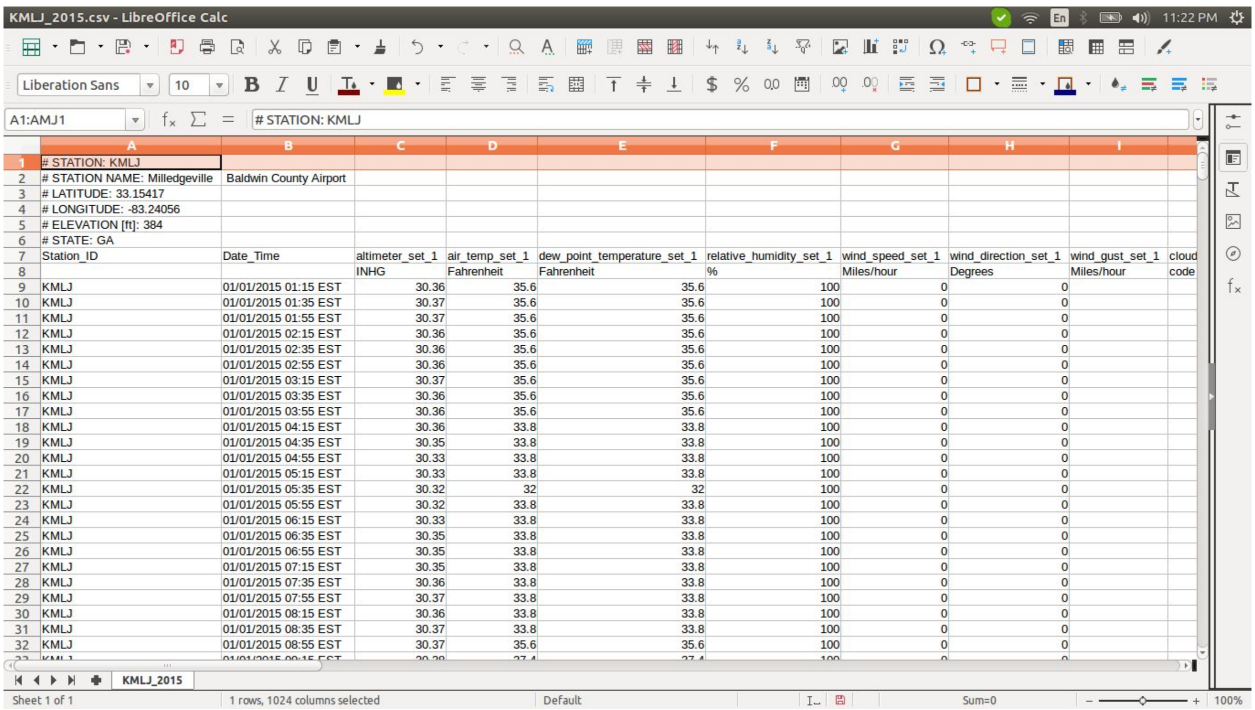
CSL 7020 - Assignment 1
Linear Regression

Vishakh.S (B16CS038) and Zaid Khan (B16CS040)

---

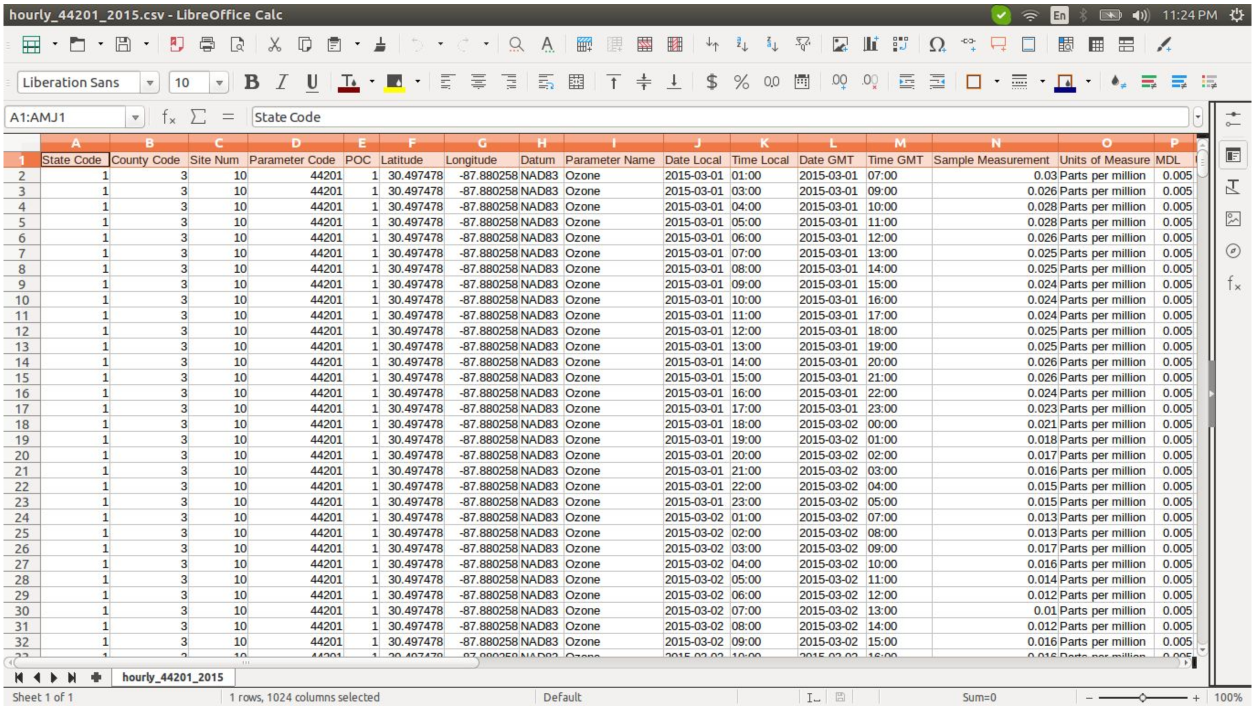
1. DATASET

Meteorological data for the station KMLJ, Georgia, USA is available for the year 2015. This dataset has been chosen over the other available ones as the other options lack a significant number of entries in the dataset, hindering the learning process.



	A	B	C	D	E	F	G	H	I	
1	# STATION: KMLJ									
2	# STATION NAME: Milledgeville	Baldwin County Airport								
3	# LATITUDE: 33.15417									
4	# LONGITUDE: -83.24056									
5	# ELEVATION [ft]: 384									
6	# STATE: GA									
7	Station_ID	Date_Time	altimeter_set_1	air_temp_set_1	dew_point_temperature_set_1	relative_humidity_set_1	wind_speed_set_1	wind_direction_set_1	wind_gust_set_1	cloud
8			INHG	Fahrenheit	Fahrenheit	%	Miles/hour	Degrees	Miles/hour	code
9	KMLJ	01/01/2015 01:15 EST	30.36	35.6	35.6	100	0	0		
10	KMLJ	01/01/2015 01:35 EST	30.37	35.6	35.6	100	0	0		
11	KMLJ	01/01/2015 01:55 EST	30.37	35.6	35.6	100	0	0		
12	KMLJ	01/01/2015 02:15 EST	30.36	35.6	35.6	100	0	0		
13	KMLJ	01/01/2015 02:35 EST	30.36	35.6	35.6	100	0	0		
14	KMLJ	01/01/2015 02:55 EST	30.36	35.6	35.6	100	0	0		
15	KMLJ	01/01/2015 03:15 EST	30.37	35.6	35.6	100	0	0		
16	KMLJ	01/01/2015 03:35 EST	30.36	35.6	35.6	100	0	0		
17	KMLJ	01/01/2015 03:55 EST	30.36	35.6	35.6	100	0	0		
18	KMLJ	01/01/2015 04:15 EST	30.36	33.8	33.8	100	0	0		
19	KMLJ	01/01/2015 04:35 EST	30.35	33.8	33.8	100	0	0		
20	KMLJ	01/01/2015 04:55 EST	30.33	33.8	33.8	100	0	0		
21	KMLJ	01/01/2015 05:15 EST	30.33	33.8	33.8	100	0	0		
22	KMLJ	01/01/2015 05:35 EST	30.32	32	32	100	0	0		
23	KMLJ	01/01/2015 05:55 EST	30.32	33.8	33.8	100	0	0		
24	KMLJ	01/01/2015 06:15 EST	30.33	33.8	33.8	100	0	0		
25	KMLJ	01/01/2015 06:35 EST	30.35	33.8	33.8	100	0	0		
26	KMLJ	01/01/2015 06:55 EST	30.35	33.8	33.8	100	0	0		
27	KMLJ	01/01/2015 07:15 EST	30.35	33.8	33.8	100	0	0		
28	KMLJ	01/01/2015 07:35 EST	30.36	33.8	33.8	100	0	0		
29	KMLJ	01/01/2015 07:55 EST	30.37	33.8	33.8	100	0	0		
30	KMLJ	01/01/2015 08:15 EST	30.36	33.8	33.8	100	0	0		
31	KMLJ	01/01/2015 08:35 EST	30.37	33.8	33.8	100	0	0		
32	KMLJ	01/01/2015 08:55 EST	30.37	35.6	35.6	100	0	0		

U.S. EPA's Air Quality System (AQS) database contains the air pollutant data for the period 1980 to 2019 measured at multiple stations across the country. For this assignment, the data for 2015 for three pollutants - ozone, particulate matter 2.5 and sulphur dioxide has been downloaded. (https://aqs.epa.gov/aqsweb/airdata/download_files.html#main-content)


hourly_44201_2015.csv - LibreOffice Calc

	State Code	County Code	Site Num	Parameter Code	POC	Latitude	Longitude	Datum	Parameter Name	Date Local	Time Local	Date GMT	Time GMT	Sample Measurement	Units of Measure	MDL
2	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	01:00	2015-03-01	07:00	0.03	Parts per million	0.005
3	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	03:00	2015-03-01	09:00	0.026	Parts per million	0.005
4	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	04:00	2015-03-01	10:00	0.028	Parts per million	0.005
5	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	05:00	2015-03-01	11:00	0.028	Parts per million	0.005
6	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	06:00	2015-03-01	12:00	0.026	Parts per million	0.005
7	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	07:00	2015-03-01	13:00	0.025	Parts per million	0.005
8	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	08:00	2015-03-01	14:00	0.025	Parts per million	0.005
9	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	09:00	2015-03-01	15:00	0.024	Parts per million	0.005
10	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	10:00	2015-03-01	16:00	0.024	Parts per million	0.005
11	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	11:00	2015-03-01	17:00	0.024	Parts per million	0.005
12	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	12:00	2015-03-01	18:00	0.025	Parts per million	0.005
13	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	13:00	2015-03-01	19:00	0.025	Parts per million	0.005
14	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	14:00	2015-03-01	20:00	0.026	Parts per million	0.005
15	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	15:00	2015-03-01	21:00	0.026	Parts per million	0.005
16	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	16:00	2015-03-01	22:00	0.024	Parts per million	0.005
17	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	17:00	2015-03-01	23:00	0.023	Parts per million	0.005
18	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	18:00	2015-03-02	00:00	0.021	Parts per million	0.005
19	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	19:00	2015-03-02	01:00	0.018	Parts per million	0.005
20	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	20:00	2015-03-02	02:00	0.017	Parts per million	0.005
21	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	21:00	2015-03-02	03:00	0.016	Parts per million	0.005
22	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	22:00	2015-03-02	04:00	0.015	Parts per million	0.005
23	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-01	23:00	2015-03-02	05:00	0.015	Parts per million	0.005
24	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	01:00	2015-03-02	07:00	0.013	Parts per million	0.005
25	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	02:00	2015-03-02	08:00	0.013	Parts per million	0.005
26	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	03:00	2015-03-02	09:00	0.017	Parts per million	0.005
27	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	04:00	2015-03-02	10:00	0.016	Parts per million	0.005
28	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	05:00	2015-03-02	11:00	0.014	Parts per million	0.005
29	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	06:00	2015-03-02	12:00	0.012	Parts per million	0.005
30	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	07:00	2015-03-02	13:00	0.01	Parts per million	0.005
31	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	08:00	2015-03-02	14:00	0.012	Parts per million	0.005
32	1	3	10	44201	1	30.497478	-87.880258	NAD83	Ozone	2015-03-02	09:00	2015-03-02	15:00	0.016	Parts per million	0.005

hourly_44201_2015

Sheet 1 of 1 | 1 rows, 1024 columns selected | Default | Sum=0 | 100%

2. APPROACH

a. PRE-PROCESSING

The meteorological data for 2015 for the station Baldwin County airport, Georgia has been taken for the study. A bunch of meteorological variables affect the concentration of pollutants in the air at any time. The following variables have been considered for the purpose of this study:

Height	Air temperature
Relative humidity	Wind speed
Wind direction	Wind gust
Cloud layer	Visibility
Dew point temperature	Wind chill
Wind cardinal direction	Pressure
Sea level pressure	Weather conditions

➔ For each numeric variable, if there were multiple observed records within an hour, then compute and use the mean value.

➔ For Date_Time object, if there were multiple observed records within an hour, then compute and use the date and current running hour of the day.

➔ There are missing values for some variables. This is not suitable for applying supervised machine learning methods. Therefore, the missing values have been inferred using constant interpolation.

.i.e. They have been given a value obtained as follows :
Let the value for interpolation be **x**
Require that the mean after interpolation is **x** itself
Then, **x** is the solution of the equation :

$$x = \frac{Sum\ of\ all\ values\ present + x + x + \ldots\ldots + x\ (for\ each\ missing\ value)}{Number\ of\ samples}$$

➔ Dummy coding has been applied for two categorical variables, the cardinal wind direction (16 values, e.g., N, S, E, W, etc) and weather conditions (31 values, e.g., sunny, rainy, windy, etc).

.i.e. Create a column N with a value of 1 when cardinal wind direction for that sample is N. Assign a value of 0, if not N. The same is done for all other columns generated for wind direction and weather conditions.

Finally, we obtained 53 features in total (18 numerical meteorological features, 16 dummy codings for wind direction, 19 dummy codings for weather conditions).
The numerical features considered are :

altimeter_set_1	air_temp_set_1
dew_point_temperature_set_1	relative_humidity_set_1
wind_speed_set_1	wind_direction_set_1
wind_gust_set_1	cloud_layer_1_code_set_1
cloud_layer_2_code_set_1	visibility_set_1
weather_cond_code_set_1	cloud_layer_3_code_set_1
ceiling_set_1	dew_point_temperature_set_1d
pressure_set_1d	sea_level_pressure_set_1d
heat_index_set_1d	wind_chill_set_1d

The number of numerical features considered .i.e. 53 is the size of the input feature space, denoted by d.

The pollutant data :

➔ hourly_44201_2015.csv - ozone data,
➔ hourly_42401_2015.csv - sulphur dioxide data and
➔ hourly_88101_2015.csv - particulate matter 2.5 data

has been preprocessed by eliminating all columns except the local date, local time and the sample measurement. The sample measurement is the ppm value which forms the target for the learning task. The local date and time have been merged into a pandas data_time object for the purpose of combining with the meteorological dataset.

Take a natural join of the pollutants dataset with the meteorological dataset along the Date_Time column. This will generate the training samples of the form (**X**, Y) where X is the input data and Y is the target ppm value. **X** is the input vector of dimension m by d, where m is the number of input samples. Y is a vector of the form m by 1 and contains the target value corresponding to each tuple of d features in **X**. The generated samples (**X**, Y) are used for training the linear regression model.

6. MODEL DESCRIPTION

Linear regression has been used for modelling the relation between meteorological Variables and the real valued target, the concentration in parts per million (ppm) of the pollutant.

We would like to learn a linear function, $h : \mathbb{R}^d \rightarrow \mathbb{R}$ that best approximates the relationship between the variables.

i.e.

$$\boxed{h_{\underline{w}}(\underline{x}) = \langle \underline{w}, \underline{x} \rangle + b}$$

where $\underline{w} \in \mathbb{R}^d$

and $b \in \mathbb{R}$

If y denotes the target values, we can define the squared-loss function for learning as follows.:

$$l(h(\underline{x}, y)) = (h(x) - y)^2$$

For this loss function, the emperical risk function is called the Mean Squared Error, namely, for a sample set of size m:

$$\boxed{L_s(h) = \frac{1}{m} \sum_{i=1}^{m} [h(\underline{x}_i) - y_i]^2}$$

Given a training set S, the goal is to find

$$\underset{\underline{w}}{\operatorname{argmin}}\ L_S(h_w) = \underset{\underline{w}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} (\langle \underline{w}, \underline{x_i} \rangle - y_i)^2$$

To solve the problem of minimization, we use gradient descent & update $\underline{w}$ as follows

$$\boxed{\underline{w}^{t+1} = \underline{w}^{t} - \eta \nabla (L_S(h_w))}$$

where $\eta > 0$ is the learning rate

And the gradient of the objective function,

$$\boxed{\nabla (L_S(h_w)) = \frac{2}{m} \sum_{i=1}^{m} (\langle \underline{w}, \underline{x_i} \rangle - y_i) x_i}$$

Eventually, after T iterations, the algorithm outputs the averaged vector,

$$\boxed{\bar{w} = 1/T \sum_{t=1}^{T} w^{(t)}}$$

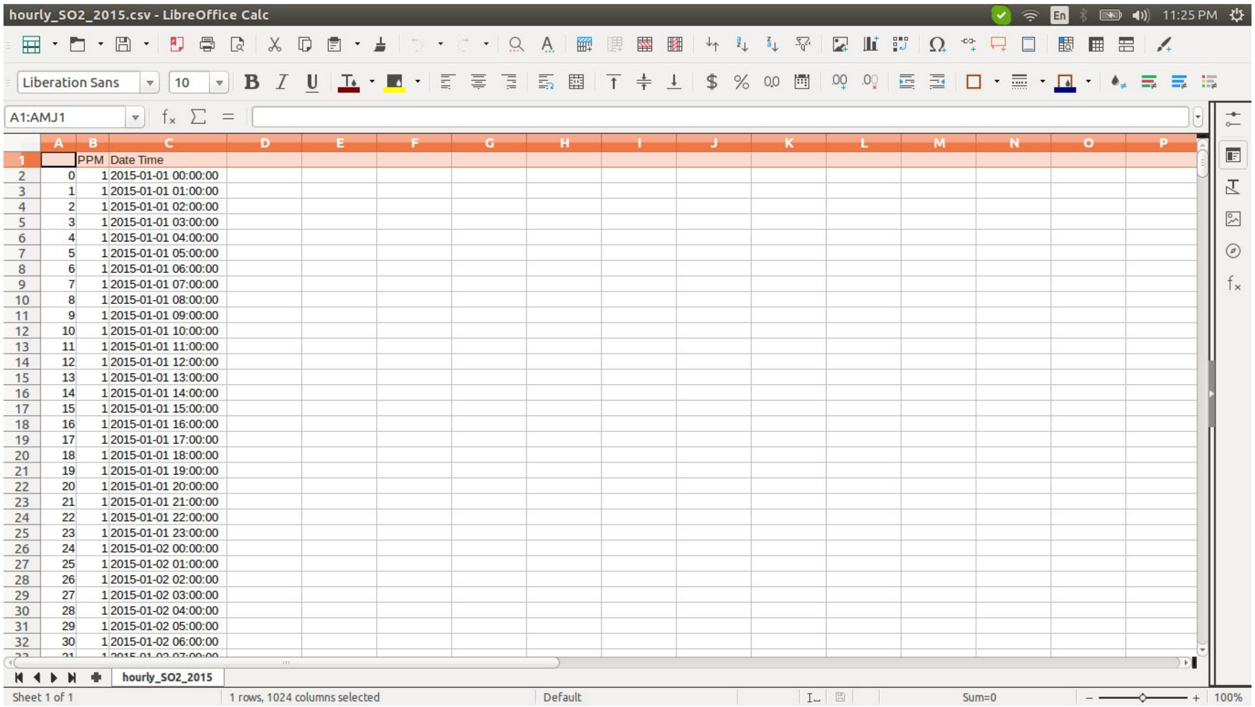
3. EXPERIMENTS

a. SETUP

- ➔ Set up Python3 virtual environment.
- ➔ Install math, numpy and pandas.
- ➔ Download the requisite datasets (pollutant - hourly_44201_2015.csv, hourly_42401_2015.csv, hourly_88101_2015.csv and meteorological - KMLJ_2015.csv).
- ➔ Preprocess the data to eliminate the unwanted information. (as described in section 2(a))
- ➔ Take a natural join of the pollutants data with the meteorological data to generate the training samples (**X,** Y)
- ➔ Train the linear regression model on the training data.
- ➔ Perform affine to homogenous conversion by appending a column of ones to X and an extra bias term in the parameter vector w.
- ➔ Use gradient descent for minimizing the regularized loss.
- ➔ After a specific number of iterations, we obtain the value of w that minimizes the emperical risk.
- ➔ Predict the output based on the learnt parameter value.

b. RESULTS

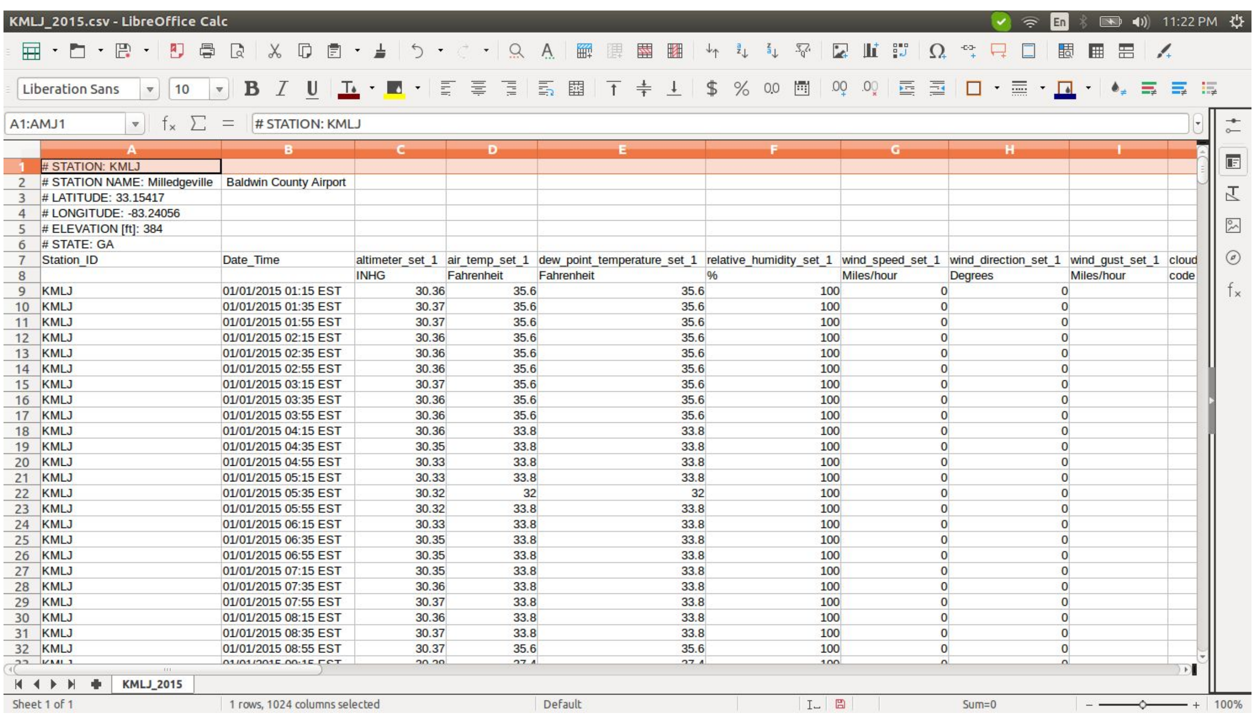
After preprocessing, the pollutant data for sulphur dioxide and the meteorological data are as shown below.



hourly_SO2_2015.csv - LibreOffice Calc

	A	B	C
1		PPM	Date Time
2	0	1	2015-01-01 00:00:00
3	1	1	2015-01-01 01:00:00
4	2	1	2015-01-01 02:00:00
5	3	1	2015-01-01 03:00:00
6	4	1	2015-01-01 04:00:00
7	5	1	2015-01-01 05:00:00
8	6	1	2015-01-01 06:00:00
9	7	1	2015-01-01 07:00:00
10	8	1	2015-01-01 08:00:00
11	9	1	2015-01-01 09:00:00
12	10	1	2015-01-01 10:00:00
13	11	1	2015-01-01 11:00:00
14	12	1	2015-01-01 12:00:00
15	13	1	2015-01-01 13:00:00
16	14	1	2015-01-01 14:00:00
17	15	1	2015-01-01 15:00:00
18	16	1	2015-01-01 16:00:00
19	17	1	2015-01-01 17:00:00
20	18	1	2015-01-01 18:00:00
21	19	1	2015-01-01 19:00:00
22	20	1	2015-01-01 20:00:00
23	21	1	2015-01-01 21:00:00
24	22	1	2015-01-01 22:00:00
25	23	1	2015-01-01 23:00:00
26	24	1	2015-01-02 00:00:00
27	25	1	2015-01-02 01:00:00
28	26	1	2015-01-02 02:00:00
29	27	1	2015-01-02 03:00:00
30	28	1	2015-01-02 04:00:00
31	29	1	2015-01-02 05:00:00
32	30	1	2015-01-02 06:00:00

hourly_SO2_2015 — Sheet 1 of 1 — 1 rows, 1024 columns selected — Default — Sum=0 — 100%

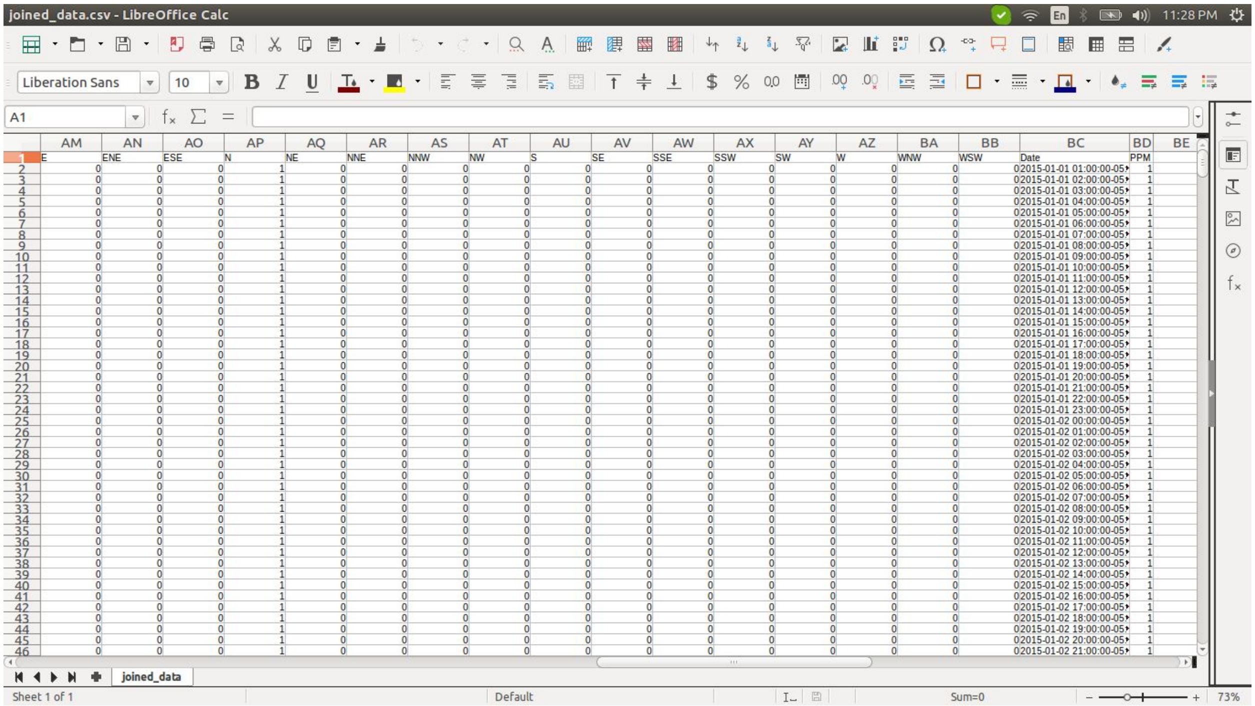


KMLJ_2015.csv - LibreOffice Calc

	A	B	C	D	E	F	G	H	I
1	# STATION: KMLJ								
2	# STATION NAME: Milledgeville	Baldwin County Airport							
3	# LATITUDE: 33.15417								
4	# LONGITUDE: -83.24056								
5	# ELEVATION [ft]: 384								
6	# STATE: GA								
7	Station_ID	Date_Time	altimeter_set_1	air_temp_set_1	dew_point_temperature_set_1	relative_humidity_set_1	wind_speed_set_1	wind_direction_set_1	wind_gust_set_1
8			INHG	Fahrenheit	Fahrenheit	%	Miles/hour	Degrees	Miles/hour
9	KMLJ	01/01/2015 01:15 EST	30.36	35.6	35.6	100	0	0	
10	KMLJ	01/01/2015 01:35 EST	30.37	35.6	35.6	100	0	0	
11	KMLJ	01/01/2015 01:55 EST	30.37	35.6	35.6	100	0	0	
12	KMLJ	01/01/2015 02:15 EST	30.36	35.6	35.6	100	0	0	
13	KMLJ	01/01/2015 02:35 EST	30.36	35.6	35.6	100	0	0	
14	KMLJ	01/01/2015 02:55 EST	30.36	35.6	35.6	100	0	0	
15	KMLJ	01/01/2015 03:15 EST	30.37	35.6	35.6	100	0	0	
16	KMLJ	01/01/2015 03:35 EST	30.36	35.6	35.6	100	0	0	
17	KMLJ	01/01/2015 03:55 EST	30.36	35.6	35.6	100	0	0	
18	KMLJ	01/01/2015 04:15 EST	30.36	33.8	33.8	100	0	0	
19	KMLJ	01/01/2015 04:35 EST	30.35	33.8	33.8	100	0	0	
20	KMLJ	01/01/2015 04:55 EST	30.33	33.8	33.8	100	0	0	
21	KMLJ	01/01/2015 05:15 EST	30.33	33.8	33.8	100	0	0	
22	KMLJ	01/01/2015 05:35 EST	30.32	32	32	100	0	0	
23	KMLJ	01/01/2015 05:55 EST	30.32	33.8	33.8	100	0	0	
24	KMLJ	01/01/2015 06:15 EST	30.33	33.8	33.8	100	0	0	
25	KMLJ	01/01/2015 06:35 EST	30.35	33.8	33.8	100	0	0	
26	KMLJ	01/01/2015 06:55 EST	30.35	33.8	33.8	100	0	0	
27	KMLJ	01/01/2015 07:15 EST	30.35	33.8	33.8	100	0	0	
28	KMLJ	01/01/2015 07:35 EST	30.36	33.8	33.8	100	0	0	
29	KMLJ	01/01/2015 07:55 EST	30.37	33.8	33.8	100	0	0	
30	KMLJ	01/01/2015 08:15 EST	30.36	33.8	33.8	100	0	0	
31	KMLJ	01/01/2015 08:35 EST	30.37	33.8	33.8	100	0	0	
32	KMLJ	01/01/2015 08:55 EST	30.37	35.6	35.6	100	0	0	

KMLJ_2015 — Sheet 1 of 1 — 1 rows, 1024 columns selected — Default — Sum=0 — 100%

After computing the natural join of the two data, the combined data is shown below.



4. ANALYSIS

When the learning rate is more than 1e-4, the cost starts overshooting to a value of 1e20. So the learning rate should be as small as possible.