Homework 4

Akhil Havaldar (ash2sfp)

Question 1

a)

```
library(tidyverse)
dat <- read.csv("fatal accidents.csv")
head(dat)
```

```
##                  State Case.number Vehicle.count People.count.IN
## 1 District of Columbia      110001             1               1
## 2 District of Columbia      110002             1               1
## 3 District of Columbia      110003             1               1
## 4 District of Columbia      110004             2               2
## 5 District of Columbia      110005             1               2
## 6 District of Columbia      110006             4               7
##   People.count.OUT Day Month Year Day.of.week Hour Minute
## 1                1  11     2 2019           2   23     34
## 2                1  20     2 2019           4   18     25
## 3                1   5     3 2019           3   21      1
## 4                0  13     5 2019           2    5     19
## 5                0   4     8 2019           1    4      7
## 6                0   5     4 2019           6    2     45
```

b)

```
vars <- unique(dat$State)
state.list <- lapply(vars, function(x){subset(dat, State == x)})
names(state.list) <- vars
```

c)

```
lapply(state.list,head,n=3)
```

```
## $`District of Columbia`
##                  State Case.number Vehicle.count People.count.IN
## 1 District of Columbia      110001             1               1
## 2 District of Columbia      110002             1               1
## 3 District of Columbia      110003             1               1
##   People.count.OUT Day Month Year Day.of.week Hour Minute
## 1                1  11     2 2019           2   23     34
## 2                1  20     2 2019           4   18     25
## 3                1   5     3 2019           3   21      1
##
## $Maryland
##       State Case.number Vehicle.count People.count.IN People.count.OUT Day
## 23 Maryland      240001             2               3                1   7
## 24 Maryland      240002             3               3                0   3
## 25 Maryland      240003             2               4                1   6
##    Month Year Day.of.week Hour Minute
## 23     1 2019           2    5     55
## 24     1 2019           5    6     43
## 25     1 2019           1   15     30
##
## $`North Carolina`
##              State Case.number Vehicle.count People.count.IN People.count.OUT
## 507 North Carolina      370001             1               1                0
## 508 North Carolina      370002             2               2                0
## 509 North Carolina      370003             2               2                0
##     Day Month Year Day.of.week Hour Minute
## 507   5     1 2019           7   23     47
## 508  17     1 2019           5    6     44
## 509  17     1 2019           5   14     54
##
## $Virginia
##         State Case.number Vehicle.count People.count.IN People.count.OUT Day
## 1791 Virginia      510001             1               1                1   1
## 1792 Virginia      510002             2               2                0   2
## 1793 Virginia      510003             1               2                0   3
##      Month Year Day.of.week Hour Minute
## 1791     1 2019           3    5     48
## 1792     1 2019           4   15     35
## 1793     1 2019           5   15      5
##
## $`West Virginia`
##              State Case.number Vehicle.count People.count.IN People.count.OUT
```

```
## 2565 West Virginia      540001             1                4                0
## 2566 West Virginia      540002             2                2                0
## 2567 West Virginia      540003             1                1                0
##      Day Month Year Day.of.week Hour Minute
## 2565   2     1 2019           4   20     30
## 2566   2     1 2019           4    6      8
## 2567   9     1 2019           4   23     36
```

d)

```
func <- function(x){
  y <- x %>%
  group_by(Day.of.week) %>%
  summarise(cnt = n()) %>%
  mutate(freq = (round(100*(cnt / sum(cnt)), 1)))
}

dotw <- lapply(state.list,func)
lapply(dotw,head,n=7)
```

```
## $`District of Columbia`
## # A tibble: 7 x 3
##   Day.of.week   cnt  freq
##         <int> <int> <dbl>
## 1           1     3  13.6
## 2           2     3  13.6
## 3           3     5  22.7
## 4           4     3  13.6
## 5           5     1   4.5
## 6           6     6  27.3
## 7           7     1   4.5
##
## $Maryland
## # A tibble: 7 x 3
##   Day.of.week   cnt  freq
##         <int> <int> <dbl>
## 1           1    82  16.9
## 2           2    65  13.4
## 3           3    72  14.9
## 4           4    52  10.7
## 5           5    59  12.2
## 6           6    70  14.5
## 7           7    84  17.4
```

```
##
## $`North Carolina`
## # A tibble: 7 x 3
##   Day.of.week   cnt  freq
##         <int> <int> <dbl>
## 1           1   186  14.5
## 2           2   157  12.2
## 3           3   169  13.2
## 4           4   172  13.4
## 5           5   170  13.2
## 6           6   207  16.1
## 7           7   223  17.4
##
## $Virginia
## # A tibble: 7 x 3
##   Day.of.week   cnt  freq
##         <int> <int> <dbl>
## 1           1   119  15.4
## 2           2    95  12.3
## 3           3   104  13.4
## 4           4   102  13.2
## 5           5   106  13.7
## 6           6   127  16.4
## 7           7   121  15.6
##
## $`West Virginia`
## # A tibble: 7 x 3
##   Day.of.week   cnt  freq
##         <int> <int> <dbl>
## 1           1    30  12.1
## 2           2    36  14.6
## 3           3    36  14.6
## 4           4    34  13.8
## 5           5    32  13
## 6           6    32  13
## 7           7    47  19
```

e)

- From the tables created in part d, we can see that in DC the majority of accidents happen on Tuesday and Friday, with the least being Thursday and Saturday. In Maryland, North Carolina and West Virginia, the most crashes are on Saturday, but are relatively evenly spread out. Virginia's crashes are the most evenly spread out among the days with Friday having the slight edge in most crashes.

f)

```
# rows are dot, cols are number of vehicles in crash

func2 <- function(x){
  y <- with(x, table(Day.of.week,Vehicle.count))
}

vehicle.ct <- lapply(state.list, func2)
lapply(vehicle.ct, head, n=7)
```

```
## $`District of Columbia`
##             Vehicle.count
## Day.of.week 1 2 3 4
##           1 2 1 0 0
##           2 1 2 0 0
##           3 3 1 1 0
##           4 2 0 1 0
##           5 1 0 0 0
##           6 4 1 0 1
##           7 1 0 0 0
##
## $Maryland
##             Vehicle.count
## Day.of.week  1  2  3  4  5  7 12
##           1 50 22  8  1  1  0  0
##           2 29 30  4  0  1  0  1
##           3 40 24  5  2  1  0  0
##           4 34 14  3  0  0  1  0
##           5 30 17 11  1  0  0  0
##           6 31 27 10  1  1  0  0
##           7 50 26  5  2  1  0  0
##
## $`North Carolina`
##             Vehicle.count
## Day.of.week   1   2   3   4   5   7
##           1 112  60  11   0   2   1
##           2  78  68   8   2   1   0
##           3  82  76  10   1   0   0
##           4 104  56   8   3   0   1
##           5  86  67  10   6   1   0
##           6 120  70  14   2   1   0
##           7 142  68  12   0   1   0
```

```
##
## $Virginia
##             Vehicle.count
## Day.of.week  1  2  3  4  5  6  8
##           1 81 30  8  0  0  0  0
##           2 62 24  5  3  1  0  0
##           3 56 37  7  3  0  1  0
##           4 59 38  4  1  0  0  0
##           5 58 40  5  0  2  1  0
##           6 79 36  7  3  1  1  0
##           7 81 32  6  1  0  0  1
##
## $`West Virginia`
##             Vehicle.count
## Day.of.week  1  2  3  4  5
##           1 21  8  0  1  0
##           2 22 13  1  0  0
##           3 21 14  1  0  0
##           4 15 15  3  1  0
##           5 19 10  1  1  1
##           6 24  5  2  1  0
##           7 25 19  2  0  1
```

g)

- Instead of the displaying the counts in a table, we can display them in a dataframe which would make it easier to subset individual observations based on a certain criterion.

Question 2

a)

```
dat$People.count <- dat$People.count.IN + dat$People.count.OUT
head(dat)
```

```
##                  State Case.number Vehicle.count People.count.IN
## 1 District of Columbia      110001             1               1
## 2 District of Columbia      110002             1               1
## 3 District of Columbia      110003             1               1
## 4 District of Columbia      110004             2               2
## 5 District of Columbia      110005             1               2
```

```
## 6 District of Columbia      110006             4               7
##   People.count.OUT Day Month Year Day.of.week Hour Minute People.count
## 1                1  11     2 2019           2   23     34            2
## 2                1  20     2 2019           4   18     25            2
## 3                1   5     3 2019           3   21      1            2
## 4                0  13     5 2019           2    5     19            2
## 5                0   4     8 2019           1    4      7            2
## 6                0   5     4 2019           6    2     45            7
```

b)

```
b <- group_by(dat, State) %>%
  summarize(avg.vehicles=mean(Vehicle.count), avg.ppl=mean(People.count))
b
```

```
## # A tibble: 5 x 3
##   State                avg.vehicles avg.ppl
##   <chr>                       <dbl>   <dbl>
## 1 District of Columbia         1.55    2.95
## 2 Maryland                     1.64    2.59
## 3 North Carolina               1.54    2.34
## 4 Virginia                     1.51    2.28
## 5 West Virginia                1.50    2.38
```

c)

```
c <- group_by(dat, State) %>%
  summarize(avg.vehicles=mean(Vehicle.count), min.vehicles=min(Vehicle.count), max.vehic
c
```

```
## # A tibble: 5 x 4
##   State                avg.vehicles min.vehicles max.vehicles
##   <chr>                       <dbl>        <int>        <int>
## 1 District of Columbia         1.55            1            4
## 2 Maryland                     1.64            1           12
## 3 North Carolina               1.54            1            7
## 4 Virginia                     1.51            1            8
## 5 West Virginia                1.50            1            5
```

d)

- From part b we can see that the average number of vehicles in a crash is very similar across the different states. This is also the case with number of people in the crash, with DC having 0.5 more than the next highest state on average. From part d, we can see that all states have a minimum of 1 vehicle per crash, but the maximums vary by state. Maryland has the most with 12, and DC has the least with 4.

e)

```
e <- subset(dat, State = "Virginia") %>%
  group_by(Month) %>%
  summarise(cnt = n())
e
```

```
## # A tibble: 12 x 2
##    Month   cnt
##    <int> <int>
##  1     1   205
##  2     2   195
##  3     3   218
##  4     4   224
##  5     5   276
##  6     6   256
##  7     7   227
##  8     8   250
##  9     9   290
## 10    10   272
## 11    11   248
## 12    12   150
```

f)

```
f <- subset(dat, State = "Virginia")
f2 <- filter(f, Month == 6 |Month == 7 | Month== 8) %>%
  group_by(Day.of.week) %>%
  summarise(mean = mean(Vehicle.count), median = median(Vehicle.count))
f2
```

```
## # A tibble: 7 x 3
##   Day.of.week  mean median
```

```
##         <int> <dbl>  <dbl>
## 1           1  1.57      1
## 2           2  1.56      1
## 3           3  1.57      2
## 4           4  1.42      1
## 5           5  1.53      1
## 6           6  1.53      1
## 7           7  1.46      1
```

g)

- From part e, we can see the most accidents in a month for 2019 in Virginia is September (290). On the contrary, the lowest number of accidents occurred in December. From part f, based on only the summer months, Tuesday had the greatest median number of number of vehicles in a crash, while the rest of the days were equal at 1. The mean number of vehicles involved in accidents is roughly even across all the days of the week, with Wednesday and Saturday having slightly lower means.