Feature Engineering: Transformations

Read in text

```
import pandas as pd

data = pd.read_csv("SMSSpamCollection.tsv", sep='\t')
data.columns = ['label', 'body_text']
```

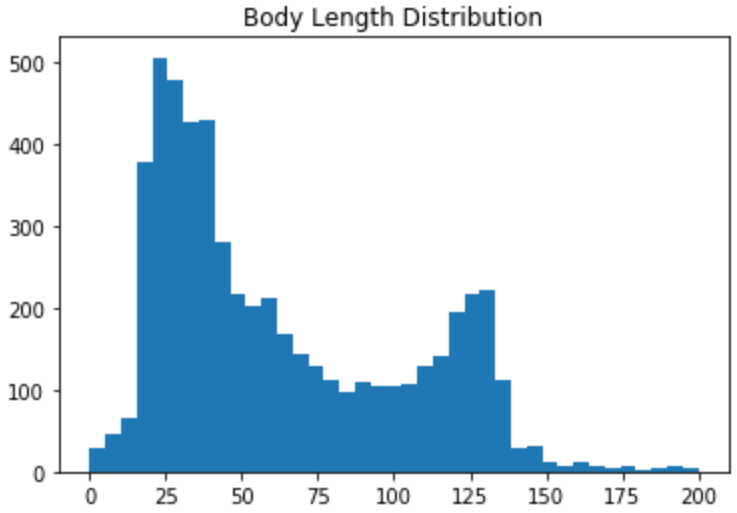
Create the two new features

```
import string

def count_punct(text):
    count = sum([1 for char in text if char in string.punctuation])
    return round(count/(len(text) - text.count(" ")), 3)*100

data['body_len'] = data['body_text'].apply(lambda x: len(x) - x.count(" "))
data['punct%'] = data['body_text'].apply(lambda x: count_punct(x))
data.head()
```

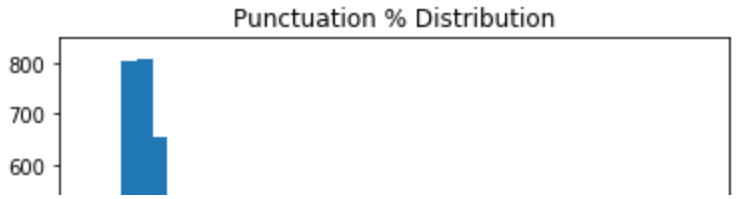
	label	body_text	body_len	punct%
0	spam	Free entry in 2 a wkly comp to win FA Cup fina...	128	4.7
1	ham	Nah I don't think he goes to usf, he lives aro...	49	4.1
2	ham	Even my brother is not like to speak with me. ...	62	3.2
3	ham	I HAVE A DATE ON SUNDAY WITH WILL!!	28	7.1
4	ham	As per your request 'Melle Melle (Oru Minnamin...	135	4.4

Plot the two new features

```
from matplotlib import pyplot
import numpy as np
%matplotlib inline


bins = np.linspace(0, 200, 40)

pyplot.hist(data['body_len'], bins)
pyplot.title("Body Length Distribution")
pyplot.show()
```




```
bins = np.linspace(0, 50, 40)

pyplot.hist(data['punct%'], bins)
pyplot.title("Punctuation % Distribution")
pyplot.show()
```



Transform the punctuation % feature

Box-Cox Power Transformation

Base Form:

$$y^x$$

X	Base Form	Transformation
-2	y^{-2}	$\frac{1}{y^2}$
-1	y^{-1}	$\frac{1}{y}$
-0.5	$y^{\frac{-1}{2}}$	$\frac{1}{\sqrt{y}}$
0	y^0	$log(y)$
0.5	$y^{\frac{1}{2}}$	$\sqrt{y}$
1	y^1	y
2	y^2	y^2

Process

1. Determine what range of exponents to test
2. Apply each transformation to each value of your chosen feature
3. Use some criteria to determine which of the transformations yield the best distribution

Colab paid products - Cancel contracts here