Movement of University Students in Australia

Kar Ng

2022-07-19

# R PACKAGES

library(tidyverse)  
library(kableExtra)  
library(tidytext)  
library(ggh4x)  
library(DT)

# DATA PREPARATION

## Data Import

Raw data set:

data <- read.csv("data.csv")  
  
datatable(data, )

## PhantomJS not found. You can install it with webshot::install\_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.

## Data Cleaning

The data cleaning:

mydata <- data %>%   
 # pivoting to change data from wider to longer format:  
 pivot\_longer(c(3:7), names\_to = "method", values\_to = "values\_per") %>%   
 # create 3 new variables:  
 mutate(metrics = case\_when(method = str\_detect(method, "Attrition") ~ "Attrition",   
 method = str\_detect(method, "Retention") ~ "Retention",  
 TRUE ~ "Success"),  
 year = str\_extract(method, "[:digit:]+"),  
 method = str\_remove(method, "\_Attrition\_2019"),  
 method = str\_remove(method, "\_Retention\_2019"),  
 method = str\_remove(method, "\_2020")) %>%   
 # change values\_per into numeric class  
 mutate(values\_per = as.double(values\_per)) %>%   
 # Remove NA  
 na.omit() %>%   
 # create a new variables "state"  
 mutate(state = fct\_collapse(Uni,  
 "New South Wales" = c("Charles Sturt University",   
 "Macquarie University",   
 "Southern Cross University",   
 "The University of New England",   
 "The University of Newcastle",   
 "The University of Sydney",   
 "University of New South Wales(b)",   
 "University of Technology Sydney",  
 "University of Wollongong",  
 "Western Sydney University"),  
 "Victoria" = c("Deakin University",  
 "Federation University Australia(c)",  
 "La Trobe University",  
 "Monash University",  
 "RMIT University",  
 "Swinburne University of Technology",  
 "The University of Melbourne",  
 "University of Divinity",  
 "Victoria University"),  
 "Queensland" = c("Bond University",  
 "CQUniversity",  
 "Griffith University",  
 "James Cook University",  
 "Queensland University of Technology",  
 "The University of Queensland",  
 "University of Southern Queensland",  
 "University of the Sunshine Coast"),  
 "Western Australia" = c("Curtin University",  
 "Edith Cowan University",  
 "Murdoch University",  
 "The University of Notre Dame Australia",  
 "The University of Western Australia"),  
 "South Australia" = c("Flinders University",  
 "The University of Adelaide",  
 "Torrens University Australia(d)",  
 "University of South Australia"),  
 "Tasmania" = c("Australian Maritime College(e)",  
 "University of Tasmania(e)"),  
 "Northern Territory" = c("Batchelor Institute of Indigenous Tertiary Education(f)",  
 "Charles Darwin University(f)"),  
 "Australian Capital Territory" = c("The Australian National University",  
 "University of Canberra"),  
 "Multi-State" = c("Australian Catholic University"),  
 "Australia" = c("National Total")  
 )) %>%   
 # change all character variables into factor  
 mutate\_if(is.character, as.factor) %>%   
 dplyr::select("year", "state", "Uni", "student", "metrics", "method", "values\_per")

**Change-log**:

* I changed the format of the table to a longer format
* I Created several new variables, which are “metrics”, “year”, and “method”
* I changed the data type of “values\_per” from character to numeric
* I created a new variable “state” to group universities
* I changed all the character variables into factor
* I shifted around (relocate) the order of variables in the data set

The cleaned data set

datatable(mydata)

## Data Exploration

Summary of the data.

summary(mydata)

## year state   
## 2019:252 New South Wales :90   
## 2020:126 Victoria :81   
## Queensland :72   
## Western Australia :45   
## South Australia :36   
## Australian Capital Territory:18   
## (Other) :36   
## Uni student metrics   
## Australian Catholic University: 9 domestic :126 Attrition:126   
## Bond University : 9 international:126 Retention:126   
## Charles Darwin University(f) : 9 overall :126 Success :126   
## Charles Sturt University : 9   
## CQUniversity : 9   
## Curtin University : 9   
## (Other) :324   
## method values\_per   
## New\_Adjusted: 84 Min. : 3.43   
## New\_Normal :168 1st Qu.:20.58   
## Success :126 Median :81.38   
## Mean :61.94   
## 3rd Qu.:88.04   
## Max. :97.64   
##

There are 42 universities involved in the dataset, other than the “National Total”

table(mydata$Uni) %>%   
 as.data.frame() %>%  
 filter(Freq != 0) %>%  
 mutate("no" = c(1:42)) %>%   
 rename("Universities" = "Var1",  
 "Sample Size" = "Freq") %>%   
 relocate(no, .before = Universities) %>%   
 kbl(align = "c") %>%   
 kable\_styling(full\_width = F, c("bordered", "stripped", "hover"))

no

Universities

Sample Size

1

Australian Catholic University

9

2

Bond University

9

3

Charles Darwin University(f)

9

4

Charles Sturt University

9

5

CQUniversity

9

6

Curtin University

9

7

Deakin University

9

8

Edith Cowan University

9

9

Federation University Australia(c)

9

10

Flinders University

9

11

Griffith University

9

12

James Cook University

9

13

La Trobe University

9

14

Macquarie University

9

15

Monash University

9

16

Murdoch University

9

17

National Total

9

18

Queensland University of Technology

9

19

RMIT University

9

20

Southern Cross University

9

21

Swinburne University of Technology

9

22

The Australian National University

9

23

The University of Adelaide

9

24

The University of Melbourne

9

25

The University of New England

9

26

The University of Newcastle

9

27

The University of Notre Dame Australia

9

28

The University of Queensland

9

29

The University of Sydney

9

30

The University of Western Australia

9

31

Torrens University Australia(d)

9

32

University of Canberra

9

33

University of Divinity

9

34

University of New South Wales(b)

9

35

University of South Australia

9

36

University of Southern Queensland

9

37

University of Tasmania(e)

9

38

University of Technology Sydney

9

39

University of the Sunshine Coast

9

40

University of Wollongong

9

41

Victoria University

9

42

Western Sydney University

9

The data structure has desired data type allocated.

glimpse(mydata)

## Rows: 378  
## Columns: 7  
## $ year <fct> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019,…  
## $ state <fct> Australia, New South Wales, New South Wales, New South Wale…  
## $ Uni <fct> National Total, Charles Sturt University, Macquarie Univers…  
## $ student <fct> domestic, domestic, domestic, domestic, domestic, domestic,…  
## $ metrics <fct> Attrition, Attrition, Attrition, Attrition, Attrition, Attr…  
## $ method <fct> New\_Adjusted, New\_Adjusted, New\_Adjusted, New\_Adjusted, New…  
## $ values\_per <dbl> 13.23, 21.01, 8.77, 21.85, 24.90, 12.92, 4.87, 3.86, 4.99, …

There is no missing value in the dataset (NA).

colSums(is.na(mydata))

## year state Uni student metrics method values\_per   
## 0 0 0 0 0 0 0

# Visualisation

## Attrition rate

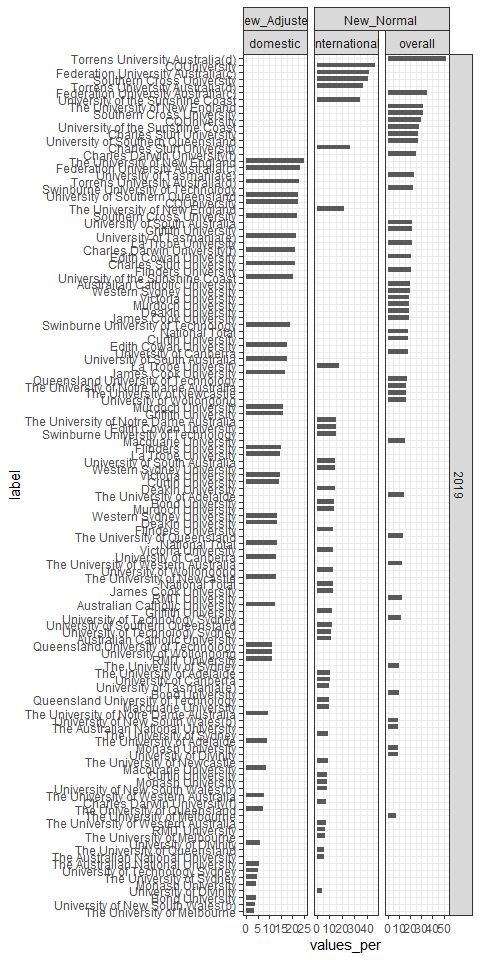
df1 <- mydata %>%   
 filter(metrics == "Attrition") %>%   
 dplyr::select(-metrics) %>%  
 mutate(label = reorder\_within(x = Uni, by = values\_per, within = student))

For New Adjusted Attrition rate,

df1

## # A tibble: 126 × 7  
## year state Uni student method value…¹ label  
## <fct> <fct> <fct> <fct> <fct> <dbl> <fct>  
## 1 2019 Australia National Total domest… New\_A… 13.2 Nati…  
## 2 2019 New South Wales Charles Sturt University domest… New\_A… 21.0 Char…  
## 3 2019 New South Wales Macquarie University domest… New\_A… 8.77 Macq…  
## 4 2019 New South Wales Southern Cross University domest… New\_A… 21.8 Sout…  
## 5 2019 New South Wales The University of New Eng… domest… New\_A… 24.9 The …  
## 6 2019 New South Wales The University of Newcast… domest… New\_A… 12.9 The …  
## 7 2019 New South Wales The University of Sydney domest… New\_A… 4.87 The …  
## 8 2019 New South Wales University of New South W… domest… New\_A… 3.86 Univ…  
## 9 2019 New South Wales University of Technology … domest… New\_A… 4.99 Univ…  
## 10 2019 New South Wales University of Wollongong domest… New\_A… 11.3 Univ…  
## # … with 116 more rows, and abbreviated variable name ¹​values\_per  
## # ℹ Use `print(n = ...)` to see more rows

ggplot(df1, aes(y = label, x = values\_per)) +  
 geom\_bar(stat = "identity", width = 0.7) +  
 facet\_nested(year ~ method+student, scales = "free") +  
 theme\_bw() +  
 scale\_y\_reordered()



## Retention rate

## Success rate