# Data Analytics Project - Checkpoint 1

1. **How does your chosen topic and identified data and supporting material satisfy each one of the 5 criterias below? Please see the explanation provided above for each criterion in the "Five Criteria for Appropriate Data " guideline above.**
    a. **Importance:** Find where to best allocate resources for drug treatment

    b. **Availability**: Data available at
        i. National Survey on Drug Use and Health, 2017 (NSDUH-2017-DS0001)https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2017-nsduh-2017-ds0001-nid17939

    c. **Documentation**
        i. Codebook for National Survey on Drug Use and Health, 2017 (NSDUH-2017-DSS0001): http://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/NSDUH-2017/NSDUH-2017-datasets/NSDUH-2017-DS0001/NSDUH-2017-DS0001-info/NSDUH-2017-DS0001-info-codebook.pdf

    d. **Support**
        i. Questionnaire Specs for National Survey on Drug Use and Health, 2017 (NSDUH-2017-DSS0001): http://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/NSDUH-2017/NSDUH-2017-datasets/NSDUH-2017-DS0001/NSDUH-2017-DS0001-info/NSDUH-2017-DS0001-info-questionnaire-specs.pdf
        ii. Questionnaire Showcards for National Survey on Drug Use and Health, 2017 (NSDUH-2017-DSS0001): http://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/NSDUH-2017/NSDUH-2017-datasets/NSDUH-2017-DS0001/NSDUH-2017-DS0001-info/NSDUH-2017-DS0001-info-questionnaire-showcards.pdf

    e. **Size: NSDUH Data from 2017**
        i. NSDUH: 2668 columns, 56,276 rows

2. **Describe your data properties, including the following, as much as possible.**
    a. **Data format (tabular, database or file format, etc.)**
        i. Downloadable in an R file
    b. **Data tables (how many, their content/organization, etc.)**
        i. Every column represents a question that was asked in a survey. Each row then represents the respondents (individuals or treatment facilities) ID number and the answers they gave. If no answer was given, then that was recorded as well.

ii. Pain Reliever Misuse in the Past Year (by Age groups and States)
iii. Needing But not received treatment at a specialty facility for Substance use in the past year (by Age groups and states)

**c. Data columns (most important ones, etc.)**

i. FIPE4: IN WHAT STATE IS THIS SAMPLE DWELLING UNIT (SDU) LOCATED?
ii. TX17: During the past 12 months, did you need treatment or counseling for your use of prescription pain relievers?
iii. TX10 [IF TX09 = 1] During the past 12 months, for which of the following drugs did you need additional treatment or counseling?
iv. TX14 [IF (HE01 = 1 OR HEREF = 1) AND TX08 = 1] During the past 12 months, did you need treatment or counseling for your use of heroin?
v. TX21: During the past 12 months, did you need treatment or counseling for your use of some other drugs besides the ones just listed?
vi. TX22B [IF ANY ENTRY IN TX22A = 10] Which of these statements explain why you did not get the treatment or counseling you needed for your use of TXFILL2 (alcohol or any other drug)?
vii. TX22A: [IF TX22 IS NOT BLANK] Which of these statements explain why you did not get the treatment or counseling you needed for your use of [TXFILL2]?
    1. NDTXNOCOV1 : NOT GET TRMT COULDN'T AFFORD - NO HLTH CARE COVER
    2. NDTXNOTPY1 : TREATMENT NOT COVERED ON HLTH CARE
    3. NDTXMIMPT1 : MOST IMPORTANT OTHER REASON DIDNT GET TRTMT

**d. Data rows (unit of observation, count, etc.)**

i. NSDUH: people (with age groups, genders, ethnicity, etc.)
ii. NSSATS: treatment facilities (including location by states or cities)

**3. Describe your data variables. Please use distribution statistics (mean, median, mode, percent missing, etc.) and distribution chats.**

**a. Categorical variables (nominal or ordinal)**

i. Location(state, urbanicity), age, gender, education, ethnicity, income

**b. Numerical variables (binary or interval)**

i. Gender (binary), age (interval), education (interval), ethnicity (binary), income (interval)

**c. Potential target variable**

i. Drug addiction/recovery

**4. Propose potential questions you will answer with or insights you will gain from your data analytics.**

**a.** What factors affect drug use (e.g. location, gender, race, economic class, etc.)?

**b.** What factors determine if a person is most likely to get treatment/stay in treatment?