Homework 4

Will Scheib

Problem 1

Part a

```
fatal_accidents <- read.csv("data/fatal accidents.csv")
head(fatal_accidents)
```

```
##                   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
```

Part b

```
state.list <- split(fatal_accidents, fatal_accidents$State)
#https://stackoverflow.com/questions/15377238/r-subsetting-a-data-frame-into-
#multiple-data-frames-based-on-multiple-column-val
```

Part 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
```

Part d

```
percent_table <- function(state) {
  round(100*table(state$Day.of.week)/nrow(state), 1)
}

lapply(state.list, percent_table)
```

```
## $‘District of Columbia‘
##
##    1    2    3    4    5    6    7
## 13.6 13.6 22.7 13.6  4.5 27.3  4.5
##
## $Maryland
##
##    1    2    3    4    5    6    7
## 16.9 13.4 14.9 10.7 12.2 14.5 17.4
##
## $‘North Carolina‘
##
##    1    2    3    4    5    6    7
## 14.5 12.2 13.2 13.4 13.2 16.1 17.4
##
## $Virginia
##
##    1    2    3    4    5    6    7
## 15.4 12.3 13.4 13.2 13.7 16.4 15.6
##
## $‘West Virginia‘
##
##    1    2    3    4    5    6    7
## 12.1 14.6 14.6 13.8 13.0 13.0 19.0
```

Part e

Accidents occur at approximately the same rate on every day in every state with the exception of spikes on Friday in the District of Columbia and weekends in the other four.

Part f

```
percent_table2 <- function(state) {
  table(state$Day.of.week, state$Vehicle.count)
}

lapply(state.list, percent_table2)
```

```
## $‘District of Columbia‘
##
##     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
##
##      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‘
##
##       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
##
##      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`
##
##      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
```

Part g

Substituting names of days for numbers would make the data much more readable.

Problem 2

Part a

```
fatal_accidents2 <- fatal_accidents %>% mutate(People.count=People.count.IN+People.count
head(fatal_accidents2)
```

```
##                  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
```

Part b

```
fatal_accidents2 %>%
  group_by(State) %>%
  summarize(avg.Vehicles=mean(Vehicle.count), avg.People=mean(People.count))
```

```
## # A tibble: 5 x 3
##   State                avg.Vehicles avg.People
##   <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
```

Part c

```
fatal_accidents2 %>%
  group_by(State) %>%
  summarize(
    min.Vehicles=min(Vehicle.count),
    avg.Vehicles=mean(Vehicle.count),
    max.Vehicles=max(Vehicle.count)
  )
```

```
## # A tibble: 5 x 4
##   State                min.Vehicles avg.Vehicles max.Vehicles
##   <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
```

Part d

There tend to be more vehicles involved in crashes in Maryland than anywhere else and more people involved in crashes in DC than anywhere else.

Part e

```
fatal_accidents2 %>%
  filter(State=="Virginia") %>%
  count(Month)
```

```
##    Month  n
## 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
```

Part f

```
fatal_accidents2 %>%
  filter(State=="Virginia", Month %in% 6:8) %>%
  group_by(Day.of.week, Month) %>%
  summarize(
    med.Vehicles=median(Vehicle.count),
    avg.Vehicles=mean(Vehicle.count)
  )
```

```
## `summarise()` has grouped output by 'Day.of.week'. You can override using the `.group
```

```
## # A tibble: 21 x 4
## # Groups:   Day.of.week [7]
##    Day.of.week Month med.Vehicles avg.Vehicles
##          <int> <int>        <dbl>        <dbl>
##  1           1     6          1           1.55
##  2           1     7          1           1.6
##  3           1     8          1.5         1.5
##  4           2     6          1           1.6
##  5           2     7          1           1.38
##  6           2     8          1.5         1.88
##  7           3     6          1           1.4
##  8           3     7          2           1.73
##  9           3     8          2           1.57
## 10           4     6          1           1.2
## # ... with 11 more rows
```

Part g

There are more accidents in the fall (September, October, November) than any other time. Also, it is very hard to interpret information from the second tibble.