Assignment2

zhaoxiuping

2023-11-15

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

library("dplyr")

##   
## Attaching package: 'dplyr'

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

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

library("ggplot2")  
library("tidyr")  
data2 <- read.csv("/Users/zhaoxiuping/R\_zxp/data-assignment/LTPP/iri.csv")  
attach(data2)  
# data2部分数据展示  
sampled\_data <- slice\_sample(data2, n=10)  
sampled\_data

## STATE\_CODE SHRP\_ID CONSTRUCTION\_NO VISIT\_DATE IRI  
## 1 20 1005 4 12/8/15, 12:00:00 AM 1.0528  
## 2 26 7072 2 4/13/99, 12:00:00 AM 1.4606  
## 3 21 3016 1 4/8/01, 12:00:00 AM 1.4282  
## 4 1 0109 1 1/29/03, 12:00:00 AM 0.7208  
## 5 29 0604 4 3/24/03, 12:00:00 AM 1.5468  
## 6 26 0115 4 8/10/10, 12:00:00 AM 1.1486  
## 7 90 0903 1 5/15/02, 12:00:00 AM 1.1492  
## 8 40 0506 3 4/11/07, 12:00:00 AM 1.1444  
## 9 13 1031 3 3/13/02, 12:00:00 AM 0.5096  
## 10 6 0505 1 2/16/91, 12:00:00 AM 1.4082

subset\_data <- data2 |>  
 subset(stringr::str\_detect(SHRP\_ID,"050") & STATE\_CODE==6)  
 # |>  
 # slice\_sample(n=4)  
# subset\_data部分数据展示  
sampled\_data <- slice\_sample(subset\_data, n=10)  
sampled\_data

## STATE\_CODE SHRP\_ID CONSTRUCTION\_NO VISIT\_DATE IRI  
## 1 6 0508 2 3/5/99, 12:00:00 AM 1.0704  
## 2 6 0503 2 2/27/97, 12:00:00 AM 1.1220  
## 3 6 0508 3 3/21/07, 12:00:00 AM 1.8240  
## 4 6 0508 3 3/12/05, 12:00:00 AM 1.1960  
## 5 6 0502 2 2/11/98, 12:00:00 AM 1.4020  
## 6 6 0503 4 3/10/00, 12:00:00 AM 1.3338  
## 7 6 0506 4 2/12/02, 12:00:00 AM 1.0618  
## 8 6 0504 3 3/5/03, 12:00:00 AM 1.2448  
## 9 6 0508 1 1/25/90, 12:00:00 AM 1.7724  
## 10 6 0507 2 4/4/95, 12:00:00 AM 0.9748

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

subset\_data2 <- data2 |>  
 group\_by(STATE\_CODE,SHRP\_ID) |>  
 summarise(  
 observation = n(),  
 # 填补缺失值  
 iri\_max = max(IRI,na.rm = TRUE),  
 iri\_min = min(IRI,na.rm = TRUE),  
 iri\_median = median(IRI,na.rm = TRUE),  
 iri\_mean = mean(IRI,na.rm = TRUE)  
 ) |>  
ungroup()

## `summarise()` has grouped output by 'STATE\_CODE'. You can override using the  
## `.groups` argument.

#部分数据展示  
sampled\_data2 <- slice\_sample(subset\_data2, n=10)  
sampled\_data2

## # A tibble: 10 × 7  
## STATE\_CODE SHRP\_ID observation iri\_max iri\_min iri\_median iri\_mean  
## <int> <chr> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 27 1016 10 3.13 0.985 1.98 1.89   
## 2 31 1030 11 2.18 1.03 1.37 1.47   
## 3 29 0703 11 3.23 0.965 1.72 1.77   
## 4 10 0112 14 0.622 0.525 0.556 0.563  
## 5 34 0902 13 1.97 0.853 0.893 0.982  
## 6 56 2017 20 1.94 0.764 1.51 1.44   
## 7 29 A603 6 1.37 1.15 1.27 1.26   
## 8 53 A809 18 1.17 0.838 0.971 0.986  
## 9 50 1004 22 2.27 0.609 1.31 1.39   
## 10 6 2053 10 1.70 1.32 1.46 1.49

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

subset\_data3 <- arrange(subset\_data2,desc(iri\_mean))  
# 部分数据展示  
sampled\_data3 <- slice\_sample(subset\_data3, n=1)  
sampled\_data3

## # A tibble: 1 × 7  
## STATE\_CODE SHRP\_ID observation iri\_max iri\_min iri\_median iri\_mean  
## <int> <chr> <int> <dbl> <dbl> <dbl> <dbl>  
## 1 83 0506 21 2.94 1.42 1.80 1.91

Generate a scatter plot for the averaged IRI against the time for a selected section, and then give your interpretation of the plot: • HINT1:meanIRIvs.date • HINT2:STATE\_CODEandSHRP\_IDtogethertoformaprimarykeythatuniquelyidentifiesasection

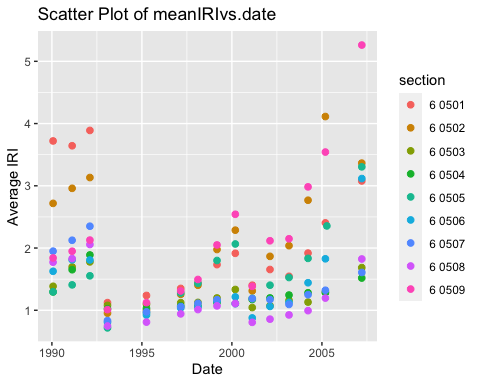
# STATE\_CODE=6 SHRP\_ID 以050 开始的section 作图，包含图表  
subset\_data4 <- subset\_data |>  
 group\_by(VISIT\_DATE,STATE\_CODE,SHRP\_ID) |>  
 summarise(  
 iri\_mean = mean(IRI,na.rm = TRUE)  
 ) |>  
ungroup()

## `summarise()` has grouped output by 'VISIT\_DATE', 'STATE\_CODE'. You can  
## override using the `.groups` argument.

subset\_data5 <- subset\_data4 |>  
 group\_by(VISIT\_DATE,STATE\_CODE,SHRP\_ID) |>  
 separate(VISIT\_DATE, c("VISIT\_DATE", NULL), sep=",") |> mutate(  
 VISIT\_DATE = as.Date(VISIT\_DATE, "%m/%d/%y"),   
 iri\_mean = iri\_mean  
 )

## Warning: Expected 1 pieces. Additional pieces discarded in 135 rows [1, 2, 3, 4, 5, 6,  
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

data <- data.frame(  
 VISIT\_DATE = subset\_data5$VISIT\_DATE,  
 IRI\_MEAN = subset\_data5$iri\_mean,  
 STATE\_CODE = subset\_data5$STATE\_CODE,  
 SHRP\_ID = subset\_data5$SHRP\_ID  
)  
library("ggplot2")  
scatter\_plot <- ggplot(data, aes(x = VISIT\_DATE, y = IRI\_MEAN)) +  
 geom\_point(aes(color = paste(STATE\_CODE, SHRP\_ID)), size = 2) +  
 labs(x = "Date", y = "Average IRI", title = "Scatter Plot of meanIRIvs.date") +  
 labs(color = "section",size = 2)+  
 theme()   
print(scatter\_plot)



#全部的section作图，无图标，图标太多无法显示完全图  
subset\_data6 <- data2 |>  
 group\_by(VISIT\_DATE,STATE\_CODE,SHRP\_ID) |>  
 summarise(  
 iri\_mean = mean(IRI,na.rm = TRUE)  
 ) |>  
ungroup()

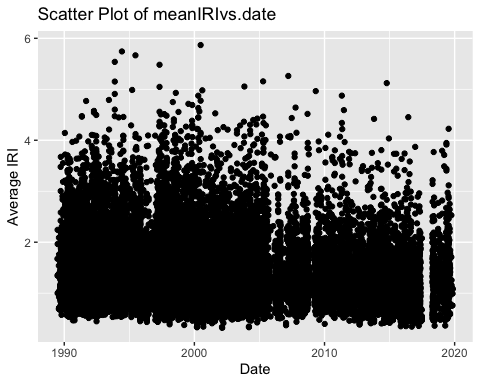
## `summarise()` has grouped output by 'VISIT\_DATE', 'STATE\_CODE'. You can  
## override using the `.groups` argument.

subset\_data7 <- subset\_data6 |>  
 group\_by(VISIT\_DATE,STATE\_CODE,SHRP\_ID) |>  
 separate(VISIT\_DATE, c("VISIT\_DATE", NULL), sep=",") |> mutate(  
 VISIT\_DATE = as.Date(VISIT\_DATE, "%m/%d/%y"),   
 iri\_mean = iri\_mean  
 )

## Warning: Expected 1 pieces. Additional pieces discarded in 29329 rows [1, 2, 3, 4, 5, 6,  
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

data <- data.frame(  
 VISIT\_DATE = subset\_data7$VISIT\_DATE,  
 IRI\_MEAN = subset\_data7$iri\_mean,  
 STATE\_CODE = subset\_data7$STATE\_CODE,  
 SHRP\_ID = subset\_data7$SHRP\_ID  
)  
  
scatter\_plot <- ggplot(data, aes(x = VISIT\_DATE, y = IRI\_MEAN)) +  
 geom\_point() +  
 labs(x = "Date", y = "Average IRI", title = "Scatter Plot of meanIRIvs.date") +  
 labs(color = "section",size = 2)+  
 theme()   
print(scatter\_plot)

## Warning: Removed 4 rows containing missing values (`geom\_point()`).



Get the intersection of the datasets accident and person:

accident <- data.frame(read.csv("/Users/zhaoxiuping/R\_zxp/data-assignment/CRSS/ACCIDENT.csv"))  
person <- data.frame(read.csv("/Users/zhaoxiuping/R\_zxp/data-assignment/CRSS/PERSON.csv"))  
inter <- intersect(colnames(accident),colnames(person))  
  
inter\_ap <- inner\_join(  
 x=accident,  
 y=person,  
 by = inter  
)  
# inter\_ap部分数据展示  
sampled\_inter\_ap <- slice\_sample(inter\_ap, n=5)  
sampled\_inter\_ap

## CASENUM REGION PSU PJ PSU\_VAR URBANICITY STRATUM VE\_TOTAL VE\_FORMS  
## 1 201701484831 3 35 2799 35 1 6 2 2  
## 2 201701431726 1 72 1723 72 1 2 1 1  
## 3 201700034905 3 59 506 59 1 9 2 2  
## 4 201700124225 4 68 299 68 1 4 2 2  
## 5 201701532200 3 40 3106 40 1 10 2 2  
## PVH\_INVL PEDS PERMVIT PERNOTMVIT NUM\_INJ MONTH YEAR DAY\_WEEK HOUR MINUTE  
## 1 0 0 3 0 1 3 2017 4 17 15  
## 2 0 1 6 1 1 5 2017 1 11 20  
## 3 0 0 2 0 0 8 2017 3 13 37  
## 4 0 0 2 0 1 9 2017 7 15 27  
## 5 0 0 2 0 0 6 2017 2 15 58  
## HARM\_EV ALCOHOL MAX\_SEV MAN\_COLL RELJCT1 RELJCT2 TYP\_INT WRK\_ZONE REL\_ROAD  
## 1 12 2 1 6 0 2 3 0 1  
## 2 9 2 1 0 8 3 98 0 1  
## 3 12 2 0 1 8 3 98 0 1  
## 4 12 1 3 1 0 1 1 0 1  
## 5 12 2 0 2 0 2 2 0 1  
## LGT\_COND WEATHER1 WEATHER2 WEATHER SCH\_BUS INT\_HWY CF1 CF2 CF3 WKDY\_IM  
## 1 1 1 0 1 0 0 0 0 0 4  
## 2 1 10 0 10 0 0 0 0 0 1  
## 3 1 1 0 1 0 0 0 0 0 3  
## 4 1 10 0 10 0 1 0 0 0 7  
## 5 1 98 0 98 0 0 0 0 0 2  
## HOUR\_IM MINUTE\_IM EVENT1\_IM MANCOL\_IM RELJCT1\_IM RELJCT2\_IM LGTCON\_IM  
## 1 17 15 12 6 0 2 1  
## 2 11 20 9 0 0 3 1  
## 3 13 37 12 1 0 3 1  
## 4 15 27 12 1 0 1 1  
## 5 15 58 12 2 0 2 1  
## WEATHR\_IM MAXSEV\_IM NO\_INJ\_IM ALCHL\_IM PSUSTRAT WEIGHT VEH\_NO PER\_NO  
## 1 1 1 1 2 13 53.37274 2 1  
## 2 10 1 1 2 1 24.63940 1 4  
## 3 1 0 0 2 10 155.60109 1 1  
## 4 10 3 1 1 23 15.55987 1 1  
## 5 10 0 0 2 15 208.24141 1 1  
## STR\_VEH MAKE BODY\_TYP MOD\_YEAR MAK\_MOD TOW\_VEH SPEC\_USE EMER\_USE ROLLOVER  
## 1 0 63 4 2016 63038 0 0 0 0  
## 2 0 49 19 2005 49499 0 0 0 0  
## 3 0 32 1 2013 32049 0 0 0 0  
## 4 0 49 4 2017 49032 0 0 0 0  
## 5 0 30 14 2012 30402 0 0 0 0  
## IMPACT1 FIRE\_EXP AGE SEX PER\_TYP INJ\_SEV SEAT\_POS REST\_USE REST\_MIS AIR\_BAG  
## 1 1 0 29 2 1 0 11 3 0 20  
## 2 12 0 1 1 2 0 98 3 0 98  
## 3 12 0 32 2 1 0 11 3 0 20  
## 4 12 0 25 1 1 3 11 20 0 9  
## 5 12 0 45 2 1 0 11 3 0 20  
## EJECTION DRINKING ALC\_STATUS ATST\_TYP ALC\_RES DRUGS DSTATUS DRUGTST1 DRUGTST2  
## 1 0 0 0 0 996 0 0 0 0  
## 2 0 8 0 0 996 8 0 0 0  
## 3 0 0 0 0 996 0 0 0 0  
## 4 0 1 2 2 997 0 8 6 0  
## 5 0 0 8 95 995 0 8 6 0  
## DRUGTST3 DRUGRES1 DRUGRES2 DRUGRES3 HOSPITAL P\_SF1 P\_SF2 P\_SF3 LOCATION  
## 1 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0 0  
## 4 0 95 0 0 1 0 0 0 0  
## 5 0 95 0 0 0 0 0 0 0  
## SEX\_IM INJSEV\_IM EJECT\_IM PERALCH\_IM SEAT\_IM AGE\_IM  
## 1 2 0 0 0 11 29  
## 2 1 0 0 0 23 1  
## 3 2 0 0 0 11 32  
## 4 1 3 0 1 11 25  
## 5 2 0 0 0 11 45

Tabulate the total number of observations in each injury severity (INJ\_SEV) • HINT:usesummarise()andgroup\_by()

observations\_injury\_severity <- inter\_ap |>  
 group\_by(INJ\_SEV) |>  
 summarise(  
 count = n()  
 )  
observations\_injury\_severity

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

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 • HINT:left\_join()

vehicle <- read.csv("/Users/zhaoxiuping/R\_zxp/data-assignment/CRSS/VEHICLE.CSV")  
inter\_2 <- intersect(colnames(accident),colnames(vehicle))  
left\_av <- left\_join(  
 x=accident,  
 y=vehicle,  
 # by = c("CASENUM","PSU")  
 by = inter\_2  
) |>  
 distinct()  
#右侧数据集的缺失值  
missing\_VEH\_NO <- sum(is.na(left\_av$VEH\_NO))  
print(missing\_VEH\_NO)

## [1] 0

#合并后的数据集维度  
dim(left\_av)

## [1] 97625 123