wheaton college massachusetts

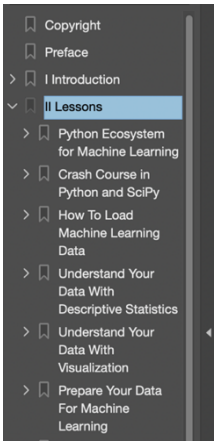
weCodeForFoo, Programming Unit
Norton, Massachusetts 02766

March 18, 2025

Dear Applicant:

Thank you for applying to join our team. At weCodeForFoo™, we "cut code" for others; we break, we patch, we do the systems engineering and programming work that helps our clients get from their "starting data" toward machine learning experiments, whether it be clustering or classification. Our specialty and focus are on the programming techniques for data cleaning.

Your challenge ("if you so choose to accept it" 😊) is for you to start with the attached file of data (Akime-O_odham-diabetes.data.csv) and produce a professionally formatted Jupyter Notebook that shows an experimental workflow; that is, you can write the Python to load the data into Pandas data frames and follow Jason Brownlee's (Data Preparation) Workflow and other techniques shown in class or from the reading; your final dataset is your stamp of approval for this data to proceed to future ML experiment. This dataset describes the medical records for the (primarily) Arizona Akimel O'odham ("River People") community and will be used to predict whether or not a patient is likely to experience onset of Type 2 diabetes (adult-onset diabetes) within five years. Later, we will attempt to predict the class variable/column (**1** means the person tested positive for diabetes, **0** means the person tested negative for diabetes) based on other features (columns, variables) in the dataset.



Specifically, re-create a workflow on the data as shown in **Chapters 2 – 7**, and demo-ed in class. You should also reference other websites on data cleaning techniques; in particular you will want to use methods from the data cleaning tutorial on KDNuggets site (https://www.kdnuggets.com/10-pandas-one-liners-quick-data-quality-checks). Note: your datafile may contain some questionable rows and columns of data, e.g., duplicate lines, missing features (column values), and/or outliers. Your Notebook should discuss your detection of any "errors", show corrections taken (in Pandas), and discuss your decisions. In particular, you should make a convincing argument for removing any outliers (but see the boss' critique of our first draft on the next page).

Please submit your work as a .zip file of a folder containing: (1) a README.txt, (2) the original and "cleaned" .csv files, and (3) a Jupyter Notebook that opens, discusses, and verifies how your code follows a "data cleaning" workflow, including changes made from our original Notebook per comments made by our expert data scientist. See the next page for more details.

We look forward to your submission by Wednesday, April 2nd. And thanks for applying!

Sincerely,

MDLeBlanc

Mark D. LeBlanc, Ph.D.
Software Engineer
weCodeForFoo.com

For each scrubbing/cleaning operation that you perform and flag an anomaly, you must "log" (document) each type of change you make (see categories below), including the row# **of the original data set**.

Your Jupyter Notebook report should include:

(1) A thorough description of the data's original rows and columns, including original (and final, after scrubbing) dimensions, data types, and distributions of each column of values (e.g., how many of each type or within a range).
(2) Pre-scrubbing, correct any incorrect data types.
(3) Pre-scrubbing, report on the descriptive statistics of your data, including whisker or other plots.
(4) Post-scrubbing, you should share the final dimensions and distributions of each column of values.
(5) Post-scrubbing, report on the descriptive statistics of your data, including plots when helpful to the reader; you should comment on any significant changes made during scrubbing as compared to the original data.
(6) Post-scrubbing, comment on the distribution of values for each column, especially any changes in the distributions of our class variable; Any cautions before heading to machine learning?

Categories checked:

a) Null (missing column) values
b) Duplicate lines
c) Suspicious zeros (only consider the **mass** and **age** columns)
d) Outliers (only consider the **mass**, **age**, and **class** columns;
discuss the IQR ranges for each column, using the conservative 2*IQR calculation)
e) Incorrect Data formats
f) Count Unique Values in a Column
g) Other (*you add more here that you think are needed)*

Sample lines from an (output) "log file" are shown below:

Log_DataCleaning.csv

```
preg,plas,pres,skin,test,mass,pedi,age,class,ERROR_TYPE,Original_Line_Number
1.0,89,66.0,23,94,28.1,0.167,21.0,,Missing Data,6
1.0,85,66.0,29,0,26.6,0.351,31.0,0.0,Duplicate Row,4
8.0,125,96.0,0,0,0.0,0.232,54.0,1.0,Invalid ZERO Mass,13
```

I showed a draft Notebook to our expert data scientist and he sent back the following comments. Please address his concerns in your upgraded report.

Fri, Mar 14, 3:42 PM (4 days ago)

From Benjamin Batorsky

Thanks for sending along. I think it's a pretty nice overview, but I have a few major gripes:

- Wholesale dropping of outliers, duplicates and errors in coding. I think it'd be sufficient to just point out that these are points worth further investigation rather than dropping them. If the issue is making a clean box-and-whisker plot, those can be excluded.
- The dataset used requires a lot of domain expertise, as demonstrated by the long paragraphs injected into the notebook. Why not just use heights and weights? Those can be intuitively understood (e.g. no one is 100 feet tall).
- "High correlation" is a term used, but not described. There's no real attention to the correlation as a metric. Pandas uses Pearson correlations by default, which makes normality assumptions on the data.
- There's much too much focus on "outliers" and not enough looking at the data distributions themselves. One useful addition would be something like a pairplot (https://seaborn.pydata.org/generated/seaborn.pairplot.html). This is probably preferable to correlations, which can be misleading.
- Calling the data being dropped "bad data" is not the right term and it's something I've seen with new data scientists, that they think anything that doesn't align with their expectation as "wrong". There's not much nuance added. Think of the recent example of the mythical 150 year old Medicare recipients. Any basic Exploratory Data Analysis (EDA) would have revealed the mistake.