Assignment 3

Michael LaVallee

9/23/2019

1.Read the titanic data set as a tibble,Redo questions 13 to 23

```
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
```

```
titanic = read.csv("C:/Users/student/Documents/Senior Year/Fall Semester/R Analytics/Data Set/titanic.cs
```

13

```
titanic %>% filter(Sex=='female')%>% summarise(mean=mean(Age,na.rm =TRUE))
```

```
##       mean
## 1 27.91571
```

14

```
titanic %>% filter(Pclass == '1') %>% summarise(median(Fare,na.rm=TRUE))
```

```
##   median(Fare, na.rm = TRUE)
## 1                    60.2875
```

15

```
titanic %>% filter (Pclass != '1')%>% filter(Sex == 'female')%>% summarise(median(Fare, na.rm = TRUE))
```

```
##   median(Fare, na.rm = TRUE)
## 1                   14.45625
```

16

```
titanic %>% filter(Pclass != '1' ) %>% filter (Sex == 'female')%>% filter(Survived == '1')%>% summarise
```

```
##   median(Age, na.rm = TRUE)
## 1                        24
```

17

```
titanic %>% filter(Age >= 13)%>% filter(Age<=19)%>% filter(Survived == '1')%>% filter (Sex == 'female')
```

```
##   mean(Fare, na.rm = TRUE)
## 1                 49.17966
```

18

```
titanic %>% filter(Age >= 13)%>% filter(Age<=19)%>% filter(Survived == '1')%>% filter (Sex == 'female')
```

```
## # A tibble: 3 x 2
##   Pclass `mean(Fare, na.rm = TRUE)`
##    <int>                      <dbl>
## 1      1                     108.
## 2      2                      20.0
## 3      3                       8.77
```

19

```
nineteen=titanic %>% filter(Fare > mean(Fare,na.rm = TRUE)) %>% group_by(Survived)%>% summarise(counting
```

20

```
twenty= titanic %>% mutate(sfare=(Fare-mean(Fare,na.rm = TRUE))/sd(Fare,na.rm = TRUE))
```

21

```
titanic1 =titanic %>% mutate (cfare = cut(Fare,breaks = c(-Inf,mean(Fare,na.rm = TRUE),Inf),label = c("
```

22

```
ages <- c(0,9.99,19.99,29.99,39.99,49.99,59.99,69.99,79.99,89.99)

label<-c(0,1,2,3,4,5,6,7,8)
titanic2=titanic1%>% mutate(cage=cut(Age,breaks = ages,labels = label))
```

23

```
frequen = titanic2%>%mutate(Embarked=replace(Embarked,Embarked=='',"S"))%>% group_by(Embarked)%>% summa
```

2.Using Dplyr and in Assignment 2, redo 4 using sample_n function, redo 5 using glimpse, redo 11, 12,and 13. For 11, 12 and 13, you may want to use the combo group_by and summarise

```
library(readxl)
c2015 = read_excel("C:/Users/student/Documents/Senior Year/Fall Semester/R Analytics/Data Set/c2015.xlsx
library(dplyr)
dim(c2015)
```

```
## [1] 80587    28
```

```
set.seed(2019)
sample2015 = sample_n(c2015, 1000)
sample2015$TRAV_SP <-substr(sample2015$TRAV_SP,1,nchar(sample2015$TRAV_SP)-4)
sample2015$TRAV_SP <- as.numeric(as.character(sample2015$TRAV_SP))
```

```
## Warning: NAs introduced by coercion
```

11

```
no=sample2015 %>% filter(INJ_SEV == 'No Apparent Injury (O)') %>% summarise(mean (TRAV_SP, na.rm = TRUE)

injury=sample2015 %>% filter(INJ_SEV != 'No Apparent Injury (O)') %>% summarise(mean (TRAV_SP, na.rm = T
```

12

```
driver=sample2015 %>% filter(SEAT_POS == "Front Seat, Left Side") %>% group_by(SEX)%>% summarise(mean(TR
```

13

```
yes=sample2015 %>% filter(DRINKING == 'Yes (Alcohol Involved)')%>%  summarise(mean(TRAV_SP, na.rm = TRU
no=sample2015 %>% filter(DRINKING == 'No (Alcohol Not Involved)')%>%  summarise(mean(TRAV_SP, na.rm = TR
```

3. Calculate the travel speed (TRAV_SP variable) by day. Compare the travel speed of the first 5 days and the last 5 days of months.

```
first5=sample2015%>% filter (DAY >= 1)%>% filter(DAY <= 5)%>% summarise(mean(TRAV_SP, na.rm = TRUE))
last5=sample2015%>% filter (DAY >= 25)%>% filter(DAY <= 30)%>% summarise(mean(TRAV_SP, na.rm = TRUE))
```

4. Calculate the travel speed (TRAV_SP variable) by day of the week. Compare the travel speed of the weekdays and weekends

```
days = sample2015 %>% group_by(DAY_WEEK) %>% summarise(mean = mean(TRAV_SP, na.rm = TRUE))

mean(days[3:4,]$mean)
```

```
## [1] 54.53541
```

```
mean(days[-3:-4,]$mean)
```

```
## [1] 48.40777
```

5 Find the top 5 states with greatest travel speed

```
states = sample2015 %>% group_by(STATE)%>% summarise(mean=mean(TRAV_SP,na.rm = TRUE))%>% arrange(desc(m

states[1:5,]
```

```
## # A tibble: 5 x 2
##   STATE         mean
##   <chr>        <dbl>
## 1 South Dakota 107
## 2 North Dakota  85
## 3 Nevada        73.5
## 4 Wyoming       66.5
## 5 Kentucky      65.4
```

Rank travel speed by month

```
month = sample2015 %>% group_by(MONTH)%>% summarise(mean=mean(TRAV_SP,na.rm = TRUE))%>% arrange(desc(mea
month
```

```
## # A tibble: 12 x 2
##    MONTH      mean
##    <chr>     <dbl>
##  1 April      59.3
##  2 December   59.0
##  3 September  54.7
##  4 June       53.4
##  5 October    52.5
##  6 November   52.5
##  7 August     48.9
##  8 May        48.3
##  9 February   46.4
## 10 March      45.4
## 11 January    45.2
## 12 July       44.9
```

7. Find the average speed of teenagers in December.

```
teen = sample2015%>% filter(AGE>12, AGE <20)%>%  filter (MONTH == 'December')%>% summarise(mean(TRAV_SP
teen
```

```
## # A tibble: 1 x 1
##   `mean(TRAV_SP, na.rm = TRUE)`
##                           <dbl>
## 1                            80
```

8 Find the month that female drivers drive fastest on average

```
fast = sample2015 %>% filter(SEX == 'Female',SEAT_POS == 'Front Seat, Left Side')%>% group_by(MONTH)%>%

fast[1,]
```

```
## # A tibble: 1 x 2
##   MONTH      mean
##   <chr>     <dbl>
## 1 September  75.7
```

9. Find the month that male driver drive slowest on average.

```
fast = sample2015 %>% filter(SEX == 'Male',SEAT_POS == 'Front Seat, Left Side')%>% group_by(MONTH)%>% su

fast[1,]
```

```
## # A tibble: 1 x 2
##   MONTH     mean
##   <chr>    <dbl>
## 1 February  36.2
```

10 Create a new column containing information about the season of the accidents. Compare the percentage of Fatal Injury by seasons.

```
sample2015 <- sample2015%>% mutate(seasons = recode(MONTH,'December' = 'Winter', 'January'='Winter', 'Fe

Fatal <- sample2015 %>% group_by(seasons)%>% summarise(percent= sum(INJ_SEV == 'Fatal Injury (K)')/n())

Fatal
```

```
## # A tibble: 4 x 2
##   seasons percent
##   <chr>     <dbl>
## 1 Autumn    0.440
## 2 Spring    0.418
## 3 Summer    0.459
## 4 Winter    0.409
```

11 Compare the percentage of fatal injuries for different type of deformations (DEFORMED variable)

```
deformed <- sample2015 %>% group_by(DEFORMED)%>% summarise(percent= sum(INJ_SEV == 'Fatal Injury (K)')/n

deformed
```

```
## # A tibble: 7 x 2
##   DEFORMED          percent
##   <chr>               <dbl>
## 1 Disabling Damage   0.477
## 2 Functional Damage  0.103
## 3 Minor Damage       0.0897
## 4 No Damage          0.125
## 5 Not Reported       0.205
## 6 Unknown            0.35
## 7 <NA>               0.895
```