Medical Insurance Charges Prediction

In [1]:

```
import warnings
warnings.filterwarnings('ignore')

#importing the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:

In [3]:

```
med = pd.read_csv('insurance.csv')
pd.set_option('display.max_columns', None)
med.head()
```

Out[3]:

	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 [4]:

```
med.shape
```

Out[4]:

```
(1338, 7)
```

In [5]:

```
med.describe()
```

Out[5]:

	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

In [4]:

```
med.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
age         1338 non-null int64
sex         1338 non-null object
bmi         1338 non-null float64
children    1338 non-null int64
smoker      1338 non-null object
region      1338 non-null object
charges     1338 non-null float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.2+ KB
```

EDA

In [5]:

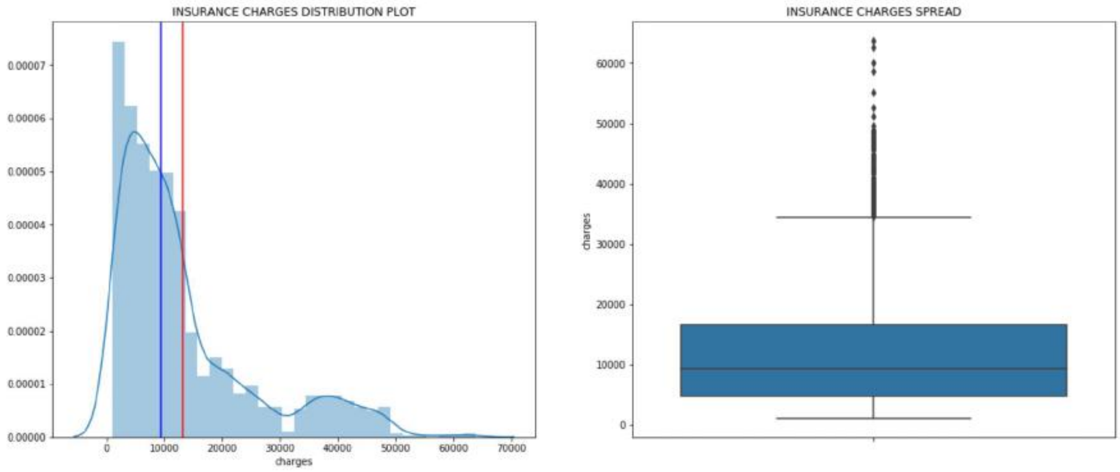
```
med.columns
```

Out[5]:

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

In [6]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('INSURANCE CHARGES DISTRIBUTION PLOT')
sns.distplot(med.charges)
plt.axvline(med.charges.mean(), color="r")
plt.axvline(med.charges.median(), color="b")

plt.subplot(1,2,2)
plt.title('INSURANCE CHARGES SPREAD')
sns.boxplot(y=med.charges)

plt.show()
```



In [7]:

```
print(med.charges.describe(percentiles = [0.25,0.50,0.75,0.85,0.90,1]))
```

```
count     1338.000000
mean     13270.422265
std      12110.011237
min       1121.873900
25%       4740.287150
50%       9382.033000
75%      16639.912515
85%      24990.166996
90%      34831.719700
100%     63770.428010
max      63770.428010
Name: charges, dtype: float64
```

In [8]:

```
print('DIFFERNCE BETWEEN MEAN AND MEDIAN :',med.charges.mean()-med.charges.median())
```

```
DIFFERNCE BETWEEN MEAN AND MEDIAN : 3888.389265141257
```

Inference :

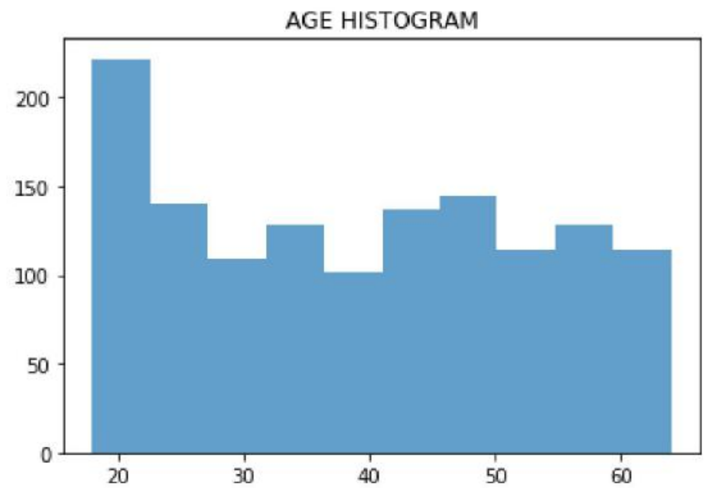
The plot seemed to be right-skewed, meaning that the most prices in the dataset are low(Below 15,000). There is a significant difference between the mean and the median of the price distribution. The data points are far spread out from the mean, which indicates a high variance in the car prices.(85% of the prices are below 18,500, whereas the remaining 15% are between 18,500 and 45,400.)

In [9]:

```
plt.title('AGE HISTOGRAM')
plt.hist(med['age'], bins=10, alpha=0.7)
```

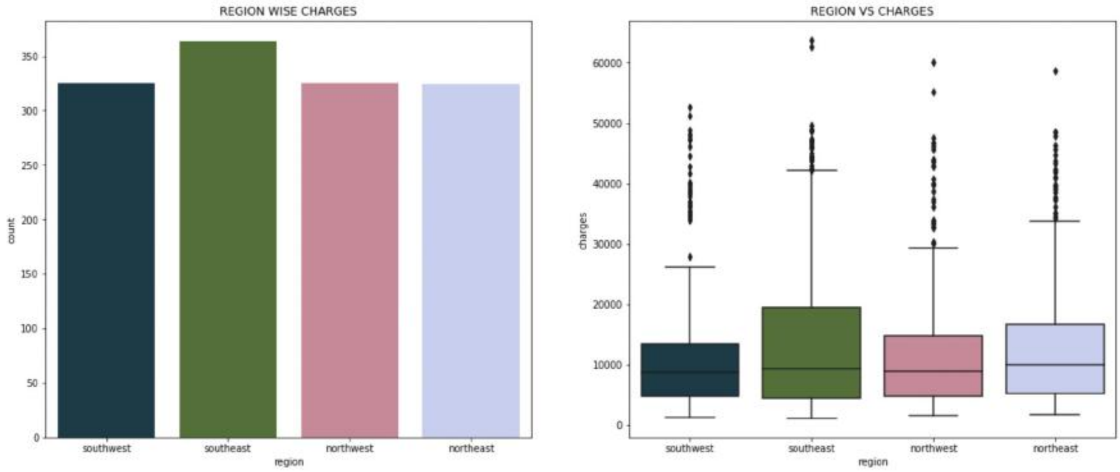
Out[9]:

```
(array([222., 140., 109., 128., 102., 137., 144., 114., 128., 114.]),
 array([18. , 22.6, 27.2, 31.8, 36.4, 41. , 45.6, 50.2, 54.8, 59.4, 64. ]),
 <a list of 10 Patch objects>)
```



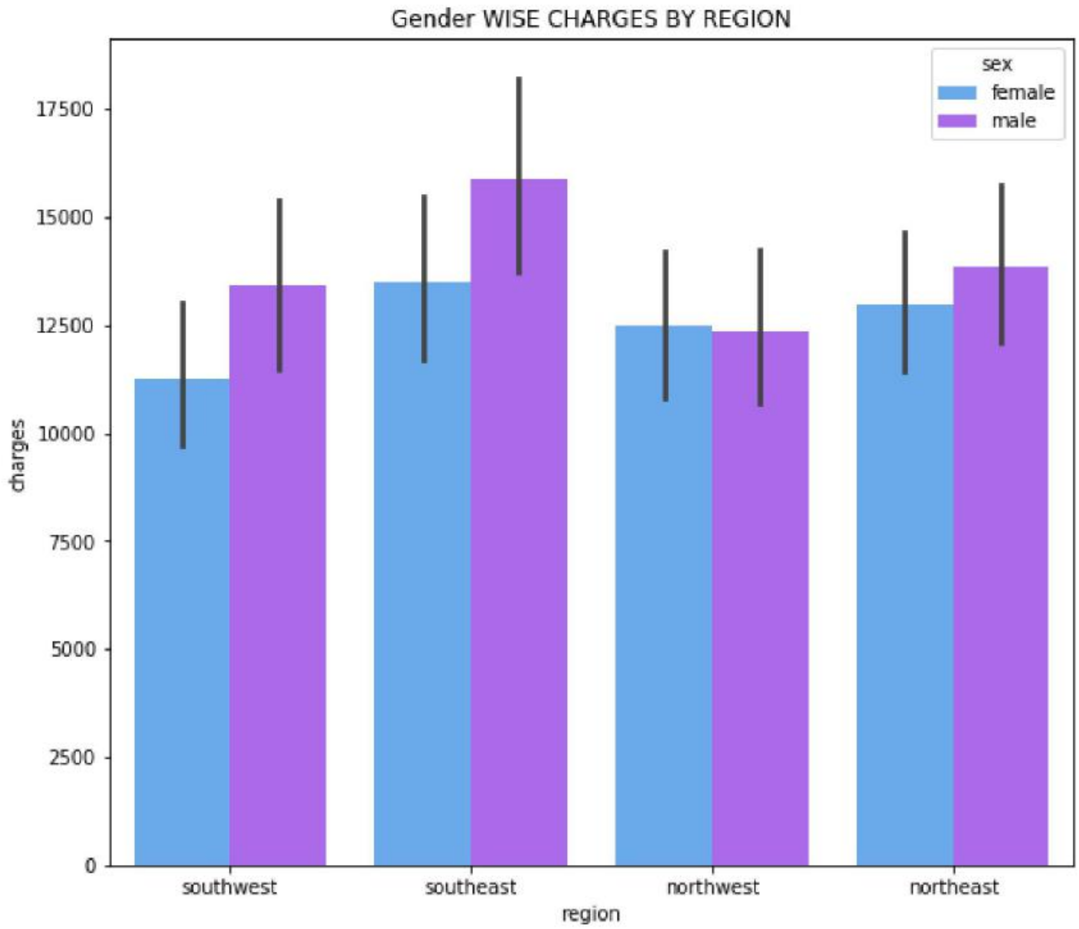
In [10]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('REGION WISE CHARGES')
sns.countplot(med.region, palette=("cubehelix"))

plt.subplot(1,2,2)
plt.title('REGION VS CHARGES')
sns.boxplot(x=med.region, y=med.charges, palette=("cubehelix"))

plt.show()
```



In [11]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('Gender WISE CHARGES BY REGION')
ax = sns.barplot(x='region', y='charges',hue='sex', data=med, palette='cool')
```

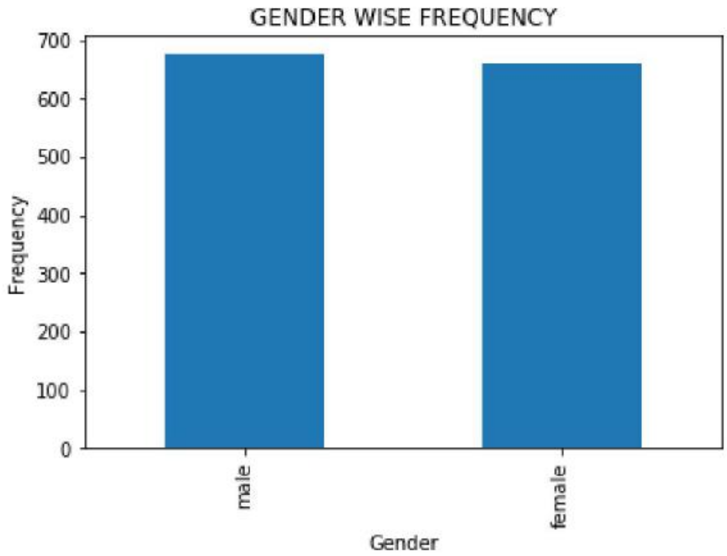


In [12]:

```
plt1 = med.sex.value_counts().plot(kind='bar')
plt.title("GENDER WISE FREQUENCY")
plt1.set(xlabel = 'Gender', ylabel='Frequency')
```

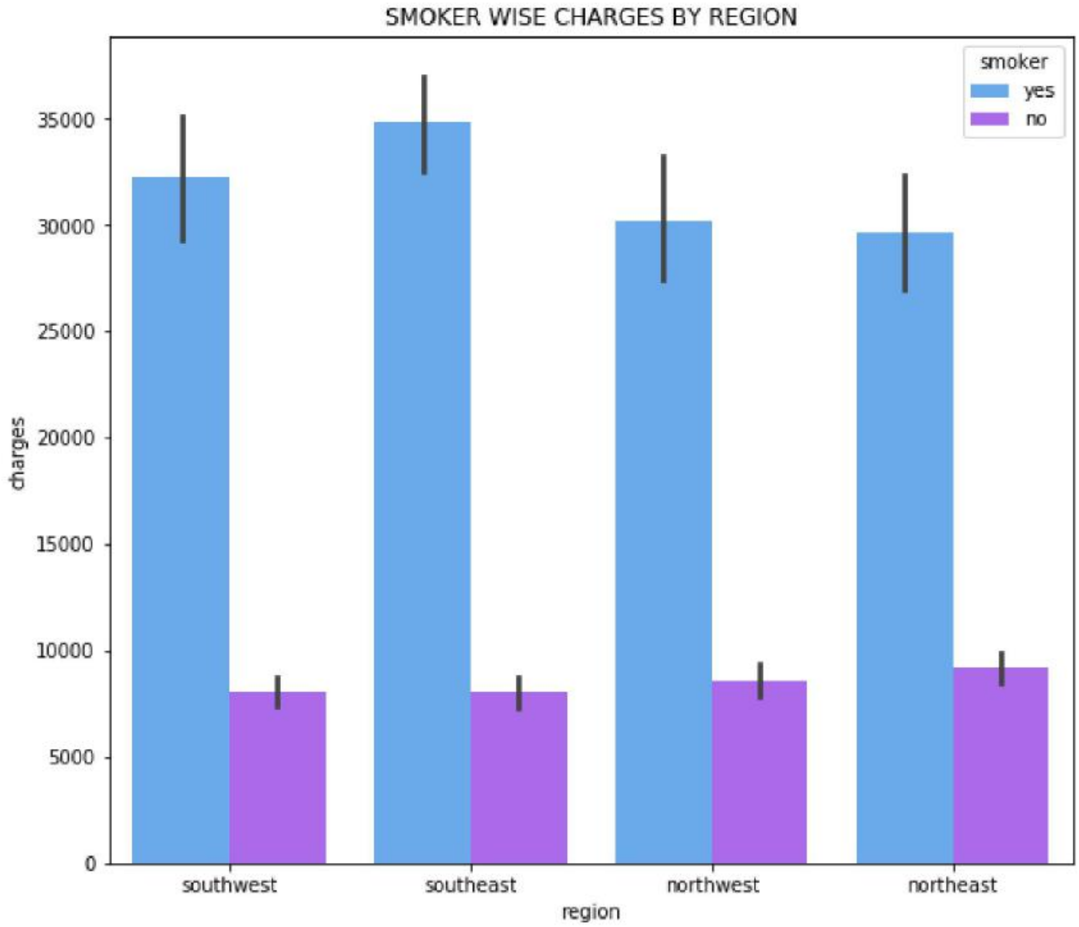
Out[12]:

```
[Text(0, 0.5, 'Frequency'), Text(0.5, 0, 'Gender')]
```



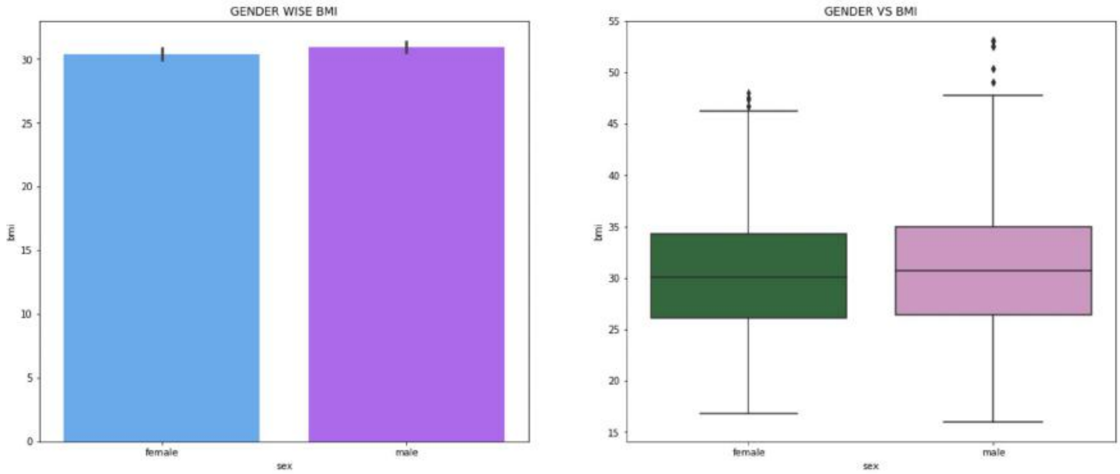
In [13]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('SMOKER WISE CHARGES BY REGION')
ax = sns.barplot(x='region', y='charges',hue='smoker', data=med, palette='cool')
```



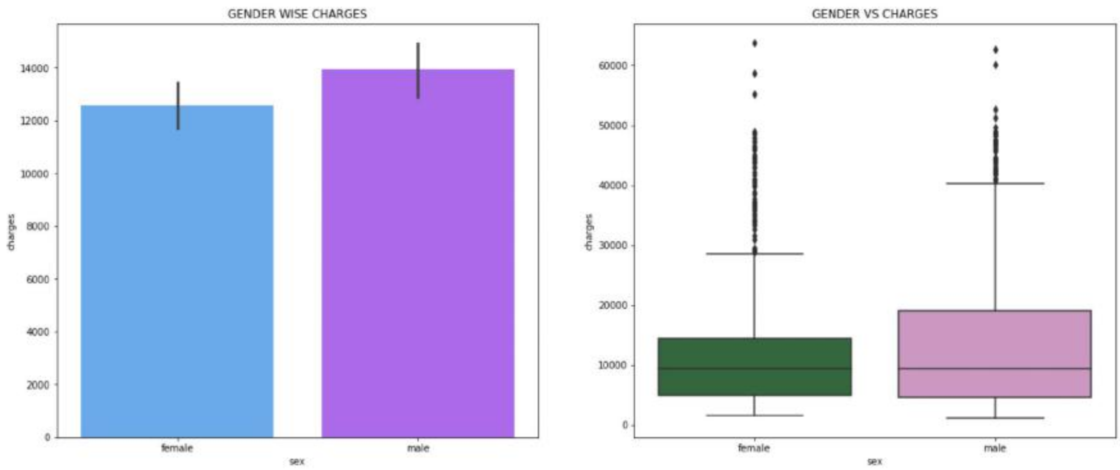
In [14]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('GENDER WISE BMI')
ax = sns.barplot(x='sex', y='bmi', data=med, palette='cool')

plt.subplot(1,2,2)
plt.title('GENDER VS BMI')
sns.boxplot(x=med.sex, y=med.bmi, palette=("cubehelix"))

plt.show()
```



In [15]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('GENDER WISE CHARGES')
ax = sns.barplot(x='sex', y='charges', data=med, palette='cool')

plt.subplot(1,2,2)
plt.title('GENDER VS CHARGES')
sns.boxplot(x=med.sex, y=med.charges, palette=("cubehelix"))

plt.show()
```

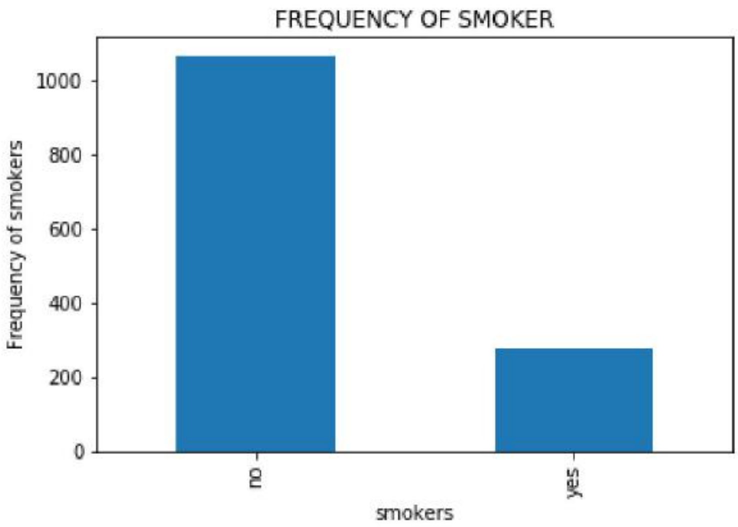


In [16]:

```
plt1 = med.smoker.value_counts().plot(kind='bar')
plt.title("FREQUENCY OF SMOKER")
plt1.set(xlabel = 'smokers', ylabel='Frequency of smokers')
```

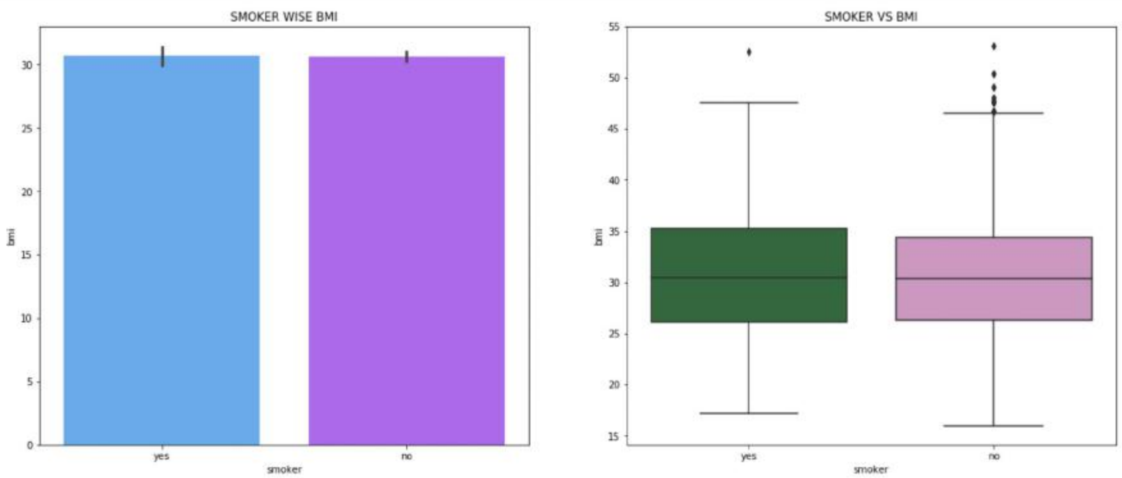
Out[16]:

```
[Text(0, 0.5, 'Frequency of smokers'), Text(0.5, 0, 'smokers')]
```



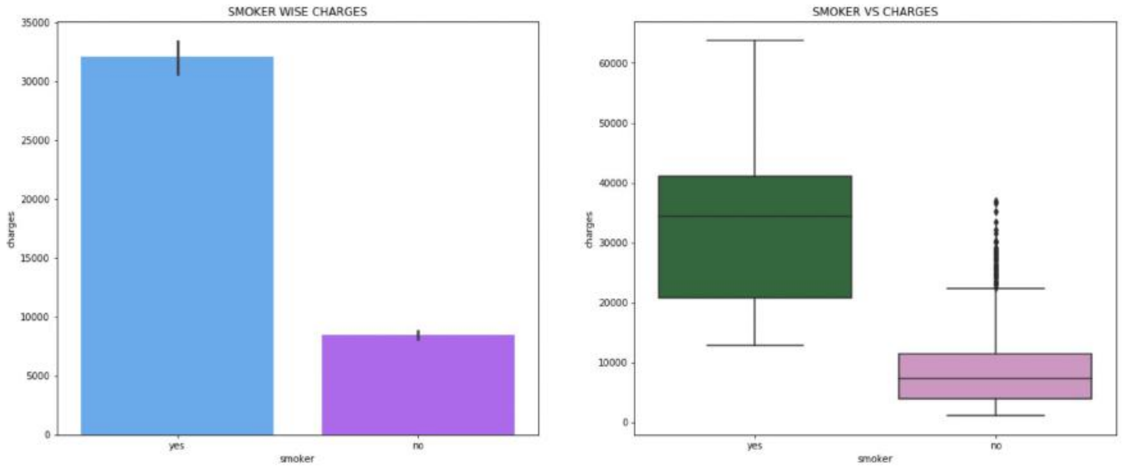
In [17]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('SMOKER WISE BMI')
ax = sns.barplot(x='smoker', y='bmi', data=med, palette='cool')

plt.subplot(1,2,2)
plt.title('SMOKER VS BMI')
sns.boxplot(x=med.smoker, y=med.bmi, palette=("cubehelix"))

plt.show()
```



In [18]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('SMOKER WISE CHARGES')
ax = sns.barplot(x='smoker', y='charges', data=med, palette='cool')

plt.subplot(1,2,2)
plt.title('SMOKER VS CHARGES')
sns.boxplot(x=med.smoker, y=med.charges, palette=("cubehelix"))

plt.show()
```

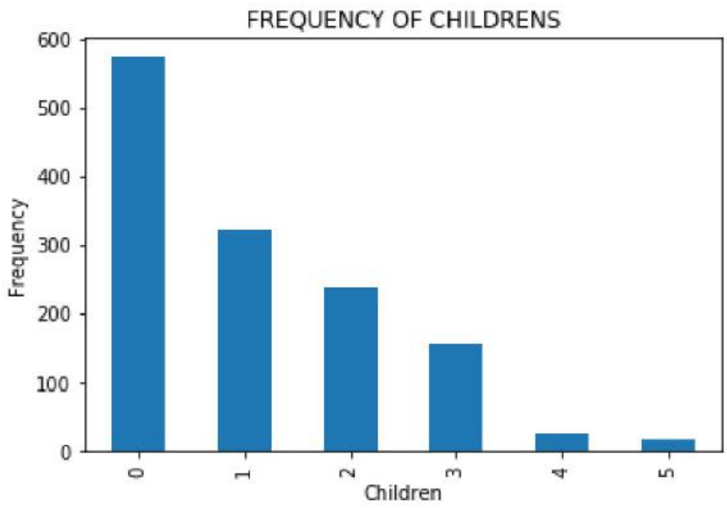


In [19]:

```
plt1 = med.children.value_counts().plot(kind='bar')
plt.title("FREQUENCY OF CHILDRENS")
plt1.set(xlabel = 'Children', ylabel='Frequency')
```

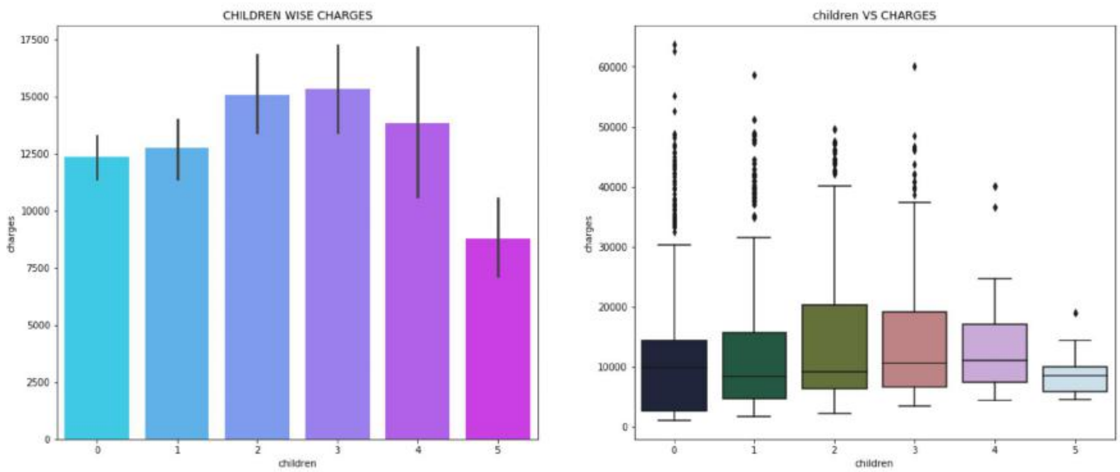
Out[19]:

```
[Text(0, 0.5, 'Frequency'), Text(0.5, 0, 'Children')]
```



In [20]:

```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
plt.title('CHILDREN WISE CHARGES')
ax = sns.barplot(x='children', y='charges', data=med, palette='cool')

plt.subplot(1,2,2)
plt.title('children VS CHARGES')
sns.boxplot(x=med.children, y=med.charges, palette=("cubehelix"))

plt.show()
```



PREPARING DATA FOR MODEL

In [21]:

```
catcols=list(med.select_dtypes(include=['object']).head())
```

In [22]:

```
med=pd.get_dummies(med, columns=catcols, drop_first=True)

med.head()
```

Out[22]:

	age	bmi	children	charges	sex_male	smoker_yes	region_northwest	region_southea
0	19	27.900	0	16884.92400	0	1	0	
1	18	33.770	1	1725.55230	1	0	0	
2	28	33.000	3	4449.46200	1	0	0	
3	33	22.705	0	21984.47061	1	0	1	
4	32	28.880	0	3866.85520	1	0	1	

In [23]:

```
ncols=['age','bmi','children','charges']
```

In [24]:

```
med.columns
```

Out[24]:

```
Index(['age', 'bmi', 'children', 'charges', 'sex_male', 'smoker_yes',
       'region_northwest', 'region_southeast', 'region_southwest'],
      dtype='object')
```

In [25]:

```
x=med.drop(['charges'], axis=1)
y=med['charges']
x.head(2)
```

Out[25]:

	age	bmi	children	sex_male	smoker_yes	region_northwest	region_southeast	region_sou
0	19	27.90	0	0	1	0	0	
1	18	33.77	1	1	0	0	1	

In [ ]:

In [26]:

```
from sklearn.model_selection import train_test_split
xt,xte,yt,yte = train_test_split(x,y, test_size = 0.3, random_state = 100)
```

In [27]:

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
num_vars = ['age','bmi']
xt[num_vars] = scaler.fit_transform(xt[num_vars])
xte[num_vars] = scaler.transform(xte[num_vars])
xte.head()
```

Out[27]:

	age	bmi	children	sex_male	smoker_yes	region_northwest	region_southeast
12	0.108696	0.496099	0	1	0	0	0
306	0.217391	0.310465	2	0	0	0	0
318	0.565217	0.314366	0	0	0	1	0
815	0.043478	0.417003	0	0	0	0	1
157	0.000000	0.247915	0	1	1	0	0

In [28]:

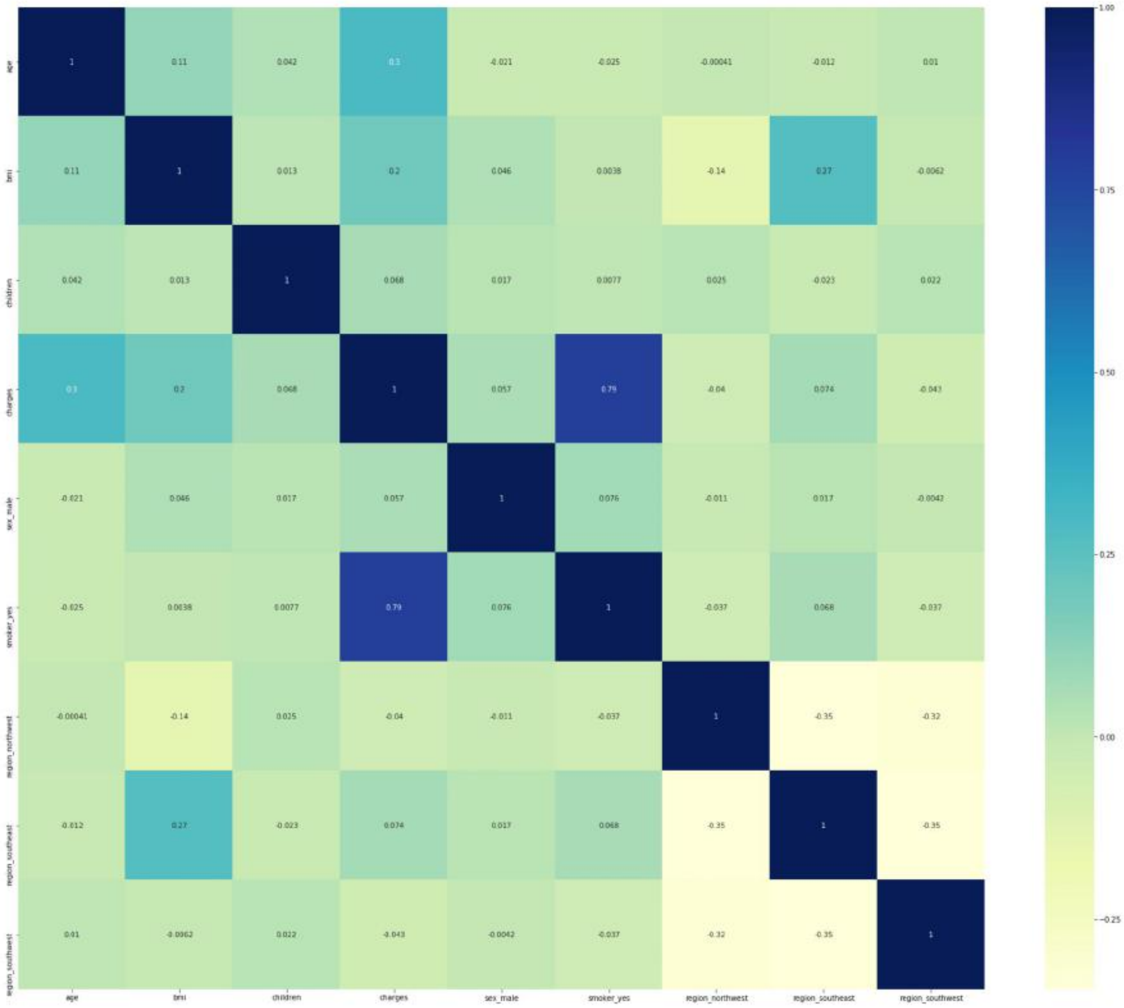
```
xt.columns
```

Out[28]:

```
Index(['age', 'bmi', 'children', 'sex_male', 'smoker_yes', 'region_northwes
t',
       'region_southeast', 'region_southwest'],
      dtype='object')
```

In [29]:

```python
#Correlation using heatmap
plt.figure(figsize = (30, 25))
sns.heatmap(med.corr(), annot = True, cmap="YlGnBu")
plt.show()
```



In [30]:

```
from sklearn.metrics import r2_score
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(xt,yt)
yp = lr.predict(xt)
```

In [31]:

```
lr.score(xt,yt)
```

Out[31]:

```
0.7378638257001522
```

In [32]:

```
r2=r2_score(yt,yp)
n=xt.shape[0]
p = xt.shape[1]
num = (1-r2)*(n-1)
den = n-p-1
ar2_train = 1 - (num/den)
ar2_train
```

Out[32]:

```
0.735601593343735
```

RFE

In [33]:

```
from sklearn.feature_selection import RFE
fe=RFE(estimator=LinearRegression(), n_features_to_select=1, step=1)
fe.fit(xt,yt)
fe.score(xte,yte)
```

Out[33]:

```
0.6401231940697194
```

In [34]:

```
xte.iloc[:,fe.support_].columns
```

Out[34]:

```
Index(['smoker_yes'], dtype='object')
```

In [35]:

```
feature_imp=pd.DataFrame({'cols':med.columns})
feature_imp
```

Out[35]:

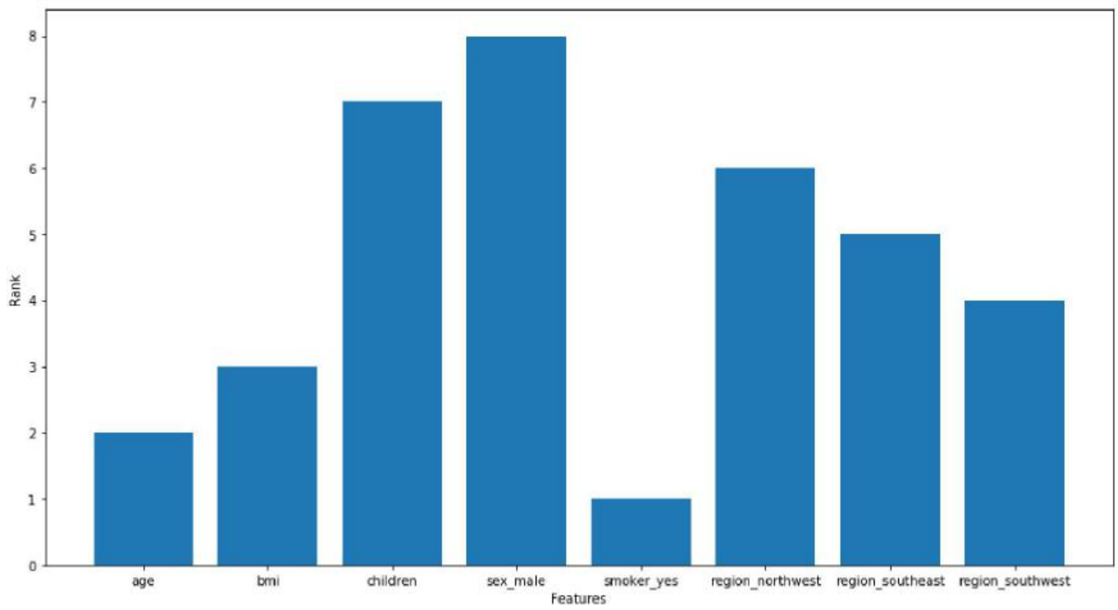
	cols
0	age
1	bmi
2	children
3	charges
4	sex_male
5	smoker_yes
6	region_northwest
7	region_southeast
8	region_southwest

In [36]:

```
plt.figure(figsize=(15,8))
plt.xlabel('Features')
plt.ylabel('Rank')
plt.bar(xt.columns,fe.ranking_)
```

Out[36]:

<BarContainer object of 8 artists>



RFECV

In [37]:

```
from sklearn.feature_selection import RFECV
fea=RFECV(estimator=LinearRegression(), min_features_to_select=1, step=1, n_jobs=-1, scorir
fea.fit(xt,yt)
fea.score(xte,yte)
```

Out[37]:

```
0.7772310511733102
```

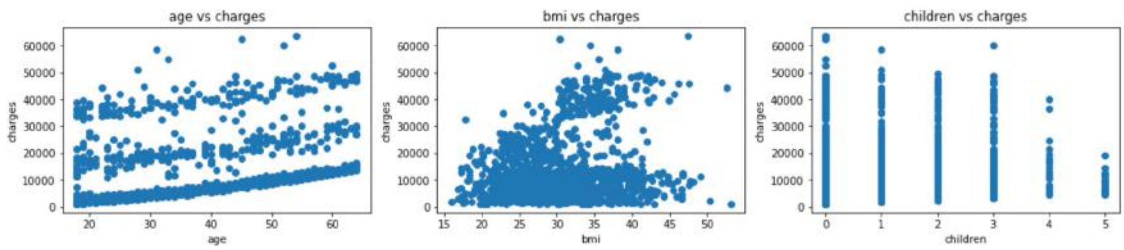
In [38]:

```
print(xt.iloc[:,fea.support_].columns)
```

```
Index(['age', 'bmi', 'children', 'sex_male', 'smoker_yes', 'region_northwes
t',
       'region_southeast', 'region_southwest'],
      dtype='object')
```

In [76]:

```
def scatter(x,fig):
    plt.subplot(5,3,fig)
    plt.scatter(med[x],med['charges'])
    plt.title(x+' vs charges')
    plt.ylabel('charges')
    plt.xlabel(x)

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

scatter('age', 1)
scatter('bmi', 2)
scatter('children', 3)


plt.tight_layout()
```



Building model using statsmodel, for the detailed statistics

In [39]:

```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

In [40]:

```
xt.columns
```

Out[40]:

```
Index(['age', 'bmi', 'children', 'sex_male', 'smoker_yes', 'region_northwes
t',
       'region_southeast', 'region_southwest'],
      dtype='object')
```

In [54]:

```
X_train_rfe = xt[xt.columns[fea.support_]]
X_train_rfe.head()
```

Out[54]:

	age	bmi	children	sex_male	smoker_yes	region_northwest	region_southeast
966	0.717391	0.237692	2	1	1	1	0
522	0.717391	0.483051	0	0	0	0	0
155	0.565217	0.633844	0	1	0	1	0
671	0.239130	0.408932	0	0	0	0	0
1173	0.434783	0.357815	2	1	0	1	0

In [55]:

```
import statsmodels.api as sm
```

In [56]:

```
def build_model(X,y):
    X = sm.add_constant(X) #Adding the constant
    lm = sm.OLS(y,X).fit() # fitting the model
    print(lm.summary()) # model summary
    return X

def checkVIF(X):
    vif = pd.DataFrame() # expty dataframe
    vif['Features'] = X.columns
    vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
    vif['VIF'] = round(vif['VIF'], 2)
    vif = vif.sort_values(by = "VIF", ascending = False)
    return(vif)
```

MODEL 1

In [57]:

```
X_train_new = build_model(X_train_rfe,yt)
```

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:                charges   R-squared:                       0.7
38
Model:                            OLS   Adj. R-squared:                  0.7
36
Method:                 Least Squares   F-statistic:                     32
6.2
Date:                Wed, 25 May 2022   Prob (F-statistic):          2.08e-2
63
Time:                        00:14:34   Log-Likelihood:                -950
3.0
No. Observations:                 936   AIC:                         1.902e+
04
Df Residuals:                     927   BIC:                         1.907e+
04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
========
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const             -955.2075    710.405     -1.345      0.179   -2349.395
438.980
age               1.196e+04    673.201     17.767      0.000    1.06e+04
1.33e+04
bmi               1.077e+04   1289.405      8.351      0.000    8237.478
1.33e+04
children           472.4266    169.779      2.783      0.006     139.231
805.622
sex_male            -0.0621    409.343     -0.000      1.000    -803.409
803.285
smoker_yes        2.399e+04    518.432     46.275      0.000     2.3e+04
2.5e+04
region_northwest  -755.4086    592.396     -1.275      0.203   -1918.000
407.183
region_southeast  -941.4032    593.535     -1.586      0.113   -2106.232
223.426
region_southwest -1601.7141    596.693     -2.684      0.007   -2772.740
-430.688
==============================================================================
==
Omnibus:                      224.957   Durbin-Watson:                   2.0
21
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              541.0
73
Skew:                           1.274   Prob(JB):                    3.22e-1
18
Kurtosis:                       5.717   Cond. No.                          1
3.2
==============================================================================
==

Warnings:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

p-value of sex_male seems to be high than significance of 0.05, therefore drop it.

In [58]:

```
X_train_new = X_train_new.drop(["sex_male"], axis = 1)
```

MODEL 2

In [59]:

```
X_train_new = build_model(X_train_new,yt)
```

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:                charges   R-squared:                       0.7
38
Model:                            OLS   Adj. R-squared:                  0.7
36
Method:                 Least Squares   F-statistic:                      37
3.2
Date:                Wed, 25 May 2022   Prob (F-statistic):          1.04e-2
64
Time:                        00:14:39   Log-Likelihood:                 -950
3.0
No. Observations:                 936   AIC:                         1.902e+
04
Df Residuals:                     928   BIC:                         1.906e+
04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
========
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const             -955.2341    687.935     -1.389      0.165   -2305.323
394.855
age               1.196e+04    672.838     17.777      0.000    1.06e+04
1.33e+04
bmi               1.077e+04   1287.671      8.362      0.000    8240.877
1.33e+04
children           472.4256    169.568      2.786      0.005     139.645
805.206
smoker_yes        2.399e+04    517.185     46.386      0.000     2.3e+04
2.5e+04
region_northwest  -755.4078    592.055     -1.276      0.202   -1917.329
406.514
region_southeast  -941.4019    593.151     -1.587      0.113   -2105.474
222.670
region_southwest -1601.7127    596.295     -2.686      0.007   -2771.957
-431.469
==============================================================================
==
Omnibus:                      224.957   Durbin-Watson:                   2.0
21
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              541.0
73
Skew:                           1.274   Prob(JB):                    3.22e-1
18
Kurtosis:                       5.717   Cond. No.                          1
2.8
==============================================================================
==

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

```
p-value of region northwest seems to be high than significance of 0.05, therefore drop
it.
```

In [60]:

```
X_train_new = X_train_new.drop(["region_northwest"], axis = 1)
```

MODEL 3

In [62]:

```
X_train_new = build_model(X_train_new,yt)
```

```
                            OLS Regression Results
============================================================================
==
Dep. Variable:                charges   R-squared:                       0.7
37
Model:                            OLS   Adj. R-squared:                  0.7
36
Method:                 Least Squares   F-statistic:                      43
4.8
Date:                Wed, 25 May 2022   Prob (F-statistic):          1.07e-2
65
Time:                        00:15:28   Log-Likelihood:                 -950
3.8
No. Observations:                 936   AIC:                         1.902e+
04
Df Residuals:                     929   BIC:                         1.906e+
04
Df Model:                           6
Covariance Type:            nonrobust
============================================================================
========
                       coef    std err          t      P>|t|      [0.025
0.975]
----------------------------------------------------------------------------
--------
const            -1328.2307    622.941     -2.132      0.033   -2550.765
-105.696
age               1.195e+04    673.001     17.755      0.000    1.06e+04
1.33e+04
bmi               1.076e+04   1288.081      8.352      0.000    8229.962
1.33e+04
children           464.6120    169.514      2.741      0.006     131.937
797.287
smoker_yes           2.4e+04   517.287     46.399      0.000      2.3e+04
2.5e+04
region_southeast  -552.8548    509.183     -1.086      0.278   -1552.136
446.427
region_southwest -1211.9786    512.291     -2.366      0.018   -2217.360
-206.597
============================================================================
==
Omnibus:                      222.427   Durbin-Watson:                   2.0
27
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              529.0
12
Skew:                           1.265   Prob(JB):                    1.34e-1
15
Kurtosis:                       5.677   Cond. No.                          1
2.8
============================================================================
==

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

p-value of region south-east seems to be high than significance of 0.05, therefore drop it.

In [64]:

```
X_train_new = X_train_new.drop(["region_southeast"], axis = 1)
```

MODEL 4

In [66]:

```
X_train_new = build_model(X_train_new,yt)
```

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:                charges   R-squared:                       0.7
37
Model:                            OLS   Adj. R-squared:                  0.7
36
Method:                 Least Squares   F-statistic:                     52
1.4
Date:                Wed, 25 May 2022   Prob (F-statistic):          8.06e-2
67
Time:                        00:21:20   Log-Likelihood:                 -950
4.4
No. Observations:                 936   AIC:                         1.902e+
04
Df Residuals:                     930   BIC:                         1.905e+
04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
========
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const            -1382.1606    621.017     -2.226      0.026   -2600.918
-163.403
age               1.199e+04    672.242     17.829      0.000    1.07e+04
1.33e+04
bmi               1.034e+04   1230.710      8.405      0.000    7929.397
1.28e+04
children           469.7370    169.465      2.772      0.006     137.159
802.315
smoker_yes        2.395e+04    515.555     46.464      0.000    2.29e+04
2.5e+04
region_southwest -1006.9214    476.254     -2.114      0.035   -1941.579
-72.264
==============================================================================
==
Omnibus:                      221.599   Durbin-Watson:                   2.0
25
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              525.2
36
Skew:                           1.262   Prob(JB):                    8.84e-1
15
Kurtosis:                       5.665   Cond. No.                          1
2.2
==============================================================================
==

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

In [67]:

```
checkVIF(X_train_new)
```

Out[67]:

	Features	VIF
0	const	9.27
1	age	1.02
2	bmi	1.02
3	children	1.00
4	smoker_yes	1.00
5	region_southwest	1.00

In [101]:

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:                charges   R-squared:                       0.7
38
Model:                            OLS   Adj. R-squared:                  0.7
36
Method:                 Least Squares   F-statistic:                      37
3.2
Date:                Tue, 24 May 2022   Prob (F-statistic):          1.04e-2
64
Time:                        02:11:27   Log-Likelihood:                 -950
3.0
No. Observations:                 936   AIC:                         1.902e+
04
Df Residuals:                     928   BIC:                         1.906e+
04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
========
                       coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const             -955.2341    687.935     -1.389      0.165   -2305.323
394.855
age               1.196e+04    672.838     17.777      0.000    1.06e+04
1.33e+04
bmi               1.077e+04   1287.671      8.362      0.000    8240.877
1.33e+04
children           472.4256    169.568      2.786      0.005     139.645
805.206
smoker_yes        2.399e+04    517.185     46.386      0.000     2.3e+04
2.5e+04
region_northwest  -755.4078    592.055     -1.276      0.202   -1917.329
406.514
region_southeast  -941.4019    593.151     -1.587      0.113   -2105.474
222.670
region_southwest -1601.7127    596.295     -2.686      0.007   -2771.957
-431.469
==============================================================================
==
Omnibus:                      224.957   Durbin-Watson:                   2.0
21
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              541.0
73
Skew:                           1.274   Prob(JB):                      3.22e-1
18
Kurtosis:                       5.717   Cond. No.                          1
2.8
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
```

In [102]:

```
X_train_new = X_train_new.drop(["region_northwest"], axis = 1)
```

In [103]:

```
checkVIF(X_train_new)
```

Out[103]:

	Features	VIF
0	const	9.33
5	region_southeast	1.26
6	region_southwest	1.16
2	bmi	1.11
1	age	1.02
4	smoker_yes	1.01
3	children	1.00

In [68]:

```
X_train_new = X_train_new.drop(["const"], axis = 1)
```

In [69]:

```
checkVIF(X_train_new)
```

Out[69]:

	Features	VIF
1	bmi	3.30
0	age	2.80
2	children	1.73
4	region_southwest	1.29
3	smoker_yes	1.21

Residual Analysis of Model

In [70]:

```
lm = sm.OLS(yt,X_train_new).fit()
y_train_price = lm.predict(X_train_new)
```

In [71]:

```
lm.summary()
```

Out[71]:

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.879
Model:	OLS	**Adj. R-squared:**	0.878
Method:	Least Squares	**F-statistic:**	1349.
Date:	Wed, 25 May 2022	**Prob (F-statistic):**	0.00
Time:	00:26:17	**Log-Likelihood:**	-9506.9
No. Observations:	936	**AIC:**	1.902e+04
Df Residuals:	931	**BIC:**	1.905e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
age	1.14e+04	620.736	18.372	0.000	1.02e+04	1.26e+04
bmi	8377.6978	858.303	9.761	0.000	6693.265	1.01e+04
children	375.5487	164.444	2.284	0.023	52.825	698.272
smoker_yes	2.377e+04	510.212	46.597	0.000	2.28e+04	2.48e+04
region_southwest	-1183.4761	470.597	-2.515	0.012	-2107.030	-259.922

Omnibus:	223.403	**Durbin-Watson:**	2.033
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	530.199
Skew:	1.271	**Prob(JB):**	7.39e-116
Kurtosis:	5.670	**Cond. No.**	8.13

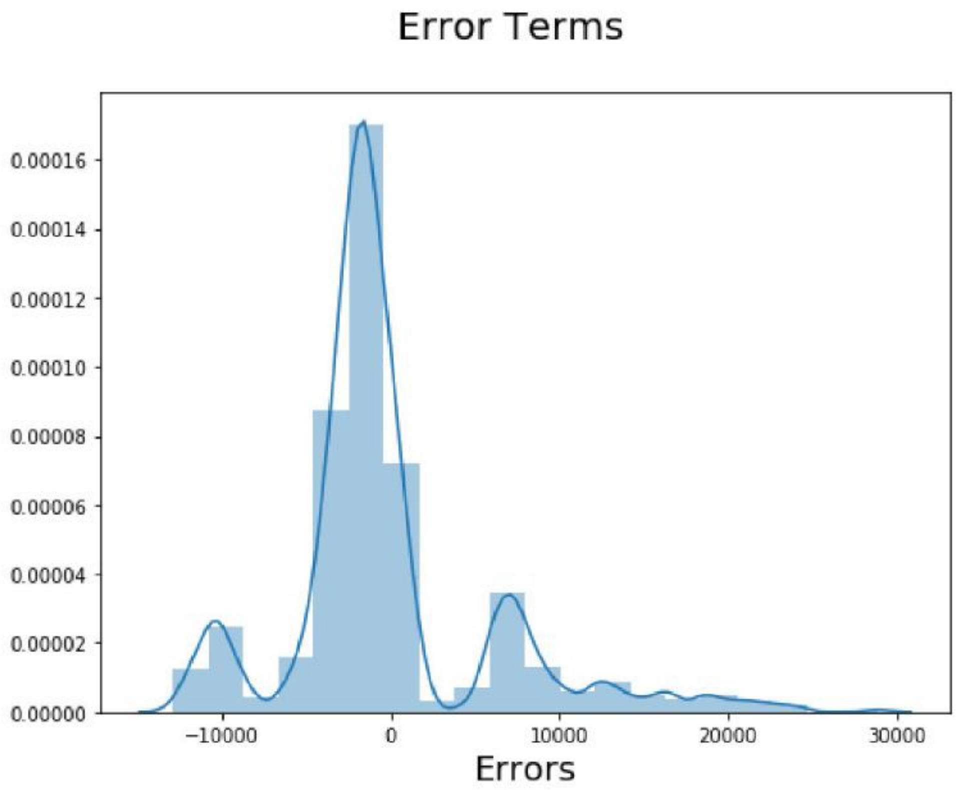
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [73]:

```
# Plot the histogram of the error terms
fig = plt.figure(figsize=(8,6))
sns.distplot((yt - y_train_price), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)
```

Out[73]:

```
Text(0.5, 0, 'Errors')
```



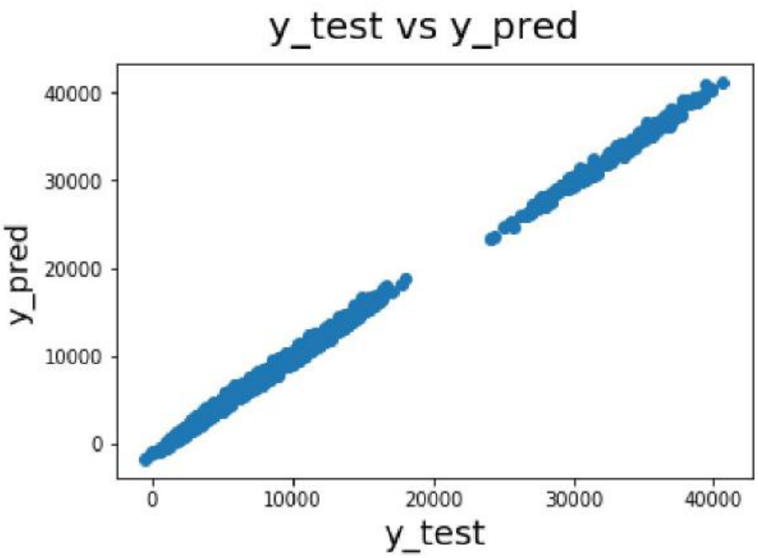
Error terms seem to be approximately normally distributed so the assumption on the linear modeling seems to be fulfilled.

In [76]:

```
fig = plt.figure()
plt.scatter(y_train_price,yp)
fig.suptitle('y_test vs y_pred', fontsize=20)            # Plot heading
plt.xlabel('y_test', fontsize=18)                        # X-label
plt.ylabel('y_pred', fontsize=16)
```

Out[76]:

```
Text(0, 0.5, 'y_pred')
```



CONCLUSION:

There are 3 factors that affects insurance charges

1. smoker
2. age
3. bmi As per result of rfe when n_features was 1 accuracy score was 64% so, we can say that smoking is the greatest factor that affects medical cost charges, then it's bmi and age.

In [ ]: