```
from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving medical_examination.txt to medical_examination.txt
```

```
import pandas as pd

# Read the uploaded TXT file as if it's a CSV (assuming it's comma-separated)
df = pd.read_csv('medical_examination.txt', sep=",", engine="python")

# Display the first few rows
df.head()
```

	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

```
# Add an 'overweight' column (1 if BMI > 25, else 0)
df['overweight'] = (df['weight'] / (df['height'] / 100) ** 2 > 25).astype(int)

# Show a few values to confirm
df[['weight', 'height', 'overweight']].head()
```

	weight	height	overweight
0	62.0	168	0
1	85.0	156	1
2	64.0	165	0
3	82.0	169	1
4	56.0	156	0

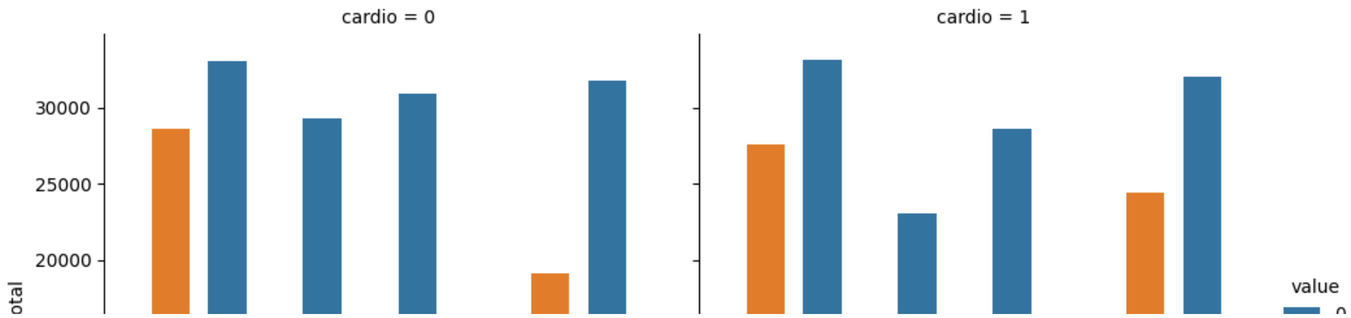
```
# Normalize data: 0 = good, 1 = bad
df['cholesterol'] = (df['cholesterol'] > 1).astype(int)
df['gluc'] = (df['gluc'] > 1).astype(int)

# Check the changes
df[['cholesterol', 'gluc']].head()
```

	cholesterol	gluc
0	0	0
1	1	0
2	1	0
3	0	0
4	0	0

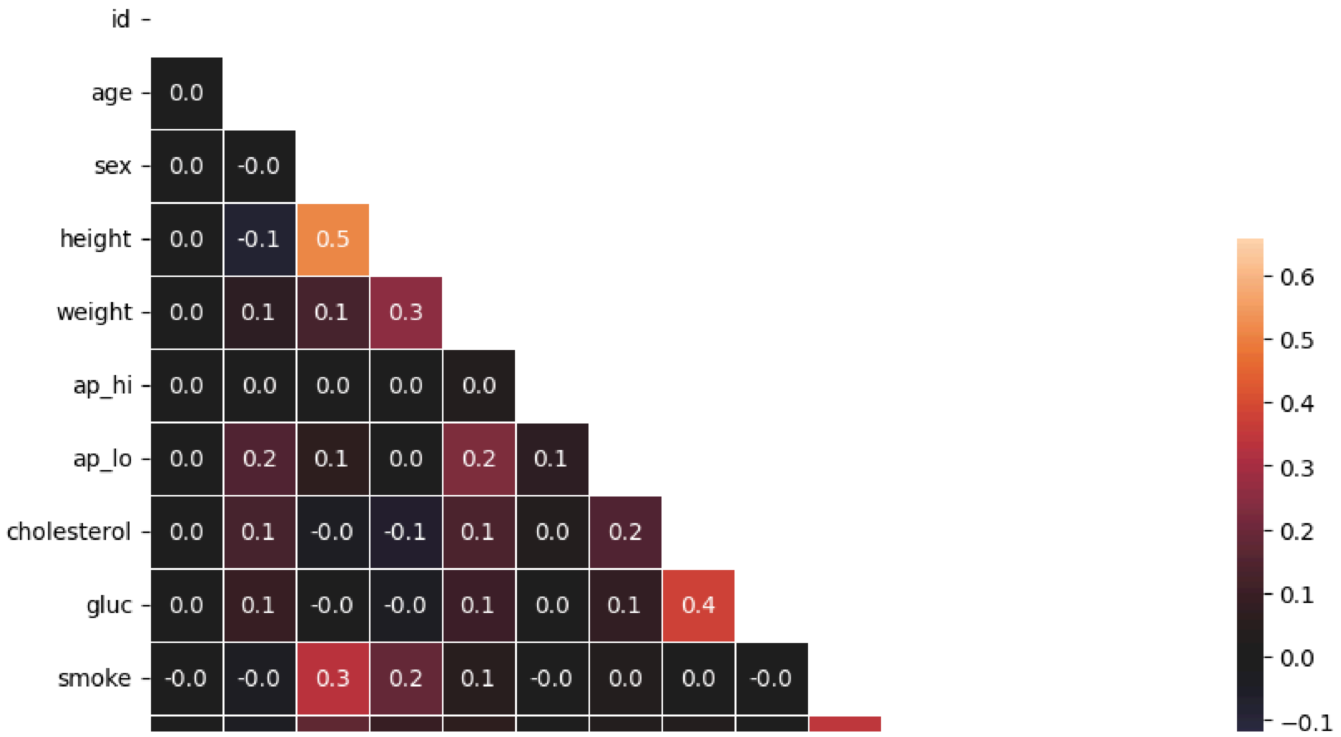
```
# Clean the data
df_cleaned = df[
    (df['ap_lo'] <= df['ap_hi']) &
    (df['height'] >= df['height'].quantile(0.025)) &
    (df['height'] <= df['height'].quantile(0.975)) &
    (df['weight'] >= df['weight'].quantile(0.025)) &
    (df['weight'] <= df['weight'].quantile(0.975))
]

# Check how many rows remain after cleaning
print("Rows before cleaning:", df.shape[0])
print("Rows after cleaning:", df_cleaned.shape[0])
```

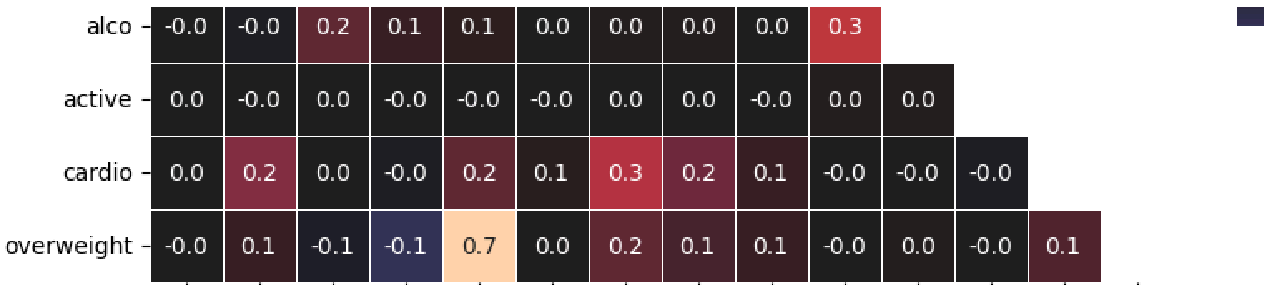
```
Rows before cleaning: 70000
Rows after cleaning: 63259
```

```
import seaborn as sns
import matplotlib.pyplot as plt

# First, prepare the data in "long" format
df_cat = pd.melt(
    df,
    id_vars=['cardio'],
    value_vars=['active', 'alco', 'cholesterol', 'gluc', 'overweight', 'smoke']
)

# Group and reformat the data
df_cat = df_cat.groupby(['cardio', 'variable', 'value'])['value'].count().reset_index(name='total')

# Draw the catplot
cat_plot = sns.catplot(
    data=df_cat,
    x='variable',
    y='total',
    hue='value',
    col='cardio',
    kind='bar'
)

# Show the plot
plt.show()
```



```
import numpy as np

# Calculate the correlation matrix
corr = df_cleaned.corr()

# Create a mask for the upper triangle
mask = np.triu(np.ones_like(corr, dtype=bool))

# Set up the matplotlib figure
fig, ax = plt.subplots(figsize=(10, 8))

# Draw the heatmap
sns.heatmap(
    corr,
    mask=mask,
    annot=True,
    fmt=".1f",
    center=0,
    square=True,
    linewidths=0.5,
    cbar_kws={"shrink": 0.5}
)

plt.show()
```




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 -0.1 0.7 0.0 0.2 0.1 0.1 -0.0 0.0 -0.0 0.1