Axis Insurance Project

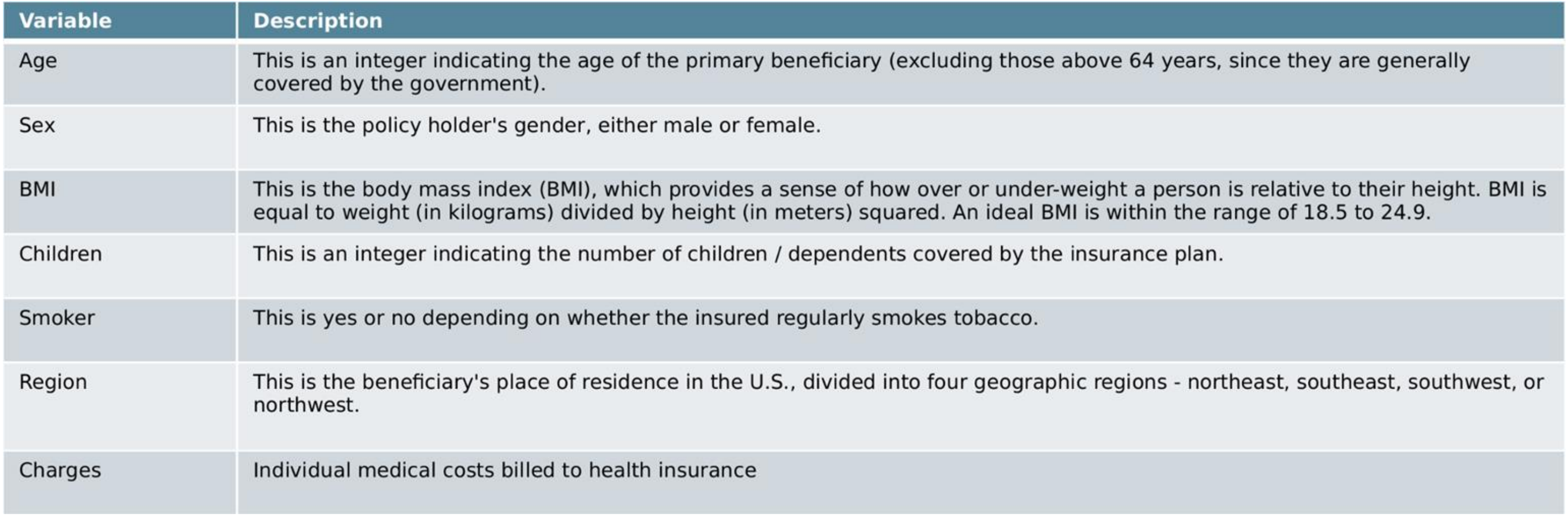
By Naeem Sufi

Objective

Statistical Analysis of Business Data. Explore the dataset and extract insights from the data. We want to get comfortable doing statistial analysis using Python.

- Explore the dataset and extract insights using Exploratory Data.
- Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?
- Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.
- Is the proportion of smokers significantly different across different regions?
- Is the mean BMI of women with no children, one child, and two children the same? Explain your answer with statistical evidence.
- Consider a significance level of 0.05 for all tests.

Dataset Description

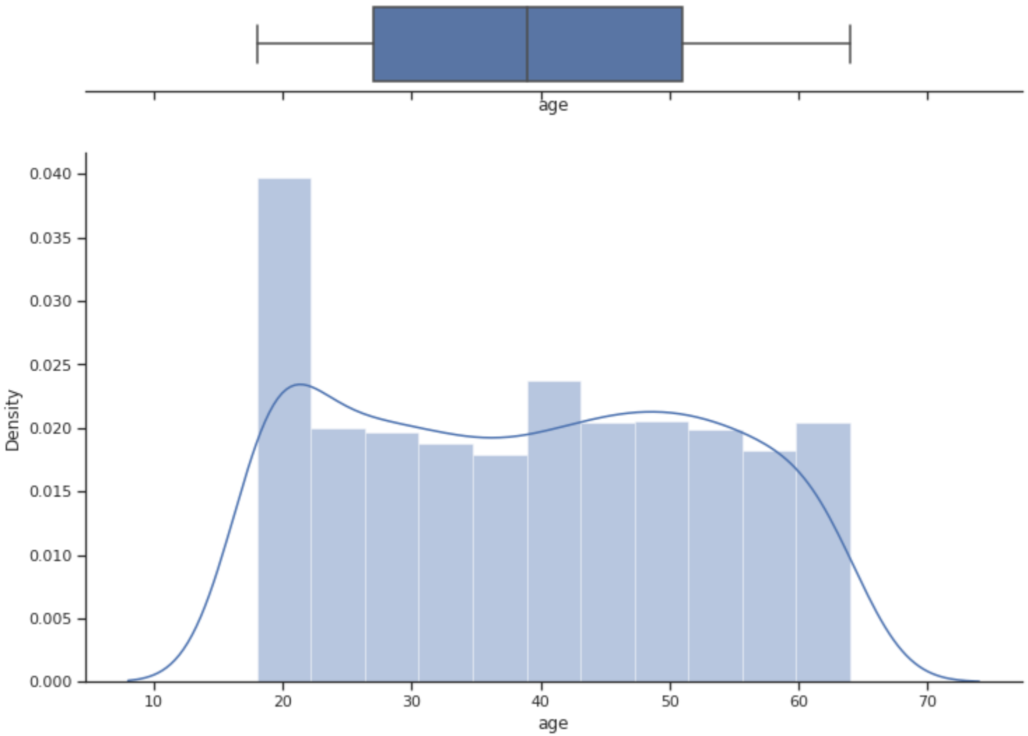


Variable	Description
Age	This is an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
Sex	This is the policy holder's gender, either male or female.
BMI	This is the body mass index (BMI), which provides a sense of how over or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
Children	This is an integer indicating the number of children / dependents covered by the insurance plan.
Smoker	This is yes or no depending on whether the insured regularly smokes tobacco.
Region	This is the beneficiary's place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest.
Charges	Individual medical costs billed to health insurance

Exploratory Analysis

Age

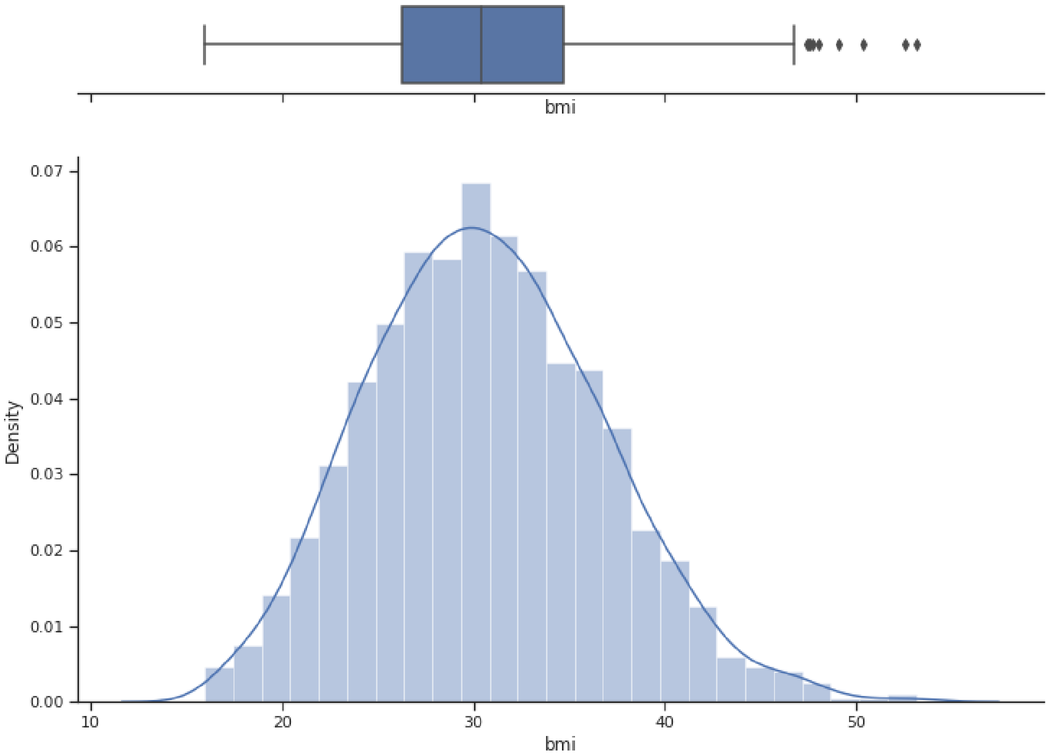
- Average age for people represented by Axis insurance is 39.
- Highest number of people represented by Axis Insurance is under the age of 22



Exploratory Analysis

BMI

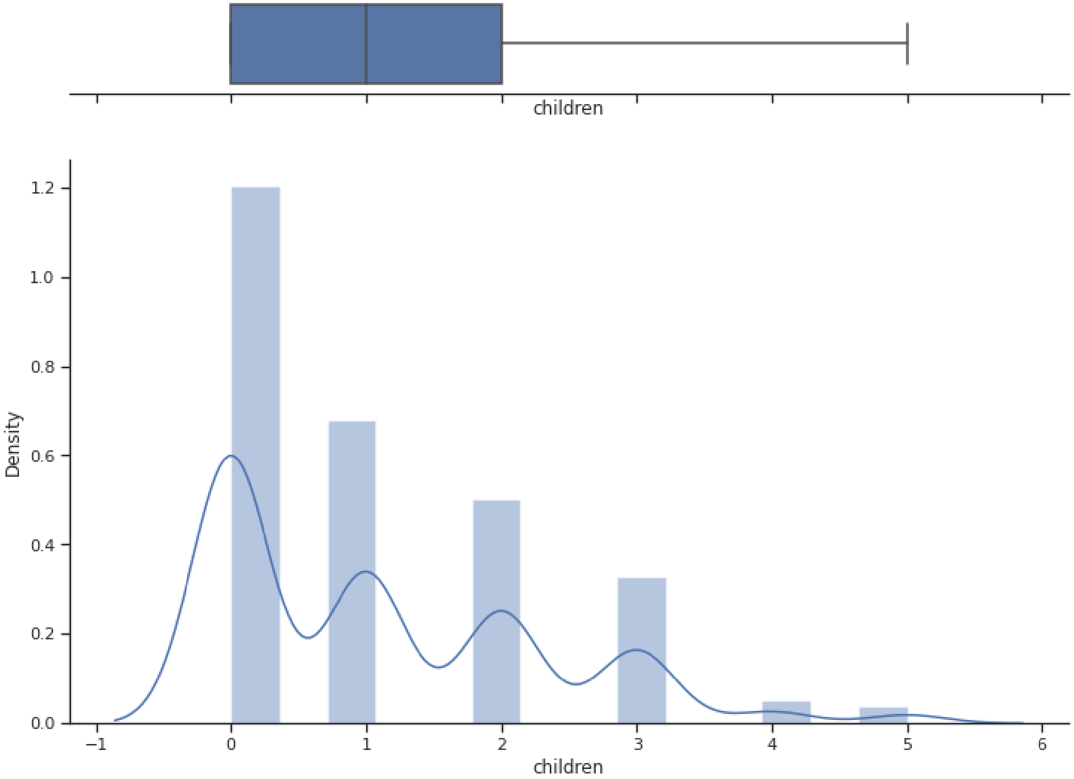
- Average BMI for the people represented by Axis insurance is 31. It means that most of the population in Axis insurance case is obese
- The highest frequency of people lie in range of 26 to 35



Exploratory Analysis

Children

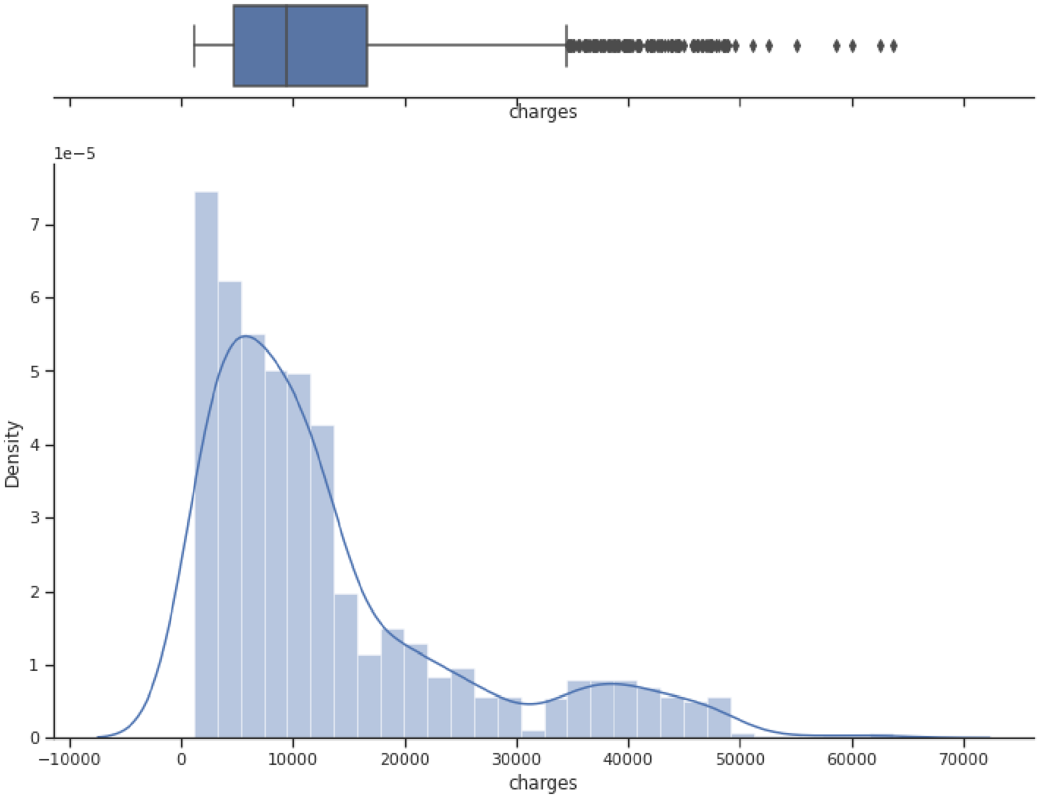
- On average each insurer have 1 child.
- There are very few outliers who have 4 to 5 children



Exploratory Analysis

Charges

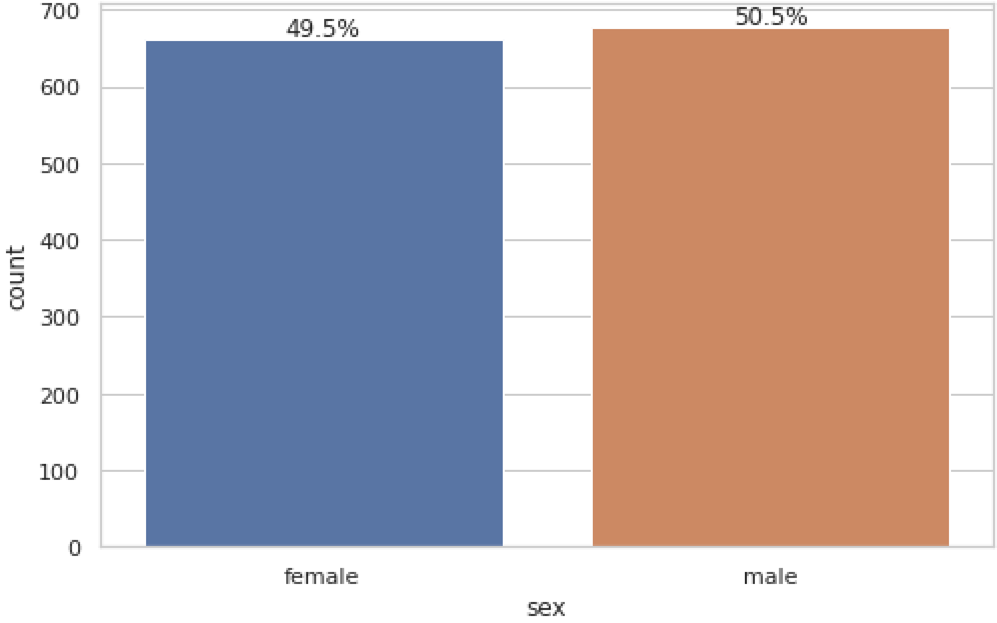
- There are more outliers towards higher end of insurance claim
- There are very few claims over $30,000
- On average $13,300 average insurance claim



Exploratory Analysis

Sex

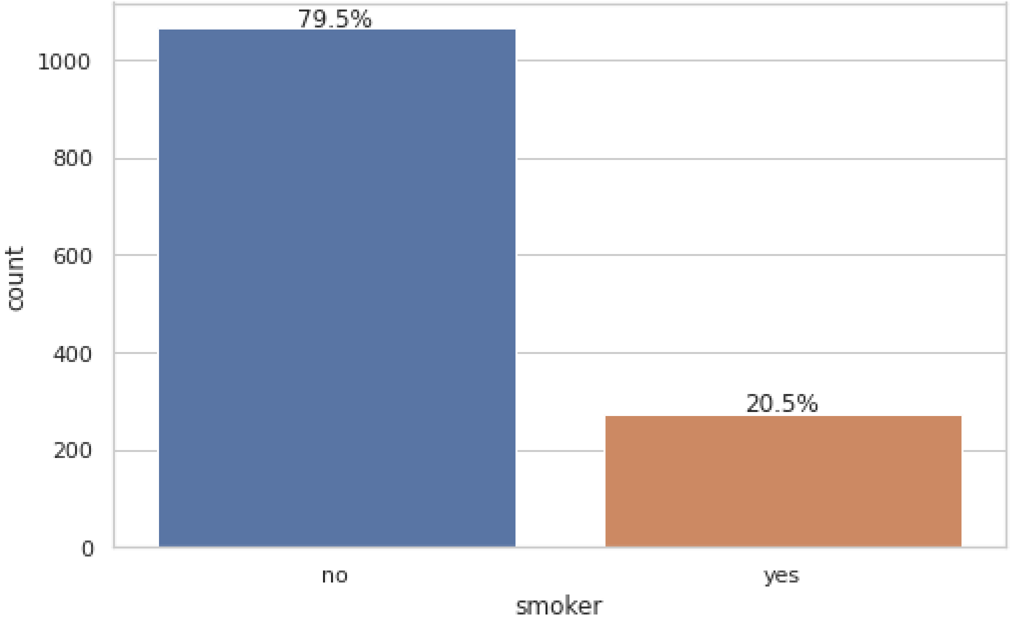
- The ratio of male to female insurer is almost the the same.
- The difference between male and female is 1%.



Exploratory Analysis

Smoker

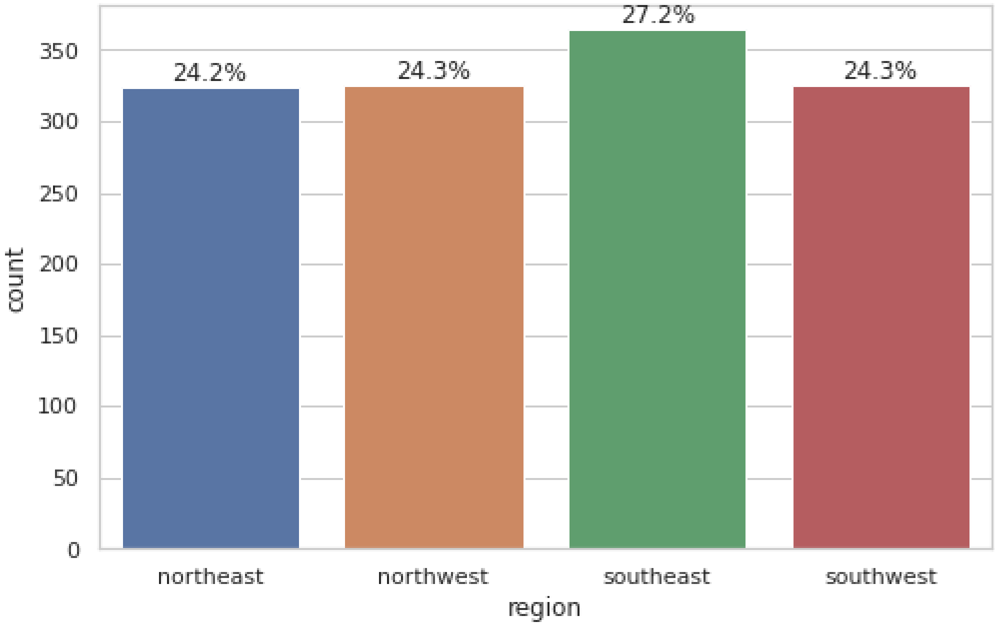
- Most of people represented by Axis are non smoker.
- The gap between smoker and non smoker is 59.5% in



Exploratory Analysis

Region

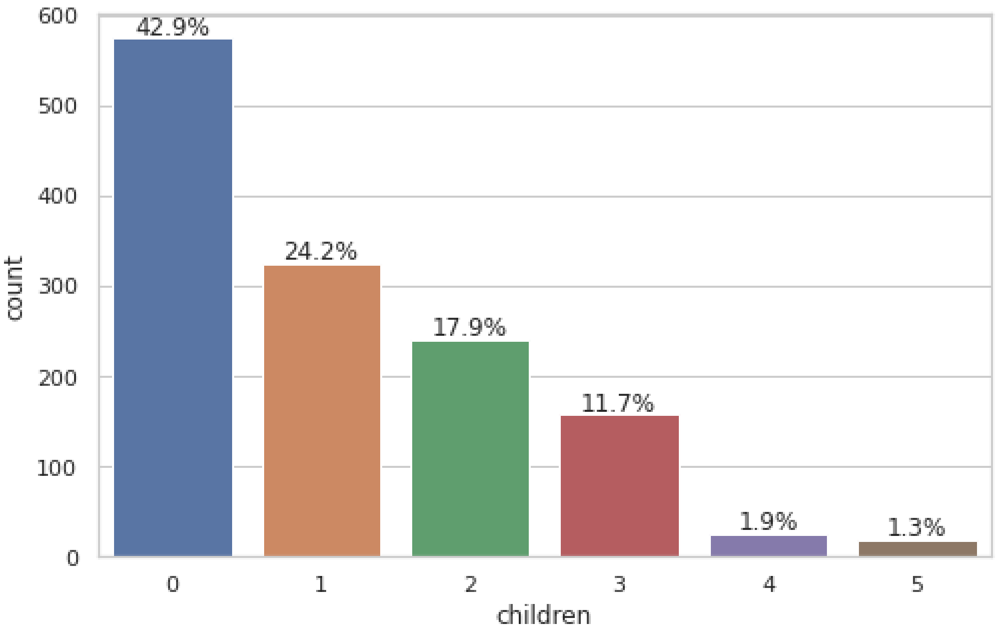
- The customer served by Axis are very evenly distributed around various regions.
- Only Southeast has 3% more customers than other regions



Exploratory Analysis

Children

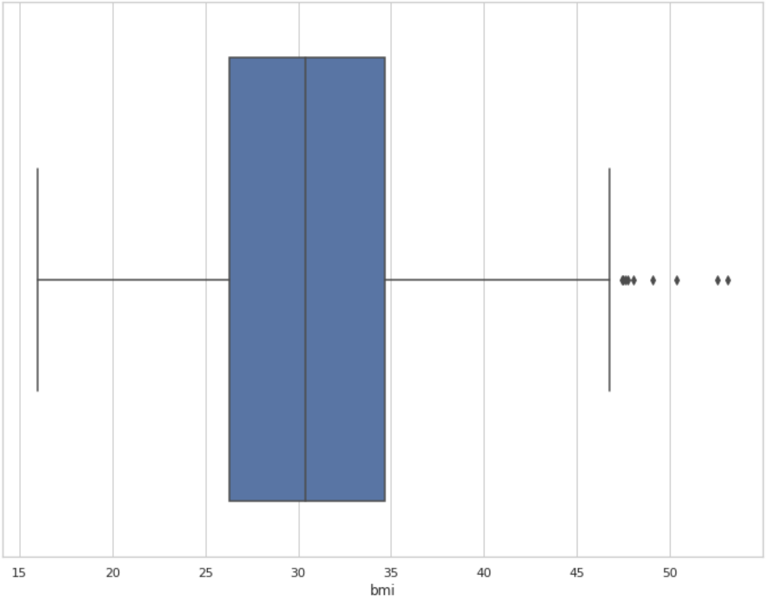
- Most of the insurers dont have children
- Very few of insurers have 4 or 5 children



Exploratory Analysis

Outliers in BMI

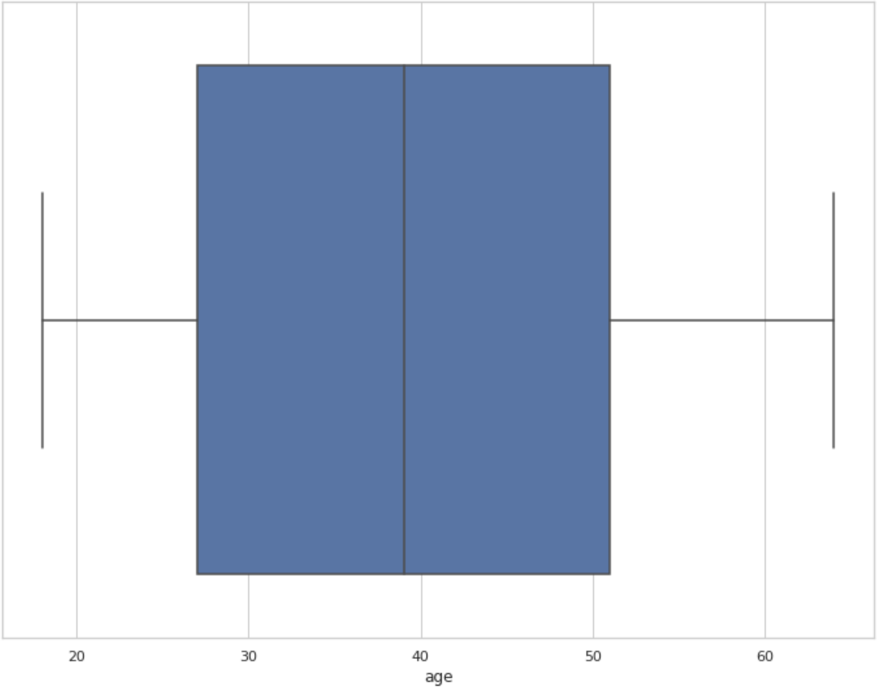
- BMI has outliers on right side of the chart on higher values



Exploratory Analysis

Outliers in Age

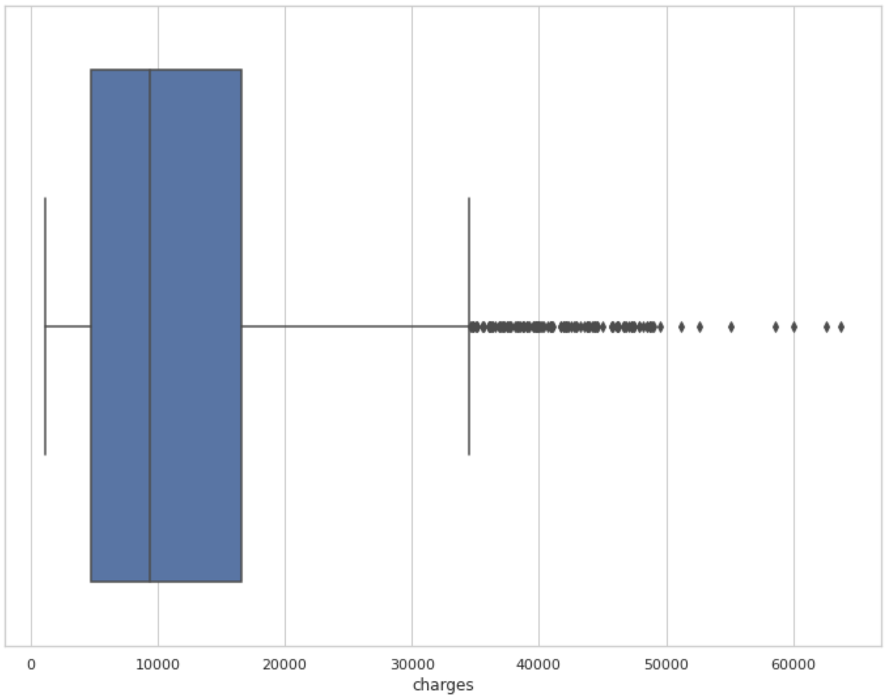
- There are no outliers in age column



Exploratory Analysis

Outliers in charge

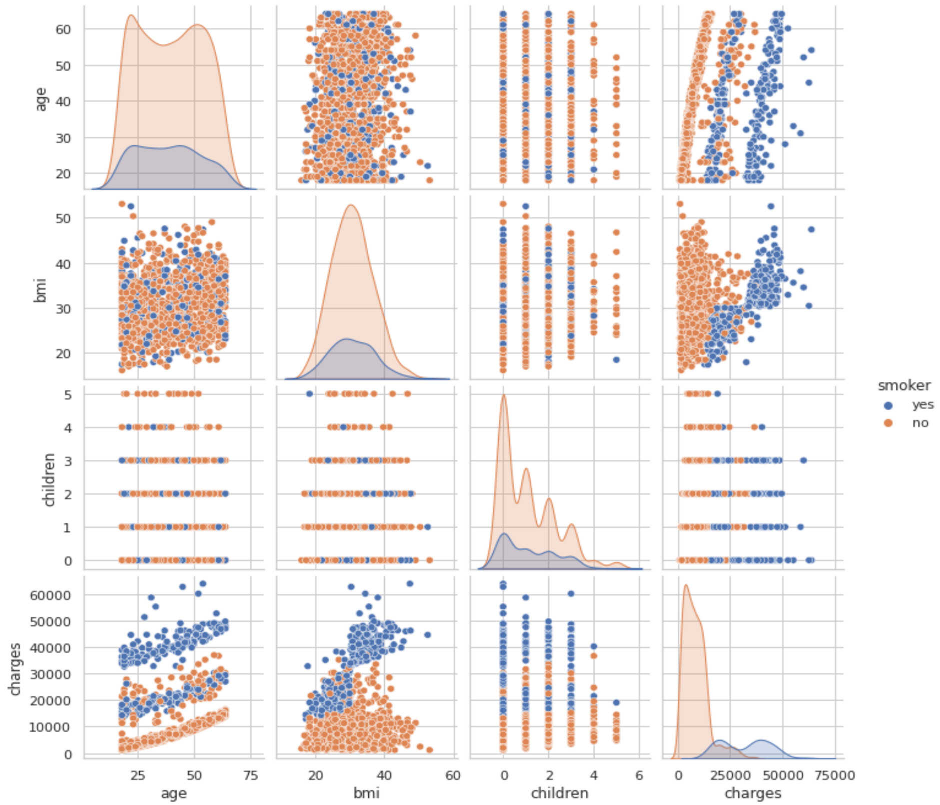
- Charges has outliers on right side of the chart on higher values



Exploratory Analysis

Pairplot with smoker hue

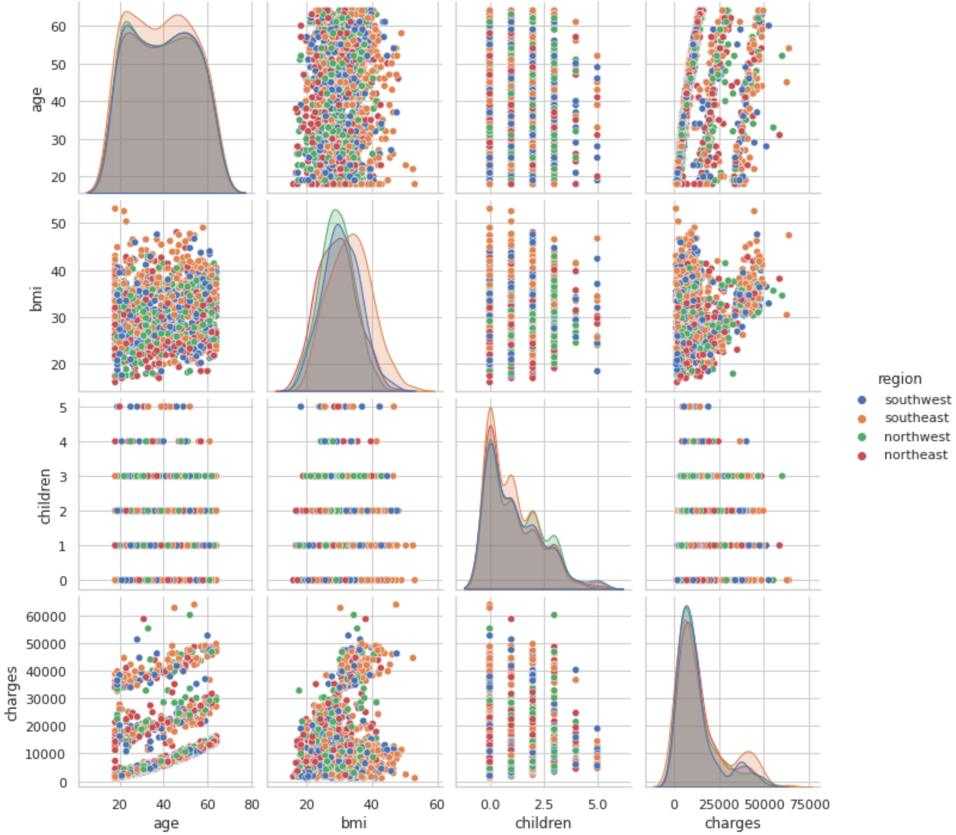
- More smokers are of young age
- BMI is higher for smokers
- Charge is higher for smokers
- Smokers have less number of children



Exploratory Analysis

Pairplot with sex hue

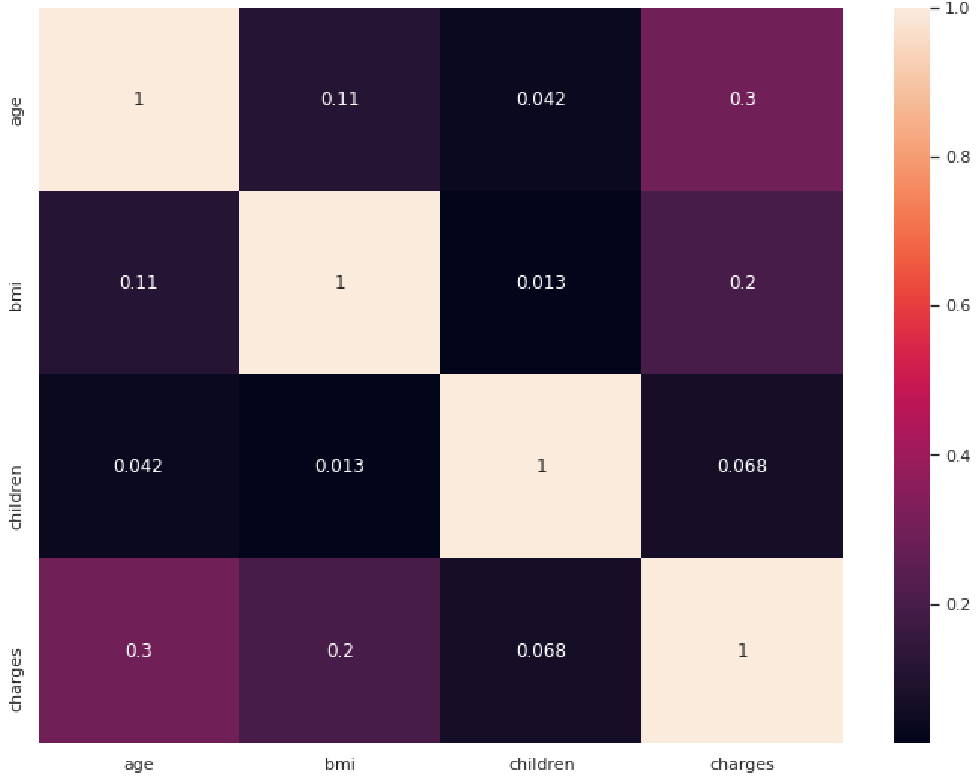
- Region is consistent with all other features.
- Southeast region shows higher values than others



Exploratory Analysis

Correlation Matrix

- Correlation Matrix shows that there is not really any strong correlations in the data.

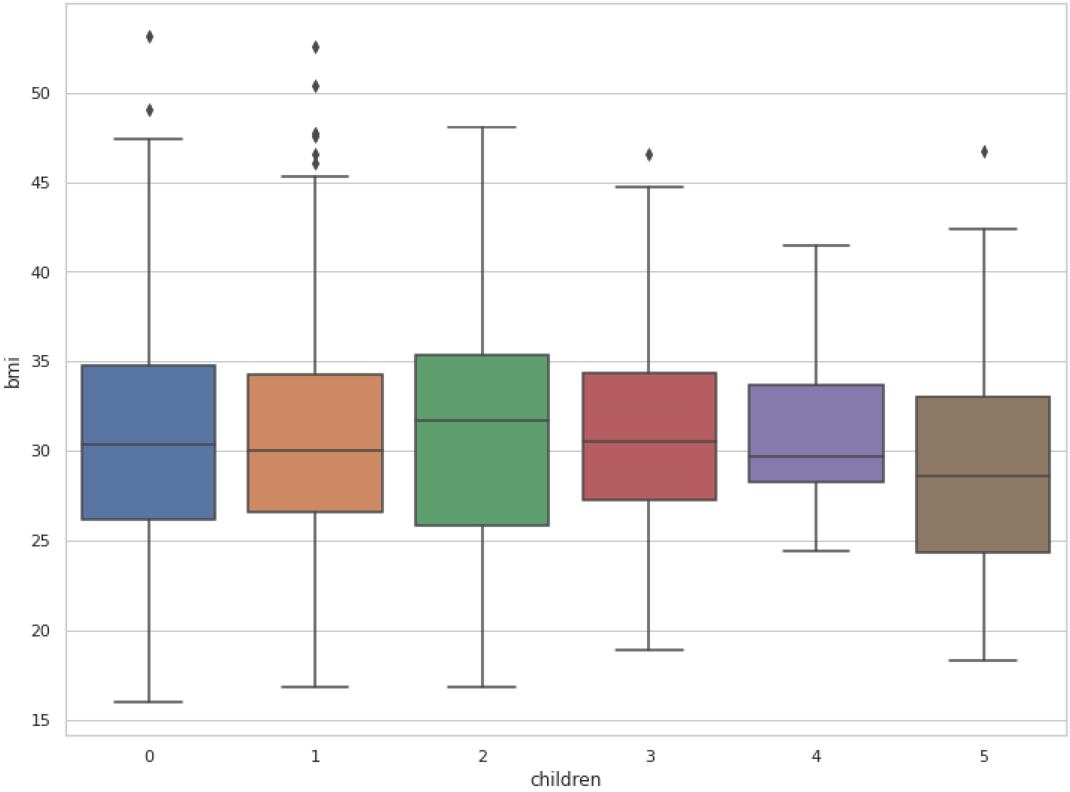


	age	bmi	children	charges
age	1	0.11	0.042	0.3
bmi	0.11	1	0.013	0.2
children	0.042	0.013	1	0.068
charges	0.3	0.2	0.068	1

Exploratory Analysis

BMI VS Children

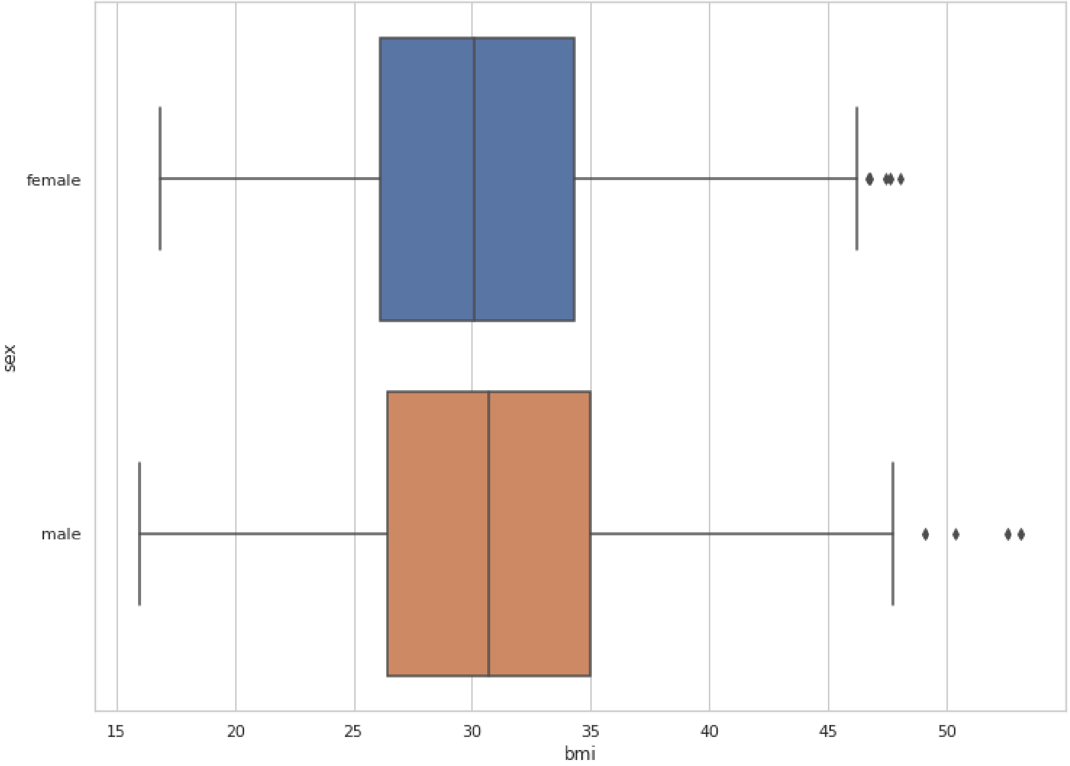
- BMI is almost same around the customer with any number of children except for small diversion in people with 4 kids



Exploratory Analysis

BMI VS Sex

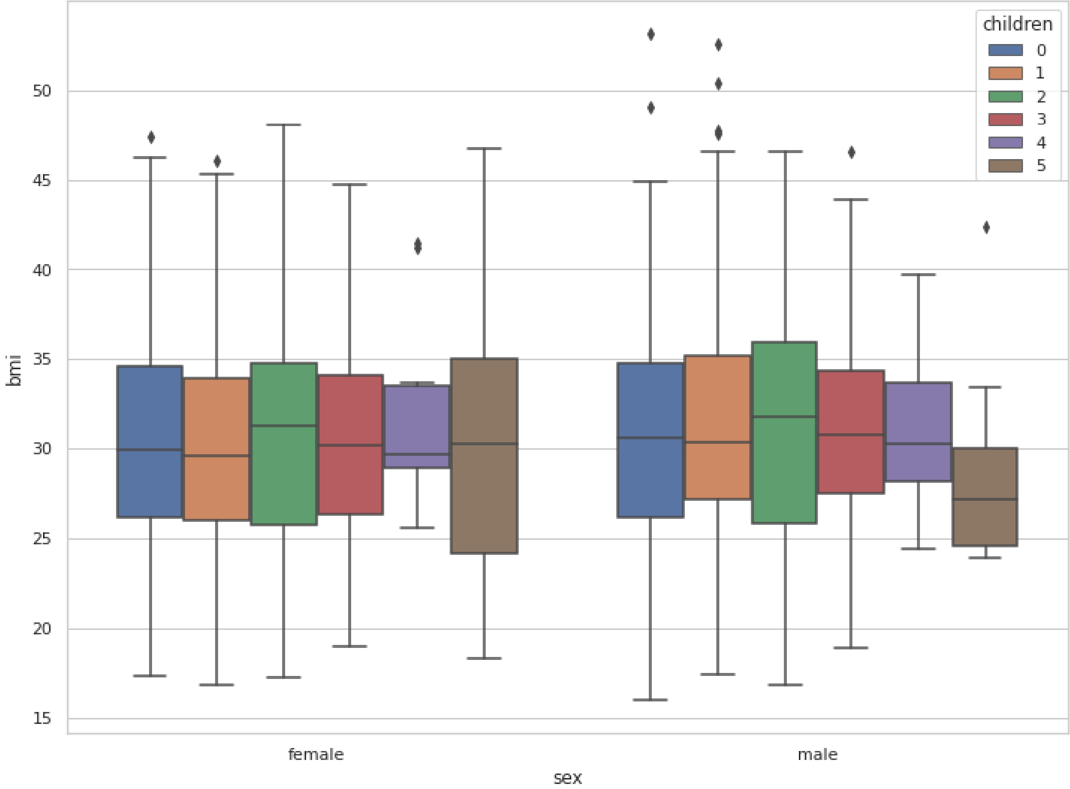
- BMI’s are consistent for both male and female insurer



Exploratory Analysis

BMI VS Sex VS Children

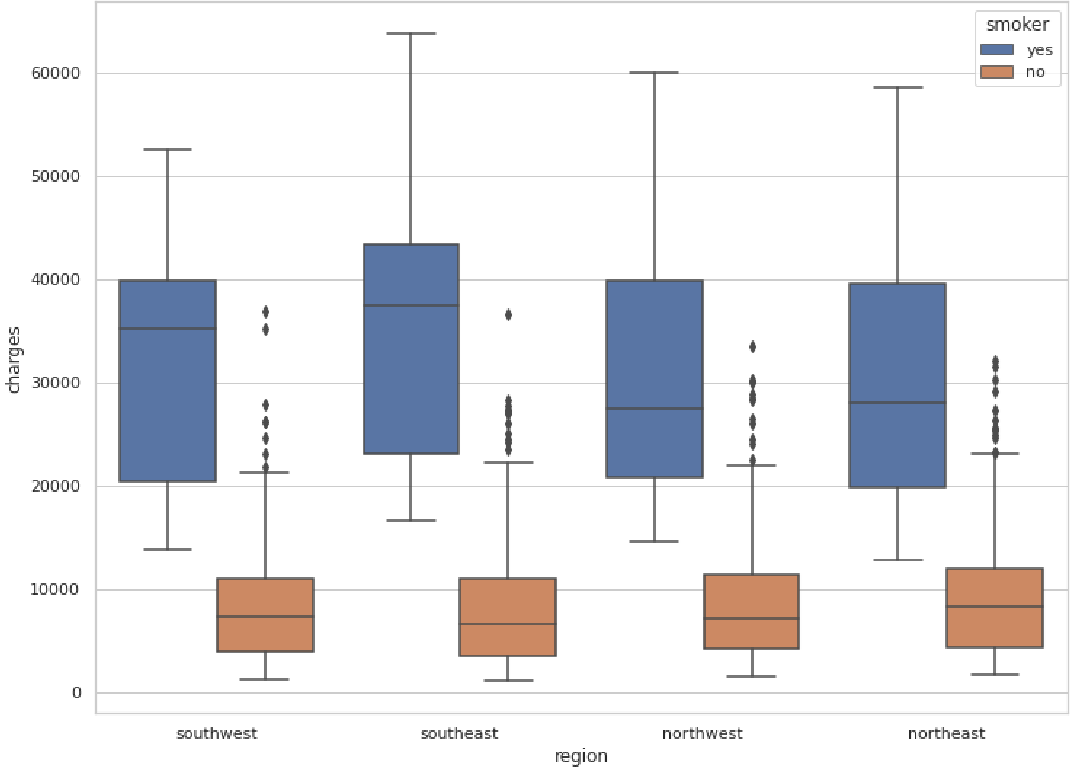
- Average BMI is in the range of 26 to 35 for both male and females
- Female with 5 children have higher range of BMI
- Males with 5 children have low values of BMI
- Except for Insurers with 5 children other BMI ranges seem pretty consistent



Exploratory Analysis

Charges VS Region VS Sex

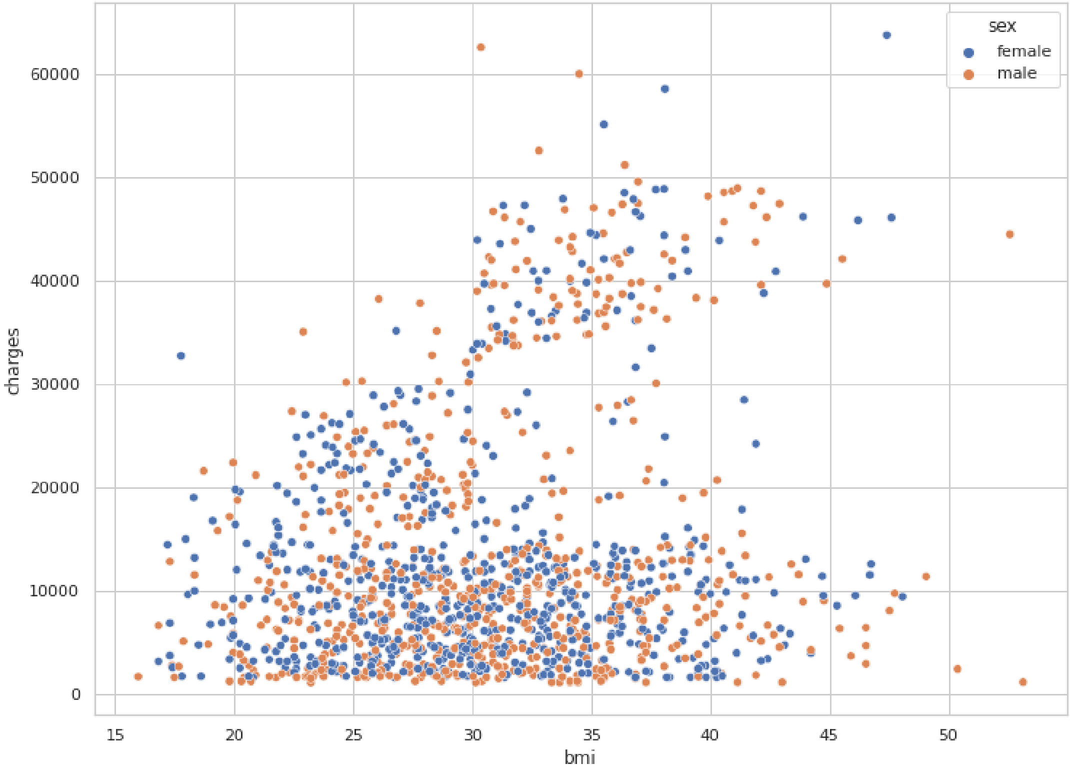
- The preference of smokers for charge is pretty consistent over the region.
- Customers from Southeast Who smoke have slightly higher charges than other regions



Exploratory Analysis

Charges VS BMI VS Sex

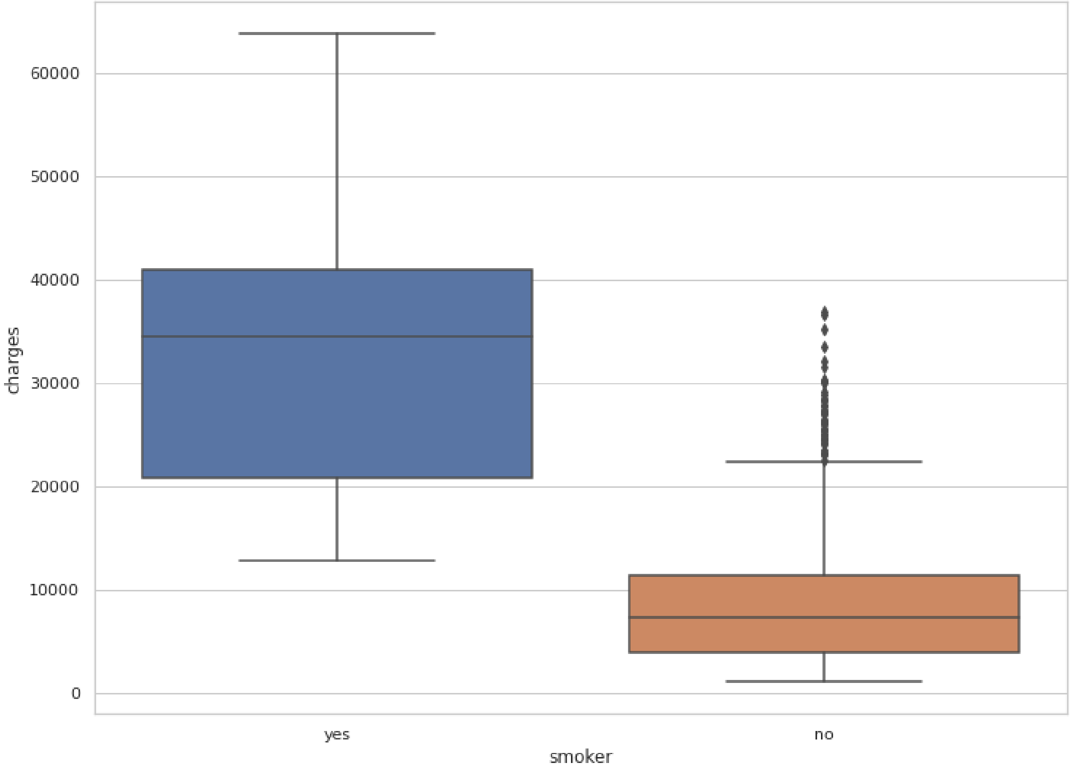
- For both male and female more charges are in low category although there are few outliers with more than 50K$



Statistical Analysis

Charges Vs Smoker

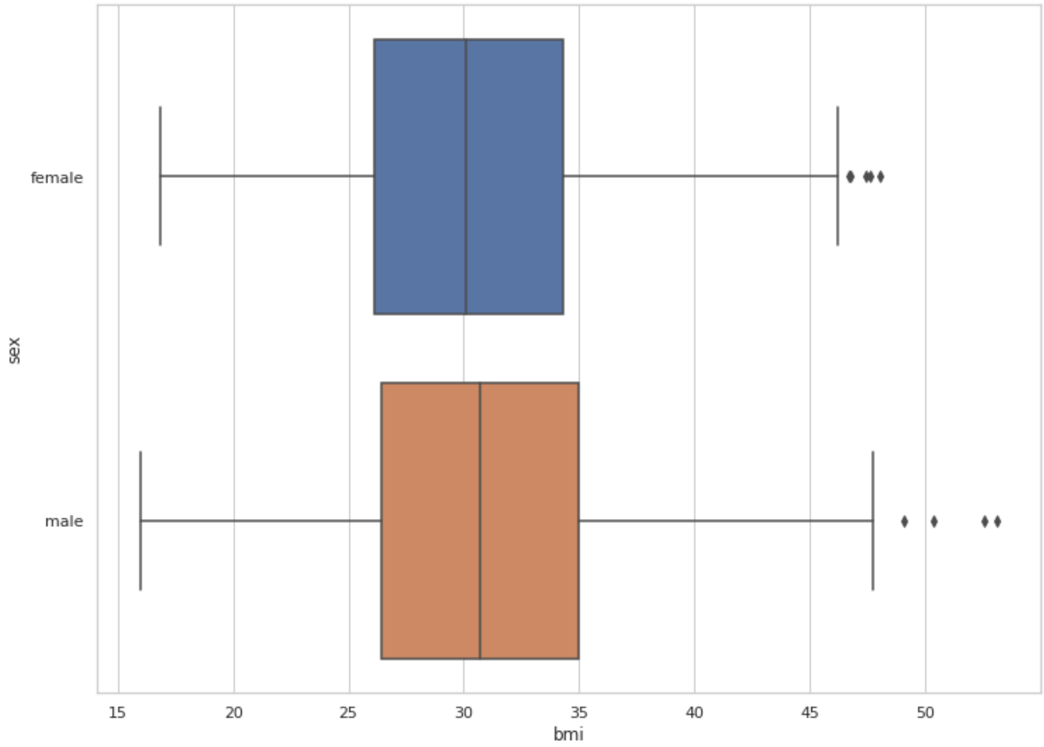
- Null Hypothesis--> H0 = "charges has no effect on smoking"
- Alternate hypothesis--> H1 = "charges has effect on smoking"
- Smoker have much higher claims than non smokers. After performing T-Test this is verified. The P value for the test is 8.271435842177219e- 283 (.00)



Statistical Analysis

BMI Vs Sex

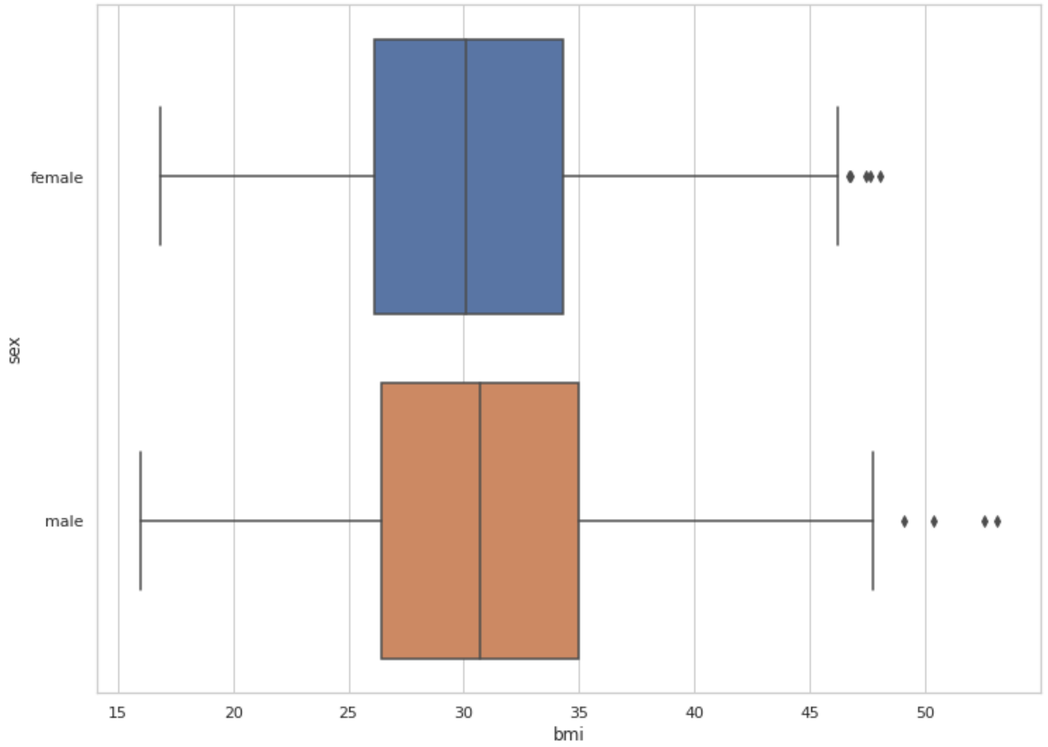
- Null Hypothesis--> H0 = "Gender has no effect on BMI"
- Alternate hypothesis--> H1 = "Gender has effect on BMI"
- From the P value of T-Test statistic that there is not much difference in BMI for male and female as it is evident from P value of T-Test is 0.089



Statistical Analysis

BMI Vs Sex

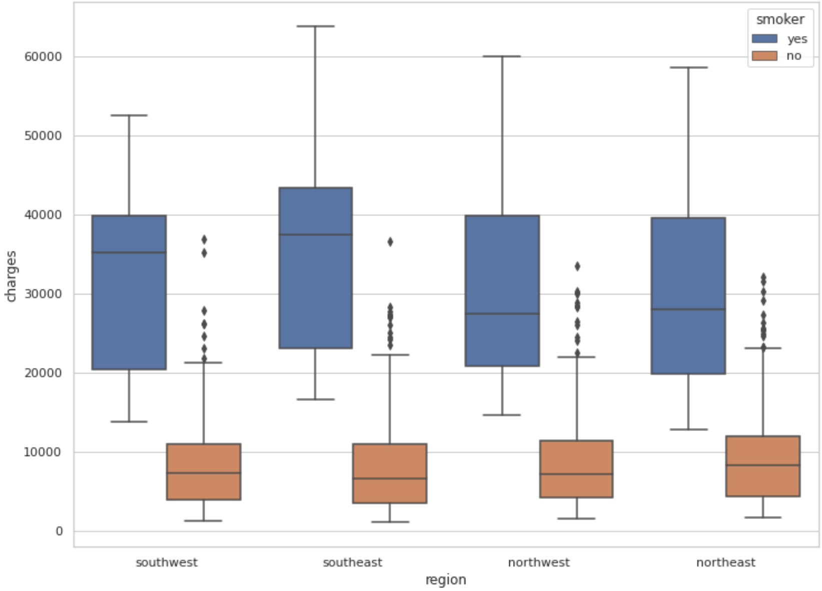
- Null Hypothesis--> H0 = "Gender has no effect on BMI"
- Alternate hypothesis--> H1 = "Gender has effect on BMI"
- From the P value of T-Test statistic that there is not much difference in BMI for male and female as it is evident from P value of T-Test is 0.089



Statistical Analysis

Smoker Vs Region

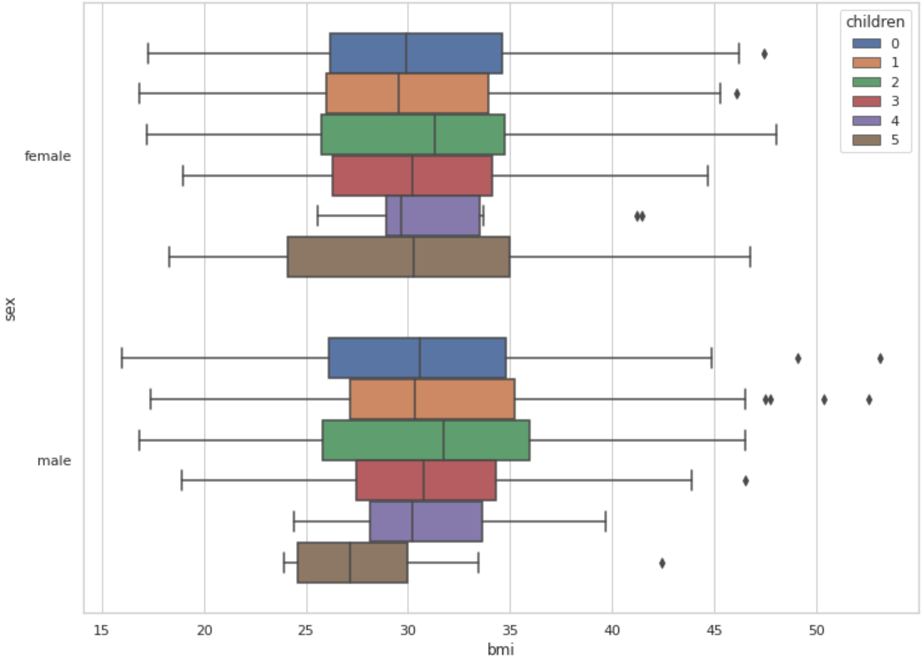
- Null Hypothesis--> H0 = "Smokers are effected by region"
- Alternate hypothesis--> H1 = "Smokers are not effected by region"
- There is no effect of region on smoker as p values of t-test is (0.99~1)



Statistical Analysis

Sex Vs BMI

- Null Hypothesis--> H0 = "Average BMI of female with 0,1,2 children is same"
- Alternate hypothesis--> H1 = "Average BMI of female with 0,1,2 children is not same"
- There is not a significant difference between BMI's of women with 0, 1, or 2 children.



Conclusion

- Majority of insurers are below the age of 23
- Most of the insurers are obese as the average BMI is 31.
- $13,300 is the average insurance claim.
- Most of the insurers are Non-Smokers.
- There are more insurers from southeast region (364). So it more for insurer to be from southeast region.
- Ratio of Men and Women as insurer is almost same. 49.5 female to 51.5 male
- There are a lot of outliers at the high end of charges. I.e. very few people are likely to claim higher charges
- The difference in regions of insurer is very small although it is more likely for insurer to be from Southeast region
- The preference to smoke is consistent across all the regions
- People who smoke have higher claims than non smokers
- Women with 4 or less children have small and similar BMI across the spectrum
- Women with 5 children are rare but have high range of BMI