Exploratory data analysis
TOTAL POINTS 8

1.Question 1

Suppose we are given a data set with features XX, YY, ZZ.

On the top figure you see a scatter plot for variables XX and YY. Variable ZZ is a function of XX and YY and on the bottom figure a scatter plot between XX and ZZ is shown. Can you recover ZZ as a function of XX and YY?

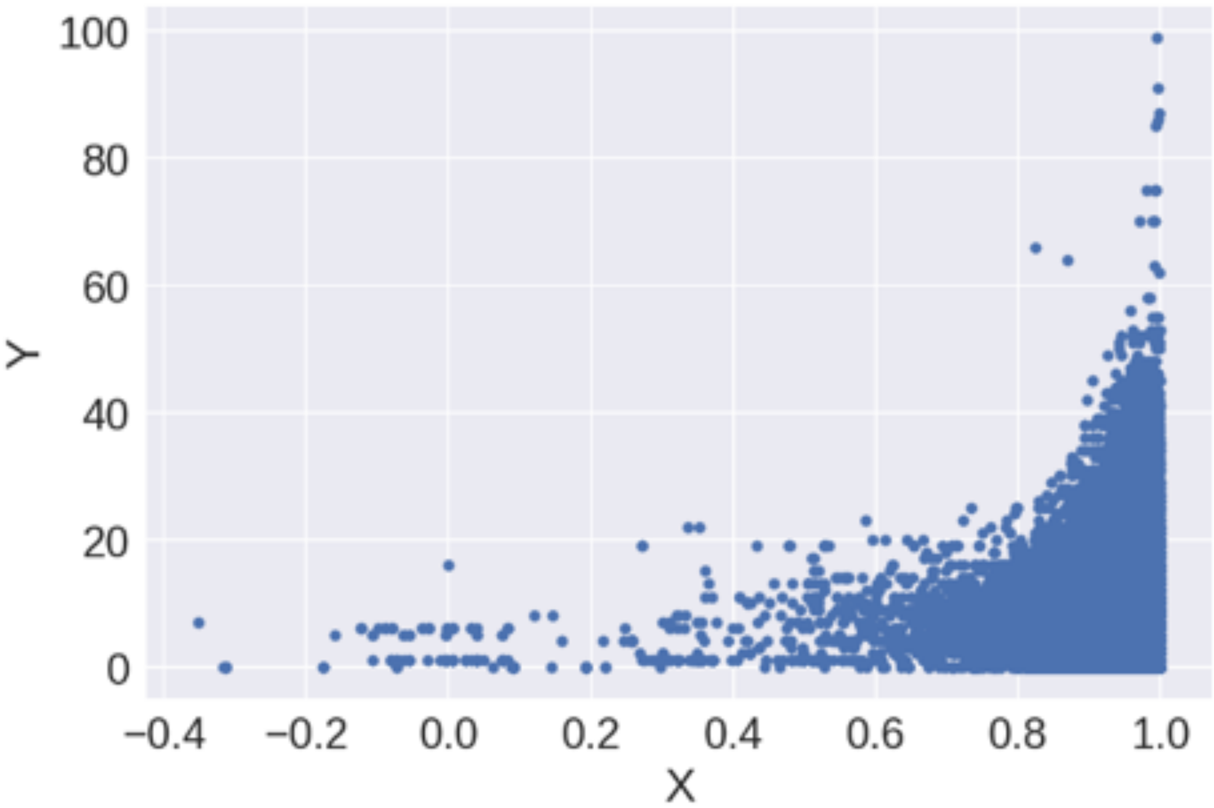
() Z = XY

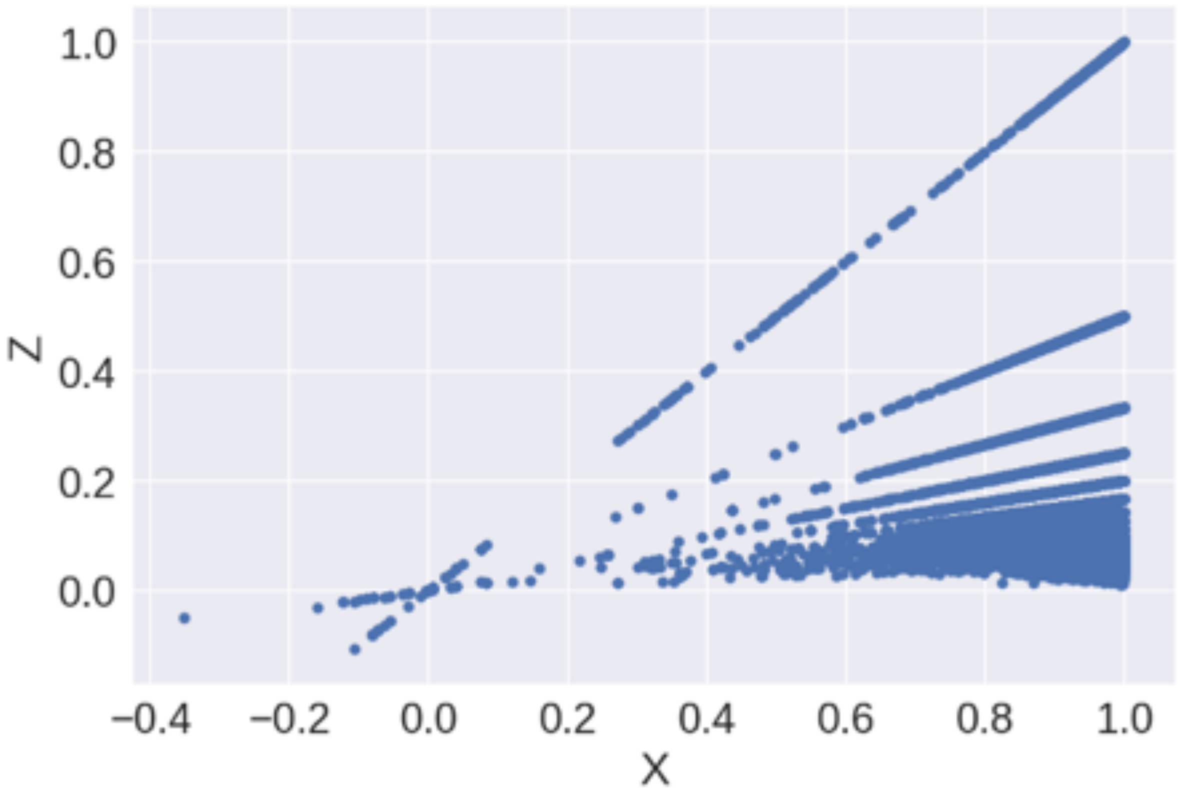
() Z = X + Y

(X) Z = X / Y

() Z = X - Y

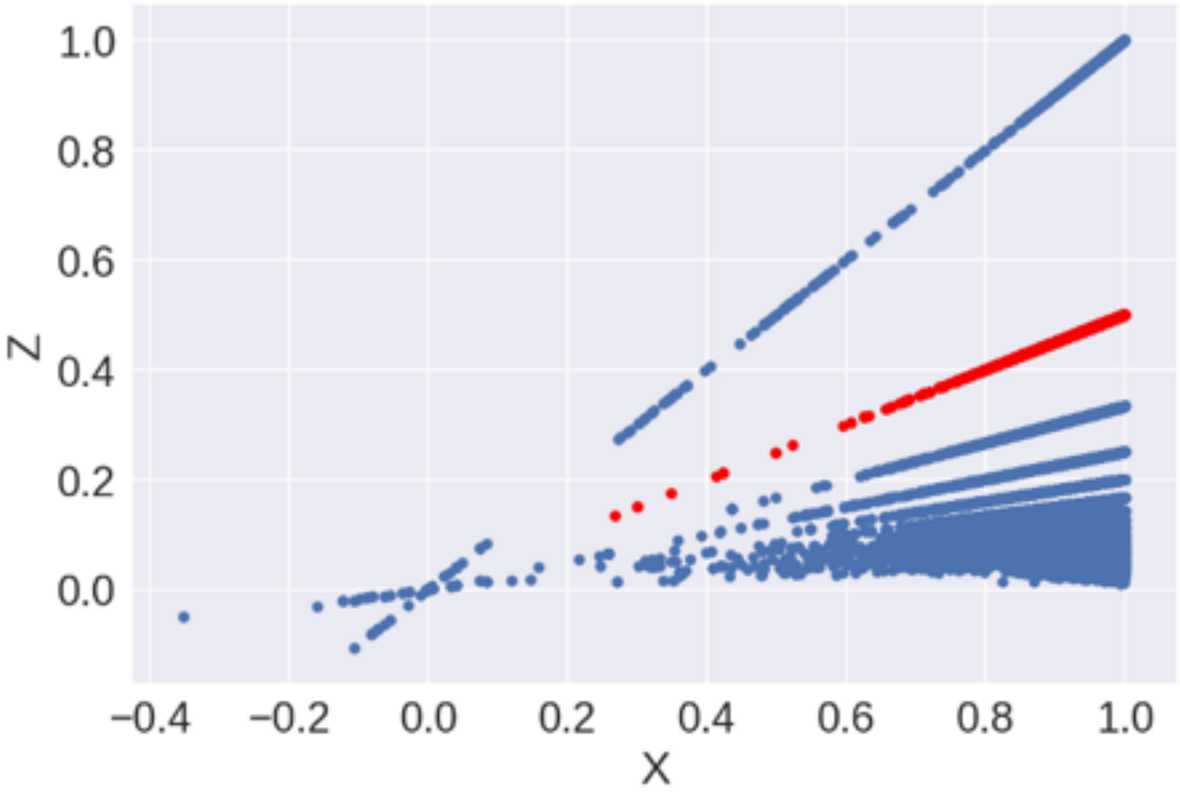
2 points





2.Question 2

What Y value do the objects colored in red have? 2



3.Question 3

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 XX as in previous questions).

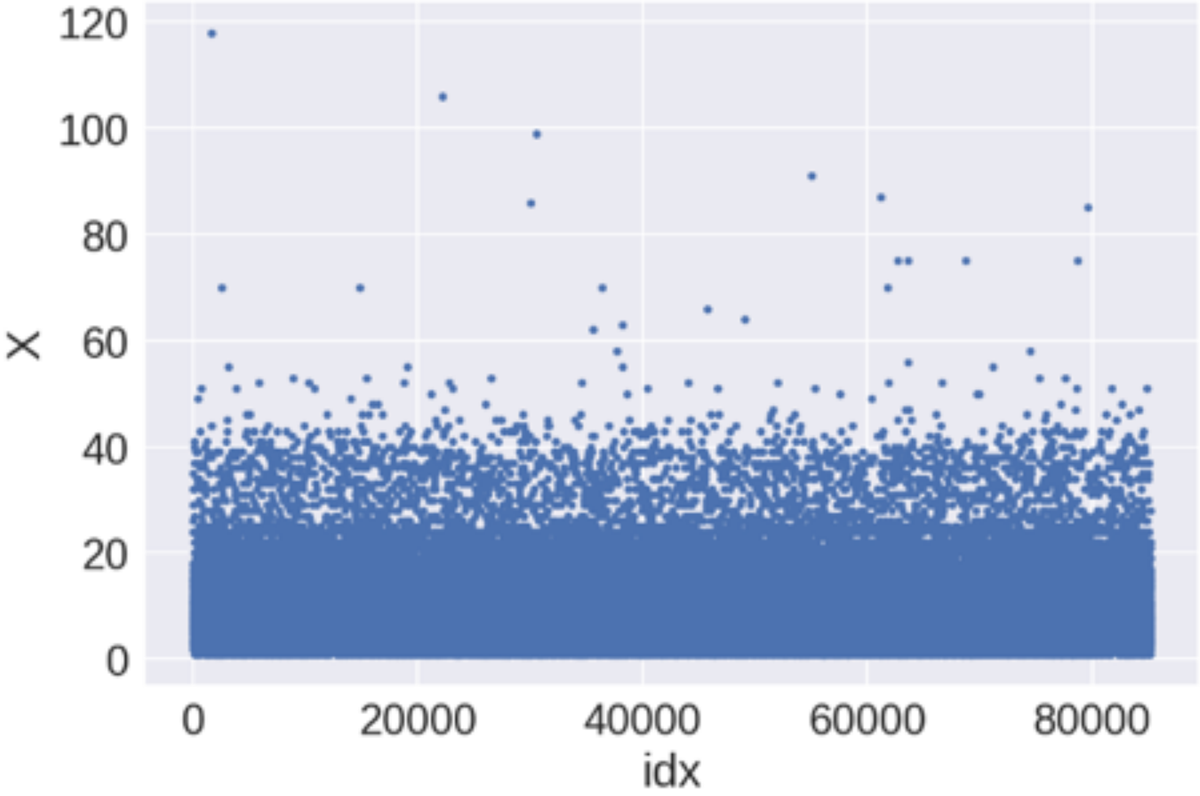
Which hypotheses about variable XX do NOT contradict with the plots? In other words: what hypotheses we can't reject (not in statistical sense) based on the plots and our intuition?

(**X**) X is a counter or label encoded categorical feature

(**X**) X takes only discrete values

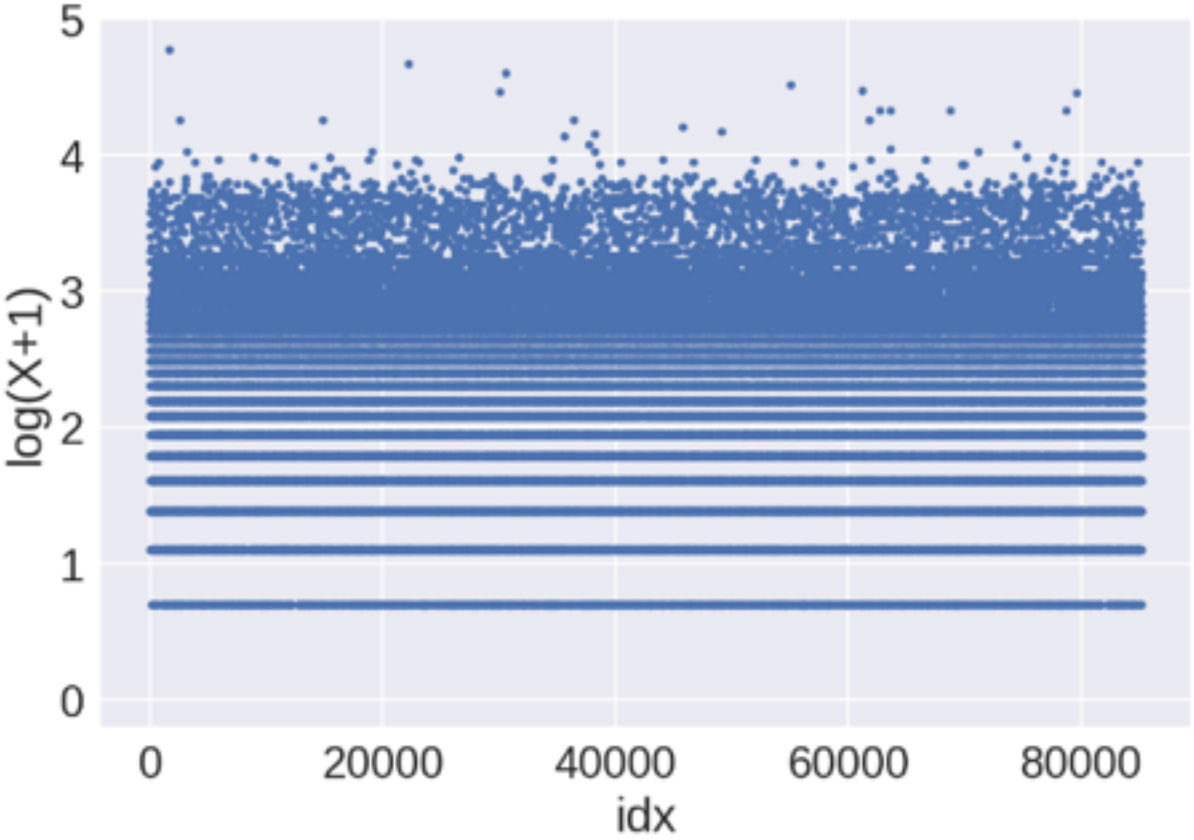
() X can take a value of zero

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



(X) 2≤X<3 happens more frequently than 3≤X<4

2 points



4.Question 4

Suppose we are given a dataset with features X and Y and need to learn to classify objects into 22 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:

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

We use target variable y to color code the points.
The other three plots were produced by jittering X and Y values:

```
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.

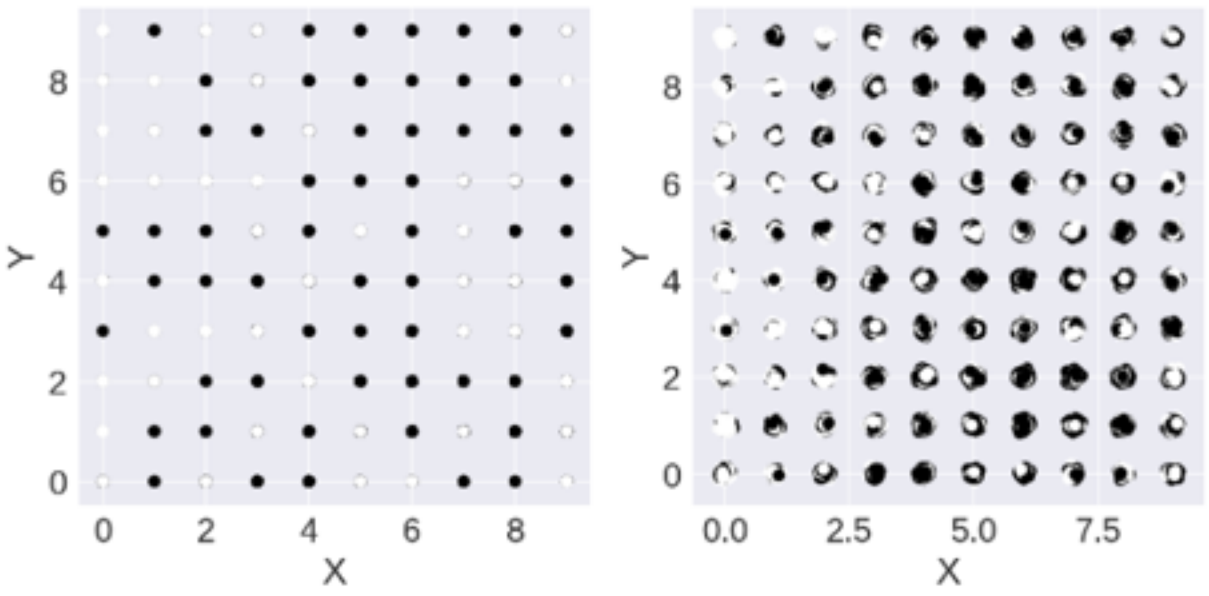
() Target is completely determined by coordinates (x,y), i.e. the label of the point is completely determined by point's position (x,y). Saying the same in other words: if we only had two features (x,y), we could build a classifier, that is accurate 100% of time.

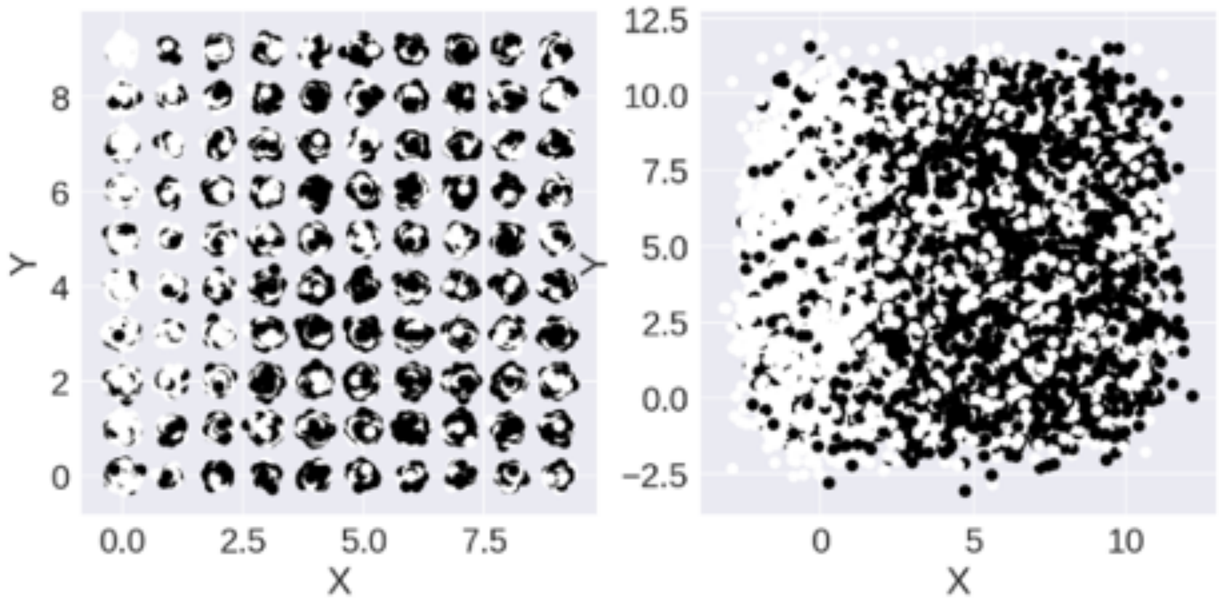
(**X**) Standard deviation for Jittering is the largest on the bottom right plot.

() It is always beneficial to jitter variables before building a scatter plot

(**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.

2 points





I, Chun-Min Jen, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account. Learn more about Coursera's Honor Code