Objective: The objective of this analysis is to determine whether smokers have statistically higher mean individual medical costs billed by health insurance than do non-smokers. Furthermore, is a person's BMI correlated with individual medical costs billed by health insurance?

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import kurtosis, skew, stats
from math import sqrt
from numpy import mean, var
```

In [2]:
```
data = pd.read_csv("../input/insurance2.csv")
print(data.head())
```

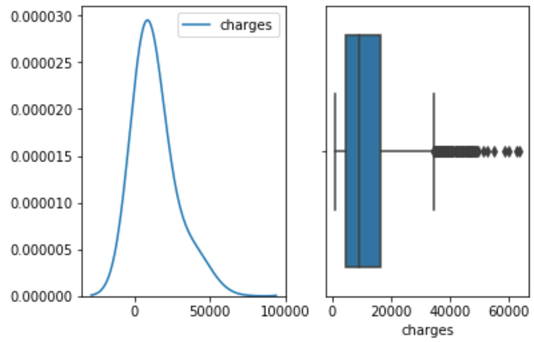
```
   age  sex     bmi       ...        region      charges  insurancecla
im
0   19    0  27.900       ...             3  16884.92400
1
1   18    1  33.770       ...             2   1725.55230
1
2   28    1  33.000       ...             2   4449.46200
0
3   33    1  22.705       ...             1  21984.47061
0
4   32    1  28.880       ...             1   3866.85520
1

[5 rows x 8 columns]
```

In [3]:
```
print("Summary Statistics of Medical Costs")
print(data['charges'].describe())
print("skew:  {}".format(skew(data['charges'])))
print("kurtosis:  {}".format(kurtosis(data['charges'])))
print("missing charges values: {}".format(data['charges'].isnull().sum
()))
print("missing smoker values: {}".format(data['smoker'].isnull().sum
()))
```

```
Summary Statistics of Medical Costs
count     1338.000000
mean     13270.422265
std      12110.011237
min       1121.873900
25%       4740.287150
50%       9382.033000
75%      16639.912515
max      63770.428010
Name: charges, dtype: float64
skew:  1.5141797118745743
kurtosis:  1.595821363956751
missing charges values: 0
missing smoker values: 0
```

In [4]:
```
f, axes = plt.subplots(1, 2)
sns.kdeplot(data['charges'], bw=10000, ax=axes[0])
```

```
sns.boxplot(data['charges'], ax=axes[1])
plt.show()
```



There are 1338 observations in this dataset. Both the boxplot and kernel density estimation plot reveal that the charges data is right skewed. Furthermore, there are some outliers but no missing charges and smoker values.

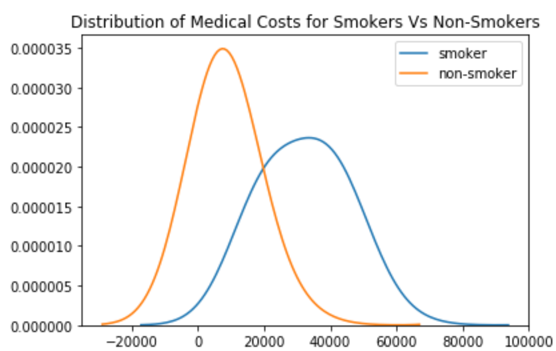
Objective Part 1: Do smokers have statistically higher mean individual medical costs billed by health insurance than do non-smokers?

In [5]:

```
#prepare our 2 groups to test
smoker = data[data['smoker']==1]
non_smoker = data[data['smoker']==0]
```

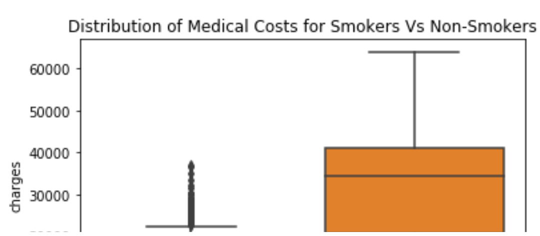
In [6]:

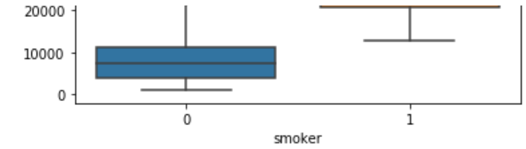
```
plt.title('Distribution of Medical Costs for Smokers Vs Non-Smokers')
ax = sns.kdeplot(smoker['charges'], bw=10000, label='smoker')
ax = sns.kdeplot(non_smoker['charges'], bw=10000, label='non-smoker')
plt.show()
```



In [7]:

```
plt.title('Distribution of Medical Costs for Smokers Vs Non-Smokers')
ax = sns.boxplot(x="smoker", y="charges", data=data)
```





The boxplots and kernel density estimation plots reveal that the 2 datasets are likely different.

In [8]:

```
statistic, pvalue = stats.ttest_ind(non_smoker['charges'], smoker['charges'], equal_var = False)
print("2 sample, 2 sided t-test pvalue:  {} t-stat: {}".format(pvalue, statistic))
```

```
2 sample, 2 sided t-test pvalue:  5.88946444671698e-103 t-stat: -32.751887766341824
```

In [9]:

```
# function to calculate Cohen's d for independent samples
def cohend(d1, d2):
        n1, n2 = len(d1), len(d2)
        s1, s2 = var(d1, ddof=1), var(d2, ddof=1)
        s = sqrt(((n1 - 1) * s1 + (n2 - 1) * s2) / (n1 + n2 - 2))
        u1, u2 = mean(d1), mean(d2)
        return (u1 - u2) / s

d = cohend(smoker['charges'], non_smoker['charges'])
print("cohen's d:  {}".format(d))
```

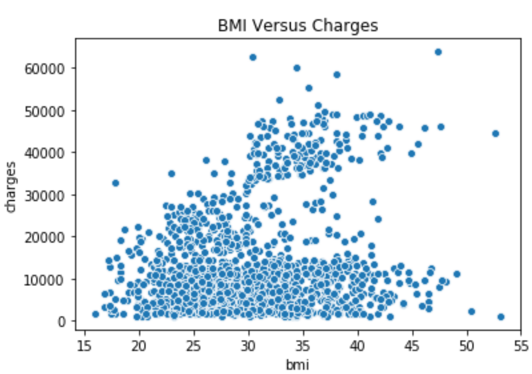
```
cohen's d:  3.1613494007377874
```

Results from the 2 sample, 2 sided t test indicate that non-smokers have significantly less mean individual medical costs billed by health insurance than do smokers. Furthermore, Cohen's D indicates that the difference between the means is more than 3 standard deviations which is interpreted as a large effect size.

Objective Part 2: Is a person's BMI correlated with individual medical costs billed by health insurance?

In [10]:

```
plt.title("BMI Versus Charges")
ax = sns.scatterplot(x="bmi", y="charges", data=data)
plt.show()
```



In [11]:
```
data.bmi.corr(data.charges)
```

Out[11]:
```
0.19834096883362895
```

The scatterplot and correlation coefficient both reveal that bmi and charges have a very weak correlation. However, for charges larger than a specified amount, there might be a stronger correlation.

In [12]:
```
def corr_converge(data=data):
    for i in range(0,60000,1000):
        data_new = data[data['charges'] >= i]
        print("lower bound: {} \t correlation coefficient: {} \t numbe
r of observations: {}".format(i,data_new.bmi.corr(data_new.charges),le
n(data_new)))
        pass

corr_converge()
```

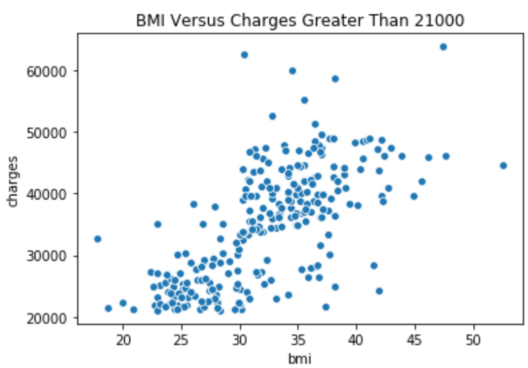
```
lower bound: 0   correlation coefficient: 0.19834096883362895    numbe
r of observations: 1338
lower bound: 1000       correlation coefficient: 0.19834096883362895
number of observations: 1338
lower bound: 2000       correlation coefficient: 0.20716424638136222
number of observations: 1246
lower bound: 3000       correlation coefficient: 0.21031560947020433
number of observations: 1147
lower bound: 4000       correlation coefficient: 0.21798470528808944
number of observations: 1069
lower bound: 5000       correlation coefficient: 0.21979606223533085
number of observations: 979
lower bound: 6000       correlation coefficient: 0.23402630972840624
number of observations: 909
lower bound: 7000       correlation coefficient: 0.24403352661404867
number of observations: 836
lower bound: 8000       correlation coefficient: 0.2542746105436367
number of observations: 768
lower bound: 9000       correlation coefficient: 0.267419418610242
number of observations: 690
lower bound: 10000      correlation coefficient: 0.30737043088673865
number of observations: 626
lower bound: 11000      correlation coefficient: 0.3423437825299707
number of observations: 566
lower bound: 12000      correlation coefficient: 0.4149410244586434
number of observations: 492
lower bound: 13000      correlation coefficient: 0.4885633430081083
number of observations: 435
lower bound: 14000      correlation coefficient: 0.556840003378273
number of observations: 390
lower bound: 15000      correlation coefficient: 0.6064606850640031
number of observations: 358
lower bound: 16000      correlation coefficient: 0.6320465902809337
number of observations: 346
lower bound: 17000      correlation coefficient: 0.6309299945456162
number of observations: 331
lower bound: 18000      correlation coefficient: 0.6309291125552199
number of observations: 313
lower bound: 19000      correlation coefficient: 0.6379910195848879
number of observations: 294
lower bound: 20000      correlation coefficient: 0.6444805352189786
number of observations: 273
```

```
lower bound: 21000      correlation coefficient: 0.6724519995614608
number of observations: 257
lower bound: 22000      correlation coefficient: 0.6429706743171245
number of observations: 239
lower bound: 23000      correlation coefficient: 0.6203827827986519
number of observations: 230
lower bound: 24000      correlation coefficient: 0.5914088710750379
number of observations: 217
lower bound: 25000      correlation coefficient: 0.5715683411333236
number of observations: 201
lower bound: 26000      correlation coefficient: 0.5328204763270952
number of observations: 193
lower bound: 27000      correlation coefficient: 0.5134706105380401
number of observations: 185
lower bound: 28000      correlation coefficient: 0.4750099406172605
number of observations: 174
lower bound: 29000      correlation coefficient: 0.4732837998538616
number of observations: 166
lower bound: 30000      correlation coefficient: 0.44738622619986773
number of observations: 162
lower bound: 31000      correlation coefficient: 0.4113823709015129
number of observations: 156
lower bound: 32000      correlation coefficient: 0.4192112299987014
number of observations: 155
lower bound: 33000      correlation coefficient: 0.3821221444535294
number of observations: 151
lower bound: 34000      correlation coefficient: 0.3539673825033866
number of observations: 144
lower bound: 35000      correlation coefficient: 0.3106595936673613
number of observations: 133
lower bound: 36000      correlation coefficient: 0.2474415793187489
number of observations: 127
lower bound: 37000      correlation coefficient: 0.22028577887801656
number of observations: 113
lower bound: 38000      correlation coefficient: 0.17956413779212585
number of observations: 103
lower bound: 39000      correlation coefficient: 0.17336244477878335
number of observations: 92
lower bound: 40000      correlation coefficient: 0.14686236534498467
number of observations: 79
lower bound: 41000      correlation coefficient: 0.11546640502754943
number of observations: 69
lower bound: 42000      correlation coefficient: 0.050027724577297025
number of observations: 62
lower bound: 43000      correlation coefficient: 0.04029457066125254
number of observations: 52
lower bound: 44000      correlation coefficient: -0.039476305494083805
       number of observations: 45
lower bound: 45000      correlation coefficient: -0.016195318940419645
       number of observations: 38
lower bound: 46000      correlation coefficient: -0.019761399614851946
       number of observations: 34
lower bound: 47000      correlation coefficient: 0.006696430167800046
number of observations: 25
lower bound: 48000      correlation coefficient: -0.14355776836491949
number of observations: 16
lower bound: 49000      correlation coefficient: 0.2717144794403384
number of observations: 8
lower bound: 50000      correlation coefficient: 0.34568453000295046
number of observations: 7
lower bound: 51000      correlation coefficient: 0.34568453000295046
number of observations: 7
lower bound: 52000      correlation coefficient: 0.42558411114120265
number of observations: 6
```

```
lower bound: 53000       correlation coefficient: 0.31694249379591577
number of observations: 5
lower bound: 54000       correlation coefficient: 0.31694249379591577
number of observations: 5
lower bound: 55000       correlation coefficient: 0.31694249379591577
number of observations: 5
lower bound: 56000       correlation coefficient: 0.33867302539091126
number of observations: 4
lower bound: 57000       correlation coefficient: 0.33867302539091126
number of observations: 4
lower bound: 58000       correlation coefficient: 0.33867302539091126
number of observations: 4
lower bound: 59000       correlation coefficient: 0.5660888211129969
number of observations: 3
```

In [13]:

```
data_new = data[data['charges']>=21000]
plt.title("BMI Versus Charges Greater Than 21000")
ax = sns.scatterplot(x="bmi", y="charges", data=data_new)
plt.show()
```



In [14]:

```
data_new.bmi.corr(data_new.charges)
```

Out[14]:

```
0.6724519995614608
```

After examining the convergence of correlation coefficients, I looked at charges larger than 21,000 USD. The

Results: Smokers have statistically higher mean individual medical costs billed by health insurance than do

Sample Insurance Claim Prediction Dataset

- insurance2.csv 8 columns
- insurance3r2.csv 9 columns

Sample Insurance Claim Prediction Dataset

Last Updated: a year ago (Version 1 of 2)

About this Dataset

Content

This is "Sample Insurance Claim Prediction Dataset" which based on "[Medical Cost Personal Datasets][1]" to update sample value on top.

age : age of policyholder sex: gender of policy holder (female=0, male=1) 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 25 steps: average walking steps per day of policyholder children: number of children / dependents of policyholder smoker: smoking state of policyholder (non-smoke=0;smoker=1) region: the residential area of policyholder in the US (northeast=0, northwest=1, southeast=2, southwest=3) charges: individual medical costs billed by health insurance insuranceclaim: yes=1, no=0

Comments (0)

Please sign in to leave a comment.

Notebook Data Comments

Our Team Terms Privacy Contact/Support