2018 DATA SCRUBBING

Linux Commands

**Columns we need:**

1159:IRCGRGDY

1195:IRABUPOSINH

1197:IRABUPOSMTH

1201:IRABUPOSTRQ

1203:IRABUPOSSTM

1205:IRABUPOSSED

**1254:UDPYOPI**

1262:BOOKED

129:HALLUCEVR

1343:CIGYRBFR

1346:TXEVRRCVD

1660:BMI2

1714:AUUNMTYR

1795:YEATNDYR

1803:YESTSCIG

1804:YESTSMJ

1805:YESTSALC

1808:YEPHLPHW

1813:YEPPROUD

1846:YEFAIACT

2527:AGE2

2567:EDUSCHLGO

2578:MILTPARNT

2584:WRKDPSTWK

2628:MEDICARE

2635:HLTINMNT

2636:HLTINNOS

2664:IRFAMSOC

2681:INCOME

Original Datafile Information:

* 2691 Columns
* 56313 rows of data in 2018
* 448 Yes's out of 56313 rows, that means there are 55,865 No's for the **UDPYOPI** column

**Number of columns in the datafiles:**

awk -F'\t' '{print NF; exit}' NSDUH\_2018\_Tab.tsv

>>2691

**Number of lines in the datafile:**

wc -l NSDUH\_2018\_Tab.tsv

>> 56,313 NSDUH\_2018\_Tab.tsv

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

Head -n 1 NSDUH\_2018\_Tab.tsv | tr “\t” “\n” | less

Find Column UDPYOPI and return column number

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "UDPYOPI"

1254:UDPYOPI

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "AGE2"

2527:AGE2

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "BOOKED"

1262:BOOKED

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "TXEVRRCVD"

1346:TXEVRRCVD

1489:TXEVRRCVD2

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "BMI2"

1660:BMI2

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "IRABUPOSSED\|IRABUPOSTRQ"

1201:IRABUPOSTRQ

1205:IRABUPOSSED

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "IRABUPOSMTH\|IRABUPOSINH\|CIGYRBFR\|HALLUCEVR"

129:HALLUCEVR

1195:IRABUPOSINH

1197:IRABUPOSMTH

1343:CIGYRBFR

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "IRCGRGDY\|IRABUPOSSTM\|YEPHLPHW\|YEPPROUD\|YEFAIACT\|MILTPARNT"

1159:IRCGRGDY

1203:IRABUPOSSTM

1808:YEPHLPHW

1813:YEPPROUD

1846:YEFAIACT

2578:MILTPARNT

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "IRFAMSOC\|INCOME\|HLTINMNT\|AUUNMTYR\|HLTINNOS\|MEDICARE"

1714:AUUNMTYR

2628:MEDICARE

2635:HLTINMNT

2636:HLTINNOS

2664:IRFAMSOC

2681:INCOME

head -n 1 NSDUH\_2018\_Tab.tsv | tr "\t" "\n" |grep -n "EDUSCHLGO\|WRKDPSTWK\|YEATNDYR\|YESTSMJ\|YESTSCIG\|YESTSALC"

1795:YEATNDYR

1803:YESTSCIG

1804:YESTSMJ

1805:YESTSALC

2567:EDUSCHLGO

2584:WRKDPSTWK

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\_2018\_Tab.tsv > Opioid\_Data\_Scrubbed.txt

Cut -f 1159,1195,1197,1201,1203,1205,1254,1262,129,1343,1346,1660,1714,1795,1803,1804,1805,1808,1813,1846,2527,2567,2578,2584,2628,2635,2636,2664,2681 NSDUH\_2018\_Tab.tsv > Opioid\_Data\_2018\_Scrubbed.txt

2019 DATA SCRUBBING

Linux Commands

**Columns we need:**

1197:IRCGRGDY

1233:IRABUPOSINH

1235:IRABUPOSMTH

1239:IRABUPOSTRQ

1241:IRABUPOSSTM

1243:IRABUPOSSED

129:HALLUCEVR

**1292:UDPYOPI**

1300:BOOKED

1381:CIGYRBFR

1384:TXEVRRCVD

1698:BMI2

1752:AUUNMTYR

1833:YEATNDYR

1841:YESTSCIG

1842:YESTSMJ

1843:YESTSALC

1846:YEPHLPHW

1851:YEPPROUD

1884:YEFAIACT

2577:AGE2

2617:EDUSCHLGO

2628:MILTPARNT

2634:WRKDPSTWK

2678:MEDICARE

2685:HLTINMNT

2686:HLTINNOS

2714:IRFAMSOC

2731:INCOME

Original Datafile Information:

* 2741 Columns
* 56,136 rows of data in 2019
* 369 Yes's out of 56,136 rows, that means there are 55,767 No's for the **UDPYOPI** column

**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

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