Assignment 2- Introduction To Data Science

Instructions:

- **Submit only one colab (.ipynb) file and one this report file (.pdf).**
- **Files should be named as yourrollnumber.ipynb (22L7521.ipynb, 22L7521.pdf)**
- **You are provided with two dataset files (Iris, Titanic) .csv files**
- **You have to provide code for all datasets of the necessary steps described in the tables of each question.**
- **Only the mentioned columns/features mentioned for each dataset should be used.**

Part A. Preprocessing

1. In this step, you are required to apply the preprocessing steps that you've covered in the course. Specifically, for each of the input dimension, fill in the following (add rows and complete the table for all input dimensions).

Iris:

Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	Std_ Dev
SepalLength	Float 64	35	0	0	4.3	7.9	5.0	5.843	5.8	0.685	0.82
SepalWidth	Float 64	23	0	4	2.0	4.4	3.0	3.05	3.0	0.188	0.43
PetalLength	Float 64	43	0	0	1.0	6.9	1.5	3.75	4.35	3.113	1.76
PetalWidth	Float 64	22	0	0	0.1	2.5	0.2	1.19	1.3	0.58	0.76
Species	object	3	0	null	Iris-Setosa	Iris-virginica	Iris-Setosa	null	null	null	null

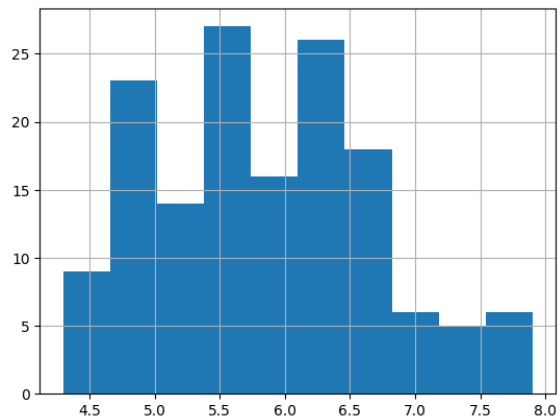
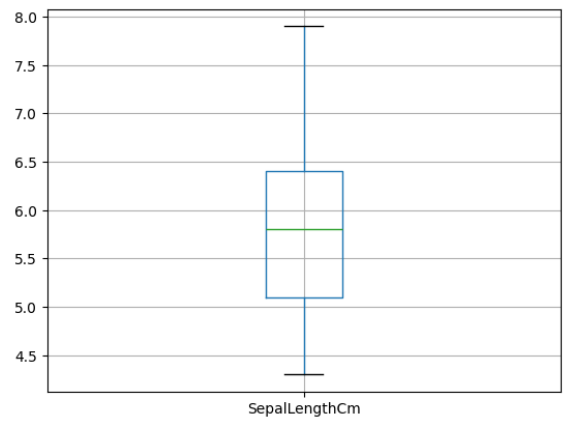
Titanic:

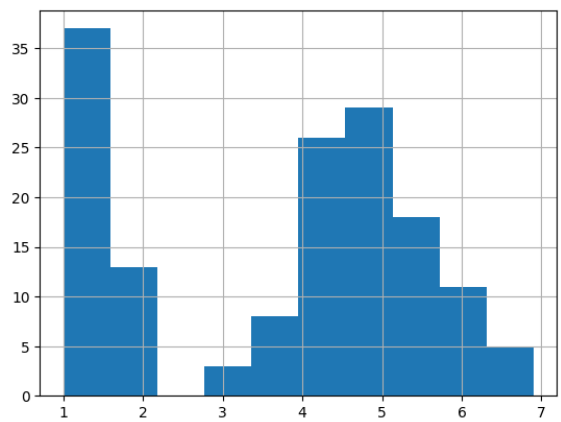
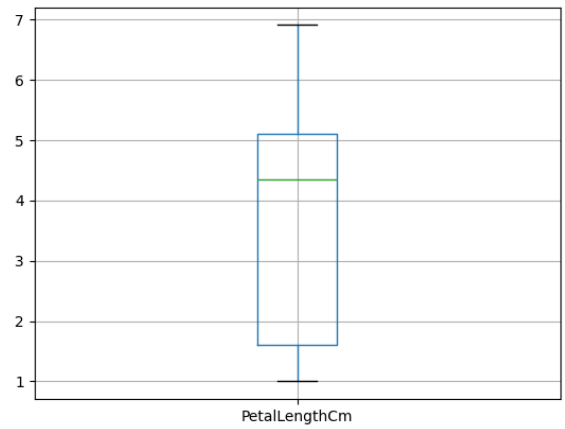
Dim Name	Data Type	Total Instances	Number of Nulls	Number of Outliers	Min. Value	Max Value	Mode	Mean	Median	Variance	Std_Dev
PassengerId	Int64	891	0	0	1	891	1	446.0	446.0	66231.0	257
Survived	Int64	2	0	0	0	1	0	0.3838	0.0	0.23	0.48
Pclass	Int64	3	0	0	1	3	3	2.3086	3.0	0.69	0.83
Name	object	891	0	null	Abbing, Mr Anthony	Van melkebebe, Mr Philemon	Abbig, Mr Anthony	null	null	null	null
Sex	object	2	0	null	female	male	male	null	null	null	null
Age	Float 64	88	177	11	0.42	80.0	24.0	29.6	28.0	211.01	14.52
SibSp	Int64	7	0	46	0	8	0	0.52	0	1.21	1.10
Parch	Int64	7	0	213	0	6	0	0.38	0	0.64	0.8
Ticket	object	681	0	null	110152	WE/P 5735	1601	null	null	null	null
Fare	Float 64	248	0	116	0.0	512.3292	8.05	32.2	14.4	2649.4	49.69
Cabin	object	147	687	null	A10	T	B96 B98	null	null	null	null
Embarked	object	3	2	null	C	S	S	null	null	null	null

2. For each of the input dimension, plot histogram and comment the type of distribution the dimension exhibits. Further, visualize each dimension using a Box Plot. Specifically, for each of the input dimension, you're required to fill the following table (duplicate it for each of the 15 dimensions).

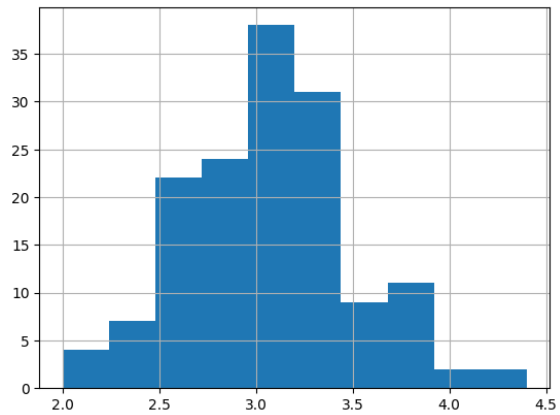
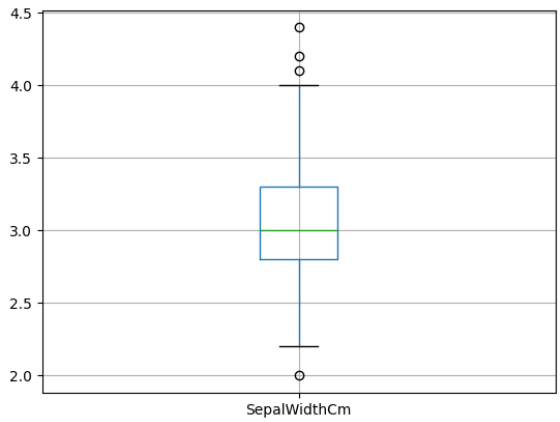
Iris:

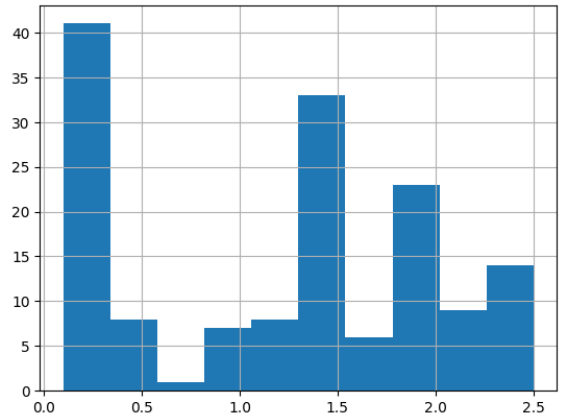
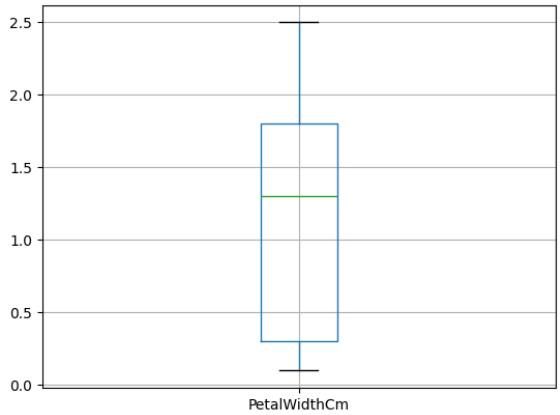
SepalLength	
Histogram	Box Plot

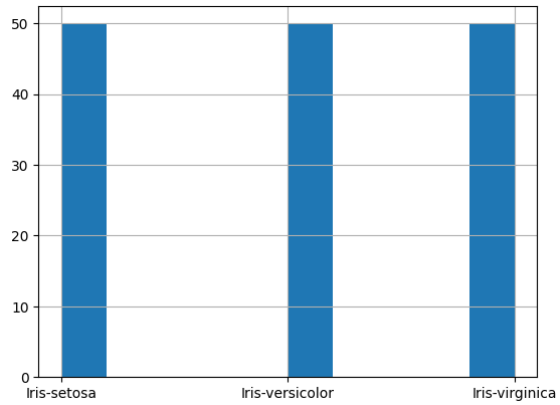
	
Comments: this shows that sepal length is skewed towards right and is not normally distributed	Comments: Sepal length has no outliers. The green line shows the median and the two whiskers show the max and min value

PetalLength	
Histogram	Box Plot
	
Comments: this data is skewed towards left and thus, not normally distributed	Comments: this data also doesn't have any outliers and the box plot also shows that data is skewed towards left.

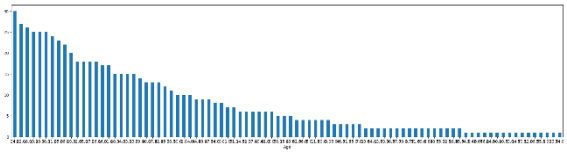
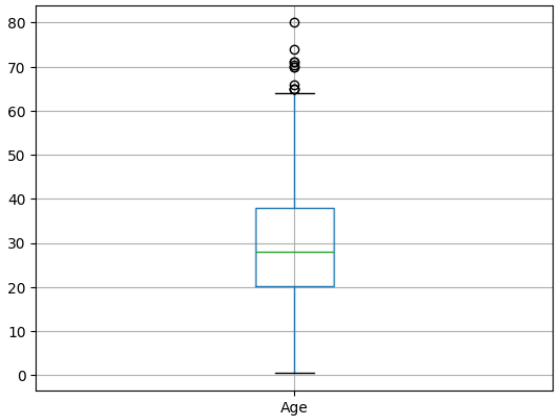
SepalWidth

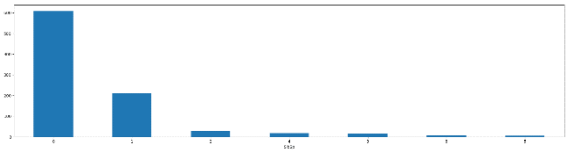
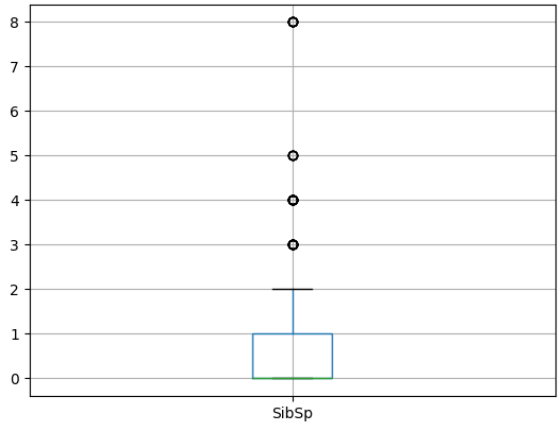
Histogram	Box Plot
	
Comments: this graph is skewed towards the right.	Comments: this data also has outliers.

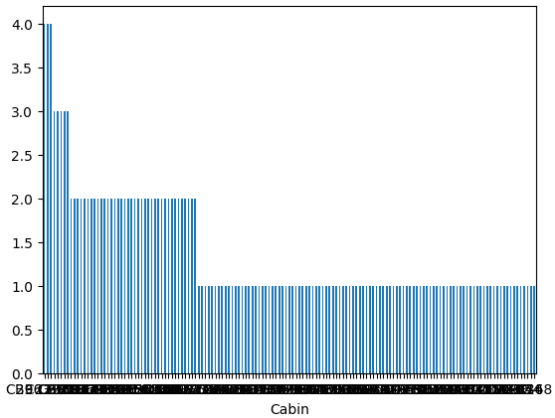
PetalWidth	
Histogram	Box Plot
	
Comments: this data is skewed towards left.	Comments: this data has no outliers

Species	
Histogram	Box Plot
	
Comments: this data is normally distributed and has no skewness which means mean = median = mode	Comments: box plot cant be made because species is categorical

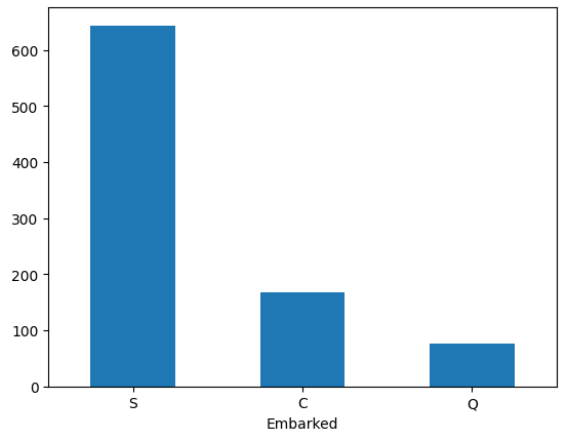
Titanic:

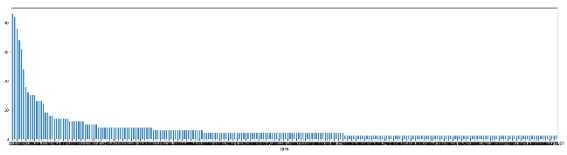
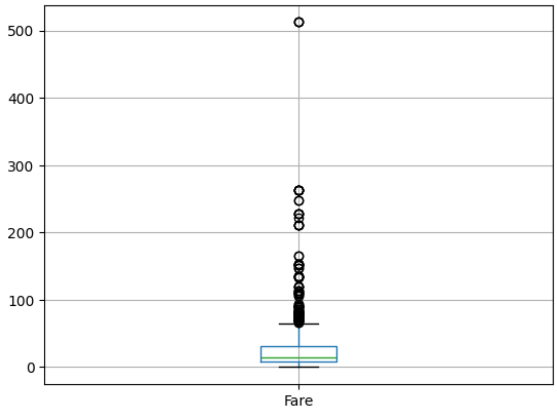
Age	
Histogram	Box Plot
	
Comments: this data is skewed towards the right and is not normally distributed.	Comments: this data also has many outliers.

SibSp	
Histogram	Box Plot
	
Comments: this data is skewed to the right.	Comments: this data also has outliers and the min value, median and Q1 have the same values.

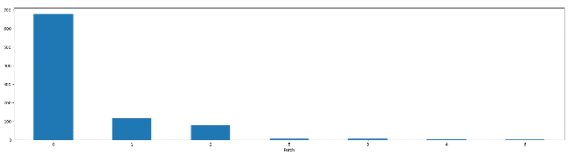
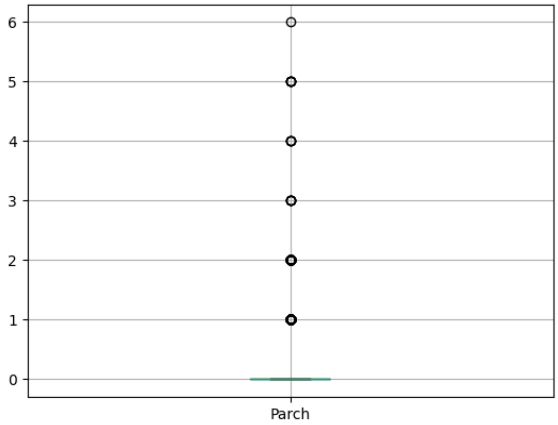
Cabin	
Histogram	Box Plot
	null

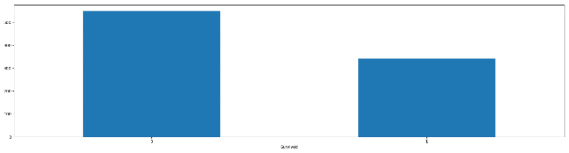
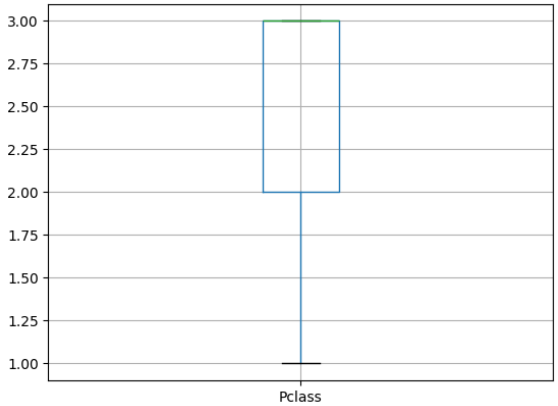
Comments: this data is skewed towards right.	Comments: this data cant be plotted into a boxplot as it is categorical

Embarked	
Histogram	Box Plot
	null
Comments: this data is skewed to the right and is imbalanced.	Comments:

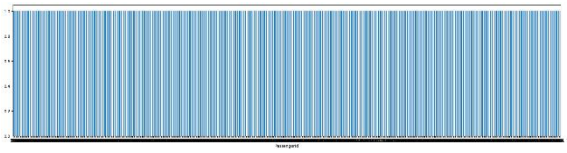
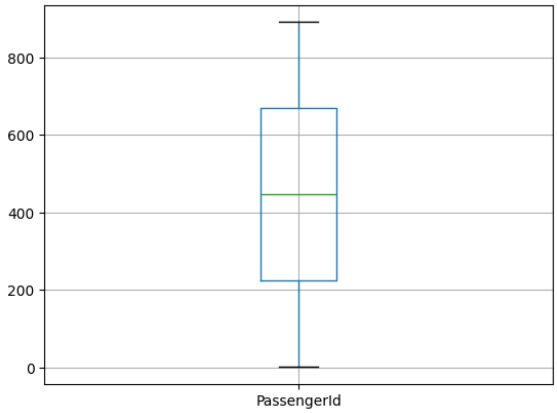
Fare	
Histogram	Box Plot
	

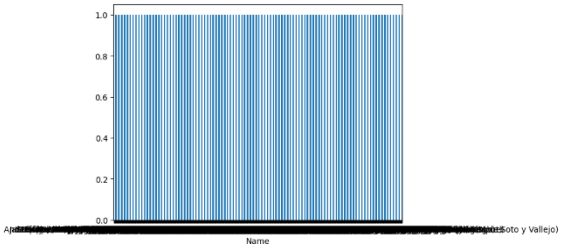
Comments: this data is highly imbalanced and is skewed to the right	Comments: this dimension also has many outliers.

Parch	
Histogram	Box Plot
	
Comments: skewed to the right.	Comments: has many outliers. Whereas,

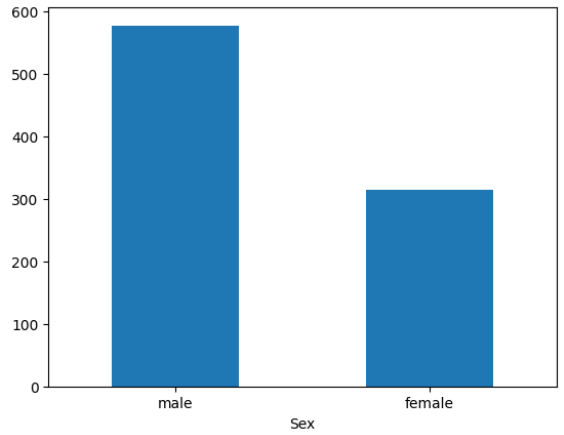
Pclass	
Histogram	Box Plot
	

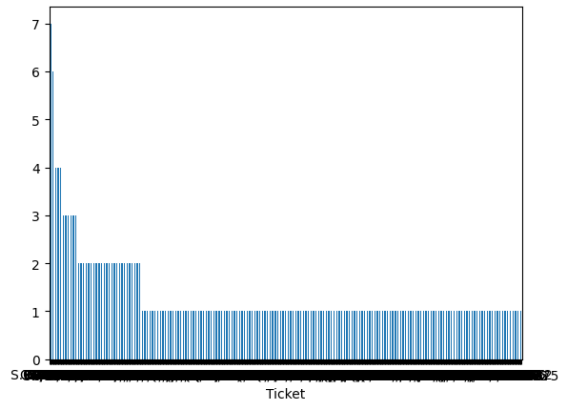
Comments: skewed to the right.	Comments: the max value and Q3, and median have the same value.

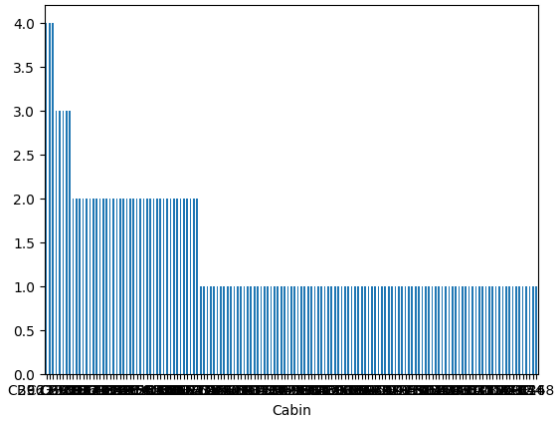
PassengerId	
Histogram	Box Plot
	
Comments: this data is normally distributed	Comments:

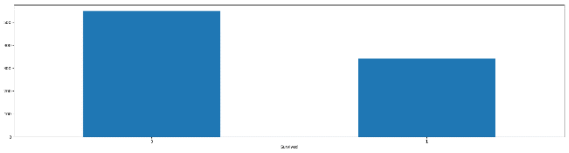
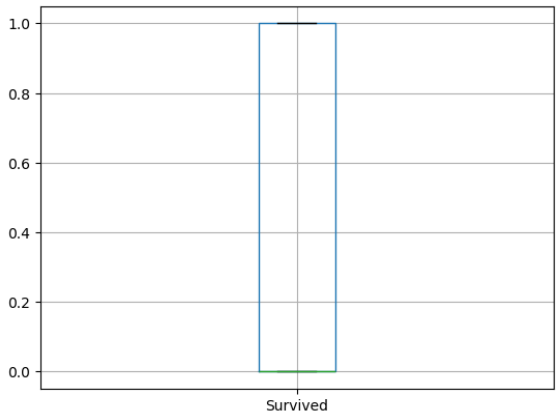
Name	
Histogram	Box Plot
	null

Comments: normally distributed.	Comments:

Sex	
Histogram	Box Plot
	null
Comments: skewed to the right.	Comments:

Ticket	
Histogram	Box Plot
	null
Comments: skewed to the right.	Comments:

Cabin	
Histogram	Box Plot
	null
Comments: skewed to the right.	Comments:

Survived	
Histogram	Box Plot
	
Comments: skewed to the right.	Comments: the Q1, min value and Q3, max value has the same values.

3. Find the missing values in each of the dimension (do this for both input and output dimensions), and fill these using an "appropriate" methodology that we've discussed in the class. You may also choose to drop a certain sample based on your analysis. Mention your approach and its justification.

Iris:

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
SepalLength	0	Mean	Because the column value is numerical and so, it cant be filled with mean
SepalWidth	0	Mean	same
PetalLength	0	Mean	same
Species	0	Mean	same
PetalWidth	0	Mean	same

Titanic:

Dim Name	Number of Missing Values	Filled using OR Dropped	Reason for selecting a certain approach
PassengerId	0	Mean	As it is a numeric value so mean can be used
Survived	0	mean	
Pclass	0	mean	
Name	0	unknown	As name of every person is distinct so it cant be filled with mode

Sex	0	mode	For categorical value we used mode
Age	177	mean	
SibSp	0	mean	
Parch	0	mean	
Ticket	0	unknown	Ticket number for every passenger is also unique so it cant be filled w mode
Fare	0	mean	
Cabin	687	unknown	Cabin number is also distinct so mode cant be used
Embarked	2	mode	