

```
In [236]: import boto3
import configparser
from datetime import date
from venv import create
from matplotlib.pyplot import table
# Importing packages needed
import pandas as pd
from sklearn import pipeline
from sqlalchemy import VARCHAR
print("Data Pipeline")
import os
import numpy as np
import configparser
import sys

Data Pipeline
```

Process needed installs needed to build pipeline

```
In [15]: pip install s3fs

Requirement already satisfied: s3fs in /opt/anaconda3/lib/python3.7/site-packages (2022.2.0)
Requirement already satisfied: fsspec==2022.02.0 in /opt/anaconda3/lib/python3.7/site-packages (from s3fs) (2022.2.0)
Requirement already satisfied: aioboto3==2.1.0 in /opt/anaconda3/lib/python3.7/site-packages (from s3fs) (2.1.1)
Requirement already satisfied: aiohttp<4 in /opt/anaconda3/lib/python3.7/site-packages (from s3fs) (3.6.2)
Requirement already satisfied: wrapt>=1.10.10 in /opt/anaconda3/lib/python3.7/site-packages (from aio botocore==2.1.0->s3fs) (1.11.2)
Requirement already satisfied: aioitertools>=0.5.1 in /opt/anaconda3/lib/python3.7/site-packages (fro m aioboto3==2.1.0->s3fs) (0.10.0)
Collecting botocore<1.23.25,>=1.23.24
  Using cached https://files.pythonhosted.org/packages/ef/05/cb9048c5f77ab9fa6970d2c88bd9247dc4ccc4 bcf5b59c645db9cbdedc/botocore-1.23.24-py3-none-any.whl
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/anaconda3/lib/python3.7/site-packages (from aio http<4->s3fs) (1.4.2)
Requirement already satisfied: async-timeout<4.0,>=3.0 in /opt/anaconda3/lib/python3.7/site-packages (from aiohttp<4->s3fs) (3.0.1)
Requirement already satisfied: attrs>=17.3.0 in /opt/anaconda3/lib/python3.7/site-packages (from aioh ttp<4->s3fs) (19.3.0)
Requirement already satisfied: chardet<4.0,>=2.0 in /opt/anaconda3/lib/python3.7/site-packages (from aiohttp<4->s3fs) (3.0.4)
Requirement already satisfied: multidict<5.0,>=4.5 in /opt/anaconda3/lib/python3.7/site-packages (fro m aiohttp<4->s3fs) (4.7.3)
Requirement already satisfied: typing_extensions>=4.0; python version < "3.10" in /opt/anaconda3/lib/ python3.7/site-packages (from aioitertools>=0.5.1->aiobotocore==2.1.0->s3fs) (4.1.1)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in /opt/anaconda3/lib/python3.7/site-packages (from botocore<1.23.25,>=1.23.24->aiobotocore==2.1.0->s3fs) (0.10.0)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in /opt/anaconda3/lib/python3.7/site-packages (f rom botocore<1.23.25,>=1.23.24->aiobotocore==2.1.0->s3fs) (1.25.7)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /opt/anaconda3/lib/python3.7/site-packa ges (from botocore<1.23.25,>=1.23.24->aiobotocore==2.1.0->s3fs) (2.8.1)
Requirement already satisfied: idna>=2.0 in /opt/anaconda3/lib/python3.7/site-packages (from yarl<2. 0,>=1.0->aiohttp<4->s3fs) (2.8)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.7/site-packages (from python-da teutil<3.0.0,>=2.1->botocore<1.23.25,>=1.23.24->aiobotocore==2.1.0->s3fs) (1.13.0)
ERROR: boto3 1.21.8 has requirement botocore<1.25.0,>=1.24.8, but you'll have botocore 1.23.24 which is incompatible.
ERROR: awscli 1.22.63 has requirement botocore==1.24.8, but you'll have botocore 1.23.24 which is inc ompatible.
Installing collected packages: botocore
  Found existing installation: botocore 1.24.8
    Uninstalling botocore-1.24.8:
      Successfully uninstalled botocore-1.24.8
Successfully installed botocore-1.23.24
Note: you may need to restart the kernel to use updated packages.

In [16]: import psycopg2 as db

In [17]: pip install fsspec==2022.02.0

Requirement already satisfied: fsspec==2022.02.0 in /opt/anaconda3/lib/python3.7/site-packages (2022. 2.0)
Note: you may need to restart the kernel to use updated packages.

In [18]: pip install touch

Requirement already satisfied: touch in /opt/anaconda3/lib/python3.7/site-packages (2020.12.3)
Requirement already satisfied: values in /opt/anaconda3/lib/pytho3.7/site-packages (from touch) (202 0.12.3)
Note: you may need to restart the kernel to use updated packages.

In [19]: pip install botocore

Requirement already satisfied: botocore in /opt/anaconda3/lib/python3.7/site-packages (1.23.24)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in /opt/anaconda3/lib/python3.7/site-packages (from botocore) (0.10.0)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in /opt/anaconda3/lib/python3.7/site-packages (f rom botocore) (1.25.7)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /opt/anaconda3/lib/python3.7/site-packa ges (from botocore) (2.8.1)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.7/site-packages (from python-da teutil<3.0.0,>=2.1->botocore) (1.13.0)
Note: you may need to restart the kernel to use updated packages.
```

Install AWS Command line Interface

```
In [20]: pip install awscli

Requirement already satisfied: awscli in /opt/anaconda3/lib/python3.7/site-packages (1.22.63)
Requirement already satisfied: colorama<0.4.4,>=0.2.5 in /opt/anaconda3/lib/python3.7/site-packages (from awscli) (0.4.3)
Requirement already satisfied: PyYAML<5.5,>=3.10 in /opt/anaconda3/lib/python3.7/site-packages (from awscli) (5.2)
Requirement already satisfied: rsa<4.8,>=3.1.2 in /opt/anaconda3/lib/python3.7/site-packages (from aw scli) (4.7.2)
Requirement already satisfied: s3transfer<0.6.0,>=0.5.0 in /opt/anaconda3/lib/python3.7/site-packages (from awscli) (0.5.2)
Collecting botocore==1.24.8
  Using cached https://files.pythonhosted.org/packages/ae/e6/f841946cf32a862ddfa0a6de1114339033b1f34a 6f94d2cb868ab68a96f0/botocore-1.24.8-py3-none-any.whl
Requirement already satisfied: docutils<0.16,>=0.10 in /opt/anaconda3/lib/python3.7/site-packages (fr om awscli) (0.15.2)
Requirement already satisfied: pyasn1>=0.1.3 in /opt/anaconda3/lib/python3.7/site-packages (from rsa< 4.8,>=3.1.2->awscli) (0.4.8)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in /opt/anaconda3/lib/python3.7/site-packages (from botocore==1.24.8->awscli) (0.10.0)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in /opt/anaconda3/lib/python3.7/site-packages (f rom botocore==1.24.8->awscli) (1.25.7)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /opt/anaconda3/lib/python3.7/site-packa ges (from botocore==1.24.8->awscli) (2.8.1)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.7/site-packages (from python-da teutil<3.0.0,>=2.1->botocore==1.24.8->awscli) (1.13.0)
ERROR: aiobotocore 2.1.1 has requirement botocore<1.23.25,>=1.23.24, but you'll have botocore 1.24.8 which is incompatible.
Installing collected packages: botocore
  Found existing installation: botocore 1.23.24
    Uninstalling botocore-1.23.24:
      Successfully uninstalled botocore-1.23.24
Successfully installed botocore-1.24.8
Note: you may need to restart the kernel to use updated packages.
```

Install configure and botocore

```
In [21]: pip install configure

Requirement already satisfied: configure in /opt/anaconda3/lib/python3.7/site-packages (0.5)
Requirement already satisfied: pyyaml in /opt/anaconda3/lib/python3.7/site-packages (from configure) (5.2)
Note: you may need to restart the kernel to use updated packages.

In [237]: pip install botocore

Requirement already satisfied: botocore in /opt/anaconda3/lib/python3.7/site-packages (1.24.8)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in /opt/anaconda3/lib/python3.7/site-packages (from botocore) (0.10.0)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /opt/anaconda3/lib/python3.7/site-packa ges (from botocore) (2.8.1)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in /opt/anaconda3/lib/python3.7/site-packages (f rom botocore) (1.25.7)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.7/site-packages (from python-da teutil<3.0.0,>=2.1->botocore) (1.13.0)
Note: you may need to restart the kernel to use updated packages.
```

Import data from Amazon S3 into python to do data transformation.

```
In [23]: client = boto3.client('s3')

In [24]: path = 's3://mod7-pipeline/raw_sales.csv'

In [25]: path

Out[25]: 's3://mod7-pipeline/raw_sales.csv'

In [26]: df = pd.read_csv(path)

In [27]: df = pd.read_csv(path)

In [238]: df.head(10)

Out[238]:
```

	datesold	postcode	price	propertyType	bedrooms
0	2007-02-07 00:00:00	2607	525000	house	4
1	2007-02-27 00:00:00	2906	290000	house	3
2	2007-03-07 00:00:00	2905	328000	house	3
3	2007-03-09 00:00:00	2905	380000	house	4
4	2007-03-21 00:00:00	2906	310000	house	3
5	2007-04-04 00:00:00	2905	465000	house	4
6	2007-04-24 00:00:00	2607	399000	house	3
7	2007-04-30 00:00:00	2606	1530000	house	4
8	2007-05-24 00:00:00	2902	359000	house	3
9	2007-05-25 00:00:00	2906	320000	house	3

Creating a condensed dataframe to import into SQLite

```
In [234]: db_1 = df[['price', 'bedrooms', 'propertyType']]

In [235]: db_1

Out[235]:
```

	price	bedrooms	propertyType
0	525000	4	house
1	290000	3	house
2	328000	3	house
3	380000	4	house
4	310000	3	house
...
29575	500000	3	unit
29576	560000	2	unit
29577	464950	2	unit
29578	589000	2	unit
29579	775000	2	unit

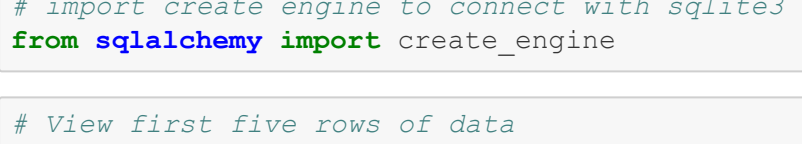
Import visualization packages to view data

```
In [29]: import matplotlib.pyplot as plt
import seaborn as sns

In [30]: sns.barplot(y='price', x='bedrooms', data=df);
```



```
In [31]: sns.barplot(y='price', x='propertyType', data=df);
```



Connect Jupyter notebook dataframe to SQLite to begin running sql queries for transformation

```
In [231]: pip install sqlalchemy

Requirement already satisfied: sqlalchemy in /opt/anaconda3/lib/python3.7/site-packages (1.3.11)
Note: you may need to restart the kernel to use updated packages.

In [232]: # import create engine to connect with sqlite3
from sqlalchemy import create_engine

In [233]: # View first five rows of data
df.head()

Out[233]:
```

	datesold	postcode	price	propertyType	bedrooms
0	2007-02-07 00:00:00	2607	525000	house	4
1	2007-02-27 00:00:00	2906	290000	house	3
2	2007-03-07 00:00:00	2905	328000	house	3
3	2007-03-09 00:00:00	2905	380000	house	4
4	2007-03-21 00:00:00	2906	310000	house	3

```
In [225]: #creating a file to import into sqlite3
engine = create_engine('sqlite:///df.db', echo=True)
sqlite_connection = engine.connect()

2022-03-02 04:38:07,109 INFO sqlalchemy.engine.base.Engine SELECT CAST('test plain returns' AS VARCHAR
R(60)) AS anon_1
2022-03-02 04:38:07,111 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:07,115 INFO sqlalchemy.engine.base.Engine SELECT CAST('test unicode returns' AS VARC
HAR(60)) AS anon_1
2022-03-02 04:38:07,116 INFO sqlalchemy.engine.base.Engine ()

In [228]: #Creating a table to perform transformation
sqlite_table = 'salesdata1'
db_1.to_sql(sqlite_table, sqlite_connection, if_exists='fail')

2022-03-02 04:38:03,664 INFO sqlalchemy.engine.base.Engine PRAGMA main.table_info("salesdata1")
2022-03-02 04:38:03,666 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:03,668 INFO sqlalchemy.engine.base.Engine PRAGMA temp.table_info("salesdata1")
2022-03-02 04:38:03,670 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:03,674 INFO sqlalchemy.engine.base.Engine
CREATE TABLE salesdata1 (
  "index" BIGINT,
  price BIGINT,
  bedrooms BIGINT,
  "propertyType" TEXT
)

2022-03-02 04:38:03,676 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:03,680 INFO sqlalchemy.engine.base.Engine COMMIT
2022-03-02 04:38:03,682 INFO sqlalchemy.engine.base.Engine CREATE INDEX ix_salesdata1_index ON salesd
atal ("index")
2022-03-02 04:38:03,683 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:03,687 INFO sqlalchemy.engine.base.Engine COMMIT
2022-03-02 04:38:03,698 INFO sqlalchemy.engine.base.Engine BEGIN (implicit)
2022-03-02 04:38:04,121 INFO sqlalchemy.engine.base.Engine INSERT INTO salesdata1 ("index", price, be
drooms, "propertyType") VALUES (?, ?, ?, ?)
2022-03-02 04:38:04,122 INFO sqlalchemy.engine.base.Engine ((0, 525000, 4, 'house'), (1, 290000, 3,
'house'), (2, 328000, 3, 'house'), (3, 380000, 4, 'house'), (4, 310000, 3, 'house'), (5, 465000, 4,
'house'), (6, 399000, 3, 'house'), (7, 1530000, 4, 'house')) ... displaying 10 of 29580 total bound p
arameter sets ... (29578, 589000, 2, 'unit'), (29579, 775000, 2, 'unit'))
2022-03-02 04:38:04,196 INFO sqlalchemy.engine.base.Engine COMMIT

In [214]: #creating a file to import into sqlite3
engine = create_engine('sqlite:///db_1.db', echo=True)
sqlite_connection = engine.connect()

2022-03-02 04:38:29,168 INFO sqlalchemy.engine.base.Engine SELECT CAST('test plain returns' AS VARCHAR
R(60)) AS anon_1
2022-03-02 04:38:29,169 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:29,173 INFO sqlalchemy.engine.base.Engine SELECT CAST('test unicode returns' AS VARC
HAR(60)) AS anon_1
2022-03-02 04:38:29,176 INFO sqlalchemy.engine.base.Engine ()

In [215]: # Send pandas dataframe from redshift to sqlite3 for transformation
sqlite_table = 'homesales'
db_1.to_sql(sqlite_table, sqlite_connection, if_exists='fail')

2022-03-02 04:38:53,566 INFO sqlalchemy.engine.base.Engine PRAGMA main.table_info("homesales")
2022-03-02 04:38:53,568 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:53,570 INFO sqlalchemy.engine.base.Engine PRAGMA temp.table_info("homesales")
2022-03-02 04:38:53,572 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:53,579 INFO sqlalchemy.engine.base.Engine
CREATE TABLE homesales (
  "index" BIGINT,
  price BIGINT,
  bedrooms BIGINT,
  "propertyType" TEXT
)

2022-03-02 04:38:53,581 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:53,760 INFO sqlalchemy.engine.base.Engine COMMIT
2022-03-02 04:38:53,763 INFO sqlalchemy.engine.base.Engine CREATE INDEX ix_homesales_index ON homesal
es ("index")
2022-03-02 04:38:53,767 INFO sqlalchemy.engine.base.Engine ()
2022-03-02 04:38:53,777 INFO sqlalchemy.engine.base.Engine COMMIT
2022-03-02 04:38:53,793 INFO sqlalchemy.engine.base.Engine BEGIN (implicit)
2022-03-02 04:38:54,169 INFO sqlalchemy.engine.base.Engine INSERT INTO homesales ("index", price, bed
rooms, "propertyType") VALUES (?, ?, ?, ?)
2022-03-02 04:38:54,170 INFO sqlalchemy.engine.base.Engine ((0, 525000, 4, 'house'), (1, 290000, 3,
'house'), (2, 328000, 3, 'house'), (3, 380000, 4, 'house'), (4, 310000, 3, 'house'), (5, 465000, 4,
'house'), (6, 399000, 3, 'house'), (7, 1530000, 4, 'house')) ... displaying 10 of 29580 total bound p
arameter sets ... (29578, 589000, 2, 'unit'), (29579, 775000, 2, 'unit'))
2022-03-02 04:38:54,275 INFO sqlalchemy.engine.base.Engine COMMIT
```

```
In [239]: # Close the database connecton
sqlite_connection.close()

In [ ]:

In [ ]:

In [ ]:
```