Dataset: heart.csv <https://www.kaggle.com/ronitf/heart-disease-uci>

Columns:

* age: The person's age in years
* sex: The person's sex (1 = male, 0 = female)
* cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
* trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
* chol: The person's cholesterol measurement in mg/dl
* fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
* restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
* thalach: The person's maximum heart rate achieved
* exang: Exercise induced angina (1 = yes; 0 = no)
* oldpeak: ST depression induced by exercise relative to rest
* slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
* ca: The number of major vessels (0-3)
* thal: A blood disorder called thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
* target: Heart disease (0 = no, 1 = yes)

Tasks:

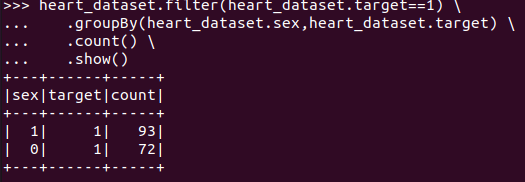
1. **Which gender is more vulnerable to heart disease**

>>> gender=heart\_dataset.filter(heart\_dataset.target==1) \

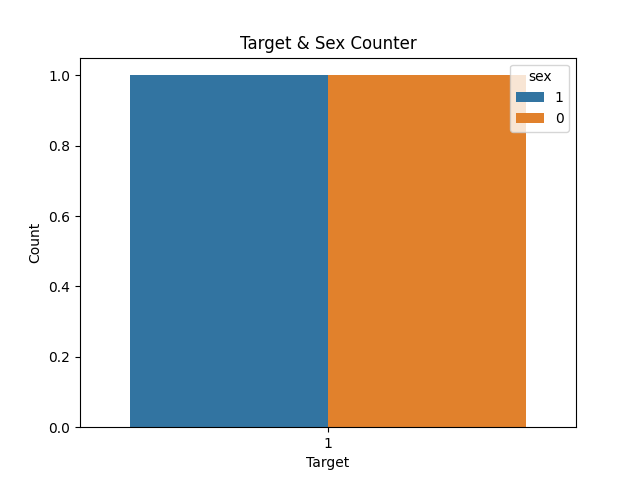
... .groupBy(heart\_dataset.sex,heart\_dataset.target) \

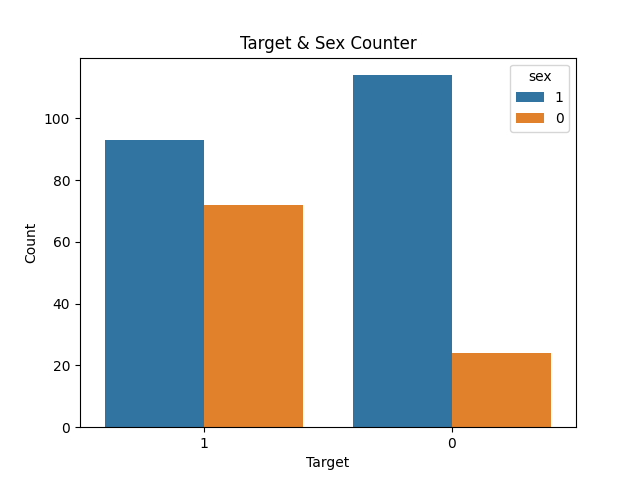
... .count() \

... .show()



pd\_gender = gender.toPandas()  
  
sns.countplot(pd\_gender.target, hue=pd\_gender.sex)  
plt.xlabel('Target')  
plt.ylabel('Count')  
plt.title('Target & Sex Counter')  
plt.show()





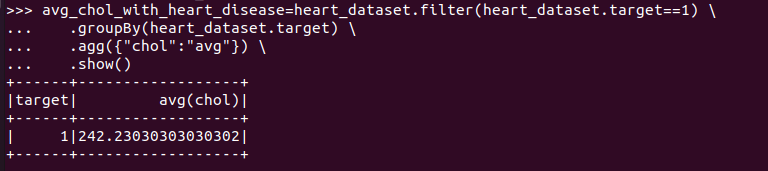
1. **What is the average cholesterol level for people with heart disease?**

>>> avg\_chol\_with\_heart\_disease=heart\_dataset.filter(heart\_dataset.target==1) \

... .groupBy(heart\_dataset.target) \

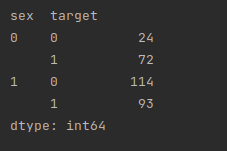
... .agg({"chol":"avg"}) \

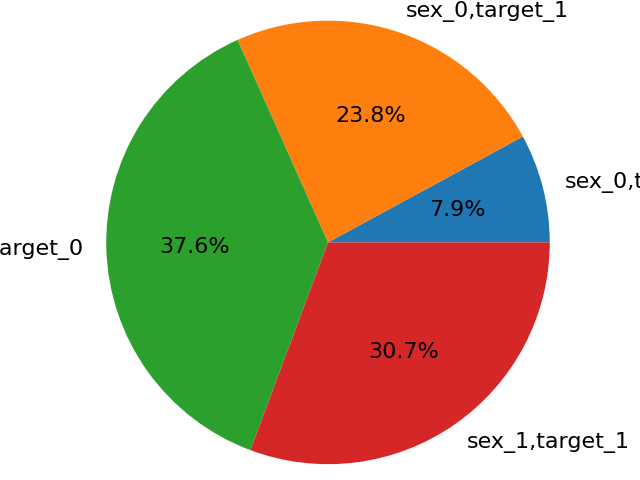
... .show()



1. **Compute the number of people with and without heart disease on the basis of gender and display them in a pie chart.**

pd\_df = heart\_dataset.toPandas()  
  
df\_sex = pd\_df.groupby(["sex", "target"]).size()   
print(df\_sex)  
  
plt.pie(df\_sex.values, labels = ["sex\_0,target\_0", "sex\_0,target\_1", "sex\_1,target\_0", "sex\_1,target\_1"],autopct='%1.1f%%',radius = 1.5, textprops = {"fontsize" : 16})  
plt.show()

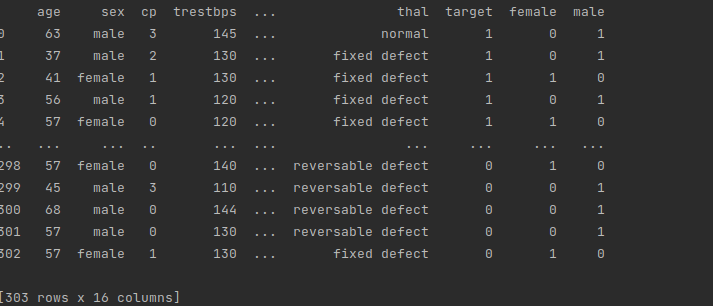




1. **Replace the name of category data and separate them into different columns such that 1=True and 0=False.**

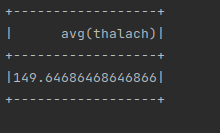
**Eg. sex column into sex\_male and sex\_female**

df.sex = df.sex.map({1:"male", 0:"female"})dummy=pd.get\_dummies(df['sex'])print(dummy)df=pd.concat([df,dummy], axis=1)print(df)

****

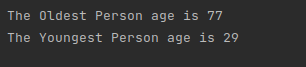
1. **What is the average maximum heart rate achieved by admitted patients?**

avg\_thalach = heart\_dataset.agg({"thalach": "avg"}) \  
 .show()



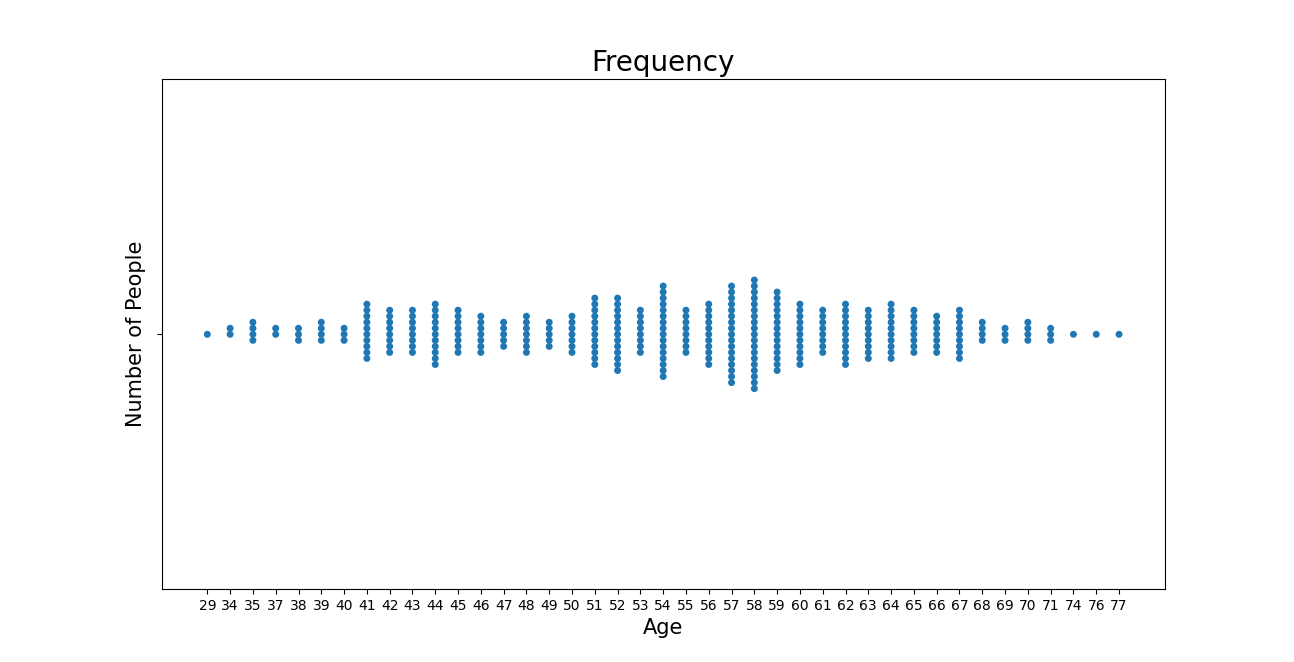
1. **What is the age of the youngest and oldest patient admitted for a check on cardiovascular disease.**

print('The Oldest Person age is', pd\_df.age.max())  
print('The Youngest Person age is', pd\_df.age.min())



1. **Does age of patient affect the heart disease on the basis of given dataset (hint: use swarmplot)**

plt.figure(figsize = (15,8))  
sns.swarmplot(x = 'age',data = pd\_df)  
plt.xlabel('Age', fontsize = 15)  
plt.ylabel('Number of People', fontsize = 15)  
plt.title('Frequency', fontsize = 20)  
plt.show()



1. **What type of chest pain in patients are more susceptible to heart disease.**

cp\_data = (pd\_df.groupby(['target']))['cp'].value\_counts(normalize=True)\  
 .mul(100).reset\_index(name = "percentage")  
  
sns.barplot(x = "target", y = "percentage", hue = "cp", data = cp\_data)  
plt.title("Chest Pain types with Heart Disease")  
plt.show()

