Capstone Project 1 Data Wrangling

Drug consumption (quantified) Data Set

Alissa Mezzacappa

For my capstone project I chose the Drug consumption (quantified) Data Set. This data set has 32 attributes collected from real people via an online survey method including: age, gender, education, country, ethnicity, several personality attributes, and usage of 19 different drugs. This data set already has false data removed by identifying individuals that claimed drug use of a fictional drug included in the survey. The data had no column labels.

First, I scraped the column labels from the data set information page with the following code:

import pandas as pd

import urllib

import re

u=urllib.urlopen('http://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29')

text=u.read()

matches=re.findall(r"\d+\.\s\w+", text)

del(matches[0:3])

del(matches[12])

for i in range(0, len(matches)):

matches[i]=re.sub(r"\d+\.\s","",matches[i])

Code above identifies the column names by identifying the pattern of digits, period, then space as preceding the column name. A few false positives were also obtained which are then deleted in the rows below. After that the digit, period, space part of the match pattern is removed to yield only the column name.

Once column names are found, data is imported and a binary data is created for drug use with the following code:

url = "http://archive.ics.uci.edu/ml/machine-learning-databases/00373/drug\_consumption.data"

data = pd.read\_csv(url,names=matches) #import data

data0=data.replace(['CL0'],0) #non drug user =CL0 -- never use

data0=data0.replace(['CL1','CL2','CL3','CL4','CL5','CL6'],1)

data1=data.replace(['CL0','CL1'],0) #non frug user = CL1 -- over a decade

data1=data1.replace(['CL2','CL3','CL4','CL5','CL6'],1)

data2=data.replace(['CL0','CL1','CL2'],0) #non drug user = CL2 -- last decade

data2=data2.replace(['CL3','CL4','CL5','CL6'],1)

data3=data.replace(['CL0','CL1','CL2','CL3'],0) #non drug user = Cl3 -- last year

data3=data3.replace(['CL4','CL5','CL6'],1)

data4=data.replace(['CL0','CL1','CL2','CL3','CL4'],0) #non drug user = CL4 -- last month

data4=data4.replace(['CL5','CL6'],1)

data5=data.replace(['CL0','CL1','CL2','CL3','CL4','CL5'],0) #non drug user = CL5 -- last week

data5=data5.replace(['CL6'],1) #CL6 -- last day is always a drug user

In the code above, data is directly imported from the online repository. Then I created binary variables where a non-user is assigned a zero value and a user is assigned a value of 1. Separate variables are created for each usage type delineation “never used”, “used over a decade ago”, “used in the last decade”, “used in the last year”, “used in the last month”, “used in the last week”, and “used in the last day”.