BUSINESS CASE: INSURANCE COVERAGE AND COST-Customer Profiling and Hypothesis Testing

About data- This dataset contains information on the relationship between personal attributes (age, gender, BMI, family size, smoking habits), geographic factors, and their impact on medical insurance charges. The columns include-

Age: The insured person's age.

Sex: Gender (male or female) of the insured.

BMI (Body Mass Index): A measure of body fat based on height and weight.

Children: The number of dependents covered.

Smoker: Whether the insured is a smoker (yes or no).

Region: The geographic area of coverage.

Charges: The medical insurance costs incurred by the insured person.

Problem Statement: Analyse the features that influence insurance cost. Do customer profiling based on different features like age, sex, BMI, number of children, smoking status or region.

Analysing basic metric and understanding the data:

Loading important libraries:

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
```

Uploading the data:

```
!gdown 1g50gpXOrA6b7J6iQVekb20jbCIEgyF5k
```

```
Downloading...
From: https://drive.google.com/uc?id=1g50gpXOrA6b7J6iQVekb20jbCIEgyF5k
To: /content/insurance.csv
100% 55.6k/55.6k [00:00<00:00, 58.0MB/s]
```

```
df = pd.read_csv('insurance.csv')
df.head()
```

	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

Exploring the data:

```
df.shape
```

```
(1338, 7)
```

```
df.columns
```

```
Index(['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges'], dtype='object')
```

```
df.isna().sum()
```

	0
age	0
sex	0
bmi	0
children	0
smoker	0
region	0
charges	0

dtype: int64

Note: It is found that the dataset doesn't have any null values. So, we now move towards statistical description of the data:

```
[ ] df[['age', 'bmi', 'charges']].describe()
```

	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

Insights:

1. Minimum age recorded is 18 years and maximum is 64 years.
2. Minimum BMI is 16 and maximum is 53. Mean value is 30. Hence BMI has outlier values. That is, some people have BMI that is outside the normal range.
3. BMI is normally distributed as most of the values are clustered near its mean value.
4. Minimum charge for an insurance plan is 1122 and maximum is 63770. Clearly charges have outlier values. That means, some people have insurance plans which are very costly, owing to their high BMI, age and diseases.

```
[ ] # let's check how many people are healthy (normal bmi)
    df[(df['bmi'] >= 18.5) & (df['bmi'] < 24.9)]
```

	age	sex	bmi	children	smoker	region	charges
3	33	male	22.705	0	no	northwest	21984.47061
15	19	male	24.600	1	no	southwest	1837.23700
17	23	male	23.845	0	no	northeast	2395.17155
26	63	female	23.085	0	no	northeast	14451.83515
35	19	male	20.425	0	no	northwest	1625.43375
...	...	...	...	...	...	...	...
1304	42	male	24.605	2	yes	northeast	21259.37795
1306	29	female	21.850	0	yes	northeast	16115.30450
1314	30	female	23.655	3	yes	northwest	18765.87545
1316	19	female	20.600	0	no	southwest	1731.67700
1328	23	female	24.225	2	no	northeast	22395.74424

222 rows × 7 columns

```
[ ] 222 * 100/1338
```

```
16.591928251121075
```

Insights: There are only 16.6% people who lie in the normal range of BMI. Hence, we will see how overweight/ underweight customers will be profiled.

Handling Outliers: Since there are no missing values in the dataset, our attention will now shift towards detecting and managing outliers, if any. This process will involve scrutinizing the data for any unusual or extreme observations that may impact the robustness of our analysis.

Outliers in BMI column:

```
[ ] # outliers in bmi column, through IQR method.
    q1 = df['bmi'].quantile(0.25)
    q3 = df['bmi'].quantile(0.75)
    iqr = q3-q1

    lower_bound = q1-1.5*iqr
    upper_bound = q3+1.5*iqr

    outliers = df[(df['bmi'] < lower_bound) | (df['bmi'] > upper_bound)]

    print(" Number of outliers-> ", len(outliers))
    print("percentage of outliers-> ", len(outliers)*100 / df.shape[0])
```

```
 Number of outliers->  9
percentage of outliers->  0.672645739910314
```

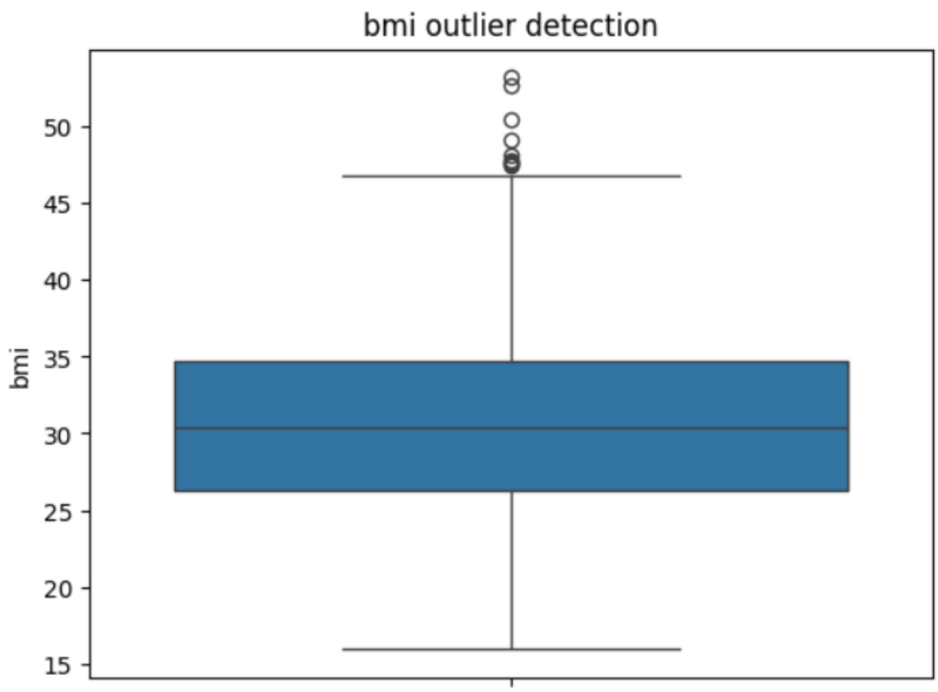
Outliers in Charges column:

```python
# outliers in charges column, through IQR method.
q1 = df['charges'].quantile(0.25)
q3 = df['charges'].quantile(0.75)
iqr = q3-q1

lower_bound = q1-1.5*iqr
upper_bound = q3+1.5*iqr

outliers = df[(df['charges'] < lower_bound) | (df['charges'] > upper_bound)]

print(" Number of outliers-> ", len(outliers))
print("percentage of outliers-> ", len(outliers)*100 / df.shape[0])
```

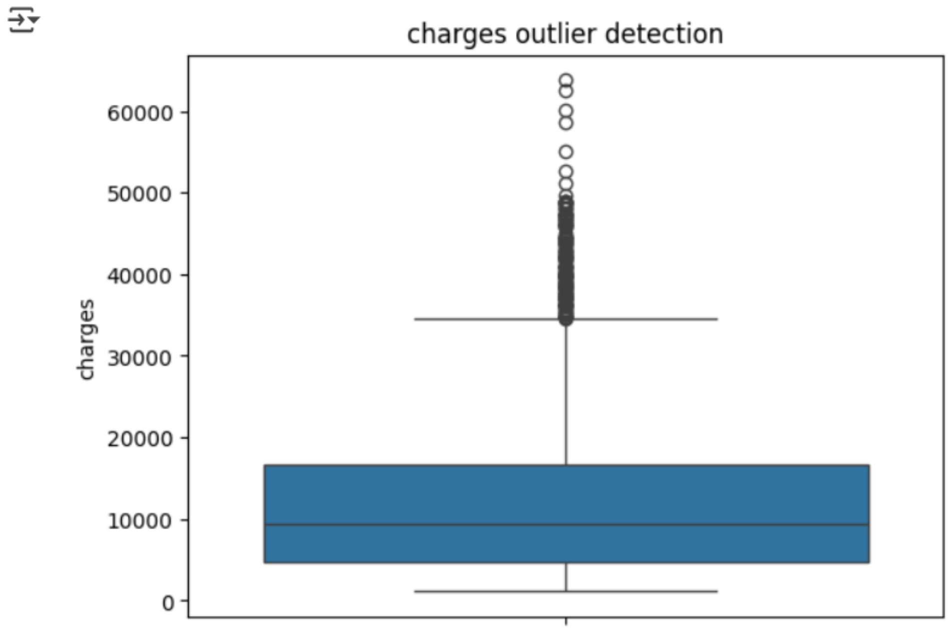
```
 Number of outliers->  139
percentage of outliers->  10.388639760837071
```

Outliers through boxplot:

```python
sns.boxplot(data = df, y = 'bmi')
plt.title('bmi outlier detection')
plt.show()
```



```
sns.boxplot(data = df, y = 'charges')
plt.title('charges outlier detection')
plt.show()
```



Insights:

1. BMI has 9 outliers (0.67%), this may be due to rare health conditions.

2. Charges have a large number of outliers (10.3%) — this often happens with cost-related data, which is typically right-skewed. A small number of patients with extremely high medical costs may drive up the outliers.

Handling Outliers:

1. Since the outlier data is good for business perspective for the insurance company, it would be wise to keep the data for the targeted customers.
2. Since medical expenses are naturally right-skewed — a small number of patients account for very high costs, if removed those, it will hide reality.
3. Extreme values in insurance cost might also help in indicating fraudulent claims or over-utilization. hence it is wise to keep them for the company business.

We will form a new column “BMI category”.

```
[ ] df['bmi category'] = pd.cut(df['bmi'], bins=[0, 25, 30, 35, 100],
                                labels=['Normal', 'Overweight', 'Obese I', 'Obese II+'])
    df.head(5)
```

	age	sex	bmi	children	smoker	region	charges	bmi category
0	19	female	27.900	0	yes	southwest	16884.92400	Overweight
1	18	male	33.770	1	no	southeast	1725.55230	Obese I
2	28	male	33.000	3	no	southeast	4449.46200	Obese I
3	33	male	22.705	0	no	northwest	21984.47061	Normal
4	32	male	28.880	0	no	northwest	3866.85520	Overweight

Categorical Analysis:

```
[ ]
    cat_cols = ['sex', 'children', 'smoker', 'region']
    for i in cat_cols:
      print(df[i].value_counts()*100/1338)
      print()
```

```
sex
male      50.523169
female    49.476831
Name: count, dtype: float64

children
0    42.899851
1    24.215247
2    17.937220
3    11.733931
4     1.868460
5     1.345291
Name: count, dtype: float64

smoker
no     79.521674
yes    20.478326
Name: count, dtype: float64

region
southeast    27.204783
southwest    24.289985
northwest    24.289985
northeast    24.215247
Name: count, dtype: float64
```

Insights:

1. almost equal number of male and female, with slightly more males
2. 42% people have no children
3. almost 80% are non-smokers, but smokers may be driving high charges. risk adjusted pricing for smokers is essential.
4. fairly balanced regional spread, with slightly more customers in the southeast.

BIVARIATE ANALYSIS:

Grouping "age" data into bins:

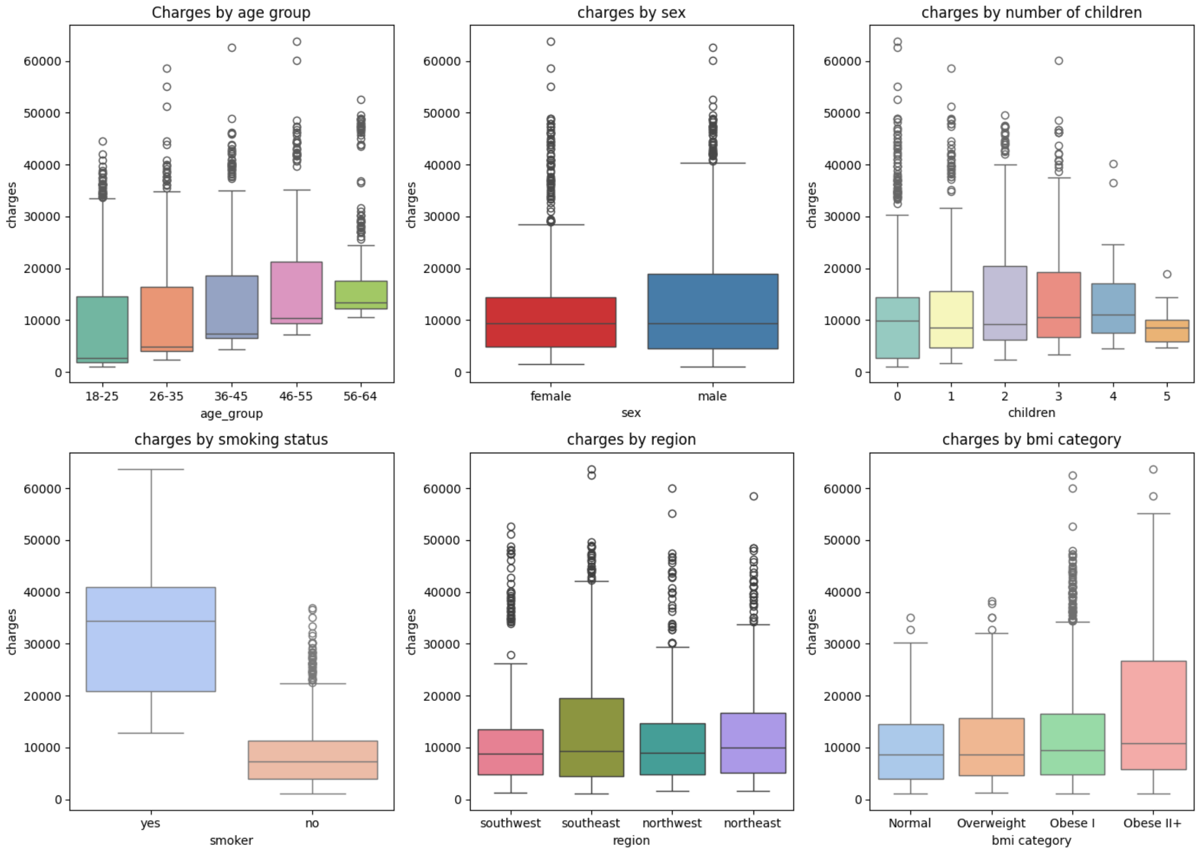
```
bins = [17, 25, 35, 45, 55, 65]
labels = ['18-25', '26-35', '36-45', '46-55', '56-64']

df['age_group'] = pd.cut(df['age'], bins=bins, labels=labels, right=True)
df.head()
```

	age	sex	bmi	children	smoker	region	charges	bmi category	age_group
0	19	female	27.900	0	yes	southwest	16884.92400	Overweight	18-25
1	18	male	33.770	1	no	southeast	1725.55230	Obese I	18-25
2	28	male	33.000	3	no	southeast	4449.46200	Obese I	26-35
3	33	male	22.705	0	no	northwest	21984.47061	Normal	26-35
4	32	male	28.880	0	no	northwest	3866.85520	Overweight	26-35

Next, exploring relationships between variables (e.g., smoker vs. charges, BMI vs. charges, etc.) using boxplots and heatmaps.

```
fig, axes = plt.subplots(2, 3, figsize = (14, 10))

#charges by age
sns.boxplot(data = df, x = 'age_group', y = 'charges', hue = 'age_group', ax = axes[0, 0], palette= 'Set2', legend = False)
axes[0, 0].set_title('Charges by age group')

#charges by sex
sns.boxplot(data = df, x = 'sex', y = 'charges', hue = 'sex', ax = axes[0, 1], palette= 'Set1', legend= False)
axes[0, 1].set_title('charges by sex')

#charges by number of children
sns.boxplot(data = df, x = 'children', y = 'charges', hue = 'children', ax = axes[0, 2], palette= 'Set3', legend = False)
axes[0, 2].set_title('charges by number of children')

#charges by smoking status
sns.boxplot(data = df, x = 'smoker', y = 'charges', hue = 'smoker', ax = axes[1, 0], palette= 'coolwarm', legend = False)
axes[1, 0].set_title('charges by smoking status')

#charges by region
sns.boxplot(data = df, x = 'region', y = 'charges', hue = 'region', ax = axes[1,1], palette= 'husl', legend= False)
axes[1,1].set_title('charges by region')

#charges by bmi category
sns.boxplot(data= df, x = 'bmi category', y = 'charges', hue= 'bmi category', ax= axes[1, 2], palette= 'pastel', legend= False)
axes[1, 2].set_title('charges by bmi category')

plt.tight_layout()
plt.show()
```

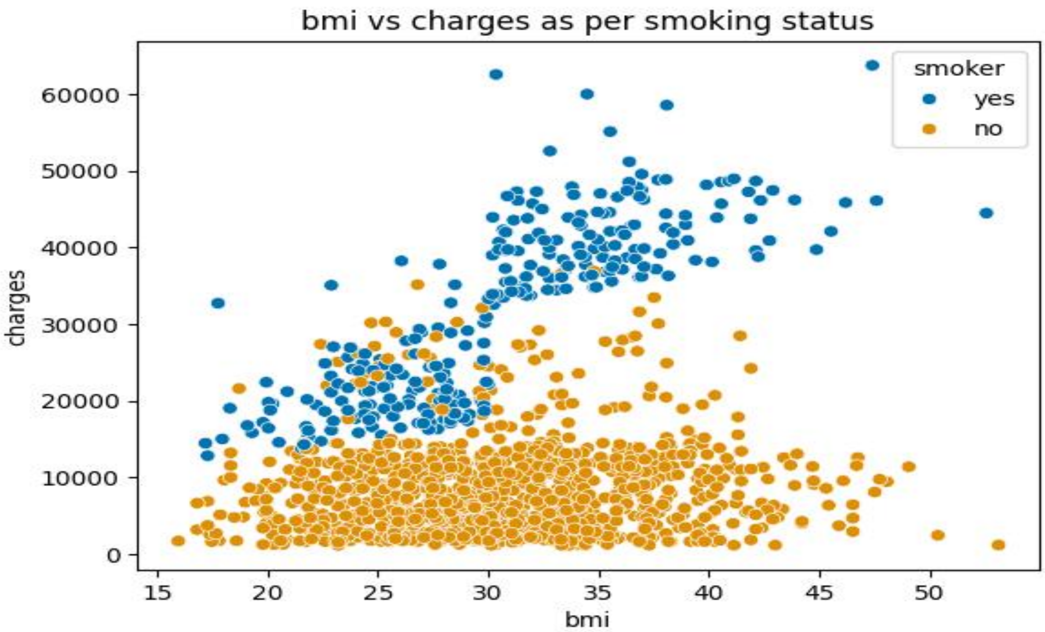
Output:



BMI vs Charges as per smoking status:

```
[ ] # bmi vs charges as per smoking status
    sns.scatterplot(data= df, x = 'bmi', y = 'charges', hue = 'smoker', palette = 'colorblind')
    plt.title('bmi vs charges as per smoking status')
    plt.show()
```

Output:



INSIGHTS:

Charges vs Age:

a. Each group has outliers shows that some individuals in each group have unusually high medical cost.

b. older groups (46–55, 56–64) have higher lower-quartile values — suggesting even the less expensive cases cost more.

Charges by sex:

a. Males have wider spread than women. it means male members have variable insurance cost than females.

b. Female members have narrow spread. It means their charges are more compact and closer to the median.

c. Outliers exist in both females and male. which means high-cost individual exist in both groups.

Charges vs number of children:

a. children alone don't affect insurance charges.

b. Parents who have 0 or 1 children have lower charges.

c. 2 or 3 children's parents have higher cost

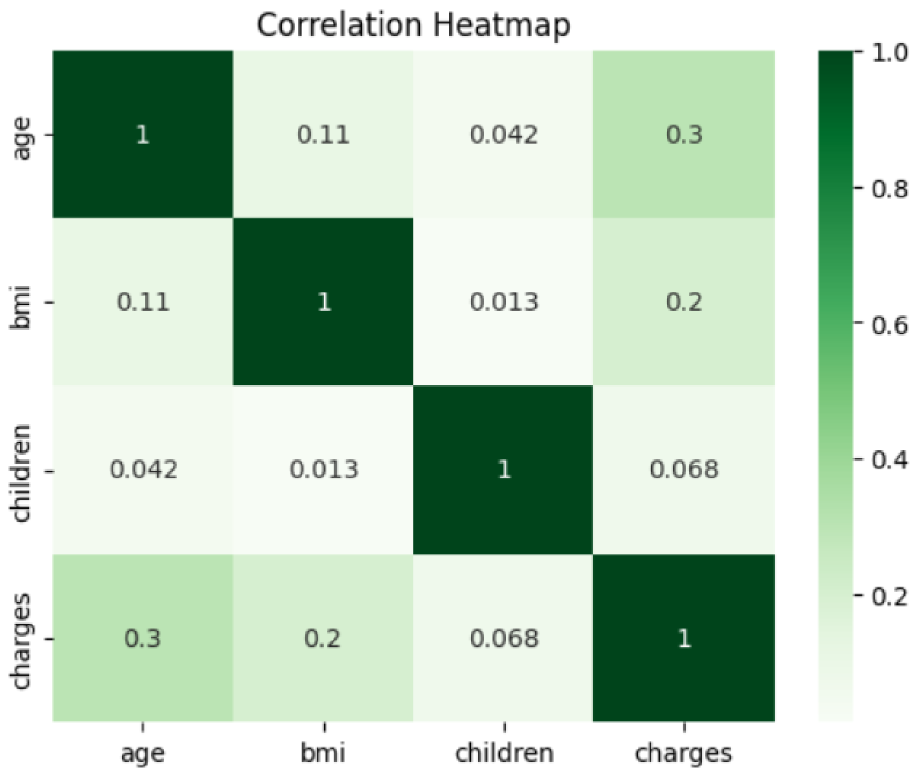
d. whereas 4 or 5 children's parents have lower charges and narrow spread. this may be due to less sample size or poor financial conditions, who can't afford the insurance cost.

Charges vs region:

a. There is no clear difference in charges in regions. However, southeast and northeast show wider boxplot than southwest and northwest. It means that southeast and northeast people have more variable insurance charges

Correlation heatmap:

```python
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap="Greens")
plt.title("Correlation Heatmap")
plt.show()
```



HYPOTHESIS TESTING:

1. **Is smoking status independence of sex?**

   Since smoking (yes or no) and sex (male or female) both are categorical columns, Chi Squared test of independence is appropriate here.

   Chi-Squared test of independence:

   a. Hypotheses:
      Null Hypothesis (H0): Smoking and sex are independent of each other.
      Alternative Hypothesis (H1): Smoking and sex are dependent on each other.
   b. Assumptions Checking: There are no specific assumptions to check for the Chi-square test.
   c. P-value and Conclusion: After conducting the Chi-square test, we'll obtain the p-value. If the p value is less than alpha, we reject the null hypothesis and conclude that there is a significant association between smoking and sex.

```python
from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(df['sex'], df['smoker'])
chi2, p, dof, expected = chi2_contingency(contingency_table)
print("p- value", p)
print("contingency table: ")
print(contingency_table)
```

```
p- value 0.006548143503580696
contingency table:
smoker   no  yes
sex
female  547  115
male    517  159
```

Insights: Since p-value turns out to be 0.006 < 0.05. Hence with 95% confidence, we reject the null hypothesis. That is, there is a relationship between sex and smoking. To find out the relationship between smoking and sex, we will find the % in the contingency table.

```python
[ ] pd.crosstab(df['sex'], df['smoker'], normalize='index') * 100
```

smoker	no	yes
sex		
female	82.628399	17.371601
male	76.479290	23.520710

Hence, it was statistically found out that males (23.5%) are more likely to be smokers than females (17.4%).

2. **Is BMI correlated with age?**
   Since BMI and age both are numerical columns, we will find out the relationship through **Pearson correlation**.
   a. Hypothesis:
      Null Hypothesis (H0): No correlation between age and BMI
      Alternative Hypothesis (H1): There is a correlation
   b. P-value and Conclusion: After conducting the Chi-square test, we'll obtain the p-value. If the p value is less than alpha, we reject the null hypothesis and conclude that there is a significant association between smoking and sex.

```
[ ] from scipy.stats import pearsonr

    corr, p_value = pearsonr(df['age'], df['bmi'])
    print(p_value)
```

```
6.194289065049117e-05
```

Insights: Since p value for BMI and age is 6.194289065049117e-05, i.e., 0.0000619 < 0.05, hence with 95% confidence we reject the Null hypothesis. **There is a statistically significant linear relationship between BMI and age.**

To find the relationship between BMI and age, we will find Pearson Coefficient.

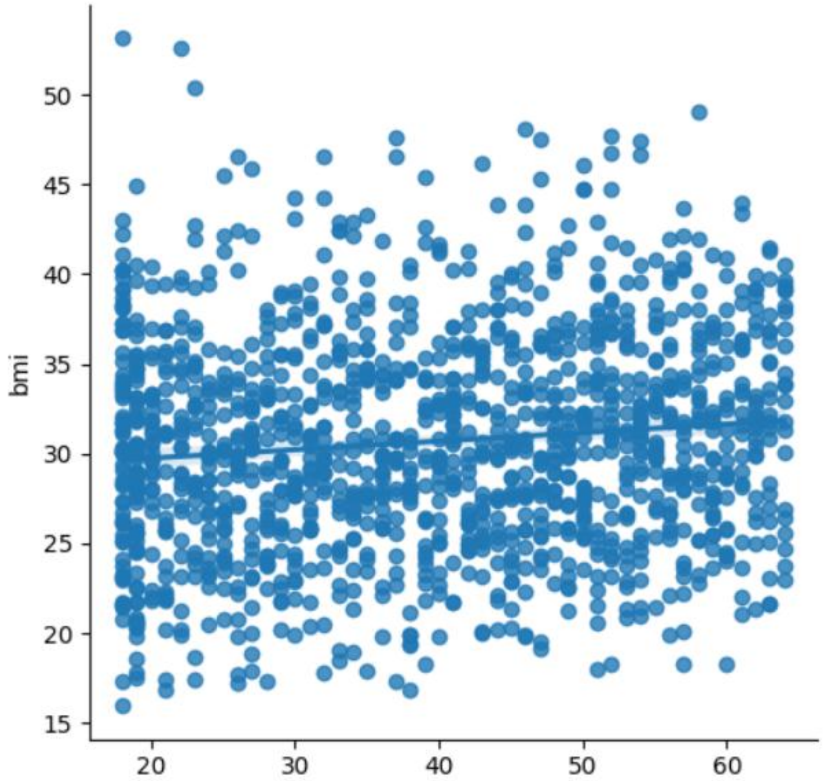
```
#Find the relationship  between BMI and age

from scipy.stats import pearsonr
r, p = pearsonr(df['age'], df['bmi'])
print("Correlation Coefficient:", r)
```

```
Correlation Coefficient: 0.10927188154853515
```

Insights: There is a very weak positive correlation between BMI and age. As age increases, BMI very slightly increases — but the effect is minimal. Let's see the relationship between them through scatterplot.

```python
sns.lmplot(x='age', y='bmi', data=df)
```

<seaborn.axisgrid.FacetGrid at 0x7fc6bcc106d0>



3. Does BMI vary across regions?

Test Type: One-way ANOVA (numeric vs categorical)

a. Hypothesis:
   Null Hypothesis (H0): Mean BMI is the same across all regions.
   Alternative Hypothesis (H1): At least one region has different BMI
b. P-value and Conclusion: After conducting one way ANOVA, we'll obtain the p-value. If the p value is less than alpha, we reject the null hypothesis and conclude that at least one region has different BMI.

```python
from scipy.stats import f_oneway

groups = [df[df['region'] == r]['bmi'] for r in df['region'].unique()]
f_stat, p_value = f_oneway(*groups)
print(p_value)
```

```
1.881838913929143e-24
```

Insights: Hence, we reject the Null Hypothesis. Hence, at least one region has different BMI.

FINAL INSIGHTS: Based on the results of the hypothesis tests and analysis conducted, we can derive the following final insights:

RECOMMENDATIONS:

Age 18-35:

1. Every group has extreme-cost individuals, not just elderly, company can launch initiatives like **'Outlier-Aware Risk Pooling'.**

Age 35-45:

1. Introduce **Preventive Health Plans (**(free checkups, fitness programs) for Mid-Age Adults
2. Focus on early detection to reduce future high-cost claims.

Age 45-65:

1. Since Older age groups consistently have higher costs, create tiered premium plans by Age bracket.
2. Consider higher base premiums or more comprehensive plans for 45+.

NOTE:

1. Offer custom pricing for combinations. e.g., BMI > 30 + smoker + age 45+ = very high-risk group.

2. People with 2 or 3 children can be targeted for family-focussed insurance plans. Some family-centric insurance plans can be introduced which includes-

    a. Covers 2 adults + 2 or 3 children

    b. Coverage for: Paediatric care, Maternity & postnatal care, Vaccinations and routine child checkups.

    c. **School-time accident protection** for kids

    d. Dental and vision care for children.

3. Offer bundled coverage for dependents, maternity, and paediatric services.

4. People with larger family, 4 or 5 children, need not be charged. Child related claims can be considered separately.

5. Since male tend to have higher and more variable cost than females, premium policy should be more gender-centric with certain age-risk brackets.

6. Since, males (23.5%) are more likely to be smokers than females (17.4%), some smoking- centric insurance plans can be introduced which includes-
    a. Higher base premium
    b. Covers: Heart disease, cancer (especially lung), COPD, stroke
    c. **Annual health checkups** included
    d. **Smoking cessation benefit**: premium discounts after 1–2 years of quitting (verified)

7. Policies in southeast region can be more robust as there is more variability and some people exist with higher medical expenses.

8. Adjust base premiums for southeast to reflect the higher expected claim amounts.

9. Introduce regional risk scoring in pricing models.