CS400 Special Topics: Machine Learning
Assignment 2: Pandas Package

Instructions:

1) **Write a python script for the following question.**
2) **Include appropriate comments to separate each question.**
3) **Each question should print the result.**
4) **Submit the script in Blackboard**

Questions:

1. Import pandas and print the version.
2. Create a pandas series from a list of numbers (created using the numpy package).
3. Convert the series created in the last question into a dataframe. This will require researching functions for series from the pandas package.
4. Create another series and combine this series with the one created in the previous question to form a dataframe. Series 1 will represent column 1 and Series 2 will represent column 2.
5. Here is an example on how to find patterns using python:

```
import pandas as pd
import re
emails = pd.Series(['buying books at amazom.com', 'someone@someone.com', 'matt@t.co', 'somebody@com'])
pattern ='[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\\.[A-Za-z]{2,4}'
emails.str.findall(pattern, flags=re.IGNORECASE)
```

re refers to the regular expression operations (https://docs.python.org/3/library/re.html)
emails is a series of variables that may or may not be emails
pattern defines the regular expression that represents an email.
The last line of code compares the series with the pattern and returns the list with the matching pattern.

Write similar code to find phone numbers that match this pattern – (XXX) XXX-XXXX
Use this series as input:
phones = pd.Series(['12345', '(610) 786-9089', '610 786-9089', '(610) 786 9089'])

The code should output the list of phone numbers that match the pattern.

6. Compute the euclidean distance between series p and q, without using a python packaged formula.
   p = pd.Series([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
   q = pd.Series([10, 9, 8, 7, 6, 5, 4, 3, 2, 1])
7. Import the dataset (Housing.csv available in Blackboard) into a dataframe **df**. Make sure the column names are used in the dataframe.
8. Check is **df** has any missing values.
9. Calculate the frequency of distinct values in df.
10. Find all rows where the values of "zn" column is greater than 15.
11. Display only the "zn" column of the dataframe.
12. Define two conditions – crim is greater than 1.0 and medv is greater than 10.0. Apply these two conditions to the dataframe and show the resulting rows.