# Case Study: Analyze data to predict who will Survive the Titanic

Contents

[Case Study: Analyze data to predict who will Survive the Titanic 1](#_Toc71468632)

[1. Load the data from the “train.csv” file into a Dataframe. 2](#_Toc71468633)

[2. Display the dimensions of the file (so you’ll have a good idea the amount of data you are working with. 3](#_Toc71468634)

[3. Display the first 5 rows of data so you can see the column headings and the type of data for each column. 4](#_Toc71468635)

[4. Think about some questions that might help you predict who will survive: 5](#_Toc71468636)

[a. What do the variables look like? For example, are they numerical or categorical data? If they are numerical, what are their distribution; if they are categorical, how many are they in different categories? 5](#_Toc71468637)

[5. Look at summary information about your data (total, mean, min, max, freq, unique, etc.) Does this present any more questions for you? Does it lead you to a conclusion yet? 9](#_Toc71468638)

[6. Make some histograms of your data (“A picture is worth a thousand words!”) 10](#_Toc71468639)

[a. Most of the passengers are around 20 to 30 years old and don't have siblings or relatives with them. A large amount of the tickets sold were less than $50. There are very few tickets sold where the fare was over $500. 10](#_Toc71468640)

[7. Make some bar charts for variables with only a few options. 12](#_Toc71468641)

[Ticket and Cabin have more than 100 variables so don’t do those! 12](#_Toc71468642)

[8. To see if the data is correlated, make some Pearson Ranking charts 13](#_Toc71468643)

[9. Use Parallel Coordinates visualization to compare the distributions of numerical variables between passengers that survived and those that did not survive. 15](#_Toc71468644)

[10. Use Stack Bar Charts to compare passengers who survived to passengers who didn’t survive based on the other variables. 16](#_Toc71468645)

[11. Some of my questions have been answered by seeing the charts but in some ways, looking at this much data has created even more questions. 18](#_Toc71468646)

[a. Now it’s time to reduce some of the features so we can concentrate on the things that matter! There features we will get rid of are: "PassengerId", "Name", "Ticket" and "Cabin". (ID doesn’t really give us any useful data, Ticket and Cabin have too many variables. Name might reflect that they are related but we’re keeping the category about siblings (for now). 18](#_Toc71468647)

[b. Age has some missing values, so I’ll fill in with the average age. Embarked also has some missing so I’ll the most common. 19](#_Toc71468648)

[12. If you go back and look at the histograms of Fare, you’ll see that it is very skewed…many low cost fares, not very many high cost fares. Log Transformation is a good method to use on highly skewed data. 20](#_Toc71468649)

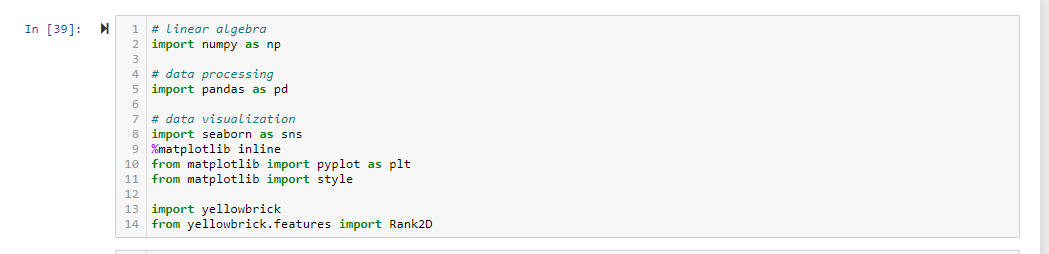
[13. Convert your categorical data into numbers (Sex, PClass, Embark) 21](#_Toc71468650)

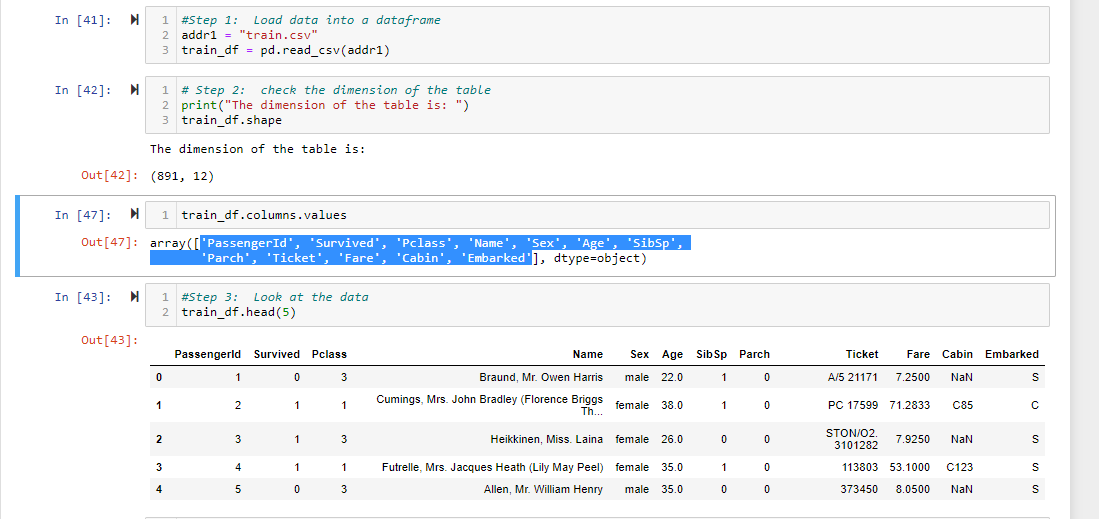
[14. Training - Split your data into two sets: Training and Testing. 22](#_Toc71468651)

[15. Evaluation – Remember, we are trying to predict if a passenger has survived or not so this is a classification problem. There are many algorithms that could be used but we’re going to use logistic regression. 23](#_Toc71468652)

## Load the data from the “train.csv” file into a Dataframe.

* Numpy for Linear Algebra and numeric
* Pandas for Data Processing
* Matplotlib for data visualization





Load the data from .csv file into train\_df Dataframe.

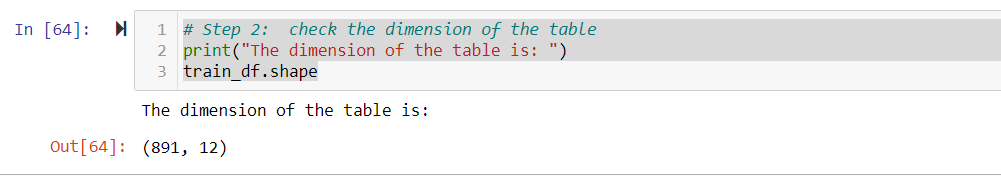
12 variables as below

'PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',

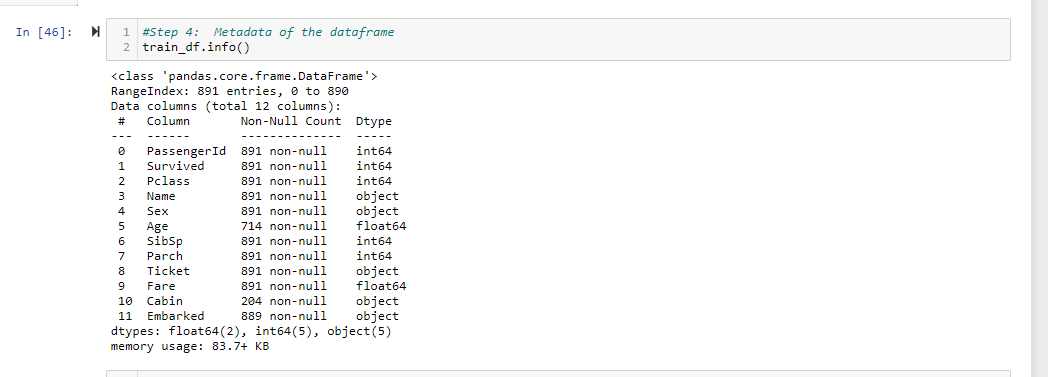
'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'

## Display the dimensions of the file (so you’ll have a good idea the amount of data you are working with.

Train\_df has 12 variables and 891 rows.



Verify the metadata of the Dataframe.



## Display the first 5 rows of data so you can see the column headings and the type of data for each column.

* 1. Notice that Survived is represented as a 1 or 0

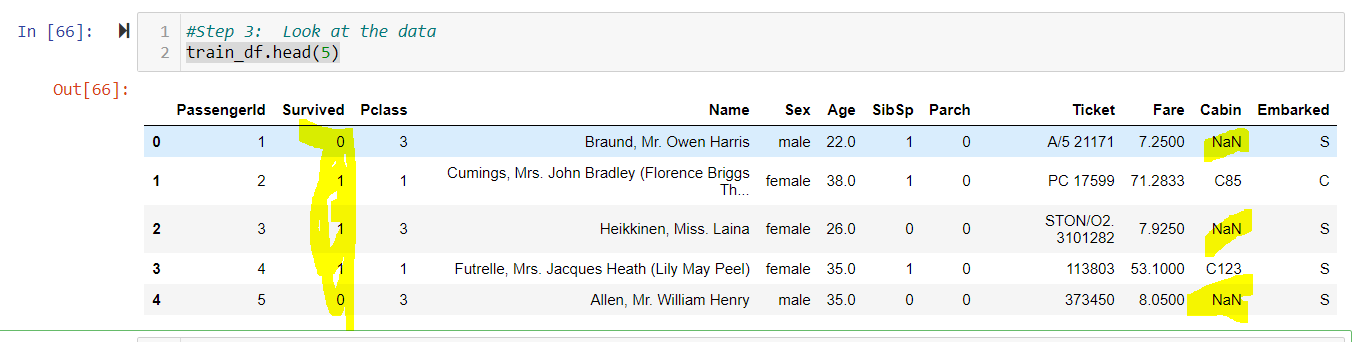
Yes, survived represents 1 and 0. I assume 1 for Survived and 0 for Not Survived

* 1. Notice that missing data is represented as “NaN”

Yes, Cabin variable contains NaN values.

* 1. The Survived variable will be the “target” and the other variables will be the “features”

Yes, we can use Survived will be target and we can verify the other variables how features.



## Think about some questions that might help you predict who will survive:

## What do the variables look like? For example, are they numerical or categorical data? If they are numerical, what are their distribution; if they are categorical, how many are they in different categories?

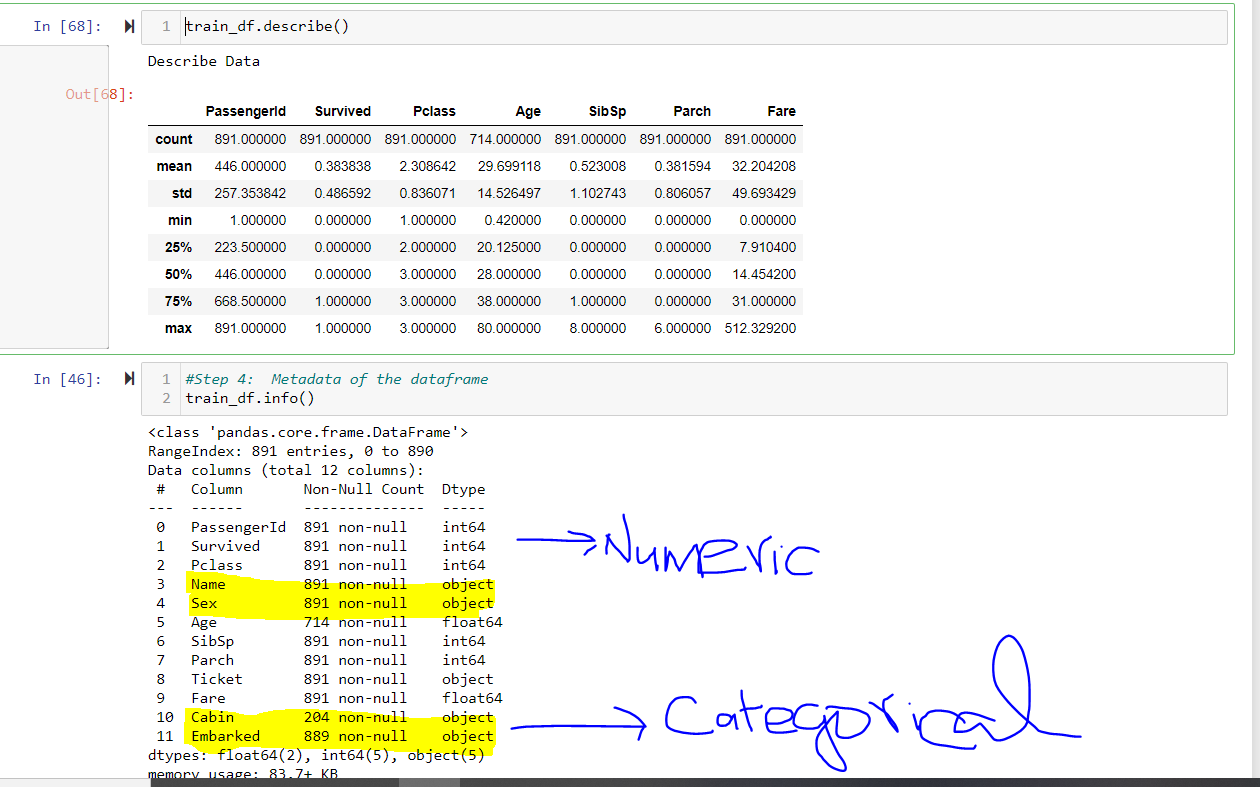
* Numerical –

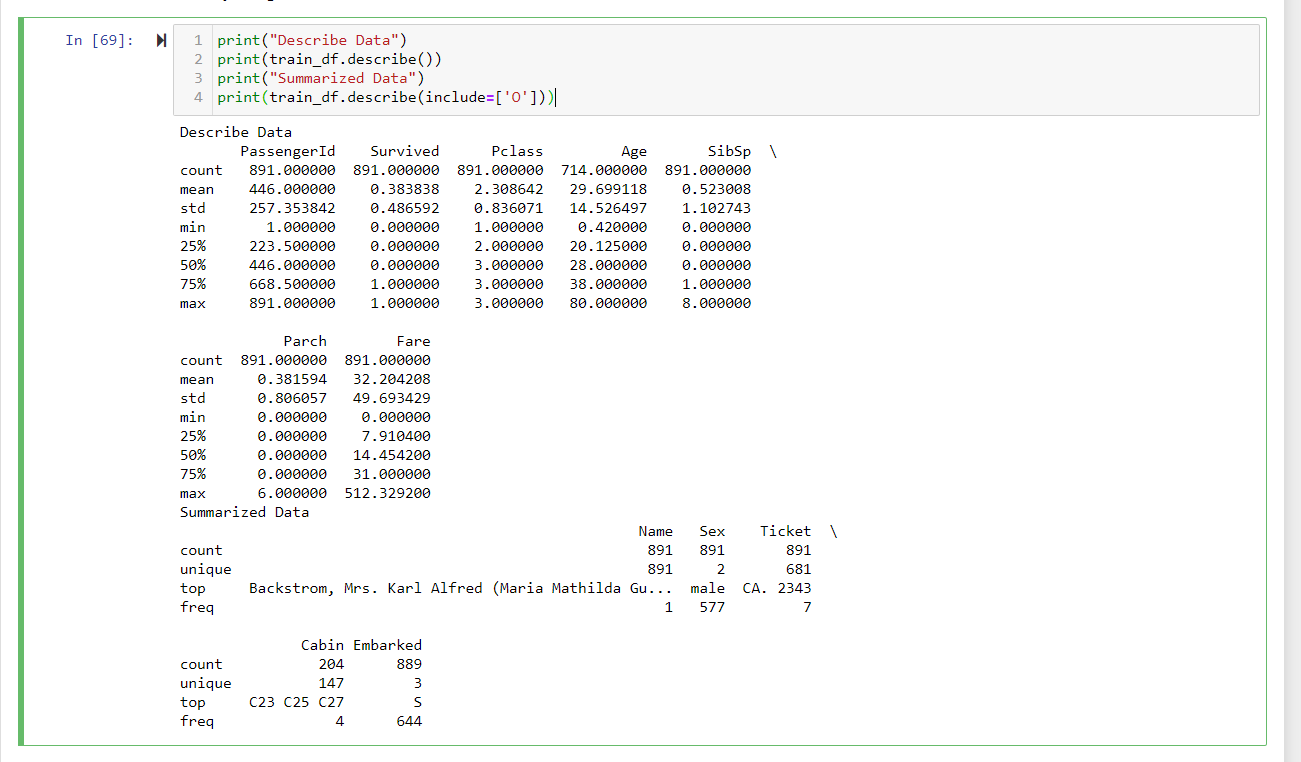
Continuous - Age , Fare (Range) and discrete - SibSp, Parch ( whole number )

* Categorical –

Nominal - Name, Sex, Ticket, Cabin, Embarked, Survived (unordered category)

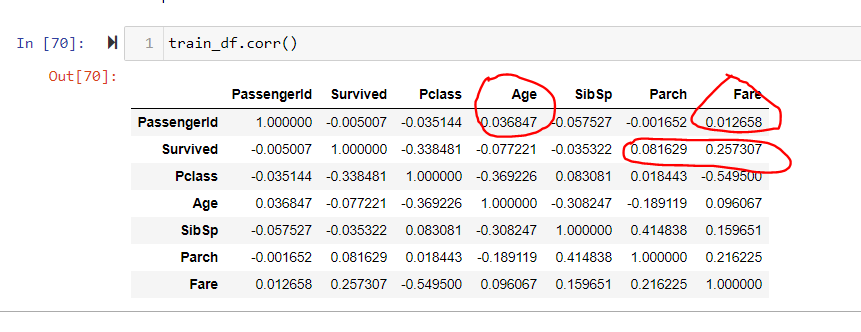
Ordinal – Pclass (ordered category)





1. Are the numerical variables correlated?

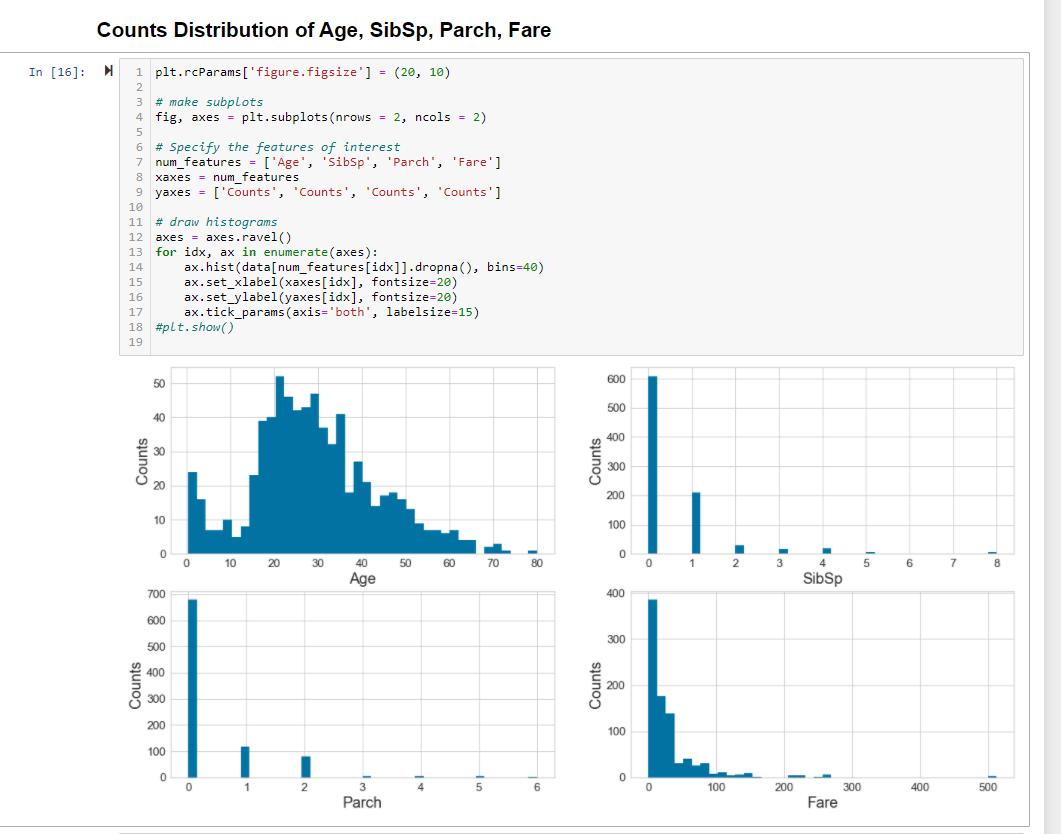
Yes, Its correlated. Example, Passengerid and Age, Fares are positive Correlated.



1. Are the distributions of numerical variables the same or different among survived and not survived? Is the survival rate different for different values? For example, were people more likely to survive if they were younger?

The Distribution of numeric variables different among survived and Not Survived.

The Survival rate different for different values. Females are survived more when compare to Males.



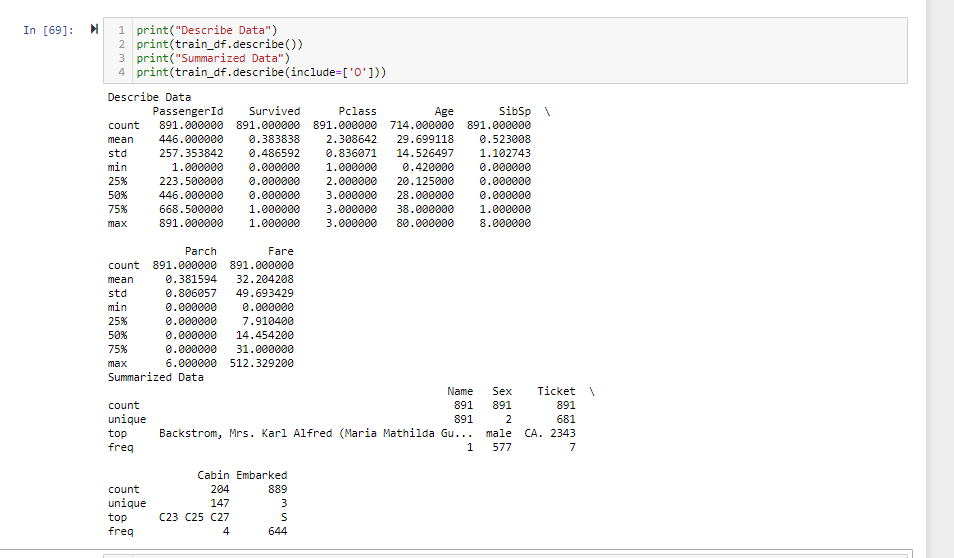
* 1. Are there different survival rates in different categories? For example, did more women survived than man?

Yes, Different Survival rates in different categories. It clearly shows that Female passengers in age of 20-40 Age Group are survived more when compare to Male passengers.



## Look at summary information about your data (total, mean, min, max, freq, unique, etc.) Does this present any more questions for you? Does it lead you to a conclusion yet?

This information clearly explains about how the data presents in each field. But with this, I could not conclude anything without seeing the relation between each variables.



## Make some histograms of your data (“A picture is worth a thousand words!”)

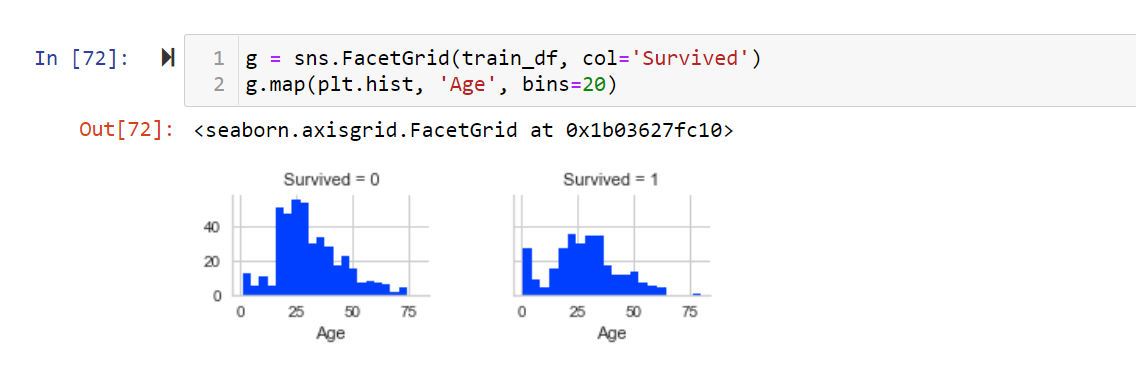
## Most of the passengers are around 20 to 30 years old and don't have siblings or relatives with them. A large amount of the tickets sold were less than $50. There are very few tickets sold where the fare was over $500.

I used both Matplot and sns library to verify the histogram of the data.

1. Yes, most of the passengers are around 20-30 and they do not have siblings.
2. Large number of tickets sold less than 50 (Around 380 counts)
3. Very less in more than 500 ( 3)



Interesting to know that infants are survived well (next of 20-30 age Group ).



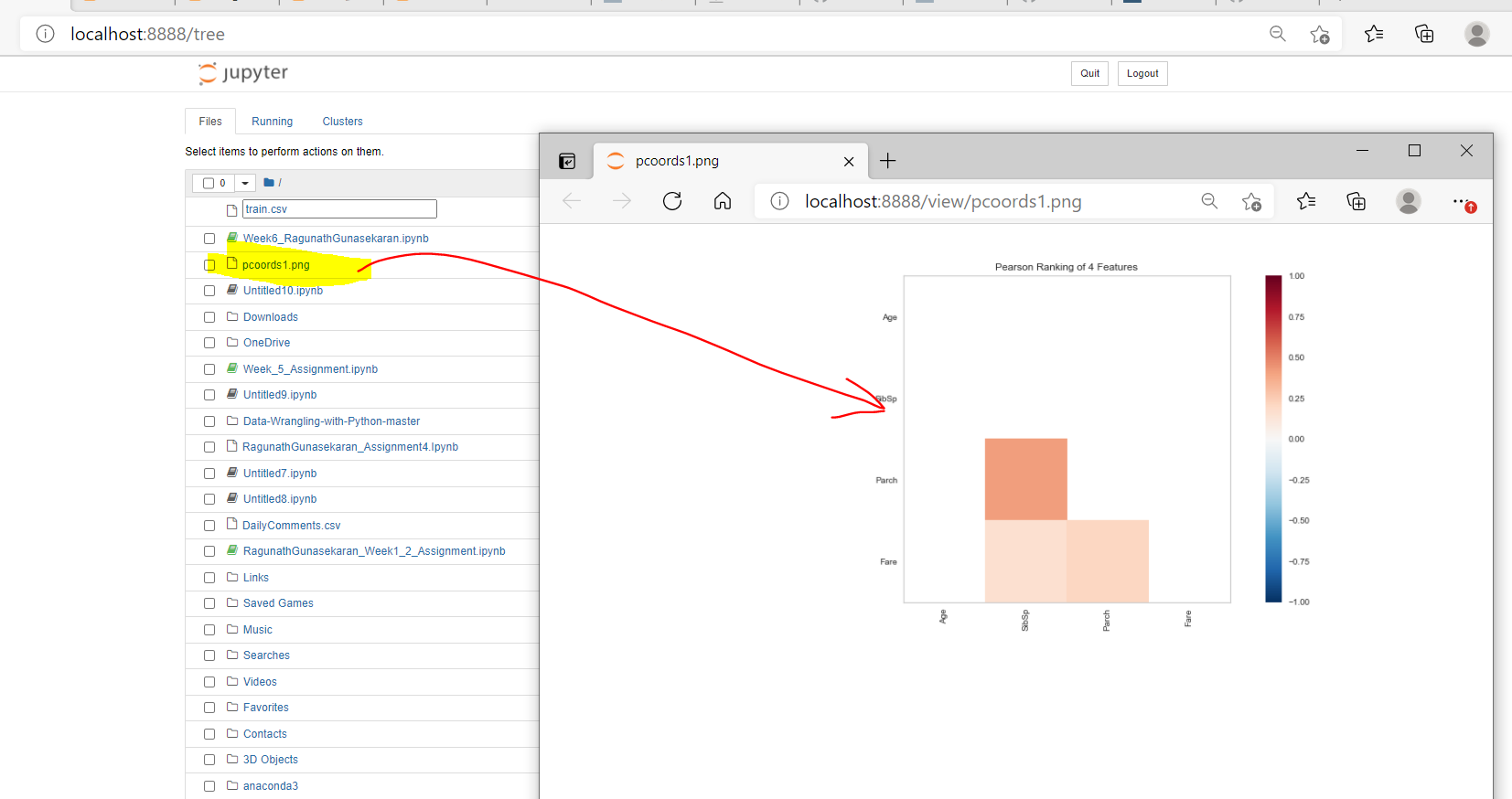
## Make some bar charts for variables with only a few options.

## Ticket and Cabin have more than 100 variables so don’t do those!

Agreed that Ticket and Cabin are having more distinct values since those are categorical values. Bar charts of Survived, Pclass, Sex, embarked explains the various distribution details. Examples, more people did not survive. 3rd class, male, S Embarked passengers are more.

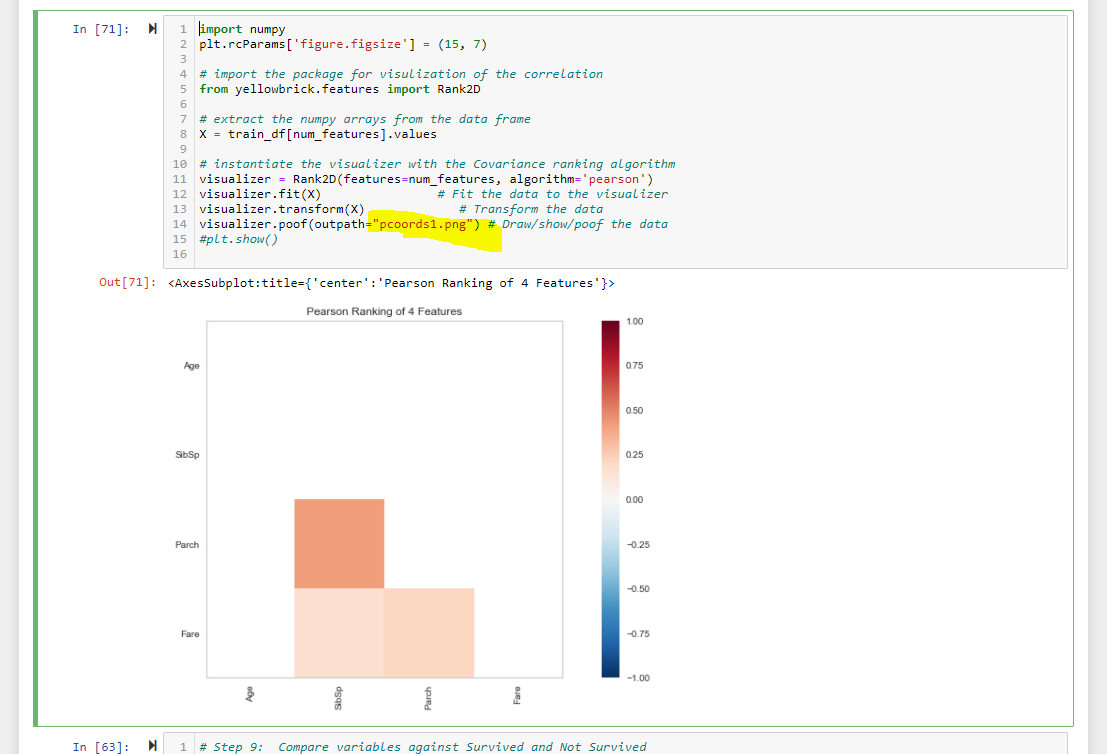
## To see if the data is correlated, make some Pearson Ranking charts

* 1. Notice that in my sample code, I have saved this png file.



* 1. The correlation between the variables is low (1 or -1 is high positive or high negative, 0 is low or no correlation) These results show there is “some” positive correlation but it’s not a high correlation.

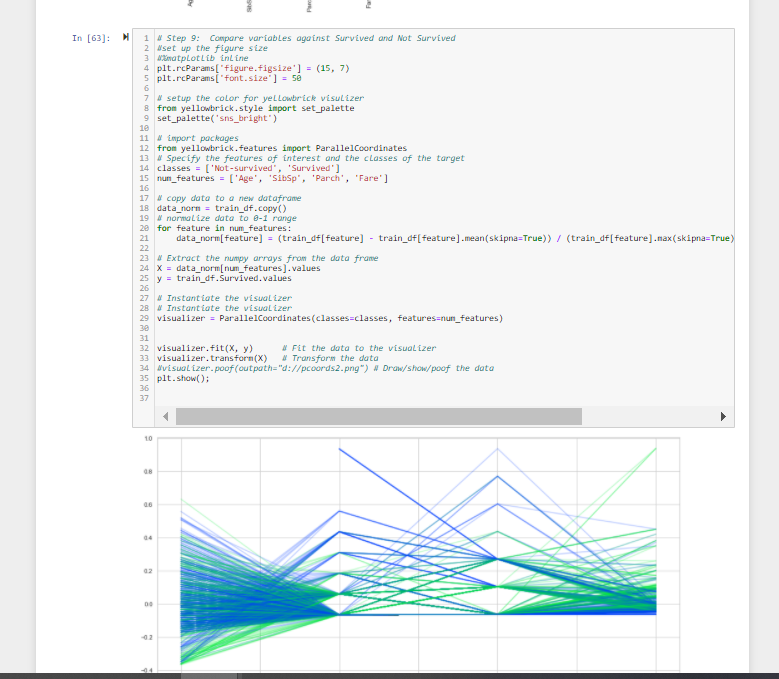
This shows that how Age, SibSp, Parch and Fare related each other. It’s shows that Parch and SibSp are related well when compare to other variables.



## Use Parallel Coordinates visualization to compare the distributions of numerical variables between passengers that survived and those that did not survive.

* 1. That’s a cool chart, isn’t it?! Passengers traveling with siblings on the boat have a higher death rate and passengers who paid a higher fare had a higher survival rate.

Yes, Passengers traveling with siblings on the boat have a higher death rate and passengers who paid a higher fare had a higher survival rate.

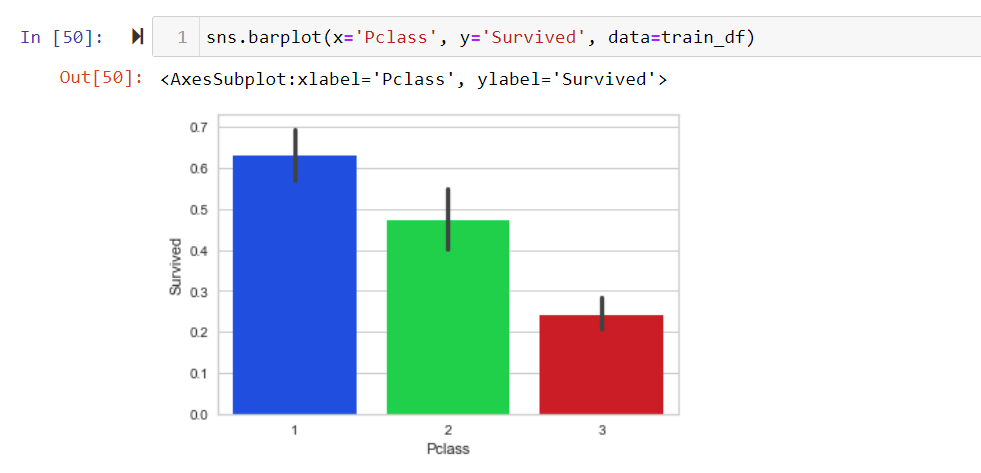
.

## Use Stack Bar Charts to compare passengers who survived to passengers who didn’t survive based on the other variables.

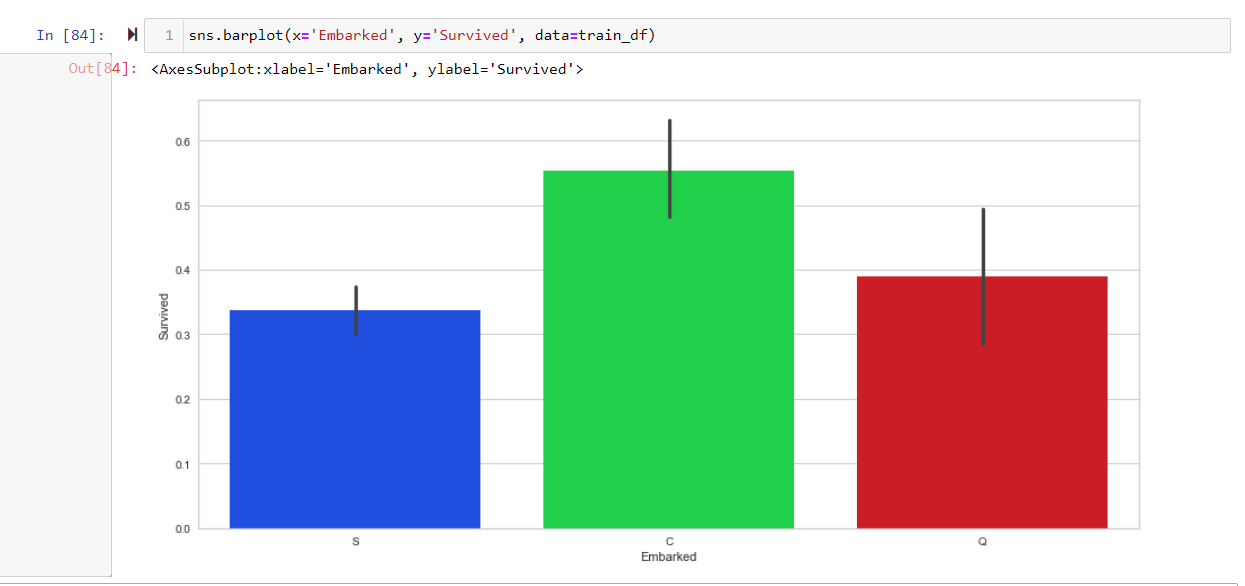
* 1. More females survived than men. 3rd Class Tickets had a lower survival rate. Also, Embarkation from Southampton port had a lower survival rate.

Yes, Female survived more than Male.



3rd class has less survival Rate

Embarkation from Southampton port had a lower survival rate

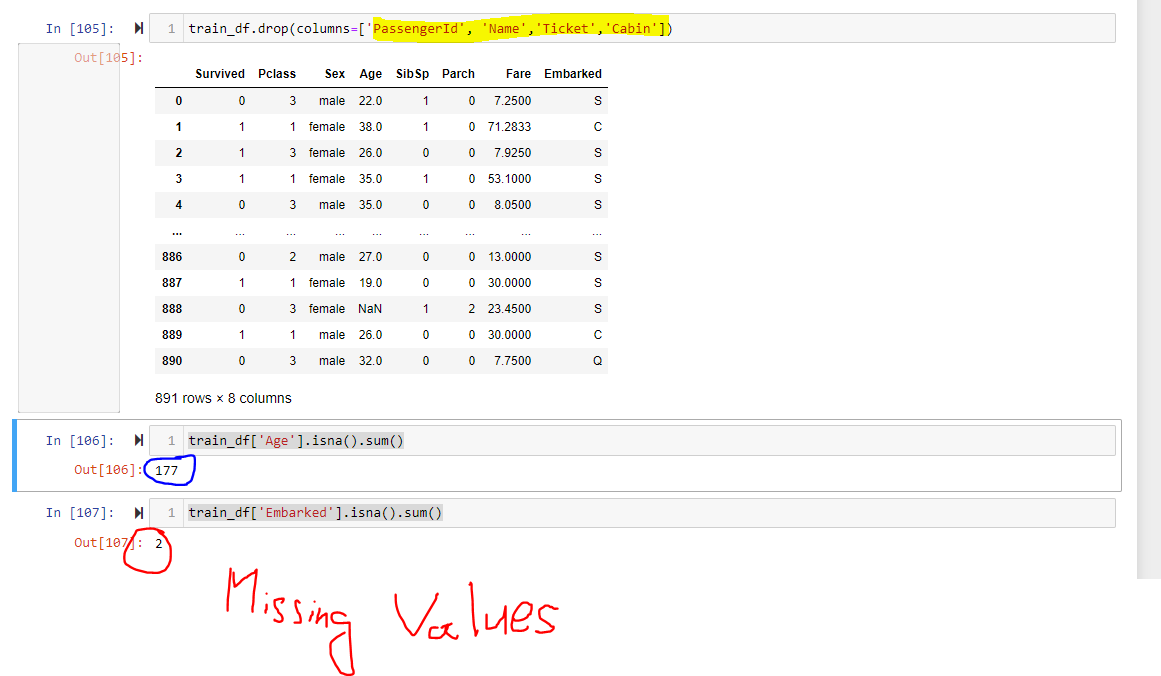


## Some of my questions have been answered by seeing the charts but in some ways, looking at this much data has created even more questions.

## Now it’s time to reduce some of the features so we can concentrate on the things that matter! There features we will get rid of are: "PassengerId", "Name", "Ticket" and "Cabin". (ID doesn’t really give us any useful data, Ticket and Cabin have too many variables. Name might reflect that they are related but we’re keeping the category about siblings (for now).

At first Step, Removed the four columns - : "PassengerId", "Name", "Ticket" and "Cabin."

Identified the NA data for Age and Embarked.



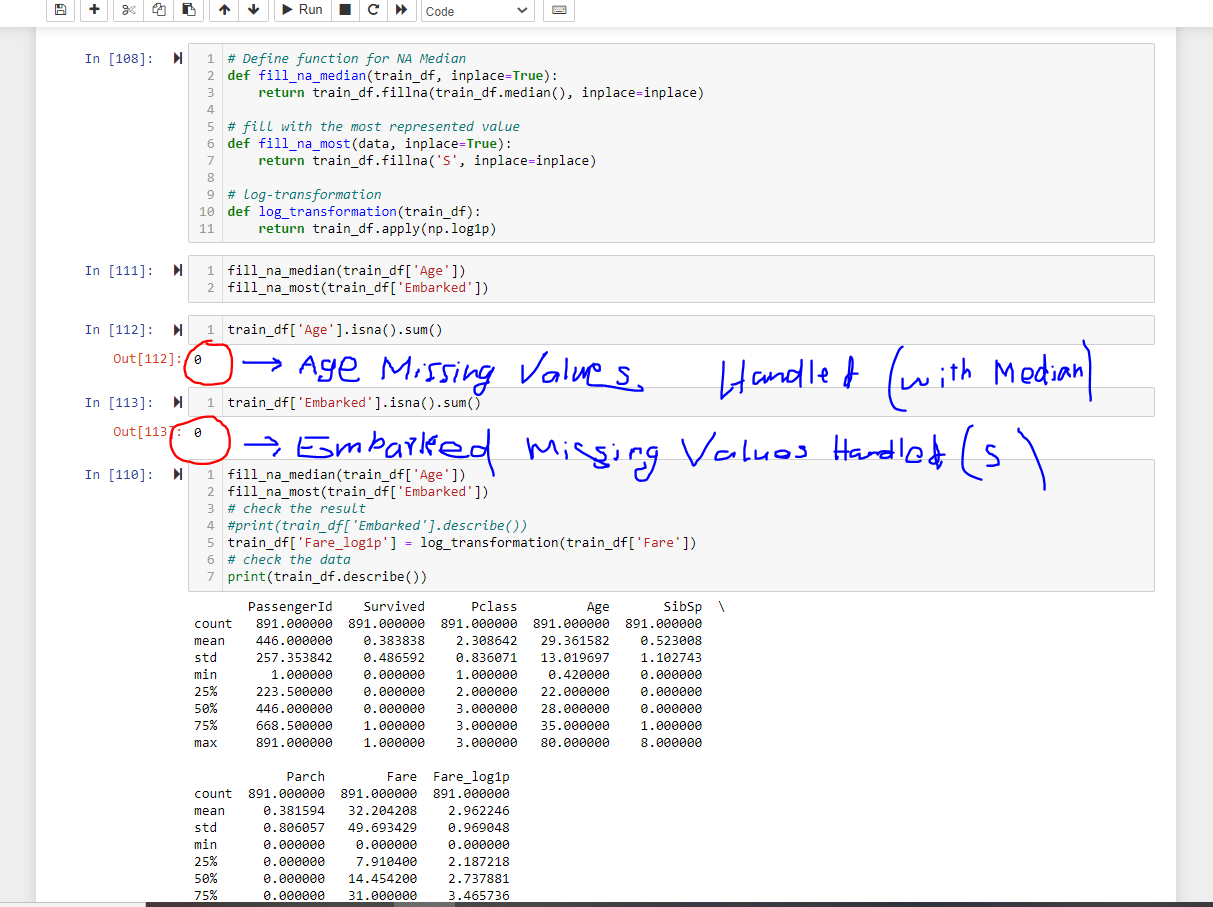
## Age has some missing values, so I’ll fill in with the average age. Embarked also has some missing so I’ll the most common.

Fill\_na\_median – If na values replaced with median value

Fill\_na\_most – If na values replaced with S

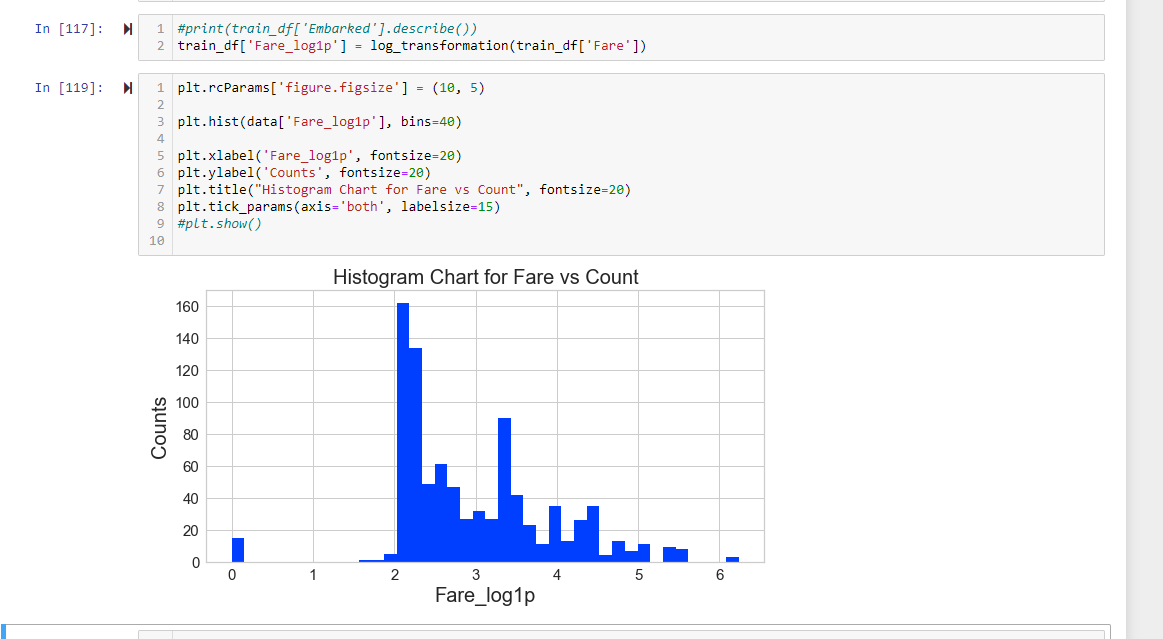
Age missing values handed ( Replaced NA values with Median age )

Embarked missing values handed ( Replaced NA values with S )



## If you go back and look at the histograms of Fare, you’ll see that it is very skewed…many low cost fares, not very many high cost fares. Log Transformation is a good method to use on highly skewed data.

Yes , Data is very skewed and many of the fare falls in the low cost (2 -3).



## Convert your categorical data into numbers (Sex, PClass, Embark)

Converted the categorical data into numbers.

Pclass – 1st, 2nd, 3rd

Sex – Male, Female

Embarked – C,Q,S



## Training - Split your data into two sets: Training and Testing.

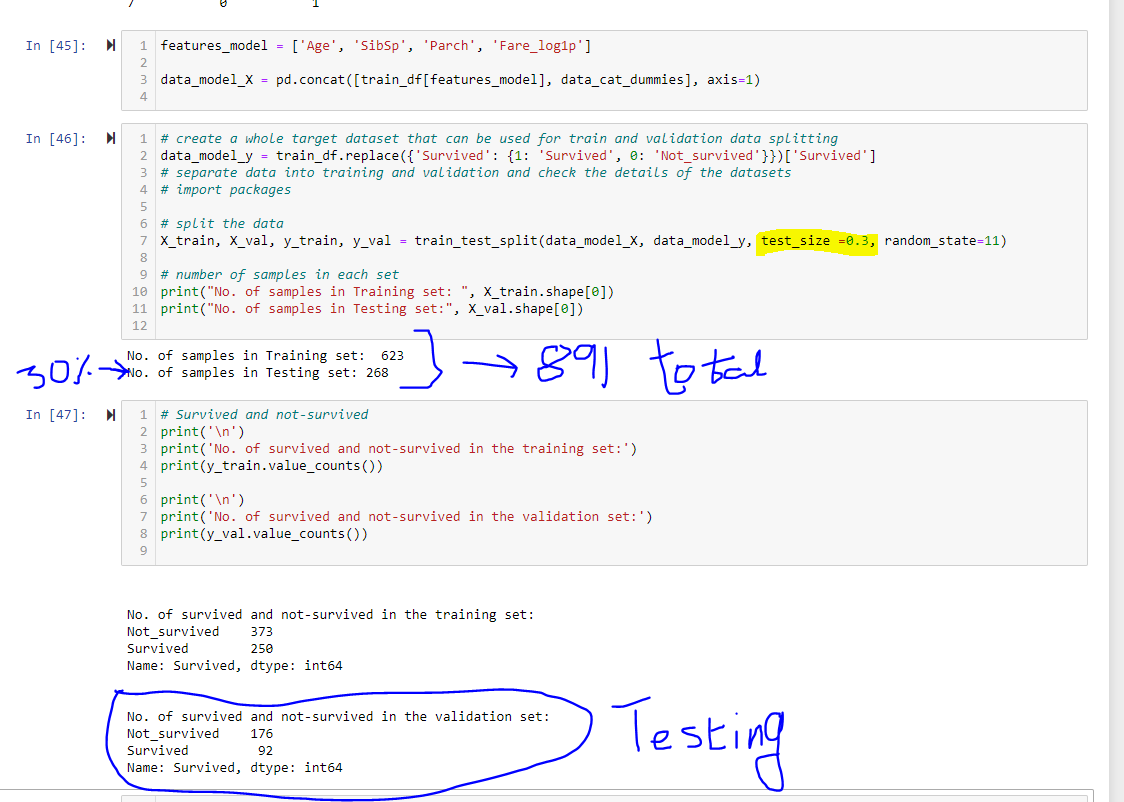
Used train\_test\_split library from sklearn.

from sklearn.model\_selection import train\_test\_split

1. **x\_train:** The training part of the first sequence (x)
2. **x\_val / test:** The test part of the first sequence (x)
3. **y\_train:** The training part of the second sequence (y)
4. **y\_val / test:** The test part of the second sequence (y)

Testing =0.3 🡪 which means 30% ( 268 rows )

Training = 0.7 which means 70% ( 623 rows )

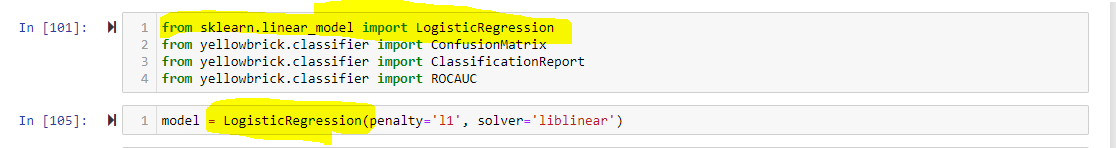


## Evaluation – Remember, we are trying to predict if a passenger has survived or not so this is a classification problem. There are many algorithms that could be used but we’re going to use logistic regression.

Metrics for the evaluation and Confusion Matrix

LogisticRegression used.

84.32% Confusion Matrix model accuracy.





Precision, Recall & F1 score

* F1 Score : used to measure a test’s accuracy.
* Precision : It is the number of correct positive results divided by the number of positive results predicted by the classifier.
* Recall : It is the number of correct positive results divided by the number of all relevant samples

Precision, Recall and F1 Score are very good. Based on the visualization, we can see that Precision, Recall and F1 Score are more than 79% and we can see that all three are looking very good. ( in Red color )



ROC curve (the dotted line is the randomly guessed so anything above that is good metric)

The dotted line is the Randomly guessed.

We can see that ROC Curve for Class 0,1

Also, we have shown the Micro and Macro Curve.

