Medical Data Visualizer

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

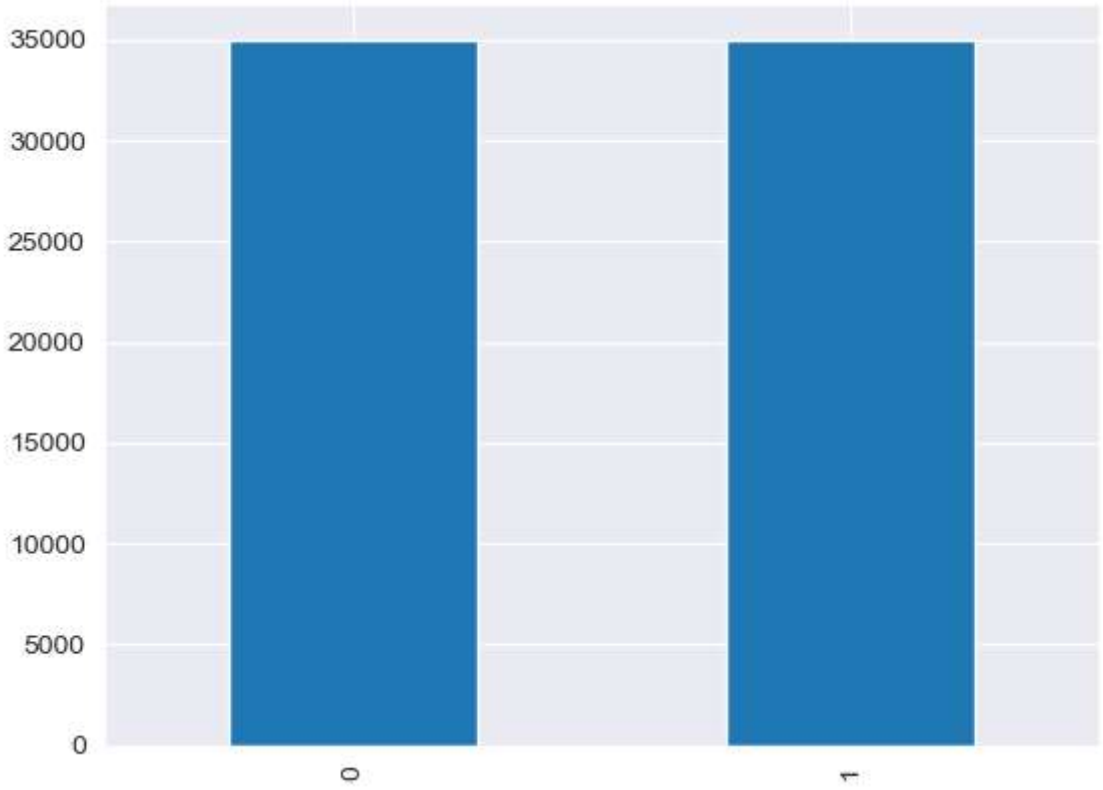
In [2]:
```
df = pd.read_csv("medical_examination.csv")
df.head()
```

Out[2]:

	id	age	sex	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

In [ ]:
```
df.cardio.value_counts().plot.bar();
```

```
<Figure size 640x480 with 1 Axes>
```

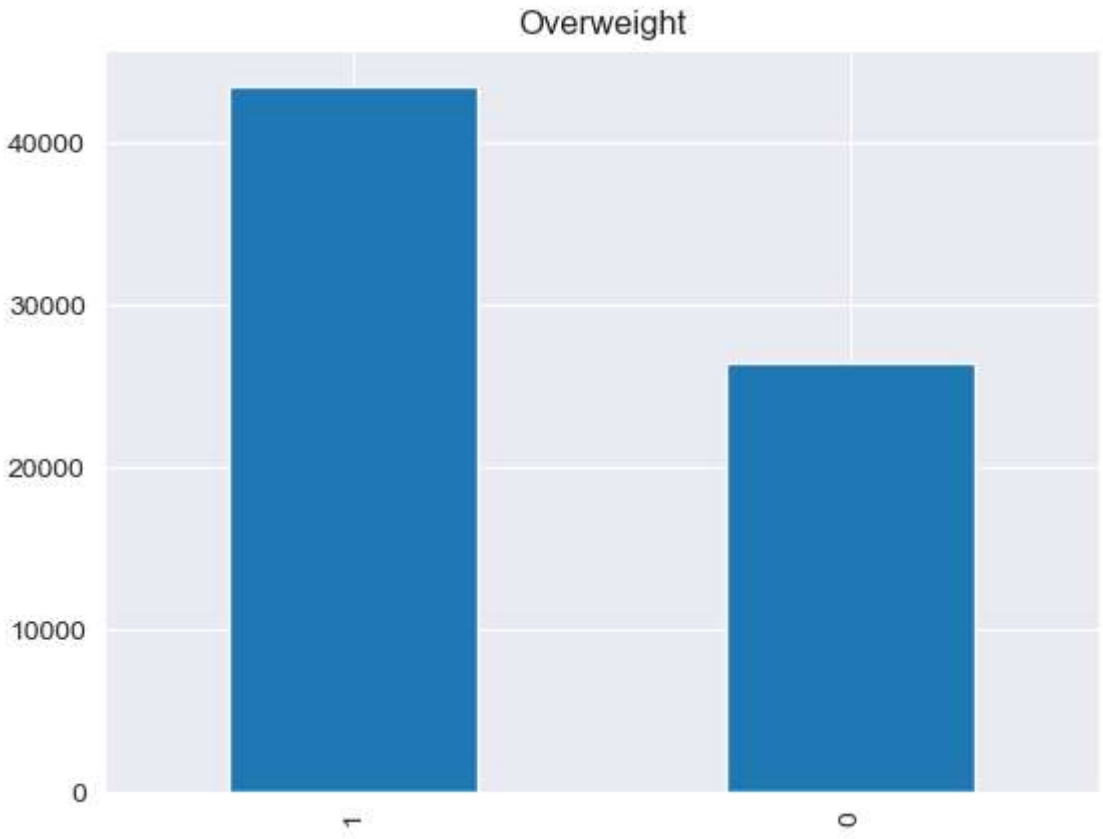


In [ ]:
```
df["overweight"]=((df.weight/np.square(df.height*0.01))>25).astype(int)
```

This chart describes overweight percentage in the data we have

In [ ]:
```
df.overweight.value_counts().plot(kind="bar",title="Overweight");
```

```
<Figure size 640x480 with 1 Axes>
```



In [ ]:
```
df.cholesterol=(df.cholesterol>1).astype(int)
df.gluc=(df.gluc>1).astype(int)
```

In [ ]:
```
df.gluc.value_counts()
```

```
0    59479
1    10521
Name: gluc, dtype: int64
```

In [ ]:
```
df
```

```
           id    age  sex  height  weight  ap_hi  ap_lo  cholesterol  gluc  \
0           0  18393    2     168    62.0    110     80            0     0
1           1  20228    1     156    85.0    140     90            1     0
2           2  18857    1     165    64.0    130     70            1     0
3           3  17623    2     169    82.0    150    100            0     0
4           4  17474    1     156    56.0    100     60            0     0
...       ...    ...  ...     ...     ...    ...    ...          ...   ...
69995   99993  19240    2     168    76.0    120     80            0     0
69996   99995  22601    1     158   126.0    140     90            1     1
69997   99996  19066    2     183   105.0    180     90            1     0
69998   99998  22431    1     163    72.0    135     80            0     1
69999   99999  20540    1     170    72.0    120     80            1     0

       smoke  alco  active  cardio  overweight
0          0     0       1       0           0
1          0     0       1       1           1
2          0     0       0       1           0
3          0     0       1       1           1
4          0     0       0       0           0
...      ...   ...     ...     ...         ...
69995      1     0       1       0           1
69996      0     0       1       1           1
69997      0     1       0       1           1
69998      0     0       0       1           1
69999      0     0       1       0           0

[70000 rows x 14 columns]
```

	id	age	sex	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	0	0	0	0	1	0
1	1	20228	1	156	85.0	140	90	1	0	0	0	1	1
2	2	18857	1	165	64.0	130	70	1	0	0	0	0	1
3	3	17623	2	169	82.0	150	100	0	0	0	0	1	1
4	4	17474	1	156	56.0	100	60	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	19240	2	168	76.0	120	80	0	0	1	0	1	0
69996	99995	22601	1	158	126.0	140	90	1	1	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	1	0	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	0	1	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	1	0	0	0	1	0

70000 rows × 14 columns

In [ ]:
```
df_cat=pd.melt(df,value_vars=['cholesterol', 'gluc', 'smoke', 'alco', 'active','overwei
df_cat
```

```
        cardio     variable  value
0            0  cholesterol      0
1            1  cholesterol      1
2            1  cholesterol      1
3            1  cholesterol      0
4            0  cholesterol      0
...        ...          ...    ...
419995       0   overweight      1
419996       1   overweight      1
419997       1   overweight      1
419998       1   overweight      1
419999       0   overweight      0

[420000 rows x 3 columns]
```

	cardio	variable	value
0	0	cholesterol	0
1	1	cholesterol	1
2	1	cholesterol	1
3	1	cholesterol	0
4	0	cholesterol	0
...	...	...	...
419995	0	overweight	1
419996	1	overweight	1
419997	1	overweight	1
419998	1	overweight	1
419999	0	overweight	0

420000 rows × 3 columns

In [ ]:
```
data_dict = {
    cardio: data.groupby(["variable", "value"]).size().reset_index(name='total')
    for cardio, data in df_cat.groupby("cardio")
}
for key,data in data_dict.items():
    data["cardio"]=key


df_cat = pd.concat([data_dict[0],data_dict[1]])
df_cat
# df_cat_1=pd.concat([data_dict["cardio_0"],data_dict["cardio_1"]])
# df_cat_1
```

	variable	value	total	cardio
0	active	0	6378	0
1	active	1	28643	0
2	alco	0	33080	0
3	alco	1	1941	0
4	cholesterol	0	29330	0
5	cholesterol	1	5691	0
6	gluc	0	30894	0
7	gluc	1	4127	0
8	overweight	0	15915	0
9	overweight	1	19106	0
10	smoke	0	31781	0
11	smoke	1	3240	0
0	active	0	7361	1
1	active	1	27618	1
2	alco	0	33156	1
3	alco	1	1823	1
4	cholesterol	0	23055	1
5	cholesterol	1	11924	1
6	gluc	0	28585	1
7	gluc	1	6394	1
8	overweight	0	10539	1
9	overweight	1	24440	1
10	smoke	0	32050	1
11	smoke	1	2929	1

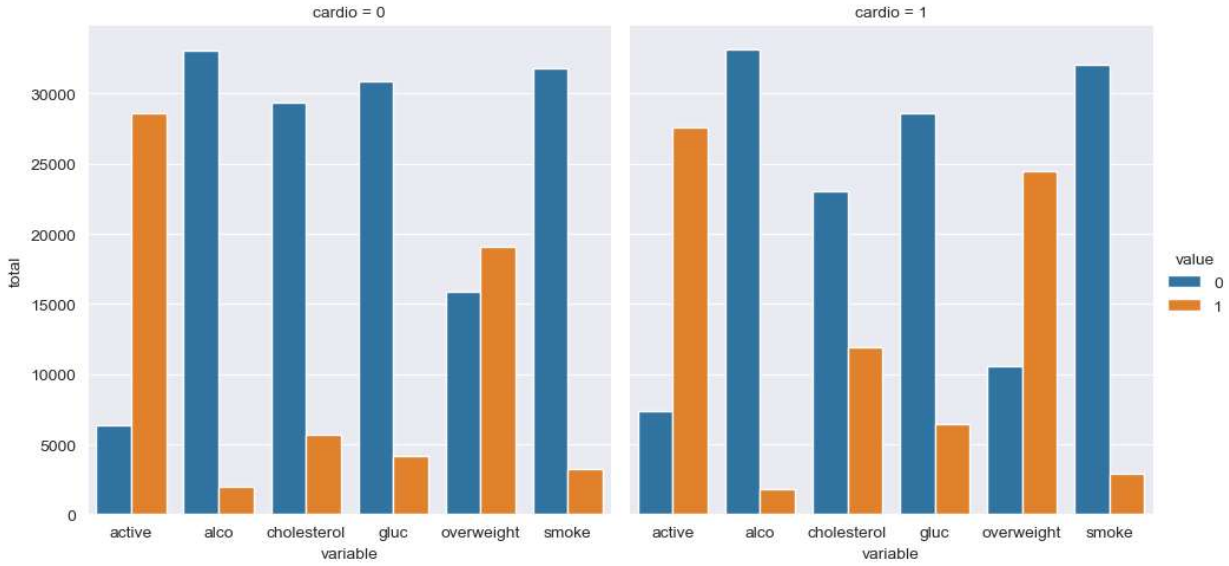
	variable	value	total	cardio
0	active	0	6378	0
1	active	1	28643	0
2	alco	0	33080	0
3	alco	1	1941	0
4	cholesterol	0	29330	0
5	cholesterol	1	5691	0
6	gluc	0	30894	0
7	gluc	1	4127	0
8	overweight	0	15915	0
9	overweight	1	19106	0
10	smoke	0	31781	0
11	smoke	1	3240	0
0	active	0	7361	1
1	active	1	27618	1
2	alco	0	33156	1
3	alco	1	1823	1
4	cholesterol	0	23055	1
5	cholesterol	1	11924	1
6	gluc	0	28585	1
7	gluc	1	6394	1
8	overweight	0	10539	1
9	overweight	1	24440	1
10	smoke	0	32050	1
11	smoke	1	2929	1

The dataset is split by 'Cardio' so there is one chart for each cardio value

In [ ]:
```
fig = sns.catplot(df_cat,kind="bar",col="cardio",x="variable",y="total",hue="value");

    # Do not modify the next two lines
fig.savefig('catplot.png')
```

```
<Figure size 1057.75x500 with 2 Axes>
```



In [ ]:
```
df=df[(df['ap_lo'] <= df['ap_hi'])]
```

```
            id    age  sex  height  weight  ap_hi  ap_lo  cholesterol  gluc  \
0            0  18393    2     168    62.0    110     80            0     0
1            1  20228    1     156    85.0    140     90            1     0
2            2  18857    1     165    64.0    130     70            1     0
3            3  17623    2     169    82.0    150    100            0     0
4            4  17474    1     156    56.0    100     60            0     0
...        ...    ...  ...     ...     ...    ...    ...          ...   ...
69995    99993  19240    2     168    76.0    120     80            0     0
69996    99995  22601    1     158   126.0    140     90            1     1
69997    99996  19066    2     183   105.0    180     90            1     0
69998    99998  22431    1     163    72.0    135     80            0     1
69999    99999  20540    1     170    72.0    120     80            1     0

       smoke  alco  active  cardio  overweight
0          0     0       1       0           0
1          0     0       1       1           1
2          0     0       0       1           0
3          0     0       1       1           1
4          0     0       0       0           0
...      ...   ...     ...     ...         ...
69995      1     0       1       0           1
69996      0     0       1       1           1
69997      0     1       0       1           1
69998      0     0       0       1           1
69999      0     0       1       0           0

[68766 rows x 14 columns]
```

	id	age	sex	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	0	0	0	0	1	0
1	1	20228	1	156	85.0	140	90	1	0	0	0	1	1
2	2	18857	1	165	64.0	130	70	1	0	0	0	0	1
3	3	17623	2	169	82.0	150	100	0	0	0	0	1	1
4	4	17474	1	156	56.0	100	60	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	19240	2	168	76.0	120	80	0	0	1	0	1	0
69996	99995	22601	1	158	126.0	140	90	1	1	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	1	0	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	0	1	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	1	0	0	0	1	0

68766 rows × 14 columns

In [ ]:
```
df = df[(df['height'] >= df['height'].quantile(0.025))]
df = df[(df['height'] <= df['height'].quantile(0.975))]
```

In [ ]:
```
df = df[(df['weight'] >= df['weight'].quantile(0.025))]
df = df[(df['weight'] <= df['weight'].quantile(0.975))]
```

In [ ]:
```
df
```

```
           id    age  sex  height  weight  ap_hi  ap_lo  cholesterol  gluc  \
0           0  18393    2     168    62.0    110     80            0     0
1           1  20228    1     156    85.0    140     90            1     0
2           2  18857    1     165    64.0    130     70            1     0
3           3  17623    2     169    82.0    150    100            0     0
4           4  17474    1     156    56.0    100     60            0     0
...       ...    ...  ...     ...     ...    ...    ...          ...   ...
69993   99991  19699    1     172    70.0    130     90            0     0
69994   99992  21074    1     165    80.0    150     80            0     0
69995   99993  19240    2     168    76.0    120     80            0     0
69998   99998  22431    1     163    72.0    135     80            0     1
69999   99999  20540    1     170    72.0    120     80            1     0

       smoke  alco  active  cardio  overweight
0          0     0       1       0           0
1          0     0       1       1           1
2          0     0       0       1           0
3          0     0       1       1           1
4          0     0       0       0           0
...      ...   ...     ...     ...         ...
69993      0     0       1       1           0
69994      0     0       1       1           1
69995      1     0       1       0           1
69998      0     0       0       1           1
69999      0     0       1       0           0

[57931 rows x 14 columns]
```

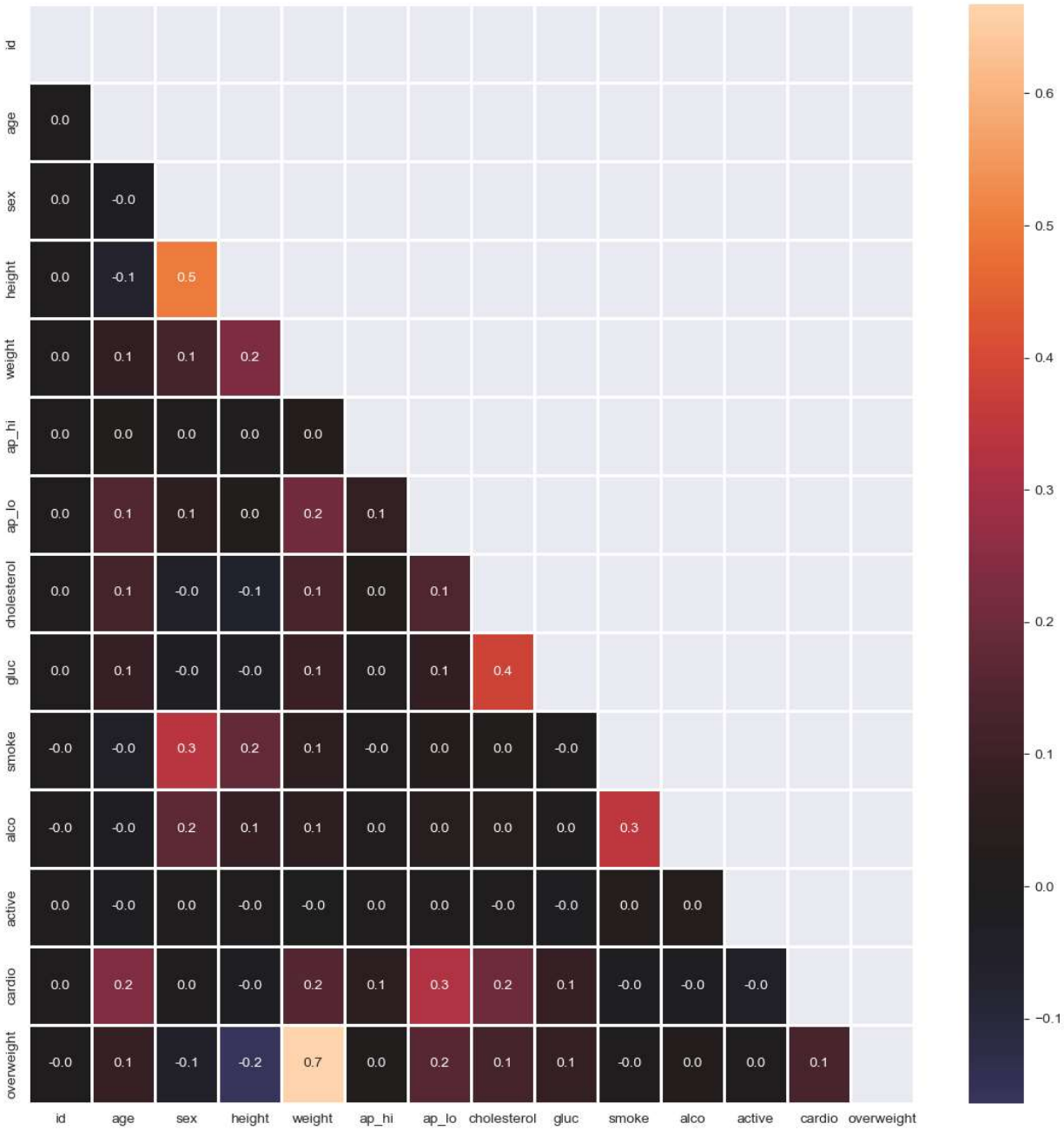
	id	age	sex	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	0	0	0	0	1	0
1	1	20228	1	156	85.0	140	90	1	0	0	0	1	1
2	2	18857	1	165	64.0	130	70	1	0	0	0	0	1
3	3	17623	2	169	82.0	150	100	0	0	0	0	1	1
4	4	17474	1	156	56.0	100	60	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69993	99991	19699	1	172	70.0	130	90	0	0	0	0	1	1
69994	99992	21074	1	165	80.0	150	80	0	0	0	0	1	1
69995	99993	19240	2	168	76.0	120	80	0	0	1	0	1	0
69998	99998	22431	1	163	72.0	135	80	0	1	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	1	0	0	0	1	0

57931 rows × 14 columns

Making a correlation matrix using the dataset and masking the upper half

In [ ]:
```
mask = np.triu(np.ones_like(df.corr()))
fig,ax = plt.subplots(figsize=(14,14))
sns.heatmap(df.corr(),annot=True,mask=mask,fmt='.1f',linewidths=1,center=0,ax=ax)
```

```
<AxesSubplot:>
<Figure size 1400x1400 with 2 Axes>
```


id
age 0.0
sex 0.0 -0.0
height 0.0 -0.1 0.5
weight 0.0 0.1 0.1 0.2
ap_hi 0.0 0.0 0.0 0.0 0.0
ap_lo 0.0 0.1 0.1 0.0 0.2 0.1
cholesterol 0.0 0.1 -0.0 -0.1 0.1 0.0 0.1
gluc 0.0 0.1 -0.0 -0.0 0.1 0.0 0.1 0.4
smoke -0.0 -0.0 0.3 0.2 0.1 -0.0 0.0 0.0 -0.0
alco -0.0 -0.0 0.2 0.1 0.1 0.0 0.0 0.0 0.0 0.3
active 0.0 -0.0 0.0 -0.0 -0.0 0.0 0.0 -0.0 -0.0 0.0 0.0
cardio 0.0 0.2 0.0 -0.0 0.2 0.1 0.3 0.2 0.1 -0.0 -0.0 -0.0
overweight -0.0 0.1 -0.1 -0.2 0.7 0.0 0.2 0.1 0.1 -0.0 0.0 0.0 0.1
id age sex height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio overweight
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1