

Assignment 4

Rachel Cardarelli

September 30, 2019

Question 1

```
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

#sin(2019)
x <- 2019
x %>% sin

## [1] 0.8644605

#sin(cos(2019))
x %>% cos %>% sin

## [1] -0.4817939

#sin(cos(tan(log(2019))))
x%>%log10%>%tan%>%cos%>%sin

## [1] 0.8340538

#log2(2019)
y <- 2
x%>% (function (x) log(x, base = y))

## [1] 10.97943
```

Question 2 - Redoing HW 2

```

library(readxl)
library(stringr)

c2015 <- read_excel("c2015.xlsx")
c2015 <-tbl_df(c2015)

c2015 <- c2015 %>%
  mutate(SEX = replace(SEX, SEX == "Unknown" | SEX == "Not Rep", "Female"))

c2015 <- c2015 %>%
  mutate(AGE = replace(AGE, AGE == "Less than 1", "0"), AGE = replace(AGE, AGE == "Unknown", "NA"), AGE = replace(AGE, AGE == "999", "NA"))

## Warning: NAs introduced by coercion

c2015 <- c2015 %>%
  mutate(TRAV_SP = str_replace(TRAV_SP, " MPH", ""), TRAV_SP = replace(TRAV_SP, TRAV_SP == "Unknown" | TRAV_SP == "999", "NA"))

## Warning: NAs introduced by coercion

```

Question 3

```

c2015 %>%
  filter(SEX == "Female", DAY_WEEK %in% c("Saturday", "Sunday")) %>%
  summarise(mean_sp = mean(TRAV_SP, na.rm = TRUE), mean_age = mean(AGE, na.rm = TRUE))

## # A tibble: 1 x 2
##   mean_sp mean_age
##     <dbl>    <dbl>
## 1     44.7     36.4

```

Question 4

```

numeric <- c2015 %>%
  select_if(is.numeric)

```

Question 5

```

c2015 %>%
  select_if(is.numeric) %>%
  summarise_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 275607.    1.39    1.63   91.7  15.5  14.0  28.4  39.1  2015    44.5
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>

```

Question 6

```
c2015 %>%
  summarise_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>
## 1 275607.   1.39   1.63  91.7 15.5 14.0  28.4 39.1 2015   44.5
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

Question 7

```
c2015 %>%
  summarise_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>
## 1 270282     1     1    71    15    15    29    37  2015     50
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

Question 8

```
c2015 %>%
  summarise_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>
## 1 163031.   1.45   1.84  95.0  8.78  9.06  17.3  20.1     0   25.1
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

Question 9

```
c2015 %>%
  summarise_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>
## 1        0      0      0      0      0      0    377      0      0   51420
## # ... with 2 more variables: LATITUDE <int>, LONGITUD <int>
```

Question 10

```
c2015 %>%
  summarise_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>
## 1     12.5  0.329  0.488  4.52  2.74  2.64  3.35  3.67  7.61  3.80
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

Question 11

```
c2015 %>%
  summarise_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>
## 1     12.5  0.329  0.488  4.52  2.74  2.64  3.35  3.67  7.61  3.80
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

Question 12

```
c2015 %>%
  summarise_if(is.character, ~sum(is.na(.)))  
  
## # 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>
## 1     0     0     0      0      0      0      0     7197  7197
## # ... with 7 more variables: MOD_YEAR <int>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, HARM_EV <int>, LGT_COND <int>,
## #   WEATHER <int>
```

Question 13

```
c2015 %>%
  summarise_all(~ifelse(is.character(.), sum(is.na(.)), 0))  
  
## # A tibble: 1 x 28
##   STATE ST_CASE VEH_NO PER_NO COUNTY  DAY MONTH  HOUR MINUTE AGE   SEX
##   <int>  <dbl>  <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
```

```

## 1      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0
## # ... with 17 more variables: PER_TYP <int>, INJ_SEV <int>,
## #   SEAT_POS <int>, DRINKING <int>, YEAR <dbl>, MAN_COLL <int>,
## #   OWNER <int>, MOD_YEAR <int>, TRAV_SP <dbl>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <int>, LGT_COND <int>, WEATHER <int>

```

Question 14

```

c2015 %>%
  summarise(length(table(STATE)))

```

```

## # A tibble: 1 x 1
##   `length(table(STATE))`<int>
## 1                               51

```

Question 15

```

c2015 %>%
  summarise_if(is.character, ~length(table(.)))

```

```

## # 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     2     11     8     29      4     11      8
## # ... with 7 more variables: MOD_YEAR <int>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, HARM_EV <int>, LGT_COND <int>,
## #   WEATHER <int>

```

Question 16

```

c2015 %>%
  summarise_all(~ifelse(is.character(.), length(table(.)), 0))

```

```

## # A tibble: 1 x 28
##   STATE ST_CASE VEH_NO PER_NO COUNTY  DAY MONTH  HOUR MINUTE  AGE  SEX
##   <int>    <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <int>
## 1      51        0      0      0      0      0     12      0      0      0      2
## # ... with 17 more variables: PER_TYP <int>, INJ_SEV <int>,
## #   SEAT_POS <int>, DRINKING <int>, YEAR <dbl>, MAN_COLL <int>,
## #   OWNER <int>, MOD_YEAR <int>, TRAV_SP <dbl>, DEFORMED <int>,
## #   DAY_WEEK <int>, ROUTE <int>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <int>, LGT_COND <int>, WEATHER <int>

```

Question 17

```
c2015 %>%
  summarise_if(~length(table(.)) > 30, ~length(table(.)))
```



```
## # A tibble: 1 x 13
##   STATE ST_CASE VEH_NO PER_NO COUNTY DAY MINUTE AGE MOD_YEAR TRAV_SP
##   <int>    <int>  <int>  <int> <int>  <int> <int>  <int>    <int>
## 1     51    32166     59     51    288     31     60    104      77     131
## # ... with 3 more variables: LATITUDE <int>, LONGITUD <int>, HARM_EV <int>
```

Question 18

```
c2015 %>%
  select_if(~is.character(.)) %>%
  summarise_if(~length(table(.)) > 30, ~length(table(.)))
```



```
## # A tibble: 1 x 3
##   STATE MOD_YEAR HARM_EV
##   <int>    <int>    <int>
## 1     51       77       50
```

Question 19

```
c2015 %>%
  select_if(~is.numeric(.)) %>%
  summarise_if(~length(table(.)) > 30, ~length(table(.)))
```



```
## # A tibble: 1 x 10
##   ST_CASE VEH_NO PER_NO COUNTY DAY MINUTE AGE TRAV_SP LATITUDE LONGITUD
##   <int>    <int>  <int>  <int> <int>  <int> <int>    <int>    <int>    <int>
## 1    32166     59     51    288     31     60    104     131    31818    31882
```

Question 20

```
c2015 %>%
  select_if(~is.numeric(.)) %>%
  summarise_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 275607.    1.39    1.63   91.7  15.5  14.0   28.4  39.1  2015    44.5
## # ... with 1 more variable: LATITUDE <dbl>
```

Question 21

```
c2015 %>%
  select_if(is.numeric) %>%
  summarise_all(~ifelse(max(.,na.rm = TRUE) > 30, mean(., na.rm = TRUE), 0))

## # 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>
## 1 275607.    1.39    1.63   91.7  15.5  14.0   28.4 39.1  2015    44.5
## # ... with 2 more variables: LATITUDE <dbl>, LONGITUD <dbl>
```

Question 22

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

Question 23

```
d <- d1 %>%
  mutate_all(~(. - mean(.,na.rm = TRUE)))
```

Question 24

```
d %>%
  mutate_all(~(. / sd(., na.rm = TRUE))) %>%
  summarise_all(~mean(., na.rm = TRUE), ~sd(., na.rm = TRUE))

## # A tibble: 1 x 6
##   ST_CASE COUNTY  MINUTE      AGE    TRAV_SP LONGITUD
##       <dbl>   <dbl>   <dbl>     <dbl>     <dbl>     <dbl>
## 1 -9.97e-17 1.15e-16 -6.85e-17 8.49e-17 -7.98e-17 -3.50e-16
```