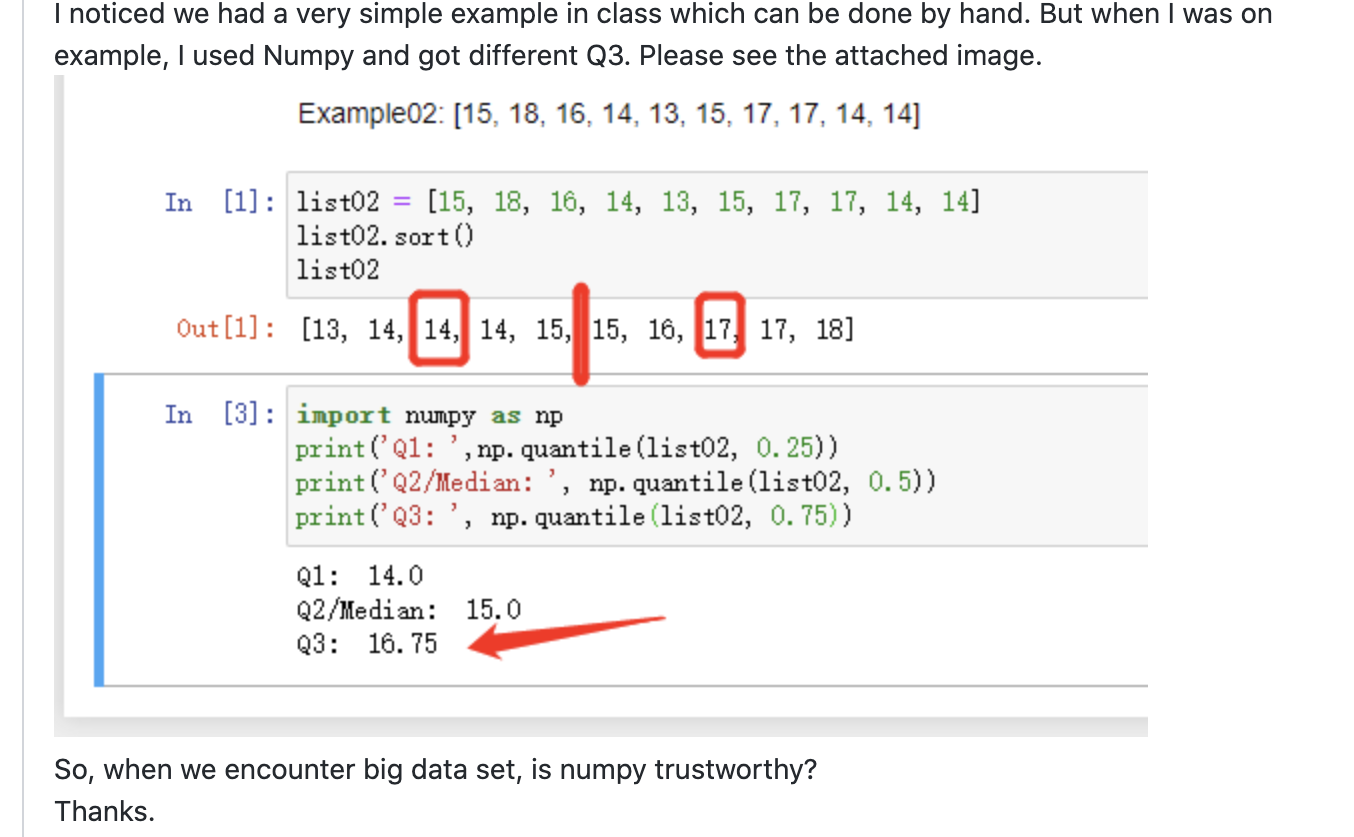
**Q & A**

1. Numpy Q3 question:



<https://www.mathsisfun.com/data/quartiles.html>

|  |  |  |  |
| --- | --- | --- | --- |
| Value | Percentile | Quantile | Interval(%) |
| 13 | 10% |  | 0-10 |
| 14 | 20% | Q1 | 10-20 |
| 14 | 30% | Q1 | 20-30 |
| 14 | 40% |  | 30-40 |
| 15 | 50% | Q2 | 40-50 |
| 15 | 60% | Q2 | 50-60 |
| 16 | 70% | Q3 | 60-70 |
| 17 | 80% | Q3 | 70-80 |
| 17 | 90% |  | 80-90 |
| 18 | 100% |  | 90-100 |

Numpy logic is looking for >70% and <80%, this is totally due to the discrete distribution, while you should also pay attention for discrete is an interval or range of distribution







Class1 homework summary:

**HW1:**

Read the data in pandas, and perform exploratory data analysis for all variables in the data:

1. Show all data types, missing rates:

df.info()

df.dtypes

State 50 non-null object

Population 50 non-null int64

Murder.Rate 50 non-null float64

Abbreviation 50 non-null object

Missing Rate

df.isnull().mean()

df.isnull().sum()/length(df)

1. Numerical variables provide mean, SD, five number summaries, histogram

df.describe()

1. Categorical variables provide frequency table and bar chart

df[variablename].value\_counts()

1. Perform at least 2 pairs of variables for 2 types of bivariate charts
2. Provide a summary to describe what you know about your data in this practice.

**HW2: df.sample usage and practice boot-strip getting the sense of CLT**

Np.random.choice

pd.plotting.bootstrap\_plot(data, size=1000)

**HW3: Answer**

A rental car company claims the mean time to rent a car on their website is 60 seconds with a standard deviation of 30 seconds. A random sample of 36 customers attempted to rent a car on the website.

1) If the mean time to rent was 65 seconds. Is this enough evidence to contradict the company's claim at 5% confidence level? What's the power?

2) If the analysts team believe the claim is not correct and the time for rent is 67 seconds, what's the sample size needed to design for a sufficient test with 85% of power

1)

Two side tests, because in the problem description, there is no significant expectation for the test to provide evidence to be larger or smaller than the claimed number. (In real world application, recommended to talk to you customer, to align your test with their expectation)

Z-test

A picture containing object, clock

Description generated with very high confidence

(65-60)/(30/sqrt(36)) = 1 < 1.96 = Z0.975

Therefore, there is not enough evidence to contradict the company claim of 60 seconds car renting average time.

Follow class1 slides page 52 formula to calculate:

Power = Pr{Z<-1.96-sqrt(36)\*(65-60)/30} + Pr(Z>1.96-sqrt(36)\*(65-60)/30)

= Pr{Z<-2.96-} + Pr(Z>0.96) = 0.00145 + 0.16853 = 0.16998

2)

Then this choice will be right-sided with design sample size calculation

Power = Pr(Z>Z0.95 - sqrt(n)\*(67-60)/30)) = 0.85

Pr(Z>1.645 - sqrt(n)\*(67-60)/30)) = 0.85

1.645 - sqrt(n)\*(7/30)) = -1.035

N= sq((1.645+1.035) \* 30/7) = 132