项目二：titanic 数据探索分析

邓剑波

问题提出：

1、样本总体的存活率是多少？

2、所给数据中，特征和乘客的生存率有什么关系？

项目描述：

项目二是对数据集 titanic-data 进行整理和分析，通过清理数据和分析数据，提出问题，得出结论。

一、数据探索

1、首先对数据集 titanic-data 打开进行查看，发现数据由乘客编号、生存状况、船舱等级、乘客姓名、乘客性别、乘客年龄、配偶和亲属状况、票号、票价、舱号、登船地点组成，大致状况如下：

	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

2、查询数据的属性：

	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

3、查询数据的状况，通过对数据的基本信息的查询，可以得知数据存在非常多的缺失，例如 Age 年龄数据，就存在非常多的缺失，具体状况如下：

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

二、数据清理

1、结合实际经验，可以假设数据中的 PassengerId（乘客编号）， Name（乘客姓名）, Embarked（登船地点），Ticket（票号）并不会对生存状况造成影响，并且 Cabin（舱号）数据存在过多的缺失，所以选择将以上特征排除出数据，清理过后得到的数据状况如下：

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 7 columns):
Survived    891 non-null int64
Pclass      891 non-null int64
Sex         891 non-null object
Age         714 non-null float64
SibSp       891 non-null int64
Parch       891 non-null int64
Fare        891 non-null float64
dtypes: float64(2), int64(4), object(1)
memory usage: 48.8+ KB
```

2、从上图可得知，Age（年龄）存在较多的缺失数据，可由统计学知识得知，中位数可以很好的描述数据的平均状况，同时不容易受到极端值影响，所以，选择 Age 特征的中位数对缺失的数据进行填充；

3、其中，SibSp 和 Parch 同样表示的都是亲属状况，将两者数量合并为 Relatives；

4、清理过后数据属性状况如下：

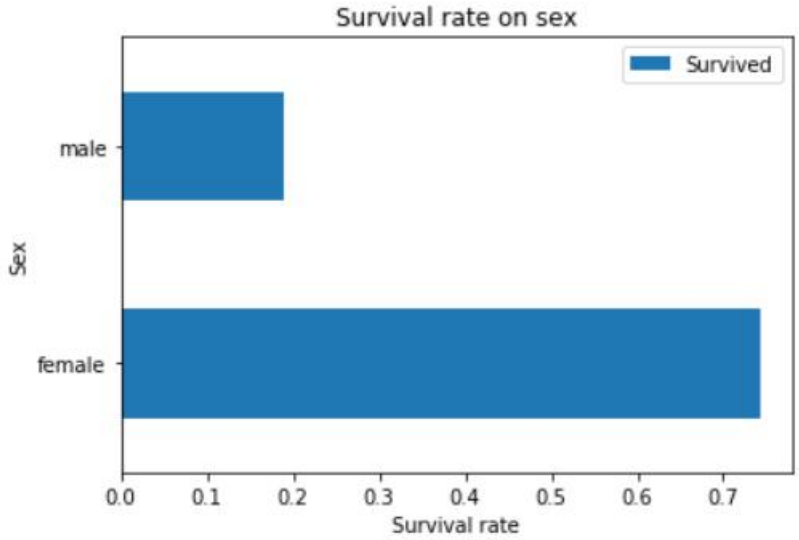
	Survived	Pclass	Age	SibSp	Parch	Fare	Relatives
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.361582	0.523008	0.381594	32.204208	0.904602
std	0.486592	0.836071	13.019697	1.102743	0.806057	49.693429	1.613459
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	22.000000	0.000000	0.000000	7.910400	0.000000
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200	0.000000
75%	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000	1.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200	10.000000

三、数据分析

1、性别对生存率的影响：

通过对数据的整合分析，可以得到，乘客中，女性乘客的生存率为 0.742038，男性乘客的生存率为 0.188908，数据中可以很明显的得知，女性乘客的生存率远高于男性；

Sex	Survival_rate
female	0.742038
male	0.188908

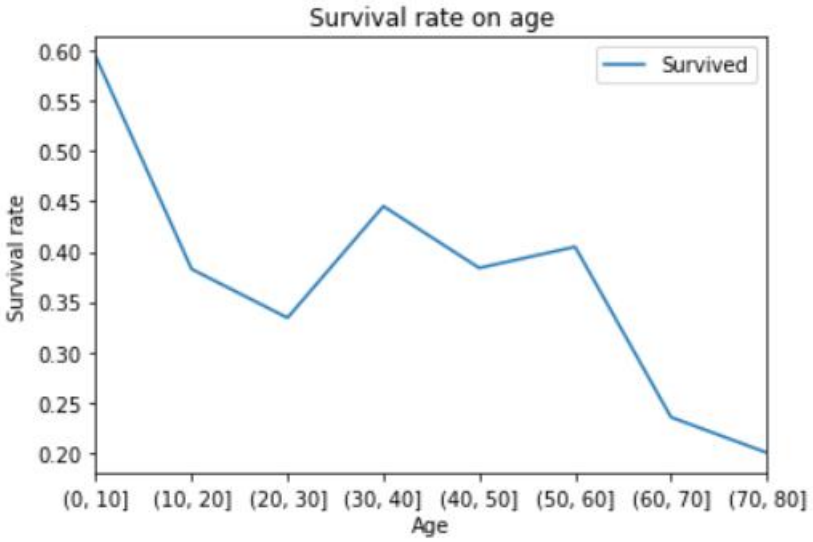


分析：从性别对生存率的影响上看，女性的生存率远高于男性，原因可能在于当时的社会氛围，普遍重视绅士风度。

2、年龄对生存率的影响：

首先，由于年龄数据较多，首先将年龄进行分组，从数据中可得知，年龄最小为 0 岁，最大为 80 岁，决定以每 10 岁进行一个分组，再计算每组的生存率，可得知随着年龄的增长，生存率逐步下降，儿童（0—10 岁）的生存率最高，在青壮年（20-60 岁）阶段，30—40 岁组别生存率最高，具体图标如下：

Age	Survival_rate
(0, 10]	0.593750
(10, 20]	0.382609
(20, 30]	0.334152
(30, 40]	0.445161
(40, 50]	0.383721
(50, 60]	0.404762
(60, 70]	0.235294
(70, 80]	0.200000



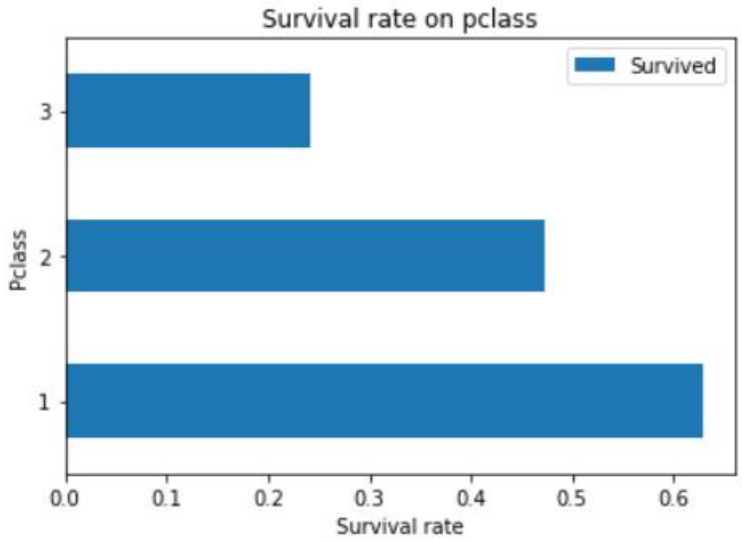
分析：生存率随着年龄的增加，普遍呈现减少的趋势，证明船上当时妇女和儿童优先的逃难的秩序较好，但是在青壮年到中年阶段，年轻人明显生存率会有所升高，原因可能是在排除妇女儿童的先后顺序后，年轻人的体力能更好的从灾难环境中求得生存。

3、船舱等级对生存率的影响：

船舱等级分为 1、2、3 三个等级，生存率数据分别为：

Pclass	Survival_rate
1	0.629630
2	0.472826
3	0.242363

从数据中可以得知，等级越高的船舱的乘客，生存率越高，生存率的柱形图如下：

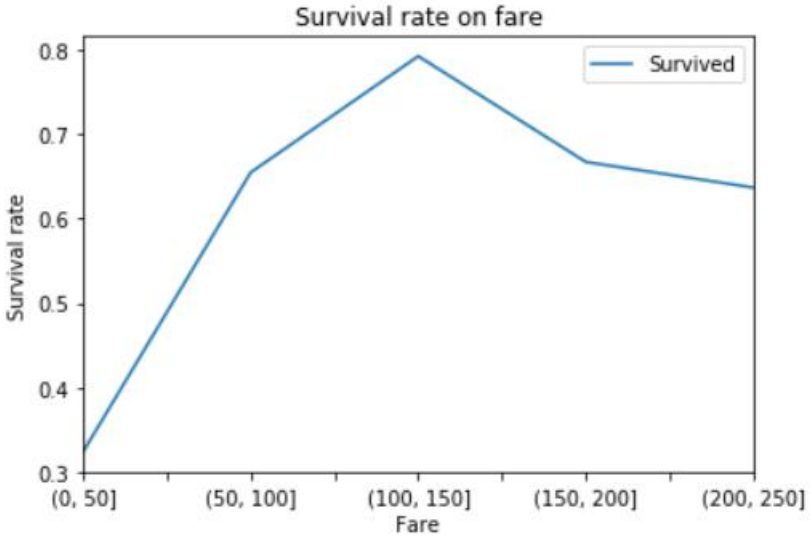


分析：从船舱等级上看，越高的船舱等级的乘客生存率越高，原因或许是因为高等级船舱都普遍分布在甲板附近，拥有更短的逃生距离，并且船舱分布不如低等级船舱紧凑，人口密度较低，也利于逃生。

4、船票价格对生存率的影响：

首先，通过对数据的观察，存在三个 512.3292 票价的远超于正常票价范围的异常值，所以去掉这三个值，通过对票价 0—300，每 30 元一组进行分组，可得：

Fare	Survival_rate
(0, 50]	0.324022
(50, 100]	0.654206
(100, 150]	0.791667
(150, 200]	0.666667
(200, 250]	0.636364



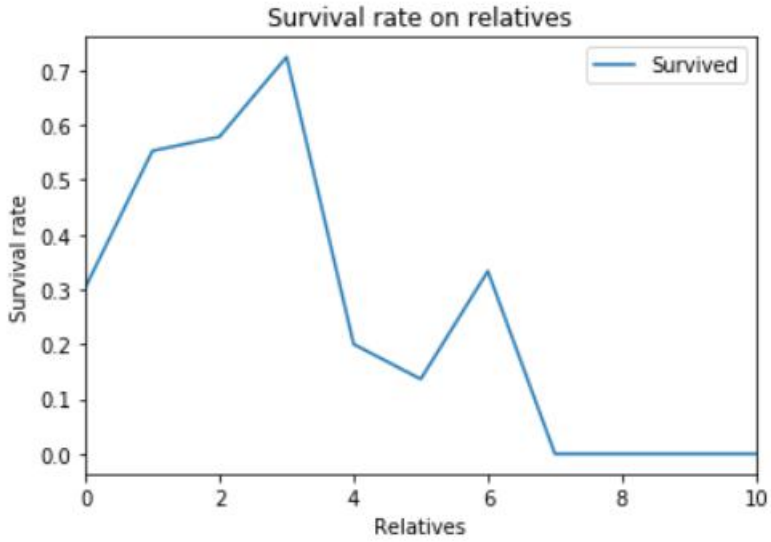
分析：从数据和图像得出，在 150 元之前，生存率随着票价的提高上升，在 150 元之后生存率却在下降，分析原因：1、首先越高的票价一般表示越高的船舱等级，所以随着票价

的升高，生存率会有所升高；2、在高票价阶段，通过对原始数据调查，发现 150 元以上票价的只有 29 人，相较于 891 人的样本，或许是小样本的偏差带来的。

5、亲属数量对生存率的影响：

从数据上看出，生存率随着亲属数量增加而上升，亲属数量在 3 个人的时候，乘客的生存率达到最高，之后随着数量的增加，生存率下降。

Relatives	Survival_rate
0	0.303538
1	0.552795
2	0.578431
3	0.724138
4	0.200000
5	0.136364
6	0.333333
7	0.000000
10	0.000000



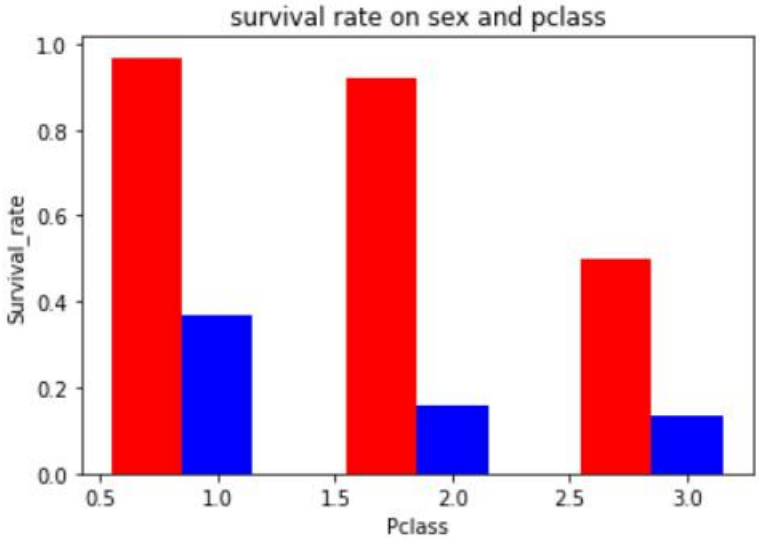
分析：从结果上看，3 口之家的生存率最高，生存率图形出现先升后降的原因，或许是因为在家庭成员较少的时候，灾难发生时，乘客无法得到足够的帮助，当有一定的亲人时，同心协力逃难使得生存率得到提高，但是过大的家庭规模，反而会因为行动不便拖了后腿。

6、性别、船舱等级对生存率的影响：

（1）除了知道性别和船舱等级分别对生存率的影响，还按照两者进行同时分类调查生存率影响，得知，无论船舱等价如何，女性的生存率还是远高于男性，只是在等级 3 的船舱，差距不如等级 1，2 船舱明显，具体数据如下：

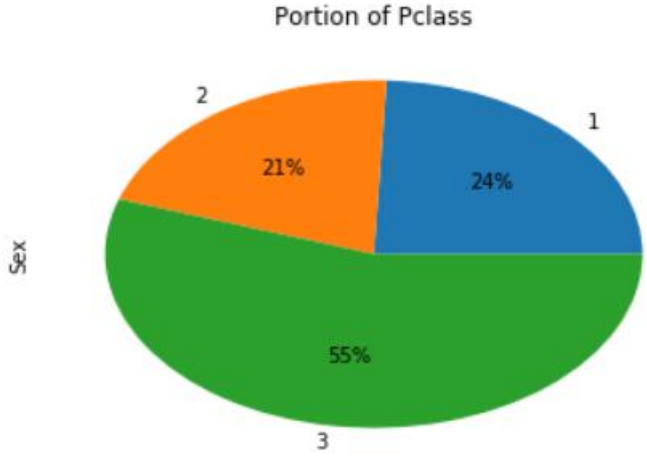
Sex	Pclass	Survival_rate
female	1	0.968085
	2	0.921053
	3	0.500000
male	1	0.368852

2	0.157407
3	0.135447

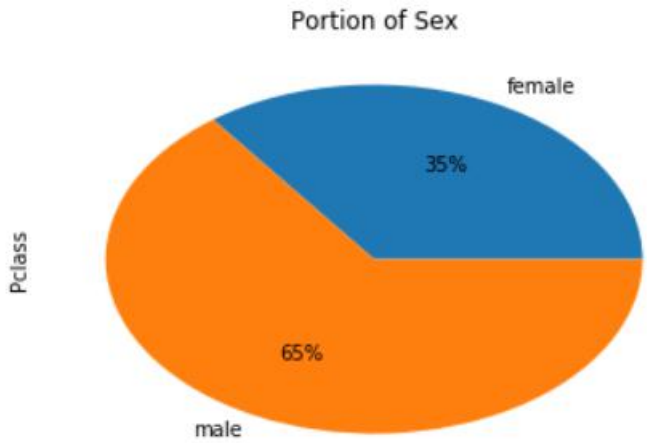


（2）除了对生存率进行探索，为了避免数据因为小样本产生偏差，分别对船舱等级的数据比例，和船上乘客的性别比例进行的探索，可以排除小样本对数据比例带来的影响，具体结果如下：

Pclass	nums
1	216
2	184
3	491



Sex	nums
female	314
male	577



（3）分析：首先从男女比例和船舱等级比例上看，可以排除某一个特征为小样本，从而干扰性别和船舱等级的分析；

从数据上可以观察到，性别对生存率的影响十分重要，无论是哪个等级的船舱女性生存率都是远高于男性的，证明上面对性别生存率影响的分析普遍适用于各个等级；另外，等级 3 的客舱男女生存比率的差距并没有等级 1、2 的大，原因可能在于等级 3 的乘客，普遍的生存率都很低，所以道德因素无法发挥作用。

四、总结

1、 通过对数据的调查和分析，可以知道，数据样本的总体存活率约为 38.39%；

可以对数据的分析，还原当时灾难发生时的大致场景，在逃生阶段，船上并没有发生完全的混乱，还是拥有一定的秩序，遵循妇女儿童优先的原则，所以妇女儿童的生存率普遍较高；但是船舱等级越好，能负担得起更高票价的社会上流，因为在船上所处位置拥有优势，所以可能更早的得到消息并且更容易疏散，生存率会高于在船底舱的贫穷乘客；同时，由于灾难发生时人人无法自顾，有着一定家庭成员的乘客，能够团结在一起互相帮助，会相较孤身一人的乘客能提高生存率，但是，过多的家庭成员反而导致逃生的行动不便，生存率明显下降。

2、 该数据分析的准确性还是受到一定的限制：

通过背景调查看，泰坦尼克号上总共有 2224 人，而数据只是不到一半的人数的状况，所以存在样本的偏差，例如该样本中可能会有过高的生存率，并不能真实的反应总体生存率的实际状况；

另外，除了所给数据特征，可能有其他特征会影响生存率，例如同样是男性，有可能船员的生存率会比普通乘客的生存率低，数据特征的不足也会限制分析的准确性；

最后，报告采用的数据分析方法，大都是单变量分析，但是实际状况中，特征之间的相互关系差异也有可能对生存率产生较大影响，所以该报告的准确性也会受到分析方法的限制。

所以，该分析在现有数据阶段，结合报告中的分析方法，结论只是暂时的，随着新的数据和特征出现，通过不同的分析方法，可能会有着不同的结论。

五、参考资料

http://pandas.pydata.org/pandas-docs/stable/index.html pandas 说明文档
http://matplotlib.org/index.html matplotlib 说明文档