Data Handling: Import, Cleaning and Visualisation

Lecture 11: Visualization and Dynamic Documents

Prof. Dr. Ulrich Matter

13/12/2018

1 Data display

- overview of last step in Data Science process
- low level: display data in R Murrell (2009) 9.10, only key aspects (use the practical aspects of this to start the workshop)
- visualization: plotting with gg (again, maybe part of the code examples in exercises)
- dynamic documents (partly last part of Murrell (2009) 9.10, rest from webmining: tables etc.), basics of markdown (focus particularly on this in exercises)

2 Workshop: Visualization with R (`ggplot2`)

2.1 'Grammer of Graphics'

2.2 add more theory here

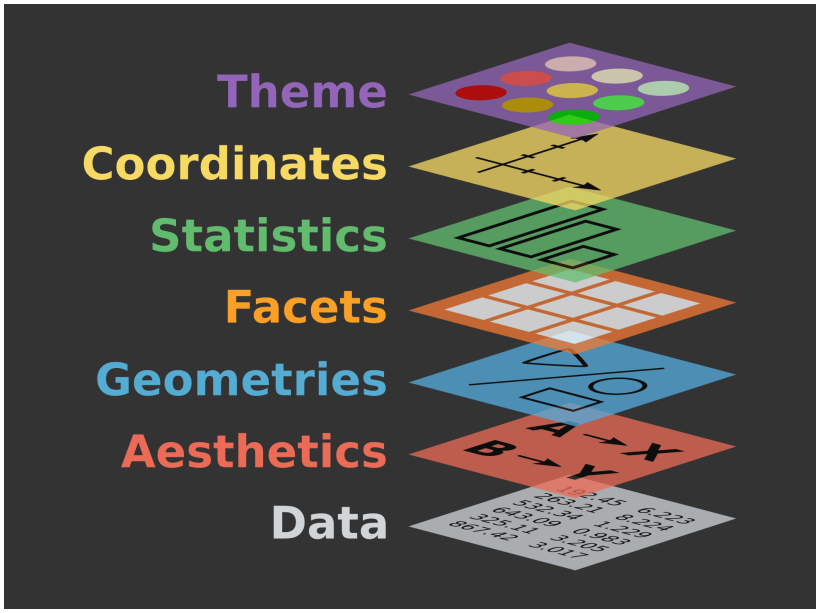


Figure 1: Source: http://bloggotype.blogspot.ch/2016/08/holiday-notes2-grammar-of-graphics.html

2.3 ggplot2

2.4 ggplot2 basics

- Data must be stored in a `data.frame`
- Basic function/starting point of a plot: `ggplot`
- First line of plot code declares the data and the 'aesthetics' (what variables are mapped to the x-/y-axes):

```
ggplot(data = my_dataframe, aes(x= xvar, y= yvar))
```

2.5 Example data set: `swiss`

```
# load the R package
library(ggplot2)
# load the data
data(swiss)
# get details about the data set
# ?swiss
# inspect the data
head(swiss)
```

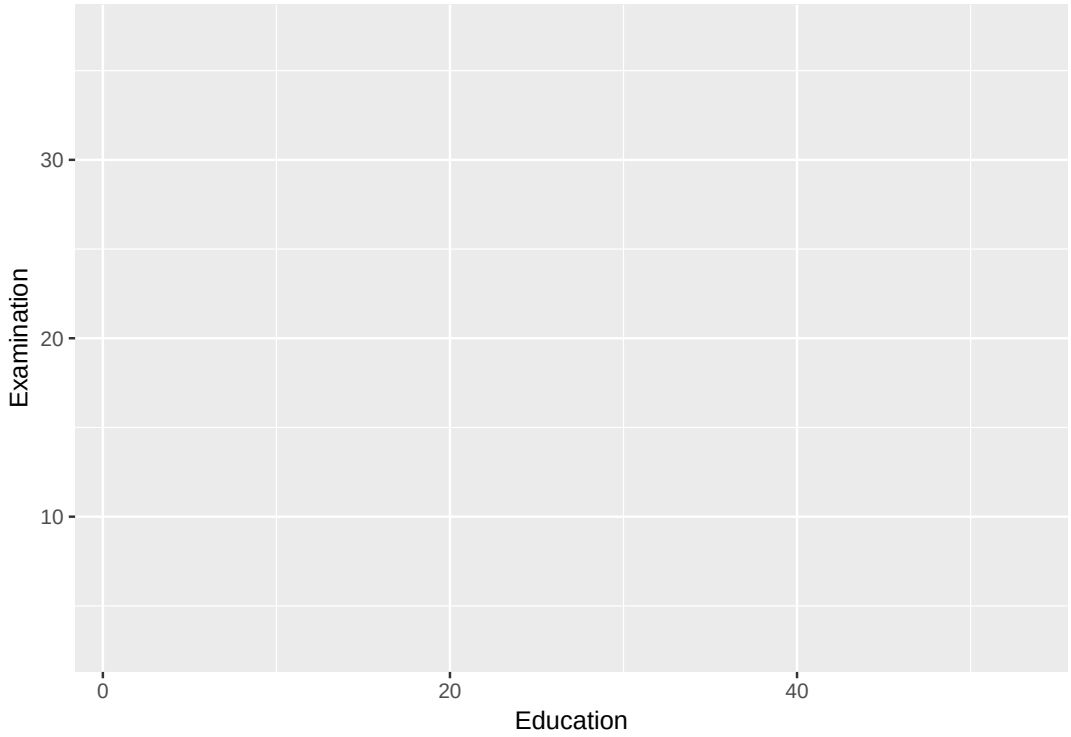
```
##              Fertility Agriculture Examination Education Catholic Infant.Mortality
## Courtelary        80.2        17.0          15        12     9.96             22.2
## Delemont          83.1        45.1           6         9    84.84             22.2
## Franches-Mnt      92.5        39.7           5         5    93.40             20.2
## Moutier           85.8        36.5          12         7    33.77             20.3
## Neuveville        76.9        43.5          17        15     5.16             20.6
## Porrentruy        76.1        35.3           9         7    90.57             26.6
```

2.6 Add indicator variable

```
# code province as 'Catholic' if more than 50% are catholic
swiss$Religion <- 'Protestant'
swiss$Religion[50 < swiss$Catholic] <- 'Catholic'
swiss$Religion <- as.factor(swiss$Religion)
```

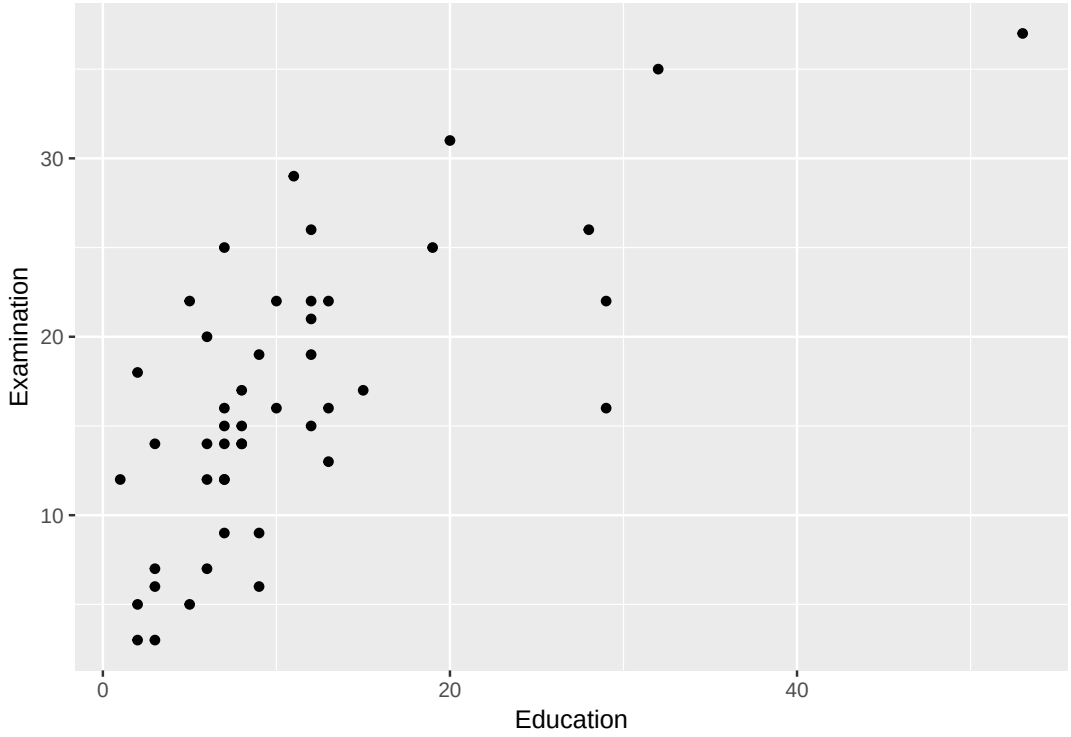
2.7 Data and aesthetics

```
ggplot(data = swiss, aes(x = Education, y = Examination))
```



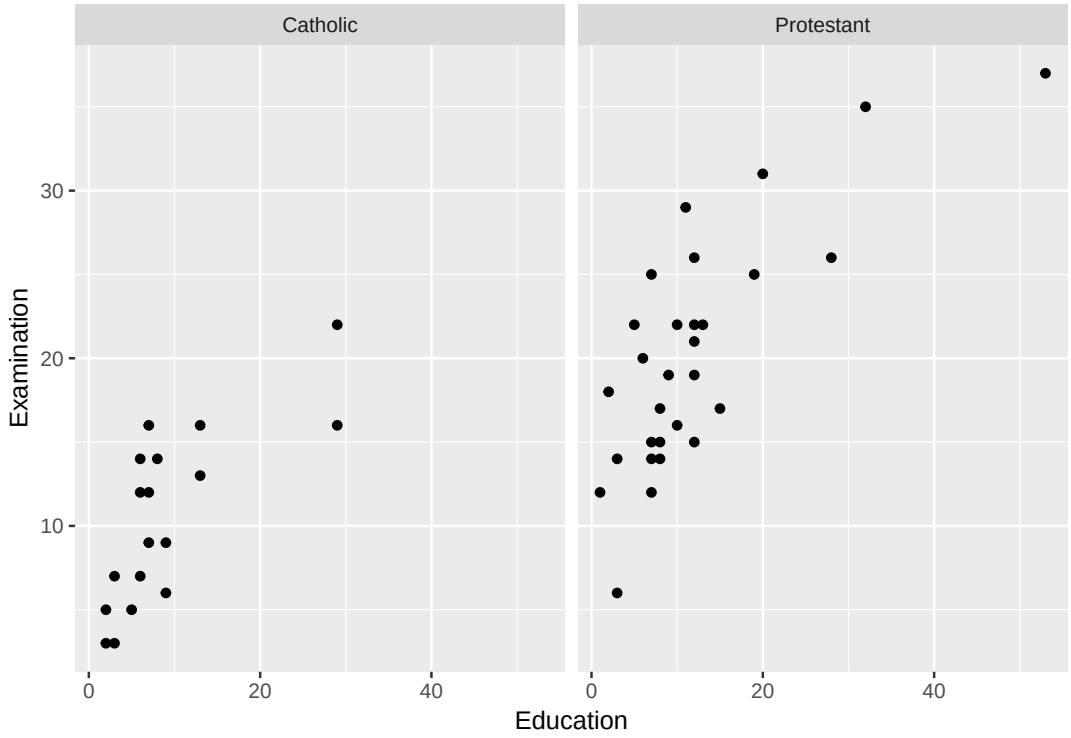
2.8 Geometries (~the type of plot)

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
    geom_point()
```



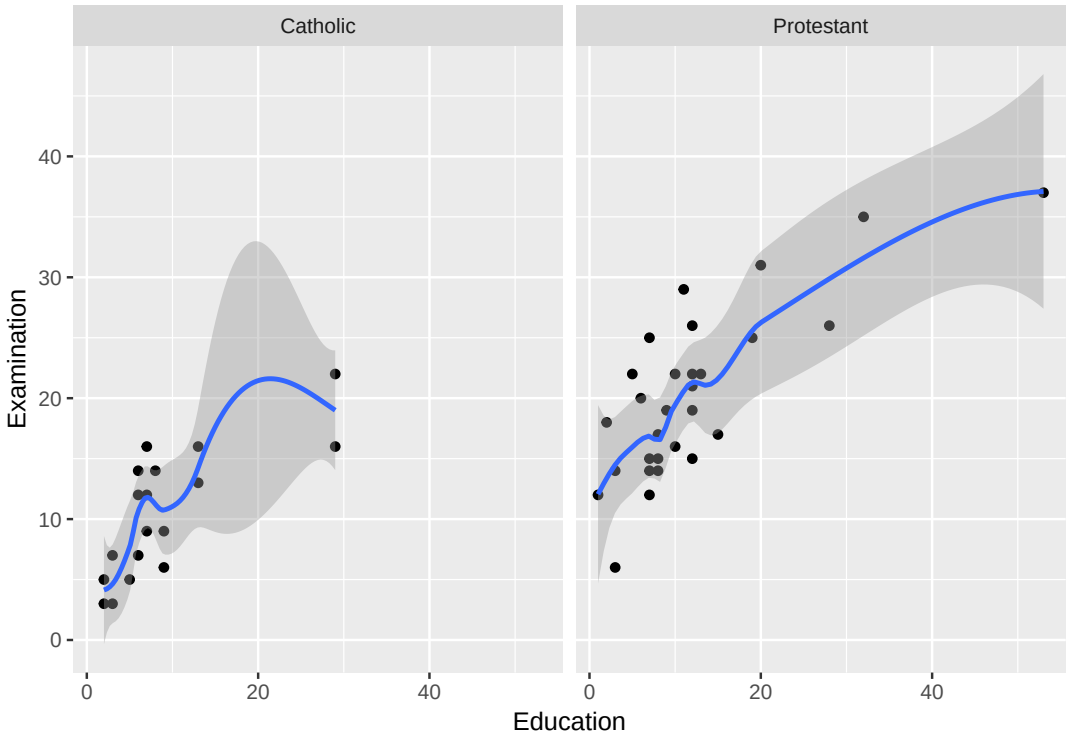
2.9 Facets

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
    geom_point() +
    facet_wrap(~Religion)
```



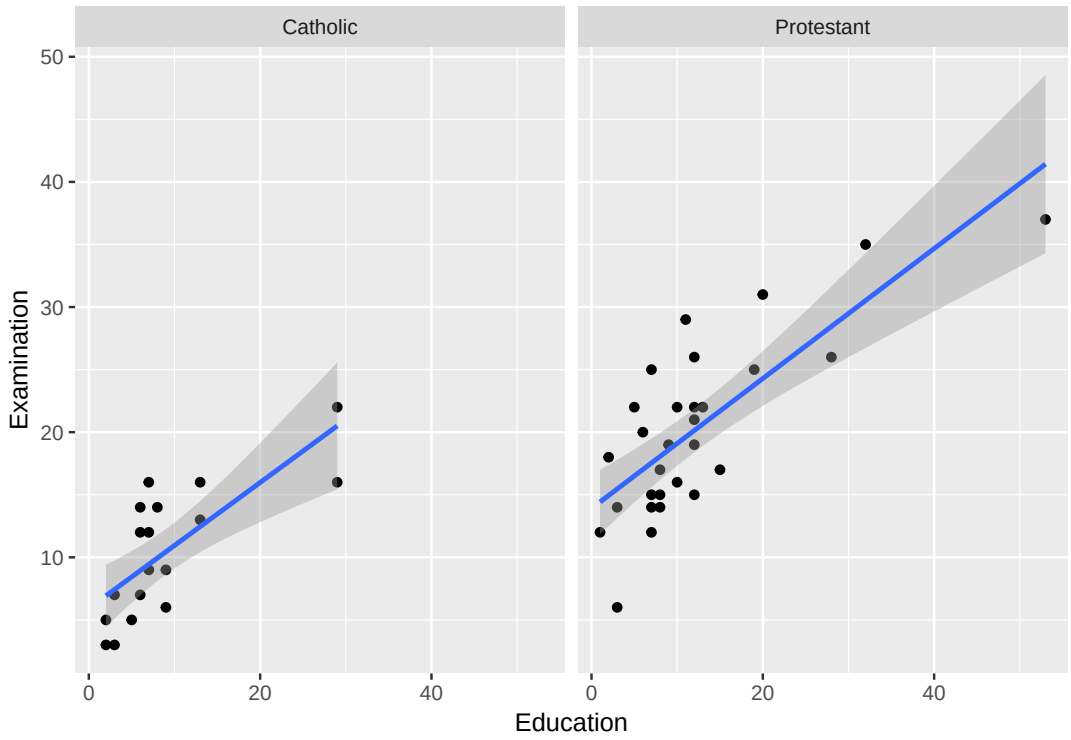
2.10 Additional layers and statistics

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
     geom_point() +
     geom_smooth(method = 'loess') +
     facet_wrap(~Religion)
```



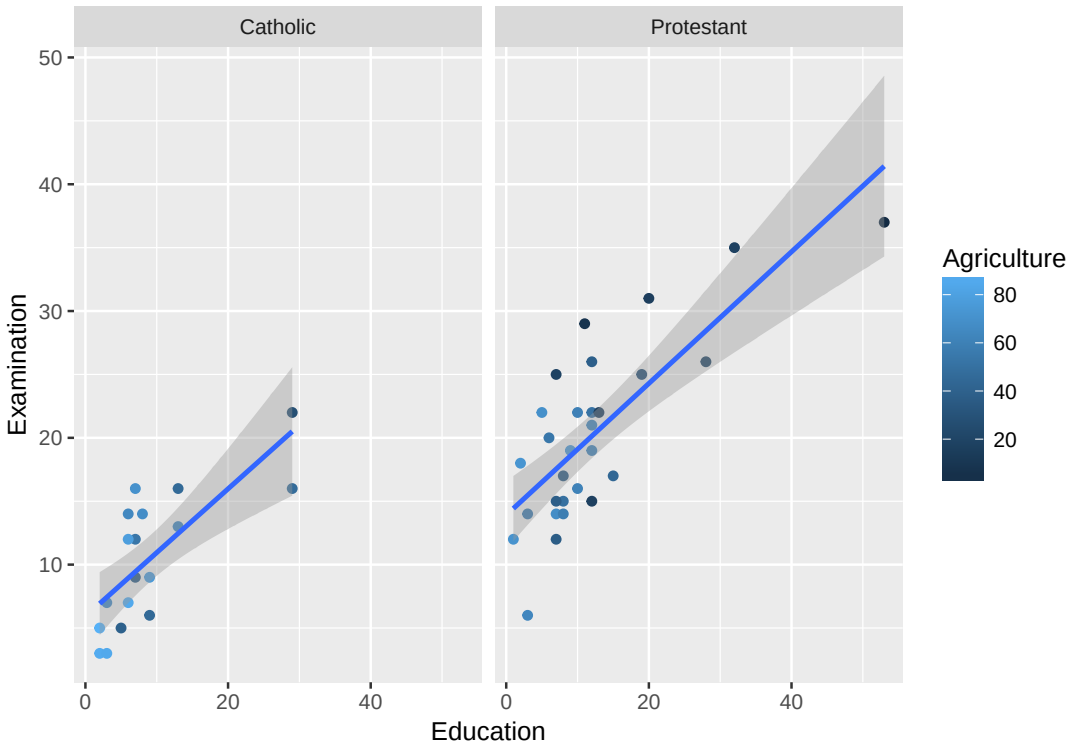
2.11 Additional layers and statistics

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
     geom_point() +
     geom_smooth(method = 'lm') +
     facet_wrap(~Religion)
```



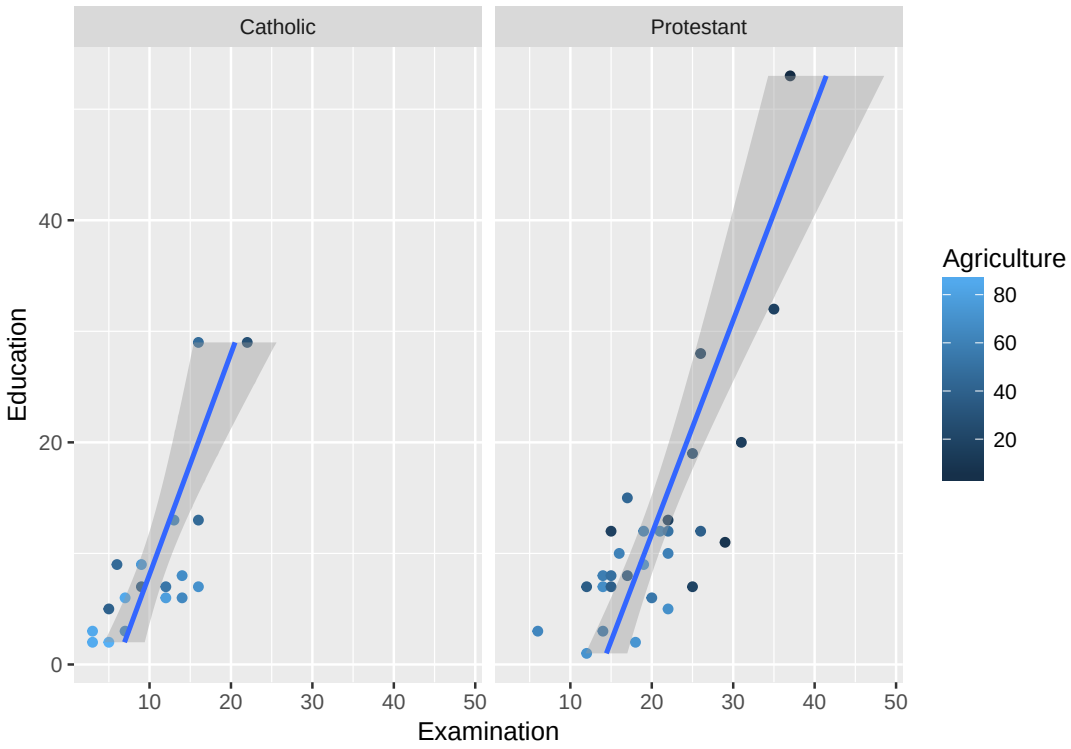
2.12 Additional aesthetics

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
     geom_point(aes(color = Agriculture)) +
     geom_smooth(method = 'lm') +
     facet_wrap(~Religion)
```



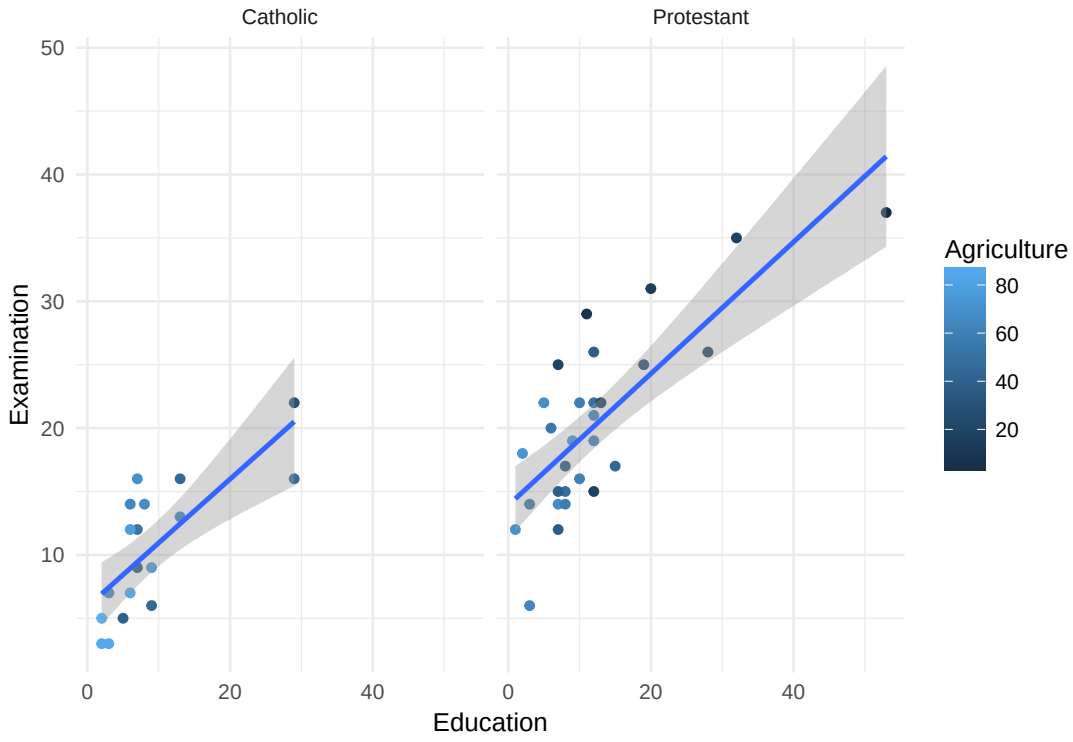
2.13 Change coordinates

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
     geom_point(aes(color = Agriculture)) +
     geom_smooth(method = 'lm') +
     facet_wrap(~Religion) +
     coord_flip()
```



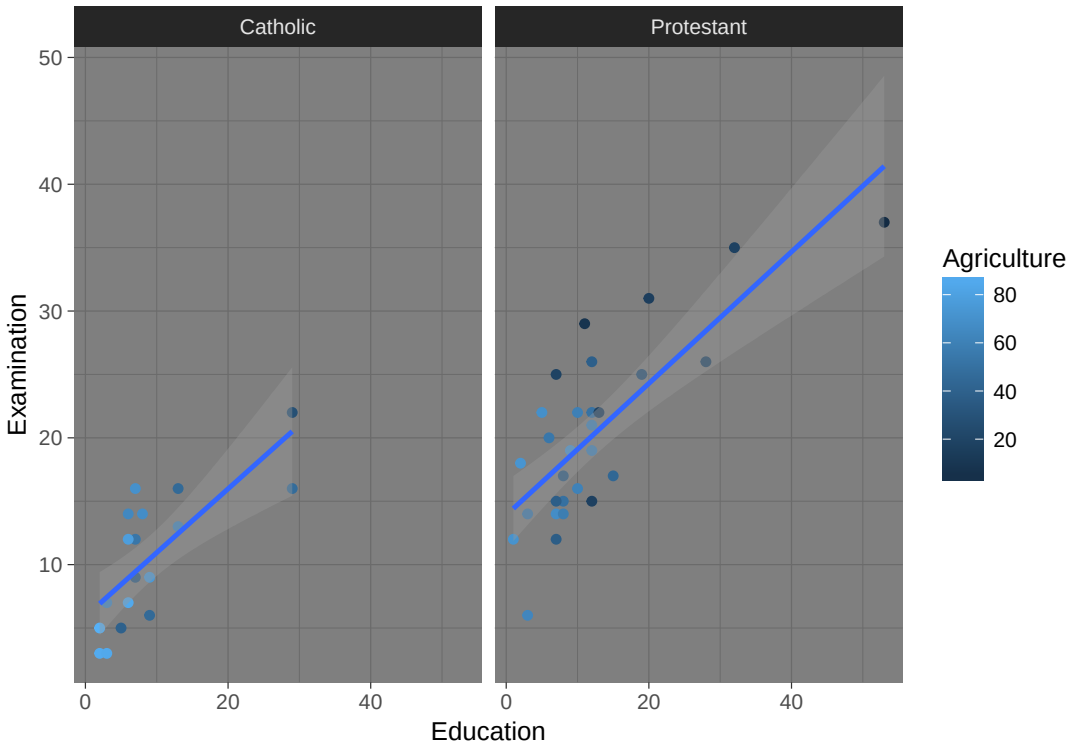
2.14 Themes

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
     geom_point(aes(color = Agriculture)) +
     geom_smooth(method = 'lm') +
     facet_wrap(~Religion) +
     theme_minimal()
```



2.15 Themes

```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
     geom_point(aes(color = Agriculture)) +
     geom_smooth(method = 'lm') +
     facet_wrap(~Religion) +
     theme_dark()
```



3 Dynamic Documents: basic idea (focus on HTML because they already know it)

References

Murrell, Paul. 2009. *Introduction to Data Technologies*. London, UK: CRC Press.