Day 2 - Visualizing time series

Introduction

The purpose of today's lecture is to practice manipulating time series data in R. To guide our exploration, we will use a data set describing the number of international passenger bookings (in thousands) per month on an airline (Pan Am) in the United States for the period of 1949-1960.

```
# load data and rename
data(AirPassengers)
ap <- AirPassengers

# inspect the ap object
str(ap)
```

```
 Time-Series [1:144] from 1949 to 1961: 112 118 132 129 121 135 148 148 136 119 ...
```

```
class(ap)
```

```
[1] "ts"
```

> **Note**
>
> The `ts` class is often used to store time series objects in R. It is a succinct class of objects that enables the use of a number of helpful functions, including `start`, `end`, `frequency`, and `aggregate`.

```
start(ap); end(ap); frequency(ap)
```

```
[1] 1949    1
```

```
[1] 1960   12
```

```
[1] 12
```

We can also create `ts` objects from existing time series. For example, the code below converts the `vt_temps` data frame into a `ts` object.

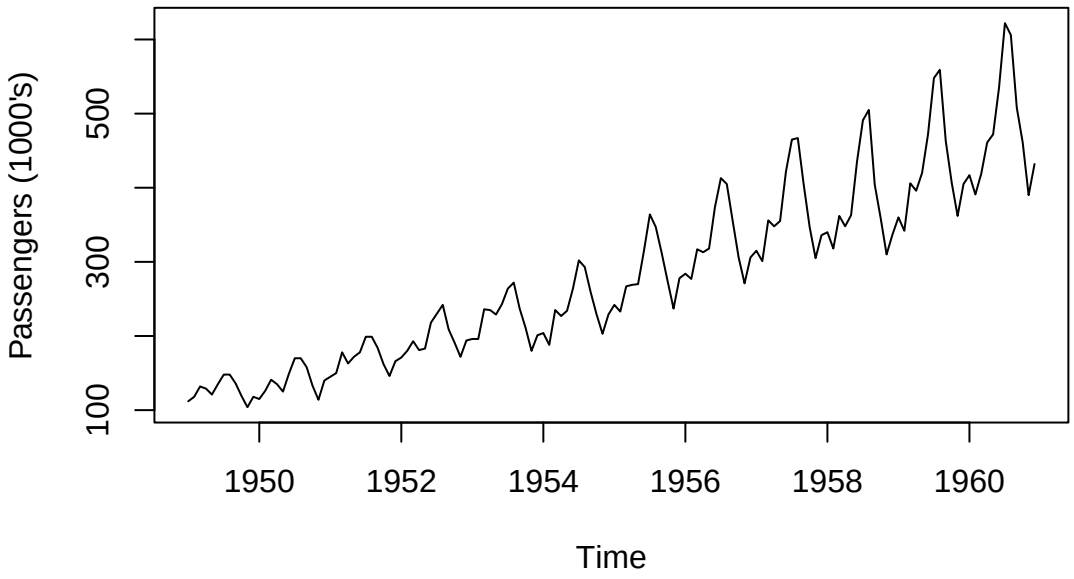
```r
# library packages
library(tidyverse);library(lubridate)

# read in data
vt_temps <- readr::read_csv("vt_temps.csv")

# create year and month variables and filter to post 1979
vt_temps <- vt_temps |>
  mutate(
    year = year(dt),
    month = month(dt)
  ) |>
  filter(year >= 1980)

# create ts object
vt_ts <- ts(
  data = vt_temps$AverageTemperature,
  start = c(1980, 1), end = c(2013, 9),
  frequency = 12
)
```

Note

The `ts` function requires specification of the start and end date for the series. For montly data, you are required to input both the starting year and month. If you do not include the month, you are likely to exclude some of the series!

Visualizing time series

```
plot(ap, ylab = "Passengers (1000's)")
```



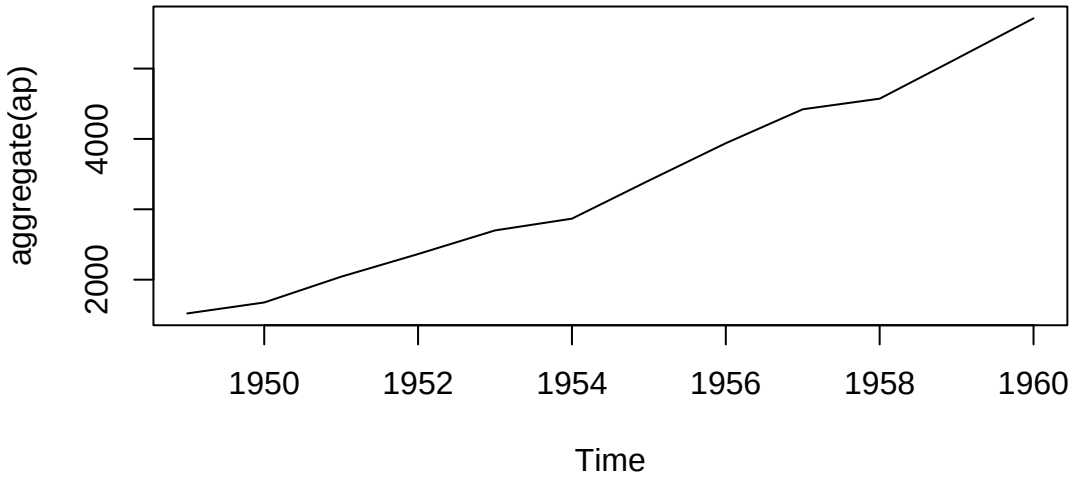
What is the first step in every time series analysis?

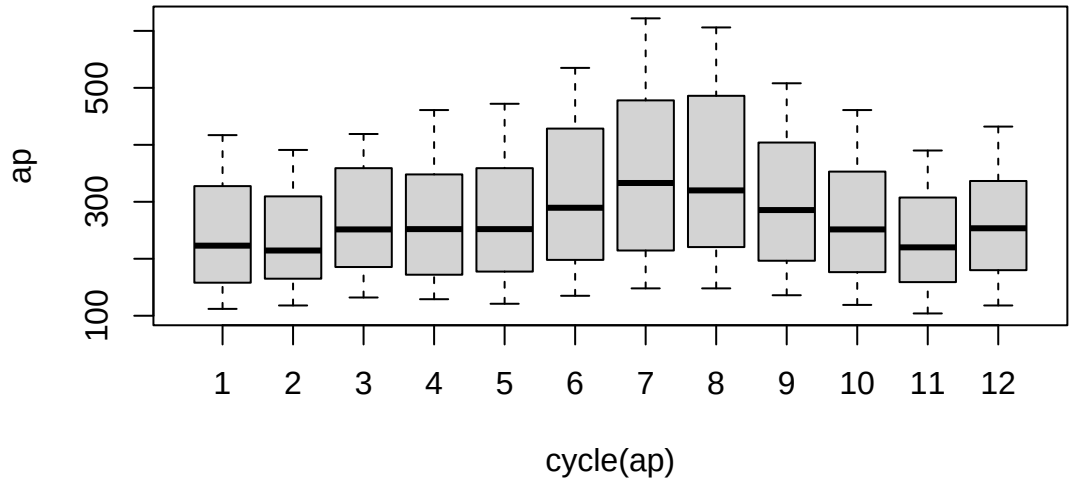
What is the difference between **trend** and **seasonal variation**?

If the trend and seasonal components of a time series are known, do we expect to be able to perfectly predict the series? Why or why not?

> **Note**
>
> The `aggregate` and `cycle` functions can be used on a `ts` object to visualize the trend and seasonal components, respectively.

```
par(mfrow = c(2, 1))
plot(aggregate(ap))
boxplot(ap ~ cycle(ap))
```

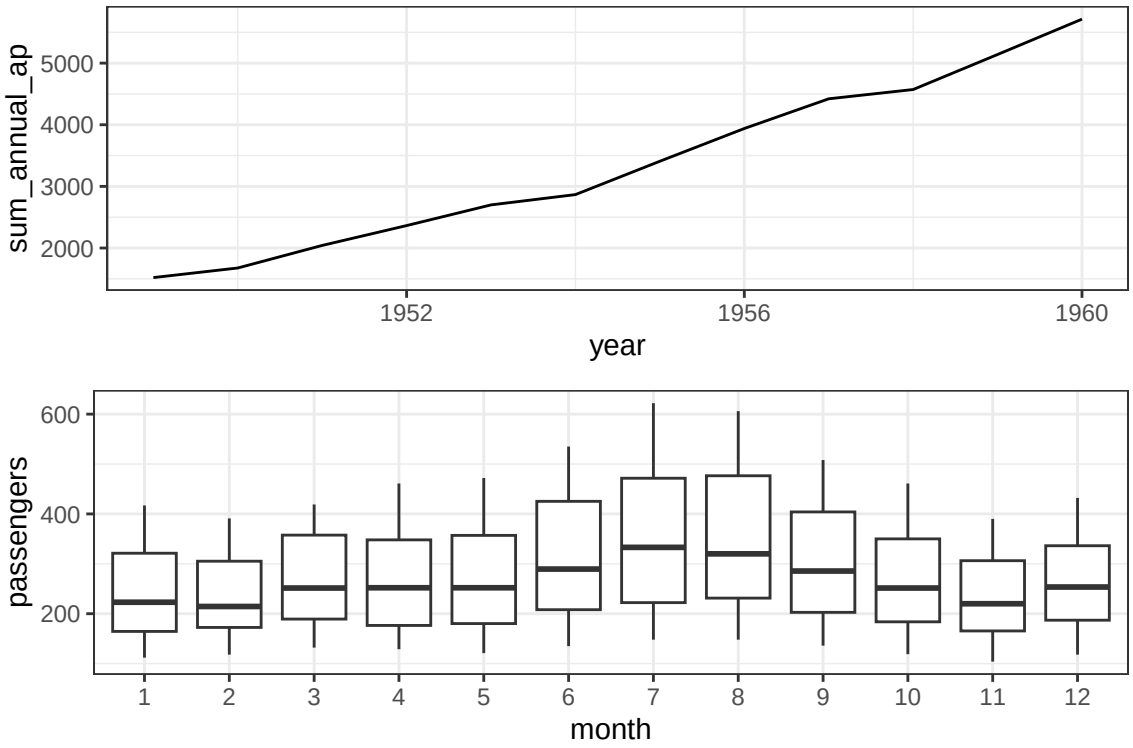




```r
par(mfrow = c(1,1))
```

What are each of the aggregate and cycle functions doing?

```r
ap_tbl <- tibble(
  passengers = c(ap),
  month = rep(1:12, length(1949:1960)) |> factor(),
  year = rep(1949:1960, each = 12)
)

p1 <- ap_tbl |>
  group_by(year) |>
  summarize(sum_annual_ap = sum(passengers)) |>
  ggplot() +
  geom_line(aes(x = year, y = sum_annual_ap)) +
  theme_bw()

p2 <- ap_tbl |>
  ggplot() +
  geom_boxplot(aes(x = month, y = passengers)) +
  theme_bw()

gridExtra::grid.arrange(p1, p2, ncol = 1)
```



Multiple series

Often times, it is of interest to compare multiple time series. To illustrate how this can be done in R, and provide some cautionary advice, we pull in data on the monthly supply of electricity (millions of kWh), beer (Ml), and chocolate-based production (tonnes) in Austrailia over the period of January 1958 to December 1990, courtesy of the Australian Bureau of Statistics.

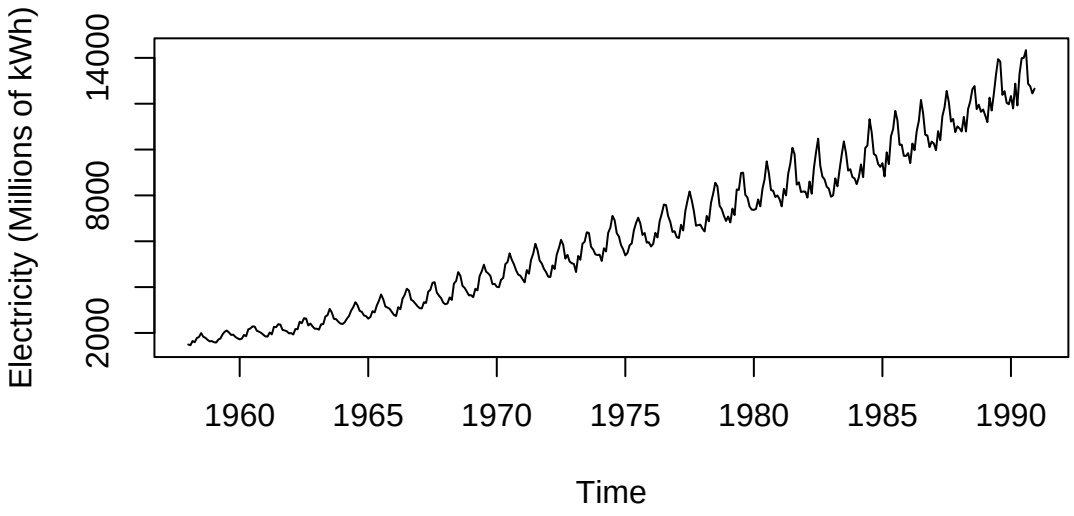
```
# install a package from Github, comment out after running
# devtools::install_github("speegled/tswrdata")
library(tswrdata)

# load data
data(cbe)
```

1. Create three `ts` objects, named `elect.ts`, `beer.ts`, and `choc.ts`, containing the electricity, beer, and chocolate time series, respectively.

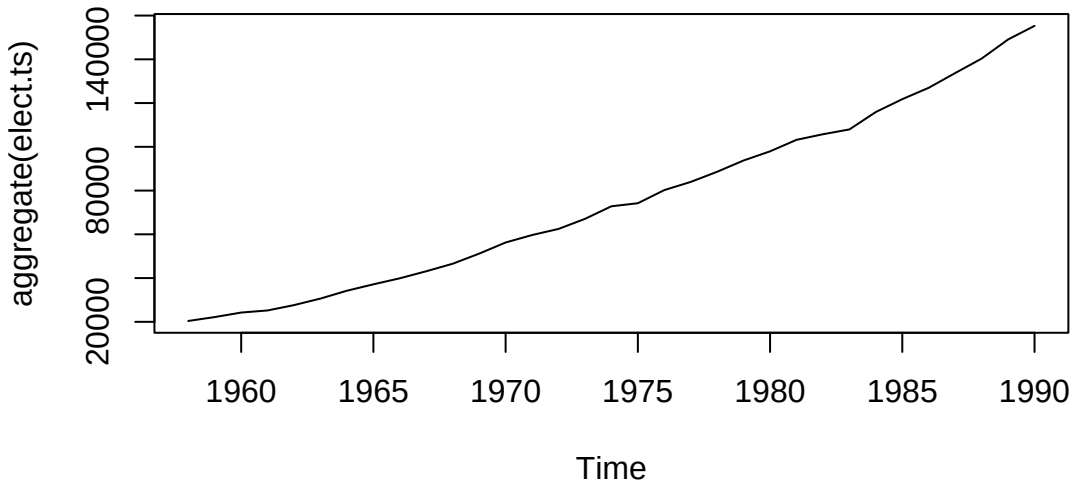
```
elect.ts <- ts(cbe[,3], start = c(1958, 1), end = c(1990, 12), freq = 12)
beer.ts <- ts(cbe[,2], start = c(1958, 1), end = c(1990, 12), freq = 12)
choc.ts <- ts(cbe[,1], start = c(1958, 1), end = c(1990, 12), freq = 12)
```

2. Plot the `elect.ts` series and describe the series in terms of trend and seasonality.

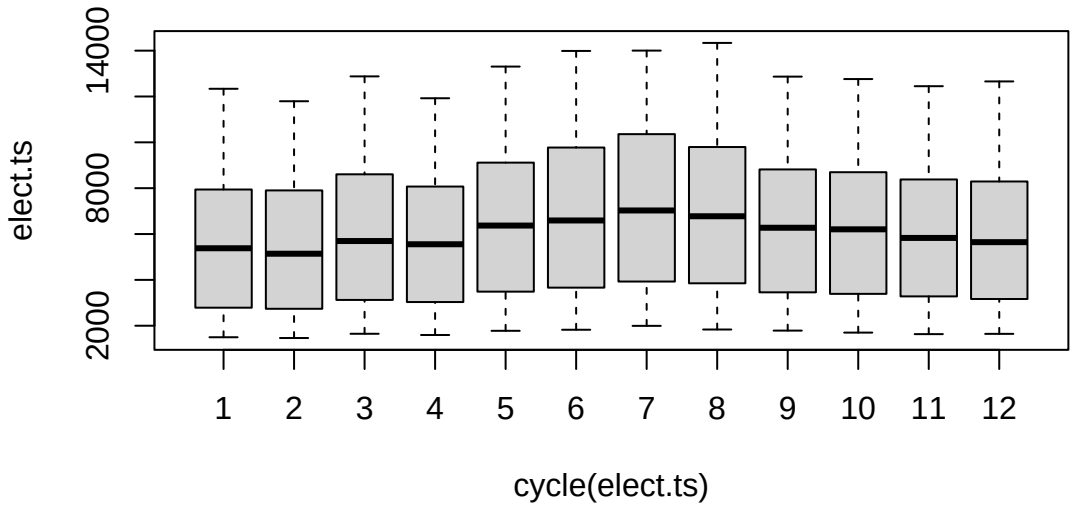
```
plot(elect.ts, ylab = "Electricity (Millions of kWh)")
```



```
plot(aggregate(elect.ts))
```




```
boxplot(elect.ts ~ cycle(elect.ts))
```

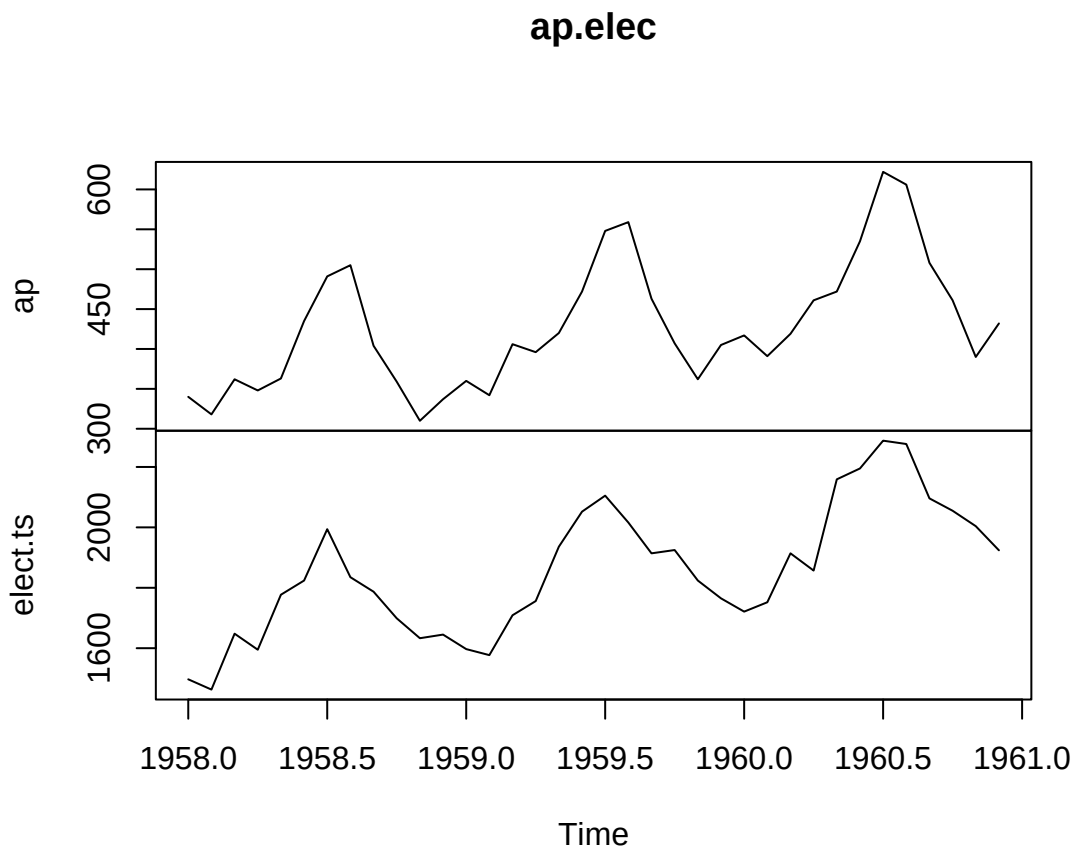


The series has a strong, positive, increasing trend and evidence of seasonal variation within each year, with greater electricity use in the winter months (recall that Australia is in the southern hemisphere).

3. Create a new object, called `ap.elec`, that represents the intersection between the air passenger data and electricity data and plot it. Describe what you see.

> **Note**
>
> The `ts.intersect` function can be used to find the intersection between two overlapping time series. Run `help("ts.intersect")` for more information.

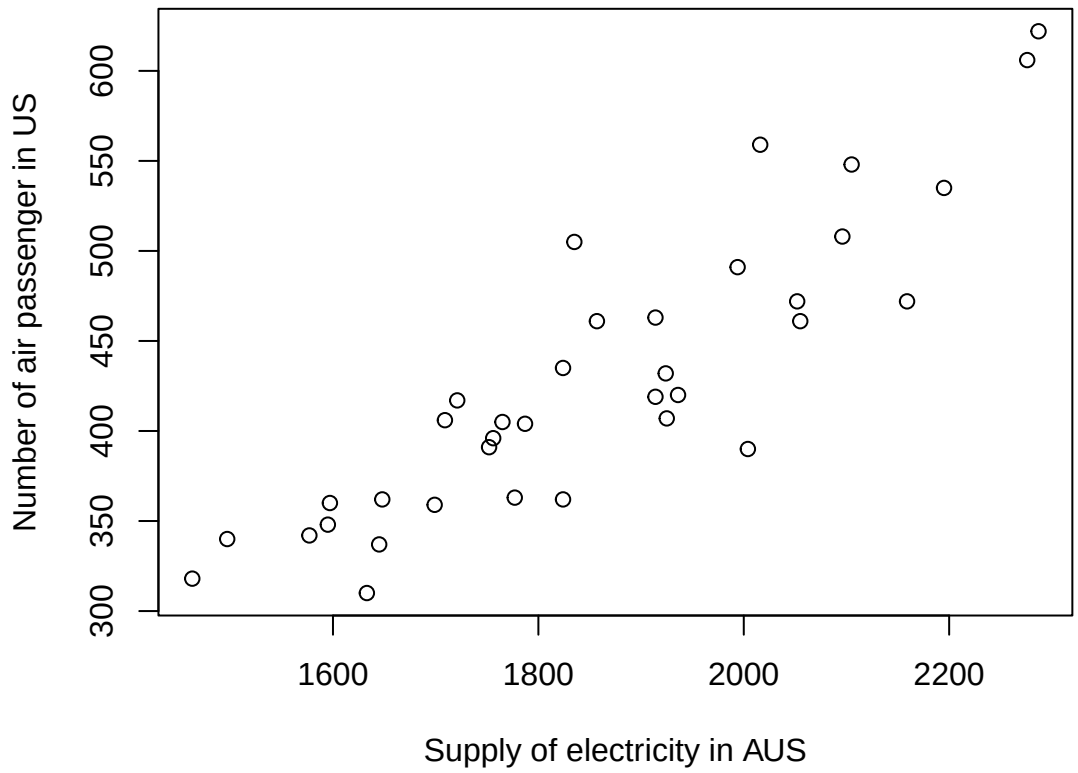
```
ap.elec <- ts.intersect(ap, elect.ts)
plot(ap.elec)
```



The two series are remarkably similar, sharing both an increasing trend and similar seasonal dynamics.

4. The code below calculates the correlation between the two series (more on this in week 3) and creates a scatterplot of the number of air passengers in the United States against the monthly supply of electricity in Australia. Both the correlation and scatterplot suggest a strong, positive, linear association between the US air passengers and AUS monthly supply of electricity. Is it reasonable to conclude that the increasing supply of electricity in Australia caused the increasing number of air passengers in the US?

```
ap_subset <- ap.elec[,1]
elect_subset <- ap.elec[,2]
plot(
  c(ap_subset) ~ c(elect_subset),
  xlab = "Supply of electricity in AUS",
  ylab = "Number of air passenger in US"
)
```




```
cor(c(ap_subset), c(elect_subset))
```

```
[1] 0.8841668
```

No! It is likely that the association is driven by the increasing prosperity and technology in both countries. The moral of this story is correlation does not imply causation, and we should be wary of building associations based on the similarity between time series.