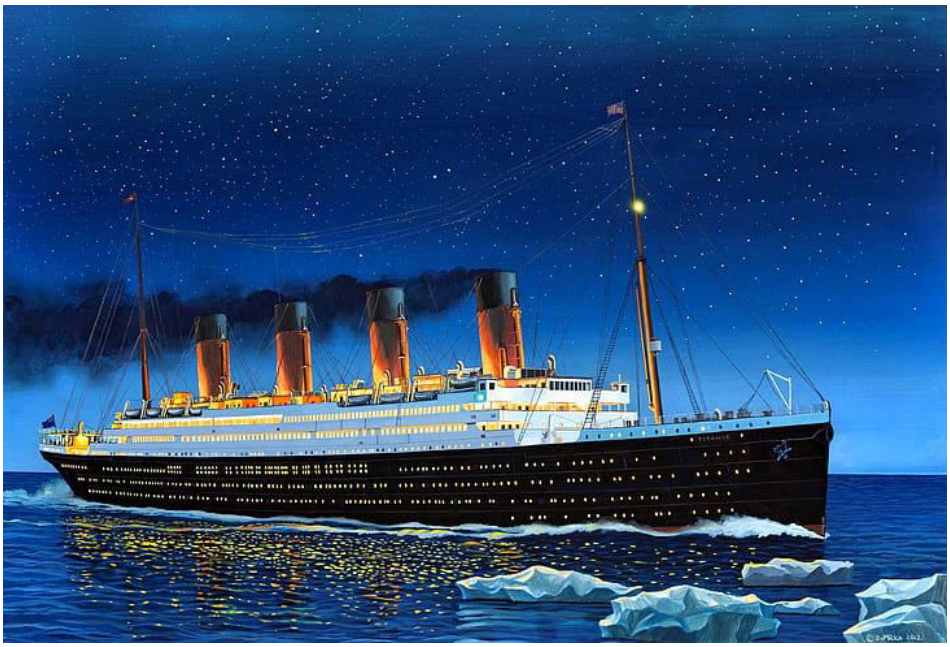
NAME: MD ATIKUJJAMAN

STUDENT ID : ATI215087

(DATA SCIENCE)

Applied Logistic Regression (Supervised learning model) on Titanic: Machine Learning from Disaster Data-set from Kaggle



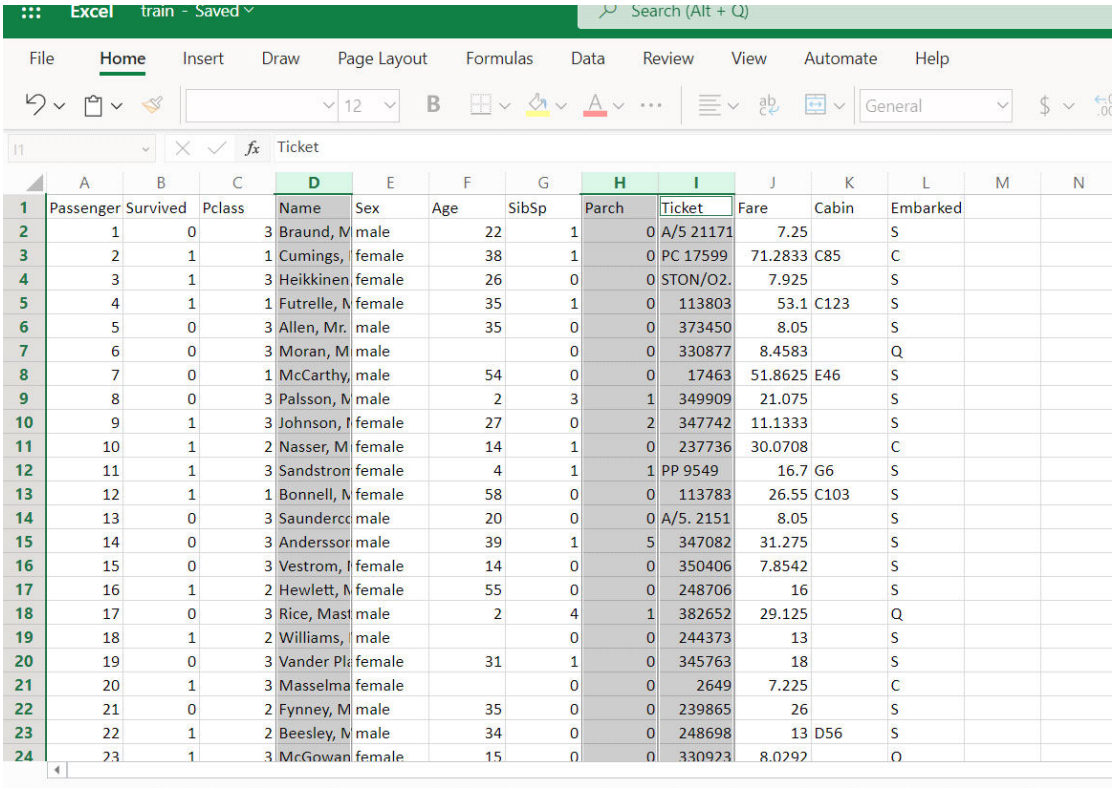
INDEX

1.INTRODUCTION

In this report , i am going to discuss about the dataset i have downloaded from this website https://www.kaggle.com/competitions/titanic/data .i am going to do analysis for different genders on the basis of different factors which are discussed below.The training set should be used to build my machine learning models. For the training set,i provide the outcome (also known as the “ground truth”) for each passenger. my model will be based on “features” like passengers’ gender and class.

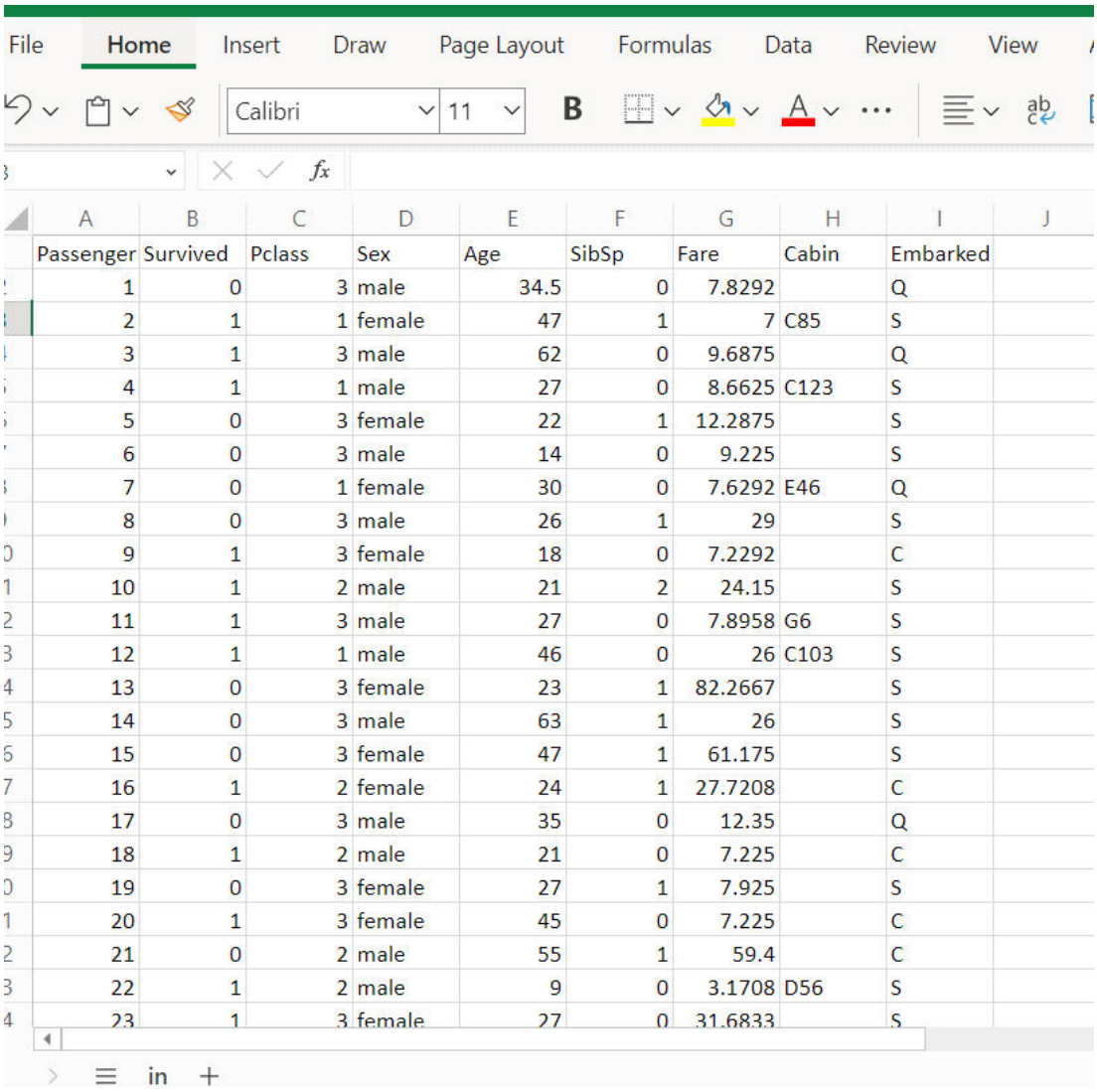
2. PRE-PROCESSING

Step1: First i delete name, Parch and ticket because i am not using this data.in the picture below



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked		
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S		
3	2	1	1	Cumings,	female	38	1	0	PC 17599	71.2833	C85	C		
4	3	1	3	Heikkinen	female	26	0	0	STON/O2.	7.925		S		
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S		
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S		
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q		
8	7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S		
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S		
10	9	1	3	Johnson, I	female	27	0	2	347742	11.1333		S		
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C		
12	11	1	3	Sandstron	female	4	1	1	PP 9549	16.7	G6	S		
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S		
14	13	0	3	Saundercc	male	20	0	0	A/5. 2151	8.05		S		
15	14	0	3	Anderssor	male	39	1	5	347082	31.275		S		
16	15	0	3	Vestrom, I	female	14	0	0	350406	7.8542		S		
17	16	1	2	Hewlett, M	female	55	0	0	248706	16		S		
18	17	0	3	Rice, Mast	male	2	4	1	382652	29.125		Q		
19	18	1	2	Williams,	male		0	0	244373	13		S		
20	19	0	3	Vander Pla	female	31	1	0	345763	18		S		
21	20	1	3	Masselma	female		0	0	2649	7.225		C		
22	21	0	2	Fynney, M	male	35	0	0	239865	26		S		
23	22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S		
24	23	1	3	McGowan	female	15	0	0	330923	8.0292		Q		

Step2: I fillup all blank data for age . i use median for fill up blank space. perfect dataset in shown below.



Passenger	Survived	Pclass	Sex	Age	SibSp	Fare	Cabin	Embarked
1	0	3	male	34.5	0	7.8292		Q
2	1	1	female	47	1	7	C85	S
3	1	3	male	62	0	9.6875		Q
4	1	1	male	27	0	8.6625	C123	S
5	0	3	female	22	1	12.2875		S
6	0	3	male	14	0	9.225		S
7	0	1	female	30	0	7.6292	E46	Q
8	0	3	male	26	1	29		S
9	1	3	female	18	0	7.2292		C
10	1	2	male	21	2	24.15		S
11	1	3	male	27	0	7.8958	G6	S
12	1	1	male	46	0	26	C103	S
13	0	3	female	23	1	82.2667		S
14	0	3	male	63	1	26		S
15	0	3	female	47	1	61.175		S
16	1	2	female	24	1	27.7208		C
17	0	3	male	35	0	12.35		Q
18	1	2	male	21	0	7.225		C
19	0	3	female	27	1	7.925		S
20	1	3	female	45	0	7.225		C
21	0	2	male	55	1	59.4		C
22	1	2	male	9	0	3.1708	D56	S
23	1	3	female	27	0	31.6833		S

USING PYTHON COMMANDS:

STEP 1: if there is Null values in the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Sex          418 non-null    object
 4   Age          418 non-null    float64
 5   SibSp        418 non-null    float64
 6   Fare         418 non-null    float64
 7   Cabin        204 non-null    object
 8   Embarked     418 non-null    object
dtypes: float64(3), int64(3), object(3)
memory usage: 62.8+ KB
```

STEP 2: Describe train dataseta

	PassengerId	Survived	Pclass	Age	SibSp	Fare
count	891.000000	891.000000	891.000000	418.000000	418.000000	418.000000
mean	446.000000	0.383838	2.308642	29.599282	0.447368	35.541956
std	257.353842	0.486592	0.836071	12.703770	0.896760	55.867684
min	1.000000	0.000000	1.000000	0.170000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	23.000000	0.000000	7.895800
50%	446.000000	0.000000	3.000000	27.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	35.750000	1.000000	31.471875
max	891.000000	1.000000	3.000000	76.000000	8.000000	512.329200

STEP 3: Checking null values

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	True	False

3.HYPOTHESIS:

Predictcting train survived passengers:

1.Survived data for Pclass and age

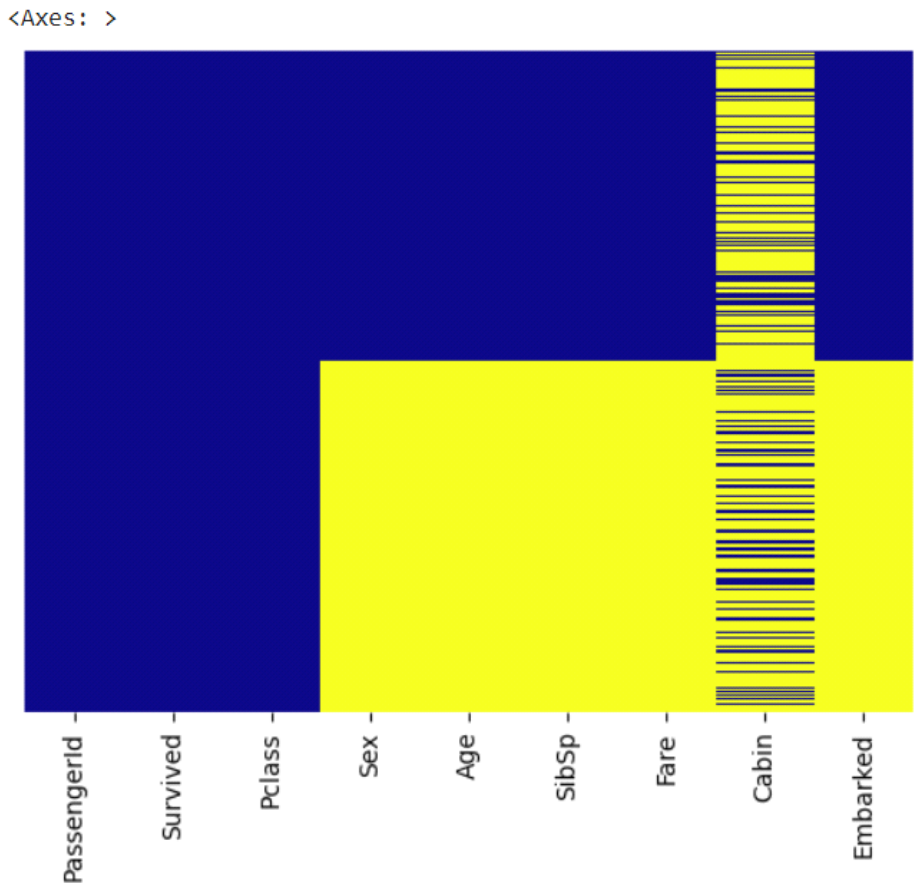
2. The Passengers in class check with age .

3.Making Submission.csv file dataset with passenger id for survived .

4.PROPOSED SOLUTIONS:

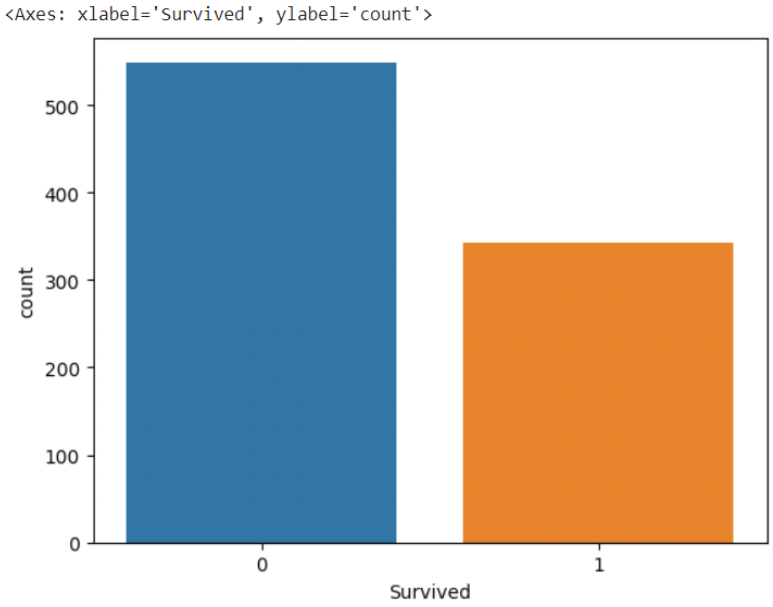
Hypothesis 1:Survived data for Pclass and age:

STEP 1: Plotting the null values on a heatmap for better visualization

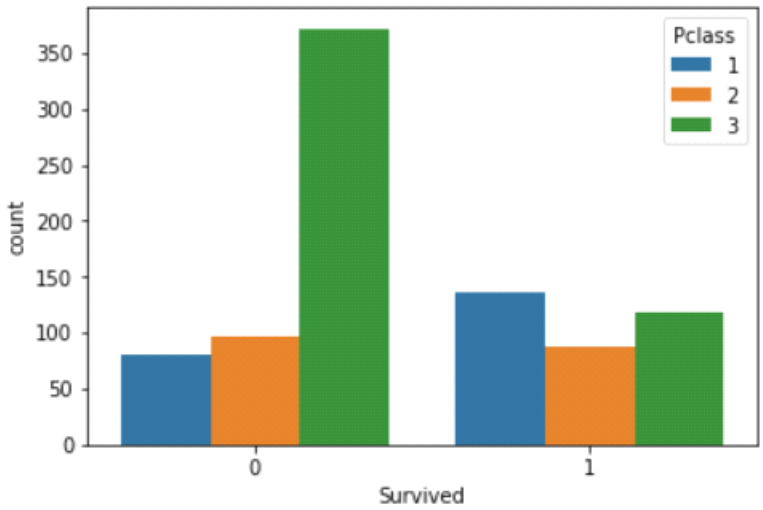


STEP 2: 1. The Age and Cabin columns lack information.

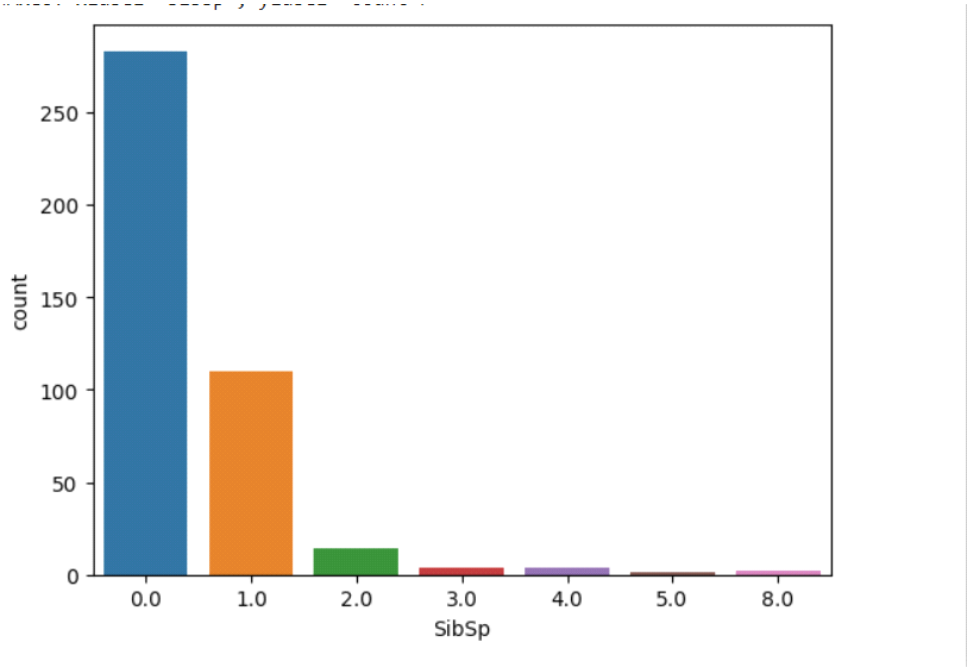
2. Compared to Cabin, the Age column contains considerably less missing values.



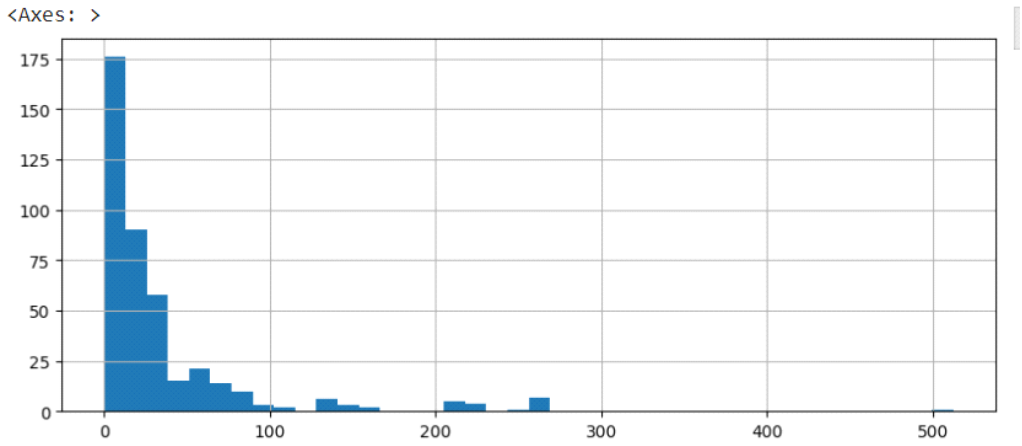
STEP 3: Survived data for Pclass



STEP 4: By looking at this plot, most people on board neither had Sibsp

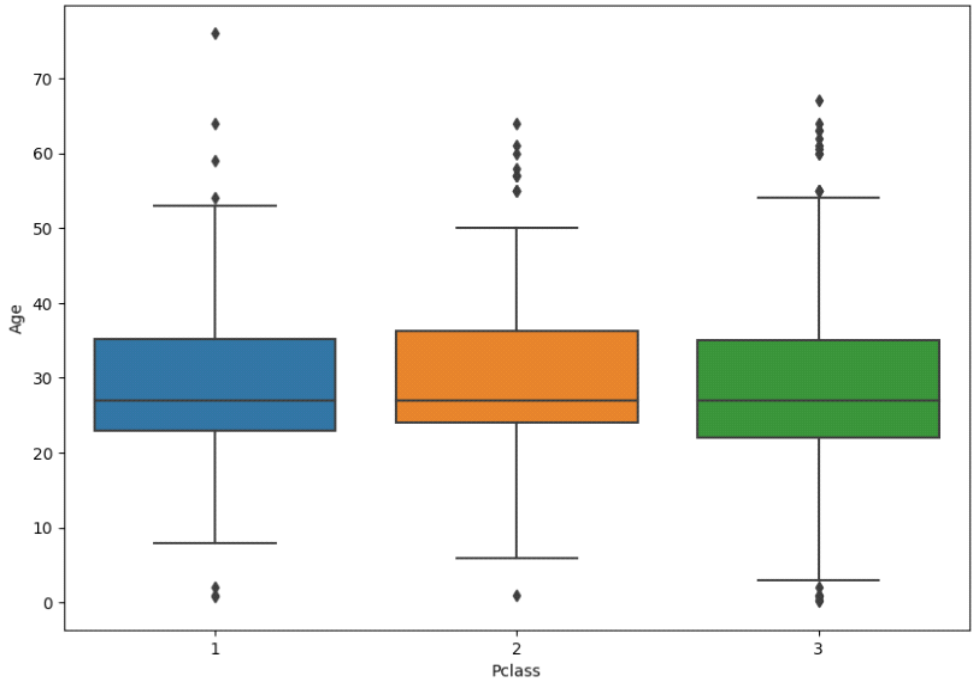


STEP 5: Most of the distribution is between 0 and 100 .

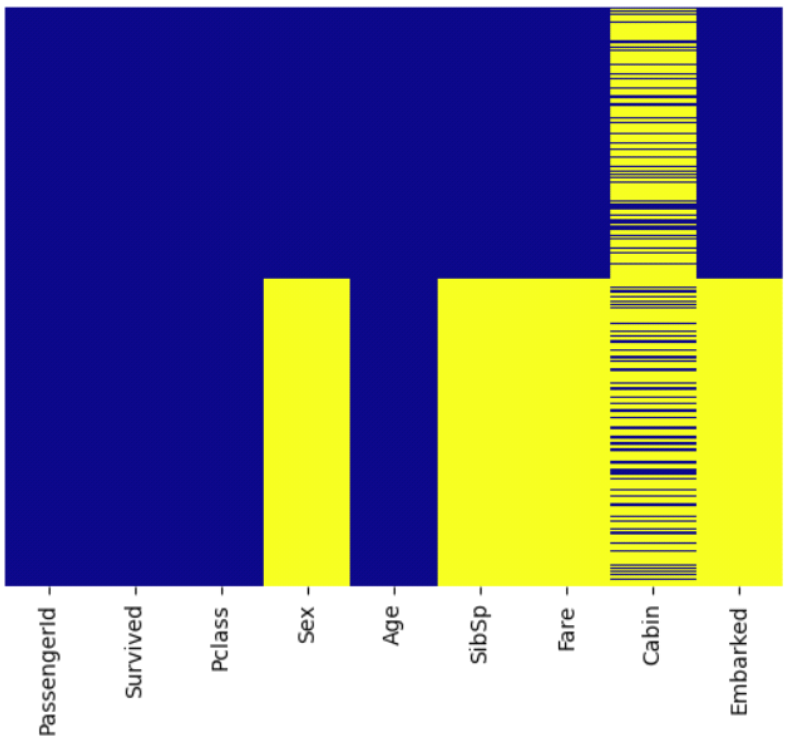


Hypothesis 2: The Passengers in class check with age and i can see there is no more null values for age.

STEP 1: The figure shows that the Passengers in class 1 have older people And younger people in lower Pclass



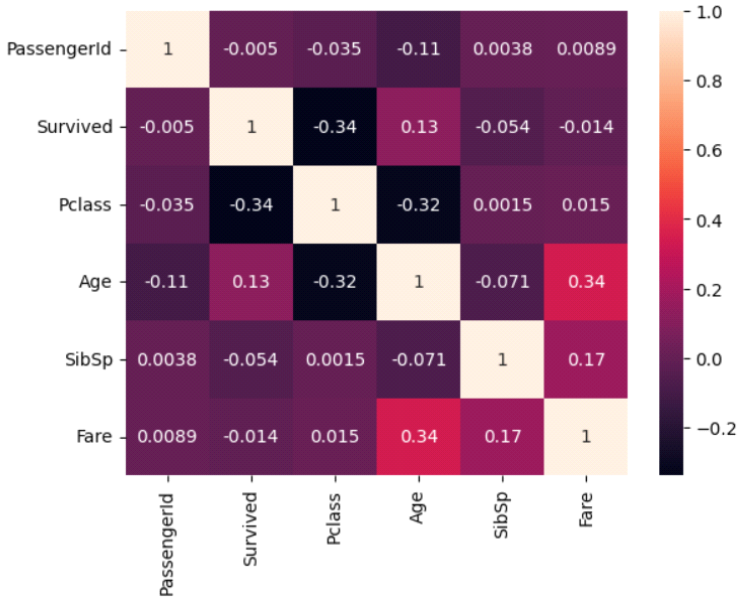
STEP 2: No more missing values in Age



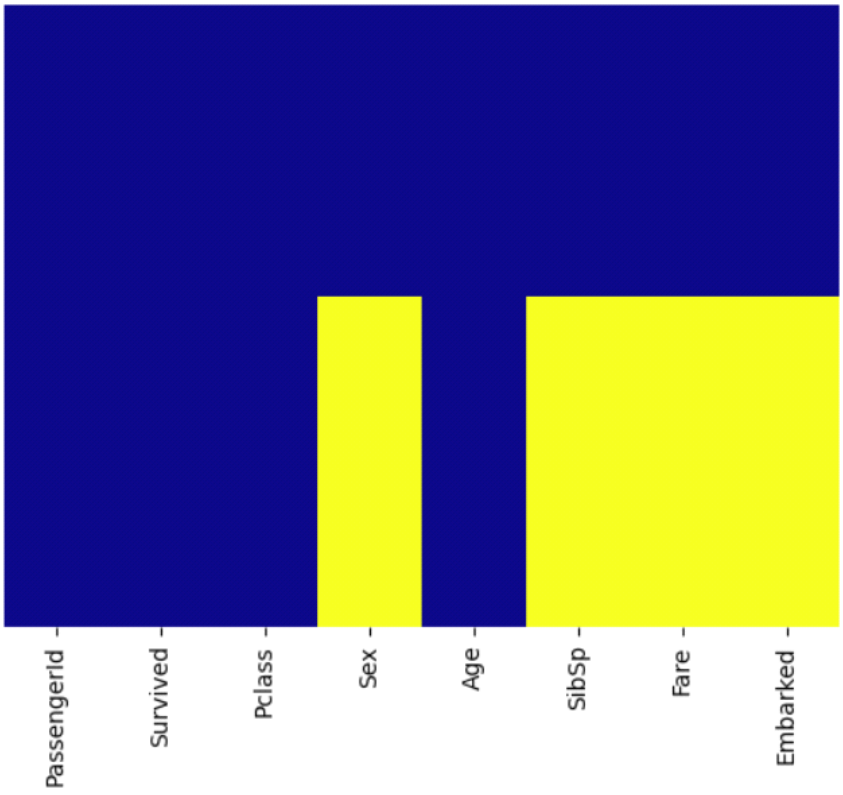
Hypothesis 3(Making Submission.csv file dataset with passenger id for survived **):** For this hypothesis i check all data

correlation between columns and i check again all null values for all data.Then i need to encoded sex and enbarked columns because the machine learning alogorithm won't be able process the data . finally i will get perfect data for machine learning algorithm and my data will be ready for csv file . I use Logistic-Regression for this one.

STEP 1: Checking for correlation between columns



STEP 2: Final check for null values



STEP 3: 1. I Need to convert the sex column,otherwise the machine learning alogorithm won't be able process the data

2. I can not feed both these columns as male and female are opposite

	female	male
0	0	1
1	1	0
2	0	1
3	0	1
4	1	0

	male
0	1
1	0
2	1
3	1
4	0

STEP 4: same process with Embarked column like sex column

	Q	S
0	1	0
1	0	1
2	1	0
3	0	1
4	0	1

STEP 5: Now, i don't need sex, embarked, plcass column because we have encoded them.

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Fare	Embarked	male	Q	S	2	3
0	1	0	3	male	34.5	0.0	7.8292	Q	1	1	0	0	1
1	2	1	1	female	47.0	1.0	7.0000	S	0	0	1	0	0
2	3	1	3	male	62.0	0.0	9.6875	Q	1	1	0	0	1
3	4	1	1	male	27.0	0.0	8.6625	S	1	0	1	0	0
4	5	0	3	female	22.0	1.0	12.2875	S	0	0	1	0	1

STEP 6: I made perfect data for my machine learning algorithm, all data is numeric

	PassengerId	Survived	Age	SibSp	Fare	male	Q	S	2	3
0	1	0	34.5	0.0	7.8292	1	1	0	0	1
1	2	1	47.0	1.0	7.0000	0	0	1	0	0
2	3	1	62.0	0.0	9.6875	1	1	0	0	1
3	4	1	27.0	0.0	8.6625	1	0	1	0	0
4	5	0	22.0	1.0	12.2875	0	0	1	0	1

STEP 7: Train columns

```
Index(['PassengerId', 'Survived', 'Age', 'SibSp', 'Fare', 'male', 'Q', 'S', 2,
       3],
      dtype='object')
```

STEP 8: 1. I made one submissin csv file with passengerid for survived.

2. I made another predictions csv file with passenger id for survived.

PassengerId	Survived
210	1.0
670	1.0
228	0.0
850	1.0
513	1.0
528	1.0
877	0.0
97	1.0
293	0.0
324	0.0
737	0.0
530	0.0
219	1.0
276	1.0
79	0.0
786	0.0
605	1.0
797	1.0
140	1.0
579	0.0
495	0.0
360	0.0
66	0.0
482	0.0
778	0.0

1.submissin.csv file

submission.csv predictions.csv X

991 to 1000 of 1070 entries Filter

PassengerId	Survived
322	0.0
546	1.0
591	0.0
292	1.0
466	0.0
815	0.0
443	0.0
671	0.0
463	1.0
287	0.0

Show 10 per page 1 10 90 99 100 101 102 103 104 105 106 107

2. predictions .csv file

5.REFLECTION:

For my coursework two i made some changes for all hypothesis . because i need to encode sex and embarked columns for mechine learning.

however , apart from this things i fullfilled all of requirement for coursework 2 and i use same data and same median value for null place for this one. i did proper

implemetation as well.

6.REFERENCES:

1. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

2. https://medium.com/analytics-vidhya/your-guide-for-logistic-regression-with-titanic-dataset-784943523994

3. https://www.analyticsvidhya.com/blog/2021/07/titanic-survival-prediction-using-machine-learning/