1 – تابع describe، تعداد اعضای یک ستون که NaN نیستند، میانگین، انحراف معیار، کمینه داده ها، چارک اول، چارک دوم (میانه)، چارک سوم و بیشنیه داده ها را نشان میدهد :

```
df.describe() output is :
        PassengerId    Survived      Pclass         Age       SibSp       Parch        Fare
count   891.000000  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std     257.353842    0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min       1.000000    0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%     446.000000    0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%     668.500000    1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

تابع tail، اطلاعات 5 ردیف آخر را به شکل پیش فرض نشان میدهد ولی میتوان با پاس دادن متغیر n به آن تعداد بیشتری ردیف از آخر ببینیم مثلا اگر n = 10 باشد، (df.tail(10)) اطلاعات 10 ردیف آخر را به ما نشان میدهد :

```
df.tail() output is :
     PassengerId  Survived  Pclass                                      Name     Sex   Age  SibSp  Parch      Ticket   Fare Cabin Embarked
886          887         0       2                     Montvila, Rev. Juozas    male  27.0      0      0      211536  13.00   NaN        S
887          888         1       1              Graham, Miss. Margaret Edith  female  19.0      0      0      112053  30.00   B42        S
888          889         0       3  Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1      2  W./C. 6607  23.45   NaN        S
889          890         1       1                     Behr, Mr. Karl Howell    male  26.0      0      0      111369  30.00  C148        C
890          891         0       3                       Dooley, Mr. Patrick    male  32.0      0      0      370376   7.75   NaN        Q
```

تابع head، اطلاعات 5 ردیف اول را به شکل پیش فرض نشان میدهد ولی میتوان با پاس دادن متغیر n به آن تعداد بیشتری ردیف از اول ببینیم مثلا اگر n = 10 باشد، (df.head(10)) اطلاعات 10 ردیف اول را به ما نشان میدهد :

```
df.head() output is :
   PassengerId  Survived  Pclass                                               Name     Sex   Age  SibSp  Parch            Ticket     Fare Cabin Embarked
0            1         0       3                            Braund, Mr. Owen Harris    male  22.0      1      0         A/5 21171   7.2500   NaN        S
1            2         1       1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1      0          PC 17599  71.2833   C85        C
2            3         1       3                             Heikkinen, Miss. Laina  female  26.0      0      0  STON/O2. 3101282   7.9250   NaN        S
3            4         1       1       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1      0            113803  53.1000  C123        S
4            5         0       3                           Allen, Mr. William Henry    male  35.0      0      0            373450   8.0500   NaN        S
```

تابع info، تعداد اعضای یک ستون که NaN نیستند، نوع اعضای یک ستون و حافظه ای که دیتا مصرف میکند را نشان میدهد :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 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   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

2 – نوع داده های هر ستون در شکل زیر نشان داده شده اند که داده هایی که نوع آن ها object است، داده های غیر عددی و دیگر داده ها، داده های عددی اند.

```
Data types are :
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
```

3 – تعداد سطرهای خالی در هر ستون :

```
Number of blank rows for each column :
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         0
```

مزایای این روش عبارتند از :

- از از دست رفتن داده ها (data loss) جلوگیری میکند که data loss باعث حذف ستون ها یا ردیف ها میشود.
- با مجموعه دیتا های کوچک به خوبی کار میکند و پیاده سازی ساده ای دارد.

معایب این روش عبارتند از :

- فقط با متغیر های عددی پیوسته کار میکند.
- میتواند باعث نشت داده (data leakage) شود.
- کوواریانس بین ویژگی ها را فاکتور نمیگیرد.

4 – ستون های PassengerId، Ticket و Name را حذف کردم.

5 – تعداد مسافران با ویژگی های خواسته شده :

```
Men count = 577
Women count = 314
Men count that embarked at Southhampton = 441
```

6 – تعداد مسافران با ویژگی خواسته شده :

```
Number of passengers whoe are older than 35 and have no compeer and their ticket type is 3 = 41
```

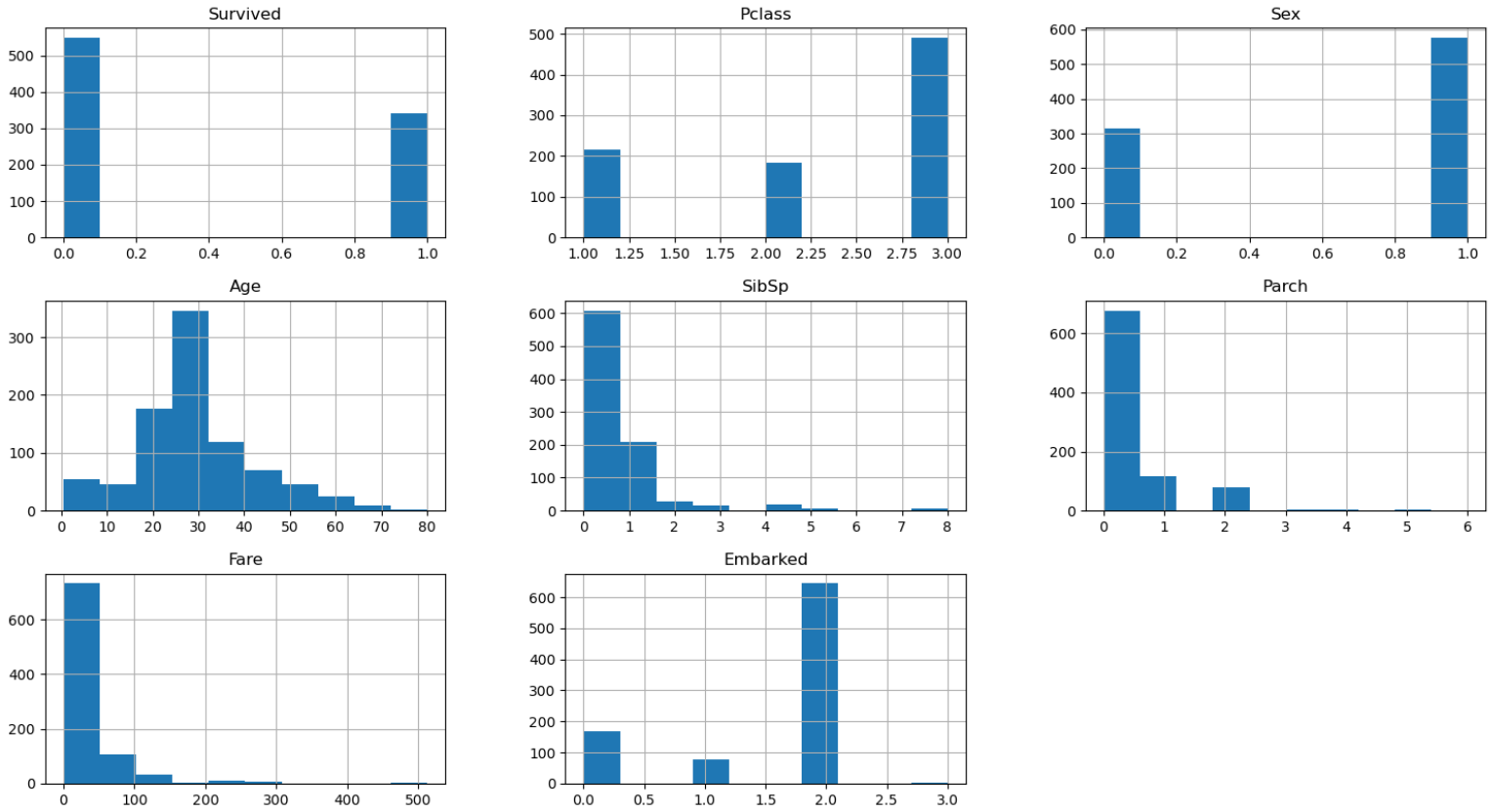
7 – میانگین کرایه بلیط مسافران با ویژگی خواسته شده :

```
Average fare for travelers who embarked at Queenstown = 13.276029870129872
```

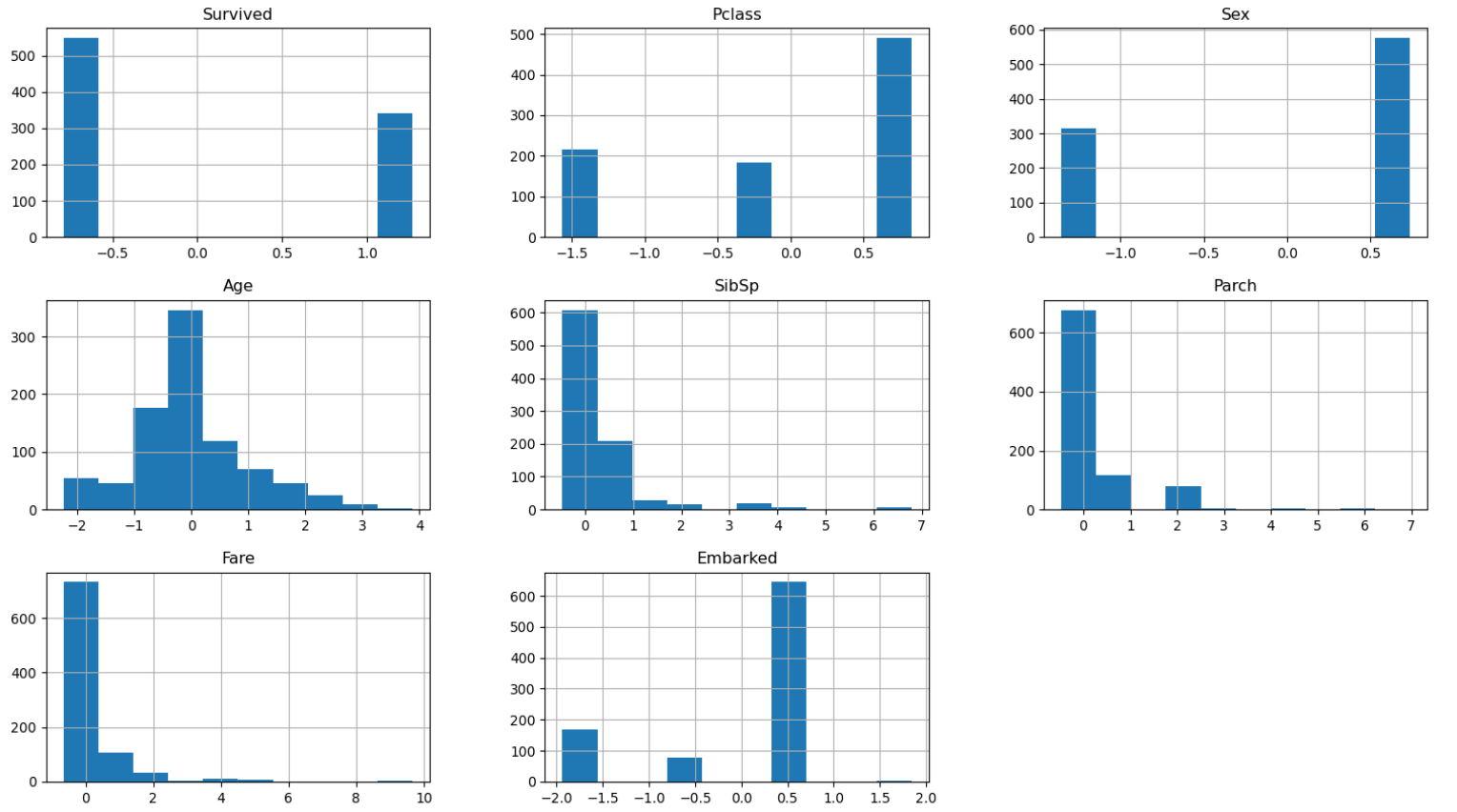
8 – زمان اجرای محاسبات بخش قبلی با استفاده از loop به جای استفاده از vectorization حدود 6.6 برابر است.

```
average_fare = 13.276029870129872
The time spent using for loop = 0.0031728744506835938
average_fare = 13.276029870129872
The time spent using vectorization = 0.0004801750183105469
```

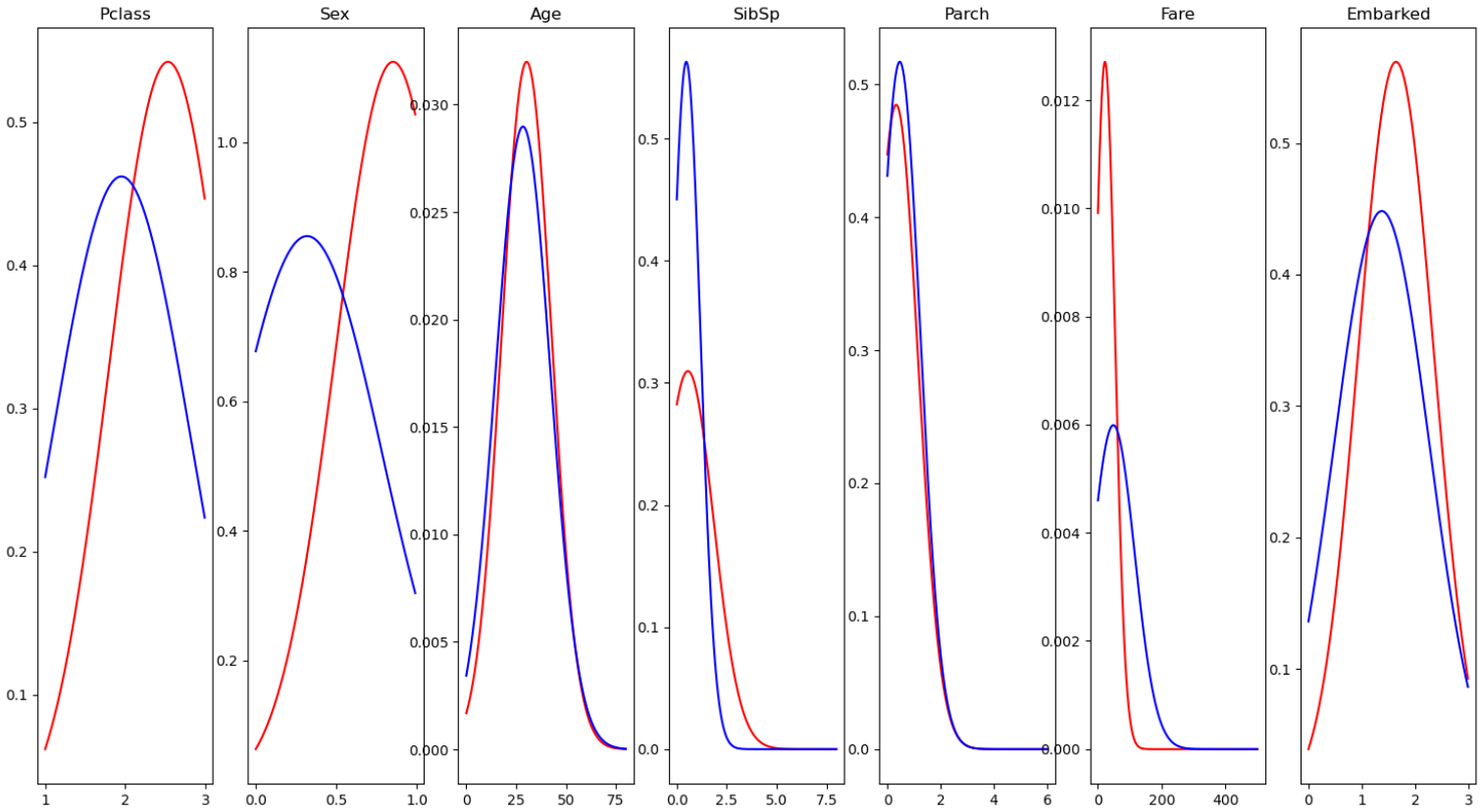
9 – شکل توزیع هر ستون از داده :



10 – شکل توزیع هر ستون از داده بعد از نرمال سازی :



11 – نمودار های تابع چگالی احتمال توزیع نرمال ویژگی های مختلف (رنگ قرمز برای افرادی است که مرده اند و رنگ آبی برای افرادی است که زنده مانده اند.) :



برای مدل کردن ورودی و پیش بینی این که یک فرد زنده میماند یا خیر سه ویژگی Pclass، Sex و Embarked را در نظر گرفتم که علت انتخابم این بود که سه ویژگی مذکور، پراکندگی بیشتری دارند و نمودار های شان برای افرادی که مرده اند و زنده مانده اند ، تفاوت خوبی دارد و از هم قابل تشخیص تر است. روش پیش بینی زنده ماندن یا نماندنی هم که به کار گرفتم به این شکل است که برای هر فرد احتمال زنده ماندن و احتمال مردنش بر اساس هر ویژگی را محاسبه میکنم و احتمالات به دست آمده را در هم ضرب میکنم. (چون از هم مستقلند) در ادامه هر احتمالی بیشتر شد ، همان را به عنوان پاسخ پیش بینی در نظر میگیرم.