Assignment 3. Data Wrangling with Dplyr

Louis Dion

September 22, 2019

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:

Operators	Discription
!=	not equal to
!x	Not x
x | y	x OR y
x & y	x AND y

Read in titanic dataset, bring in dplyr

```
titanic<-read.csv(file='C:/Users/student/Documents/MATH421/data/titanic.csv', header=TRUE, sep=',')
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
```

#13. Mean age female passengers

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

```
##    femmean
## 1 27.91571
```

#14. Median Fare class 1 passengers

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

```
##   onefare
## 1 60.2875
```

#15. Median Fare non-class 1 Female passengers

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

```
##    femfare
## 1 14.45625
```

#16. Median age of survived passengers who are female and Class 1 or Class 2

```
titanic%>%filter(Survived==1 & Sex=='female' & Pclass!=3)%>%summarize(femage=median(Age,na.rm=TRUE))
```

```
##   femage
## 1     31
```

#17. Mean fare of female teenagers survived passengers

```
titanic%>%filter(Sex=='female' & Survived==1 & (Age>=13 & Age<20))%>%summarize(faremean=mean(Fare,na.rm
```

```
##   faremean
## 1 49.17966
```

#18. Mean fare of female teenagers survived passengers for each class

```
titanic%>%filter(Sex=='female' & Survived==1 & (Age>=13 & Age<20))%>%group_by(Pclass)%>%summarize(fareme
```

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

#19. Ratio of Survived and not Survived for passengers who are who pays more than the average fare

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

```
##      ratio
## 1 1.482353
```

```
#number of survived over number of not survived
#1.48 survived to 1 not survived
```

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

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

#21. 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.

```
newtitanic1<-newtitanic%>%mutate(cfare=cut(Fare,breaks=c(-Inf,mean(Fare,na.rm=TRUE),Inf),labels=c("Chea
```

#22. 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

```
ages<-c(0,9.99,19.99,29.99,39.99,49.99,59.99,69.99,79.99,89.99)
#allows lower limit of each group to be multiple of 10 - group "1" starts with 10, etc.
labels<-c(0,1,2,3,4,5,6,7,8)
newtitanic2<-newtitanic1%>%mutate(cage=cut(Age,breaks=ages,labels=labels))
```

#23. Show the frequency of Ports of Embarkation. It appears that there are two missing values in the Embarked variable. Assign the most frequent port to the missing ports.

```
frequency<-newtitanic2%>%mutate(Embarked=replace(Embarked,Embarked=='',"S"))%>%group_by(Embarked)%>%summ
frequency
```

```
## # A tibble: 3 x 2
##   Embarked  freq
##   <fct>    <int>
## 1 C          168
## 2 Q           77
## 3 S          646
```

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` #Read in c2015 dataset

```
library(readxl)
c2015<-read_excel('C:/Users/student/Documents/MATH421/data/c2015.xlsx')
```

#Number 4. Check dimension of data. Make new dataset with 1000 random observations. Use seed of 2019 so everyone in the class had the same dataset

```
dim(c2015)
```

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

```
set.seed(2019)
sample2015<-sample_n(c2015,1000)
```

#Number 5. Look at the data. One variable is a constant. Remove that variable from the data.

```
glimpse(sample2015)
```

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

```
sample2015<-sample2015%>%select(-YEAR)
glimpse(sample2015)
```

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

#Transform the Speed variable for the next set of questions

```
library(stringr)
names<-rownames(sample2015)
sample2015<-sample2015%>%filter(TRAV_SP!="Unknown",TRAV_SP!='Not Rep')%>%mutate(TRAV_SP=sapply(strsplit
```

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

```
sample2015
```

```
## # A tibble: 371 x 27
##    STATE ST_CASE VEH_NO PER_NO COUNTY   DAY MONTH  HOUR MINUTE AGE   SEX
##    <chr>   <dbl>  <dbl>  <dbl>  <dbl> <dbl> <chr> <dbl>  <dbl> <chr> <chr>
##  1 Ariz~   40327      1      1     13     7 May      22     15 47    Fema~
##  2 Minn~  270119      2      4     59    16 May      21     59 15    Fema~
##  3 Miss~  290576      1      1    201     2 Octo~    15     38 55    Male
##  4 Cali~   62865      1      1     19     6 June     15     20 56    Male
##  5 Sout~  450153      1      1     29    19 March    14     15 54    Male
##  6 Alab~   10239      1      5     61     9 May      18     55 10    Fema~
##  7 Nort~  370294      1      2    183     4 April    10     14 15    Fema~
##  8 Cali~   60153      1      1     53    29 Janu~    22     15 56    Male
##  9 Wisc~  550300      2      1      7    21 Augu~    16     10 79    Male
## 10 Flor~  121999      2      1     57    21 Octo~     6     28 53    Male
## # ... with 361 more rows, and 16 more variables: PER_TYP <chr>,
## #   INJ_SEV <chr>, SEAT_POS <chr>, DRINKING <chr>, MAN_COLL <chr>,
## #   OWNER <chr>, MOD_YEAR <chr>, TRAV_SP <dbl>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

#Number 11. Compare the average speed of those who had “No Apprent Injury” and the rest. What do you observe?

```
noinjury<-sample2015%>%filter(INJ_SEV=='No Apparent Injury (O)')%>%summarize(mean=mean(TRAV_SP,na.rm=TRU
injury<-sample2015%>%filter(INJ_SEV!='No Apparent Injury (O)')%>%summarize(mean=mean(TRAV_SP,na.rm=TRUE)
noinjury
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  44.6
```

```
#no injury went an average speed of 44.63 mph
injury
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  53.1
```

```
#injury went an average speed of 53.09 mph.
#We observe that accidents with injuries have higher average car speeds than accidents without injuries
```

#Number 12. 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.

```
maledriver<-sample2015%>%filter(SEAT_POS=="Front Seat, Left Side",SEX=="Male")%>%summarize(mean=mean(TRA
femaledriver<-sample2015%>%filter(SEAT_POS=="Front Seat, Left Side",SEX=="Female")%>%summarize(mean=mean
maledriver
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  51.7
```

```
#The average speed of male drivers is 51.65 mph.
femaledriver
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  46.1
```

```
#The average speed of female drivers is 46.07 mph.
#We observe that male drivers in accidents drive faster on average than female drivers in accidents.
```

#Number 13. Compare the average speed of drivers who drink and those who do not. Comment on the results.

```
nodrink<-sample2015%>%filter(SEAT_POS=="Front Seat, Left Side",DRINKING=="No (Alcohol Not Involved)")%>%
drink<-sample2015%>%filter(SEAT_POS=="Front Seat, Left Side",DRINKING=="Yes (Alcohol Involved)")%>%summa
nodrink
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  44.9
```

```
#The average speed for non-drinkers is 44.94 mph.
drink
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  68.2
```

```
#The average speed for drinkers is 68.25 mph.
#We observe that people in accidents who had drank alcohol drive faster on average than people in accid
```

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.

```
days<-sample2015%>%group_by(DAY)%>%summarize(mean=mean(TRAV_SP,na.rm=TRUE))
mean(days[1:5,]$mean)
```

```
## [1] 52.41238
```

```
#The mean driving speed for the first five days of the month is 52.41 mph.
mean(days[27:31,]$mean)
```

```
## [1] 53.74722
```

```
#The mean driving speed for the last five days of the month is 53.74 mph.
#We observe people in accidents tend to drive faster on the last five days of the month on average comp
```

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

```
weekday=c('Monday','Tuesday','Wednesday','Thursday','Friday')
weekend=c('Saturday','Sunday')
week<-sample2015%>%group_by(DAY_WEEK)%>%summarize(mean=mean(TRAV_SP,na.rm=TRUE))
week #shows where Saturday and Sunday are in the dataframe
```

```
## # A tibble: 7 x 2
##   DAY_WEEK   mean
##   <chr>     <dbl>
## 1 Friday     50.7
## 2 Monday     48.6
## 3 Saturday   53.3
## 4 Sunday     55.8
## 5 Thursday   50.8
## 6 Tuesday    47.2
## 7 Wednesday  44.7
```

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

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

```
#The mean driving speed for my definition of weekend is 54.53 mph.
mean(week[c(1:2,5:7),]$mean)
```

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

```
#The mean driving speed for my definition of weekday is 48.40 mph.
#We observe that people in accidents on weekend days drive faster on average than people in accidents o
```

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

```
states<-sample2015%>%group_by(STATE)%>%summarize(mean=mean(TRAV_SP,na.rm=TRUE))%>%arrange(desc(mean))
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
```

```
#The top 5 states with greatest travel speeds are South Dakota, North Dakota, Nevada, Wyoming, and Kent
```

6. Rank the travel speed by MONTH.

```
month<-sample2015%>%group_by(MONTH)%>%summarize(mean=mean(TRAV_SP,na.rm=TRUE))%>%arrange(desc(mean))
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.

```
decteen<-sample2015%>%filter(AGE>=13,AGE<20,MONTH=='December')%>%summarize(mean=mean(TRAV_SP,na.rm=TRUE
decteen
```

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

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

```
femmonth<-sample2015%>%filter(SEX=='Female',SEAT_POS=="Front Seat, Left Side")%>%group_by(MONTH)%>%summa
femmonth
```

```
## # A tibble: 12 x 2
##    MONTH      mean
##    <chr>     <dbl>
##  1 September  75.7
##  2 July       65
##  3 April      54
##  4 December   53.3
##  5 May        52
##  6 June       49.6
##  7 October    41.5
##  8 March      38.7
##  9 November   37.7
## 10 August     37.6
## 11 January    35
## 12 February  NaN
```

```
#September is the month female drivers drive fastest on average
```

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

```
malemonth<-sample2015%>%filter(SEX=='Male',SEAT_POS=="Front Seat, Left Side")%>%group_by(MONTH)%>%summar
malemonth
```

```
## # A tibble: 12 x 2
##    MONTH      mean
##    <chr>     <dbl>
##  1 February   36.2
##  2 July       38
##  3 March      42.1
##  4 January    48.2
##  5 May        50.1
##  6 December   50.6
##  7 September  52.3
##  8 June       54.5
##  9 October    56.5
## 10 November   57
## 11 August     57.5
## 12 April      61.4
```

```
#February is the month male drivers drive slowest on average.
```

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','February'=
fatal<-sample2015%>%group_by(seasons)%>%summarize(percentage=sum(INJ_SEV=='Fatal Injury (K)')/n())
fatal
```

```
## # A tibble: 4 x 2
##   seasons percentage
##   <chr>        <dbl>
## 1 Autumn       0.432
## 2 Spring       0.268
## 3 Summer       0.330
## 4 Winter       0.25
```

```
#The season with the most percentage of fatal accidents is Autumn, and winter is the season with the le
```

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

```
deformfatal<-sample2015%>%group_by(DEFORMED)%>%summarize(percentage=sum(INJ_SEV=='Fatal Injury (K)')/n(
deformfatal
```

```
## # A tibble: 6 x 2
##   DEFORMED          percentage
##   <chr>                  <dbl>
## 1 Disabling Damage      0.435
## 2 Functional Damage     0.0833
## 3 Minor Damage          0.0303
## 4 No Damage             0
## 5 Not Reported          0.2
## 6 Unknown               0
```

```
#Accidents with Disabling Damage have the highest percentage of fatal injuries. As the amount of damage
```