Congratulations! You passed!

TO PASS 75% or higher

Keep Learning

GRADE
100%

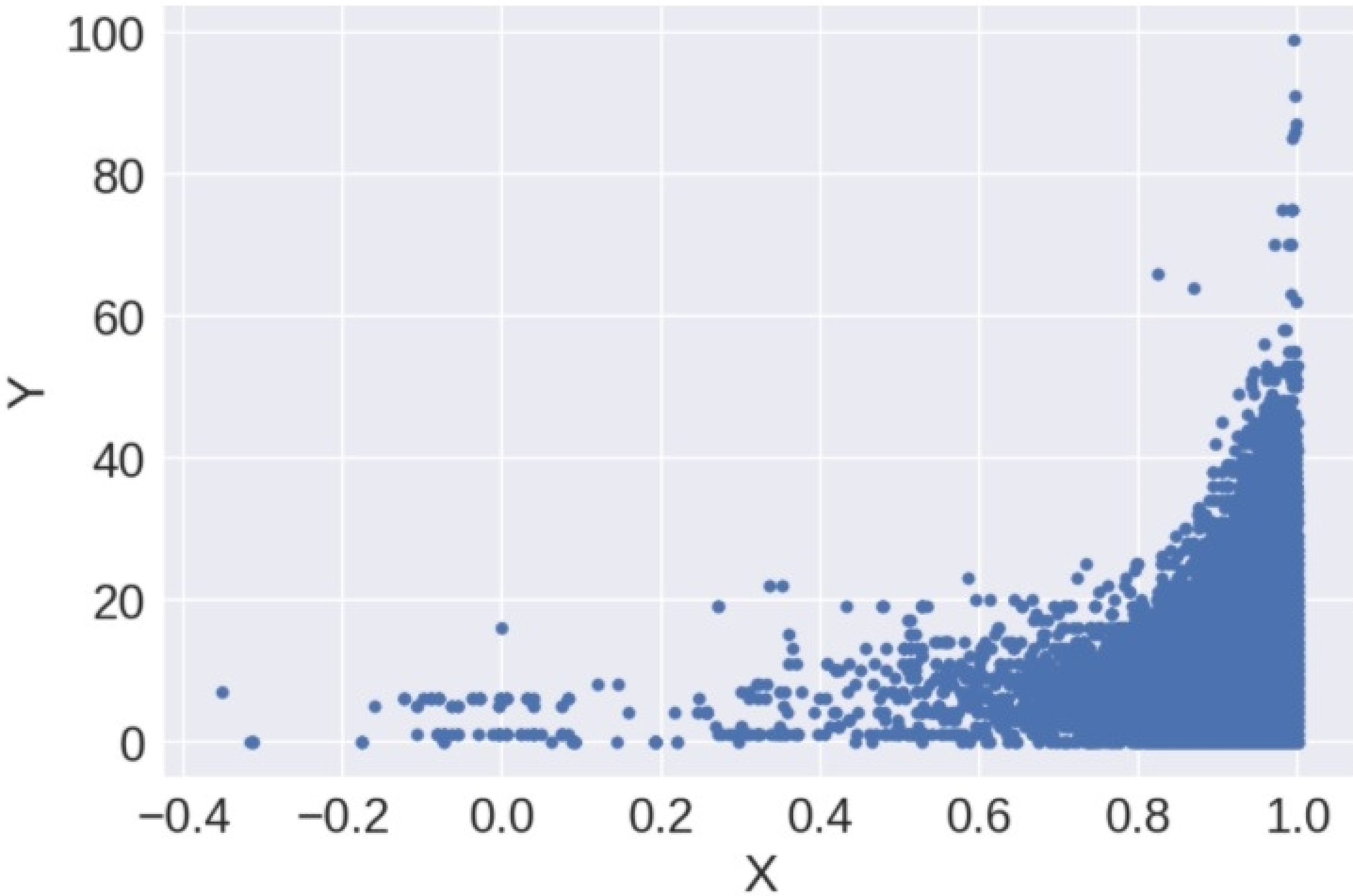
Exploratory data analysis

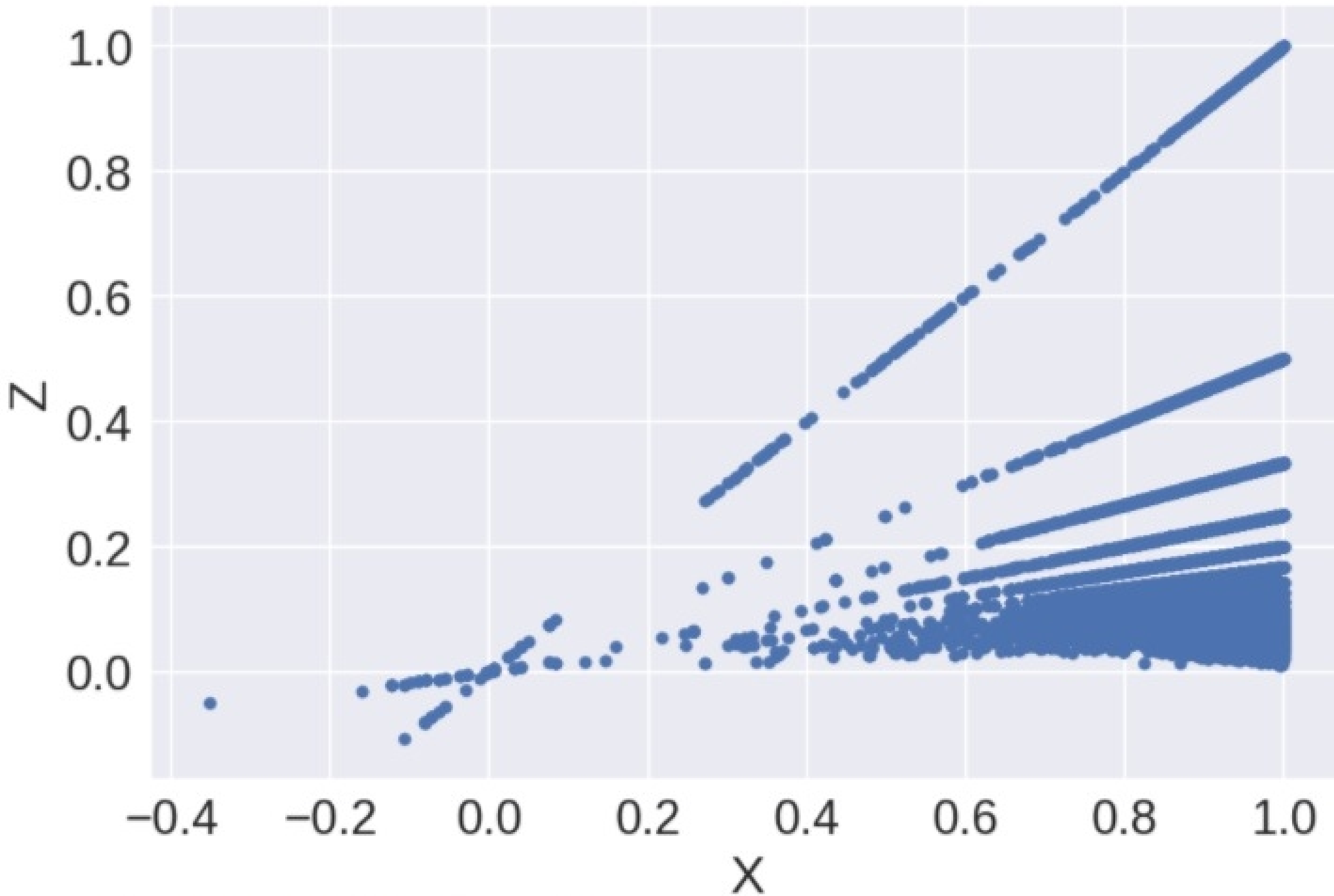
LATEST SUBMISSION GRADE

100%

1.

2 / 2 points





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?

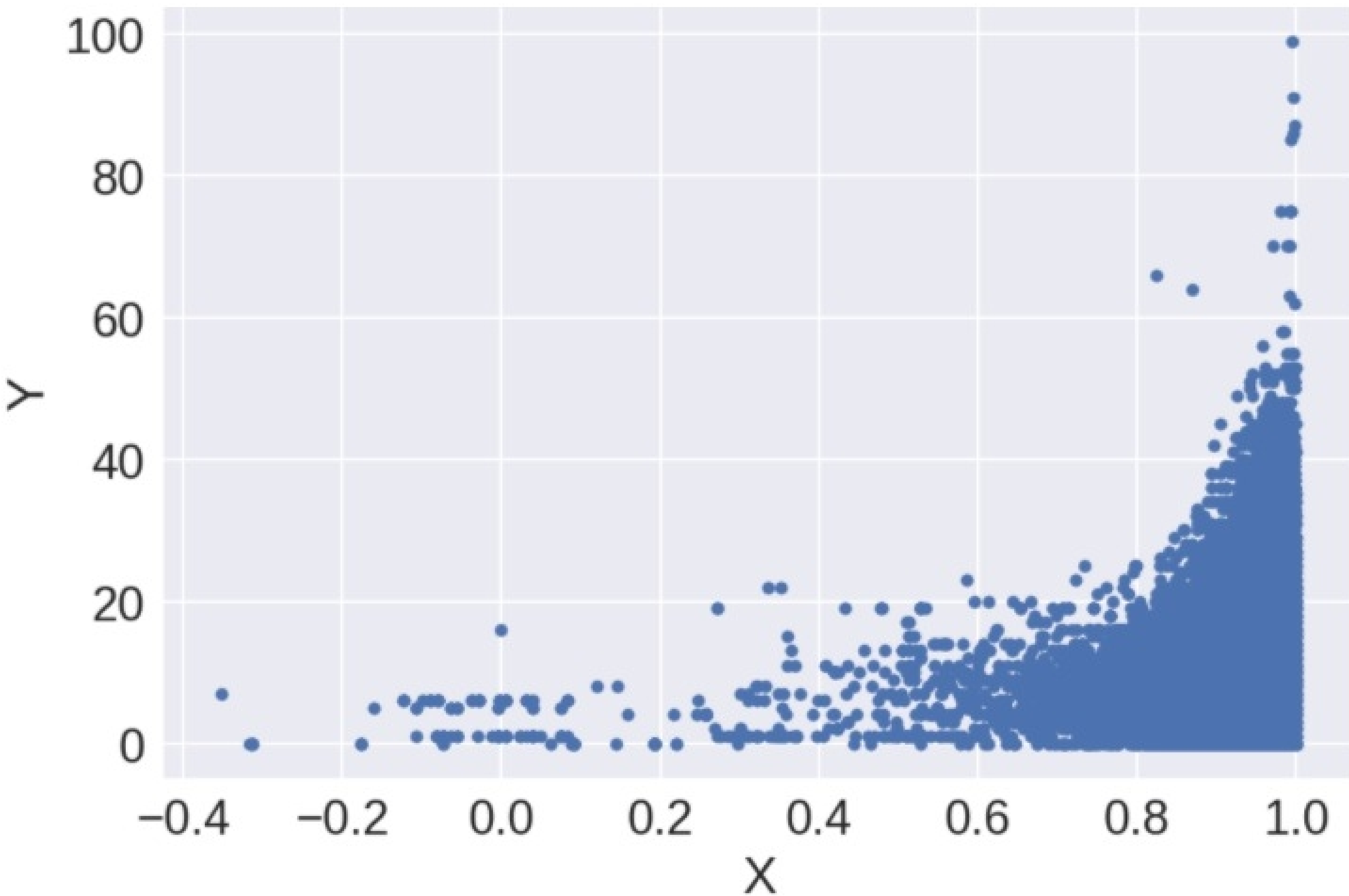
- ○ $Z = X - Y$
- ○ $Z = X + Y$
- ◉ $Z = X/Y$
- ○ $Z = XY$

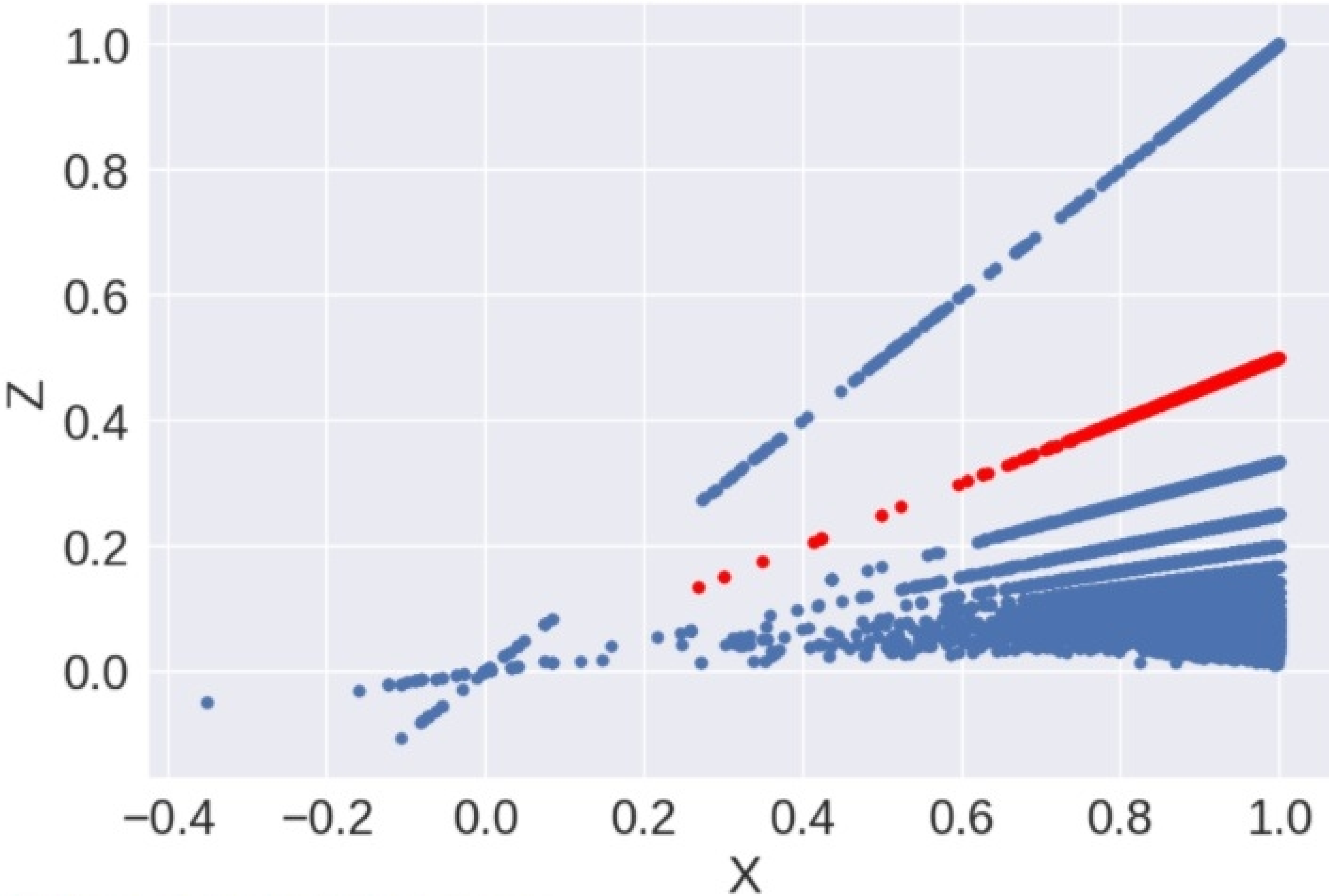
Correct

Correct!

2.

2 / 2 points



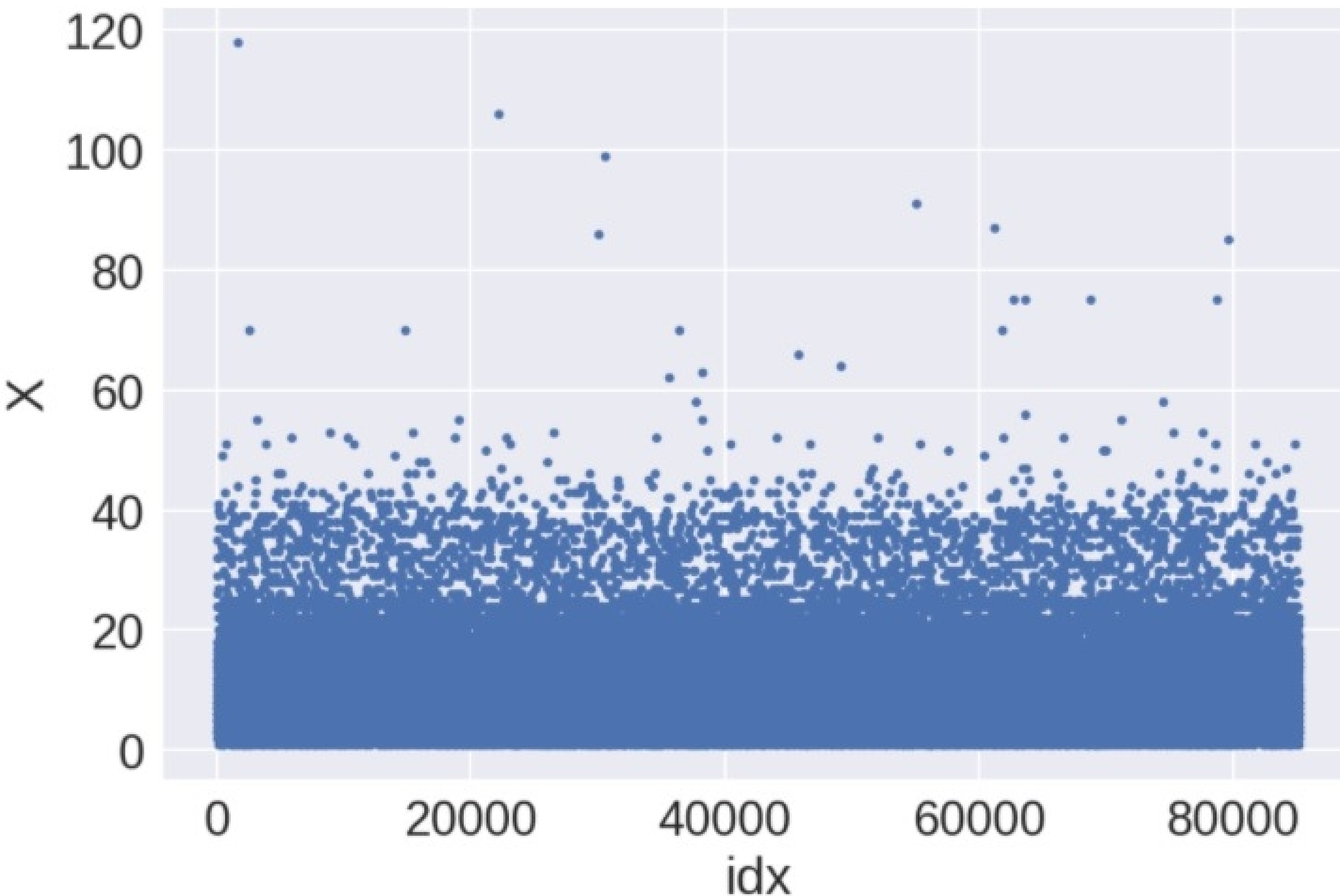


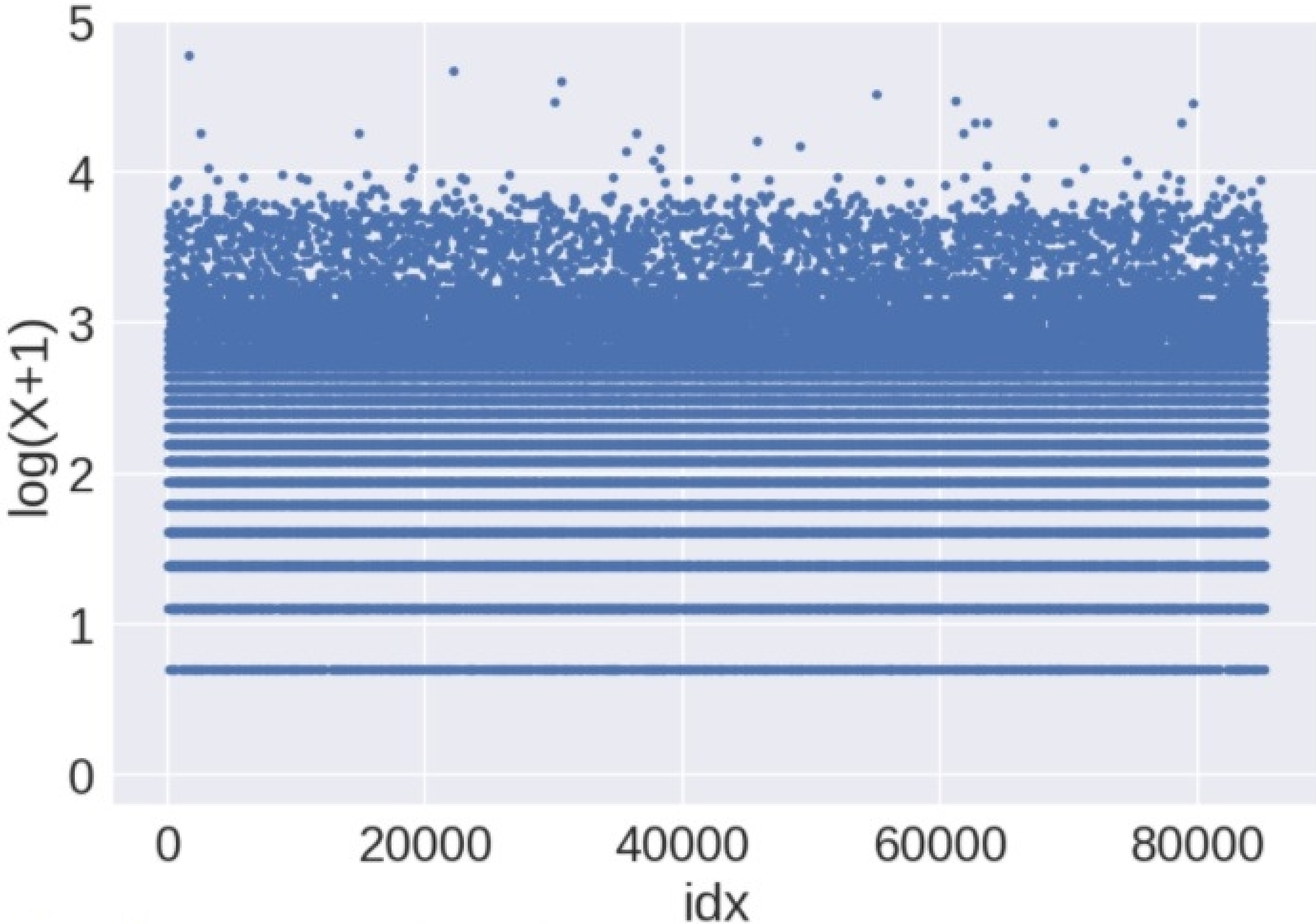
What Y value do the objects colored in red have?

2

3.

2 / 2 points





The following code was used to produce these two plots:

```
# top plot
plt.plot(x, '.')

# bottom plot
logX = np.log1p(x) # no NaNs after this operation
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] 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 a categorical features encoded with label encoder, which starts with label 1 (in pandas and sklearn label encoders usually start with 0).

- [ ] X can take a value of zero

- [x] $2 \leq X < 3$ 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.

- [x] X takes only discrete values

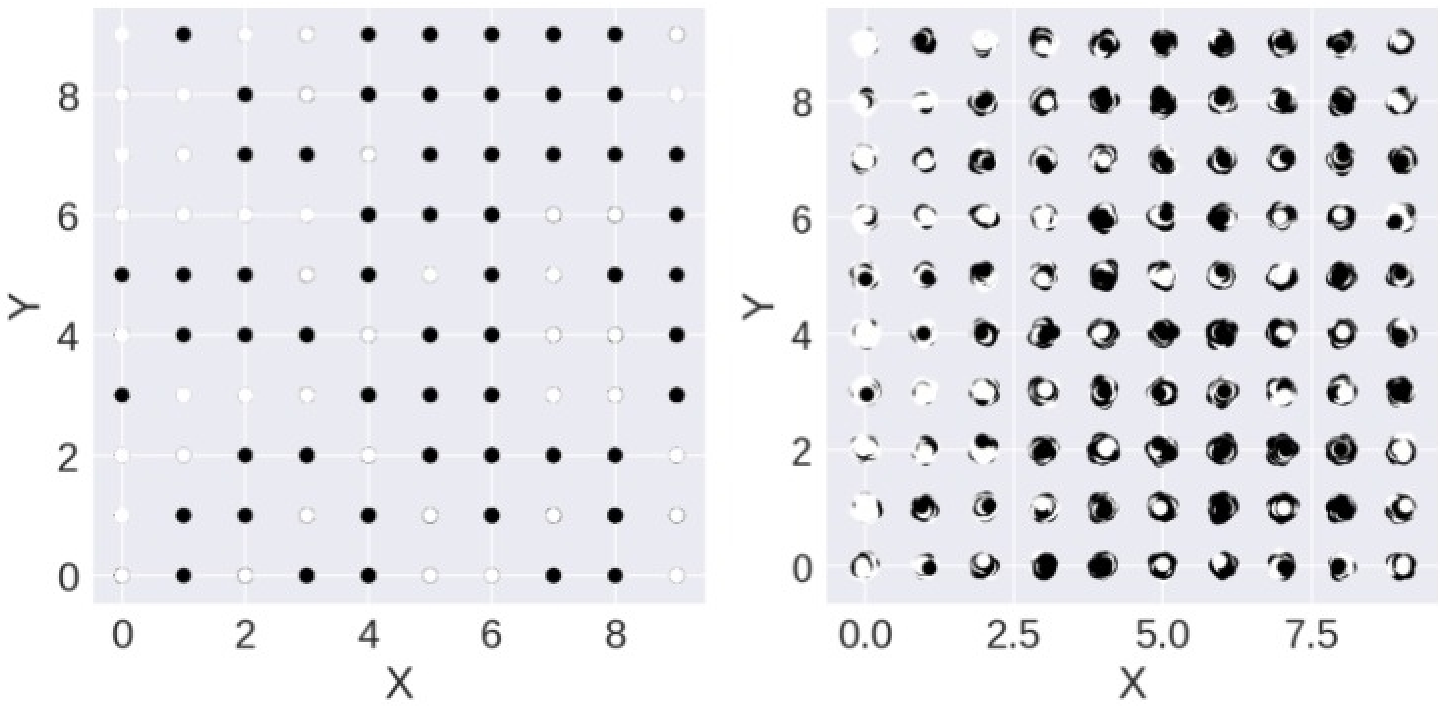
✓ **Correct**

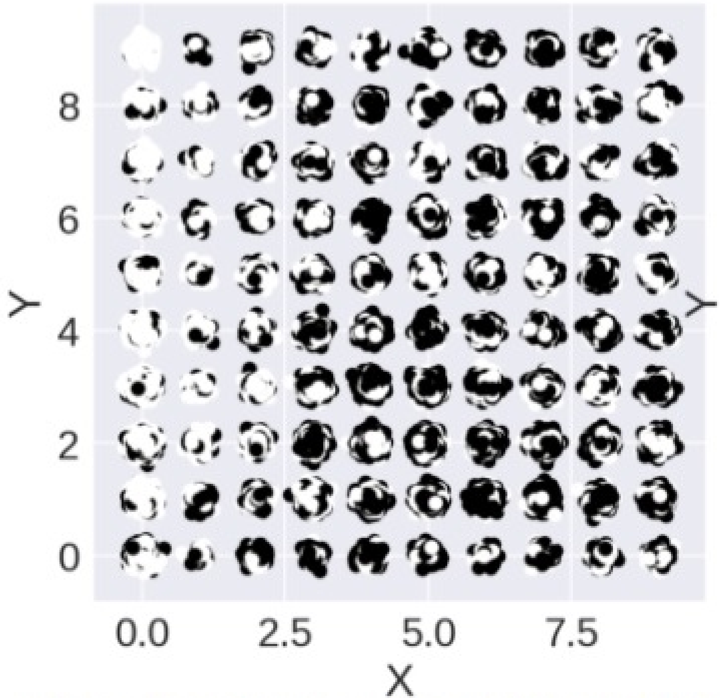
In fact, horizontal lines indicate a lot or repeated values. The most bottom horizontal line on $log(X+1)$ plot corresponds to the value 1, the next to the value 2 and so on.

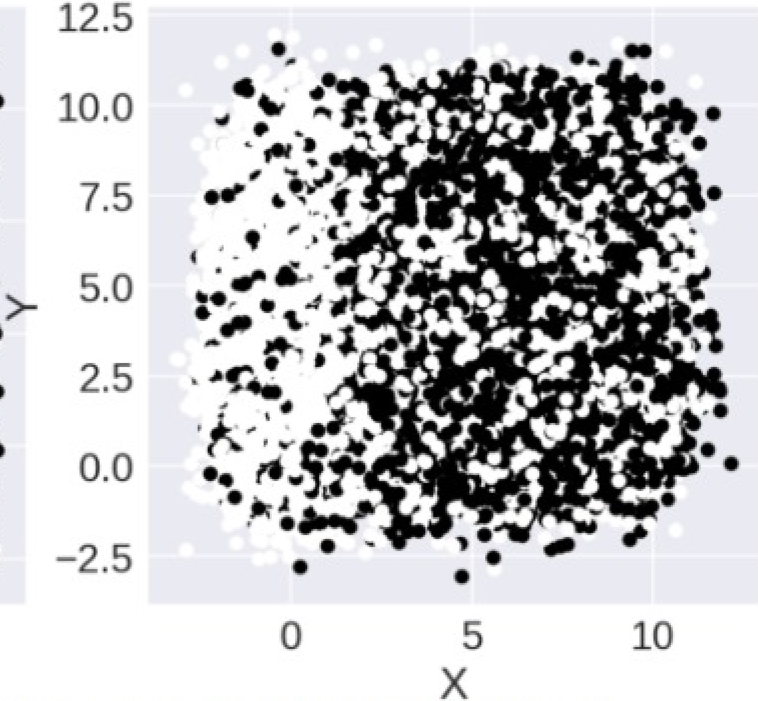
- [ ] X can be the temperature (in Celsius) in different cities at different times

4.

2 / 2 points







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:

```python
# y is a target vector
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:

```python
def jitter(data, stdev):
    N = len(data)
    return data + np.random.randn(N) * stdev

# sigma is a given std. dev. for Gaussian distribution
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.

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

- [x] 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.

- [ ] It is *always* beneficial to jitter variables before building a scatter plot

- [x] 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.