به نام خدا

تمرین سری سوم کامپیوتری - یادگیری ماشین

دانشجو : مهرسا پوریا (95101247)

شروع کار با دادگان

1. همانگونه که در زیر مشاهده میکنید، ویژگی های کابین، سن، محل سوار شدن و قیمت بلیط به ترتیب بیشترین درصد خانه خالی را دارند.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Number of Nulls in Train Data	0	0	0	0	0	177	0	0	0	0	687	2

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Percentage of Nulls in Train Data	0.0	0.0	0.0	0.0	0.0	19.86532	0.0	0.0	0.0	0.0	77.104377	0.224467

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Number of Nulls in Test Data	0	0	0	0	86	0	0	0	1	327	0

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Percentage of Nulls in Test Data	0.0	0.0	0.0	0.0	20.574163	0.0	0.0	0.0	0.239234	78.229665	0.0

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Percentage of Nulls in Test Data	0.0	0.0	0.0	0.0	20.574163	0.0	0.0	0.0	0.239234	78.229665	0.0

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Percentage of Nulls in mergedData	0.0	0.0	0.0	0.0	20.091673	0.0	0.0	0.0	0.076394	77.463713	0.152788

2. به کمک جدا کردن اسم این کار را انجام داده و ستون Title را افزودیم.
3. همانگونه که در پیاتزای درس مطرح شده است اگر عنوان ها را 3 دسته بگیریم کار چندان درستی نمیشود چون برای مثال اعضای دو دسته ی Miss و Mrs با هم اختلاف سن دارند و بهتر است که جدا در نظر گرفته شوند. بنابراین القاب را 5 دسته (Miss, Mrs, Mr, Maste, Other) میکنیم و تحت عنوان myTitle به ستون های dataFrame داده اضافه می‌کنیم. که تعداد هر لقب قبل و بعد تغییرات در جداول زیر قابل مشاهده است.

	Mr	Miss	Mrs	Master	Dr	Rev	Mlle	Major	Col	Mme	Don	Sir	the Countess	Lady	Capt	Ms	Jonkheer
Number of Each Title in Train Data	517	182	125	40	7	6	2	2	2	1	1	1	1	1	1	1	1

	Mr	Miss	Mrs	Master	Col	Rev	Dona	Dr	Ms
Number of Each Title in Test Data	240	78	72	21	2	2	1	1	1

	Mr	Miss	Mrs	Master	Other
Number of Each newTitle in Train Data	517	184	127	40	23

	Mr	Miss	Mrs	Master	Other
Number of Each newTitle in Test Data	240	78	73	21	6

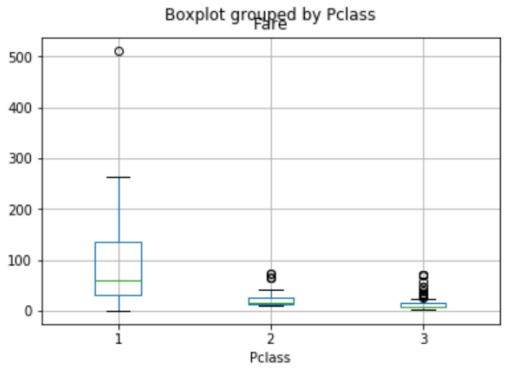
4. از آنجا که داده تست نیز بیانگر ارتباط بین سن و لقب است داده تست و آموزش را با هم ترکیب میکنیم و ستونی تحت عنوان myAge به داده می‌افزاییم که در آن سن‌های خالی را با میانگین هر دسته لقبی‌اش پر کرده ایم. میانگین سنهای جایگذاری شده برای هر دسته در جدول زیر ملاحضه میشود :

mean of Age in each myTitle for merged data:

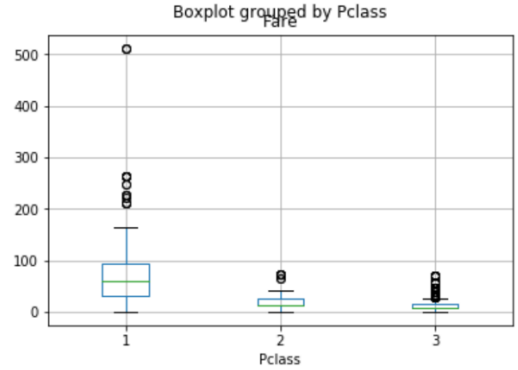
myTitle	Master	Miss	Mr	Mrs	Other
Age	5.482642	21.795236	32.252151	36.866279	45.178571

5. Boxplot خواسته شده برای داده آموزش، تست و ادغام این دو رسم شده است و همچنین ستونی تحت عنوان myFare به داده افزوده شده که خانه خالی را با توجه به میانگین Fare در Pclass مربوط آن در داده ادغامی پر کرده ایم، میانگین ها نیز در جداول زیر گزارش شده است.

Test Data:



Train Data :



Train Data :

Pclass	1	2	3
Fare	84.154687	20.662183	13.67555

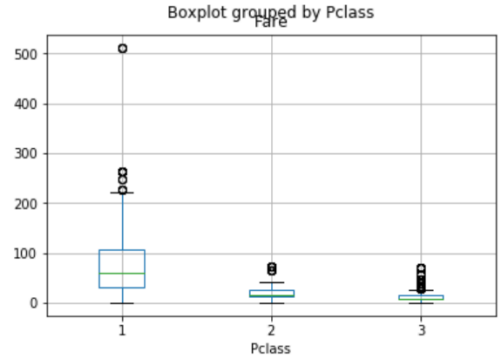
Test Data :

Pclass	1	2	3
Fare	94.280297	22.202104	12.459678

Merged Data:

Pclass	1	2	3
Fare	87.508992	21.179196	13.302889

Merged Data:



6. مشکل نسبت دادن 0، 1، 2، 3 آن است که باعث برداشت نابه‌جا از داده categorical می‌شود و ترتیب آن می تواند غلط انداز باشد برای مثال 2 برابر بودن عدد نسبت داده شده به یک دسته نسبت به عدد دیگر هیچ مفهومی در واقع ندارد اما در دیدگاه عددی میتواند داشته باشد و همین در روش‌های یادگیری مشکل ایجاد میکند. به همین علت از indicator ها استفاده می‌کنیم که به ازای K گروه داده categorical،K-1 بردارد باینری می سازیم که یک بودن هر معادل وقوع هرکدام از دسته ها باشد. (در متن تمرین گفته شده برای 4 حالت از 2 ستون استفاده کنیم که چنین چیزی باز هم مشکل دارد، زیرا در حالت 1 1 که معادل حالت سوم است نمیتوان تاثیر هر دسته را مستقلا اثر داد و به بیان دیگر اینکار یکی از درجه آزادی‌ها را از بین می‌برد.)

برای ویژگی Embarked خانه‌های خالی را با ’S‘ که بیشترین تکرار را دارد پر می‌کنیم. حال با سه دسته روبه‌رو خواهیم بود که 3 بردار باینری تعریف میکنیم که یک بودن هریک برابر یک گروه آن است. (البته 2 ستون هم کافیست ولی برای بهتر بودن کوریلشن ها در قسمت های بعد سه دسته لحاظ میکنیم.)

در این بخش سه ویژگی Embarked و Sex و Title را کمی میکنیم و در دیتافریم های جدیدی نگه میداریم. ویژگی Ticket با توجه به تعداد دسته های زیاد آن بهتر است در قسمت بعد حذف شود و به کمی سازی ویژگی Cabin در سوالهای بعد میپردازیم.

```
Embarked Groups: ['S' 'C' 'Q']
Sex groups : ['male' 'female']
myTitle groups : ['Mr' 'Mrs' 'Miss' 'Master' 'Other']
```

ستون های افزوده شده :

Embarked_C	Embarked_Q	Embarked_S	sex_num	myTitle_Master	myTitle_Miss	myTitle_Mr	myTitle_Mrs	myTitle_Other
0	0	1	0	0	0	1	0	0
1	0	0	1	0	0	0	1	0
0	0	1	1	0	1	0	0	0
0	0	1	1	0	0	0	1	0
0	0	1	0	0	0	1	0	0

۷. در این قسمت در دیتافریم های عددی ویژگی های بخش قبل که ذکر کردیم را حذف کردیم.

آشنایی با تحلیل آماری

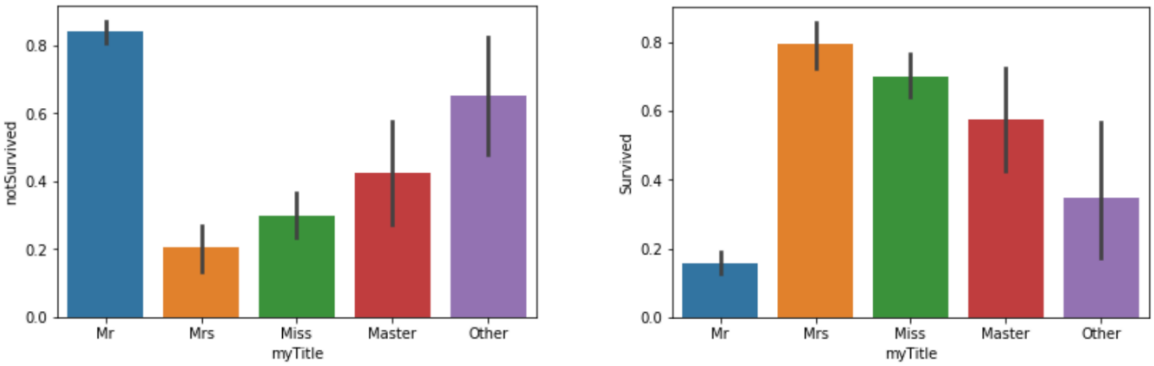
شاخص‌های خواسته شده به صورت زیر محاسبه شده اند:

	variable	mean	median	mode	variance
0	Total Parch	0.381594	0.0	0.000000	0.649728
1	Survived Parch	0.329690	0.0	0.000000	0.677602
2	Not Survied Parch	0.464912	0.0	0.000000	0.595539
3	Total SibSp	0.523008	0.0	0.000000	1.216043
4	Survived SibSp	0.553734	0.0	0.000000	1.659972
5	Not Survied SibSp	0.473684	0.0	0.000000	0.502238
6	Total Sex	0.352413	0.0	0.000000	0.228475
7	Survived Sex	0.147541	0.0	0.000000	0.126002
8	Not Survied Sex	0.681287	1.0	1.000000	0.217772
9	Total Age	29.756420	30.0	32.368090	176.334162
10	Survived Age	30.694453	32.0	32.368090	160.624892
11	Not Survied Age	28.250630	28.0	21.804054	198.405991

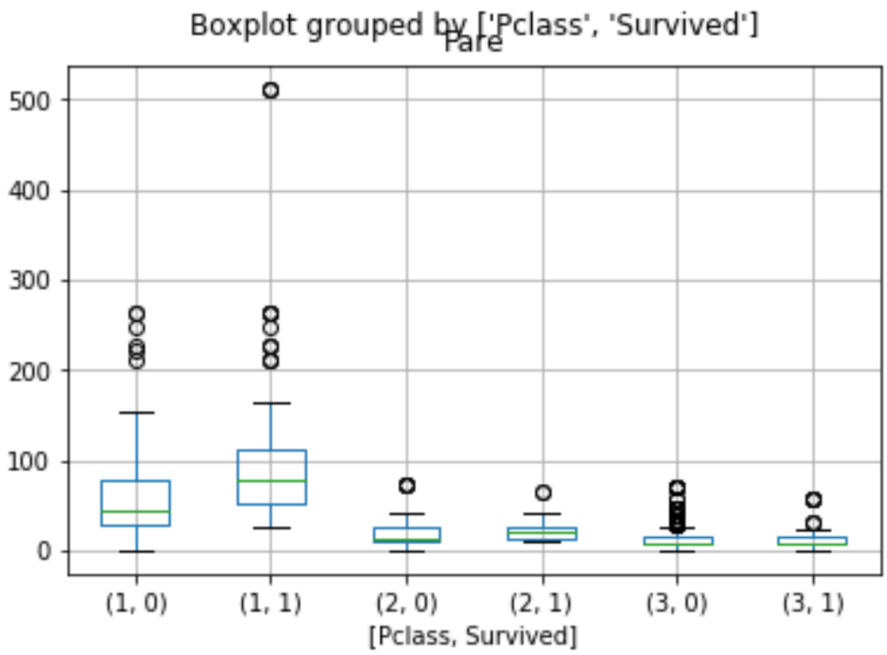
با توجه به این شاخش ها میتوان گفت جنسیت عامل بسیار مهمی است زیرا میانگین (معادل درصد) نجات یافتگان خانم از آقایان بیشتر است.

در Parch هم تفاوتاتی ملاحظه میشود و برای SibSp و Age این تفاوت‌ها کمتر است.

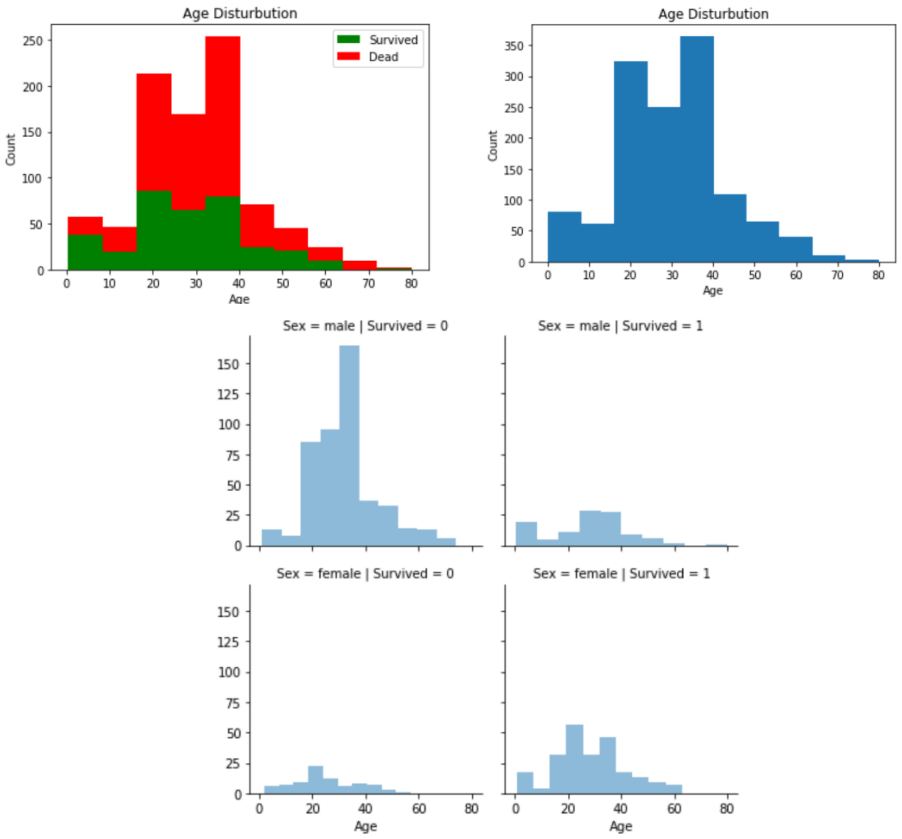
.1



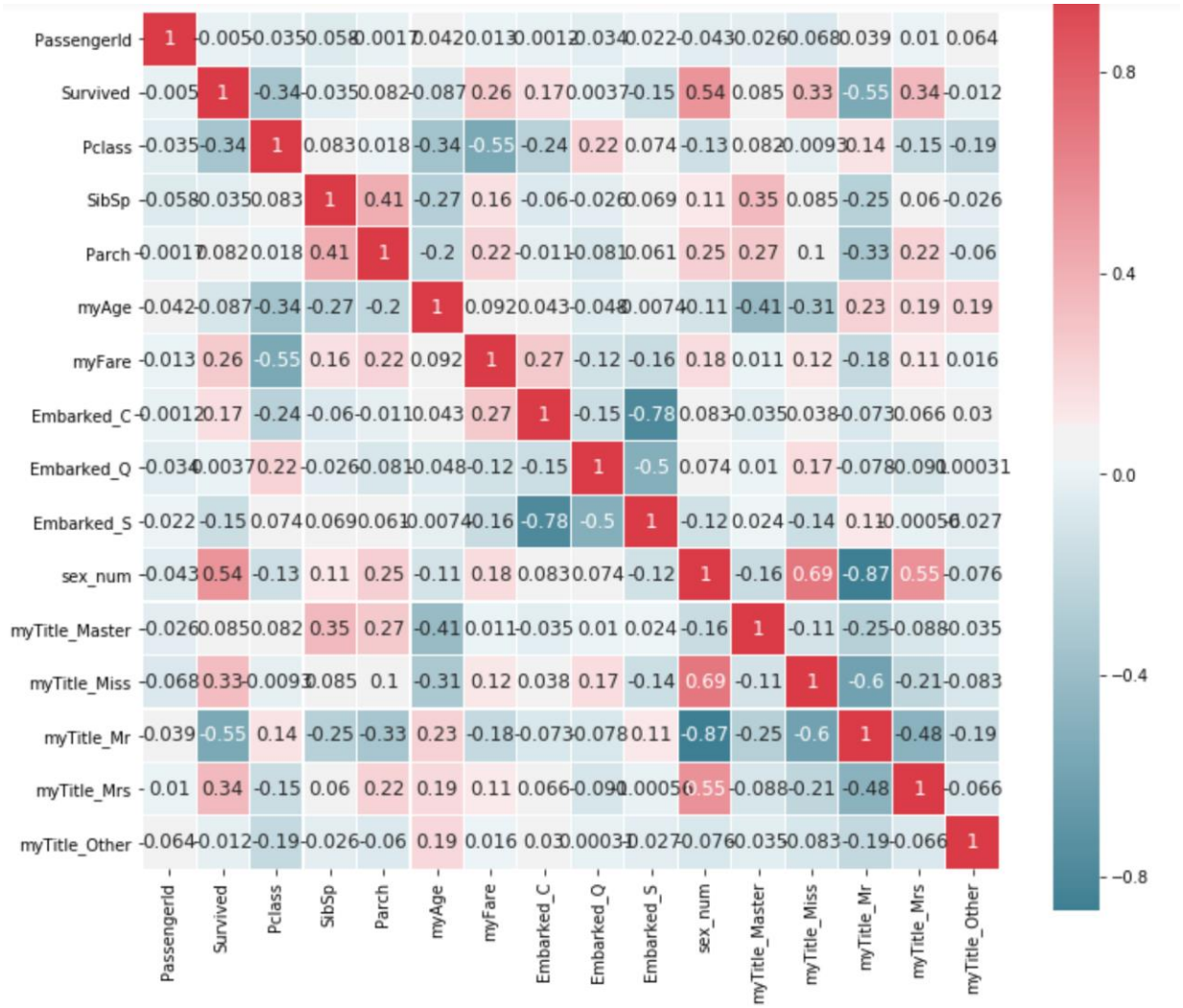
2. می توان نتیجه گرفت قیمت کلاس اول بیشتر از کلاس دوم و قسمت کلاس دوم بیشتر از کلاس اول است. همچنین پراکندگی قیمت ها در کلاس اول بیشتر است. (نمودار ها در صفحه بعد)



3. می توان مشاهده کرد تعداد نجات یافتگان خانم در هر سن بیشتر از مردگان خانم و برای آقاها این امر برعکس است. و در هر سن تعداد آقایان مرده بیش از تعداد خانم های مرده است.



4. در اینجا هم مشاهده میشود جنسیت با زنده‌ماندن همبستگی زیادی دارد. همچنین کوریلشن قیمت و زنده ماندن هم قابل توجه است. در مورد ویژگی ها هم تایتل ها و سن و جنسیت در گروه غیر Other کورلیشن زیادی دارند که میتوانیم بعضی القاب را به دلیل اینکه توسط سن و جنسیت مورد پوشش قرار میگیرند (مثل Miss) را حذف کنیم. بقیه همبستگی ها آنقدر بالا نیست که موجب حذف گروهی شود.



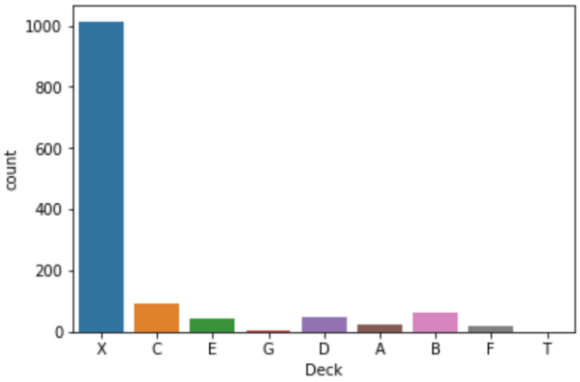
استخراج ویژگی جدید

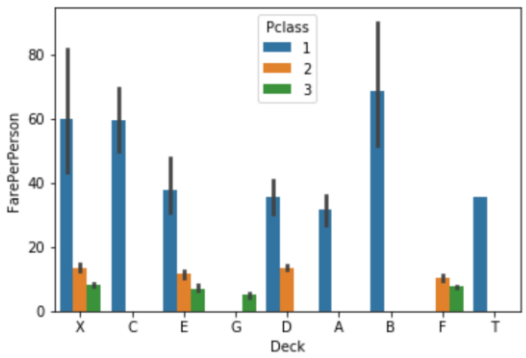
1. برای اینکار با شرط گذاشتن روی سن کمتر از 18 سال و همچنین جنسیت ستون ویژگی جدید را ساختیم.

2. اینکار با جمع زدن دو متغییر Parch و SibSp انجام شد.

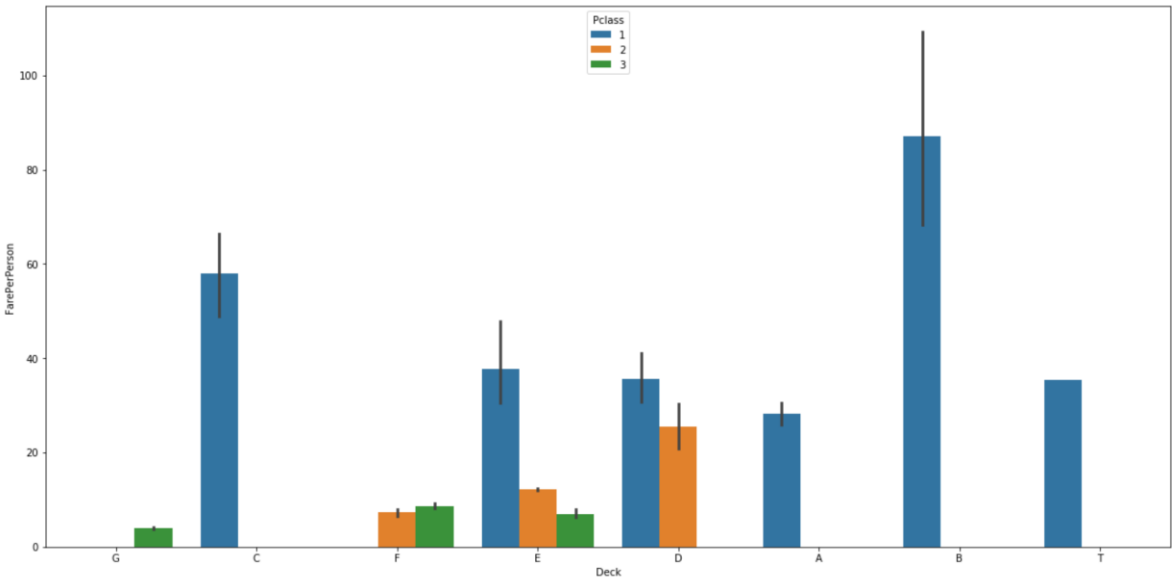
3. خود کابین ها بسیار زیادند و غیرقابل پیش بینی اما با استفاده از حرف ابتدا هر کابین که نشان دهنده عرشه ای است که آن کابین در آن بوده میتوانیم عرشه های افراد را پیش بینی کنیم. برای اینکار از قیمت برای هر نفر (قیمت تقسیم بر تعداد نفرات) با توجه به نمودار های زیر با تعیین مرزهایی میتوانیم عرشه را نسبت دهیم.

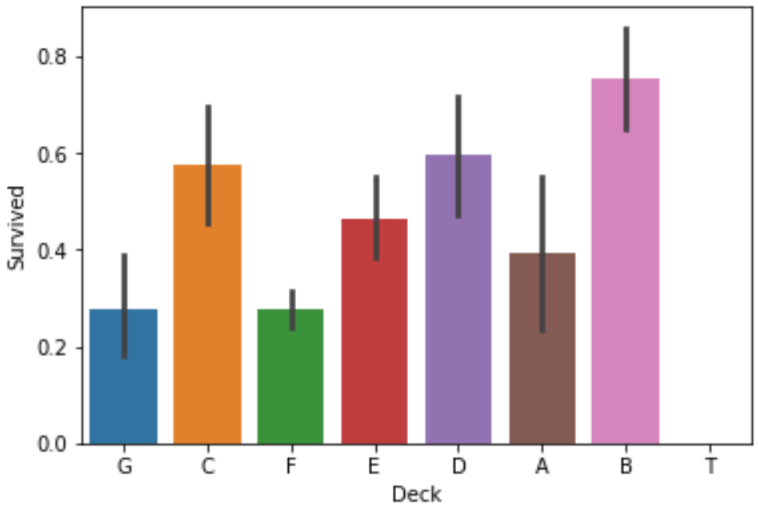
نمودار میله ای عرشه ها (X نماینده عرشه ای که نمیدانیم) قبل اعمال شروط :





بعد :

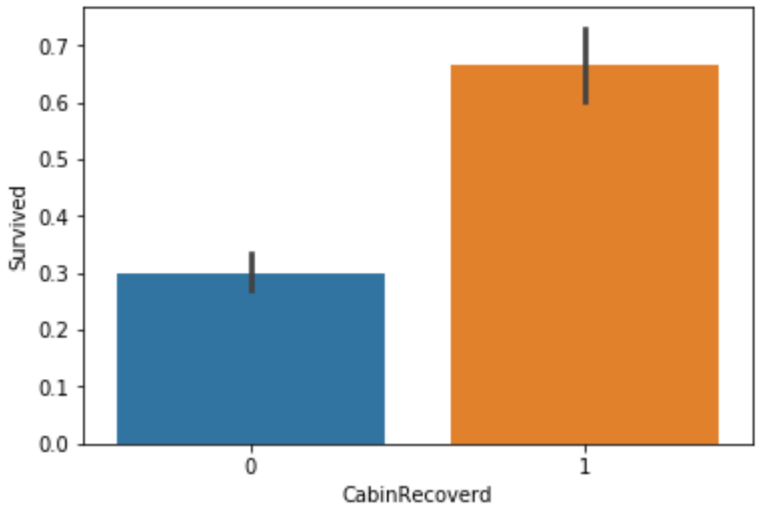


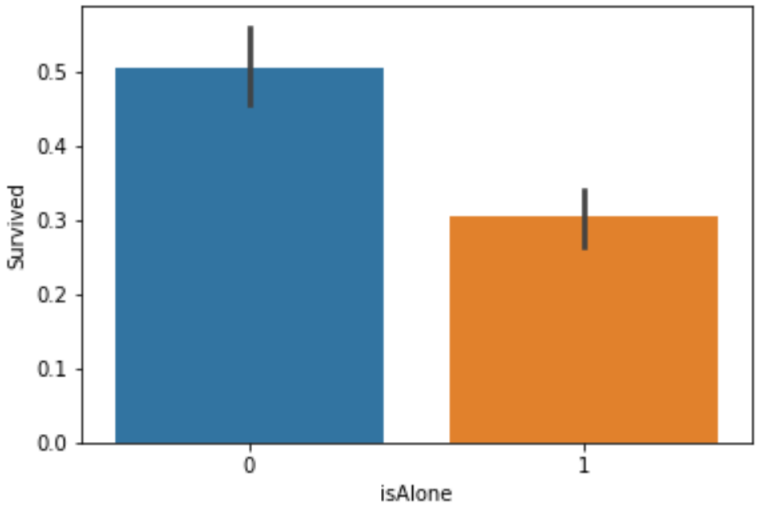


ادامه‌ی راه

دو ویژگی جدید اضافه میکنیم که به ترتیب عبارتند از همراه داشتن یا نداشتن کسی و همچنین موجود بودن اطلاعات کابین یا نه.

باتوجه به نمودار های زیر مشخص است این ویژگی ها خوب هستند چون تفاوت زنده ماندن در دو دسته آنها قابل ملاحضه است.





در نهایت همه ی ویژگی ها را کمی میکنیم و بار دیگر correlation map را رسم میکنیم. که در آنجا هم مشاهده میکنیم موجود بودن کابین همبستگی زیادی با زنده ماندن دارد که منطقی هم هست چون احتمالا افرادی که نجات یافته اند کابینشان مشخص شده است.

نمودار بزرگ است و در نوت بوک بهتر نمایش داده میشود.