Department of Data Science - Data and Visual Analytics Lab

Lab9. EDA on Cardiovascular Data

Objectives

In this lab, you will perform Exploratory Data Analysis on Cardiovascular data.

- You will understand the features of the dataset, its size, shape, basic information and datatypes of each feature.

- Then you will perform data cleaning, data wrangling and data visualization on the dataset.

- Further, you will answer several questions about a dataset on cardiovascular disease by writing code in Pandas and visualization.

The machine learning problem requires to predict the presence or absence of cardiovascular disease (CVD) using the patient examination results, which is beyond the scope of your course. You will simply perform EDA on the dataset.

Dataset Description

```
age int (days)
height  int (cm)
weight  float (kg)
gender  categorical code  # 1-male, 2-female
ap_hi   int  # Systolic blood pressure
ap_lo   int  # Diastolic blood pressure
cholesterol 1: normal, 2: above normal, 3: well above normal
gluc    1: normal, 2: above normal, 3: well above normal
smoke   binary  # smoking or not, 0-no, 1-yes
alco    binary # alcohol intake or not
active  binary  # physically active or not
cardio  binary  # presence or absence of cardiovascular discese
```

Import necessary packages

```
In [1]: # import all required modules
        import pandas as pd
        import numpy as np
        # Disable warnings
        import warnings
        warnings.filterwarnings('ignore')
        # Import plotting modules
        import matplotlib.pyplot as plt
        import seaborn as sns
        # import statistical module
```

Import dataset into DataFrame

```
In [2]: df = pd.read_csv("mlbootcamp5_train.csv", sep=';')
        df.head()
```

Out[2]:

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

Print the size

```
In [3]: data.shape
Dataset Size:  (70000, 13)
```

Count Values

How many people smoke?

In [4]:
```
data.smoke.value_counts()
```

Out[4]:
```
0    63831
1     6169
Name: smoke, dtype: int64
```

How many people consume alcohol?

In [5]:
```
df.alco.value_counts()
```

Out[5]:
```
0    66236
1     3764
Name: alco, dtype: int64
```

What are the difference glucose levels?

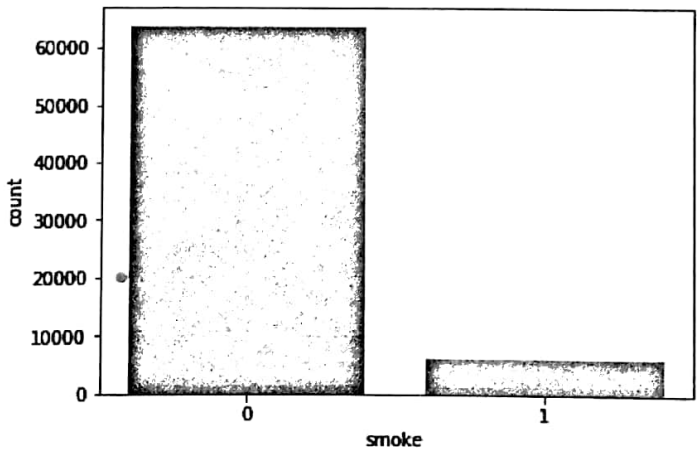
In [6]:
```
data.gluc.value_counts()
```

Out[6]:
```
1    59479
3     5331
2     5190
Name: gluc, dtype: int64
```

Draw bar chart for smoke column

In [7]:
```
sns.countplot(x='smoke', data=data)
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1458000eda0>
```



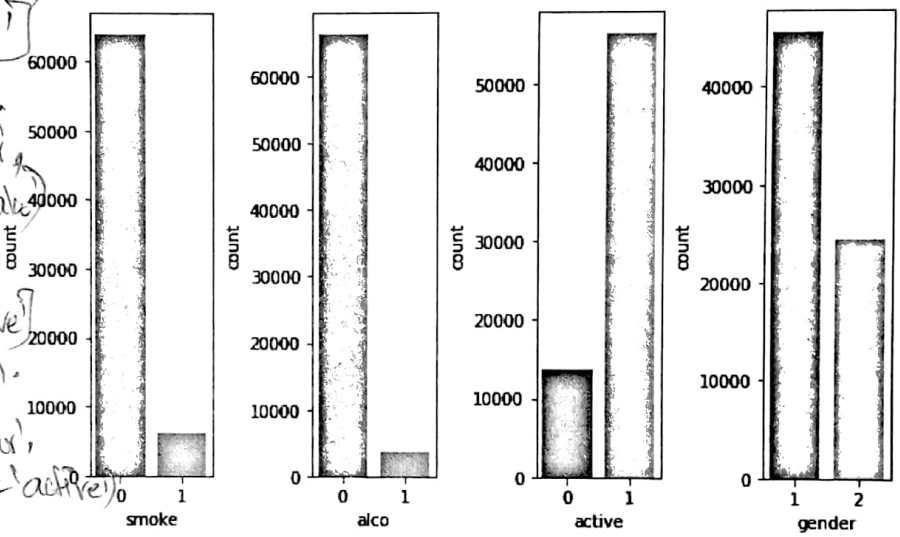
Draw 4 count plots for gender, smoke, alco and active columns respectively in 1 row, 4 columns

In [8]:
```
# First extract all 4 columns into a dataframe, binary_df
binary_df = data[['gender', 'smoke', 'alco', 'active']]
```

Then, plot count plots

In [9]:
```
plt.subplot(231)
binary_df['gender'].value_counts().plot(kind='bar', title='gender', figsize=(20,10))
plt.subplot(232)
binary_df['smoke'].value_counts().plot(kind='bar', title='smoke')
plt.subplot(233)
binary_df['alco'].value_counts().plot(kind='bar', title='alco')
plt.subplot(234)
binary_df['active'].value_counts().plot(kind='bar', title='active')
plt.legend()
plt.show()
```

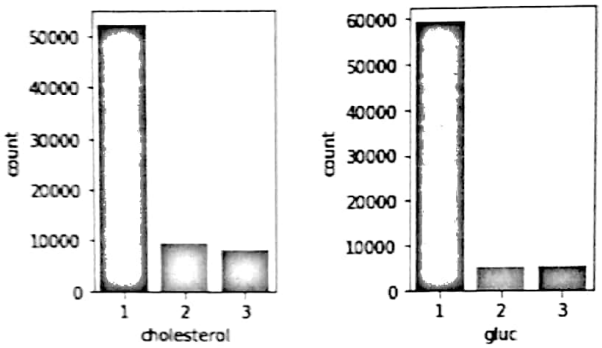
```
C:\Users\Rajkumar\Anaconda3\lib\site-packages\matplotlib\figure.py:445: UserWarning: Matplotlib is currently using module://ipykernel.pylab.backend_inline, which is a non-GUI backend, so cannot show the figure.
  % get_backend())
```



Draw a count plot for cholesterol and gluc columns

In [10]:
```
sns.countplot(x="variable", hue="value", data=pd.melt(data
```

```
C:\Users\Rajkumar\Anaconda3\lib\site-packages\matplotlib\figure.py:445: UserW
arning: Matplotlib is currently using module://ipykernel.pylab.backend_inlin
e, which is a non-GUI backend, so cannot show the figure.
  % get_backend())
```

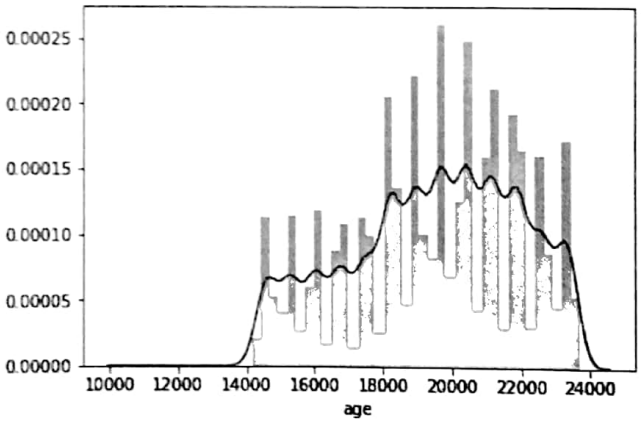


Plot Data Distribution

Show the distribution of age values as histogram

In [11]:
```
sns.distplot(data.age)
```

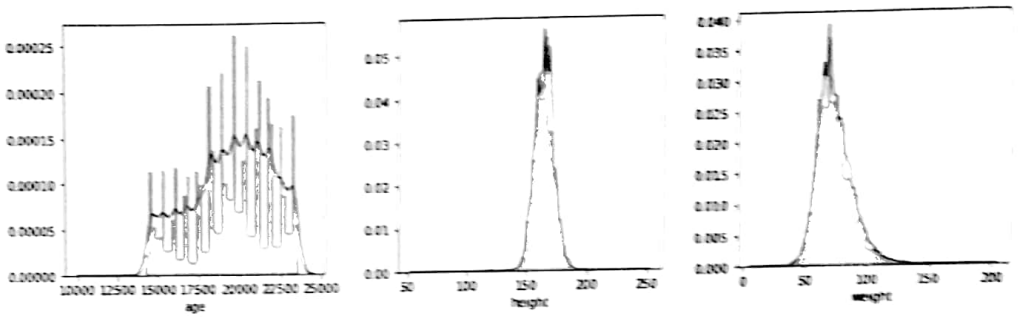
```
<matplotlib.axes._subplots.AxesSubplot at 0x14582dd7f28>
```



Show the distribution of age, height and weight values as 3 histograms in one plot

In [12]:

```
C:\Users\Rajkumar\Anaconda3\lib\site-packages\matplotlib\figure.py:445: UserW
arning: Matplotlib is currently using module://ipykernel.pylab.backend_inlin
e, which is a non-GUI backend, so cannot show the figure.
  % get_backend())
```



Calculate Summary Statistics Using Pandas

1. How many men and women are present in this dataset?

In [13]:

```
#Now, count gender column
```

```
Out[13]: 1    45530
         2    24470
         Name: gender, dtype: int64
```

But, we do not know if 1 means male or female. Similarly, 2 means male or female. We need to somehow find it out. How to do that?. When we inspect other columns, we can find out that there is a column "height" in centimeters. So, we can assume that men are more taller than women, generally.

So, we can compute the average height for gender=1 and gender=2. The largest average value will denote "male".

In [14]:

```
Out[14]: gender
         1    161.355612
         2    169.947895
         Name: height, dtype: float64
```

161 cm and almost 170 cm on average, so we make a conclusion that gender=1 represents females, and gender=2 – males.

Therefore, looking at the value_counts() of gender column, we can conclude that the dataset contains 45530 women and 24470 men.

2. Which gender more often reports consuming alcohol - men or women?

```
In [15]: temp['alco'].mean()
Out[15]: gender
         1    0.025500
         2    0.106375
         Name: alco, dtype: float64
```

Here, larger value is 2, which denotes men

3. Which gender is more physically active - men or women?

```
In [16]: temp['active'].mean()
Out[16]: gender
         1    0.802021
         2    0.806906
         Name: active, dtype: float64
```

Here, larger values denotes 2, so answer is men

4. What is the the rounded difference between the percentages of smokers among men and women (rounded)?

First, let us find who smokes more.

```
In [17]: temp['smoke'].mean()
Out[17]: gender
         1    0.017856
         2    0.218880
         Name: smoke, dtype: float64
```

So, men smokes more tha women. Now, let us find out what percentage men smokes more than women

```
In [18]: round((data[data['smoke']==0]['age'].median() - data[data
         ['smoke']==1]['age'].mean))
Out[18]: 20
```

5. What is the difference between median values of age for smokers and non-smokers (in months, rounded)? You'll need to figure out the units of feature age in this dataset

In the dataset, age is given in terms of days. Therefore, you should divide by 365 to convert age into years. First, find the median age in years of smoke category.

In [19]:
```
data['yearly'] = data['age'].apply(lambda x : x / 365)
temp1 = data.groupby('smoke'), temp1['yearly'].median()
```

Out[19]:
```
smoke
0    53.995893
1    52.361396
Name: age, dtype: float64
```

Median age of smokers is 52.4 years, for non-smokers it's 54. We see that the correct answer is 20 months.

Now, subtract the median age to find out the difference.

In [20]:
```
(data[data['smoke'] == 0]['yearly'].median() - data[data['smoke']
                                                == 1]['yearly'].mean
```

Out[20]:
```
19.613963039014372
```

Perform Risk Analysis

Calculate a new feature, age_years

The age variable represents age in days. You need to transform each age into years rounded as integer and store in new column, age_years

In [21]:
```
data = data.drop(['yearly'], axis=1)
data.head()
```

Check age_years column using head()

In [22]: df.head()

Out[22]:

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

What is maximum age_years?

In [23]: df['age_years'].max()

Out[23]: 64

What is minimum age_years?

In [24]: data.age_years.min()

Out[24]: 29

Risk Factors for Cardio Vascular Discese

Men who are 50 and above
Men who are smokers
Men whose cholesterol level > 1
Men whose systolic pressure is from 160 to 180 (both inclusive)

How many risky men are in the dataset?

How many people who are 50 and above?

In [25]:
```
data['old_df'] = data['age_years']
data.loc[data.age_years >=50, 'old_df'] = True
data.loc['data.age_years < 50, 'old_df'] = False
```

In [26]:
```
# Show its head()

df_old.head()
```

Out[26]:
```
0     True
1     True
2     True
3    False
4    False
Name: age_years, dtype: bool
```

Now, count its unique values

In [27]:
```
data.old_data.value_counts()
```

Out[27]:
```
True     48591
False    21409
Name: age_years, dtype: int64
```

Therefore, there are 48591 people who are 50 years and above

How many are 50 years and above and men and smokers?

In [28]:
```
df_smoke_old_men = data.loc[(data.gender==2) & (data.smoke==1) & (data.age_years)]
```

In [29]:
```
# prit top-5 from df_smoke_old_men
data_smoke_old_men.head()
```

Out[29]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active
19	29	21755	2	162	56.0	120	70	1	1	1	0	1
38	52	23388	2	162	72.0	130	80	1	1	1	0	1
67	90	22099	2	171	97.0	150	100	3	1	1	0	1
105	140	20627	2	168	78.0	140	90	2	1	1	0	1
121	166	19507	2	174	77.0	120	80	1	1	1	0	1

How many old men have their cholesterol level > 1 and systolic pressure is from 160 to 180 too ?

In [30]:
```
risky_men = (data_smoke_old_men.cholesterol>=1) & (data_smoke_old_men.ap_hi>=160)
```

In [31]:
```
# Print its head risky_men.head()
```

Out[31]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active
230	318	23376	2	175	75.0	180	100	3	1	1	1	1
732	1032	21652	2	167	70.0	160	90	2	1	1	1	1
2786	3930	21799	2	171	94.0	160	100	2	2	1	0	1
4099	5807	19749	2	183	85.0	180	110	2	1	1	0	1
4216	5950	19063	2	175	94.0	170	110	3	3	1	0	0

What is the size of risky_men ?

In [32]:
```
risky_men.shape
```

Out[32]:
```
(130, 14)
```

Therefore, ther are 136 risky men are in the dataset

How many risky men have cardiovascular discese out of these 136 samples?

In [33]:
```
risky_men.cardio.value_counts()
```

Out[33]:
```
True     116
False     14
Name: cardio, dtype: int64
```

Conclusion: There are 122 cardiovascular discese men in the dataset

Compute Body Mass Index

Create a new feature – BMI. To do this, divide weight in kilogramms by the square of the height in meters. Normal BMI values are said to be from 18.5 to 25.

In our dataset, height is in centimeters. So, while you are computing BMI, you have to convert into meters by dividing it by 100

Create a column bmi and store the bmi values

In [34]:
```
data['height'] = data['height'].apply(lambda x: x/100)
data['BMI'] = data.apply(lambda x: x.weight/(x.height*x.height), axis=1)
```

In [35]:
```
df.head()
```

Out[35]:

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

How many people have ideal BMI values?

We already know that ideal BMI values are said to be from 18.5 to 25.

Compute ideal bmi values using bmi column and store the result in a new column, ideal_bmi

In [36]:
```
ideal_bmi = data[(data.BMI>18.5)&(data.BMI<25)]
```

In [37]:
```
ideal_bmi.shape
```

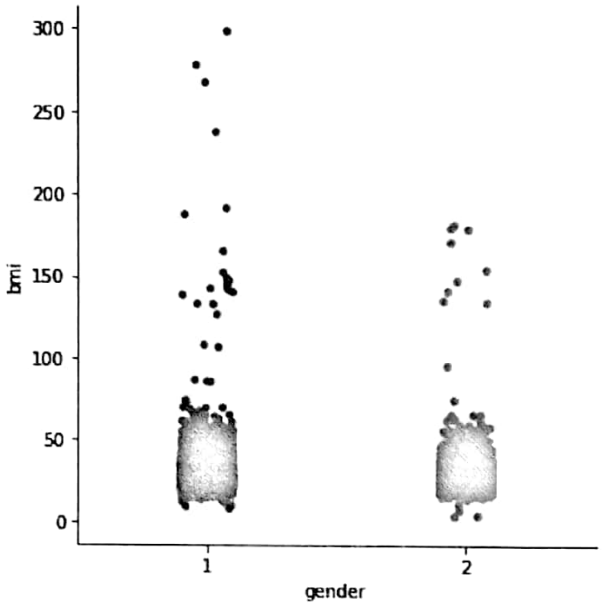
Out[37]:
```
(25804, 15)
```

25804 people have ideal BMI values

Draw catplot between gender and bmi values

In [38]: sns.catplot(x='gender', y='BMI', data=data)

Out[38]: <seaborn.axisgrid.FacetGrid at 0x145831a63c8>



Looking at catplot, is BMI of male is larger than BMI of female (we know 1-female, 2-male already) ?

From the plot, we can conclude Female bmi is greater than Male bmi

Is median value of Men's BMI is higher then women's BMI?

Compute median bmi for gender

In [39]: data.groupby('gender')['BMI'].median()

Out[39]:
```
gender
1    26.709402
2    25.910684
Name: bmi, dtype: float64
```

From the above values, we conclude that Female have higher BMI values than male

Consider the output of the following query and answer the questions

```
In [40]: df.groupby(['gender', 'alco', 'cardio'])['bmi'].median().to_frame()
```

Out[40]:

			bmi
gender	**alco**	**cardio**	
1	**0**	**0**	25.654372
		1	27.885187
	1	**0**	27.885187
		1	30.110991
2	**0**	**0**	25.102391
		1	26.674874
	1	**0**	25.351541
		1	27.530797

Is it true?. Healthy people have, on average, a higher BMI than the people with CVD.

Is it true?. For healthy, non-drinking men, BMI is closer to the norm than for healthy, non-drinking women

Data Cleaning

Remove the following people, that we consider to have erroneous data, from the dataset

- diastilic pressure is higher then systolic
- height is strictly less than 2.5%-percentile
- height is strictly more than 97.5%-percentile
- weight is strictly less then 2.5%-percentile
- weight is strictly more than 97.5%-percentile

Here, we will retain those records which do not satisfy the above conditions

```
In [41]: filtered_df =
```

```
print(filtered_df.shape[0] / df.shape[0])
```

```
0.9037
```

So, what percentage of people do you remove from dataset?

Visual Data Analytics

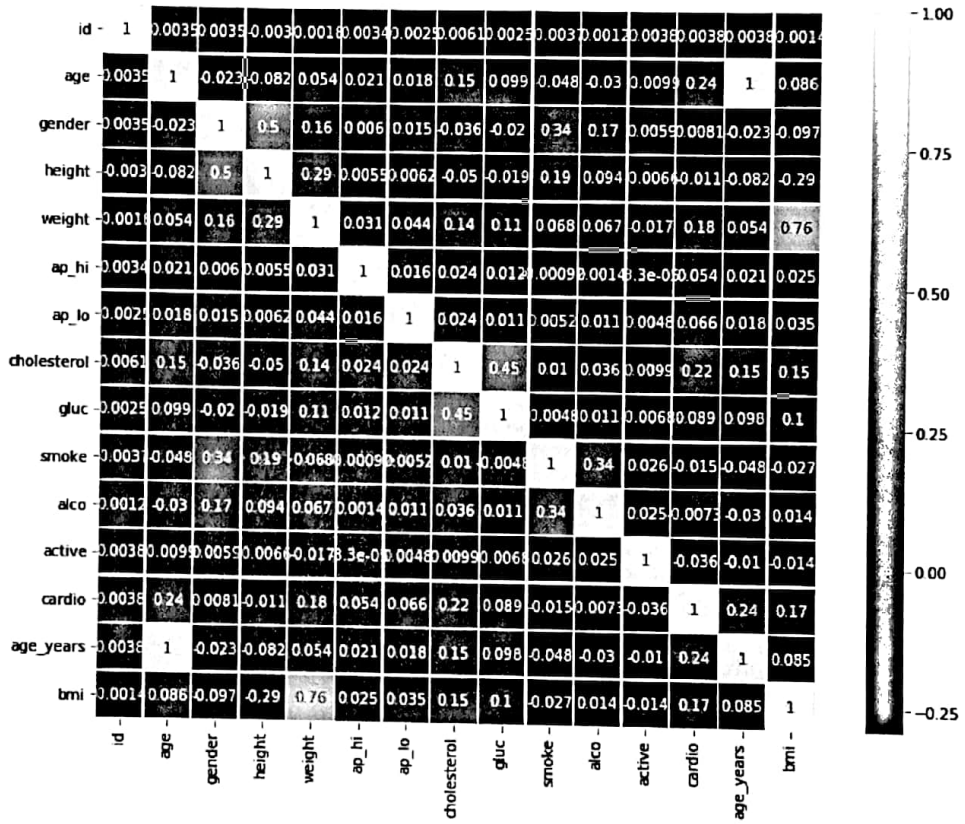
Correlation matrix visualization

To understand the features better, you can create a matrix of the correlation coefficients between the features. Use the initial dataset (non-filtered).

Plot a correlation matrix using heatmap().

In [42]: `plt.figure(figsize=(20,10)), sns.heatmap(df.corr(), annot=True)`

Out[42]: `<matplotlib.axes._subplots.AxesSubplot at 0x145834951d0>`



From the Heatmap, find out top two features that have strongest Pearson's correlation with the gender feature.

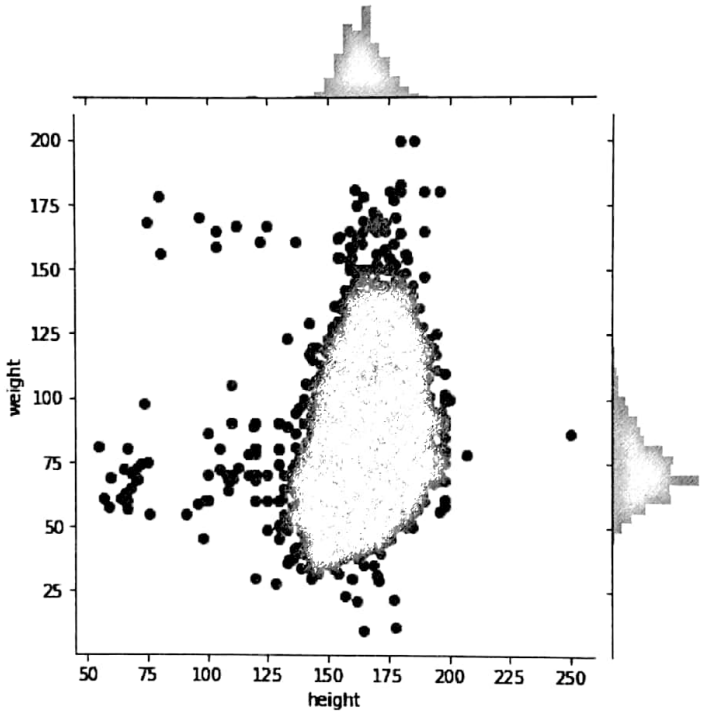
In the Heatmap, which feature strongly correlates to weight?

Height and Weight Distribution

Joint Plot between height and weight columns

Let us see how two independent variables, height and weight, are distributed in the dataset using Joint Plot.
Draw a Joint Plot

```
In [43]: sns.jointplot(x='height', y='weight', data=data)
Out[43]: <seaborn.axisgrid.JointGrid at 0x14582d05080>
```

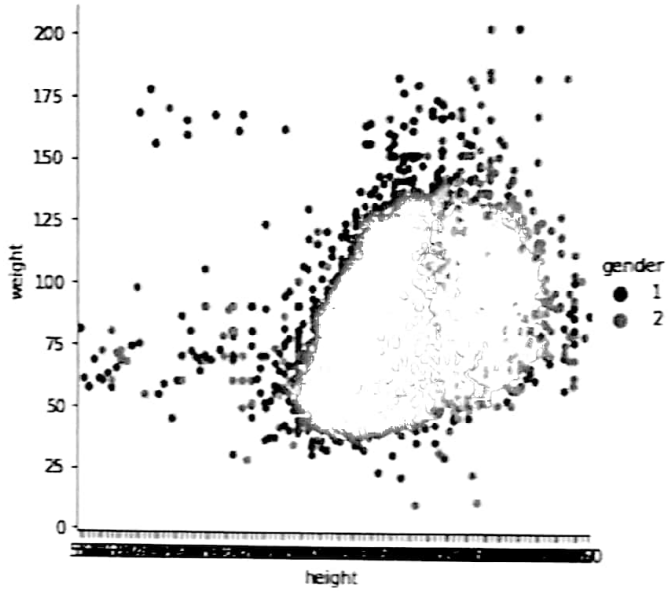


Distribution of height and weight for gender

Draw a catplot between height and weight with hue as "gender"

In [44]:
```
sns.catplot(x='height', y='weight', data=data, hue='gender')
```

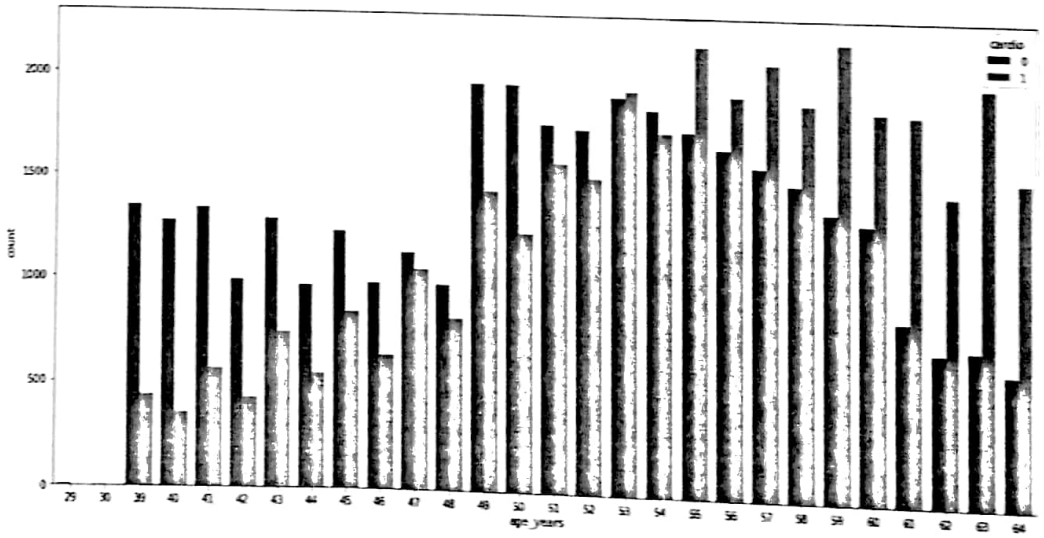
Out[44]:
```
<seaborn.axisgrid.FacetGrid at 0x14582d5fdd8>
```



Find relationship between age_years and Cardio discese. Draw countplot with hue as "cardio"

In [45]:
```
plt.figure(figsize=(15,9))
sns.countplot(x='age_years', hue='cardio', data=data)
```

Out[45]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x14585afdeb8>
```



From the above figure, we know critical age for cardio discese is between 50 and 60.

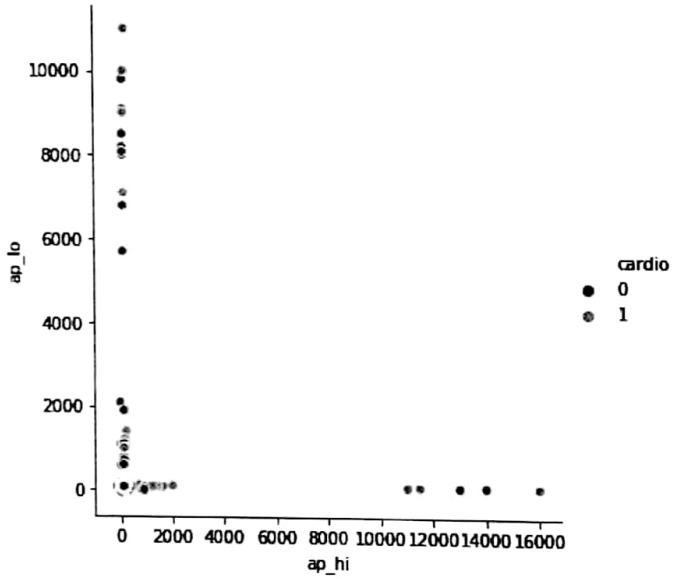
Note: You should use plt.rcParams to modify figure size.

How diastilic and systolic values affect cardio patients?

Draw Boxen plot

for plotting a large number of quantiles, which provides more insights about the shape of the distribution

```
In [46]: sns.relplot(x='ap_hi', y='ap_lo', hue='cardio', data=data)
Out[46]: <seaborn.axisgrid.FacetGrid at 0x145860f4160>
```



Since, the range of ap_hi and ap_lo values very large, the plot appears too contensed.

Now, print max and min values and justify.

```
In [47]: data.ap_hi.max()
Out[47]: 16020

In [48]: data.ap_hi.min()
Out[48]: -150
```

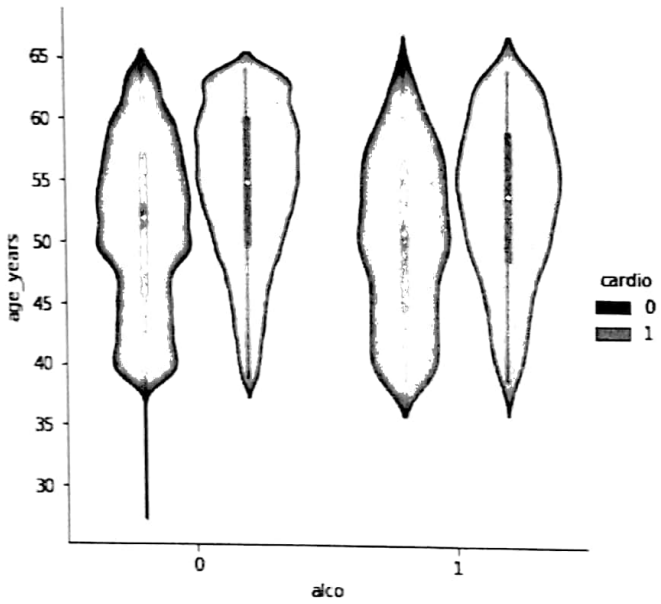
```
In [49]: data.ap_lo.max()
Out[49]: 11000

In [50]: data.ap_lo.min()
Out[50]: -70
```

How alcohol intake and age affect cardios?

Draw Violin Plot to represent relationship between alcohol intake and age_years with hue as "cardio"

```
In [51]: sns.catplot(x='alco', y='age_years', data=data, hue='cardio',
                     kind='violin')
Out[51]: <seaborn.axisgrid.FacetGrid at 0x145863e6278>
```



From this plot, we can understand the distribution of age values among alcohol consumers for cardio discese

1. For Non alcoholic category (ie., alco=0), what is the 50th percentile value for Non-Cardio (ie., cardio=0) people?

```
In [ ]: gdp=.data.groupby(['alco','cardio'])['age_years']
```

2. For Non alcoholic category (ie., alco=0), what is the 50th percentile value for Cardio (ie., cardio=1) people?

In [ ]: pd.loc[0,1]['50%']

3. For alcoholic category (ie., alco=1), what is the 25th percentile value for Non-Cardio (ie., cardio=0) people?

In [ ]: pd.loc[1,0]['25%']

4. For alcoholic category (ie., alco=1), what is the 25th percentile value for Cardio (ie., cardio=1) people?

In [ ]: pd.loc[1,1]['25%']