MScFE 610 Econometrics (C19-S4)

Group Work Assignment Submission 1 M3

Kwadwo Amo-Addai (kwadwoamoad@gmail.com)

David Sabater Dinter (david.sabater@gmail.com)

Ruben Ohayon (rubensacha.ohayon@gmail.com)

Andrea Chello (chelloandrea@gmail.com)

Pasin Marupanthorn (oporkabbb@hotmail.com)

Basic Analysis

Download JP Morgan stock historical prices from Yahoo Finance
Period: February 1, 2018 – December 30, 2018
Frequency: Daily
Price considered in the analysis: Close price adjusted for dividends and splits

Calculate in R:

- Average stock value
- Stock volatility
- Daily stock return

```
install.packages("quantmod")
library(quantmod)

getSymbols("JPM", src = "yahoo", from = "2018-02-01", to = "2018-12-30")
df <- data.frame(JPM); jpm_adjusted <- df$JPM.Adjusted

plot(jpm_adjusted, main = "JP Morgan Adjusted Close", lwd = 2, col = " red")
chartSeries(JPM, type = "line", subset = "2018", theme = chartTheme("white"))


# 1.1. Get Average Stock Value

mean(jpm_adjusted)

# 1.2. Get Stock Volatility

## 231 because 252 are the normal trading days in a year but we lose one month
## as we start the analysis on February

log_returns_jpm <- periodReturn(JPM, period = "daily")
annualized_volatility_jpm <- sd(log_returns_jpm) * sqrt(231) * 100
daily_volatility_jpm <- sd(log_returns_jpm)
paste("Annualized Volatility -> ", annualized_volatility_jpm)
paste("Daily Volatility -> ", daily_volatility_jpm)

# 1.3. Get Daily Stock Return

daily_returns_jpm <- dailyReturn(JPM)
print(daily_returns_jpm)
```

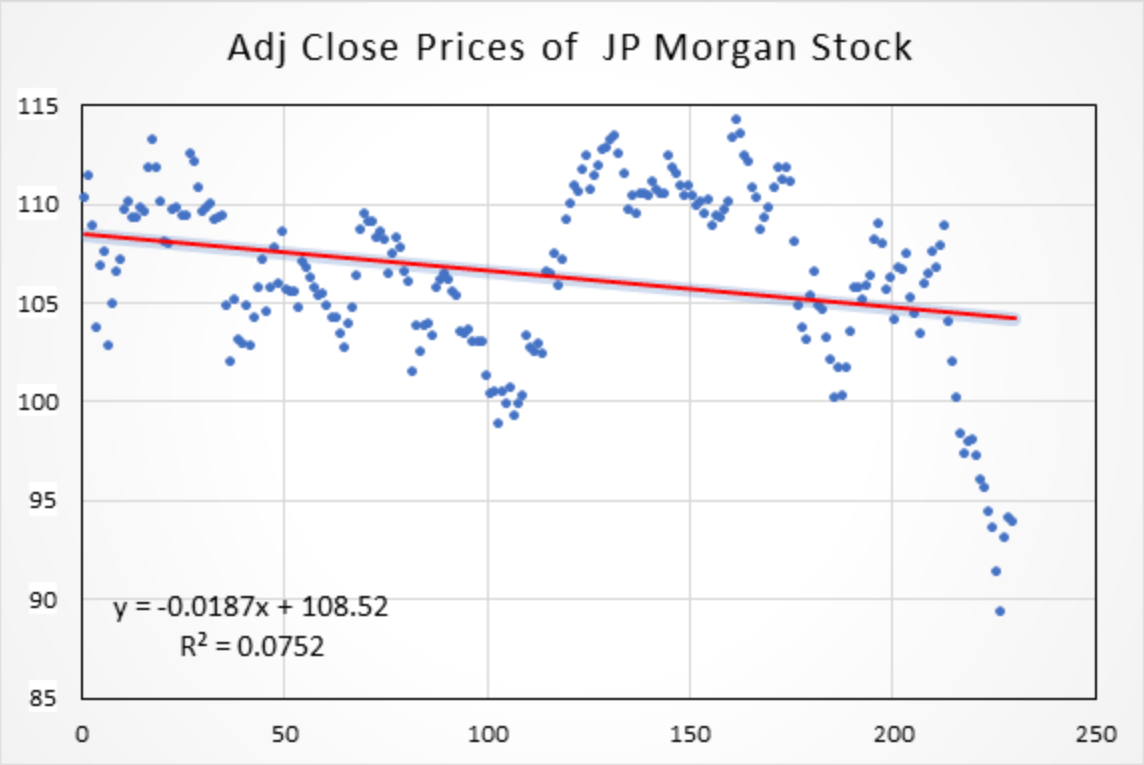
Basic Statistics by Excel

Let xi be the stock prices for i th day, r is the log return and be the number of days. The formulas and values of the given measures are shown in the following table.

	Average stock value	**Daily stock log return**	**Stock Volatility**
Formula	$xbar = \Sigma x/N$	$r = log(x_i/x_{i-1})$	$s = \sqrt{\sum_{i=1}^{N} \frac{(x_i - xbar)^2}{N-1}}$
Formula in Excel	AVERAGE	Log(cell I th/cell (i-1) th)	STDEV.S(Log return)
Value	106.3664348 $	-4.53737E-05 (Average)	0.006256267

The average of stock prices is 106.37 $ with the standard deviation of log returns or volatility is 0.006. The stock has relatively low volatility. Therefore, it is not surprising that the average return is low (approach to zero).

The scatter plot, trendline with the coefficient of determine are shown in the following figure. As can be seen in the figure, the slope of the regression is negative thus the downward trend line is shown. In the short future, the stock price is predicted to drop on average.



Linear Regression

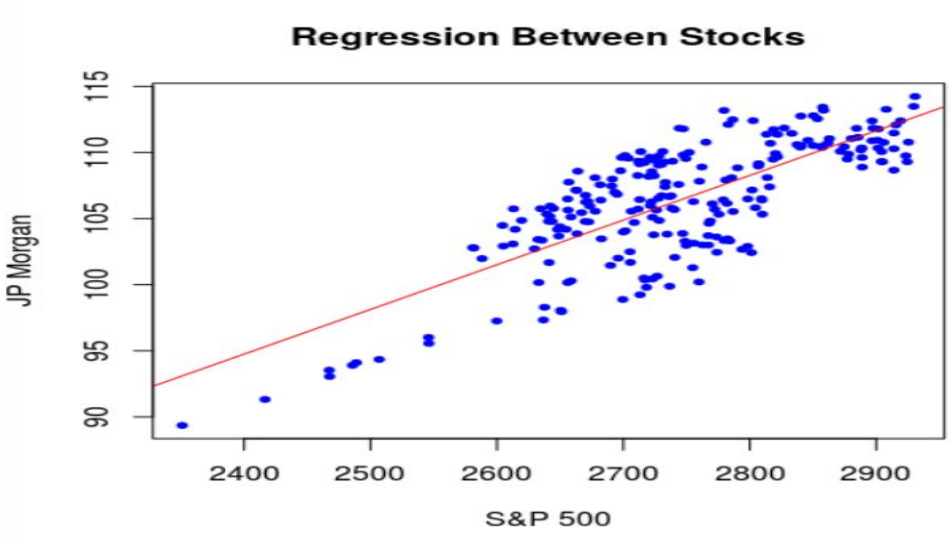
Implement a two-variable regression in R

Explained variable: JP Morgan stock (adjusted close price)
Explanatory variable: S&P500
Period: February 1, 2018 – December 30, 2018
Frequency: Daily

Regression in R

```
getSymbols("^GSPC", src = "yahoo", from = "2018-02-01", to = "2018-12-30")

df <- data.frame(GSPC); head(df)
```

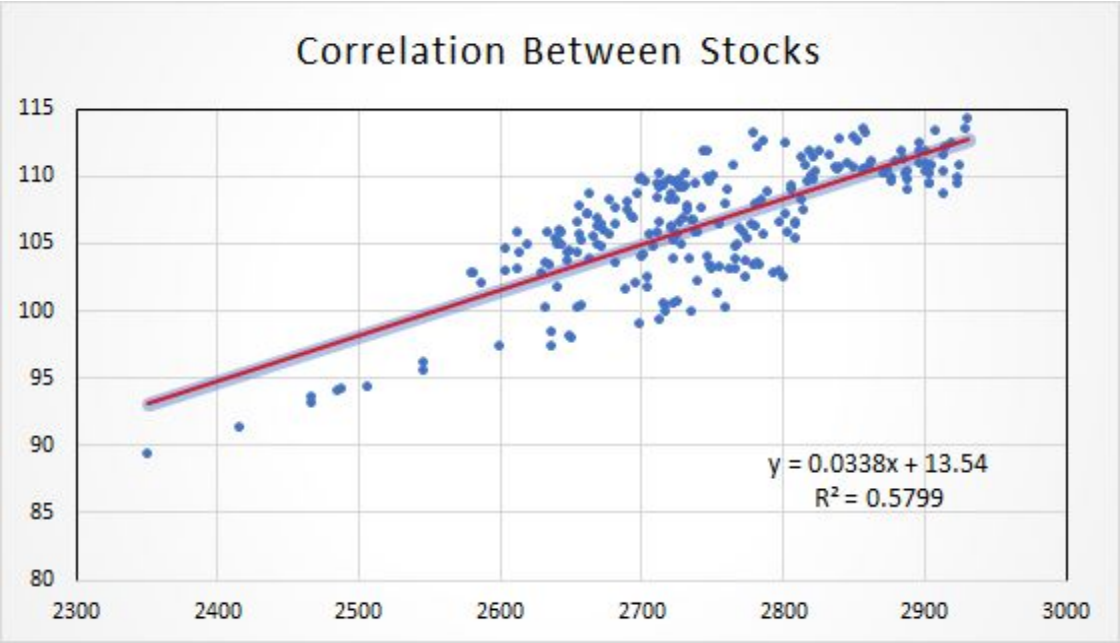
```
# Get the Adjusted Close Prices

sp500_adjusted <- df$GSPC.Adjusted; head(sp500_adjusted)


# Regression Plot

plot(y = jpm_adjusted, x = sp500_adjusted, main ="Regression Between Stocks", col = "blue",

    pch = 20, xlab = "S&P 500", ylab = "JP Morgan")

abline(lm(jpm_adjusted ~ sp500_adjusted), col = "red")


# Regression Summary

model <- lm(formula = jpm_adjusted ~ sp500_adjusted)

summary(model)
```



```
Call:
lm(formula = jpm_adjusted[, 1] ~ sp500_adjusted[, 1])

Residuals:
    Min      1Q  Median      3Q     Max
-6.7015 -2.3783  0.4797  2.3649  5.6034

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        13.641947   5.252654   2.597     0.01 *
sp500_adjusted[, 1]  0.033795   0.001913  17.662   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.947 on 227 degrees of freedom
Multiple R-squared:  0.5788,    Adjusted R-squared:  0.5769
F-statistic: 311.9 on 1 and 227 DF,  p-value: < 2.2e-16
```

Regression by Excel

We use the stock prices of the JP Morgan as the dependent variable (Y) and those of S&P500 as the independent variable (X). The LINEST function in Excel was applied to calculate the slope and the constant for the regression. The slope of the line is 0.033833625 and the constant is 13.54008218. We also investigate the correlation between two stock prices via plotting in the following figure.



It is clear that both stocks have the linear correlation (to be specific, this correlation is called concordance not linear). The positive slope refers to similar direction correlation (either both increase or decrease).

Next, the add-in package in excel called Analysis ToolPak is applied to the same data sets. It provides the following table.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.760791
R Square	0.578803
Adjusted R Square	0.576947
Standard Error	2.946659
Observations	229

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2708.51	2708.51	311.9397	1.66E-44

Residual	227	1970.996	8.682801
Total	228	4679.506	

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	13.64195	5.252654	2.597153	0.010014	3.291753	23.99214	3.291753	23.99214
	0.033795	0.001913	17.66181	1.66E-44	0.030024	0.037565	0.030024	0.037565

Regression Statistics

Multiple R. It is the Correlation Coefficient that measures the strength of a linear relationship between two variables. Note here that it is the square root of R (Coefficient of Determination). It ranges from -1 to 1. The absolute value of Multiple R approached 1 referring to strongly linear relation between variables while it approached 0 referring uncorrelated between variables. The negative sign indicates the opposite direction correlation while the positive sign indicates similar direction correlation.

R Square. It is the Coefficient of Determination, which is used as an indicator of the goodness of fit. It approaches to 1 referring to perfect fitting by linear function. While it goes to zero referring to inappropriate fitting by a linear model.

Adjusted R Square. It is used for multiple regression analysis instead of R square. In other words, it is the R square adjusted for the number of independent variables in the model.

Standard Error. It is an average distance (How far) between predicted value obtained by the model and the given data. The lower number indicates more accurate prediction.

Observations. It is the number of data used in the model.

ANOVA table

df is the degrees of freedom, **SS** is the sum of squares, **MS** is the mean square and **F** is the F statistic of the overall significance is that the fit of the regression model that contains no predictors and the linear model are equal. According to the hypothesis, if the p- value for the F-test is less than the given significance level, that means the prediction linear model provides a better fit than the regression model that contains no predictors **Significance F** is the P-value of F. If Significance F is less than 0.05 (5%), the model is appropriate. In contrast, it is greater than 0.05, the independent variable is not appropriate to the model.

Regression Table

Coefficients is the constant b in the linear model y = ax + b (first row) and it is the slope a of the linear model (second row). **Standard Error** is the error for testing. **Lower and Upper Bounds** refer to the confidence interval of the estimation.

Univariate Time Series Analysis

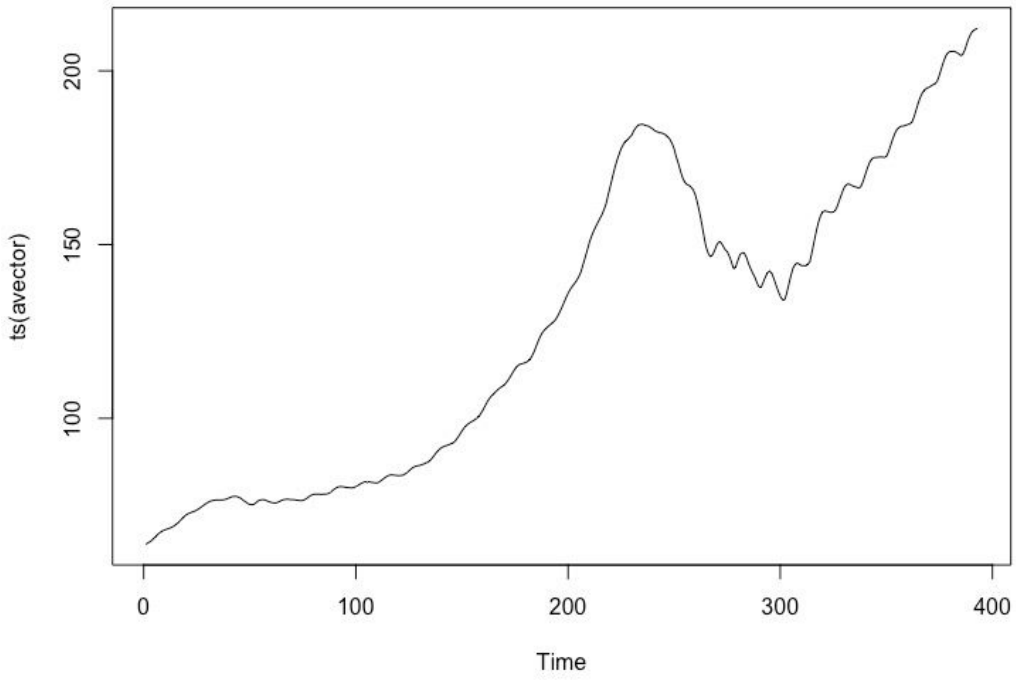
Forecast S&P/Case-Shiller U.S. National Home Price Index using an ARMA model.

Data source: https://fred.stlouisfed.org/series/CSUSHPINSA
Period considered in the analysis: January 1978 – latest data
Frequency: monthly data

1. Implement the Augmented Dickey-Fuller Test for checking the existence of a unit root in Case-Shiller Index series

```
library(tseries)
csv_path = "/Users/Ruben/Desktop/ADF/CSUSHPINSA.csv"
df <- read.table(csv_path, sep = ',', header=TRUE)
avector <- as.numeric(unlist(df["CSUSHPINSA"]))
plot(ts(avector))
adf.test(avector)
```

Cash Shiller Index



ADF

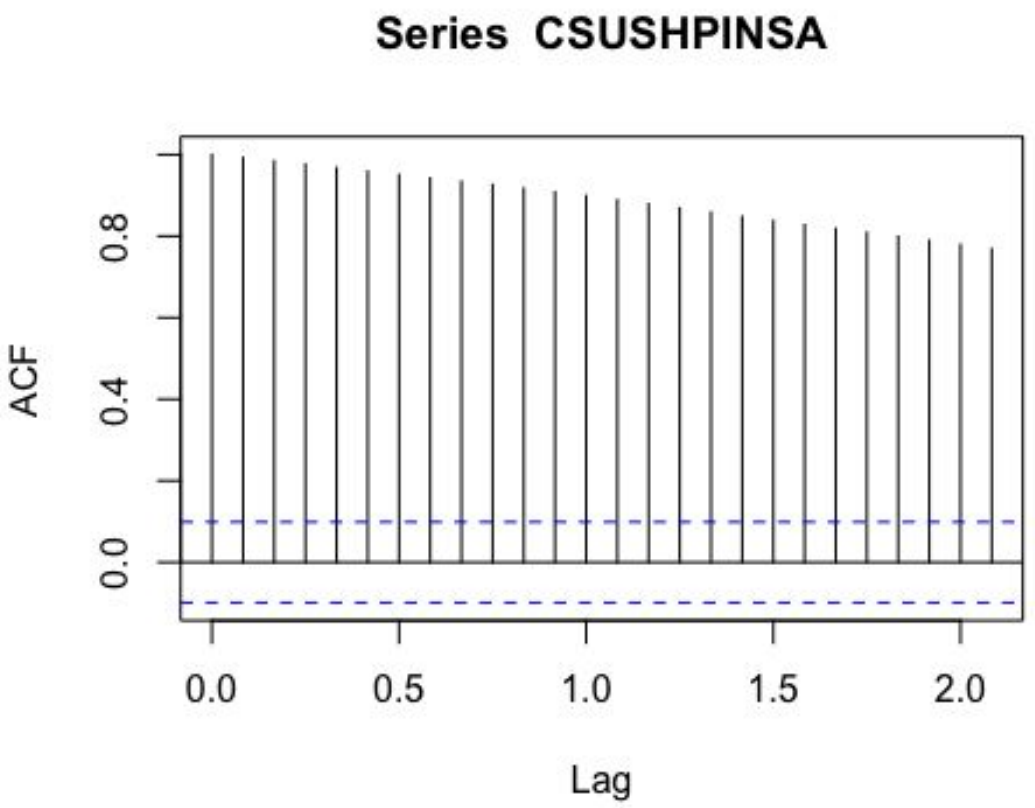
Results

```
        Augmented Dickey-Fuller Test

data:  avector
Dickey-Fuller = -2.2758, Lag order = 7, p-value =
0.4607
alternative hypothesis: stationary
```

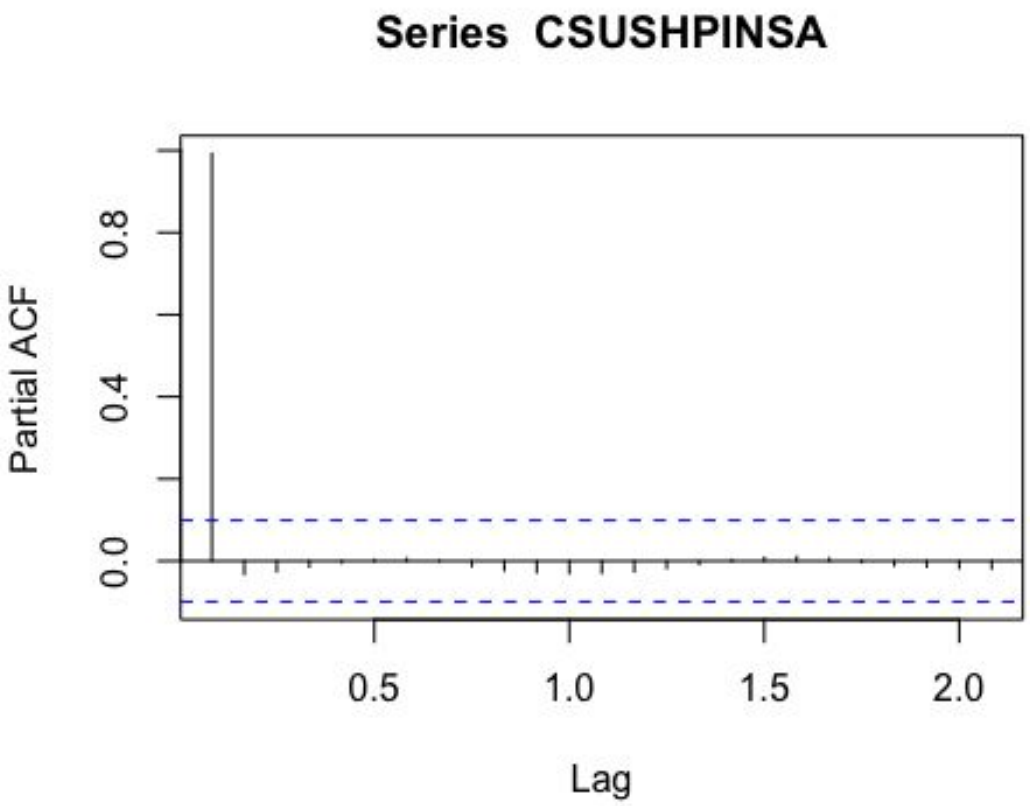
In the test output above, Dickey-Fuller is the test statistic. The more negative the number, the higher the p-value significance and hence the more we want to reject the null hypothesis. Here the p-value is displayed as 0.46 which is really high. It's not lower than 0.05 (95% level). Assuming significance α=0.05, so at the 95 percent level the null hypothesis of a unit root will be rejected.
The Cash Shriller Index is likely not stationary.

2. Implement an ARIMA(p,d,q) model. Determine p, d, q using Information Criterion or Box-Jenkins methodology. Comment results
   First step of Box-Jenkins method is to determine p and q orders by using the correlogram and the partial correlogram.

```
acf(CSUSHPINSA)
```

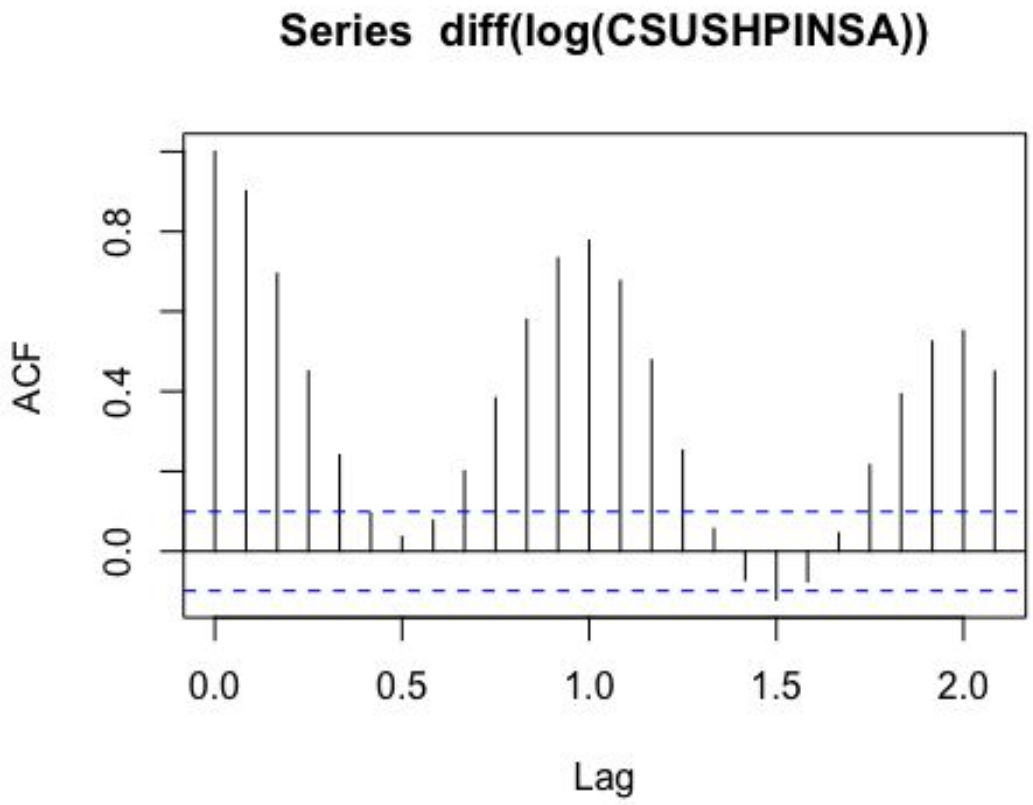



```
pacf(CSUSHPINSA)
```

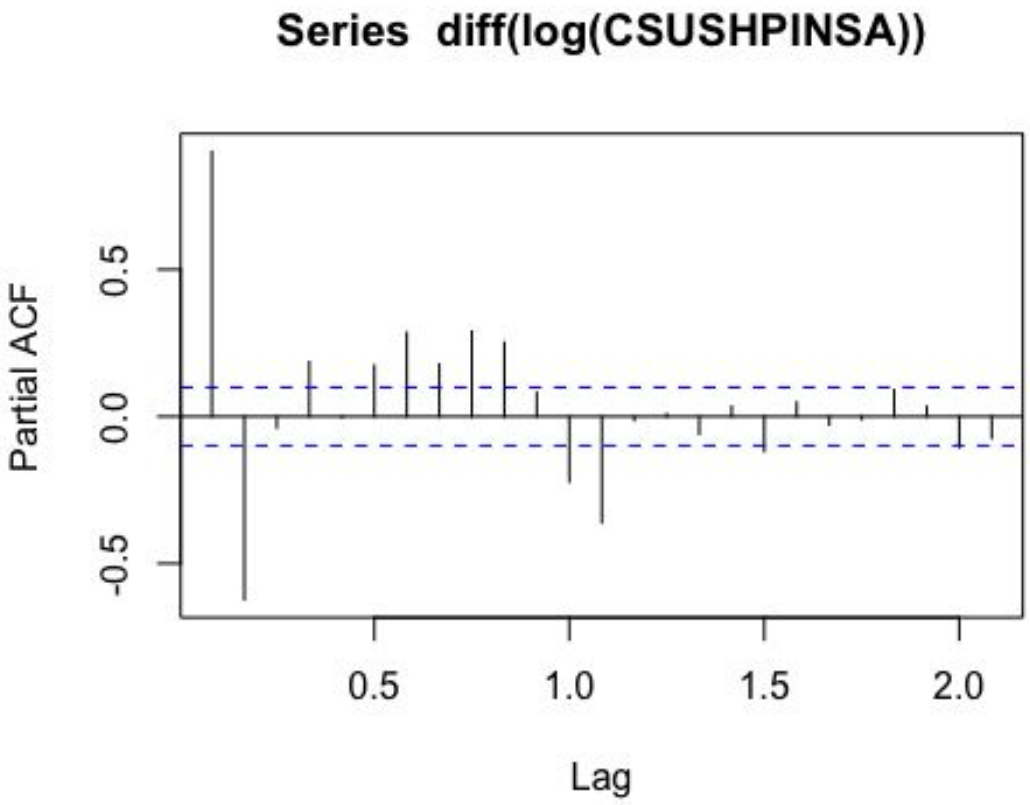


We can see that PACF displays a sharp cut-off while ACF decays more slowly (i.e., has significant spikes at higher lags), we say that the series displays an “AR signature”. Usually, a large number of significant auto-correlations or partial auto-correlations indicates non-stationarity and a need for further differencing.
The lag at which the PACF cuts off is the indicated number of AR terms, 0 in this case. We know that we need to address two issues before we test stationary series, as indicated already in previous question. One, we need to remove unequal variances. We do this using log of the series. Two, we need to address the trend component. We do this by taking the difference of the series. Now, let’s test the resultant series using the diff() and log() of the time series.

```
acf(diff(log(CSUSHPINSA)))
```




```
pacf(diff(log(CSUSHPINSA)))
```



We can see now what is called a damp sine wave, ACF and PACF graph may also suggest ARIMA(0,1,0), a.k.a., ARMA(0,0). Let's fit the model with this orders and using log() and diff() transformations.

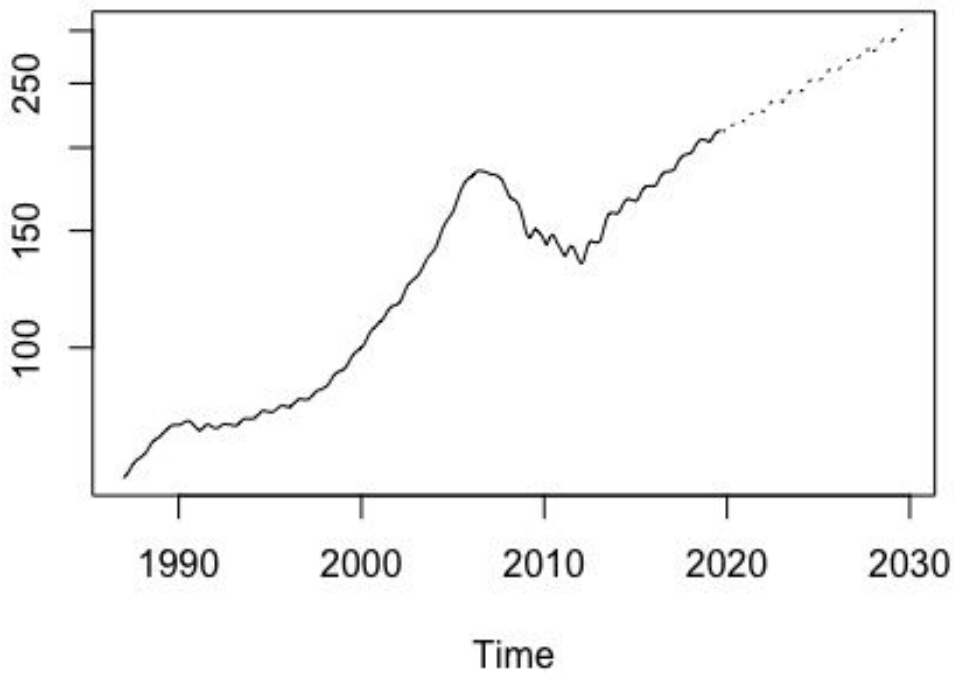
After some iterations we selected ARMA(0,1) as it showed the least AIC (-3350.91).

```
(fit <- arima(log(CSUSHPINSA), order = c(0, 1, 1),seasonal = list(order = c(0, 1, 1),
period = 12)))

##

## Call:

## arima(x = log(CSUSHPINSA), order = c(0, 1, 1), seasonal = list(order = c(0,

##     1, 1), period = 12))

##
```

```
## Coefficients:
##          ma1     sma1
##       0.7045  -0.1320
## s.e.  0.0291   0.0598
##
## sigma^2 estimated as 8.51e-06:  log likelihood = 1678.46,  aic = -3350.91
```

3. Forecast the future evolution of Case-Shiller Index using the ARMA model. Test model using in-sample forecasts

```
pred <- predict(fit, n.ahead = 10*12)
ts.plot(CSUSHPINSA,2.718^pred$pred, log = "y", lty = c(1,3))
```



```r
Box.test(fit$residuals, lag = 12)
```

```
##
##  Box-Pierce test
##
## data:  fit$residuals
## X-squared = 562.36, df = 12, p-value < 2.2e-16
```

4. Suggest exogenous variables that can improve forecasts

You can use a macro factor such as the Effective Federal Funds Rate (https://fred.stlouisfed.org/series/FEDFUNDS) which give a general overview of the US bank treasury. It can be statistically significant when we want to forecast an economic US index such as Real Estate.