**Report On Titanic Dataset**

**Introduction**

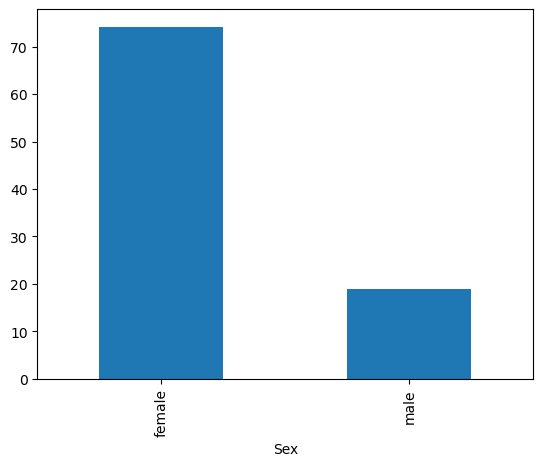
The Titanic dataset is made up of Four hundred and eighteen (1309) rows with Thirteen (13) columns. This simply means that the ship had 1309 passengers that have been described using 13 columns. The columns are: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked. Some of the columns had missing data, which were Survived (with 418 missing values), Age (with 263 missing values), Fare (1 missing value), Cabin (with 1014 missing values) and Embarked (with 2 missing values). The passengerId and Tickets was later dropped in the course of the analysis. The “Name” column had eighteen (18) titles attached to names. The titles are: Capt, Col, Countess, Don, Dona, Dr., Jonkheer, Lady, Major, Master, Mr, Miss, Mrs, Mme, Mlle, Rev., Ms, Sir. Each of these titles had 757,260,197, 61, 8, 8, 4, 2, 2, 2,1,1,1,1,1,1,1,1 frequency respectively. The Passengers’ Class, ‘Pclass’ has three classes namely: 1,2 and 3. The 1309 passengers were made up of 843 males and 466 females who boarded using 3 ports (Embarked) namely: S, C and Q, 914, 270 and 123 passengers respectively. A summary of the columns that were either float or integers are shown on Table 1 (all the data presented are corrected to two decimal places). For the ‘Survived’ column, the minimum value was 0 while the maximum value was 1. The oldest person on the ship was 76.00 years old. The proportion of passengers with family members were 39.6486% and people that constituted family members were siblings and spouses (SibSp), and parents with children (Parch)

Table 1. Summary of Columns with floats and Integer data types data

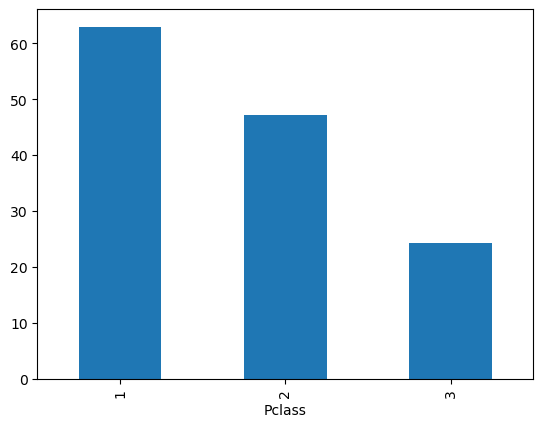
|  | **Unnamed: 0** | **PassengerId** | **Survived** | **Pclass** | **Age** | **SibSp** | **Parch** | **Fare** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1309.000000 | 1309.000000 | 891.000000 | 1309.000000 | 1046.000000 | 1309.000000 | 1309.000000 | 1308.000000 |
| **mean** | 369.478992 | 655.000000 | 0.383838 | 2.294882 | 29.881138 | 0.498854 | 0.385027 | 33.295479 |
| **std** | 248.767105 | 378.020061 | 0.486592 | 0.837836 | 14.413493 | 1.041658 | 0.865560 | 51.758668 |
| **min** | 0.000000 | 1.000000 | 0.000000 | 1.000000 | 0.170000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 163.000000 | 328.000000 | 0.000000 | 2.000000 | 21.000000 | 0.000000 | 0.000000 | 7.895800 |
| **50%** | 327.000000 | 655.000000 | 0.000000 | 3.000000 | 28.000000 | 0.000000 | 0.000000 | 14.454200 |
| **75%** | 563.000000 | 982.000000 | 1.000000 | 3.000000 | 39.000000 | 1.000000 | 0.000000 | 31.275000 |
| **max** | 890.000000 | 1309.000000 | 1.000000 | 3.000000 | 80.000000 | 8.000000 | 9.000000 | 512.329200 |

**The Survival Rate**

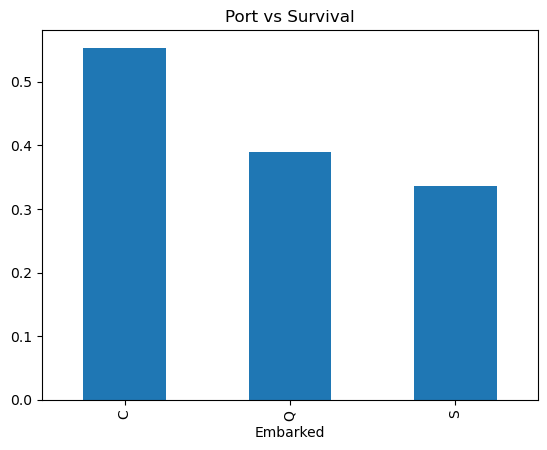
342 passengers survived which was about 38.38% of the entire people that boarded the ship while 549 people died which also was about 61.62%. Analysis shows that 18.890815 male survived. Also 74.203822 female survived.



62.962963% of the people that survived were in the ‘Pclass 1’, first class, 47.282609% in 2nd class and 24.236253 in the third class. Of all the people that survived, 50.56 % were with family member(s) while 30.35 % that survived were without family.



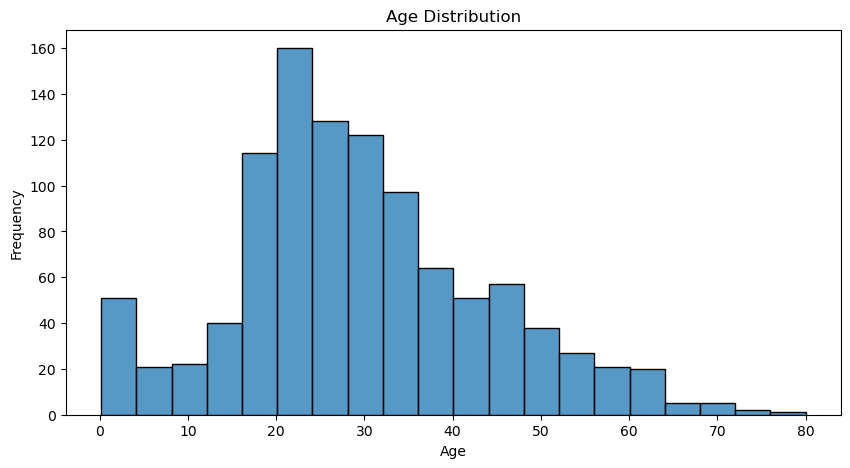
There were 3 three ports for the boarding of passengers into the ship namely C, Q and S. the survival rate by port were 0.553571, 0.389610 and 0.336957 respectively.



**Age Distribution**

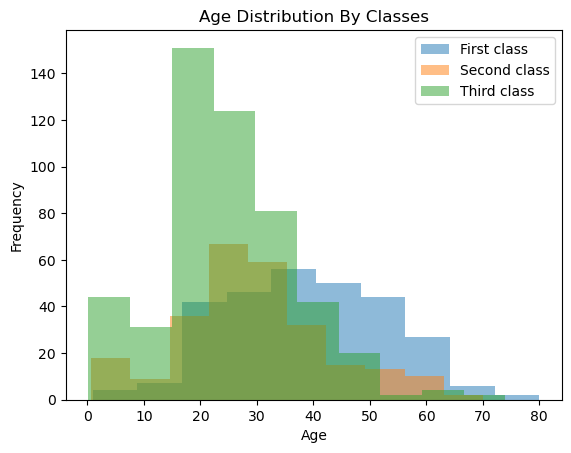
The ages of people that boarded the ship spanned from zero years to about 80

Ages with highest frequency was between 20 and 25 years with the lowest frequency at about Age 82.

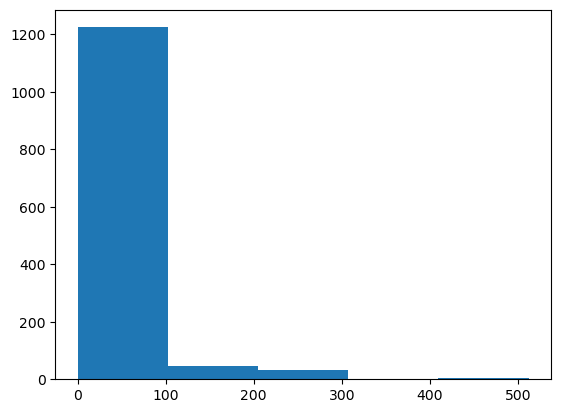


**Age distribution by passengers’ class.**

First class spanned from about 0 yrs to 80 years, second class spanned from a little over age 0yrs to about 64 years old and third-class spans from zero to about 75 years old. For classes 1 and 2, they have the lowest frequency while class 3 had ages 15-25 with the highest frequency seconded by ages 25-30.



The fare spanned from zero to 500 with no fare between 310 and 410. The fare paid with the lowest frequency were within the range of 412 and 512. Fare with the highest frequency was between zero and 100.



**Correlation**

The following were correlations with survived: Fare with 0.26, family with 0.017 and Parch with 0.082

Correlations with PClass were age(-0.41), SibSp (0.061), Parch(0.019), Fare (-0.56), family (0.05)

Correlations with Age were Family (-0.24), Fare (0.18), Parch (-0.15), SibSp (-0.24) Pclass (-0.41), Survived (-0.077)

The highest correlation was Family(1) which constituted SibSp and Parch. SibSp was(0.86) followed by parents with children (Parch 0.79). This is because they are the units that constituted the family.



**Conclusion**

The dataset was made up of 1309 rows and 10 columns which were directly 1309 passengers that boarded the ship with 13 attributes to describe them. The attributes were, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Fare, Cabin, Embarked and then another column Family was added which constituted the SibSp and Parch. The type of data on the dataset were integers, float and strings. Total number of male that boarded the ship was 843 and female was 466. 74.203822 of the survivors were female and 18.890815 of the survivors were male. Three hundred and forty two people survived which was about 38.4 % of the entire people that boarded the ship while 549 people died which was about 61.6 %. People in the third class had the highest rate of survival. The oldest person on the ship was 76.00 years old. The proportion of passengers with family members were 50.56% and people that constituted family members were siblings and spouses (SibSp), and parents with children(Parch). The fare spanned from zero to 512 with no fare between 310 and 410. The fare paid with the lowest frequency were within the range of 412 and 512. Fare with the highest frequency was between zero and 100. Correlation between the attributes showed that the factors that had the highest correlation (0.83)was between parents with children (Parch) with family followed by SibSp with family. This is because they are the units that constituted the family.