S K Somaiya College (SKSC)

DEPARTMENT OF STATISTICS

131P18E301: Time Series Analysis

MSc GROUP ASSIGNMENT

(2021 – 2022)

Reference link: https://archive.ics.uci.edu/ml/datasets/Air+quality

Roll Number	Name	Email ID
31031820020	Varsha Jadhav	jadhav.vm@somaiya.edu
31031820027	Niki Mehta	niki.mehta@somaiya.edu
31031820038	Jyothi Puligilla	jyothi.puligilla@somaiya.edu
31031820041	Vedang Sawant	vedang.s@somaiya.edu
31031820044	Moheed Tai	mohammedmoheed.t@somaiya.edu
31031820046	Vishal Yadav	vishal05@somaiya.edu

Prof. Mayur More
Professor Incharge

INDEX

Abstract:

The total of 9358 instances of hourly average responses from an array of 5 metal oxide chemical sensors are collected which are regarded as air pollutants and are observed, analyzed and conclusion is given on the effect of their presence in the air after doing some statistical time series analysis. Identifying the problem statement and preparing a basic methodology for the dataset is the first step carried out. Data dictionary and structure of the data is described for better understanding of the analyzed data. The next step is data analysis which involved finding missing values and visualize the data. In final step we forecasted regression line. Every dataset has limitation this dataset has large number of missing values, the dataset is 15 years old and the geographical location of some data points were missing.

Introduction:

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metallic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value. This dataset can be used exclusively for research purposes.

Literature review:

Anikender Kumar, Pramila Goyal (2011) presented the study that forecasts the daily AQI (Air Quality Index) value for the city Delhi, India using previous record of AQI and meteorological parameters with the help of Principal Component Regression (PCR) and Multiple Linear Regression Techniques. They perform the prediction of daily AQI of the year 2006 using previous records of the year 2000-2005 and different equations. After that this predicted value then compared with observed value of AQI of 2006 for the seasons summer, Monsoon, Post Monsoon and winter using Multiple Linear Regression Technique. Principal Component Analysis is used to find the collinearity among the independent variables. The Principal components were used in Multiple Linear Regression to eliminate collinearity among the predictor variables and also reduce the number of predictors. The Principal Component Regression gives the better performance for predicting the AQI in winter season than any other seasons. In this study only meteorological parameters were considered or used while forecasting the future AQI but they have not considered the ambient air pollutants that may cause the adverse health effects.

Objective:

- Checking the normality of the variables
- Forecasting Air Quality features using Exponential Smoothing, ARIMA, and Linear regression and finding which is the best fit.

Data dictionary:

1.Date(DD/MM/YYYY)
2.Time(HH.MM.SS)
3.CO.R=True hourly averaged concentration CO in mg/m^3 (reference analyzer)
4.CO.T=PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
5.NMHC.R=True hourly averaged overall Non Metanic Hydro Carbons concentration in microg/m^3(reference analyzer)
6.NMHC.T=PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7.NOx.R=True hourly averaged NOx concentration in ppb (reference analyzer)
8.NOx.T=PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9.NO2.R=True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10.NO2.T=PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11.C6H6.R=True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
12.O3.T=PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
13.T=Temperature in Â°C
14.RH= Relative Humidity (%)
15.AH=Absolute Humidity

Exploratory Data Analysis:

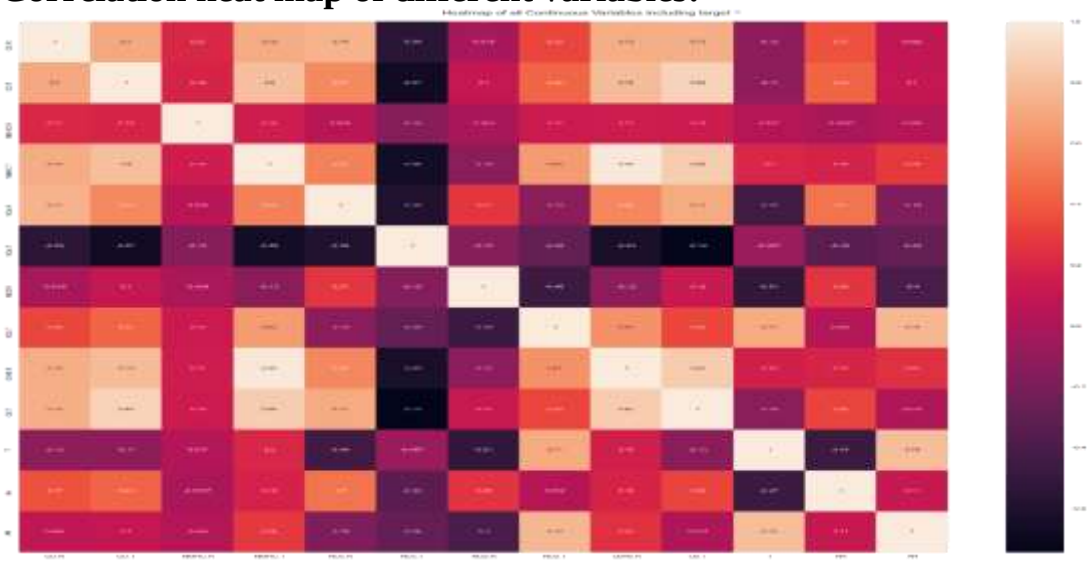
Description of the variables

```
In [17]: M ddf.describe().round(2)
```

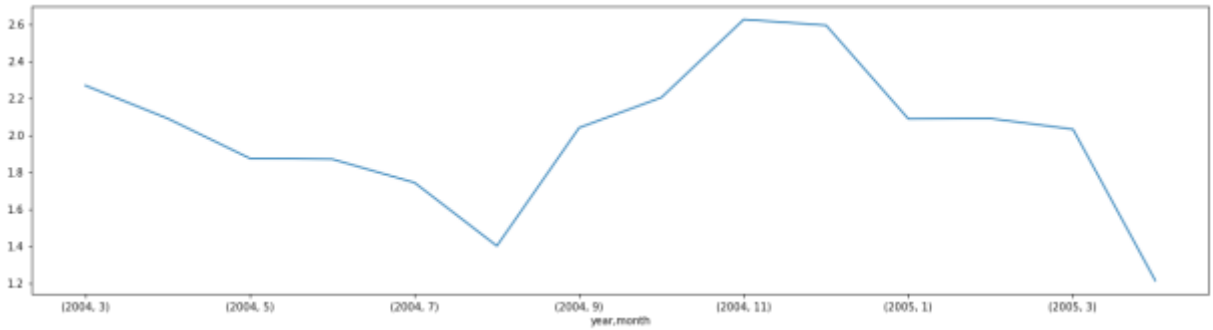
Out[17]:

	DATE	CO.R	CO.T	NMHC.R	NMHC.T	NOx.R	NOx.T	NO2.R	NO2.T	C6H6.R	O3.T	T	RH	AH
count	391.00	391.00	391.00	391.00	391.00	391.00	391.00	391.00	391.00	391.00	391.00	391.00	391.00	391.00
mean	38251.00	2.15	1102.87	219.14	941.17	243.37	833.77	109.66	1452.52	10.15	1031.72	18.22	49.20	1.02
std	113.02	0.93	145.47	43.84	154.36	155.06	173.93	32.89	273.45	4.28	288.42	7.88	12.95	0.39
min	38056.00	0.11	724.70	48.62	524.49	33.12	495.30	45.23	898.94	1.43	379.96	1.43	19.29	0.24
25%	38153.50	1.47	992.82	218.00	829.17	125.30	705.78	85.58	1276.51	6.83	815.49	11.99	38.96	0.73
50%	38251.00	2.08	1097.72	218.00	943.85	194.58	823.82	105.77	1504.72	9.98	1013.89	17.79	48.74	1.00
75%	38348.50	2.62	1201.56	218.00	1053.28	330.52	939.58	129.02	1668.39	13.11	1220.84	24.77	58.97	1.30
max	38446.00	5.80	1513.21	552.72	1426.50	824.31	1678.69	237.91	1959.21	24.98	2092.37	33.00	81.10	2.00

Correlation heat map of different variables:

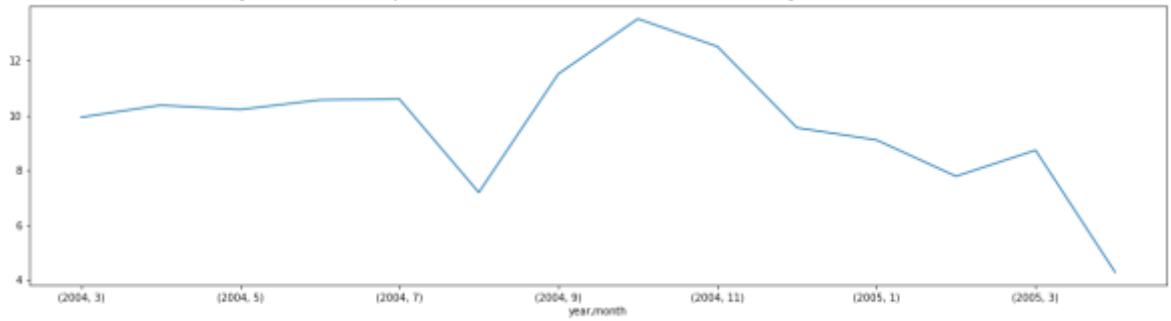


CO average monthly concentration in mg/m^3



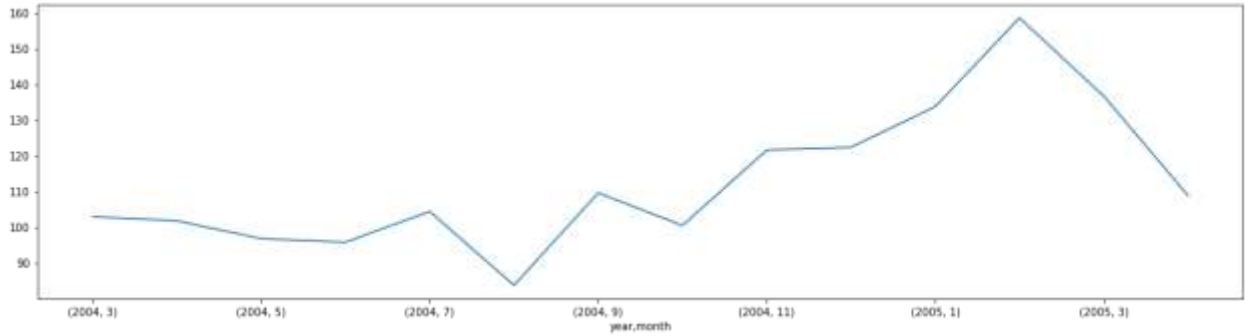
From the above graph, we can see that concentration of the CO decreases In the Month august, increases in the Month November to December & again decreases in the Month April.

Benzene average monthly concentration in microg/m^3



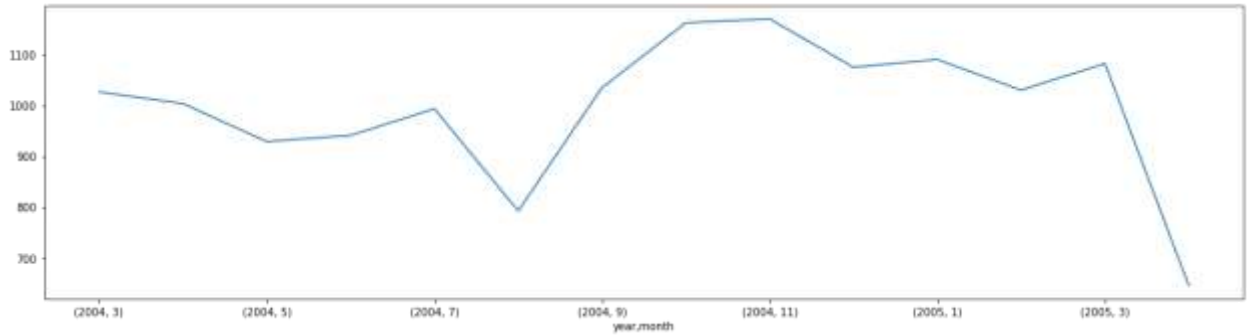
From the above graph, we can see that the concentration of Benzene decreases in the Month August, increases in the Month October & again decreases in the Month April.

Nitrogen oxide average monthly sensor response



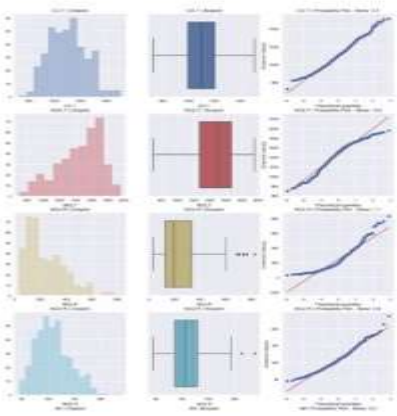
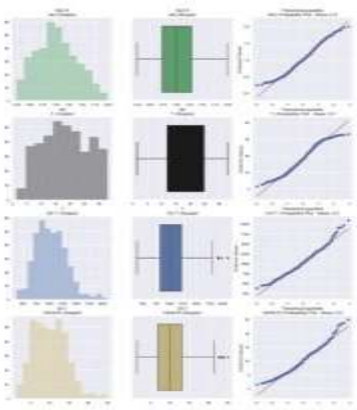
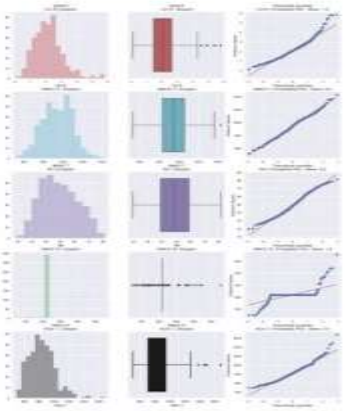
From the above graph, we can see that the sensor response of Nitrogen oxide decreases in the Month August, increases in the Month February.

Ozone average monthly sensor response



From the above graph, we can see that the sensor response of Ozone decreases in the Month August, increases in the Month October & again decreases in the Month April.

Visualizing the Data using Normality plot, Box plot, and Q-Q plot:

Normally Distributed :

If the histogram is bell shape and symmetrically distributed.
If the median line in the box plot is at the center.
In Q-Q plot if all the point lies on the straight line.

Right Skewed :

If in the histogram there is long tail in the positive direction(right direction).
If the median line in the box plot is on the left side.
In the Q-Q plot if the upper end of the Q-Q plot deviates from the straight line but the bottom end is not.

Left Skewed :

If in the histogram there is long tail in the negative direction(left direction).
If the median line in the box plot is on the right side.
In the Q-Q plot if the bottom end of the Q-Q plot deviates from the straight line but the upper end is not.

Methodology:

1. Data preprocessing
2. ARIMA
3. Simple Exponential Smoothing
4. Double Exponential Smoothing
5. Linear Regression

Analysis:

1.Data preprocessing:

The air quality data was extracted from UCI. There were 15 variables in our data with 9358 observations. We found missing values in every variable so we treated them by moving average method and the variable NMHC.R had around 90% missing values so it was difficult to apply moving average method and hence we replaced it by column mean. Also the variables were in character form and hence we converted them into the alphanumeric form. The data was taken on hourly basis for each day so we took the average of each day by hours. By this result, we got the dataset of 391 values.

a) ARIMA Model

In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity of the mean function (i.e., the trend). When the seasonality shows in a time series, the seasonal-differencing could be applied to eliminate the seasonal component. Since the ARMA model, according to the Wald's decomposition theorem, is theoretically sufficient to describe a regular (a.k.a. purely nondeterministic) wide-sense stationary time series, we are motivated to make stationary a non-stationary time series, e.g., by using differencing, before we can use the ARMA model. Note that if the time series contains a predictable sub-process (a.k.a. pure sine or complex-valued exponential process), the predictable component is treated as a non-zero-mean but periodic (i.e., seasonal) component in the ARIMA framework so that it is eliminated by the seasonal differencing.

ARIMA Model

$$Y_t = \beta_1 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p}$$

b) Simple Exponential Smoothing:

Single Exponential Smoothing, SES for short, also called Simple Exponential Smoothing, is a time series forecasting method for univariate data without a trend or seasonality. It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.It requires a single parameter, called alpha (a), also called the smoothing factor or smoothing coefficient.

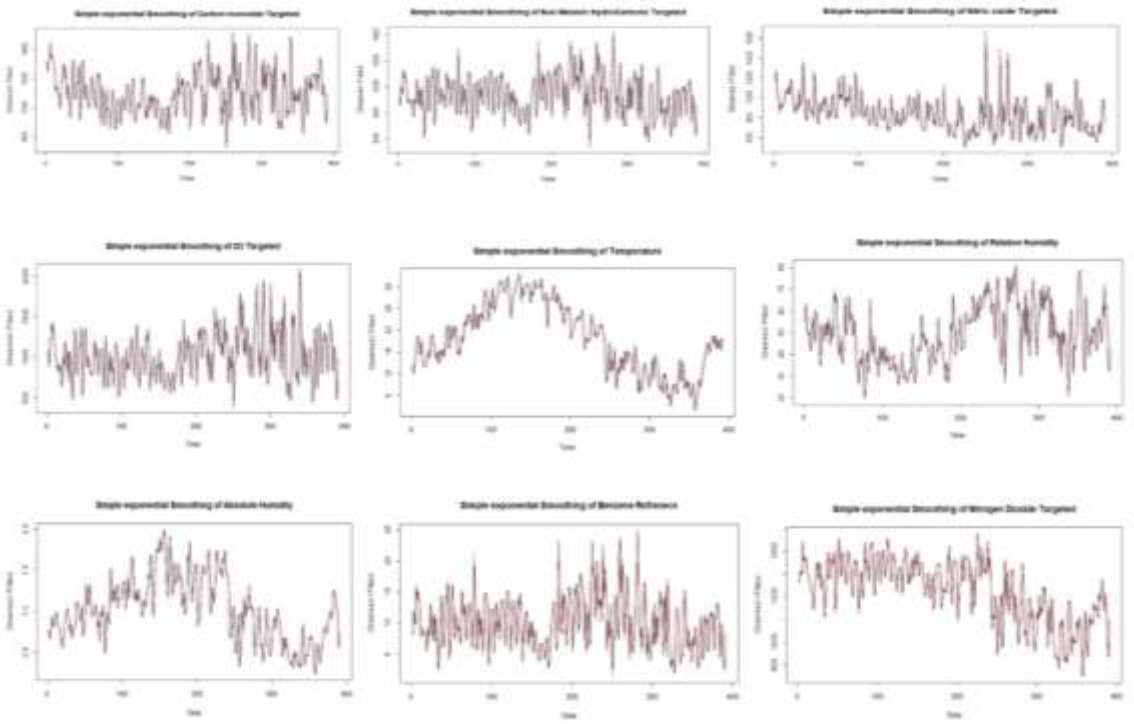
This parameter controls the rate at which the influence of the observations at prior time steps decay exponentially. Alpha is often set to a value between 0 and 1. Large values mean that the model

pays attention mainly to the most recent past observations, whereas smaller values mean more of the history is taken into account when making a prediction.

$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$$

where: F_t = new forecast
F_{t-1} = previous period forecast
A_{t-1} = previous period *actual* demand
α = smoothing (weighting) constant

Analysis Of Simple Exponential Smoothing model



c) Double Exponential Smoothing:

Simple exponential smoothing does not do well when there is a trend in the data. In such situations, several methods were devised under the name "double exponential smoothing" or "second-order exponential smoothing," which is the recursive application of an exponential filter twice, thus being termed "double exponential smoothing". This nomenclature is similar to quadruple exponential smoothing, which also references its recursion depth.The basic idea behind double exponential smoothing is to introduce a term to take into account the possibility of a series exhibiting some form of trend.

Double exponential Smoothing model

$$C_t = \alpha y_t + (1-\alpha)(C_{t-1} + T_{t-1})$$

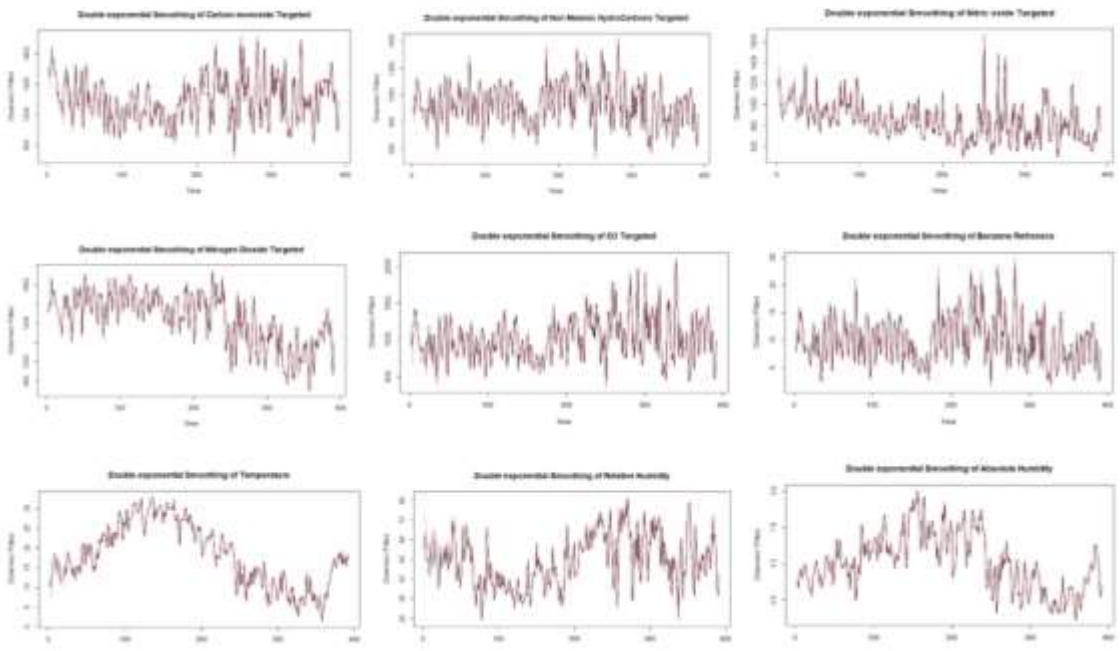
$$T_t = \beta(C_t - C_{t-1}) + (1-\beta)T_{t-1}$$

$$F_{t+1} = C_t + T_t$$

where:

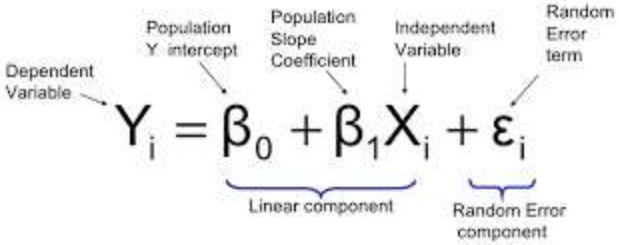
y_t = actual value in time t
α = constant-process smoothing constant
β = trend-smoothing constant
C_t = smoothed constant-process value for period t
T_t = smoothed trend value for period t
F_{t+1} = forecast value for period t + 1
t = current time period

Analysis Of Double Exponential Smoothing model



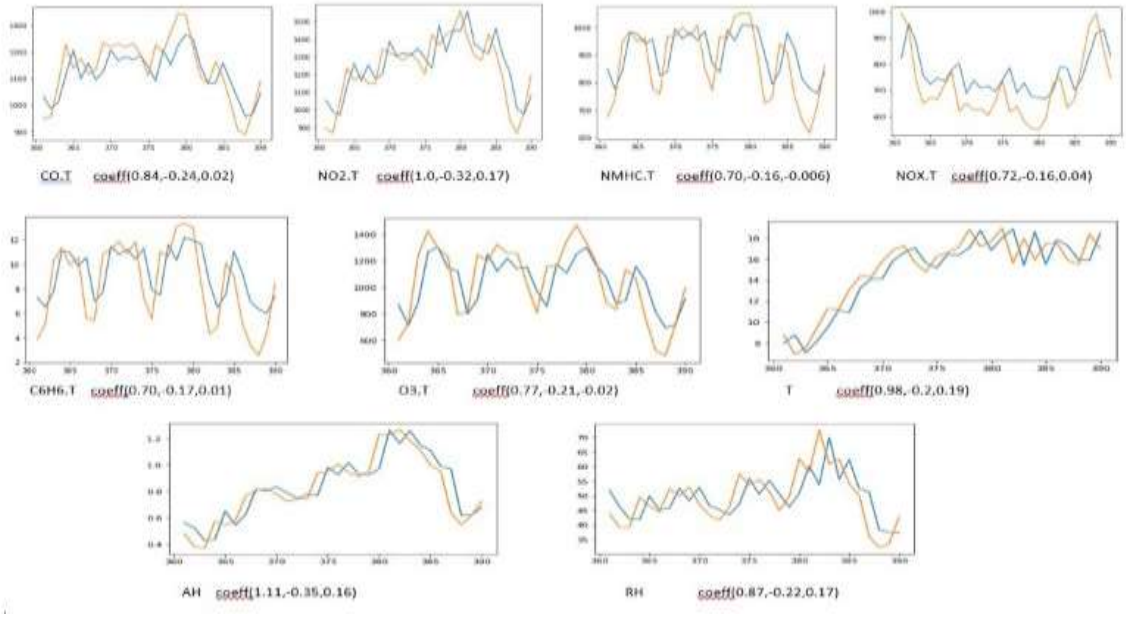
d) Forecasting using Linear Regression:

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.



Here for this data dependent variable is target variable (eg CO.T) and independent variables are lag variables. Last 30 observations were taken for prediction.

Analysis Of Linear Regression Model



Yellow line represents predicted and blue shows trained data, here also more weightage is given to the immediate previous values and from the above diagrams we can say that the model fitted has a good accuracy.

e) Table of RMSE values of the models fitted using ARIMA, Linear Regression and Exponential Smoothing:

Variable	RMSE of ARIMA	Best fit of ARIMA model	RMSE of linear regression	RMSE of single exponential	RMSE of double exponential
CO.T	150.814	Best model: ARIMA(2,0,0)(0,0,0)[0]	71.67	113.40	115.98
NMHC.T	209.76	Best model: ARIMA(2,1,1)(0,0,0)[0]	99.59	135.74	138.26
NOx.T	163.27	Best model: ARIMA(2,1,2)(0,0,0)[0]	97.86	143.32	147.99
NO2.T	248.84	Best model: ARIMA(2,1,1)(0,0,0)[0]	110.94	144.08	145.05
C6H6	4.56	Best model: ARIMA(2,0,0)(0,0,0)[0]	2.55	3.80	3.83
O3	393.08	Best model: ARIMA(2,1,1)(0,0,0)[0]	188.42	246.86	254.56
T			1.51	2.00	2.08
RH			**7.16**	**8.24**	**8.59**
AH			0.10	0.15	0.15

For the variables T, RH and AH the best model for ARIMA was not predicted since they did not have any seasonality

Conclusion:

- From the Exploratory data analysis we get that NO2.T, CO.R, and NOx.T are negatively skewed while CO.T is positively skewed.
- In simple exponential smoothing the data gets smoothed and in the double exponential smoothing the data and the trend both get smoothed. Here more weightage is given to the immediate previous values of the data. In our case we get similar result for both simple exponential and double exponential smoothing. Also there was no seasonality in our data and hence we did not proceed for triple exponential smoothing.
- The RMSE values of the Linear regression model are the least for all the components. Hence models fitted using Linear Regression is the best compared to ARIMA and Exponential Smoothing.

Software Used: R, Python, Excel.

Limitations:

- Due to time constraint we were restricted to use the analysis we were aware about and were not able to explore more types of analysis suitable for the data.
- Due to the current restriction we were only able to interact virtually leading to some drawbacks which could have overcomed with one to one interaction with each other.

References:

https://archive.ics.uci.edu/ml/datasets/Air+quality

https://ijisrt.com/assets/upload/files/IJISRT20AUG683.pdf

https://medium.com/analytics-vidhya/how-to-guide-on-exploratory-data-analysis-for-time-series-data-34250ff1d04f

https://en.wikipedia.org/wiki/Exponential_smoothing

https://en.wikipedia.org/wiki/Exponential_smoothing#Double_exponential_smoothing

https://en.wikipedia.org/wiki/Exponential_smoothing#Basic_(simple)_exponential_smoothing_(Holt_linear)

https://en.wikipedia.org/wiki/Linear_regression

https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average

https://en.wikipedia.org/wiki/Moving_average

THANK YOU!!!