Questions:

1 - In lesion 3 section 2 of ‘Working with Databases’ I don’t understand what the following code is doing:

cols = [desc[0] for desc in cur.description]

Can you please explain this, thanks.

List comprehension. List of names from data frame. Df.columns

2 – In Unit 2 Lesson 1 subsection ‘Overview of Statistics’ can you explain this code and syntax please.

data = [i.split(',') for i in data]

list comprehension. It means split the elements of I by ‘,’ for all i’s in data

I did it like this instead, but I would like to know how the above works.

data1 = []

for i in range(len(data)):

var = data[i]

var = var.split(',')

data1.append(var)

print(data1)

3 - Same module as question 2. Why is python returning this instead of a simple number? How do I convert this to a simple numeric value?

from scipy import stats

stats.mode(data['Alcohol'])

Python returns the following:

ModeResult(mode=array([ 4.02]), count=array([1]))

I fixed it like this:

the\_mode = stats.mode(data['Alcohol'])

the\_mode[0][0]

Is there a better way?

Not really:

alc\_mode = stats.mode(df['Alcohol'])

tob\_mode = stats.mode(df['Tobacco'])

print "The mode of the alcohol and tobacco is %.2f %.2f" %(float(alc\_mode[0]), float(tob\_mode[0]))

4 - For the Statistics Overview challenge, I want to write a function that will take any data frame and calculate those statistics for each variable.

A - How would I exclude variables that are not numeric? How do I test if a column is some numeric type? Like I don’t want to have to test is dtype == float64 or dtype == int32. I want to test if it is just some kind of numeric or not?

import pandas as pd

df = pd.DataFrame({'A':[1,2,3],'B':[1.2,3.4,5.6], 'C':['a','b','c']})

df1 = df.\_get\_numeric\_data()

5 - Please check my stats.ipynb (on my github - https://github.com/realrbird) script and let me know if there is anything that I could have done better or if there is a more efficient way to write any of that code.

6 - My plots don’t print to an external file when I use the plt.savefig() command. Is this because I use the ‘%matplotlib inline’ in my jupyter notebook? Is there a way to do both?

7 - The chi-squared didn’t work for me, python threw an exception that Tuple was out of range. Freq.values() comes out like this:

dict\_values([24, 60, 106, 153, 232, 216, 262, 237, 185, 187, 153, 158, 138, 96, 66, 58, 51, 30, 23, 26, 8, 11, 7, 4, 3, 1, 1, 1, 1])

so I just did this for a fix:

chi,p = stats.chisquare(list(freq.values()))

Is there a better way to do this? Why is this happening?

Is it python 2 or 3 ? python 3 does not return a list so in order to make it work you need to convert it into a list like you did here.

8 - In the chi-squared script what is the following line of code doing?

freq = collections.Counter(loansData['Open.CREDIT.Lines'])

A [**Counter**](https://docs.python.org/2/library/collections.html#collections.Counter) is a **[dict](https://docs.python.org/2/library/stdtypes.html" \l "dict" \o "dict)** subclass for counting hashable objects. It is an unordered collection where elements are stored as dictionary keys and their counts are stored as dictionary values.

9 - The Logistic Regression gives me only the coefficients. How do I get a full table with standard Errors p-vale etc. like I did with Linear Regression?

result.summary() like linear regression

10 - When creating categorical variables, can I use the C() function of the pd.Categorical function on factor variables that have multiple levels? Or only on Dummy 0 1 variabbles? What if I have a categorical variable that is “Low”, “Medium”, “High”?

http://fastml.com/converting-categorical-data-into-numbers-with-pandas-and-scikit-learn/