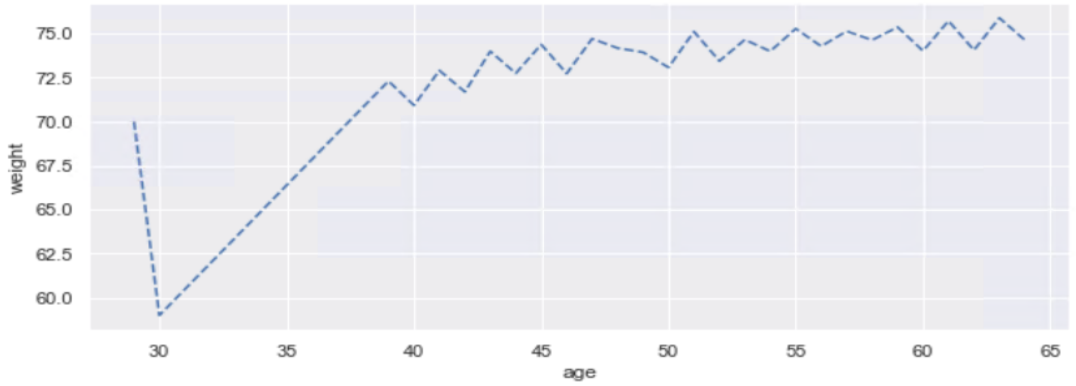
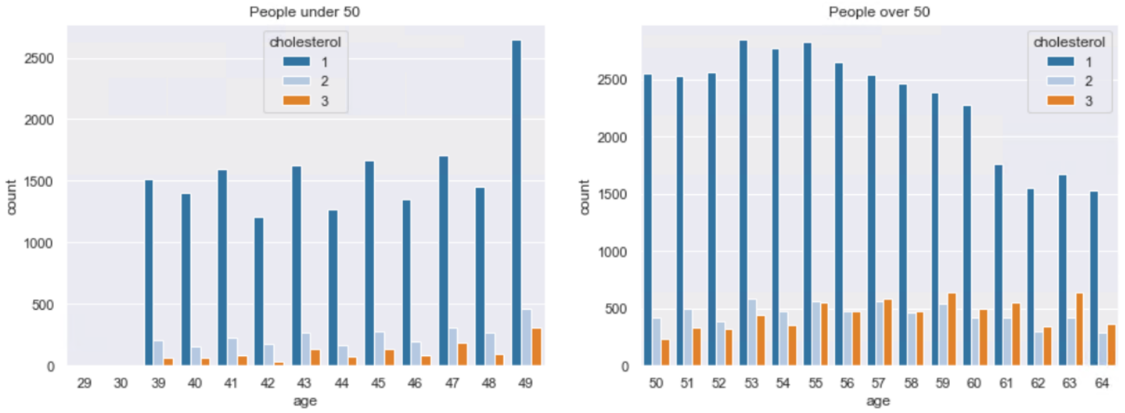
1) Please load "cardio_base.csv". This is a health dataset. Each row represents a person and corresponding attributes like age, height, weight, blood pressure, cholesterol level etc. When asked about age, please calculate with age in years rounded down. How much heavier is the age group with the highest average weight than the age group with the lowest weight?

As we can see on the graphic, the age group with the highest average is 63 years with an average of 75,8 and the lowest weight is 30 years with 59. The difference between these groups is 16,8.

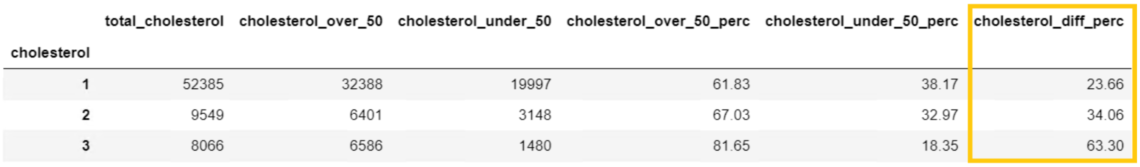


2) Do people over 50 have higher cholesterol levels than the rest? If so, what is the percentage of that difference?



It can be observed that people over 50 have higher cholesterol levels than the rest considering the levels 2 and 3 are higher than 1.

We can summarize level 2 and 3 and classify them as high level of cholesterol. After calculating, the percentage of people over 50 is around **47,45%** and If we want to check the percentage of differences per level of cholesterol, we need to review the following table.

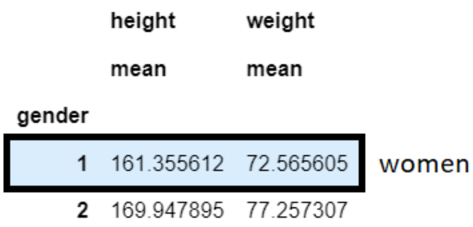


	total_cholesterol	cholesterol_over_50	cholesterol_under_50	cholesterol_over_50_perc	cholesterol_under_50_perc	cholesterol_diff_perc
cholesterol						
1	52385	32388	19997	61.83	38.17	23.66
2	9549	6401	3148	67.03	32.97	34.06
3	8066	6586	1480	81.65	18.35	63.30

% difference per level of cholesterol

3) Are men more likely to be a smoker than women? If so, how many times more? The data contains information to identify gender IDs.

To define which value belongs to every sex, it is necessary to analyze the mean of height and weight between women and men. As well as we assume that men are heavier than women. Considering 1 as women and 2 as men.



	height	weight
	mean	mean
gender		
1	161.355612	72.565605
2	169.947895	77.257307

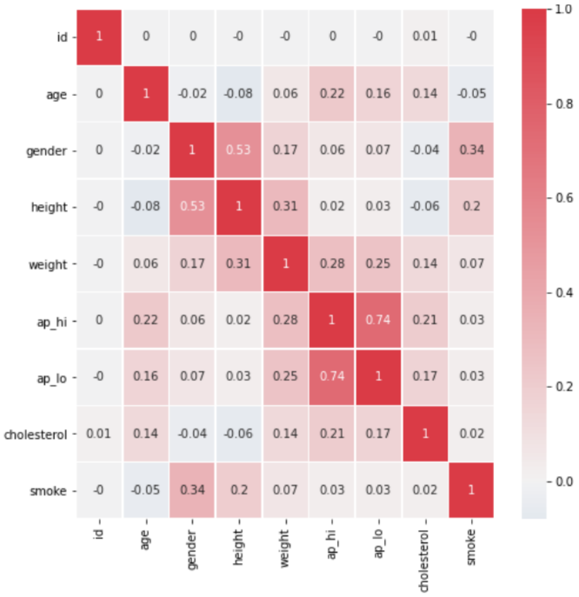
After processing we can see that men are around 6,6 times more smokers than women.

4) How tall are the tallest 1% of the people?

To get how tall is the tallest 1% of people we should get the 99 percentile and max value of height. The 1 % of the tallest people is between **184 and 250**

5) Which two features have the highest spearman rank correlation?

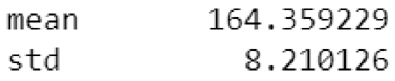
It is important to analyze the correlation between variables using spearman method. As a result we can notice that **ap_lo** and **ap_hi** show the highest correlation.



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	smoke
id	1	0	0	-0	-0	0	-0	0.01	-0
age	0	1	-0.02	-0.08	0.06	0.22	0.16	0.14	-0.05
gender	0	-0.02	1	0.53	0.17	0.06	0.07	-0.04	0.34
height	-0	-0.08	0.53	1	0.31	0.02	0.03	-0.06	0.2
weight	-0	0.06	0.17	0.31	1	0.28	0.25	0.14	0.07
ap_hi	0	0.22	0.06	0.02	0.28	1	0.74	0.21	0.03
ap_lo	-0	0.16	0.07	0.03	0.25	0.74	1	0.17	0.03
cholesterol	0.01	0.14	-0.04	-0.06	0.14	0.21	0.17	1	0.02
smoke	-0	-0.05	0.34	0.2	0.07	0.03	0.03	0.02	1

6) What percentage of people are more than 2 standard deviations far from the average height?

The average of people height is 164.35 and 2 standard deviations is 16,4. The height 2 std far from avg is 180,75. We need to calculate the percentage of people who has over 180,75. Using SciPy we can get this information. The percentage 2std far is around 2,27 % of people.



```
mean        164.359229
std           8.210126
```

7) What percentage of the population over 50 years old consume alcohol? Also use the cardio_alco.csv and merge the datasets on ID. Ignore those persons, where we have no alcohol consumption information.

The percentage of the population over 50 years old that consume alcohol is around 3,99 % and 5% when you ignore people with no alcohol information over 50 years old.

8) Which of the following statements is true with 95% confidence?

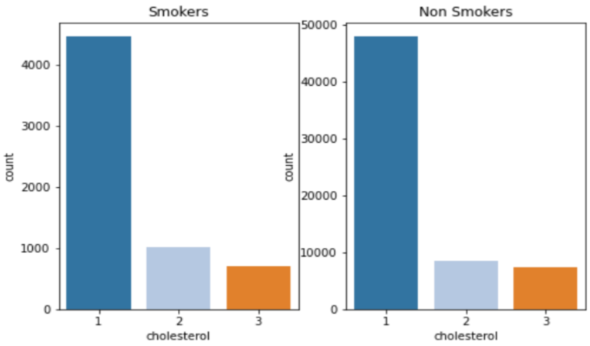
We can use Hypothesis Testing with 95% of confidence to validate each statement

a) Smokers have higher blood pressure than non-smokers (False)

I used the function ttest_ind from scipy to apply the hypothesis. As a result, the null hypothesis is accepted. It means that smokers do not have higher blood pressure than non-smokers.

b) Smokers have higher cholesterol level than non-smokers (True)

These variables are categorical, and we need to compare the level of cholesterol per smokers and non-smokers. The following graphic shows that non-smokers have fewer levels (taking level 2 and 3 as high levels) of cholesterol. We consider accept the affirmation, but it is important to check the hypothesis using chi square



As a result the alternative hypothesis is accepted, so smokers have higher cholesterol level than non-smokers

c) Smokers weight less than non-smokers (False)

The alternative hypothesis was rejected, smokers do not weight less than non-smokers

d)Men have higher blood pressure than women (True)

As a result, the alternative hypothesis was accepted. Men have higher blood pressure than women.

9) Second Dataset, Covid19 cases. This dataset contains daily covidl9 cases for all countries in the world. Each row represents a calendar day. The rows also contain some simple information about the countries, like population, percentage of the population over 65, GDP and hospital beds per thousand inhabitants. Please use this dataset to answer the following questions.

When did the difference in the total number of confirmed cases between Italy and Germany become more than 10 000?

The difference in the total number of confirmed cases between Italy and Germany becme more than 10 000 at 2020-03-12.

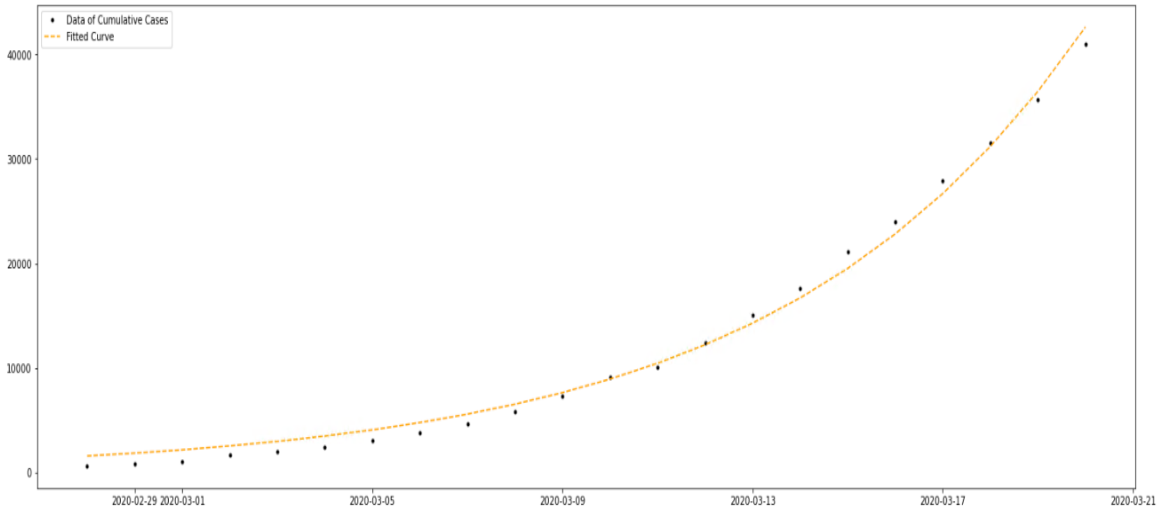
10) Look at the cumulative number of confirmed cases in Italy between 2020-02-28 and 2020-03-20. Fit an exponential function (y = Ae^(Bx)) to this set to express cumulative cases as a function of days passed, by minimizing squared loss. What is the difference between the exponential curve and the total number of real cases on 2020-03-20?

Fit Exponential curve

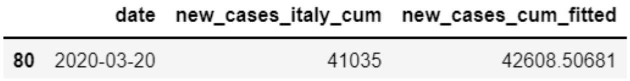
To fit the cumulative cases in function of days passed using exponential formula to achieve the task, it is necessary following the next steps:

- define the independent and dependent variables, in this case x and y would be the total of days and the cumulative cases respectably
- implement a function that represents exponential formula
- apply curvefit function from scipy to fit data, this function uses leastsq

The graphic after fitting the date, represents real data points and fitted curve based on dates



After that we get the information about fitted points that were calculated using the coefficients b and a. The following table shows the difference between real point and fit point.



	date	new_cases_italy_cum	new_cases_cum_fitted
80	2020-03-20	41035	42608.50681

The difference between the exponential curve and the total number of real cases on 2020-03-20 is 1573.51

11) Which country has the 3rd highest death rate? Death rate: total number of deaths per million inhabitants

After calculating the death rate per country, Andorra has the 3rd highest death rate.

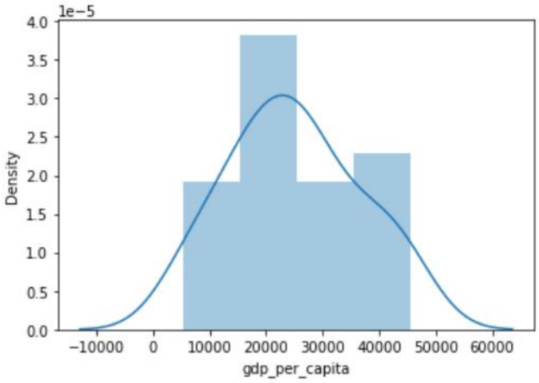
12) What is the F1 score of the following statement: Countries, where more than 20% of the population is over 65 years old, have death rates over 50 per million inhabitants. Ignore countries, where any of the necessary information is missing!

This question is not clear enough, F1 is a popular performance measure for classification, combines the precision and recall of classification models.

13) What is the probability that a country has GDP over $10 000, if we know that they have at least 5 hospital beds per 1000 inhabitants.

we need to calculate P(X>10000) per country that have at least 5 hospitals per 10000 inhab.

The distribution of gdp_per_capita



It is important calculate z score using mean and std, after that I used scipy.stats.norm.cdf to get the probability. The result was 90.047 %