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 urllib

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

text=u.read()

import re

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)

data=data.replace(['CL0','CL1','CL2','CL3'],0)

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

In the code above, data is directly imported from the online repository. Then I create a binary variable where a non-user is assigned a zero value and is classified as “never used”, “used over a decade ago”, “used in the last decade”, or “used in the last year”. A user is assigned a value of 1 and is classified as “used in the last month”, “used in the last week”, or “used in the last day”. I made the distinction of user vs. non user based on a month frequency or less because I want this analysis to represent drug use that has a dramatic affect.