Axis Insurance

Nibu Kuriakose

Mar 2021

Content

No	Item	Slide #
1.	Exec Summary	3
2.	Dataset Information	4
3.	Analysis 3.1 Univariate Analysis 3.2 Multivariate Analysis	 5-7 8-15
4.	Hypothesis Testing	16-19

1. Executive Summary

Background

- Axis Insurance has gathered key data of its customers.
- The company wants to use this data for making business decisions.

Purpose

- Perform exploratory data analysis to derive insights, and prove or disprove certain hypothesis using the data provided

Key Findings

Exploratory Data Analysis

- Customer base not having ideal weight- 83.4% of customers doesn’t have an ideal weight, with 81.9% over ideal bmi and rest under.
- 75% of the customers are between 18 and 51 years old.
- 43% of customers do not have dependents. The rest have up to 5 dependents.
- BMI and charges highest for southeast, followed by southwest. Males have more charges than females.
- **Smoking**
  - ✓ 20% of customers are smokers. Southeast has highest number of smokers.
  - ✓ Smoker have much higher charges than non-smokers.
  - ✓ As BMI increases, charges increases for smokers.
  - ✓ Male smokers have higher bmi than male non-smokers, but for females it is the other way around. Male smokers have slightly higher bmi than female smokers.

Hypothesis Testing- Following can be confirmed with 95% confidence:

- Medical claims made by smokers is greater than claims by non-smokers.
- BMI of males and females are equal
- Proportion of smokers is not significantly different across regions i.e. proportion of smokers independent of region.
- Mean BMI of women with zero, one and two children are equal

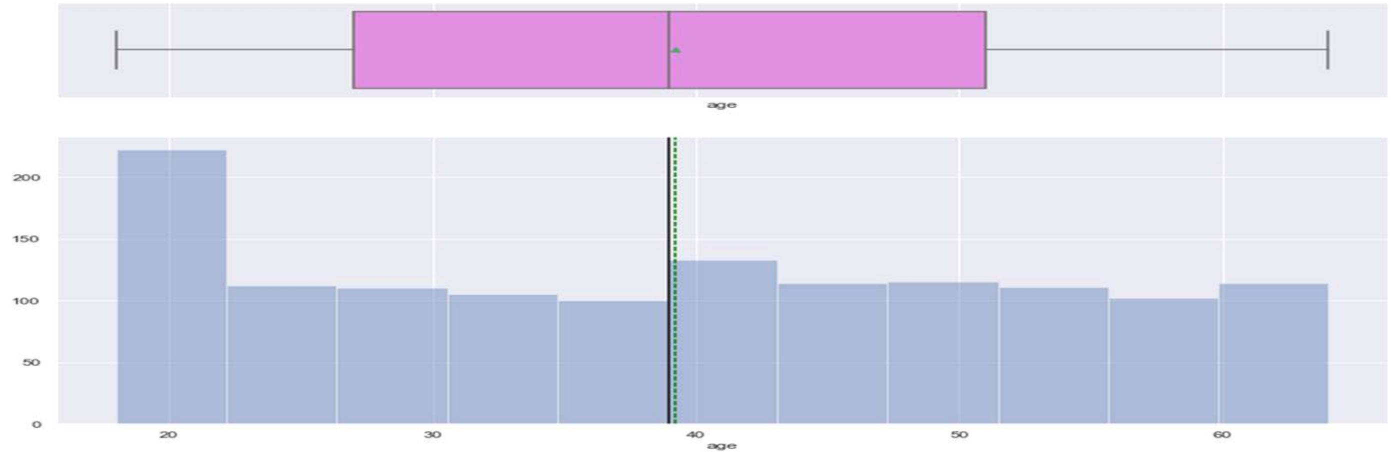
2. Dataset Information

- There are 1,338 samples
- Each sample has 7 attributes which are given below
- There is no null value in the dataset

Attribute	Description	Type
Age	Age of the primary beneficiary	Integer
Sex	Policy holder's gender	Object
BMI	Body mass index (ideal BMI is within the range of 18.5 to 24.9)	Float
Children	Number of children / dependents covered by the insurance plan	Integer
Smoker	Yes or No depending on whether the insured regularly smokes tobacco.	Object
Region	Place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest	Object
Charges	Individual medical costs billed to health insurance	Float

3.1 Univariate Analysis (I/III)

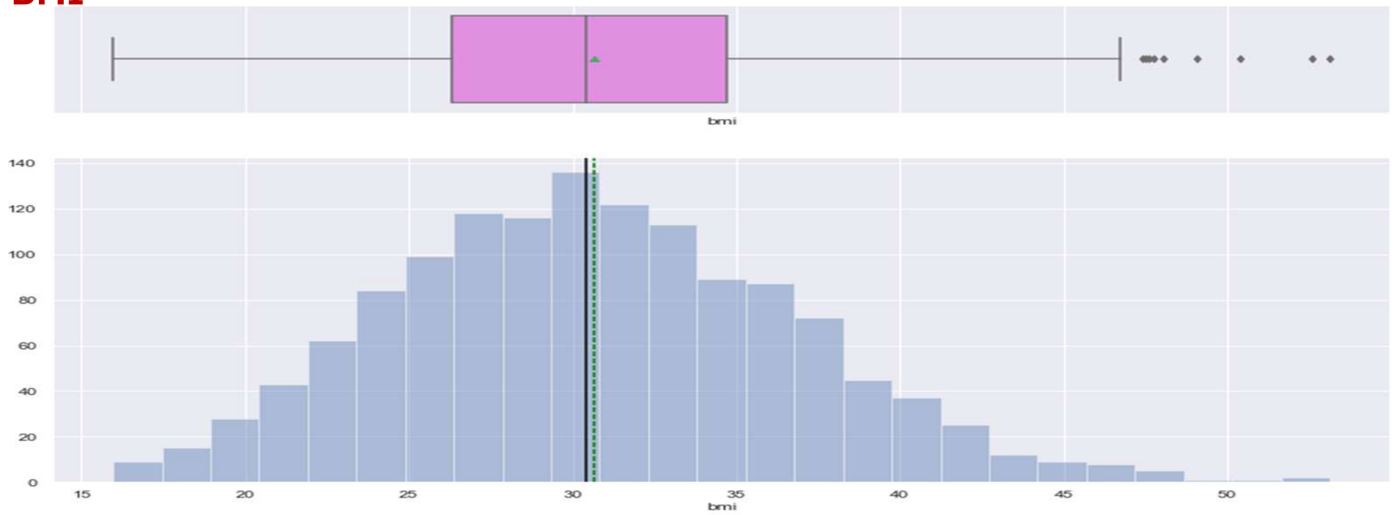
AGE



count	1338.0
mean	39.2
std	14.0
min	18.0
25%	27.0
50%	39.0
75%	51.0
max	64.0
mode	18.0
IQR	24.0
Range	46.0
Coeff. of variance	36%

- 75% of the customers are between 18 and 51 years old, and most bought by 18 year old.
- No notable skewness (mean and median very close) with a range of 46 years and coeff. of variation of 36%

BMI

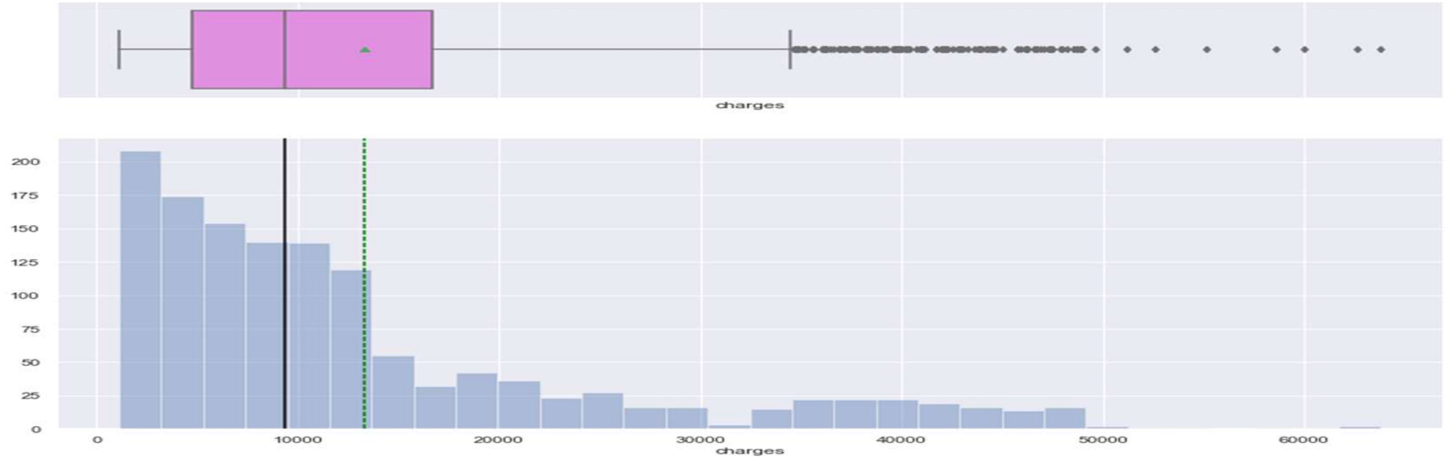


count	1338.0
mean	30.7
std	6.1
min	16.0
25%	26.3
50%	30.4
75%	34.7
max	53.1
mode	32.3
IQR	8.4
Range	37.2
Coeff. of variance	20%

- 83.4% of customers doesn’t have an ideal weight- 81.9% over 24.9 bmi and rest under 18.5.
- No notable skewing (mean and median very close) with range of 37.2 and coeff. of variation of 20%.

3.1 Univariate Analysis (II/III)

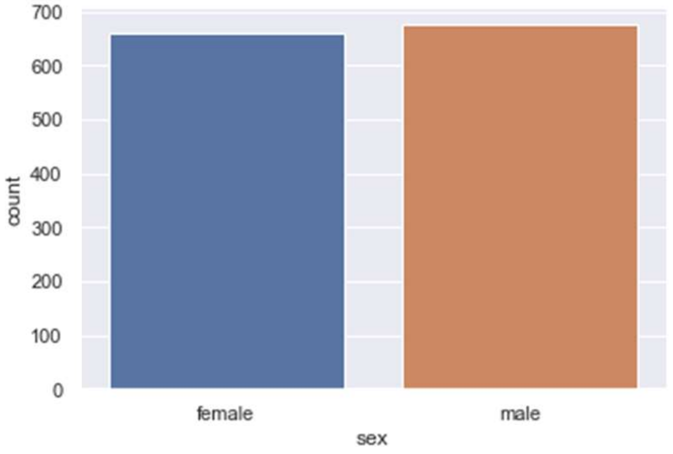
CHARGES

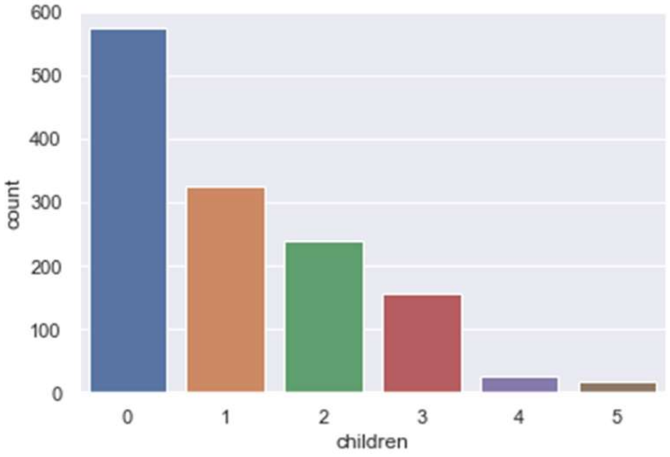


count	1,338
mean	13,270
std	12,110
min	1,122
25%	4,740
50%	9,382
75%	16,640
max	63,770
mode	1,640
IQR	11,900
Range	62,649
Coeff. of variance	91%

- 75% of the customers charges are less than $16.6k and 50% below $9.3k.
- Right skewed data. High spread with coefficient of variation of 91%.
- There are a number of outliers above the upper whisker.

GENDER AND CHILDREN/DEPENDENTS

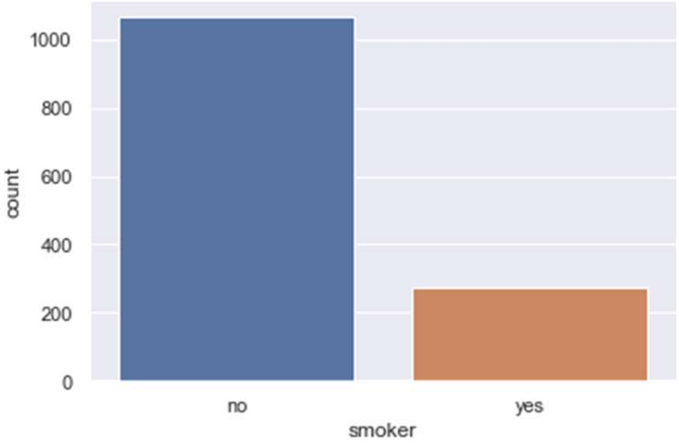


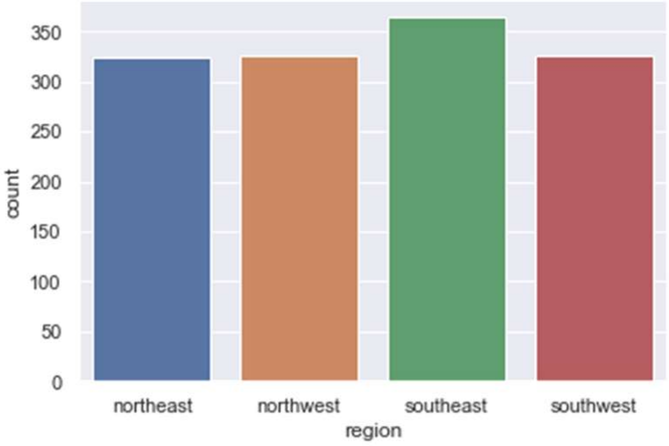


- Almost same number of men (676) and women (662)
- Customers with no children/dependents is the highest (43%).
- The highest number of children/dependent is 5.

3.1 Univariate Analysis (III/III)

SMOKER AND REGIONS

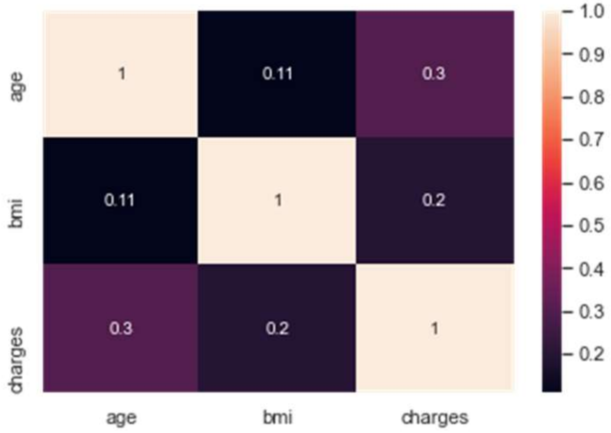




- 20% (274/1,388)of customers smoke.
- The number of customers is almost equally distributed between the four regions, with Southeast having slightly higher number of customers than the rest.

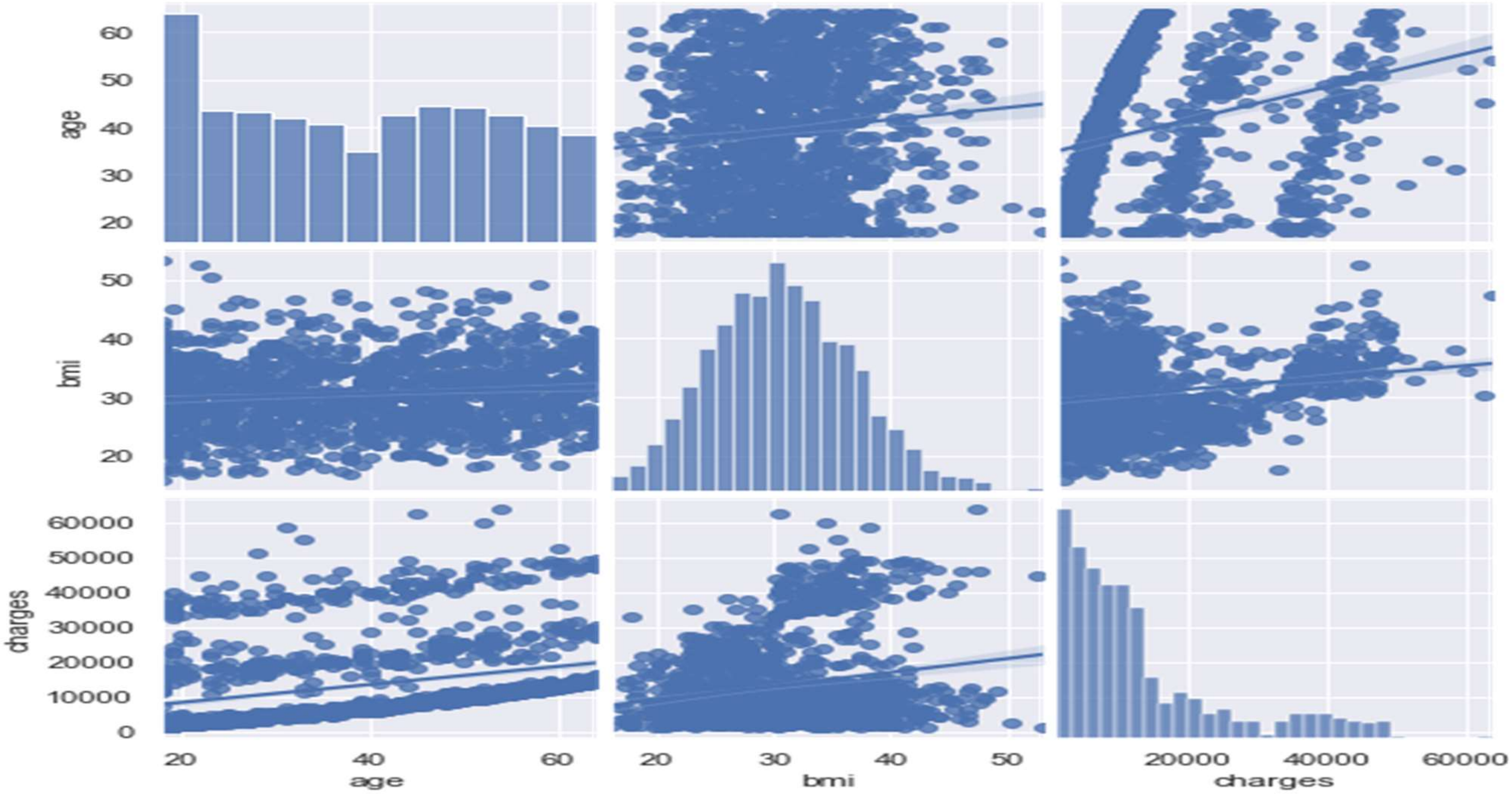
3.2 Multivariate Analysis

HEATMAP



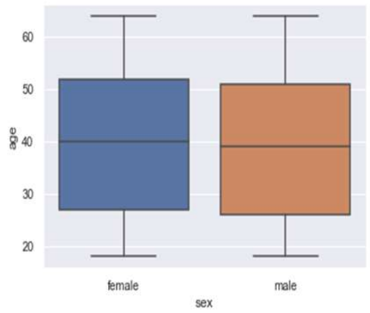
- No high co-relation between attributes.

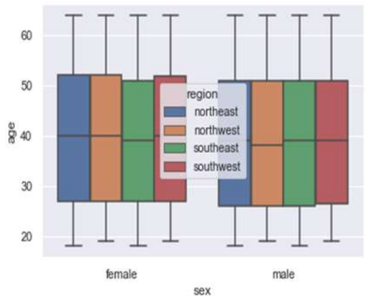
3.2 Multivariate Analysis

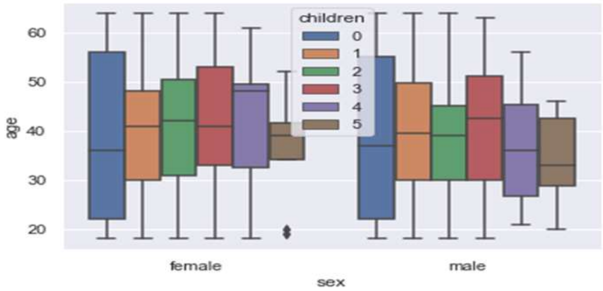


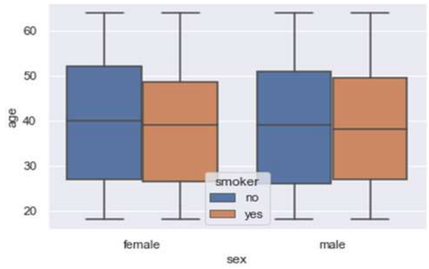
- There are young customers with high income

3.2 Multivariate Analysis

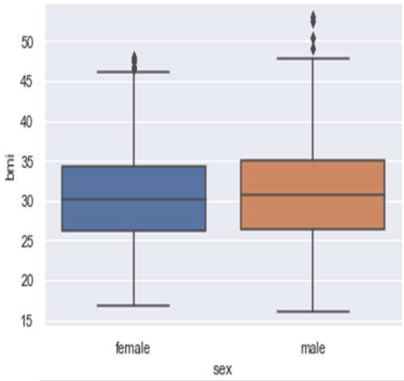


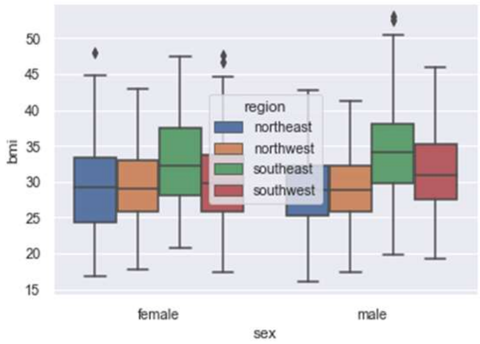


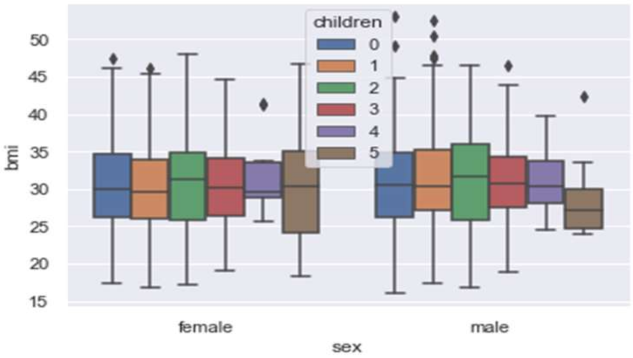


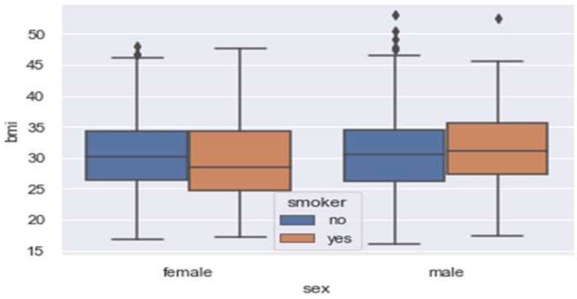


- Male and female customers have almost similar age across smokers and across regions.
- Men and female customers have almost similar age when they have 0 to 2 dependents. Men with 3 children have average age higher than women and the other way round for 4 or 5 children.



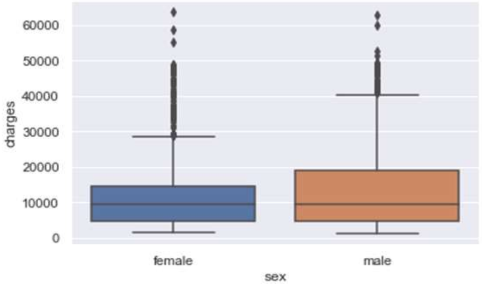


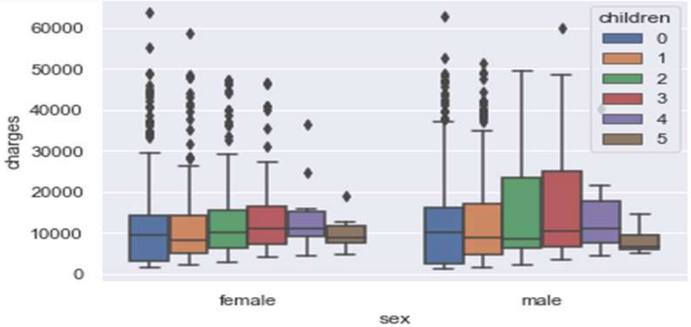


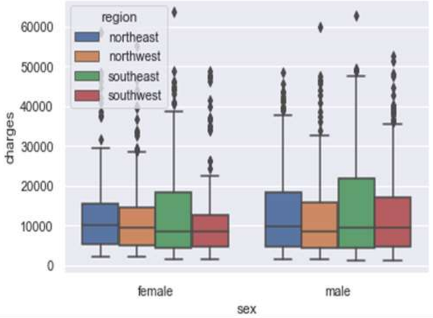


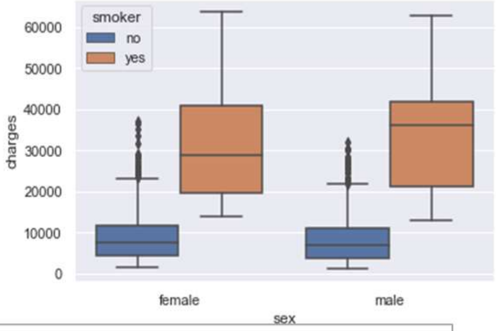
- Male smokers have higher bmi than male non-smokers, but for females it is the other way around. Male smokers have slightly higher bmi than female smokers.
- BMI for southeast is highest, followed by southwest. Males and females have similar bmi within a region.
- BMI for customers with 2 children seem to be the highest
- Female BMI with 5 dependents is much higher than men.

3.2 Multivariate Analysis

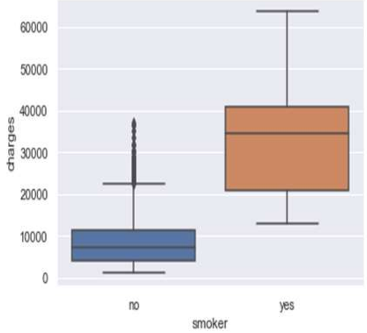


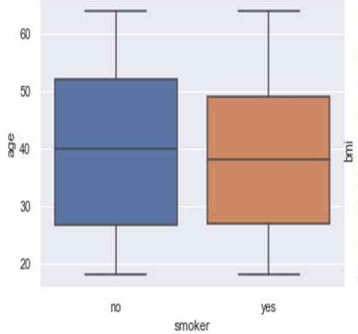


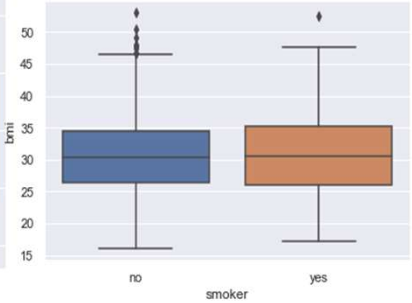


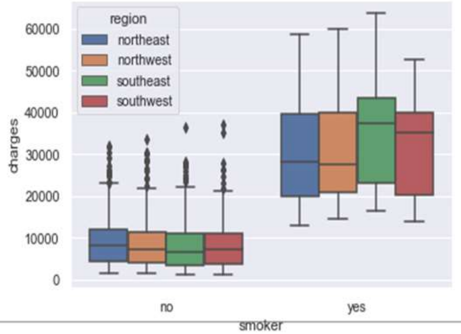


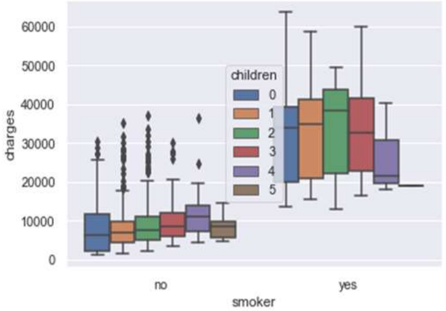
- Males have more charges than female.
- Charges go up until 3 dependents and then falls
- Charges highest in southest compared to other regions
- Smoker have much higher charges than non-smokers. For non-smokers, males have higher charges than female.





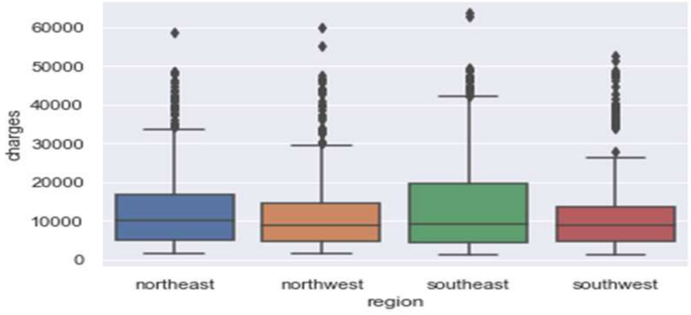


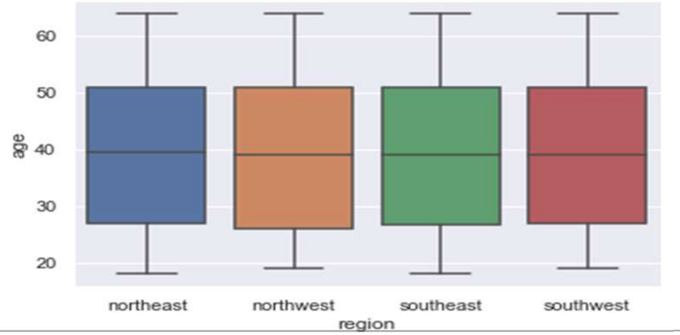


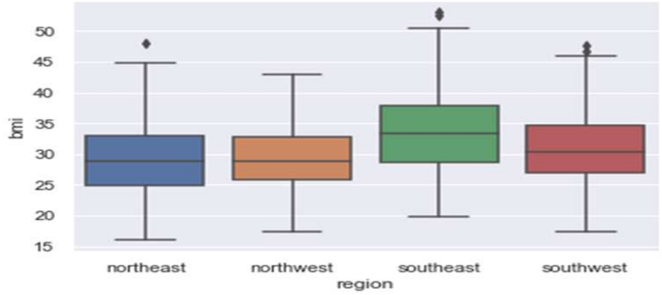


- Smokers have higher charges than non-smokers
- Smokers slightly younger than non-smokers
- Smokers have slightly higher bmi
- Smokers in southeast has highest charges followed by southwest
- Smokers with 2 dependents have the highest charges, and least with 5 children.

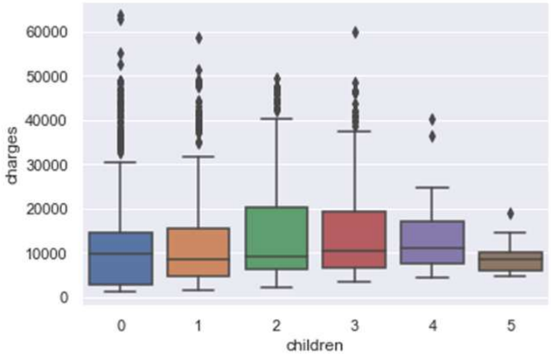
3.2 Multivariate Analysis

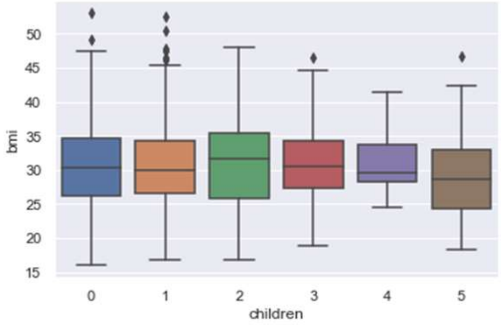


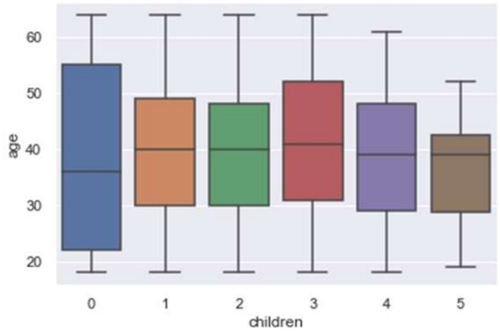




- Charges and bmi highest in southeast. Rest have similar bmi and charges.

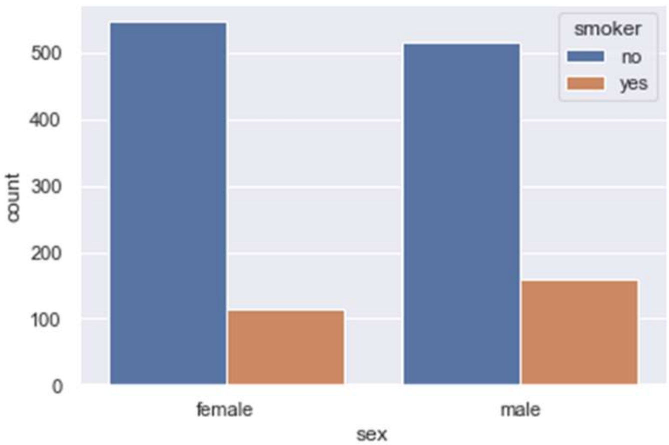


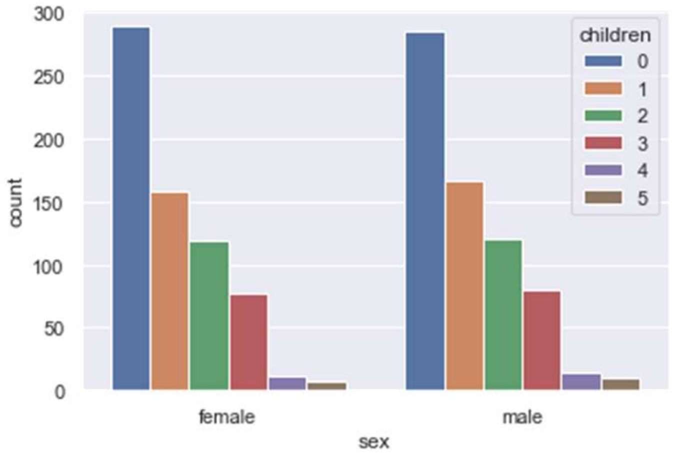


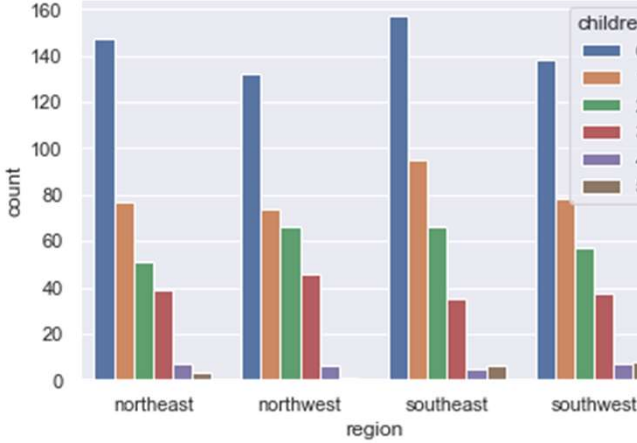


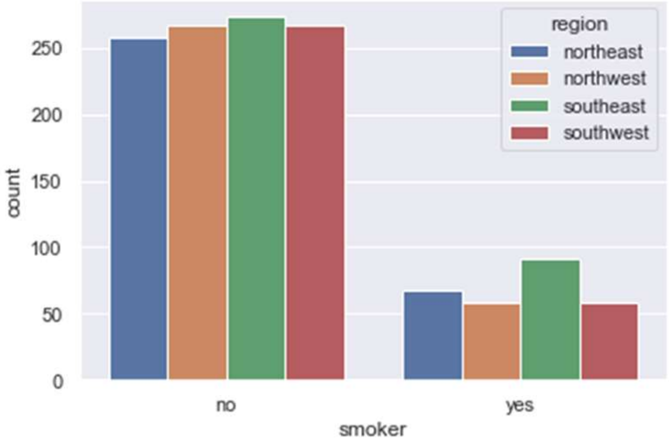
- Charges go up until 3 dependents and then falls
- Age of customers with 3 dependents is the highest

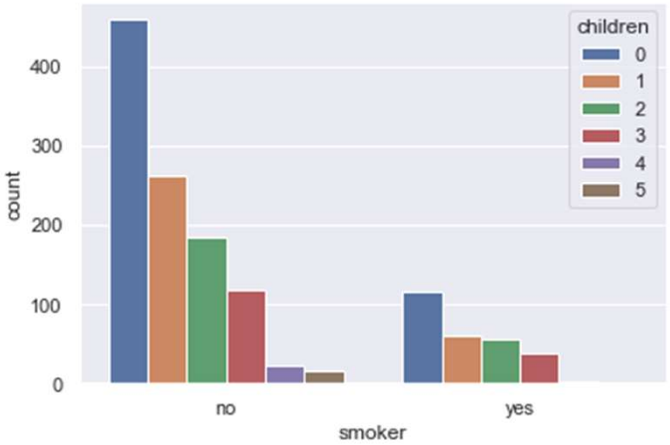
3.2 Multivariate Analysis

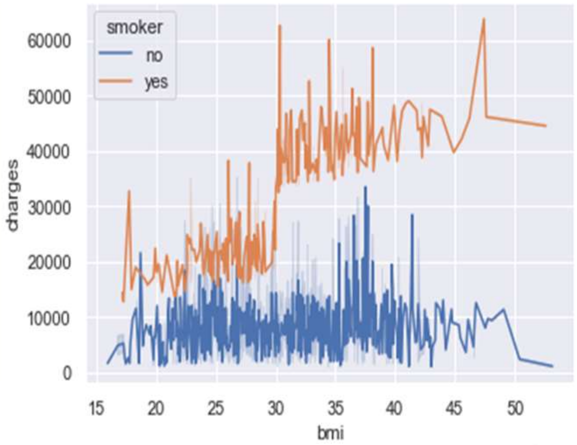






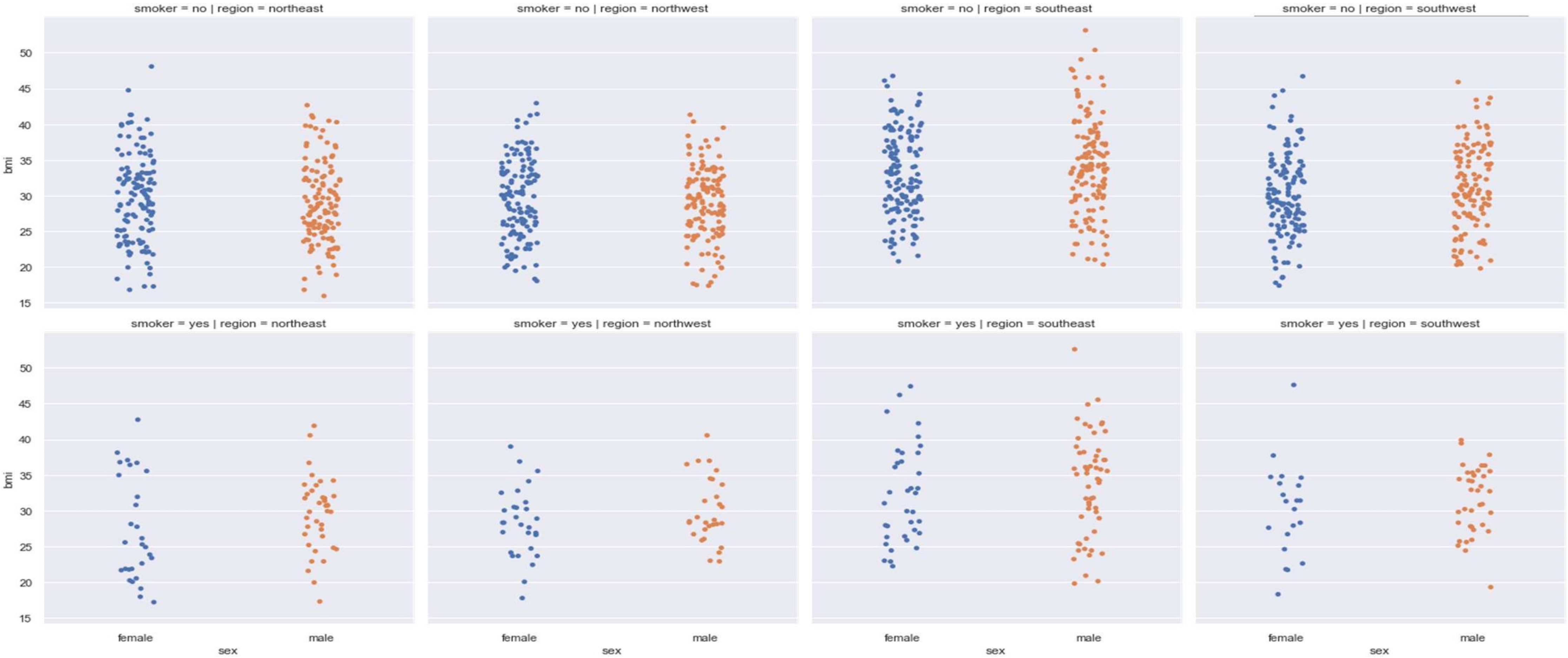






- Slightly more male smokers than females
- Southeast has highest number of smokers
- As BMI increases, charges increases for smokers

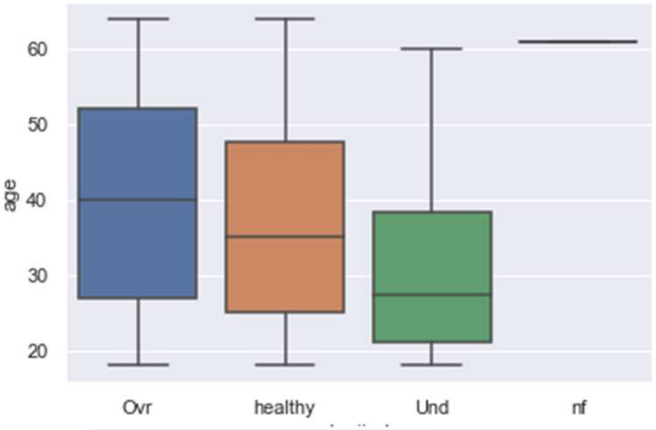
3.2 Multivariate Analysis

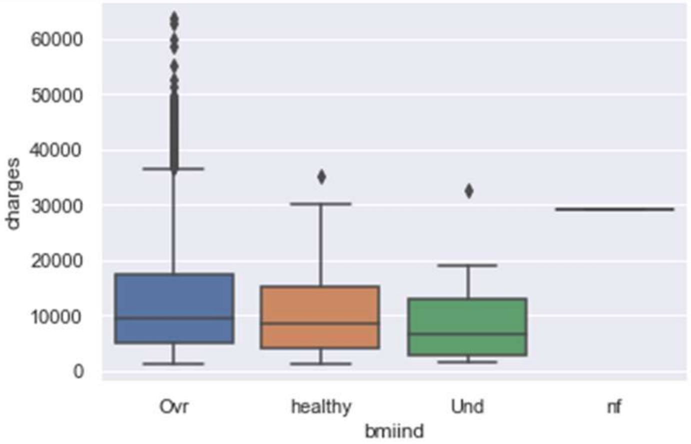


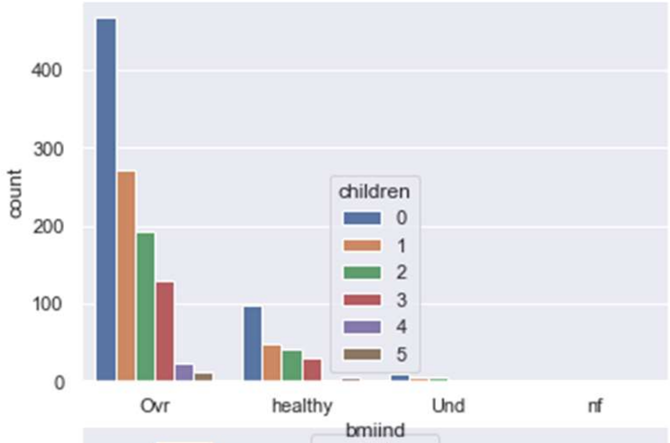
- BMI highest for southeast irrespective of their gender or whether they smoke or not, compared to other regions

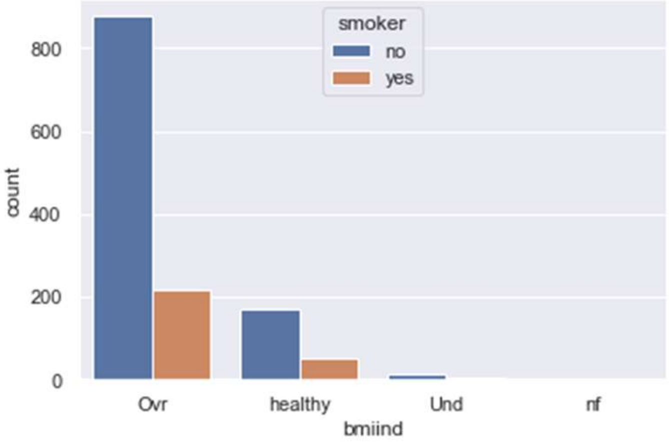
3.2 Multivariate Analysis

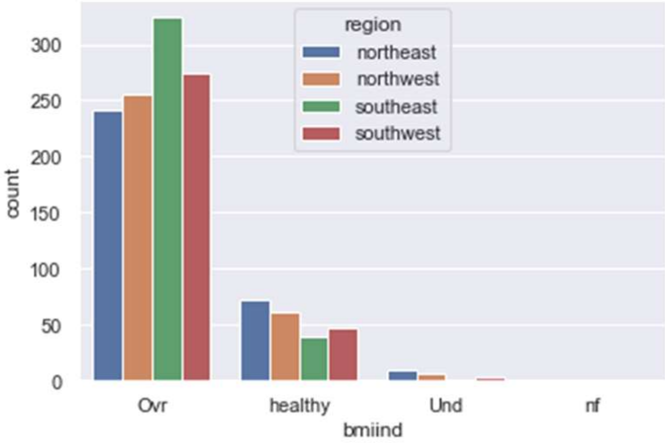
Ovr- BMI above 24.9
Healthy- BMI between 18.5 and 24.9
Und- BMI below 18.5

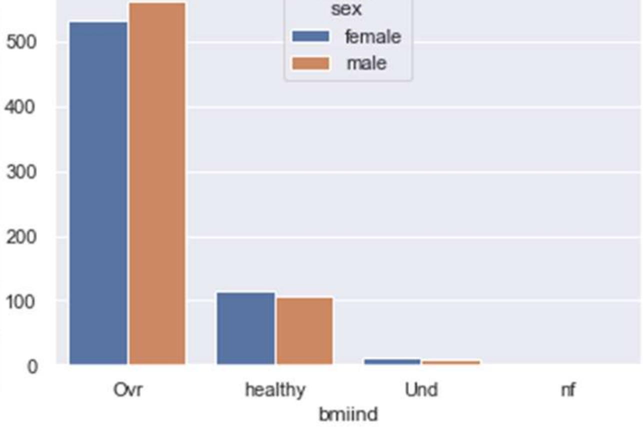








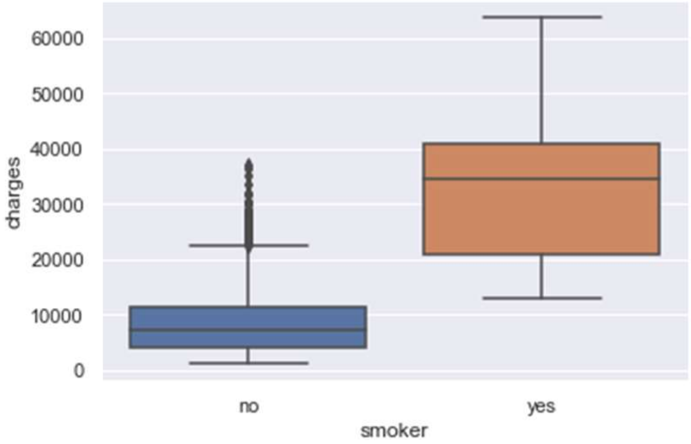




- The charges decrease with BMI
- BMI increases with age
- Smokers are not below ideal BMI
- South east highest number of high BMI followed by southwest
- Slightly higher males with high BMI than females. It is other way round for healthy and low BMI.

4.1 Hypothesis Testing (question # 2)

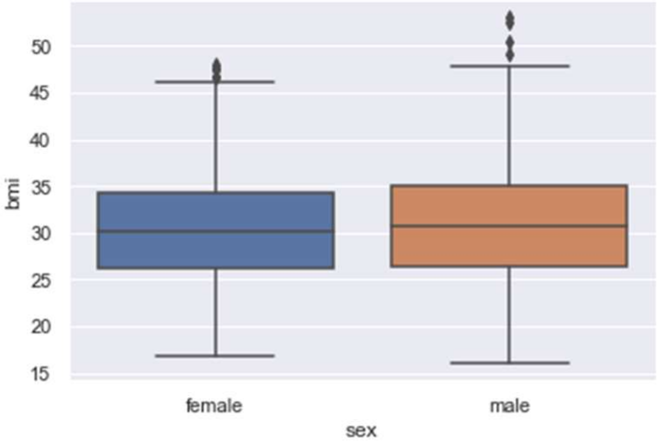
Steps	Summary
Hypothesis	H_0- Mean medical claims made by smokers is less than or equal to claims by non-smokers H_a- Mean medical claims made by smokers is greater than claims by non-smokers
Test	Two sample t-test (one tail)
Significance Level (alpha)	0.05
Decision	If p< alpha, reject H_0 If p >= alpha, fail to reject H_0
Result and Inference	
P value	8.27 x10^{-283}
Pvalue/2 (as one tail test)	4.13 x10^{-283}
Result	P (4.13 x10^{-283}) < alpha (0.05). Hence, reject H_0
Inference	Support the alternate hypothesis that medical claims made by smokers is greater than claims by non-smokers



- This box plot supports the above inference as the mean and distribution of the charges of smokers are different from that of the non-smokers.

4.2 Hypothesis Testing (question # 3)

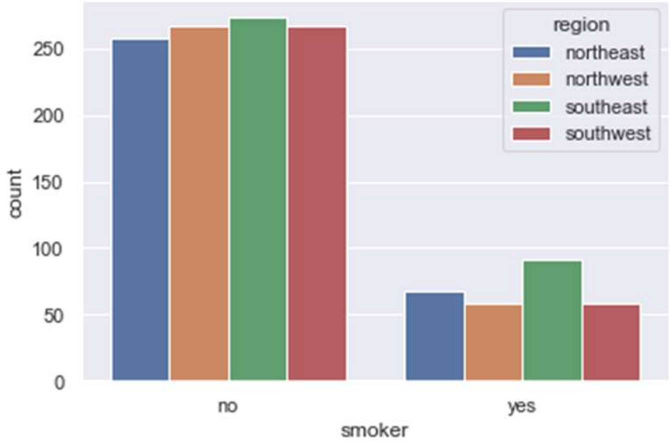
Steps	Summary
Hypothesis	H_0- BMI of males and females are equal H_a- BMI of males and females are not equal
Test	Two sample t-test (two tail)
Significance Level (alpha)	0.05
Decision	If p< alpha, reject H_0 If p >= alpha, fail to reject H_0
Result and Inference	
P value	0.08997637178984932
Result	P (0.089) > alpha (0.05). Hence, fail to reject H_0
Inference	We cannot reject that BMI of males and females are equal



- This box plot supports the above inference as the mean and distribution of the bmi is similar for men and women

4.3 Hypothesis Testing (question # 4)

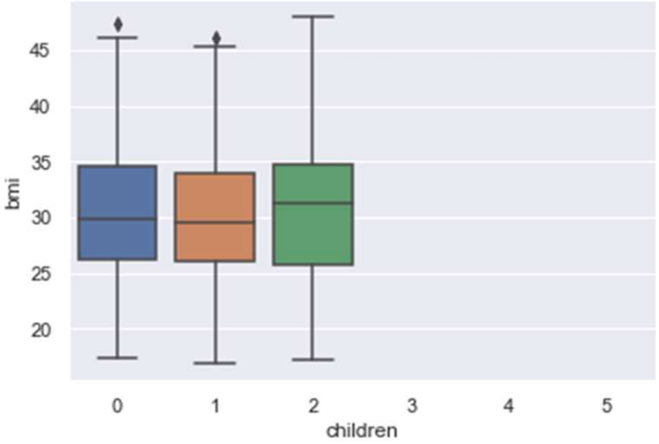
Steps	Summary
Hypothesis	H_0- Proportion of smokers is not significantly different across regions (proportion of smokers independent of regions) H_a- Proportion of smokers is significantly different across regions (proportion of smokers dependent of regions)
Test	Chi2 contingency
Significance Level (alpha)	0.05
Decision	If p< alpha, reject H_0 If p >= alpha, fail to reject H_0
Result and Inference	
P value	0.06171954839170547
Result	P (0.061) > alpha (0.05). Hence, fail to reject H_0
Inference	We cannot reject that the ‘proportion of smokers is not significantly different across regions’ (proportion of smokers independent of region)



- This box plot supports the above inference as the proportion of smokers across regions are broadly the same.

4.4 Hypothesis Testing (question # 5)

Steps	Summary
Hypothesis	H_0- Mean BMI of women with zero, one and two children are equal H_a- Mean BMI of women with zero, one and two children differs
Test	One way ANOVA test
Significance Level (alpha)	0.05
Decision	If p< alpha, reject H_0 If p >= alpha, fail to reject H_0
Result and Inference	
P value	0.715858
Result	P (0.715) > alpha (0.05). Hence, fail to reject H_0
Inference	We cannot reject that mean BMI of women with zero, one and two children are equal



- This box plot supports the above inference as the mean and distribution of the bmi of women with 0 to 2 dependents are the same.