Pinpointing rare categories in categorical variables

Different labels appear in a variable with different frequencies. Some categories of a variable appear a lot, that is, they are very common among the observations, whereas other categories appear only in a few observations. In fact, categorical variables often contain a few dominant labels that account for the majority of the observations and a large number of labels that appear only seldom. Categories that appear in a tiny proportion of the observations are rare. Typically, we consider a label to be rare when it appears in less than 5% or 1% of the population.so here we will learn how to identify infrequent labels in a categorical variable.

Data set: Car Evaluation dataset from the UCI Machine Learning Repository

```
# Importing the python libraries
import pandas as pd
import matplotlib.pyplot as plt

# load the dataset

data = pd.read_csv('data/car.data',header=None)
data.columns =  ['buying', 'maint', 'doors', 'persons',
'lug_boot','safety', 'class']
data.head()
```

```
  buying  maint doors persons lug_boot safety  class
0  vhigh  vhigh     2       2    small    low  unacc
1  vhigh  vhigh     2       2    small    med  unacc
2  vhigh  vhigh     2       2    small   high  unacc
3  vhigh  vhigh     2       2      med    low  unacc
4  vhigh  vhigh     2       2      med    med  unacc
```

TIP: By default, pandas read_csv() uses the first row of the data as the column names. If the column names are not part of the raw data, we need to specifically tell pandas not to assign the column names by adding the header = None argument.

```
# Let's display the unique categories of the variable class
data['class'].unique()
```

```
array(['unacc', 'acc', 'vgood', 'good'], dtype=object)
```

Let's calculate the number of cars per category of the class variable and then divide them by the total number of cars in the dataset to obtain the percentage of cars per category.

```
label_freq = data['class'].value_counts()/ len(data)
label_freq
```

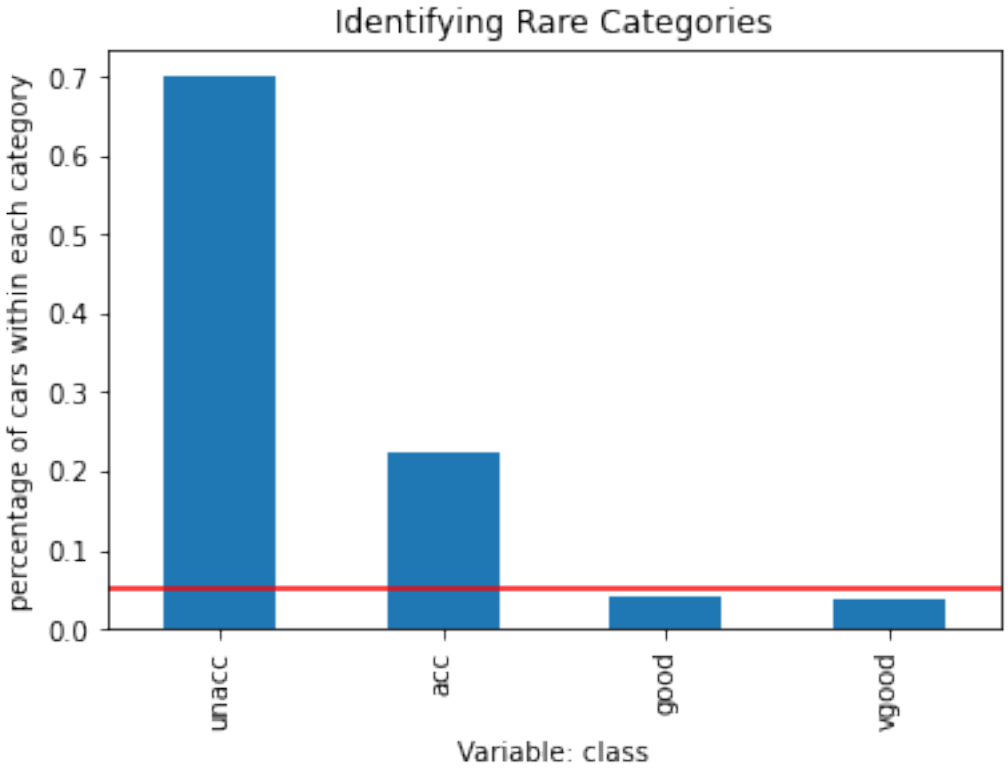
```
unacc    0.700231
acc      0.222222
good     0.039931
```

```
vgood    0.037616
Name: class, dtype: float64
```

Observation: Here we can observe unacc category has more frequency compared with other categories. we can also plot using matplotlib.

Let's make a bar plot showing the frequency of each category and highlight the 5% mark with a red line:

```
fig = label_freq.sort_values(ascending=False).plot.bar()
fig.axhline(y=0.05, color='red')
fig.set_ylabel('percentage of cars within each category')
fig.set_xlabel('Variable: class')
fig.set_title('Identifying Rare Categories')
plt.show()
```



The good and vgood categories are present in less than 5% of cars, as indicated by the red line in the preceding plot.