Titanic

2022-11

Titanic

Trabajaremos con el Dataset **Titanic**

Table 1: Primeros registros de la base de datos Titanic

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	NA	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	NA	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	NA	S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583	NA	Q

Data Frame Summary

TitanicT Dimensions: 891 x 12
Duplicates: 0

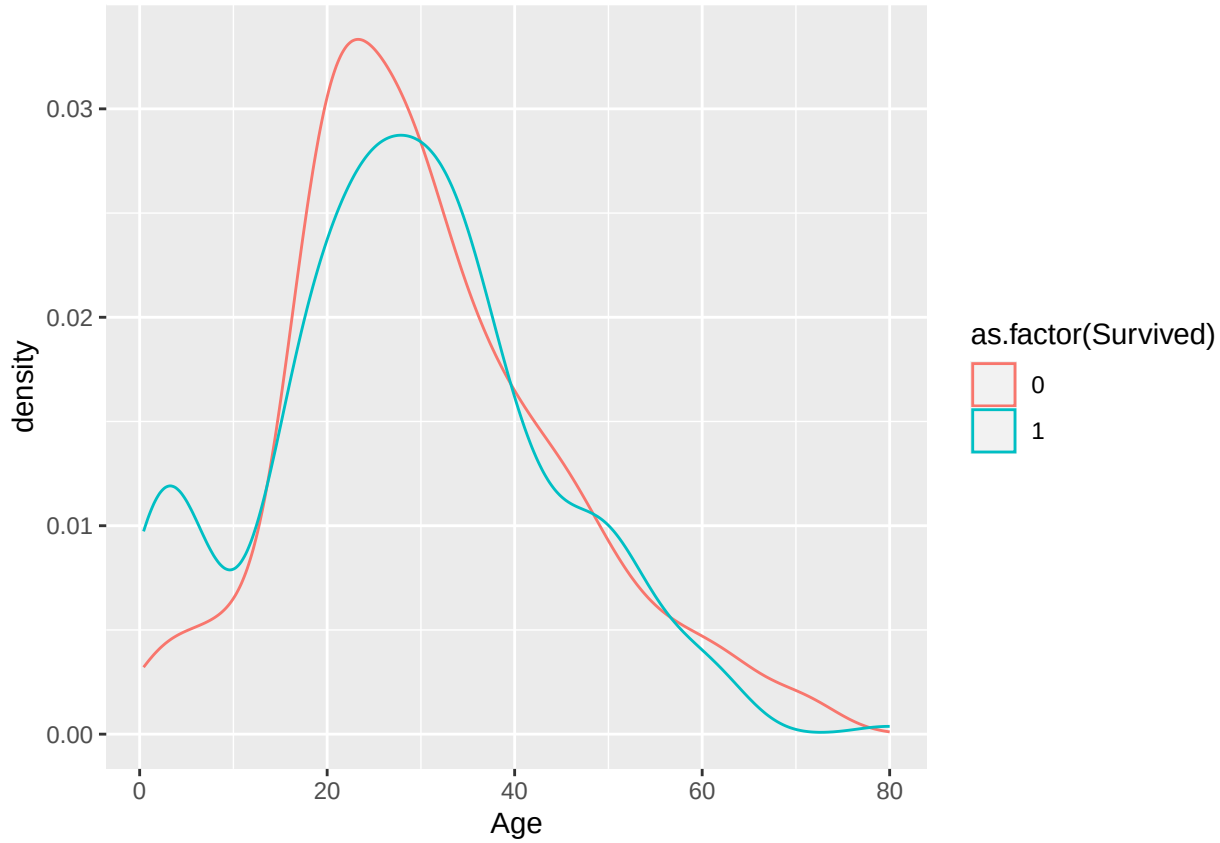
No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	PassengerId [numeric]	Mean (sd) : 446 (257.4) min < med < max: 1 < 446 < 891 IQR (CV) : 445 (0.6)	891 distinct values	891 (100.0%)	0 (0.0%)
2	Survived [numeric]	Min : 0 Mean : 0.4 Max : 1	0 : 549 (61.6%) 1 : 342 (38.4%)	891 (100.0%)	0 (0.0%)
3	Pclass [numeric]	Mean (sd) : 2.3 (0.8) min < med < max: 1 < 3 < 3 IQR (CV) : 1 (0.4)	1 : 216 (24.2%) 2 : 184 (20.7%) 3 : 491 (55.1%)	891 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
4	Name	1. Abbing, Mr. Anthony	1 (0.1%)	891	0
	[character]	2. Abbott, Mr. Rossmore	1 (0.1%)	(100.0%)	(0.0%)
		Edwa	1 (0.1%)		
		3. Abbott, Mrs. Stanton (Ros	1 (0.1%)		
		4. Abelson, Mr. Samuel	1 (0.1%)		
		5. Abelson, Mrs. Samuel (Han	886 (99.4%)		
		[886 others]			
5	Sex	1. female	314 (35.2%)	891	0
	[character]	2. male	577 (64.8%)	(100.0%)	(0.0%)
6	Age	Mean (sd) : 29.7 (14.5)	88 distinct values	714	177
	[numeric]	min < med < max:		(80.1%)	(19.9%)
		0.4 < 28 < 80			
		IQR (CV) : 17.9 (0.5)			
7	SibSp	Mean (sd) : 0.5 (1.1)	7 distinct values	891	0
	[numeric]	min < med < max:		(100.0%)	(0.0%)
		0 < 0 < 8			
		IQR (CV) : 1 (2.1)			
8	Parch	Mean (sd) : 0.4 (0.8)	7 distinct values	891	0
	[numeric]	min < med < max:		(100.0%)	(0.0%)
		0 < 0 < 6			
		IQR (CV) : 0 (2.1)			
9	Ticket	1. 1601	7 (0.8%)	891	0
	[character]	2. 347082	7 (0.8%)	(100.0%)	(0.0%)
		3. CA. 2343	7 (0.8%)		
		4. 3101295	6 (0.7%)		
		5. 347088	6 (0.7%)		
		[676 others]	858 (96.3%)		
10	Fare	Mean (sd) : 32.2 (49.7)	248 distinct values	891	0
	[numeric]	min < med < max:		(100.0%)	(0.0%)
		0 < 14.5 < 512.3			
		IQR (CV) : 23.1 (1.5)			
11	Cabin	1. B96 B98	4 (2.0%)	204	687
	[character]	2. C23 C25 C27	4 (2.0%)	(22.9%)	(77.1%)
		3. G6	4 (2.0%)		
		4. C22 C26	3 (1.5%)		
		5. D	3 (1.5%)		
		[142 others]	186 (91.2%)		
12	Embarked	1. C	168 (18.9%)	889	2
	[character]	2. Q	77 (8.7%)	(99.8%)	(0.2%)
		3. S	644 (72.4%)		

1. ¿La varianza de las edades de quienes sobrevivieron es diferente para ambos grupos?
2. Con base en la respuesta anterior, prueba si las edades promedio son iguales o diferentes para quienes sobrevivieron o no?

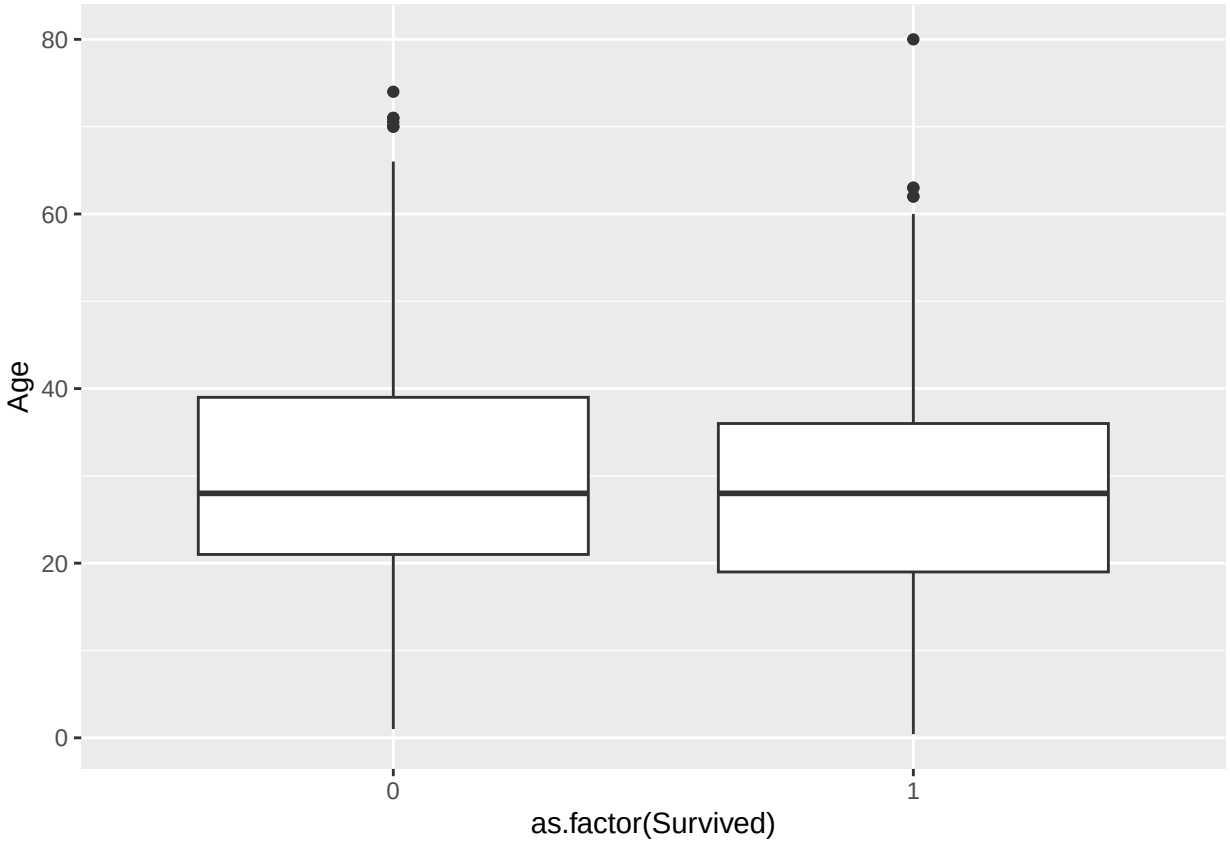
```
TitanicT %>% ggplot(aes(x=Age,colour=as.factor(Survived),group=Survived)) +
  geom_density()
```

```
## Warning: Removed 177 rows containing non-finite values (‘stat_density()‘).
```




```
TitanicT %>% ggplot(aes(x=as.factor(Survived), y=Age)) +
  geom_boxplot()
```

```
## Warning: Removed 177 rows containing non-finite values (‘stat_boxplot()‘).
```




```
TitanicT %>% group_by(Survived) %>%  summarise(varAge = var(Age,na.rm = T))
```

```
## # A tibble: 2 x 2
##   Survived varAge
##      <dbl>  <dbl>
## 1        0   201.
## 2        1   224.
```

```
stats::var.test( Age ~ as.factor(Survived), data = TitanicT, conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  Age by as.factor(Survived)
## F = 0.89853, num df = 423, denom df = 289, p-value = 0.317
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7253979 1.1082231
## sample estimates:
## ratio of variances
##          0.8985274
```

$H_0 : \mu_{age0} = \mu_{age1}, \mu_{age0} - \mu_{age1} = 0$

```r
t.test(Age ~ Survived, data = TitanicT, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Survived
## t = 2.046, df = 598.84, p-value = 0.04119
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.09158472 4.47339446
## sample estimates:
## mean in group 0 mean in group 1
##        30.62618        28.34369
```

```r
TitanicT %>% t_test(Age ~ Survived)
```

```
## Warning: The statistic is based on a difference or ratio; by default, for
## difference-based statistics, the explanatory variable is subtracted in the
## order "0" - "1", or divided in the order "0" / "1" for ratio-based statistics.
## To specify this order yourself, supply `order = c("0", "1")`.
```

```
## # A tibble: 1 x 7
##   statistic  t_df p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      2.05  599.  0.0412 two.sided       2.28   0.0916     4.47
```

Se dice que aproximadamente una tercera parte de la gente sobrevivió en el Titanic, estos datos respaldan esta afirmación

```r
prop.test(sum(TitanicT$Survived),891,p=1/3)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(TitanicT$Survived) out of 891, null probability 1/3
## X-squared = 10.001, df = 1, p-value = 0.001564
## alternative hypothesis: true p is not equal to 0.3333333
## 95 percent confidence interval:
##  0.3519194 0.4167722
## sample estimates:
##         p
## 0.3838384
```

```r
TitanicT %>% mutate(Survived=as.factor(Survived)) %>% prop_test(Survived ~ NULL,p=1/3)
```

```
## # A tibble: 1 x 4
##   statistic chisq_df  p_value alternative
##       <dbl>    <int>    <dbl> <chr>
## 1      319.        1 1.90e-71 two.sided
```

¿Sobrevivió la misma proporción de hombres y de mujeres o se aplicó lo de "mujeres y niños primero"?

```r
prop.test(sum(TitanicT$Sex=="female"),891,p=.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(TitanicT$Sex == "female") out of 891, null probability 0.5
## X-squared = 77.042, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3211923 0.3849235
## sample estimates:
##        p
## 0.352413
```

```r
TitanicT %>%  prop_test(Sex ~ NULL,p=.5, order = c("female","male"))
```

```
## # A tibble: 1 x 4
##   statistic chisq_df  p_value alternative
##       <dbl>    <int>    <dbl> <chr>
## 1      77.0        1 1.67e-18 two.sided
```

```r
TitanicT %>%  prop_test(Sex ~ NULL,p=.5, success =  "male", z=TRUE)
```

```
## # A tibble: 1 x 3
##   statistic  p_value alternative
##       <dbl>    <dbl> <chr>
## 1      8.81 1.24e-18 two.sided
```

```r
(tabla<- with(TitanicT,addmargins(table(Sex,Survived))))
```

```
##         Survived
## Sex        0   1 Sum
##   female  81 233 314
##   male   468 109 577
##   Sum    549 342 891
```

```r
with(TitanicT,prop.table(table(Sex,Survived),margin = 1))
```

```
##         Survived
## Sex              0         1
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

```r
prop.test(x=c(233,109),n=c(314,577),alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(233, 109) out of c(314, 577)
```

```
## X-squared = 260.72, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.5020113 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7420382 0.1889081
```

```
TitanicT %>%
  group_by(Sex) %>%
  summarise(
    p_hat=mean(Survived),
    n=n()
  )
```

```
## # A tibble: 2 x 3
##   Sex    p_hat     n
##   <chr>  <dbl> <int>
## 1 female 0.742   314
## 2 male   0.189   577
```

```
TitanicT %>%  mutate(Survived=as.factor(Survived)) %>%
  prop_test(Survived ~ Sex,
            order = c("female","male"),
            success = "1",
            alternative = "greater",
            correct = F)
```

```
## # A tibble: 1 x 6
##   statistic chisq_df  p_value alternative lower_ci upper_ci
##       <dbl>    <dbl>    <dbl> <chr>          <dbl>    <dbl>
## 1      263.        1 1.86e-59 greater        0.485        1
```

```
(tabla<- with(TitanicT,addmargins(table(Pclass,Survived))))
```

```
##       Survived
## Pclass   0   1 Sum
##    1    80 136 216
##    2    97  87 184
##    3   372 119 491
##    Sum 549 342 891
```

```
TitanicT %>%
  mutate(Survived=as.factor(Survived),
         Pclass = as.factor(Pclass)) %>%
  chisq_test(Survived ~ Pclass)
```

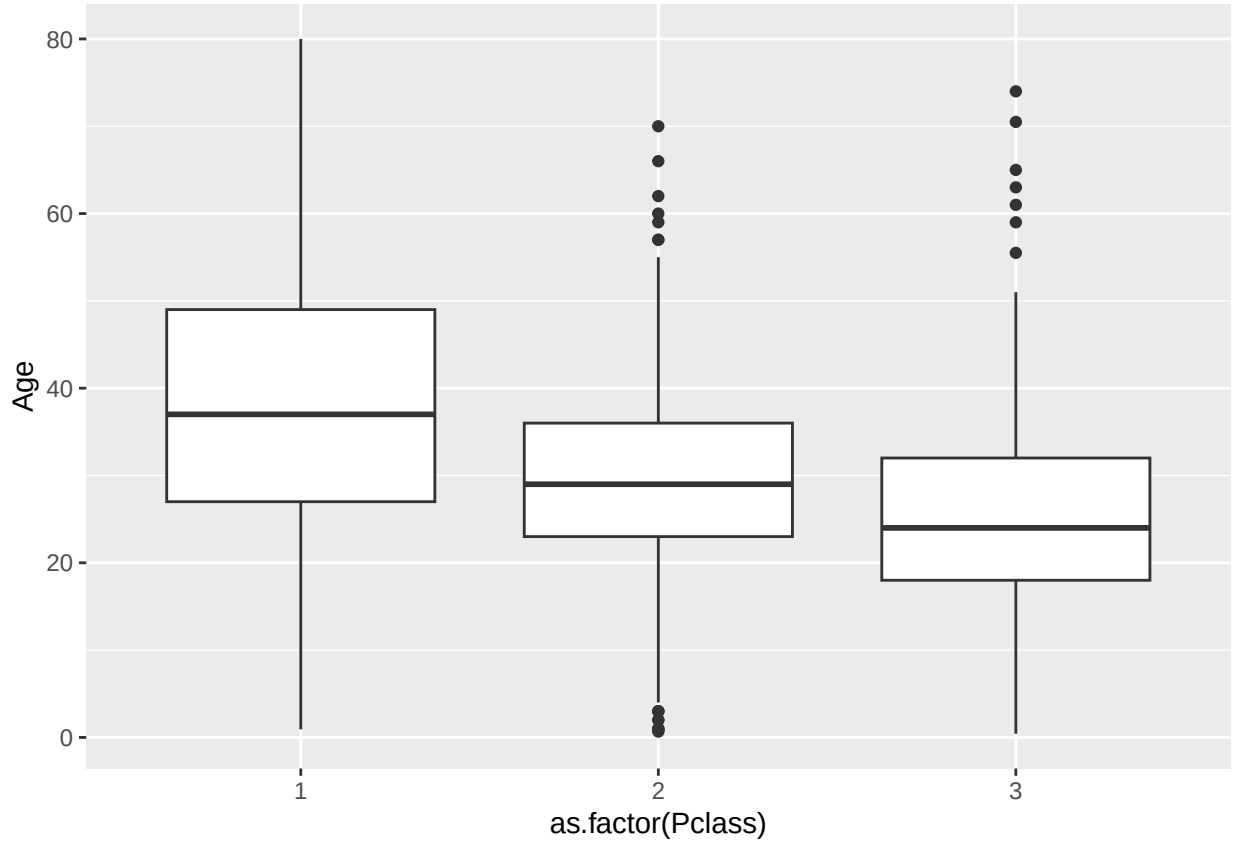
```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      103.        2 4.55e-23
```

```
TitanicT %>%
  mutate(Survived=as.factor(Survived),
         Pclass = as.factor(Pclass)) %>%
  chisq_test(Pclass ~ Survived)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <int>    <dbl>
## 1      103.        2 4.55e-23
```

```
TitanicT %>%
  ggplot(aes(x=as.factor(Pclass), y = Age)) +
  geom_boxplot()
```

```
## Warning: Removed 177 rows containing non-finite values (‘stat_boxplot()‘).
```




```
TitanicT %>% group_by(Pclass) %>%  summarise(meanAge = mean(Age,na.rm = T))
```

```
## # A tibble: 3 x 2
##   Pclass meanAge
##    <dbl>   <dbl>
## 1      1    38.2
## 2      2    29.9
## 3      3    25.1
```

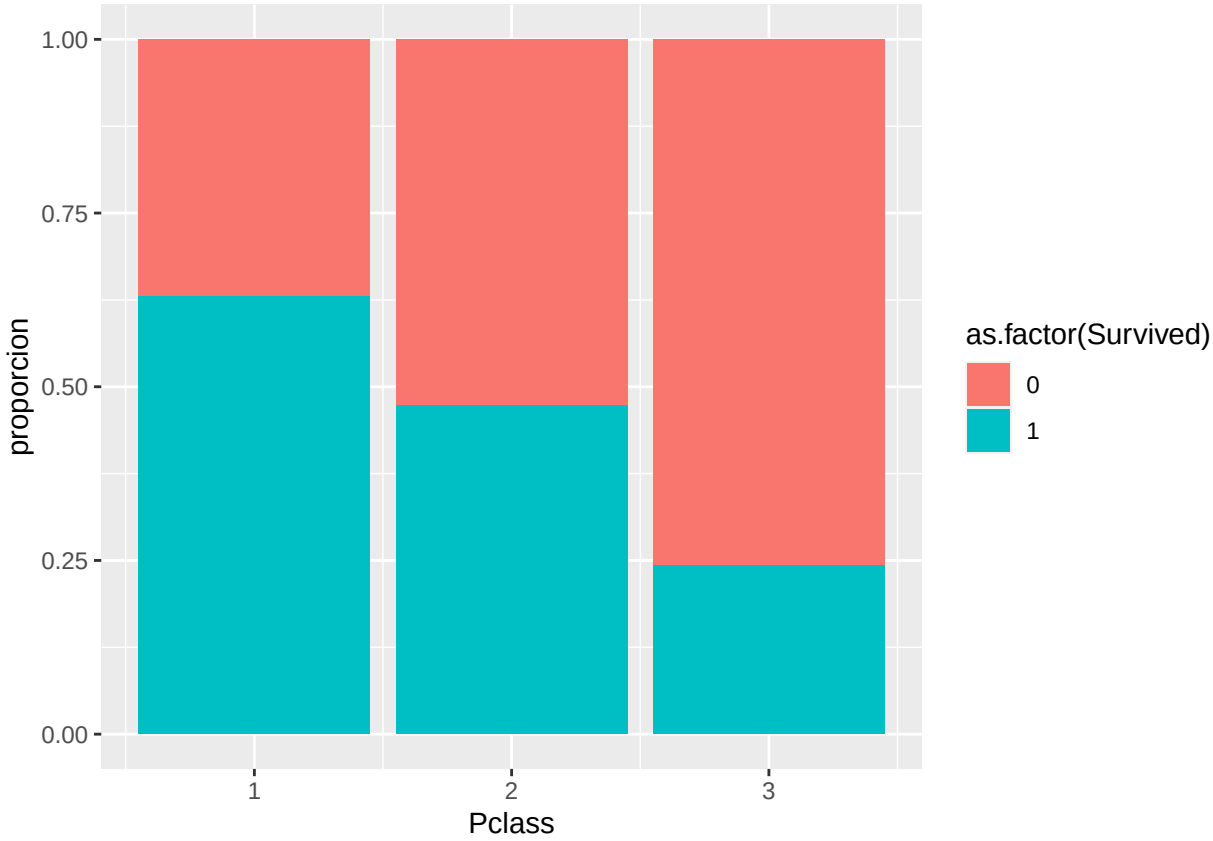
```r
mdl_Age_Pclass <- lm(Age ~ Pclass, data = TitanicT)  # variable_num ~ variable_cat
anova(mdl_Age_Pclass)
```

```
## Analysis of Variance Table
##
## Response: Age
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Pclass      1  20511 20511.4  112.39 < 2.2e-16 ***
## Residuals 712 129945   182.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
pairwise.t.test(TitanicT$Age,TitanicT$Pclass, p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  TitanicT$Age and TitanicT$Pclass
##
##   1       2
## 2 7e-09   -
## 3 < 2e-16 0.00017
##
## P value adjustment method: none
```

```r
TitanicT %>%
  ggplot(aes(Pclass,fill = as.factor(Survived)))+
  geom_bar(position = "fill")+
  ylab("proporcion")
```




```
TitanicT %>% group_by(Pclass) %>%  summarise(meanSurv = mean(Survived,na.rm = T))
```

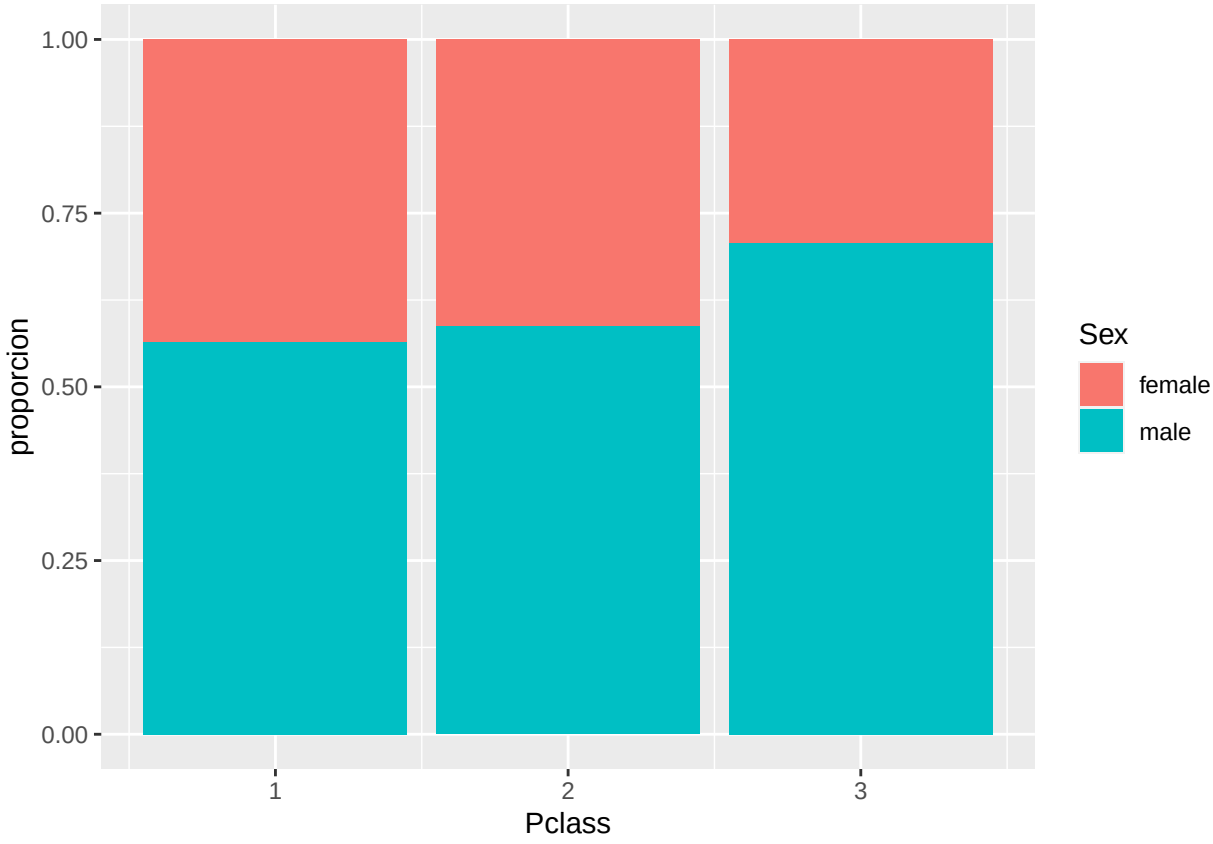
```
## # A tibble: 3 x 2
##   Pclass meanSurv
##    <dbl>    <dbl>
## 1      1    0.630
## 2      2    0.473
## 3      3    0.242
```

```
pairwise.t.test(TitanicT$Survived,TitanicT$Pclass, p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  TitanicT$Survived and TitanicT$Pclass
##
##   1       2
## 2 0.00068 -
## 3 < 2e-16 8.2e-09
##
## P value adjustment method: none
```

```
TitanicT %>%
  ggplot(aes(Pclass,fill = Sex ))+
```

```
  geom_bar(position = "fill")+
  ylab("proporcion")
```




```
TitanicT %>% mutate(Sex=(Sex=="female")) %>%  group_by(Pclass) %>%  summarise(meanSex = mean(Sex,na.rm =
```

```
## # A tibble: 3 x 2
##   Pclass meanSex
##    <dbl>   <dbl>
## 1      1   0.435
## 2      2   0.413
## 3      3   0.293
```

```
TitanicT %>% mutate(Sex=(Sex=="female")) %>% with(pairwise.t.test(Sex,Pclass, p.adjust.method = "none"))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Sex and Pclass
##
##   1       2
## 2 0.64156 -
## 3 0.00026 0.00355
##
## P value adjustment method: none
```