CSCI5408 – ASSIGNMENT 4

Dhrumil Amish Shah (B00857606)
dh416386@dal.ca

Problem 1

Task Description

Business Intelligence Reporting using Cognos

Measurable fields and dimensions

This dataset ctontains climate data from a total of 112 weather stations recorded on an hourly basis. Weather stations are in the southeast region of Brazil. The source of this dataset is INMET - National Meteorological Institute situated in Brazil [1].

Dimensions

Dimensions selected for the given dataset with the reason for selecting them are described in the table below:

Table 1: Dimensions in the weather dataset

Dimension	Selection Reason
Weather station	Precipitation, temperature, dew point temperature, solar radiation, pressure, humidity and wind can be derived for every weather station from the dataset which makes weather station, one of the dimensions necessary for calculating key measurable fields. Fields "wsid", "wsnm", "elvt", "lat", "lon" and "inme" describes a weather station.
City	Facts like precipitation, temperature, dew point temperature, solar radiation, pressure, humidity and wind can be calculated from the dataset for every city. Hence the city is chosen as a dimension.
Province	Precipitation, temperature, dew point temperature, solar radiation, pressure, humidity and wind can be derived for every province from the dataset. Hence province is one of the key factors for deriving facts.
Observation datetime	Observation datetime is chosen as a dimension because measurable fields like precipitation, temperature, dew point temperature, solar radiation, pressure, humidity and wind can be analyzed for a particular date and time.
Date	Facts can be derived for a particular date which makes date a dimension.
Year	Facts can be derived for a particular year which makes the year a dimension.
Month	Precipitation, temperature, dew point temperature, solar radiation, pressure, humidity and wind can be derived on monthly basis.
Day	Weather analysis can be done on a specific day. Hence the day is considered as one of the factors for measuring data.
Hour	Weather analysis can be done on an hourly basis. Hence hour is considered as one of the factors for measuring weather data.

Facts

The measurable fields for this dataset are precipitation, air pressure, solar radiation, air temperature, dew point temperature, humidity, and wind parameters. These fields can be derived from the dimensions mentioned in Table 1. Analysis of weather data can be done for each of these measurable fields with respect to the dimensions. For example, facts can be calculated on an hourly, monthly, and yearly basis, for a particular day or date and even for a particular datetime. Further, analysis can also be done for a particular weather station, city and province.

Dataset cleaning

1. Deleted rows with 0 and blank readings as it provides no information. Total records deleted – 47,364.

2. For records where data is not available, I added -1 that indicates not available during analysis.
    - Filled 911395 cells for the column name "prcp".
    - Filled 420664 cells for the column name "gbrd".
    - Filled 17 cells for the column name "dewp".
    - Filled 30 cells for the column name "dmin".
    - Filled 23569 cells for the column name "wdsp".
    - Filled 4979 cells for the column name "gust".

3. The given dataset is broken down into nine identified dimensions for which a separate CSV file is created. Further, cleaning is performed by retaining only the unique values for the identified dimensions of the weather dataset.
    - Created a separate CSV file for the dimension weather station and removed duplicate entries. Total unique entries – 13.
    - Created a separate CSV file for the dimension city and removed duplicate entries. Total unique entries – 13.
    - Created a separate CSV file for the dimension province and removed duplicate entries. Total unique entries – 3.
    - Created a separate CSV file for the dimension observation datetime and removed duplicate entries. Total unique entries – 120321.
    - Created a separate CSV file for the dimension date and removed duplicate entries. Total unique entries – 5028.
    - Created a separate CSV file for the dimension year and removed duplicate entries. Total unique entries – 15.
    - Created a separate CSV file for the dimension month and removed duplicate entries. Total unique entries – 12.
    - Created a separate CSV file for the dimension day and removed duplicate entries. Total unique entries – 31.
    - Created a separate CSV file for the dimension hour and removed duplicate entries. Total unique entries – 24.

Star Schema

The star schema for the dataset has been created using the dimensions and facts as listed above. The dimensions are selected after analysing data to derive entities that can help us gain measurable fields that are called facts.

Each of the dimensions have a 1:N relationship with the fact table as there can be 1 or more rows for a dimension in the fact table. The fact table used here is **sudeste.csv** which consists of measurable data for all the dimensions whereas the data for dimensions in uploaded using the CSV files.

Figure 1 displays the star schema for the given weather dataset.

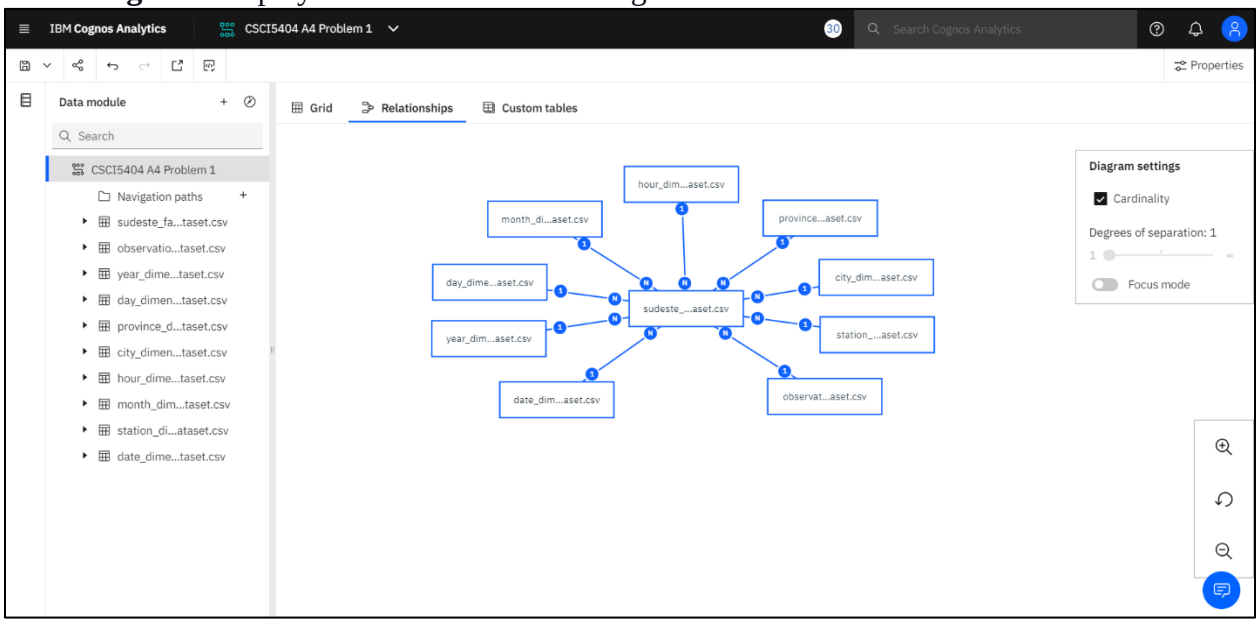

Figure 1 – Star Schema [2]

After uploading the data and establishing relationships, using the **Exploration** option, we can visualize the uploaded data into bar graphs, charts, line graphs, and so on.

Visualizations

1. Bar graph

   Figure 2 displays a bar graph visualizing the average temperature, average minimum temperature and average maximum temperature by province and year displaying the statistics for the years 2015 and 2016 for every province.

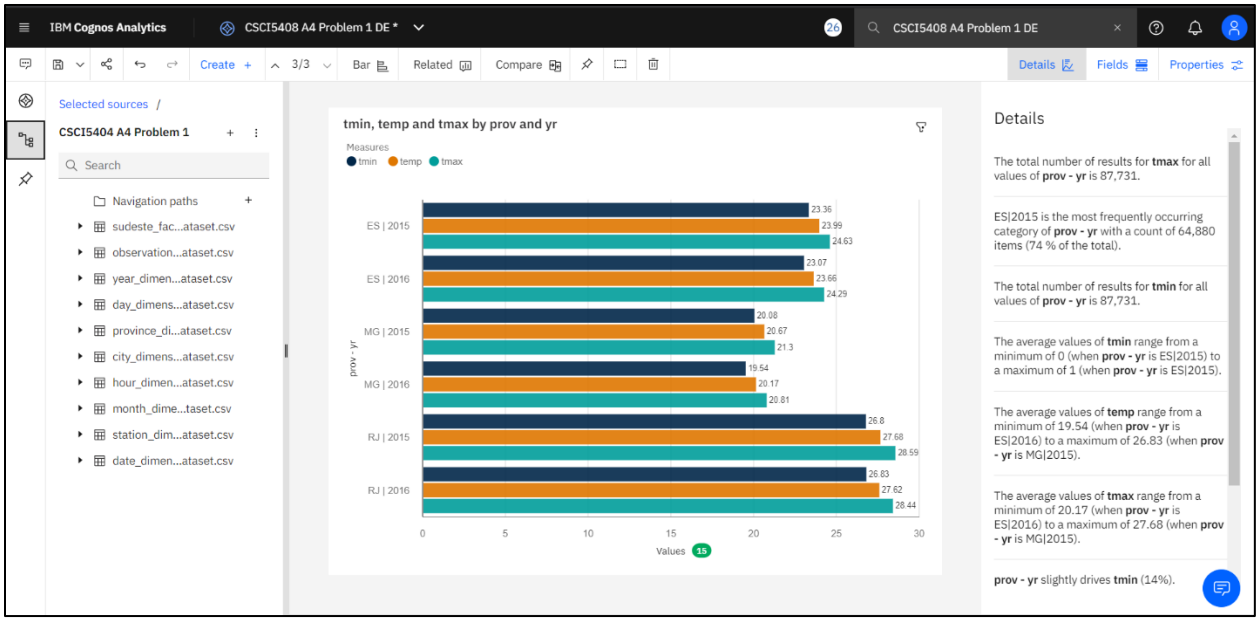

Figure 2 – Average tmin(minimum air temperature in °C), temp(instant air temperature in °C), and tmax(maximum air temperature in °C) for all three provinces(ES, MG, and RJ) in the year 2015 and 2016 [2]

Figure 3 displays a bar graph visualizing the average temperature, average minimum temperature and average maximum temperature by province ES and year 2016 displaying the statistics for for every city in the province.

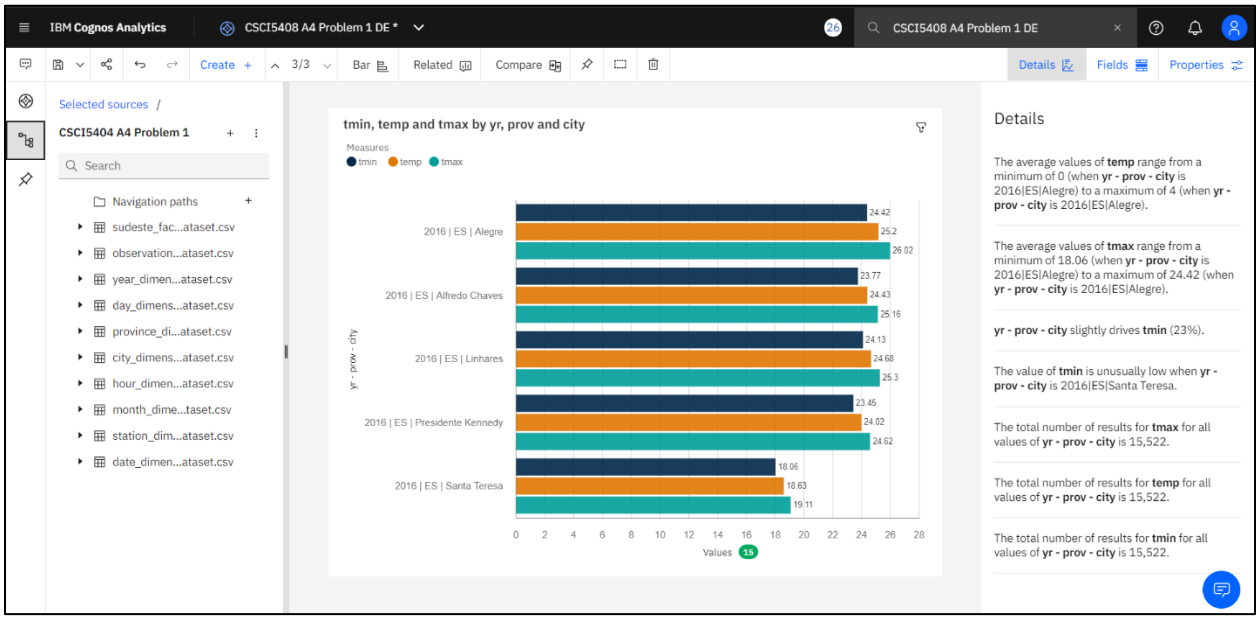

Figure 3 – Average tmin(minimum air temperature in °C), temp(instant air temperature in °C), and tmax(maximum air temperature in °C) for all cities in province ES and the year 2016 [2]

2. Line graph

Figure 4 displays a line graph visualizing the average humidity, average minimum humidity and average maximum humidity by date (date range: 01-Dec-2006 to 31-Dec-2006) and city Alegre.

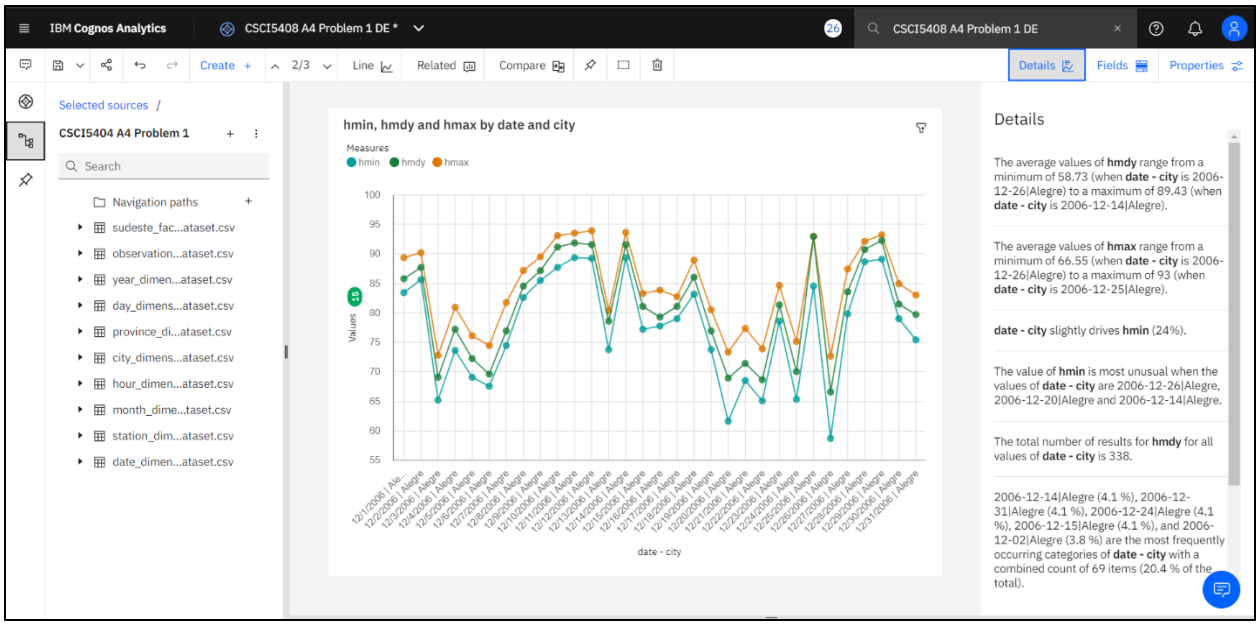

Figure 4 - Average hmin(minimum relative humid temperature in %), hmdy(relative humid temperature in %), and hmax(maximum relative humid temperature in %) for city Alegre between dates Dec 01, 2006, and Dec 31, 2006 [2]

3. Stacked Bar graph

Figure 5 displays a bar graph visualizing the average wind gust and average wind speed for all cities in the year 2016.

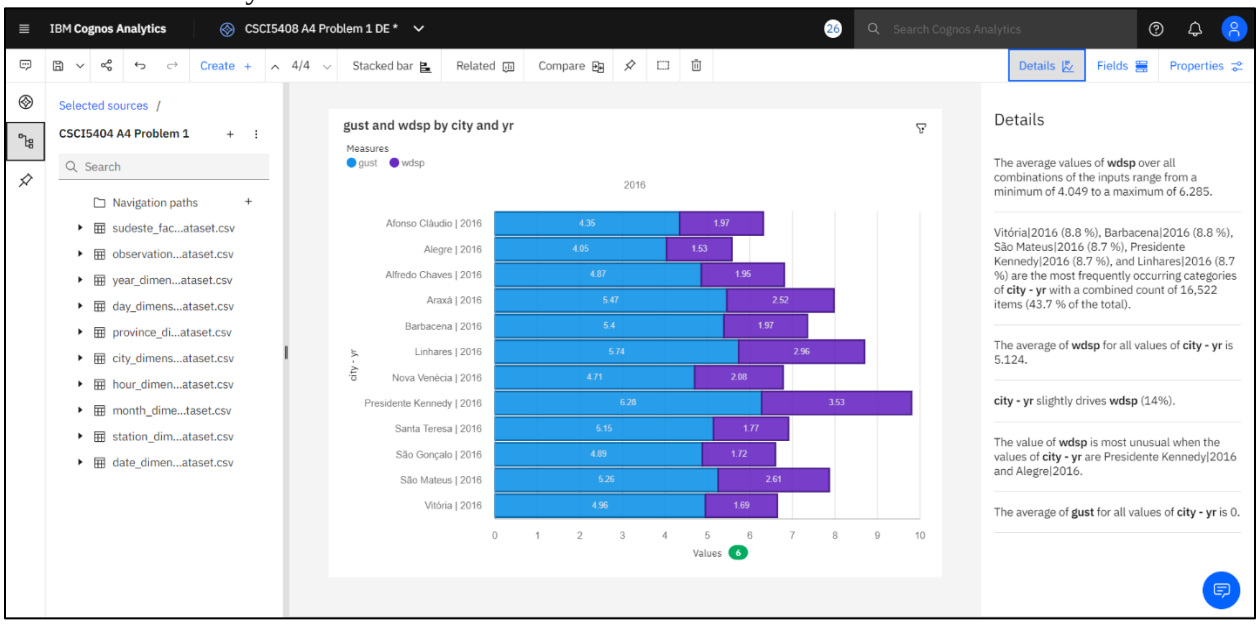

Figure 5 – Average gust(gust speed in m/s) and wdsp(wind speed in m/s) for all cities in the year 2016 [2]

Main Java Code
https://git.cs.dal.ca/dashah/csci-5408-s2021-b00857606-dhrumil-amish-shah/-/tree/master/A4

Problem 2

Task Description
Sentiment Analysis

Link to the JAVA files
https://git.cs.dal.ca/dashah/csci-5408-s2021-b00857606-dhrumil-amish-shah/-/tree/master/A4/src/main/java/problem2

Link to the output file
https://git.cs.dal.ca/dashah/csci-5408-s2021-b00857606-dhrumil-amish-shah/-/blob/master/A4/src/main/java/problem2/problem_2_output.txt

Problem 3

Task Description
Semantic Analysis

Link to the JAVA files
https://git.cs.dal.ca/dashah/csci-5408-s2021-b00857606-dhrumil-amish-shah/-/tree/master/A4/src/main/java/problem3

Link to the output file
https://git.cs.dal.ca/dashah/csci-5408-s2021-b00857606-dhrumil-amish-shah/-/blob/master/A4/src/main/java/problem3/problem_3_output.txt

References

[1] INMET, "Hourly Weather Surface - Brazil (Southeast region)," National Meteorological Institute - Brazil, [Online]. Available: https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region. [Accessed 27 July 2021].

[2] IBM, "IBM Cognos Analytics," IBM, [Online]. Available: https://www.ibm.com/ca-en/products/cognos-analytics. [Accessed 27 July 2021].