Deliverable 2

GitHub Link:

https://github.com/saikrishnasanda/bigdata

Data Understanding:

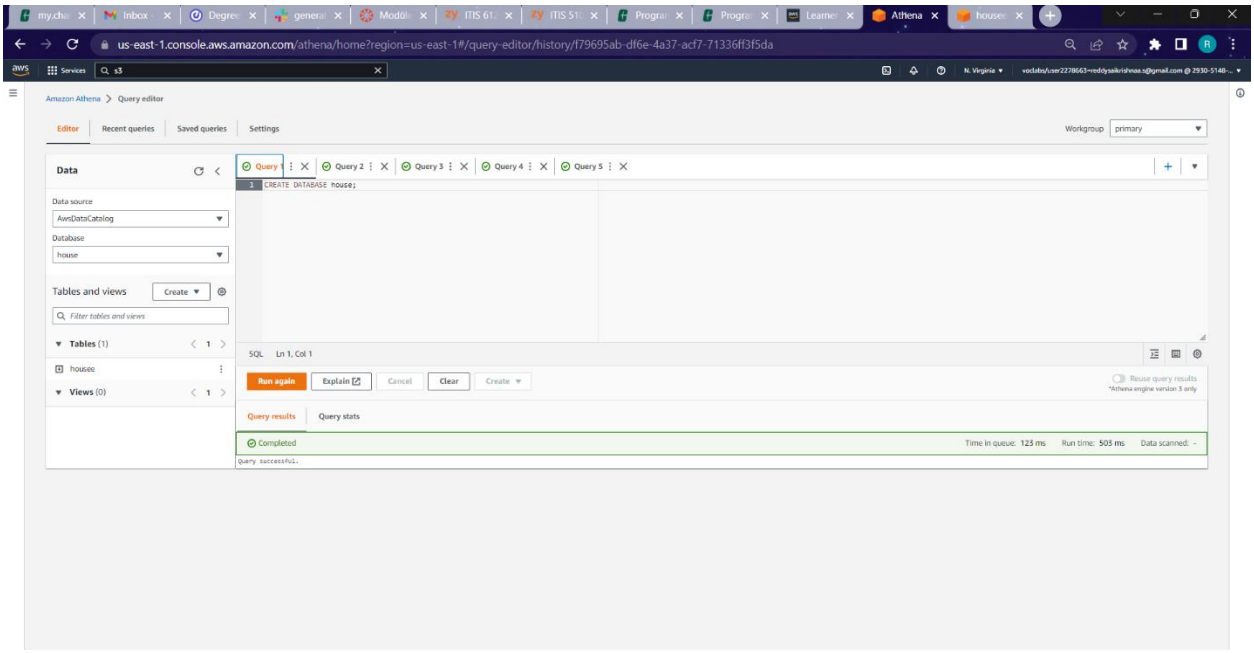
a) Exploratory Data Analysis

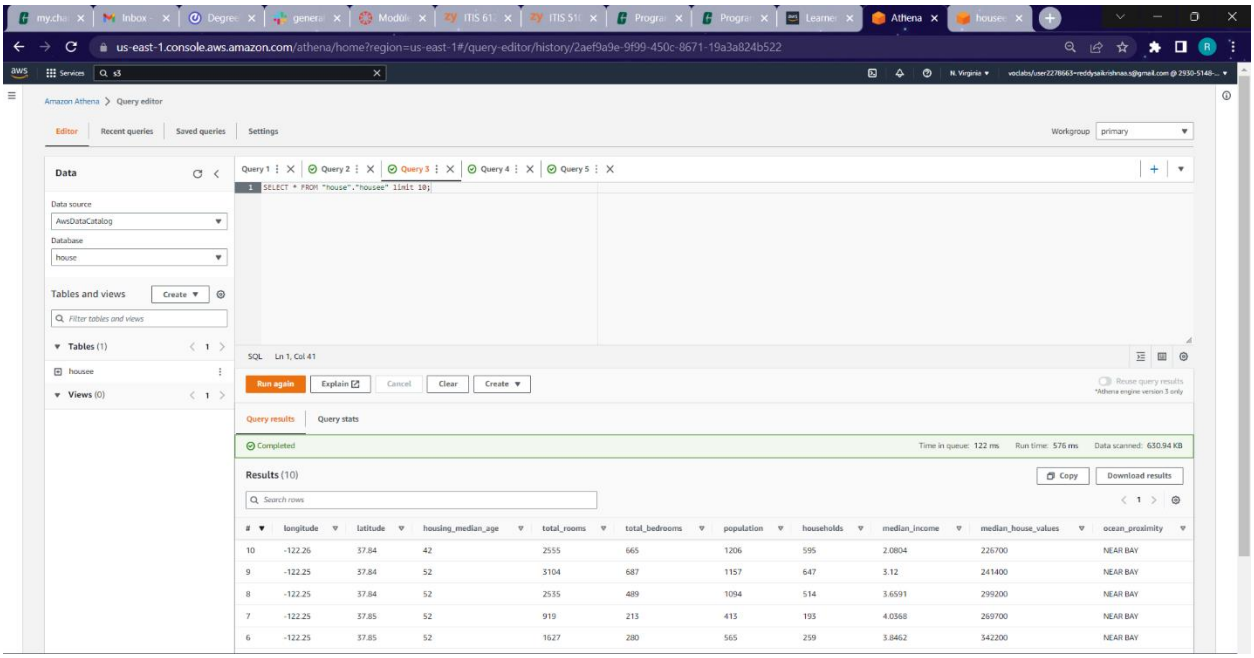
AWS Components Used:

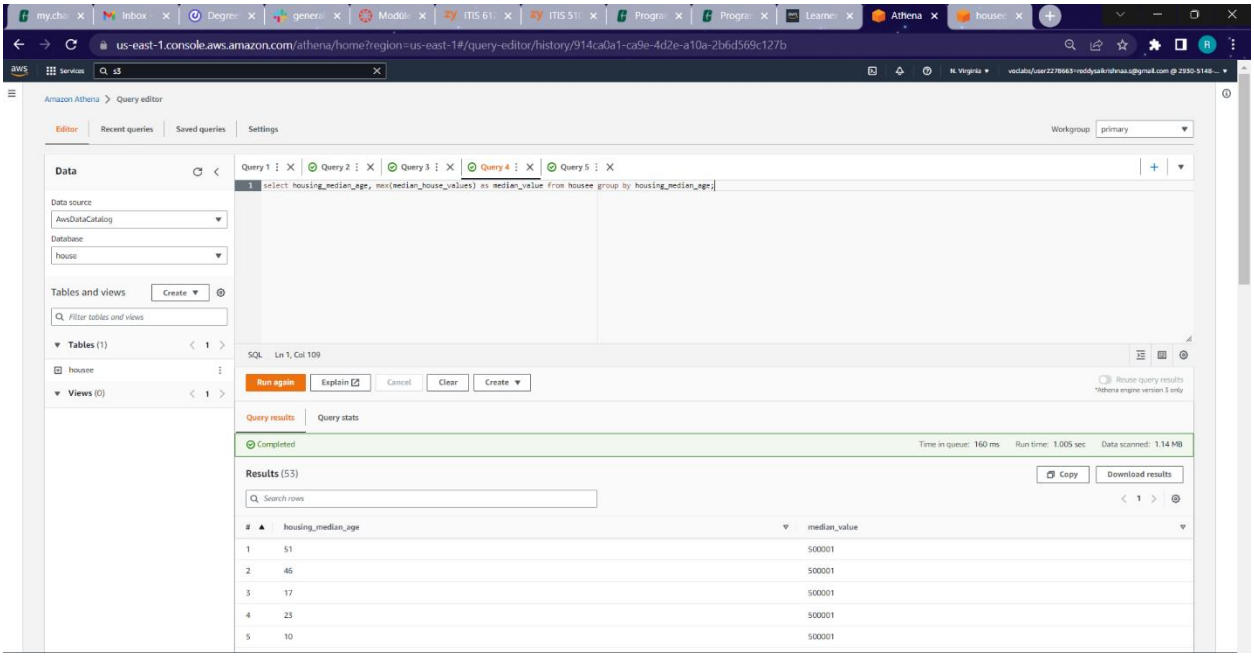
Amazon S3 Bucket

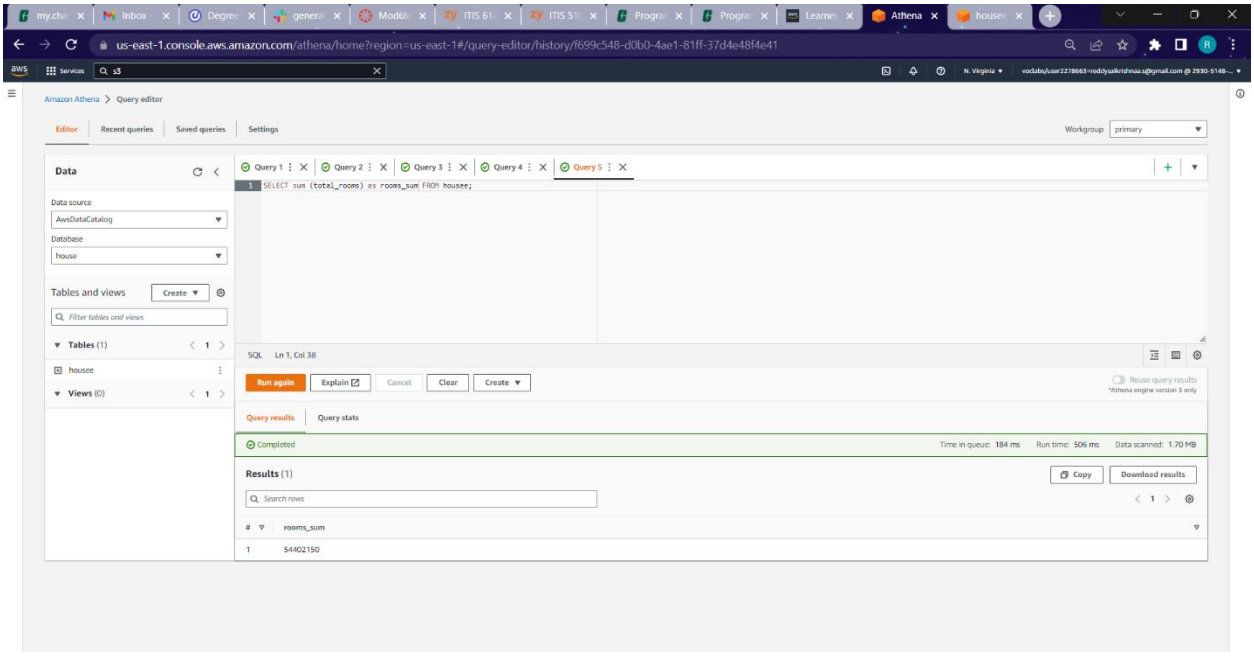
https://housee.s3.amazonaws.com/housing.csv

Amazon Athena:

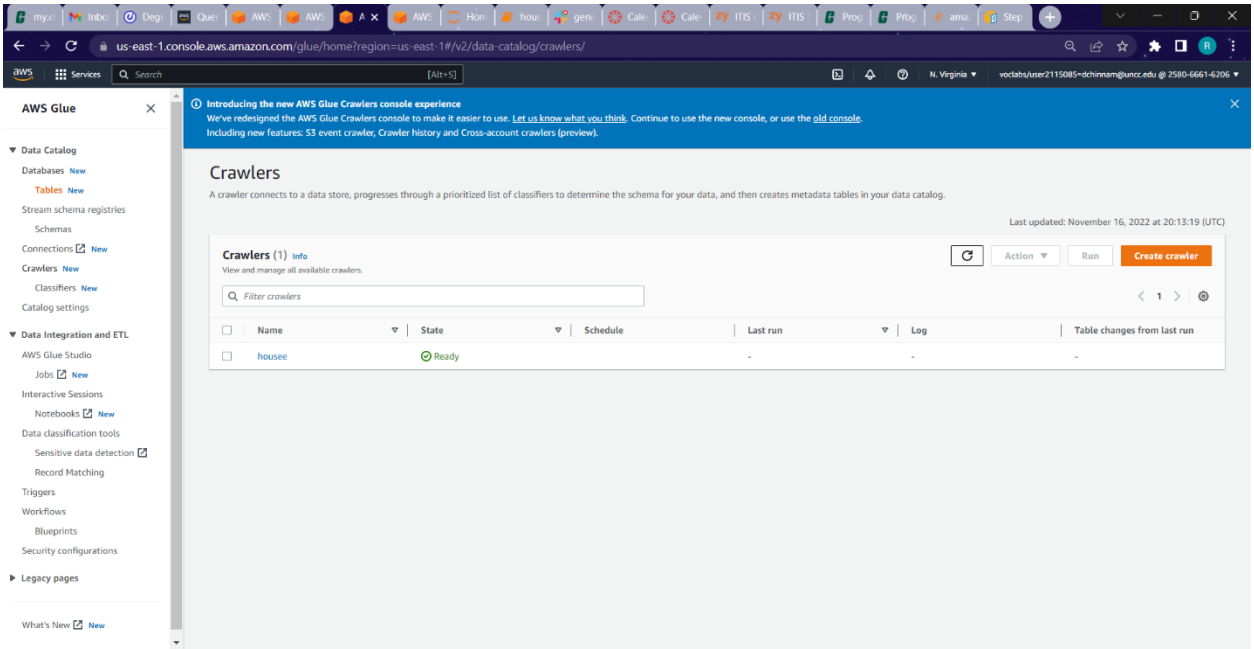


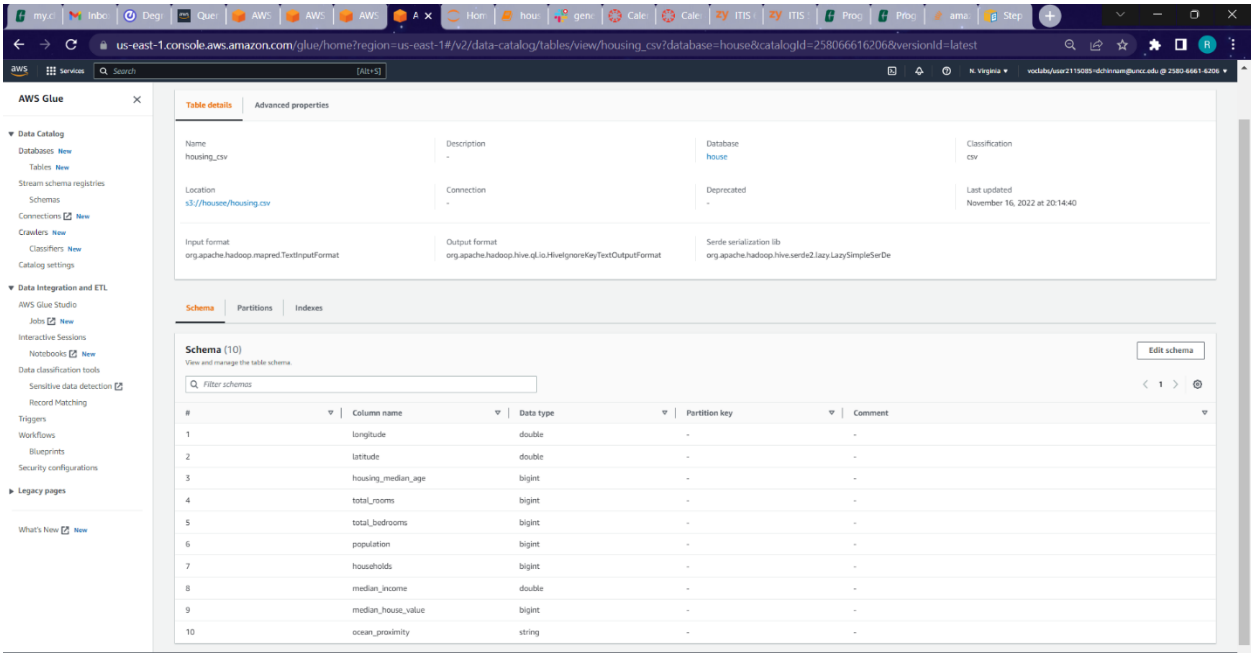

us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/2aef9a9e-9f99-450c-8671-19a3a824b522
Amazon Athena > Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup primary
Data
Data source
AwsDataCatalog
Database
house
Tables and views
Create
Filter tables and views
Tables (1)
housee
Views (0)
Query 1
Query 2
Query 3
Query 4
Query 5
1 SELECT * FROM "house"."housee" limit 10;
SQL Ln 1, Col 41
Run again
Explain
Cancel
Clear
Create
Reuse query results
*Athena engine version 3 only
Query results
Query stats
Completed
Time in queue: 122 ms
Run time: 576 ms
Data scanned: 630.94 KB
Results (10)
Copy
Download results
Search rows
longitude latitude housing_median_age total_rooms total_bedrooms population households median_income median_house_values ocean_proximity
10 -122.26 37.84 42 2555 665 1206 595 2.0804 226700 NEAR BAY
9 -122.25 37.84 52 3104 687 1157 647 3.12 241400 NEAR BAY
8 -122.25 37.84 52 2535 489 1094 514 3.6591 299200 NEAR BAY
7 -122.25 37.85 52 919 213 413 193 4.0368 269700 NEAR BAY
6 -122.25 37.85 52 1627 280 565 259 3.8462 342200 NEAR BAY


us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/914ca0a1-ca9e-4d2e-a10a-2b6d569c127b
Amazon Athena > Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup primary
Data
Data source
AwsDataCatalog
Database
house
Tables and views
Create
Filter tables and views
Tables (1)
housee
Views (0)
Query 1
Query 2
Query 3
Query 4
Query 5
1 select housing_median_age, max(median_house_values) as median_value from housee group by housing_median_age;
SQL Ln 1, Col 109
Run again
Explain
Cancel
Clear
Create
Reuse query results
*Athena engine version 3 only
Query results
Query stats
Completed
Time in queue: 160 ms
Run time: 1.005 sec
Data scanned: 1.14 MB
Results (53)
Copy
Download results
Search rows
housing_median_age median_value
1 51 500001
2 46 500001
3 17 500001
4 23 500001
5 10 500001


Athena
us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/f699c548-d0b0-4ae1-81ff-37d4e48f4e41
Amazon Athena > Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup primary
Data
Data source
AwsDataCatalog
Database
house
Tables and views
Create
Tables (1)
housee
Views (0)
Query 1
Query 2
Query 3
Query 4
Query 5
1 SELECT sum (total_rooms) as rooms_sum FROM housee;
SQL Ln 1, Col 38
Run again
Explain
Cancel
Clear
Create
Query results
Query stats
Completed
Time in queue: 184 ms
Run time: 506 ms
Data scanned: 1.70 MB
Results (1)
Copy
Download results
Search rows
rooms_sum
1 54402150

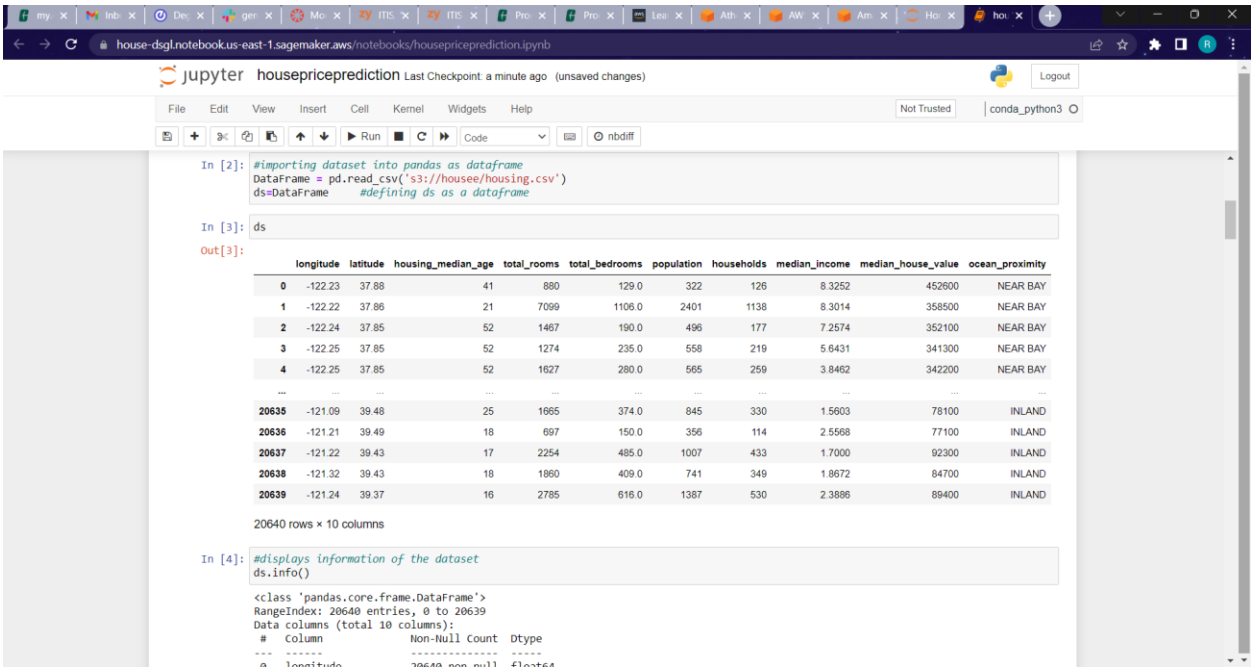
Amazon Glue:


us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/crawlers/
AWS Glue
Data Catalog
Databases New
Tables New
Stream schema registries
Schemas
Connections New
Crawlers New
Classifiers New
Catalog settings
Data Integration and ETL
AWS Glue Studio
Jobs New
Interactive Sessions
Notebooks New
Data classification tools
Sensitive data detection
Record Matching
Triggers
Workflows
Blueprints
Security configurations
Legacy pages
What's New New
Introducing the new AWS Glue Crawlers console experience
We've redesigned the AWS Glue Crawlers console to make it easier to use. Let us know what you think. Continue to use the new console, or use the old console.
Including new features: S3 event crawler, Crawler history and Cross-account crawlers (preview).
Crawlers
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.
Last updated: November 16, 2022 at 20:13:19 (UTC)
Crawlers (1) Info
View and manage all available crawlers.
Action
Run
Create crawler
Filter crawlers
Name
State
Schedule
Last run
Log
Table changes from last run
housee
Ready



Amazon SageMaker:

Dashboards:



house-dsgl.notebook.us-east-1.sagemaker.aws/notebooks/housepriceprediction.ipynb

jupyter housepriceprediction Last Checkpoint: a minute ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | conda_python3

Run Code nbdiff

In [5]:
```
# display top 5 rows in the dataset
ds.head()
```

Out[5]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41	880	129.0	322	126	8.3252	452600	NEAR BAY
1	-122.22	37.86	21	7099	1106.0	2401	1138	8.3014	358500	NEAR BAY
2	-122.24	37.85	52	1467	190.0	496	177	7.2574	352100	NEAR BAY
3	-122.25	37.85	52	1274	235.0	558	219	5.6431	341300	NEAR BAY
4	-122.25	37.85	52	1627	280.0	565	259	3.8462	342200	NEAR BAY

In [6]:
```
# display last 5 rows in the dataset
ds.tail()
```

Out[6]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
20635	-121.09	39.48	25	1665	374.0	845	330	1.5603	78100	INLAND
20636	-121.21	39.49	18	697	150.0	356	114	2.5568	77100	INLAND
20637	-121.22	39.43	17	2254	485.0	1007	433	1.7000	92300	INLAND
20638	-121.32	39.43	18	1860	409.0	741	349	1.8672	84700	INLAND
20639	-121.24	39.37	16	2785	616.0	1387	530	2.3886	89400	INLAND

In [7]:
```
# display the dimentions of the dataset
ds.shape
```

Out[7]: (20640, 10)

In [8]:
```
#displays the columns in the dataset
list(ds.columns)
```

Out[8]: ['longitude',

house-dsgl.notebook.us-east-1.sagemaker.aws/notebooks/housepriceprediction.ipynb

jupyter housepriceprediction Last Checkpoint: 2 minutes ago (unsaved changes)

Logout

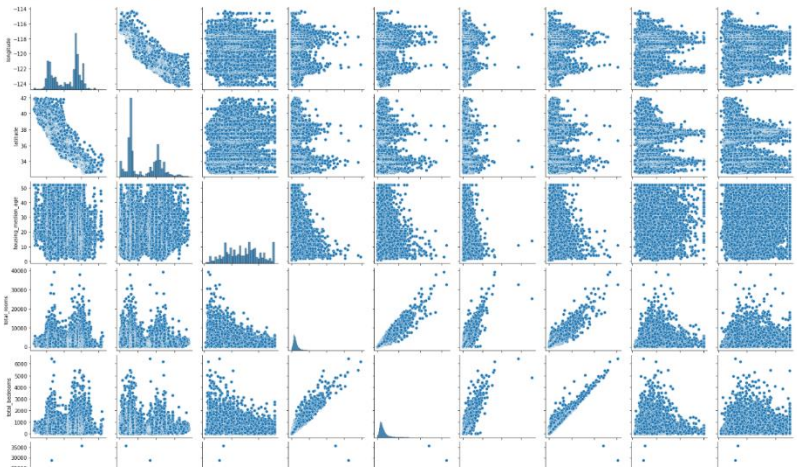
File Edit View Insert Cell Kernel Widgets Help

Not Trusted | conda_python3

Run Code nbdiff

In [13]:
```
sns.pairplot(ds, diag_kind="hist")
```

Out[13]: <seaborn.axisgrid.PairGrid at 0x7f839523dc10>

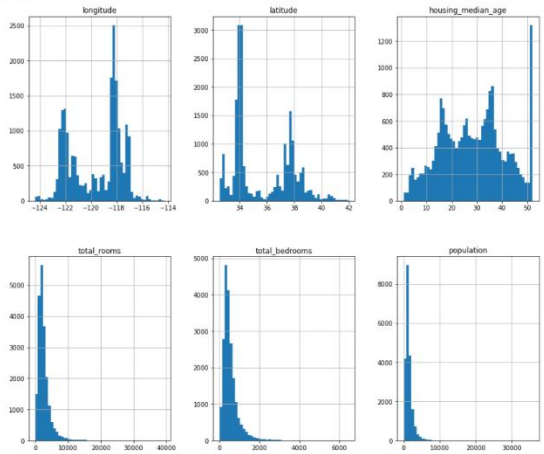


house-dsgl.notebook.us-east-1.sagemaker.aws/notebooks/housepriceprediction.ipynb

jupyter housepriceprediction Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [14]:
```
%matplotlib inline
import matplotlib.pyplot as plt
ds.hist(bins=50,figsize=(15,20))
plt.show()
```

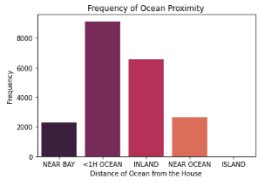


house-dsgl.notebook.us-east-1.sagemaker.aws/notebooks/housepriceprediction.ipynb

jupyter housepriceprediction Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [15]:
```
from matplotlib import pyplot as plt
sns.countplot(data=ds, x="ocean_proximity",  palette="rocket").set(
    title = "Frequency of Ocean Proximity",
    xlabel = "Distance of Ocean from the House",
    ylabel = "Frequency"
)
plt.figure(figsize=(25,15))
plt.show()
```




```
<Figure size 1800x1080 with 0 Axes>
```

Data Preparation

Converting Unstructured Data into Structured Data

In [16]:
```
# displays missing values in the dataset
missing_values_ds = ds.isnull().sum()
print(missing_values_ds)
```

```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms      207
population            0
households            0
```

house-dsgl.notebook.us-east-1.sagemaker.aws/notebooks/housepriceprediction.ipynb

jupyter housepriceprediction Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Data Preparation

Converting Unstructured Data into Structured Data

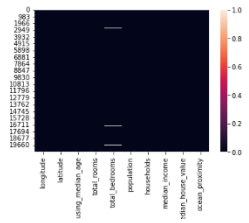
In [16]:
```
# displays missing values in the dataset
missing_values_ds = ds.isnull().sum()
print(missing_values_ds)
```

```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms      207
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

total_bedrooms is the only Column to have missing values in our dataset

In [17]:
```
sns.heatmap(ds.isnull())
```

Out[17]: <AxesSubplot:>



house-dsgl.notebook.us-east-1.sagemaker.aws/notebooks/housepriceprediction.ipynb

jupyter housepriceprediction Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [18]:
```
ds = ds[ds['total_bedrooms'].notna()]
```

In [19]:
```
# displays missing values after dropping
ds.isnull().sum()
```

Out[19]:
```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms        0
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

In [20]:
```
# displays duplicate values in the dataset
ds.duplicated().sum()
```

Out[20]: 0

No duplicate values found in the dataset

In [21]:
```
# displays heatmap for null values
sns.heatmap(ds.isnull())
```

Out[21]: <AxesSubplot:>