Axis Insurance

EDA & Statistical Analysis

April 2021

Brian Hall – DSBA – Project 2

03 | Objective

04 | Data Overview

05 | Member Overview & Key Observations

06 | Statistical Analysis (Hypothesis Testing)

Appendix

11 | Uni-Variate Data Analysis

16 | Multi-Variate Data Analysis

April 2021

Objective

Axis Insurance would like to leverage its customer data to gain insights on its members and answer a few key questions that will prove useful for making business decisions.

Axis Insurance would like to answer the following questions with statistical significance:

1. Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don’t?

2. Prove (or disprove) that the BMI of females is different from that of males.

3. Is the proportion of smokers significantly different across different regions?

4. Is the mean BMI of women with no children, one child, and two children the same?

A significance level of 0.05 will be used for all tests.

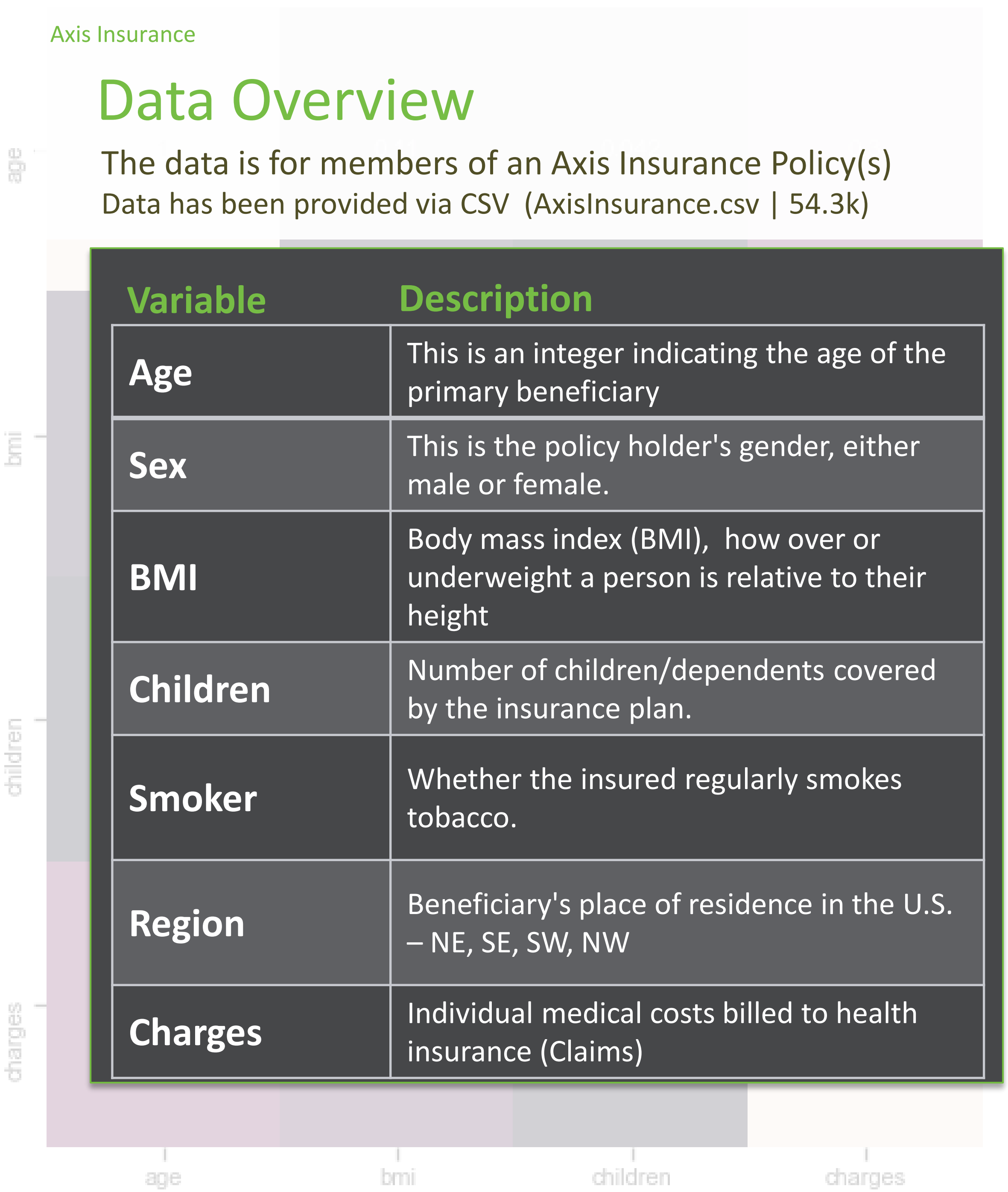
Approach

Explore the dataset provided by Axis Insurance and extract insights from the data, perform a statistical analysis of key business questions and provide conclusions.

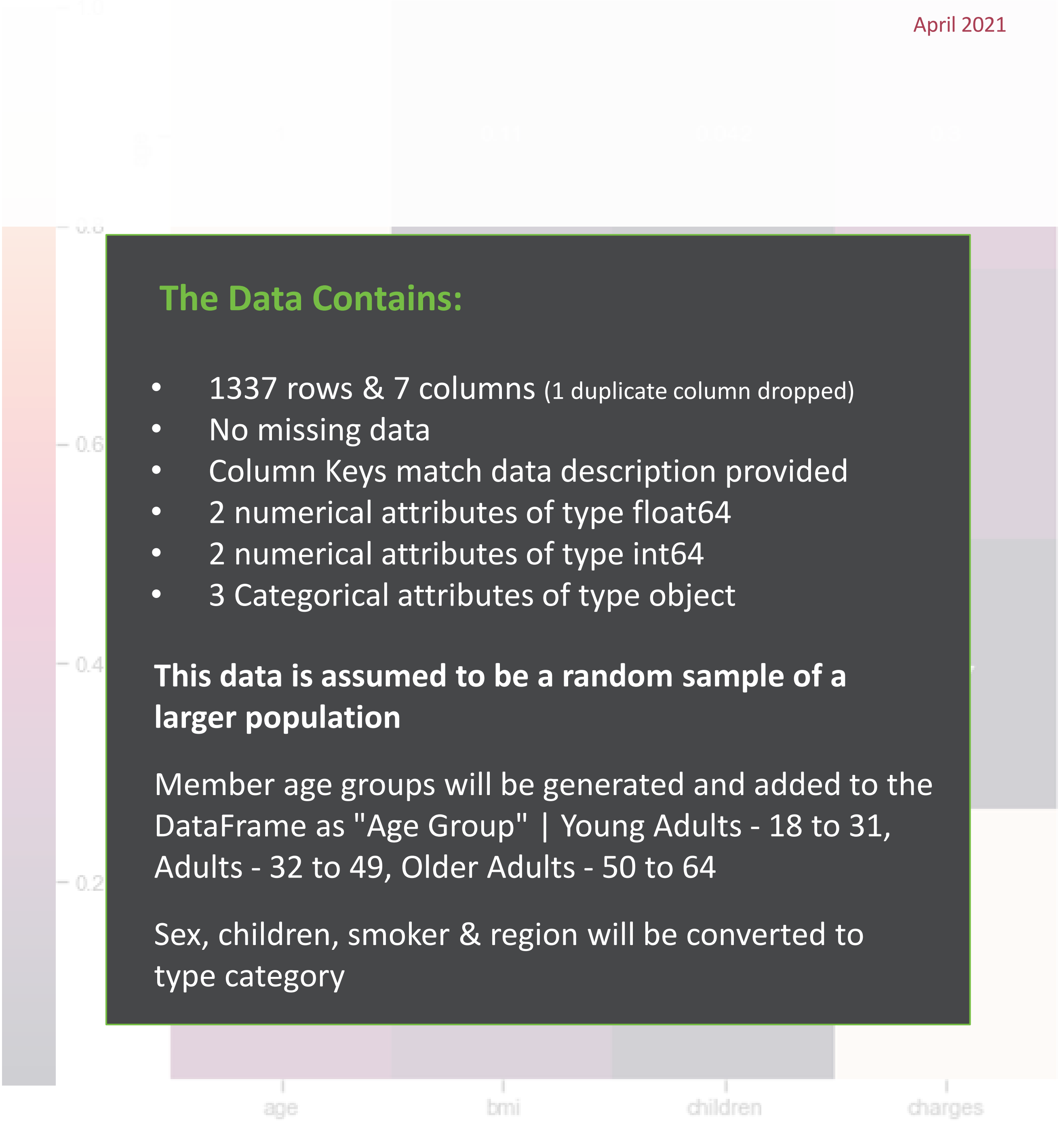
Data Overview

The data is for members of an Axis Insurance Policy(s)

Data has been provided via CSV (AxisInsurance.csv | 54.3k)



Variable	Description
Age	This is an integer indicating the age of the primary beneficiary
Sex	This is the policy holder's gender, either male or female.
BMI	Body mass index (BMI), how over or underweight a person is relative to their height
Children	Number of children/dependents covered by the insurance plan.
Smoker	Whether the insured regularly smokes tobacco.
Region	Beneficiary's place of residence in the U.S. – NE, SE, SW, NW
Charges	Individual medical costs billed to health insurance (Claims)



The Data Contains:

- 1337 rows & 7 columns (1 duplicate column dropped)
- No missing data
- Column Keys match data description provided
- 2 numerical attributes of type float64
- 2 numerical attributes of type int64
- 3 Categorical attributes of type object

This data is assumed to be a random sample of a larger population

Member age groups will be generated and added to the DataFrame as "Age Group" | Young Adults - 18 to 31, Adults - 32 to 49, Older Adults - 50 to 64

Sex, children, smoker & region will be converted to type category

Axis Insurance

April 2021

Member Overview

1337 observations

Young Adult (age 18 - 31)	37%
Adult (age 32 - 49)	36%
Older Adult (age 50 - 64)	27%
Mean Age	39
Age Range	18 - 64

Males	50.5%
Females	49.5%

Average BMI	31
BMI Range	16 - 53

Children – None	43%
Children Range	1 - 5

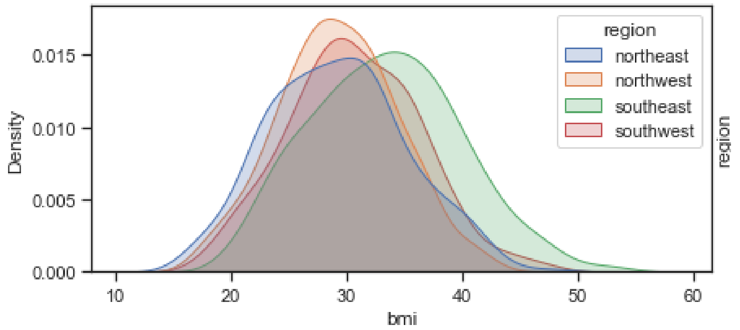
Smoker Yes – 20.5%	Smoker No – 79.5%

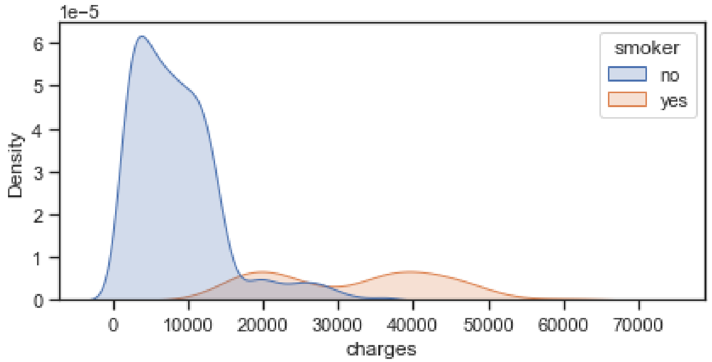
Average Claims	$13,297
Claims Range	$1,122 - $63,770

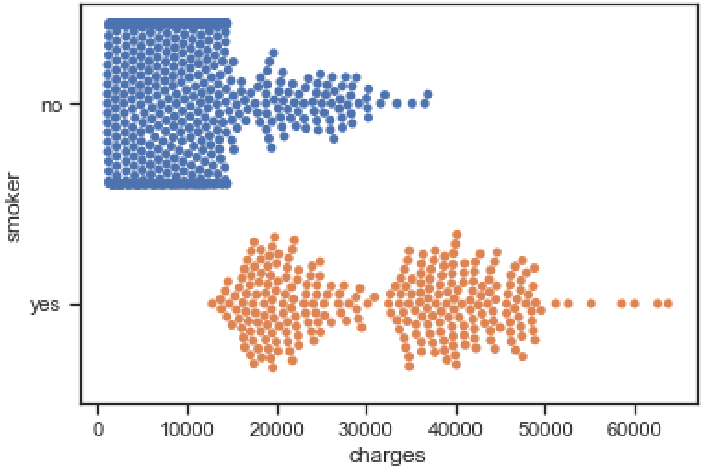
See appendix for data analysis and distributions

Key Member Observations

- There are no members over the age of 64, likely due to Medicare / Medicaid
- Appx 75% of members are below the age of ~ 52
- **Nearly 75% of members are greater than the upper range for normal BMI (18.5 - 24.9)**
- Members are distributed across the regions nearly evenly with slightly more, 3%, in the Southeast
- There are many outliers above $50k in claims and could likely be considered catastrophic cases
- Most charges are below 15,000.
- **There is clear correlation with smokers and charges**
- BMI seems to be evenly distributed across age groups
- **There is surprisingly low correlation among age, bmi & charges**
- Charges primarily overlap across age groups in the 12k to 18k range
- **Charges above 30k are primarily to BMI above 30**
- Median for charges is approximately the same, ~ $9k across all regions
- **Smokers' median charges is apx $25k higher than non-smokers**
- Most charges below $18k are to non-smokers
- **Smokers account for all charges above $39k**
- Smoker charges median is higher then nearly all non-smoker outliers
- Members with no children have more smokers, they also represent 50% of the population
- The Southeast has a higher mean BMI by 4 - 6 points than other regions
- **No regions IQR is within the normal BMI range**
- Southeast is the only region with BMIs higher than 48
- **Southeast has no BMIs lower then 20**
- **The highest BMI's > 50 are male young adults**







Axis Insurance

Statistical Analysis (Hypothesis Testing)

April 2021

Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?

Let $\mu1,\mu2$μ1,μ2 be the means of smokers and non-smokers respectively.

Testing the null hypothesis
$H0$: $\mu1$ = $\mu2$

against the alternative hypothesis
Ha: $\mu1$ > $\mu2$

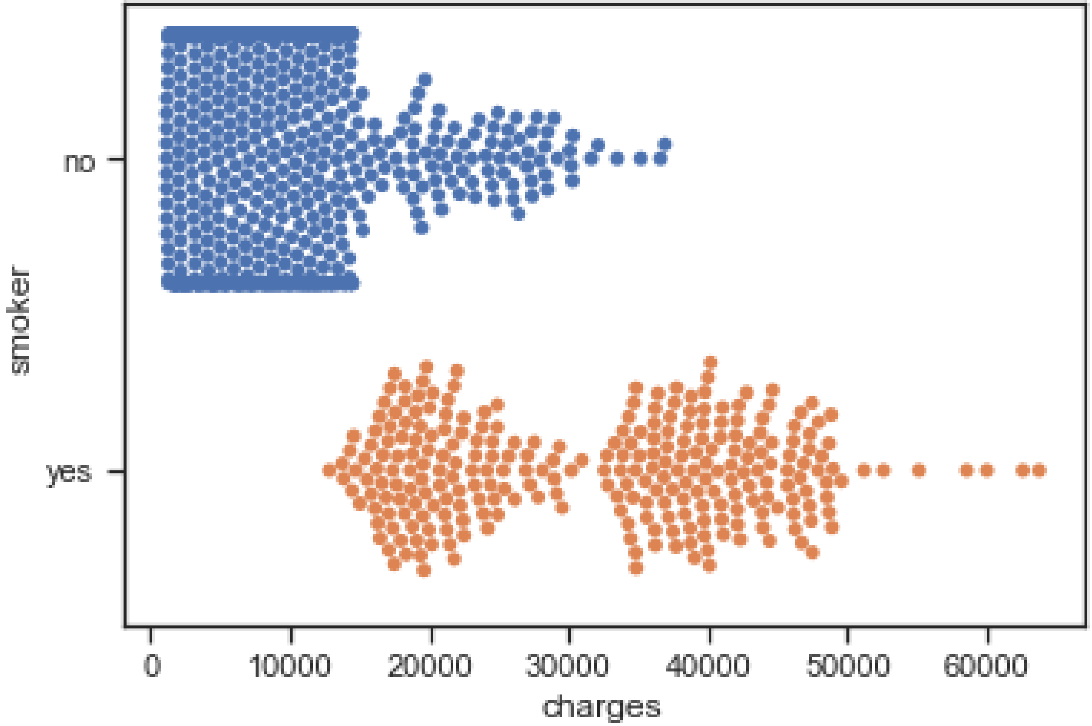
Test Method:
Paired t-test & 2 Sample Independent t-test (unknown std dev)

p-values - 2.42 e-65 & 1.35 e-78

p-values < alpha of 0.05
Fail to accept the Null Hypothesis

Conclusion

There is statistically significant evidence that the claims of members who smoke is greater than those who do not smoke based on provided data



Axis Insurance

April 2021

Statistical Analysis (Hypothesis Testing)

Prove (or disprove) that the BMI of females is different from that of males.?

Let $\mu 1$,$\mu 2$ be the means of females and males BMI's respectively.

Testing the null hypothesis
$H0$: $\mu 1=\mu 2$

against the alternative hypothesis
Ha: $\mu 1\neq\mu 2$

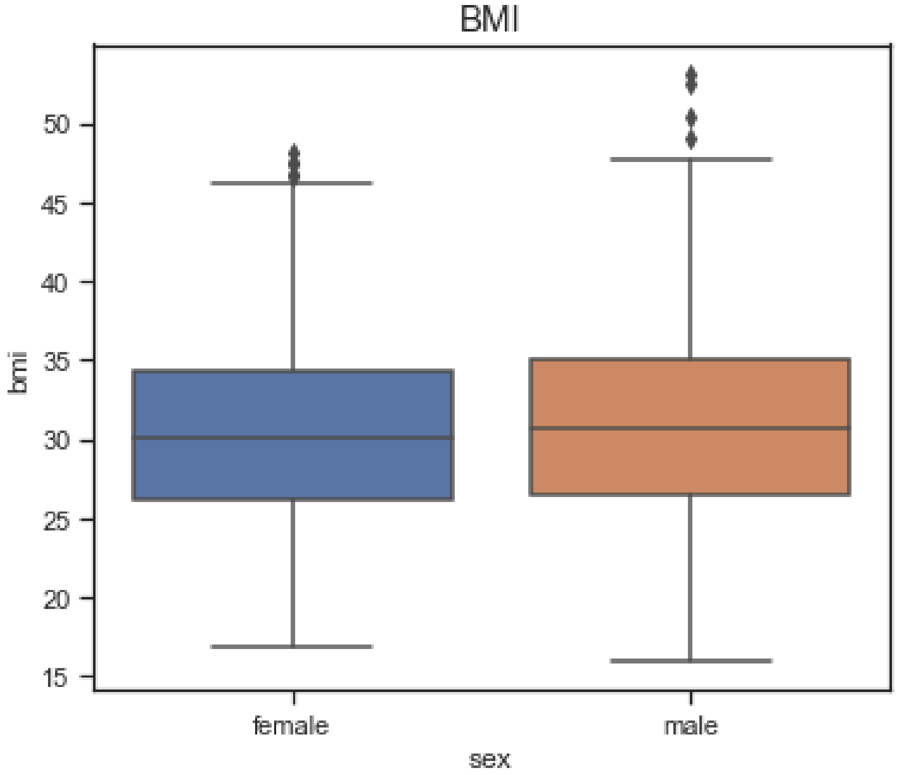
Test Method:
One-way ANOVA & 2 Sample Independent t-test

p-values - 0.072 & 0.072

p-values > alpha of 0.05
We accept the Null Hypothesis

Conclusion

There is statistically significant evidence that the BMI of females and males are NOT different based on the provided data.



Axis Insurance

Statistical Analysis (Hypothesis Testing)

April 2021

Is the proportion of smokers significantly different across different regions?

Testing the null hypothesis
$H0$: The proportion of smokers is equal across regions

against the alternative hypothesis
Ha: The proportion of smokers is not equal across regions

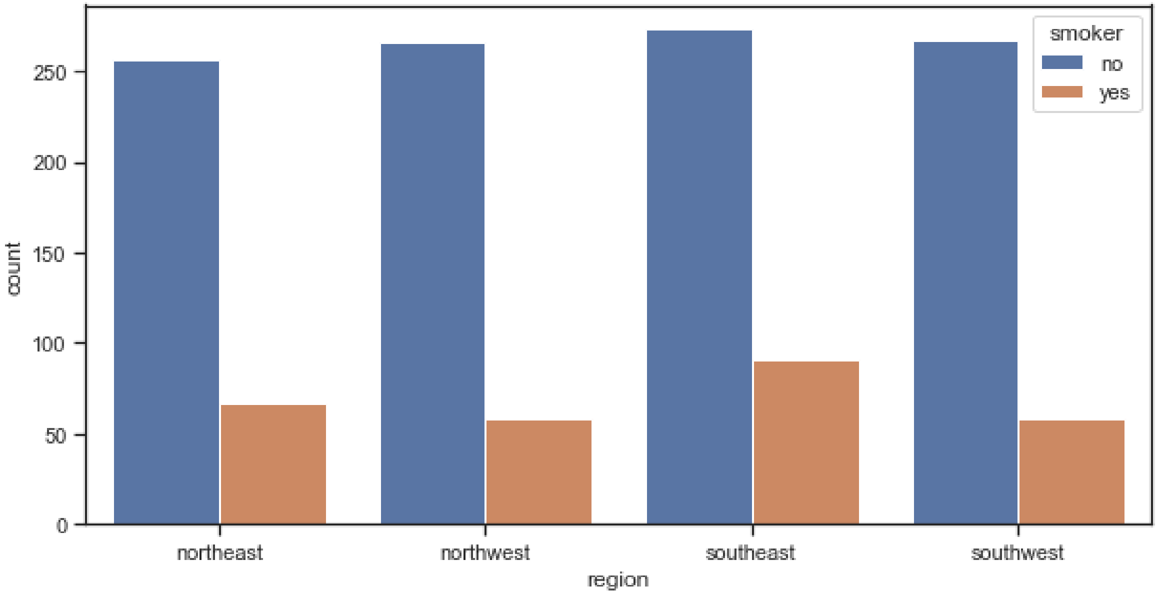
Test Method:
Chi Squared Test for Independence

p-value - 0.063

p-value > alpha of 0.05
We accept the Null Hypothesis

Conclusion

There is statistically significant evidence that the proportion of smokers is NOT different across different regions based on provided data.



Statistical Analysis (Hypothesis Testing)

Is the mean BMI of women with no children, one child, and two children the same?

Let $\mu1,\mu2,\mu3$ be the means of female BMIs with 0, 1, 2 children, respectively.

Testing the null hypothesis
$H0$: $\mu1 = \mu2 = \mu3$

against the alternative hypothesis
Ha: $\mu1 \neq \mu2 \neq \mu3$

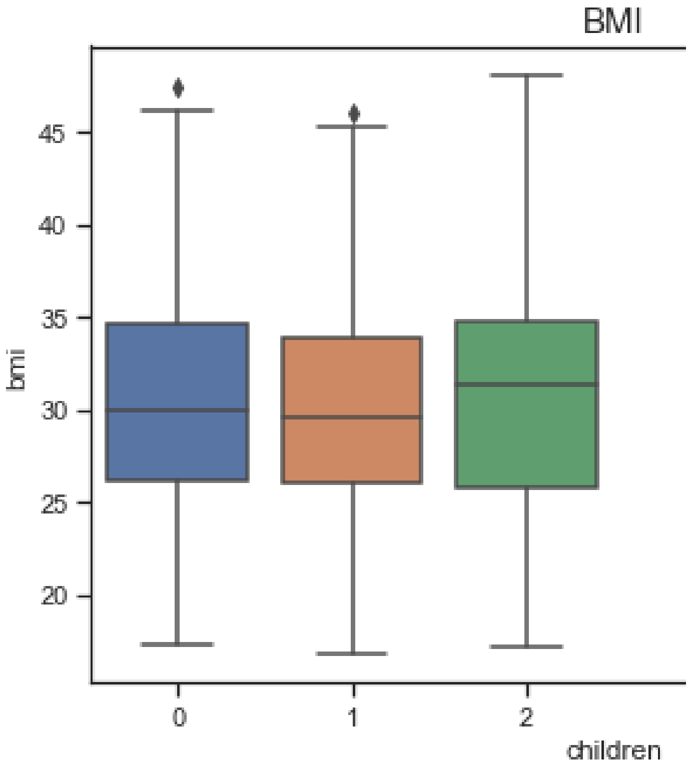
Test Method:
One-way ANOVA & Kruskal-Wallis

p-value - 0.715 & 0.699

p-value > alpha of 0.05
We accept the Null Hypothesis

Conclusion

There is statistically significant evidence that the BMI of females with no children, one child and two children are the same based on the provided data.



Appendix

Axis Insurance

April 2021

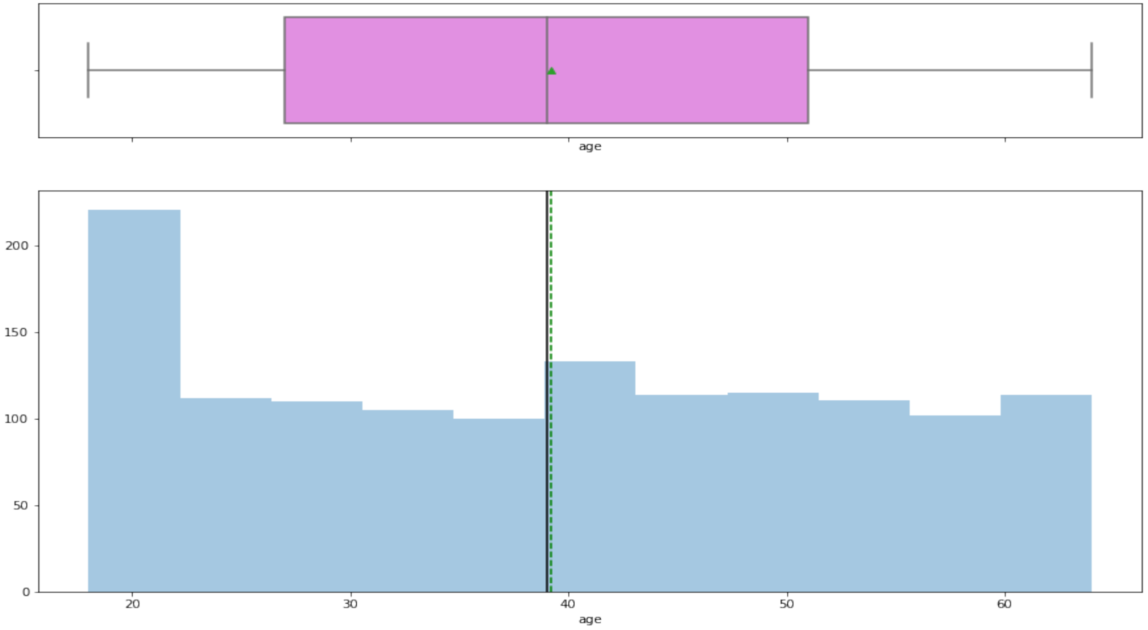
Uni-Variate Analysis

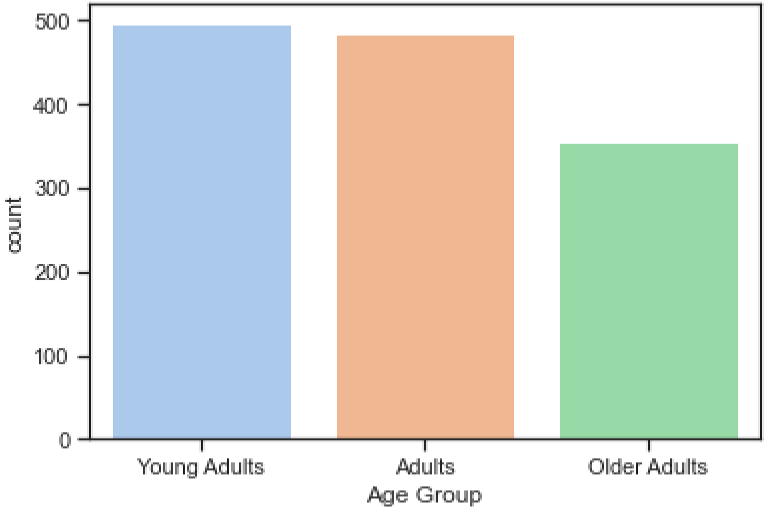
Axis Insurance

April 2021

Univariate Analysis

Age & Age Group Distribution





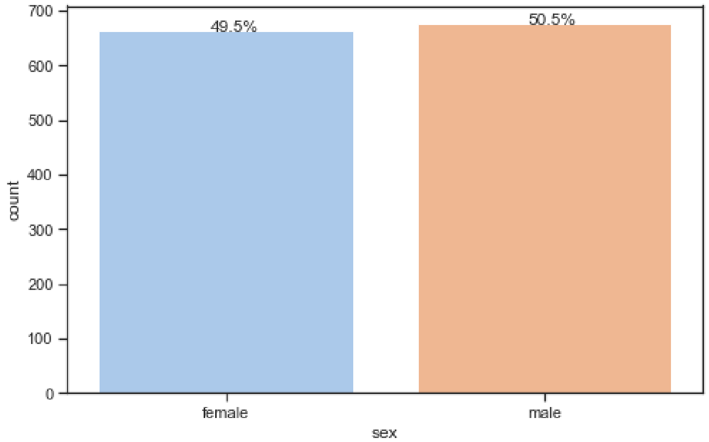
Observations

- Mean and median are aproximately the same suggesting normal distribution despite a spike around age 20
- There seem to be few if any outliers
- There are many more members around the age of 20 than any other age group
- Aproximately 75% of members are below the age of ~ 52
- There are no members over the age of 64, likely due to Medicare / Medicaid
- Create age groups to more easily visualize age. Young Adults - 18 to 31, Adults - 32 to 49, Older Adults - 50 to 64

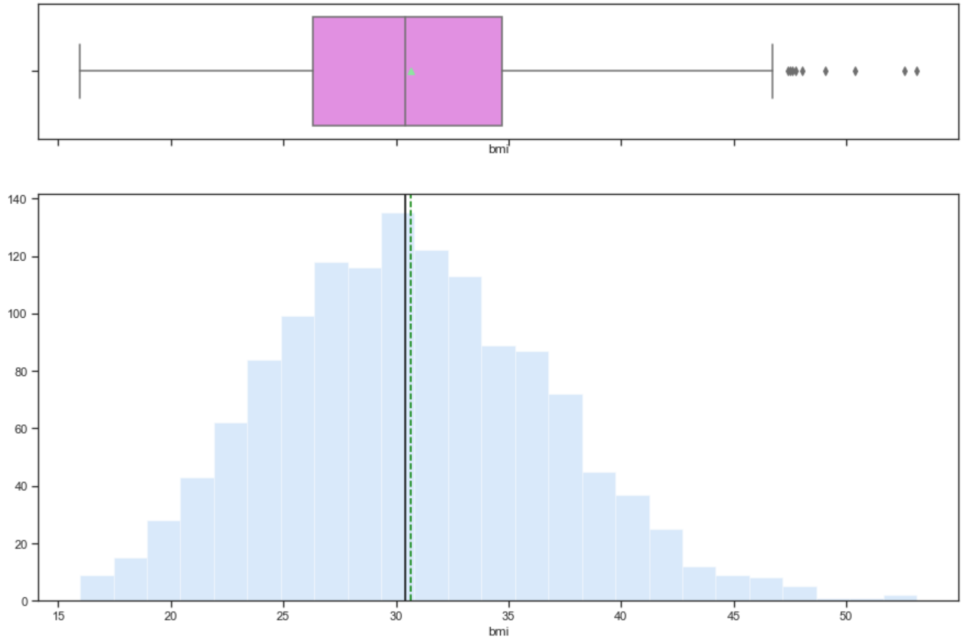
- 496 Young Adults - 37.1%
- 485 Adults - 36.2%
- 356 Older Adults - 26.6%

Univariate Analysis

Sex / Gender



BMI



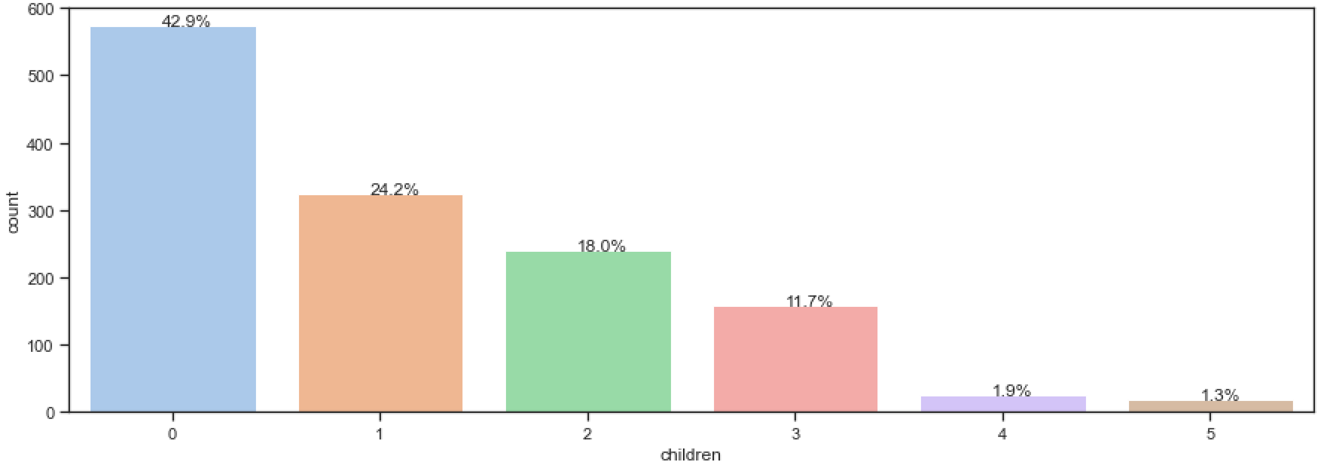
Observations

- Males and Females are nearly evenly distributed
- Females 49.5% -- Males 50.5%

- Mean and Median are nearly the same, ~30, suggesting a normal distribution. However there does seem to be a very slight right skew
- Nearly 75% of members are greater than the upper range for normal BMI (18.5 - 24.9)
- There are several outliers above a BMI of 46

Univariate Analysis

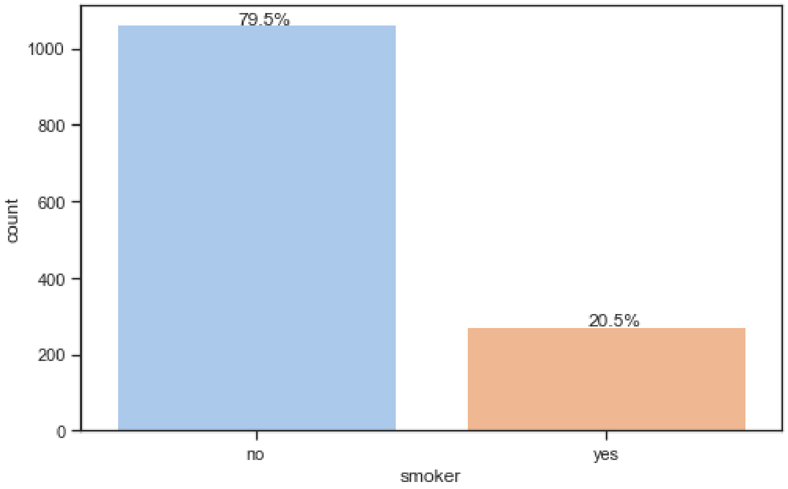
Children



Observations

- Nearly half, 43%, of members have 0 children
- Members with >3 children account for only 3.2%

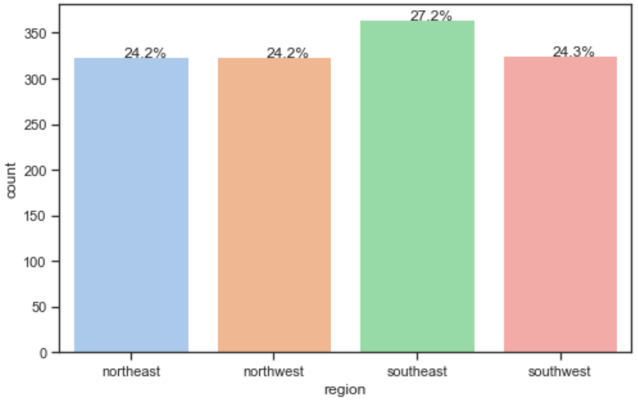
Smoker



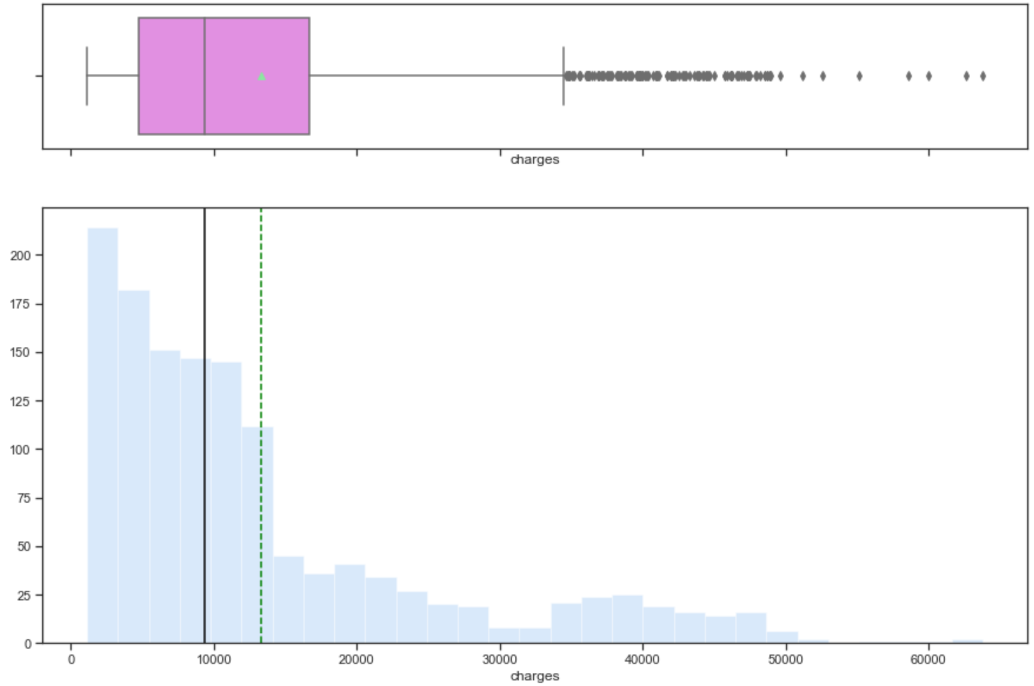
- Nearly 80% of members do not smoke

Univariate Analysis

Region



Charges



Observations

- Members are distributed across the regions nearly evenly with slightly more, 3%, in the Southeast

- Charges appears to be a F Distribution

- Charges distribution is highly right skewed

- There are many outliers between 34,000 and 50,000

- There are several outliers above 50,000 and could likely be considered catestrophic cases

- The majority of charges are below 15,000

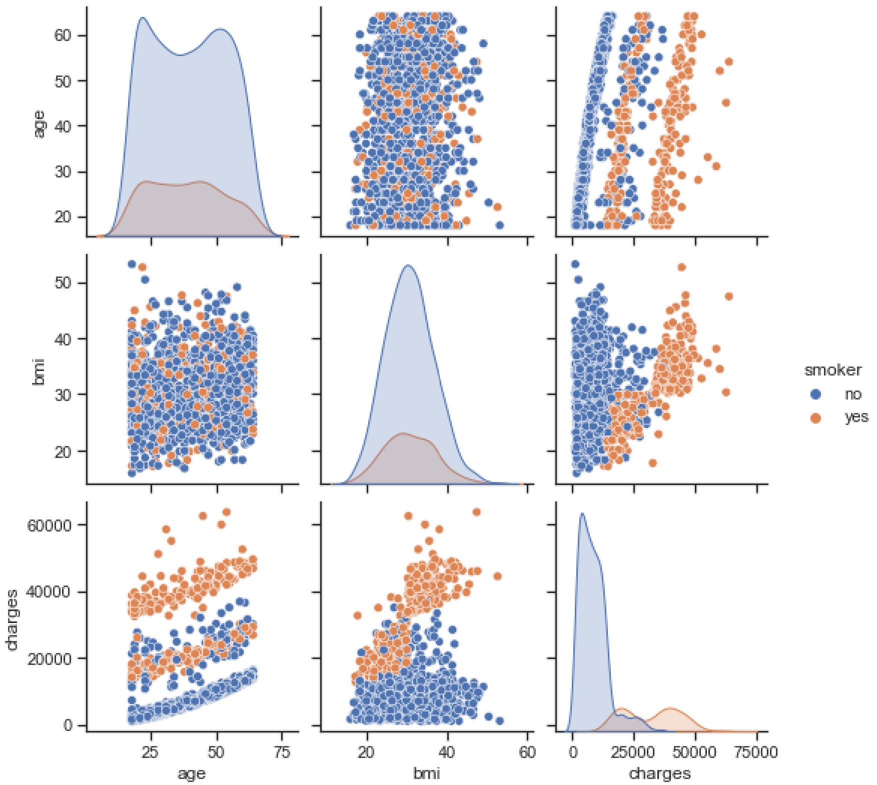
Multi-Variate Analysis

Axis Insurance

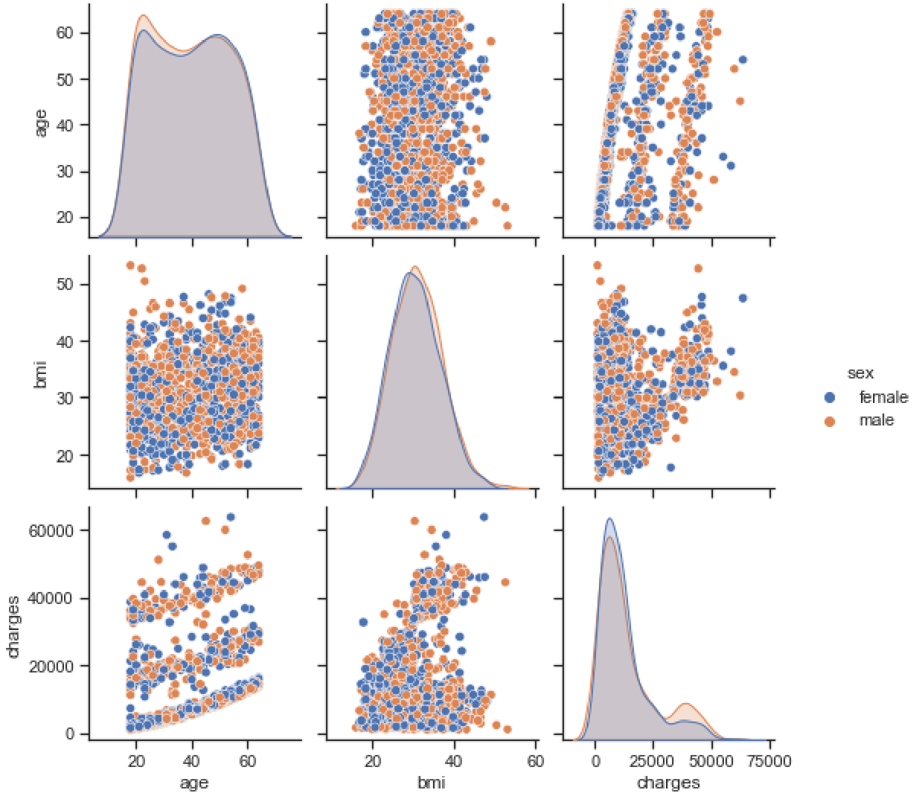
April 2021

Multi-Variate Analysis

Correlation w/ Smoker



Correlation w/ Sex / Gender



Observations

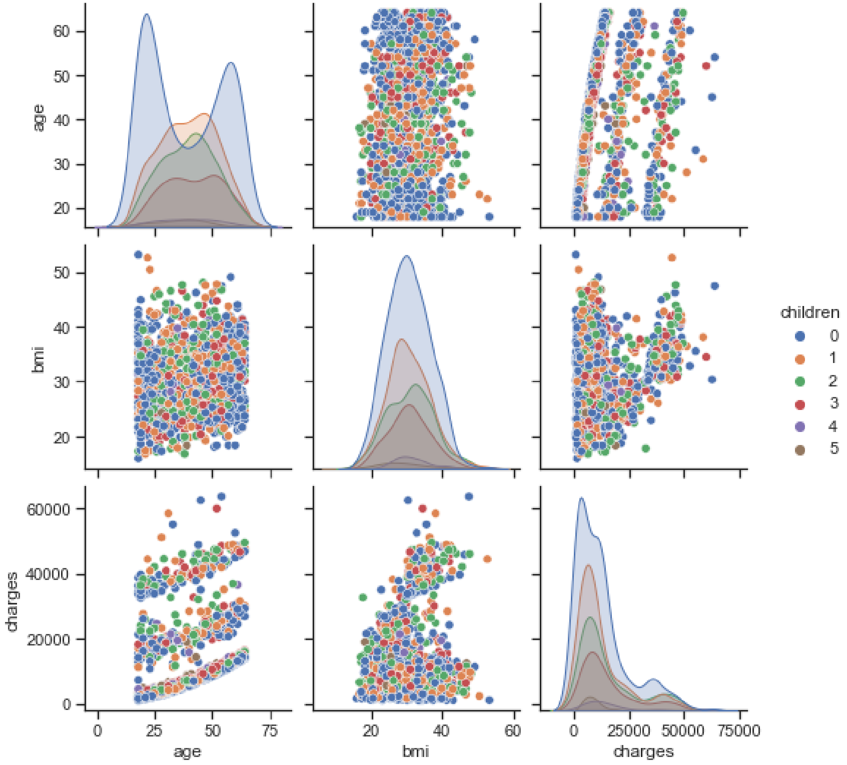
- There Seems to be a high correlation with smokers and charges
- Smokers seem to be evenly distributed within bmi and age
- Age & BMI seem to be evenly distributed

Observations

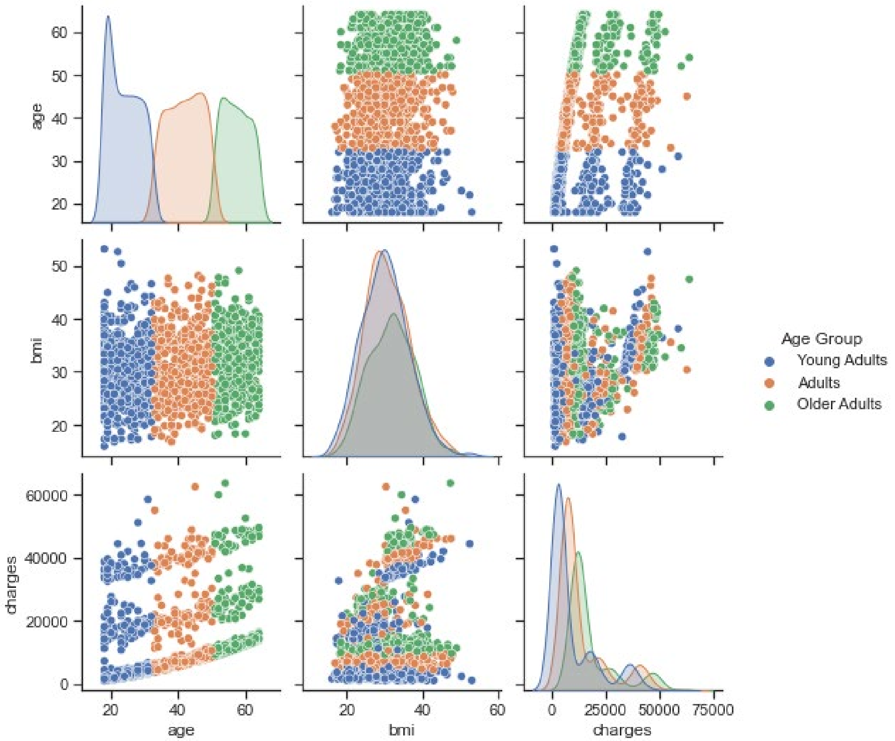
- Male and female seem to be evenly distributed across attributes
- Males have slightly higher charges in the 25k to 52k range
- Females have slightly higher charges in the 1k to 20k range

Multi-Variate Analysis

Correlation w/ Children



Correlation w/ Age Group



Observations

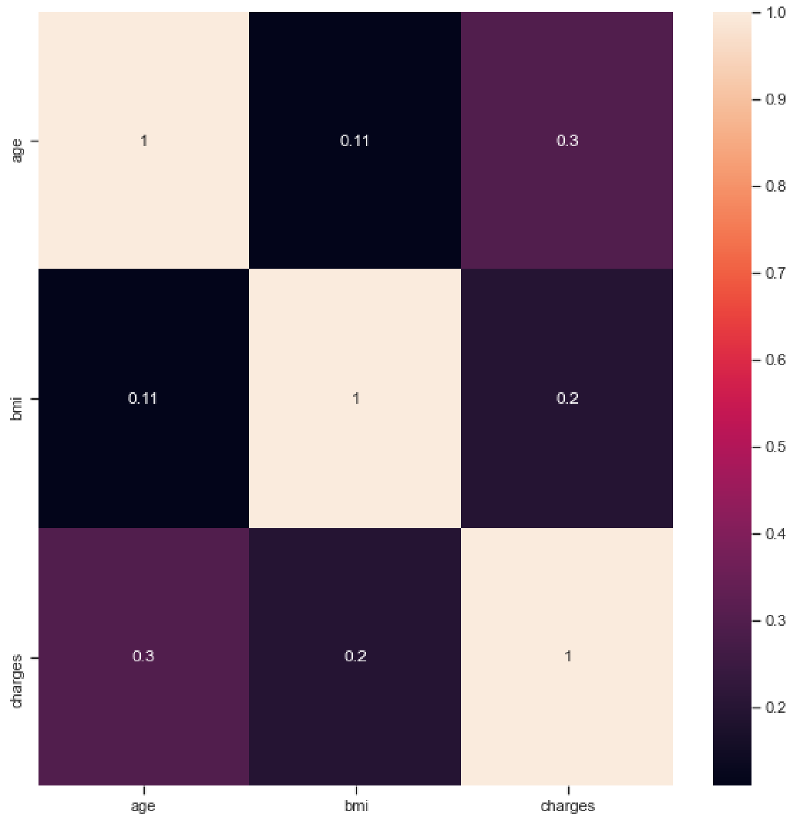
- BMI, Charges & Age seems to be evenly distributed regardless of number of children

Observations

- BMI seems to be evenly distributed across age groups
- There may be some correlation between age and charges

Multi-Variate Analysis

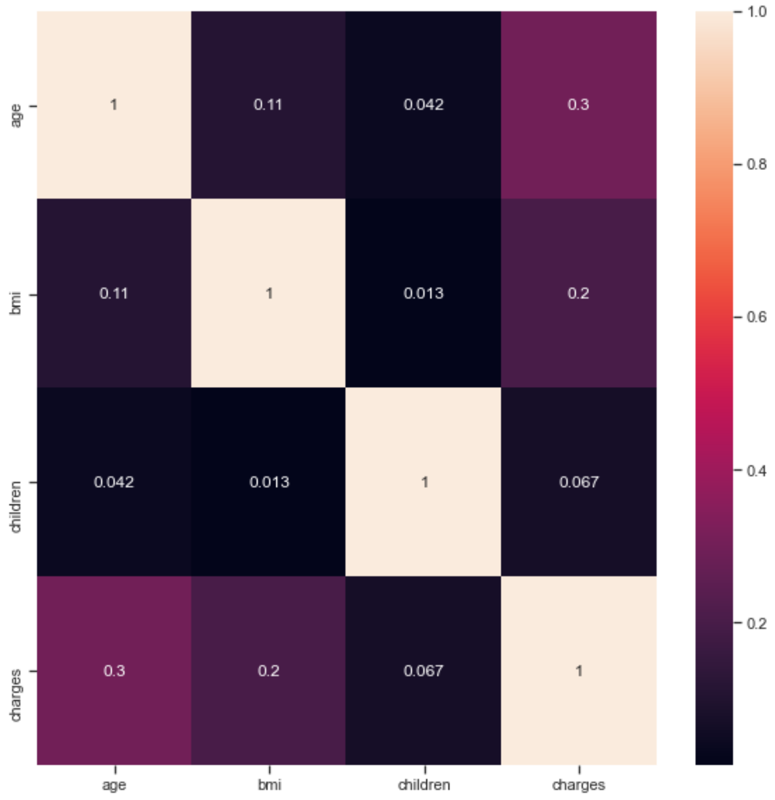
Heatmap



Observations

- There is surprisingly low correlation among age, BMI & charges
- Age and Charges have the highest correlation of 0.3
- Age and BMI have the lowest correlation of 0.11

Heatmap w/ Children

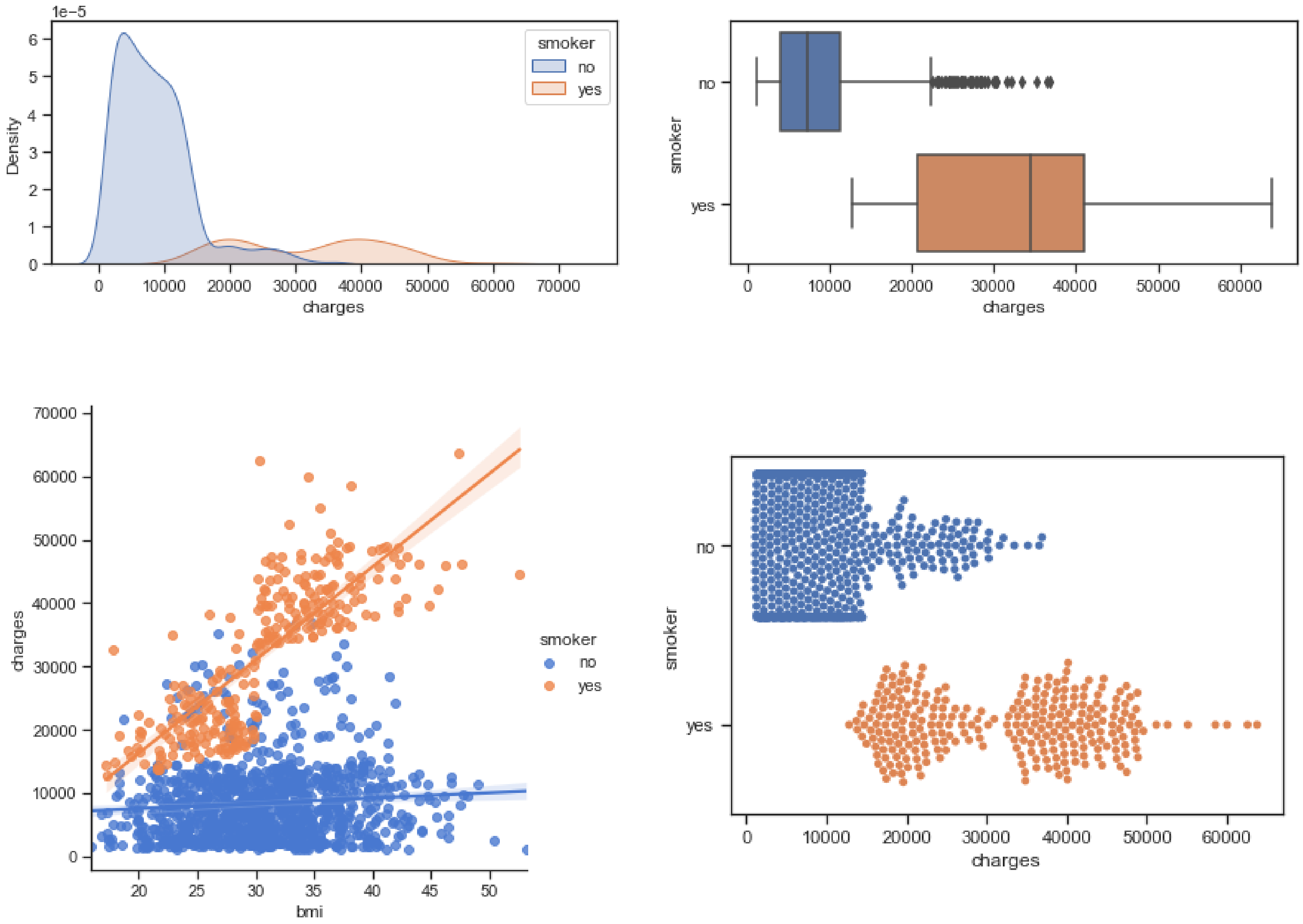


Observations

- There is very low correlation for all attributes and number of children, 0.013 to 0.068

Multi-Variate Analysis

Smoking, BMI & Charges

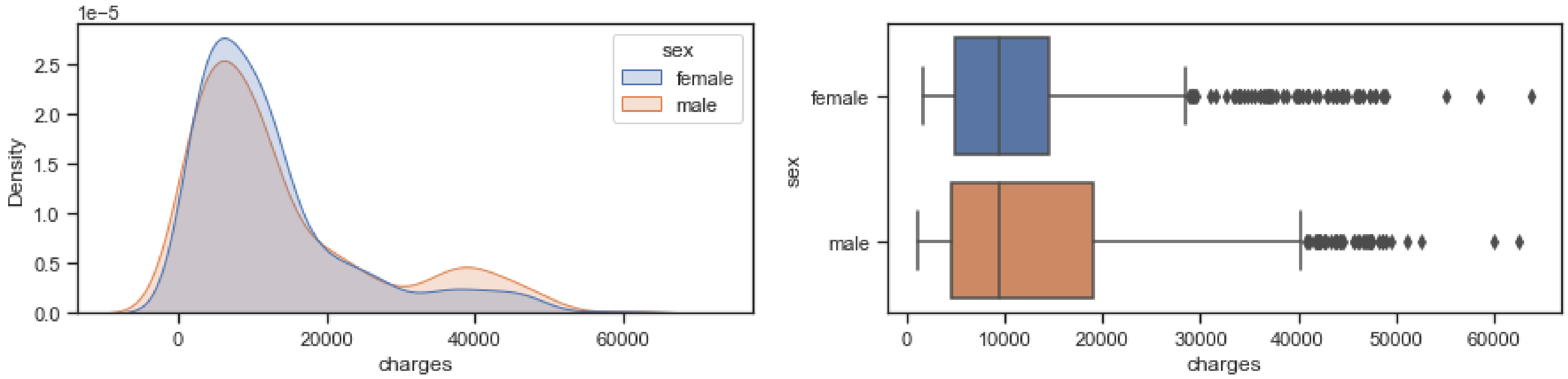


Observations

- There seems to be a high correlation between smokers, bmi and charges
- Smokers median is apx 25k higher than non-smokers
- The vast majority of charges below 18k are to non-smokers
- The vast majority of charges above 18k are to smokers
- Smokers account for all charges above 39k
- Smoker charges median is higher then nearly all smoker outliers
- Nearly all of the smokers IQR is higher then the entire non-smokers distribution
- Smokers 75th percentile and above are higher than any non-somker outliers

Multi-Variate Analysis

Charges with Sex / Gender

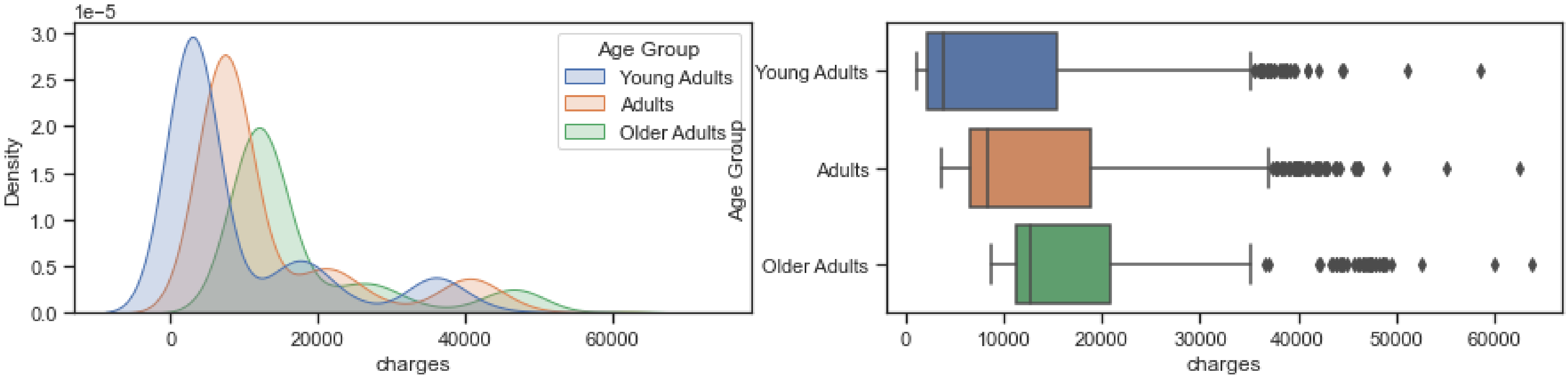


Observations

- Male and female have nearly the same mean of 9.5k and 25% of about 5k however males have higher charges in the 75%
- Males and females have many higher outliers in charges with females starting about 30k and males about 40k
- Males have a wider distribution across charges

Multi-Variate Analysis

Charges with Age Group

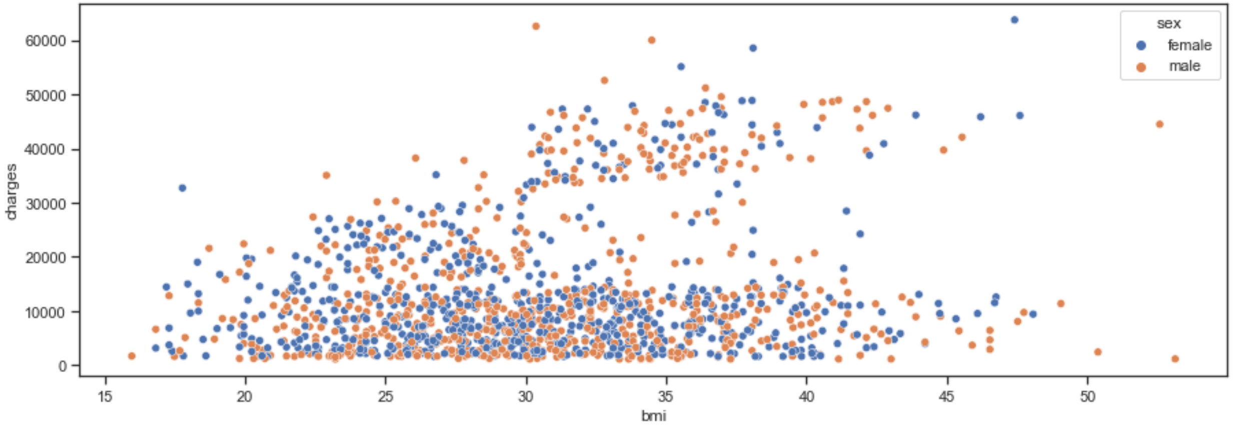


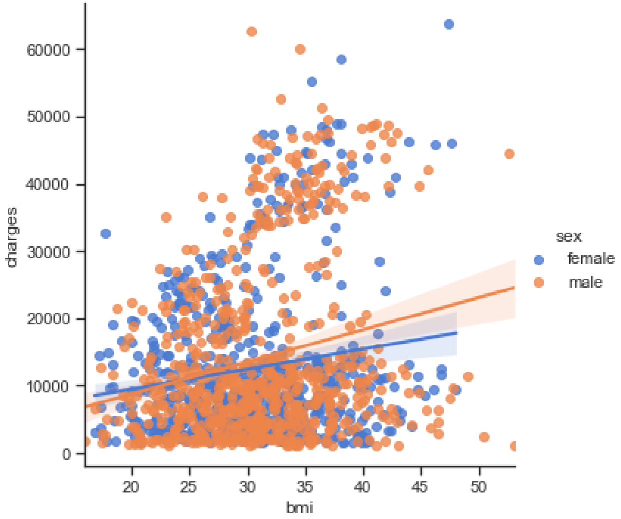
Observations

- As would be expected Older Adults have higher average charges and Young Adults have lower average charges
- Surprisingly Older Adults do not have as high a distribution as Young Adults and Adults across charges
- Charges primarily overlap across age groups in the 12k to 18k range
- All age groups have higher charges outliers
- All age ranges have a significantly larger Q3

Multi-Variate Analysis

Charges, BMI, Sex



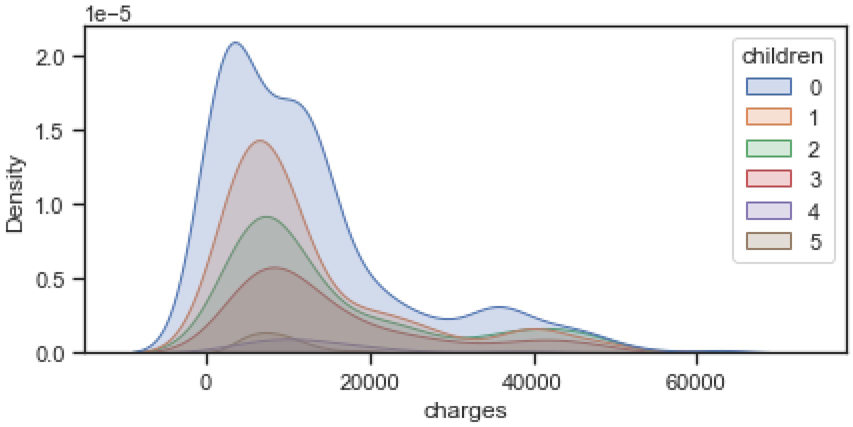


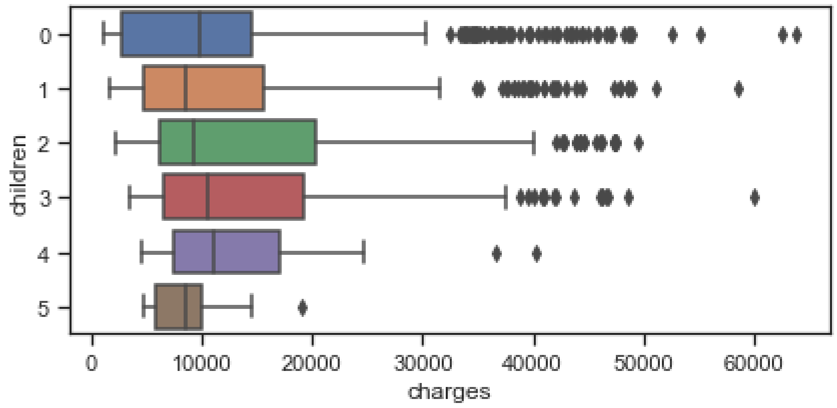
Observations

- Note: Normal bmi 18.5 to 25
- Normal bmi has slightly lower density in the lower charges
- Normal bmi has nearly no charges above 30k
- Charges above 30k are primarily to bmi above 30
- Charges between 15k and 30k are for bmi 20 to 30
- Highest density is 20 to 40 bmi and 1k to 15k
- BMI and charges are mildly correlated

Multi-Variate Analysis

Charges & Children

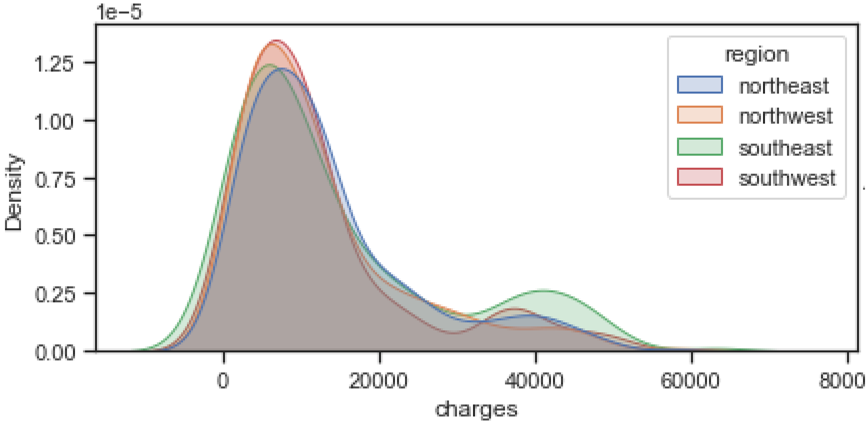


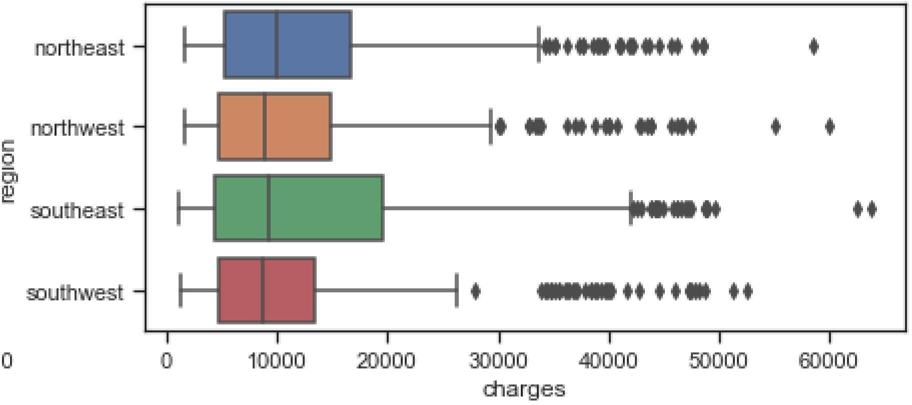


Observations

- Members with 2 children have the widest distribution and highest charges within the IQR despite making up only 17.9% of the population
- Median charges for members are close across # of children with 5 children haveing the lowest (5 children is only 1.7% of the population

Charges & Region

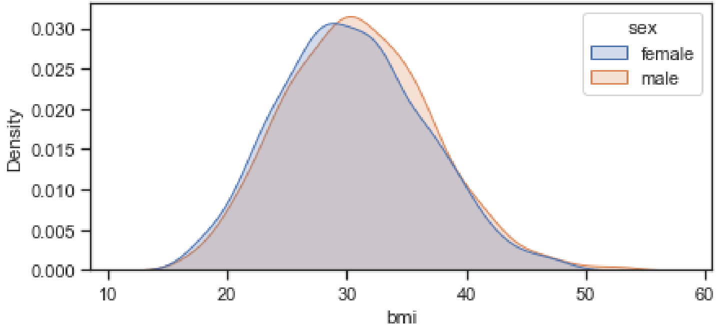


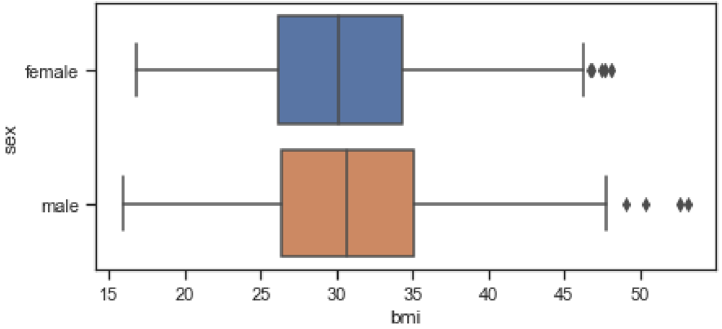


- Median for charges is approximately the same, ~ 9k across all regions
- Southeast region has the widest distribution and highest charges with only 3% more members than othe regions
- Swouthwest region has the most narrow distribution of charges but still has many higher charges outliers
- Southeast has higher charges in the 30k to 55k range

Multi-Variate Analysis

BMI & Gender

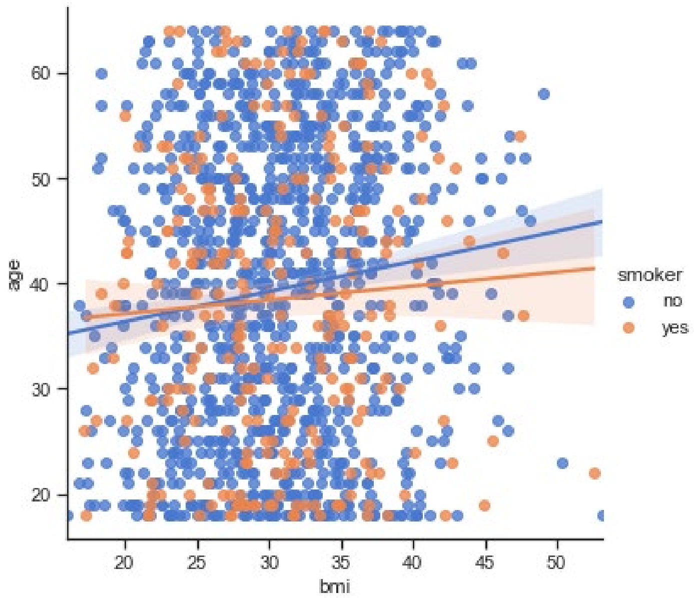




Observations

- BMI distribution and gender have nearly identical distributions and means and do not appear to be correlated
- Males have a few higher outliers

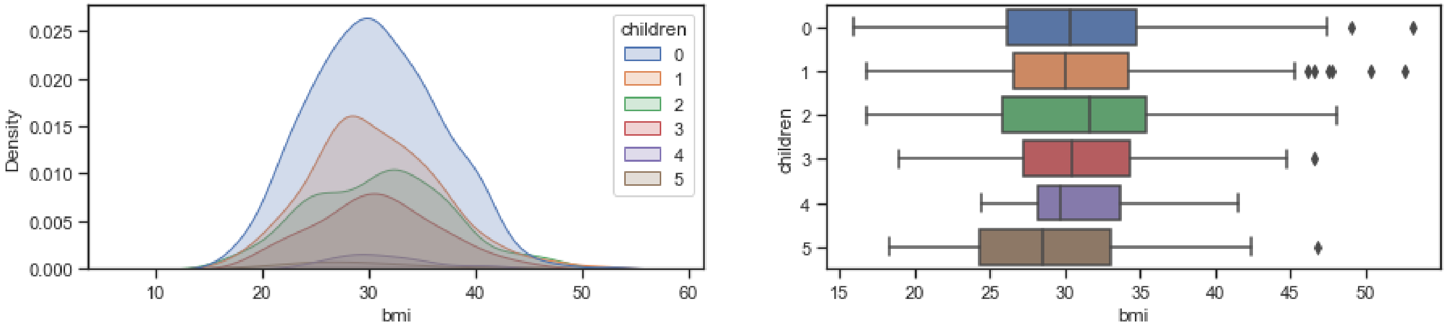
BMI & Smoker



- BMI for smokers and non-smokers are nearly evenly distributed
- Non-Smokers have more outliers but account for nearly 50% of the population
- There appears to be little to no correlation between bmi, age & smoking

Multi-Variate Analysis

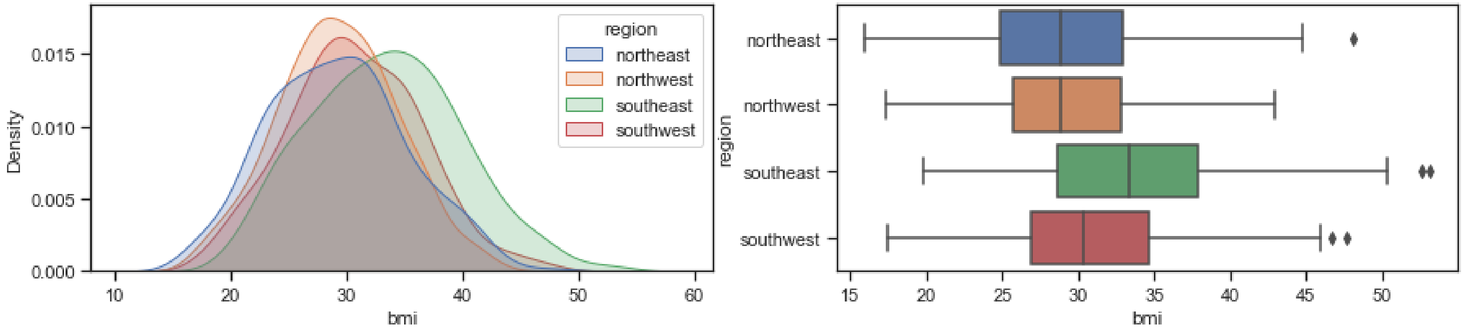
BMI & Children



Observations

- BMI and children seem to have little correlation
- BMI distribution across children is faily even

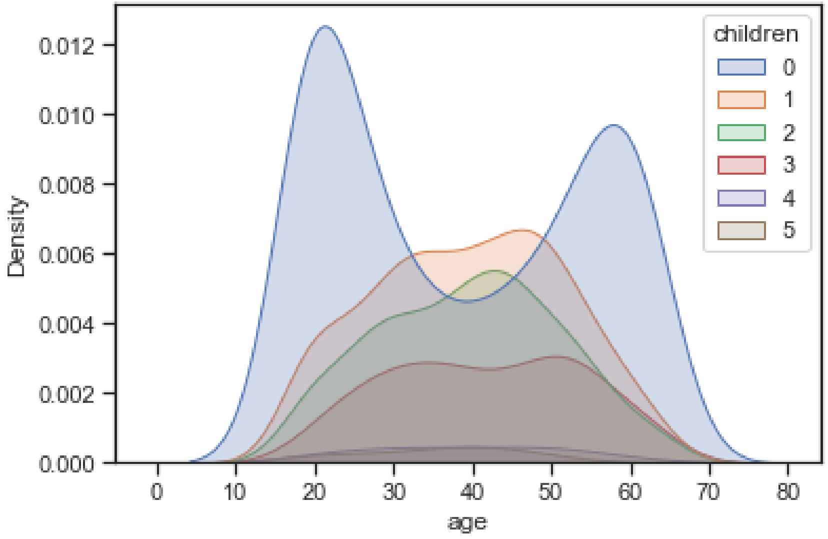
BMI & Region



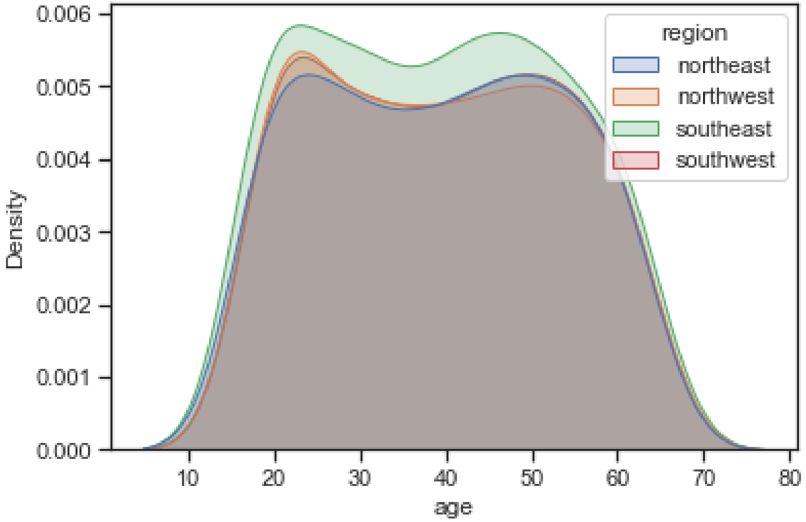
- The Southeast has a higher mean BMI by 4 - 6 points than other regions
- Southeast's BMI distribution is the highest of the regions
- No regions IQR is within the normal bmi range
- Southeast is the only region with BMIs higher than 48
- Southeast has no BMIs lower then 20

Multi-Variate Analysis

Age & Children



Age & Region



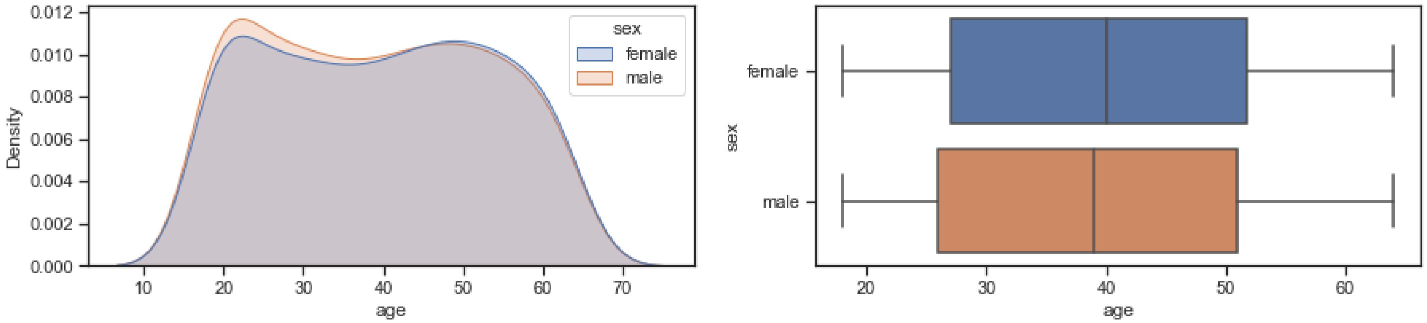
Observations

- For 1 or more children there seems to be a normal distribution across ages
- For members with no children the age distribution seems to be bimodal with peaks around 22 and 60
- The bimodal age distribution for members with no children is slightly reflected in other bi-variate analyses, though due to the normal peak for other age / children groups it starts to look more like a normal distribution, it could be intteresting to further study why no children is different
- Speculation: People may not be having children until they are in their 20's and start losing some children when they are in their 40's

- Age seems to be fairly evenly distributed across regions with slightly more in the SouthEast region

Multi-Variate Analysis

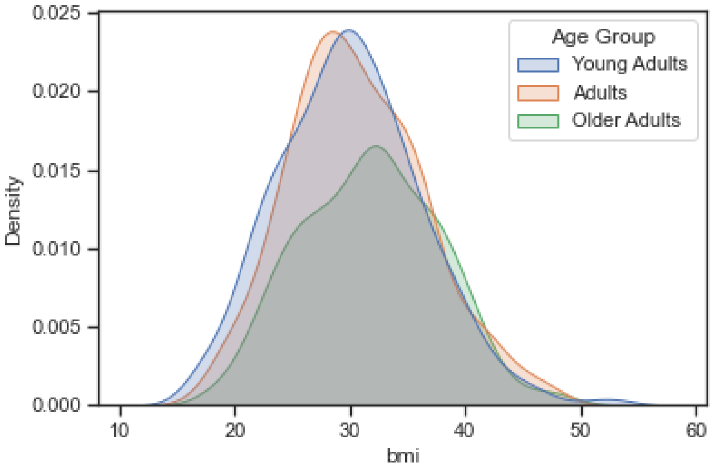
Age & Sex / Gender



Observations

- Male and Female are evenly distribut across age
- Females have a slightly higher mean and 75th percentile

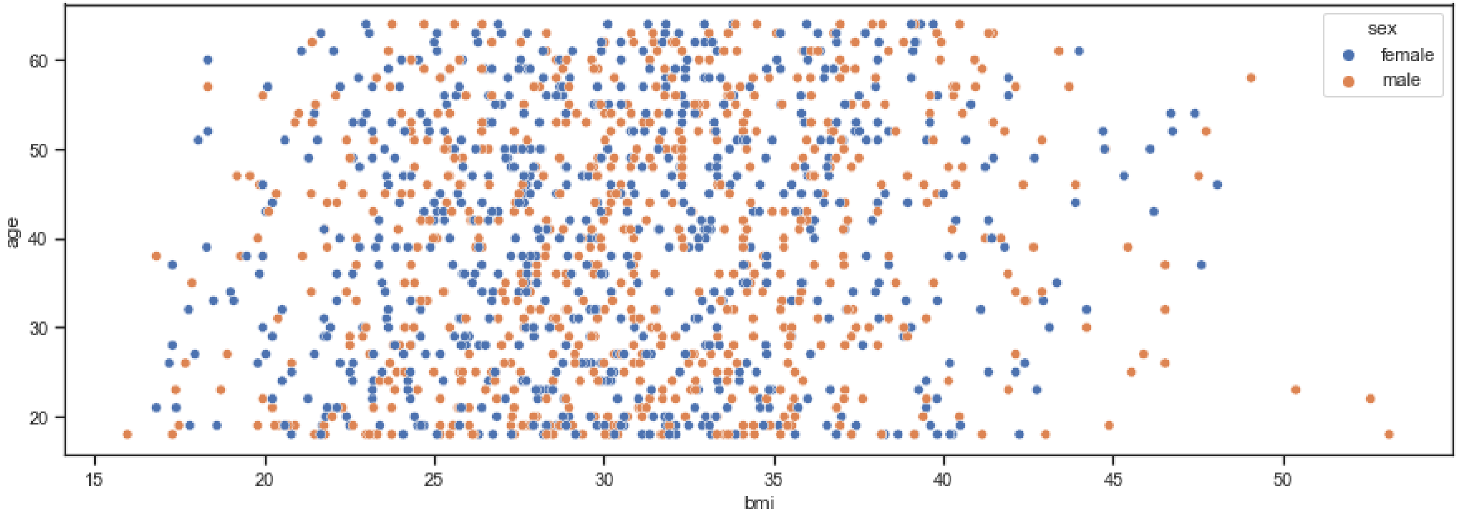
BMI & Age Group



- Older Adults have a higher mean at about 33
- Young Adults and Adults have a similar mean about 28

Multi-Variate Analysis

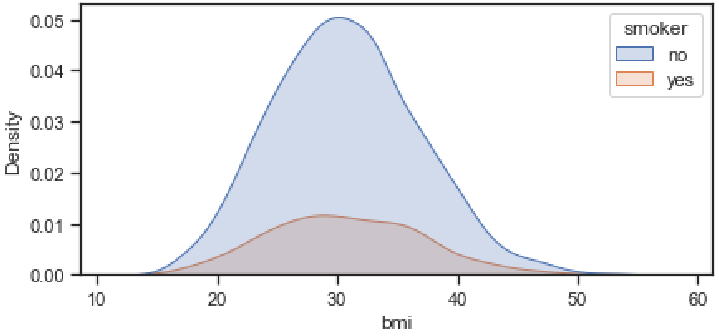
BMI, Age & Sex / Gender

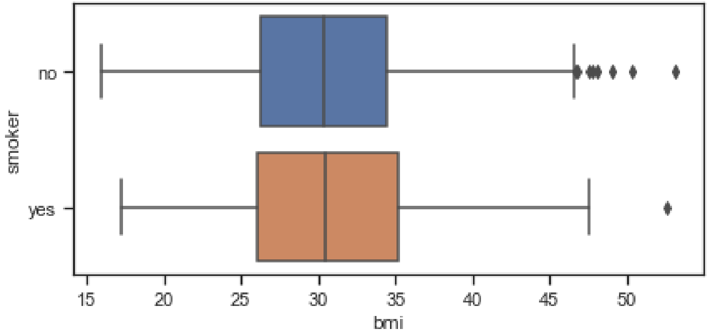


Observations

- There seems to be no correlation between BMI, age & gender
- The highest BMI's > 50 are male young adults

BMI & Smoking

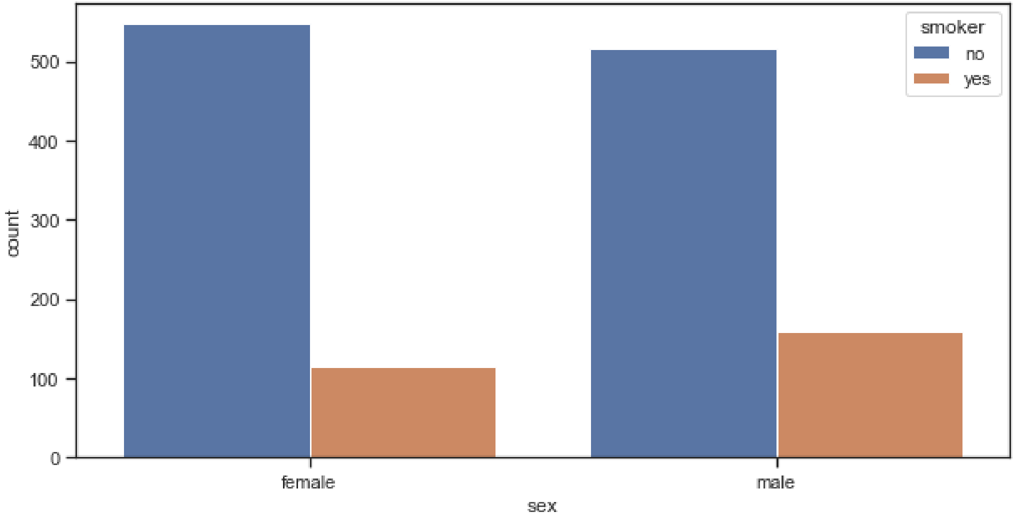




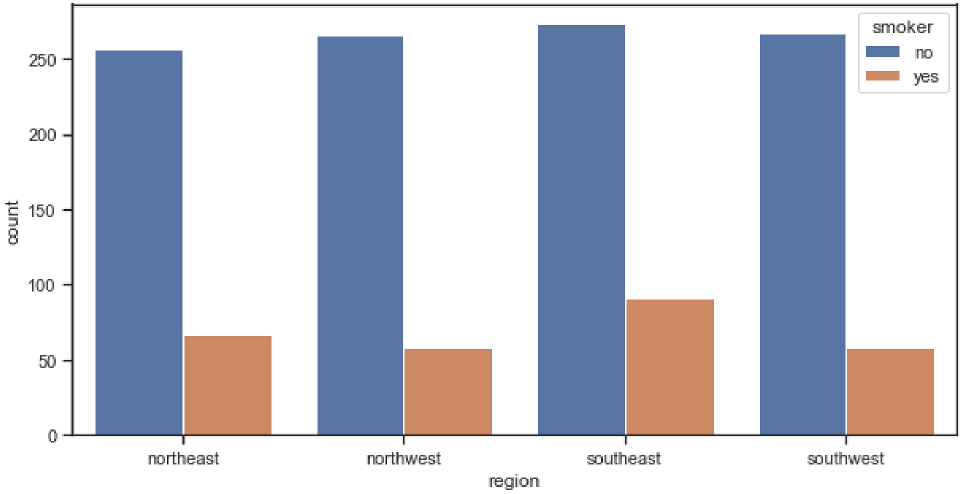
- BMI distribution and mean are very similar for smokers and non-smokers

Multi-Variate Analysis

Smoker & Sex / Gender



Smoker & Region



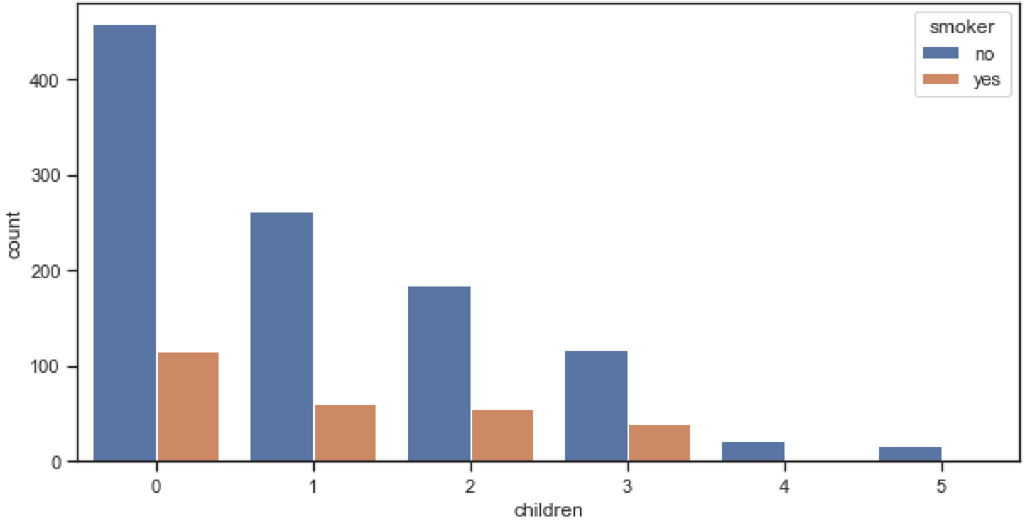
Observations

- There are slightly more male somkers than female smokers

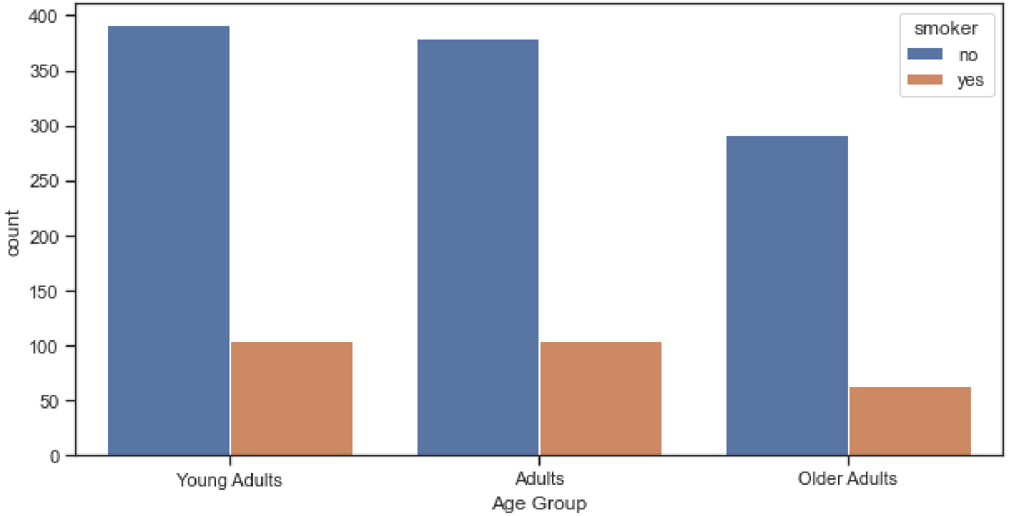
- Smokers are fairly evenly distributed across regions
- Non- Smokers are fairly evenly distributed across regions
- There are slightly more smokers in the SouthEast Region

Multi-Variate Analysis

Smoker & Children



Smoker & Age Group



Observations

- Members with no children have more smokers, they also represent 50% of the population

- There are very few smokers among members who have 4 or 5 children

- Smokers are fairly evenly distributed across age group with Older Adults being slightly less