```
import base64
import pandas as pd
from IPython.display import HTML

train = pd.read_csv("https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv")

def create_download_link( df, title = "Download CSV file", filename = "data.csv"):
    csv = df.to_csv()
    b64 = base64.b64encode(csv.encode())
    payload = b64.decode()
    html = '<a download="{data4435546}" href="data:text/csv base64,{payload}" target="_blank">{title}</a>'
    html = html.format(payload=payload,title=title,filename=filename)
    return HTML(html)

def get_my_dataset(id):4435546
    to_add = id % 100
    to_sample = 790 + to_add
    df = train.sample(n=to_sample, random_state=to_add)
    return create_download_link(df, title="Download Final Project Dataset for "+str(id), filename="data"+str(id)+".csv")


#IMPORTANT!!!!
#call the get_my_data_set function with your PeopleSoft ID, i.e. get_my_dataset(1234567)
get_my_dataset(4435546)
```

```
Error in parse(text = x, srcfile = src): <text>:1:8: unexpected symbol
1: import base64
           ^
Traceback:
```

SEARCH STACK OVERFLOW

```
data <- read.csv('/content/data4435546.csv')
attach(data)
```

```
install.packages("plyr"); library(plyr)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependency ‘Rcpp’
```

Data Preparation #1

```
data$SexNum <- as.factor(ifelse(data$Sex == 'male',1,0))
```

```
data
```

Data Preparation #2

```
data$EmbarkedNum <- as.factor(ifelse(data$Embarked=='S',1,
                                     ifelse(data$Embarked=='C',2,3)))
```

```
data
```

A data.frame: 857 × 15

X	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticke
<int>	<int>	<int>	<int>	<chr>	<chr>	<dbl>	<int>	<int>	<chr
695	696	0	2	Chapman, Mr. Charles Henry	male	52	0	0	2487
82	83	1	3	McDermott, Miss. Brigdet Delia	female	NA	0	0	3309
765	766	1	1	Hogeboom, Mrs. John C (Anna Andrews)	female	51	1	0	135
27	28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	199
844	845	0	3	Culumovic, Mr. Jeso	male	17	0	0	3150
712	713	1	1	Taylor, Mr. Elmer Zebley	male	48	1	0	199
875	876	1	3	Najib, Miss. Adele Kiamie "Jane"	female	15	0	0	26
408	409	0	3	Birkeland, Mr. Hans Martin Monsen	male	21	0	0	3129
465	466	0	3	Goncalves, Mr. Manuel Estanslas	male	38	0	0	SOTON/O.Q. 31013
131	132	0	3	Coelho, Mr. Domingos Fernandeo	male	20	0	0	SOTON/O.Q. 31013
266	267	0	3	Panula, Mr. Ernesti Arvid	male	16	4	1	31012
808	809	0	2	Meyer, Mr. August	male	39	0	0	2487
294	295	0	3	Mineff, Mr. Ivan	male	24	0	0	3492
174	175	0	1	Smith, Mr. James Clinch	male	56	0	0	177
336	337	0	1	Pears, Mr. Thomas Clinton	male	29	1	0	1137
20	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	2398
501	502	0	3	Canavan, Miss. Mary	female	21	0	0	3648
575	576	0	3	Patchett, Mr. George	male	19	0	0	3585
429	430	1	3	Pickard, Mr. Berk (Berk Trembisky)	male	32	0	0	SOTON/O.Q. 3920
635	636	1	2	Davis, Miss. Mary	female	28	0	0	2376
786	787	1	3	Sjoblom, Miss. Anna Sofia	female	18	0	0	31012

```
install.packages("ggplot2"); library(ggplot2)
```

```
    Installing package into ‘/usr/local/lib/R/site-library’
    (as ‘lib’ is unspecified)
```

```
t = table(Pclass,Survived)
```

62 63 0 1 Harris, Mr. Henry Birkhardt male 45 1 0 369

```
t
```

```
         Survived
    Pclass   0   1
         1  79 128
         2  94  84
         3 358 114
```

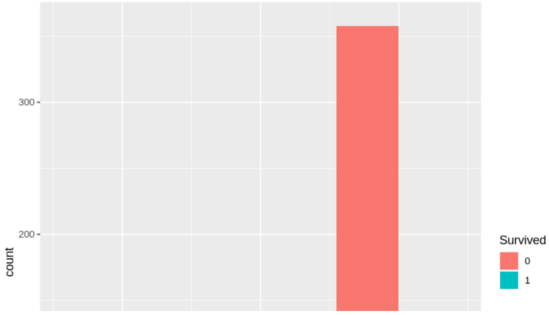
87 88 0 3 Slocovski, Mr. Selman Francis male NA 0 0 SOTON/OQ 3920

```
data.frame()
```

397 398 0 2 McKane, Mr. Peter David male 46 0 0 284

Exploratory Data Analysis#1: Create a bar graph of the Pclass variable with Survived overlay

```
ggplot(data, aes(Pclass, fill = Survived)) + geom_bar(position = 'dodge')
```



Madill, Miss. Georgette Alexandra female 15 0 1 2410

Exploratory Data Analysis #2: Create a normalized bar graph of Pclass variable with Survived over lay. Describe the raltionship between Pclass and Survived.

Levy, Mr. Rene Jacques male 36 0 0 SC/Paris 21

```
data$Survived <- as.factor(data$Survived)
```

Double-click (or enter) to edit

Pclass

Panula, Mr. Jaako Arnold male 14 4 1 31012

```
t = table(Pclass,Survived)
```

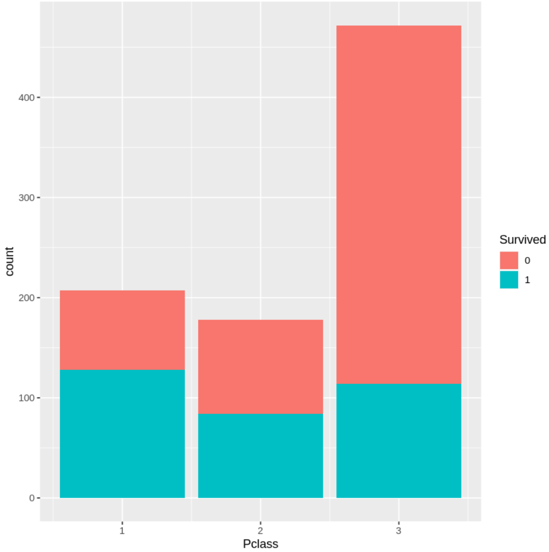
156 157 1 3 Gilnagh, Miss. Katherine "Katie" female 16 0 0 3593

```
as.data.frame.table(t/t[,'0'])
```

A data.frame: 6 × 3

Pclass <fct>	Survived <fct>	Freq <dbl>
1	0	1.0000000
2	0	1.0000000
3	0	1.0000000
1	1	1.6202532
2	1	0.8936170
3	1	0.3184358

```
ggplot(data, aes(Pclass)) + geom_bar(aes(fill = Survived))
```



According to the normalized bar graph, we could observed that Pclass 1 tend to have more people survived. Thus, we could came up with the conclusion that the higher the Pclass, the easier to survive. Pclass and Survived rate has a positive correlation.

Exploratory Data Analysis #3: Create a histogram of age with Survived overlay.

```
data
```

```
ggplot(data, aes(Age, fill = Survived)) + geom_histogram(position = 'identity', bins=30, lwd=0.2)
```

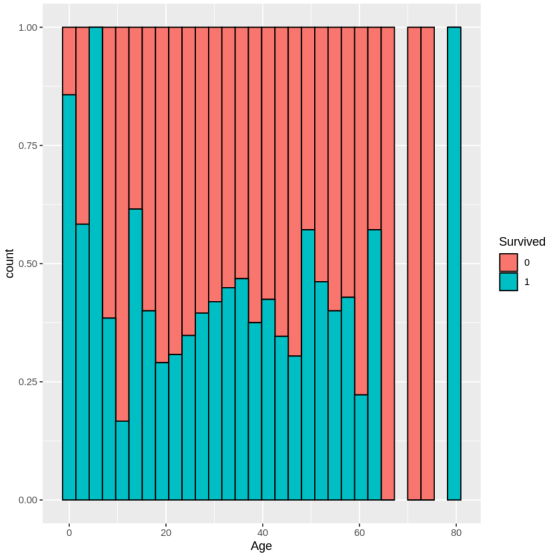
```
Warning message:
“Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.”
Warning message:
“Removed 172 rows containing non-finite values (`stat_bin()`).”
```

McDermott, Miss. Bridget Delia female NA 0 0 3309

Exploratory Data Analysis: Create a normalized histogram of age wuth Survived overlay.

```
ggplot(data, aes(Age)) + geom_histogram(aes(fill
= Survived), color="black", position = "fill")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
Warning message:
“Removed 172 rows containing non-finite values (`stat_bin()`).”
Warning message:
“Removed 4 rows containing missing values (`geom_bar()`).”
```



Exploratory Data Analysis #5: based on the standard and normalized histogram, we could make a conclusion that toldlers aged from 0-5 and young adults are more likely to survive than older age. People age between 20 to 40 are most likelyh to survive. Older people tend to have less opportunity to survive. Thus, Age and Survived rate has a negative relationship.

455 456 1 3 Jalsevac, Mr. Ivan male 29 0 0 3492

Data Partition #1

195 196 1 1 Lurette, Miss. Elise female 58 0 0 PC 175

```
set.seed(4435546)

i <- sample(nrow(data),nrow(data)*0.8)

train = data[i,]
test = data[-i,]

dim(train)
```

685 · 15

812 813 0 2 Slemen, Mr. Richard James male 35 0 0 282

```
dim(test)
```

172 · 15

59 60 0 3 Goodwin, Master. William Frederick male 11 5 2 CA 21

Data Partition #2: show descriptive stastics of age for each subset

```
train
```

A data.frame: 685 × 15

	X	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
	<int>	<int>	<fct>	<int>	<chr>	<chr>	<dbl>	<int>	<int>	<chr>
782	142	143	1	3	Hakkarainen, Mrs. Pekka Pietari (Elin Matilda Dolck)	female	24.00	1	0	STON/O2. 3101279
711	761	762	0	3	Nirva, Mr. Iisakki Antino Aijo	male	41.00	0	0	SOTON/O2 3101272
568	506	507	1	2	Quick, Mrs. Frederick Charles (Jane Richards)	female	33.00	0	2	26360
89	697	698	1	3	Mullens, Miss. Katherine "Katie"	female	NA	0	0	35852
85	806	807	0	1	Andrews, Mr. Thomas Jr	male	39.00	0	0	112050
476	341	342	1	1	Fortune, Miss. Alice Elizabeth	female	24.00	3	2	19950
351	374	375	0	3	Palsson, Miss. Stina Viola	female	3.00	3	1	349909
228	750	751	1	2	Wells, Miss. Joan	female	4.00	1	1	29103
396	733	734	0	2	Berriman, Mr. William John	male	23.00	0	0	28425
349	691	692	1	3	Karun, Miss. Manca	female	4.00	0	1	349256
614	740	741	1	1	Hawksford, Mr. Walter James	male	NA	0	0	16988
820	45	46	0	3	Rogers, Mr. William John	male	NA	0	0	S.C./A.4. 23567
636	177	178	0	1	Isham, Miss. Ann Elizabeth	female	50.00	0	0	PC 17595
353	454	455	0	3	Peduzzi, Mr. Joseph	male	NA	0	0	A/5 2817
768	83	84	0	1	Carrau, Mr. Francisco M	male	28.00	0	0	113059
705	256	257	1	1	Thorne, Mrs. Gertrude Maybelle	female	NA	0	0	PC 17585
2	82	83	1	3	McDermott, Miss. Brigdet Delia	female	NA	0	0	330932
685	192	193	1	3	Andersen-Jensen, Miss. Carla Christine Nielsine	female	19.00	1	0	350046
714	338	339	1	3	Dahl, Mr. Karl Edwart	male	45.00	0	0	7598
281	428	429	0	3	Flynn, Mr. James	male	NA	0	0	364851
741	789	790	0	1	Guggenheim, Mr. Benjamin	male	46.00	0	0	PC 17593
3	765	766	1	1	Hogeboom, Mrs. John C (Anna Andrews)	female	51.00	1	0	13502
593	521	522	0	3	Vovk, Mr. Janko	male	22.00	0	0	349252
99	137	138	0	1	Futrelle, Mr. Jacques Heath	male	37.00	1	0	113803
439	753	754	0	3	Jonkoff, Mr. Lalio	male	23.00	0	0	349204
452	623	624	0	3	Hansen, Mr. Henry Damsgaard	male	21.00	0	0	350029
224	395	396	0	3	Johansson, Mr. Erik	male	22.00	0	0	350052

```
summary(train$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.42   20.25   28.00   29.45   38.00   80.00     131 
```

```
test
```

A data.frame: 172 × 15

	X	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ti
	<int>	<int>	<fct>	<int>	<chr>	<chr>	<dbl>	<int>	<int>	<
1	695	696	0	2	Chapman, Mr. Charles Henry	male	52.0	0	0	24
6	712	713	1	1	Taylor, Mr. Elmer Zebley	male	48.0	1	0	1
7	875	876	1	3	Najib, Miss. Adele Kiamie "Jane"	female	15.0	0	0	
14	174	175	0	1	Smith, Mr. James Clinch	male	56.0	0	0	1
23	232	233	0	2	Sjostedt, Mr. Ernst Adolf	male	59.0	0	0	23
26	455	456	1	3	Jalsevac, Mr. Ivan	male	29.0	0	0	34
28	195	196	1	1	Lurette, Miss. Elise	female	58.0	0	0	PC 1
34	491	492	0	3	Windelov, Mr. Einar	male	21.0	0	0	SOTON/OQ 310
35	96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 1
36	57	58	0	3	Novel, Mr. Mansouer	male	28.5	0	0	
40	563	564	0	3	Simmons, Mr. John	male	NA	0	0	SOTON/OQ 39
49	322	323	1	2	Slayter, Miss. Hilda Mary	female	30.0	0	0	23
50	638	639	0	3	Panula, Mrs. Juha (Maria Emilia Ojala)	female	41.0	0	5	310
56	146	147	1	3	Andersson, Mr. August Edvard ("Wennerstrom")	male	27.0	0	0	35
64	417	418	1	2	Silven, Miss. Lyyli Karoliina	female	18.0	0	2	25
65	641	642	1	1	Sagesser, Mlle. Emma	female	24.0	0	0	PC 1
66	19	20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	

```
summary(test$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.83   22.00   30.00   31.63   41.00   71.00      41 
```

Data Partition #3: Create separate histograms and normalized histograms of age with Survived overlay for training and test subsets.

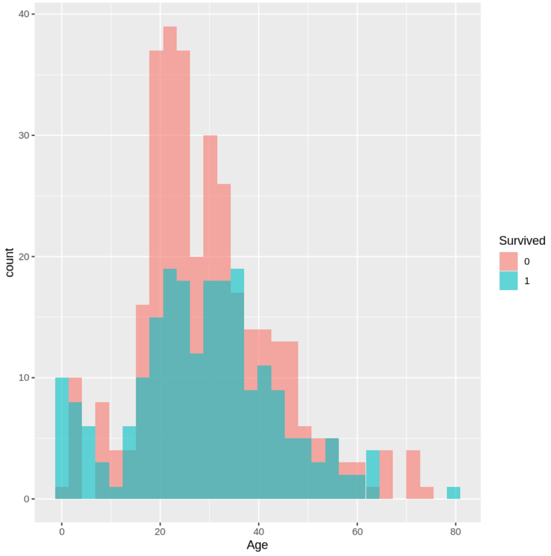
```
train
```

A data.frame: 685 × 15

	X	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
	<int>	<int>	<fct>	<int>	<chr>	<chr>	<dbl>	<int>	<int>	<chr>
782	142	143	1	3	Hakkarainen, Mrs. Pekka Pietari (Elin Matilda Dolck)	female	24.00	1	0	STON/O2. 3101279
711	761	762	0	3	Nirva, Mr. Iisakki Antino Aijo	male	41.00	0	0	SOTON/O2 3101272
568	506	507	1	2	Quick, Mrs. Frederick Charles (Jane Richards)	female	33.00	0	2	26360
89	697	698	1	3	Mullens, Miss. Katherine "Katie"	female	NA	0	0	35852
85	806	807	0	1	Andrews, Mr. Thomas Jr	male	39.00	0	0	112050
476	341	342	1	1	Fortune, Miss. Alice Elizabeth	female	24.00	3	2	19950
351	374	375	0	3	Palsson, Miss. Stina Viola	female	3.00	3	1	349909
228	750	751	1	2	Wells, Miss. Joan	female	4.00	1	1	29103
396	733	734	0	2	Berriman, Mr. William John	male	23.00	0	0	28425
452	623	624	0	3	Hansen, Mr. Henry Damsgaard	male	21.00	0	0	350029

```
ggplot(train,aes(Age,group=Survived,fill = Survived))+
  geom_histogram(bins = 30, lwd=0.2,position="identity",alpha=0.6)
```

```
Warning message:
“Removed 131 rows containing non-finite values (`stat_bin()`).”
```



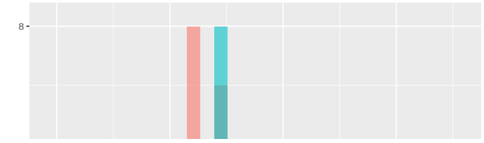
```
ggplot(train, aes(Age)) + geom_histogram(aes(fill
= Survived), color="black", position = "fill")
```

```
    `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.    Keefe, Mr. Arthur    male    NA    0    0    323592
```

```
test
```

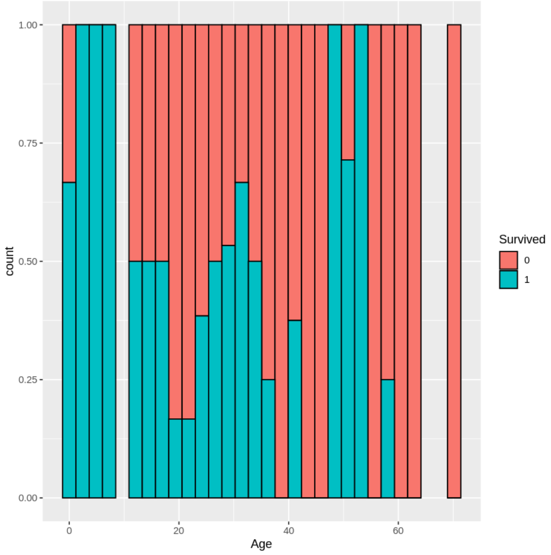
```
ggplot(test,aes(Age,group=Survived,fill = Survived))+
  geom_histogram(bins = 30, lwd=0.2,position="identity",alpha=0.6)
```

```
Warning message:
“Removed 41 rows containing non-finite values (`stat_bin()`).”
```



```
ggplot(test, aes(Age)) + geom_histogram(aes(fill
= Survived), color="black", position = "fill")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
Warning message:
“Removed 41 rows containing non-finite values (`stat_bin()`).”
Warning message:
“Removed 6 rows containing missing values (`geom_bar()`).”
```



Data Partition #4: Identify the total number of records in the training data set and how many records in the training data set have 1 for a survived variable value.

```
table(train$Survived)
```

```
  0   1 
423 262 
```

There are 423+262=785 total records in the traning data set. There are 262 records have 1 dor their survived variable value.

Decision Tree: Build a CART decison tree using ***R*** or python based on the traning data set above. (Used R)

```
install.packages(c("rpart", "rpart.plot"))
library(rpart); library(rpart.plot)
```

```
Installing packages into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

1: Age and Sex(numeric)

```
cart01 = rpart(Survived ~ Age + SexNum,method="class", data = train)
```

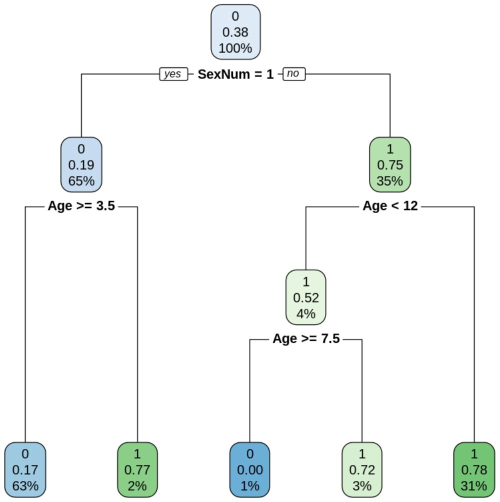
```
cart01
```

```
n= 685

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 685 262 0 (0.6175182 0.3824818)
   2) SexNum=1 446  83 0 (0.8139013 0.1860987)
     4) Age>=3.5 433  73 0 (0.8314088 0.1685912) *
     5) Age< 3.5 13   3 1 (0.2307692 0.7692308) *
   3) SexNum=0 239  60 1 (0.2510460 0.7489540)
```

```
rpart.plot(cart01)
```



2 Pclass, Age, Fare, and Sex

Since Fare has null values, I ruled out Fare in order to generate the decison tree.

```
cart02 = rpart(Survived ~ Age + SexNum + Pclass + Age,method="class", , data = train)
```

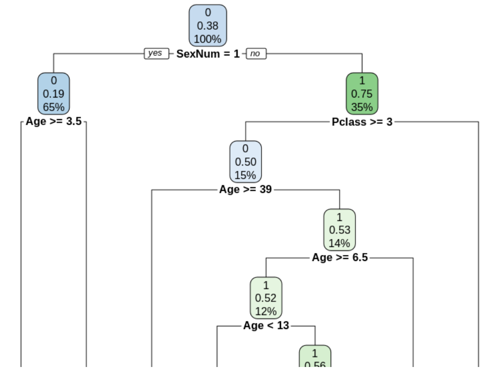
```
cart02
```

```
n= 685

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 685 262 0 (0.61751825 0.38248175)
    2) SexNum=1 446  83 0 (0.81390135 0.18609865)
      4) Age>=3.5 433  73 0 (0.83140878 0.16859122) *
      5) Age< 3.5 13   3 1 (0.23076923 0.76923077) *
    3) SexNum=0 239  60 1 (0.25104603 0.74895397)
      6) Pclass>=2.5 105  52 0 (0.50476190 0.49523810)
       12) Age>=38.5 9   1 0 (0.88888889 0.11111111) *
       13) Age< 38.5 96  45 1 (0.46875000 0.53125000)
         26) Age>=6.5 85  41 1 (0.48235294 0.51764706)
           52) Age< 12.5 7   0 0 (1.00000000 0.00000000) *
           53) Age>=12.5 78  34 1 (0.43589744 0.56410256)
            106) Age>=27.5 13   5 0 (0.61538462 0.38461538) *
            107) Age< 27.5 65  26 1 (0.40000000 0.60000000) *
         27) Age< 6.5 11   4 1 (0.36363636 0.63636364) *
      7) Pclass< 2.5 134   7 1 (0.05223881 0.94776119) *
```

```
rpart.plot(cart02)
```



3 SibSp, Parch, and Embarked(numeric)

```
cart03 = rpart(Survived ~ SibSp + Parch + EmbarkedNum, method = "class", data = train)
```

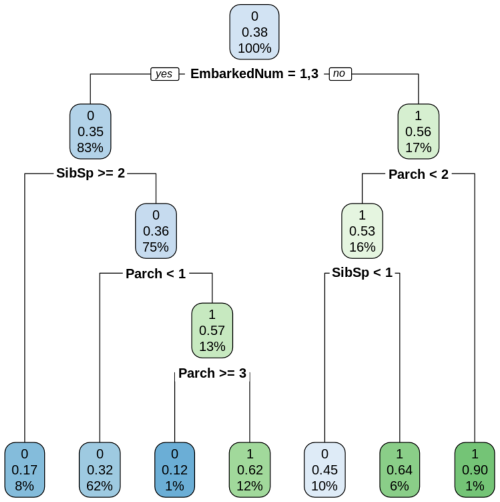
```
cart03
```

```
n= 685

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 685 262 0 (0.6175182 0.3824818)
   2) EmbarkedNum=1,3 567 196 0 (0.6543210 0.3456790)
     4) SibSp>=1.5 54   9 0 (0.8333333 0.1666667) *
     5) SibSp< 1.5 513 187 0 (0.6354776 0.3645224)
      10) Parch< 0.5 426 137 0 (0.6784038 0.3215962) *
      11) Parch>=0.5 87  37 1 (0.4252874 0.5747126)
        22) Parch>=2.5 8   1 0 (0.8750000 0.1250000) *
        23) Parch< 2.5 79  30 1 (0.3797468 0.6202532) *
   3) EmbarkedNum=2 118  52 1 (0.4406780 0.5593220)
     6) Parch< 1.5 108  51 1 (0.4722222 0.5277778)
      12) SibSp< 0.5 66  30 0 (0.5454545 0.4545455) *
      13) SibSp>=0.5 42  15 1 (0.3571429 0.6428571) *
     7) Parch>=1.5 10   1 1 (0.1000000 0.9000000) *
```

```
rpart.plot(cart03)
```


0
0.38
100%
yes
EmbarkedNum = 1,3
no
0
0.35
83%
SibSp >= 2
1
0.56
17%
Parch < 2
0
0.36
75%
Parch < 1
1
0.53
16%
SibSp < 1
1
0.57
13%
Parch >= 3
0
0.17
8%
0
0.32
62%
0
0.12
1%
1
0.62
12%
0
0.45
10%
1
0.64
6%
1
0.90
1%