《数据分析与智能计算》课程报告

学号：21122926 姓名：夏磊 学院：计算机工程与科学学院

**1 需求分析**

本次课程设计，我选择了美国1994年人口普查数据库中的部分数据，用来预测居民收入（是否超过50 k）。这能够帮助我们更好的了解居民的收入水平和平均生活水平，并对居民生活水平的提升有所预测，帮助有关部门进行经济、生活等政策的指定。

**2 功能介绍**

本次数据分析的目标是分析美国1994年人口普查数据库中的部分数据，展示了该数据中的部分统计指标，并通过图形可视化的方式展示了部分变量间的关系，同时使用决策树分类对居民收入建立了机器学习模型，并在测试集上进行了性能评估。

**3 关键技术/算法说明**

本次数据分析中，我使用了函数定义的方法：

def replace\_num(a):

j = 1

for i in data[a].unique():

data.loc[data[a] == i, a] = j

j = j + 1

简化了对DataFrame中不同字符串进行量化的操作，节省了数据预处理的操作与时间。

同时，我使用了箱型图、柱状图、饼状图、折线图4种共计5个图对数据进行了可视化，其中绘制不同收入水平与学历的关系的柱状图时，我使用了子图将图分为了高收入人群学历分布与低收入人群学历分布两个部分，并通过pd.value\_counts()统计了学历分布情况，依此绘制了柱状图。

**4 完整代码**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn import model\_selection, metrics,tree

def replace\_num(a):

j = 1

for i in data[a].unique():

data.loc[data[a] == i, a] = j

j = j + 1

# 数据准备

data = pd.read\_csv('income\_census\_train.csv', index\_col=0)

data = data.dropna()

data = data.drop\_duplicates()

data.drop('fnlwgt',axis=1,inplace=True)

replace\_num('occupation')

replace\_num('relationship')

replace\_num('marital\_status')

replace\_num('workclass')

replace\_num('gender')

data.loc[data['native\_country'] != 'United-States', 'native\_country'] = 'Non-United-States'

# 数据探索

pd.set\_option('display.max\_columns', None)

print(data.describe())

data['age'].plot(kind='box', figsize=(6, 6), title='Age')

plt.show()

data['income\_bracket'].value\_counts().plot(kind='pie',startangle=45,autopct='%1.1f%%',title='Income distribution',ylabel='income\_bracket')

plt.show()

data[['hours\_per\_week', 'income\_bracket']].groupby('hours\_per\_week').aggregate(np.mean).plot(marker='o',linestyle='dashed', title='Income varied by working time',rot=0)

plt.show()

data1=data.loc[data['income\_bracket']==1,'education'].value\_counts()

data2=data.loc[data['income\_bracket']==0,'education'].value\_counts()

fig=plt.figure(figsize=(8,4))

ax1=fig.add\_subplot(1,2,1)

ax2=fig.add\_subplot(1,2,2)

data1.plot(kind='bar',ax=ax1,title='High-income',ylabel='Number of person')

data2.plot(kind='bar',ax=ax2,title='Low-income')

plt.subplots\_adjust(bottom=0.3,wspace=0.3)

plt.show()

data[['native\_country', 'income\_bracket']].groupby('native\_country').aggregate(np.mean).plot(kind='bar', title='Income varied by Native-Country',rot=0)

plt.show()

# 建模预测

replace\_num('education')

replace\_num('race')

data.loc[data['native\_country'] == 'United-States', 'native\_country'] = 1

data.loc[data['native\_country'] != 1, 'native\_country'] = 2

y = data['income\_bracket']

X = data.drop('income\_bracket',axis=1)

X\_train, X\_test, y\_train, y\_test = model\_selection.train\_test\_split(X, y, test\_size=0.35, random\_state=1)

clf=tree.DecisionTreeClassifier()

clf.fit(X\_train, y\_train)

y\_train\_pred=clf.predict(X\_train)

y\_test\_pred=clf.predict(X\_test)

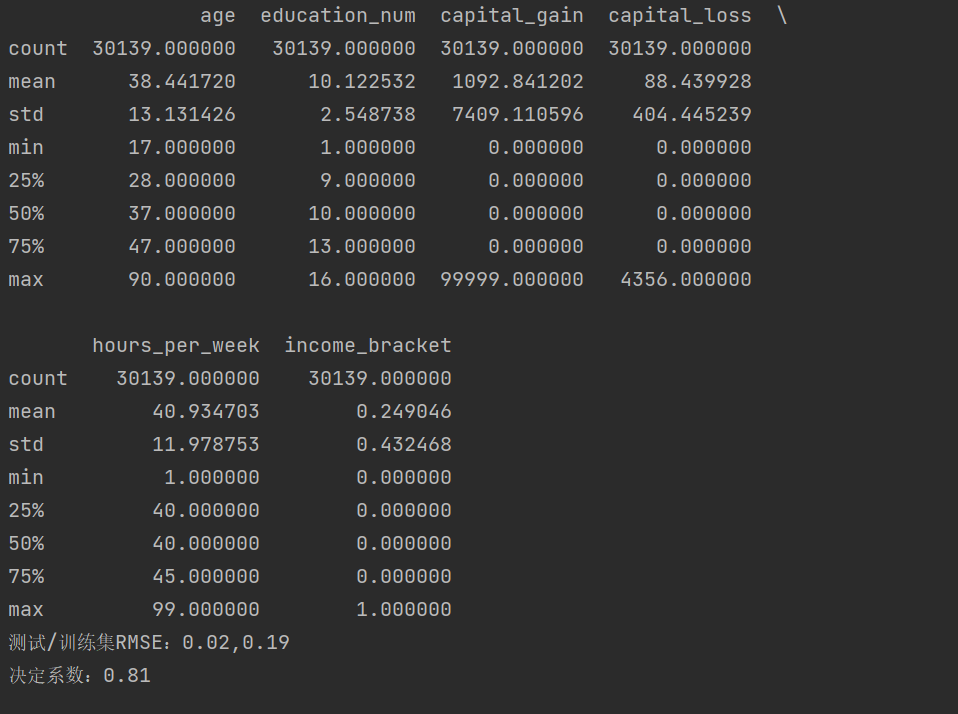
train\_err=metrics.mean\_squared\_error(y\_train,y\_train\_pred)

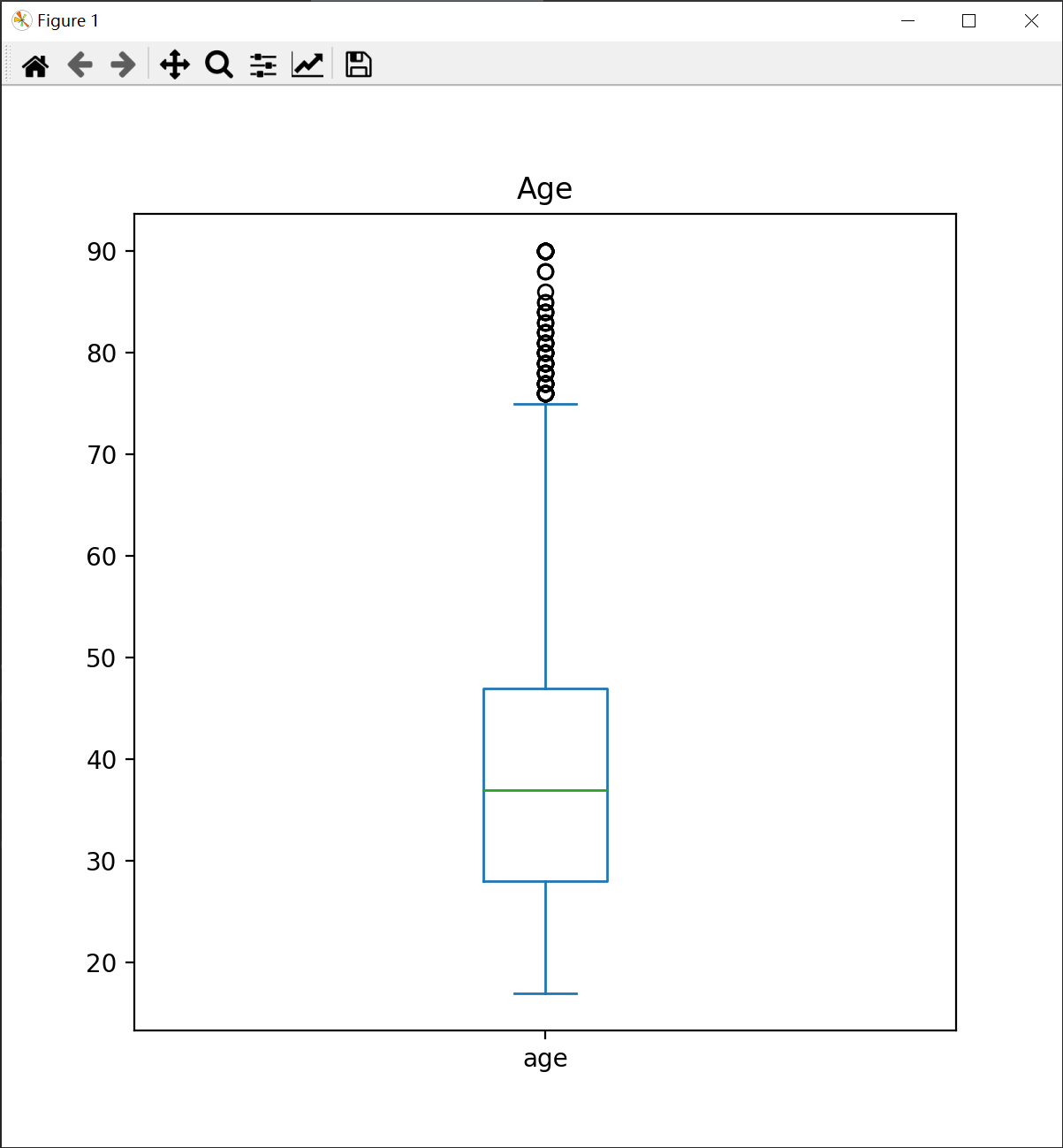
test\_err=metrics.mean\_squared\_error(y\_test,y\_test\_pred)

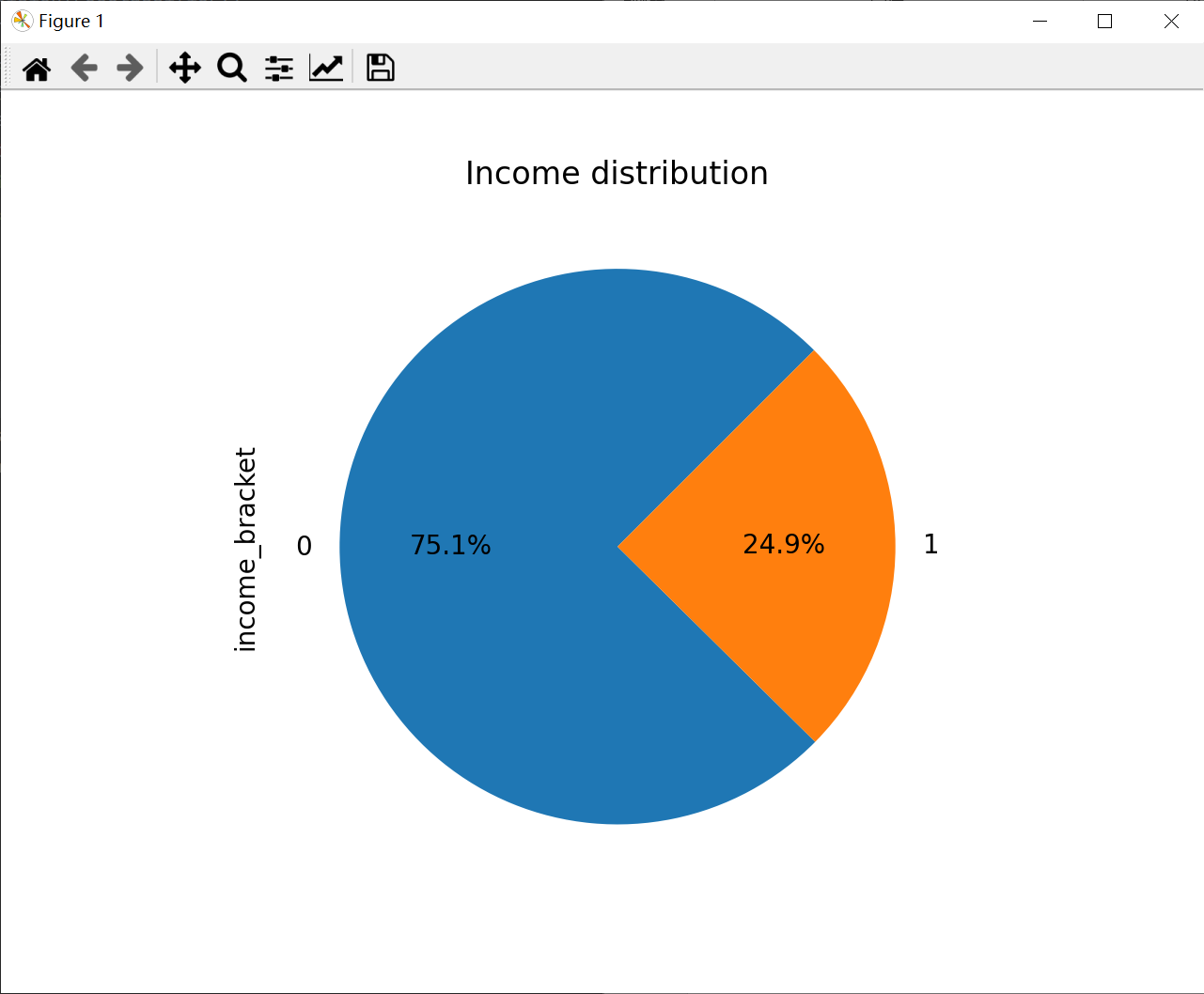
print('测试/训练集RMSE：{:.2f},{:.2f}'.format(train\_err,test\_err))

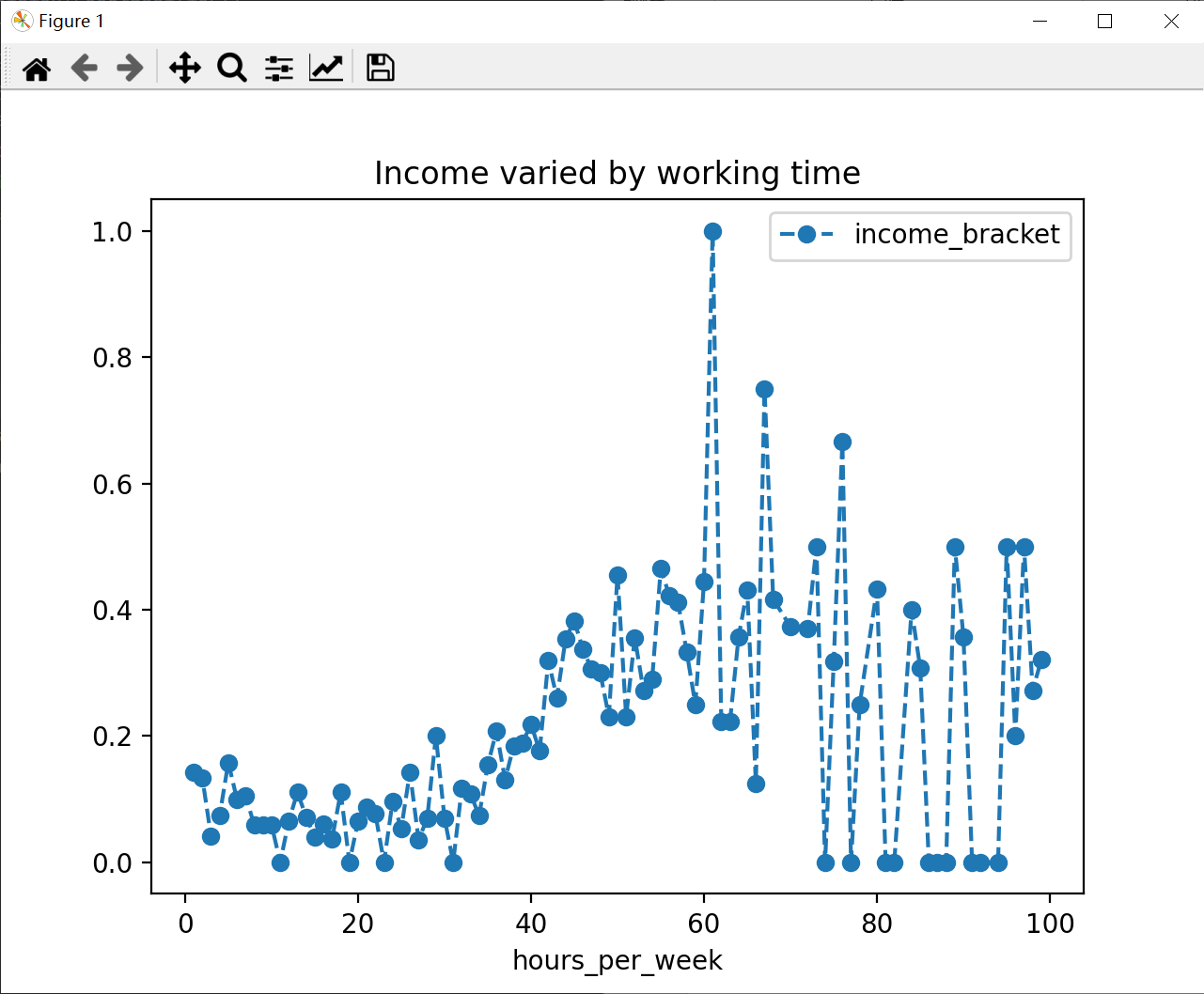
print('决定系数：{:.2f}'.format(clf.score(X\_test, y\_test)))

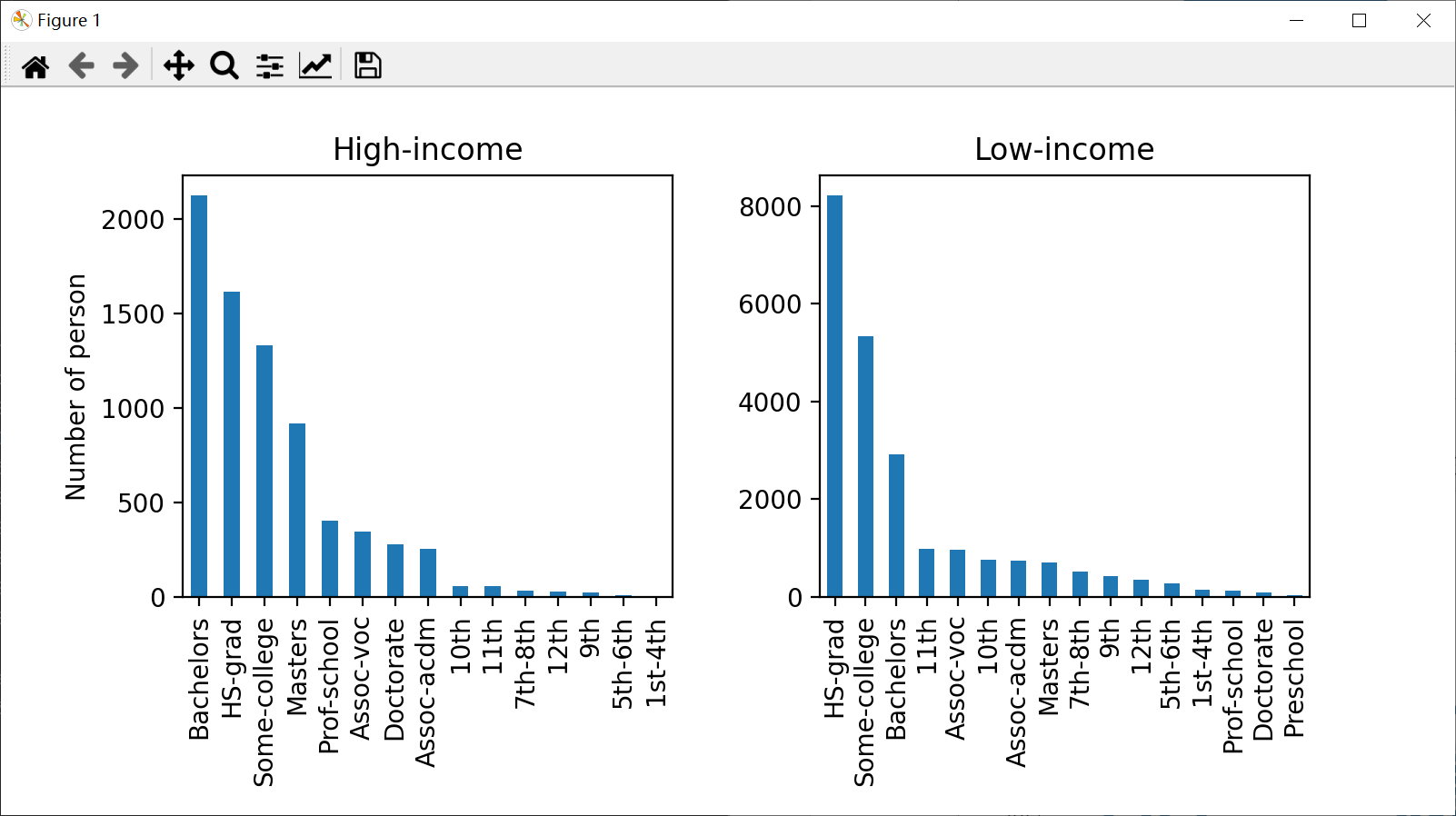
5 结果展示/性能分析

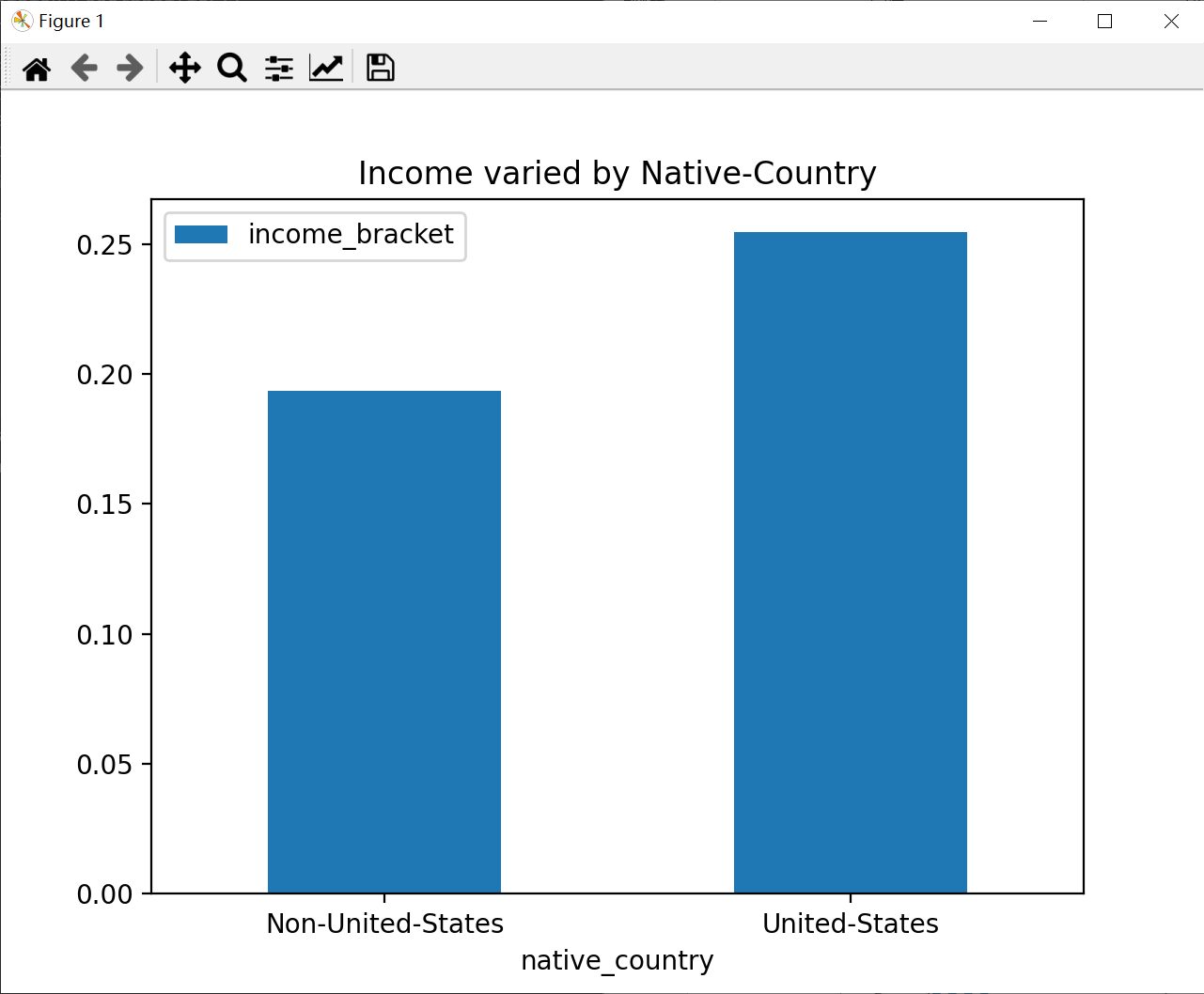












该数据分析中，首先对各数据进行了整体的统计描述，判断是否有不合理的数据列；接着在数据可视化环节中，对年龄绘制了箱型图，确定了年龄的合理性；并对收入分布绘制了饼状图，判断高收入人群分布较少；对工作时长与收入绘制了折线图，发现高收入人群主要集中于工作时长每周60-80小时的人群；对学历与收入关系绘制了两张柱状图，发现高收入人群中学士比重较高，而高中辍学人群在该组中的比重很低，低收入人群中高中学历比重较高，专科、博士及学前学历的人数则较少；最后对国籍与收入关系绘制了柱状图，发现美国人的平均收入水平比非美国人相对较高。

在机器学习建模环节中，基于训练集进行了建模，并在测试集上进行了测试，最后计算了RMSE与决定系数，计算得出测试/训练集RMSE：0.02,0.19，决定系数：0.81，故该模型拟合程度较好。

**6 认识与思考**

我认为，人工智能是通过机器的强大算力，为人类提供预测与重复性劳动的机器，同时具备一定的自主学习能力。该领域的研究包括机器人、语言识别、图像识别、自然语言处理、博弈、人工神经网络等。如今，像小爱同学、天猫精灵等人工智能产品已经深入千家万户，而机器翻译、个性化推荐、医学图像处理等技术也已经应用于我们的生活。目前，人工智能的主要目的是用来处理庞大的数据以及重复与大量的工作，它是出色的执行者，但不足以称为拥有自主意识的领导者。它替代劳动的能力将是空前的，但在智能原理发现前，未来的世界仍不足以由人类和人工智能共同承担，但我认为这并不会太过遥远。可当未来的人工智能发展出自我，人类也必须站在伦理角度思考该如何对待它们。