THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

DATA to insight
An Introduction to Data Analysis

4.11 Exercise: Advanced scatterplots for deeper analysis – *R version*

This exercise will enable you to explore more complicated relationships between variables and the explore the effects of a third and fourth variable, enabling you to view changes over time.

The skills addressed are:

1. Create a scatterplot of two numeric variables, subset by a 3rd variable.
2. Explore the effect of a third and fourth variable using colour and size.

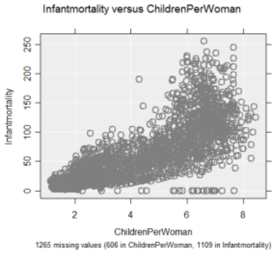
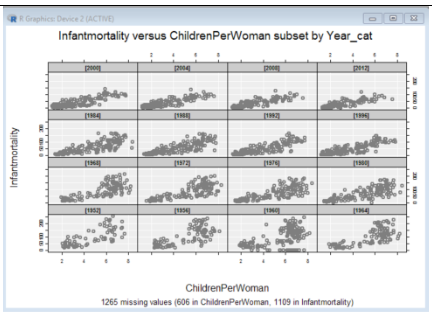
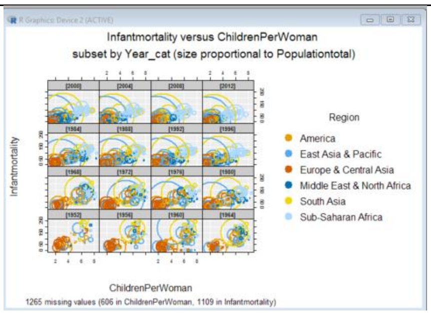
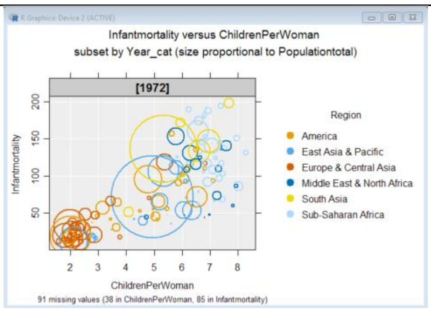
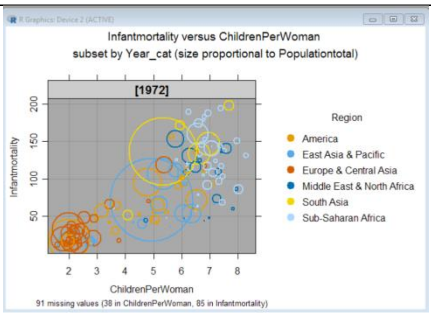
We will use the **gapminder** dataset (but **not** *gapminder_2008*).

Create a scatterplot of two numeric variables, subset by a 3rd variable

We are going to explore the relationship between the variables **Infantmortality** and **ChildrenPerWoman** of countries in the **Gapminder** dataset over time.

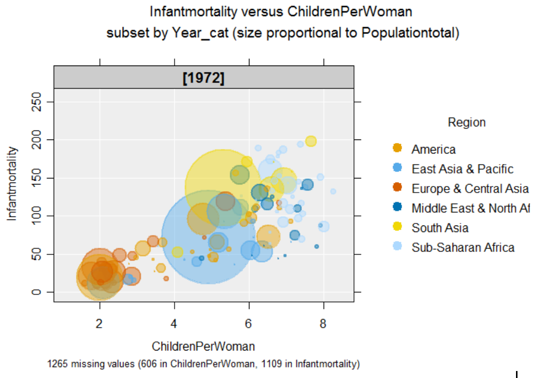
#R Code	Output and/or Commentary

# Setup 	
library(iNZightPlots)
library(FutureLearnData)
data(gapminder)	

***# Scatterplot of* Infantmortaility *against* ChildrenPerWoman** iNZightPlot(**ChildrenPerWoman,Infantmortality** , data=**gapminder**)	
# Subset by Year_cat iNZightPlot(ChildrenPerWoman,Infantmortality, **g1=Year_cat**, data=gapminder)	 R Graphics: Device 2 (ACTIVE) Infantmortality versus ChildrenPerWoman subset by Year_cat [2000] [2004] [2008] [2012] [1984] [1988] [1992] [1996] [1968] [1972] [1976] [1980] [1952] [1956] [1960] [1964] Infantmortality ChildrenPerWoman 1265 missing values (606 in ChildrenPerWoman, 1109 in Infantmortality)
# Change size and colour of points iNZightPlot(ChildrenPerWoman,Infantmortality,g1=Year_cat, data=gapminder, **colby=Region, sizeby=Populationtotal**)	 R Graphics: Device 2 (ACTIVE) Infantmortality versus ChildrenPerWoman subset by Year_cat (size proportional to Populationtotal) [2000] [2004] [2008] [2012] [1984] [1988] [1992] [1996] [1968] [1972] [1976] [1980] [1952] [1956] [1960] [1964] Infantmortality Region America East Asia & Pacific Europe & Central Asia Middle East & North Africa South Asia Sub-Saharan Africa ChildrenPerWoman 1265 missing values (606 in ChildrenPerWoman, 1109 in Infantmortality)
# Show results for 1972 only iNZightPlot(ChildrenPerWoman,Infantmortality,**g1=Year_cat,** **g1.level="[1972]"**,data=gapminder, colby=Region, sizeby=Populationtotal)	 R Graphics: Device 2 (ACTIVE) Infantmortality versus ChildrenPerWoman subset by Year_cat (size proportional to Populationtotal) [1972] Infantmortality 50 100 150 200 2 3 4 5 6 7 8 ChildrenPerWoman Region America East Asia & Pacific Europe & Central Asia Middle East & North Africa South Asia Sub-Saharan Africa 91 missing values (38 in ChildrenPerWoman, 85 in Infantmortality)
***# Darker background** (often easier to see some of the lighter dots)* iNZightPlot(ChildrenPerWoman,Infantmortality, g1=Year_cat, g1.level="[1972]",data=gapminder, colby=Region, sizeby=Populationtotal, **bg="darkgray"**)	 R Graphics: Device 2 (ACTIVE) Infantmortality versus ChildrenPerWoman subset by Year_cat (size proportional to Populationtotal) [1972] Infantmortality 50 100 150 200 2 3 4 5 6 7 8 ChildrenPerWoman Region America East Asia & Pacific Europe & Central Asia Middle East & North Africa South Asia Sub-Saharan Africa 91 missing values (38 in ChildrenPerWoman, 85 in Infantmortality)

# Try transparency and smaller points *(removed garkgray)*

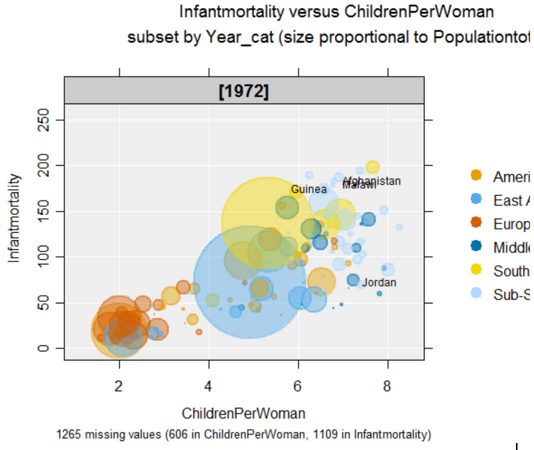
```
iNZightPlot(ChildrenPerWoman,Infantmortality, g1=Year_cat,
        g1.level="[1972]",data=gapminder, colby=Region,
        sizeby=Populationtotal, alpha=.45, cex.dotpt=.5)
```



# Try subsetting by different years, e.g. g1.level="[1976]",

# Label some of the extreme points *(ask for 4)*

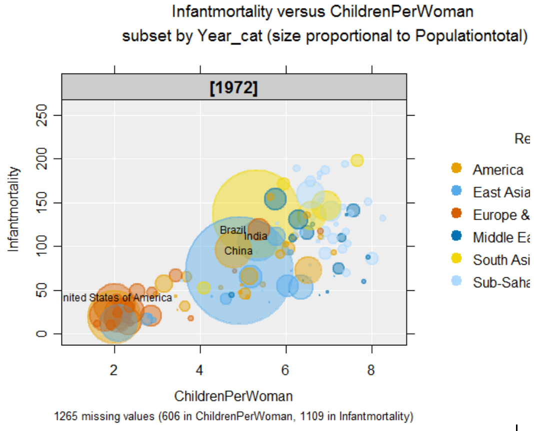
```
iNZightPlot(ChildrenPerWoman,Infantmortality, g1=Year_cat,
        g1.level="[1972]",data=gapminder, colby=Region,
        sizeby=Populationtotal, alpha=.45, cex.dotpt=.5,
        locate.extreme=4, locate=Country)
```



# Label some specific countries

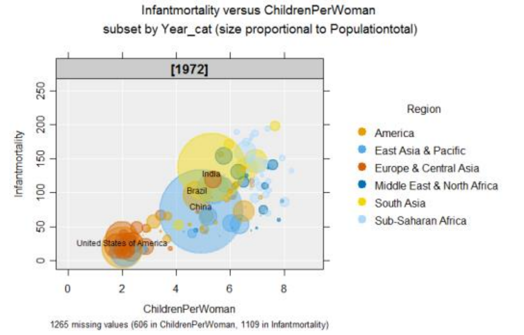
```
ids = (1:nrow(gapminder))[gapminder$Country %in%
                c("United States of America","China","Brazil", "India")]

iNZightPlot(ChildrenPerWoman,Infantmortality, g1=Year_cat,
        g1.level="[1972]",data=gapminder, colby=Region,
        sizeby=Populationtotal, alpha=.45, cex.dotpt=.5,
        locate.id=ids, locate=Country)
```



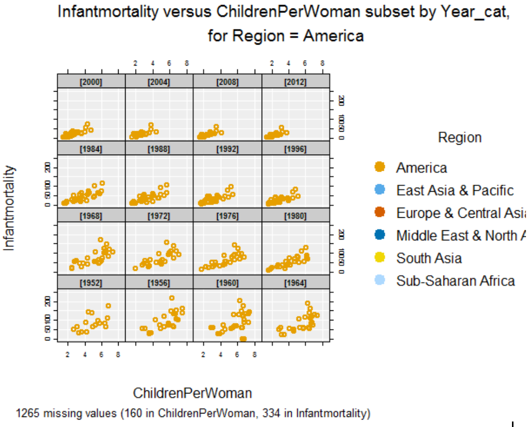
# Allow a little more room on left to accommodate label

```
iNZightPlot(ChildrenPerWoman,Infantmortality, g1=Year_cat,
        g1.level="[1972]",data=gapminder, colby=Region,
        sizeby=Populationtotal, alpha=.45, cex.dotpt=.5,
        locate.id=ids, locate=Country, xlim=c(0,9))
```



# Subset by a fourth variable (Region)

```
iNZightPlot(ChildrenPerWoman, Infantmortality, g1=Year_cat,
g2=Region, g2.level="America", data=gapminder, colby=Region)
```



# Play through the years

```
for (k in levels(gapminder$Year_cat)) {
iNZightPlot(ChildrenPerWoman,Infantmortality, g1=Year_cat,
        g1.level=k, data=gapminder, colby=Region,
        sizeby=Populationtotal, alpha=.45, cex.dotpt=.5,
        locate.id=ids, locate=Country)
        Sys.sleep(1)
}
```

- **Play some more with these settings and try other variables**
- For even more settings, type **?inzpar** into R to get help on the inzpar, or type **inzpar** to just get a complete list (last time I looked the help file wasn't entirely complete)

To discuss issues related to this Exercise,

go to **https://gitter.im/iNZightVIT/d2i-R-discussion**

*To be able to post to the list you will have to set up a (free) account on **Github***
https://github.com/login

If your question relates to an Exercise, say which one you are talking about!