交通数据分析与应用 R assignment2

学号：2331630

姓名：冯蕊

2023-11-14

# Task1: Using the iri dateset

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

library(dplyr)  
  
# Read csv. data  
iri <- read.csv("LTPP\\iri.csv")  
  
# Filter the data that meets the conditions  
iri\_sub <- subset(iri, STATE\_CODE == 6 & substr(SHRP\_ID,1,3) == '050')  
  
# Show 10 lines  
iri\_sub |> dim()

## [1] 135 5

iri\_sub |>  
 slice\_sample(n=10)

## STATE\_CODE SHRP\_ID CONSTRUCTION\_NO VISIT\_DATE IRI  
## 1 6 0507 2 3/12/05, 12:00:00 AM 1.3214  
## 2 6 0504 2 2/27/97, 12:00:00 AM 1.0700  
## 3 6 0505 4 3/5/03, 12:00:00 AM 1.5264  
## 4 6 0501 2 3/5/99, 12:00:00 AM 1.7332  
## 5 6 0502 2 2/27/97, 12:00:00 AM 1.2565  
## 6 6 0501 2 2/11/98, 12:00:00 AM 1.4376  
## 7 6 0503 2 2/11/98, 12:00:00 AM 1.1262  
## 8 6 0504 3 3/20/07, 12:00:00 AM 1.5176  
## 9 6 0507 2 3/5/99, 12:00:00 AM 1.1696  
## 10 6 0502 4 2/12/02, 12:00:00 AM 1.8666

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

# Calculate the summary statistics  
iri\_stat <- iri |>  
 group\_by(STATE\_CODE, SHRP\_ID) |>  
 summarise(  
 observation = n(),  
 iri\_min = min(IRI),  
 iri\_max = max(IRI),  
 iri\_mean = mean(IRI)  
 ) |>  
 ungroup()  
  
# Show 10 lines  
iri\_stat |> dim()

## [1] 2576 6

iri\_stat |>  
 slice\_sample(n=10)

## # A tibble: 10 × 6  
## STATE\_CODE SHRP\_ID observation iri\_min iri\_max iri\_mean  
## <int> <chr> <int> <dbl> <dbl> <dbl>  
## 1 29 0506 5 1.05 1.13 1.09   
## 2 40 0504 14 0.655 2.08 1.16   
## 3 48 0001 12 0.674 1.80 0.997  
## 4 32 AA03 6 0.650 0.993 0.718  
## 5 5 B430 6 1.49 1.71 1.58   
## 6 48 1123 4 0.757 0.918 0.816  
## 7 35 0802 17 0.913 1.60 1.42   
## 8 89 3015 39 0.783 3.88 1.80   
## 9 47 3101 10 0.766 1.47 1.01   
## 10 16 1007 19 0.832 1.41 1.04

## 3. Sort the summarized data by the averaged IRI in a descending order.

# Sort in a descending order  
iri\_sort <- iri\_stat[order(iri\_stat$iri\_mean, decreasing = T),]  
  
# Show 10 lines  
iri\_sort |> dim()

## [1] 2576 6

iri\_sort |>  
 slice\_sample(n=10)

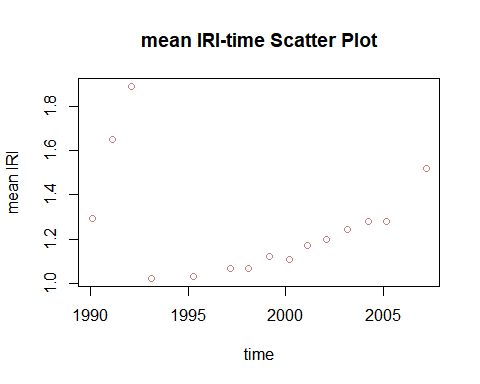
## # A tibble: 10 × 6  
## STATE\_CODE SHRP\_ID observation iri\_min iri\_max iri\_mean  
## <int> <chr> <int> <dbl> <dbl> <dbl>  
## 1 4 0264 26 1.63 2.28 2.00   
## 2 42 1608 11 1.61 1.76 1.69   
## 3 48 1087 12 1.10 1.40 1.28   
## 4 27 2018 6 1.60 3.25 2.54   
## 5 8 A455 6 1.27 1.46 1.34   
## 6 35 6401 5 0.586 2.57 1.55   
## 7 26 A350 8 1.07 2.12 1.59   
## 8 4 A330 6 0.790 1.57 1.11   
## 9 31 7017 7 1.71 3.60 2.57   
## 10 13 0508 9 0.642 1.02 0.825

## 4.1. Generate a scatter plot for the averaged IRI against the time for a selected section.

# Generate a subset with STATE\_CODE=6 and SHRP\_ID=0504  
sel <- subset(iri, STATE\_CODE == 6 & SHRP\_ID == '0504')  
  
# Calculate the summary statistics grouped by VISIT\_DATE  
sel\_stat <- sel |>  
 group\_by(VISIT\_DATE) |>  
 summarise(  
 iri\_mean = mean(IRI)  
 ) |>  
 ungroup()  
  
# Split VISIT\_DATE  
library(tidyr)   
sel\_stat <- separate(sel\_stat, VISIT\_DATE, c("DATE","TIME"), sep=",") |>  
 mutate(  
 DATE = as.Date(DATE,"%m/%d/%y")  
 )  
sel\_stat <- sel\_stat[order(sel\_stat$DATE),]  
  
# Show  
sel\_stat

## # A tibble: 15 × 3  
## DATE TIME iri\_mean  
## <date> <chr> <dbl>  
## 1 1990-01-25 " 12:00:00 AM" 1.29  
## 2 1991-02-16 " 12:00:00 AM" 1.65  
## 3 1992-02-11 " 12:00:00 AM" 1.89  
## 4 1993-02-02 " 12:00:00 AM" 1.02  
## 5 1995-04-04 " 12:00:00 AM" 1.03  
## 6 1997-02-27 " 12:00:00 AM" 1.07  
## 7 1998-02-11 " 12:00:00 AM" 1.07  
## 8 1999-03-05 " 12:00:00 AM" 1.12  
## 9 2000-03-10 " 12:00:00 AM" 1.11  
## 10 2001-02-17 " 12:00:00 AM" 1.17  
## 11 2002-02-12 " 12:00:00 AM" 1.20  
## 12 2003-03-05 " 12:00:00 AM" 1.24  
## 13 2004-03-25 " 12:00:00 AM" 1.28  
## 14 2005-03-12 " 12:00:00 AM" 1.28  
## 15 2007-03-20 " 12:00:00 AM" 1.52

# Generate a scatter plot  
plot(x=sel\_stat$DATE,  
 y=sel\_stat$iri\_mean,   
 col=adjustcolor("darkred", 1/2),  
 xlab="time", ylab="mean IRI",  
 main = "mean IRI-time Scatter Plot")



## 4.2. Give your interpretation of the plot.

如图所示，STATE\_CODE=6且SHRP\_ID=0504的这一子类的IRI均值随时间变化趋势明显。在1990年至1992年间，IRI均值激增，年增长率15%~30%；在1993年断崖式下跌后，1993年至2006年间呈现基本增长趋势，增长幅度较缓，年增长率保持在6%以下；2007年增速显著提高，年增长19%。

# Task2: Using the CRSS datasets in 2017

## 1. Get the intersection of the datasets accident and person.

# Read csv. data  
accident <- read.csv("CRSS\\ACCIDENT.csv")   
person <- read.csv("CRSS\\PERSON.csv")  
vehicle <- read.csv("CRSS\\VEHICLE.csv")  
  
# Get the intersection  
intersection <-   
 inner\_join(  
 x = accident,  
 y = person  
 )

## Joining with `by = join\_by(CASENUM, REGION, PSU, PJ, PSU\_VAR, URBANICITY,  
## STRATUM, VE\_FORMS, MONTH, HOUR, MINUTE, HARM\_EV, MAN\_COLL, SCH\_BUS, PSUSTRAT,  
## WEIGHT)`

# Show 3 lines  
intersection |> dim()

## [1] 138913 96

intersection |>  
 slice\_sample(n=3)

## CASENUM REGION PSU PJ PSU\_VAR URBANICITY STRATUM VE\_TOTAL VE\_FORMS  
## 1 201701231252 1 76 2364 76 1 8 1 1  
## 2 201700359904 3 70 138 70 2 6 2 2  
## 3 201700145453 3 51 1098 51 1 3 1 1  
## PVH\_INVL PEDS PERMVIT PERNOTMVIT NUM\_INJ MONTH YEAR DAY\_WEEK HOUR MINUTE  
## 1 0 0 1 0 1 3 2017 2 23 5  
## 2 0 0 5 0 2 12 2017 7 12 11  
## 3 0 0 1 0 1 9 2017 1 15 3  
## HARM\_EV ALCOHOL MAX\_SEV MAN\_COLL RELJCT1 RELJCT2 TYP\_INT WRK\_ZONE REL\_ROAD  
## 1 30 1 1 0 0 1 1 0 4  
## 2 12 2 2 6 0 2 2 0 1  
## 3 1 2 2 0 0 1 1 0 1  
## LGT\_COND WEATHER1 WEATHER2 WEATHER SCH\_BUS INT\_HWY CF1 CF2 CF3 WKDY\_IM  
## 1 3 98 0 98 0 0 0 0 0 2  
## 2 1 10 0 10 0 0 0 0 0 7  
## 3 1 2 0 2 0 0 0 0 0 1  
## HOUR\_IM MINUTE\_IM EVENT1\_IM MANCOL\_IM RELJCT1\_IM RELJCT2\_IM LGTCON\_IM  
## 1 23 5 30 0 0 1 3  
## 2 12 11 12 6 0 2 1  
## 3 15 3 1 0 0 1 1  
## WEATHR\_IM MAXSEV\_IM NO\_INJ\_IM ALCHL\_IM PSUSTRAT WEIGHT VEH\_NO PER\_NO  
## 1 10 1 1 1 2 206.42075 1 1  
## 2 10 2 2 2 16 52.06651 1 2  
## 3 2 2 1 2 10 24.94295 1 1  
## STR\_VEH MAKE BODY\_TYP MOD\_YEAR MAK\_MOD TOW\_VEH SPEC\_USE EMER\_USE ROLLOVER  
## 1 0 37 4 2009 37032 0 0 0 9  
## 2 0 37 4 2004 37031 0 0 0 0  
## 3 0 53 80 2005 53706 0 0 0 0  
## IMPACT1 FIRE\_EXP AGE SEX PER\_TYP INJ\_SEV SEAT\_POS REST\_USE REST\_MIS AIR\_BAG  
## 1 2 0 32 1 1 1 11 3 0 9  
## 2 11 0 28 1 2 0 13 3 0 20  
## 3 0 0 33 2 1 2 11 19 0 20  
## EJECTION DRINKING ALC\_STATUS ATST\_TYP ALC\_RES DRUGS DSTATUS DRUGTST1 DRUGTST2  
## 1 0 1 2 1 160 0 0 0 0  
## 2 0 8 0 0 996 8 0 0 0  
## 3 8 0 0 0 996 0 0 0 0  
## DRUGTST3 DRUGRES1 DRUGRES2 DRUGRES3 HOSPITAL P\_SF1 P\_SF2 P\_SF3 LOCATION  
## 1 0 0 0 0 2 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0 0  
## SEX\_IM INJSEV\_IM EJECT\_IM PERALCH\_IM SEAT\_IM AGE\_IM  
## 1 1 1 0 1 11 32  
## 2 1 0 0 0 13 28  
## 3 2 2 8 0 11 33

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

# Calculate the summary statistics grouped by INJ\_SEV  
injury\_severity <- person |>  
 group\_by(INJ\_SEV) |>  
 summarise(  
 obversation = n()  
 ) |>  
 ungroup()  
  
# Show  
injury\_severity

## # A tibble: 8 × 2  
## INJ\_SEV obversation  
## <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

## 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)  
# Merge ACCIDENT with the VEHICLE  
merge <- left\_join(  
 x = accident,  
 y = vehicle  
)  
  
# Report the dimension of results & number of missing values  
merge |> 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

# Task3: Push your homework to a git repository

## Provide a link to your repository

### Assignment1

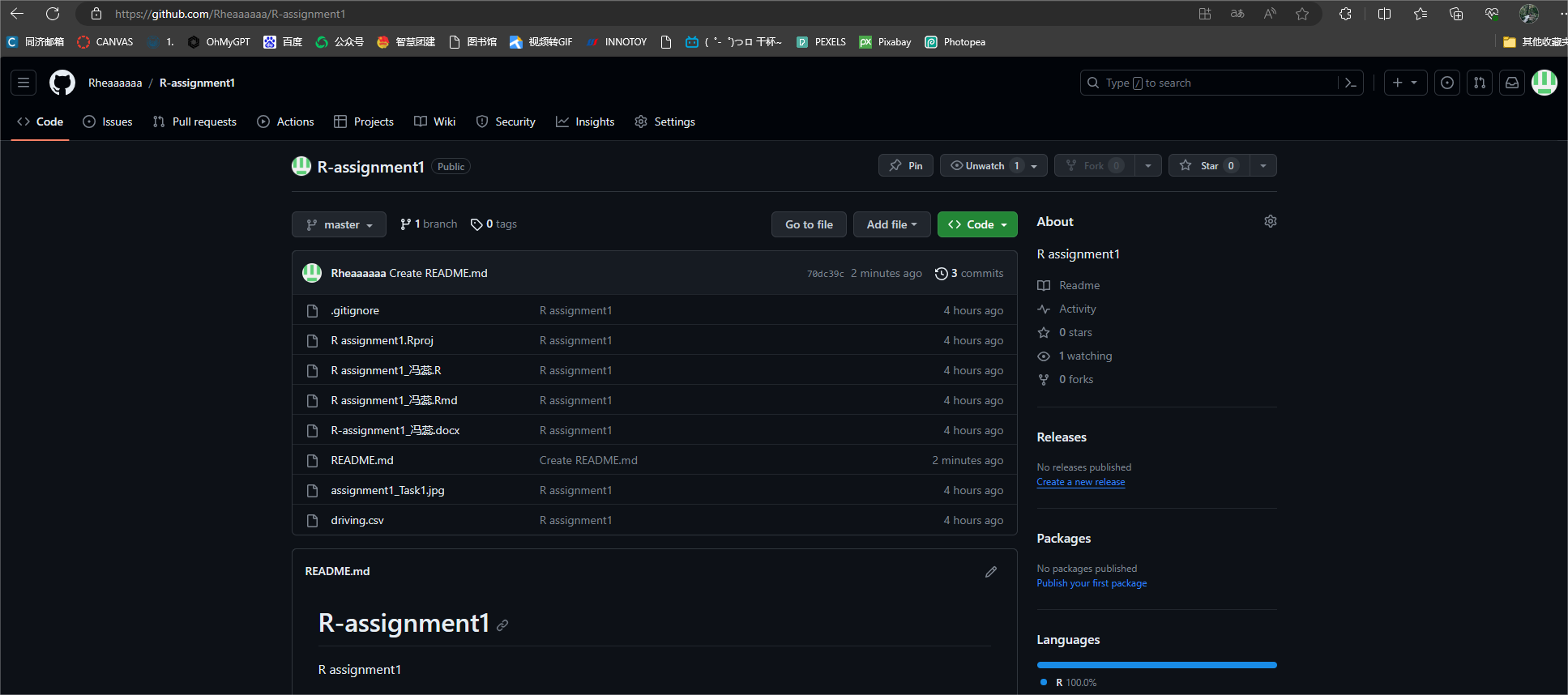
Assignment1: <https://github.com/Rheaaaaaa/R-assignment1>

### Assignment2

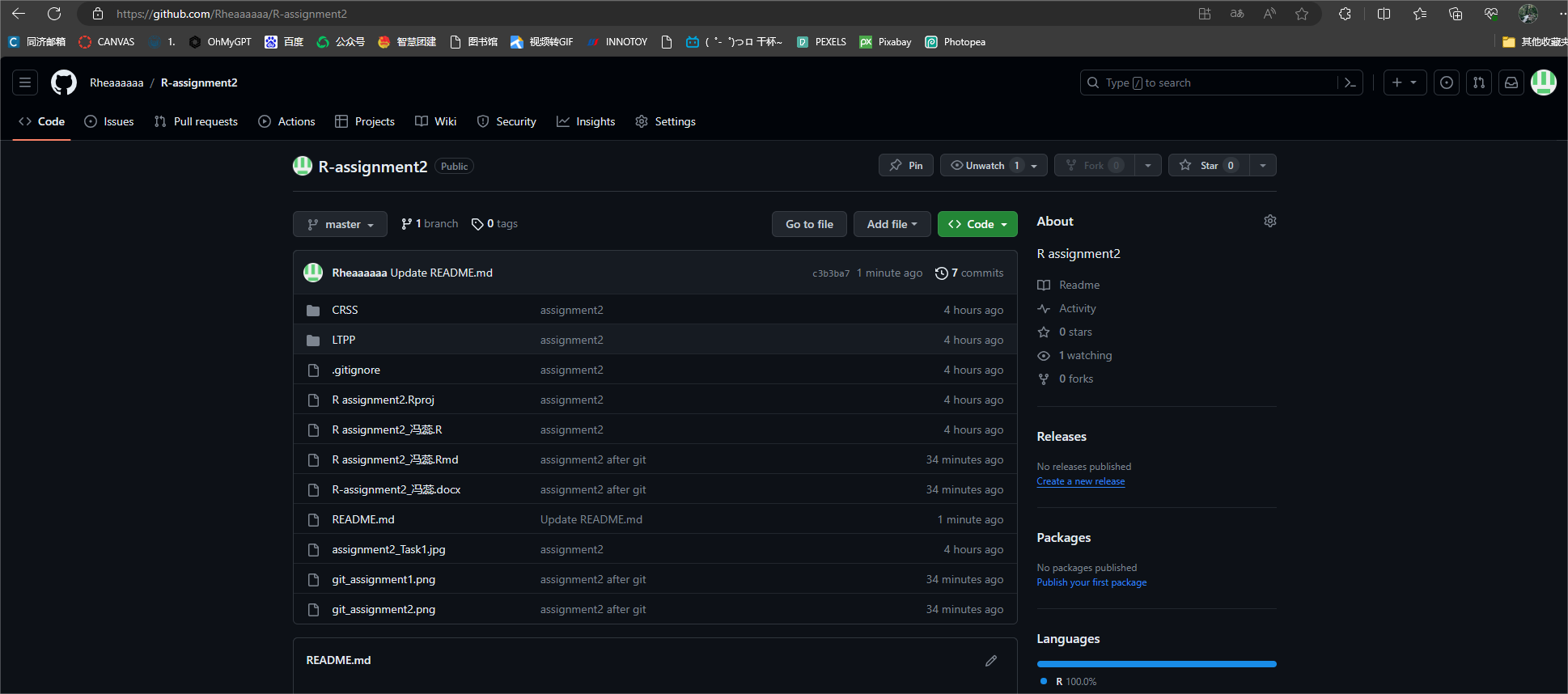
Assignment2: <https://github.com/Rheaaaaaa/R-assignment2>

## Provide a screenshot of your repository

### Assignment1



### Assignment2



# Task4: 关于本课程的建议

## 本课程的内容

我认为r语言对于我自己来说还是有一定难度的，如果能在课堂上有更多的时间动手实操，可能会对我的理解有一些帮助。但老师的ppt非常详细，在课后我对照ppt并在网络的帮助下，还是能够比较顺利地完成这次作业。

## 编程语言与编程工具

在课后自学markdown的时候我觉得它很有趣也很方便，如果老师能在课堂上讲述一些rmd相关内容的话，可能对我们理解和应用r语言有所帮助。