Data Exploration Through Python

John Smutny

Homework 2

ECE5984 Applications of ML

02/15/2022

Table of Contents

Abstract

Create various statistical summaries using python to better understand what bodily features contribute to a patient's risk of being diagnosed with Heart Disease. A Data Quality Report, Covariance Matrix and Correlation Matrix was created and included in this document (as well as the python code used to generate them). Python code used was started by Dr Creed Jones of Virginia Tech University and then added to by the submitting party.

Heart Disease Statistics Report

Instructions Provided: *“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):”*

Below shows two images of the console output of the python code used to generate a Data Quality Report summarizing data from the relevant ‘Heart Disease.xlsx’ file. The console output is divided into two images for improved readability.

Figures 1a & Figure 1b will be analyzed in the next section to better understand each data feature.

	stat	age	sex	cp	trestbps	chol	fbs	restecg
0	cardinality	41	2.000000	4.000000	49	152	3.000000	3.000000
1	mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.152027	0.528053
2	median	55.0	1.000000	1.000000	130.0	240.0	0.000000	1.000000
3	n_at_median	8	207.000000	50.000000	36	4	251.000000	152.000000
4	mode	58	1.000000	0.000000	120	197	0.000000	1.000000
5	n_at_mode	19	207.000000	143.000000	37	6	251.000000	152.000000
6	stddev	9.082101	0.466011	1.032052	17.538143	51.830751	0.359655	0.525860
7	min	29	0.000000	0.000000	94	126	0.000000	0.000000
8	max	77	1.000000	3.000000	200	564	1.000000	2.000000
9	n_zero	N/A	96.000000	143.000000	N/A	N/A	251.000000	147.000000
10	n_missing	0	0.000000	0.000000	0	0	7.000000	0.000000

Figure 1a: 1 of 2.

	thalach	exang	oldpeak	slope	ca	thal	target
0	91	2.000000	40.000000	3.000000	6.000000	4.000000	2.000000
1	149.646865	0.326733	1.039604	1.399340	0.741611	2.313531	0.544554
2	153.0	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
3	3	204.000000	13.000000	140.000000	170.000000	166.000000	165.000000
4	162	0.000000	0.000000	2.000000	0.000000	2.000000	1.000000
5	11	204.000000	99.000000	142.000000	170.000000	166.000000	165.000000
6	22.905161	0.469794	1.161075	0.616226	1.026753	0.612277	0.498835
7	71	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	202	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000
9	N/A	204.000000	99.000000	21.000000	170.000000	2.000000	138.000000
10	0	0.000000	0.000000	0.000000	5.000000	0.000000	0.000000

Figure 1b: 2 of 2.

Excel Statistics Report

Instructions Provided: "*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 😊😊):* "

Please see the appendix or the included xlsx file "HeartDisease-DataQualityReport.xlsx" for a divided table with a more readable font.

stat	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
cardinality	41	2	4	49	152	3	3	91	2	40	3	6	4	2
mean	54.36634	0.683168	0.966997	131.6238	246.264	0.152027	0.528053	149.6469	0.326733	1.039604	1.39934	0.741611	2.313531	0.544554
median	55	1	1	130	240	0	1	153	0	0.8	1	0	2	1
N_at_ median	8	207	50	36	4	251	152	3	204	13	140	170	166	165
mode	58	1	0	120	197	0	1	162	0	0	2	0	2	1
N_at_ mode	19	207	143	37	6	251	152	11	204	99	142	170	166	165
stddev	9.082101	0.466011	1.032052	17.53814	51.83075	0.359655	0.52586	22.90516	0.469794	1.161075	0.616226	1.026753	0.612277	0.498835
min	29	0	0	94	126	0	0	71	0	0	0	0	0	0
max	77	1	3	200	564	1	2	202	1	6.2	2	4	3	1
N_ zero	N/A	96	143	N/A	N/A	251	147	N/A	204	99	21	170	2	138
N_ missing	0	0	0	0	0	7	0	0	0	0	0	5	0	0

Table 1: Table showing various statistical values of measured features and the target variable (a patient's risk of Heart Disease).

Interpreting the Statistics Report

Instructions Provided: "*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, ordinal, text.)

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)"

The desired answers are provided below for all features in tables of three as well as the definition of each feature. Please see the appendix for a list view of all features and definitions.

Data Members: 1-3

	age	**sex**	**cp (chest pain type) { typical angina, atypical angina, non-anginal pain, asymptomatic }**
a. Feature Types & Type of Feature?	Feature - Continuous	Feature - Categorical	Feature - Categorical
b. How many values are missing and how many are invalid?	Missing Values: 0 Invalid Values: 0	Missing Values: 0 Invalid Values: 0 - Cardinality = 2	Missing Values: 0 Invalid Values: 0 - Cardinality = 4
c. What should be done about the missing values and how?	No action needed. There are zero missing values.	No action needed. There are zero missing values.	No action needed. There are zero missing values.
d. Whether the feature contains a significant number of outliers?	This feature does not contain outliers. A range of ages from 29 to 77 is realistic for humans.	This feature does not contain outliers. The data measures only two sexs {Male & Female}.	This feature does not contain outliers. The data lists four types of pain. The data has cardinality of four.
e. Whether the feature should be ignored in modeling (by removing the column)"?	This feature **should** be used in modeling.	This feature **should** be used in modeling.	This feature **should** be used in modeling.

Data Member(s): 4-6

	trestbps (resting blood pressure)	**chol (cholesterol in mg/dl)**	**fbs (resting blood sugar) {1 = fbs > 120mg/dl, 0 = false }**
a. Feature Types & Type of Feature?	Feature - Continuous	Feature - Continuous	Feature - Categorical
b. How many values are missing and how many are invalid?	Missing Values: 0 Invalid Values: 0	Missing Values: 0 Invalid Values: 0	Missing Values: 7 Invalid Values: 0
c. What should be done about the missing values and how?	No action needed. There are zero missing values.	No action needed. There are zero missing values.	No action is taken. Only 7/303 (2.3%) of samples are missing. Due to the binary nature of 'fbs' we can make predictions of the boolean value based on other features like 'trestbps', 'chol' and 'thalach' through KNN Imputation. But due to the amount of values missing, imputation should not be necessary.
d. Whether the feature contains a significant number of outliers?	This feature does not contain outliers. A range of blood pressure of 94 to 200 is normal, but the max of 200 is high.	This feature does not contain outliers. A range of cholesterol of 126 to 564 is normal but the 564 mg/dl max is noticeably high. Given the data studies heart disease. No action is taken.	This feature does not contain outliers. A boolean range of 0 to 1 is present.
e. Whether the feature should be ignored in modeling (by removing the column)"?	This feature **should** be used in modeling.	This feature **should** be used in modeling.	This feature **should** be used in modeling.

Data Member(s): 7-9

	restecg (resting electrocardiographic results) { 0 = normal, 1 = wave abnormality }	**thalach (maximum heart rate achieved)**	**exang (exercise induced angina) { 1 = yes, 0 = no}**
a. Feature Types & Type of Feature?	Feature - Categorical	Feature - Continuous	Feature - Categorical
b. How many values are missing and how many are invalid?	Missing Values: 147 Invalid Values: 4	Missing Values: 0 Invalid Values: 0	Missing Values: 0 Invalid Values: 0 - Cardinality = 2
c. What should be done about the missing values and how?	147/303 (48%) of samples are missing. Since 'restecg' is a complex medical test, and there is a significant number of values missing; it may be unrealistic to predict values for the 149 entries.	No action needed. There are zero missing values.	No action needed. There are zero missing values.
d. Whether the feature contains a significant number of outliers?	This feature does contain outliers (invalid values) but not a significant number (4/303). It is believed that these four entries were the result of 'data entry mistakes', that the '2' should have been a '1'.	This feature does not contain a significant number of outliers. A max heart range of 71 to 202 bpm is realistic, but the min of 71 is low. The patient was a 61 year old male without chest pain. No action is explicitly required, thus no action is taken.	This feature does not contain outliers. A boolean range of 0 to 1 is present.
e. Whether the feature should be ignored in modeling (by removing the column)"?	This feature **should not** be used in modeling. As described above in row C), this column should be removed.	This feature **should** be used in modeling.	This feature **should** be used in modeling.

Data Member(s): 10-12

	oldpeak (ST depression induced by exercise relative to rest)	**slope (slope of peak exercise ST segment) { 1 = upsloping, 2 = flat, 3 = downsloping }**	**ca (number of major vessels colored by fluoroscopy) { 0, 1, 2, 3 }**
a. Feature Types & Type of Feature?	Feature - Continuous	Feature - Ordinal	Feature - Ordinal
b. How many values are missing and how many are invalid?	Missing Values: 0 Invalid Values: 0	Missing Values: 0 Invalid Values: 0	Missing Values: 5 Invalid Values: 5
c. What should be done about the missing values and how?	No action needed. There are zero missing values.	No action needed. There are zero missing values.	5/303 (0.017%) of samples are missing. Similar to the 'fbs' data feature; stratified imputation (based on age & sex) or KNN Imputation can be used to fill empty values if deemed necessary.
d. Whether the feature contains a significant number of outliers?	This feature does not contain a significant amount of outliers. There are entries above 4.5 (away from the median (0.8) and mean (1.04)), but the occurrence of other values above 3.0 suggests a highly right-skewed distribution of values..	This feature does not contain outliers.	5/303 (0.017%) of samples are outliers/invalid. Since the outliers are all entries of value '4', it is assumed that this is a data entry error. However, since there are multiple categories, it cannot be precisely judged what value '4' should represent; maybe the largest category ('3') or the standard keyboard value just below '4' ('1').
e. Whether the feature should be ignored in modeling (by removing the column)"?	This feature **should** be used in modeling.	This feature **should** be used in modeling.	This feature **should** be used in modeling.

Data Member(s): 13-14

	thal **{ 3 = normal,** **6 = fixed defect,** **7 = reversible defect }**	**target** **(diagnosis of heart disease)** **{ 0 = < 50% diameter narrowing,** **1 = > 50% diameter narrowing }**
a. Feature Types & Type of Feature?	Feature - Categorical	Target - Continuous
b. How many values are missing and how many are invalid?	Missing Values: 0 Invalid Values: 303	Missing Values: 0 Invalid Values: 0
c. What should be done about the missing values and how?	No action needed. There are zero missing values.	No action needed. There are zero missing values.
d. Whether the feature contains a significant number of outliers?	303/303 (100%) of values are outside of the defined acceptable values { 3,6,7 }. The majority (301/303 entries) are in the range of { 1,2,3 }. It is believed that this is a Feature Definition error, that the actual possible values should be { 1,2,3 } instead of the defined { 3,6,7 }. Therefore, 1) all '0' entries will be deemed 'invalid values' and replaced with blank values since they are an insignificant part of the data set. 2) The Feature's description will be re-documented as '{ 1 = normal, 2 = fixed defect, 3 = reversible defect }**'.** These actions will lead to an outlier/invalid rate of 2/303 (0.006%).	The target does not contain outliers. A boolean range of 0 to 1 is present.
e. Whether the feature should be ignored in modeling (by removing the column)"?	This feature **should** be used in modeling.	The target (by definition) **should** be used in modeling.

Covariance & Correlation Matrices from Report

Instructions Provided: ”*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.*”

Both the Covariance and Correlation matrices of various features and heart health are listed on the following four pages. Both of the matrices were created using pandas.dataframe built in functions .cov and .corr respectively.

Covariance Matrix

Please see the appendix or the included xlsx file “HeartDisease-CovarianceMatrix.xlsx” for a divided table with a more readable font.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	82.48456	-0.41666	-0.6435	44.4959	100.5851	0.384963	-0.55501	-82.9033	0.413022	2.214583	-0.94479	2.565035	0.378139	-1.02134
sex	-0.41666	0.217166	-0.02374	-0.46397	-4.78031	0.007249	-0.01426	-0.46987	0.031014	0.051993	-0.00882	0.056267	0.05993	-0.06531
cp	-0.6435	-0.02374	1.065132	0.861714	-4.11377	0.035147	0.024108	6.991618	-0.19117	-0.17882	0.076137	-0.18889	-0.1022	0.22333
trestbps	44.4959	-0.46397	0.861714	307.5865	111.9672	1.124347	-1.05232	-18.7591	0.557111	3.934486	-1.31283	1.816227	0.668022	-1.26795
chol	100.5851	-4.78031	-4.11377	111.9672	2686.427	0.286887	-4.1167	-11.8005	1.631991	3.246794	-0.12896	3.914232	3.135488	-2.20386
fbs	0.384963	0.007249	0.035147	1.124347	0.286887	0.129352	-0.01547	-0.00638	0.003733	0.002806	-0.01335	0.051147	-0.00828	-0.00346
restecg	-0.55501	-0.01426	0.024108	-1.05232	-4.1167	-0.01547	0.276528	0.531462	-0.01747	-0.03588	0.030151	-0.03849	-0.00386	0.035998
thalach	-82.9033	-0.46987	6.991618	-18.7591	-11.8005	-0.00638	0.531462	524.6464	-4.07629	-9.15352	5.459369	-4.82319	-1.35249	4.818766
exang	0.413022	0.031014	-0.19117	0.557111	1.631991	0.003733	-0.01747	-4.07629	0.220707	0.157216	-0.07462	0.054957	0.059472	-0.10235
oldpeak	2.214583	0.051993	-0.17882	3.934486	3.246794	0.002806	-0.03588	-9.15352	0.157216	1.348095	-0.41322	0.277342	0.149462	-0.24945
slope	-0.94479	-0.00882	0.076137	-1.31283	-0.12896	-0.01335	0.030151	5.459369	-0.07462	-0.41322	0.379735	-0.05385	-0.03953	0.106321
ca	2.565035	0.056267	-0.18889	1.816227	3.914232	0.051147	-0.03849	-4.82319	0.054957	0.277342	-0.05385	1.054222	0.090254	-0.1975
thal	0.378139	0.05993	-0.1022	0.668022	3.135488	-0.00828	-0.00386	-1.35249	0.059472	0.149462	-0.03953	0.090254	0.374883	-0.10508
target	-1.02134	-0.06531	0.22333	-1.26795	-2.20386	-0.00346	0.035998	4.818766	-0.10235	-0.24945	0.106321	-0.1975	-0.10508	0.248836

Table 2: Table showing covariance values between measured features and the target variable (a patient’s risk of Heart Disease).

Correlation Matrix

Please see the appendix or the included xlsx file “HeartDisease-CorrelationMatrix.xlsx” for a divided table with a more readable font.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.09845	-0.06865	0.279351	0.213678	0.117935	-0.11621	-0.39852	0.096801	0.210013	-0.16881	0.275932	0.068001	-0.22544
sex	-0.09845	1	-0.04935	-0.05677	-0.19791	0.043348	-0.0582	-0.04402	0.141664	0.096093	-0.03071	0.117733	0.210041	-0.28094
cp	-0.06865	-0.04935	1	0.047608	-0.0769	0.094015	0.044421	0.295762	-0.39428	-0.14923	0.119717	-0.17849	-0.16174	**0.433798**
trestbps	0.279351	-0.05677	0.047608	1	0.123174	0.176699	-0.1141	-0.0467	0.067616	0.193216	-0.12147	0.100268	0.06221	-0.14493
chol	0.213678	-0.19791	-0.0769	0.123174	1	0.015392	-0.15104	-0.00994	0.067023	0.053952	-0.00404	0.073564	0.098803	-0.08524
fbs	0.117935	0.043348	0.094015	0.176699	0.015392	1	-0.08168	-0.00077	0.022021	0.006707	-0.06054	0.139262	-0.03761	-0.01924
restecg	-0.11621	-0.0582	0.044421	-0.1141	-0.15104	-0.08168	1	0.044123	-0.07073	-0.05877	0.093045	-0.07123	-0.01198	0.13723
thalach	-0.39852	-0.04402	0.295762	-0.0467	-0.00994	-0.00077	0.044123	1	-0.37881	-0.34419	0.386784	-0.20528	-0.09644	0.421741
exang	0.096801	0.141664	-0.39428	0.067616	0.067023	0.022021	-0.07073	-0.37881	1	0.288223	-0.25775	0.11374	0.206754	**-0.43676**
oldpeak	0.210013	0.096093	-0.14923	0.193216	0.053952	0.006707	-0.05877	-0.34419	0.288223	1	-0.57754	0.233027	0.210244	**-0.4307**
slope	-0.16881	-0.03071	0.119717	-0.12147	-0.00404	-0.06054	0.093045	0.386784	-0.25775	-0.57754	1	-0.0863	-0.10476	0.345877
ca	0.275932	0.117733	-0.17849	0.100268	0.073564	0.139262	-0.07123	-0.20528	0.11374	0.233027	-0.0863	1	0.143738	-0.38511
thal	0.068001	0.210041	-0.16174	0.06221	0.098803	-0.03761	-0.01198	-0.09644	0.206754	0.210244	-0.10476	0.143738	1	-0.34403
target	-0.22544	-0.28094	0.433798	-0.14493	-0.08524	-0.01924	0.13723	0.421741	-0.43676	-0.4307	0.345877	-0.38511	-0.34403	1

Table 3B: Table 2 of 2. Table showing correlation values between measured features and the target variable (a patient’s risk of Heart Disease). **Bold** numbers represent the greatest correlations.

Predicting Modeling Features and Correlation

Instructions Provided: "*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.*"

Features Correlated to the Target

After analyzing the Heart-Disease data for relationships between measured features and the target (whether or not a patient had Heart Disease); several correlations became more visible. The three features that were most correlated to the target variable were;

1. exang and target: -0.43676
2. cp and target: 0.433798
3. oldpeak and target: -0.4307

These three features infer that there is a correlation of an increased narrowing diameter of vanes if you did not have exercise induced angina, have abnormal chest pain, and a lower ST induced depression.

Features Correlated to One Another

After analyzing the Heart-Disease data for relationships between measured features and the target (whether or not a patient had Heart Disease); several correlations became more visible between features. The three features that were most correlated to the each other were;

1. slope and oldpeak: -0.57754
2. thalach and age: -0.39852
3. thalach and slope: 0.386784

These three features infer that they are significantly more influenced by one another than other features. Correlation 1's (slope&oldpeak) mild correlation (negative or positive) is reasonable since both are measures of a ST Measure done (oldpeak is the measure minimum and slope is related to the rate of the measurement). Correlation 2 (thalach (maximum heart rate) and age) is also reasonable to be negatively correlated, since younger individuals will have stronger hearts capable of more vigorous activity than older individuals.

References

Data Set 1

- Name: Heart Disease.xlsx
- Source: Dr Creed Jones
- Format: xlsx for Excel
- Applications Used: Excel

Appendix

Abbreviation Definitions

List of feature abbreviation definitions in the “Heart Disease.xlsx” file supplied from the University of California Irvine’s Machine Learning Repository.

- https://archive.ics.uci.edu/ml/index.php

Please see the link provided for the full list of features. The list of relevant feature definitions are provided below from the dataset’s author: Andras Janosi MD, William Steinbrunn MD, Matthias Pfisterer MD, and Robert Detrano MD PhD

- https://archive.ics.uci.edu/ml/datasets/Heart+Disease

3 **age**: age in years
4 **sex**: sex (1 = male; 0 = female)
9 **cp**: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
10 **trestbps**: resting blood pressure (in mm Hg on admission to the hospital)
12 **chol**: serum cholestoral in mg/dl
16 **fbs**: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
19 **restecg**: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
32 **thalach**: maximum heart rate achieved
38 **exang**: exercise induced angina (1 = yes; 0 = no)
40 **oldpeak** = ST depression induced by exercise relative to rest
41 **slope**: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
44 **ca**: number of major vessels (0-3) colored by flourosopy
51 **thal**: 3 = normal; 6 = fixed defect; 7 = reversable defect
58 **num**: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing

Python Code

```
#########################################################################
#   File:   hw_dataQualityReport.py
#   Name:   John Smutny
#   Course: ECE-5984: Applications of Machine Learning
#   Date:   02/15/2022
#   Description:
#       Use numpy to perform statistical analysis on a datasets.
#       Then use Panda DataFrames to create a Data Quality Report.
#########################################################################


import pandas
import numpy as np
from Libraries.DataExploration.DataQualityReport import DataQualityReport

####################################
# Initial loading of data
filename = 'C:/Data/HeartDisease.xlsx'
df = pandas.read_excel(filename)  # read an Excel spreadsheet

####################################
# Dissect the data into text labels, features, and the desired target
variable.
print('File {0} is of size {1}'.format(filename, df.shape))
labels = df.columns
featureLabels = labels.drop('target').values  # get just the predictors
xFrame = df[featureLabels]
yFrame = df['target']  # and the target variable
predictors = xFrame.to_numpy(np.float64)  # convert them to numpy arrays
target = yFrame.to_numpy(np.float64)

####################################
# Create an organized data set summary for the console using a data frame.
report = DataQualityReport()

for thisLabel in labels:  # for each column, report basic stats
   thisCol = df[thisLabel]
   report.addCol(thisLabel, thisCol)

print(report.to_string())


####################################
# Print all reports to statistics report to excel
# 1) Statistics Report
outFilename = "C:/Data/HeartDisease-DataQualityReport.xlsx"
report.statsdf.to_excel(sheet_name='DataQualityReport',
```

```
                        excel_writer=outFilename)

# 2) Covariance Matrix (using .cov() dataframe function)
outFilename = "C:/Data/HeartDisease-CovarianceMatrix.xlsx"
df.cov().to_excel(sheet_name='CovarianceMatrix', excel_writer=outFilename)

# 3) Correlation Matrix (using .corr() dataframe function)
outFilename = "C:/Data/HeartDisease-CorrelationMatrix.xlsx"
df.corr().to_excel(sheet_name='CorrelationMatrix', excel_writer=outFilename)



#############################################################################
#   Contributor: John Smutny
#   Original Author: Dr Creed Jones
#
#   statsdf     data frame summarizing the Data Quality Report from a dataset.
#############################################################################
import pandas

class DataQualityReport:
   # Contructor. Define what values are being calculated for the stats report.
   def __init__(self):
       self.statsdf = pandas.DataFrame()
       self.statsdf['stat'] = ['cardinality',
                               'mean',
                               'median',
                               'n_at_median',
                               'mode',
                               'n_at_mode',
                               'stddev',
                               'min',
                               'max',
                               'n_zero',
                               'n_missing']
       pass

   # Add a 'feature' to be included on the report. Append columns by 1.
   def addCol(self, label, data):
       cardinalityV = len(data.unique())
       meanV = data.mean()
       medianV = data.median()
       modeV = data.mode()[0]
       stdDev = data.std()
       minV = data.min()
       maxV = data.max()
       n_missing = data.isnull().sum()

       # Various qualities throw errors when trying to calculate them.
```

```
        #   1) 'value_counts' cannot handle situations where there are multiple
        #   median or mode values.
        try:
            n_medianV = data.value_counts()[medianV]
        except(TypeError, ValueError, KeyError):
            n_medianV = "N/A"

        try:
            n_modeV = data.value_counts()[modeV]
        except(TypeError, ValueError, KeyError):
            n_modeV = "N/A"

        #   2) 'value_counts()[0]' to find zero fields cannot handle blank
        #   values.
        try:
            n_zeros = data.value_counts()[0]
        except(TypeError, ValueError, KeyError):
            n_zeros = "N/A"

        self.statsdf[label] = [cardinalityV,
                               meanV,
                               medianV,
                               n_medianV,
                               modeV,
                               n_modeV,
                               stdDev,
                               minV,
                               maxV,
                               n_zeros,
                               n_missing]

    def to_string(self):
        return self.statsdf.to_string()
```

Data Quality Report (Improved Readability)

	age	sex	cp	trestbps	chol	fbs	restecg
cardinality	41	2	4	49	152	3	3
mean	54.36634	0.683168	0.966997	131.6238	246.264	0.152027	0.528053
median	55	1	1	130	240	0	1
N_at_ median	8	207	50	36	4	251	152
mode	58	1	0	120	197	0	1
N_at_mode	19	207	143	37	6	251	152
stddev	9.082101	0.466011	1.032052	17.53814	51.83075	0.359655	0.52586
min	29	0	0	94	126	0	0
max	77	1	3	200	564	1	2
N_zero	N/A	96	143	N/A	N/A	251	147
N_missing	0	0	0	0	0	7	0

Table 1A: Table 1 of 2. Table showing various statistical values of measured features.

	thalach	exang	oldpeak	slope	ca	thal	target
cardinality	91	2	40	3	6	4	2
mean	149.6469	0.326733	1.039604	1.39934	0.741611	2.313531	0.544554
median	153	0	0.8	1	0	2	1
N_at_ median	3	204	13	140	170	166	165
mode	162	0	0	2	0	2	1
N_at_mode	11	204	99	142	170	166	165
stddev	22.90516	0.469794	1.161075	0.616226	1.026753	0.612277	0.498835
min	71	0	0	0	0	0	0
max	202	1	6.2	2	4	3	1
N_zero	N/A	204	99	21	170	2	138
N_missing	0	0	0	0	5	0	0

Table 1B: Table 2 of 2. Table showing various statistical values of measured features and the target variable (a patient's risk of Heart Disease).

Correlation Matrix (Improved Readability)

	age	sex	cp	trestbps	chol	fbs	restecg
age	82.48456	-0.41666	-0.6435	44.4959	100.5851	0.384963	-0.55501
sex	-0.41666	0.217166	-0.02374	-0.46397	-4.78031	0.007249	-0.01426
cp	-0.6435	-0.02374	1.065132	0.861714	-4.11377	0.035147	0.024108
trestbps	44.4959	-0.46397	0.861714	307.5865	111.9672	1.124347	-1.05232
chol	100.5851	-4.78031	-4.11377	111.9672	2686.427	0.286887	-4.1167
fbs	0.384963	0.007249	0.035147	1.124347	0.286887	0.129352	-0.01547
restecg	-0.55501	-0.01426	0.024108	-1.05232	-4.1167	-0.01547	0.276528
thalach	-82.9033	-0.46987	6.991618	-18.7591	-11.8005	-0.00638	0.531462
exang	0.413022	0.031014	-0.19117	0.557111	1.631991	0.003733	-0.01747
oldpeak	2.214583	0.051993	-0.17882	3.934486	3.246794	0.002806	-0.03588
slope	-0.94479	-0.00882	0.076137	-1.31283	-0.12896	-0.01335	0.030151
ca	2.565035	0.056267	-0.18889	1.816227	3.914232	0.051147	-0.03849
thal	0.378139	0.05993	-0.1022	0.668022	3.135488	-0.00828	-0.00386
target	-1.02134	-0.06531	0.22333	-1.26795	-2.20386	-0.00346	0.035998

Table 2A: Table 1 of 2. Table showing correlation values between measured features.

	thalach	exang	oldpeak	slope	ca	thal	target
age	-82.9033	0.413022	2.214583	-0.94479	2.565035	0.378139	-1.02134
sex	-0.46987	0.031014	0.051993	-0.00882	0.056267	0.05993	-0.06531
cp	6.991618	-0.19117	-0.17882	0.076137	-0.18889	-0.1022	0.22333
trestbps	-18.7591	0.557111	3.934486	-1.31283	1.816227	0.668022	-1.26795
chol	-11.8005	1.631991	3.246794	-0.12896	3.914232	3.135488	-2.20386
fbs	-0.00638	0.003733	0.002806	-0.01335	0.051147	-0.00828	-0.00346
restecg	0.531462	-0.01747	-0.03588	0.030151	-0.03849	-0.00386	0.035998
thalach	524.6464	-4.07629	-9.15352	5.459369	-4.82319	-1.35249	4.818766
exang	-4.07629	0.220707	0.157216	-0.07462	0.054957	0.059472	-0.10235
oldpeak	-9.15352	0.157216	1.348095	-0.41322	0.277342	0.149462	-0.24945
slope	5.459369	-0.07462	-0.41322	0.379735	-0.05385	-0.03953	0.106321
ca	-4.82319	0.054957	0.277342	-0.05385	1.054222	0.090254	-0.1975
thal	-1.35249	0.059472	0.149462	-0.03953	0.090254	0.374883	-0.10508
target	4.818766	-0.10235	-0.24945	0.106321	-0.1975	-0.10508	0.248836

Table 2B: Table 2 of 2. Table showing covariance values between measured features and the target variable (a patient's risk of Heart Disease).

Correlation Matrix (Improved Readability)

	age	sex	cp	trestbps	chol	fbs	restecg
age	1	-0.09845	-0.06865	0.279351	0.213678	0.117935	-0.11621
sex	-0.09845	1	-0.04935	-0.05677	-0.19791	0.043348	-0.0582
cp	-0.06865	-0.04935	1	0.047608	-0.0769	0.094015	0.044421
trestbps	0.279351	-0.05677	0.047608	1	0.123174	0.176699	-0.1141
chol	0.213678	-0.19791	-0.0769	0.123174	1	0.015392	-0.15104
fbs	0.117935	0.043348	0.094015	0.176699	0.015392	1	-0.08168
restecg	-0.11621	-0.0582	0.044421	-0.1141	-0.15104	-0.08168	1
thalach	-0.39852	-0.04402	0.295762	-0.0467	-0.00994	-0.00077	0.044123
exang	0.096801	0.141664	-0.39428	0.067616	0.067023	0.022021	-0.07073
oldpeak	0.210013	0.096093	-0.14923	0.193216	0.053952	0.006707	-0.05877
slope	-0.16881	-0.03071	0.119717	-0.12147	-0.00404	-0.06054	0.093045
ca	0.275932	0.117733	-0.17849	0.100268	0.073564	0.139262	-0.07123
thal	0.068001	0.210041	-0.16174	0.06221	0.098803	-0.03761	-0.01198
target	-0.22544	-0.28094	0.433798	-0.14493	-0.08524	-0.01924	0.13723

Table 3A: Table 1 of 2. Table showing correlation values between measured features.

	thalach	exang	oldpeak	slope	ca	thal	target
age	-0.39852	0.096801	0.210013	-0.16881	0.275932	0.068001	-0.22544
sex	-0.04402	0.141664	0.096093	-0.03071	0.117733	0.210041	-0.28094
cp	0.295762	-0.39428	-0.14923	0.119717	-0.17849	-0.16174	**0.433798**
trestbps	-0.0467	0.067616	0.193216	-0.12147	0.100268	0.06221	-0.14493
chol	-0.00994	0.067023	0.053952	-0.00404	0.073564	0.098803	-0.08524
fbs	-0.00077	0.022021	0.006707	-0.06054	0.139262	-0.03761	-0.01924
restecg	0.044123	-0.07073	-0.05877	0.093045	-0.07123	-0.01198	0.13723
thalach	1	-0.37881	-0.34419	0.386784	-0.20528	-0.09644	**0.421741**
exang	-0.37881	1	0.288223	-0.25775	0.11374	0.206754	**-0.43676**
oldpeak	-0.34419	0.288223	1	-0.57754	0.233027	0.210244	-0.4307
slope	0.386784	-0.25775	-0.57754	1	-0.0863	-0.10476	0.345877
ca	-0.20528	0.11374	0.233027	-0.0863	1	0.143738	-0.38511
thal	-0.09644	0.206754	0.210244	-0.10476	0.143738	1	-0.34403
target	0.421741	-0.43676	-0.4307	0.345877	-0.38511	-0.34403	1

Table 3B: Table 2 of 2. Table showing correlation values between measured features and the target variable (a patient's risk of Heart Disease). **Bold** numbers represent the greatest correlations.