Homework 4

Noah McIntire

Problem 1

a)

```
fatal<-read.csv("/Users/noahmcintire/Desktop/STAT 3080/fatal accidents.csv")
head(fatal)
```

```
##                   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)

```
datasub<-function(x){
  x<-fatal[fatal$State==x,]
  x
}
state.list<-lapply(unique(fatal$State),datasub)
```

c)

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

```
## [[1]]
##                   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
##
## [[2]]
##       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
##
## [[3]]
##              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
##
## [[4]]
##         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
##
## [[5]]
```

```
##              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)

```
newtable <- function(x){
  x1<-group_by(x, Day.of.week)
  accident<- count(x1, name="num.acc")
  sum.1<- sum(accident$num.acc)
  perc<- summarize(accident, percentage.of.accidents= round((100*sum(num.acc)/sum.1), 1)
  perc
}
lapply(state.list, newtable)
```

```
## [[1]]
## # A tibble: 7 x 2
##   Day.of.week percentage.of.accidents
##         <int>                   <dbl>
## 1           1                    13.6
## 2           2                    13.6
## 3           3                    22.7
## 4           4                    13.6
## 5           5                     4.5
## 6           6                    27.3
## 7           7                     4.5
##
## [[2]]
## # A tibble: 7 x 2
##   Day.of.week percentage.of.accidents
##         <int>                   <dbl>
## 1           1                    16.9
## 2           2                    13.4
## 3           3                    14.9
## 4           4                    10.7
## 5           5                    12.2
## 6           6                    14.5
## 7           7                    17.4
##
```

```
## [[3]]
## # A tibble: 7 x 2
##   Day.of.week percentage.of.accidents
##         <int>                   <dbl>
## 1           1                    14.5
## 2           2                    12.2
## 3           3                    13.2
## 4           4                    13.4
## 5           5                    13.2
## 6           6                    16.1
## 7           7                    17.4
##
## [[4]]
## # A tibble: 7 x 2
##   Day.of.week percentage.of.accidents
##         <int>                   <dbl>
## 1           1                    15.4
## 2           2                    12.3
## 3           3                    13.4
## 4           4                    13.2
## 5           5                    13.7
## 6           6                    16.4
## 7           7                    15.6
##
## [[5]]
## # A tibble: 7 x 2
##   Day.of.week percentage.of.accidents
##         <int>                   <dbl>
## 1           1                    12.1
## 2           2                    14.6
## 3           3                    14.6
## 4           4                    13.8
## 5           5                    13
## 6           6                    13
## 7           7                    19
```

e)

By viewing each state, we can see that in most cases, there is a higher percentage of accidents that occur on Fridays, Saturdays, and Sundays.

f)

```
newtable1 <- function(x){
  x1<-group_by(x, Day.of.week)
  accident<- count(x1, name="Total Number of Accidents")
  accident
  x2<-summarize(x1, Total.Vehicle= sum(Vehicle.count))
  x3<-merge(accident, x2)
  x3
  # merge() comes from reference 3
}
lapply(state.list,newtable1)
```

```
## [[1]]
##   Day.of.week Total Number of Accidents Total.Vehicle
## 1           1                         3             4
## 2           2                         3             5
## 3           3                         5             8
## 4           4                         3             5
## 5           5                         1             1
## 6           6                         6            10
## 7           7                         1             1
##
## [[2]]
##   Day.of.week Total Number of Accidents Total.Vehicle
## 1           1                        82           127
## 2           2                        65           118
## 3           3                        72           116
## 4           4                        52            78
## 5           5                        59           101
## 6           6                        70           124
## 7           7                        84           130
##
## [[3]]
##   Day.of.week Total Number of Accidents Total.Vehicle
## 1           1                       186           282
## 2           2                       157           251
## 3           3                       169           268
## 4           4                       172           259
## 5           5                       170           279
## 6           6                       207           315
## 7           7                       223           319
##
## [[4]]
##   Day.of.week Total Number of Accidents Total.Vehicle
```

```
## 1           1                        119           165
## 2           2                         95           142
## 3           3                        104           169
## 4           4                        102           151
## 5           5                        106           169
## 6           6                        127           195
## 7           7                        121           175
##
## [[5]]
##   Day.of.week Total Number of Accidents Total.Vehicle
## 1           1                        30            41
## 2           2                        36            51
## 3           3                        36            52
## 4           4                        34            58
## 5           5                        32            51
## 6           6                        32            44
## 7           7                        47            74
```

g)

Being able to include all the states in one table, in which the state is used as a categorical variable, would allow the viewer to more easily assertain observations about the data.

Problem 2

a)

```
People.count<- fatal$People.count.IN + fatal$People.count.OUT
fatal1<-mutate(fatal,People.count)
head(fatal1,3)
```

```
##                  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 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
```

```
# mutate comes from reference 1 (dplyr)
```

b)

```
fatal1<-filter(fatal1, Year == 2019)
fatal2<-group_by(fatal1, State)
sum1<-summarise(fatal2, avgVehicle=mean(Vehicle.count), avgPeople=mean(People.count))
sum1
```

```
## # A tibble: 5 x 3
##   State                avgVehicle avgPeople
##   <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)

```
fatal3<-group_by(fatal1,State)

sum2<-summarise(fatal3, minVehicle=min(Vehicle.count), avgVehicle=mean(Vehicle.count), m
sum2
```

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

d)

In all 5 states, the average number of vehicles involved in an accident was between one and two vehicles, and the average number of people in a crash was between 2 and three people. Additionally, each state has had a single-vehicle accident, while some states have had a higher max number of vehicles involved in a crash (see DC).

e)

```
fatal4<- filter(fatal1, State == "Virginia")
fatal4<- group_by(fatal4, Month)
sum3<- count(fatal4, name="Total Number of Accidents")
sum3
```

```
## # A tibble: 12 x 2
## # Groups:   Month [12]
##    Month `Total Number of Accidents`
##    <int>                       <int>
##  1     1                          63
##  2     2                          55
##  3     3                          57
##  4     4                          60
##  5     5                          66
##  6     6                          62
##  7     7                          55
##  8     8                          69
##  9     9                          79
## 10    10                          78
## 11    11                          70
## 12    12                          60
```

```
# count() comes from reference 2 (dplyr)
```

f)

```
## fatal4 is already subsetted to Virginia and grouped by months
fatal5<-filter(fatal4, Month>5 & Month<9)
sum4<- summarize(fatal5, mean.Vehicles=mean(Vehicle.count), median.Vehicles=median(Vehic
sum4
```

```
## # A tibble: 3 x 3
##   Month mean.Vehicles median.Vehicles
##   <int>         <dbl>           <dbl>
## 1     6          1.58               1
## 2     7          1.47               1
## 3     8          1.57               1
```

g)

Based of the table in e, we can see the total number of accidents per month stay somewhat consistent, but the number of accidents do see to trend upwards as fall approaches (August, September, and October). From table f, we can see that the average number of vehicles involved in accidents does not vary by much during summer months, and that the majority of accidents involve 1 vehicle, as the median for all three months is 1.

References

1. https://dplyr.tidyverse.org/reference/mutate.html
2. https://dplyr.tidyverse.org/reference/count.html
3. https://r-lang.com/how-to-combine-two-data-frames-in-r/