Project Three

Dylan Kruse (djk2382)

5/6/2021

Introduction

In [1]:

```
# Running this chunk lets you have multiple outputs from a single chunk
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

In [2]:

```
# Import packages
import numpy as np
import pandas as pd
import seaborn as sns
import scipy.stats as stats
import matplotlib.pyplot as plt
```

Features of the Dataset

In [3]:

```
# open insurance dataset
insurance = pd.read_csv("insurance.csv")
# head the dataset
insurance.info()
# recieve information from the dataset
insurance.head()
```

```
<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
```

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

Within the *insurance* dataset, there are 1338 observations each with seven different variables. The numeric variables in the *insurance* dataset are age (years), BMI (kg/m^2), children (number), and annual insurance charges ($). The categorical variables in the *insurance* dataset are sex, smoking status, and region of inhabitance.

Explanatory Data Analysis

In [4]:
```
# use describe to return summary statistics for the variables
insurance.describe()
```

Out[4]:

	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

The mean values for age, BMI, number of children, and annual insurance charges are 39.21 years old, 30.66 kg/m^2, 1.10 children, and (
)13, 270.42.*Thehighestamountofmoneyanindividualpayedfortheirannualinsurancechargesinthisdatase*
63,770.43. The oldest individual included in this dataset was 64 years old.

In [5]:
```
# describe the sex variable
insurance['sex'].value_counts()
# describe the smoker variable
insurance['smoker'].value_counts()
# describe the region variable
insurance['region'].value_counts()
```

Out[5]:
```
male      676
female    662
Name: sex, dtype: int64
```

Out[5]:
```
no     1064
yes     274
Name: smoker, dtype: int64
```

Out[5]:
```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

Within the *insurance* dataset, there are 676 men and 662 women. Of the 1338 individuals, there are 1064 who reported not smoking and 274 reported being a smoker. The number of observations in the southeast, northwest, southwest, and northeast regions are 364, 325, 325, and 324 respectively.

```
In [6]: # find descriptive statistics for the annual insurance charges variable
        insurance['charges'].describe()
        # create a histogram for the annual insurance charges variable
        insurance['charges'].plot(kind = "hist")
        plt.title("Distribution of Annual Insurance Charges in Dollars")
        plt.xlabel("Price ($)")
        # find IQR
        16639.912515 - 4740.287150
```

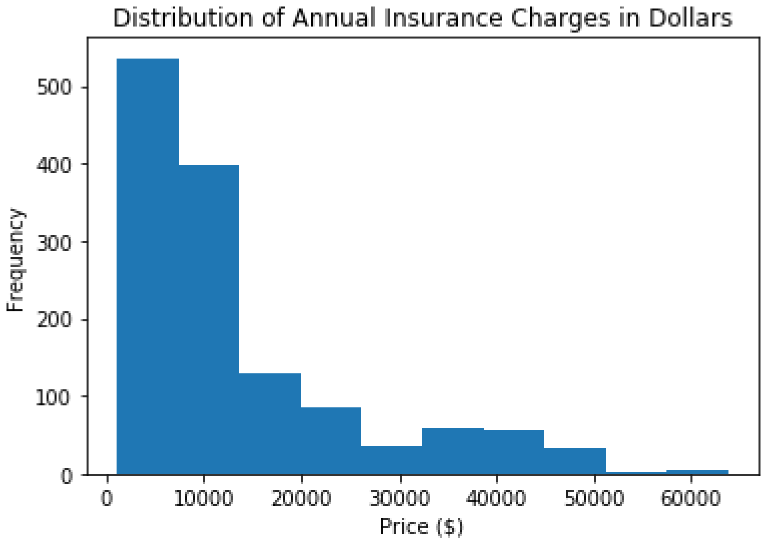
```
Out[6]: count     1338.000000
        mean     13270.422265
        std      12110.011237
        min       1121.873900
        25%       4740.287150
        50%       9382.033000
        75%      16639.912515
        max      63770.428010
        Name: charges, dtype: float64

Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7efeba26bf60>

Out[6]: Text(0.5,1,'Distribution of Annual Insurance Charges in Dollars')

Out[6]: Text(0.5,0,'Price ($)')

Out[6]: 11899.625365
```



As previously discussed, the mean annual insurance cost in the *insurance* dataset is ($)13, 270.42.Asshownbytheabovehistogramofannualinsurancecost, however, thedistributionispositively$ $anappropriatemeasureofcentralityforthisvariablewouldbethemedianwhichis($ 9,382.03. The interquartile range for the annual insurance charge variable is ($) 11,899.63.

In [7]:
```
# find counts for smoker variable
insurance['smoker'].value_counts()
# find proportions of smokers and non-smokers
1064 / (1064 + 274)
274 / (1064 + 274)
# create a pie chart to illustrate the proportion of smokers in dataset
insurance['smoker'].value_counts() \
.plot(kind = "pie") \
.axis('equal') # equal aspect ratio
plt.title("Proportions of Smokers and Non-Smokers")
```

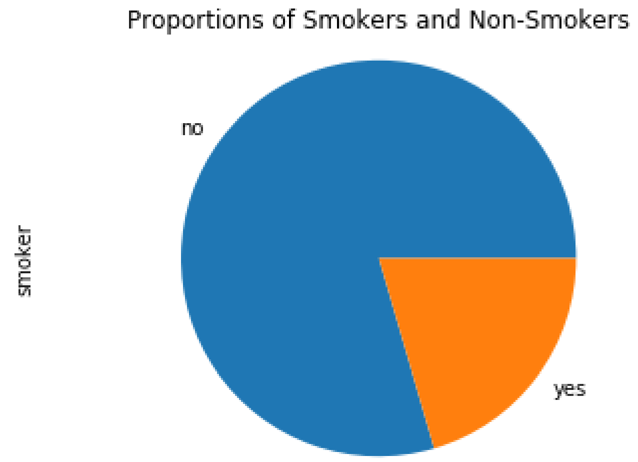
Out[7]:
```
no     1064
yes     274
Name: smoker, dtype: int64
```

Out[7]:
```
0.7952167414050823
```

Out[7]:
```
0.20478325859491778
```

Out[7]:
```
(-1.1018718496698334,
 1.1000891356985636,
 -1.1027154146109623,
 1.1011093550812814)
```

Out[7]:
```
Text(0.5,1,'Proportions of Smokers and Non-Smokers')
```



Within the dataset, 1064 individuals identified themselves as non-smokers while 274 identified themselves as smokers. These counts correlate to proportions of 79.52 % non-smokers and 20.48 % smokers. The above pie-chart illustrates these proportions.

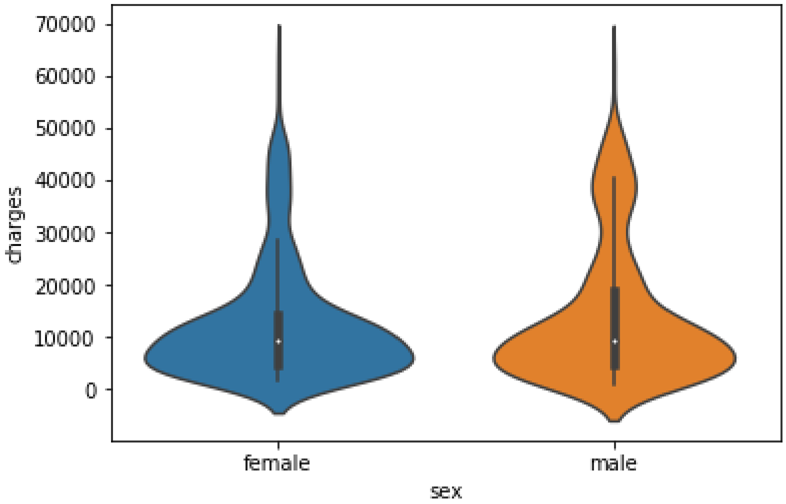
Hypothesis Test

Let's determine if there appears to be a significant difference in the annual insurance charges between men and women. To do so, a two-sample t test will be conducted. The null hypothesis in this test will be that the mean annual insurance charges are the same for men and women. The alternative hypothesis in this test will be that the mean annual insurance charges are different for men and women.

```
In [8]: # create violin plot to assess assumptions
        sns.violinplot(data = insurance, x = "sex", y = "charges")
        # run two sample t-test
        stats.ttest_ind(insurance['charges'][insurance['sex'] == 'male'],
                        insurance['charges'][insurance['sex'] == 'female'])
```

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7efeb811c358>
```

```
Out[8]: Ttest_indResult(statistic=2.097546590051688, pvalue=0.0361327210059297
        6)
```



After creating a violin plot to assess the test's assumptions, it was determined that each distribution was positively skewed indicating that the normality assumption was possibly violated. Nonetheless, after conducting the two-sample t test, it was determined that there was a significant difference in the mean annual insurance charges between men and women (p-val = 0.036).