Assignment 4

Hyungue Lim

9/29/2019

```r
library("tidyverse")
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1

```r
#a
2019 %>% sin
```

```
## [1] 0.8644605
```

```r
#b
2019 %>% cos %>% sin
```

```
## [1] -0.4817939
```

```r
#c
2019 %>% log %>% tan %>% cos %>% sin
```

```
## [1] -0.5939393
```

```r
#d
2019 %>% log(base = 2)
```

```
## [1] 10.97943
```

2

```
library(readxl)
c2015 <- read_excel("~/Math 421/c2015.xlsx")
c2015$SEX[c2015$SEX == "Unknown"] <- "Female"

c2015$AGE <- c2015$AGE %>% recode("Less than 1" = "0") %>% as.numeric
```

```
## Warning in function_list[[k]](value): NAs introduced by coercion
```

```
c2015$AGE <- c2015$AGE %>% replace_na(mean(c2015$AGE, na.rm=TRUE))


library("stringr")
c2015$TRAV_SP <- c2015$TRAV_SP %>% str_replace(" MPH", "") %>% str_replace("Stopped", "0") %>% as.numer
```

```
## Warning in function_list[[k]](value): NAs introduced by coercion
```

```
c2015 <- c2015 %>% filter(!is.na(TRAV_SP))
```

3

```
c2015 <- c2015 %>% mutate(date = paste(sep = "/",YEAR,MONTH,DAY))

c2015$date <- as.Date(c2015$date, format = "%Y/%B/%d")
c2015$weekdays <- weekdays(c2015$date)

c2015 %>% filter(weekdays == "Saturday" | weekdays == "Sunday", SEX == "Female") %>% summarize(avg_spee
```

```
## # A tibble: 1 x 2
##   avg_speed avg_age
##       <dbl>   <dbl>
## 1      44.7    36.2
```

```
#Realized day of the week is provided
c2015 <- c2015[,-c(29,30)]
```

4

```
num_v <- c2015 %>%  select_if(is.numeric)
names(num_v)
```

```
##  [1] "ST_CASE"  "VEH_NO"   "PER_NO"   "COUNTY"   "DAY"      "HOUR"
##  [7] "MINUTE"   "AGE"      "YEAR"     "TRAV_SP"  "LATITUDE" "LONGITUD"
```

5

```
c2015 %>% select_if(is.numeric) %>% summarize_all(~mean(., na.rm = TRUE))
```

```
## # A tibble: 1 x 12
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 251487.   1.63   1.66   76.2  15.4  13.8   28.6  38.8  2015    44.5
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

6

```
c2015 %>% summarize_if(is.numeric, ~mean(., na.rm= TRUE))
```

```
## # A tibble: 1 x 12
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 251487.   1.63   1.66   76.2  15.4  13.8   28.6  38.8  2015    44.5
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

7

```
c2015 %>% summarize_if(is.numeric, ~median(., na.rm= TRUE))
```

```
## # A tibble: 1 x 12
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1  220480      1      1     67    15    15     30    36  2015      50
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

8

```
c2015 %>% summarize_if(is.numeric, ~sd(., na.rm= TRUE))
```

```
## # A tibble: 1 x 12
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 169431.   1.52   1.63   75.6  8.79  7.63   17.4  20.2     0    25.1
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

9

```
c2015 %>% summarize_if(is.numeric, ~sum(is.na(.)))
```

```
## # A tibble: 1 x 12
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <int>  <int>  <int>  <int> <int> <int>  <int> <int> <int>   <int>
## 1       0      0      0      0     0     0     47     0     0       0
## # ... with 2 more variables: LATITUDE <int>, LONGITUD <int>
```

10

```
c2015 %>% summarize_if(is.numeric, ~log(mean(., na.rm= TRUE)))
```

```
## Warning in log(mean(., na.rm = TRUE)): NaNs produced
```

```
## # A tibble: 1 x 12
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1    12.4  0.486  0.506   4.33  2.73  2.62   3.35  3.66  7.61    3.80
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

11

```
c2015 %>% summarize_if(is.numeric, ~log(abs(mean(., na.rm= TRUE))))
```

```
## # A tibble: 1 x 12
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1    12.4  0.486  0.506   4.33  2.73  2.62   3.35  3.66  7.61    3.80
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

12

```
c2015 %>% summarize_if(is.character, ~sum(.=="Unknown"))
```

```
## # A tibble: 1 x 16
##   STATE MONTH   SEX PER_TYP INJ_SEV SEAT_POS DRINKING MAN_COLL OWNER
##   <int> <int> <int>   <int>   <int>    <int>    <int>    <int> <int>
## 1     0     0     0       0     124      156        0       33   257
## # ... with 7 more variables: MOD_YEAR <int>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, HARM_EV <int>, LGT_COND <int>,
## #   WEATHER <int>
```

13

```
c2015 %>% select_if(is.character) %>% summarize_all(~sum(.=="Unknown"))
```

```
## # A tibble: 1 x 16
##   STATE MONTH   SEX PER_TYP INJ_SEV SEAT_POS DRINKING MAN_COLL OWNER
##   <int> <int> <int>   <int>   <int>    <int>    <int>    <int> <int>
## 1     0     0     0       0     124      156        0       33   257
## # ... with 7 more variables: MOD_YEAR <int>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, HARM_EV <int>, LGT_COND <int>,
## #   WEATHER <int>
```

14

```
length(table(c2015$STATE))
```

```
## [1] 51
```

15

```
c2015 %>% summarize_if(is.character, ~n_distinct(.,na.rm=TRUE))
```

```
## # A tibble: 1 x 16
##   STATE MONTH   SEX PER_TYP INJ_SEV SEAT_POS DRINKING MAN_COLL OWNER
##   <int> <int> <int>   <int>   <int>    <int>    <int>    <int> <int>
## 1    51    12     3       3       8       26        4       10     8
## # ... with 7 more variables: MOD_YEAR <int>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, HARM_EV <int>, LGT_COND <int>,
## #   WEATHER <int>
```

16

```
c2015 %>% select_if(is.character) %>% summarize_all( ~n_distinct(.,na.rm=TRUE))
```

```
## # A tibble: 1 x 16
##   STATE MONTH   SEX PER_TYP INJ_SEV SEAT_POS DRINKING MAN_COLL OWNER
##   <int> <int> <int>   <int>   <int>    <int>    <int>    <int> <int>
## 1    51    12     3       3       8       26        4       10     8
## # ... with 7 more variables: MOD_YEAR <int>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, HARM_EV <int>, LGT_COND <int>,
## #   WEATHER <int>
```

17

```r
c2015 %>% select_if(~n_distinct(., na.rm=TRUE) > 30) %>% names
```

```
##  [1] "STATE"    "ST_CASE"  "VEH_NO"   "PER_NO"   "COUNTY"   "DAY"
##  [7] "MINUTE"   "AGE"      "MOD_YEAR" "TRAV_SP"  "LATITUDE" "LONGITUD"
## [13] "HARM_EV"
```

18

```r
c2015 %>% select_if(is.character) %>% select_if(~n_distinct(., na.rm=TRUE) > 30) %>% names
```

```
## [1] "STATE"    "MOD_YEAR" "HARM_EV"
```

19

```r
c2015 %>% select_if(is.numeric) %>% select_if(~max(., na.rm=TRUE) > 30) %>% names
```

```
##  [1] "ST_CASE"  "VEH_NO"   "PER_NO"   "COUNTY"   "DAY"      "HOUR"
##  [7] "MINUTE"   "AGE"      "YEAR"     "TRAV_SP"  "LATITUDE"
```

20

```r
c2015 %>% select_if(is.numeric) %>% summarize_if(~max(., na.rm=TRUE) > 30, ~mean(., na.rm = TRUE))
```

```
## # A tibble: 1 x 11
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 251487.   1.63   1.66   76.2  15.4  13.8   28.6  38.8  2015    44.5
## # ... with 1 more variable: LATITUDE <dbl>
```

21

```r
c2015 %>% select_if(is.numeric) %>% select_if(~max(., na.rm=TRUE) > 30) %>% summarize_all(~mean(., na.r
```

```
## # A tibble: 1 x 11
##   ST_CASE VEH_NO PER_NO COUNTY   DAY  HOUR MINUTE   AGE  YEAR TRAV_SP
##     <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1 251487.   1.63   1.66   76.2  15.4  13.8   28.6  38.8  2015    44.5
## # ... with 1 more variable: LATITUDE <dbl>
```

22

```
d1 <- c2015 %>% select_if(is.numeric) %>% select_if(~sd(., na.rm=TRUE) > 10)
```

23

```
b <- d1 %>% mutate_all(function(x, na.rm = FALSE) {x - mean(x, na.rm = TRUE)})
colMeans(b, na.rm=TRUE)
```

```
##       ST_CASE        COUNTY        MINUTE           AGE       TRAV_SP 
## -5.194126e-11  4.783075e-15  1.179032e-15 -1.565450e-15 -2.480700e-15 
##      LONGITUD 
##  2.127210e-15
```

24

```
c <- d1 %>% mutate_all(function(x, na.rm = FALSE) {(x - mean(x, na.rm = TRUE)) / sd(x, na.rm=TRUE)})

c %>% summarize_all(~mean(., na.rm=TRUE))
```

```
## # A tibble: 1 x 6
##     ST_CASE   COUNTY   MINUTE       AGE   TRAV_SP LONGITUD
##       <dbl>    <dbl>    <dbl>     <dbl>     <dbl>    <dbl>
## 1 -8.06e-17 5.05e-17 6.45e-17 -7.27e-17 -7.98e-17 1.40e-16
```

```
c %>% summarize_all(~sd(., na.rm=TRUE))
```

```
## # A tibble: 1 x 6
##   ST_CASE COUNTY MINUTE   AGE TRAV_SP LONGITUD
##     <dbl>  <dbl>  <dbl> <dbl>   <dbl>    <dbl>
## 1   1.000  1.000  1.000    1.   1.000       1.
```