In [573]:

```
!python --version
```

Python 3.7.10

In [574]:

```
!pip install --disable-pip-version-check -q sagemaker==2.38.0
!pip install --disable-pip-version-check -q smdebug==1.0.4
!pip install --disable-pip-version-check -q sagemaker-experiments==0.1.28
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [575]:

```
!pip install --disable-pip-version-check -q tensorflow==2.3.1
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [576]:

```
!pip install --disable-pip-version-check -q tensorflow==2.3.1
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [577]:

```
!pip install --disable-pip-version-check -q transformers==3.5.1
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead

from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [578]:

```
!pip install --disable-pip-version-check -q PyAthena==2.1.1
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [579]:

```
!pip install --disable-pip-version-check -q SQLAlchemy==1.3.23
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [580]:

```
!conda install -q -y zip
```

Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

All requested packages already installed.

In [581]:

```
!pip install --disable-pip-version-check -q matplotlib==3.1.3
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [582]:

```
!pip install --disable-pip-version-check -q seaborn==0.10.0
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [583]:

```
!pip list
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes

/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes i
nstead
  from cryptography.utils import int_from_bytes
Package   Version
--------------------------------- -----------------
absl-py   1.0.0
aiobotocore   2.0.1
aiohttp   3.8.1
aioitertools   0.8.0
aiosignal   1.2.0
alabaster   0.7.12
anaconda-client   1.7.2
anaconda-project   0.8.3
argh   0.26.2
argon2-cffi   21.3.0
argon2-cffi-bindings   21.2.0
asn1crypto   1.3.0
astroid   2.9.0
astropy   4.0
astunparse   1.6.3
async-timeout   4.0.1
asynctest   0.13.0
atomicwrites   1.3.0
attrs   19.3.0
autopep8   1.4.4
autovizwidget   0.19.1
awscli   1.22.23
Babel   2.9.1
backcall   0.1.0
backports.shutil-get-terminal-size 1.0.0
beautifulsoup4   4.8.2
bitarray   1.2.1
bkcharts   0.2
bleach   4.1.0
bokeh   1.4.0
boto   2.49.0
boto3   1.20.23
botocore   1.23.23
Bottleneck   1.3.2
brotlipy   0.7.0
cached-property   1.5.2
cachetools   5.0.0
certifi   2021.10.8
cffi   1.14.6
chardet   3.0.4
charset-normalizer   2.0.4
Click   7.0
cloudpickle   2.0.0
clyent   1.2.2
colorama   0.4.3
conda   4.12.0
conda-package-handling   1.7.3
contextlib2   0.6.0.post1
cryptography   36.0.0
cycler   0.10.0
Cython   0.29.15
cytoolz   0.10.1
dask   2021.12.0
decorator   4.4.1
defusedxml   0.6.0
diff-match-patch   20181111
dill   0.3.4
distributed   2021.12.0
distro   1.6.0
docker   5.0.0
docker-compose   1.29.2
dockerpty   0.4.1
docopt   0.6.2
docutils   0.15.2
dparse   0.5.1
entrypoints   0.3
et-xmlfile   1.0.1
fastcache   1.1.0
filelock   3.0.12
flake8   3.7.9
Flask   1.1.1
frozenlist   1.2.0
fsspec   2021.11.1
future   0.18.2
gast   0.3.3
gevent   1.4.0
glob2   0.7
gmpy2   2.0.8

google-auth	2.6.2
google-auth-oauthlib	0.4.6
google-pasta	0.2.0
greenlet	0.4.15
grpcio	1.44.0
h5py	2.10.0
hdijupyterutils	0.19.1
HeapDict	1.0.1
html5lib	1.0.1
hypothesis	5.5.4
idna	2.8
imageio	2.6.1
imagesize	1.2.0
importlib-metadata	4.11.3
intervaltree	3.0.2
ipykernel	5.1.4
ipython	7.12.0
ipython_genutils	0.2.0
ipywidgets	7.5.1
isort	4.3.21
itsdangerous	1.1.0
jdcal	1.4.1
jedi	0.14.1
jeepney	0.4.2
Jinja2	3.0.3
jmespath	0.10.0
joblib	0.14.1
json5	0.9.1
jsonschema	3.2.0
jupyter	1.0.0
jupyter-client	5.3.4
jupyter-console	6.1.0
jupyter-core	4.6.1
jupyterlab	1.2.21
jupyterlab-server	1.0.6
Keras-Preprocessing	1.1.2
keyring	21.1.0
kiwisolver	1.1.0
lazy-object-proxy	1.4.3
libarchive-c	2.8
lief	0.9.0
llvmlite	0.37.0
locket	0.2.0
lxml	4.6.4
Markdown	3.3.6
MarkupSafe	2.0.1
matplotlib	3.1.3
mccabe	0.6.1
mistune	0.8.4
mkl-fft	1.0.15
mkl-random	1.1.0
mkl-service	2.3.0
mock	4.0.1
more-itertools	8.2.0
mpmath	1.1.0
msgpack	0.6.1
multidict	5.2.0
multipledispatch	0.6.0
multiprocess	0.70.12.2
nbconvert	5.6.1
nbformat	5.0.4
nest-asyncio	1.5.4
networkx	2.4
nltk	3.4.5
nose	1.3.7
notebook	6.4.6
numba	0.54.1
numexpr	2.7.1
numpy	1.18.5
numpydoc	0.9.2
oauthlib	3.2.0
olefile	0.46
openpyxl	3.0.3
opt-einsum	3.3.0
packaging	20.1
pandas	1.0.1
pandocfilters	1.4.2
parso	0.5.2
partd	1.1.0
path	13.1.0
pathlib2	2.3.5
pathos	0.2.8

pathtools 0.1.2
patsy 0.5.1
pep8 1.7.1
pexpect 4.8.0
pickleshare 0.7.5
Pillow 8.4.0
pip 21.3.1
pkginfo 1.5.0.1
platformdirs 2.4.0
plotly 5.4.0
pluggy 0.13.1
ply 3.11
pox 0.3.0
ppft 1.6.6.4
prometheus-client 0.7.1
prompt-toolkit 3.0.3
protobuf 3.19.1
protobuf3-to-dict 0.1.5
psutil 5.6.7
ptyprocess 0.6.0
pure-sasl 0.6.2
py 1.11.0
pyarrow 6.0.1
pyasn1 0.4.8
pyasn1-modules 0.2.8
pyathena 2.1.1
pycodestyle 2.5.0
pycosat 0.6.3
pycparser 2.19
pycrypto 2.6.1
pycurl 7.43.0.5
pydocstyle 4.0.1
pyflakes 2.1.1
pyfunctional 1.4.3
Pygments 2.5.2
PyHive 0.6.4
pyinstrument 4.1.1
pykerberos 1.2.1
pylint 2.12.2
pyodbc 4.0.0-unsupported
pyOpenSSL 19.1.0
pyparsing 2.4.6
pyrsistent 0.15.7
PySocks 1.7.1
pytest 5.3.5
pytest-arraydiff 0.3
pytest-astropy 0.8.0
pytest-astropy-header 0.1.2
pytest-doctestplus 0.5.0
pytest-openfiles 0.4.0
pytest-remotedata 0.3.2
python-dateutil 2.8.1
python-dotenv 0.19.2
python-jsonrpc-server 0.3.4
python-language-server 0.31.7
pytz 2019.3
PyWavelets 1.1.1
pyxdg 0.26
PyYAML 6.0
pyzmq 18.1.1
QDarkStyle 2.8
QtAwesome 0.6.1
qtconsole 4.6.0
QtPy 1.9.0
regex 2022.3.15
requests 2.26.0
requests-kerberos 0.12.0
requests-oauthlib 1.3.1
rope 0.16.0
rsa 4.8
Rtree 0.9.3
ruamel_yaml 0.15.87
s3fs 2021.11.1
s3transfer 0.5.0
sacremoses 0.0.49
sagemaker 2.38.0
sagemaker-experiments 0.1.28
sagemaker-studio-analytics-extension 0.0.4
sagemaker-studio-sparkmagic-lib 0.1.3
sasl 0.2.1
scikit-image 0.16.2
scikit-learn 0.22.1
scipy 1.4.1

```
seaborn                    0.10.0
SecretStorage              3.1.2
Send2Trash                 1.8.0
sentencepiece              0.1.91
setuptools                 59.5.0
simplegeneric              0.8.1
singledispatch             3.4.0.3
six                        1.14.0
sklearn                    0.0
smclarify                  0.2
smdebug                    1.0.4
smdebug-rulesconfig        1.0.1
snowballstemmer            2.0.0
sortedcollections          1.1.2
sortedcontainers           2.1.0
soupsieve                  1.9.5
sparkmagic                 0.19.1
Sphinx                     2.4.0
sphinxcontrib-applehelp    1.0.1
sphinxcontrib-devhelp      1.0.1
sphinxcontrib-htmlhelp     1.0.2
sphinxcontrib-jsmath       1.0.1
sphinxcontrib-qthelp       1.0.2
sphinxcontrib-serializinghtml 1.1.3
sphinxcontrib-websupport   1.2.0
spyder                     4.0.1
spyder-kernels             1.8.1
SQLAlchemy                 1.3.23
statsmodels                0.11.0
sympy                      1.5.1
tables                     3.6.1
tabulate                   0.8.9
tblib                      1.6.0
tenacity                   8.0.1
tensorboard                2.8.0
tensorboard-data-server    0.6.1
tensorboard-plugin-wit     1.8.1
tensorflow                 2.3.1
tensorflow-estimator       2.3.0
termcolor                  1.1.0
terminado                  0.8.3
testpath                   0.4.4
texttable                  1.6.4
thrift                     0.13.0
thrift-sasl                0.4.3
tokenizers                 0.9.3
toml                       0.10.2
toolz                      0.10.0
tornado                    6.1
tqdm                       4.42.1
traitlets                  4.3.3
transformers               3.5.1
typed-ast                  1.5.1
typing_extensions          4.0.1
ujson                      1.35
unicodecsv                 0.14.1
urllib3                    1.26.7
watchdog                   0.10.2
wcwidth                    0.1.8
webencodings               0.5.1
websocket-client           0.59.0
Werkzeug                   1.0.0
wheel                      0.34.2
widgetsnbextension         3.5.1
wrapt                      1.11.2
wurlitzer                  2.0.0
xlrd                       1.2.0
XlsxWriter                 1.2.7
xlwt                       1.3.0
yapf                       0.28.0
yarl                       1.7.2
zict                       1.0.0
zipp                       2.2.0
WARNING: You are using pip version 21.3.1; however, version 22.0.4 is available.
You should consider upgrading via the '/opt/conda/bin/python -m pip install --upgrade pip' command.
```

In [584]:

```
setup_dependencies_passed = True
```

In [585]:

```
%store
```

```
Stored variables and their in-db values:
autopilot_train_s3_uri                      -> 's3://sagemaker-us-east-1-189468192453/data/amazon
ingest_create_athena_db_passed                -> True
ingest_create_athena_table_parquet_passed     -> True
s3_private_path_tsv                         -> 's3://sagemaker-us-east-1-189468192453/ads-508-azh
s3_public_path_tsv                          -> 's3://ads-508-azhang/finalproject/'
setup_dependencies_passed                     -> True
setup_iam_roles_passed                      -> True
setup_instance_check_passed                   -> True
setup_s3_bucket_passed                      -> True
```

In [586]:

```
from pyathena import connect
import pandas as pd
```

In [587]:

```
%store -r setup_dependencies_passed
```

In [588]:

```
try:
 setup_dependencies_passed
except NameError:
 print("++++++++++++++++++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN THE PREVIOUS NOTEBOOK ")
 print("You did not install the required libraries. ")
 print("++++++++++++++++++++++++++++++++++++++++++++++++")
```

In [589]:

```
print(setup_dependencies_passed)
```

```
True
```

In [590]:

```
if not setup_dependencies_passed:
 print("++++++++++++++++++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN THE PREVIOUS NOTEBOOK ")
 print("You did not install the required libraries. ")
 print("++++++++++++++++++++++++++++++++++++++++++++++++")
else:
 print("[OK]")
```

```
[OK]
```

In [591]:

```
%store -r setup_iam_roles_passed
```

In [592]:

```
try:
 setup_iam_roles_passed
except NameError:
 print("++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup IAM Roles.")
 print("++++++++++++++++++++++++++++++++")
```

In [593]:

```
print(setup_iam_roles_passed)
```

```
True
```

In [594]:

```
import boto3
region = boto3.Session().region_name
session = boto3.session.Session()
ec2 = boto3.Session().client(service_name="ec2", region_name=region)
sm = boto3.Session().client(service_name="sagemaker", region_name=region)
```

In [595]:

```
import json
notebook_instance_name = None
try:
 with open("/opt/ml/metadata/resource-metadata.json") as notebook_info: data = json.load(notebook_info)
 domain_id = data["DomainId"]
 resource_arn = data["ResourceArn"]
 region = resource_arn.split(":")[3]
 name = data["ResourceName"]
 print("DomainId: {}".format(domain_id))
 print("Name: {}".format(name))
except:
 print("+++++++++++++++++++++++++++++++++++++++++++")
 print("[ERROR]: COULD NOT RETRIEVE THE METADATA.")
 print("+++++++++++++++++++++++++++++++++++++++++++")
```

```
DomainId: d-2ywrjjiz4zpt
Name: datascience-1-0-ml-t3-medium-1abf3407f667f989be9d86559395
```

In [596]:

```
%store -r setup_instance_check_passed
```

In [597]:

```
try:
 setup_instance_check_passed
except NameError:
 print("+++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. Youare missing Instance Check.")
 print("+++++++++++++++++++++++++++++++++")
```

In [598]:

```
%store -r setup_dependencies_passed
```

In [599]:

```
try:
 setup_dependencies_passed
except NameError:
 print("+++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. You are missing Setup Dependencies.")
 print("+++++++++++++++++++++++++++++++++")
```

In [600]:

```
print(setup_dependencies_passed)
```

```
True
```

In [601]:

```
%store -r setup_s3_bucket_passed
```

In [602]:

```
try:
 setup_s3_bucket_passed
except NameError:
 print("+++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. Youare missing Setup Dependencies.")
 print("+++++++++++++++++++++++++++++++++")
```

In [603]:

```
print(setup_s3_bucket_passed)
```

True

In [604]:

```
if not setup_instance_check_passed:
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. Youare missing Instance Check.")
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
if not setup_dependencies_passed:
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. Youare missing Setup Dependencies.")
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
if not setup_s3_bucket_passed:
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. Youare missing Setup S3 Bucket.")
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
if not setup_iam_roles_passed:
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
 print("[ERROR] YOU HAVE TO RUN ALL NOTEBOOKS IN THE SETUP FOLDER FIRST. Youare missing Setup IAM Roles.")
 print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

In [605]:

```
!aws s3 ls s3://ads-508-azhang/finalproject/
```

```
2022-03-21 23:41:03          0
2022-03-22 02:21:16   11743774 NHSDA-1979-DS0001-data-excel.tsv
2022-03-21 23:41:22   22765996 NHSDA-1988-DS0001-data-excel.tsv
2022-03-21 23:41:22   58394554 NHSDA-1995-DS0001-data-excel.tsv
```

In [606]:

```
import boto3
import sagemaker
import pandas as pd
sess = sagemaker.Session()
bucket = sess.default_bucket()
role = sagemaker.get_execution_role()
region = boto3.Session().region_name
account_id = boto3.client("sts").get_caller_identity().get("Account")
sm = boto3.Session().client(service_name="sagemaker", region_name=region)
```

In [607]:

```
s3_public_path_tsv = "s3://ads-508-azhang/finalproject/"
```

In [608]:

```
%store s3_public_path_tsv
```

Stored 's3_public_path_tsv' (str)

In [609]:

```
s3_private_path_tsv = "s3://{}/ads-508-azhang/finalproject".format(bucket)
```

In [610]:

```
print(s3_private_path_tsv)
```

s3://sagemaker-us-east-1-189468192453/ads-508-azhang/finalproject

In [611]:

```
%store s3_private_path_tsv
```

Stored 's3_private_path_tsv' (str)

In [612]:

```
!aws s3 cp --recursive $s3_public_path_tsv/ $s3_private_path_tsv/ --exclude "*"--include "NHSDA-1988-DS0001-data-excel.tsv"
!aws s3 cp --recursive $s3_public_path_tsv/ $s3_private_path_tsv/ --exclude "*"--include "NHSDA-1995-DS0001-data-excel.tsv"
!aws s3 cp --recursive $s3_public_path_tsv/ $s3_private_path_tsv/ --exclude "*"--include "NHSDA-1979-DS0001-data-excel.tsv"
```

```
Unknown options: NHSDA-1988-DS0001-data-excel.tsv

Unknown options: NHSDA-1995-DS0001-data-excel.tsv

Unknown options: NHSDA-1979-DS0001-data-excel.tsv
```

In [613]:

```
print(s3_private_path_tsv)
```

```
s3://sagemaker-us-east-1-189468192453/ads-508-azhang/finalproject
```

In [614]:

```
!aws s3 ls $s3_private_path_tsv/
```

```
                           PRE NHSDA-1979-DS0001-data-excel/
                           PRE NHSDA-1988-DS0001-data-excel/
                           PRE NHSDA-1995-DS0001-data-excel/
                           PRE staging/
```

In [615]:

```
session = boto3.Session()
#Then use the session to get the resource
s3 = session.resource('s3')
my_bucket = s3.Bucket('ads-508-azhang')
for my_bucket_object in my_bucket.objects.all():
 print(my_bucket_object.key)
```

```
ads-508-azhang/finalproject/staging/03acdc84-f89a-45cb-a0c4-a9968b1fbd94.txt
ads-508-azhang/finalproject/staging/0ced616a-da39-4987-b490-b73016c207f2.txt
ads-508-azhang/finalproject/staging/0f98869d-8c4b-4b71-be72-1a3c58d4981d.txt
ads-508-azhang/finalproject/staging/1206ca5c-5811-4e65-b5eb-be93509b6863.csv
ads-508-azhang/finalproject/staging/1206ca5c-5811-4e65-b5eb-be93509b6863.csv.metadata
ads-508-azhang/finalproject/staging/161aaaaa-506d-4a3e-a44c-35d056b791aa.txt
ads-508-azhang/finalproject/staging/19f23706-b202-44e7-9c2a-bbf20f2dc6c8.csv
ads-508-azhang/finalproject/staging/19f23706-b202-44e7-9c2a-bbf20f2dc6c8.csv.metadata
ads-508-azhang/finalproject/staging/22bc1d08-cbd0-4a7c-9a3e-9d7d26ffa1bf.txt
ads-508-azhang/finalproject/staging/38378900-1d99-45b3-b89b-7dec3002686f.txt
ads-508-azhang/finalproject/staging/5165bdfe-9958-49c4-b579-8d2b1d15b1bb.csv
ads-508-azhang/finalproject/staging/5165bdfe-9958-49c4-b579-8d2b1d15b1bb.csv.metadata
ads-508-azhang/finalproject/staging/593190d9-e0a1-41f2-a475-befe1f2a50a3.txt
ads-508-azhang/finalproject/staging/60a34b1a-e6f8-4a73-87ca-3326dd664edc.txt
ads-508-azhang/finalproject/staging/6391eeb7-9dd6-4eb9-be54-60401aba5638.csv
ads-508-azhang/finalproject/staging/6391eeb7-9dd6-4eb9-be54-60401aba5638.csv.metadata
ads-508-azhang/finalproject/staging/6dd89416-182a-4798-8e97-b90533c6c96d.txt
ads-508-azhang/finalproject/staging/6f0335d1-764a-4311-a356-246b923f0f04.csv
ads-508-azhang/finalproject/staging/6f0335d1-764a-4311-a356-246b923f0f04.csv.metadata
ads-508-azhang/finalproject/staging/7c0d2eb9-d28b-4039-b5b7-edc2bc6f85fa.csv
ads-508-azhang/finalproject/staging/7c0d2eb9-d28b-4039-b5b7-edc2bc6f85fa.csv.metadata
ads-508-azhang/finalproject/staging/7d6d3f06-b176-44dc-ad4e-5ccde2d229c1.csv
ads-508-azhang/finalproject/staging/7d6d3f06-b176-44dc-ad4e-5ccde2d229c1.csv.metadata
ads-508-azhang/finalproject/staging/859ebbac-7d7f-40b4-8178-4a6ce23c2e34.txt
ads-508-azhang/finalproject/staging/88a91ac9-369e-43f1-894a-0763585ae553.txt
ads-508-azhang/finalproject/staging/8ba7e925-0e74-4816-b469-2c60e105e963.txt
ads-508-azhang/finalproject/staging/9107a5dc-b32d-4a1a-9896-80f8b411c6cd.txt
ads-508-azhang/finalproject/staging/9107a5dc-b32d-4a1a-9896-80f8b411c6cd.txt.metadata
ads-508-azhang/finalproject/staging/b4e8ba24-acbb-425f-bfe8-33127ecdb0b0.csv
ads-508-azhang/finalproject/staging/b4e8ba24-acbb-425f-bfe8-33127ecdb0b0.csv.metadata
ads-508-azhang/finalproject/staging/bdcb2511-4999-476f-be54-b64c016fbb4a.txt
ads-508-azhang/finalproject/staging/c1900a9d-9de1-49e7-be42-0641a73d08a3.csv
ads-508-azhang/finalproject/staging/c1900a9d-9de1-49e7-be42-0641a73d08a3.csv.metadata
ads-508-azhang/finalproject/staging/c24239bb-cb44-42df-8053-434a5c0e5e31.txt
ads-508-azhang/finalproject/staging/cd84dac9-c7f8-475c-98ba-e90db03f65b6.csv
ads-508-azhang/finalproject/staging/cd84dac9-c7f8-475c-98ba-e90db03f65b6.csv.metadata
ads-508-azhang/finalproject/staging/dee2dcbc-b9c4-447a-b0ec-9cbad91ae0c9.csv
ads-508-azhang/finalproject/staging/dee2dcbc-b9c4-447a-b0ec-9cbad91ae0c9.csv.metadata
ads-508-azhang/finalproject/staging/f5c0b7a8-59e2-4e6f-a14d-143418195a47.txt
```

ads-508-azhang/finalproject/staging/fcf313be-d649-4c21-89a5-203387fb72d3.txt
ads-508-azhang/finalproject/staging/fd5c01f0-ea81-4ebf-b8a8-f9ee9264e5f5.csv
ads-508-azhang/finalproject/staging/fd5c01f0-ea81-4ebf-b8a8-f9ee9264e5f5.csv.metadata
finalproject/
finalproject/NHSDA-1979-DS0001-data-excel.tsv
finalproject/NHSDA-1988-DS0001-data-excel.tsv
finalproject/NHSDA-1995-DS0001-data-excel.tsv

In [616]:

```
from IPython.core.display import display, HTML
display(HTML('<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/s3/buckets/sagemaker-{}-{}/ads-508-azhang/finalproject/?region={}&tab=overview">S3 Bucket</a></b>'.format(region, account_id, region)))
```

Review S3 Bucket

In [617]:

```
%store
```

Stored variables and their in-db values:
autopilot_train_s3_uri -> 's3://sagemaker-us-east-1-189468192453/data/amazon
ingest_create_athena_db_passed -> True
ingest_create_athena_table_parquet_passed -> True
s3_private_path_tsv -> 's3://sagemaker-us-east-1-189468192453/ads-508-azh
s3_public_path_tsv -> 's3://ads-508-azhang/finalproject/'
setup_dependencies_passed -> True
setup_iam_roles_passed -> True
setup_instance_check_passed -> True
setup_s3_bucket_passed -> True

In [618]:

```
#!aws s3 cp s3://ads-508-azhang/finalproject/NHSDA-1979-DS0001-data-excel.tsv -| head
```

In [619]:

```
s3_client = boto3.client("s3")
```

In [620]:

```
#Create Athena DB Schema
```

In [621]:

```
import boto3
import sagemaker
sess = sagemaker.Session()
bucket = sess.default_bucket()
role = sagemaker.get_execution_role()
region = boto3.Session().region_name
```

In [622]:

```
ingest_create_athena_db_passed = False
```

In [623]:

```
get_ipython().run_line_magic('store', '-r s3_public_path_tsv')
```

In [624]:

```
print(s3_public_path_tsv)
```

s3://ads-508-azhang/finalproject/

In [625]:

```
get_ipython().run_line_magic('store', '-r s3_private_path_tsv')
```

In [626]:

```
print(s3_private_path_tsv)
```

s3://sagemaker-us-east-1-189468192453/ads-508-azhang/finalproject

In [627]:

```
#import PyAthena
get_ipython().system('pip install --disable-pip-version-check -q PyAthena==2.1.0')
from pyathena import connect
```

/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25: CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes instead
  from cryptography.utils import int_from_bytes
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

In [628]:

```
database_name = "drugs"
```

In [629]:

```
# Set S3 staging directory -- this is a temporary directory used for Athena queries
s3_staging_dir = "s3://{0}/ads-508-azhang/finalproject/staging".format(bucket)
```

In [630]:

```
conn = connect(region_name=region, s3_staging_dir=s3_staging_dir)
```

In [631]:

```
statement = "CREATE DATABASE IF NOT EXISTS {}".format(database_name)
print(statement)
```

CREATE DATABASE IF NOT EXISTS drugs

In [632]:

```
import pandas as pd
pd.read_sql(statement, conn)
```

Out[632]:

—

In [633]:

```
statement = "SHOW DATABASES"
df_show = pd.read_sql(statement, conn)
df_show.head(5)
```

Out[633]:

	database_name
0	default
1	drugs
2	dsoaws

In [634]:

```
drug_dir = 's3://sagemaker-us-east-1-189468192453/ads-508-azhang/finalproject'
```

In [635]:

```
table_name ='NHSDA_1979'
pd.read_sql(f'DROP TABLE IF EXISTS {database_name}.{table_name}', conn)
file_name1 = 'NHSDA-1979-DS0001-data-excel.tsv'
file_name2 = 'NHSDA-1988-DS0001-data-excel.tsv'
file_name3 = 'NHSDA-1995-DS0001-data-excel.tsv'
```

In [636]:

```
create_table = f"""
CREATE EXTERNAL TABLE IF NOT EXISTS {database_name}.{table_name}(
             CASEID  float,
RESPID  float,
ENCPSU  float,
ENCSEG  float,
ENCCASE  float,
CIGMORLS  float,
CIGTRY  float,
CIG5PK  float,
CIGREC  float,
AVCIG  float,
HRDHER  float,
HRDMJ  float,
HRDCOC  float,
HRDLSD  float,
HRDBAR  float,
HRDTRN  float,
HRDAMP  float,
ADDHER  float,
ADDALC  float,
ADDMJ  float,
ADDTOB  float,
ADDBAR  float,
ADDTRN  float,
ADDAMP  float,
ADDLSD  float,
ADDCOC  float,
ADDNONE  float,
SEDLIKE  float,
SEDFEEL  float,
SEDNEED  float,
SEDREC  float,
SED30MOA  float,
SED30MOB  float,
SED30MOC  float,
SEDDAL30  float,
BUTISOL  float,
BUTICAPS  float,
AMYTAL  float,
ESKABARB  float,
LUMINAL  float,
MEBARAL  float,
AMOBARB  float,
PHENOBAR  float,
ALURATE  float,
PLACIDYL  float,
DORIDEN  float,
NOLUDAR  float,
SOPOR  float,
QUAALUDE  float,
PAREST  float,
NOCTEC  float,
METHAQ  float,
CHHYD  float,
NEMBUTAL  float,
CARBTAL  float,
SECONAL  float,
TUINAL  float,
PENTOB  float,
SECOB  float,
DALMANE  float,
SEDDKNAM  float,
NOSEDAT  float,
SEDAGE  float,
TRNLIKE  float,
TRNFEEL  float,
TRNNEED  float,
TRANREC  float,
TRN30MOA  float,
TRN30MOB  float,
TRN30MOC  float,
```

TRNBEN30 float,
VALIUM float,
LIBRIUM float,
LIBRITAB float,
SKLY float,
SERAX float,
TRANXENE float,
ATIVAN float,
VERSTRAN float,
MEPRSPAN float,
MILTOWN float,
EQUANIL float,
MEPROB float,
VISTAR float,
ATARAX float,
BENADRYL float,
TRDKNAM float,
NOTRANQ float,
TRANAGE float,
STIMLIKE float,
STIMFEEL float,
STIMNEED float,
STIMREC float,
STM30MOA float,
STM30MOB float,
STMRIT30 float,
STMCYL30 float,
DEXED float,
DEXAMYL float,
ESKAT float,
BENZ float,
BIPHET float,
DESOXYN float,
DETAMP float,
METHI float,
OBLA float,
TENUATE float,
TEPANIL float,
DIDREX float,
PLEGINE float,
PRELUDIN float,
PRESATE float,
IONAMIN float,
PONDIMIN float,
VORANIL float,
SANOREX float,
RITALIN float,
CYLERT float,
STMDKNAM float,
NOSTIMS float,
STIMAGE float,
ANALLIKE float,
ANALFEEL float,
ANALNEED float,
ANALREC float,
ANL30MOA float,
ANL30MOB float,
ANL30MOC float,
ANLTAL30 float,
DARVON float,
DOLENE float,
SK65A float,
PROPOXY float,
LERITINE float,
LEVODRO float,
PERCODAN float,
DEMEROL float,
DILAUD float,
TYLCOD float,
CODEINE float,
DOLOP float,
WESTODON float,
METHDON float,
TALWIN float,
ANLDKNAM float,
ANALNONE float,
ANALAGE float,
ALCFIRST float,
ALCTRY float,
ALCREC float,
ALCDAYS float,
MODR30A float,

```
MODR30DY float,
UNDSTAS1 float,
VRA7AS1 float,
MRKEAAS1 float,
VRA8AS1 float,
MJKNOWN float,
MJOPP float,
MJFIRST float,
MJAGE float,
MJLIVE float,
MJREC float,
MJDAY30A float,
MJTOT float,
UNDSTAS2 float,
VRM9AS2 float,
MRKEAAS2 float,
VRM10AS2 float,
INHREAD float,
INHOPP float,
INHFIRST float,
INHAGE float,
GAS float,
SPPAINT float,
AEROS float,
GLUE float,
SOLVENT float,
AMYLNIT float,
ETHER float,
NITOXID float,
ODORIZER float,
INHNEVER float,
GAS30A float,
SPPAN30A float,
AEROS30A float,
GLUE30A float,
SOLVN30A float,
AMLNT30A float,
ETHER30A float,
NOX30A float,
ODR30A float,
INH30NO float,
INHREC float,
INHTOT float,
INHODRHR float,
INHODRUS float,
UNDSTAS3 float,
VRG10AS3 float,
MRKEAAS3 float,
VRG11AS3 float,
HALLOPP float,
HALFIRST float,
HALLAGE float,
HALLREC float,
HAL30USE float,
HALLTOT float,
HALPCPHR float,
PCP float,
HALPCP30 float,
UNDSTAS4 float,
VRL10AS4 float,
MRKEAAS4 float,
VRL11AS4 float,
COCOPP float,
COCFIRST float,
COCAGE float,
COCREC float,
COCUS30A float,
COCTOT float,
UNDSTAS5 float,
VRC7AS5 float,
MRKEAAS5 float,
VRC8AS5 float,
HERKNOW float,
HEROPP float,
HERFIRST float,
HERAGE float,
HERREC float,
HER30USE float,
HERTOT float,
HERFRNDS float,
HERNOADR float,
HERNEEDL float,
UNDSTAS6 float,
```

VRH11AS6 float,
MRKEAAS6 float,
VRH12AS6 float,
SPLCOC float,
SPLHAL float,
SPLCIG float,
SPLHER float,
SPLBEER float,
SPLLQR float,
SPLMJR float,
SPLPILLS float,
SPLINH float,
GMJNOHO float,
GMJNONE float,
GMJMED float,
GMJJOB float,
GMJFUN float,
GMJRELAX float,
GMJAWARE float,
GMJCNFDN float,
GMJDEAL float,
GMJSLEEP float,
GMJSEX float,
GMJAPPET float,
GMJDK float,
GMJMISC float,
GMJREF1 float,
BMJCONTR float,
BMJMEMRY float,
BMJNONE float,
BMJHABIT float,
BMJSTRGR float,
BMJHLTH float,
BMJDIZZY float,
BMJREFLX float,
BMJMOOD float,
BMJHALLU float,
BMJAPTHY float,
BMJJOB float,
BMJDRIVE float,
BMJILLEG float,
BMJCRIME float,
BMJEXPNS float,
BMJDK float,
BMJMISC float,
BMJREF1 float,
MJHIGH float,
MJDRHIGH float,
MJOTHDR float,
MJPUFFS float,
MJDRPUFF float,
MJOTHPUF float,
MJINVOLV float,
MJCAREMR float,
MJCRMORE float,
MJOTHMOR float,
MJCARELS float,
MJCRLESS float,
MJOTHLES float,
MJWKEND float,
MJCRWKEN float,
MJOTHWKN float,
ALHIGH float,
ALDRHIGH float,
ALOTHDR float,
ALSOME float,
ALDRSOME float,
ALOTHSOM float,
ALOTHDRK float,
ALYOUDRK float,
CLOSFRNS float,
FRNSHER float,
FRNSEX float,
FRNAGE float,
FRNTRYH float,
FRNRECH float,
SEENUSE float,
CONFESS float,
TESTMNY float,
TRACKMRK float,
ARREST float,
UNPRREF float,

UNPRREP float,
UNPRBEH float,
UNPROTH float,
AMBULANC float,
DETECOTH float,
GIVESELL float,
TREATMNT float,
OTHKNOW float,
LVDHEREA float,
LVDHEREB float,
EVRLIVEA float,
AGEINA1 float,
AGEOUTA1 float,
AGEINA2 float,
AGEOUTA2 float,
AGEINA3 float,
AGEOUTA3 float,
ALLLIFEA float,
EVRLIVEB float,
AGEINB1 float,
AGEOUTB1 float,
AGEINB2 float,
AGEOUTB2 float,
AGEINB3 float,
AGEOUTB3 float,
ALLLIFEB float,
EVRLIVEC float,
AGEINC1 float,
AGEOUTC1 float,
AGEINC2 float,
AGEOUTC2 float,
AGEINC3 float,
AGEOUTC3 float,
ALLLIFEC float,
SEX float,
RESPAGE float,
HISPANIC float,
HISPGRP float,
RESPRACE float,
RAGEGRP float,
ENRLCOLL float,
TYPESCHL float,
STUDFTPT float,
EDUC float,
TOTPEOP float,
UNDAGE18 float,
UNDAGE6 float,
AGE612 float,
AGE1217 float,
HHPAREN float,
NUMPAREN float,
HHSPOUS float,
NUMSPOUS float,
HHSIBLN float,
NUMSIBLN float,
HHOTREL float,
NUMOTREL float,
HHFRNDS float,
NUMFRNDS float,
HHOTPER float,
NUMOTPER float,
MARITAL float,
EMPLOYED float,
ROCCUP2 float,
NOLABOR float,
CWE float,
CWEOCC2 float,
INCOME float,
ESTHHIN float,
YTHSTUD float,
YSTDFTPT float,
YTHEDUC float,
YTOTPEOP float,
MOTHER float,
FATHER float,
OLDSIBS float,
NUMOSIBS float,
YNGSIBS float,
NUMYSIBS float,
YTHOTREL float,
NUMYOREL float,
YTHOTPER float,
NUMYOPER float,

OTHSIBS float,
YTHEMPLD float,
YTHOCCU2 float,
YNOLABOR float,
HHAREA float,
MILINSTA float,
LOGCAMP float,
COLLEGE float,
RESORT float,
CONSTR float,
RANCH float,
MIGRANTS float,
TEMPRES float,
HHTYPE float,
UNDINT float,
COOPINT float,
PRIVACY float,
ADULTYTH float,
PAREXAMQ float,
ADLTQCD float,
QUEXTYPE float,
INTVLEN float,
FIID float,
TOTHHVIS float,
FINLRES1 float,
VSADLTCM float,
PHADLTCM float,
FINLRES2 float,
VSYTHCM float,
PHYTHCM float,
YTHINHH float,
RES1825 float,
RES2649 float,
RES50OVR float,
AGR1REL1 float,
AGR1SEX1 float,
AGR1AGE1 float,
AGR1RSP1 float,
AGR1REL2 float,
AGR1SEX2 float,
AGR1AGE2 float,
AGR1RSP2 float,
AGR1REL3 float,
AGR1SEX3 float,
AGR1AGE3 float,
AGR1RSP3 float,
AGR1REL4 float,
AGR1SEX4 float,
AGR1AGE4 float,
AGR1RSP4 float,
AGR2REL1 float,
AGR2SEX1 float,
AGR2AGE1 float,
AGR2RSP1 float,
AGR2REL2 float,
AGR2SEX2 float,
AGR2AGE2 float,
AGR2RSP2 float,
AGR2REL3 float,
AGR2SEX3 float,
AGR2AGE3 float,
AGR2RSP3 float,
AGR2REL4 float,
AGR2SEX4 float,
AGR2AGE4 float,
AGR2RSP4 float,
AGR3REL1 float,
AGR3SEX1 float,
AGR3AGE1 float,
AGR3RSP1 float,
AGR3REL2 float,
AGR3SEX2 float,
AGR3AGE2 float,
AGR3RSP2 float,
AGR3REL3 float,
AGR3SEX3 float,
AGR3AGE3 float,
AGR3RSP3 float,
AGR3REL4 float,
AGR3SEX4 float,
AGR3AGE4 float,
AGR3RSP4 float,

YTH1217 float,
YTH1REL float,
YTH1SEX float,
YTH1AGE float,
YTH1RSP float,
YTH2REL float,
YTH2SEX float,
YTH2AGE float,
YTH2RSP float,
YTH3REL float,
YTH3SEX float,
YTH3AGE float,
YTH3RSP float,
YTH4REL float,
YTH4SEX float,
YTH4AGE float,
YTH4RSP float,
REGION float,
DIVISION float,
POPDENX float,
IRAGE float,
IIAGE float,
IRSEX float,
IISEX float,
IRRACEX float,
IIRACEX float,
IRHOIND float,
IIHOIND float,
IRHOGRP float,
IIHOGRP float,
IRMARIT float,
IIMARIT float,
IREDUC float,
IIEDUC float,
IRALCRC float,
IIALCRC float,
IRMJRC float,
IIMJRC float,
IRCOCRC float,
IICOCRC float,
IRSEDRC float,
IISEDRC float,
IRTRANRC float,
IITRANRC float,
IRSTIMRC float,
IISTIMRC float,
IRANALRC float,
IIANALRC float,
IRCIGRC float,
IICIGRC float,
IRINHRC float,
IIINHRC float,
IRHALLRC float,
IIHALLRC float,
IRHERRC float,
IIHERRC float,
CATAGE float,
CATAG2 float,
CATAG3 float,
RACE float,
HISPRACE float,
EDUCCAT2 float,
HALFLAG float,
HALYR float,
HALMON float,
STMFLAG float,
STMYR float,
STMMON float,
SEDFLAG float,
SEDYR float,
SEDMON float,
TRQFLAG float,
TRQYR float,
TRQMON float,
ANLFLAG float,
ANLYR float,
ANLMON float,
ALCFLAG float,
ALCYR float,
ALCMON float,
CIGFLAG float,
CIGYR float,
CIGMON float,

```
HERFLAG  float,
HERYR  float,
HERMON  float,
MRJFLAG  float,
MRJYR  float,
MRJMON  float,
COCFLAG  float,
COCYR  float,
COCMON  float,
INHFLAG  float,
INHYR  float,
INHMON  float,
PSYFLAG2  float,
PSYYR2  float,
PSYMON2  float,
SUMFLAG  float,
SUMYR  float,
SUMMON  float,
MJOFLAG  float,
MJOYR2  float,
MJOMON2  float,
IEMFLAG  float,
IEMYR  float,
IEMMON  float,
VESTR  float,
VEREP  float,
ANALWT float,
CANALWT float,
NANALWT float,
INITWT float,
WT1 float,
WT2 float,
CINITWT float,
CWT1 float,
CWT2 float,
NINITWT float,
NWT1 float,
NWT2 float
)

            ROW FORMAT DELIMITED
            FIELDS TERMINATED BY ' '
            LINES TERMINATED BY '\n'
            LOCATION '{drug_dir}/NHSDA-1979-DS0001-data-excel'
            TBLPROPERTIES ('skip.header.line.count'='1')
"""

pd.read_sql(create_table, conn)
```

Out[636]:

—

In [637]:

```
pd.read_sql(f'SELECT count(*) FROM {database_name}.{table_name} LIMIT 5', conn)
```

Out[637]:

	_col0
0	7224

In [638]:

```
table_name2 ='NHSDA_1988'
pd.read_sql(f'DROP TABLE IF EXISTS {database_name}.{table_name2}', conn)

create_table = f"""
CREATE EXTERNAL TABLE IF NOT EXISTS {database_name}.{table_name2}(
            CASEID  float,
RESPID  float,
ENCPSU  float,
ENCSEG  float,
ENCCASE  float,
CIGMORLS  float,
CIGTRY  float,
CIG5PK  float,
CIGREC  float,
```

AVCIG float,
HRDHER float,
HRDMJ float,
HRDCOC float,
HRDLSD float,
HRDBAR float,
HRDTRN float,
HRDAMP float,
ADDHER float,
ADDALC float,
ADDMJ float,
ADDTOB float,
ADDBAR float,
ADDTRN float,
ADDAMP float,
ADDLSD float,
ADDCOC float,
ADDNONE float,
SEDLIKE float,
SEDFEEL float,
SEDNEED float,
SEDREC float,
SED30MOA float,
SED30MOB float,
SED30MOC float,
SEDDAL30 float,
BUTISOL float,
BUTICAPS float,
AMYTAL float,
ESKABARB float,
LUMINAL float,
MEBARAL float,
AMOBARB float,
PHENOBAR float,
ALURATE float,
PLACIDYL float,
DORIDEN float,
NOLUDAR float,
SOPOR float,
QUAALUDE float,
PAREST float,
NOCTEC float,
METHAQ float,
CHHYD float,
NEMBUTAL float,
CARBTAL float,
SECONAL float,
TUINAL float,
PENTOB float,
SECOB float,
DALMANE float,
SEDDKNAM float,
NOSEDAT float,
SEDAGE float,
TRNLIKE float,
TRNFEEL float,
TRNNEED float,
TRANREC float,
TRN30MOA float,
TRN30MOB float,
TRN30MOC float,
TRNBEN30 float,
VALIUM float,
LIBRIUM float,
LIBRITAB float,
SKLY float,
SERAX float,
TRANXENE float,
ATIVAN float,
VERSTRAN float,
MEPRSPAN float,
MILTOWN float,
EQUANIL float,
MEPROB float,
VISTAR float,
ATARAX float,
BENADRYL float,
TRDKNAM float,
NOTRANQ float,
TRANAGE float,
STIMLIKE float,
STIMFEEL float,
STIMNEED float,

STIMNEED float,
STIMREC float,
STM30MOA float,
STM30MOB float,
STMRIT30 float,
STMCYL30 float,
DEXED float,
DEXAMYL float,
ESKAT float,
BENZ float,
BIPHET float,
DESOXYN float,
DETAMP float,
METHI float,
OBLA float,
TENUATE float,
TEPANIL float,
DIDREX float,
PLEGINE float,
PRELUDIN float,
PRESATE float,
IONAMIN float,
PONDIMIN float,
VORANIL float,
SANOREX float,
RITALIN float,
CYLERT float,
STMDKNAM float,
NOSTIMS float,
STIMAGE float,
ANALLIKE float,
ANALFEEL float,
ANALNEED float,
ANALREC float,
ANL30MOA float,
ANL30MOB float,
ANL30MOC float,
ANLTAL30 float,
DARVON float,
DOLENE float,
SK65A float,
PROPOXY float,
LERITINE float,
LEVODRO float,
PERCODAN float,
DEMEROL float,
DILAUD float,
TYLCOD float,
CODEINE float,
DOLOP float,
WESTODON float,
METHDON float,
TALWIN float,
ANLDKNAM float,
ANALNONE float,
ANALAGE float,
ALCFIRST float,
ALCTRY float,
ALCREC float,
ALCDAYS float,
MODR30A float,
MODR30DY float,
UNDSTAS1 float,
VRA7AS1 float,
MRKEAAS1 float,
VRA8AS1 float,
MJKNOWN float,
MJOPP float,
MJFIRST float,
MJAGE float,
MJLIVE float,
MJREC float,
MJDAY30A float,
MJTOT float,
UNDSTAS2 float,
VRM9AS2 float,
MRKEAAS2 float,
VRM10AS2 float,
INHREAD float,
INHOPP float,
INHFIRST float,
INHAGE float,
GAS float,

SPPAINT float,
AEROS float,
GLUE float,
SOLVENT float,
AMYLNIT float,
ETHER float,
NITOXID float,
ODORIZER float,
INHNEVER float,
GAS30A float,
SPPAN30A float,
AEROS30A float,
GLUE30A float,
SOLVN30A float,
AMLNT30A float,
ETHER30A float,
NOX30A float,
ODR30A float,
INH30NO float,
INHREC float,
INHTOT float,
INHODRHR float,
INHODRUS float,
UNDSTAS3 float,
VRG10AS3 float,
MRKEAAS3 float,
VRG11AS3 float,
HALLOPP float,
HALFIRST float,
HALLAGE float,
HALLREC float,
HAL30USE float,
HALLTOT float,
HALPCPHR float,
PCP float,
HALPCP30 float,
UNDSTAS4 float,
VRL10AS4 float,
MRKEAAS4 float,
VRL11AS4 float,
COCOPP float,
COCFIRST float,
COCAGE float,
COCREC float,
COCUS30A float,
COCTOT float,
UNDSTAS5 float,
VRC7AS5 float,
MRKEAAS5 float,
VRC8AS5 float,
HERKNOW float,
HEROPP float,
HERFIRST float,
HERAGE float,
HERREC float,
HER30USE float,
HERTOT float,
HERFRNDS float,
HERNOADR float,
HERNEEDL float,
UNDSTAS6 float,
VRH11AS6 float,
MRKEAAS6 float,
VRH12AS6 float,
SPLCOC float,
SPLHAL float,
SPLCIG float,
SPLHER float,
SPLBEER float,
SPLLQR float,
SPLMJR float,
SPLPILLS float,
SPLINH float,
GMJNOHO float,
GMJNONE float,
GMJMED float,
GMJJOB float,
GMJFUN float,
GMJRELAX float,
GMJAWARE float,
GMJCNFDN float,
GMJDEAL float,
GMJSLEEP float,

GMJSEX float,
GMJAPPET float,
GMJDK float,
GMJMISC float,
GMJREF1 float,
BMJCONTR float,
BMJMEMRY float,
BMJNONE float,
BMJHABIT float,
BMJSTRGR float,
BMJHLTH float,
BMJDIZZY float,
BMJREFLX float,
BMJMOOD float,
BMJHALLU float,
BMJAPTHY float,
BMJJOB float,
BMJDRIVE float,
BMJILLEG float,
BMJCRIME float,
BMJEXPNS float,
BMJDK float,
BMJMISC float,
BMJREF1 float,
MJHIGH float,
MJDRHIGH float,
MJOTHDR float,
MJPUFFS float,
MJDRPUFF float,
MJOTHPUF float,
MJINVOLV float,
MJCAREMR float,
MJCRMORE float,
MJOTHMOR float,
MJCARELS float,
MJCRLESS float,
MJOTHLES float,
MJWKEND float,
MJCRWKEN float,
MJOTHWKN float,
ALHIGH float,
ALDRHIGH float,
ALOTHDR float,
ALSOME float,
ALDRSOME float,
ALOTHSOM float,
ALOTHDRK float,
ALYOUDRK float,
CLOSFRNS float,
FRNSHER float,
FRNSEX float,
FRNAGE float,
FRNTRYH float,
FRNRECH float,
SEENUSE float,
CONFESS float,
TESTMNY float,
TRACKMRK float,
ARREST float,
UNPRREF float,
UNPRREP float,
UNPRBEH float,
UNPROTH float,
AMBULANC float,
DETECOTH float,
GIVESELL float,
TREATMNT float,
OTHKNOW float,
LVDHEREA float,
LVDHEREB float,
EVRLIVEA float,
AGEINA1 float,
AGEOUTA1 float,
AGEINA2 float,
AGEOUTA2 float,
AGEINA3 float,
AGEOUTA3 float,
ALLLIFEA float,
EVRLIVEB float,
AGEINB1 float,
AGEOUTB1 float,
AGEINB2 float,

```
AGEOUTB2 float,
AGEINB3 float,
AGEOUTB3 float,
ALLLIFEB float,
EVRLIVEC float,
AGEINC1 float,
AGEOUTC1 float,
AGEINC2 float,
AGEOUTC2 float,
AGEINC3 float,
AGEOUTC3 float,
ALLLIFEC float,
SEX float,
RESPAGE float,
HISPANIC float,
HISPGRP float,
RESPRACE float,
RAGEGRP float,
ENRLCOLL float,
TYPESCHL float,
STUDFTPT float,
EDUC float,
TOTPEOP float,
UNDAGE18 float,
UNDAGE6 float,
AGE612 float,
AGE1217 float,
HHPAREN float,
NUMPAREN float,
HHSPOUS float,
NUMSPOUS float,
HHSIBLN float,
NUMSIBLN float,
HHOTREL float,
NUMOTREL float,
HHFRNDS float,
NUMFRNDS float,
HHOTPER float,
NUMOTPER float,
MARITAL float,
EMPLOYED float,
ROCCUP2 float,
NOLABOR float,
CWE float,
CWEOCC2 float,
INCOME float,
ESTHHIN float,
YTHSTUD float,
YSTDFTPT float,
YTHEDUC float,
YTOTPEOP float,
MOTHER float,
FATHER float,
OLDSIBS float,
NUMOSIBS float,
YNGSIBS float,
NUMYSIBS float,
YTHOTREL float,
NUMYOREL float,
YTHOTPER float,
NUMYOPER float,
OTHSIBS float,
YTHEMPLD float,
YTHOCCU2 float,
YNOLABOR float,
HHAREA float,
MILINSTA float,
LOGCAMP float,
COLLEGE float,
RESORT float,
CONSTR float,
RANCH float,
MIGRANTS float,
TEMPRES float,
HHTYPE float,
UNDINT float,
COOPINT float,
PRIVACY float,
ADULTYTH float,
PAREXAMQ float,
ADLTQCD float,
QUEXTYPE float,
INTVLEN float,
```

INTVLEN float,
FIID float,
TOTHHVIS float,
FINLRES1 float,
VSADLTCM float,
PHADLTCM float,
FINLRES2 float,
VSYTHCM float,
PHYTHCM float,
YTHINHH float,
RES1825 float,
RES2649 float,
RES50OVR float,
AGR1REL1 float,
AGR1SEX1 float,
AGR1AGE1 float,
AGR1RSP1 float,
AGR1REL2 float,
AGR1SEX2 float,
AGR1AGE2 float,
AGR1RSP2 float,
AGR1REL3 float,
AGR1SEX3 float,
AGR1AGE3 float,
AGR1RSP3 float,
AGR1REL4 float,
AGR1SEX4 float,
AGR1AGE4 float,
AGR1RSP4 float,
AGR2REL1 float,
AGR2SEX1 float,
AGR2AGE1 float,
AGR2RSP1 float,
AGR2REL2 float,
AGR2SEX2 float,
AGR2AGE2 float,
AGR2RSP2 float,
AGR2REL3 float,
AGR2SEX3 float,
AGR2AGE3 float,
AGR2RSP3 float,
AGR2REL4 float,
AGR2SEX4 float,
AGR2AGE4 float,
AGR2RSP4 float,
AGR3REL1 float,
AGR3SEX1 float,
AGR3AGE1 float,
AGR3RSP1 float,
AGR3REL2 float,
AGR3SEX2 float,
AGR3AGE2 float,
AGR3RSP2 float,
AGR3REL3 float,
AGR3SEX3 float,
AGR3AGE3 float,
AGR3RSP3 float,
AGR3REL4 float,
AGR3SEX4 float,
AGR3AGE4 float,
AGR3RSP4 float,
YTH1217 float,
YTH1REL float,
YTH1SEX float,
YTH1AGE float,
YTH1RSP float,
YTH2REL float,
YTH2SEX float,
YTH2AGE float,
YTH2RSP float,
YTH3REL float,
YTH3SEX float,
YTH3AGE float,
YTH3RSP float,
YTH4REL float,
YTH4SEX float,
YTH4AGE float,
YTH4RSP float,
REGION float,
DIVISION float,
POPDENX float,
IRAGE float,
IIAGE float,

IRSEX float,
IISEX float,
IRRACEX float,
IIRACEX float,
IRHOIND float,
IIHOIND float,
IRHOGRP float,
IIHOGRP float,
IRMARIT float,
IIMARIT float,
IREDUC float,
IIEDUC float,
IRALCRC float,
IIALCRC float,
IRMJRC float,
IIMJRC float,
IRCOCRC float,
IICOCRC float,
IRSEDRC float,
IISEDRC float,
IRTRANRC float,
IITRANRC float,
IRSTIMRC float,
IISTIMRC float,
IRANALRC float,
IIANALRC float,
IRCIGRC float,
IICIGRC float,
IRINHRC float,
IIINHRC float,
IRHALLRC float,
IIHALLRC float,
IRHERRC float,
IIHERRC float,
CATAGE float,
CATAG2 float,
CATAG3 float,
RACE float,
HISPRACE float,
EDUCCAT2 float,
HALFLAG float,
HALYR float,
HALMON float,
STMFLAG float,
STMYR float,
STMMON float,
SEDFLAG float,
SEDYR float,
SEDMON float,
TRQFLAG float,
TRQYR float,
TRQMON float,
ANLFLAG float,
ANLYR float,
ANLMON float,
ALCFLAG float,
ALCYR float,
ALCMON float,
CIGFLAG float,
CIGYR float,
CIGMON float,
HERFLAG float,
HERYR float,
HERMON float,
MRJFLAG float,
MRJYR float,
MRJMON float,
COCFLAG float,
COCYR float,
COCMON float,
INHFLAG float,
INHYR float,
INHMON float,
PSYFLAG2 float,
PSYYR2 float,
PSYMON2 float,
SUMFLAG float,
SUMYR float,
SUMMON float,
MJOFLAG float,
MJOYR2 float,
MJOMON2 float,
IFMFLAG float,

```
IEMYR  float,
IEMMON  float,
VESTR  float,
VEREP  float,
ANALWT float,
CANALWT float,
NANALWT float,
INITWT float,
WT1 float,
WT2 float,
CINITWT float,
CWT1 float,
CWT2 float,
NINITWT float,
NWT1 float,
NWT2 float
)

            ROW FORMAT DELIMITED
            FIELDS TERMINATED BY ' '
            LINES TERMINATED BY '\n'
            LOCATION '{drug_dir}/NHSDA-1988-DS0001-data-excel'
            TBLPROPERTIES ('skip.header.line.count'='1')
"""

pd.read_sql(create_table, conn)
```

Out[638]:

—

In [639]:

```
pd.read_sql(f'SELECT count(*) FROM {database_name}.{table_name2} LIMIT 5', conn)
```

Out[639]:

	_col0
0	8814

In [640]:

```
table_name3 ='NHSDA_1995'
pd.read_sql(f'DROP TABLE IF EXISTS {database_name}.{table_name3}', conn)

create_table = f"""
CREATE EXTERNAL TABLE IF NOT EXISTS {database_name}.{table_name3}(
            CASEID  float,
RESPID  float,
ENCPSU  float,
ENCSEG  float,
ENCCASE  float,
CIGMORLS  float,
CIGTRY  float,
CIG5PK  float,
CIGREC  float,
AVCIG  float,
HRDHER  float,
HRDMJ  float,
HRDCOC  float,
HRDLSD  float,
HRDBAR  float,
HRDTRN  float,
HRDAMP  float,
ADDHER  float,
ADDALC  float,
ADDMJ  float,
ADDTOB  float,
ADDBAR  float,
ADDTRN  float,
ADDAMP  float,
ADDLSD  float,
ADDCOC  float,
ADDNONE  float,
SEDLIKE  float,
SEDFEEL  float,
SEDNEED  float,
SEDREC  float,
```

SED30MOA float,
SED30MOB float,
SED30MOC float,
SEDDAL30 float,
BUTISOL float,
BUTICAPS float,
AMYTAL float,
ESKABARB float,
LUMINAL float,
MEBARAL float,
AMOBARB float,
PHENOBAR float,
ALURATE float,
PLACIDYL float,
DORIDEN float,
NOLUDAR float,
SOPOR float,
QUAALUDE float,
PAREST float,
NOCTEC float,
METHAQ float,
CHHYD float,
NEMBUTAL float,
CARBTAL float,
SECONAL float,
TUINAL float,
PENTOB float,
SECOB float,
DALMANE float,
SEDDKNAM float,
NOSEDAT float,
SEDAGE float,
TRNLIKE float,
TRNFEEL float,
TRNNEED float,
TRANREC float,
TRN30MOA float,
TRN30MOB float,
TRN30MOC float,
TRNBEN30 float,
VALIUM float,
LIBRIUM float,
LIBRITAB float,
SKLY float,
SERAX float,
TRANXENE float,
ATIVAN float,
VERSTRAN float,
MEPRSPAN float,
MILTOWN float,
EQUANIL float,
MEPROB float,
VISTAR float,
ATARAX float,
BENADRYL float,
TRDKNAM float,
NOTRANQ float,
TRANAGE float,
STIMLIKE float,
STIMFEEL float,
STIMNEED float,
STIMREC float,
STM30MOA float,
STM30MOB float,
STMRIT30 float,
STMCYL30 float,
DEXED float,
DEXAMYL float,
ESKAT float,
BENZ float,
BIPHET float,
DESOXYN float,
DETAMP float,
METHI float,
OBLA float,
TENUATE float,
TEPANIL float,
DIDREX float,
PLEGINE float,
PRELUDIN float,
PRESATE float,
IONAMIN float,

PONDIMIN float,
VORANIL float,
SANOREX float,
RITALIN float,
CYLERT float,
STMDKNAM float,
NOSTIMS float,
STIMAGE float,
ANALLIKE float,
ANALFEEL float,
ANALNEED float,
ANALREC float,
ANL30MOA float,
ANL30MOB float,
ANL30MOC float,
ANLTAL30 float,
DARVON float,
DOLENE float,
SK65A float,
PROPOXY float,
LERITINE float,
LEVODRO float,
PERCODAN float,
DEMEROL float,
DILAUD float,
TYLCOD float,
CODEINE float,
DOLOP float,
WESTODON float,
METHDON float,
TALWIN float,
ANLDKNAM float,
ANALNONE float,
ANALAGE float,
ALCFIRST float,
ALCTRY float,
ALCREC float,
ALCDAYS float,
MODR30A float,
MODR30DY float,
UNDSTAS1 float,
VRA7AS1 float,
MRKEAAS1 float,
VRA8AS1 float,
MJKNOWN float,
MJOPP float,
MJFIRST float,
MJAGE float,
MJLIVE float,
MJREC float,
MJDAY30A float,
MJTOT float,
UNDSTAS2 float,
VRM9AS2 float,
MRKEAAS2 float,
VRM10AS2 float,
INHREAD float,
INHOPP float,
INHFIRST float,
INHAGE float,
GAS float,
SPPAINT float,
AEROS float,
GLUE float,
SOLVENT float,
AMYLNIT float,
ETHER float,
NITOXID float,
ODORIZER float,
INHNEVER float,
GAS30A float,
SPPAN30A float,
AEROS30A float,
GLUE30A float,
SOLVN30A float,
AMLNT30A float,
ETHER30A float,
NOX30A float,
ODR30A float,
INH30NO float,
INHREC float,
INHTOT float,
INHODRHR float,

INHODRUS float,
UNDSTAS3 float,
VRG10AS3 float,
MRKEAAS3 float,
VRG11AS3 float,
HALLOPP float,
HALFIRST float,
HALLAGE float,
HALLREC float,
HAL30USE float,
HALLTOT float,
HALPCPHR float,
PCP float,
HALPCP30 float,
UNDSTAS4 float,
VRL10AS4 float,
MRKEAAS4 float,
VRL11AS4 float,
COCOPP float,
COCFIRST float,
COCAGE float,
COCREC float,
COCUS30A float,
COCTOT float,
UNDSTAS5 float,
VRC7AS5 float,
MRKEAAS5 float,
VRC8AS5 float,
HERKNOW float,
HEROPP float,
HERFIRST float,
HERAGE float,
HERREC float,
HER30USE float,
HERTOT float,
HERFRNDS float,
HERNOADR float,
HERNEEDL float,
UNDSTAS6 float,
VRH11AS6 float,
MRKEAAS6 float,
VRH12AS6 float,
SPLCOC float,
SPLHAL float,
SPLCIG float,
SPLHER float,
SPLBEER float,
SPLLQR float,
SPLMJR float,
SPLPILLS float,
SPLINH float,
GMJNOHO float,
GMJNONE float,
GMJMED float,
GMJJOB float,
GMJFUN float,
GMJRELAX float,
GMJAWARE float,
GMJCNFDN float,
GMJDEAL float,
GMJSLEEP float,
GMJSEX float,
GMJAPPET float,
GMJDK float,
GMJMISC float,
GMJREF1 float,
BMJCONTR float,
BMJMEMRY float,
BMJNONE float,
BMJHABIT float,
BMJSTRGR float,
BMJHLTH float,
BMJDIZZY float,
BMJREFLX float,
BMJMOOD float,
BMJHALLU float,
BMJAPTHY float,
BMJJOB float,
BMJDRIVE float,
BMJILLEG float,
BMJCRIME float,
BMJEXPNS float,

BMJDK float,
BMJMISC float,
BMJREF1 float,
MJHIGH float,
MJDRHIGH float,
MJOTHDR float,
MJPUFFS float,
MJDRPUFF float,
MJOTHPUF float,
MJINVOLV float,
MJCAREMR float,
MJCRMORE float,
MJOTHMOR float,
MJCARELS float,
MJCRLESS float,
MJOTHLES float,
MJWKEND float,
MJCRWKEN float,
MJOTHWKN float,
ALHIGH float,
ALDRHIGH float,
ALOTHDR float,
ALSOME float,
ALDRSOME float,
ALOTHSOM float,
ALOTHDRK float,
ALYOUDRK float,
CLOSFRNS float,
FRNSHER float,
FRNSEX float,
FRNAGE float,
FRNTRYH float,
FRNRECH float,
SEENUSE float,
CONFESS float,
TESTMNY float,
TRACKMRK float,
ARREST float,
UNPRREF float,
UNPRREP float,
UNPRBEH float,
UNPROTH float,
AMBULANC float,
DETECOTH float,
GIVESELL float,
TREATMNT float,
OTHKNOW float,
LVDHEREA float,
LVDHEREB float,
EVRLIVEA float,
AGEINA1 float,
AGEOUTA1 float,
AGEINA2 float,
AGEOUTA2 float,
AGEINA3 float,
AGEOUTA3 float,
ALLLIFEA float,
EVRLIVEB float,
AGEINB1 float,
AGEOUTB1 float,
AGEINB2 float,
AGEOUTB2 float,
AGEINB3 float,
AGEOUTB3 float,
ALLLIFEB float,
EVRLIVEC float,
AGEINC1 float,
AGEOUTC1 float,
AGEINC2 float,
AGEOUTC2 float,
AGEINC3 float,
AGEOUTC3 float,
ALLLIFEC float,
SEX float,
RESPAGE float,
HISPANIC float,
HISPGRP float,
RESPRACE float,
RAGEGRP float,
ENRLCOLL float,
TYPESCHL float,
STUDFTPT float,
EDUC float,

TOTPEOP float,
UNDAGE18 float,
UNDAGE6 float,
AGE612 float,
AGE1217 float,
HHPAREN float,
NUMPAREN float,
HHSPOUS float,
NUMSPOUS float,
HHSIBLN float,
NUMSIBLN float,
HHOTREL float,
NUMOTREL float,
HHFRNDS float,
NUMFRNDS float,
HHOTPER float,
NUMOTPER float,
MARITAL float,
EMPLOYED float,
ROCCUP2 float,
NOLABOR float,
CWE float,
CWEOCC2 float,
INCOME float,
ESTHHIN float,
YTHSTUD float,
YSTDFTPT float,
YTHEDUC float,
YTOTPEOP float,
MOTHER float,
FATHER float,
OLDSIBS float,
NUMOSIBS float,
YNGSIBS float,
NUMYSIBS float,
YTHOTREL float,
NUMYOREL float,
YTHOTPER float,
NUMYOPER float,
OTHSIBS float,
YTHEMPLD float,
YTHOCCU2 float,
YNOLABOR float,
HHAREA float,
MILINSTA float,
LOGCAMP float,
COLLEGE float,
RESORT float,
CONSTR float,
RANCH float,
MIGRANTS float,
TEMPRES float,
HHTYPE float,
UNDINT float,
COOPINT float,
PRIVACY float,
ADULTYTH float,
PAREXAMQ float,
ADLTQCD float,
QUEXTYPE float,
INTVLEN float,
FIID float,
TOTHHVIS float,
FINLRES1 float,
VSADLTCM float,
PHADLTCM float,
FINLRES2 float,
VSYTHCM float,
PHYTHCM float,
YTHINHH float,
RES1825 float,
RES2649 float,
RES50OVR float,
AGR1REL1 float,
AGR1SEX1 float,
AGR1AGE1 float,
AGR1RSP1 float,
AGR1REL2 float,
AGR1SEX2 float,
AGR1AGE2 float,
AGR1RSP2 float,
AGR1REL3 float,

AGR1SEX3 float,
AGR1AGE3 float,
AGR1RSP3 float,
AGR1REL4 float,
AGR1SEX4 float,
AGR1AGE4 float,
AGR1RSP4 float,
AGR2REL1 float,
AGR2SEX1 float,
AGR2AGE1 float,
AGR2RSP1 float,
AGR2REL2 float,
AGR2SEX2 float,
AGR2AGE2 float,
AGR2RSP2 float,
AGR2REL3 float,
AGR2SEX3 float,
AGR2AGE3 float,
AGR2RSP3 float,
AGR2REL4 float,
AGR2SEX4 float,
AGR2AGE4 float,
AGR2RSP4 float,
AGR3REL1 float,
AGR3SEX1 float,
AGR3AGE1 float,
AGR3RSP1 float,
AGR3REL2 float,
AGR3SEX2 float,
AGR3AGE2 float,
AGR3RSP2 float,
AGR3REL3 float,
AGR3SEX3 float,
AGR3AGE3 float,
AGR3RSP3 float,
AGR3REL4 float,
AGR3SEX4 float,
AGR3AGE4 float,
AGR3RSP4 float,
YTH1217 float,
YTH1REL float,
YTH1SEX float,
YTH1AGE float,
YTH1RSP float,
YTH2REL float,
YTH2SEX float,
YTH2AGE float,
YTH2RSP float,
YTH3REL float,
YTH3SEX float,
YTH3AGE float,
YTH3RSP float,
YTH4REL float,
YTH4SEX float,
YTH4AGE float,
YTH4RSP float,
REGION float,
DIVISION float,
POPDENX float,
IRAGE float,
IIAGE float,
IRSEX float,
IISEX float,
IRRACEX float,
IIRACEX float,
IRHOIND float,
IIHOIND float,
IRHOGRP float,
IIHOGRP float,
IRMARIT float,
IIMARIT float,
IREDUC float,
IIEDUC float,
IRALCRC float,
IIALCRC float,
IRMJRC float,
IIMJRC float,
IRCOCRC float,
IICOCRC float,
IRSEDRC float,
IISEDRC float,
IRTRANRC float,
IITRANRC float,

```
IRSTIMRC  float,
IISTIMRC  float,
IRANALRC  float,
IIANALRC  float,
IRCIGRC  float,
IICIGRC  float,
IRINHRC  float,
IIINHRC  float,
IRHALLRC  float,
IIHALLRC  float,
IRHERRC  float,
IIHERRC  float,
CATAGE  float,
CATAG2  float,
CATAG3  float,
RACE  float,
HISPRACE  float,
EDUCCAT2  float,
HALFLAG  float,
HALYR  float,
HALMON  float,
STMFLAG  float,
STMYR  float,
STMMON  float,
SEDFLAG  float,
SEDYR  float,
SEDMON  float,
TRQFLAG  float,
TRQYR  float,
TRQMON  float,
ANLFLAG  float,
ANLYR  float,
ANLMON  float,
ALCFLAG  float,
ALCYR  float,
ALCMON  float,
CIGFLAG  float,
CIGYR  float,
CIGMON  float,
HERFLAG  float,
HERYR  float,
HERMON  float,
MRJFLAG  float,
MRJYR  float,
MRJMON  float,
COCFLAG  float,
COCYR  float,
COCMON  float,
INHFLAG  float,
INHYR  float,
INHMON  float,
PSYFLAG2  float,
PSYYR2  float,
PSYMON2  float,
SUMFLAG  float,
SUMYR  float,
SUMMON  float,
MJOFLAG  float,
MJOYR2  float,
MJOMON2  float,
IEMFLAG  float,
IEMYR  float,
IEMMON  float,
VESTR  float,
VEREP  float,
ANALWT float,
CANALWT float,
NANALWT float,
INITWT float,
WT1 float,
WT2 float,
CINITWT float,
CWT1 float,
CWT2 float,
NINITWT float,
NWT1 float,
NWT2 float
)

        ROW FORMAT DELIMITED
        FIELDS TERMINATED BY ' '
        LINES TERMINATED BY '\n'
        LOCATION '[drug_dir]/NHSDA_1995_DS0001_data_excel'
```

```
            LOCATION '{drug_dir}/NHSDA-1995-DS0001-data-excel'
            TBLPROPERTIES ('skip.header.line.count'='1')
"""

pd.read_sql(create_table, conn)
```

Out[640]:

—

In [641]:

```
pd.read_sql(f'SELECT count(*) FROM {database_name}.{table_name3} LIMIT 5', conn)
```

Out[641]:

	_col0
0	17747

In [642]:

```
pd.read_sql(f'SELECT * FROM {database_name}.{table_name3} LIMIT 2', conn)
```

Out[642]:

	caseid	respid	encpsu	encseg	enccase	cigmorls	cigtry	cig5pk	cigrec	avcig	...	nanalwt	initwt	wt1	wt2	cinitwt	cwt1	cwt2	ninitwt	nv
0	8883.0	89789.0	9404.0	2.0	59.0	705.0	9462.0	2.0	1.0	1.0	...	99.0	99.0	99.0	99.0	99.0	99.0	99.0	2.0	
1	8884.0	89797.0	9415.0	1.0	39.0	795.0	1548.0	2.0	3.0	1.0	...	99.0	99.0	99.0	99.0	99.0	99.0	99.0	2.0	

2 rows × 603 columns

In [643]:

```
pd.read_sql(f'SELECT * FROM {database_name}.{table_name2} LIMIT 2', conn)
```

Out[643]:

	caseid	respid	encpsu	encseg	enccase	cigmorls	cigtry	cig5pk	cigrec	avcig	...	nanalwt	initwt	wt1	wt2	cinitwt	cwt1	cwt2	ninitwt	nwt1
0	1.0	224.0	65.0	1397.0	6360.0	23.0	1.0	5.0	99.0	99.0	...	9.0	1.0	5.0	1.0	9.0	1.0	9.0	1.0	9.0
1	2.0	1032.0	91.0	524.0	260.0	14.0	1.0	1.0	2.0	4.0	...	9.0	1.0	1.0	1.0	1.0	1.0	9.0	1.0	9.0

2 rows × 603 columns

In [644]:

```
pd.read_sql(f'SELECT * FROM {database_name}.{table_name} LIMIT 2', conn)
```

Out[644]:

	caseid	respid	encpsu	encseg	enccase	cigmorls	cigtry	cig5pk	cigrec	avcig	...	nanalwt	initwt	wt1	wt2	cinitwt	cwt1	cw
0	1.0	1214.0	63.0	151.0	2040.0	3.0	16.0	1.0	4.0	99.0	...	0.0	3.2720	24964.414	1.0096	3.2720	31239.281	0.992
1	2.0	1478.0	63.0	151.0	5931.0	3.0	14.0	2.0	19.0	99.0	...	0.0	1.5764	24964.414	1.0278	1.5764	31239.281	1.055

2 rows × 603 columns

In [645]:

```
pd.read_sql(f'DROP VIEW IF EXISTS all_record', conn)
```

Out[645]:

—

In [646]:

```
pd.read_sql(f'create view all_record as SELECT * FROM {database_name}.{table_name} union all SELECT * FROM {database_name}.{table_name
2} union all SELECT * FROM {database_name}.{table_name3}', conn)
```

```
2} union all SELECT * FROM {database_name}.{table_names3}", conn)
```

Out[646]:

—

In [647]:

```
pd.read_sql(f'SELECT count(*) FROM all_record', conn)
```

Out[647]:

	_col0
0	33785

In [648]:

```
df = pd.read_sql(f'SELECT * FROM all_record', conn)
```

In [649]:

```
print('Number of Rows:', df.shape[0])
print('Number of Columns:', df.shape[1], '\n')
data_types = df.dtypes
data_types = pd.DataFrame(data_types)
data_types = data_types.assign(Null_Values = df.isnull().sum())
data_types.reset_index(inplace = True)
data_types.rename(columns={0:'Data Type',
 'index': 'Column/Variable',
 'Null_Values': "# of Nulls"})
```

Number of Rows: 33785
Number of Columns: 603

Out[649]:

	Column/Variable	Data Type	# of Nulls
0	caseid	float64	0
1	respid	float64	0
2	encpsu	float64	0
3	encseg	float64	0
4	enccase	float64	0
...	...	...	...
598	cwt1	float64	0
599	cwt2	float64	0
600	ninitwt	float64	0
601	nwt1	float64	0
602	nwt2	float64	0

603 rows × 3 columns

In [650]:

```
df.corr
```

Out[650]:

```
<bound method DataFrame.corr of       caseid  respid  encpsu  encseg  enccase  cigmorls  cigtry  cig5pk  \
0      175.0   1625.0    26.0   337.0    908.0       1.0    91.0    91.0
1      176.0   1628.0    26.0   337.0   4023.0       1.0    11.0     1.0
2      177.0   2394.0    26.0   337.0   3445.0       1.0    91.0    91.0
3      178.0   4666.0    26.0   337.0   1244.0       1.0     9.0     1.0
4      179.0   6081.0    26.0   337.0   5800.0       3.0    15.0     1.0
...      ...      ...     ...     ...      ...       ...     ...     ...
33780  8878.0  89722.0  9425.0     1.0     34.0    1659.0  3225.0     2.0
33781  8879.0  89730.0  9519.0     1.0     93.0     666.0  6209.0     2.0
33782  8880.0  89748.0  9527.0     1.0     54.0    1132.0  2673.0     2.0
```

```
33783  8881.0  89755.0  9435.0    2.0    97.0    883.0  8452.0    1.0
33784  8882.0  89763.0  9433.0    1.0    66.0    314.0  7070.0    2.0

      cigrec  avcig  ...    nanalwt   initwt       wt1      wt2  cinitwt  \
0       91.0   91.0  ...       0.00   3.5219  24964.414   1.0096   3.5219
1        1.0    6.0  ...       0.00   1.3868  24964.414   1.9451   1.3868
2       91.0   91.0  ...       0.00   4.1575  24964.414   0.7451   4.1575
3        1.0    6.0  ...       0.00   0.8016  24964.414   0.9488   0.8016
4        1.0    4.0  ...  129973.12   0.9311  24964.414   0.9780   0.0000
...      ...    ...  ...        ...      ...        ...      ...      ...
33780    1.0    1.0  ...      99.00  99.0000     99.000  99.0000  99.0000
33781    4.0   99.0  ...      99.00  99.0000     99.000  99.0000  99.0000
33782    4.0   99.0  ...      99.00  99.0000     99.000  99.0000  99.0000
33783   99.0   99.0  ...      99.00  99.0000     99.000  99.0000  99.0000
33784    3.0    1.0  ...      99.00  99.0000     99.000  99.0000  99.0000

          cwt1     cwt2  ninitwt       nwt1    nwt2
0    31239.281   0.9923   0.0000  24964.168  1.0105
1    31239.281   1.8513   0.0000  24964.168  2.2689
2    31239.281   0.7352   0.0000  24964.168  0.7331
3    31239.281   0.8522   0.0000  24964.168  1.6097
4    31239.281   0.9473   4.9778  24964.168  1.0459
...        ...      ...      ...        ...     ...
33780   99.000  99.0000   2.0000      2.000  2.0000
33781   99.000  99.0000   2.0000      2.000  2.0000
33782   99.000  99.0000   2.0000      2.000  2.0000
33783   99.000  99.0000   2.0000      2.000  2.0000
33784   99.000  99.0000   2.0000      2.000  2.0000

[33785 rows x 603 columns]>
```

In [651]:

```
print(df.shape)
```

```
(33785, 603)
```

In [652]:

```
print(df.columns)
```

```
Index(['caseid', 'respid', 'encpsu', 'encseg', 'enccase', 'cigmorls', 'cigtry',
       'cig5pk', 'cigrec', 'avcig',
       ...
       'nanalwt', 'initwt', 'wt1', 'wt2', 'cinitwt', 'cwt1', 'cwt2', 'ninitwt',
       'nwt1', 'nwt2'],
      dtype='object', length=603)
```

In [653]:

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33785 entries, 0 to 33784
Columns: 603 entries, caseid to nwt2
dtypes: float64(603)
memory usage: 155.4 MB
None
```

In [654]:

```
df.describe()
```

Out[654]:

	caseid	respid	encpsu	encseg	enccase	cigmorls	cigtry	cig5pk	cigrec	avci
count	33785.000000	33785.000000	33785.000000	33785.000000	33785.000000	33785.000000	33785.000000	33785.000000	33785.000000	33785.00000
mean	6583.728963	60815.072991	4998.708865	253.390824	1741.831760	509.685837	4181.251591	15.828415	33.627793	63.35980
std	4720.702219	53405.573841	4701.591161	397.733889	2317.336228	608.026076	5152.547735	32.193161	42.712597	45.18596
min	1.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000
25%	2816.000000	9375.000000	56.000000	1.000000	61.000000	11.000000	14.000000	2.000000	2.000000	3.00000
50%	5631.000000	49692.000000	9404.000000	2.000000	109.000000	95.000000	762.000000	2.000000	4.000000	91.00000

	caseid	respid	encpsu	encseg	enccase	cigmorls	cigtry	cig5pk	cigrec	avci
75%	9301.000000	94029.000000	9501.000000	368.000000	3321.000000	997.000000	8320.000000	3.000000	91.000000	99.00000
max	17747.000000	182295.000000	9536.000000	1532.000000	8214.000000	1908.000000	15855.000000	98.000000	99.000000	99.00000

8 rows × 603 columns

In [655]:

```
#drop columns with missing values
df = df.dropna(axis='columns')
```

In [656]:

```
print(df.shape)
```

(33785, 599)

In [657]:

```
#Calculate correlation between all columns and remove highly correlated columns
#import numpy as np

# Create correlation matrix
#corr_matrix = df.corr().abs()

# Select upper triangle of correlation matrix
#upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool))

# Find features with correlation greater than 0.8
#to_drop = [column for column in upper.columns if any(upper[column] > 0.90)]

# Drop features
#df.drop(to_drop, axis=1, inplace=True)
```

In [658]:

```
from pandas import read_csv
from numpy import set_printoptions
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
```

In [659]:

```
#get index of target column
df.columns.get_loc("herneedl")
```

Out[659]:

233

In [660]:

```
# Remove all columns between column index 234 to 603
df.drop(df.iloc[:, 234:599], inplace = True, axis = 1)
```

```
/opt/conda/lib/python3.7/site-packages/pandas/core/frame.py:3997: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  errors=errors,
```

In [661]:

```
print(df.shape)
```

(33785, 234)

In [662]:

```
print(df.columns)
```

```
Index(['caseid', 'respid', 'encpsu', 'encseg', 'enccase', 'cigmorls', 'cigtry',
```

```
Index(['caseid', 'respid', 'encpsu', 'encseg', 'enccase', 'cigmoris', 'cigtry',
       'cig5pk', 'cigrec', 'avcig',
       ...
       'herknow', 'heropp', 'herfirst', 'herage', 'herrec', 'her30use',
       'hertot', 'herfrnds', 'hernoadr', 'herneedl'],
      dtype='object', length=234)
```

In [663]:

```
array = df.values
X = array[:,0:233]
Y = array[:,233]
```

In [664]:

```
# feature extraction
test = SelectKBest(score_func=f_classif, k=15)
fit = test.fit(X, Y)
```

In [665]:

```
# summarize scores
set_printoptions(precision=3)
print(fit.scores_)
```

```
[1.686e+03 2.554e+03 7.021e+05 3.927e+03 6.628e+03 6.384e+03 6.698e+03
 1.317e+03 1.164e+03 1.484e+02 1.801e+03 1.555e+02 3.034e+03 8.049e+02
 2.346e+01 2.961e+03 1.023e+03 2.823e+02 1.303e+03 4.166e+03 4.162e+02
 5.109e+02 1.408e+03 4.049e+02 1.417e+01 2.377e+03 1.305e+03 1.593e+03
 7.336e+02 4.646e+03 4.059e+03 5.235e+03 3.919e+02 5.032e+03 4.144e+03
 4.895e+03 6.234e+02 1.727e+02 2.318e+03 5.813e+02 1.054e+03 1.195e+03
 2.857e+02 5.082e+04 3.659e+03 7.201e+02 7.234e+02 1.211e+02 7.524e+01
 4.248e+04 2.590e+04 5.059e+04 1.550e+04 1.139e+02 1.138e+02 1.231e+02
 1.827e+02 4.453e+03 4.426e+03 5.024e+04 6.517e+04 8.245e+01 1.115e+02
 2.078e+03 2.128e+03 3.036e+03 2.992e+04 1.478e+05 1.442e+05 5.892e+01
 5.852e+01 1.245e+02 1.337e+02 4.426e+04 1.602e+03 1.842e+03 9.344e+01
 8.602e+01 8.632e+01 8.770e+01 8.625e+01 8.645e+01 8.380e+01 2.922e+05
 5.751e+05 6.800e+05 6.794e+05 7.005e+05 1.830e+04 1.087e+03 1.071e+03
 1.898e+03 2.225e+02 1.398e+02 1.354e+02 5.331e+04 4.505e+03 7.016e+01
 7.046e+01 5.205e+01 6.444e+01 6.808e+01 3.932e+01 3.932e+01 3.951e+01
 3.930e+01 3.939e+01 3.928e+01 3.855e+01 3.179e+05 6.412e+05 7.026e+05
 7.048e+05 7.053e+05 2.789e+04 1.148e+02 1.121e+02 2.669e+02 3.287e+02
 2.712e+04 2.203e+04 1.375e+03 1.394e+03 1.393e+03 1.187e+02 1.641e+02
 1.647e+02 1.636e+02 1.637e+02 1.145e+02 1.163e+02 1.158e+02 2.091e+05
 4.852e+05 6.383e+05 6.686e+05 7.024e+05 4.449e+03 1.412e+03 3.326e+03
 3.337e+03 2.014e+02 3.505e+04 2.612e+04 1.373e+02 1.445e+02 5.371e+01
 7.051e+02 1.041e+03 1.165e+03 1.105e+03 1.186e+03 1.066e+03 4.717e+05
 6.901e+05 6.906e+05 7.007e+05 7.013e+05 1.240e+03 5.595e+04 3.118e+02
 3.316e+02 5.229e+02 1.058e+04 7.989e+03 1.617e+03 1.610e+03 1.290e+03
 1.606e+03 1.209e+03 2.757e+02 1.306e+02 1.304e+02 1.241e+02 1.263e+02
 1.248e+02 2.240e+05 4.841e+05 6.104e+05 4.035e+05 4.078e+05 3.191e+04
 4.709e+04 9.650e+01 1.200e+02 2.550e+02 3.468e+04 2.264e+04 8.884e+01
 8.887e+01 8.783e+01 8.891e+01 8.848e+01 8.641e+01 8.683e+01 1.458e+03
 1.135e+03 3.247e+03 4.994e+05 6.704e+05 7.026e+05 7.014e+05 7.010e+05
 3.467e+04 9.036e+04 4.041e+02 3.940e+02 2.514e+03 1.542e+04 1.494e+04
 3.257e+02 2.403e+03 2.687e+03 2.402e+03 1.580e+04 2.685e+03 7.349e+04
 5.320e+04 5.793e+02 4.268e+04 5.140e+02 2.350e+03 1.084e+03 2.898e+03
 1.959e+03 7.243e+04 4.572e+03 4.588e+03 1.432e+06 5.595e+05 1.212e+06
 7.149e+06 1.823e+06]
```

In [666]:

```
features = fit.transform(X)
```

In [667]:

```
# summarize selected features
print(features[0:16,:])
```

```
[[26. 91. 91. 91. 91. 91.  1.  2.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
```

```
 [26.  0.  0.  0.  0.  1.  1.  1.  1. 12.  7.  1.  4. 30. 30.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91.  0.  0.  0. 91.  1.  1.  1. 18. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 91. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91.  0.  1.  1.  1. 20.  6. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 16. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 14. 91. 91. 91. 91. 91.]
 [26. 91. 91. 91. 91. 91.  1.  1.  1. 21. 91. 91. 91. 91. 91.]]
```

In [668]:

```
print(features)
```

```
[[2.600e+01 9.100e+01 9.100e+01 ... 9.100e+01 9.100e+01 9.100e+01]
 [2.600e+01 9.100e+01 9.100e+01 ... 9.100e+01 9.100e+01 9.100e+01]
 [2.600e+01 9.100e+01 9.100e+01 ... 9.100e+01 9.100e+01 9.100e+01]
 ...
 [9.527e+03 9.991e+03 9.991e+03 ... 9.000e+00 1.000e+00 9.000e+00]
 [9.435e+03 9.991e+03 9.991e+03 ... 9.000e+00 1.000e+00 9.000e+00]
 [9.433e+03 9.991e+03 9.991e+03 ... 9.000e+00 1.000e+00 9.000e+00]]
```

In [669]:

```
from sklearn.decomposition import PCA
```

In [670]:

```
#PCA
pca = PCA(n_components=10)
fit = pca.fit(X)
```

In [671]:

```
# summarize components
print("Explained Variance: %s" % fit.explained_variance_ratio_)
```

```
Explained Variance: [8.702e-01 1.238e-01 2.943e-03 1.460e-03 6.078e-04 2.095e-04 1.165e-04
 9.773e-05 8.662e-05 8.168e-05]
```

In [672]:

```
print(fit.components_)
```

```
[[ 7.262e-02  9.367e-01  5.778e-02 ... -5.208e-04 -5.449e-04 -5.011e-04]
 [-2.878e-02 -3.412e-01  1.588e-01 ... -1.440e-03 -1.502e-03 -1.384e-03]
 [ 1.281e-03  1.005e-03 -2.448e-02 ...  2.140e-04  2.308e-04  2.090e-04]
 ...
 [-4.211e-04  7.076e-05  5.745e-03 ... -1.045e-03 -9.159e-05 -3.310e-04]
 [-1.045e-03 -9.990e-05  2.830e-02 ...  2.363e-03 -2.295e-04  2.649e-04]
 [ 9.384e-04  3.202e-05 -9.671e-03 ... -2.476e-04  1.528e-04  8.493e-05]]
```

In [673]:

```
from sklearn.ensemble import ExtraTreesClassifier
# feature extraction
model = ExtraTreesClassifier(n_estimators=10)
model.fit(X, Y)
print(model.feature_importances_)
```

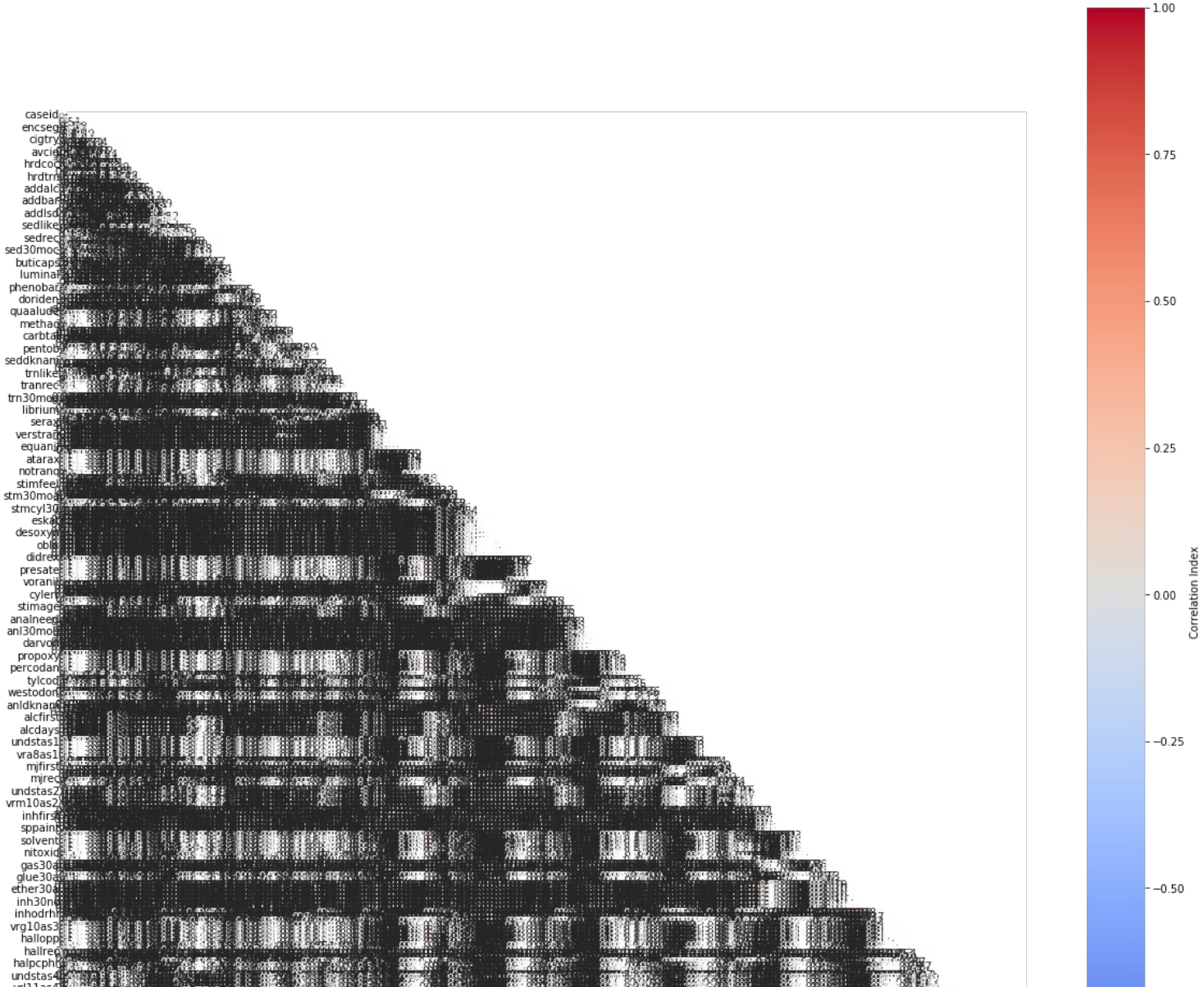
```
[1.851e-04 1.599e-04 1.980e-04 9.787e-05 1.619e-04 2.054e-04 1.438e-04
 9.779e-05 1.297e-04 1.405e-04 4.860e-05 1.898e-04 5.595e-05 1.409e-04
 8.707e-05 1.230e-04 9.956e-05 9.051e-05 1.699e-04 1.684e-04 1.664e-04
 3.780e-04 1.292e-04 9.461e-05 1.223e-04 1.279e-04 1.190e-04 1.547e-04
 2.267e-03 3.331e-04 1.204e-04 4.274e-03 2.364e-03 4.271e-05 3.830e-05
 3.089e-05 1.700e-04 2.816e-04 1.612e-04 8.798e-05 1.173e-04 1.300e-04
 1.196e-04 4.371e-04 7.458e-05 8.740e-05 5.710e-05 5.074e-05 1.283e-04
 6.032e-05 3.085e-05 7.927e-05 2.933e-04 5.250e-05 3.349e-05 2.114e-05
 4.883e-05 7.507e-05 3.916e-04 1.222e-04 8.761e-05 2.478e-05 6.436e-05
 8.815e-05 1.273e-04 8.353e-05 5.283e-05 2.957e-05 2.241e-05 1.260e-05
 1.553e-05 6.783e-05 6.523e-05 5.805e-05 1.122e-04 5.025e-05 7.703e-05
 2.569e-05 4.564e-05 5.047e-05 3.292e-05 2.711e-05 1.601e-05 1.794e-05
 9.034e-02 7.891e-06 9.047e-02 7.396e-05 4.227e-05 1.119e-04 6.172e-05
 2.020e-03 5.671e-05 4.857e-05 8.542e-05 2.062e-05 8.451e-04 1.753e-05
 3.559e-05 4.403e-05 5.453e-05 5.110e-05 1.492e-05 1.129e-05 2.562e-06
```

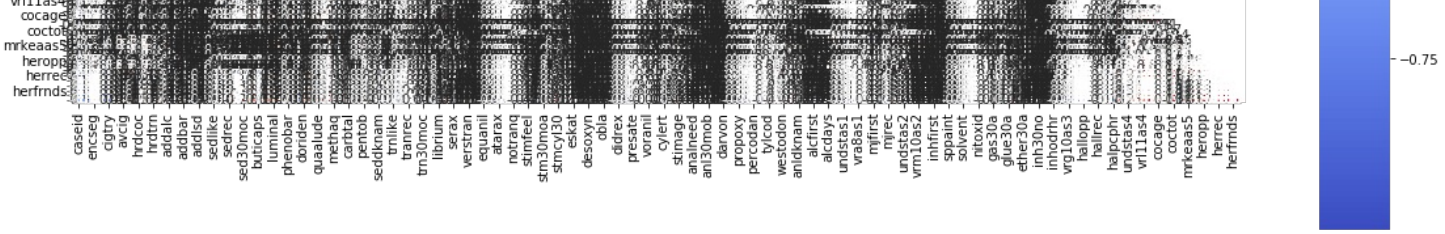
3.253e-05 4.129e-05 2.323e-05 4.159e-05 9.048e-02 1.182e-04 9.047e-02
5.435e-05 1.800e-05 2.519e-05 2.164e-05 2.936e-05 2.548e-05 4.341e-05
3.360e-05 3.960e-05 2.994e-05 2.220e-05 4.612e-05 8.110e-05 5.986e-04
6.006e-06 8.259e-04 4.259e-06 2.927e-05 3.325e-05 4.123e-04 7.575e-07
3.777e-05 1.890e-05 1.354e-05 2.570e-04 2.656e-03 4.694e-05 2.123e-03
4.872e-05 6.198e-05 3.535e-05 1.076e-05 7.630e-05 4.068e-04 5.073e-05
2.248e-05 3.540e-05 1.843e-05 8.507e-05 6.947e-04 7.380e-05 4.684e-05
2.065e-04 1.527e-05 9.048e-02 4.820e-07 3.456e-05 3.113e-05 7.679e-05
6.433e-05 5.584e-05 2.423e-05 1.599e-03 2.269e-05 6.976e-08 1.942e-05
4.869e-05 2.011e-05 5.201e-05 1.937e-04 3.598e-05 4.821e-05 1.289e-05
8.094e-06 7.575e-06 2.998e-05 2.966e-05 9.323e-02 9.065e-02 1.604e-05
1.531e-05 2.404e-05 3.982e-05 2.625e-05 2.149e-05 4.609e-05 1.655e-05
4.187e-05 2.158e-05 2.023e-05 7.950e-05 2.612e-04 3.757e-05 4.616e-05
1.913e-04 6.768e-05 4.809e-05 1.814e-01 4.935e-05 4.684e-05 1.130e-04
7.367e-05 6.763e-05 4.701e-05 5.747e-04 7.843e-05 1.257e-04 1.143e-04
8.480e-05 4.741e-05 1.755e-04 3.364e-05 1.165e-04 8.474e-05 3.717e-04
2.913e-03 4.974e-03 3.154e-03 1.032e-03 4.176e-04 7.478e-04 8.737e-05
2.779e-03 3.541e-03 2.182e-04 4.356e-03 1.420e-02 6.230e-03 2.792e-03
9.523e-02 3.089e-03]

In [674]:

```
import matplotlib.pyplot as plt

import seaborn as sns
corr = df.corr()
matrix = np.triu(corr) # for triangular matrix
plt.figure(figsize=(20,20))
# parse corr variable intro triangular matrix
sns.heatmap(df.corr(method='pearson'),
            annot=True, linewidths=.9,
            cmap="coolwarm", mask=matrix,
            square = True,
            cbar_kws={'label': 'Correlation Index'})
plt.show()
```





In [675]:

```
print(df.shape)
```

(33785, 234)

In [676]:

```
# Remove all columns between column index 234 to 603
df.drop(df.iloc[:, 6:35], inplace = True, axis = 1)
```

/opt/conda/lib/python3.7/site-packages/pandas/core/frame.py:3997: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
errors=errors,

In [677]:

```
df.shape
```

Out[677]:

(33785, 205)

In [683]:

```
#Calculate correlation between all columns and remove highly correlated columns
import numpy as np

# Create correlation matrix
corr_matrix = df.corr().abs()

# Select upper triangle of correlation matrix
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool))

# Find features with correlation greater than 0.8
to_drop = [column for column in upper.columns if any(upper[column] > 0.65)]

# Drop features
df.drop(to_drop, axis=1, inplace=True)
```

In [684]:

```
df.shape
```

Out[684]:

(33785, 21)

In [685]:

```
df.columns
```

Out[685]:

```
Index(['caseid', 'encpsu', 'buticaps', 'amytal', 'eskabarb', 'doriden',
       'sopor', 'nembutal', 'seconal', 'trn30moc', 'valium', 'tranxene',
       'stimrec', 'stm30mob', 'analrec', 'mjage', 'inhopp', 'amlnt30a',
       'inhodrus', 'hal30use', 'cocus30a'],
      dtype='object')
```

In [ ]: