00a_S3_Setup_Final

April 14, 2023

1 ADS-508-01-SP23 Team 8: Final Project

2 Setup Database and Athena Tables

Much of the code is modified from `Fregly, C., & Barth, A. (2021). Data science on AWS: Implementing end-to-end, continuous AI and machine learning pipelines. O'Reilly.`

2.1 Install missing dependencies

PyAthena is a Python DB API 2.0 (PEP 249) compliant client for Amazon Athena.

[2]:
```
!pip install --disable-pip-version-check -q PyAthena==2.1.0
```

```
WARNING: The directory '/root/.cache/pip' or its parent directory is not
owned or is not writable by the current user. The cache has been disabled. Check
the permissions and owner of that directory. If executing pip with sudo, you
should use sudo's -H flag.
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
```

2.2 Globally import libraries

[3]:
```
import boto3
from botocore.client import ClientError
import sagemaker
import pandas as pd
from pyathena import connect
from IPython.core.display import display, HTML

%matplotlib inline
```

2.3 Instantiate AWS SageMaker session

[4]:
```
session = boto3.session.Session()
region = session.region_name
sagemaker_session = sagemaker.Session()
def_bucket = sagemaker_session.default_bucket()
bucket = 'sagemaker-us-east-ads508-sp23-t8'

s3 = boto3.Session().client(service_name="s3",
                            region_name=region)

role = sagemaker.get_execution_role()
account_id = boto3.client("sts").get_caller_identity().get("Account")

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

[5]:
```
setup_s3_bucket_passed = False
ingest_create_athena_db_passed = False
ingest_create_athena_table_tsv_passed = False
```

[6]:
```
print(f"Default bucket: {def_bucket}")
print(f"Public T8 bucket: {bucket}")
```

```
Default bucket: sagemaker-us-east-1-657724983756
Public T8 bucket: sagemaker-us-east-ads508-sp23-t8
```

2.4 Verify S3 Bucket Creation

[7]:
```
%%bash

aws s3 ls s3://${bucket}/
```

```
2023-03-16 17:05:02 aws-athena-query-results-657724983756-us-east-1
2023-03-02 16:56:48 sagemaker-studio-657724983756-5nh7ydsouq7
2023-03-02 17:25:41 sagemaker-studio-657724983756-7yc8bp8xk0b
2023-03-02 17:01:51 sagemaker-us-east-1-657724983756
2023-03-17 05:19:31 sagemaker-us-east-ads508-sp23-t8
```

[8]:
```
response = None

try:
    response = s3.head_bucket(Bucket=bucket)
    print(response)
    setup_s3_bucket_passed = True
except ClientError as e:
    print(f"[ERROR] Cannot find bucket {bucket} in {response} due to {e}.")
```

```
{'ResponseMetadata': {'RequestId': 'SVMXN6X37MEBHKWD', 'HostId':
```

```
'RXgRFVHFnPcOr1MLGT3ZVZzuS5ZyT09xU8/1usp6nCDCivyey/7QG2Q/A5Z9fyNsL0et/C5+S2g=',
'HTTPStatusCode': 200, 'HTTPHeaders': {'x-amz-id-2':
'RXgRFVHFnPcOr1MLGT3ZVZzuS5ZyT09xU8/1usp6nCDCivyey/7QG2Q/A5Z9fyNsL0et/C5+S2g=',
'x-amz-request-id': 'SVMXN6X37MEBHKWD', 'date': 'Thu, 13 Apr 2023 16:48:46 GMT',
'x-amz-bucket-region': 'us-east-1', 'x-amz-access-point-alias': 'false',
'content-type': 'application/xml', 'server': 'AmazonS3'}, 'RetryAttempts': 0}}
```

[9]:
```
%store setup_s3_bucket_passed
```

```
Stored 'setup_s3_bucket_passed' (bool)
```

3 Set S3 Source Location (Public S3 Bucket)

[10]:
```
s3_public_path_tsv = f"s3://{bucket}"
```

[11]:
```
%store s3_public_path_tsv
```

```
Stored 's3_public_path_tsv' (str)
```

4 Set S3 Destination Location (Our Private S3 Bucket)

[12]:
```
s3_private_path_tsv = f"s3://{def_bucket}/team_8_data"
print(s3_private_path_tsv)
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data
```

[13]:
```
%store s3_private_path_tsv
```

```
Stored 's3_private_path_tsv' (str)
```

5 Copy Data From the Public S3 Bucket to our Private S3 Bucket in this Account

[14]:
```
!aws s3 cp --recursive $s3_public_path_tsv/ $s3_private_path_tsv/
```

```
copy: s3://sagemaker-us-east-
ads508-sp23-t8/raw_data/census_block/census_block_loc.csv to s3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/census_block/census_block_loc.csv
copy: s3://sagemaker-us-east-
ads508-sp23-t8/raw_data/census/nyc_census_tracts.csv to s3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/census/nyc_census_tracts.csv
copy: s3://sagemaker-us-east-
ads508-sp23-t8/raw_data/grad_outcomes/2005-2010_Graduation_Outcomes_-
_School_Level.tsv to s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/
grad_outcomes/2005-2010_Graduation_Outcomes_-_School_Level.tsv
copy: s3://sagemaker-us-east-
ads508-sp23-t8/raw_data/hs_dir/2014_-_2015_DOE_High_School_Directory.tsv to
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/hs_dir/2014_-_2015_DO
```

```
E_High_School_Directory.tsv
copy: s3://sagemaker-us-east-ads508-sp23-t8/raw_data/jobs/NYC_Jobs.tsv to
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/jobs/NYC_Jobs.tsv
copy: s3://sagemaker-us-east-ads508-sp23-t8/raw_data/evictions/Evictions.tsv to
s3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/evictions/Evictions.tsv
copy: s3://sagemaker-us-east-
ads508-sp23-t8/raw_data/crime/NYPD_Complaint_Data_Historic (1).tsv.gz to
s3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/crime/NYPD_Complaint_Data_Historic
(1).tsv.gz
```

6 List Files in our Private S3 Bucket in this Account

[15]:
```
print(s3_private_path_tsv)
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data
```

[16]:
```
!aws s3 ls $s3_private_path_tsv/
```

```
                           PRE raw_data/
```

[17]:
```
from IPython.core.display import display, HTML

display(
    HTML(
        f'<b>Review <a target="blank" href="https://s3.console.aws.amazon.com/
 ↪s3/buckets/sagemaker-{region}-{account_id}/amazon-reviews-pds/?
 ↪region={region}&tab=overview">S3 Bucket</a></b>'
    )
)
```

```
<IPython.core.display.HTML object>
```

6.1 Create Athena Database and Tables

[18]:
```
database_name = "ads508_t8"
```

[19]:
```
# Set S3 staging directory -- this is a temporary directory used for Athena␣
 ↪queries
s3_staging_dir = f"s3://{def_bucket}/team_8_data/athena/staging"
print(s3_staging_dir)
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data/athena/staging
```

[20]:
```
conn = connect(region_name=region,
               s3_staging_dir=s3_staging_dir)
```

[21]:
```
create_db_stmnt = f"CREATE DATABASE IF NOT EXISTS {database_name}"
print(create_db_stmnt)
```

```
CREATE DATABASE IF NOT EXISTS ads508_t8
```

[22]:
```
pd.read_sql(create_db_stmnt,
            conn)
```

[22]:
```
Empty DataFrame
Columns: []
Index: []
```

6.1.1 Verify The Database Has Been Created Succesfully

[23]:
```
show_db_stmnt = "SHOW DATABASES"

df_show = pd.read_sql(show_db_stmnt,
                      conn)

df_show.head(17)
```

[23]:

	database_name
0	ads508_t8
1	default
2	dsoaws
3	sagemaker_featurestore

[24]:
```
if database_name in df_show.values:
    ingest_create_athena_db_passed = True
```

[25]:
```
%store ingest_create_athena_db_passed
```

```
Stored 'ingest_create_athena_db_passed' (bool)
```

6.2 Define custom function to create tables in existing database

[26]:
```
def create_athena_tbl_tsv(conn=None,
                          db=None,
                          tbl_name=None,
                          fields='',
                          s3_path=None,
                          delim=',',
                          ret='',
                          comp='',
                          skip=''):
    # Set Athena parameters

    # SQL statement to execute
```

```
    drop_tsv_tbl_stmnt = f"""DROP TABLE IF EXISTS {db}.{tbl_name}"""

    create_tsv_tbl_stmnt = f"""
        CREATE EXTERNAL TABLE IF NOT EXISTS {db}.{tbl_name}({fields})
        ROW FORMAT DELIMITED
            FIELDS
                TERMINATED BY '{delim}'
            LINES
                TERMINATED BY '{ret}\\n'
        LOCATION '{s3_path}'
        TBLPROPERTIES ({comp}{skip})
        """

    print(f'Create table statement:\n{create_tsv_tbl_stmnt}')

    pd.read_sql(drop_tsv_tbl_stmnt,
                conn)

    pd.read_sql(create_tsv_tbl_stmnt,
                conn)

    # Verify The Table Has Been Created Succesfully
    show_tsv_tbl_stmnt = f"SHOW TABLES IN {db}"

    df_show = pd.read_sql(show_tsv_tbl_stmnt,
                          conn)
    display(df_show.head(17))

    if tbl_name in df_show.values:
        ingest_create_athena_table_tsv_passed = True

    print(f'\nDataframe contains records:␣
↪{ingest_create_athena_table_tsv_passed}')
```

6.3 Create Athena Table from Local TSV File - 2005-2010_Graduation_Outcomes_-_School_Level.tsv

[27]:
```
grd_tsv_tbl_name = 'grad_outcomes'
grd_tsv_field_list = """
demographic string,
dbn string,
school_name string,
cohort string,
total_cohort string,
total_grads_n string,
total_grads_perc_cohort string,
total_regents_n string,
```

```
total_regents_perc_cohort string,
total_regents_perc_grads string,
advanced_regents_n string,
advanced_regents_perc_cohort string,
advanced_regents_perc_grads string,
regents_wo_advanced_n string,
regents_wo_advanced_perc_cohort string,
regents_wo_advanced_perc_grads string,
local_n string,
local_perc_cohort string,
local_perc_grads string,
still_enrolled_n string,
still_enrolled_perc_cohort string,
dropped_out_n string,
dropped_out_perc_cohort string
"""
grd_tsv_s3_raw_data_path = f"s3://{def_bucket}/team_8_data/raw_data/
 ↪grad_outcomes"
print(grd_tsv_s3_raw_data_path)

create_athena_tbl_tsv(conn=conn,
                      db=database_name,
                      tbl_name=grd_tsv_tbl_name,
                      fields=grd_tsv_field_list,
                      s3_path=grd_tsv_s3_raw_data_path,
                      delim='\\t',
                      comp='',
                      skip="'skip.header.line.count'='1'")
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/grad_outcomes
Create table statement:

        CREATE EXTERNAL TABLE IF NOT EXISTS ads508_t8.grad_outcomes(
demographic string,
dbn string,
school_name string,
cohort string,
total_cohort string,
total_grads_n string,
total_grads_perc_cohort string,
total_regents_n string,
total_regents_perc_cohort string,
total_regents_perc_grads string,
advanced_regents_n string,
advanced_regents_perc_cohort string,
advanced_regents_perc_grads string,
regents_wo_advanced_n string,
regents_wo_advanced_perc_cohort string,
```

```
regents_wo_advanced_perc_grads string,
local_n string,
local_perc_cohort string,
local_perc_grads string,
still_enrolled_n string,
still_enrolled_perc_cohort string,
dropped_out_n string,
dropped_out_perc_cohort string
)
        ROW FORMAT DELIMITED
            FIELDS
                TERMINATED BY '\t'
            LINES
                TERMINATED BY '\n'
        LOCATION 's3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/grad_outcomes'
        TBLPROPERTIES ('skip.header.line.count'='1')


        tab_name
0         census
1   census_block
2          crime
3      crime_pqt
4      evictions
5  grad_outcomes
6        hs_info
7           jobs


Dataframe contains records: True
```

6.3.1 Run A Sample Query

[28]:
```
grd_dbn_id01 = "01M448"

grd_select_dbn_stmnt = f"""
SELECT * FROM {database_name}.{grd_tsv_tbl_name}
WHERE dbn = '{grd_dbn_id01}'
LIMIT 17
"""

print(grd_select_dbn_stmnt)

grd_df01_s01 = pd.read_sql(grd_select_dbn_stmnt,
                           conn)

grd_df01_s01.head(17)
```

```
SELECT * FROM ads508_t8.grad_outcomes
WHERE dbn = '01M448'
LIMIT 17
```

[28]:

	demographic	dbn	school_name	\
0	Total Cohort	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
1	Total Cohort	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
2	Total Cohort	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
3	Total Cohort	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
4	Total Cohort	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
5	Total Cohort	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
6	Total Cohort	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
7	English Language Learners	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
8	English Language Learners	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
9	English Language Learners	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
10	English Language Learners	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
11	English Language Learners	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
12	English Language Learners	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
13	English Language Learners	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
14	English Proficient Students	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
15	English Proficient Students	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	
16	English Proficient Students	01M448	UNIVERSITY NEIGHBORHOOD HIGH SCHOOL	

	cohort	total_cohort	total_grads_n	total_grads_perc_cohort	\
0	2001	64	46	71.900000000000006	
1	2002	52	33	63.5	
2	2003	87	67	77	
3	2004	112	75	67	
4	2005	121	64	52.9	
5	2006	124	53	42.7	
6	2006 Aug	124	60	48.4	
7	2002	1	s		
8	2001	5	s		
9	2003	1	s		
10	2004	9	s		
11	2005	7	s		
12	2006	3	s		
13	2006 Aug	3	s		
14	2001	59	44	74.599999999999994	
15	2002	51	32	62.7	
16	2003	86	67	77.900000000000006	

	total_regents_n	total_regents_perc_cohort	total_regents_perc_grads	…	\
0	32	50	69.599999999999994	…	
1	19	36.5	57.6	…	

```
2                39                    44.8                    58.2  …
3                36                    32.1                      48  …
4                35                    28.9                    54.7  …
5                42                    33.9                    79.2  …
6                42                    33.9                      70  …
7                 s                                                    …
8                 s                                                    …
9                 s                                                    …
10                s                                                    …
11                s                                                    …
12                s                                                    …
13                s                                                    …
14               31                    52.5                    70.5  …
15               19       37.299999999999997                    59.4  …
16               39                    45.3                    58.2  …

   regents_wo_advanced_n regents_wo_advanced_perc_cohort  \
0                     25                            39.1
1                     11                            21.2
2                     28          32.200000000000003
3                     30                            26.8
4                     31                            25.6
5                     34                            27.4
6                     34                            27.4
7                      s
8                      s
9                      s
10                     s
11                     s
12                     s
13                     s
14                    24          40.700000000000003
15                    11                            21.6
16                    28                            32.6

   regents_wo_advanced_perc_grads local_n   local_perc_cohort  \
0                            54.3      14                21.9
1               33.299999999999997      14                26.9
2                            41.8      28  32.200000000000003
3                              40      39  34.799999999999997
4                            48.4      29                  24
5                            64.2      11                 8.9
6                            56.7      18                14.5
7                                       s
8                                       s
9                                       s
10                                      s
```

```
11                                    s
12                                    s
13                                    s
14                          54.5      13                  22
15                          34.4      13                25.5
16                          41.8      28                32.6

   local_perc_grads still_enrolled_n still_enrolled_perc_cohort dropped_out_n  \
0              30.4               10                       15.6             6
1              42.4               16                       30.8             1
2              41.8                9                       10.3            11
3                52               33                       29.5             4
4              45.3               41                       33.9            11
5              20.8               46                       37.1            20
6                30               39                       31.5            20
7                                    s                                        s
8                                    s                                        s
9                                    s                                        s
10                                   s                                        s
11                                   s                                        s
12                                   s                                        s
13                                   s                                        s
14             29.5                8                       13.6             5
15             40.6               16                       31.4             1
16             41.8                9                       10.5            10

   dropped_out_perc_cohort
0                      9.4
1                      1.9
2                     12.6
3                      3.6
4                      9.1
5       16.100000000000001
6       16.100000000000001
7
8
9
10
11
12
13
14                     8.5
15                       2
16                    11.6

[17 rows x 23 columns]
```

[29]:
```
if not grd_df01_s01.empty:
    print("[OK]")
else:
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN REGISTERED WITH ATHENA. LOOK IN␣
 ↪PREVIOUS CELLS TO FIND THE ISSUE.")
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

6.4 Create Athena Table from Local TSV File - 2014_-_2015_DOE_High_School_Directory.tsv

[30]:
```
hsi_tsv_tbl_name = 'hs_info'
hsi_tsv_field_list = """
dbn string,
school_name string,
borough string,
building_code string,
phone_number string,
fax_number string,
grade_span_min string,
grade_span_max string,
expgrade_span_min string,
expgrade_span_max string,
bus string,
subway string,
primary_address_line_1 string,
city string,
state_code string,
postcode string,
website string,
total_students string,
campus_name string,
school_type string,
overview_paragraph string,
program_highlights string,
language_classes string,
advancedplacement_courses string,
online_ap_courses string,
online_language_courses string,
extracurricular_activities string,
psal_sports_boys string,
psal_sports_girls string,
psal_sports_coed string,
school_sports string,
partner_cbo string,
```

```
partner_hospital string,
partner_highered string,
partner_cultural string,
partner_nonprofit string,
partner_corporate string,
partner_financial string,
partner_other string,
addtl_info1 string,
addtl_info2 string,
start_time string,
end_time string,
se_services string,
ell_programs string,
school_accessibility_description string,
number_programs string,
priority01 string,
priority02 string,
priority03 string,
priority04 string,
priority05 string,
priority06 string,
priority07 string,
priority08 string,
priority09 string,
priority10 string,
location_1 string,
community_board string,
council_district string,
census_tract string,
bin string,
bbl string,
nta string
"""
hsi_tsv_s3_raw_data_path = f"s3://{def_bucket}/team_8_data/raw_data/hs_dir"
print(hsi_tsv_s3_raw_data_path)

create_athena_tbl_tsv(conn=conn,
                      db=database_name,
                      tbl_name=hsi_tsv_tbl_name,
                      fields=hsi_tsv_field_list,
                      s3_path=hsi_tsv_s3_raw_data_path,
                      delim='\\t',
                      comp='',
                      skip="'skip.header.line.count'='1'")
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/hs_dir
Create table statement:
```

```
        CREATE EXTERNAL TABLE IF NOT EXISTS ads508_t8.hs_info(
dbn string,
school_name string,
borough string,
building_code string,
phone_number string,
fax_number string,
grade_span_min string,
grade_span_max string,
expgrade_span_min string,
expgrade_span_max string,
bus string,
subway string,
primary_address_line_1 string,
city string,
state_code string,
postcode string,
website string,
total_students string,
campus_name string,
school_type string,
overview_paragraph string,
program_highlights string,
language_classes string,
advancedplacement_courses string,
online_ap_courses string,
online_language_courses string,
extracurricular_activities string,
psal_sports_boys string,
psal_sports_girls string,
psal_sports_coed string,
school_sports string,
partner_cbo string,
partner_hospital string,
partner_highered string,
partner_cultural string,
partner_nonprofit string,
partner_corporate string,
partner_financial string,
partner_other string,
addtl_info1 string,
addtl_info2 string,
start_time string,
end_time string,
se_services string,
ell_programs string,
school_accessibility_description string,
number_programs string,
```

```
priority01 string,
priority02 string,
priority03 string,
priority04 string,
priority05 string,
priority06 string,
priority07 string,
priority08 string,
priority09 string,
priority10 string,
location_1 string,
community_board string,
council_district string,
census_tract string,
bin string,
bbl string,
nta string
)
        ROW FORMAT DELIMITED
            FIELDS
                TERMINATED BY '\t'
            LINES
                TERMINATED BY '\n'
        LOCATION 's3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/hs_dir'
        TBLPROPERTIES ('skip.header.line.count'='1')


        tab_name
0         census
1   census_block
2          crime
3      crime_pqt
4      evictions
5  grad_outcomes
6        hs_info
7           jobs


Dataframe contains records: True
```

6.4.1 Run A Sample Query

[31]:
```
hsi_dbn_id01 = "01M448"

hsi_select_dbn_stmnt = f"""
SELECT * FROM {database_name}.{hsi_tsv_tbl_name}
WHERE dbn = '{hsi_dbn_id01}'
```

```
LIMIT 17
"""

print(hsi_select_dbn_stmnt)

hsi_df01_s01 = pd.read_sql(hsi_select_dbn_stmnt,
                           conn)

hsi_df01_s01.head(17)
```

```
SELECT * FROM ads508_t8.hs_info
WHERE dbn = '01M448'
LIMIT 17
```

[31]:
```
       dbn                          school_name    borough building_code  \
0  01M448  University Neighborhood High School  Manhattan          M446

   phone_number    fax_number grade_span_min grade_span_max expgrade_span_min  \
0  212-962-4341  212-267-5611              9             12

  expgrade_span_max  … priority08 priority09 priority10          location_1  \
0                    …                                  "200 Monroe Street

  community_board council_district census_tract   bin   bbl   nta
0            None             None         None  None  None  None

[1 rows x 64 columns]
```

[32]:
```
if not hsi_df01_s01.empty:
    print("[OK]")
else:
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN REGISTERED WITH ATHENA. LOOK IN␣
 ↪PREVIOUS CELLS TO FIND THE ISSUE.")
    print("+++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

6.5 Create Athena Table from Local CSV File - `nyc_census_tracts.csv`

[33]:
```
cen_tsv_tbl_name = 'census'
cen_tsv_field_list = """
censustract string,
county string,
borough string,
totalpop int,
```

```
men int,
women int,
hispanic double,
white double,
black double,
native double,
asian double,
citizen int,
income int,
incomeerr int,
incomepercap int,
incomepercaperr int,
poverty double,
childpoverty double,
professional double,
service double,
office double,
construction double,
production double,
drive double,
carpool double,
transit double,
walk double,
othertransp double,
workathome double,
meancommute double,
employed int,
privatework double,
publicwork double,
selfemployed double,
familywork double,
unemployment double
"""
cen_tsv_s3_raw_data_path = f"s3://{def_bucket}/team_8_data/raw_data/census"
print(cen_tsv_s3_raw_data_path)

create_athena_tbl_tsv(conn=conn,
                      db=database_name,
                      tbl_name=cen_tsv_tbl_name,
                      fields=cen_tsv_field_list,
                      s3_path=cen_tsv_s3_raw_data_path,
                      comp='',
                      skip="'skip.header.line.count'='1'")
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/census
Create table statement:

        CREATE EXTERNAL TABLE IF NOT EXISTS ads508_t8.census(
```

```
censustract string,
county string,
borough string,
totalpop int,
men int,
women int,
hispanic double,
white double,
black double,
native double,
asian double,
citizen int,
income int,
incomeerr int,
incomepercap int,
incomepercaperr int,
poverty double,
childpoverty double,
professional double,
service double,
office double,
construction double,
production double,
drive double,
carpool double,
transit double,
walk double,
othertransp double,
workathome double,
meancommute double,
employed int,
privatework double,
publicwork double,
selfemployed double,
familywork double,
unemployment double
)
        ROW FORMAT DELIMITED
            FIELDS
                TERMINATED BY ','
            LINES
                TERMINATED BY '\n'
        LOCATION 's3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/census'
        TBLPROPERTIES ('skip.header.line.count'='1')


        tab_name
```

```
0         census
1   census_block
2          crime
3      crime_pqt
4      evictions
5  grad_outcomes
6        hs_info
7           jobs
```

```
Dataframe contains records: True
```

6.5.1 Run A Sample Query

[34]:
```
cen_bourough_id01 = "Bronx"

cen_select_dbn_stmnt = f"""
SELECT * FROM {database_name}.{cen_tsv_tbl_name}
WHERE borough = '{cen_bourough_id01}'
LIMIT 17
"""

print(cen_select_dbn_stmnt)

cen_df01_s01 = pd.read_sql(cen_select_dbn_stmnt,
                           conn)

cen_df01_s01.head(17)
```

```
SELECT * FROM ads508_t8.census
WHERE borough = 'Bronx'
LIMIT 17
```

[34]:

	censustract	county	borough	totalpop	men	women	hispanic	white	black	\
0	36005000100	Bronx	Bronx	7703	7133	570	29.9	6.1	60.9	
1	36005000200	Bronx	Bronx	5403	2659	2744	75.8	2.3	16.0	
2	36005000400	Bronx	Bronx	5915	2896	3019	62.7	3.6	30.7	
3	36005001600	Bronx	Bronx	5879	2558	3321	65.1	1.6	32.4	
4	36005001900	Bronx	Bronx	2591	1206	1385	55.4	9.0	29.0	
5	36005002000	Bronx	Bronx	8516	3301	5215	61.1	1.6	31.1	
6	36005002300	Bronx	Bronx	4774	2130	2644	62.3	0.2	36.5	
7	36005002400	Bronx	Bronx	150	109	41	0.0	52.0	48.0	
8	36005002500	Bronx	Bronx	5355	2338	3017	76.5	1.5	18.9	
9	36005002701	Bronx	Bronx	3016	1375	1641	68.0	0.0	31.2	
10	36005002702	Bronx	Bronx	4778	2427	2351	71.3	1.6	26.2	
11	36005002800	Bronx	Bronx	5299	2292	3007	23.0	0.2	71.4	

```
12  36005003100  Bronx   Bronx      1466   769    697      72.3    0.6   24.6
13  36005003300  Bronx   Bronx      3912  1824   2088      65.6    1.0   30.6
14  36005003500  Bronx   Bronx      3948  1921   2027      73.5    0.7   25.9
15  36005003700  Bronx   Bronx       246   128    118      57.7   24.4   17.9
16  36005003800  Bronx   Bronx      1193   542    651      53.1    1.8   42.7

    native  …  walk  othertransp  workathome  meancommute  employed  \
0      0.2  …   NaN          NaN         NaN          NaN         0
1      0.0  …   2.9          0.0         0.0         43.0      2308
2      0.0  …   1.4          0.5         2.1         45.0      2675
3      0.0  …   8.6          1.6         1.7         38.8      2120
4      0.0  …   3.0          2.4         6.2         45.4      1083
5      0.3  …   4.3          1.0         0.0         46.0      2508
6      1.0  …  14.0          1.5         4.1         42.7      1191
7      0.0  …   0.0          0.0         0.0          NaN       113
8      0.0  …  17.7          1.8         2.7         35.5      1691
9      0.0  …  18.0          0.0         1.6         42.8      1102
10     0.0  …   7.1          0.7         0.5         44.0      1559
11     0.0  …   2.0          0.6         2.7         47.3      2394
12     0.0  …  14.6          3.5         0.0         40.1       722
13     1.7  …  13.5          0.8         1.8         42.5      1113
14     0.0  …   7.4          1.0         4.7         41.6      1360
15     0.0  …  12.5          0.0         0.0         33.0        96
16     0.0  …   6.9          1.3         1.1         41.3       476

    privatework  publicwork  selfemployed  familywork  unemployment
0           NaN         NaN           NaN         NaN           NaN
1          80.8        16.2           2.9         0.0           7.7
2          71.7        25.3           2.5         0.6           9.5
3          75.0        21.3           3.8         0.0           8.7
4          76.8        15.5           7.7         0.0          19.2
5          71.0        21.3           7.7         0.0          17.2
6          74.2        16.1           9.7         0.0          18.9
7          62.8        37.2           0.0         0.0           0.0
8          85.1         8.3           6.1         0.5           9.4
9          86.9         8.5           4.5         0.0          15.2
10         75.0        14.0          11.0         0.0          10.6
11         61.9        37.4           0.6         0.0          12.8
12         79.2        10.2          10.5         0.0           6.6
13         77.2        16.9           5.9         0.0          18.5
14         83.2        13.4           3.4         0.0          11.8
15        100.0         0.0           0.0         0.0          11.1
16         71.8        27.1           1.1         0.0          13.1

[17 rows x 36 columns]
```

[35]:
```
if not cen_df01_s01.empty:
    print("[OK]")
else:
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN REGISTERED WITH ATHENA. LOOK IN␣
 ↪PREVIOUS CELLS TO FIND THE ISSUE.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

6.6 Create Athena Table from Local TSV File - NYPD_Complaint_Data_Historic (1).csv

[36]:
```
cri_tsv_tbl_name = 'crime'
cri_tsv_field_list = """
cmplnt_num string,
cmplnt_fr_dt string,
cmplnt_fr_tm string,
cmplnt_to_dt string,
cmplnt_to_tm string,
addr_pct_cd string,
rpt_dt string,
ky_cd string,
ofns_desc string,
pd_cd string,
pd_desc string,
crm_atpt_cptd_cd string,
law_cat_cd string,
borough string,
loc_of_occur_desc string,
prem_typ_desc string,
juris_desc string,
jurisdiction_code string,
parks_nm string,
hadevelopt string,
housing_psa string,
x_coord_cd string,
y_coord_cd string,
susp_age_group string,
susp_race string,
susp_sex string,
transit_district string,
latitude string,
longitude string,
lat_lon string,
patrol_boro string,
station_name string,
```

```
vic_age_group string,
vic_race string,
vic_sex string
"""
cri_tsv_s3_raw_data_path = f"s3://{def_bucket}/team_8_data/raw_data/crime"
print(cri_tsv_s3_raw_data_path)

create_athena_tbl_tsv(conn=conn,
                      db=database_name,
                      tbl_name=cri_tsv_tbl_name,
                      fields=cri_tsv_field_list,
                      s3_path=cri_tsv_s3_raw_data_path,
                      delim='\\t',
                      comp="'compressionType'='gzip', ",
                      skip="'skip.header.line.count'='1'")
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/crime
Create table statement:

        CREATE EXTERNAL TABLE IF NOT EXISTS ads508_t8.crime(
cmplnt_num string,
cmplnt_fr_dt string,
cmplnt_fr_tm string,
cmplnt_to_dt string,
cmplnt_to_tm string,
addr_pct_cd string,
rpt_dt string,
ky_cd string,
ofns_desc string,
pd_cd string,
pd_desc string,
crm_atpt_cptd_cd string,
law_cat_cd string,
borough string,
loc_of_occur_desc string,
prem_typ_desc string,
juris_desc string,
jurisdiction_code string,
parks_nm string,
hadevelopt string,
housing_psa string,
x_coord_cd string,
y_coord_cd string,
susp_age_group string,
susp_race string,
susp_sex string,
transit_district string,
latitude string,
```

```
longitude string,
lat_lon string,
patrol_boro string,
station_name string,
vic_age_group string,
vic_race string,
vic_sex string
)
        ROW FORMAT DELIMITED
            FIELDS
                TERMINATED BY '\t'
            LINES
                TERMINATED BY '\n'
        LOCATION 's3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/crime'
        TBLPROPERTIES ('compressionType'='gzip', 'skip.header.line.count'='1')


        tab_name
0         census
1   census_block
2          crime
3      crime_pqt
4      evictions
5  grad_outcomes
6        hs_info
7           jobs


Dataframe contains records: True
```

6.6.1 Run A Sample Query

[37]:
```
cri_law_cat_cd01 = "misdemeanor"
cri_borough01 = "bronx"

cri_select_dbn_stmnt01 = f"""
SELECT * FROM {database_name}.{cri_tsv_tbl_name}
WHERE LOWER(law_cat_cd) = '{cri_law_cat_cd01}'
    AND LOWER(borough) = '{cri_borough01}'
LIMIT 17
"""

print(cri_select_dbn_stmnt01)

cri_df01_s01 = pd.read_sql(cri_select_dbn_stmnt01,
                           conn)
```

```
cri_df01_s01.head(17)
```

```
SELECT * FROM ads508_t8.crime
WHERE LOWER(law_cat_cd) = 'misdemeanor'
    AND LOWER(borough) = 'bronx'
LIMIT 17
```

[37]:

	cmplnt_num	cmplnt_fr_dt	cmplnt_fr_tm	cmplnt_to_dt	cmplnt_to_tm	addr_pct_cd	\
0	629632833	02/06/2018	23:15:00			52	
1	377132404	08/04/2018	22:15:00			44	
2	584276892	02/11/2018	17:30:00	02/12/2018	06:00:00	41	
3	599398393	05/23/2018	23:30:00	05/24/2018	02:00:00	47	
4	955332763	02/23/2018	13:55:00			43	
5	412087799	05/07/2018	15:00:00	05/19/2018	18:00:00	47	
6	692539256	08/30/2018	17:01:00	08/31/2018	17:41:00	52	
7	763109503	05/03/2018	16:55:00			44	
8	472961714	08/01/2018	11:30:00	08/01/2018	11:33:00	49	
9	249426294	06/14/2018	14:50:00	06/14/2018	14:55:00	49	
10	828720525	04/06/2018	17:25:00	04/06/2018	17:25:00	40	
11	783013551	04/28/2018	09:40:00			40	
12	847047309	10/06/2018	08:01:00			45	
13	746565217	07/14/2018	22:00:00			44	
14	700974838	12/09/2018	17:00:00	12/09/2018	18:30:00	43	
15	140793872	10/10/2018	00:20:00	10/10/2018	00:35:00	40	
16	544404081	05/10/2018	11:00:00	05/10/2018	11:50:00	46	

	rpt_dt	ky_cd	ofns_desc	pd_cd	…	susp_sex	\
0	02/07/2018	341	PETIT LARCENY	333	…	F	
1	08/04/2018	344	ASSAULT 3 & RELATED OFFENSES	101	…	M	
2	02/12/2018	351	CRIMINAL MISCHIEF & RELATED OF	254	…	U	
3	05/24/2018	351	CRIMINAL MISCHIEF & RELATED OF	254	…		
4	02/23/2018	351	CRIMINAL MISCHIEF & RELATED OF	259	…	F	
5	05/21/2018	361	OFF. AGNST PUB ORD SENSBLTY &	639	…	U	
6	09/01/2018	341	PETIT LARCENY	313	…		
7	05/03/2018	341	PETIT LARCENY	333	…	M	
8	08/16/2018	361	OFF. AGNST PUB ORD SENSBLTY &	639	…	M	
9	06/14/2018	351	CRIMINAL MISCHIEF & RELATED OF	259	…	M	
10	04/06/2018	235	DANGEROUS DRUGS	567	…	M	
11	04/28/2018	341	PETIT LARCENY	333	…	F	
12	10/06/2018	341	PETIT LARCENY	333	…	M	
13	07/21/2018	344	ASSAULT 3 & RELATED OFFENSES	101	…	M	
14	12/10/2018	341	PETIT LARCENY	321	…	U	
15	10/10/2018	344	ASSAULT 3 & RELATED OFFENSES	101	…	M	
16	05/18/2018	361	OFF. AGNST PUB ORD SENSBLTY &	639	…	M	

	transit_district	latitude	longitude	\
0		40.87367103500002	-73.90801364899994	
1		40.82616961200006	-73.91683070899995	
2		40.827049319000025	-73.89499419099997	
3		40.882615325000074	-73.85194765899996	
4		40.82870937100006	-73.87776995499998	
5		40.881300913000075	-73.85433733899998	
6		40.86840712200007	-73.89260767699994	
7		40.83778161800007	-73.91945797099999	
8		40.846705615000076	-73.86472139499993	
9		40.844996090000045	-73.85167356799997	
10		40.80806113500005	-73.92248131399998	
11		40.80719918600005	-73.91835350199995	
12		40.84306602000004	-73.83703668899994	
13		40.83857962500008	-73.92663072799998	
14		40.82667327000007	-73.88353694199996	
15		40.82142768200004	-73.91436893199995	
16		40.85036729300003	-73.91725589399994	

	lat_lon	patrol_boro	station_name	\
0	(40.873671035, -73.908013649)	PATROL BORO	BRONX	
1	(40.826169612, -73.916830709)	PATROL BORO	BRONX	
2	(40.827049319, -73.894994191)	PATROL BORO	BRONX	
3	(40.882615325, -73.851947659)	PATROL BORO	BRONX	
4	(40.828709371, -73.877769955)	PATROL BORO	BRONX	
5	(40.881300913, -73.854337339)	PATROL BORO	BRONX	
6	(40.868407122, -73.892607677)	PATROL BORO	BRONX	
7	(40.837781618, -73.919457971)	PATROL BORO	BRONX	
8	(40.846705615, -73.864721395)	PATROL BORO	BRONX	
9	(40.84499609, -73.851673568)	PATROL BORO	BRONX	
10	(40.808061135, -73.922481314)	PATROL BORO	BRONX	
11	(40.807199186, -73.918353502)	PATROL BORO	BRONX	
12	(40.84306602, -73.837036689)	PATROL BORO	BRONX	
13	(40.838579625, -73.926630728)	PATROL BORO	BRONX	
14	(40.82667327, -73.883536942)	PATROL BORO	BRONX	
15	(40.821427682, -73.914368932)	PATROL BORO	BRONX	
16	(40.850367293, -73.917255894)	PATROL BORO	BRONX	

	vic_age_group	vic_race	vic_sex
0	UNKNOWN	UNKNOWN	D
1	25-44	WHITE HISPANIC	F
2	45-64	BLACK	F
3	25-44	ASIAN / PACIFIC ISLANDER	F
4	UNKNOWN	UNKNOWN	D
5	<18	WHITE HISPANIC	F
6	65+	BLACK HISPANIC	F
7	UNKNOWN	UNKNOWN	D

```
8          25-44  ASIAN / PACIFIC ISLANDER       M
9          45-64                     WHITE       M
10       UNKNOWN                   UNKNOWN       E
11       UNKNOWN                   UNKNOWN       D
12       UNKNOWN                   UNKNOWN       D
13         45-64                     WHITE       M
14       UNKNOWN                   UNKNOWN       D
15         45-64            BLACK HISPANIC       F
16         18-24            WHITE HISPANIC       F

[17 rows x 35 columns]
```

[38]:
```
if not cri_df01_s01.empty:
    print("[OK]")
else:
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN REGISTERED WITH ATHENA. LOOK IN␣
 ↪PREVIOUS CELLS TO FIND THE ISSUE.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

6.7 Create Athena Table from Local TSV File - Evictions.tsv

[39]:
```
evi_tsv_tbl_name = 'evictions'
evi_tsv_field_list = """
court_index_number string,
docket_number string,
eviction_address string,
eviction_apartment_number string,
executed_date string,
marshal_first_name string,
marshal_last_name string,
residential_or_commercial string,
borough string,
eviction_postcode string,
ejectment string,
eviction_or_legal_possession string,
latitude string,
longitude string,
community_board string,
council_district string,
census_tract string,
bin string,
bbl string,
nta string
"""
evi_tsv_s3_raw_data_path = f"s3://{def_bucket}/team_8_data/raw_data/evictions"
```

```
print(evi_tsv_s3_raw_data_path)

create_athena_tbl_tsv(conn=conn,
                      db=database_name,
                      tbl_name=evi_tsv_tbl_name,
                      fields=evi_tsv_field_list,
                      s3_path=evi_tsv_s3_raw_data_path,
                      delim='\\t',
                      comp='',
                      skip="'skip.header.line.count'='1'")
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/evictions
Create table statement:

        CREATE EXTERNAL TABLE IF NOT EXISTS ads508_t8.evictions(
court_index_number string,
docket_number string,
eviction_address string,
eviction_apartment_number string,
executed_date string,
marshal_first_name string,
marshal_last_name string,
residential_or_commercial string,
borough string,
eviction_postcode string,
ejectment string,
eviction_or_legal_possession string,
latitude string,
longitude string,
community_board string,
council_district string,
census_tract string,
bin string,
bbl string,
nta string
)
        ROW FORMAT DELIMITED
            FIELDS
                TERMINATED BY '\t'
            LINES
                TERMINATED BY '\n'
        LOCATION 's3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/evictions'
        TBLPROPERTIES ('skip.header.line.count'='1')


        tab_name
0         census
```

```
1   census_block
2          crime
3     crime_pqt
4     evictions
5  grad_outcomes
6       hs_info
7          jobs
```

```
Dataframe contains records: True
```

6.7.1 Run A Sample Query

[40]:
```
evi_borough01 = "BRONX"

evi_select_dbn_stmnt = f"""
SELECT * FROM {database_name}.{evi_tsv_tbl_name}
WHERE borough = '{evi_borough01}'
LIMIT 17
"""

print(evi_select_dbn_stmnt)

evi_df01_s01 = pd.read_sql(evi_select_dbn_stmnt,
                           conn)

evi_df01_s01.head(17)
```

```
SELECT * FROM ads508_t8.evictions
WHERE borough = 'BRONX'
LIMIT 17
```

[40]:

	court_index_number	docket_number
0	56037/17	339568
1	B047517/19	409031
2	15068/17	334442
3	14866/19A	097278
4	66703/18BX	090391
5	B806500/18	396012
6	54026/17	341956
7	69137/18	10335
8	18348/16	324092
9	75943/16A	060118
10	44987/17	069826
11	B55293/17	101573
12	B54741/17	083789

```
13            8911/17        062195
14           72383/16        292013
15            7096/18        077442
16          902063/14        067290

                                    eviction_address  \
0                                547 EAST 168TH STREET
1                                4014 CARPENTER AVENUE
2                                655 EAST 224TH STREET
3                                  718 PENFIELD STREET
4   2032 EAST 177TH ST A /K/A 2032 CROSS BRONX EXP…
5                                281 EAST 143RD STREET
6                               1211 SOUTHERN BOULEVARD
7                             1351 BOSTON ROAD - APT 201
8                                2280 LORING PLACE NORTH
9                           1551    WILLIAMSBRID GE ROAD
10                               1514 SEDGWICK AVENUE
11                               2800 SEDGWICK AVENUE
12                              810 EAST 152ND STREE T
13                                   40 RICHMAN PLAZA
14                                 5 METROPOLITAN OVAL
15                                 3319 BAYCHESTER AVE
16  3488 JEROME AVENUE GROUND FLOOR STORE PREMISES…

   eviction_apartment_number executed_date marshal_first_name  \
0                         3H    02/26/2018             Thomas
1                         4B    11/16/2022            Richard
2                          1    09/29/2017             Thomas
3                        2-F    10/24/2019             Justin
4                         1E    07/30/2019             Justin
5                        07A    01/17/2019            Richard
6                        301    11/19/2018             Thomas
7                        201    07/15/2019             Robert
8                         4B    05/22/2017             Thomas
9                        4-B    08/10/2017             Justin
10                        7C    05/22/2018             Justin
11                        5I    01/19/2018            Darlene
12                       705    07/31/2018             Ileana
13                       11J    05/23/2017             Justin
14                        8D    10/13/2017             George
15                1ST FLOOR    08/21/2018             Justin
16                             02/07/2017              Henry

   marshal_last_name residential_or_commercial borough eviction_postcode  \
0                Bia                Residential   BRONX             10456
1              McCoy                Residential   BRONX             10466
2                Bia                Residential   BRONX             10467
```

```
3          Grossman               Residential   BRONX            10470
4          Grossman               Residential   BRONX            10472
5             McCoy               Residential   BRONX            10451
6               Bia               Residential   BRONX            10459
7          Renzulli               Residential   BRONX            10456
8               Bia               Residential   BRONX            10468
9          Grossman               Residential   BRONX            10461
10         Grossman               Residential   BRONX            10453
11           Barone               Residential   BRONX            10468
12           Rivera               Residential   BRONX            10455
13         Grossman               Residential   BRONX            10453
14           Essock               Residential   BRONX            10462
15         Grossman               Residential   BRONX            10469
16            Daley                Commercial   BRONX            10467

           ejectment eviction_or_legal_possession   latitude   longitude  \
0   Not an Ejectment                   Possession  40.830857  -73.905191
1   Not an Ejectment                   Possession  40.889878  -73.862686
2   Not an Ejectment                   Possession  40.887599  -73.862391
3   Not an Ejectment                   Possession  40.904888  -73.849089
4   Not an Ejectment                   Possession  40.831685  -73.856168
5   Not an Ejectment                   Possession  40.814845  -73.924083
6   Not an Ejectment                   Possession  40.828949  -73.891897
7   Not an Ejectment                   Possession  40.832166  -73.898808
8   Not an Ejectment                   Possession  40.861277  -73.908723
9   Not an Ejectment                   Possession
10  Not an Ejectment                   Possession  40.846731  -73.924961
11  Not an Ejectment                   Possession  40.871918  -73.902287
12  Not an Ejectment                   Possession  40.815435  -73.905199
13  Not an Ejectment                   Possession  40.852084  -73.922436
14  Not an Ejectment                   Possession  40.838831  -73.860013
15  Not an Ejectment                   Possession  40.878064  -73.836946
16  Not an Ejectment                   Possession

   community_board council_district census_tract      bin         bbl  \
0                3                16          145  2004227  2026100065
1               12                12          408  2063060  2048280031
2               12                12          394  2062985  2048260028
3               12                11          442  2071873  2051130039
4                9                18           78  2026230  2038030019
5                1                 8           51  2091116  2023240001
6                3                17          125  2113777  2029750037
7                3                16          151  2128618  2029340050
8                7                14          255  2014918  2032250015
9
10               5                16        20501  2114714  2028800009
11               8                14        26702  2015379  2032490202
```

```
12                1                 8           79  2094310  2026640061
13                5                16           53  2113629  2028820229
14                9                18        21001  2096599  2039447501
15               12                12        46202  2065413  2048810067
16
```

```
                                nta
0                  Claremont-Bathgate
1             Williamsbridge-Olinville
2             Williamsbridge-Olinville
3                   Woodlawn-Wakefield
4                Westchester-Unionport
5               Mott Haven-Port Morris
6                   Morrisania-Melrose
7                   Morrisania-Melrose
8                  Kingsbridge Heights
9
10  University Heights-Morris Heights
11               Van Cortlandt Village
12     Melrose South-Mott Haven North
13  University Heights-Morris Heights
14                         Parkchester
15                          Co-op City
16
```

[41]:
```
if not evi_df01_s01.empty:
    print("[OK]")
else:
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN REGISTERED WITH ATHENA. LOOK IN␣
 ↪PREVIOUS CELLS TO FIND THE ISSUE.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

6.8 Create Athena Table from Local TSV File - NYC _Jobs.tsv

[42]:
```
job_tsv_tbl_name = 'jobs'
job_tsv_field_list = """
job_id string,
agency string,
posting_type string,
num_of_positions string,
business_title string,
civil_service_title string,
title_classification string,
title_code_no string,
```

```
level string,
job_category string,
fulltime_or_parttime_indicator string,
career_level string,
salary_range_from string,
salary_range_to string,
salary_frequency string,
work_location string,
division_or_work_unit string,
job_description string,
minimum_qual_requirements string,
preferred_skills string,
additional_information string,
to_apply string,
hours_or_shift string,
work_location_1 string,
recruitment_contact string,
residency_requirement string,
posting_date string,
post_until string,
posting_updated string,
process_date string
"""
job_tsv_s3_raw_data_path = f"s3://{def_bucket}/team_8_data/raw_data/jobs"
print(job_tsv_s3_raw_data_path)

create_athena_tbl_tsv(conn=conn,
                      db=database_name,
                      tbl_name=job_tsv_tbl_name,
                      fields=job_tsv_field_list,
                      s3_path=job_tsv_s3_raw_data_path,
                      delim='\\t',
                      comp='',
                      skip="'skip.header.line.count'='1'")
```

```
s3://sagemaker-us-east-1-657724983756/team_8_data/raw_data/jobs
Create table statement:

        CREATE EXTERNAL TABLE IF NOT EXISTS ads508_t8.jobs(
job_id string,
agency string,
posting_type string,
num_of_positions string,
business_title string,
civil_service_title string,
title_classification string,
title_code_no string,
level string,
```

```
job_category string,
fulltime_or_parttime_indicator string,
career_level string,
salary_range_from string,
salary_range_to string,
salary_frequency string,
work_location string,
division_or_work_unit string,
job_description string,
minimum_qual_requirements string,
preferred_skills string,
additional_information string,
to_apply string,
hours_or_shift string,
work_location_1 string,
recruitment_contact string,
residency_requirement string,
posting_date string,
post_until string,
posting_updated string,
process_date string
)
        ROW FORMAT DELIMITED
            FIELDS
                TERMINATED BY '\t'
            LINES
                TERMINATED BY '\n'
        LOCATION 's3://sagemaker-us-
east-1-657724983756/team_8_data/raw_data/jobs'
        TBLPROPERTIES ('skip.header.line.count'='1')


        tab_name
0         census
1   census_block
2          crime
3      crime_pqt
4      evictions
5  grad_outcomes
6        hs_info
7           jobs

Dataframe contains records: True
```

6.8.1 Run A Sample Query

[43]:
```
job_agency01 = "HOUSING"

job_select_dbn_stmnt = f"""
SELECT * FROM {database_name}.{job_tsv_tbl_name}
WHERE agency LIKE '%{job_agency01}%'
LIMIT 17
"""

print(job_select_dbn_stmnt)

job_df01_s01 = pd.read_sql(job_select_dbn_stmnt,
                           conn)

job_df01_s01.head(17)
```

```
SELECT * FROM ads508_t8.jobs
WHERE agency LIKE '%HOUSING%'
LIMIT 17
```

[43]:

	job_id	agency	posting_type	num_of_positions	business_title
0	573469	HOUSING PRESERVATION & DVLPMNT	External	1	Strategic Program Development Analyst for the …
1	568091	HOUSING PRESERVATION & DVLPMNT	External	5	Case Manager for the Division of Tenant Resources
2	576376	NYC HOUSING AUTHORITY	Internal	1	CARETAKER X
3	571769	HOUSING PRESERVATION & DVLPMNT	Internal	2	Case Manager for the Division of Tenant Resources
4	575854	HOUSING PRESERVATION & DVLPMNT	External	1	Data & Analytics Manager, Division of Strategi…
5	554300	NYC HOUSING AUTHORITY	External	1	
6	575870	HOUSING PRESERVATION & DVLPMNT	Internal	1	
7	440244	NYC HOUSING AUTHORITY	External	1	
8	570222	NYC HOUSING AUTHORITY	Internal	1	
9	576674	NYC HOUSING AUTHORITY	External	1	
10	576328	HOUSING PRESERVATION & DVLPMNT	External	1	
11	566107	HOUSING PRESERVATION & DVLPMNT	External	1	
12	563049	NYC HOUSING AUTHORITY	Internal	1	
13	550123	NYC HOUSING AUTHORITY	Internal	1	
14	576272	HOUSING PRESERVATION & DVLPMNT	External	1	
15	576321	NYC HOUSING AUTHORITY	Internal	1	
16	573720	HOUSING PRESERVATION & DVLPMNT	External	1	

```
5   RESIDENT RELOCATION SERVICES COMMUNITY COORDIN…
6   Director of Manhattan Planning for the Divisio…
7                                      Senior Writer
8   SENIOR PROJECT MANAGER, ADULT EDUCATION & TRAI…
9                     SUPERVISOR OF HOUSING CARETAKER
10  Executive Director of Inclusionary Housing for…
11  Production Specialist for the Division of Tena…
12             SUPERVISING HOUSING GROUNDSKEEPER (HA)
13                                  Treasury Analyst
14  Director of the Supportive Housing Loan Progra…
15              ASSISTANT RESIDENT BUILDING SUPT (HA)
16  Housing Connect Project Manager,  Division of …

               civil_service_title title_classification title_code_no level  \
0          CITY RESEARCH SCIENTIST    Non-Competitive-5         21744    02
1              COMMUNITY ASSOCIATE    Non-Competitive-5         56057    00
2                   CARETAKER (HA)              Labor-3         90645    00
3              COMMUNITY ASSOCIATE    Non-Competitive-5         56057    00
4          CITY RESEARCH SCIENTIST    Non-Competitive-5         21744    02
5            COMMUNITY COORDINATOR    Non-Competitive-5         56058    00
6                     CITY PLANNER        Competitive-1         22122    03
7                  AGENCY ATTORNEY    Non-Competitive-5         30087    03
8   ASSOCIATE JOB OPPORTUNITY SPEC        Competitive-1         52316    02
9   SUPERVISOR OF HOUSING CARETAKE        Competitive-1         82011    00
10  ADMINISTRATIVE PROJECT DIRECTO    Non-Competitive-5         95566    M1
11             COMMUNITY ASSOCIATE    Non-Competitive-5         56057    00
12  SUPERVISING HOUSING GROUNDSKEE        Competitive-1         81350    00
13   ADMINISTRATIVE CLAIM EXAMINER        Competitive-1         1004E    00
14   ASSOCIATE HOUSING DEVELOPMENT        Competitive-1         22508    00
15  ASSISTANT RESIDENT BUILDING SU        Competitive-1         80305    00
16           COMMUNITY COORDINATOR    Non-Competitive-5         56058    00

                                       job_category  …  \
0                       Policy, Research & Analysis  …
1            Constituent Services & Community Programs  …
2                    Building Operations & Maintenance  …
3            Constituent Services & Community Programs  …
4                       Policy, Research & Analysis  …
5            Constituent Services & Community Programs  …
6                Engineering, Architecture, & Planning  …
7            Legal Affairs Policy, Research & Analysis  …
8            Constituent Services & Community Programs  …
9   Building Operations & Maintenance Public Safet…  …
10                  Finance, Accounting, & Procurement  …
11           Constituent Services & Community Programs  …
12                   Building Operations & Maintenance  …
13                  Finance, Accounting, & Procurement  …
```

```
14              Engineering, Architecture, & Planning  …
15                  Building Operations & Maintenance  …
16          Constituent Services & Community Programs  …

                                 additional_information  \
0   We engage New Yorkers to build and sustain nei…
1   Determination and verification of eligibility â ¢
2   Prepare apartments for move outs.   Please rea…
3   Determination and verification of eligibility â ¢
4   We engage New Yorkers to build and sustain nei…
5                                                 "1.
6   We engage New Yorkers to build and sustain nei…
7   Conducting operational analysis to understand …
8                                                 "1.
9                          Handle tenant lockouts.  4.
10  We engage New Yorkers to build and sustain nei…
11  We engage New Yorkers to build and sustain nei…
12  Train others in gardening techniques, grounds …
13  Monitor and project the daily flow of funds & …
14  We engage New Yorkers to build and sustain nei…
15  Monitor inventory supply and arrange for reple…
16  We engage New Yorkers to build and sustain nei…

                                               to_apply  \
0   Continue to work on the implementation of Loca…
1   Client briefings {internal and external meetin…
2   Qualification Requirements  There are no forma…
3   Client briefings {internal and external meetin…
4   Gather, prepare, and merge large datasets from…
5   Preference will be given to employees who have…
6   Planning & Predevelopment (P&P) is central to …
7   Conducting independent research of varying dif…
8   Preference will be given to employees who have…
9   Fill out work orders as a result of apartment …
10  1. A baccalaureate degree from an accredited c…
11  Reviewing recertification packages to ensure a…
12  Ensure the proper maintenance of all assigned …
13  Prepare, approve and release the daily Intra A…
14  In collaboration with the Assistant Commission…
15  Conduct building inspections and follow-up on …
16  Facilitating marketing and compliance meetings…

                                           hours_or_shift  \
0   Retrieve and review affordable housing regulat…
1   May perform community outreach to assist Secti…
2
3   May perform community outreach to assist Secti…
```

```
4   Create performance metrics for lottery and hom…
5           NYCHA residents are encouraged to apply."
6   Neighborhood Development & Stabilization (ND&S…
7   Organizing complex text and processes in seque…
8           NYCHA residents are encouraged to apply."
9   Report any hazardous conditions observed in an…
10  Candidates should have a record of achieving r…
11  Sort, record time/date and assign incoming ele…
12  Check repair work being performed by contracto…
13  Familiarity with Cash Management Systems (see …
14  Managing the development pipeline for SHLP and…
15  Monitor work orders in Maximo and deployment o…
16  Monitoring lotteries and reviewing lottery log…

                                      work_location_1  \
0   Research initiatives in other jurisdictions or…
1   Prepare and send appropriate correspondence, t…
2                                                 "1.
3   Prepare and send appropriate correspondence, t…
4   Manage and analyze eviction filing data to und…
5                          Click the Apply Now button.
6   Promote HPD and City policy objectives across …
7   Leading meetings with subject matter experts t…
8                          Click the Apply Now button.
9   One year of permanent service in the title of …
10  This position is also open to qualified person…
11  Maintain electronic files by uploading relevan…
12  Assistant Resident Building Superintendent in …
13  Prepare, approve and release third party wire …
14  Leading negotiations relating to deal structur…
15  Oversee the repair work done by Maintenance Wo…
16  Reviewing applicantsâ  files and required doc…

                                   recruitment_contact  \
0   Understand and leverage existing Agency datase…
1   Document case files and electronic records, fi…
2   Possession of a valid driver's license is requ…
3   Document case files and electronic records, fi…
4   Prepare analytic reports to inform program des…
5
6   Define, manage, and track team priorities and …
7   Editing documents of a high degree of difficul…
8
9
10                                     Apply online
11  Tracking packages and correspondence for accur…
12  Assist tenants in community development projec…
```

```
13  Responsible for making recurring payments to o…
14  Tracking and reporting of key project issues, …
15  Supervise the preparation of move-outs.  NOTE:…
16  Answering the Housing Connect hotline to assis…

                                residency_requirement  \
0   Summarizing and communicating findingsâ quali…
1                              Rent calculations  â ¢
2   Preference will be given to employees who have…
3                              Rent calculations  â ¢
4   Support the implementation of data-driven prog…
5
6   Meet regularly with individual staff members a…
7   Working with the Compliance Integration Report…
8
9                                                 "1.
10
11  Ensuring documents submitted to unit are distr…
12  1. A four-year high school diploma or its educ…
13  Prepare for more than 10 account activities su…
14  Supervising a team of Project Managers who are…
15  1. One year of permanent service in the title …
16  Reviewing and drafting correspondence in respo…

                                         posting_date  \
0   Contributing to the rollout of new initiatives…
1   Review of yearly recertificationâ s of househ…
2           NYCHA residents are encouraged to apply."
3   Review of yearly recertificationâ s of househ…
4   Develop strategies for data integration and au…
5
6   Identify staffing needs and advocate for resou…
7   Working with the Compliance Monitoring Unit to…
8
9   For NYCHA employees: This position is open as …
10                                    100 Gold Street
11  Entering data into Elite or another unit datab…
12  Three years of satisfactory full-time gardenin…
13  Check the daily Wire Register Reports and prep…
14  Identifying opportunities to train and build o…
15
16  1. A baccalaureate degree from an accredited c…

                                            post_until  \
0   Conducting special research, analytical, or co…
1   Demonstrate ability to manage multiple cases w…
2                         Click the Apply now button.
```

```
3   Demonstrate ability to manage multiple cases w…
4   Assist SOA colleagues with quantitative analys…
5                NYCHA has no residency requirements.
6   Ensure that all projects move efficiently thro…
7   Working with Compliance Inquiry Review and Ass…
8                NYCHA has no residency requirements.
9   For NYCHA employees: Preference will be given …
10
11  Contacting participants to inquire about recer…
12                                                 "1.
13  Prepare daily Reconciliation for the Net Chang…
14  Managing special projects, including developme…
15                                                 "1.
16                                                "â ¢

                                      posting_updated  \
0   Adhering to work plans and internal and extern…
1                          Attend mandatory trainings"
2
3                          Attend mandatory trainings"
4   Respond to ad hoc data requests from programs,…
5                                           10/28/2022
6   Identify risks and troubleshoot problems, invo…
7   Coordinating with NYCHA department heads regar…
8                                           02/14/2023
9           NYCHA residents are encouraged to apply."
10  New York City residency is generally required …
11  Assisting clients with inquiries regarding doc…
12  For NYCHA employees, this position is open as …
13  Work closely with bank personnel to resolve an…
14  Communicating with elected officials, other Ci…
15  For NYCHA employees, this position is open as …
16  Strong analytical ability and attention to det…

                                        process_date
0   Participating in meetings, presentations, and …
1   Qualification Requirements  1. High school gra…
2
3   Qualification Requirements  1. High school gra…
4   1.  For Assignment Level I (only physical, bio…
5
6   Create, implement, and maintain consistent, ef…
7   Drafting complex documents based on important …
8
9                          Click the Apply Now button.
10                                          02/24/2023
11  Meeting with participants to assist with compl…
```

```
12  For NYCHA employees, preference will be given …
13  Assist with procurement of cash management ser…
14  Other responsibilities and initiatives as may …
15  For NYCHA employees, preference will be given …
16  Strong time management skills, demonstrated ab…

[17 rows x 30 columns]
```

[44]:
```
if not job_df01_s01.empty:
    print("[OK]")
else:
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN REGISTERED WITH ATHENA. LOOK IN␣
 ↪PREVIOUS CELLS TO FIND THE ISSUE.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

7 Create Parquet Files from TSV Table

[45]:
```
ingest_create_athena_table_parquet_passed = False
```

[46]:
```
%store -r ingest_create_athena_table_tsv_passed
```

[47]:
```
try:
    ingest_create_athena_table_tsv_passed
except NameError:
    print("++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL PREVIOUS NOTEBOOKS.  You did not␣
 ↪register the TSV Data.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++")
```

[48]:
```
print(ingest_create_athena_table_tsv_passed)
```

```
True
```

[49]:
```
if not ingest_create_athena_table_tsv_passed:
    print("++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOU HAVE TO RUN ALL PREVIOUS NOTEBOOKS.  You did not␣
 ↪register the TSV Data.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++")
else:
    print("[OK]")
```

```
[OK]
```

[50]:
```
# Set S3 path to Parquet data
cri_pqt_s3_data_path = f"s3://{def_bucket}/team_8_data/columnar"
```

8 Execute Statement

[51]:
```
cri_pqt_tbl_name = 'crime_pqt'
drop_pqt_tbl_stmnt = f"""DROP TABLE IF EXISTS {database_name}.
 ↪{cri_pqt_tbl_name}"""

# SQL statement to execute
create_pqt_tble_stmnt = f"""
CREATE TABLE IF NOT EXISTS {database_name}.{cri_pqt_tbl_name}
WITH (
    format = 'PARQUET',
    external_location = '{cri_pqt_s3_data_path}',
    partitioned_by = ARRAY['law_cat_cd', 'borough']
    )
AS
SELECT
    cmplnt_num,
    cmplnt_fr_dt,
    cmplnt_fr_tm,
    cmplnt_to_dt,
    cmplnt_to_tm,
    addr_pct_cd,
    rpt_dt,
    ky_cd,
    ofns_desc,
    pd_cd,
    pd_desc,
    crm_atpt_cptd_cd,
    loc_of_occur_desc,
    prem_typ_desc,
    juris_desc,
    jurisdiction_code,
    parks_nm,
    hadevelopt,
    housing_psa,
    x_coord_cd,
    y_coord_cd,
    susp_age_group,
    susp_race,
    susp_sex,
    transit_district,
    latitude,
    longitude,
    lat_lon,
    patrol_boro,
    station_name,
    vic_age_group,
```

```
    vic_race,
    vic_sex,
    law_cat_cd,
    borough
FROM {database_name}.{cri_tsv_tbl_name}
TABLESAMPLE BERNOULLI(2)
"""

print(f'Create table statement:\n{create_pqt_tble_stmnt}')

pd.read_sql(drop_pqt_tbl_stmnt,
            conn)

pd.read_sql(create_pqt_tble_stmnt,
            conn)
```

```
Create table statement:

CREATE TABLE IF NOT EXISTS ads508_t8.crime_pqt
WITH (
    format = 'PARQUET',
    external_location = 's3://sagemaker-us-
east-1-657724983756/team_8_data/columnar',
    partitioned_by = ARRAY['law_cat_cd', 'borough']
    )
AS
SELECT
    cmplnt_num,
    cmplnt_fr_dt,
    cmplnt_fr_tm,
    cmplnt_to_dt,
    cmplnt_to_tm,
    addr_pct_cd,
    rpt_dt,
    ky_cd,
    ofns_desc,
    pd_cd,
    pd_desc,
    crm_atpt_cptd_cd,
    loc_of_occur_desc,
    prem_typ_desc,
    juris_desc,
    jurisdiction_code,
    parks_nm,
    hadevelopt,
    housing_psa,
    x_coord_cd,
    y_coord_cd,
```

```
    susp_age_group,
    susp_race,
    susp_sex,
    transit_district,
    latitude,
    longitude,
    lat_lon,
    patrol_boro,
    station_name,
    vic_age_group,
    vic_race,
    vic_sex,
    law_cat_cd,
    borough
FROM ads508_t8.crime
TABLESAMPLE BERNOULLI(2)
```

[51]:
```
Empty DataFrame
Columns: [rows]
Index: []
```

9 Load partitions by running MSCK REPAIR TABLE

[52]:
```
partition_pqt_stmnt = f"MSCK REPAIR TABLE {database_name}.{cri_pqt_tbl_name}"

print(partition_pqt_stmnt)
```

```
MSCK REPAIR TABLE ads508_t8.crime_pqt
```

[53]:
```
cri_df02 = pd.read_sql(partition_pqt_stmnt,
                       conn)

cri_df02.head(17)
```

[53]:
```
Empty DataFrame
Columns: []
Index: []
```

10 Show the Partitions

[54]:
```
show_part_stmnt = f"SHOW PARTITIONS {database_name}.{cri_pqt_tbl_name}"

print(show_part_stmnt)
```

```
SHOW PARTITIONS ads508_t8.crime_pqt
```

[55]:
```
cri_df02_part = pd.read_sql(show_part_stmnt,
                            conn)

cri_df02_part.head(31)
```

[55]:

	partition
0	law_cat_cd=MISDEMEANOR/borough=BROOKLYN
1	law_cat_cd=VIOLATION/borough=MANHATTAN
2	law_cat_cd=MISDEMEANOR/borough=__HIVE_DEFAULT_…
3	law_cat_cd=FELONY/borough=__HIVE_DEFAULT_PARTI…
4	law_cat_cd=FELONY/borough=BROOKLYN
5	law_cat_cd=MISDEMEANOR/borough=MANHATTAN
6	law_cat_cd=MISDEMEANOR/borough=STATEN ISLAND
7	law_cat_cd=VIOLATION/borough=QUEENS
8	law_cat_cd=MISDEMEANOR/borough=BRONX
9	law_cat_cd=VIOLATION/borough=__HIVE_DEFAULT_PA…
10	law_cat_cd=FELONY/borough=MANHATTAN
11	law_cat_cd=FELONY/borough=BRONX
12	law_cat_cd=VIOLATION/borough=STATEN ISLAND
13	law_cat_cd=FELONY/borough=STATEN ISLAND
14	law_cat_cd=VIOLATION/borough=BROOKLYN
15	law_cat_cd=FELONY/borough=QUEENS
16	law_cat_cd=VIOLATION/borough=BRONX
17	law_cat_cd=MISDEMEANOR/borough=QUEENS

11 Show the Tables

[56]:
```
show_tbl_stmnt = f"SHOW TABLES in {database_name}"
```

[57]:
```
df_tables = pd.read_sql(show_tbl_stmnt,
                        conn)

df_tables.head(17)
```

[57]:

	tab_name
0	census
1	census_block
2	crime
3	crime_pqt
4	evictions
5	grad_outcomes
6	hs_info
7	jobs

[58]:
```
if cri_pqt_tbl_name in df_tables.values:
    ingest_create_athena_table_parquet_passed = True
```

[59]:
```
%store ingest_create_athena_table_parquet_passed
```

```
Stored 'ingest_create_athena_table_parquet_passed' (bool)
```

12 Run Sample Query

[60]:
```
cri_select_dbn_stmnt02 = f"""
SELECT * FROM {database_name}.{cri_pqt_tbl_name}
WHERE LOWER(law_cat_cd) = '{cri_law_cat_cd01}'
    AND LOWER(borough) = '{cri_borough01}'
LIMIT 17
"""

print(cri_select_dbn_stmnt02)

cri_df02_s01 = pd.read_sql(cri_select_dbn_stmnt02,
                           conn)

cri_df02_s01.head(17)
```

```
SELECT * FROM ads508_t8.crime_pqt
WHERE LOWER(law_cat_cd) = 'misdemeanor'
    AND LOWER(borough) = 'bronx'
LIMIT 17
```

[60]:

	cmplnt_num	cmplnt_fr_dt	cmplnt_fr_tm	cmplnt_to_dt	cmplnt_to_tm	addr_pct_cd	\
0	590499848	07/14/2013	05:30:00	07/14/2013	05:43:00	43	
1	675597625	05/27/2017	21:15:00	05/27/2017	21:25:00	45	
2	336036648	03/21/2012	13:00:00			45	
3	388744844	05/25/2015	04:25:00	05/25/2015	04:30:00	52	
4	706022394	02/07/2016	16:00:00			47	
5	281742597	08/29/2013	00:25:00	08/29/2013	00:30:00	40	
6	411962262	04/09/2013	16:55:00			46	
7	545346999	10/01/2017	18:00:00	10/01/2017	18:15:00	52	
8	818313819	05/30/2017	15:15:00	05/30/2017	15:26:00	46	
9	890406291	03/11/2013	21:25:00	03/12/2013	12:00:00	45	
10	973380835	03/31/2015	19:25:00	03/31/2015	19:34:00	42	
11	625369326	03/28/2016	16:10:00	03/28/2016	16:15:00	40	
12	532182401	04/18/2013	21:30:00			42	
13	251257334	04/26/2016	08:30:00	04/26/2016	08:30:00	44	
14	863725090	03/16/2011	18:45:00	03/16/2011	18:50:00	43	
15	127246599	01/18/2011	01:00:00			48	
16	593364190	06/15/2015	09:45:00			43	

	rpt_dt	ky_cd	ofns_desc	pd_cd	…	latitude	\

```
0   07/14/2013   340                          FRAUDS   707  …  40.823101299
1   05/27/2017   235                 DANGEROUS DRUGS   511  …  40.848632895
2   03/21/2012   351  CRIMINAL MISCHIEF & RELATED OF   254  …  40.830889993
3   05/25/2015   351  CRIMINAL MISCHIEF & RELATED OF   259  …  40.868812402
4   02/10/2016   344    ASSAULT 3 & RELATED OFFENSES   114  …  40.886936175
5   08/29/2013   344    ASSAULT 3 & RELATED OFFENSES   101  …  40.811116426
6   04/09/2013   358        OFFENSES INVOLVING FRAUD   705  …  40.861886273
7   10/02/2017   341                   PETIT LARCENY   339  …  40.865155015
8   05/30/2017   352              CRIMINAL TRESPASS   205  …  40.849496743
9   03/12/2013   359  OFFENSES AGAINST PUBLIC ADMINI   748  …  40.827532802
10  03/31/2015   233                       SEX CRIMES   681  …  40.822569916
11  03/28/2016   344    ASSAULT 3 & RELATED OFFENSES   101  …  40.812053484
12  04/18/2013   359  OFFENSES AGAINST PUBLIC ADMINI   748  …   40.82935201
13  04/26/2016   235                 DANGEROUS DRUGS   511  …   40.82386862
14  03/16/2011   235                 DANGEROUS DRUGS   567  …  40.830447407
15  01/18/2011   359  OFFENSES AGAINST PUBLIC ADMINI   749  …   40.84139516
16  06/15/2015   361   OFF. AGNST PUB ORD SENSBLTY &   639  …  40.836612833

       longitude                      lat_lon        patrol_boro  \
0   -73.869690461  (40.823101299, -73.869690461)  PATROL BORO BRONX
1     -73.8279976    (40.848632895, -73.8279976)  PATROL BORO BRONX
2    -73.82728462   (40.830889993, -73.82728462)  PATROL BORO BRONX
3   -73.888723856  (40.868812402, -73.888723856)  PATROL BORO BRONX
4    -73.85249861   (40.886936175, -73.85249861)  PATROL BORO BRONX
5   -73.927329309  (40.811116426, -73.927329309)  PATROL BORO BRONX
6    -73.89320749   (40.861886273, -73.89320749)  PATROL BORO BRONX
7   -73.892996163  (40.865155015, -73.892996163)  PATROL BORO BRONX
8   -73.909315789  (40.849496743, -73.909315789)  PATROL BORO BRONX
9   -73.821613113  (40.827532802, -73.821613113)  PATROL BORO BRONX
10  -73.911307169  (40.822569916, -73.911307169)  PATROL BORO BRONX
11   -73.90950047   (40.812053484, -73.90950047)  PATROL BORO BRONX
12   -73.89188659    (40.82935201, -73.89188659)  PATROL BORO BRONX
13  -73.919402547   (40.82386862, -73.919402547)  PATROL BORO BRONX
14  -73.875790162  (40.830447407, -73.875790162)  PATROL BORO BRONX
15  -73.885520622   (40.84139516, -73.885520622)  PATROL BORO BRONX
16  -73.864315481  (40.836612833, -73.864315481)  PATROL BORO BRONX

   station_name vic_age_group        vic_race vic_sex   law_cat_cd borough
0                                    UNKNOWN       E  MISDEMEANOR   BRONX
1                     UNKNOWN        UNKNOWN       E  MISDEMEANOR   BRONX
2                       45-64  WHITE HISPANIC       M  MISDEMEANOR   BRONX
3                                    UNKNOWN       D  MISDEMEANOR   BRONX
4                       25-44          BLACK       F  MISDEMEANOR   BRONX
5                       18-24          BLACK       F  MISDEMEANOR   BRONX
6                                    UNKNOWN       D  MISDEMEANOR   BRONX
7                         65+  WHITE HISPANIC       F  MISDEMEANOR   BRONX
8                     UNKNOWN  WHITE HISPANIC       M  MISDEMEANOR   BRONX
```

```
9                       25-44           WHITE       F  MISDEMEANOR   BRONX
10                                    UNKNOWN       E  MISDEMEANOR   BRONX
11                      25-44  WHITE HISPANIC       F  MISDEMEANOR   BRONX
12                      18-24           BLACK       F  MISDEMEANOR   BRONX
13                    UNKNOWN         UNKNOWN       E  MISDEMEANOR   BRONX
14                                    UNKNOWN       E  MISDEMEANOR   BRONX
15                      25-44           BLACK       M  MISDEMEANOR   BRONX
16                      25-44           BLACK       F  MISDEMEANOR   BRONX

[17 rows x 35 columns]
```

[61]:
```
if not cri_df02_s01.empty:
    print("[OK]")
else:
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
    print("[ERROR] YOUR DATA HAS NOT BEEN CONVERTED TO PARQUET. LOOK IN␣
 ↪PREVIOUS CELLS TO FIND THE ISSUE.")
    print("++++++++++++++++++++++++++++++++++++++++++++++++++++++++++")
```

```
[OK]
```

12.1 Review the New Athena Table in the Glue Catalog

[62]:
```
display(
    HTML(
        f'<b>Review <a target="top" href="https://console.aws.amazon.com/glue/
 ↪home?region={region}#">AWS Glue Catalog</a></b>'
    )
)
```

```
<IPython.core.display.HTML object>
```

12.2 Store Variables for the Next Notebooks

[63]:
```
%store
```

```
Stored variables and their in-db values:
balance_dataset                                 -> True
balanced_bias_data_jsonlines_s3_uri             -> 's3://sagemaker-us-
east-1-657724983756/bias-detect
balanced_bias_data_s3_uri                       -> 's3://sagemaker-us-
east-1-657724983756/bias-detect
bias_data_s3_uri                                -> 's3://sagemaker-us-
east-1-657724983756/bias-detect
experiment_name                                 -> 'Amazon-Customer-
Reviews-BERT-Experiment-168013737
feature_group_name                              -> 'reviews-feature-
group-1680137375'
feature_store_offline_prefix                    -> 'reviews-feature-
```

```
store-1680137375'
ingest_create_athena_db_passed                          -> True
ingest_create_athena_table_parquet_passed               -> True
ingest_create_athena_table_tsv_passed                   -> True
max_seq_length                                          -> 64
processed_test_data_s3_uri                              -> 's3://sagemaker-us-
east-1-657724983756/sagemaker-s
processed_train_data_s3_uri                             -> 's3://sagemaker-us-
east-1-657724983756/sagemaker-s
processed_validation_data_s3_uri                        -> 's3://sagemaker-us-
east-1-657724983756/sagemaker-s
raw_input_data_s3_uri                                   -> 's3://sagemaker-us-
east-1-657724983756/amazon-revi
s3_private_path_tsv                                     -> 's3://sagemaker-us-
east-1-657724983756/team_8_data
s3_public_path_tsv                                      -> 's3://sagemaker-us-
east-ads508-sp23-t8'
setup_dependencies_passed                               -> True
setup_iam_roles_passed                                  -> True
setup_instance_check_passed                             -> True
setup_s3_bucket_passed                                  -> True
test_split_percentage                                   -> 0.05
train_split_percentage                                  -> 0.9
trial_name                                              -> 'trial-1680137374'
validation_split_percentage                             -> 0.05
```

12.3 Release Resources

[64]:
```
%%html

<p><b>Shutting down your kernel for this notebook to release resources.</b></p>
<button class="sm-command-button" data-commandlinker-command="kernelmenu:
 ↪shutdown" style="display:none;">Shutdown Kernel</button>

<script>
try {
    els = document.getElementsByClassName("sm-command-button");
    els[0].click();
}
catch(err) {
    // NoOp
}
</script>
```

```
<IPython.core.display.HTML object>
```

[65]:
```
%%javascript
```

```
try {
    Jupyter.notebook.save_checkpoint();
    Jupyter.notebook.session.delete();
}
catch(err) {
    // NoOp
}
```

```
<IPython.core.display.Javascript object>
```