CSC425 – Time series analysis and forecasting
Homework 1
Total points: 70
YunTzu,Yu

For those of you who are new to R

- **See the tutorials and links posted in the "Module 0" content section on D2L**
- Useful resources: http://www.statmethods.net/ and http://www.ats.ucla.edu/stat/R/
- Install the following packages that will be used throughout the course: a. zoo
  b. forecast
  c. fBasics
  d. ggplot2
  e. ggfortify
  f. fpp2 (for one of the datasets we will use in this homework)

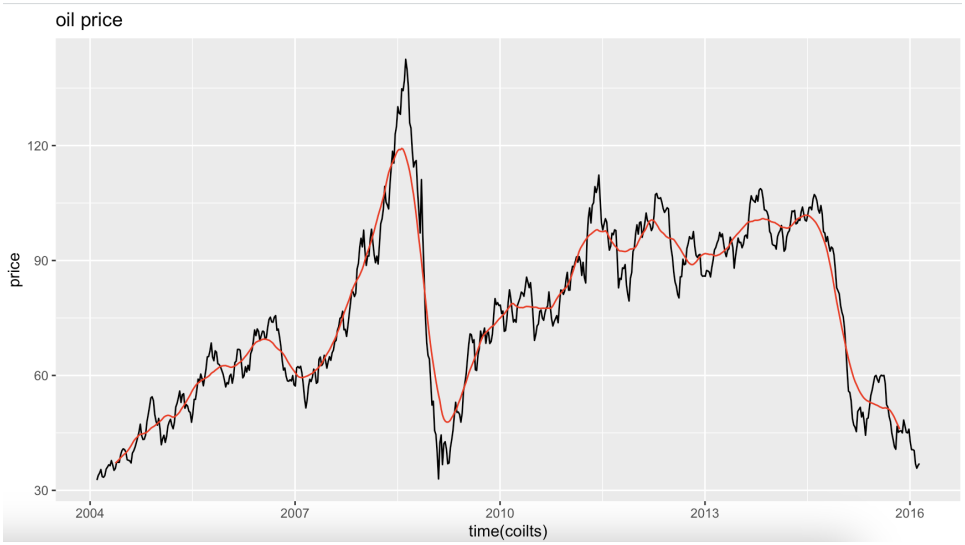
Problem 1 [15 points]
Consider the weekly spot prices for crude oil (dollars per gallon) from January 2004 to January 2016. The data file is crudeoil.csv and contains dates (date) and prices (price) separated by commas. In this problem, you are asked to analyze the *percentage change rate in oil price*.

a) Compute a 30-day moving average for the series of spot prices and plot it along with the series plot. Analyze the time trend displayed by the plot, and discuss if data show any striking pattern, such as upward/downward trends or seasonality.

From the graph, we could see mild upward trends,which means the oil price would mildly goes up among ten years.

We find a repeating pattern from 2012 to 2017 which could mean that the oil price is pretty stable at that time. We could see an outlier status in 2008 which means that oil price could be affected by the financial crsis then.

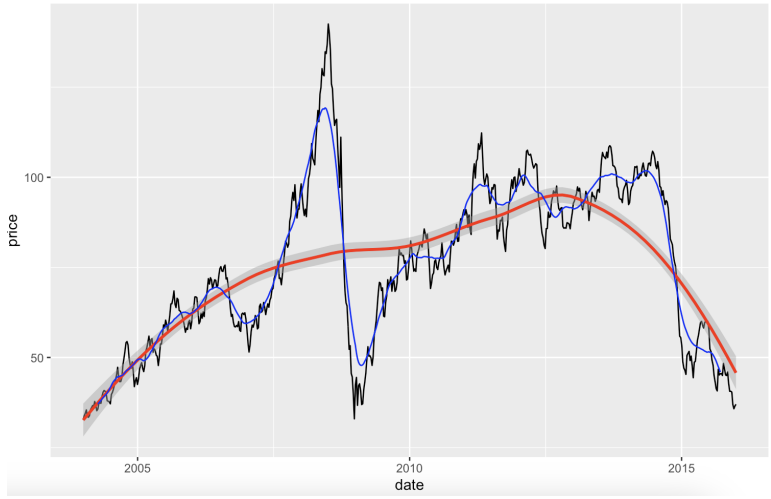
```
coil <- read_csv("Desktop/dsc 425/crudeoil.csv")
coil$date=as.Date(coil$date,"%d-%b-%y")
f30 = rep(1/30, 30)
f30
coilts = ts(coil$price,start = 2004.1,frequency = 52)
coilts
mAve = filter(coilts, f30, sides=2)

qplot(time(coilts), coilts, geom="line",ylab = "price",main = "oil price") +
 geom_line(aes(x=time(mAve), y=mAve), col="red")
```



b) Plot the series of spot prices along with a LOESS smoothing of the series (you may use ggplot to do this, see the notes about ggplot and LOESS in the lecture), evaluate and compare the results with the graph in a).

LOESS smoothing is a weighted linear regression at each point, here we try to use the window size = 30% to see the trend, we could see that it has a clear upward trend from 2003 to 2013. The LOESS smoothing mitigate the outlier affect so we could see more clearly the long-term trend here.

```
price=coil$price
date=coil$date
ds = data.frame(price,date)
ds
loess30 = loess(coil$price~ as.numeric(coil$date),data = ds, span =0.30)
smoothed30 = predict(loess30)

ggplot(data=ds, aes(x=date, y=price)) + geom_line() + geom_smooth(col="red")+
  geom_line(aes(x=date, y=mAve), col="blue")
```

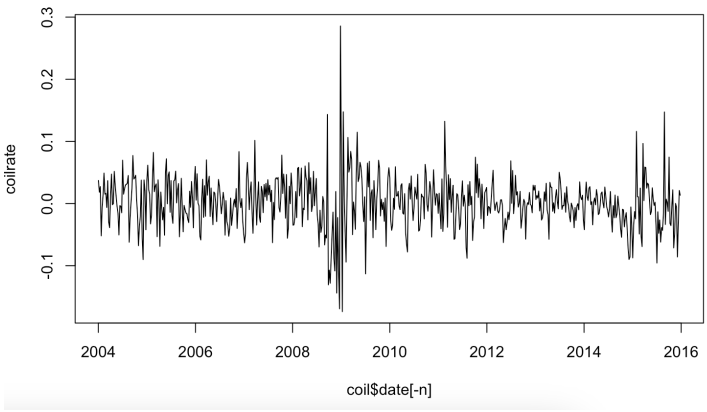


c) Compute the percentage change rate of spot prices using the formula

rate = $(p_t - p_{t-1}) / p_{t-1}$, where p_t is the oil price

Remember that if you make *p* into a time series as we did in class, you can use the lag function to compute p_{t-1}. Plot the *rate* series vs. time and discuss what the plot reveals about the rate series.

```
n = nrow(coil)

coilrate=diff(coil$price)/coil$price[-n]

plot(coil$date[-n],coilrate,type = "l")
```

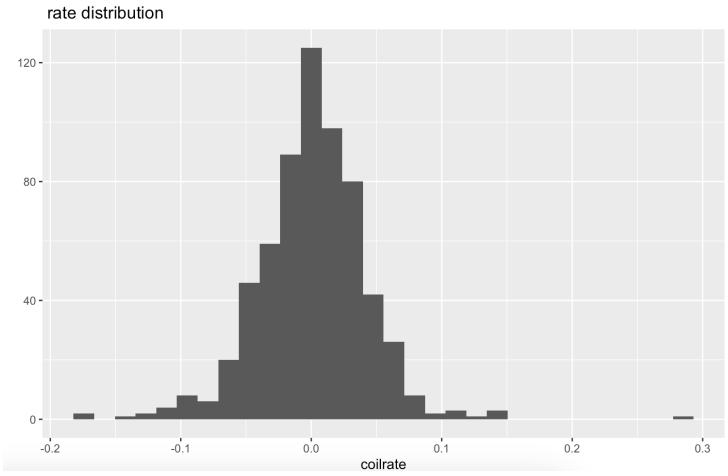


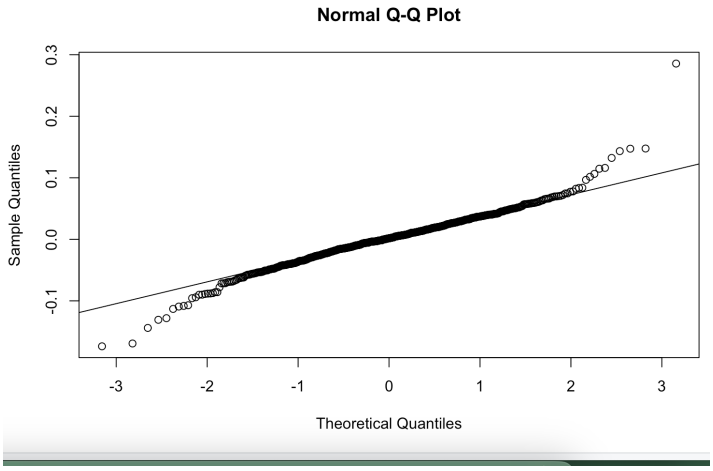
d) Analyze the distribution of the *rate* using a normal quantile plot. Discuss the results. Compute the symmetry and kurtosis of the rate distribution. Is it close to a normal distribution? Test the normality for the distribution of *rate* (possibly transformed) using the Jarque-Bera test

at a 95% level and discuss the result. (NormalTest from the fBasics is a good option).

From the histograms, we could see that the rate is a little bit skew to the left, we could also see a little skew at the low end and at the high end.
The result of the kurtosis is around 4.63 here, which is higher than 2 that the distribution of it should be a heavy-tail compared too normal.
The result of the skewness is 0.26 which shows that it is mildly skewed.
Then we check Jarque-Bera test, the result is pretty close to 0 , so the sample data could be a normal distribution.

```
qplot(coilrate,geom="histogram",main=" rate distribution")
qqnorm(coilrate)
qqline(coilrate)
```






```
rate = data.frame(coil$date[-n],coilrate)
names(rate)=c("date","rate")
library(fBasics)
```

```
kurtosis(rate$rate)
4.627621
attr(,"method")
[1] "excess"
skewness(rate$rate)
[1] 0.2586747
attr(,"method")
[1] "moment"
 library(tseries)
jarque.bera.test(rate$rate)

        Jarque Bera Test

data:  rate$rate
X-squared = 571.5, df = 2, p-value < 2.2e-16
```

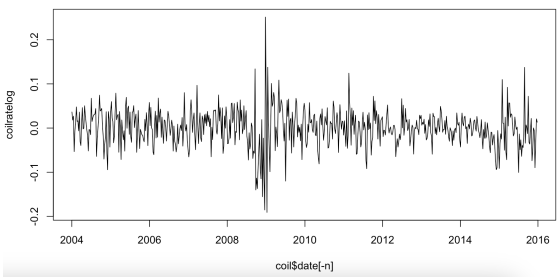
e) Compute the log-rate of change of the series (i.e. same as the log-return, compute the difference of the logs of the prices). Test the normality of the log-rate and compare to the last result. Discuss how taking the log changed the distribution.

If we have a multiplicatives series outcome here, the log transformation would help to reduce the variance growing with the value. However, the return rate right here tends to be a additive series, so log transformation just make the series be shifted a little bit ,it did not help a lot with this analysis

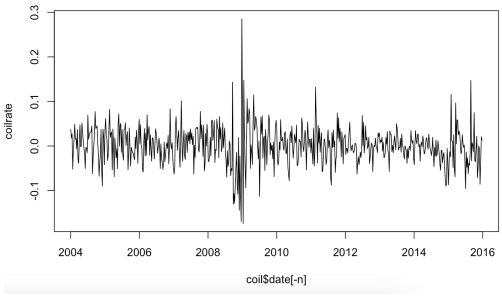
Log -rate is the differenced of the log of the x.
R = diff(log(x))

```
coilratelog=diff(log(coil$price))

plot(coil$date[-n],coilratelog,type = "l")
```

Log transformation



No log transformation



Problem 2 [15 points]

The dataset "groceries.csv" contains weekly units sold for three grocery items: ToothPaste (100ml container of toothpaste), peanut butter (340g. jat of crunchy peanut butter), and Biscuits (200g., 10 finger package of shortbread cookies). The dataset contains the variable Date defined as the first day of the week for the sales period. For this problem, you will analyze the weekly sales data for **ToothPaste**.

a) Create a "ts" time series for the object. What start and frequency should you use to correctly display the date? Explain your choices.

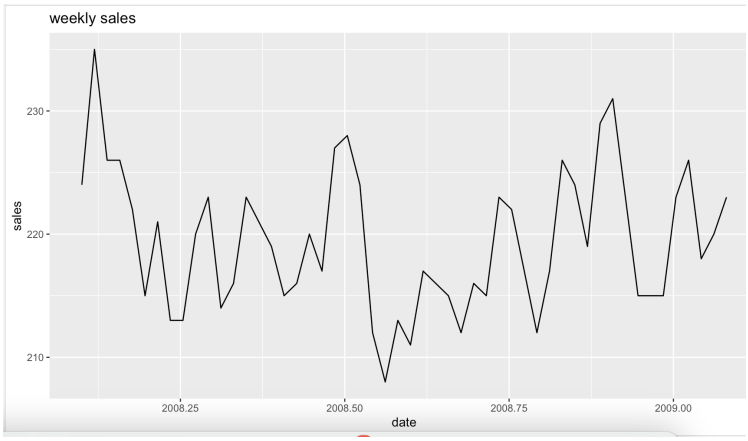
I would use 2008.1 as the start and 52 as frequency, because 2008.1 means it is the begging of the year 2008 ,and 52 means that there are 52 samples per unit of the time and in the year of 2008 we have 52 weeks.

```
groceries <- read_csv("Desktop/dsc 425/module 1/groceries.csv")
groceries$Date=as.Date(groceries$Date,"%d-%b-%y")
groceriestst <-ts(groceries$ToothPaste,start= 2008.1,frequency = 52)
groceriestsp <- ts(groceries$PeanutButter,2008.1,frequency = 52)
groceriestsb <- ts(groceries$Biscuits,2008.1,frequency = 52)
```

b) Create a time plot for the time series of ToothPaste weekly sales. Make sure the plot is correctly labeled and titled. Analyze the graph of the series, and discuss if the data show any striking patterns, such as trends or seasonality.

From the graph, we could tell that there is no a obvious upward or down trade for the ToothPaste weekly sales in 2008, and the size of swing did not change drastically, so we could say the weekly sales for ToothPasteis pretty stable.

```
qplot(time(groceriestst),groceriestst,geom="line",main="weekly
sales",xlab="date",ylab="sales")
```



c) Is the series additive or multiplicative? Justify your answer.

It should be additive,because the variance tends to stay the salem no matter what size the actual value is.

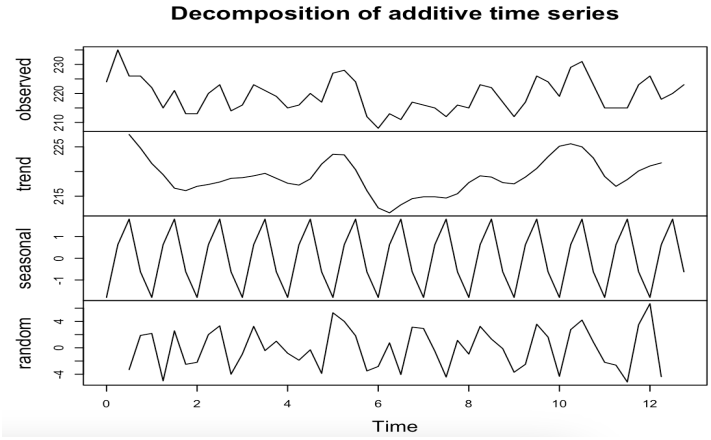
d) Use the “decompose” function, with the proper series type to plot the series in such a way as to highlight any seasonality present. You may need to try several different “frequency” values when converting the ToothPaste column to a time series (don’t worry if the dates do not align with years, or if you wish to keep the years present, you will have to use the zoo data type).

Remember that the decomposition should have a clean trend (no obvious seasonality in the trend), the seasonal variation should be as large as possible (look at the range on the y-axis) and the “Random” component should be as unstructured as possible. Analyze your final answer for each of the components of the decomposition.

Here I apply frequency = 4 here to see if there are seasonal pattern month to month. We could see the original one and find that there might be a seasonal pattern from February to May.

We could look at the Y- axis, the variation on the seasonal term in units is from -2 to 2, it’s about 4 units and the random term is between -4 and 3, the seasonal term is a bit comparable to the random term.

```
groceriestst <-ts(groceries$ToothPaste,0,frequency = 4)
dec = decompose(groceriestst)
plot(dec)
```



Problem 3 (15 points)

Load the R "fpp2" library (note that there are two R packages, fpp, and fpp2. Use the latter). For this problem, we will be using the "auscafe" dataset from this library, which contains data on the total monthly expenditure on cafes, restaurants, and takeout food services in Australia (in billions of Australian dollars) from April 1982 - September 2017. Note that it is already a "ts" object, so you do not need to convert it.

a) Print the head of the time series, including 20 samples. What is the frequency of this "ts" dataset? Explain why you can conclude this from the "head" output. Confirm your deduction by using the "frequency" function.

The frequency should be 12 for this dataset, because by applying frequence = 12, we will get 12 number of monthly samples per unit of time.

The table is in a form of period of month in the time series

```
> head(auscafe,20)
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
1982                        0.3424 0.3421 0.3287 0.3385 0.3315 0.3419 0.3584 0.3747 0.4331
1983 0.3686 0.3481 0.3658 0.3511 0.3605 0.3471 0.3645 0.3760 0.3776 0.3741 0.3906
```

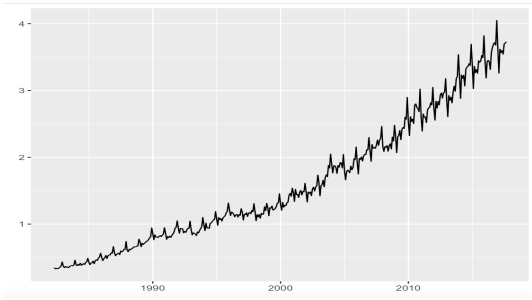
b) Create a time plot for the series. Make sure the plot is correctly labeled and titled. Analyze the time trend displayed by the plot, and discuss if the data show any striking pattern, such as upward/downward trends, obvious seasonality, and multiplicative behavior.

From the graph, we could see upward trends,which means monthly expenditure expotienally going up.

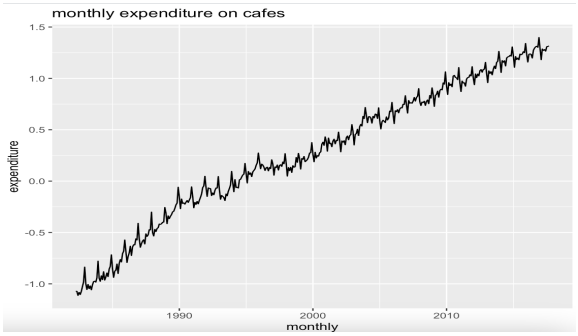
This is a multiplicative series, because the the bigger value is the bigger the swing are gonna to be, so we need to take the log to make it more stable.

qplot(time(auscafe), log(auscafe),xlab = "monthly",ylab = "expenditure", geom="line",main = "monthly expenditure on cafes")

Before log



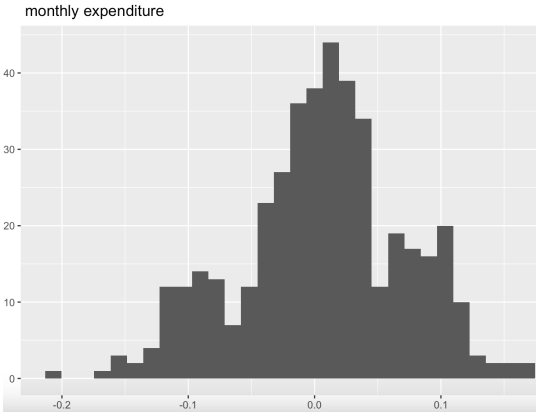
After log

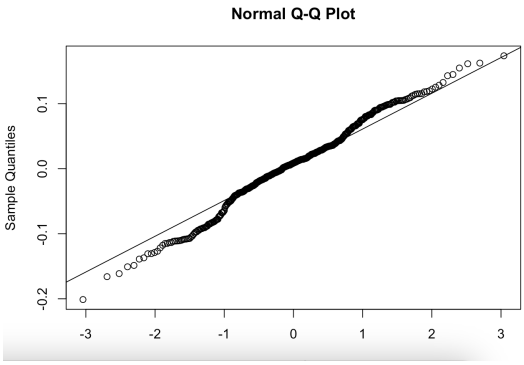


c) Compute and analyze the distribution of the returns or the log returns for the series (depending on your choice in b) using a histogram and a normal qq-plot. Is it close to a normal distribution? Is the distribution symmetric? How bad is its kurtosis?

From the histogram, we could tell it is close to a normal distribution and the distribution is nearly symmetric, it just mildly skewed to the left side.
From the qq-plot, we could see that it has heavy tail.

```
auscafediff=diff(log(auscafe))

qplot(auscafediff,geom="histogram",main=" monthly expenditure")
qqnorm(auscafediff)
qqline(auscafediff)
```





d) Test the hypothesis of normality for the distribution of rate using the Jarque-Bera test at the 5% level. You may use the NormalTest function from the fBasics package in R.

The Jarque-Bera test tells us that the test statistic is 3.1393 and the p-value of the test is 0.2081.
In this case, we would fail to reject the null hypothesis that the data is normally distributed.

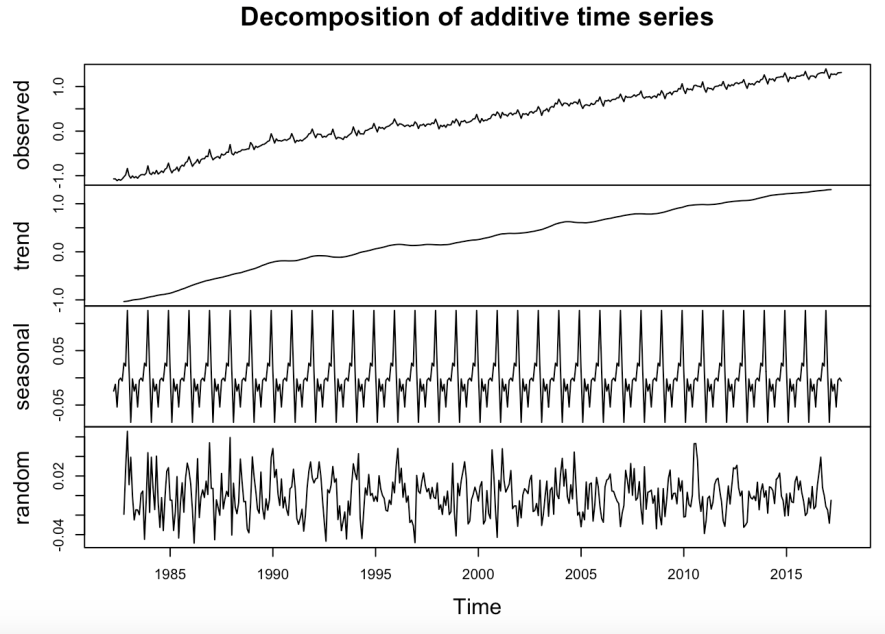
```
> jarqueberaTest(auscafediff)

Title:
 Jarque - Bera Normalality Test

Test Results:
```

STATISTIC:
X-squared: 3.1393
P VALUE:
Asymptotic p Value: 0.2081

e) Use the appropriate (additive or multiplicative) decompose function with the default frequency and evaluate the trend and seasonality that it computes.

We decompose it with the seasonality in month, we could tell from the graph it has an upward trend.We do not see the strong pattern that repeat monthly here.

```
decauscafe = decompose(log(auscafe))
plot(decauscafe)
```



Problem 4: Article review and reflection [20 pts]
Read the article "SEASONALITY IN TOURISM: A REVIEW" posted in the documents for homework 1 and answer the following questions in detail with short paragraphs (at least three sentences/four or five lines) developing your answer to each question.

a) What is the importance of understanding seasonality in tourism (or any other kind) time series data?

Since seasonality is a measurable feature in time series, we could predict what the resource that we need for the future.. We could explore some factors that could cause seasonal fluctuation in tourism time series, so we could adjust some

resources like labor or electronic machines for tourism to prevent unnecessary waste in some off-peak seasons. Understanding the main factors can also help to modify the imbalance in the phenomenon of tourism to prevent severe economic and social issues. These are the reason why it is important to understand seasonality in tourism.

b) What negative consequences of seasonal patterns do the authors explore?

The negative consequences of seasonal patterns would cause a seasonal loss in employment, investment, and the environment for the tourism industry.
Along with the fluctuation of seasonality, for the employer is hard to hire people as full-time employees. During the peak tourism season, it could result in the environment being overcrowded or overused and damage the environment. In terms of investment, it is difficult to maintain the asset consistently so it could lead to a certain unstable economic status.

c) Evaluate the author's exploration of positive side-effects or consequences of seasonality.

In spite of the seasonal fluctuation in the tourism industry would cause some damage either to the economy or to the environment, the off-peak season could definitely help people or the environment to be recovered. So we should consider the off-peak season as a chance to stop consuming the resource and let them have a chance to recover by themselves, after all, continuous use of natural resources without stopping may cause irreversible damage to the environment.

d) Thinking as a reviewer for the study in the article, come up with another possible cause of seasonality for tourism that the authors did not explore. Explain that cause and why it might be worth studying.

The political election may cause seasonality for tourism as well, for example, if the candidate would hold the election campaign in the city, the event may attract a lot of supporters to visit the place, so we could take the political factor into account to know when could be the peak that people would visit the place.

e) In some other data domain (i.e. subject or source of data), perhaps one that you work within your job or research field, explain a practical cause of seasonality that might occur in the data and explain the effect it may have on the study of time series for that type of data.

There are a lot of online courses nowadays on different websites. The student may apply for these online courses seasonally as a supplement tool for preparing for their test. So by analyzing the time series of the test, we could let these providers know when to release their promotion to stimulate the volume of the user for their website

and to increase the company's revenue.

Problem 5: "Reflection" Problem [5 pts]

Post a message on the Homework 1 discussion board on the D2L site reflecting on the topics in week 1. Indicate the topic or assignment in this module you found to be the easiest, and the one you found to be the hardest, and why.