باسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

درس یادگیری ماشین

گزارش تمرین شماره سه

علی محرابیان96102331

استاد: دکتر صالح کلیبر

پاییز 1398

ابتدا ستون هایی که دارای مقادیر NaN هستند را می یابیم.

```
Embarked has 2 NaN values-> 0.22446689113355783%
Age has 177 NaN values-> 19.865319865319865%
Cabin has 687 NaN values-> 77.10437710437711%
```

ویژگی های cabin و age بیشترین درصد را دارا می باشند.

حال تکرار title های موجود در در ستون name را می یابیم.

```
we have 517 Mr title
we have 182 Miss title
we have 127 Mrs title
we have 40 Master title
we have 7 Dr title
we have 6 Rev title
we have 2 Col title
```

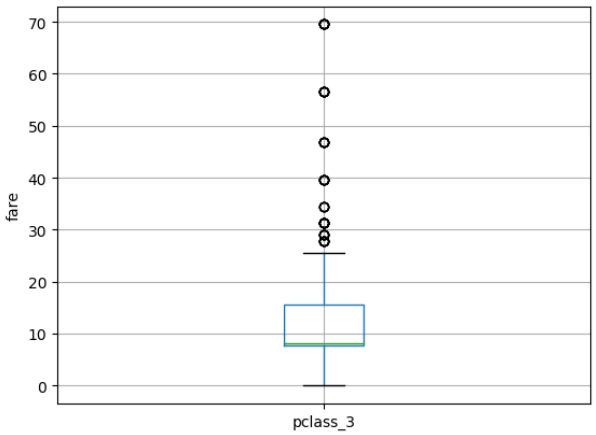
البته title های دیگری هم موجود بودند ولی تکرر آن ها بسیار کم بود.حال این title هارا به سه دسته تقسیم می کنیم.به title Mr عدد 1،به Missو Mrs عدد 2 و به بقیه عدد 3 را نسبت می دهیم.

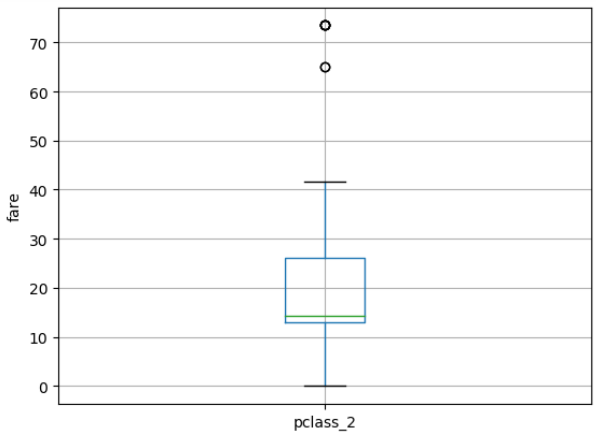
مقادیر NaN در ستون age را با میانگین خانه هایی که title یکسان با خانه NaN دارند پر می کنیم.

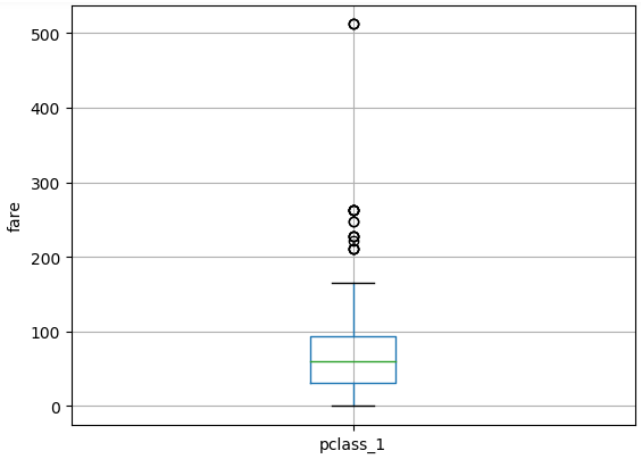
مقادیر این میانگین ها به صورت زیر هستند.

```
for Mr title,mean is 24.91779497098646
for Mrss title,mean is 23.067961165048544
for others title,mean is 18.379538461538463
```

در ابتدا boxplot pclass های مختلف را برای fare می کشیم.







پراکندگی برای کلاس های مختلف در تصاویر بالا گویا است.برخی از مقادیر ستون fare،مقدار 0 دارند.احتمالا مقادیر این خانه ها مربوط به خدمه کشتی باشد.جای این مقادیر،میانگین fare هایی که با آن خانه در کلاس بلیط یکسان هستند را می گذاریم.

برای متغیر embarked،از one hot encoding استفاده می کنیم.

	Embarked1	Embarked2	Embarked3	Embarked4
S	1	0	0	0
C	0	1	0	0
Q	0	0	1	0
NaN	0	0	0	1

دلیل آن که از اعداد ترتیبی استفاده نکردیم این است که ممکن است در الگوریتم Learning،با این اعداد شبیه با وزن رفتار شود در این صورت ممکن است به موردی که اهمیت کمتری دارد،توجه بیشتری شود.

آمارگان برای feature های ذکر شده به صورت زیر است.

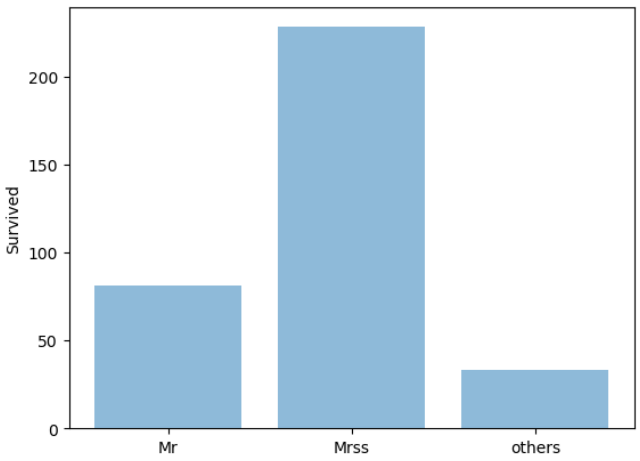
	Features	mean	mean_Nsurvived	mean_survived	median	median_Nsurvived	median_survived
0	age	0 28.607374 dtype: float64	29.244991	27.583830	0	25.0	25.0
1	sex	0 0.647587 dtype: float64	0.852459	0.318713	0	1.0	0.0
2	sibsp	0.523008	0.553734	0.473684	0	0.0	0.0
3	parch	0.381594	0.329690	0.464912	0	0.0	0.0

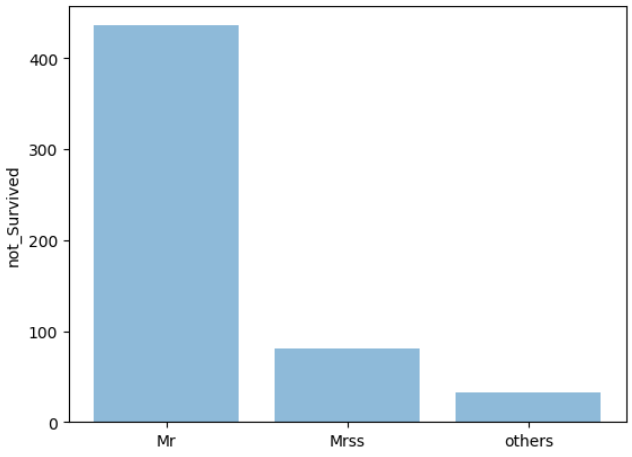
mode	mode_Nsurvived	mode_survived	var	var_Nsurvived	var_survived
0 0 25.0	25.0	23.0	0 174.253103 dtype: float64	161.570672	192.401962
0 0 1.0	1.0	0.0	0 0.228475 dtype: float64	0.125773	0.217135
0 0 dtype: int64	0.0	0.0	1.21604	1.656949	0.500769
0 0 dtype: int64	0.0	0.0	0.649728	0.676368	0.593798

برای متغیر age مشاهده می شود که واریانس زیاد است که نشان از پراکندگی زیاد سن افراد دارد.برای متغیر Sex،میانه و مد برابر با 1 بوده که نشان می دهد بیشتر افراد مرد بوده اند.در این متغیر مشاهده می شود که میانگین افرادی که مرده اند،به 1 نزدیک تر است که حاکی از آن است که زن ها بیشتر زنده مانده اند.

برای متغیر های sibsp و parch،میانه و مد برابر با صفر بوده که نشان می دهد بیشتر افراد دارای قوم و خویش کمتری بوده اند.

نمودار های میله ای به صورت زیر است.

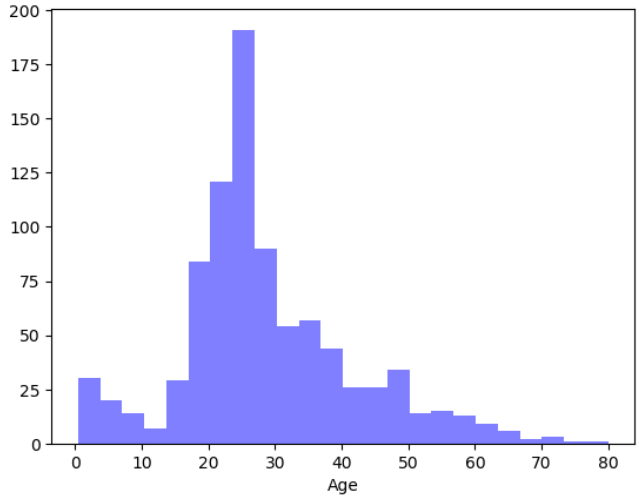


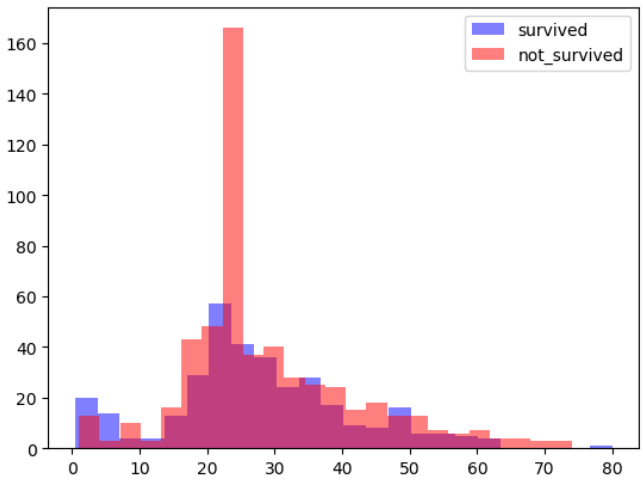


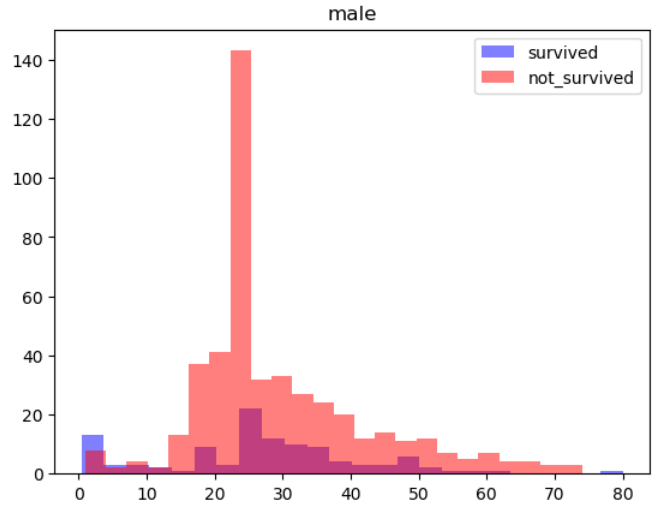
همان طور که مشاهده می شود،تعداد بیشتری از زن ها زنده مانده اند و تعداد بیشتری از مردها مرده اند.

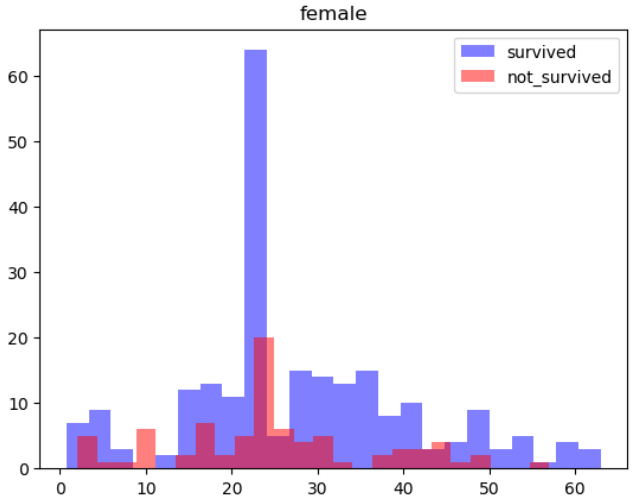
به کمک قانون sturge،تعداد دسته های مطلوب برای نمودار هیستوگرام از رابطه زیر به دست می آید.

$$K = 1 + 3.322log_{10}^{N}$$



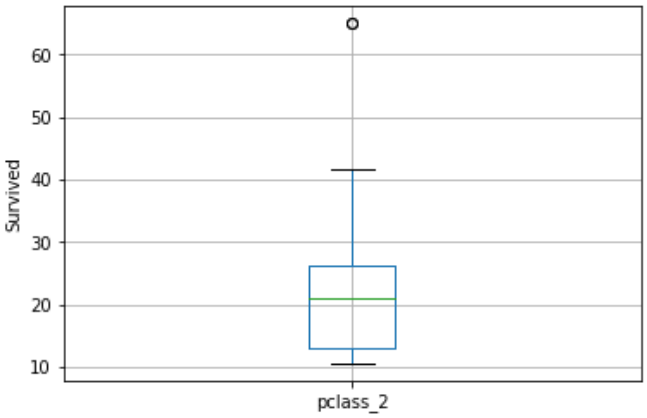


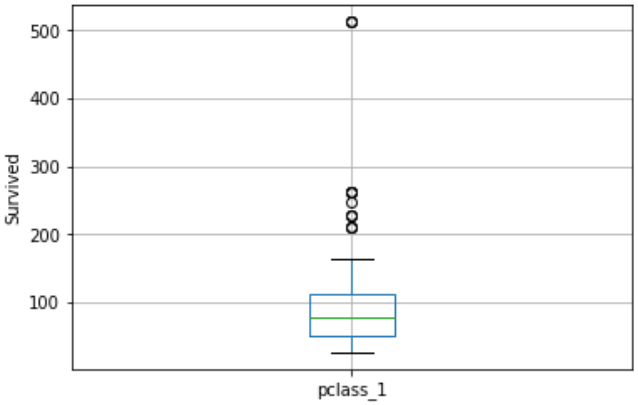


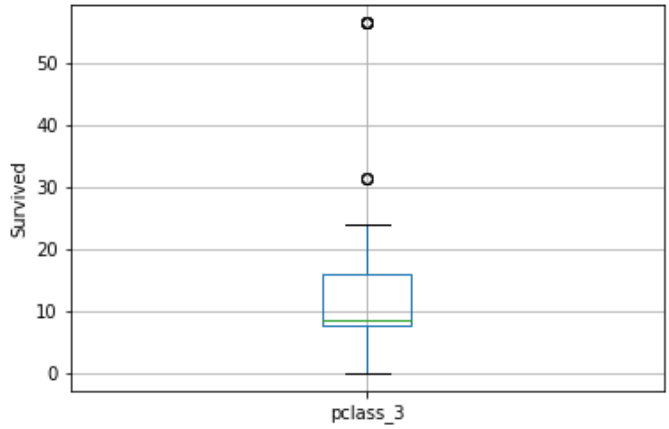


مردهای کوچکتر از 10 سال و بین 25 تا 35 سال شانس بیشتری برای زنده ماندن دارند.زن های بین 15 تا 40 سال هم شانس بیشتری برای زنده ماندن دارند.به طور کلی می توان دید که زن ها شانس بیشتری برای زنده ماندن دارند.

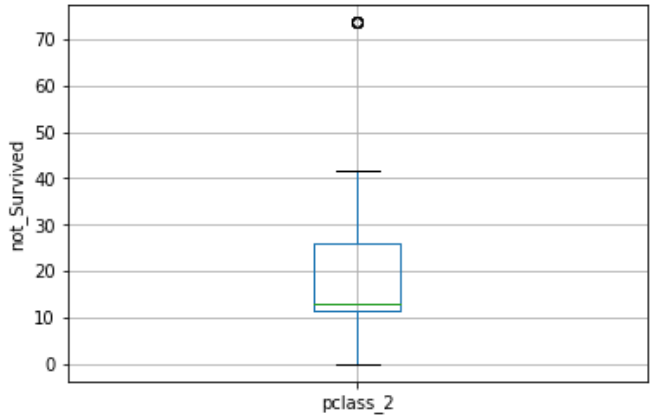
حال نمودار های boxplot های خواسته شده را رسم می کنیم.برای افراد زنده مانده به صورت زیر است.

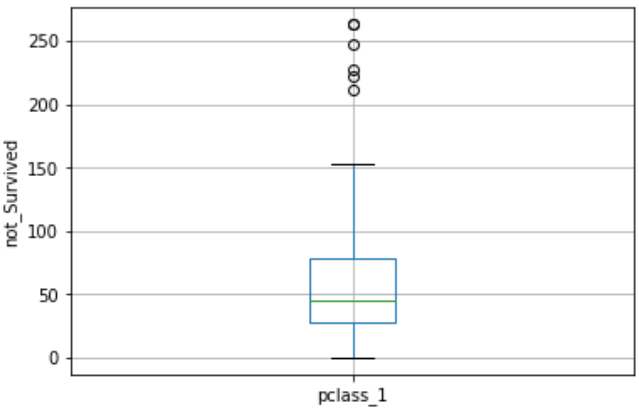


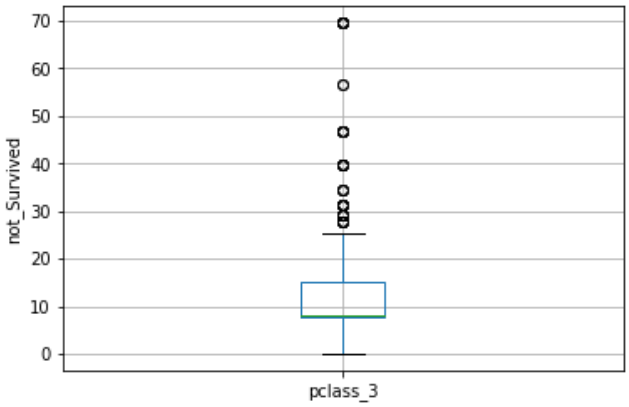




برای افرادی که زنده نماندند،نمودار ها به صورت زیر است.

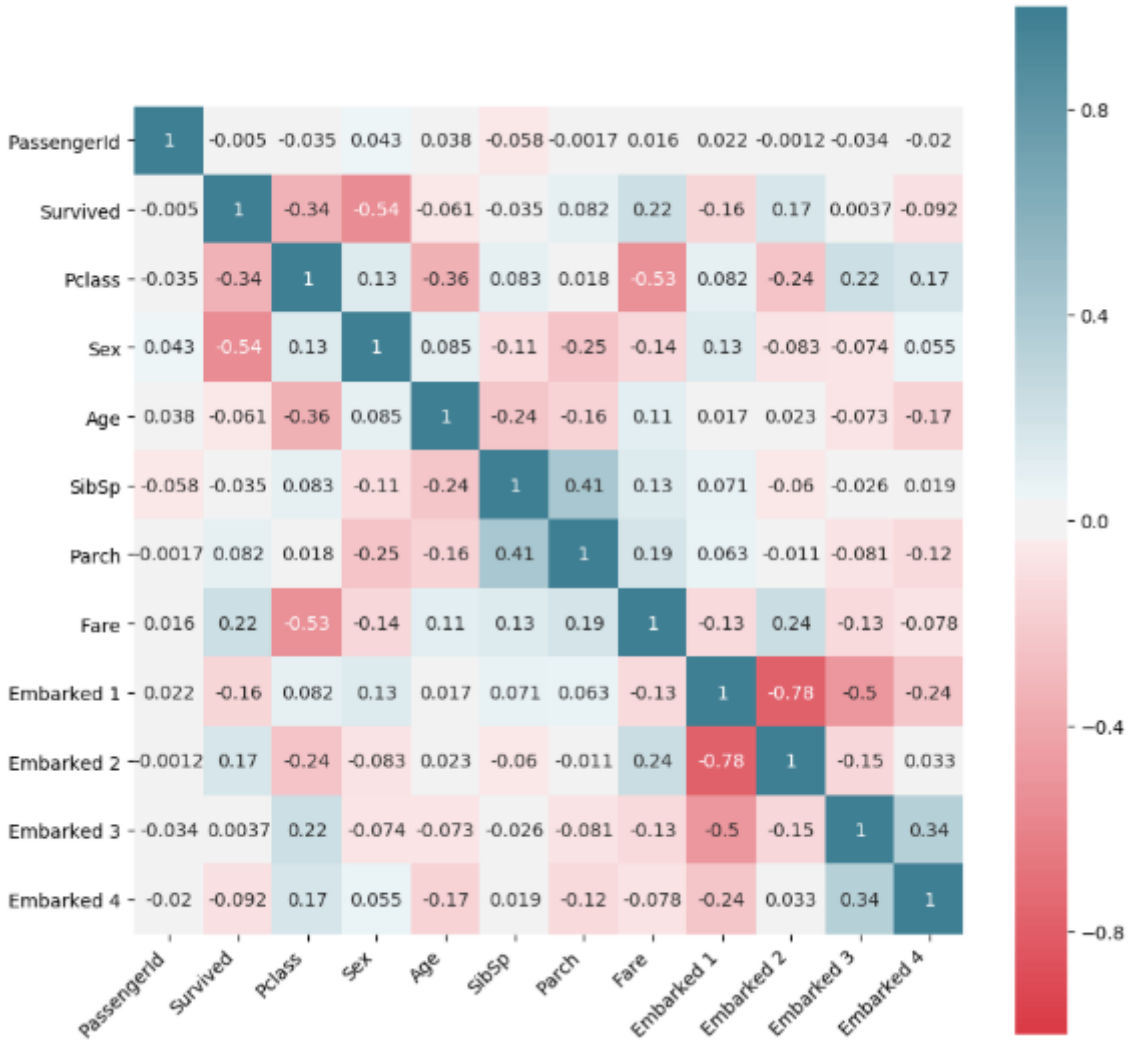






در حالت زنده ماندن با توجه به این که میانه pclass=1 از بقیه بالاتر است،پس می توان فهمید که شانس زنده ماندن برای افراد صاحب این کلاس بیشتر است.

نمودار heatmap به همراه correlation به صورت زیر است.



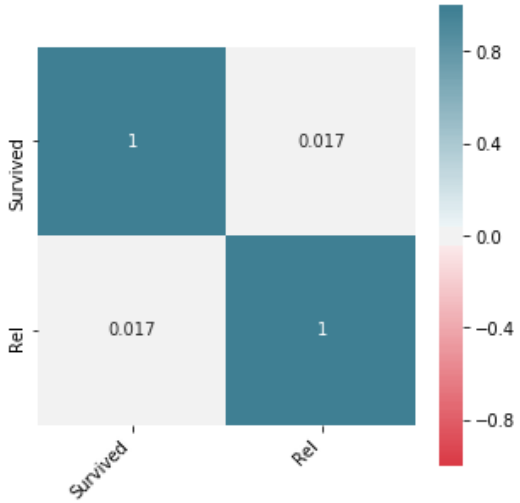
پر رنگ تر بودن خانه ها به این معنی است که داده ها همبستگی زیادی دارند.منفی بودن به این معنی است که دو متغیر رابطه عکس دارند.به طور مثال دو متغیر sex و survival،همبستگی زیادی با مقدار منفی دارند که نشان می دهد زن ها بیشتر زنده مانده اند.متغیر survived که بیشتر مدنظر ما است،با embarked 2 و مقدار fare رابطه بیشتری دارد.بعضی متغیر ها مانند parch نیز شاید به تنهایی قابل صرف نظر کردن باشند.

با ترکیب دو ویژگی age و sex،ویژگی kind را ایجاد می کنیم.به افرادی که سن کمتر از 10 سال دارند را در دسته child در نظر می گیریم.بقیه افراد را با توجه به جنسیت آن ها به دسته های man و woman تقسیم می کنیم.چون دوباره به داده categorical برخورد می کنیم،به روش one hot encoding آن را عددی می کنیم.

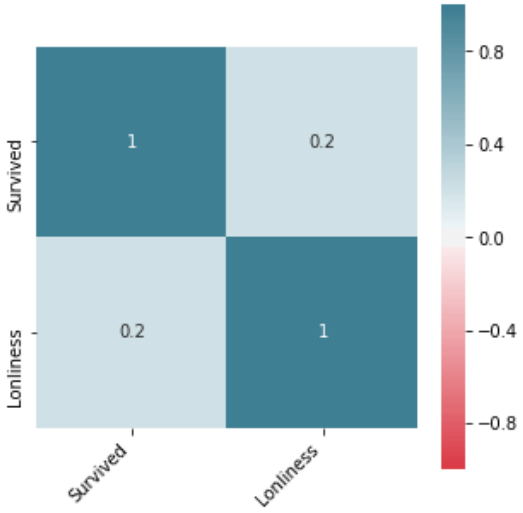
	Kind1	Kind2	Kind3
Child	0	0	1
Woman	0	1	0
Man	1	0	0

برای کابین ها،متغیر deck را تعریف می کنیم و به حروف از A تا G،اعداد 1 تا 7 را نسبت می دهیم.برای NaN ها هم حرف U را که متناظر با عدد 8 است،در نظر می گیریم.حال مکان حروف را در cabin پیدا کرده و به اعداد map می کنیم.ولی به نظر می آید که بهتر است اثر متغیر کابین حذف شود.

متغیر Rel را از جمع کردن دو متغیر sibsp و parch ایجاد می کنیم.heatmap آن با متغیر survival به صورت زیر است.



متغیر Lonliness را این گونه تعریف می کنیم که اگر Rel مقدار بزرگتر از صفر داشت،به آن 1 نسبت می دهیم اگر مقدار آن برابر صفر بود،به آن صفر نسبت می دهیم.heatmap آن با survival به صورت زیر است.



می توان دید که همبستگی خوبی بین دو متغیر برقرار است.