

## Exploratory data analysis

Quiz, 4 questions

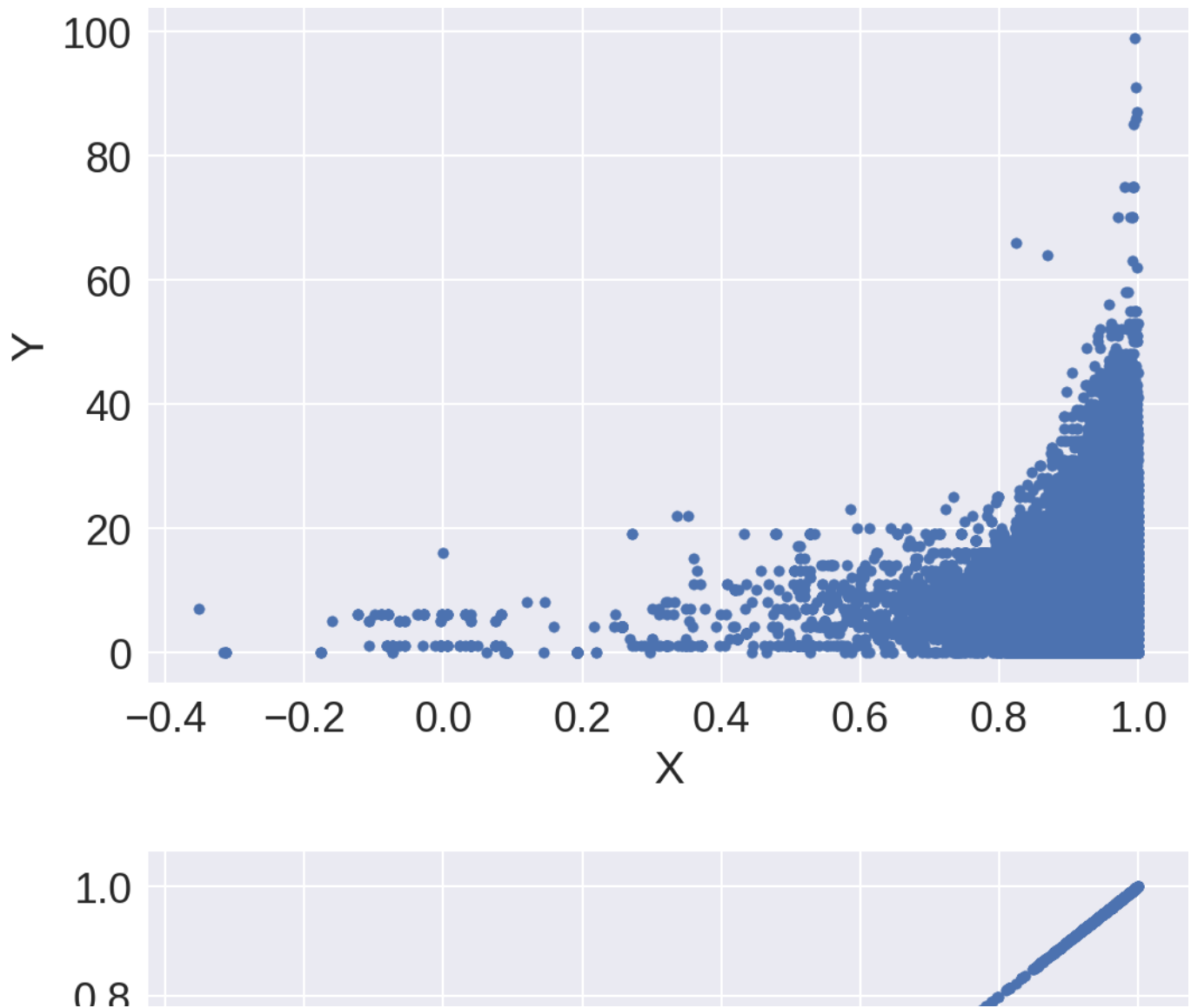
✓ **Congratulations! You passed!**

Next Item



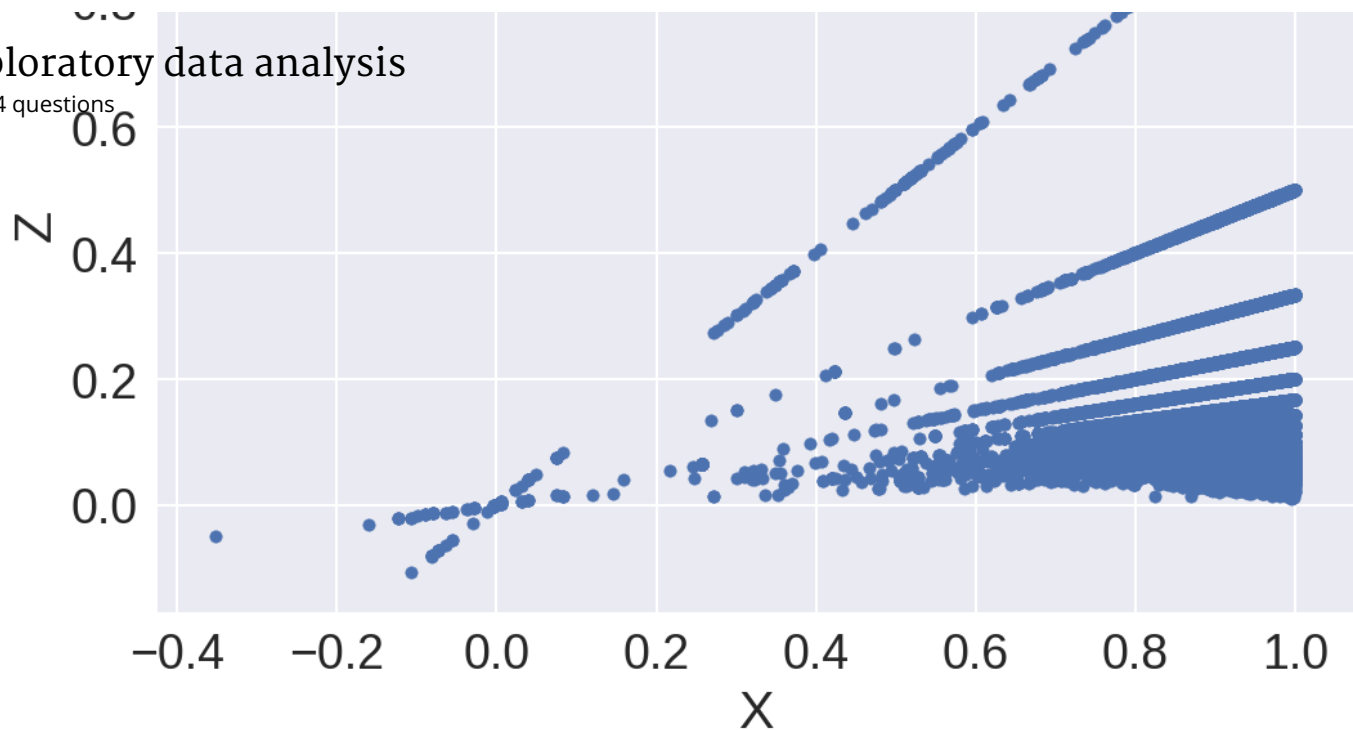
2 / 2  
points

1.



## Exploratory data analysis

Quiz, 4 questions



Suppose we are given a data set with features  $X$ ,  $Y$ ,  $Z$ .

On the top figure you see a scatter plot for variables  $X$  and  $Y$ . Variable  $Z$  is a function of  $X$  and  $Y$  and on the bottom figure a scatter plot between  $X$  and  $Z$  is shown. Can you recover  $Z$  as a function of  $X$  and  $Y$ ?

☐  $Z = X - Y$

☒  $Z = X/Y$



**Correct**

Correct!

☐  $Z = XY$

☐  $Z = X + Y$

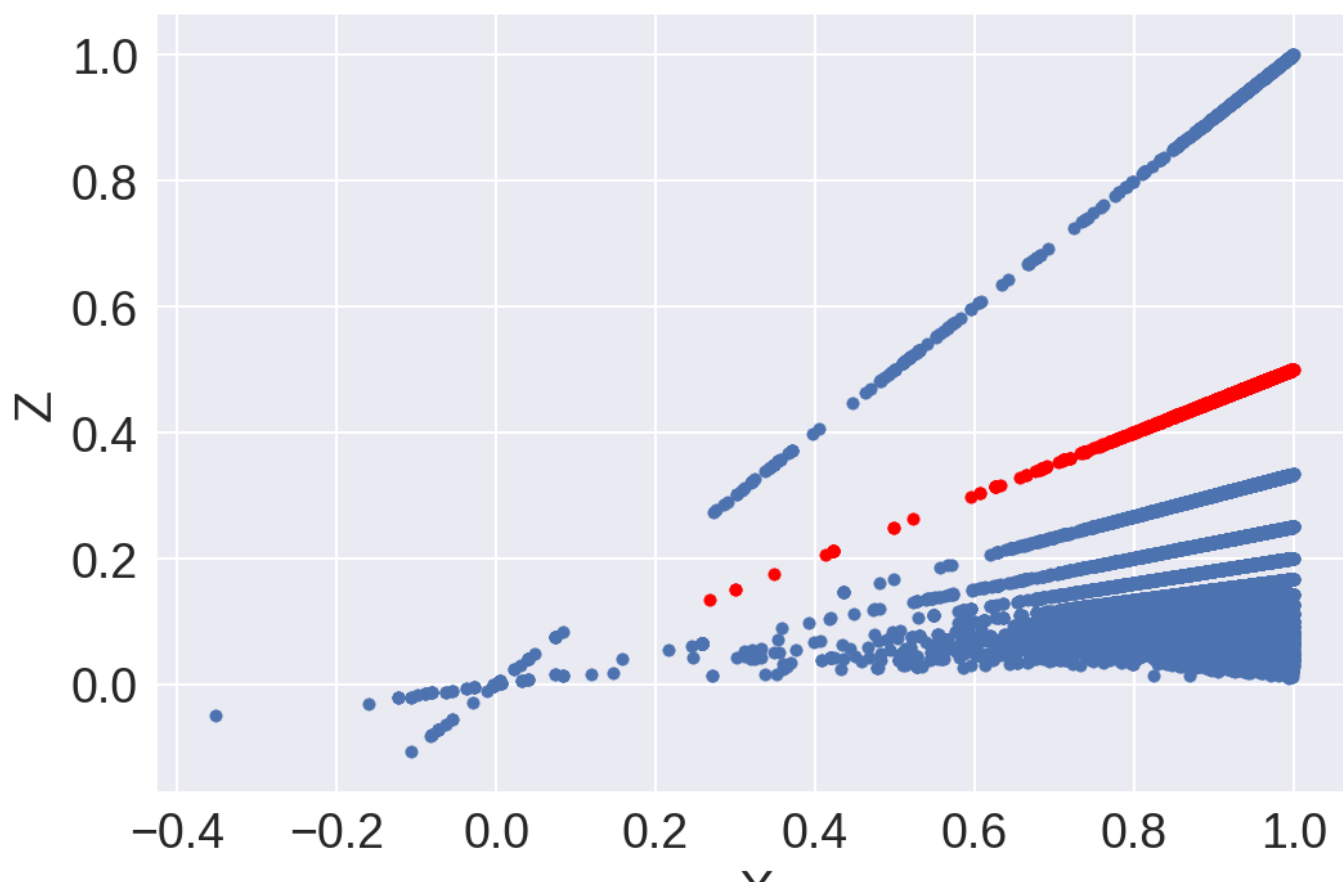
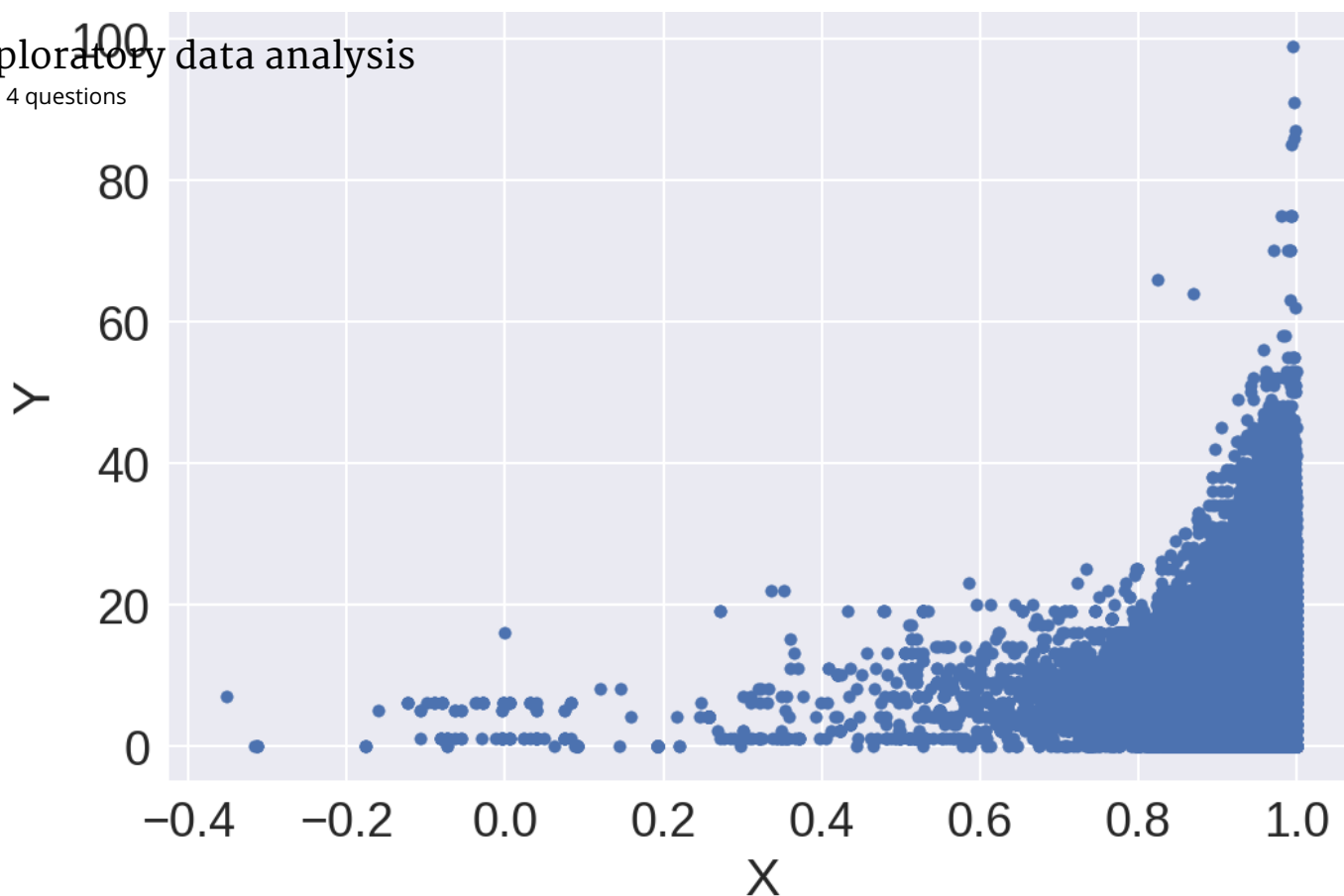


2 / 2  
points

2.

## Exploratory data analysis

Quiz, 4 questions



X

## Exploratory data analysis

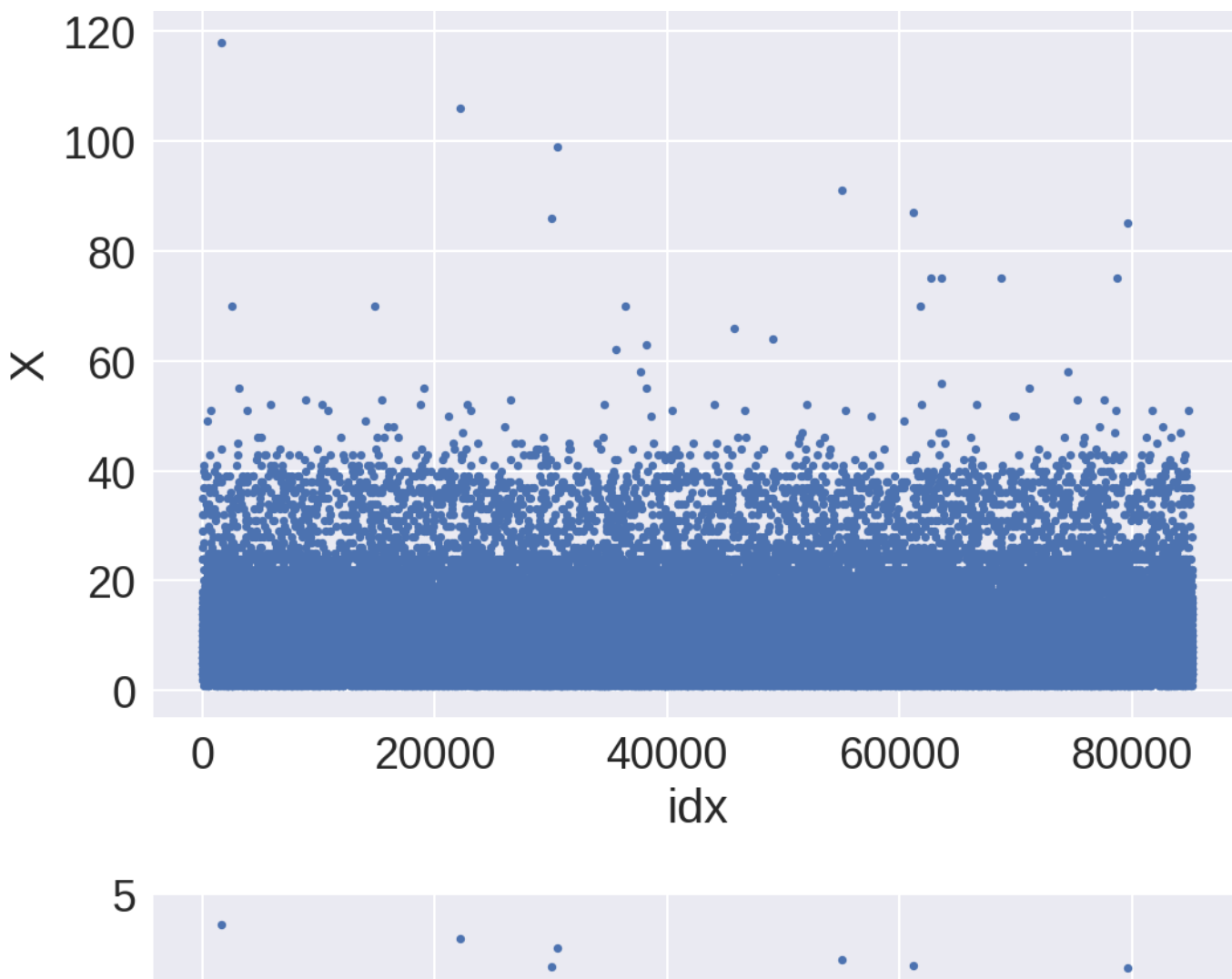
Quiz, 4 questions  
What Y value do the objects colored in red have?

### Correct Response

The equation for a line, built through red points is  $Z = X/2$ , now recalling that  $Z = X/Y$  we conclude  $Y = 2$ .

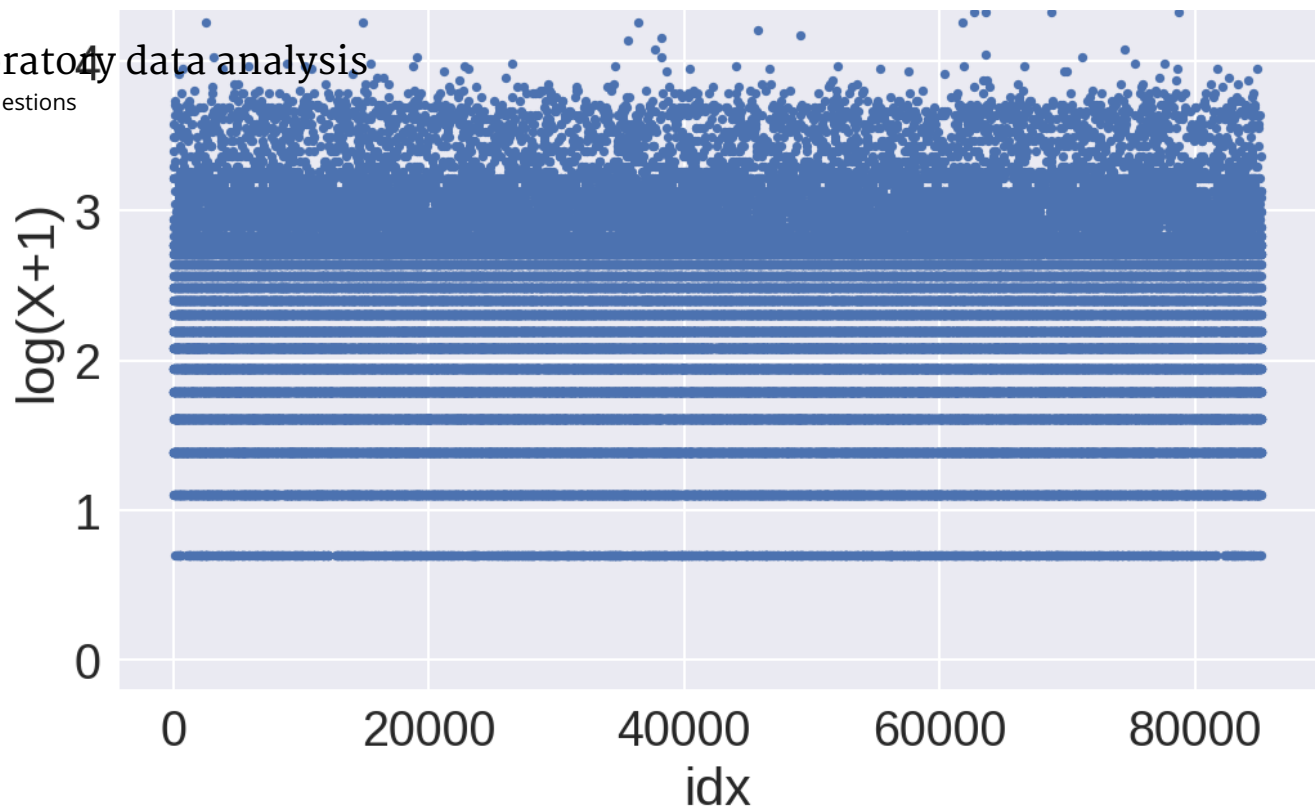
1.60 / 2  
points

3.



## Exploratory data analysis

Quiz, 4 questions



The following code was used to produce these two plots:

```
1 # top plot
2 plt.plot(x, '.')
3
4 # bottom plot
5 logX = np.log1p(x) # no NaNs after this operation
6 plt.plot(logX, '.')
```

(note that it is not the same variable  $X$  as in previous questions).

Which hypotheses about variable  $X$  do NOT contradict with the plots? In other words: what hypotheses we can't reject (not in statistical sense) based on the plots and our intuition?

☐

$X$  can be the temperature (in Celsius) in different cities at different times



**Un-selected is correct**

☐

$X$  is a counter or label encoded categorical feature



**Correct**

Yes! The values are integers and start from 1. It could be e.g. a counter how many times a used opened web-site. Or it could be a categorical features encoded with label encoder, which starts with label 1 (in pandas and sklearn label encoders usually start with 0).

## Exploratory data analysis

Quiz, 4 questions ☒  $X$  can take a value of zero

Un-selected is correct

☐  $X$  takes only discrete values

This should be selected

☒  $2 \leq X < 3$  happens more frequently than  $3 \leq X < 4$

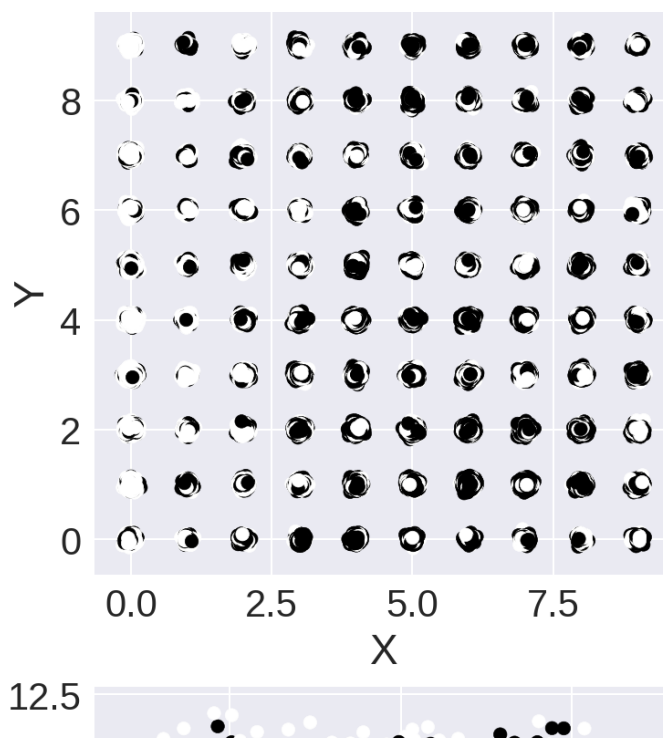
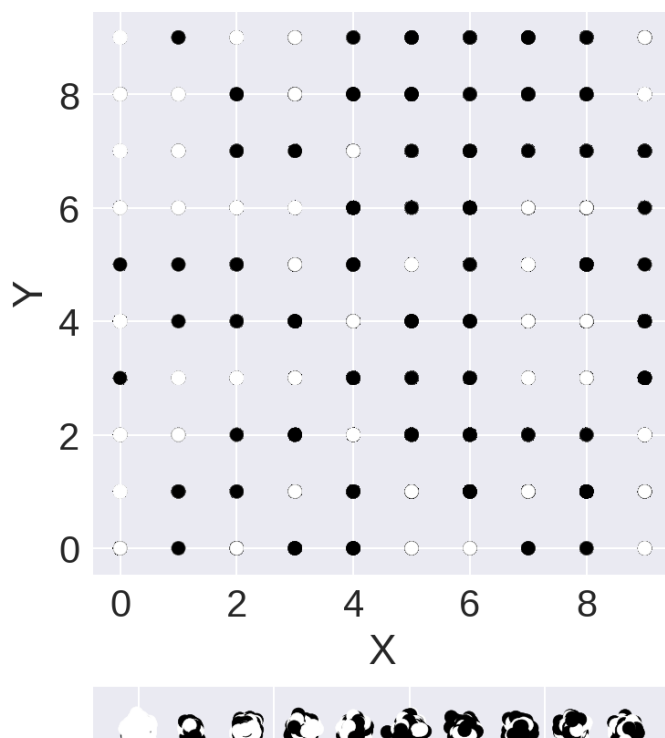
Correct

Yes! It can be the case, we cannot understand it from these plots, more exploration is needed, but such hypothesis does not contradicts with the plots.



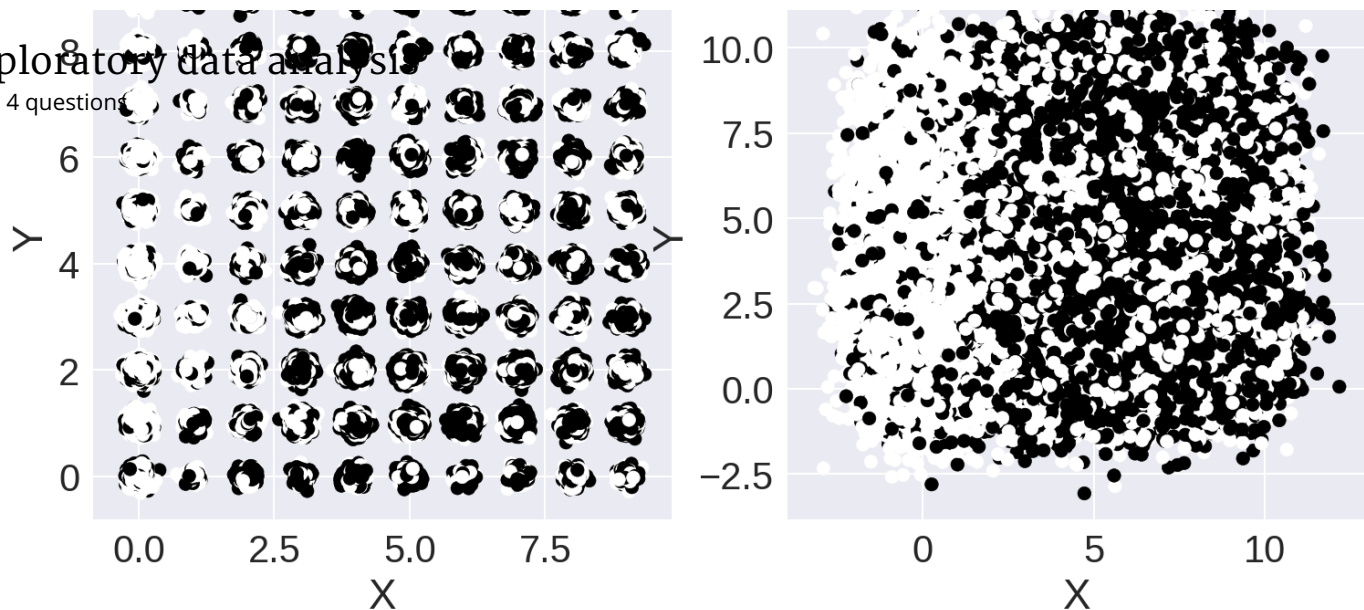
2 / 2  
points

4.



## Exploratory data analysis

Quiz, 4 questions



Suppose we are given a dataset with features  $X$  and  $Y$  and need to learn to classify objects into 2 classes. The corresponding targets for the objects from the dataset are denoted as  $y$ .

Top left plot shows  $X$  vs  $Y$  scatter plot, produced with the following code:

```
1 # y is a target vector
2 plt.scatter(X, Y, c = y)
```

We use target variable  $y$  to colorcode the points.

The other three plots were produced by *jittering*  $X$  and  $Y$  values:

```
1 def jitter(data, stdev):
2     N = len(data)
3     return data + np.random.randn(N) * stdev
4
5 # sigma is a given std. dev. for Gaussian distribution
6 plt.scatter(jitter(X, sigma), jitter(Y, sigma), c = y)
```

That is, we add Gaussian noise to the features before drawing scatter plot.

Select the correct statements.



Standard deviation for Jittering is the largest on the bottom right plot.



**Correct**

Yes! We can't even see, that  $X$ ,  $Y$  originally have small number of unique values.

## Exploratory data analysis

Quiz, 4 questions ☒ It is *always* beneficial to jitter variables before building a scatter plot

Un-selected is correct



We need to jitter variables not only for a sake of visualization, but also because it is beneficial for a model.

Un-selected is correct



Top right plot is "better" than top left one. That is, every piece of information we can find on the top left we can also find on the top right, but not vice versa.

Correct

Yes! On the top left plot we only see, that pairs  $(x, y)$  lie on the grid. Top right also shows target distribution for each  $(x, y)$  and density in  $(x, y)$ .



Target is completely determined by coordinates  $(x, y)$ , i.e. the label of the point is *completely determined* by point's position  $(x, y)$ . Saying the same in other words: if we only had two features  $(x, y)$ , we could build a classifier, that is accurate 100% of time.

Un-selected is correct

