DataVizA Tutorial: Plotting many variables: Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 5

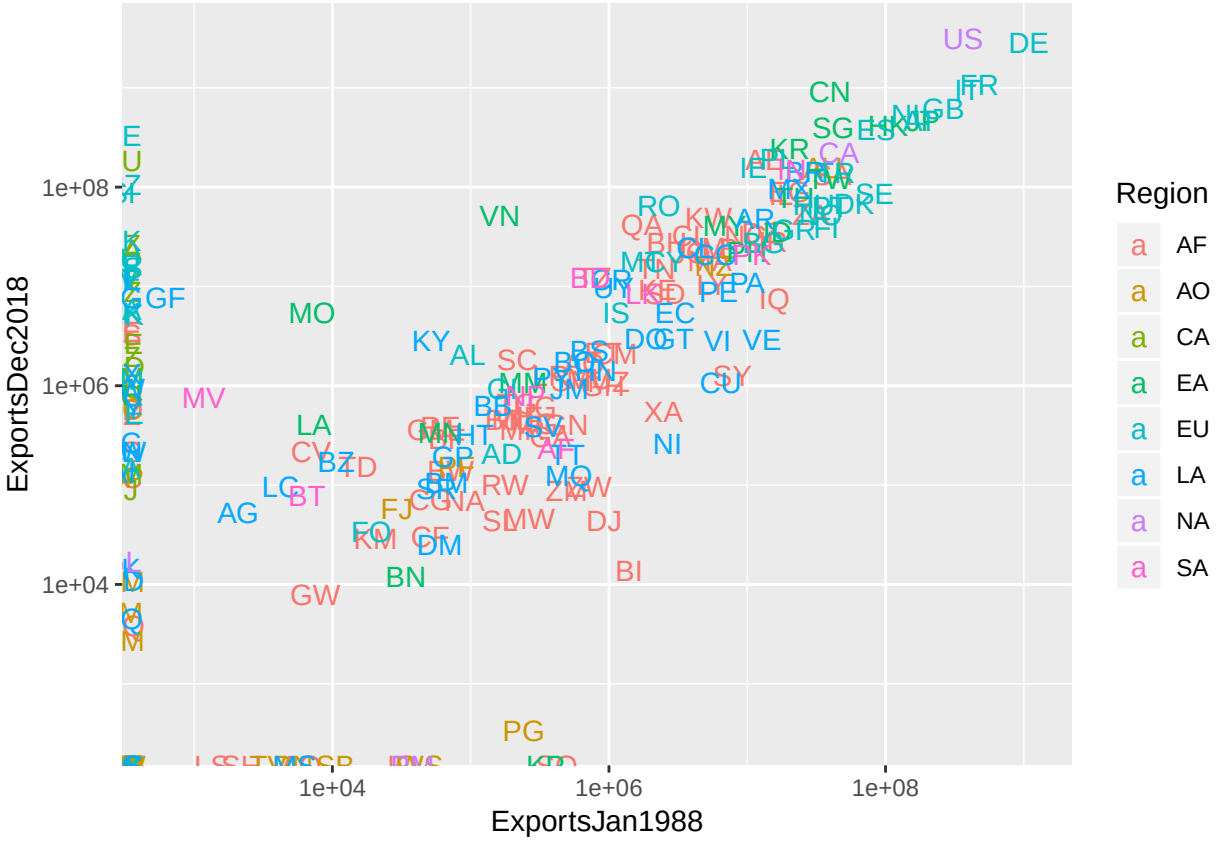
Swiss Data

1. Repeat question 3 from last week but use text to represent the country and color to represent the region. Note these regions are very loosely defined for instance Mexico could easily be in North America, and the Central Asian (CA) could be better named the former Soviet region, etc.

```
library(tidyverse)
SwissE<-readRDS('SwissExport.rds')
ggplot(SwissE,
       aes(x=ExportsJan1988,
           y=ExportsDec2018,
           col=Region, #Region mapped to color
           label=Country))+ #Country mapped to label
         geom_text()+ #use geom_text not geom_point
  scale_x_log10()+scale_y_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```



2. Discuss this plot.

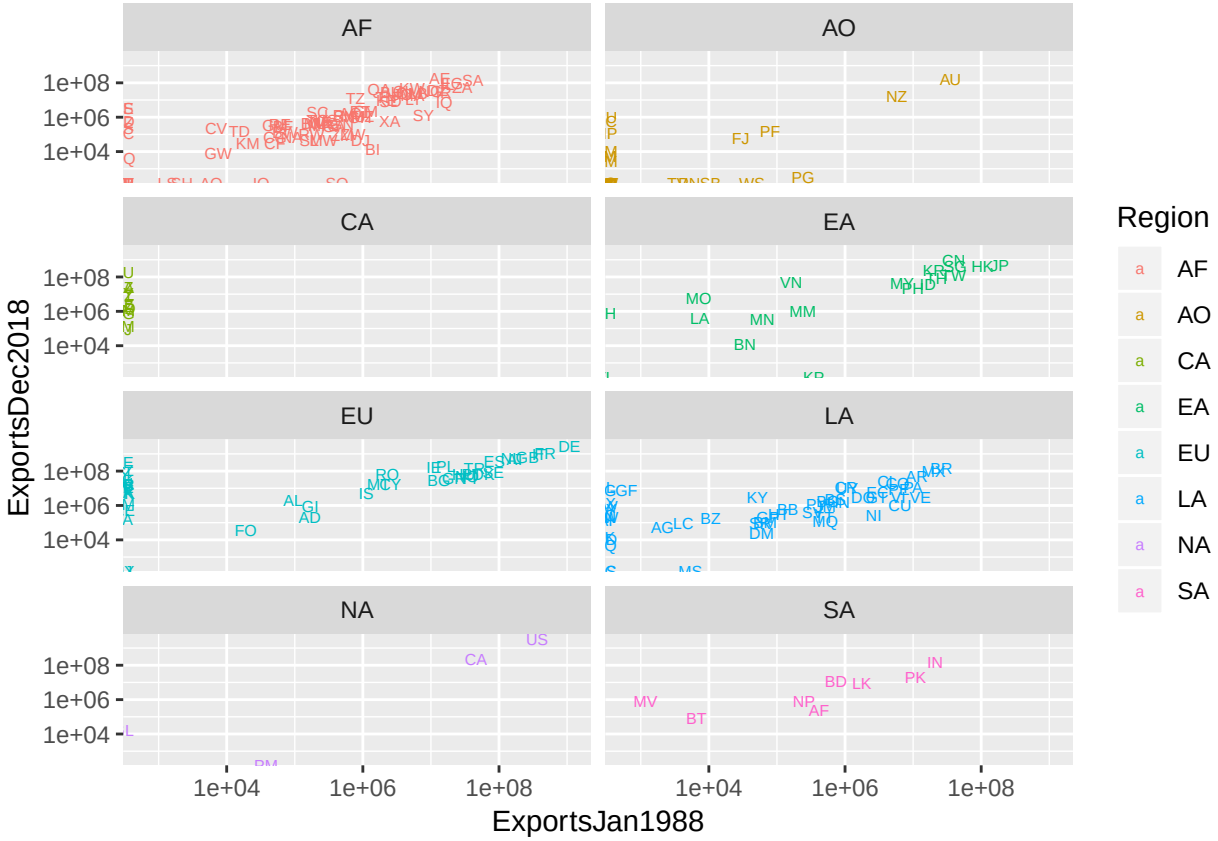
Some interesting things to note are that the many EU nations occupy the top right hand side of the plot. This is to be expected since Switzerland is in the middle of Europe and exports mostly to its neighbours. Note that Frech Guyana (GF) went from having almost no exports in January 1988 to a moderately high level of exports in December 2018, while Papua New Guinea (PG) went in the opposite direction. It should be noted that monthly data are more volatile and a more accurate understanding could be obtained by looking at a yearly average.

3. Since the previous plots suffer from a little overplotting, repeat the analysis but facet by region.

```
library(tidyverse)
SwissE<-readRDS('SwissExport.rds')
ggplot(SwissE,
       aes(x=ExportsJan1988,
           y=ExportsDec2018,
           col=Region, #Region mapped to color
           label=Country))+ #Country mapped to label
         geom_text(size=2)+ #use geom_text not geom_point
  scale_x_log10()+scale_y_log10()+
  facet_wrap(~Region,nrow = 4)
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

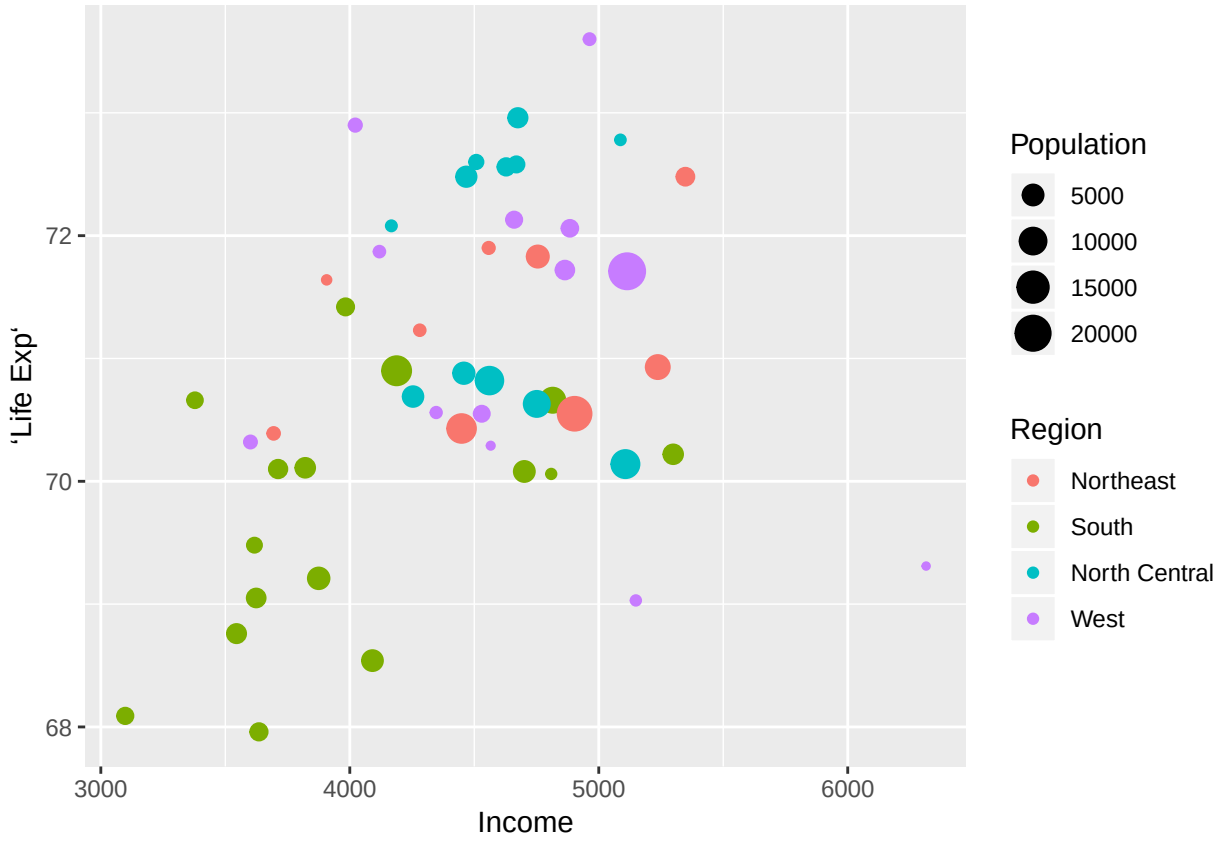


U.S. States Data

The data in the file *USStateData.rds* contains information on demography and geography for the states of the United States. These data were measured in 1977.

4. Construct a bubble chart of income per capita (x axis) against life expectancy (y axis). The size of the bubble should reflect the State population, and the color should reflect the region.

```
USState<-readRDS('USStateData.rds')
ggplot(USState,
       aes(x=Income,
           y=`Life Exp`, #Note use of ``
           col=Region, #Region mapped to color
           size=Population))+ #Country mapped to label
         geom_point()
```



5. Discuss this plot.

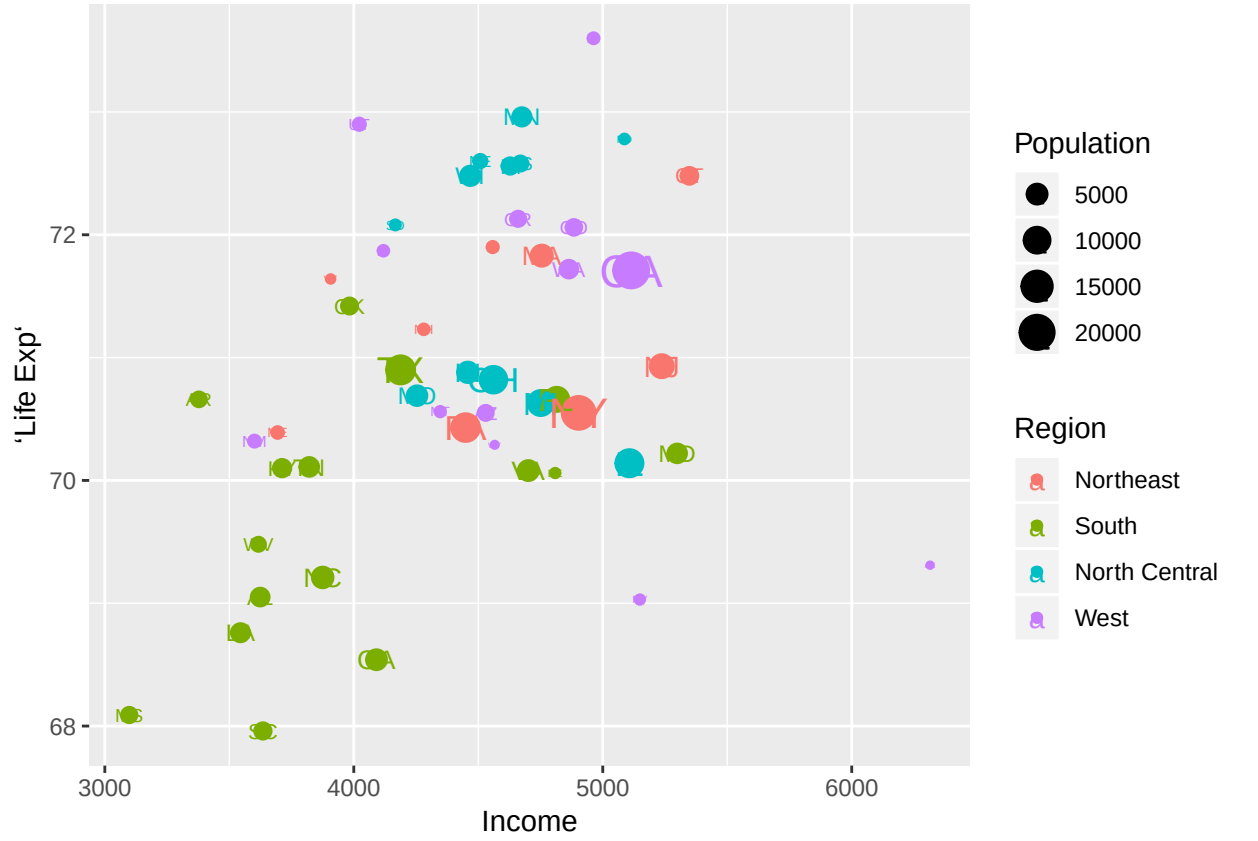
Some things to notice are the following. The South region has mostly lower income states with lower life expectancy. Generally the larger states have higher income and moderately high life expectancy. There is a small state that is an outlier for having the highest per capita income.

6. Add the state abbreviations as text on the plot

```
#The obvious thing to do would be to take the previous code
#and add geom_text and a label aesthetic

ggplot(USState,
       aes(x=Income,
           y=`Life Exp`, #Note use of ``
```

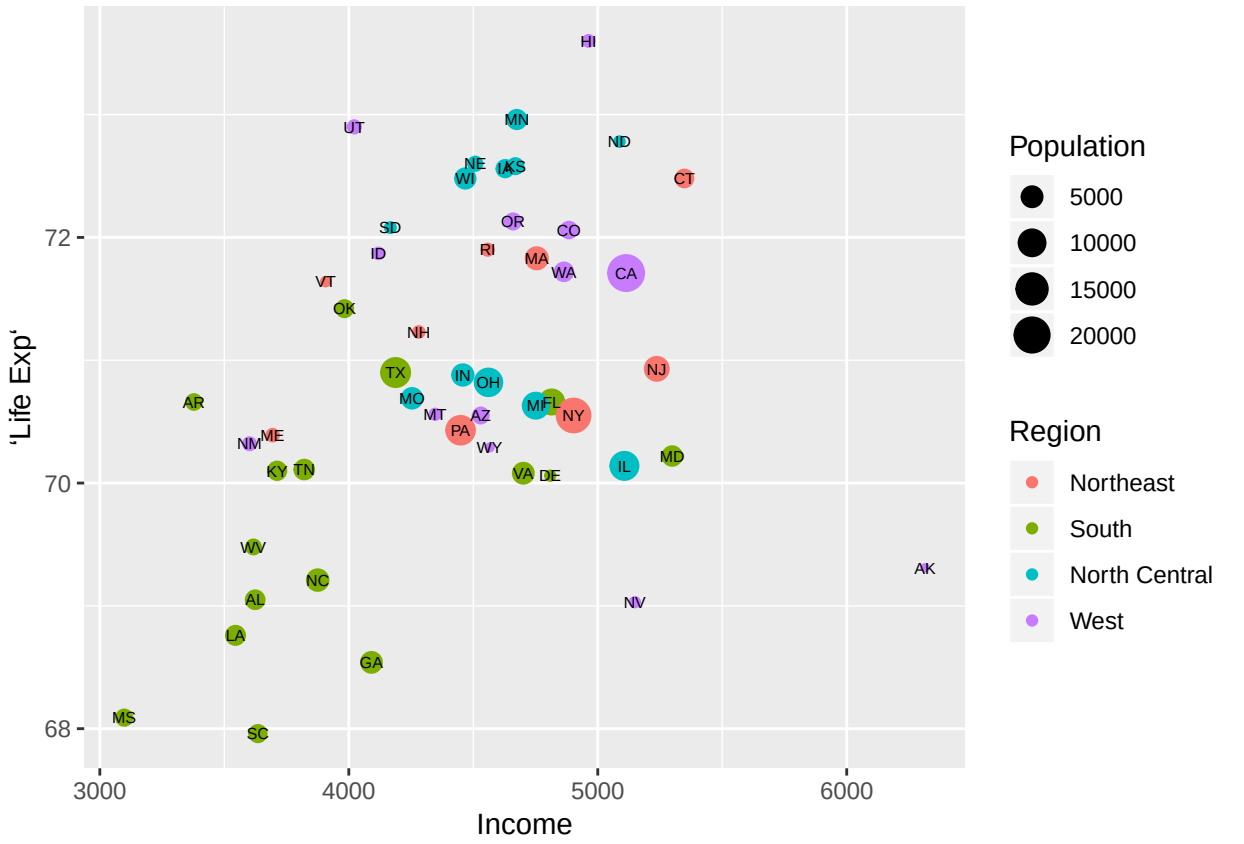
```
        col=Region, #Region mapped to color
        size=Population,
        label=Abbreviation))+ #Country mapped to label
    geom_point()+
    geom_text()
```




```
#This looks bad since the text labels do not need to inherit color
#or size.
```

7. If you are not satisfied with your anwser to question 6, redo the question. Hint: you may need to specify some aesthetics within a geom.

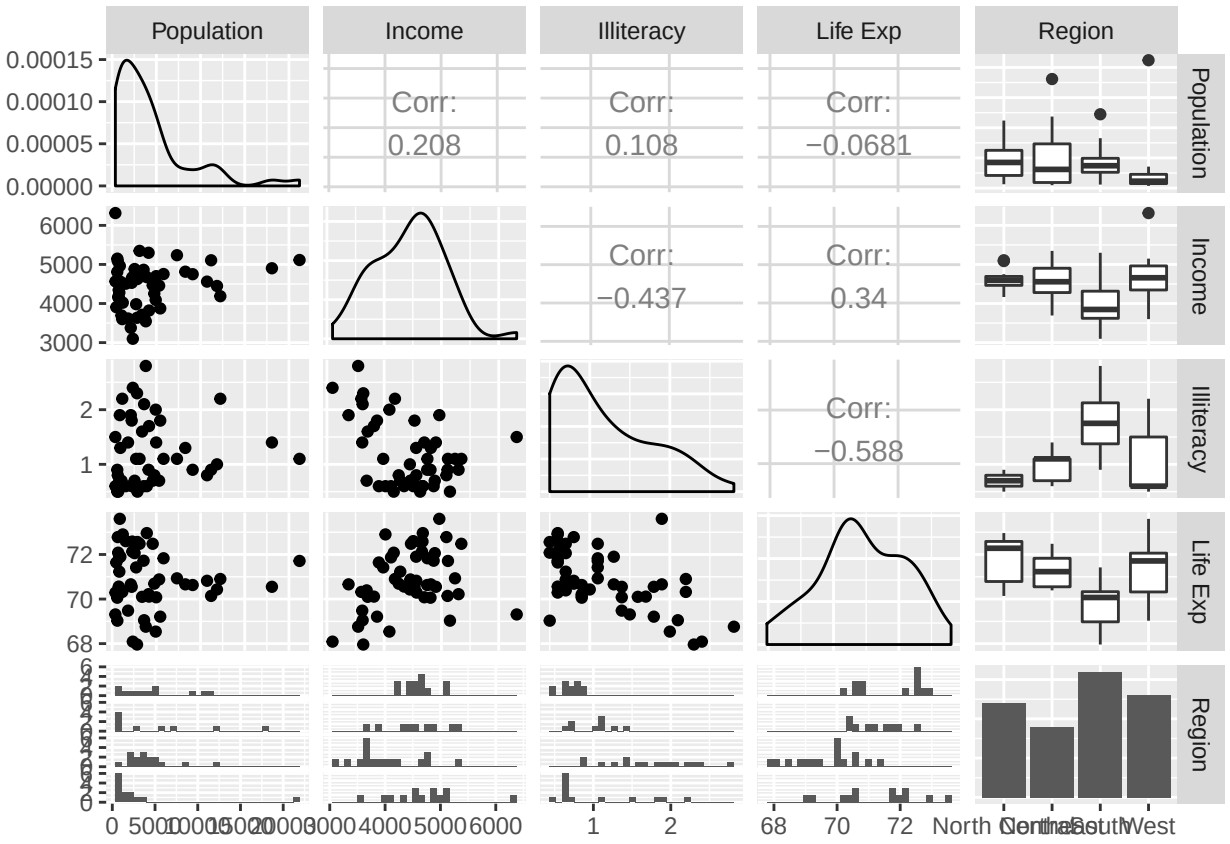
```
ggplot(USState,
       aes(x=Income,
           y=`Life Exp`))+ #Country mapped to label
       geom_point(aes(col=Region,
           size=Population))+ #color only in point
       geom_text(aes(label=Abbreviation),size=2)
```




```
#Note we can now clearly see more things.  For instance the
#outlying state for income is Alaska (AK).  It is not too suprising
#since Alaska produces oil and has a small population.  Similarly
#it is not too surprising that life expectancy is lower since
#Alaska is very cold.
```

8. The file *USStateRed.rds* contains the same data but with the variables `State`, `Abbreviation`, `Frost`, `Murder` and `HS Grad` stripped out. Load this data and plot a ggpairs plot.

```
library(GGally)
StateRed<-readRDS('USStateRed.rds')
ggpairs(StateRed)
```



9. Discuss some interesting features of the plot.

Just a few things are the following. The marginal distributions (densities) show that state population is right skewed, most states are small but there are a few big ones. Life expectancy is correlated highly with other socioeconomic variables such as income and illiteracy but not as much with either State size or area. The correlation between state area and state population is surprisingly close to zero. However the scatterplot suggests that an outlier (Alaska again) may be having an impact on this result. The Western states tend to have skewed distributions for illiteracy and life expectancy, most Western states do well (low illiteracy, high life expectancy) but there are a few western states with extreme values in the opposite direction.