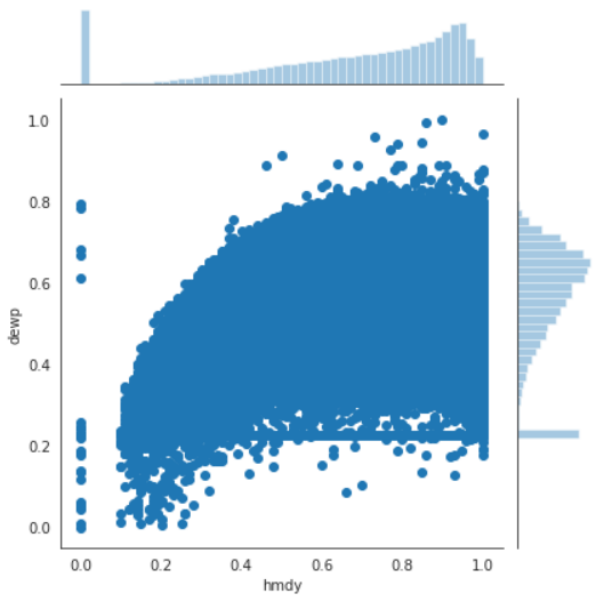


During the year of 2001 and 2006 the fluctations are notised and observed a constant sloppy decrease in temperature from 2004 to 2006

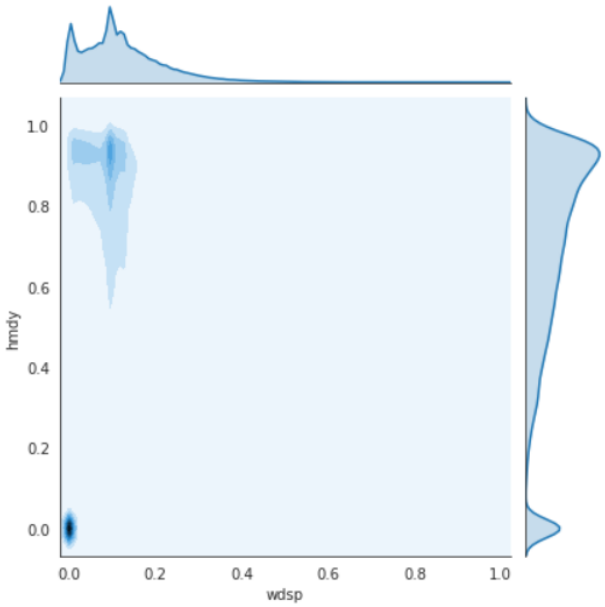


2013 – 16 is the period of most ups and downs and average temp is in the range of 20.5 to 16.5 from 2013 to 2014 and increased to 22 on average

DEW AND HUMIDITY



HUMIDITY AND WIND SPEED



V. CONCLUSION

We analysed the weather data and visualized using graphs to have patterns.Which helps to predict the weather using the patterns.We have done with different correlation between attributes in given data to classify the effects of one to the other.Python is much better to analyse the data visually and represents it in a systematically way. Having number of open source modules makes the analysis ease.Matlibplot, seaborn, plotly, clufflinks are the visualization libraries.Sklearn is to analyse the data.

VI. INFERENCES

[1] Lam, Joseph C., C. L. Tsang, L. Yang, and Danny HW Li. "Weather data analysis and design implications for different climatic zones in China." *Building and Environment* 40, no. 2 (2005): 277-296.

[2] Basak, Jayanta, Anant Sudarshan, Deepak Trivedi, and M. S. Santhanam. "Weather data mining using independent component analysis." *The Journal of Machine Learning Research* 5 (2004): 239-253.

[3] Leverich and C. Kozyrakis, “On the energy (in) efficiency of hadoop clusters,” ACM SIGOPS Operating Systems Rev., vol. 44, issue 1, pp. 61–65, 2010

[4] Riyaz P.A., Surekha Mariam Varghese, “Leveraging Map Reduce With Hadoop for Weather Data Analytics” IOSR Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 3, Ver. II (May – Jun. 2015), PP. 06- 12

[5] Basvanth Reddy and Prof B. A. Patil, “Weather Prediction on Big Data Using Hadoop Map Reduce Technique”, IJARCCE, ISSN: 2278-1021 Volume-05, Issue-06, Page No (643-647), June, 2016

[6] Ye Ding, Yanhua Li, “Detecting and Analyzing Urban Regions with High Impact of Weather Change on Transport”, IEEE Transactions on Big Data -2016

[7] Mr. Sunil Navadia, “Weather Prediction: A novel approach for measuring and analyzing weather data”, IEEE International conference on ISMAC-2017

[8] E Sreehari, J. Velmurugan and Dr. M. Venkatesan, “A Survey Paper on Climate Changes Prediction Using Data Mining”, IJARCCE, ISSN: 2278-1021 Volume05, Issue-02, Page No (294296), February, 2016

[9] “Hadoop-based ARIMA Algorithm and its Application in Weather Forecasting”, Authors: Leixiao Li, Zhiqiang Ma, Limin Liu, Yuhong Fan, International Journal of Database Theory and Application Vol.6, No.5 (2013), pp.119-132.