**weather prediction**

**case study:**

**Rain prediction in Australia**

**Aim:**

The aim of our work is to realize a model that predict if there will be rain fall tomorrow from data we have today. So, we will build a supervised learning data learning model on a historic dataset which is made available online.

**Requirements:**

For this project to come to completion we will use the following tools:

**\*QuickTable**

**\*Vscode**

**\*Python 3.7**

**\*Zoom**

**\*Bandicam**

**Project plan**

1. **Data acquisition**
2. **Data preprocessing**
3. **Data Modeling**
4. **User interface**

**Process**

**Data acquisition**

Our dataset is the **weatherAUS.csv** dataset gotten from <http://www.bom.gov.au/climate/data>. This dataset contains about 10 years of daily weather observations from numerous Australian weather stations. RainTomorrow is the target variable to predict. That is, is it going to rain the next day Yes or No.

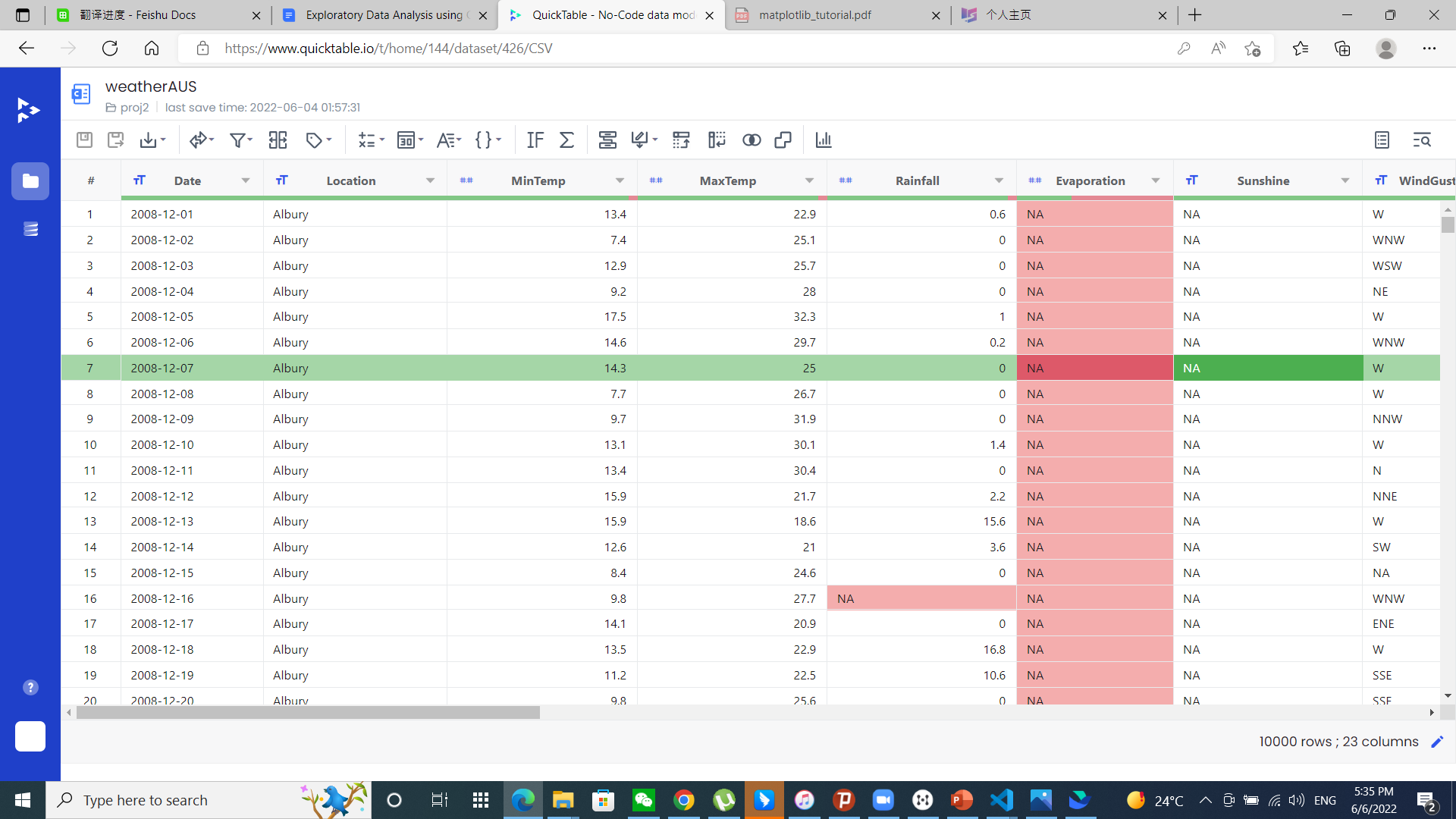
This column is Yes if the rain for that day was 1mm or more. The dataset contains about 145460rows and 23columns. We download the dataset and in a csv format.

**Data preprocessing**

Our preprocessing is divided into 3 phases:

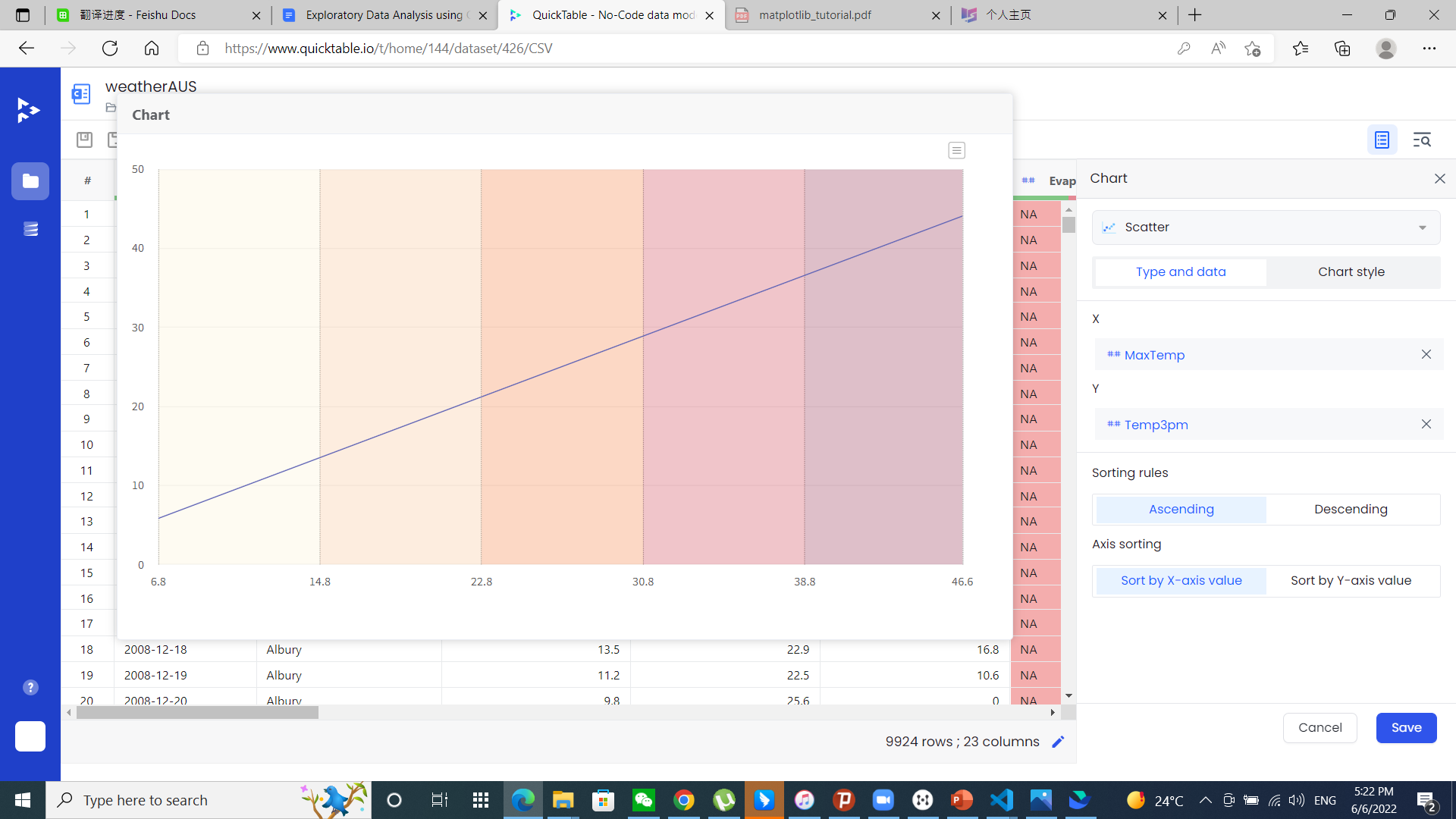
1. **Exploratory data analysis**

Here we take a look at the data and try to figure out the relationships between data columns. This was done with the help of **QuickTable** after uploading our dataset to QuickTable we have the following:



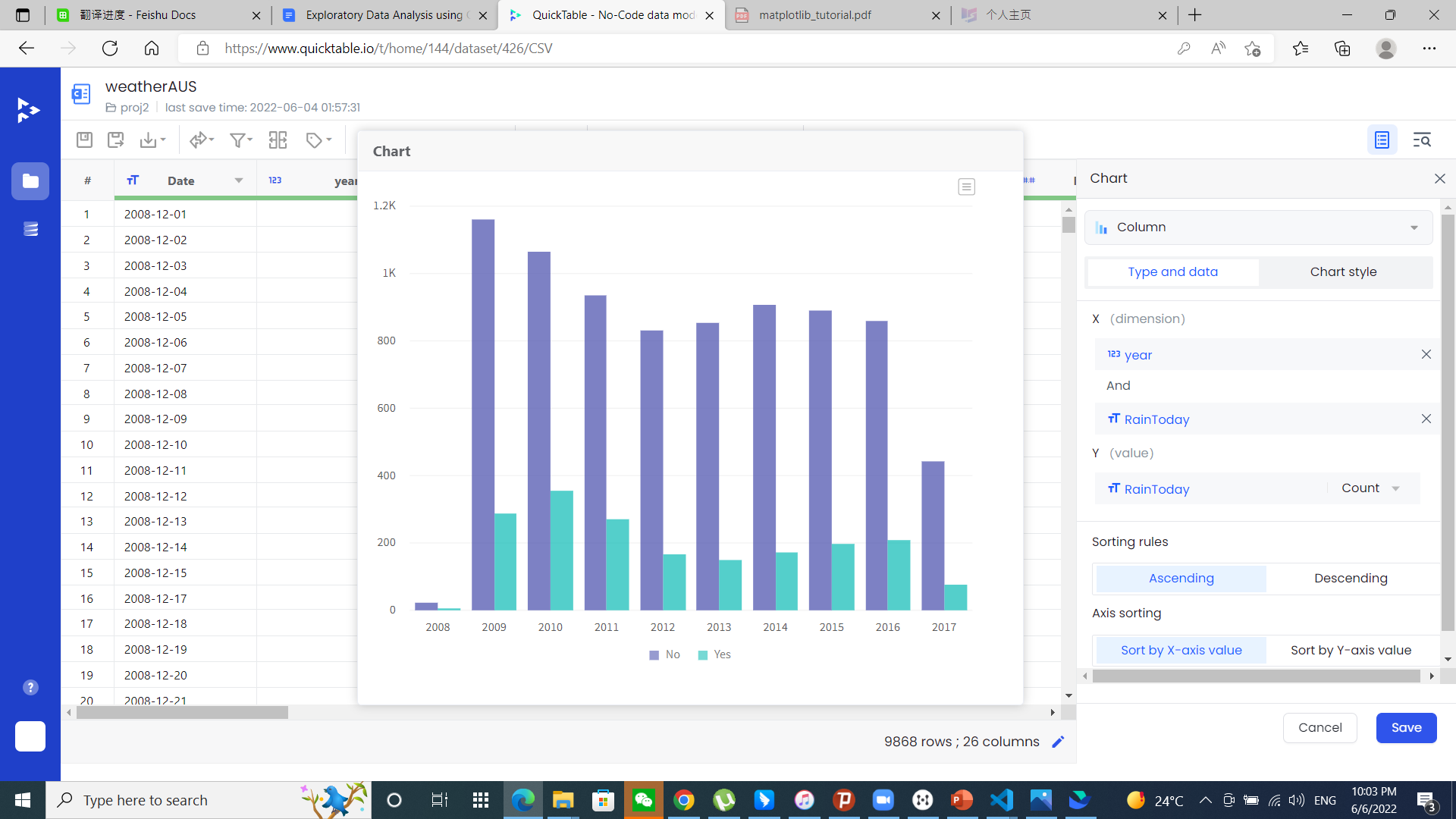
We can already see a chunk of missing values in our dataset.

Firstly, we do a line plot of **MaxTemp** and **Temp3pm**



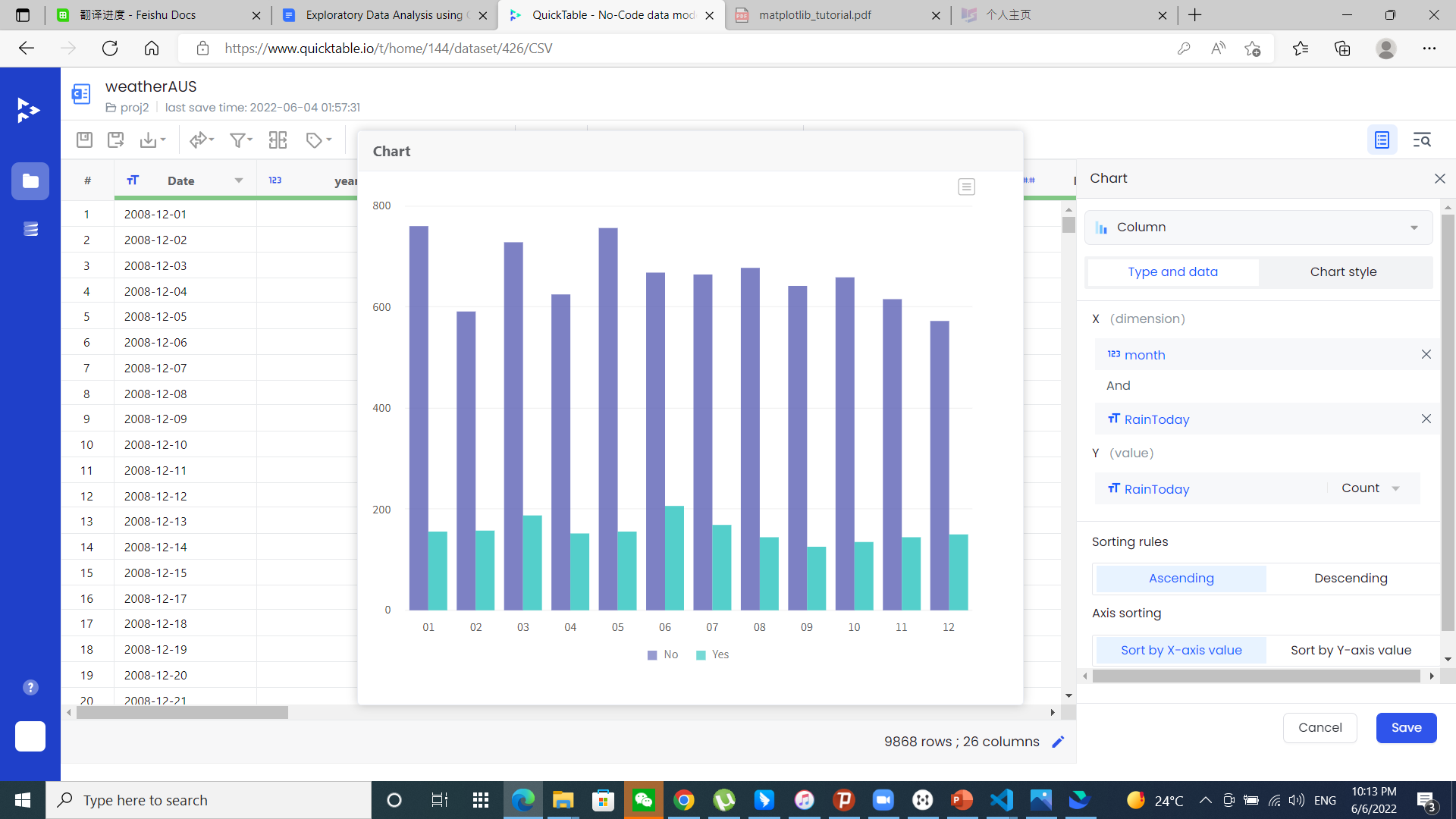
We can clearly see a linear relation between **MaxTemp** and **Temp3pm.**

Then a histogram of year and RainToday count

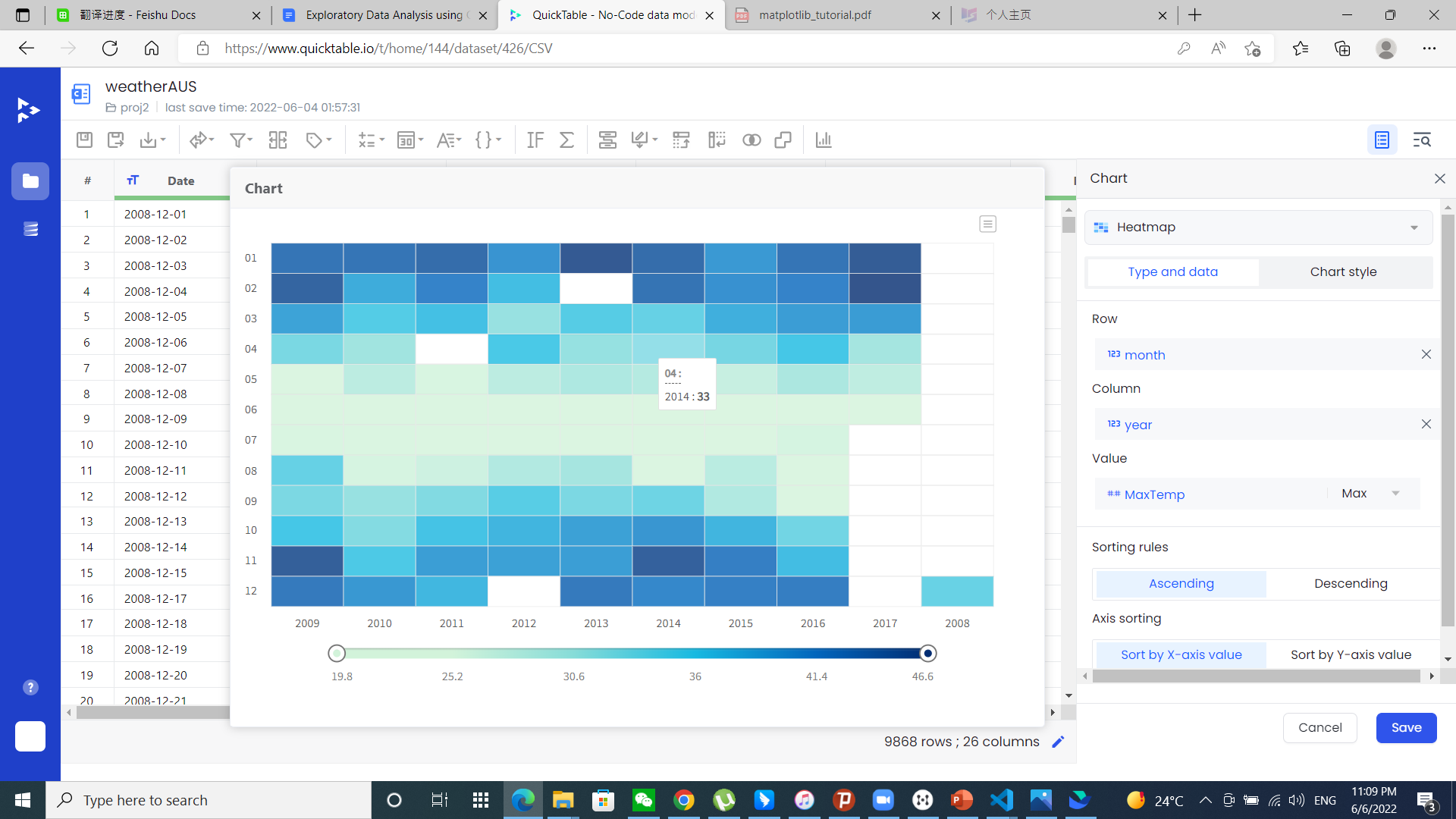


we can see that the average of RainToday count Yes and No is relative same for different years with respect to amount of data available.

Then a histogram of month and RainToday count



Here, we were trying to figure out if there is a defined rainy season and dry season but it was not really clear so we plotted a heatmap of month and MaxTemp

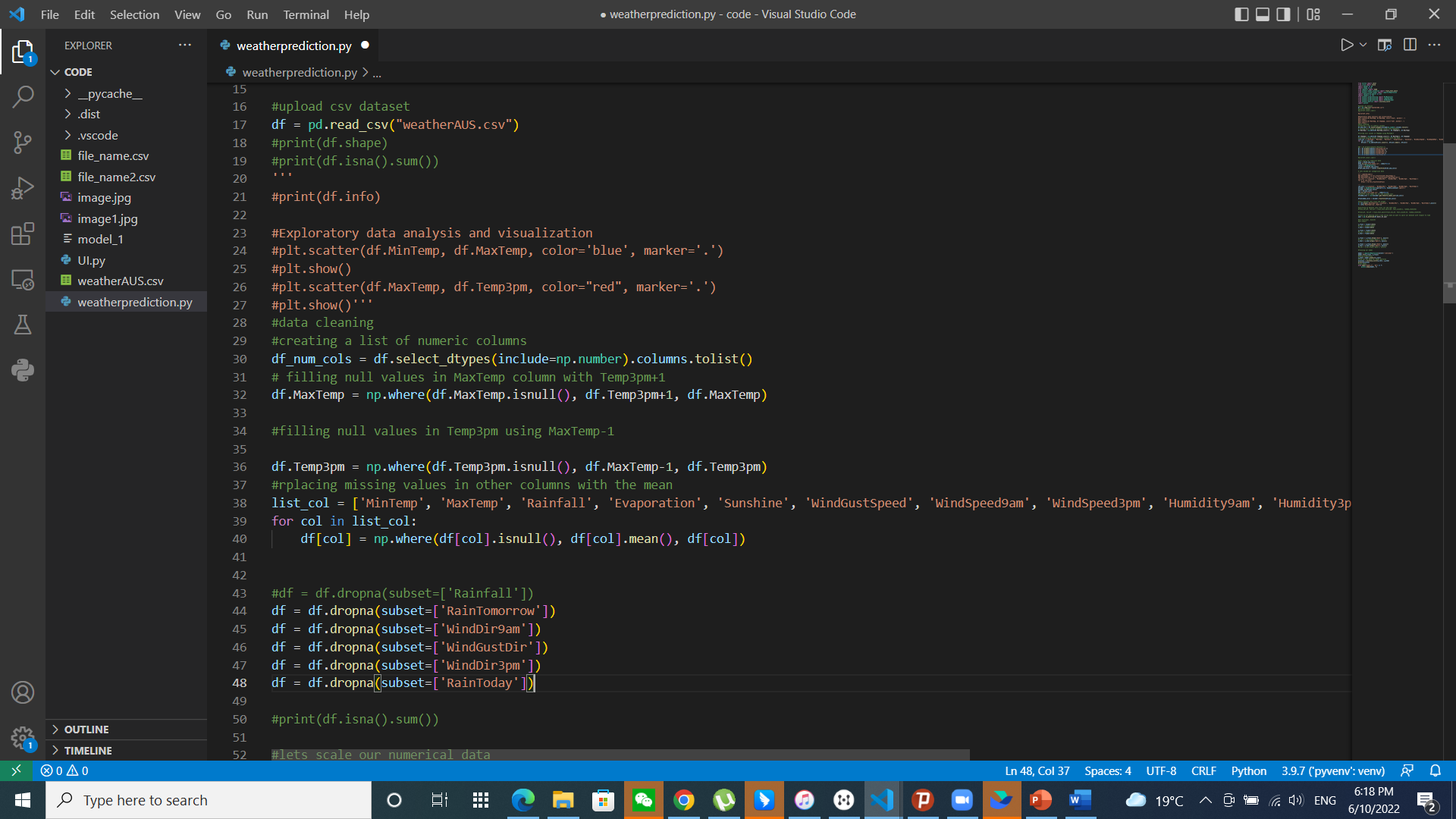


Assuming months with a high average of maxTemp is during the dry seasons while others will be during the raining season it was very clear from the heat map.

1. **Data cleaning**

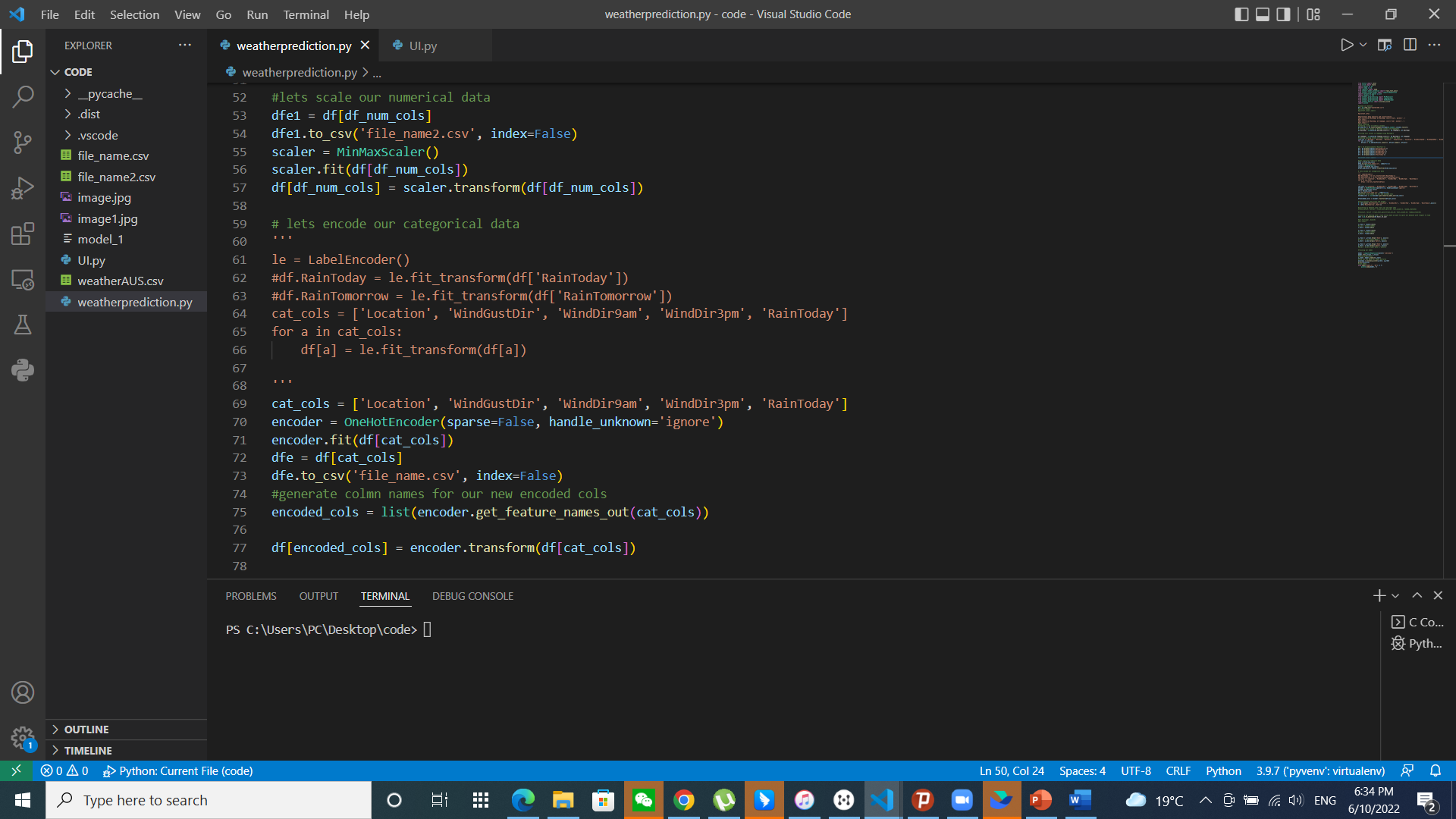
Data cleaning was done using vscode and the main programming language here was python 3.

The cleaning was done as shown below:



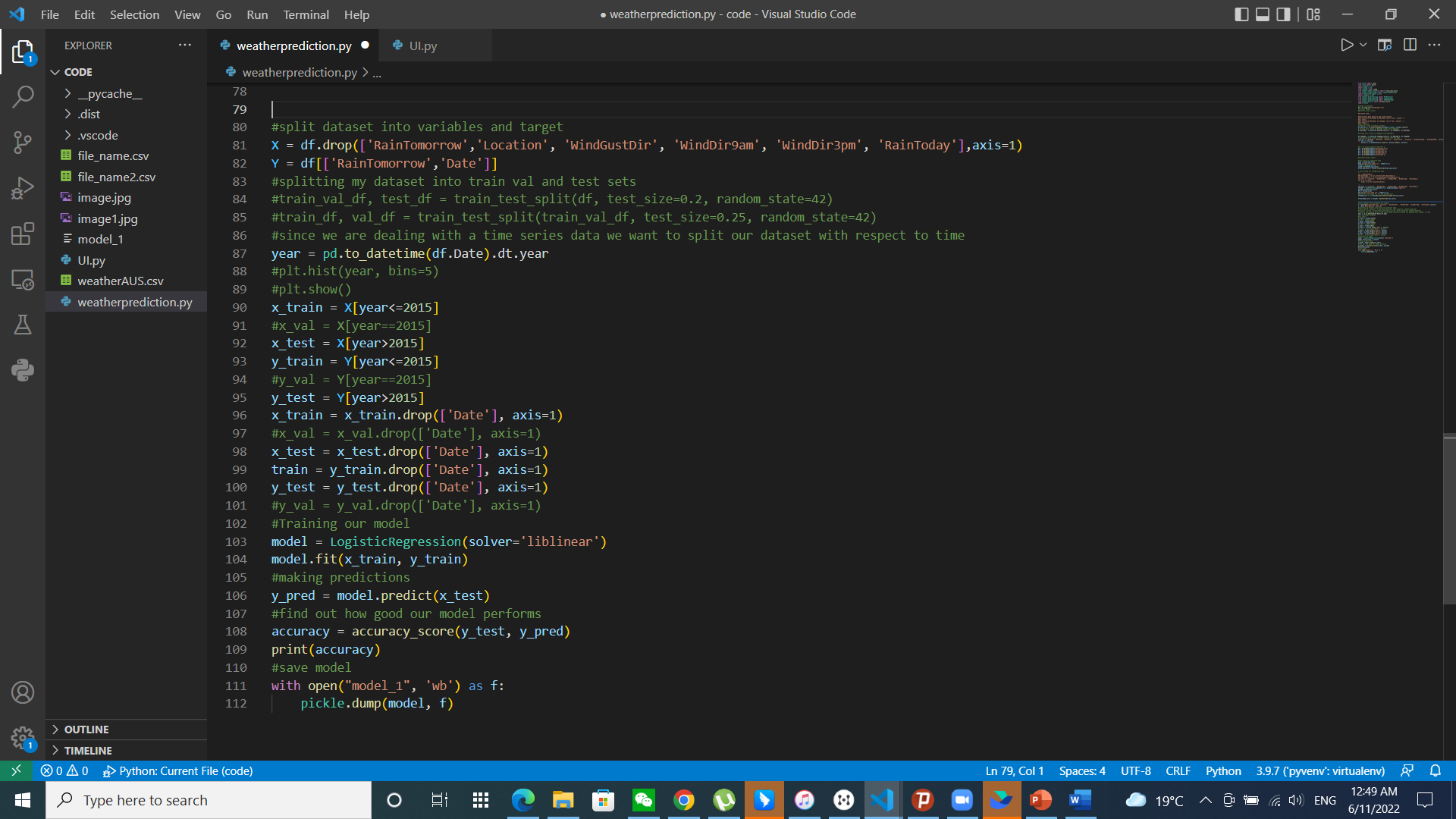
1. **Data scaling, encoding and normalization**

We also scaled our numerical data so every data point has a fairly equal weight on our model. We will also encode our non-numeric data



**Data Modeling**

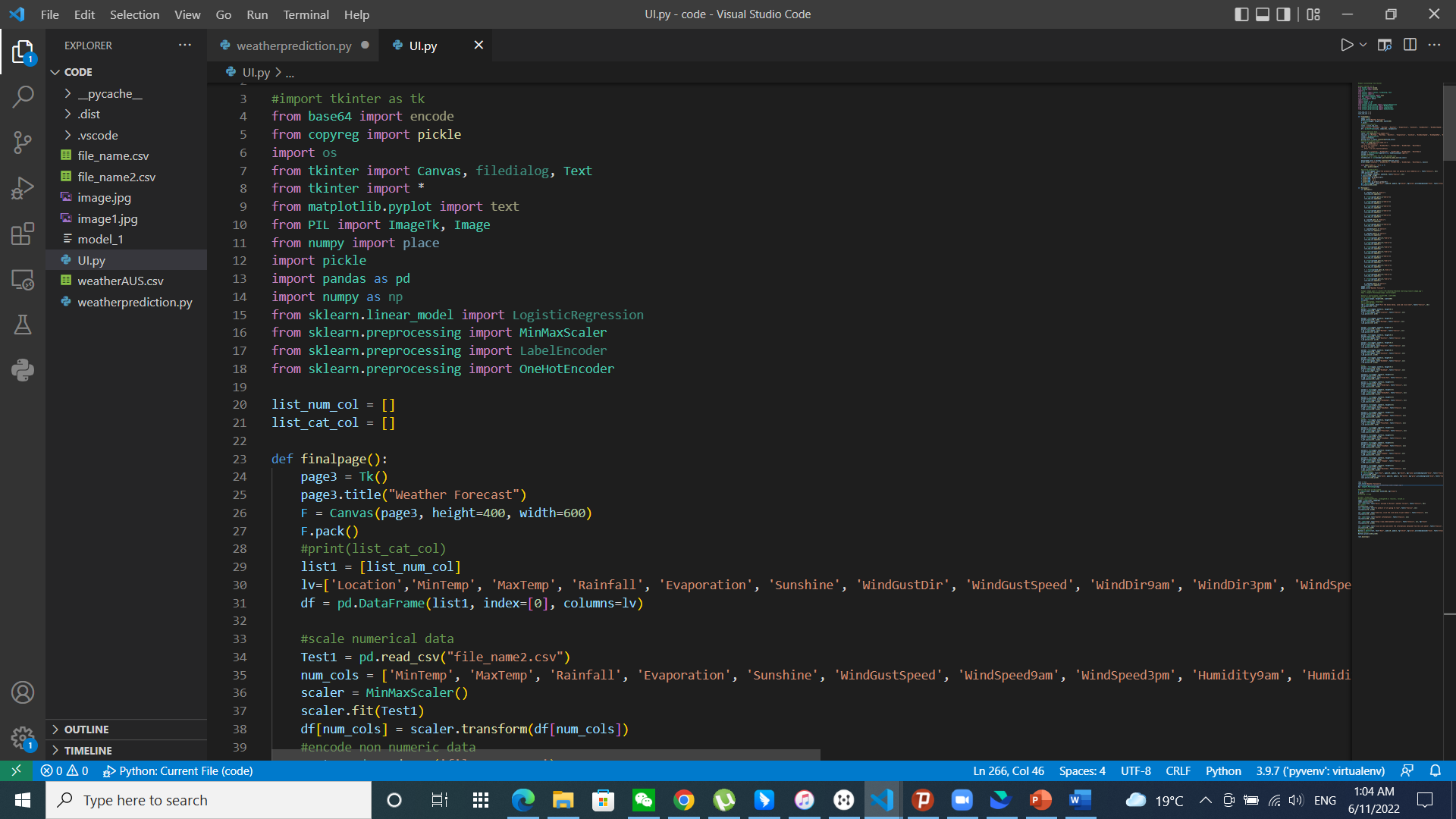
We used sklearns’ Logistic regression to build our model since our target column is categorical.



The model was trained once and saved for further use.

**User interface**

We made a simple user interface using pythons’ tkinter package.



Below a photo of the welcome window.

