

```

> #Perform the below operations:
> # a. Pre-process the passenger names to come up with a list of
> #titles that represent families and represent using appropriate
> #visualization graph.
>
> #Importing the titanic dataset into R
> library(xlsx)
> titanicdf<-read.xlsx("titanic3.xls",1)
> #Converting the name column to char
> titanicdf$name<-as.character(titanicdf$name)
>
> #Extracting the family name (first name) and adding it to a new column family
> for (i in seq (1:1309)) {
+   titanicdf$family[i]<- strsplit(titanicdf$name, ",")[[i]][1]
+ }
> View(titanicdf$family)
>
> #Count the frequency of common names in family column.
> library(plyr)
> family_frequency<- count(titanicdf,"family")
> barplot(family_frequency$freq)
>
> #b. Represent the proportion of people survived by family size
> #using a graph.
> library(sqldf)
>
> survived<- sqldf("SELECT *
+   FROM titanicdf
+   WHERE survived = '1'")
> survived_family_freq <- count(survived$family)
> barplot(survived_family_freq$freq)
>
> #c. Impute the missing values in Age variable using Mice library,
> #create two different graphs showing Age distribution before
> #and after imputation
> library(mice)
Loading required package: lattice
Attaching package: 'mice'
The following object is masked from 'package:tidyr':
  complete
The following objects are masked from 'package:base':
  cbind, rbind

>
> #Removing columns 1,3,8,9,10,12,13,14,15
>
> mini_titanic <- titanicdf[-c(1,3,8,9,10,12,13,14,15)]
>
> md.pattern(mini_titanic)
      survived sex sibsp parch embarked age
1044         1   1     1     1         1   1   0
263          1   1     1     1         1   0   1
2            1   1     1     1         0   1   1
1            0   0     0     0         0   0   6
            1   1     1     1         3 264 271
>
> library(dplyr)

```

Attaching package: 'dplyr'

The following objects are masked from 'package:plyr':

arrange, count, desc, failwith, id, mutate, rename, summarise, summarize

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
> mini_titanic <- mini_titanic %>%
+   mutate(
+     survived = as.factor(survived),
+     sex = as.factor(sex),
+     age = as.numeric(age),
+     sibsp = as.factor(sibsp),
+     parch = as.factor(parch),
+     embarked = as.factor(embarked)
+   )
> str(mini_titanic)
'data.frame': 1310 obs. of 6 variables:
 $ survived: Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...
 $ sex      : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
 $ age      : num 29 0.917 2 30 25 ...
 $ sibsp    : Factor w/ 7 levels "0","1","2","3",...: 1 2 2 2 2 1 2 1 3 1 ...
 $ parch    : Factor w/ 8 levels "0","1","2","3",...: 1 3 3 3 3 1 1 1 1 1 ...
 $ embarked: Factor w/ 3 levels "C","Q","S": 3 3 3 3 3 3 3 3 3 1 ...
>
>
> #running the mice function
> temp_mini_titanic <- mice(mini_titanic,m=5,maxit=50,seed=500)
```

iter	imp	variable					
1	1	survived	sex	age	sibsp	parch	embarked
1	2	survived	sex	age	sibsp	parch	embarked
1	3	survived	sex	age	sibsp	parch	embarked
1	4	survived	sex	age	sibsp	parch	embarked
1	5	survived	sex	age	sibsp	parch	embarked
2	1	survived	sex	age	sibsp	parch	embarked
2	2	survived	sex	age	sibsp	parch	embarked
2	3	survived	sex	age	sibsp	parch	embarked
2	4	survived	sex	age	sibsp	parch	embarked
2	5	survived	sex	age	sibsp	parch	embarked
3	1	survived	sex	age	sibsp	parch	embarked
3	2	survived	sex	age	sibsp	parch	embarked
3	3	survived	sex	age	sibsp	parch	embarked
3	4	survived	sex	age	sibsp	parch	embarked
3	5	survived	sex	age	sibsp	parch	embarked
4	1	survived	sex	age	sibsp	parch	embarked
4	2	survived	sex	age	sibsp	parch	embarked
4	3	survived	sex	age	sibsp	parch	embarked
4	4	survived	sex	age	sibsp	parch	embarked
4	5	survived	sex	age	sibsp	parch	embarked
5	1	survived	sex	age	sibsp	parch	embarked
5	2	survived	sex	age	sibsp	parch	embarked
5	3	survived	sex	age	sibsp	parch	embarked
5	4	survived	sex	age	sibsp	parch	embarked
5	5	survived	sex	age	sibsp	parch	embarked

[illegible]

[illegible]

[illegible]

```

43 5 survived sex age sibsp parch embarked
44 1 survived sex age sibsp parch embarked
44 2 survived sex age sibsp parch embarked
44 3 survived sex age sibsp parch embarked
44 4 survived sex age sibsp parch embarked
44 5 survived sex age sibsp parch embarked
45 1 survived sex age sibsp parch embarked
45 2 survived sex age sibsp parch embarked
45 3 survived sex age sibsp parch embarked
45 4 survived sex age sibsp parch embarked
45 5 survived sex age sibsp parch embarked
46 1 survived sex age sibsp parch embarked
46 2 survived sex age sibsp parch embarked
46 3 survived sex age sibsp parch embarked
46 4 survived sex age sibsp parch embarked
46 5 survived sex age sibsp parch embarked
47 1 survived sex age sibsp parch embarked
47 2 survived sex age sibsp parch embarked
47 3 survived sex age sibsp parch embarked
47 4 survived sex age sibsp parch embarked
47 5 survived sex age sibsp parch embarked
48 1 survived sex age sibsp parch embarked
48 2 survived sex age sibsp parch embarked
48 3 survived sex age sibsp parch embarked
48 4 survived sex age sibsp parch embarked
48 5 survived sex age sibsp parch embarked
49 1 survived sex age sibsp parch embarked
49 2 survived sex age sibsp parch embarked
49 3 survived sex age sibsp parch embarked
49 4 survived sex age sibsp parch embarked
49 5 survived sex age sibsp parch embarked
50 1 survived sex age sibsp parch embarked
50 2 survived sex age sibsp parch embarked
50 3 survived sex age sibsp parch embarked
50 4 survived sex age sibsp parch embarked
50 5 survived sex age sibsp parch embarked

```

Warning message:

Number of logged events: 250

> summary(temp_mini_titanic)

Class: mids

Number of multiple imputations: 5

Imputation methods:

```

survived sex age sibsp parch embarked
"logreg" "logreg" "pmm" "polyreg" "polyreg" "polyreg"

```

PredictorMatrix:

```

      survived sex age sibsp parch embarked
survived      0  1  1      1      1      1
sex           1  0  1      1      1      1
age           1  1  0      1      1      1
sibsp         1  1  1      0      1      1
parch         1  1  1      1      0      1
embarked      1  1  1      1      1      0

```

Number of logged events: 250

```

  it im dep meth out
1 1 1 age pmm parch9
2 1 2 age pmm parch9
3 1 3 age pmm parch9
4 1 4 age pmm parch9
5 1 5 age pmm parch9
6 2 1 age pmm parch9

```

>

> #check for imputed data in a field

> temp_mini_titanic\$imp\$age

	1	2	3	4	5
16	26.0000	63.0000	30.0000	57.0000	23.0
38	37.0000	70.5000	33.0000	24.0000	22.0
41	36.0000	71.0000	57.0000	34.0000	17.0
47	20.0000	21.0000	28.0000	57.0000	22.0
60	21.0000	26.0000	26.0000	61.0000	35.0
70	0.3333	14.0000	36.0000	1.0000	53.0
71	18.0000	27.0000	27.0000	28.0000	18.0
75	19.0000	44.0000	42.0000	38.0000	24.0
81	20.0000	26.0000	17.0000	30.0000	24.0
107	31.0000	60.0000	40.0000	23.0000	51.0
108	16.0000	29.0000	30.0000	16.0000	14.0
109	24.0000	43.0000	29.0000	62.0000	32.0
119	40.0000	31.0000	23.0000	30.0000	23.0
122	66.0000	23.0000	35.0000	30.0000	35.0
126	20.0000	31.0000	40.5000	22.0000	35.0
135	36.0000	45.0000	31.0000	35.0000	43.0
148	27.0000	40.0000	30.0000	20.0000	45.0
153	37.0000	40.5000	33.0000	33.0000	16.0
158	18.0000	21.0000	30.0000	22.0000	24.0
167	36.0000	58.0000	32.5000	36.0000	36.0
177	60.0000	14.0000	29.0000	22.0000	31.0
180	31.0000	21.0000	30.0000	29.0000	24.0
185	46.0000	36.0000	14.5000	37.0000	50.0
197	35.0000	17.0000	33.0000	36.0000	25.0
205	14.0000	25.0000	54.0000	20.0000	43.0
220	26.0000	55.0000	40.0000	20.0000	23.0
224	26.0000	40.0000	19.0000	45.0000	29.0
236	37.0000	31.0000	18.0000	33.0000	22.0
238	26.5000	40.0000	39.0000	37.0000	36.0
242	29.0000	31.0000	29.0000	22.0000	24.0
255	30.5000	35.0000	30.5000	24.0000	30.0
257	18.5000	31.0000	35.0000	24.0000	30.0
270	26.0000	40.0000	33.0000	23.0000	39.0
278	36.0000	40.0000	28.0000	40.0000	25.0
284	48.0000	71.0000	27.0000	44.0000	48.0
294	21.0000	30.0000	20.0000	35.0000	62.0
298	64.0000	47.0000	22.0000	61.0000	32.0
319	30.0000	25.0000	29.0000	57.0000	29.0
321	37.0000	61.0000	30.5000	33.0000	15.0
364	60.0000	47.0000	16.0000	20.0000	35.0
383	35.0000	21.0000	22.0000	31.0000	18.0
385	31.0000	26.0000	47.0000	30.0000	76.0
411	20.0000	60.0000	26.0000	22.0000	20.0
470	48.0000	18.0000	9.0000	21.0000	22.0
474	27.0000	20.0000	16.0000	38.0000	63.0
478	20.0000	23.0000	15.0000	22.0000	21.0
484	16.0000	21.0000	16.0000	9.0000	32.0
492	25.0000	25.0000	32.5000	29.0000	17.0
496	21.0000	34.5000	50.0000	63.0000	36.0
525	25.0000	41.0000	30.0000	45.0000	32.0
529	47.0000	25.0000	29.0000	20.0000	18.0
532	23.0000	24.0000	57.0000	21.0000	14.0
582	30.0000	26.0000	22.0000	24.0000	29.0
596	27.0000	60.0000	30.0000	21.0000	24.0
598	30.5000	31.0000	37.0000	33.0000	30.0
673	35.0000	25.0000	18.0000	44.0000	19.0
681	30.0000	28.5000	29.0000	45.0000	17.0
682	0.8333	36.0000	22.0000	50.0000	25.0
683	18.0000	20.0000	20.0000	7.0000	59.0
706	36.0000	64.0000	29.0000	14.5000	28.0
707	27.0000	27.0000	26.5000	18.0000	64.0
757	58.0000	30.0000	60.0000	31.0000	17.0

758	14.0000	18.0000	22.0000	33.0000	19.0
768	20.0000	47.0000	22.0000	57.0000	24.0
769	27.0000	26.0000	47.0000	25.0000	19.0
776	33.0000	40.0000	28.0000	31.0000	50.0
790	33.0000	33.0000	26.0000	30.0000	19.0
796	27.0000	33.0000	26.0000	25.0000	28.0
799	37.0000	35.0000	18.0000	37.0000	62.0
801	49.0000	19.0000	65.0000	19.0000	25.0
802	25.0000	32.5000	50.0000	30.0000	35.0
803	36.5000	27.0000	37.0000	16.0000	57.0
805	25.0000	33.0000	17.0000	30.0000	62.0
806	37.0000	21.0000	18.0000	33.0000	16.0
809	27.0000	31.0000	33.0000	20.0000	29.0
813	25.0000	21.0000	15.0000	15.0000	21.0
814	44.0000	44.0000	32.5000	18.0000	19.0
816	27.0000	47.0000	48.0000	22.0000	24.0
817	71.0000	33.0000	47.0000	27.0000	27.0
820	39.0000	29.0000	39.0000	39.0000	21.0
836	19.0000	27.0000	24.0000	38.0000	24.0
843	46.0000	54.0000	17.0000	35.0000	30.0
844	71.0000	52.0000	43.0000	35.0000	30.0
853	19.0000	25.0000	25.0000	21.0000	35.0
855	24.0000	9.0000	50.0000	15.0000	26.0
857	39.0000	28.0000	40.0000	39.0000	30.0
859	33.0000	21.0000	33.0000	21.0000	16.0
866	6.0000	40.0000	65.0000	39.0000	25.0
872	27.0000	33.0000	15.0000	22.0000	25.0
873	30.0000	30.0000	15.0000	29.0000	45.0
875	37.0000	65.0000	37.0000	21.0000	16.0
877	29.0000	27.0000	25.0000	30.0000	35.0
880	60.0000	30.0000	22.0000	37.0000	24.0
883	40.0000	60.0000	20.0000	30.0000	22.0
887	9.0000	3.0000	24.0000	23.0000	29.0
888	30.5000	40.5000	33.0000	33.0000	38.0
901	0.9167	64.0000	8.0000	44.0000	58.0
902	27.0000	40.5000	13.0000	55.0000	18.0
903	0.3333	29.0000	4.0000	18.0000	26.0
904	24.0000	0.3333	4.0000	59.0000	18.0
919	46.0000	20.0000	14.5000	50.0000	52.0
921	24.0000	9.0000	37.0000	22.0000	25.0
922	40.0000	27.0000	17.0000	24.0000	20.0
923	27.0000	28.0000	27.0000	39.0000	29.0
924	39.0000	30.0000	27.0000	36.0000	24.0
927	9.0000	27.0000	30.0000	7.0000	45.0
928	39.0000	55.0000	18.0000	34.5000	14.5
929	30.0000	27.0000	40.0000	26.0000	18.0
930	14.0000	27.0000	36.0000	21.0000	27.0
931	54.0000	27.0000	31.0000	34.0000	26.0
932	42.0000	34.0000	31.0000	30.0000	35.0
941	28.5000	30.0000	55.0000	42.0000	14.0
943	39.0000	34.5000	47.0000	36.0000	36.0
945	30.0000	25.0000	32.0000	29.0000	35.0
946	21.0000	65.0000	18.0000	33.0000	22.0
947	60.0000	40.0000	50.0000	30.0000	24.0
949	27.0000	27.0000	17.0000	30.0000	31.0
955	8.0000	9.0000	18.0000	18.0000	10.0
956	14.5000	10.0000	20.0000	9.0000	4.0
957	14.5000	14.5000	24.0000	18.0000	14.5
958	19.0000	4.0000	12.0000	9.0000	9.0
959	43.0000	29.0000	45.0000	41.0000	29.0
962	30.0000	31.0000	15.0000	17.0000	22.0
963	15.0000	31.0000	45.0000	35.0000	26.0
972	27.0000	5.0000	49.0000	22.0000	61.0

974	26.0000	44.0000	32.0000	20.0000	51.0
977	40.0000	18.0000	18.5000	23.0000	24.0
983	18.0000	60.0000	45.5000	22.0000	24.0
984	40.0000	47.0000	40.0000	20.0000	35.0
985	24.0000	3.0000	24.0000	21.0000	29.0
988	15.0000	27.0000	65.0000	39.0000	25.0
989	25.0000	30.0000	37.0000	30.0000	19.0
990	26.0000	44.0000	19.0000	22.0000	18.0
992	29.0000	24.0000	28.5000	20.0000	38.0
994	39.0000	30.0000	50.0000	36.0000	24.0
995	71.0000	20.0000	64.0000	29.0000	60.0
998	42.0000	54.0000	31.0000	31.0000	39.0
999	30.0000	21.0000	32.5000	27.0000	50.0
1000	39.0000	18.0000	19.0000	23.0000	20.0
1001	21.0000	50.0000	22.0000	60.0000	21.0
1002	13.0000	9.0000	0.3333	8.0000	36.0
1003	21.0000	3.0000	36.5000	0.8333	36.0
1004	7.0000	20.0000	49.0000	11.0000	10.0
1005	50.0000	1.0000	40.0000	0.4167	45.0
1006	38.0000	63.0000	6.0000	16.0000	62.0
1007	40.0000	28.0000	19.0000	36.0000	21.0
1010	20.0000	27.0000	31.0000	22.0000	34.5
1013	31.0000	40.0000	65.0000	27.0000	70.5
1014	31.0000	20.0000	52.0000	18.0000	24.0
1015	35.0000	23.0000	23.0000	19.0000	18.5
1017	27.0000	19.0000	37.0000	22.0000	57.0
1019	60.0000	20.0000	22.0000	22.0000	29.0
1023	27.0000	21.0000	26.0000	25.0000	24.0
1024	39.0000	1.0000	39.0000	36.0000	45.0
1028	31.0000	18.0000	34.0000	30.0000	45.0
1029	19.0000	40.0000	9.0000	6.0000	18.0
1030	35.0000	25.0000	27.0000	66.0000	35.0
1031	24.0000	35.0000	17.0000	30.0000	24.0
1033	25.0000	50.0000	15.0000	30.0000	65.0
1034	37.0000	31.0000	33.0000	18.5000	16.0
1035	29.0000	36.0000	22.0000	16.0000	60.0
1036	29.0000	0.4167	16.0000	16.0000	20.0
1037	21.0000	0.7500	0.3333	31.0000	58.0
1038	32.0000	29.0000	53.0000	62.0000	45.0
1039	19.0000	25.0000	22.0000	23.0000	19.0
1040	24.0000	1.0000	40.0000	23.0000	38.0
1042	40.0000	30.0000	14.0000	28.0000	18.0
1043	26.0000	39.0000	50.0000	49.0000	29.0
1044	32.0000	40.0000	16.0000	50.0000	47.0
1045	50.0000	29.0000	39.0000	23.0000	22.0
1053	18.0000	21.0000	24.0000	22.0000	35.0
1054	71.0000	48.0000	18.0000	17.0000	50.0
1055	37.0000	48.0000	24.0000	16.0000	40.5
1056	47.0000	31.0000	18.0000	22.0000	18.0
1070	15.0000	38.0000	32.0000	20.0000	20.0
1071	36.5000	5.0000	6.0000	22.0000	26.0
1072	25.0000	40.0000	39.0000	49.0000	30.0
1073	38.0000	30.0000	17.0000	30.0000	62.0
1074	38.0000	26.0000	6.0000	16.0000	31.0
1075	27.0000	55.0000	50.0000	22.0000	61.0
1077	49.0000	39.0000	7.0000	45.0000	57.0
1078	48.0000	26.0000	39.0000	21.0000	21.0
1079	27.0000	3.0000	48.0000	21.0000	29.0
1081	9.0000	50.0000	15.0000	45.0000	21.0
1082	39.0000	3.0000	40.0000	23.0000	29.0
1086	20.0000	21.0000	29.0000	23.0000	76.0
1096	31.0000	40.0000	65.0000	18.0000	61.0
1110	22.0000	30.0000	64.0000	30.0000	49.0

```
[ reached getOption("max.print") -- omitted 64 rows ]
```

```
> #Now we can get back the completed dataset using the complete() function.
> completed_mini_titanic <- complete(temp_mini_titanic,1)
> md.pattern(completed_mini_titanic)
```

$\Rightarrow V \leq$ No need for mice. This data set is completely observed.

```
>
> #Inspecting the distribution of original and imputed data
> #before imputation - Age
>
> #The density of the imputed data for each imputed dataset
> #is showed in magenta while the density of the observed data is showed in blue
> #after imputation of age
> densityplot(temp_mini_titanic)
> #before imputation of age
> densityplot(mini_titanic$age)
```