AXIS INSURANCE PROJECT

2021

OBJECTIVES

- Statistical Analysis of Business Data
- Explore the dataset and extract insights using Exploratory Data Analysis.
- Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't? [Hint- Formulate a hypothesis and prove/disprove it]
- 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? [Hint : Create a contingency table/cross tab, Use the function : stats.chi2_contingency()]
- Is the mean BMI of women with no children, one child, and two children the same? Explain your answer with statistical evidence.
- Given: Level of Significance : 0.05 for all tests.

DATA INFORMATION

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 underweight 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

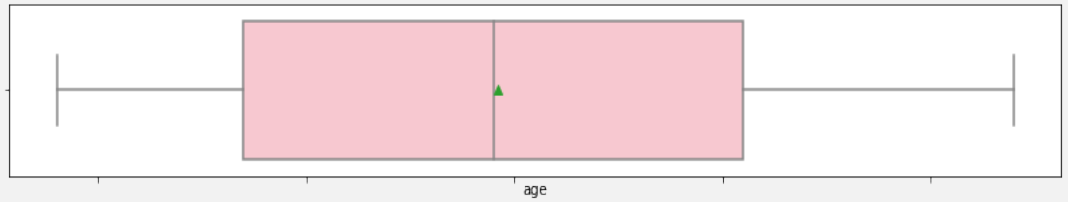
Observations	Variables
1338	7

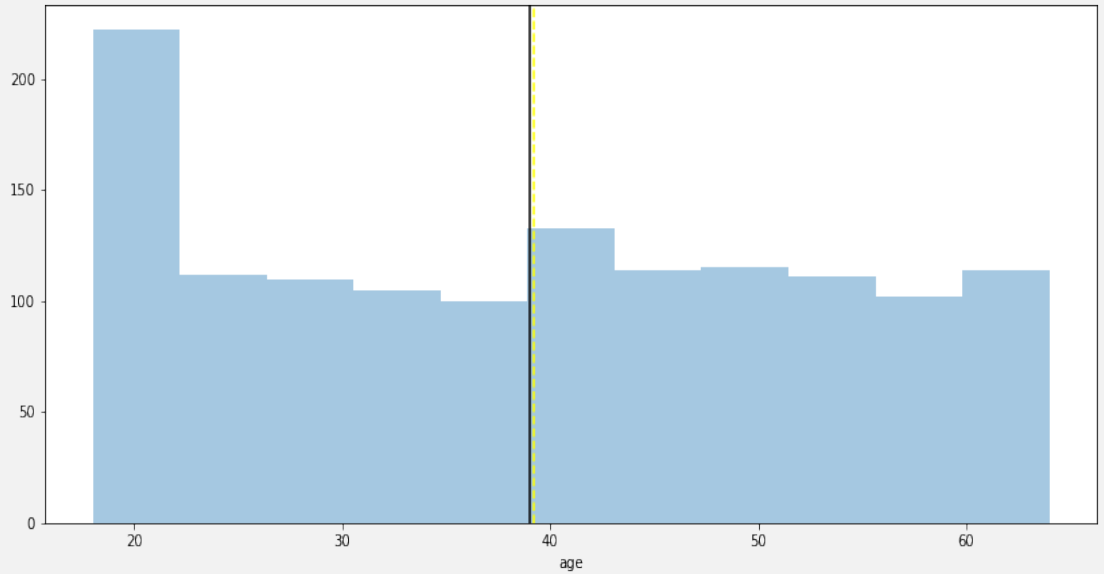
Numerical	Categorical
4 Age BMI Children Charges	3 (converted) Sex Smoker Region

No Missing Values in the Dataset

EXPLORATORY DATA ANALYSIS

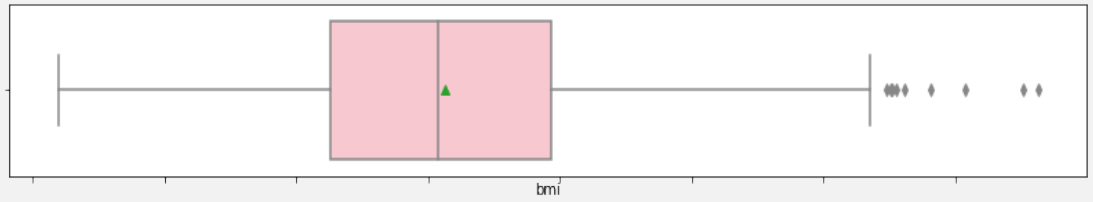
AGE

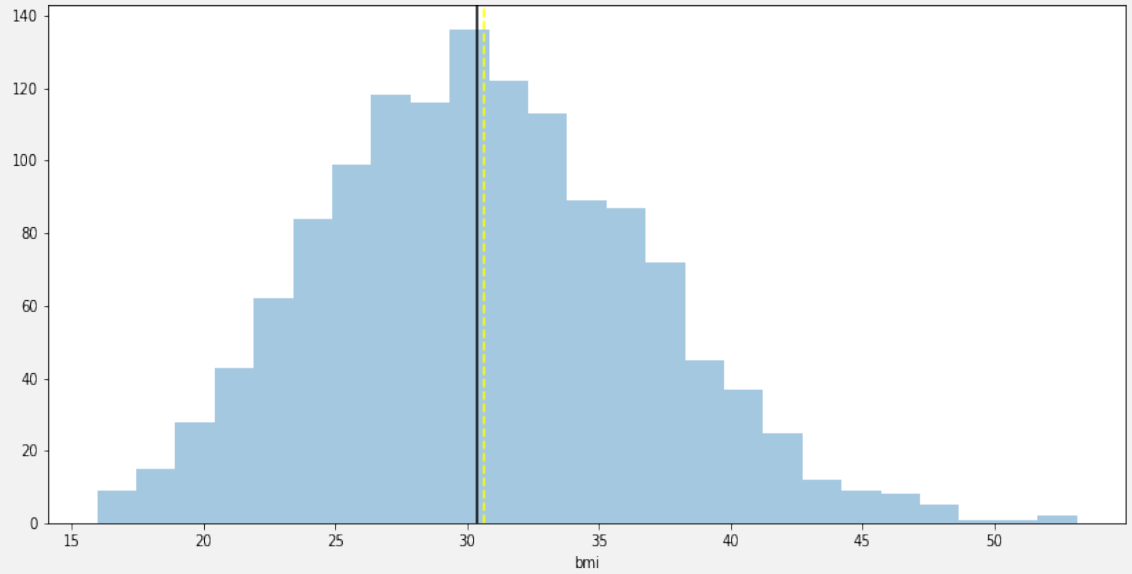




- There are no outliers
- Age is uniformly distributed with Zero Skewness.
- Mean (39.2) and Median (39) are near equal as shown
- 25% of the customers are below 27 years of age while 75% of them are below 51 years of age

BMI

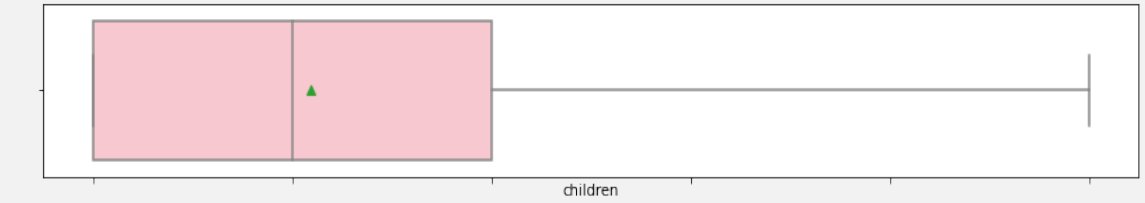


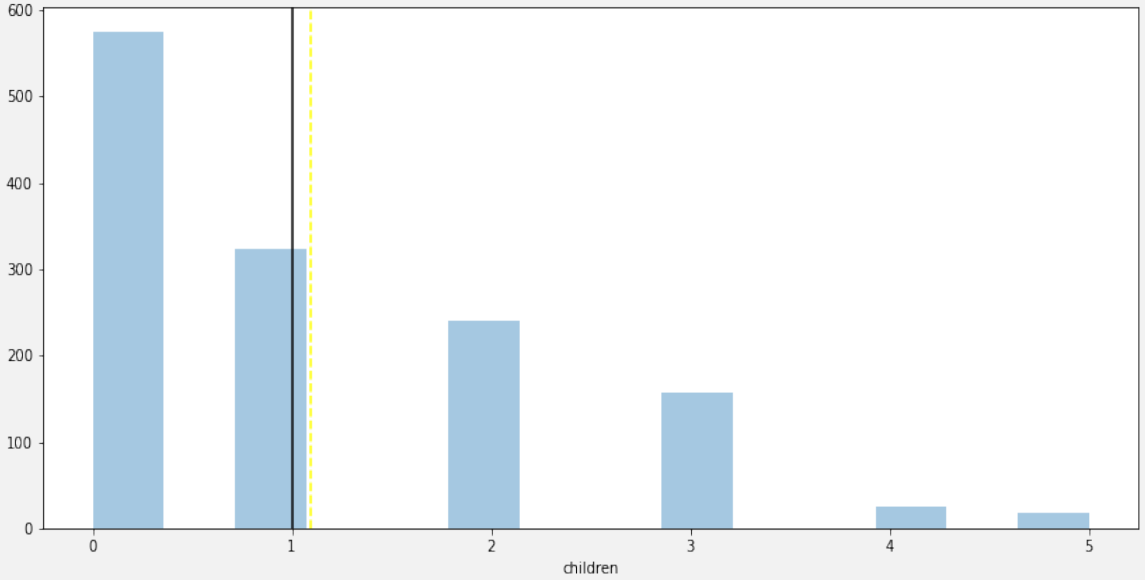


- There are significant outliers which triggered a form of minimal skewness.
- The BMI Index is a normal distribution. Mean (30.6) is approx. equal to Median (30.4)
- We can infer that the insurance company's customer distribution is heavily overweight
- There is a presence of outliers which may indicate variability in data collection

EXPLORATORY DATA ANALYSIS

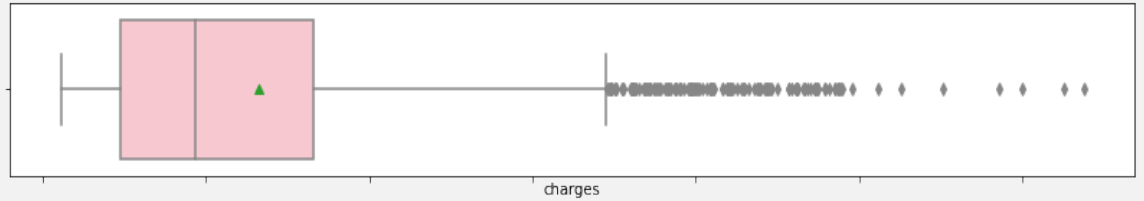
CHILDREN

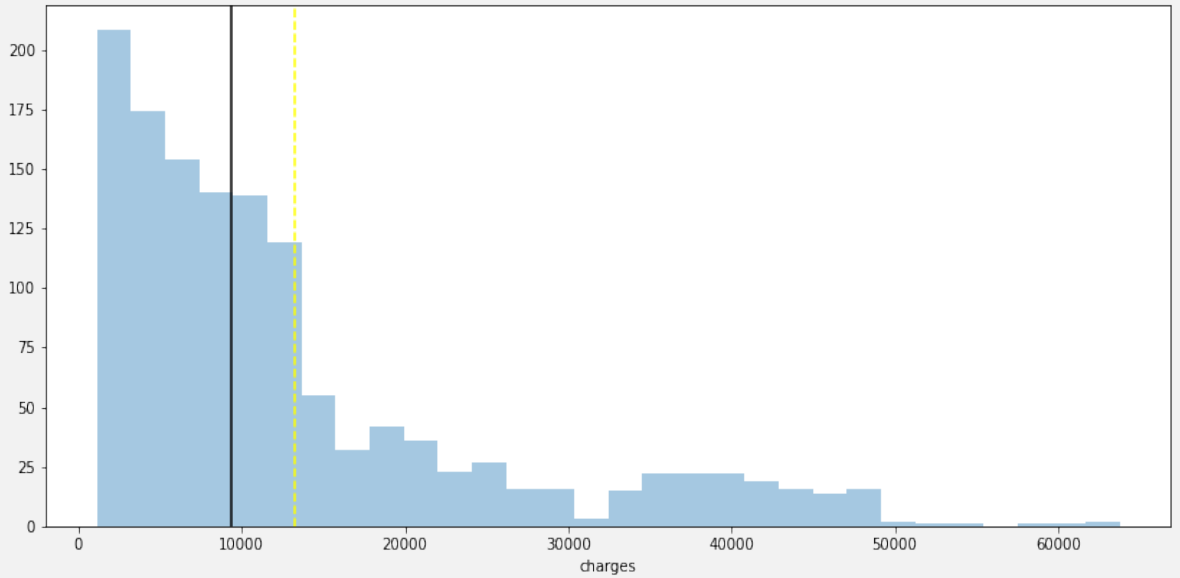




- The distribution is right skewed
- The mean is fairly greater than the median by a little margin.
- 75% of the no. of policy holders have below 2 dependents
- The presence of outliers is obvious

CHARGES

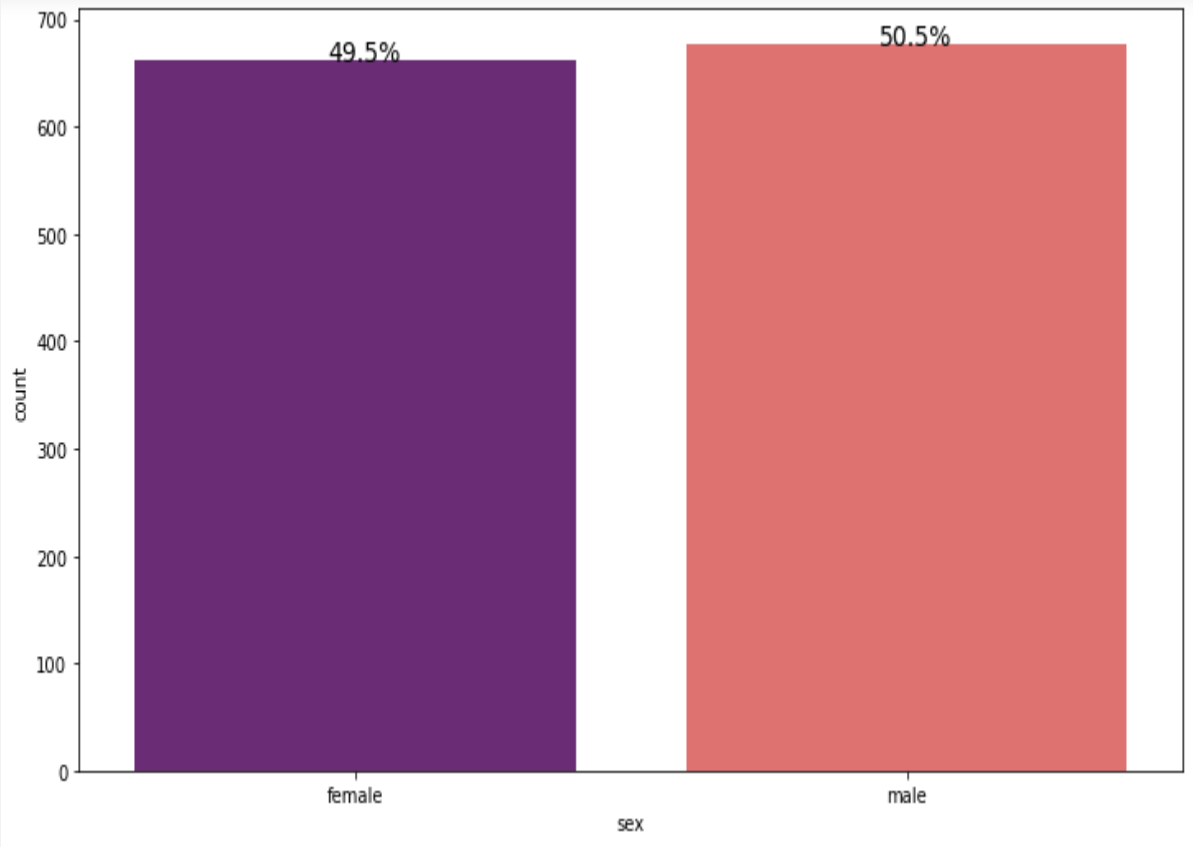




- The distribution of charges/medical claims is right skewed
- There is a presence of a lot of outliers which greatly impacts on the mean position
- 75% of the insurance medical claims is below $16,640
- There is an obvious possibility of error in the data collection owing to the tremendous amount of outliers present

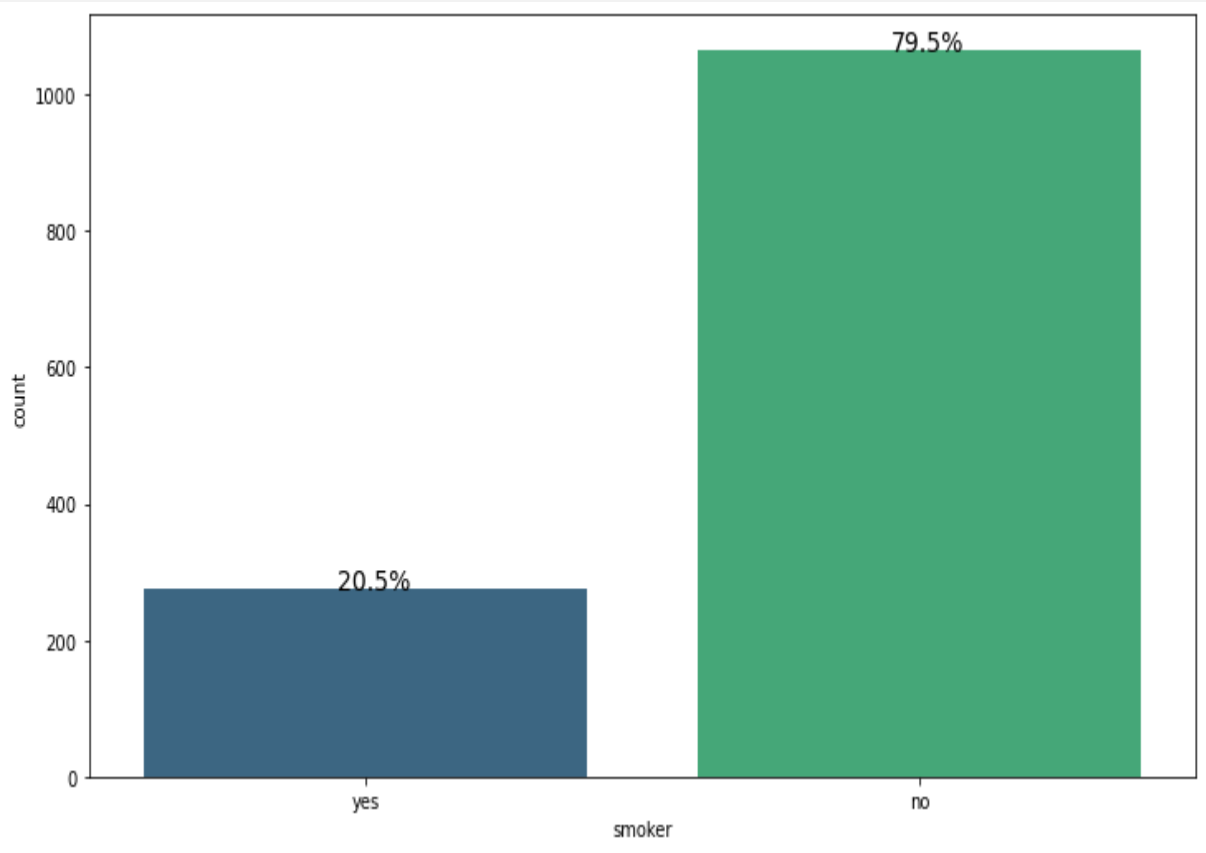
EXPLORATORY DATA ANALYSIS

Sex



The Policy Holders are made up of Male(50.5%) and Female(49.5%)

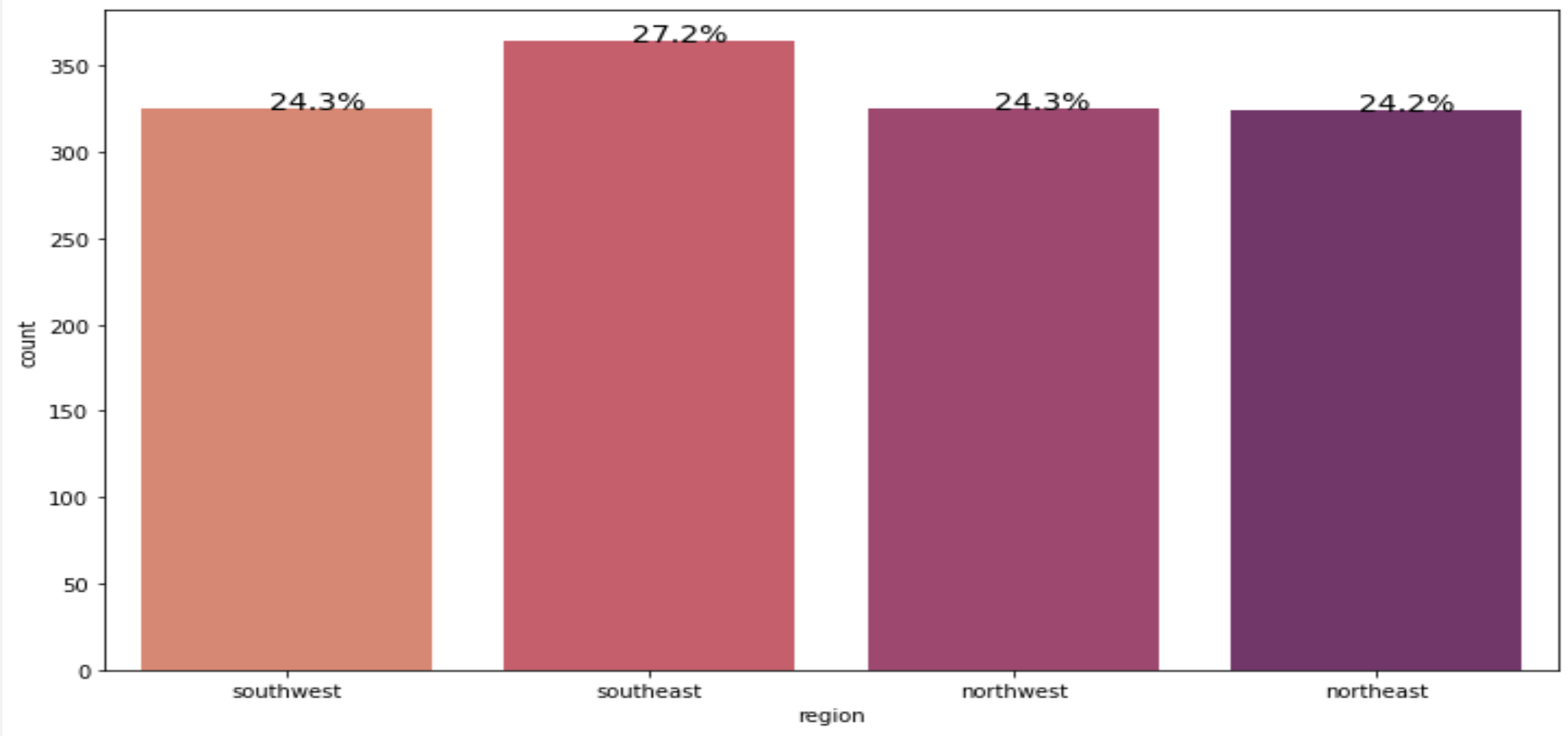
Smoker



The Insurance Policy Holders are made up of 20.5% Smokers and 79.5% Non-smokers.

EXPLORATORY DATA ANALYSIS

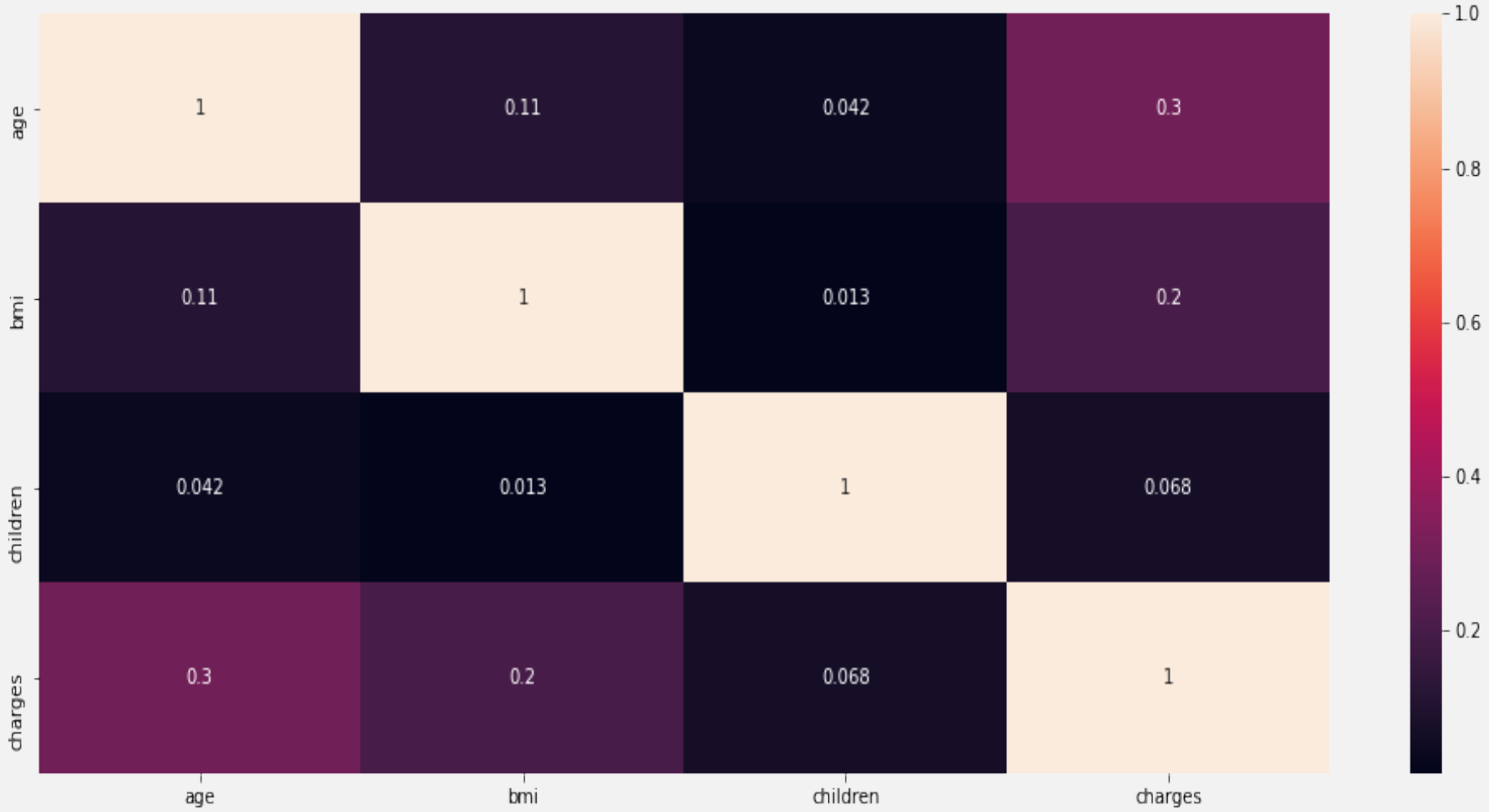
Region



The Southeast has the highest percentage of Policy holders with 27.2%
The Southwest and Northwest pair come second with 24.3% of the policy holders respectively
The Northeast has the lowest percentage of policy holders at 24.2%

EXPLORATORY DATA ANALYSIS

Heat Map



It would appear as though there is no relationship of any sort between the numerical variables.

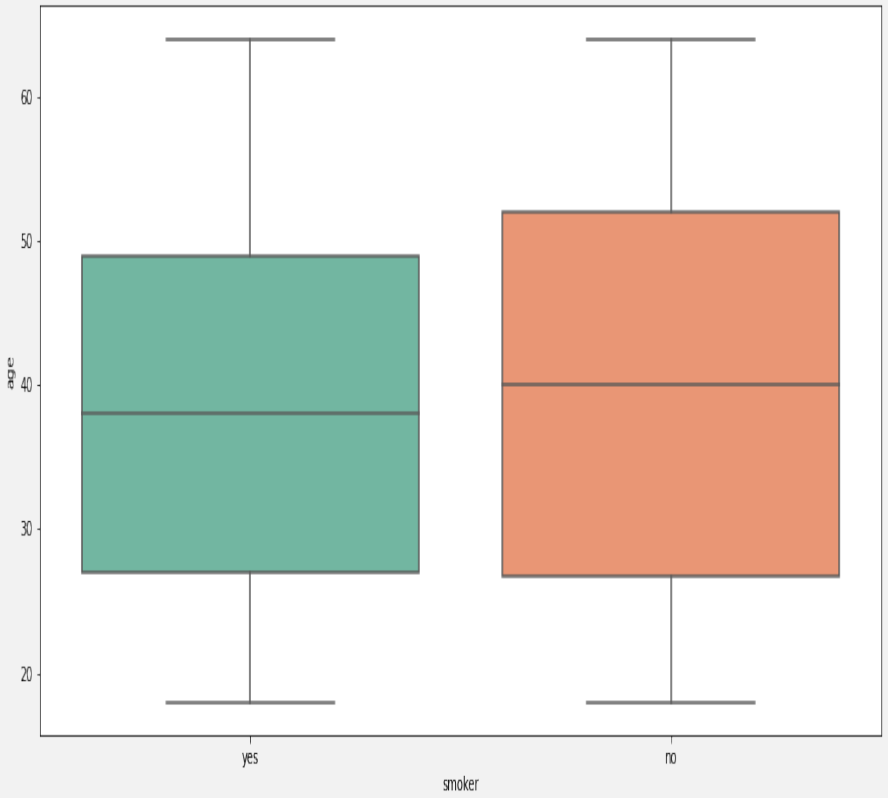
EXPLORATORY DATA ANALYSIS

PAIR PLOT

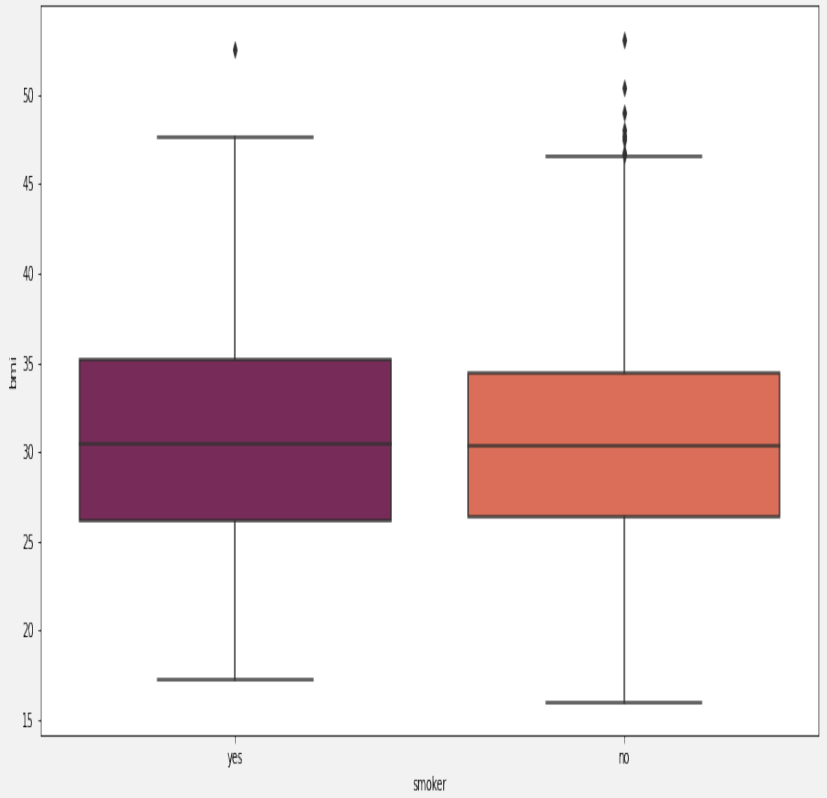
The closest to a correlation is between Age and Charges though faintly marginal

EXPLORATORY DATA ANALYSIS

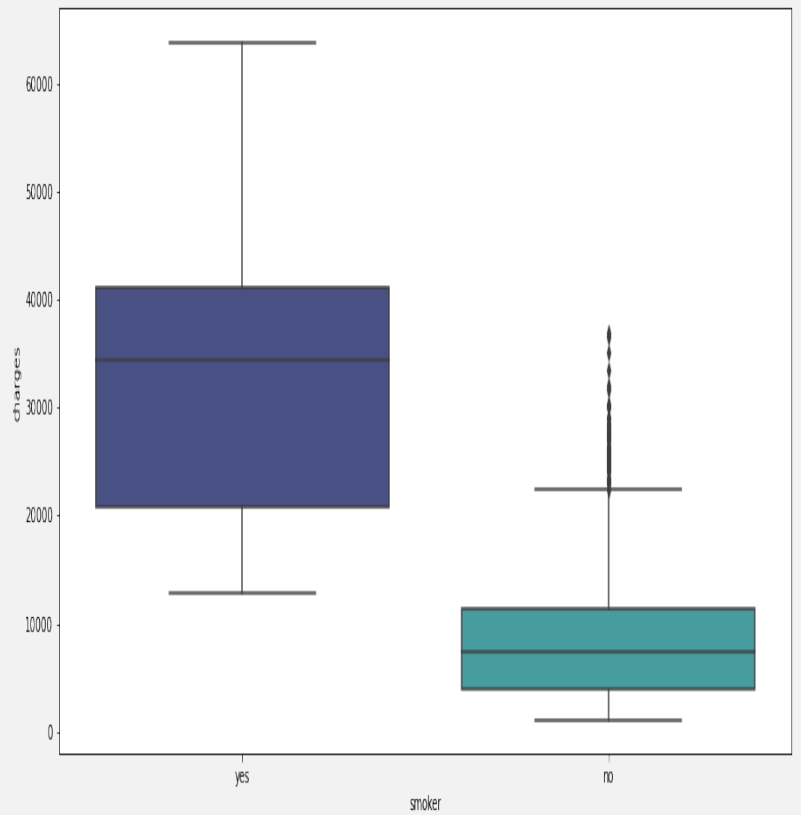
SMOKERS w.r.t AGE - BMI - CHARGES



- There are more smokers between ages 27-49yrs
- There are more non-smokers in the population
- There are no outliers across the smokers and non-smokers



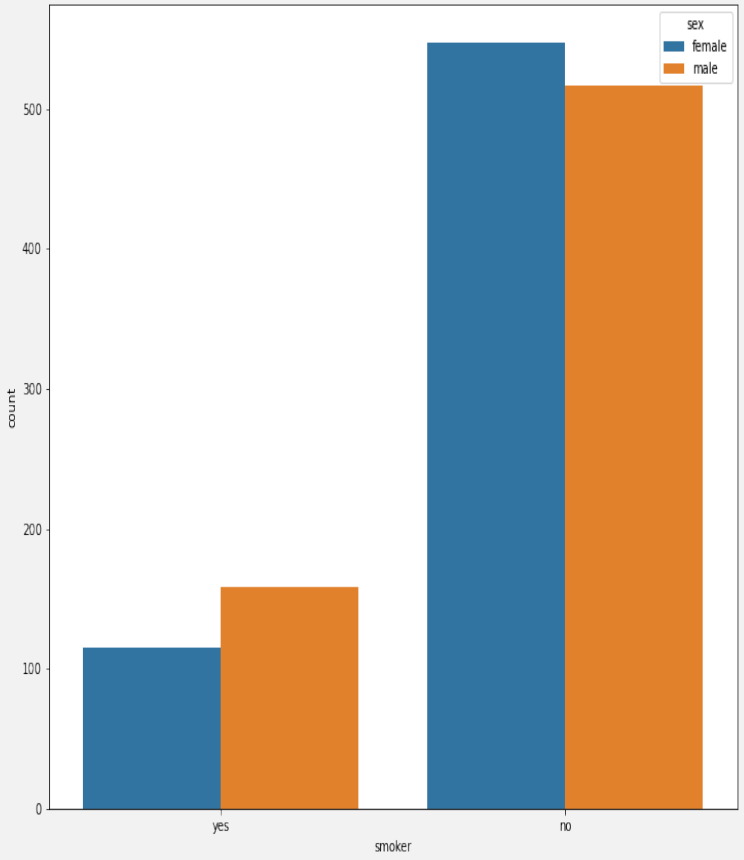
- Presence of outliers across both groups
- The smokers have a higher BMIs
- Median BMI for both groups is equal to 30



- Smokers pay higher charges than non smokers
- Mean/Median charge of 35,000 for smokers
- as against 5,000 for non-smokers
- Presence of significant outliers

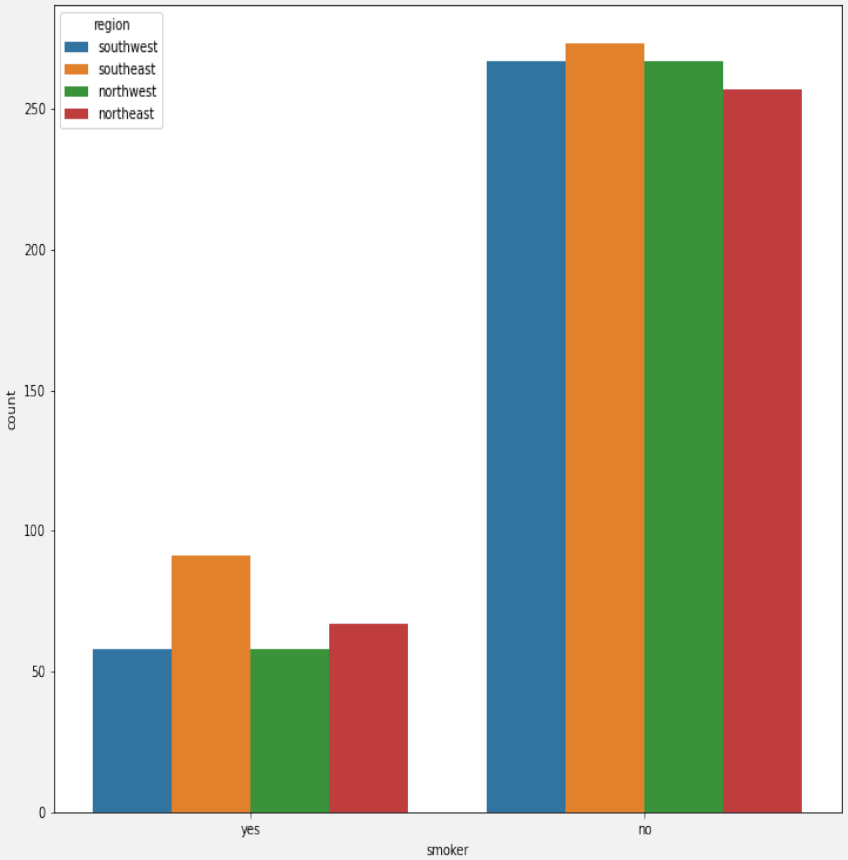
EXPLORATORY DATA ANALYSIS

SMOKER VS SEX



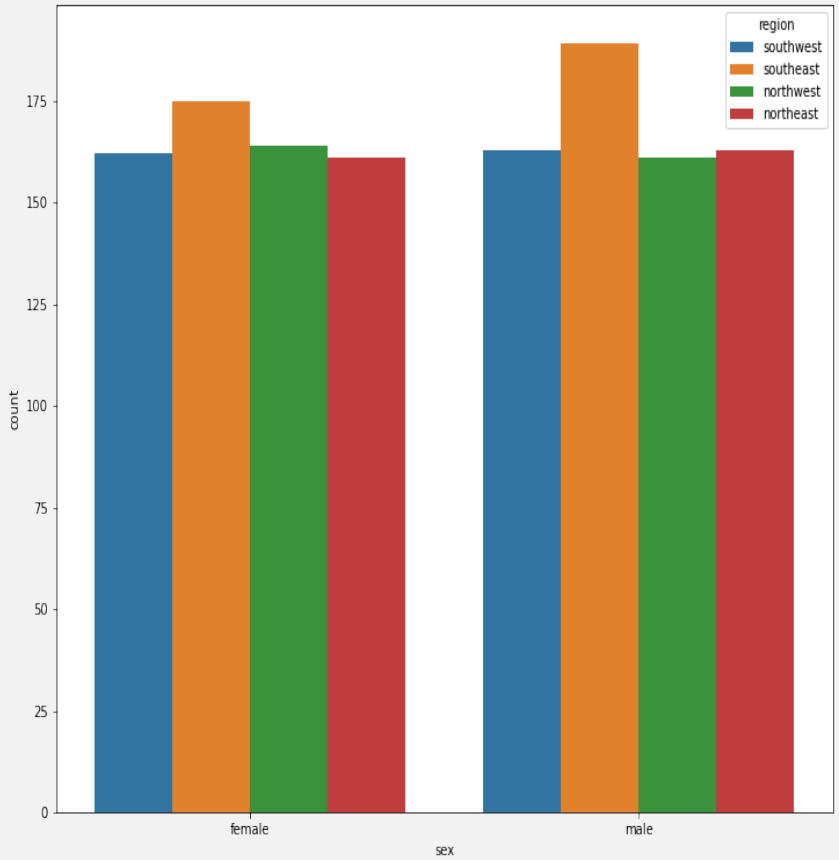
- Female non-smokers (NS) > female smokers (S).
- Male non-smokers (NS) > male smokers

SMOKER VS REGION



- SE has the most smokers/non-smokers
- NE is next but with the least non-smokers
- NW and SW have the least smokers

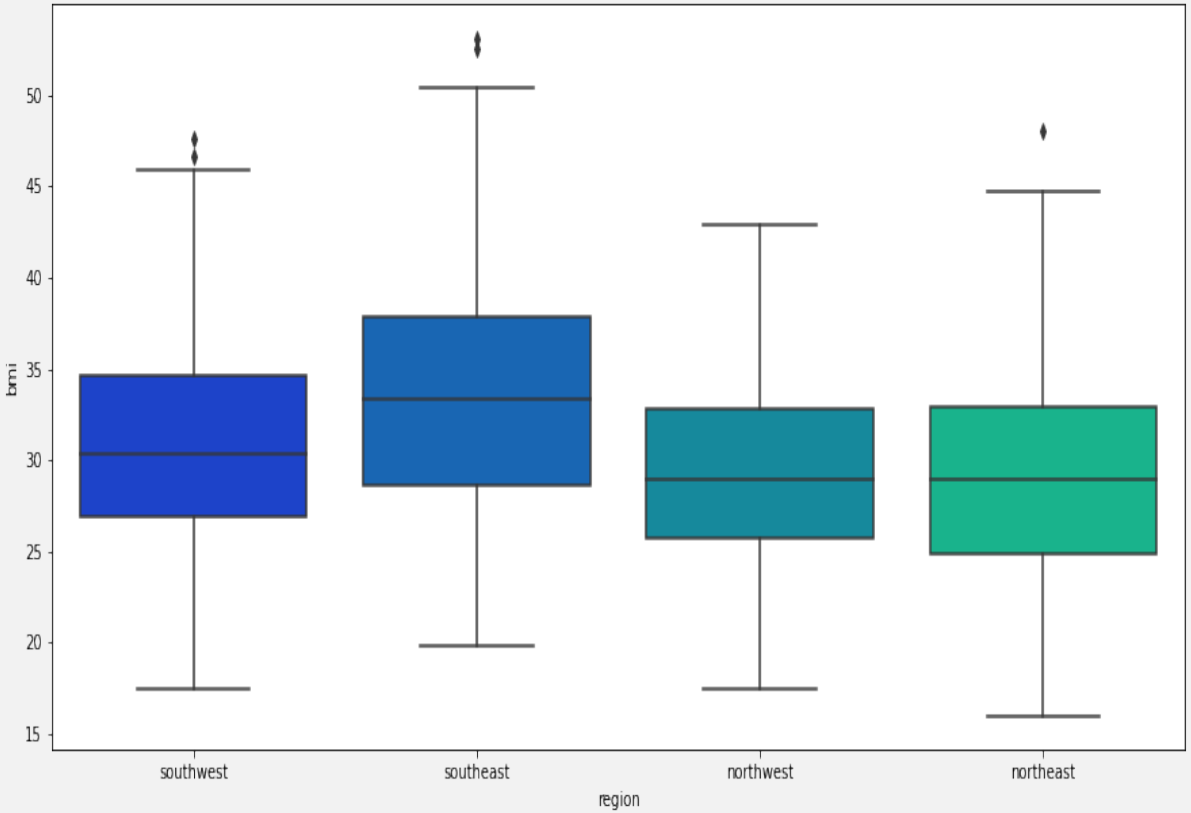
SEX VS REGION



- The SE has the most policy holders (Male/Female)
- Males in the NE > males in NW and SW respectively
- Females in NE < females in NW and SW respectively

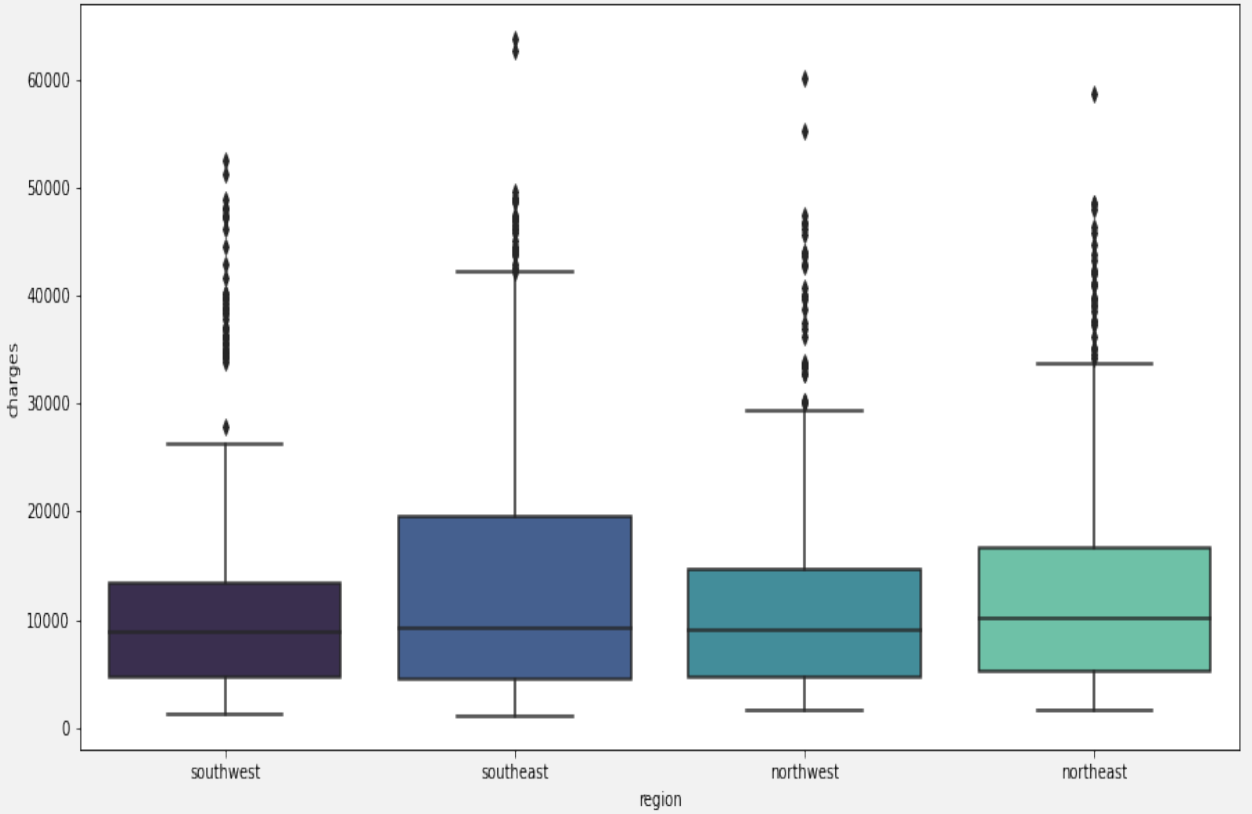
EXPLORATORY DATA ANALYSIS

REGION VS BMI



- The SE has the most obese policy holders across the regions.
- SW comes next and then the NE
- The NW has the least obese in the population
- Note also the presence of outliers in this regions

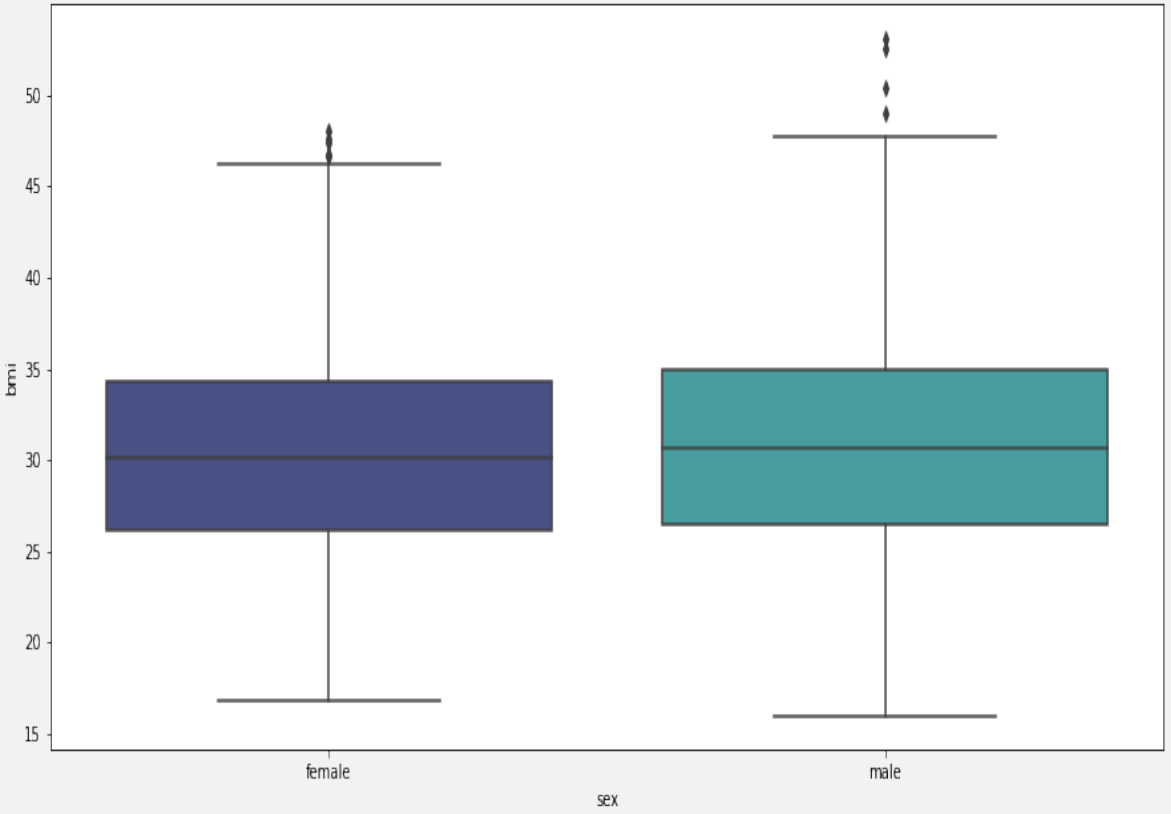
REGION VS CHARGES



- SE made the highest medical insurance claim
- NE is next but with the least non-smokers
- NW is the next in terms of claims and SW is the least

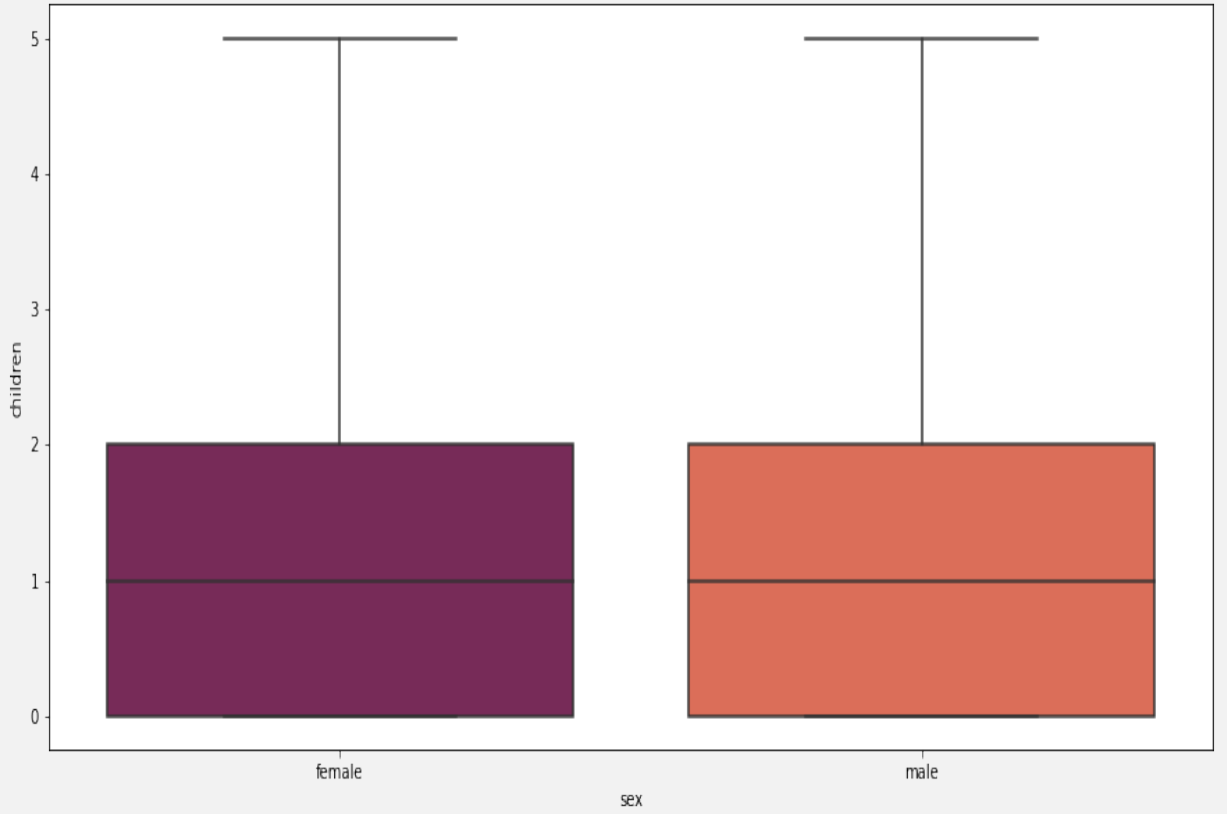
EXPLORATORY DATA ANALYSIS

SEX VS BMI



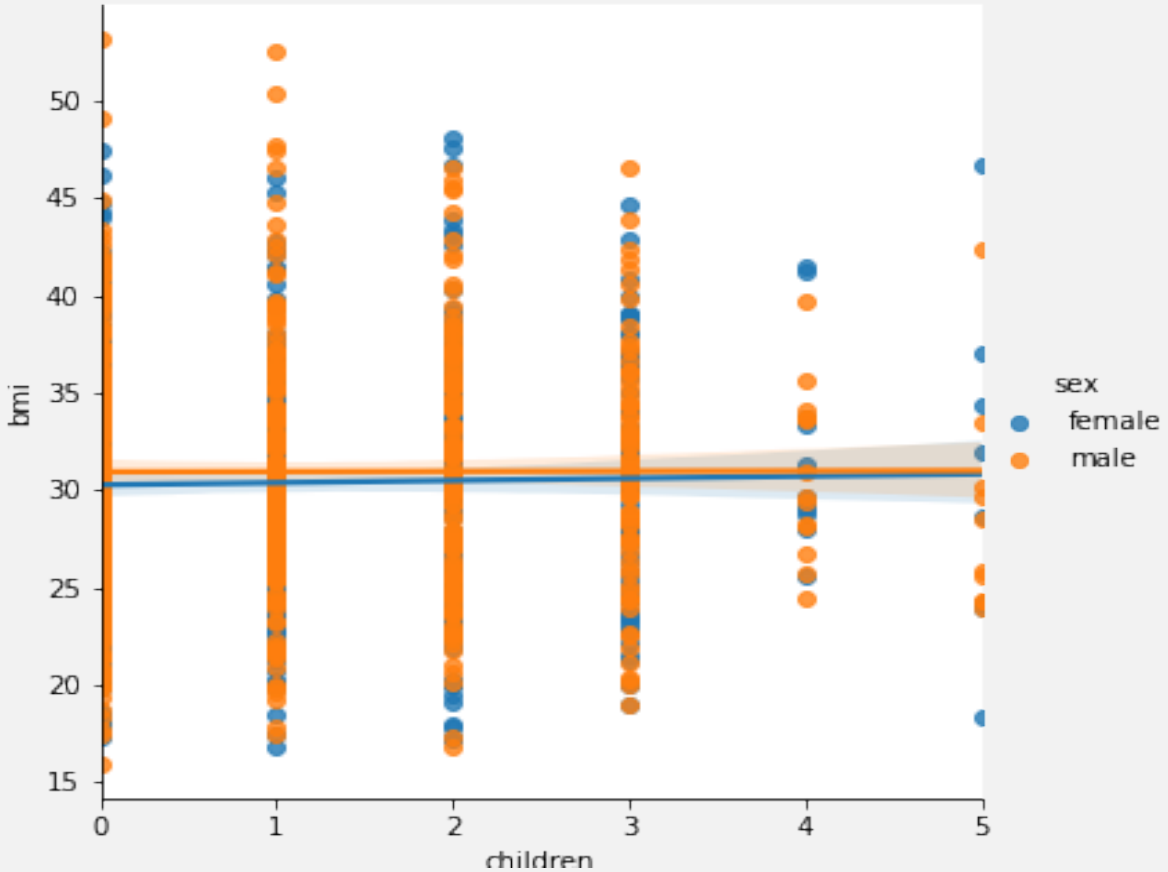
- The males are more overweight compared to the females
- There are presence of outliers

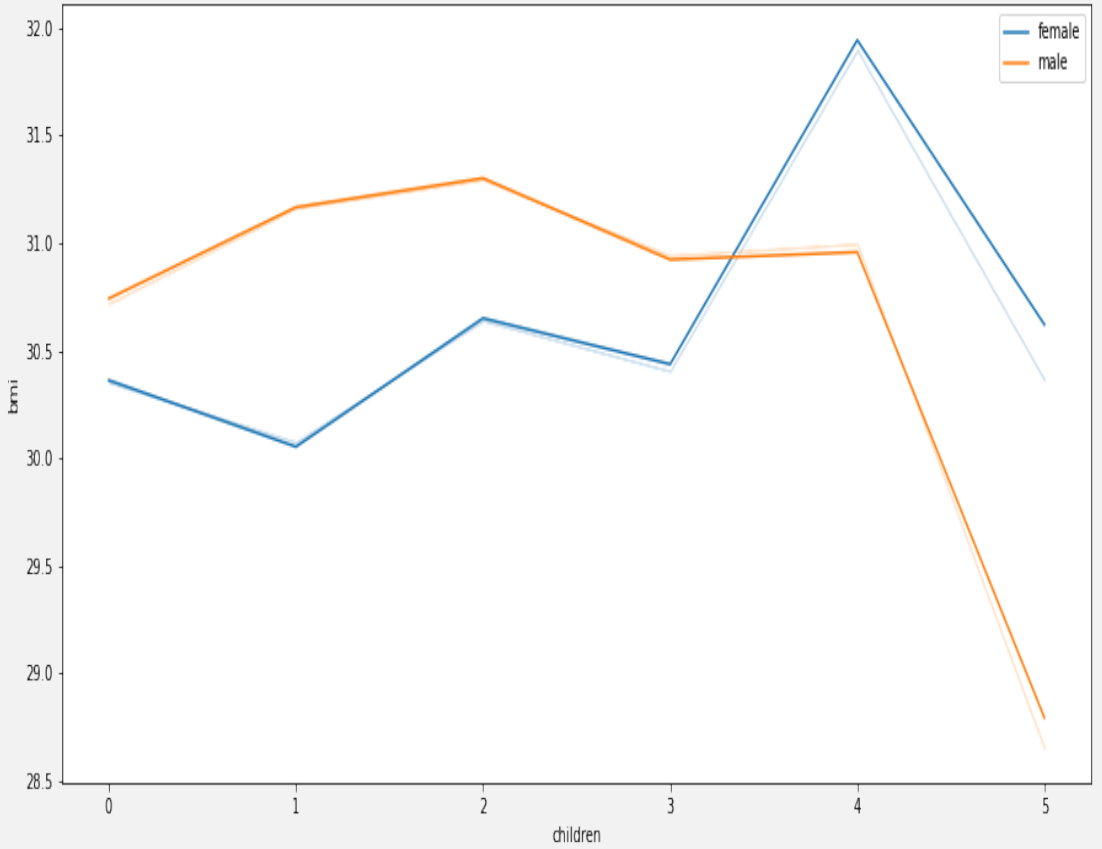
SEX VS DEPENDANTS



Both Male and female policy holders seem to have predominantly 2 or less dependants respectively

EXPLORATORY DATA ANALYSIS





Both plots validate the obvious that a significant proportion of the population are the most independent. This holds true for data points at zero but decreases over the distribution

HYPOTHESIS TESTING- INSIGHTS

1. To Prove/Disprove that the medical claims by people who smoke are greater than those who don't.

   TEST : One Tailed T-test for Two Independent Samples with un-equal standard deviations

   RESULT : p-value < 0.05(Level of Significance), Ho was rejected , Ha was accepted. Statement holds True

2. To Prove/Disprove that the BMI of females is different from that of males

   TEST : This is a Two-tailed T-test for Two Independent Samples with equal standard deviations

   RESULT : p-value > 0.05(Level of Significance), Ho was accepted , Ha was rejected. Statement not valid

3. To validate if the proportion of smokers is significantly different across the regions

   TEST : This is a Chi Square Test of Independence

   RESULT : p-value > 0.05(Level of Significance), Ho was accepted , Ha was rejected. Statement not valid

4. To establish if the mean BMI of women with no children, one child, and two children the same

   TEST : This is the ANOVA Test for equality of means for more than Two(2) samples

   RESULT : p-value > 0.05(Level of Significance), Ho was accepted , Ha was rejected. Statement holds True

CONCLUSIONS-KEY INSIGHTS

POLICY HOLDER PROFILES BASED ON SEX

MALES

- There are 676 males with a mean age of approx. 39yrs with the SouthEast region having the most males with 189
- Most Male Policy holders are between the ages of 26 and 51 or mid-20s and early 50s
- Most of them have their BMIs reading between 26kg to 35kg
- Majority of the males have at most 2 dependants (0-2) per policy
- There are 517 non-smoking males
- The medical insurance claims posted by most males ranges predominantly from 4600 and 19,000 approx.

FEMALES

- There are 662 females with a mean age of approx. 40yrs with the SouthEast region having the highest population of females with 175
- Most female Policy holders are between the ages of 27 and 52 or late 20s and early 50s
- Most of them have their BMIs reading between 26kg to 34kg. the females are grossly overweight
- Majority of the females have at most 2 dependants (0-2)per policy
- There are 547 non-smoking females
- The medical insurance claims posted by most females is predominantly from roughly 4900 and 14,500

CONCLUSIONS-KEY INSIGHTS

POLICY HOLDER PROFILES BASED ON SMOKERS

SMOKERS

- The males score the highest smokers across the distribution at 159 with a mean age of approx. 39yrs and the Southeast region having the most smoking males with 91
- Most smoking male policy holders are between the ages of 27 and 49 or late 20s and late 40s
- Most of them have their BMIs reading between 26kg to 35kg
- Majority of them have at most 2 dependents (0-2) per policy
- There are 274 smoking policy holders in all
- The medical insurance claims posted by most smoking males ranges predominantly between 21,000 and 41,000 approx.

NON-SMOKERS

- The females score the highest non-smokers across the distribution at 547 with a mean age of approx. 39yrs and the Southeast region having the most non-smoking females with 273
- Most non-smoking female policy holders are between the ages of 27 and 52 or late 20s and early 50s
- Most of them have their BMIs reading between 26kg to 34kg
- Majority of them have at most 2 dependents (0-2) per policy
- There are 1064 non-smoking policy holders in all
- The medical insurance claims posted by most non-smoking females ranges predominantly between 4,000 and 11,000 approx.

CONCLUSIONS-KEY INSIGHTS

POLICY HOLDER PROFILES BASED ON REGIONS

SOUTHEAST

- There is a total of 364 policy holders living in the South East region,273 of which are smokers in all with 189 male smokers and a mean age of approx. 39yrs
- Age range of 27 and 51 or late 20s and early 50s
- Most of them have their BMIs reading between 29kg to 38kg. This is an over weighted distribution
- The medical insurance claims posted by most non-smoking males ranges predominantly between 4,400 and 20,000 approx.

NORTHEAST

- There is a total of 324 policy holders living in the Northeast region,257 of which are non-smokers in all with 163 male non-smokers and a mean age of approx. 39yrs
- Age range of 27 and 51 or late 20s and early 50s
- Most of them have their BMIs reading between 25kg to 33kg.
- The medical insurance claims posted by most non-smoking males ranges predominantly between 5,100 and 17,000 approx.

CONCLUSIONS-KEY INSIGHTS

POLICY HOLDER PROFILES BASED ON REGIONS

SOUTHWEST

- There is a total of 325 policy holders living in the Southwest region,267 of which are non-smokers in all with 163 males and a mean age of approx. 39yrs
- Age range of 27 and 51 or late 20s and early 50s
- Most of them have their BMIs reading between 27 to 35
- The medical insurance claims posted by most non-smoking males ranges predominantly between 4,800 and 13,500 approx.

NORTHWEST

- There is a total of 325 policy holders living in the Northwest region,267 of which are non-smokers in all with 164 females and a mean age of approx. 39yrs
- Age range of 26 and 51 or late 20s and early 30s
- Most of them have their BMIs reading between 26 to 33
- The medical insurance claims posted by most non-smoking females ranges predominantly between 4,800 and 15,000 approx.

BUSINESS RECOMMENDATIONS

Based on the key insights regarding the policy holder profiles generated, the following are ideal recommendations to the board;

- With 79.5% as Non-smokers compared to 20.5% smokers, there is an obvious untapped opportunity regarding the non-smoking population. New product developments or alternatives such as Life Insurance policies, Education policies, Vehicle insurance or contingent liability insurance could form the basis of a formidable marketing vis-à-vis advert campaigns to gain a competitive edge, more market share and ultimately increased revenue
- Given that the ideal BMI is between 18.5 to 24.9, from the analysis this is clearly an over weighted population . Besides premium payments by policy holders, the board could look at diversifying into nutritional or weight loss startups opportunities to grow the group's bottom line
- With the smoking population accounting for a chunk of the medical insurance claims, it is inevitable to channel resources to other less costly product initiatives that ensures payment of more premiums by policy holders as compared to heavy claims
- Incentives such as number of dependents per policy holder can be reduced to discourage smokers from subscribing to a policy
- As evidenced in the analysis, an extensive an aggressing marketing initiative ought to be considered in the Northwest and Southwest regions to take advantage of a more healthy population with less risks and thus less claims through new product offerings target and incentives
- The fittest by BMI as well as the youngest age band of the policy holders are from the Northwest and Southwest; as such it is pertinent to develop products along the lines of Educational loans Insurance for certification programs or other credit risk insurance products to enable them take advantage of new opportunities