Titanic: Machine Learning from Disaster

Predict survival on the Titanic

- Defining the problem statement
- Collecting the data
- Exploratory data analysis
- Feature engineering
- Modelling
- Testing

1. Defining the problem statement

Complete the analysis of what sorts of people were likely to survive.
In particular, we ask you to apply the tools of machine learning to predict which passengers survived the Titanic tragedy.

2. Collecting the data

training data set and testing data set are given by Kaggle you can download from
or you can download from kaggle directly kaggle

load train, test dataset using Pandas

3. Exploratory data analysis

Printing first 5 rows of the train dataset.

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarkec
0	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.00	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	C
6	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.8542	NaN	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	16.0000	NaN	S
16	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250	NaN	C

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	NaN	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.00	1	0	345763	18.0000	NaN	S
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7.2250	NaN	C
20	21	0	2	Fynney, Mr. Joseph J	male	35.00	0	0	239865	26.0000	NaN	S
21	22	1	2	Beesley, Mr. Lawrence	male	34.00	0	0	248698	13.0000	D56	S
22	23	1	3	McGowan, Miss. Anna "Annie"	female	15.00	0	0	330923	8.0292	NaN	Q
23	24	1	1	Sloper, Mr. William Thompson	male	28.00	0	0	113788	35.5000	A6	S
24	25	0	3	Palsson, Miss. Torborg Danira	female	8.00	3	1	349909	21.0750	NaN	S
25	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia...	female	38.00	1	5	347077	31.3875	NaN	S
26	27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	2631	7.2250	NaN	C
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.00	3	2	19950	263.0000	C23 C25 C27	S
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NaN	0	0	330959	7.8792	NaN	Q
29	30	0	3	Todoroff, Mr. Lalio	male	NaN	0	0	349216	7.8958	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
50	51	0	3	Panula, Master. Juha Niilo	male	7.00	4	1	3101295	39.6875	NaN	S
51	52	0	3	Nosworthy, Mr. Richard Cater	male	21.00	0	0	A/4. 39886	7.8000	NaN	S
52	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.00	1	0	PC 17572	76.7292	D33	C
53	54	1	2	Faunthorpe, Mrs. Lizzie (Elizabeth Anne Wilkin...	female	29.00	1	0	2926	26.0000	NaN	S

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
54	55	0	1	Ostby, Mr. Engelhart Cornelius	male	65.00	0	1	113509	61.9792	B30	C
55	56	1	1	Woolner, Mr. Hugh	male	NaN	0	0	19947	35.5000	C52	S
56	57	1	2	Rugg, Miss. Emily	female	21.00	0	0	C.A. 31026	10.5000	NaN	S
57	58	0	3	Novel, Mr. Mansouer	male	28.50	0	0	2697	7.2292	NaN	C
58	59	1	2	West, Miss. Constance Mirium	female	5.00	1	2	C.A. 34651	27.7500	NaN	S
59	60	0	3	Goodwin, Master. William Frederick	male	11.00	5	2	CA 2144	46.9000	NaN	S
60	61	0	3	Sirayanian, Mr. Orsen	male	22.00	0	0	2669	7.2292	NaN	C
61	62	1	1	Icard, Miss. Amelie	female	38.00	0	0	113572	80.0000	B28	NaN
62	63	0	1	Harris, Mr. Henry Birkhardt	male	45.00	1	0	36973	83.4750	C83	S
63	64	0	3	Skoog, Master. Harald	male	4.00	3	2	347088	27.9000	NaN	S
64	65	0	1	Stewart, Mr. Albert A	male	NaN	0	0	PC 17605	27.7208	NaN	C
65	66	1	3	Moubarek, Master. Gerios	male	NaN	1	1	2661	15.2458	NaN	C
66	67	1	2	Nye, Mrs. (Elizabeth Ramell)	female	29.00	0	0	C.A. 29395	10.5000	F33	S
67	68	0	3	Crease, Mr. Ernest James	male	19.00	0	0	S.P. 3464	8.1583	NaN	S
68	69	1	3	Andersson, Miss. Erna Alexandra	female	17.00	4	2	3101281	7.9250	NaN	S
69	70	0	3	Kink, Mr. Vincenz	male	26.00	2	0	315151	8.6625	NaN	S
70	71	0	2	Jenkin, Mr. Stephen Curnow	male	32.00	0	0	C.A. 33111	10.5000	NaN	S
71	72	0	3	Goodwin, Miss. Lillian Amy	female	16.00	5	2	CA 2144	46.9000	NaN	S
72	73	0	2	Hood, Mr. Ambrose Jr	male	21.00	0	0	S.O.C. 14879	73.5000	NaN	S
73	74	0	3	Chronopoulos, Mr. Apostolos	male	26.00	1	0	2680	14.4542	NaN	C

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
74	75	1	3	Bing, Mr. Lee	male	32.00	0	0	1601	56.4958	NaN	S
75	76	0	3	Moen, Mr. Sigurd Hansen	male	25.00	0	0	348123	7.6500	F G73	S
76	77	0	3	Staneff, Mr. Ivan	male	NaN	0	0	349208	7.8958	NaN	S
77	78	0	3	Moutal, Mr. Rahamin Haim	male	NaN	0	0	374746	8.0500	NaN	S
78	79	1	2	Caldwell, Master. Alden Gates	male	0.83	0	2	248738	29.0000	NaN	S
79	80	1	3	Dowdell, Miss. Elizabeth	female	30.00	0	0	364516	12.4750	NaN	S

80 rows × 12 columns

Data Dictionary

- Survived: 0 = No, 1 = Yes
- pclass: Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
- sibsp: # of siblings / spouses aboard the Titanic
- parch: # of parents / children aboard the Titanic
- ticket: Ticket number
- cabin: Cabin number
- embarked: Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

Total rows and columns

We can see that there are 891 rows and 12 columns in our training dataset.

Out[4]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

```
Out[5]: (891, 12)
```

```
Out[6]: (418, 11)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
```

```
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
PassengerId    418 non-null int64
Pclass         418 non-null int64
Name           418 non-null object
Sex            418 non-null object
Age            332 non-null float64
SibSp          418 non-null int64
Parch          418 non-null int64
Ticket         418 non-null object
Fare           417 non-null float64
Cabin          91 non-null object
Embarked       418 non-null object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
```

We can see that *Age* value is missing for many rows.

Out of 891 rows, the *Age* value is present only in 714 rows.

Similarly, *Cabin* values are also missing in many rows. Only 204 out of 891 rows have *Cabin* values.

```
Out[9]:  PassengerId      0
         Survived         0
         Pclass           0
         Name             0
         Sex              0
         Age            177
         SibSp            0
         Parch            0
         Ticket           0
         Fare             0
         Cabin          687
         Embarked         2
         dtype: int64
```

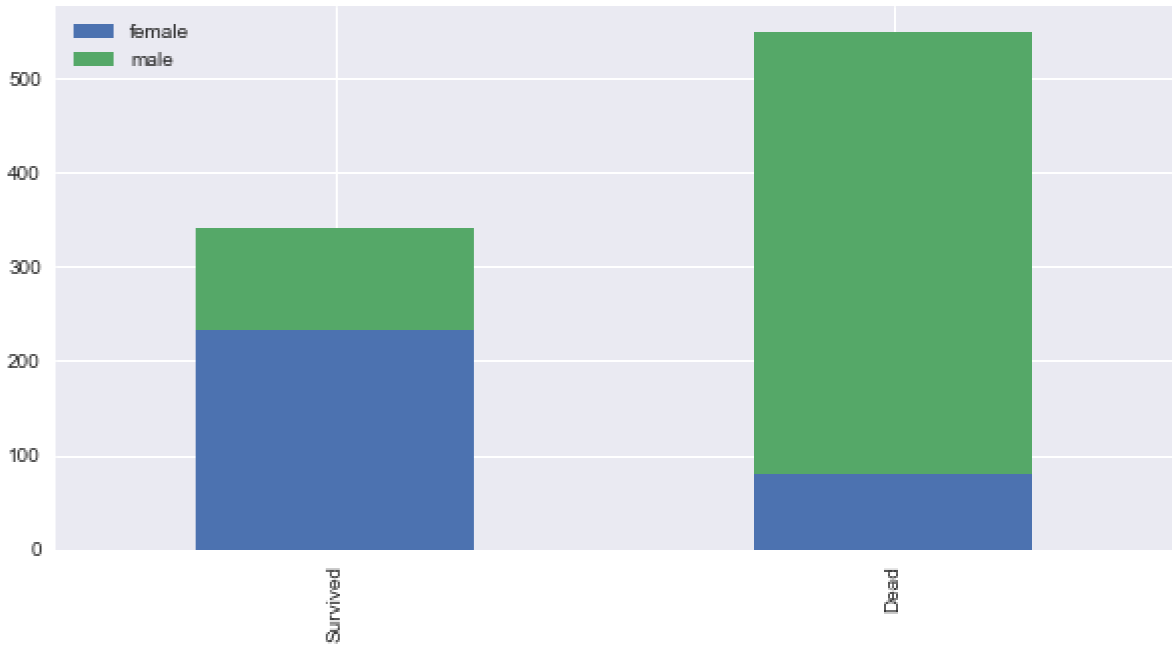
```
Out[10]: PassengerId      0
         Pclass           0
         Name             0
         Sex              0
         Age             86
         SibSp            0
         Parch            0
         Ticket           0
         Fare             1
         Cabin          327
         Embarked         0
         dtype: int64
```

There are 177 rows with missing *Age*, 687 rows with missing *Cabin* and 2 rows with missing *Embarked* information.

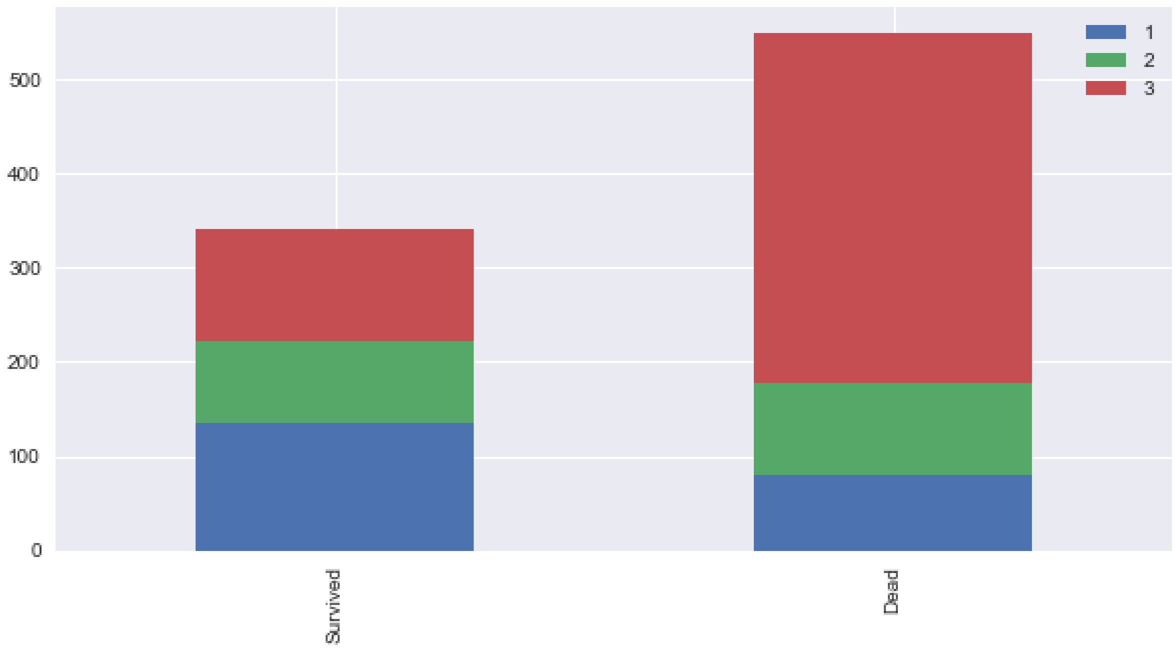
import python lib for visualization

Bar Chart for Categorical Features

- Pclass
- Sex
- SibSp (# of siblings and spouse)
- Parch (# of parents and children)
- Embarked
- Cabin

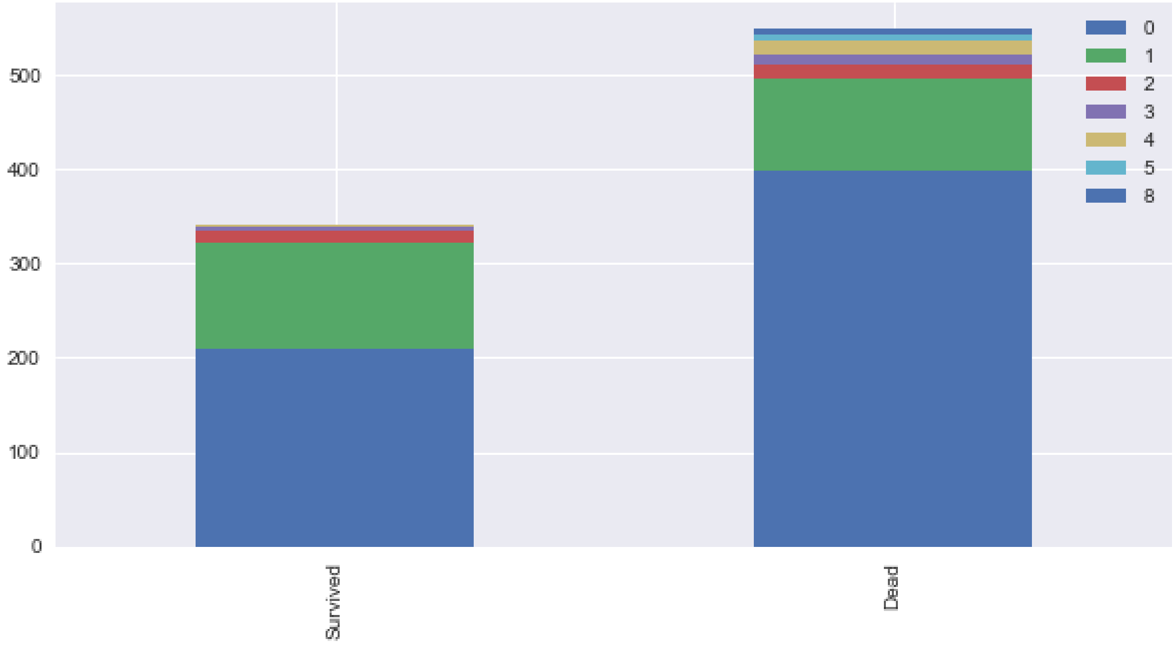


The Chart confirms **Women** more likely survivied than **Men**

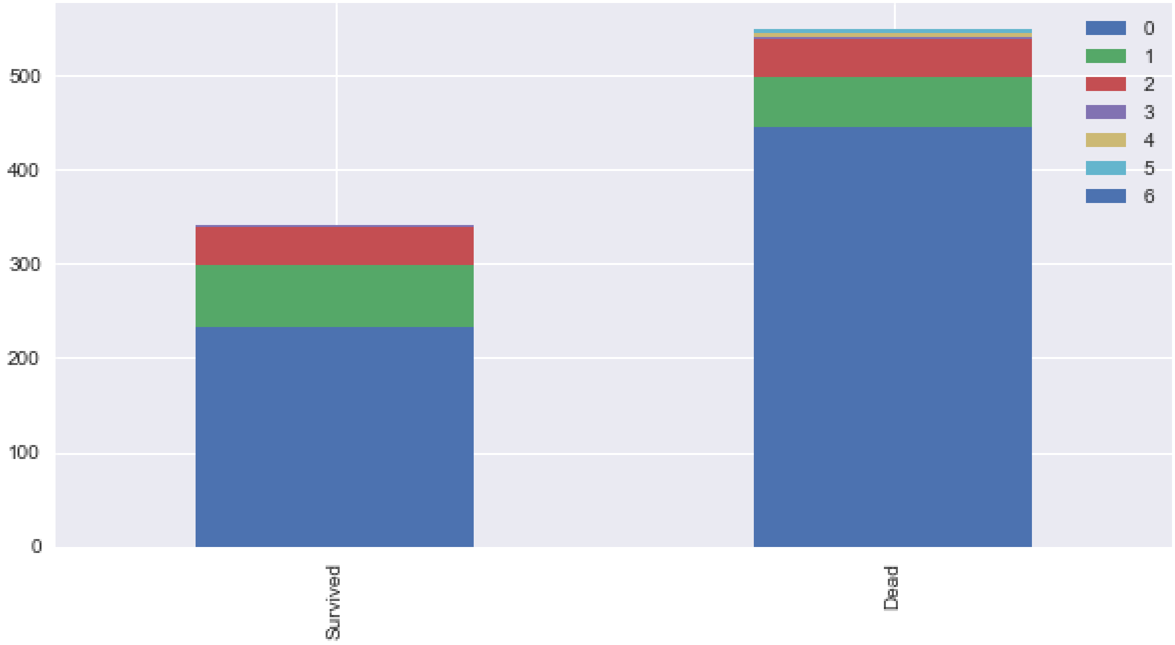


The Chart confirms **1st class** more likely survivied than **other classes**

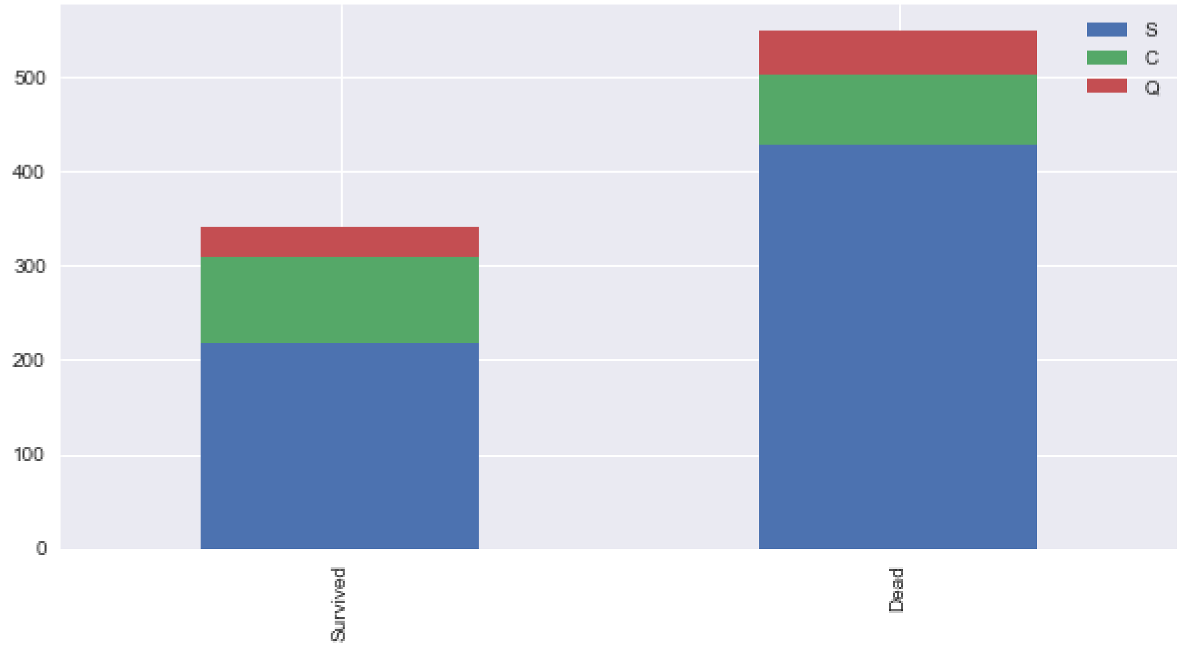
The Chart confirms **3rd class** more likely dead than **other classes**



The Chart confirms **a person aboarded with more than 2 siblings or spouse** more likely survived
The Chart confirms **a person aboarded without siblings or spouse** more likely dead



The Chart confirms **a person aboarded with more than 2 parents or children** more likely survived
The Chart confirms **a person aboarded alone** more likely dead



The Chart confirms **a person aboarded from C** slightly more likely survived
The Chart confirms **a person aboarded from Q** more likely dead
The Chart confirms **a person aboarded from S** more likely dead

4. Feature engineering

Feature engineering is the process of using domain knowledge of the data
to create features (**feature vectors**) that make machine learning algorithms work.

feature vector is an n-dimensional vector of numerical features that represent some object.
Many algorithms in machine learning require a numerical representation of objects,
since such representations facilitate processing and statistical analysis.

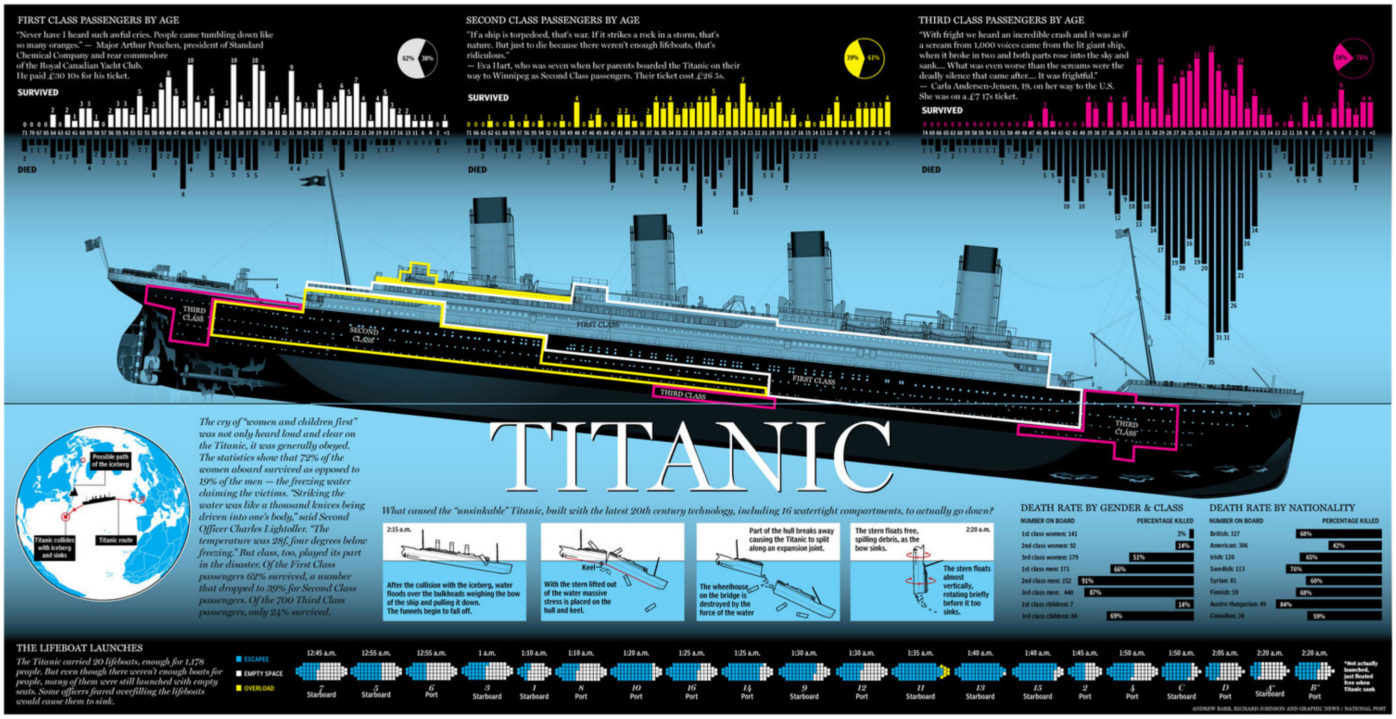
Out[18]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

4.1 how titanic sank?

sank from the bow of the ship where third class rooms located

conclusion, Pclass is key feature for classifier



Out[20]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

4.2 Name

Out[22]:

```
Mr          517
Miss        182
Mrs         125
Master       40
Dr            7
Rev           6
Col           2
Major         2
Mlle          2
Countess      1
Ms            1
Lady          1
Jonkheer      1
Don           1
Mme           1
Capt          1
Sir           1
Name: Title, dtype: int64
```

Out[23]:

```
Mr        240
Miss       78
Mrs        72
Master     21
Col         2
Rev         2
Dona        1
Ms          1
Dr          1
Name: Title, dtype: int64
```

Title map

Mr : 0

Miss : 1

Mrs: 2

Others: 3

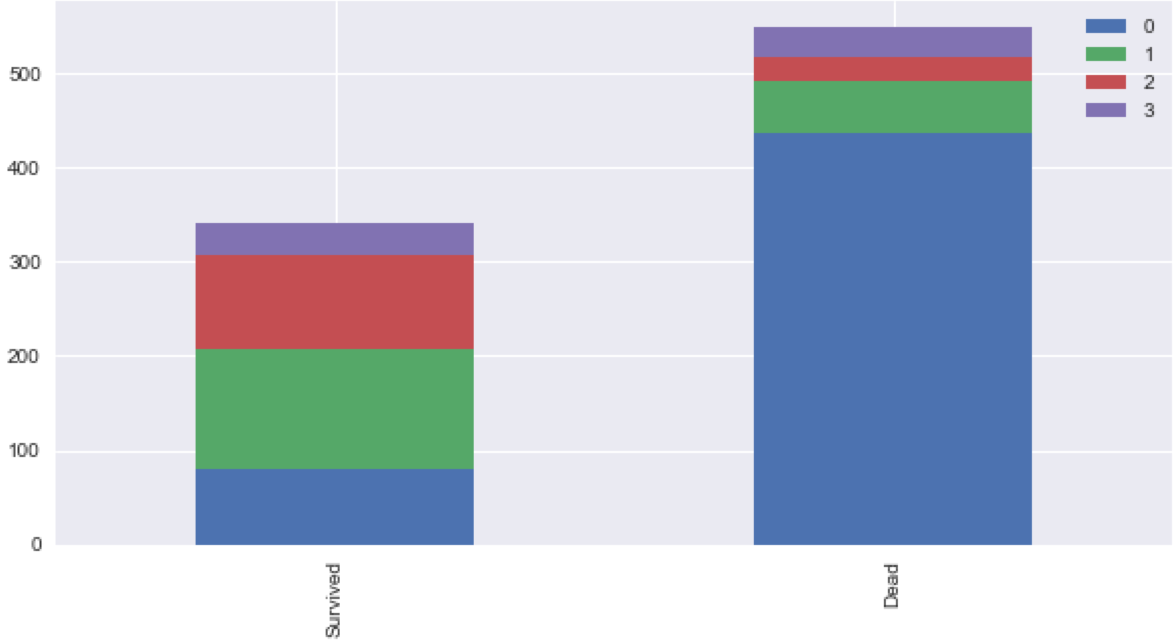
Out[25]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	0

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	2
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	0

Out[26]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q	0
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S	2
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	0
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S	0
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S	2



Out[29]:

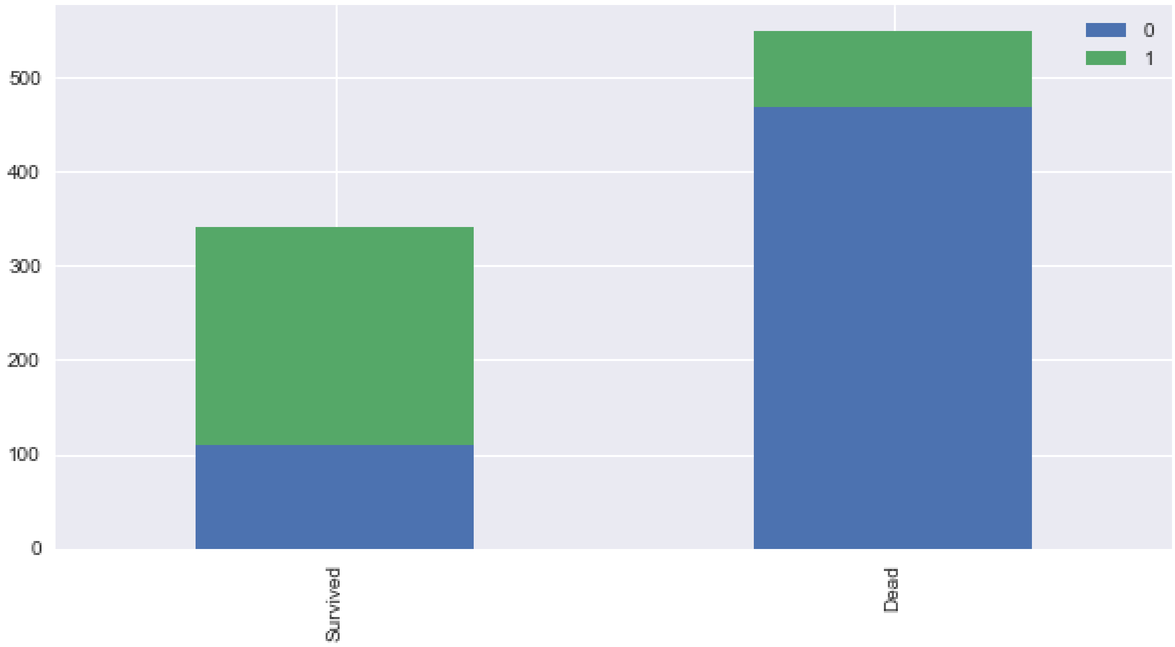
	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	male	22.0	1	0	A/5 21171	7.2500	NaN	S	0
1	2	1	1	female	38.0	1	0	PC 17599	71.2833	C85	C	2
2	3	1	3	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
3	4	1	1	female	35.0	1	0	113803	53.1000	C123	S	2
4	5	0	3	male	35.0	0	0	373450	8.0500	NaN	S	0

Out[30]:

	PassengerId	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	892	3	male	34.5	0	0	330911	7.8292	NaN	Q	0
1	893	3	female	47.0	1	0	363272	7.0000	NaN	S	2
2	894	2	male	62.0	0	0	240276	9.6875	NaN	Q	0
3	895	3	male	27.0	0	0	315154	8.6625	NaN	S	0
4	896	3	female	22.0	1	1	3101298	12.2875	NaN	S	2

4.3 Sex

male: 0 female: 1



4.4 Age

4.4.1 some age is missing

Let's use Title's median age for missing Age

Out[33]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	0	22.00	1	0	A/5 21171	7.2500	NaN	S	0
1	2	1	1	1	38.00	1	0	PC 17599	71.2833	C85	C	2

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
2	3	1	3	1	26.00	0	0	STON/O2. 3101282	7.9250	NaN	S	1
3	4	1	1	1	35.00	1	0	113803	53.1000	C123	S	2
4	5	0	3	0	35.00	0	0	373450	8.0500	NaN	S	0
5	6	0	3	0	NaN	0	0	330877	8.4583	NaN	Q	0
6	7	0	1	0	54.00	0	0	17463	51.8625	E46	S	0
7	8	0	3	0	2.00	3	1	349909	21.0750	NaN	S	3
8	9	1	3	1	27.00	0	2	347742	11.1333	NaN	S	2
9	10	1	2	1	14.00	1	0	237736	30.0708	NaN	C	2
10	11	1	3	1	4.00	1	1	PP 9549	16.7000	G6	S	1
11	12	1	1	1	58.00	0	0	113783	26.5500	C103	S	1
12	13	0	3	0	20.00	0	0	A/5. 2151	8.0500	NaN	S	0
13	14	0	3	0	39.00	1	5	347082	31.2750	NaN	S	0
14	15	0	3	1	14.00	0	0	350406	7.8542	NaN	S	1
15	16	1	2	1	55.00	0	0	248706	16.0000	NaN	S	2
16	17	0	3	0	2.00	4	1	382652	29.1250	NaN	Q	3
17	18	1	2	0	NaN	0	0	244373	13.0000	NaN	S	0
18	19	0	3	1	31.00	1	0	345763	18.0000	NaN	S	2
19	20	1	3	1	NaN	0	0	2649	7.2250	NaN	C	2
20	21	0	2	0	35.00	0	0	239865	26.0000	NaN	S	0
21	22	1	2	0	34.00	0	0	248698	13.0000	D56	S	0
22	23	1	3	1	15.00	0	0	330923	8.0292	NaN	Q	1
23	24	1	1	0	28.00	0	0	113788	35.5000	A6	S	0
24	25	0	3	1	8.00	3	1	349909	21.0750	NaN	S	1
25	26	1	3	1	38.00	1	5	347077	31.3875	NaN	S	2
26	27	0	3	0	NaN	0	0	2631	7.2250	NaN	C	0
27	28	0	1	0	19.00	3	2	19950	263.0000	C23 C25 C27	S	0
28	29	1	3	1	NaN	0	0	330959	7.8792	NaN	Q	1
29	30	0	3	0	NaN	0	0	349216	7.8958	NaN	S	0
...	...	...	...	...	...	...	...	...	...	...	...	...
70	71	0	2	0	32.00	0	0	C.A. 33111	10.5000	NaN	S	0
71	72	0	3	1	16.00	5	2	CA 2144	46.9000	NaN	S	1
72	73	0	2	0	21.00	0	0	S.O.C. 14879	73.5000	NaN	S	0
73	74	0	3	0	26.00	1	0	2680	14.4542	NaN	C	0
74	75	1	3	0	32.00	0	0	1601	56.4958	NaN	S	0
75	76	0	3	0	25.00	0	0	348123	7.6500	F G73	S	0

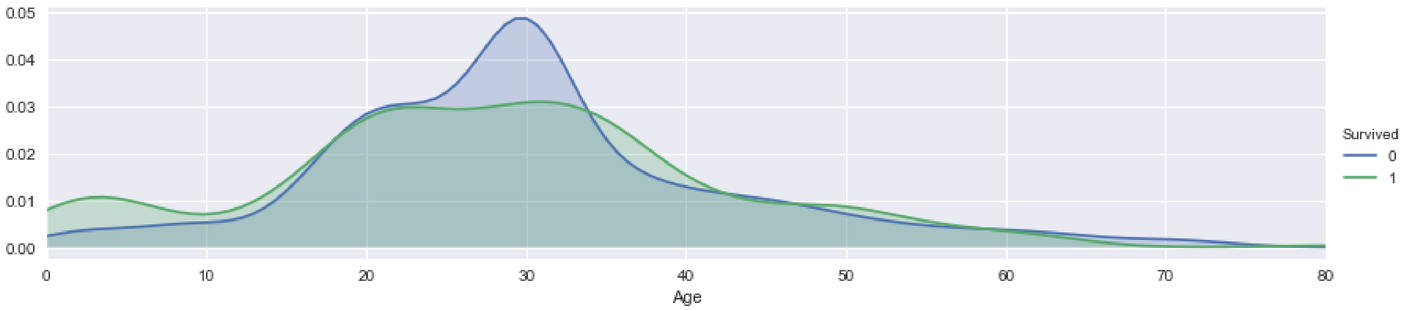
	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
76	77	0	3	0	NaN	0	0	349208	7.8958	NaN	S	0
77	78	0	3	0	NaN	0	0	374746	8.0500	NaN	S	0
78	79	1	2	0	0.83	0	2	248738	29.0000	NaN	S	3
79	80	1	3	1	30.00	0	0	364516	12.4750	NaN	S	1
80	81	0	3	0	22.00	0	0	345767	9.0000	NaN	S	0
81	82	1	3	0	29.00	0	0	345779	9.5000	NaN	S	0
82	83	1	3	1	NaN	0	0	330932	7.7875	NaN	Q	1
83	84	0	1	0	28.00	0	0	113059	47.1000	NaN	S	0
84	85	1	2	1	17.00	0	0	SO/C 14885	10.5000	NaN	S	1
85	86	1	3	1	33.00	3	0	3101278	15.8500	NaN	S	2
86	87	0	3	0	16.00	1	3	W./C. 6608	34.3750	NaN	S	0
87	88	0	3	0	NaN	0	0	SOTON/OQ 392086	8.0500	NaN	S	0
88	89	1	1	1	23.00	3	2	19950	263.0000	C23 C25 C27	S	1
89	90	0	3	0	24.00	0	0	343275	8.0500	NaN	S	0
90	91	0	3	0	29.00	0	0	343276	8.0500	NaN	S	0
91	92	0	3	0	20.00	0	0	347466	7.8542	NaN	S	0
92	93	0	1	0	46.00	1	0	W.E.P. 5734	61.1750	E31	S	0
93	94	0	3	0	26.00	1	2	C.A. 2315	20.5750	NaN	S	0
94	95	0	3	0	59.00	0	0	364500	7.2500	NaN	S	0
95	96	0	3	0	NaN	0	0	374910	8.0500	NaN	S	0
96	97	0	1	0	71.00	0	0	PC 17754	34.6542	A5	C	0
97	98	1	1	0	23.00	0	1	PC 17759	63.3583	D10 D12	C	0
98	99	1	2	1	34.00	0	1	231919	23.0000	NaN	S	2
99	100	0	2	0	34.00	1	0	244367	26.0000	NaN	S	0

100 rows × 12 columns

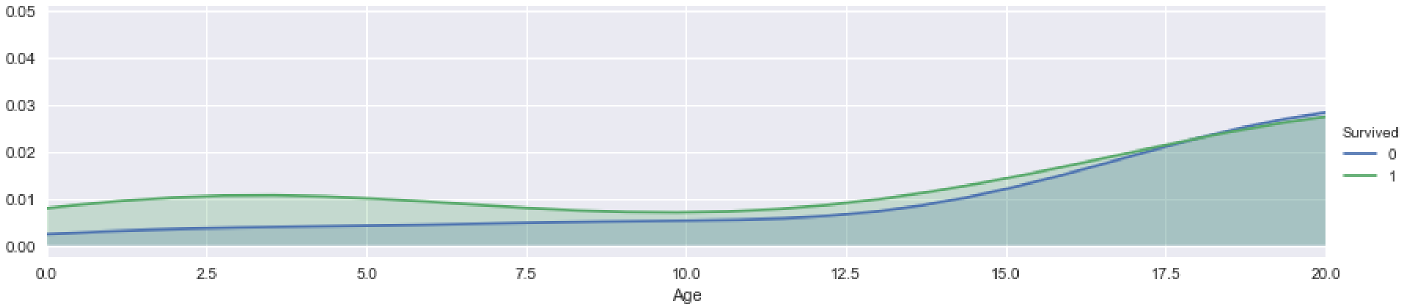
Out[35]:

```
0       30.0
1       35.0
2       21.0
3       35.0
4       30.0
5       30.0
6       30.0
7        9.0
8       35.0
9       35.0
10      21.0
11      21.0
12      30.0
13      30.0
14      21.0
```

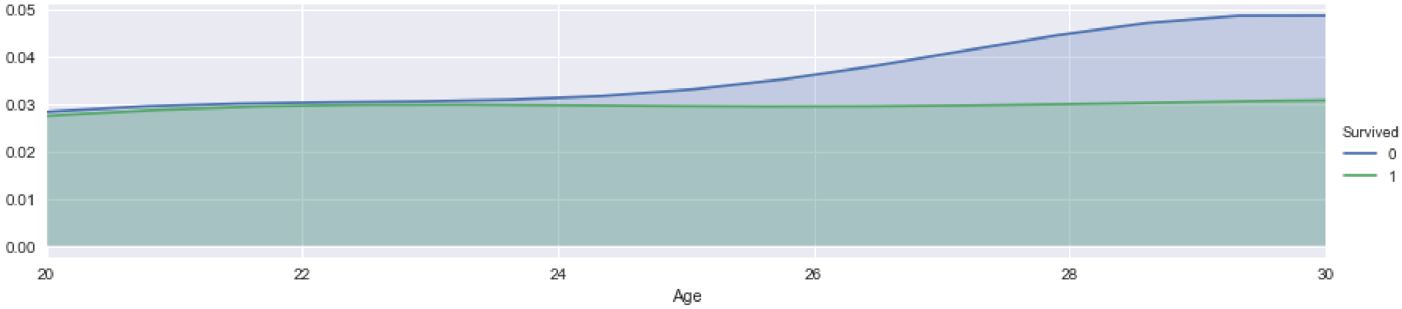
```
15     35.0
16      9.0
17     30.0
18     35.0
19     35.0
20     30.0
21     30.0
22     21.0
23     30.0
24     21.0
25     35.0
26     30.0
27     30.0
28     21.0
29     30.0
       ...
861    30.0
862    35.0
863    21.0
864    30.0
865    35.0
866    21.0
867    30.0
868    30.0
869     9.0
870    30.0
871    35.0
872    30.0
873    30.0
874    35.0
875    21.0
876    30.0
877    30.0
878    30.0
879    35.0
880    35.0
881    30.0
882    21.0
883    30.0
884    30.0
885    35.0
886     9.0
887    21.0
888    21.0
889    30.0
890    30.0
Name: Age, Length: 891, dtype: float64
```



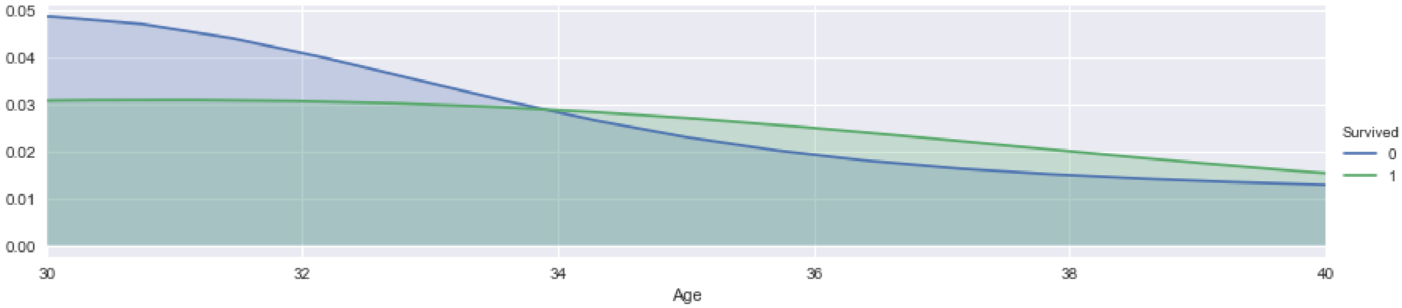
Out[37]:
```
(0, 20)
```



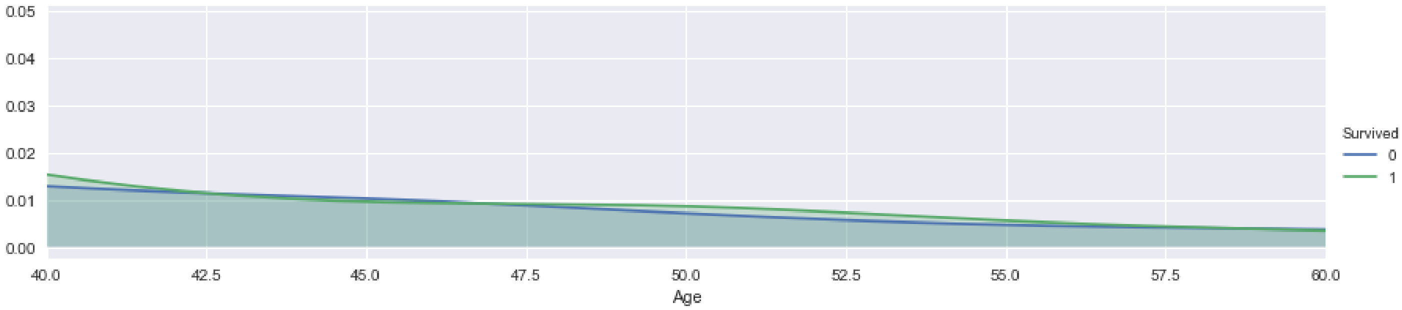
```
Out[38]: (20, 30)
```



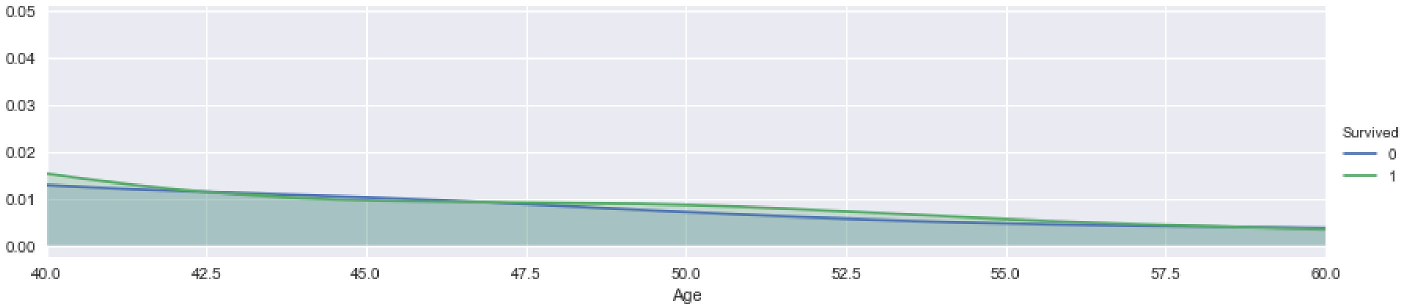
```
Out[39]: (30, 40)
```



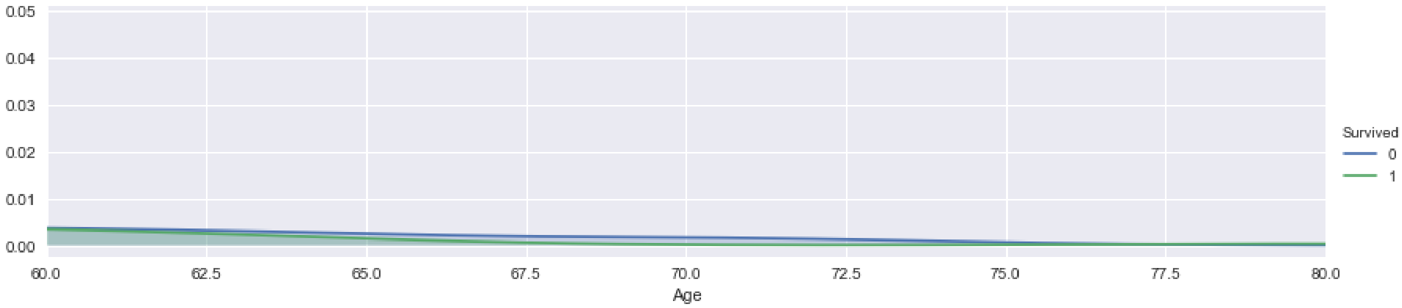
```
Out[40]: (40, 60)
```



```
Out[41]: (40, 60)
```



```
Out[42]: (60, 80.0)
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Sex            891 non-null int64
Age            891 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
Title          891 non-null int64
dtypes: float64(2), int64(7), object(3)
memory usage: 83.6+ KB

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
PassengerId    418 non-null int64
Pclass         418 non-null int64
Sex            418 non-null int64
Age            418 non-null float64
SibSp          418 non-null int64
Parch          418 non-null int64
Ticket         418 non-null object
Fare           417 non-null float64
Cabin          91 non-null object
Embarked       418 non-null object
Title          418 non-null int64
dtypes: float64(2), int64(6), object(3)
memory usage: 36.0+ KB
```

4.4.2 Binning

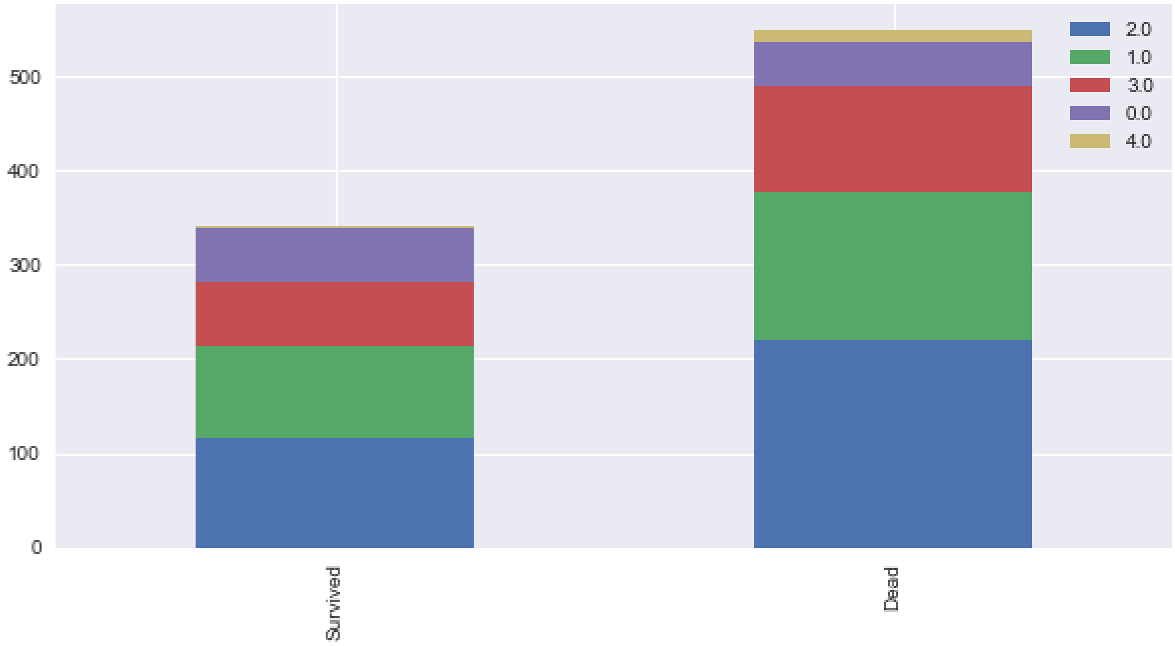
Binning/Converting Numerical Age to Categorical Variable

feature vector map:
child: 0
young: 1
adult: 2
mid-age: 3
senior: 4

Out[46]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	0	1.0	1	0	A/5 21171	7.2500	NaN	S	0
1	2	1	1	1	3.0	1	0	PC 17599	71.2833	C85	C	2

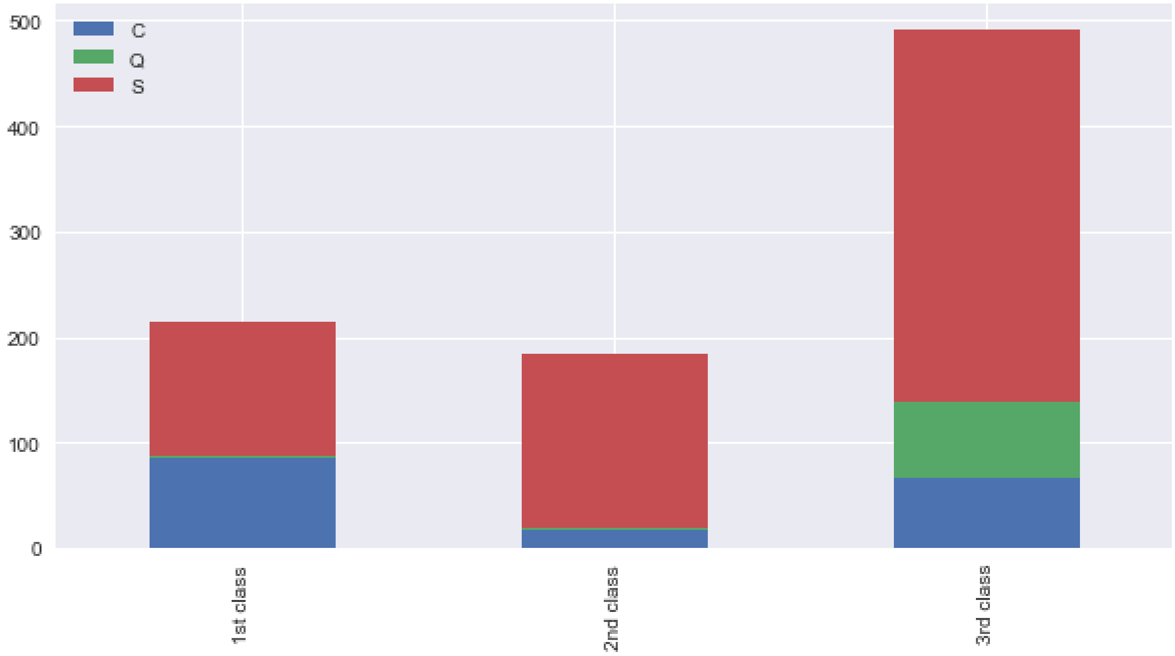
	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
2	3	1	3	1	1.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
3	4	1	1	1	2.0	1	0	113803	53.1000	C123	S	2
4	5	0	3	0	2.0	0	0	373450	8.0500	NaN	S	0



4.5 Embarked

4.5.1 filling missing values

```
Out[48]:  <matplotlib.axes._subplots.AxesSubplot at 0x1113ee790>
```



more than 50% of 1st class are from S embark
more than 50% of 2nd class are from S embark
more than 50% of 3rd class are from S embark

fill out missing embark with S embark

Out[50]:

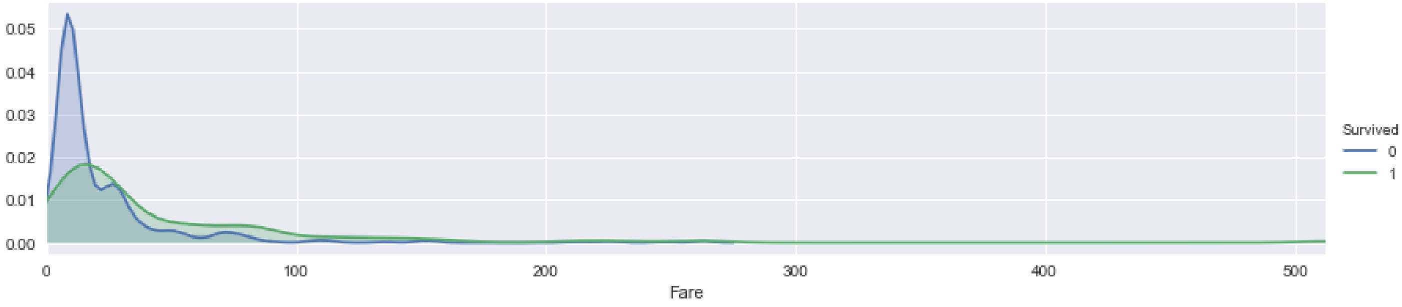
	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	0	1.0	1	0	A/5 21171	7.2500	NaN	S	0
1	2	1	1	1	3.0	1	0	PC 17599	71.2833	C85	C	2
2	3	1	3	1	1.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
3	4	1	1	1	2.0	1	0	113803	53.1000	C123	S	2
4	5	0	3	0	2.0	0	0	373450	8.0500	NaN	S	0

4.6 Fare

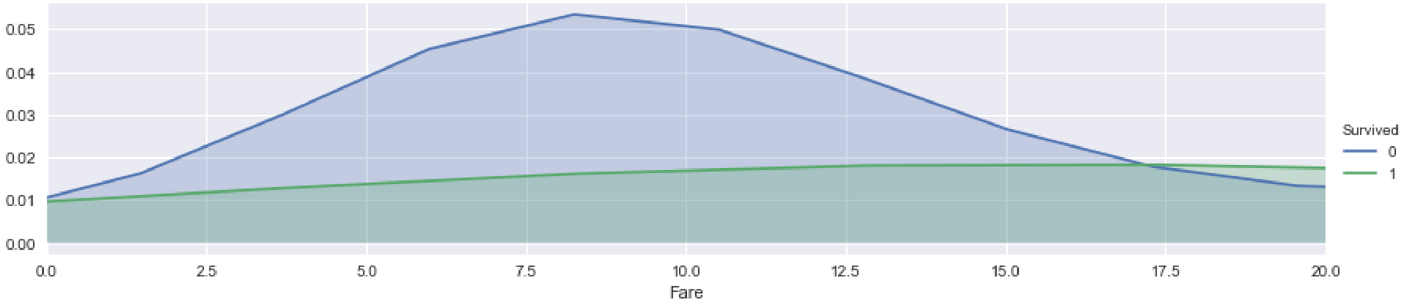
Out[52]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	0	1.0	1	0	A/5 21171	7.2500	NaN	0	0
1	2	1	1	1	3.0	1	0	PC 17599	71.2833	C85	1	2
2	3	1	3	1	1.0	0	0	STON/O2. 3101282	7.9250	NaN	0	1
3	4	1	1	1	2.0	1	0	113803	53.1000	C123	0	2
4	5	0	3	0	2.0	0	0	373450	8.0500	NaN	0	0
5	6	0	3	0	2.0	0	0	330877	8.4583	NaN	2	0
6	7	0	1	0	3.0	0	0	17463	51.8625	E46	0	0
7	8	0	3	0	0.0	3	1	349909	21.0750	NaN	0	3
8	9	1	3	1	2.0	0	2	347742	11.1333	NaN	0	2
9	10	1	2	1	0.0	1	0	237736	30.0708	NaN	1	2
10	11	1	3	1	0.0	1	1	PP 9549	16.7000	G6	0	1
11	12	1	1	1	3.0	0	0	113783	26.5500	C103	0	1
12	13	0	3	0	1.0	0	0	A/5. 2151	8.0500	NaN	0	0
13	14	0	3	0	3.0	1	5	347082	31.2750	NaN	0	0
14	15	0	3	1	0.0	0	0	350406	7.8542	NaN	0	1
15	16	1	2	1	3.0	0	0	248706	16.0000	NaN	0	2
16	17	0	3	0	0.0	4	1	382652	29.1250	NaN	2	3
17	18	1	2	0	2.0	0	0	244373	13.0000	NaN	0	0
18	19	0	3	1	2.0	1	0	345763	18.0000	NaN	0	2
19	20	1	3	1	2.0	0	0	2649	7.2250	NaN	1	2
20	21	0	2	0	2.0	0	0	239865	26.0000	NaN	0	0
21	22	1	2	0	2.0	0	0	248698	13.0000	D56	0	0
22	23	1	3	1	0.0	0	0	330923	8.0292	NaN	2	1
23	24	1	1	0	2.0	0	0	113788	35.5000	A6	0	0
24	25	0	3	1	0.0	3	1	349909	21.0750	NaN	0	1
25	26	1	3	1	3.0	1	5	347077	31.3875	NaN	0	2

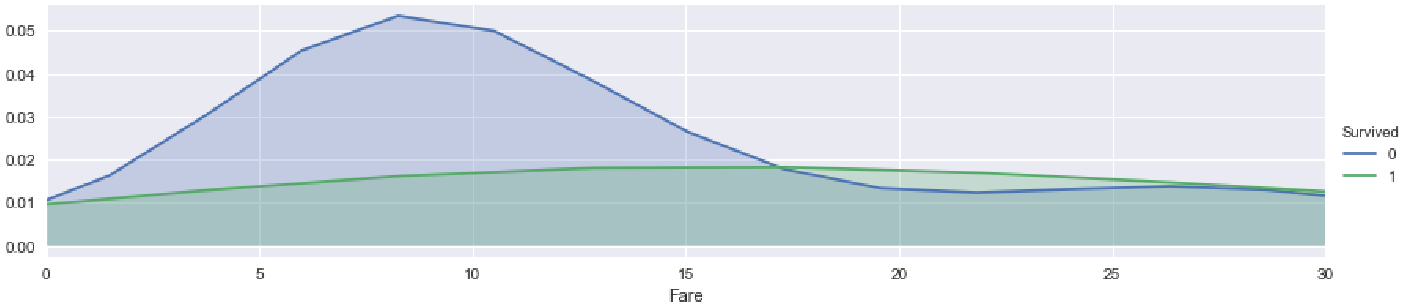
	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
26	27	0	3	0	2.0	0	0	2631	7.2250	NaN	1	0
27	28	0	1	0	1.0	3	2	19950	263.0000	C23 C25 C27	0	0
28	29	1	3	1	1.0	0	0	330959	7.8792	NaN	2	1
29	30	0	3	0	2.0	0	0	349216	7.8958	NaN	0	0
30	31	0	1	0	3.0	0	0	PC 17601	27.7208	NaN	1	3
31	32	1	1	1	2.0	1	0	PC 17569	146.5208	B78	1	2
32	33	1	3	1	1.0	0	0	335677	7.7500	NaN	2	1
33	34	0	2	0	4.0	0	0	C.A. 24579	10.5000	NaN	0	0
34	35	0	1	0	2.0	1	0	PC 17604	82.1708	NaN	1	0
35	36	0	1	0	3.0	1	0	113789	52.0000	NaN	0	0
36	37	1	3	0	2.0	0	0	2677	7.2292	NaN	1	0
37	38	0	3	0	1.0	0	0	A./5. 2152	8.0500	NaN	0	0
38	39	0	3	1	1.0	2	0	345764	18.0000	NaN	0	1
39	40	1	3	1	0.0	1	0	2651	11.2417	NaN	1	1
40	41	0	3	1	3.0	1	0	7546	9.4750	NaN	0	2
41	42	0	2	1	2.0	1	0	11668	21.0000	NaN	0	2
42	43	0	3	0	2.0	0	0	349253	7.8958	NaN	1	0
43	44	1	2	1	0.0	1	2	SC/Paris 2123	41.5792	NaN	1	1
44	45	1	3	1	1.0	0	0	330958	7.8792	NaN	2	1
45	46	0	3	0	2.0	0	0	S.C./A.4. 23567	8.0500	NaN	0	0
46	47	0	3	0	2.0	1	0	370371	15.5000	NaN	2	0
47	48	1	3	1	1.0	0	0	14311	7.7500	NaN	2	1
48	49	0	3	0	2.0	2	0	2662	21.6792	NaN	1	0
49	50	0	3	1	1.0	1	0	349237	17.8000	NaN	0	2



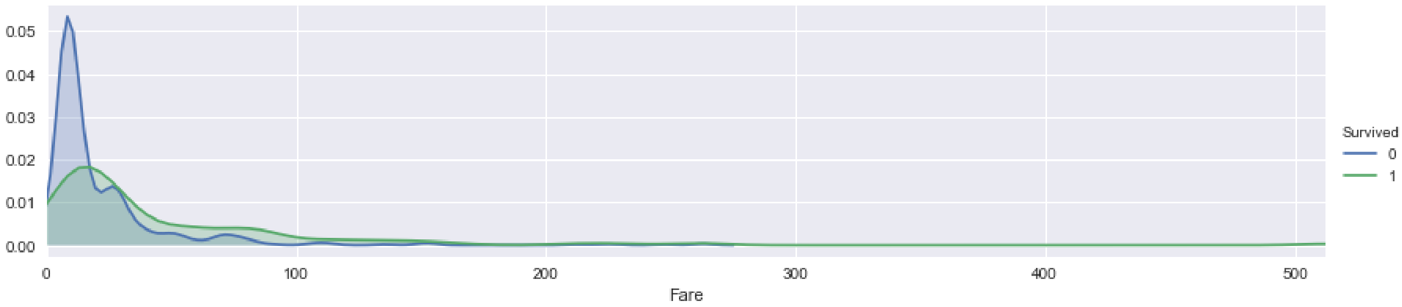
```
Out[54]: (0, 20)
```



```
Out[55]: (0, 30)
```



```
Out[56]: (0, 512.32920000000001)
```



Out[58]:

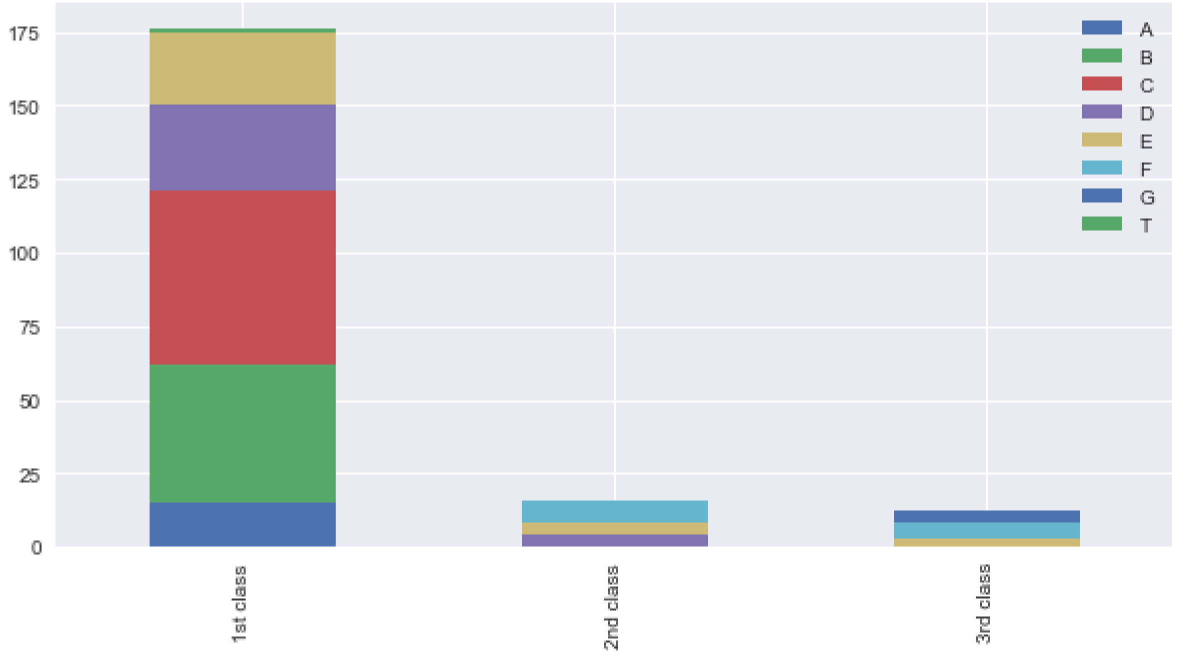
	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
0	1	0	3	0	1.0	1	0	A/5 21171	0.0	NaN	0	0
1	2	1	1	1	3.0	1	0	PC 17599	2.0	C85	1	2
2	3	1	3	1	1.0	0	0	STON/O2. 3101282	0.0	NaN	0	1
3	4	1	1	1	2.0	1	0	113803	2.0	C123	0	2
4	5	0	3	0	2.0	0	0	373450	0.0	NaN	0	0

4.7 Cabin

```
Out[59]: C23 C25 C27            4
         G6                     4
         B96 B98                4
         D                      3
         C22 C26                3
         E101                   3
         F2                     3
         F33                    3
         B57 B59 B63 B66        2
         C68                    2
         B58 B60                2
         E121                   2
         D20                    2
```

```
E8                   2
E44                  2
B77                  2
C65                  2
D26                  2
E24                  2
E25                  2
B20                  2
C93                  2
D33                  2
E67                  2
D35                  2
D36                  2
C52                  2
F4                   2
C125                 2
C124                 2
                    ..
F G63                1
A6                   1
D45                  1
D6                   1
D56                  1
C101                 1
C54                  1
D28                  1
D37                  1
B102                 1
D30                  1
E17                  1
E58                  1
F E69                1
D10 D12              1
E50                  1
A14                  1
C91                  1
A16                  1
B38                  1
B39                  1
C95                  1
B78                  1
B79                  1
C99                  1
B37                  1
A19                  1
E12                  1
A7                   1
D15                  1
Name: Cabin, Length: 147, dtype: int64
```

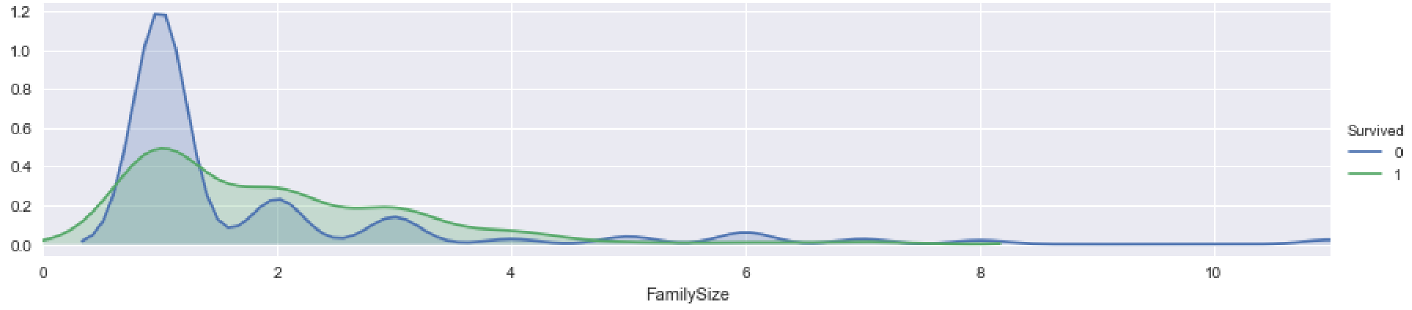
```
Out[61]: <matplotlib.axes._subplots.AxesSubplot at 0x1121b2d10>
```



4.8 FamilySize

Out[65]:

```
(0, 11.0)
```



Out[67]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	FamilySize
0	1	0	3	0	1.0	1	0	A/5 21171	0.0	2.0	0	0	0.4
1	2	1	1	1	3.0	1	0	PC 17599	2.0	0.8	1	2	0.4
2	3	1	3	1	1.0	0	0	STON/O2. 3101282	0.0	2.0	0	1	0.0
3	4	1	1	1	2.0	1	0	113803	2.0	0.8	0	2	0.4
4	5	0	3	0	2.0	0	0	373450	0.0	2.0	0	0	0.0

Out[68]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	FamilySize
0	1	0	3	0	1.0	1	0	A/5 21171	0.0	2.0	0	0	0.4
1	2	1	1	1	3.0	1	0	PC 17599	2.0	0.8	1	2	0.4
2	3	1	3	1	1.0	0	0	STON/O2. 3101282	0.0	2.0	0	1	0.0
3	4	1	1	1	2.0	1	0	113803	2.0	0.8	0	2	0.4
4	5	0	3	0	2.0	0	0	373450	0.0	2.0	0	0	0.0

```
Out[70]: ((891, 8), (891,))
```

Out[71]:

	Pclass	Sex	Age	Fare	Cabin	Embarked	Title	FamilySize
0	3	0	1.0	0.0	2.0	0	0	0.4
1	1	1	3.0	2.0	0.8	1	2	0.4
2	3	1	1.0	0.0	2.0	0	1	0.0
3	1	1	2.0	2.0	0.8	0	2	0.4
4	3	0	2.0	0.0	2.0	0	0	0.0
5	3	0	2.0	0.0	2.0	2	0	0.0
6	1	0	3.0	2.0	1.6	0	0	0.0
7	3	0	0.0	1.0	2.0	0	3	1.6
8	3	1	2.0	0.0	2.0	0	2	0.8
9	2	1	0.0	2.0	1.8	1	2	0.4

5. Modelling

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
Survived      891 non-null int64
Pclass        891 non-null int64
Sex           891 non-null int64
Age           891 non-null float64
Fare          891 non-null float64
Cabin         891 non-null float64
Embarked      891 non-null int64
Title         891 non-null int64
FamilySize    891 non-null float64
dtypes: float64(4), int64(5)
memory usage: 62.7 KB
```

6.2 Cross Validation (K-fold)

6.2.1 kNN

```
[ 0.82222222  0.76404494  0.80898876  0.83146067  0.87640449  0.82022472
  0.85393258  0.79775281  0.84269663  0.84269663]
```

```
Out[78]: 82.6
```

6.2.2 Decision Tree

```
[ 0.76666667  0.82022472  0.78651685  0.76404494  0.88764045  0.76404494
  0.82022472  0.82022472  0.74157303  0.79775281]
```

```
Out[80]: 79.69
```

6.2.3 Ramdom Forest

```
[ 0.77777778  0.80898876  0.82022472  0.76404494  0.86516854  0.82022472
  0.79775281  0.80898876  0.76404494  0.83146067]
```

Out[82]:

```
80.59
```

6.2.4 Naive Bayes

```
[ 0.85555556  0.73033708  0.75280899  0.75280899  0.70786517  0.80898876
  0.76404494  0.80898876  0.86516854  0.83146067]
```

Out[84]:

```
78.78
```

6.2.5 SVM

```
[ 0.83333333  0.80898876  0.83146067  0.82022472  0.84269663  0.82022472
  0.84269663  0.85393258  0.83146067  0.86516854]
```

Out[86]:

```
83.5
```

7. Testing

Out[89]:

	PassengerId	Survived
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1