Analysis Report

Domain: Healthcare, Insurance

Prepared By:

Tohfa Siddika Barbhuiya,

EXL Analytics

10 Jan’2021

EXECUTIVE SUMMARY:

BACKGROUND:

Leveraging customer information is paramount for most businesses. In the case of an insurance company, 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:

To dive deep into this data to find some valuable insights.

Attribute Information:

age: age of primary beneficiary

sex: insurance contractor gender, female, male

bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

children: Number of children covered by health insurance / Number of dependents

smoker: Smoking

region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

charges: Individual medical costs billed by health insurance.

KEY FINDINGS:

Q1)Import the necessary libraries

import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

import statsmodels.api as sm

import scipy.stats as stats

from sklearn.preprocessing import LabelEncoder

import copy

Q2)Read the data as a data frame (3 marks)

df = pd.read_csv("C:/Users/user/Downloads/insurance.csv")

df.head(3) #Reading 3 values

OUTPUT:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.90	0	yes	southwest	16884.9240
1	18	male	33.77	1	no	southeast	1725.5523
2	28	male	33.00	3	no	southeast	4449.4620

Q3) Perform basic EDA which should include the following and print out your insights at every step.

Q)3.a. Shape of the data

df.shape

Answer: The dataframe Insurance.csv has 1338 row and 7 columns

OUTPUT: (1338, 7)

Q) 3.b. Data type of each attribute

df.info()

Answer: The attributes age and children are integer type, sex, smoker, region are object type, bmi and charges are float datatype

OUTPUT:

```
 #   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
```

Q)3.c. Checking the presence of missing values

df.isnull().sum()

Answer: From the output, it can be inferred that there are no missing values present in the given dataframe

OUTPUT:

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

Q) 3.d. 5 point summary of numerical attributes

df.describe()

Answer:

1. Mostly the people are in the age group of 51 years(75%)
2. Number of children is mostly not more than 2 children with maximum of upto 5 children
3. Charges are highly skewed

4. Mostly the bmi is 34.67
5. 18years upto 27 years age group are mostly not having children.

OUTPUT:

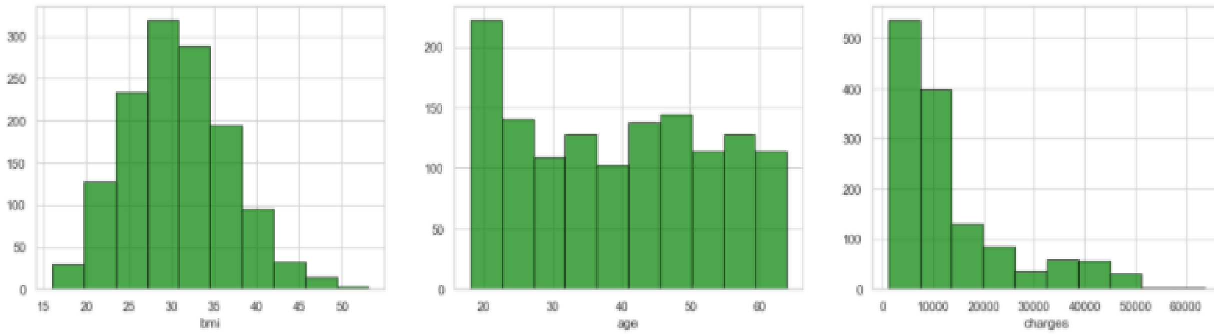
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000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

Q) 3.e. Distribution of 'bmi', 'age' and 'charges' columns.

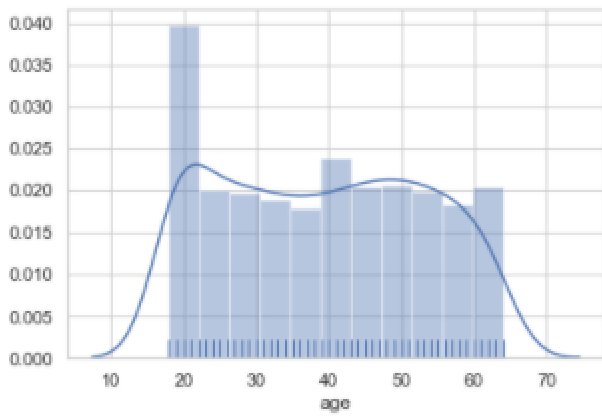
```
plt.figure(figsize= (20,15))

plt.subplot(3,3,1)

plt.hist(df.bmi, color='green', edgecolor = 'black', alpha = 0.7)

plt.xlabel('bmi')


plt.subplot(3,3,2)

plt.hist(df.age, color='green', edgecolor = 'black', alpha = 0.7)

plt.xlabel('age')


plt.subplot(3,3,3)

plt.hist(df.charges, color='green', edgecolor = 'black', alpha = 0.7)

plt.xlabel('charges')


plt.show()
```

Answer: bmi is slightly left skewed and charges are highly left skewed but age is almost uniformly distributed.

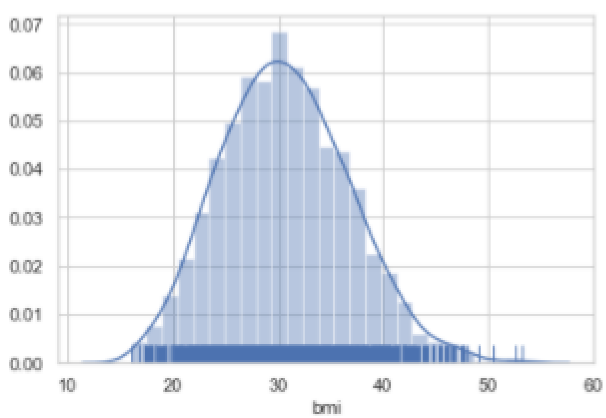
OUTPUT:



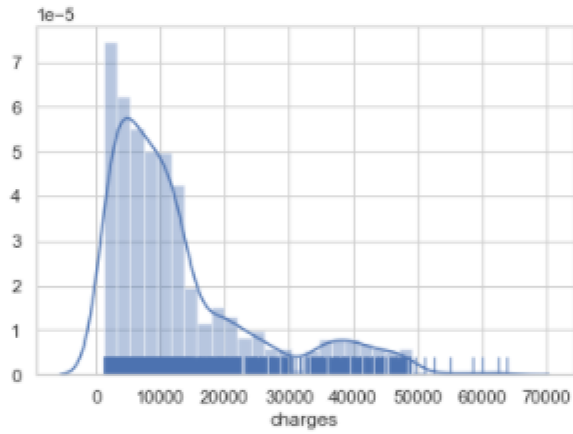

```
sns.distplot(df['age'], kde=True, rug=True)
```



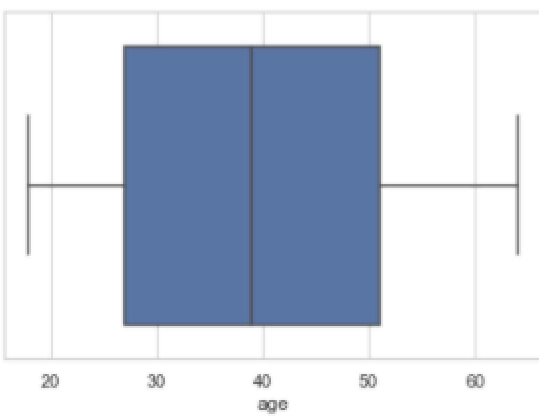

```
sns.distplot(df['bmi'], kde=True, rug=True);
```



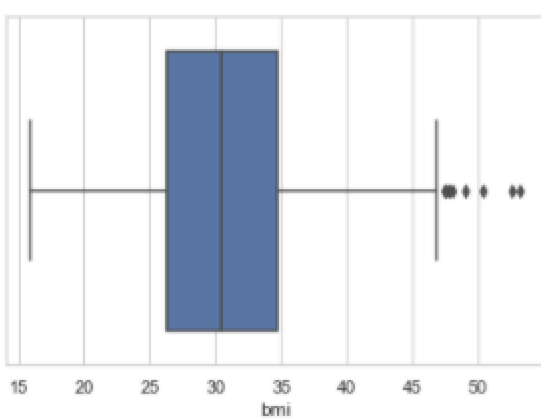
```python
sns.distplot(df['charges'], kde=True, rug=True);
```



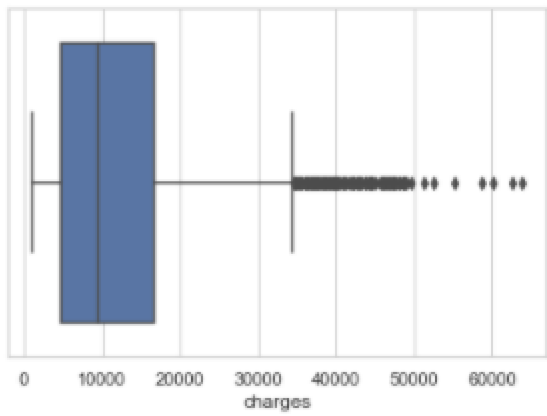

```python
sns.set(style="whitegrid")
sns.boxplot(df["age"])
```




```python
sns.boxplot(df["bmi"])
```



sns.boxplot(df["charges"])



Q) 3.f. Measure of skewness of 'bmi', 'age' and 'charges' columns

```
Skewness = pd.DataFrame({'Skewness' : [stats.skew(df.bmi),
                                        stats.skew(df.age),stats.skew(df.charges)]},
                         index=['bmi','age','charges'])
Skewness
```

Answer: Skewness of bmi is 0.28 which is very less, age is 0.05 which represents almost negligible skewness but of charges, skewness value is 1.51 which is high.

OUTPUT:

	Skewness
bmi	0.283729
age	0.055610
charges	1.514180

Q) 3.g. Checking the presence of outliers in 'bmi', 'age' and 'charges columns

```
Interquartile_range = np.subtract(*np.percentile(df['charges'], [75, 25]))
print(Interquartile_range)
```

OUTPUT: 11899.625365

```
q25, q75 = np.percentile(df['charges'], 25), np.percentile(df['charges'], 75)
Interquartile_range = q75 - q25
cut_off = Interquartile_range * 1.5
lower, upper = q25 - cut_off, q75 + cut_off

outliers = [x for x in df['charges'] if x < lower or x > upper]
print('Number of outliers for charges in 1338 data are- %d' % len(outliers))
```

OUTPUT: Number of outliers for charges in 1338 data are- 139

```
q25, q75 = np.percentile(df['bmi'], 25), np.percentile(df['bmi'], 75)
Interquartile_range = q75 - q25
cut_off = Interquartile_range * 1.5
lower, upper = q25 - cut_off, q75 + cut_off

outliers = [x for x in df['bmi'] if x < lower or x > upper]
print('Number of outliers for bmi in 1338 data are- %d' % len(outliers))
```

OUTPUT: Number of outliers for bmi in 1338 data are- 9

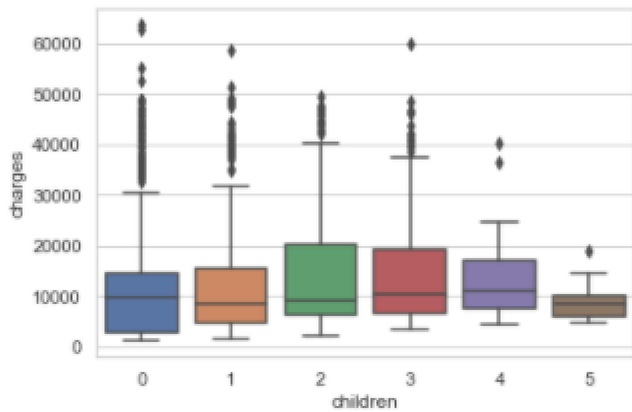
```
q25, q75 =
np.percentile(df['age'], 25), np.percentile(df['age'], 75)
Interquartile_range = q75 - q25
cut_off = iqr * 1.5
lower, upper = q25 - cut_off, q75 + cut_off

outliers = [x for x in df['age'] if x < lower or x > upper]
print('Number of outliers for age in 1338 data are- %d' % len(outliers))
```

OUTPUT: `Number of outliers for age in 1338 data are- 0`

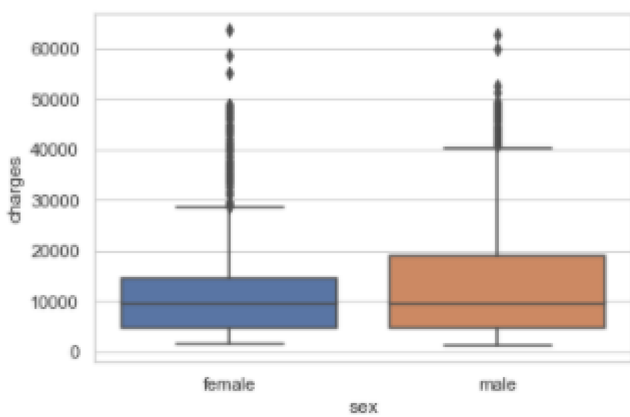
Answer: Number of outliers out of 1338 data for each of charges, bmi, age are 139,9 and 0 respectively.

#Q)3.h. Distribution of categorical columns (include children)

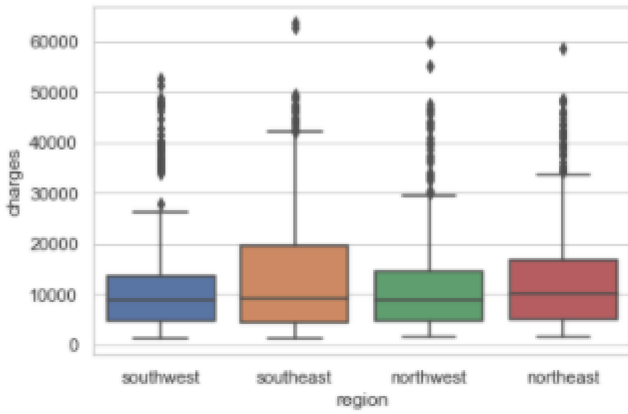
sns.boxplot(x='children', y='charges', data= df)



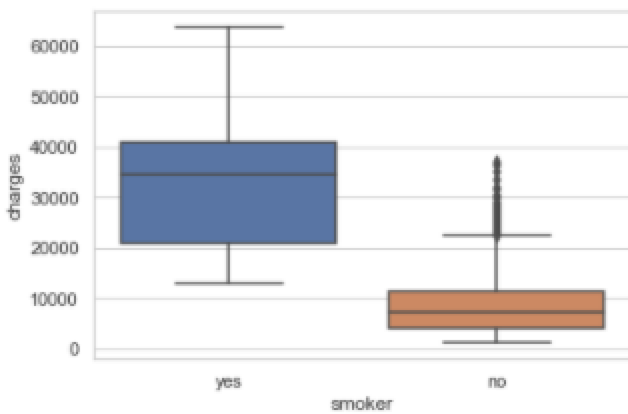
sns.boxplot(x='sex', y='charges', data= df)



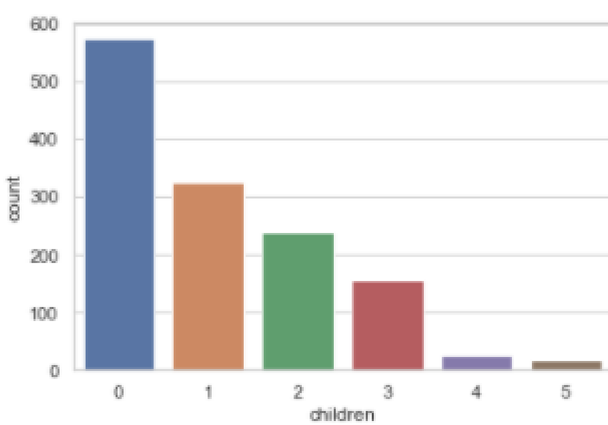
sns.boxplot(x='region', y='charges', data= df)



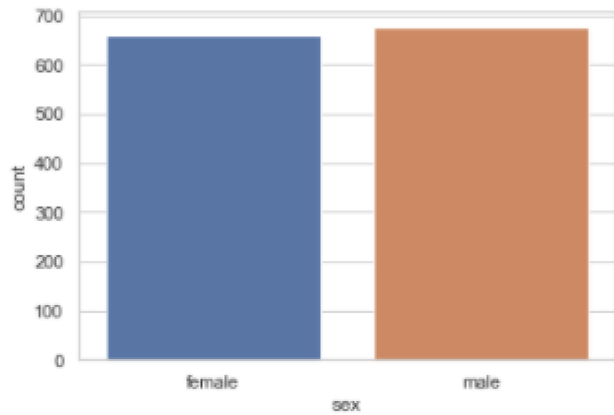
sns.boxplot(x='smoker', y='charges', data= df)



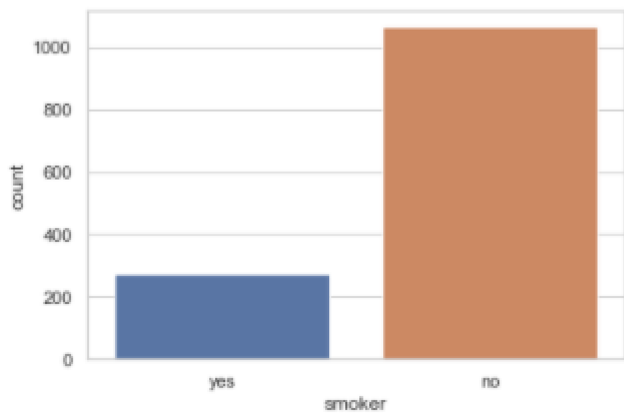
sns.countplot(df['children'])



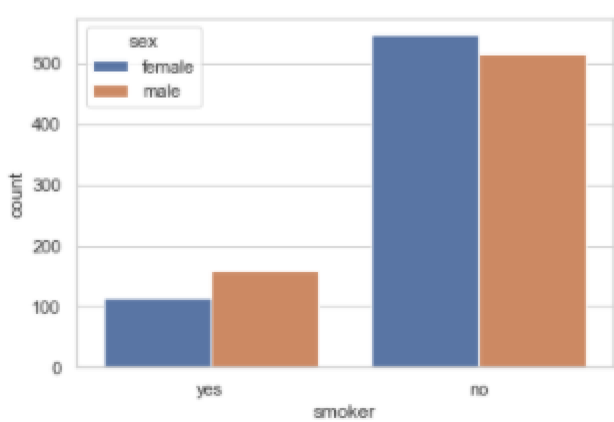
sns.countplot(df['sex'])



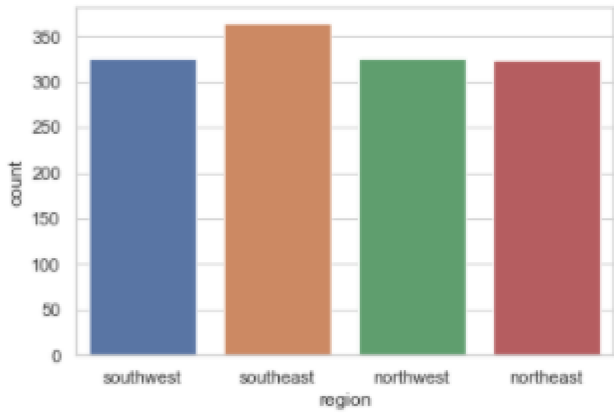
sns.countplot(df['smoker'])



sns.countplot(df['smoker'],hue = df['sex'])



sns.countplot(df['region'])



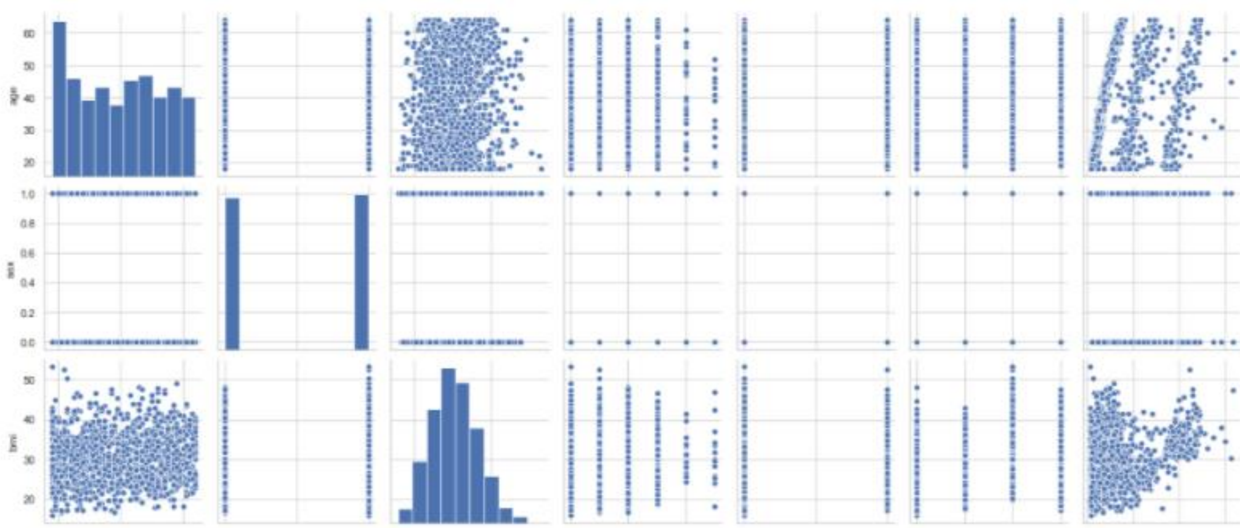
Q)3.i. Pair plot that includes all the columns of the data frame

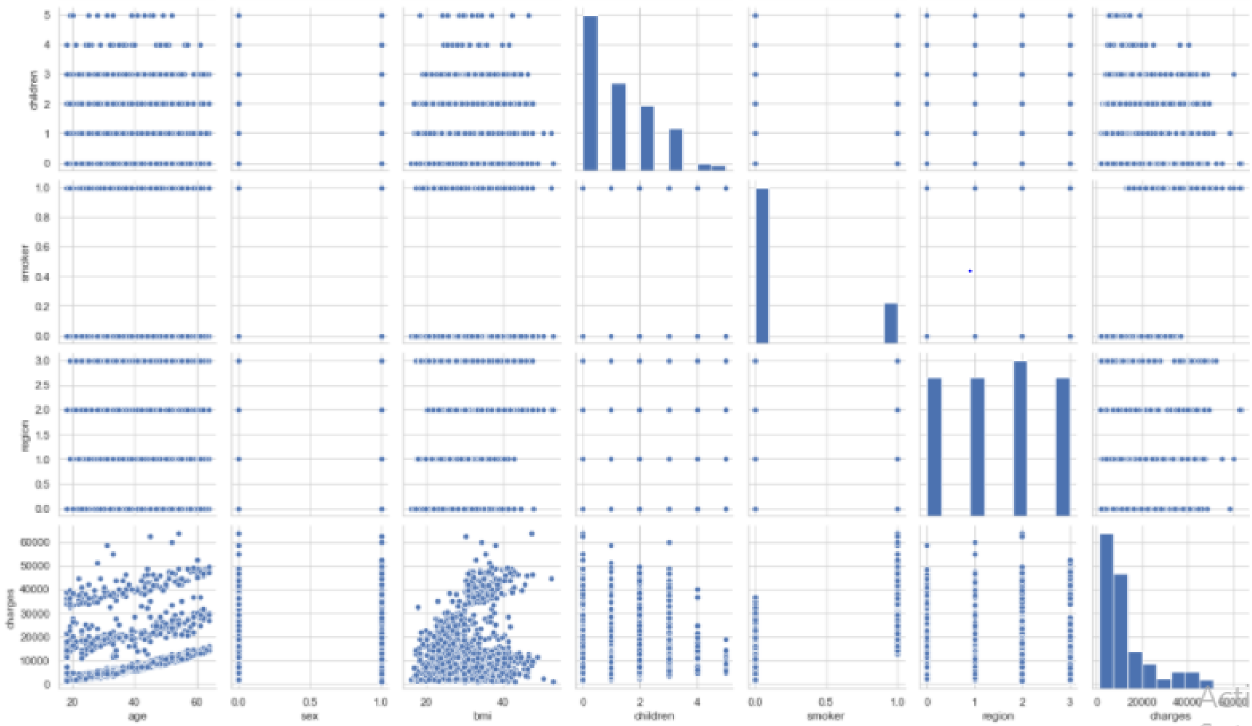
insurance_df_encoded = copy.deepcopy(df)

insurance_df_encoded.loc[:,['sex', 'smoker', 'region']] = df.loc[:,['sex', 'smoker', 'region']].apply(LabelEncoder().fit_transform)

sns.pairplot(insurance_df_encoded)

plt.show()





Q04.a. Do charges of people who smoke differ significantly from the people who don't?

df.smoker.value_counts()

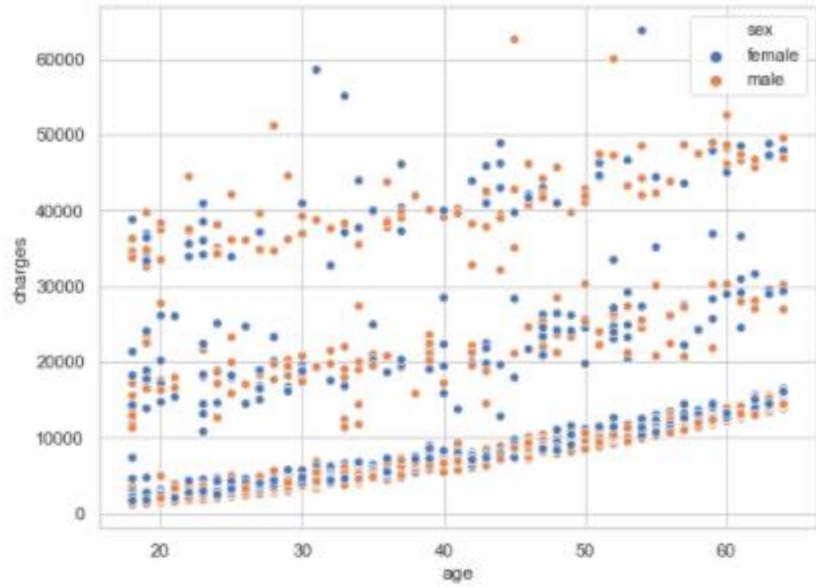
```
male      676
female    662
Name: sex, dtype: int64
```

Answer: Yes, from the scatterplot we can see that people who smoke differ significantly from the people who don't

plt.figure(figsize=(8,6))

sns.scatterplot(df.age, df.charges,hue=df.smoker,palette= ['red','green'] ,alpha=0.6)

plt.show()



T-test

s = "Same charges for smokers and non smokers"

ns = "Not same charges for smokers and non smokers"

x = np.array(df[df.smoker == 'yes'].charges)

y = np.array(df[df.smoker == 'no'].charges)

t, p_value = stats.ttest_ind(x,y, axis = 0)

print(p_value)

OUTPUT: 0.08997637178984932

Answer: as p_value<0.05, therefore people who smoke differ significantly from the people who don't

Q)4.b. Does bmi of males differ significantly from that of females?

df.sex.value_counts()

plt.figure(figsize=(8,6))

```
sns.scatterplot(df.age, df.charges,hue=df.sex  )
plt.show()

y = "bmi is affected by gender"
n = "bmi is not affected by gender"

x = np.array(df[df.sex == 'male'].bmi)
y = np.array(df[df.sex == 'female'].bmi)

t, p_value  = stats.ttest_ind(x,y, axis = 0)

print(p_value)
```

OUTPUT: 0.006548143503580696

Answer: As p_value>0.05, therefore bmi is not getting affected by gender

Q) 4.c. Is the proportion of smokers significantly different in different genders?

```
y = "smoking habits are affected by Gender"
n = "smoking habits are not affected by Gender"

crosstab = pd.crosstab(df['sex'],df['smoker'])
chi, p_value, dof, expected =  stats.chi2_contingency(crosstab)
print(p_value)
```

OUTPUT: 0.7158579926754841

Answer: Smoking habits are affected by gender.

Q)4.d. Is the distribution of bmi across women with no children, one child and two children, the same ?

Anova test

y = "Women bmi are affected by number of children"

n = "Women bmi are not affected by number of children"

df_women = copy.deepcopy(df[df['sex'] == 'female'])

zero = df_women[df_women.children == 0]['bmi']

one = df_women[df_women.children == 1]['bmi']

two = df_women[df_women.children == 2]['bmi']

f_stat, p_value = stats.f_oneway(zero,one,two)

print(p_value)

Answer: It can be inferred from null hypothesis that women bmi are not affected by number of children.

CONCLUSION:

Following can be concluded from the analysis:

1. Mostly the people are in the age group of 51 years(75%)
2. Number of children is mostly not more than 2 children with maximum of upto 5 children
3. Charges are highly skewed
4. Mostly the bmi is 34.67
5. 18years upto 27 years age group are mostly not having children.
6. People who smoke differ significantly from the people who don't
7. Smoking habits are affected by gender
8. Women bmi are not affected by number of children.