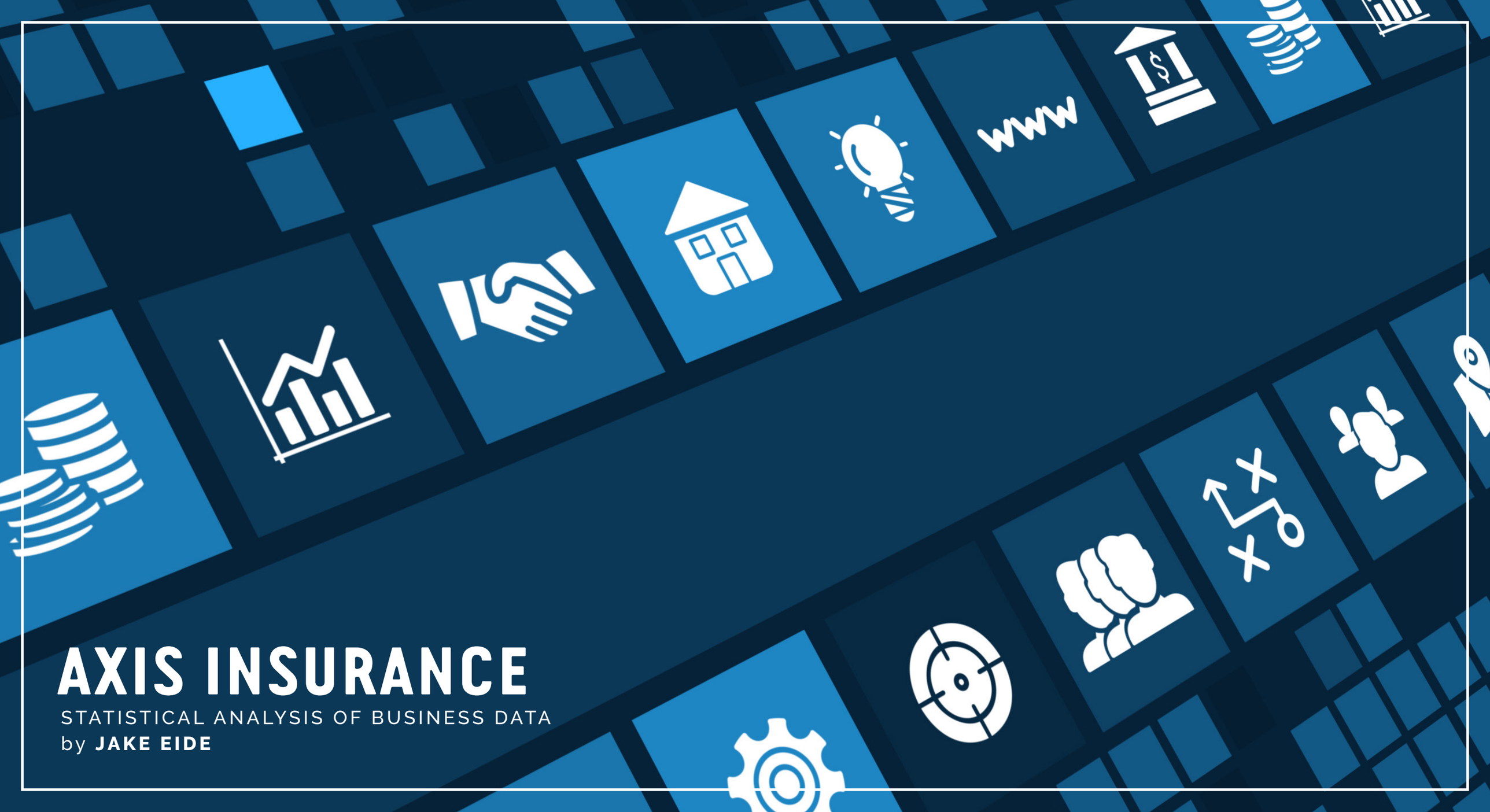


AXIS INSURANCE

STATISTICAL ANALYSIS OF BUSINESS DATA

by **JAKE EIDE**

BACKGROUND & OBJECTIVE

Axis Insurance is a business seeking to leverage existing customer information in making future business decisions. The objective of this analysis is to explore the company's dataset and extract insights using Exploratory Data Analysis. Insights will be analyzed further via Hypothesis Testing.

KEY QUESTIONS

1) What insights can be extracted from the Axis Insurance dataset?

2) Smokers: Are there more medical claims made by Axis customers who smoke than those who don't smoke?

3) Smokers: Is the proportion of smokers significantly different across different regions?

4) BMI: Is the BMI of females different than the BMI of males?

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

DATA INFORMATION

The data contains information about 1338 Axis Insurance customers.

VARIABLE	DESCRIPTION	DATA TYPE	RANGE
Age	an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government)	Integer	18 - 64 years
Sex	the policy holder's gender, either male or female	Category	male, female
BMI	the body mass index, 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.	Float	15.96 - 53.13
Children	an integer indicating the number of children / dependents covered by the insurance plan	Integer	0 - 5
Smoker	yes or no depending on whether the insured regularly smokes tobacco	Category	yes, no
Region	the beneficiary's place of residence in the U.S., divided into four geographic regions	Category	northeast, southeast, southwest, northwest
Charges	Individual medical costs billed to health insurance	Float	1121 - 63770

EXPLORATORY DATA ANALYSIS - UNIVARIATE: **AGE**

Axis customers range in age from 18 to 64. Looking at the histogram and box plots to the right, we see the age distibution of Axis customers is fairly uniform, except for a larger number of 18 and 19 year olds.

Observations:

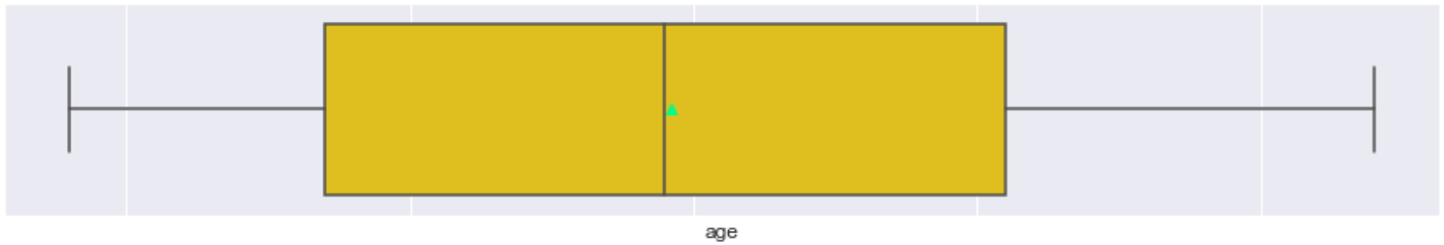
• The distribution of age is right skewed, caused by the large number of customers aged 18 and 19 years.

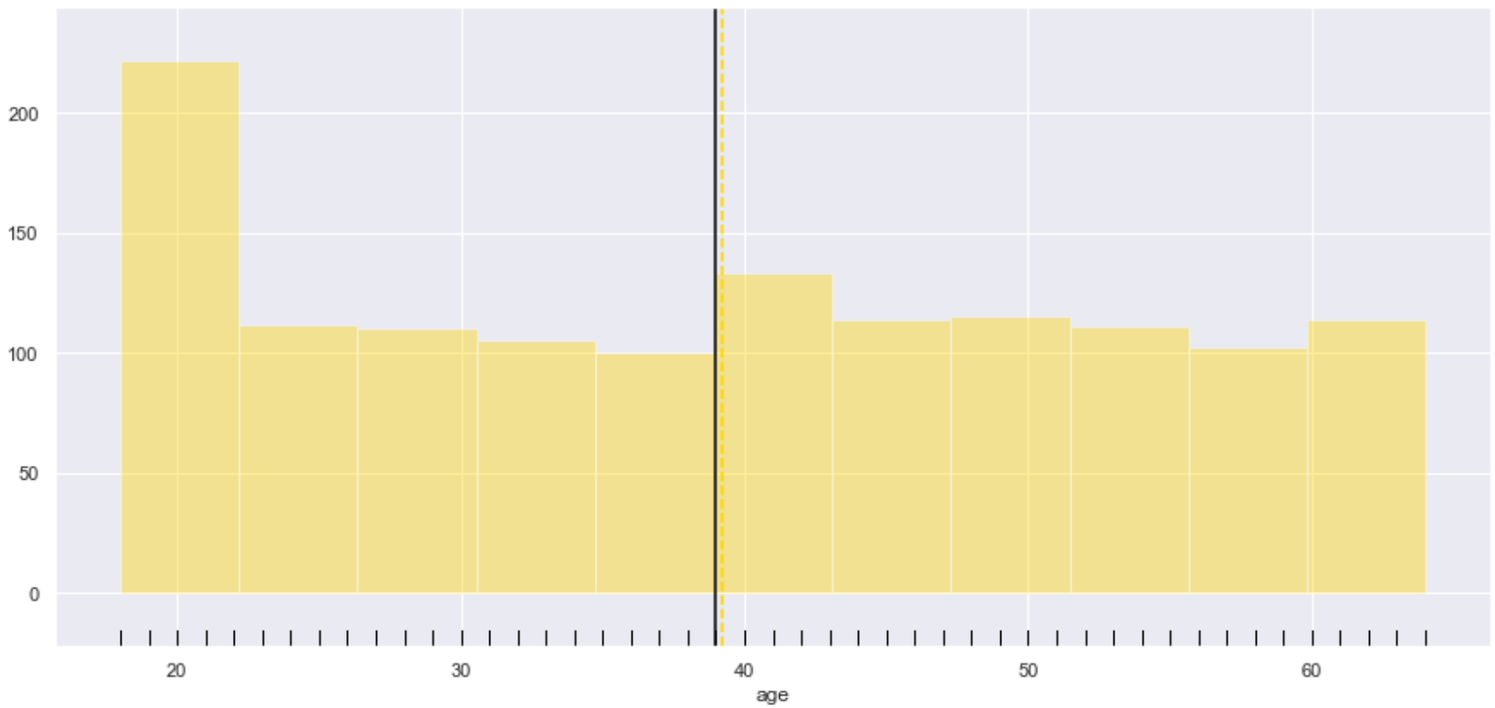
Observations on Central Tendency:

• Mean Age is 39.2 years old
• Median Age is 39 years old
• Mode is 18 years old
• The middle 50% of customers are 27 to 51 years old

Standard Deviation:

• The standard deviation for age is 14.05, indicating that the data points are spread out and are not clustered around the mean





EXPLORATORY DATA ANALYSIS - UNIVARIATE: **BMI**

The body mass index provides a sense of how over or under-weight a person is relative to their height. An ideal BMI is within the range of 18.5 to 24.9.

Observations:

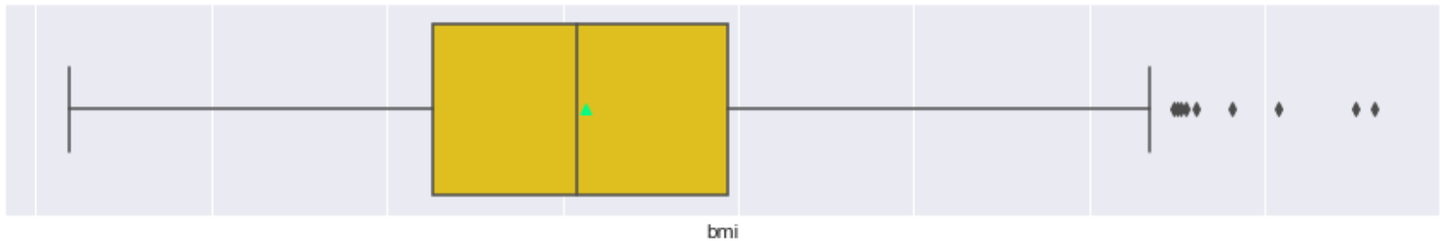
• The BMI distibution for Axis customers appears mound-shaped and symmetrical around 30.5, with a slight amount of right-skew.
• The BMI values range from 15.96 to 53.13.

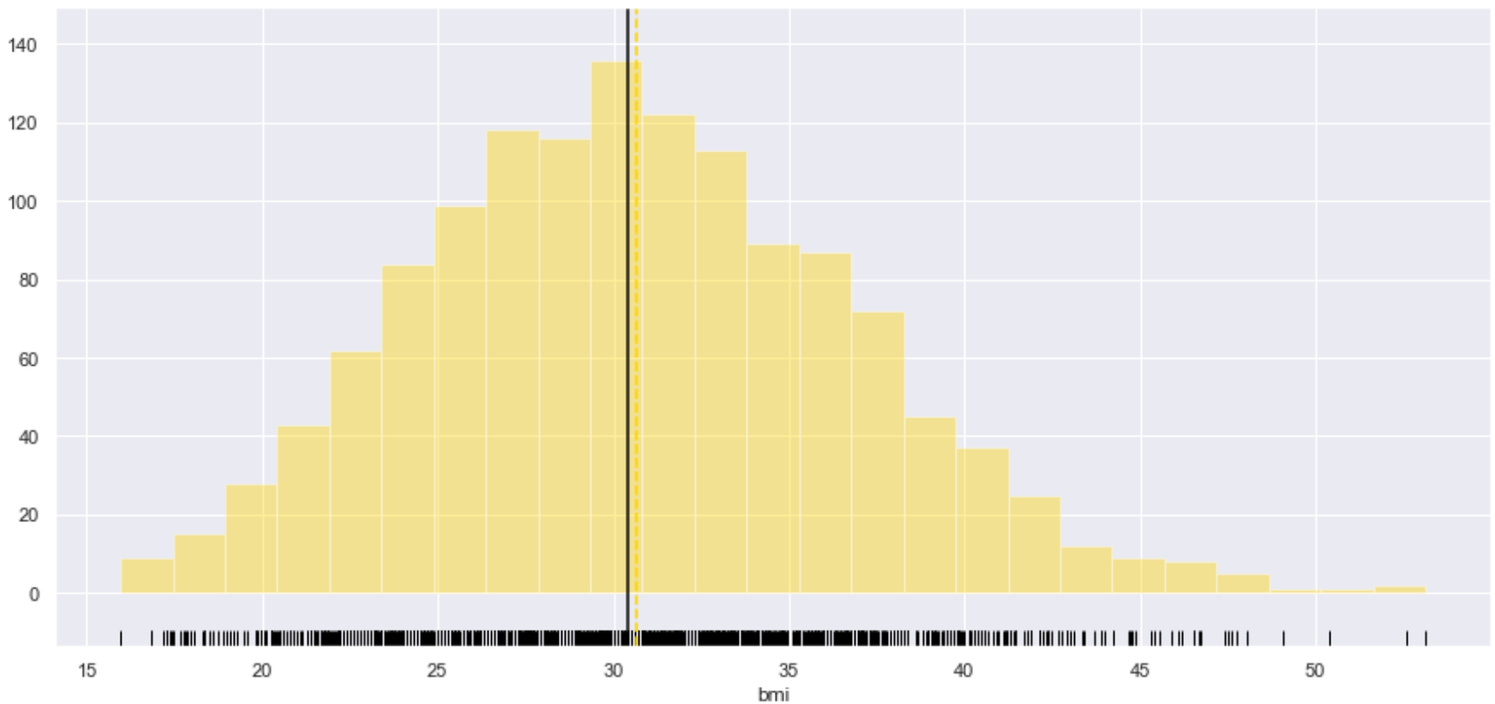
Observations on Central Tendency:

• Mean BMI is 30.7 kg/m²
• Median BMI is 30.4 kg/m²
• The middle 50% of customers are between 26.3 and 34.7 kg/m²

Insight:

• An ideal BMI is within the range of 18.5 to 24.9, and the average Axis customer has a BMI around 30.5 – higher that the ideal range.





EXPLORATORY DATA ANALYSIS - UNIVARIATE: **CHILDREN**

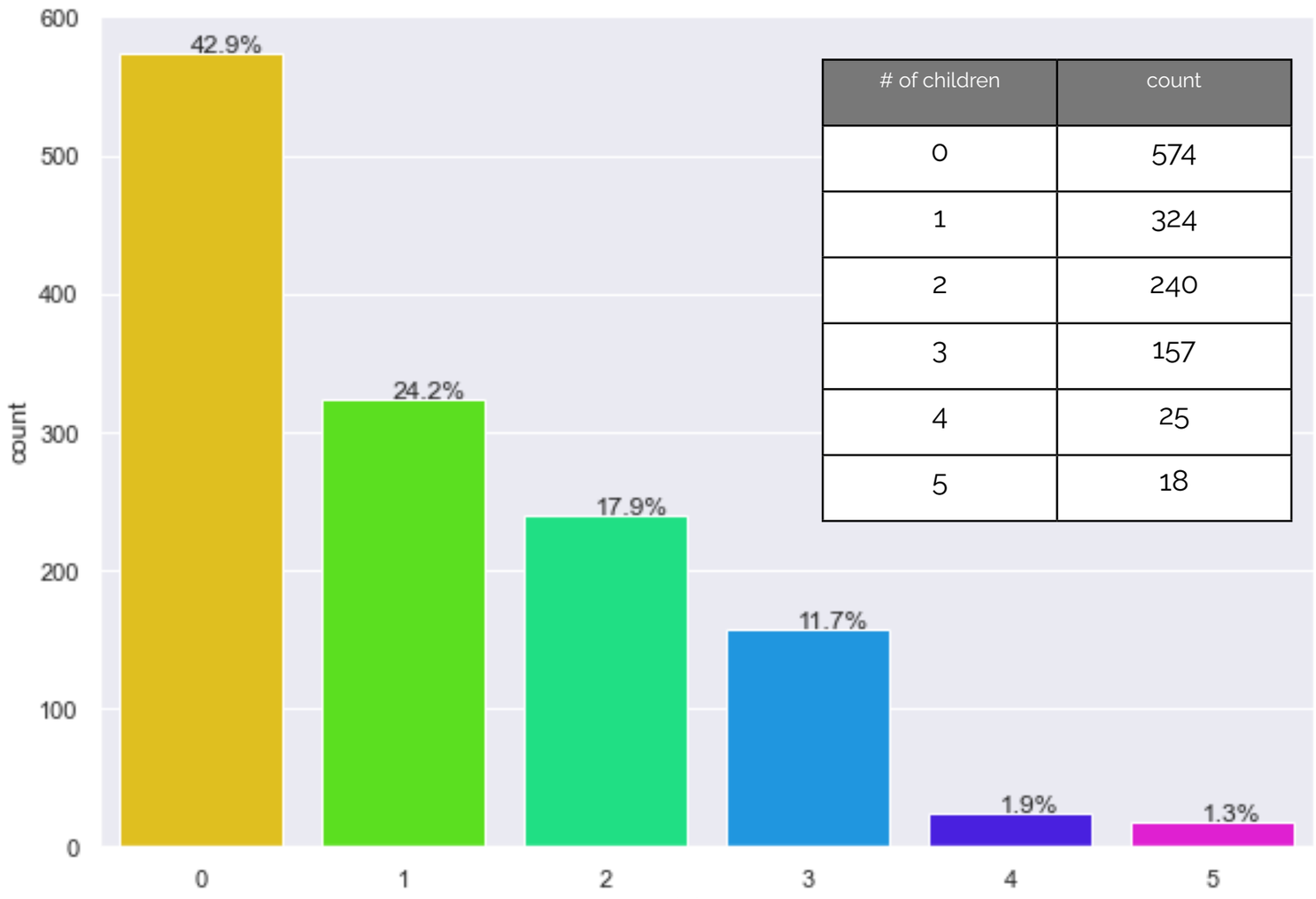
This distribution shows the number of children each Axis customer has – ranging from 0 to 5 children.

Observations:

- The distibution is heavily right skewed.
- The percentage of customers gets progressively smaller with each child
- 67% of customers have either no children or only 1 child
- 3% of customers have either 4 or 5 children

Observations on Central Tendency:

- Mean is 1.1 child
- Median 1 child
- Mode is 0 children



# of children	count
0	574
1	324
2	240
3	157
4	25
5	18

EXPLORATORY DATA ANALYSIS - UNIVARIATE: **CHARGES**

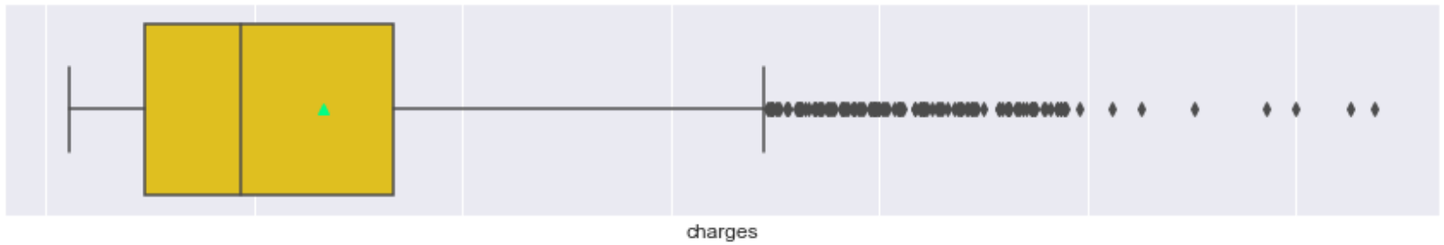
This distribution shows the medical costs billed to health insurance by Axis customers.

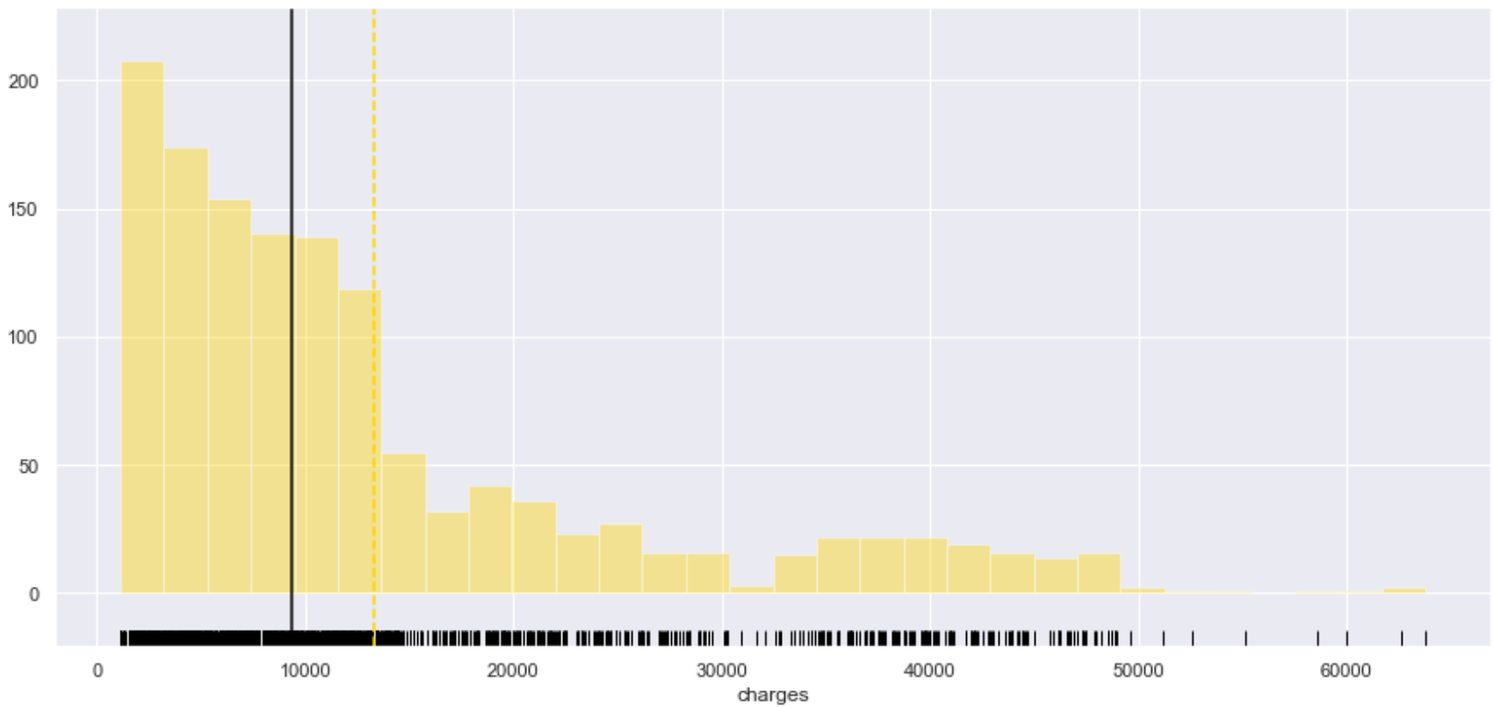
Observations:

• The distibution is heavily right skewed.
• There are outliers with higher charges: the largest charge was $63,770
• From the graph above we can see that there is a large falloff in the number of charges above $14,000

Observations on Central Tendency:

• Mean is $13,270
• Median is $9,382
• The middle 50% of charges were between $4,740 and $16,640



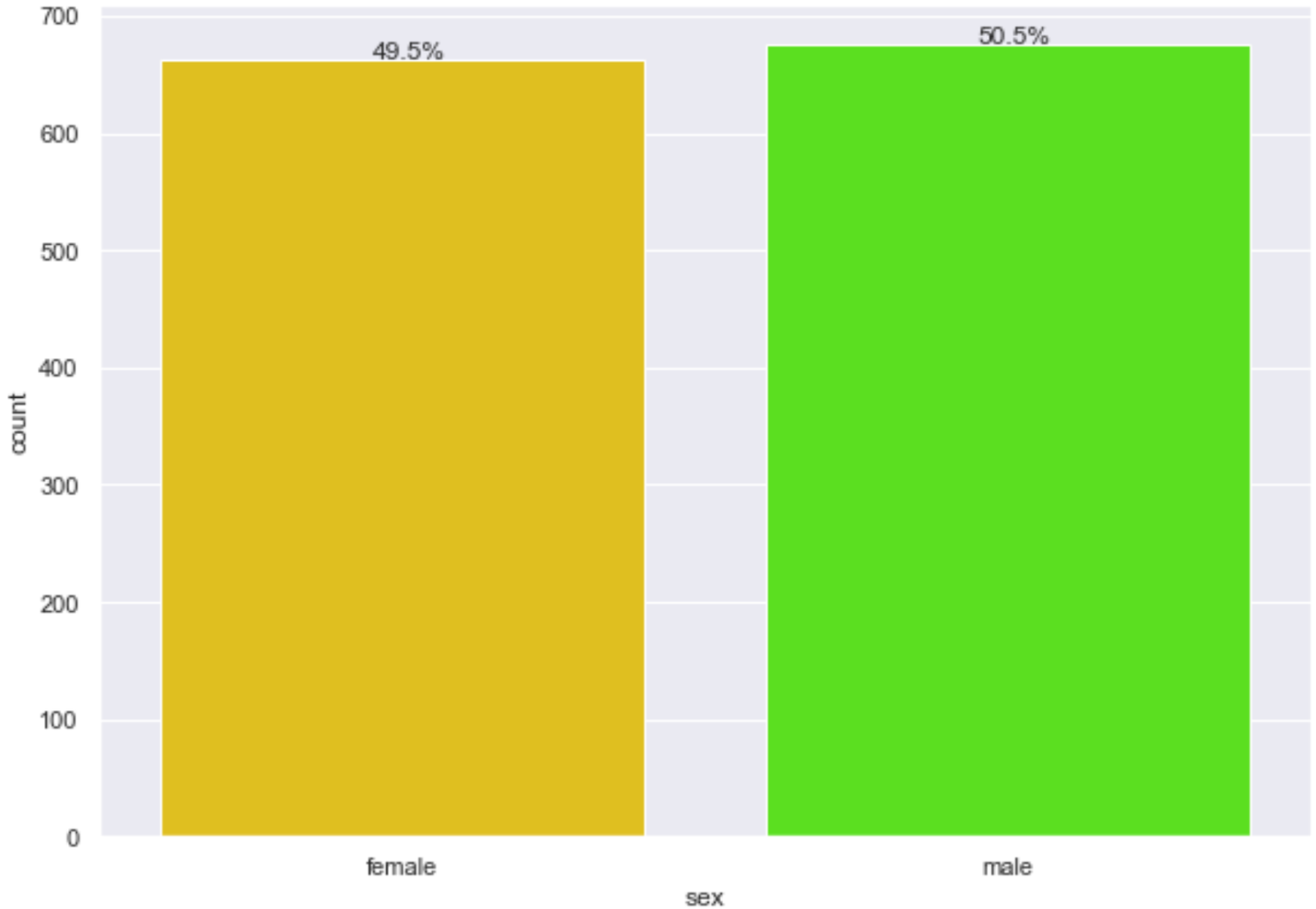


EXPLORATORY DATA ANALYSIS - UNIVARIATE: **SEX**

This chart shows the gender distribution of customers.

Observations:

• Axis customers are divided nearly equally among two genders, with only slightly more male customers.



EXPLORATORY DATA ANALYSIS - UNIVARIATE: **SMOKING STATUS**

This chart shows the distribution of customers who smoke tobacco.

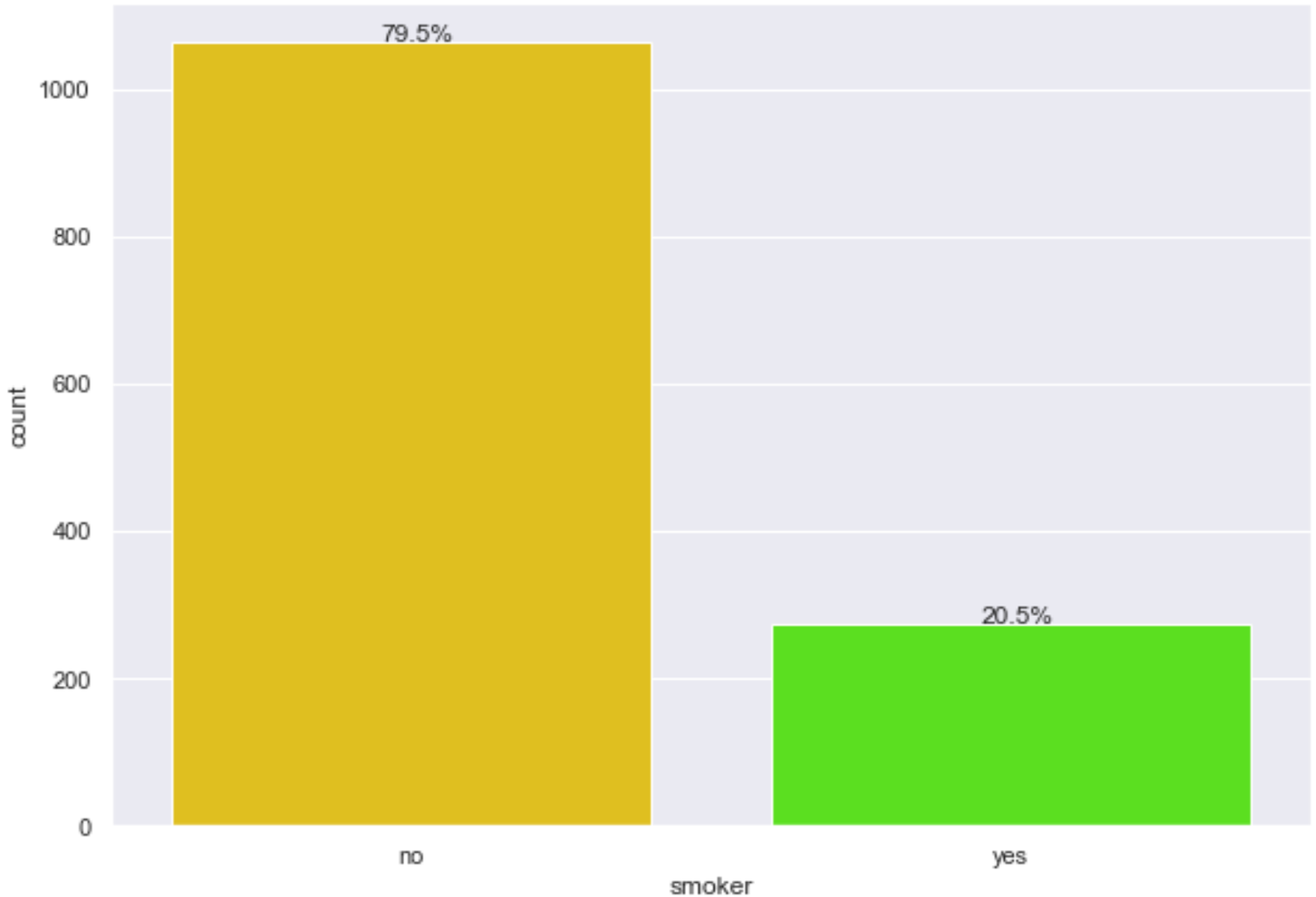
Observations:

• Axis customers are heavily skewed towards non-smokers.

• Smokers account for roughly 1 in 5 (20%) of Axis customers.

Insight:

• One might read this statistic and conclude that Axis customers are overwhelmingly concerned with the health consequences of smoking. This may be the case, or Axis customers may follow the same distribution of the U.S. population. According to the CDC*, in 2019, 14% of adults in the U.S. currently smoked cigarettes. It would be wise to compare the Axis dataset to the U.S. population from the same year.



***Centers for Disease Control and Prevention. "Current Cigarette Smoking Among Adults in the United States" www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm [accessed 2021 April 15]**

EXPLORATORY DATA ANALYSIS - UNIVARIATE: **REGION**

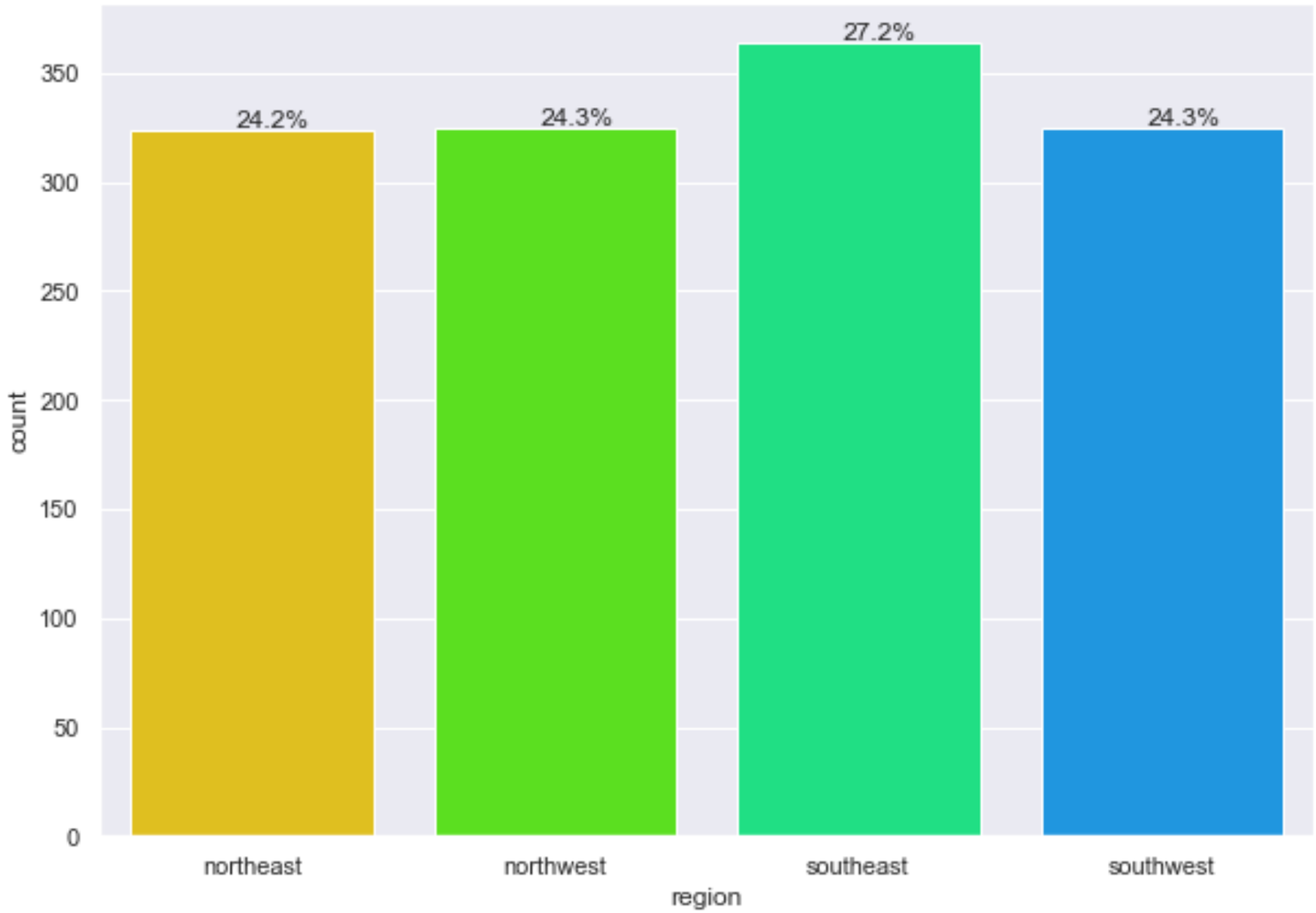
This is the beneficiary's place of residence in the U.S..

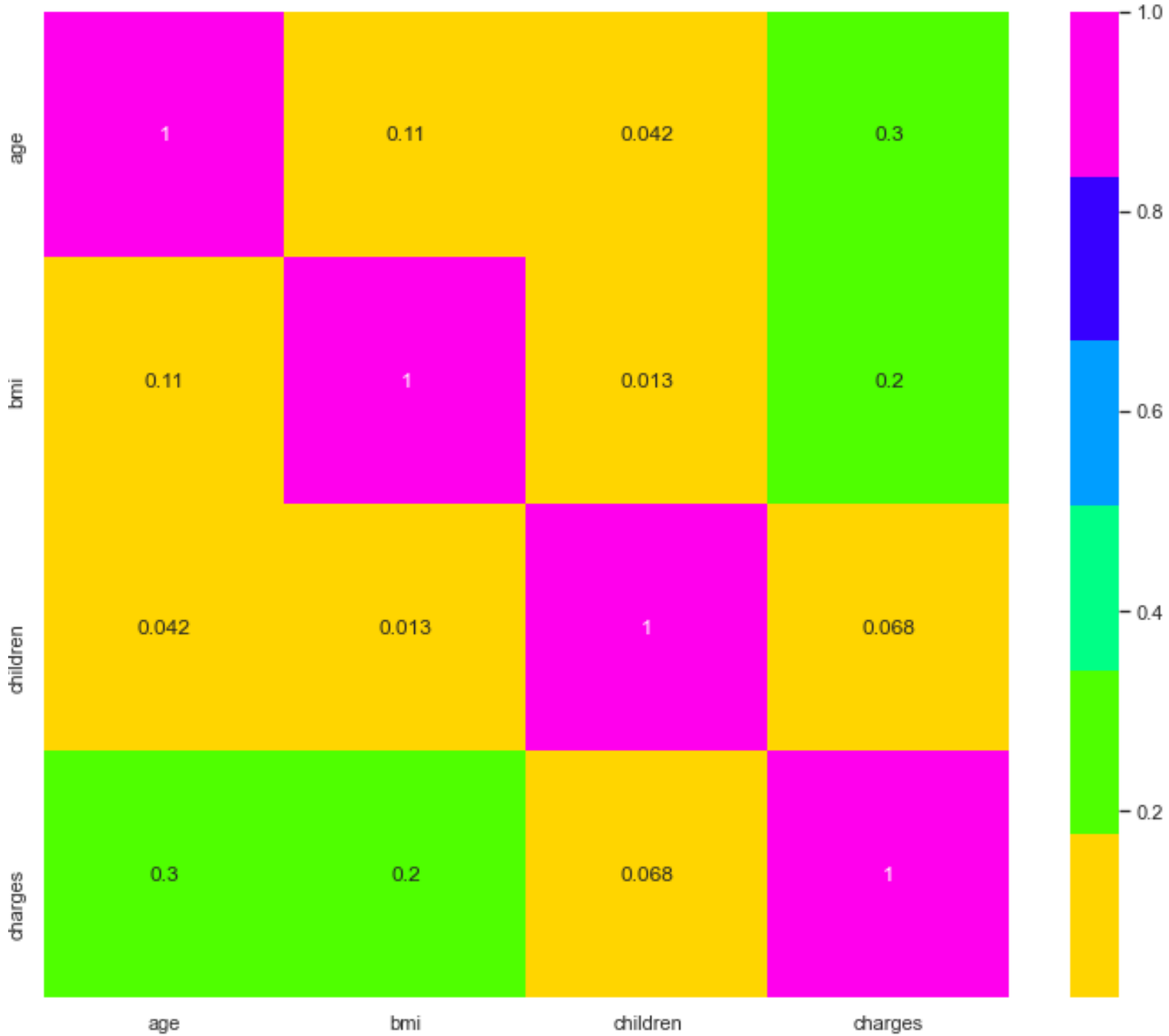
Observations:

• Axis customers are fairly evenly divided between the four regions.

• The Southeast has a slightly higher percentage of the total customers than the other three regions.

• The other three regions (Southwest, Northwest, and Northeast) are very evenly distributed.





EXPLORATORY DATA ANALYSIS:
BIVARIATE CORRELATION

Observations:

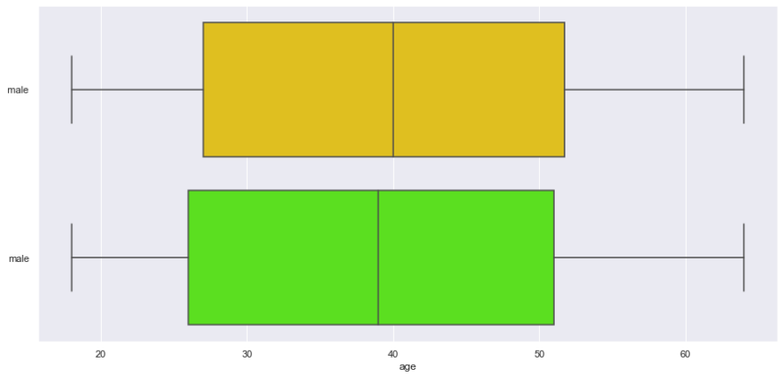
• The correlation levels between these four variables is fairly low (less than 0.3)

• The highest correlation is between age and charges, but at 0.042, even this is a moderate level of correlation.

Insight:

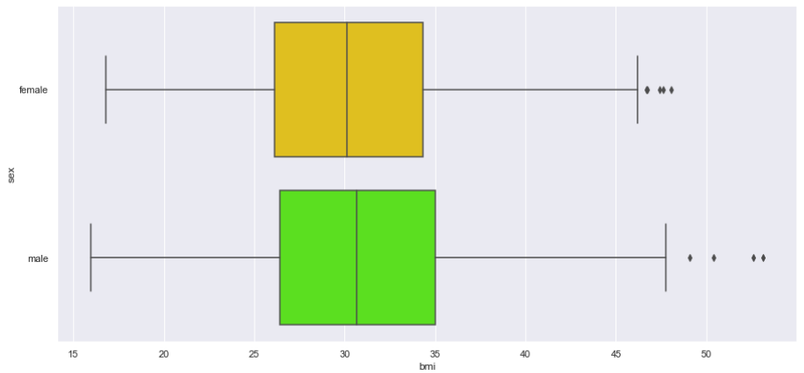
• This heatmap only shows correlation between the four numerical variables. With further bivariate analysis including categorical variables, we may be able find more relationships in the data.

EXPLORATORY DATA ANALYSIS - BIVARIATE:
SEX VS. OTHER VARIABLES



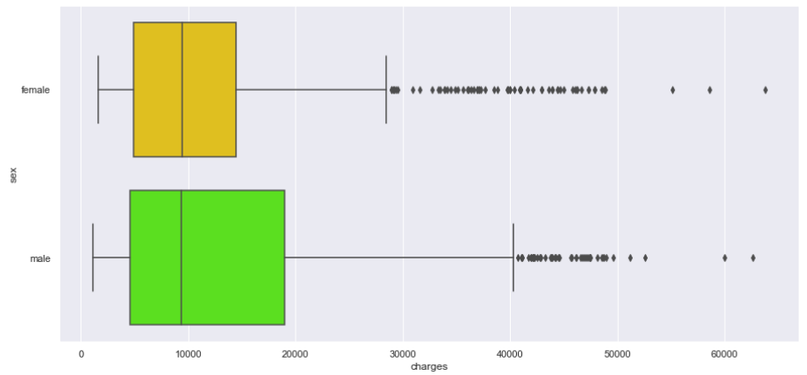
Sex vs. Age

• Female customers are slightly older than their male counterparts



Sex vs. BMI

• Male customers have a slightly higher BMI – and a larger range of BMI values – than female customers



Sex vs. Charges

• Median charges are very similar for both genders

• Males have a higher right skew than females in terms of charges

EXPLORATORY DATA ANALYSIS - BIVARIATE: **SMOKERS VS. CHARGES**

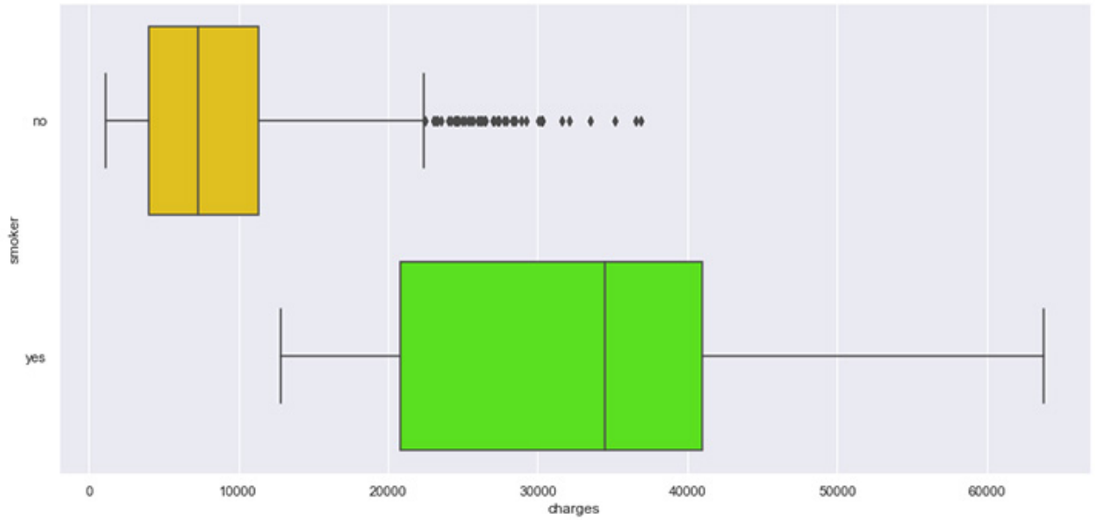
Key Question:

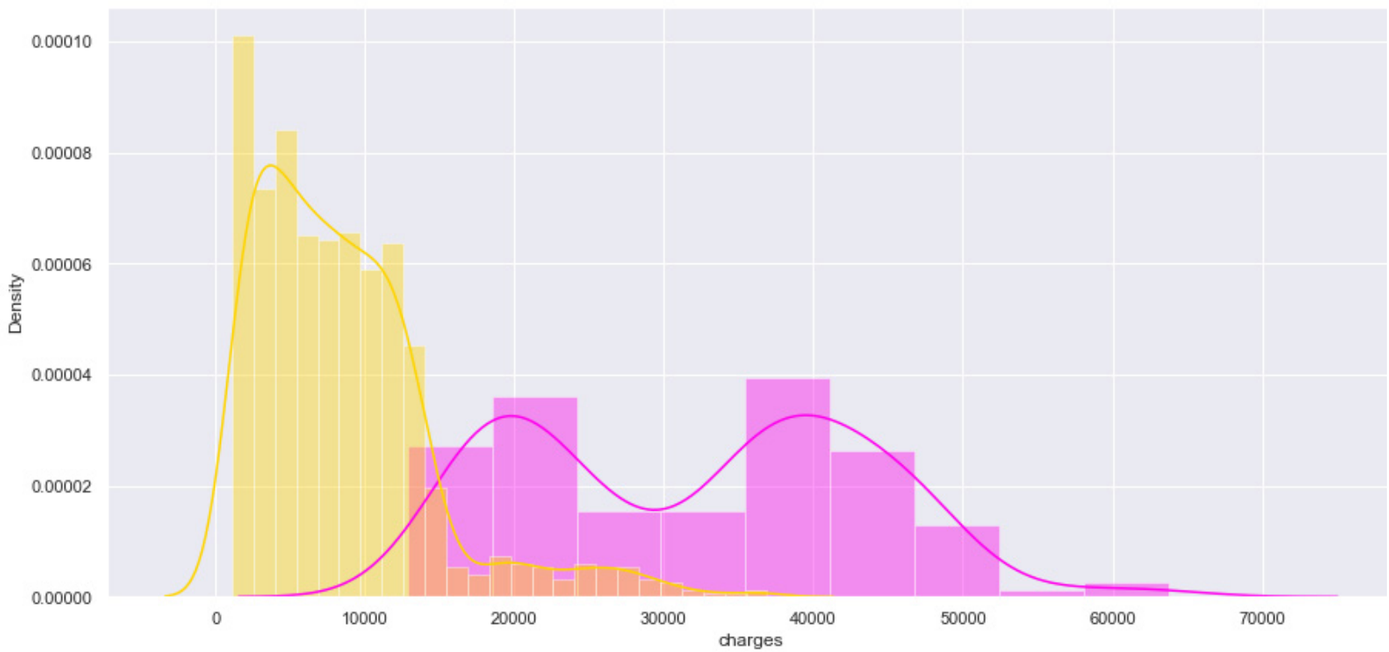
• Are there more medical claims made by people who smoke than those who don't smoke?

Statistical Test:

• We tested the above question using a 95% confidence level

• The test concludes that there is a difference between the mean charges made by smokers vs. non-smokers





EXPLORATORY DATA ANALYSIS - BIVARIATE: **SMOKERS VS. REGION**

Key Question:

• Is the proportion of smokers significantly different across regions?

• The graphs below show the proportion of smokers to non-smokers in each of the four regions.

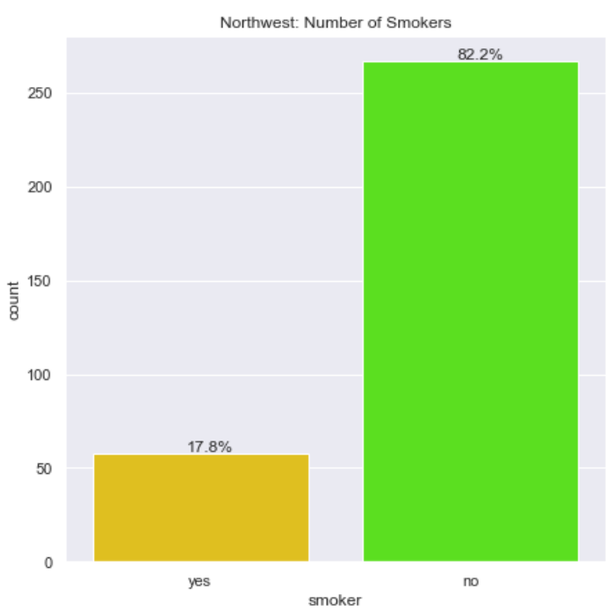
• The ratios vary from a 64.4% difference in the two western regions, to a 75% difference in the Southeast. Is this percentage difference significant enough to draw any conclusions?

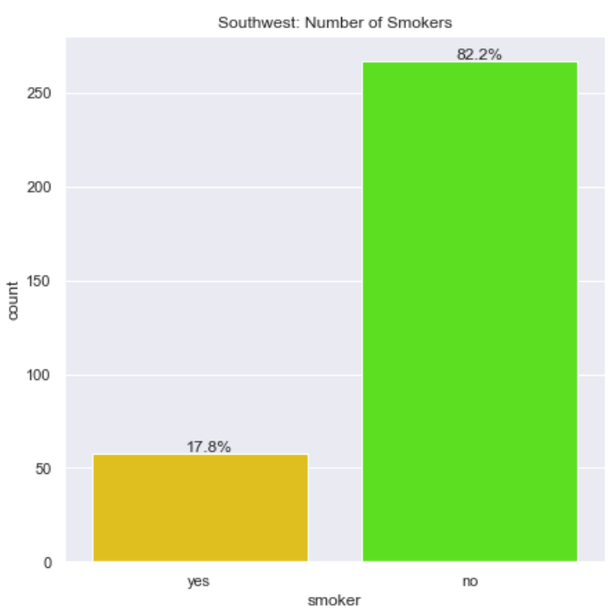
Statistical Test:

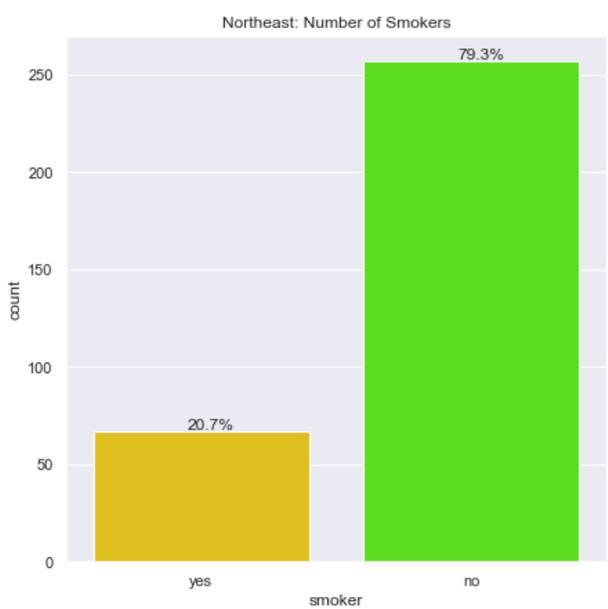
• We tested the above question using a 95% confidence level

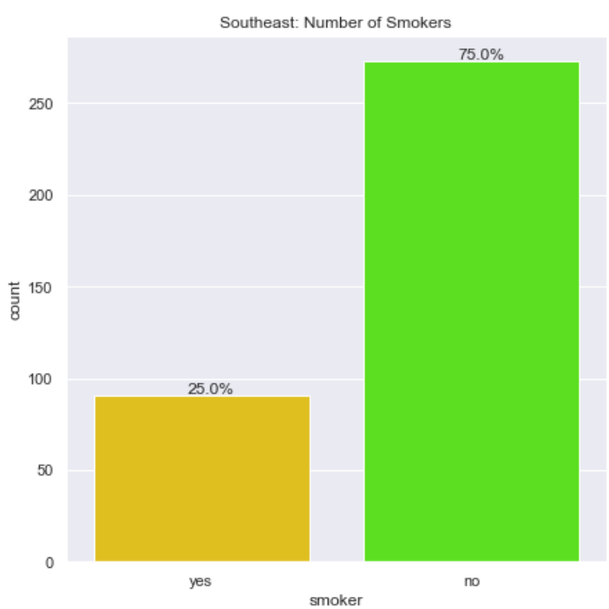
• The test concludes that the proportion of smokers is **NOT** significantly different across different regions.

• This means that even though there is a higher percentage of smokers in the Southeast U.S., this percentage is not high enough to say that there is a *significant* difference.









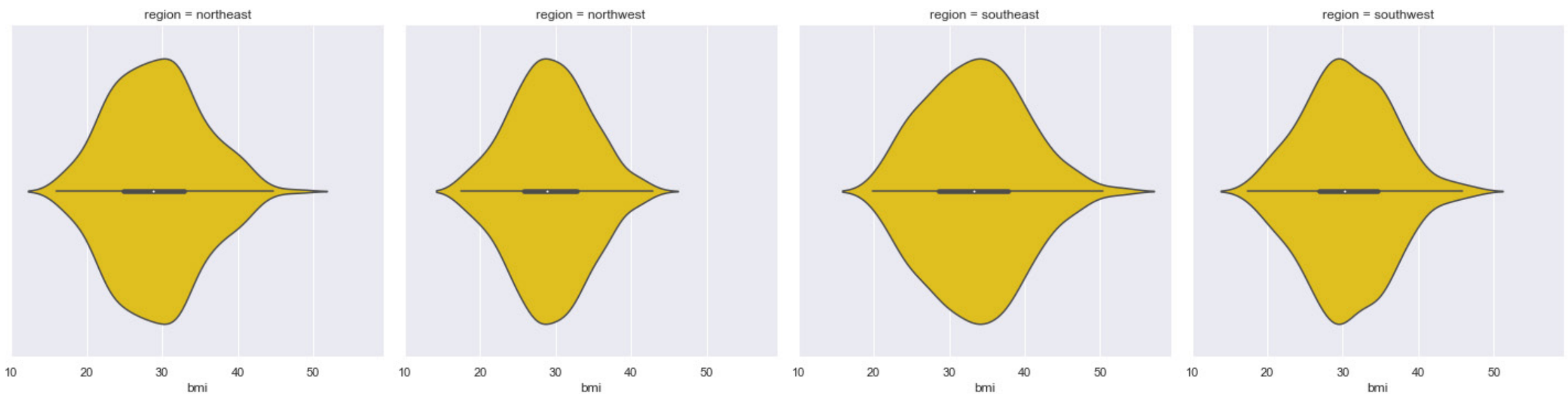
EXPLORATORY DATA ANALYSIS - BIVARIATE: **REGION VS. BMI**

Observations:

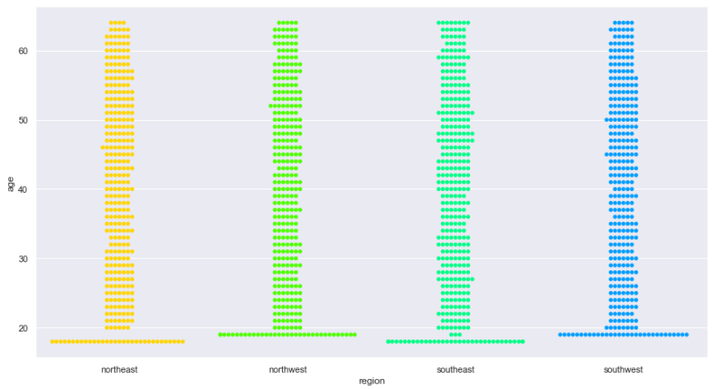
• The Southeast has the highest median BMI, as well as the highest overall BMI.

• The Northwest has the most clustered BMI.

• The Northeast and Southwest have similar medians and spreads to each other.

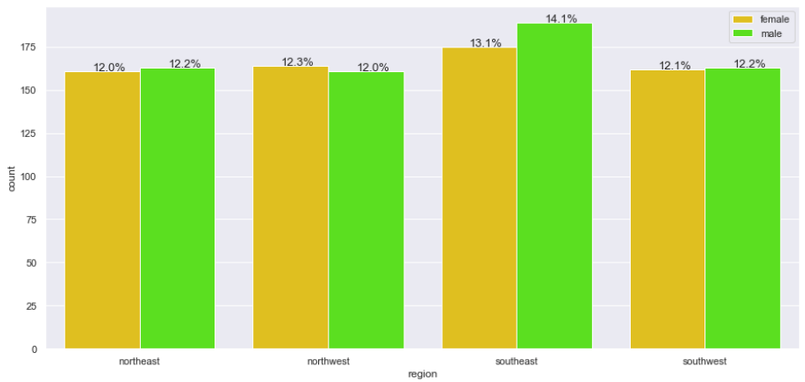


EXPLORATORY DATA ANALYSIS - BIVARIATE:
REGION VS. OTHER VARIABLES



Region vs. Age

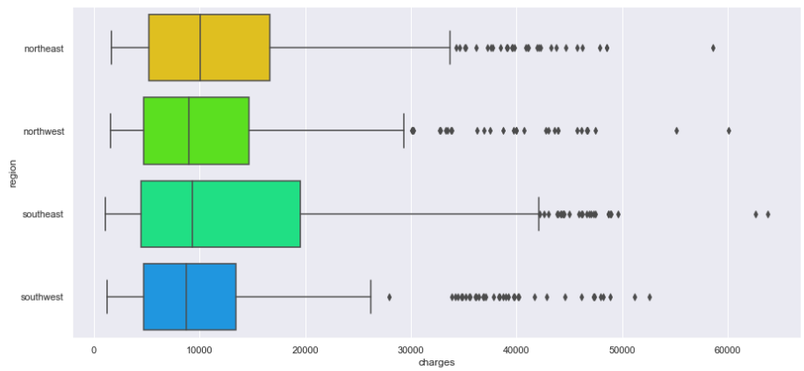
• The age distribution is very similar across the four regions



Region vs. Sex

• The two genders have almost equal proportions in the Northeast, Northwest, and the Southwest

• The overall male skew shows up more in the Southeast than any other region



Region vs. Charges

• The median charges do not appear to be significantly different across regions

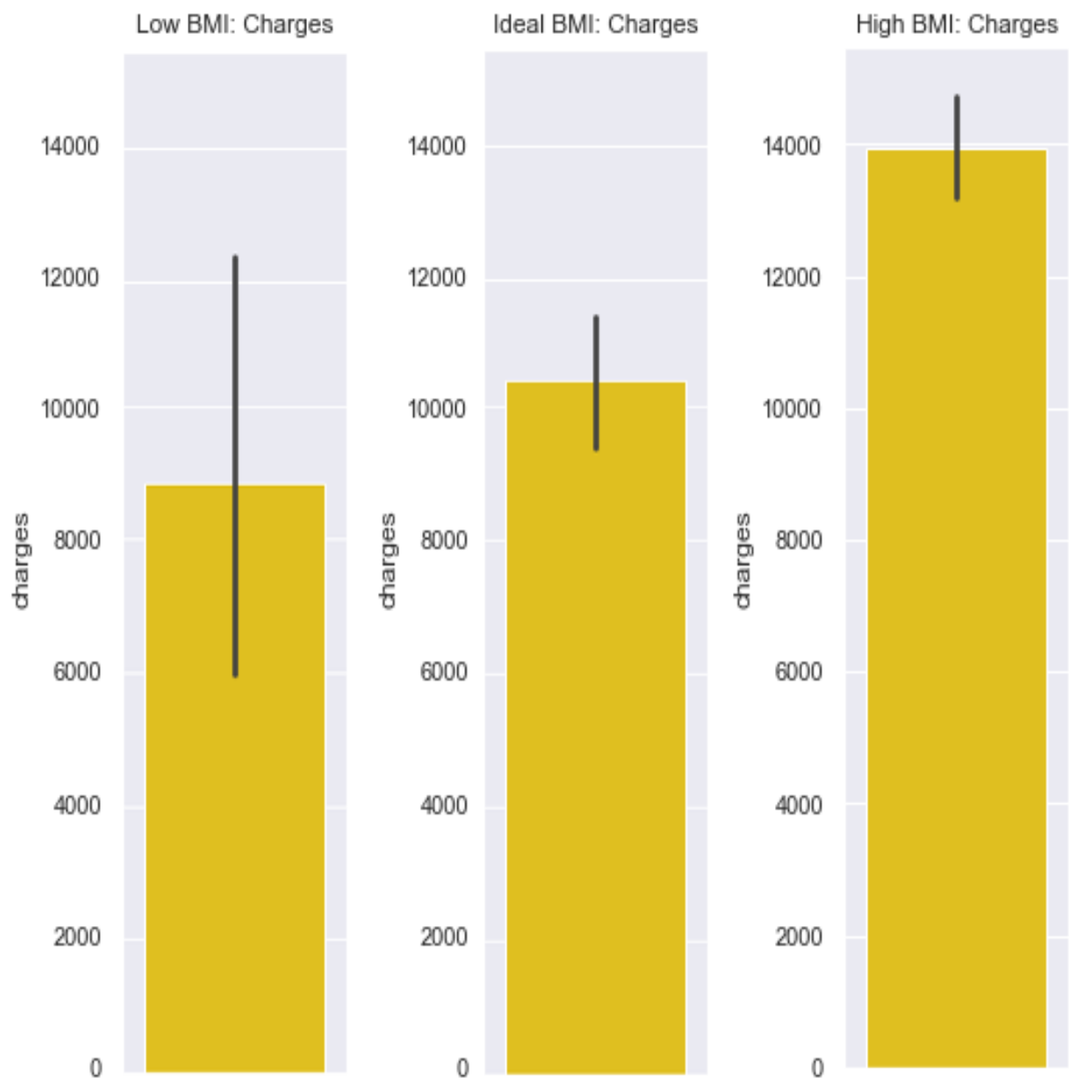
• The spread of charges is noticably different across the regaions, with the Southeast having the largest spread, and the Southwest having the smallest spread

EXPLORATORY DATA ANALYSIS - BIVARIATE: **BMI VS. CHARGES**

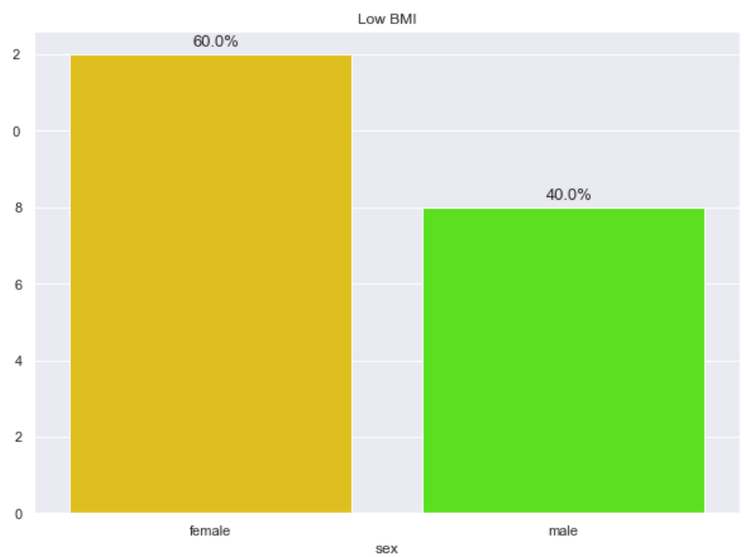
Low BMI – Charges
- Customers with a BMI less than 18.5
- Mean charges = $8852
- Median charges = $6759

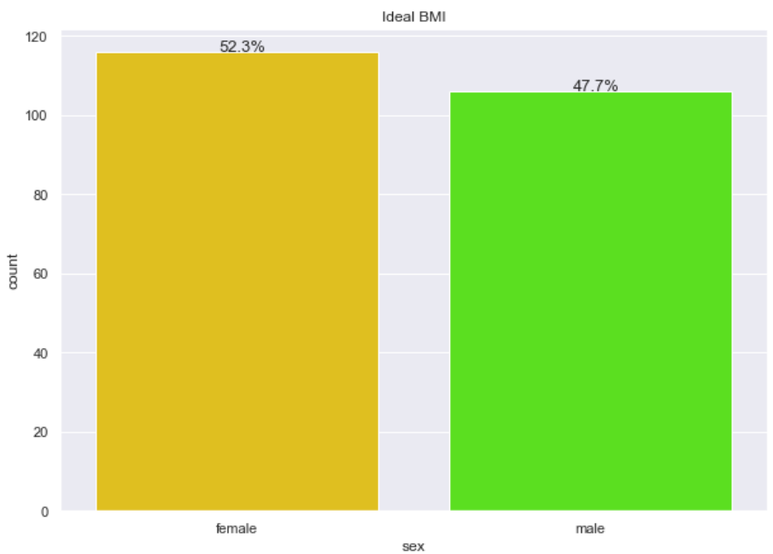
Ideal BMI – Charges
- BMI between 18.5 and 24.9
- Mean charges = $10379
- Median charges = $8604

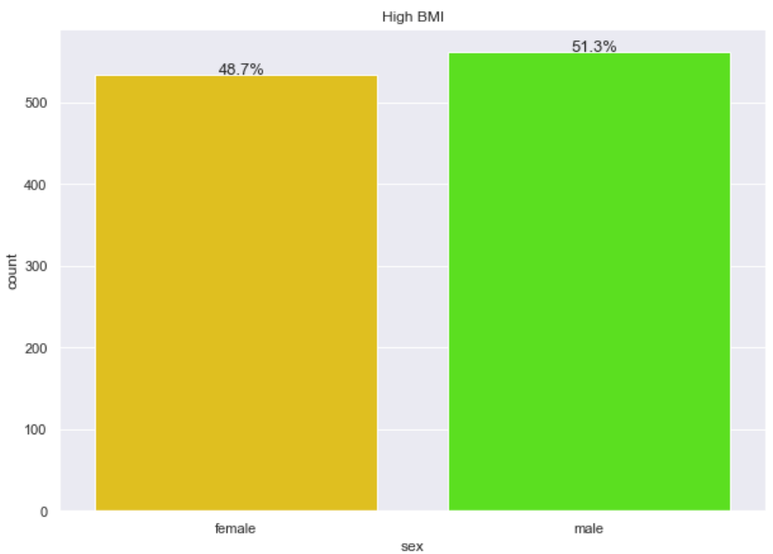
High BMI – Charges
- Customers with a BMI above 24.9
- Mean charges = $13936
- Median charges = $9556



EXPLORATORY DATA ANALYSIS - BIVARIATE: **BMI VS. SEX**







Low BMI – Gender Distribution

- Customers with a BMI less than 18.5
- A higher proportion of females have a low BMI as compared to males

Ideal BMI – Gender Distribution

- BMI between 18.5 and 24.9
- Looking at customers with a BMI that is considered in the ideal range, the proportion is skewed slightly female

High BMI – Gender Distribution

- Customers with a BMI above 24.9
- A higher proportion of males have a higher BMI as compared to females

EXPLORATORY DATA ANALYSIS - BIVARIATE: **BMI VS. SEX**

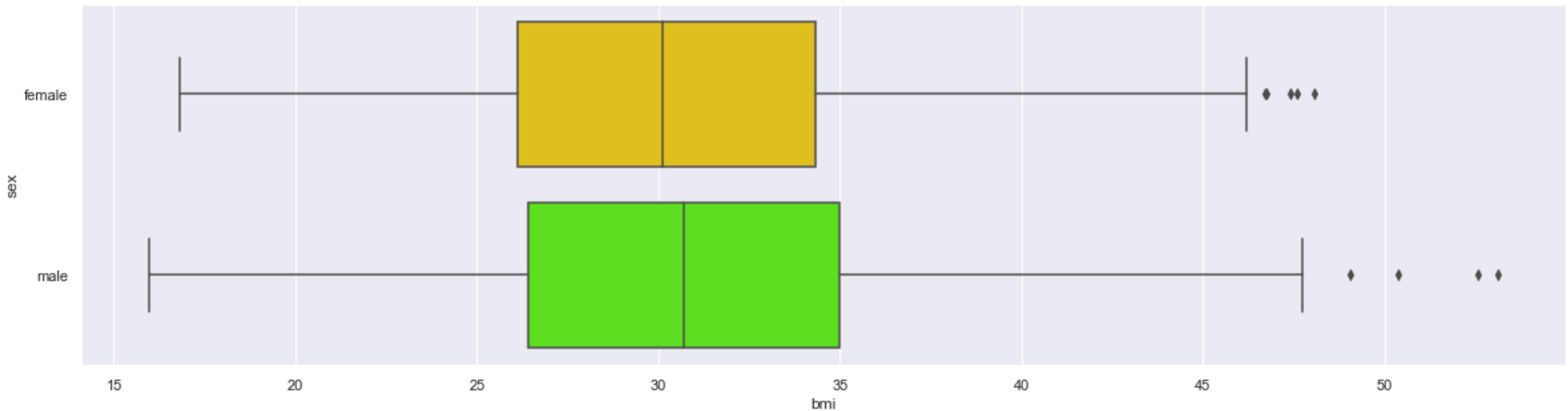
Key Question:

• Is the BMI of females different than the BMI of males?

Statistical Test:

• We tested the above question using a 95% confidence level.

• The test concludes that there is not a significant difference between the BMI of men vs. women.



EXPLORATORY DATA ANALYSIS - BIVARIATE: FEMALE BMI VS. # OF CHILDREN

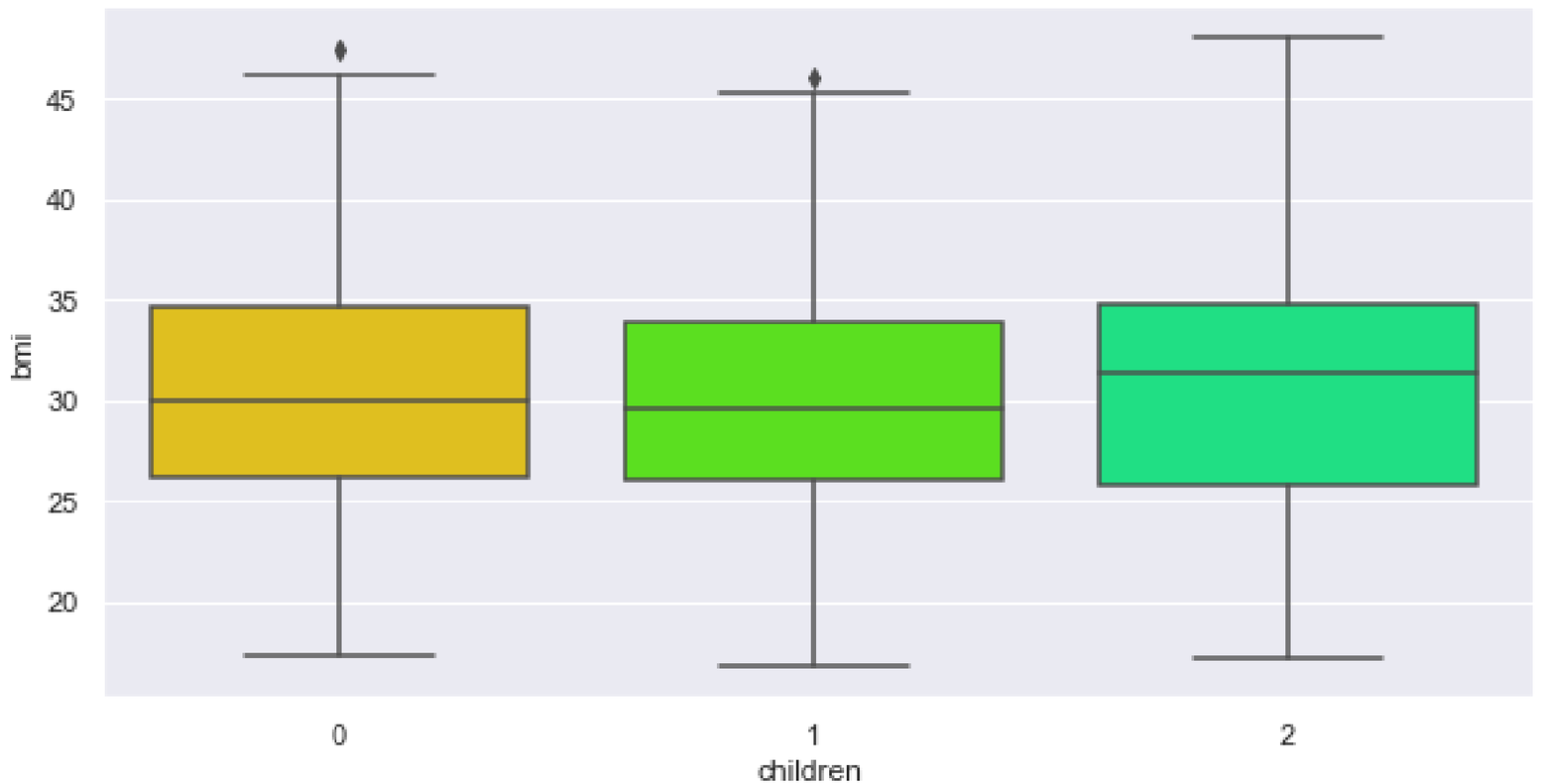
Key Question:

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

Statistical Test:

• We tested the above question using a 95% confidence level.

• The test concludes that there is not a significant difference between the mean BMI of women with no children, one child, and two children.



SUMMARY

1. The age distibution of Axis customers is fairly uniform, except for a larger number of 18 and 19 year olds.

2. An ideal BMI is within the range of 18.5 to 24.9, and the average Axis customer has a BMI around 30.5 – higher that the ideal range. Customers with a high BMI had higher charges on average when compared to those with ideal or low BMIs.

3. 67% of Axis customers have either no children or only 1 child; only 3% of customers have either 4 or 5 children.

4. The middle 50% of charges were between $4,740 and $16,640, but there are higher charges extending up to $63,770.

5. Axis customers are divided nearly equally among two genders, with only slightly more male customers.

6. Smokers account for roughly 1 in 5 (20%) of Axis customers. **Statistical testing concluded that there is a difference between the mean charges made by smokers vs. non-smokers.**

7. Axis customers are fairly evenly divided between the four regions, with the Southeast having a slightly higher percentage of the total customers than the other three regions.

8. The Southeast shows a higher percentage of smokers, but **statistical testing concludes that the proportion of smokers is not significantly different across different regions.**

9. The Southeast also shows signs of having higher BMIs and potentially higher charges.

10. **Statistical testing concludes that there is not a significant difference between the BMI of men vs. women.**

11. **Statistical testing concludes that there is not a significant difference between the mean BMI of women with no children, one child, and two children.**

THANK YOU

STATISTICAL ANALYSIS OF BUSINESS DATA

by **JAKE EIDE**