algorithmica

# Programming Language Learning Series
# Mastery of Python Language
## (Interview Questions/Assignment-Functional Style)

Q1: Write a function to compute 1/2+2/3+3/4+...+n/n+1 with a given n (n>0).

Q2: Write a function to find the sum of all the multiples of 3 or 5 below 1000.

Q3: A palindromic number reads the same both ways. The largest palindrome made from the product of two 2-digit numbers is 9009 = 91 × 99. Write a function to find the largest palindrome made from the product of two 3-digit numbers.

Q4: We count 35 heads and 94 legs among the chickens and rabbits in a farm. Write a python function that returns how many rabbits and how many chickens do we have.

Q5: Given a text file as input, we are interested to computing the following text analytics on that input:

- Compute the number of words in the given file
- Find the 10 most frequent words in the given file
- Find the number of times a given word appears in the file

Assuming that we want to develop a solution for the required text analytics using procedural abstractions. Which abstraction do you prefer and why?

Using the `iris` dataset from http://goo.gl/3b3439, answer the following questions:

How many rows and columns are there?

What is the type of each column?

Show all unique values for the state column

- Show the first 3 rows
- Show the last 3 rows
- Show 3 random rows without repetition

Show rows 5 to 10 (inclusive)

Show only rows where the state is treated and the rate is more than 100

Show only the conc and rate columns

Show only the columns whose type is numeric

Convert all column names to UPPERCASE

Rearrange the columns in the order state, conc, rate

```
In [ ]:
```

Drop the state column

create a new column rate2 that is the square of rate

Using the `iris` data set,

- Find the mean value of all 4 measurements
- Find the mean value of all 4 measurements for each Species

Using the `iris` data set,

- Sort the observations by Sepal.Width in decreasing order.

Using the `iris` data`m set,

- Count the number of flowers of each Species

Using the `iris` data set,

- Count the number of observations where Petal.Length is longer than Sepal.Width

sing the `iris` data set,

- Find the Species with the most number of observations where the Sepal.Length is less then the mean Sepal.Length of all observations

```
In [1]:
```

**7**. Using the `iris` data set,

- Convert the data frame from the current wide format to a tall format, with just 3 columns: Species, Measurement, Value.

Using the `iris` data set

- Create a new data frame `df` that has only 3 columns (`Species`, `Measure`, `Value`) where `Measure` takes on the values `Sepal.Length`, `Sepal.Width`, `Petal.Length` or `Petal.Width`. Show the first 5 rows.
- Show the mean value and counts for each Species and Measure of `df`

- Find the mean, min and max values of all four measurements (sepal.length, sepal.width, petal.length, petal.width) for each species
- Find the average petal.width for rows where the petal.length is less than the sepal.width

The `heart` dataframe at https://goo.gl/CbJwQM contains information about the survival of patients on the waiting list for the Stanford heart transplant program.

```
start, stop, event:  Entry and exit time and status for this interval
of time
age:                 age-48 years
year:                year of acceptance (in years after 1 Nov 1967)
surgery:             prior bypass surgery 1=yes
transplant:          received transplant 1=yes
id:                  patient id
```

Answer the following questions with respect to the `heart` data set:

- Sort the data frame by age in descending order (oldest at top) without making a copy
- How many patients received a transplant?
- What is the average age for transplanted patients under the age of 70?
- Find the mean and standard deviation of age for each value of the `transplant` variable.

algorithmica

# Programming Language Learning Series
# Mastery of Python Language
## (Interview Questions/Assignment-Functional Style)

Read the flights data at https://raw.githubusercontent.com/mwaskom/seaborn-data/master/flights.csv into a `pnadas` data frame. Find the average number of passengers per quarter (Q1, Q2, Q3,Q4) across the years 1950-1959 (inclusive of 1950 and 1959), where

- Q1 = Jan, Feb, Mar
- Q2 = Apr, May, Jun
- Q3 = Jul, Aug, Sep
- Q4 = Oct, Nov, Dec

Read the following data sets into DataFrames.

- url1 = "https://raw.github.com/vincentarelbundock/Rdatasets/master/csv/DAAG/hills.csv"
- url2 = "https://raw.github.com/vincentarelbundock/Rdatasets/master/csv/DAAG/hills2000.csv"

Create a new DataFraem only containing the names present in both DataFrames. Drop the `timef` column and have a single column for `dist` , `climb` and `time` that shows the average value of the two DataFrames. The final DtataFrame will thus have 4 columns (name, dist, climb, time).

This data contains the survival time after receiving a heart transplant, the age of the patient and whether or not the survival time was censored

- Number of Observations - 69
- Number of Variables - 3

Variable name definitions:

```
-  survival - Days after surgery until death
-  censors - indicates if an observation is censored. 1 is uncensored
-  age - age at the time of surgery
```

Answer the following questions with respect to the `heart` data set:

- How many patients were censored?
- What is the correlation coefficient between age and survival for uncensored patients?

# Programming Language Learning Series
# Mastery of Python Language
## (Interview Questions/Assignment-Functional Style)

- What is the average age for censored and uncensored patients?
- What is the average survival time for censored and uncensored patients under the age of 45?
- What is the survival time of the youngest and oldest uncensored patient?
- `import statsmodels.api as sm`
- `heart = sm.datasets.heart.load_pandas().data`
- `heart.head(n=6)`
- `Out[5]:`

| | survival | censors | age |
|---|---|---|---|
| **0** | 15 | 1 | 54.3 |
| **1** | 3 | 1 | 40.4 |
| **2** | 624 | 1 | 51.0 |
| **3** | 46 | 1 | 42.5 |
| **4** | 127 | 1 | 48.0 |
| **5** | 64 | 1 | 54.6 |

- `In [6]:`
- `# How many patients were censored?`
- 
- `print('# censroed', sum(heart.censors == 0), '\n')`
- 
- `# What is the correlation coefficient between age and survival for uncensored patients?`
- 
- `uncensored = heart[heart.censors == 1]`
- `print('Correlation coefficient', np.corrcoef(uncensored.age, uncensored.survival)[0,1], '\n')`
- 
- `# What is the average age for censored and uncensored patients?`
- 
- `print(heart.groupby('censors')['age'].mean(), '\n')`
- 
- `# What is the average survival time for censored and uncensored patients under the age of 45?`
- 
- `young = heart[heart.age < 45]`
- `print(young.groupby('censors')['survival'].mean(), '\n')`
- 
- `# What is the survival time of the youngest and oldest uncensored patient?`

# Programming Language Learning Series
# Mastery of Python Language
## (Interview Questions/Assignment-Functional Style)

- 
- `print('Survival of youngest', uncensored.survival[np.argmin(uncensored.age)])`
- `print('Survival of olderst', uncensored.survival[np.argmax(uncensored.age)])`

df = pd.DataFrame(2by2 array, columns=list('QR'))

df - df.iloc[0]

data = pd.DataFrame(2d array 4 by 4, columns=['C1',..,'C4'], index=['r1',..'r4'])

data.loc['r1':'r3', 'c2':'c4']

data.loc[:'r4', :'c3']

data.iloc[0:2,1:3]

data.iloc[0:2,::2]

df1 = pd.DataFrame(d1, columns=list('ABC'))

df2 = pd.DataFrame(d1, columns=list('ACD'))

pd.merge(df1, df2, on='A')

how do you handle duplicate names when merge?

pd.concat([df1,df2], axis = 0)

how do you get the index right after concatenation?

df = pd.DataFrame([10,20,30,40], columns=['c1'], index=[1,4,6,8])

df[1]

df[1:4]

df[df.c1>30]

why?

s = pd.Series([10,20,30,40], index=[1,4,6,8])

s[1]

s[1:4]

why?

convert a series to float type

s3 = Series([1,2,3,4])

s4 = Series([1,2,3,4], index=range(1, 5))

s3+s4

Differene between pd.merge and pd.join operations?