**Introduction**

The Intelligence Data Analysis is presented in this report. The data analysis has many tools that include a few main parts: data cleaning, data analyse and data prediction. The standard science approaches can be greatly improved using these tools.

**Initial data**

The initial minute-by-minute data was used (from <https://www.ruoa.unam.mx/> ‘**Meteorología’**). The more data we have, the better it is when using data science methods.

Data info:

58966128 raws

12 columns:

0 TIMESTAMP datetime64[ns] yyyy-mm-dd hh:mm:ss

1 Temp\_Avg float32 C

2 RH\_Avg float32 %

3 WSpeed\_Avg float32 m/s,

4 WSpeed\_Max float32 m/s,

5 WDir\_Avg float32 deg,

6 WDir\_SD float32 deg,

7 Rain\_Tot float32 mm,

8 Press\_Avg float32 hPa,

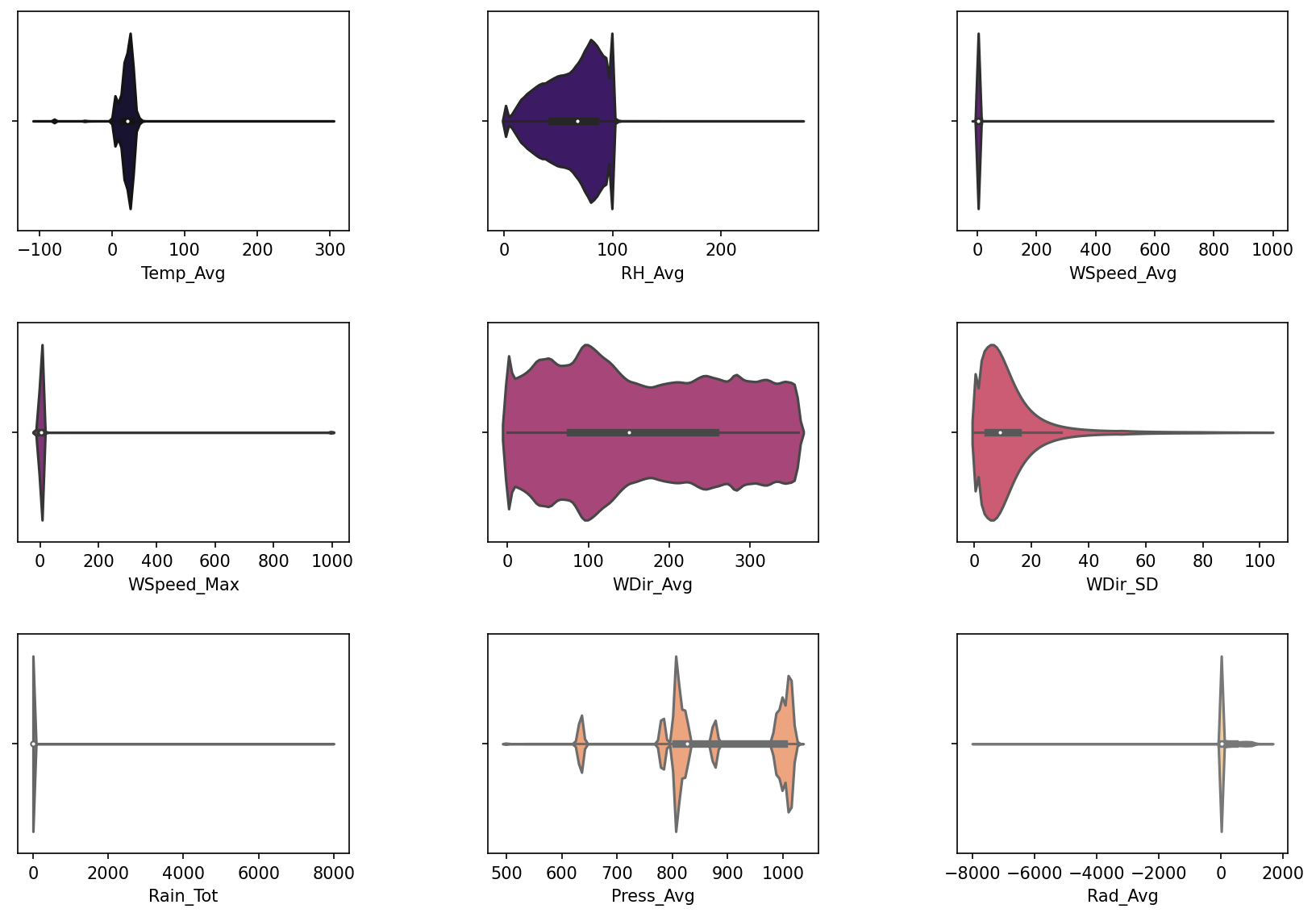
9 Rad\_Avg float32 W/m^2,

10 region category name of region

11 Visibility float32 m

12 Visibility\_Avg float32 m

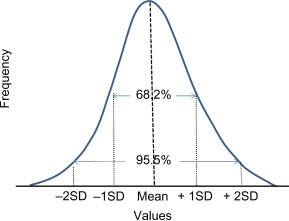
Data contains outliers as extremal temperature (lower than -100 and higher than 300) (**Figure 1**). Outliers are data that are clearly out of the picture. They are usually not used in general analysis, because they can greatly distort the results. To exclude them, it is necessary to clean the data. In **Figure 1**, outliers are represented by long thin lines that lie outer of a 95% confidence interval.



**Fig. 1.** Violinplot - black box demonstrated 50% of data (from 25 to 75 percentile) white point is median. Black lines are 95% сonfidence interval. Plot for data before cleaning. Temp\_Avg - average temperature, RH\_avg - …, Wspeed\_Avg - mean speed of wind, WSpeed\_Max - max speed of wind, WDir\_avg - mean direction of speed, WDir\_SD - …, Rain\_Tot - total rain, Press\_avg - mean pressure, Rad\_avg - mean …

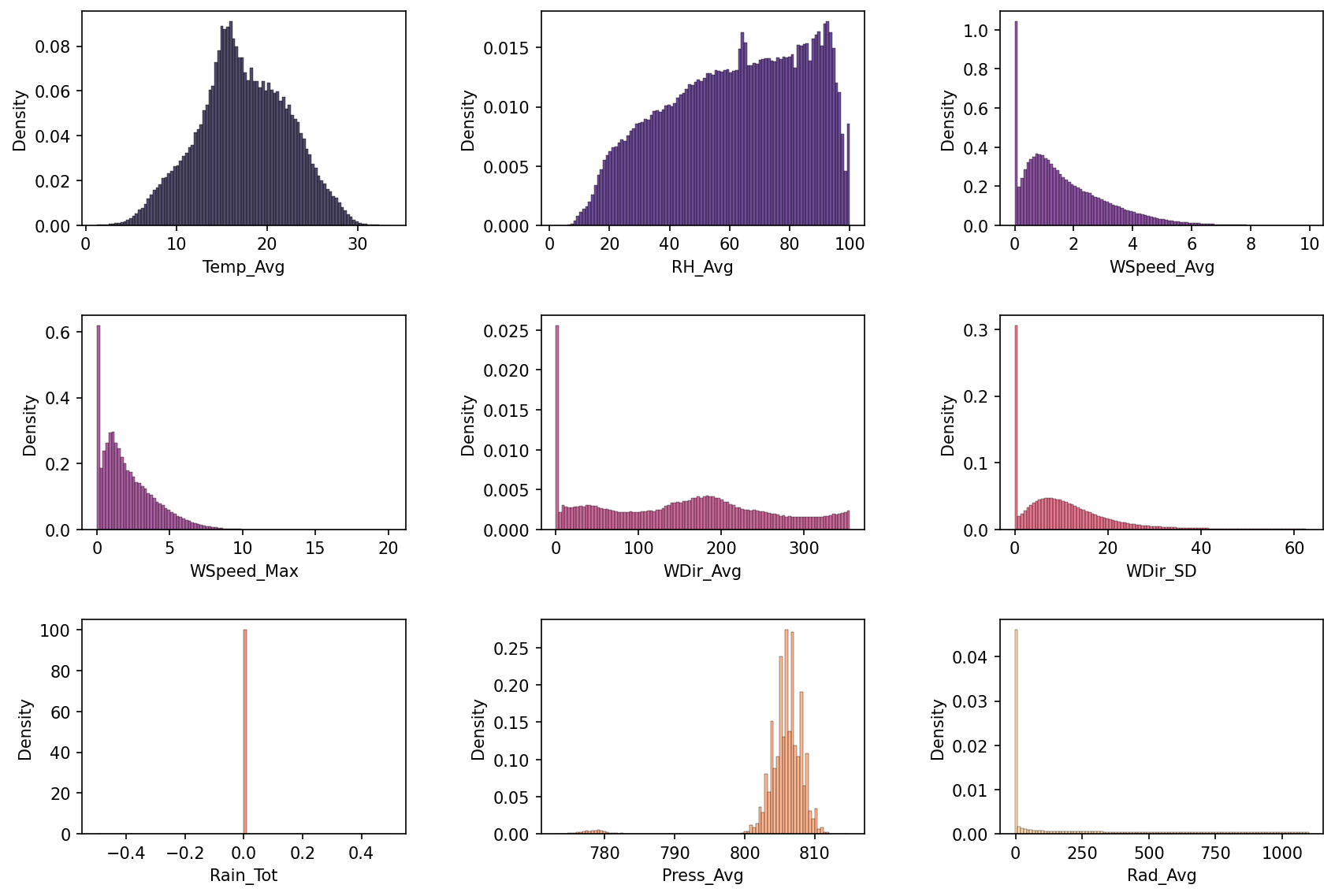
There are two basic ways to clean data:

**1.** Based on gaussian distribution



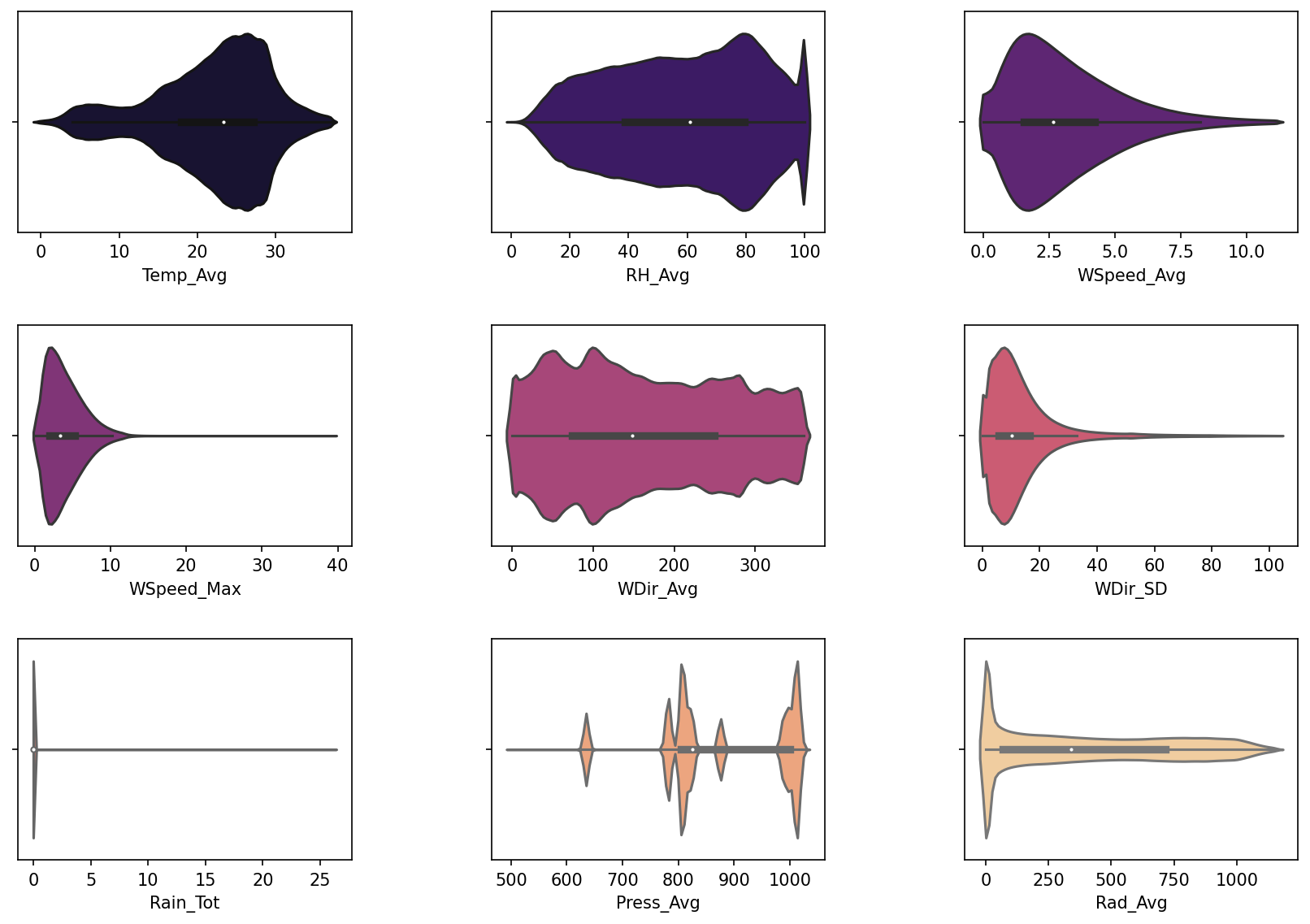
**2.** Based on quantilles (cut data into equal parts) For example q=0.99 means that 99% of data are lower and 1% of data are higher.

To show distribution of our data we choose and analyze one region (Mexico City) (**Figure 2**). As can be seen from **Figure 2**, only the temperature and pressure data are close to the Gaussian distribution. Therefore, it would be more correct to use a purification method based on quantiles.



**Figure 2.** Distributions of the data for Mexico City.

The data after cleaning gives a more physical picture (I am doubt, that there was +300 C in Mexico…) and can be analyzed and interpreted (**Figure 3**).

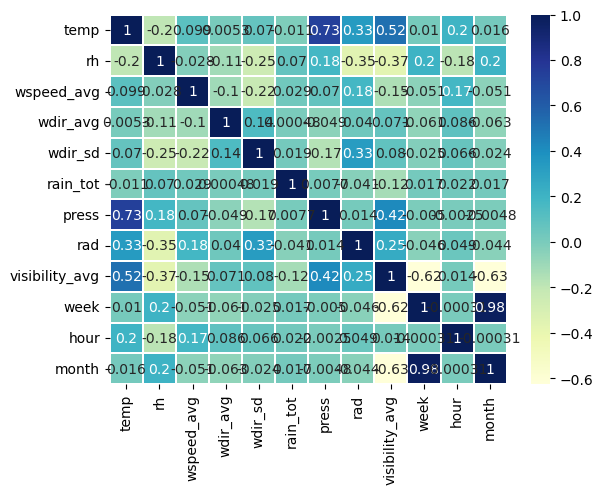


**Figure 3.** Violinplot of data after cleaning.

**Data analysis**

Cleaned data was grouped by hours to fastes calculation. 3 new columns were added to dataframe - week (week of year iso 8601), hour - hour of day 1-24, month 1-12.

Linear dependence is not the most informative type of correlations, but it allows us to get the big picture with a quick look at the data (**Figure 4**). For example, we see that pressure and temperature correlate with each other with a correlation coefficient r\*\*2 = 0.73, and visibilit with temperature has r\*\*2=0.52. Visibility depends on month with r\*\*2 = -0.63 (anticorrelation).



**Figure 4.** Correlation Matrix.

some columns has missing values. Especially visibility.

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 time 826288 non-null datetime64[ns]

1 region 826288 non-null object

2 temp 817077 non-null float64

3 rh 823748 non-null float64

4 wspeed\_avg 822742 non-null float64

5 wdir\_avg 824490 non-null float64

6 wdir\_sd 826081 non-null float64

7 rain\_tot 822033 non-null float64

8 press 792434 non-null float64

9 rad 811892 non-null float64

10 visibility\_avg 11788 non-null float64

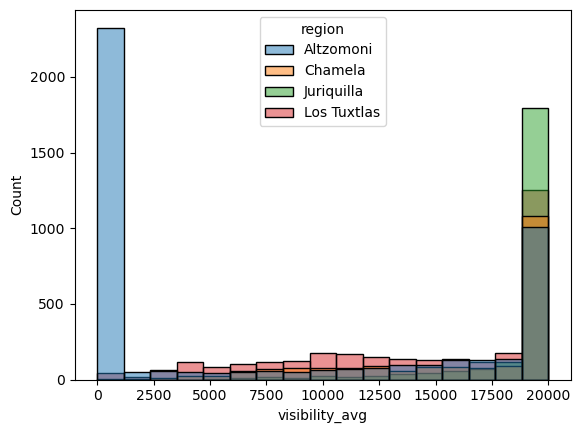
11 week 826288 non-null int64

12 hour 826288 non-null int64

13 month 826288 non-null int64

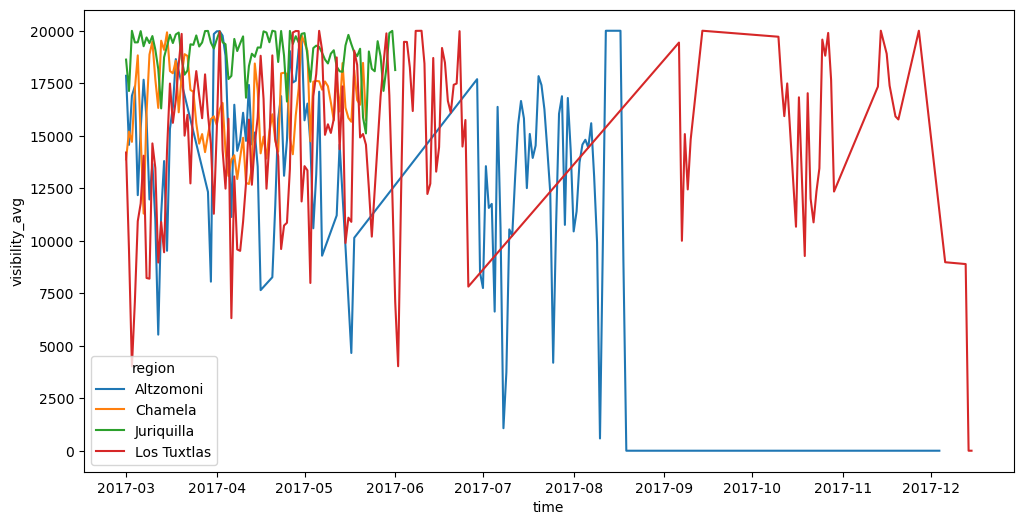
**Visibility analyse**

Only 4 regions contains info about visibility and region Altzomoni contains a lot of zero visibility. It can by outliers.



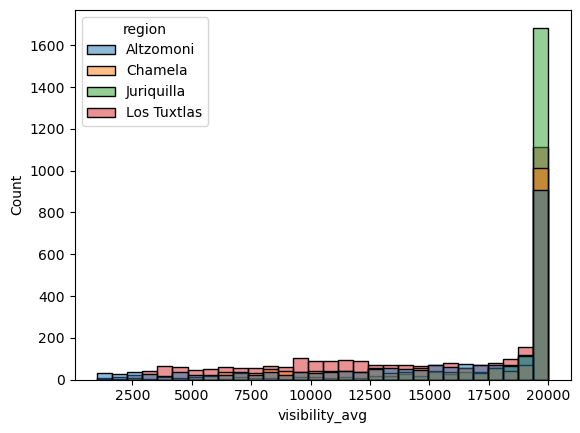
**Figure 5.** Histogramm of visibility for different regions.

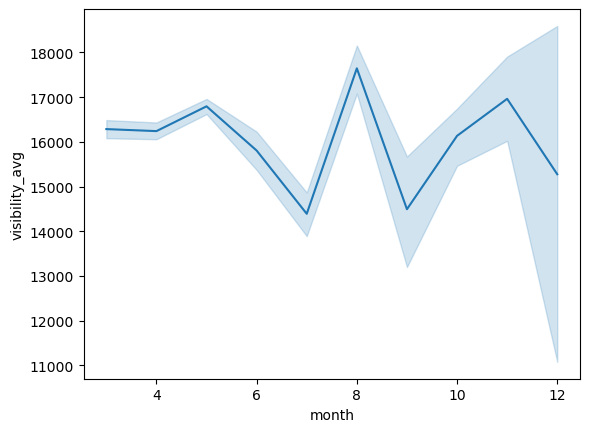
To analyse what time Altzomoni has 0 visibility we can show visibility on time. It looks as if someone soiled the lens of the device in Altzomoni (please punish him).



**Figure 6.** Visibility for different regions on time.

Lets fix it.

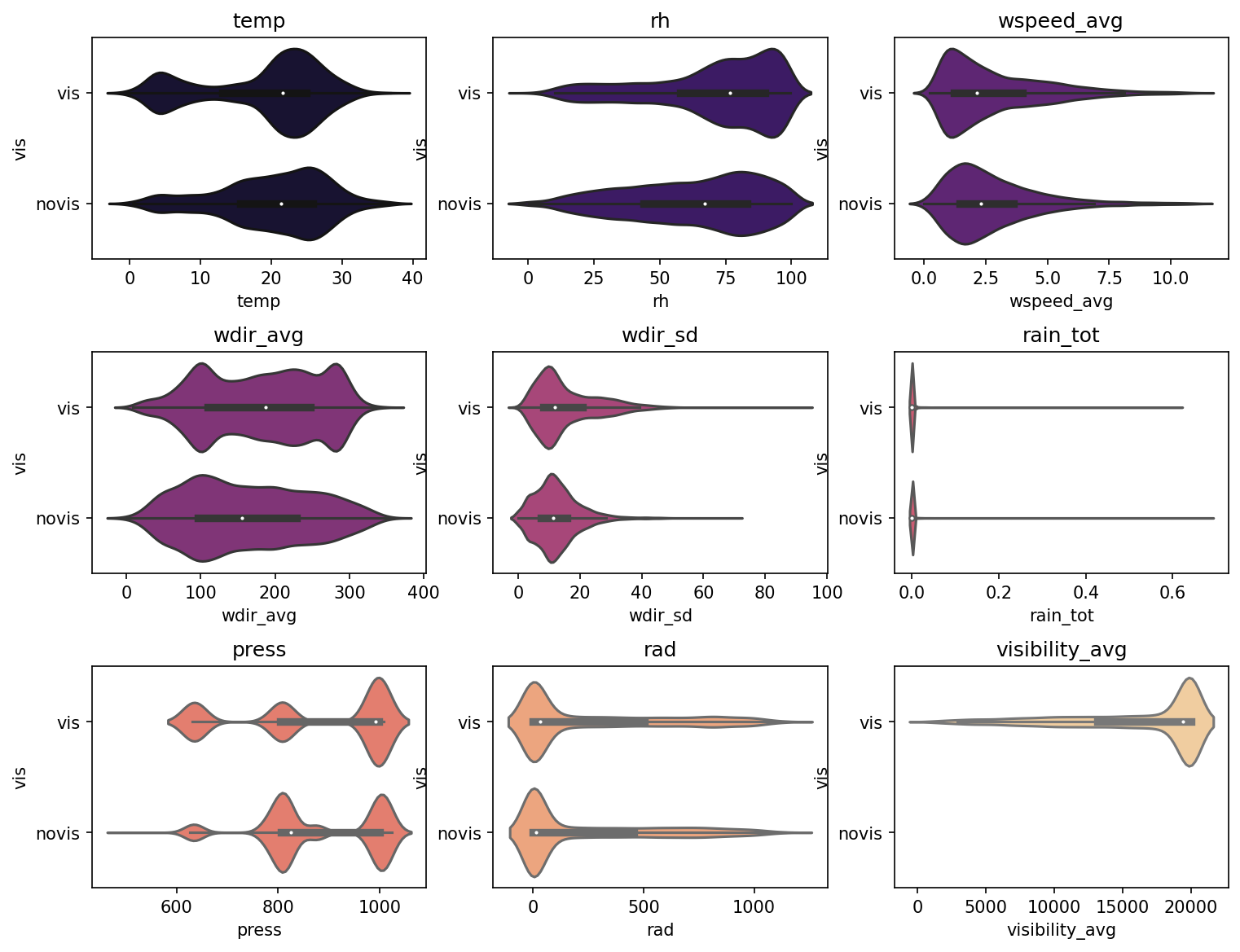




**Figure 7.** Visibility after cleaning

**Visibility predictions**

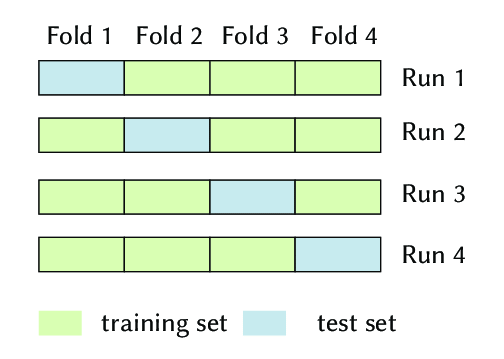
Similare distribution in regions with visibility data and without it allows us train Machine learning algorithm to predict visibility (**Figure 8**).



**Figure 8.** Distribution of variables in data with and without visibility.

To predict visibility we have used lightGBM python library. It is based on <https://en.wikipedia.org/wiki/Gradient_boosting> algorythm.

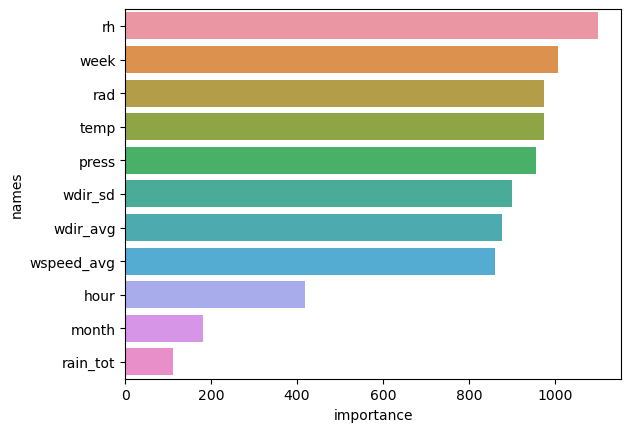
Train data was splitted into 10 parts and 10 different gbm trained on 9 parts and predict the other 1 part. It is called cross validation (**Figure 9**)



**Figure 8.** k fold cross validation

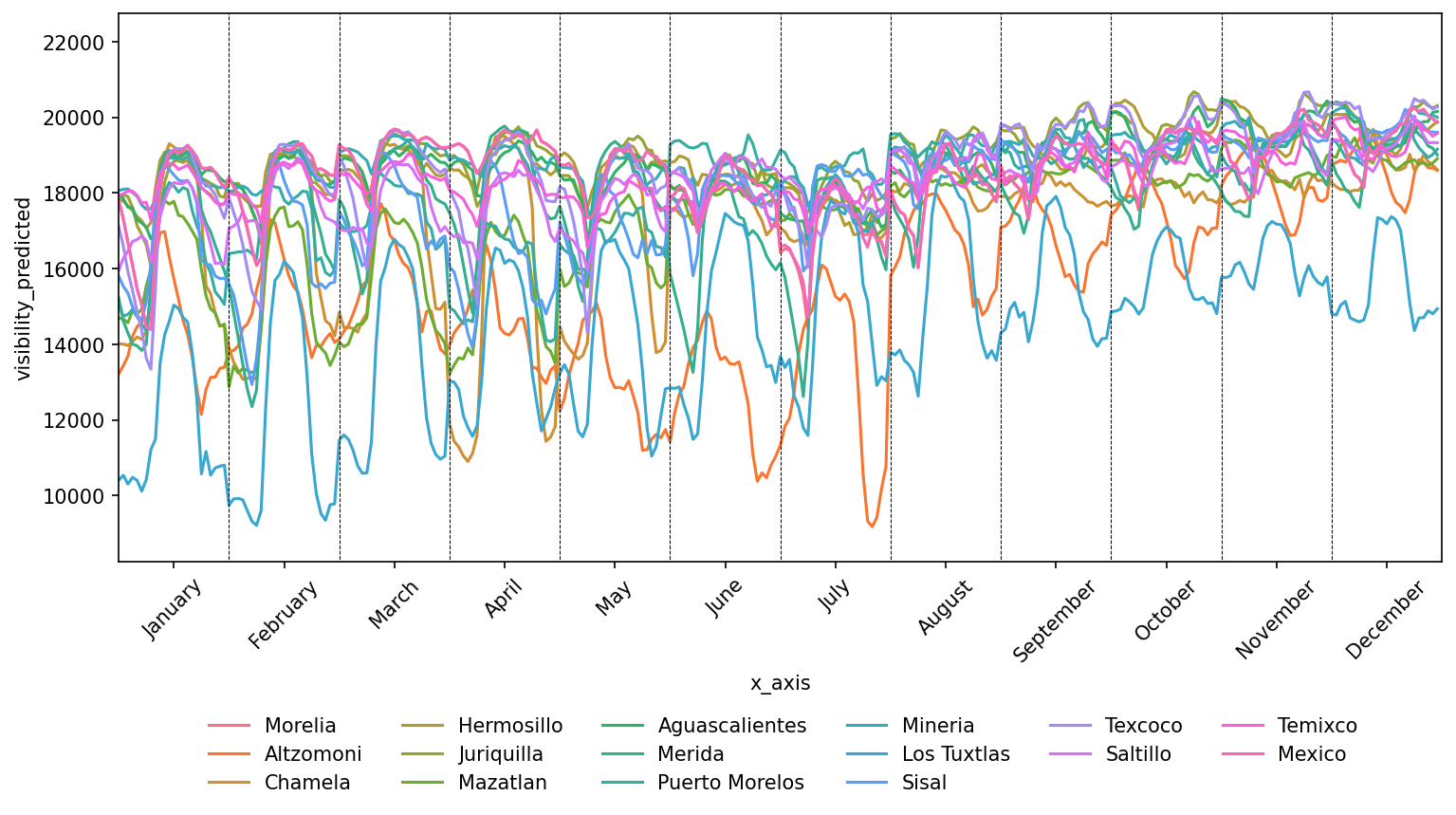
mean RMSE value is 2692 m.

GBM can show us feature importance. It is better than correlation matrix because it shows non linear importance between variables (**Figure 10**).



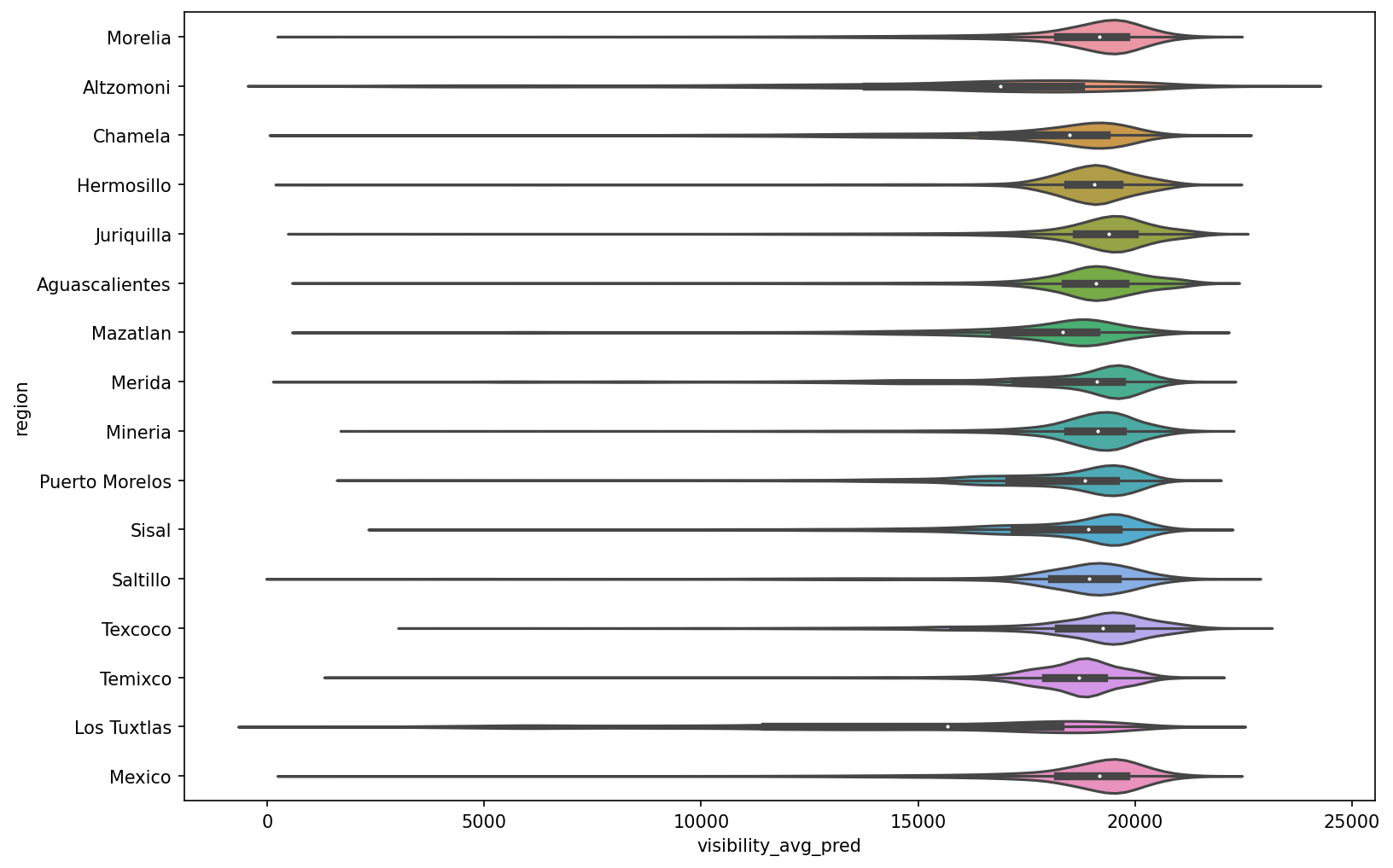
**Figure 10.** Feauture importance of GBM.

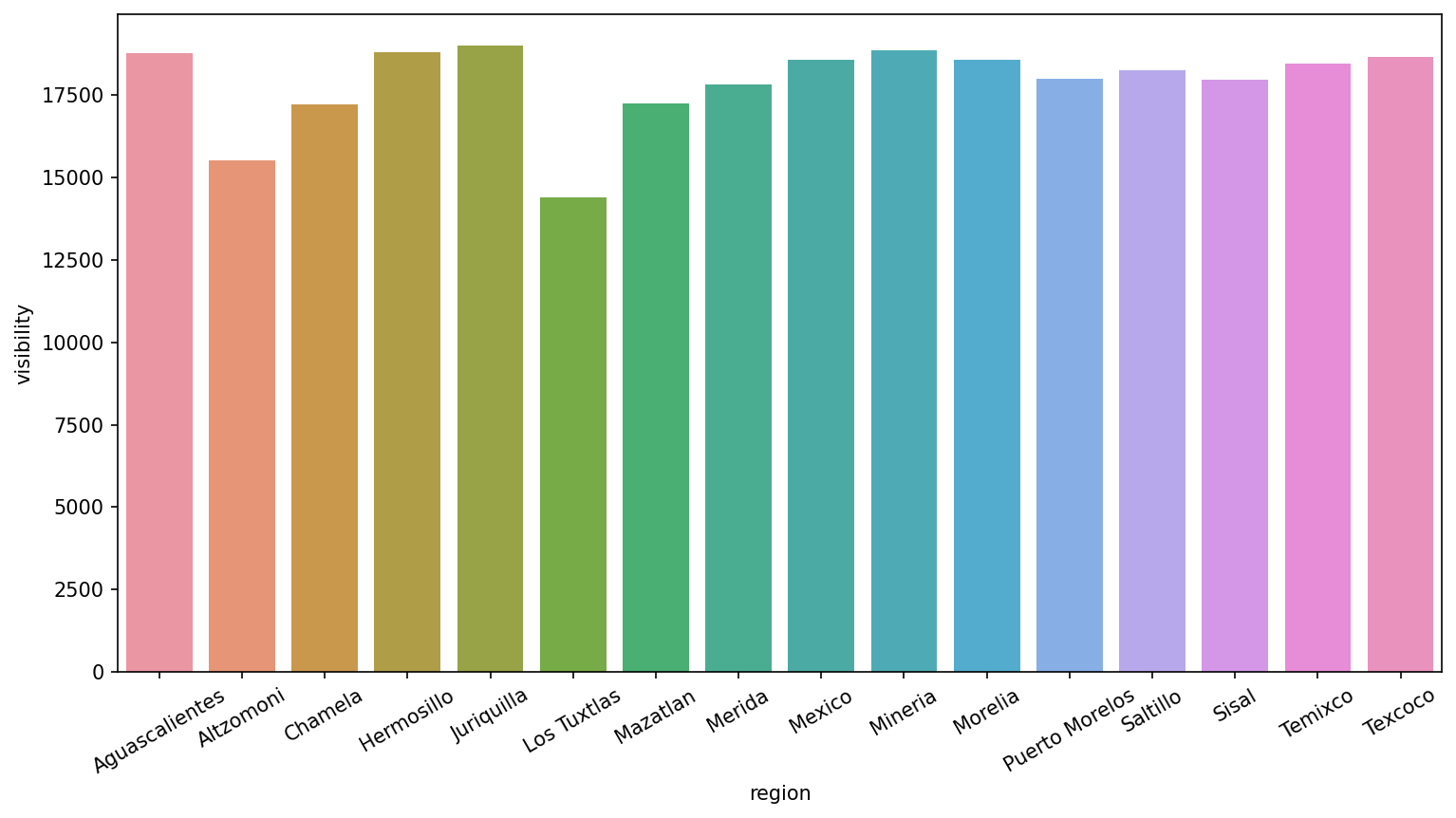
It is strange that hour of day stays on the last positions… I dont know is it reasonable or not. But, it depends on hour of day.



Month on the last position because we have week of year, it is correlated variables. But, week more useful for GBM so it uses week number istead of month.

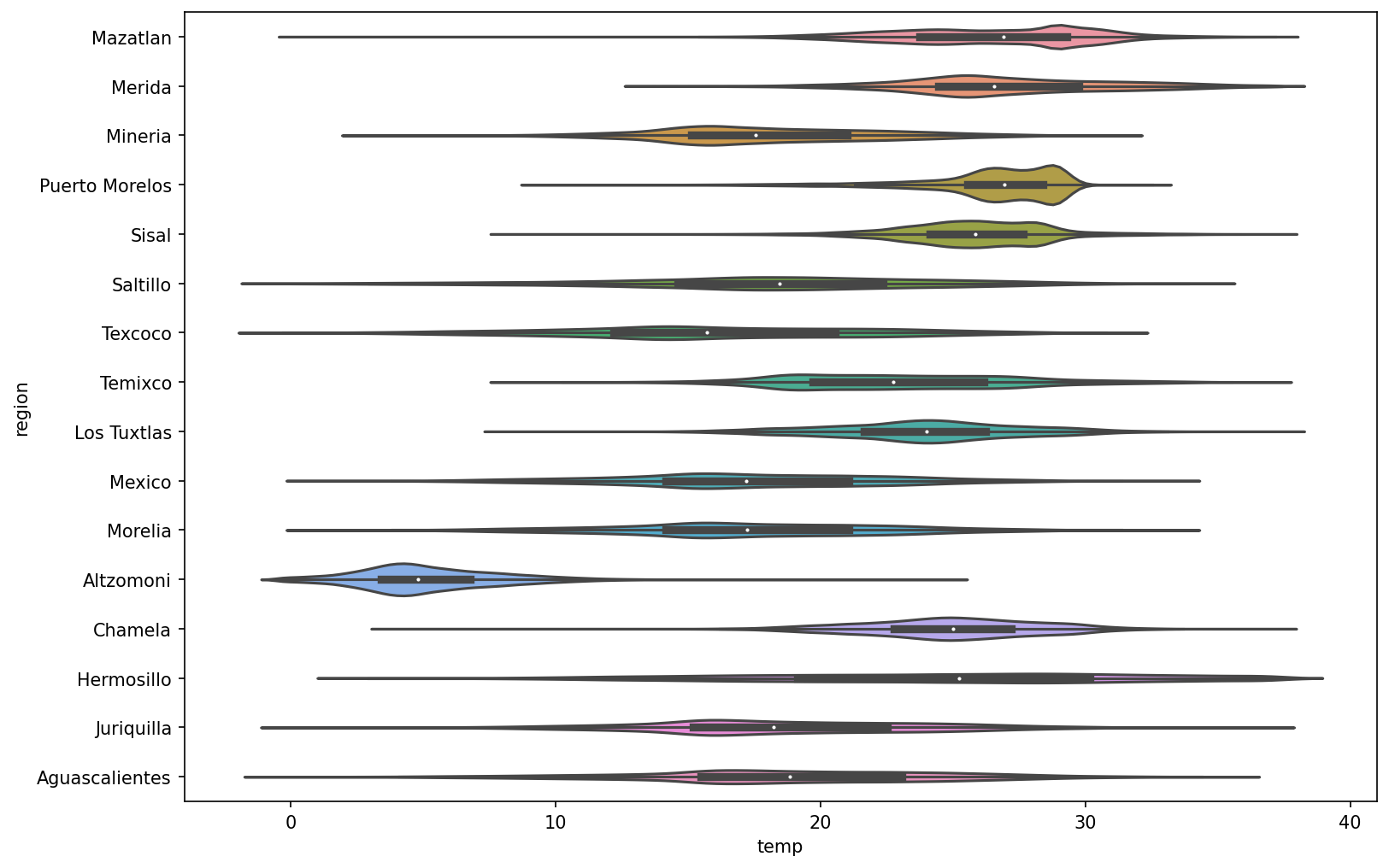
So, now we can see visibility in all regions:



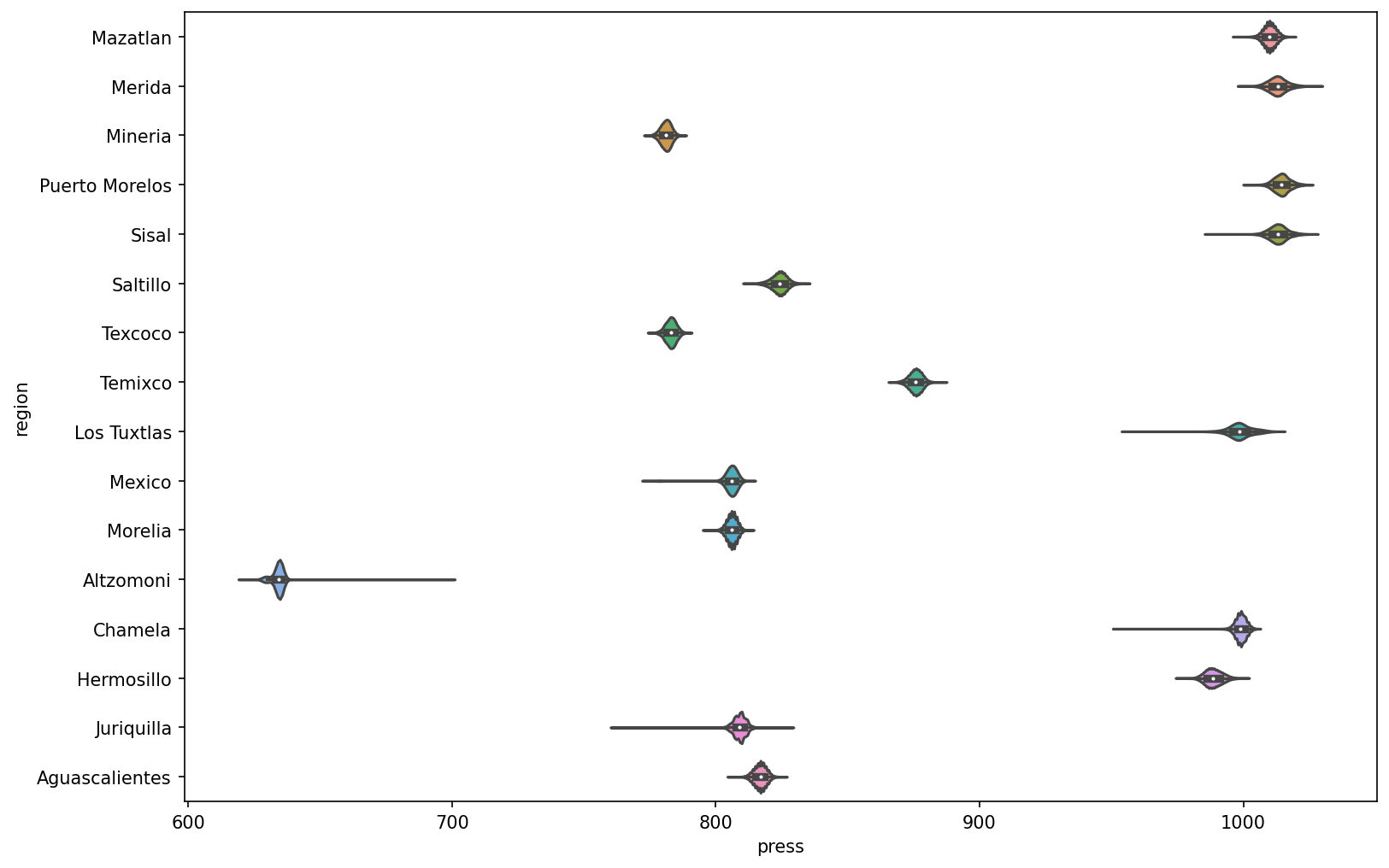


**Other data analyse**

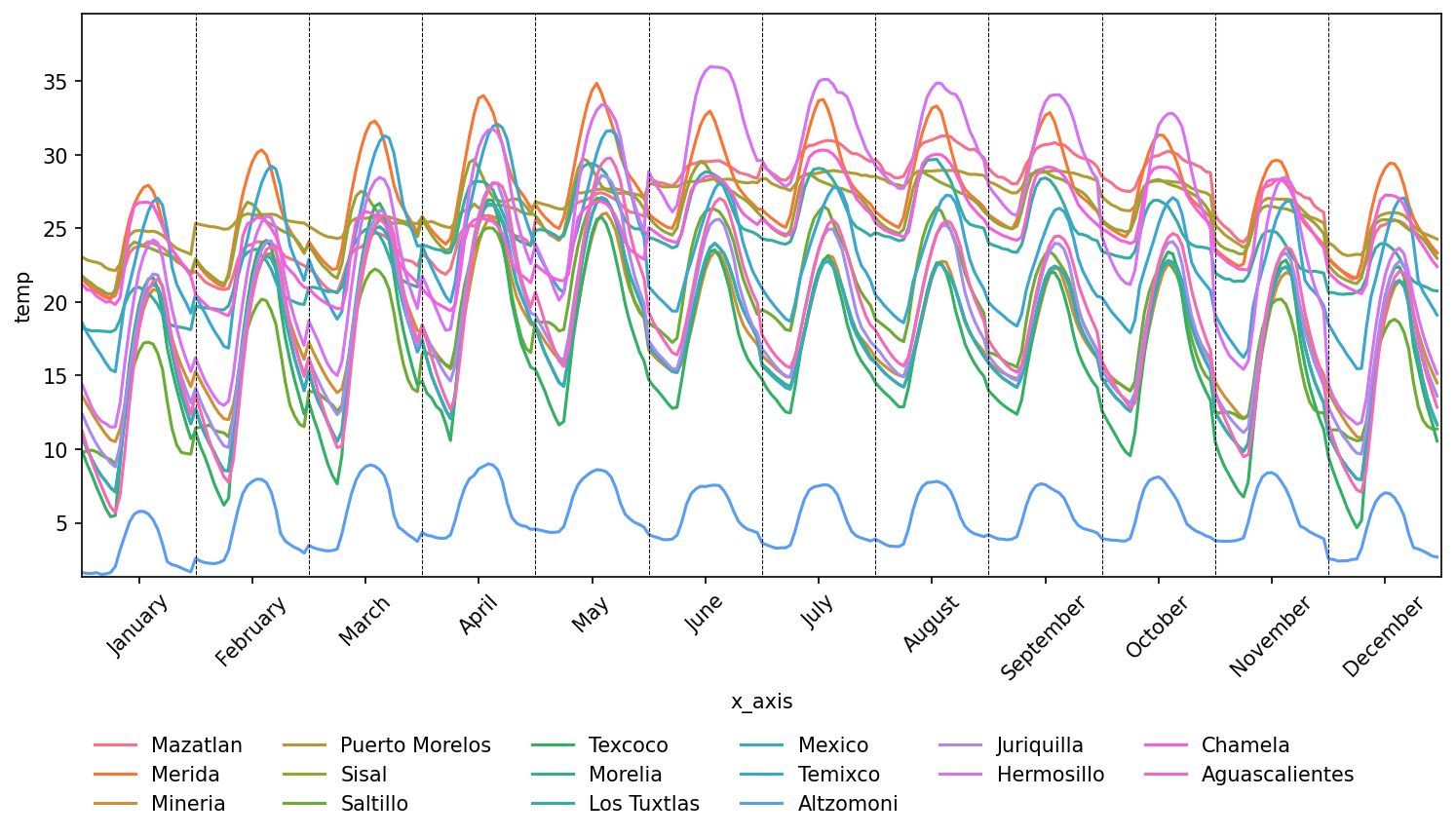
We can analyse distribution of different variables in different region. For example, we cant detect temperature in different regions.



Distribution of pressure has lower standart deviation



Also, we can show seasonality. For example mean temperature during a day in different month. Temperature was grouped by hour and month so, we have mean temperature during the day at certain month. It is seen that temperature rises at the morning and decreases at the evening. Also, difference of temperature during a day more, than difference during a year.



**Afterword**

Here's how you can do an exploratory data analysis. How can you predict the missing values. There are many ways to improve, for example - cleaning data by region, doing analysis and predictions by the minute and not by the hour - more data, more accurate predictions. Add other air pollution data. Apply more complex models - for example, neural networks.