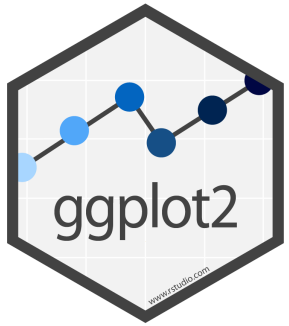


Introduction to ggplot2

Saghir Bashir

ilustat & GitHub: @saghirb

28th Feb 2019

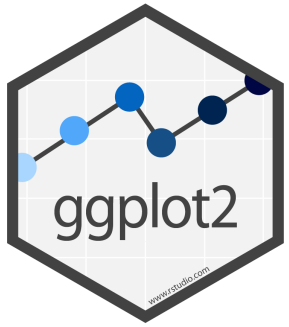
Relax

Experiment

Make Mistakes

Learn

Enjoy



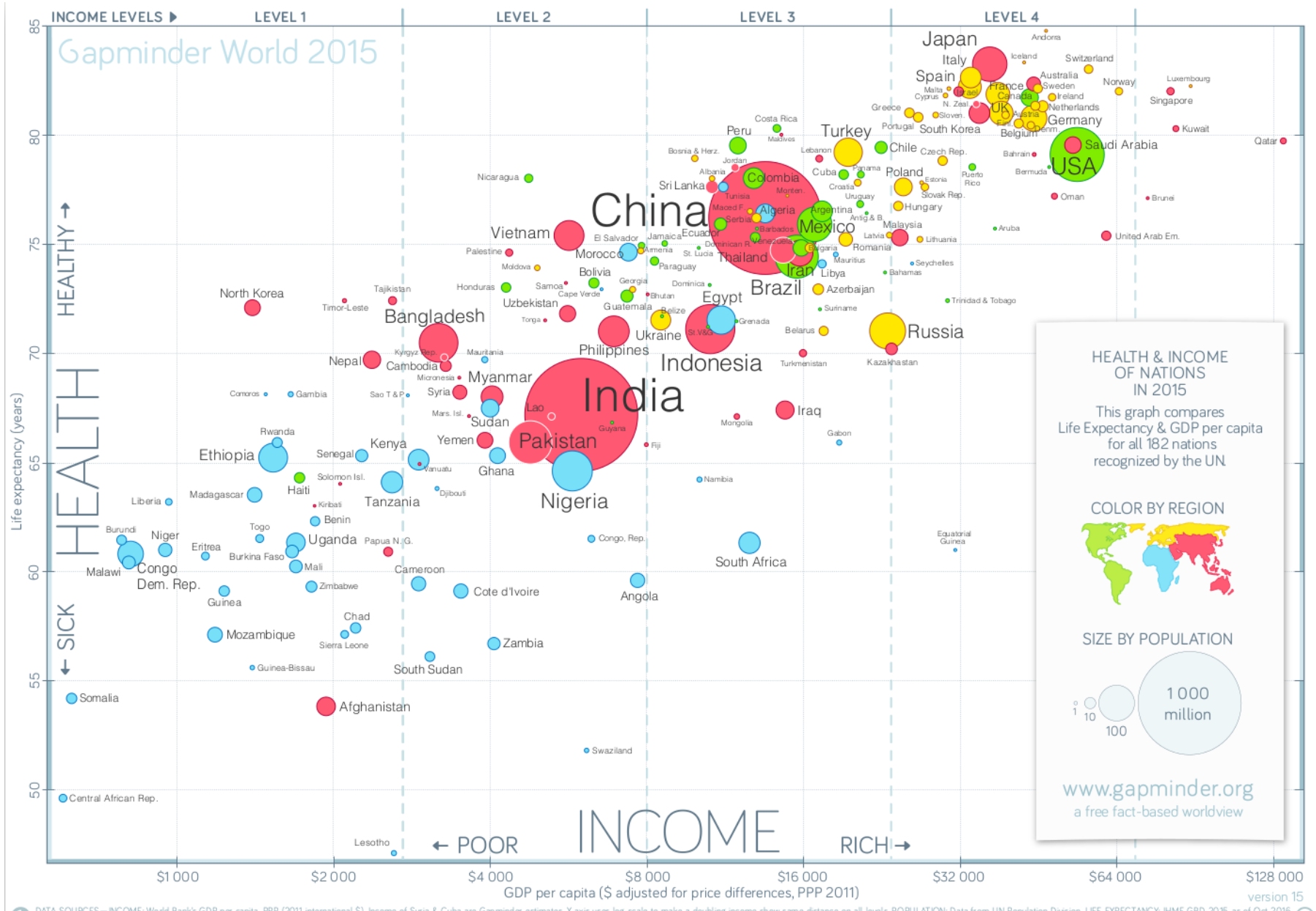
Outline

Gapminder Data

Plotting Data

Facetting Plots

Summary



DATA SOURCES — INCOME: World Bank's GDP per capita, PPP (2011 international $). Income of Syria & Cuba are Gapminder estimates. X-axis uses log-scale to make a doubling income show same distance on all levels. POPULATION: Data from UN Population Division. LIFE EXPECTANCY: IHME GBD-2015, as of Oct 2016.
ANIMATING GRAPH: Go to www.gapminder.org/tools to see how this graph changed historically and compare 500 other indicators. LICENSE: Our charts are freely available under Creative Commons Attribution License. Please copy, share, modify, integrate and even sell them, as long as you mention "Based on a free chart from www.gapminder.org".

Gapminder Data

- Population (Total)

- GDP per capita (US$, inflation-adjusted)

- Life expectancy at birth, in years

- Infant Mortality per 1000 births

- Total Fertility (children per woman)

First 10 Observations

continent	country	year	pop	gdpPercap	lifeexp	infmort	fertility
Europe	Albania	1960	1636054	NA	62.87	115.4	6.19
Europe	Albania	1965	1896125	NA	66.59	94.1	5.59
Europe	Albania	1970	2150602	NA	67.83	76.8	5.05
Europe	Albania	1975	2411229	NA	69.77	63.1	4.39
Europe	Albania	1980	2681245	1056.75	71.39	64.0	3.68
Europe	Albania	1985	2966799	1056.50	72.71	45.9	3.23
Europe	Albania	1990	3281453	980.16	73.30	35.1	2.97
Europe	Albania	1995	3106727	909.74	73.70	29.1	2.72
Europe	Albania	2000	3121965	1180.87	74.70	23.2	2.38
Europe	Albania	2005	3082172	1555.24	76.20	18.3	1.92

Last 10 Observations

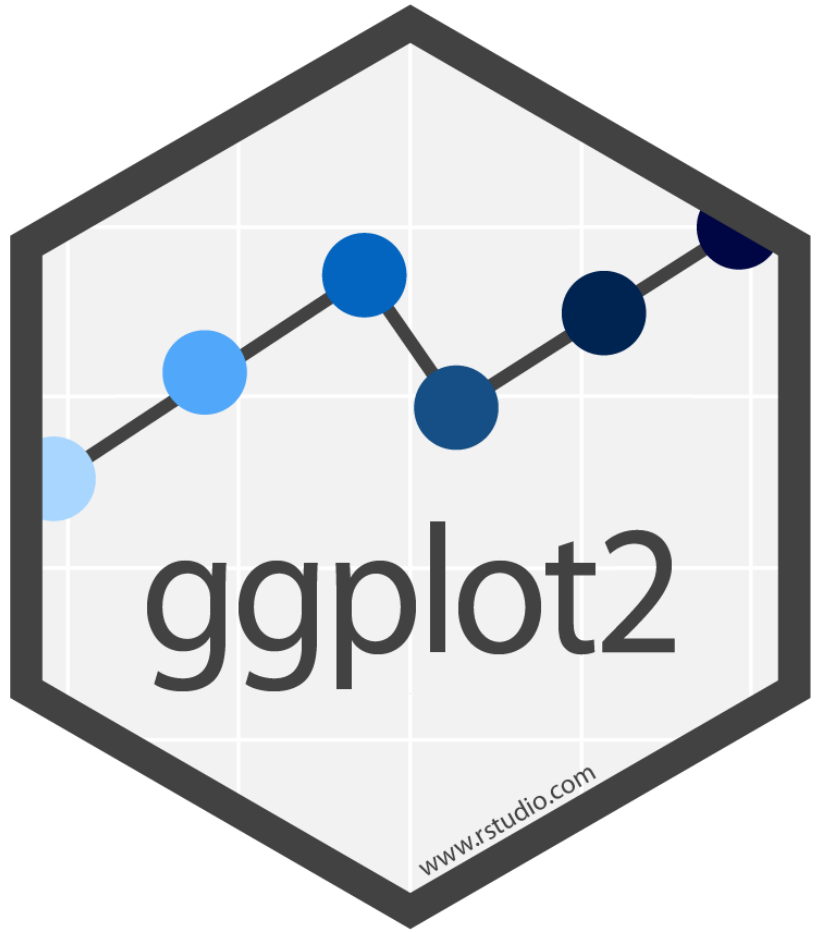
continent	country	year	pop	gdpPercap	lifeexp	infmort	fertility
Africa	Zimbabwe	1970	5206311	515.23	57.22	72.4	7.42
Africa	Zimbabwe	1975	6170284	550.34	59.41	70.3	7.40
Africa	Zimbabwe	1980	7289083	501.28	62.48	66.4	7.10
Africa	Zimbabwe	1985	8862601	507.41	64.86	53.6	6.22
Africa	Zimbabwe	1990	10484771	536.20	63.00	51.2	5.18
Africa	Zimbabwe	1995	11683136	510.82	56.00	60.1	4.43
Africa	Zimbabwe	2000	12499981	535.20	47.90	63.5	4.07
Africa	Zimbabwe	2005	12984418	351.02	45.10	61.0	3.97
Africa	Zimbabwe	2010	13973897	288.57	49.10	55.8	3.72
Africa	Zimbabwe	2015	15602751	NA	59.30	46.6	3.35

Questions of Interest

What are the time trends for Portugal?

How does Portugal compare to other European countries?

How does Portugal perform on health and wealth?


ggplot2
www.rstudio.com

```
install.packages("ggplot2")

install.packages("magrittr")

install.packages("data.table")

install.packages("here")
```

Imported Gapminder Data


```
library(data.table)
library(here)

gm <- fread(here("data", "gapminder.csv"))

# Create factor (categorical) variables to be used later.

gm[, continent:=as.factor(continent)]

gm[, country:=as.factor(country)]

gm[, fyear:=as.factor(year)]
```

Look at Imported Data

```
gm
```

```
##       continent  country year      pop gdpPercap lifeexp infmort fertility fyear
##    1:    Europe  Albania 1960  1636054        NA   62.87   115.4      6.19  1960
##    2:    Europe  Albania 1965  1896125        NA   66.59    94.1      5.59  1965
##    3:    Europe  Albania 1970  2150602        NA   67.83    76.8      5.05  1970
##    4:    Europe  Albania 1975  2411229        NA   69.77    63.1      4.39  1975
##    5:    Europe  Albania 1980  2681245 1056.7504   71.39    64.0      3.68  1980
##   ---
## 2216:    Africa Zimbabwe 1995 11683136  510.8200   56.00    60.1      4.43  1995
## 2217:    Africa Zimbabwe 2000 12499981  535.1974   47.90    63.5      4.07  2000
## 2218:    Africa Zimbabwe 2005 12984418  351.0233   45.10    61.0      3.97  2005
## 2219:    Africa Zimbabwe 2010 13973897  288.5683   49.10    55.8      3.72  2010
## 2220:    Africa Zimbabwe 2015 15602751        NA   59.30    46.6      3.35  2015
```

Structure of Gapminder Data

```
str(gm)
```

```
## Classes 'data.table' and 'data.frame':    2220 obs. of  9 variables:
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ country  : Factor w/ 185 levels "Albania","Algeria",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year     : int  1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 ...
##  $ pop      : int  1636054 1896125 2150602 2411229 2681245 2966799 3281453 3106727 3121965 3082172 ...
##  $ gdpPercap: num  NA NA NA NA 1057 ...
##  $ lifeexp  : num  62.9 66.6 67.8 69.8 71.4 ...
##  $ infmort  : num  115.4 94.1 76.8 63.1 64 ...
##  $ fertility: num  6.19 5.59 5.05 4.39 3.68 3.23 2.97 2.72 2.38 1.92 ...
##  $ fyear    : Factor w/ 12 levels "1960","1965",..: 1 2 3 4 5 6 7 8 9 10 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

What are the time trends for Portugal?

Portuguese Data

Create a dataset for Portugal

```
# Portuguese Data
options(width = 100)
gmPT <- gm[country == "Portugal"]
gmPT
```

```
##     continent  country year      pop gdpPercap lifeexp infmort fertility fyear
##  1:    Europe Portugal 1960  8875311  2363.976   64.23    84.6      3.16  1960
##  2:    Europe Portugal 1965  8888635  3212.569   66.17    66.4      3.18  1965
##  3:    Europe Portugal 1970  8670352  4615.613   67.14    55.4      2.99  1970
##  4:    Europe Portugal 1975  9185876  5398.299   68.90    36.0      2.71  1975
##  5:    Europe Portugal 1980  9755635  6518.493   71.71    22.8      2.29  1980
##  6:    Europe Portugal 1985  9929014  6693.814   73.22    16.5      1.78  1985
##  7:    Europe Portugal 1990  9890319  8854.322   74.20    11.5      1.54  1990
##  8:    Europe Portugal 1995 10078431  9455.050   75.50     7.4      1.48  1995
##  9:    Europe Portugal 2000 10278542 11412.078   76.80     5.5      1.47  2000
## 10:    Europe Portugal 2005 10480085 11663.632   78.40     3.7      1.41  2005
## 11:    Europe Portugal 2010 10584837 11805.935   79.90     3.1      1.33  2010
## 12:    Europe Portugal 2015 10349803        NA   80.80     3.0      1.31  2015
```

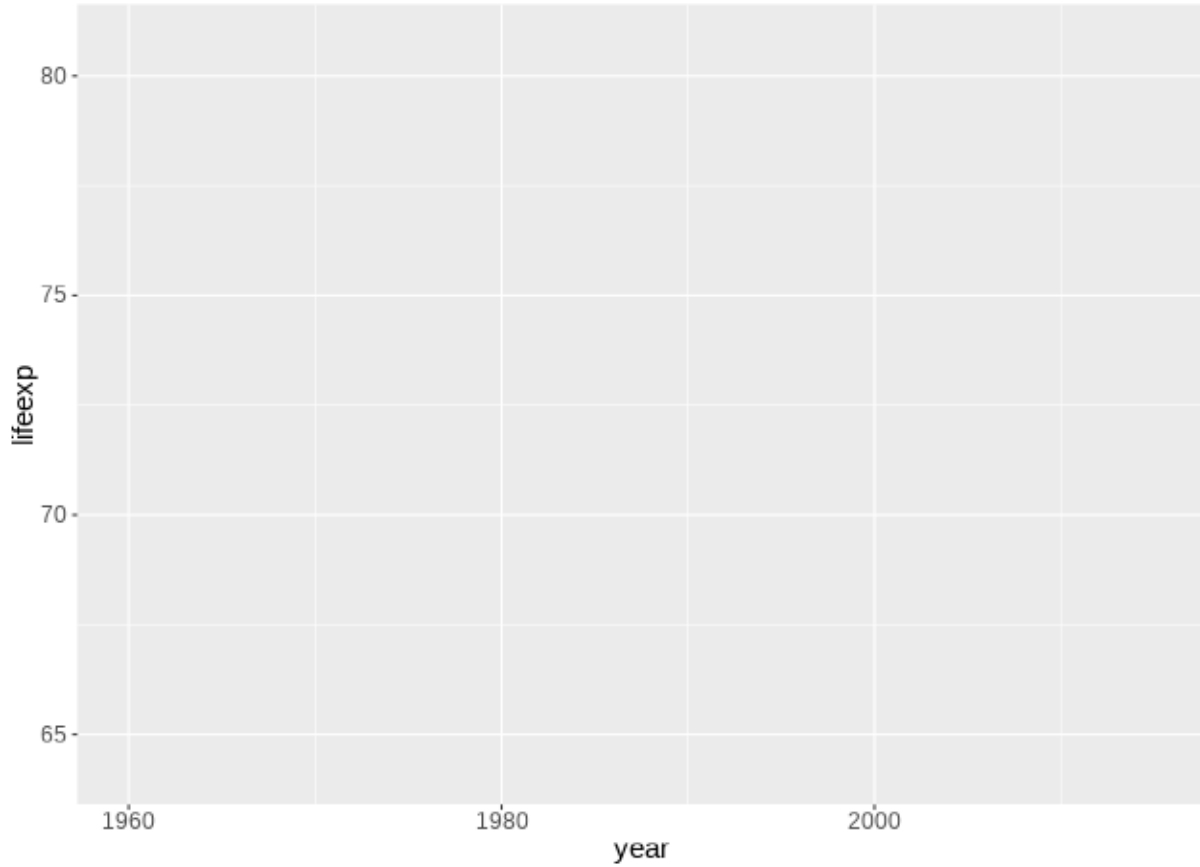
```
options(width = widthDefault)
```

Life Expectancy


```
ggplot(gmPT)
```

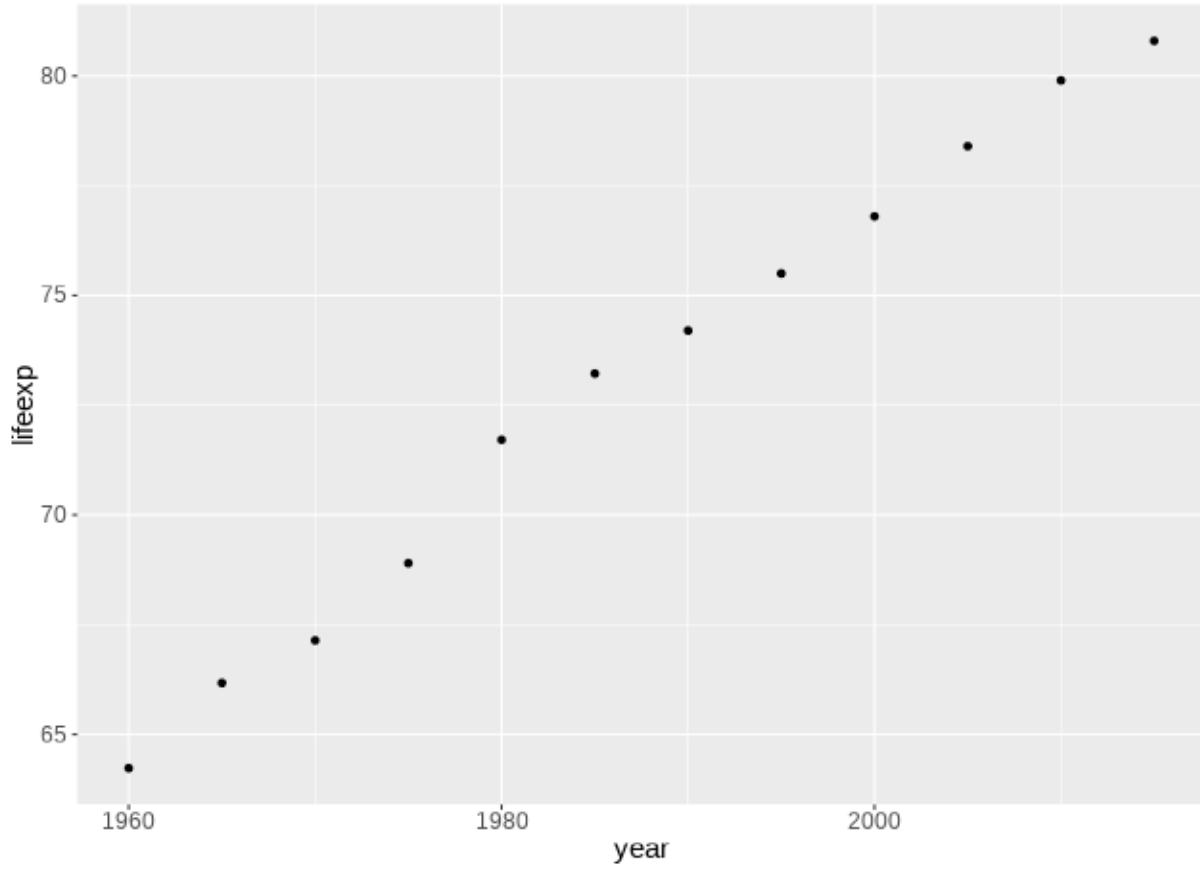
Life Expectancy


```
ggplot(gmPT, aes(year, lifeexp))
```



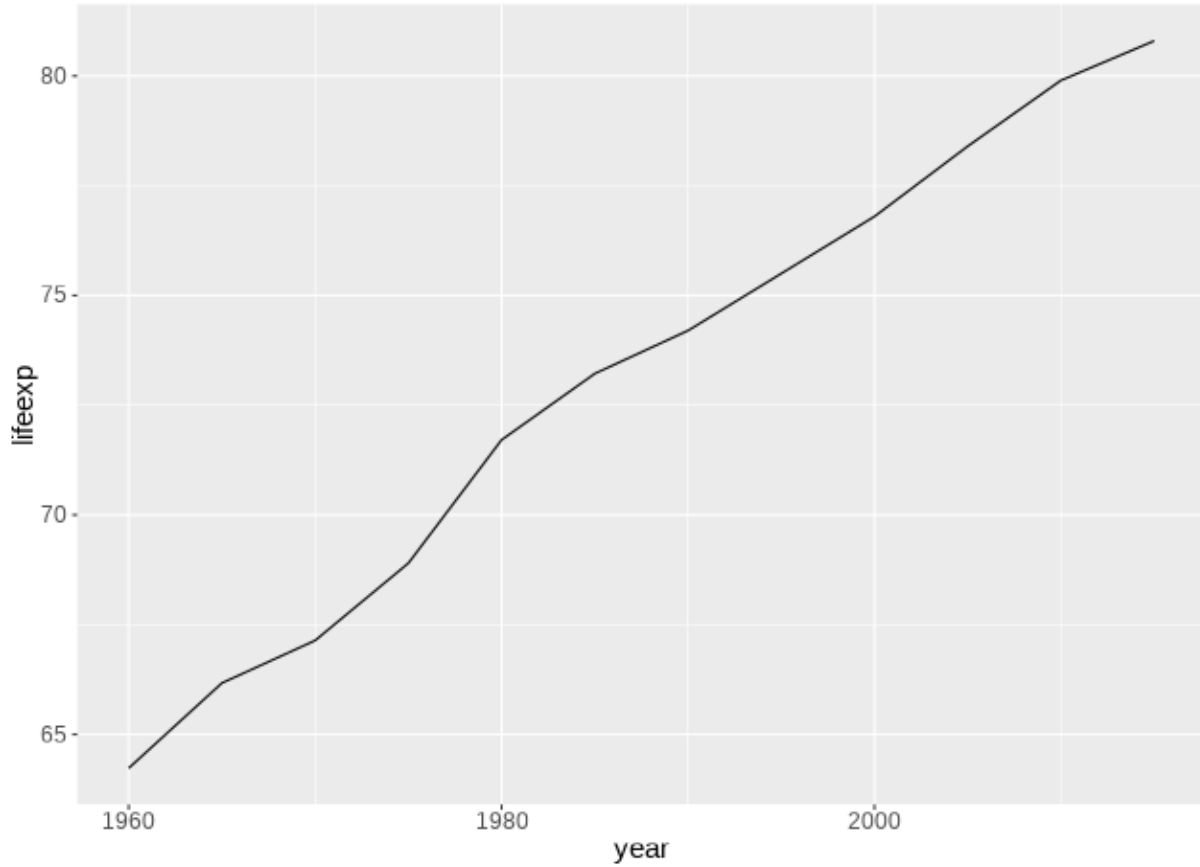
Life Expectancy - Points


```
ggplot(gmPT, aes(year, lifeexp)) +
  geom_point()
```



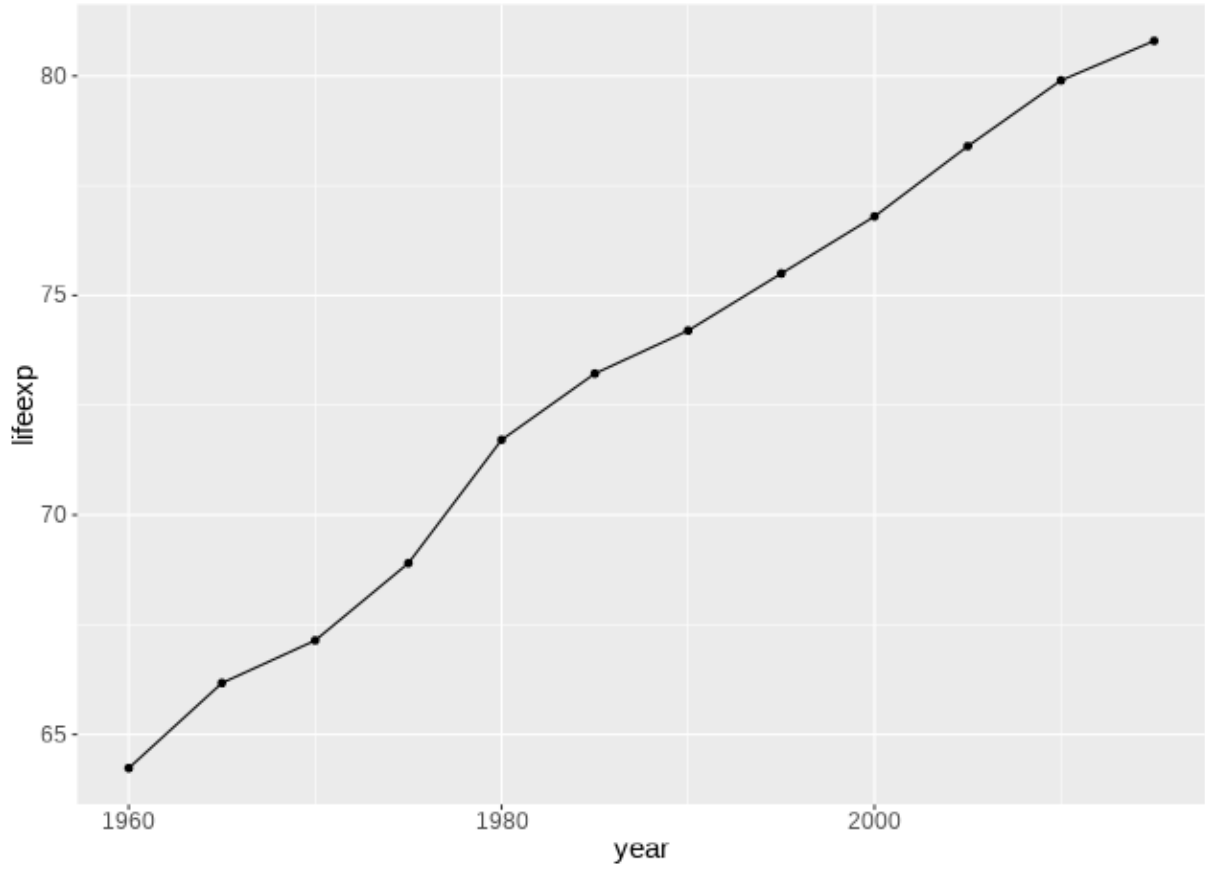
Life Expectancy - Line


```
ggplot(gmPT, aes(year, lifeexp)) +
  geom_line()
```



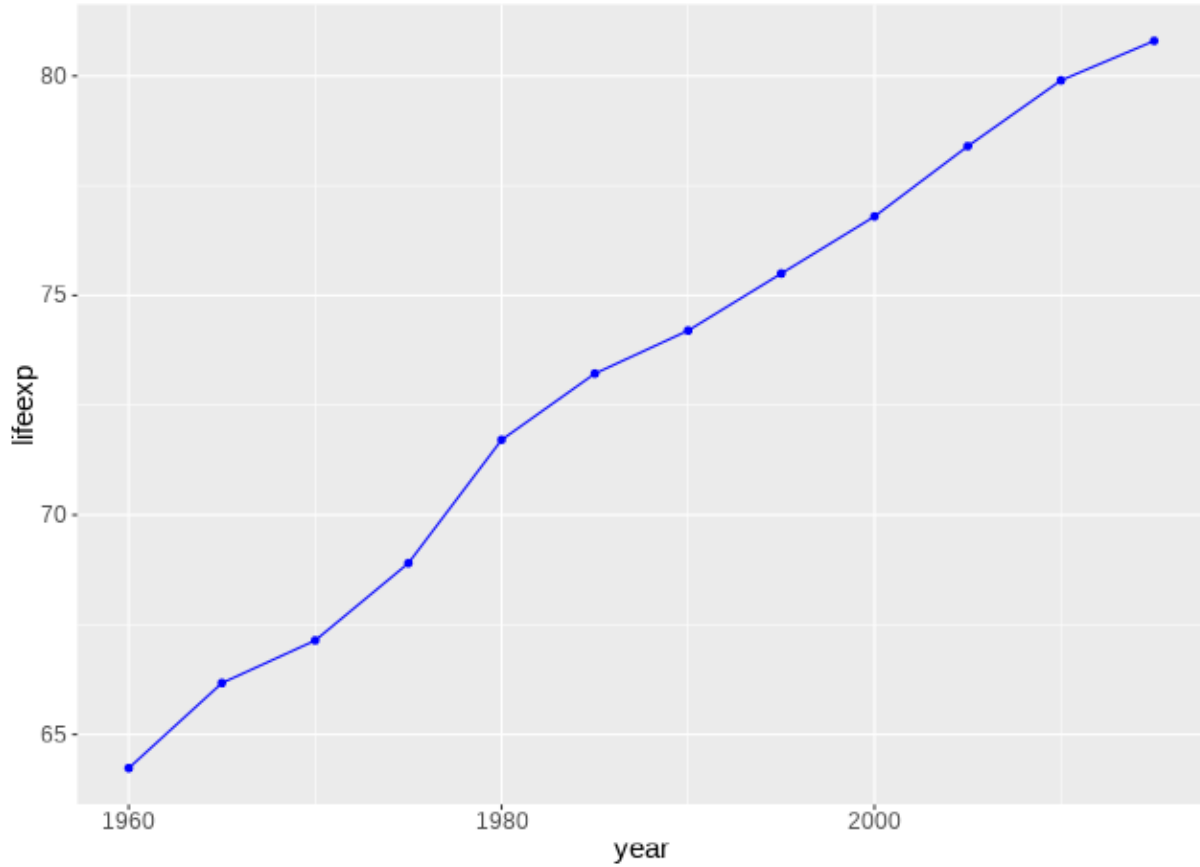
Life Expectancy - Points & Line

```
ggplot(gmPT, aes(year, lifeexp)) +
  geom_point() +
  geom_line()
```



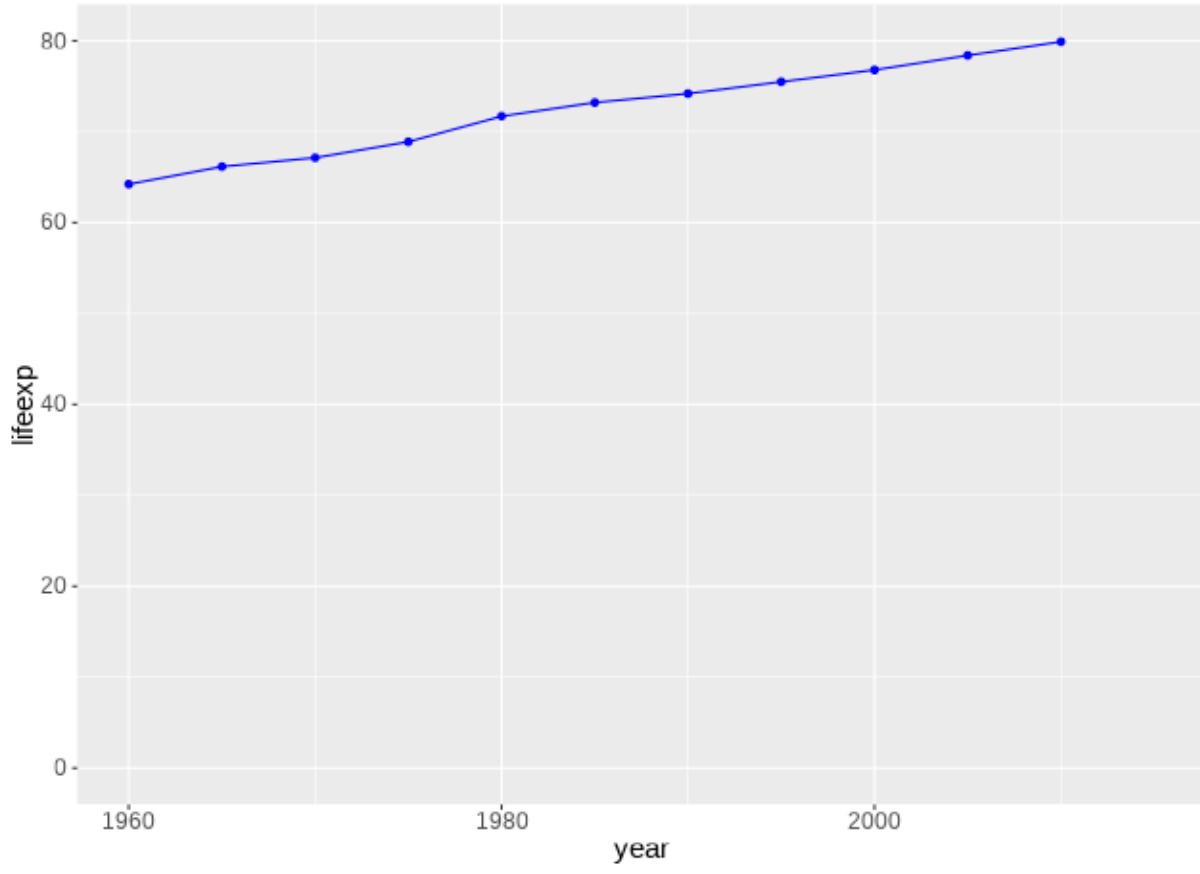
Life Expectancy - Colour


```
ggplot(gmPT, aes(year, lifeexp)) +
  geom_point(colour="blue") +
  geom_line(colour="blue")
```



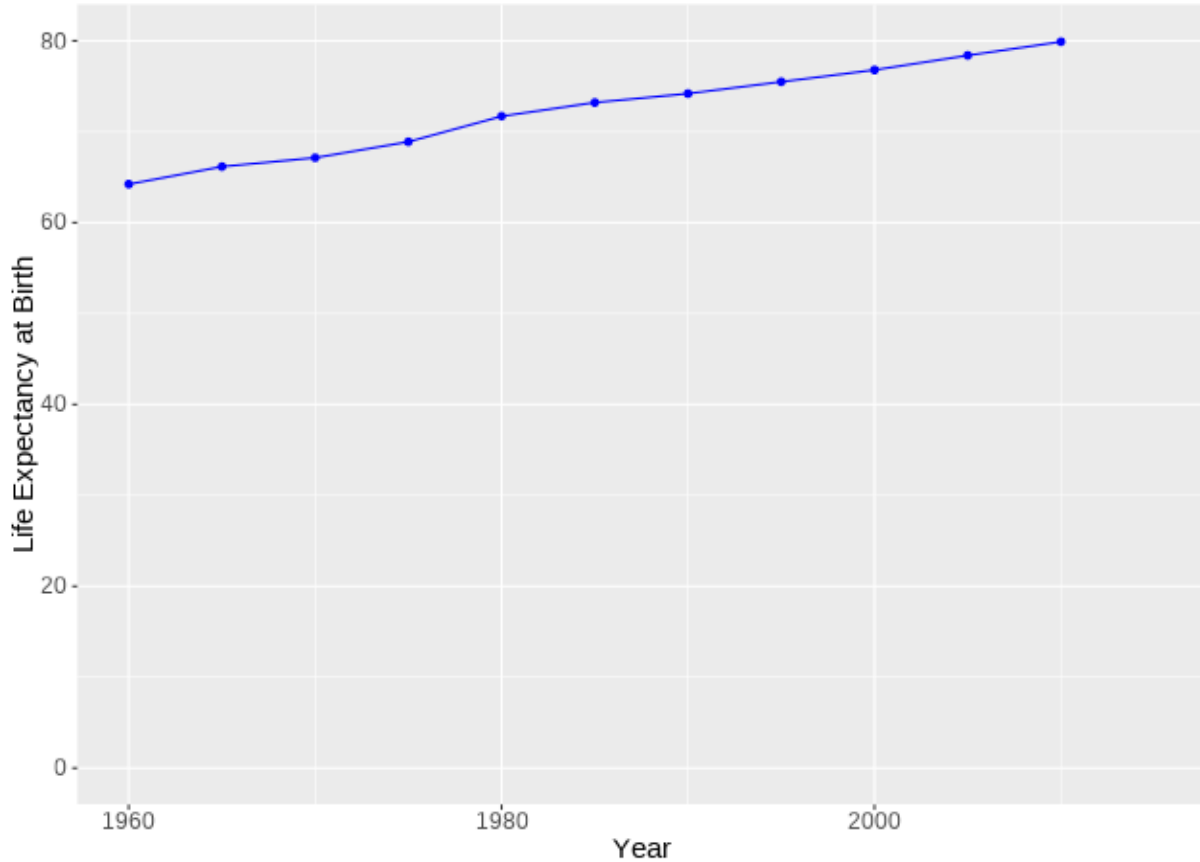
Life Expectancy - Y-axis


```
ggplot(gmPT, aes(year, lifeexp)) +
  geom_point(colour="blue") +
  geom_line(colour="blue") +
  scale_y_continuous(
                limits=c(0, 80))
```



Life Expectancy - Axis labels


```
ggplot(gmPT, aes(year, lifeexp)) +
  geom_point(colour="blue") +
  geom_line(colour="blue") +
  scale_y_continuous(
                limits=c(0, 80)) +
  xlab("Year") +
  ylab("Life Expectancy at Birth")
```



Interpret this plot.

Exercises

Download: https://ilustat.com/shared/ggplot2-Intro.zip

Double Click on "ggplot2-Exercises.Rproj"

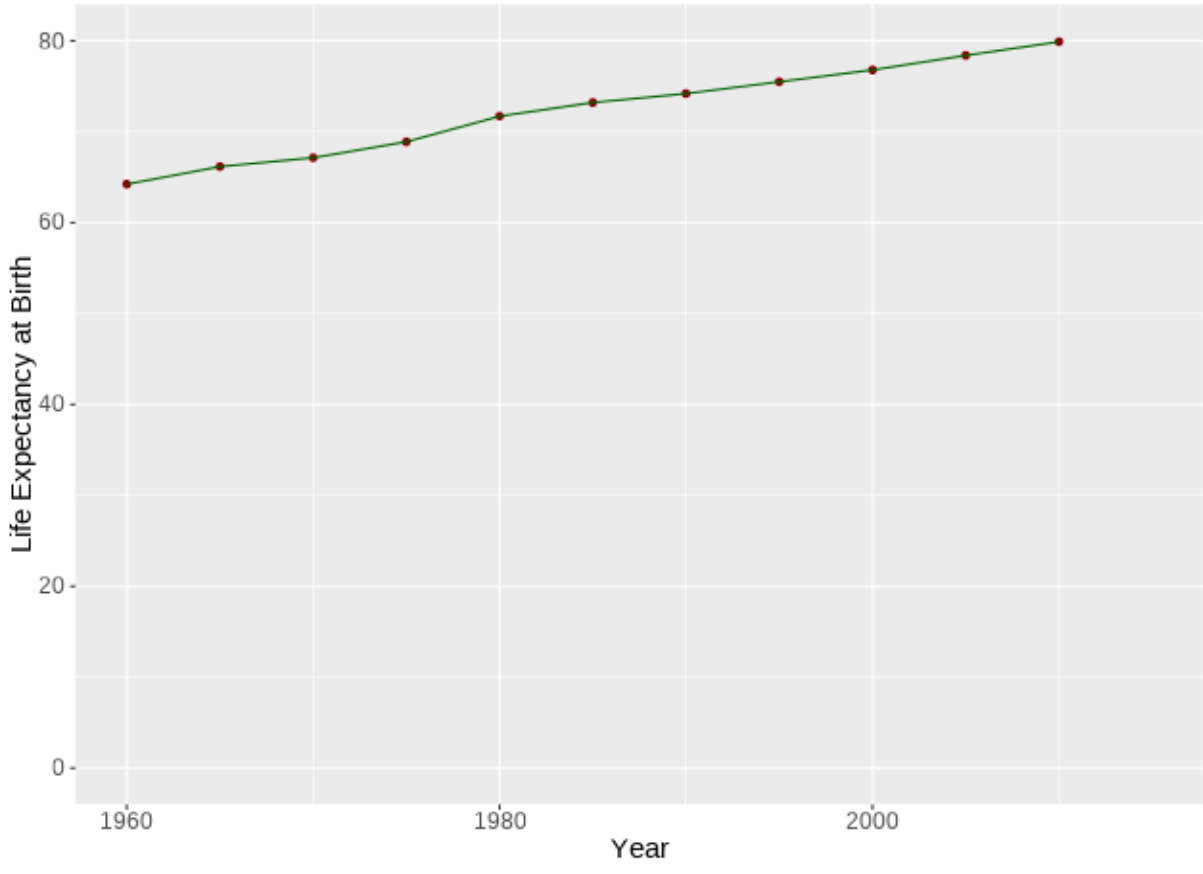
Open file "ggplot2-Exercises.Rmd"

Complete "Exercise 1 - Portugal"

Some comments (i)

Piping data to ggplot2

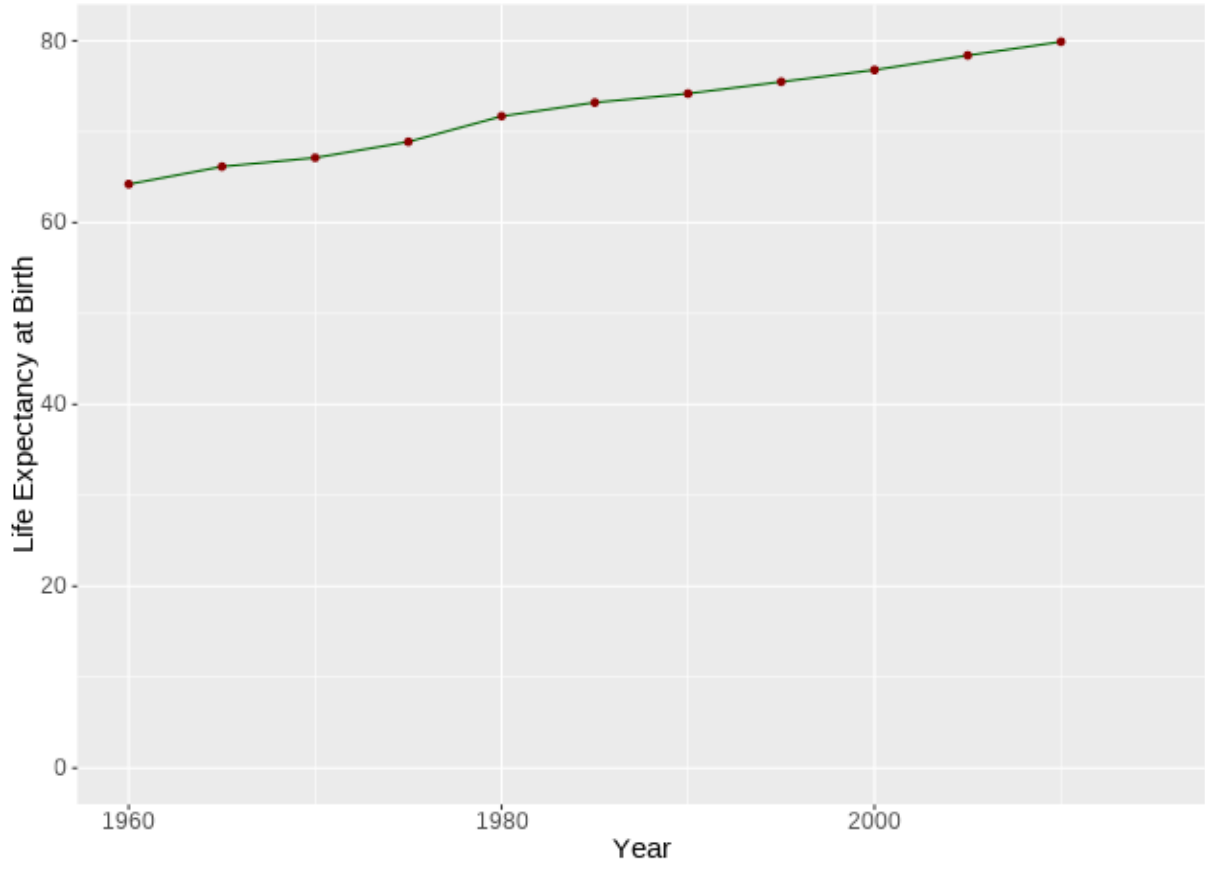
```
library(magrittr)

gm[country=="Portugal"] %>%
ggplot(aes(year, lifeexp)) +
  geom_point(colour="darkred") +
  geom_line(colour="darkgreen") +
  scale_y_continuous(
                limits=c(0, 80)) +
  xlab("Year") +
  ylab("Life Expectancy at Birth")
```



Some comments (ii)

Order matters

```
gm[country=="Portugal"] %>%
ggplot(aes(year, lifeexp)) +
  geom_line(colour="darkgreen") +
  geom_point(colour="darkred") +
  scale_y_continuous(
                limits=c(0, 80)) +
  xlab("Year") +
  ylab("Life Expectancy at Birth")
```

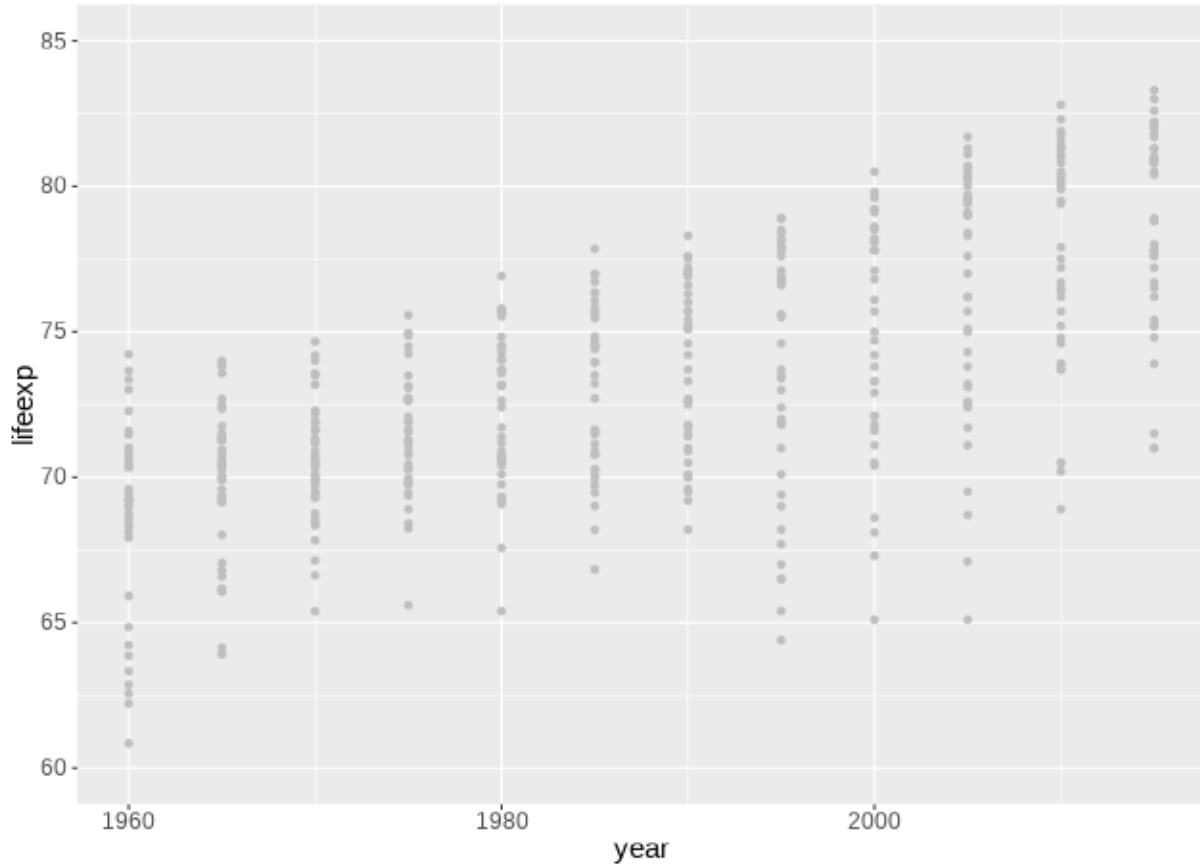


How does Portugal compare to other European countries?

Life Expectancy - Points

European Countries

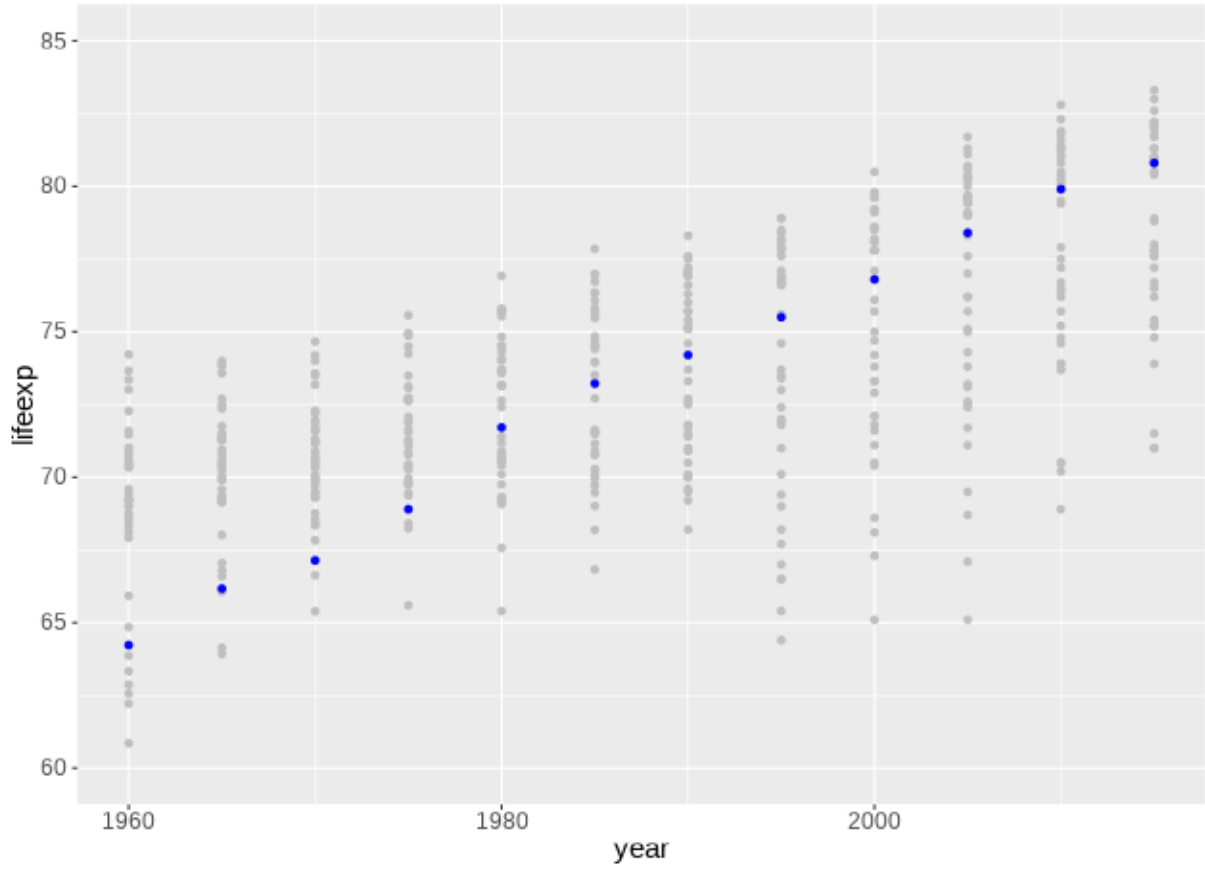
```
gm[continent=="Europe"] %>%
ggplot(aes(year, lifeexp)) +
geom_point(colour="grey75") +
scale_y_continuous(limits=c(60,85))
```



Life Expectancy - Points

Add Portugal (blue)

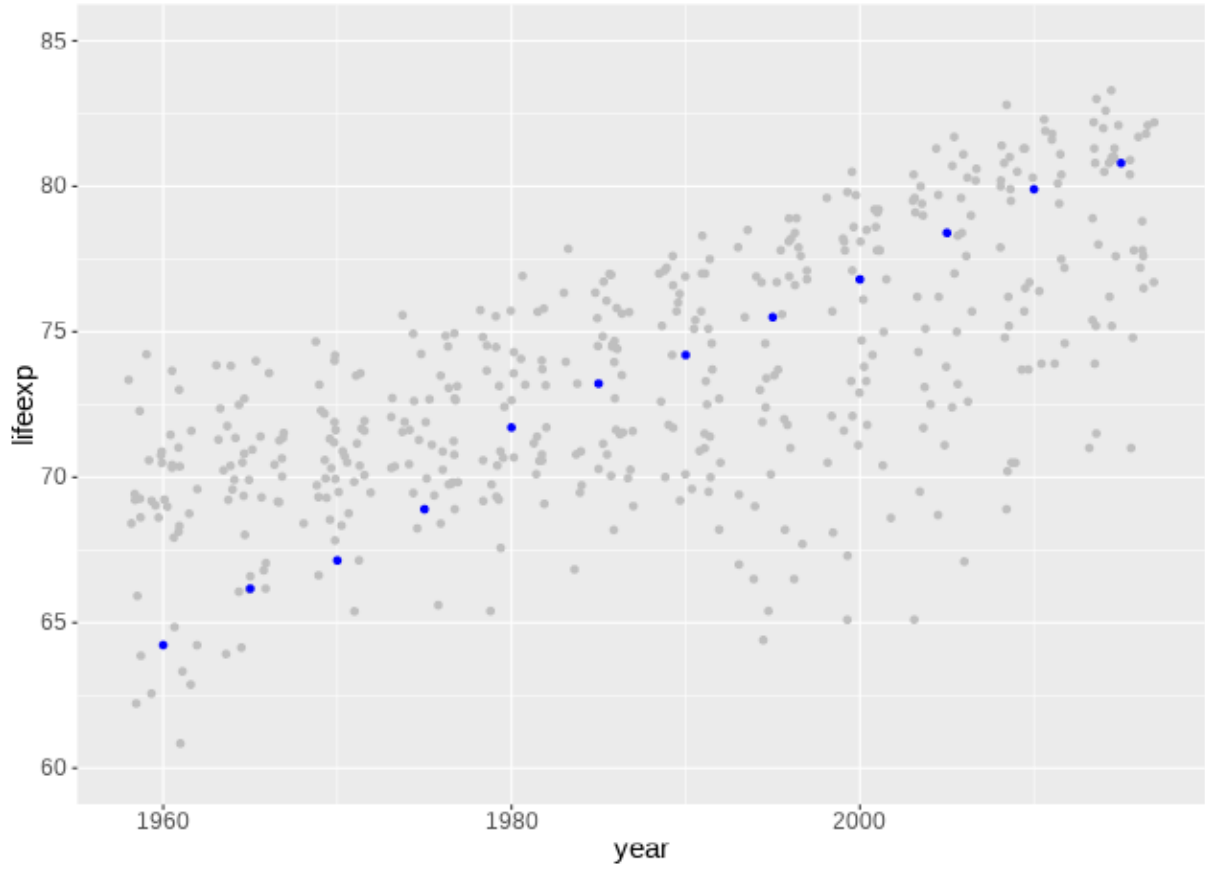
```
gm[continent=="Europe"] %>%
ggplot(aes(year, lifeexp)) +
geom_point(colour="grey75") +
geom_point(data=gmPT,
           colour="blue") +
scale_y_continuous(limits=c(60,85))
```



Life Expectancy - Jitter

Jitter overlapping points

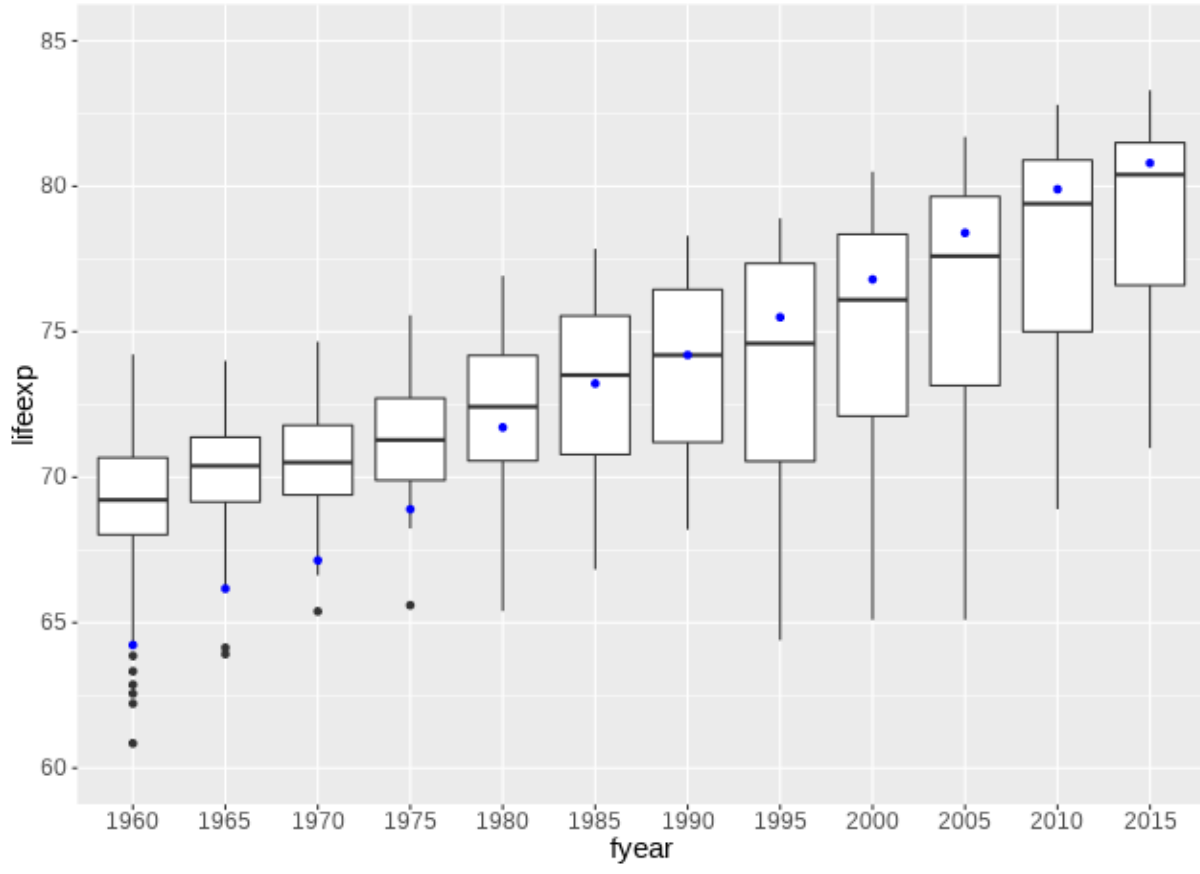
```
gm[continent=="Europe"] %>%
ggplot(aes(year, lifeexp)) +
geom_jitter(colour="grey75") +
geom_point(data=gmPT,
           colour="blue") +
scale_y_continuous(limits=c(60,85))
```



Life Expectancy - Boxplot

```
gm[continent=="Europe"] %>%
ggplot(aes(fyear, lifeexp)) +
geom_boxplot() +
geom_point(data=gmPT,
           colour="blue") +
scale_y_continuous(limits=c(60,85))
```

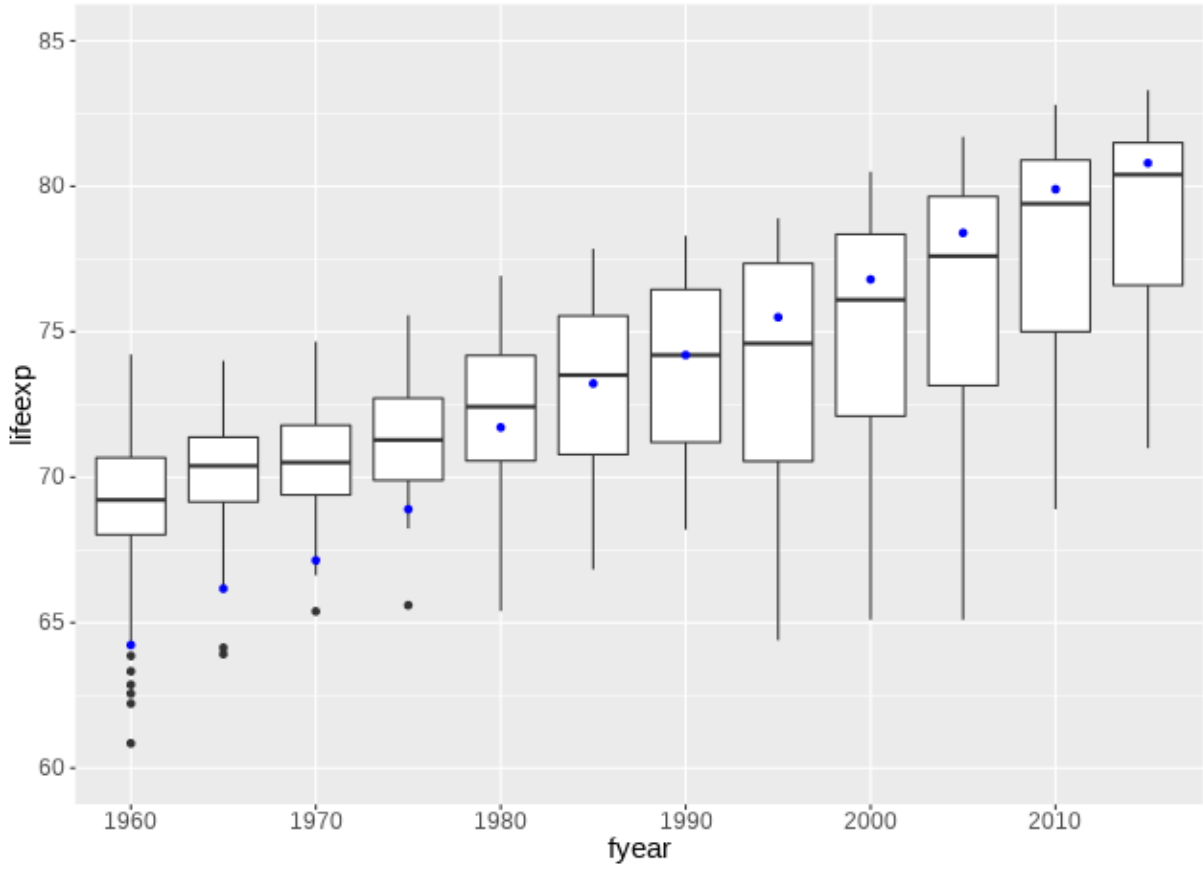
`fyear` is a factor variable



Life Expectancy - Boxplot

Cleaner x-axis

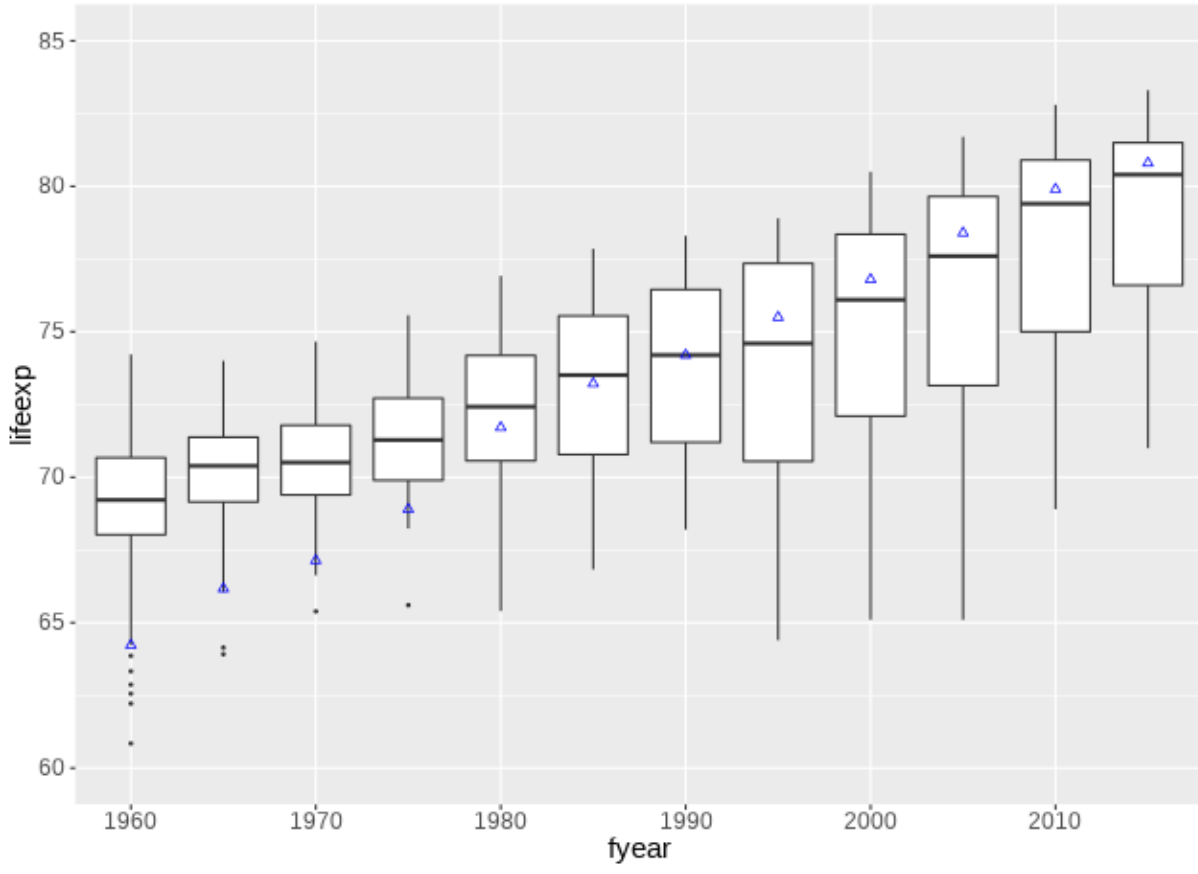
```
gm[continent=="Europe"] %>%
ggplot(aes(fyear, lifeexp)) +
geom_boxplot() +
geom_point(data=gmPT,
           colour="blue") +
scale_x_discrete(
  breaks=seq(1960, 2010, 10)) +
scale_y_continuous(limits=c(60,85))
```



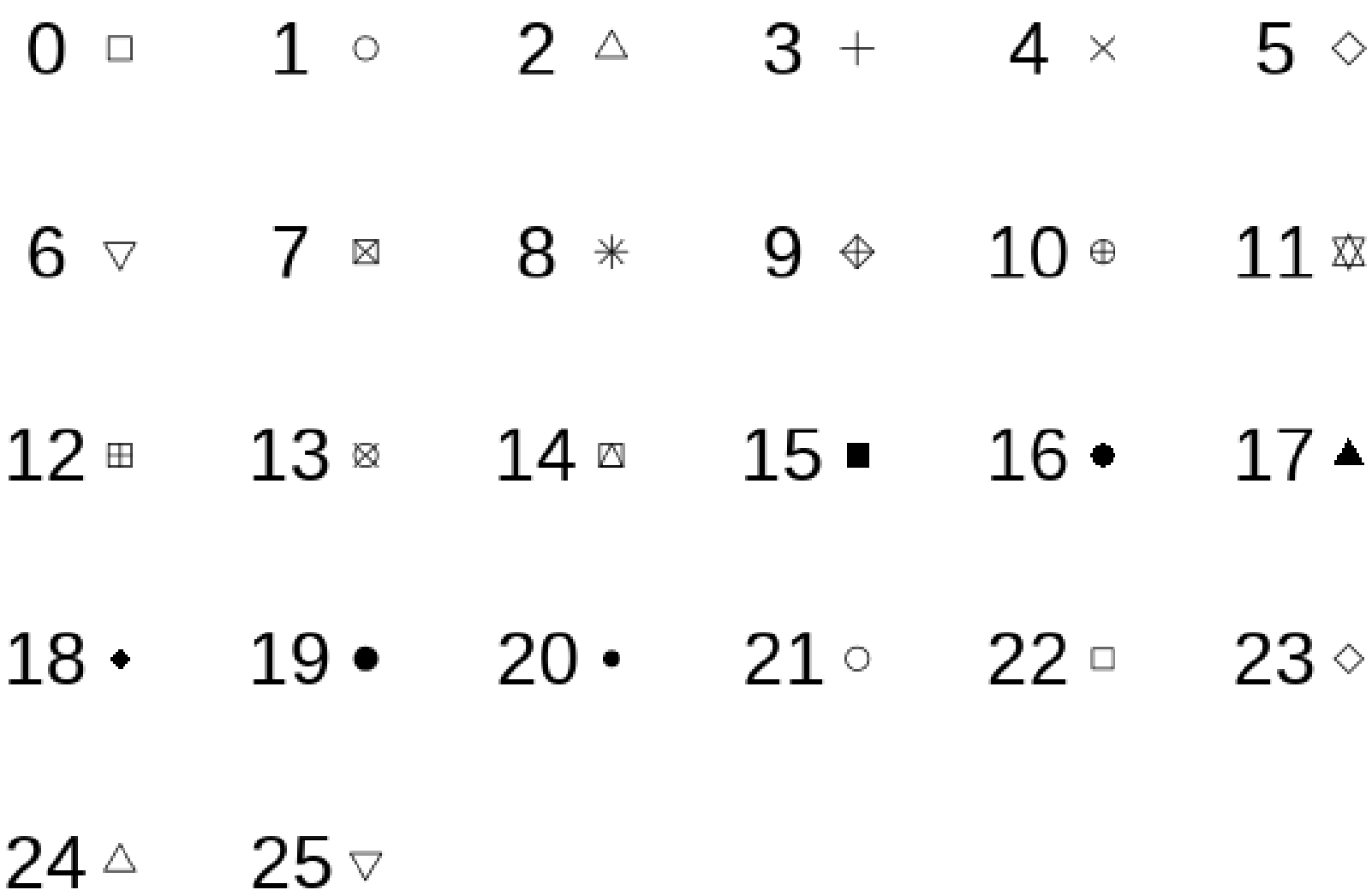
Life Expectancy - Boxplot

Modify points

```
gm[continent=="Europe"] %>%
ggplot(aes(fyear, lifeexp)) +
geom_boxplot(outlier.size = .5) +
geom_point(data=gmPT,
           shape = 2,
           colour="blue") +
scale_x_discrete(
  breaks=seq(1960, 2010, 10)) +
scale_y_continuous(limits=c(60,85))
```



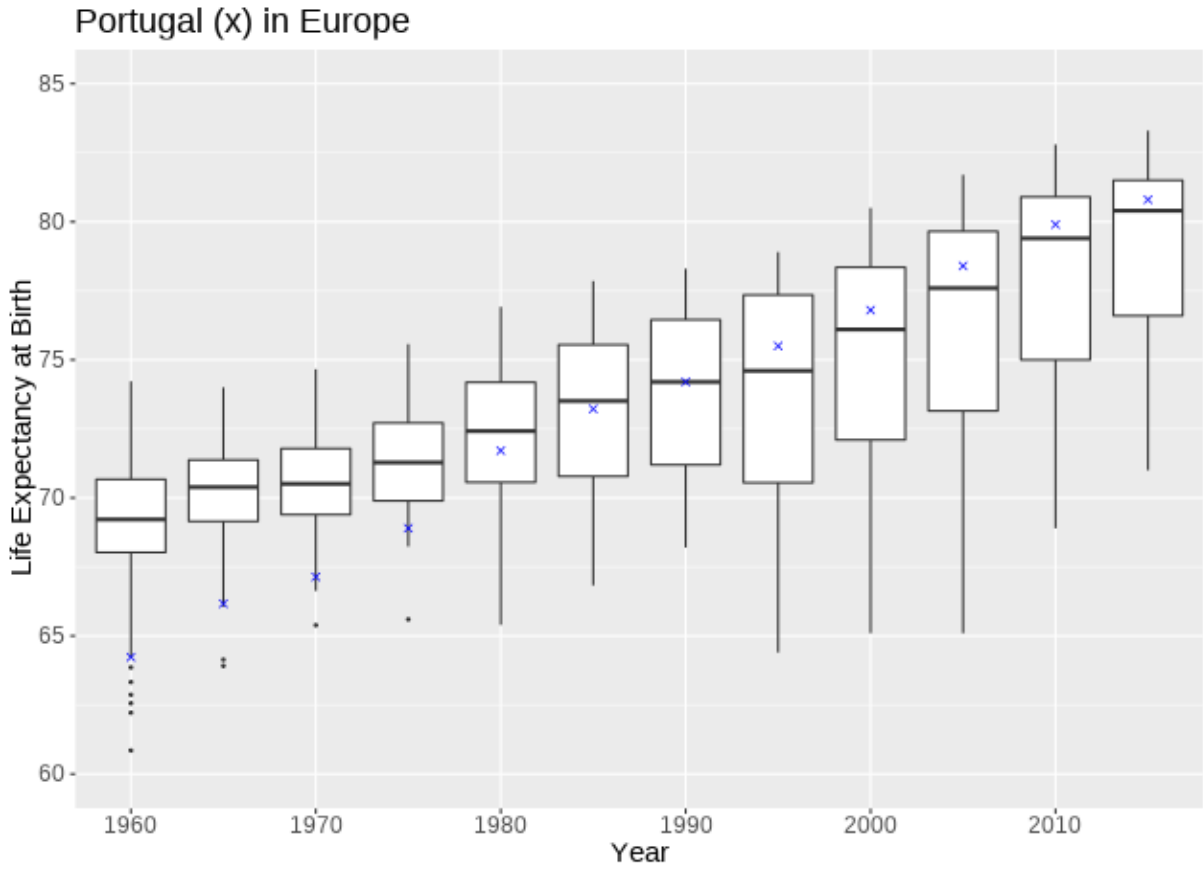
Shapes & Values



Life Expectancy - Boxplot

Final

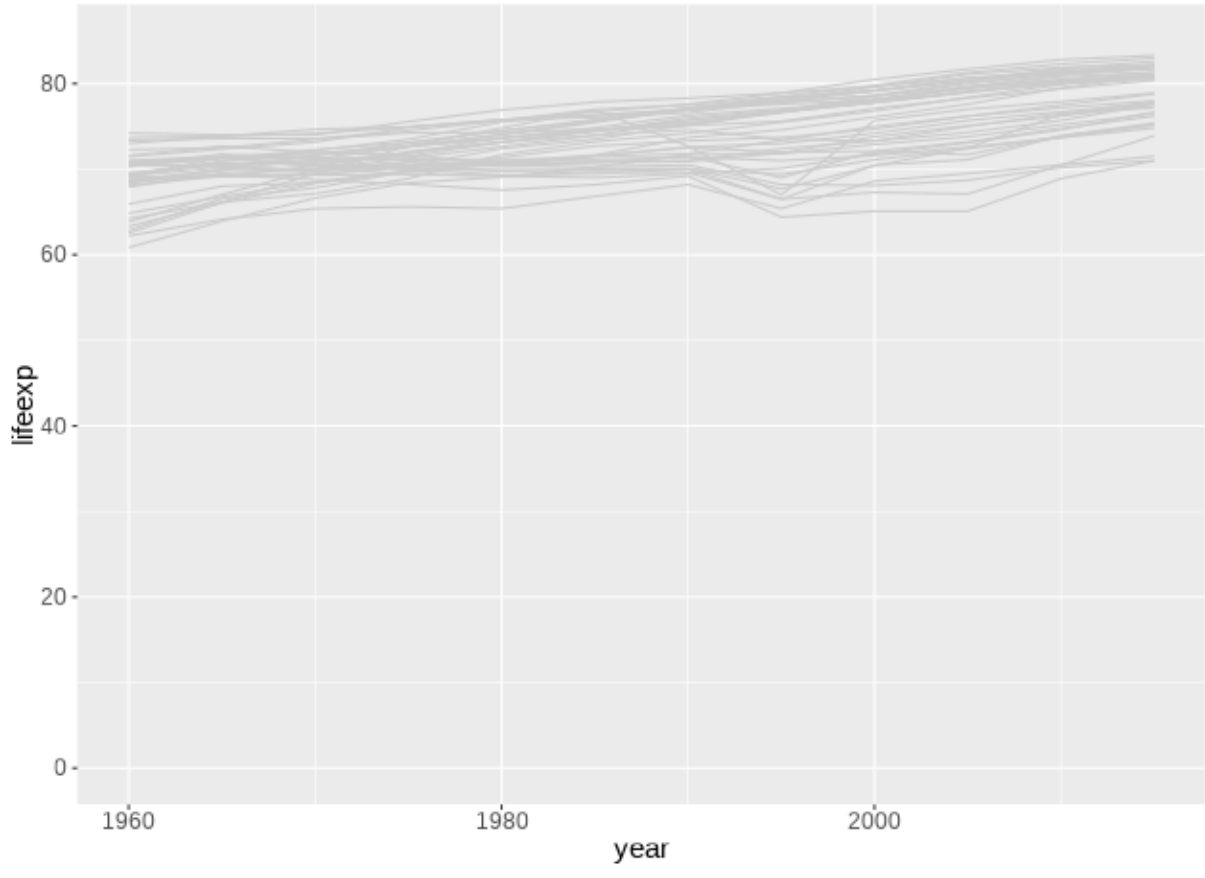
```
gm[continent=="Europe"] %>%
ggplot(aes(fyear, lifeexp)) +
geom_boxplot(outlier.size = .5) +
geom_point(data=gmPT,
           shape = 4,
           colour="blue") +
scale_x_discrete(
  breaks=seq(1960, 2010, 10)) +
scale_y_continuous(limits=c(60,85)) +
ggtitle("Portugal (x) in Europe") +
xlab("Year") +
ylab("Life Expectancy at Birth")
```



Life Expectancy - Line Plot

Line for each country

```
gm[continent=="Europe"] %>%
ggplot(aes(year, lifeexp,
           group=country)) +
geom_line(colour="grey80") +
scale_y_continuous(limits=c(0, 85))
```

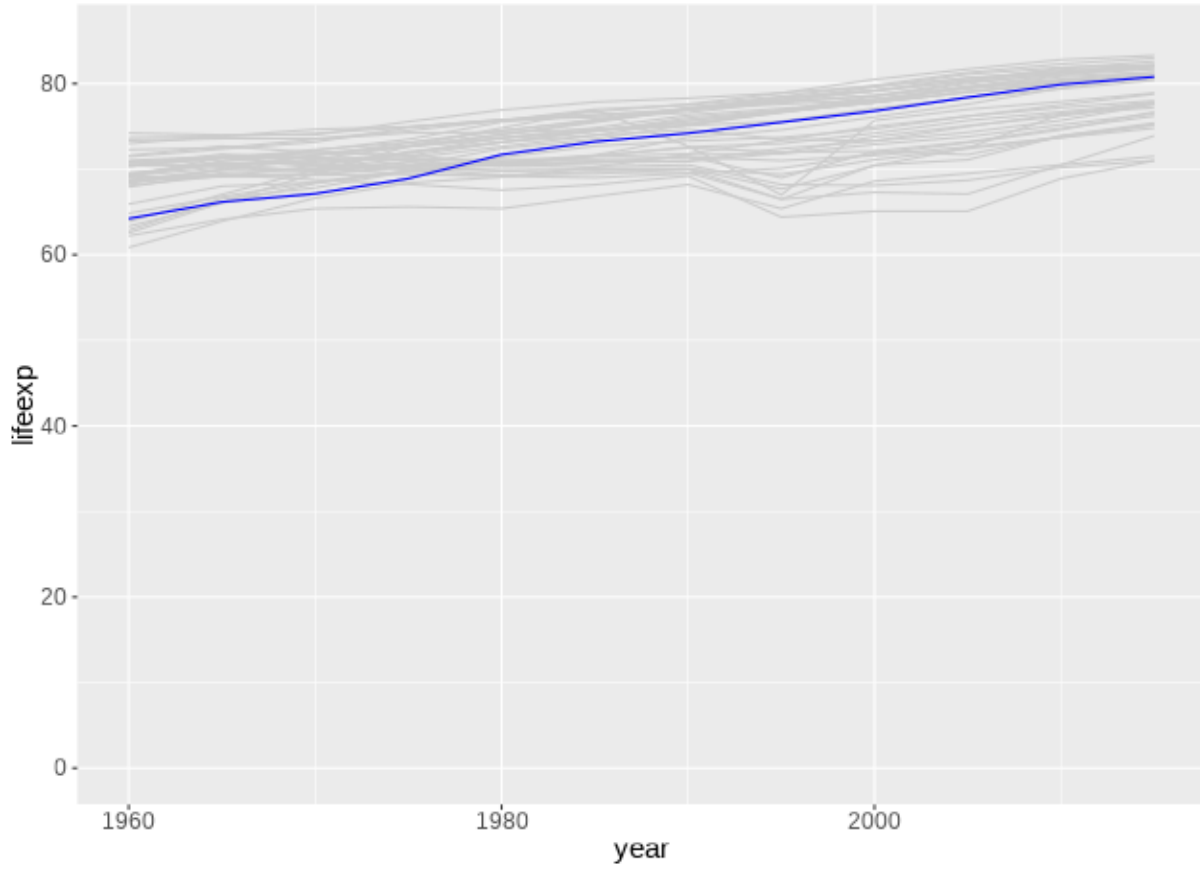


Life Expectancy - Line Plot


```
gm[continent=="Europe"] %>%
ggplot(aes(year, lifeexp,
           group=country)) +
geom_line(colour="grey80") +
geom_line(data=gmPT,
          colour="blue") +
scale_y_continuous(limits=c(0, 85))
```

Used `geom_line()` twice

See "gghighlight" package

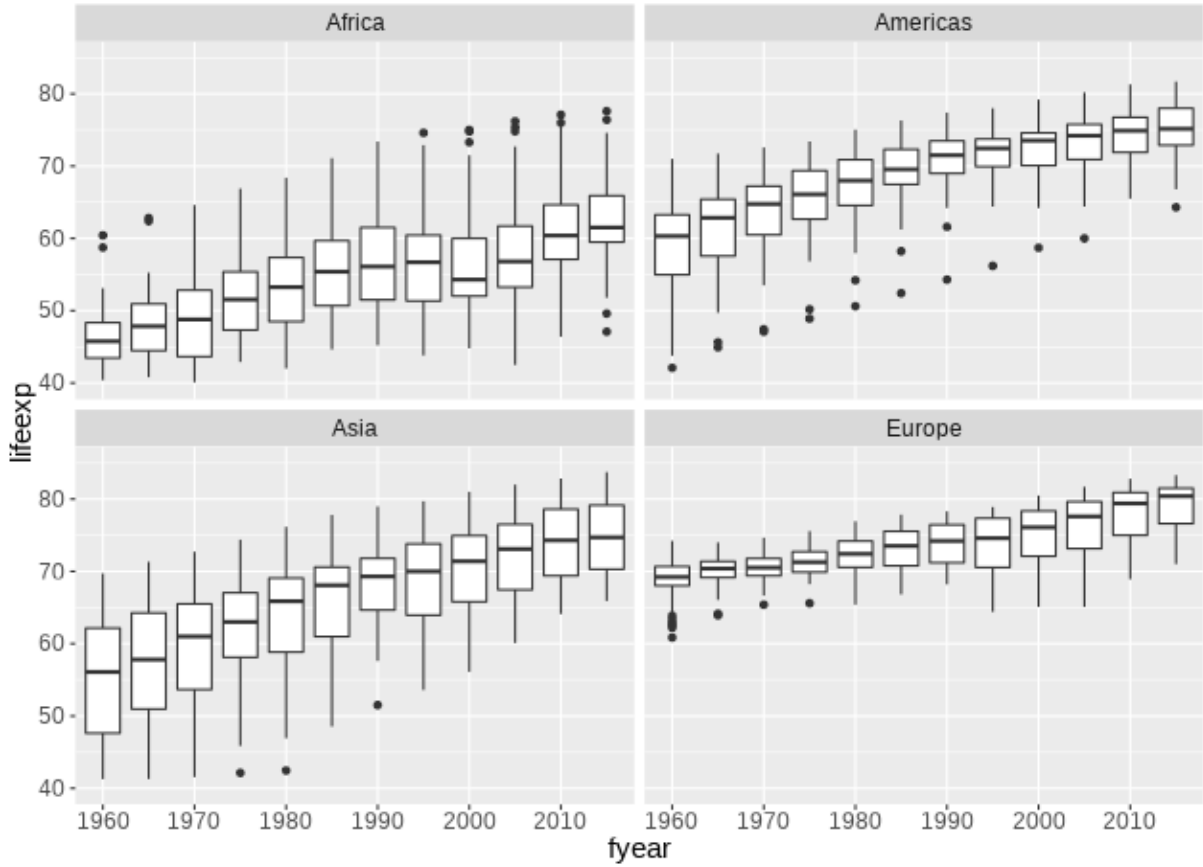


Facetting Plots

Life Expectancy - By Continent

Facetting by continent

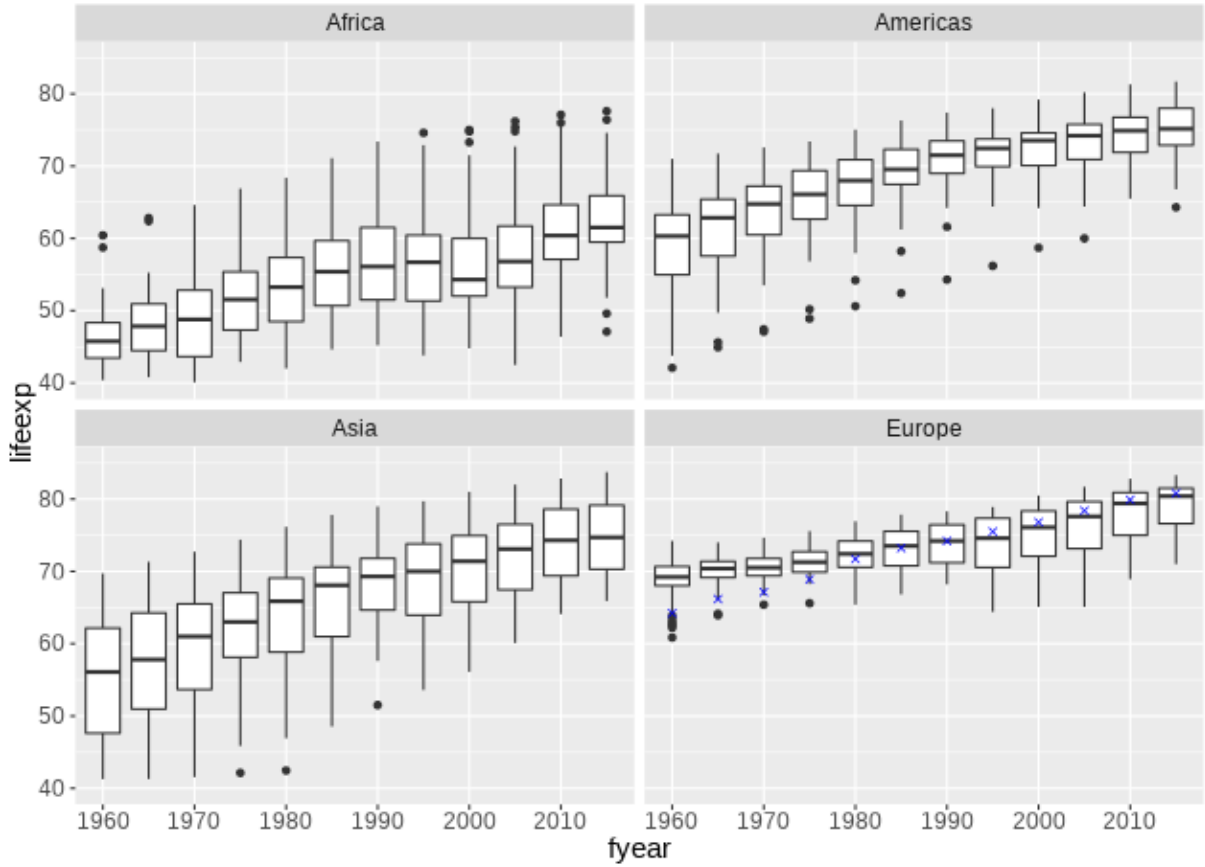
```
gm[continent!="Oceania"] %>%
ggplot(aes(fyear, lifeexp)) +
geom_boxplot() +
scale_x_discrete(
  breaks=seq(1960, 2010, 10)) +
scale_y_continuous(limits=c(40,85)) +
facet_wrap(~continent)
```



Life Expectancy - By Continent

Adding Portugal

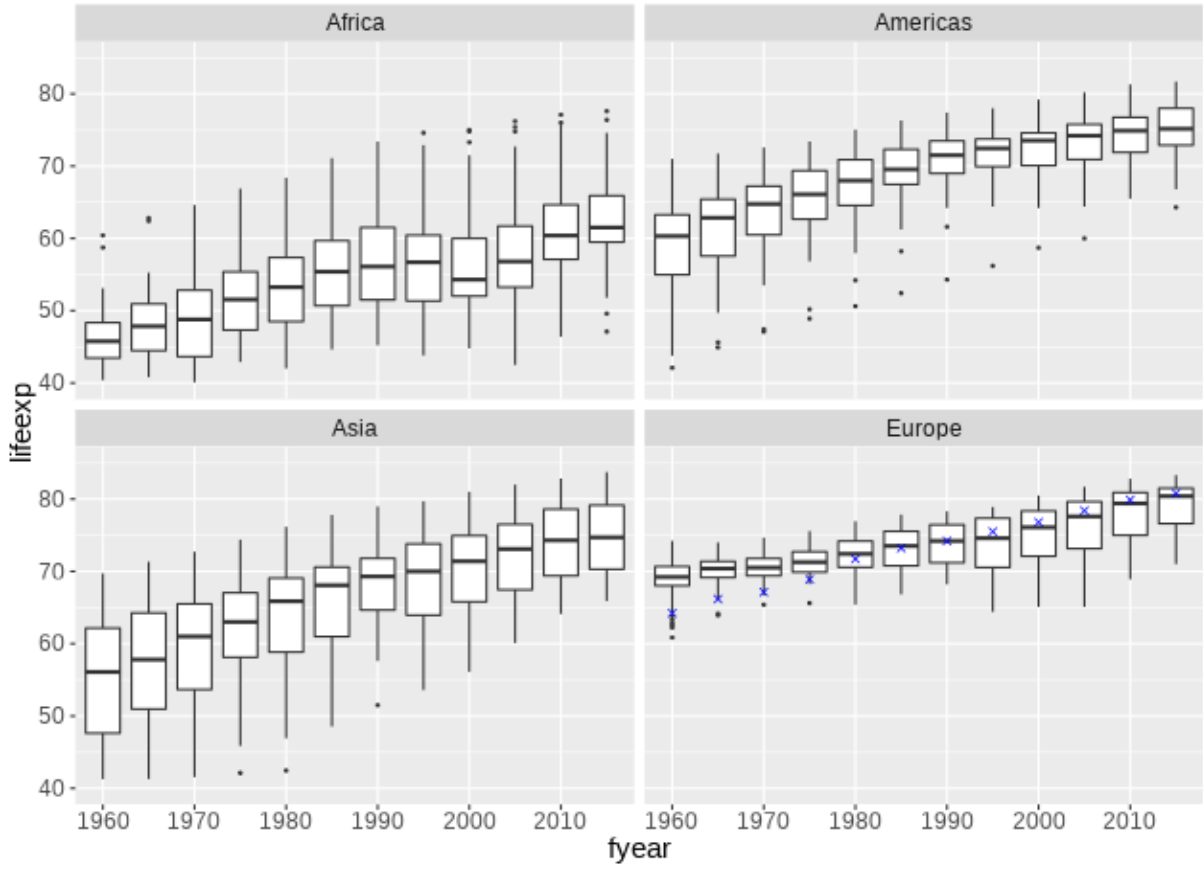
```
gm[continent!="Oceania"] %>%
ggplot(aes(fyear, lifeexp)) +
geom_boxplot() +
geom_point(data=gmPT,
           shape = 4,
           colour="blue") +
scale_x_discrete(
  breaks=seq(1960, 2010, 10)) +
scale_y_continuous(limits=c(40,85)) +
facet_wrap(~continent)
```



Life Expectancy - By Continent

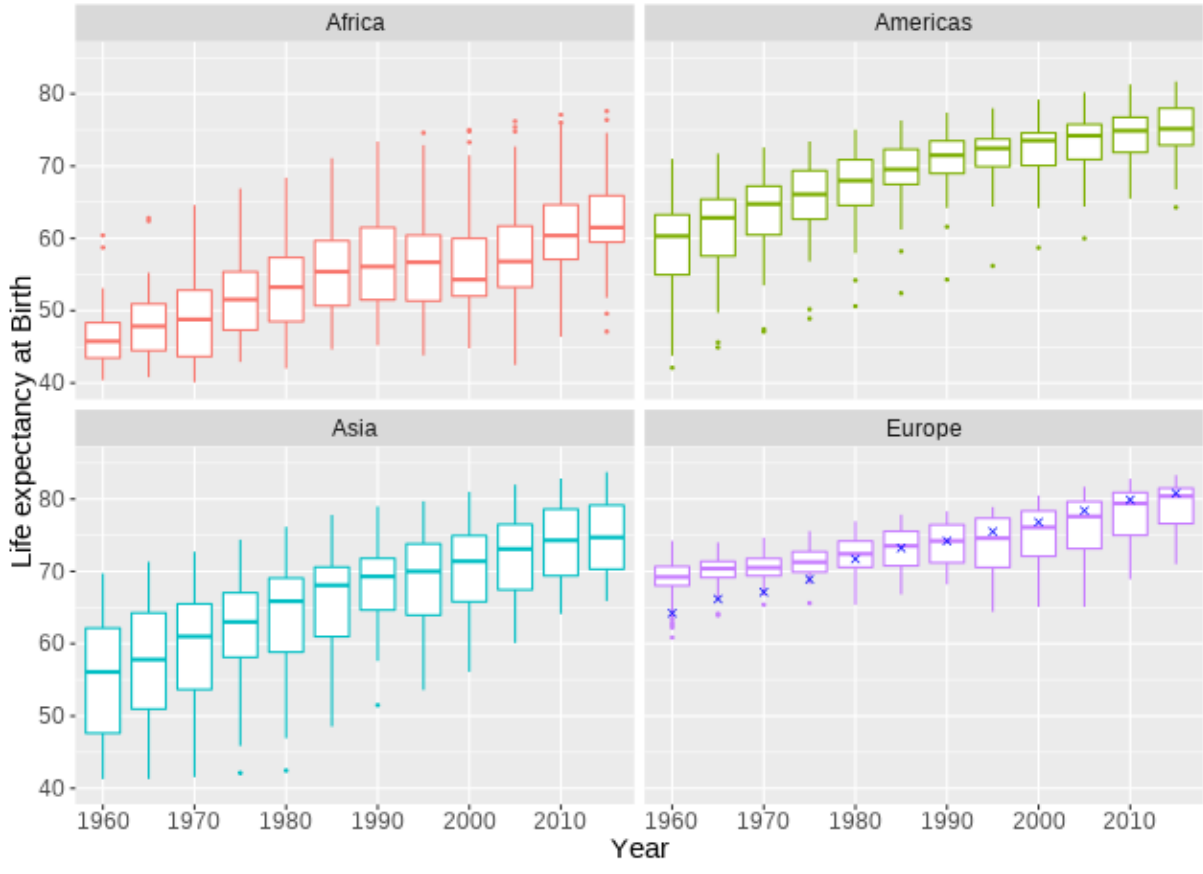
Reduce outlier size

```
gm[continent!="Oceania"] %>%
ggplot(aes(fyear, lifeexp)) +
geom_boxplot(outlier.size = .5) +
geom_point(data=gmPT,
           shape = 4,
           colour="blue") +
scale_x_discrete(
  breaks=seq(1960, 2010, 10)) +
scale_y_continuous(limits=c(40,85)) +
facet_wrap(~continent)
```



Life Expectancy - By Continent

```
gm[continent!="Oceania"] %>%
ggplot(aes(fyear, lifeexp,
           colour = continent)) +
geom_boxplot(outlier.size = .5) +
geom_point(data=gmPT,
           shape = 4,
           colour="blue") +
scale_x_discrete(
  breaks=seq(1960, 2010, 10)) +
scale_y_continuous(limits=c(40,85)) +
facet_wrap(~continent) +
xlab("Year") +
ylab("Life expectancy at Birth") +
theme(legend.position = "none")
```



Exercises

Double Click on "`ggplot2-Exercises.Rproj`"

Open file "`ggplot2-Exercises.Rmd`"

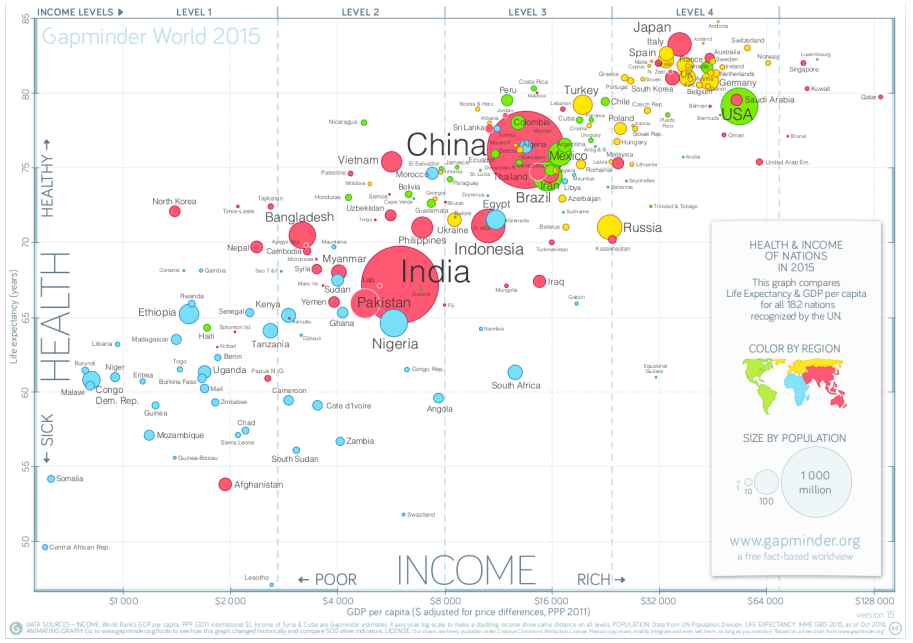
Complete "Exercise 2 - Europe"

How does Portugal perform on health and wealth?

Health versus Wealth over Time

We will only use data for 1960, 1980, 2000 and 2010.

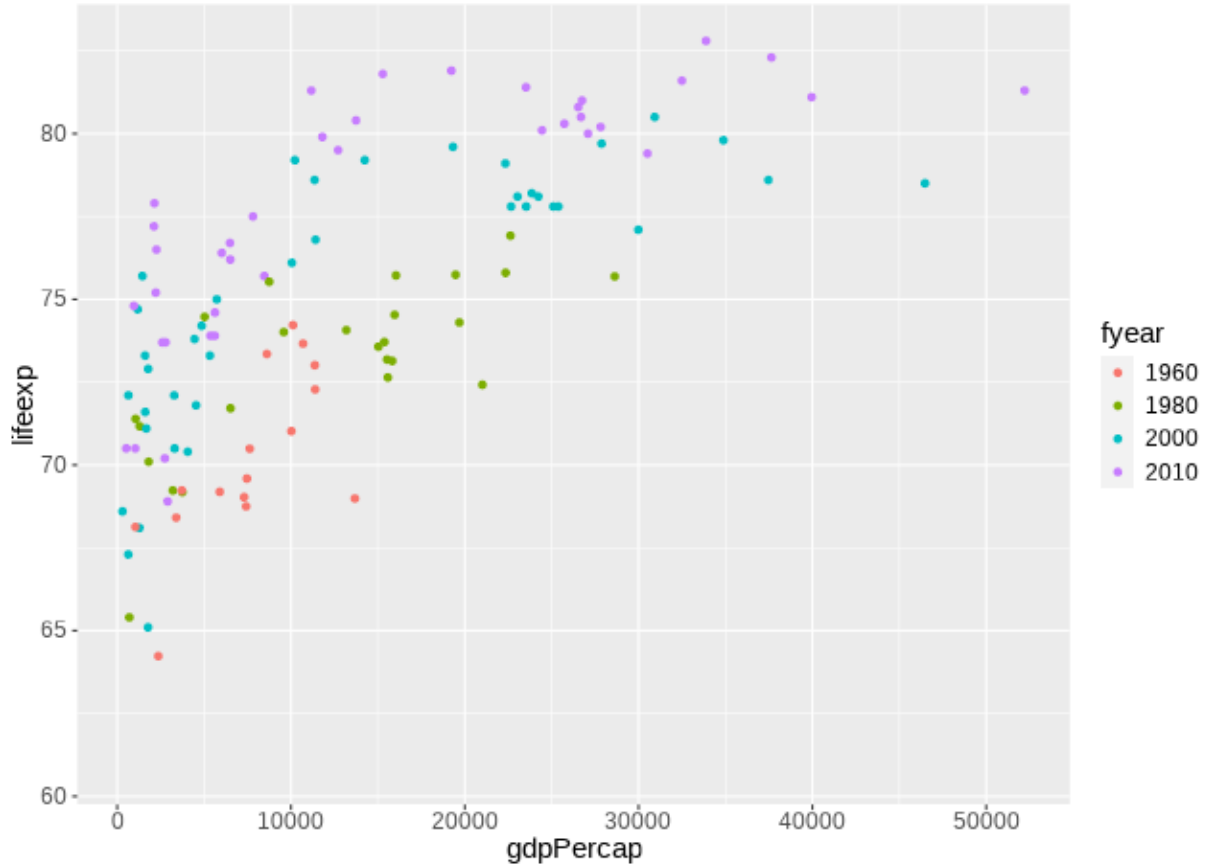
This is essentialy a subset of the famous gapminder plot



Health vs Wealth

Colour code by year

```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp,
       colour=fyear)) +
geom_point()
```

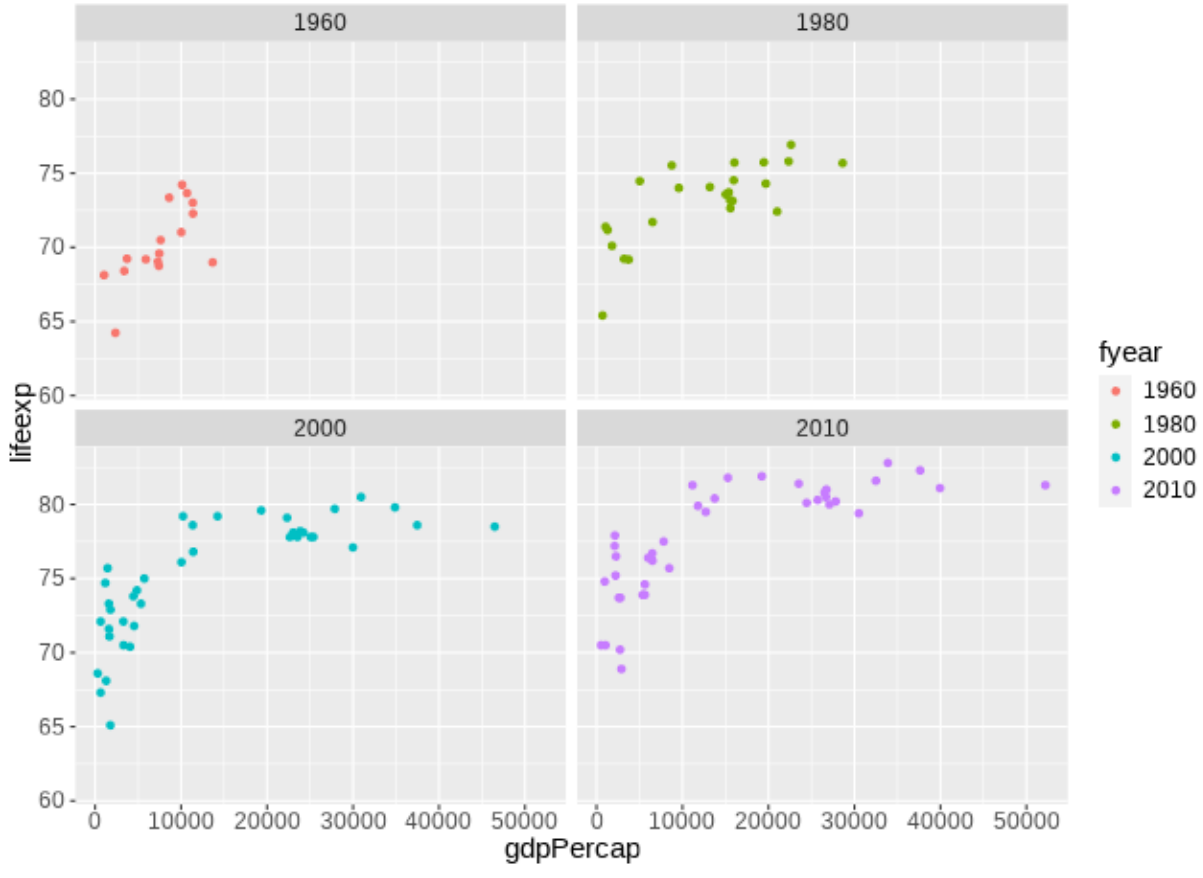


Not easy to interpret

Health vs Wealth - Facet

Plot years separately

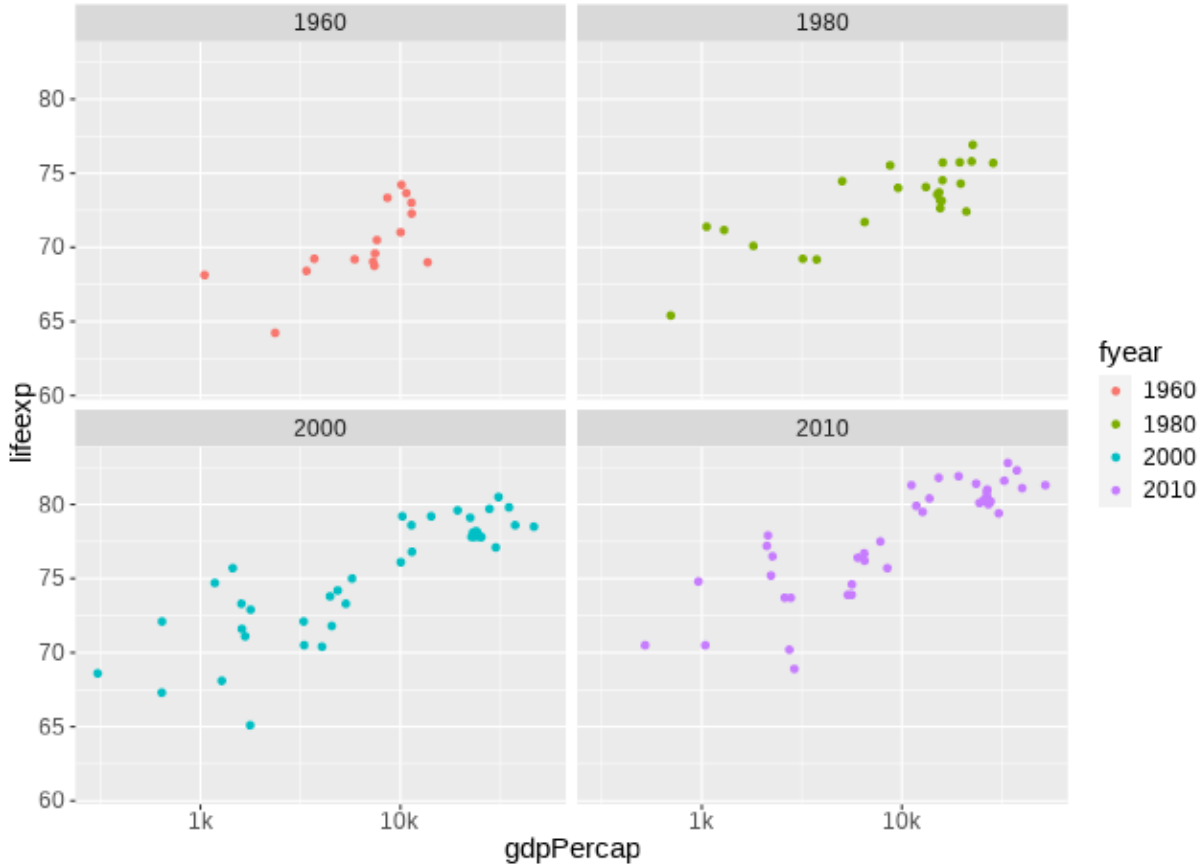
```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp,
       colour=fyear)) +
geom_point() +
facet_wrap(~fyear)
```



Health vs Wealth - Log Scale

Log scale x-axis

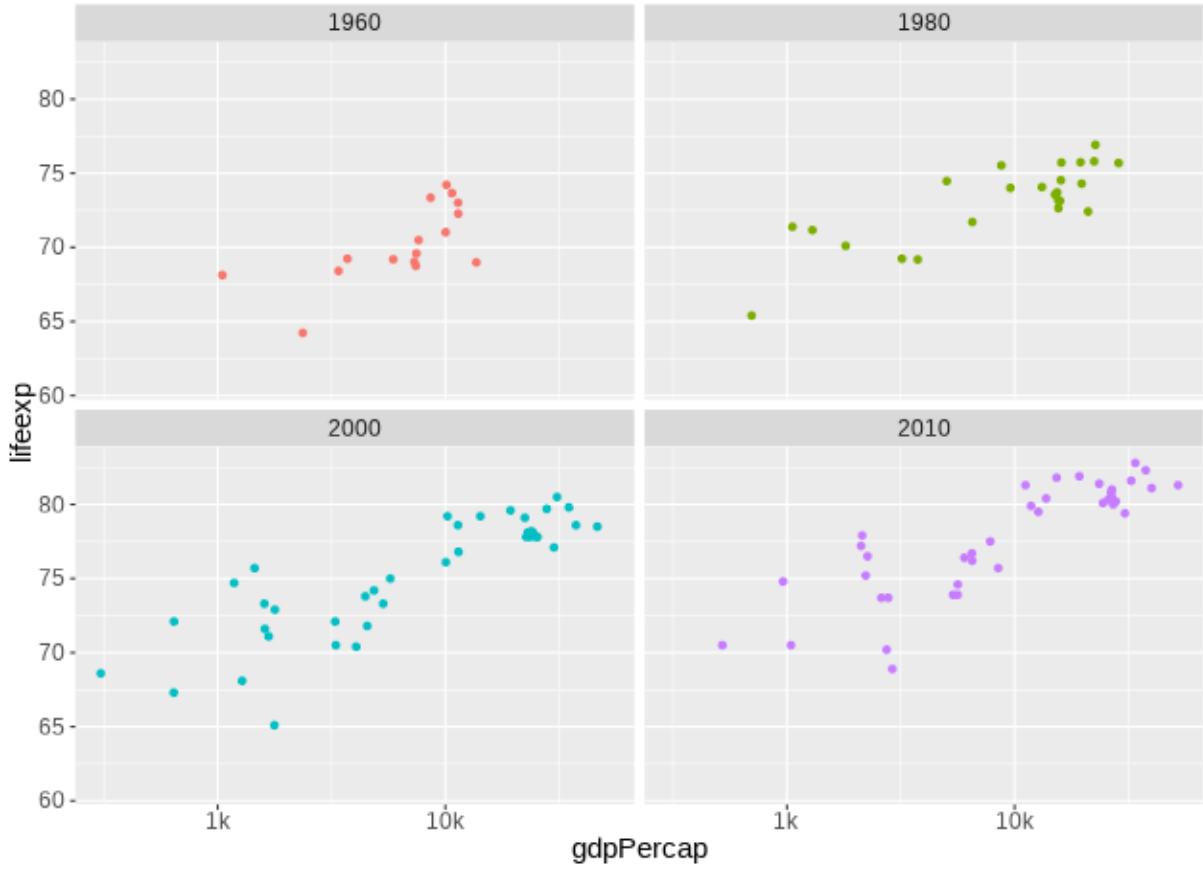
```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp,
       colour=fyear)) +
geom_point() +
scale_x_log10(
breaks = c(10^3, 10^4, 10^5),
labels = c("1k","10k","100k")) +
facet_wrap(~fyear)
```



Health vs Wealth - Log Scale

Remove Legend

```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp,
       colour=fyear)) +
geom_point() +
scale_x_log10(
 breaks = c(10^3, 10^4, 10^5),
 labels = c("1k","10k","100k")) +
facet_wrap(~fyear) +
theme(legend.position = "none")
```



Health vs Wealth - Log Scale

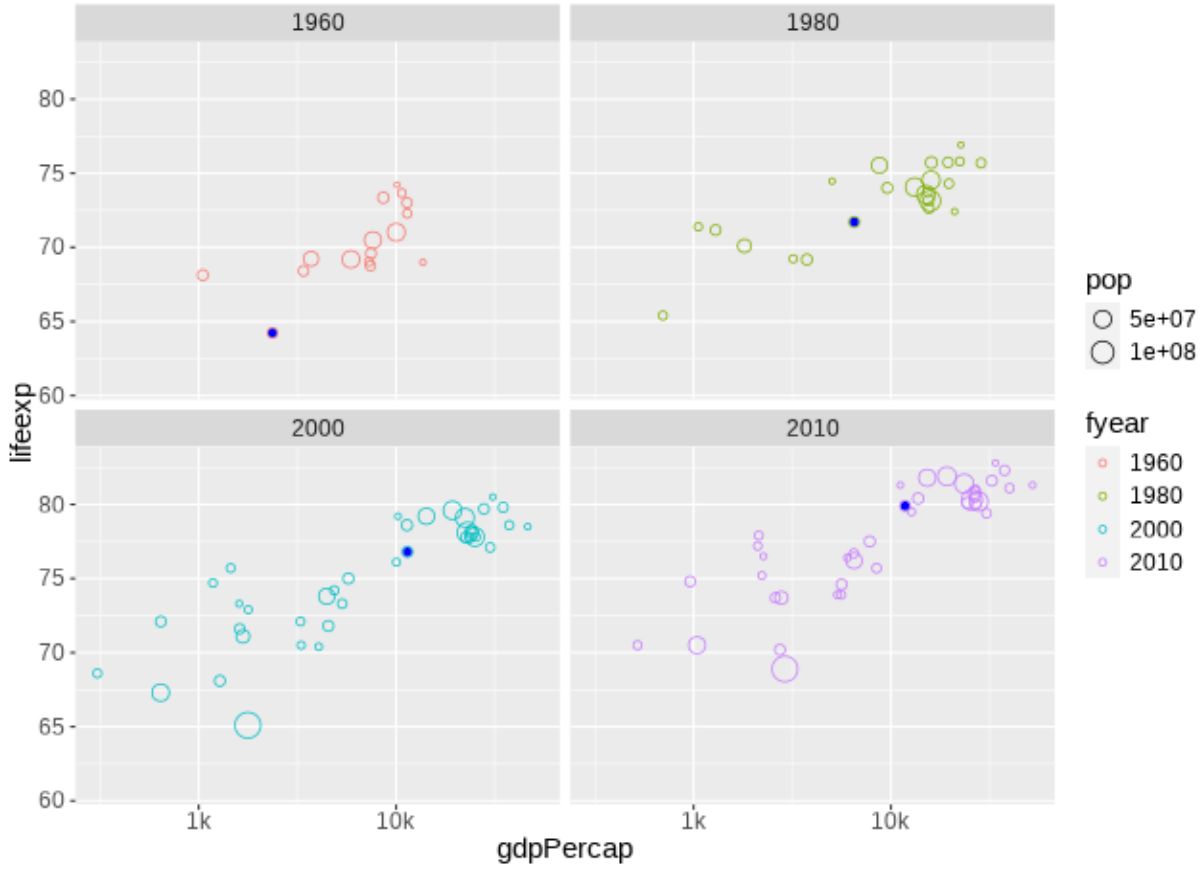
Point size = population

```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp)) +
geom_point(aes(size=pop,
               colour=fyear),
               shape = 21) +
scale_x_log10(
 breaks = c(10^3, 10^4, 10^5),
 labels = c("1k","10k","100k")) +
facet_wrap(~fyear)
```

Health vs Wealth - Portugal

Add Portugal

```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp)) +
geom_point(aes(size=pop,
               colour=fyear),
               shape = 21) +
scale_x_log10(
 breaks = c(10^3, 10^4, 10^5),
 labels = c("1k","10k","100k")) +
facet_wrap(~fyear) +
geom_point(data =
           gmPT[year %in% yrs],
           colour="blue")
```



Health vs Wealth - Portugal

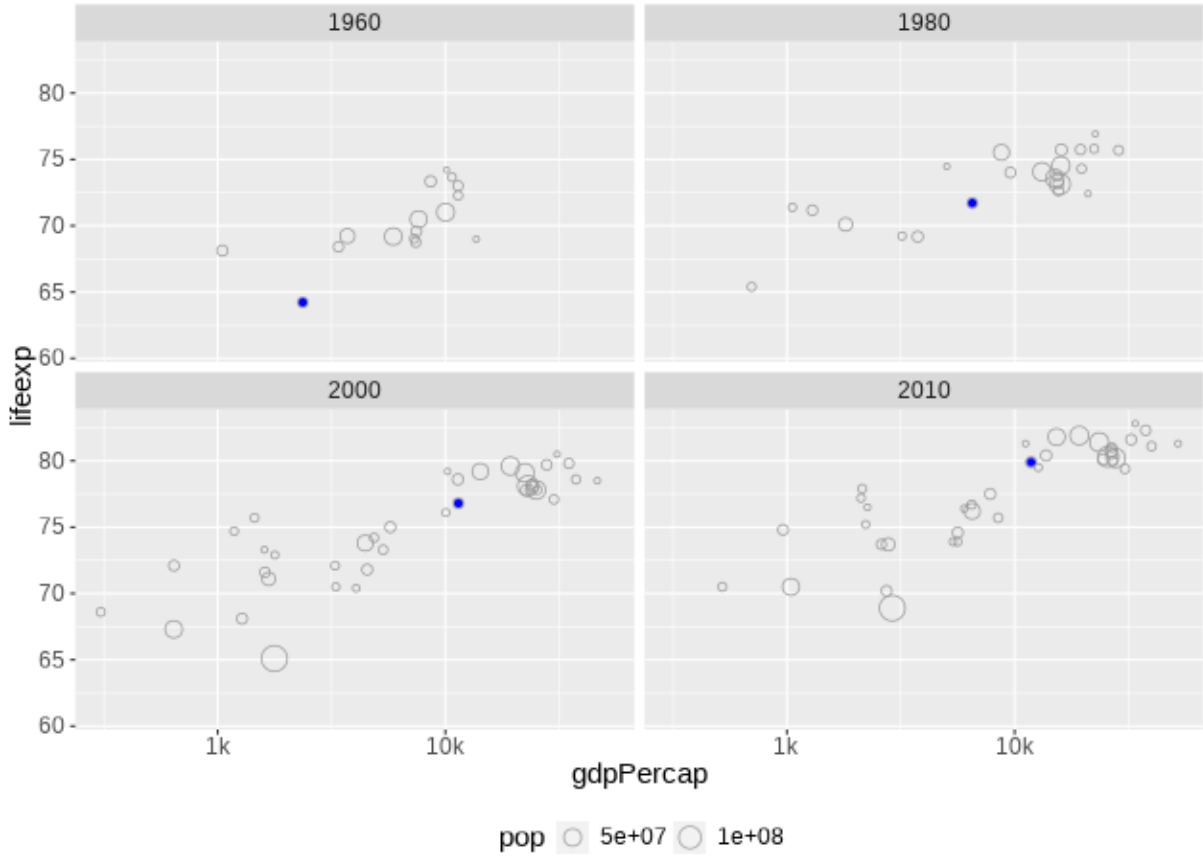
No need for year colours

```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp)) +
geom_point(aes(size=pop),
               colour="grey60",
               shape = 21) +
scale_x_log10(
 breaks = c(10^3, 10^4, 10^5),
 labels = c("1k","10k","100k")) +
facet_wrap(~fyear) +
geom_point(data =
           gmPT[year %in% yrs],
           colour="blue")
```

Health vs Wealth

Move the legend below

```
yrs <- c(1960,1980,2000,2010)
gm[continent=="Europe" &
   year %in% yrs] %>%
ggplot(aes(gdpPercap, lifeexp)) +
geom_point(aes(size=pop),
               colour="grey60",
               shape = 21) +
scale_x_log10(
 breaks = c(10^3, 10^4, 10^5),
 labels = c("1k","10k","100k")) +
facet_wrap(~fyear) +
geom_point(data =
           gmPT[year %in% yrs],
           colour="blue") +
theme(legend.position = "bottom")
```



Exercises

Double Click on "`ggplot2-Exercises.Rproj`"

Open file "`ggplot2-Exercises.Rmd`"

Complete "Exercise 3: Gapminder plot"

Summary

- Answer questions with data visualisations

- Define your plots and `ggplot2` does the rest

- Integrates well with R markdown

- Lots more than this short workshop could show

- Try it out and experiment to learn more!