In [1]:
```r
install.packages("OECD")
library(tidyverse)
library(httr)
library(jsonlite)
library(dplyr)
library(OECD)
library(ggplot2)
library(gridExtra)
```

```
Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
Registered S3 methods overwritten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
Registered S3 method overwritten by 'rvest':
  method            from
  read_xml.response xml2
── Attaching packages ─────────────────────────────────────── tidyverse 1.
2.1 ──
✔ ggplot2 3.1.1       ✔ purrr   0.3.2
✔ tibble  2.1.1       ✔ dplyr   0.8.0.1
✔ tidyr   0.8.3       ✔ stringr 1.4.0
✔ readr   1.3.1       ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts
() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()

Attaching package: ‘jsonlite’

The following object is masked from ‘package:purrr’:

    flatten


Attaching package: ‘gridExtra’

The following object is masked from ‘package:dplyr’:

    combine
```

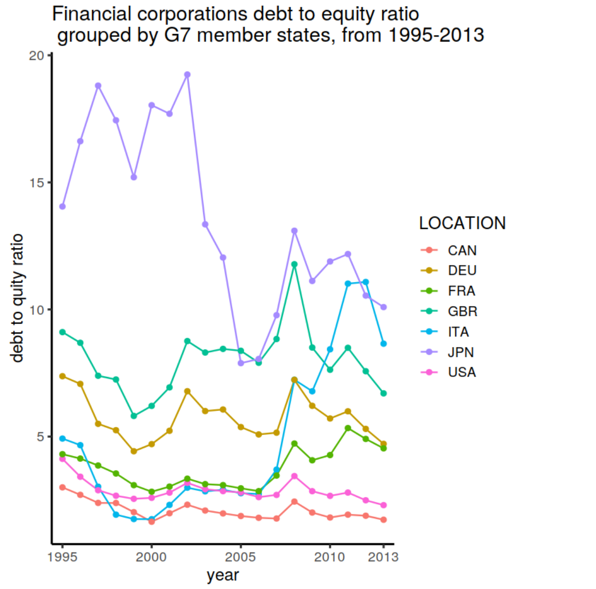
Data source: OECD (2021), "National Accounts at a Glance", OECD National Accounts Statistics (database), https://doi.org/10.1787/data-00369-en (accessed on 10 March 2021).

In [2]:
```r
df <- get_dataset("NAAG", "CAN+FRA+DEU+ITA+JPN+GBR+USA.DBTEQS12", start_ti
```

In [3]:
```r
print(df)
```

```
# A tibble: 133 x 6
   LOCATION INDICATOR TIME_FORMAT POWERCODE obsTime obsValue
 * <chr>    <chr>     <chr>       <chr>     <chr>      <dbl>
 1 ITA      DBTEQS12  P1Y         0         1995        4.92
 2 ITA      DBTEQS12  P1Y         0         1996        4.66
 3 ITA      DBTEQS12  P1Y         0         1997        3.03
 4 ITA      DBTEQS12  P1Y         0         1998        1.93
 5 ITA      DBTEQS12  P1Y         0         1999        1.76
 6 ITA      DBTEQS12  P1Y         0         2000        1.75
 7 ITA      DBTEQS12  P1Y         0         2001        2.31
 8 ITA      DBTEQS12  P1Y         0         2002        2.99
 9 ITA      DBTEQS12  P1Y         0         2003        2.84
10 ITA      DBTEQS12  P1Y         0         2004        2.91
# … with 123 more rows
```

In [4]:
```r
title = "Financial corporations debt to equity ratio \n grouped by G7 membe
df %>%
    ggplot( aes(x=obsTime, y=obsValue, group=LOCATION, color=LOCATION)) +
        geom_line() +
        geom_point() +
        ggtitle(title) +
        theme(plot.title = element_text(size = 20))+
        xlab("year") +
        ylab("debt to quity ratio") +
        theme_classic(base_size = 14) +
        scale_x_discrete(breaks=c(1995, 2000, 2005, 2010, 2013))
```

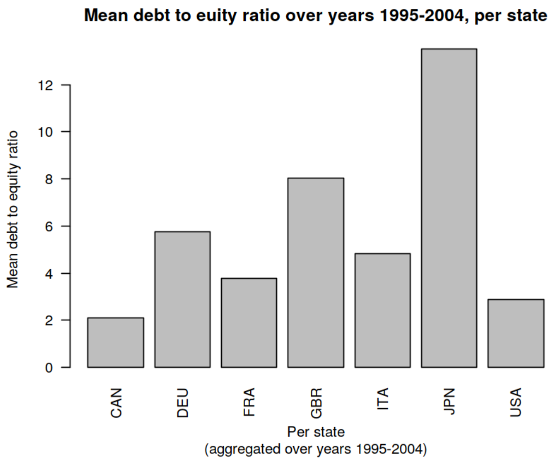


In [5]:
```r
table(df$LOCATION)
```

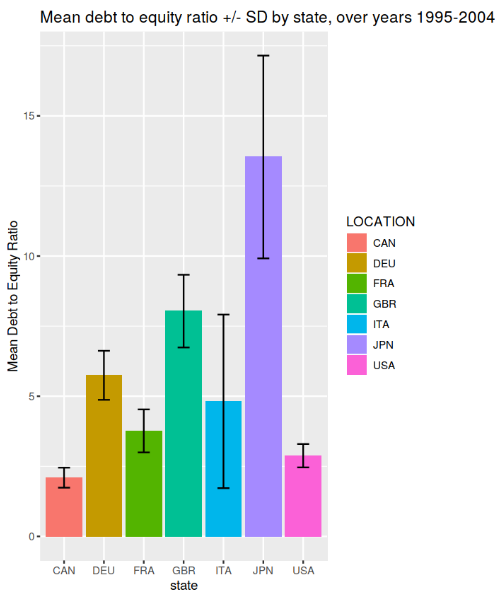
```
CAN DEU FRA GBR ITA JPN USA
 19  19  19  19  19  19  19
```

1) Computation of Mean over years per state

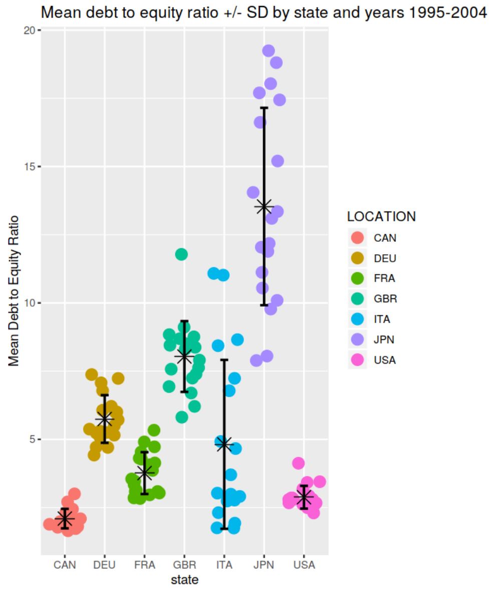
In [6]:
```r
loc.means <- tapply(df$obsValue, df$LOCATION, mean)
par(mar=c(12, 4, 4, 2))

barplot(loc.means, las=2, main="Mean debt to euity ratio over years 1995-2
        ylab="Mean debt to equity ratio",
        xlab="Per state",
        sub="(aggregated over years 1995-2004)")
#loc.sd <- tapply(df$obsValue, df$LOCATION, sd)
```



In [7]:
```r
ggplot(df) +
    aes(x=LOCATION, y=obsValue, fill=LOCATION) +
    stat_summary(aes(), geom = "bar", fun.y = "mean") +
    stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "e
    scale_x_discrete(name = "state") + scale_y_continuous(name = "Mean Deb
    ggtitle("Mean debt to equity ratio +/- SD by state, over years 1995-20
```



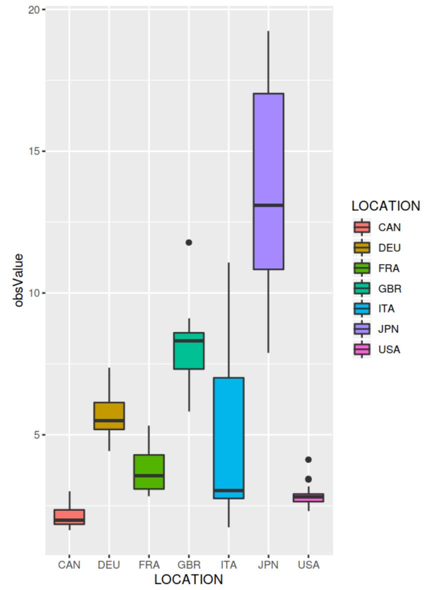
In [8]:
```r
ggplot(df) +
 aes(x=LOCATION, y=obsValue, color=LOCATION) +
 geom_jitter(position = position_jitter(0.4), size = 4) +
 stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "erro
 size = .75, color = "black") +
 stat_summary(fun.y=mean, colour="black", geom="point", shape=8, size=5) +
 scale_x_discrete(name = "state") + scale_y_continuous(name = "Mean Debt t
 ggtitle("Mean debt to equity ratio +/- SD by state and years 1995-2004")
```



Both representations (barplot and jitter) have advantages and disadvantages:

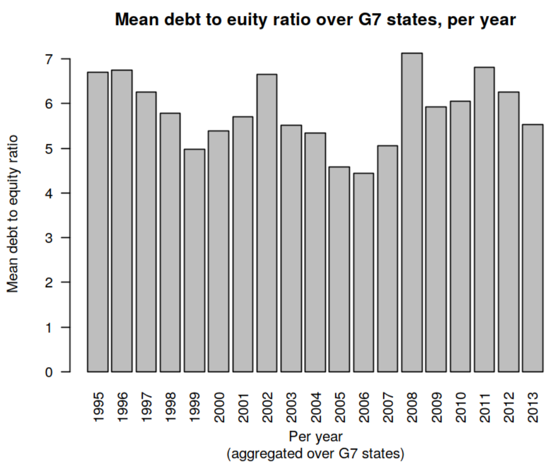
The barplot does not provide visual information about the distribution of the data, but is very unambiguous.
The jitterplot visualizes the distribution of the data, but the horizontal shift of the dots can not be interpreted and can thus be confusing to the observer.

In [9]:
```
ggplot(df) +
    geom_boxplot(aes(x=LOCATION, y=obsValue, fill=LOCATION))
```

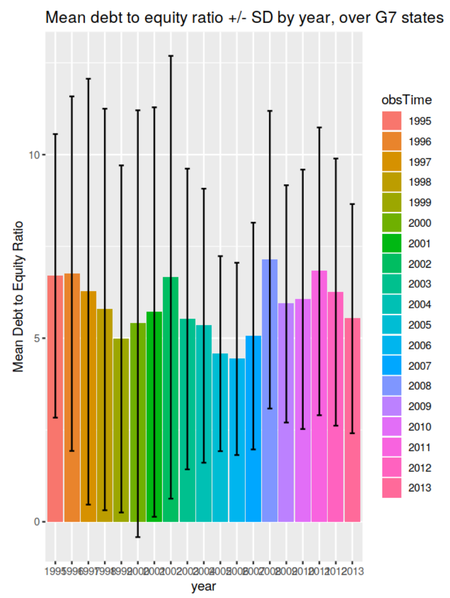


2) Computation of Mean over states and per year

In [10]:
```
yrs.means <- tapply(df$obsValue, df$obsTime, mean)
par(mar=c(12, 4, 4, 2))

barplot(yrs.means, las=2, main="Mean debt to euity ratio over G7 states, p
        ylab="Mean debt to equity ratio",
        xlab="Per year",
        sub="(aggregated over G7 states)")
```

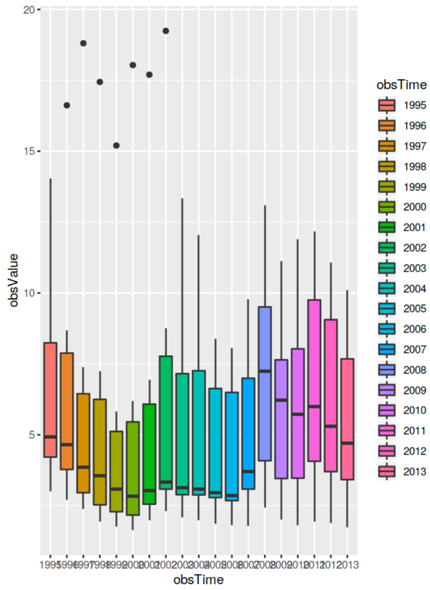


In [11]:
```
ggplot(df) +
    aes(x=obsTime, y=obsValue, fill=obsTime) +
    stat_summary(aes(), geom = "bar", fun.y = "mean") +
    stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "e
    scale_x_discrete(name = "year") + scale_y_continuous(name = "Mean Debt
    ggtitle("Mean debt to equity ratio +/- SD by year, over G7 states")
```



In [12]:
```
ggplot(df) +
 aes(x=obsTime, y=obsValue, color=obsTime) +
 geom_jitter(position = position_jitter(0.4), size = 4) +
 stat_summary(fun.data = mean_sdl, fun.args = list(mult = 1), geom = "erro
 size = .75, color = "black") +
 stat_summary(fun.y=mean, colour="black", geom="point", shape=8, size=5) +
 scale_x_discrete(name = "year") + scale_y_continuous(name = "Mean Debt to
 ggtitle("Mean debt to equity ratio +/- SD by year and over G7 states")
```

In [13]:
```
ggplot(df) +
    geom_boxplot(aes(x=obsTime, y=obsValue, fill=obsTime))
```



3) ANOVA

Grouped by state: differ the mean values of the states significantly?

In [14]:
```
print("normality test on all groups:")
shapiro.test(df$obsValue)
print("\nnormality test per group:")
locations <- list("ITA", "DEU", "FRA", "JPN", "GBR", "USA", "CAN")
for(l in locations) {
    print(l)
    df_cur <- df[df$LOCATION == l,]
    shapiro.test(df_cur$obsValue)
}
shapiro.test((df[df$LOCATION == "ITA",])$obsValue)
shapiro.test((df[df$LOCATION == "DEU",])$obsValue)
shapiro.test((df[df$LOCATION == "FRA",])$obsValue)
shapiro.test((df[df$LOCATION == "GBR",])$obsValue)
shapiro.test((df[df$LOCATION == "USA",])$obsValue)
shapiro.test((df[df$LOCATION == "CAN",])$obsValue)
shapiro.test((df[df$LOCATION == "JPN",])$obsValue)
```

```
[1] "normality test on all groups:"
        Shapiro-Wilk normality test

data:  df$obsValue
```

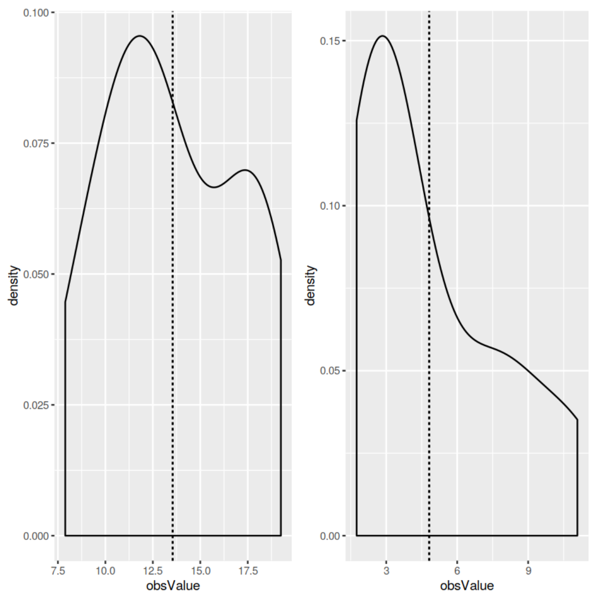
```
W = 0.84172, p-value = 1.215e-10
[1] "\nnormality test per group:"
[1] "ITA"
[1] "DEU"
[1] "FRA"
[1] "JPN"
[1] "GBR"
[1] "USA"
[1] "CAN"
        Shapiro-Wilk normality test

data:  (df[df$LOCATION == "ITA", ])$obsValue
W = 0.83963, p-value = 0.004569
        Shapiro-Wilk normality test

data:  (df[df$LOCATION == "DEU", ])$obsValue
W = 0.93899, p-value = 0.2529
        Shapiro-Wilk normality test

data:  (df[df$LOCATION == "FRA", ])$obsValue
W = 0.92445, p-value = 0.1368
        Shapiro-Wilk normality test

data:  (df[df$LOCATION == "GBR", ])$obsValue
W = 0.91344, p-value = 0.08562
        Shapiro-Wilk normality test

data:  (df[df$LOCATION == "USA", ])$obsValue
W = 0.85024, p-value = 0.006769
        Shapiro-Wilk normality test

data:  (df[df$LOCATION == "CAN", ])$obsValue
W = 0.89593, p-value = 0.04108
        Shapiro-Wilk normality test

data:  (df[df$LOCATION == "JPN", ])$obsValue
W = 0.94463, p-value = 0.319
```

Caution: p-value not < 0.05 for all groups (USA, CAN, ITA)! This means that the distribution of the data significantly differs from the normal distribution. We continue anyways.

In [15]:
```
#df_deu <- df[df$LOCATION == "DEU",]
#head(df_deu)
```

In [16]:
```
df_jpn <- df[df$LOCATION == "JPN",]
jpn_density <- ggplot(df_jpn, aes(x = obsValue)) +
    geom_density() +
    geom_vline(aes(xintercept=mean(obsValue)),
               linetype="dashed", size=0.6)

df_ita <- df[df$LOCATION == "ITA",]
ita_density <- ggplot(df_ita, aes(x = obsValue)) +
    geom_density() +
    geom_vline(aes(xintercept=mean(obsValue)),
               linetype="dashed", size=0.6)
```

In [17]:
```
grid.arrange(jpn_density, ita_density, nrow=1)
```



In [18]:
```
vals <- df$obsValue
groups <- df$LOCATION
loc.aov <- aov(vals ~ groups)
```

In [19]:
```
summary(loc.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
groups        6   1751  291.91   78.59 <2e-16 ***
Residuals   126    468    3.71
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Inter group variability (row groups):

\n

Intra group variability (row residuals):

In [20]:
```
f_test <- df(78.59, 6, 126)
print(f_test)
```

```
[1] 2.116975e-40
```

Levene Test: test Homogenity of Variance withing each group

In [21]:
```
library(car)
leveneTest(vals ~ groups, data=df)
```

```
Loading required package: carData

Attaching package: ‘car’

The following object is masked from ‘package:dplyr’:

    recode

The following object is masked from ‘package:purrr’:

    some

Warning message in leveneTest.default(y = y, group = grou
p, ...):
“group coerced to factor.”
```

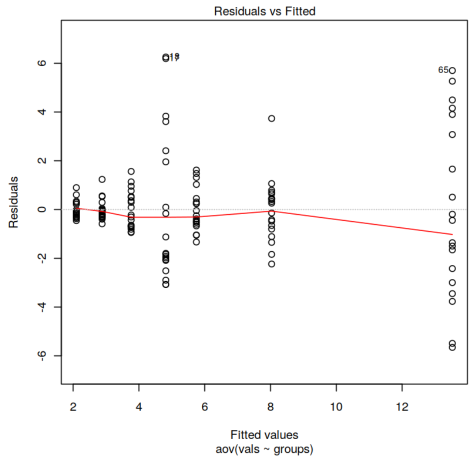
	Df	F value	Pr(>F)
group	6	12.18508	8.945817e-11
	126	NA	NA

In [22]:
```
print("Manually compute Levene Test (TODO)") # TODO
difference?
group.means <- tapply(vals, groups, mean)
vals_new <- abs(vals-group.means[groups]) # abs deviation
from true group mean
summary(aov(vals_new ~ groups))
```

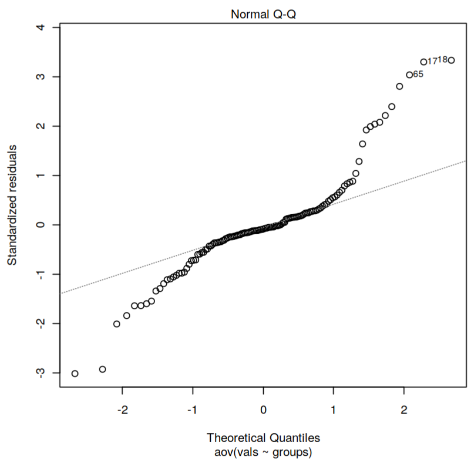
```
[1] "Manually compute Levene Test (TODO)"
             Df Sum Sq Mean Sq F value Pr(>F)    
groups        6  142.8   23.79   23.11 <2e-16 ***
Residuals   126  129.7    1.03                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1
```

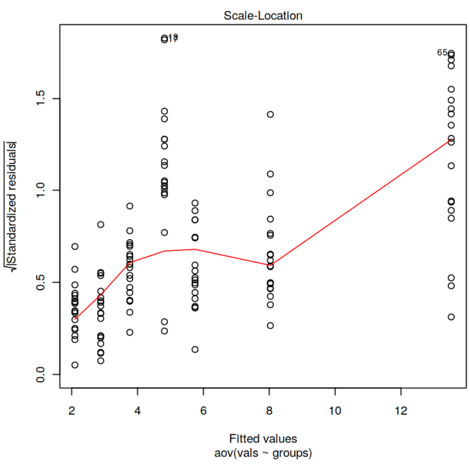
We clearly have a problem with the intra group variance (see p-value in Levene Test).

In [23]:
```
plot(loc.aov)
```




```
hat values (leverages) are all = 0.05263158
 and there are no factor predictors; no plot no. 5
```





In [24]:
```
summary.lm(loc.aov)
```

```
Call:
aov(formula = vals ~ groups)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6447 -0.6725 -0.1542  0.5095  6.2652 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.0947     0.4421   4.738 5.73e-06 ***
groupsDEU     3.6514     0.6253   5.840 4.18e-08 ***
groupsFRA     1.6679     0.6253   2.667  0.00865 ** 
groupsGBR     5.9414     0.6253   9.502  < 2e-16 ***
groupsITA     2.7217     0.6253   4.353 2.75e-05 ***
groupsJPN    11.4386     0.6253  18.294  < 2e-16 ***
groupsUSA     0.7836     0.6253   1.253  0.21247    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 1.927 on 126 degrees of freedom
Multiple R-squared:  0.7891,	Adjusted R-squared:  0.77
91 
F-statistic: 78.59 on 6 and 126 DF,  p-value: < 2.2e-16
```

Reference group: LOCATION="CAN"; mean "debt to equity ratio" from DEU, FRA, GBR, ITA, JPN each compared to CAN can be seen as significantly different (p-value < 0.05).

CAN is randomly chosen as reference group. There are, most certainly, better approaches than that. TODO

In [25]:
```
pairwise.t.test(vals, groups, p.adj = "holm")
```

```
	Pairwise comparisons using t tests with pooled SD 

data:  vals and groups 

    CAN     DEU     FRA     GBR     ITA     JPN    
DEU 5.0e-07 -       -       -       -       -      
FRA 0.04323 0.01330 -       -       -       -      
GBR 2.9e-15 0.00293 4.1e-09 -       -       -      
ITA 0.00025 0.41858 0.37764 1.1e-05 -       -      
JPN < 2e-16 < 2e-16 < 2e-16 1.4e-13 < 2e-16 -      
USA 0.41858 0.00011 0.41858 2.5e-12 0.01434 < 2e-16
```