Titanic Dataset Analysis

Creation Date:21/07/2017

Author: Mamadou Diallo

Source code for analysis (separate file): '***Titanic Dataset Notebook Coding.ipynb***'

Dataset selected

In the frame of Udacity Data Analyst Nanodegree this project is conducted using the Titanic Dataset found here (https://d17h27t6h515a5.cloudfront.net/topher/2016/September/57e9a84c_titanic-data/titanic-data.csv)

The kaggle site - cf. [kaggleTitanic] in documentation section - contains the original source of the data and its description.

Questions

Q1. What are the most important factors to survival (e.g. Sex, Class, ...)

Q2. Did they apply the protocol "Women and Children first"?

Q3. Is there any linear correlation among factors?

Q4. How to deal with missing values?

Approach to investigate the questions

- Q1: Use of grouping techniques by categories (e.g. Sex, Embarked, etc.). We will use programming based on document [Hdbk]. We'll use plots with the help of document [Plot].
- Q2: The computation of the survival numbers by category (e.g. Sex, Embarked) and investigation (such as Document [Wikipedia]) if the protocol was enforced.
- Q3: The use of the Pearson's correlation matrix would be usefull. And an example is provided in document [HM]
- Q4: Measure the % of missing values per feature (i.e. variable). Use of articles on the subject such as document [MV].

Data Description

The following Data Dictionary is given from Document [kaggleTitanic]

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	male/female
Age	Age	in years
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

pclass is A proxy for socio-economic status (SES). 1st = Upper, 2nd = Middle, 3rd = Lower

Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

Data Wrangling

To convert and Split nominal features, we use methodology proposed in document [Edx]. Pandas .get_dummies() method allows us to completely replace a single, nominal feature with multiple boolean indicator features.

In addition, when analysing the correlation between features it is preferrable to convert these categorical features into boolean features.

- Sex split into 2 boolean features: 'Sex_female','Sex_male'. Done in SOURCE CODE *§ Change feature representation for Sex*
- Embarked split into 3 variables boolean features: 'Embarked_C', 'Embarked_Q','Embarked_S'. Done in SOURCE CODE *§ Change feature representation for Embarked*
- We create a new Title feature from Name feature as proposed in Document [FE]. Then we split it into boolean features: Title_Mr, Title_Master, Title_Mrs, Title_Miss. Done in SOURCE CODE *§ Change feature representation for Title*
- We create the boolean Child feature (passenger with age under 18) to answer the question relate Women and Children first protocol. Done in SOURCE CODE *§ Build new feature Child*

We don't need to consider the following variables for analysing the relations between features (correlation):

- 'PassengerId': since it is replication (shifted by 1) with the line number or the data frame index.
- 'Cabin': since there are too little data, we'll skip this feature.
- 'Name': Each name is unique
- 'Ticket': as such, the textual feature is no use for analysing relation with other features.

About missing values:

We had 3 features - Cabin, Age and Embarked - concerned by missing values and since the size of the dataset is not large, we did not consider the deletion of entire observations (rows) containing missing values. Dealing completely with missing values is part of the questions (Q4)

Features Study

Number of passengers (i.e. observations) in the data source: 891

On April 15, 1912, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew according to document [Wikipedia].

The data provided in document [kaggleTitanic] shows that it is the training data. It represents 40% (i.e. 891/2224) of the overall population of the Titanic.

In the file, the typical passenger (given by the computation of the mode of the dataframe):

- did not survived
- was male
- had the title Mr
- was 30 year old.
- was in the 3rd class
- paid a fare of 8.05
- embarked at Southampton
- had no sibling and no spouse
- had no parent

Definitions: Continuous Features vs Categorical Features

Given from Document [Edx]:

Throughout the document, we'll use the word '*feature*' instead of the word '*variable*'. They can be used interchangeably.

- **Continuous Features**

In the case of continuous features, there exist a measurable difference between possible feature values. Feature values usually are also a subset of all real numbers:

- **Categorical Features**

With categorical features, there is a specified number of discrete, possible feature values. These values may or may not have ordering to them. If they do have a natural ordering, they are called ordinal categorical features. Otherwise if there is no intrinsic ordering, they are called nominal categorical features.

Example	**Type**
Distance	Continuous
Time	Continuous
Cost	Continuous
Temperature	Continuous
Car Models	Nominal
Colors	Nominal
TV Shows	Nominal
High-Medium-Low	Ordinal
1-10 Years Old, 11-20 Years Old, 30-40 Years Old	Ordinal
Happy, Neutral, Sad	Ordinal

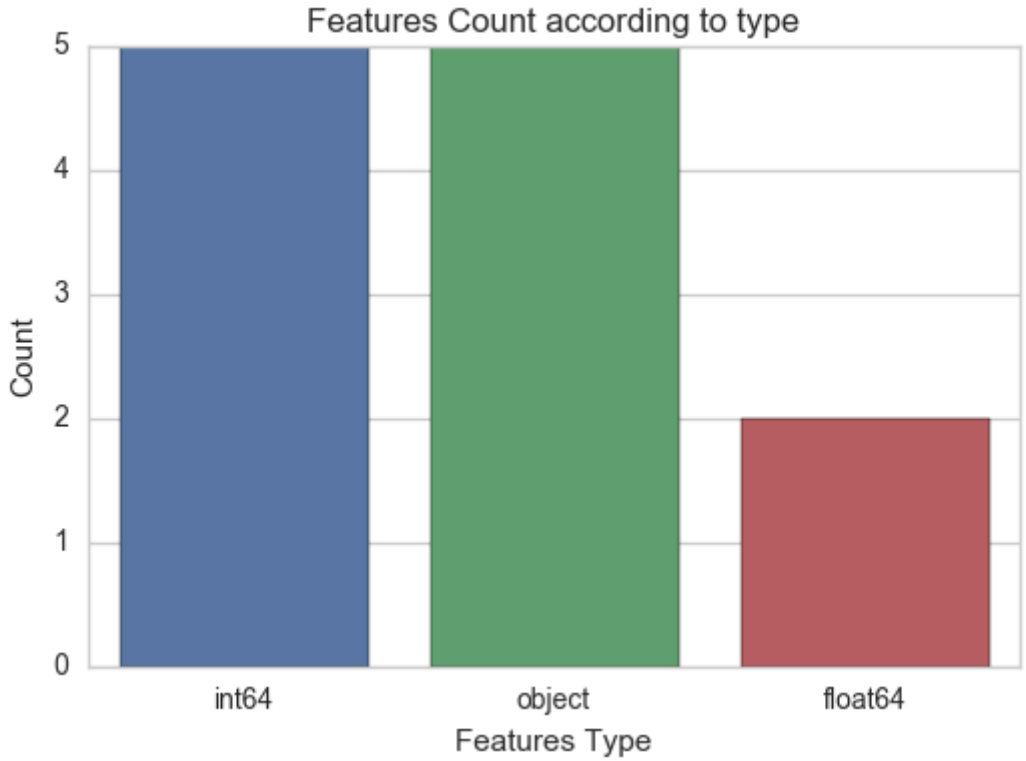
Features Study

We have 12 features - i.e. variables - which are either **continuous or categorical**.

The dependant variable is 'Survived' and the rest are independant variables.

Categorical features could be split into two sub-types:

ordinal feature or nominal.

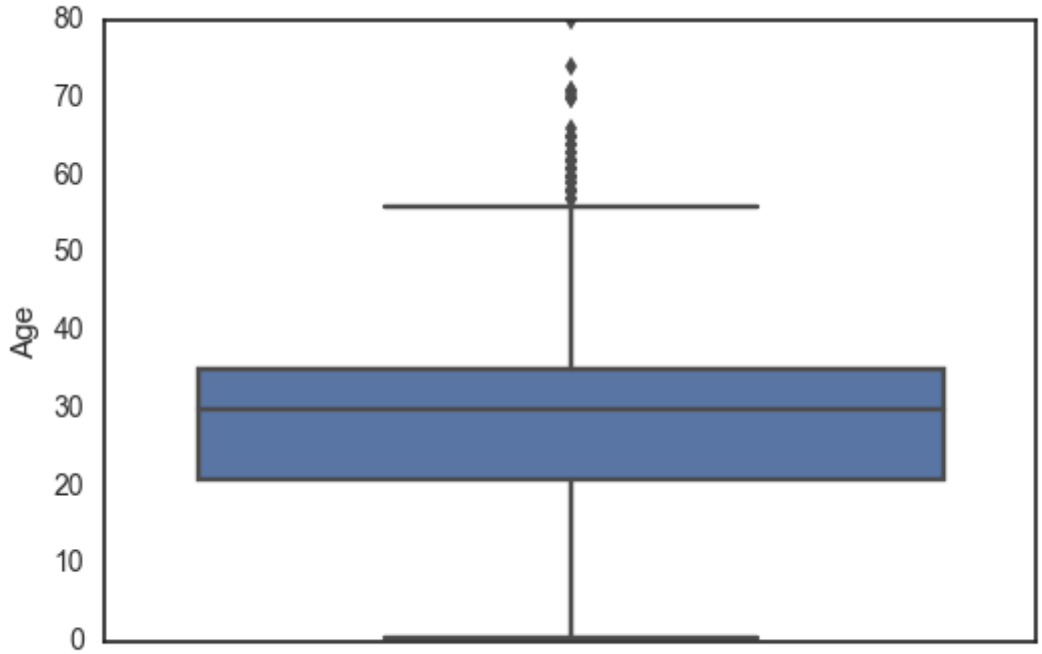

Features Count according to type
Count
5
4
3
2
1
0
int64
object
float64
Features Type

Two Continuous features are: Age and Fare

Age:

Its type is float64.

Variable	Value
mode	30.0
median	30.0
mean	29.377295
std	13.254246
min	0.420000
25%	21.000000
50%	30.000000
75%	35.000000
max	80.000000

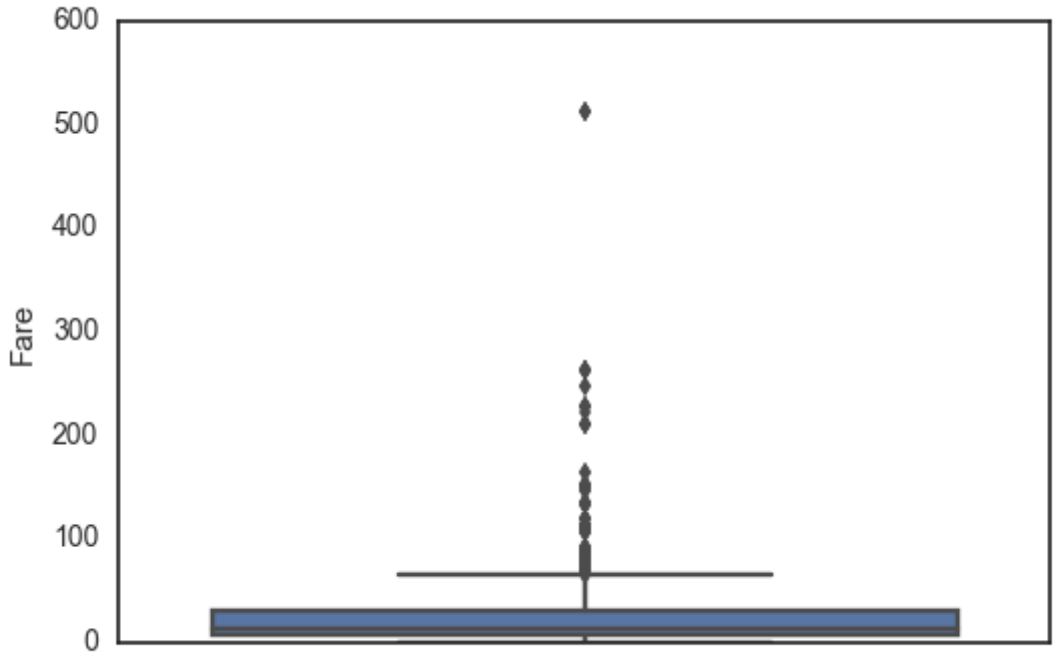


Fare:

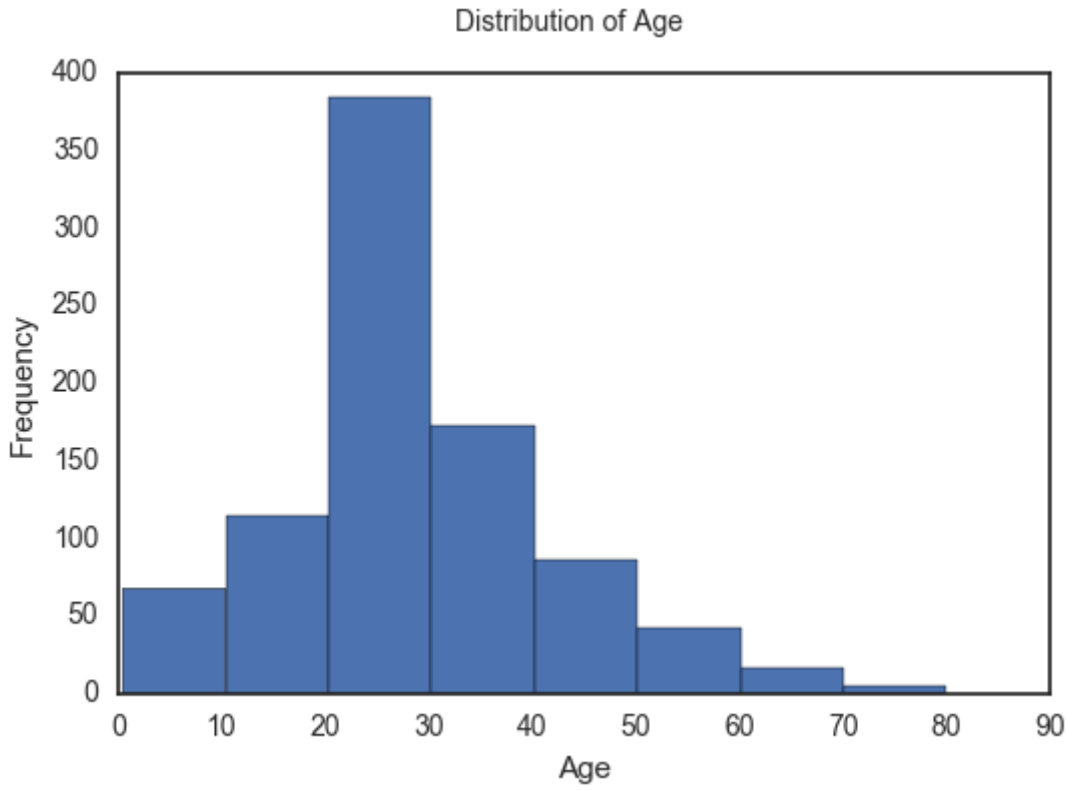
Its type is float64.

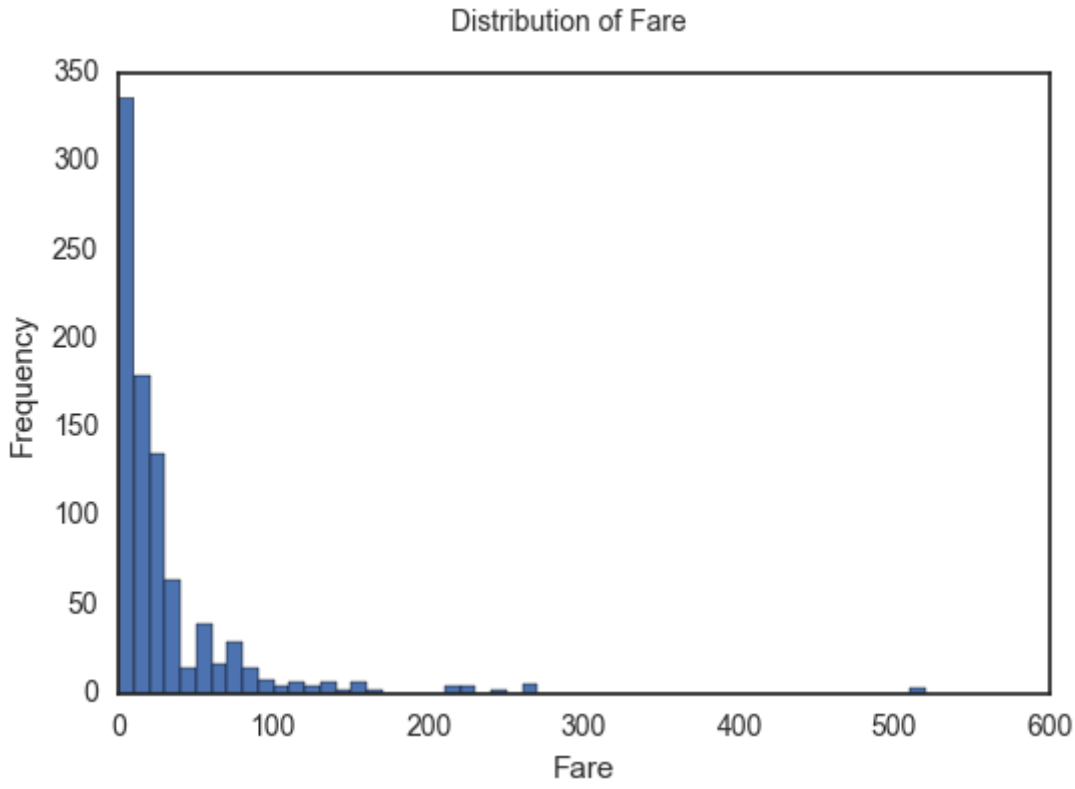
Variable	Value
mode	8.05
mean	32.2042079686
std	49.6934285972

Variable	Value
min	0.0
25%	7.910400
50%	14.454200
75%	31.000000
max	512.3292



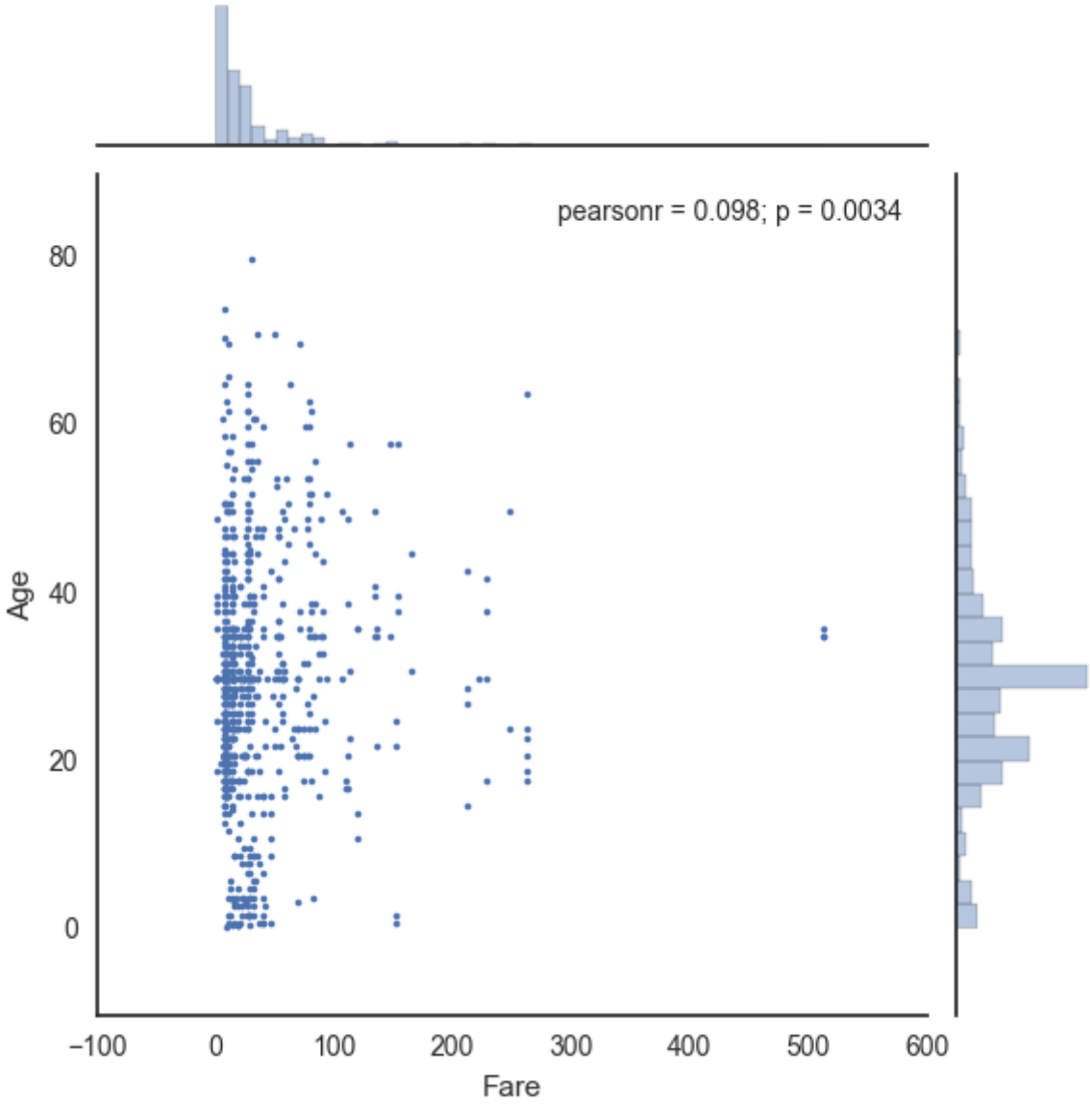
Here below are their distribution:





The Fare feature distribution is left-skewed (see Fare distribution here above)

Relation Between *Age* and *Fare* features



There is no relation between the two continuous features (Fare and Age) as shown here above

Categorical features are:

'*Pclass*', Ordinal feature (1st class > 2nd Class > 3rd Class in terms of confort)

'*SibSp*', Ordinal feature.

'*Parch*', Ordinal feature.

'*Survived*', Nominal feature. Two possible values (0 = No, 1 = Yes)

'*PassengerId*', Nominal feature (no rank in numbers) as identifier.

'*Name*', Nominal feature (Textual feature).

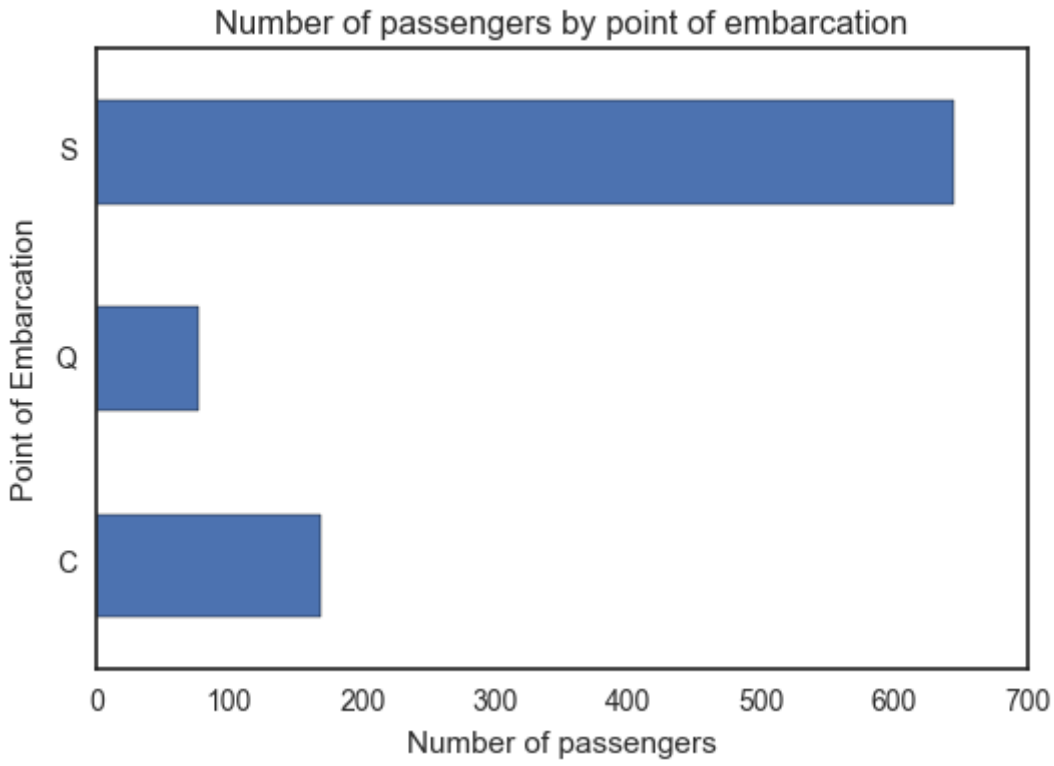
'*Sex*', Nominal feature. Two possible values: male or female.

'*Ticket*', Nominal feature (Textual feature).

'*Cabin*', Nominal feature.

'*Embarked*', Nominal feature. Only 3 points of Embarcation.

Distribution of Embarked feature:



KEYS: C = Cherbourg, Q = Queenstown, S = Southampton

About Duplicated data:

The code shows that the following categorical features have duplicates:

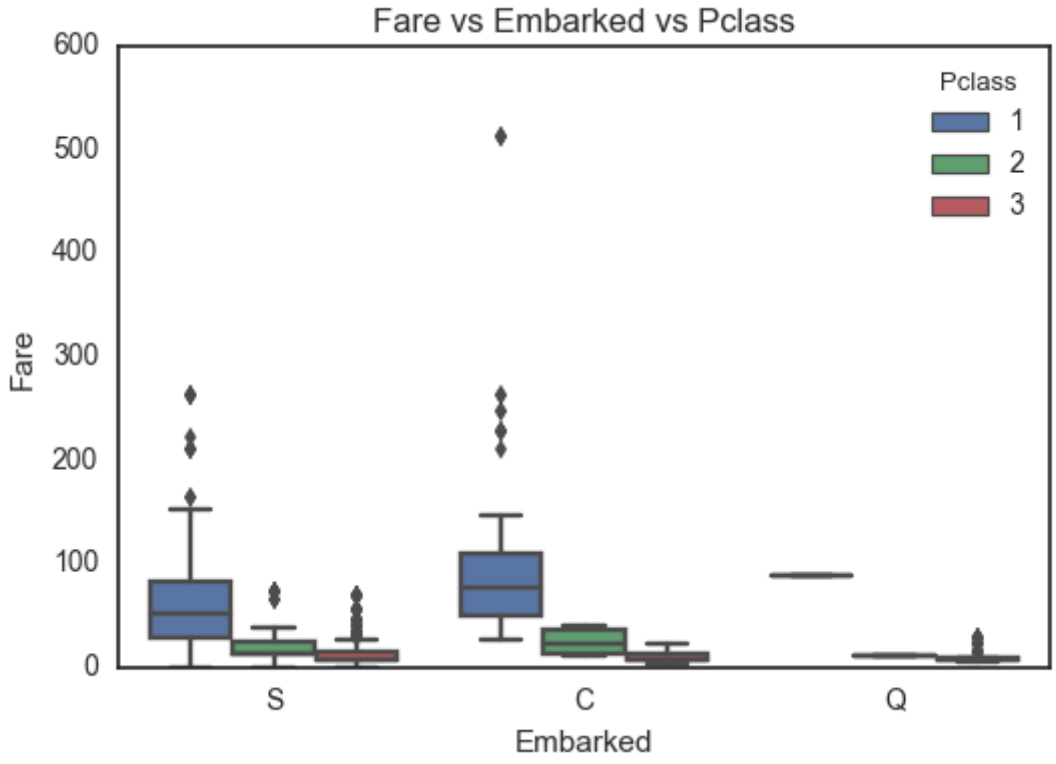
- **Cabin**: meaning some passengers shared the same cabin. Moreover, some passengers have many cabins
- **Ticket**: meaning some passengers were registered on the same ticket.

The following categorical features don't have duplicates:

- **Name**: meaning this feature could be considered as unique and as an identifier
- **PassengerId**: this feature is redundant with the index

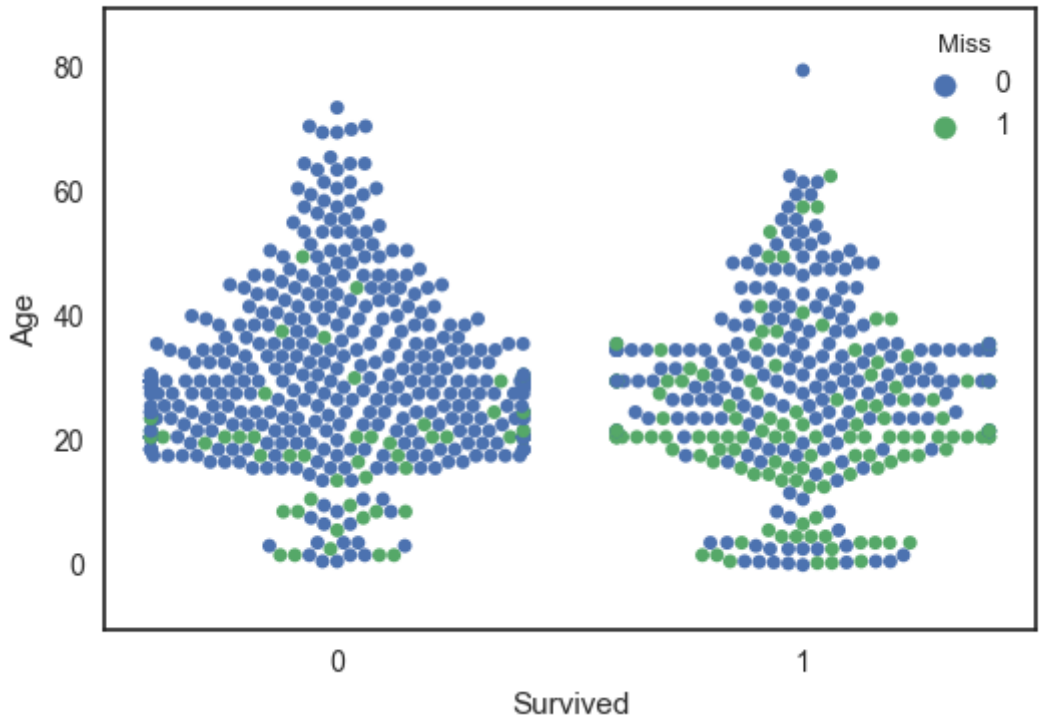
Relation between features

Relation Fare vs Embarked vs Pclass



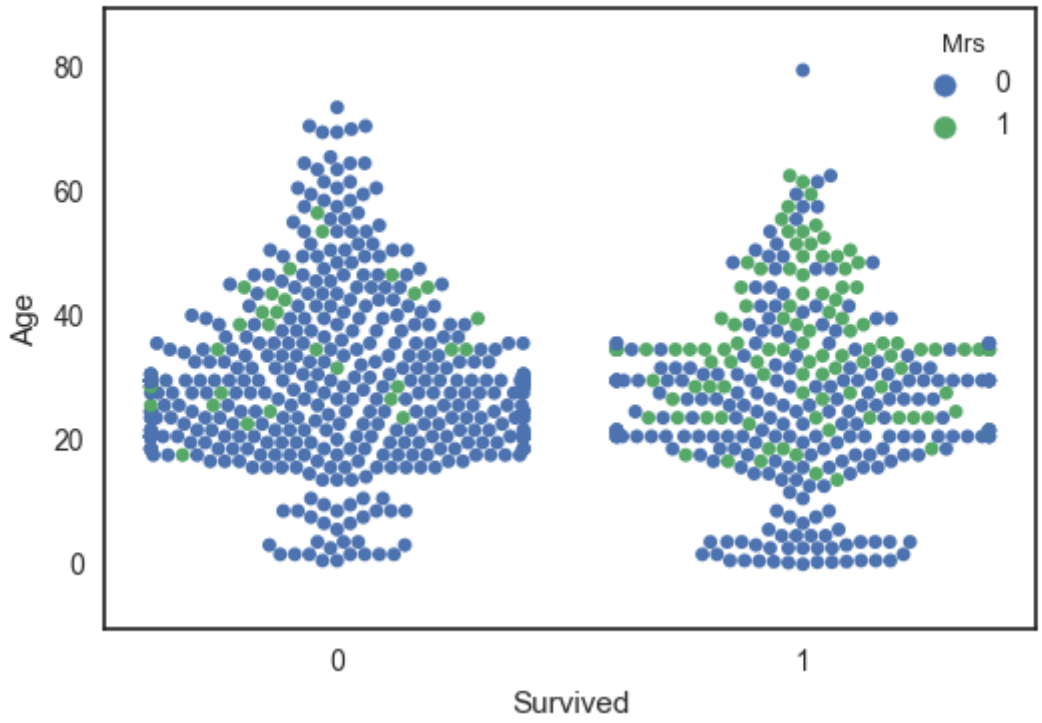
Relation Survived vs Age vs Title_Miss: It gives clues on:

- Age distribution according to Survival and
- relation between Title Miss, Survival and Age features

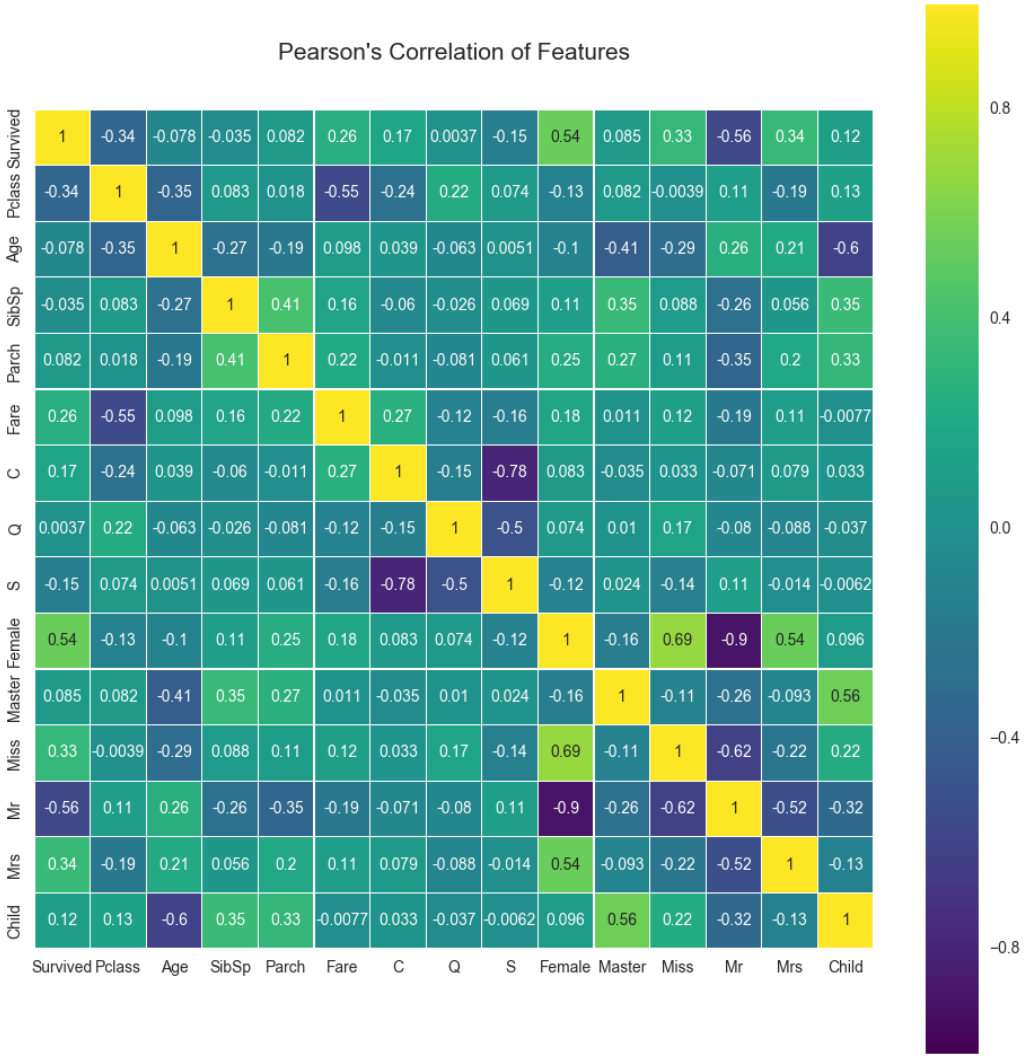


Relation Survived vs Age vs Title_Mrs: It gives clues on:

- Age distribution according to Survival and
- relation between Title Mrs, Survival and Age features



Reuse of document [HM] for building the Correlation Diagram.



Linear Correlation deduced from Pearson's Correlation Diagram:

Survived and Sex (female) have a linear relationship (increasing).

Survived and Title (Mr) have a linear relationship (decreasing). It is not a surprise since Title (Mr) and Sex (female) are strongly correlated.

Pclass and Fare have a linear relationship (decreasing).

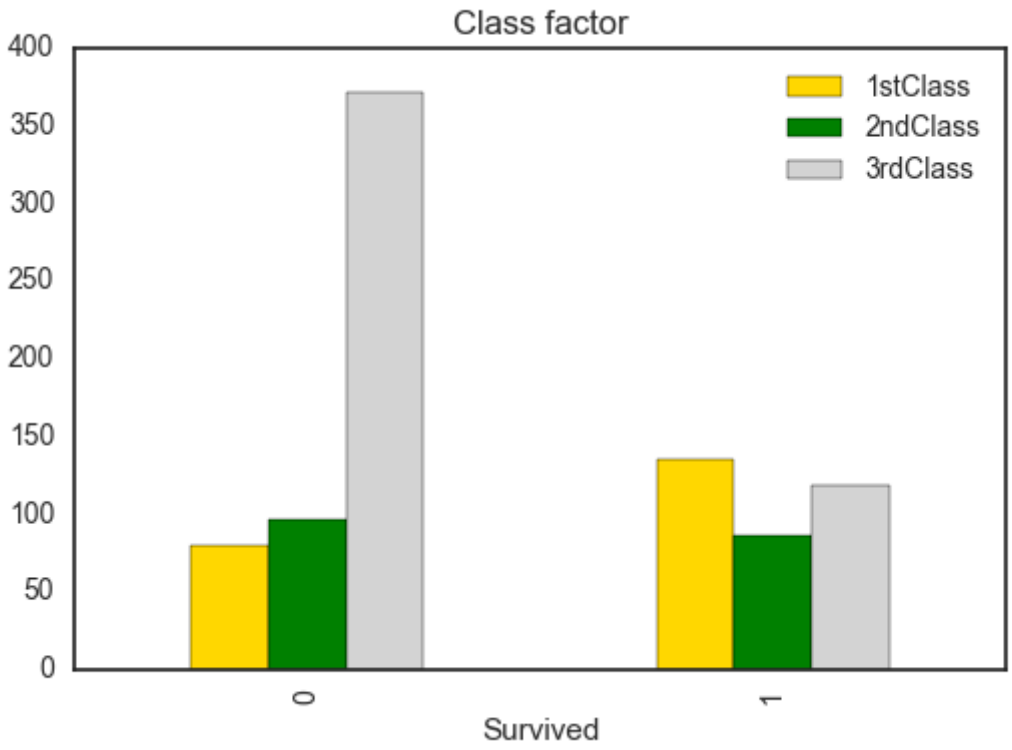
Master and Age have a linear relationship (decreasing).

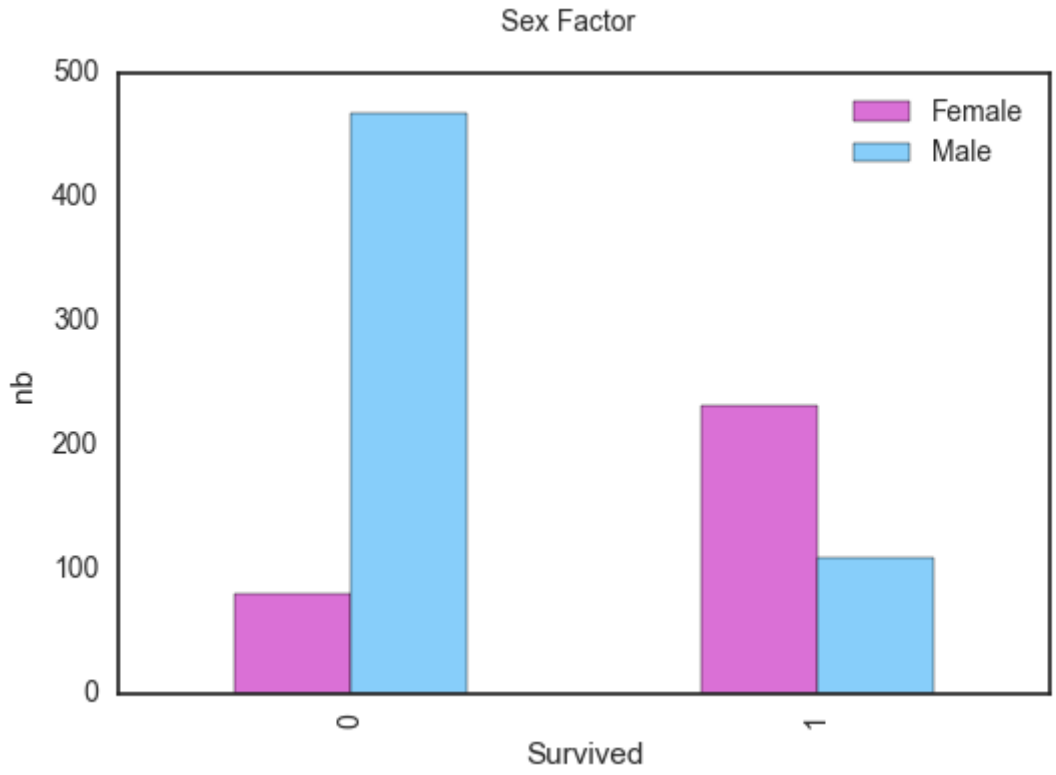
Point of Embarcation C and S have a strong linear relationship. It is normal since Point of Embarcation C, S, Q are built from Embarked.

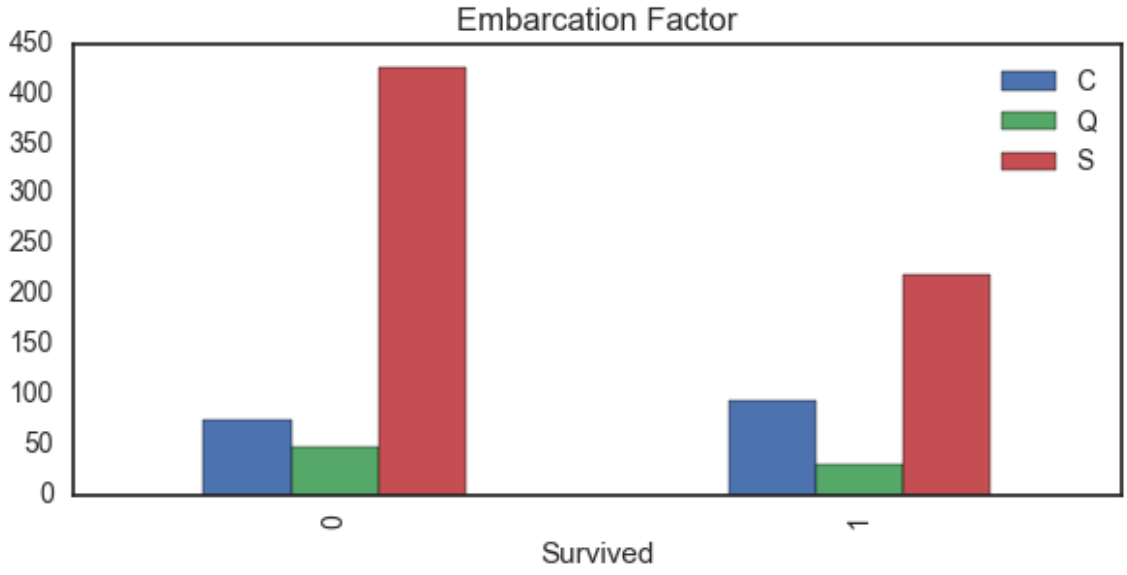
Age and Child have a significant linear relationship. It is normal since child was built from Age.

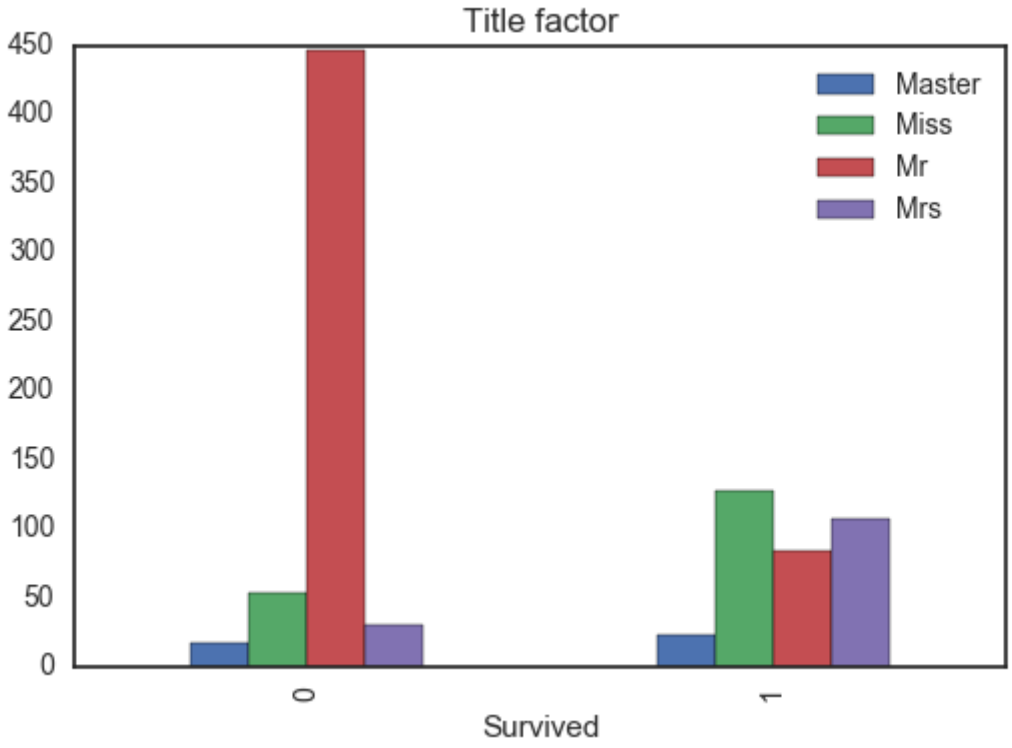
SibSp and Parch have a linear relationship.

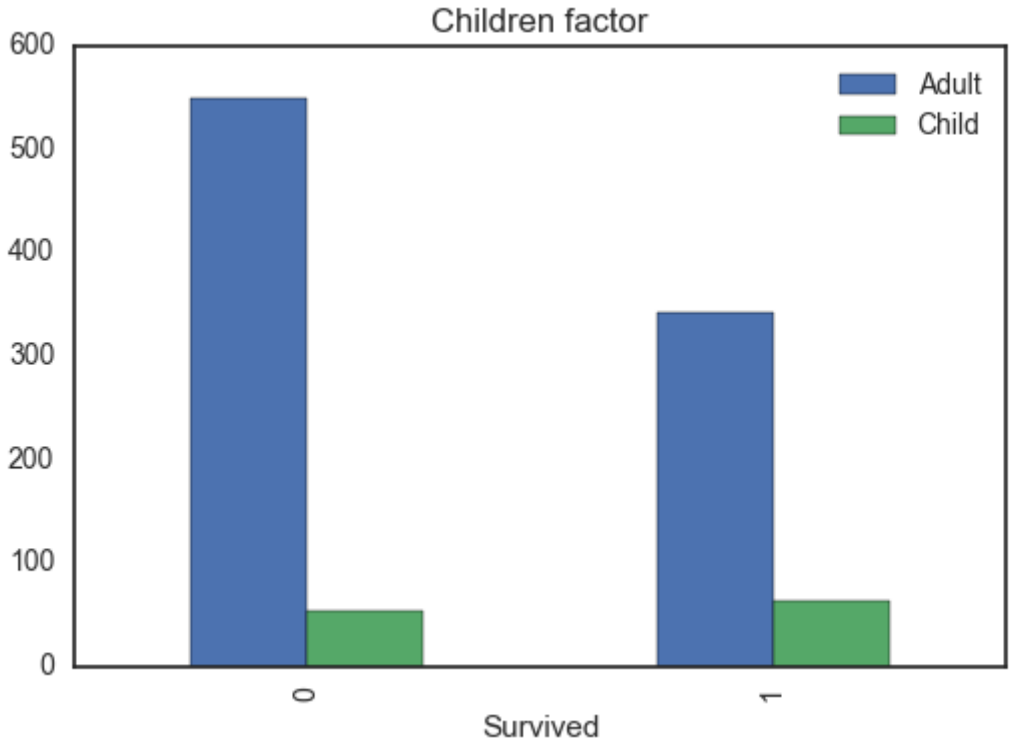
Importance of factors on survival




Sex Factor
Female
Male
nb
500
400
300
200
100
0
0
1
Survived


Embarcation Factor
C
Q
S
450
400
350
300
250
200
150
100
50
0
0
1
Survived





Answer to questions

- **Q1. What are the most important factors to survival (e.g. Sex, Class, ...)?**

Title: Significantly More survival chance for Mrs, Miss and slight more chance for Master. Significantly less survival chance for Mr.

Sex: Significantly more survival chance for Sex is female. Significantly less survival chance for sex is male.

Class: The first class has higher survival chance than other classes. The third class has less chance than the two others.

Embarked: Point of Embarcation influences the survival chance and Cherbourg is favored.

Children: Being a child (under 18) gave slightly more survival chance.

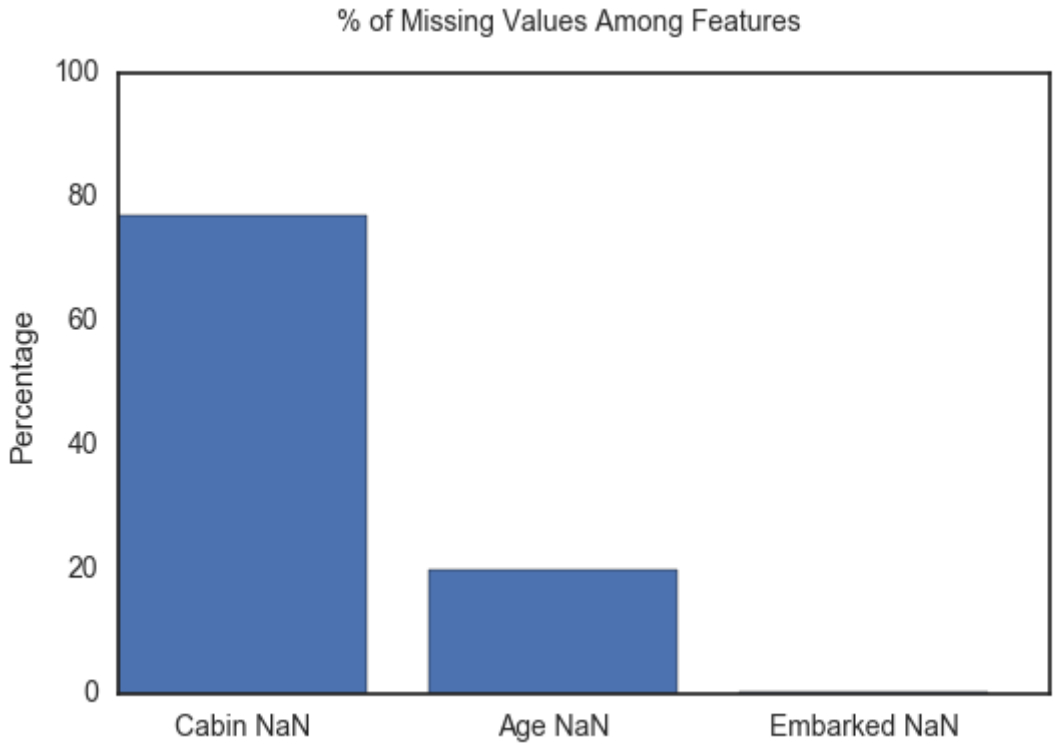
- **Q2. Did they apply the protocol "Women and Children first"?**

In document [Wikipedia] "A disproportionate number of men were left aboard because of a "women and children first" protocol for loading lifeboats". It has been applied and the figures reflect it for Women. **The figures confirm strongly for women and slightly for children.**

- **Q3. Is there any linear correlation among factors?**

To establish it is it practical to use numerical values. See here above in Section *'Relation between features'* The linear correlation among features has been established numerically and in figure 'Pearson Correlation of Features.png' provided here above

- **Q4. How to deal with missing values?**



High percentage of missing values for Cabin feature. **Nothing can be done for Cabin feature.**

Significant percentage of missing values for Age feature but something can be done for this feature. **Replace missing values for Age with the median value of the group split by title.** See Coding in *§ Processing missing values for Age*

Title	Age (median)
Master	3.5
Miss	21.0
Mr	30.0
Mrs	35.0

There is a very small percentage (0.2 %) of missing values for Embarked feature. **For the Embarked features, replacement of the actual values is possible (i.e. Southampton). They have been found in document [DB] from the name of the passengers.**

Way forward:

It would be useful but not in the scope of the assignment:

- MAKE PREDICTION: to apply machine learning (e.g. Logistic Regression) to build a model and make prediction.
- GET MORE DATA: to get more data and split it into training and test data. The information about crew passengers might give more insight.

Documentation/References

Here is the list of references - including Web sites, books, blog posts - used for my submission.

[Edx] Web site: DAT210x Programming with Python for Data Science: EdxSite (https://courses.edx.org/courses/course-v1:Microsoft+DAT210x+2T2017/info)

[FE] Feature Engineering: site (https://triangleinequality.wordpress.com/2013/09/08/basic-feature-engineering-with-the-titanic-data/5)

[DB] Encyclopedia Titanica database: site (https://www.encyclopedia-titanica.org/)

[Wikipedia] Wikipedia article - See §Survivors and victims: wikipedia (https://en.wikipedia.org/wiki/RMS_Titanic)

[kaggleTitanic] kaggle web site: kaggleTitanic (https://www.kaggle.com/c/titanic/data)

[Plot] Matplotlib resources from Blog: blog (http://www.datasciencecentral.com/profiles/blogs/matplotlib-cheat-sheet)

[Hdbk] Python for Data Science Handbook from Blog: blog (http://www.datasciencecentral.com/profiles/blogs/book-python-data-science-handbook?utm_content=buffer09a5c&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer)

[HM] Heatmap example: kaggle (https://www.kaggle.com/arthurtok/introduction-to-ensembling-stacking-in-python)

[MV] How to treat missing values in you data from Article: Article (https://clevertap.com/blog/how-to-treat-missing-values-in-your-data-part-i/)