Step 1: Importing of Dataset

Downloaded the data for S&P 500 for Year 1967 (1950 + 12 +5) Jan-June.

```
* Importing Data ;
FILENAME REFFILE2 '/folders/myfolders/TimeSeries/table2.csv';

PROC IMPORT DATAFILE=REFFILE2
	DBMS=CSV
	OUT=sp500csv3;
	GETNAMES=YES;
RUN;
```

Step 2: Cleaning Data

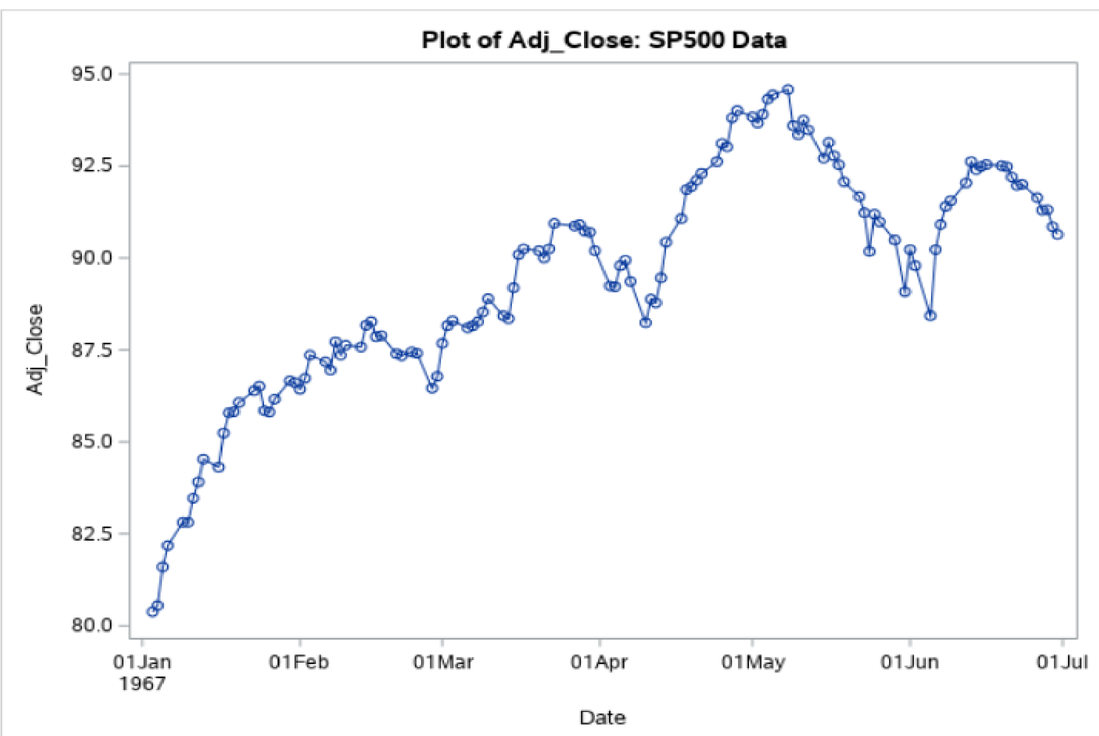
After importing the data, there were many blank observations that might have been due to downloading data from source. Removing observations with no values in it. This doesn't result in any loss of data but only helps to clean the data.

```
* Removing Null Values from Data ;

data sp500csv3 ; set sp500csv3;
if compress(cats(of _all_),'.')=' ' then delete;
run;
```

Step 3: Visualizing Data

```
title "Plot of Adj_Close: SP500 Data";
   proc sgplot data=sp500csv3;
      series x=Date y=Adj_Close  / markers;
   run;
```



Above is the graph for the data points for the Adj Close variable from S&P data. The mean of the graph is increasing while the variance seems to be constant or fluctuating at points. This calls for one to check for the non-stationarity in mean for the data.

Step 4: Stationarity Check

```
proc arima data=sp500csvtransformed3;
identify var=Adj_Close nlag=31 stationarity=(adf) ;
run;
```

- The mean of the working series is 89.43 while the standard deviation is 3.09
- From the table Autocorrelation check for white noise, the p-values are significant to reject the null hypothesis that the series is a white noise
- From the ADF test table, since none of the p-values are not significant enough to reject the null-hypothesis that the series has a Unit Root i.e Non-stationarity in mean
- Below are the ACF and PACF plots corresponding to the original series. But since there is a non-stationarity involved. It would be better to look at the plots after differencing the series.

The ARIMA Procedure

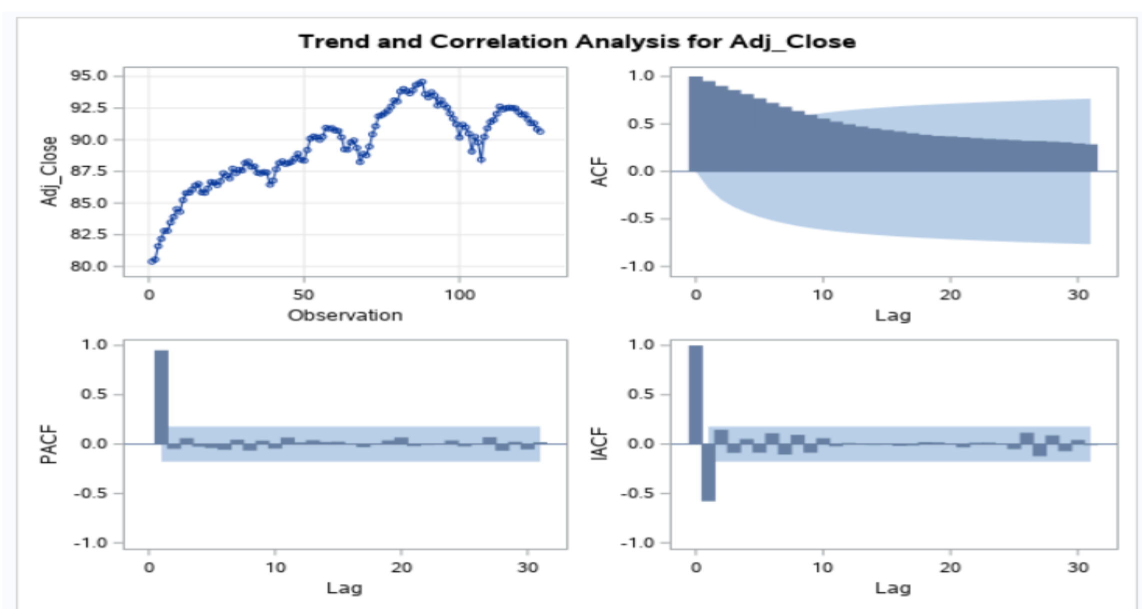
Name of Variable = Adj_Close	
Mean of Working Series	89.43127
Standard Deviation	3.097327
Number of Observations	126

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	554.85	6	<.0001	0.951	0.900	0.857	0.814	0.769	0.722
12	838.78	12	<.0001	0.681	0.637	0.597	0.557	0.524	0.497
18	999.93	18	<.0001	0.473	0.453	0.436	0.421	0.403	0.386
24	1117.03	24	<.0001	0.373	0.368	0.361	0.351	0.342	0.337
30	1214.40	30	<.0001	0.330	0.322	0.321	0.313	0.307	0.296

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	0.1074	0.7063	1.63	0.9748		
	1	0.1047	0.7056	1.43	0.9618		
	2	0.0948	0.7033	1.38	0.9580		
Single Mean	0	-5.9731	0.3424	-3.28	0.0181	7.02	0.0010
	1	-6.4871	0.3029	-3.24	0.0203	6.52	0.0010
	2	-5.5681	0.3763	-2.92	0.0465	5.42	0.0260
Trend	0	-7.7434	0.5928	-2.30	0.4327	5.55	0.0917
	1	-9.2718	0.4728	-2.48	0.3382	5.60	0.0890
	2	-7.2893	0.6301	-2.04	0.5728	4.38	0.3022



Step 5: Estimating the Non-Stationarity Parameters

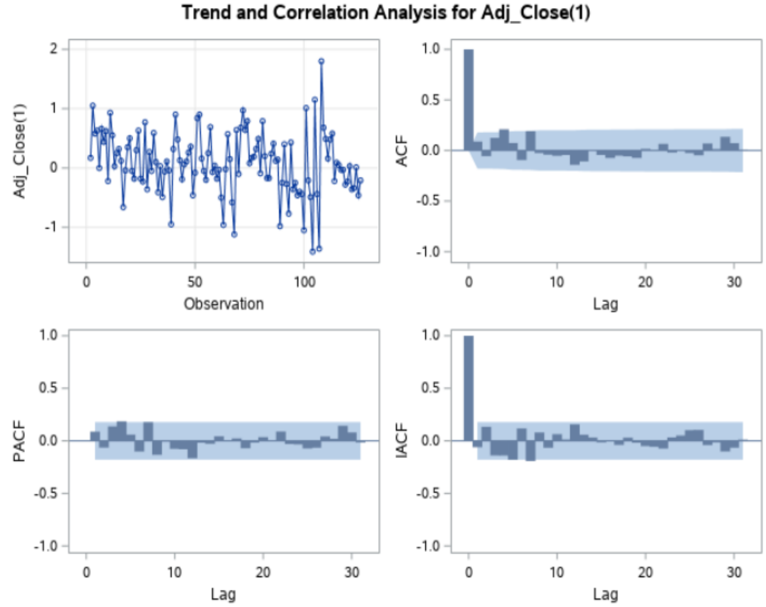
```
proc arima data=sp500csvtransformed3;
identify var=Adj_Close(1) nlag=31 minic esacf scan ;
run;
```

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-110.330	0.0001	-9.93	<.0001		
	1	-120.055	0.0001	-7.80	<.0001		
	2	-78.6863	<.0001	-5.52	<.0001		
Single Mean	0	-113.067	0.0001	-10.10	<.0001	51.00	0.0010
	1	-127.826	0.0001	-7.98	<.0001	31.84	0.0010
	2	-86.4694	0.0011	-5.64	<.0001	15.96	0.0010
Trend	0	-118.130	0.0001	-10.49	<.0001	55.03	0.0010
	1	-142.428	0.0001	-8.36	<.0001	34.96	0.0010
	2	-103.053	0.0001	-5.95	<.0001	17.72	0.0010

Name of Variable = Adj_Close	
Period(s) of Differencing	1
Mean of Working Series	0.08208
Standard Deviation	0.52363
Number of Observations	125
Observation(s) eliminated by differencing	1

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	11.07	6	0.0861	0.088	-0.056	0.125	0.209	0.074	-0.094
12	19.80	12	0.0710	0.191	-0.025	-0.041	-0.053	-0.044	-0.142
18	23.43	18	0.1746	-0.108	0.001	-0.044	-0.074	-0.051	-0.058
24	25.01	24	0.4051	-0.074	0.016	0.006	0.065	-0.021	-0.003
30	30.16	30	0.4573	-0.025	-0.047	0.069	0.014	0.135	0.072

Trend and Correlation Analysis for Adj_Close(1)



- After inspecting the first order difference series, ADF test reveals that the series is now stationary. The p-values are significant and can reject the null hypothesis that there is Unit root

- The mean of the working series is 0.08 while the standard deviation is 0.52

- From the table Autocorrelation check for white noise, the p-values are significant (alpha=0.1) to reject the null hypothesis that the series is a white noise

- The plot for ACF suggests that it gets cut off at a lag of suggesting MA(4)

- Using MINIC, SCAN and ESCAF functionality, tried determining the best fit for the model

- BIC best suggests that ARIMA(0,1,0); While SCAN and ESCAF best suggests a ARIMA(0,1,1). On combining information available by looking at the ACF plots and test results. I suggest looking at the model ARIMA(0,1,4) for the original series.

ESACF Probability Values						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	0.3255	0.5358	0.1671	0.0225	0.4411	0.3275
AR 1	<.0001	0.8566	0.5580	0.1064	0.3191	0.9704
AR 2	<.0001	0.1423	0.4579	0.8325	0.4711	0.6696
AR 3	<.0001	0.0098	0.8721	0.9176	0.2174	0.2836
AR 4	0.0004	0.0101	0.9275	0.0279	0.2614	0.7046
AR 5	<.0001	0.7004	0.0842	0.8510	0.0022	0.5212

SCAN Chi-Square[1] Probability Values						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	0.3239	0.5289	0.1590	0.0185	0.4284	0.3127
AR 1	0.4785	0.7125	0.2481	0.2487	0.2245	0.8720
AR 2	0.1213	0.2226	0.1815	0.8631	0.1676	0.4009
AR 3	0.0316	0.3877	0.9669	0.4186	0.0529	0.5277
AR 4	0.4389	0.2291	0.1848	0.0956	0.1235	0.5916
AR 5	0.2862	0.8857	0.5327	0.9530	0.4348	0.2241

Minimum Information Criterion						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	-1.3875	-1.36419	-1.34286	-1.31243	-1.29356	-1.27626
AR 1	-1.35962	-1.33003	-1.30546	-1.27403	-1.25793	-1.26767
AR 2	-1.33284	-1.3	-1.26914	-1.23611	-1.22302	-1.22911
AR 3	-1.3111	-1.27525	-1.24135	-1.20657	-1.18746	-1.20159
AR 4	-1.29613	-1.27023	-1.24097	-1.20348	-1.1737	-1.18219
AR 5	-1.26186	-1.26017	-1.22179	-1.18705	-1.15651	-1.15435

Error series model: AR(10)

Minimum Table Value: BIC(0,0) = -1.3875

ARMA(p+d,q) Tentative Order Selection Tests					
SCAN			ESACF		
p+d	q	BIC	p+d	q	BIC
1	1	-1.33003	1	1	-1.33003
4	0	-1.29613	0	4	-1.29356
0	4	-1.29356			

(5% Significance Level)

```
proc arima data=sp500csvtransformed3;
identify var=Adj_Close(1) nlag=31 ;
estimate q=4;
run;
```

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.08419	0.05804	1.45	0.1495	0
MA1,1	-0.04673	0.08985	-0.52	0.6040	1
MA1,2	0.03160	0.09011	0.35	0.7264	2
MA1,3	-0.06022	0.09013	-0.67	0.5053	3
MA1,4	-0.17864	0.09027	-1.98	0.0501	4

- Estimates for mu, theta1, theta2, theta3 are non-significant as per the p-value. Only theta4 is significant.

```
proc arima data=sp500csvtransformed3;
identify var=Adj_Close(1) nlag=31 ;
estimate q=(4) noconstant;
run;
```

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	-0.22622	0.08791	-2.57	0.0113	4

Variance Estimate	0.268421
Std Error Estimate	0.518094
AIC	191.331
SBC	194.1593
Number of Residuals	125

* AIC and SBC do not include log determinant.

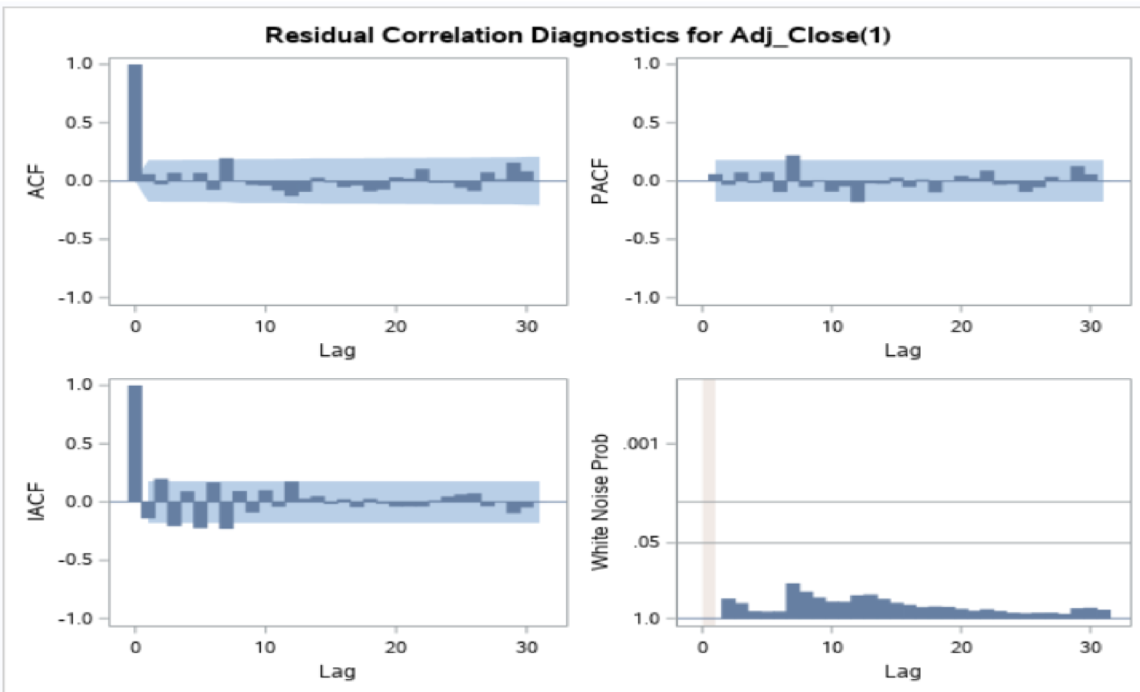
Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	3.11	5	0.6835	0.075	-0.014	0.084	0.010	0.084	-0.061
12	11.66	11	0.3901	0.207	0.019	-0.020	-0.026	-0.070	-0.116
18	13.97	17	0.6695	-0.078	0.038	0.001	-0.041	-0.027	-0.077
24	16.80	23	0.8191	-0.068	0.038	0.029	0.108	-0.006	-0.005
30	25.45	29	0.6547	-0.048	-0.078	0.084	0.023	0.167	0.094

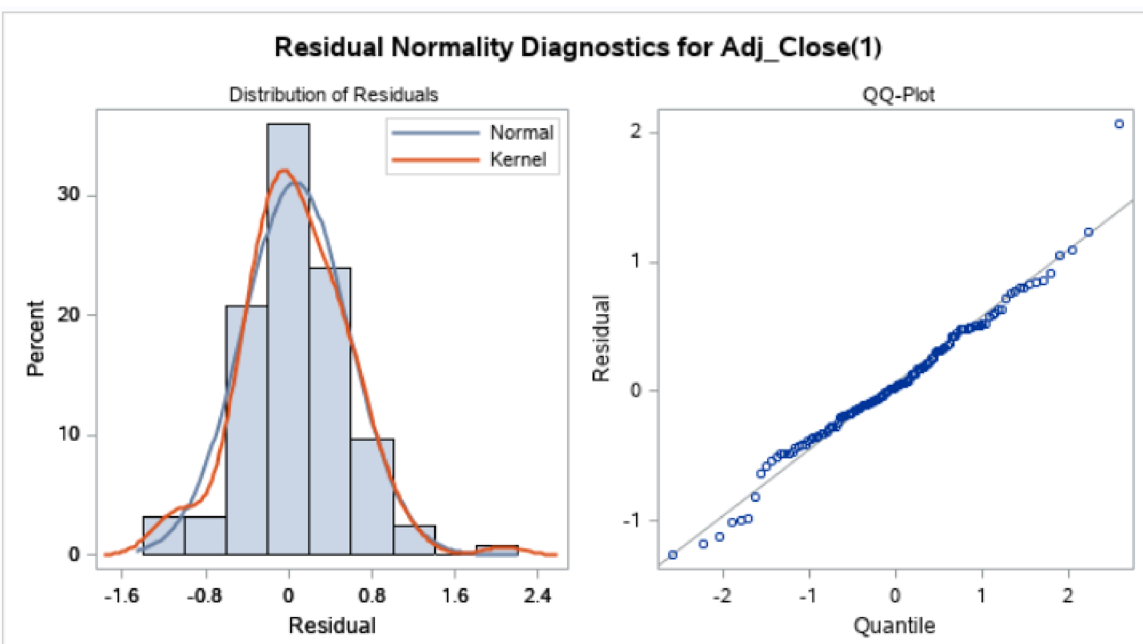
Model for variable Adj_Close	
Period(s) of Differencing	1

No mean term in this model.

Moving Average Factors	
Factor 1:	1 + 0.22622 B**(4)

- P-value corresponding to the hypothesis theta4=0 is significant. Hence rejecting it. theta4 of the model is -0.22622
- AIC and BIC for the model are 191.33 and 194.15 respectively
- From the table Autocorrelation check for residuals, the P-values are insignificant and hence fails to reject the null hypothesis that the residuals follows a white noise
- Both ACF and PACF plot falls between the bands and looks fine for the residuals.
- The below plots verify that the residuals follows the normal distribution similar to a white noise. Also provides and estimates of theta values for the model.





Step 6: Forecast

```
proc arima data=sp500csvtransformed3;
identify var=Adj_Close(1) nlag=31 ;
estimate q=(4) noconstant;
run;
forecast lead=31 out=output;
quit;
```

- The forecast for the next thirty days is most constant around the mean value 90.45. The variance increases as the timeperiod from the observed month increases. i.e the variance explodes.
- The final model for the series is ARIMA(0,1,4) with theta4= -0.22622

Forecasts for variable Adj_Close				
Obs	Forecast	Std Error	95% Confidence Limits	
127	90.5773	0.5181	89.5618	91.5927
128	90.5919	0.7327	89.1558	92.0279
129	90.4847	0.8974	88.7259	92.2435
130	90.4562	1.0362	88.4253	92.4871
131	90.4562	1.2154	88.0740	92.8384
132	90.4562	1.3715	87.7682	93.1442
133	90.4562	1.5115	87.4938	93.4186
134	90.4562	1.6395	87.2427	93.6696
135	90.4562	1.7583	87.0099	93.9024
136	90.4562	1.8696	86.7919	94.1205
137	90.4562	1.9746	86.5861	94.3263
138	90.4562	2.0742	86.3907	94.5216
139	90.4562	2.1694	86.2043	94.7080
140	90.4562	2.2605	86.0257	94.8866
141	90.4562	2.3480	85.8541	95.0583
142	90.4562	2.4325	85.6886	95.2237
143	90.4562	2.5141	85.5287	95.3837
144	90.4562	2.5931	85.3738	95.5385
145	90.4562	2.6698	85.2235	95.6889
146	90.4562	2.7443	85.0774	95.8350
147	90.4562	2.8169	84.9351	95.9772
148	90.4562	2.8877	84.7965	96.1159
149	90.4562	2.9567	84.6611	96.2512
150	90.4562	3.0242	84.5289	96.3835
151	90.4562	3.0902	84.3995	96.5129
152	90.4562	3.1548	84.2728	96.6395
153	90.4562	3.2182	84.1487	96.7637
154	90.4562	3.2803	84.0270	96.8854
155	90.4562	3.3412	83.9075	97.0049
156	90.4562	3.4011	83.7902	97.1222
157	90.4562	3.4599	83.6749	97.2375