Intro to Data Science - Lab 4

DATA 1501 — Dr. Mihail
Department of Computer Science
Valdosta State University

September 22, 2021

1 Introduction

In this lab, you will compute basic univariate descriptive statistics for several datasets with various types of attributes. The submission for this lab will be a Word document with answers to the questions in each part of the lab. Please read the entire document and make sure you don't miss any parts.

1.1 Part 1 (25 points)

In this part, you will download a dataset consisting of names that babies were named from 1880 to 2008. Create a new Colab notebook and run the following code in a code cell:

```
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab4/baby-names.csv
```

Confirm that the file was downloaded and no errors reported from mispellings. You should see something like:

```
baby-names.csv      100%[===================>]   7.10M  32.2MB/s    in 0.2s
```

Next, create a code cell and run the following code:

```
import pandas as pd
df = pd.read_csv('baby-names.csv')
df
```

Confirm you can see a few records in the dataset. Now, you will execute Python code that counts the occurence of each name value in the dataset and displays them in descending order. To accomplish that, run the following command:

```
df["name"].value_counts()
```

Given the above computation, answer the following questions:

1. What measure of central tendency can you compute for the **name** attribute in the dataset?

2. Assuming there is a measure, indicate the value of the central tendency measure, or say Not Applicable if no such measure can be computed.

3. How many records are there in the dataset?

1.2 Part 2 (50 points)

In this part of the lab you will download a dataset consisting of answers to Rosenberg Self-Esteem Scale, a 10-item scale that measures global self-worth by measuring both positive and negative feelings about the self. The scale is believed to be uni-dimensional. All items are answered using a 4-point Likert scale format ranging from strongly agree to strongly disagree.

First, download the data legend (codebook) that explains the attributes:

```
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab4/RSE_codebook.txt
!cat RSE_codebook.txt
```

Please read the description of the dataset and questions carefully. Next, download the dataset, and load into a Pandas dataframe using the following code:

```
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab4/RSE_data.csv
import pandas as pd
df = pd.read_csv('RSE_data.csv', delimiter='\t')
```

Answer the following questions:

- What type are attributes Q1 through Q10?
- What central tendency and dispersion measures are valid to compute?

In order to compute the mode, median, mean and standard deviation for a given attribute (in the example below **Q1** was used, the following syntax should be used in python:

```
print("Mode for attribute Q1 is: ", df['Q1'].mode()[0])
print("Median for attribute Q1 is: ", df['Q1'].median())
print("Mean for attribute Q1 is: ", df['Q1'].mean())
print("Standard deviation for attribute Q1 is: ", df['Q1'].std())
```

Create a table with all the descriptive statistics allowed for all the attributes in the dataset given what you know about their types. This table will be part of your lab submission. The Microsoft Word table should have as columns the attribute names, and as rows each of the descriptive statistics. If any measure is not applicable, fill in **N/A**

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	gender	age	source	country
mode														
median														
mean														
st dev														

1.3 Part 3 (25 points)

When filling out the table in Part 2, you may have noticed something odd about the mean value of age. The following code will list all the records where age is above 100 years:

```
print(df[df['age']>100])
```

Answer the following question:

- How many records have likely errors?

Next, we will compute the mode, median and mean for all records where age is less than 100.

```
print(df[df['age']<100]['age'].mode()[0])
print(df[df['age']<100]['age'].median())
print(df[df['age']<100]['age'].mean())
```

Explain, in your own words, what you observed when computing the mode, median, mean and standard deviation for the "age" attribute in the dataset. Explain the very large difference between mode, median and mean.

Due Date: Before Midnight on Sunday, September 19th.
Submit the Word document via Blazeview/Assignments/Lab 4.