TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

-----o0o-----

BÁO CÁO LAB 01
NHẬP MÔN HỌC MÁY

Giáo viên hướng dẫn:

Nguyễn Tiến Huy

I. Tổng quan

1. Thông tin nhóm, phân công công việc.

STT	Họ và tên	MSSV	Công việc
1	Trần Ngọc Tịnh	18120597	Tiền xử lý
2	Nguyễn Ngọc Năng Toàn	18120600	Thực hiện các thuật toán học máy cơ bản
3	Nguyễn Đức Trực	18120621	Thực hiện các thuật toán học máy nâng cao
4	Trần Luật Vy	18120656	Trực quan dữ liệu + insight

2. Mức độ hoàn thành tổng thể mỗi yêu cầu.

STT	Yêu cầu	Đánh giá
1	Trực quan dữ liệu + insight	100%
2	Tiền xử lý	100%
3	Thực hiện các thuật toán học máy cơ bản	100%
4	Thực hiện các thuật toán học máy nâng cao	100%

II. Khám phá dữ liệu:

❖ Tổng thể về dữ liệu:

	age	sex	bmi	children	smoker	region	charges
0	24	male	23.655	0	no	northwest	2352.96845
1	28	female	26.510	2	no	southeast	4340.44090
2	51	male	39.700	1	no	southwest	9391.34600
3	47	male	36.080	1	yes	southeast	42211.13820
4	46	female	28.900	2	no	southwest	8823.27900
5	63	female	26.220	0	no	northwest	14256.19280
6	38	female	19.950	2	no	northeast	7133.90250
7	28	female	26.315	3	no	northwest	5312.16985
8	25	male	26.800	3	no	southwest	3906.12700
9	18	female	30.115	0	no	northeast	2203.47185

❖ Kiểu dữ liệu thuộc tính:

```
data_df.dtypes
```

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

❖ Dữ liệu có thiếu không:

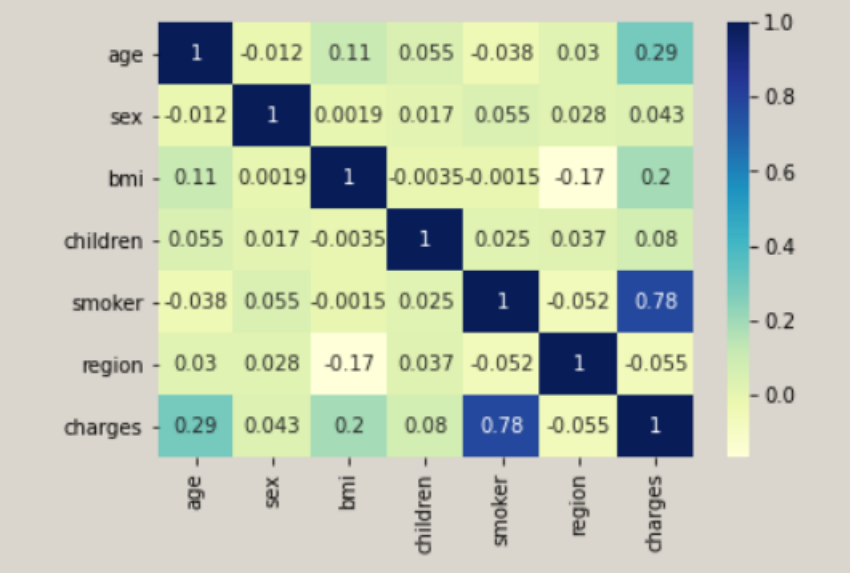
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

❖ Thống kê thuộc tính:

	age	bmi	children	charges
count	1003.000000	1003.000000	1003.000000	1003.000000
mean	39.255234	30.511780	1.104686	13267.935817
std	14.039105	6.013107	1.204619	12051.356547
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.210000	0.000000	4780.839400
50%	39.000000	30.200000	1.000000	9447.382400
75%	51.000000	34.430000	2.000000	16840.667970
max	64.000000	53.130000	5.000000	62592.873090

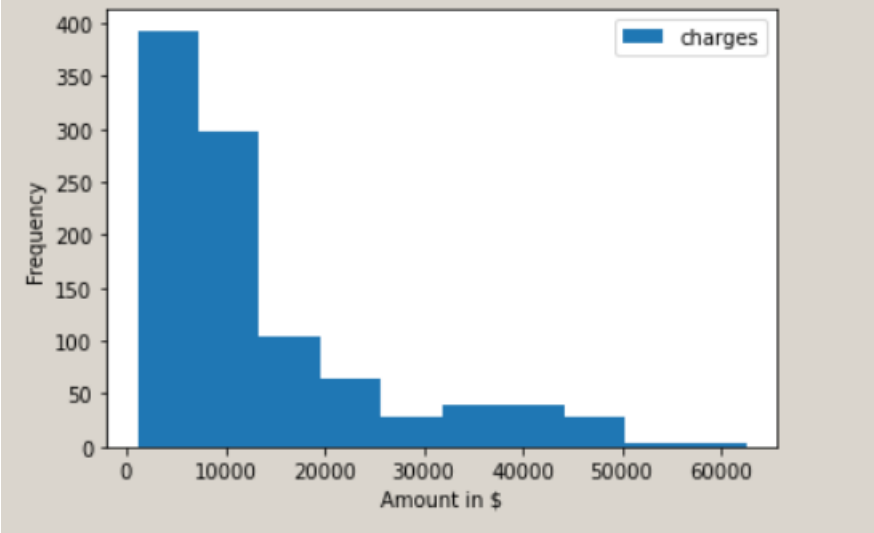
III. Trực quan và trình bày các thông tin hữu ích

❖ Heatmap thể hiện sự tương quan giữa các thuộc tính:



❖ **Tiến hành trực quan các trường và tìm ra sự tương quan giữa chúng với nhau**:

1. Sự phân bố của charges như thế nào?

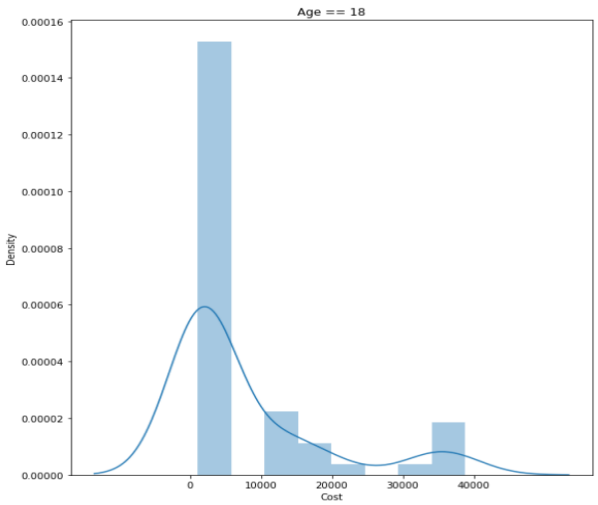


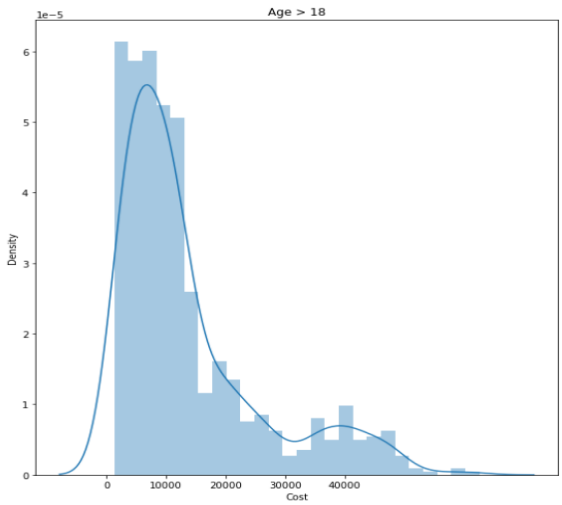
❖ **Từ đồ thị ta có thể nhận thấy phần lớn chi phí y tế rơi vào khoảng 1000 - 13500 dollar**

2. Sự tương quan giữa các thuộc tính

a. Age có ảnh hưởng như thế nào đến sự phân bố của Charges?

- Min(age) = 18
- Max(age) = 64
- **Dự đoán:** ở độ tuổi 18 đang là độ tuổi phát triển và khỏe mạnh nhất thế chi phí y tế bỏ ra cũng có thể ít.
- Ta sẽ xem xét sự phân phối của **Charges** với **age = 18** và **age > 18** sẽ như thế nào?

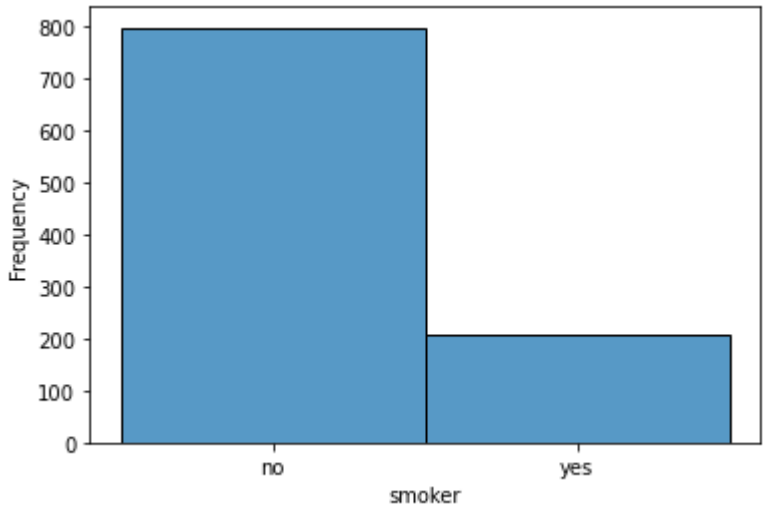




- ❖ Với age = 18 thì phân phối của charges có giá trị trung bình khoảng 3000.
- ❖ Với age > 18 thì phân phối của charges có giá trị trung bình khoảng 8500.

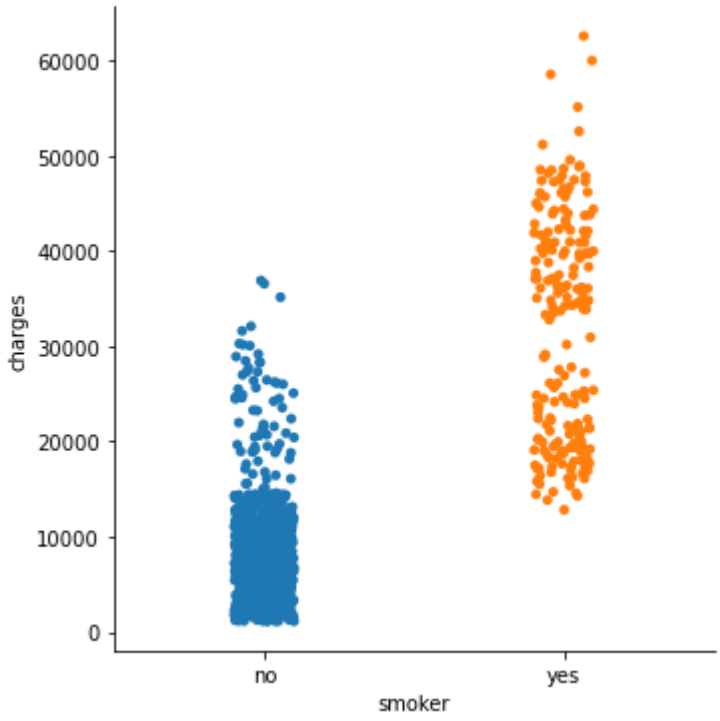
b. Giữa smoker với Charges có ảnh hưởng với nhau như thế nào?

- Số lượng người hút thuốc và không hút thuốc là bao nhiêu?



- ❖ Số lượng người hút thuốc: 797
- ❖ Số lượng người không hút thuốc: 206

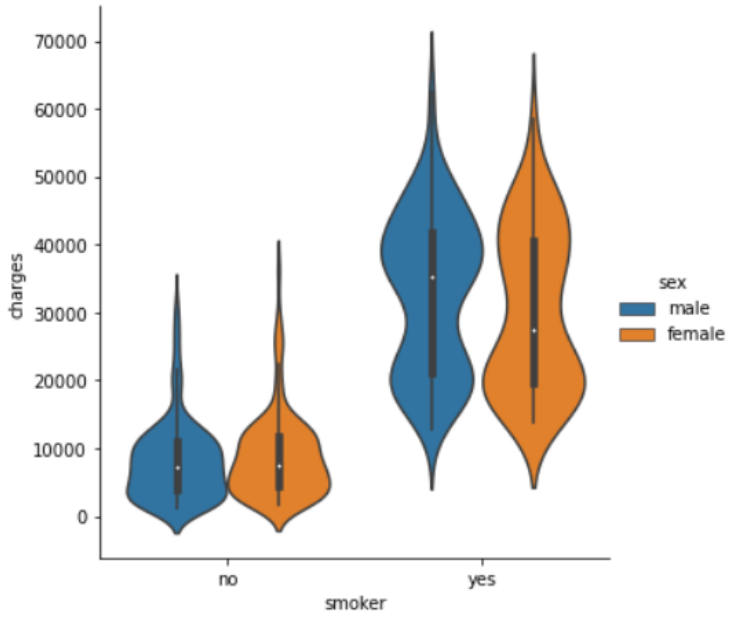
- Sự tương quan giữa smoker với charges



- Nhận xét:

- ❖ Số lượng người không hút thuốc cao gấp 3 lần số lượng người hút thuốc.
- ❖ Người hút thuốc sẽ phải tốn nhiều chi phí y tế hơn so với người không hút thuốc.
- ❖ Thế nên **smoker có tương quan mạnh đối với charges**

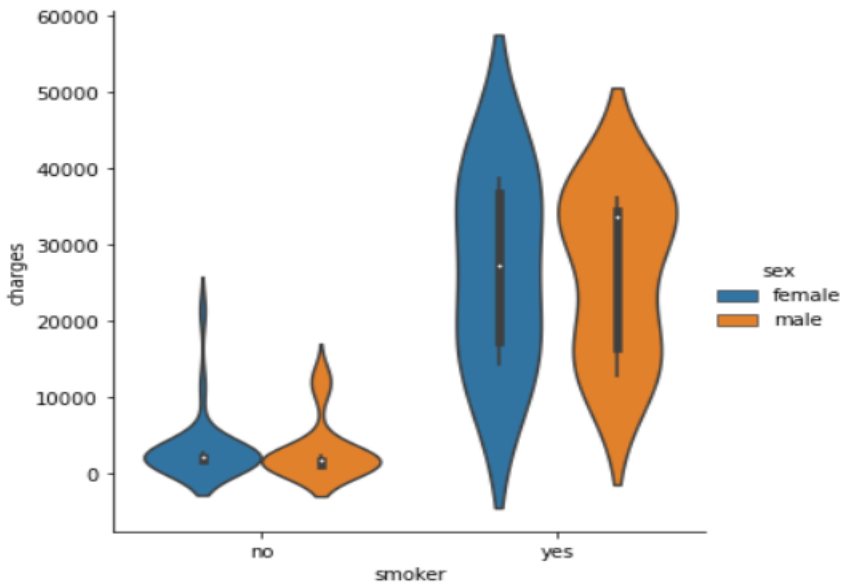
c. Sự tương quan giữa smoker, sex và charges:



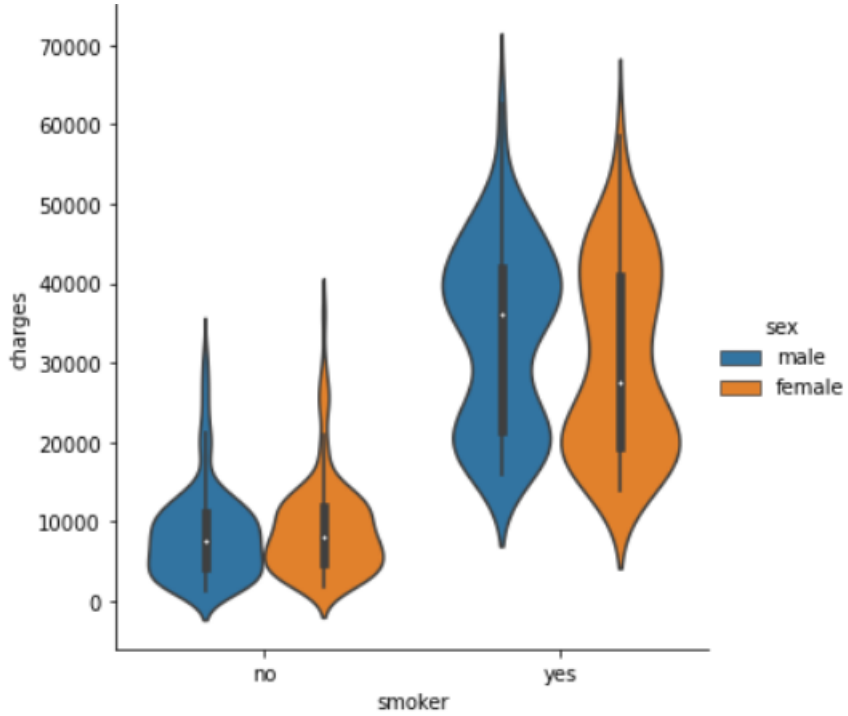
- ❖ Ở phần trên ta có thể thấy **smoker = yes** thì chi phí y tế sẽ phải trả nhiều hơn
- ❖ **Bất kì giới tính nào nếu hút thuốc thì đều phải tốn nhiều chi phí y tế hơn**
- ❖ Khi thêm thuộc tính **sex** vào thì cũng không có sự thay đổi gì.
- ❖ Chưa nhận thấy được sự tương quan mạnh giữa **sex** với **charges**

d. Sự tương quan giữa smoker, sex, age và charges:

- Với **age = 18**



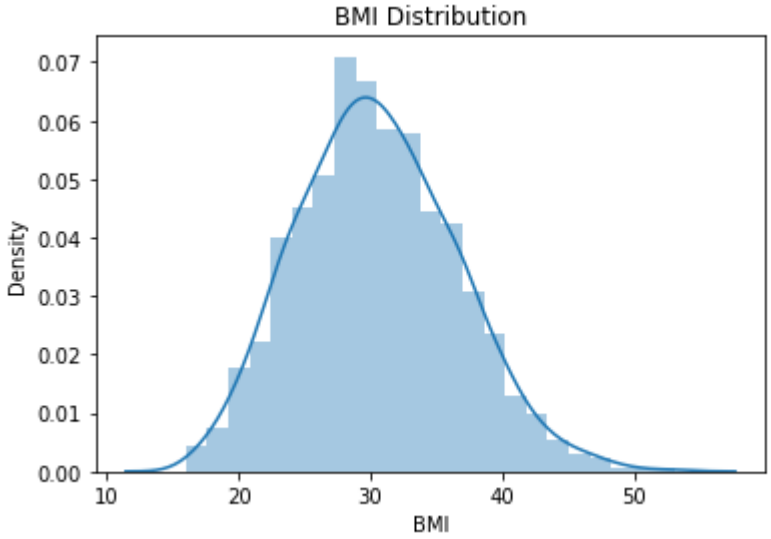
- Với **age > 18**



❖ Không có sự khác biệt đáng kể nào, bất kỳ giới tính hay lứa tuổi nào nếu hút thuốc thì sẽ phải tốn nhiều chi phí y tế hơn.

e. Sự tương quan giữa BMI với charges:

- Xét sự phân phối của chỉ số BMI

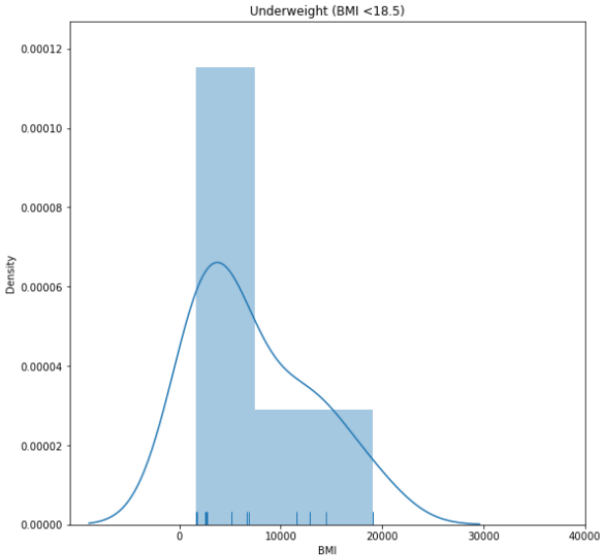


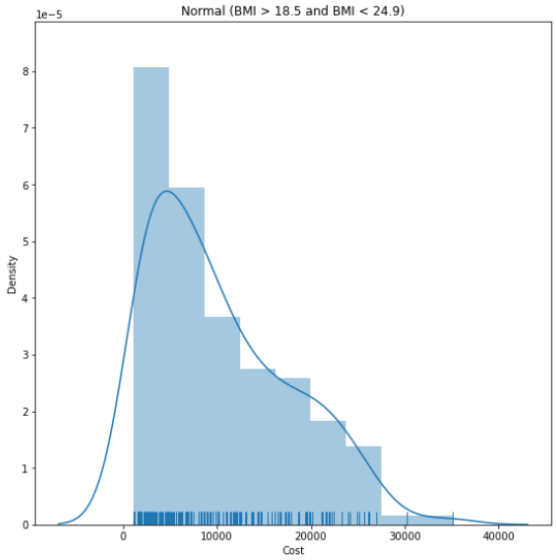
❖ **Nhận xét:** ta thấy **BMI** trung bình rời vào khoảng 30 điều này thật sự không tốt
❖ Theo **WHO** ta có thể chia thang đo BMI thành 4 loại như sau:

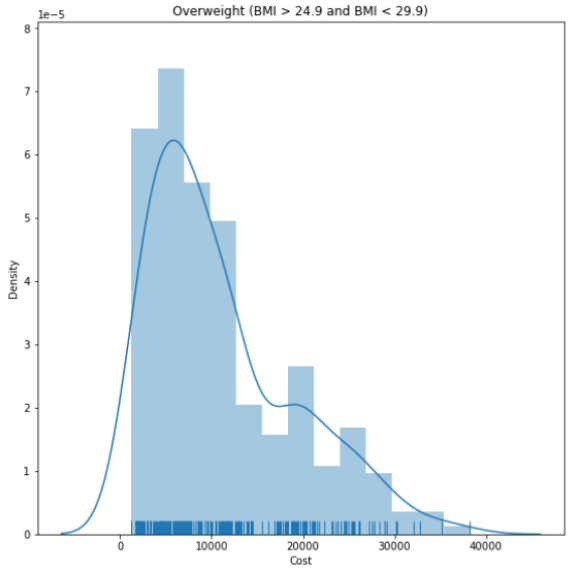
1. Underweight - BMI < 18.5

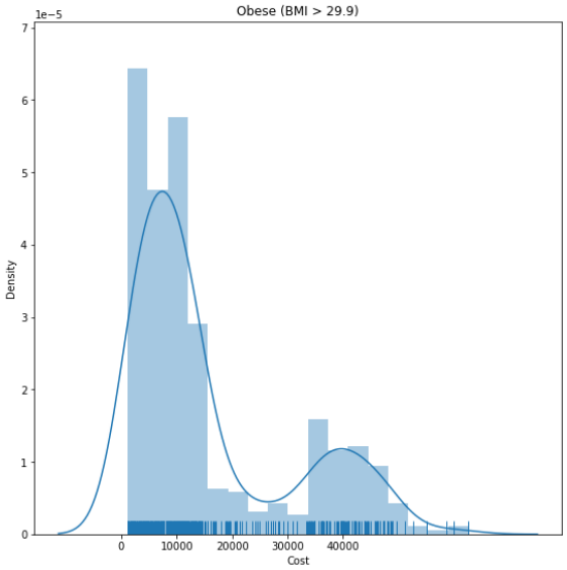
2. Normal - BMI >= 18.5 and BMI < 24.9
3. Overweight - BMI >= 24.9 and BMI < 29.9
4. Obese - BMI >= 30

- Sự tương quan giữa BMI với charges





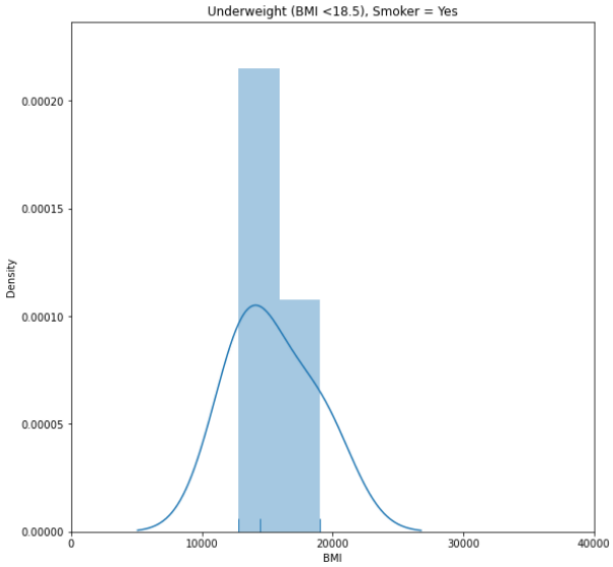


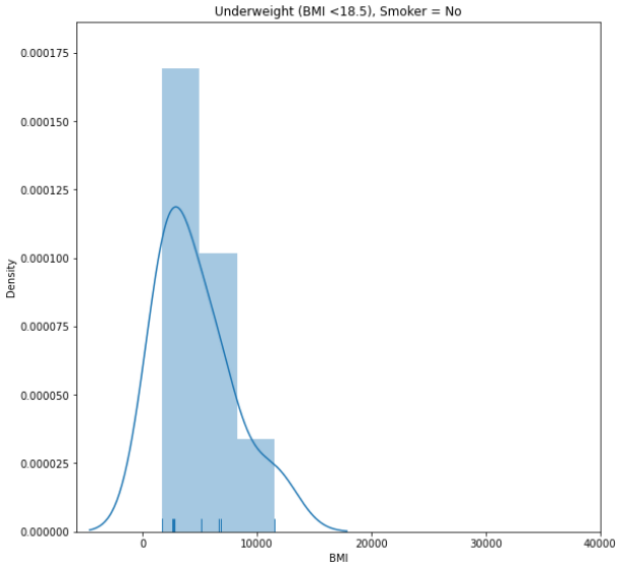


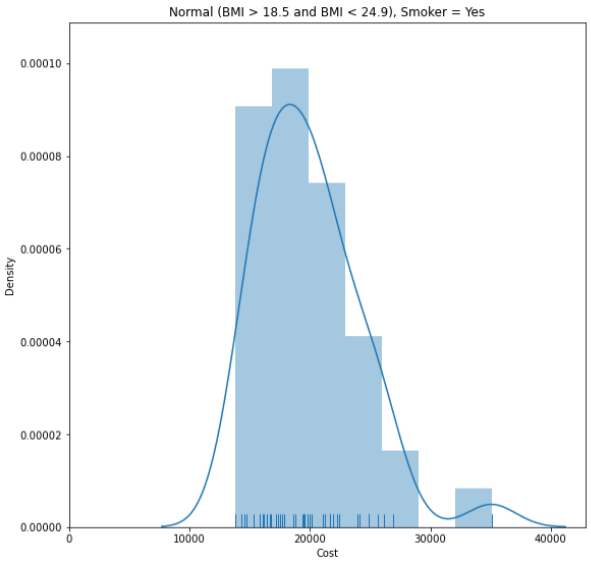
- Nhận xét:

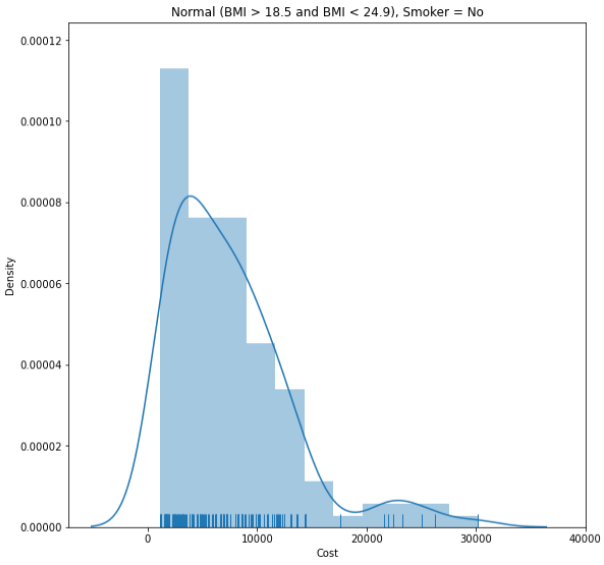
- ❖ Hình dáng đồ thì có sự thay đổi ở sườn bên phải
- ❖ Chi phí sẽ tăng nhẹ đối với mỗi loại **BMI** khác nhau

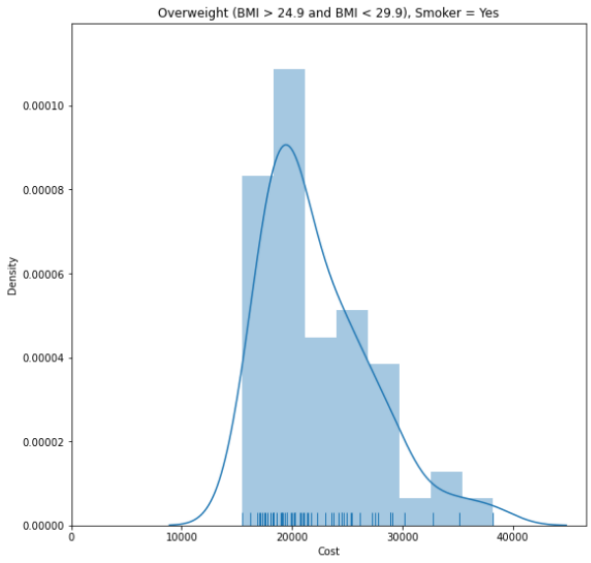
f. Giữa BMI, smoker với charges:

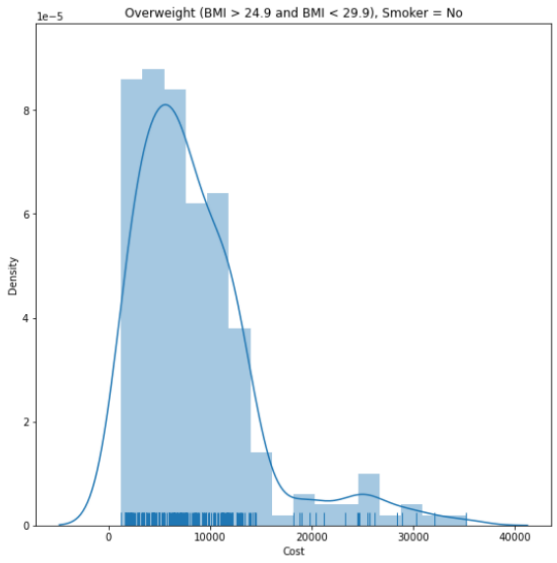


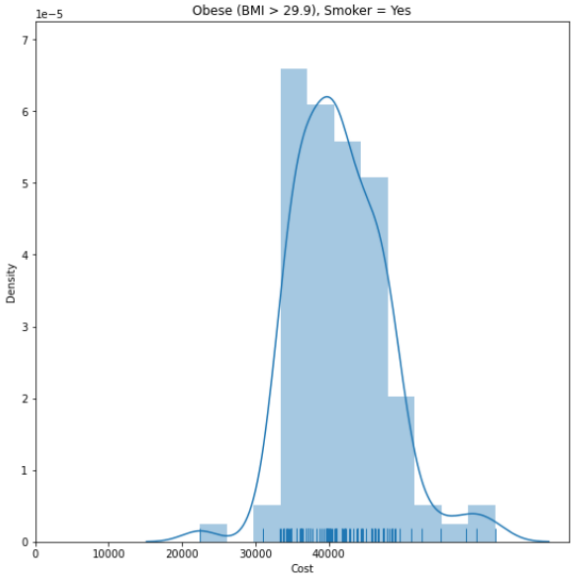


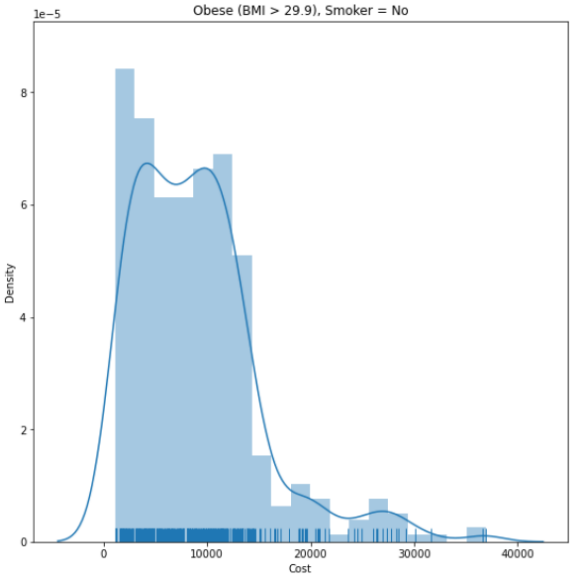








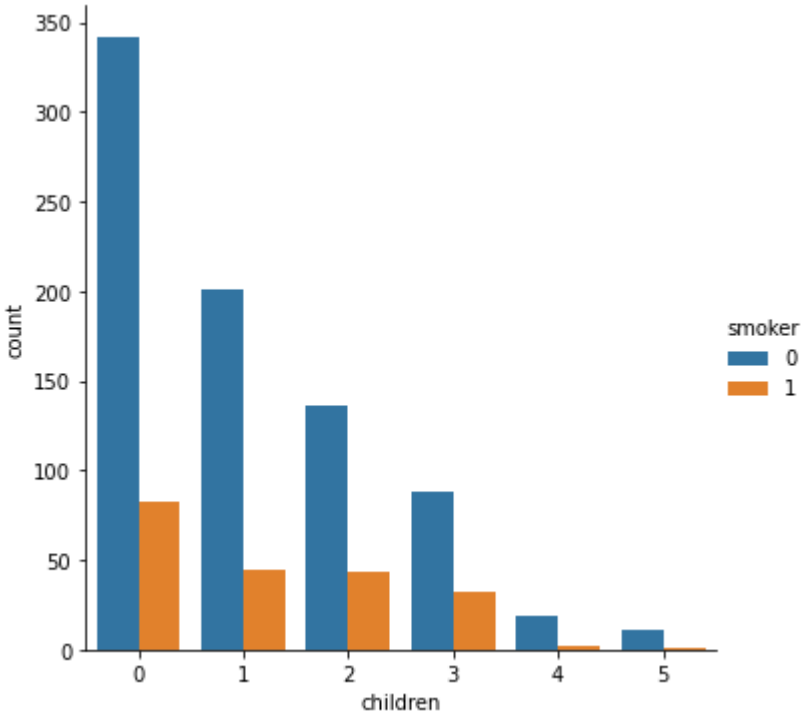




- Bất kỳ chỉ số **BMI** như thế nào nếu **có hút thuốc** thì sẽ phải trả nhiều chi phí y tế hơn

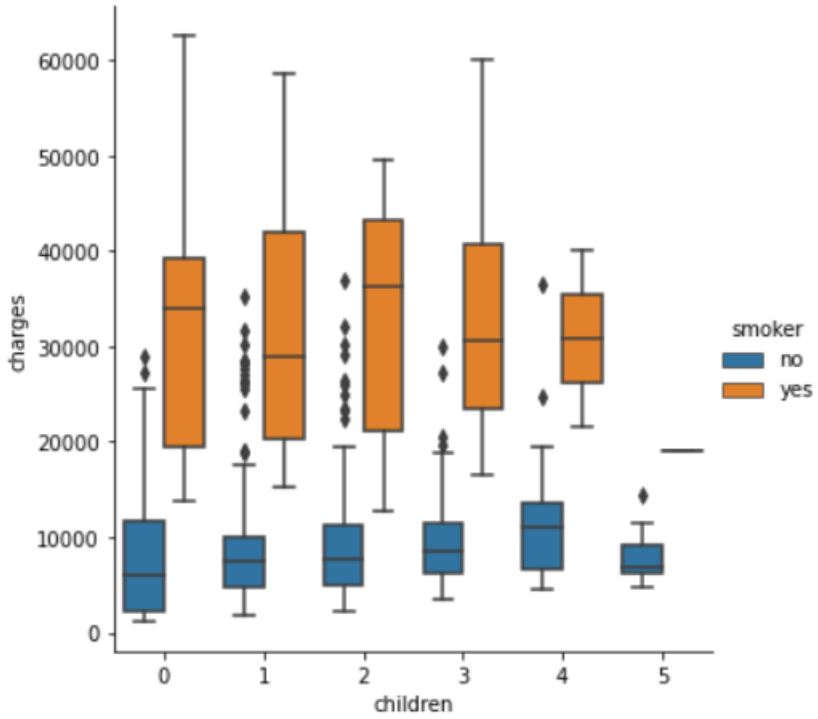
g. Sự tương quan giữa smoker và children

- **Dự đoán:** thông thường nếu gia đình có trẻ con thì người lớn sẽ ít hút thuốc hơn vì sợ ảnh hưởng đến sức khỏe của trẻ nhỏ.



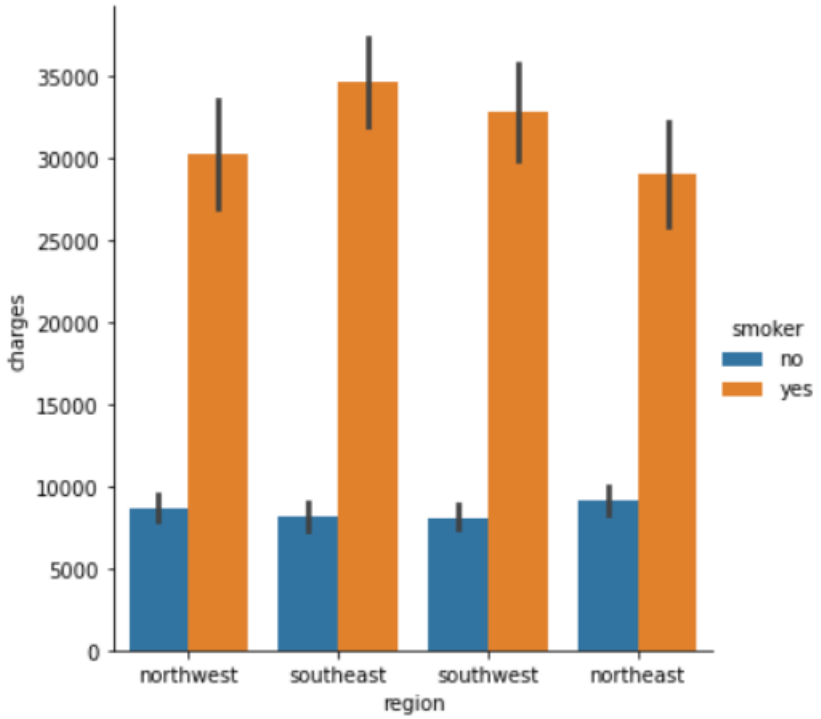
❖ **Nhận xét:** ta thấy số children càng lớn thì số smoker càng giảm

h. Sự tương quan giữa smoker, children và charges:



❖ Bất kể nó bao nhiêu children đi nữa nếu hút thuốc thì sẽ phải tốn nhiều chi phí y tế hơn.

i. Giữa Region và Charges:



- Vùng miền không ảnh hưởng nhiều đến charges
- Ở **southeast** sẽ nhiều hơn các vùng còn lại một tí

- **Kết luận:**

Những yếu tố quan trọng quyết định đến chi phí y tế:

+ **Smoker**: một người có hút thuốc hay không.

+ **Age**: tuổi tác, tuổi tác càng già thì cần được chăm sóc y tế nhiều hơn.

+ **BMI**: chỉ số BMI càng lớn (>30) có nguy cơ béo phì nên chi phí y tế sẽ tăng.

+ Những yếu tố như **Region, Children, Sex** không ảnh hưởng nhiều đến việc tăng giảm chi phí y tế.

IV. *Thực hiện tiền xử lý và áp dụng các mô hình máy học*:

Gồm bốn bước chính:

- **Xử lý các số liệu nhiễu**
- **Chuyển các dữ liệu thành dạng số**
- **Chuẩn hóa**
- **Áp dụng mô hình máy học**

1. Tiền xử lý:

❖ Chia tập train thành 2 tập train và validation với tỷ lệ 7/3

❖ Kiểu dữ liệu các cột:

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
dtype: object
```

- Nhận xét về tập dữ liệu
  - Dữ liệu có 6 thuộc tính.
  - Các thuộc tính có kiểu dữ liệu có vẻ phù hợp.

❖ Như đã phân tích các nhân tố của từng thuộc tính ảnh hưởng đến mức độ chi phí, do đó ta sẽ sắp xếp và chuyển đổi các nhân tố của từng thuộc tính đó theo thứ tự dạng numeric, tùy theo độ mức độ ảnh hưởng.

❖ Ở đây, ví dụ: Với thuộc tính smoker, sau khi phân tích ở trên, ta có thể nhận thấy chi phí trung bình của người có hút thuốc cao gấp 4 lần so với người không hút thuốc nên chuyển yes sang 4 và no sang 1. Tương tự cho các thuộc tính khác, sẽ được chuyển đổi sao cho phù hợp.

❖ Class ColAdderDropper sẽ thực hiện các bước ở trên.

❖ Ngoài ra, class ColAdderDropper được kế thừa từ 2 class của Sklearn là BaseEstimator và TransformerMixin. Việc kế thừa này giúp class của ta tự động có các phương thức như set_params, get_params, fit_transform.

	age	sex	bmi	children	smoker	region	charges
0	24	male	23.655	2	1	northwest	2352.96845
1	28	female	26.510	1	1	southeast	4340.44090
2	51	male	39.700	3	1	southwest	9391.34600
3	47	male	36.080	3	4	southeast	42211.13820
4	46	female	28.900	1	1	southwest	8823.27900
...	...	...	...	...	...	...	...
998	18	female	31.350	4	1	northeast	4561.18850
999	39	female	23.870	1	1	southeast	8582.30230
1000	58	male	25.175	2	1	northeast	11931.12525
1001	37	female	47.600	1	4	southwest	46113.51100
1002	55	male	29.900	2	1	southwest	10214.63600

1003 rows × 7 columns

❖ Bây giờ ta chuyển tất cả các cột về dạng số như sau:

+ Do các cột không chứa các giá trị thiếu nên ta không cần điền giá trị thiếu vào.

+ Với các cột dạng số, ta giữ nguyên.

+ Các cột không phải dạng số và không có thứ tự, ta sẽ mã hóa bằng one-hot.

❖ Cuối cùng tạo một pipeline sử dụng các thao tác trên cùng lúc.

```
unorder_cate_cols = ['sex','region']

mode_unordercols = make_pipeline(OneHotEncoder(handle_unknown='ignore'))

col_transform = ColumnTransformer([('unorder_cate_cols', mode_unordercols,unorder_cate_cols)],remainder='passthrough')

preprocess_pipeline = make_pipeline(col_adderdropper,col_transform,StandardScaler())
preprocessed_train_X = preprocess_pipeline.fit_transform(data_df)
preprocessed_train_X
```

❖ Độ lỗi: Sử dụng độ do R-Squared cho mô hình hồi quy.

2. Áp dụng các mô hình học máy:

> Đối với mô hình cơ bản

❖ Gồm 3 bước:

- Tạo pipeline hoàn chỉnh cho cả mô hình và tìm tham số tốt nhất.
- Huấn luyện lại mô hình với tham số tốt nhất.
- Dự đoán trên tập test và tính độ chính xác của mô hình.

❖ Mô hình sử dụng:

1. **Linear Regression.**
2. **Linear Regression với Gradient Descent**

❖ Nhận xét cũng như độ chính xác của mô hình:

Hai mô hình có độ chính xác xấp xỉ gần bằng nhau: 76.793518%

> Đối với mô hình nâng cao

❖ **Gồm 3 bước:**

- Tạo pipeline hoàn chỉnh cho mô hình.
- Huấn luyện lại mô hình với các tham số trên tập validation.
- Chọn mô hình tốt nhất để dự đoán trên tập test.

❖ **Mô hình sử dụng:**

1. **MLPRegresson:**
   Với các tham số: activation='logistic', solver='lbfgs', random_state=0, max_iter=2500, alpha và hidden_layer_sizes.
2. **RainForestRegression.**
   Với các tham số: max_depth, n_estimators.

❖ **Lựa chọn mô hình tốt nhất trên tập validation**:

Chọn mô hình RandomForestRegresson vì cho độ lỗi trên tập validation thấp nhất với 15%.

3. Tổng kết:

Dự đoán cho tập test sau khi huấn luyện trên tập train.

```
print("Độ chính xác")
print("Linear regression:                          ",linear_score)
print("Linear regression sử dụng Gradient Decent: ",GDLR_score)
print("RandomForestRegressor:                      ",RFR_score)
```

```
Độ chính xác
Linear regression:                          76.39518336880413
Linear regression sử dụng Gradient Decent:  76.39518336884825
RandomForestRegressor:                      85.03163260852679
```

- **Vậy qua các mô hình trên, ta tìm được mô hình tốt nhất cho việc huấn luyện. Với mô hình RandomForestRegresson ở mô hình nâng cao cho ra dự đoán với độ chính xác lên đến 85% (khá cao).**

V. Tài liệu tham khảo

1) https://github.com/namas191297/medical_cost_eda
2) https://api.rpubs.com/msahmed3/493560?fbclid=IwAR1aNDQAh_vMWIKZLyPyMkGU_YiDDz_0GcsRTO49s_Pwomq053ZjRH0h-CU
3) https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
4) https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

---------- HẾT ----------