# Assignment 3. Data Wrangling with Dplyr

This assignment assumes that you have taken the `Introduction to the Tidyverse` and `Data Manipulation with dplyr in R` course at Datacamp. You can use base R functions and dplyr functions in the assignment.

***Submission Instruction***. You will need to submit on **Blackboard**, in the **Assignment** section, the follows:

- A knitted pdf
- A link to the markdown document in your Github
- A link to the pdf document in your Github

## Questions

1. Read the `titanic` data set as a tibble. Redo questions 13 to 23 in the Assignment 1 using `dplyr`. **Notice:** you may want to use logical operators such as:

```r
df <- read.csv(file="C:\\Users\\student\\Documents\\R\\titanic.csv")
str(df)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

13. Calculate the mean age of female passengers

```r
library("dplyr")          ## load
```

```
## 
## 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
```

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

```
##   mean(Age, na.rm = 1)
## 1             27.91571
```

14. Calculate the median fare of the passengers in Class 1

```
df %>%
  filter(df$Pclass== "1") %>%
  summarize(median(Fare, na.rm = 1))
```

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

Calculate the median fare of the female passengers that are not in Class 1

```
df %>%
  filter(Sex == "female",Pclass!= "1" ) %>%
  summarize(median(Fare, na.rm = 1))
```

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

Calculate the median age of survived passengers who are female and Class 1 or Class 2,

```
df %>%
  filter(Sex == "female",Pclass== "1"|Pclass=="2",Survived=="1" ) %>%
  summarize(median(Age, na.rm = 1))
```

```
##   median(Age, na.rm = 1)
## 1                      31
```

Calculate the mean fare of female teenagers survived passengers

```
df %>%
  filter(Sex == "female", Age>12 & Age<20 ,Survived=="1" ) %>%
  summarize(mean(Fare, na.rm = 1))
```

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

Calculate the mean fare of female teenagers survived passengers for each class

```
df %>%
  filter(Sex == "female", Age>12 & Age<20 ,Survived=="1" ) %>%
  group_by(Pclass) %>%
  summarize(mean(Fare, na.rm = 1))
```

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

Calculate the ratio of Survived and not Survived for passengers who are who pays more than the average fare

```
meanFare<-mean(df$Fare)
titanicsub= subset(df,Fare>meanFare)

df %>%
  filter(Fare>meanFare) %>% group_by(Survived) %>% summarise(n=n()) %>% mutate(f = n/sum(n))
```

```
## # A tibble: 2 x 3
##   Survived     n     f
##      <int> <int> <dbl>
## 1        0    85 0.403
## 2        1   126 0.597
```

Add column that standardizes the fare (subtract the mean and divide by standard deviation) and name it ‘sfare

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

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"    "sfare"
```

Add categorical variable named `cfare` that takes value **cheap** for passengers paying less the average fare and takes value **expensive** for passengers paying more than the average fare.

```
avgFare<- mean(df$Fare)

df = df %>%
  mutate(cfare = case_when(
    Fare < avgFare ~ "cheap",
    Fare > avgFare ~ "expensive"))
names(df)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"    "sfare"       "cfare"
```

Add categorical variable named `cage` that takes value 0 for age 0-10, 1 for age 10-20, 2 for age 20-30, and so on

```
df= mutate(
  df,
  cage = case_when(
    Age <11 ~ "1",
    Age %in% 11:20 ~ "2",
     Age %in% 21:30 ~ "3",
     Age %in% 31:40 ~ "4",
     Age %in% 41:50 ~ "5",
     Age %in% 51:60 ~ "6",
    Age %in% 61:70 ~ "7",
    Age > 70 ~ "8"

  )
)
names(df)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"    "sfare"       "cfare"       "cage"
```

```
|Operators|Discription|
|-------|--------------|
| !=    | not equal to |
| !x    | Not x        |
| x \text{|} y | x OR y|
| x & y | x AND y      |

2. Using Dplyr and in Assignment 2, redo  4 using `sample_n` function, redo 5 using `glimpse`, redo 11,

 ----Use `dim` function to check the dimension of the data. Since this data is quite big, a common prac
 dim(data_excel)
set.seed(2019)
c2015sample<-data_excel[sample(1:80587,1000),]
dim(c2015sample) ---
```

```
path <- "C:/Users/student/Documents/RStudio/c2015.xlsx"
library(readxl)
data_excel=read_excel(path)
class(data_excel)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
set.seed(2019)
df1<-sample_n(data_excel,1000, replace=TRUE)
```

Use `summary` function to have a quick look at the data. You will notice there is one variable is actually a constant. Remove that variable from the data. summary(c2015sample) data_excel2 = subset(c2015sample, select = -c(YEAR) ) summary(data_excel2)

```
glimpse(df1)
```

```
## 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", ...
```

```
df1<-select (df1,-c("YEAR"))
glimpse(df1)
```

```
## Observations: 1,000
## Variables: 27
## $ 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...
## $ 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", ...
```

Redo 11 from assignment 2 Compare the average speed of those who had `"No Apprent Injury"` and the rest. What do you observe? data_excel2<- subset(data_excel2,data_excel2$INJ_SEV ==' NoApparentInjury(O)')mean(data_excel2$TRAV_SP, na.rm=TRUE)

```r
library(stringr)
df1$TRAV_SP[df1$TRAV_SP=='Stopped'] <- '0'
df1$TRAV_SP[df1$TRAV_SP=='Not Rep' | df1$TRAV_SP=='Unknown'] <- NA
df1$TRAV_SP<- stringr::str_replace(df1$TRAV_SP," MPH", "")
df1$TRAV_SP <- as.numeric(df1$TRAV_SP)
avgSpeed<-mean(df1$TRAV_SP, na.rm = TRUE)

df1 %>%
group_by(INJ_SEV) %>%
  summarize(mean(TRAV_SP, na.rm = 1))
```

```
## # A tibble: 7 x 2
##   INJ_SEV                          `mean(TRAV_SP, na.rm = 1)`
##   <chr>                                                  <dbl>
## 1 Fatal Injury (K)                                        53.0
## 2 Injured, Severity Unknown                               35
## 3 No Apparent Injury (O)                                  33.3
## 4 Possible Injury (C)                                     34.9
## 5 Suspected Minor Injury(B)                               46.7
## 6 Suspected Serious Injury(A)                             50.8
## 7 Unknown                                                 35
```

Redo 12 from assignment 2 Use the `SEAT_POS` variable to filter the data so that there is only **drivers** in the dataset. Compare the average speed of man drivers and woman drivers. Comment on the results.

table(data_excel2$SEAT_POS)data_excel2 < -subset(data_excel2, data_excel2$SEAT_POS==‘Front Seat, Left Side’) table(data_excel2$SEAT_POS)

MaleSet<-(subset(data_excel2,data_excel2$SEX ==' Male'))FemaleSet < -(subset(data_excel2, data_excel2$SEX==‘Femal
mean(MaleSet$TRAV_SP, na.rm = TRUE)mean(FemaleSet$TRAV_SP,na.rm = TRUE)

```r
library(dplyr)
df1 %>%
  filter(SEAT_POS=="Front Seat, Left Side") %>%
  group_by(SEX) %>%
  summarize(mean(TRAV_SP, na.rm=1))
```

```
## # A tibble: 3 x 2
##   SEX     `mean(TRAV_SP, na.rm = 1)`
##   <chr>                        <dbl>
## 1 Female                        37.6
## 2 Male                          46.0
## 3 Unknown                       36.7
```

redo 13 from assignment 2 Compare the average speed of drivers who drink and those who do not. Comment on the results. **Hint:** This calculation can be done manually or by using the `aggregate` function or `by` function in base R. For example: aggregate(data_excel2$TRAV_SP, by = list(data_excel2$DRINKING),FUN=mean, na.rm=TRUE)

```
df1 %>%
  group_by(DRINKING) %>%
  summarize(mean(TRAV_SP, na.rm=1))
```

```
## # A tibble: 4 x 2
##   DRINKING                  `mean(TRAV_SP, na.rm = 1)`
##   <chr>                                          <dbl>
## 1 No (Alcohol Not Involved)                       37.3
## 2 Not Reported                                    44.8
## 3 Unknown (Police Reported)                       51.3
## 4 Yes (Alcohol Involved)                          67.0
```

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.

```
library(dplyr)

df1 %>%
  mutate(
   timeofmonth = case_when(
    DAY=="1" ~ "first 5",
        DAY=="2" ~ "first 5",
        DAY=="3" ~ "first 5",
        DAY=="4" ~ "first 5",
        DAY=="5" ~ "first 5",
    DAY=="27" ~ "last 5",
      DAY=="28" ~ "last 5",
      DAY=="29" ~ "last 5",
      DAY=="30" ~ "last 5",
      DAY=="31" ~ "last 5",)) %>%
  group_by(timeofmonth) %>%
  summarize(mean(TRAV_SP, na.rm=1))
```

```
## # A tibble: 3 x 2
##   timeofmonth `mean(TRAV_SP, na.rm = 1)`
##   <chr>                            <dbl>
## 1 first 5                           44.4
## 2 last 5                            52.8
## 3 <NA>                              42.3
```

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

```
df1 %>%
  mutate(
   timeofweek = case_when(
    DAY_WEEK=="Monday" ~ "weekday",
        DAY_WEEK=="Tuesday" ~ "weekday",
        DAY_WEEK=="Wednesday" ~ "weekday",
        DAY_WEEK=="Thursday" ~ "weekday",
        DAY_WEEK=="Friday" ~ "weekday",
    DAY_WEEK=="Saturday" ~ "weekend",
      DAY_WEEK=="Sunday" ~ "weekend")) %>%
  group_by(timeofweek) %>%
  summarize(mean(TRAV_SP, na.rm=1))
```

```
## # A tibble: 2 x 2
##   timeofweek `mean(TRAV_SP, na.rm = 1)`
##   <chr>                           <dbl>
## 1 weekday                          41.4
## 2 weekend                          48.8
```

5. Find the top 5 states with greatest travel speed.

```
df1 %>%
  group_by(STATE) %>%
  summarize(mean(TRAV_SP, na.rm=1)) %>%
  top_n(5)
```

```
## Selecting by mean(TRAV_SP, na.rm = 1)
```

```
## # A tibble: 5 x 2
##   STATE        `mean(TRAV_SP, na.rm = 1)`
##   <chr>                             <dbl>
## 1 Alabama                            57.6
## 2 Nevada                             73.5
## 3 North Dakota                       85
## 4 Rhode Island                       57
## 5 Wisconsin                          64
```

6. Rank the travel speed by MONTH.

```
df1 %>%
  mutate(avgTrav=mean(TRAV_SP),rank= dense_rank(desc(avgTrav))) %>%
   group_by(rank, MONTH) %>%
               summarize(mean(TRAV_SP, na.rm=1))
```

```
## # A tibble: 12 x 3
## # Groups:   rank [1]
##     rank MONTH     `mean(TRAV_SP, na.rm = 1)`
##    <int> <chr>                          <dbl>
##  1    NA April                           49.7
##  2    NA August                          44.6
##  3    NA December                        51.8
##  4    NA February                        36.4
##  5    NA January                         34.3
##  6    NA July                            35.7
##  7    NA June                            47.7
##  8    NA March                           35.8
##  9    NA May                             43.1
## 10    NA November                        49.4
## 11    NA October                         47.0
## 12    NA September                       48.0
```

7. Find the average speed of teenagers in December.

```
df1 %>%
  filter(MONTH=="December",AGE>12 & AGE<20) %>%
  summarize(mean(TRAV_SP, na.rm=1))
```

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

8. Find the month that female drivers drive fastest on average.

```
df1 %>%
  filter(SEX=="Female") %>%
  group_by(MONTH) %>%
  summarize(mean(TRAV_SP, na.rm=1)) %>%
  top_n(1)
```

```
## Selecting by mean(TRAV_SP, na.rm = 1)
```

```
## # A tibble: 1 x 2
##   MONTH    `mean(TRAV_SP, na.rm = 1)`
##   <chr>                         <dbl>
## 1 December                       60.3
```

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

```
df1 %>%
  filter(SEX=="Male") %>%
  group_by(MONTH) %>%
  summarize(mean(TRAV_SP, na.rm=1)) %>%
  top_n(-1)
```

```
## Selecting by mean(TRAV_SP, na.rm = 1)
```

```
## # A tibble: 1 x 2
##   MONTH   `mean(TRAV_SP, na.rm = 1)`
##   <chr>                        <dbl>
## 1 January                         34
```

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

```
unique(df1$INJ_SEV)
```

```
## [1] "Unknown"                     "No Apparent Injury (O)"
## [3] "Suspected Minor Injury(B)"   "Fatal Injury (K)"
## [5] "Suspected Serious Injury(A)" "Injured, Severity Unknown"
## [7] "Possible Injury (C)"
```

```
df1 %>%
  mutate(
   season = case_when(
      MONTH =="January" ~ "Winter",
       MONTH=="February" ~ "Winter",
       MONTH=="March" ~ "Spring",
        MONTH=="April" ~ "Spring",
        MONTH=="May" ~ "Spring",
        MONTH=="June" ~ "Summer",
        MONTH=="July" ~ "Summer",
        MONTH=="August" ~ "Summer",
        MONTH=="September" ~ "Fall",
        MONTH=="October" ~ "Fall",
       MONTH=="November" ~ "Fall",
       MONTH=="December" ~ "Winter")) %>%
  group_by(season) %>%
  summarize(prop.table(table(INJ_SEV))[4])
```

```
## # A tibble: 4 x 2
##   season `prop.table(table(INJ_SEV))[4]`
##   <chr>                            <dbl>
## 1 Fall                            0.0833
## 2 Spring                          0.101
## 3 Summer                          0.0860
## 4 Winter                          0.117
```

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

```
unique(df1$DEFORMED)
```

```
## [1] "Unknown"           "Functional Damage" "Minor Damage"
## [4] "Disabling Damage"  NA                  "Not Reported"
## [7] "No Damage"
```

```
df1 %>%
  group_by(DEFORMED) %>%
  summarize(prop.table(table(INJ_SEV))[4])
```

```
## # A tibble: 7 x 2
##   DEFORMED          `prop.table(table(INJ_SEV))[4]`
##   <chr>                                       <dbl>
## 1 Disabling Damage                           0.0980
## 2 Functional Damage                          0.0690
## 3 Minor Damage                               0.0132
## 4 No Damage                                 NA
## 5 Not Reported                               0.0455
## 6 Unknown                                    0.1
## 7 <NA>                                       0.0319
```