

- Exploratory data analysis
- Reading: Week 2 overview

10 min
- Video: Exploratory data analysis

7 min
- Video: Building intuition about the data

6 min
- Homework: Reading material for video 2

22 min
- Video: Exploring univariate data

15 min
- Homework: Notebook for video 3 (optional)
- Video: Visualizations

11 min
- Video: Dataset cleaning and other things to check

7 min
- Quiz: Exploratory data analysis

4 questions
- Reading: Additional material and links

10 min
- EDA examples
- Validation
- Data packages

QUIZ • 13 MIN

Exploratory data analysis

Submit your assignment

Due Sep 9, 2:59 PM CDT

Attempts 3 (every 8 hours)

Try again

Receive grade

To PASS (75% or higher)

Grade 100%

View Feedback

Go keep your highest score

Share

Flag

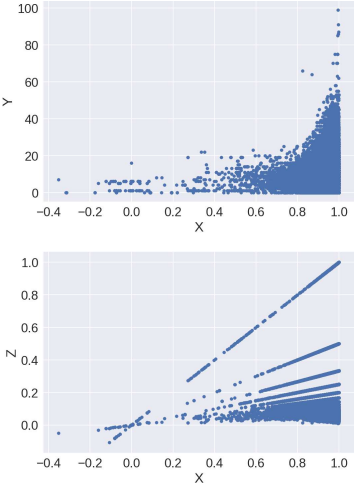
Report

Exploratory data analysis

TOTAL POINTS 0

1. 

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  $Y$ ?

☐  $Z = X + Y$

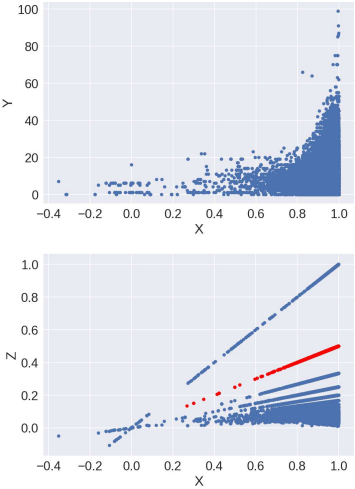
☐  $Z = X/Y$

☐  $Z = X - Y$

☐  $Z = XY$

2. 

2 points

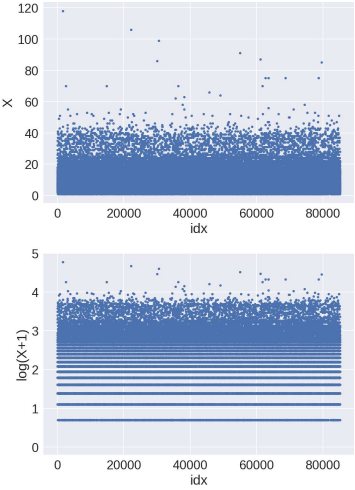


What  $Y$  value do the objects colored in red have?

Enter answer here

3. 

2 points



The following code was used to produce these two plots:

```
# Top plot
plt.scatter(X, "X")

# Bottom plot
logX = np.log(X+1) # no NaNs after this operation
plt.scatter(logX, "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?

☐  $X$  is a counter or label encoded categorical feature

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

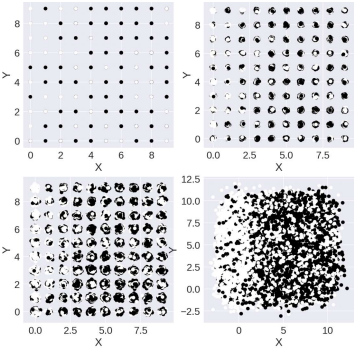
☐  $X$  takes only discrete values

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

☐  $X$  can take a value of zero

4. 

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:

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

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

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

```
# Jitter scatter, noisy
X = np.random.randn(1000)
y = np.random.randn(1000) * 10
# Jitter to y given old, dev. for Gaussian distribution
plt.scatter(jitter(X), jitter(y), alpha=.1, color=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

☐ Standard deviation for jittering is the largest on the bottom right plot.

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

☐ 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.

☐ 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.

☐ I Hung-Yuan Lin, understand that submitting work that isn't my own may result in permanent failure

Share

Flag

Report