Medical Cost In USA

Get dataset

In [3]:
```
df=pd.read_csv('insurance.csv')
print_df(df.head())
```

Out[3]:

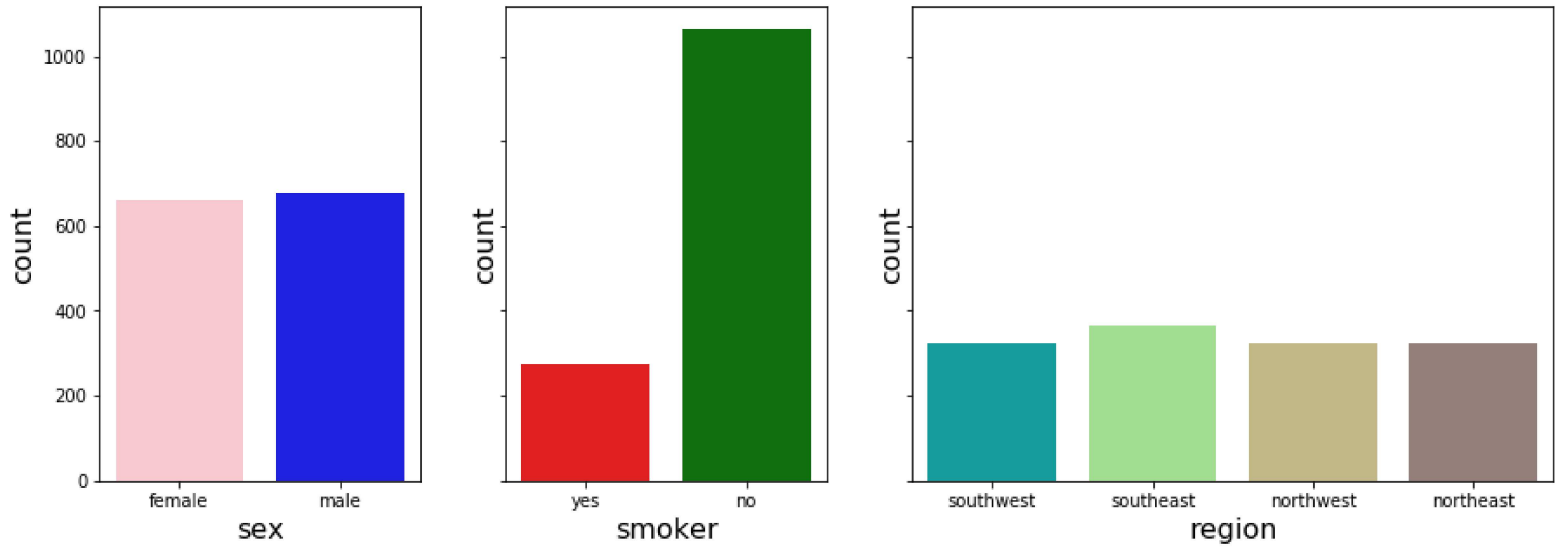
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.9	0	yes	southwest	16884.9
1	18	male	33.77	1	no	southeast	1725.55
2	28	male	33	3	no	southeast	4449.46
3	33	male	22.705	0	no	northwest	21984.5
4	32	male	28.88	0	no	northwest	3866.86

- **bmi**: Body Mass Index (kg/m^2)
- **charges**: Individual medical costs billed by health insurance
- **1338 rows**

Exploring the data

In [6]:
```
f
```

Out[6]:



From this, it seems that the dataset sample is representative of the US population:

- male and female equally represented
- 4 regions equally represented
- a majority of non smokers

Exploring the data

In [7]:
```
print_df(round(df.describe(),1))
```

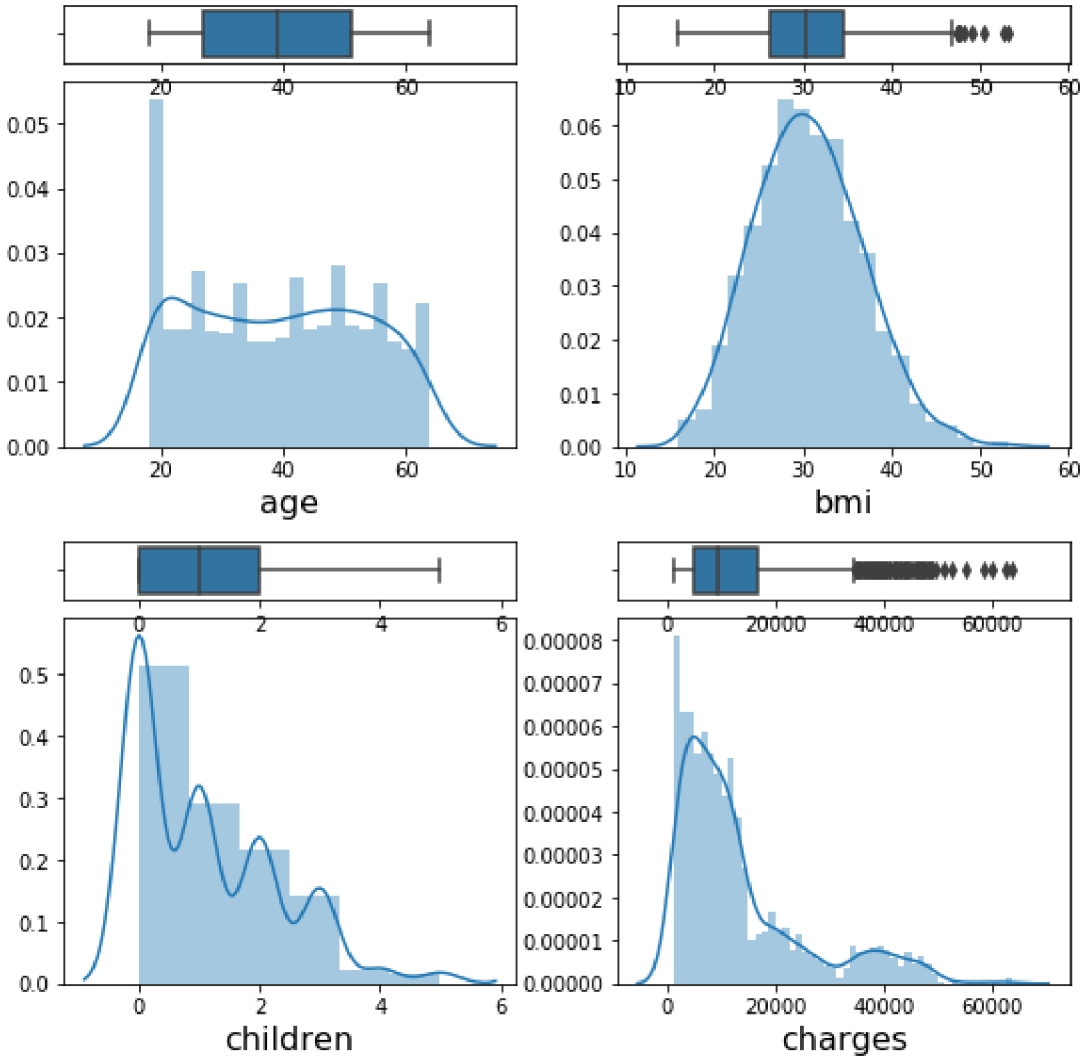
Out[7]:

	age	bmi	children	charges
count	1338	1338	1338	1338
mean	39.2	30.7	1.1	13270.4
std	14	6.1	1.2	12110
min	18	16	0	1121.9
25%	27	26.3	0	4740.3
50%	39	30.4	1	9382
75%	51	34.7	2	16639.9
max	64	53.1	5	63770.4

- For age, bmi and number of children : $mean \simeq median$
- For charges : $mean > median$

In [12]:
```
fig
```

Out[12]:

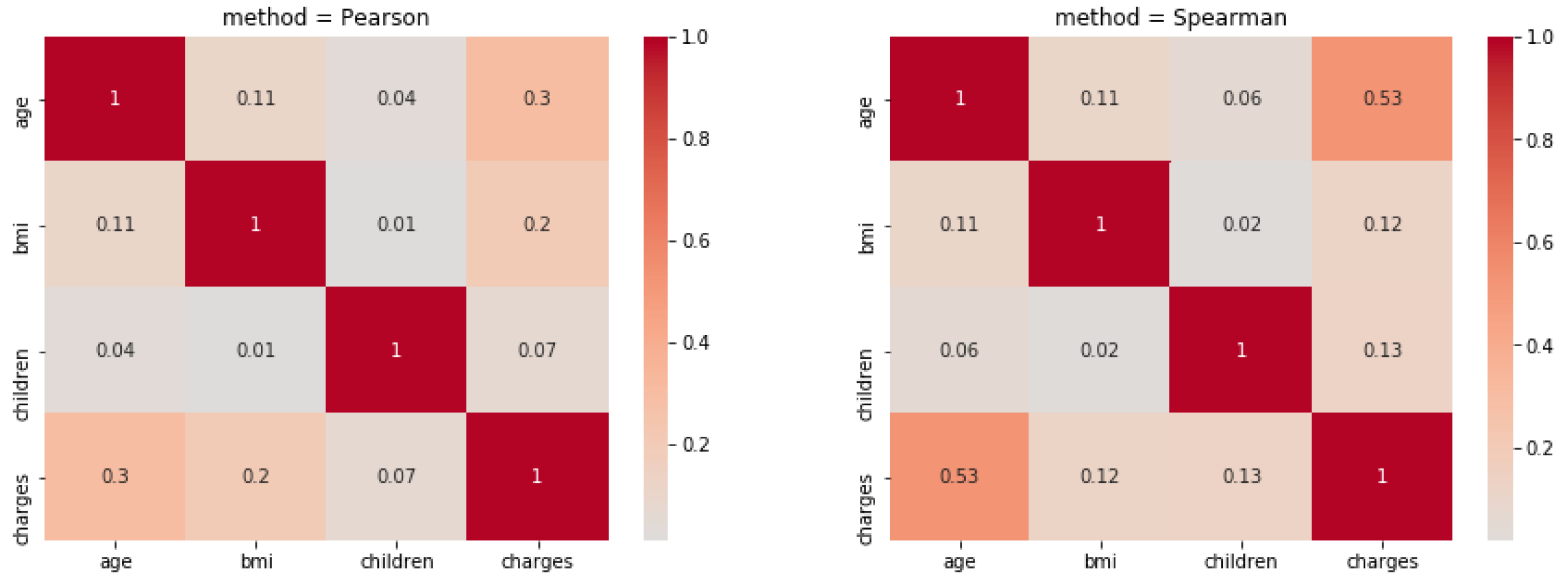


- For charges : $mean > median \Rightarrow$ a lot of outliers (high charges)

Are there any correlations between numerical columns ?

In [96]:
```
f2
```

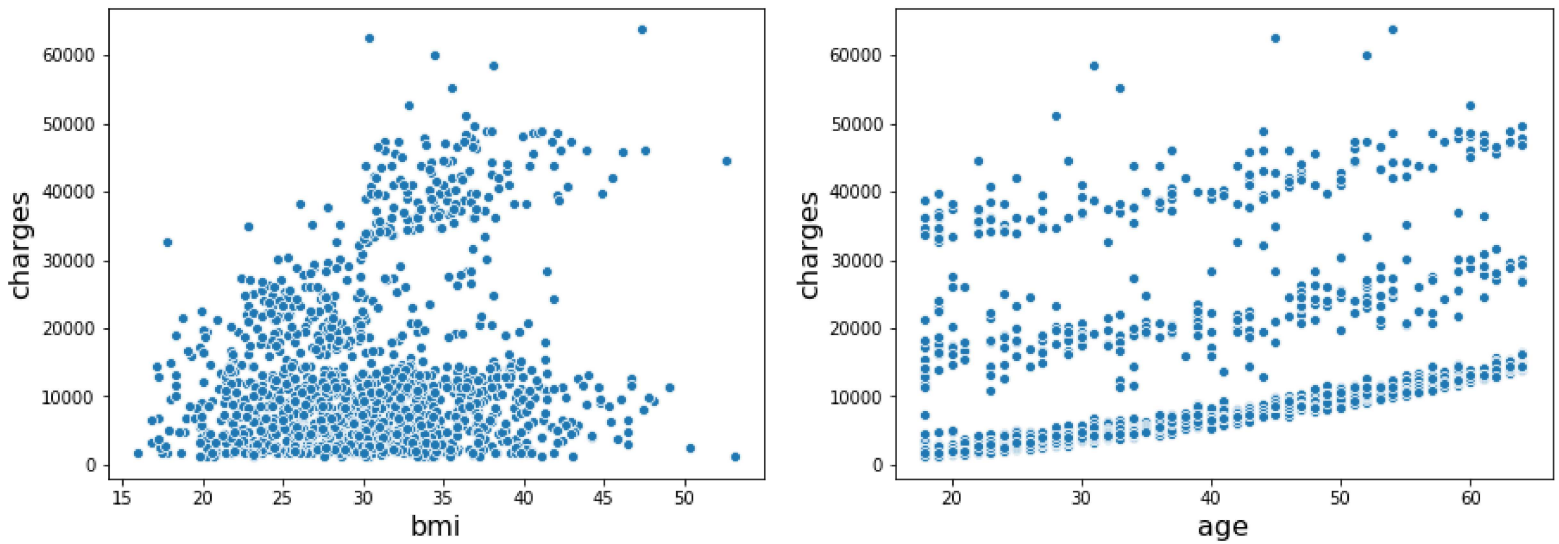
Out[96]:



- Charges correlated with age ?
- Charges correlated with BMI ?

In [98]:
```
f3
```

Out[98]:



- In the charges=f(bmi) plot, there seem to be 2 groups on points (one with low charges, one with high charges)
- In the charges=f(age) plot, there seem to be 3 groups of points forming 3 lines with a positive slope

Scraping BMI intervals depending on gender and age

ncbi.nlm.nih.gov/pmc/articles/PMC4681110/

younger ages (Figure 1).

Table 1.

Apparent optimal[a] and acceptable[a] ranges of body mass index for longevity

	Men			Women	
Age (years)	Optimal BMI[b]	Acceptable BMI[b]	Age (years)	Optimal BMI[b]	Acceptable BMI[b]
18–34	23.0–25.9	21.0–28.9	18–34	15.5–24.9[c]	15.5–25.9[c]
35–44	23.0–26.9	22.0–28.9	35–44	19.0–23.9	17.5–25.9
45–54	24.0–27.9	23.0–28.9	45–49	20.0–25.9	19.0–26.9
			50–54	22.0–26.9	21.0–27.9
55–64	24.0–28.9	23.0–31.4	55–64	23.0–27.9	22.0–29.9
65–74	25.0–28.9	23.0–31.4	65–74	24.0–28.9	22.0–31.4
75–99	25.0–32.9	24.0–34.9	75–99	24.0–29.9	22.0–36.4[d]

Int J Epidemiol

In [39]:
```
print_df(BMIs)
```

Out[39]:

	age	male.optimal	male.acceptable	female.optimal	female.acceptable
0	[18.0, 34.0]	[23.0, 25.9]	[21.0, 28.9]	[15.5, 24.9]	[15.5, 25.9]
1	[35.0, 44.0]	[23.0, 26.9]	[22.0, 28.9]	[19.0, 23.9]	[17.5, 25.9]
2	[45.0, 49.0]	[24.0, 27.9]	[23.0, 28.9]	[20.0, 25.9]	[19.0, 26.9]
3	[50.0, 54.0]	[24.0, 27.9]	[23.0, 28.9]	[22.0, 26.9]	[21.0, 27.9]
4	[55.0, 64.0]	[24.0, 28.9]	[23.0, 31.4]	[23.0, 27.9]	[22.0, 29.9]
5	[65.0, 74.0]	[25.0, 28.9]	[23.0, 31.4]	[24.0, 28.9]	[22.0, 31.4]
6	[75.0, 99.0]	[25.0, 32.9]	[24.0, 34.9]	[24.0, 29.9]	[22.0, 36.4]

adding a column that gives BMI range

(too low, optimal, acceptable, too high)

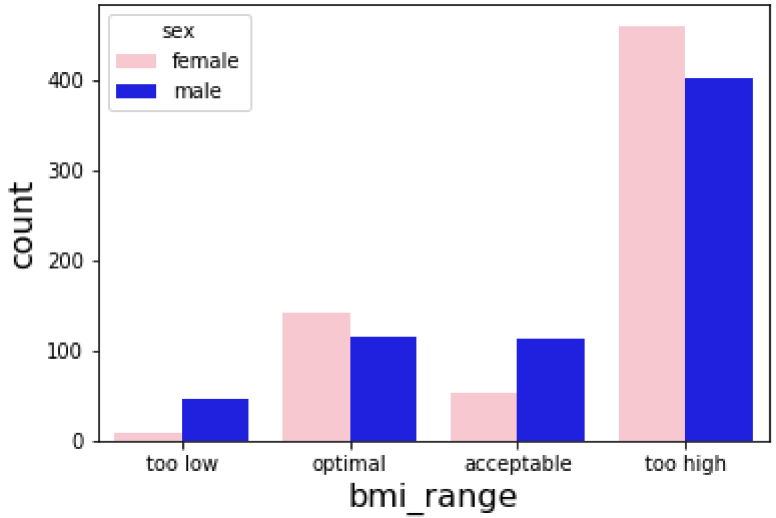
In [45]:
```
print_df(df.sample(10))
```

Out[45]:

	age	sex	bmi	children	smoker	region	charges	bmi_range
965	35	male	27.1	1	no	southwest	4746.34	acceptable
1153	35	female	35.815	1	no	northwest	5630.46	too high
426	38	female	27.265	1	no	northeast	6555.07	too high
198	51	female	18.05	0	no	northwest	9644.25	too low
1160	43	female	34.58	1	no	northwest	7727.25	too high
709	36	female	27.74	0	no	northeast	5469.01	too high
76	29	female	29.59	1	no	southeast	3947.41	too high
347	46	male	33.345	1	no	northeast	8334.46	too high
724	50	female	27.075	1	no	northeast	10106.1	acceptable
1196	19	female	30.02	0	yes	northwest	33307.6	too high

In [47]:
```
fig2.figure
```

Out[47]:



- A majority of the people in the sample have a too high BMI!

Influence of smoking

In [50]:
```
print_df(round(df.groupby(df.smoker).mean(),1))
```

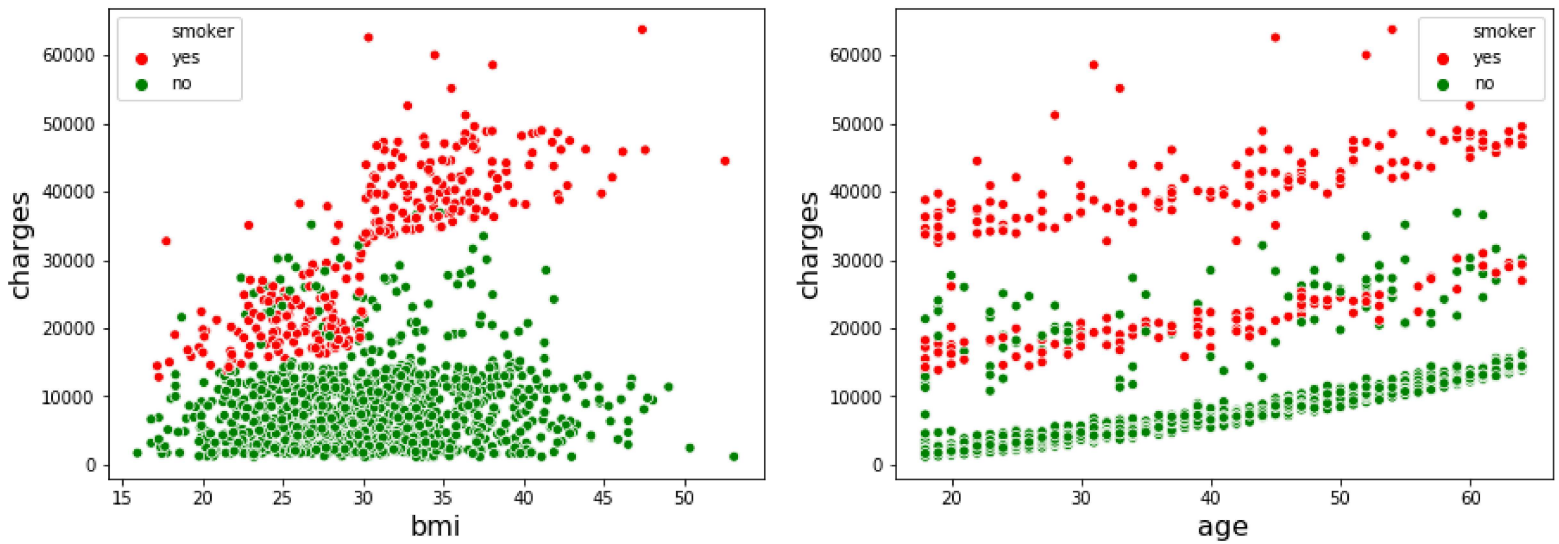
Out[50]:

smoker	age	bmi	children	charges
no	39.4	30.7	1.1	8434.3
yes	38.5	30.7	1.1	32050.2

- Smokers are charged much more: smoking has negative effect on health!

In [54]:
```
fig3
```

Out[54]:

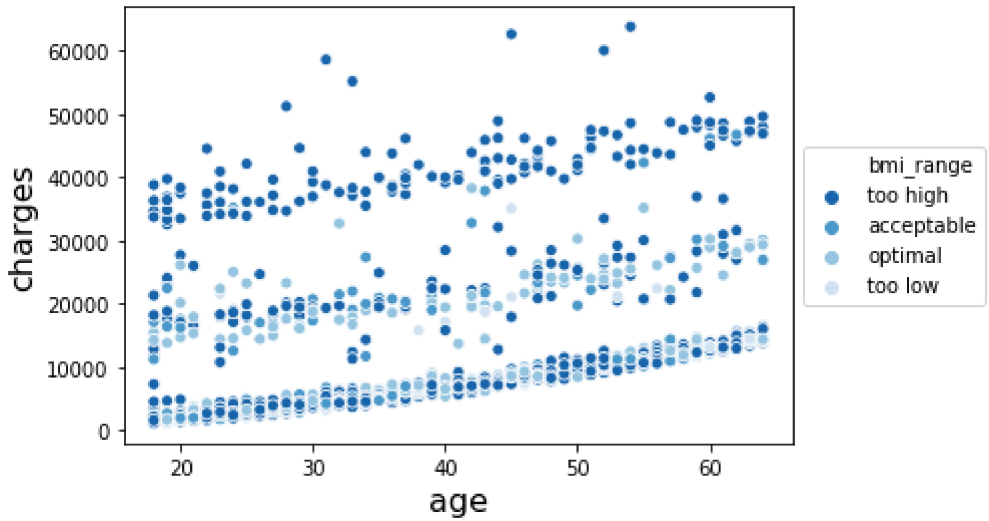


- In the charges=f(bmi) plot, the 2 groups on points (one with low charges, one with high charges) can be explained by smokers/non smokers
- In the charges=f(age) plot, the 3 groups of points can be *partially* explained by smokers/non smokers

Influence of BMI range ?

In [62]:
```
fig4.figure
```

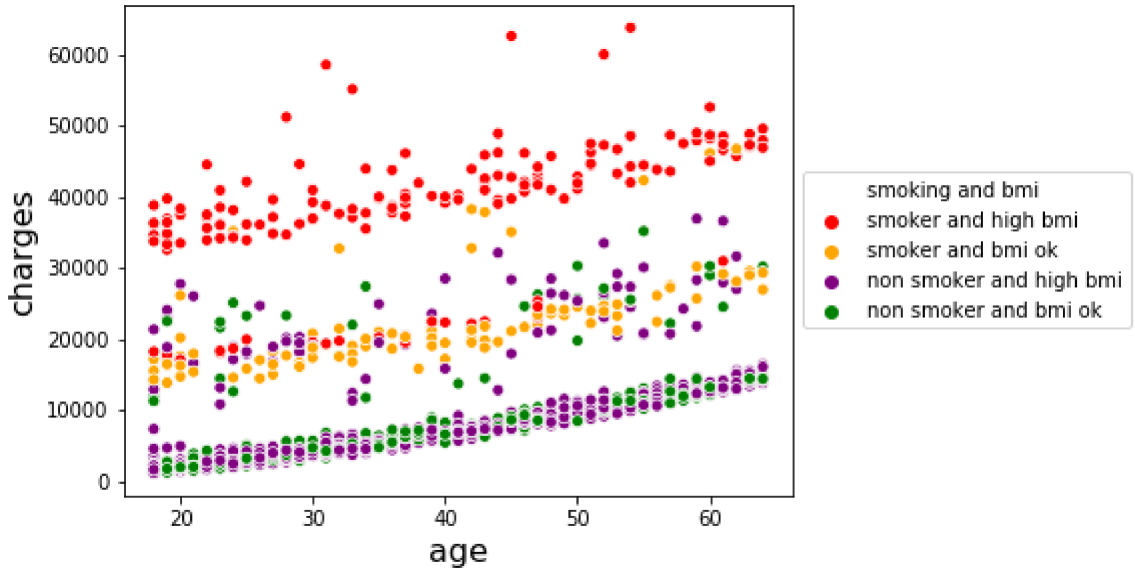
Out[62]:



Influence of smoking AND BMI

In [87]:
```
fig5
```

Out[87]:



In the charges=f(age) plot, the 3 groups of points can be explained by:

- low charges: non smokers (independant of BMI)
- medium charges: smokers with too low, optimal or acceptable BMI
- high charges: smokers with too high BMI