# 实验 1 描述性统计

关于本次实验中用到的有关中风数据集（stroke.csv）的说明:

Context（背景）

According to the World Health Organization(WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11%of total deaths.This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relavant information about the patient.

Attribute Information（特征列信息）

1) **id**: unique identifier
2) **gender**: "Male", "Female" or "Other"
3) **age**: age of the patient
4) **hypertension**（高血压）: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) **heart_disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) **ever_married**: "No" or "Yes"
7) **work_type**: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) **Residence_type**: "Rural" or "Urban"
9) **avg_glucose_level**（平均血糖水平）: average glucose level in blood
10) **bmi**: body mass index
11) **smoking_status**: "formerly smoked", "never smoked", "smokes" or "Unknown"
*Note: "Unknown" in smoking_status means that the information is unavailable for this patient
12) **stroke**: 1 if the patient had a stroke or 0 if not

## 一、（必做）数值型描述性统计

1.请根据数据集 stroke.csv，完成中风患者（stroke 特征列指是否中风，1 为中风，0 为未中风）年龄的平均值,中位数，第 25 百分位数，众数的相关程序代码。

```
import pandas as pd
import numpy as np
import os
print(os.getcwd())    #当前工作目录
my_data=pd. ________("stroke.csv")      # stroke 文件必须先拷贝到当前工作目录中
stro_ple=my_data[my_data.stroke==________]#根据 stroke 特征值筛选出中风人群（stro_ple）
print('中风患者年龄的位置性测度:')
stro_age=stro_ple.age#取出中风人群的年龄
print('均值：\t\t',stro_age.________)
print('中位数：\t',________.median()[0])
print('第 25 百分位数：\t',stro_ple[['age']].quantile(q=________)[0])#注意此处获取中风人群年龄的另一种方法
print('众数：\t\t',________.mode().values[________])
```

结果对照：

```
中风患者年龄的位置性测度：
均值：          67.72819277108434
中位数：        71.0
第25百分位数：  59.0
众数：          78.0
```

2.完成对中风患者的 avg_glucose_level（平均血糖水平）的离散性测度统计(包含方差,标准差,变异系数)。

```
print('对平均血糖水平的离散性测度统计结果：')
print('方差：\t\t',stro_ple[['avg_glucose_level']]._____[0])
print('标准差：\t',_____.std()[0])
print('变异系数：\t',
stro_ple[['avg_glucose_level']]. _____[0]/
stro_ple[['avg_glucose_level']]._____[0])
#注意上面三行为一个整句，代码写在一行
```

结果对照：

```
对平均血糖水平的离散性测度统计结果：
方差：          3834.2171242259365
标准差：        61.92105558068222
变异系数：      0.46717097991584794
```

## 二、（必做）可视化描述性统计

1.为了更加明了地看出患者的年龄分布，请绘制以患者年龄为特征的直方图(dataframe.hist())，x 轴、y 轴标注尺寸均为 16，图片大小为长 10，宽 8，分成 10 个区间。

hist()函数参数补充说明：

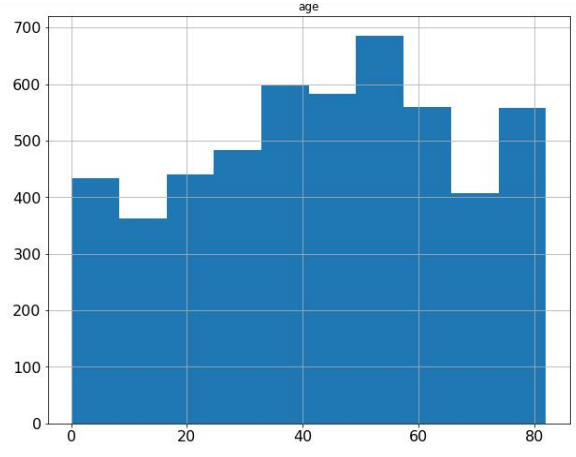
bins:类型是 int 或者 sequence，就是指定显示多少竖条或者多少个区间，默认为 10。

xlabelsize：类型 int 如果指定了这个值，则可以改变 x 轴的标注尺寸,ylabelsize 类似

figsize：类型是（tuple），单位是英寸，表示要创建的图的大小，即长，宽。

```
my_data[['age']].hist(bins=_____,figsize=(_____,_____),xlabelsize=_____,yla
belsize=_____)
```

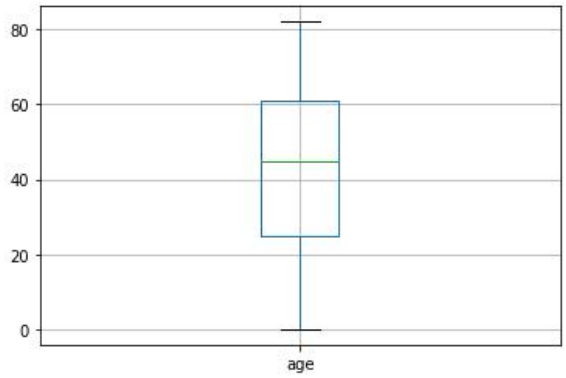
结果图：



2. 为了直观的看出患者年龄的中位数，第一四分位数和最大、最小值等，请绘制以患者年龄为特征的箱型图，了解箱型图每一位置的含义并想一想为什么此图没有离群点（异常值）。

```
my_data[['age']].________    #绘制箱型图
```

结果图:

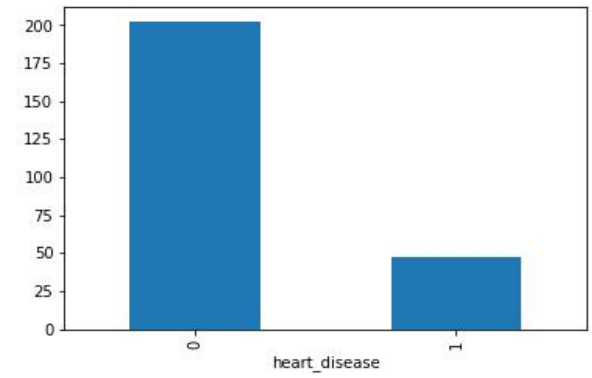



3. 简单的考虑中风和心脏病的联系，根据是否患有心脏疾病（即 heart_disease=1 或 0）来分组，看中风患者中患有心脏疾病和不患心脏疾病的数量各有多少，并绘制直方图。

```
my_plot_data=stro_ple[['heart_disease']].________(['heart_disease']).size()#对heart_disease 进行分组
print(my_plot_data)
my_plot_data.plot(kind='________')     #绘制柱状图
```

结果：

```
heart_disease
0    202
1     47
dtype: int64

<AxesSubplot:xlabel='heart_disease'>
```



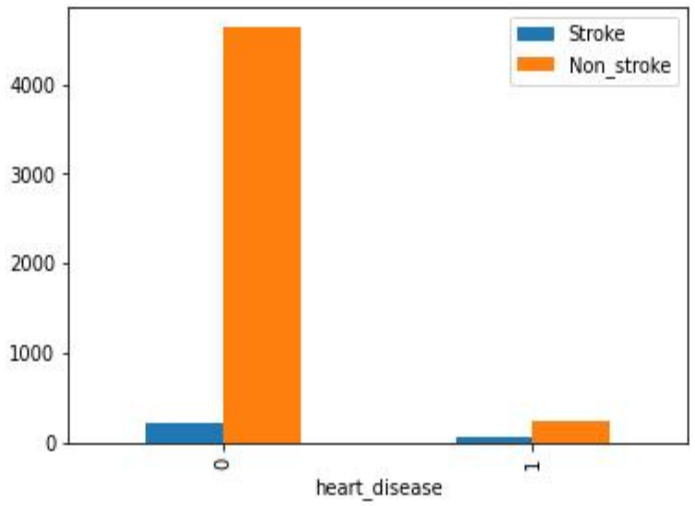

**三、（必做）**进一步了解患心脏病和中风人数的关系，进行分组统计，你能得到什么结论？有心脏

病的人更容易中风还是反之？

```
nstro_ple=my_data[_____]#dataframe 中对于未中风患者（nstro_ple）的筛选
hd_stro=_____[['heart_disease']]._____(['heart_disease']).size()#中风患者中是否患有心脏病的分组人数(hd_stro)
print('中风患者中是否患有心脏病的分组人数')
print(hd_stro,'\n')
hd_nostro=_____.size()#未中风患者中是否患有心脏病进行分组人数（hd_nostro）
print('未中风患者中是否患有心脏病进行分组人数')
print(hd_nostro,'\n')
tol_data=pd.concat([_____, _____],axis=1)#将两组分组数据进行连接
my_plot_data=tol_data.rename(columns={0:'Stroke',1:'Non_stroke'})#将连接后的的数据列重命名
print(my_plot_data)
my_plot_data.plot(kind='bar')#绘制相应的条形图
```

结果图：

```
中风患者中是否患有心脏病的分组人数
heart_disease
0    202
1     47
dtype: int64

未中风患者中是否患有心脏病进行分组人数
heart_disease
0    4632
1     229
dtype: int64

               Stroke  Non_stroke
heart_disease
0                 202        4632
1                  47         229
```




四、 **（必做）**根据 diabetes.csv 文件，我们想知道 BMI 和 BloodPressure 对糖尿病相关性的影响，先根据提示完成对缺失值的处理，然后再完成散点图的绘制，并了解散点图各个颜色的含义。（散点图可视化相关参数和枚举函数用法请参考文末）

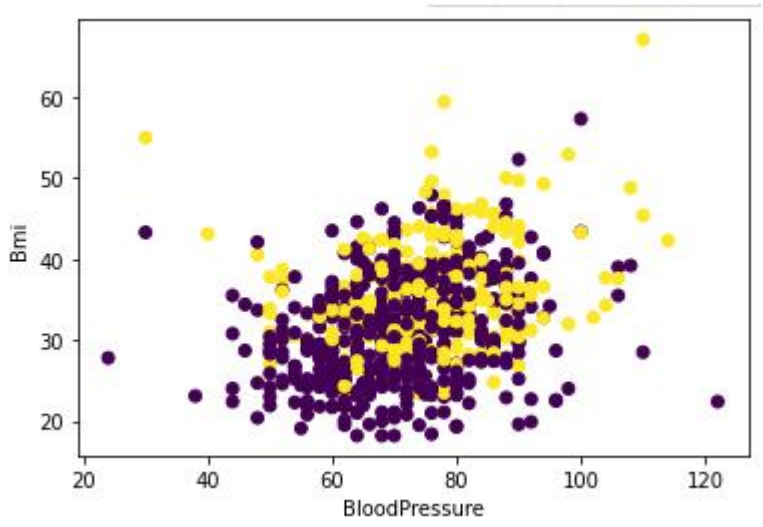
## Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**Pregnancies：** Number of times pregnant

**Glucose（血糖）：** Plasma glucose concentration a 2 hours in an oral glucose tolerance test
**BloodPressure：** Diastolic blood pressure (mm Hg)
**SkinThickness（肱三头肌皮肤褶皱厚度）：** Triceps skin fold thickness (mm)
**Insulin（胰岛素）：** 2-Hour serum insulin (mu U/ml)
**BMI：** Body mass index (weight in kg/(height in m)^2)
**DiabetesPedigreeFunction（糖尿病谱系功能）：** Diabetes pedigree function
一种根据家族史对糖尿病可能性进行评分的功能，综合研究对象的亲属糖尿病史及亲属间的遗传关系研究中使用的一个特别有趣的属性，这种遗传影响的度量使我们对糖尿病发病可能存在的遗传风险有了一个认识。
**Age:** Age (years)
**Outcome:** Class variable (0 or 1) 268 of 768 are 1, the others are 0

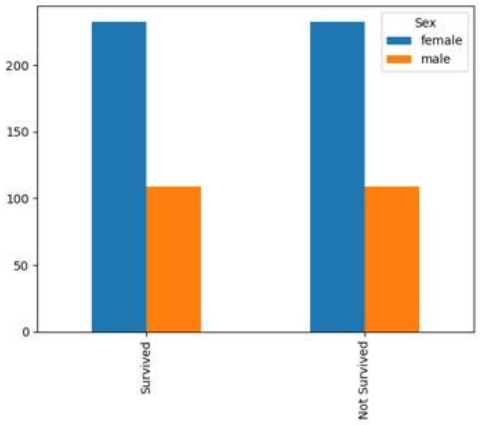
```
import pandas as pd
from matplotlib import pyplot as plt
my_data1=pd.read_csv("diabetes.csv",usecols=['BloodPressure','BMI','Outcome'])  #只读取这三列
my_data1._____    #查看数据的总记录数以及缺失情况

my_fil_data1=my_data1._____  #由于有缺失值的样本数量不多，所以丢弃这些样本

x1=my_fil_data1.BloodPressure   #横坐标
y1=my_fil_data1.BMI             #纵坐标
plt._____(x1,y1,c=my_fil_data1.Outcome)     #根据 Outcome 给散点染色
plt.gca().set_xlabel('BloodPressure')
plt.gca()._____('BMI')
plt.show
```



五、选做题（共 3 小题，提交到学习通选做题中）

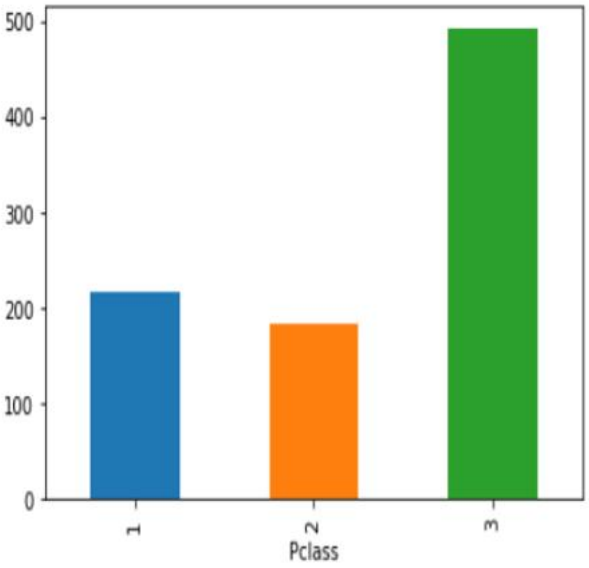
# 作业1

根据Titanic.csv，做一个横轴按是否生还分为两组，纵轴是两组各性别人数的柱状图（参考下图）




# 作业2

根据Titanic.csv，人数最少的是二等舱而不是一等舱，有一种解释是因为一等舱活下来的人更多，所以留下的资料就更全。
请做一个描述性统计，来证实或证伪这种说法，并给出你的结论。



# 作业3

根据Titanic.csv，diabetes.csv或stroke.csv进行探索性数据分析，看能不能有一些发现。

作业1-3可以做到一个ipynb文件中