Homework1

R26131141 Yenwei, Luo

2025-02-26

```
library(Hmisc)

titanic$Survived<-factor(titanic$Survived)
titanic$Pclass<-factor(titanic$Pclass)
titanic$Sex<-factor(titanic$Sex)
titanic$Embarked<-factor(titanic$Embarked)
titanic$CabinType <- factor(substr(titanic$Cabin, 1, 1))

describe(titanic)
```

```
titanic

 13  Variables      891  Observations
--------------------------------------------------------------------------------
PassengerId
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     891        0      891        1      446      446    297.3     45.5
     .10      .25      .50      .75      .90      .95
    90.0    223.5    446.0    668.5    802.0    846.5

lowest :   1   2   3   4   5, highest: 887 888 889 890 891
--------------------------------------------------------------------------------
Survived
       n  missing distinct
     891        0        2

Value          0     1
Frequency    549   342
Proportion 0.616 0.384
--------------------------------------------------------------------------------
Pclass
       n  missing distinct
```

```
     891        0        3

Value          1     2     3
Frequency    216   184   491
Proportion 0.242 0.207 0.551
---------------------------------------------------------------------------
Name
       n  missing distinct
     891        0      891

lowest : Abbing, Mr. Anthony                    Abbott, Mr. Rossmore Edward           Abbott,
highest: Yousseff, Mr. Gerious                  Yrois, Miss. Henriette ("Mrs Harbeck") Zabour,
---------------------------------------------------------------------------
Sex
       n  missing distinct
     891        0        2

Value      female   male
Frequency     314    577
Proportion  0.352  0.648
---------------------------------------------------------------------------
Age
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     714      177       88    0.999     29.7       29    16.21     4.00
     .10      .25      .50      .75      .90      .95
   14.00    20.12    28.00    38.00    50.00    56.00

lowest : 0.42 0.67 0.75 0.83 0.92, highest: 70   70.5 71   74   80
---------------------------------------------------------------------------
SibSp
       n  missing distinct     Info     Mean  pMedian      Gmd
     891        0        7    0.669    0.523      0.5    0.823

Value          0     1     2     3     4     5     8
Frequency    608   209    28    16    18     5     7
Proportion 0.682 0.235 0.031 0.018 0.020 0.006 0.008

For the frequency table, variable is rounded to the nearest 0
---------------------------------------------------------------------------
Parch
       n  missing distinct     Info     Mean  pMedian      Gmd
     891        0        7    0.556   0.3816        0   0.6259

Value          0     1     2     3     4     5     6
Frequency    678   118    80     5     4     5     1
Proportion 0.761 0.132 0.090 0.006 0.004 0.006 0.001

For the frequency table, variable is rounded to the nearest 0
```

```
---------------------------------------------------------------------------
Ticket
       n  missing distinct
     891        0      681

lowest : 110152      110413      110465      110564      110813
highest: W./C. 6608  W./C. 6609  W.E.P. 5734 W/C 14208   WE/P 5735
---------------------------------------------------------------------------
Fare
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     891        0      248        1     32.2     19.6    36.78    7.225
     .10      .25      .50      .75      .90      .95
   7.550    7.910   14.454   31.000   77.958  112.079

lowest : 0       4.0125  5       6.2375  6.4375
highest: 227.525 247.521 262.375 263     512.329
---------------------------------------------------------------------------
Cabin
       n  missing distinct
     204      687      147

lowest : A10 A14 A16 A19 A20, highest: F33 F38 F4  G6  T
---------------------------------------------------------------------------
Embarked
       n  missing distinct
     889        2        3

Value          C     Q     S
Frequency    168    77   644
Proportion 0.189 0.087 0.724
---------------------------------------------------------------------------
CabinType
       n  missing distinct
     204      687        8

Value          A     B     C     D     E     F     G     T
Frequency     15    47    59    33    32    13     4     1
Proportion 0.074 0.230 0.289 0.162 0.157 0.064 0.020 0.005
---------------------------------------------------------------------------
```

```r
summary(titanic)
```

```
  PassengerId    Survived Pclass      Name              Sex
 Min.   :  1.0   0:549    1:216   Length:891         female:314
 1st Qu.:223.5   1:342    2:184   Class :character   male  :577
 Median :446.0            3:491   Mode  :character
 Mean   :446.0
```

```
3rd Qu.:668.5
Max.   :891.0

     Age            SibSp           Parch           Ticket
Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891
1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   Class :character
Median :28.00   Median :0.000   Median :0.0000   Mode  :character
Mean   :29.70   Mean   :0.523   Mean   :0.3816
3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
Max.   :80.00   Max.   :8.000   Max.   :6.0000
NA's   :177
     Fare           Cabin           Embarked     CabinType
Min.   :  0.00   Length:891         C   :168   C      : 59
1st Qu.:  7.91   Class :character   Q   : 77   B      : 47
Median : 14.45   Mode  :character   S   :644   D      : 33
Mean   : 32.20                      NA's:  2   E      : 32
3rd Qu.: 31.00                                 A      : 15
Max.   :512.33                                 (Other): 18
                                               NA's   :687
```

Short conclusion for the titanic dataset

[Let the “Survived”, “Pclass”, “Sex”, and “Embarked” variables become factors, and consider retaining only the first letter of the cabin (representing the cabin category)]

As we can see in the summary of the dataset “titanic”, there are many missing values in the variables “Embarked”, “Age” and “Cabin”. Therefore, we need to do the subsequent analysis to get more information for this dataset