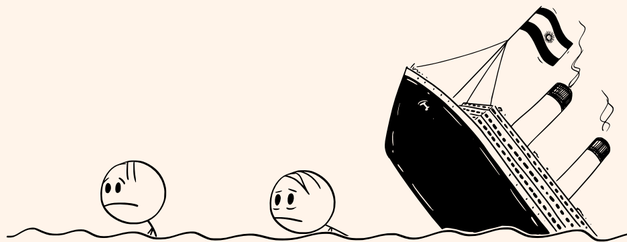
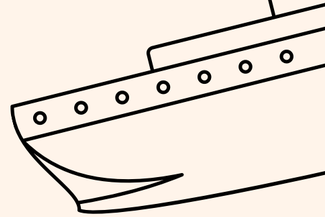
Titanic Data Analysis: Solving Queries and Plotting Graphs

Abstract

- The Titanic dataset is a popular dataset used for data analysis and machine learning.
- In this project, we will explore the dataset and perform various analyses to gain insights into the passengers on board in Titanic.
- Using Python and its data analysis libraries, we will clean and preprocess the data, visualize the data through plots and charts, and draw conclusions.
- Overall, this project will give us a better understanding of the passengers on board the Titanic and their survival compared with various other factors.

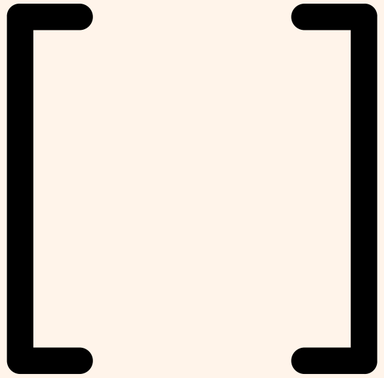




Introduction

- Titanic dataset is taken, imported it to notebook/python environment using panda's library.
- the imported data set is observed and cleaned/preprocessed to obtain meaningful data.
- This data is used to analyze relations between different columns/ features.
- This is also arrange,pick and solve user queries.
- Different Python modules like matplotlib,seaborn,scikitlearn along with numpy and pandas are used.

Libraries Used:

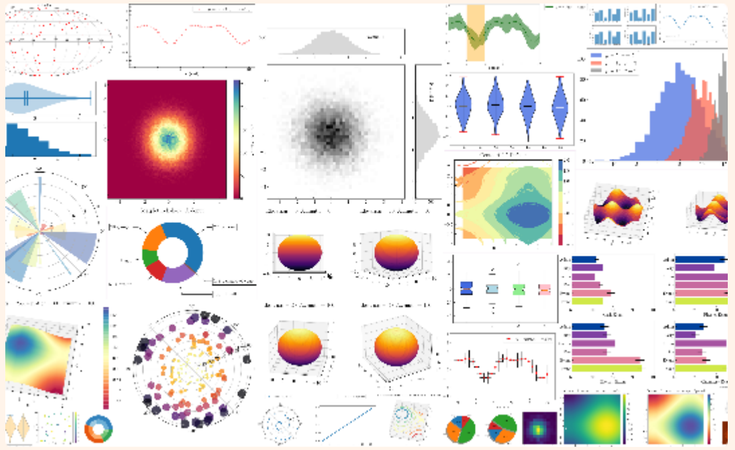


Pandas

A Python library for data manipulation and analysis. Used for reading and Doing operations easily on Dataset.

ScikitLearn

More Advanced Version of Matplotlib. plots relations between features easily, auto scales the data.



Matplotlib

A Python library for creating visualizations. We will use it to plot data and explore the dataset.

Queries

- Distributions of features
- Query to Replace something in data
- Searching a record
- Features vs Features (plots)
- which ages survied most?
- Records of people of age greater than given input
- Youngest, Oldest, Max, Min
- Total and mean
- Age Distributions?
- Names starting with given input
- Correlation matrix

DataSet

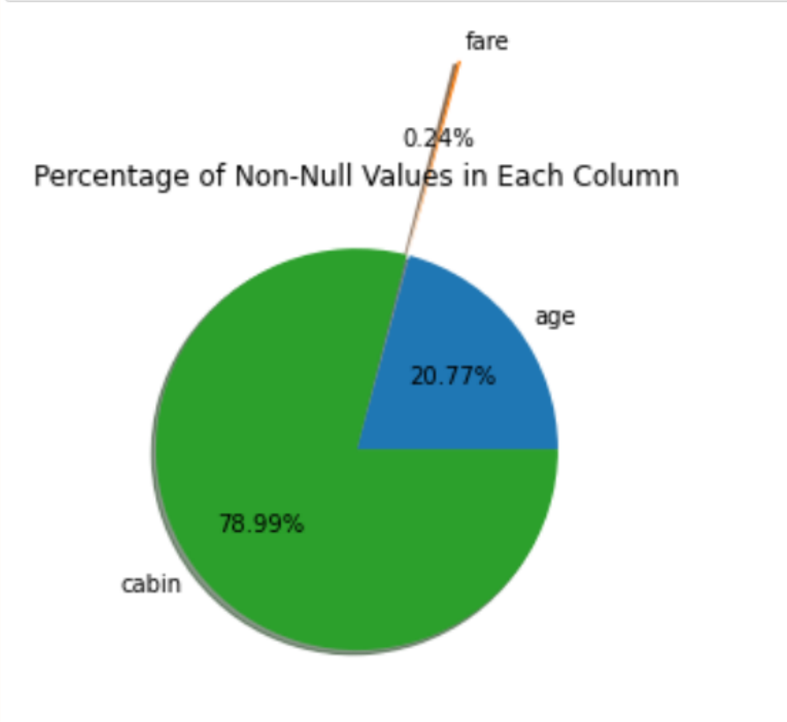
	id	survived	class	name	gender	age	siblings	parch	ticket	fare	cabin	location
0	892	1	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
413	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	1	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

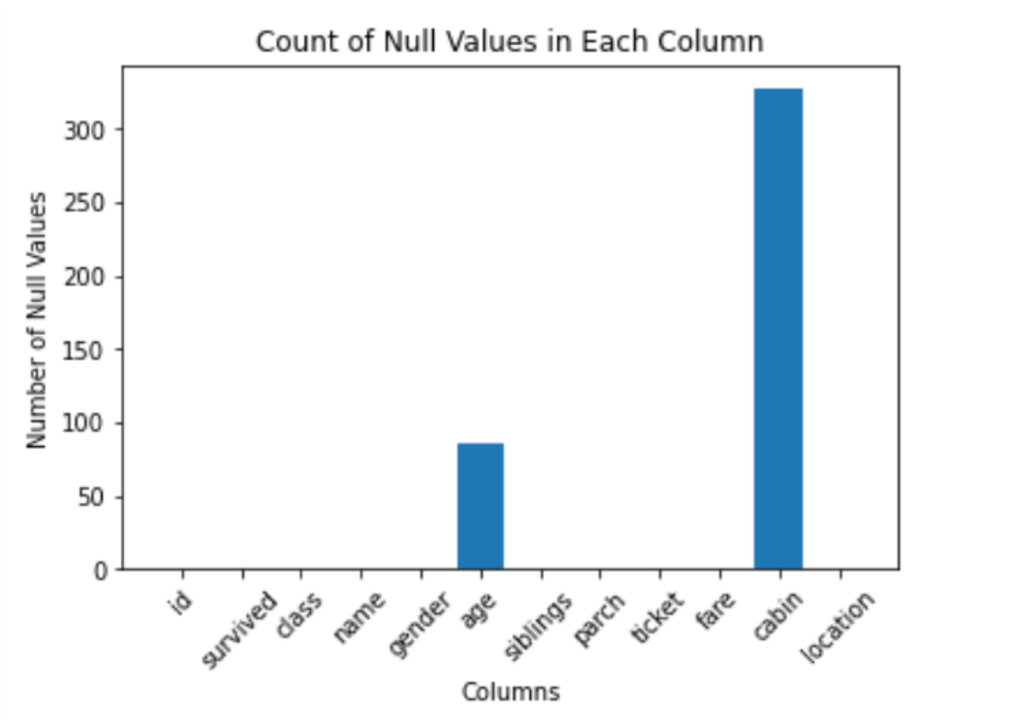
418 rows × 12 columns

Data Cleaning

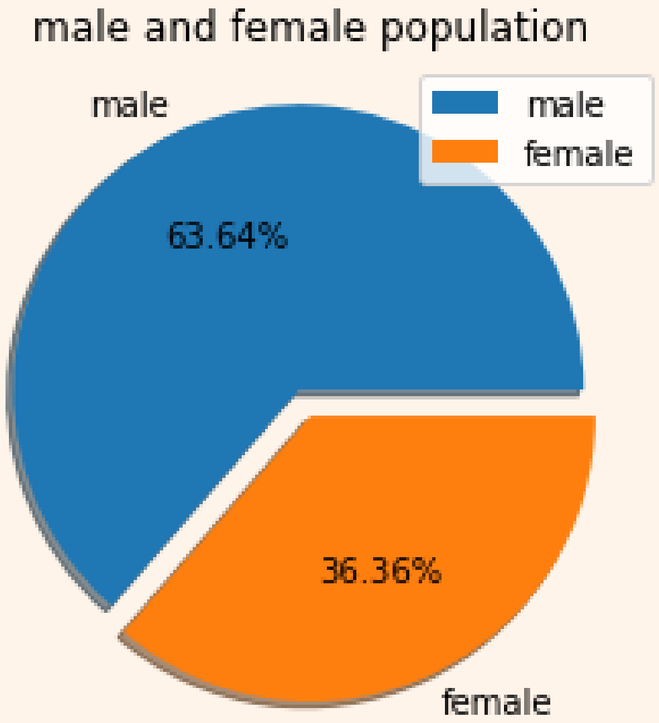
```
df.isnull().sum()  #
```

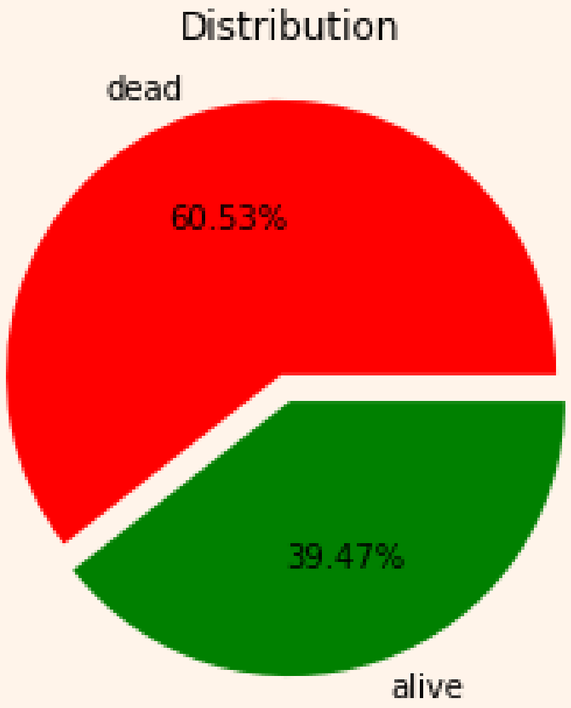
```
id              0
survived        0
class           0
name            0
gender          0
age            86
siblings        0
parch           0
ticket          0
fare            1
cabin         327
location        0
dtype: int64
```

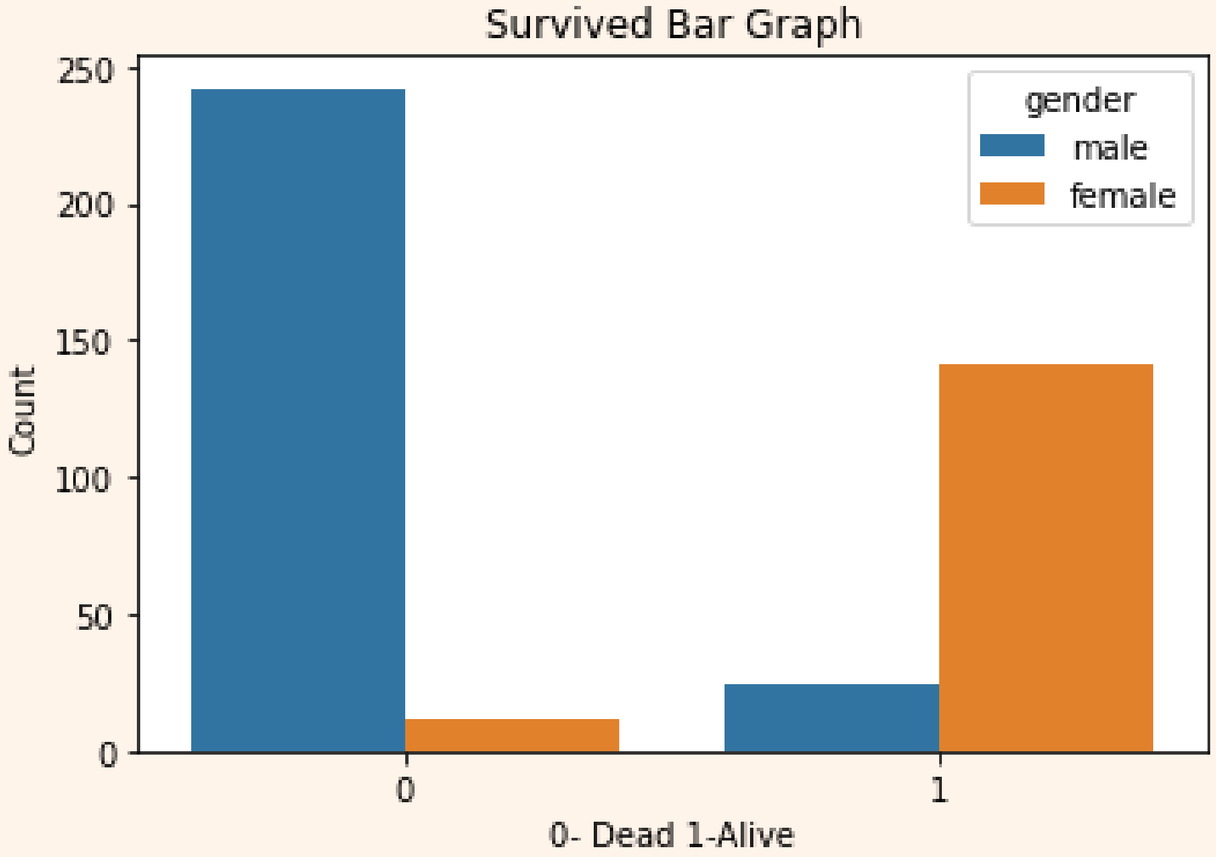




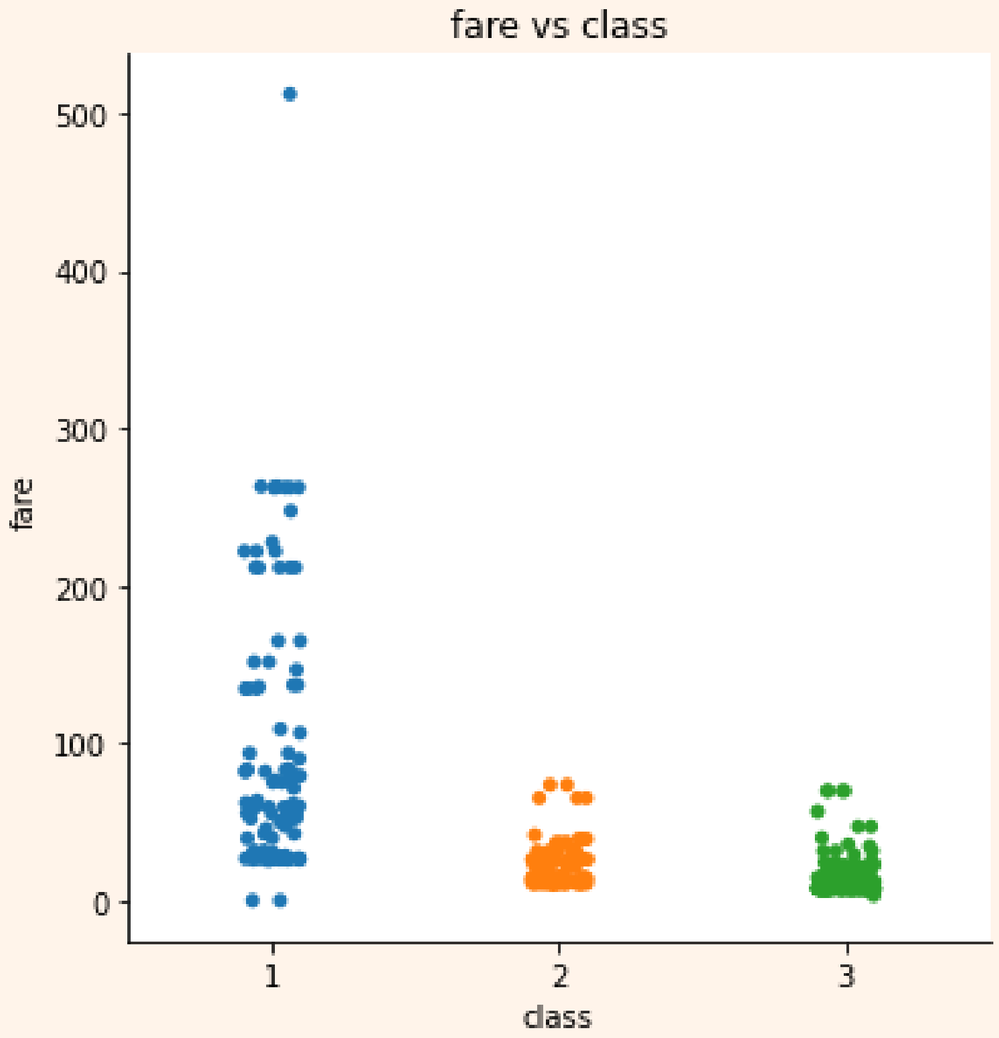
Distribution Graphs

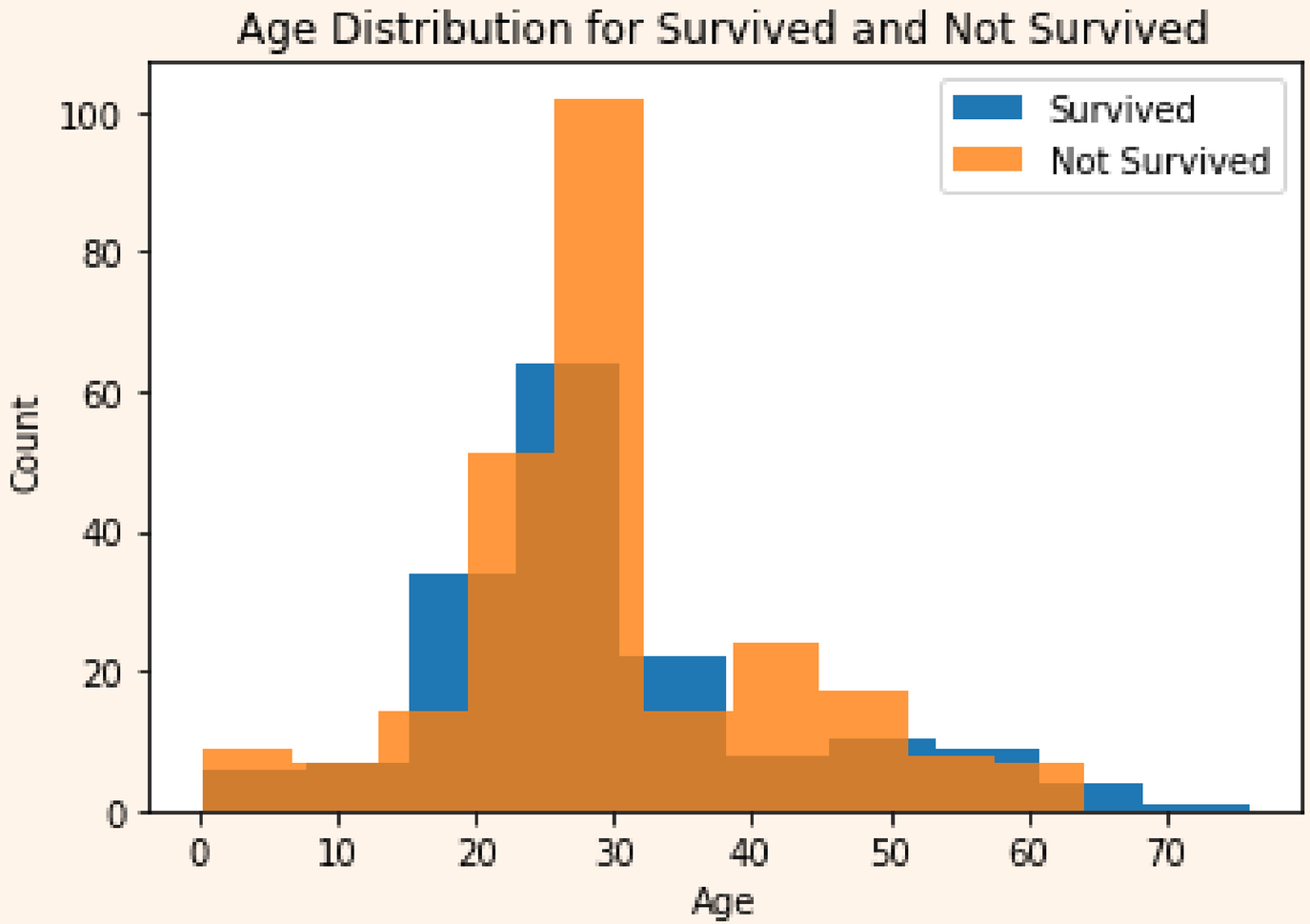




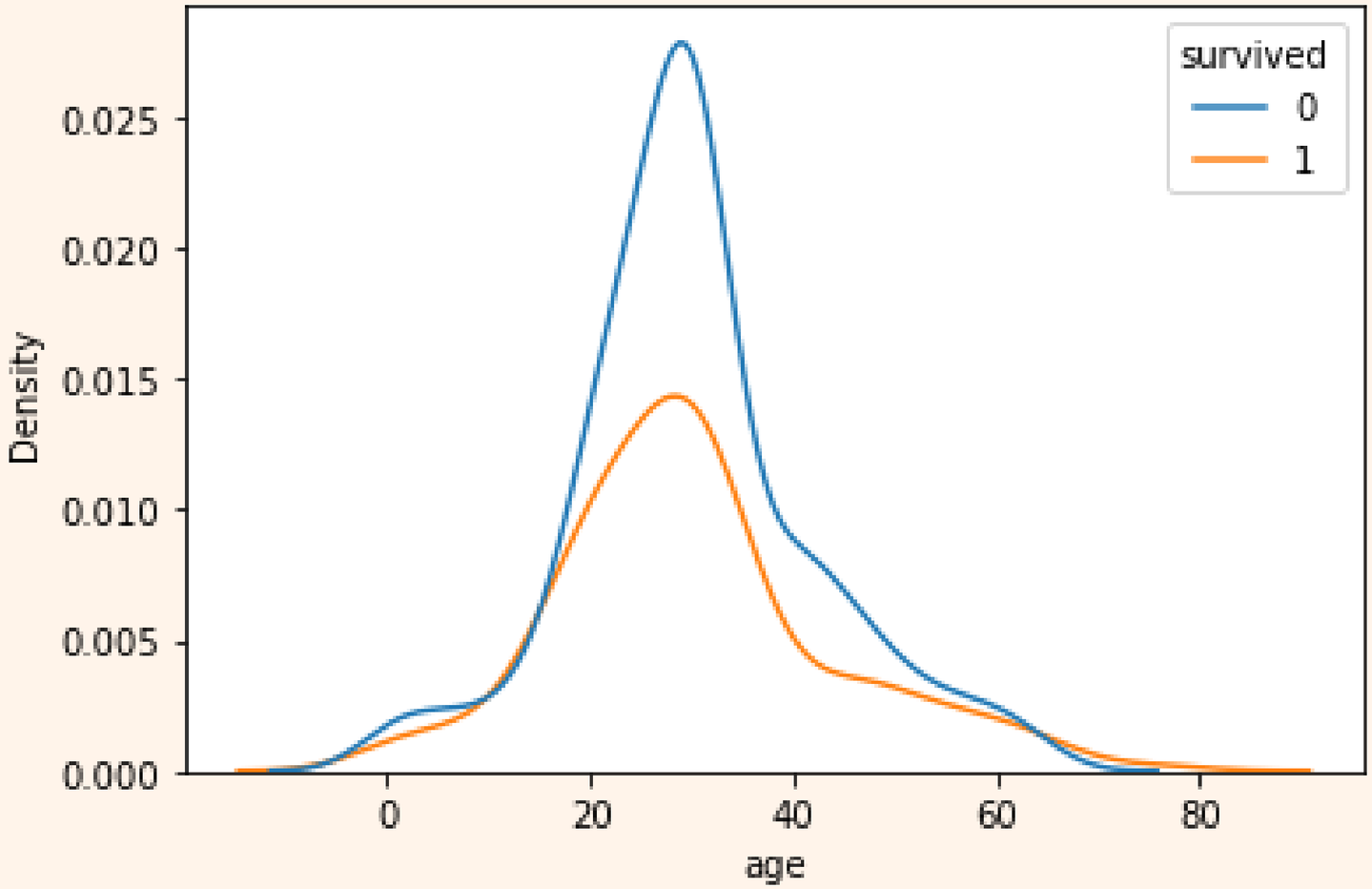


Corelation Graphs





Same age vs Survived but using sns



* Among males how many survived , among females how many survived?

```
g=df.groupby("gender")["survived"].value_counts()
g
```

```
gender  survived
female  1           141
        0            11
male    0           242
        1            24
Name: survived, dtype: int64
```

```
f=len(df[df['gender']=='female'])
m=len(df[df['gender']=='male'])
print(f"Percentage of female survivors = {g[0]/f*100}")
print(f"Percentage of female death = {g[1]/f*100}")
print(f"Percentage of male survivors = {g[-1]/m*100}")
print(f"Percentage of male death = {g[2]/m*100}")
```

```
Percentage of female survivors = 92.76315789473685
Percentage of female death = 7.236842105263158
Percentage of male survivors = 9.022556390977442
Percentage of male death = 90.97744360902256
```

Advantages

- Visualize the Titanic dataset to discover patterns, correlations.
- Can Apply Similar procedures to many other Datasets
- Learning Libraries to implement on real data sets.
- Can also be used later for predictions/machine learning.

Conclusion

- The survival rate for women and children was significantly higher than for men.
- Passengers in first,second class had a better chance of survival.
- Above conclusions were solely drawn due to visualization instead of studying 892 rows

References

https://www.kaggle.com/datasets/brendan45774/test-file

Thank you !