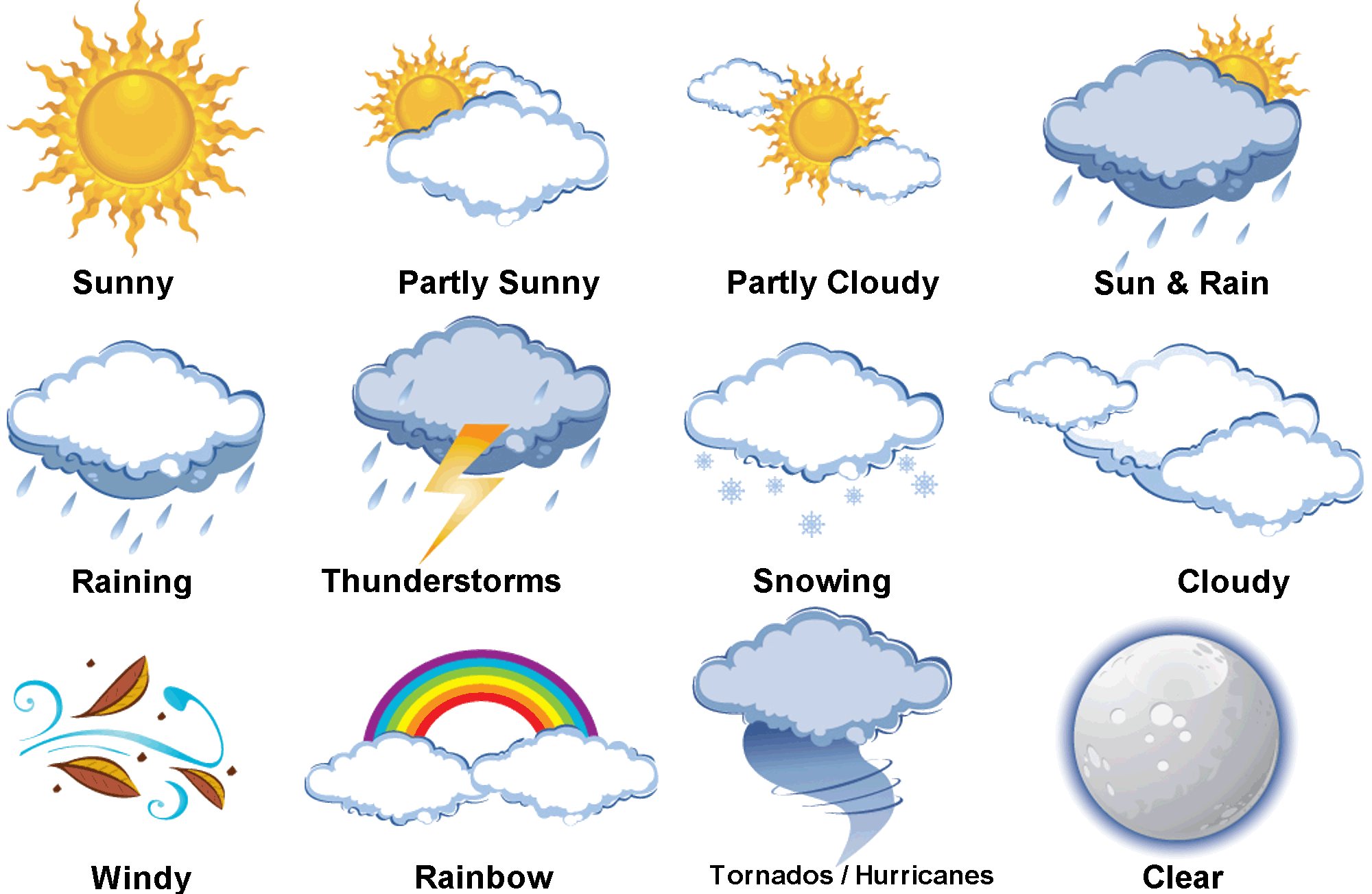
Data Analyst Nanodegree





Explore Weather Trends

By

Anuj Dutt

**Extract Data from Database** (using SQL)

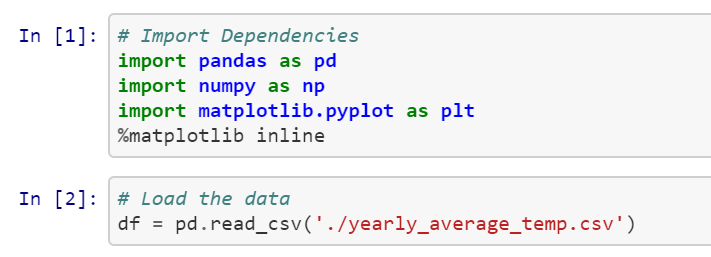
|  |  |
| --- | --- |
| Data Type | SQL Command for Extracting Data |
| To select the Country Name in “city\_list”. | SELECT \*  FROM city\_list  WHERE country = 'India'; |
| To Alter the name for “avg\_temp” in  City\_data to “city\_avg\_temp”. | ALTER TABLE city\_data  RENAME COLUMN avg\_temp to city\_avg\_temp; |
| To Alter the name for “avg\_temp” in  Global\_data to “global\_avg\_temp”. | ALTER TABLE global\_data  RENAME COLUMN avg\_temp to global\_avg\_temp; |
| To join two tables “global\_data” and “city\_data” and extract three columns  of relevant data i.e. “year”, “global\_avg\_temp” and “city\_avg\_temp”  for a city in India. | SELECT gd.year, gd.global\_avg\_temp, cd.city\_avg\_temp  FROM global\_data AS gd  INNER JOIN city\_data as cd  ON gd.year = cd.year  WHERE city = 'Delhi'; |

**Approach to extract the desired data:**

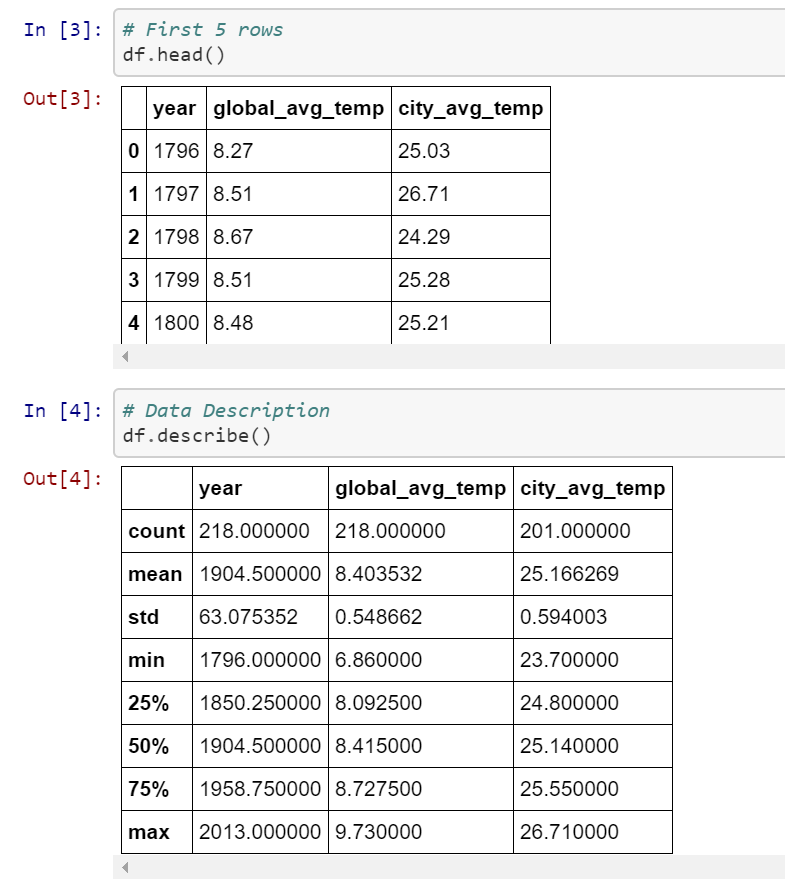
* I used the above SQL commands to get the final CSV file containing the columns “year”, “global\_average\_temp” and “city\_average\_temp”.
* After downloading the CSV file, I performed analysis in three ways i.e. using pandas with Python, Excel and Tableau.

**Analysis using Python and Pandas**

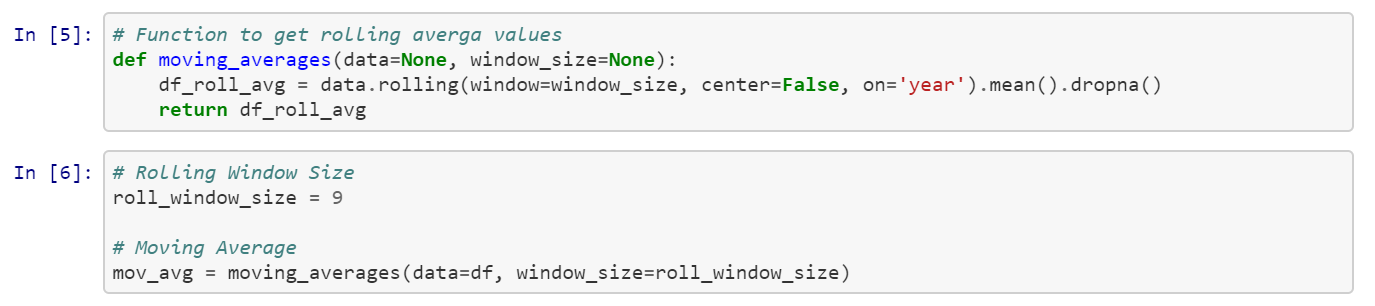
* For this step, I downloaded the CSV file and imported that in my Jupyter Notebook using Pandas as follows:



* After reading in the data, I first checked the data for any imbalance in classes i.e. missing values in data using the following:



* From the above tables, I could see that the “year” and “global\_avg\_temp” had same number of values but the column for “city\_avg\_temp” had 7 missing values. So, in order to balance the data, I decided to drop the rows with “NaN” for the “city\_avg\_temp” values.
* I wrote a function to calculate the **Moving Averages** for “global\_avg\_temp” and “city\_avg\_temp” columns to see the trend in the trends in temperature readings over years using rolling function from pandas as follows:



* In above function, the inputs to the “moving\_averages” function are “data” i.e. the original dataframe and “window\_size” i.e. size of the moving window. This is the number of observations used for calculating the statistic. Each window is of a fixed size.
* Additionally, this function takes care of dropping the rows for which the “city\_avg\_temp” is a “NaN” by using “dropna()”.
* For this dataset, I tried different values for the rolling window sizes but settled down for a value of 9.