Assignment 3

Rachel Cardarelli

9/30/2019

Assignment 1 Using Dplyr

```
titanic <- read.csv("C:/Users/student/Documents/Honors Thesis/titanic.csv")
titanic <- subset(titanic, select = -c(Name, PassengerId, Ticket, Cabin))

#install.packages("dplyr")
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 <-tbl_df(titanic)
```

Question 13

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

```
## # A tibble: 1 x 1
##   mean_Age
##      <dbl>
## 1     27.9
```

Question 14

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

```
## # A tibble: 1 x 1
##   median_fare
##         <dbl>
## 1        60.3
```

Question 15

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

```
## # A tibble: 1 x 1
##   median_fare
##         <dbl>
## 1        14.5
```

Question 16

```
titanic %>%
  filter(Survived == 1, Sex == "female", Pclass != 3) %>%
  summarise(median_age = median(Age, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   median_age
##        <dbl>
## 1         31
```

Question 17

```
titanic %>%
  filter(Sex == 'female', Survived == 1, Age>12, Age<20) %>%
  summarise(mean_fare = mean(Fare, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   mean_fare
##       <dbl>
## 1      49.2
```

Question 18

```
titanic %>%
  filter(Sex == 'female', Survived == 1, Age>12, Age<20) %>%
  group_by(Pclass) %>%
  summarise(mean_fare = mean(Fare, na.rm = TRUE))
```

```
## # A tibble: 3 x 2
##   Pclass mean_fare
##    <int>     <dbl>
## 1      1    108.
## 2      2     20.0
## 3      3      8.77
```

Question 19

```
titanic %>%
  filter(Fare > mean(Fare, na.rm = TRUE)) %>%
  summarize(ratio = sum(Survived == 1)/sum(Survived == 0))
```

```
## # A tibble: 1 x 1
##   ratio
##   <dbl>
## 1  1.48
```

Question 20

```
titanic %>%
  mutate(sfare = (Fare - mean(Fare))/sd(Fare)) %>%
  names
```

```
## [1] "Survived" "Pclass"   "Sex"      "Age"      "SibSp"    "Parch"
## [7] "Fare"     "Embarked" "sfare"
```

Question 21

```
titanic %>%
  na.omit %>%
  mutate(cfare=cut(Fare, breaks=c(-Inf,mean(Fare),Inf), labels=c("cheap","expensive"))) %>%
  names
```

```
## [1] "Survived" "Pclass"   "Sex"      "Age"      "SibSp"    "Parch"
## [7] "Fare"     "Embarked" "cfare"
```

Question 22

```
max(titanic$Age, na.rm = TRUE)
```

```
## [1] 80
```

```
titanic %>%
  na.omit %>%
  mutate(cage=cut(Age, breaks=seq(0,80,by=10), labels=c(1:8))) %>%
  names
```

```
## [1] "Survived" "Pclass"   "Sex"      "Age"      "SibSp"    "Parch"   
## [7] "Fare"     "Embarked" "cage"    
```

Question 23

```
titanic %>%
  group_by(Embarked) %>%
  count(Embarked)
```

```
## # A tibble: 4 x 2
## # Groups:   Embarked [4]
##   Embarked     n
##   <fct>    <int>
## 1 ""           2
## 2 C          168
## 3 Q           77
## 4 S          644
```

```
levels(titanic$Embarked)[1] <- "S"
```

Assignment 2 Using Dplyr

Question 4

```
library(dplyr)
library(readxl)

c2015 <- read_excel("c2015.xlsx")
c2015 <-  tbl_df(c2015)
set.seed(2019)
c2015 <- sample_n(c2015,1000, replace = FALSE)
```

Question 5

```
glimpse(c2015)
```

```
## Observations: 1,000
## Variables: 28
## $ STATE    <chr> "New Jersey", "Arizona", "Tennessee", "Minnesota", "M...
## $ ST_CASE  <dbl> 340336, 40327, 470789, 270119, 290576, 62865, 330095,...
## $ VEH_NO   <dbl> 1, 1, 1, 2, 1, 1, 0, 0, 2, 5, 1, 2, 1, 0, 1, 1, 2, 1,...
## $ PER_NO   <dbl> 1, 1, 1, 4, 1, 1, 1, 1, 4, 1, 1, 1, 5, 1, 1, 2, 1, 1,...
## $ COUNTY   <dbl> 27, 13, 163, 59, 201, 19, 15, 127, 13, 115, 29, 141, ...
## $ DAY      <dbl> 19, 7, 2, 16, 2, 6, 3, 30, 17, 30, 19, 12, 9, 30, 9, ...
## $ MONTH    <chr> "September", "May", "December", "May", "October", "Ju...
## $ HOUR     <dbl> 3, 22, 8, 21, 15, 15, 14, 20, 7, 14, 14, 17, 18, 6, 4...
## $ MINUTE   <dbl> 17, 15, 26, 59, 38, 20, 32, 20, 41, 36, 15, 50, 55, 4...
## $ AGE      <chr> "Unknown", "47", "23", "15", "55", "56", "26", "63", ...
## $ SEX      <chr> "Unknown", "Female", "Male", "Female", "Male", "Male"...
## $ PER_TYP  <chr> "Driver of a Motor Vehicle In-Transport", "Driver of ...
## $ INJ_SEV  <chr> "Unknown", "No Apparent Injury (O)", "Unknown", "Susp...
## $ SEAT_POS <chr> "Front Seat, Left Side", "Front Seat, Left Side", "Fr...
## $ DRINKING <chr> "Not Reported", "No (Alcohol Not Involved)", "Unknown...
## $ YEAR     <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,...
## $ MAN_COLL <chr> "Not a Collision with Motor Vehicle In-Transport", "N...
## $ OWNER    <chr> "Unknown", "Driver (in this crash) Not Registered Own...
## $ MOD_YEAR <chr> "Unknown", "2003", "1994", "2011", "2000", "2013", NA...
## $ TRAV_SP  <chr> "Unknown", "048 MPH", "Not Rep", "055 MPH", "055 MPH"...
## $ DEFORMED <chr> "Unknown", "Functional Damage", "Minor Damage", "Disa...
## $ DAY_WEEK <chr> "Saturday", "Thursday", "Wednesday", "Saturday", "Fri...
## $ ROUTE    <chr> "State Highway", "Local Street", "County Road", "Stat...
## $ LATITUDE <dbl> 40.95270, 33.41048, 36.57834, 45.42841, 37.13481, 36....
## $ LONGITUD <dbl> -74.59644, -112.06459, -82.27889, -93.36788, -89.5946...
## $ HARM_EV  <chr> "Pedestrian", "Pedestrian", "Pedalcyclist", "Motor Ve...
## $ LGT_COND <chr> "Dark - Not Lighted", "Dark - Lighted", "Dark - Not L...
## $ WEATHER  <chr> "Clear", "Clear", "Clear", "Rain", "Cloud", "Clear", ...
```

```
c2015 <- select(c2015,-YEAR)
```

Question 11

```
library("stringr")

c2015$TRAV_SP <- str_replace(c2015$TRAV_SP," MPH","")

c2015$TRAV_SP[c2015$TRAV_SP == "Unknown"] <- "NA"
c2015$TRAV_SP[c2015$TRAV_SP == "Not Rep"] <- "NA"
c2015$TRAV_SP[c2015$TRAV_SP == "Greater"] <- "NA"
c2015$TRAV_SP[c2015$TRAV_SP == "Stopped"] <- "0"

c2015$TRAV_SP <- as.numeric(c2015$TRAV_SP)
```

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

```
mean(c2015$TRAV_SP,na.rm=TRUE)
```

```
## [1] 43.79245
```

```
#Those with no apparent injuries had lower travel speeds on average
c2015 %>%
    group_by(INJ_SEV) %>%
    summarise(mean_speed = mean(TRAV_SP, na.rm = TRUE))
```

```
## # A tibble: 7 x 2
##   INJ_SEV                      mean_speed
##   <chr>                             <dbl>
## 1 Fatal Injury (K)                   52.5
## 2 Injured, Severity Unknown          35
## 3 No Apparent Injury (O)             33.6
## 4 Possible Injury (C)                34.9
## 5 Suspected Minor Injury(B)          46.7
## 6 Suspected Serious Injury(A)        51.5
## 7 Unknown                            35
```

Question 12

```
c2015$SEX[c2015$SEX == "Unknown"] <- "Female"
c2015$SEX[c2015$SEX == "Not Rep"] <- "Female"

c2015 %>%
  filter(SEAT_POS == "Front Seat, Left Side") %>%
  group_by(SEX) %>%
  summarise(mean_speed = mean(TRAV_SP, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   SEX    mean_speed
##   <chr>       <dbl>
## 1 Female       37.1
## 2 Male         45.6
```

Question 13

```
c2015 %>%
  group_by(DRINKING) %>%
  summarise(mean_speed = mean(TRAV_SP, na.rm = TRUE))
```

```
## # A tibble: 4 x 2
##   DRINKING                    mean_speed
##   <chr>                            <dbl>
## 1 No (Alcohol Not Involved)         37.2
## 2 Not Reported                      45.0
## 3 Unknown (Police Reported)         50.8
## 4 Yes (Alcohol Involved)            66.4
```

Assignment 3 Questions

Question 3

```
c2015 %>%
  mutate(day_of_month = ifelse(DAY %in% 1:5, "First", ifelse(DAY %in% 26:30,"Last", NA))) %>%
  group_by(day_of_month) %>%
  summarise(mean_by_day = mean(TRAV_SP, na.rm = TRUE))
```

```
## # A tibble: 3 x 2
##   day_of_month mean_by_day
##   <chr>              <dbl>
## 1 First               44.4
## 2 Last                51.2
## 3 <NA>                42.4
```

Question 4

```
c2015 %>%
  group_by(DAY_WEEK %in% c("Saturday", "Sunday")) %>%
  summarise(mean_by_day_of_week = mean(TRAV_SP, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   `DAY_WEEK %in% c("Saturday", "Sunday")` mean_by_day_of_week
##   <lgl>                                                 <dbl>
## 1 FALSE                                                  41.3
## 2 TRUE                                                   48.5
```

Question 5

```
c2015 %>%
  group_by(STATE) %>%
  summarise(mean_sp = mean(TRAV_SP, na.rm = TRUE)) %>%
  arrange(desc(mean_sp)) %>%
  top_n(5)
```

```
## Selecting by mean_sp
```

```
## # A tibble: 5 x 2
##   STATE        mean_sp
##   <chr>          <dbl>
## 1 North Dakota    85
## 2 Nevada          73.5
## 3 Wyoming         66.5
## 4 Alabama         57.6
## 5 Rhode Island    57
```

Question 6

```r
c2015 %>%
  group_by(MONTH) %>%
  summarize(mean_sp = mean(TRAV_SP, na.rm = TRUE)) %>%
  arrange(desc(mean_sp))
```

```
## # A tibble: 12 x 2
##    MONTH     mean_sp
##    <chr>       <dbl>
##  1 December     51.9
##  2 April        49.4
##  3 September    48.0
##  4 June         47.7
##  5 November     47.1
##  6 October      46.8
##  7 August       43.9
##  8 May          43.1
##  9 July         37.4
## 10 March        37.0
## 11 February     36.4
## 12 January      34.3
```

Question 7

```r
c2015 %>%
  filter(MONTH == "December", AGE > 12, AGE < 20) %>%
  summarize(mean_sp_teenagers_December = mean(TRAV_SP, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   mean_sp_teenagers_December
##                        <dbl>
## 1                         80
```

Question 8

```r
c2015 %>%
  filter(SEX == "Female") %>%
  group_by(MONTH) %>%
  summarize(mean_sp_female = mean(TRAV_SP, na.rm = TRUE)) %>%
  top_n(1)
```

```
## Selecting by mean_sp_female
```

```
## # A tibble: 1 x 2
##   MONTH    mean_sp_female
##   <chr>             <dbl>
## 1 December           60.3
```

Question 9

```
c2015 %>%
  filter(SEX == "Male") %>%
  group_by(MONTH) %>%
  summarize(mean_sp_male = mean(TRAV_SP, na.rm = TRUE)) %>%
  top_n(-1)
```

```
## Selecting by mean_sp_male
```

```
## # A tibble: 1 x 2
##   MONTH   mean_sp_male
##   <chr>          <dbl>
## 1 January           34
```

Question 10

```
Spring <-  c("March", "April", "May")
Summer <-  c("June", "July", "August")
Fall <- c("September", "October", "November")
Winter <- c("December", "January", "February")

c2015 %>%
  na.omit %>%
  mutate(SEASONS= ifelse(MONTH %in% Spring, "Spring", ifelse(MONTH %in% Summer, "Summer", ifelse(MONTH %
  group_by(SEASONS) %>%
  summarise(fatal_injury = sum(INJ_SEV == "Fatal Injury (K)")/n())
```

```
## # A tibble: 4 x 2
##   SEASONS fatal_injury
##   <chr>          <dbl>
## 1 Fall           0.432
## 2 Spring         0.268
## 3 Summer         0.330
## 4 Winter         0.254
```

Question 11

```
c2015 %>%
  group_by(DEFORMED) %>%
  summarise(fatal_injury = sum(INJ_SEV == "Fatal Injury (K)")/n())
```

```
## # A tibble: 7 x 2
##   DEFORMED          fatal_injury
##   <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
```