“Impact of Transformations in Modeling and Forecasting with ARIMA: Time Series Analysis & Forecasting of UK Drivers Death Dataset”

Developed by

MD. ZUBAYER

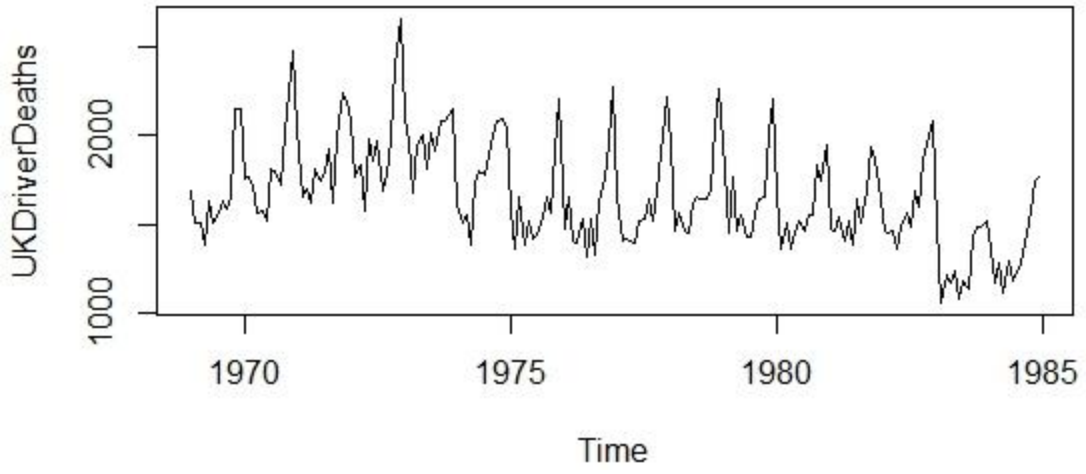
Table of Contents

Section 1: Introduction

In this project, the impact of transformations in modeling and forecasting with ARIMA in time series analysis using the R programming language is explored. The objective is to analyze and forecast the UK Driver Deaths dataset, which contains monthly data on the number of road traffic fatalities in the UK from 1969 to 1984. The dataset is divided into training and testing sets, and several ARIMA models are fitted to the training data. The performance of these models is evaluated using various diagnostic tests and accuracy measures.

Section 2: Time Series Components Analysis

2.1: Time plot of UKDriverDeaths



From the time plot, we may be able to identify the presence of time series components such as Trend, Seasonality, Cyclical Fluctuation and Irregular Fluctuation. For this particular time series plot-

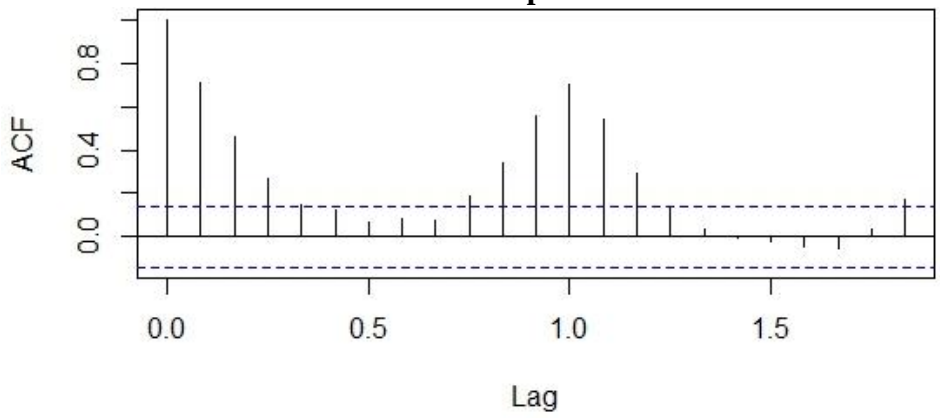
Trend: From the time plot, it is evident that there is no trend.

Seasonality: There are seasonal fluctuations spotted from the plot. Thus, there is seasonality.

Cyclical Fluctuation: This graph does not show any cyclical fluctuations.

Irregular Fluctuation: There are irregular fluctuations between 1973 to 1974 and around 1983 to 1984.

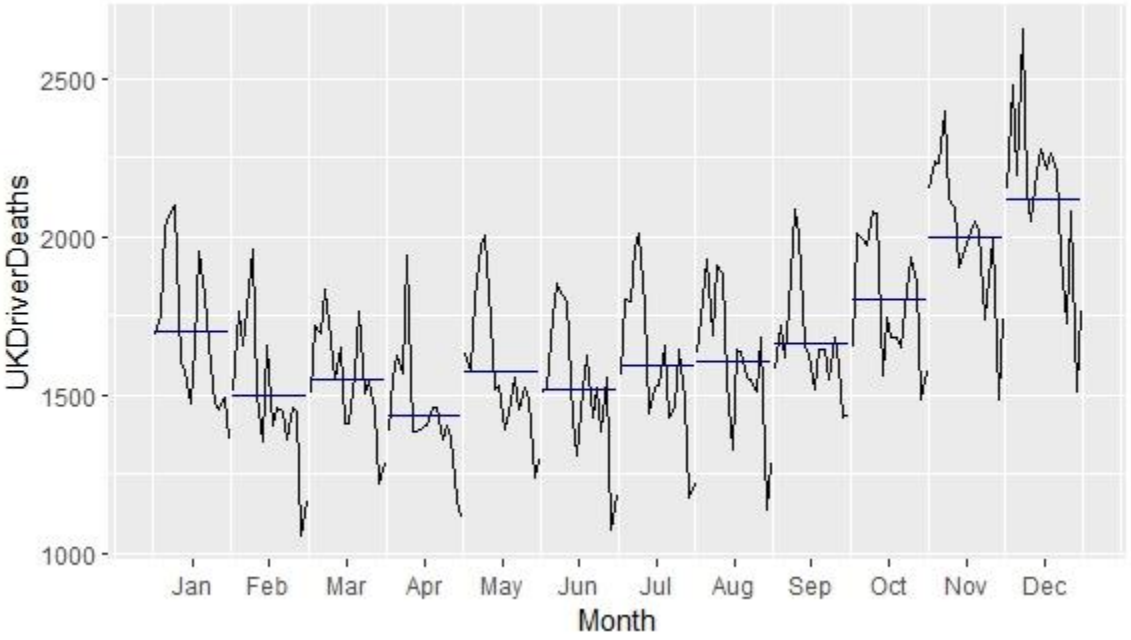
2.2: ACF Plot of UKDriverDeaths and Interpretation



Interpretation of ACF plot:

- There is no evidence of trend.
- There is a spike on the seasonal value – 12th value, this suggests that there is seasonality.
- Since the ACF values crosses 95% of CI, thus the series is not stationary.
- This is not an altering series.
- There may not be any evidence of outliers from this ACF plot.

2.3: UKDriverDeaths- ggsubseriesplot

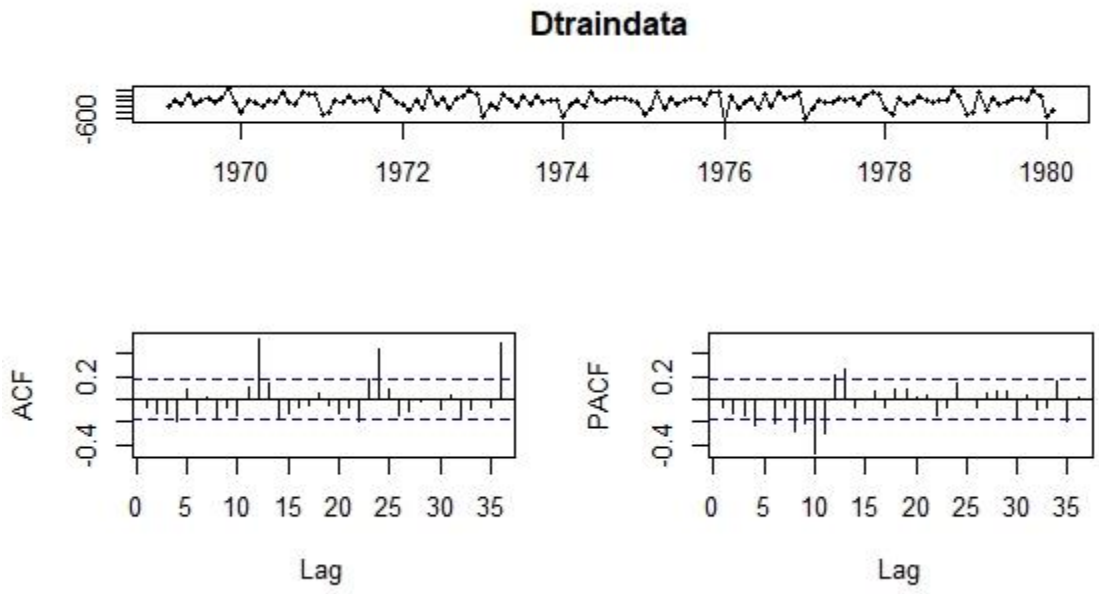


From the ggsubseries plot of UKdriverdeath, it can be interpreted that – the death count of the drivers rises high during October, November and December months. December month shows the highest driver deaths. On contrary, in February, March and April months rate of driver death is relatively lower. April month shows the lowest driver deaths.

Section 3: Model 1

In the first step, stationary check process done with Augmented Dickey-Fuller Test (using adf.test function in tseries package) and The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test (using kpss.test function in tseries package).

After taking differenc, on the training data, the Dtraindata plot illustrates as follows:



Here from the plot, it has been identified that the series becomes stationary. Also, ARIMA model is selected as **M1 =** ARIMA(1,1,0)(1,1,1)[12]

3.1: Model 1 Output

ARIMA(1,1,0)(1,1,1)[12]

Coefficients:

	ar1	sar1	sma1
	-0.4624	0.0975	-0.9999
s.e.	0.0804	0.0984	0.3439

sigma^2 = 18882: log likelihood = -779.34
AIC=1566.68 AICc=1567.03 BIC=1577.87

3.2: Mathematical Expression of the M1 model

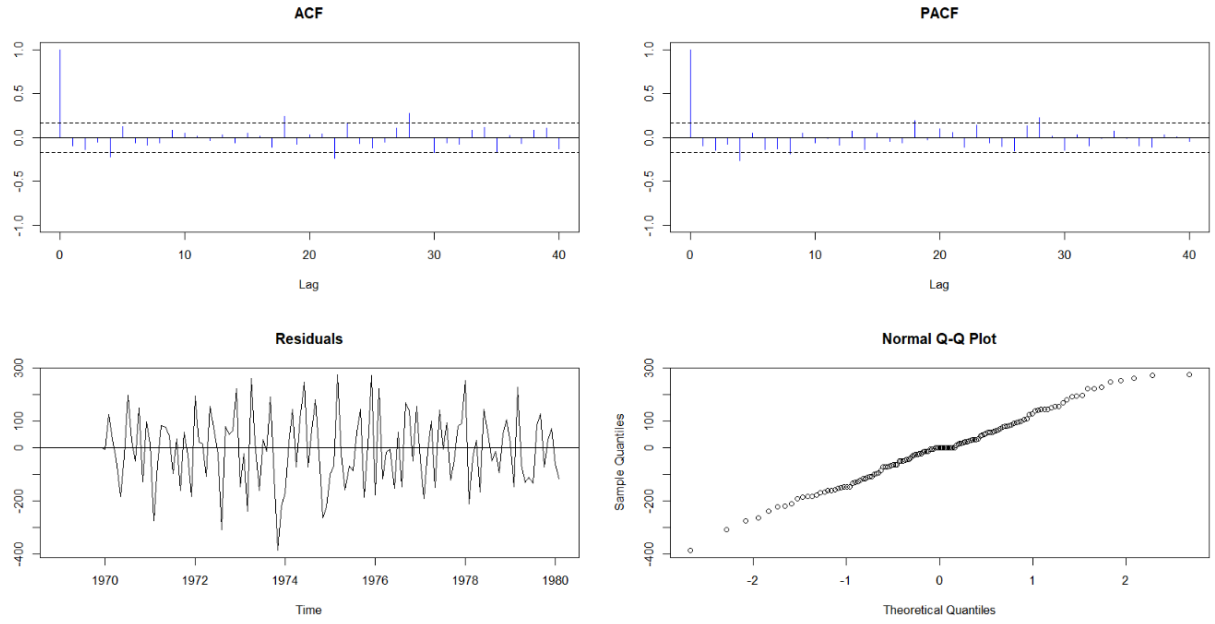
$\phi 1 = -0.4624$ (ar1), $\Phi 1 = 0.0975$ (sar1), $\Theta 1 = -0.9999$ (sma1)

Based on these values, the equation will be

$$(1-(-0.4624)B)(1-0.0975B^{12})(1-B)(1-(-0.9999)B^{12})\, Xt = Zt$$

Where {Zt} ~(0,18882)

3.3: Residual Diagnosis of M1



From the ACF and PACF plots we can see there is no significant spikes. Also, from the Q-Q plot, errors are following approximately 45° straight line. Which means Errors are following normal distribution with mean zero and variance sigma square.

Section 4: Model 2

By analyzing the time plot, the suitable transformation is Logarithm transformation. Thus, the train data has gone through logarithmic transformation. Then the same steps applied as for the model 1.

ARIMA model is selected as **M2** = ARIMA(2,1,1)(1,0,1)[12]

4.1: Model 2 Output

Coefficients:

	ar1	ar2	ma1	sar1	sma1
	0.2757	0.2014	-0.879	0.9945	-0.8580
s.e.	0.1137	0.1023	0.066	0.0115	0.1465

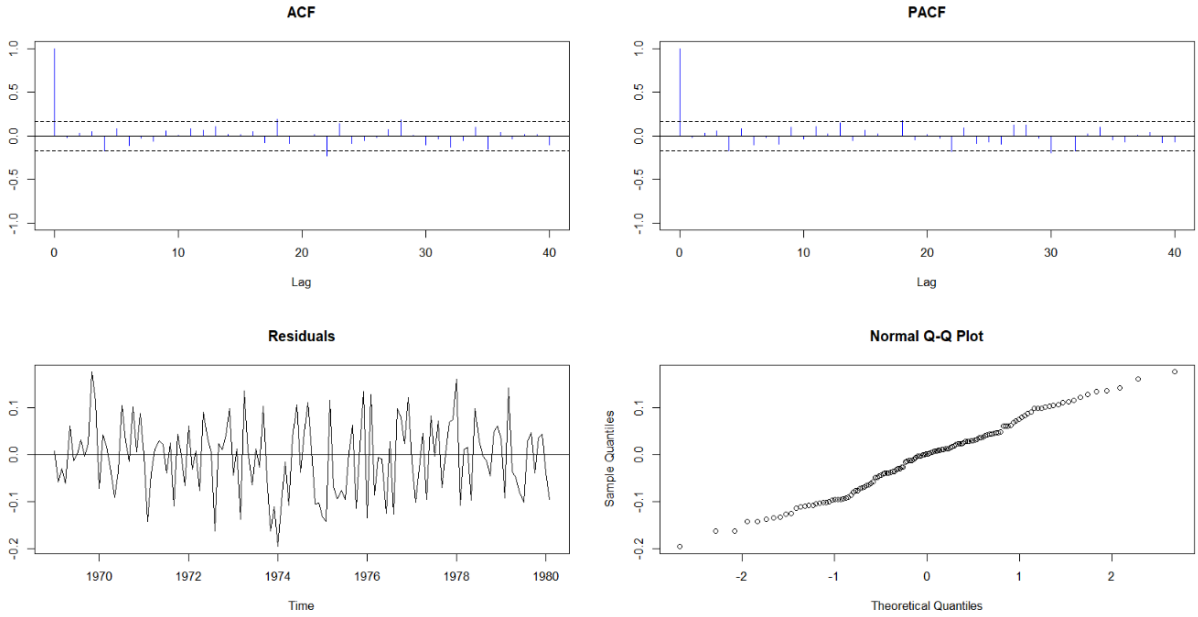
sigma^2 = 0.006422: log likelihood = 137.32
AIC=-262.65 AICc=-261.98 BIC=-245.31

4.2: Mathematical Expression of the M2 model

$(1-0.2757B-0.2014B^2)(1-0.9945B^{12})\ Xt=(1+0.879B)(1+0.8580B^{12})Zt$

Where {Zt} ~(0,.006422)

4.3: Residual Diagnosis of M2



From the ACF and PACF plots we can see there are no significant spikes. Also, from the Q-Q plot, errors are following approximately 45° straight line. That suggests Errors are following normal distribution with mean zero and variance sigma square.

Section 5: Model 3

In the model 3 Boxcox transformation has used. Thus, the train data has gone through Boxcox transformation with optimum lambda using -

BoxCox(traindata,lambda = BoxCox.lambda(traindata)). Then similar steps applied like the Model 1.

ARIMA model is selected as **M3** =ARIMA(1,1,1)(1,0,3)[12]

5.1: Model 3 Output

Coefficients:

	ar1	ma1	sar1	sma1	sma2	sma3
	0.2745	-0.8209	0.9987	-0.826	-0.198	0.0952
s.e.	0.1269	0.0766	0.0054	0.151	0.121	0.0954

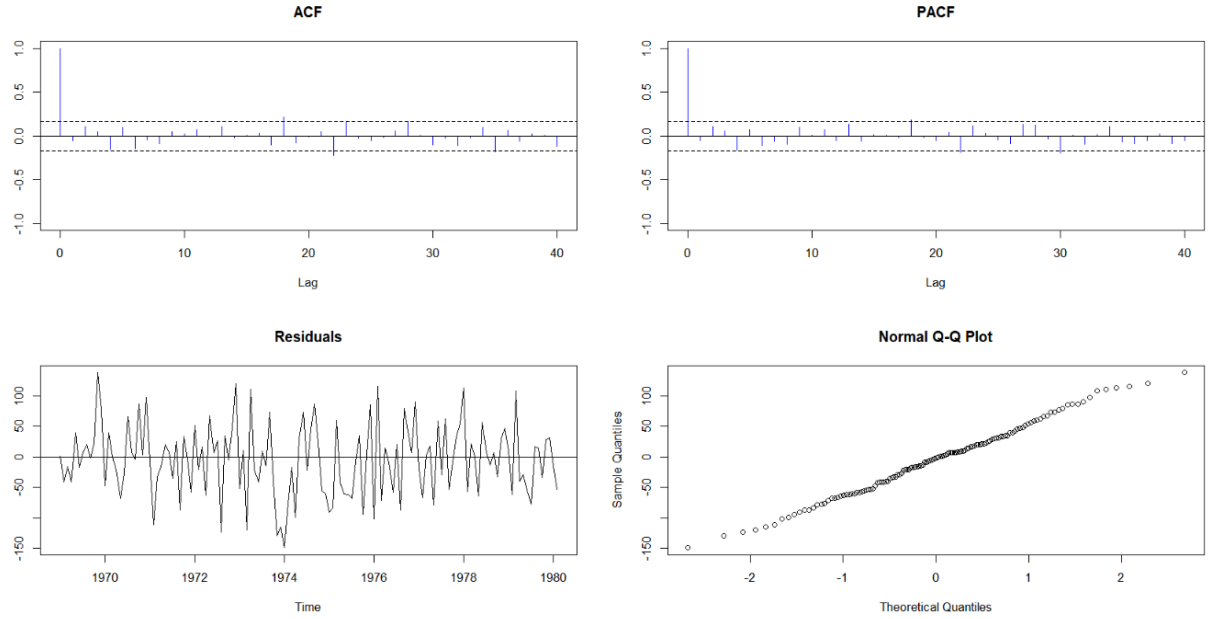
sigma^2 = 3633: log likelihood = -746.26
AIC=1506.52 AICc=1507.42 BIC=1526.76

5.2: Mathematical Expression of the M3 model

$(1-B)(1-0.9987B^{12})(1+0.826B)(1+0.198B^{12})(1-0.0952B^{12})Xt=(1-0.2745B)(1-0.8209B)\ Zt$

Where {Zt} ~(0,3633)

5.3: Residual Diagnosis of M3



From the ACF and PACF plots we can see there are no significant spikes other than lag 0. Also, from the Q-Q plot, errors are following approximately 45° straight line. That suggests Errors are following normal distribution with mean zero and variance sigma square.

Section 6: Model 4

Model 4 is developed by auto.arima function of R. Using this, the Arima model:

M4=ARIMA(0,1,1)(2,1,0)[12]

6.1: Model 3 Output

Coefficients:

	ma1	sar1	sar2
	-0.5921	-0.5506	-0.3847
s.e.	0.0982	0.0859	0.0827

sigma^2 = 22949: log likelihood = -780.83
AIC=1569.65 AICc=1570 BIC=1580.84

6.2: Mathematical Expression of the M4 model

$(1+0.5506B)(1-0.5506B^{12})(1+0.3847B^{12})Xt=Zt$

Where {Zt} ~(0,22949)

Section 7: Model Selection Criteria

Model	AIC	AICc	BIC
M1	1566.68	1567.03	1577.87
M2	-262.65	-261.98	-245.31
M3	1506.52	1507.42	1526.76
M4	1569.65	1570	1580.84
Minimum Value	Model 2		

To measure the prediction error, AIC, AICc and BIC values of the models are compared. In this case among the 4 models the model 2 has lower AIC, AICc and BIC values. However, selecting M2 as the best model will be inadequate in this because, since it is a logarithmic transformed model. Hence, the best model will be selected based on forecasting accuracy – RMSE value.

Section 8: Residual Diagnosis Comparison

Test	Null hypothesis: Residuals are iid noise; P Value				
	M1	M2	M3	M4	Highest Value
Ljung-Box Q	0.0658	0.496	0.2456	0.4422	0.4617
McLeod-Li Q	0.0773	0.5292	0.2019	0.1081	0.269
Turning points T	0.4093	1	0.6799	0.8366	
Diff signs S	0.8815	0.2967	0.2967	0.6547	
Rank P	0.7277	0.858	0.7744	0.7191	

To assess which model best fulfills the underlying assumptions of residuals, typically p-values greater than 0.05 for the diagnostic tests are checked. This indicates that the null hypothesis of the tests (that the residuals are independent and identically distributed noise) cannot be rejected, suggesting that the model's residuals meet the assumptions.

Now, by analyzing the p-values for the Ljung-Box Q test and McLeod-Li Q test for each model, it is found that, Model 2 has the highest p-values for both the Ljung-Box Q test and the McLeod-Li Q test, indicating that its residuals are closest to meeting the assumptions of independent and identically distributed noise among the four models. However, it also exhibits more fluctuations in the residuals compared to the other models. Model 1 has

Therefore, the selection of the best-fit model should consider both the p-values of diagnostic tests and the characteristics of the residuals.

Section 9: Forecasting Using Four Models

In this section, forecasted values of each model will be illustrated via plot. The forecast is done based on the test data.

Test Data												
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980	-	-	1506	1360	1453	1522	1460	1552	1548	1827	1737	1941
1981	1474	1458	1542	1404	1522	1385	1641	1510	1681	1938	1868	1726
1982	1456	1445	1456	1365	1487	1558	1488	1684	1594	1850	1998	2079
1983	1494	1057	1218	1168	1236	1076	1174	1139	1427	1487	1483	1513
1984	1357	1165	1282	1110	1297	1185	1222	1284	1444	1575	1737	1763

M1

Model 1 Forecasted Values on Test Data												
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980			1457	1308	1458	1390	1468	1486	1521	1623	1911	2065
1981	1591	1369	1408	1291	1439	1379	1464	1472	1501	1612	1893	2044
1982	1576	1363	1396	1282	1429	1371	1456	1463	1492	1604	1884	2034
1983	1567	1355	1387	1273	1421	1363	1448	1455	1483	1595	1875	2026
1984	1559	1346	1379	1265	1413	1354	1440	1446	1475	1587	1867	2018

M2

Model 2 Forecasted Values on Test Data												
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980	-	-	1481	1374	1495	1462	1529	1545	1585	1668	1924	2073
1981	1647	1436	1494	1388	1497	1460	1524	1538	1576	1657	1910	2056
1982	1636	1427	1485	1380	1487	1451	1514	1528	1566	1646	1895	2039
1983	1624	1418	1475	1371	1478	1442	1504	1518	1555	1634	1880	2023
1984	1613	1409	1466	1363	1468	1433	1494	1508	1545	1623	1866	2006

M3

Model 3 Forecasted Values on Test Data												
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980			1485	1378	1527	1449	1519	1548	1582	1681	1947	2091
1981	1641	1437	1453	1370	1510	1466	1551	1545	1573	1683	1937	2079
1982	1642	1444	1477	1371	1505	1453	1530	1538	1570	1671	1934	2080
1983	1637	1429	1467	1363	1498	1446	1523	1531	1562	1663	1926	2072
1984	1629	1422	1460	1356	1491	1439	1516	1524	1555	1656	1919	2064

M4

Model 4 Forecasted Values on Test Data												
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1980			1530	1375	1412	1433	1440	1538	1531	1607	1952	2155
1981	1738	1350	1518	1358	1385	1442	1458	1516	1529	1580	1936	2141
1982	1691	1325	1550	1336	1392	1372	1379	1470	1510	1549	1906	2105
1983	1625	1279	1473	1291	1334	1344	1352	1440	1458	1512	1865	2067
1984	1616	1250	1439	1260	1300	1322	1333	1411	1430	1481	1835	2038

Section 10: Forecasting Accuracy Measures

Model	MSE	RMSE	MAE	MPE	MAPE
M1	31312.58	176.954	143.174	-3.90927	10.0205
M2	37069.17	192.534	150.406	-7.86949	10.88216
M3	41198.59	202.974	157.784	-8.66503	11.41585
M4	34051.62	184.531	143.991	-4.14357	9.862876
Minimum Error	31312.58	**176.954**	143.174	-8.66503	9.862876
Preferred Model	M1				

Based on the Forecasting Accuracy Measures table:

- Model 1 appears to perform the best overall, having relatively lower error measures- RMSE, MSE compared to the other models.
- Model 4 also shows competitive performance, especially in terms of MAPE and MPE, making it a close contender to Model 1.
- Model 2 and Model 3 demonstrate inferior performance compared to Models 1 and 4, with higher error measures across the board.

Section 11: Selecting the Best Model

Among the four models, M1(model 1) outperforms other models in terms of forecasting accuracy measures. At the same time, residual diagnostic results for Model 1:

- The Ljung-Box Q and McLeod-Li Q tests are used to assess the presence of autocorrelation in the residuals. In this case, both tests have p-values greater than 0.05, suggesting that there is no strong evidence to reject the null hypothesis that the residuals are iid noise.
- The Turning points T test and Diff signs S test assess normality and symmetry of the residuals, respectively. Both tests have high p-values (0.4093 and 0.8815), suggesting no significant departure from normality or symmetry.
- The Rank P test evaluates randomness in the residuals. The p-value of 0.7277 indicates that there is no significant deviation from randomness.

Overall, the diagnostic tests do not provide strong evidence of violation of the white noise assumption.

Thus, **M1(Model 1)** is selected for forecasting the 10 points ahead.

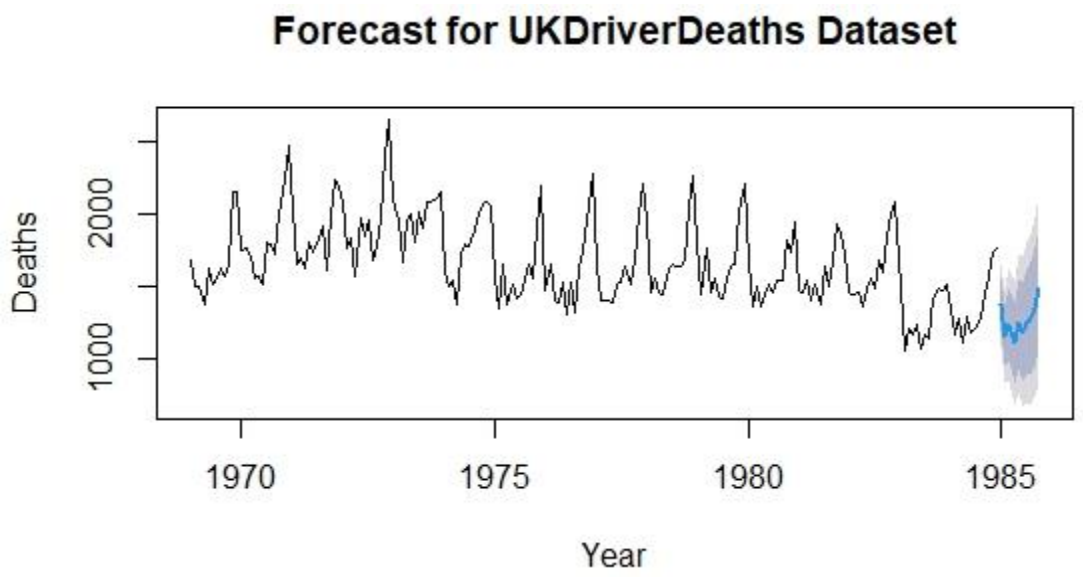
Section 12: Forecasting the Next 10 points

Using **M1(Model 1)**, the next 10 points of the UKDriverDeaths Dataset is forecasted.

Forecasted Values

10 Forecasted Values Ahead, Using Best Model for Entire Dataset										
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
1985	1381	1162	1238	1116	1249	1191	1257	1281	1356	1496

The plot for the 10 points ahead of December 1984 for the UKDriverDeaths dataset illustrated as following:



Section 13: Conclusion

In conclusion, the project provides valuable insights into the application of ARIMA models for time series forecasting. Throughout the project it has been identified the best-performing model based on prediction errors (AIC, AICc, BIC), forecasting accuracy measures and residual diagnostic tests and utilizes it to forecast future values of the UK Driver Deaths dataset. The findings contribute to the understanding of ARIMA modeling techniques and their potential applications in forecasting real-world datasets. Also, this comparison can be further extended with ETS models of Time series analysis.