Project #3: Python EDA with Vietnam Insurance Dataset

Hieu Pham

5/10/2021

The "Vietnam Insurance" Dataset was drawn from the vincentarebundock website. The data was collected from a cross sectional study which observed from a number of 27,765 participants ranging from multiple regions in Vietnam. Variables in this dataset includes pharmacy visits, medical expenses, age, sex, marriage, education, illness, injury, illdays, active days, insurance status and commune size. This python portion is a continuation of project 1 and 2 in SDS348 - Bioinformatics course.

1.) Import Dataset

In [4]:

```
# Import package pandas
import pandas as pd
```

In [5]:

```
# Import dataset
VN = pd.read_csv("https://raw.githubusercontent.com/HieuxPham/myrepo/main/VietNamI%20(1
```

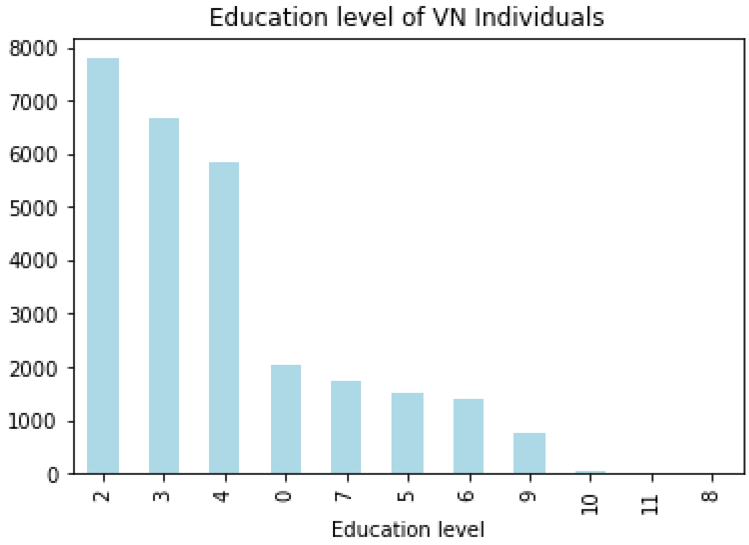
2.) Age of VN Individuals (Categorical)

In [23]:

```
# Graph
VN['educ'].value_counts().plot(kind = "bar", color = 'lightblue') # bar plot
plt.xlabel('Education level') # x-axis
plt.title('Education level of VN Individuals') # title
```

Out[23]:

```
Text(0.5, 1.0, 'Education level of VN Individuals')
```



In [11]:

```
# Statistic
VN['educ'].value_counts() # frequency of education levels
```

Out[11]:

```
2     7803
3     6653
4     5833
```

```
0     2048
7     1720
5     1495
6     1411
9      756
10      25
11      14
8        7
Name: educ, dtype: int64
```

According to the histogram, a majority of the individuals in the study have an education level of between 2nd - 4th grade. Furthermore, the frequency table indicates that the numbers of individuals decrease with higher education levels.

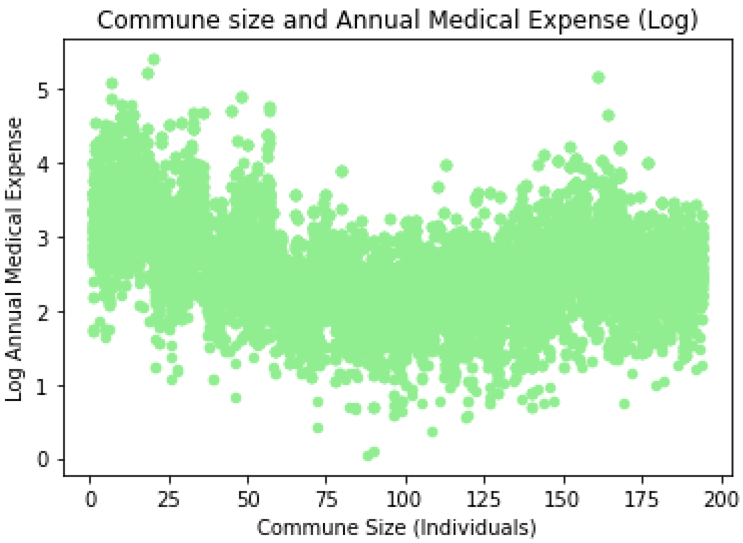
3.) Commune size and Annual Medical Expenses (Numerical)

In [17]:

```
# Graph
VN.plot.scatter(x = 'commune', y = 'lnhhexp', color = 'lightgreen' )
plt.xlabel('Commune Size (Individuals)') # x-axis
plt.ylabel('Log Annual Medical Expense') # y-axis
plt.title('Commune size and Annual Medical Expense (Log)') # title
```

Out[17]:

```
Text(0.5, 1.0, 'Commune size and Annual Medical Expense (Log)')
```



In [21]:

```
# Statistic
VN.filter(['commune']) \ # Mean and standard deviation of commune size
.agg(['mean', 'std'])
```

Out[21]:

	commune
mean	101.526598
std	56.283344

In [22]:

```
# Statistic
VN.filter(['lnhhexp']) \ # Mean and standard deviation of Log annual medical expenses
.agg(['mean', 'std'])
```

Out[22]:

	lnhhexp
mean	2.602610

	lnhhexp
std	0.624414

The scatterplot indicates that there is a generally negative correlation between commune size and individual medical expenses. Therefore, as commune size increases, an individual's medical expense will likely decrease. The mean commune size is 102 individuals, while the standard deviation is 56.3. Likewise, the mean log annual medical expense is 2.6 units with a standard deviation of 0.624.