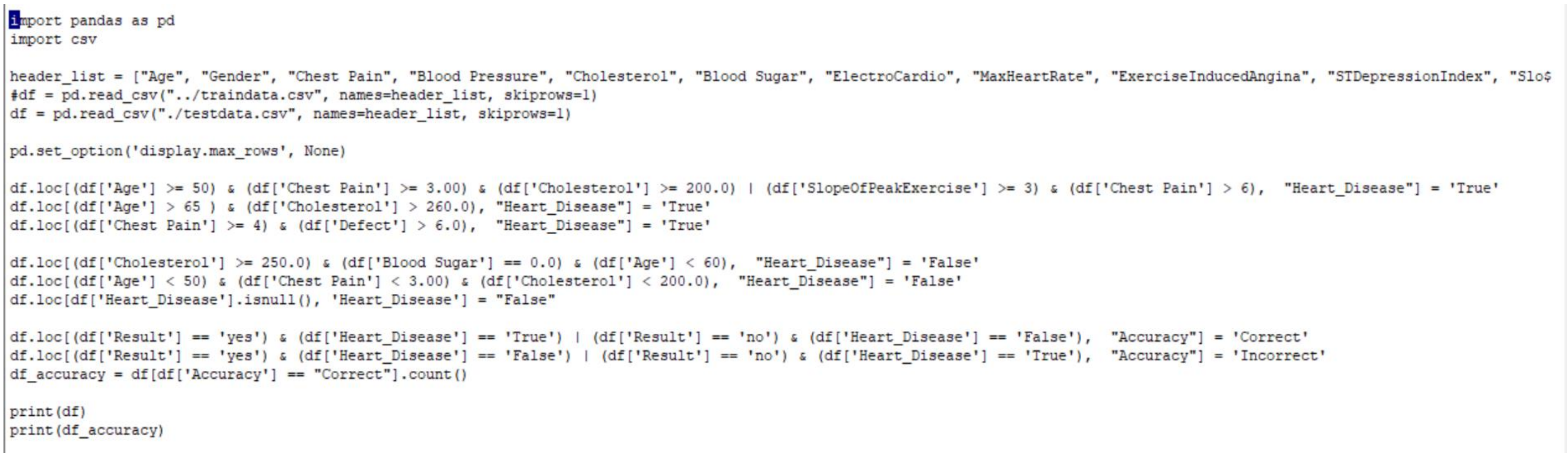
Shane Irons

Intro to Data Science

Assignment 2

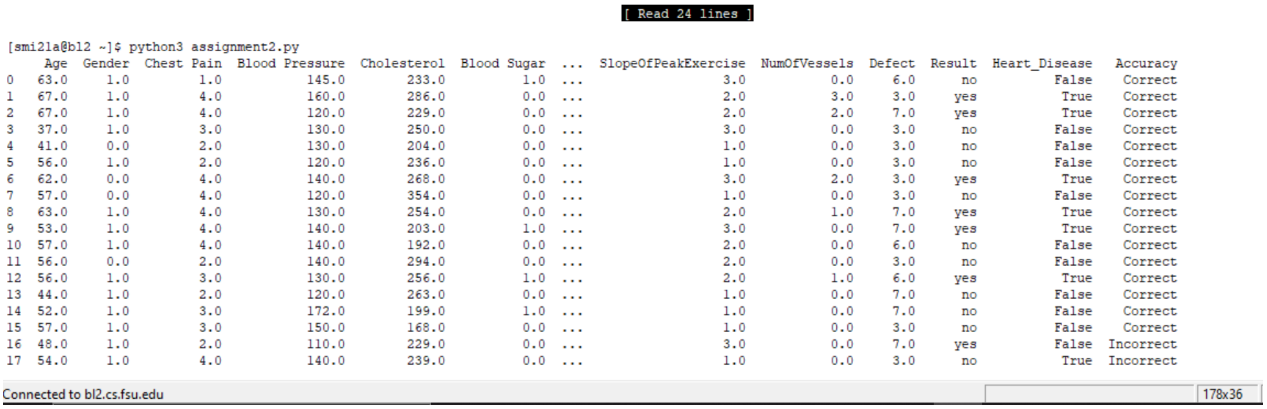
Dr. Gubanov

1. Below is my rule to label the data rows and determine if a patient may have heart disease or not.



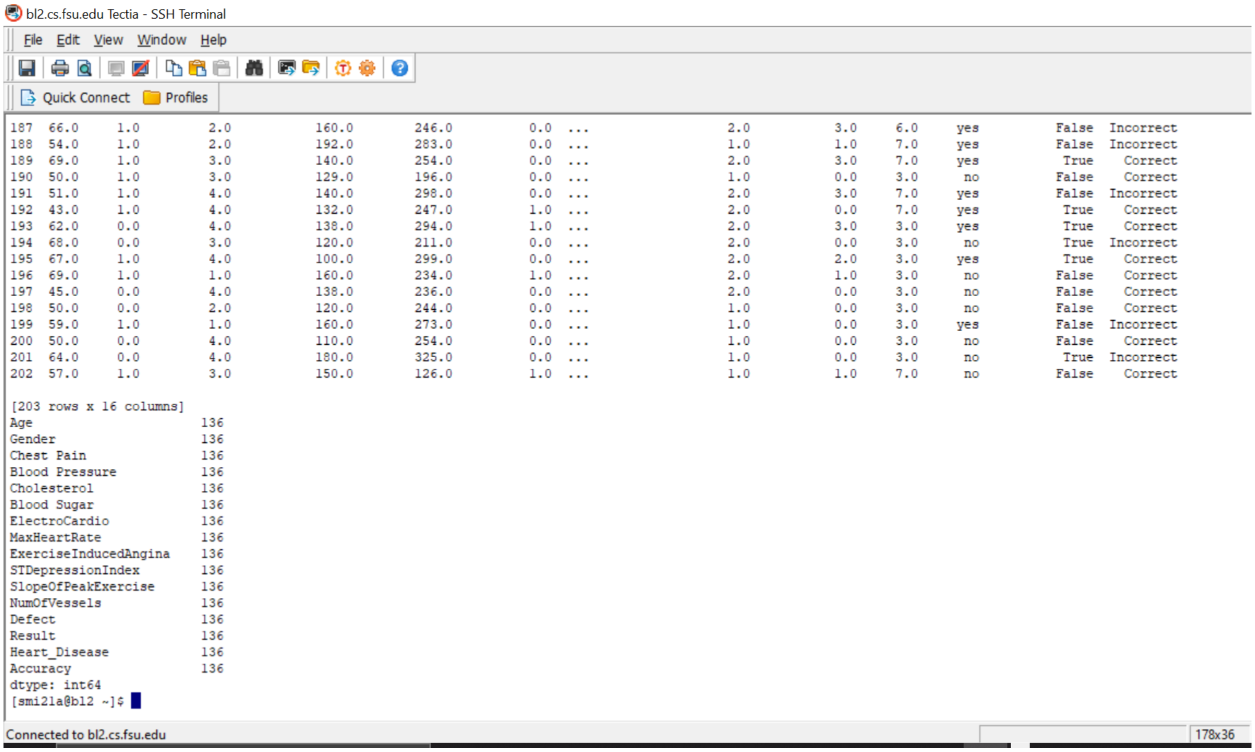
```
import pandas as pd
import csv

header_list = ["Age", "Gender", "Chest Pain", "Blood Pressure", "Cholesterol", "Blood Sugar", "ElectroCardio", "MaxHeartRate", "ExerciseInducedAngina", "STDepressionIndex", "Slo$
#df = pd.read_csv("../traindata.csv", names=header_list, skiprows=1)
df = pd.read_csv("./testdata.csv", names=header_list, skiprows=1)

pd.set_option('display.max_rows', None)

df.loc[(df['Age'] >= 50) & (df['Chest Pain'] >= 3.00) & (df['Cholesterol'] >= 200.0) | (df['SlopeOfPeakExercise'] >= 3) & (df['Chest Pain'] > 6),  "Heart_Disease"] = 'True'
df.loc[(df['Age'] > 65 ) & (df['Cholesterol'] > 260.0), "Heart_Disease"] = 'True'
df.loc[(df['Chest Pain'] >= 4) & (df['Defect'] > 6.0),  "Heart_Disease"] = 'True'

df.loc[(df['Cholesterol'] >= 250.0) & (df['Blood Sugar'] == 0.0) & (df['Age'] < 60),  "Heart_Disease"] = 'False'
df.loc[(df['Age'] < 50) & (df['Chest Pain'] < 3.00) & (df['Cholesterol'] < 200.0),  "Heart_Disease"] = 'False'
df.loc[df['Heart_Disease'].isnull(), 'Heart_Disease'] = "False"

df.loc[(df['Result'] == 'yes') & (df['Heart_Disease'] == 'True') | (df['Result'] == 'no') & (df['Heart_Disease'] == 'False'),  "Accuracy"] = 'Correct'
df.loc[(df['Result'] == 'yes') & (df['Heart_Disease'] == 'False') | (df['Result'] == 'no') & (df['Heart_Disease'] == 'True'),  "Accuracy"] = 'Incorrect'
df_accuracy = df[df['Accuracy'] == "Correct"].count()

print(df)
print(df_accuracy)
```

2. **Explanation of Rule**: above is my designed rule for the dataset. I import pandas and create a header list that copies the headers from the CSV file. Then I provide the file that is to be read and I skip the first row (the first row in the csv file is the list of headers). Following this the ruleset begins and works as follows: df.loc (locate in the dataset) where age >= 50 & chest pain >= 3.00 & Cholesterol >= 200.0 OR where Slope of Peak Exercise >= 3 and if Chest Pain > 6 then to determine heart disease as True. Also, if Age > 65 and if Cholesterol is > 260 then to determine heart disease as true. One more True case if Chest Pain >= 4 and if the Defect is > 6.0. Opposite of this, if Cholesterol is >= 250 and if blood sugar = 0 and if the age is < 60, then determine heart disease as False. Also, if age is < 50 & chest pain < 3.00 & cholesterol < 200.0 then determine heart disease as False. Then for all the NaN cases, default them to False. Under this, I have another column called Accuracy that tests for the correlation between 'Result' and my created 'Heart_Disease' columns. If Result = Heart_Disease then Correct and if not then Incorrect. Running this python program shows the following output:



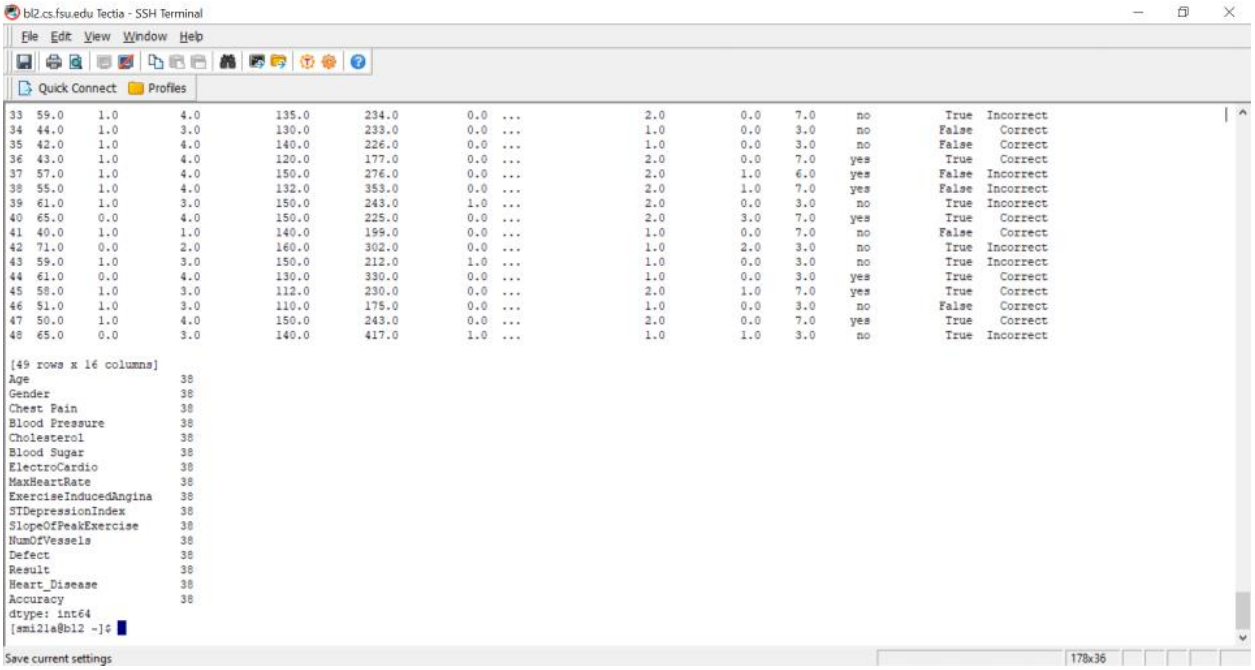
```
                                                                                 [ Read 24 lines ]

[smi21a@bl2 ~]$ python3 assignment2.py
     Age  Gender  Chest Pain  Blood Pressure  Cholesterol  Blood Sugar  ...  SlopeOfPeakExercise  NumOfVessels  Defect  Result  Heart_Disease   Accuracy
0   63.0     1.0         1.0           145.0        233.0          1.0  ...                  3.0           0.0     6.0      no          False    Correct
1   67.0     1.0         4.0           160.0        286.0          0.0  ...                  2.0           3.0     3.0     yes           True    Correct
2   67.0     1.0         4.0           120.0        229.0          0.0  ...                  2.0           2.0     7.0     yes           True    Correct
3   37.0     1.0         3.0           130.0        250.0          0.0  ...                  3.0           0.0     3.0      no          False    Correct
4   41.0     0.0         2.0           130.0        204.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
5   56.0     1.0         2.0           120.0        236.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
6   62.0     0.0         4.0           140.0        268.0          0.0  ...                  3.0           2.0     3.0     yes           True    Correct
7   57.0     0.0         4.0           120.0        354.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
8   63.0     1.0         4.0           130.0        254.0          0.0  ...                  2.0           1.0     7.0     yes           True    Correct
9   53.0     1.0         4.0           140.0        203.0          1.0  ...                  3.0           0.0     7.0     yes           True    Correct
10  57.0     1.0         4.0           140.0        192.0          0.0  ...                  2.0           0.0     6.0      no          False    Correct
11  56.0     0.0         2.0           140.0        294.0          0.0  ...                  2.0           0.0     3.0      no          False    Correct
12  56.0     1.0         3.0           130.0        256.0          1.0  ...                  2.0           1.0     6.0     yes           True    Correct
13  44.0     1.0         2.0           120.0        263.0          0.0  ...                  1.0           0.0     7.0      no          False    Correct
14  52.0     1.0         3.0           172.0        199.0          1.0  ...                  1.0           0.0     7.0      no          False    Correct
15  57.0     1.0         3.0           150.0        168.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
16  48.0     1.0         2.0           110.0        229.0          0.0  ...                  3.0           0.0     7.0     yes          False  Incorrect
17  54.0     1.0         4.0           140.0        239.0          0.0  ...                  1.0           0.0     3.0      no           True  Incorrect
```

3. Accuracy number on the main file is 136/202 correctly determined heart disease rate. This is roughly a 67% accuracy for the first iteration of this rule. Running this on the main file was just to test



bl2.cs.fsu.edu Tectia - SSH Terminal
File Edit View Window Help
Quick Connect
Profiles

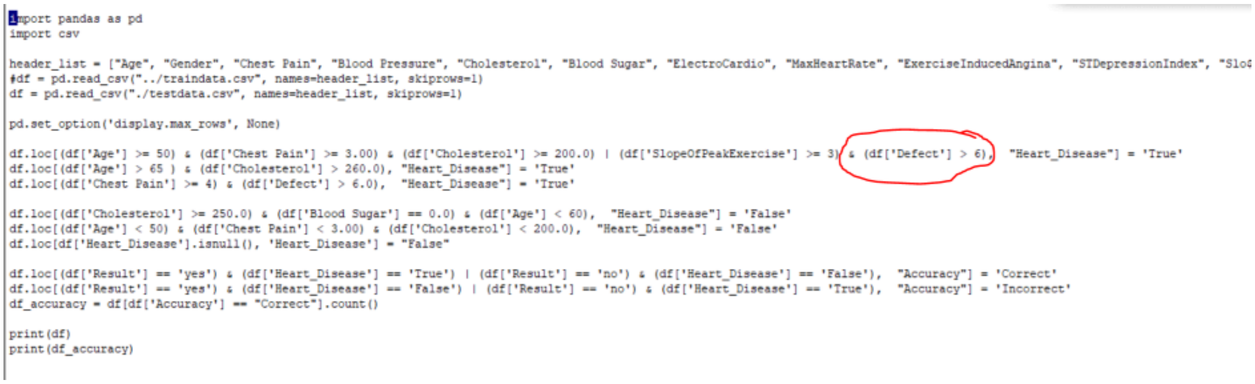
```
187  66.0     1.0         2.0           160.0        246.0          0.0  ...                  2.0           3.0     6.0     yes          False  Incorrect
188  54.0     1.0         2.0           192.0        283.0          0.0  ...                  1.0           1.0     7.0     yes          False  Incorrect
189  69.0     1.0         3.0           140.0        254.0          0.0  ...                  2.0           3.0     7.0     yes           True    Correct
190  50.0     1.0         3.0           129.0        196.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
191  51.0     1.0         4.0           140.0        298.0          0.0  ...                  2.0           3.0     7.0     yes          False  Incorrect
192  43.0     1.0         4.0           132.0        247.0          1.0  ...                  2.0           0.0     7.0     yes           True    Correct
193  62.0     0.0         4.0           138.0        294.0          1.0  ...                  2.0           3.0     3.0     yes           True    Correct
194  68.0     0.0         3.0           120.0        211.0          0.0  ...                  2.0           0.0     3.0      no           True  Incorrect
195  67.0     1.0         4.0           100.0        299.0          0.0  ...                  2.0           2.0     3.0     yes           True    Correct
196  69.0     1.0         1.0           160.0        234.0          1.0  ...                  2.0           1.0     3.0      no          False    Correct
197  45.0     0.0         4.0           138.0        236.0          0.0  ...                  2.0           0.0     3.0      no          False    Correct
198  50.0     0.0         2.0           120.0        244.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
199  59.0     1.0         1.0           160.0        273.0          0.0  ...                  1.0           0.0     3.0     yes          False  Incorrect
200  50.0     0.0         4.0           110.0        254.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
201  64.0     0.0         4.0           180.0        325.0          0.0  ...                  1.0           0.0     3.0      no           True  Incorrect
202  57.0     1.0         3.0           150.0        126.0          1.0  ...                  1.0           1.0     7.0      no          False    Correct

[203 rows x 16 columns]
Age                      136
Gender                   136
Chest Pain               136
Blood Pressure           136
Cholesterol              136
Blood Sugar              136
ElectroCardio            136
MaxHeartRate             136
ExerciseInducedAngina    136
STDepressionIndex        136
SlopeOfPeakExercise      136
NumOfVessels             136
Defect                   136
Result                   136
Heart_Disease            136
Accuracy                 136
dtype: int64
[smi21a@bl2 ~]$
```


bl2.cs.fsu.edu Tectia - SSH Terminal
File Edit View Window Help
Quick Connect
Profiles
Save current settings
178x36

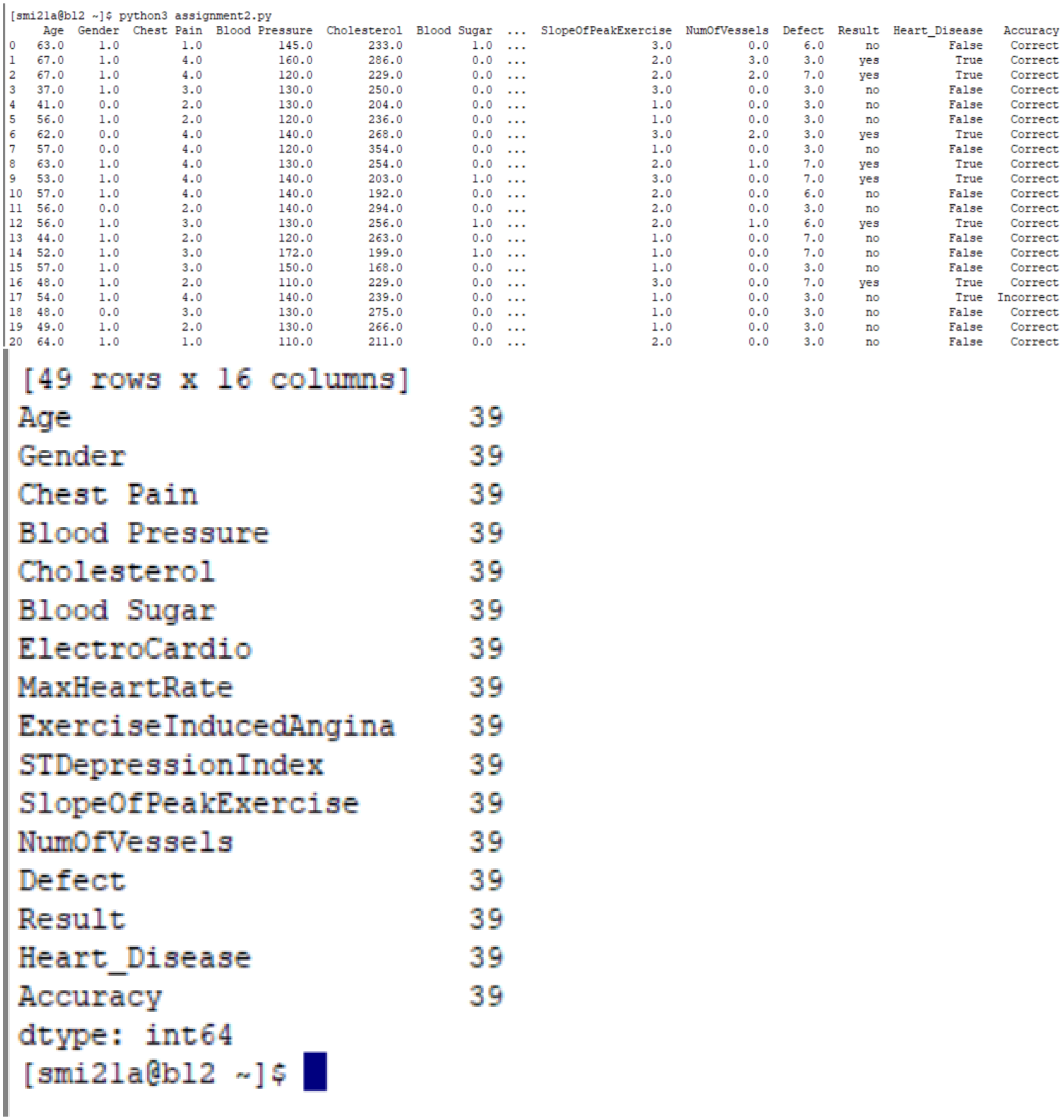
```
33  59.0     1.0         4.0         135.0       234.0       0.0  ...             2.0          0.0   7.0    no          True  Incorrect
34  44.0     1.0         3.0         130.0       233.0       0.0  ...             1.0          0.0   3.0    no         False    Correct
35  42.0     1.0         4.0         140.0       226.0       0.0  ...             1.0          0.0   3.0    no         False    Correct
36  43.0     1.0         4.0         120.0       177.0       0.0  ...             2.0          0.0   7.0   yes          True    Correct
37  57.0     1.0         4.0         150.0       276.0       0.0  ...             2.0          1.0   6.0   yes         False  Incorrect
38  55.0     1.0         4.0         132.0       353.0       0.0  ...             2.0          1.0   7.0   yes         False  Incorrect
39  61.0     1.0         3.0         150.0       243.0       1.0  ...             2.0          0.0   3.0    no          True  Incorrect
40  65.0     0.0         4.0         150.0       225.0       0.0  ...             2.0          3.0   7.0   yes          True    Correct
41  40.0     1.0         1.0         140.0       199.0       0.0  ...             1.0          0.0   7.0    no         False    Correct
42  71.0     0.0         2.0         160.0       302.0       0.0  ...             1.0          2.0   3.0    no          True  Incorrect
43  59.0     1.0         3.0         150.0       212.0       1.0  ...             1.0          0.0   3.0    no          True  Incorrect
44  61.0     0.0         4.0         130.0       330.0       0.0  ...             1.0          0.0   3.0   yes          True    Correct
45  58.0     1.0         3.0         112.0       230.0       0.0  ...             2.0          1.0   7.0   yes          True    Correct
46  51.0     1.0         3.0         110.0       175.0       0.0  ...             1.0          0.0   3.0    no         False    Correct
47  50.0     1.0         4.0         150.0       243.0       0.0  ...             2.0          0.0   7.0   yes          True    Correct
48  65.0     0.0         3.0         140.0       417.0       1.0  ...             1.0          1.0   3.0    no          True  Incorrect

[49 rows x 16 columns]
Age                      38
Gender                   38
Chest Pain               38
Blood Pressure           38
Cholesterol              38
Blood Sugar              38
ElectroCardio            38
MaxHeartRate             38
ExerciseInducedAngina    38
STDepressionIndex        38
SlopeOfPeakExercise      38
NumOfVessels             38
Defect                   38
Result                   38
Heart_Disease            38
Accuracy                 38
dtype: int64
[smi21a@bl2 ~]$ 
```

When running the program on the testdata.csv file for question 3, which is a csv file only containing the first 50 rows of data from the original traindata.csv file, the accuracy jumps up to 38/49 or 77.5%. This is likely because towards the end of the main file, there are a lot of cases that slip by the rule.

Edit: I realized that I had an error in my code regarding the first rule. Previously, I had put "… | (df['SlopeOfPeakExercise'] >= 3) & **(df['Chest Pain'] > 6),** "Heart_Disease"] = 'True'" where chest pain here should be 'Defect'. I changed this and ran the program again to receive better accuracy than before.



```
import pandas as pd
import csv

header_list = ["Age", "Gender", "Chest Pain", "Blood Pressure", "Cholesterol", "Blood Sugar", "ElectroCardio", "MaxHeartRate", "ExerciseInducedAngina", "STDepressionIndex", "Slo
#df = pd.read_csv("../traindata.csv", names=header_list, skiprows=1)
df = pd.read_csv("./testdata.csv", names=header_list, skiprows=1)

pd.set_option('display.max_rows', None)

df.loc[(df['Age'] >= 50) & (df['Chest Pain'] >= 3.00) & (df['Cholesterol'] >= 200.0) | (df['SlopeOfPeakExercise'] >= 3) & (df['Defect'] > 6),  "Heart_Disease"] = 'True'
df.loc[(df['Age'] > 65 ) & (df['Cholesterol'] > 260.0), "Heart_Disease"] = 'True'
df.loc[(df['Chest Pain'] >= 4) & (df['Defect'] > 6.0),  "Heart_Disease"] = 'True'

df.loc[(df['Cholesterol'] >= 250.0) & (df['Blood Sugar'] == 0.0) & (df['Age'] < 60),  "Heart_Disease"] = 'False'
df.loc[(df['Age'] < 50) & (df['Chest Pain'] < 3.00) & (df['Cholesterol'] < 200.0),  "Heart_Disease"] = 'False'
df.loc[df['Heart_Disease'].isnull(), 'Heart_Disease'] = "False"

df.loc[(df['Result'] == 'yes') & (df['Heart_Disease'] == 'True') | (df['Result'] == 'no') & (df['Heart_Disease'] == 'False'),  "Accuracy"] = 'Correct'
df.loc[(df['Result'] == 'yes') & (df['Heart_Disease'] == 'False') | (df['Result'] == 'no') & (df['Heart_Disease'] == 'True'),  "Accuracy"] = 'Incorrect'
df_accuracy = df[df['Accuracy'] == "Correct"].count()

print(df)
print(df_accuracy)
```



```
[smi21a@bl2 ~]$ python3 assignment2.py
      Age  Gender  Chest Pain  Blood Pressure  Cholesterol  Blood Sugar  ...  SlopeOfPeakExercise  NumOfVessels  Defect  Result  Heart_Disease   Accuracy
0   63.0     1.0         1.0           145.0        233.0          1.0  ...                  3.0           0.0     6.0      no          False    Correct
1   67.0     1.0         4.0           160.0        286.0          0.0  ...                  2.0           3.0     3.0     yes           True    Correct
2   67.0     1.0         4.0           120.0        229.0          0.0  ...                  2.0           2.0     7.0     yes           True    Correct
3   37.0     1.0         3.0           130.0        250.0          0.0  ...                  3.0           0.0     3.0      no          False    Correct
4   41.0     0.0         2.0           130.0        204.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
5   56.0     1.0         2.0           120.0        236.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
6   62.0     0.0         4.0           140.0        268.0          0.0  ...                  3.0           2.0     3.0     yes           True    Correct
7   57.0     0.0         4.0           120.0        354.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
8   63.0     1.0         4.0           130.0        254.0          0.0  ...                  2.0           1.0     7.0     yes           True    Correct
9   53.0     1.0         4.0           140.0        203.0          1.0  ...                  3.0           0.0     7.0     yes           True    Correct
10  57.0     1.0         4.0           140.0        192.0          0.0  ...                  2.0           0.0     6.0      no          False    Correct
11  56.0     0.0         2.0           140.0        294.0          0.0  ...                  2.0           0.0     3.0      no          False    Correct
12  56.0     1.0         3.0           130.0        256.0          1.0  ...                  2.0           1.0     6.0     yes           True    Correct
13  44.0     1.0         2.0           120.0        263.0          0.0  ...                  1.0           0.0     7.0      no          False    Correct
14  52.0     1.0         3.0           172.0        199.0          1.0  ...                  1.0           0.0     7.0      no          False    Correct
15  57.0     1.0         3.0           150.0        168.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
16  48.0     1.0         2.0           110.0        229.0          0.0  ...                  3.0           0.0     7.0     yes           True    Correct
17  54.0     1.0         4.0           140.0        239.0          0.0  ...                  1.0           0.0     3.0      no           True  Incorrect
18  48.0     0.0         3.0           130.0        275.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
19  49.0     1.0         2.0           130.0        266.0          0.0  ...                  1.0           0.0     3.0      no          False    Correct
20  64.0     1.0         1.0           110.0        211.0          0.0  ...                  2.0           0.0     3.0      no          False    Correct

[49 rows x 16 columns]
Age                      39
Gender                   39
Chest Pain               39
Blood Pressure           39
Cholesterol              39
Blood Sugar              39
ElectroCardio            39
MaxHeartRate             39
ExerciseInducedAngina    39
STDepressionIndex        39
SlopeOfPeakExercise      39
NumOfVessels             39
Defect                   39
Result                   39
Heart_Disease            39
Accuracy                 39
dtype: int64
[smi21a@bl2 ~]$
```

This was run using the testdata.csv file and produced 39/49 correct predictions (accuracy of 79.5%). This is the highest accuracy I was able to achieve.