Basic Data Processing and Visualization

August 23, 2020

0.0.1 Dataset Description

Our dataset is from a study of heart disease that has been open to the public for many years. The study collects various measurements on patient health and cardiovascular statistics, and of course makes patient identities anonymous.

There are 14 columns in the dataset, where the patient_id column is a unique and random identifier. The remaining 13 features are described in the section below.

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

0.0.2 Attributes types

Real: 1,4,5,8,10,12

Ordered:11,

Binary: 2,6,9

Nominal:7,3,13

Data is provided courtesy of the Cleveland Heart Disease Database via the UCI Machine Learning repository.

Aha, D., and Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database." University of California 3.1 (1988): 3-2.

0.0.3 Data preparation and/or cleaning was required

Data is loaded and checked for any missing values. Data Exploration and Graphs are drawn out to see data distributions and basic statistics are done.

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import datetime

%matplotlib inline
sns.set_style('dark')
sns.set(font_scale=1.2)

import warnings
warnings.filterwarnings('ignore')

pd.set_option('display.max_columns',None)
#pd.set_option('display.max_rows',None)
```

[2]:
```
df = pd.read_csv("heart.csv")
```

[3]:
```
df
```

[3]:

	age	sex	chestpain	bloodpressure	serum	bloodsugar	\
0	70	1	4	130	322	0	
1	67	0	3	115	564	0	
2	57	1	2	124	261	0	
3	64	1	4	128	263	0	
4	74	0	2	120	269	0	
..	…	…	…	…	…	…	
265	52	1	3	172	199	1	
266	44	1	2	120	263	0	
267	56	0	2	140	294	0	
268	57	1	4	140	192	0	
269	67	1	4	160	286	0	

	electrocardiographic	heartrate	angina	depression	slope	vessels	\
0	2	109	0	2.4	2	3	

```
1                    2        160       0         1.6      2        0
2                    0        141       0         0.3      1        0
3                    0        105       1         0.2      2        1
4                    2        121       1         0.2      1        1
..                 …        …      …         …     …        …
265                  0        162       0         0.5      1        0
266                  0        173       0         0.0      1        0
267                  2        153       0         1.3      2        0
268                  0        148       0         0.4      2        0
269                  2        108       1         1.5      2        3

     thal  disease
0       3        1
1       7        0
2       7        1
3       7        0
4       3        0
..    …       …
265     7        0
266     7        0
267     3        0
268     6        0
269     3        1

[270 rows x 14 columns]
```

0.0.4 Exploratory Data Analysis

[4]:
```
df.info() #No missing values
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 270 entries, 0 to 269
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   age                   270 non-null    int64
 1   sex                   270 non-null    int64
 2   chestpain             270 non-null    int64
 3   bloodpressure         270 non-null    int64
 4   serum                 270 non-null    int64
 5   bloodsugar            270 non-null    int64
 6   electrocardiographic  270 non-null    int64
 7   heartrate             270 non-null    int64
 8   angina                270 non-null    int64
 9   depression            270 non-null    float64
 10  slope                 270 non-null    int64
 11  vessels               270 non-null    int64
```

```
 12  thal                 270 non-null    int64
 13  disease              270 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 29.7 KB
```

[5]:
```
df.describe() #Statistics
```

[5]:

	age	sex	chestpain	bloodpressure	serum
count	270.000000	270.000000	270.000000	270.000000	270.000000
mean	54.433333	0.677778	3.174074	131.344444	249.659259
std	9.109067	0.468195	0.950090	17.861608	51.686237
min	29.000000	0.000000	1.000000	94.000000	126.000000
25%	48.000000	0.000000	3.000000	120.000000	213.000000
50%	55.000000	1.000000	3.000000	130.000000	245.000000
75%	61.000000	1.000000	4.000000	140.000000	280.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000

	bloodsugar	electrocardiographic	heartrate	angina	depression
count	270.000000	270.000000	270.000000	270.000000	270.00000
mean	0.148148	1.022222	149.677778	0.329630	1.05000
std	0.355906	0.997891	23.165717	0.470952	1.14521
min	0.000000	0.000000	71.000000	0.000000	0.00000
25%	0.000000	0.000000	133.000000	0.000000	0.00000
50%	0.000000	2.000000	153.500000	0.000000	0.80000
75%	0.000000	2.000000	166.000000	1.000000	1.60000
max	1.000000	2.000000	202.000000	1.000000	6.20000

	slope	vessels	thal	disease
count	270.000000	270.000000	270.000000	270.000000
mean	1.585185	0.670370	4.696296	0.444444
std	0.614390	0.943896	1.940659	0.497827
min	1.000000	0.000000	3.000000	0.000000
25%	1.000000	0.000000	3.000000	0.000000
50%	2.000000	0.000000	3.000000	0.000000
75%	2.000000	1.000000	7.000000	1.000000
max	3.000000	3.000000	7.000000	1.000000

[6]:
```
df.columns
```

[6]:
```
Index(['age', 'sex', 'chestpain', 'bloodpressure', 'serum', 'bloodsugar',
       'electrocardiographic', 'heartrate', 'angina', 'depression', 'slope',
       'vessels', 'thal', 'disease'],
      dtype='object')
```

0.0.5 Data Visualization

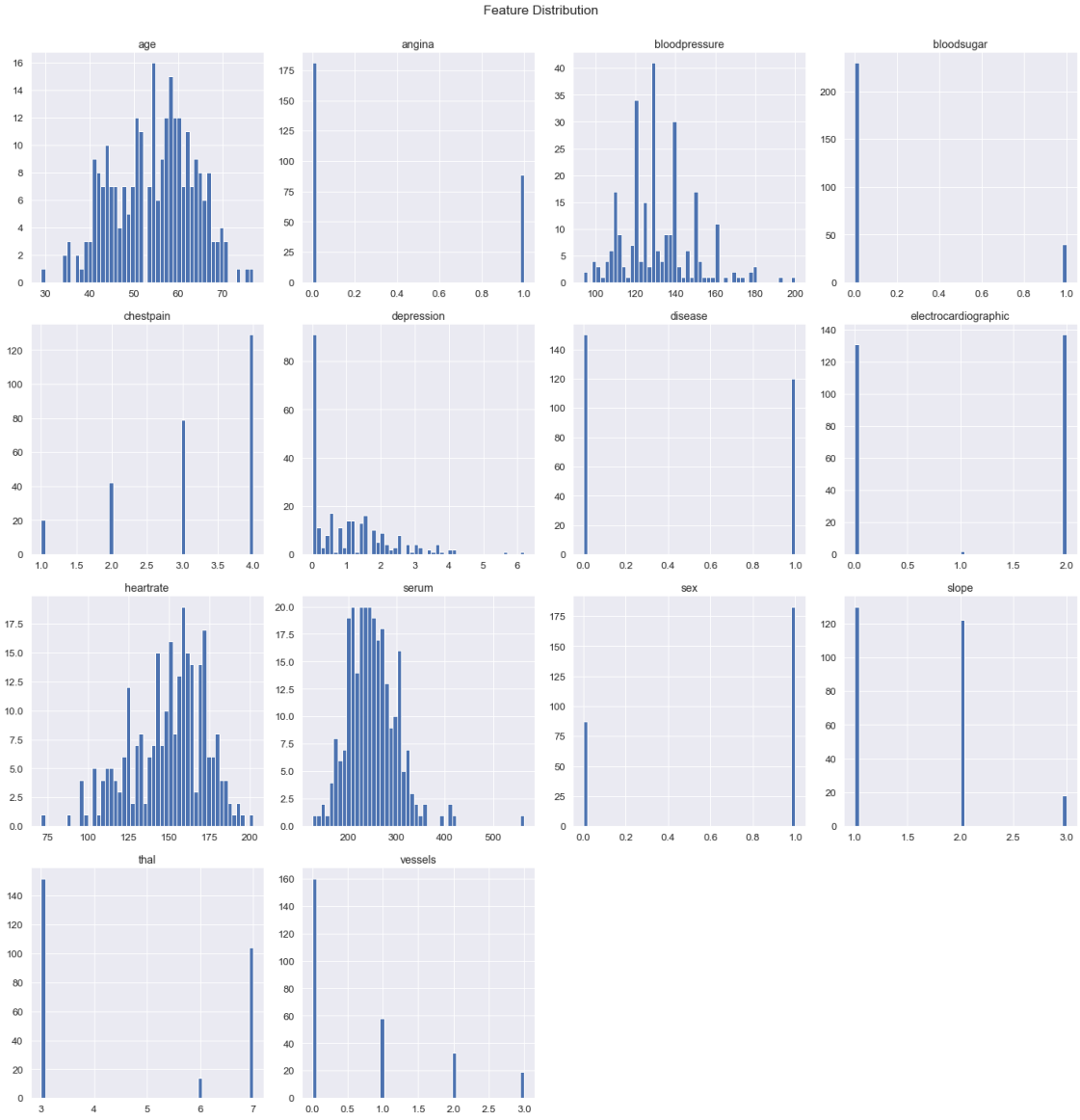
0.0.6 Univariate Data Exploration

[7]:
```
df.hist(bins=50, figsize=(20,20))

plt.suptitle('Feature Distribution', x=0.5, y=1.02, ha='center',␣
 ↪fontsize='large')

plt.tight_layout()

plt.show();
```



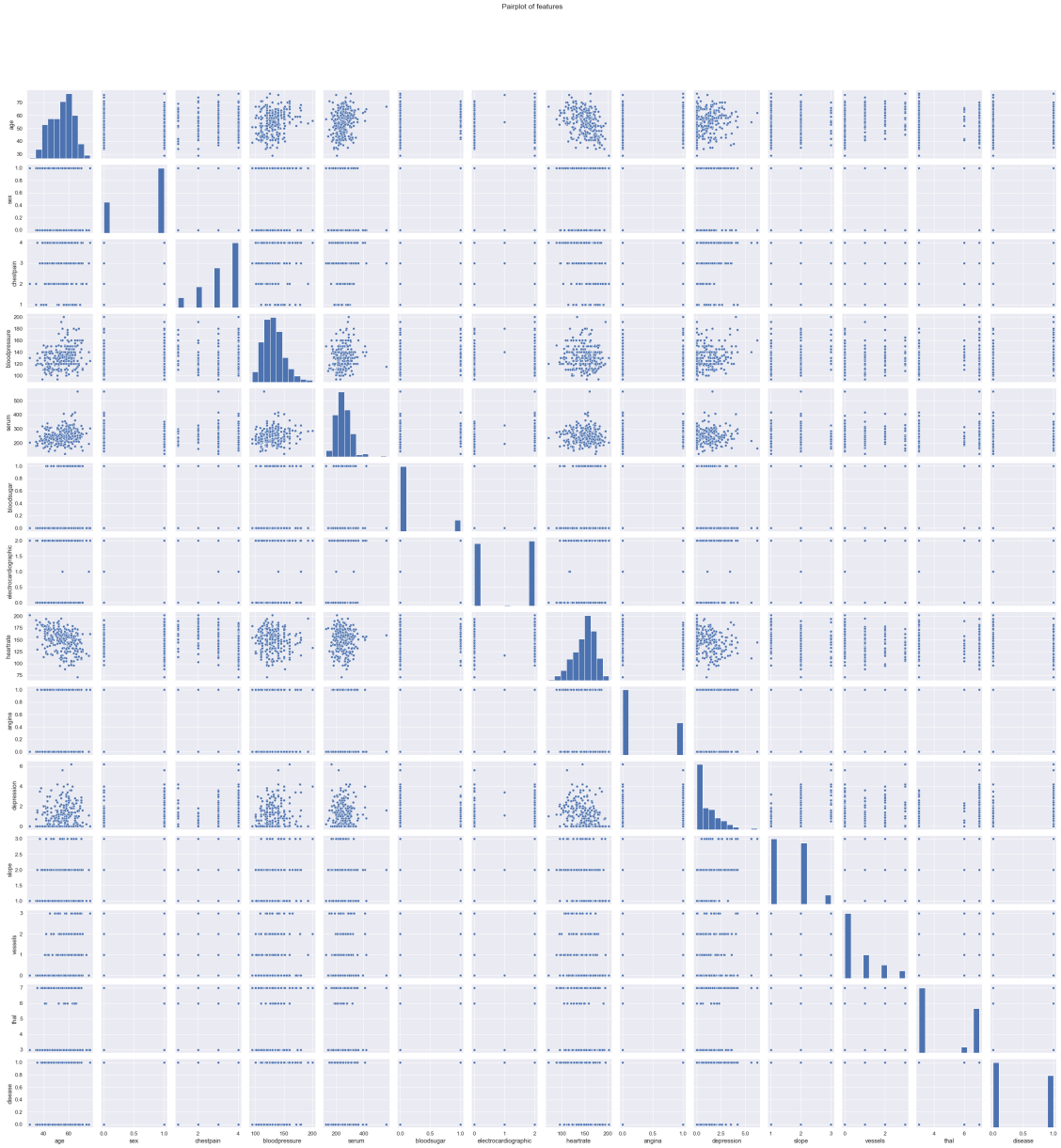
[8]:
```
plt.figure(figsize=(20,20))

g = sns.pairplot(df)

g.fig.suptitle("Pairplot of features", y=1.08, ha='center', fontsize='large')

plt.show()
```

```
<Figure size 1440x1440 with 0 Axes>
```



0.0.7 Correlation

[9]:
```
df.corr()
```

[9]:

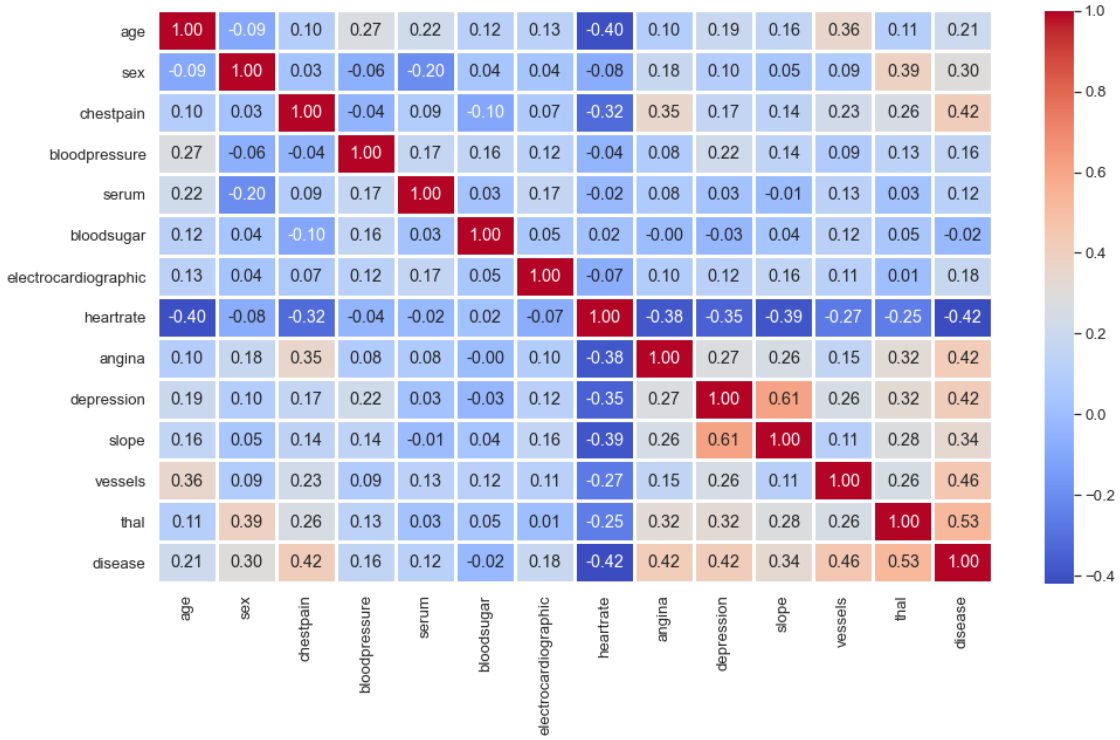
	age	sex	chestpain	bloodpressure	serum
age	1.000000	-0.094401	0.096920	0.273053	0.220056
sex	-0.094401	1.000000	0.034636	-0.062693	-0.201647
chestpain	0.096920	0.034636	1.000000	-0.043196	0.090465
bloodpressure	0.273053	-0.062693	-0.043196	1.000000	0.173019
serum	0.220056	-0.201647	0.090465	0.173019	1.000000
bloodsugar	0.123458	0.042140	-0.098537	0.155681	0.025186
electrocardiographic	0.128171	0.039253	0.074325	0.116157	0.167652
heartrate	-0.402215	-0.076101	-0.317682	-0.039136	-0.018739
angina	0.098297	0.180022	0.353160	0.082793	0.078243
depression	0.194234	0.097412	0.167244	0.222800	0.027709
slope	0.159774	0.050545	0.136900	0.142472	-0.005755
vessels	0.356081	0.086830	0.225890	0.085697	0.126541
thal	0.106100	0.391046	0.262659	0.132045	0.028836
disease	0.212322	0.297721	0.417436	0.155383	0.118021

	bloodsugar	electrocardiographic	heartrate	angina
age	0.123458	0.128171	-0.402215	0.098297
sex	0.042140	0.039253	-0.076101	0.180022
chestpain	-0.098537	0.074325	-0.317682	0.353160
bloodpressure	0.155681	0.116157	-0.039136	0.082793
serum	0.025186	0.167652	-0.018739	0.078243
bloodsugar	1.000000	0.053499	0.022494	-0.004107
electrocardiographic	0.053499	1.000000	-0.074628	0.095098
heartrate	0.022494	-0.074628	1.000000	-0.380719
angina	-0.004107	0.095098	-0.380719	1.000000
depression	-0.025538	0.120034	-0.349045	0.274672
slope	0.044076	0.160614	-0.386847	0.255908
vessels	0.123774	0.114368	-0.265333	0.153347
thal	0.049237	0.007337	-0.253397	0.321449
disease	-0.016319	0.182091	-0.418514	0.419303

	depression	slope	vessels	thal	disease
age	0.194234	0.159774	0.356081	0.106100	0.212322
sex	0.097412	0.050545	0.086830	0.391046	0.297721
chestpain	0.167244	0.136900	0.225890	0.262659	0.417436
bloodpressure	0.222800	0.142472	0.085697	0.132045	0.155383
serum	0.027709	-0.005755	0.126541	0.028836	0.118021
bloodsugar	-0.025538	0.044076	0.123774	0.049237	-0.016319
electrocardiographic	0.120034	0.160614	0.114368	0.007337	0.182091
heartrate	-0.349045	-0.386847	-0.265333	-0.253397	-0.418514
angina	0.274672	0.255908	0.153347	0.321449	0.419303
depression	1.000000	0.609712	0.255005	0.324333	0.417967

slope	0.609712	1.000000	0.109498	0.283678	0.337616
vessels	0.255005	0.109498	1.000000	0.255648	0.455336
thal	0.324333	0.283678	0.255648	1.000000	0.525020
disease	0.417967	0.337616	0.455336	0.525020	1.000000

[10]:
```
plt.figure(figsize=(16,9))
sns.heatmap(df.corr(),cmap="coolwarm",annot=True,fmt='.2f',linewidths=2)
plt.show()
```



There are no strong correlations between features.

[ ]: