CSCI 5408: Assignment 4

Problem 1

Task:

Business Intelligence (BI) Reporting using Cognos

Analysing the Dataset:

The dataset has been downloaded from the **Hourly Weather Surface – Brazil (Southeast region) – Kaggle** website [1]. The size of the dataset is almost around 2 GB. The dataset contains the weather data from 122 weather stations of southwest regions of Brazil which has been recorded and maintained by the INMET (National Meteorological Institute – Brazil) and been provided to Kaggle. The states included are Rio de Janeiro, São Paulo, Minas Gerais e Espirito Santo. The factors covered in the dataset are elevation, weather station, city, province, date, year, month, day, hour, precipitation etc. considering which a detailed analysis can be performed using any BI tool to derive insightful facts. These derived insightful facts in turn help predicting or even making informed decisions from business perspective.

Dimensions in the Dataset:

For analysing the data, we need to decide upon the dimensions as they are the driving factors that help derive useful insights. So, the dimensions and the reason for choosing them from the weather dataset are described in detail in the below table:

Table 1. Identifies dimensions in weather dataset.

Sr. No.	Dimension	Reason for selecting Dimension
1	Dew point temperature	The reason for selecting dew point temperature as a dimension is because we can derive facts based on max dew, min dew, instant dew on a particular day, hour or even month. It can also be done for city or province.
2	Humidity	The reason for selecting humidity as a dimension is because we can derive various facts based on relative humidity, max humidity and min humidity on a particular day, hour or month. The facts of city and province can also be used to derive insights along with the amount of precipitation caused due to humidity.
3	Location	Considering location as a dimension is because it consists of elevation, latitude and longitude. So, using the facts mentioned above, some interesting insights can be derived like the precipitation at a specified location or temperature at a specific latitude and longitude. The wind speed fact can also derive good information from this dimension.

4	Precipitation	The dimension of precipitation just consists of the precipitation in millimetres (any unit can be used). The facts that could be used to derive this dimension are precipitation on an hourly, daily or monthly basis. What is the precipitation when the wind is very high or low or when solar radiation is high. All those facts give useful insights about the precipitation.
5	Pressure	The air pressure dimension consists of the instant air pressure, max air pressure and min air pressure. The derivations that could be done based on the below mentioned facts could be based on the time i.e. hourly, daily or weekly. It can also be derived based on the solar radiation and wind speed plus humidity.
6	Station	Weather station is considered as a dimension because a detailed analysis can be done based on various facts mentioned below. For example, we can calculate the wind speed or even temperature based on different stations. The precipitation or even humidity can be measured. This dimension is described using weather station id, station name and station number.
7	Temperature	Temperature consists of instant air temperature, max temperature and min temperature which is identified as a dimension because we can derive the average temperature at high, medium or low wind speed. For the precipitation fact, we can derive the different ranges of temperature.
8	Time	Time as a dimension incudes date, time, hr, day and month. So, time is considered because facts can be derived on a yearly, hourly, daily or even monthly basis. This analysis would be the most important from a long-term standpoint.
9	Wind	The wind dimension is identified because it consists of wind speed, wind direction and wind gust. The derivations could be done based on the time fact, dew fact and even the solar radiation fact.

Transforming and Cleaning the dataset:

1. All the blank values in columns precipitation, air pressure, solar radiation and air temperature were replaced with 0.
2. There are no NULL or NAN values present in the dataset considered.
3. I will create separate CSV files for the Dimensions considered from the weather dataset provided.
4. Separate CSV file for facts is also created.

Facts:

The Facts are those measurable fields that are derived for various different dimensions present in the database. Selecting these facts assist in having a clear idea of what parameters are we going to perform analysis that will help in making informed decisions or predictions. Facts are nothing but metrics used for dimensions in various BI tools. The various facts which I have identified for the available weather dataset are as follows:

1. Precipitation
2. Air pressure
3. Solar radiation
4. Air temperature
5. Dew point temperature
6. Humidity
7. Wind speed
8. Wind direction
9. Wind gust
10. Time

The reason for choosing these fields as facts from the dataset is because various calculations can be performed for a particular city, state, province, year, month, day, hour etc. based on them.

NOTE:
After analysing the dataset I took a subset of the dataset with 500,000 rows and performed all the operations on it. Taking a subset of the database was necessary as IBM Cognos does not allow the entire file to be uploaded.

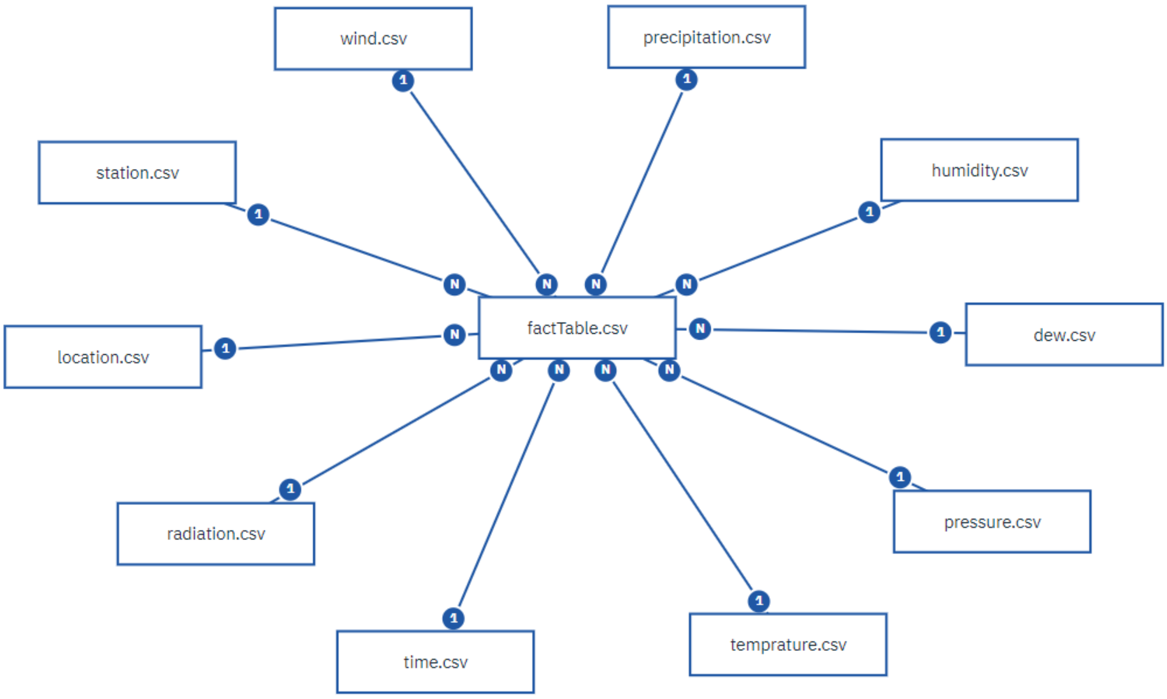

Figure 1. Star Schema

The above Star diagram showcases the relationship, cardinality and the dimensions selected. The **factTable** is a new file that consists of the ids of all the dimensions being considered. All the dimensions have a one-to-many relationship with the central main file of **factTable.** The Star Schema is a schema where there is a single central node connected to multiple nodes, as it can be seen in the above image. The model is mainly created by keeping in mind the above-mentioned dimensions and measurable facts.

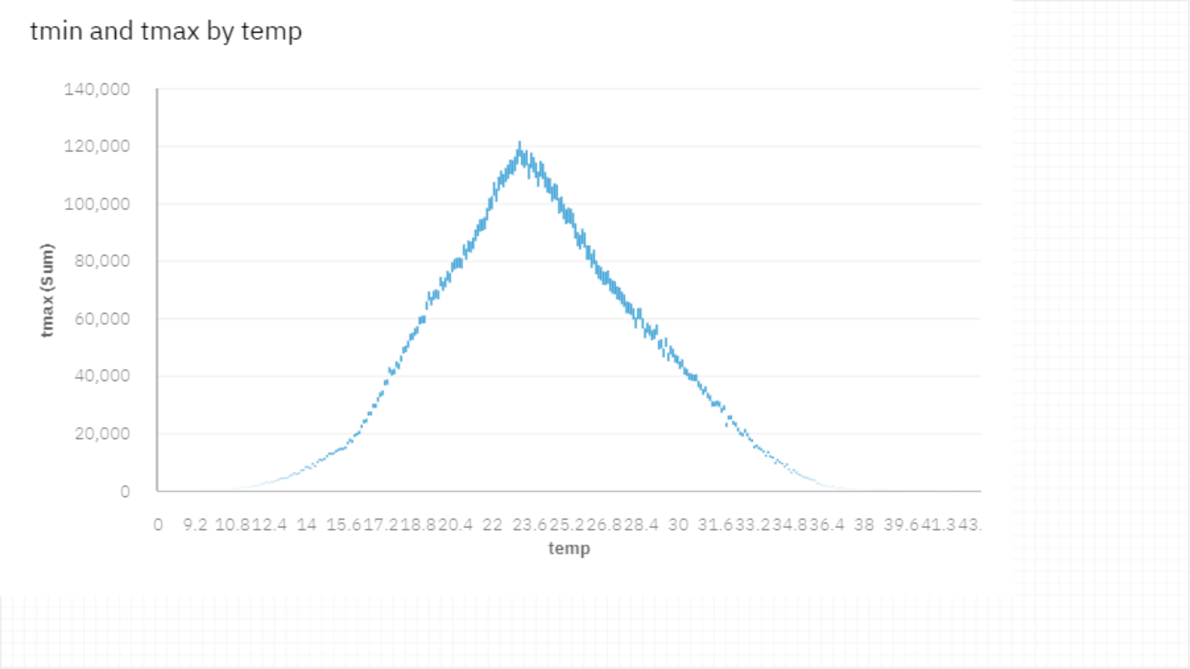

Figure 2. Visualisation 1

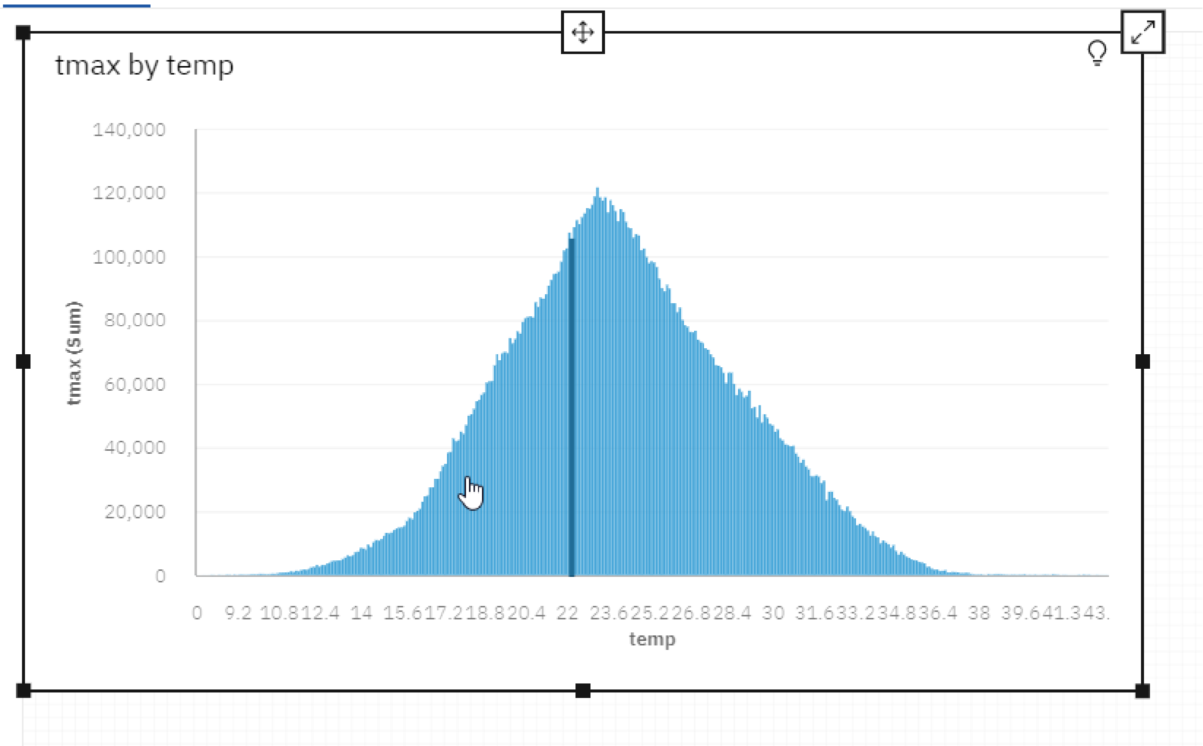

Figure 3. Visualisation 2

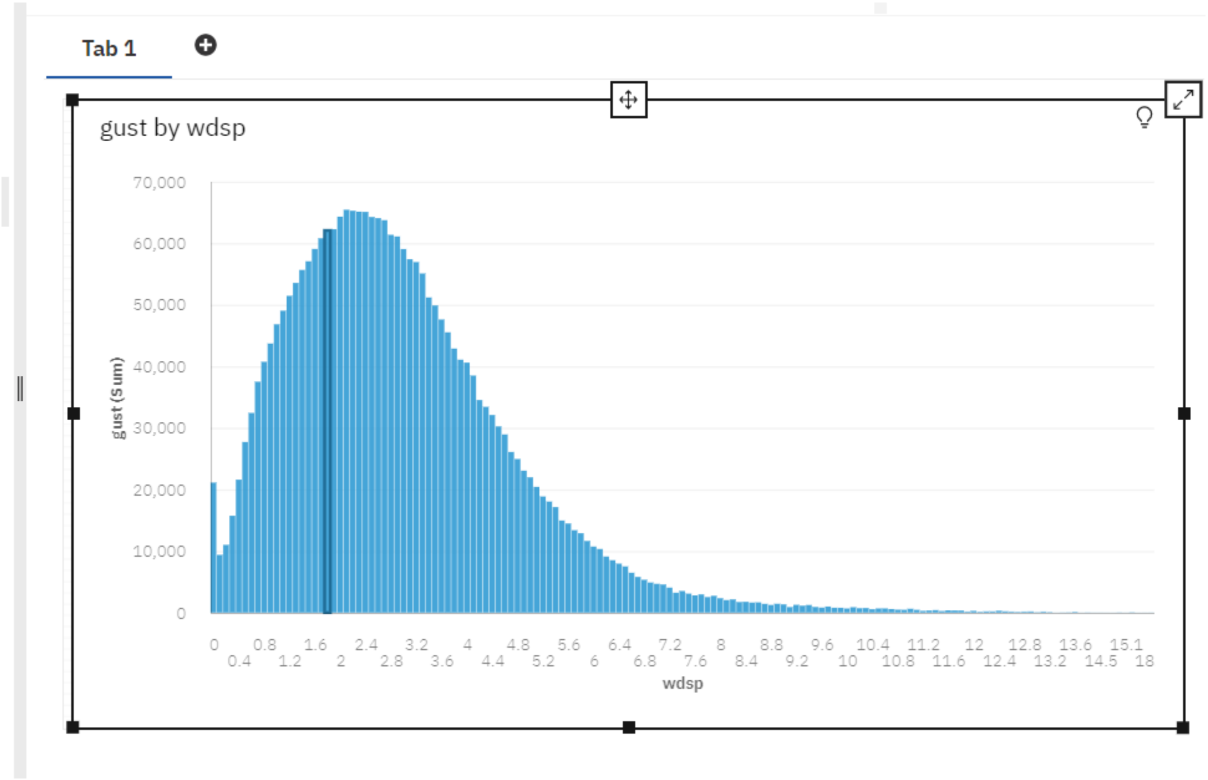

Figure 4. Visualisation 3

Problem 2 & 3:

GitLab Link: https://git.cs.dal.ca/drshah/csci-5408-s2021-b00870600-dhrumil-shah/-/tree/master/A4

The Assignment 4 coding solutions are available inside the **A4** folder in **master** branch in the above GitLab link

References:

[1] Kaggle, "Hourly Weather Surface - Brazil (Southeast region)," NMET (National Meteorological Institute - Brazil), 2021. [Online]. Available: https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region.