

# hw1 (data description)

Yuzhen Liu

## 目錄

Variable definition . . . . .	1
Data discription . . . . .	1

## Variable definition

According to the information from Kaggle (<https://www.kaggle.com/c/titanic/overview>), the detailed data coding book is presented in Figure 1.

Variable	Data Type	Definition	Note
PassengerId	character	A unique identifier for each passenger in the dataset.	
survival	factor	Survival	0 = No, 1 = Yes
pclass	ordinal	Ticket class	1 = upper, 2 = middle, 3 = lower
sex	factor	Sex	female; male
Name	character	Passengers' names	
Age	numerical	Age in years	Age is fractional if less than 1.
sibsp	numerical	# of siblings / spouses aboard the Titanic	<ul style="list-style-type: none"><li>Only legally recognized spouses are considered.</li><li>Siblings, including brother, sister, stepbrother, and stepsister, are considered.</li></ul>
parch	numerical	# of parents / children aboard the Titanic	Some children travelled only with a nanny, therefore parch=0 for them.
ticket	character	Ticket number	
fare	numerical	Passenger fare	
cabin	character	Cabin number	
embarked	factor	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

圖 1: Data dictionary

## Data discription

In this dataset, 61.6% of passengers did not survive, while 38.4% did. The socio-economic status, represented by `Pclass`, indicates that 55.1% of passengers belonged to the lower class, 20.7% to the middle class, and 24.2% to the upper class.

Regarding gender distribution, 64.8% of the passengers were male, while 35.2% were female. The average age of passengers was approximately 29.7 years; however, there were 177 missing values for age.

Additionally, the `Sibsp` and `Parch` variables were treated as factor variables. Among the passengers, 68.2% traveled alone without siblings or spouses, while 23.5% had one sibling or spouse. Regarding parents or children, 76.1% of passengers traveled without them, and 13.2% had one parent or child accompanying them.

Ticket prices varied significantly, with an average fare of 32.2 and a maximum fare of 512.33. Most passengers embarked at the port of Southampton, while fewer departed from Cherbourg and Queenstown.

```
# R Interface to Python
library(reticulate)
library(Hmisc)
library(dplyr)
df<-read.csv("titanic.csv")
df <- df %>% mutate(across(c(Survived, Pclass, Sex, Embarked), as.factor))
df$Pclass <- factor(df$Pclass, levels = c(3, 2, 1), ordered = TRUE)
df <- df %>% mutate(across(c( PassengerId, Name, Ticket, Cabin), as.character))
latex(describe(df), descript = "Descriptive Statistics",
      file = '', caption.placement = "top")
```

12 Variables												df	891 Observations																																																			
PassengerId																																																																
n					missing					distinct																																																						
891					0					891																																																						
lowest : 1 10 100 101 102, highest: 95 96 97 98 99																																																																
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					3					2					1																																																	
Frequency					491					184					216																																																	
Proportion					0.551					0.207					0.242																																																	
Name																																																																
n					missing					distinct																																																						
891					0					891																																																						
lowest : Abbing, Mr. Anthony Abbott, Mr. Rossmore Edward Abbott, Mrs. Stanton (Rosa Hunt)																																																																
highest: Yousseff, Mr. Gerious Yrois, Miss. Henriette ("Mrs Harbeck") Zabour, Miss. Hileni																																																																
Sex																																																																
n					missing					distinct																																																						
891					0					2																																																						
Value					female					male																																																						
Frequency					314					577																																																						
Proportion					0.352					0.648																																																						
Age																																																																
n					missing					distinct					Info					Mean					Gmd					.05					.10					.25					.50					.75					.90					.95				
714					177					88					0.999					29.7					16.21					4.00					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	Gmd		
	891	0	7	0.669	0.523	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	Gmd		
	891	0	7	0.556	0.3816	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
highest:	W./C. 6608		W./C. 6609
			110465
			W.E.P. 5734
			110564
			W/C 14208
			110813
			WE/P 5735

## Fare

	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	891	0	248	1	32.2	36.78	7.225	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
	891	0	4
Value			C Q S
Frequency		2	168 77 644
Proportion		0.002	0.189 0.086 0.723