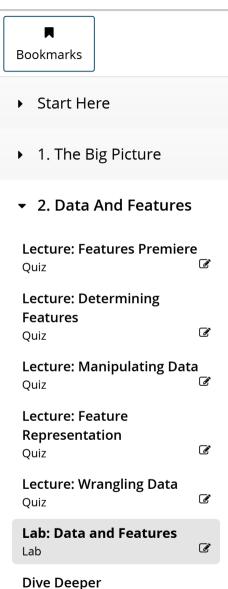edX   **Microsoft:** DAT210x Programming with Python for Data Science

Hel

---

2. Data And Features > Lab: Data and Features > Assignment 5

Assignment 5

🔖 **Bookmark this page**

Lab Assignment 5

Barry Becker extracted a reasonably clean subset of the 1994, U.S. Census database, with a goal of running predictions to determine whether a person makes over 50K a year. The dataset is hosted on the University of California, Irvine's Machine Learning Repository and includes features such as the person's age, occupation, and hours worked per week, etc.

As clean as the data is, it still isn't quite ready for analysis by SciKit-Learn! Using what you've learned in this chapter, clean up the various columns by encode them *properly* using the best practices so that they're ready to be examined. We've included a modifies *subset* of the dataset at Module2/Datasets/**census.data** and also have some started code to get you going located at Module2/**assignment5.py**.
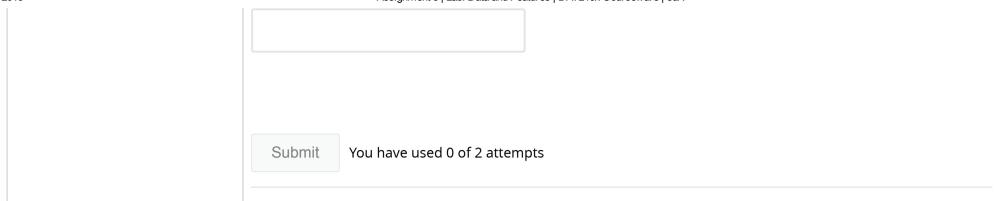
1. Load up the dataset and set header label names to:

```
['education', 'age', 'capital-gain', 'race', 'capital-loss', 'hours-per-week', 'sex',
'classification']
```

Ensure you use the *right* command to do this, as there is more than one command! To verify you used the correct one, open the dataset in a text editor like SublimeText or Notepad, and double check your `df.head()` to ensure the first values match up.

2. Make sure any value that needs to be replaced with a NAN is set as such. There are at least three ways to do this. One is *much* easier than the other two.

Sidebar navigation

3. Look through the dataset and ensure all of your columns have appropriate data types. Numeric columns should be float64 or int64, and textual columns should be object.

4. Properly encode any ordinal features using the method discussed in the chapter.

5. Properly encode any nominal features by exploding them out into new, separate, boolean features.

Lab Question 1

1 point possible (graded)

Please enter a numeric value (e.g. 0, 1, 2, 3, etc) which correctly answers the question(s) below:

Before you made any changes to the downloaded dataset, how many of the columns were ordinal?

Submit   You have used 0 of 2 attempts

Lab Question 2

1 point possible (graded)

Please enter a numeric value (e.g. 0, 1, 2, 3, etc) which correctly answers the question(s) below:

Before you made any changes to the downloaded dataset, how many of the columns were nominal?

Submit　　　You have used 0 of 2 attempts

POWERED BY
OPENedX