浙江大学

本科实验报告

课程名称：数据挖掘导论

姓　　名：张溢弛

学　　院：计算机科学与技术学院

专　　业：软件工程

学　　号：3180103772

指导教师：李石坚

2021 年 5 月 11 日

浙江大学 实验报告

课程名称：数据挖掘导论　　实验类型：综合

实验项目名称：数据可视化

学生姓名：张溢弛　专业：软件工程　学号：3180103772

同组学生姓名：无　指导老师：李石坚

实验地点：曹西-503　实验日期：2021 年 5 月 11 日

目录

一　实验基本信息

1.1　实验内容

选取 Kaggle 或者 KDD Cup 中的公开数据集，并选用若干种方法对数据集进行可视化与简要的分析。

1.2　实验环境

实验的基本环境如下：

- 操作系统：Windows 10
- 编程环境：Anaconda 4.10+Jupyter Notebook 4.4.0
- 编程语言：Python 3.6.5
- Python 库版本：Pandas 0.23.0，seaborn 0.8.1，matplotlib 2.2.2

二　数据集介绍

本次实验我选用了 Kaggle 入门级比赛中的 Titanic 数据集，该数据集包含一系列泰坦尼克号乘客的基本信息，而目标则是预测乘客是否从泰坦尼克号事件中存活下来，测试集和训练集一共有 1309 条数据。该数据集主要有如下几个字段：

- Pclass 乘客所在的船舱等级
- Survived 乘客是否存活
- Name 乘客的名字
- Sex 乘客性别
- Age 乘客的年龄
- SibSp 乘客的兄弟姐妹和配偶数量
- Parch 乘客的父母与子女数量
- Ticket 乘客的票的编号
- Fare 票价
- Cabin 乘客的座位号
- Embarked 乘客的登船码头

数据的大致格式如下图所示，其中一小部分数据是空缺的：

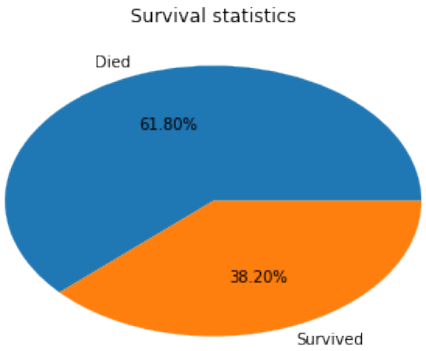
	Pclass	Survived	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	1	Hays, Miss. Margaret Bechstein	female	24.0	0	0	11767	83.1583	C54	C
1	3	0	Holm, Mr. John Fredrik Alexander	male	43.0	0	0	C 7075	6.4500	NaN	S
2	3	0	Hansen, Mr. Claus Peter	male	41.0	2	0	350026	14.1083	NaN	S
3	3	0	Keane, Mr. Andrew "Andy"	male	NaN	0	0	12460	7.7500	NaN	Q
4	2	0	Milling, Mr. Jacob Christian	male	48.0	0	0	234360	13.0000	NaN	S

三　数据可视化结果

我使用 pandas 与 matplotlib 等 Python 数据科学库对数据集进行了读取、预处理、查询与可视化，具体的代码可以在 code 目录下的 homework1.ipynb 代码文件中查看 (需要先安装好 jupyter notebook 环境)，下面主要展示我在可视化过程中产生的一些结果。整个过程中用到了扇形统计图、条形统计图、直方图、小提琴图、箱线图等形式的图表。

3.1　总体存活情况统计

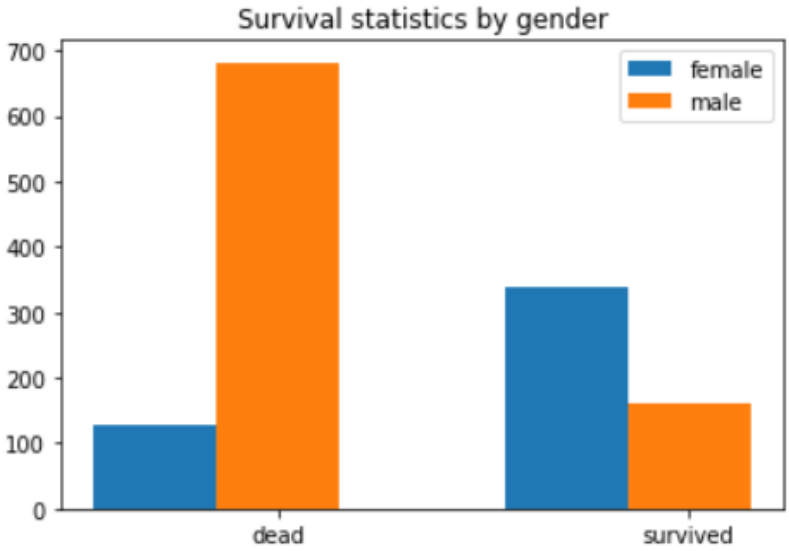
首先进行最简单的一项可视化工作，统计整个数据集 (训练集和测试集) 中的存活和死亡的人数，并绘制成如下扇形统计图：



可以看到总体上来说，样本的数据分布中，死亡的人数占了多数。

3.2　按性别统计存活情况

然后我们按照性别分别统计男性和女性存活以及死亡的人数，并用条形统计图的方式来呈现人数的对比情况，最终的结果如下图所示：

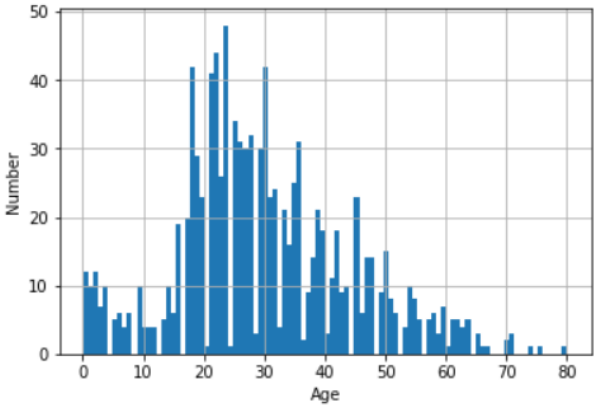


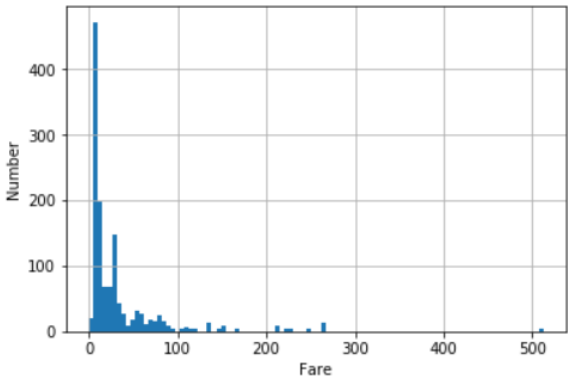
女性存活率：339/466=0.7274678111587983
男性存活率：161/843=0.19098457888493475

从这一可视化的结果中很容易就可以分析出，女性的存活率要远高于男性，遇难乘客主要都是男性而存活的乘客主要都是女性，**这说明在是否存活的预测任务中，性别是一个很重要的特征。**

3.3 年龄与票价的分布

下面两幅直方图展示了数据集中的乘客的年龄（Age）分布情况和船票票价（Fare）的分布情况：



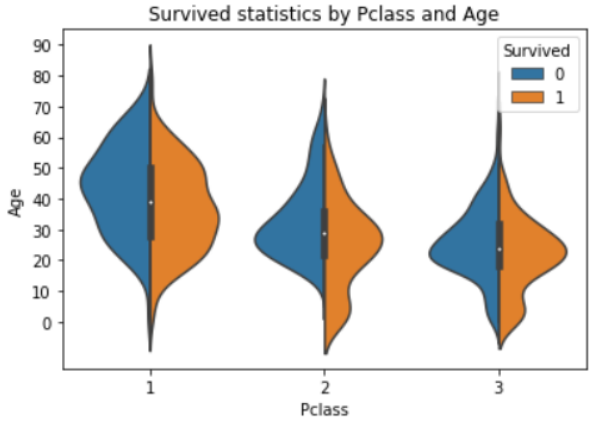


我们可以发现乘客主要以青年人为主，而船票价格基本趋于一致，仅有少部分高价票，接下来我们将进一步分析年龄与票价和存活情况之间的关系。

3.4 按船舱等级统计存活情况

小提琴图 (Violin Plot) 是用来展示多组数据的分布状态以及概率密度。这种图表结合了箱形图和密度图的特征，主要用来显示数据的分布形状。跟箱形图类似，但是在密度层面展示更好。在数据量非常大不方便一个一个展示的时候小提琴图特别适用。

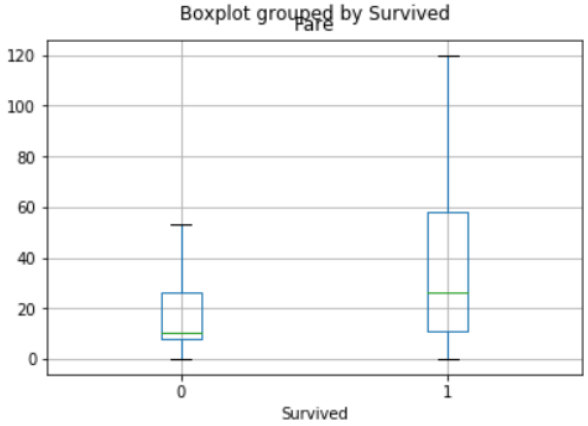
这里我们采用小提琴图来对不同船舱 (共有 1,2,3 三种，Pclass 字段) 的不同年龄段的存活情况进行可视化，得到了下面的小提琴图：



从这个图中我们可以分析出，相对而言，年轻乘客，特别是 10 岁及以下的儿童的存活概率更高，说明年龄也是影响存活与否的重要因素，同时船舱的不同也会影响不同年龄段乘客的存活概率。

3.5 按票价统计存活率

我使用箱线图对票价和存活率的关系进行了可视化，得到的结果下图：



存活人乘客票价信息

count	500.000000
mean	49.361184
std	68.648795
min	0.000000
25%	11.214600
50%	26.000000
75%	57.750000
max	512.329200

Name: Fare, dtype: float64

死亡乘客票价信息

count	808.000000
mean	23.353831
std	34.145096
min	0.000000
25%	7.854200
50%	10.500000
75%	26.000000
max	263.000000

Name: Fare, dtype: float64

从这一结果中我们可以分析得到，买高价票的乘客存活概率高于低价票的乘客，从箱线图反映出的各种统计指标来看，存活的乘客的票价都比死亡的乘客的要高，我推测票价与消费水平和身份地位成一定的关系，身份地位高的人更容易活下来。

四 实验总结

本次实验中我实践了数据挖掘中的数据预处理和可视化分析等环节的具体操作，初步学习了 pandas，seaborn 和 matplotlib 等数据挖掘常用 Python 库的基本用法，并采用多种图表对 Titanic 数据集进行了初步的分析，收获颇丰。