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 35 0101 1 6/9/04, 12:00:00 AM 1.15400  
## 2 35 AA62 1 9/10/14, 12:00:00 AM 0.69060  
## 3 29 B310 3 4/26/01, 12:00:00 AM 1.17980  
## 4 16 A350 2 8/14/91, 12:00:00 AM 0.67925  
## 5 4 1022 3 3/12/04, 12:00:00 AM 0.57200  
## 6 51 2564 1 12/11/89, 12:00:00 AM 0.95580  
## 7 29 0960 3 12/18/96, 12:00:00 AM 1.03600  
## 8 20 3060 1 2/15/00, 12:00:00 AM 1.73480  
## 9 39 4018 3 6/28/10, 12:00:00 AM 0.85320  
## 10 8 0213 1 10/4/02, 12:00:00 AM 1.16580

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 0507 2 3/5/99, 12:00:00 AM 1.1696  
## 2 6 0505 3 3/10/00, 12:00:00 AM 2.0654  
## 3 6 0506 5 3/12/05, 12:00:00 AM 1.8272  
## 4 6 0508 2 2/11/98, 12:00:00 AM 1.0112  
## 5 6 0507 1 2/16/91, 12:00:00 AM 2.1252  
## 6 6 0509 2 4/4/95, 12:00:00 AM 1.1222  
## 7 6 0503 5 2/17/01, 12:00:00 AM 1.0436  
## 8 6 0506 1 2/11/92, 12:00:00 AM 1.8082  
## 9 6 0501 2 2/27/97, 12:00:00 AM 1.3510  
## 10 6 0508 1 1/25/90, 12:00:00 AM 1.7724

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 1 C310 6 1.18 0.916 1.04 1.05   
## 2 6 A340 10 2.20 1.57 1.90 1.89   
## 3 46 0603 17 2.61 0.431 1.34 1.44   
## 4 42 B350 9 3.21 2.11 2.48 2.62   
## 5 4 A442 8 1.19 0.639 0.765 0.832  
## 6 4 1017 19 1.35 0.545 0.798 0.863  
## 7 30 7076 23 1.26 0.616 0.972 0.941  
## 8 4 6053 4 1.60 1.22 1.37 1.39   
## 9 51 0121 26 1.79 0.981 1.12 1.17   
## 10 12 0506 17 0.970 0.458 0.564 0.590

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 13 4118 7 0.672 0.576 0.598 0.611

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

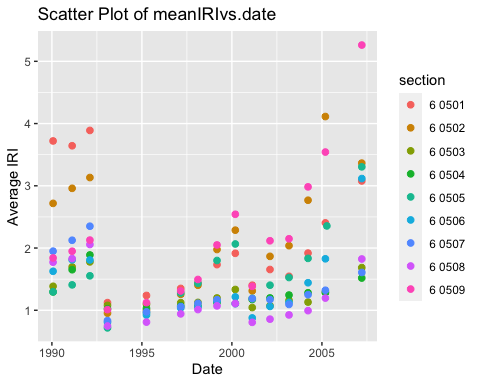
# STATE\_CODE=6 SHRP\_ID 以050 开始的section 作图，包含图表  
subset\_data4 <- subset\_data |>  
 group\_by(VISIT\_DATE,STATE\_CODE,SHRP\_ID) |>  
 summarise(  
 iri\_mean = mean(IRI,na.rm = TRUE)  
 ) |>  
ungroup()

## `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  
)  
  
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()

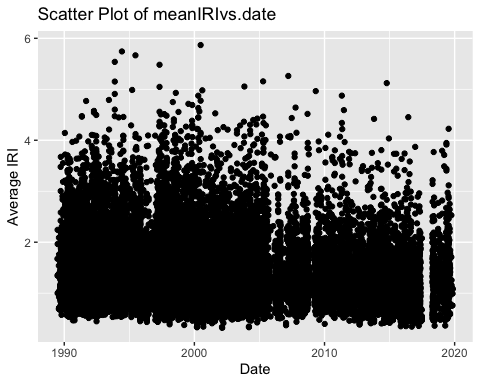
## `summarise()` has grouped output by 'VISIT\_DATE', 'STATE\_CODE'. You can  
## override using the `.groups` argument.

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 201700363813 4 41 163 41 1 6 3 3  
## 2 201700083300 4 41 159 41 1 10 1 1  
## 3 201701428417 1 31 1714 31 1 5 2 2  
## 4 201701272592 2 47 2001 47 2 10 2 2  
## 5 201700038409 3 30 3296 30 1 10 2 2  
## PVH\_INVL PEDS PERMVIT PERNOTMVIT NUM\_INJ MONTH YEAR DAY\_WEEK HOUR MINUTE  
## 1 0 0 5 0 2 10 2017 3 17 16  
## 2 0 0 2 0 0 7 2017 6 17 39  
## 3 0 0 2 0 1 3 2017 4 14 30  
## 4 0 0 3 0 0 3 2017 5 11 33  
## 5 0 0 2 0 0 5 2017 3 14 20  
## HARM\_EV ALCOHOL MAX\_SEV MAN\_COLL RELJCT1 RELJCT2 TYP\_INT WRK\_ZONE REL\_ROAD  
## 1 12 2 1 1 0 3 2 0 1  
## 2 59 2 0 0 1 3 2 0 4  
## 3 12 9 3 6 8 2 3 0 1  
## 4 12 2 0 1 8 3 3 0 1  
## 5 12 2 0 7 0 1 1 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 3  
## 2 1 1 0 1 0 0 0 0 0 6  
## 3 1 1 0 1 0 0 0 0 0 4  
## 4 1 1 0 1 0 0 0 0 0 5  
## 5 1 10 0 10 0 0 0 0 0 3  
## HOUR\_IM MINUTE\_IM EVENT1\_IM MANCOL\_IM RELJCT1\_IM RELJCT2\_IM LGTCON\_IM  
## 1 17 16 12 1 0 3 1  
## 2 17 39 59 0 1 3 1  
## 3 14 30 12 6 0 2 1  
## 4 11 33 12 1 0 3 1  
## 5 14 20 12 7 0 1 1  
## WEATHR\_IM MAXSEV\_IM NO\_INJ\_IM ALCHL\_IM PSUSTRAT WEIGHT VEH\_NO PER\_NO  
## 1 1 2 3 2 22 56.80253 3 2  
## 2 1 0 0 2 22 223.42874 1 1  
## 3 1 3 1 2 2 28.77843 2 1  
## 4 1 0 0 2 9 205.30186 2 2  
## 5 10 0 0 2 14 217.35132 2 1  
## STR\_VEH MAKE BODY\_TYP MOD\_YEAR MAK\_MOD TOW\_VEH SPEC\_USE EMER\_USE ROLLOVER  
## 1 0 58 14 2014 58403 0 0 0 0  
## 2 0 49 4 1999 49040 0 0 0 0  
## 3 0 20 4 2002 20399 0 0 0 0  
## 4 0 18 4 2011 18019 0 0 0 0  
## 5 0 22 4 2004 22018 0 0 0 0  
## IMPACT1 FIRE\_EXP AGE SEX PER\_TYP INJ\_SEV SEAT\_POS REST\_USE REST\_MIS AIR\_BAG  
## 1 6 0 15 2 2 0 13 3 0 98  
## 2 12 0 61 2 1 0 11 3 0 98  
## 3 82 0 55 2 1 3 11 3 0 98  
## 4 6 0 998 8 2 0 98 98 0 98  
## 5 9 0 43 2 1 0 11 3 0 20  
## EJECTION DRINKING ALC\_STATUS ATST\_TYP ALC\_RES DRUGS DSTATUS DRUGTST1 DRUGTST2  
## 1 7 8 0 0 996 8 0 0 0  
## 2 0 0 8 95 995 0 8 6 0  
## 3 0 0 8 95 995 0 8 6 0  
## 4 7 8 0 0 996 8 0 0 0  
## 5 0 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 0 0 0 0 0  
## 2 0 95 0 0 0 0 0 0 0  
## 3 0 95 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0 0  
## SEX\_IM INJSEV\_IM EJECT\_IM PERALCH\_IM SEAT\_IM AGE\_IM  
## 1 2 0 0 0 13 15  
## 2 2 0 0 0 11 61  
## 3 2 3 0 0 11 55  
## 4 2 0 0 0 13 59  
## 5 2 0 0 0 11 43

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

关于本课程的建议： 从课程内容方面，建议丰富利用R语言进行数据可视化方面的讲解，通过真实案例进行演示和练习；也希望能够更多的学习有关数据应用的算法例如回归分析、聚类及机器学习算法等的应用。 关于编程语言和编程工具，我认为龚红仁老师的讲解已经足够深刻和细致，课堂的讲解已经能够帮助我深入了解 R 语言的编程技巧以及Rstudio的使用。

本次作业已经上传至Git\_hub，地址是：<https://github.com/zxppp-wp/R_zxp>。

