***Titanic***

Kaggle

Predict survival of passengers in test set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Variables |  |  |  |
| 1. | survival | Survival | 0 = No, 1 = Yes | Catergorical |
| 2. | pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd | Integer/Float |
| 3. | sex | Sex/Gender | female,male | String |
| 4. | Age | Age in years |  | Integer/Float |
| 5. | sibsp | # of siblings / spouses aboard |  | Integer/Float |
| 6. | parch | # of parents / children aboard |  | Integer/Float |
| 7. | ticket | Ticket number |  | String |
| 8. | fare | Passenger fare |  | Integer/Float |
| 9. | cabin | Cabin number |  | String |
| 10. | embarked | Port of Embarkation | C = Cherbourg,  Q = Queenstown,  S = Southampton | String |

**Pclass**

A proxy for socio-economic status (SES)

1st = Upper, 2nd = Middle, 3rd = Lower

**Age**

Age is fractional if less than 1.

If the age is estimated, is it in the form of xx.5

**Sibsp**

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

**Parch**

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them.

**Univariate analysis**

**Sex/Gender**

* Categorical variable
* 577 male / 314 female
* Replace with 0 if female, 1 if male

A close up of a map

Description generated with high confidence**Age**

* Continuous variable
* Has missing variables
* Fill with median (as data is right-skewed)

|  |  |
| --- | --- |
| count | 714.000000 |
| mean | 29.699118 |
| median | 28.0 |
| std | 14.526497 |
| min | 0.420000 |
| 25% | 20.125000 |
| 50% | 28.000000 |
| 75% | 38.000000 |
| max | 80.000000 |

**SibSp**