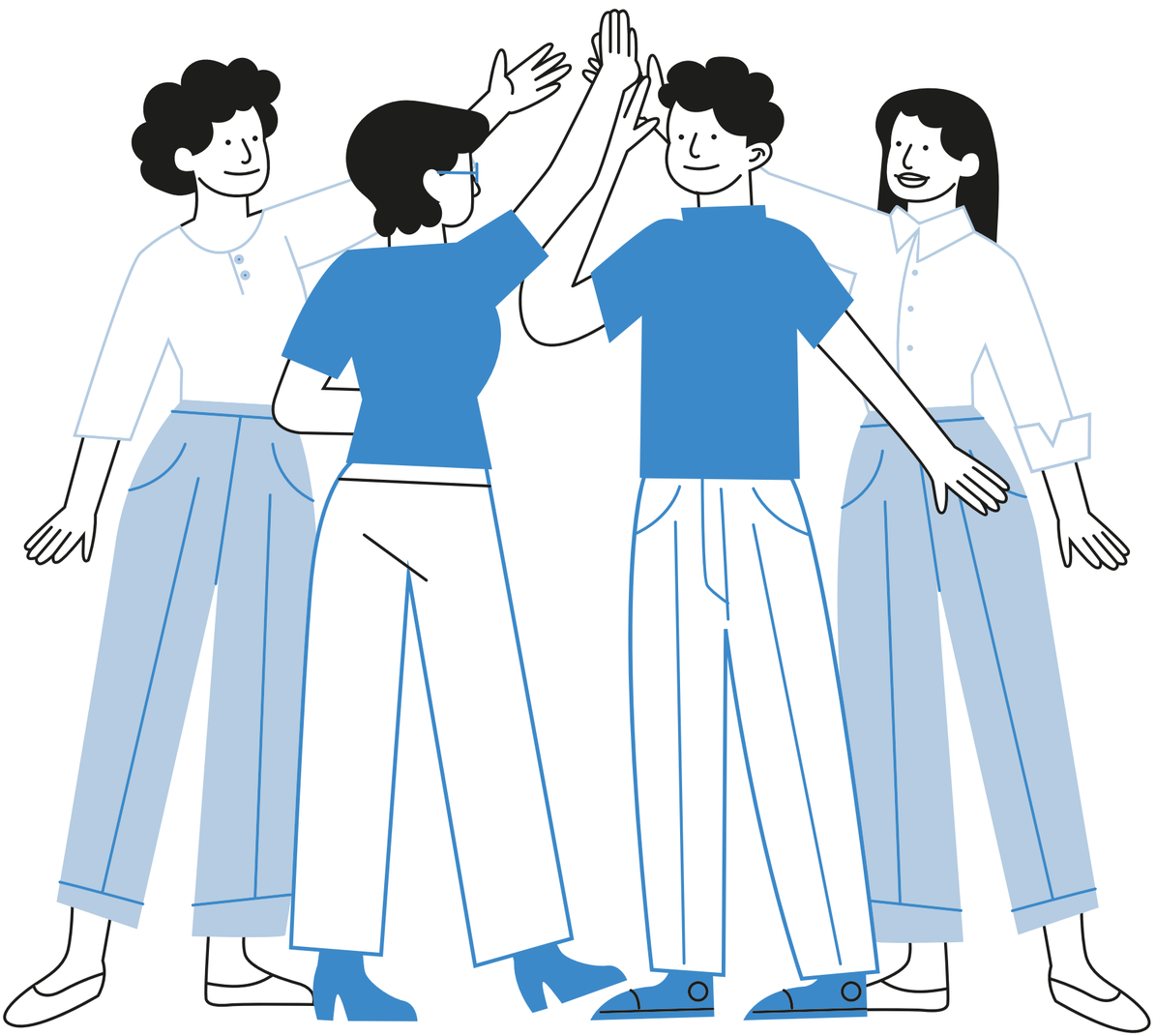
Group 07

Weather Forecast Project

Presentated by Group 07



Group’s Infomation

Bùi Nguyên Hanh - 21127606

Nguyễn Cao Sơn - 21127159

Cao Nguyễn Khánh - 21127627

Nguyễn Minh Hiếu - 21127742

Introduction

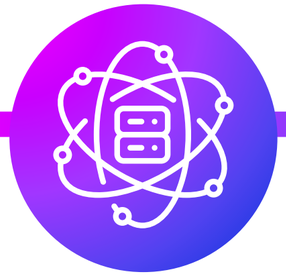
Analise of weather data in Ho Chi Minh City

Gain a comprehensive understanding of significant weather changes, including temperature fluctuations and humidity.

Apply data analysis methods to predict future weather conditions

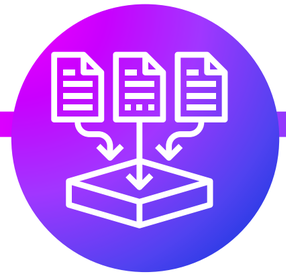
Assist in making decisions related to agriculture, transportation, ...

Implementation steps



Crawl data

We use the API of the website https://openweathermap.org/ to crawl data, this way we will extract data fields through json file.



Preprocess data

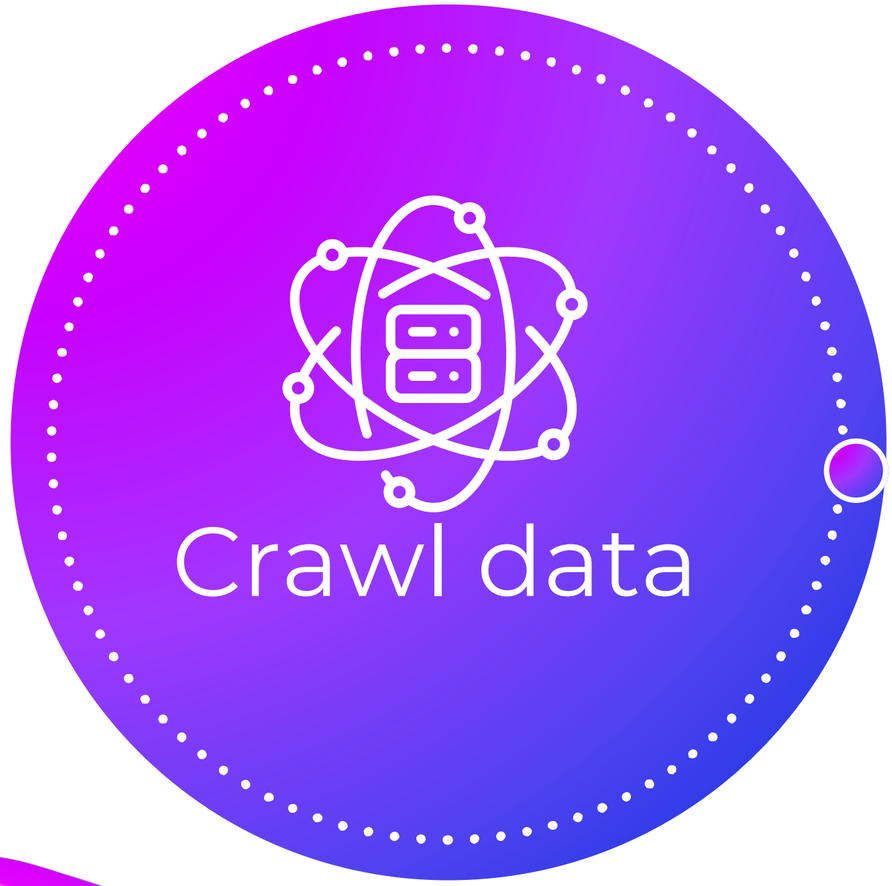
In the data preprocessing step, we will briefly filter through the data (delete error lines, duplicate rows; remove unnecessary columns...)

Exploring data

We dissect the data in more detail (what type of data is it, is that type reasonable or not, ..) and ask some questions to understand more about the data

Modeling

We use a number of machine learning methods to train the model so it can predict future weather.



#

Crawling data using web API

Crawl data

Data is crawled from:

https://openweathermap.org/.

LIMITED

Problem : Limited data crawling access.

Because our OpenWeatherMap account is designated as a student account, we can only collect data within the past year up to the crawl start date.

API

Collect Collect data using API.

In order to collect data for each weather index at 1 location, you can use a URL with the following format:

Collect data using API

Format:

https://history.openweathermap.org/data/2.5/history/city?id={**id**}&type=hour&start={**dt**}&appid={**API_key**}

Key information to note when requesting this API.

- **id**: City ID. (ID of Ho Chi Minh City = “1566083”)
- **dt**: is a timestamp counted from the Epoch (usually 1/1/1970).
-**API_key**: Our API key is “626e8ec21c8de03a592d15a0f2dca7f9”.

Example:

https://history.openweathermap.org/data/2.5/history/city?id=**1566083**&type=hour&start=**1685811600**&appid=**626e8ec21c8de03a592d15a0f2dca7f9**

Response from the website

When requesting the URL, you will receive a JSON-formatted response containing information about the weather at the specified timestamp (dt) and the next 23 hours. (total 24 hours).

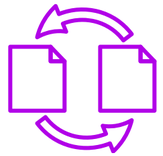
Response form

{
  "message": "Count: 24",
  "cod": "200",
  "city_id": 1566083,
  "calctime": 0.026848065,
  "cnt": 24,
  "**list**": [{
      "**dt**": 1685811600,
      "**main**": { "**temp**": 302.16, "**feels_like**": 309.16, "**pressure**": 1009, "**humidity**": 89, "**temp_min**": 302.16, "**temp_max**": 302.16},
      "**wind**": {"**speed**": 5.14, "**deg**": 240 },
      "**clouds**": { "**all**": 40 },
      "**weather**": [{ "**id**": 802, "**main**": "Clouds", "**description**": "scattered clouds", "**icon**": "03n"}]
    }, ... (There are 23 similar "list")]
}

Preprocess data

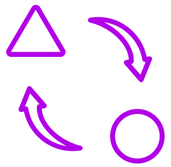
Prune out unnecessary columns, rows, and other things to eliminate missing or noisy data

Preprocess data



Checking shape and duplicate

- Ensuring Data Consistency
- Reduce the data size



Converting inappropriate data types to appropriate types

- Easier to calculate
- Compatibility with Algorithms

Dropping all columns that are irrelevant or too many null values

- Reduce the data size
- Maintaining the accuracy of the model

01

Basic level: Provide some basic information about the data (number of rows, columns, distribution...)

02

Ask and answer some questions related to data to get a more detailed view of the data

Objectives

EDA is the first door that opens up a deep understanding of data and helps create a solid foundation for future data analysis and processing decisions.

Exploring data in a meaningful way

Helps us better prepare for the following stages in the data analysis and processing process

Bring a visual perspective to the data with questions

Problems

Numerical analysis

Use the statistical function to list statistical parameters to make data processing more convenient

Attribute's imformation & description

Explains in detail the meaning of each attribute column, as well as indicating the type of each column and checking to see if any rows are null or not.

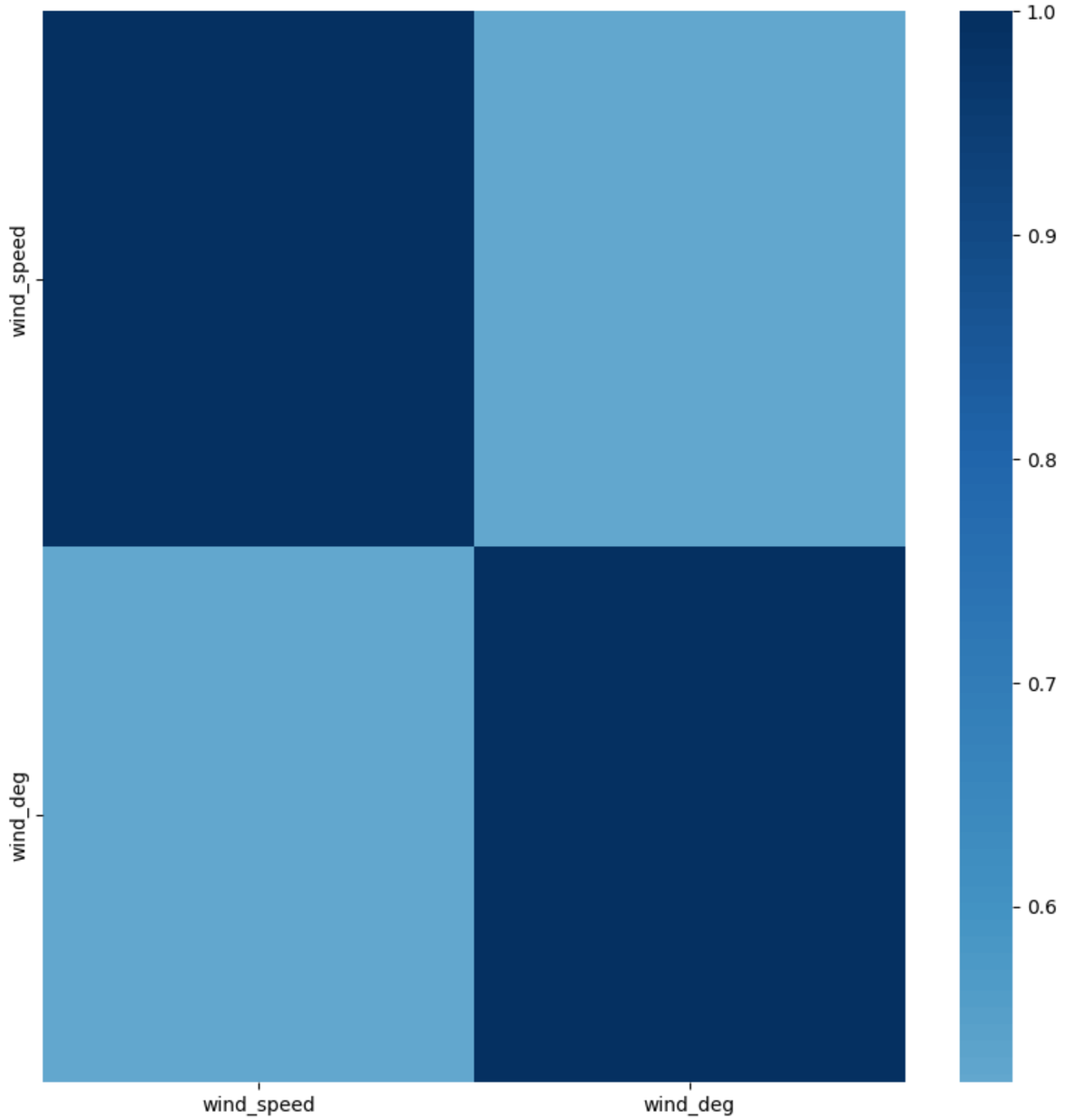
Answer the question

Ask and answer questions with visual templates to give a more intuitive view of the data

Answer some questions

Question 01:
How are wind speed and wind direction related, and what can be drawn from it?

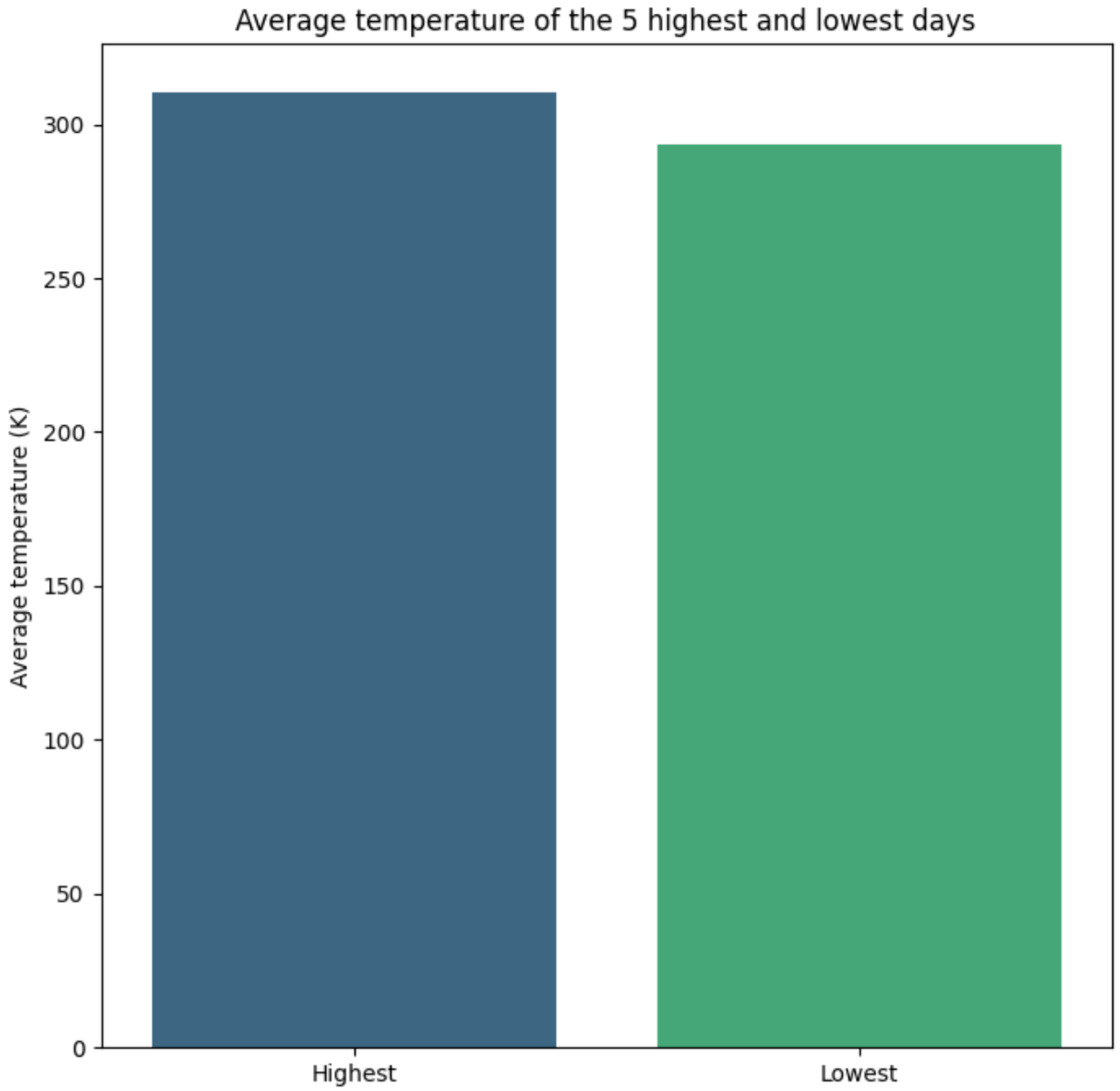
Answer:
We can conclude that for jobs that require wind speed, we should choose a suitable wind direction.



Answer some questions

Question 02:
Compare the temperature of the 5 days with the highest temperature and the 5 days with the lowest temperature ?

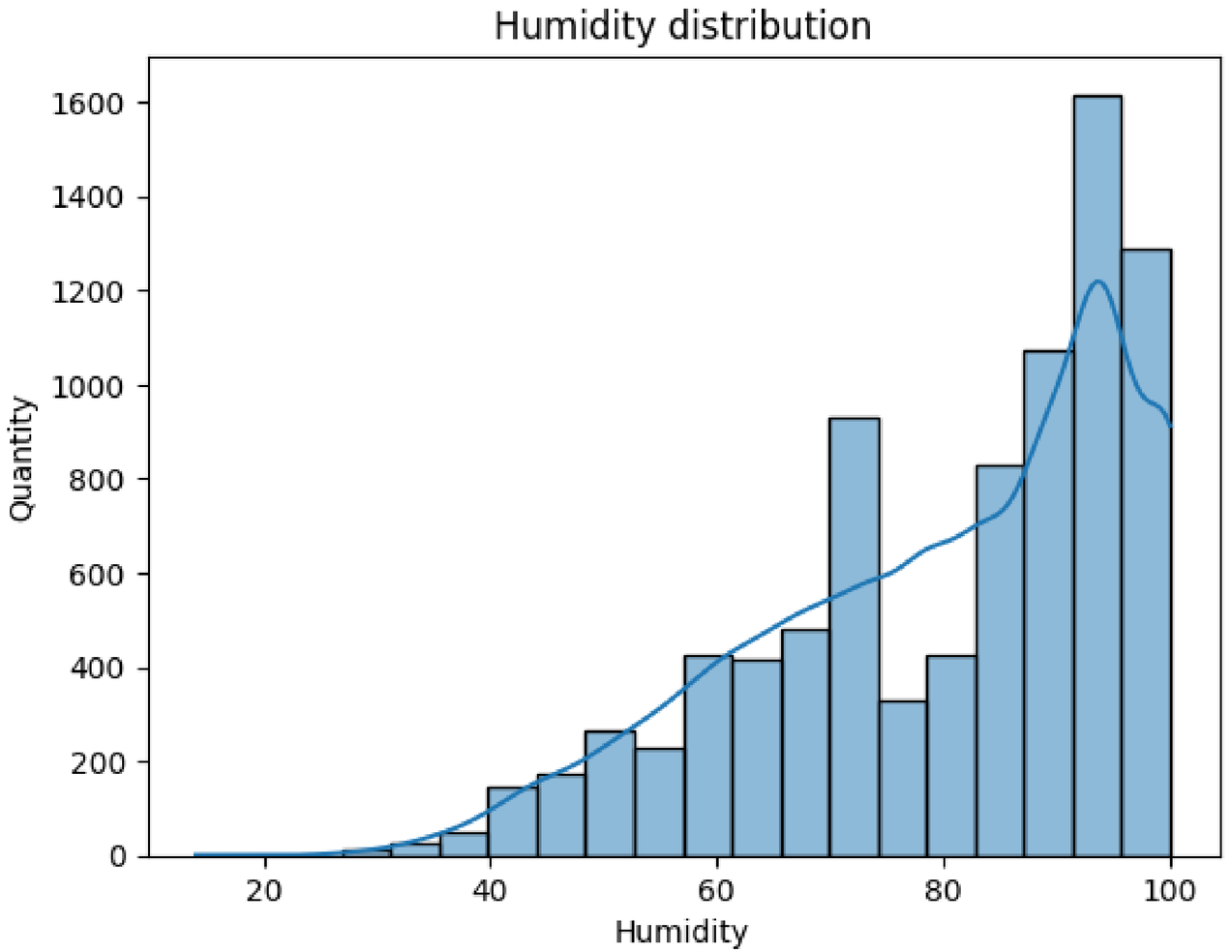
Answer:
Avg Highest temperature: 37.3 ℃
Avg Lowest temperature: 20 ℃



Answer some questions

Question 03:
Provides information about the distribution of humidity ?

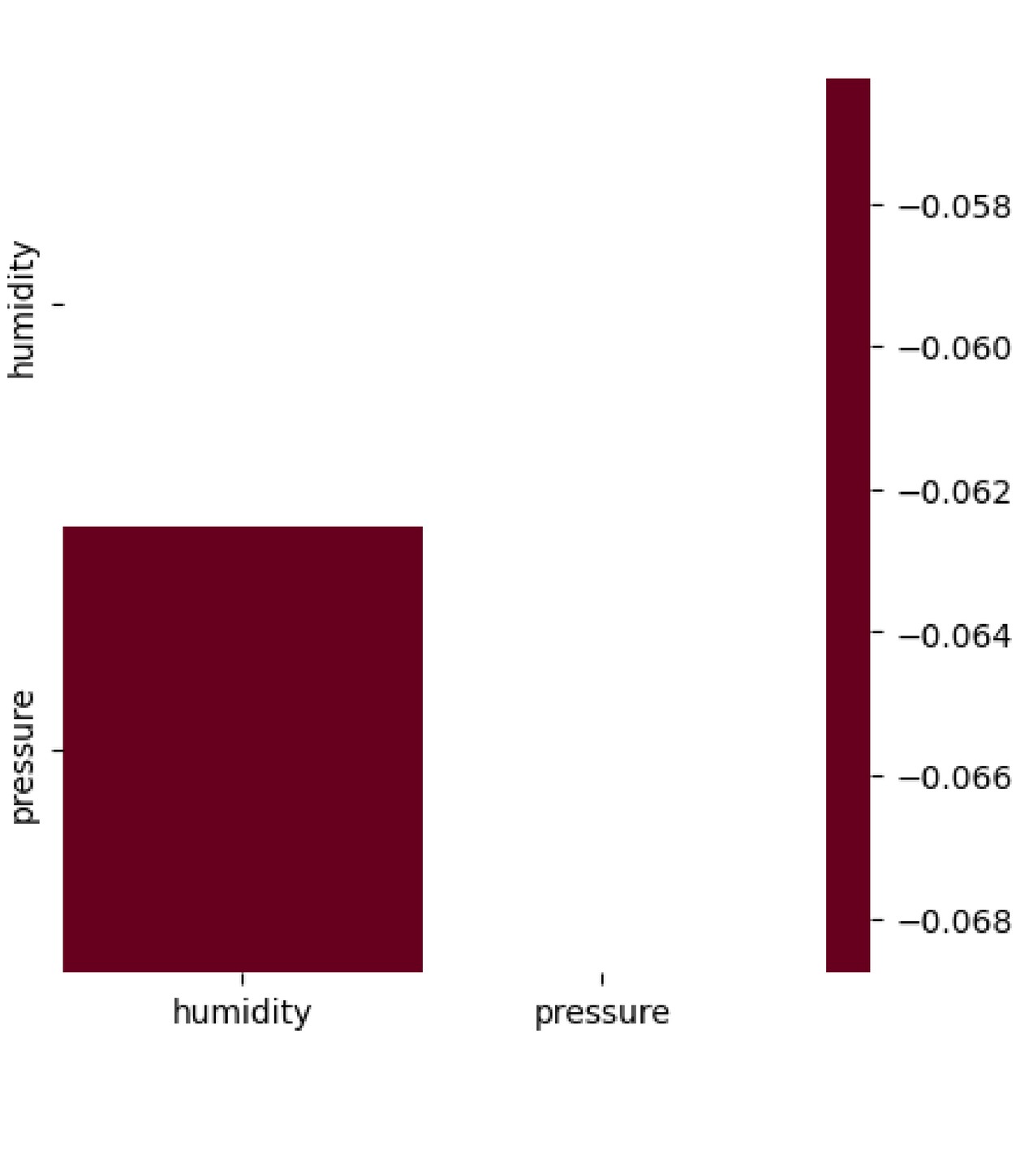
Answer:
As we can see in the histogram, the humidity is in the very high range of 80 - 100, this partly helps us be proactive in protecting our health, and at the same time shows us the typical climate in Ho Chi Minh City.



Answer some questions

Question 04:
Is there a relationship between humidity and air pressure?

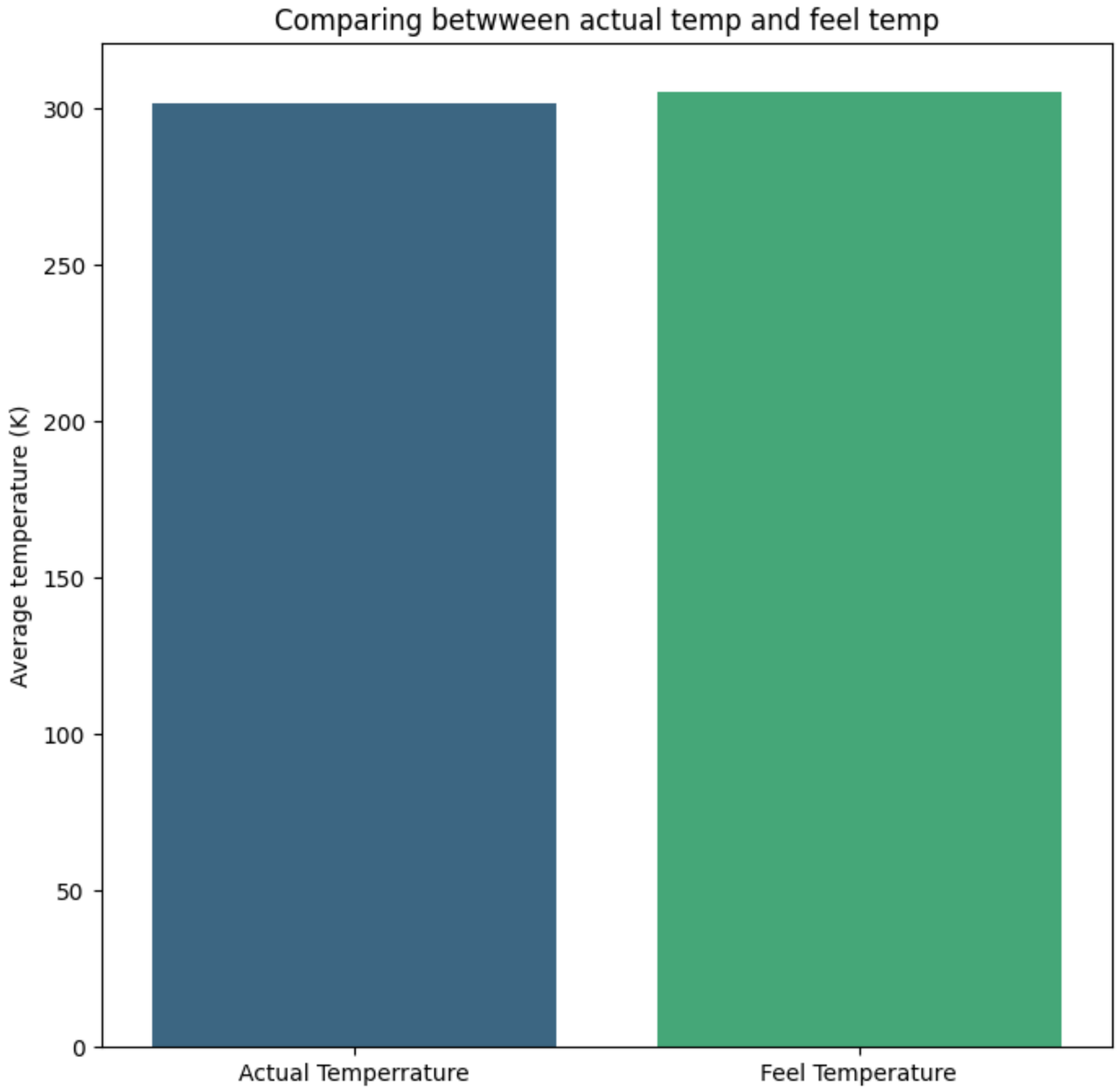
Answer: As we can see in the heatmap, these two attributes have approximately zero correlation, so we cannot use these two attributes as a measure or a tool for research.

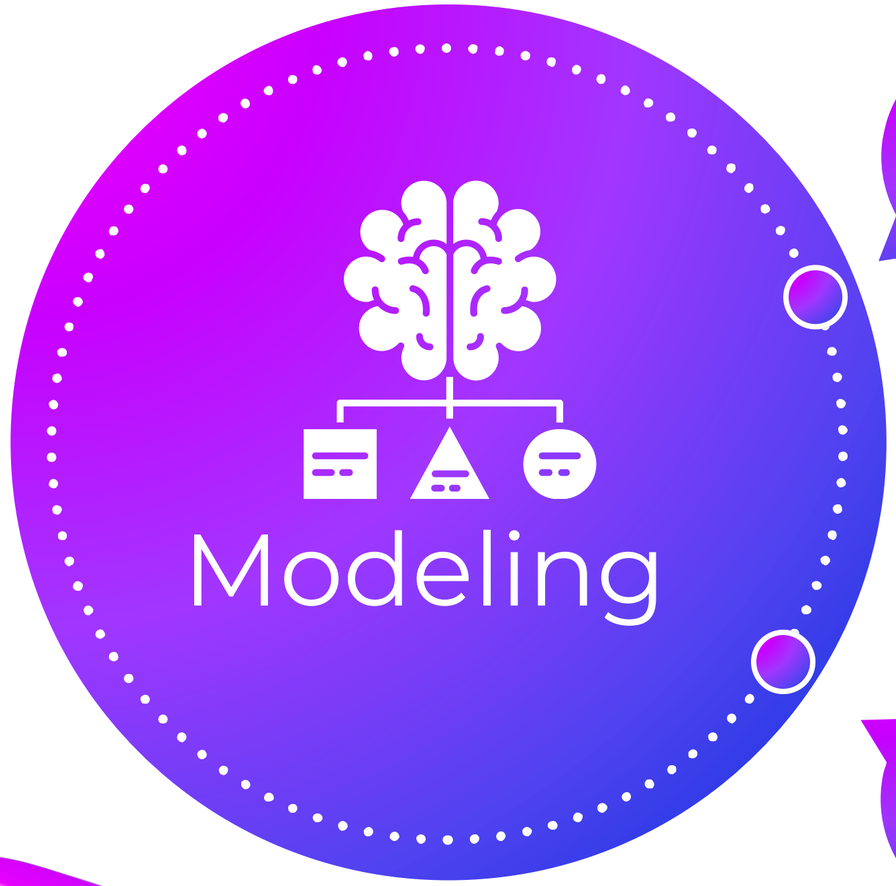


Answer some questions

Question 05:
Is there a difference between the actual temperature and the feeling?

Answer: As you can see in the bar chart, there’s a nimor different between the actual temperature and feel temperature, this difference can be considered non-existent.





01

Linear Approach: In Modeling session, we had been taught that all the problems we encounter, we should try the linear stuffs first although we probability that we can solve the problem is quite low, but we should try it first.

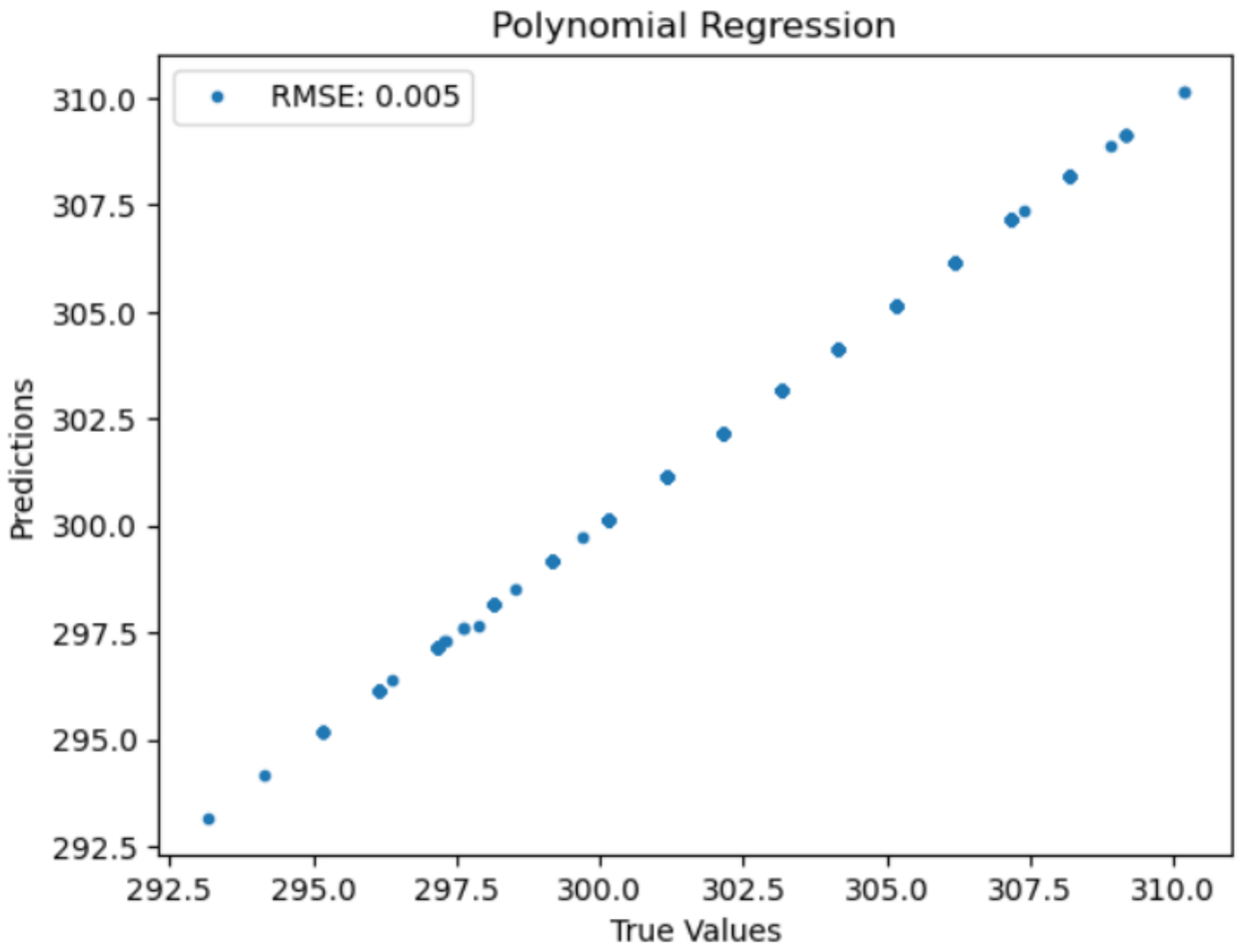
02

The ARIMA (Autoregressive Integrated Moving Average) model is a popular time series forecasting model in the field of forecasting and analyzing time series data. ARIMA combines autoregressive (AR), integrated (I), and moving average (MA) components to model variations and trends in time series data.

Polynomial Regression

- Polynomial regression is a type of linear regression where the regression function is a polynomial of degree higher than 1 (in this case we choose 2).

- Advantages:
  - Can model nonlinear relationships between variables.
  - Can provide more accurate predictions than low-degree linear regression models.
- Disadvantages:
  - Can lead to overfitting if the degree of the regression function is too high.

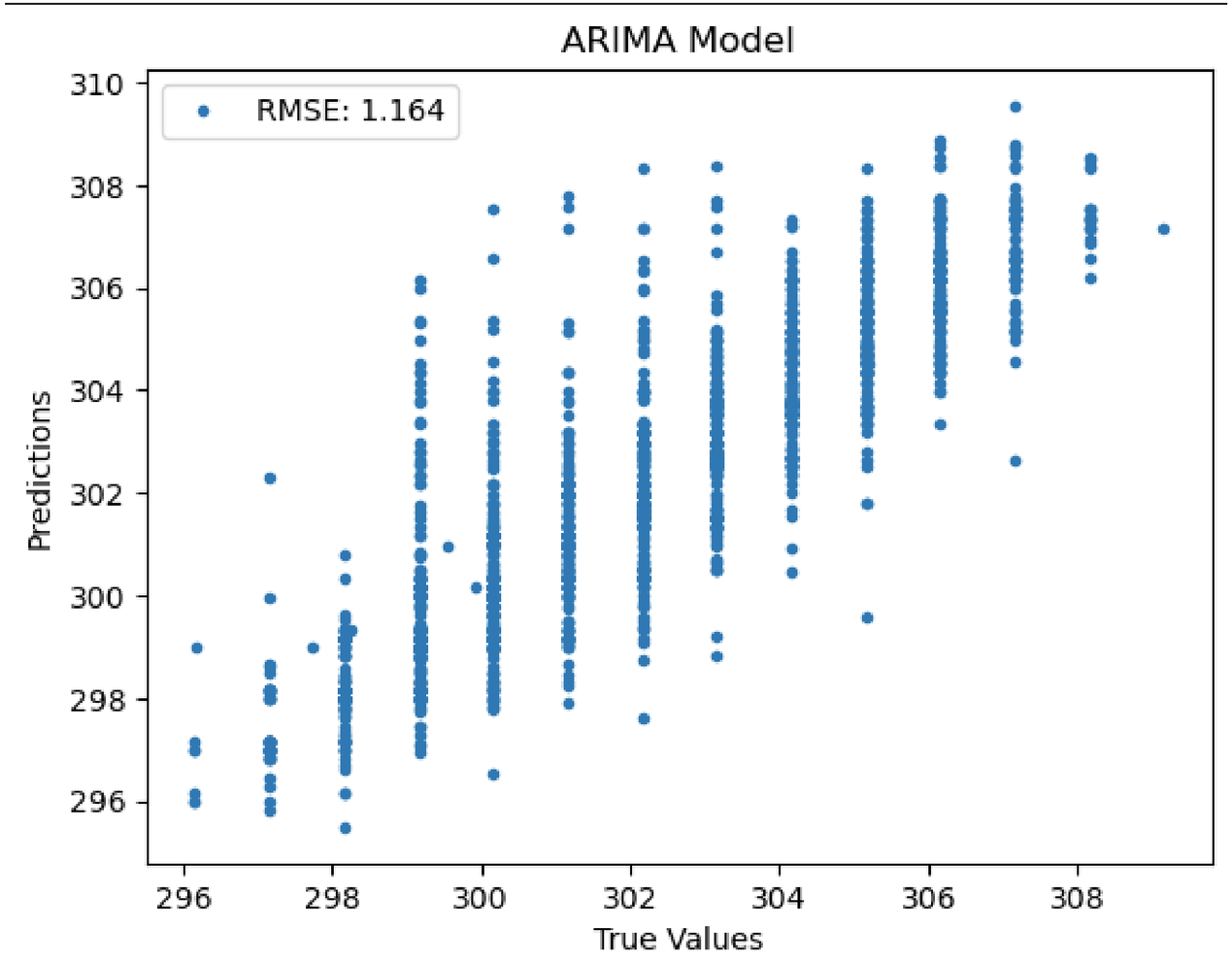


temp = 0.333 * feels_like + 0.333 * temp_max + 0.333 * wind_speed

ARIMA Model

- We use ARIMA(2,1,0): This parameter means that the model has two autoregressive terms, one integrated term, and no moving average terms.

- Advantages:
  - Can forecast non-linear time series.
  - Can forecast time series with trend and seasonality.
- Disadvantages:
  - Can lead to overfitting if the parameters are not chosen properly.



y(t) = 1.278 + 0.181 * y(t-1) + 0.175 * y(t-2) + ε(t)

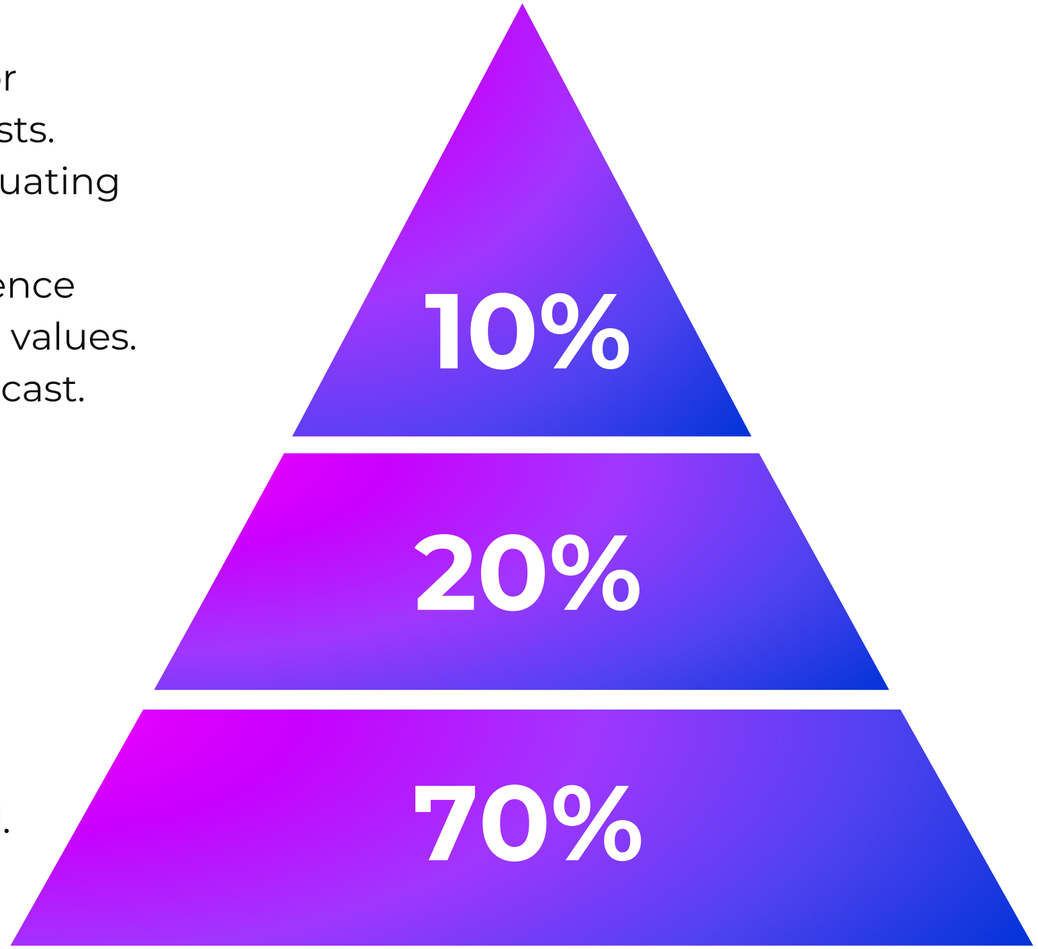
Metric: MRSE

- Root mean square error (RMSE) is a metric for evaluating the accuracy of time series forecasts.
- Mean squared error (MSE) is a metric for evaluating the accuracy of time series forecasts.
- It is calculated as the average squared difference between the predicted values and the actual values.
- A lower RMSE indicates a more accurate forecast.

POLYNOMIAL **0.005**

ARIMA **1.164**

Then we choose Polynomial regression model.



Thanks for listening

Raise some questions if you concern about our topic