Introduction: Health Insurance Cost Analysis (EDA Project)

1. Introduction

Health insurance plays a crucial role in managing medical expenses, and its costs vary significantly based on several factors. This project aims to analyze health insurance charges and understand the impact of key factors such as age, BMI, region, smoking status, and other demographic attributes. By leveraging Exploratory Data Analysis (EDA), we will identify trends, correlations, and insights that influence insurance premiums.

2. Objectives

The main objectives of this analysis include:

- Understanding the distribution of insurance costs across different demographics.
- Exploring the relationship between smoking status and insurance charges.
- Analyzing the effect of age and BMI on insurance costs.
- Comparing insurance charges across different regions.
- Identifying outliers that significantly affect cost distribution.

3. Methodology

To conduct this analysis, we will use Python and its data analytics libraries:

- pandas – For data manipulation and preprocessing.
- numpy – For numerical computations.
- matplotlib & seaborn – For data visualization to uncover trends and relationships.

We will perform the following EDA techniques:

- Data Cleaning – Handling missing values, duplicates, and data inconsistencies.
- Descriptive Statistics – Understanding data distribution and key statistics.
- Data Visualization – Using histograms, boxplots, scatterplots, and correlation heatmaps.
- Outlier Detection – Identifying extreme values that might affect analysis.

Step 1: import the libraries

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

Step 2: Importing Data Set

```
Health_Insurance_Cost = pd.read_csv(r"C:\Users\ASUS\OneDrive\Desktop\
Project\insurance.csv")
print(Health_Insurance_Cost)
```

```
      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
...   ...     ...     ...       ...    ...        ...          ...
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
1337   61  female  29.070         0    yes  northwest  29141.36030

[1338 rows x 7 columns]
```

Step 3: Dataset Overview

The dataset typically includes the following columns:

Column Name Description

- age - Age of the individual
- sex - Gender (male/female)
- bmi - Body Mass Index
- children - Number of dependents
- smoker - Whether the person is a smoker (yes/no)
- region - Geographic region (northeast, northwest, southeast, southwest)
- charges - Insurance cost (target variable)

show First 5 row form top

```
Health_Insurance_Cost.head(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
```

show Last 5 row form top

```
Health_Insurance_Cost.tail(5)
```

```
      age     sex    bmi  children smoker     region      charges
1333   50    male  30.97         3     no  northwest  10600.5483
1334   18  female  31.92         0     no  northeast   2205.9808
1335   18  female  36.85         0     no  southeast   1629.8335
1336   21  female  25.80         0     no  southwest   2007.9450
1337   61  female  29.07         0    yes  northwest  29141.3603
```

Step 4 : Analys the shape , Dimenssion , Rows , Data Types and all the meta data information

```
Health_Insurance_Cost.shape
```

```
(1338, 7)
```

```
Health_Insurance_Cost.ndim
```

```
2
```

```
Health_Insurance_Cost.columns
```

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

```
Health_Insurance_Cost.dtypes
```

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

```
Health_Insurance_Cost.info()
```

```
<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
```

Step 5 : Derive the 5 Number Summary

```
Health_Insurance_Cost.describe()
```

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Step 6 : Analys the Null values and Duplicate values

```
Health_Insurance_Cost.nunique()
```

```
age           47
sex            2
bmi          548
children       6
smoker         2
region         4
charges     1337
dtype: int64
```

```
Health_Insurance_Cost.isna().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

```
Health_Insurance_Cost.duplicated().sum()
```

```
1
```

```
Health_Insurance_Cost[Health_Insurance_Cost.duplicated()]
```

	age	sex	bmi	children	smoker	region	charges
581	19	male	30.59	0	no	northwest	1639.5631

Here we have 1 Duplicate value in our data set (i.e. in the line no 581). So we have to remove this Duplicate value by using drop function .

```
Health_Insurance_Cost.drop_duplicates(inplace=True)
Health_Insurance_Cost.duplicated().sum()
```

```
0
```

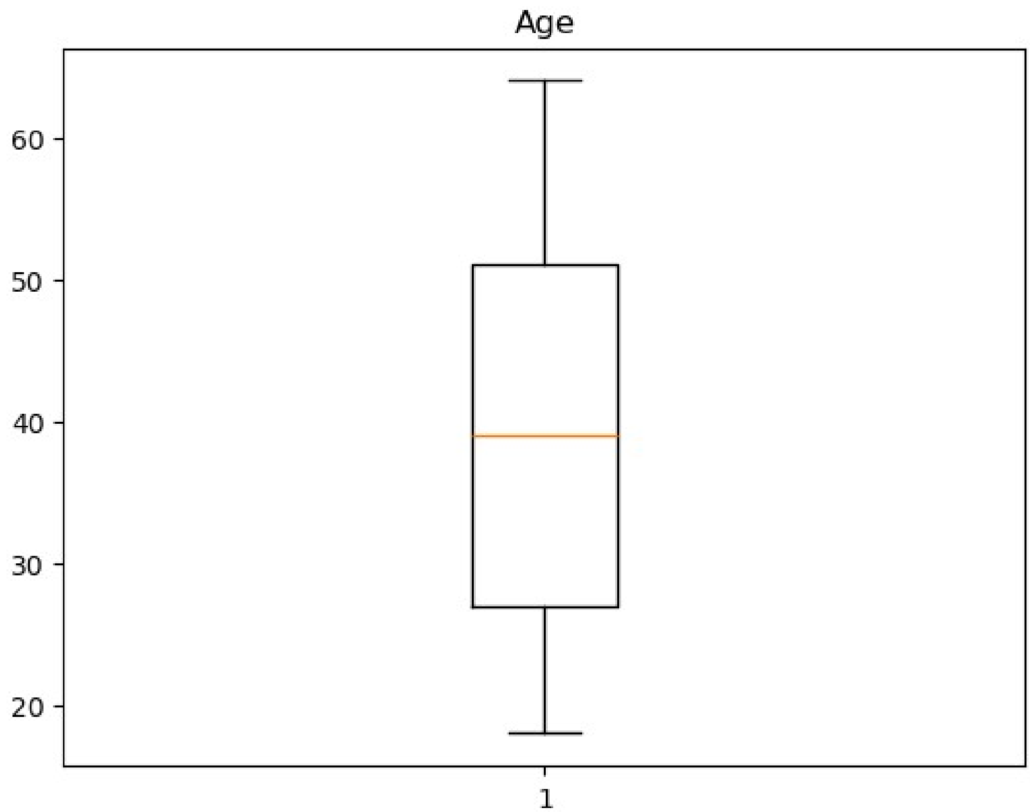
Step 7 : Analys the Outlier Values

```
Health_Insurance_Cost['age']
```

```
0       19
1       18
2       28
3       33
4       32
        ..
1333    50
1334    18
1335    18
1336    21
1337    61
Name: age, Length: 1337, dtype: int64
```

- By ploting the Boxplot Digram we can see the Outlier Values of 'Age' coloumn

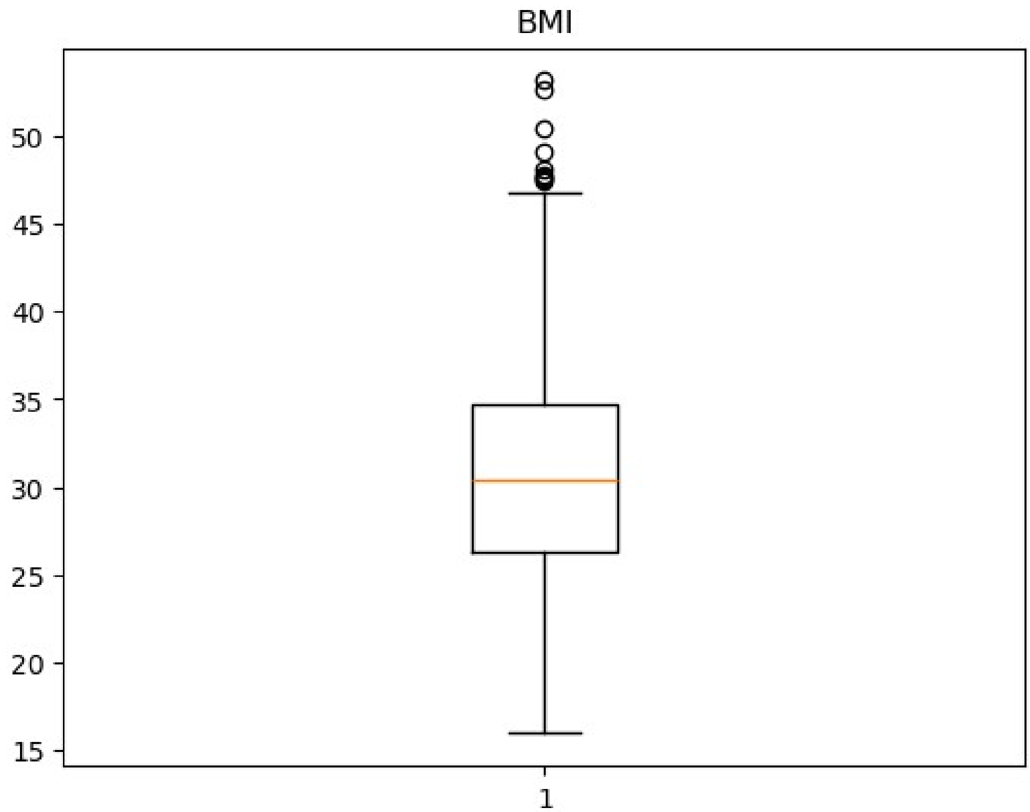
```
plt.boxplot(Health_Insurance_Cost['age'])
plt.title("Age")
plt.show()
```



From the above digram i cam see that the 'Age' coloumn has no outlier value

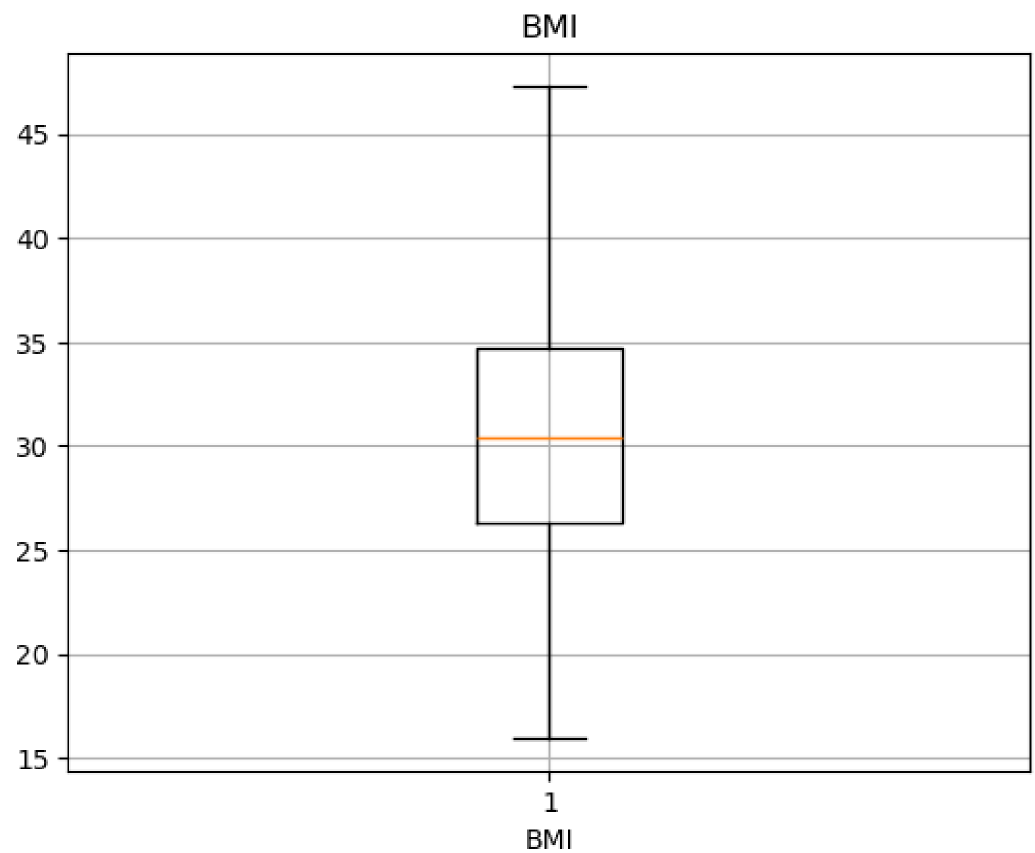
- By ploting the Boxplot Digram we can see the Outlier Values of 'BMI' coloumn

```
plt.boxplot(Health_Insurance_Cost['bmi'])
plt.title("BMI")
plt.show()
```



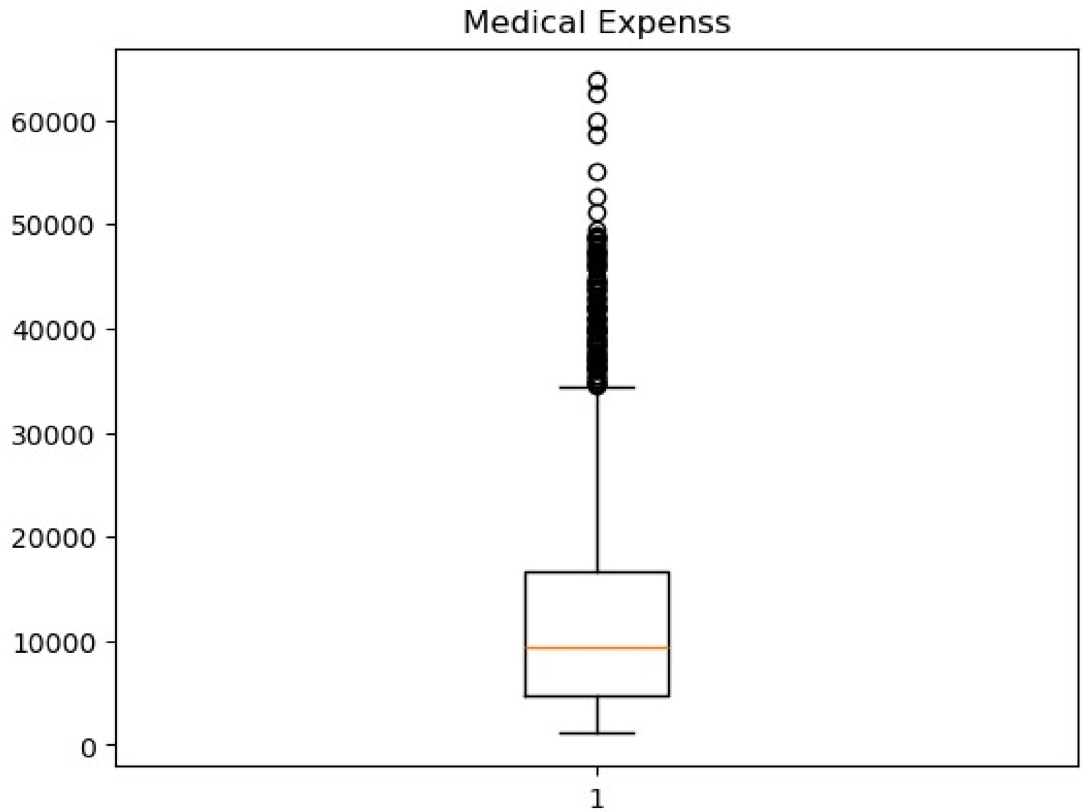
In this Boxplot Digram there are some outlier values are presented. In order to remove those Outlier values we can use capping the Value by it's 'IQR' Range ,in this way we can remove the outlier with out dropping any values from the coloumn

```
def remove_outliers(Health_Insurance_Cost, col):
    Q1 = Health_Insurance_Cost[col].quantile(0.25)
    Q3 = Health_Insurance_Cost[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

Health_Insurance_Cost_cleaned=np.clip(Health_Insurance_Cost[col],lower
_bound,upper_bound)
    return Health_Insurance_Cost_cleaned

Health_Insurance_Cost['bmi']=remove_outliers(Health_Insurance_Cost,
'bmi')

plt.boxplot(Health_Insurance_Cost['bmi'])
plt.xlabel('BMI')
plt.grid(True)
plt.title("BMI")
plt.show()
```



- By ploting the Boxplot Digram we can see the Outlier Values of 'Charge' coloumn

```
plt.boxplot(Health_Insurance_Cost['charges'])
plt.title("Medical Expenss")
plt.show()
```

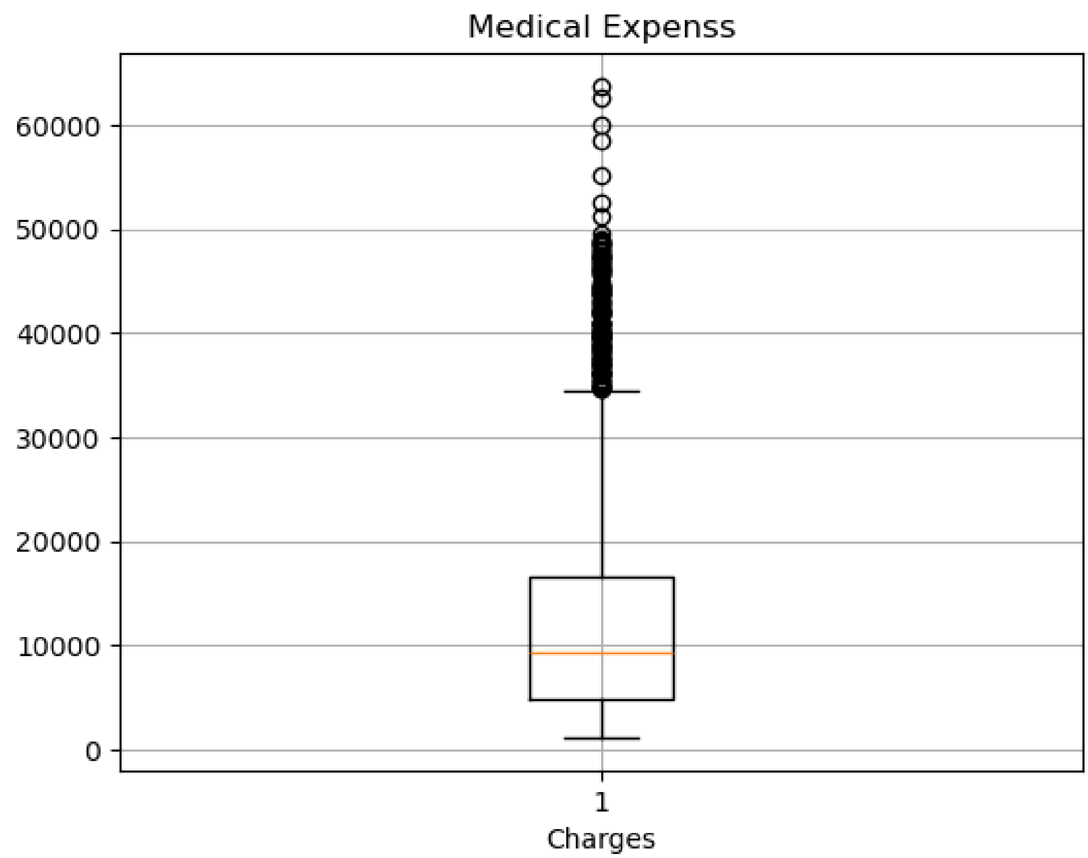


In this Boxplot Digram there are some outlier values are presented. In order to remove those Outlier values we can use above capping function the Value by it's 'IQR' Range ,in this way we can remove the outlier with out dropping any values from the coloumn

```
Q1 = Health_Insurance_Cost['charges'].quantile(0.25)
Q3 = Health_Insurance_Cost['charges'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
np.where(Health_Insurance_Cost['charges']>upper_bound,Health_Insurance
_Cost['charges'].mean(),Health_Insurance_Cost['charges'])
```

```
array([16884.924 ,  1725.5523,  4449.462 , ...,  1629.8335,
2007.945 ,
       29141.3603])
```

```
plt.boxplot(Health_Insurance_Cost['charges'])
plt.xlabel('Charges')
plt.grid(True)
plt.title("Medical Expenss")
plt.show()
```



Step 8 : Categorical Feature Analysis

1. Factors Influencing Insurance Costs

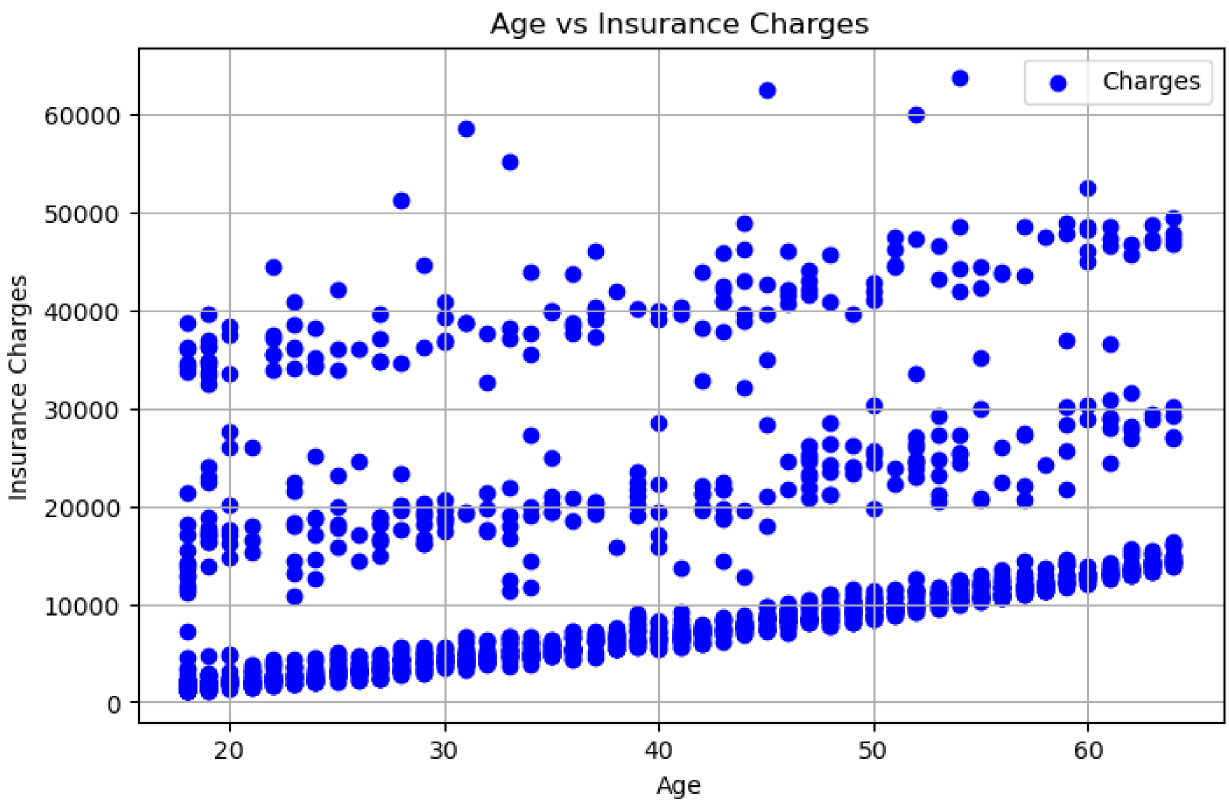
Age:

```
pd.pivot_table(Health_Insurance_Cost,index='age',values='charges',sort
=True)
```

age	charges
18	7086.217556
19	9868.929428
20	10159.697736
21	4730.464330
22	10012.932802
23	12419.820040
24	10648.015962
25	9838.365311
26	6133.825309
27	12184.701721
28	9069.187564
29	10430.158727

```
30    12719.110358
31    10196.980573
32     9220.300291
33    12351.532987
34    11613.528121
35    11307.182031
36    12204.476138
37    18019.911877
38     8102.733674
39    11778.242945
40    11772.251310
41     9653.745650
42    13061.038669
43    19267.278653
44    15859.396587
45    14830.199856
46    14342.590639
47    17653.999593
48    14632.500445
49    12696.006264
50    15663.003301
51    15682.255867
52    18256.269719
53    16020.930755
54    18758.546475
55    16164.545488
56    15025.515837
57    16447.185250
58    13878.928112
59    18895.869532
60    21979.418507
61    22024.457609
62    19163.856573
63    19884.998461
64    23275.530837
```

```
plt.figure(figsize=(8, 5))
plt.scatter(Health_Insurance_Cost['age'],
Health_Insurance_Cost['charges'], marker='o', linestyle='-',
color='b', label='Charges')

# Labels & Title
plt.xlabel('Age')
plt.ylabel('Insurance Charges')
plt.title('Age vs Insurance Charges')
plt.legend()
plt.grid(True)

# Show Plot
plt.show()
```

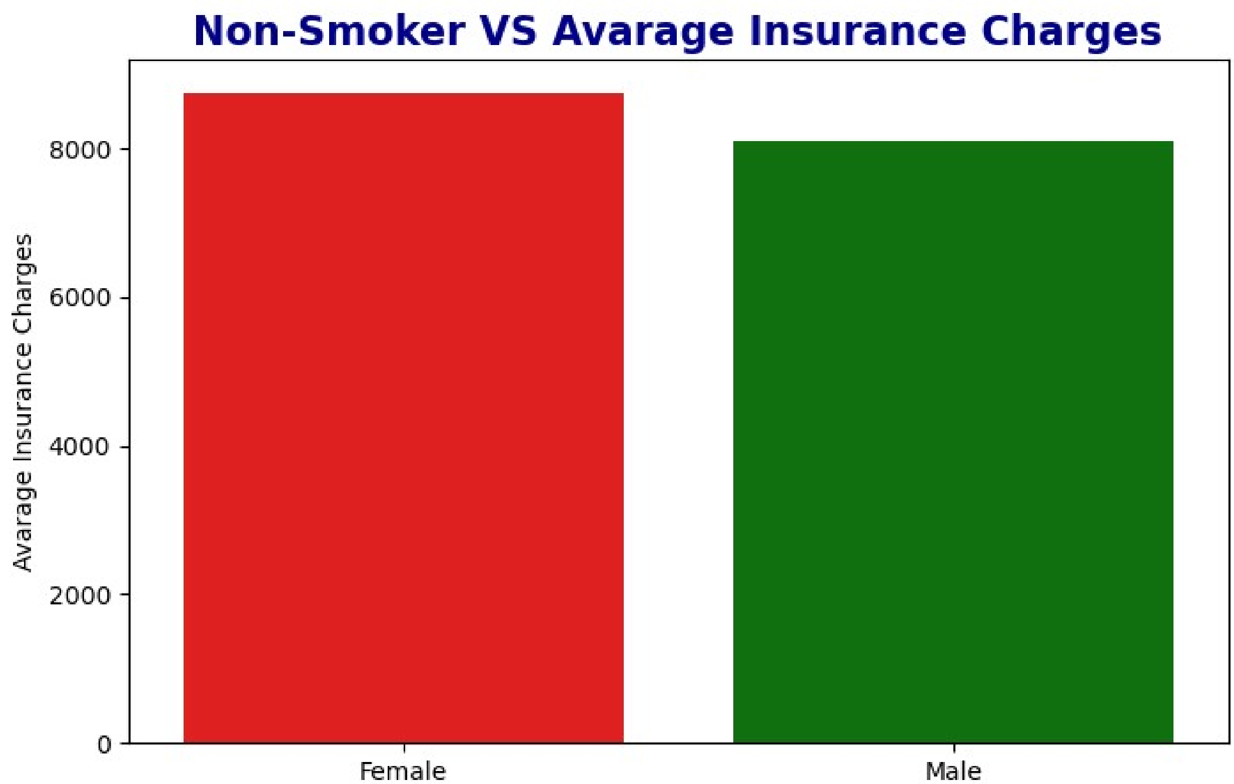


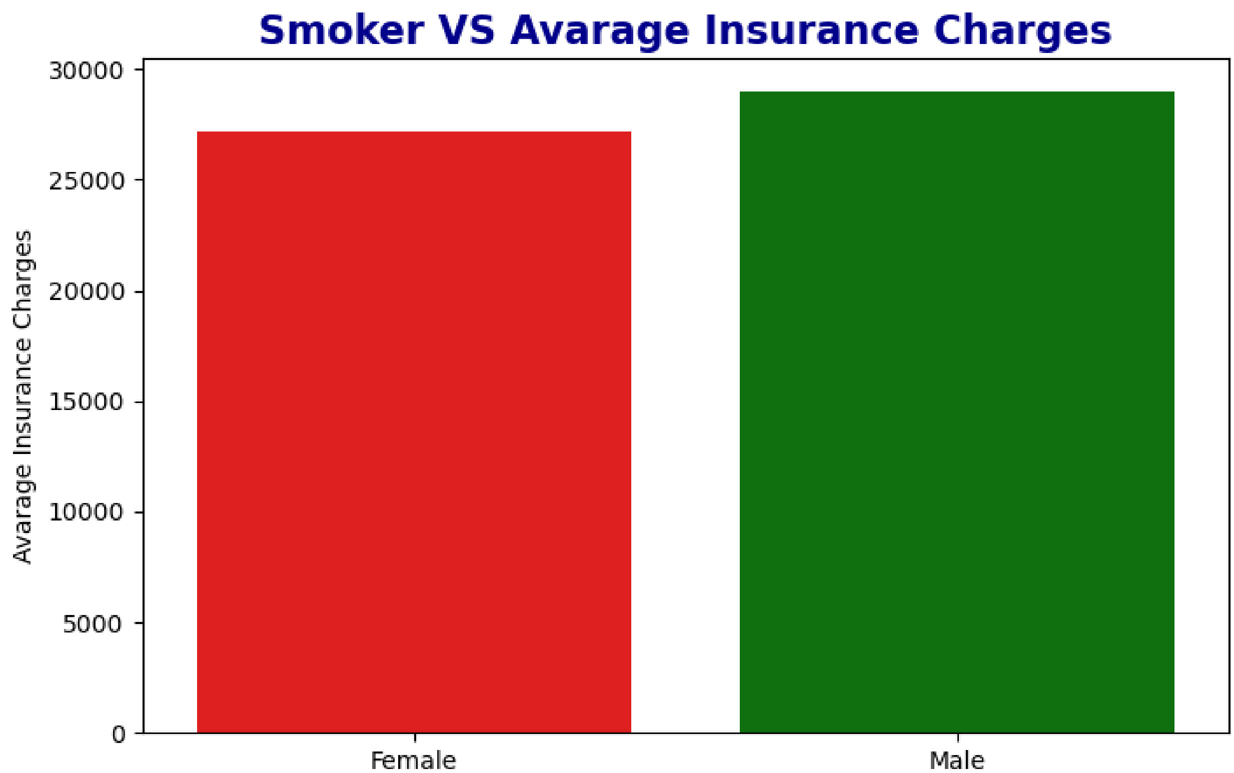
Age: By exploring the above "Scatter chart" & "Pivote Table" I can relate the insurance costs vary with age. Older individuals typically pay higher premiums due to increased health risks.

Smoker:

```
Health_Insurance_Cost.groupby(['smoker','sex']).charges.mean()

smoker  sex
no      female     8762.297300
        male       8099.700161
yes     female    30678.996276
        male      33042.005975
Name: charges, dtype: float64

x = np.array(['Female','Male'])
y = np.array([8753.016327,8099.700161])
colors=np.array(['red','green'])
plt.figure(figsize=(8, 5))
ax = sns.barplot(x=x,y= y, palette=colors)
plt.ylabel('Avarage Insurance Charges')
plt.title('Non-Smoker VS Avarage Insurance Charges', fontsize=16,
color='darkblue', fontweight='bold')
plt.show()
x = np.array(['Female','Male'])
y = np.array([27128.684566,29015.955452])
plt.figure(figsize=(8, 5))
```

```
ax = sns.barplot(x=x,y= y, palette=colors)
plt.ylabel('Avarage Insurance Charges')
plt.title('Smoker VS Avarage Insurance Charges', fontsize=16,
color='darkblue', fontweight='bold')
plt.show()
```





Smoker: By exploring the above "Bar chart" & "Pivote Table" I can relate the insurance costs vary with Smoker. Smokers generally have higher insurance costs, due to increased health risks.

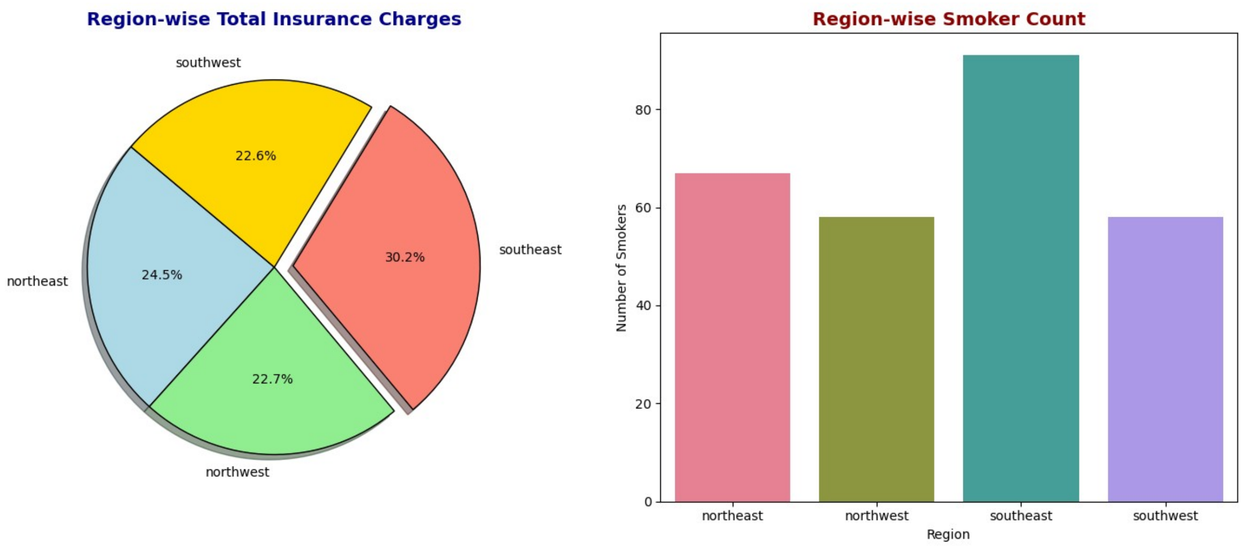
Region:

```
Health_Insurance_Cost.groupby(["region"]).charges.sum()
```

```
region
northeast    4.343669e+06
northwest    4.034072e+06
southeast    5.363690e+06
southwest    4.012755e+06
Name: charges, dtype: float64
```

```
region_charges = Health_Insurance_Cost.groupby('region')
['charges'].sum()
region_smokers=Health_Insurance_Cost[Health_Insurance_Cost['smoker']
== 'yes'].groupby('region').size()
explode_values = [0.1 if charge == region_charges.max() else 0 for
charge in region_charges]
fig,axes = plt.subplots(1, 2, figsize=(14, 6))
colors = ['lightblue', 'lightgreen', 'salmon', 'gold']
axes[0].pie(region_charges, labels=region_charges.index,
autopct='%1.1f%%', colors=colors,startangle=140,
wedgeprops={'edgecolor': 'black'}, explode=explode_values,
shadow=True)
axes[0].set_title('Region-wise Total Insurance Charges', fontsize=14,
```

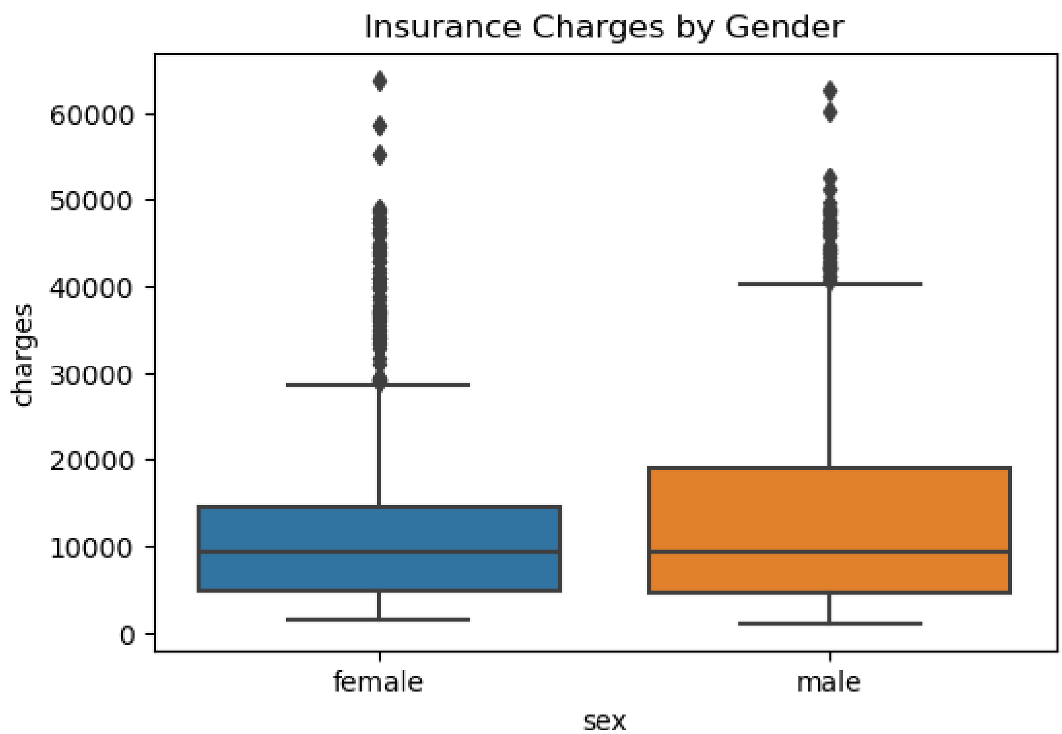
```
color='darkblue', fontweight='bold')

sns.barplot(x=region_smokers.index, y=region_smokers.values,
ax=axes[1], palette='husl')
axes[1].set_title('Region-wise Smoker Count', fontsize=14,
color='darkred', fontweight='bold')
axes[1].set_xlabel('Region')
axes[1].set_ylabel('Number of Smokers')

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



From the above 2 graph i can tell that the South-East are people are taking more life insurance than other area (Due to number of Smoker in that area is high).If we devide the total into 2 part i.e. "Eastern Side" and "Western Side" then we can tell that Eastern Side has more Smoker.

Gender

```
Health_Insurance_Cost.groupby(["sex"]).charges.sum()

sex
female    8.321061e+06
male      9.433124e+06
Name: charges, dtype: float64

plt.figure(figsize=(6, 4))
sns.boxplot(x='sex', y='charges', data=Health_Insurance_Cost)
plt.title('Insurance Charges by Gender')
plt.show()
```



From the above graph we can see that Male has slightly higher charge in comparigion to Female

```
Heatmap_Dataset=Health_Insurance_Cost
Heatmap_Dataset['sex']=Heatmap_Dataset.sex.apply(lambda x: 1 if x ==
'male' else 0)
Heatmap_Dataset['smoker']=Heatmap_Dataset.smoker.apply(lambda x: 1 if
x == 'yes' else 0)
Heatmap_Dataset = pd.concat((Heatmap_Dataset,
pd.get_dummies(Heatmap_Dataset['region'], dtype = int)), axis = 1)
Heatmap_Dataset =Heatmap_Dataset.drop(columns=['region'])
Heatmap_Dataset
```

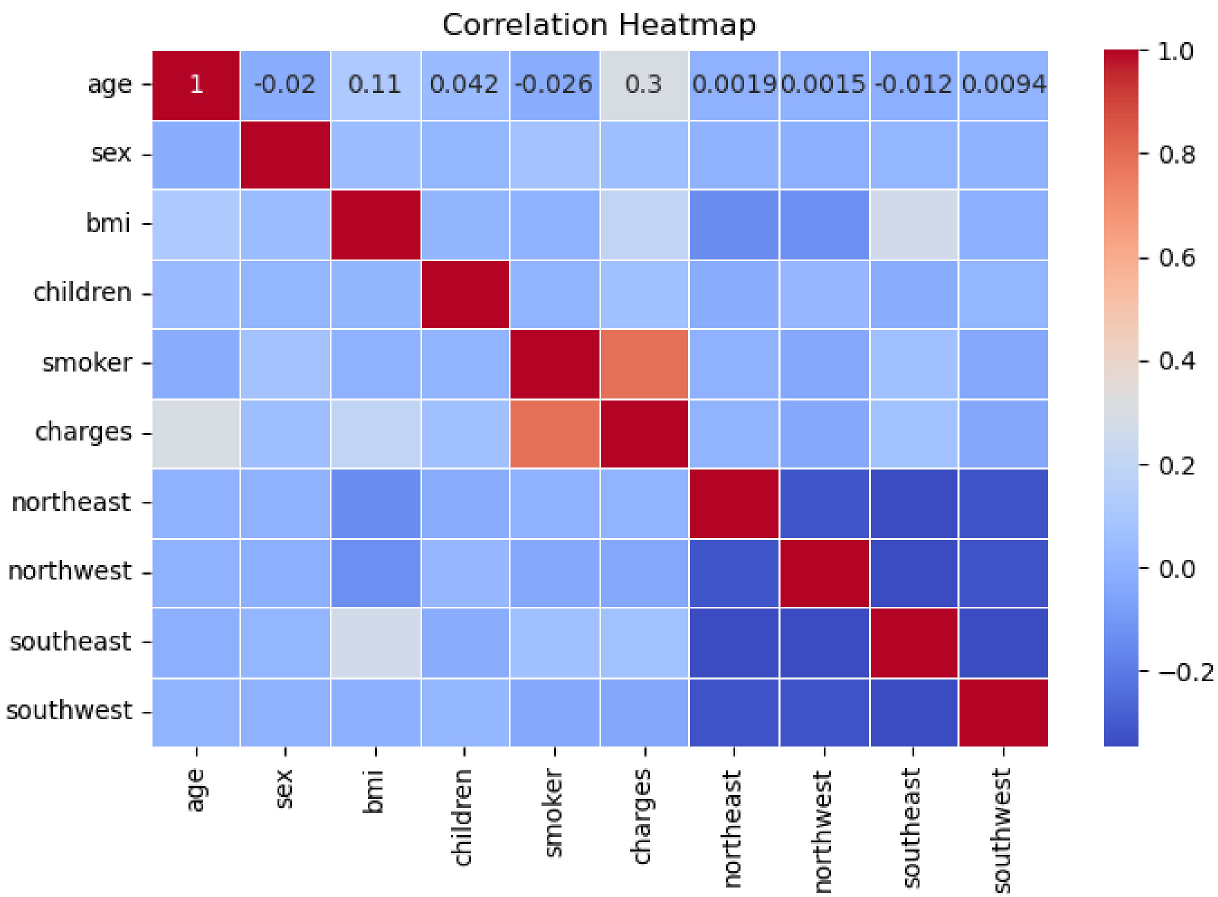
```
      age  sex     bmi  children  smoker      charges  northeast  
northwest  \
0      19    0  27.900         0       1  16884.92400          0  
0  
1      18    1  33.770         1       0   1725.55230          0  
0  
2      28    1  33.000         3       0   4449.46200          0  
0  
3      33    1  22.705         0       0  21984.47061          0  
1  
4      32    1  28.880         0       0   3866.85520          0  
1  
...   ...  ...     ...       ...     ...          ...        ...  
...  
1333   50    1  30.970         3       0  10600.54830          0  
1  
```

```
1334    18     0  31.920          0        0   2205.98080           1
0
1335    18     0  36.850          0        0   1629.83350           0
0
1336    21     0  25.800          0        0   2007.94500           0
0
1337    61     0  29.070          0        1  29141.36030           0
1

      southeast  southwest
0             0          1
1             1          0
2             1          0
3             0          0
4             0          0
...         ...        ...
1333          0          0
1334          0          0
1335          1          0
1336          0          1
1337          0          0

[1337 rows x 10 columns]
```

Step 9 : Multivariate Analysis

Correlation Heatmap

```
plt.figure(figsize=(8, 5))
sns.heatmap(Heatmap_Dataset.corr(), annot=True, cmap='coolwarm',
linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```



From this Heatmap we can say that some of coloums have Stronger Bond & some of have Weaker Bond with each other

Step 10 : Conclusion

- Age and Charges: Insurance cost increases with age.
- BMI and Charges: Higher BMI may result in higher charges.
- Smoking Impact: Smokers have significantly higher insurance costs.
- Regional Variations: Some regions have slightly higher insurance costs.