DW-Exercise2

Diane K

September 7, 2016

0. Load data to data frame

```
titanic_dataset <- read.csv(file="titanic_original.csv"
                        , head = TRUE, sep=",")
str(titanic_dataset)
```

```
## 'data.frame':    1310 obs. of  14 variables:
##  $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : Factor w/ 1308 levels "","Abbing, Mr. Anthony",..: 23 25 26 27 28 32 47 48 52 56 ...
##  $ sex      : Factor w/ 3 levels "","female","male": 2 3 2 3 2 3 2 3 2 3 ...
##  $ age      : num  29 0.917 2 30 25 ...
##  $ sibsp    : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch    : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : Factor w/ 930 levels "","110152","110413",..: 189 51 51 51 51 126 94 17 78 827 ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...
##  $ embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...
##  $ boat     : Factor w/ 28 levels "","1","10","11",..: 13 4 1 1 1 14 3 1 28 1 ...
##  $ body     : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...
```

```
kable(head(titanic_dataset[,1:5]) , format = "markdown", caption = "Titanic dataset")
```

pclass	survived	name	sex	age
1	1	Allen, Miss. Elisabeth Walton	female	29.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
1	0	Allison, Miss. Helen Loraine	female	2.0000
1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000
1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000
1	1	Anderson, Mr. Harry	male	48.0000

```
kable(head(titanic_dataset[,6:14]) , format = "markdown")
```

sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	0	24160	211.3375	B5	S	2	NA	St Louis, MO
1	2	113781	151.5500	C22 C26	S	11	NA	Montreal, PQ / Chesterville, O
1	2	113781	151.5500	C22 C26	S		NA	Montreal, PQ / Chesterville, O
1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville, O
1	2	113781	151.5500	C22 C26	S		NA	Montreal, PQ / Chesterville, O
0	0	19952	26.5500	E12	S	3	NA	New York, NY

1. Replace missing values in embarked with S

```
if( is.factor(titanic_dataset$embarked)){
  levels(titanic_dataset$embarked)[1] <- "S"
}else{
  titanic_dataset$embarked[ is.na(titanic_dataset$embarked) |
                              titanic_dataset$embarked == ""] <- "S"
}
```

2.a Replace missing age with mean age

```
av_age <- mean( titanic_dataset$age, na.rm = TRUE)
titanic_dataset$age[ is.na(titanic_dataset$age) ] <- av_age
```

2.b

- Could not replace with 0, because it's a wrong age for someone alive
- Could not replace with a negative value because it might impact calculations on the age column

3. Replace missing values in LifeBoat column with "None"

```
if( is.factor(titanic_dataset$boat)){
  levels(titanic_dataset$boat)[1] <- "None"
}else{
  titanic_dataset$boat[is.na(titanic_dataset$boat) ] <- "None"
}
```

4. Add flag has_cabin_number

We should not replace missing cabin numbers because not all passengers purchased cabins

```
titanic_dataset <- titanic_dataset %>% mutate(has_cabin_number = ifelse(
  is.na(titanic_dataset$cabin) | (titanic_dataset$cabin) == "", 0, 1) )
```

Output

```
str(titanic_dataset)
```

```
## 'data.frame':    1310 obs. of  15 variables:
##  $ pclass          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived        : int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name            : Factor w/ 1308 levels "","Abbing, Mr. Anthony",..: 23 25 26 27 28 32 47 48 52 5
##  $ sex             : Factor w/ 3 levels "","female","male": 2 3 2 3 2 3 2 3 2 3 ...
##  $ age             : num  29 0.917 2 30 25 ...
```

```
##  $ sibsp           : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch           : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket          : Factor w/ 930 levels "","110152","110413",..: 189 51 51 51 51 126 94 17 78 827
##  $ fare            : num  211 152 152 152 152 ...
##  $ cabin           : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...
##  $ embarked        : Factor w/ 3 levels "S","C","Q": 1 1 1 1 1 1 1 1 1 2 ...
##  $ boat            : Factor w/ 28 levels "None","1","10",..: 13 4 1 1 1 14 3 1 28 1 ...
##  $ body            : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest       : Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 23
##  $ has_cabin_number: num  1 1 1 1 1 1 1 1 1 0 ...
```

```
kable(head(titanic_dataset[,1:5]) , format = "markdown", caption = "Titanic dataset cleaned")
```

pclass	survived	name	sex	age
1	1	Allen, Miss. Elisabeth Walton	female	29.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
1	0	Allison, Miss. Helen Loraine	female	2.0000
1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000
1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000
1	1	Anderson, Mr. Harry	male	48.0000

```
kable(head(titanic_dataset[,6:12]) , format = "markdown")
```

sibsp	parch	ticket	fare	cabin	embarked	boat
0	0	24160	211.3375	B5	S	2
1	2	113781	151.5500	C22 C26	S	11
1	2	113781	151.5500	C22 C26	S	None
1	2	113781	151.5500	C22 C26	S	None
1	2	113781	151.5500	C22 C26	S	None
0	0	19952	26.5500	E12	S	3

```
kable(head(titanic_dataset[,13:15]) , format = "markdown")
```

body	home.dest	has_cabin_number
NA	St Louis, MO	1
NA	Montreal, PQ / Chesterville, ON	1
NA	Montreal, PQ / Chesterville, ON	1
135	Montreal, PQ / Chesterville, ON	1
NA	Montreal, PQ / Chesterville, ON	1
NA	New York, NY	1

```
write.csv(titanic_dataset, file = "titanic_clean.csv"
          , row.names = FALSE, append = FALSE)
```

```
## Warning in write.csv(titanic_dataset, file = "titanic_clean.csv", row.names
## = FALSE, : attempt to set 'append' ignored
```