DATA SCRUBBING

Linux Commands

**Columns we need:**

1292:UDPYOPI

1300:BOOKED

1384:TXEVRRCVD

1698:BMI2

1239:IRABUPOSTRQ

1243:IRABUPOSSED

129:HALLUCEVR

1233:IRABUPOSINH

1235:IRABUPOSMTH

1381:CIGYRBFR

1197:IRCGRGDY

1241:IRABUPOSSTM

1846:YEPHLPHW

1851:YEPPROUD

1884:YEFAIACT

2628:MILTPARNT

1752:AUUNMTYR

2678:MEDICARE

2685:HLTINMNT

2686:HLTINNOS

2714:IRFAMSOC

2731:INCOME

1833:YEATNDYR

1841:YESTSCIG

1842:YESTSMJ

1843:YESTSALC

2617:EDUSCHLGO

2634:WRKDPSTWK

**Number of columns in the datafiles:**

awk '{print split($0,a,\t); exit}' NSDUH\_2019\_Tab2.txt

>>2741

**Number of lines in the datafile:**

wc -l NSDUH\_2019\_Tab2.txt

>>56137 NSDUH\_2019\_Tab2.txt

**List columns in file**

Head -n 1 filename

**List columns in file on a new line by replacing each tab with the newline character**

Head -n 1 NSDUH\_2019\_Tab2.txt | tr “\t” “\n” | less

Find Column UDPYOPI and return column number

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "UDPYOPI"

1292:UDPYOPI

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "AGE"

>>>2577:AGE2

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "BOOKED"

1300:BOOKED

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "TXEVRRCVD"

1384:TXEVRRCVD

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "BMI2"

1698:BMI2

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "IRABUPOSSED\|IRABUPOSTRQ"

1239:IRABUPOSTRQ

1243:IRABUPOSSED

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "IRABUPOSMTH\|IRABUPOSINH\|CIGYRBFR\|HALLUCEVR"

129:HALLUCEVR

1233:IRABUPOSINH

1235:IRABUPOSMTH

1381:CIGYRBFR

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "IRCGRGDY\|IRABUPOSSTM\|YEPHLPHW\|YEPPROUD\|YEFAIACT\|MILTPARNT"

1197:IRCGRGDY

1241:IRABUPOSSTM

1846:YEPHLPHW

1851:YEPPROUD

1884:YEFAIACT

2628:MILTPARNT

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "IRFAMSOC\|INCOME\|HLTINMNT\|AUUNMTYR\|HLTINNOS\|MEDICARE"

1752:AUUNMTYR

2678:MEDICARE

2685:HLTINMNT

2686:HLTINNOS

2714:IRFAMSOC

2731:INCOME

head -n 1 NSDUH\_2019\_Tab2.txt | tr "\t" "\n" |grep -n "EDUSCHLGO\|WRKDPSTWK\|YEATNDYR\|YESTSMJ\|YESTSCIG\|YESTSALC"

1833:YEATNDYR

1841:YESTSCIG

1842:YESTSMJ

1843:YESTSALC

2617:EDUSCHLGO

2634:WRKDPSTWK

List the head of NSDUH\_2019\_Tab2.txt only columns 7 & 4

awk 'BEGIN {FS=\t}{print $7 \t $4}' NSDUH\_2019\_Tab2.txt | head

CIGYFU CIGOFRSM

9999 99

9991 99

9999 99

9991 99

9991 99

9999 99

9991 99

9999 99  
9991 99

Create scrubbed data file by selecting only the columns we want

\*tab delimeter is the default in the cut command remove -d “\t”

Cut -f 129, 1197 -d “\t” NSDUH\_2019\_Tab2.txt > Opioid\_Data\_Scrubbed.txt

Cut -f 129,1197,1233,1235,1239,1241,1243,1292,1300,1381,1384,1698,1752,1833,1841,1842,1843,1846,1851,1884,2577,2617,2628,2634,2678,2685,2686,2714,2731 NSDUH\_2019\_Tab2.txt > Opioid\_Data\_Scrubbed3.txt

Cut -f 2577 NSDUH\_2019\_Tab2.txt >> Opioid\_Data\_Scrubbed2.txt