exercise_1_code+output.R

Emmanuel

Wed Mar 27 12:50:12 2019

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse
1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tibble)
library(ggplot2)
library(dplyr)
```

```
cwur <- read.csv("cwurData.csv")
cwurdf <- data.frame(cwur)
```

```
names(cwurdf)
```

```
##  [1] "world_rank"           "institution"          "country"
##  [4] "national_rank"        "quality_of_education" "alumni_employment"
##  [7] "quality_of_faculty"   "publications"         "influence"
## [10] "citations"            "broad_impact"         "patents"
## [13] "score"                "year"
```

```
times <- read.csv("timesData.csv")
timesdf <- data.frame(times)
```

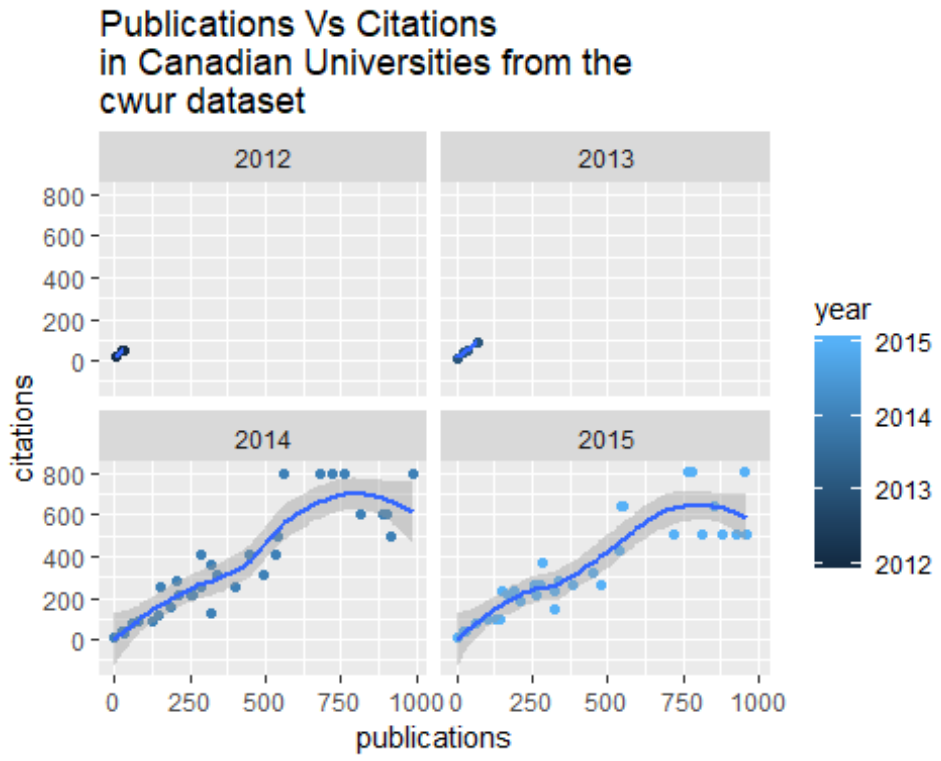
```
names(timesdf)
```

```
##  [1] "world_rank"             "university_name"
##  [3] "country"                "teaching"
##  [5] "international"          "research"
##  [7] "citations"              "income"
##  [9] "total_score"            "num_students"
## [11] "student_staff_ratio"    "international_students"
## [13] "female_male_ratio"      "year"
```

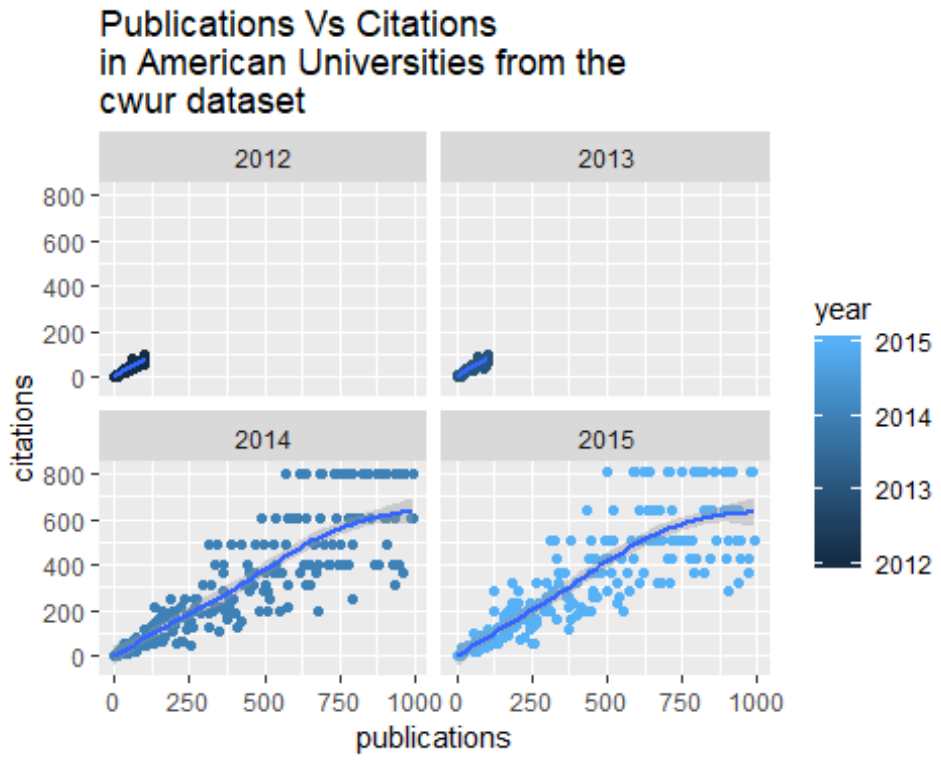
```r
#comparing the research(publications) and citations from both the center for
#world ranking dataset and the times dataset between canadian universities
#and American universities only.

#canadian universities from the cwur dataset
cwurfilter <- cwurdf %>% filter(country == "Canada")
ggplot(cwurfilter, aes(x=publications, y=citations)) + geom_point(aes(color =
year)) +
  geom_smooth(method = "loess", formula = y~x) + facet_wrap(year~.) +
labs(title = "Publications Vs Citations
in Canadian Universities from the
cwur dataset")
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 6.87

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 21.13

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 26.317

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too small.
## fewer data values than degrees of freedom.

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used
## at 6.87

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood
radius
## 21.13

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal
## condition number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
```

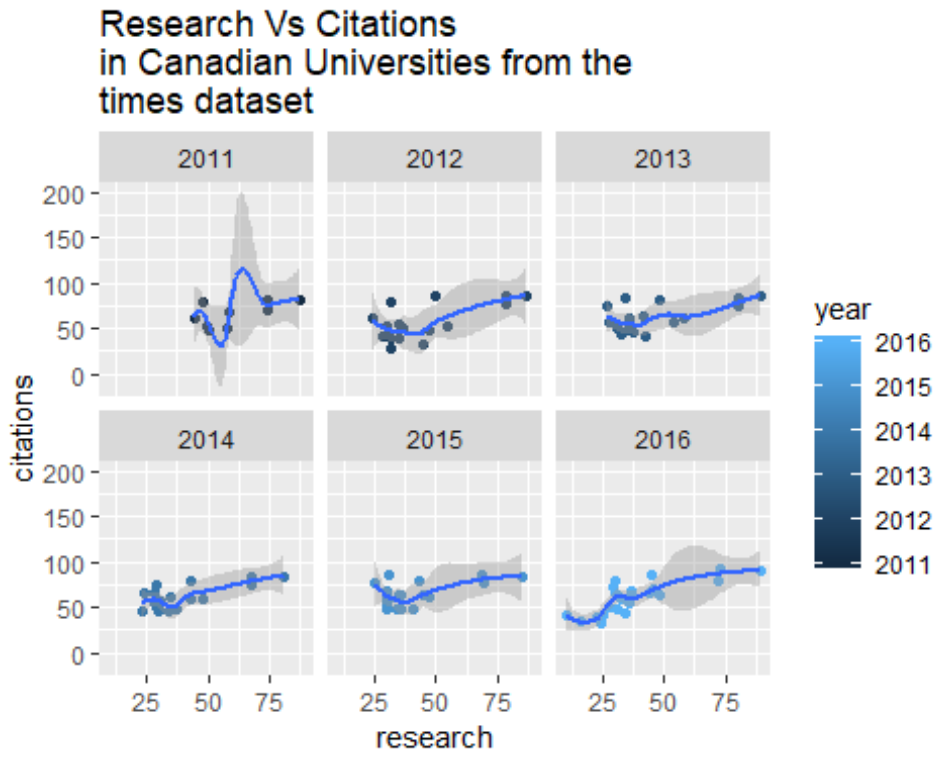
```
## as.matrix(model.frame(delete.response(terms(object)), : There are other
## near singularities as well. 26.317

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 1.67

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 32.33

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1877.5

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too small.
## fewer data values than degrees of freedom.

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used
## at 1.67

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood
radius
## 32.33

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal
## condition number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other
## near singularities as well. 1877.5
```



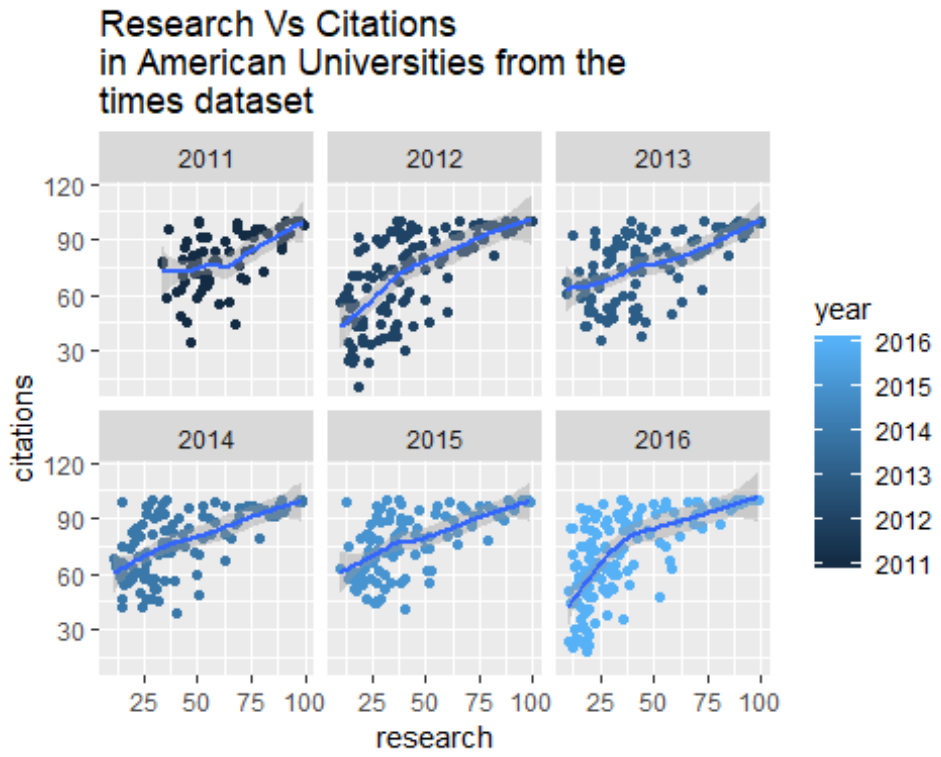

```
#american universities from the cwur dataset
amerfilter <- cwurdf %>% filter(country =="USA")
ggplot(amerfilter, aes(x=publications, y=citations)) + geom_point(aes(color =
year)) +
  geom_smooth(method = "loess", formula = y~x) + facet_wrap(year~.) +
labs(title = "Publications Vs Citations
in American Universities from the
cwur dataset")
```




```r
#canadian universities from the times dataset
timesfilter <- timesdf %>% filter(country == "Canada")
ggplot(timesfilter, aes(x=research, y=citations)) + geom_point(aes(color =
year)) +
  geom_smooth(method = "loess", formula = y~x) + facet_wrap(year~.) +
labs(title = "Research Vs Citations
in Canadian Universities from the
times dataset")
```




```r
#american universities from the times dataset
amertimes <- timesdf %>% filter(country == "United States of America")
ggplot(amertimes, aes(x=research, y=citations)) + geom_point(aes(color =
year)) +
  geom_smooth(method = "loess", formula = y~x) + facet_wrap(year~.) +
labs(title = "Research Vs Citations
in American Universities from the
times dataset")
```


Research Vs Citations
in American Universities from the
times dataset
2011
2012
2013
2014
2015
2016
citations
120
90
60
30
25
50
75
100
research
year
2016
2015
2014
2013
2012
2011

exercise_2_code+output.R

Emmanuel

Wed Mar 27 14:29:46 2019

```
library(tidyverse)

## -- Attaching packages ---------------------------------------- tidyverse
1.2.1 --

## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tibble)
library(ggplot2)
library(dplyr)

times <- read.csv("timesData.csv")
timesdf <- data.frame(times)

names(timesdf)

##  [1] "world_rank"             "university_name"
##  [3] "country"                "teaching"
##  [5] "international"          "research"
##  [7] "citations"              "income"
##  [9] "total_score"            "num_students"
## [11] "student_staff_ratio"    "international_students"
## [13] "female_male_ratio"      "year"

shangai <- read.csv("shanghaiData.csv")
shangaidf <- data.frame(shangai)

names(shangaidf)

##  [1] "world_rank"      "university_name" "national_rank"
##  [4] "total_score"     "alumni"          "award"
##  [7] "hici"            "ns"              "pub"
## [10] "pcp"             "year"
```

```r
cwur <- read.csv("cwurData.csv")
cwurdf <- data.frame(cwur)

names(cwurdf)
```

```
##  [1] "world_rank"           "institution"          "country"
##  [4] "national_rank"        "quality_of_education" "alumni_employment"
##  [7] "quality_of_faculty"   "publications"         "influence"
## [10] "citations"            "broad_impact"         "patents"
## [13] "score"                "year"
```

```r
#comparing the mean of the total scores from three datasets(cwur, shangai and
times)
#used to determine the world ranking of universities in the year 2014.


timesfilter <- timesdf %>% filter(year == "2014")
timesnumeric <- as.numeric(timesfilter$total_score)
timesomit <- na.omit(timesnumeric)
timesmean <- mean(timesomit)
timesmean
```

```
## [1] 71.3275
```

```r
cwurfilter <- cwurdf %>% filter(year == "2014")
cwurnumeric <- as.numeric(cwurfilter$score)
cwuromit <- na.omit(cwurnumeric)
cwurmean <- mean(cwuromit)
cwurmean
```

```
## [1] 47.27141
```

```r
shangaifilter <- shangaidf %>% filter(year == "2014")
shangainumeric <- as.numeric(shangaifilter$total_score)
shangaiomit <- na.omit(shangainumeric)
shangaimean <- mean(shangaiomit)
shangaimean
```

```
## [1] 36.172
```

exercise_3_code+output.R

Emmanuel

Wed Mar 27 15:24:34 2019

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse
1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tibble)
library(ggplot2)
library(dplyr)
```

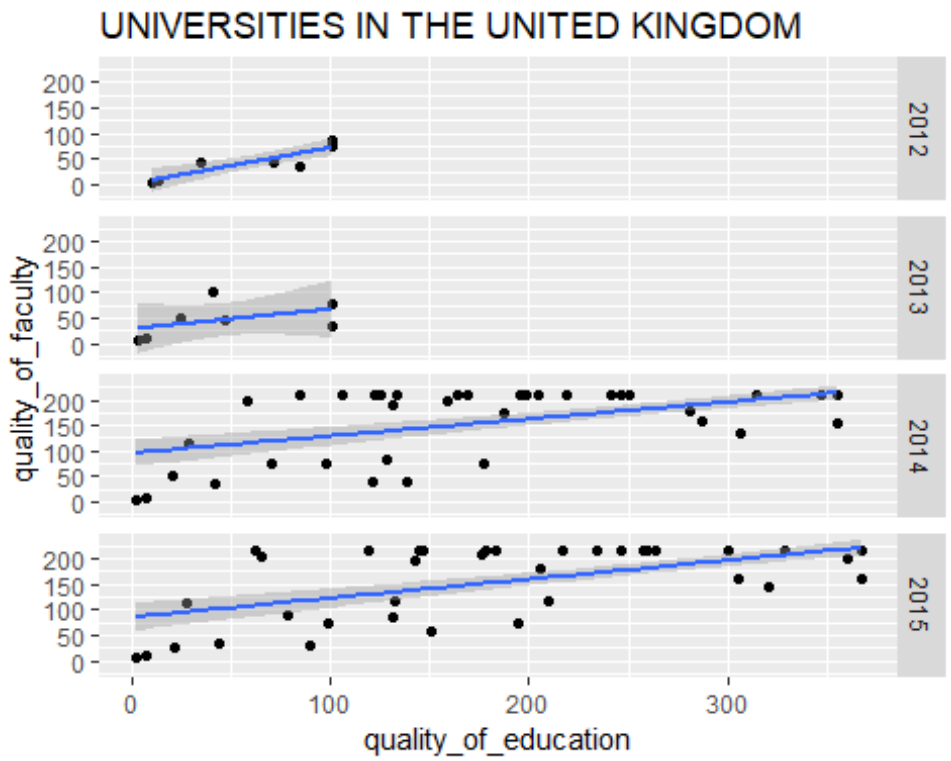
```
cwur <- read.csv("cwurData.csv")
cwurdf <- data.frame(cwur)
```

```
names(cwurdf)
```

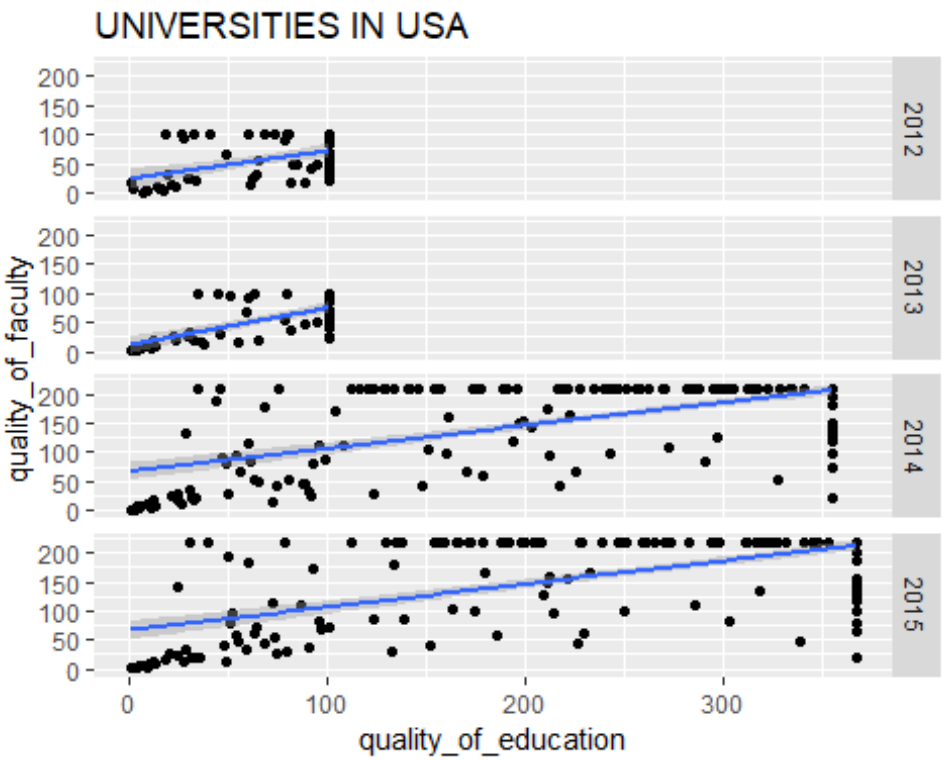
```
##  [1] "world_rank"           "institution"          "country"
##  [4] "national_rank"        "quality_of_education" "alumni_employment"
##  [7] "quality_of_faculty"   "publications"         "influence"
## [10] "citations"            "broad_impact"         "patents"
## [13] "score"                "year"
```

```
#To display the quality of education and quality of faculty using the cwur
dataset
#Using universities in the United Kingdom, USA, France,Switzerland and
Canada.
#between the year 2012-2015
```

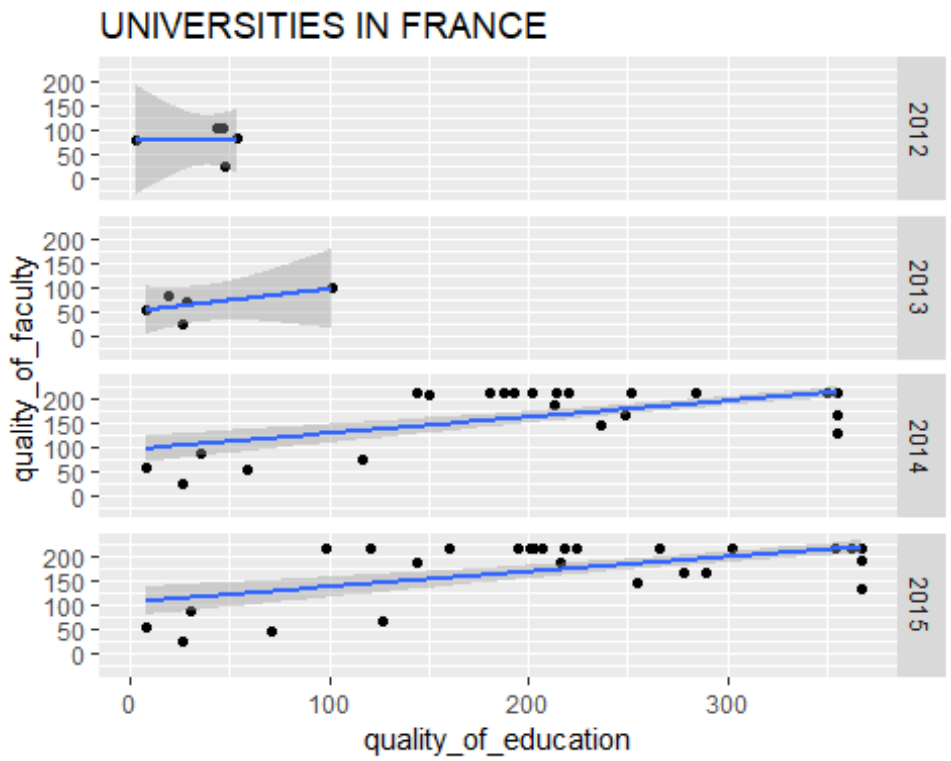
```
unitedkingdom <- cwurdf %>% filter(country == "United Kingdom")
ggplot(unitedkingdom, aes(x=quality_of_education, y=quality_of_faculty)) +
  geom_point() + geom_smooth(method="lm") + facet_grid(year~.) + labs(title =
"UNIVERSITIES IN THE UNITED KINGDOM")
```



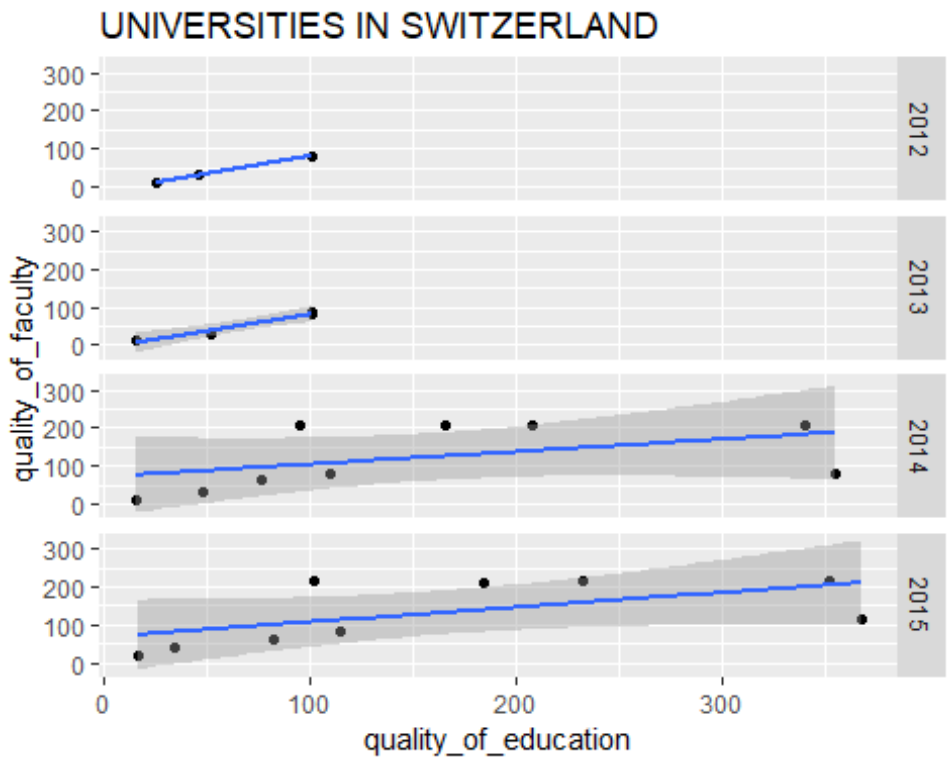

```
usa <- cwurdf %>% filter(country == "USA")
ggplot(usa, aes(x=quality_of_education, y=quality_of_faculty)) +
  geom_point() + geom_smooth(method="lm") + facet_grid(year~.) + labs(title =
"UNIVERSITIES IN USA")
```



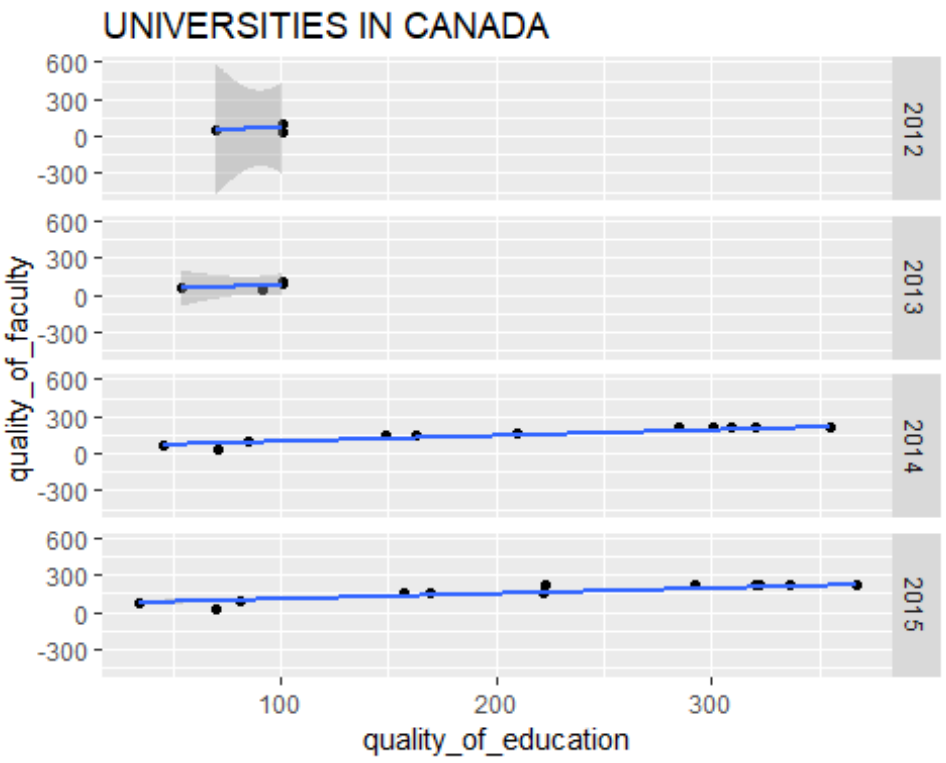

```
france <- cwurdf %>% filter(country == "France")
ggplot(france, aes(x=quality_of_education, y=quality_of_faculty)) +
  geom_point() + geom_smooth(method="lm") + facet_grid(year~.) + labs(title =
"UNIVERSITIES IN FRANCE")
```




```
switzerland <- cwurdf %>% filter(country == "Switzerland")
ggplot(switzerland, aes(x=quality_of_education, y=quality_of_faculty)) +
  geom_point() + geom_smooth(method="lm") + facet_grid(year~.) + labs(title =
"UNIVERSITIES IN SWITZERLAND")
```




```
switzerland <- cwurdf %>% filter(country == "Canada")
ggplot(switzerland, aes(x=quality_of_education, y=quality_of_faculty)) +
  geom_point() + geom_smooth(method="lm") + facet_grid(year~.) + labs(title =
"UNIVERSITIES IN CANADA")
```

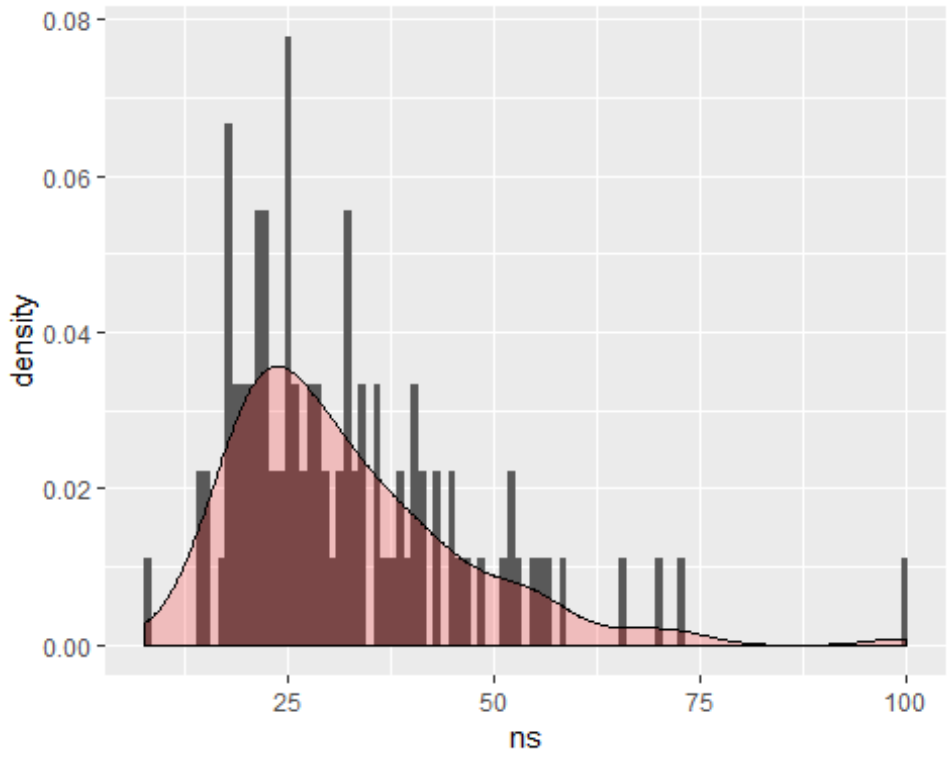

UNIVERSITIES IN CANADA
quality_of_faculty
quality_of_education
600
300
0
-300
100
200
300
2012
2013
2014
2015

exercise_4_code+output.R

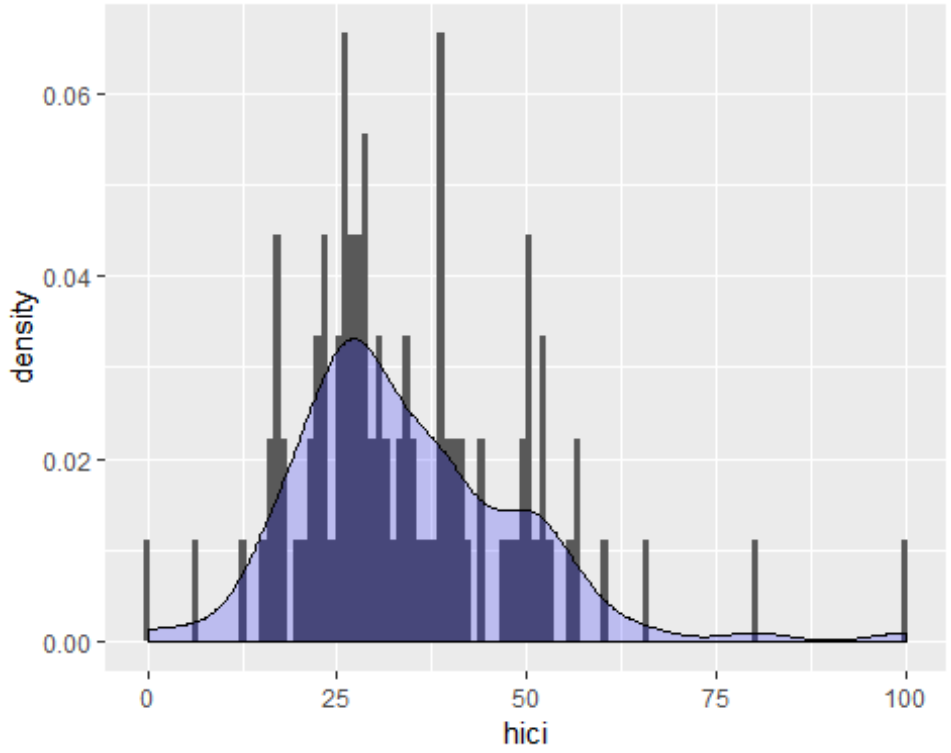
Emmanuel

Wed Mar 27 19:15:46 2019

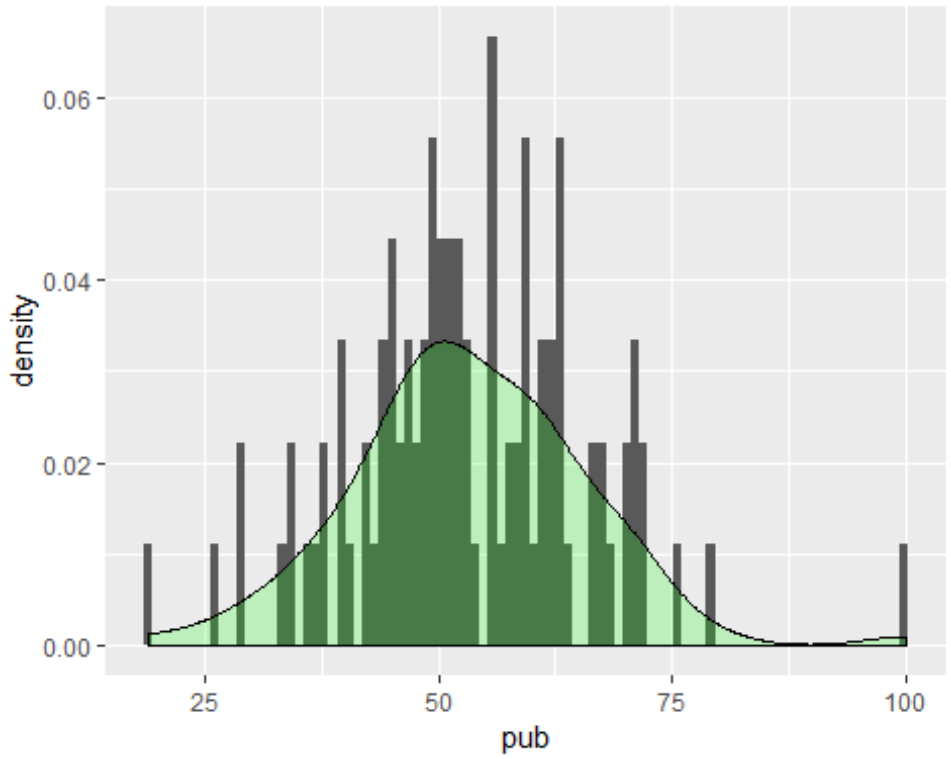
```
library(tidyverse)

## -- Attaching packages ---------------------------------------- tidyverse
1.2.1 --

## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tibble)
library(ggplot2)
library(dplyr)

shangai <- read.csv("shanghaiData.csv")
shangaidf <- data.frame(shangai)

names(shangaidf)

##  [1] "world_rank"      "university_name" "national_rank"
##  [4] "total_score"     "alumni"          "award"
##  [7] "hici"            "ns"              "pub"
## [10] "pcp"             "year"

#To show the N & S scores based on the number of papers published in Nature
#and science...
#using the top 100 univerities in the year 2015

shangai2015 <- shangaidf %>% filter(year == "2015")
shangaihead <- head(shangai2015,100)

ggplot(shangaihead, aes(x=ns)) + geom_histogram(aes(y=..density..), binwidth
= .9) +
  geom_density(alpha=.2, fill = "red")
```



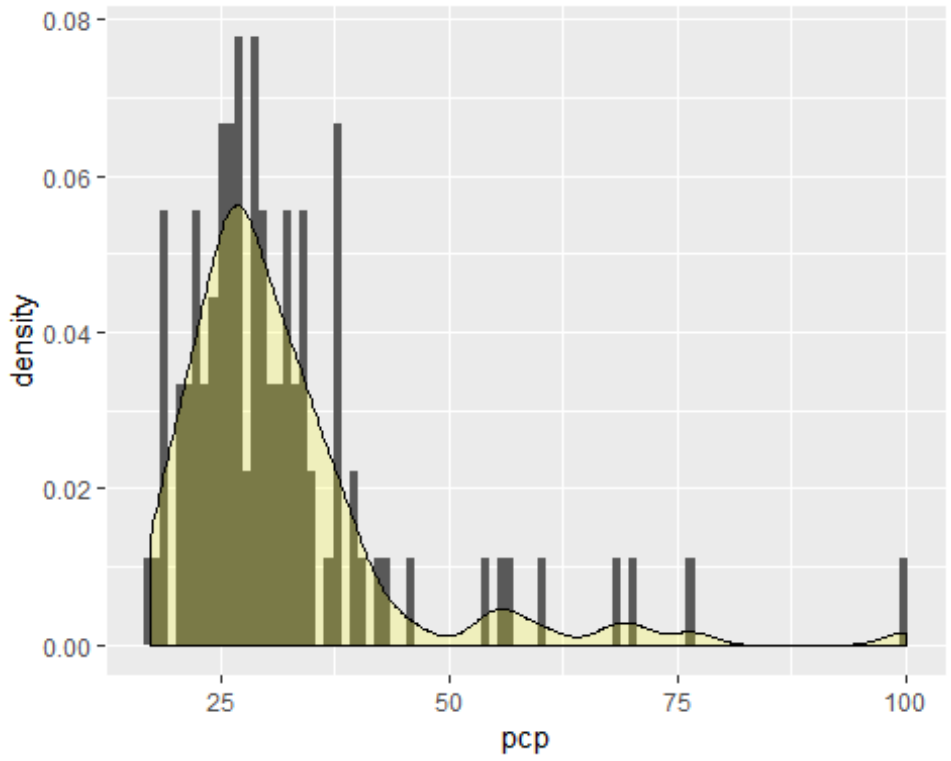

```
#To show the hici scores based on the number of Highly Cited Researchers
selected
#by Thomson Reuters
#using the top 100 univerities in the year 2015

ggplot(shangaihead, aes(x=hici)) + geom_histogram(aes(y=..density..),
binwidth = .9) +
  geom_density(alpha=.2, fill = "blue")
```




```
#To show the pub scores based on total number of papers indexed in the
Science
#Citation Index-Expanded and Social Science Citation Index
#using the top 100 univerities from the year 2015

ggplot(shangaihead, aes(x=pub)) + geom_histogram(aes(y=..density..), binwidth
= .9) +
  geom_density(alpha=.2, fill = "green")
```




```
#To show the pcp scores the weighted scores of the above five indicators
#divided by the number of full time academic staff
#using the top 100 univerities from the year 2015

ggplot(shangaihead, aes(x=pcp)) + geom_histogram(aes(y=..density..), binwidth
= .9) +
  geom_density(alpha=.2, fill = "yellow")
```

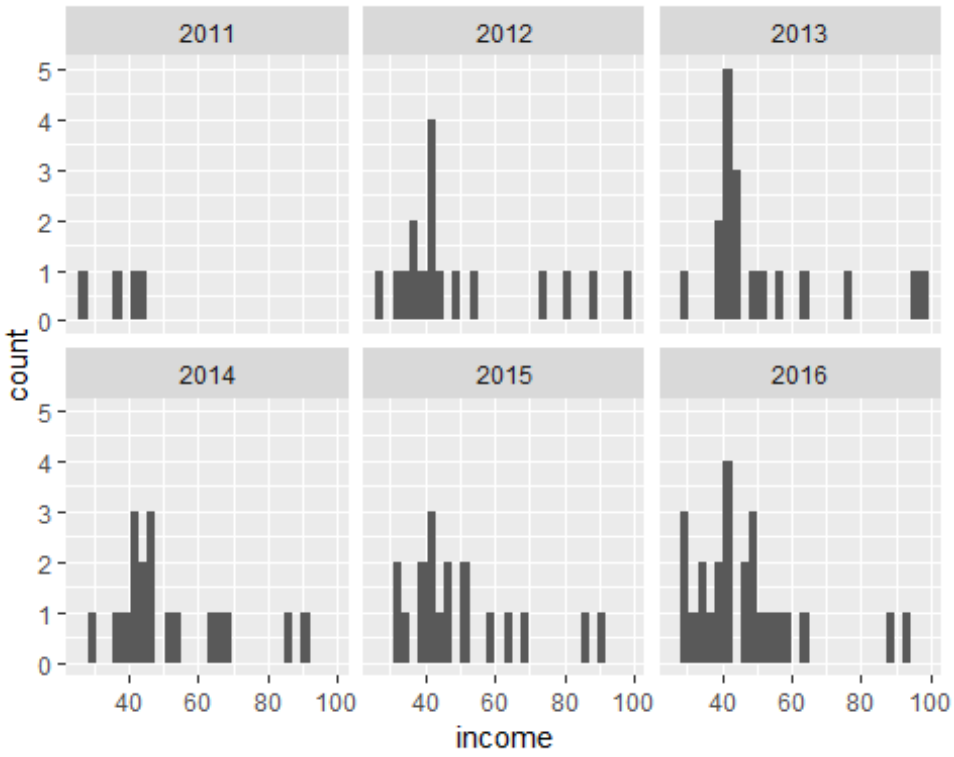

0.08
0.06
0.04
0.02
0.00
density
25
50
75
100
pcp

exercise_5_code+output.R

Emmanuel

Wed Mar 27 20:48:47 2019

```
library(tidyverse)

## -- Attaching packages ---------------------------------------- tidyverse
1.2.1 --

## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tibble)
library(ggplot2)
library(dplyr)

times <- read.csv("timesData.csv")
timesdf <- data.frame(times)


#To show the incomes in canadian universities between 2011-2016
b <- timesdf %>% mutate(income = as.character(income)) %>%
  mutate(income = as.numeric(income))%>% filter(country == "Canada")

## Warning in evalq(as.numeric(income), <environment>): NAs introduced by
## coercion

v <- na.omit(b)

ggplot(v, aes(x=income)) + geom_histogram() + facet_wrap(year~.)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#To show the incomes in german universities between 2011-2016
b1 <- timesdf %>% mutate(income = as.character(income)) %>%
  mutate(income = as.numeric(income))%>% filter(country == "Germany")

## Warning in evalq(as.numeric(income), <environment>): NAs introduced by
## coercion

v1 <- na.omit(b1)

ggplot(v1, aes(x=income)) + geom_histogram() + facet_wrap(year~.)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


2011
2012
2013
2014
2015
2016
count
income
0
2
4
6
8
40
60
80
100

exercise_6_code+output.R

Emmanuel

Wed Mar 27 21:01:27 2019

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.2.1 --

## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tibble)
library(ggplot2)
educationcsv <- read.csv(file =
"education_expenditure_supplementary_data.csv")
timesdf <- read.csv(file = "timesData.csv")

sub <- function(data, vars){
  i = length(vars)
  x = 1
  df = NA
  while(i >= x){
    df = rbind(df, filter(data, country==vars[x]))
    x = x + 1
  }
  na.omit(df)
}

educ <- select(educationcsv, country, direct_expenditure_type, X2011)
educ[order(educ$X2011, decreasing = TRUE),] %>% na.omit() -> educ
filter(educ, direct_expenditure_type=="Total") -> educ
countries <- educ[1:20,]$country

times_data <- filter(timesdf, year==2015)
clist <- str_replace_all(countries, "United States", "United States of
America")
sub(times_data, clist) -> times_data
ggplot(data = na.omit(times_data), mapping = aes(country)) +
```

```
  geom_bar() +
  xlab(label = "Country") +
  ylab(label = "Total Institions") +
  theme(axis.text.x = element_text(angle=60, vjust=0.6))
```