Chapter 3
Time-Series Data

1 Introduction

At the end of the last chapter, we discussed a few summary statistics, such as mean and standard deviation, that you probably learned about in high school (and almost certainly covered in STAT 220). While these statistics will almost always provide meangingful summaries if we are using cross-sectional data, things get more complicated if we use time series data. When we have time series data, it will always still be possible to perform the calculations for these summary statistics. However, much of the time-series data that we analyze will be *non-stationary*, meaning that its features such as the mean and standard deviation are changing over time. Therefore, at least when we first load in data, plotting the data will be more informative than calculating simple summary statistics.

In this chapter, we will detail different types of plots for time-series data. The purpose of the plots is to identify propoerties of the data which may make it non-stationary, and therefore require us to transform the data. Some common properties that make a time-series non-stationary are:

- Trend
- Seasonality
- Unit Root

We will begin the chapter by discussing the most common plot for time series data — the time plot. We will then move to discussing data transformations, and cover plots that can help us identify unit roots or seasonality in the data.

2 Identifying Trend

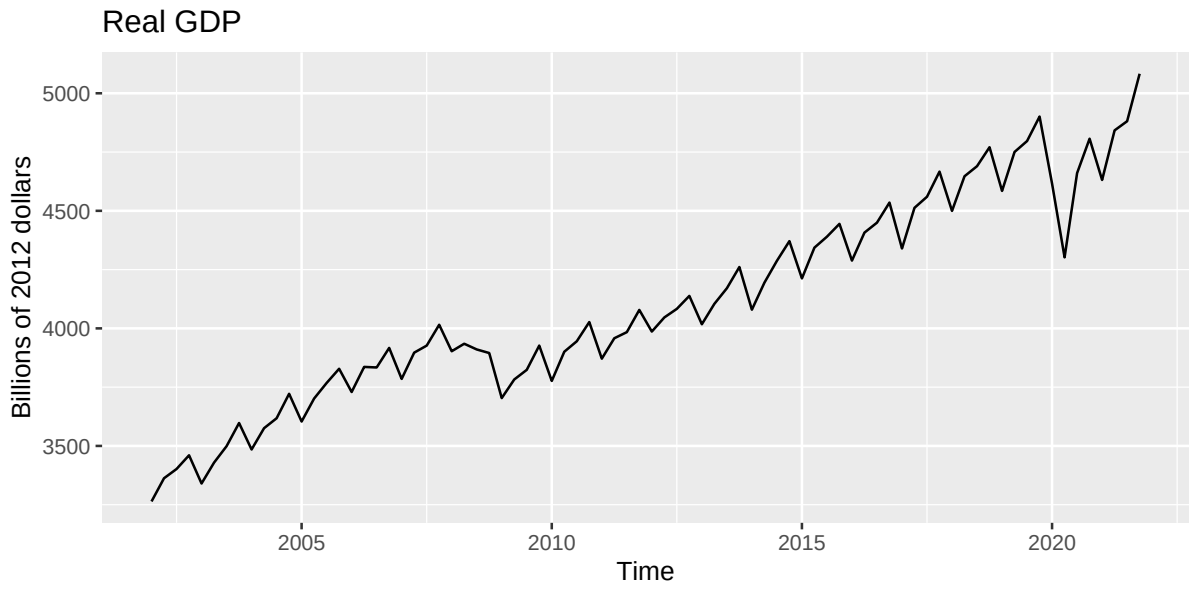
The first feature to look for is a long-run trend. Long-run trends are common in many raw economic and business data sets, since historically there has been population growth, growth in prices (inflation), and growth in real incomes over time. This means that, on average, series such as the total dollar value of sales have been increasing throughout the economy. Of course, at any given company, sales could be flat or declining over time, but at the majority of companies and for the majority of products, the total dollar value of sales tends to grow over time.

As you know from ECON 251, the total dollar value of everything produced within a country is called Gross Domestic Product (GDP). When held at constant prices to strip out the effect of inflation, it is called Real Gross Domestic Product (RGDP). This measure allows us to see how total production in a given country is changing over time.

2.1 Time Plot

We will use RGDP in the United States to illustrate most of the concepts in this chapter, starting with the *time plot*. A time plot plots the value of the data on the y-axis with time on the x-axis. This makes it easy to see how the data has been changing over time. Here is the time-plot for US RGDP beginning in the first quarter of 2002:

```
data_set <- read.csv("RGDP_NSA.csv")
Y <- ts(data_set[,2],start=c(2002,1),frequency=4)
autoplot(Y) + ggtitle("Real GDP") + ylab("Billions of 2012 dollars")
```

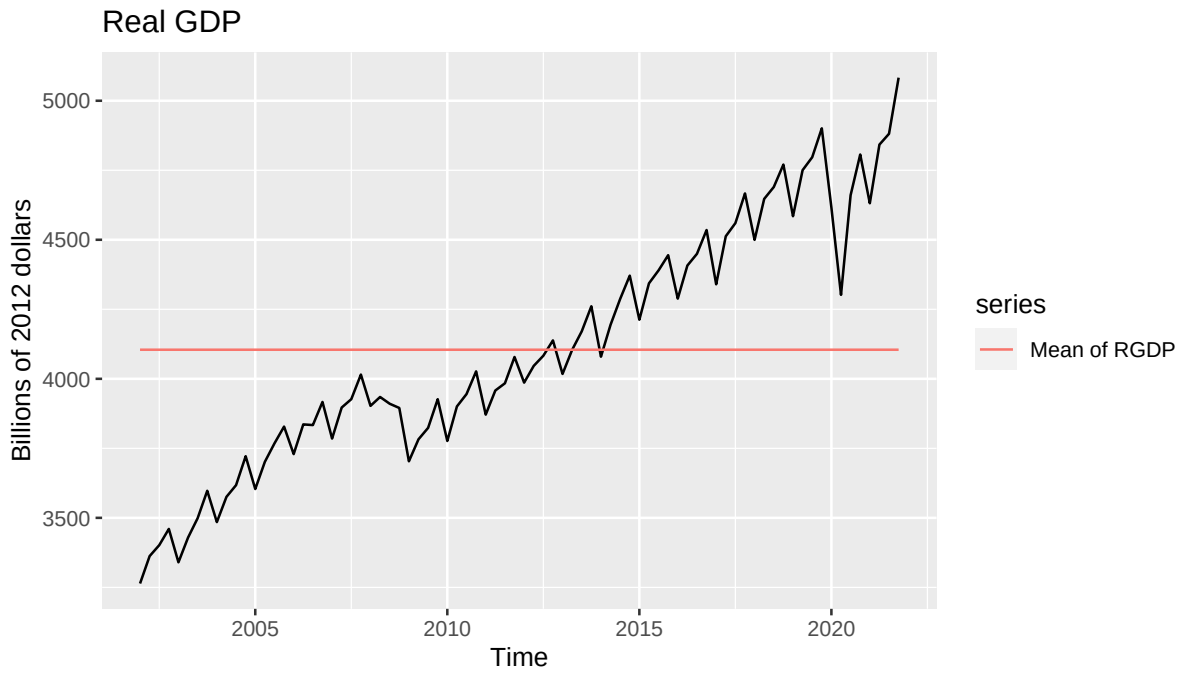


Although it was relatively flat between 2008 and 2011, we can see a clear upward trend in US RGDP over time. This long-run increase is called a "postive trend." Positive simply denotes that it is increasing over time, and does not signify if this is "good" or "bad" (for instance, the number of homicides in a given area may have a "positive trend," but that would be a very bad thing!). We also see some fluctuations around this trend. These fluctuations look fairly predictable, as they appear to happen at regular time intervals (for instance, RGDP looks like it usually declines at the beginning of each year). We will analyze these fluctuations in more later, when we investigate this series for "seasonality."

2.2 Transforming Trending Data

When time series data has either a positive or negative trend, the summary statistics such as mean and standard deviation are not informative. For example, take a look at the time plot for US RGDP with the mean superimposed:

```
Ym <- rep(mean(Y),length(Y))
Ym <- ts(Ym,start=c(2002,1),frequency=4)
autoplot(Y) + ggtitle("Real GDP") + ylab("Billions of 2012 dollars") +
  autolayer(Ym,series="Mean of RGDP")
```



While we can calculate the mean, we see that it tells us very little about the data, as the data in the first half of the sample is below the mean, and the data in the second half of the sample is above the mean. Simply knowing when the data was collected will tell us much more about the likely value of the data than knowing the mean. For trending data, knowing the mean is essentially worthless.

In addition to making our summary statistics meaningless, trends can also hide other features of the data. For instance, if there is a very strong trend but only mild seasonality, it can be nearly impossible to see the seasonality when the raw data is plotted on a time plot. Therefore, when data has a trend, we will transform it before continuing with our analysis.

2.3 Transforming Trending Data

When data has a trend, we will transform it to remove the trend. There are two types of trends that we will commonly encounter: *linear* and *exponential*. The difference is the type of growth that the data exhibits. While some series grow by a roughly constant number of *units* over time, others grow by a roughly constant *percentage* over time. Linear and exponential trends will require a slightly different transformation.

If the data grow by a roughly constant number of units over time (e.g. \$100 per quarter), then the data has a *linear trend*. To transform data with a linear trend, we will take the *first difference* of the data. That is, at each point in time, we will compute $y_t - y_{t-1}$ so the data is

transformed to be the change per period:

$$\Delta Y = \begin{bmatrix} y_2 - y_1 \\ y_3 - y_2 \\ \vdots \\ y_T - y_{T-1} \end{bmatrix}$$

This will remove the trend, and the data will be relatively flat, moving around a roughly constant mean. Our summary statistics for ΔY will be meaningful, unlike the summary statistics for Y.

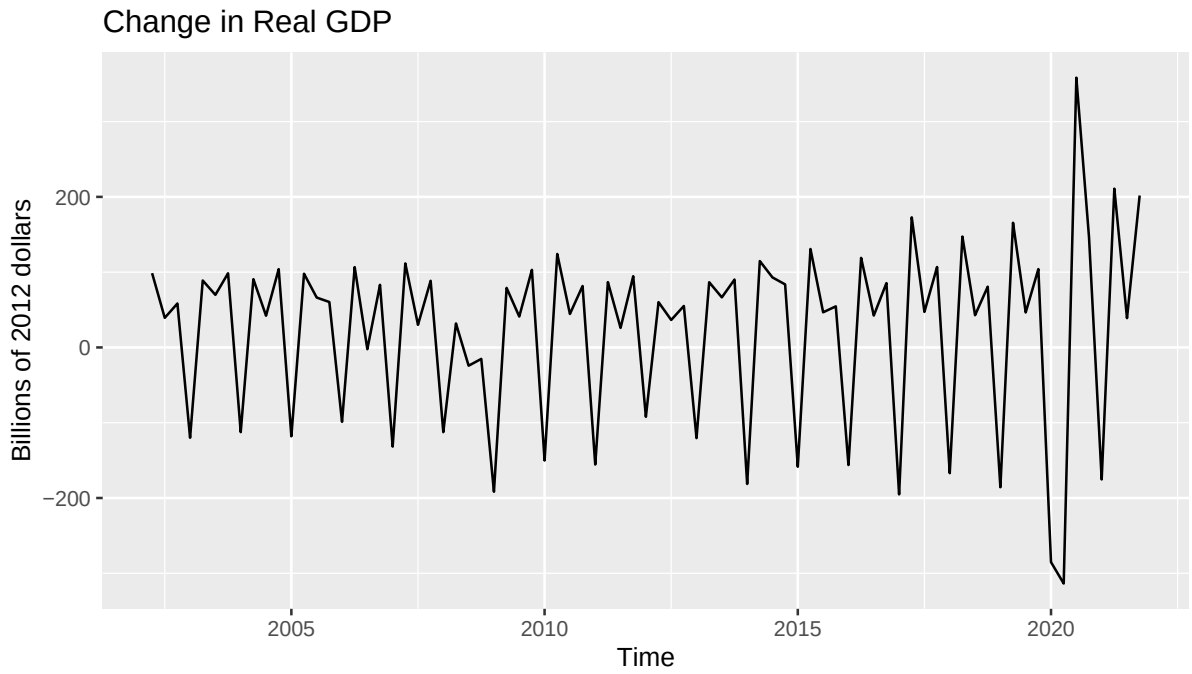
If the data grow by a roughly constant percentage over time (e.g. 1% per quarter), then the data has an *exponential trend*. To transform data with an exponential trend, we will take the *first difference of the natural log* of the data. This will convert the series to be in percent change per period. That is, at each point in time, we will compute $\ln(y_t) - \ln(y_{t-1})$ so the data is transformed to be the *precent* change per period:

$$\%\Delta Y = \begin{bmatrix} \ln(y_2) - \ln(y_1) \\ \ln(y_3) - \ln(y_2) \\ \vdots \\ \ln(y_T) - \ln(y_{T-1}) \end{bmatrix}$$

2.4 Distinguishing Linear from Exponential Trend

So if our data appears to be trending, how do we know if our data has a linear or exponential trend? One way is to plot the first differenced data, ΔY, on a time plot. If the variance of the first difference seems to be increasing over time, so the highs get higher and the lows get lower as time progresses, then the series likely has an exponential trend. On the other hand, if the variance of the first differce appears roughly constant, then the data most likely has a linear trend. Let's return to our US RGDP example:

```
DY <- diff(Y)
autoplot(DY) + ggtitle("Change in Real GDP") +
  ylab("Billions of 2012 dollars")
```

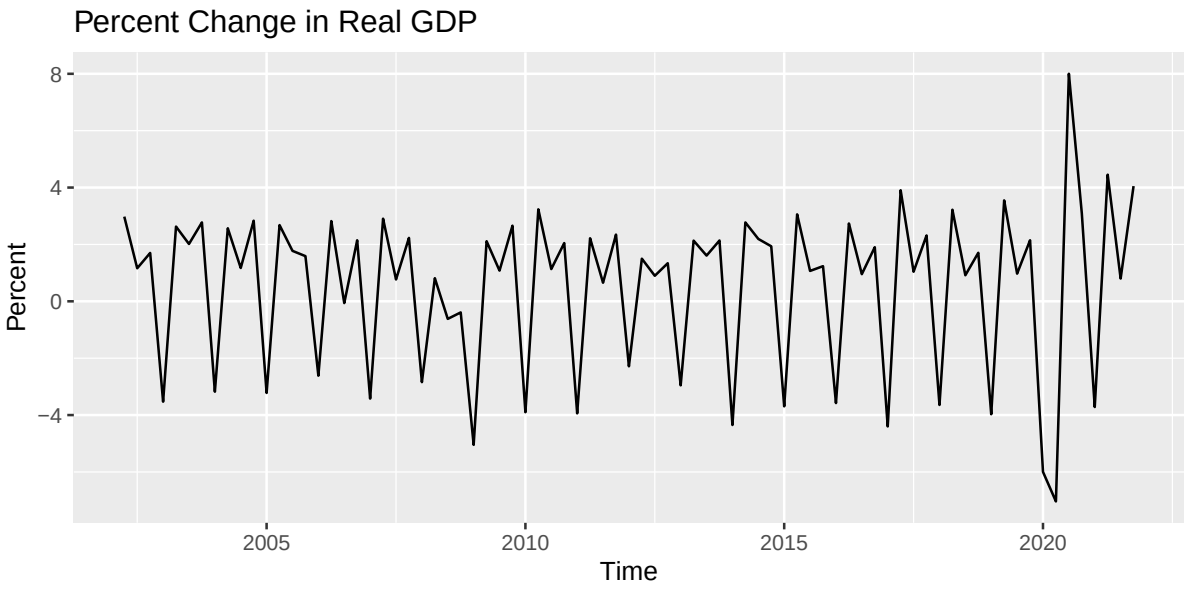


At the end of the series, the highest values are substantially higher than they were at the beginning of the series (roughly $150 billion in the late 2010's compared to $100 billion in the early 2000's), and the lowest values are substantially lower (roughly -$175 billion in the late 2010's compared to roughly -$100 in the early 2000's). The increasing spread in the first difference signals that this series likely has an exponential trend.[1] Therefore, we should take the natural log of the series prior to differencing it. This will put the series in "percent change".[2]

```
lnY <- log(Y)
Ystar <- diff(lnY)
Ystar <- 100*Ystar
autoplot(Ystar) + ggtitle("Percent Change in Real GDP") + ylab("Percent")
```

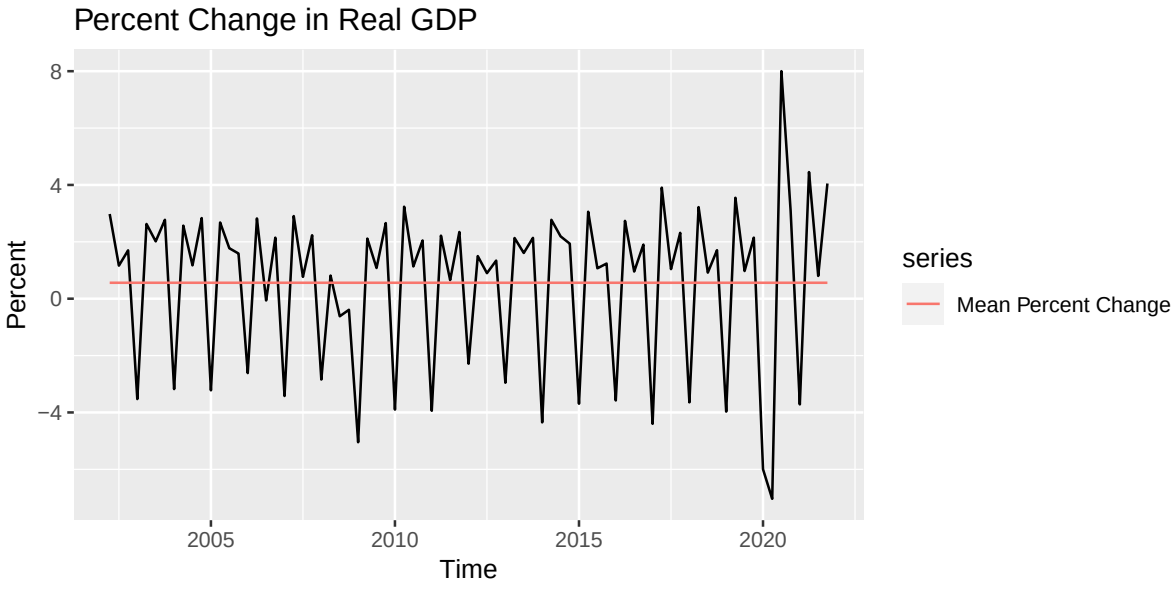
[1] In addition, we almost always hear about GDP discussed in "percentage change" rather than the change in dollars, signaling that most economists and financial jouralists treat RGDP as having an exponential trend.

[2] If you are reading the R code closely, you will notice that I also multiplied it by 100 (to convert to percent from decimal).



The swings still appear slightly larger towards the end of the sample, but to a much lesser extent then before. Although this data exhibits large fluctuations, appearing to decrease sharply at the beginning of every year, the mean now serves as a much more coherent measure of central tendency, as the data regularly cross the mean at the beginning, middle, and end of the sample:

```
Ystarm <- rep(mean(Ystar),length(Ystar))
Ystarm <- ts(Ystarm,start=c(2002,2),frequency=4)
autoplot(Ystar) + ggtitle("Percent Change in Real GDP") + ylab("Percent") +
  autolayer(Ystarm,series="Mean Percent Change")
```



2.5 Summarizing Trend Transformations

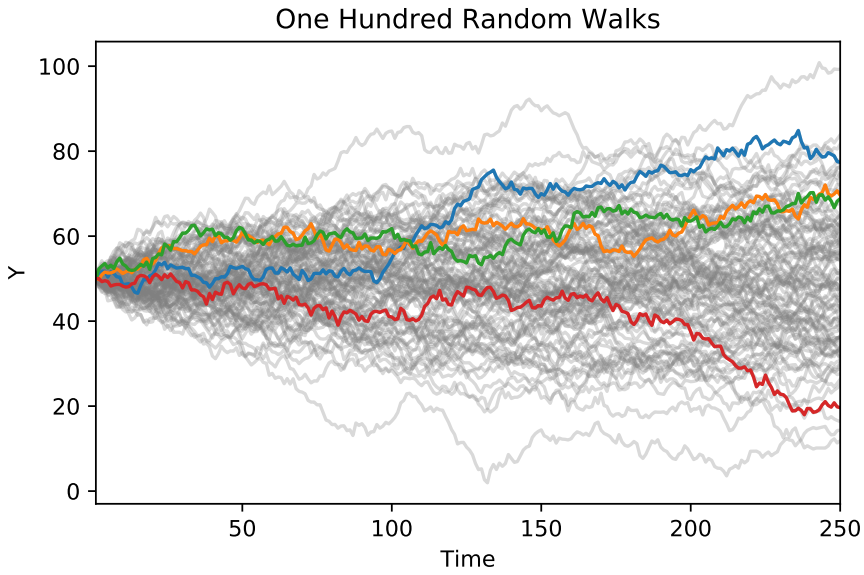
After performing the time plot on the original data, you should be able to distinguish whether or not the data has a long-run trend. This will inform your choice of transformation. Before proceeding to the next steps of checking the data for a unit root or seasonality, you should transform the data to remove the trend. In the next section, we will act as if you have already done this. Let the (possibly) transformed data be given by Y^*, where:

$$Y^* = \begin{cases} Y & \text{if series has no trend} \\ \Delta Y & \text{if series has a linear trend} \\ \Delta \ln(Y) & \text{if series has an exponential trend} \end{cases}$$

We will proceed by working with Y^* to find other time series properties of the data.

3 Transforming Unit Root Data

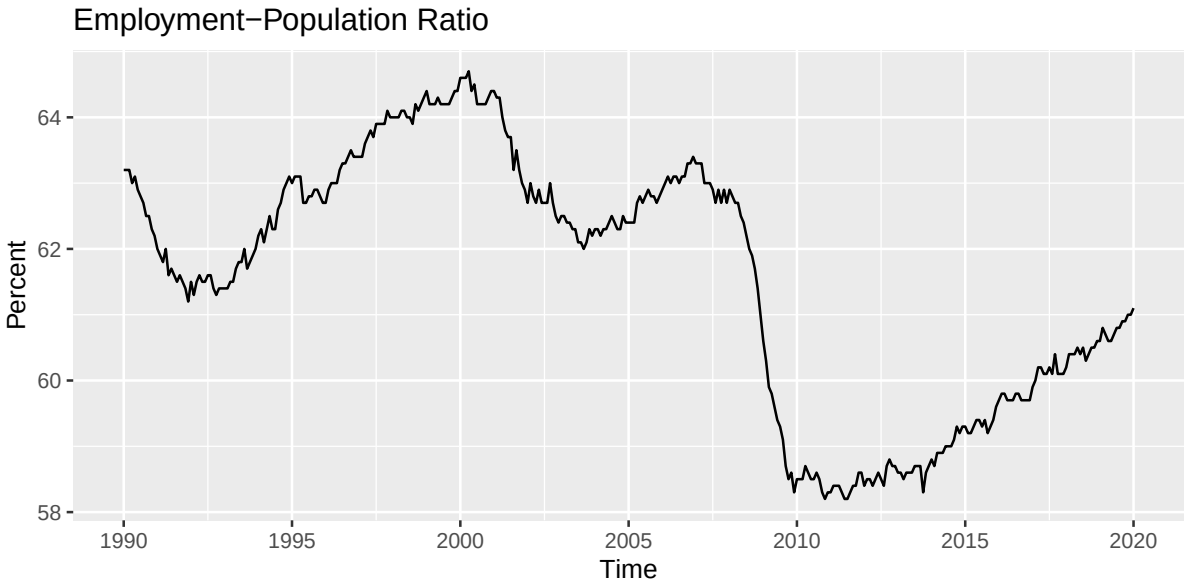
Assume you have already analyzed your raw data, Y, for a long-run trend and made any transformations as necessary. Next, you should analyze the possibly transformed data, Y^* for a "unit root." In data with a unit root, the data appears to wander around randomly. Unit root data may also appear to have short-run trends that then reverse. For example, here are time plots of simulated data series, each of which has a unit root:



For real life data that displays this behavior, here is the time plot of the employment-population ratio for the United States. This statistic represents the percentage of adults in the

United States who are working:

```
data_set <- read.csv("EMRATIO.csv")
EPOP <- ts(data_set[,2],start=c(1990,1),frequency=12)
autoplot(EPOP) + ggtitle("Employment-Population Ratio") + ylab("Percent")
```



We essentially see six separate short-run trends in this data, with a long-term downward trend lurking in the background after the year 2000. In the early 1990's, the economy was in recession, so fewer people were employed. From the mid 1990's through 2001, the economy was booming, and therefore adding jobs, and the ratio began increasing steadily. Then, there was a recession in 2001, followed by a slow but steady increase from roughly 2004-2008. Then the Great Recession hit and there was a precipitious dropoff between 2008-2010. Finally, since 2010, this ratio has been growing. However, it is still below its historical peak in 2000, as the population has aged and a larger fraction of Americans are retired today than in 2000. This complicated behavior is clearly not best summarized by a long-run trend, but in this context, summary statistics like the mean do not really seem appropriate, as the data has changed so much over time.

3.1 Identifying a Unit Root

To identify the presence of a unit root, we use the autocorrelation function. This function computes the correlation of data with past lags of itself. For example, autocorrelation at the first "lag" tells this historical correlation from every adjacent pair of data points, (y_t, y_{t-1}) and

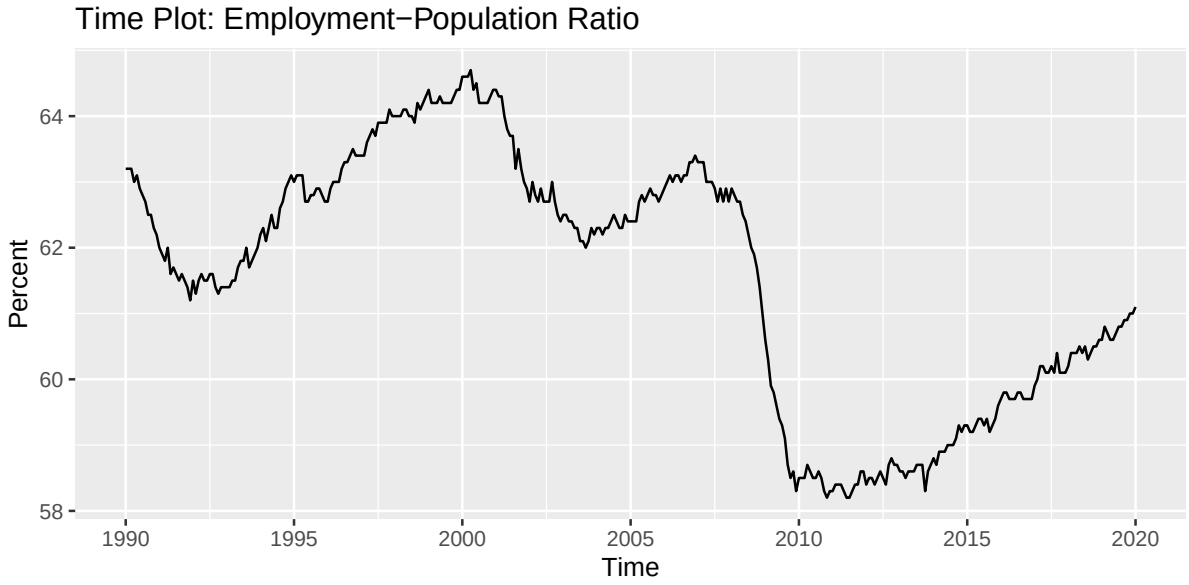
autocorrelation at the second "lag" tells the historical correlation from every pair of data points separated by two time periods, (y_t, y_{t-2}). If the data are strongly related through time, then this correlation will be very high. If the data are completely random so the past tells us nothing about the data today, this correlation will be very low. Mathematically, autocorrelation at the k^{th} lag can be computed as:

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}(y_t - \bar{y})^2}$$

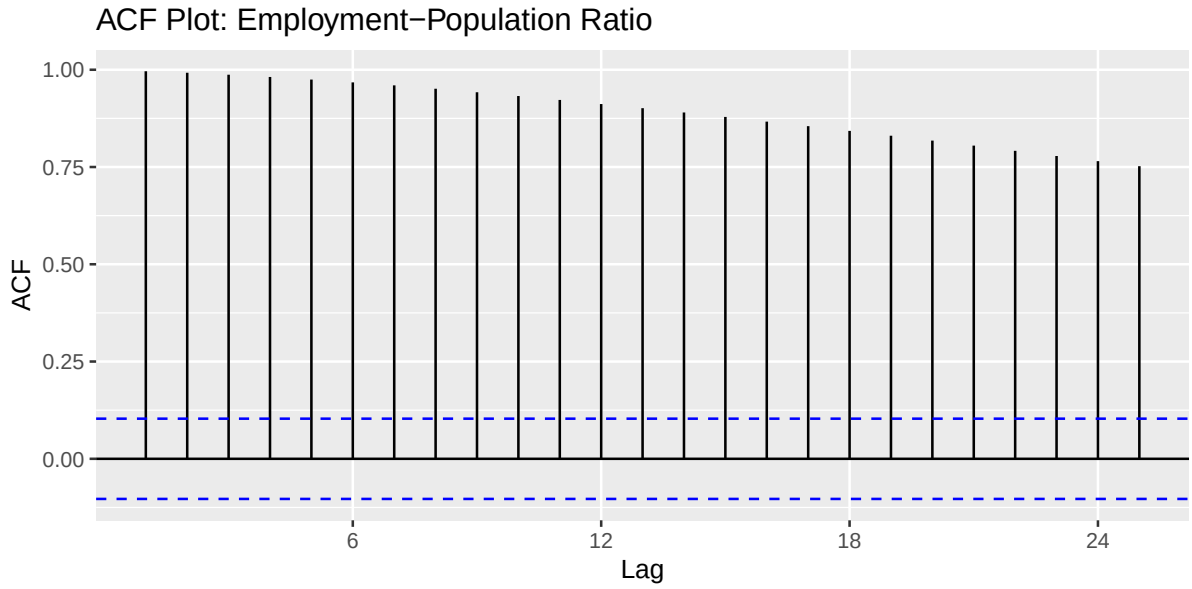
Like all correlation coefficients, the value will always lie between -1 and +1, with a value of 0 indicating no autocorrelation at the k^{th} lag. If a series has a unit root, the correlation will be high at ALL past lags. If the series does NOT have a unit root, the correlation may be high at first (or not), but will quickly drop to zero.

One easy way to visually inspect data for a unit root is to look at its autocorrelation function plot (ACF Plot). This plots the correlation coefficient at many lags (typically 16-30 depending on the length of the series and the frequency of the data). For example, for the employment population ratio data:

```
autoplot(EPOP) + ggtitle("Time Plot: Employment-Population Ratio") +
  ylab("Percent")
```



```
ggAcf(EPOP) + ggtitle("ACF Plot: Employment-Population Ratio")
```



The correlation coefficient is very high all the way out to the 25th lag. Since it always lies outside of the blue dashed lines, it is always statiscially significantly different from 0 at the 95% confidence level. This strongly indicates the presence of a unit root.

3.2 Transforming Data to Remove Unit Root

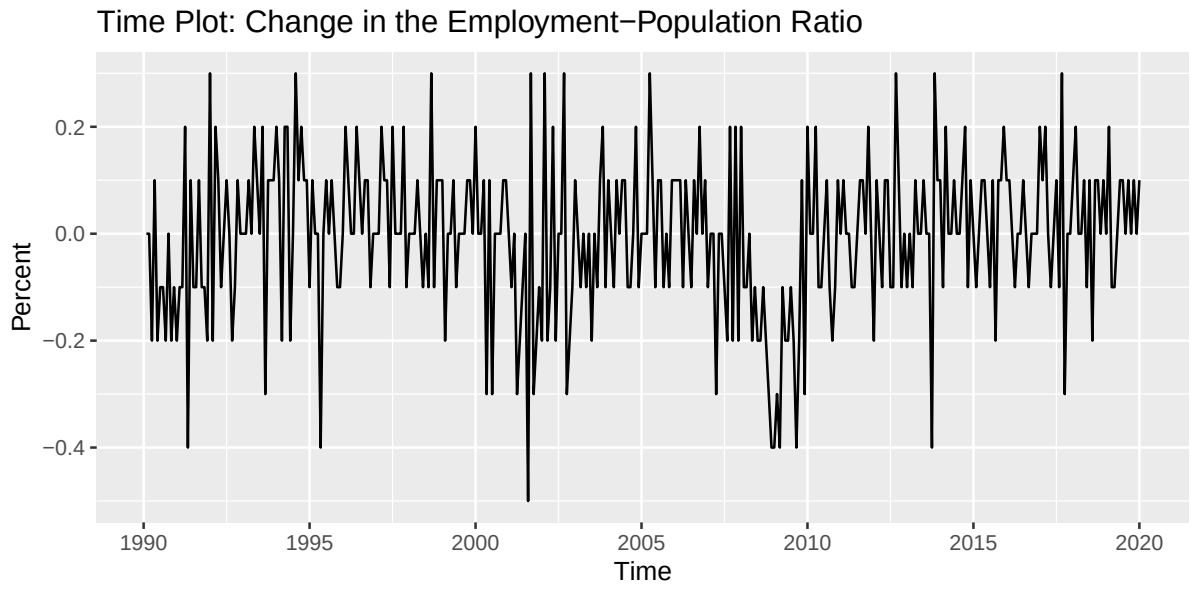
Again, if data has a unit root, its summary statistics are not meaningful. In order to proceed, we should transform data that has a unit root to remove it. To remove a unit root, we can simply first difference the series, as we did if there were a linear trend:

$$\Delta Y^* = \begin{bmatrix} y_2^* - y_1^* \\ y_3^* - y_2^* \\ \vdots \\ y_T^* - y_{T-1}^* \end{bmatrix}$$

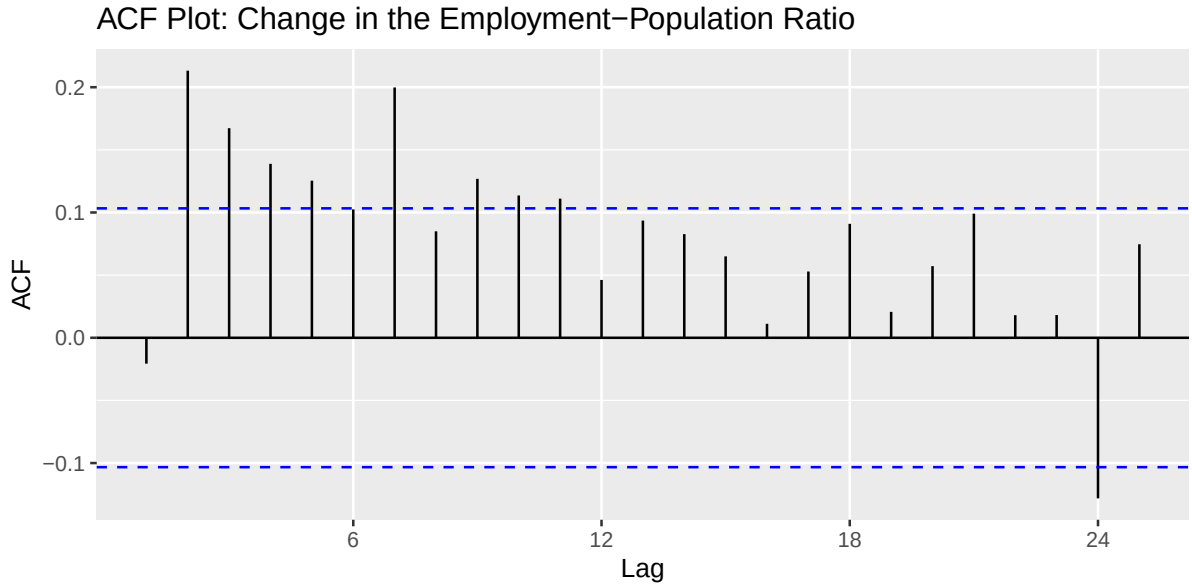
where we are using y^* to represent the fact that the data may have already been transformed to remove the trend.

For the employment-population ratio data, we now have:

```
DEPOP <- diff(EPOP)
autoplot(DEPOP) +
  ggtitle("Time Plot: Change in the Employment-Population Ratio") +
  ylab("Percent")
```




```
ggAcf(DEPOP) +
  ggtitle("ACF Plot: Change in the Employment-Population Ratio")
```



Although the correlation is still fairly high at the 2nd through 11th lags, the correlation is much lower, never rising above 0.2, and drops toward 0 for lags 12 any beyond. Therefore this transformed data does NOT appear to have a unit root. Once we first-difference the data, this is typical. However, in some rare cases you will need to continue differencing the data further

in order to remove the unit root (e.g. compute the double difference or the "change in the change").

3.3 Summarizing Unit Root Transformations

Let Y^* be the data after transforming to remove any trend. Then, let $Y^{*'}$ denote the data after transforming to remove both trend and unit root. We have:

$$Y^{*'} = \begin{cases} Y^* & \text{if series has no unit root} \\ \Delta Y^* & \text{if series has a unit root} \end{cases}$$

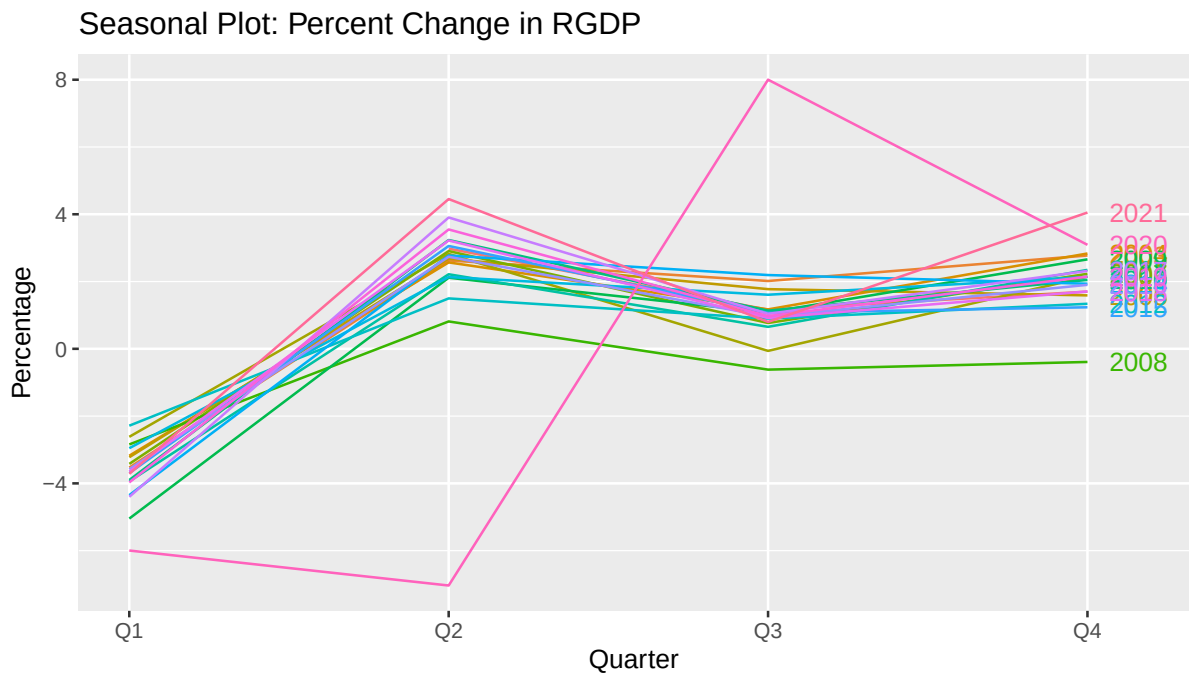
4 Identifying Seasonality

Many data sets in business and economics have seasonality. That is, they are consistently higher in some periods within the year than in other periods within the year. For example, retail sales are typically higher during the run-up to Christmas, so production of goods and services usually ramps up during the fourth quarter (October-December). Then, from January-March, total sales typically decline, and therefore production typically declines during this period as well.

While seasonality is sometimes easy to spot in a time plot, sometimes it is hard to tell if the fluctuations occur regularly (during the same time periods) or irregularly. Therefore, we consider two additional types of plots: seasonal plots and seasonal subseries plots. A seasonal plot plots the data from each year across all the time periods within that year. Each year gets its own line. A seasonal subseries plot plots all the values from one period across all years, with each period getting its own line. Seeing the different plots might help if the verbal description is confusing.

First, a seasonal plot for the percent change in US RGDP:

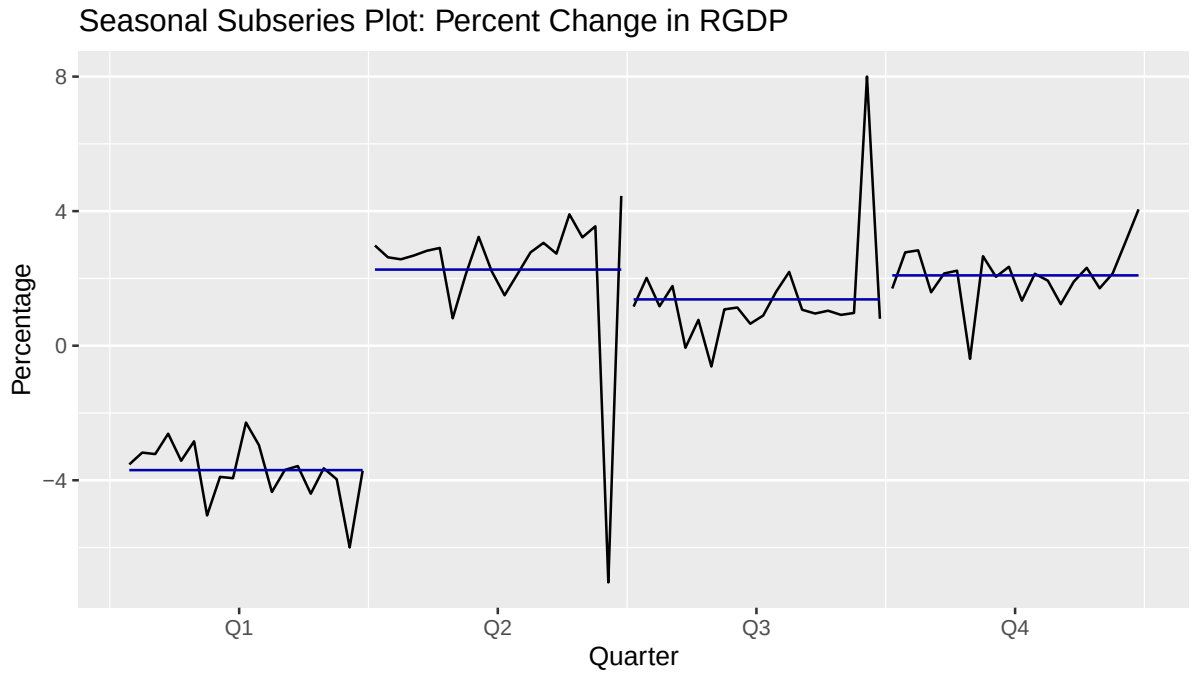
```
Ystarp <- Ystar # No transformation for unit root since no unit root
ggseasonplot(Ystarp,year.labels = TRUE) +
  ggtitle("Seasonal Plot: Percent Change in RGDP") +
  ylab("Percentage")
```



Again, each year gets its own line. For example, 2002 is one line, 2003 is another, etc. Seasonal plots can become hard to read as the data accumulates and the numer of years in your sample becomes large.

Let's look at the same data on a seasonal subseries plot instead:

```
ggsubseriesplot(Ystarp) +
  ggtitle("Seasonal Subseries Plot: Percent Change in RGDP") +
  ylab("Percentage")
```

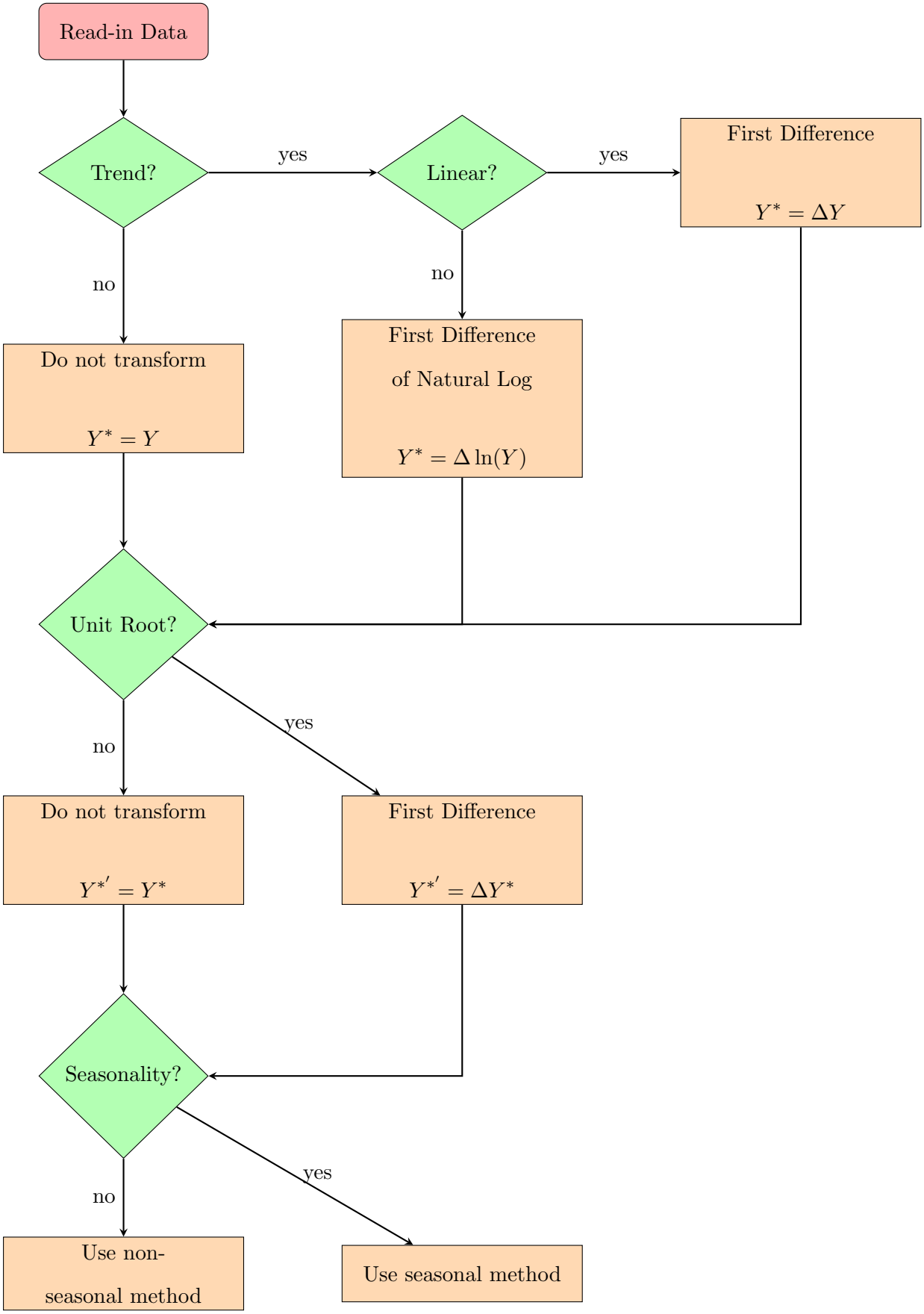


In the plot above, the first "subseries" is the first quarter of every year, all plotted together over time. These subseries for each period are like mini time plots — one for each period of data across all years. The horizontal line during each period is the average value in that period. For example, we can see that the average change in RGDP in the first quarter is rouhgly -3%. Since the mean in the first quarter is so far below the mean in any other quarter, this indicates that the data has seasonality — it is predictably different in at least one quarter compared to the rest of the year.

4.1 Transformations for Seasonality

The methods we learn in this class can be generalized to handle data with seasonality. Therefore, we will not need to transform seasonal data. Instead, we will make a note of whether or not the data appears to have seasonality. This will inform our choice of forecasting model. In other words, if the data has seasonality we need to use the version of the forecasting model that allows for seasonal data.

5 Overview


Read-in Data
Trend?
yes
Linear?
yes
First Difference
$Y^* = \Delta Y$
no
no
Do not transform
$Y^* = Y$
First Difference
of Natural Log
$Y^* = \Delta \ln(Y)$
Unit Root?
yes
no
Do not transform
$Y^{*'} = Y^*$
First Difference
$Y^{*'} = \Delta Y^*$
Seasonality?
yes
no
Use non-
seasonal method
Use seasonal method

6 Exercises

1. Use FRED to download the monthly unemployment rate in Minnesota beginning in 1990 (NOT seasonally adjusted).

   (a) Use this data to create a Time Plot. Does the series have a trend? If so, what type?

   (b) If you found the data has a trend, transform the data before continuing.

   (c) Use an ACF plot to plot the (possibly) tranformed series. Does it appear the (transformed) series has a unit root?

   (d) If you found the (transformed) data has a unit root, further transform it.

   (e) Use both types of seasonality plots on the (transformed) data. Does the series appear to have seasonality?

   (f) Summarize your findings about trend, unit root, and seasonality.

2. Suppose you had data on a monthly economic time series, and you found that the autocorrelation at the first lag was -0.52

   (a) Intuitively, what does negative autocorrelation at the first lag mean? (i.e. explain in "plain" language)

   (b) Given that this is economic data, are you surprised to see a negative autocorrelation?

   (c) Suppose the autocorrelation at the 12^{th} lag was 0.80. What might you be able to infer about the data?