

Report

Assignment No. – 4

Problem – 1:

Step -1:

I downloaded the weather dataset available on <https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region?select=sudeste.csv>

After downloading the dataset, there were various datasets based on different regions as given below:

Datasets:

1. Central_west
2. North
3. Northeast
4. Southeast
5. Stations

The dataset I chose was for the region “central_west”.

Step – 2:

I explored the dataset and identified the data fields that could be measured by certain factors or dimensions. Below I mentioned dimensions and facts.

Identifying facts and dimensions:

Dimensions:

Below is the dimension selected for the central_west dataset:

Table 1: Dimesions with its reason of selection

Dimension	Reason for Selection
Date	Date provides descriptive information of the facts like temperature, precipitation, dew point temperature on that particular date.
Hour	Calculations about temperature or weather conditions can be calculated based on the hour. That means, information regarding weather for that particular hour can be determined.

Longitude	Longitude gives the description of the place where the weather condition took place. The units of weather like temperature, solar radiation etc can be analyzed based on the longitude of that particular place.
Latitude	Latitude gives the description of the place where the weather condition took place. The units of weather like temperature, solar radiation etc can be analyzed based on the latitude of that particular place.
Region	Region gives the description of the place where the weather condition took place. The units of weather like temperature, solar radiation etc can be derived for that region.
State	Region gives the description of the place where the weather condition took place. The units of weather like temperature, precipitation, dew point temperature etc can be derived for that region.
Station_code	Station_code gives the description of the place where the weather condition took place. The units of weather like temperature, precipitation, dew point temperature etc can be derived for that particular place having the station_code.
Station	Station gives the description of the place where the weather condition took place. The units of weather like temperature, precipitation, dew point temperature etc can be derived for that Station.

Facts:

After analyzing the dataset, I understood that the data provided is for different weather conditions and because it's a brazil's dataset, the language of the column names is in brazil. So I provided below the column description. Facts provide us with the measurements or metrics like values for temperature, dew point temperature, solar radiation , relative humidity etc for the particular dimension.

Column Description:

Table 2: Column Description for Central west Brazil Dataset

Column_index	columns_pt-br	columns_en	abbreviation	description
0	data	date	Date	date (YYYY-MM-DD)
1	hora	hour	Hr	hour (HH:00)
2	precipitacao total,horario (mm)	total precipitation (mm)	Prcp	Amount of precipitation in millimetres (last hour)
3	pressao atmosferica ao nivel da estacao (mb)	atmospheric pressure at station height (mb)	Stp	Atmospheric pressure at station level (mb)
4	pressao atmosferica max. na hora ant. (aut) (mb)	atmospheric pressure max. in the previous hour (mb)	Smax	Maximum air pressure for the last hour in hPa to tenths
5	pressao atmosferica min. na hora ant. (aut) (mb)	atmospheric pressure min. in the previous hour (mb)	smin	Minimum air pressure for the last hour in hPa to tenths
6	radiation (kj/m2)	radiation (kj/m2)	Gbrd	Solar radiation KJ/m2
7	temperatura do ar - bulbo seco ($^{\circ}\text{C}$)	air temperature - dry bulb ($^{\circ}\text{C}$)	Temp	Air temperature (instant) in celsius degrees
8	temperatura do ponto de orvalho ($^{\circ}\text{C}$)	dew point temperature ($^{\circ}\text{C}$)	Dewp	Dew point temperature (instant) in celsius degrees

9	temperatura maxima na hora ant. (aut) ($\hat{A}^{\circ}\text{C}$)	max. temperature in the previous hour ($\hat{A}^{\circ}\text{C}$)	Tmax	Maximum temperature for the last hour in celsius degrees
10	temperatura minima na hora ant. (aut) ($\hat{A}^{\circ}\text{C}$)	min. temperature in the previous hour ($\hat{A}^{\circ}\text{C}$)	Tmin	Minimum temperature for the last hour in celsius degrees
11	temperatura orvalho max. na hora ant. (aut) ($\hat{A}^{\circ}\text{C}$)	dew temperature max. in the previous hour ($\hat{A}^{\circ}\text{C}$)	Dmax	Maximum dew point temperature for the last hour in celsius degrees
12	temperatura orvalho min. na hora ant. (aut) ($\hat{A}^{\circ}\text{C}$)	dew temperature min. in the previous hour ($\hat{A}^{\circ}\text{C}$)	Dmin	Minimum dew point temperature for the last hour in celsius degrees
13	umidade rel. max. na hora ant. (aut) (%)	relative humidity max. in the previous hour (%)	Hmax	Maximum relative humid temperature for the last hour in %
14	umidade rel. min. na hora ant. (aut) (%)	relative humidity min. in the previous hour (%)	Hmin	Minimum relative humid temperature for the last hour in %
15	umidade relativa do ar, horaria (%)	air relative humidity (%)	Hmdy	Relative humid in % (instant)
16	vento direcao horaria (gr) (\hat{A}° (gr))	wind direction (\hat{A}° (gr))	Wdct	Wind direction in radius degrees (0-360)

17	vento rajada maxima (m/s)	wind rajada maxima (m/s)	gust	Wind gust in metres per second
18	vento velocidade horaria (m/s)	wind speed (m/s)	Wdsp	Wind speed in metres per second
19	Region	region		Brazilian geopolitical regions
20	State	State	Prov	State (Province)
21	Station	Station	Wsnm	Name station (usually city location or nickname)
22	Station_code	Station_code	Inme	Station number (INMET number) for the location
23	Latitude	Latitude	Lat	Latitude
24	Longitude	Longitude	Lon	Longitude
25	Height	height	elvt	Elevation

Step-3)

Cleaning and Spreadsheet Filtration for Facts:

- The total records in the central_west datasets was 1048576. I noticed there were -9999 values under the columns from column_index : 2 to 18 (Refer table 1).
- I deleted the entire row containing the values -9999 from column having column_index: 2 to 18 because the values for temperature can not be -9999. So to maintain the consistency of the data, I removed the value.
- So total records removed were 116070 of 1048576.
- I removed blanks from 932506 records, because there was no value in it and the data consistency needs to be maintained.
- There were scattered -9999 values across columns from column_index: 2 to 18 where the corresponding values were not -9999. So in this case, we can't remove the entire row because data for that particular column is necessary. So I replaced -9999 values with 0.
- Now, the total records of the filtered dataset is 932506.
- So the filtered CSV file we received for central_west is the fact.

I created separate CSV files for dimensions and facts.

Files for Dimensions: I created file for each dimension as mentioned below:

1. Date_dimention.csv
2. Hour_dimension.csv

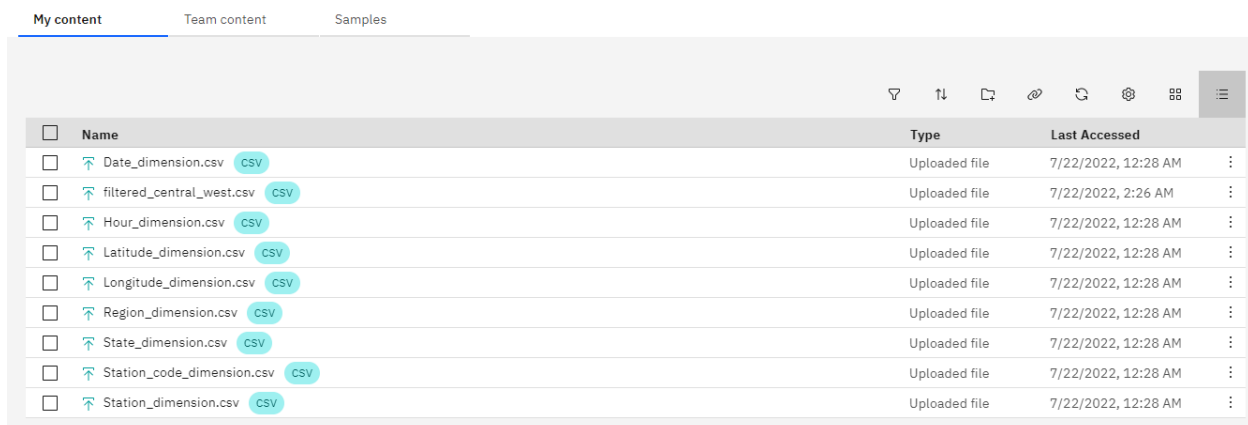
3. Latitude_dimension.csv
4. Longitude_dimension.csv
5. Region_dimension.csv
6. State_dimension.csv
7. Station_code_dimension.csv
8. Station_dimension.csv

File for facts: I created is the file I obtained after cleaning and filtering datasets of the spreadsheet (central_west).

1. Filtered_central_west.csv

Step-4)

After creating the dimension and fact CSV files, I created my Cognos account on <https://myibm.ibm.com/dashboard/>. I imported the above dimension and fact files as shown in the



The screenshot shows the IBM Cognos dashboard interface. At the top, there are tabs for 'My content', 'Team content', and 'Samples'. Below the tabs, there is a table listing imported files. The table has columns for 'Name', 'Type', and 'Last Accessed'. Each row represents a CSV file that has been uploaded to the system.

Name	Type	Last Accessed
Date_dimension.csv	Uploaded file	7/22/2022, 12:28 AM
filtered_central_west.csv	Uploaded file	7/22/2022, 2:26 AM
Hour_dimension.csv	Uploaded file	7/22/2022, 12:28 AM
Latitude_dimension.csv	Uploaded file	7/22/2022, 12:28 AM
Longitude_dimension.csv	Uploaded file	7/22/2022, 12:28 AM
Region_dimension.csv	Uploaded file	7/22/2022, 12:28 AM
State_dimension.csv	Uploaded file	7/22/2022, 12:28 AM
Station_code_dimension.csv	Uploaded file	7/22/2022, 12:28 AM
Station_dimension.csv	Uploaded file	7/22/2022, 12:28 AM

Figure 1: Imported files to Cognos IBM

Star Schema:

I created star schema using fact and dimension. In star schema, Fact table is always placed at the centre. Whereas dimension table is placed at the edge of the star or snowflake schema. In the figure 2 & 3, you'll see that the dimensions table is having 1:N relationship with the fact table. Because the primary key in fact table is mapped as foreign key in dimensions table [1].

Before creating the star schema on IBM Cognos, I build the (Star schema) dimension modelling using a drawing tool like draw.io. Refer Figure 2.

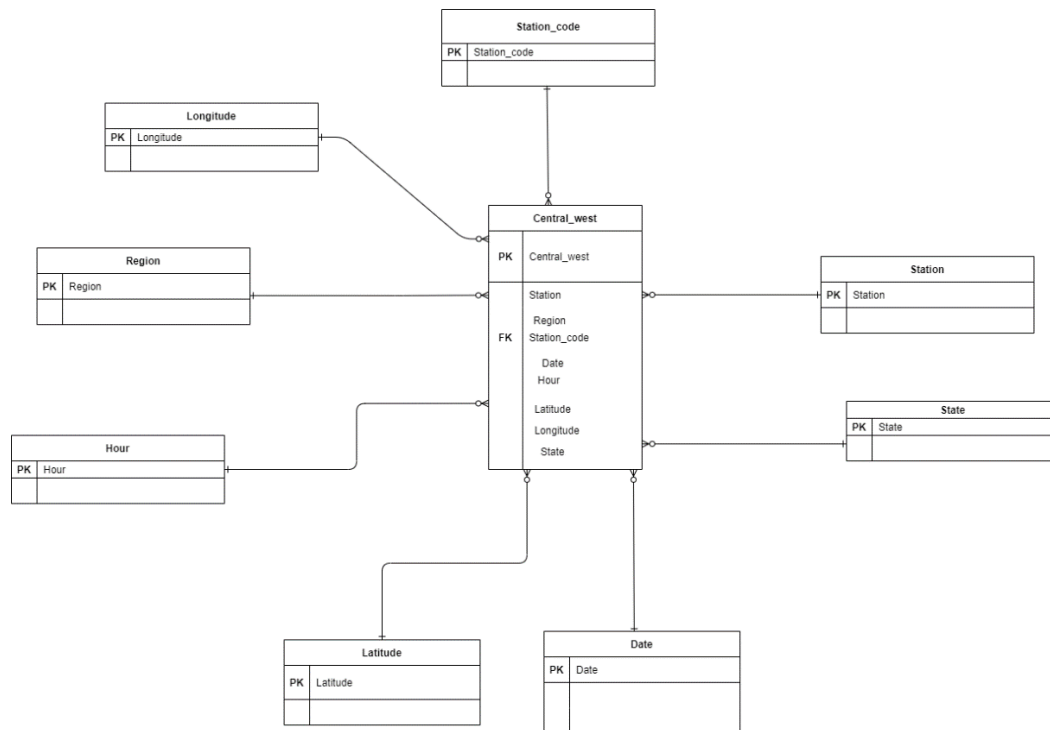


Figure 2: Star schema using draw.io

After analyzing and mapping the relationship between the dimension and fact on draw.io. I build the schema on IBM Cognos as shown in Figure 4 & 5. I created the relationship between the fact table and dimension table as shown in the figure 3.

Create relationship

Table 1

Station_code_dimension.csv

station_code
A047
A757
A929
A002

Table 2

filtered_central_west.csv

#	Row Id	station_code
1	138998	A047
2	138999	A757
3	139000	A929

Match selected columns

Row Id	station_code
1	A047
2	A757
3	A929

Row Id	index	Data	Hora
1	138998	2017-12-20	14:00:00
2	138999	2017-12-20	15:00:00
3	139000	2017-12-20	16:00:00

Figure 3: Mapping relationship between Fact and Dimension.

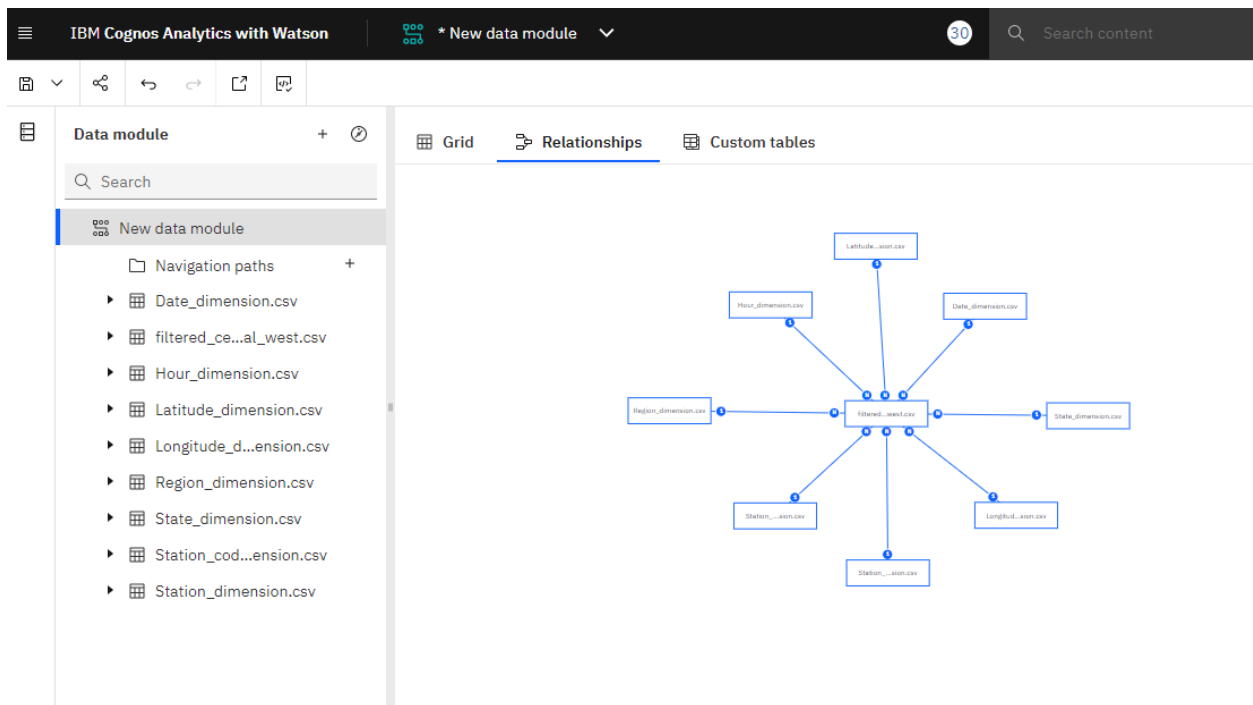


Figure 4: Star Schema showing relationship between dimensions and facts on IBM Cognos

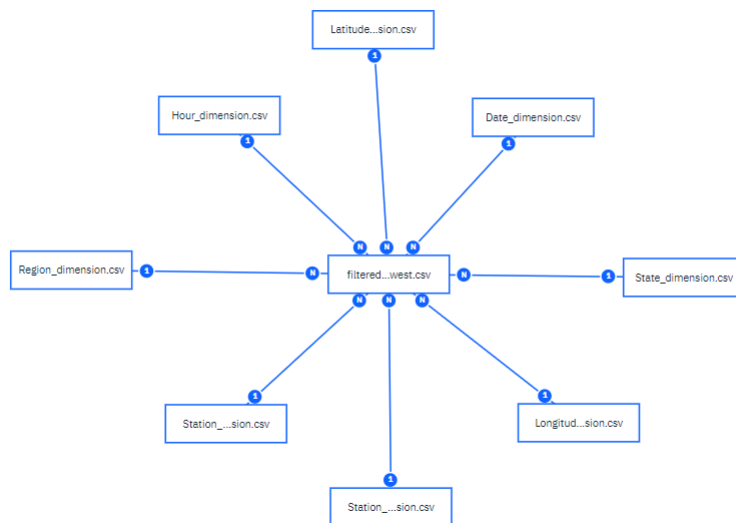


Figure 5: Zoomed image for Star Schema

After creating Star Schema, we can explore the data visualisations. I created few data visuals using bar chart, pie charts etc through “Present Data” tab. Refer Figure 6.

^ Quick launch

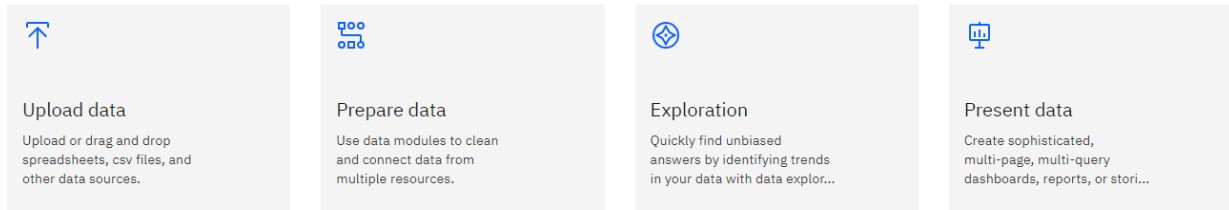


Figure 6: Present data tab on IBM Cognos

Visual Analytics:

1. Bar Charts:

Figure 7 represents the facts dew temperature max. in the previous hour ($\hat{A}^{\circ}\text{C}$) and dew temperature min. in the previous hour ($\hat{A}^{\circ}\text{C}$) by the dimension station.

Also Refer figure 8 & 9.

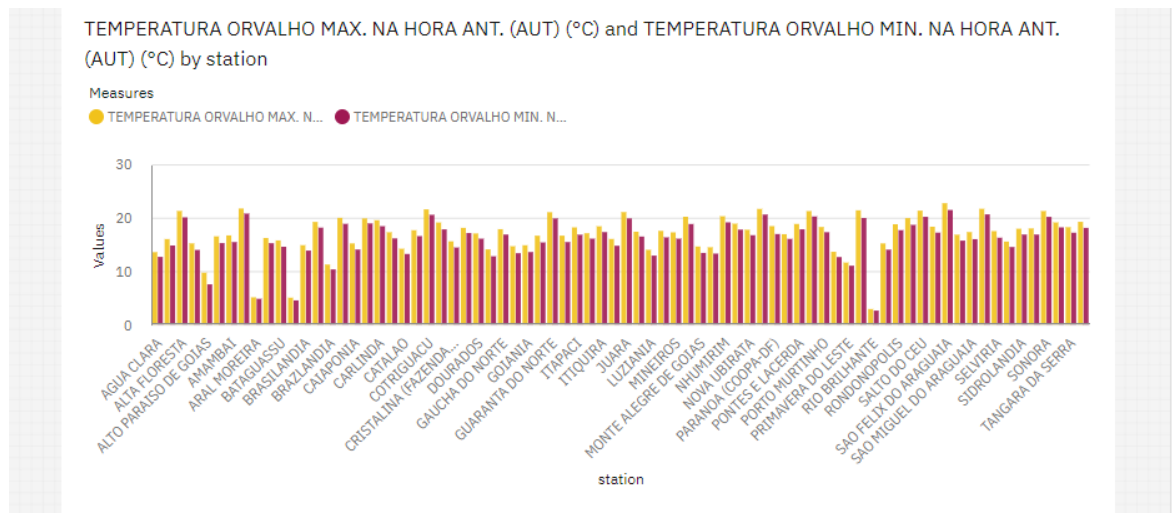


Figure 7: Represents the measures of the Max dew temperature and Min dew temperature by station

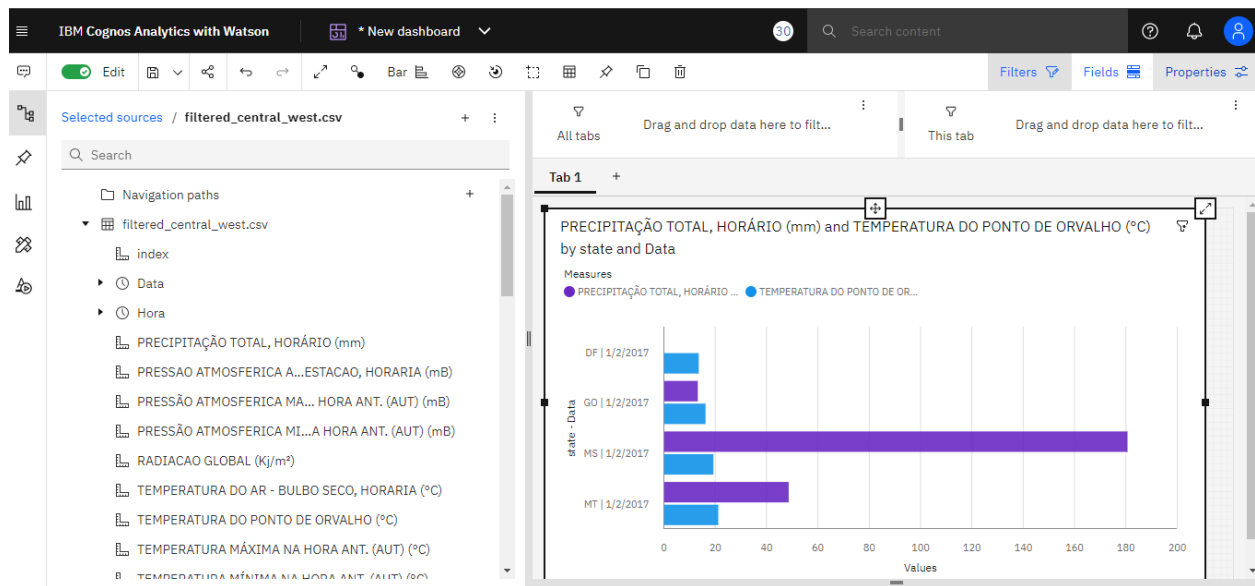


Figure 8: Represents the measures total precipitation (mm) and dew point temperature ($^{\circ}\text{C}$) by state and data.

2. Pie Chart:

Refer figure 9

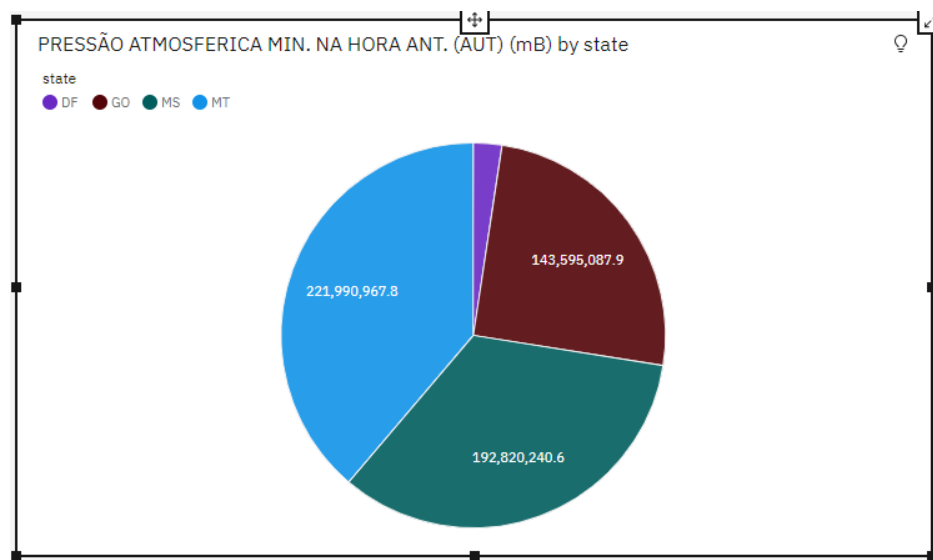


Figure 9: Represents the measure measure of atmospheric pressure min. in the previous hour (mb) by station

3. Stacked Bar:

Refer Figure 10

VENTO, DIREÇÃO HORARIA (gr) (° (gr)), UMIDADE REL. MAX. NA HORA ANT. (AUT) (%) and UMIDADE REL. MIN. NA HORA ANT. (AUT) (%) by Hora

Measures

● VENTO, DIREÇÃO HORARIA (gr) (° ... ● UMIDADE REL. MAX. NA HORA AN...
● UMIDADE REL. MIN. NA HORA AN...

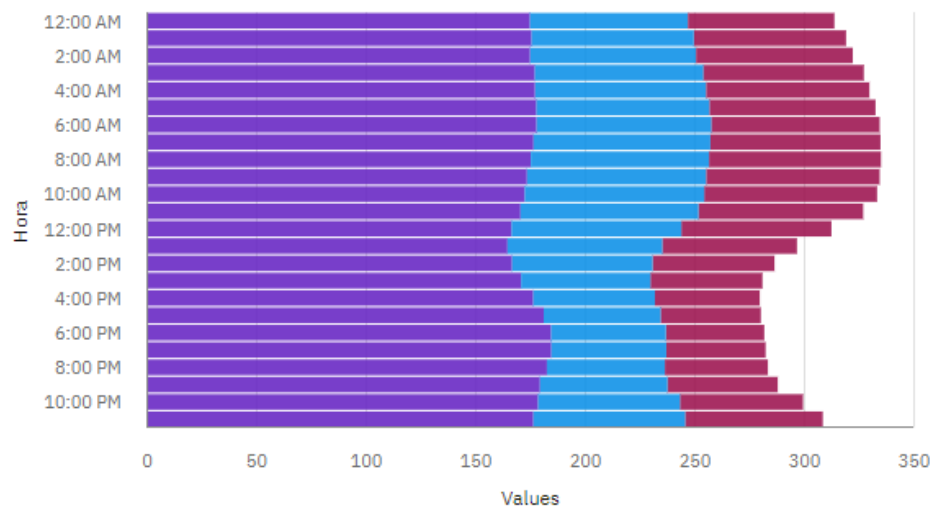


Figure 10: Represents measurements for wind direction (\hat{A}° (gr)), relative humidity max. in the previous hour (%) and relative humidity min. in the previous hour (%) by hour.

References:

- [1] D. Taylor, "Difference between Fact Table and dimension table," *Guru99*, 20-Apr-2020. [Online]. Available: <https://www.guru99.com/fact-table-vs-dimension-table.html>. [Accessed: 24-Jul-2022].
- [2] Brightspace, "Lab8_BI_IBM_Cognos," Brightspace. [Online]. Available: <https://dal.brightspace.com/d2l/le/content/221749/viewContent/3063666/View>. [Accessed: 20-Jul-2022].