Homework 3.1

The first 7 questions require the use of a dataset on air quality. The dataset is available via the following link:

https://docs.google.com/spreadsheets/d/1kIQlTMT871e9REV5UABr-PTcCJ-6BMb7TteMxYBOYeo/pub?gid=1225723766&single=true&output=csv.

The dataset is obtained from the UC Irvine Machine Learning Repository, as described in the References. It contains hourly average responses of a gas multisensor device deployed on the field in an Italian city. For a detailed description of the dataset, please follow the provided reference. The precise dataset we use for Part I is a curated subset of the Air Quality dataset.

The easiest way to bring the data into AzureML is to use the “Import Data” feature taught in class, and copy and paste the link in AzureML. If you like, you can also download the csv data file, and upload it to AzureML (this has also been shown in class), or use it on your own computer with some other software like Excel or R. Once you have the data uploaded, conduct the descriptive analytics exercises below.

Each question is worth 1 point.

Question 1

1/1 point (graded)

What is the average NOx concentration in ppb? Round your answer to the first decimal place.

- ○ 128.0
- ○ 81.8
- ◉ 143.5 ✔
- ○ 478.0

Submit You have used 2 of 2 attempts

Question 2

1/1 point (graded)

You can use this template as a guide to the simple editor markdown and OLX markup to use for multiple choice problems. Edit this component to replace this template with your own assessment.

What is the standard deviation of Benzene (labeled as C6H6(GT) in the data) concentration in microg./m^3? Round your answer to the first decimal place.

- ○ 9.1
- ◉ 7.4 ✔
- ○ 10.8
- ○ 0.5

Submit You have used 2 of 2 attempts

Question 3

1/1 point (graded)

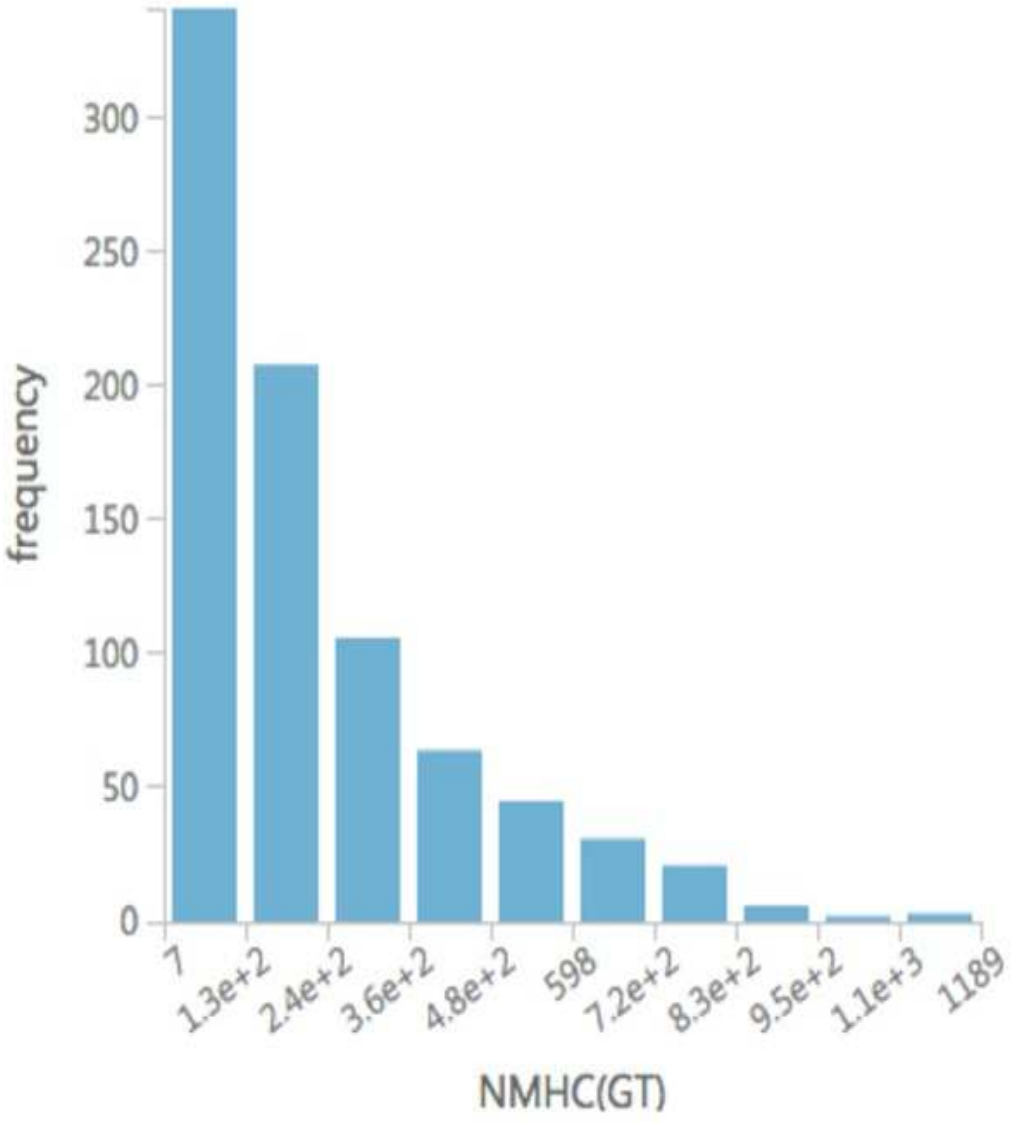
Which of the following best describes the shape of the histogram of the Non Metanic HydroCarbons (labeled NMHC(GT) in data) in microg./m^3?

- ○ Normal distribution
- ○ Uniform distribution
- ◉ Exponential distribution ✔
- ○ Bimodal distribution
- ○ None of the above

Answer

Correct:

Answer explanation: The resulting histogram is shown below, and you can see that it is highly asymmetric.



Submit You have used 3 of 3 attempts

Question 4

1/1 point (graded)

Which of the following best describes the relationship between CO concentration (labeled CO(GT) in the data) and NO2 concentration (labeled NO2(GT))?

- ○ Positive, and strongly linear
- ○ Negative, and strongly linear
- ◉ Positive, but plausibly non-linear ✔
- ○ Negative, but plausibly non-linear

Submit You have used 2 of 2 attempts

Question 5

1/1 point (graded)

The coefficient of correlation between NOx concentration (labeled NOX(GT)) and NO2 concentration (labeled NO2(GT)) lies in which range?

- ○ -1 to -0.5
- ○ -0.5 to 0
- ○ 0 to 0.5
- ◉ 0.5 to 1 ✔

Submit You have used 2 of 2 attempts

Question 6

1/1 point (graded)

Which of the following is the most tightly correlated with PT08.S2 sensor response (labeled PT08.S2(NMHC))?

- Non Metanic HydroCarbons concentration (labeled NMHC(GT))
- PT08.S3 sensor response (labeled PT08.S3(NOx))
- NO2 concentration (labeled NOx(GT)) ✔
- Temperature (labeled T)

Answer
Correct:
Answer explanation: The coefficient of correlation between PT08.S2 sensor response and the four options are 0.88 for NMHC, -0.91 for PT08.S3, 0.93 for NO2, and 0.44for temperature, respectively. Note that using scatter plot may not be feasible here, since the visual "tightness" of some scatter plots are indistinguishable.

Submit You have used 4 of 4 attempts

Question 7

1/1 point (graded)
Suppose that the sensor that detect the Benzene concentration (6-th column with header "C6H6(GT)") breaks down while other parts work fine. To predict the Benzene concentration, the value of which of the following options is the most useful?

- CO concentration (3-rd column with header "CO(GT)") ✔
- Non Metanic HydroCarbons concentration (5-th column with header "NMHC(GT)")
- Temperature (13-th column with header "T")
- Relative humidity (14-th column with header "RH")

Answer
Correct:
Answer explanation: Simply looking at the scatter plots can provide the answer. Alternatively, we can compare the coefficients of correlation; they are 0.97 for CO, 0.89 for NMHC, 0.41 for temperature, and -0.12 for relative humidity, respectively.

Submit You have used 2 of 2 attempts

Question 8

1/1 point (graded)
Which of the following term refers to the most frequently occurring value of a data set?

- ○ Maxima
- ○ Mean
- ○ Median
- ○ Standard deviation
- ◉ None of the above answers is correct. ✔

Answer
Correct: Answer explanation: The most likely value is referred to as the Mode.

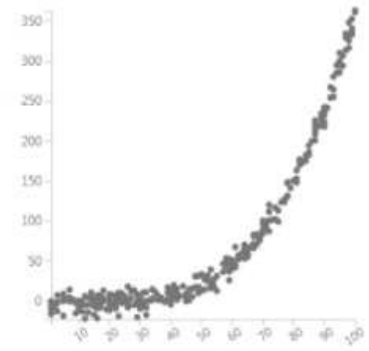
Submit You have used 2 of 2 attempts

Question 9

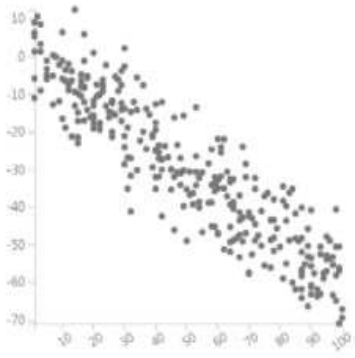
1/1 point (graded)

We know that the coefficient of dataset X and dataset Y equals to -0.87. Which of the following scatter plots is most likely to be the scatter plot of (X, Y).

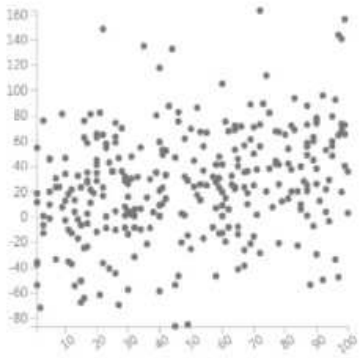
a.



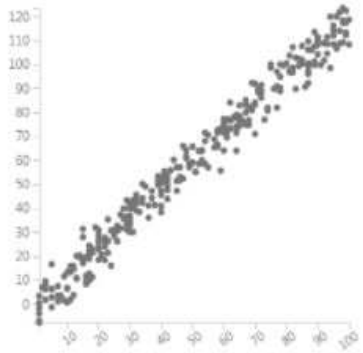
b.



c.



d.



- ○ a
- ◉ b ✔
- ○ c
- ○ d

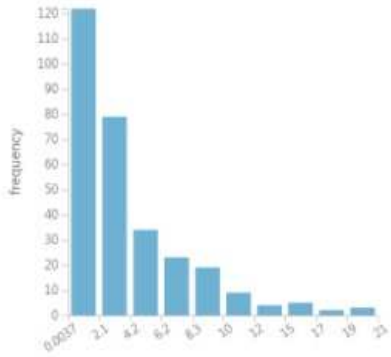
Submit You have used 2 of 2 attempts

Question 10

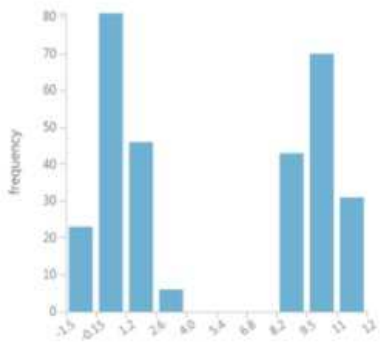
1/1 point (graded)

Which of the following datasets would you consider most close to a normal distribution?

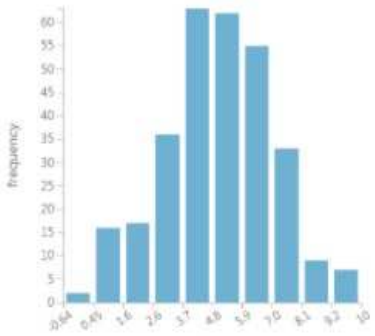
a.



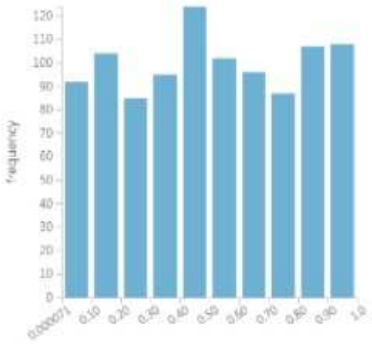
b.



c.



d.



- a
- b
- c ✔
- d

Answer
Correct:
Answer explanation: Histogram (c) most closely shows the relevant features of a normal distribution: a single peak near the center, roughly symmetric distribution, with tails falling off.

Submit You have used 2 of 2 attempts

Question 11

0/1 point (graded)

Which of the following statement is true for any set of data whose mean equals half of the sum of the maxima and the minima?

- ○ The mean equals to the median
- ◉ The median equals to the half of the sum of the maxima and the minima ✖
- ○ The standard deviation is no greater than the half of the difference between the maxima and the minima
- ○ The standard deviation is less than the minima.

Submit You have used 2 of 2 attempts