Name: Waqar Hassan
Email: Waqar_comsat@yahoo.com

Question 1

For descriptive analysis, first the dataset is preprocessed and missing values in rend_r and massa_r (real income and mass income) columns are filled using the mean value of the corresponding column. After that, dataset is group according to sector and region wise to see the energy consumption trends in the 5 regions and three sectors i.e., Industrial, commercial and residential.

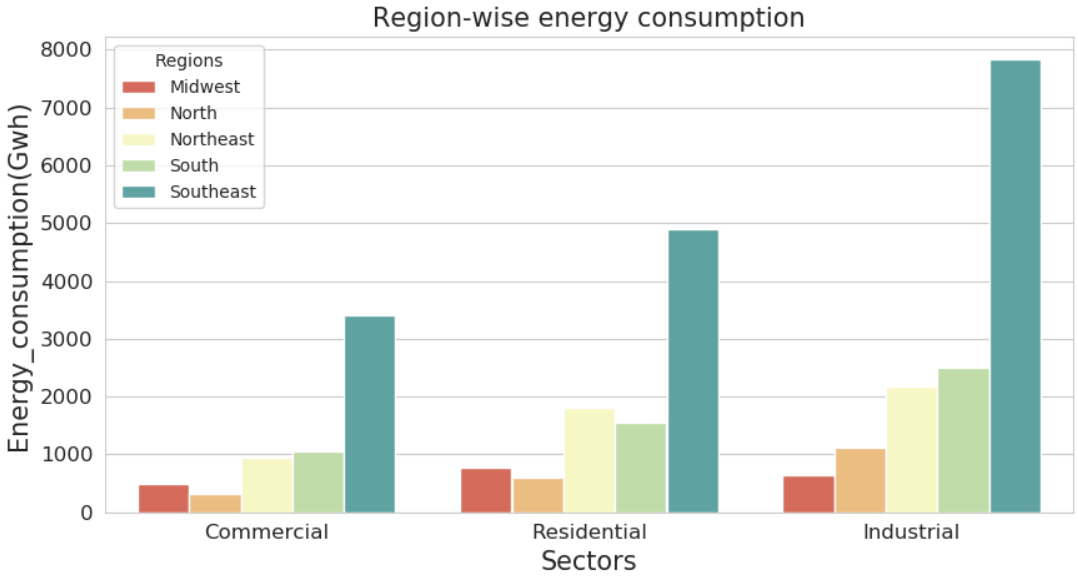

Figure 1. Regional energy consumption (Gwh) in commercial, residential and industrial sectors in Brazil.

Figure 1, gives an insight to the amount of energy consumed by the commercial, industrial, and residential sectors in the five regions of Brazil i.e., Midwest, North, Northeast, South and Southeast, respectively.

It can be clearly seen that the highest amount of energy consumed by all three sectors is in the Southeast region, with industrial sector consuming the highest amount of energy followed by residential and commercial sector. It can be inferred from here that this region is industrially well developed with most of the population concentrated in this region contributing to more businesses and therefore the energy consumption is highest. Similarly, South region follows the same trend as in Southeast region and becomes the second most energy consumption region in industrial and commercial sectors. However, in residential sector, Northeast region consumes more energy than South region.

Furthermore, Midwest region is the least energy consumption region in industrial sector. However, in residential and commercial sectors, Midwest consumes more energy than north region.

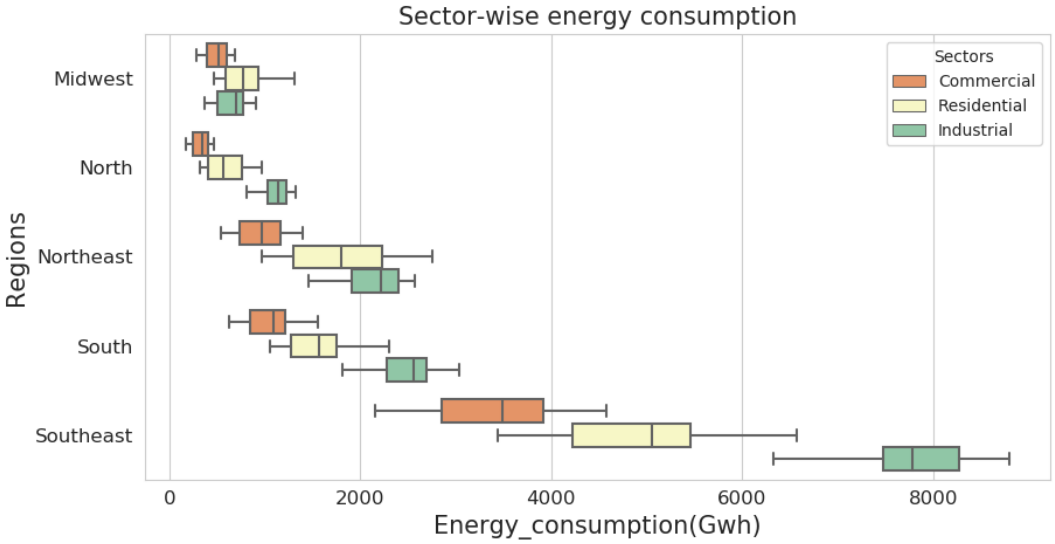

Figure 2. Sectoral energy consumption (Gwh) in five geographic regions of Brazil.

The above boxplot in Figure 2 shows the descriptive statistical information for energy consumption in commercial, residential and industrial sectors of Brazil. The plot indicates that all the geographic regions of Brazil except Southeastern region consumes energy in the range of (0 – 3000 Gwh) for all the three sectors. Whereas, the energy consumption of Southeastern region ranges from 3000-9000 (Gwh) respectively, indicating highest development in the country. Similar trend is shown in South region but with lower ranges (500-3000 Gwh).

For all regions, Industrial sector consumes the highest energy followed by residential and commercial sector. As commercial sector consumes least energy but energy consumption of this sector for Southeast region is more than all the sectors of other regions.

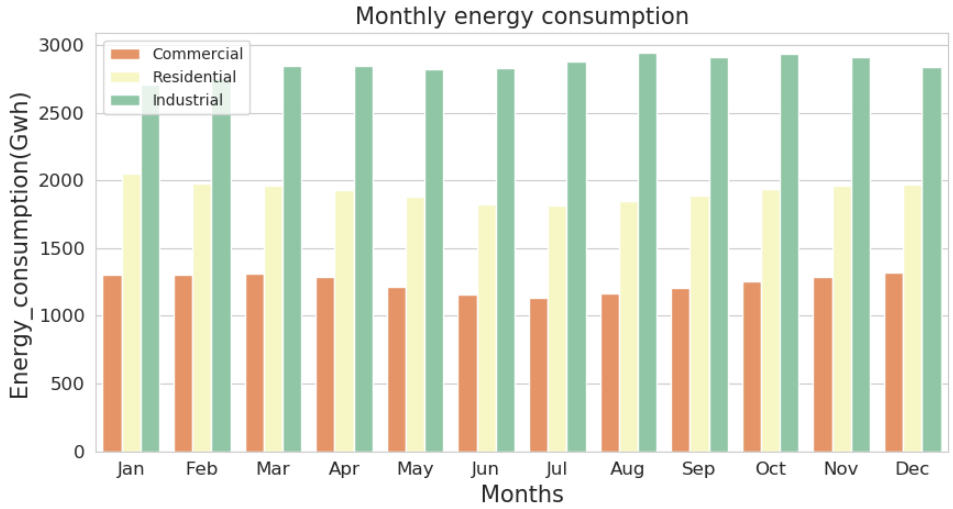

Figure 3. Monthly consumption of energy (Gwh) in commercial, residential, and industrial sectors in Brazil.

Figure 3, represents the average monthly energy consumption (Gwh) in commercial, residential, and industrial sectors in Brazil. The plot indicates that the energy consumption in residential and commercial sector somehow varies with the months. In these sectors, there is less energy consumption in the mid of the year while in start and end of the year it has maximum energy consumption. This trend can relate with the temperature of Brazil too. As the mid of the year is mostly winter season, so there is less consumption of energy in winter season because of the little cold temperature and days are short too, while, in summer days are long and temperature is hot too, so more consumption of energy is expected.. However, industrial sector consumes has a similar trend all over the year.

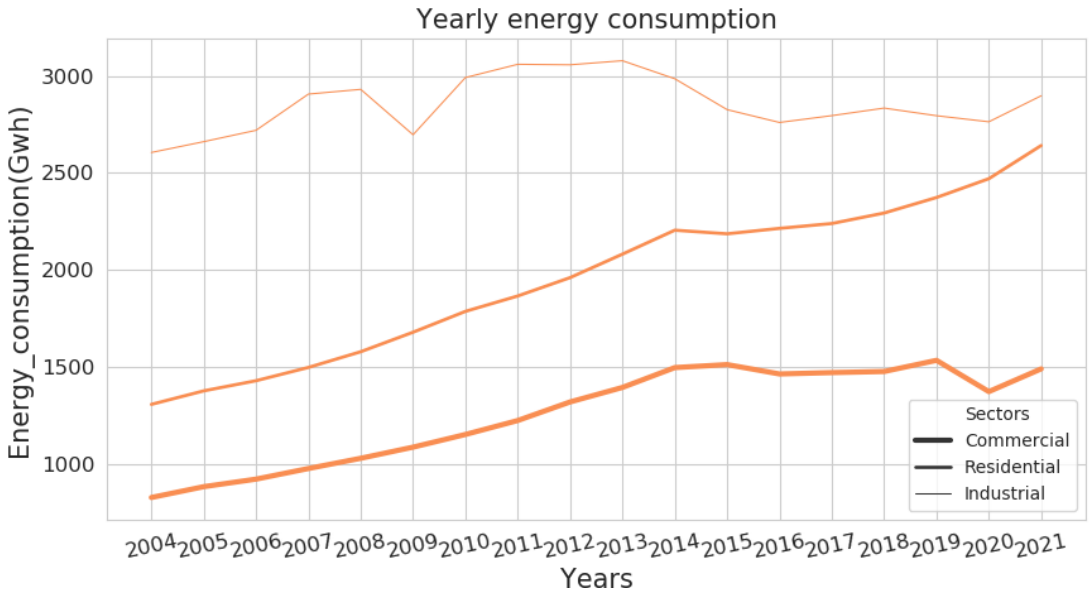

Figure 4. Year wise consumption of energy (Gwh) in commercial, residential, and industrial sectors of Brazil.

Figure 4, represents year wise trend in the energy consumption in the three sectors of Brazil from 2004-2020. It can be clearly seen that industrial sector always consumed highest amount of energy, followed by residential and commercial sectors.

In industrial sector, energy consumption gradually increase from 2004 to 2008, however, it significantly decreases in 2009 may be because of economic resection in this year or presence of outliers in the data, and then again significantly increases in 2010. After 2011, energy consumption decline gradually till 2020.

The trend in residential sector shows a gradual increase in energy consumption from 2004 - 2021. The commercial sector faced increase in energy consumption from 2004- 2014, remained stable from 2014-2020 with a little decline in 2020.

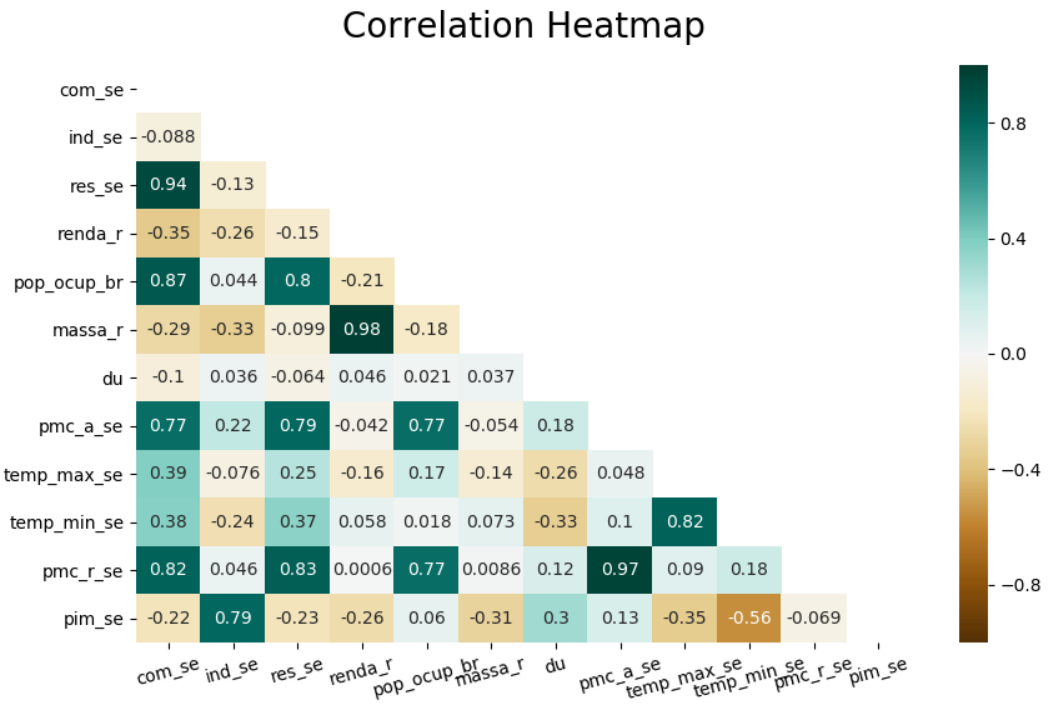

Figure 5. Correlation matrix explaining correlation between features of Southeast region of Brazil.

The Figure 5, shows the correlation matrix between energy consumption of commercial, industrial, and residential sectors with independent variables in Southeastern region of Brazil, respectively. It can be observed that energy consumption in the commercial sector is positively correlated with occupied population (r = 0.87), restricted trade (r = 0.82) and expanded monthly trade survey (r = 0.77) features. Similarly, residential sector has the same correlation trend with the features.

The energy consumption in industrial sector of southeast is positively correlate with the industrial production dependent feature with a coefficient of r= 0.79. While, other features have a very low correlation coefficient. Moreover, Table 1, represents the descriptive statistics that summarize the central tendency, dispersion and shape of the features.

Table 1. Central tendencies of features related to Southeast region.

Descriptive Measures	com_se	ind_se	res_se	renda_r	pop_ocup_br	massa_r	du	pmc_a_se	temp_max_se	temp_min_se	pmc_r_se	pim_se
count	206	206	206	228	228	228	228	228	228	228	228	228
mean	3399.49	7828.91	4903.59	1787.27	87294.79	142963.40	20.93	85.32	27.54	19.04	84.21	90.90
std	622.48	519.73	765.95	188.05	4235.77	21399.06	1.31	17.27	2.00	2.04	18.97	10.07
min	2159.48	6331.12	3433.44	1400.51	75777.70	101690.06	18.00	46.53	22.31	14.89	45.08	60.32
25%	2848.94	7477.52	4224.06	1708.65	84519.19	135533.16	20.00	74.64	25.74	17.27	70.20	84.20
50%	3488.67	7783.61	5048.88	1787.27	87671.69	142963.40	21.00	89.45	27.57	19.43	89.67	91.07
75%	3914.28	8272.47	5463.76	1787.27	90707.75	143367.74	22.00	96.39	29.06	20.87	95.22	97.19
max	4571.72	8795.55	6571.31	2335.00	94552.00	201815.00	23.00	126.26	32.02	23.13	137.33	112.05

Question 2 and 3

Note: This section consists of two parts. First, univariate time series analysis and second, multiple feature models are used for prediction.

Single Feature Time Series Analysis

Time series analysis is different as compared with the regular modeling task. The analysis is divided into the following parts:

Visualization:

In the first step, Industrial energy consumption of Southeast region is visualized using python to see the consumption pattern like trend and seasonality in the data.

The date column in provided dataset is set to index and plot the energy consumption columns to see the trend on different years, as shown below:

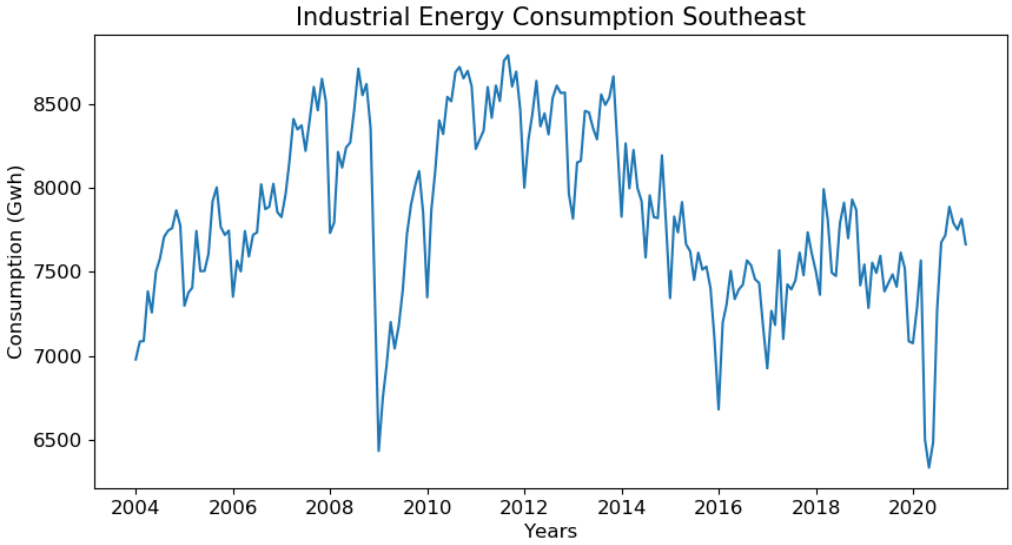

Figure 6. Energy consumption of industrial sector in southeast region.

It is seen from Figure 6 that in the start of the year consumption is less and it grows gradually till the end of the year. Moreover, energy consumption increases significantly from 2004 to 2008, but in 2009 there is a significant fall in consumption. May be this fall is because of the presence of outliers in the dataset or something serious may happened in 2009 in industrial sector of the region. Furthermore, the energy consumption is gradually decreases after 2011.

Time series consists of 3 systematic components such as, level, trend and seasonality and the other nonsystematic component is random noise. The level in series considers the average values, trend represents increasing or decreasing behavior of series, seasonality shows repeating short term pattern or cycle in the series and noise depicts randomness in the series.

From Figure 7, it is clearly seen that trend is increasing till 2008 and then decreasing till the end of 2020. Moreover, seasonality is also significant where consumption shoots maximum till the end of every year.

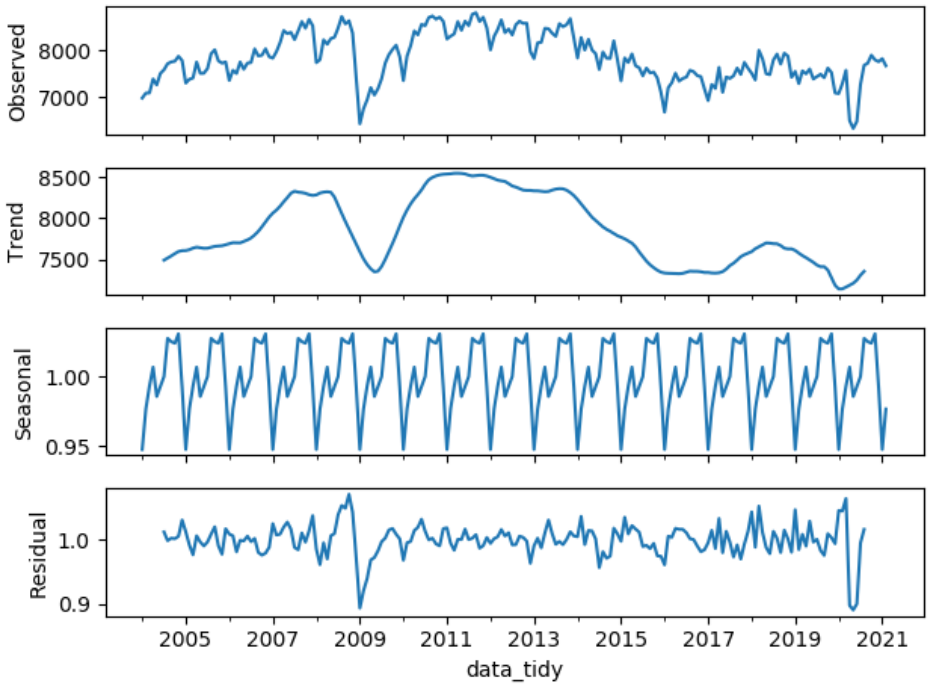

Figure 7. Seasonal decompose of consumption data.

Stationarity:

For time series analysis and forecasting, it is necessary that data needs to be stationary i.e., it has a constant mean, variance and autocorrelation over the time. Therefore, the statistical test i.e., Augmented Dickey-Fuller (ADF), is applied to see whether the energy consumption series data is stationary or not, by determining the unit root presence in the series. The Null hypothesis be the series has unit root and alternative hypothesis be, it has no root. If Null hypothesis is not rejected, the series is non-stationary and if mean and standard deviation are constant the series is stationary.

Table 2. Dickey fuller test on original data

Test statistics	-2.50583
p-value	0.114061
No. of lags used	12
Number of observations used	193
Critical value (1%)	-3.46469
Critical value (5%)	-2.87663
Critical value (10%)	-2.57482

As p-value in Table 2 is greater than 0.05, means it does not rejects the null hypothesis and test statistics is greater than critical values thus series is not stationary. Moreover, it can be clearly seen in the Figure 8, that mean and standard deviation are also not constant. Rolling mean is calculated by taking input for past 12 months.

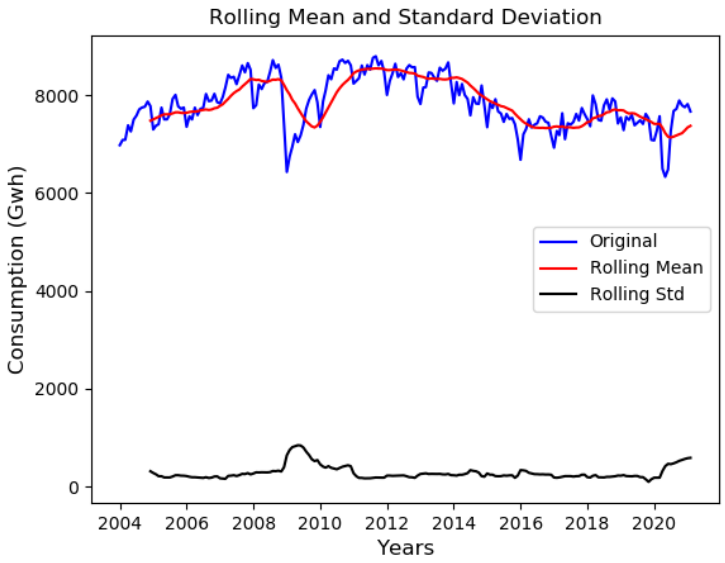

Figure 8. Rolling mean and standard deviation.

To make the series stationary, trends and seasonality in it needs to be eliminated. Therefore, in the first step, log is applied for smoothing and decreasing the magnitude of the values in series as shown in Figure 9.

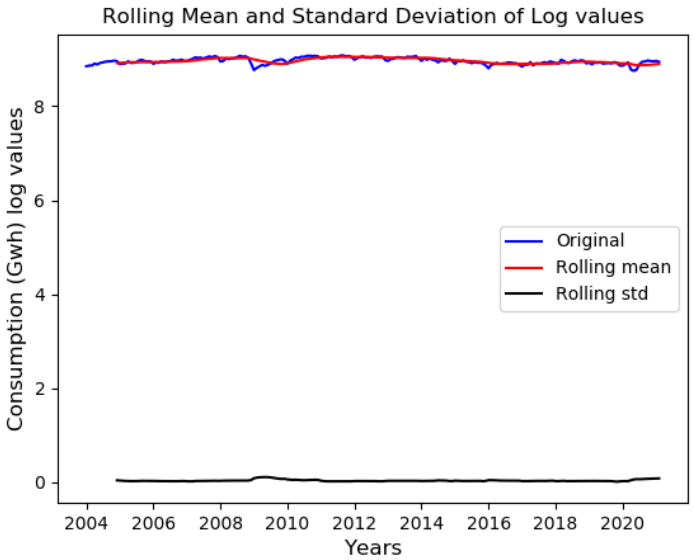

Figure 9. Rolling mean and standard deviation of Log values.

After that, the difference of the new log value series and the rolling average mean is calculated to eliminate the trends and make it stationary and also ADF test is applied to see the series become stationary or not.

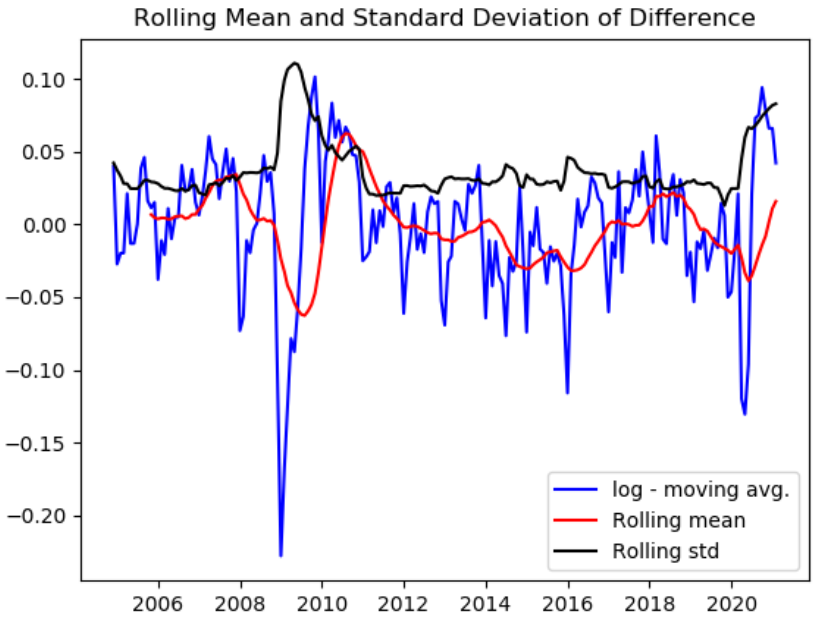

Figure 10. Rolling mean and standard deviation of difference between log and moving average.

Table 3: Dickey fuller test after logging

Test statistics	-4.30727
p-value	0.000432
No. of lags used	12
Number of observations used	182
Critical value (1%)	-3.4668
Critical value (5%)	-2.87756
Critical value (10%)	-2.57531

From Table 3 and Figure 10, it is observed that p-value is less than 0.05 and data attained stationarity. However, from Figure 7, it is observed that there is high seasonality in the data. Therefore, differencing task is applied to eliminate the seasonality.

Differencing: Differencing helps to stabilize the mean by removing changes in the level of a time series. It is applied by subtracting the previous observation from the current observation. After differencing, ADF test is again applied to see the stationarity of the data.

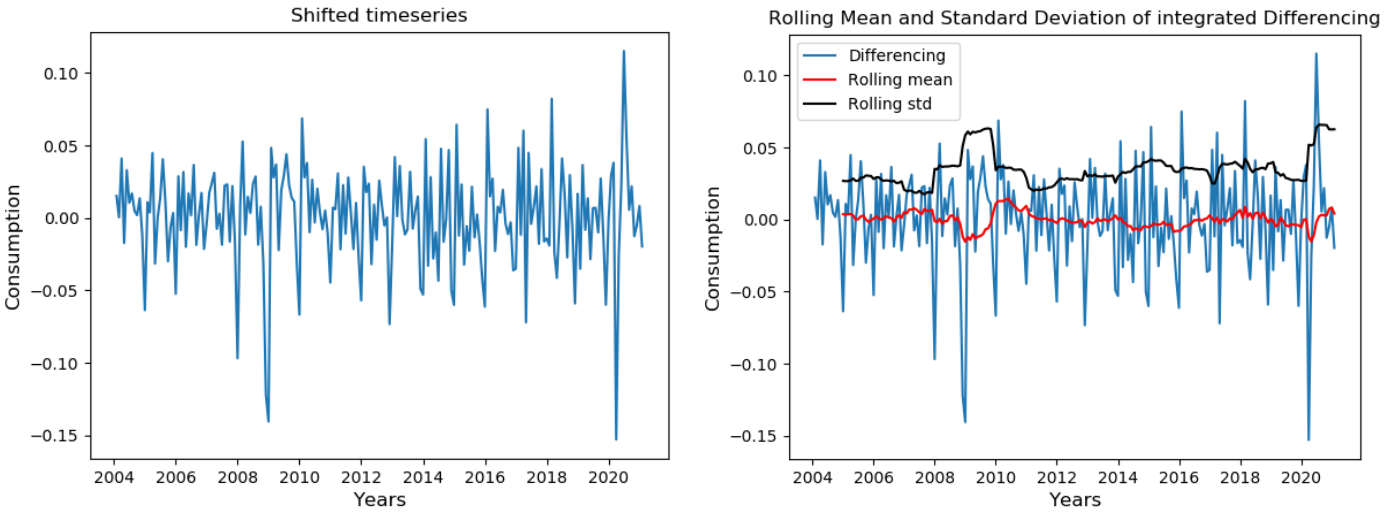

Figure 11. Shifted time series and differencing with rolling mean and standard deviation.

Table 4. Dickey fuller test after differencing

Test statistics	-3.6644
p-value	0.004643
No. of lags used	11
Number of observations used	193
Critical value (1%)	-3.46469
Critical value (5%)	-2.87663
Critical value (10%)	-2.57482

It is clearly observed from the Figure 11 that mean is stable than before and ADF test also tells that series is stationary, as p-value is less 0.05. In next step, decomposition is perform that provides the structural way of thinking about a time series forecasting problem. For decomposition, series is separated in the components once again as done before in Figure 7, and also performs ADF test once again.

Table 5. Dickey fuller test after decomposition

Test statistics	-5.32052
p-value	0.00000496
No. of lags used	11
Number of observations used	193
Critical value (1%)	-3.46469
Critical value (5%)	-2.87663
Critical value (10%)	-2.57482

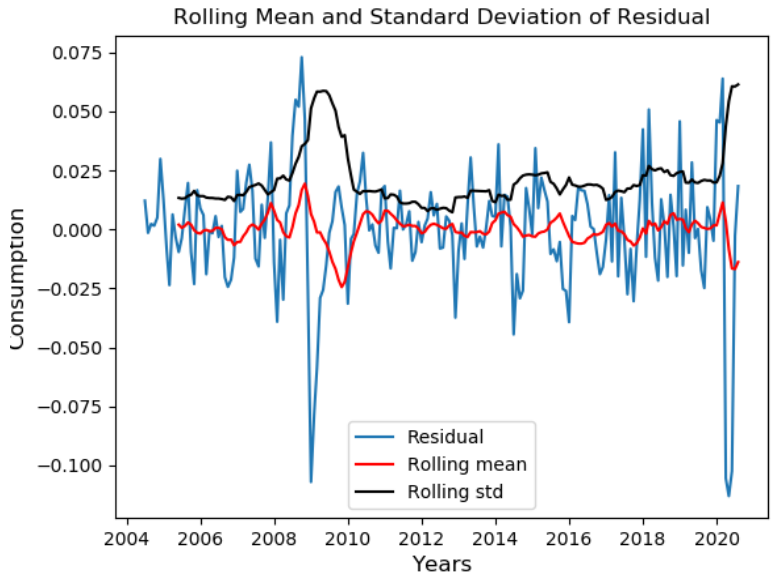

Figure 12. Rolling mean and standard deviation of residual.

After the decomposition, it is clearly seen that series is stationary and p-value is much less than before i.e., 0.000004.96 which represents the stationarity.

In next step, the optimal parameters for the model are estimated using the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) plots.

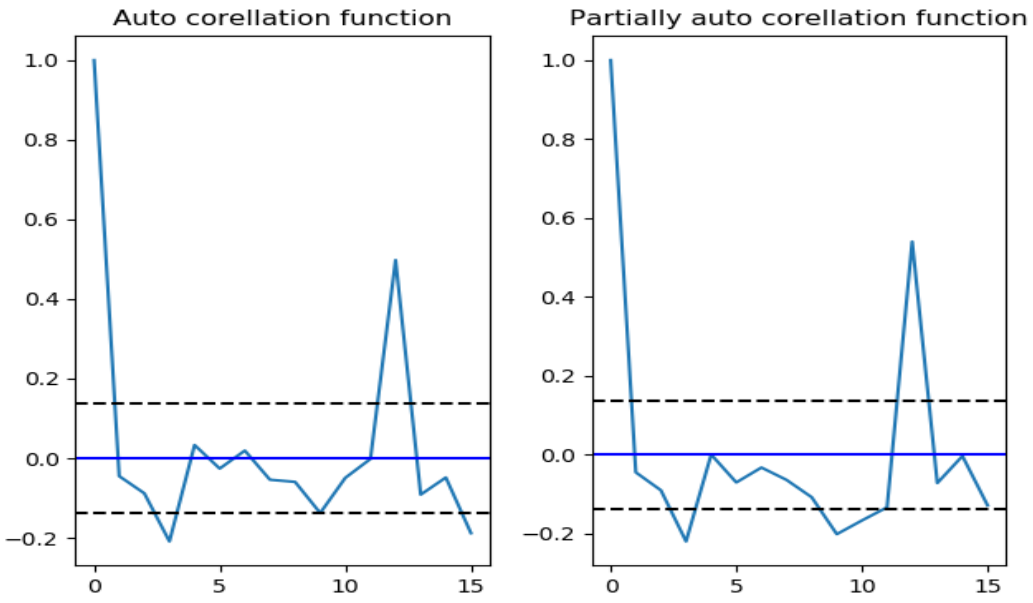

Figure 13. ACF and PACF.

Model Fitting and prediction

In order to fit the ARIMA (p,d,q) model, it requires the p, q values which will be find from the PACF and ACF plots in Figure 13, where the graph cuts the origin for the first time. Thus, from the graph, at point 1 graph cuts or drop to zero first time so p and q is 1. Thus, ARIMA (1,1,1) model is fitted on the data to predict the future series as shown in Figure 8. The model is well fitted on the history values, as shown in Figure 14. However, prediction represents the downward trend means energy consumption will gradually slow in the next 24 months, but just shown by constant line instead of showing peaks as history values. To evaluate the model performance, root mean square error (RMSE), mean square error (MSE) and r2 score are used, i.e., RMSE =0.034, MSE = 0.001and r2 = 0.005. The model is fitted well on the history data.

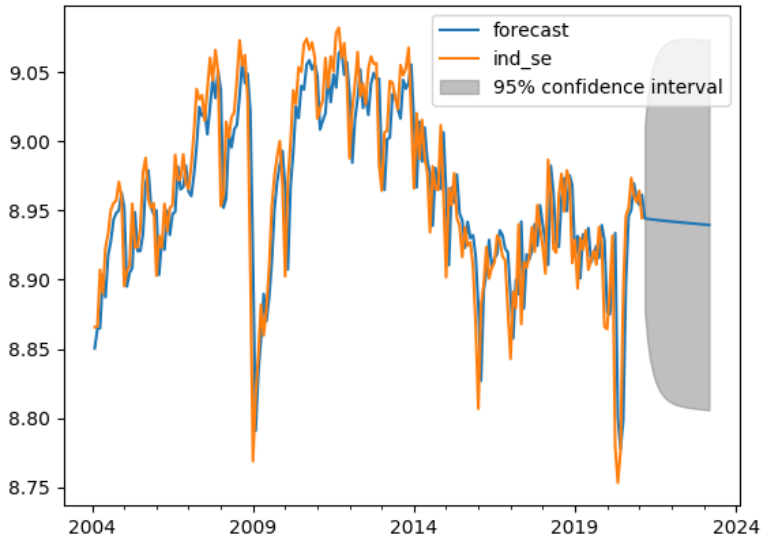

Figure 14. Forecasting next 24 months.

Using Multiple Feature Models

In the next part, prediction are estimated using the multiple independent features. This part uses 5 models i.e., Linear regression, Support vector regression, Random forest, Decision trees and Bayesian Ridge regression model. In the experimental setup, first the data is split into training and test set. Then, 5 fold cross validation is applied to minimize the overfitting and biasedness of the models. For each fold all the models are trained or fitted together on the same training sample and the model is evaluated on the same validation set for each fold.

The model is evaluated using the root mean square error (RMSE), mean square error (MSE) and r2 score. R2 score are good to understand the explanatory power of the independent variable or features. Mean square error is a measure that tells how close a fitted line is to data

points, which takes the difference between the points and takes the square of the difference so negative values do not cancel the positive values. RMSE is the square root of MSE. The model is run manually for each dependent features and evaluation measures are saved for each model, as shown in Table 6. After comparing the model errors, optimal model is selected for future prediction.

Table 6. Different model error for each feature.

Model	Error Measure	du	Massa_r	pim_se	pmc_a_se	pmc_r_se	renda_r	temp_max_se	temp_min_se
Linear Regression	**RMSE**	596.18	**555.34**	**351.78**	**587.35**	624.76	798.34	599.33	581.09
	MSE	369861.55	321180.19	125109.30	364984.32	406824.20	800529.66	373400.45	350645.68
	R2	-2.40	**-1.89**	**-0.05**	-2.48	-3.05	-6.04	-2.37	-2.20
Support Vector Regression	**RMSE**	583.27	584.43	577.31	585.05	583.69	583.47	583.26	583.71
	MSE	356156.76	357541.18	348429.27	358659.70	356894.79	356392.18	356122.27	356749.86
	R2	-2.37	-2.37	-2.25	**-2.42**	**-2.41**	-2.37	-2.35	-2.36
Random Forest	**RMSE**	597.18	574.03	501.15	688.64	647.32	608.23	600.58	600.06
	MSE	376171.01	342060.58	273284.60	494307.26	442036.96	384118.11	385075.51	388908.33
	R2	**-2.36**	-2.02	-0.16	-3.88	-3.76	-2.71	**-2.19**	-2.19
Decision Tree	**RMSE**	**596.50**	564.41	510.38	699.21	638.31	590.27	610.48	599.11
	MSE	376629.75	333773.21	283095.56	514473.59	434685.75	361080.15	413690.60	387076.98
	R2	**-2.36**	-1.92	-0.26	-3.99	-3.71	**-2.36**	-2.46	**-2.14**
Bayesian Regression	**RMSE**	712.56	1278.26	767.98	1912.64	2099.39	989.91	760.79	984.60
	MSE	602903.60	2069849.38	725710.56	4720245.89	5720333.54	1140992.04	694872.75	1197119.15
	R2	-4.48	-16.97	-2.40	-38.67	-52.14	-8.78	-6.44	-11.70

The Table 6, depicts the three error measures (RMSE, MSE and R2) results for each model across different independent features of Southeast regions. According to error measures, the pim_se (industrial production) feature represent least error for all models when compared with features. For industrial production feature, linear regression has the least errors compared with other models. For pmc_a_se and pmc_r_se features, all the models give very high errors in comparison with errors for using other features, and for these features Bayesian regression gives extremely high errors.

Table 7. Average of R2 score across features for each model

Linear Regression	-2.56
Support Vector Regression	**-2.36**
Random Forest	-2.41
Decision Tree	-2.40
Bayesian Regression	-17.69

Table 7 represents the mean of r2 score across all features for each model. According to the Table 7, support vector regression has the least mean r2 scores compared with other models. However, Decision Tress and random forest ranks have almost same performance and ranks

2 and 3 position, followed by linear regression model. Bayesian regression has poor performance when compared with other models.

From Table 7, support vector regression model is used to predict the future values. To predict the future values, model is again trained on all training dataset to predict test set.

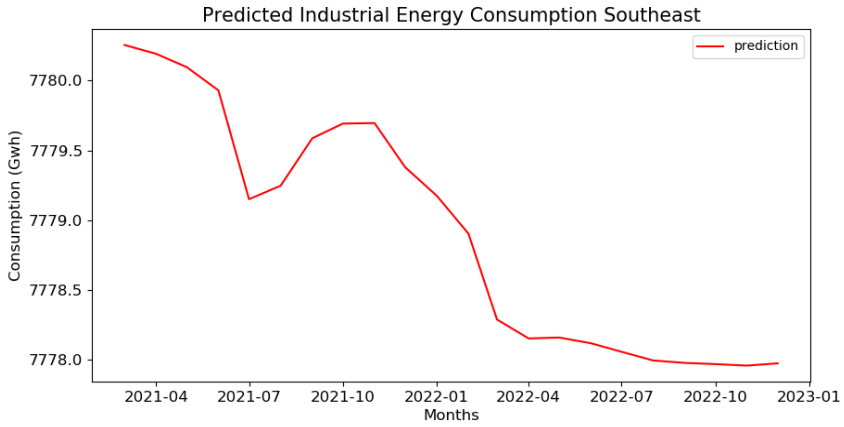

Figure 15. Prediction future values for 22 months.

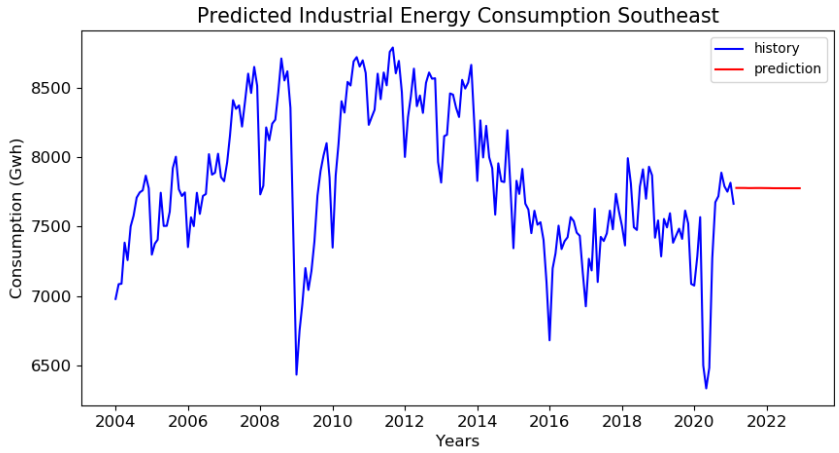

Figure 16. History and Prediction future values for 22 months.

Figure 15, represents the predicted trend of energy consumption for next 22 months. The energy consumption for next 2 years is gradually decline following the history trend but with very less variation, as shown in time series analysis and also in Figure 16.

Question 4

The provided data gives insightful information about the energy consumption in 5 different region of Brazil divided among 3 sectors (Industrial, Commercial and Residential). As Southeast and South regions are the most developed areas in Brazil, thus, the results shows the expectations that this region consumes more energy compared with other regions. Moreover, industry sector consumes significant energy in all the regions. After that, residential is the second most energy consumption sector. In industrial sector, North has more energy consumption than Midwest region and has less energy consumption in commercial and residential sector. This shows that North has less population than Midwest region but North is more developed in Industrial sector than Midwest region.

The plots depict that energy consumption for residential and industrial sector gradually increase with the years, however, the consumption trend for industrial sector is not linear. It shows increasing trend until 2008 and then after 2011 it gradually decreases. This decline is maybe because of some economic recession or outliers in the provided data.

The results of ARIMA model show the data fit properly because it give small errors. May be the reason is data is processed well and non-stationarity is eliminated from the series. The errors for multiple feature models are very high because of not applying preprocessing to the data. However, both time series and multiple feature models predict the gradually decline trend for the next 2 years.

Note:

The codes are implemented using python Language in VSCode. The codes are presented well but they are not in optimized shape. For running the codes please see the **Readme.txt** file.