Data Preprocessing Project

New Attempt

Due Oct 1 by 11:59pm **Points** 100 **Submitting** a file upload
Available Aug 31 at 12am - Dec 13 at 11:59pm

Dear students,

Please see the following instructions. This is a data preprocessing project assignment. You may choose a language of your choice. Data Preprocessing currently involves 80% of the efforts in any commercial data science or data engineering project in the industry.

Python is very popular and used extensively in this course. However, you are permitted to choose an alternate language (such as Julia) only if you have the experience and confidence, and if you will not need any code samples given to you in the alternate language you choose over Python. In general, the entire course depends mainly on Python and is used extensively for all course demos and code samples (tutorials). You may refer to these artifacts and piggyback your code on top of existing tutorials so that you are successful in your class projects.

1. Find the tutorial (jupyter notebook) on Data Preprocessing on CANVAS.

Files --> lab help --> labs --> Tutorial_4_Data_Preprocessing.ipynb

2. Clearly understand each step of the tutorial by executing the tutorial in Python. Export to PDF and submit the PDF file for this part.
3. Next, you have two choices. Use choice (b), **only if choice (a) is not possible.**

3(a) You may choose a data set that is already unclean and fix it with data pre-processing.

for 3(a) check to see if the below link has unclean datasets:

The following link "perhaps" has unclean datasets ready for preprocessing. Check it out. If not, you may use the link to choose a dataset and use simple techniques to make it unclean first.

For Example, one of the links is (London Air dataset download link):

https://www.londonair.org.uk/london/asp/datasite.asp?CBXSpecies1=COm&CBXSpecies2=NOm&CBXSpecies3=NO2m&CBXSpecies4=NOXm&CBXSpecies5=O3m&day1=1&month1=jan&year1=2018&day2=1&month2=jan&year2=2019&period=15min&ratidate=&site=HI0&res=6&Submit=Plot graphLinks to an external site.
Links to an external site.

3(b) You can re-engineer any dataset on **CANVAS** so that you can change it from a good dataset to a bad dataset by clever use of instructions in the language you are programming in. So, it may be all about becoming more proficient in the language. Then apply data preprocessing.

Apply cleaning (preprocessing) tasks to your data. Make sure you cover all techniques of data pre-processing.

For detailed rubric explanations and FAQs, please refer to the below document.

CSC177_Data_Preprocessing_Rubric_Explanation&FAQ's.pdf

CSC177-rubric-1

Criteria	Ratings	Pts
Test using in-class Jupyter notebook (tutorial4 on data preprocessing)		40 pts
Chose domain area		5 pts
find dataset for preprocessing		5 pts
Make dataset ready for preprocessing		5 pts
document techniques for preprocessing		5 pts
arrange steps for preprocessing in correct order and explain why		20 pts
document insights after preprocessing		10 pts
split dataset using the split function and calculate mean and standard deviation		5 pts
compare the split datasets and develop intuition on datasets		5 pts
Total Points: 100		