Weather Classification Model

Introduction to Data Science

Tran Dinh Quang	21127406
Nguyen Hong Hanh	21127503
Do Quoc Tri	21127556
Nguyen Khanh Nhan	21127657

Introduction

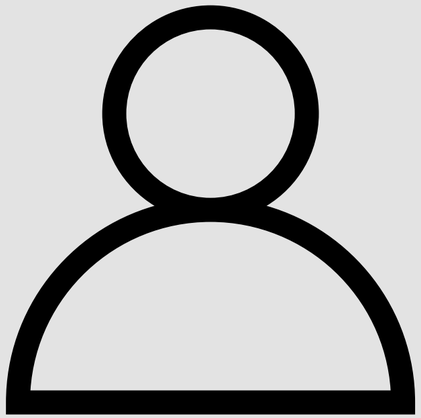
Weather Classification Model

Ho Chi Minh City is one of the largest cities in Vietnam, with a high population density and numerous industrial and urban areas. The weather in this city is not overly complex, mainly consisting of two seasons: sunny and rainy. However, due to its dense and developed nature, weather forecasting and classification are necessary for the residents to better prepare for their activities. The model will utilize the surrounding environmental conditions to classify and predict weather, providing useful information for forecasting applications, analysis, risk assessment, and optimizing operations in the field of weather and environment

Member Roles

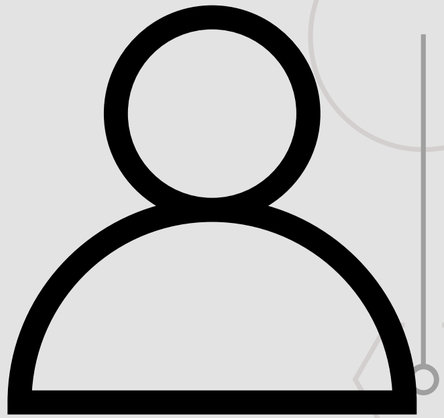
Tran Dinh Quang

Data Collecting
Data Exploration



Nguyen Hong Hanh

Data preprocessing
Make presentation slides



Do Quoc Tri

Choosing topic
Data Exploration

Nguyen Khanh Nhan

Data Modeling
Check and summarize members' work

Stages

1

Data Colecting

2

Data Exploration

3

Data Preprocessing

4

Data Modeling

Data Collecting

visualcrossing

Visual Crossing is a website that provides users with easy, quick, and efficient weather search tools, serving as a crucial resource for the community interested in weather and climate.

Data Collecting

Location


```
{'latitude': 10.7764,
 'longitude': 106.701,
 'resolvedAddress': 'Quận 1, Hồ Chí Minh, Việt Nam',
 'address': 'Ho Chi Minh',
 'timezone': 'Asia/Ho_Chi_Minh',
 'tzoffset': 7.0}
```

Data Collecting

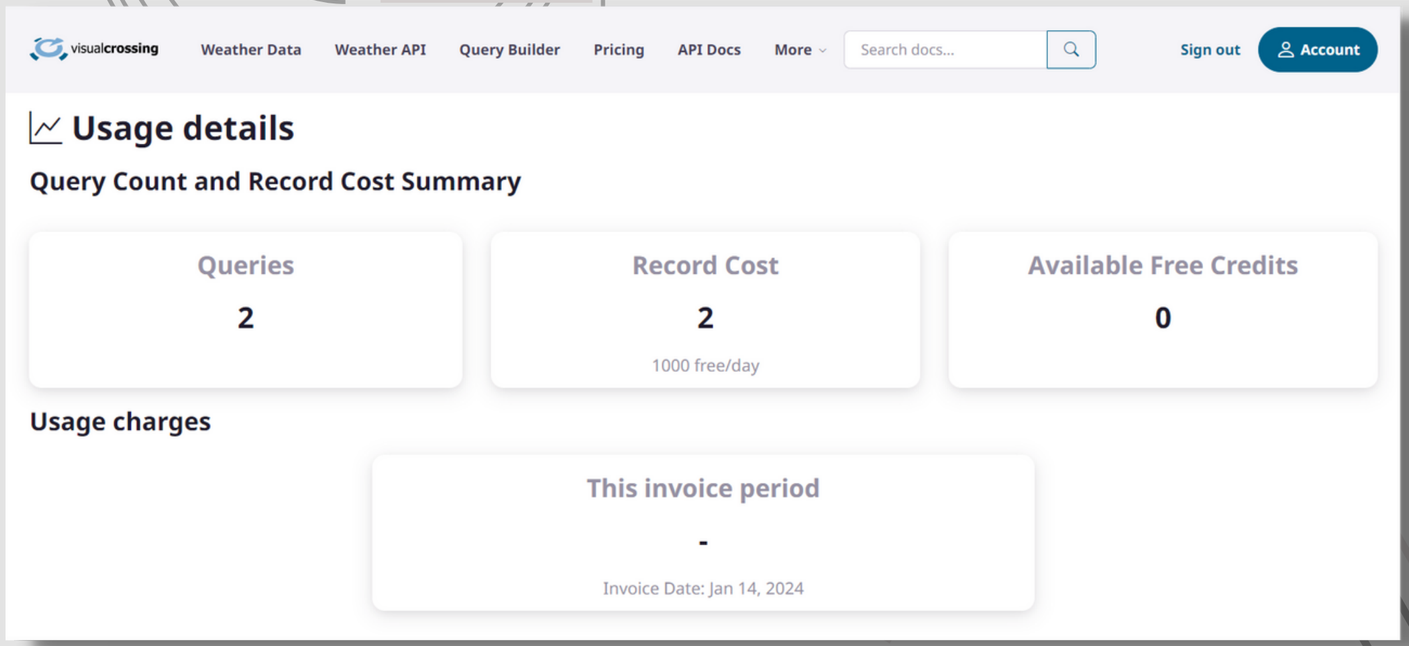
Query


```
{"queryCost":1,"latitude":10.7764,"longitude":106.701,"resolvedAddress":"Quận 1, Hồ Chí Minh, Việt
Nam","address":"Ho Chi Minh","timezone":"Asia/Ho_Chi_Minh","tzoffset":7.0,"description":"Similar temperatures
continuing with a chance of rain Sunday & Monday.","days":[{"datetime":"2023-12-
14","datetimeEpoch":1702486800,"tempmax":35.4,"tempmin":25.1,"temp":29.0,"feelslikemax":37.2,"feelslikemin":25.1,"f
eelslike":30.5,"dew":21.1,"humidity":64.7,"precip":0.0,"precipprob":0.0,"precipcover":0.0,"preciptype":null,"snow":
0.0,"snowdepth":0.0,"windgust":34.9,"windspeed":19.1,"winddir":117.0,"pressure":1011.0,"cloudcover":73.9,"visibilit
y":23.5,"solarradiation":246.6,"solarenergy":21.3,"uvindex":9.0,"severerisk":10.0,"sunrise":"06:02:38","sunriseEpoc
h":1702508558,"sunset":"17:32:37","sunsetEpoch":1702549957,"moonphase":0.04,"conditions":"Partially
cloudy","description":"Partly cloudy throughout the day.","icon":"partly-cloudy-day","stations":
["VVTS"],"source":"comb","hours":
[{"datetime":"00:00:00","datetimeEpoch":1702486800,"temp":29.0,"feelslike":34.4,"humidity":79.05,"dew":25.0,"precip
":0.0,"precipprob":0.0,"snow":0.0,"snowdepth":0.0,"preciptype":null,"windgust":16.2,"windspeed":9.4,"winddir":130.0
,"pressure":1013.0,"visibility":10.0,"cloudcover":50.0,"solarradiation":0.0,"solarenergy":0.0,"uvindex":0.0,"severe
risk":10.0,"conditions":"Partially cloudy","icon":"partly-cloudy-night","stations":["VVTS"],"source":"obs"},
{"datetime":"01:00:00","datetimeEpoch":1702490400,"temp":25.9,"feelslike":25.9,"humidity":75.78,"dew":21.3,"precip"
:0.0,"precipprob":0.0,"snow":0.0,"snowdepth":0.0,"preciptype":null,"windgust":16.6,"windspeed":7.9,"winddir":57.9,"
pressure":1011.0,"visibility":24.1,"cloudcover":73.1,"solarradiation":0.0,"solarenergy":0.0,"uvindex":0.0,"severeri
sk":10.0,"conditions":"Partially cloudy","icon":"partly-cloudy-night","stations":null,"source":"fcst"},
{"datetime":"02:00:00","datetimeEpoch":1702494000,"temp":25.7,"feelslike":25.7,"humidity":77.16,"dew":21.4,"precip"
:0.0,"precipprob":0.0,"snow":0.0,"snowdepth":0.0,"preciptype":null,"windgust":12.6,"windspeed":6.8,"winddir":46.3,"
pressure":1011.0,"visibility":24.1,"cloudcover":80.0,"solarradiation":0.0,"solarenergy":0.0,"uvindex":0.0,"severeri
sk":10.0,"conditions":"Partially cloudy","icon":"partly-cloudy-night","stations":null,"source":"fcst"},
```

Data Collecting

visualcrossing

Each free account is entitled to retrieve 1000 data queries from the website in a single day.



Data Collecting

Save data

Data Exploration

Location	Value
Latitude	10.7764
Longitude	106.701
Address	Quan 1, Ho Chi Minh, Viet Nam, Asia
Timezone	Asia/Ho Chi Minh city
Date start	27/9/2009
Date end	5/6/2023

Data Context

Data Exploration

Data Context

Column	Meaning
`datetime`	ISO 8601 formatted date, time, or datetime value indicating the date and time of the weather data in local time.
`datetimeEpoch`	Number of seconds since 1st January 1970 in UTC time.

Data Exploration

Data Context

Column	Meaning
`tempmax`	Maximum temperature at the location (°F).
`tempmin`	Minimum temperature at the location (°F).
`temp`	Average temperature at the location during the day (°F).
`feelslikemax`	Maximum feels-like temperature at the location.
`feelslikemin`	Minimum feels-like temperature at the location.
`feelslike`	Average feels-like temperature, accounting for heat index or wind chill. Daily values are mean values for the day.
`dew`	Dew point temperature.
`humidity`	Relative humidity in %.

Data Exploration

Data Context

Column	Meaning
`precip`	Amount of liquid precipitation (mm).
`precipprob`	Likelihood of measurable precipitation (0-100%).
`precipcover`	Proportion of hours with non-zero precipitation in one day.
`preciptype`	Types of precipitation expected or occurred (rain, snow, freezing rain, ice).
`snow`	Amount of snowfall (mm).
`snowdepth`	Depth of snow on the ground.
`windgust`	Instantaneous wind speed at a location. Daily values are the maximum hourly value for the day.
`windspeed`	Sustained wind speed measured as the average windspeed over the preceding one to two minutes.
`winddir`	Direction from which the wind is blowing.
`pressure`	Sea level atmospheric or barometric pressure in millibars (or hectopascals).
`cloudcover`	Percentage of sky covered in clouds (0-100%).
`visibility`	Visibility distance of distant objects.

Data Exploration

Data Context

Column	Meaning
`solarradiation`	Solar radiation power at the instantaneous moment of observation or forecast prediction (W/m^2).
`solarenergy`	Total energy from the sun that builds up over an hour or day (MJ/m^2).
`uvindex`	Level of ultraviolet (UV) exposure on a scale of 0 to 10.
`sunrise`	Formatted time of sunrise.
`sunriseEpoch`	Sunrise time specified as the number of seconds since 1st January 1970 in UTC time.
`sunset`	Formatted time of sunset.
`sunsetEpoch`	Sunset time specified as the number of seconds since 1st January 1970 in UTC time.
`moonphase`	Fractional portion through the current moon lunation cycle (0 to 1).

Data Exploration

Data Context

Column	Meaning
`conditions`	Textual representation of weather conditions.
`description`	Longer text descriptions suitable for weather displays.
`icon`	Machine-readable summary for displaying an icon.
`stations`	Weather stations used when collecting a historical observation record.
`source`	Type of weather data used for this weather object (obs, fcst, histfcst, stats, comb).
`severerisk`	Risk of convective storms on a scale of 0 to 100, indicating the severity of risk.

Data Exploration

Data Type

9 columns with **object** datatype
3 columns with **int64** datatype
24 columns with **float64** datatype

datetime	object
datetimeEpoch	int64
tempmax	float64
tempmin	float64
temp	float64
feelslikemax	float64
feelslikemin	float64
feelslike	float64
dew	float64
humidity	float64
precip	float64
precipprob	float64
precipcover	float64
preciptype	object
snow	float64
snowdepth	float64
windgust	float64
windspeed	float64

winddir	float64
pressure	float64
cloudcover	float64
visibility	float64
solarradiation	float64
solarenergy	float64
uvindex	float64
sunrise	object
sunriseEpoch	int64
sunset	object
sunsetEpoch	int64
moonphase	float64
conditions	object
description	object
icon	object
stations	object
source	object
severerisk	float64

Data Exploration

Data Type

datetime

object → datetime

sunrise - sunset

object → int64

datetime	object
datetimeEpoch	int64
tempmax	float64
tempmin	float64
temp	float64
feelslikemax	float64
feelslikemin	float64
feelslike	float64
dew	float64
humidity	float64
precip	float64
precipprob	float64
precipcover	float64
preciptype	object
snow	float64
snowdepth	float64
windgust	float64
windspeed	float64

winddir	float64
pressure	float64
cloudcover	float64
visibility	float64
solarradiation	float64
solarenergy	float64
uvindex	float64
sunrise	object
sunriseEpoch	int64
sunset	object
sunsetEpoch	int64
moonphase	float64
conditions	object
description	object
icon	object
stations	object
source	object
severerisk	float64

Data Exploration

Data Quality Check

Duplicate Rows

The data has 0 duplicate rows

```
# Check for duplicate rows
duplicates_series = raw_data.duplicated()

print(f'The data has {sum(duplicates_series)} duplicate rows.')
if sum(duplicates_series) > 0:
    print(f'The duplicate rows are: {[i for i, d in enumerate(duplicates_series) if d == True]}')
```

```
The data has 0 duplicate rows.
```

Data Exploration

Data Quality Check

Missing values

```
The number of rows missing 0 values: 369 rows (7%).
The number of rows missing 1 values: 143 rows (3%).
The number of rows missing 3 values: 356 rows (7%).
The number of rows missing 4 values: 2611 rows (52%).
The number of rows missing 5 values: 1424 rows (28%).
The number of rows missing 6 values: 1 rows (0%).
The number of rows missing 9 values: 96 rows (2%).

The number of rows with missing data: 4631 (92.62)%
```

Almost all rows have missing values

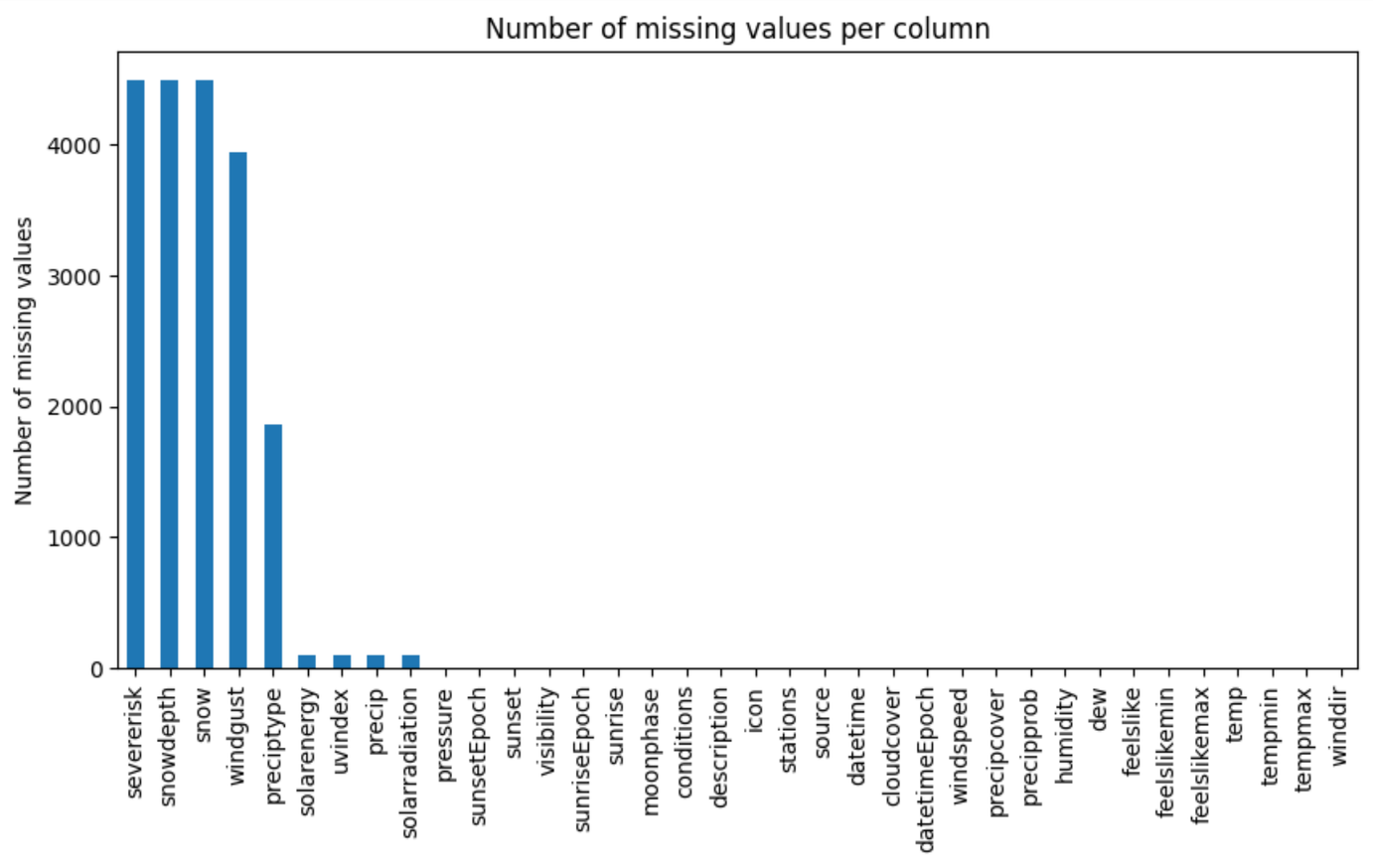
Almost rows have 4-5 missing values

The largest number of missing values in a row is 9 (1/4 number of columns)

Data Exploration

Data Quality Check

Missing values



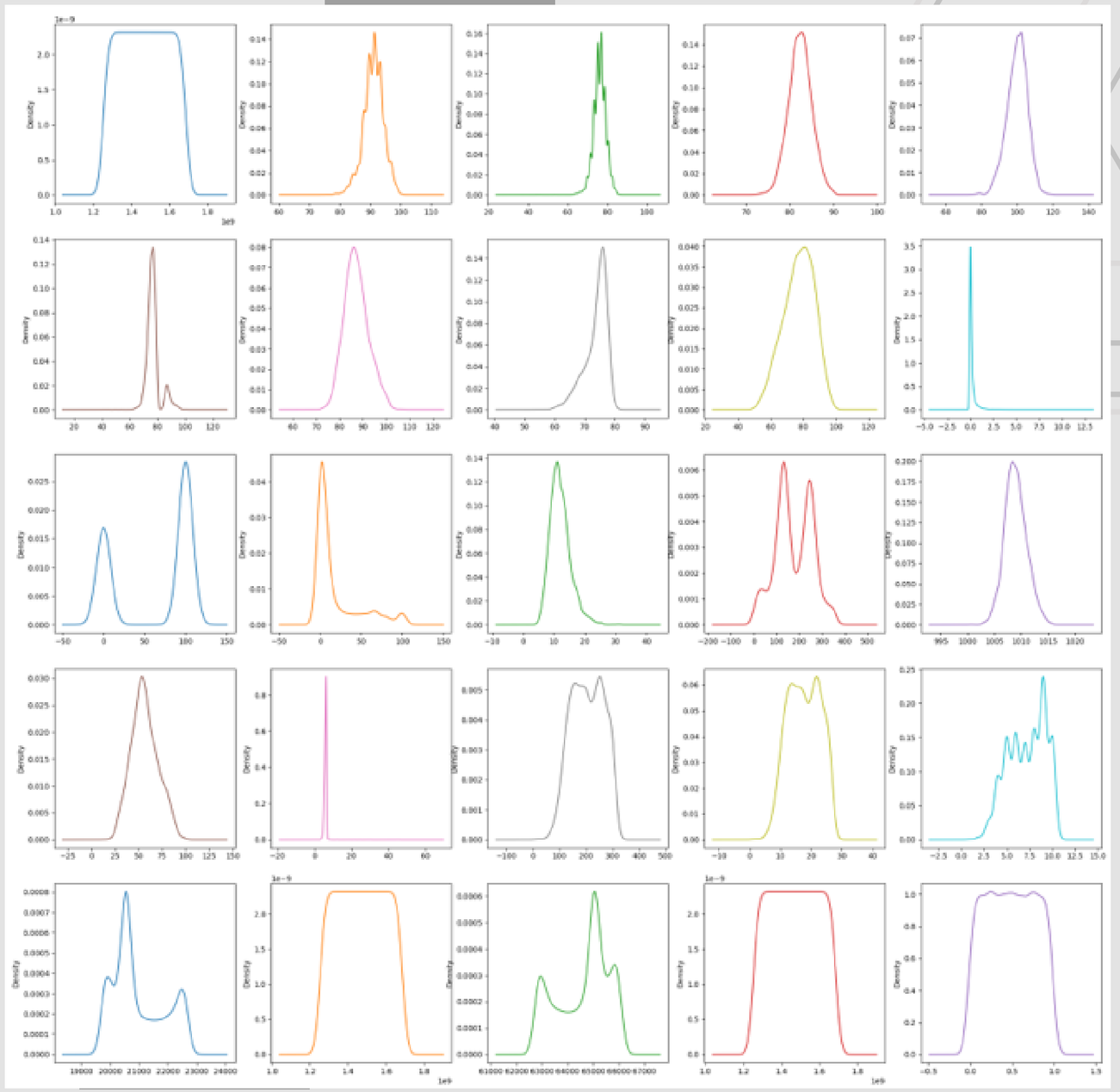
There are total 10 columns that have missing values

There are 5 columns that have a number of missing values exceeding the allowable threshold (33%):

- preciptype
- snowdepth
- snow
- windgust
- severerisk

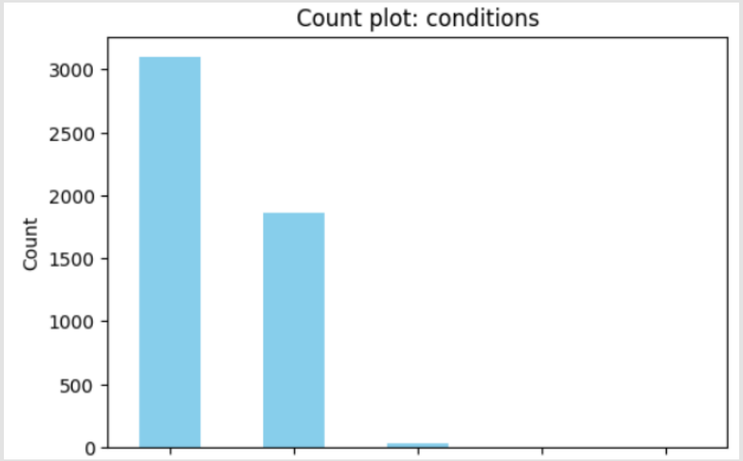
Data Exploration

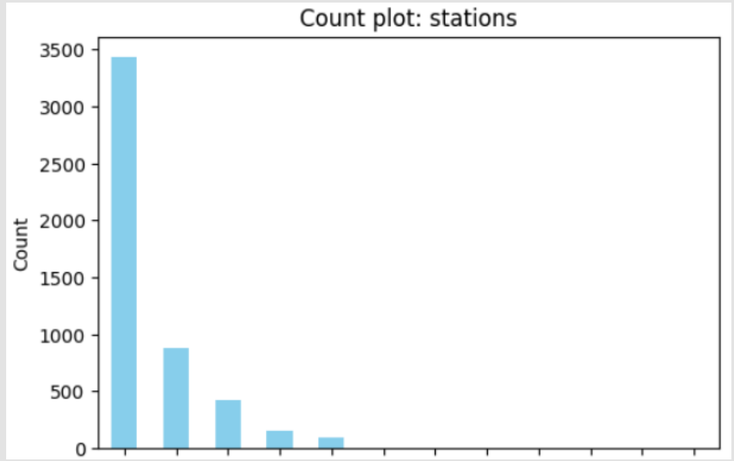
Distribution

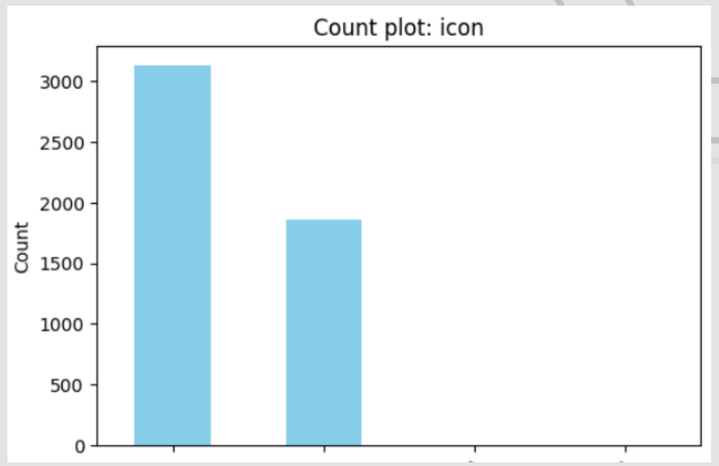


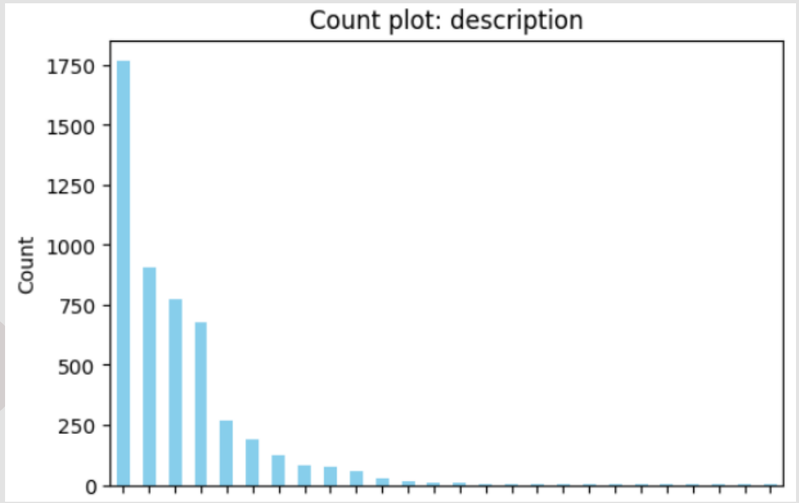
Data Exploration

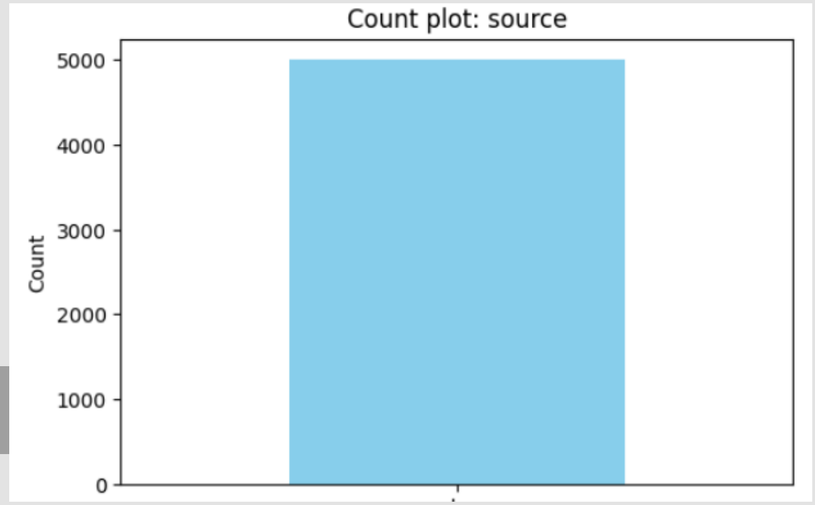
Data Distribution





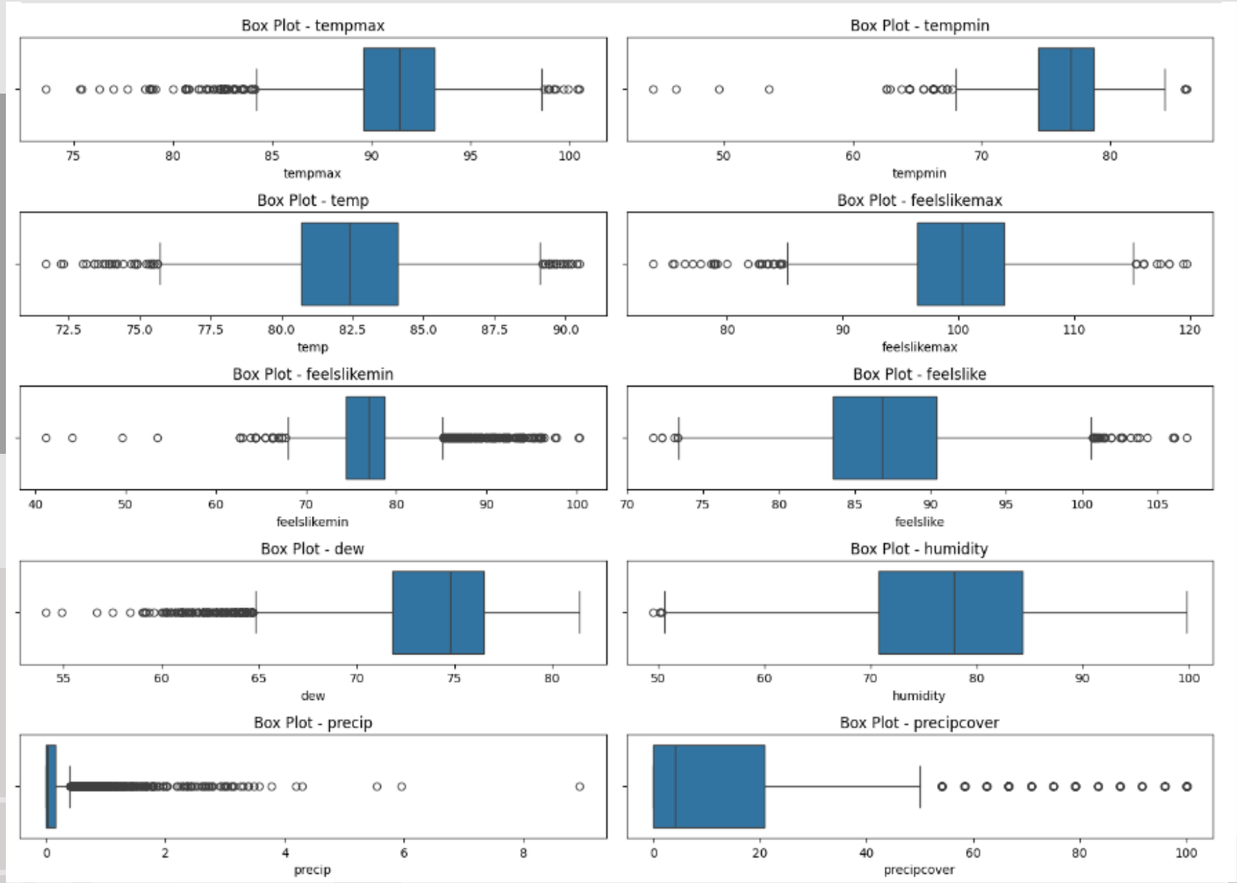


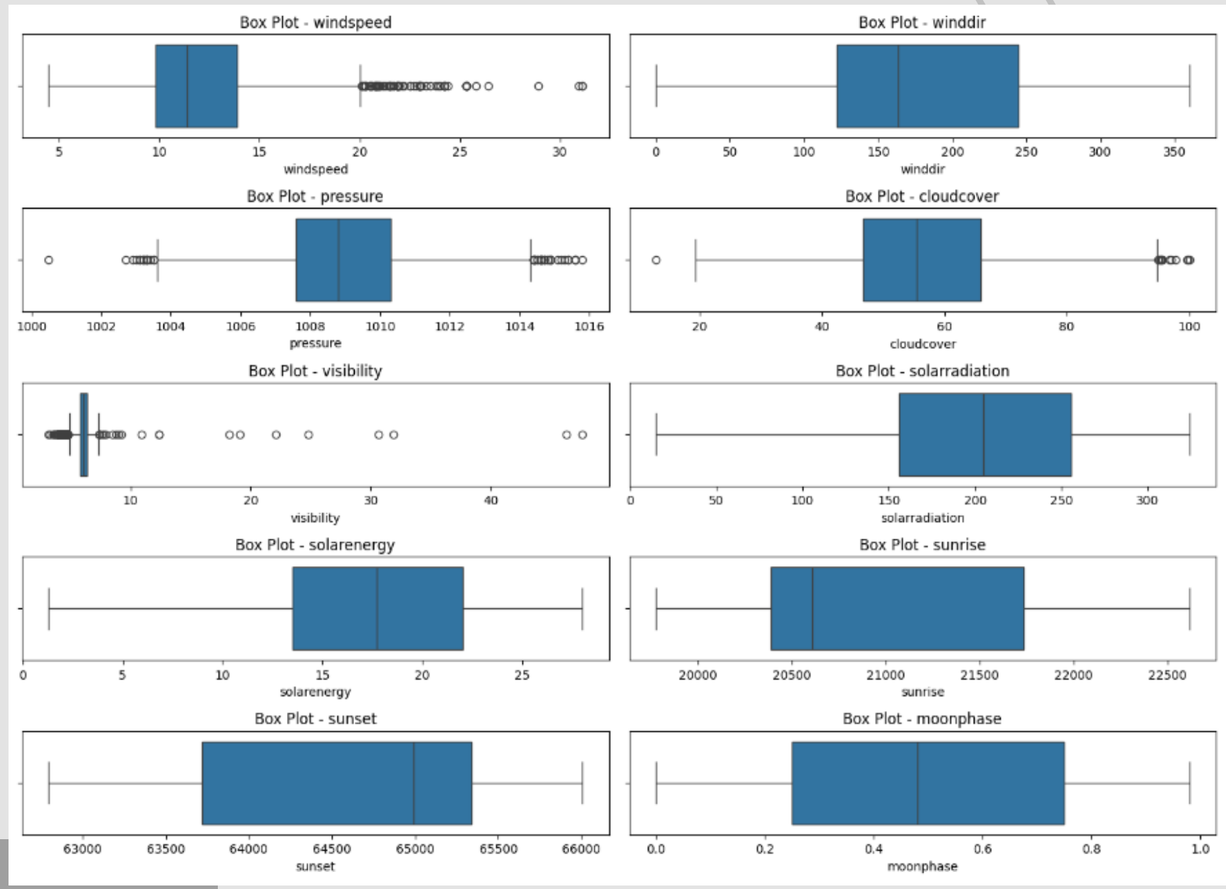




Data Exploration

Outliers



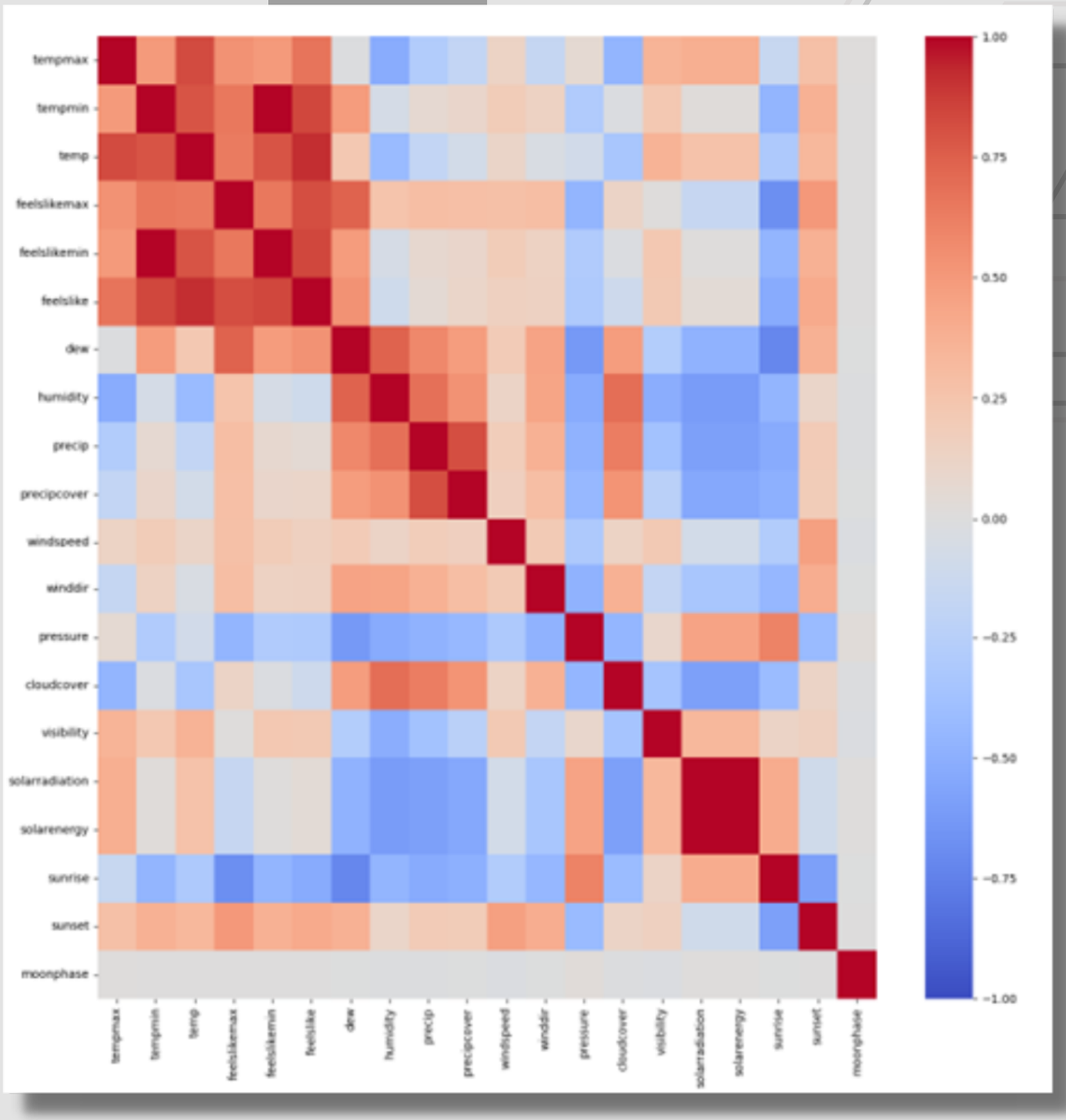


Data Exploration

Correlation

moonphase has weak correlation with the other attributes

tempmin, tempmax, temp, feelslikemax, feelslikemin, and feelslike have strong correlations with each other



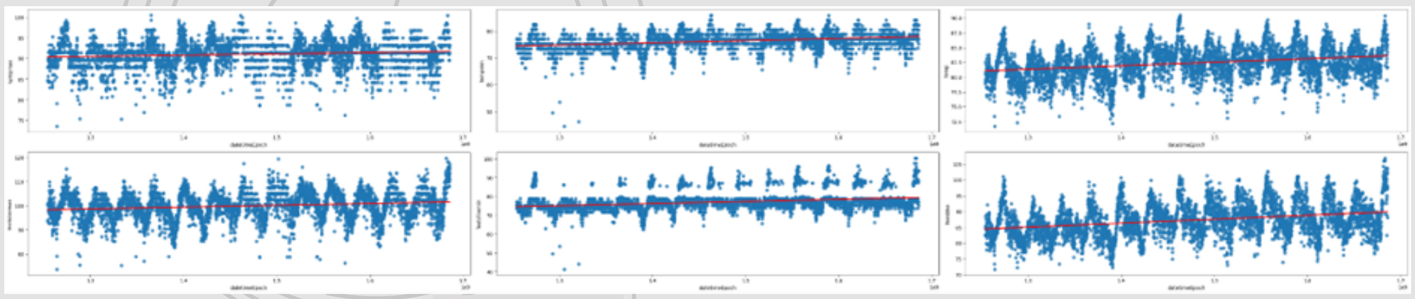
Data Exploration

Make & Answer questions

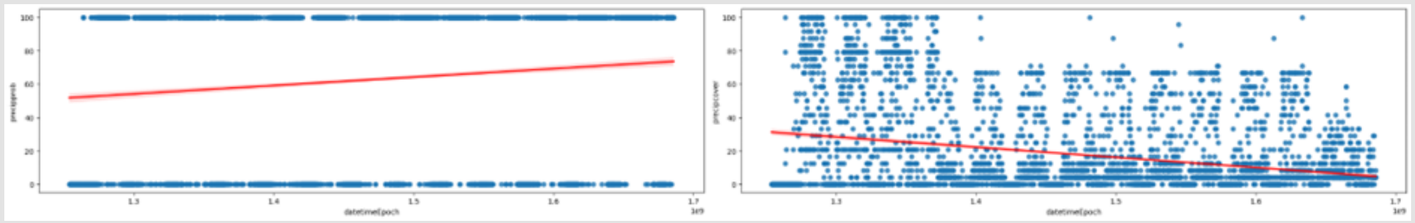
Question 1: What is the trend of changes in features over time (datetimeEpoch)?

Benefit: Discovering the patterns of weather changes over many years helps us predict the weather

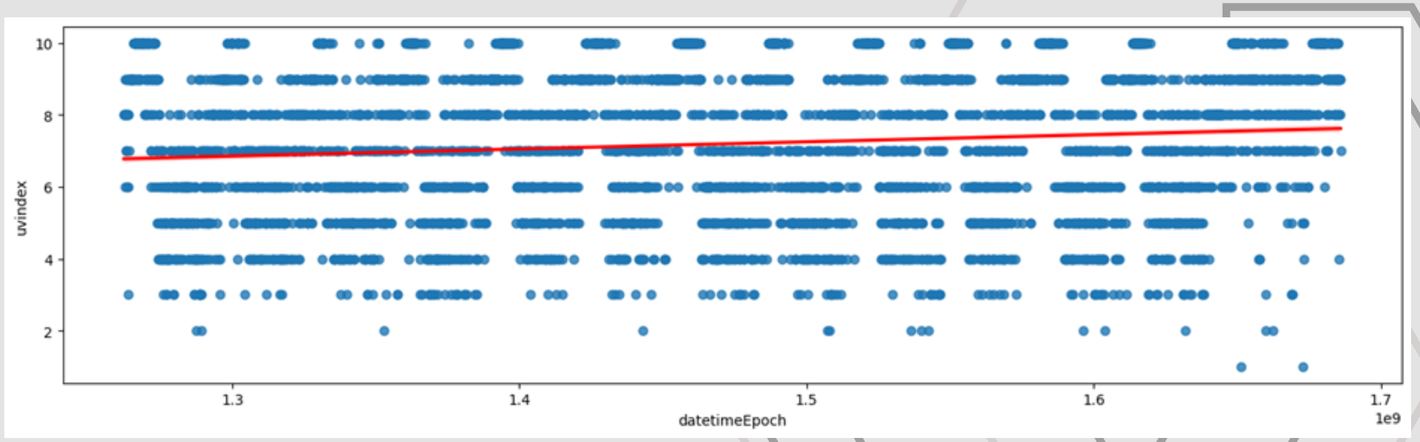
Temperature



precipprob, precipcover



uv index



Data Exploration

Make & Answer questions

Question 1: What is the trend of changes in features over time (datetimeEpoch)?

Benefit: Discovering the patterns of weather changes over many years helps us predict the weather

Conclusion

Over time, Ho Chi Minh City is experiencing a `noticeable warming trend`, evident in the rising temperature-related indices like 'tempmax,' 'tempmin,' 'temp,' 'feelslikemax,' 'feelslikemin,' and 'feelslike.' This warming, likely linked to climate change, poses concerns for the region.

The `increasing temperatures` are accompanied by a corresponding `rise` in the `dew` index, indicating heightened humidity levels. This interplay of factors has implications for various sectors, including agriculture and public health.

A unique observation is the simultaneous `increase in precipitation` probability and `decrease in precipitation cover`. This might signify changing precipitation patterns, impacting water resource management and flood control strategies.

Additionally, the `'uvindex' shows an upward trend`, signaling intensified ultraviolet radiation. This has implications for public health, emphasizing the need for increased awareness and protective measures.

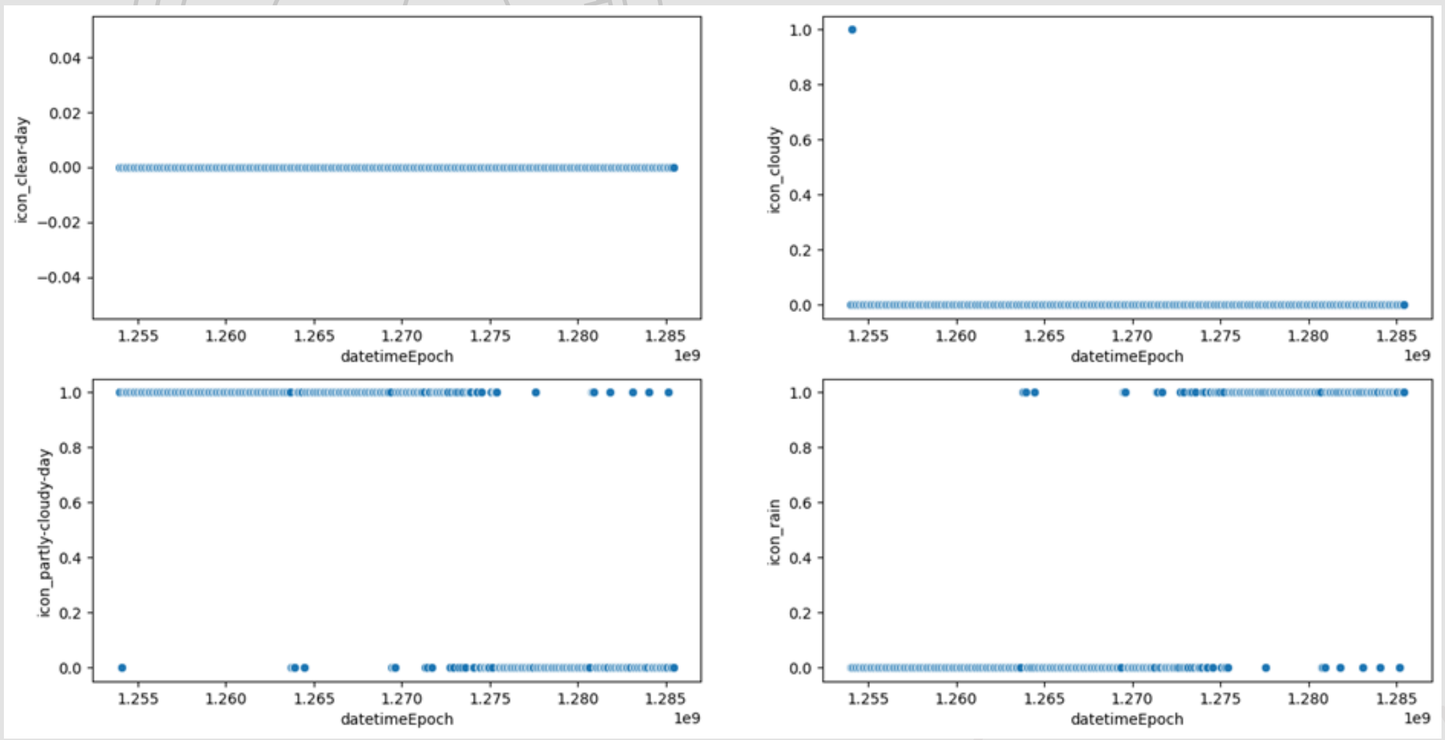
In summary, the data suggests a complex scenario in Ho Chi Minh City, driven by `climate change` and local dynamics, with potential impacts on multiple aspects of life. Further research is crucial for a comprehensive understanding and effective adaptation strategies.

Data Exploration

Make & Answer questions

Question 2: What is the weather condition of a year?

Benefit: We can predict the condition(rain, not rain)



Conclusion

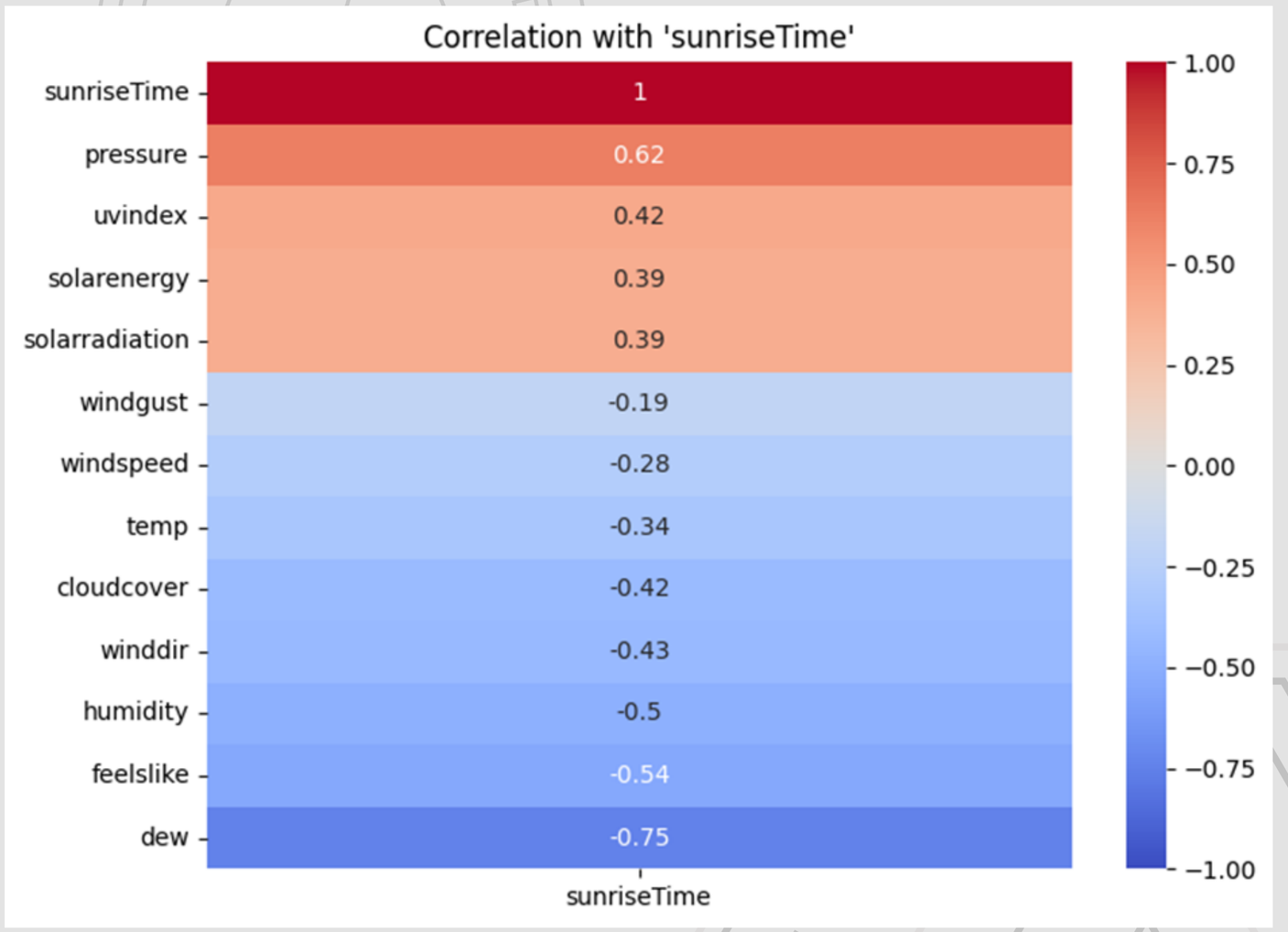
From the above data, we can observe that the weather in Ho Chi Minh City is distinctly divided into two seasons: the rainy season and the dry season. During the dry season, the sky is generally clear with occasional clouds. The remaining period constitutes the rainy season.

Data Exploration

Make & Answer questions

Question 3: Can we predict the weather for the day based on the time of sunrise?

Benefit: We can predict the weather at the beginning of the day

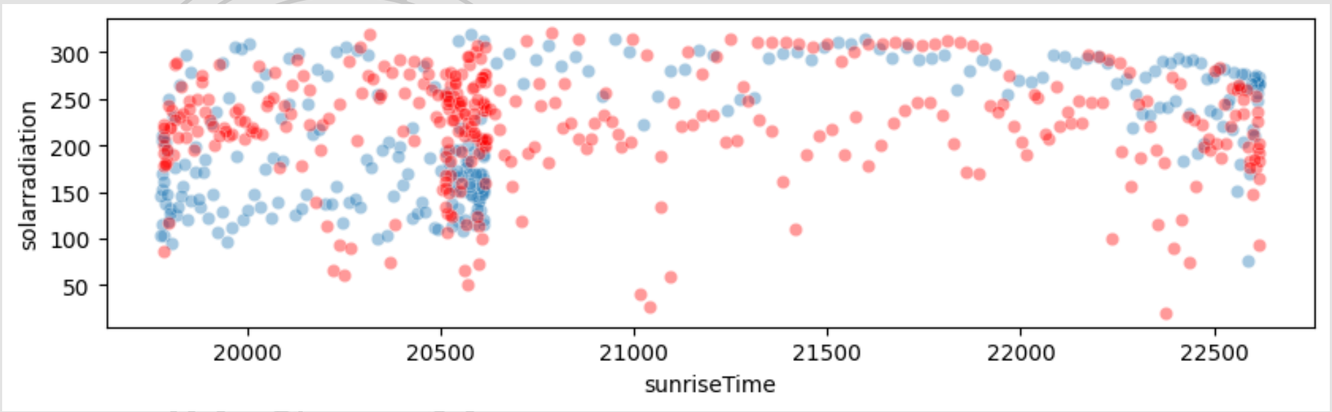


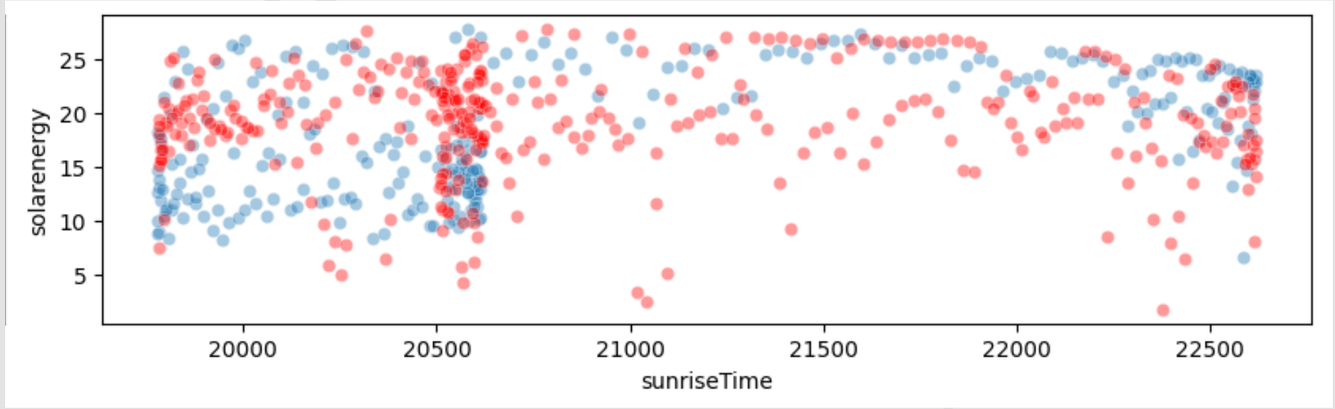
Data Exploration

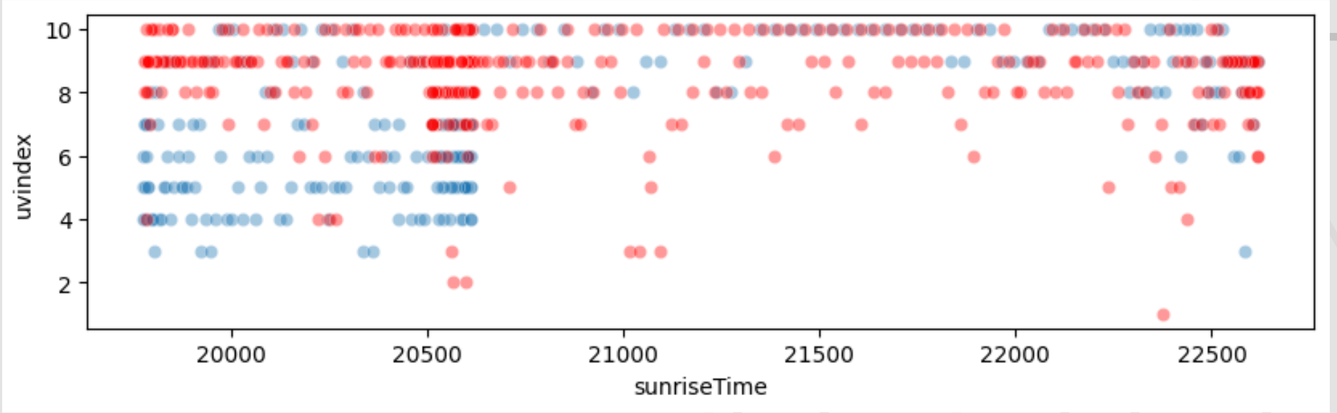
Make & Answer questions

Question 3: Can we predict the weather for the day based on the time of sunrise?

Benefit: We can predict the weather at the beginning of the day







Data Exploration

Make & Answer questions

Question 3: Can we predict the weather for the day based on the time of sunrise?

Benefit: We can predict the weather at the beginning of the day

Conclusion

The `temperature`, `humidity`, `wind speed`, and `cloud cover` tend to `decrease` as the `sunrise time increases`. Conversely, the indices related to `pressure, solar radiation, solar energy`, as well as the `level of UV` rays, tend to `increase`. The `wind direction` also undergoes a `significant change` before and after the threshold time of 20500.

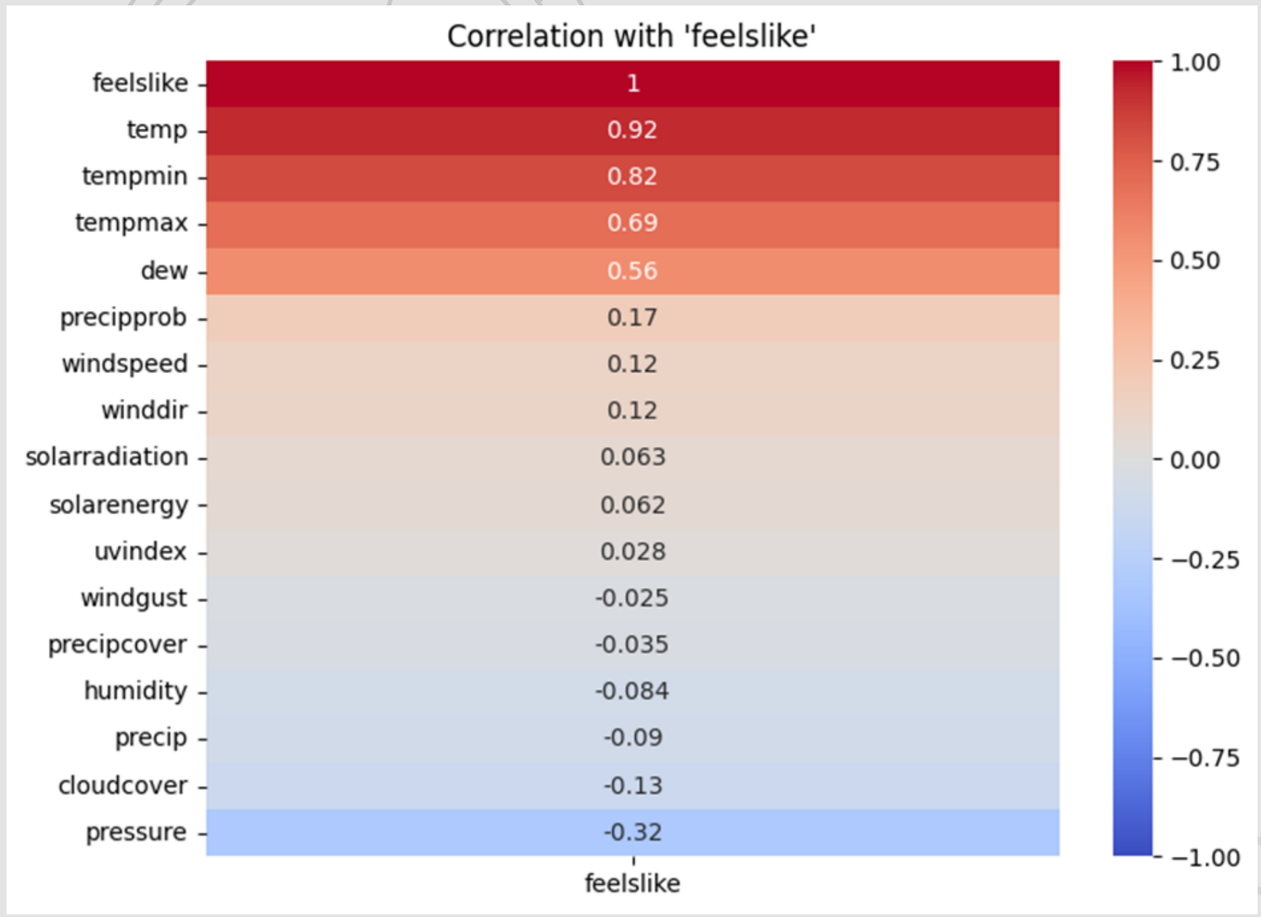
However, the impact of direct sunlight on 'solarradiation', 'solarenergy', 'uvindex' has significantly decreased between the first and last years. In the most recent year survey, it was observed that the `early sunrise` (before 5:41 AM) `did not contribute` to an increase in `solar radiation`, `solar energy`, and `UV index` as it did in the past (the first year of the survey).

Data Exploration

Make & Answer questions

Question 4: What factors influence the sensation of temperature during the day?

Benefit: Select appropriate factors to predict the perceived temperature



Conclusion

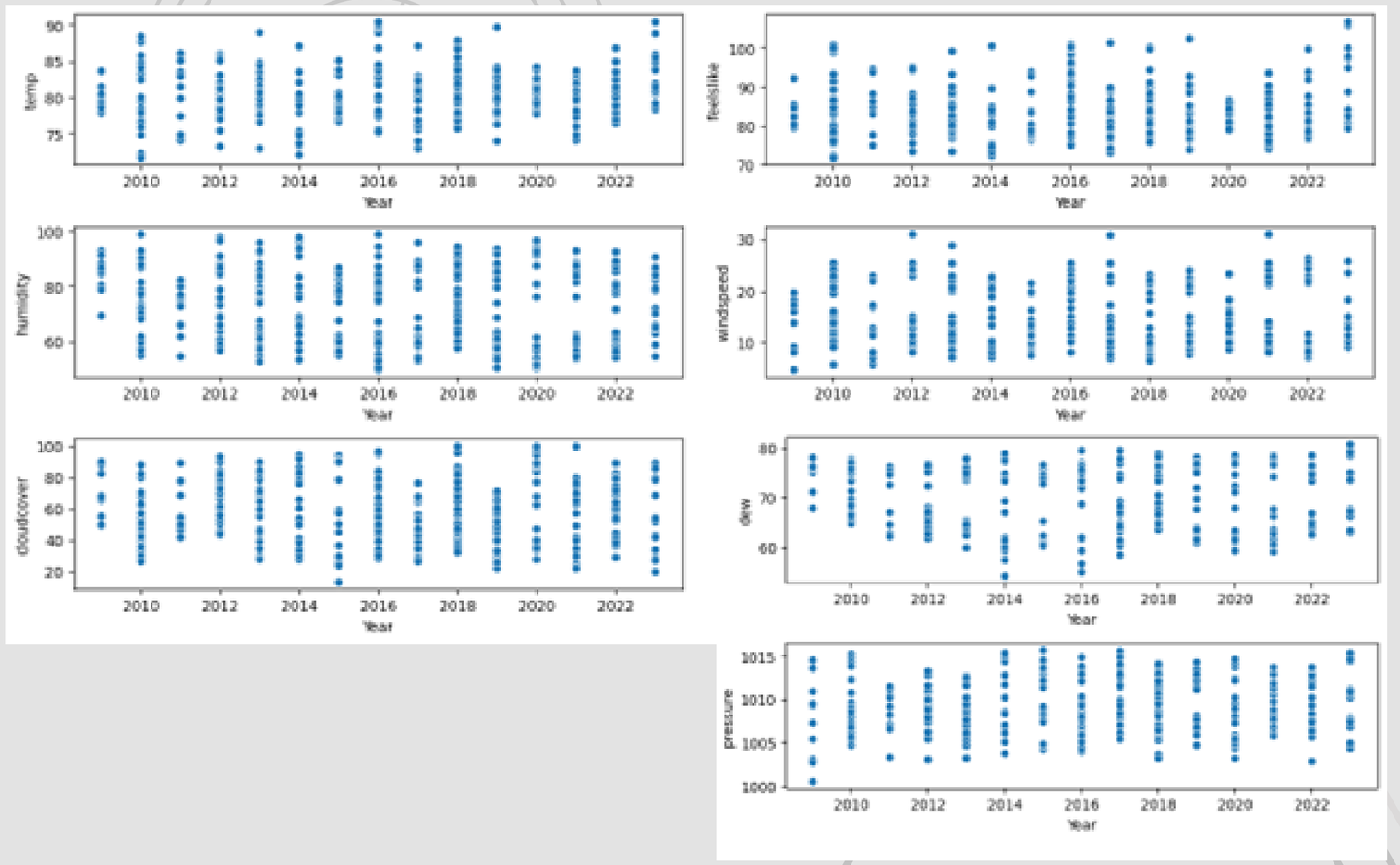
The sensation of temperature during the day is heavily affected by temperature, obviously. Surprisingly, **solar energy** and **solar radiation** show a low level of correlation. We can also observe that dew and pressure show a high correlation with **feelslike**. Surprisingly, **tempmin** exhibits a higher correlation with **feelslike** compared to **tempmax**.

Data Exploration

Make & Answer questions

Question 5: How are outliers distributed over time?

Benefit: Achieving a reasonable approach in handling outliers



Conclusion

Outliers are `evenly distributed` over time, so we `can removed them from the raw data`

Data Preprocessing

Data cleaning

Remove **datetime** and **dateEpoch**, use a column for storing month instead

Remove **sunriseEpoch, sunsetEpoch**, store **sunrise, sunset** as the number of seconds since 00:00

Remove **snow, snowdepth** and **windgust, severerisk**

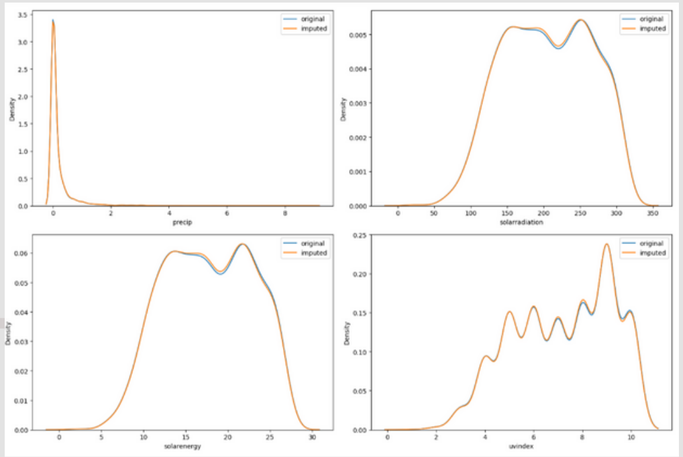
Remove **precipprob** and encode **preciptype** to 1 and 0

Remove **description, conditons**, keep only **icon**

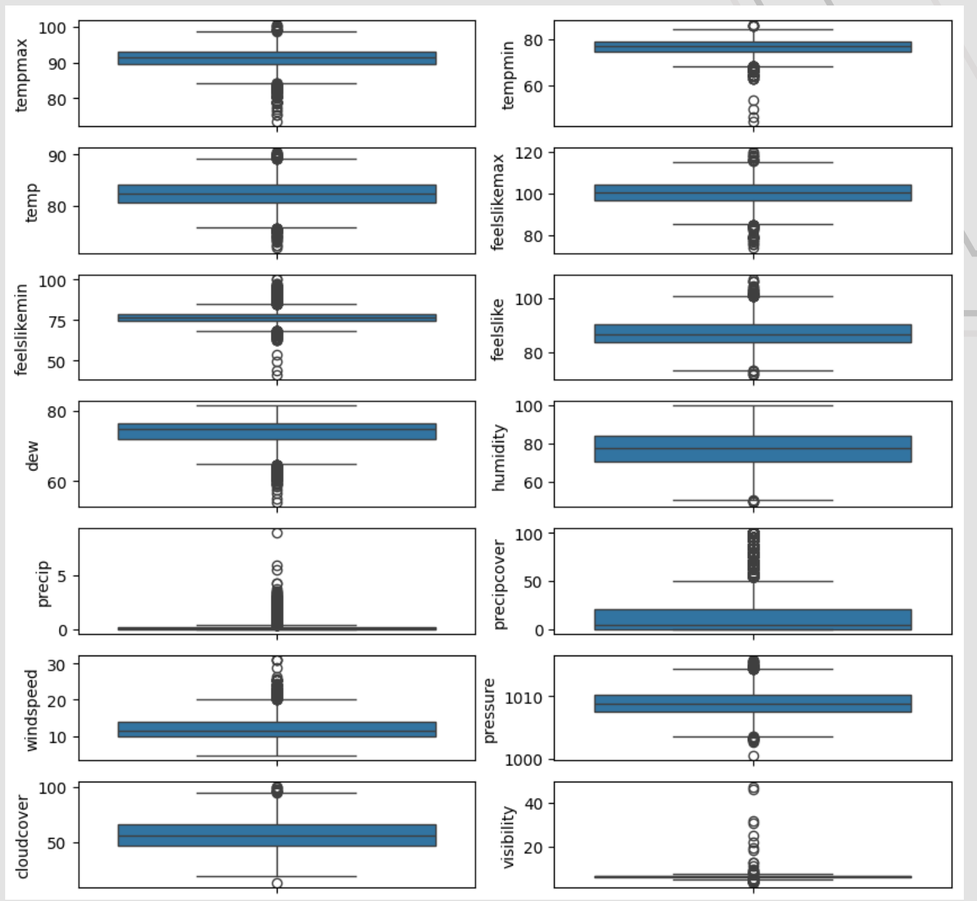
Remove **stations** and **source**

Imput missing values

Using **K-nearest neighbor imputation** to imput missing values for 5 columns **precip, pressure, solarradiation, solarenergy** and **uvindex**



Detect outliers & Normalize



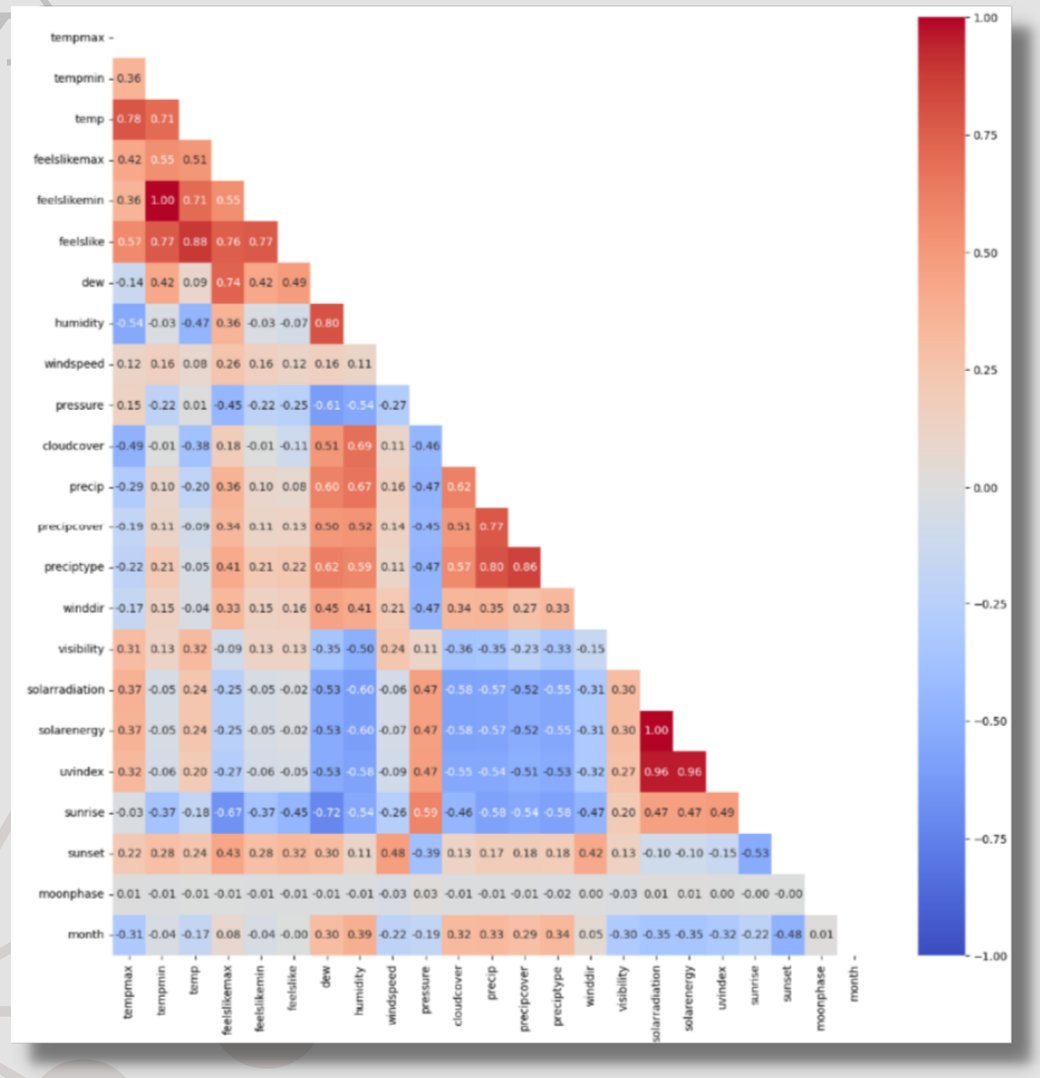
There are quite many features with outliers

Using **interquartile range (IQR)** to remove extreme outliers

After removing extreme outliers, scale columns to transforming it into normal distribution using **z-score** normalization

Data Preprocessing

Multicollinearity handling



For temperature values, **temp, tempmax, tempmin, feelslikemax, feelslikemin** and **feelslike** exhibit relatively high correlation with each other. Keep only **temp** as it represents the average actual temperature

Related to moisture content in the air, **humidity** and **dew** have a relatively high correlation with each other. Keep only **humidity** as its distribution is closer to a normal distribution compared to **dew**

Data Preprocessing

Multicollinearity handling

	Feature	VIF
10	solarradiation	2477.340006
11	solarenergy	2453.600076
12	uvindex	12.665446
13	sunrise	3.912138
1	humidity	3.866148
14	sunset	3.378962
16	month	3.041676
0	temp	2.446150
7	preciptype	2.353095
4	cloudcover	2.325132
3	pressure	1.984554
8	winddir	1.526897
6	precipcover	1.458997
2	windspeed	1.369087
5	precip	1.220756
9	visibility	1.063483
15	moonphase	1.005671

Using the **Variance Inflation Factor (VIF)** to evaluate the correlation between the columns

Removing **solarradiation** and **solarenergy** as **solarradiation, solarenergy** and **uvindex** are all related to solar radiation and **uvindex** has a significantly lower VIF value compared to solarradiation and solarenergy

Data Preprocessing

Multicollinearity handling

	Feature	VIF
11	sunrise	3.866514
1	humidity	3.853825
12	sunset	3.347349
14	month	3.035989
0	temp	2.443007
4	cloudcover	2.291703
7	preciptype	2.282592
3	pressure	1.977023
10	uvindex	1.755391
8	winddir	1.523650
6	precipcover	1.456573
2	windspeed	1.367488
5	precip	1.220107
9	visibility	1.063388
13	moonphase	1.005153

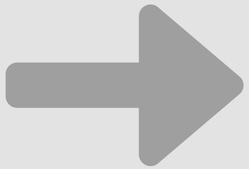
After all, all the columns in the dataframe have a VIF value of less than 5, which means that there is no multicollinearity between the columns

The processed data is stored into processed_data.csv file

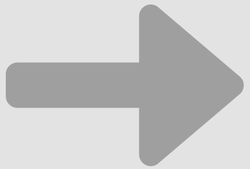
Data Modeling

Process label

icon	count
rain	2569
partly-cloudy-day	1441
clear-day	1



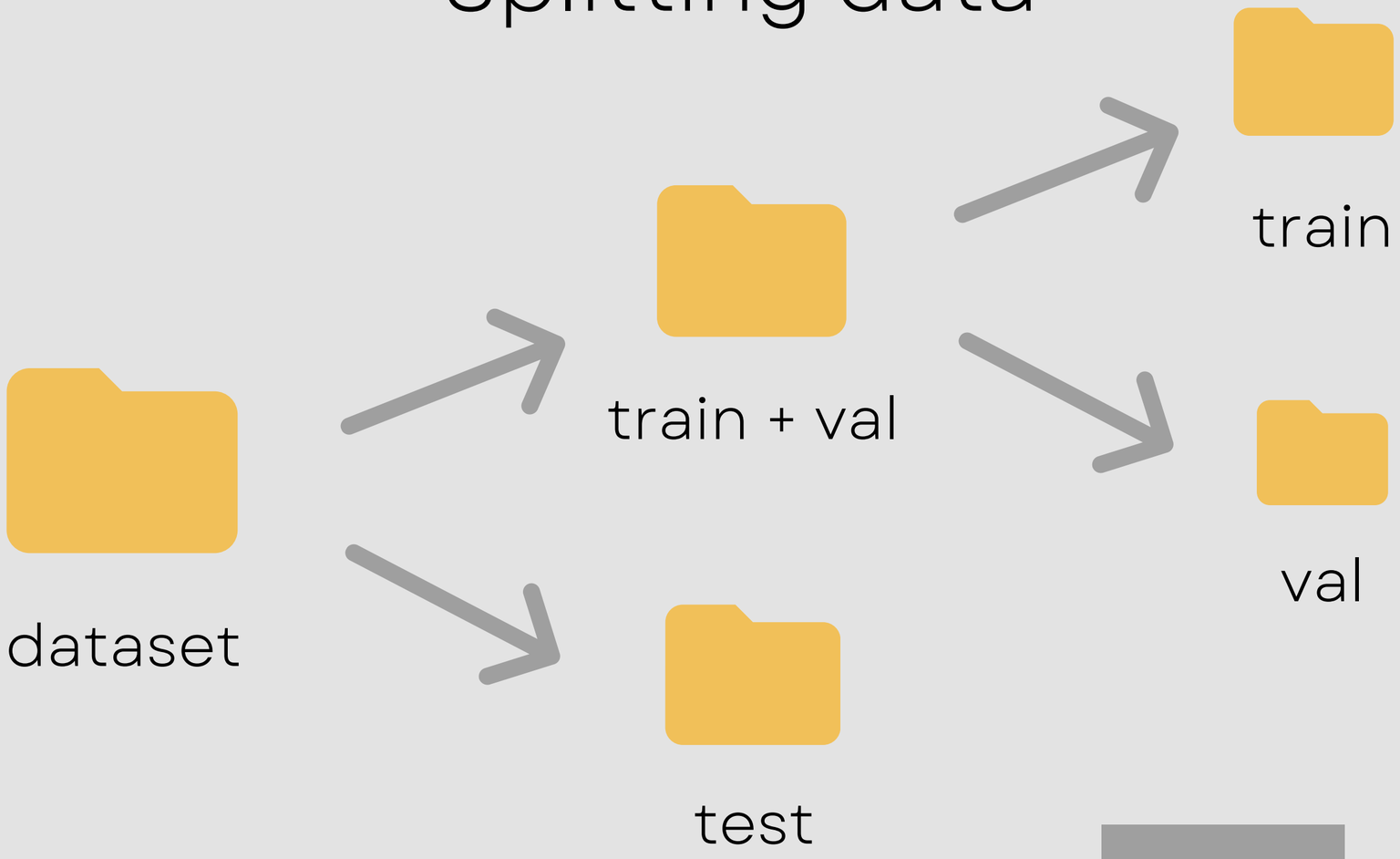
icon	count
rain	2569
partly-cloudy-day	1441



label_encoded	count
1	2569
0	1441

Data Modeling

Splitting data



Data Modeling

Classification Models

Decision Tree

Logistic Regression

Naive Bayes

Random Forest Classifier

KNN Classifier

XGBoost Classifier.

Support Vector Machine

Data Modeling

Metrics

Accuracy

The ratio of the total number of correct predictions and the total number of predictions

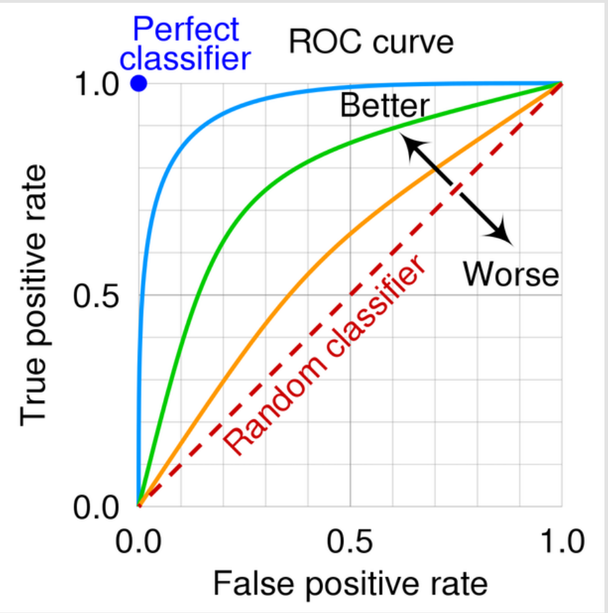
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall

The measure of our model correctly identifying True Positives

$$recall = \frac{TP}{TP + FN}$$

ROC curve and AUC

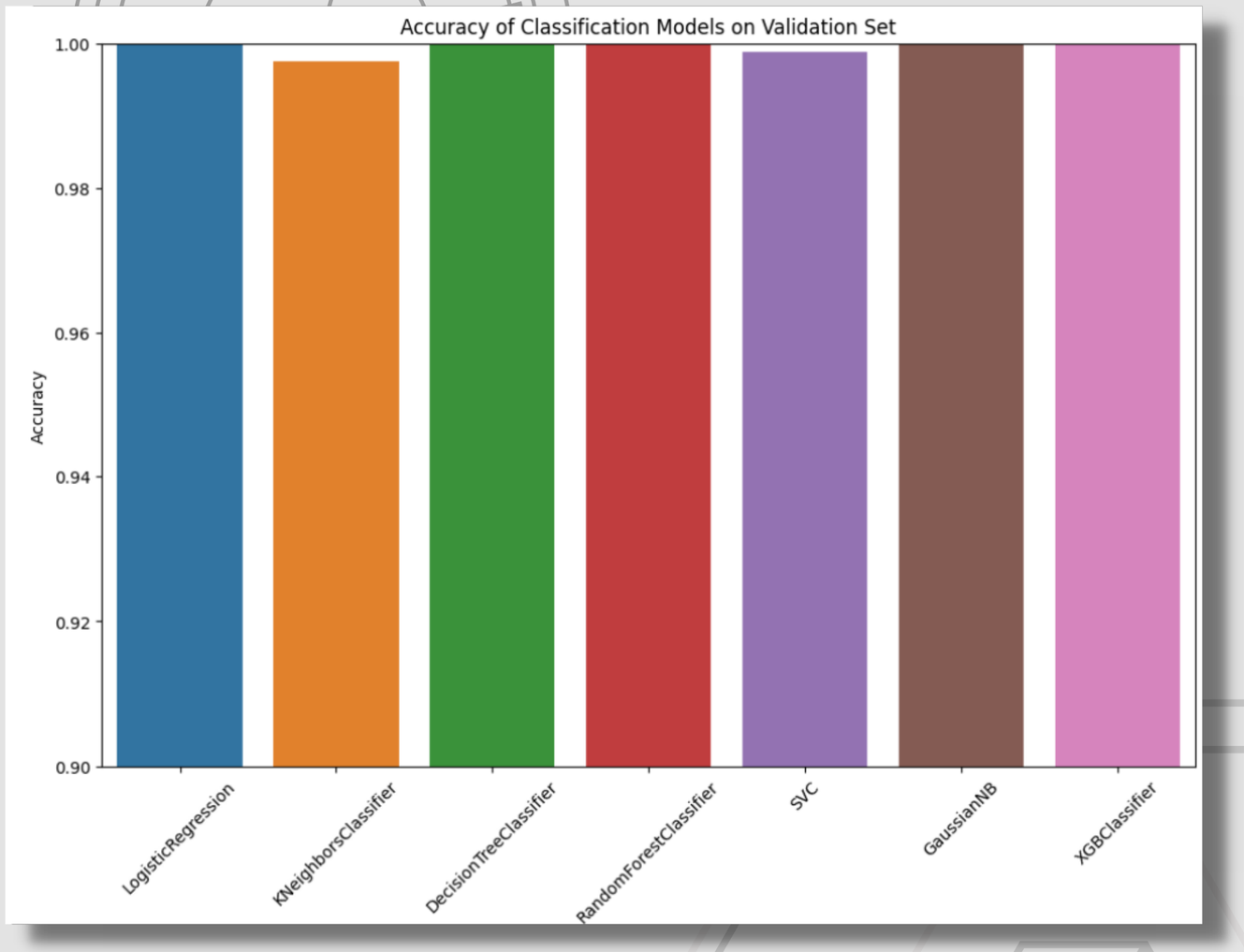


Data Modeling

Model Selection

Basic approach: train all the model on the train dataset and evaluate on the validation set

Base on their accuracy on the validation set, the models with high accuracy will be chosen

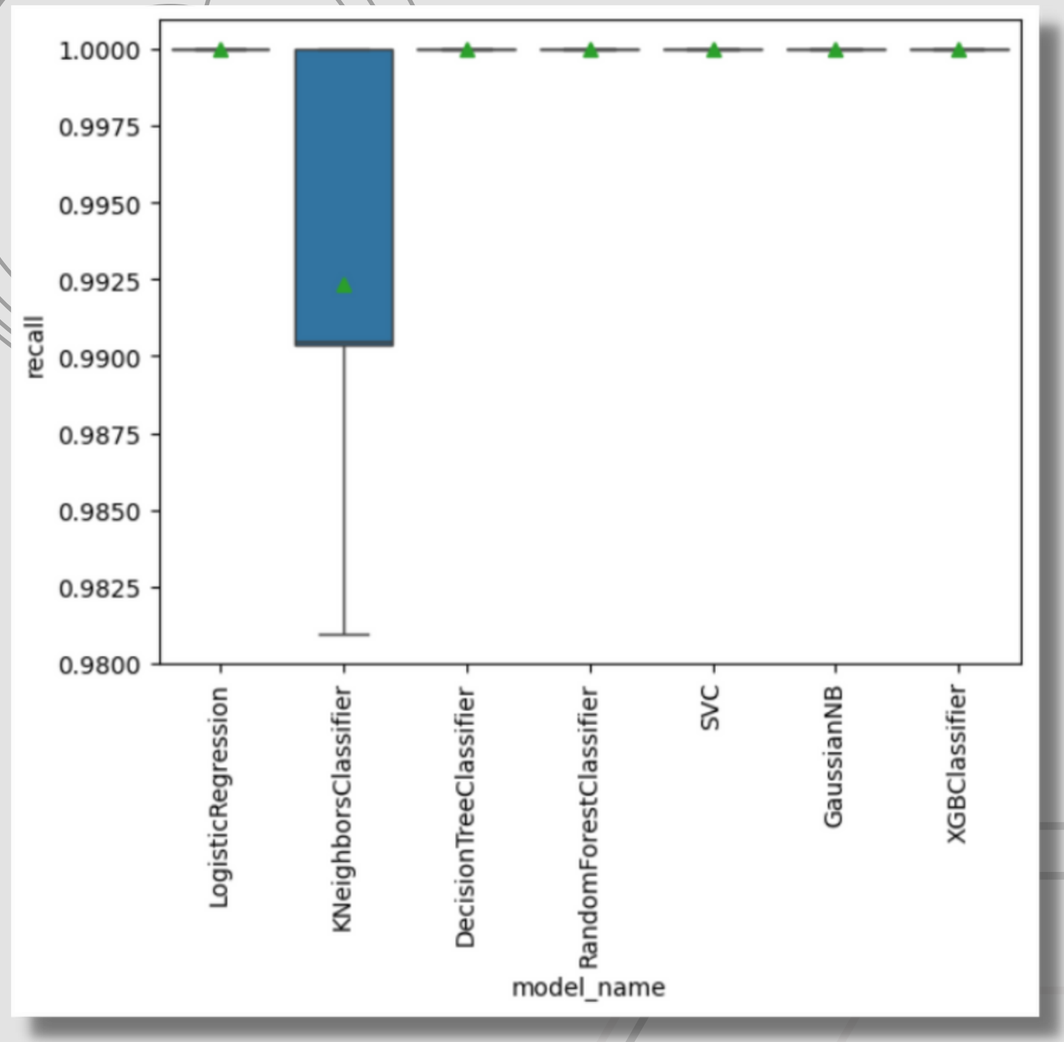


Data Modeling

Model Selection

Using the **k-fold cross-validation** to select the best and most representative model to use

The validation set is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time



recall metric is used in this case as we want to accurately predict rainy days

Data Modeling

Fine-tunning

Models

KNN Classifier

Random Forest Classifier

XGBoost Classifier

Technique

A technique for finding the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made

GridSearchCV

Data Modeling

Fine-tunning

Models

KNN Classifier

Random Forest Classifier

XGBoost Classifier

Hyperparameters

'metric': 'euclidean',
'n_neighbors': 3,
'weights': 'uniform'

'criterion': 'entropy',
'max_depth': 5,
'min_samples_split': 2,
'n_estimators': 100

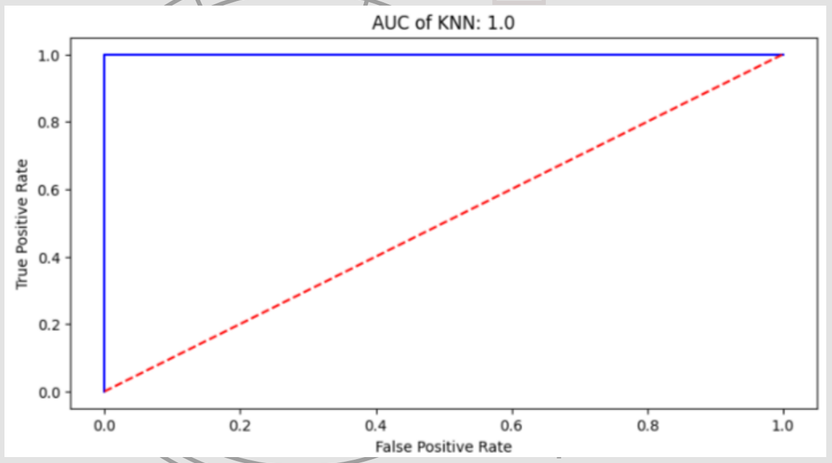
'learning_rate': 0.1,
'max_depth': 5,
'min_child_weight': 1,
'n_estimators': 100

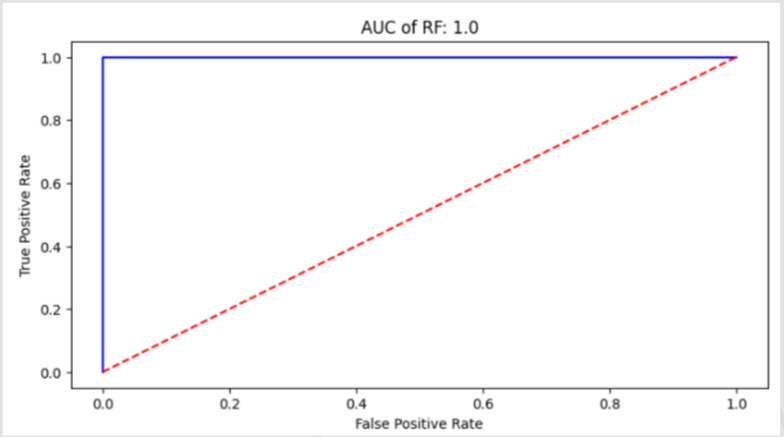
Data Modeling

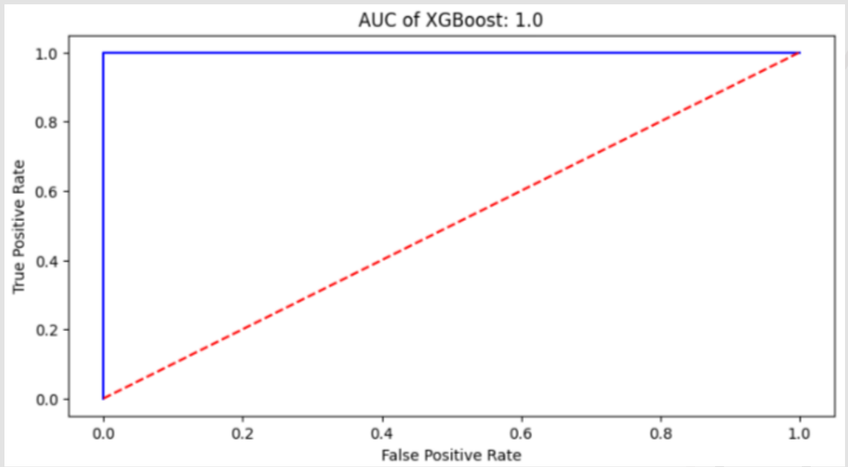
Fine-tunning

Use ROC curve and AUC metric to re-evaluate 3 models after fine-tunning

ROC curves of all three models are close to the upper-left corner, indicating high TPR and low FPR at various threshold values







Data Modeling

Basic ensemble method

Weighted Average

Weighted Average technique is used to predict results by combining the predictions of multiple models

$$pred_{final} = \sum_{i=1}^{n} (pred_i * w_i)$$

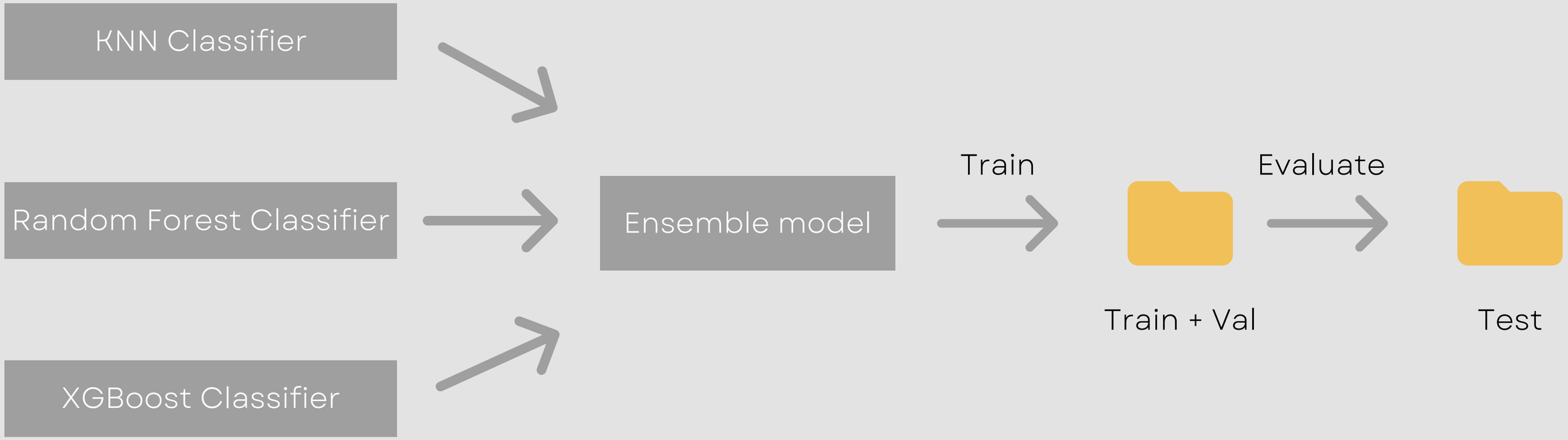
The weight of each model can be calculated by taking the **recall** of that model on the **validation set** and dividing it by the sum of the **recall** of all models on the **validation set**

Based on the results of the **k-folds cross-validation** method on the **validation set**, we define the weights as follows

Model	Weight
K-Nearest Neighbors Classifier	0.3
Random Forest Classifier	0.35
XGBoost Classifier	0.35

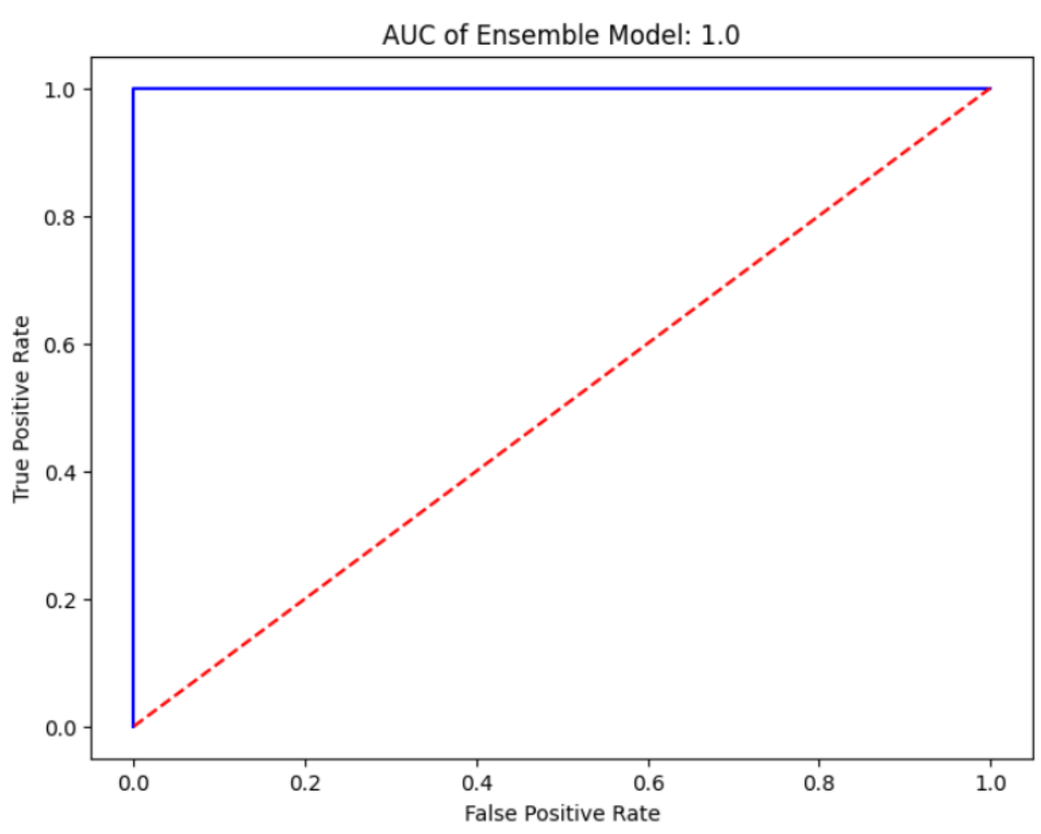
Data Modeling

Retrain model



Data Modeling

Model Evaluation




```
Accuracy of the ensemble model: 1.0
Confusion matrix of the ensemble model:
[[288   0]
 [  0 514]]
Recall of the ensemble model: 1.0
```

Result of ensemble model

Conclusion

Our final model has perfect accuracy and recall on the test set. The ROC curve and AUC also indicate that the model has the ability to classify data points accurately

The reason for achieving such high accuracy is due to our dataset being weather data in Ho Chi Minh City, where the weather is relatively simple and does not have many fluctuations. If we were to replace it with weather data from a different location with more diverse climate patterns, we may not achieve the same high accuracy

However, we still go through all the steps, from applying k-folds cross-validation to select the best models, then performing fine-tuning to find the best hyperparameters for the models, and finally using Weighted Average to combine and improve the model's accuracy. With these steps, we can confidently say that even if the dataset changes, the final model will still be improved compared to using individual models without these steps

Lastly, with the high accuracy that the model achieves on the weather dataset in Ho Chi Minh City, we can use this final model for classification and weather prediction based on environmental information surrounding Ho Chi Minh City

THANK YOU FOR LISTENING