DS202 lab 5

Andrew Fahmy

April 16, 2021

```
    acc <- read.csv("https://raw.githubusercontent.com/xdaiISU/ds202materials/master/hwlabs/fars2017/ac
    person <- read.csv("https://raw.githubusercontent.com/xdaiISU/ds202materials/master/hwlabs/fars2017
```

1. Are there some days of the week where more accidents happen than the others (see FARS manual, use variable DAY_WEEK)?

```
    library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

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

```
    names(acc)
```

```
##  [1] "STATE"      "ST_CASE"    "VE_TOTAL"   "VE_FORMS"   "PVH_INVL"
##  [6] "PEDS"       "PERNOTMVIT" "PERMVIT"    "PERSONS"    "COUNTY"
## [11] "CITY"       "DAY"        "MONTH"      "YEAR"       "DAY_WEEK"
## [16] "HOUR"       "MINUTE"     "NHS"        "RUR_URB"    "FUNC_SYS"
## [21] "RD_OWNER"   "ROUTE"      "TWAY_ID"    "TWAY_ID2"   "MILEPT"
## [26] "LATITUDE"   "LONGITUD"   "SP_JUR"     "HARM_EV"    "MAN_COLL"
## [31] "RELJCT1"    "RELJCT2"    "TYP_INT"    "WRK_ZONE"   "REL_ROAD"
## [36] "LGT_COND"   "WEATHER1"   "WEATHER2"   "WEATHER"    "SCH_BUS"
## [41] "RAIL"       "NOT_HOUR"   "NOT_MIN"    "ARR_HOUR"   "ARR_MIN"
## [46] "HOSP_HR"    "HOSP_MN"    "CF1"        "CF2"        "CF3"
## [51] "FATALS"     "DRUNK_DR"
```

```
    names(person)
```

```
##  [1] "STATE"      "ST_CASE"    "VE_FORMS"   "VEH_NO"     "PER_NO"
##  [6] "STR_VEH"    "COUNTY"     "DAY"        "MONTH"      "HOUR"
## [11] "MINUTE"     "RUR_URB"    "FUNC_SYS"   "HARM_EV"    "MAN_COLL"
## [16] "SCH_BUS"    "MAKE"       "MAK_MOD"    "BODY_TYP"   "MOD_YEAR"
## [21] "TOW_VEH"    "SPEC_USE"   "EMER_USE"   "ROLLOVER"   "IMPACT1"
## [26] "FIRE_EXP"   "AGE"        "SEX"        "PER_TYP"    "INJ_SEV"
## [31] "SEAT_POS"   "REST_USE"   "REST_MIS"   "AIR_BAG"    "EJECTION"
## [36] "EJ_PATH"    "EXTRICAT"   "DRINKING"   "ALC_DET"    "ALC_STATUS"
```

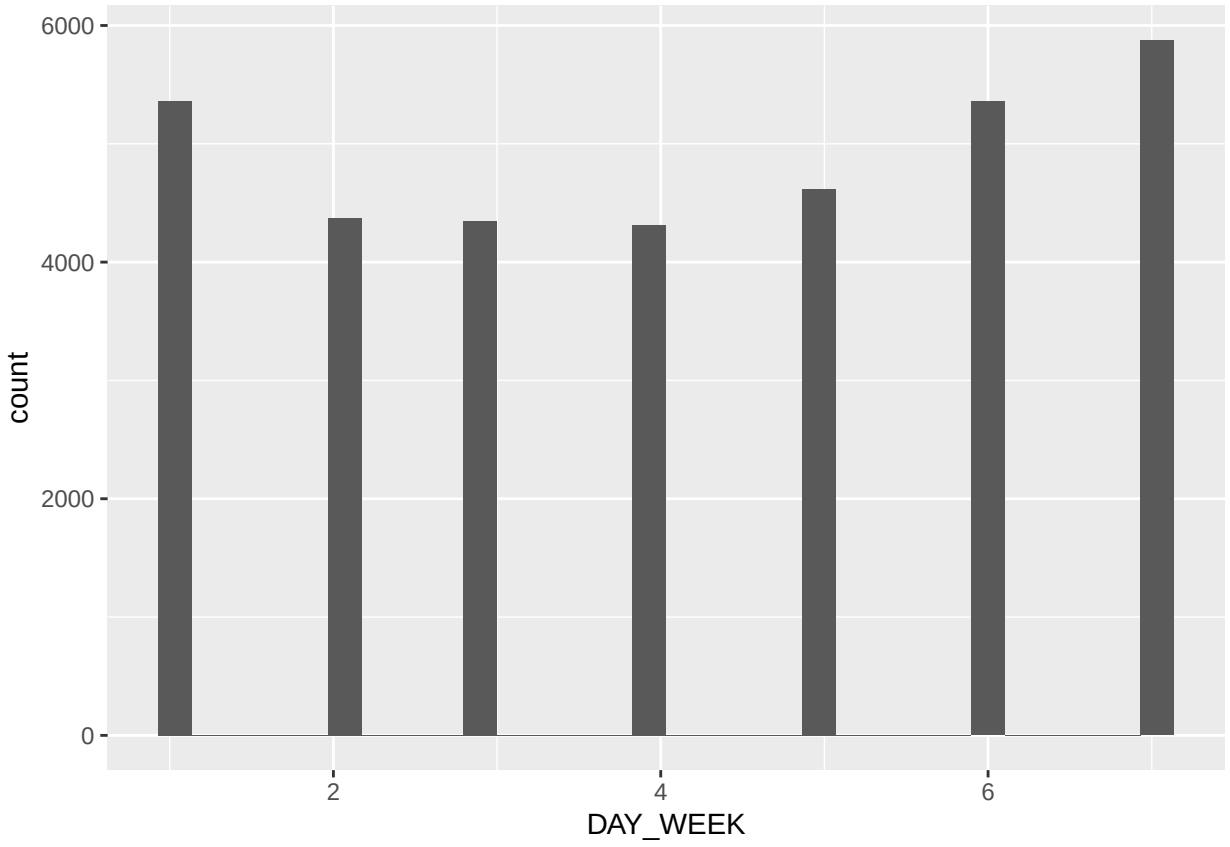
```
## [41] "ATST_TYP"   "ALC_RES"    "DRUGS"      "DRUG_DET"   "DSTATUS"
## [46] "DRUGTST1"   "DRUGTST2"   "DRUGTST3"   "DRUGRES1"   "DRUGRES2"
## [51] "DRUGRES3"   "HOSPITAL"   "DOA"        "DEATH_DA"   "DEATH_MO"
## [56] "DEATH_YR"   "DEATH_HR"   "DEATH_MN"   "DEATH_TM"   "LAG_HRS"
## [61] "LAG_MINS"   "P_SF1"      "P_SF2"      "P_SF3"      "WORK_INJ"
## [66] "HISPANIC"   "RACE"       "LOCATION"
```

```
    acc %>% group_by(DAY_WEEK) %>% summarise(n = n())
```

```
## # A tibble: 7 x 2
##   DAY_WEEK     n
##      <int> <int>
## 1        1  5360
## 2        2  4374
## 3        3  4347
## 4        4  4314
## 5        5  4621
## 6        6  5358
## 7        7  5873
```

```
    acc %>% ggplot(aes(x = DAY_WEEK)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. Create a data frame containing the persons who are fatally hurt in the accidents (look up variable INJ_SEV)

```
fatally_hurt <- person %>% filter(INJ_SEV == 4)
#names(fatally_hurt)
#head(fatally_hurt)
```

3. Create a data frame containing the most dangerous vehicle make in each state. The number of persons fatally hit in the vehicle make is used to assess the (non-)safety of a make. Make sure to handle the missing values appropriately. (look up variable MAKE)

```
library(readxl)
glcs_us <- read_xlsx("GLCs_US.xlsx")
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J1216 / R1216C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J1574 / R1574C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J3583 / R3583C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J3749 / R3749C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J3947 / R3947C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J4211 / R4211C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J6308 / R6308C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J15239 / R15239C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J22086 / R22086C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J24679 / R24679C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J26724 / R26724C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J32616 / R32616C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J36166 / R36166C10: got a date

## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting logical in J36432 / R36432C10: got a date
```

```
    names(glcs_us)
```

```
##  [1] "Territory"         "State Name"        "State Code"
##  [4] "City Code"         "City Name"         "County Code"
##  [7] "County Name"       "Country Code"      "Old City Name"
## [10] "Date Record Added" "Duty Station Code"
```

```
    dangerous_vehicle <- fatally_hurt %>%
                             mutate(MAKE = ifelse(is.na(MAKE), 99, MAKE)) %>%
                             select(STATE, MAKE) %>%
                             group_by(STATE) %>%
                             count(MAKE) %>%
                             top_n(1)
```

```
## Selecting by n
```

```
    #head(dangerous_vehicle)
```

Accourding to the FARS manual, 99 is unknown make. We replace NA with 99 to get rid of NA for the MAKE

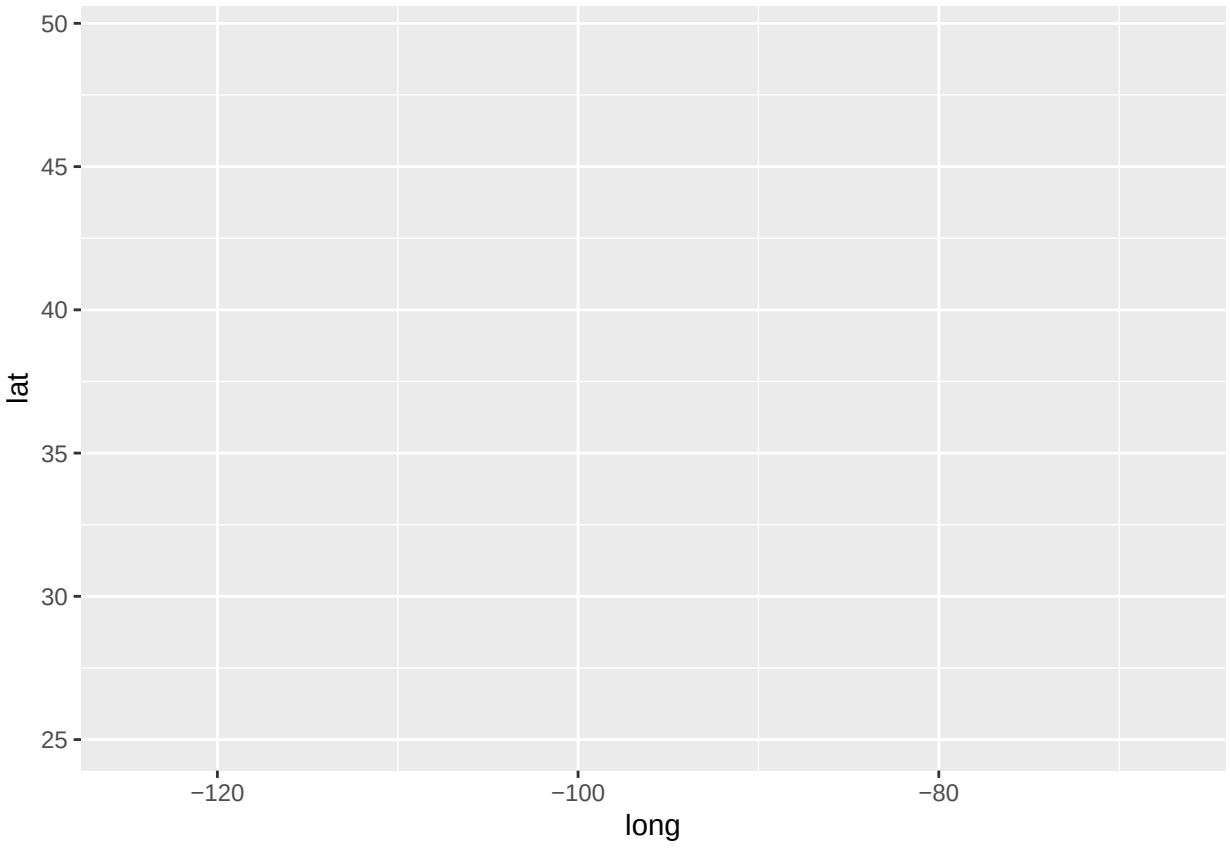
4. Create a map, and label each state with the most dangerous vehicle. Discuss the definition of the most dangerous vehicle, and what you find from the map. (Hint: Read the description for the STATE and COUNTY columns in the FARS manual. The state & county codes are Geographic Locator Codes (GLCs) from the General Services Administration's (GSA) publication. Use readxl::read_xlsx to read in the GLCs.)

```
    #install.packages("maps")
    library(maps)
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
    states <- map_data("state")
    states %>% ggplot(aes(x = long, y = lat, group = group, fill = region))
```



5. Join the accident and person table (work out which variable(s) to use)

```
    dat <- acc %>% inner_join(person, on = ST_CASE)
```

```
## Joining, by = c("STATE", "ST_CASE", "VE_FORMS", "COUNTY", "DAY", "MONTH", "HOUR", "MINUTE", "RUR_URB
```

```
    dim(dat)
```

```
## [1] 84921   107
```

```
    head(dat)
```

```
##   STATE ST_CASE VE_TOTAL VE_FORMS PVH_INVL PEDS PERNOTMVIT PERMVIT PERSONS
## 1     1   10001        1        1        0    0          0       1       1
## 2     1   10002        1        1        0    0          0       1       1
## 3     1   10003        3        3        0    0          0       3       3
## 4     1   10003        3        3        0    0          0       3       3
## 5     1   10003        3        3        0    0          0       3       3
## 6     1   10004        1        1        0    0          0       1       1
##   COUNTY CITY DAY MONTH YEAR DAY_WEEK HOUR MINUTE NHS RUR_URB FUNC_SYS RD_OWNER
## 1     73  330  19     2 2017        1   23     35   1       2        1        1
## 2     89 1730  14     2 2017        3   14     59   1       2        1        1
## 3    101 2130  31     1 2017        3   20     31   1       2        1        1
## 4    101 2130  31     1 2017        3   20     31   1       2        1        1
## 5    101 2130  31     1 2017        3   20     31   1       2        1        1
## 6     73  350   1     1 2017        1   16     55   0       2        4        4
```

```
##   ROUTE        TWAY_ID       TWAY_ID2 MILEPT LATITUDE  LONGITUD SP_JUR HARM_EV
## 1     1          I-459                   10 33.33566 -87.00709      0      38
## 2     1          I-565                   70 34.66153 -86.78685      0       1
## 3     1           I-85 CHANTILLY PKWY    100 32.36652 -86.14528      0      12
## 4     1           I-85 CHANTILLY PKWY    100 32.36652 -86.14528      0      12
## 5     1           I-85 CHANTILLY PKWY    100 32.36652 -86.14528      0      12
## 6     6 20TH ST ENSLEY          AVE I      0 33.51017 -86.89400      0      30
##   MAN_COLL RELJCT1 RELJCT2 TYP_INT WRK_ZONE REL_ROAD LGT_COND WEATHER1 WEATHER2
## 1        0       0       1       1        0        3        2        1        0
## 2        0       0       1       1        0        3        1        1        0
## 3        1       0       1       1        0        1        2        1        0
## 4        1       0       1       1        0        1        2        1        0
## 5        1       0       1       1        0        1        2        1        0
## 6        0       0       3       2        0        4        3        2        0
##   WEATHER SCH_BUS    RAIL NOT_HOUR NOT_MIN ARR_HOUR ARR_MIN HOSP_HR HOSP_MN CF1
## 1       1       0 725409J       99      99       99      99      88      88   0
## 2       1       0 0000000       15       0       15       9      88      88   0
## 3       1       0 0000000       99      99       99      99      88      88   0
## 4       1       0 0000000       99      99       99      99      88      88   0
## 5       1       0 0000000       99      99       99      99      88      88   0
## 6       2       0 0000000       99      99       16      58      88      88  20
##   CF2 CF3 FATALS DRUNK_DR VEH_NO PER_NO STR_VEH MAKE MAK_MOD BODY_TYP MOD_YEAR
## 1   0   0      1        0      1      1       0   20   20421       15     2004
## 2   0   0      1        0      1      1       0   37   37402       14     2005
## 3   0   0      1        0      1      1       0   82   82881       66     2007
## 4   0   0      1        0      2      1       0    2    2404       14     2003
## 5   0   0      1        0      3      1       0   84   84884       66     2015
## 6   0   0      1        0      1      1       0   30   30046        4     2014
##   TOW_VEH SPEC_USE EMER_USE ROLLOVER IMPACT1 FIRE_EXP AGE SEX PER_TYP INJ_SEV
## 1       1        0        0        0      12        0  42   1       1       4
## 2       0        0        0        9       0        0  43   1       1       4
## 3       1        0        0        0      12        1  63   1       1       0
## 4       0        0        0        0       6        0  47   1       1       4
## 5       1        0        0        0       6        0  64   1       1       0
## 6       0        0        0        1      11        0  18   1       1       4
##   SEAT_POS REST_USE REST_MIS AIR_BAG EJECTION EJ_PATH EXTRICAT DRINKING ALC_DET
## 1       11       20        0       1        1       9        0        0       9
## 2       11        3        0       8        0       0        0        9       9
## 3       11        3        0      20        0       0        0        0       9
## 4       11        3        0      99        0       0        0        0       9
## 5       11        3        0      20        0       0        0        0       9
## 6       11        3        0       8        0       0        0        9       9
##   ALC_STATUS ATST_TYP ALC_RES DRUGS DRUG_DET DSTATUS DRUGTST1 DRUGTST2 DRUGTST3
## 1          0        0     996     0        8       0        0        0        0
## 2          2        1       0     9        8       2        1        1        1
## 3          2       10       0     0        8       0        0        0        0
## 4          2        1       0     0        8       2        1        0        0
## 5          0        0     996     0        8       0        0        0        0
## 6          0        0     996     9        8       0        0        0        0
##   DRUGRES1 DRUGRES2 DRUGRES3 HOSPITAL DOA DEATH_DA DEATH_MO DEATH_YR DEATH_HR
## 1        0        0        0        0   7       19        2     2017       23
## 2      605      600      996        0   7       14        2     2017       14
## 3        0        0        0        0   0       88       88     8888       88
## 4        1        0        0        0   7       31        1     2017       20
```

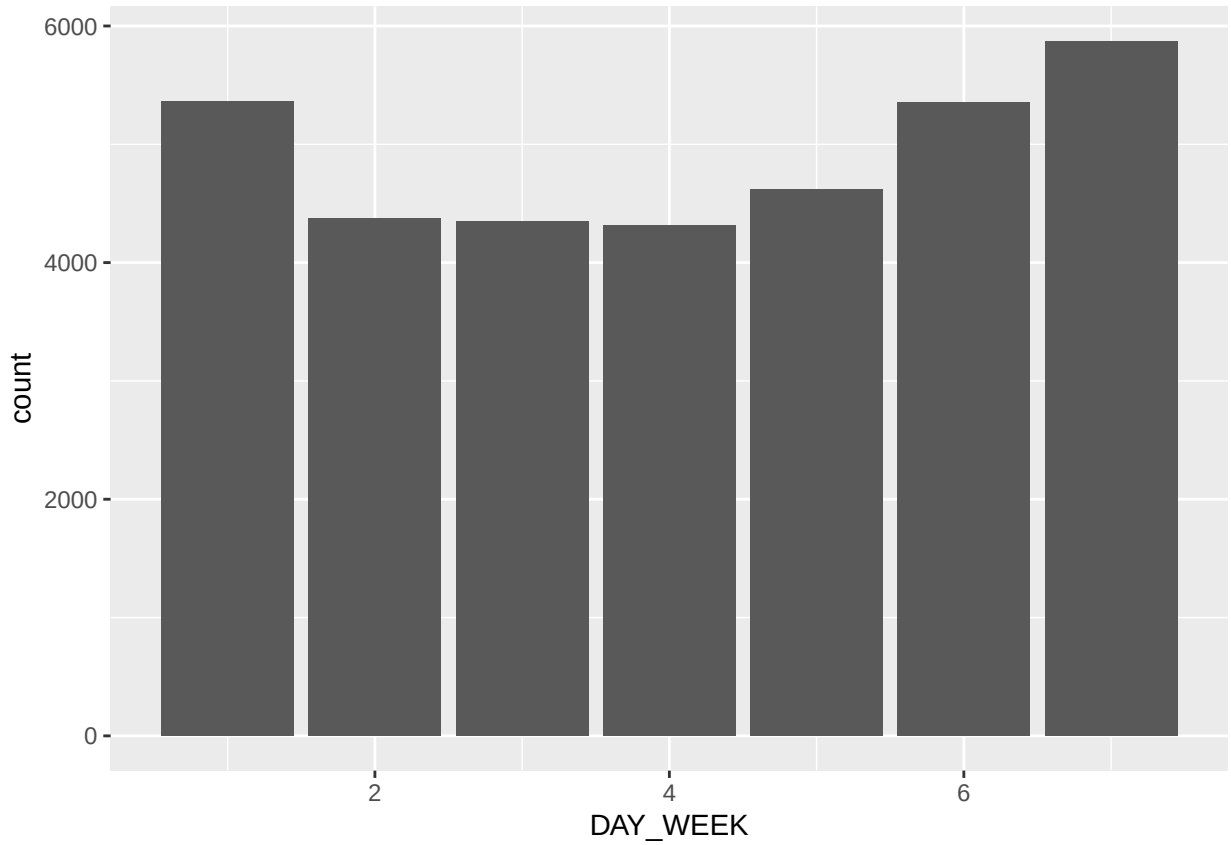
```
## 5        0        0        0        0   0       88       88     8888       88
## 6        0        0        0        0   7        1        1     2017       16
##   DEATH_MN DEATH_TM LAG_HRS LAG_MINS P_SF1 P_SF2 P_SF3 WORK_INJ HISPANIC RACE
## 1       35     2335       0        0     0     0     0        0        7    1
## 2       59     1459       0        0     0     0     0        0        7    1
## 3       88     8888     999       99     0     0     0        8        0    0
## 4       31     2031       0        0     0     0     0        0        7    2
## 5       88     8888     999       99     0     0     0        8        0    0
## 6       55     1655       0        0     0     0     0        0        7    2
##   LOCATION
## 1        0
## 2        0
## 3        0
## 4        0
## 5        0
## 6        0
```

6. Tally the number of accidents by day of the week (DAY_WEEK), hour of the day (HOUR) and gender (SEX). Visualize the results and explain what you find.

```
    # Day of week
    acc %>% group_by(DAY_WEEK) %>% summarise(n = n())
```

```
## # A tibble: 7 x 2
##   DAY_WEEK     n
##      <int> <int>
## 1        1  5360
## 2        2  4374
## 3        3  4347
## 4        4  4314
## 5        5  4621
## 6        6  5358
## 7        7  5873
```

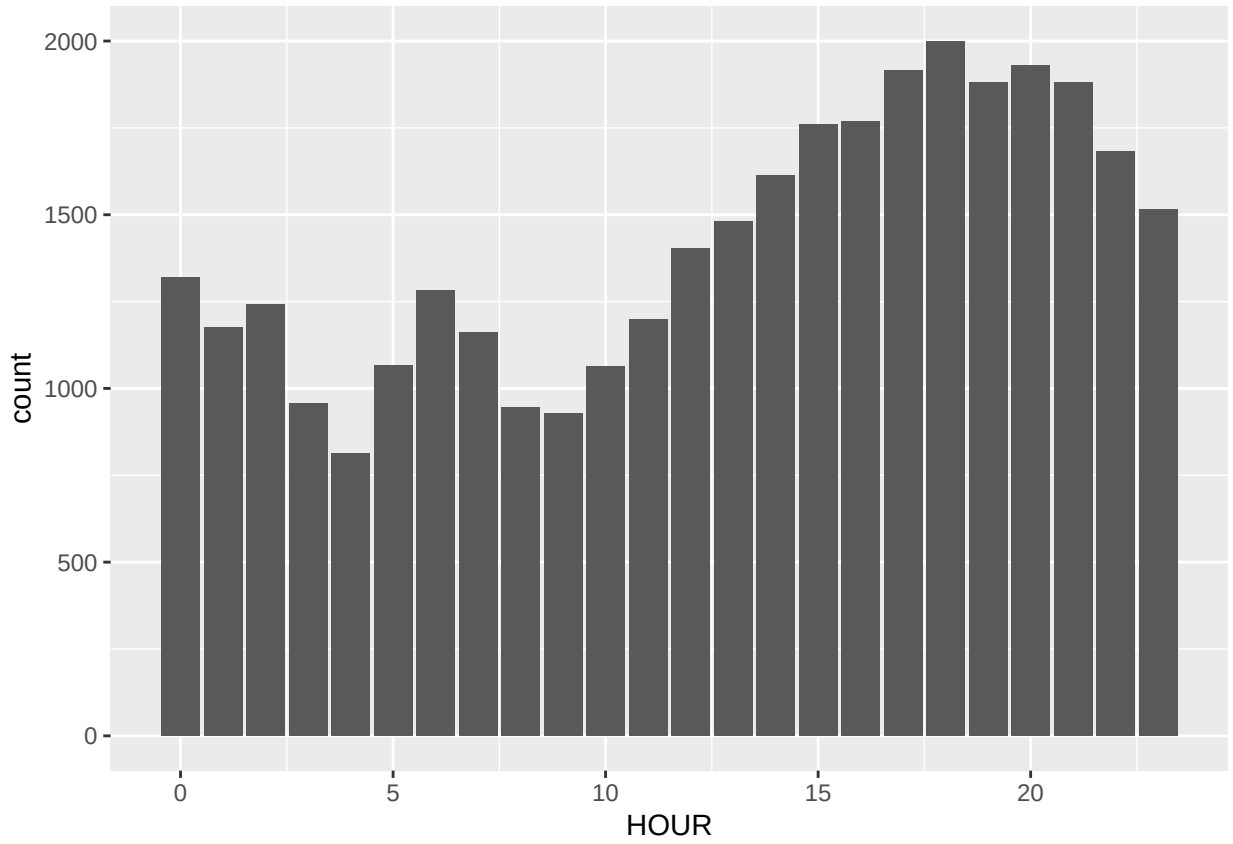
```
    acc %>% ggplot(aes(x = DAY_WEEK)) + geom_bar()
```




```
    # Hour of day
    acc %>% group_by(HOUR) %>% summarise(n = n())
```

```
## # A tibble: 25 x 2
##     HOUR     n
##    <int> <int>
##  1     0  1321
##  2     1  1177
##  3     2  1241
##  4     3   957
##  5     4   813
##  6     5  1065
##  7     6  1282
##  8     7  1162
##  9     8   945
## 10     9   929
## # ... with 15 more rows
```

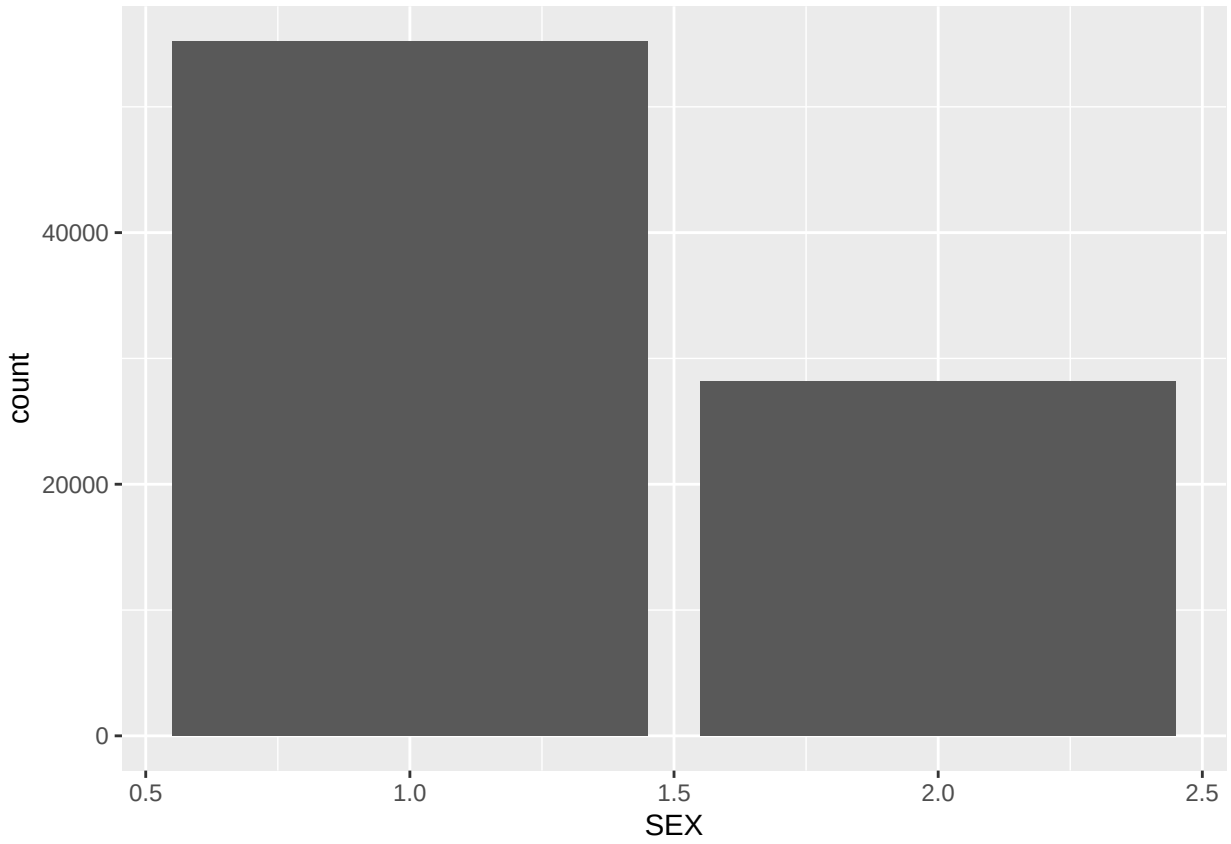
```
    acc %>% filter(HOUR != 99) %>% ggplot(aes(x = HOUR)) + geom_bar()
```




```
    # Sex
    dat %>% group_by(SEX) %>% summarise(n = n())
```

```
## # A tibble: 4 x 2
##     SEX     n
##   <int> <int>
## 1     1 55230
## 2     2 28149
## 3     8   473
## 4     9  1069
```

```
    dat %>% filter(SEX <= 2) %>% ggplot(aes(x = SEX)) + geom_bar()
```



7. Now plot a choropleth map of the number of deaths on a county level. Also explain what you find.

8. Is summer or winter more dangerous? Does this depend on states? Explore and explain.

```
    acc %>% group_by(MONTH) %>% summarise(n = n())
```

```
## # A tibble: 12 x 2
##    MONTH     n
##    <int> <int>
##  1     1  2616
##  2     2  2302
##  3     3  2686
##  4     4  2743
##  5     5  2896
##  6     6  3015
##  7     7  3226
##  8     8  2964
##  9     9  3068
## 10    10  3064
## 11    11  2852
## 12    12  2815
```

```
acc %>% ggplot(aes(x = MONTH)) + geom_bar()
```