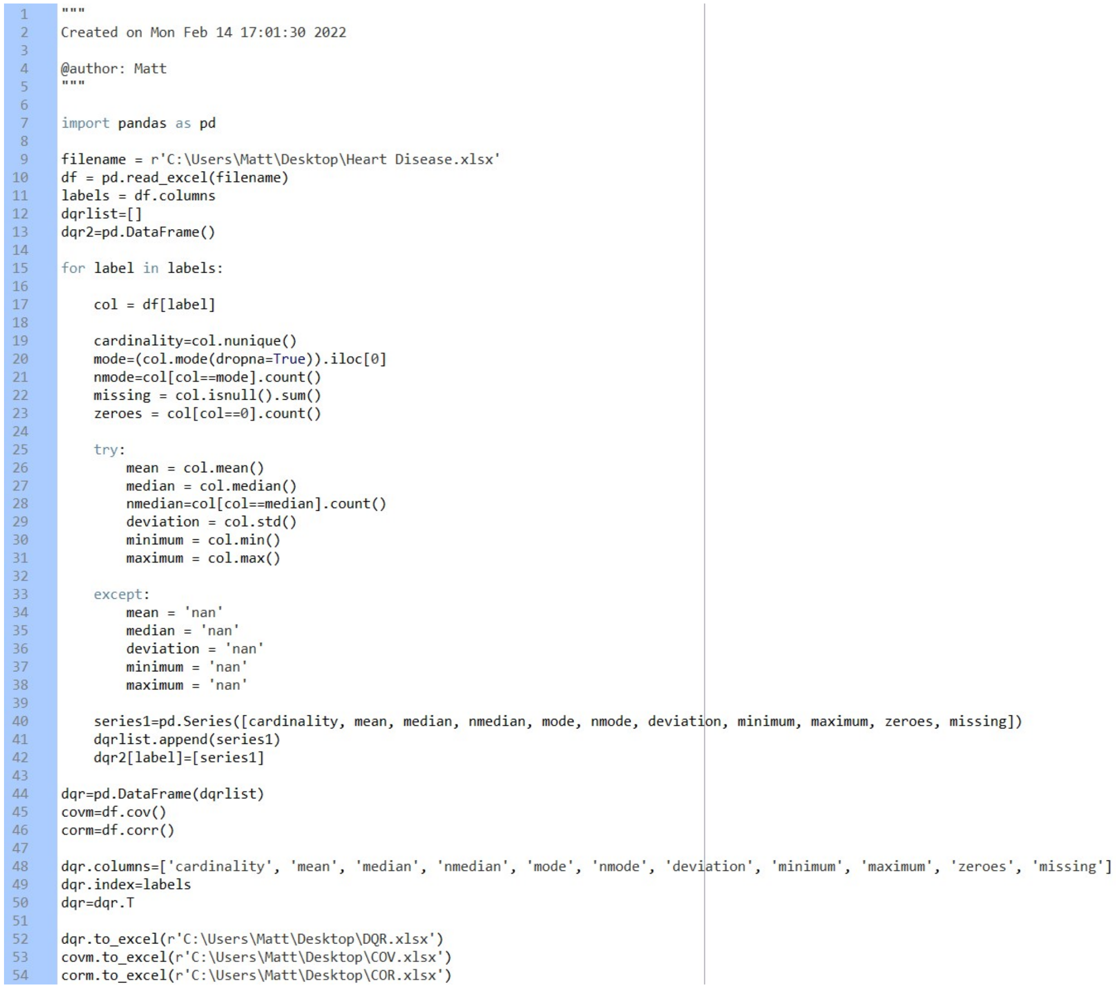
ECE 5984 MACHINE LEARNING HOMEWORK 2

MATT FERGUSON

Python Code

Here we can see my python code. It loads the heart disease data and then loops through each column and calculates the quantitative and qualitative statistics and adds them to a new data frame we name as our Data Quality Report.

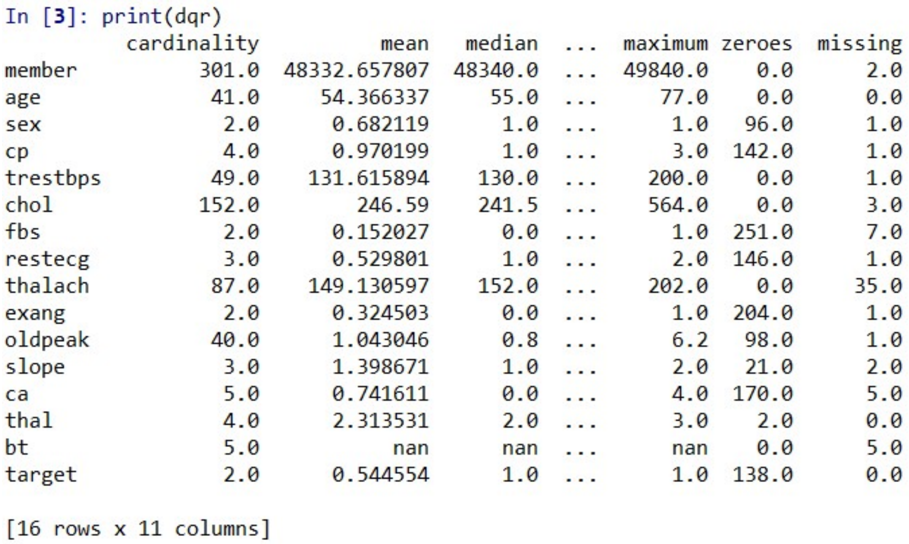
To compute quantitative and qualitative statistics a try/except logic is employed. If quantitative statistics fail to compute we handle the exception by placing the values of statistics appropriate for qualitative labels ('nan').



```
"""
Created on Mon Feb 14 17:01:30 2022

@author: Matt
"""

import pandas as pd

filename = r'C:\Users\Matt\Desktop\Heart Disease.xlsx'
df = pd.read_excel(filename)
labels = df.columns
dqrlist=[]
dqr2=pd.DataFrame()

for label in labels:

    col = df[label]

    cardinality=col.nunique()
    mode=(col.mode(dropna=True)).iloc[0]
    nmode=col[col==mode].count()
    missing = col.isnull().sum()
    zeroes = col[col==0].count()

    try:
        mean = col.mean()
        median = col.median()
        nmedian=col[col==median].count()
        deviation = col.std()
        minimum = col.min()
        maximum = col.max()

    except:
        mean = 'nan'
        median = 'nan'
        deviation = 'nan'
        minimum = 'nan'
        maximum = 'nan'

    series1=pd.Series([cardinality, mean, median, nmedian, mode, nmode, deviation, minimum, maximum, zeroes, missing])
    dqrlist.append(series1)
    dqr2[label]=[series1]

dqr=pd.DataFrame(dqrlist)
covm=df.cov()
corm=df.corr()

dqr.columns=['cardinality', 'mean', 'median', 'nmedian', 'mode', 'nmode', 'deviation', 'minimum', 'maximum', 'zeroes', 'missing']
dqr.index=labels
dqr=dqr.T

dqr.to_excel(r'C:\Users\Matt\Desktop\DQR.xlsx')
covm.to_excel(r'C:\Users\Matt\Desktop\COV.xlsx')
corm.to_excel(r'C:\Users\Matt\Desktop\COR.xlsx')
```

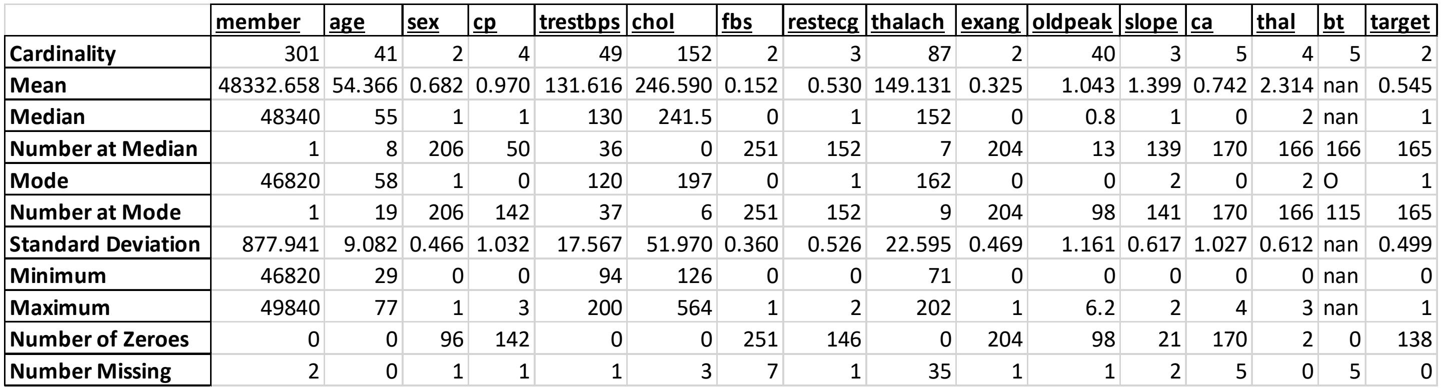
Python Console Output

This is the output the console makes when given a print(dqr) command. We can see that the data has been successfully operated on to produce the Data Quality Report.



```
In [3]: print(dqr)
          cardinality          mean   median  ...  maximum zeroes  missing
member          301.0  48332.657807  48340.0  ...  49840.0    0.0      2.0
age              41.0     54.366337     55.0  ...     77.0    0.0      0.0
sex               2.0      0.682119      1.0  ...      1.0   96.0      1.0
cp                4.0      0.970199      1.0  ...      3.0  142.0      1.0
trestbps         49.0    131.615894    130.0  ...    200.0    0.0      1.0
chol            152.0        246.59    241.5  ...    564.0    0.0      3.0
fbs               2.0      0.152027      0.0  ...      1.0  251.0      7.0
restecg           3.0      0.529801      1.0  ...      2.0  146.0      1.0
thalach          87.0    149.130597    152.0  ...    202.0    0.0     35.0
exang             2.0      0.324503      0.0  ...      1.0  204.0      1.0
oldpeak          40.0      1.043046      0.8  ...      6.2   98.0      1.0
slope             3.0      1.398671      1.0  ...      2.0   21.0      2.0
ca                5.0      0.741611      0.0  ...      4.0  170.0      5.0
thal              4.0      2.313531      2.0  ...      3.0    2.0      0.0
bt                5.0           nan      nan  ...      nan    0.0      5.0
target            2.0      0.544554      1.0  ...      1.0  138.0      0.0

[16 rows x 11 columns]
```

Data Quality Report



	member	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	bt	target
Cardinality	301	41	2	4	49	152	2	3	87	2	40	3	5	4	5	2
Mean	48332.658	54.366	0.682	0.970	131.616	246.590	0.152	0.530	149.131	0.325	1.043	1.399	0.742	2.314	nan	0.545
Median	48340	55	1	1	130	241.5	0	1	152	0	0.8	1	0	2	nan	1
Number at Median	1	8	206	50	36	0	251	152	7	204	13	139	170	166	166	165
Mode	46820	58	1	0	120	197	0	1	162	0	0	2	0	2	O	1
Number at Mode	1	19	206	142	37	6	251	152	9	204	98	141	170	166	115	165
Standard Deviation	877.941	9.082	0.466	1.032	17.567	51.970	0.360	0.526	22.595	0.469	1.161	0.617	1.027	0.612	nan	0.499
Minimum	46820	29	0	0	94	126	0	0	71	0	0	0	0	0	nan	0
Maximum	49840	77	1	3	200	564	1	2	202	1	6.2	2	4	3	nan	1
Number of Zeroes	0	0	96	142	0	0	251	146	0	204	98	21	170	2	0	138
Number Missing	2	0	1	1	1	3	7	1	35	1	1	2	5	0	5	0

Above we can see the full Data Quality Report, on the next page we will assess this data.

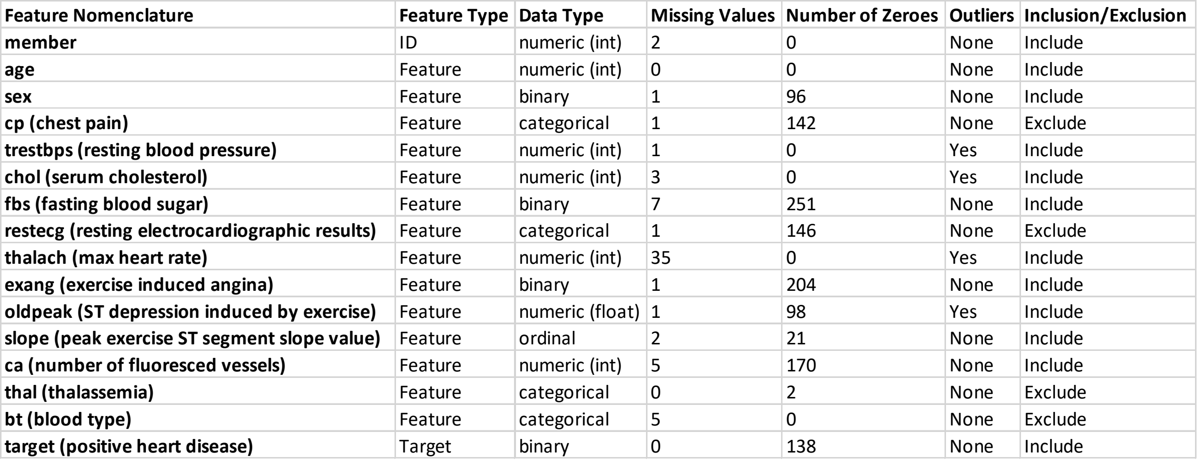
Data Quality Report Assessment

Outliers were computed as being greater than 3 standard deviations from the mean.

These items marked as excluded are not included with the correlation matrix we present. The exclusions by and large are made due to research of the metric showing it is categorical in nature. The exception being slope which we designate as ordinal due to high values of slope indicating down sloping during the peak ST exercise segment.

Thalach (maximum heart rate) has the most missing values and could be computed using SMOTE or another technique which determines which records are most similar to the ones with missing thalach values and weight those nearest neighbors to compute/interpolate the missing thalach.

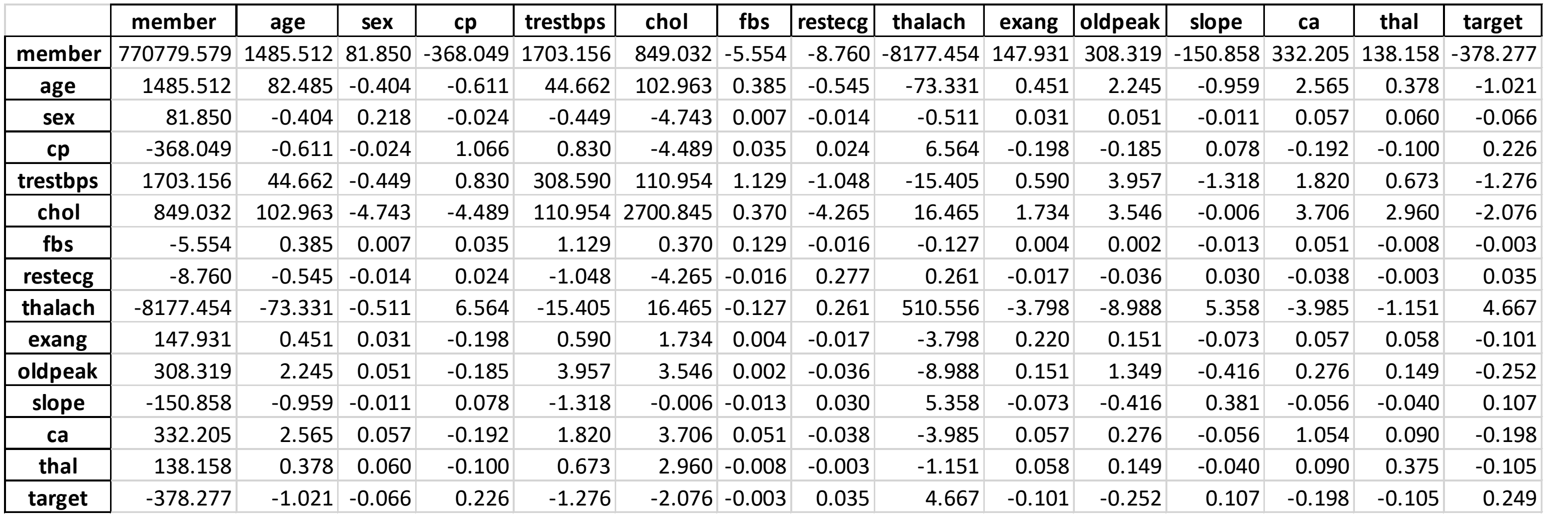
For the two member IDs which are missing we can simply generate new member ids.



Feature Nomenclature	Feature Type	Data Type	Missing Values	Number of Zeroes	Outliers	Inclusion/Exclusion
member	ID	numeric (int)	2	0	None	Include
age	Feature	numeric (int)	0	0	None	Include
sex	Feature	binary	1	96	None	Include
cp (chest pain)	Feature	categorical	1	142	None	Exclude
trestbps (resting blood pressure)	Feature	numeric (int)	1	0	Yes	Include
chol (serum cholesterol)	Feature	numeric (int)	3	0	Yes	Include
fbs (fasting blood sugar)	Feature	binary	7	251	None	Include
restecg (resting electrocardiographic results)	Feature	categorical	1	146	None	Exclude
thalach (max heart rate)	Feature	numeric (int)	35	0	Yes	Include
exang (exercise induced angina)	Feature	binary	1	204	None	Include
oldpeak (ST depression induced by exercise)	Feature	numeric (float)	1	98	Yes	Include
slope (peak exercise ST segment slope value)	Feature	ordinal	2	21	None	Include
ca (number of fluoresced vessels)	Feature	numeric (int)	5	170	None	Include
thal (thalassemia)	Feature	categorical	0	2	None	Exclude
bt (blood type)	Feature	categorical	5	0	None	Exclude
target (positive heart disease)	Target	binary	0	138	None	Include

The number of zeros for oldpeak is potentially concerning as we see about a third of the population experiences zero ST depression in response to exercise. Oldpeak and ca are the only numeric features with large amounts of zeroes, and zero values for ca make sense intuitively as zero fluorescence of vessels likely indicates a blockage of some kind. We converted sex to binary so every zero represents a female, and the missing value for male/female may be someone who didn't disclose their sex. Fasting blood sugar, presences of heart disease, and exercise induced angina are also binary so the presences of many zeroes makes sense. Slope is ordinal so the presence of zeroes means that slope is at a maximum, increased values representing neutral and downslope progressively. The remaining zeroes are categorical in nature and simply represent the first category value of the feature.

Covariance Matrix



	member	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
member	770779.579	1485.512	81.850	-368.049	1703.156	849.032	-5.554	-8.760	-8177.454	147.931	308.319	-150.858	332.205	138.158	-378.277
age	1485.512	82.485	-0.404	-0.611	44.662	102.963	0.385	-0.545	-73.331	0.451	2.245	-0.959	2.565	0.378	-1.021
sex	81.850	-0.404	0.218	-0.024	-0.449	-4.743	0.007	-0.014	-0.511	0.031	0.051	-0.011	0.057	0.060	-0.066
cp	-368.049	-0.611	-0.024	1.066	0.830	-4.489	0.035	0.024	6.564	-0.198	-0.185	0.078	-0.192	-0.100	0.226
trestbps	1703.156	44.662	-0.449	0.830	308.590	110.954	1.129	-1.048	-15.405	0.590	3.957	-1.318	1.820	0.673	-1.276
chol	849.032	102.963	-4.743	-4.489	110.954	2700.845	0.370	-4.265	16.465	1.734	3.546	-0.006	3.706	2.960	-2.076
fbs	-5.554	0.385	0.007	0.035	1.129	0.370	0.129	-0.016	-0.127	0.004	0.002	-0.013	0.051	-0.008	-0.003
restecg	-8.760	-0.545	-0.014	0.024	-1.048	-4.265	-0.016	0.277	0.261	-0.017	-0.036	0.030	-0.038	-0.003	0.035
thalach	-8177.454	-73.331	-0.511	6.564	-15.405	16.465	-0.127	0.261	510.556	-3.798	-8.988	5.358	-3.985	-1.151	4.667
exang	147.931	0.451	0.031	-0.198	0.590	1.734	0.004	-0.017	-3.798	0.220	0.151	-0.073	0.057	0.058	-0.101
oldpeak	308.319	2.245	0.051	-0.185	3.957	3.546	0.002	-0.036	-8.988	0.151	1.349	-0.416	0.276	0.149	-0.252
slope	-150.858	-0.959	-0.011	0.078	-1.318	-0.006	-0.013	0.030	5.358	-0.073	-0.416	0.381	-0.056	-0.040	0.107
ca	332.205	2.565	0.057	-0.192	1.820	3.706	0.051	-0.038	-3.985	0.057	0.276	-0.056	1.054	0.090	-0.198
thal	138.158	0.378	0.060	-0.100	0.673	2.960	-0.008	-0.003	-1.151	0.058	0.149	-0.040	0.090	0.375	-0.105
target	-378.277	-1.021	-0.066	0.226	-1.276	-2.076	-0.003	0.035	4.667	-0.101	-0.252	0.107	-0.198	-0.105	0.249

Above we can see the covariance matrix output by our data exploration python script. Excluded features can be seen in this covariance matrix but are not meaningful. We do not make any inferences from this covariance matrix for this homework, and a correlation matrix is presented on the next slide with excluded features removed for cleanliness.

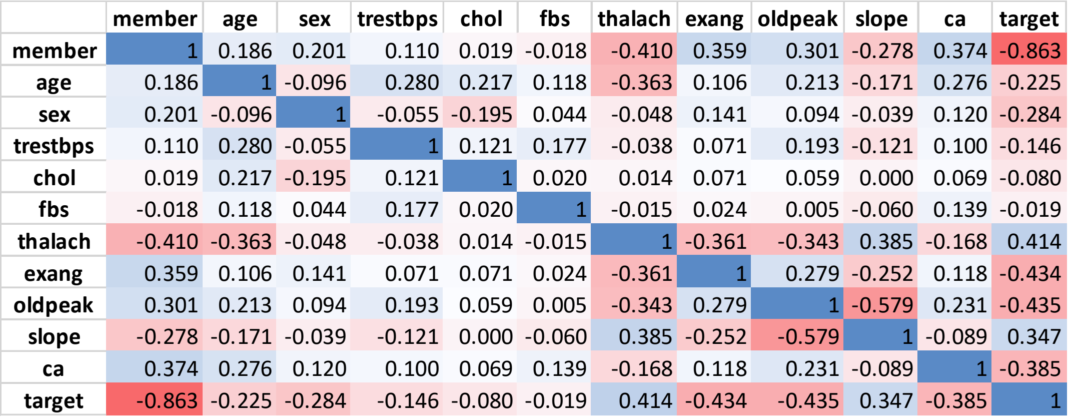
Correlation Matrix

We see the anticorrelation of -0.863 between our target of and member ID. This is because member ID's above 48460 are negative for heart disease and those at or below are positive for heart disease. This is a false correlation!

In terms of the remaining significant correlations we see a correlation of 0.414 between heart disease and maximum heart rate. The slope of the peak ST segment during exercise was correlated at 0.347. Slope is ordinal and thus heart disease correlates positively with down sloping of the peak exercise ST.

Anticorrelations with heart disease include oldpeak (ST depression induced by exercise relative to rest) at -0.435, exang(if angina was induced by exercise or not) at -0.434, and ca(number of fluoresced vessels) at -0.385.

Inter-feature correlations of significance are slope and maximum heart rate (0.385), resting blood pressure and age (0.280), and exang vs oldpeak (0.278). The largest anticorrelations are slope downwardness and oldpeak (-0.579), maximum heart rate and age (-0.363), and maximum heart rate and exang (-0.361).



	member	age	sex	trestbps	chol	fbs	thalach	exang	oldpeak	slope	ca	target
member	1	0.186	0.201	0.110	0.019	-0.018	-0.410	0.359	0.301	-0.278	0.374	-0.863
age	0.186	1	-0.096	0.280	0.217	0.118	-0.363	0.106	0.213	-0.171	0.276	-0.225
sex	0.201	-0.096	1	-0.055	-0.195	0.044	-0.048	0.141	0.094	-0.039	0.120	-0.284
trestbps	0.110	0.280	-0.055	1	0.121	0.177	-0.038	0.071	0.193	-0.121	0.100	-0.146
chol	0.019	0.217	-0.195	0.121	1	0.020	0.014	0.071	0.059	0.000	0.069	-0.080
fbs	-0.018	0.118	0.044	0.177	0.020	1	-0.015	0.024	0.005	-0.060	0.139	-0.019
thalach	-0.410	-0.363	-0.048	-0.038	0.014	-0.015	1	-0.361	-0.343	0.385	-0.168	0.414
exang	0.359	0.106	0.141	0.071	0.071	0.024	-0.361	1	0.279	-0.252	0.118	-0.434
oldpeak	0.301	0.213	0.094	0.193	0.059	0.005	-0.343	0.279	1	-0.579	0.231	-0.435
slope	-0.278	-0.171	-0.039	-0.121	0.000	-0.060	0.385	-0.252	-0.579	1	-0.089	0.347
ca	0.374	0.276	0.120	0.100	0.069	0.139	-0.168	0.118	0.231	-0.089	1	-0.385
target	-0.863	-0.225	-0.284	-0.146	-0.080	-0.019	0.414	-0.434	-0.435	0.347	-0.385	1