```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df1 = pd.read_csv('/content/The Titanic dataset.csv')
df2 = pd.read_csv('/content/Titanic Dataset.csv')
```

```
def dfinfo(df):
  datainf = {
      'dtp': df.dtypes,
      'cnt':df.count(),
      'unq': df.nunique(),
      'nul': df.isna().sum(),
      'dup': df.duplicated().sum()
      }
  res = pd.DataFrame(datainf)
  return res
```

```
df1
```

	1	2	3	4	5	6	7	8	9	10
0	sn	pclass	survived	NaN	gender	age	family	fare	embarked	date
1	1	3	0	Mr. Anthony	male	42	0	7.55	NaN	1-Jan-90
2	1	3	0	Mr. Anthony	male	42	0	7.55	NaN	1-Jan-90
3	2	3	0	Master. Eugene Joseph	male	?	2	20.25	S	2-Jan-90
4	3	2	0	Abbott, Mr. Rossmore Edward	NaN	NaN	2	**	S	3-Jan-90
...	...	...	...	...	...	...	...	...	...	...
1297	1296	2	0	Yrois, Miss. Henriette ("Mrs Harbeck")	female	24	0	13	S	19-Jul-93
1298	1297	3	0	Zabour, Miss. Hileni	female	14.5	1	14.4542	C	20-Jul-93
1299	1298	3	0	Zakarian, Mr. Mapriededer	male	26.5	0	7.225	C	21-Jul-93
1300	1299	3	0	Zakarian, Mr. Ortin	male	27	0	7.225	C	22-Jul-93
1301	1300	3	0	Zimmerman, Mr. Leo	male	29	0	7.875	S	23-Jul-93

1302 rows × 10 columns

```
header= df1.iloc[0]
df1 = df1.iloc[1:]
df1.columns = header
```

```
df1.columns = df1.columns.fillna('name')
```

```
dfinfo(df1)
```

0	dtp	cnt	unq	nul	dup
sn	object	1301	1300	0	1
pclass	object	1301	3	0	1
survived	object	1301	2	0	1
name	object	1301	1297	0	1
gender	object	1300	2	1	1
age	object	1044	97	257	1
family	object	1299	9	2	1
fare	object	1299	282	2	1
embarked	object	1295	3	6	1
date	object	1301	1300	0	1

```
df1.drop_duplicates(inplace = True)
```

```
<ipython-input-42-b8e706a2be12>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df1.drop_duplicates(inplace = True)
```

```
dfmrg = pd.merge(left = df1, right = df2, how = 'left', left_on = 'name', right_on = 'name')
```

```
dfinfo(dfmrg)
```

	dtp	cnt	unq	nul	dup
sn	object	1304	1300	0	0
pclass_x	object	1304	3	0	0
survived_x	object	1304	2	0	0
name	object	1304	1297	0	0
gender	object	1303	2	1	0
age_x	object	1047	97	257	0
family	object	1302	9	2	0
fare_x	object	1302	282	2	0
embarked_x	object	1299	3	5	0
date	object	1304	1300	0	0
pclass_y	float64	1301	3	3	0
survived_y	float64	1301	2	3	0
sex	object	1301	2	3	0
age_y	float64	1047	97	257	0
sibsp	float64	1301	7	3	0
parch	float64	1301	8	3	0
ticket	object	1301	921	3	0
fare_y	float64	1300	281	4	0
cabin	object	295	186	1009	0
embarked_y	object	1299	3	5	0
boat	object	485	27	819	0
body	float64	123	121	1181	0
home.dest	object	745	369	559	0

```
dfmrg.drop(columns = ['pclass_y', 'survived_y', 'age_y', 'fare_x', 'embarked_y', 'sex'], inplace = True)
```

```
dfmrg['gender'].unique()
```

```
array(['male', nan, 'female'], dtype=object)
```

```
np.where(dfmrg['gender'].isna())
```

```
(array([2]),)
```

```
dfmrg.drop(2, inplace = True)
```

```
dfmrg['age_x'].unique()
```

```
array(['42', '?', '35', '16', '25', '30', '28', '20', '18', '26', '40',
       '0.83', '24', '29', '0.92', '2', '32', '19', '48', '4', '6', '17',
       '38', '9', '11', '39', '27', '63', '34', '36', '53', '71', '57',
       '5', '3', '13', '23', '45', '21', '47', '33', '0.75', '80', '22',
       '51', nan, '50', '1', '12', '37', '58', '41', '15', '60', '44',
       '59', '18.5', '14', '54', '49', '76', '46', '52', '8', '31', '64',
       '70.5', '43', '55', '70', '22.5', '36.5', '65', '40.5', '10',
       '0.67', '23.5', '62', '7', '32.5', '34.5', '61', '20.5', '30.5',
       '55.5', '0.17', '28.5', '45.5', '56', '38.5', '14.5', '24.5',
       '60.5', '67', '74', '11.5', '66', '26.5'], dtype=object)
```

```
dfmrg["age_x"] = dfmrg["age_x"].replace(["?"], None)
dfmrg['age_x'] = dfmrg['age_x'].astype(float)
meanage = dfmrg['age_x'].mean()
dfmrg['age_x'] = dfmrg['age_x'].fillna(meanage)
```

```
dfmrg['family'].nunique()
```

```
9
```

```
modefamily = dfmrg['family'].mode()[0]
modefamily
```

```
'0'
```

```
dfmrg['family'] = dfmrg['family'].fillna(modefamily)
```

```
dfinfo(dfmrg)
```

	dtp	cnt	unq	nul	dup
sn	object	1303	1299	0	0
pclass_x	object	1303	3	0	0
survived_x	object	1303	2	0	0
name	object	1303	1297	0	0
gender	object	1303	2	0	0
age_x	float64	1303	97	0	0
family	object	1303	9	0	0
embarked_x	object	1298	3	5	0
date	object	1303	1299	0	0
sibsp	float64	1300	7	3	0
parch	float64	1300	8	3	0
ticket	object	1300	921	3	0
fare_y	float64	1299	281	4	0
cabin	object	295	186	1008	0
boat	object	485	27	818	0
body	float64	122	121	1181	0
home.dest	object	744	369	559	0

```
dfmrg['embarked_x'].unique()
```

```
array([nan, 'S', 'C', 'Q'], dtype=object)
```

```
modeembarked_x = dfmrg['embarked_x'].mode()[0]
dfmrg['embarked_x'] = dfmrg['embarked_x'].fillna(modeembarked_x)
```

```
dfmrg['sibsp'].unique()
```

```
array([nan,  1.,  0.,  4.,  2.,  3.,  5.,  8.])
```

```
dfmrg['sibsp'] = dfmrg['sibsp'].astype(float)
modesibsp = dfmrg['sibsp'].mode()[0]
#Using mode instead of mean as only 3 null values
```

```
dfmrg['sibsp'] = dfmrg['sibsp'].fillna(modesibsp)
```

```
dfmrg['parch'].unique()
```

```
array([nan,  1.,  0.,  2.,  5.,  3.,  4.,  6.,  9.])
```

```
modeparch = dfmrg['parch'].mode()[0]
#Using mode instead of mean as only 3 null values
dfmrg['parch'] = dfmrg['parch'].astype(float)
dfmrg['parch'] = dfmrg['parch'].fillna(modesibsp)
```

```
dfmrg['ticket'] = dfmrg['ticket'].fillna('Not_Known')
dfmrg['cabin'] = dfmrg['cabin'].fillna("Not_Known")
```

```
modefare_y = dfmrg['fare_y'].mode()[0]
dfmrg['fare_y'] = dfmrg['fare_y'].fillna(modefare_y)
```

```
dfmrg['boat'].unique()
```

```
array([nan, '16', 'A', '10', '15', 'C', '11', '13', '2', '3', 'D', '4',
       '9', '6', 'B', '8', '5', '12', '7', '14', '13 15 B', '5 9', '1',
       'C D', '15 16', '5 7', '13 15', '8 10'], dtype=object)
```

```
dfmrg['boat'] = dfmrg['boat'].fillna("Not_Known")
```

```
dfmrg['body'] = dfmrg['body'].fillna(0)
```

```
dfmrg['home.dest'] = dfmrg['home.dest'].fillna("Not_Known")
```

```
dfmrg['family'] = dfmrg['family'].astype(int)
dfmrg['age_x'] = dfmrg['age_x'].astype(int)
dfmrg['date'] = pd.to_datetime(dfmrg['date'])
dfmrg['sibsp'] = dfmrg['sibsp'].astype(int)
dfmrg['parch'] = dfmrg['parch'].astype(int)
dfmrg['body'] = dfmrg['body'].astype(int)
```

```
<ipython-input-68-8992c5047f84>:3: UserWarning: Could not infer format, so each element will be parsed individually, falling back to `dateutil`. To ens
  dfmrg['date'] = pd.to_datetime(dfmrg['date'])
```

```
dfmrg2 = dfmrg.rename(columns =
 {'sn': 'Serial Number',
  'pclass_x': 'Passenger Class',
  'survived_x': 'Survived',
  'name': 'Name',
  'gender': 'Gender',
  'age_x': 'Age',
  'family': 'Family',
  'embarked_x': 'Embarked Port',
  'date': 'Date',
  'sibsp': 'Siblings Or Spouces',
  'parch': 'Parent Or Children',
  'ticket': 'Ticket Number',
  'fare_y': 'Fare',
  'cabin': 'Cabin Number',
  'boat': 'Lifeboat Number',
  'body': 'Body Number',
  'home.dest': 'Home Destination'})
```

```
dfinfo(dfmrg2)
```

	dtp	cnt	unq	nul	dup
Serial Number	object	1303	1299	0	0
Passenger Class	object	1303	3	0	0
Survived	object	1303	2	0	0
Name	object	1303	1297	0	0
Gender	object	1303	2	0	0
Age	int64	1303	73	0	0
Family	int64	1303	9	0	0
Embarked Port	object	1303	3	0	0
Date	datetime64[ns]	1303	1299	0	0
Siblings Or Spouces	int64	1303	7	0	0
Parent Or Children	int64	1303	8	0	0
Ticket Number	object	1303	922	0	0
Fare	float64	1303	281	0	0
Cabin Number	object	1303	187	0	0
Lifeboat Number	object	1303	28	0	0
Body Number	int64	1303	122	0	0
Home Destination	object	1303	370	0	0