CSCI 5408 – Data Management, Warehousing, Analytics

Assignment 5

Work done by,

Name: Guturu Rama Mohan Vishnu

Banner ID: B00871849

Email: rm286720@dal.ca

DECLARATION

I, Guturu Rama Mohan Vishnu, declare that in assignment 3 of CSCI 5408 course, I declare that all the work done was done by myself and I have not collaborated with anyone for the assignment.

Problem #1

Business Intelligence reporting using Cognos

Step 3 : Measurable Facts and Dimensions

Facts are numeric values that represent a specific business aspect. A fact is a metric that can be justified by a number of different factors. A foreign key relationship will exist between each fact table and the dimension table. Fact table is the center of the star schema. Dimensions are the qualifying characteristics that provide additional perspectives to a given fact. Dimension table holds the data in a way that can be used to analyze the data in fact table.

From the dataset I have downloaded from Kaggle, it is identified that the dataset is about the climate weather which covers hourly weather data from various weather stations in Brazil. So, the factor in the schema is the fact_table and the dimension tables are

- air relative humidity,
- air temperature,
- atmospheric pressure at station height,
- atmospheric pressure max in the previous hour,
- atmospheric pressure min in the previous hour,
- date.
- hour,
- dew point temperature,
- index,
- location,
- radiation,
- relative humidity max in the previous hour,

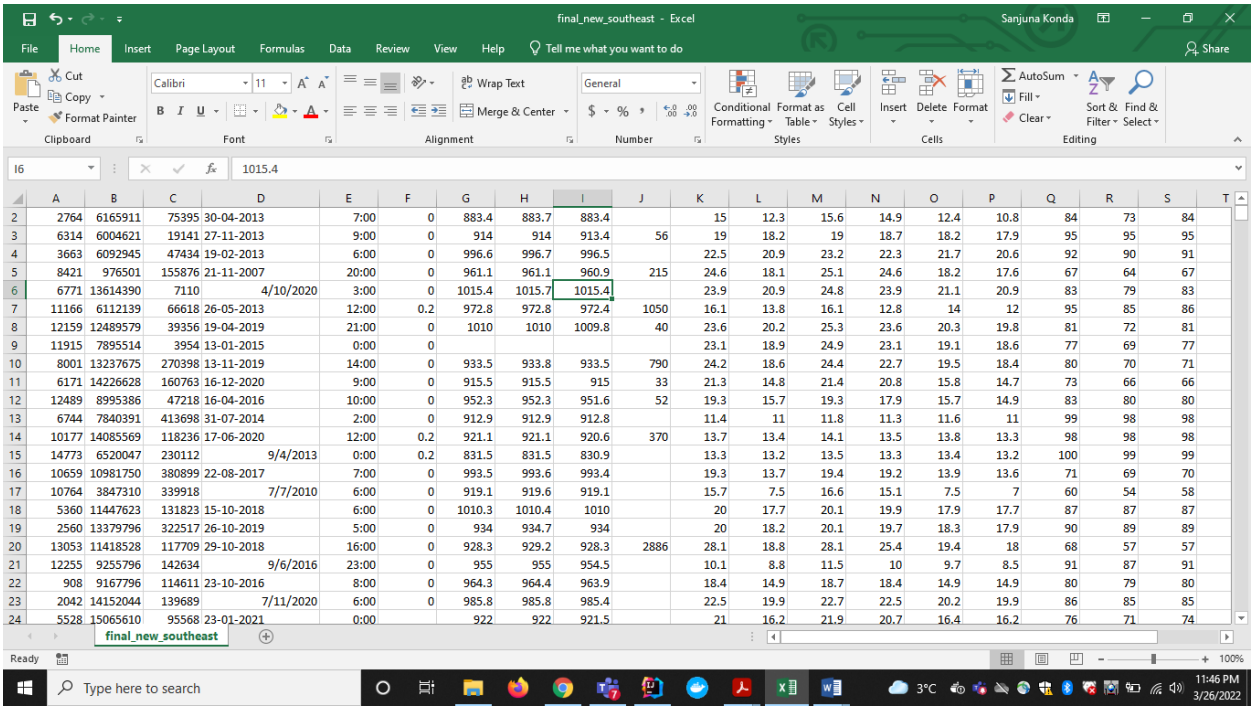
- relative humidity min in the previous hour,
- max temperature in the previous hour,
- min temperature in the previous hour,
- dew temperature max in the previous hour,
- dew temperature min in the previous hour,
- total precipitation,
- wind direction,
- wind rajada maxima and
- wind speed,

Step 4 : Cleaning and Formatting

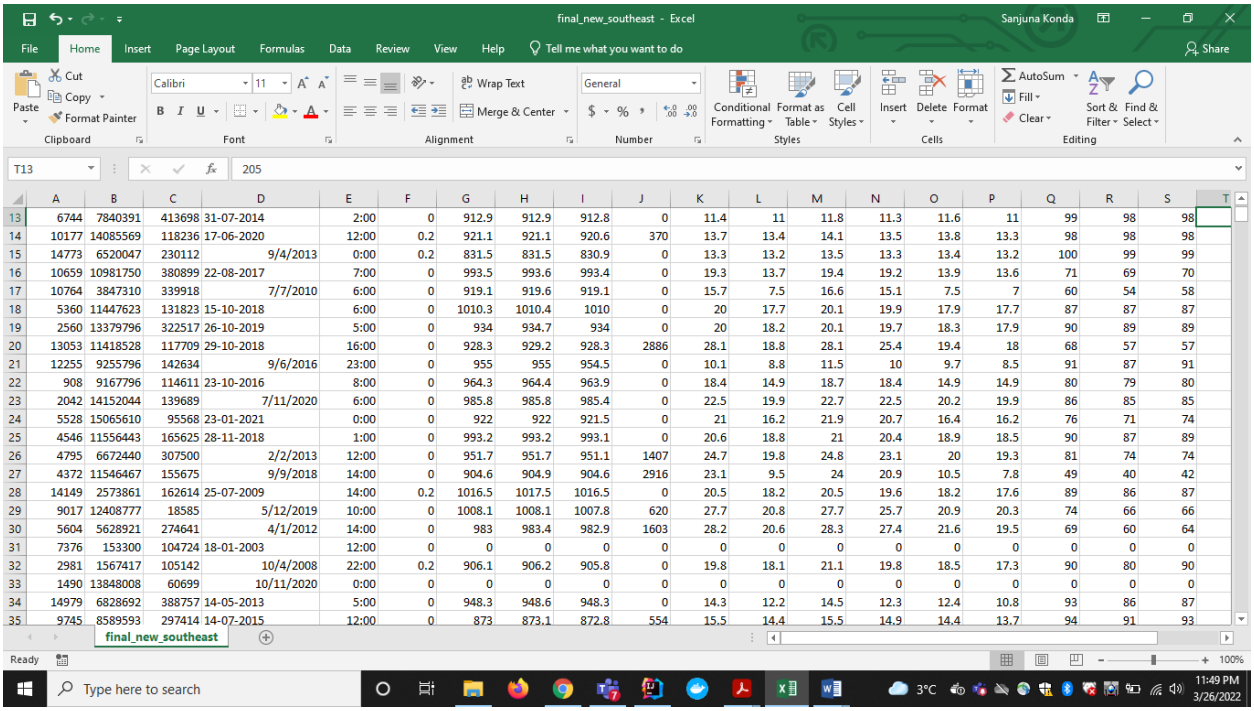
Cleaning the dataset requires multiple steps to be processed and requires multiple ways of cleaning the data. Since the dataset is to large, I have combined all the 5 files and randomized the 50000 rows and extracted 5000 rows out of it and made a new excel sheet.

- To start with, there are many rows with integer type values in it and empty cells in those rows can be replaced by value ‘0’ to maintain the consistency.

  Before cleaning:



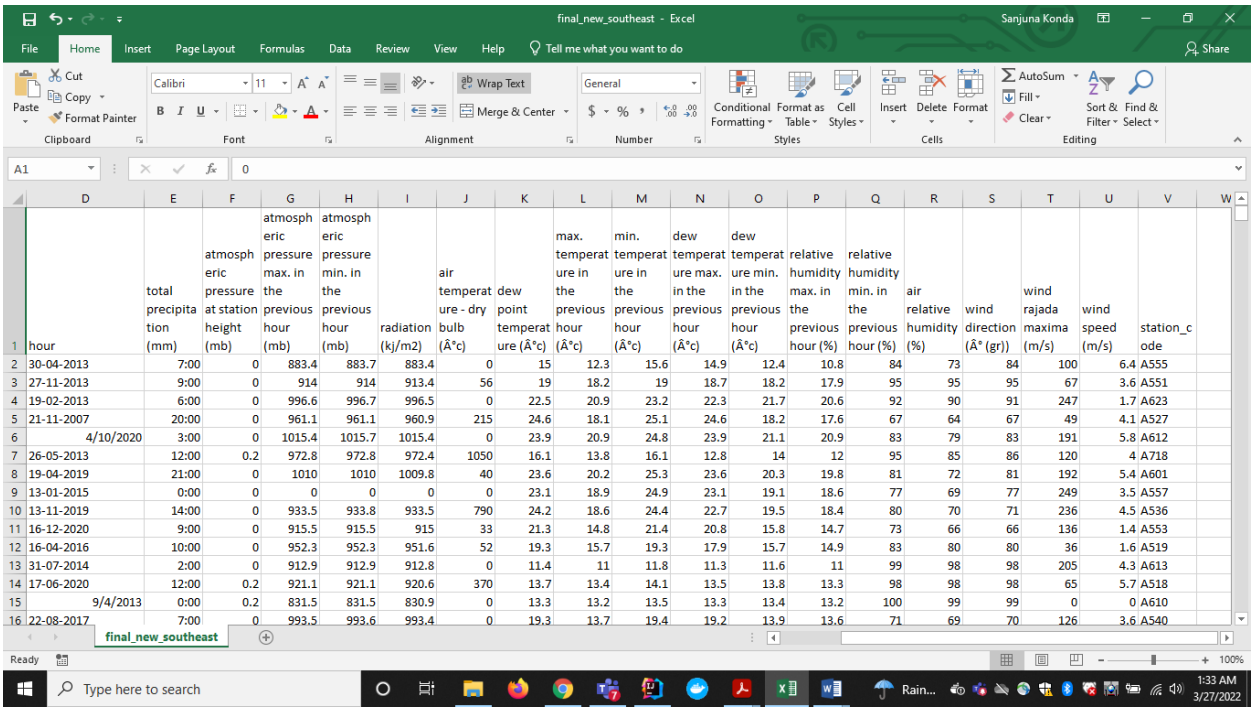
  After cleaning:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
13	6744	7840391	413698	31-07-2014	2:00	0	912.9	912.9	912.8	0	11.4	11	11.8	11.3	11.6	11	99	98	98
14	10177	14085569	118236	17-06-2020	12:00	0.2	921.1	921.1	920.6	370	13.7	13.4	14.1	13.5	13.8	13.3	98	98	98
15	14773	6520047	230112	9/4/2013	0:00	0.2	831.5	831.5	830.9	0	13.3	13.2	13.5	13.3	13.4	13.2	100	99	99
16	10659	10981750	380899	22-08-2017	7:00	0	993.5	993.6	993.4	0	19.3	13.7	19.4	19.2	13.9	13.6	71	69	70
17	10764	3847310	339918	7/7/2010	6:00	0	919.1	919.6	919.1	0	15.7	7.5	16.6	15.1	7.5	7	60	54	58
18	5360	11447623	131823	15-10-2018	6:00	0	1010.3	1010.4	1010	0	20	17.7	20.1	19.9	17.9	17.7	87	87	87
19	2560	13379796	322517	26-10-2019	5:00	0	934	934.7	934	0	20	18.2	20.1	19.7	18.3	17.9	90	89	89
20	13053	11418528	117709	29-10-2018	16:00	0	928.3	929.2	928.3	2886	28.1	18.8	28.1	25.4	19.4	18	68	57	57
21	12255	9255796	142634	9/6/2016	23:00	0	955	955	954.5	0	10.1	8.8	11.5	10	9.7	8.5	91	87	91
22	908	9167796	114611	23-10-2016	8:00	0	964.3	964.4	963.9	0	18.4	14.9	18.7	18.4	14.9	14.9	80	79	80
23	2042	14152044	139689	7/11/2020	6:00	0	985.8	985.8	985.4	0	22.5	19.9	22.7	22.5	20.2	19.9	86	85	85
24	5528	15065610	95568	23-01-2021	0:00	0	922	922	921.5	0	21	16.2	21.9	20.7	16.4	16.2	76	71	74
25	4546	11556443	165625	28-11-2018	1:00	0	993.2	993.2	993.1	0	20.6	18.8	21	20.4	18.9	18.5	90	87	89
26	4795	6672440	307500	2/2/2013	12:00	0	951.7	951.7	951.1	1407	24.7	19.8	24.8	23.1	20	19.3	81	74	74
27	4372	11546467	155675	9/9/2018	14:00	0	904.6	904.9	904.6	2916	23.1	9.5	24	20.9	10.5	7.8	49	40	42
28	14149	2573861	162614	25-07-2009	14:00	0.2	1016.5	1017.5	1016.5	0	20.5	18.2	20.5	19.6	18.2	17.6	89	86	87
29	9017	12408777	18585	5/12/2019	10:00	0	1008.1	1008.1	1007.8	620	27.7	20.8	27.7	25.7	20.9	20.3	74	66	66
30	5604	5628921	274641	4/1/2012	14:00	0	983	983.4	982.9	1603	28.2	20.6	28.3	27.4	21.6	19.5	69	60	64
31	7376	153300	104724	18-01-2003	12:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	2981	1567417	105142	10/4/2008	22:00	0.2	906.1	906.2	905.8	0	19.8	18.1	21.1	19.8	18.5	17.3	90	80	90
33	1490	13848008	60699	10/11/2020	0:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	14979	6828692	388757	14-05-2013	5:00	0	948.3	948.6	948.3	0	14.3	12.2	14.5	12.3	12.4	10.8	93	86	87
35	9745	8589593	297414	14-07-2015	12:00	0	873	873.1	872.8	554	15.5	14.4	15.5	14.9	14.4	13.7	94	91	93

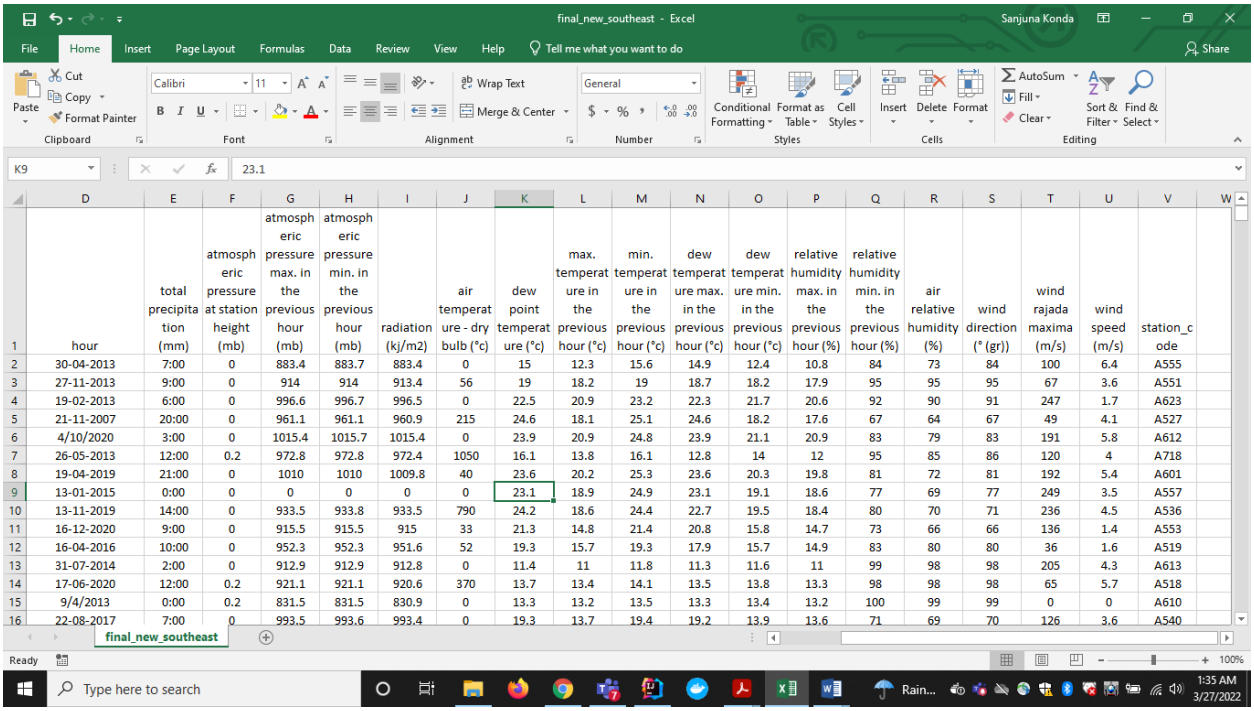
- Next, there are few miscellaneous symbols in the data sheet like (Â). I am removing these symbols in the process of cleaning.

  Before cleaning:



	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	hour	total precipitation (mm)	atmospheric pressure at station height (mb)	atmospheric pressure max. in the previous hour (mb)	atmospheric pressure min. in the previous hour (mb)	radiation (kj/m2)	air temperature - dry bulb (Â°c)	dew point temperature (Â°c)	max. temperature in the previous hour (Â°c)	min. temperature in the previous hour (Â°c)	dew temperature max. in the previous hour (Â°c)	dew temperature min. in the previous hour (Â°c)	relative humidity max. in the previous hour (%)	relative humidity min. in the previous hour (%)	air relative humidity (%)	wind direction (Â° (gr))	wind rajada maxima (m/s)	wind speed (m/s)	station_code
2	30-04-2013	7:00	0	883.4	883.7	883.4	0	15	12.3	15.6	14.9	12.4	10.8	84	73	84	100	6.4	A555
3	27-11-2013	9:00	0	914	914	913.4	56	19	18.2	19	18.7	18.2	17.9	95	95	95	67	3.6	A551
4	19-02-2013	6:00	0	996.6	996.7	996.5	0	22.5	20.9	23.2	22.3	21.7	20.6	92	90	91	247	1.7	A623
5	21-11-2007	20:00	0	961.1	961.1	960.9	215	24.6	18.1	25.1	24.6	18.2	17.6	67	64	67	49	4.1	A527
6	4/10/2020	3:00	0	1015.4	1015.7	1015.4	0	23.9	20.9	24.8	23.9	21.1	20.9	83	79	83	191	5.8	A612
7	26-05-2013	12:00	0.2	972.8	972.8	972.4	1050	16.1	13.8	16.1	12.8	14	12	95	85	86	120	4	A718
8	19-04-2019	21:00	0	1010	1010	1009.8	40	23.6	20.2	25.3	23.6	20.3	19.8	81	72	81	192	5.4	A601
9	13-01-2015	0:00	0	0	0	0	0	23.1	18.9	24.9	23.1	19.1	18.6	77	69	77	249	3.5	A557
10	13-11-2019	14:00	0	933.5	933.8	933.5	790	24.2	18.6	24.4	22.7	19.5	18.4	80	70	71	236	4.5	A536
11	16-12-2020	9:00	0	915.5	915.5	915	33	21.3	14.8	21.4	20.8	15.8	14.7	73	66	66	136	1.4	A553
12	16-04-2016	10:00	0	952.3	952.3	951.6	52	19.3	15.7	19.3	17.9	15.7	14.9	83	80	80	36	1.6	A519
13	31-07-2014	2:00	0	912.9	912.9	912.8	0	11.4	11	11.8	11.3	11.6	11	99	98	98	205	4.3	A613
14	17-06-2020	12:00	0.2	921.1	921.1	920.6	370	13.7	13.4	14.1	13.5	13.8	13.3	98	98	98	65	5.7	A518
15	9/4/2013	0:00	0.2	831.5	831.5	830.9	0	13.3	13.2	13.5	13.3	13.4	13.2	100	99	99	0	0	A610
16	22-08-2017	7:00	0	993.5	993.6	993.4	0	19.3	13.7	19.4	19.2	13.9	13.6	71	69	70	126	3.6	A540

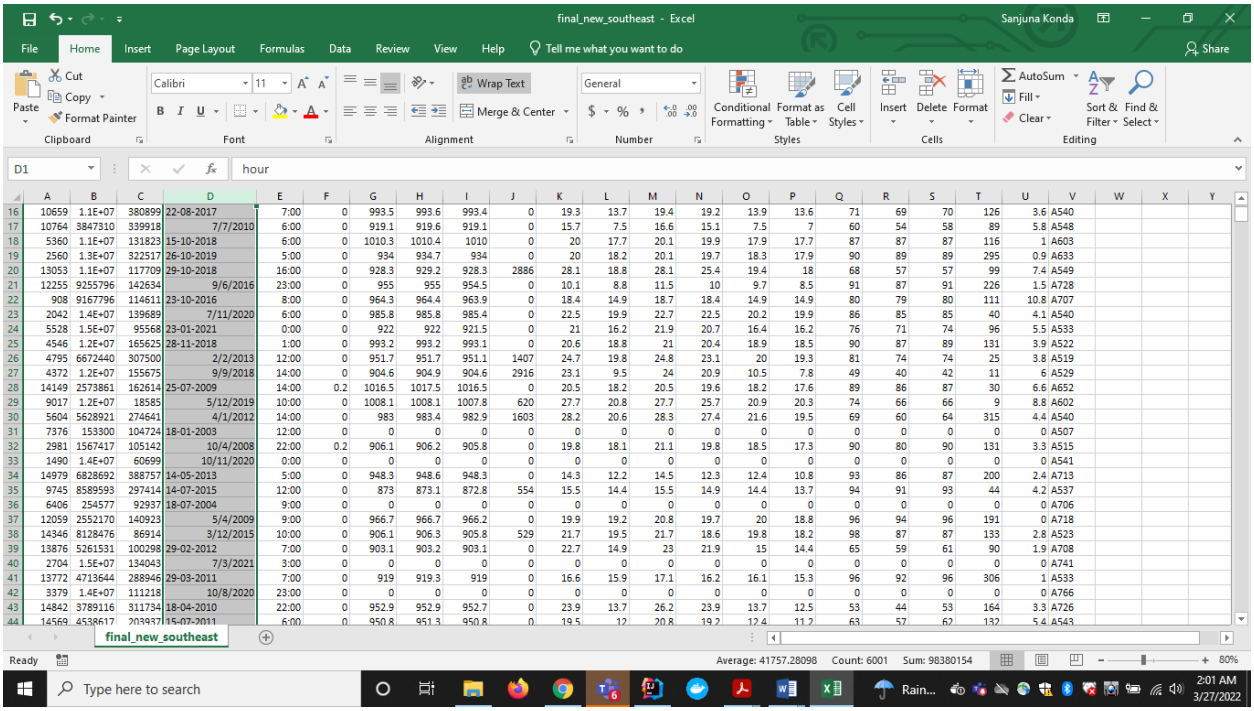
After cleaning:



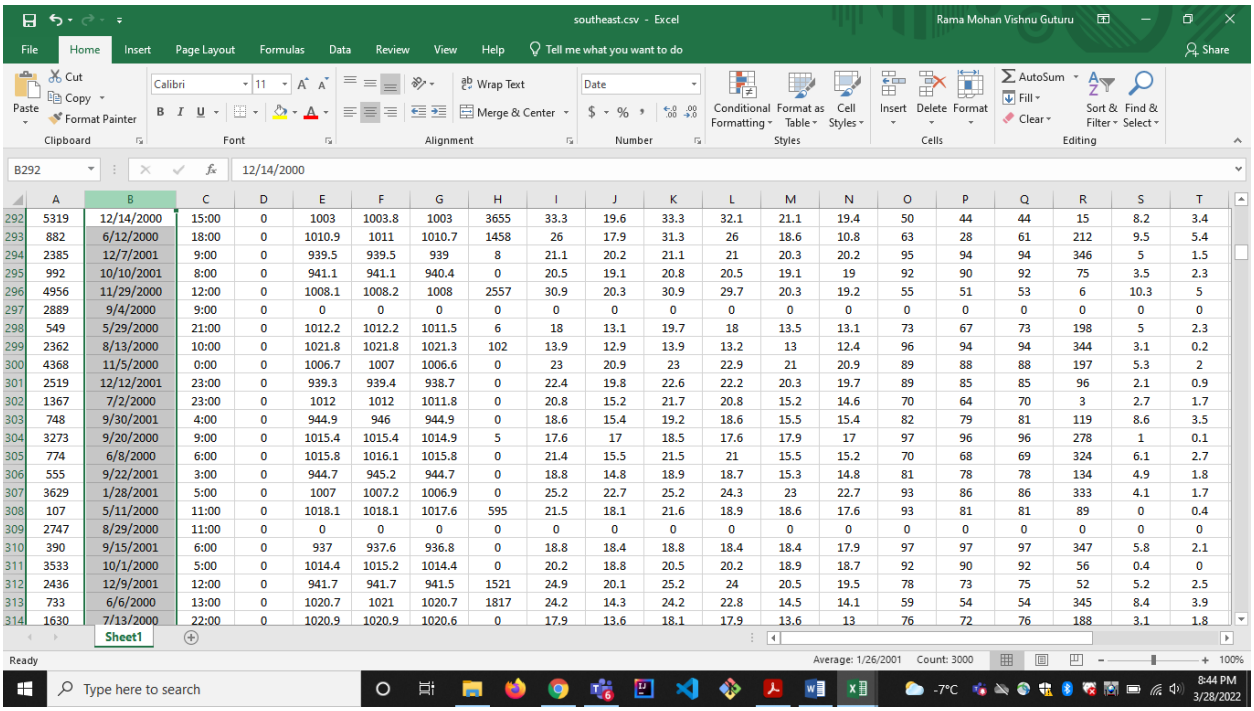
	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	hour	total precipitation (mm)	atmospheric pressure at station height (mb)	atmospheric pressure max. in the previous hour (mb)	atmospheric pressure min. in the previous hour (mb)	radiation (kj/m2)	air temperature - dry bulb (°c)	dew point temperature (°c)	max. temperature in the previous hour (°c)	min. temperature in the previous hour (°c)	dew temperature max. in the previous hour (°c)	dew temperature min. in the previous hour (°c)	relative humidity max. in the previous hour (%)	relative humidity min. in the previous hour (%)	air relative humidity (%)	wind direction (° (gr))	wind rajada maxima (m/s)	wind speed (m/s)	station_code
2	30-04-2013	7:00	0	883.4	883.7	883.4	0	15	12.3	15.6	14.9	12.4	10.8	84	73	84	100	6.4	A555
3	27-11-2013	9:00	0	914	914	913.4	56	19	18.2	19	18.7	18.2	17.9	95	95	95	67	3.6	A551
4	19-02-2013	6:00	0	996.6	996.7	996.5	0	22.5	20.9	23.2	22.3	21.7	20.6	92	90	91	247	1.7	A623
5	21-11-2007	20:00	0	961.1	961.1	960.9	215	24.6	18.1	25.1	24.6	18.2	17.6	67	64	67	49	4.1	A527
6	4/10/2020	3:00	0	1015.4	1015.7	1015.4	0	23.9	20.9	24.8	23.9	21.1	20.9	83	79	83	191	5.8	A612
7	26-05-2013	12:00	0.2	972.8	972.8	972.4	1050	16.1	13.8	16.1	12.8	14	12	95	85	86	120	4	A718
8	19-04-2019	21:00	0	1010	1010	1009.8	40	23.6	20.2	25.3	23.6	20.3	19.8	81	72	81	192	5.4	A601
9	13-01-2015	0:00	0	0	0	0	0	23.1	18.9	24.9	23.1	19.1	18.6	77	69	77	249	3.5	A557
10	13-11-2019	14:00	0	933.5	933.8	933.5	790	24.2	18.6	24.4	22.7	19.5	18.4	80	70	71	236	4.5	A536
11	16-12-2020	9:00	0	915.5	915.5	915	33	21.3	14.8	21.4	20.8	15.8	14.7	73	66	66	136	1.4	A553
12	16-04-2016	10:00	0	952.3	952.3	951.6	52	19.3	15.7	19.3	17.9	15.7	14.9	83	80	80	36	1.6	A519
13	31-07-2014	2:00	0	912.9	912.9	912.8	0	11.4	11	11.8	11.3	11.6	11	99	98	98	205	4.3	A613
14	17-06-2020	12:00	0.2	921.1	921.1	920.6	370	13.7	13.4	14.1	13.5	13.8	13.3	98	98	98	65	5.7	A518
15	9/4/2013	0:00	0.2	831.5	831.5	830.9	0	13.3	13.2	13.5	13.3	13.4	13.2	100	99	99	0	0	A610
16	22-08-2017	7:00	0	993.5	993.6	993.4	0	19.3	13.7	19.4	19.2	13.9	13.6	71	69	70	126	3.6	A540

- The date column has values with different formats which might be an issue later, in case of querying or retrieving the data. So, I have changed that column to a single format of all values to maintain consistency.

Before cleaning:

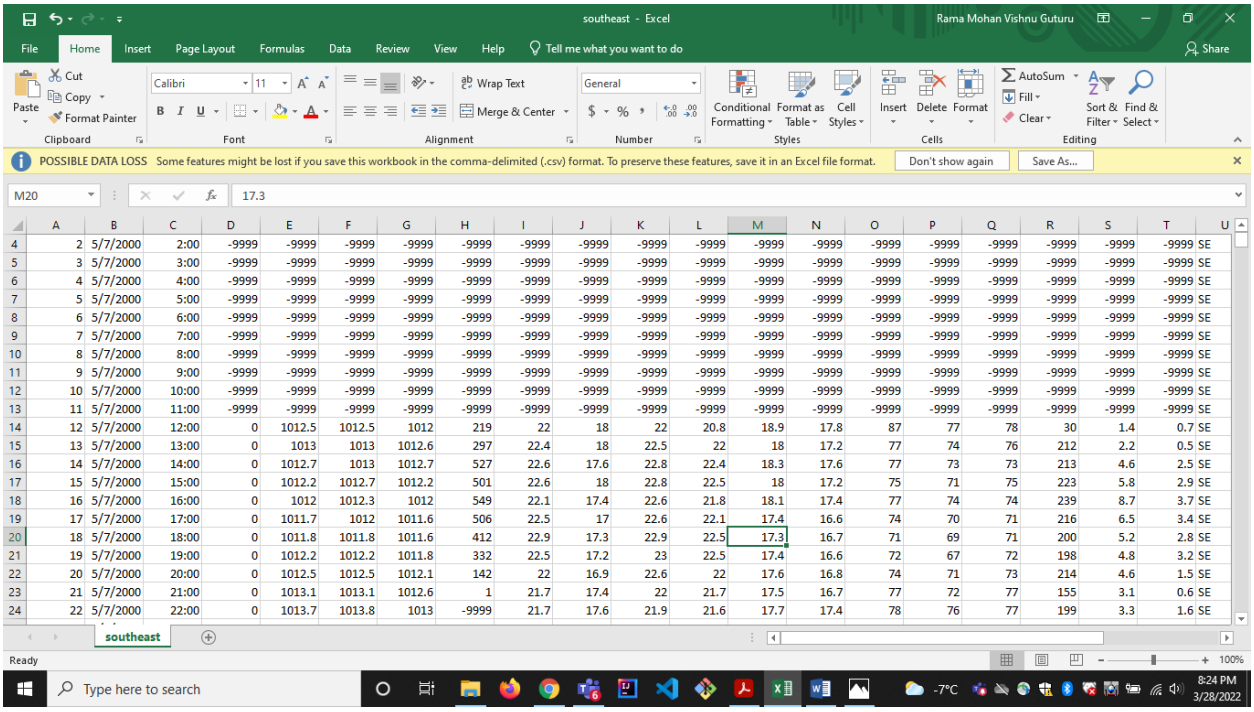


After cleaning:

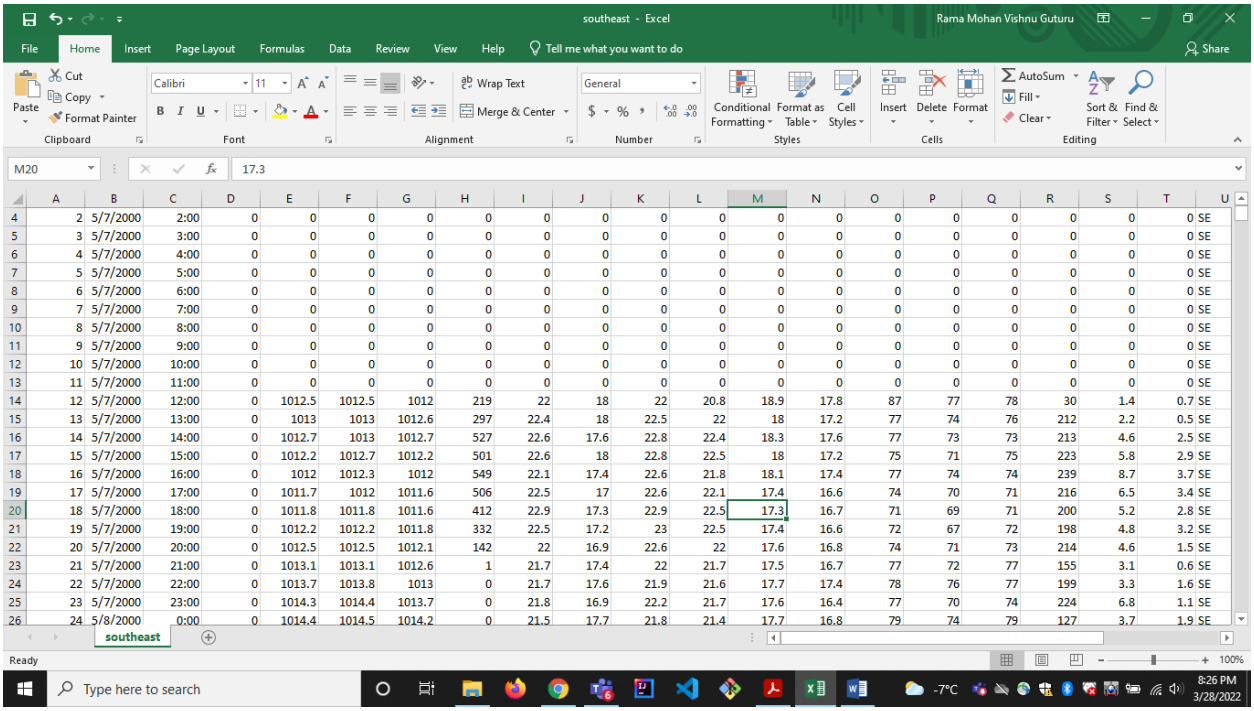


- There were many values of -9999 in the dataset. So, I removed them and replaced them with 0.

  Before cleaning:



  After cleaning:

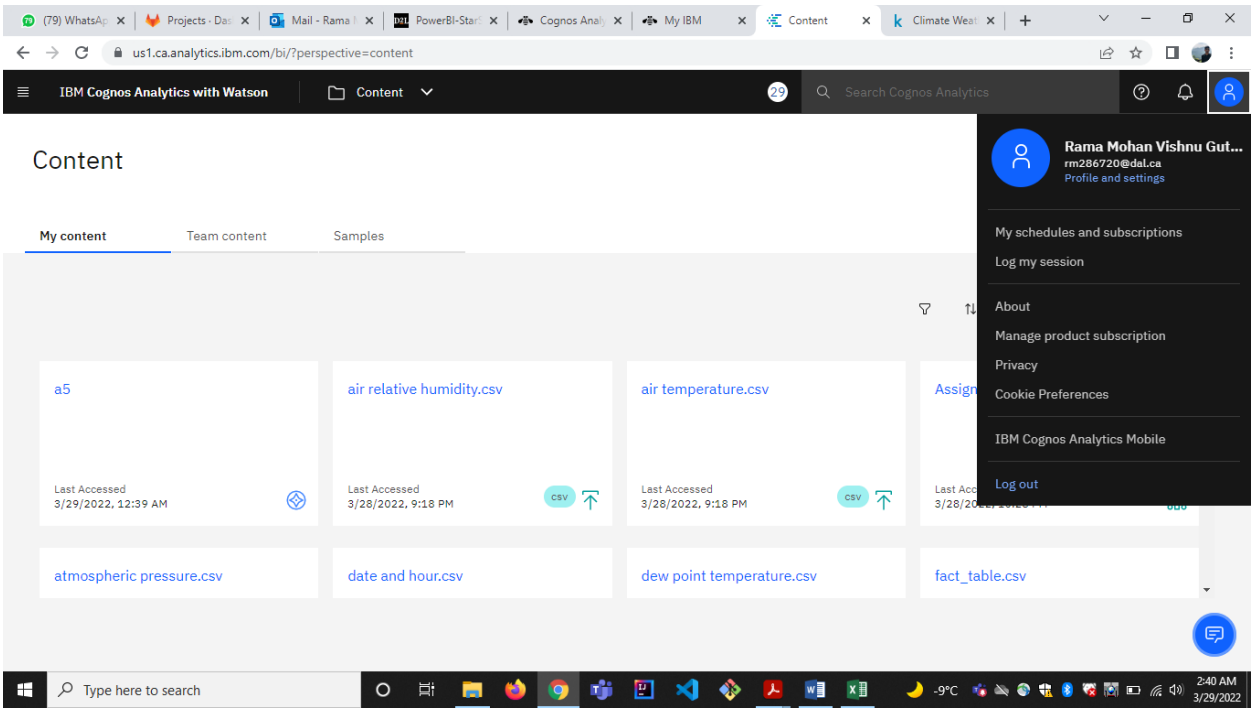


- After the above step, if a row has all 0 values in it, then I am deleting the row from the dataset.
- There are no duplicate values in the rows I extracted from the original dataset. So, there is no need to delete the duplicate rows.

After performing all these steps, the dataset is cleaned and it doesn't have any unnecessary or unwanted data in it.

Step 5 : Create Cognos account

I have created the IBM Cognos account and have imported all the dimension tables along with the fact table.



Step 6 : Create Star Schema/Snowflake Schema

I have created a star schema based on my understanding of the domain. As already said, there are multiple dimensions for the fact table and all these dimensions have 1 : N relationship with the fact table.

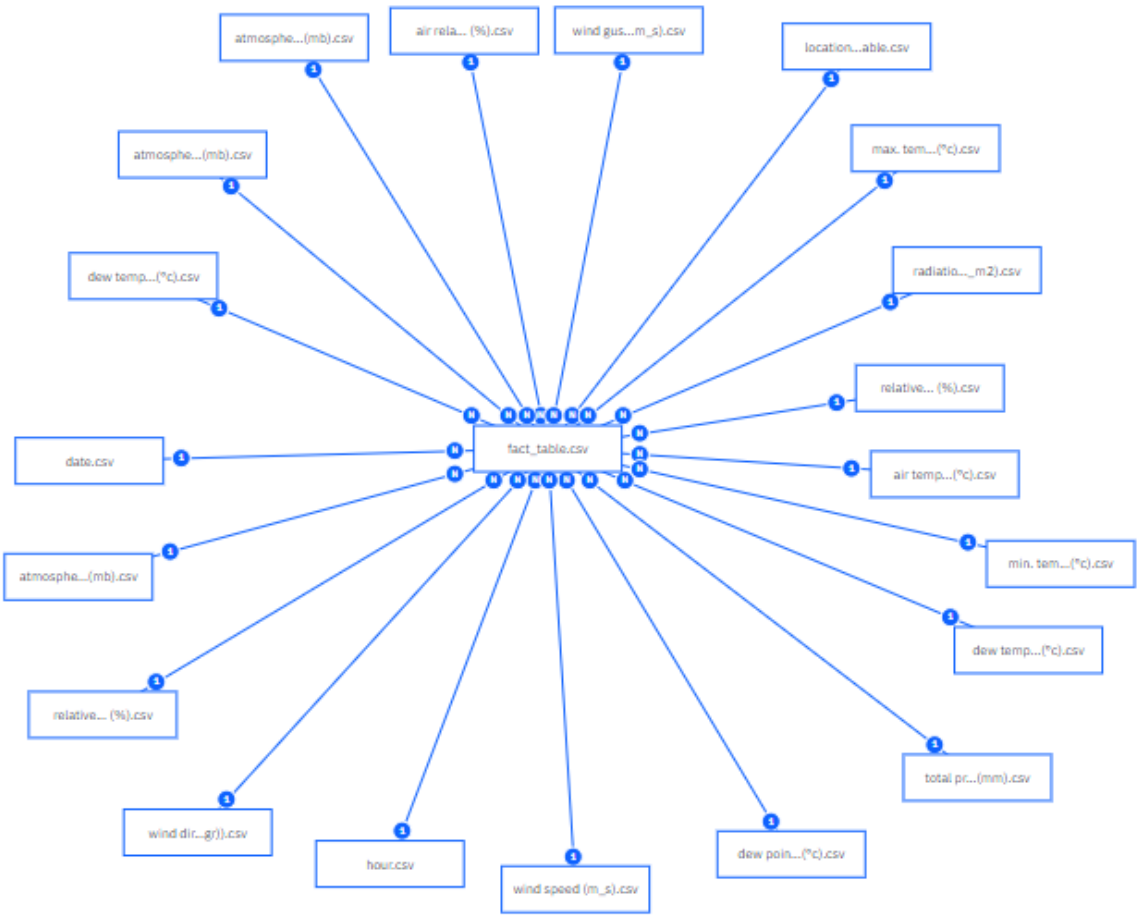
- ✓ Location – A region's climate is inextricably tied to the weather that occurs there. If a location experiences very cold weather for the majority of the year, it is considered to have a cold climate since the average temperature over a lengthy period of time is primarily cold.
- ✓ Wind – Wind transports moisture, as well as hot and cold air, into the atmosphere, influencing weather patterns. As a result, a change in wind causes a change in weather.
- ✓ Temperature – This is the main factor of all the factors for any weather. To be brief, temperature causes climate and climate is the result of temperature. So, they are bonded very closely.
- ✓ Relative Humidity – If the temperature increases, the humidity increases and climate gets hot. If the temperature decreases, humidity decreases and climate gets cold.
- ✓ Total Precipitation – Climate change has the potential to alter precipitation intensity and frequency. Warmer oceans cause more water to evaporate into the atmosphere. Heavy rain and snowstorms, for example.
- ✓ Atmospheric Pressure – Atmospheric pressure is a weather indicator. Cloudiness, wind, and precipitation are common when a low-pressure system moves into a region. Fair and quiet weather is frequently associated with high-pressure systems.
- ✓ Date and Hour – Weather changes every single day due to winds and storms. Weather also changes with seasons. In summer season, the weather is hot; in

winter, the weather is too cold and in rainy season, it is just below the room temperature.

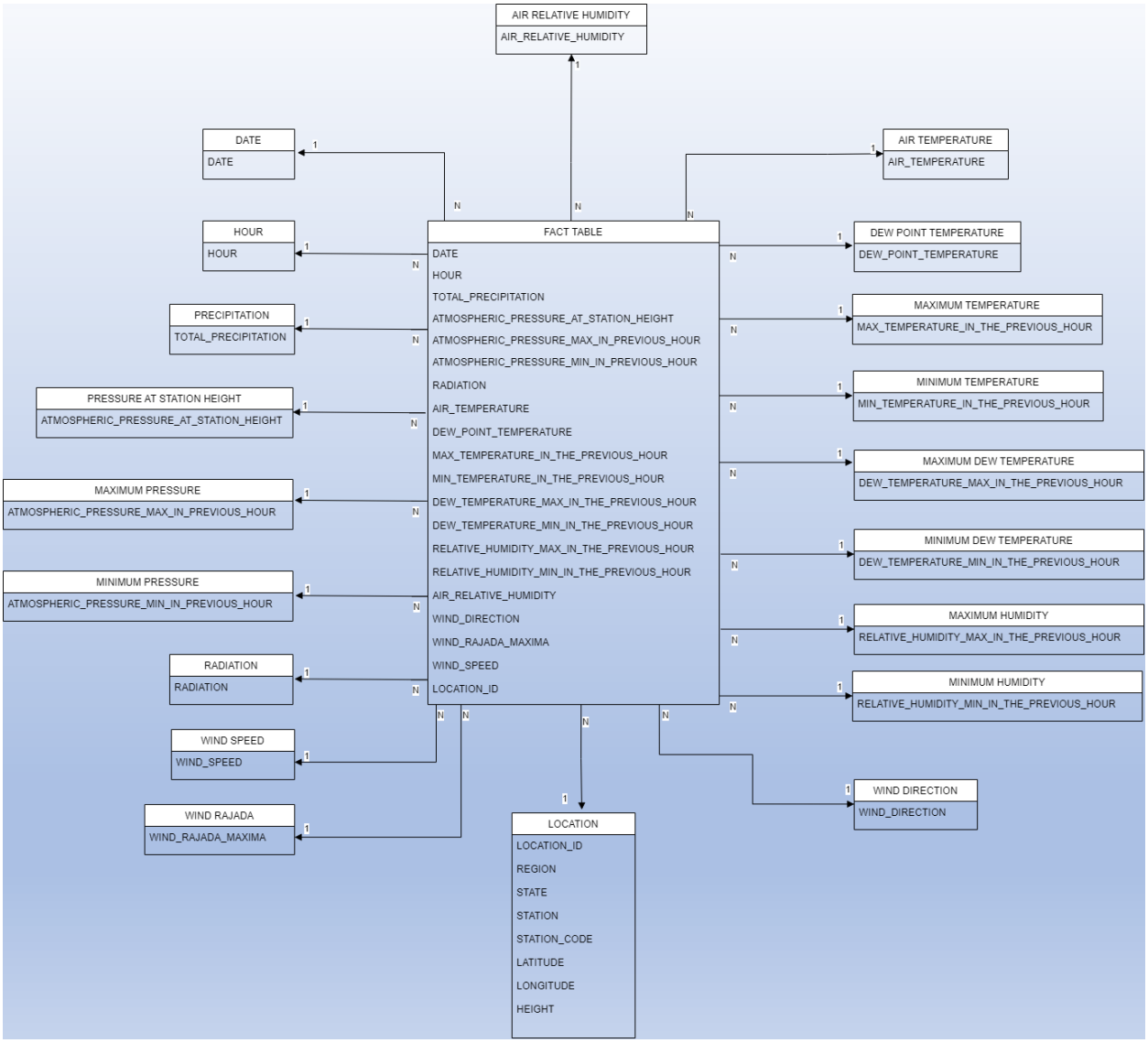
- ✓ Radiation – Radiation is very important to weather. The amount of radiation defines the temperature of that area of that city and the temperature directly affects the climate at that particular time.

Step 7:

The schema I created using IBM Cognos is provided below.



I have also built the dimensional modelling for this dataset using draw.io. Below is the model I built.

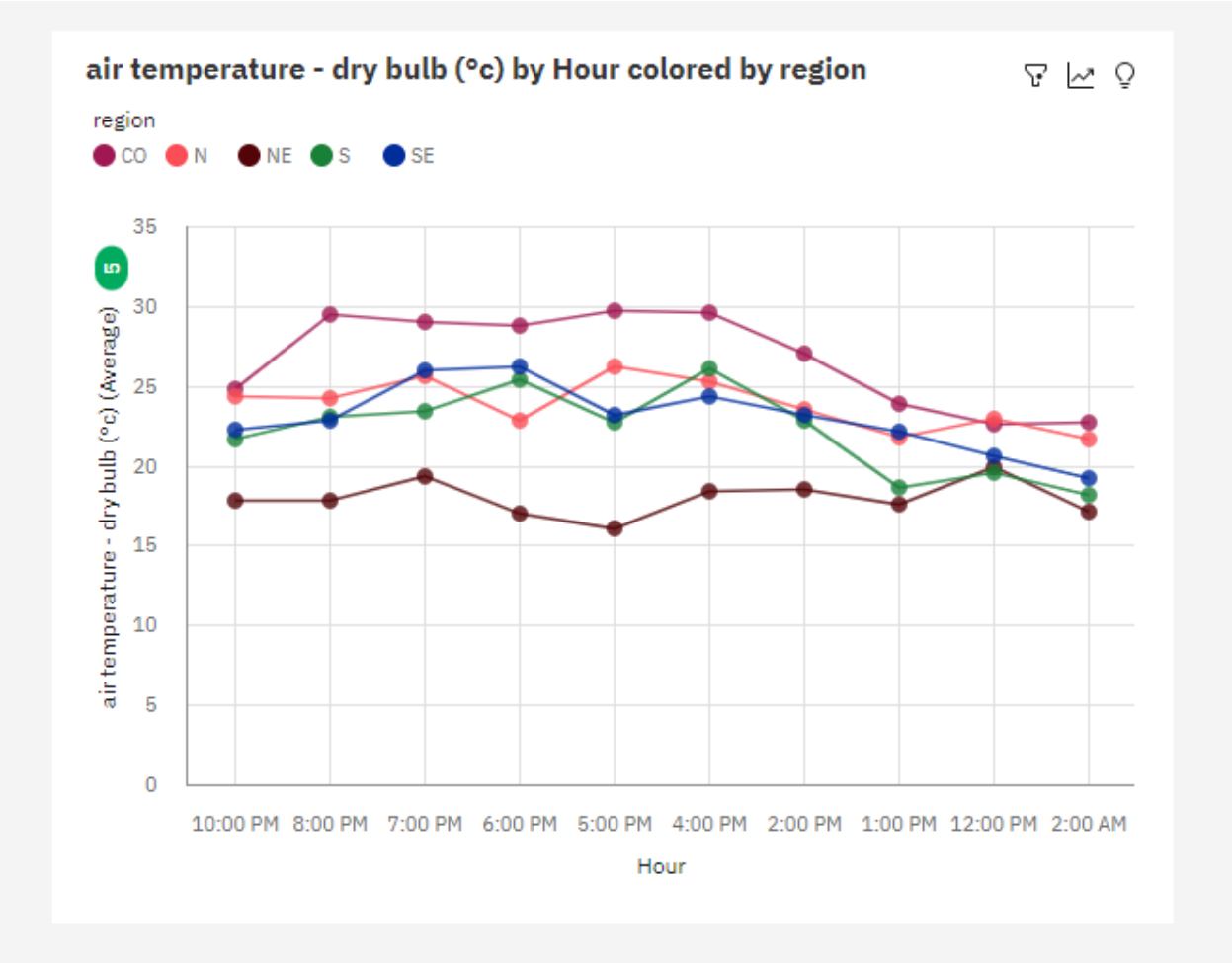


Step 8 : Visual Analysis of the data

In order for anyone to understand the dataset easily and in a better way, I created few visualizations which might help to look and understand better, rather than looking at thousands of rows of data in excel sheets.

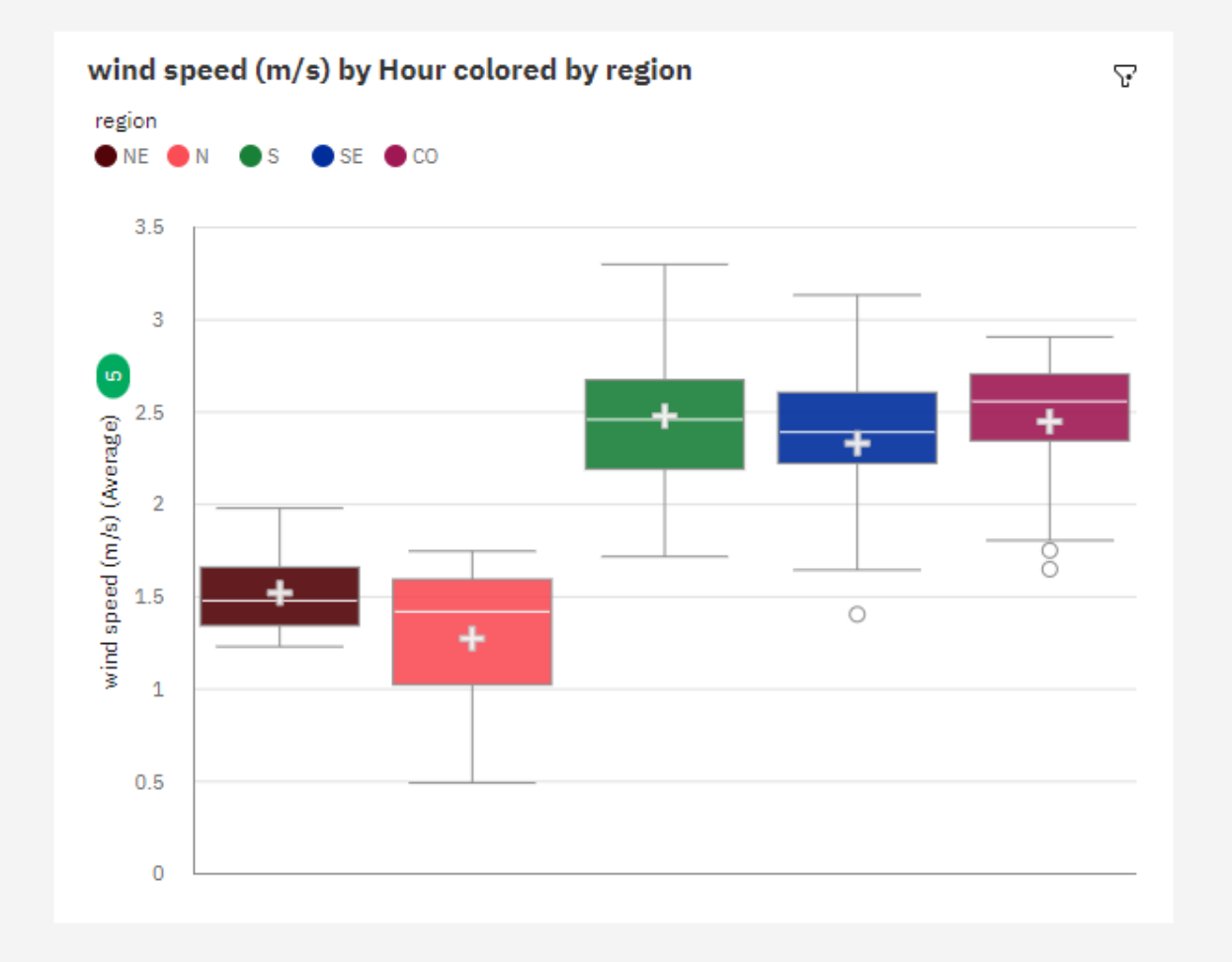
Visual 1 – Air temperature by Hour colored by Region

This visual represents the air temperature of all the regions based on hour by intervals. From this visual, we can understand that the value of air temperature is most unusual when the values of region are NE and CO. And also, the values of air temperature are most unusual when the values of hour are 02:00, 13:00 and 19:00.



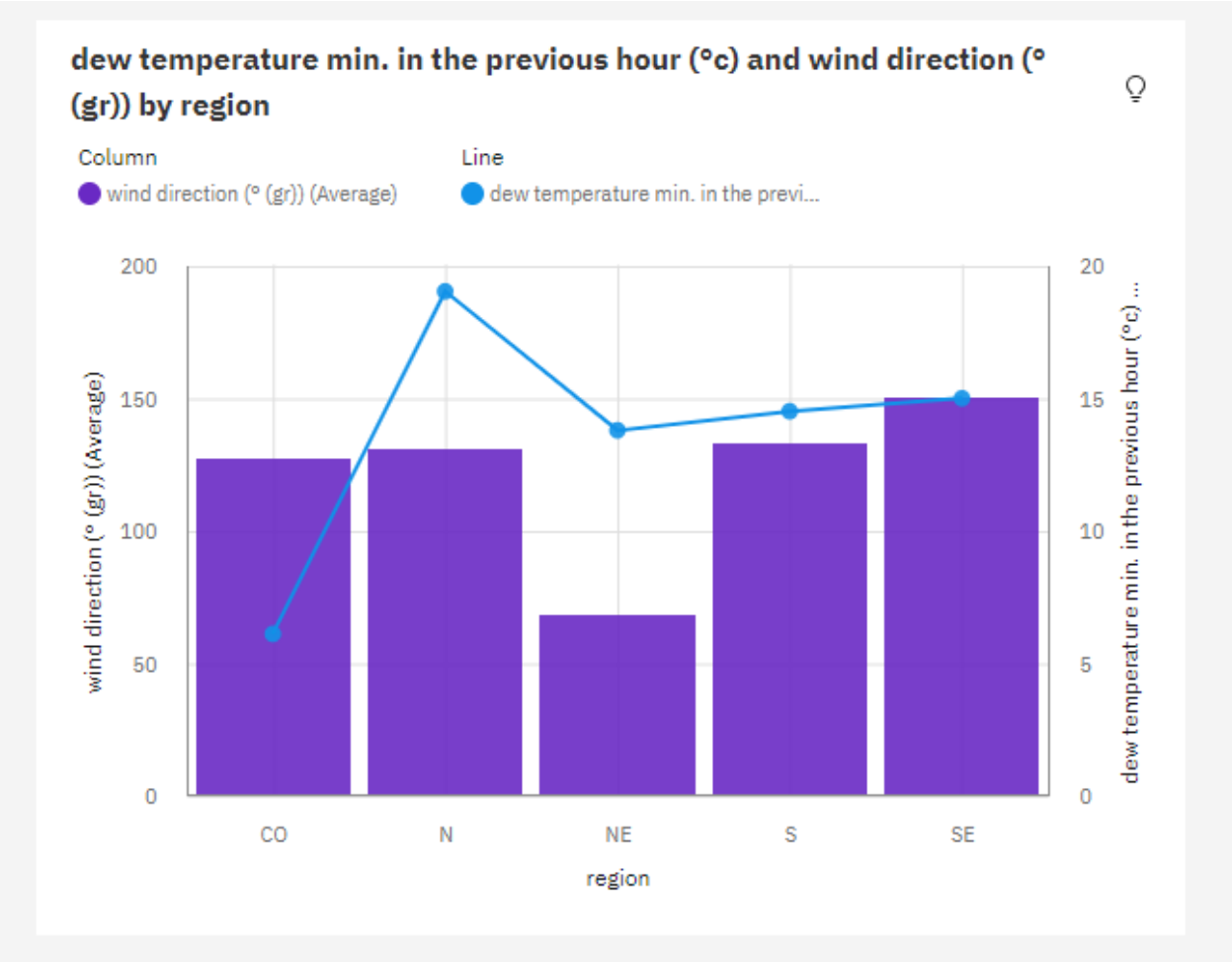
Visual 2 – Wind speed by Hour colored by Region

This visual represents the relative wind speeds separated by hours and differentiated by regions. As we can see, the average wind speed is higher in CO region and lower in N region. CO, SE and S have almost the same average wind speed with minor differences.



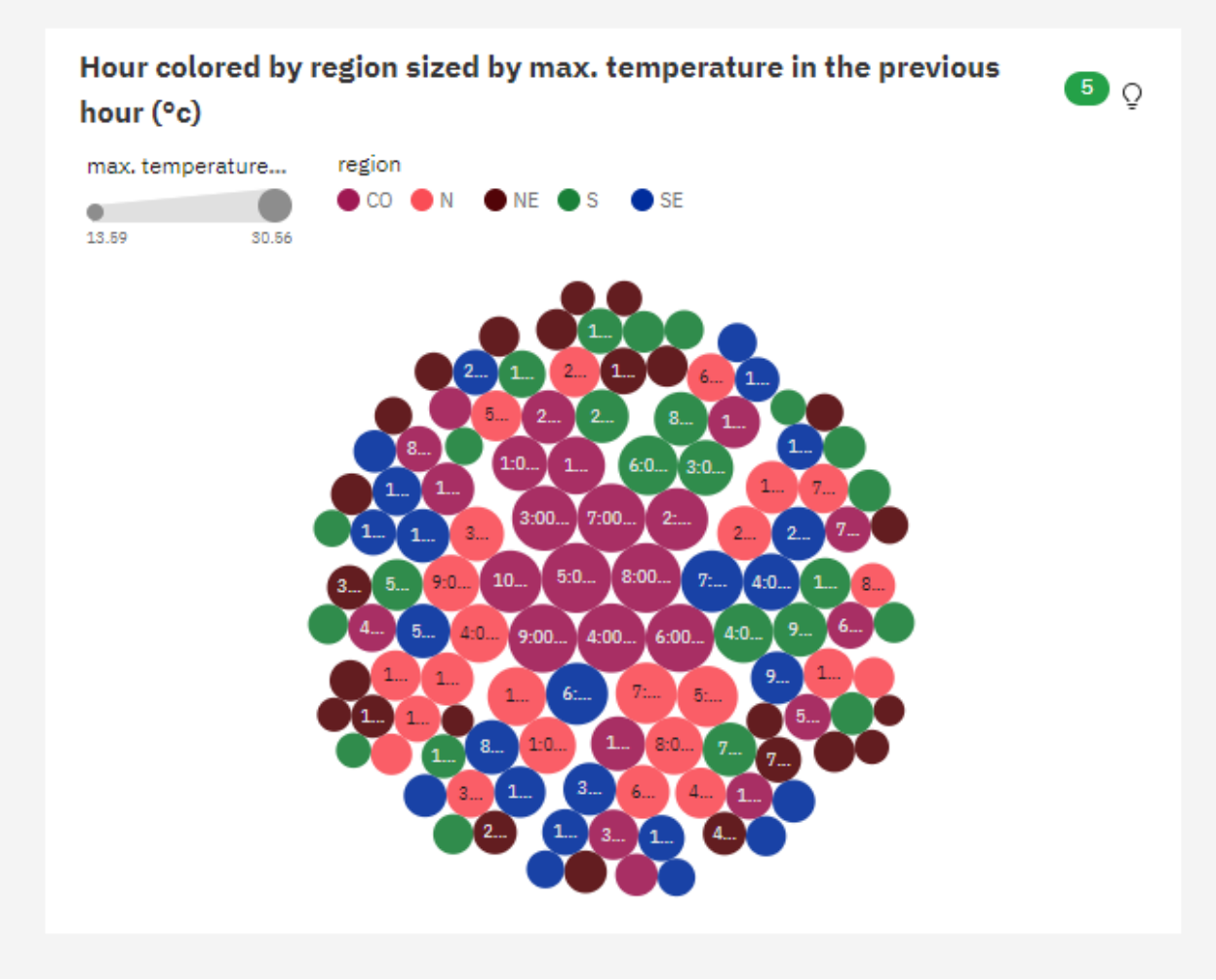
Visual 3 – Dew temperature in the previous hour and wind direction by region

As the heading says, this bar graph depicts the minimum dew temperature in all the regions in the previous hour and it also depicts the average wind direction in all the regions. The average of wind direction for all values of region is 121.6. The value of wind direction is unusually low when region is NE. The average of minimum dew temperature in the previous hour of all values in all regions is 13.79.



Visual 4 – Hour colored by region sized by maximum temperature

This bubble diagram shows the information of hours differentiated by region and sized by maximum temperature. As we can observe, the maximum temperature has been mostly in the CO region and the lowest of maximum temperature has mostly been in NE region.



Visual 5 – Air relative humidity by hour

The pie charts in the diagram below helps us to understand the relative humidity of each region differentiated by each hour from the dataset. Some of the facts we can get from these are that N and NE are the most frequently occurring regions with a combined total of 41.9% of humidity. On the other hand, the value of relative humidity is unusually low when it comes to CO region. Regarding the time, 13:00 is the most frequently occurring hour with a percentage of 4.7 of the total relative humidity.