Multiple Regression and Time Series Model Analysis

Anand Basant Haritwal
MSC in Data analytics
National College of Ireland
Dublin-1
Student Id- x19174489@student@ncirl.ie

Abstract

The report is divided into two parts.

Part A – Multiple regression:

The aim is to understand the relationship between the total deaths occurred in the year 2008 to different causes like cancer, chronic and all. The death count due to each factor is compared to the total death count and multiple regression is used to predict the total death count using individual factors.

Part B- Time series analysis:

The aim is to understand the changes in production of Natural gas, for UK, over the years. Price of natural gas has a wide impact over the revenue of the country. The plan is to analyse the production over the years, visualize any trend in data and try to fit a suitable predictive model on the time series.

Part A

Data set- Multiple regression:

The data for the age standard mortatilty rate was obtained from WHO. This data set contains the death count off 193 countries for the year 2008. It contains total 8 variables. The variables 1-8 are independent variables and the 9^{th} variable is the dependent variable.

Serial number	Variable	Description
1	Cancer value	The no. of people died due to cancer
2	Cardiovascular value	The no. of people died due to heart failure
3	Chronic respiratory Value	The no. of people died due to Lung disease
4	Communicable Value	The no. of people died due to common cold, HIV, transferable disease
5	inzuries Value	The no. of people died due to accidents, chronic or illness
6	Non-communicable Value	The no. of people died due to non transferrable disease like parkinson, diabetes etc.
7	due to border issue Value	The no. of people died due to cross border transmission
8	all Value	The total no. of people died

Reasearch question

1. Analyse the variables and suggest suitable independent variables for the model.
2. How well do the Deaths due to various factors predict the total death count? How much variance in total death count can be explained by the individual factors.
3. Which is the best predictor of total death count?

Data set analysis

- Data screening and cleaning

The data types and errors were checked before starting the analysing. The table below shows that there is no missing value present and data type for each variable is accurate.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
cancer Value	193	100.0%	0	0.0%	193	100.0%
cardiovascular Value	193	100.0%	0	0.0%	193	100.0%
chronic respiratory Value	193	100.0%	0	0.0%	193	100.0%
communicable Value	193	100.0%	0	0.0%	193	100.0%
inzuries Value	193	100.0%	0	0.0%	193	100.0%
non communicableValue	193	100.0%	0	0.0%	193	100.0%
due to border issueValue	193	100.0%	0	0.0%	193	100.0%
all Value	193	100.0%	0	0.0%	193	100.0%

1. Multiple regression
   It is used to explore the predictive ability of a set of independent variables on one continuous dependent measure. It also allows user to determine the statistical significance of the outcome, with respect to the model and the individual independent variables.

   Assumptions : The major assumptions for multiple regression are as follow

a. Sample size:
   The problem with small sample is repetability. The results obtain cannot be repeated with other samples.
   Formula : N> 50+8m (where m= no. of Independent variable)
   For this project the sample size is 193 which is way above then 106(50 + 8 * 7) as we have 7 independent variable. Hence this assumption is satisfied for our sample.
b. Multicollinearity and singularity:
   This gives the relationship between the independent variables. For a good model the correlation between the independent variables should be less then 0.7. As observed in the table below all the independent variable has high correlation with the dependent variable denoted by "Pink box" except the **Cancer value. The correlation table also give high relationship between some independent variables denoted by "brown box". This is taken into consideration while designing the multiple regression model and hence the variables with high correlation are not used together in the model.**

Correlations[b]

		cancer Value	cardiovascular Value	chronic respiratory Value	communicable Value	inzuries Value	non communicableValue	due to border issueValue	all Value
cancer Value	Pearson Correlation	1	.155*	.015	-.061	.081	.260**	-.115	.103
	Sig. (2-tailed)		.031	.833	.403	.263	.000	.112	.153
cardiovascular Value	Pearson Correlation	.155*	1	.680**	.447**	.460**	.903**	.450**	.73**
	Sig. (2-tailed)	.031		.000	.000	.000	.000	.000	.000
chronic respiratory Value	Pearson Correlation	.015	.680**	1	.810**	.547**	.737**	.786**	.85**
	Sig. (2-tailed)	.833	.000		.000	.000	.000	.000	.000
communicable Value	Pearson Correlation	-.061	.447**	.810**	1	.548**	.586**	.899**	.90**
	Sig. (2-tailed)	.403	.000	.000		.000	.000	.000	.000
inzuries Value	Pearson Correlation	.081	.460**	.547**	.548**	1	.546**	.478**	.65**
	Sig. (2-tailed)	.263	.000	.000	.000		.000	.000	.000
non communicableValue	Pearson Correlation	.260**	.903**	.737**	.586**	.546**	1	.599**	.79**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
due to border issueValue	Pearson Correlation	-.115	.450**	.786**	.899**	.478**	.599**	1	.77**
	Sig. (2-tailed)	.112	.000	.000	.000	.000	.000		.000
all Value	Pearson Correlation	.103	.729**	.848**	.898**	.654**	.786**	.772**	1
	Sig. (2-tailed)	.153	.000	.000	.000	.000	.000	.000	

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

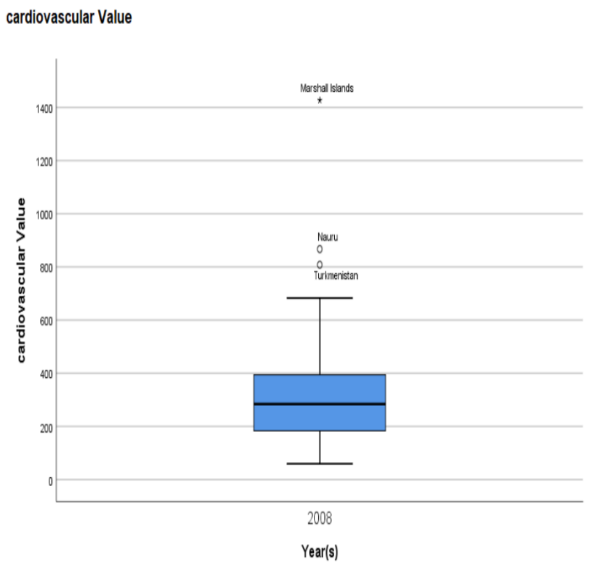
c. <u>Outliners:</u>

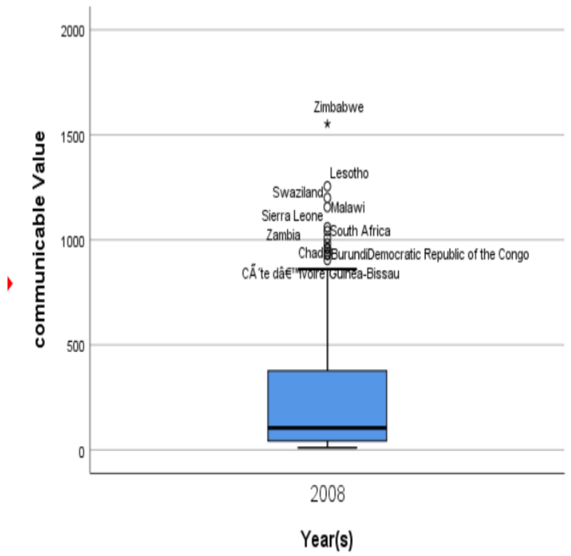
Multiple regression is very delicate when it comes to outliners. Outliners can make the regression model less accurate as it can change the slope of the regression line making the predictions inaccurate. **Box plot, Scatterplot** can be used to check outliners. All the independent variables where checked for outliners using **mean and 5% trimmed mean**. The Variables with significance difference between the two were further viewed using boxplot, normal Q-Q plot.

Descriptives

		cancer Value	cardiovascular Value	chronic respiratory Value	communicable Value	inzuries Value	non communicableValue	due to border issueValue	all Value
		Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Mean		144.18	303.69	44.77	269.03	72.99	628.15	34.61	939.99
95% Confidence Interval for Mean	Lower Bound	138.90	278.48	39.41	223.38	66.02	601.17	30.74	854.88
	Upper Bound	149.45	328.91	50.14	314.67	79.96	655.12	38.48	1025.11
5% Trimmed Mean		143.39	291.27	41.59	235.65	67.85	622.36	33.59	881.84
Median		140.00	284.00	29.00	105.00	58.00	637.00	26.00	774.00
Variance		1380.885	31533.182	1428.364	103344.223	2410.448	36100.385	743.843	359379.807
Std. Deviation		37.160	177.576	37.794	321.472	49.096	190.001	27.273	599.483
Minimum		59	59	2	11	14	273	3	220
Maximum		284	1427	195	1552	347	1289	90	3147
Range		225	1368	193	1541	333	1016	87	2927
Interquartile Range		44	213	54	341	55	268	53	708
Skewness		.481	1.679	1.199	1.501	2.013	.323	.539	1.420
Kurtosis		.999	7.662	.939	1.478	6.157	-.085	-1.204	2.069

As observed from the descriptive table there is a huge difference between mean and 5% trimmed mean for cardiovascular and communicable values indicating presence of outliners. The boxplot of these indicated Marshall Islands as an outliner for Cardiovascular value. Zimbabwe is the significant outliner for Communicable value,but as our data is taken from WHO we assume there is no mistake in the values and we will still continue with the data without deleting them. If we delete the outliner the accuracy can further be increased for the model.





d. <u>Normality, Linearity, Homoscedasticity, Independence of residuals:</u>

i) Skewness and Kurtosis: The value of skweness and kurtosis should be close to zero for a normal distribution. If the graph is **Skewed take median** rather the mean for statistics. If the graph has **high kurtosis it can result to underestimate of variance**. The Descriptive table above shows that **Cardiovascular with 7.662 and Inzuries with 6.157 have high kurtosis value.** This may be due to presence of outliner as seen in boxplot.

ii) Test of normality: The kolmogorov-smirnov significance value was used to test for normality. As seen in the table below **Cancer and non-communicable have significance value higher then standard 0.05**. The histogram of all variables were viewed which represented nearly normal distribution. The distribution further can be made normal by using sample of means or taking log of the values. We will keep the data same to find the adequate model from the real values available. Outliners were responsible for this.

Tests of Normality

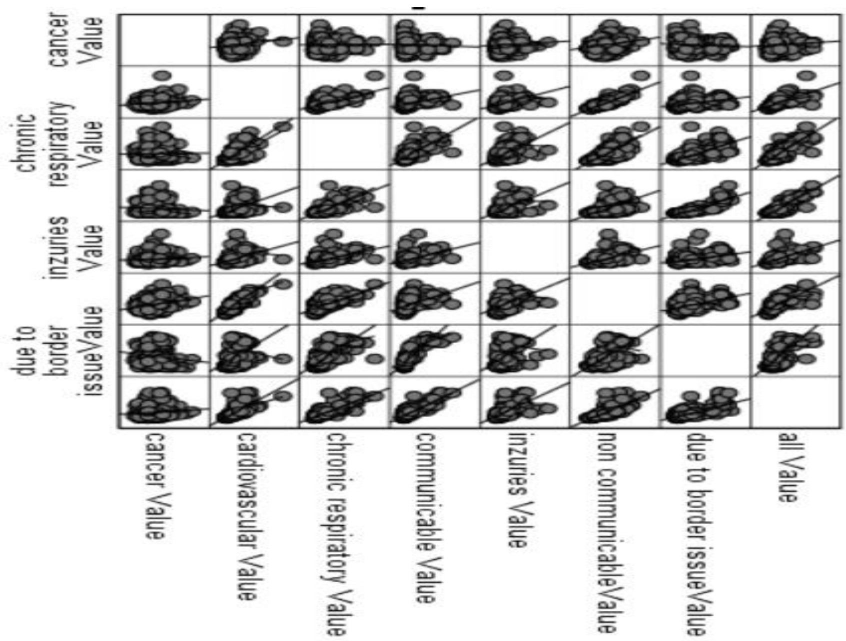
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
cancer Value	.072	193	.016	.979	193	.006
cardiovascular Value	.084	193	.002	.894	193	.000
chronic respiratory Value	.181	193	.000	.860	193	.000
communicable Value	.226	193	.000	.768	193	.000
inzuries Value	.141	193	.000	.825	193	.000
non communicableValue	.073	193	.015	.977	193	.003
due to border issueValue	.138	193	.000	.874	193	.000
all Value	.142	193	.000	.872	193	.000

a. Lilliefors Significance Correction

iii) Normal Q-Q plot: the plot of observed value vs the expected value. All the variables displayed a straight line from bottom right to top left throughout the plot.

iv) Detrended Normal Q-Q plot: the plot of actual deviation from the straight line. No clustering was observed and maximum values were near ground zero line for all the variables.

v) Linearity: **The scatterplot matrix shown below shows that the variables are linear. The straight line drawn in each scatter plot of variables prove that they are linear and thus the linearity assumption is follwed.**



e. <u>Direction, strength of measurement and coefficient of determination:</u>

variable	Direction of correlation	Strength of correlation	Coefficient of determination	comment
Cancer	Positive	Small	0.0106	It helps to explains nearly **01.06%** of variance in total death
Cardiovascular	Positive	Large	0.5314	It helps to explains nearly **53.14%** of variance in total death
Chronic respiratory	Positive	Large	0.7191	It helps to explains nearly **71.91%** of variance in total death
Communicable	Positive	Large	0.8064	It helps to explains nearly **80.64%** of variance in total death
inzuries	Positive	Large	0.4277	It helps to explains nearly **42.77%** of variance in total death
Non communicable	Positive	Large	0.6178	It helps to explains nearly **61.78%** of variance in total death
Border issues	Positive	Large	0.5959	It helps to explains nearly **59.59%** of variance in total death

Statistics for Data Analytics Individual Project A

Estimation of multiple regression model in SPSS:

- **Model Zero:**
  This model contains the prediction of total deaths by taking all other factors as independent variables.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984[a]	.968	.967	108.599

a. Predictors: (Constant), due to border issueValue, cancer Value, cardiovascular Value, inzuries Value, chronic respiratory Value, communicable Value, non communicableValue

b. Dependent Variable: all Value

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	66819081.63	7	9545583.090	809.377	.000[b]
	Residual	2181841.366	185	11793.737		
	Total	69000922.99	192			

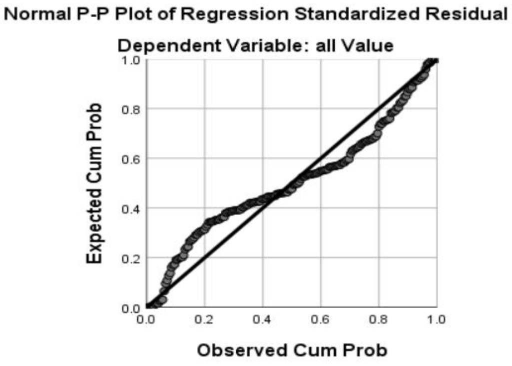
a. Dependent Variable: all Value

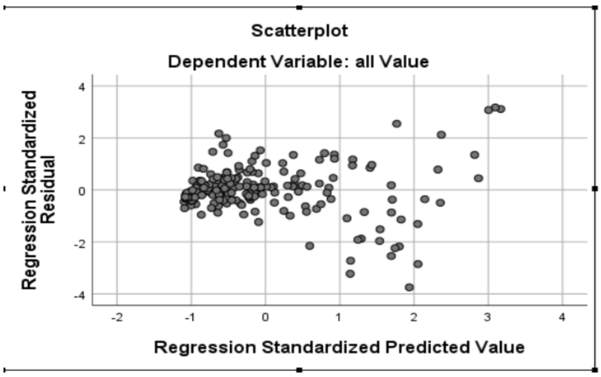
b. Predictors: (Constant), due to border issueValue, cancer Value, cardiovascular Value, inzuries Value, chronic respiratory Value, communicable Value, non communicableValue

The ANOVA significance value is less than 0.05 indicating the test of null hypothesis that multiple R in the population equals 0. The model in this example reaches statistical significance. The Adjusted R square value of 0.967 is very good that means the model can explain 96.7% of the variance in the total death count but this is due to the presence of high multicollinearity in the independence variable.

Interpretation of the model output:

- Multicollinearity- The tolerance value in **coefficient table** gives an indication of Multicollinearity. If the **tolerance value is less than 0.10 it indicates multicollinearity** is present, but as seen in the table **we have no value below 0.10.** there are some values near it indicating presence of multicollinearity in some independent variable. **V.F value above 10 also indicates presence of multicollinearity we have some cases like Cardiovascular, non-communicable, border and communicable have values near 10 indicating presence of multicollinearity.**
- Normalization- The normal PP plot of regression standardized residual shows they all lie in a straight line. Indicating no major deviation from Normality.





- Regression Standardized predicted value scatter plot- the residual scatter plot shows most of the values near zero line and no clustering of values indicating **a NON-CONSTANT residual plot.**
- Outliners – The Mahalanobis distance is used to find the outliners in the model. For 6 independent variables the critical value is 22.46 but as we can observe, we have a value of 84.14 indicating presence of outliners. Sorting according to Mahal. Distance was used in SPSS to find the outliner. Marshal Island act as a major outliner, if removed it can make the model more accurate.
- Case-wise diagnostics- This tells us about the unusual cases. The cases in which the standardized residual value is above -3 or +3. From the model output seen the cases

102,105,154,165,193 are unusual means they have large error between the predicted and the actual value. To check whether the unusual cases affects the overall result of model, check the Cooks Distance. If the value is larger than 1 that means these cases are a potential problem but for the model it is 0.698 indicating the unusual cases are not a potential problem.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	292.59	2808.99	939.99	589.929	193
Std. Predicted Value	-1.097	3.168	.000	1.000	193
Standard Error of Predicted Value	9.453	72.317	20.540	8.204	193
Adjusted Predicted Value	294.51	2777.97	939.09	587.145	193
Residual	-407.116	344.319	.000	106.601	193
Std. Residual	-3.749	3.171	.000	.982	193
Stud. Residual	-3.843	3.657	.004	1.024	193
Deleted Residual	-427.925	472.843	.910	116.476	193
Stud. Deleted Residual	-3.996	3.786	.003	1.040	193
Mahal. Distance	.460	84.144	6.964	8.812	193
Cook's Distance	.000	.698	.013	.061	193
Centered Leverage Value	.002	.438	.036	.046	193

a. Dependent Variable: all Value

Evaluating the independent variable:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	103.030	42.794		2.408	.017	18.603	187.457					
	cancer Value	1.226	.246	.076	4.982	.000	.741	1.712	.103	.344	.065	.734	1.362
	cardiovascular Value	1.516	.119	.449	12.792	.000	1.282	1.750	.729	.685	.167	.139	7.211
	chronic respiratory Value	-.095	.453	-.006	-.210	.834	-.990	.799	.848	-.015	-.003	.209	4.780
	communicable Value	1.689	.063	.906	26.938	.000	1.566	1.813	.898	.893	.352	.151	6.617
	inzuries Value	1.270	.204	.104	6.218	.000	.867	1.672	.654	.416	.081	.611	1.636
	non communicableValue	-.269	.128	-.085	-2.109	.036	-.521	-.017	.786	-.153	-.028	.105	9.558
	due to border issueValue	-5.034	.721	-.229	-6.983	.000	-6.457	-3.612	.772	-.457	-.091	.159	6.293

a. Dependent Variable: all Value

The regression equation for the model is

Summary:

All death count = 103.030 + 1.226(cancer) + 1.516(cardiovascular) – 0.095(chronic respiratory) + 1.689(communicable) + 1.270(inzuries) – 0.269(non-communicable) – 5.034(border)

- **The constant is 103.030 indicating even there is no deaths due to the given factors (all the beta value as 0) there still will be 103 total deaths. Which makes sense as there may be death due to other reasons.**
- **The largest standardized beta value is 0.906, that is by communicable, that means communicable makes the most unique contribution in explaining the total death value.**
- **The only significance value above 0.05 is for chronic respiratory value indicating it is not making a significant contribution to the prediction of the total death value. This may be due to overlap with other independent variables. Rest all the variables are making significant contribution.**
- **The total death count will increase by 0.906 standard deviation if there is an increase of 1 standard deviation in communicable count. Similarly, the total death count will decrease by 0.229 standard deviation if the border death count increases by 1 standard deviation.**
- **The square of the part correlation gives the indication of the contribution of the variable to the total R square. To say the part correlation for cardiovascular is 0.167, the square is 0.027 which means cardiovascular count uniquely explains 2.7% of variance in the total death count.**

Part B Time Series Analysis:

Introduction:

Time series is a statistical method used to understand the time data, analyze it and predict the future value by forecasting. Time data is a record of any parameter with respect to time. It can be annually, monthly, daily etc. It is simple record of a parameter with respect to any systematic time interval.

Research question:

A. To analyze the time series, identify its components and apply a suitable forecasting model on it.
B. Analyze the Forecasting model and check its accuracy.

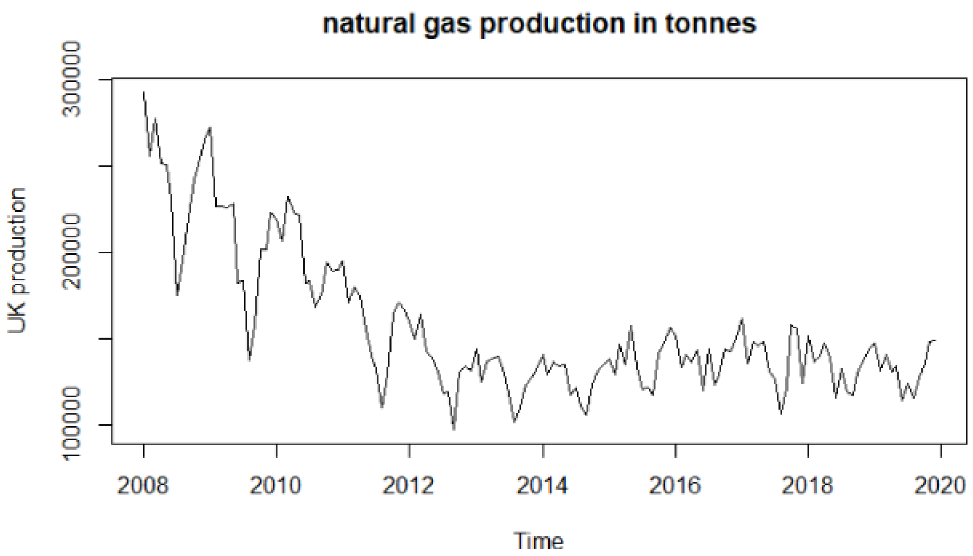
1. Data set
   The data is obtained from Eurostat. It contains the monthly production of natural gas in UK from 2008 January to 2019 December.

2. Data cleaning:
   The data was checked for any NA value using is.na () function in R. There are no NA values in the data set.
3. Implementation:
   - Creation of time series:
     To create a time series in R, ts() function was used. The data used is monthly hence frequency of 12 and start time as 2008 is used. The production is in tonnes.

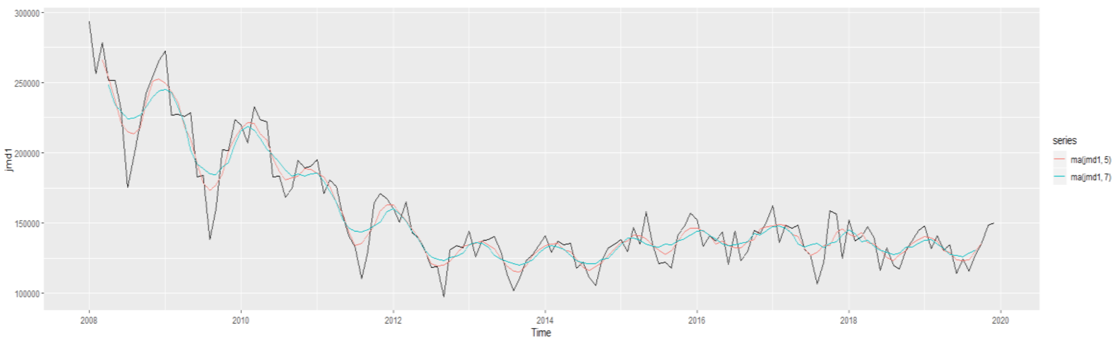
```
library(readxl)
# importing the file in R
time_series_final <- read_excel("D:/stat/time series final.xlsx")
View(time_series_final)
# creating a time series function
jmd1<-ts(time_series_final[3],start = c(2008,1),frequency = 12)
View(jmd1)
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2008	293174	256390	278125	251681	251472	231360	175393	196207	220693	243142	254175	265319
2009	272486	226636	227334	225849	228380	183105	183688	138137	159039	201990	201768	223627
2010	219796	206995	232933	223517	221882	182867	183346	168618	174967	194581	189151	190599
2011	195088	171221	180761	175640	156075	140449	133187	110509	129246	164579	171268	167227
2012	160070	150635	164882	143001	139480	131485	118512	119217	97688	130916	134274	132317
2013	144148	125742	137073	138446	140598	129969	114262	101714	110161	123400	127484	133787
2014	141241	129204	137254	134456	135561	118024	122072	111578	105837	123336	132379	135019
2015	138575	129515	146638	135170	158215	134299	121194	122421	118188	141591	148110	157179
2016	152239	133753	141103	137424	143657	120471	144156	123485	129753	144501	142717	151045
2017	162280	136213	148674	146263	148390	131100	126927	106634	122377	158917	156381	124685
2018	152174	137369	140321	147490	139481	116591	132558	119844	117570	130685	137218	144907
2019	147783	131873	141055	130985	134586	114115	124497	115850	127724	135792	148421	149888

   - Plotting the time series:
     plot.ts(jmd1, main="natural gas production in tonnes")
     (a) <u>Trend-</u> Decreasing exponential trend is observed. This indicates the **production** of natural gas in UK has been **decreasing over the years**.
     (b) <u>Seasonality-</u> there is a presence of **multiplicative seasonality** as the amplitude is decreasing over time and even the width is decreasing. The **production at start of the year is high but it decreases as the year comes to end**. This seasonality is observed in all the years except some noise at 2013 and mid- end of 2017
     (c) <u>Level</u>- The average production of natural gas in UK has decreased.
     (d) <u>Noise-</u> There is presence of noise (unexplained component in each year).

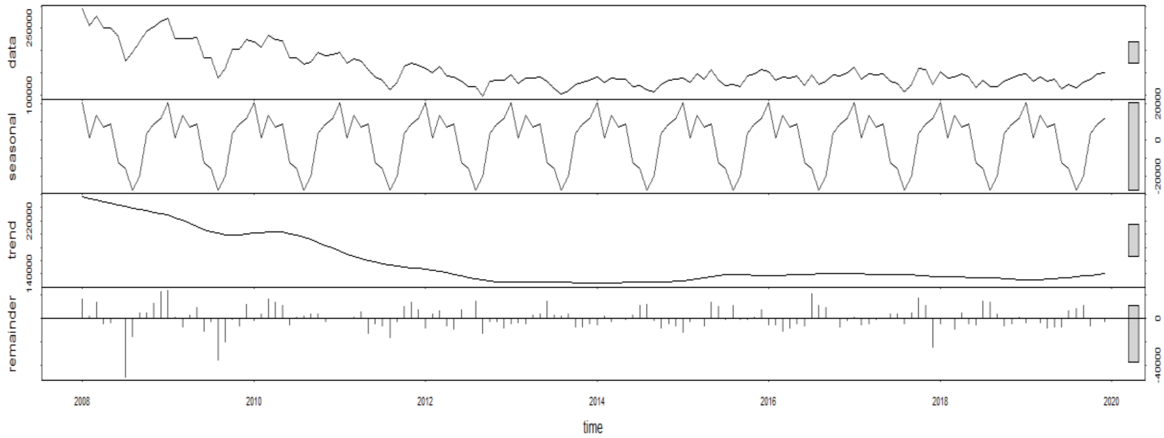


- Smoothing the curve:
  using ma() function reduces the data at both ends but it smoothens the curve by decreasing the noise in it.
  jmd2<-autoplot(jmd1) + autolayer(ma(jmd1,5)) + autolayer(ma(jmd1,7))

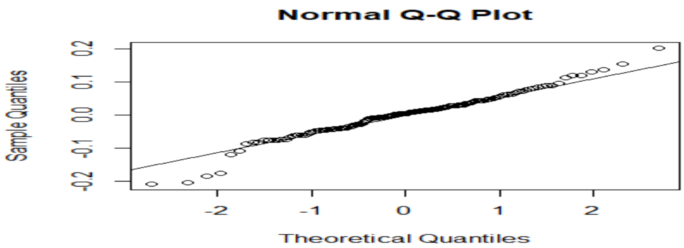


- Decomposing the time series:
  Using stl() function we will decompose the curve in its trends, seasonality and residual plots. This will help us to deeply analyze the components of time series and compare the results to our initial findings.
  (a) Seasonality – The seasonality explains the initial decrease in production of natural gas in the first quarter, in the second quarter it was overall same, the third and fourth quarter further shows decreases in the production of natural gas.
  (b) Trend – The production decreases from 2008 to 2014. The production increases for the year 2015 and 2016 then it remains constant till 2019 end.
  (c) Remainder – this explains the noise in the plot, overall the noise plot is good as it has less variation.

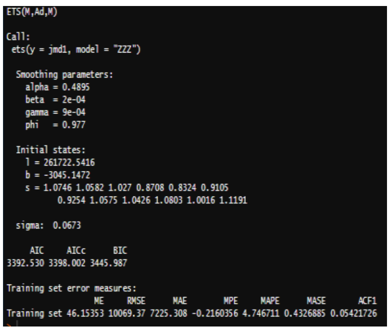
Statistics for Data Analytics Individual Project A



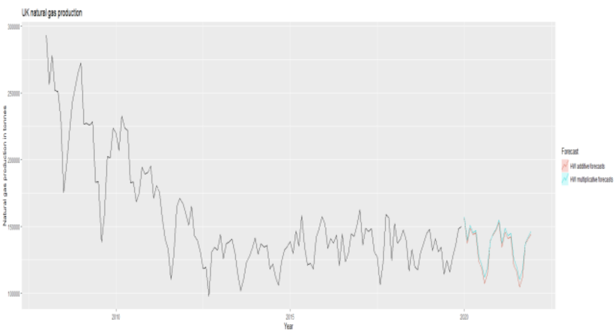
- Forecasting –
  The format of the ets() function is: ets(ts, model="ZZZ") (where ts is a time series and the model is specified by three letters. The first letter denotes the error type, the second letter denotes the trend type, and the third letter denotes the seasonal type. Allowable letters are A for additive, M for multiplicative, N for none, and Z for automatically selected)
  As the model in this report has trend and seasonality the best model is Holt-winter model. The automated ets() function gave (M,Ad,M) as the best model.
  (a) The AICc value is lowest from all the models.
  (b) The significance value is greater then 0.05 suggesting its not significant at 95% confidence interval.
  (c) The Mean error (ME) value is lowest.
  (d) The qq plot of residual is normal with the mean around zero.



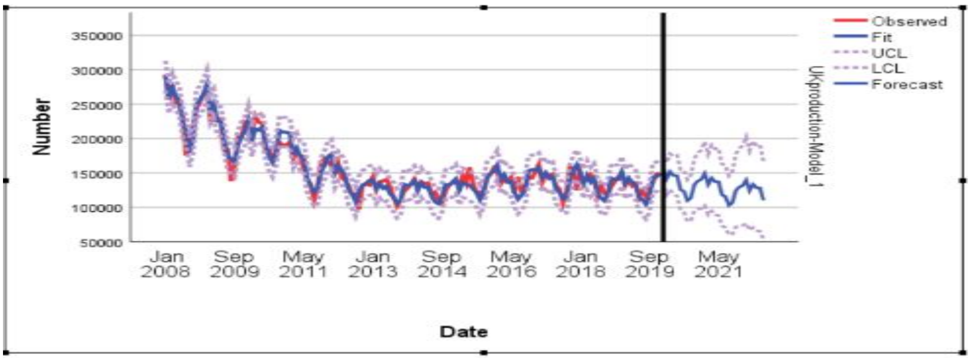
  (e) The parameter estimates are α^=0.4895, β^=0.0002, and γ^=0.0009. The output also returns the estimates for the initial states ℓ0, b0, s0, s−1, s−2 and s−3.
  (f) The small values of β and γ mean that the slope and seasonal components change very little over time. The narrow prediction intervals indicate that the series is relatively easy to forecast due to the strong trend and seasonality.
  (g) The multiplicative Holt- winter fits the forecast better.



```
ETS(M,Ad,M)

Call:
 ets(y = jmd1, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.4895
    beta  = 2e-04
    gamma = 9e-04
    phi   = 0.977

  Initial states:
    l = 261722.5416
    b = -3045.1472
    s = 1.0746 1.0582 1.027 0.8708 0.8324 0.9105
           0.9254 1.0575 1.0426 1.0803 1.0016 1.1191

  sigma:  0.0673

     AIC     AICc      BIC
3392.530 3398.002 3445.987

Training set error measures:
                   ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set 46.15353 10069.37 7225.308 -0.2160356 4.746711 0.4326885 0.05421726
```



- Forecast plot:
  The SPSS was used to summarize the model and find the best suitable forecast plot. It automatically created a forecast model, only start time and frequency was entered in it. The model created by SPSS is significant.



Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile 5	10	25	50	75	90	95
Stationary R-squared	.597	.	.597	.597	.597	.597	.597	.597	.597	.597	.597
R-squared	.933	.	.933	.933	.933	.933	.933	.933	.933	.933	.933
RMSE	10815.928	.	10815.928	10815.928	10815.928	10815.928	10815.928	10815.928	10815.928	10815.928	10815.928
MAPE	5.187	.	5.187	5.187	5.187	5.187	5.187	5.187	5.187	5.187	5.187
MaxAPE	30.031	.	30.031	30.031	30.031	30.031	30.031	30.031	30.031	30.031	30.031
MAE	7838.028	.	7838.028	7838.028	7838.028	7838.028	7838.028	7838.028	7838.028	7838.028	7838.028
MaxAE	37443.831	.	37443.831	37443.831	37443.831	37443.831	37443.831	37443.831	37443.831	37443.831	37443.831
Normalized BIC	18.681	.	18.681	18.681	18.681	18.681	18.681	18.681	18.681	18.681	18.681

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
UK production-Model_1	0	.597	.933	10815.928	5.187	7838.028	30.031	37443.831	18.681	30.284	15	.011	0

Summary:

A. **The time series analysis shows the production of natural gas in UK over the years has a lower level, decreasing exponential trend. The seasonality present is very less and of multiplicative nature as the production quantity is decreasing in both amplitude and width. There are also some noise present at the beginning years, but it is not significant. The presence of all three component suggests using of holt-winter model for forecasting.**
B. **The forecast model presented in R is not that significant but has linear residual plot. SPSS gave a Winter multiplicative model which is significant as the sig. value is less then 0.05. R square value of 0.597 indicates it can explain 59.7% od variation in the production of natural gas which is good. The MAPE of 5% indicates low error in the prediction of production. There are no outliners present hence RMSE is important here and not the MAE value.**

Conclusion:

Multiple Regression: The model used to predict total deaths can is highly accurate it explains 96.7% variance in total deaths. There are some outliners present which can be removed, to increase the R square further.

Time Series: The time series analysis shows the production of natural gas (tonnes) in UK over the years is decreasing exponentially with a little seasonal decrease in production every quarter. The production seasonality also decreases over the years becoming constant at the end. The prediction model is Winter Multiplicative, which explains 59.7% variance in the production of natural gas. There are no outliners present. The prediction is, the production will remain in range of 10000 – 15000 tonnes in the coming years.