AXIS INSURANCE BUSINESS STATISTICS

Project 2

Authored by C. Kamakani Kahunahana

DATASET OVERVIEW

The dataset contains information about the company's products, customers, and usage patterns.

Variable	Type	Description
Age	Integer	Age of the primary beneficiary
Sex	Category	Male or Female
BMI	Float	Body Mass Index
Children	Integer	Number of children/dependents covered
Smoker	Category	Yes or No
Region	Category	Northwest, Northeast, Southwest, Southeast
Charges	Float	Self-rated score (5-very fit, 1-very unfit)

Details	Amount
Observations	1338
Variables	7
Null	0

DATASET OVERVIEW
KEY STATISTICS

NUMERICAL VARIABLES

	Age	BMI	Children	Charges
Count	1338	1338	1338	1338
Mean	39.21	30.66	1.09	13270.42
Standard Deviation	14.05	6.10	1.21	12110.01
Minimum	18.00	15.96	0.00	1121.87
25%	27.00	26.30	0.00	4740.29
50%	39.00	30.40	1.00	9382.03
75%	51.00	34.69	2.00	16639.91
Maximum	64.00	53.13	5.00	63770.43

- Age, BMI, and Children have approximately the same Mean and Median
- Mean is higher than the median for Charges suggesting a left skew in the dataset.
- The max of $63,770 seems to indicate a long tail toward the upper end with some customer having high medical charges

CATEGORICAL VARIABLES

	Sex	Smoker	Region
Count	1338	1338	1338
Unique	2	2	4
Top	male	no	southeast
Frequency	676	1064	364

- Males (676) make up the majority of participants in the study
- The majority of customers are non-smokers (1064)
- The southeast is the most common region with 364 customers

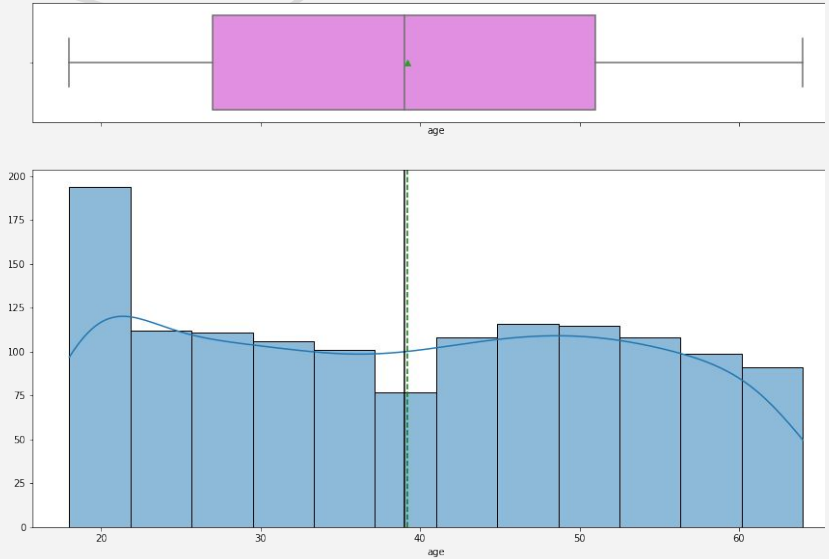
01

BUSINESS OVERVIEW

EXPLORATORY DATA ANALYSIS

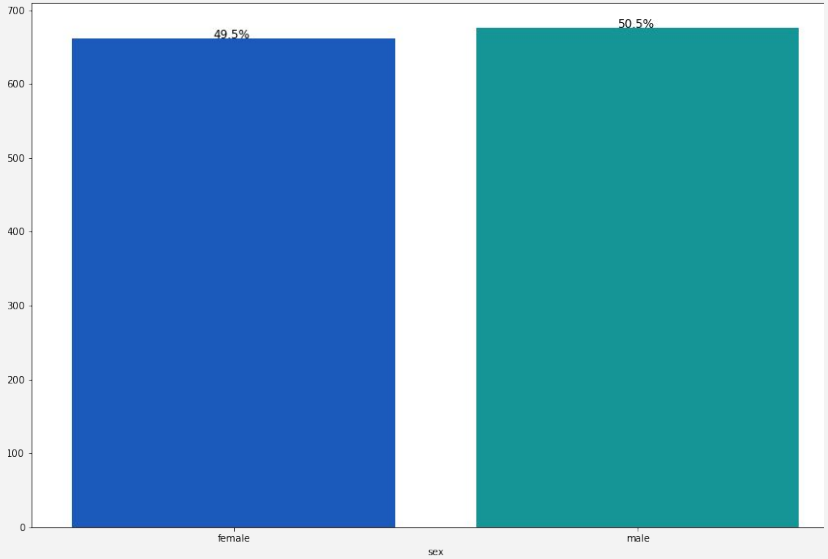
CUSTOMER DEMOGRAPHICS
AGE & SEX

CUSTOMER AGE



- The distributed of customer ages is fairly even with a slight left skew
- There does not appear to be outliers in this data
- Mean and median are approximately equal

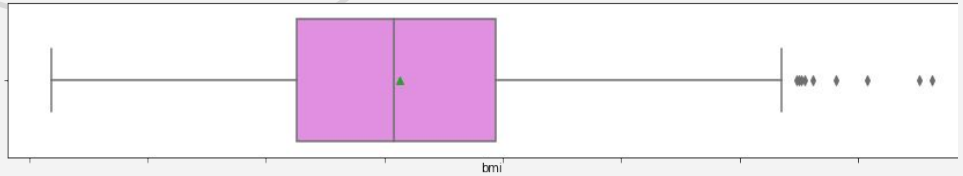
CUSTOMER SEX

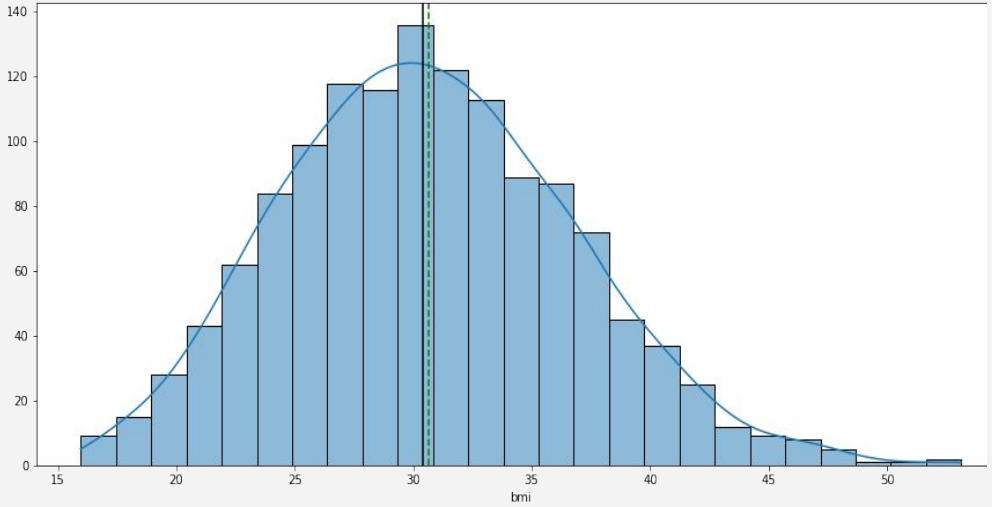


- The sex of customers in the dataset are almost evenly split between Male (50.5%) and Female (40.5%).

CUSTOMER DEMOGRAPHICS
BODY MASS INDEX

CUSTOMER BODY MASS INDEX (BMI)





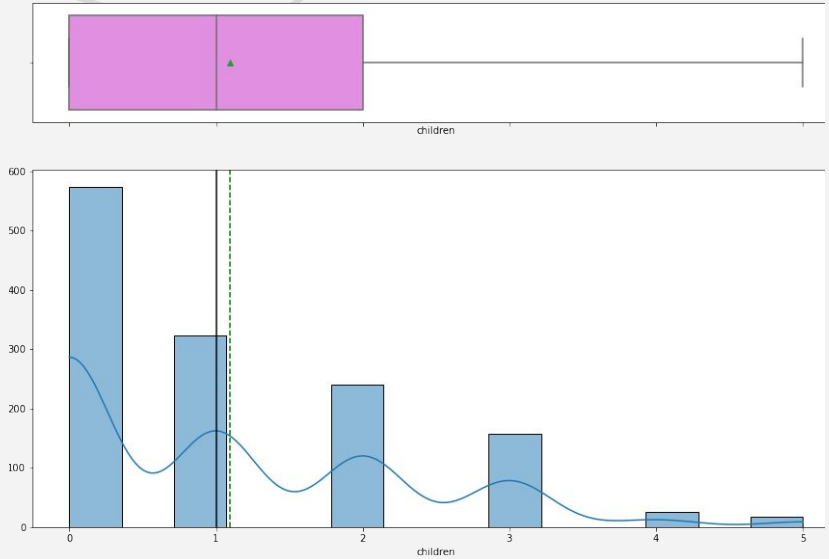
- The data appears normally distributed
- Mean (30,66) and Median (30.40) are close in value
- There are multiple outliers in the higher range of the BMI data with a maximum value of 53.13

Sex	Mean BMI
Female	30.38
Male	30.94

BUSINESS OVERVIEW

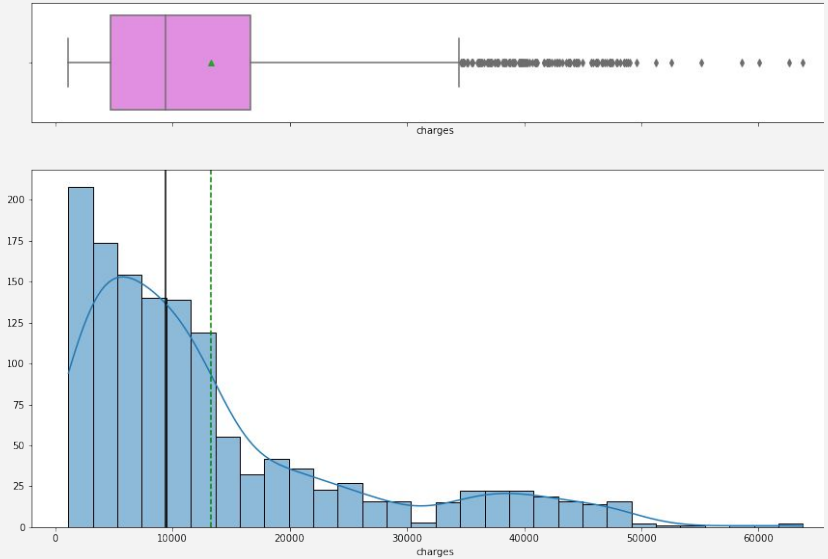
NO. OF CHILDREN & AMOUNT OF MEDICAL CHARGES

NUMBER OF CHILDREN OR DEPENDENTS



- Most customers have no children/dependents
- Number of children per customer declines as the number increases with the maximum being 5.
- 25% of customers have 2 or more children

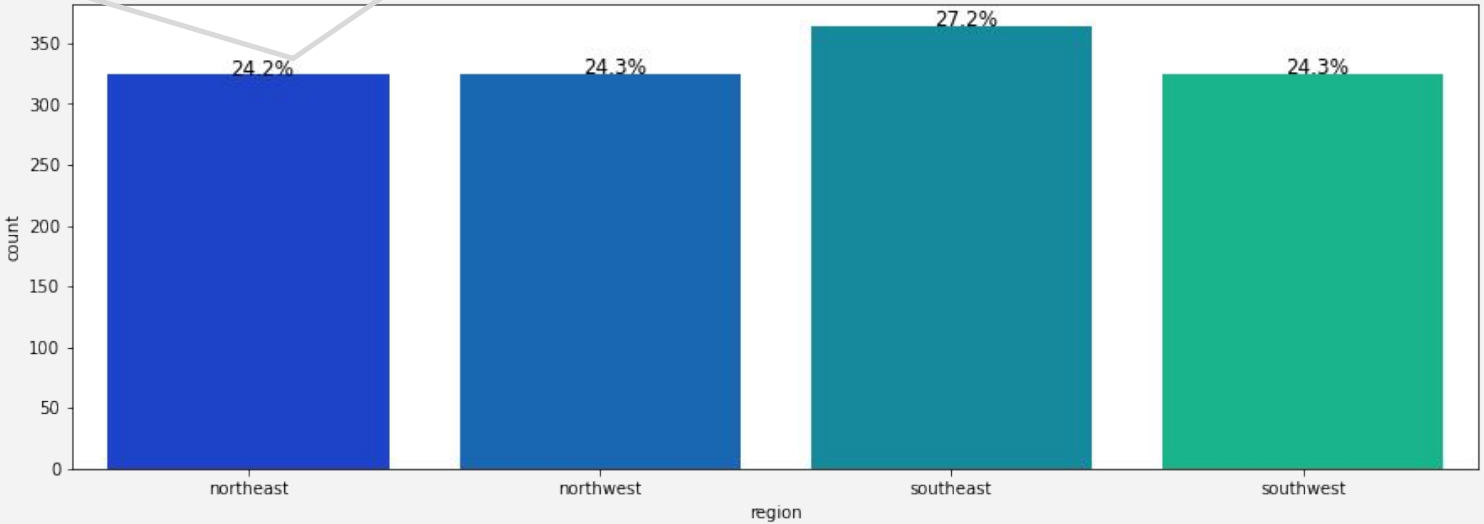
AMOUNT OF MEDICAL CHARGES



- The dataset for charges is left skewed indicating most customers have low or no medical costs but there are a fair amount of outliers in the long-tail with higher charges
- The average amount of charges is $13,270 and the median is $9,382

BUSINESS OVERVIEW
REGION

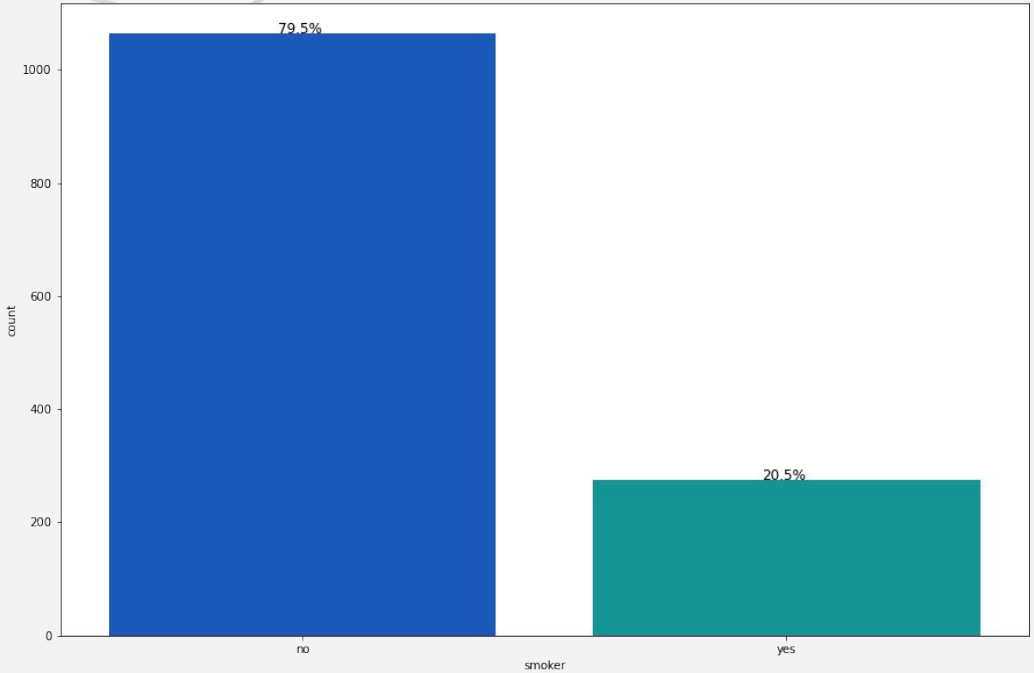
CUSTOMERS BY REGION



- There are four regions reported in the dataset Northeast (24.2%), Northwest (24.3%), Southeast (27.2%), Southwest (24.3%)

BUSINESS OVERVIEW
SMOKING STATUS

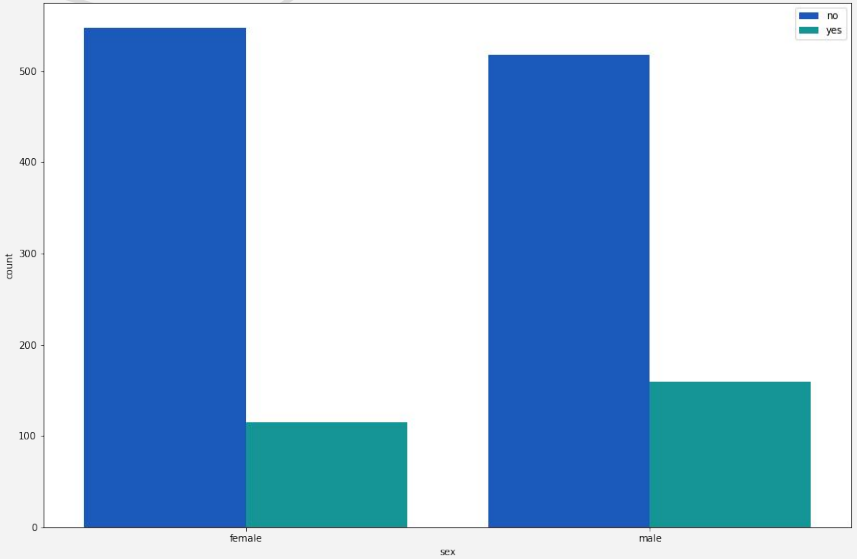
CUSTOMER SMOKING STATUS



- The majority of those surveyed identified as non-smokers 79.5% to smokers at 20.5%

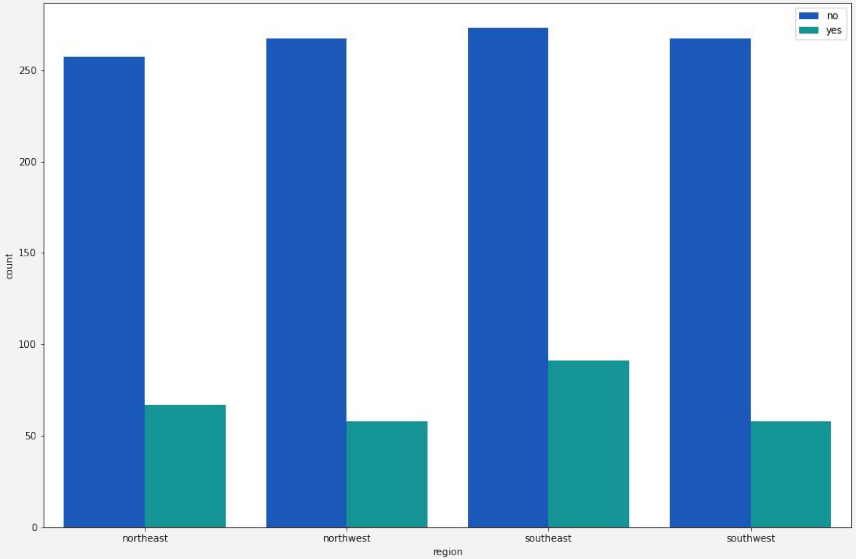
SMOKING STATUS ANALYSIS
SEX AND REGION

SEX VS. SMOKING STATUS



- More male customers smoke than do females

REGIONS VS. SMOKING STATUS



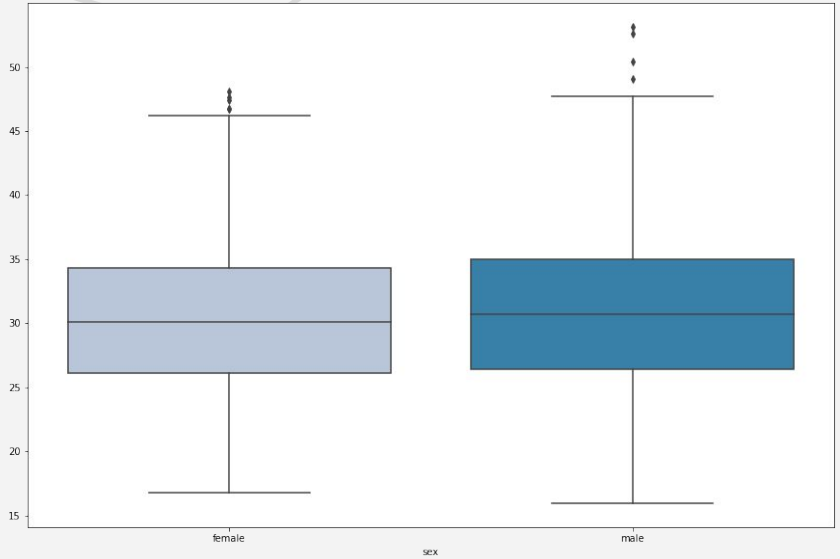
- Southeast have the most smokers

02

BIVARIATE & MULTIVARIATE ANALYSIS

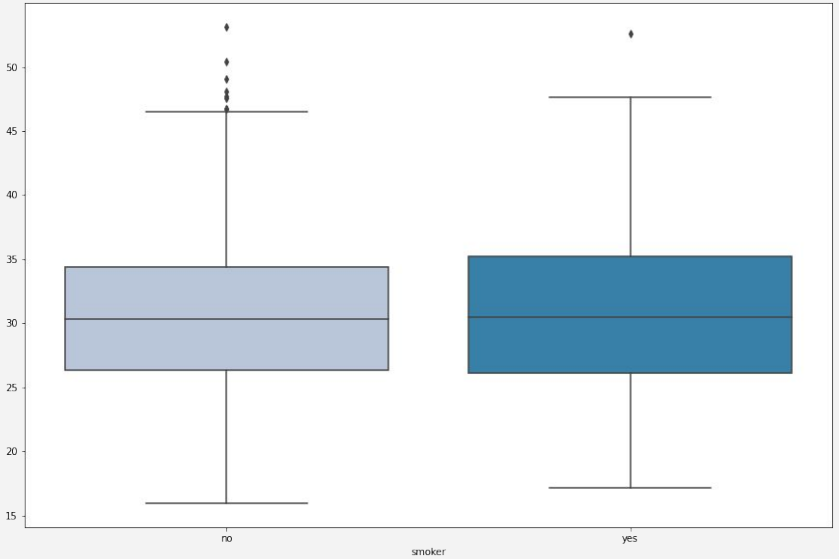
BODY MASS INDEX ANALYSIS
SEX AND SMOKING STATUS

SEX VS. BMI



- The BMI of both sexes seems to be similar with a few outliers in the upper range for both sexes but slightly more for men
- Visual observation of the above boxplot does not seem to indicate a substantial difference in BMI when accounting for sex

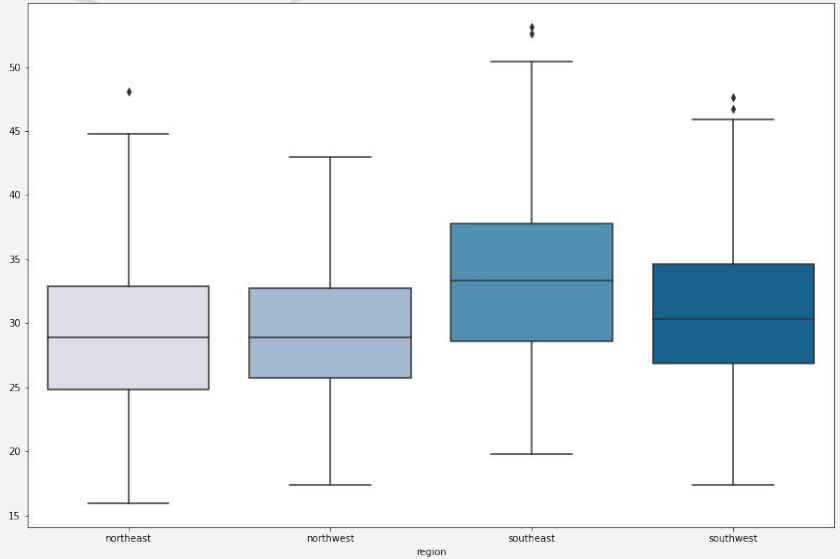
SMOKING STATUS VS. BMI



- The BMI for smokers and non-smokers appear similar with the exception of a few outliers in non-smokers and one outlier in smokers

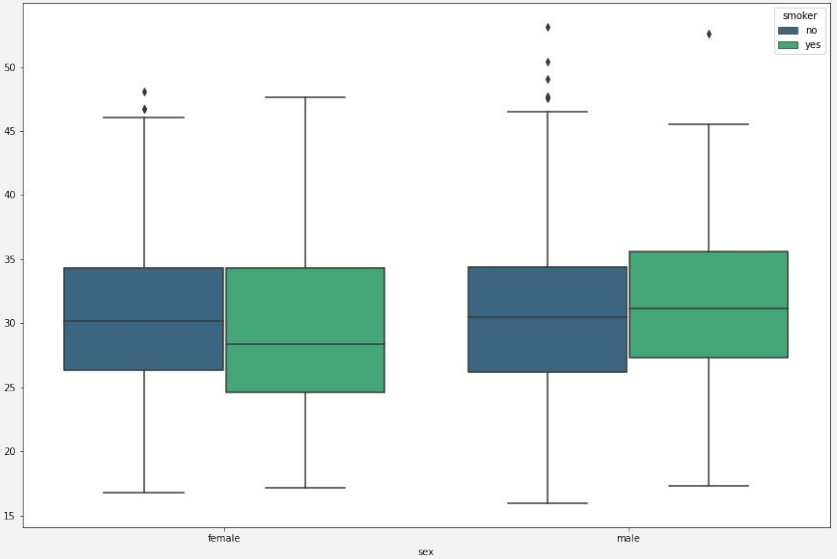
BODY MASS INDEX ANALYSIS
REGION AND SEX + SMOKING STATUS

REGION VS. BMI



- Customers in the Southeast region tend to have higher BMIs than customers in the other three regions
- Mean BMI: [NE:29.17, NW:29.10, SE:33.36, SW:30.60]

SEX VS. SMOKING STATUS VS. BMI

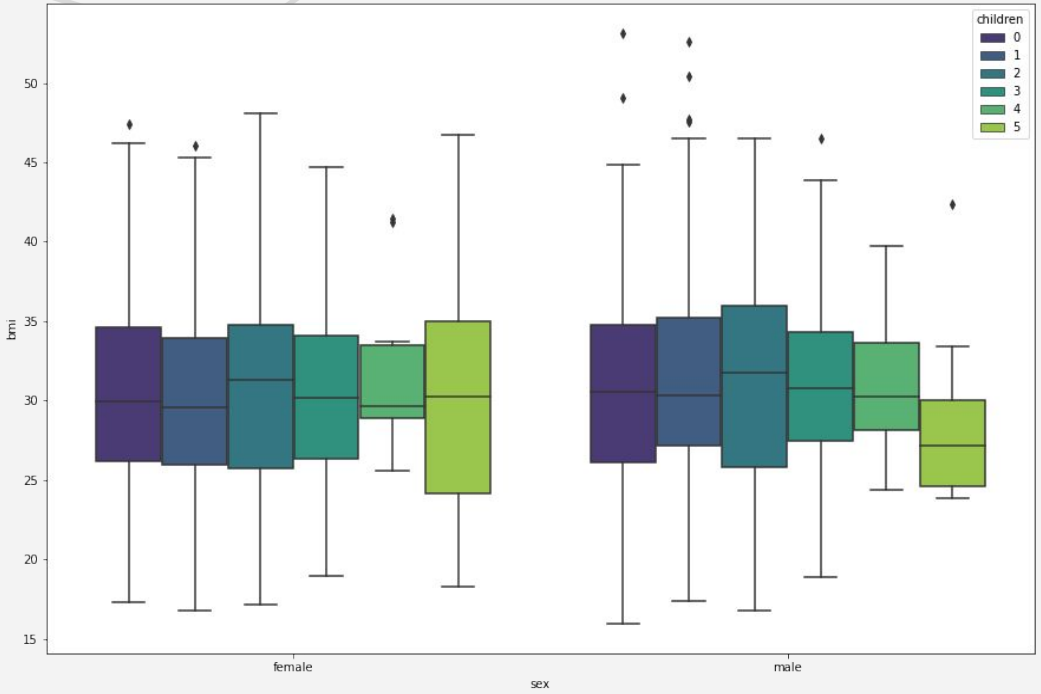


- The median BMI of females who smoke is lower than male smokers (there is one male outlier)
- The median BMI of non-smokers seems to be about the same regardless of sex (there are several outliers for both sexes)

BODY MASS INDEX ANALYSIS

SEX + NO. OF CHILDREN

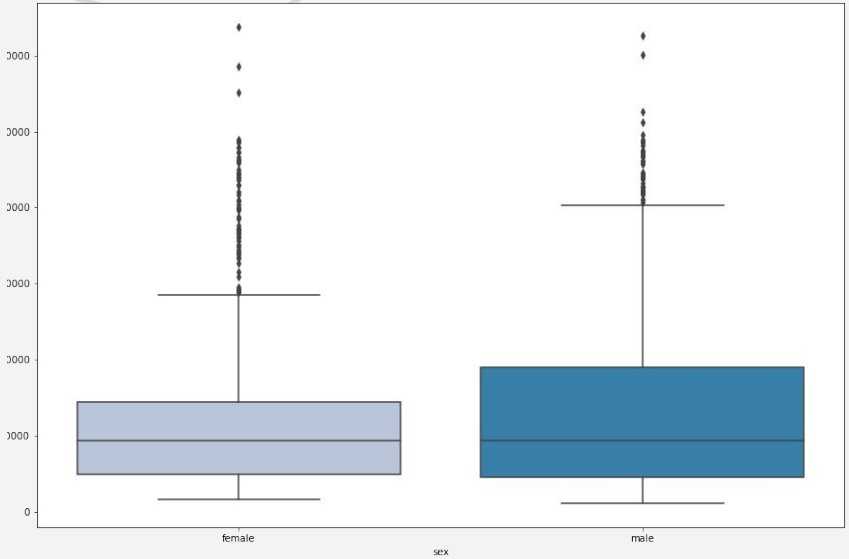
SEX VS. NO. OF CHILDREN VS. BMI



- BMI increases with the second child but declines from that level with 3 or more children
- Females increase BMI with 2-4 children but with 5 children return to about the same level as women with no children
- Men with 5 children have the lowest BMI of any group in this plot

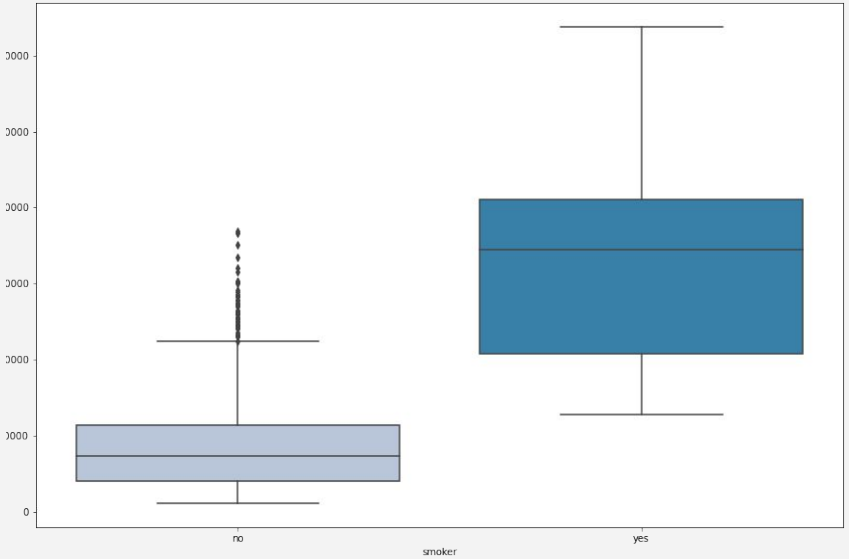
MEDICAL CHARGES ANALYSIS
SEX AND SMOKING STATUS

SEX VS. MEDICAL CHARGES



- Males have a wider range of charges than females but median charges appear the same regardless of sex
- Both sexes have many outliers at the upper end of charges

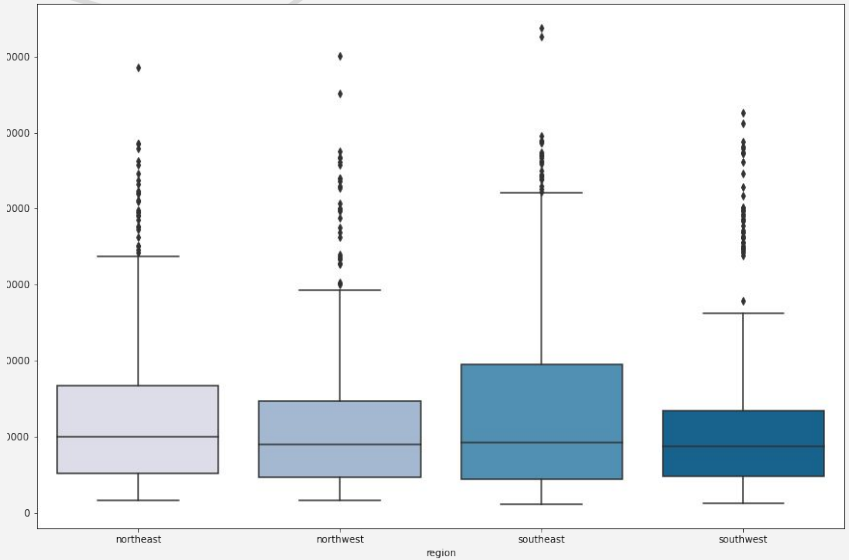
SMOKING STATUS VS. MEDICAL CHARGES



- Customers who smoke account for much higher medical charges than non-smoking customers
- The mean amount of medical charges for smokes is $32,050 compared to just $8,234 for non smoking customers

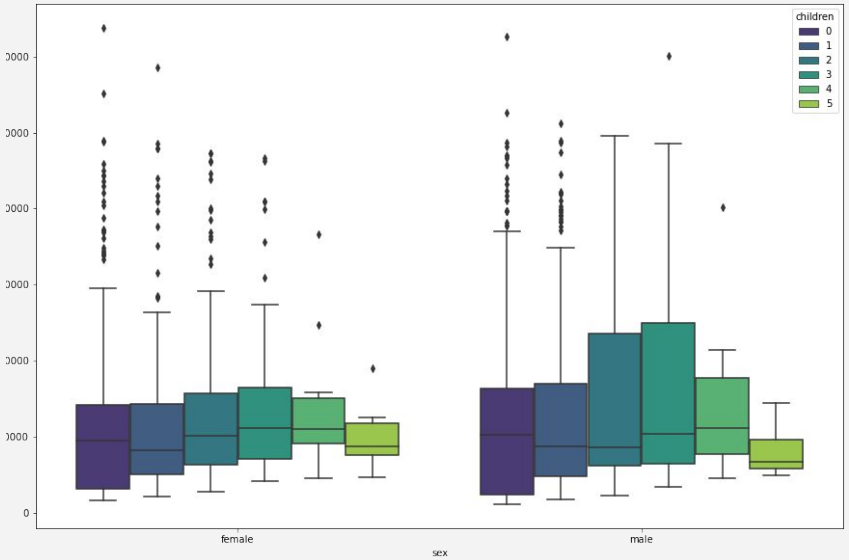
MEDICAL CHARGES ANALYSIS
REGION AND SEX + CHILDREN/DEPENDENTS

REGION VS. MEDICAL CHARGES



- Median charges across regions is similar however, the southeast seems to have a wider range and slightly higher amount of charges per customer.
- The Southeast also reports more smokers than the other regions which could account for the increase in medical charges

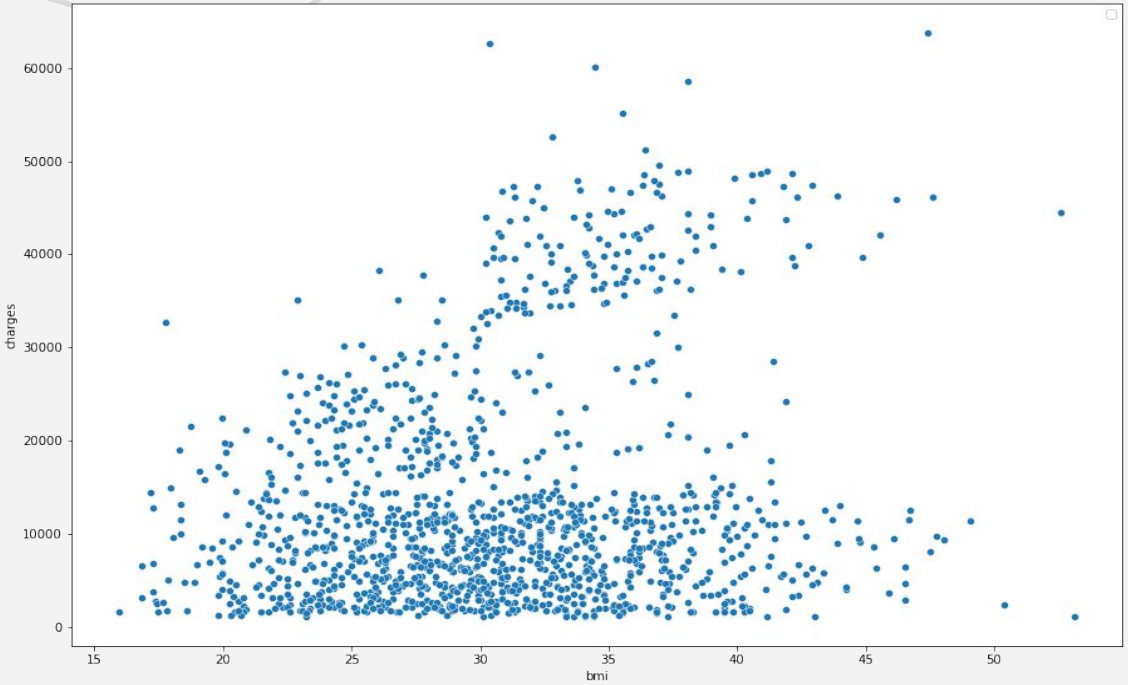
SEX VS. VS. CHILDREN/DEPENDENTS VS. MEDICAL CHARGES



- Women with 3-4 children/dependents tend to have higher medical charges as one would expect

MEDICAL CHARGES ANALYSIS
BODY MASS INDEX

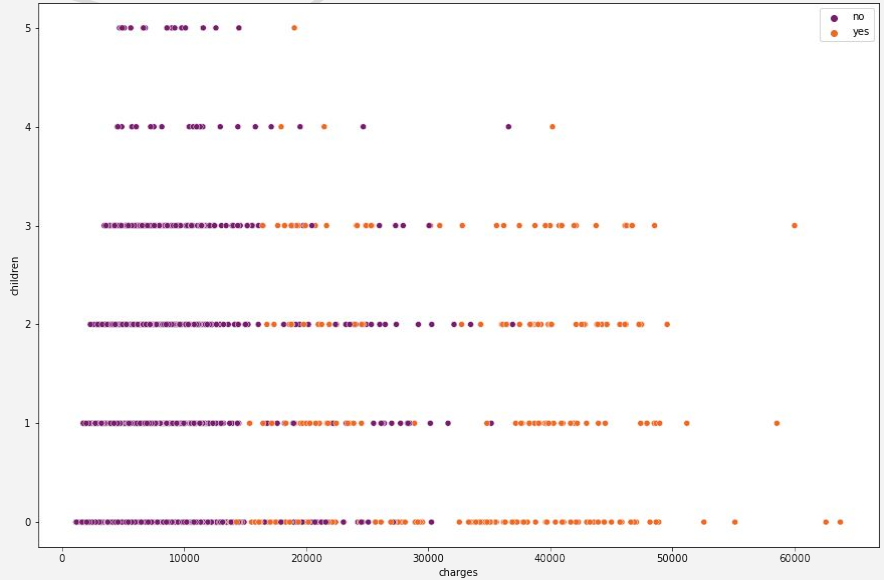
BMI VS. MEDICAL CHARGES

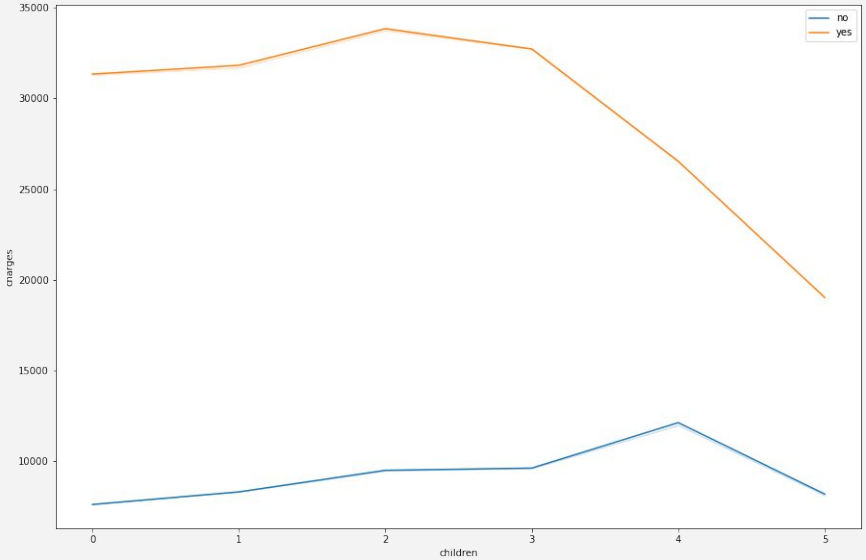


- Customers with a higher body mass index seem to have higher medical charges

SMOKING STATUS ANALYSIS
NO. CHILDREN/DEPENDENTS + MEDICAL CHARGES

SMOKING STATUS VS. CHILDREN/DEPENDENTS VS. MEDICAL CHARGES

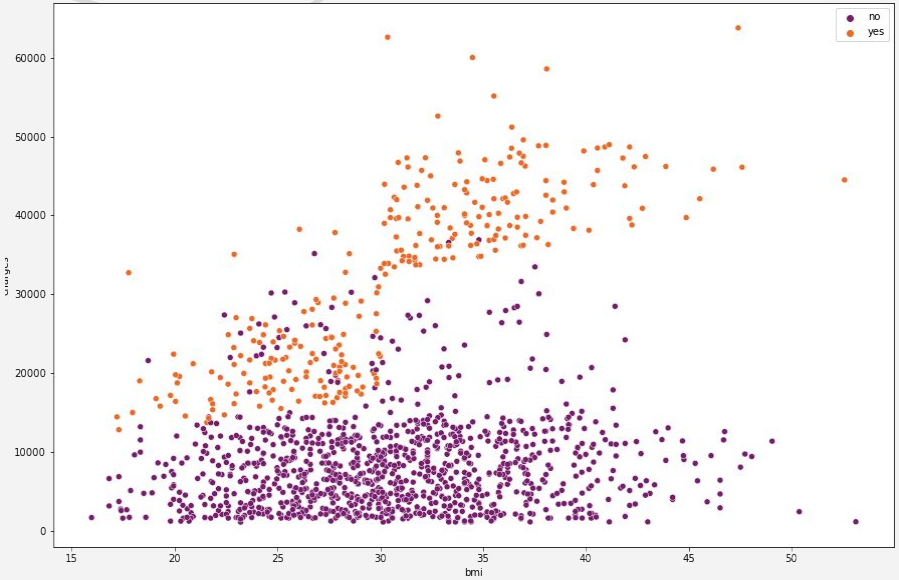


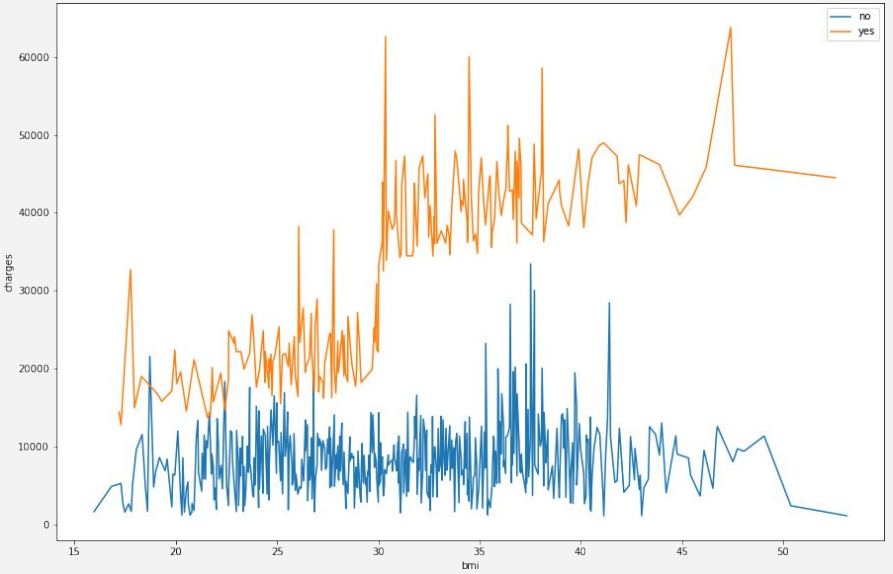


- Smoking status is more responsible for higher medical charges than are number of children/dependents

SMOKING STATUS ANALYSIS
BMI + MEDICAL CHARGES

SMOKING STATUS VS. BMI VS. MEDICAL CHARGES





- Smoking status is more responsible for higher medical charges than body mass index

03

KEY QUESTIONS

HYPOTHESIS TESTING

COMPARE MEDICAL CHARGES OF SMOKERS TO NON-SMOKERS

HYPOTHESIS TESTING

Key Question

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

Details

The level of significance (α) = 0.05.

The sample size , Non-Smokers (N = 1064) and Smokers (N = 274) but since the population standard deviation (σ) is unknown, we have to use a t-test.

Degree of Freedom: We have N-1 degrees of freedom : 1063 and 273 respectively

Since the purpose of the test is to determine if medical claims by people who smoke is 'greater' than those who don't we would compare the means of both samples by using a two-sample, right tailed, t-test

Hypothesis Formulation

$H_0: \mu_{smoker\ charges} = \mu_{non-smoker\ charges}$ (Smokers have same medical claims as non-smokers)

$H_a: \mu_{smoker\ charges} > \mu_{non-smoker\ charges}$ (Smokers have significantly greater medical claims compared to non-smokers)

Key Insight: Customer that smoke have higher medical charges than non-smoking customers

As the p-value (~ 2.94473222335849e-103) is less than the level of significance we can reject the null hypothesis. Hence, we do have enough evidence to support the claim that the population of customers who smoke average more medical charges than customers that do not smoke.

COMPARE BODY MASS INDEX (BMI) OF FEMALE CUSTOMERS TO MALE CUSTOMERS

HYPOTHESIS TESTING

Key Question

Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.

Details

The level of significance (α) = 0.05.

The sample size , N = 1338 but since the population standard deviation (σ) is unknown, we have to use a t-test.

Degree of Freedom: We have N-1 degrees of freedom : 1337

Since the purpose of the test is to determine if the BMI of Females is 'not equal' to the BMI of Males we would compare the means of both samples by using a two-sample, two-tailed, t-test

Hypothesis Formulation

$H_0: \mu_{female\ bmi} = \mu_{male\ bmi}$ (Female BMI is the same as Male BMI)

$H_a: \mu_{female\ bmi} \neq \mu_{male\ bmi}$ (Females BMI differs significantly compared to Male BMI)

Key Insight: The BMI of female customers does not differ significantly from that of male customers

As the p-value (~0.90) is not less than the level of significance we cannot reject the null hypothesis. Hence, we do not have enough evidence to support the claim that the BMI of Females is significantly different than that of Males.

COMPARE THE PROPORTION OF SMOKERS ACROSS DIFFERENT REGIONS

HYPOTHESIS TESTING

Key Question

Is the proportion of smokers significantly different across the four different regions (Southwest, Southeast, Northwest, Northeast)?

Details

The level of significance (α) = 0.05.

Since the purpose of the test is to determine if the proportion of smokers is independent vs dependent on region we will create a contingency crosstab table and use the chi-squared contingency test to test for independence

Hypothesis Formulation

H_0: Proportion of smokers is independent of region
H_a: Proportion of smokers depends on region

Key Insight: The proportion of smokers is not dependent upon region

Although the p-value (~0.062) is close, it is not less than the level of significance therefore we cannot reject the null hypothesis. Hence, we do not have enough evidence to support the claim that proportion of smokers depends on region.

COMPARE THE BODY MASS INDEX (BMI) OF FEMALE CUSTOMERS WITH 0, 1, OR 2 CHILDREN

HYPOTHESIS TESTING

Key Question

Prove (or disprove) with statistical evidence that the BMI of women 0, 1, or 3 children is the same or different.

Details

The level of significance (α) = 0.05.

The sample size , N = 662 but since the population standard deviation (σ) is unknown, we have to use a t-test.

Degree of Freedom: We have N-1 degrees of freedom : 661

Since the purpose of the test is to determine if the BMI of Females with 0, 1, or 2 children is the same or significantly different we will compare the means of the three samples by using a one-way ANOVA test.

Hypothesis Formulation

H_0: $\mu_1 = \mu_2 = \mu_3$ (The mean BMI of each sample are equal to each other)
H_a: The mean BMI of at least one sample is different

Key Insight: The BMI of female customers with 0, 1, or 2 children is not significantly different

As the p-value (~ 0.7158) is not less than the significance level (0.05), we cannot reject the null hypothesis. Hence, we do not have enough statistical significance to conclude that the BMI of women with 0, 1, or 2 children is different at a 5% significance level.

04

KEY INSIGHTS & RECOMMENDATIONS

KEY INSIGHTS

CUSTOMER DEMOGRAPHICS
This dataset consisted mostly of non-smoking males with one-child or dependent. Customers hailed almost evenly from all three regions with the Southeast having small advantage in numbers.

BODY MASS INDEX
Average BMI is 30.66 which is well above the ideal range of 18.5 to 24.9. Neither sex (male 30.94, female 30.38) nor smoking status (no 30.65, yes 30.71) had a meaningful impact on BMI. Customers from the Southeast (33.36) had the highest average BMI and women who smoke had among the lowest.

SMOKING STATUS & MEDICAL CHARGES
Smoking status more than any other factor (BMI, Sex, Region, Children) had the most profound impact on the amount of medical charges a customer would accumulate. Men smoke more than women. 79.5% of customer do not smoke.

RECOMMENDATIONS

- Smokers account for the majority of medical charges. The company should charge higher premiums for customers that smoke and look to implement programs to help customers quit smoking.

- A higher BMI also has a loose relationship with higher medical charges. Many of the company's customers have a BMI above the ideal range. The company should institute programs or benefits to customers which fall within the acceptable BMI range.

- Women with 3-4 dependents have among the highest medical charges . The company could explore raising premiums for parents with more than two children or dependents.

SUMMARY OF KEY QUESTION RESULTS - HYPOTHESIS TESTING

1. **Key Question - Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?**
   Key Insight: Customer that smoke have higher medical charges than non-smoking customers

2. **Key Question - Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.**
   Key Insight: The BMI of female customers does not differ significantly from that of male customers

3. **Key Question - Is the proportion of smokers significantly different across the four different regions (Southwest, Southeast, Northwest, Northeast)?**
   Key Insight: The proportion of smokers is not dependent upon region

4. **Key Question - Prove (or disprove) with statistical evidence that the BMI of women 0, 1, or 3 children is the same or different.**
   Key Insight: The BMI of female customers with 0, 1, or 2 children is not significantly different