

ECE5984 SP22 - Prof. Jones – HW 2


Due Tuesday, Feb 15, 2022 – 11:59 PM via Canvas

In the Datasets folder in the Files area of our Canvas site, you will find an Excel spreadsheet called “Heart Disease.xlsx”. This contains the dataset for this homework assignment. You are to modify the simple Python program presented in lecture 6 to complete the data report for this dataset. Your completed program should do the following.

1. Print to the Python console a more complete set of statistics on the data. You may use numpy operations on a numpy array, or operations on a Pandas DataFrame (or Series) for this functionality. The output that I want is as follows (not all columns are shown):

```
File C:/Data/Heart Disease.xlsx is of size (303, 15)
```

	stat	member	age	...
0	cardinality	301.000000	41.000000	...
1	mean	48332.657807	54.366337	...
2	median	48340.000000	55.000000	...
3	n_at_median	1.000000	8.000000	...
4	mode	46820.000000	58.000000	...
5	n_at_mode	1.000000	19.000000	...
6	stddev	877.940533	9.082101	...
7	min	46820.000000	29.000000	...
8	max	49840.000000	77.000000	...
9	nzero	0.000000	0.000000	...
10	nmissing	2.000000	0.000000	...

2. Write a similar report to an Excel workbook. I used operations on a pandas DataFrame for this functionality. Note: “cardinality” is the number of distinct values. The output that I want in the spreadsheet is as follows (note, yours will have numbers ):

	stat	member	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	bt	target
0																	
1																	
2																	
3																	
4																	
5																	
6																	
7																	
8																	
9																	
10																	

3. From this DQR, determine a few things about each column in the data set. For every column in the data, tell me each of the following:
 - a. The type of the feature (ID, target, or feature – and what type of feature – continuous, binary, interval, categorical, etc.)
 - b. How many values are missing and how many are invalid
 - c. What should be done about the missing values – BE SPECIFIC (don’t just say “replace missing values” but tell me how)
 - d. Whether the feature contains a significant number of outliers
 - e. Whether the feature should be ignored in modeling (by removing the column).
4. Calculate and write to Excel workbooks the covariance and correlation matrices for the numeric values in this data set. Create data frames with a row and a column for each numeric value; the entries in the cells are the covariances and the correlations for each pair of numeric features. Include the target value if it’s numeric.
5. From these workbooks, determine the three feature values (predictors) that are most highly correlated with the target; list them and their correlation. Note that either a large positive or a large negative correlation with the target indicates a good predictor. Also, find the three predictors that are the most highly correlated with each other; list them and their cross-correlation.

Tips:

- I used Pandas DataFrame and Series operations for nearly all of this assignment, but there are many ways to do it.
- Consider the exception processing in Python. Think about the following:
try:
 some operation that might fail if I try to run it on non-numeric data
except (TypeError, ValueError):
 what to do instead – set calculated stats to “N/A”, maybe
- Use the debugger in Python!

- I suggest creating a tiny test.xlsx file, to let you debug your code.

Your submission:

For your submission, please paste your Python code, your console output, and the contents of your three Excel workbooks into a single Word file for submission. Include your answers for questions 3 and 5 above. When you paste your code into Word, paste as text; don't paste a screenshot or use dark mode. Also, attach your .py file and the three output workbooks as separate file attachments. Please don't put them into a zip file! Submit your work via Canvas.