Assessment 2

Exploratory Data Analysis and Visualization

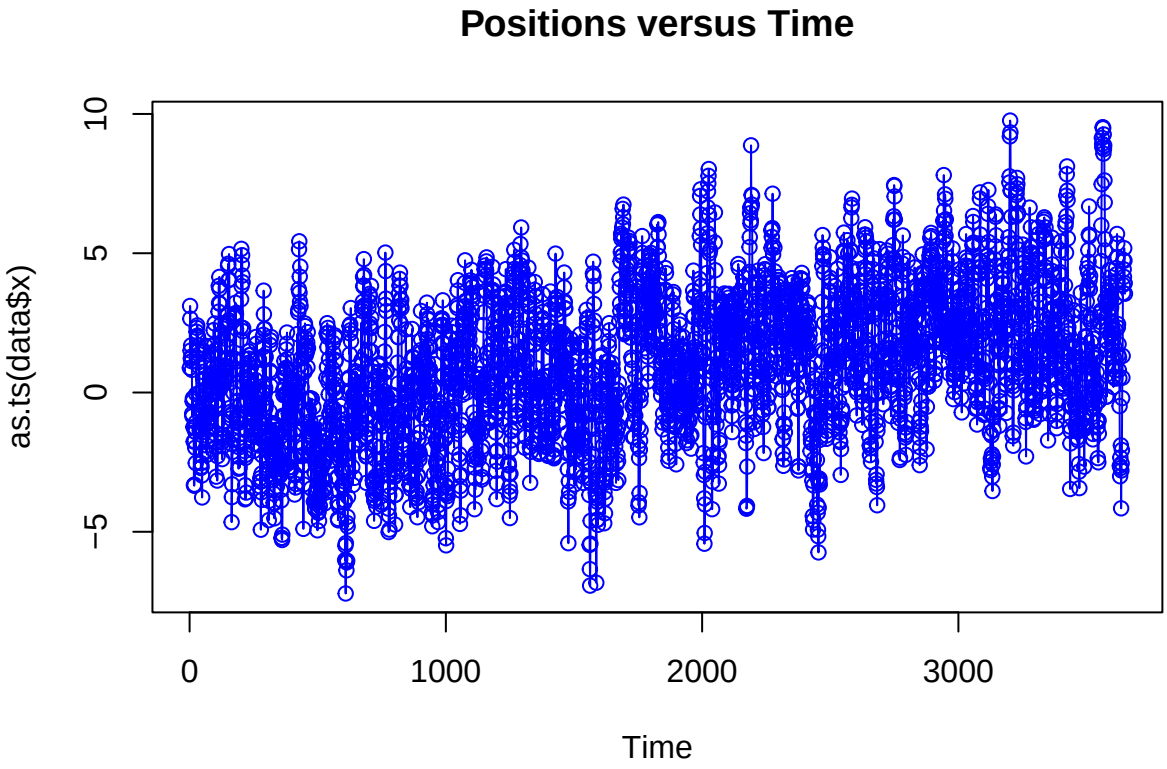
Ziyang Lin

Feb 15, 2022

Question 1

(a). We first load the data set into R and create a time series plot to visualize it.

```
data <- read.csv("~/Desktop/time-series.csv")
plot.ts(as.ts(data$x), col="blue", type="o", main="Positions versus Time")
```



In this plot we see the data has some random variation throughout all observations but still have some visual evidence of an increasing trend. In order to detect and remove the linear trend from the series, we first want to fit a linear model to estimate the trend component parameters a and b.

We can write this series in the form:

$$Y_t = a + bt + X_t,\ t \in 1, 2, \ldots, 3650$$

where X_t is the detrended (trend-free) component of the series. We can fit the model using the below code.

```r
time <- 1:length(data$x)
lm_fit <- lm(x~time, data=data)
knitr::kable(summary(lm_fit)$coefficients, caption="Trend Component Estimates")
```

Table 1: Trend Component Estimates

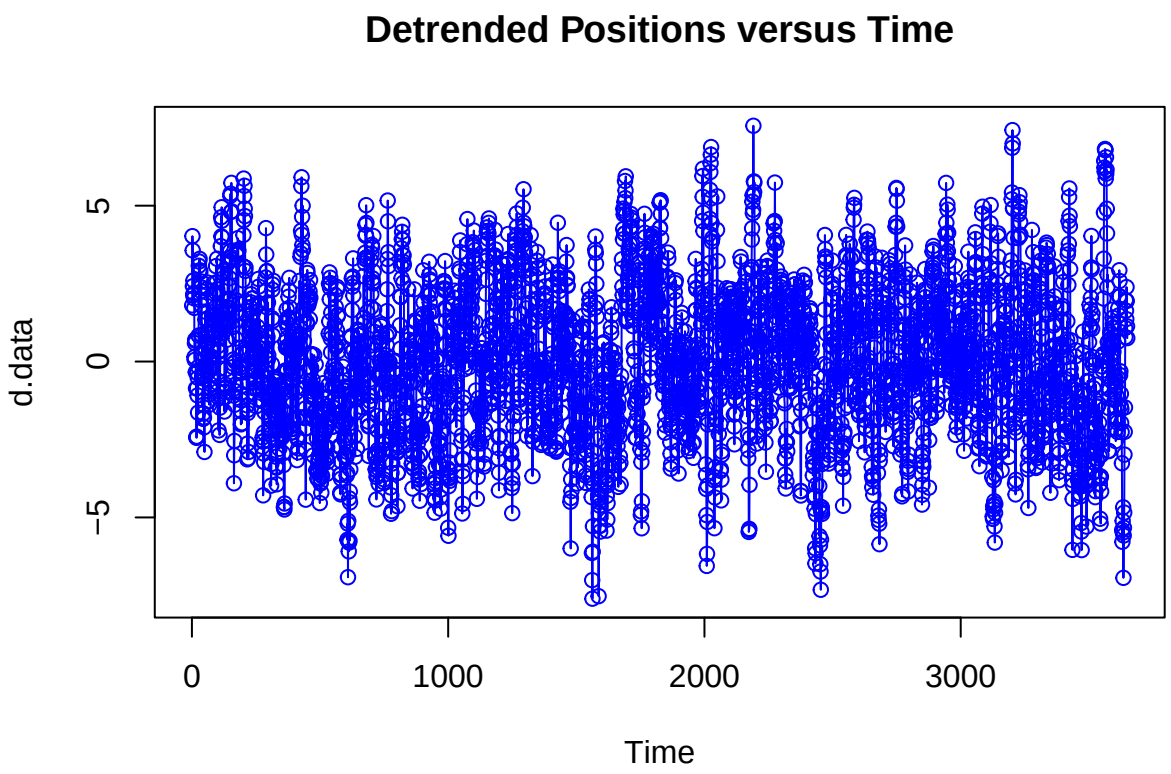
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.9175246	0.0792387	-11.57924	0
time	0.0010170	0.0000376	27.05334	0

From the above table, we get the estimates $\hat{a} = -0.9175, \hat{b} = 0.0010$. These two coefficients all have very low p-values, indicating that the data suggests the trend component does exist in the time series. Specifically, we can write the trend component as:

$$\hat{a} + \hat{b}t = -0.9175 + 0.0010t$$

Next, we detrend and plot the data using the below code.

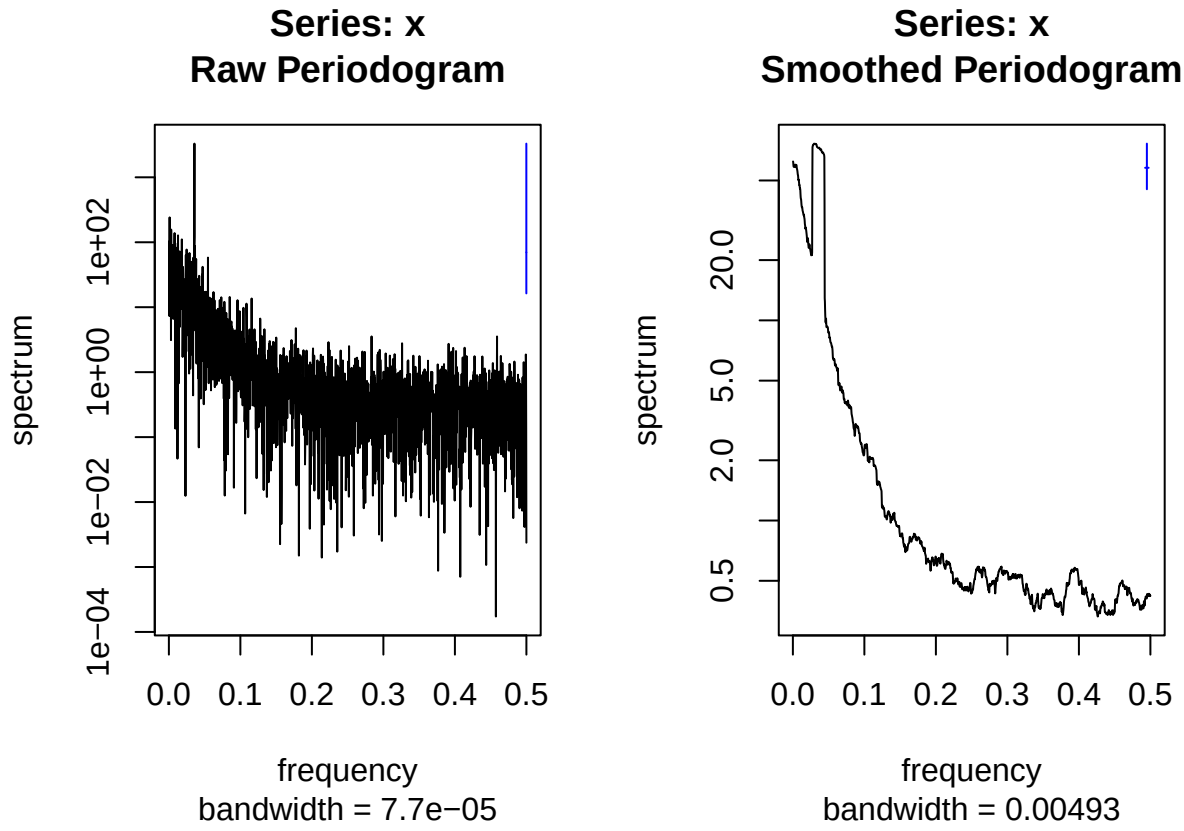
```r
trend <- lm_fit$fitted.values
d.data <- as.ts(data$x - trend)
plot(d.data, type="o", col="blue", main="Detrended Positions versus Time")
```



We now see that the data has no obvious linear trend.

(a). We will do spectrum analysis to detect seasonality in this series. Consider the below two periodogram, the left one is hard to identify, and the right one is smoothed with `span=65`. We can see that the spectrum value decay as frequency gets larger except the frequencies below 0.1.

```
par(mfrow=c(1, 2))
spectrum(d.data)
d.data.spec <- spectrum(d.data, span=65)
```



We can also output the top five frequencies and their specturm values.

```
spec.df <- data.frame(spec=d.data.spec$spec, freq=d.data.spec$freq)
knitr::kable(head(spec.df[order(spec.df$spec, decreasing=TRUE), c(1, 2)]),
             caption="Top Five Frequencies and Spectrums")
```

Table 2: Top Five Frequencies and Spectrums

	spec	freq
112	76.37934	0.0298667
111	76.30227	0.0296000
120	76.27207	0.0320000
113	76.18380	0.0301333
114	76.13085	0.0304000
115	76.10579	0.0306667

There are indeed some seasonality contained in the series. Specifically, we can use the below code to find the dominant frequency as printed.

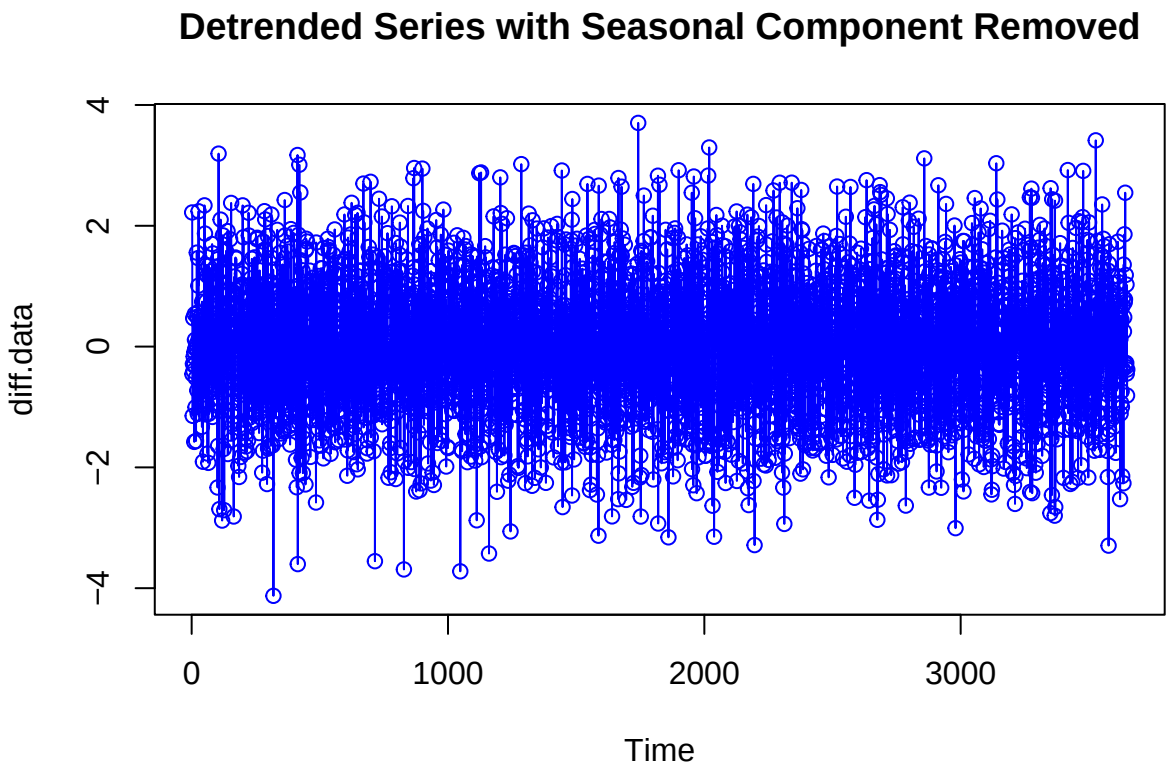
```
max_freq <- d.data.spec$freq[d.data.spec$spec == max(d.data.spec$spec)]
period <- 1/max_freq
sprintf("period = %.1f days", period)
```

```
## [1] "period = 33.5 days"
```

This implies that the dominant seasonality in this time series is 33.5 days, this value is very close to the number of days in a month. In fact, we see that the top five frequencies are all pretty close to $f = 0.03$. We can suspect that this time series has approximately monthly seasonality.

To remove it, we can simply do a seasonal differencing as below.

```
diff.data <- diff(d.data, differences=1)
plot(diff.data, type="o", col="blue",
     main="Detrended Series with Seasonal Component Removed")
```



(c). We can simply use the test function from `tseries` package.

```
adf.test(diff.data, k=1)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff.data
## Dickey-Fuller = -42.094, Lag order = 1, p-value = 0.01
## alternative hypothesis: stationary
```

This test assumes an $AR(1)$ model without shift of the form:

$$Y_t = \phi Y_{t-1} + \epsilon_t$$

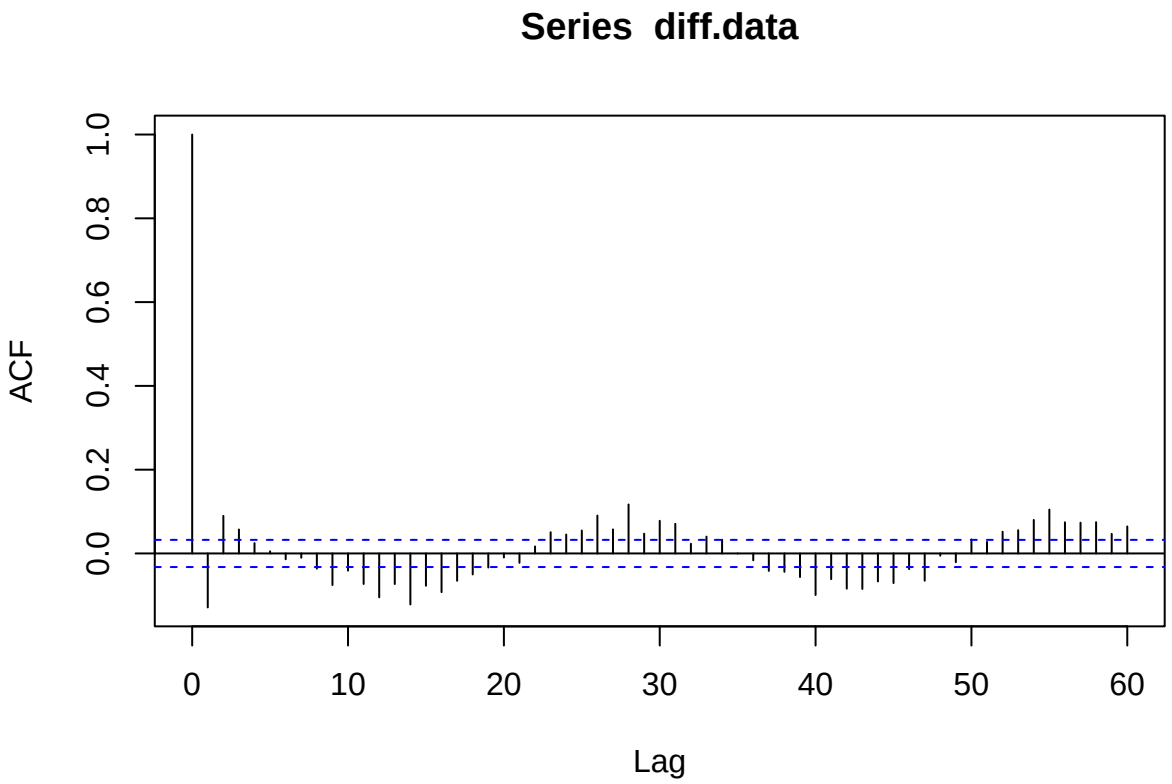
And the hypothesis set-up is:

$$H_0 : |\phi| = 1 \text{ (the process is not stationary) versus} \quad H_1 : |\phi| < 1 \text{ (the process is stationary)}$$

is used to test the stationarity of a time series, and the reported (smaller than) 0.01 p-value means that there is evidence in the data suggesting that we should reject the null hypothesis. That is, we have evidence against non-stationarity, i.e. our residual series is stationary.

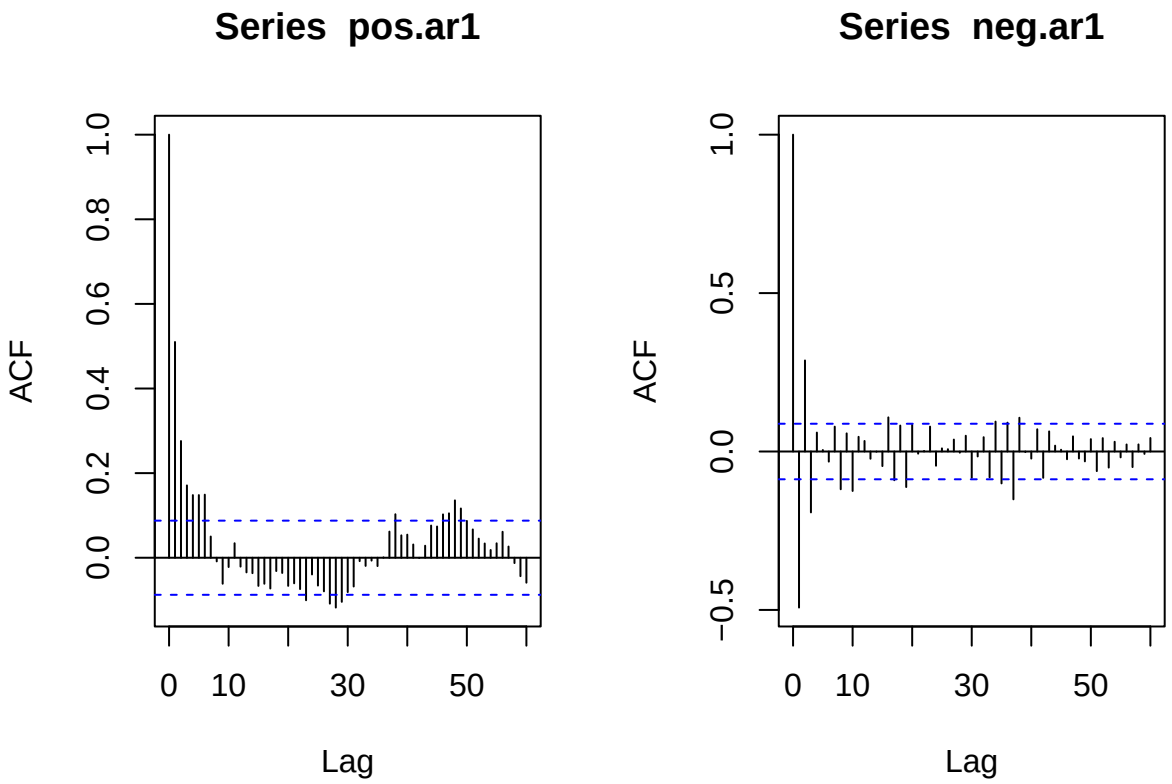
(d). The Dickey-Fuller test highly relies on the assumption that the series is an $AR(1)$ process, and we can verify this by getting the dependence structure via ACF plot.

```
acf(diff.data, lag.max = 60)
```



We see that the ACF tails off at the first few lags, but then grow larger at lags around $h = 14$, and then keep going up and down. This ACF does not look like an stationary $AR(1)$ model with $|\phi| < 1$. In fact, consider the below two simulation ACF plots for $AR(1)$ model each with $\phi = 0.5$ and $\phi = -0.5$:

```
par(mfrow=c(1, 2))
pos.ar1 <- arima.sim(list(order=c(1, 0, 0), ar=0.5), n=500)
neg.ar1 <- arima.sim(list(order=c(1, 0, 0), ar=-0.5), n=500)
acf(pos.ar1, lag.max = 60)
acf(neg.ar1, lag.max = 60)
```



We can see in both these simulated ACF plots, the ACF tails off after lag 1 and do not have a recurring structure. This implies that our residual series is not really an $AR(1)$ model, which means the assumption of DF test is not satisfied, so the DF test is not suitable for our residual series and we shall not be confident in its conclusion.