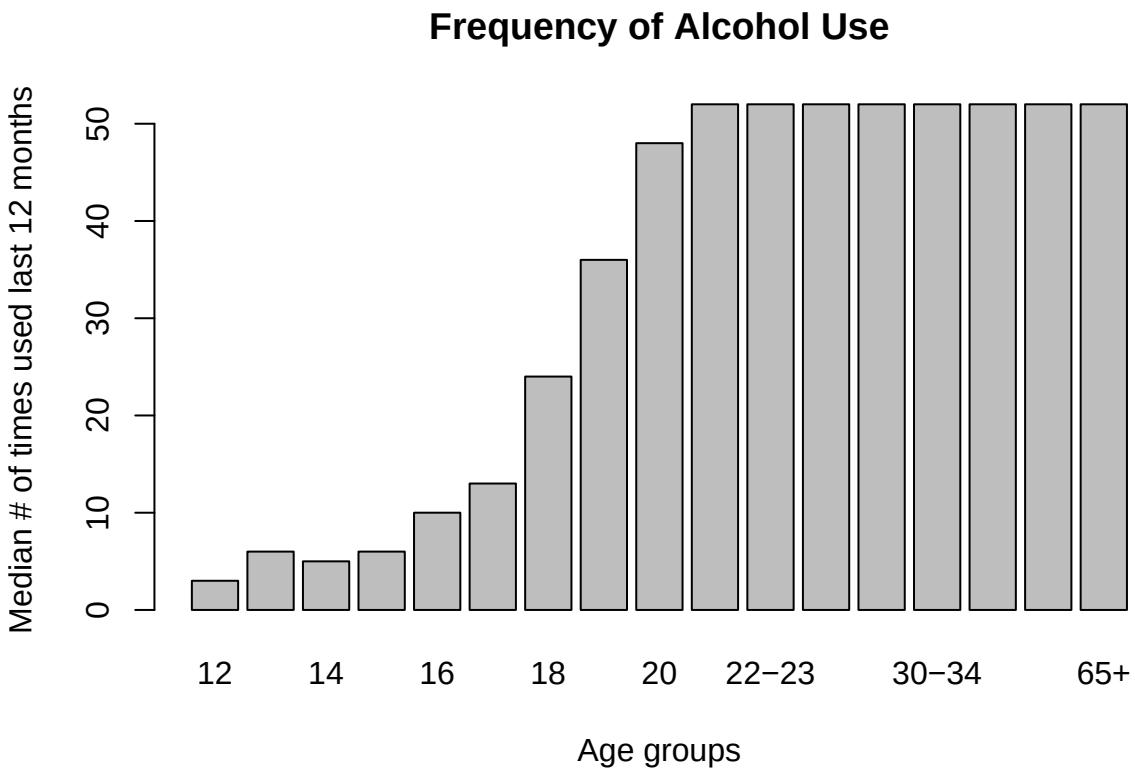
1st Pair Assignment

Chris Kardish and Marie Agosta

September 30, 2016

This is the first pair assignment from Chris Kardish and Marie Agosta. The first data set comes via FiveThirtyEight; the second is core R and is sourced from a second script. First, the FiveThirtyEight stuff. It is a survey of drug use by age group carried out by the US Substance Abuse and Mental Health Services Administration.

There are many drugs to choose from, but for purposes of brevity we'll only cover the descriptive statistics of a few, starting with the lighter stuff.



You notice immediately from the barplot that this particular variable is a bit odd: the median rises steadiliy with age before completely leveling off at 52 at age 21. That's likely a problem with SAMHSA's or FiveThirtyEight's reporting and it will affect the following descriptive statistics.

Mean:

```
## [1] 33.35294
```

Median:

```
## [1] 48
```

We immediately notice that the mean is far smaller than the median, indicating a strong left skew. Standard Deviation:

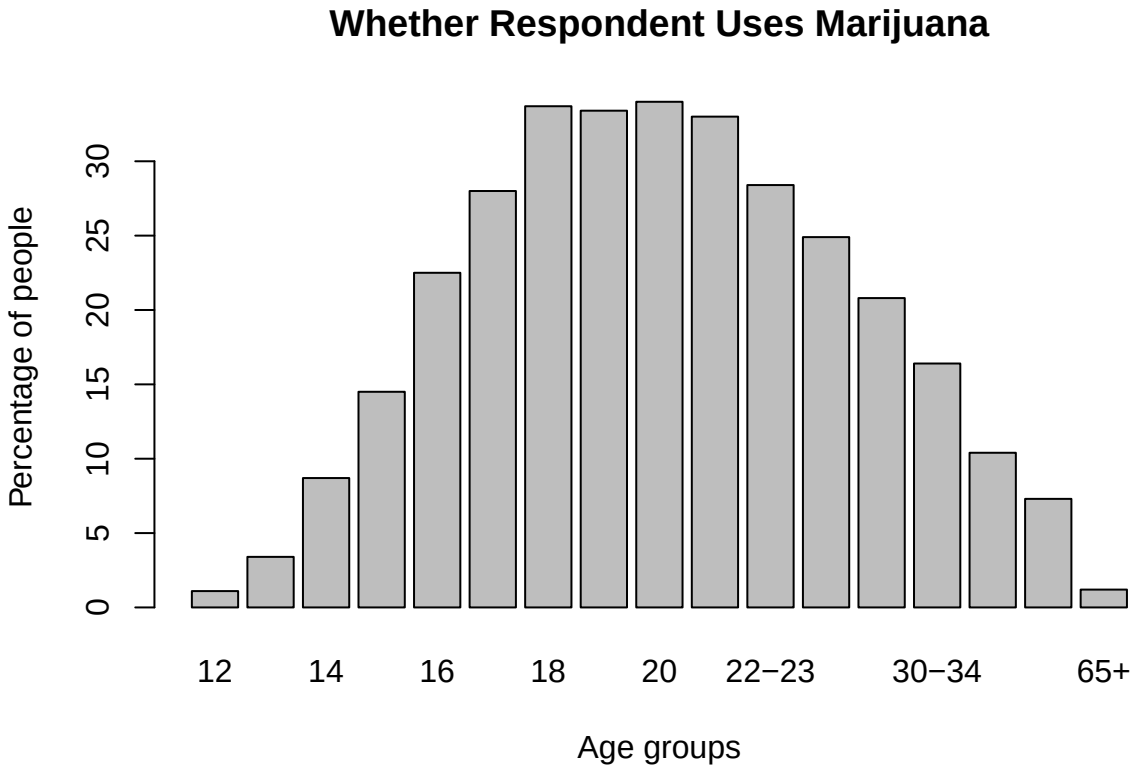
```
## [1] 21.31883
```

That seems like a large standard deviation, but perhaps not, given the units and clear differences between age groups that you would naturally expect (12 year olds don't tend to do a lot of drinking)

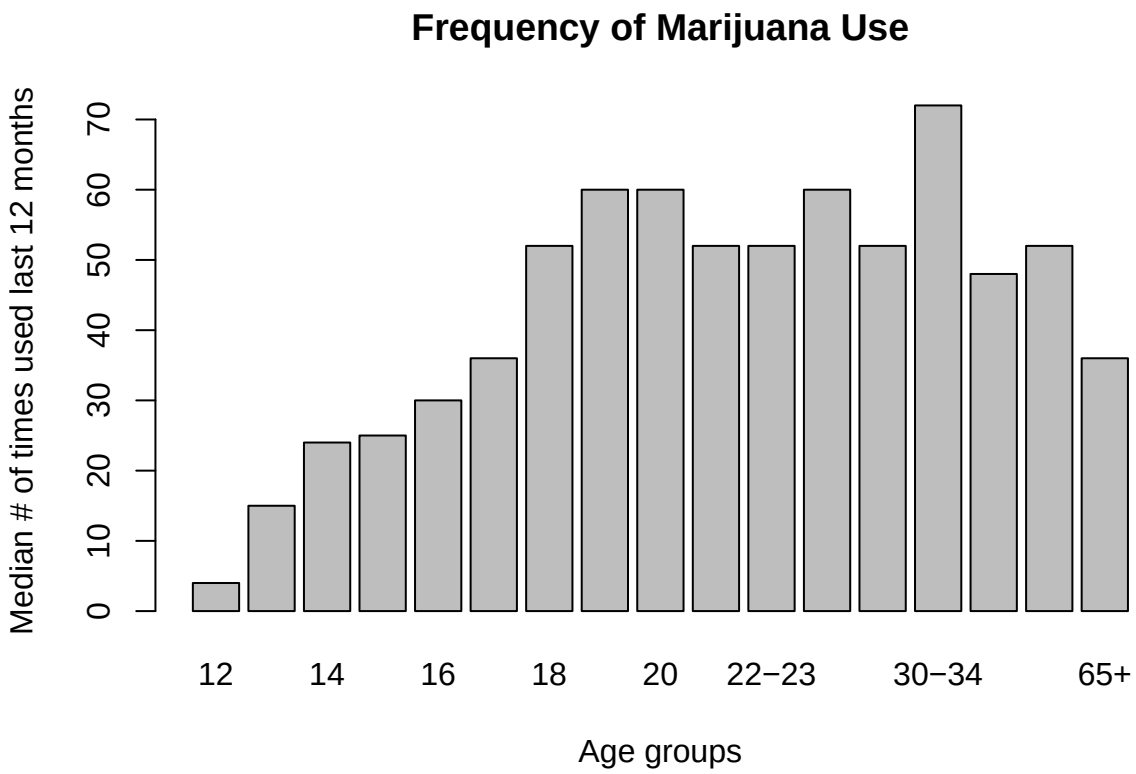
Range:

```
## [1]  3 52
```

Moving on to marijuana. Here's a plot for percentage of people in a given age group who have used marijuana in the past 12 months.



Now here's a barplot for marijuana use by age, defined as frequency of use in the last 12 months.This of course only applies to people who identify as users.



Initial visual hypothesis: as you age you're less likely to be a marijuana user, but if you do remain a user into your 30s you're likely to smoke as much as the kids or even more. But obviously this idea would have to be tested for statistical significance.

Let's close things out by answering the following question: is the average number of heroin users larger than the average number of meth users? Let's find out!

Percentage of those in any age group who used heroin in the past 12 months:

```
## [1] 0.3529412
```

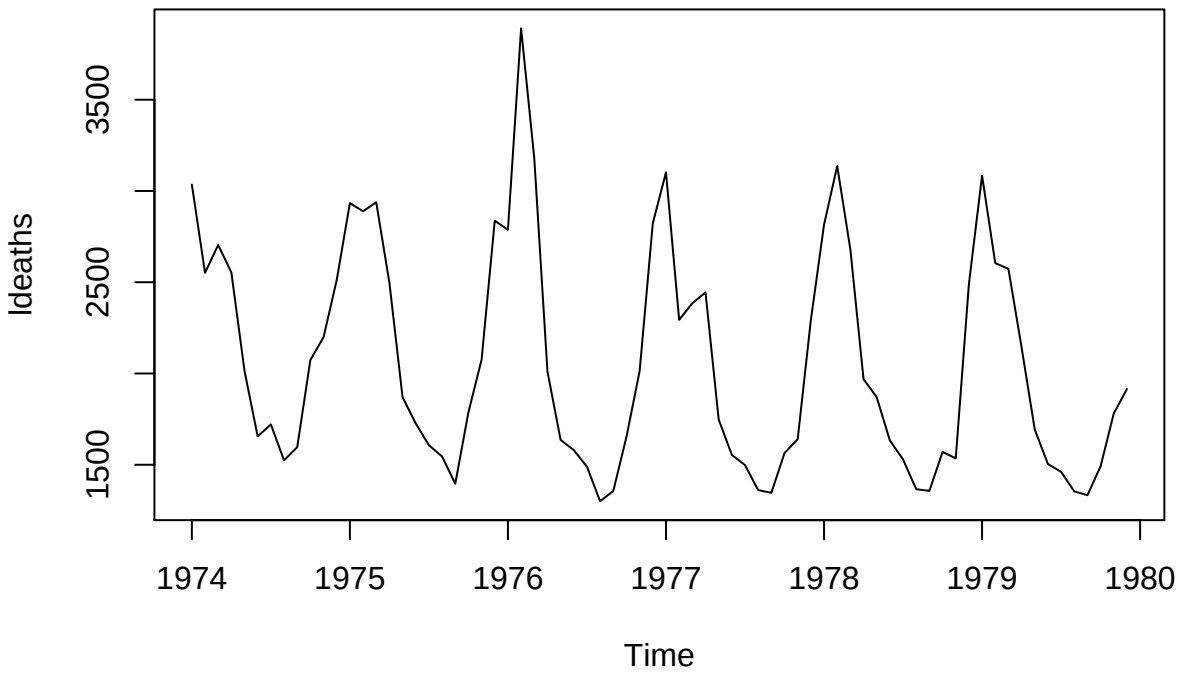
Percentage of those in any age group who used meth in the past 12 months:

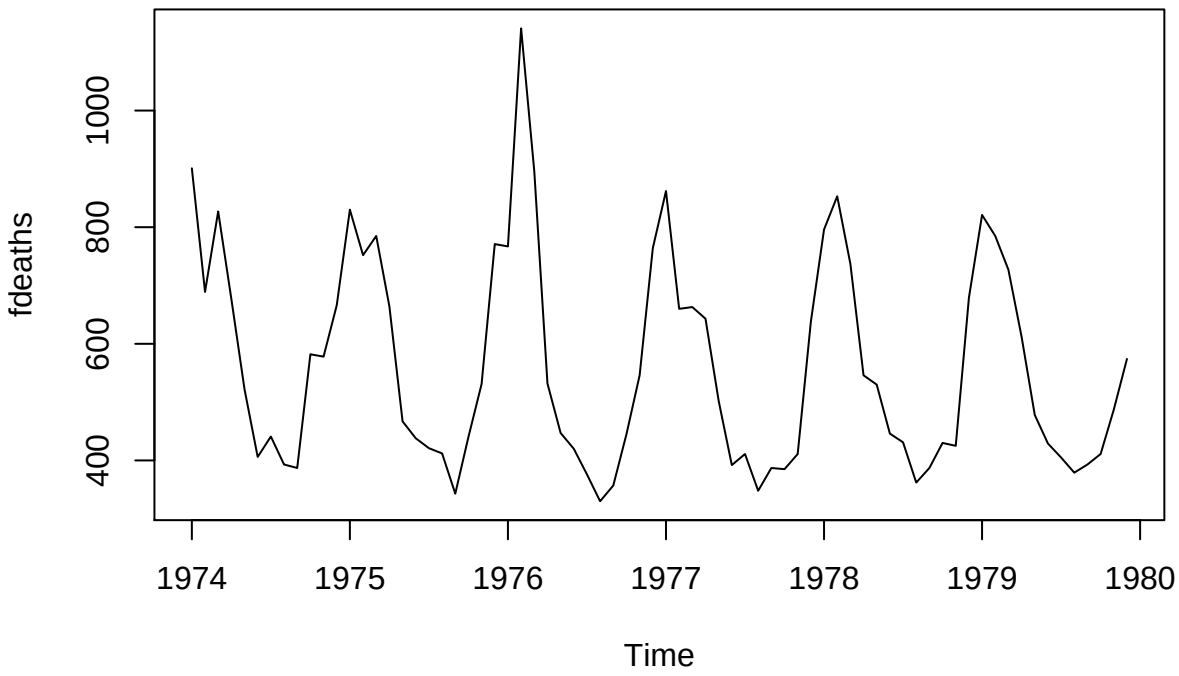
```
## [1] 0.3823529
```

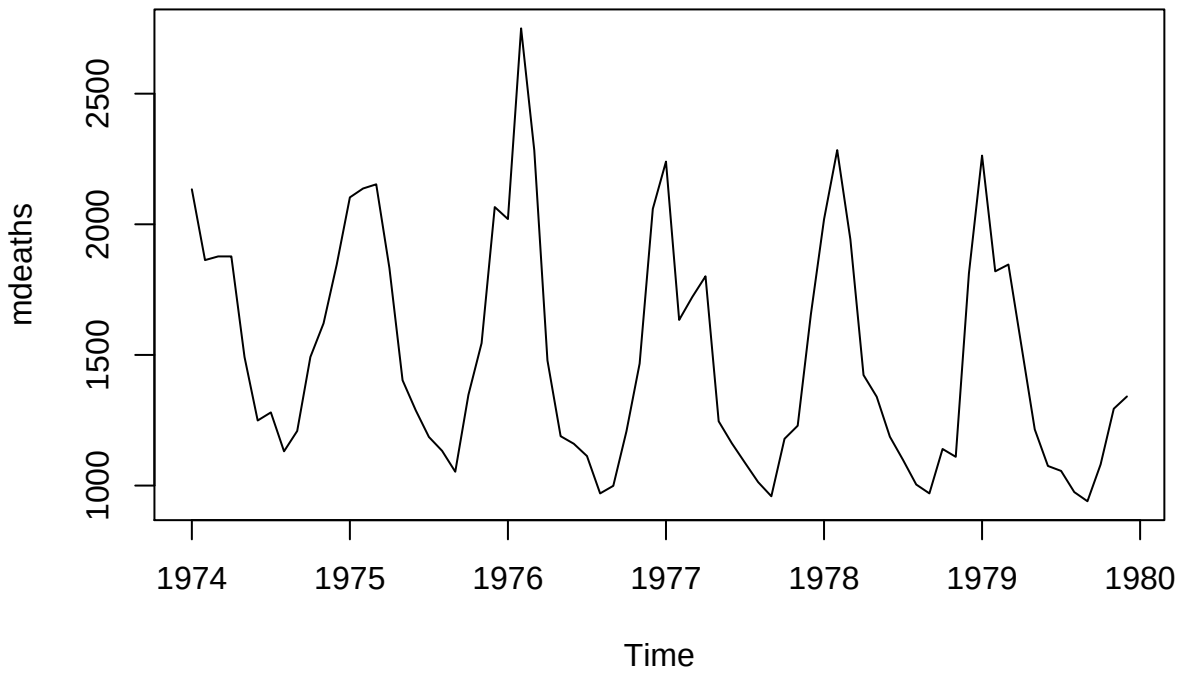
It appears meth remains slightly more popular, at least from this vantage point.

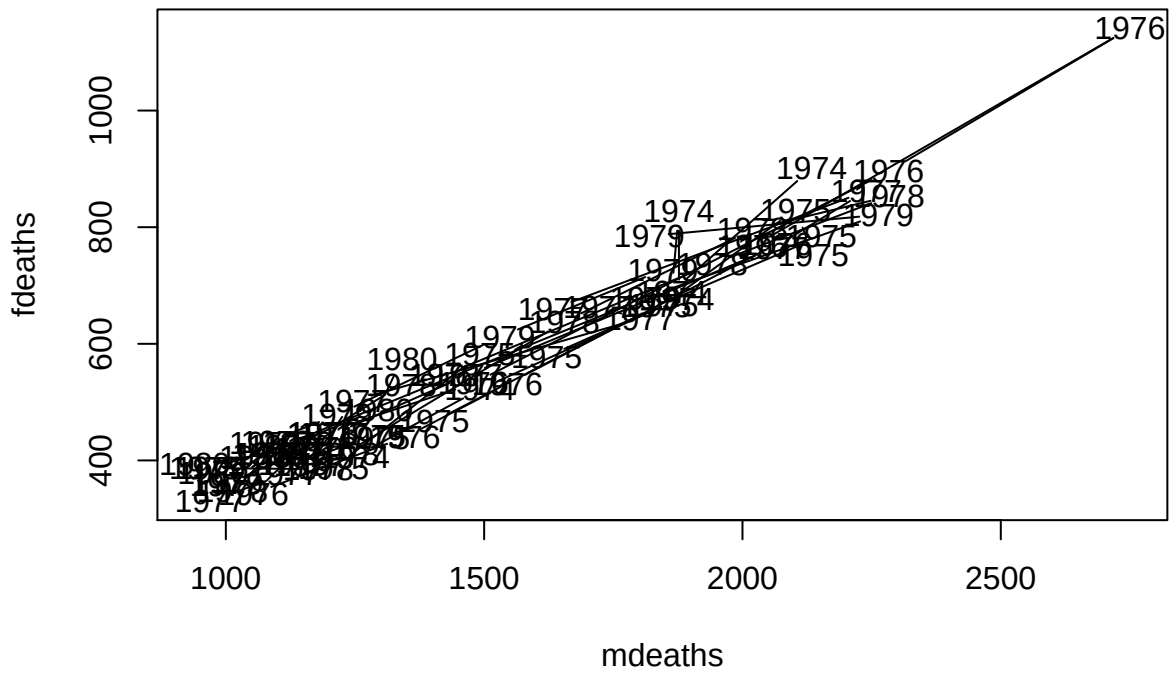
Lastly, we call up another R script that also explores descriptive statistics using the "source" function.Doing so also brings the underlying data set for it (deaths from various lung diseases in the UK)

```
source("MCAgosta.R")
```


ldeaths
1500
2500
3500
1974
1975
1976
1977
1978
1979
1980
Time


fdeaths
400
600
800
1000
1974
1975
1976
1977
1978
1979
1980
Time


mdeaths
1000
1500
2000
2500
1974
1975
1976
1977
1978
1979
1980
Time



What follows is code from the script above.

First step is looking at data sets on R in order to choose one

```
data()
```

Then loading UKLungDeaths data, as it seems doable and fairly complete

```
data(UKLungDeaths)
```

Looking at UKLungDeaths data on the tab that layouts the data in R

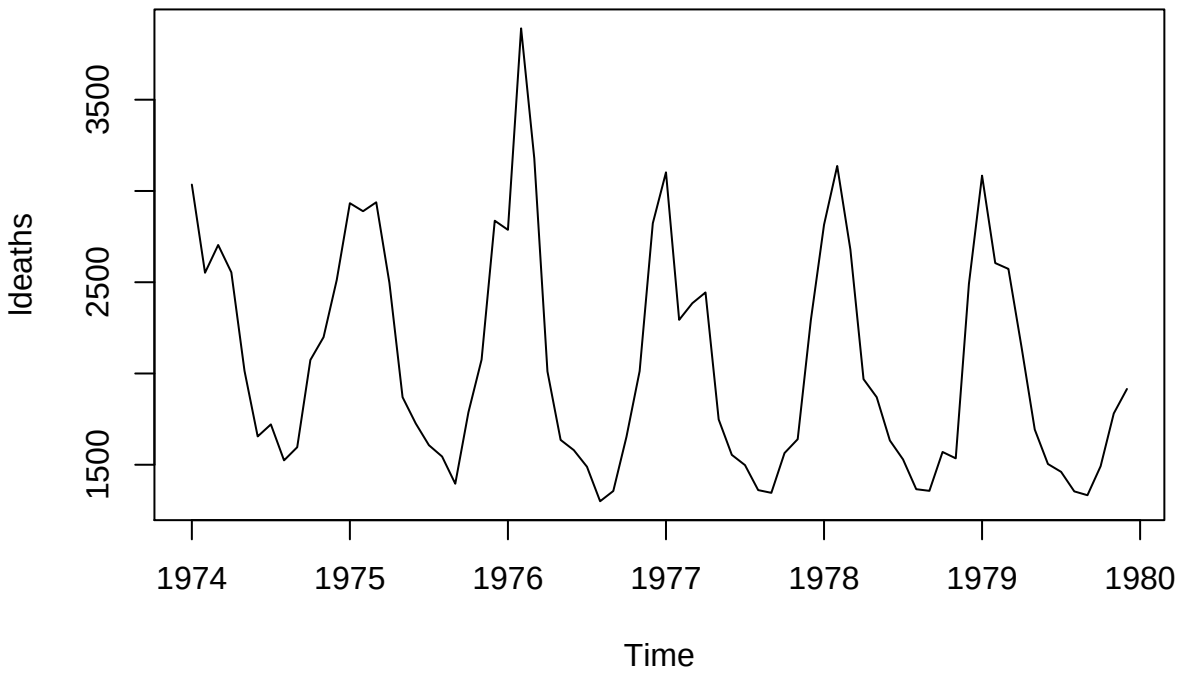
```
?UKLungDeaths
```

```
## starting httpd help server ...
```

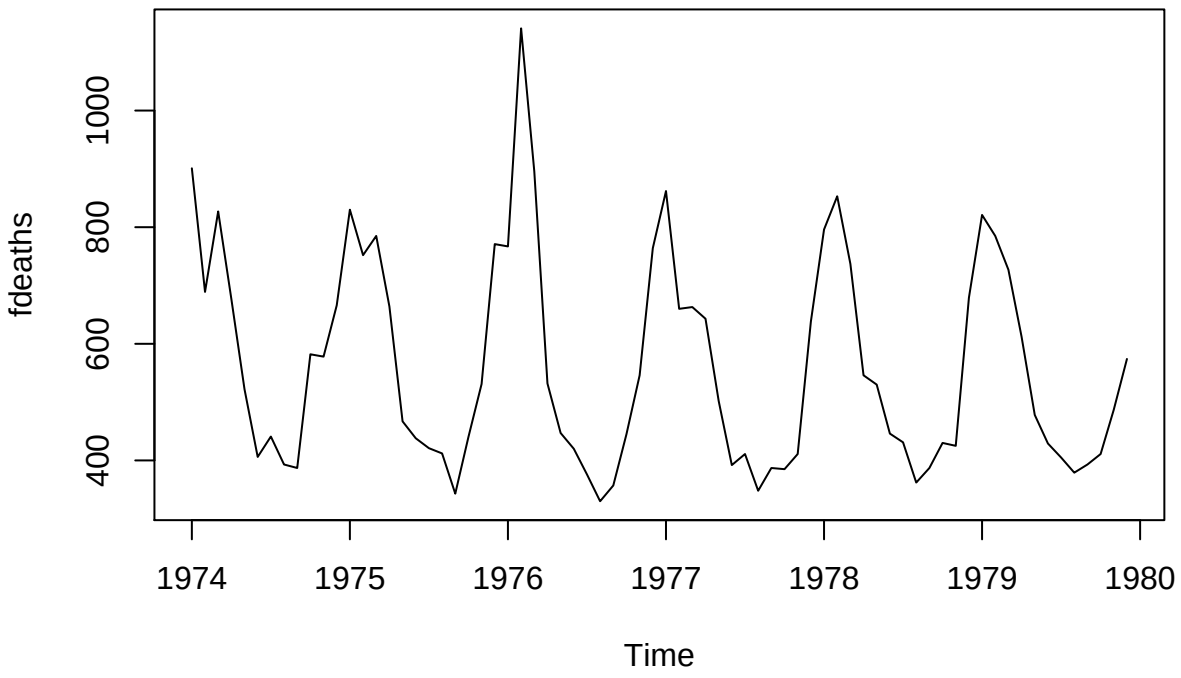
```
##  done
```

The data is broken down into three different time-series objects: both sexes (ldeaths), males (mdeaths) and females (fdeaths). It shows monthly deaths from bronchitis, emphysema and asthma in the UK, 1974-1979

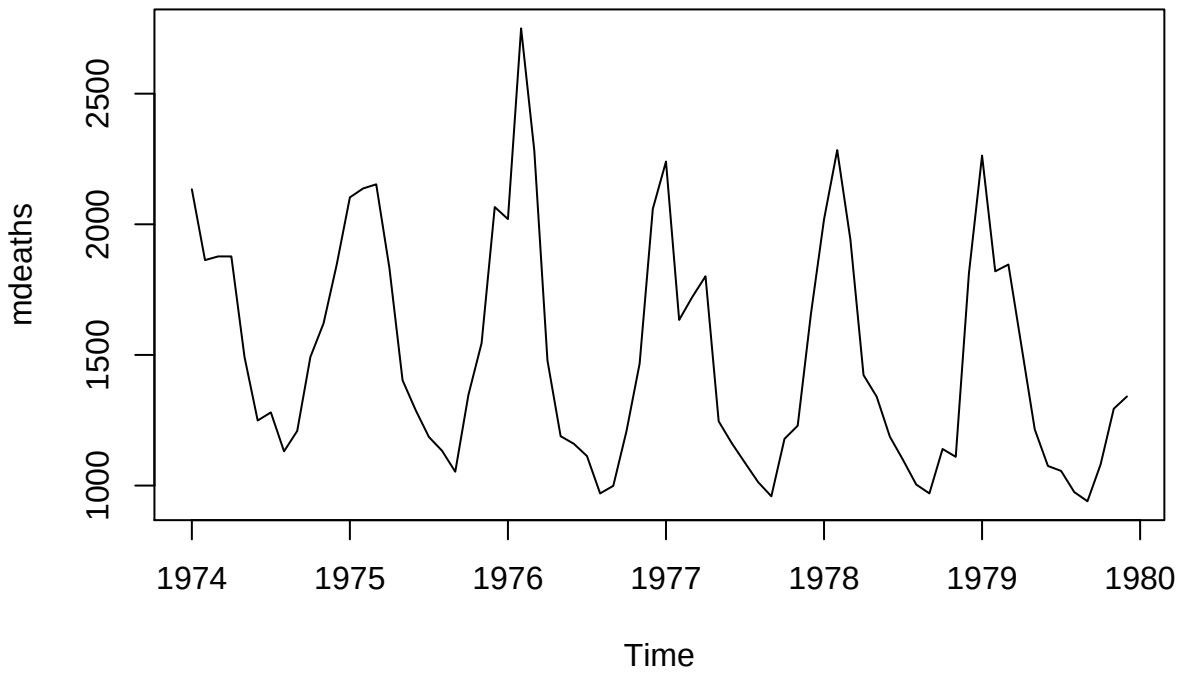
```
plot(ldeaths)
```




```
plot(fdeaths)
```

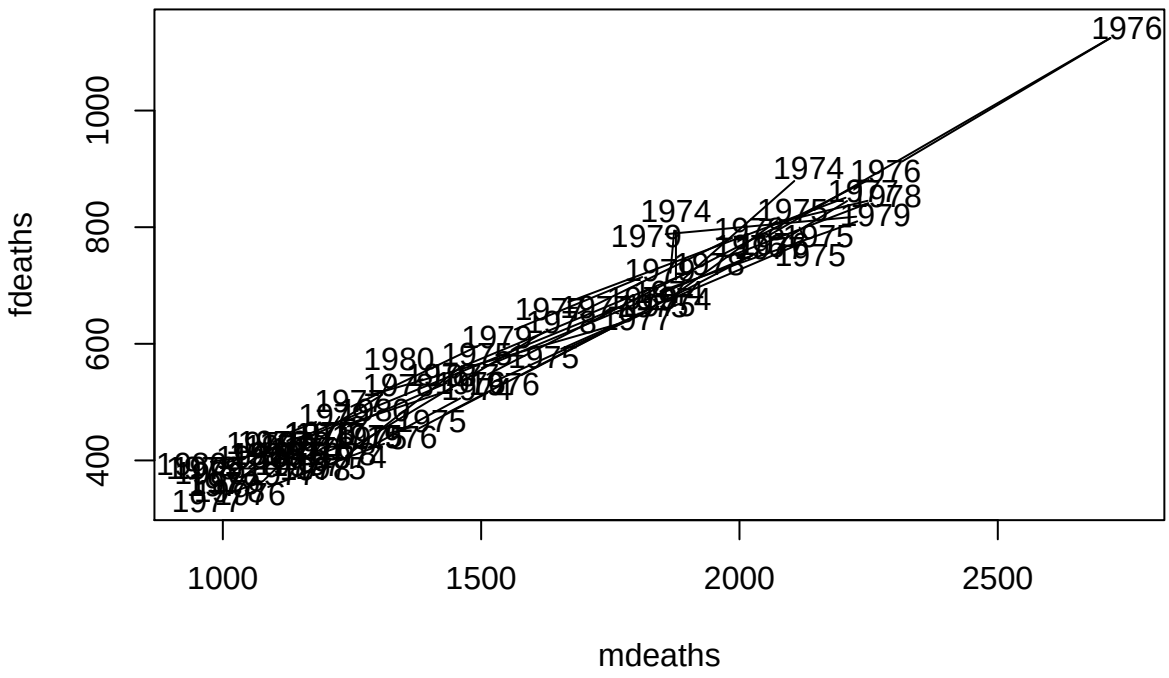



```
plot(mdeaths)
```



Plotting male and female deaths (below) to see if there is a trend or if there are large disparities

```
plot(mdeaths, fdeaths)
```



The overall trend is the same. Look at range first for female and male deaths

```
range(mdeaths)
```

```
## [1]  940 2750
```

Much larger range for males, compared to females.

```
range(fdeaths)
```

```
## [1]  330 1141
```

```
#[1]  330 1141
```

Looking at overall summary statistics to see if male numbers are larger across the board

```
summary(mdeaths)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     940    1138    1344    1496    1846    2750
```

```
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#940    1138    1344    1496    1846    2750

summary(fdeaths)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   330.0   411.0   512.0   560.7   681.5  1141.0
```

```
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#330.0   411.0   512.0   560.7   681.5  1141.0
```

There is an equal amount of data for both men and women, therefore the higher numbers in for men speak to potential enviromental factors (working in mines) and potential cultural factors (perhaps men smoked more cigerettes than women..?) These are merely guesstimates as to some of the trends.

```
sd (mdeaths)
```

```
## [1] 433.1509
```

```
# [1] 433.1509

sd (fdeaths)
```

```
## [1] 179.72
```

```
# [1] 179.72

cor.test(mdeaths, fdeaths)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mdeaths and fdeaths
## t = 37.694, df = 70, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9621846 0.9851124
## sample estimates:
##       cor
## 0.9762413
```

```
#Pearson's product-moment correlation

#data:  mdeaths and fdeaths
#t = 37.694, df = 70, p-value < 2.2e-16
#alternative hypothesis: true correlation is not equal to 0
#95 percent confidence interval:
# 0.9621846 0.9851124
#sample estimates:
#      cor
# 0.9762413
```

Combining the time-series objects into one object

```
deaths <- data.frame(fdeaths, ldeaths, mdeaths)
```