greatlearning
Learning for Life

Axis Insurance Case Study

Objective

To conduct Exploratory Data Analysis on the Axis Insurance customer dataset and extract insights.

We will be focusing on the following:

- Come up with initial hypothesis from the data set through insights from data exploration
- Perform statistical hypothesis testing on those insights

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

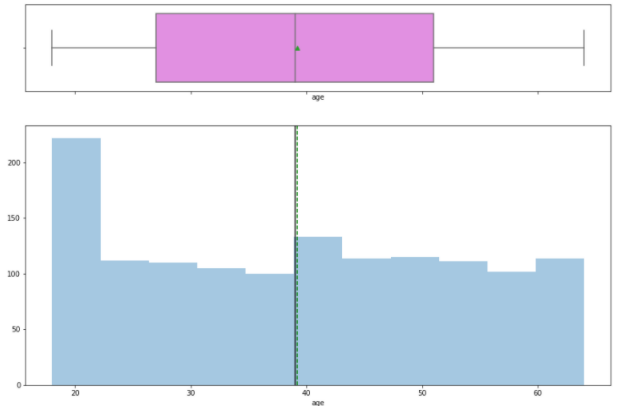
Observations	Variables
1338	7

Note:

- There are no missing values in the data set.
- The sex, smoker and region columns have been converted to category from object type.

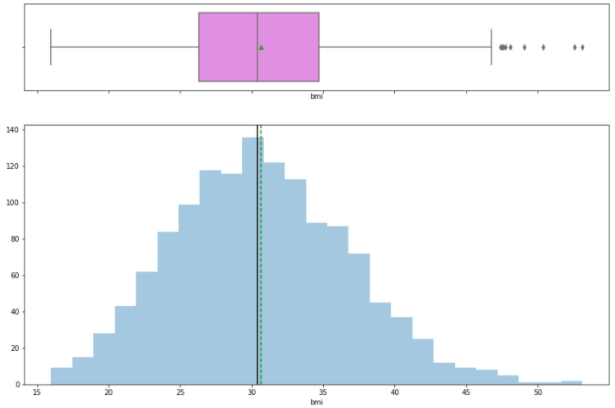
1. Exploratory Data Analysis – Age, BMI & Children

Age



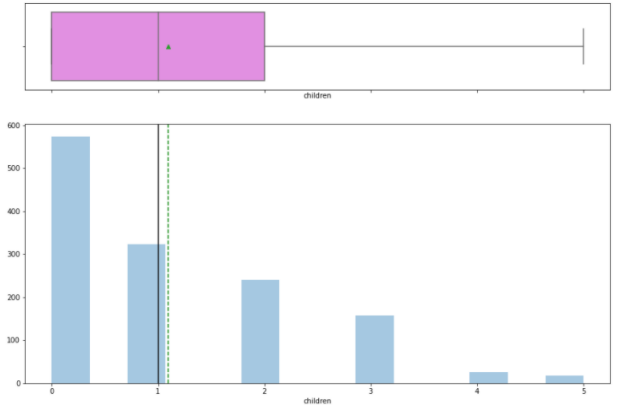
- The distribution of age is approximately normal except for high numbers of insurance customers in the early 20s.
- There are no outliers in this variable.
- From boxplot we can see that the third quartile(Q3) is equal to 52 which means 75% of customers are below the age of 52 and 1st quartile (Q1) is equal to 27 which means 25% of customers are below the age of 27.

BMI



- The mean and median BMI is 30.
- The distribution of customer BMI is normal and slightly skewed to the right.
- There are handful of outliers towards the higher BMI end.

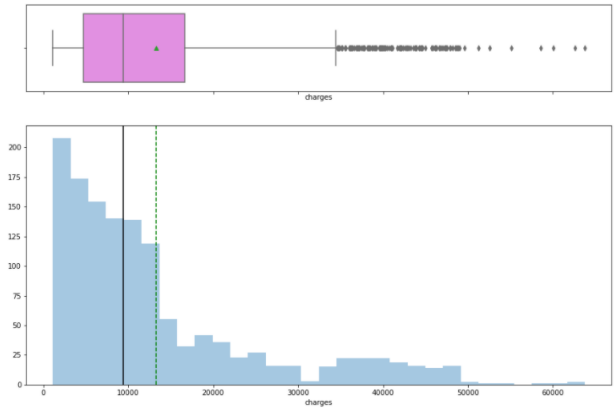
Children



- Customers in the sample mostly have 0 to 3 children.
- There are no outliers.

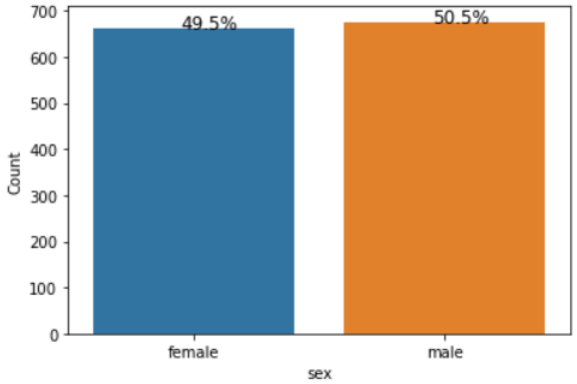
1. Exploratory Data Analysis – Charges, Sex & Smoker

Charges



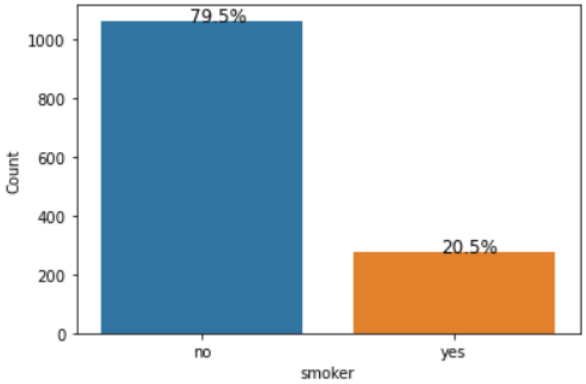
- Mean medical claims in the insurance sample is around 14000 dollars and Median medical claims is around 9500 dollars.
- Medical claims are right skewed.
- It has outliers towards the higher claim end.

Sex



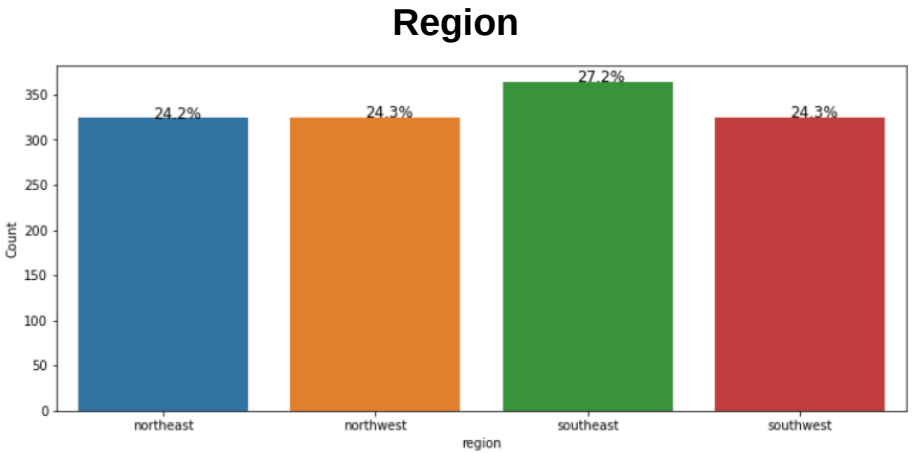
- Males and Females are about similar in numbers at 676 (50.5%) and 662 (49.5%) respectively in the insurance sample.

Smoker



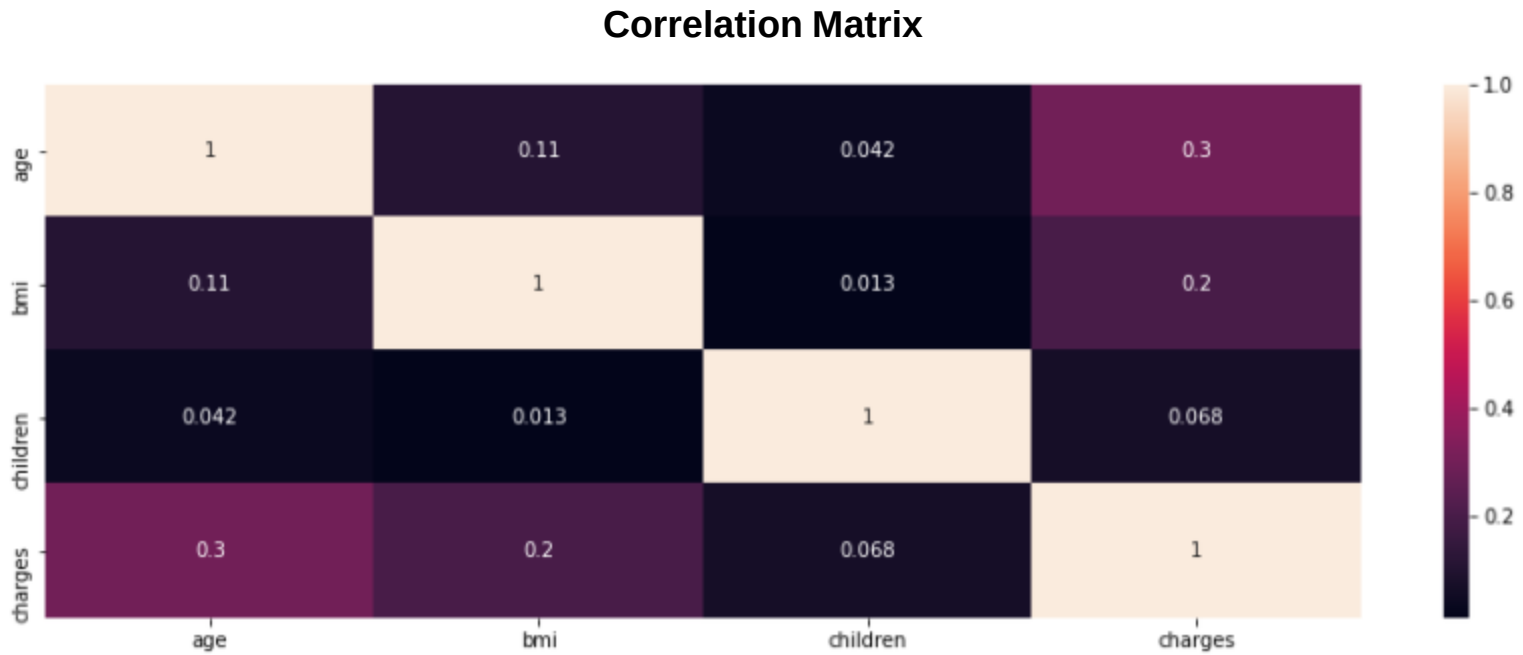
- Non smokers outnumber smokers 4:1 ratio at 1064 (79.5%) and 274 (20.5%) respectively in the insurance sample.

1. Exploratory Data Analysis – Region



- Customers are quite evenly spread across the regions with southeast region having slightly more share (27.2%) at 364 customers.

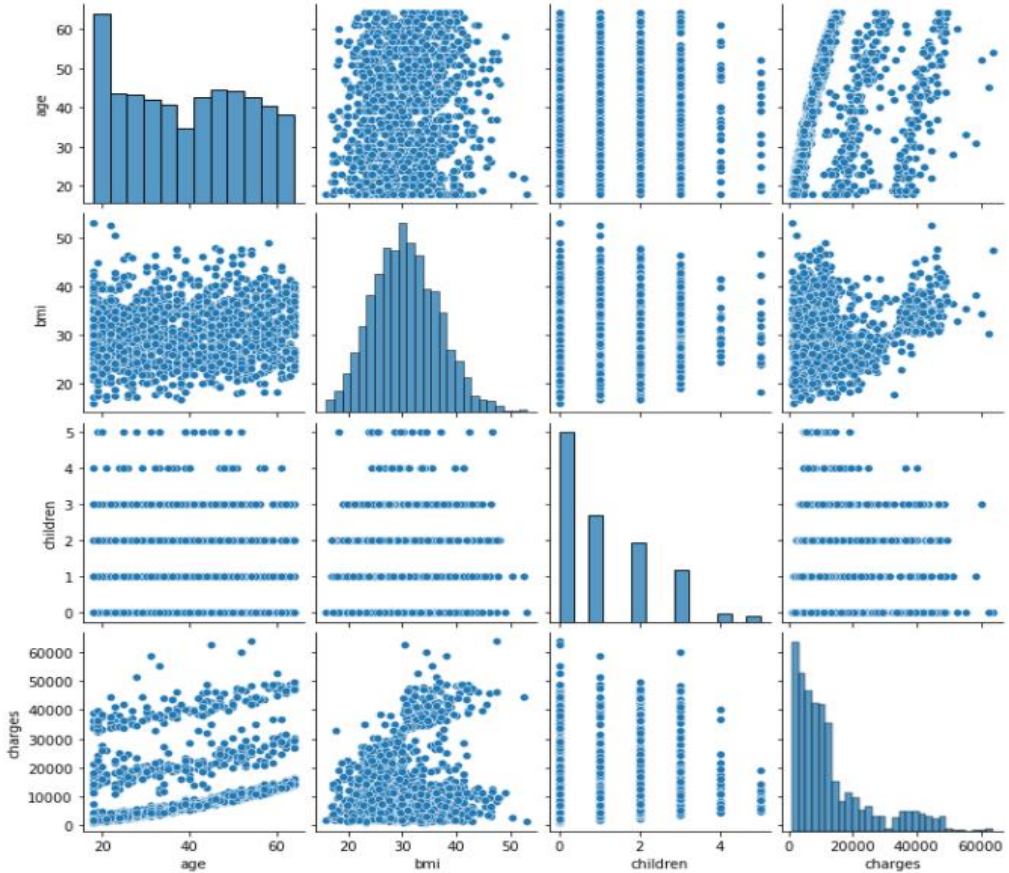
1. Exploratory Data Analysis – Correlation Matrix



Observations:

- There is no striking correlation found among these variables.
- There is a light positive correlation between age and medical charges.

1. Exploratory Data Analysis – Pairplot

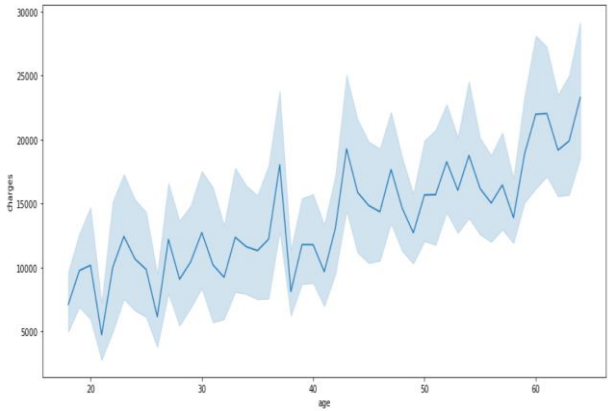


Observations:

- We can see varying distributions in variables with no discerning pattern, we should investigate further.
- BMI is normal distributed with slight right skew.

1. Exploratory Data Analysis – Medical Charges with Age, BMI & Children

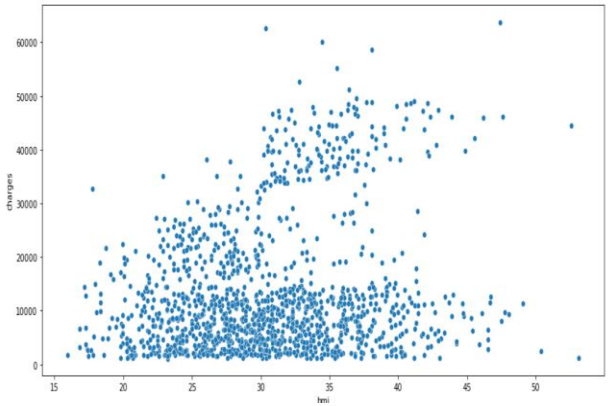
Age Vs Charges



Observations:

- Medical charges tend to trend higher as customer age increases.

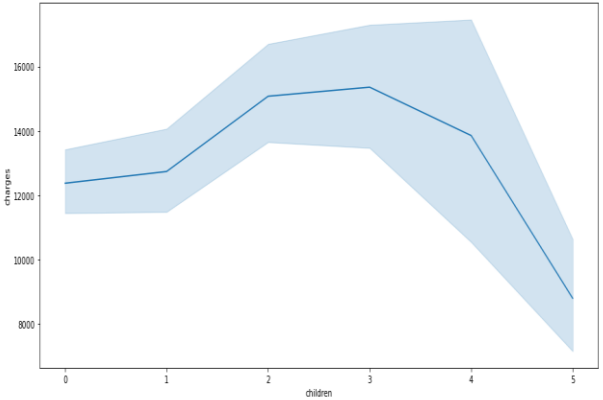
BMI Vs Charges



Observations:

- Not much of a clear discerning relation can be found between BMI and medical charges.

Number of Children Vs Charges

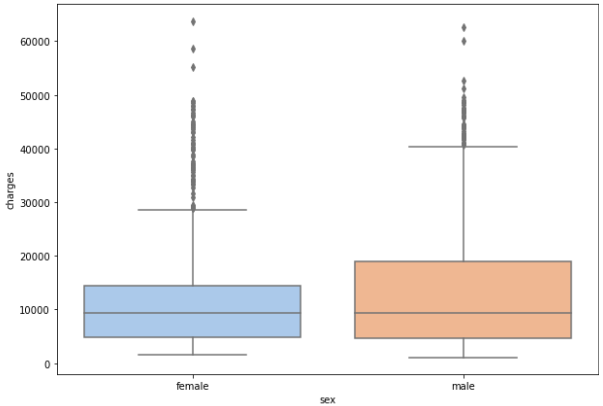


Observations:

- Medical charges tend to trend downward with customers having more children.

1. Exploratory Data Analysis – Medical Charges with Sex, Smoker & Region

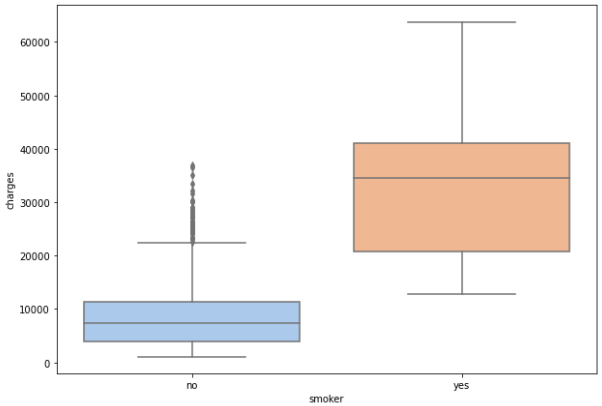
Sex Vs Charges



Observations:

- Median and minimum medical charges are about the same for both sexes.
- Male medical charges range higher than female medical charges at the higher range of 40K dollars compared to 30K dollars with both having outliers.

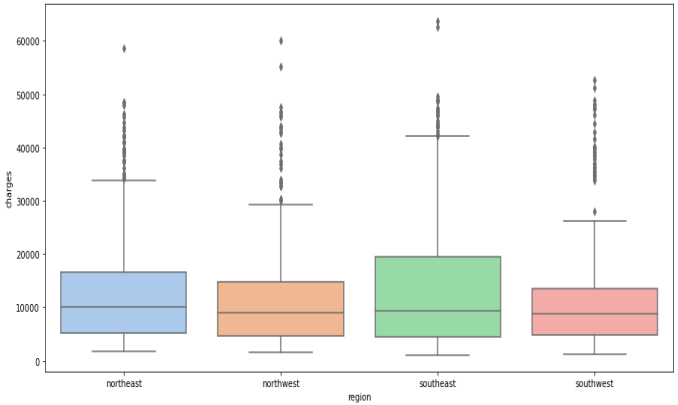
Smoker Vs Charges



Observations:

- Smokers have higher medical charge ranges on average than non smokers at ~ 4 times.
- Non Smokers exhibit some medical charge outliers.

Region Vs Charges

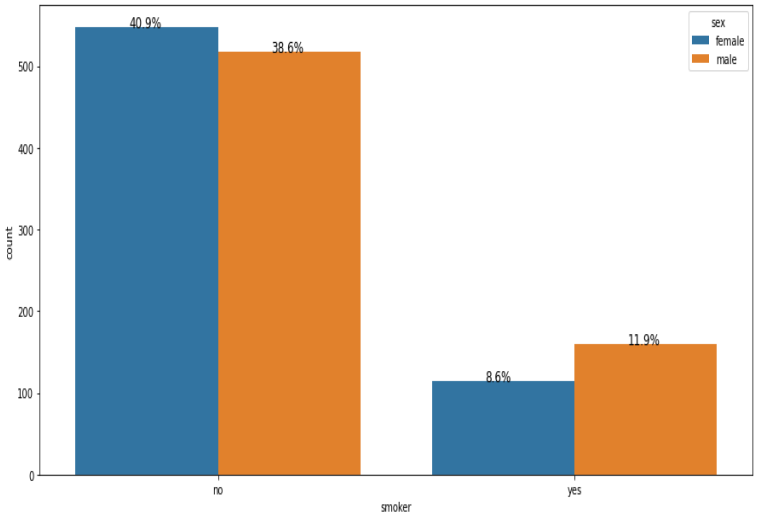


Observations:

- Customers from all residential regions have somewhat similar median and range of medical charges.
- Southeast region exhibit a slightly higher range of medical charges on the higher end at above 40K dollars.

1. Exploratory Data Analysis – Smoker with Sex & Region

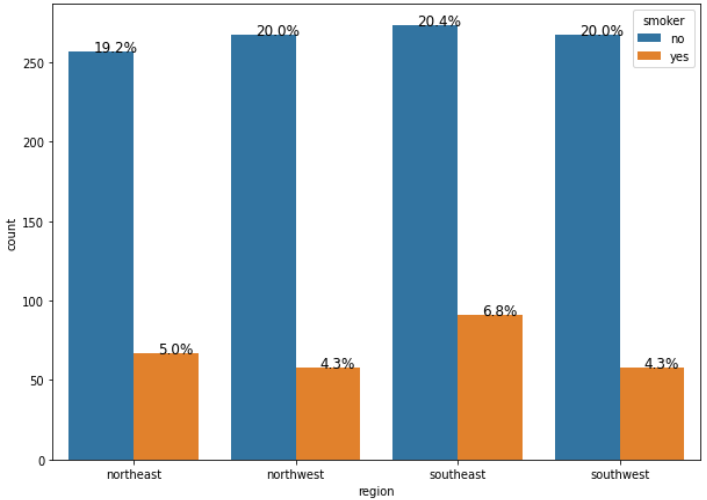
Smoker Vs Sex



Observations:

- Male smokers outnumber female smokers.
- Non Smokers are the vast majority in the sample and by extension the population.

Smoker Vs Region

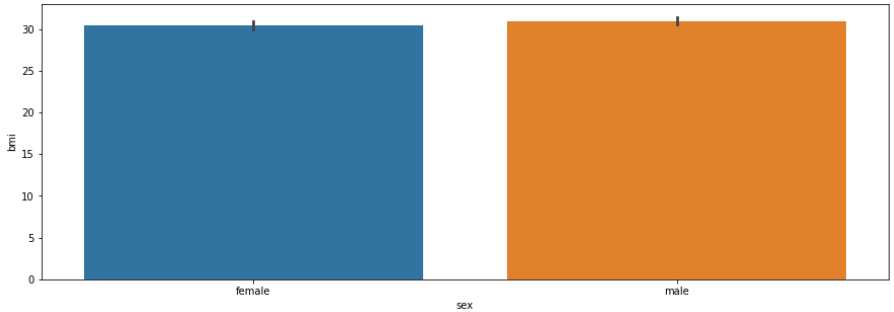


Observations:

- Non-smoker numbers are comparable across regions but smoker numbers are markedly higher and could be the main driver of higher medical charges in the Southeast region.

1. Exploratory Data Analysis – BMI with Sex & Children

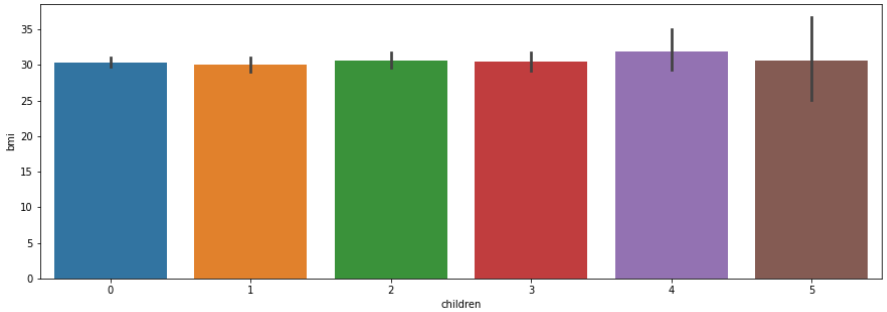
BMI Vs Sex



Observations:

- Male and female do not differ much in their BMI.

BMI Vs Number of Children



Observations:

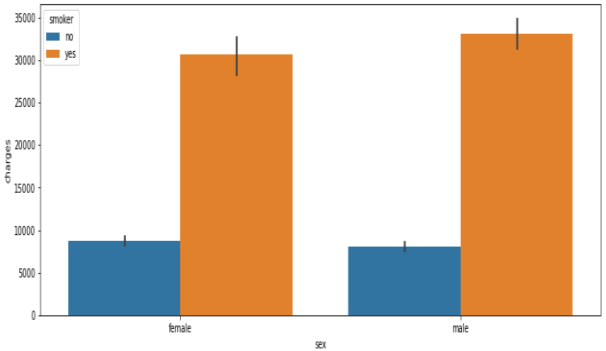
- There is not much significant difference in BMI between women who have different number of children other than women with 4 to 5 children have higher BMI variance.

1. Customer Profiles

- Medical charges tend to trend higher as customer age increases which is expected.
- Male medical charges range higher than female medical charges at the higher range of 40K dollars compared to 30K dollars.
- Smokers have higher medical charge ranges on average than non smokers at ~ 4 times and Male smokers outnumber female smokers.
- Southeast region exhibit a slightly higher range of medical charges on the higher end at above 40K dollars.
- Non-smoker numbers are comparable across regions but smoker numbers are markedly higher and could be the main driver of higher medical charges in the Southeast region.
- Male and female do not differ much in their BMI.
- There is not much significant difference in BMI between women who have different number of children other than women with 4 to 5 children have higher BMI variance.

1. Customer Segmentation

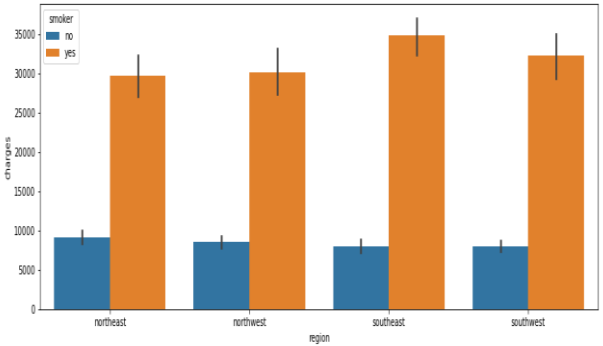
Sex Vs Smoker Vs Charges



Observations:

- Male smokers have higher medical charges at 32.5K dollars compared to female 30K dollars.

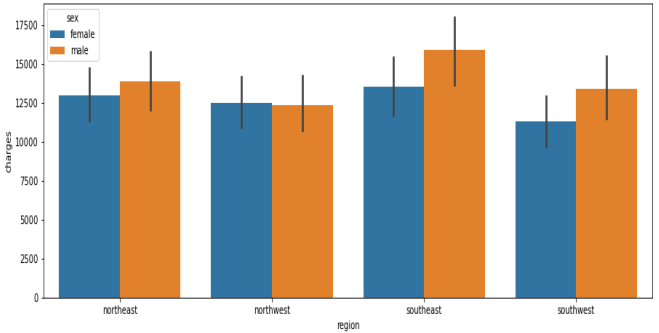
Region Vs Smoker Vs Charges



Observations:

- Southeast region exhibit a higher range of medical charges largely driven by their smoker segment.

Region Vs Sex Vs Charges



Observations:

- Males constitute a higher delta than females in medical charges in Southeast region than most other regions.

1. Conclusion

After all the analysis, we have been able to conclude that

- Smokers claim higher medical charges at ~ 4 times than non smokers. Male smokers also claim higher medical charges than female smokers and they outnumber female smokers.

- The Southeast region boosts a higher medical charge per capita largely driven by their larger number of smokers than other regions with the insurance package. It also has a higher proportion of smokers at ~ 33% of the insurance customer base.

- Male and female do not differ much in their BMI. There is also not much significant difference in BMI between women who have different number of children other than women with 4 to 5 children have higher BMI variance.

1. Recommendations

Based on the analysis, the following recommendations can improve the sales and profits

- Marketing the insurance packages to non smokers and females can help in increasing sales and profits as these 2 segments of customers claim less than male and/or smokers.
- Marketing to potential customers in the west of the country is recommended as these regions have less of a proportion of smokers.
- Potentially exploring reducing premiums for non smokers and increasing for smokers as smokers on average claim much higher medical chargers than non smokers. This can entice more non smokers to join the insurance packages and entice existing customers to consider quitting smoking to enjoy premium reduction benefits.

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

Let μ_1, μ_2 be the mean medical claims of smokers and non-smokers respectively.

We will test the null hypothesis

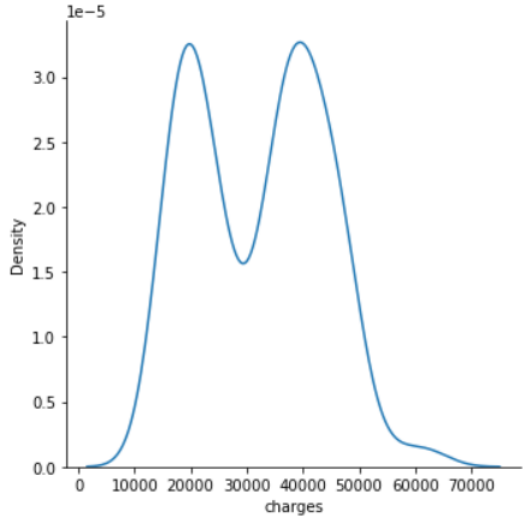
> $H_0 : \mu_1 \leq \mu_2$

against the alternate hypothesis

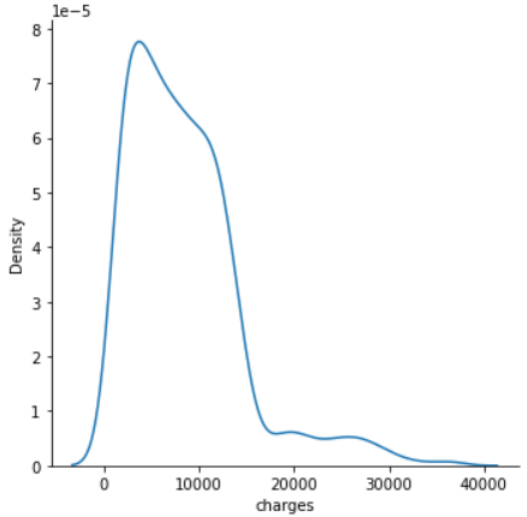
> $H_a : \mu_1 > \mu_2$

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

Smoker Medical Claim Charges Distribution



Non Smoker Medical Claim Charges Distribution



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

- Sample Statistics
  - The mean medical claims of smokers is 32050.23
  - The mean medical claims of non smokers is 8434.27
  - The standard deviation of medical claims of smokers is 11541.55
  - The standard deviation of medical claims of non smokers is 5993.78
- Let's test whether the T-test assumptions are satisfied or not
  - Continuous data - Yes, the medical claims are measured on a continuous scale.
  - Normally distributed populations - Yes, we assume that the populations are assumed to be normal.
  - Independent populations - As we are taking random samples for two different groups, the two samples are from two independent populations.
  - Unequal population standard deviations - As the sample standard deviations are different, the population standard deviations may be assumed to be different.
  - Random sampling from the population - Yes, we assume that the collected sample a simple random sample.
- We can use two sample T-test (Unequal Std Dev) for this problem.

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

- The p-value is 2.94473222335849e-103
- Insight
    - As the p-value is much less than the level of significance of 0.05, we can reject the null hypothesis. Hence, we do have enough evidence to support the claim that medical claims made by the people who smoke is greater than those who don't.

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

Let μ_1, μ_2 be the mean BMI of females and males respectively.

We will test the null hypothesis

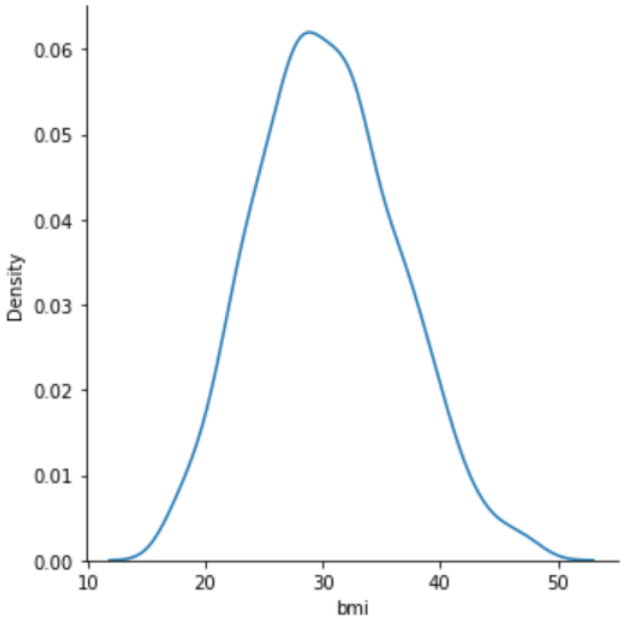
$$H_0 : \mu_1 = \mu_2$$

against the alternate hypothesis

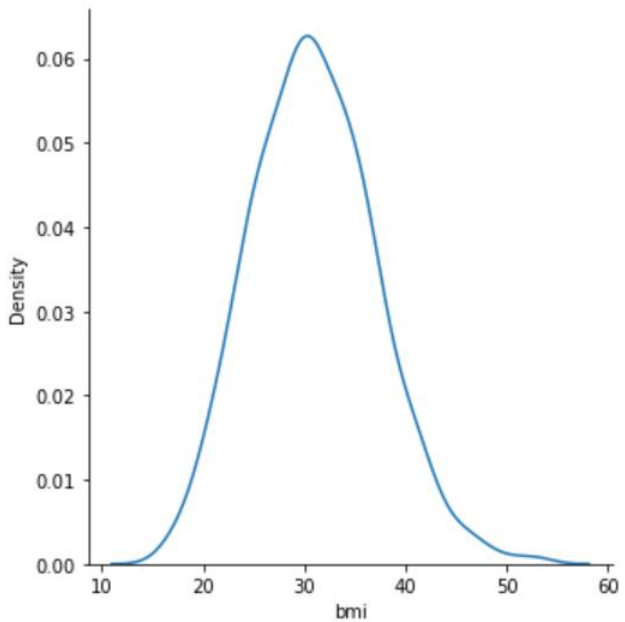
$$H_a : \mu_1 \neq \mu_2$$

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

Female BMI Distribution



Male BMI Distribution



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

- Sample Statistics
  - The mean BMI of females is 30.38
  - The mean BMI of males is 30.94
  - The standard deviation of medical claims of smokers is 6.05
  - The standard deviation of medical claims of non smokers is 6.14
- Let's test whether the T-test assumptions are satisfied or not
  - Continuous data - Yes, the BMI is measured on a continuous scale.
  - Normally distributed populations - Yes, we assume that the populations are assumed to be normal.
  - Independent populations - As we are taking random samples for two different groups, the two samples are from two independent populations.
  - Equal population standard deviations - As the sample standard deviations are almost equal, the population standard deviations may be assumed to be equal.
  - Random sampling from the population - Yes, we assume that the collected sample is a simple random sample.
- We can use two sample T-test (Equal Std Dev) for this problem.

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

- The p-value is 0.08997637178984934
- Insight
    - As the p-value(~0.09) is greater than the level of significance, we can not reject the null hypothesis. Hence, we do not have enough significance to conclude that BMI of females is different from that of males at 0.05 significance level.

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

- We have to test the whether 2 categorical variables are independent

smoker region	no	yes
northeast	257	67
northwest	267	58
southeast	273	91
southwest	267	58

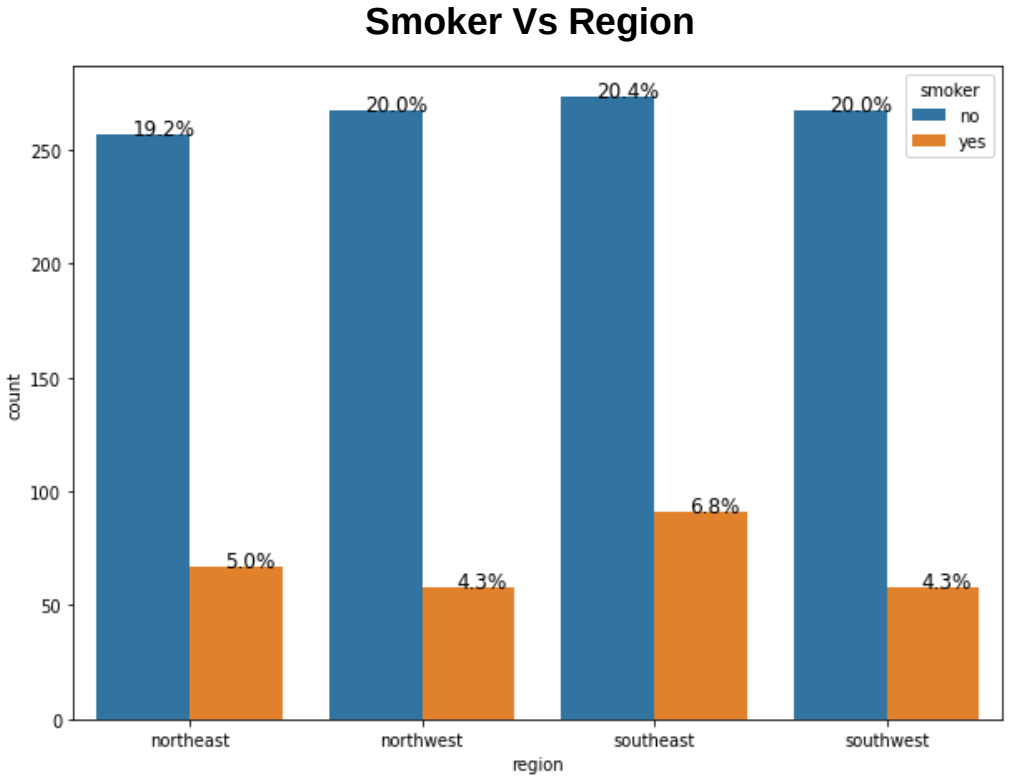
We will test the null hypothesis

> H_0 : Smoker numbers are independent of region.

against the alternate hypothesis

> H_a : Smoker numbers are dependant on region

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



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

- Let's test whether the assumptions are satisfied or not (chi2 contingency test for sample independence)
  - Categorical variables – Yes
  - Expected value of the number of sample observations in each level of the variable is at least 5 - Yes, the number of observations in each level is greater than 5.
  - Random sampling from the population - Yes, collected sample is a simple random sample of the population.
- The p-value is 0.06171954839170541
- Insight
  - As the p-value is just above the significance level, we fail to reject the null hypothesis. Hence, we do have enough statistical significance to conclude that smoker numbers are independent of region at 5% significance level. Z Test for 2 proportions can be used as the populations are independent.

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

A: Test North East vs North West Region for proportion of smokers

Let's write the null and alternative hypothesis

Let p_1, p_2 be the proportions of smokers in North East and North West regions respectively.

> $H_0 : p_1 = p_2$

against the alternate hypothesis

> $H_a : p_1 \neq p_2$

Let's test whether the Z-test assumptions are satisfied or not

- Binomally distributed population - Yes, a person is either smoker or non smoker.
- Random sampling from the population - Yes, we assume that the collected sample is a simple random sample.
- Can the binomial distribution approximated to normal distribution - Yes. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10.

```
The p-value is 0.3601698716632562
```

Insight

As the p-value is greater than the significance level 0.05, we can not reject the null hypothesis. Thus, we do not have enough statistical significance to conclude that the proportion of smokers in North East and North West regions are significantly different.

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

B: Test North East vs South East Region for proportion of smokers

Let's write the null and alternative hypothesis

Let p_1, p_2 be the proportions of smokers in North East and South East regions respectively.

> $H_0 : p_1 = p_2$

against the alternate hypothesis

> $H_a : p_1 \neq p_2$

Let's test whether the Z-test assumptions are satisfied or not

- Binomally distributed population - Yes, a person is either smoker or non smoker.
- Random sampling from the population - Yes, we assume that the collected sample is a simple random sample.
- Can the binomial distribution approximated to normal distribution - Yes. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10.

```
The p-value is 0.17861355436256732
```

Insight

As the p-value is greater than the significance level 0.05, we can not reject the null hypothesis. Thus, we do not have enough statistical significance to conclude that the proportion of smokers in North East and South East regions are significantly different.

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

C: Test North East vs South West Region for proportion of smokers

Let's write the null and alternative hypothesis

Let p_1, p_2 be the proportions of smokers in North East and South West regions respectively.

> $H_0 : p_1 = p_2$

against the alternate hypothesis

> $H_a : p_1 \neq p_2$

Let's test whether the Z-test assumptions are satisfied or not

- Binomally distributed population - Yes, a person is either smoker or non smoker.
- Random sampling from the population - Yes, we assume that the collected sample is a simple random sample.
- Can the binomial distribution approximated to normal distribution - Yes. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10.

```
The p-value is 0.3601698716632562
```

Insight

As the p-value is greater than the significance level 0.05, we can not reject the null hypothesis. Thus, we do not have enough statistical significance to conclude that the proportion of smokers in North East and South West regions are significantly different.

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

D: Test North West vs South East Region for proportion of smokers

Let's write the null and alternative hypothesis

Let p_1, p_2 be the proportions of smokers in North West and South East regions respectively.

> $H_0 : p_1 = p_2$

against the alternate hypothesis

> $H_a : p_1 \neq p_2$

Let's test whether the Z-test assumptions are satisfied or not

- Binomally distributed population - Yes, a person is either smoker or non smoker.
- Random sampling from the population - Yes, we assume that the collected sample is a simple random sample.
- Can the binomial distribution approximated to normal distribution - Yes. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10.

```
The p-value is 0.022789815463035743
```

Insight

As the p-value is less than the significance level 0.05, we can reject the null hypothesis. Thus, we do have enough statistical significance to conclude that the proportion of smokers in North West and South East regions are significantly different.

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

E: Test North West vs South West Region for proportion of smokers

Let's write the null and alternative hypothesis

Let p_1, p_2 be the proportions of smokers in North West and South West regions respectively.

> $H_0 : p_1 = p_2$

against the alternate hypothesis

> $H_a : p_1 \neq p_2$

Let's test whether the Z-test assumptions are satisfied or not

- Binomally distributed population - Yes, a person is either smoker or non smoker.
- Random sampling from the population - Yes, we assume that the collected sample is a simple random sample.
- Can the binomial distribution approximated to normal distribution - Yes. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10.

```
The p-value is 1.0
```

Insight

As the p-value is much much greater than the significance level 0.05, we cannot reject the null hypothesis. Thus, we do not have enough statistical significance to conclude that the proportion of smokers in North West and South West regions are significantly different.

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

F: Test South East vs South West Region for proportion of smokers

Let's write the null and alternative hypothesis

Let p_1, p_2 be the proportions of smokers in South East and South West regions respectively.

> $H_0 : p_1 = p_2$

against the alternate hypothesis

> $H_a : p_1 \neq p_2$

Let's test whether the Z-test assumptions are satisfied or not

- Binomally distributed population - Yes, a person is either smoker or non smoker.
- Random sampling from the population - Yes, we assume that the collected sample is a simple random sample.
- Can the binomial distribution approximated to normal distribution - Yes. For binary data, CLT works slower than usual. The standard thing is to check whether np and n(1-p) are greater than or equal to 10.

```
The p-value is 0.022789815463035743
```

Insight

As the p-value is less than the significance level 0.05, we can reject the null hypothesis. Thus, we do have enough statistical significance to conclude that the proportion of smokers in South East and South West regions are significantly different.

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

Final Statement

Proportion of smokers are significantly different for North West and South West regions compared to South East region.

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

Let μ_1, μ_2, μ_3 be the means of BMI for women with no children, one child and two children respectively.

We will test the null hypothesis

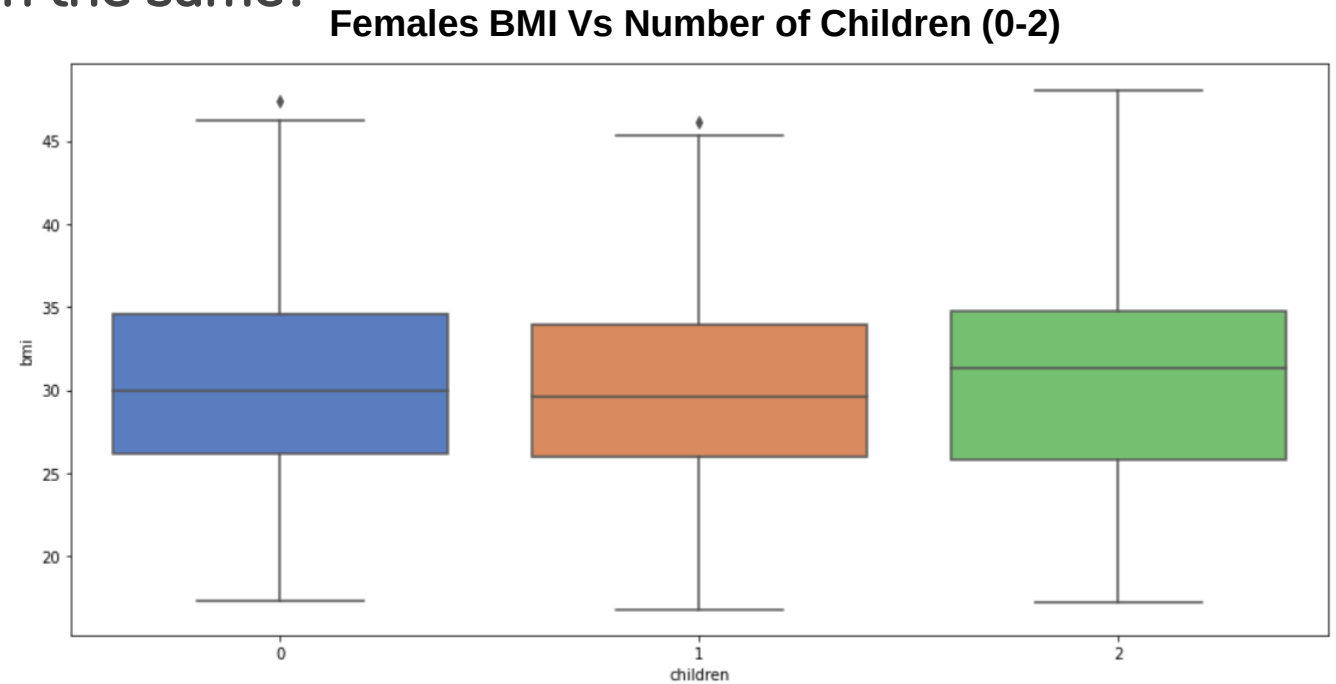
> $H_0 : \mu_1 = \mu_2 = \mu_3$

against the alternative hypothesis

> H_a : At least one BMI mean is different from the rest.

We will be using the ANOVA (Analysis Of Variance) Test to determine whether the means of 2 or more independent populations are significantly different.

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



Now, the normality and equality of variance assumptions need to be checked.

- For testing of normality, Shapiro-Wilk's test is applied to the response variable.
- For equality of variance, Levene test is applied to the response variable.

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

Shapiro-Wilk's test

We will test the null hypothesis

> H_0 : BMI follows a normal distribution against

against the alternative hypothesis

> H_a : BMI does not follow a normal distribution

```
The p-value is 0.010864038951694965
```

Since p-value of the test is large, we fail to reject the null hypothesis that the response follows the normal distribution.

Levene's test

We will test the null hypothesis

> H_0: All the population variances are equal

against the alternative hypothesis

> H_a: At least one variance is different from the rest

```
The p-value is 0.3899432394522804
```

Since the p-value is large, we fail to reject the null hypothesis of homogeneity of variances.

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

Let's test whether the assumptions are satisfied or not

- The populations are normally distributed - Yes, the normality assumption is verified using the Shapiro-Wilk's test.
- Samples are independent simple random samples - Yes, we assume that the collected sample is a simple random sample.
- Population variances are equal - Yes, the homogeneity of variance assumption is verified using the Levene's test.

```
The p-value is 0.7158579926754841
```

Insight

As the p-value is much larger than the significance level, we fail to reject the null hypothesis. Hence, we do not have enough statistical significance to conclude that at least one means of BMI for women with no children, one child and two children is different from the rest at 5% significance level.

greatlearning
Learning for Life

Happy Learning !