Background:

Leveraging customer information is of paramount importance for most businesses. In the case of an insurance company, the attributes of customers like the ones mentioned below can be crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

Objective:

Statistical Analysis of Business Data. Explore the dataset and extract insights from the data.

1. Explore the dataset and extract insights using Exploratory Data Analysis.
2. Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?
3. Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.
4. Is the proportion of smokers significantly different across different regions?
5. Is the mean BMI of women with no children, one child, and two children the same? Explain your answer with statistical evidence. *Consider a significance level of 0.05 for all tests.

Data:

1.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).

1. Sex - This is the policy holder's gender, either male or female.
2. 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.
3. Children - This is an integer indicating the number of children/dependents covered by the insurance plan.
4. Smoker - This is yes or no depending on whether the insured regularly smokes tobacco.
5. Region - This is the beneficiary's place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest.
6. Charges - Individual medical costs billed to health insurance

Import the necessary libraries - pandas, numpy, seaborn, matplotlib.pyplot, scipy

In [2]:
```
#import the important packages
import warnings
warnings.filterwarnings('ignore')
import pandas as pd #library used for data manipulation and analysis
import numpy as np # library used for working with arrays.
import matplotlib.pyplot as plt # library for plots and visualisations
import seaborn as sns # library for visualisations
import random
%matplotlib inline

import scipy.stats as stats # this library contains a large number of probability distributions as well as a grow
```

In [3]:
```
!pip install scipy==1.6.1
import scipy
scipy.__version__
```

```
Requirement already satisfied: scipy==1.6.1 in d:\anaconda\lib\site-packages (1.6.1)
Requirement already satisfied: numpy>=1.16.5 in d:\anaconda\lib\site-packages (from scipy==1.6.1) (1.19.2)
```

Out[3]: '1.6.1'

Read in the dataset

In [4]:
```
data = pd.read_csv('AxisInsurance.csv') #reading the data
```

In [5]:
```
data.head() #first 5 rows of the data
```

Out[5]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230

2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

In [6]:
```
data.info() #checking the data types of each column
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

In [7]:
```
data.shape #checking the shape of the data
```

Out[7]:
```
(1338, 7)
```

In [8]:
```
data.isnull().sum()  #checking the total number of null values
```

Out[8]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Observations:

1. There are 1338 rows of data.
2. There are 7 variables in total.
3. There are no null values in any of the variables.

Converting Objects into Categorical Variables

In [9]:
```
data['sex'] = data['sex'].astype('category') #converting the data types into categorical types
data['smoker'] = data['smoker'].astype('category')
data['region'] = data['region'].astype('category')
data['children'] = data['children'].astype('category')
```

In [10]:
```
data.info() #checking if the data types have changed
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype   
---  ------    --------------  -----   
 0   age       1338 non-null   int64   
 1   sex       1338 non-null   category
 2   bmi       1338 non-null   float64 
 3   children  1338 non-null   category
 4   smoker    1338 non-null   category
 5   region    1338 non-null   category
 6   charges   1338 non-null   float64 
dtypes: category(4), float64(2), int64(1)
memory usage: 37.3 KB
```

In [11]:
```
print(data.describe()) #checking the statistics for the data
```

```
               age          bmi       charges
```

```
count  1338.000000  1338.000000   1338.000000
mean     39.207025    30.663397  13270.422265
std      14.049960     6.098187  12110.011237
min      18.000000    15.960000   1121.873900
25%      27.000000    26.296250   4740.287150
50%      39.000000    30.400000   9382.033000
75%      51.000000    34.693750  16639.912515
max      64.000000    53.130000  63770.428010
```

In [12]:
```
statistic = data.describe(include = 'category')
print(statistic) #checking all the statistics for the cateogrical data
```

```
         sex  children smoker     region
count   1338      1338   1338       1338
unique     2         6      2          4
top     male         0     no  southeast
freq     676       574   1064        364
```

Observations:

1. The charges column have a big spread in its data, with 75% being 1139 while the max being 63770
2. Both Age and BMI seem to be well distributed without much spread
3. The standard deviation for the charges column is huge.
4. There are 4 categorical variables.
5. Most occuring sex is male with a count of 676
6. There are 6 unique values for children with no children as the most recurring with 574 values.
7. Non-smokers are most recurring with 1064 values.
8. Southeast region is most reccuring with 364 values.

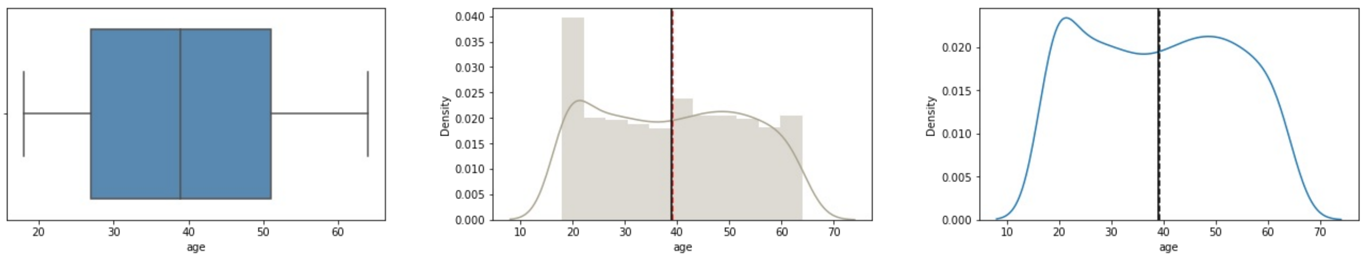
EDA

In [13]:
```
def histogram_boxplot(feature):
    """ Boxplot and histogram combined
    feature: 1-d feature array
    """
    figure, (ax_box2, ax_hist2, ax_hist3) = plt.subplots(
        nrows = 1, ncols=3,# Number of rows of the subplot grid= 2
        figsize  = (20,5)) # creating the 2 subplots
    figure.tight_layout(pad = 7)
    sns.boxplot(x = feature,ax=ax_box2, color = '#4B8BBE', orient = 'v') # boxplot will be created
    sns.distplot(feature, kde=True, ax=ax_hist2, color = '#a9a38f') # For histogram
    sns.distplot(feature, kde= True, ax=ax_hist3, hist = False) #Making an outline of the histogram
    ax_hist2.axvline(np.mean(feature), color='r', linestyle='--') # Add mean to the histogram
    ax_hist2.axvline(np.median(feature), color='black', linestyle='-') # Add median to the histogram
    ax_hist3.axvline(np.mean(feature), color = 'black', linestyle = '--') #Adding mean to second histogram
    ax_hist3.axvline(np.median(feature), color='black', linestyle='-') #Adding median to second histogram
```

Univariate Analysis

Observations on Age

In [14]:
```
histogram_boxplot(data['age']) #plotting using the function made above
```

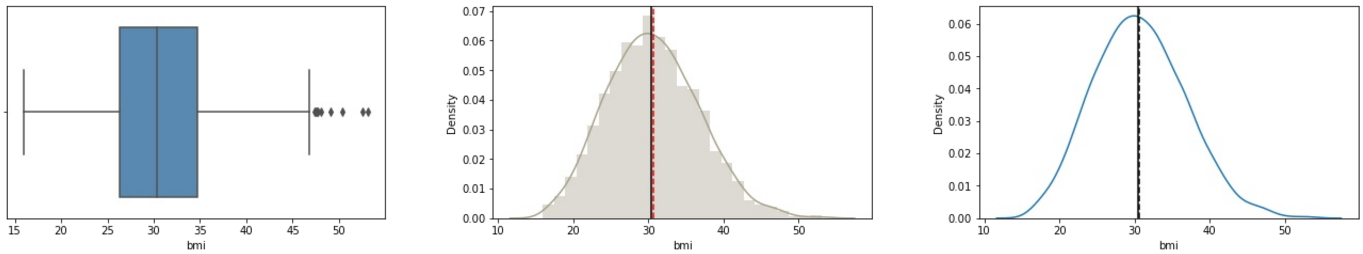


Observations:

1. There seems to be an even distribution of the data
2. The maximum value for Age is 64 while the minimum value is 18.
3. The mean of age is 39 with a standard deviation of 14.049.

Observations on BMI

In [15]:
```
histogram_boxplot(data['bmi'])
```

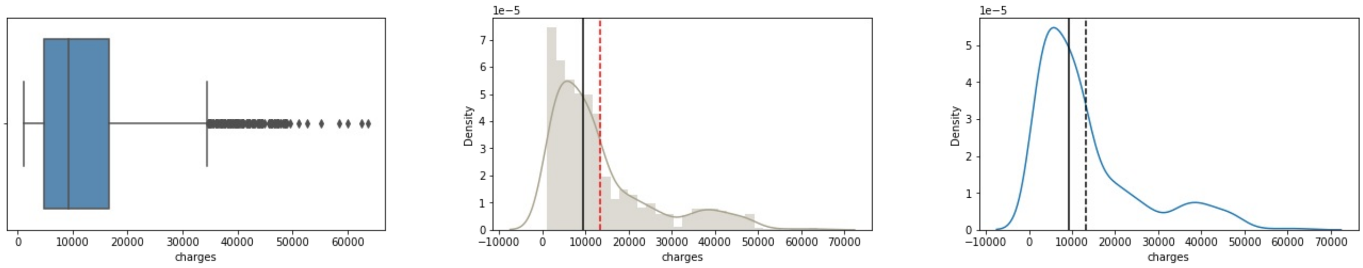


Observations:

1. There are a lot of outliers
2. The histograms show a bell-curve with a mean of 30 and standard deviation of 6.
3. The maximum vale is 53 while the minimum being 15.96.

Observations on Charges

In [16]:
```
histogram_boxplot(data['charges'])
```



Observations:

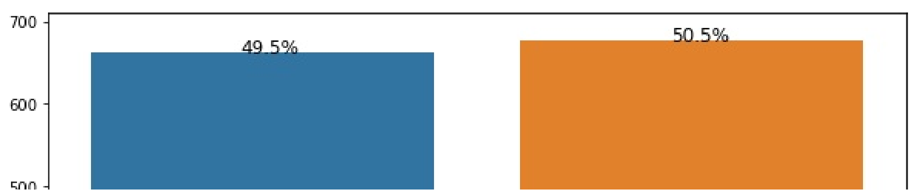
1. Charges have more outliers than any of the other variables.
2. The graph seems to be positively/right skewed.
3. The mean is 13270 and it has a high standard deviation of 12110

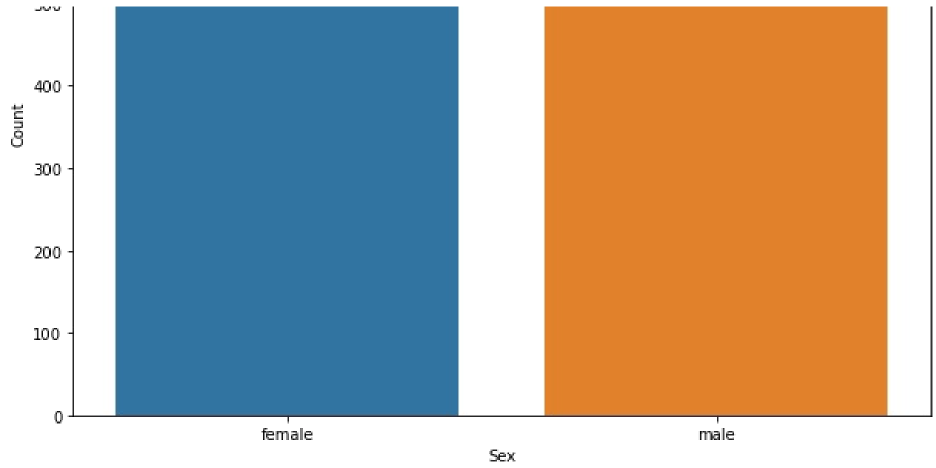
Categorical Variables

In [17]:
```
def bar_perc(plot, feature):
    '''
    plot
    feature: 1-d categorical feature array
    '''
    total = len(feature) # length of the column
    for p in ax.patches:
        percentage = '{:.1f}%'.format(100 * p.get_height()/total) # percentage of each class of the category
        x = p.get_x() + p.get_width() / 2 - 0.05 # width of the plot
        y = p.get_y() + p.get_height()           # hieght of the plot
        ax.annotate(percentage, (x, y), size = 12) # annotate the percantage
```

Observation on Sex

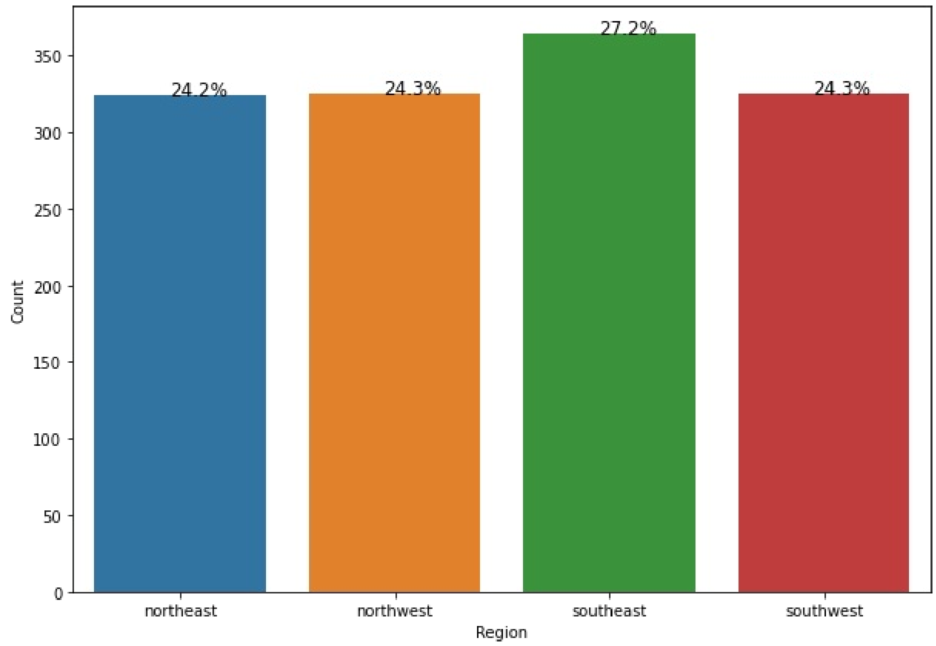
In [18]:
```
plt.figure(figsize=(10,7))
ax = sns.countplot(data['sex']) #count plot for Gender
plt.xlabel('Sex')
plt.ylabel('Count')
bar_perc(ax,data['sex'])
```





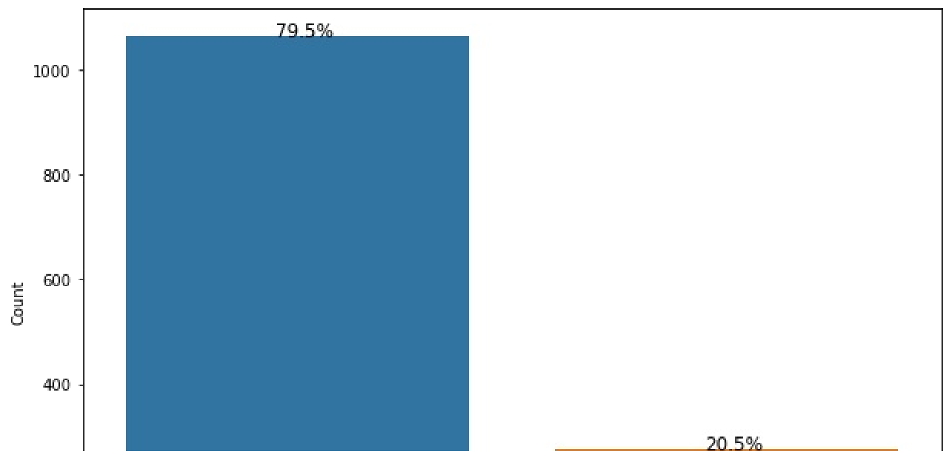
Observation on Region

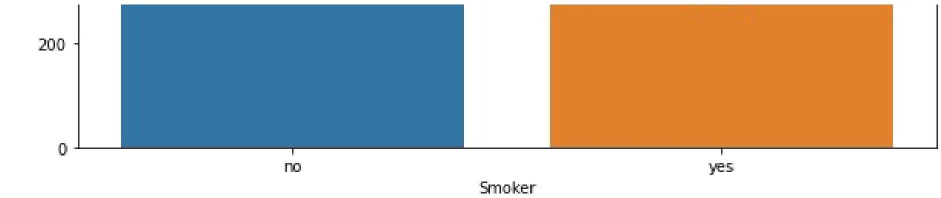
In [19]:
```
plt.figure(figsize=(10,7))
ax = sns.countplot(data['region']) #count plot for Gender
plt.xlabel('Region')
plt.ylabel('Count')
bar_perc(ax,data['region'])
```



Observation on Smoker

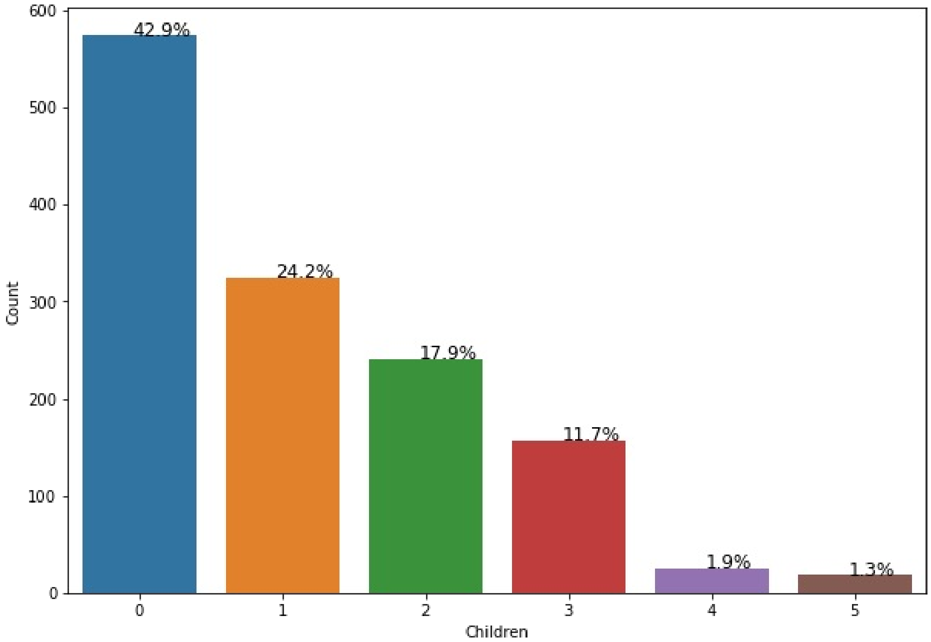
In [20]:
```
plt.figure(figsize=(10,7))
ax = sns.countplot(data['smoker']) #count plot for Gender
plt.xlabel('Smoker')
plt.ylabel('Count')
bar_perc(ax,data['smoker'])
```





Observation on Children

In [21]:
```
plt.figure(figsize=(10,7))
ax = sns.countplot(data['children']) #count plot for Gender
plt.xlabel('Children')
plt.ylabel('Count')
bar_perc(ax,data['children'])
```



Observations:

1. The female to male ratio is 49.5% to 50.5%.
2. Southeast is the most recurring region with 27.2% while the least recurring region is northeast with 24.2%.
3. There seems to be an even spread across each of the regions.
4. 79.5% of the individuals don't smoke while 20.5% do smoke.
5. Most individuals have no children with 42.9% while, some individuals have 5 children with 1.3%

Bivariate Analysis

Correlation and Covariance

In [22]:
```
data.corr() #correlation of data
```

Out[22]:

	age	bmi	charges
age	1.000000	0.109272	0.299008
bmi	0.109272	1.000000	0.198341
charges	0.299008	0.198341	1.000000

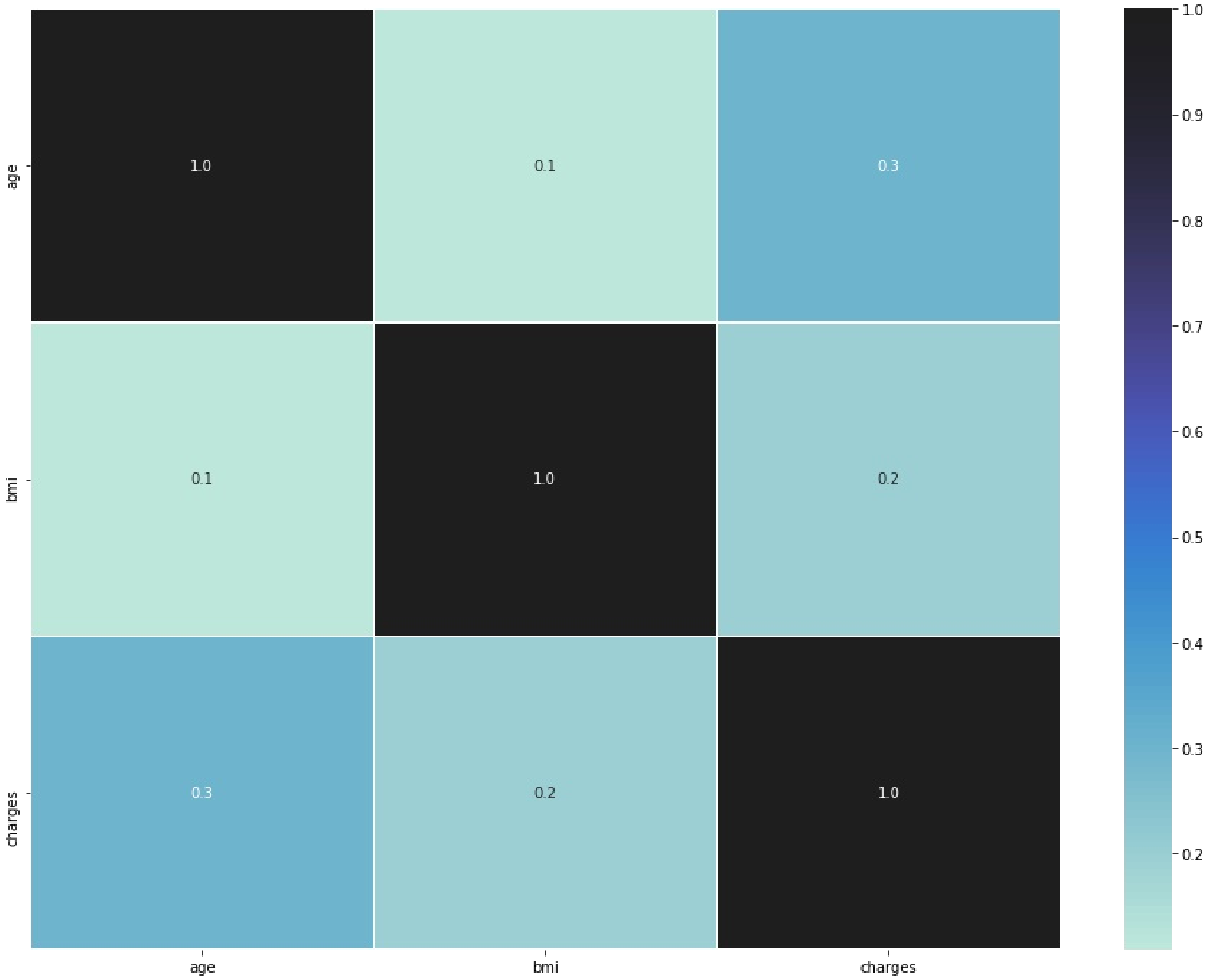
In [23]:
```
data.cov() #covariance of data
```

Out[23]:

	age	bmi	charges
age	197.401387	9.362337	5.087480e+04
bmi	9.362337	37.187884	1.464730e+04

charges	50874.802298	14647.304426	1.466524e+08

In [24]:
```python
plt.figure(figsize=(16,12))
sns.heatmap(data.corr(), annot=True, linewidths=.5, fmt= '.1f', center = 1 )  # heatmap
plt.show()
```

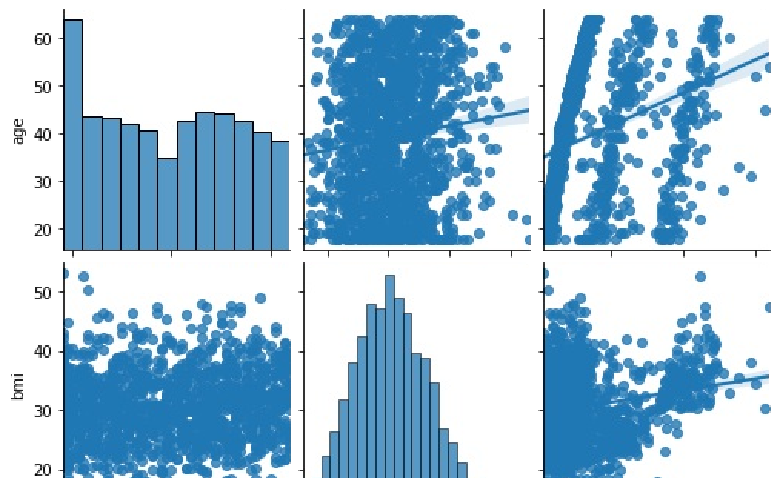


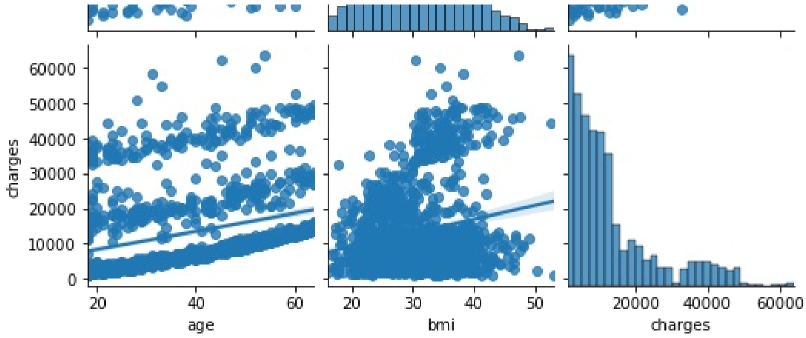
Observation:

1. As indicated in the correlation statistic most of the variables have very low correlation amongst each other.
2. The two variables with highest correlation are age and charges with 0.3 correlation.
3. The lowest correlation is between age and bmi with 0.10 correlation.
4. Most of these variables have no connection between each other as indicated by the heatmap and correlation statistic.
5. Correlation does not imply casuation.

In [25]:
```python
plt.figure(figsize = (20,20))
sns.pairplot(data = data, kind = 'reg')
plt.show()
```

```
<Figure size 1440x1440 with 0 Axes>
```





Observation:

1. There is a huge spread of data in age, bmi acharges.
2. Every single variables shows a positive correlation towards each other.
3. Correlation between each variables are very low as indicated by gentle slope.
4. BMI against itself indicates a bell-curve.
5. As there seems to be no relationship between each of the variables as indicated by the scatter plots and correlation statistics there need not be any further bivariate analysis.

Question #1

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

Null and alternative hypothesis

We will test the null hypothesis

> $H_0 : \mu1 = \mu2$

```
μ1 - being the mean of the medical claim of smokers.
μ2 - being the mean of the medical claim of non-smokers.
```

against the alternate hypothesis

> $H_a : \mu1 > \mu2$

Finding the appropriate data

In [26]:

```python
smoker_data = data[data['smoker'] == 'yes'] #only taking in the data of smokers
non_smoker_data = data[data['smoker'] == 'no'] #only taking in the data of non-smokers
print(smoker_data.head())
print(non_smoker_data.head())
```

```
    age     sex    bmi  children smoker     region      charges
0    19  female  27.90         0    yes  southwest  16884.9240
11   62  female  26.29         0    yes  southeast  27808.7251
14   27    male  42.13         0    yes  southeast  39611.7577
19   30    male  35.30         0    yes  southwest  36837.4670
23   34  female  31.92         1    yes  northeast  37701.8768
   age     sex     bmi  children smoker     region       charges
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
5   31  female  25.740         0     no  southeast   3756.62160
```

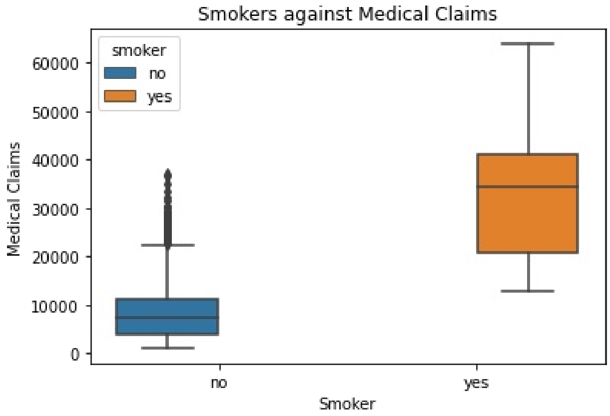
In [27]:

```python
print('The mean medical claim for Smokers is ' + str(round(smoker_data['charges'].mean(), 2)))
print('The mean medical claim for High Non-smokers group is ' + str(round(non_smoker_data['charges'].mean(), 2)))
print('The standard deviation of medical claim score for Smokers is ' + str(round(smoker_data['charges'].std(), 2
print('The standard deviation of medical claim for Non-smokers group is ' + str(round(non_smoker_data['charges'].
```

```
The mean medical claim for Smokers is 32050.23
The mean medical claim for High Non-smokers group is 8434.27
The standard deviation of medical claim score for Smokers is 11541.55
The standard deviation of medical claim for Non-smokers group is 5993.78
```

In [28]:
```
b = sns.boxplot(x= "smoker", y = 'charges' , data = data, hue = 'smoker') #boxplot
b.set_title('Smokers against Medical Claims')
plt.ylabel('Medical Claims')
plt.xlabel('Smoker')
```

Out[28]:
```
Text(0.5, 0, 'Smoker')
```



Observations:

1. There are many outliers for medical claims of non-smokers.
2. The mean for medical claims of smokers appear to be closer to the 75% percentile.
3. The mean of medical claims of smokers is much higher than the medical claim of non-smokers.

Assumptions:

1. Continuous data - Yes, the medical claims are measured on a continuous scale.
2. Independent populations - As we are taking random samples for two different groups, the two samples are from two independent populations.
3. Unequal population standard deviations - As the sample standard deviations are different, the population standard deviations may be assumed to be different.
4. Random sampling from the population - Yes, we are informed that the collected sample a simple random sample.

Two sample independent T-test

In [29]:
```
#import the required functions
from scipy.stats import ttest_ind, norm
test_stat, p_value = ttest_ind(smoker_data['charges'], non_smoker_data['charges'], equal_var = False, alternative
print('The p-value is ', p_value)
```

```
The p-value is  2.94473222335849e-103
```

Insight

As the p-value 2.944e-103 is significantly lower than the level of significance, we can reject the null hypothesis. We have enough evidence to state that the mean of medical claim of smokers is much greater than of those that don't smoke at a 0.05 level of significance.

--

Question #2

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

Null and alternative hypothesis

We will test the null hypothesis

> $H_0 : \mu 1 = \mu 2$

```
μ1 - being the mean of the BMI of females.
μ2 - being the mean of the BMI of Males.
```

against the alternate hypothesis

> $H_a : \mu1 \neq \mu2$

Finding the appropriate data

In [30]:
```python
bmi_males = data[data['sex'] == 'male' ]
bmi_females = data[data['sex'] == 'female']
print(bmi_males.head())
print(bmi_females.head())
```

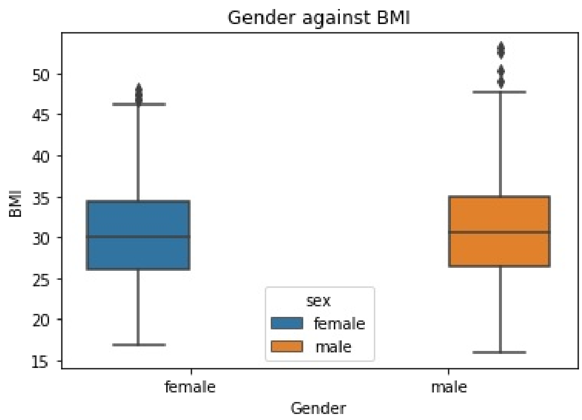
```
   age   sex     bmi  children smoker     region      charges
1   18  male  33.770         1     no  southeast   1725.55230
2   28  male  33.000         3     no  southeast   4449.46200
3   33  male  22.705         0     no  northwest  21984.47061
4   32  male  28.880         0     no  northwest   3866.85520
8   37  male  29.830         2     no  northeast   6406.41070
   age     sex    bmi  children smoker     region      charges
0   19  female  27.90         0    yes  southwest  16884.92400
5   31  female  25.74         0     no  southeast   3756.62160
6   46  female  33.44         1     no  southeast   8240.58960
7   37  female  27.74         3     no  northwest   7281.50560
9   60  female  25.84         0     no  northwest  28923.13692
```

In [31]:
```python
print('The mean of BMI for females is ' + str(round(bmi_females['bmi'].mean(), 2)))
print('The mean of BMI for males is ' + str(round(bmi_males['bmi'].mean(), 2)))
print('The standard deviation of BMI for females is ' + str(round(bmi_females['bmi'].std(), 2)))
print('The standard deviation of BMI for males is ' + str(round(bmi_males['bmi'].std(), 2)))
```

```
The mean of BMI for females is 30.38
The mean of BMI for males is 30.94
The standard deviation of BMI for females is 6.05
The standard deviation of BMI for males is 6.14
```

In [32]:
```python
plot = sns.boxplot(x = 'sex', y = 'bmi', data = data, hue = 'sex')
plot.set_title('Gender against BMI')
plt.ylabel('BMI')
plt.xlabel('Gender')
```

Out[32]: Text(0.5, 0, 'Gender')



Observation:

1. There are outliers for both female and male BMI values.
2. The mean for both female and male values seem to be close.

Assumptions

1. Continuous data - Yes, the BMI values are measured on a continuous scale.
2. Independent populations - As we are taking random samples for two different groups, the two samples are from two independent populations.
3. Equal population standard deviations - As the sample standard deviations are different, the population standard deviations may be

assumed to be different.

4. Random sampling from the population - Yes, we are informed that the collected sample a simple random sample.

Two sample independent t-test with equal standard deviations

In [33]:
```
#import the required functions
from scipy.stats import ttest_ind

# find the p-value
test_stat, p_value = ttest_ind(bmi_females['bmi'], bmi_males['bmi'], equal_var = True, alternative = 'two-sided')
print('The p-value is ' + str(p_value))
```

```
The p-value is 0.08997637178984934
```

Insights

As the p-value is 0.089 which is higher than the level of significance, we fail to reject the null hypothesis. Hence, we have enough evidence to prove that the mean bmi of females is equal to that of males at 0.05 level of significance.

Question #3

Is the proportion of smokers significantly different across different regions?

Let's write the null and alternative hypothesis

We will test the null hypothesis

> H_0 : Smoking habit is independent of the region.

against the alternate hypothesis

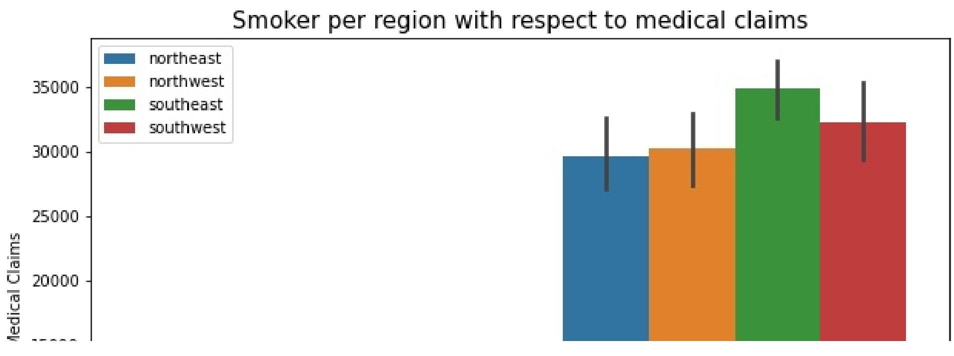
> H_a : Smoking habit is not independent of the region.

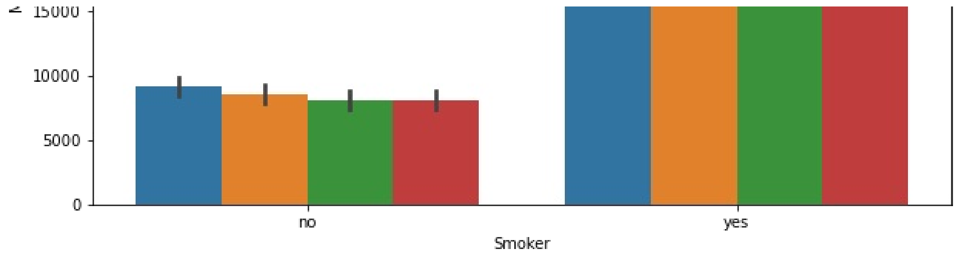
Finding the appropriate data

In [46]:
```
data_crosstab = pd.crosstab(columns = data['region'], index = data['smoker'], margins = True) #  Making Contigenc
print(data_crosstab)
```

```
region  northeast  northwest  southeast  southwest   All
smoker
no            257        267        273        267  1064
yes            67         58         91         58   274
All           324        325        364        325  1338
```

In [35]:
```
# draw the barplot for visualization
fig, ax = plt.subplots(figsize = (10,6))
ax = sns.barplot( x = 'smoker', y = 'charges' , data = data, hue = 'region') #barplots
plt.legend()
plt.xlabel(xlabel = 'Smoker')
plt.ylabel(ylabel = 'Medical Claims')
ax.set_title("Smoker per region with respect to medical claims", fontsize=15)
plt.show()
```





Observation:

1. Based on the grpah, medical claims is higher for non smokers in the northeast region followed by northwest, southeast and southwest respectivily.
2. Medical claims for smokers is highest in Southeast region, and lowest in northeast region.

Assumptions:

1. Categorical variables - Yes
2. 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.
3. Random sampling from the population - Yes, we are informed that the collected sample is a simple random sample.

Chi-squared test

In [36]:
```python
chi2, p_value, dof, expected = stats.chi2_contingency(observed =  data_crosstab) #the appropriate function
print(f'The p-value is {p_value}')
```

```
The p-value is 0.5000675325877666
```

Insight

As the p-value 0.5 is greater than the significance level, we fail to reject the null hypothesis. Hence, we do not have enough evidence to conclude that smoking habit is not independent based on the region at a 0.05 level of significance.

Question #4

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

Null and alternative hypothesis

Let μ_1, μ_2, μ_3 be the means of BMI for females with 0, 1 and 2 children respectively.

We will test the null hypothesis

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

against the alternative hypothesis

> H_a : At least 1 mean of BMI from females is differnt from the others.

In [37]:
```python
new_data =  pd.read_csv('AxisInsurance.csv')    #making a new data set
children_data = new_data[new_data['children'] == 0] #reading all the data of women and male containing 0 childrer
children_data1 = new_data[new_data['children'] == 1] #reading all the data of women and male containing 1 childre
children_data2 = new_data[new_data['children'] == 2]
children = children_data.append(children_data1) #appending the data to a new data frame
all_children_data = children.append(children_data2) #appending 0, 1 and 2 children to a new data frame
print(all_children_data)
```

```
      age     sex     bmi  children smoker     region      charges
0      19  female  27.900         0    yes  southwest  16884.92400
3      33    male  22.705         0     no  northwest  21984.47061
```

```
4       32    male  28.880         0     no  northwest   3866.85520
5       31  female  25.740         0     no  southeast   3756.62160
9       60  female  25.840         0     no  northwest  28923.13692
...    ...     ...     ...       ...    ...        ...          ...
1319    39  female  26.315         2     no  northwest   7201.70085
1323    42  female  40.370         2    yes  southeast  43896.37630
1328    23  female  24.225         2     no  northeast  22395.74424
1329    52    male  38.600         2     no  southwest  10325.20600
1330    57  female  25.740         2     no  southeast  12629.16560

[1138 rows x 7 columns]
```

In [38]:

```python
female_data = all_children_data[all_children_data['sex'] != 'male'] #dropping all the male values from the sex cc
print(female_data)
```

```
      age     sex     bmi  children smoker     region      charges
0      19  female  27.900         0    yes  southwest  16884.92400
5      31  female  25.740         0     no  southeast   3756.62160
9      60  female  25.840         0     no  northwest  28923.13692
11     62  female  26.290         0    yes  southeast  27808.72510
13     56  female  39.820         0     no  southeast  11090.71780
...   ...     ...     ...       ...    ...        ...          ...
1313   19  female  34.700         2    yes  southwest  36397.57600
1319   39  female  26.315         2     no  northwest   7201.70085
1323   42  female  40.370         2    yes  southeast  43896.37630
1328   23  female  24.225         2     no  northeast  22395.74424
1330   57  female  25.740         2     no  southeast  12629.16560

[566 rows x 7 columns]
```

In [39]:

```python
female_data['children'].value_counts() #counting the number of unique values for the children value
```

Out[39]:

```
0    289
1    158
2    119
Name: children, dtype: int64
```

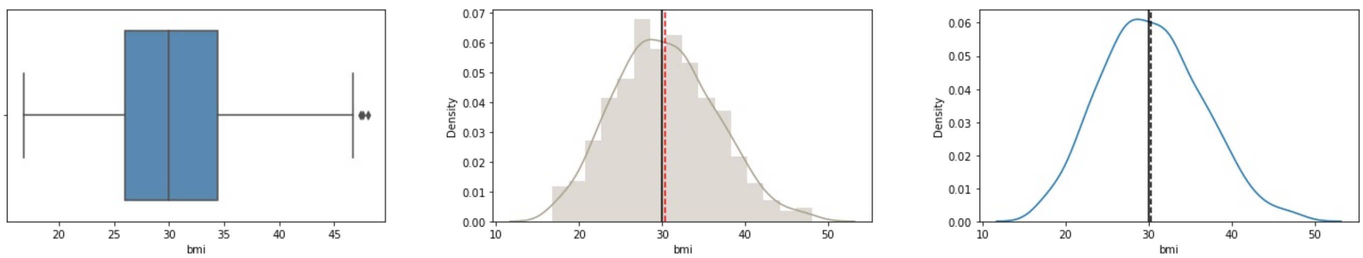
In [40]:

```python
print('The mean BMI for women is ' + str(round(female_data['bmi'].mean(), 2)))
print('The mean BMI for women with no children is ' + str(round(children_data['bmi'].mean(), 2)))
print('The mean BMI for women with one children is ' + str(round(children_data1['bmi'].mean(), 2)))
print('The mean BMI for women with two children is ' + str(round(children_data2['bmi'].mean(), 2)))
```

```
The mean BMI for women is 30.34
The mean BMI for women with no children is 30.55
The mean BMI for women with one children is 30.62
The mean BMI for women with two children is 30.98
```

In [41]:

```python
histogram_boxplot(female_data['bmi'])
```



Observation:

1. The mean of BMI for females appear to be around 30
2. There are few outliers
3. The histogram seem to represent a bell-shaped curve.

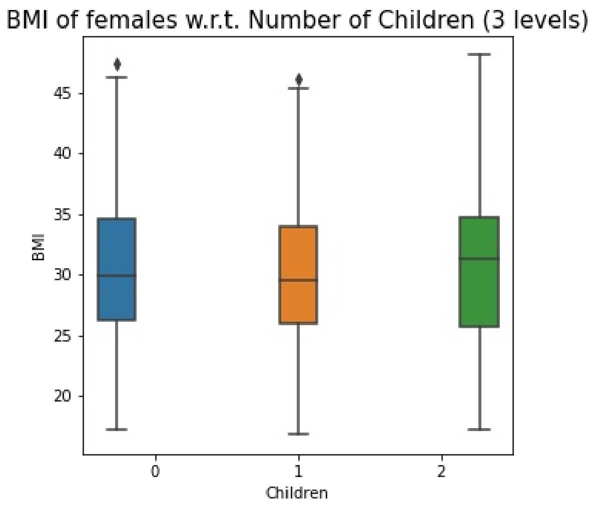
In [42]:

```python
# mean of carbon emission at different levels of the fuel_type factor
print(female_data.groupby("children")["bmi"].mean())

# draw the boxplot for visualization
fig, ax = plt.subplots(figsize = (5,5))
a = sns.boxplot(x= "children", y = 'bmi' , data = female_data, hue = 'children')
plt.legend().remove()
```

```
plt.xlabel(xlabel = 'Children')
plt.ylabel(ylabel = 'BMI')
a.set_title("BMI of females w.r.t. Number of Children (3 levels)", fontsize=15)
plt.show()
```

```
children
0    30.361522
1    30.052658
2    30.649790
Name: bmi, dtype: float64
```



Observation:

1. There are few outliers for BMI of females with 0 and 1 children.
2. The mean for BMI of 0, 1 and 2 children seem to be very close.

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 varaince, Levene test is applied to the response variable.

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

In [43]:
```
# find the p-value
w, p_value = stats.shapiro(female_data['bmi'])
print('The p-value is', p_value)
```

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

Insight

Since the p_value 0.10 is less than the level of significance, we reject the null hypothesis. We don't have enough evidence to show that the BMI follows a 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

In [44]:
```
statistic, p_value = stats.levene( female_data['bmi'][female_data['children']== 0],
                                   female_data['bmi'][female_data['children']== 1],
                                   female_data['bmi'][female_data['children']== 2])
# find the p-value
print('The p-value is', p_value)
```

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

Insight

Since the p-value 0.389, is greater than the significance value, we fail to rejects the null hypothesis of homogeneity of variances.

Assumptions:

1. Though the sample data does not follow the normal distribution, you can still use one-way ANOVA as it is quite robust against the normality assumption. It tolerates violations of the normality assumption rather well.
2. Samples are independent simple random samples - Yes, we are informed that the collected sample is a simple random sample.
3. Population variances are equal - Yes, the homogeneity of variance assumption is verified using the Levene's test.

Finding P-value using One-way ANOVA F-test

In [45]:
```
test_stat, p_value = stats.f_oneway(female_data.loc[female_data['children'] == 0, 'bmi'],
                              female_data.loc[female_data['children'] == 1, 'bmi'],
                              female_data.loc[female_data['children'] == 2, 'bmi'])
print('The p-value is ' + str(p_value))
print(test_stat)
```

```
The p-value is 0.7158579926754841
0.3344720147757968
```

Insight

Since the p-value 0.715, is higher than the level of significance, we fail to reject the null hypothesis. We have enough data to confirm that the BMI of females with no children, 1 children and 2 children are the same at a 0.05 level of significance.

Conclusion and Recommendation

Conclusion

1. As indicated in the correlation statistic most of the variables have very low correlation amongst each other.
2. The two variables with highest correlation are age and charges with 0.3 correlation. The lowest correlation is between age and BMI with 0.10 correlation. Most of these variables have no connection between each other as indicated by the heatmap and correlation statistic.
3. As there seems to be no relationship between each of the variables as indicated by the scatter plots and correlation statistics there need not be any further bivariate analysis.
4. As the p-value 2.944 x 10-103 is significantly lower than the level of significance, we can reject the null hypothesis. We have enough evidence to state that the mean of medical claim of smokers is much greater than of those that don't smoke at a 0.05 level of significance.
5. As the p-value is 0.089 which is higher than the level of significance, we fail to reject the null hypothesis. Hence, we have enough evidence to prove that the mean BMI of females is equal to that of males at 0.05 level of significance.
6. As the p-value 0.5 is greater than the significance level, we fail to reject the null hypothesis. Hence, we do not have enough evidence to conclude that smoking habit is not independent based on the region at a 0.05 level of significance.
7. Since the p-value 0.715, is higher than the level of significance, we fail to reject the null hypothesis. We have enough data to confirm that the BMI of females with no children, 1 children and 2 children are the same at a 0.05 level of significance.

Recommendation

1. Most of the variables do not correlate well with each other, so to conduct a better study and to make better business decisions it will be better to find variables that correlate well with each other based on the bivariate analysis.
2. Medical claim made by smokers is much greater than medical claim made by non-smokers as per the two-sample independent t-tests.
3. Mean BMI of females is equal to the mean BMI of males as per the Two-sample independent t-test.
4. Not enough evidence to conclude that the smoking habit is not independent of the region based on Chi-squared test.
5. Based on the One-Way ANOVA test, BMI of females with no children, 1 children and 2 children are the same.