AXIS INSURANCE

Data Analysis

Objective

- To extract actionable insights from the available data by leveraging customer information .
- Explore and generate value out of company's data.
- Perform uni-variate and multi-variate analysis.
- Help business decisions by answering key questions brought by sales / marketing team.
- Prove with hypothesis test each and every question raised by the team by demonstrating statistical evidences.

Data information

Variables	Description
Age	age of the primary beneficiary, (excluding those above 64 years)
Sex	policy holder's gender
BMI	body mass index (BMI), ideal range between 18.5 and 24.9.
Children	number of children / dependents
Smoker	Y / N if regularly smokes tobacco
Region	beneficiary's place of residence in the U.S
Charges	Individual medical costs billed to health insurance

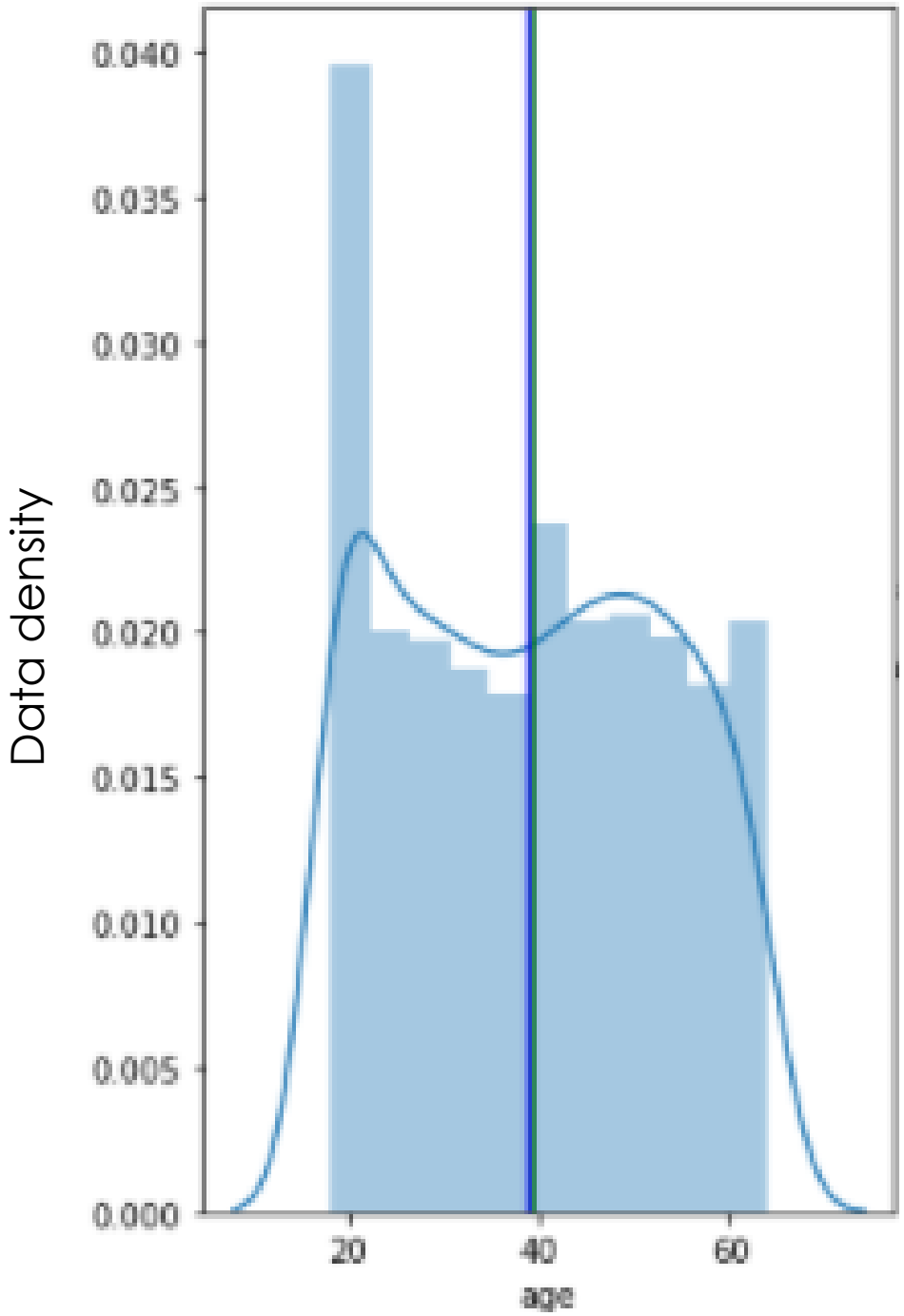
Observations	Variables
9359	7

Notes:

- No missing values were found on the data
- 7 duplicated observations were found on the data

Mean

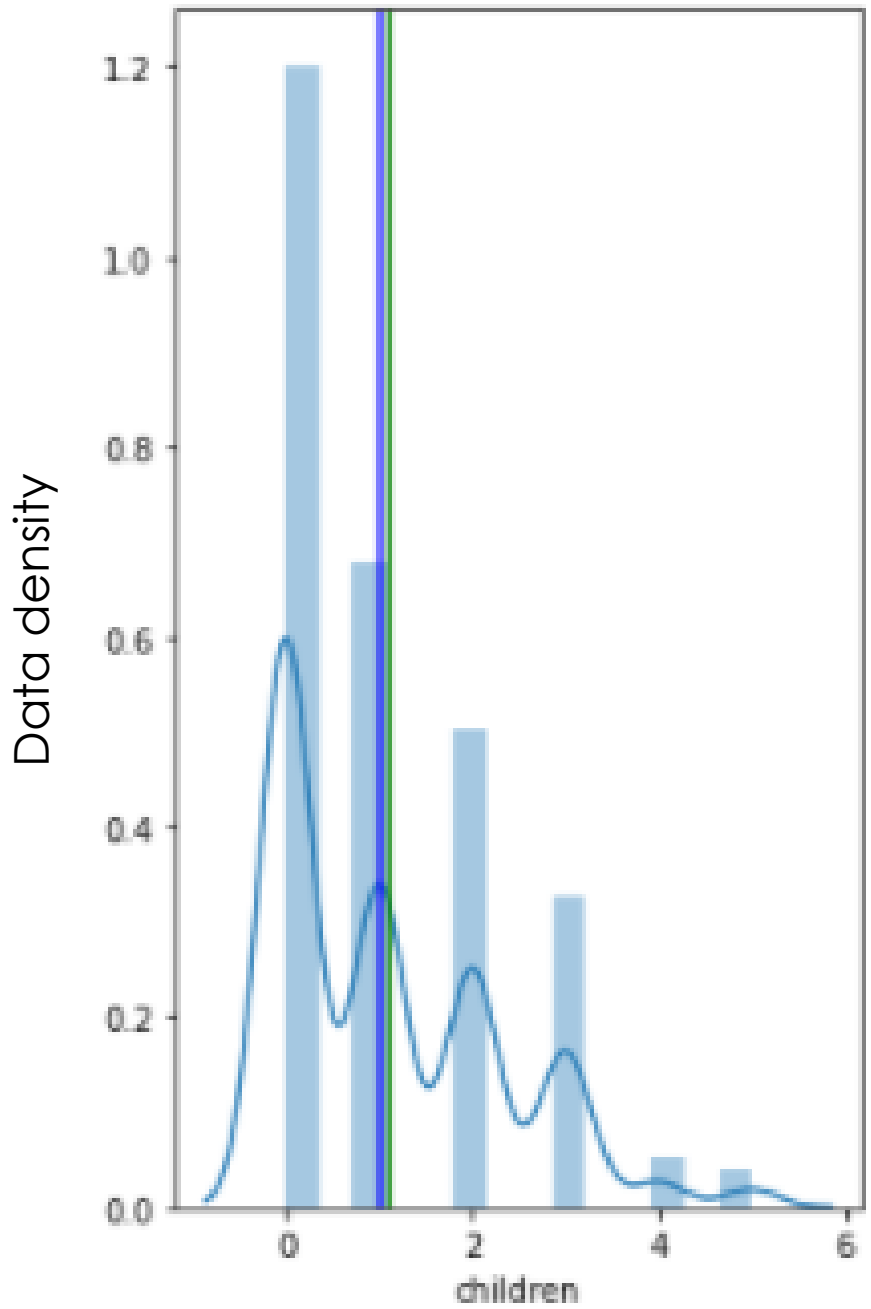
Median



EDA – Univariate Analysis

- Most customers are between 20-60 years old.
- There is a pick of young customers with age around 20y old.

Mean
Median

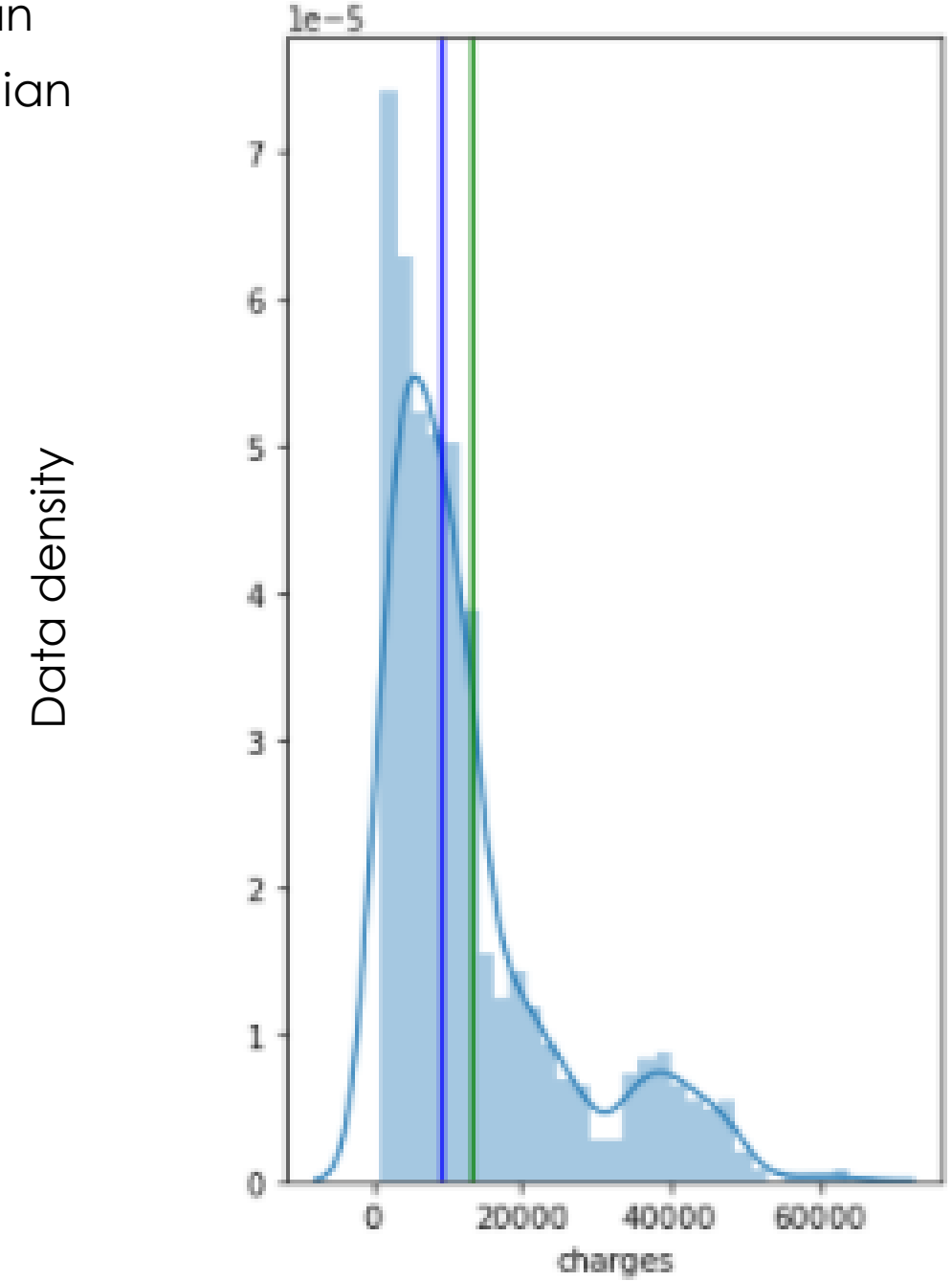


EDA – Univariate Analysis

- Most customers do not have children.
- There is a descent linear trend as the number of dependents gets higher the number of customers reduce.

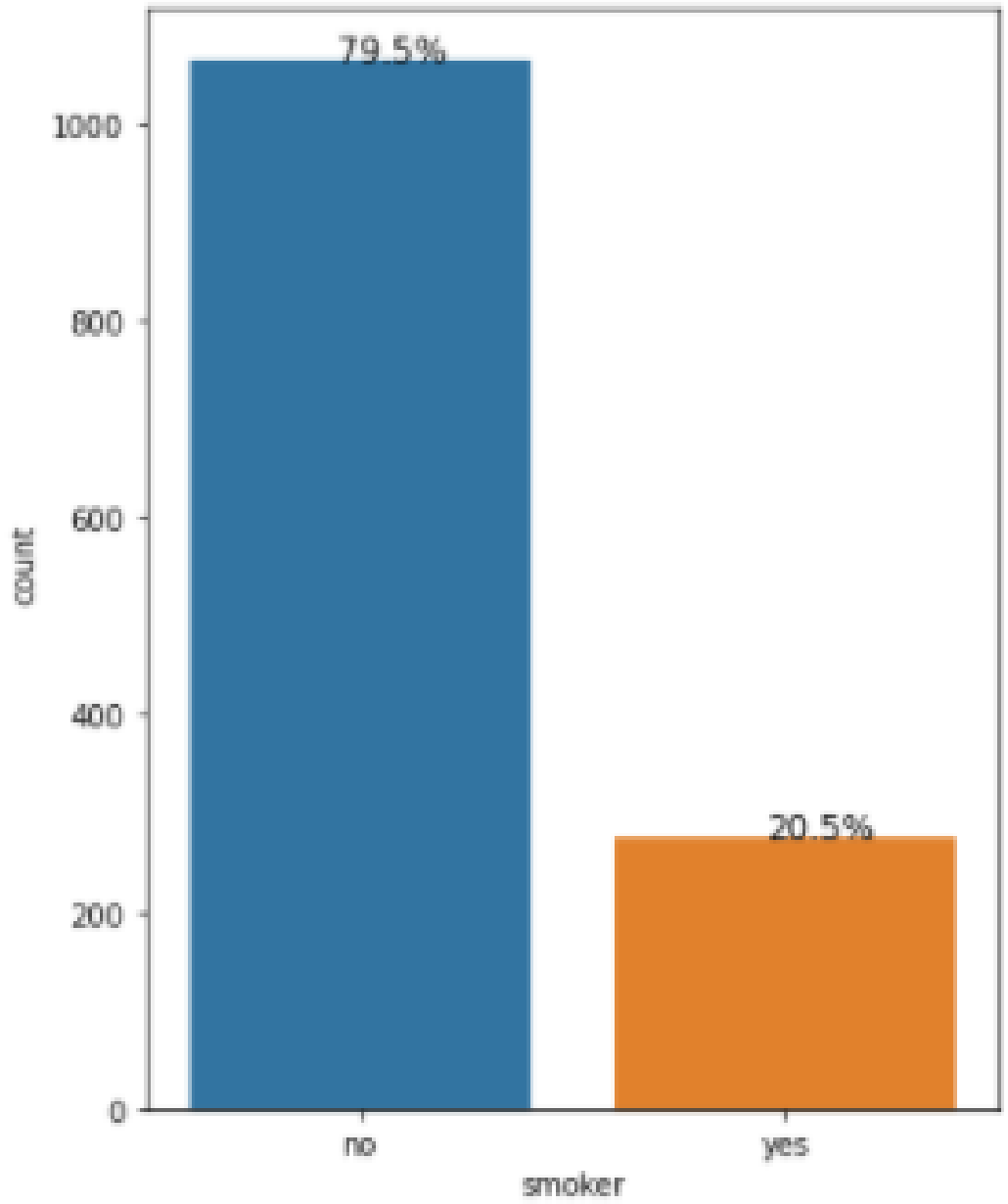
Mean

Median



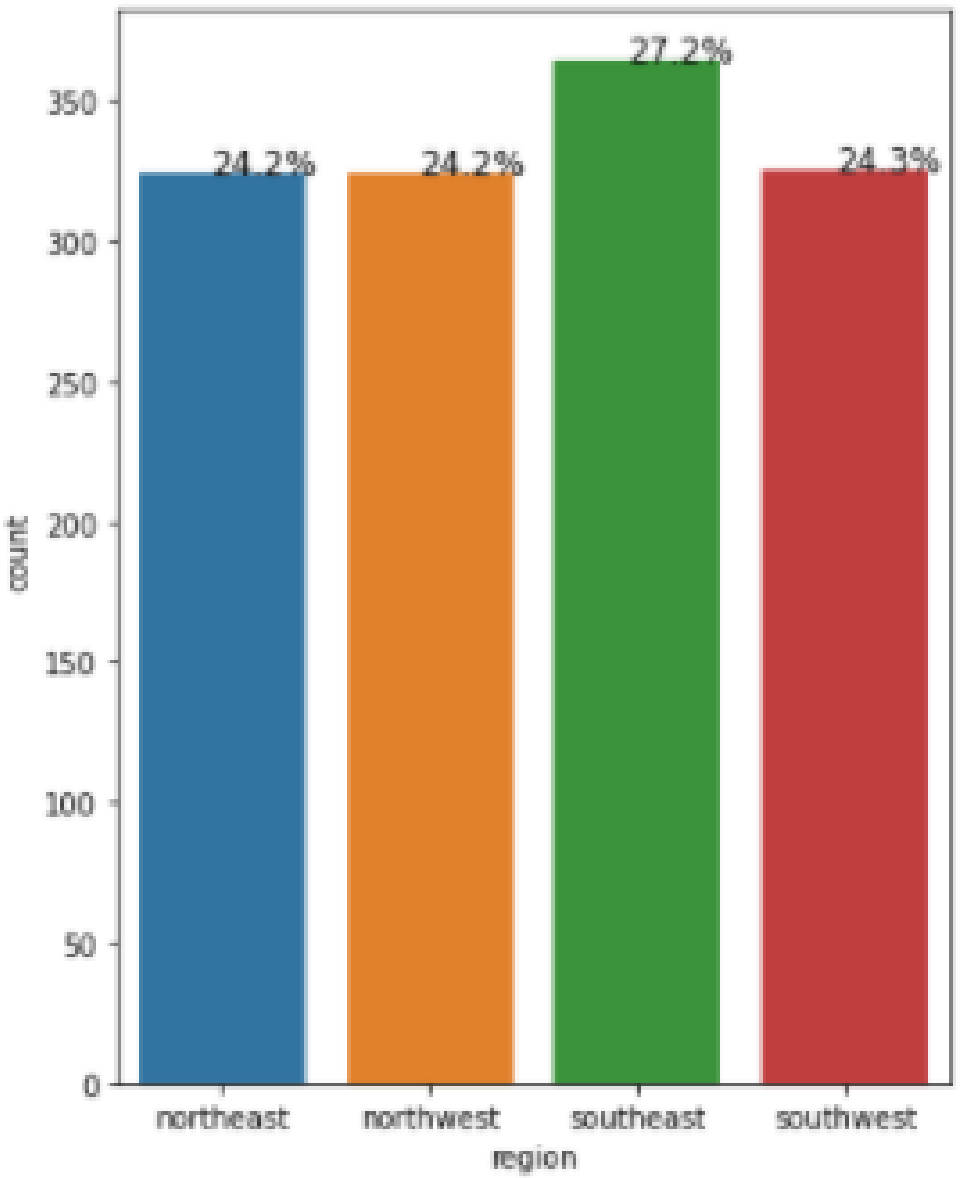
EDA – Univariate Analysis

- Most charges are concentrated between 0 and 20,000 USD.
- There is a right skewed distribution.



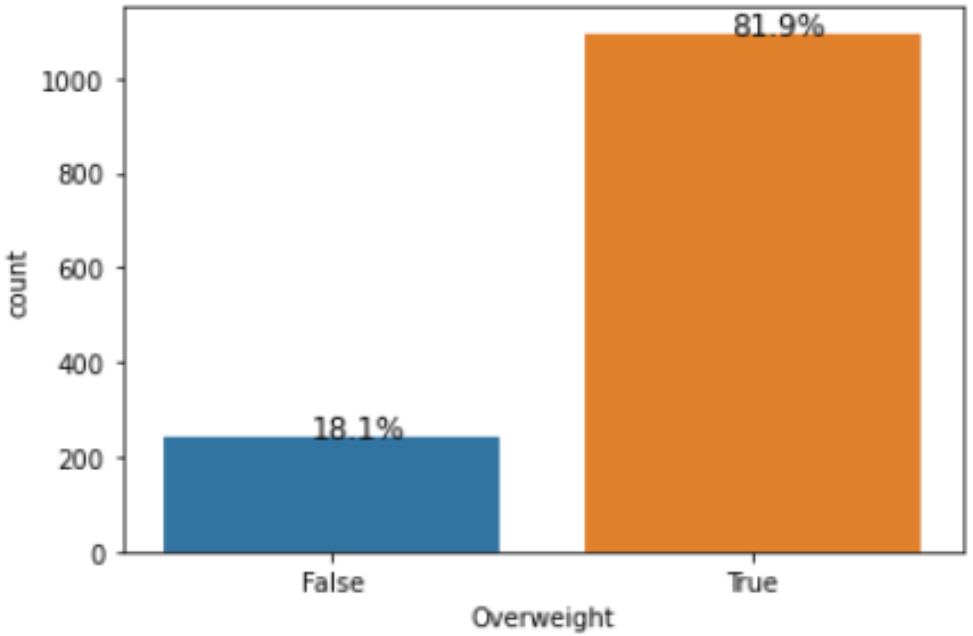
EDA – Univariate Analysis

- precisely 79.5% of policy holders are non-smokers.



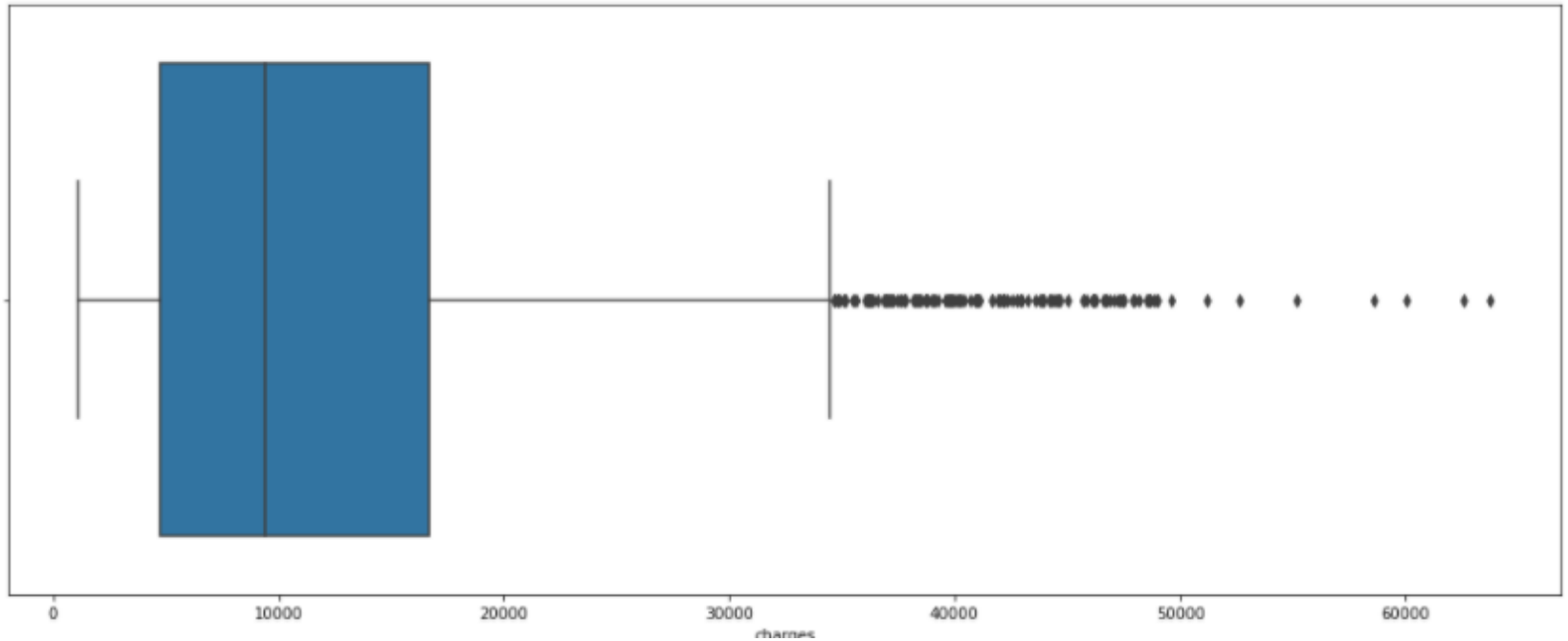
EDA – Univariate Analysis

- The percentage of customers are evenly distributed with only region Southeast standing out with 27.2 %.



EDA – Univariate Analysis

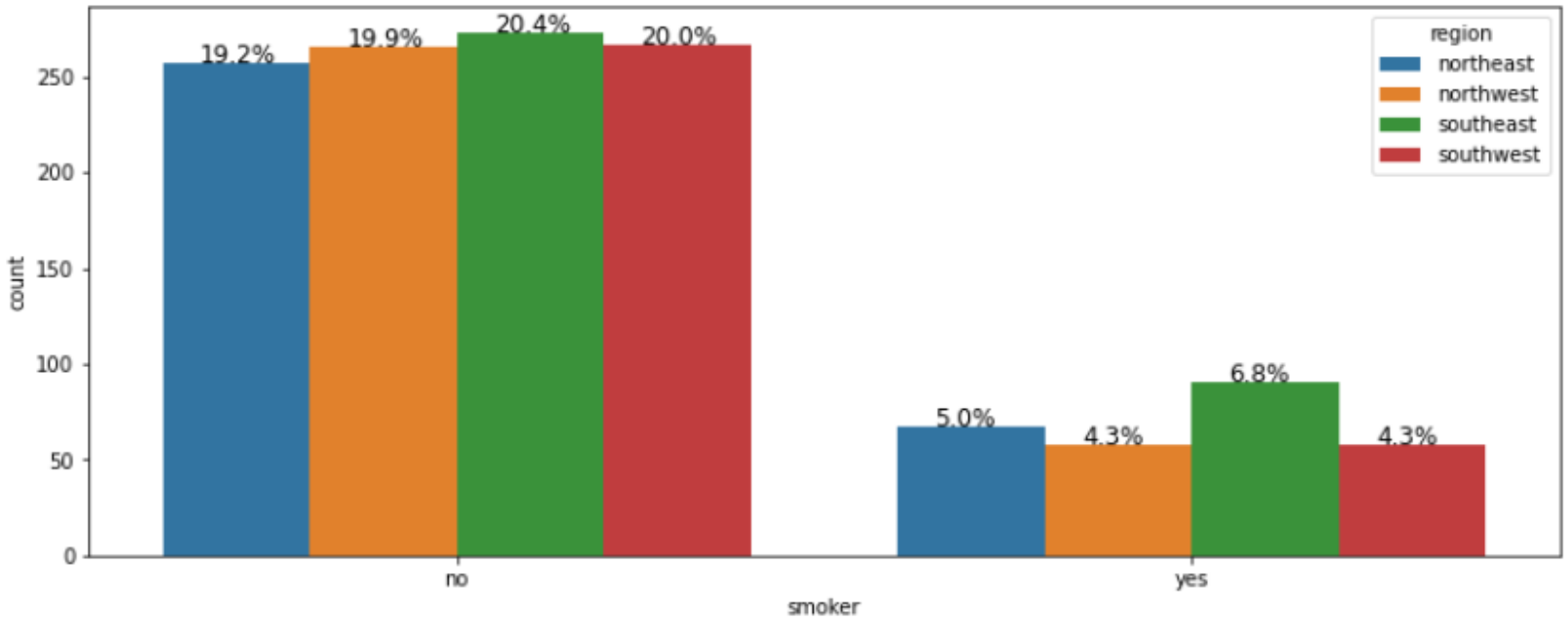
- Base on the fact that over-weight people are the ones who have BMI over 24.9, this chart shows 81.9% of policy holders are over-weight.



EDA - Boxplot

- A big number of outliers are present in the variable “Charges”.

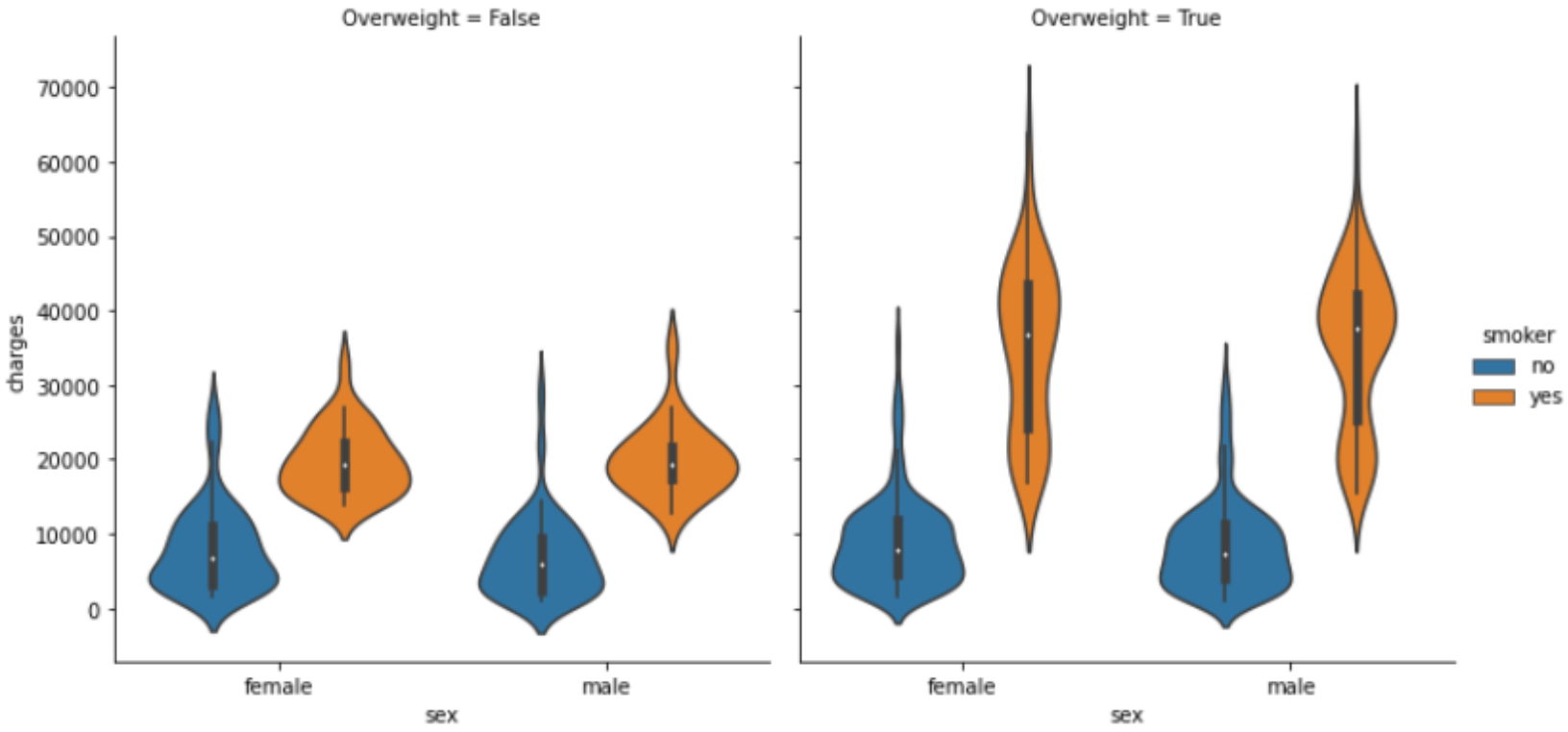
- Further pre-processing task might be needed to deal with outliers.



EDA – Multivariate Analysis

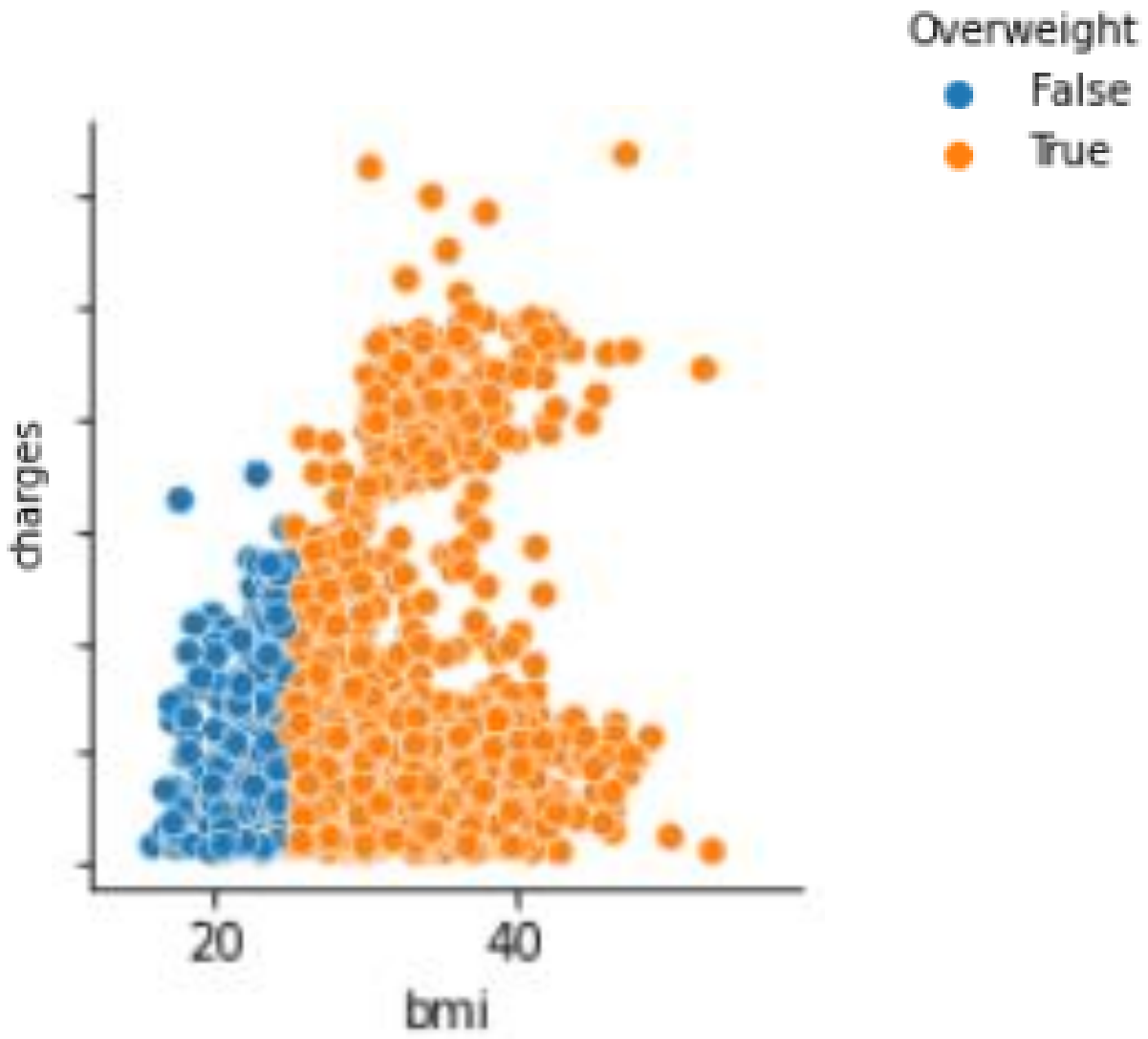
- Non-Smoker are spread out evenly across regions.

-Smokers are in bigger number in Southeast region but this data might be driven by the fact this region has the greatest number of customers.



EDA – Multivariate Analysis

- Smokers have higher bills representing 2X more than non-smokers.
- Average "charges" of a people who is over-weight and smoker is 4X higher in comparison with people in a normal weight and non-smoker.



EDA – Multivariate Analysis

- This is a visual representation of the percentage of customers who is over-weight.
- Over-weight people pays higher bills.

Conclusions

1. The majority of customers are between 20-60 years old with a pick in 20's.
2. 79.5% of policy holders are non-smokers.
3. Southeast region has the greatest number of customers 27.2 %.
4. 81.9% of policy holders are over-weight.
5. Smokers pay 2X more bills than non-smokers.
6. Over-weight customers which are smoker pay 4X more bills than a regular person (non-smoker)

Recommendations

1. 81.9% of our policy holders are overweight. This data makes the company aware about the risk to keep those customers however as they represent almost 4/5 of the total number of the customers a marketing campaign to inform and motivate people to keep their fit might be an idea. Incentives to lower their premium as they reduce their BMI might be one idea. Additional discounts and a marketing strategy to attract healthier people may help to reduce the high percentage of over-weight people.
2. 33% of total of smokers are concentrated in southeast region along with the fact they pay 2X more than non-smoker. A strategy would be lower the sales representation in southeast region and increase in the others. A better understanding about the number of smokers being higher in Southeast due to Southeast has the greatest number of customers should be checked.
3. An interesting fact is that sex "male" showed the one who pay the higher bills in comparison to "female". A marketing strategy to attract women might be something to consider.
4. A recommendation would be to create a sort of score to rate people by the risk they offer to the business and whether adjusting their premium accordingly based on their risk level or try to stay away from such customers. (An example of score based on the insights from this study would look like:

    - in a scale 1-5 (where 1 representing low risk for the business and 5 representing high risk for the business)

    - (1) - Female, Non-Smokers, BMI lower than 24.9

    ....intermediate scores should be defined

    - (5) - Male, Smokers, BMI > 24.9

Questions from the Business

1. Are the medical claims made by the people who smoke greater than those who don't?

Response: With 95% of confidence level this affirmation is correct. The medical claims made by people who smoke is greater than those who don't.

2. Is the BMI of females different from that of males?

Response: With 95% of confidence level BMI of females are the same as Males.

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

Response: With 95% of confidence level the proportion of smokers are equal across regions.

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

Response: With 95% of confidence level the mean BMI of women with no children, one child, and two children are the same.