Assignment 3

Hyungue Lim

9/30/2019

1

```
library("tidyverse")
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
titanic <- read_csv("titanic.csv")
```

```
## Parsed with column specification:
## cols(
##   PassengerId = col_double(),
##   Survived = col_double(),
##   Pclass = col_double(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_double(),
##   SibSp = col_double(),
##   Parch = col_double(),
##   Ticket = col_character(),
##   Fare = col_double(),
##   Cabin = col_character(),
##   Embarked = col_character()
## )
```

```
#13
str(titanic)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 891 obs. of  12 variables:
##  $ PassengerId: num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : num  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : num  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : num  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : num  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  NA "C85" NA "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   PassengerId = col_double(),
##   ..   Survived = col_double(),
##   ..   Pclass = col_double(),
##   ..   Name = col_character(),
##   ..   Sex = col_character(),
##   ..   Age = col_double(),
##   ..   SibSp = col_double(),
##   ..   Parch = col_double(),
##   ..   Ticket = col_character(),
##   ..   Fare = col_double(),
##   ..   Cabin = col_character(),
##   ..   Embarked = col_character()
##   .. )
```

```r
titanic %>%
  filter(Sex == "female") %>%
    summarize(female_mean_age = mean(Age, na.rm = 1))
```

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

```r
#14
titanic %>%
  filter(Pclass == 1) %>%
    summarize(Class1_median_fare = median(Fare, na.rm = 1))
```

```
## # A tibble: 1 x 1
##   Class1_median_fare
##                <dbl>
## 1               60.3
```

```r
#15
titanic %>%
  filter(Sex == "female", Pclass != 1) %>%
    summarize(female_median_fare_not_class1 = median(Fare, na.rm =1))
```

```
## # A tibble: 1 x 1
##   female_median_fare_not_class1
##                           <dbl>
## 1                          14.5
```

```
#16
titanic %>%
  filter(Survived == 1, Sex == "female", Pclass != 3) %>%
    summarize(median_age = median(Age, na.rm = 1))
```

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

```
#17
titanic %>%
  filter(Survived == 1, Sex == "female", Age <20, Age>=10) %>%
    summarize(mean_fare = mean(Fare, na.rm = 1))
```

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

```
#18
titanic %>%
  filter(Survived == 1, Sex == "female", Age <20, Age>=10) %>%
  group_by(Pclass) %>%
  summarize(mean_fare = mean(Fare, na.rm = 1))
```

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

```
#19
avg_fare <- mean(titanic$Fare, na.rm = 1)
titanic %>%
  filter(Fare > avg_fare) %>%
    summarize(ratio = sum(Survived==1)/sum(Survived==0))
```

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

```
#20
titanic <- titanic %>%
  mutate(sfare = (Fare - avg_fare) / sd(Fare, na.rm = 1))

#21
titanic <- titanic %>%
  mutate(cfare = ifelse(Fare < avg_fare,'cheap','expensive'))
```

```r
#22
titanic <- titanic %>%
  mutate(cage = Age/10 - Age%%10/10)

#23
table(titanic$Embarked)
```

```
##
##   C   Q   S
## 168  77 644
```

```r
titanic$Embarked <- titanic$Embarked %>%
  replace_na("S")
table(titanic$Embarked)
```

```
##
##   C   Q   S
## 168  77 646
```

2

```r
library(readxl)

c2015 <- read_excel("c2015.xlsx")

# 4
set.seed(2019)
c2015_sample <- sample_n(c2015, 1000)

# 5
glimpse(c2015_sample)
```

```
## 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_sample <- c2015_sample[,-16]

# 11
library("stringr")
c2015_sample$TRAV_SP <- str_replace(c2015_sample$TRAV_SP, " MPH", "")
c2015_sample$TRAV_SP <- str_replace(c2015_sample$TRAV_SP, "Stopped", "0")

c2015_sample$TRAV_SP <- as.numeric(c2015_sample$TRAV_SP)
```

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

```
c2015_sample %>% group_by(INJ_SEV) %>% summarize(mean(TRAV_SP, na.rm=TRUE)) #People with no apparent in
```

```
## # A tibble: 7 x 2
##   INJ_SEV                      `mean(TRAV_SP, na.rm = TRUE)`
##   <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
```

```
# 12
c2015_sample %>% filter(SEAT_POS == "Front Seat, Left Side") %>% group_by(SEX) %>% summarize(mean(TRAV_S
```

```
## # A tibble: 3 x 2
##   SEX     `mean(TRAV_SP, na.rm = TRUE)`
##   <chr>                           <dbl>
## 1 Female                           37.1
## 2 Male                             45.6
## 3 Unknown                          36.7
```

```
#Man were driving faster than women on average

# 13
c2015_sample %>% group_by(DRINKING) %>% summarize(mean(TRAV_SP, na.rm=TRUE))
```

```
## # A tibble: 4 x 2
##   DRINKING                     `mean(TRAV_SP, na.rm = TRUE)`
##   <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
```

```
#People who were involved with alcohol were driving faster than others on average.
```

3

```
c2015_sample %>% group_by(DAY) %>% summarize(mean(TRAV_SP, na.rm=TRUE))
```

```
## # A tibble: 31 x 2
##      DAY `mean(TRAV_SP, na.rm = TRUE)`
##    <dbl>                         <dbl>
##  1     1                          49.2
##  2     2                          49.5
##  3     3                          45.6
##  4     4                          38.2
##  5     5                          42
##  6     6                          40.6
##  7     7                          40.1
##  8     8                          42.7
##  9     9                          50.8
## 10    10                          47.1
## # ... with 21 more rows
```

```
c2015_sample <- c2015_sample %>% mutate(day_group = ifelse(DAY <=5, "first five", ifelse(DAY >= 27, "la
c2015_sample %>% group_by(day_group) %>% summarize(mean(TRAV_SP, na.rm=TRUE))  %>% filter(day_group ==
```

```
## # A tibble: 2 x 2
##   day_group  `mean(TRAV_SP, na.rm = TRUE)`
##   <chr>                              <dbl>
## 1 first five                          44.4
## 2 last five                           52.6
```

4

```
c2015_sample %>% group_by(DAY_WEEK) %>% summarize(mean(TRAV_SP, na.rm=TRUE))
```

```
## # A tibble: 7 x 2
##   DAY_WEEK  `mean(TRAV_SP, na.rm = TRUE)`
##   <chr>                             <dbl>
## 1 Friday                             42.6
## 2 Monday                             40.8
## 3 Saturday                           48.0
## 4 Sunday                             49.2
## 5 Thursday                           47.8
## 6 Tuesday                            39.7
## 7 Wednesday                          33.8
```

```
c2015_sample <- c2015_sample %>% mutate(day_week_group = ifelse(DAY_WEEK == "Saturday" | DAY_WEEK == "Su
c2015_sample %>% group_by(day_week_group) %>% summarize(mean(TRAV_SP, na.rm=TRUE))
```

```
## # A tibble: 2 x 2
##   day_week_group `mean(TRAV_SP, na.rm = TRUE)`
##   <chr>                                  <dbl>
## 1 Weekday                                 41.3
## 2 Weekend                                 48.5
```

5

```
c2015_sample %>% select(STATE, TRAV_SP) %>% top_n(5, TRAV_SP) %>% arrange(desc(TRAV_SP))
```

```
## # A tibble: 9 x 2
##   STATE        TRAV_SP
##   <chr>          <dbl>
## 1 Kentucky         113
## 2 South Dakota     107
## 3 Florida          100
## 4 Pennsylvania     100
## 5 Florida           90
## 6 Virginia          90
## 7 Florida           90
## 8 Alabama           90
## 9 Pennsylvania      90
```

6

```
c2015_sample %>% group_by(MONTH) %>% summarize(speed = mean(TRAV_SP, na.rm=TRUE)) %>% arrange(desc(spee
```

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

7

```
c2015_sample$AGE <- c2015_sample$AGE %>% recode("Less than 1" = "0") %>% as.numeric
```

```
## Warning in function_list[[k]](value): NAs introduced by coercion
```

```
c2015_sample %>% filter(AGE < 20, MONTH == "December") %>% summarize(mean(TRAV_SP, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   `mean(TRAV_SP, na.rm = TRUE)`
##                           <dbl>
## 1                          62.5
```

8

```
c2015_sample %>% filter(SEX == "Female") %>% group_by(MONTH) %>% summarize(speed = mean(TRAV_SP, na.rm
```

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

9

```
c2015_sample %>% filter(SEX == "Male") %>% group_by(MONTH) %>% summarize(speed = mean(TRAV_SP, na.rm =
```

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

10

```
c2015_sample <- c2015_sample %>% mutate(for_season=paste("2012",MONTH,DAY, Sep=""))

c2015_sample$for_season <- c2015_sample$for_season %>% as.Date(format = "%Y %b %d")

getSeason <- function(DATES) {
    Winter <- as.Date("2012-12-15", format = "%Y-%m-%d")
    Spring <- as.Date("2012-3-15",  format = "%Y-%m-%d")
    Summer <- as.Date("2012-6-15",  format = "%Y-%m-%d")
    Fall <- as.Date("2012-9-15",  format = "%Y-%m-%d")

    ifelse (DATES >= Winter | DATES < Spring, "Winter",
      ifelse (DATES >= Spring & DATES < Summer, "Spring",
        ifelse (DATES >= Summer & DATES < Fall, "Summer", "Fall")))
}
c2015_sample <- c2015_sample %>% mutate(SEASON = getSeason(for_season))


c2015_sample %>% group_by(SEASON) %>% summarize(por = prop.table(table(INJ_SEV))[1])
```

```
## # A tibble: 4 x 2
##   SEASON   por
##   <chr>  <dbl>
## 1 Fall   0.459
## 2 Spring 0.414
## 3 Summer 0.448
## 4 Winter 0.402
```

11

```
c2015_sample %>% group_by(DEFORMED) %>% summarize(por = prop.table(table(INJ_SEV))[1])
```

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