

*Toronto*

**Data Proposal: Finalised**

Soni Manana (002952645)

Parth Savaliaa(002982302)

Dhairya Davea(002110382)

a *College of Professional Studies****,*** *Master of Professional Studies in Analytics.*

***Subject****: ALY6040*

Under the guidance of

**Prof. Kamran Hootan**

**Introduction**

This data is from [Kaggle](https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package?datasetId=6012&sortBy=voteCount) with titles “Rain in Australia”. By gathering quantitative information about the atmosphere's current condition at a specific location and utilising meteorology to predict how the atmosphere will evolve, weather forecasts are created. To forecast whether or not it will rain tomorrow, use the Rain Dataset. About 10 years' worth of daily weather measurements from several Australia locales are included in the dataset. As of seeing this, this will be use for classification purposes which can be said as weather forecasting. Speaking of weather predictions, they might have an impact on daily activities as well as industries like the food sector, tourism, emergency healthcare, etc. There is a target variable that can be either "Yes" or "No”. Yes, if there was at least 1mm of rain that day. The variables in our dataset that are most likely to cause rain to fall are pressure, humidity, pressure, clouds, and sunlight. Finding a link between them that is supported by statistical evidence is therefore essential for predicting when it will rain.

Problem statement for this dataset are as follows:

a) Use machine learning techniques to create a prediction model that predicts whether or not it will rain tomorrow.

b) Use machine learning techniques to create a prediction model that estimates the likelihood of rainfall.

**Characteristics of Dataset**

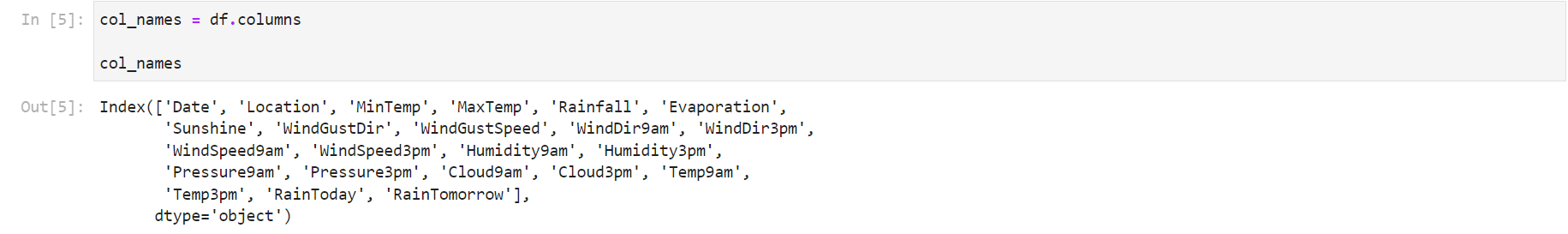
**Variables**

As shown in the Fig.1, These are the names of variables. There are two *RainToday* and *RainTomorrow* will be helpful to develop model because they have values such as *Yes* or *No* for Rain of specific days.

**Number of samples**

As show in the below picture there are 145,460 samples which are enough for a model not so less not so many.

**Fig. 1 Features**

****

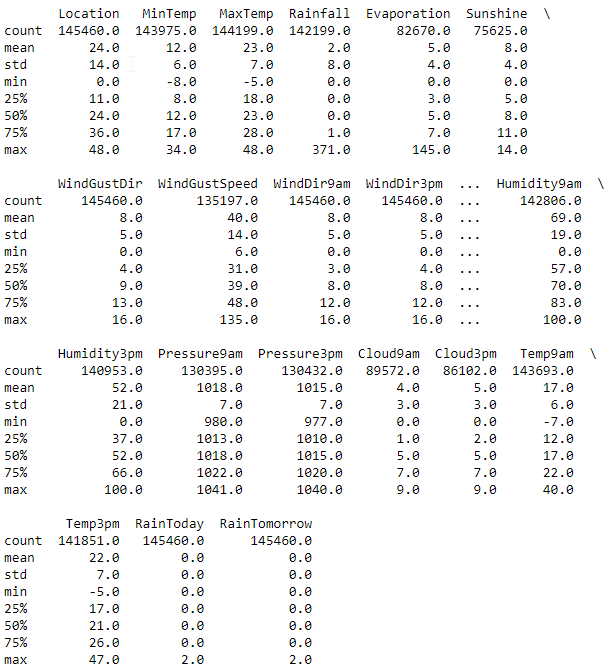
**Number of samples**

This project will be in Pythonand as shown below in the figure is a dataframe of pandas. Variables consumes 25.5+ memory which we can reduce using providing dtypes as float64 to float 32. There are missing values in many features like *sunshine.* Image beside showing null counts according to features. Dtype of *object* are having strings so that they are categorical variable.

**Description of Data**

There are some minor differences can be seen such as slight fall in mean of Humidity9am and Humidity3am this same can be seen in other variables. Max temperature recorded is 47. Almost no difference in pressure variables. Sunshine varies between 0 to 14 hours like according to season.

**Fig. 2 Information of Dataframe Fig.3 Description of Data**

Fig. 2 Information of Dataframe


**References**

Young, & Young. (2020). *Rain in Australia*. Kaggle. Retrieved October 6, 2022, from <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package?datasetId=6012&sortBy=voteCount>