Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа № 1
по дисциплине
«Методы машинного обучения»
на тему

«Создание "истории о данных" (Data Storytelling)»

Выполнил:
студент группы ИУ5И-23М
Ван Тяньшо

Москва — 2024 г.

1. Цель лабораторной работы

Цель лабораторной работы: изучение различных методов визуализация данных и создание истории на основе данных.

Краткое описание. Построение графиков, помогающих понять структуру данных, и их интерпретация.

2. Задание

Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.

На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.

Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.

Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.

История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

Сформировать отчет и разместить его в своем репозитории на github.

Средства и способы визуализации данных можно посмотреть здесь.

Мой набор данных："Titanic"

3. Текст программы

Агент реализуется с помощью метода итераций стратегии, который используется для решения задачи интенсивного обучения Taxi-v3.Повторение стратегии включает в себя два основных этапа: оценку стратегии и совершенствование стратегии до тех пор, пока стратегия не станет стабильной.

Шаг 1: Распределение пассажирского пространства -гистограмма

```
import seaborn as sns
import matplotlib.pyplot as plt

# 加载数据集
titanic = sns.load_dataset('titanic')

# 绘制直方图
plt.figure(figsize=(10,6))
sns.histplot(titanic['pclass'], bins=3, kde=False)
plt.title('乘客舱位分布')
plt.xlabel('舱位')
plt.ylabel('乘客数量')
plt.show()
```

Шаг 2: Диаграмма распределения пассажиров по возрасту - оценка ядерной плотности (KDE)

```
plt.figure(figsize=(10,6))
sns.kdeplot(titanic['age'], shade=True)
plt.title('乘客年龄分布')
plt.xlabel('年龄')
plt.ylabel('密度')
plt.show()
```

Шаг 3: Выживаемость в разбивке по полу -гистограмма

```
plt.figure(figsize=(10,6))
sns.barplot(x='sex',  y='survived',  data=titanic)
plt.title('按性别的生还情况')
plt.xlabel('性别')
plt.ylabel('生还率')
plt.show()
```

Шаг 4: Количество родственников на борту-boxplot

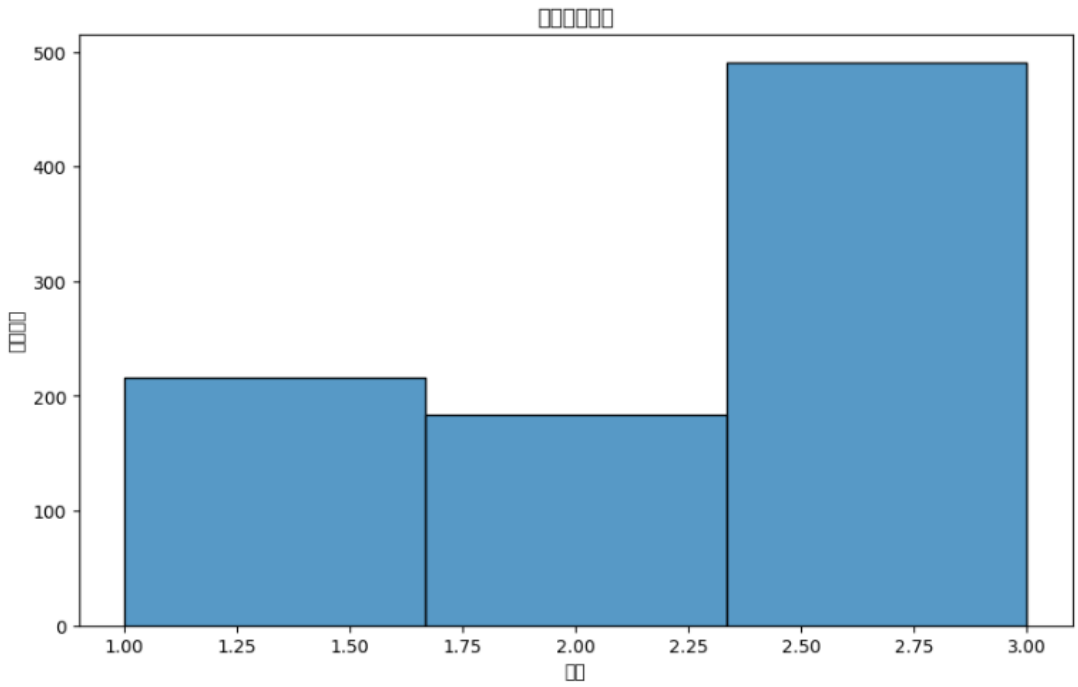
```
[4] plt.figure(figsize=(10,6))
    sns.boxplot(x='survived',  y='sibsp',  data=titanic)
    plt.title('船上兄弟姐妹/配偶数量')
    plt.xlabel('生还')
    plt.ylabel('兄弟姐妹/配偶数量')
    plt.show()

    plt.figure(figsize=(10,6))
    sns.boxplot(x='survived',  y='parch',  data=titanic)
    plt.title('船上父母/子女数量')
    plt.xlabel('生还')
    plt.ylabel('父母/子女数量')
    plt.show()
```

Шаг 5:Влияние тарифа на ситуацию с погашением кредита -точечная диаграмма

```
plt.figure(figsize=(10,6))
sns.scatterplot(x='fare',  y='survived',  data=titanic)
plt.title('票价对生还情况的影响')
plt.xlabel('票价')
plt.ylabel('生还情况')
plt.show()
```

4. Результат выполнения кода

Шаг 1: Распределение пассажирского пространства -гистограмма



Как видно из карты распределения пассажиров по классам, наибольшее количество пассажиров приходится на третий класс, за ним следуют первый класс и второй класс с наименьшим количеством пассажиров.Это говорит о том, что тариф третьего класса ниже и его легче принять большинству пассажиров.

Шаг 2: Диаграмма распределения пассажиров по возрасту - оценка ядерной плотности (KDE)

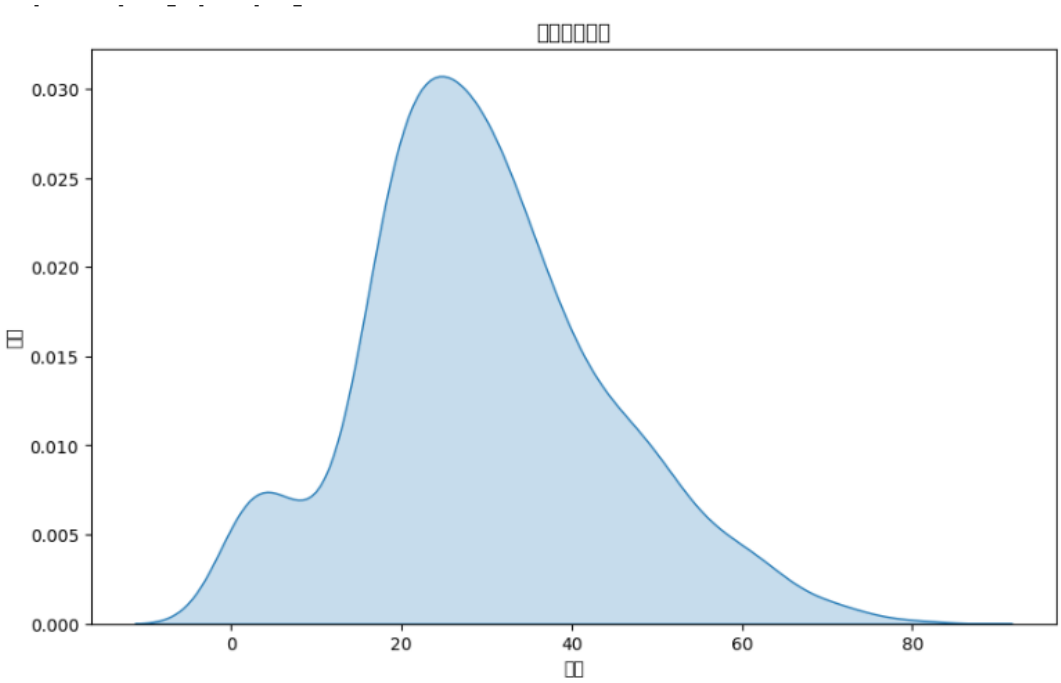
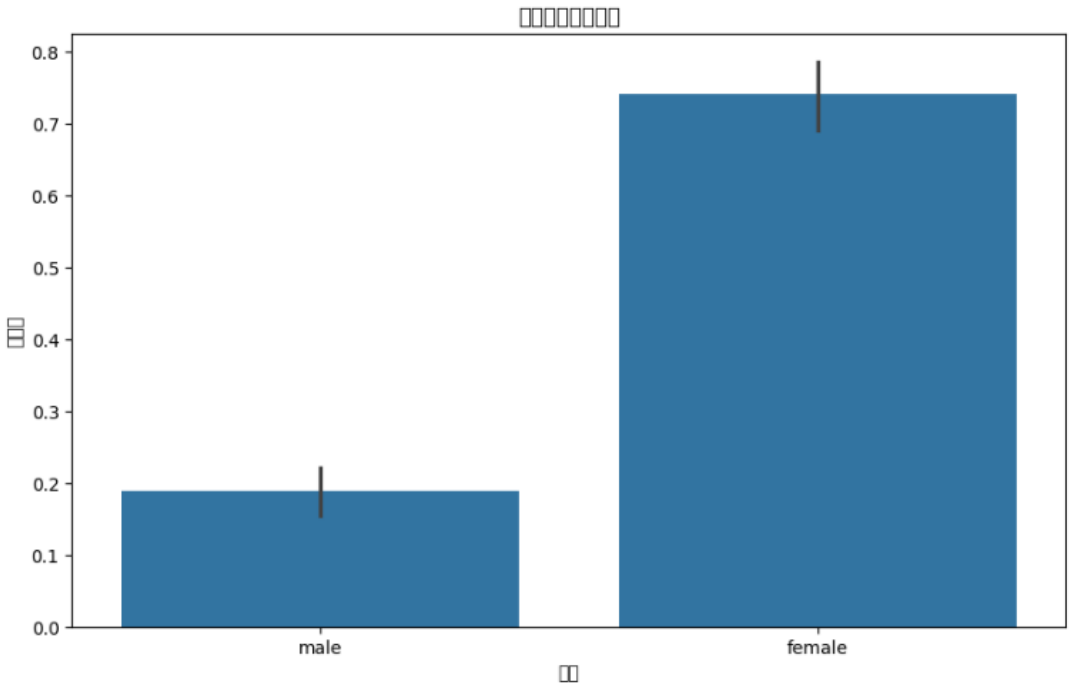


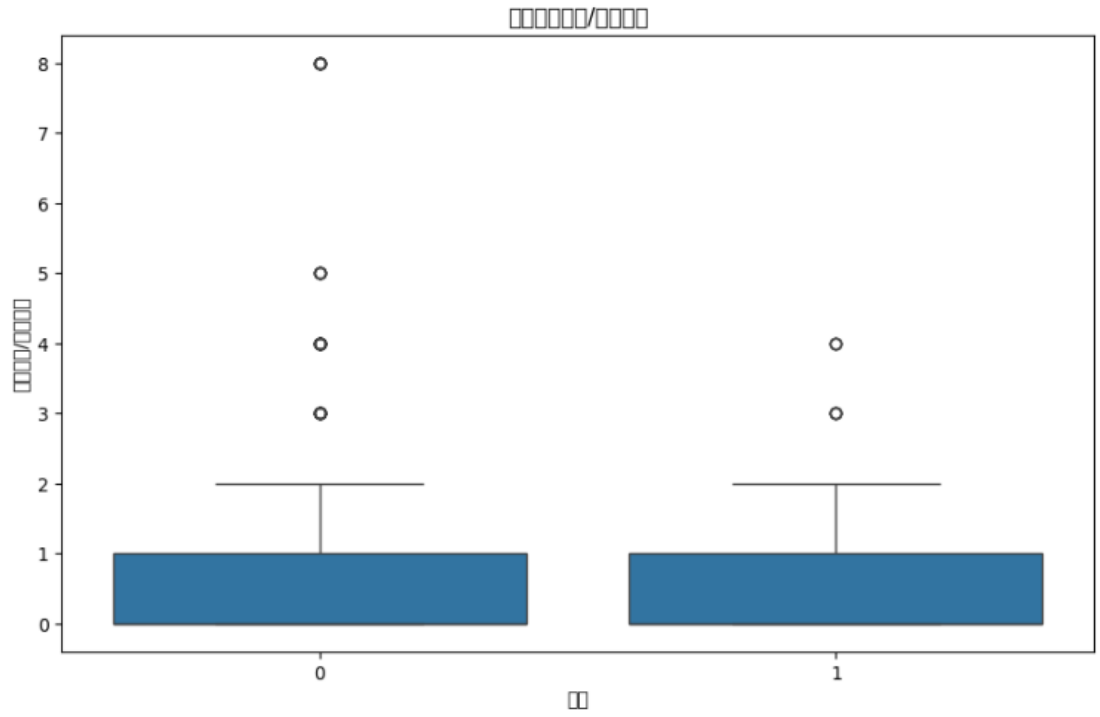
Таблица выживаемости в разбивке по полу показывает, что уровень выживаемости женщин-пассажиров значительно выше, чем у мужчин.Возможно, это связано с правилом, согласно которому в случае бедствия женщины и дети первыми садятся в спасательную шлюпку.

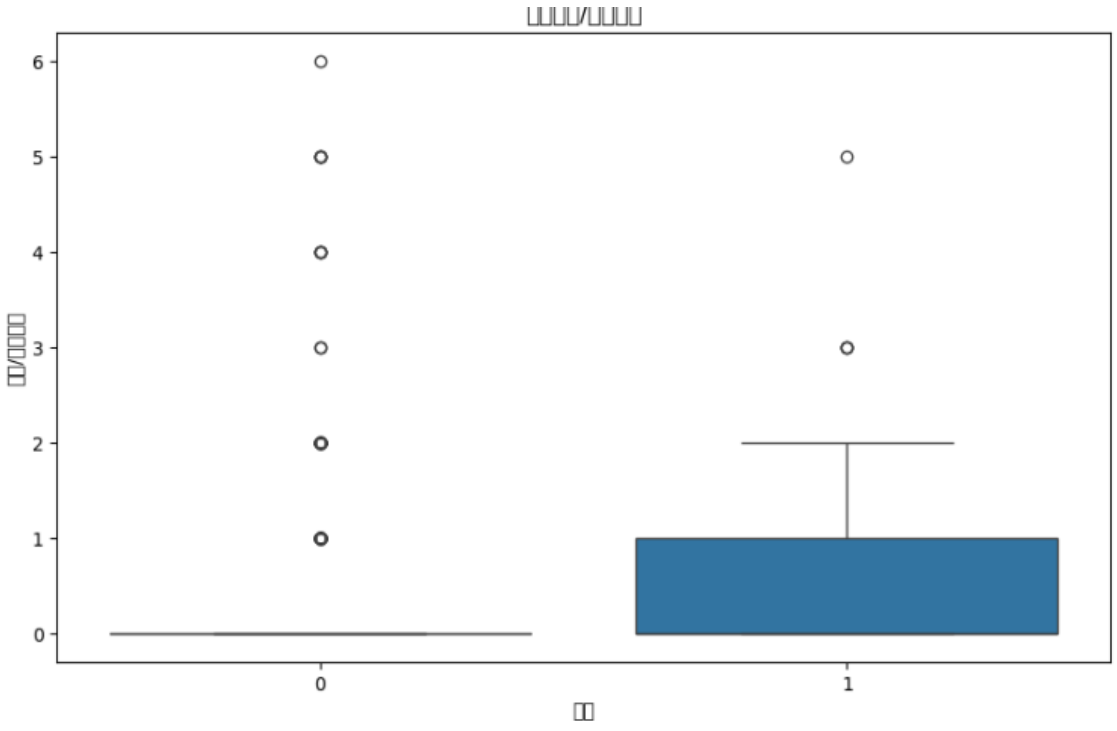
Шаг 3: Выживаемость в разбивке по полу -гистограмма



Карта распределения по возрасту показывает, что возраст большинства пассажиров составляет от 20 до 40 лет.Пассажиры этой возрастной группы составляют наибольшую долю, что свидетельствует о том, что пассажирами этого судна являются в основном молодые люди.

Шаг 4: Количество родственников на борту-boxplot





Как видно из рисунка, среди выживших пассажиров меньше случаев, когда среди них больше братьев и сестер/супругов.Это может быть связано с

тем, что большему числу членов семьи труднее выжить в чрезвычайной ситуации.

Распределение числа родителей/детей среди выживших и не выживших пассажиров примерно одинаковое, что указывает на то, что количество родителей/детей мало влияет на вероятность выживания.

Шаг 5:Влияние тарифа на ситуацию с погашением кредита -точечная диаграмма

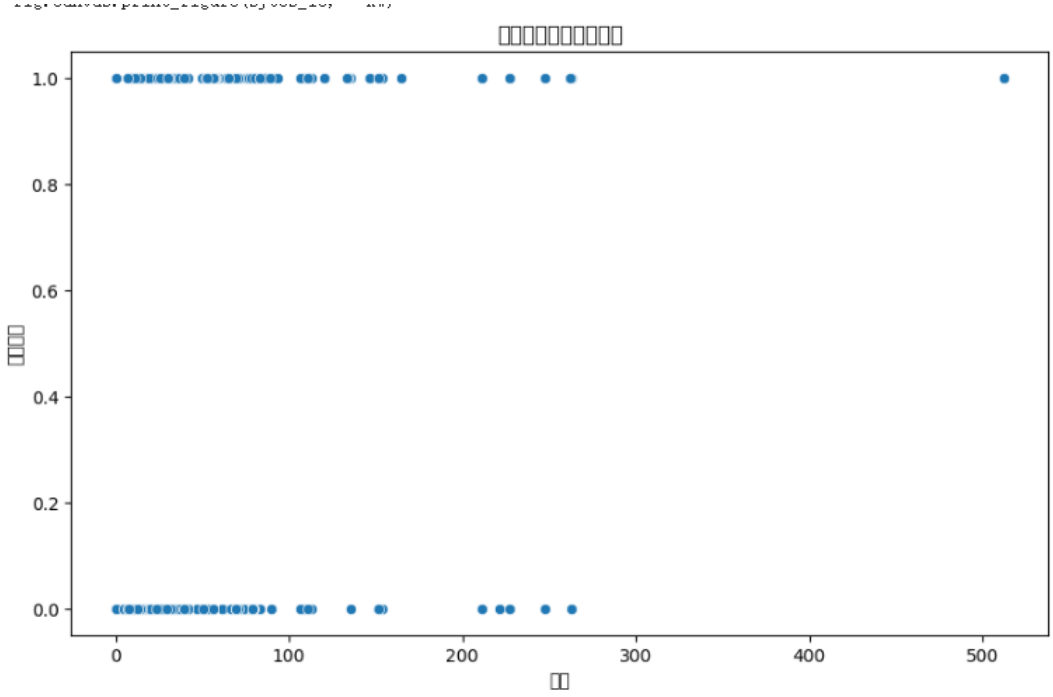


График влияния стоимости билета на выживаемость показывает, что пассажиры с более дорогими билетами выживали чаще. Это может быть связано с тем, что пассажиры с дорогими билетами находились в первом или втором классе, которые были ближе к спасательным средствам.

5.Выводы:

Анализ данных пассажиров Титаника позволил сделать следующие основные выводы:

Большинство пассажиров находились в третьем классе, но их шансы на выживание были ниже.

Основная масса пассажиров была в возрасте от 20 до 40 лет.

Женщины выживали значительно чаще, чем мужчины.

Пассажиры с большим количеством братьев/сестер/супругов на борту выживали реже.

Количество родителей/детей на борту не сильно влияло на шансы выживания.

Пассажиры с более дорогими билетами выживали чаще.

Эти выводы помогают лучше понять ключевые факторы, влияющие на выживаемость пассажиров Титаника.

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование
и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа:
https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата
обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets