Assignment 2

Hoyu (Ariel) Li

4/19/2020

```r
datapath <- "/Users/arielsmac/Desktop/Spring20/TimeSeries/Assignment2/"
x <- read.csv(paste(datapath,'data_akbilgic.csv',sep = '/'), header=TRUE)
library(tseries)
```

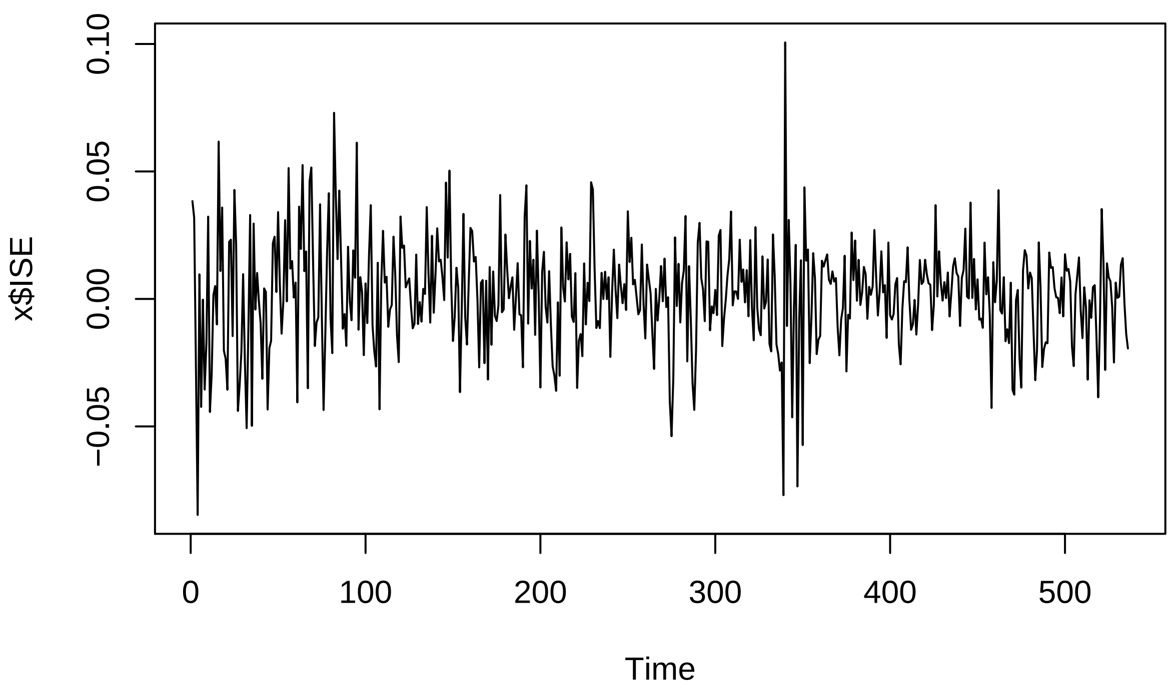
```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

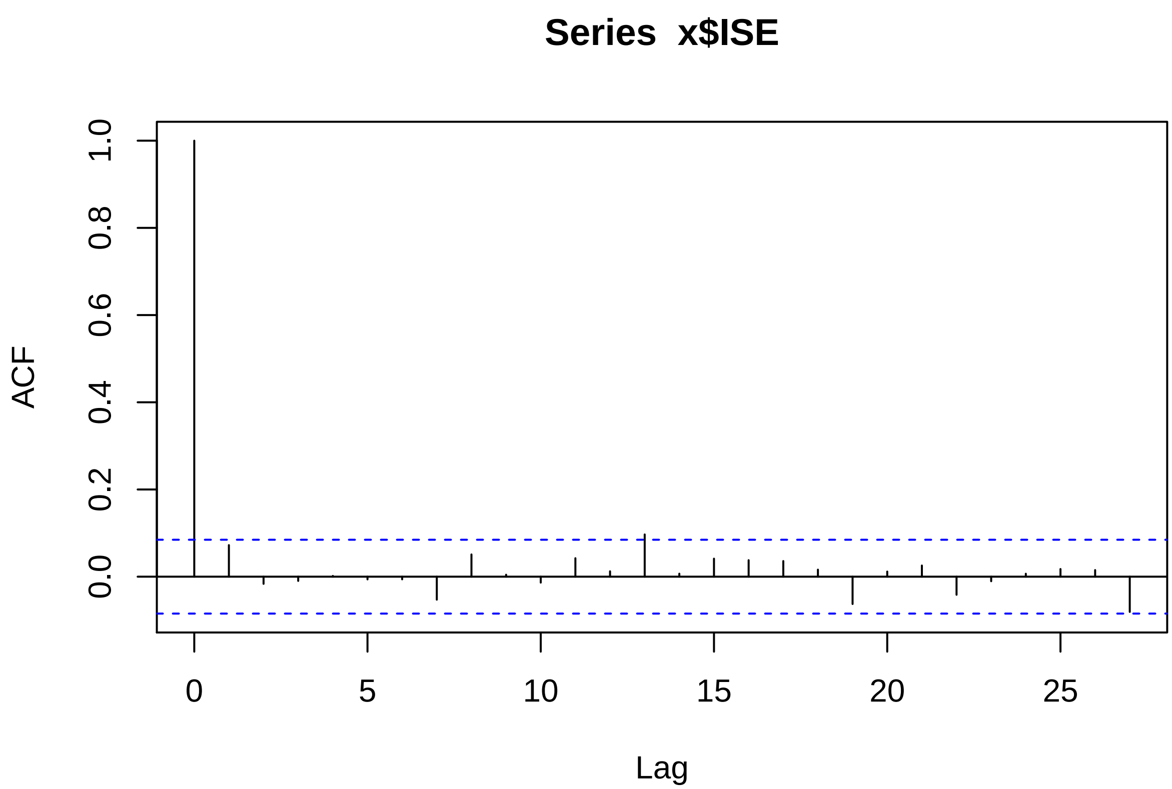
1) Determine if all the TS are stationary qualitatively and quantitatively

ISE

```r
plot.ts(x$ISE)
```




```r
acf(x$ISE)
```




```r
adf.test(x$ISE)
```

```
## Warning in adf.test(x$ISE): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  x$ISE
## Dickey-Fuller = -7.9492, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(x$ISE)
```

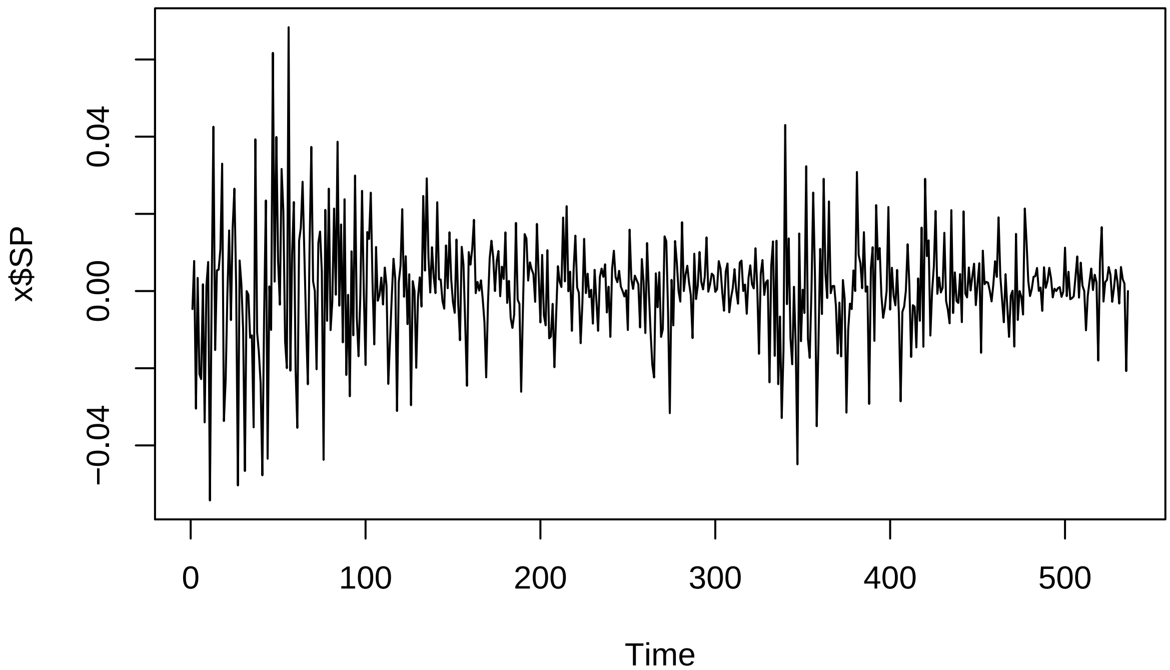
```
## Warning in kpss.test(x$ISE): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  x$ISE
## KPSS Level = 0.1967, Truncation lag parameter = 6, p-value = 0.1
```

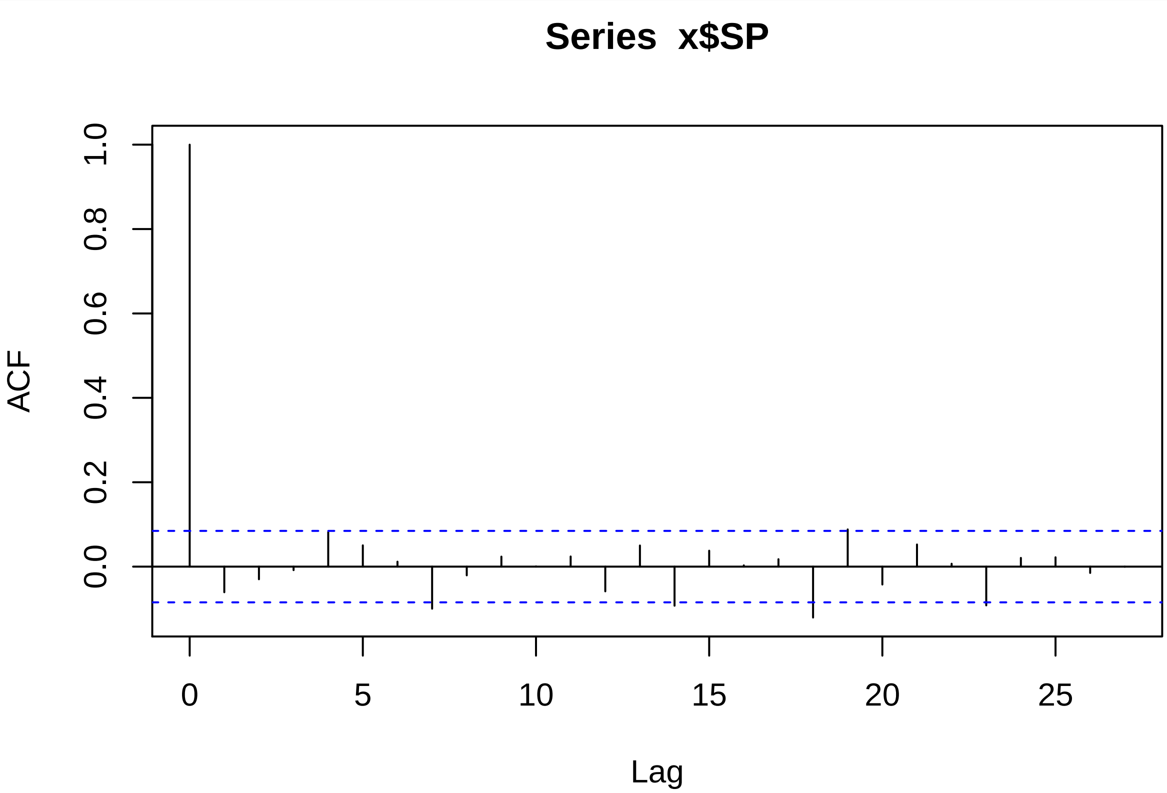
The ISE seems to be stationary in the mean. The ACF dies down quickly and follows a sinusoidal pattern about 0, which is typical of stationary time series. The p-value from the KPSS test is greater than 0.05, thus we fail to reject the null hypothesis of stationarity. From the ADF test, since the p-value is less than 0.05, we reject the null hypothesis of non-stationarity. The ADF and KPSS tests suggest that ISE series is stationary.

SP

```r
plot.ts(x$SP)
```




```
acf(x$SP)
```




```
adf.test(x$SP)
```

```
## Warning in adf.test(x$SP): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
```

```
## data:  x$SP
## Dickey-Fuller = -8.1353, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(x$SP)
```

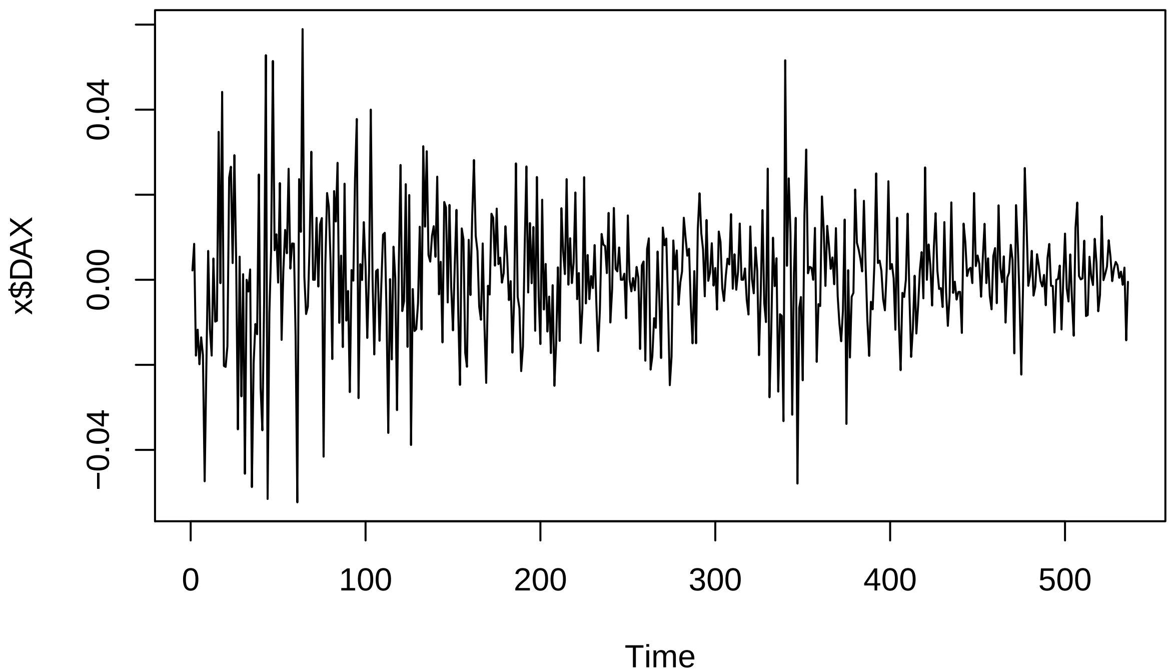
```
## Warning in kpss.test(x$SP): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  x$SP
## KPSS Level = 0.08007, Truncation lag parameter = 6, p-value = 0.1
```

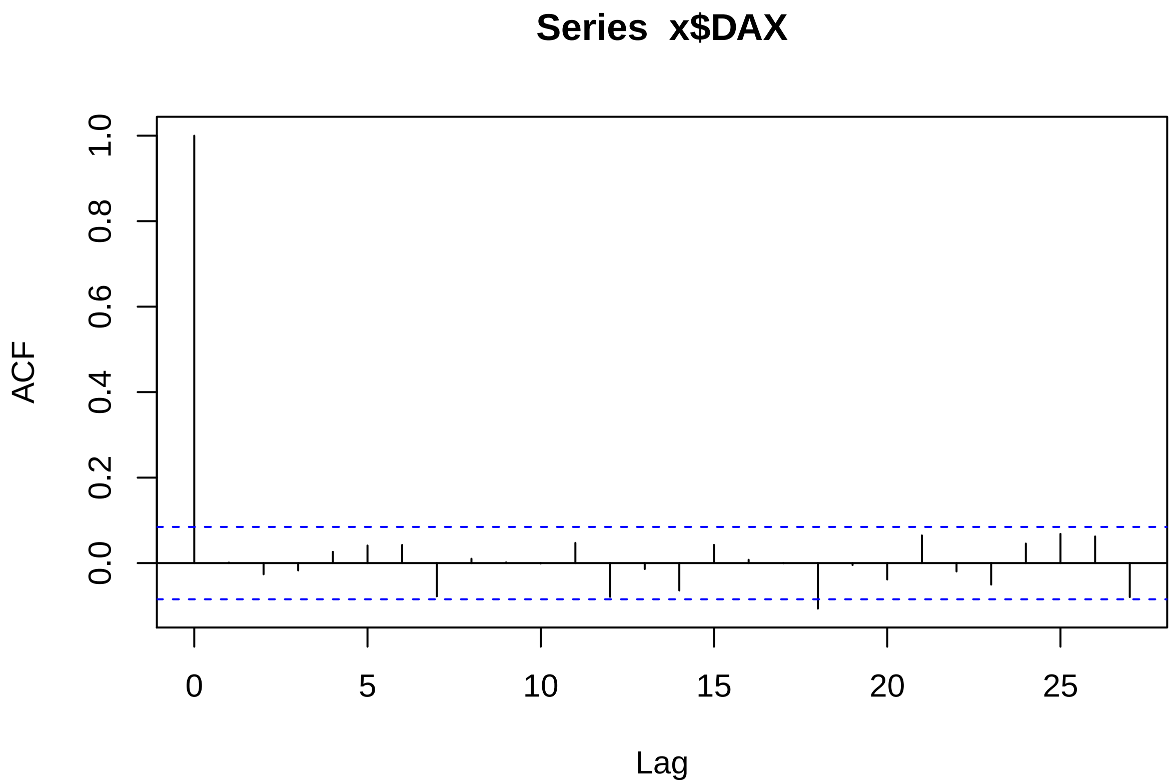
The SP seems to be stationary in the mean. The ACF dies down quickly and follows a sinusoidal pattern about 0, which is typical of stationary time series. The p-value from the KPSS test is greater than 0.05, thus we fail to reject the null hypothesis of stationarity. From the ADF test, since the p-value is less than 0.05, we reject the null hypothesis of non-stationarity. The ADF and KPSS tests suggest that SP series is stationary.

DAX

```r
plot.ts(x$DAX)
```




```r
acf(x$DAX)
```




```
adf.test(x$DAX)
```

```
## Warning in adf.test(x$DAX): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  x$DAX
## Dickey-Fuller = -8.147, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(x$DAX)
```

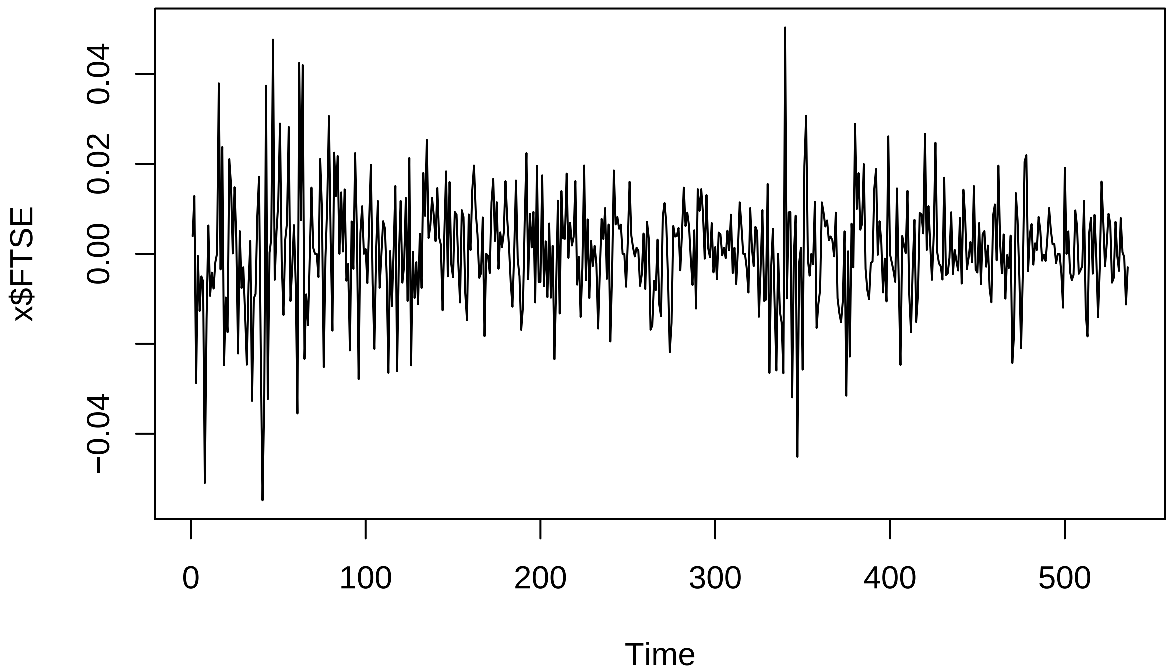
```
## Warning in kpss.test(x$DAX): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  x$DAX
## KPSS Level = 0.079887, Truncation lag parameter = 6, p-value = 0.1
```

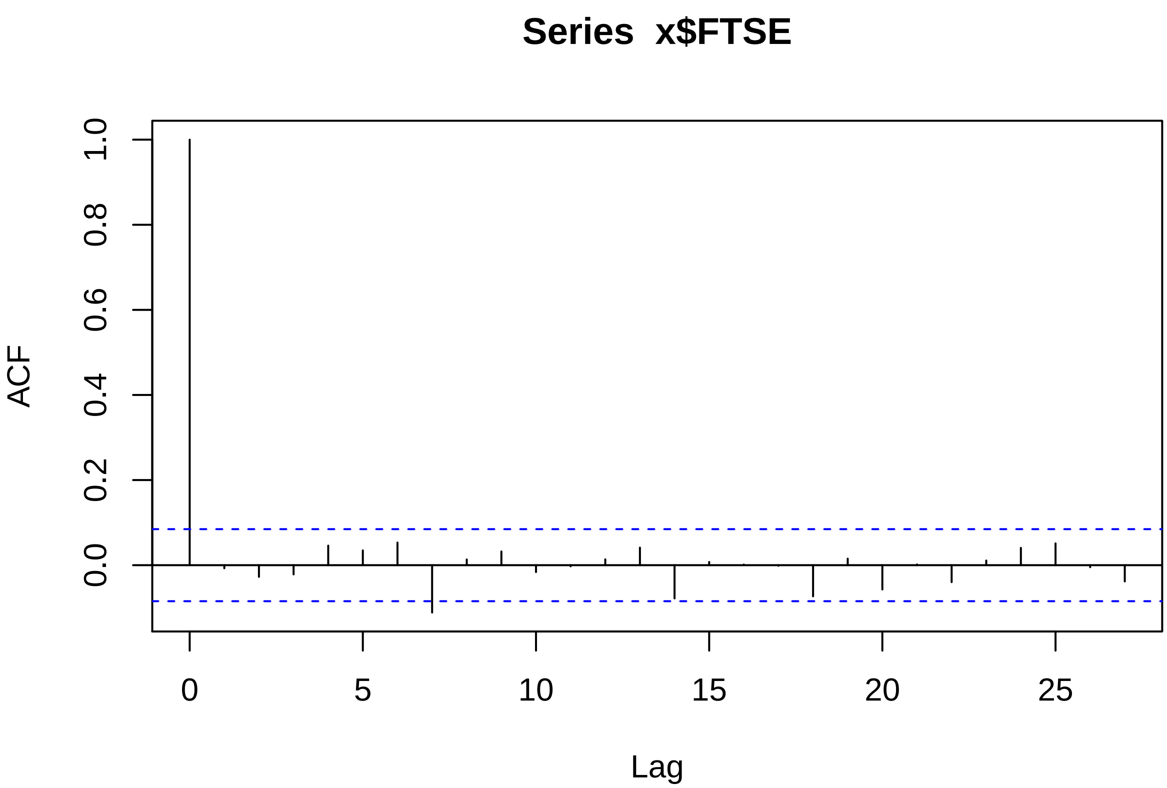
The DAX seems to be stationary in the mean. The ACF dies down quickly and follows a sinusoidal pattern about 0, which is typical of stationary time series. The p-value from the KPSS test is greater than 0.05, thus we fail to reject the null hypothesis of stationarity. From the ADF test, since the p-value is less than 0.05, we reject the null hypothesis of non-stationarity. The ADF and KPSS tests suggest that DAX series is stationary.

FTSE

```
plot.ts(x$FTSE)
```




```r
acf(x$FTSE)
```




```r
adf.test(x$FTSE)
```

```
## Warning in adf.test(x$FTSE): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
```

```
## data:  x$FTSE
## Dickey-Fuller = -7.8456, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(x$FTSE)
```

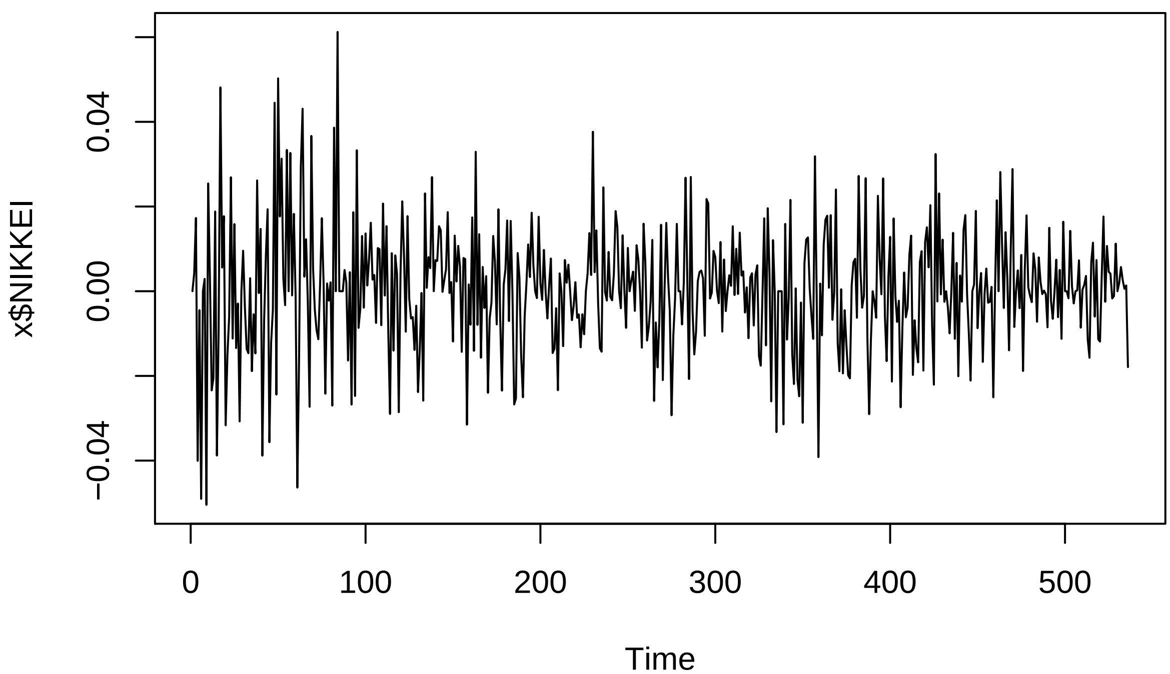
```
## Warning in kpss.test(x$FTSE): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  x$FTSE
## KPSS Level = 0.078836, Truncation lag parameter = 6, p-value = 0.1
```

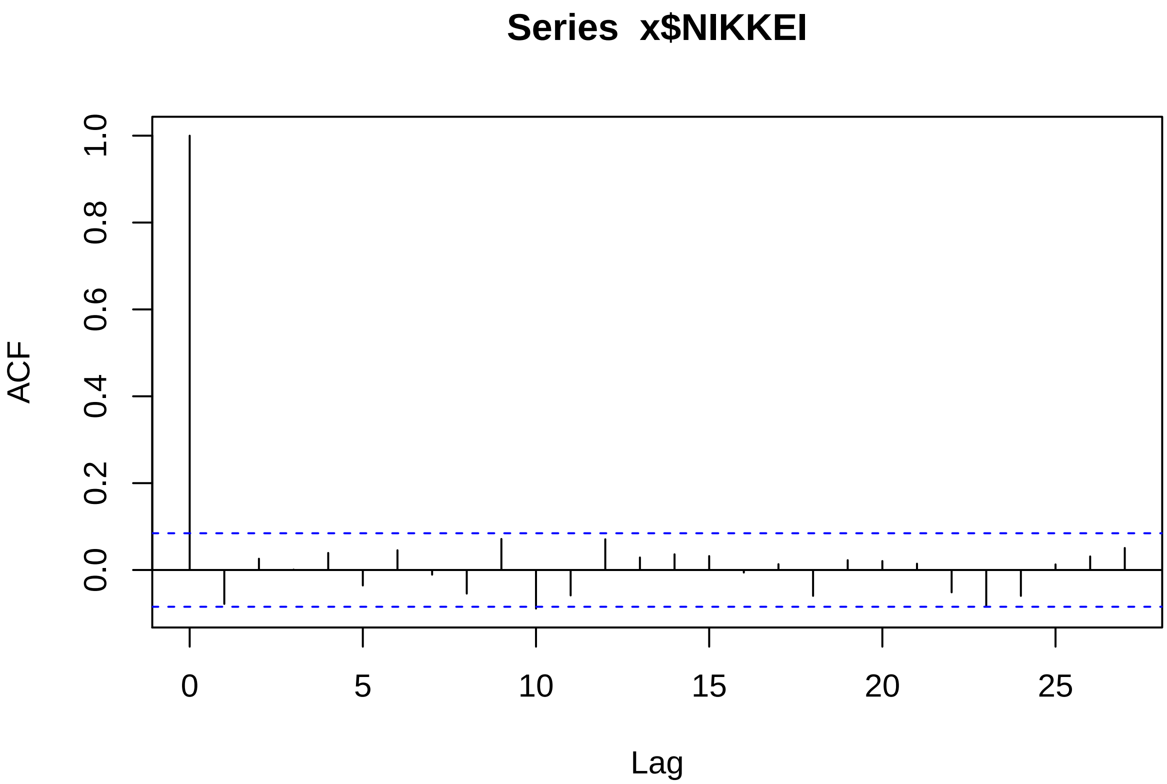
The FTSE seems to be stationary in the mean. The ACF dies down quickly and follows a sinusoidal pattern about 0, which is typical of stationary time series. The p-value from the KPSS test is greater than 0.05, thus we fail to reject the null hypothesis of stationarity. From the ADF test, since the p-value is less than 0.05, we reject the null hypothesis of non-stationarity. The ADF and KPSS tests suggest that FTSE series is stationary.

NIKKEI

```r
plot.ts(x$NIKKEI)
```




```r
acf(x$NIKKEI)
```




```r
adf.test(x$NIKKEI)
```

```
## Warning in adf.test(x$NIKKEI): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  x$NIKKEI
## Dickey-Fuller = -7.7679, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(x$NIKKEI)
```

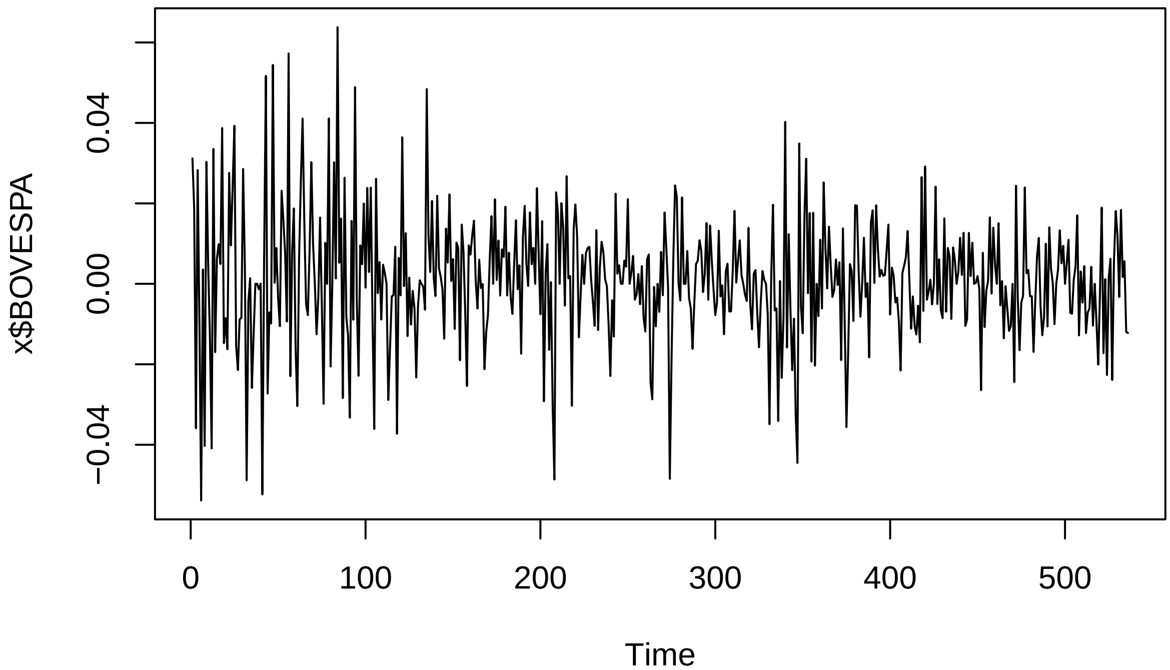
```
## Warning in kpss.test(x$NIKKEI): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  x$NIKKEI
## KPSS Level = 0.061547, Truncation lag parameter = 6, p-value = 0.1
```

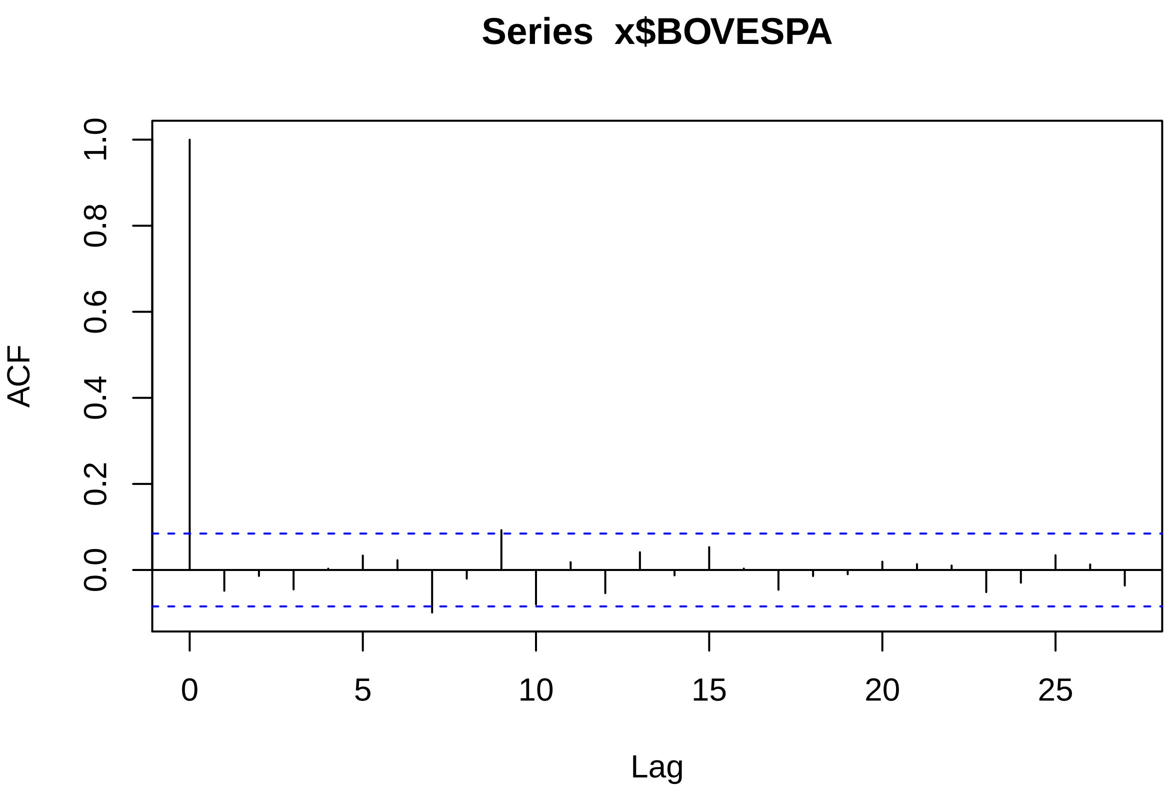
The NIKKEI seems to be stationary in the mean. The ACF dies down quickly and follows a sinusoidal pattern about 0, which is typical of stationary time series. The p-value from the KPSS test is greater than 0.05, thus we fail to reject the null hypothesis of stationarity. From the ADF test, since the p-value is less than 0.05, we reject the null hypothesis of non-stationarity. The ADF and KPSS tests suggest that NIKKEI series is stationary.

BOVESPA

```r
plot.ts(x$BOVESPA)
```




```
acf(x$BOVESPA)
```




```
adf.test(x$BOVESPA)
```

```
## Warning in adf.test(x$BOVESPA): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
```

```
## data:  x$BOVESPA
## Dickey-Fuller = -7.7122, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(x$BOVESPA)
```

```
## Warning in kpss.test(x$BOVESPA): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  x$BOVESPA
## KPSS Level = 0.26752, Truncation lag parameter = 6, p-value = 0.1
```

The BOVESPA seems to be stationary in the mean. The ACF dies down quickly and follows a sinusoidal pattern about 0, which is typical of stationary time series. The p-value from the KPSS test is greater than 0.05, thus we fail to reject the null hypothesis of stationarity. From the ADF test, since the p-value is less than 0.05, we reject the null hypothesis of non-stationarity. The ADF and KPSS tests suggest that BOVESPA series is stationary.

2) Split the data into train and test, keeping only the last 10 rows for test (from date 9-Feb-11). Remember to use only train dataset for #3 to #6.

```r
n <- dim(x)[1]
train <- x[1:(n-10),]
test <- tail(x,10)
```

3) Linearly regress ISE against the remaining 5 stock index returns - determine which coefficients are equal or better than 0.02 (*) level of significance?

```r
mod <- lm(ISE ~ SP + DAX + FTSE + NIKKEI + BOVESPA, data = train)
summary(mod)
```

```
##
## Call:
## lm(formula = ISE ~ SP + DAX + FTSE + NIKKEI + BOVESPA, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.071180 -0.009248  0.000083  0.009304  0.051863
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0008833  0.0006640   1.330 0.183979
## SP          -0.0607521  0.0770823  -0.788 0.430970
## DAX          0.3417440  0.0961243   3.555 0.000412 ***
## FTSE         0.6033493  0.1077621   5.599 3.50e-08 ***
## NIKKEI       0.3266529  0.0462163   7.068 5.09e-12 ***
## BOVESPA      0.1117630  0.0626647   1.784 0.075087 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0152 on 520 degrees of freedom
## Multiple R-squared:  0.493,  Adjusted R-squared:  0.4881
## F-statistic: 101.1 on 5 and 520 DF,  p-value: < 2.2e-16
```

DAX, FTSE and NIKKEI are significant at the 0.001 level, better than 0.02 level of significance.

4) For the non-significant coefficients, continue to lag by 1 day until all coefficients are better than 0.02 (*) level of significance. Use slide() function from package DataCombine. Remember you will need to lag, so you slideBy = -1 each step. How many lags are needed for each independent variable?

```r
library(DataCombine)
slide_sp <- slide(data = train, Var = "SP", TimeVar='date', NewVar='sp2', slideBy = -1)
```

```
##
## Lagging SP by 1 time units.
```

```r
head(slide_sp)
```

```
##        date          ISE           SP          DAX         FTSE
## 1  5-Jan-09  0.038376187 -0.004679315  0.002193419  0.003894376
## 2  6-Jan-09  0.031812743  0.007786738  0.008455341  0.012865611
## 3  7-Jan-09 -0.026352966 -0.030469134 -0.017833062 -0.028734593
## 4  8-Jan-09 -0.084715902  0.003391364 -0.011726277 -0.000465999
## 5  9-Jan-09  0.009658112 -0.021533208 -0.019872754 -0.012709717
## 6 12-Jan-09 -0.042361155 -0.022822626 -0.013525735 -0.005025533
##         NIKKEI     BOVESPA          sp2
## 1  0.000000000  0.03119023           NA
## 2  0.004162452  0.01891958 -0.004679315
## 3  0.017292932 -0.03589858  0.007786738
## 4 -0.040061309  0.02828315 -0.030469134
## 5 -0.004473502 -0.00976388  0.003391364
## 6 -0.049038532 -0.05384947 -0.021533208
```

```r
mod.sp.slide <- lm(ISE ~ sp2 + DAX + FTSE + NIKKEI + BOVESPA, data = slide_sp )
summary(mod.sp.slide)
```

```
##
## Call:
## lm(formula = ISE ~ sp2 + DAX + FTSE + NIKKEI + BOVESPA, data = slide_sp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.069334 -0.009554  0.000174  0.009197  0.053806
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0007035  0.0006537   1.076 0.282376
## sp2         0.2233447  0.0564961   3.953 8.78e-05 ***
## DAX         0.3106823  0.0912562   3.405 0.000714 ***
```

```
## FTSE        0.5946201  0.1056108   5.630 2.95e-08 ***
## NIKKEI      0.2095742  0.0545883   3.839 0.000139 ***
## BOVESPA     0.1073495  0.0531139   2.021 0.043780 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01493 on 519 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5086, Adjusted R-squared:  0.5039
## F-statistic: 107.4 on 5 and 519 DF,  p-value: < 2.2e-16
```

```
slide_bovespa <- slide(data = slide_sp, Var = "BOVESPA", TimeVar='date', NewVar='bovespa2', slideBy = -
```

```
##
## Lagging BOVESPA by 1 time units.
```

```
head(slide_bovespa)
```

```
##        date          ISE           SP          DAX         FTSE
## 1  5-Jan-09  0.038376187 -0.004679315  0.002193419  0.003894376
## 2  6-Jan-09  0.031812743  0.007786738  0.008455341  0.012865611
## 3  7-Jan-09 -0.026352966 -0.030469134 -0.017833062 -0.028734593
## 4  8-Jan-09 -0.084715902  0.003391364 -0.011726277 -0.000465999
## 5  9-Jan-09  0.009658112 -0.021533208 -0.019872754 -0.012709717
## 6 12-Jan-09 -0.042361155 -0.022822626 -0.013525735 -0.005025533
##         NIKKEI     BOVESPA          sp2    bovespa2
## 1  0.000000000  0.03119023           NA          NA
## 2  0.004162452  0.01891958 -0.004679315  0.03119023
## 3  0.017292932 -0.03589858  0.007786738  0.01891958
## 4 -0.040061309  0.02828315 -0.030469134 -0.03589858
## 5 -0.004473502 -0.00976388  0.003391364  0.02828315
## 6 -0.049038532 -0.05384947 -0.021533208 -0.00976388
```

```
mod.bovespa.slide <- lm(ISE ~ sp2 + DAX + FTSE + NIKKEI + bovespa2, data = slide_bovespa )
summary(mod.bovespa.slide)
```

```
##
## Call:
## lm(formula = ISE ~ sp2 + DAX + FTSE + NIKKEI + bovespa2, data = slide_bovespa)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064107 -0.009396  0.000428  0.009102  0.052397
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0006939  0.0006508   1.066  0.28679
## sp2         0.0668540  0.0731440   0.914  0.36114
## DAX         0.3349829  0.0894254   3.746  0.00020 ***
## FTSE        0.6419932  0.1027909   6.246 8.79e-10 ***
## NIKKEI      0.2196822  0.0541083   4.060 5.66e-05 ***
## bovespa2    0.1737584  0.0596447   2.913  0.00373 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01487 on 519 degrees of freedom
```

```
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5127, Adjusted R-squared:  0.508
## F-statistic: 109.2 on 5 and 519 DF,  p-value: < 2.2e-16
```

```r
slide_sp2 <- slide(data = slide_bovespa, Var = "sp2", TimeVar='date', NewVar='sp', slideBy = -1)
```

```
##
## Lagging sp2 by 1 time units.
```

```r
head(slide_sp2)
```

```
##        date          ISE           SP          DAX         FTSE
## 1  5-Jan-09  0.038376187 -0.004679315  0.002193419  0.003894376
## 2  6-Jan-09  0.031812743  0.007786738  0.008455341  0.012865611
## 3  7-Jan-09 -0.026352966 -0.030469134 -0.017833062 -0.028734593
## 4  8-Jan-09 -0.084715902  0.003391364 -0.011726277 -0.000465999
## 5  9-Jan-09  0.009658112 -0.021533208 -0.019872754 -0.012709717
## 6 12-Jan-09 -0.042361155 -0.022822626 -0.013525735 -0.005025533
##         NIKKEI     BOVESPA          sp2    bovespa2           sp
## 1  0.000000000  0.03119023           NA          NA           NA
## 2  0.004162452  0.01891958 -0.004679315  0.03119023           NA
## 3  0.017292932 -0.03589858  0.007786738  0.01891958 -0.004679315
## 4 -0.040061309  0.02828315 -0.030469134 -0.03589858  0.007786738
## 5 -0.004473502 -0.00976388  0.003391364  0.02828315 -0.030469134
## 6 -0.049038532 -0.05384947 -0.021533208 -0.00976388  0.003391364
```

```r
mod.sp.slide2 <- lm(ISE ~ sp + DAX + FTSE + NIKKEI + bovespa2, data = slide_sp2 )
summary(mod.sp.slide2)
```

```
##
## Call:
## lm(formula = ISE ~ sp + DAX + FTSE + NIKKEI + bovespa2, data = slide_sp2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.063412 -0.009491  0.000468  0.008739  0.050599
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0007513  0.0006491   1.157 0.247635
## sp          -0.1082670  0.0455658  -2.376 0.017861 *
## DAX          0.3355329  0.0890788   3.767 0.000184 ***
## FTSE         0.6368064  0.1024472   6.216 1.05e-09 ***
## NIKKEI       0.2395311  0.0489366   4.895 1.32e-06 ***
## bovespa2     0.2057244  0.0452856   4.543 6.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01481 on 518 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.516,  Adjusted R-squared:  0.5113
## F-statistic: 110.4 on 5 and 518 DF,  p-value: < 2.2e-16
```

For all variables to be significant at the 0.02 level, 2 lags are needed for variable SP and 1 lag is needed for BOVESPA.

5) Find correlations between ISE and each independent variable. Sum the square of the correlations. How does it compare to R-squared from #4?

```
sum_corr <- cor(train$ISE,train$SP)^2 + cor(train$ISE,train$DAX)^2 + cor(train$ISE,train$FTSE)^2 + cor(
sum_corr
```

```
## [1] 1.374691
```

```
summary(mod.sp.slide2)$r.squared
```

```
## [1] 0.5159921
```

The R-squared of the train data is much less than the sum of square of the correlations between ISE and each independent variable in the training data.

6) Concept question 1 - why do you think the R-squared in #4 is so much less than the sum of square of the correlations?

In linear regression, when all the variables are uncorrelated with one another, r-squared would equal the sum of square of the correlations. In our example, the sum of square of the correlations being much greater than the r-squared suggests there's overlap (multicollinearity) betweeen the variables.

7) Take the test dataset - perform the same lags from #4 and call predict() function using the lm regression object from #4. Why do you need to use the lm function object from #4?

```
test_slide1 <- slide(data = test, Var = "SP", TimeVar='date', NewVar = "sp", slideBy = -2)
```

```
##
## Lagging SP by 2 time units.
```

```
test_slides <- slide(data = test_slide1, Var = "BOVESPA", TimeVar='date', NewVar = 'bovespa2', slideBy =
```

```
##
## Lagging BOVESPA by 1 time units.
```

```
prediction <- predict(mod.sp.slide2,newdata = test_slides)
prediction
```

```
##           527           528           529           530           531
##            NA            NA  0.0108283060  0.0052504983  0.0008552345
##           532           533           534           535           536
##  0.0069010978  0.0053305872  0.0010620771 -0.0100557223 -0.0082730581
```

We use the lm function object from #4, which is the trained linear regression model on slided data from Jan 5th, 2009, to Feb 8th, 2011, to generate predictions for ISE in the weekdays between Feb 9th and Feb 22nd, 2011.

8) Concept question 2 - what do you find in #1 and why?

Through the qualitative analysis of time series data, I find that the time series plot for stationary data, despite the zig zags, follow a stationary mean. Another interesting feature of stationary data is that the ACF plot of stationary data typically dies down quickly and appears to mostly lie within the bounds of the blue dotted lines.

ADF and KPSS tests together provide a nice way of quantitatively determining the stationarity of time series data. Since the null hypothesis in ADF and KPSS are the opposite, when the null hypothesis is not rejected in one test, we need to see the null rejected in another test to conclude decisively the stationarity of data.