ĐẠI HỌC QUỐC GIA TP.HCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Môn học: Nhập môn học máy

Báo cáo đồ án thực hành 1: Regression

Nhóm sinh viên

Nguyễn Huy Hải – 18120023

Phạm Công Minh – 18120058

Nguyễn Thanh Tùng – 18120104

Lê Minh Đức – 18120164

Trần Đại Tài – 18120543

Contents

1 Phân tích các đặc trưng trong hai tập dữ liệu

- Kiểu dữ liệu của các cột.

Tên cột	Loại dữ liệu
age	int64
sex	object
bmi	float64
children	int64
smoker	object
region	object
charges	float64

Nhận xét:

  - Ở cột children, kiểu dữ liệu là dạng Numerical nhưng thật ra cột này có kiểu Ordinal categorical nên để cột này dạng Numerical sẽ không có ảnh hưởng.
  - Các cột còn lại không có gì bất thường.

- Dữ liệu có 1338 dòng, 7 cột (trong đó 1003 dòng ở tập train, 335 dòng ở tập test).

- Thống kê chi tiết của các cột có kiểu dữ liệu Categorical.

	sex	smoker	region
count	1338	1338	1338
unique	2	2	4
top	male	no	southeast
freq	676	1064	364

- Thống kê chi tiết của các cột có kiểu dữ liệu Numerical.

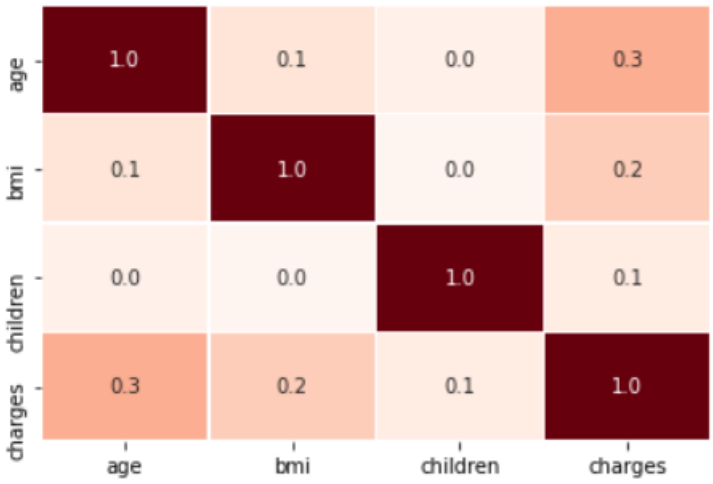
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

- Không có cột nào bị thiếu dữ liệu.
- Dữ liệu có bị lặp ở 1 dòng.

2 Phân tích, trình bày các thông tin hữu ích tác động đến chi phi y tế.

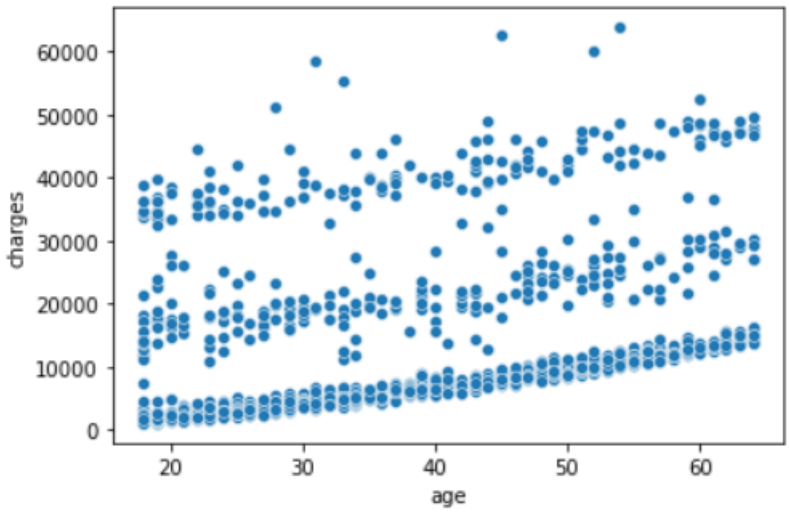
2.1 Mối tương quan giữa các thuộc tính dạng số với chi phí y tế.

- Tính Correlation giữa các thuộc tính dạng số.

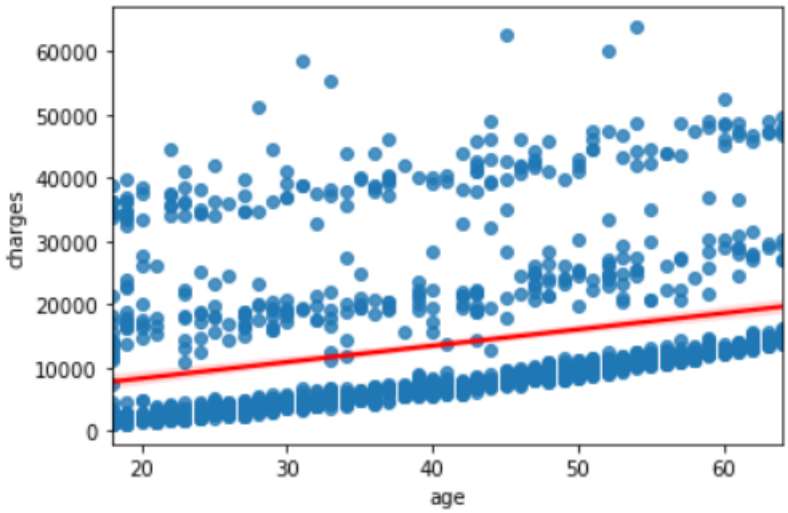


Nhận xét: Age và Charges, BMI và Charges, Children và Charges có mối tương quan thuận nhưng không mạnh (chỉ 0.3, 0.2 và 0.1). Ta sẽ vẽ biểu đồ để thể hiện rõ sự tương quan này.

- Biểu đồ thể hiện mối quan hệ giữa Age và Charges

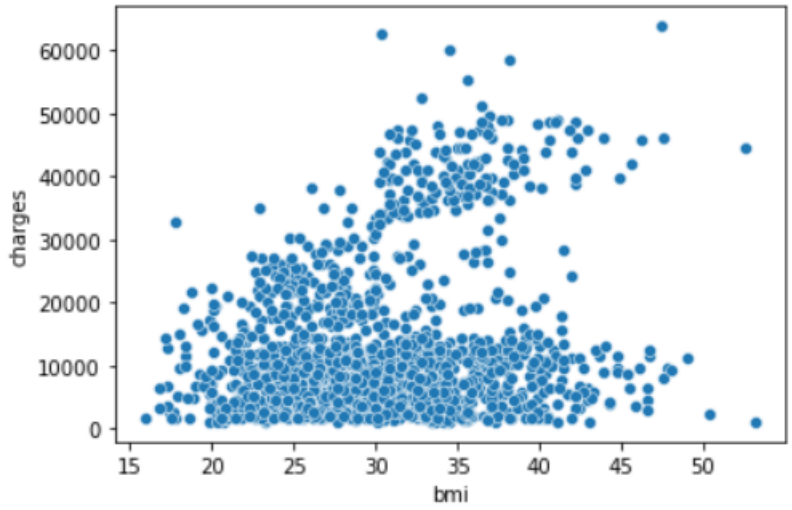


Để quan sát rõ hơn, ta sẽ dùng Linear Regression để tìm đường thẳng đi qua các điểm này.

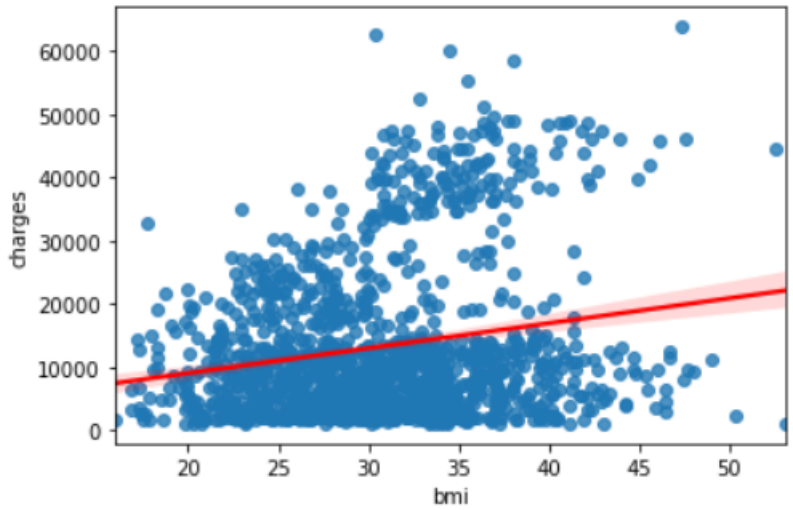


Nhận xét: Khi độ tuổi tăng thì chi phi y tế cũng tăng theo.

- Biểu đồ thể hiện mối quan hệ giữa BMI và Charges

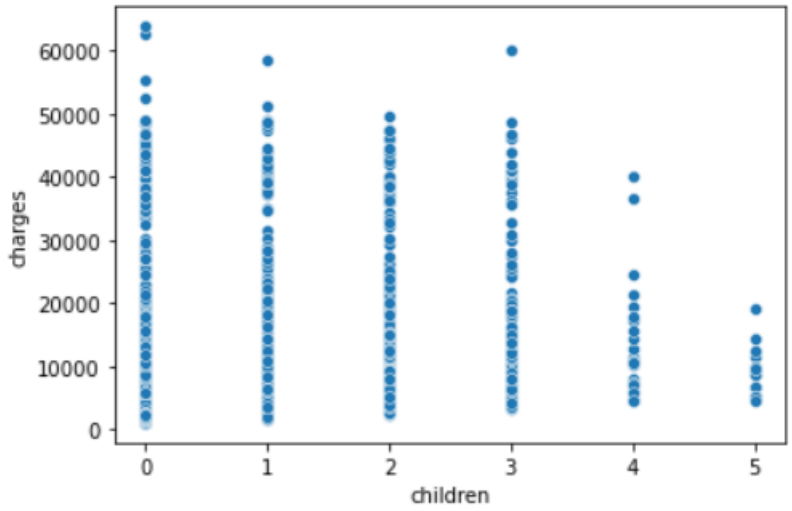


Để quan sát rõ hơn, ta sẽ dùng Linear Regression để tìm đường thẳng đi qua các điểm này.

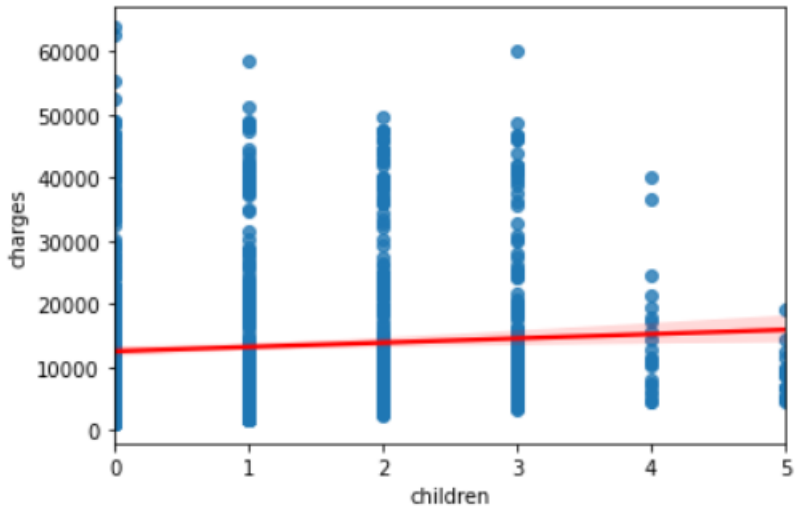


Nhận xét: Khi chỉ số BMI tăng thì chi phi y tế cũng tăng theo.

- Biểu đồ thể hiện mối quan hệ giữa Children và Charges



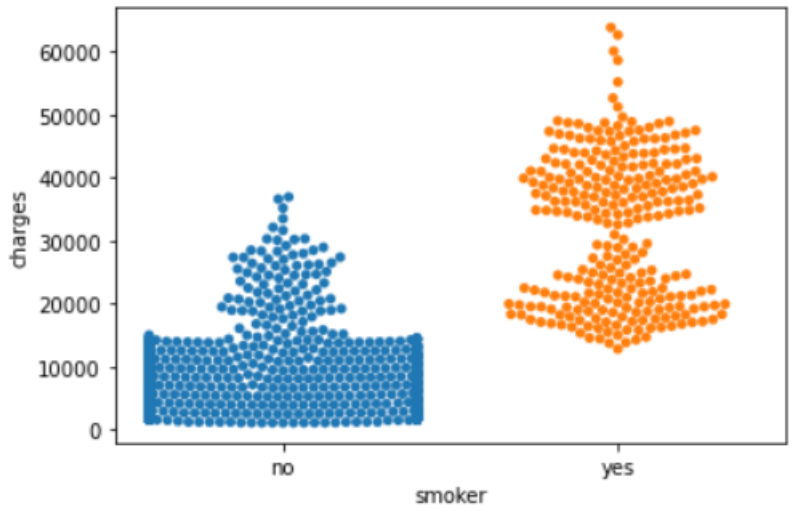
Để quan sát rõ hơn, ta sẽ dùng Linear Regression để tìm đường thẳng đi qua các điểm này.



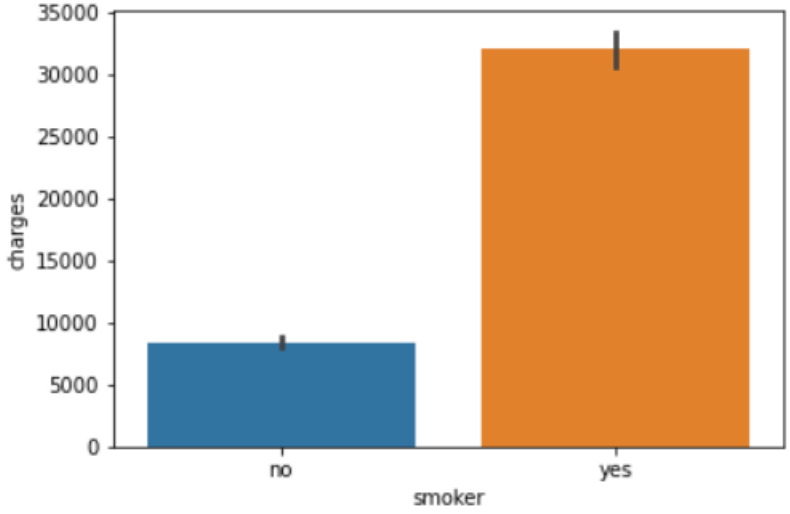
Nhận xét: Khi số trẻ con càng nhiều thì chi phi y tế cũng tăng theo

2.2 Mối tương quan giữa các thuộc tính dạng phân loại với chi phí y tế.

- Biểu đồ thể hiện mối quan hệ giữa Smoker và Charge.



Biểu đồ thể hiện sự phân bố của Smoker và Charges.



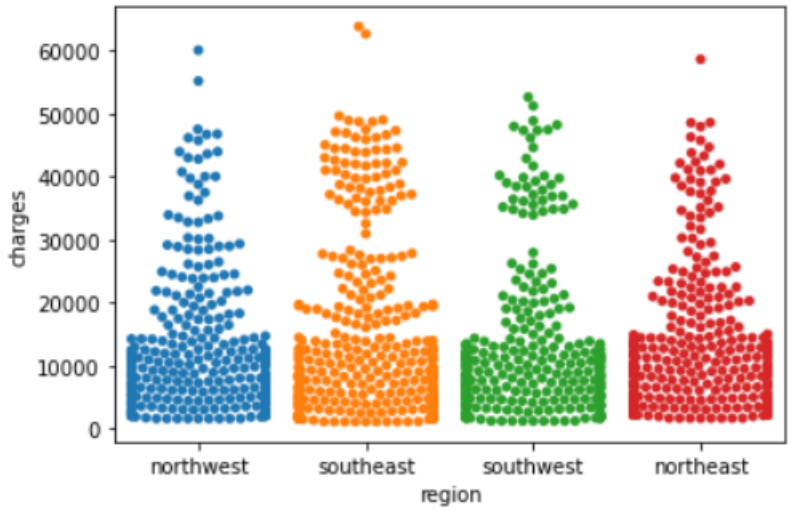
Biểu đồ thể hiện giá trị trung bình của Charges với từng giá trị của thuộc tính trong Smoker.

Nhận xét:

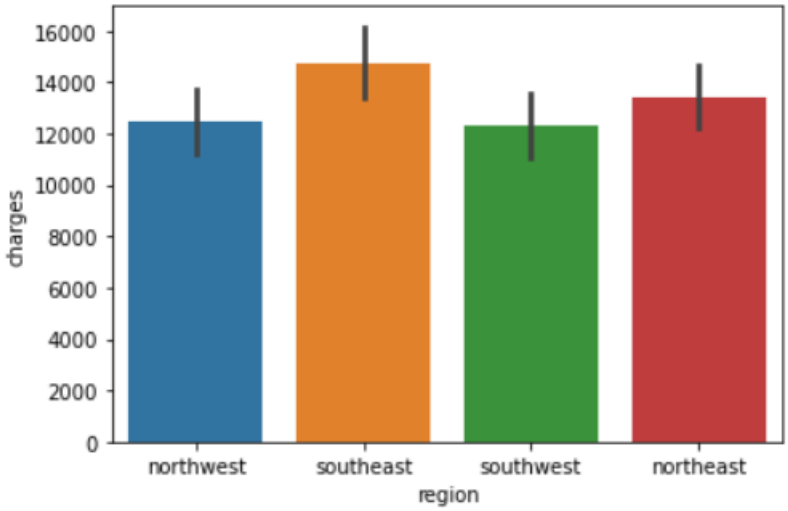
- Những người hút thuốc sẽ có chi phí y tế cao hơn những người không hút thuốc.
- Nhìn biểu đồ giá trị trung bình ta có thể thấy độ chênh lệch này là rất nhiều (khoảng hơn 20000).

Vậy Smoker có ý nghĩa rất quan trọng trong việc dự đoán Charges.

- Biểu đồ thể hiện mối quan hệ giữa Region và Charges.



Biểu đồ thể hiện sự phân bố của Region và Charges.



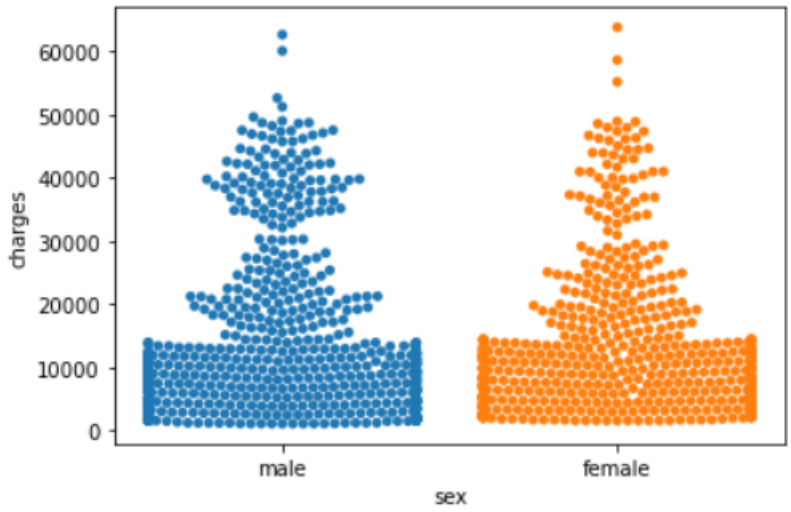
Biểu đồ thể hiện giá trị trung bình của Charges với từng giá trị của thuộc tính trong Region.

Nhận xét:

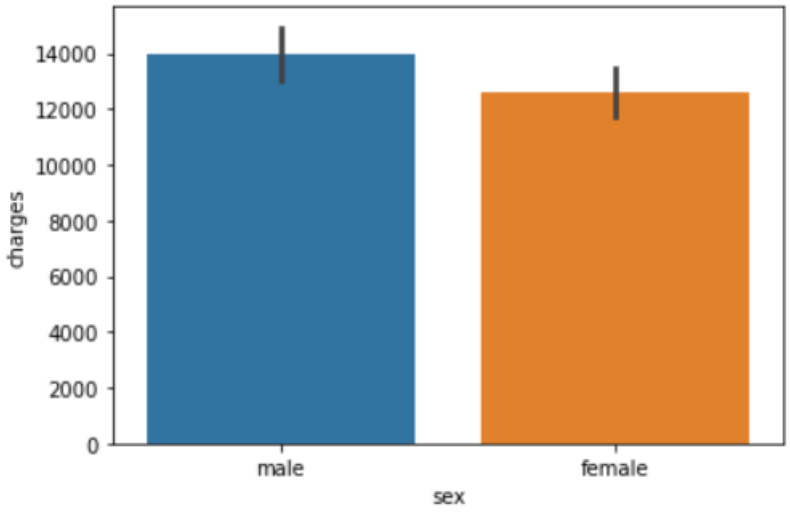
- Những người ở vùng Đông Nam sẽ có chi phí y tế cao hơn tuy không nhiều.
- Tuy nhiên độ chênh lệch chi phí y tế giữa các vùng với nhau không nhiều, gần như tương đương nhau.

Vì vậy có khả năng Region sẽ không ảnh hưởng nhiều đến việc dự đoán Charges.

- Biểu đồ thể hiện mối quan hệ giữa Sex và Charges.



Biểu đồ thể hiện sự phân bố của Sex và Charges.



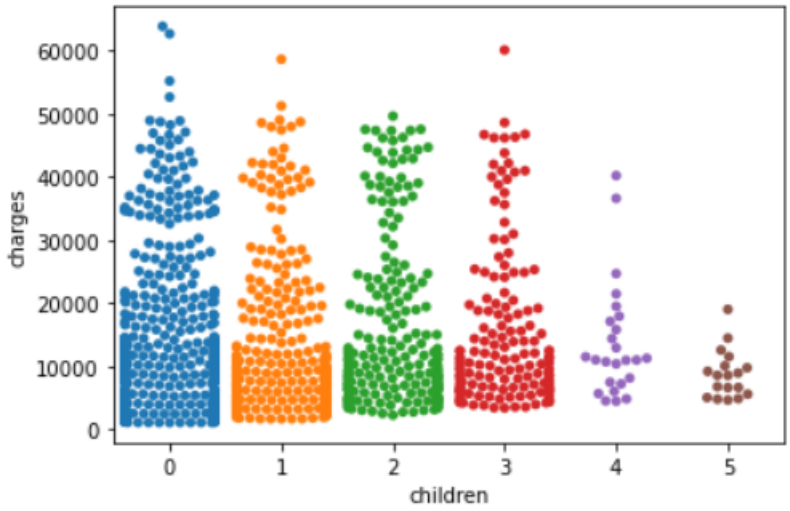
Biểu đồ thể hiện giá trị trung bình của Charges với từng giá trị của thuộc tính trong Sex.

Nhận xét:

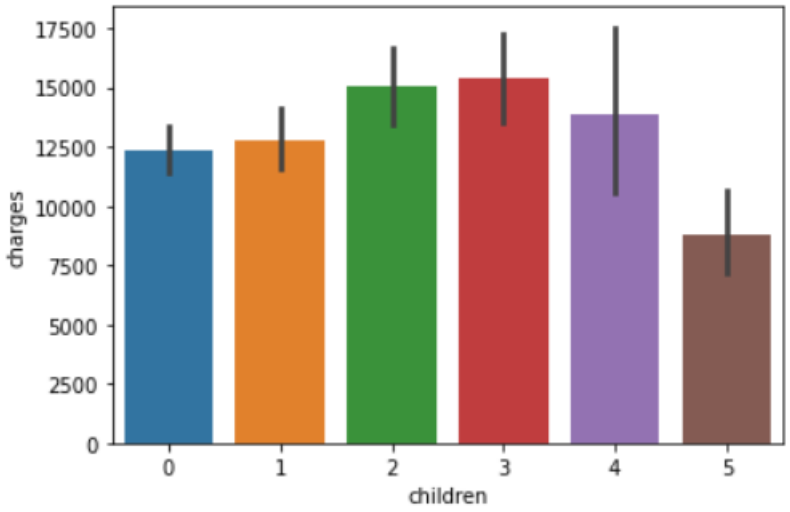
- Những người nam sẽ có chi phí y tế cao hơn tuy không nhiều.
- Độ chênh lệch chi phí y tế giữa giới tính không nhiều, gần như tương đương nhau.

Vì vậy có khả năng cũng giống như Region, Sex sẽ không ảnh hưởng nhiều đến việc dự đoán Charges.

- Biểu đồ thể hiện mối quan hệ giữa Children và Charges.



Biểu đồ thể hiện sự phân bố của Children và Charges.



Biểu đồ thể hiện giá trị trung bình của Charges với từng giá trị của thuộc tính trong Children.

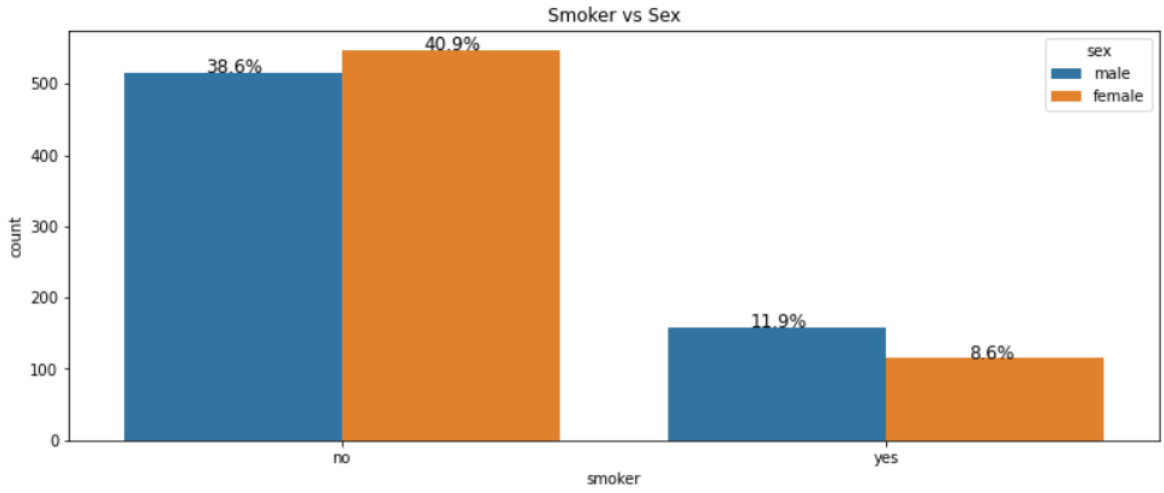
Nhận xét:

- Những người có 2-3 con có chi phí dịch vụ cao hơn những người khác.
- Những người có 5 con có chi phí dịch vụ thấp nhất.

3 Phân tích thêm về cột Smoker.

Giải thích: Vì Smoker có ảnh hưởng lớn đến Charges nên ta sẽ khảo sát thử mối quan hệ giữa nó với các thuộc tính khác trong dữ liệu.

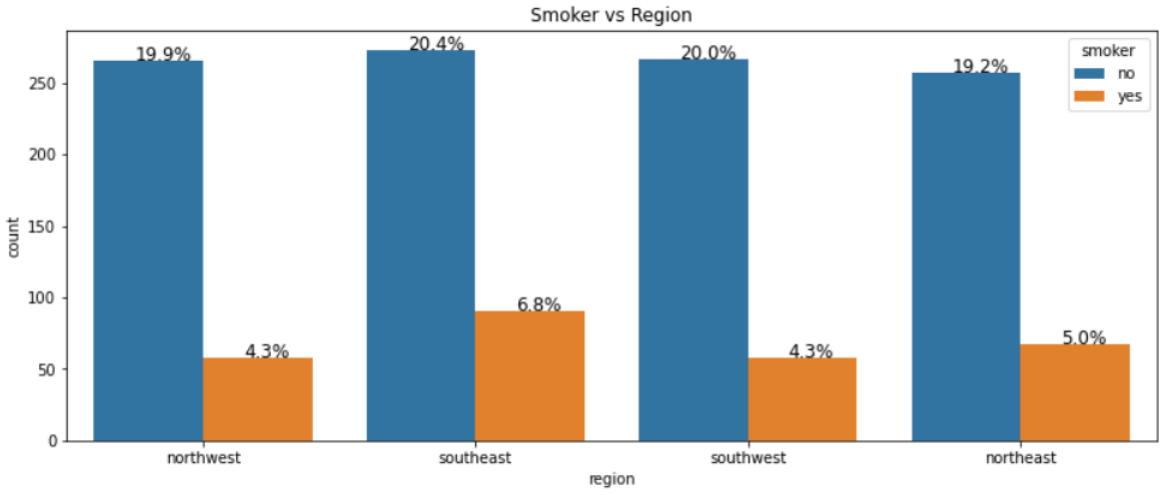
- Tỉ lệ hút thuốc và không hút thuốc ở từng giới tính.



Nhận xét:

- Tỉ lệ nam hút thuốc nhiều hơn nữ nhưng không nhiều.
- Tỉ người không hút thuốc gấp gần 4 lần người hút thuốc ở cả nam và nữ.

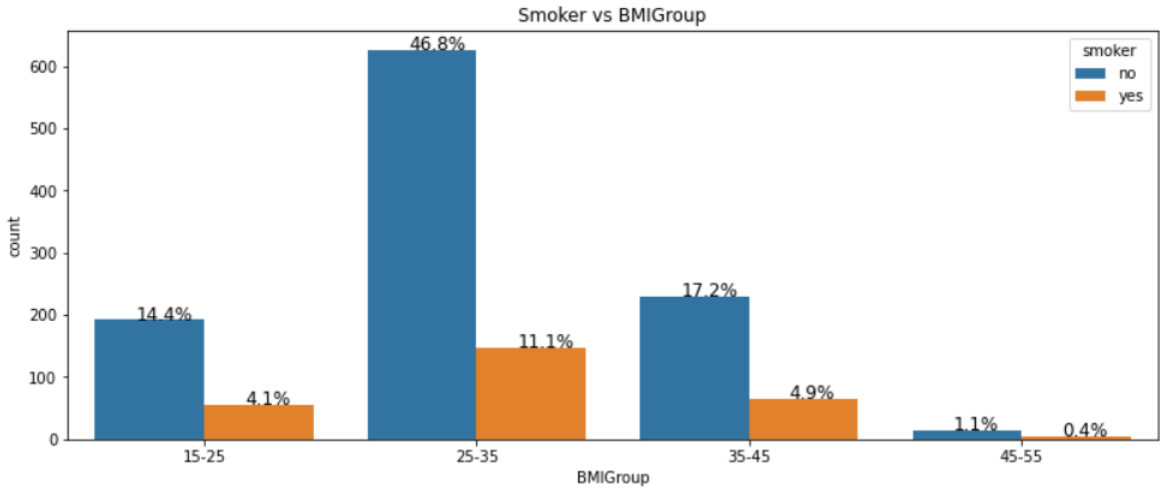
- Tỉ lệ hút thuốc và không hút thuốc ở từng vùng miền.



Nhận xét:

- Tỉ lệ giữ người không hút thuốc giữa các vùng khá đồng đều.
- Vùng Đông Nam có lượng người hút thuốc đông nhất.

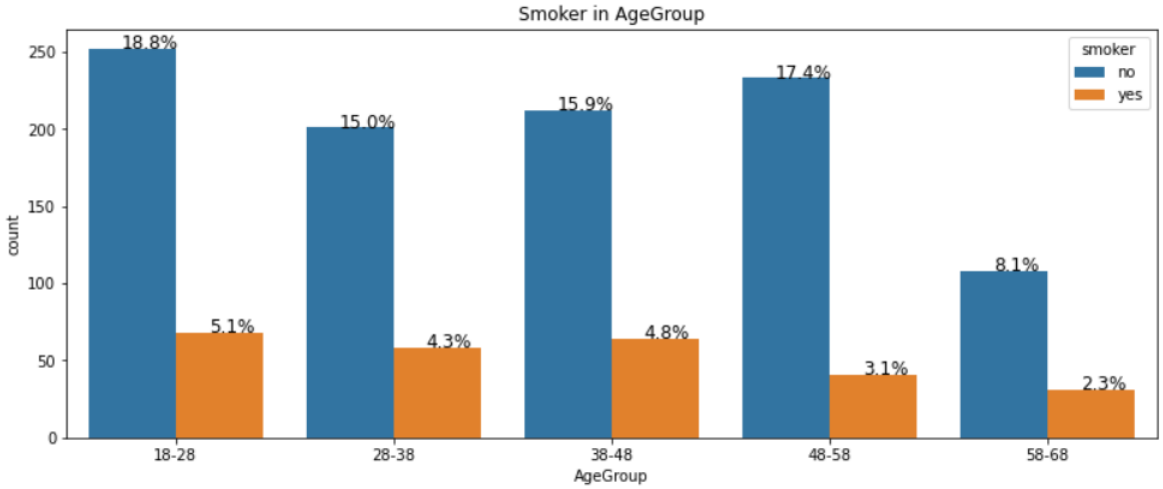
- Tỉ lệ hút thuốc và không hút thuốc ở từng nhóm BMI.



Nhận xét:

- Những người có chỉ số BMI từ 25-35 có tỉ lệ hút thuốc và không hút thuốc cao nhất.
- Những người có chỉ số BMI từ 45-55 có tỉ lệ hút thuốc và không hút thuốc thấp nhất (thấp hơn các nhóm kia tương đối nhiều).

- Tỉ lệ hút thuốc và không hút thuốc ở từng nhóm tuổi.



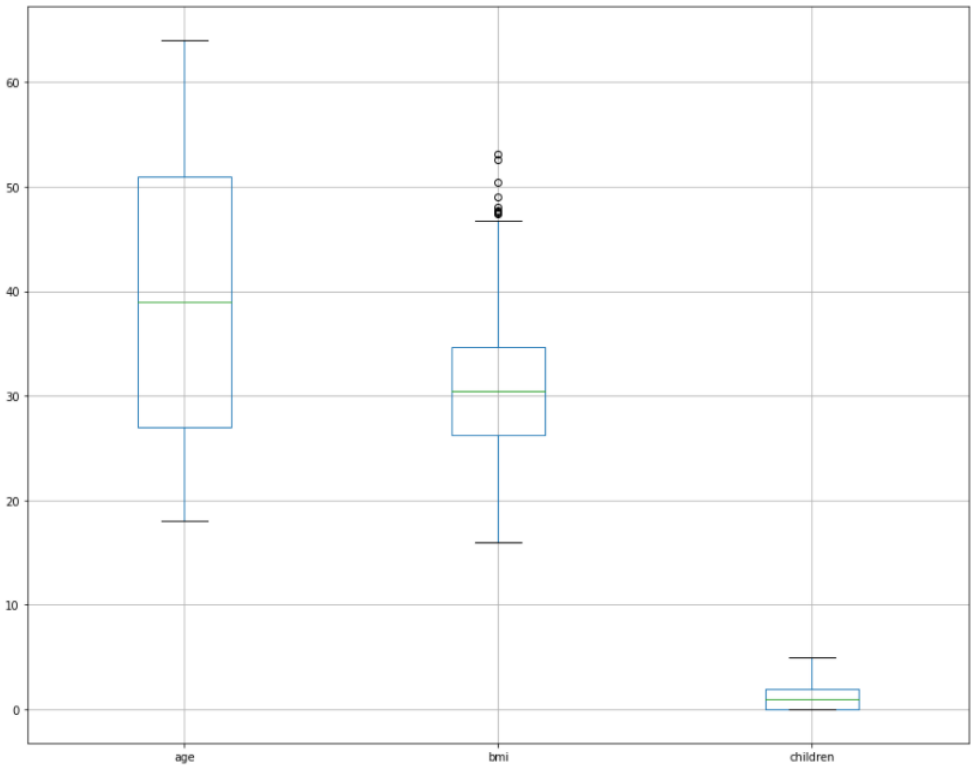
Nhận xét:

- Chỉ có nhóm người có độ tuổi từ 58-68 có tỉ lệ hút thuốc và không hút thuốc thấp nhất.
- Còn các nhóm kia có tỉ lệ không chênh lệch nhiều.

4 Cài đặt thuật toán dự đoán chi phí y tế.

4.1 Tiền xử lý dữ liệu.

4.1.1 Xử lý dữ liệu nhiễu.



- Quan sát biểu đồ ta có thể thấy ở cột BMI có tương đối nhiều dữ liệu nhiễu. Vì vậy ta cần phải xử lý nhiễu này bằng cách thay thế những dữ liệu này bằng giá trị mean của cột BMI.
- Sau khi tính toán, thì những giá trị lớn hơn 47.351 là dữ liệu nhiễu, sẽ được thay thế bằng 30.5112 là giá trị mean của cột BMI.

4.1.2 Xóa những cột không ảnh hưởng đến dự đoán chi phí y tế bằng cách kiểm định thống kê.

- Trước khi kiểm định ta cần chuyển các cột dạng Categorical về dạng Numerical bằng phương pháp OneHotEncoder.
- Ta có bảng phân tích thống kê như sau:

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.751
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	500.0
Date:	Thu, 13 May 2021	Prob (F-statistic):	0.00
Time:	23:16:39	Log-Likelihood:	-13538.
No. Observations:	1337	AIC:	2.709e+04
Df Residuals:	1328	BIC:	2.714e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-295.0121	430.721	-0.685	0.494	-1139.980	549.956
age	256.7646	11.912	21.555	0.000	233.396	280.133
bmi	339.2504	28.611	11.857	0.000	283.122	395.379
children	474.8205	137.897	3.443	0.001	204.301	745.340
smoker_no	-1.207e+04	282.517	-42.727	0.000	-1.26e+04	-1.15e+04
smoker_yes	1.178e+04	313.644	37.546	0.000	1.12e+04	1.24e+04
sex_female	-82.7653	269.326	-0.307	0.759	-611.116	445.585
sex_male	-212.2468	275.197	-0.771	0.441	-752.115	327.622
region_northeast	512.3904	300.468	1.705	0.088	-77.053	1101.834
region_northwest	163.1638	301.884	0.540	0.589	-429.058	755.386
region_southeast	-522.8752	330.901	-1.580	0.114	-1172.020	126.270
region_southwest	-447.6910	311.058	-1.439	0.150	-1057.909	162.527

Omnibus:	299.816	Durbin-Watson:	2.107
Prob(Omnibus):	0.000	Jarque-Bera (JB):	716.552
Skew:	1.211	Prob(JB):	2.53e-156
Kurtosis:	5.646	Cond. No.	5.74e+17

Với mức ý nghĩa là 0.05, ta có:

- Ở phần sex, region có p-value > 0.05, nghĩa là sex và region không có ảnh hưởng đến chi phí y tế. Vì vậy, ta có thể loại bỏ 2 cột này.
- Ở phần age, bmi, smoker, children có p-value < 0.05, nghĩa là có ảnh hưởng đến chi phí y tế. Vì vậy ta giữ nguyên 4 cột này.

Vậy: ta sẽ xóa 2 cột là sex và region.

4.1.3 Chuẩn hóa các cột dữ liệu.

- Với các cột dữ liệu dạng Numerical, ta sẽ dùng phương pháp StandardScaler để chuẩn hóa về khoảng [0,1].
- Với các cột dữ liệu dạng Categorical, ta sẽ dùng phương pháp OneHotEncoder.

4.2 Xây dựng mô hình.

4.2.1 Linear Regression.

Sau khi huẩn luyện mô hình trên tập train, ta có:

- Hệ số intercept: 20230.4933
- Hệ số coef lần lượt là: 3648.4665, 1968.7819, 524.0369 , -11816.32, 11816.32

Lý do vì sao khi huấn luyện có 4 cột nhưng lại có 5 tham số? Vì khi đó ta đã OneHotEncoder thuộc tính dạng Categorical là Smoker mà Smoker có 2 giá trị là yes và no nên sẽ thành 2 cột là 5.

4.2.2 Mô hình Multi-Layer Perceptron

- Chọn các siêu tham số để huấn luyện mô hình.
  - Hàm kích hoạt: relu.
  - Phương pháp đạo hàm: lbfgs.
  - Hệ số alpha: 0.05, 0.1, 1.
  - Số lớp ẩn: (80,), (100,) (80,2)

  Sử dụng GridSearchCV của sklearn để chọn ra siêu tham số alpha và số lớp ẩn tốt nhất để huấn luyện mô hình bằng cách sử dụng CrossValidation để tính toán độ lỗi từ đó chọn ra siêu tham số cho độ lỗi tốt nhất.
- Sau khi huấn luyện mô hình xong, ta có siêu tham số tốt nhất là:
  - alpha: 0.05.
  - Số lớp ẩn: (80,).

4.2.3 Random Forest Regressor.

- Chọn các siêu tham số để huấn luyện mô hình.
  - Số lượng cây: 200, 500
  - Độ sâu lớn nhất của cây: 4, 5, 6, 7, 8.
  - Lấy tham số mặc định của mô hình.

Sử dụng GridSearchCV của sklearn để chọn ra siêu tham số để huấn luyện mô hình bằng cách sử dụng CrossValidation để tính toán độ lỗi từ đó chọn ra siêu tham số cho độ lỗi tốt nhất.

- Sau khi huấn luyện mô hình xong, ta có siêu tham số tốt nhất là:
  - Số lượng cây: 200.
  - Độ sâu lớn nhất của cây: 4.

5 Báo cáo kết quả đạt được và nhận xét.

- Kết quả sau phân tích
  - Giữ lại cột Age, BMI, Children.
  - Xóa cột Sex, Region.
  - Xử lý giá trị nhiễu ở cột BMI.
- Kết quả huấn luyện mô hình.
  Dùng công thức R^2 để tính độ lỗi của từng mô hình.

Độ lỗi trên tập train	**Độ lỗi trên tập test**
0.7441585658576718	0.7653688584061046

Table 1: Mô hình Linear Regression

Độ lỗi trên tập train	**Độ lỗi trên tập test**
0.8874236601959798	0.8585035765971413

Table 2: Mô hình Multi-Layer Perceptron

Độ lỗi trên tập train	**Độ lỗi trên tập test**
0.8756666079774867	0.8581463981930906

Table 3: Mô hình Random Forest Regressor

- Nhận xét

- Cột Age, BMI, Children có mối tương quan thuận với chi phí y tế nhưng không mạnh (lần lượt 0.3, 0.2, 0.1 theo độ đo Correlation.)
- Bằng phương pháp kiểm định thống kê ta có thể thấy cột Sex, Region không ảnh hưởng nhiều đến chi phí y tế nên ta có thể loại bỏ 2 cột đó.
- Cột Smoker có ảnh hưởng rất lớn đến chi phí y tế.
  * Nếu có hút thuốc thì chi phi y tế sẽ rất cao.
  * Nếu không hút thuốc thì chi phí y tế sẽ thấp hơn.
- Thấy được cột BMI có giá trị nhiễu, từ đó có thể tiền xử lý dữ liệu giúp huấn luyện mô hình cho kết quả tốt hơn.

6 Tài liệu tham khảo.

- Thư viện sử dụng trong đồ án:
  - pandas.
  - seaborn.
  - matplotlib.
  - statsmodels.
  - numpy.