Trans. data analysis(R) Assignment 2

2331609 刘畅

2023-11-10

## 1 Using the iri dataset

### 1-1

Get a subset with STATE\_CODE 6 and SHRP\_ID with 050.

library("dplyr")

##   
## 载入程辑包：'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library("tidyr")  
iri <- read.csv("data/iri.csv")  
sub\_iri <- subset(iri, STATE\_CODE == 6 & substr(SHRP\_ID, 1, 3) == '050')  
  
head(sub\_iri, n = 10)

## STATE\_CODE SHRP\_ID CONSTRUCTION\_NO VISIT\_DATE IRI  
## 3380 6 0501 1 1/25/90, 12:00:00 AM 3.7206  
## 3381 6 0501 1 2/11/92, 12:00:00 AM 3.8896  
## 3382 6 0501 1 2/16/91, 12:00:00 AM 3.6434  
## 3383 6 0501 2 2/11/98, 12:00:00 AM 1.4376  
## 3384 6 0501 2 2/2/93, 12:00:00 AM 1.1230  
## 3385 6 0501 2 2/27/97, 12:00:00 AM 1.3510  
## 3386 6 0501 2 3/5/99, 12:00:00 AM 1.7332  
## 3387 6 0501 2 4/4/95, 12:00:00 AM 1.2378  
## 3388 6 0501 3 3/10/00, 12:00:00 AM 1.9154  
## 3389 6 0501 4 2/12/02, 12:00:00 AM 1.6550

### 1-2

Obtain the summary statistics of IRI of each section: min, max, and mean.

sta\_sec <- iri |>  
 group\_by(STATE\_CODE, SHRP\_ID) |>  
 summarize(min = min(IRI), max = max(IRI), mean = mean(IRI))  
  
head(sta\_sec, n = 10)

## # A tibble: 10 × 5  
## # Groups: STATE\_CODE [1]  
## STATE\_CODE SHRP\_ID min max mean  
## <int> <chr> <dbl> <dbl> <dbl>  
## 1 1 0101 0.657 0.810 0.716  
## 2 1 0102 0.897 3.10 1.34   
## 3 1 0103 0.760 0.834 0.803  
## 4 1 0104 0.594 0.684 0.644  
## 5 1 0105 0.614 0.694 0.648  
## 6 1 0106 0.582 0.764 0.688  
## 7 1 0107 0.628 0.963 0.747  
## 8 1 0108 0.731 0.875 0.766  
## 9 1 0109 0.679 0.775 0.736  
## 10 1 0110 0.672 0.764 0.705

### 1-3

Sort the summarized data by the averaged IRI in a descending order (report results for one section only).

sta\_sec\_desc <- sta\_sec |>  
 arrange(desc(mean))  
  
head(sta\_sec\_desc)

## # A tibble: 6 × 5  
## # Groups: STATE\_CODE [5]  
## STATE\_CODE SHRP\_ID min max mean  
## <int> <chr> <dbl> <dbl> <dbl>  
## 1 72 4121 4.06 5.05 4.44  
## 2 90 A350 3.06 5.15 3.96  
## 3 6 0662 3.06 4.29 3.94  
## 4 34 1030 2.77 5.74 3.91  
## 5 6 B441 3.06 5.87 3.82  
## 6 89 3002 3.48 4.27 3.81

### 1-4

Generate a scatter plot for the averaged IRI against the time for a selected section, and then give your interpretation of the plot.

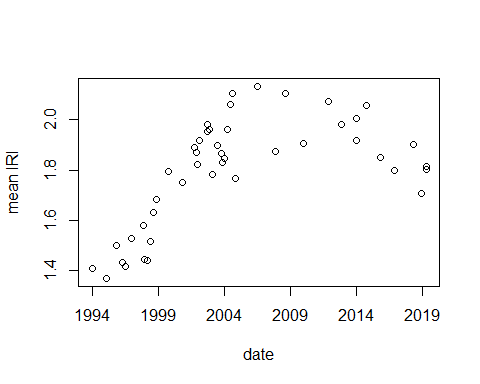
Firstly, select a section with a relatively large sample size.

iri\_count <- iri |>  
 group\_by(STATE\_CODE, SHRP\_ID) |>  
 summarise(num = n()) |>  
 arrange(desc(num))  
  
head(iri\_count, n = 5)

## # A tibble: 5 × 3  
## # Groups: STATE\_CODE [5]  
## STATE\_CODE SHRP\_ID num  
## <int> <chr> <int>  
## 1 4 0215 45  
## 2 50 1002 40  
## 3 89 3015 39  
## 4 42 1606 38  
## 5 30 0114 37

Then, perform statistics for section:(4, 0215).

iri\_4\_0215 <- subset(iri, STATE\_CODE == 4 & SHRP\_ID == '0215') |>  
 separate(col = VISIT\_DATE,   
 into = c("DATE", "TIME"),  
 sep = ",") |>  
 mutate(DATE = as.Date(DATE, "%m/%d/%y"))  
  
plot(iri\_4\_0215$DATE, iri\_4\_0215$IRI, xaxt = "n", xlab = "date", ylab = "mean IRI")  
axis.Date(1, at = seq(min(iri\_4\_0215$DATE), max(iri\_4\_0215$DATE), by = "5 years"), format = "%Y")



The average IRI increased gradually from 1994 to 2004. From 2004 to 2014, the average IRI remained at a stable high level. After 2014, the average IRI gradually decreased. This indicates that the section experienced a gradual deterioration of road roughness in the early stage, which has been improved during the past 10 years.

## 2 Using the CRSS datasets in 2017

### 2-1

Get the intersection of the datasets accident and person.

accident <- read.csv("data/CRSS/ACCIDENT.csv")  
person <- read.csv("data/CRSS/PERSON.csv")  
vehicle <- read.csv("data/CRSS/VEHICLE.csv")  
  
intersection <- intersect(colnames(accident), colnames(person))  
  
print(intersection)

## [1] "CASENUM" "REGION" "PSU" "PJ" "PSU\_VAR"   
## [6] "URBANICITY" "STRATUM" "VE\_FORMS" "MONTH" "HOUR"   
## [11] "MINUTE" "HARM\_EV" "MAN\_COLL" "SCH\_BUS" "PSUSTRAT"   
## [16] "WEIGHT"

### 2-2

Tabulate the total number of observations in each injury severity (INJ\_SEV).

count\_inj\_sev <- person |>  
 group\_by(INJ\_SEV) |>  
 summarise(observation\_num = n())  
  
count\_inj\_sev

## # A tibble: 8 × 2  
## INJ\_SEV observation\_num  
## <int> <int>  
## 1 0 91720  
## 2 1 21248  
## 3 2 12303  
## 4 3 7230  
## 5 4 1096  
## 6 5 510  
## 7 6 4  
## 8 9 4802

### 2-3

Merge the accident dataset with the vehicle dataset, and report the dimension of your results and number of missing values in one variable of the right dataset.

library("naniar")  
joined <- left\_join(accident, vehicle)  
dim(joined)

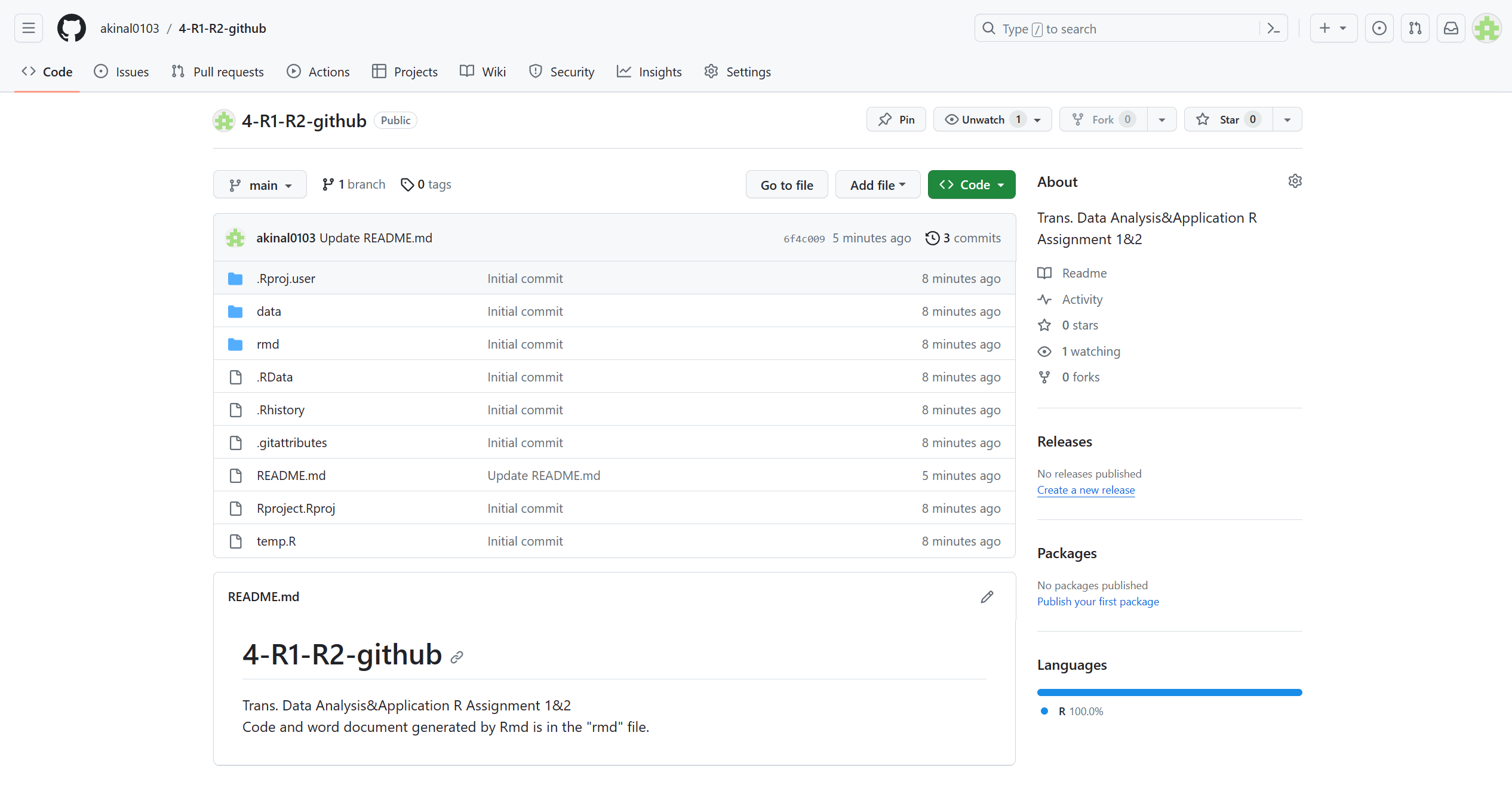
## [1] 97625 123

joined |>  
 miss\_var\_summary()

## # A tibble: 123 × 3  
## variable n\_miss pct\_miss  
## <chr> <int> <dbl>  
## 1 CASENUM 0 0  
## 2 REGION 0 0  
## 3 PSU 0 0  
## 4 PJ 0 0  
## 5 PSU\_VAR 0 0  
## 6 URBANICITY 0 0  
## 7 STRATUM 0 0  
## 8 VE\_TOTAL 0 0  
## 9 VE\_FORMS 0 0  
## 10 PVH\_INVL 0 0  
## # ℹ 113 more rows

## 3 Push your homework to a git repository

Link: <https://github.com/akinal0103/4-R1-R2-github>



## 4 关于本课程的建议

希望老师可以介绍更多交通领域常用的公共数据集，以及利用这些公共数据集做的研究工作。