Titanic Analysis with SAS

Chenxing Li

December 2, 2020

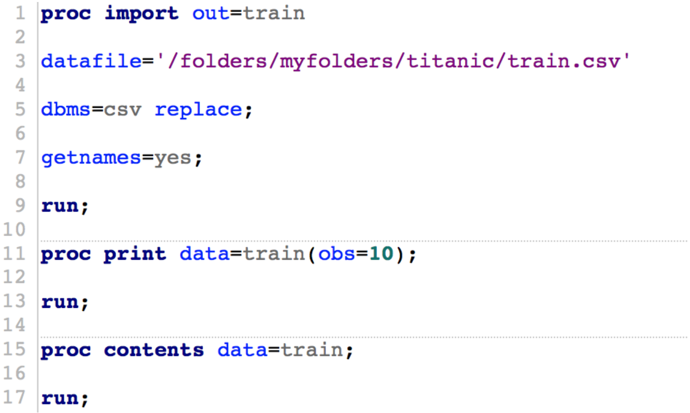
Contents

- Understand data
- Factor analysis

Data Sources

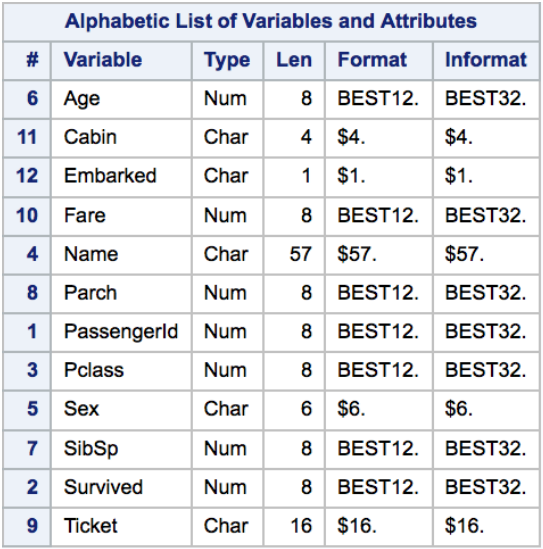
- KAGGLE Titanic survival prediction competition: https://www.kaggle.com/c/titanic/overview
- The data contains two groups:
  - training set (train.csv): rescued status and basic information of the passengers
  - test set (test.csv): only basic information of the passengers

View Data



```
proc import out=train

datafile='/folders/myfolders/titanic/train.csv'

dbms=csv replace;

getnames=yes;

run;

proc print data=train(obs=10);

run;

proc contents data=train;

run;
```

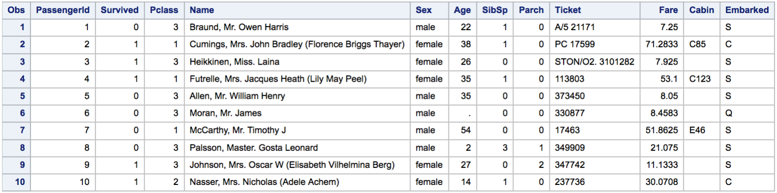
View Data



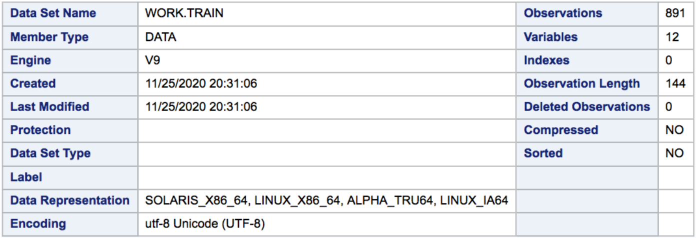
Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
6	Age	Num	8	BEST12.	BEST32.
11	Cabin	Char	4	$4.	$4.
12	Embarked	Char	1	$1.	$1.
10	Fare	Num	8	BEST12.	BEST32.
4	Name	Char	57	$57.	$57.
8	Parch	Num	8	BEST12.	BEST32.
1	PassengerId	Num	8	BEST12.	BEST32.
3	Pclass	Num	8	BEST12.	BEST32.
5	Sex	Char	6	$6.	$6.
7	SibSp	Num	8	BEST12.	BEST32.
2	Survived	Num	8	BEST12.	BEST32.
9	Ticket	Char	16	$16.	$16.

View Data

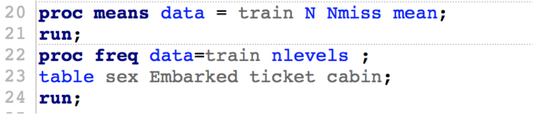


Obs	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	6	0	3	Moran, Mr. James	male	.	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

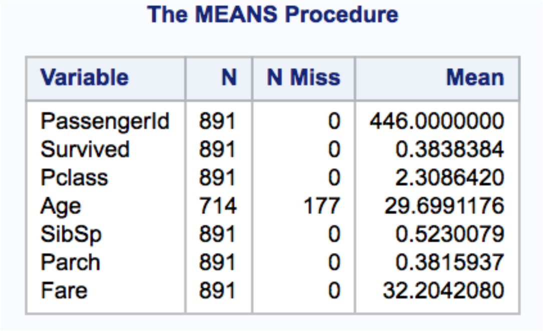


Data Set Name	WORK.TRAIN	**Observations**	891
Member Type	DATA	**Variables**	12
Engine	V9	**Indexes**	0
Created	11/25/2020 20:31:06	**Observation Length**	144
Last Modified	11/25/2020 20:31:06	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Clean Data



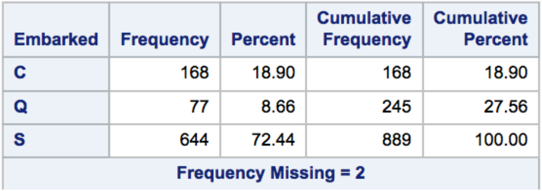
```
proc means data = train N Nmiss mean;
run;
proc freq data=train nlevels ;
table sex Embarked ticket cabin;
run;
```



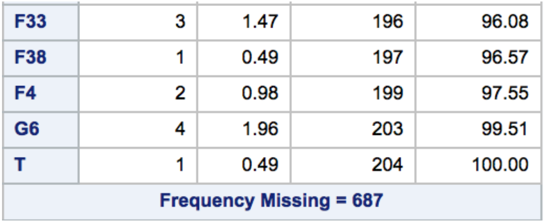
The MEANS Procedure

Variable	N	N Miss	Mean
PassengerId	891	0	446.0000000
Survived	891	0	0.3838384
Pclass	891	0	2.3086420
Age	714	177	29.6991176
SibSp	891	0	0.5230079
Parch	891	0	0.3815937
Fare	891	0	32.2042080

Clean Data

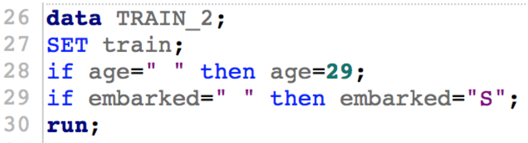


Embarked	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	168	18.90	168	18.90
Q	77	8.66	245	27.56
S	644	72.44	889	100.00
Frequency Missing = 2				



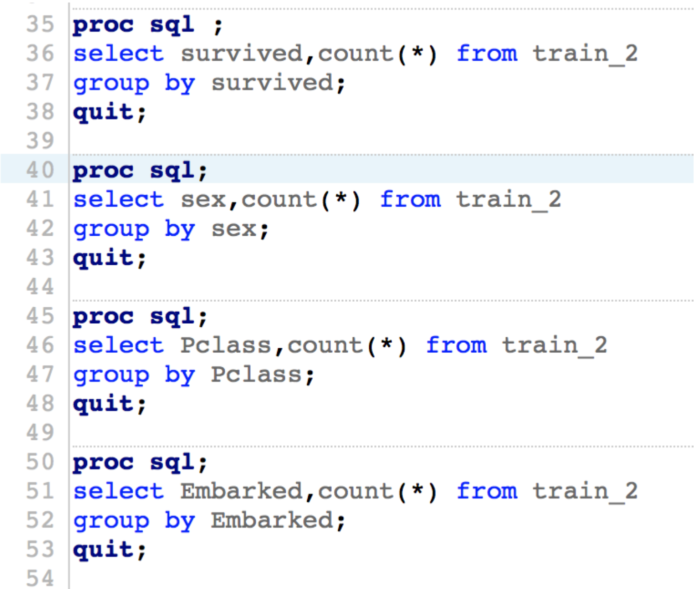
F33	3	1.47	196	96.08
F38	1	0.49	197	96.57
F4	2	0.98	199	97.55
G6	4	1.96	203	99.51
T	1	0.49	204	100.00
Frequency Missing = 687				

Clean Data



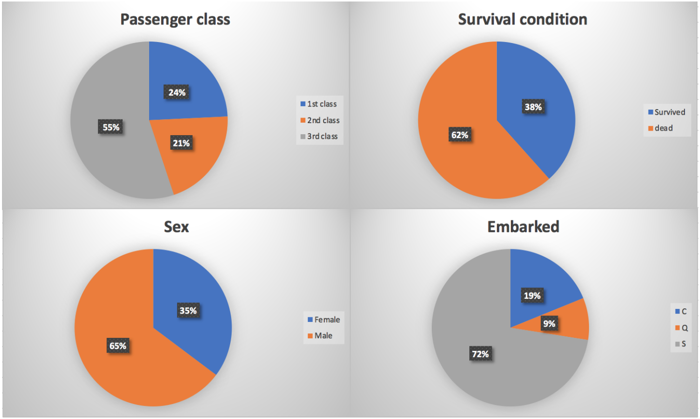
```
data TRAIN_2;
SET train;
if age=" " then age=29;
if embarked=" " then embarked="S";
run;
```

Overall situation

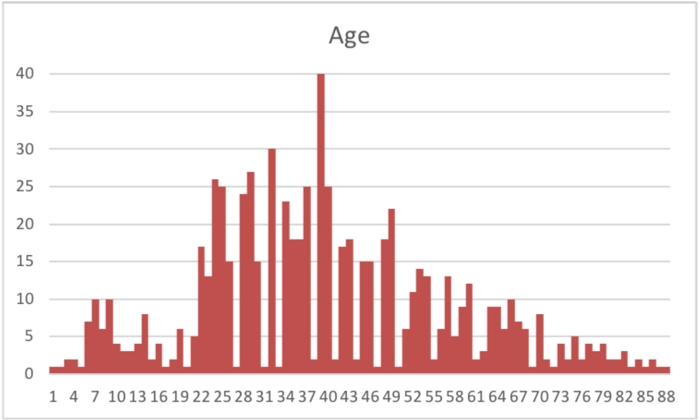


```
proc sql ;
select survived,count(*) from train_2
group by survived;
quit;

proc sql;
select sex,count(*) from train_2
group by sex;
quit;

proc sql;
select Pclass,count(*) from train_2
group by Pclass;
quit;

proc sql;
select Embarked,count(*) from train_2
group by Embarked;
quit;

```

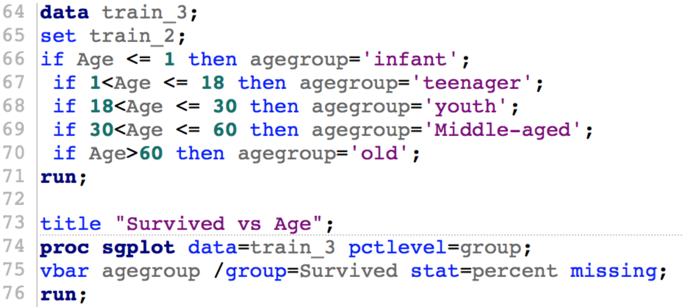
Overall situation



Overall situation

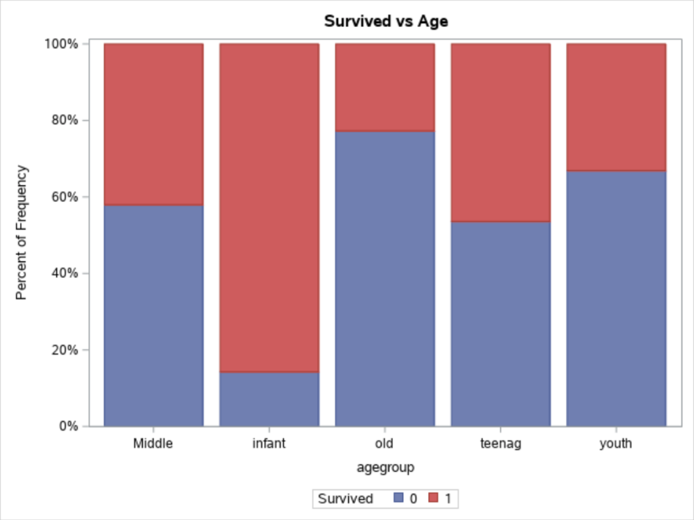


Survived VS Age

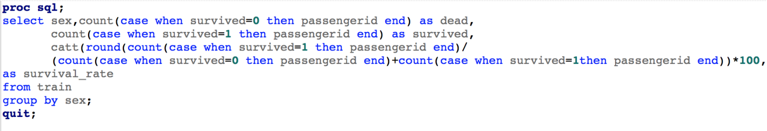


```
data train_3;
set train_2;
if Age <= 1 then agegroup='infant';
 if 1<Age <= 18 then agegroup='teenager';
 if 18<Age <= 30 then agegroup='youth';
 if 30<Age <= 60 then agegroup='Middle-aged';
 if Age>60 then agegroup='old';
run;

title "Survived vs Age";
proc sgplot data=train_3 pctlevel=group;
vbar agegroup /group=Survived stat=percent missing;
run;
```

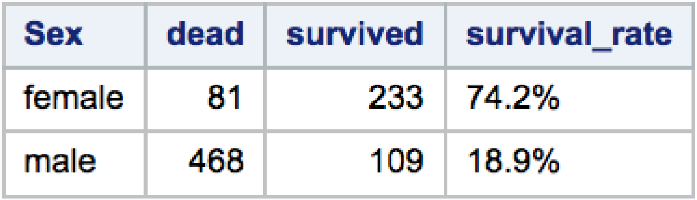
Survived VS Age



Survived VS Gender

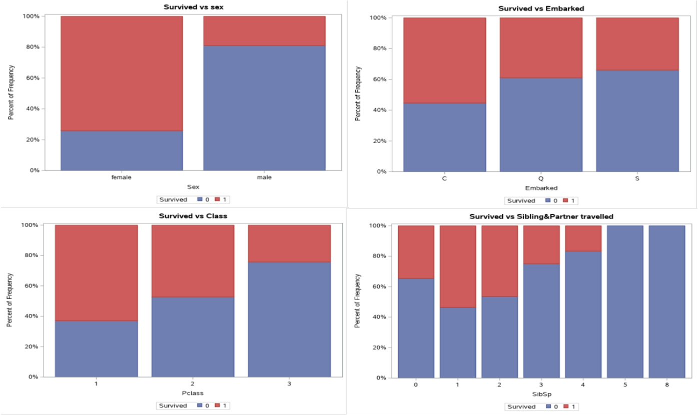


```
proc sql;
select sex,count(case when survived=0 then passengerid end) as dead,
       count(case when survived=1 then passengerid end) as survived,
       catt(round(count(case when survived=1 then passengerid end)/
       (count(case when survived=0 then passengerid end)+count(case when survived=1then passengerid end))*100,
as survival_rate
from train
group by sex;
quit;
```



Sex	dead	survived	survival_rate
female	81	233	74.2%
male	468	109	18.9%

Survived VS Other variables



Summary

- Survival rate decreases with age
- More females survived than males
- The higher the cabin level, the higher the survival rate.
- Port S has the most passengers on board, but the survival rate is the lowest. More than half of the passengers in Port C were rescued.
- Passengers with family members of 3 or less have a higher survival rate.

Future work

- **Building model**. Logistic regression
- **Testing test dataset**