MKT6971
Time Series Project
Example

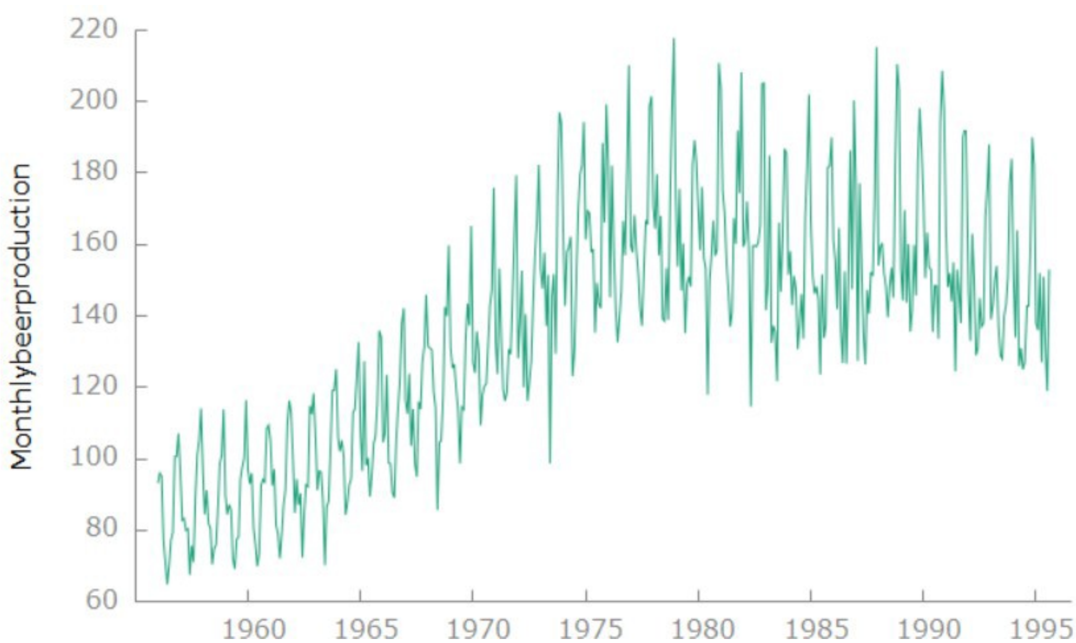
The objective of this project is for you to practice what you have learned about time series analysis and interpreting data. I suggest you use GRETL for this project.

1. Select a scientific, biomedical, business or other issue that appeals to you and go looking online for relevant time series data sets. The good news here is that there are tons of free and interesting time series data sets online. If you have problems locating them let me know and I will help. **Be sure that it looks like there is little or no seasonality to it.**

   I picked a dataset from Kaggle - https://www.kaggle.com/shenba/time-series-datasets?select=monthly-beer-production-in-austr.csv

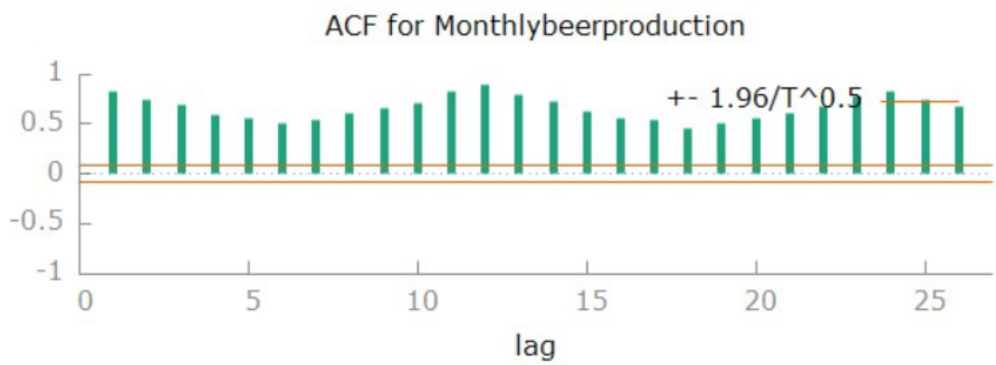
   This dataset has information about monthly production of the beer count from the year 1956 to 1995.

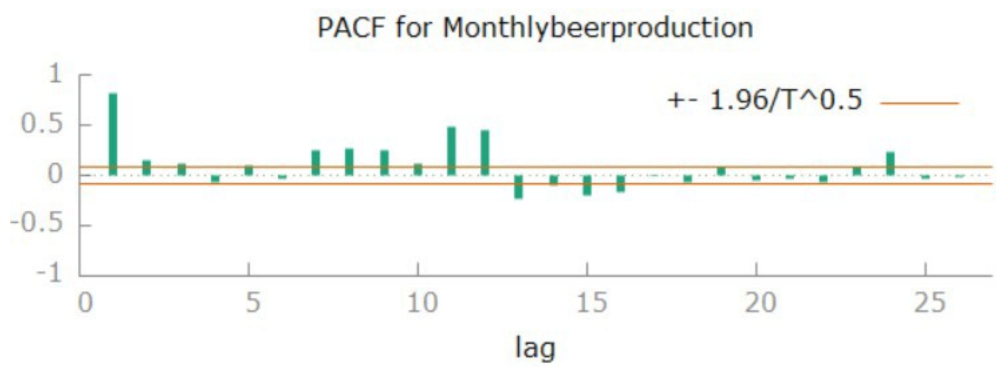
2. Plot out your time series variable. Tell me using your Mark I eyeball whether or not you think the time series data set is stationary in terms of **constant mean** and also **constant variance**.



Using Mark I Eyeball, if we draw an imaginary line through the center of the data, we can see that there is a trend / there is a non-constant mean. This also implies that the variance is not constant.

3. Plot the ACF for the time series data set. Looking at ACF, does it look like there may be a trend or non-constant mean for each time series?





It looks like the ACF exhibits a ski slope behavior which indicates that there is a trend / non-constant mean.

Autocorrelation function for Monthlybeerproduction
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0.5

LAG	ACF		PACF		Q-stat. [p-value]
1	0.8287	***	0.8287	***	328.9573 [0.000]
2	0.7358	***	0.1564	***	588.8089 [0.000]
3	0.6818	***	0.1236	***	812.3937 [0.000]
4	0.5952	***	-0.0739		983.1381 [0.000]
5	0.5604	***	0.1087	**	1134.8737 [0.000]
6	0.5071	***	-0.0356		1259.3709 [0.000]
7	0.5318	***	0.2609	***	1396.5744 [0.000]
8	0.5995	***	0.2639	***	1571.3239 [0.000]

9	0.6603	***	0.2566	***	1783.7160	[0.000]
10	0.7068	***	0.1263	***	2027.6532	[0.000]
11	0.8197	***	0.4869	***	2356.3907	[0.000]
12	0.8964	***	0.4620	***	2750.3754	[0.000]
13	0.7847	***	-0.2309	***	3052.9741	[0.000]
14	0.7185	***	-0.0981	**	3307.2263	[0.000]
15	0.6303	***	-0.1990	***	3503.2988	[0.000]
16	0.5537	***	-0.1706	***	3654.9230	[0.000]
17	0.5305	***	-0.0089		3794.4548	[0.000]
18	0.4598	***	-0.0703		3899.5017	[0.000]
19	0.5030	***	0.0814	*	4025.4858	[0.000]
20	0.5613	***	-0.0488		4182.6685	[0.000]
21	0.6000	***	-0.0344		4362.7176	[0.000]
22	0.6673	***	-0.0661		4585.8575	[0.000]
23	0.7708	***	0.0883	*	4884.2929	[0.000]
24	0.8324	***	0.2368	***	5233.0423	[0.000]
26	0.7470	***	-0.0377		5514.5723	[0.000]
26	0.6750	***	-0.0206		5744.9664	[0.000]

4. Now let's examine the time series data set using unit root tests. First use the KPSS test for the time series data set and tell me if the test suggests if there is a constant mean or not. Then see if you can confirm your KPSS evaluation using the Augmented Dickey Fuller (ADF) or the ADF-GLS test and tell me what the ADF test suggests is the case.

   KPSS Test:

   H0: There is a constant mean / no trend.
   Halt: There is a non-constant mean / there is a trend.

```
KPSS test for Monthlybeerproduction

T = 476
Lag truncation parameter = 5
Test statistic = 5.74095

                    10%      5%      1%
Critical values: 0.348   0.462   0.742
P-value < .01
```

From the results the p - value is less than 0.01 which is less than the significance value 0.05 so we can reject the H0 hypothesis and conclude that there is a non-constant mean / there is a trend.

   ADF Test:

   H0: There is a non-constant mean / there is a trend.
   Halt: There is a constant mean / no trend.

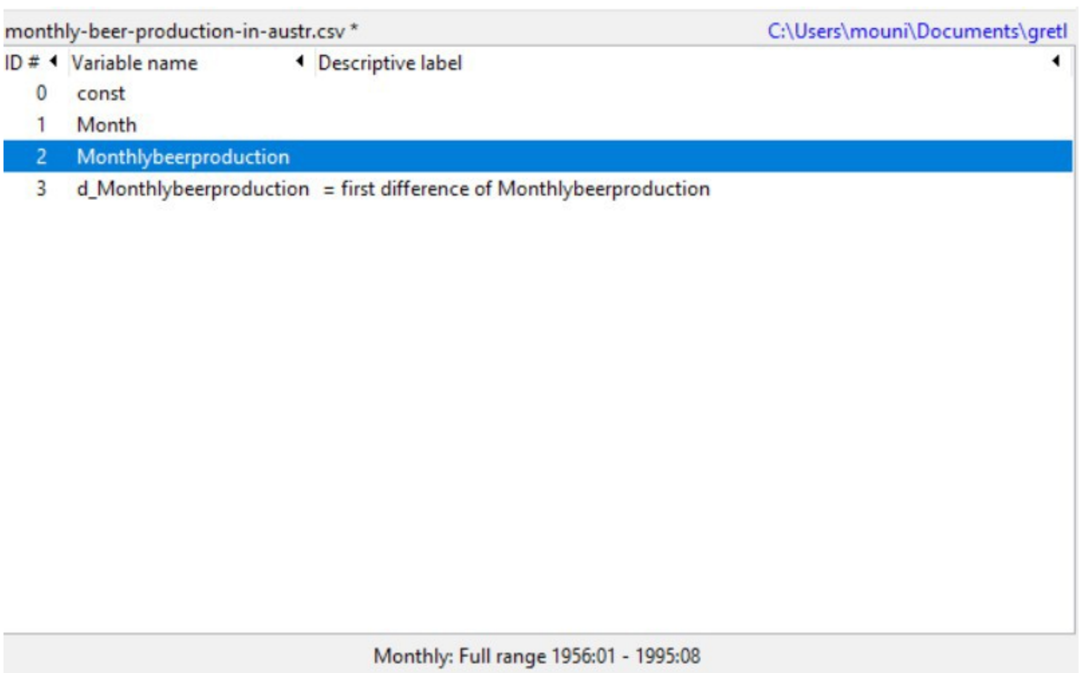
```
Augmented Dickey-Fuller test for Monthlybeerproduction
testing down from 17 lags, criterion AIC
sample size 458
unit-root null hypothesis: a = 1

  test with constant
  including 17 lags of (1-L)Monthlybeerproduction
  model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  estimated value of (a - 1): -0.0417511
  test statistic: tau_c(1) = -2.28266
  asymptotic p-value 0.1777
  1st-order autocorrelation coeff. for e: -0.002
  lagged differences: F(17, 439) = 61.983 [0.0000]
```

From the results above the P-value is 0.177 which is greater than the significance value 0.05 so we can not reject the H0 hypothesis and can conclude that there is a trend thus non-constant mean.

5. Summarize the results of steps 2 through 4 and tell what your decision is regarding constant mean in the time series data set.

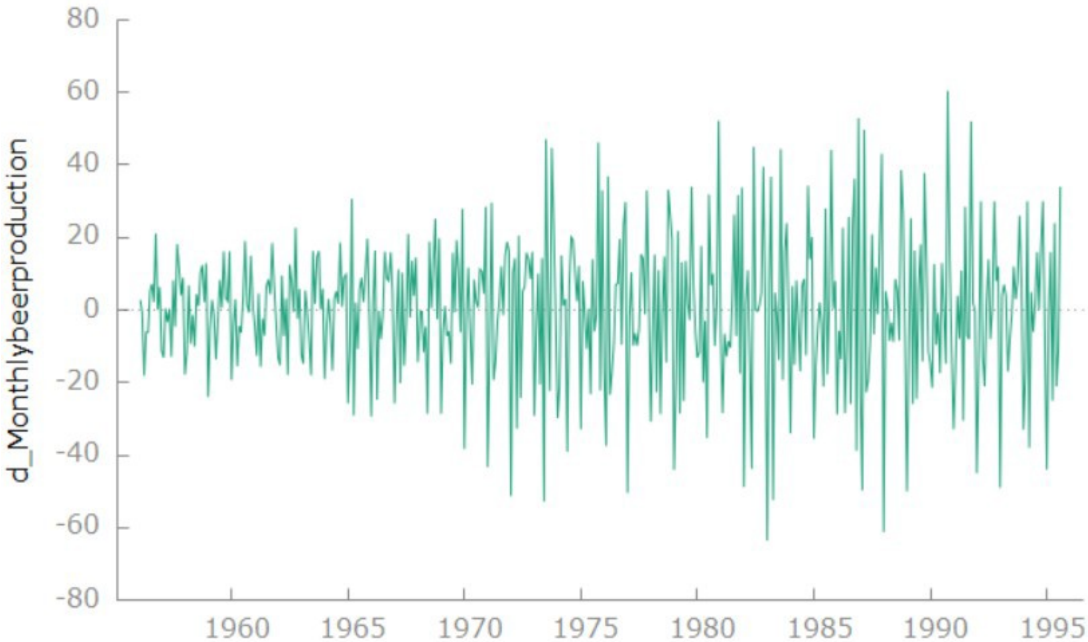
In summary, the Mark I Eyeball, the correlogram, the KPSS test and the ADF test all suggest there is a non-constant mean / there is a trend. Therefore, we perform differencing to find out if the trend recedes.



The d_Monthlybeerproduction is the differenced variable.

6. Review the decision in step #5. If the test suggests that there is a non-constant mean then use differencing to create a new differenced variable for the time series data set. Otherwise proceed to steps a through c below:

- Plot out the data for the new differenced data set. Tell me if it looks like the differencing got rid of the trend or non-constant mean.
  The plot for the new differenced dataset is showed below. The plot seems to have gotten rid of the trend or non-constant mean by Mark I Eyeball.



- Plot the ACF for the differenced time series. Tell me if this new ACF plot looks like there now is no trend.

  The ski slope behavior has reduced a little bit and there are flip flops in the plot. This indicates that there is no longer a trend / there is a constant mean. But there are other tests that can be performed to come to a strong conclusion

Autocorrelation function for d_Monthlybeerproduction
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0.5

LAG	ACF		PACF		Q-stat. [p-value]
1	-0.2277	***	-0.2277	***	24.7732 [0.000]
2	-0.1162	**	-0.1773	***	31.2464 [0.000]
3	0.0901	**	0.0203		35.1441 [0.000]
4	-0.1526	***	-0.1570	***	46.3512 [0.000]
5	0.0512		-0.0100		47.6173 [0.000]
6	-0.2277	***	-0.2942	***	72.6715 [0.000]
7	-0.1256	***	-0.2881	***	80.3147 [0.000]
8	0.0172		-0.2839	***	80.4575 [0.000]
9	0.0460		-0.1565	***	81.4844 [0.000]
10	-0.1930	***	-0.5333	***	99.6264 [0.000]
11	0.1097	**	-0.5877	***	105.5064 [0.000]
12	0.5536	***	0.0062		255.4582 [0.000]
13	-0.1359	***	-0.0949	**	264.5195 [0.000]
14	0.0653		0.0571		266.6177 [0.000]
15	-0.0356		0.1107	**	267.2421 [0.000]
16	-0.1575	***	0.0054		279.4902 [0.000]
17	0.1340	***	0.1039	**	288.3728 [0.000]
18	-0.3318	***	-0.0240		342.9431 [0.000]
19	-0.0471		0.1107	**	344.0445 [0.000]
20	0.0587		0.1281	***	345.7585 [0.000]

21	-0.0806	*	0.1783	***	349.0026	[0.000]
22	-0.1040	**	0.0360		354.4150	[0.000]
23	0.1269	***	-0.1951	***	362.4914	[0.000]
24	0.4297	***	0.0149		455.2735	[0.000]
25	-0.0412		-0.0247		456.1264	[0.000]
26	0.0738		0.0339		458.8730	[0.000]

- Apply the KPSS test and the ADF or ADF-GLS test to the differenced data – does the trend disappear?

KPSS Test:

H0: There is a constant mean / no trend.
Halt: There is a non-constant mean / there is a trend.

```
KPSS test for d_Monthlybeerproduction

T = 475
Lag truncation parameter = 5
Test statistic = 0.00921453

                   10%      5%      1%
Critical values: 0.348   0.462   0.742
P-value > .10
```

The p-value is > 0.10 which is greater than the significance level of 0.05. Hence, we do not have evidence to reject the null hypothesis and can conclude that there is a constant mean / no trend.

ADF Test:

H0: There is a non-constant mean / there is a trend.
Halt: There is a constant mean / no trend.

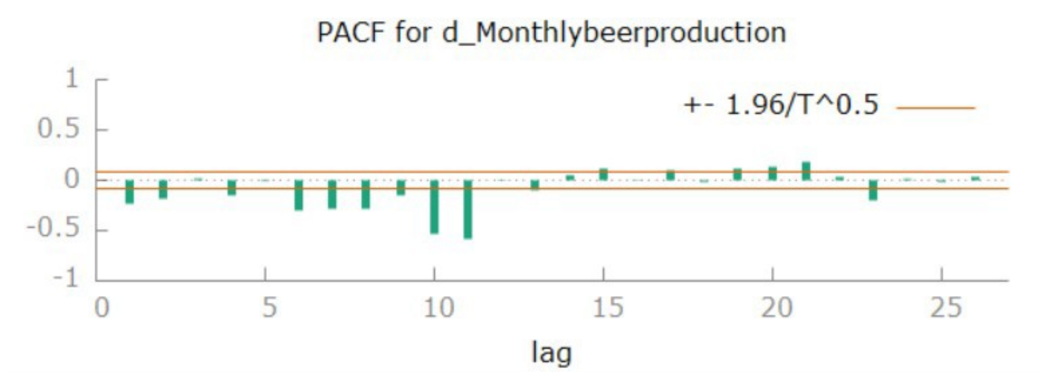
```
Augmented Dickey-Fuller test for d_Monthlybeerproduction
testing down from 17 lags, criterion AIC
sample size 458
unit-root null hypothesis: a = 1

  test with constant
  including 16 lags of (1-L)d_Monthlybeerproduction
  model: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  estimated value of (a - 1): -8.02706
  test statistic: tau_c(1) = -6.35707
  asymptotic p-value 1.581e-008
  1st-order autocorrelation coeff. for e: -0.003
  lagged differences: F(16, 440) = 69.354 [0.0000]
```

From the results above p- value is 1.581e-008 smaller than the significance value of 0.05. Hence, we reject the null hypothesis and can conclude that there is a constant mean / no trend.

Note: From this point onward through step 9, if the time series was differenced, use the differenced time series data set for all the rest of the questions. Otherwise you can use the undifferenced data set.

7. Plot the PACF for the time series data set. Using the combined information from the ACF you plotted earlier along with the information in the PACF, tell me if you see autoregressive and/or moving average processes in the data set. To help with interpretation you may want to refer to online resources – here is a decent resource from Duke University **https://people.duke.edu/~rnau/411arim3.htm** or Penn State https://onlinecourses.science.psu.edu/stat510/node/64



From the PACF plot, the green bars do not go above the red significance line (positive), hence we do not see any evidence of autoregressive processes. The green bars do go below the red significance line (negative), hence that suggests moving average processes at lags of order 1, 2, 4, 6,7,8,9 ,10, 11 and 23

8. For your time series data set, experiment with different ARIMA models for them. Try at least four models. As you try them, list out the results of the various models and
   a) Comment on how each one is working and compare it to the previous model using various metrics such as AIC, BIC (Schwartz) and Box Leung Most students end up creating a small table with these statistics across the models tried so it is easy to compare them.

MODEL 1 – ARIMA (1,0,1)

Model 1: ARMA, using observations 1956:02-1995:08 (T = 475)
Dependent variable: d_Monthlybeerproduction
Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*	
const	0.140377	0.0715253	1.963	0.0497	**
phi_1	0.480096	0.0428108	11.21	<0.0001	***
theta_1	−0.955166	0.0104243	−91.63	<0.0001	***

Mean dependent var	0.125895	S.D. dependent var	19.65357
Mean of innovations	−0.048053	S.D. of innovations	17.25440
R-squared	0.227631	Adjusted R-squared	0.225998
Log-likelihood	−2027.562	Akaike criterion	4063.124
Schwarz criterion	4079.778	Hannan-Quinn	4069.673

	Real	*Imaginary*	*Modulus*	*Frequency*
AR				
Root 1	2.0829	0.0000	2.0829	0.0000
MA				
Root 1	1.0469	0.0000	1.0469	0.0000

Model Interpretation:

Since the p-values for phi_1 is < 0.05 we can say that we found the first order auto regressive process in the model

Since the p-values for theta_1 is < 0.05 we can say that we found moving average in the model

So, our objective of moving average and auto regressive variables that are significant is met.

Adjusted $R^2 = 0.225998$
AIC = 4063.124
BIC = 4079.778
Hannan-Quinn = 4069.673

Ljung-Box:

H0: There is no evidence of serial autocorrelation in the residuals.
Halt:There is serial autocorrelation in the residuals.

Output:

Test for autocorrelation up to order 12

Ljung-Box Q' = 322.92,
with p-value = P(Chi-square(10) > 322.92) = 2.196e-063

p-value = 2.196e-063 which is < 0.05. Hence, we can reject the null hypothesis and can conclude that there is more variance left in the residuals that we might be able to find with some additional ARIMA models.

MODEL 2 – ARIMA (1,0,2)

Model 4: ARMA, using observations 1956:02-1995:08 (T = 475)
Dependent variable: d_Monthlybeerproduction
Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*	
const	0.140137	0.0710108	1.973	0.0484	**
phi_1	0.415720	0.0923283	4.503	<0.0001	***
theta_1	−0.872780	0.100489	−8.685	<0.0001	***
theta_2	−0.0770792	0.0932954	−0.8262	0.4087	

Mean dependent var	0.125895	S.D. dependent var	19.65357
Mean of innovations	−0.047565	S.D. of innovations	17.24043
R-squared	0.228882	Adjusted R-squared	0.225615
Log-likelihood	−2027.196	Akaike criterion	4064.391
Schwarz criterion	4085.208	Hannan-Quinn	4072.577

		Real	*Imaginary*	*Modulus*	*Frequency*
AR					
	Root 1	2.4055	0.0000	2.4055	0.0000
MA					
	Root 1	1.0486	0.0000	1.0486	0.0000
	Root 2	-12.3718	0.0000	12.3718	0.5000

Model Interpretation:

Since the p-values for phi_1 is < 0.05 we can say that we found the first order auto regressive process in the model

Since the p-values for theta_1 is < 0.05 we can say that we found moving average in the model
But for the p-value for theta_2 is not significant we do not find moving average in the model for second order

Adjusted R^2 = 0.225615
AIC = 4064.391
BIC =4085.208
Hannan-Quinn = 4072.577

The R-square value is decreased a bit in the second model

Ljung-Box Test:

H0: There is no evidence of serial autocorrelation in the residuals.
Halt:There is serial autocorrelation in the residuals.

Test for autocorrelation up to order 12

Ljung-Box Q' = 318.765,

with p-value = P(Chi-square(9) > 318.765) = 2.714e-063

p-value = 2.714e-063 which is < 0.05. Hence, we can reject the null hypothesis and can conclude that there is more variance left in the residuals that we might be able to find with some additional ARIMA models.

MODEL 3 – ARIMA (1,0,3)

Model 5: ARMA, using observations 1956:02-1995:08 (T = 475)
Dependent variable: d_Monthlybeerproduction
Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*	
const	0.140173	0.0709954	1.974	0.0483	**
phi_1	0.400257	0.117707	3.400	0.0007	***
theta_1	−0.860071	0.118101	−7.283	<0.0001	***
theta_2	−0.0728883	0.0973125	−0.7490	0.4538	
theta_3	−0.0155843	0.0682808	−0.2282	0.8195	

Mean dependent var	0.125895	S.D. dependent var	19.65357
Mean of innovations	−0.047382	S.D. of innovations	17.23952
R-squared	0.228971	Adjusted R-squared	0.224060
Log-likelihood	−2027.170	Akaike criterion	4066.340
Schwarz criterion	4091.320	Hannan-Quinn	4076.163

		Real	*Imaginary*	*Modulus*	*Frequency*
AR					
	Root 1	2.4984	0.0000	2.4984	0.0000
MA					
	Root 1	1.0486	0.0000	1.0486	0.0000
	Root 2	-2.8628	-7.2799	7.8225	-0.3096
	Root 3	-2.8628	7.2799	7.8225	0.3096

Model Interpretation:

Since the p-values for phi_1 is < 0.05 we can say that we found the first order auto regressive process in the model
Since the p-values for theta_1 is < 0.05 we can say that we found moving average in the model
But for the p-value for theta_2 and theta_3 is not significant we do not find moving average in the model for second order and third order

Adjusted R^2 = 0.224060
AIC = 4066.340
BIC = 4091.320
Hannan-Quinn = 4076.163

The R-square value is decreased a bit in the third model

Ljung-Box Test:

H0: There is no evidence of serial autocorrelation in the residuals.
Halt:There is serial autocorrelation in the residuals.

Output:

Test for autocorrelation up to order 12

Ljung-Box Q' = 318.069,

with p-value = P(Chi-square(8) > 318.069) = 5.844e-064

p-value = 5.844e-064 which is < 0.05. Hence, we can reject the null hypothesis and can conclude that there is more variance left in the residuals that we might be able to find with some additional ARIMA models.

MODEL 4 – ARIMA (2,0,1)

Model 7: ARMA, using observations 1956:02-1995:08 (T = 475)
Dependent variable: d_Monthlybeerproduction
Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*	
const	0.102318	0.650941	0.1572	0.8751	
phi_1	−0.967187	0.0803003	−12.04	<0.0001	***
phi_2	−0.317391	0.0435931	−7.281	<0.0001	***
theta_1	0.739956	0.0751460	9.847	<0.0001	***

Mean dependent var	0.125895	S.D. dependent var	19.65357
Mean of innovations	0.003281	S.D. of innovations	18.61698
R-squared	0.100814	Adjusted R-squared	0.097004
Log-likelihood	−2063.040	Akaike criterion	4136.080
Schwarz criterion	4156.897	Hannan-Quinn	4144.266

		Real	*Imaginary*	*Modulus*	*Frequency*
AR					
	Root 1	-1.5237	-0.9106	1.7750	-0.4143
	Root 2	-1.5237	0.9106	1.7750	0.4143
MA					
	Root 1	-1.3514	0.0000	1.3514	0.5000

Model Interpretation:

Since the p-values for phi_1 and phi_2 < 0.05 we can say that we found the first order auto regressive process in the model
Since the p-values for theta_1 is < 0.05 we can say that we found moving average in the model

Adjusted $R^2 = 0.097004$
AIC = 4136.080
BIC = 4156.897
Hannan-Quinn = 4144.266

The R-square value is increased slightly in the fifth model

Ljung-Box Test:

H0: There is no evidence of serial autocorrelation in the residuals.
Halt:There is serial autocorrelation in the residuals.

Output:

Test for autocorrelation up to order 12

Ljung-Box Q' = 286.714,

with p-value = P(Chi-square(9) > 286.714) = 1.712e-056

p-value = 1.712e-056 which is < 0.05. Hence, we can reject the null hypothesis and can conclude that there is more variance left in the residuals that we might be able to find with some additional ARIMA models.

MODEL 5 – ARIMA (1,0,6)

Model 10: ARMA, using observations 1956:02-1995:08 (T = 475)
Dependent variable: d_Monthlybeerproduction

Standard errors based on Hessian

	Coefficient	*Std. Error*	*z*	*p-value*	
const	0.119344	0.170227	0.7011	0.4832	
phi_1	0.560998	0.0647623	8.662	<0.0001	***
theta_1	−1.19640	0.0668474	−17.90	<0.0001	***
theta_2	−0.0161808	0.0833356	−0.1942	0.8460	
theta_3	0.346475	0.0767010	4.517	<0.0001	***
theta_4	−0.366460	0.0952916	−3.846	0.0001	***
theta_5	0.144122	0.0607482	2.372	0.0177	**
theta_6	0.191888	0.0617497	3.108	0.0019	***

Mean dependent var	0.125895	S.D. dependent var	19.65357
Mean of innovations	−0.056080	S.D. of innovations	15.82952
R-squared	0.359527	Adjusted R-squared	0.351316
Log-likelihood	−1989.095	Akaike criterion	3996.189
Schwarz criterion	4033.659	Hannan-Quinn	4010.924

		Real	*Imaginary*	*Modulus*	*Frequency*
AR					
	Root 1	1.7825	0.0000	1.7825	0.0000
MA					
	Root 1	0.9873	0.1875	1.0050	0.0299
	Root 2	0.9873	-0.1875	1.0050	-0.0299
	Root 3	-1.6942	-0.3178	1.7237	-0.4705
	Root 4	-1.6942	0.3178	1.7237	0.4705
	Root 5	0.3313	-1.2755	1.3178	-0.2095
	Root 6	0.3313	1.2755	1.3178	0.2095

Model Interpretation:

Since the p-values for phi_1 < 0.05 we can say that we found the first order auto regressive process in the model
Since the p-values for theta_1 , theta_3 , theta_4 , theta_5 , theta_6 are < 0.05 we can say that we found moving average in the model except for theta_2

Adjusted R^2 = 0.351316
AIC = 3996.189
BIC = 4033.659
Hannan-Quinn = 4010.924

The R-square value is increased when compared to the other models.

Ljung-Box Test:

H0: There is no evidence of serial autocorrelation in the residuals.
Halt:There is serial autocorrelation in the residuals.

Output:

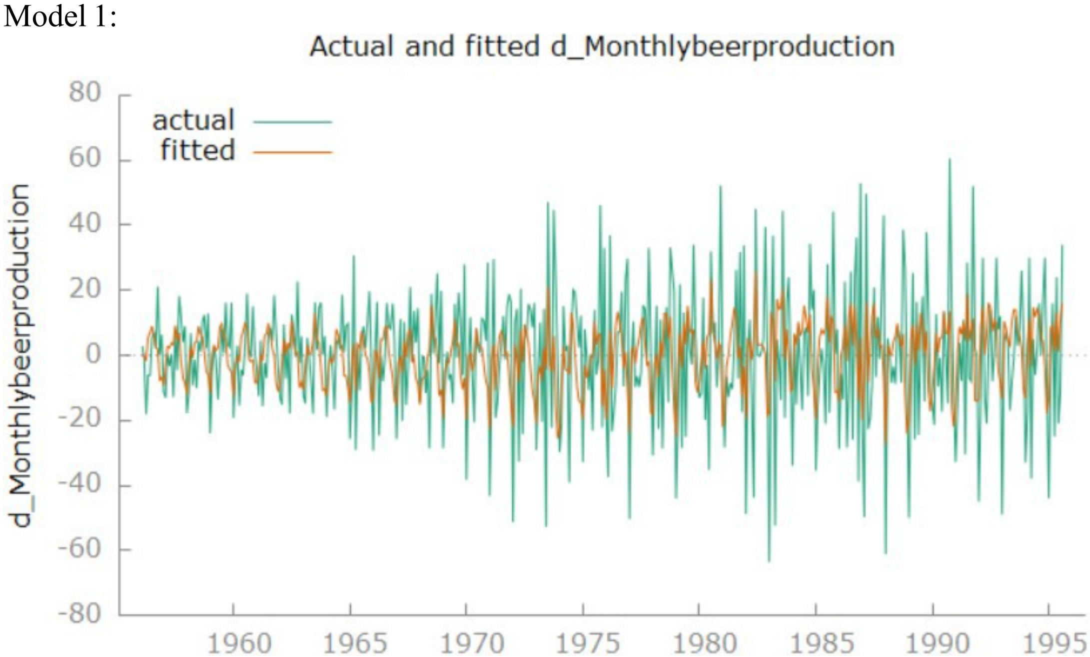
Test for autocorrelation up to order 12

Ljung-Box Q' = 246.704,

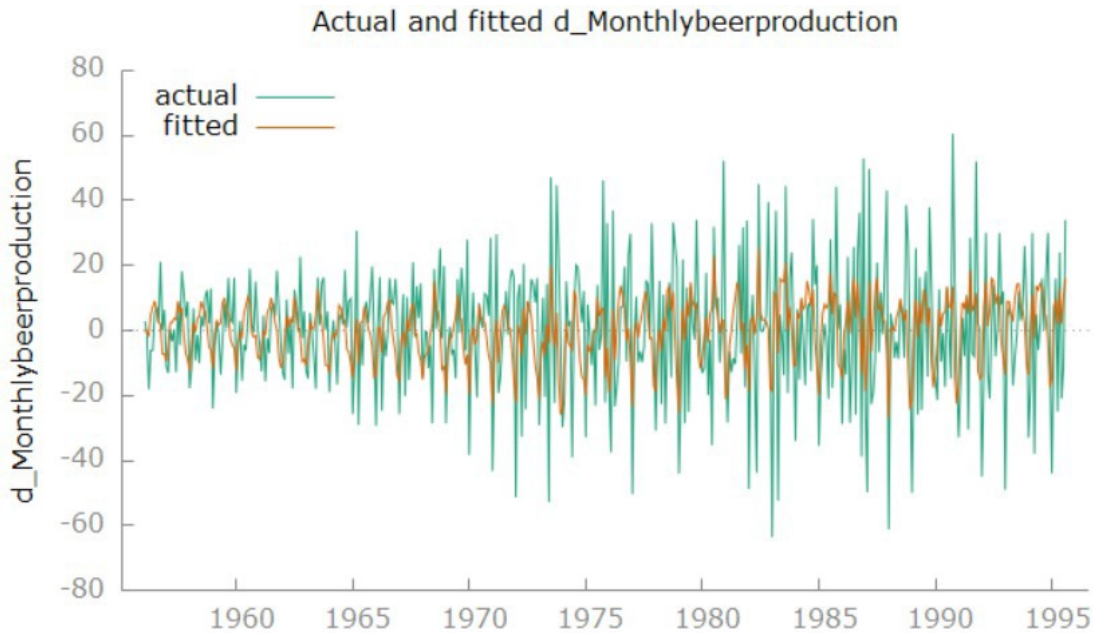
with p-value = P(Chi-square(5) > 246.704) = 2.801e-051

p-value = 2.801e-051 which is < 0.05. Hence, we can reject the null hypothesis and can conclude that there is more variance left in the residuals that we might be able to find with some additional ARIMA models.

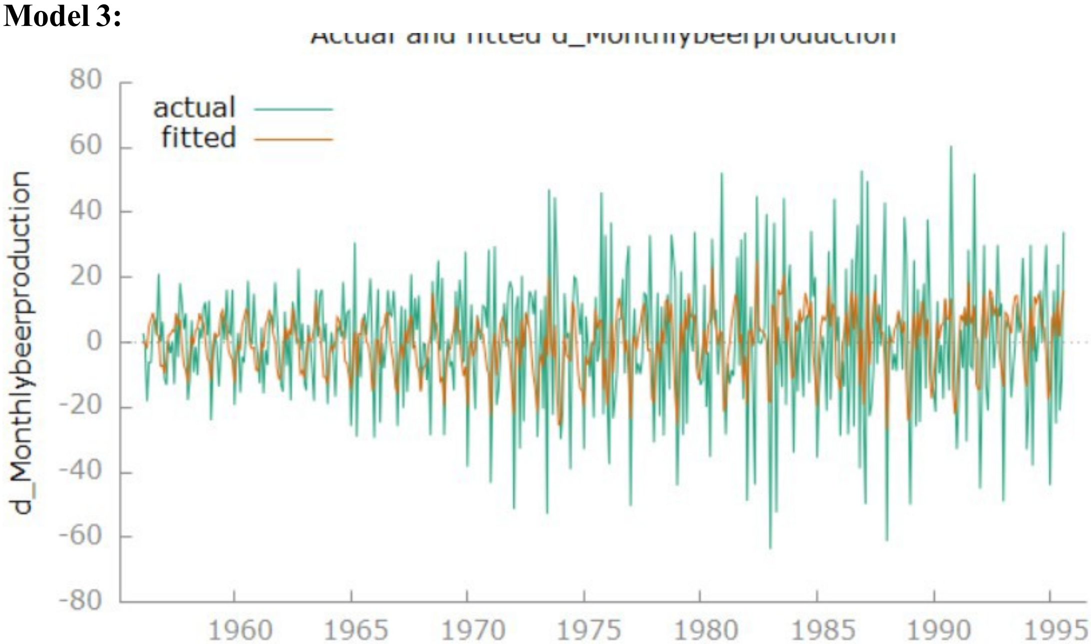
b. Plot the observed versus fitted data for the time series data set **for each model** and comment on how well the model seems to be working
Model 1:



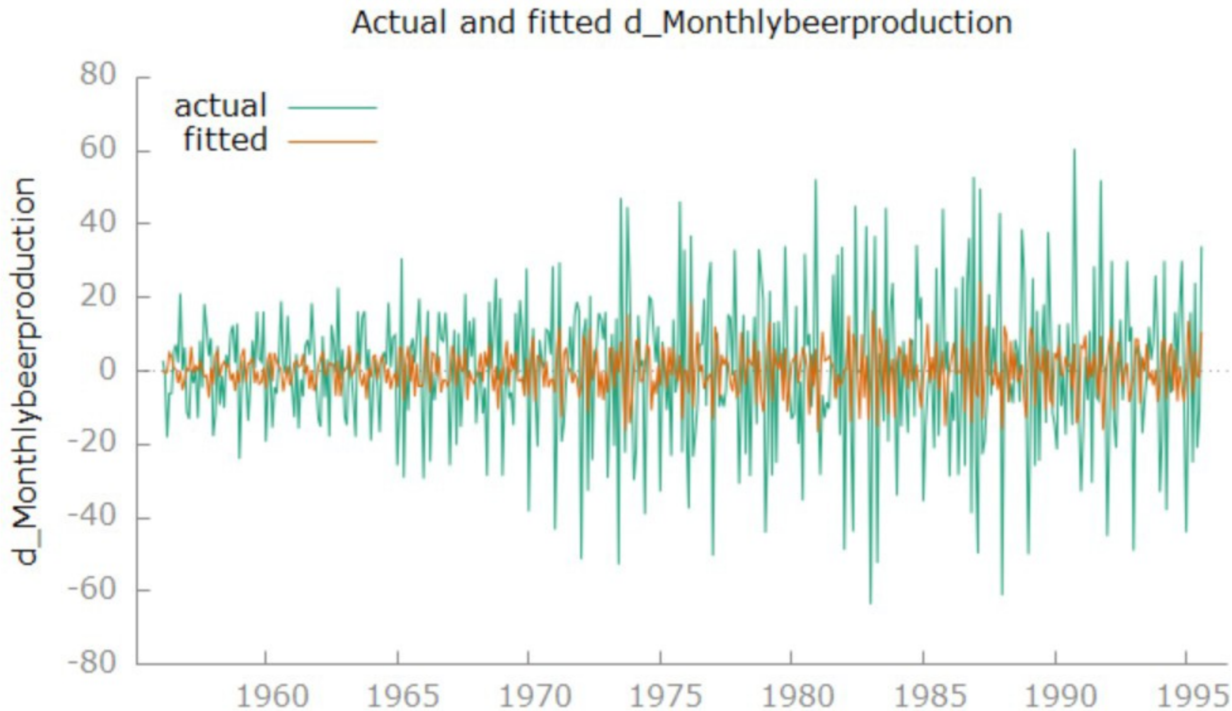
Model 2:



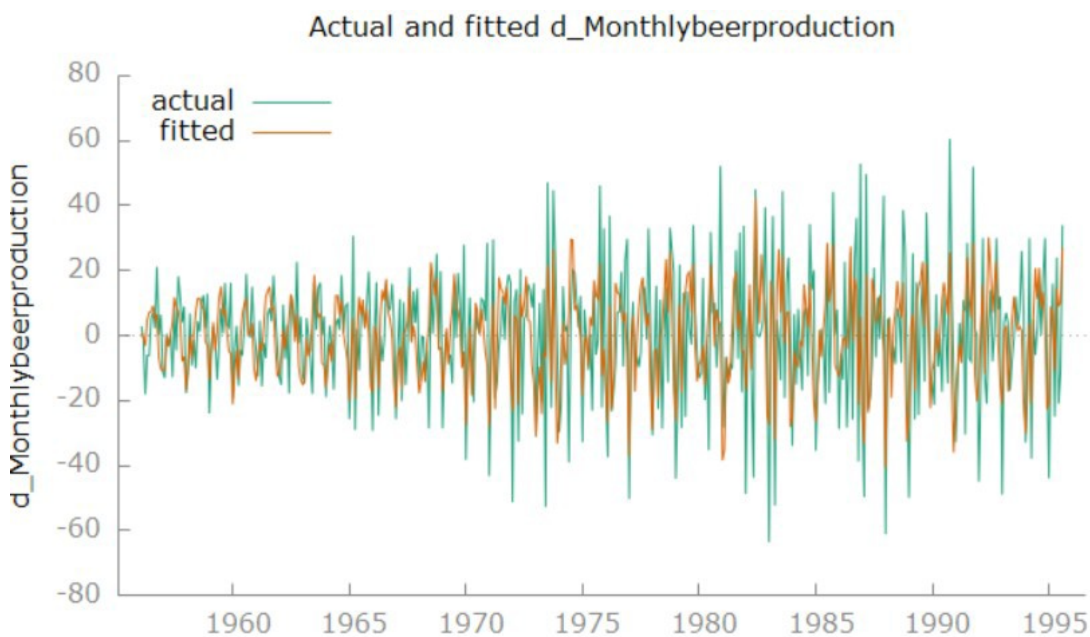
Model 3:



Model 4:



Model 5:



c. Pick one of the models as your favorite and tell me why you like that one the best.

Model	**Adj. R-square**	**AIC**	**BIC**	**Hannan-Quinn**	**Ljung-Box**
MODEL 1 – (1,0,1)	0.225998	4063.124	4079.778	4069.673	2.196e-063

MODEL 2 – (1,0,2)	0.225615	4064.391	4085.208	4072.577	2.714e-063
MODEL 3 – (1,0,3)	0.224060	4066.340	4091.320	4076.163	5.844e-064
MODEL 4 – (2,0,1)	0.097004	4136.080	4156.897	4144.266	1.712e-056
MODEL 5 – (1,0,6)	0.351316	3996.189	4033.659	4010.924	2.801e-051

I will choose the MODEL 5 – ARIMA (1,0,6) as the best model because the adjusted R^2 is the highest of all and the goodness-of-fit metrics are the lowest of all.

9. Test the time series data set for constant variance using the ARCH test (GRETL does this nicely). Tell me which ones might have issues with constant variance and so not be so nicely stationary. Note that we will not do anything about this issue for the moment, but it's good to know.

H0: There is no evidence of non-constant variance across time.
Halt: There is non-constant variance across time.

Test for ARCH of order 12

```
             coefficient    std. error   t-ratio    p-value
  ---------------------------------------------------------
  alpha(0)    108.973       35.3881       3.079    0.0022    ***
  alpha(1)    −0.0126051     0.0439170   −0.2870 0.7742
  alpha(2)    −0.0254559     0.0438394   −0.5807 0.5618
  alpha(3)     0.0131112     0.0439593    0.2983 0.7656
  alpha(4)     0.0222075     0.0439732    0.5050 0.6138
  alpha(5)    −0.0518275     0.0439757   −1.179    0.2392
  alpha(6)     0.182111      0.0438737    4.151    3.96e-05  ***
  alpha(7)     0.0851837     0.0438156    1.944    0.0525    *
  alpha(8)     0.00775876    0.0439507    0.1765 0.8600
  alpha(9)    −0.0308454     0.0439516   −0.7018 0.4832
  alpha(10)   −0.0270759     0.0440376   −0.6148 0.5390
  alpha(11)    0.0548225     0.0440510    1.245    0.2140
  alpha(12)    0.364961      0.0441212    8.272    1.51e-015 ***

  Null hypothesis: no ARCH effect is present
  Test statistic: LM = 100.066
  with p-value = P(Chi-square(12) > 100.066) = 5.40413e-016
```

The p-value is 5.40413e-016 which is less than the significance level of 0.05. Hence, we hence we can reject the H0 and conclude that there is a non-constant variance. So we do the GARCH Test now.