Tittle; predict rainfall data;

Steps:

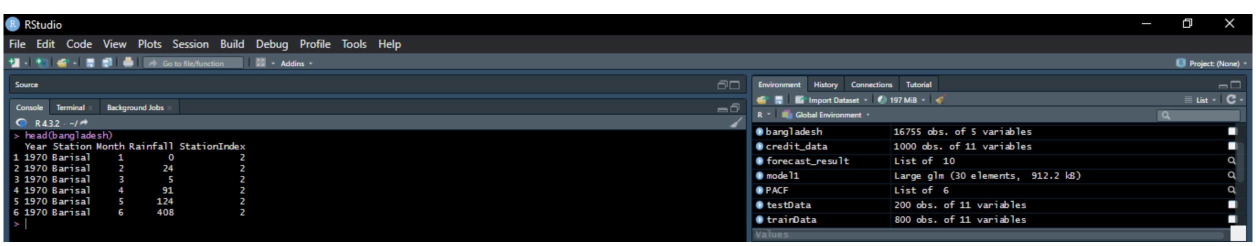
- make a ts plot
- make a acf-pacf plot
- identify possible models
- fit and check diagnosis
- choose the best model
- make a forecast

Data source; the data set provided in the link below contains the year, station, month, station index and the amount of rainfall experienced in Bangladesh. The data is derived from a GitHub website that is associated with Bangladesh.

https://raw.githubusercontent.com/arnavgarg123/Bangladesh-Rainfall/master/historical-rainfall-data-in-bangladesh/data_monthly_rainfall.csv

Overview of dataset

This shows the corresponding categories and there values in our dataset; the screenshot below shows a glance view of how our data is tabled.



Procedure

- Change the given data into time series data

Code

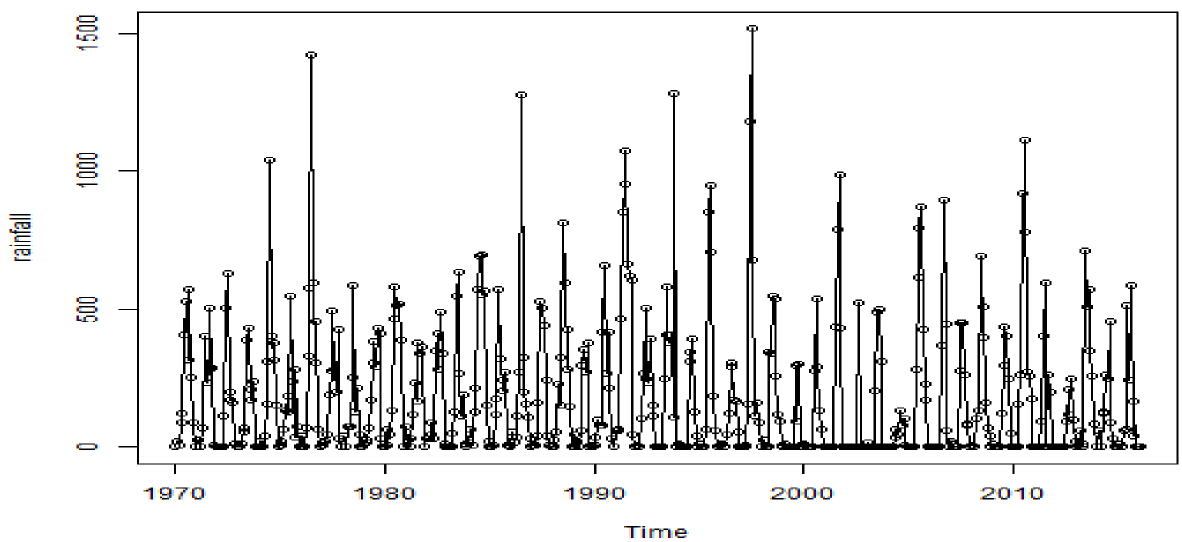
```
rainfall <- ts (bangladesh$Rainfall, start = 1970, end = 2016, frequency = 12)

rainfall
```

- Identify whether the time series data is stationery or non-stationary .the time series data is non-stationary though there is no observable trend but there is seasonality hence there will be need for differencing

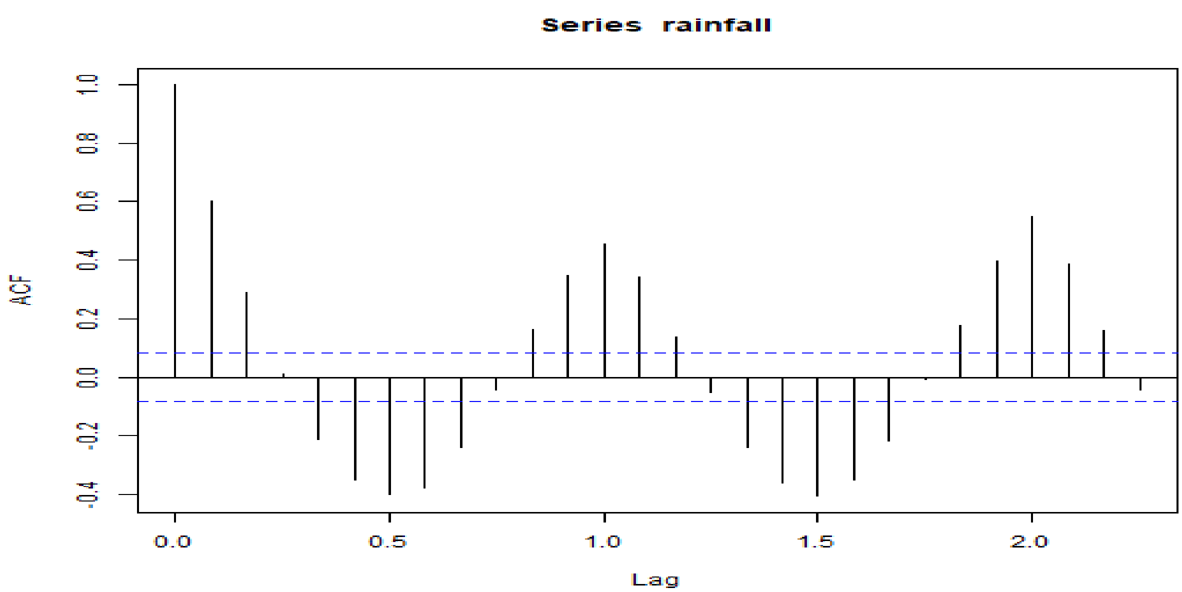
  Reasons for differencing

  Helps to eliminate trend and seasonality; a time series data that is associated with trend and seasonality is considered as a compromised data and hence predictive or forecasting analysis is biased and do not have a strong background to put a claim on a result. This data is differenced first to eliminate trend and seasonality components, differencing makes the data stationary hence predictive and forecasting analysis is possible and is also considered unbiased.

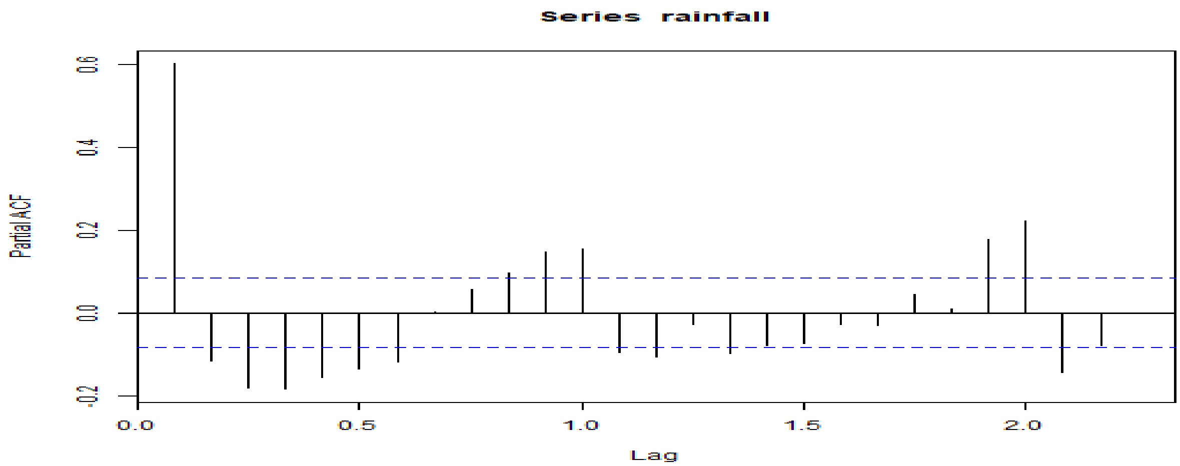


- acf-pacf plot of the time series data before differencing acf plot

acf (rainfall) #there is a strong autocorrelation hence the time series data is non-stationary

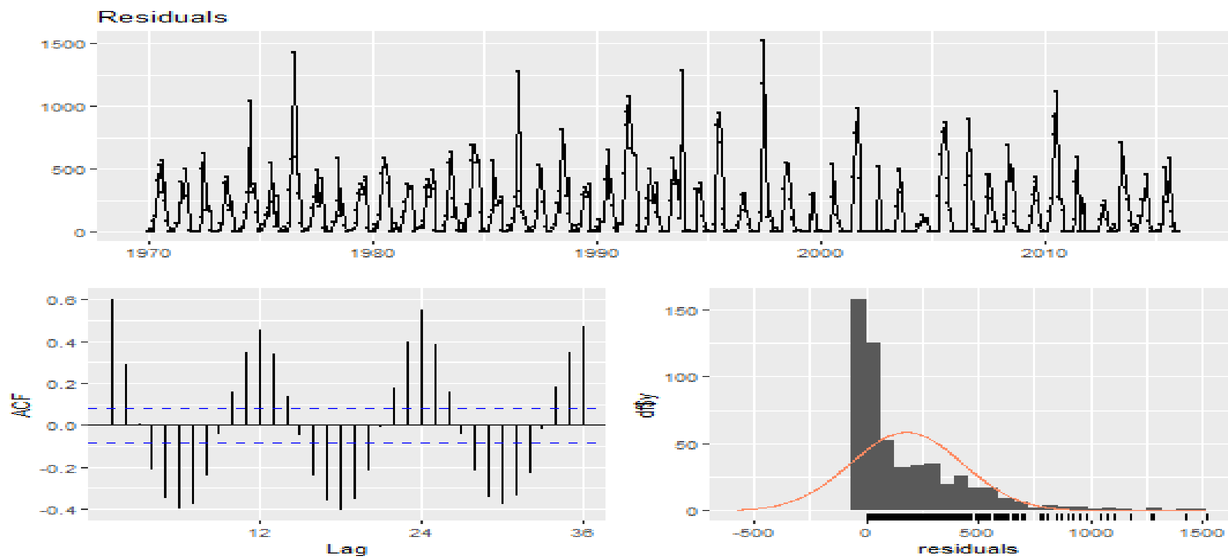


Pacf (rainfall) # there is a strong autocorrelation hence the time series data is non-stationary



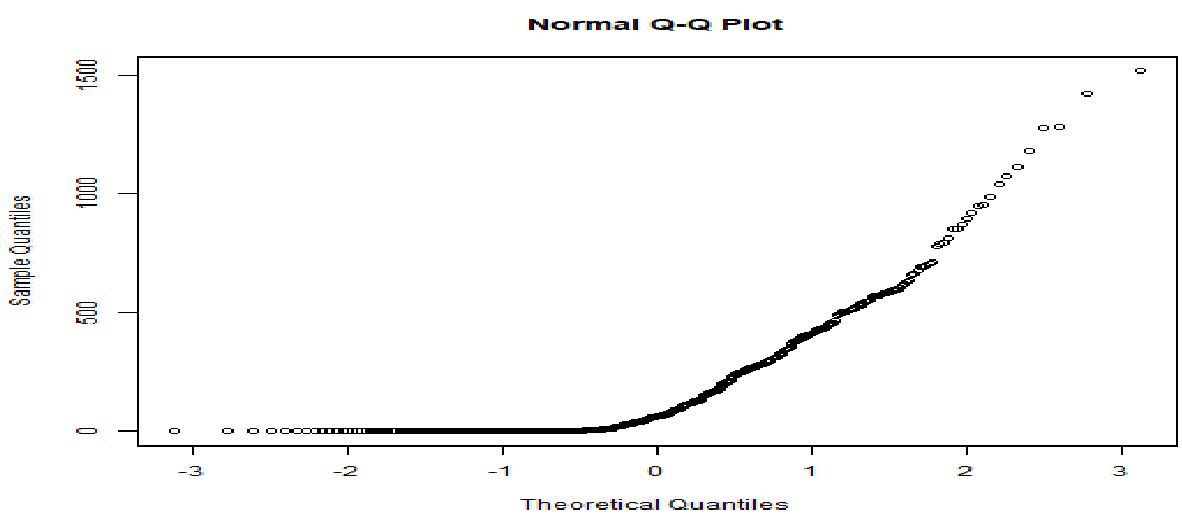
- fitting and checking diagnosis of the non-stationary time series by this the plot should show autocorrelation and also partial autocorrelation but in general the residual should show very little or no autocorrelation or partial autocorrelation in order to be considered as a good model

forecast::checkresiduals (rainfall) #as it is observable in the visualized diagram below the normal distribution plot is not normally distributed this is due to non-stationarity of the time series data

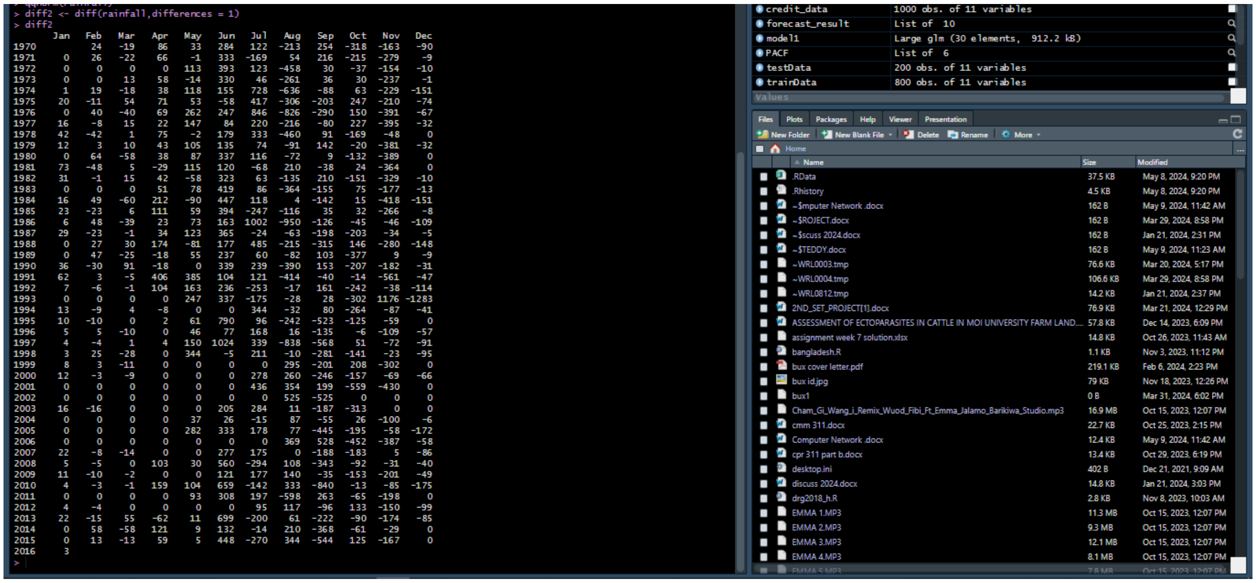


- Checking for normality of the time series data as seen in the plot the time series data is not normally distributed major reason being the data is not stationary

qqnorm(rainfall)



- Differencing the data due to seasonality; this helps to convert non-stationary time series data to stationery thus enabling forecasting of the time series data.

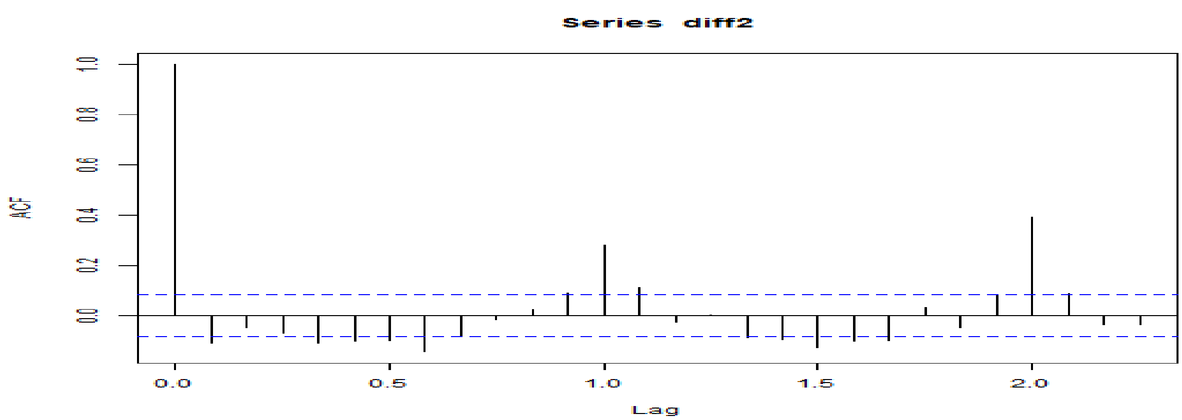


```
> diff2 <- diff(rainfall,differences = 1)
> diff2
```

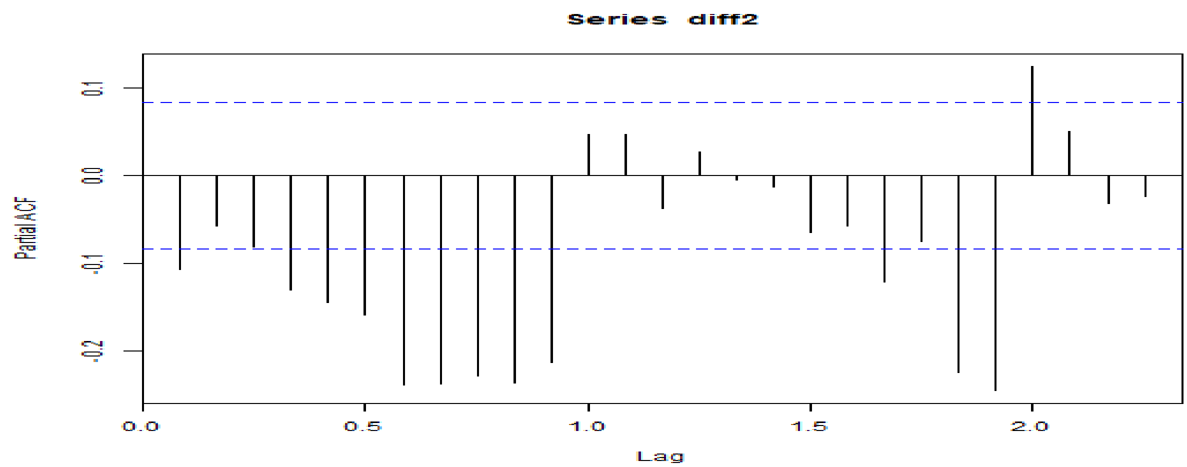
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1970		24	-19	86	33	284	122	-213	254	-318	-163	-90
1971	0	26	-22	66	-1	333	-169	54	216	-215	-279	-9
1972	0	0	0	0	113	393	123	-458	30	-37	-154	-10
1973	0	0	13	58	-14	330	46	-261	36	30	-237	-1
1974	1	19	-18	38	118	155	728	-636	-88	63	-229	-151
1975	20	-11	54	71	53	-58	417	-306	-203	247	-210	-74
1976	0	40	-40	69	262	247	846	-826	-290	150	-391	-67
1977	16	-8	15	22	147	84	220	-216	-80	227	-395	-32
1978	42	-42	1	75	-2	179	333	-460	91	-169	-48	0
1979	12	3	10	43	105	135	74	-91	142	-20	-381	-32
1980	0	64	-58	38	87	337	116	-72	9	-132	-389	0
1981	73	-48	5	-29	115	120	-68	210	-38	24	-364	0
1982	31	-1	15	42	-58	323	63	-135	210	-151	-329	-10
1983	0	0	0	51	78	419	86	-364	-155	75	-177	-13
1984	16	49	-60	212	-90	447	118	4	-142	15	-418	-151
1985	23	-23	6	111	59	394	-247	-116	35	32	-266	-8
1986	6	48	-39	23	73	163	1002	-950	-126	-45	-46	-109
1987	29	-23	-1	34	123	365	-24	-63	-198	-203	-34	-5
1988	0	27	30	174	-81	177	485	-215	-315	146	-280	-148
1989	0	47	-25	-18	55	237	60	-82	103	-377	9	-9
1990	36	-30	91	-18	0	339	239	-390	153	-207	-182	-31
1991	62	3	-5	406	385	104	121	-414	-40	-14	-561	-47
1992	7	-6	-1	104	163	236	-253	-17	161	-242	-38	-114
1993	0	0	0	0	247	337	-175	-28	28	-302	1176	-1283
1994	13	-9	4	-8	0	0	344	-32	80	-264	-87	-41
1995	10	-10	0	2	61	790	96	-242	-523	-125	-59	0
1996	5	5	-10	0	46	77	168	16	-135	-6	-109	-57
1997	4	-4	1	4	150	1024	339	-838	-568	51	-72	-91
1998	3	25	-28	0	344	-5	211	-10	-281	-141	-23	-95
1999	8	3	-11	0	0	0	0	295	-201	208	-302	0
2000	12	-3	-9	0	0	0	278	260	-246	-157	-69	-66
2001	0	0	0	0	0	0	436	354	199	-559	-430	0
2002	0	0	0	0	0	0	0	525	-525	0	0	0
2003	16	-16	0	0	0	205	284	11	-187	-313	0	0
2004	0	0	0	0	37	26	-15	87	-55	26	-100	-6
2005	0	0	0	0	282	333	178	77	-445	-195	-58	-172
2006	0	0	0	0	0	0	0	369	528	-452	-387	-58
2007	22	-8	-14	0	0	277	175	0	-188	-183	5	-86
2008	5	-5	0	103	30	560	-294	108	-343	-92	-31	-40
2009	11	-10	-2	0	0	121	177	140	-35	-153	-201	-49
2010	4	-3	-1	159	104	659	-142	333	-840	-13	-85	-175
2011	0	0	0	0	93	308	197	-598	263	-65	-198	0
2012	4	-4	0	0	0	0	95	117	-96	133	-150	-99
2013	22	-15	55	-62	11	699	-200	61	-222	-90	-174	-85
2014	0	58	-58	121	9	132	-14	210	-368	-61	-29	0
2015	0	13	-13	59	5	448	-270	344	-544	125	-167	0
2016	3											

- acf-pacf plot of the time series data after differencing acf there is a significance at lag 1, 1.5, 2 hence its semi-annual seasonality

acf (diff2)

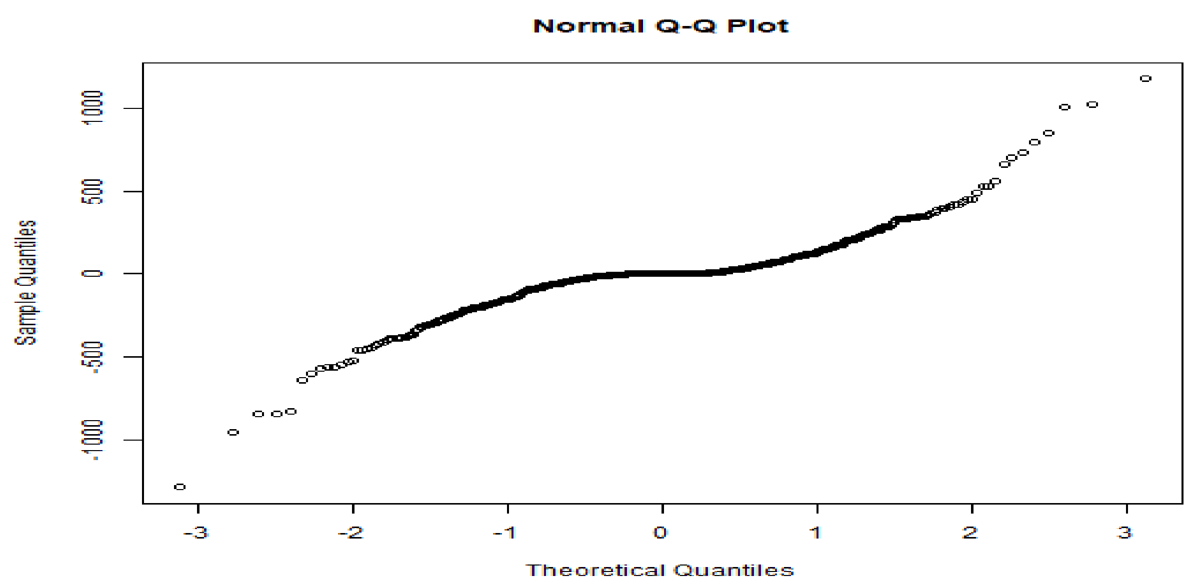


pacf(diff2)



- Checking for normality of the differenced data the time series data is normally distributed

  qqnorm (diff2)

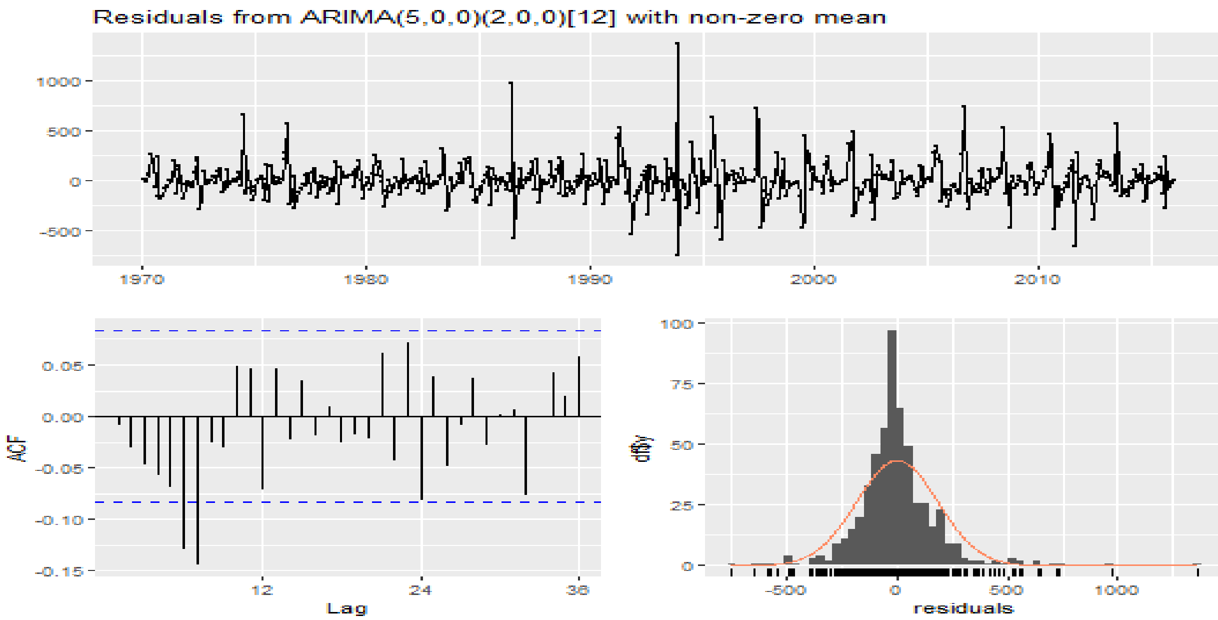


- get the best arima model for our differenced data

  arima1 <- forecast:: auto.arima (diff2)

  arima1

forecast::checkresiduals (arima1); In general the residual shows a very little autocorrelation hence we would consider ARIMA(5,0,0)(2,0,0)[12] as the best arima model.



- Forecast the arima model; enables us to use our model as a predictive model. It outlays how the future of the time series data may be structured; provides the clue on how the future of time series data may look like