Import necessary packages

In [1]:
```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
import seaborn as sns
```

Import dataset into DataFrame

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

Out[2]:

	id	age	gender	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

Print the size

In [3]:
```
data.shape
```

Out[3]:
```
(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?

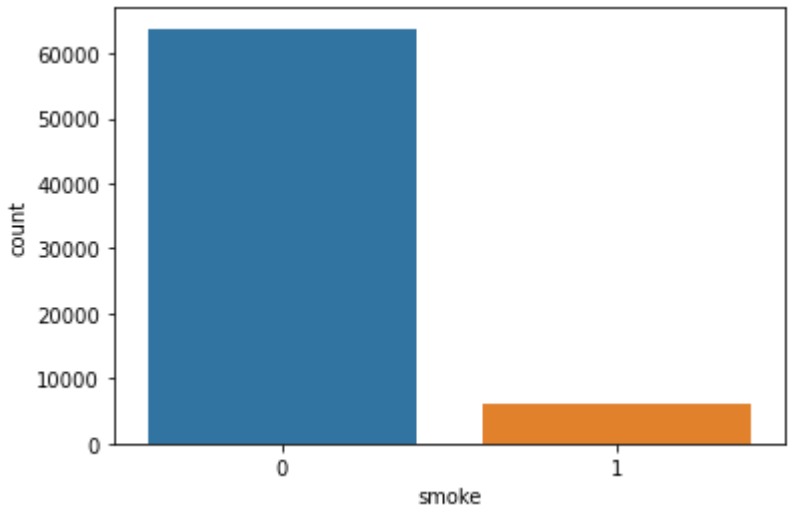
In [5]:
```
data.gluc.value_counts()
```

Out[5]:
```
1    59479
3     5331
2     5190
Name: gluc, dtype: int64
```

Draw bar chart for smoke column

In [6]:
```
sns.countplot(x='smoke',data=data)
```

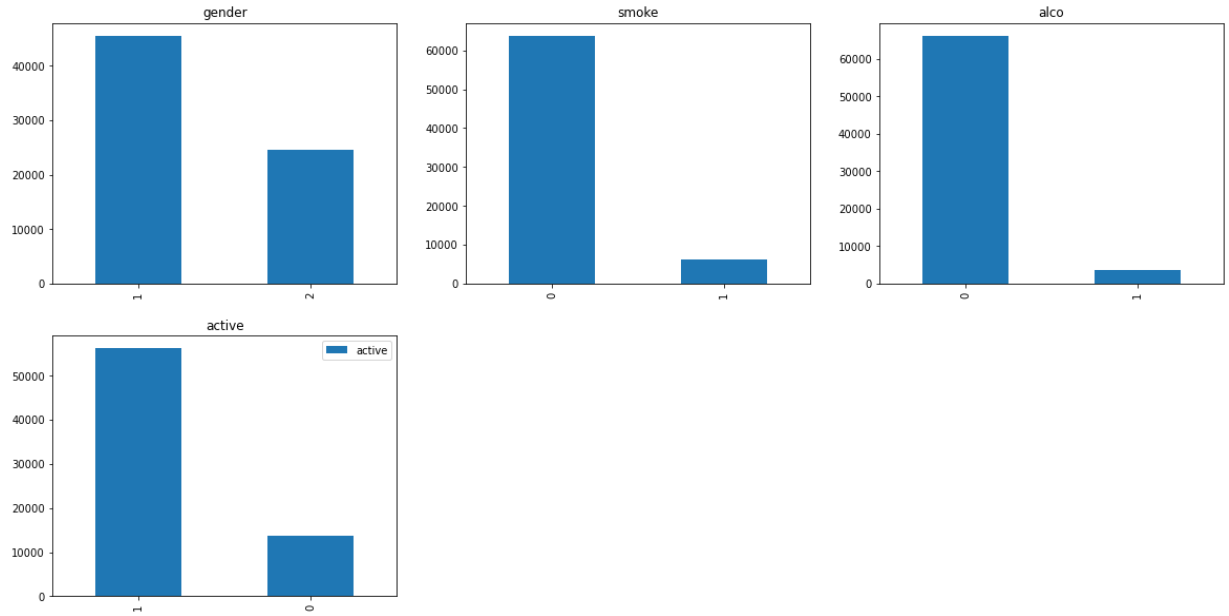
Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xb631bb0>
```



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

In [9]:
```
binary_df = data[['gender','smoke','alco','active']]
```

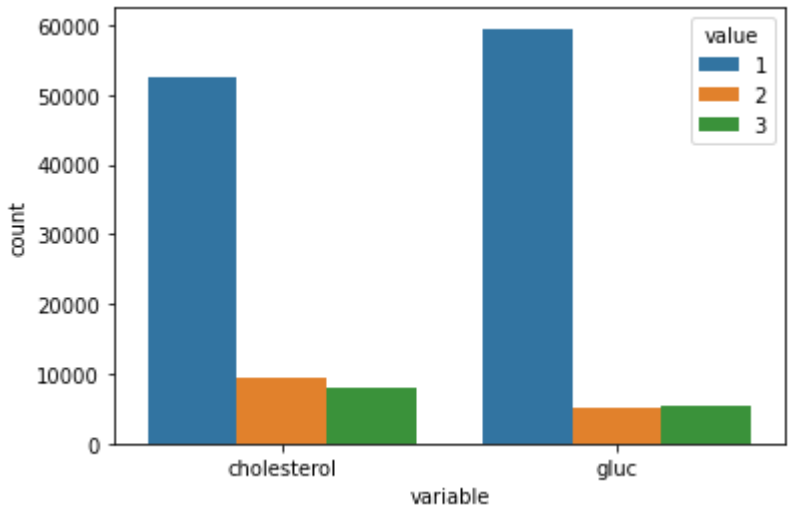
In [15]:
```
plt.subplot(231)
binary_df['gender'].value_counts().plot(kind='bar',title='gender',figsize = (20,1

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()
```



Draw a count plot for cholesterol and gluc columns

In [16]:
```
sns.countplot(x="variable", hue='value', data = pd.melt(data[['cholesterol','gluc
```

Out[16]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xbbda640>
```

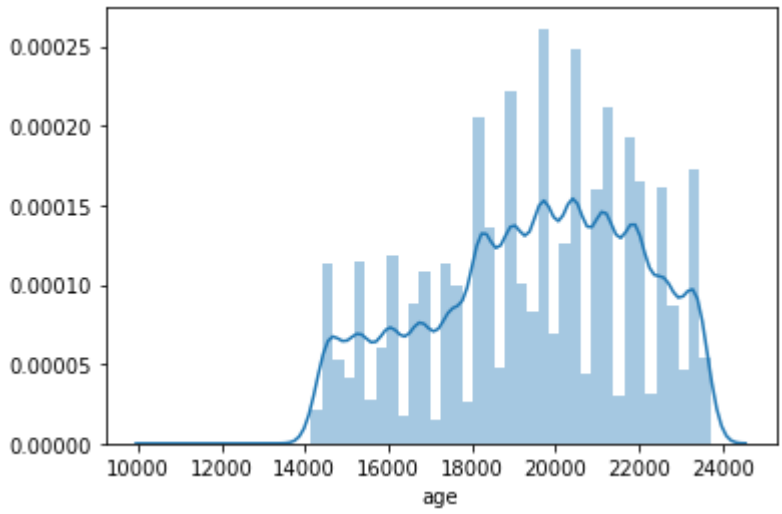


Plot Data Distribution

Show the distribution of age values as histogram

In [17]:
```
sns.distplot(data.age)
```

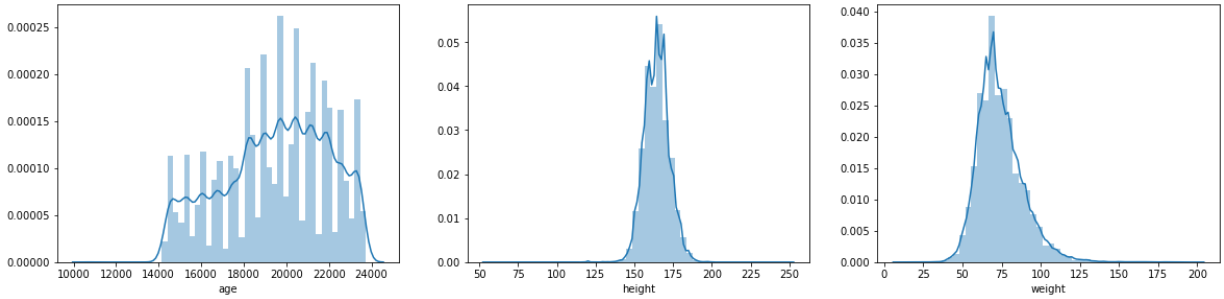
Out[17]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xbc3c6d0>
```



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

In [18]:
```
plt.figure(figsize = (20,10))

plt.subplot(231)
sns.distplot(data.age)

plt.subplot(232)
sns.distplot(data.height)

plt.subplot(233)
sns.distplot(data.weight)
```

Out[18]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x578fa60>
```



Calculate Summary Statistics Using Pandas

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

In [19]:
```
data.gender.value_counts()
```

Out[19]:
```
1    45530
2    24470
Name: gender, dtype: int64
```

In [20]:
```
temp = data.groupby('gender')
```

In [21]:
```
temp['height'].mean()
```

Out[21]:
```
gender
1    161.355612
2    169.947895
Name: height, dtype: float64
```

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

In [22]:
```
temp['alco'].mean()
```

Out[22]:
```
gender
1    0.025500
2    0.106375
Name: alco, dtype: float64
```

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

In [23]:
```
temp['active'].mean()
```

Out[23]:
```
gender
1    0.802021
2    0.806906
Name: active, dtype: float64
```

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

In [24]:
```
temp['smoke'].mean()
```

Out[24]:
```
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 [25]:
```
round((data[data['smoke']==0]['age'].median() - data[data['smoke']==1]['age'].med
```

Out[25]:
```
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 [26]:
```
data['yearly'] = data['age'].apply(lambda x : x/365)
temp1 = data.groupby('smoke')
temp1['yearly'].median()
```

Out[26]:
```
smoke
0    54.032877
1    52.397260
Name: yearly, 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 [27]:
```
(data[data['smoke']==0]['yearly'].median() - data[data['smoke']==1]['yearly'].mec
```

Out[27]:
```
19.62739726027391
```

In [22]:
```
data = data.drop(['yearly'],axis=1)
data.head()
```

Out[22]:

	id	age	gender	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

Perform Risk Analysis

Calculate a new feature, age_years

In [28]:
```
data['age_years'] =data['age'].apply(lambda x:int(x/365))
```

Check age_years column using head()

In [29]:
```
data.head()
```

Out[29]:

	id	age	gender	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

What is maximum age_years?

In [25]:
```
df.age_years.max()
```

Out[25]:
```
64
```

What is minimum age years?

In [26]:
```
data.age_years.min()
```

Out[26]:
```
29
```

How many risky men are in the dataset?

In [30]:
```
data['risky'] = data[data['gender']==2]['age_years'].apply(lambda x:1 if x >50 el
data['risky'] = data[data['gender']==2]['smoke'].apply(lambda x:1 if x==1 else 0)
data['risky'] = data[data['gender']==2]['cholesterol'].apply(lambda x:1 if x>1 el
data['risky'] = data[data['gender']==2]['ap_hi'].apply(lambda x:1 if x>=160 and x
```

How many people who are 50 and above?

In [31]:
```
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 [33]:
```
data['old_df'].head()
```

Out[33]:
```
0     True
1     True
2     True
3    False
4    False
Name: old_df, dtype: object
```

Now, count its unique values

In [30]:
```
data.old_data.value_counts()
```

Out[30]:
```
True     48591
False    21409
Name: old_df, dtype: int64
```

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

In [34]:
```
data_smoke_old_men=data.loc[(data.gender==2) & (data.smoke ==1) & (data.age_years
```

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

Out[35]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	ca
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 [36]:
```
ta_smoke_old_men.cholesterol == 1) & (data_smoke_old_men.ap_hi >=160) & (data_smol
```

In [37]:
```
risky_men.head()
```

Out[37]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active
697	986	22615	2	171	108.0	161	73	1	1	1	0	1
1434	2014	21143	2	169	85.0	180	100	1	1	1	1	0
2693	3799	21906	2	170	94.0	160	90	1	1	1	1	1
3093	4362	18755	2	185	108.0	160	90	1	1	1	0	1
3116	4396	19631	2	173	79.0	160	100	1	1	1	0	1

What is the size of risky_men ?

In [38]:
```
risky_men.shape
```

Out[38]:
```
(173, 17)
```

How many risky men have cardiovascular discese out of these 130 samples?

In [39]:
```
risky_men.cardio.value_counts()
```

Out[39]:
```
1    153
0     20
Name: cardio, dtype: int64
```

Compute Body Mass Index

Create a column bmi and store the bmi values

In [37]:
```
data['height'] = data['height'].apply(lambda x:x/100)
```

In [40]:
```
data['BMI'] = data.apply(lambda x : x.weight/(x.height*x.height),axis=1)
```

How many people have ideal BMI values?

In [45]:
```
ideal_bmi = data[(data.BMI>18.5) & (data.BMI>25)]
```

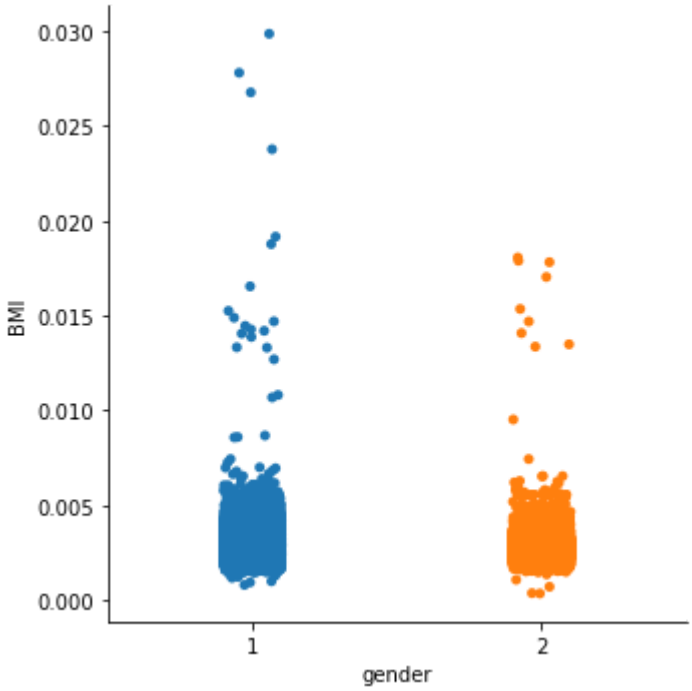
In [46]:
```
ideal_bmi.shape
```

Out[46]:
```
(0, 18)
```

Draw catplot between gender and bmi values

In [47]:
```
sns.catplot(x='gender',y='BMI',data=data)
```

Out[47]:
```
<seaborn.axisgrid.FacetGrid at 0xc40f0a0>
```



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

In [48]:
```
data.groupby('gender')['BMI'].median()
```

Out[48]:
```
gender
1    0.002671
2    0.002591
Name: BMI, dtype: float64
```

Consider the output of the following query and answer the questions

In [49]:
```
data.groupby(['gender','alco','cardio'])['BMI'].median().to_frame()
```

Out[49]:

			BMI
gender	**alco**	**cardio**	
1	**0**	**0**	0.002565
		1	0.002789
	1	**0**	0.002789
		1	0.003011
2	**0**	**0**	0.002510
		1	0.002667
	1	**0**	0.002535
		1	0.002753

Data Cleaning

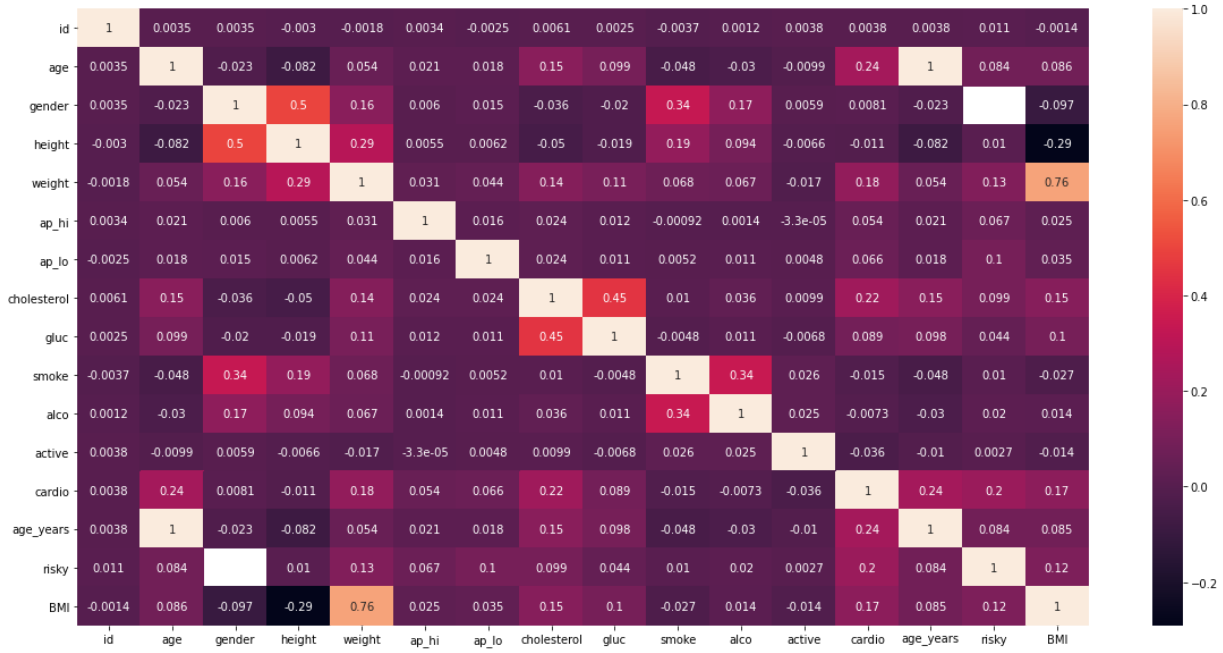
In [ ]:
```

```

Visual Data Analytics

Correlation matrix visualization

In [44]:
```
plt.figure(figsize = (20,10))
sns.heatmap(df.corr(),annot=True)
```

Out[44]: <AxesSubplot:>

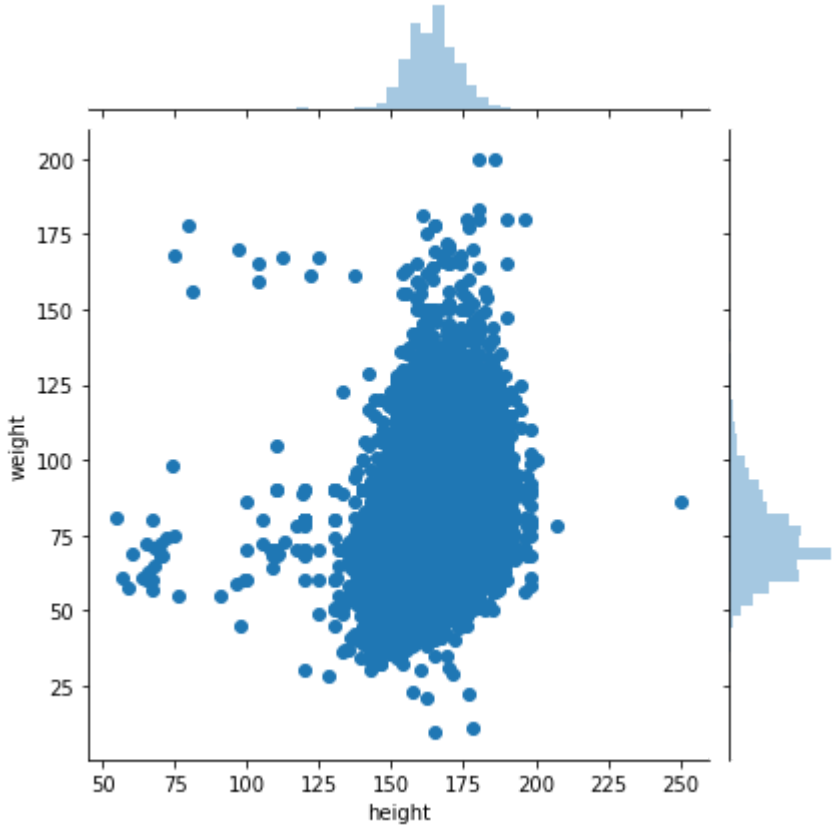


Height and Weight Distribution

Joint Plot between height and weight columns

In [50]:
```
sns.jointplot(x='height',y='weight',data=data)
```

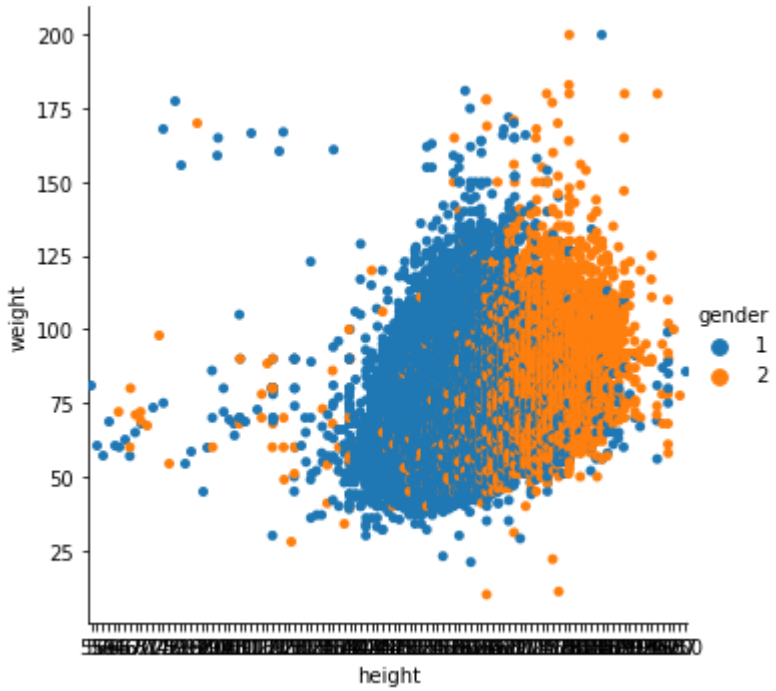
Out[50]:
```
<seaborn.axisgrid.JointGrid at 0x5698a30>
```



Distribution of height and weight for gender

In [51]:
```
sns.catplot(x='height',y='weight',data=data,hue='gender')
```

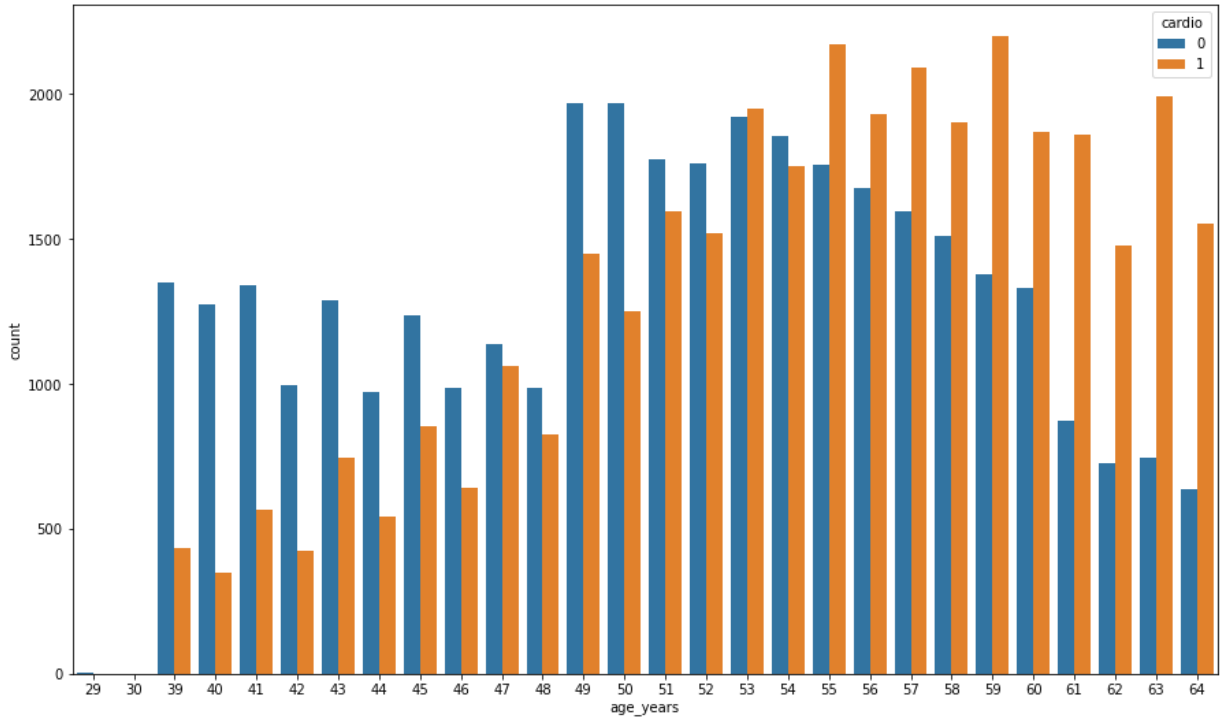
Out[51]:
```
<seaborn.axisgrid.FacetGrid at 0xcba9490>
```



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

In [53]:
```
plt.figure(figsize = (15,9))
sns.countplot(x='age_years',hue='cardio',data=data)
```

Out[53]: <matplotlib.axes._subplots.AxesSubplot at 0x104f6640>

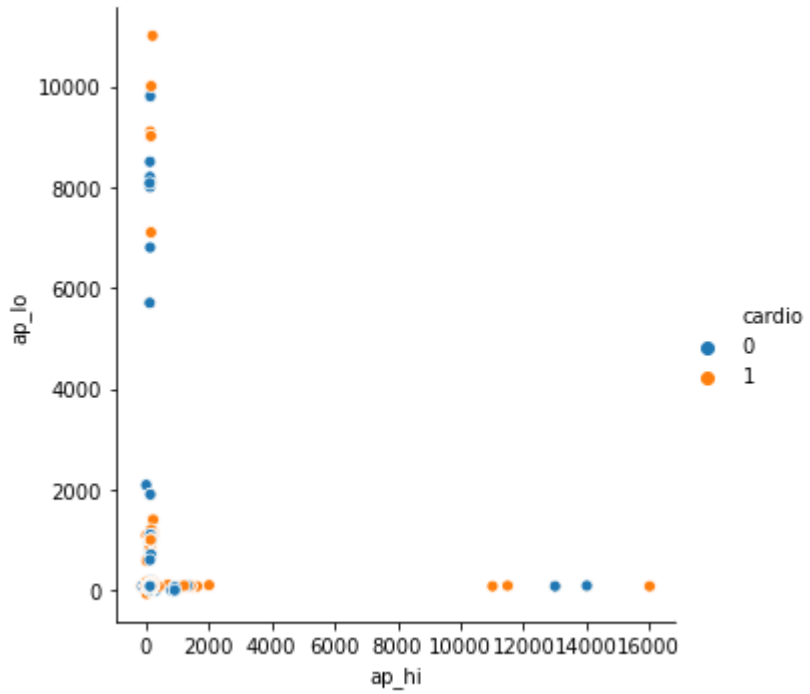


How diastilic and systolic values affect cardio patients?

Draw Boxen plot

In [54]:
```
sns.relplot(x='ap_hi',y='ap_lo',hue='cardio',data=data)
```

Out[54]: <seaborn.axisgrid.FacetGrid at 0xbf8c5b0>



Now, print max and min values and justify.

In [49]:
```
data.ap_hi.max()
```

Out[49]: 16020

In [50]:
```
data.ap_hi.min()
```

Out[50]: -150

In [51]:
```
data.ap_lo.max()
```

Out[51]:
```
11000
```

In [52]:
```
data.ap_lo.min()
```

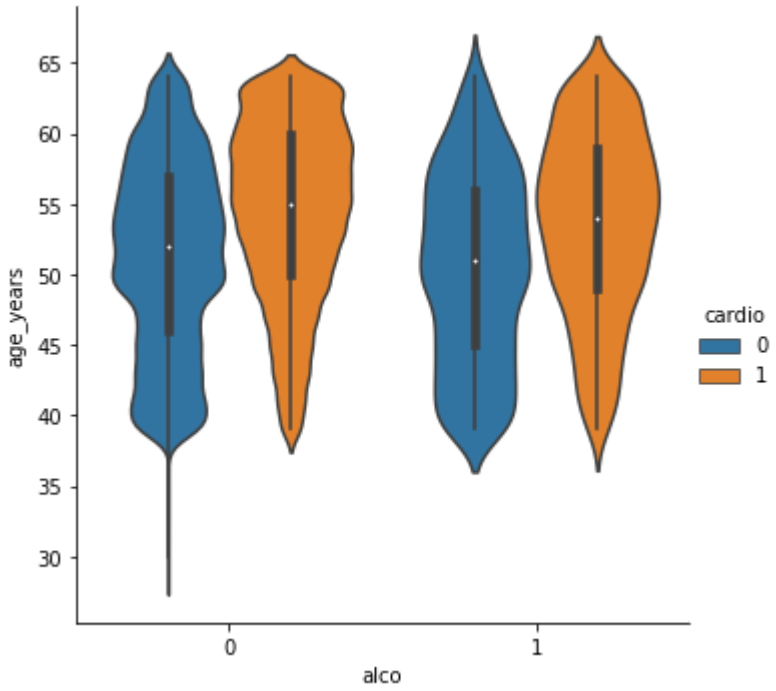
Out[52]:
```
-70
```

How alcohol intake and age affect cardios?

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

In [55]:
```
sns.catplot(x='alco',y='age_years',data=data,hue='cardio',kind='violin')
```

Out[55]:
```
<seaborn.axisgrid.FacetGrid at 0xbcb54c0>
```



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

In [56]:
```
gdp=data.groupby(['alco','cardio'])['age_years']
```

In [57]:
```
pol=gdp.describe()
pol
```

Out[57]:

		count	mean	std	min	25%	50%	75%	max
alco	**cardio**								
0	**0**	33080.0	51.272642	6.781394	29.0	46.0	52.0	57.0	64.0
	1	33156.0	54.500995	6.343918	39.0	50.0	55.0	60.0	64.0
1	**0**	1941.0	50.526018	6.777005	39.0	45.0	51.0	56.0	64.0
	1	1823.0	53.561163	6.478578	39.0	49.0	54.0	59.0	64.0

In [58]:
```
pol.loc[0,0]['50%']
```

Out[58]: 52.0

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

In [59]:
```
pol.loc[0,1]['50%']
```

Out[59]: 55.0

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

In [60]:
```
pol.loc[1,0]['25%']
```

Out[60]: 45.0

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

pol.loc[1,1]['25%']

In [ ]:
```

```