DATA 100: Vitamin 4 Solutions

September 13, 2019

1 Pandas Data Structures

Link the following definitions to their corresponding Pandas data structure:

1. A sequence of row labels
2. Two-dimensional (tabular data)
3. One-dimensional (column data)

☐ Data Frame: 1, Series: 2, Index: 3

☐ Data Frame: 2, Series: 1, Index: 3

☑ Data Frame: 2, Series: 3, Index: 1

☐ Data Frame: 3, Series: 2, Index: 1

Explanation: A Data Frame is a two-dimensional structure in tabular format. A Series is one-dimensional, and is used to represent column data. An Index is a sequence of row labels. Data frames and series have Indices. We can think of a Data Frame as a collection of Series that share the same Index.

2 Pandas Indices

Which of the following statements about Pandas Indices are false?

☑ Indices must integers

☑ Indices may be non-numeric, and are always unique

☑ Indices need not be unique, but must be numeric

☐ Indices need not be unique, and can be non-numeric

Explanation: Pandas Indices do not have to be unique, and can consist of non-numeric values.

3 loc Vs. iloc

Which of the following statements regarding `iloc` are true?

- ☐ It is harder to make mistakes with iloc than with loc
- ☐ It is easier to read iloc code than loc code
- ☑ iloc doesn't use labels
- ☑ iloc is vulnerable to changes in the ordering of rows and columns in a Data Frame

Explanation: Because iloc works with numerical positions, it is often harder to read than loc, which uses labels. This also means that iloc is vulnerable to changes in the ordering of rows and columns in a Data Frame, and is therefore more prone to produce errors than loc.

4 groupby

Fill in the blank: The result of a groupby operation applied to a DataFrame object is a ____.

- ☐ DataFrame object
- ☑ DataFrameGroupBy object
- ☐ Series object

Explanation: Applying a groupby operation to a Data Frame produces a DataFrameGroupBy object. Functions can then be applied to this new object to create new Data Frames and Series.

5 Average Absolute Loss

Given an even number of data points $x_1, \ldots, x_n$ without ties, which of the following values will minimize the mean absolute loss?

- ☑ median($\{x_1, \ldots, x_n\}$)
- ☐ x_1
- ☐ mode$\{x_1, \ldots, x_n\}$
- ☐ $\bar{x}$
- ☐ $\frac{x_1+x_n}{2}$

Explanation: The median will minimize the average absolute loss. See discussion 3's solutions for a derivation.