Завдання 1: Завантажте дані `medical-charges.csv` в пандас датафрейм і виведіть перші 5 записів. Напишіть, як ви можете підійти до вирішення задачі прогнозування колонки `charges` на основі інших колонок виходячи з наявних на даний момент знань (без ML методів, чисто з використанням аналітики). Запишіть 3 або більше ідей, які приходять вам на думку нижче:

```
import pandas as pd
```

```
medical_df=pd.read_csv('medical-charges.csv')
```

```
medical_df.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Далее: Создать код с переменной `medical_df` New interactive sheet

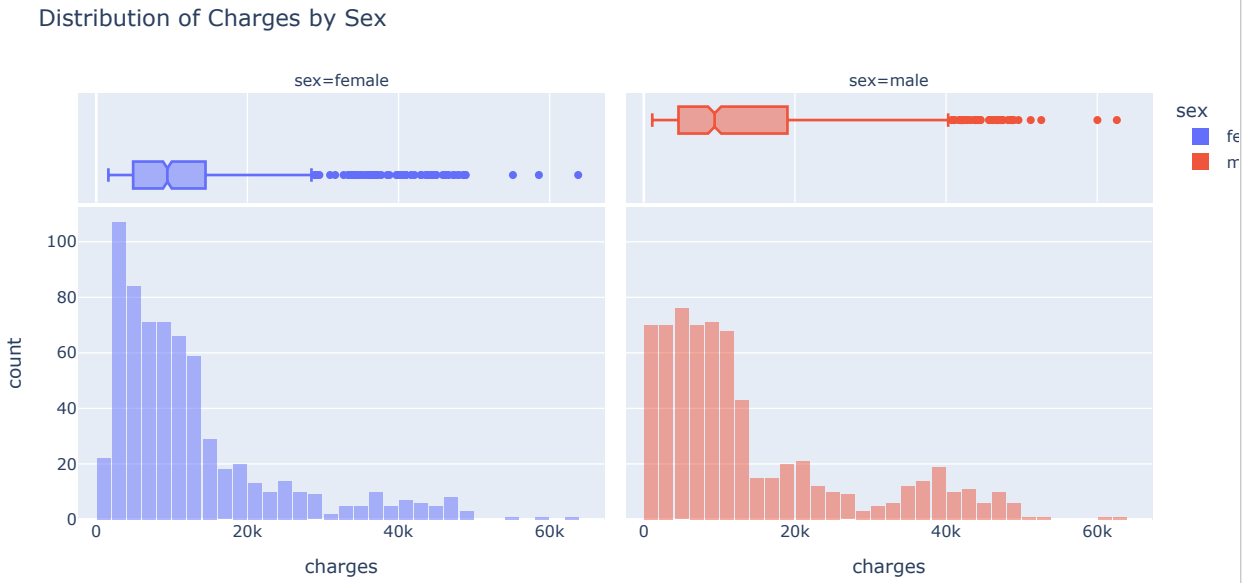
Завдання 2: Візуалізуйте розподіл медичних зборів (`charges`) у вигляді інтерактивної гістограми plotly з розбиттями за категоріями ознак

1. `sex`
2. `region`

Додайте маржинальний графік у вигляді бокс-плота вгорі по дикретним категоріям ознак. Скористайтесь прикладом візуалізації з лекції. Опишіть свої спостереження.

```
import matplotlib.pyplot as plt
import plotly.express as px
```

```
fig = px.histogram(medical_df,
                   x='charges',
                   marginal='box',
                   color='sex',
                   facet_col='sex',
                   barmode="overlay",
                   title='Distribution of Charges by Sex')
fig.update_layout(bargap=0.1, width=1000,height=500)
fig.show()
```



Висновок 2.1

Чоловіки та жінки мають схожі розподіли, майже однакові qv25 і median (4.5K та 9,5K відповідно). А от qv75 і далі (правий хвіст) більші у чоловіків. Багато outliers.

Sex сам по собі слабкий предиктор.

```
fig = px.histogram(medical_df,
                   x='charges',
                   marginal='box',
                   color='region',
                   barmode="overlay",
                   title='Distribution of Sex')
fig.update_layout(width=900,height=450, bargap=0.1)
fig.show()
```


```
fig = px.histogram(medical_df,
                   x='charges',
                   marginal='box',
                   color='region',
                   facet_col='region',
                   title='Distribution of Sex')
fig.update_layout(bargap=0.1, width=900,height=450)
fig.show()
```

Висновок 2.2

Розподіли схожі: qv25, median майже однакові. southeast має найбільший правий хвіст. Нема доказів, що region має сильний предиктовий фактор.

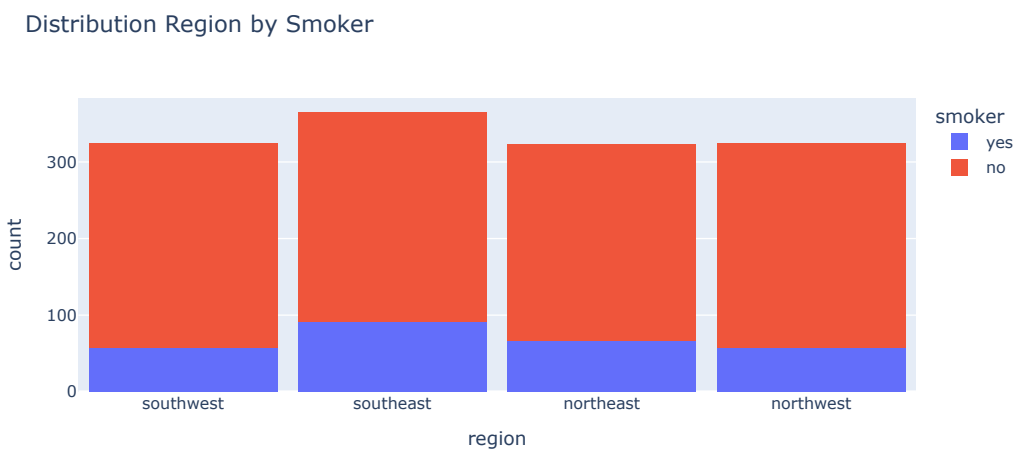
Завдання 3: Візуалізуйте з `plotly` розподіл кожного з наступних стовпців відносно того, чи є людина курцем (`smoker`)

- `region`
- `children`

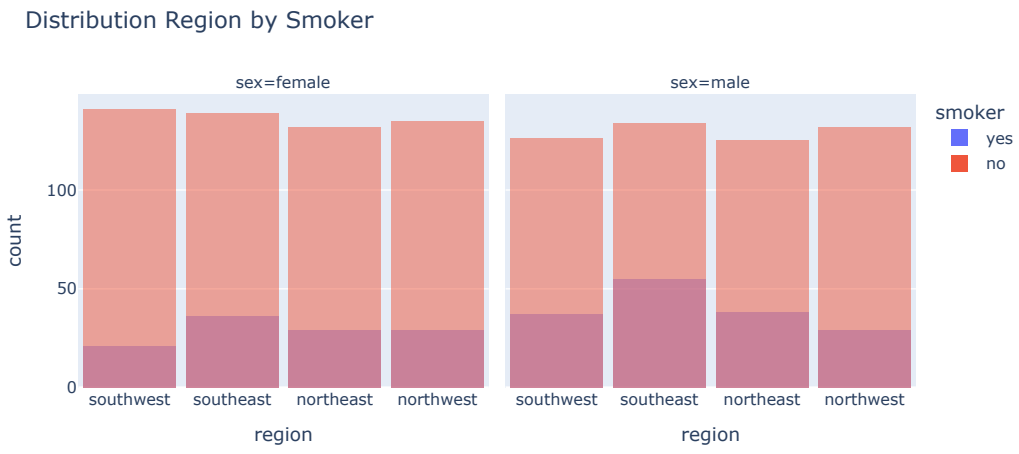
та опишіть коротко свої спостереження.

3.1 Distribution Region Count by Smoker

```
fig = px.histogram(medical_df,
                   x='region',
                   color='smoker',
                   title='Distribution Region by Smoker')
fig.update_layout(bargap=0.1, width=800, height=400)
fig.show()
```




```
fig = px.histogram(medical_df,
                   x='region',
                   color='smoker',
                   facet_col='sex',
                   barmode='overlay',
                   title='Distribution Region by Smoker')
fig.update_layout(bargap=0.1, width=800, height=400)
fig.show()
```

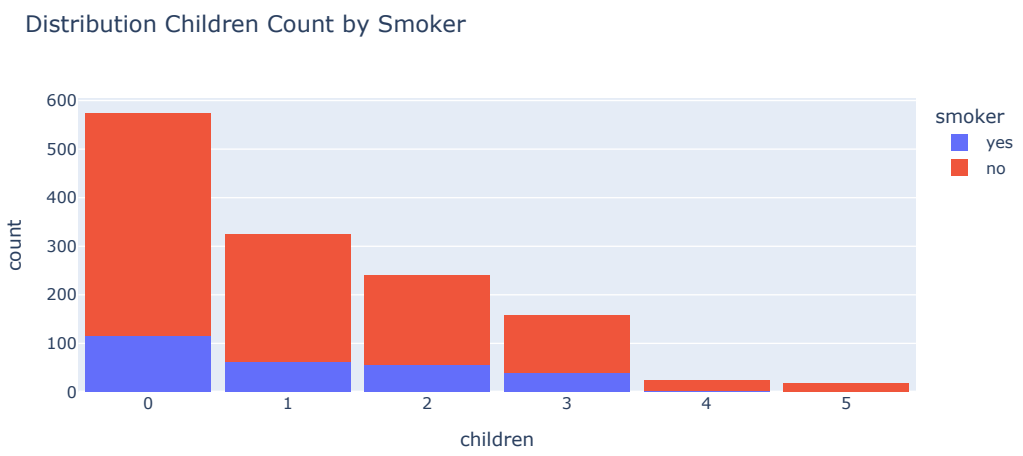


Висновок 3.1

Серед southeast більш високий %курців, в чоловіків більший, аніж серед жінок (можливо через те, вони і мали більший правий хвіст в обидвох попередніх зрізах).

3.2 Distribution Children Count by Smoker

```
fig = px.histogram(medical_df,
                   x='children',
                   color='smoker',
                   title='Distribution Children Count by Smoker'
                   )
fig.update_layout(bargap=0.1, width=800, height=400)
fig.show()
```

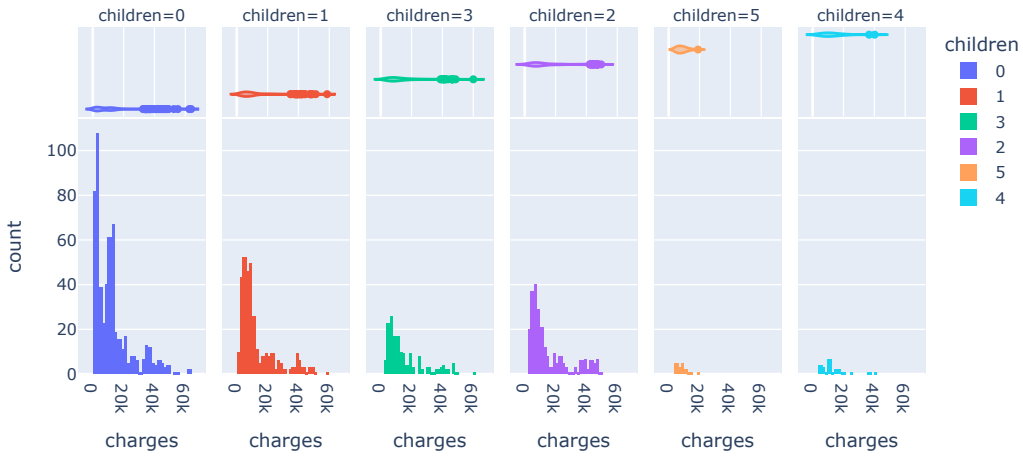


Висновок 3.2

% курців однаковий серед людей з різною кількістю дітей

Завдання 4: Візуалізуйте зв'язок між стовпцем `charges` та `children` використовуючи графіки-скрипки (`px.violin`). Опишіть свої спостереження.

```
fig=px.histogram(medical_df,
                 x='charges',
                 color='children',
                 marginal='violin',
                 facet_col='children'
                 )
fig.update_layout(width=800, height=400)
fig.show()
```




```
fig=px.scatter(medical_df,
                 color='children',
                 x='charges',
                 opacity=0.8,
                 title='Children Count vs. Charges')
fig.update_layout(width=800, height=400)
```

Висновок 4

Усі розподіли є правозмішені та мають довгі праві хвости. Високі charges зустрічаються як серед людей без дітей, так і в інших випадках. Немає ознак лінійної залежності

Завдання 5. Розглянемо модель для користувачів, які не є курцями (`no_smoker_df`):

$$charges = w \times age + b$$

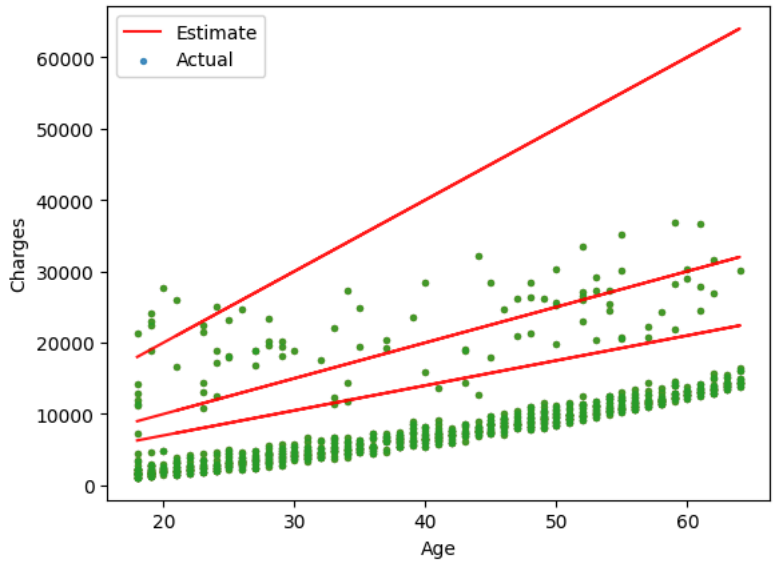
Спробуйте 3 різні пари параметрів `w` та `b` аби вручну підігнати лінію під дані використовуючи наведені допоміжні функції `estimate_charges` та `try_parameters`. Опишіть спостереження. Пари параметрів мають бути НЕ такі, як були наведені в лекції.

```
non_smoker_df = medical_df[medical_df.smoker == 'no']
```

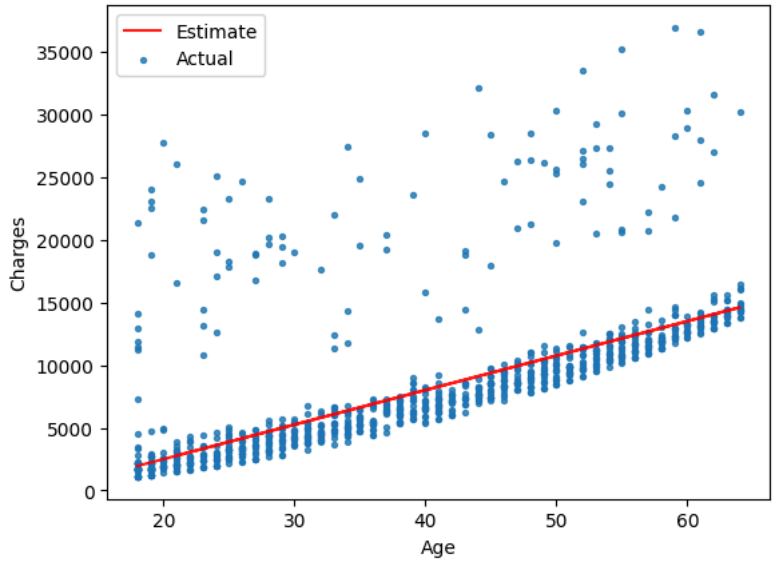
```
def estimate_charges(age, w, b):
    return w * age + b
```

```
def try_parameters(df, w, b):
    ages = df.age
    target = df.charges

    estimated_charges = estimate_charges(ages, w, b)

    plt.plot(ages, estimated_charges, 'r', alpha=0.9);
    plt.scatter(ages, target, s=8,alpha=0.8);
    plt.xlabel('Age');
    plt.ylabel('Charges')
    plt.legend(['Estimate', 'Actual']);
```

```
try_parameters (non_smoker_df, 1000, 0)
try_parameters (non_smoker_df, 500, 0)
try_parameters (non_smoker_df, 350, 0) #найкращий коєфіцієнт
```




```
# try_parameters (non_smoker_df, 350, -5000)
# try_parameters (non_smoker_df, 350, -3000)
# try_parameters (non_smoker_df, 300, -3000)
# try_parameters (non_smoker_df, 265, -3000)
try_parameters (non_smoker_df, 275, -3000)
```



- y=w*x+b
- x=20 y~=2500 w=275
- 2500=275*20+b
- b=2500-275*20
- b~=-3000

Завдання 6: Напишіть фукнцію для обчислення root mean squared error згідно з формулою цієї метрики точності моделі з використанням `numpy`.

Обчисліть RMSE для тих пар пареметрів, які Ви спробували в завданні 5.

Яке найнижче значення втрат ви зможете досягти? Чи можете ви придумати загальну стратегію для знаходження кращих значень w та b методом проб та помилок?

```
import numpy as np
```

```
def rmse(targets, predicted):
    return np.sqrt(np.mean(np.square(targets - predicted)))
```

```
w=275
b=-3000
```

```
targets = non_smoker_df['charges']
predicted = estimate_charges(non_smoker_df.age, w, b)
```

```
rmse(targets, predicted)
```

```
np.float64(4702.6414178153345)
```

Висновок 6

Мені подобається, як червона лінія описує дані, опосередовані внизу. Хоча я розумію, що "виброси" вона не описує зовсім. Ці значення і дають великі відхилення. Припускаю, що має бути ще параметр виду c*z, де c буде приймати значення 0 та 1.