

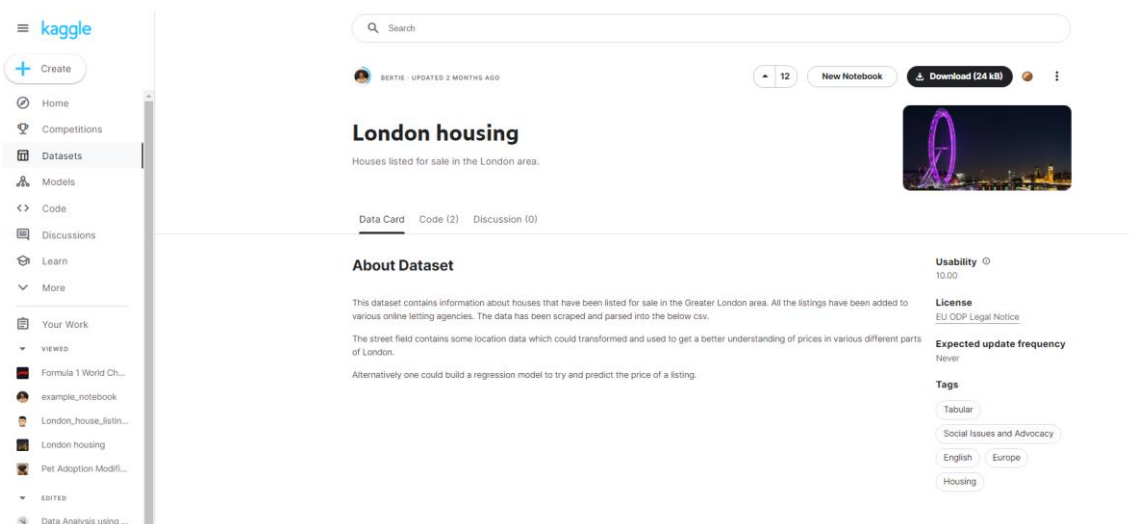
MVP 3 – Engenharia de Dados

Aluno: João Victor Barbosa de Araújo

Objetivo (Perguntas)

- Qual o total de casas a venda?
- Qual a região com mais casas a venda?
- Qual o preço médio das casas das Top 5 ruas mais caras? E as mais baratas?
- Qual a contagem de casas pelo número de quartos?
- Qual a contagem de casas pelo número de banheiros?
- Qual a contagem do tipo de posse da propriedade?
- Quantas casas possuem jardins? E quantas não possuem?
- Qual a maior casa a venda?

1. Busca pelos dados



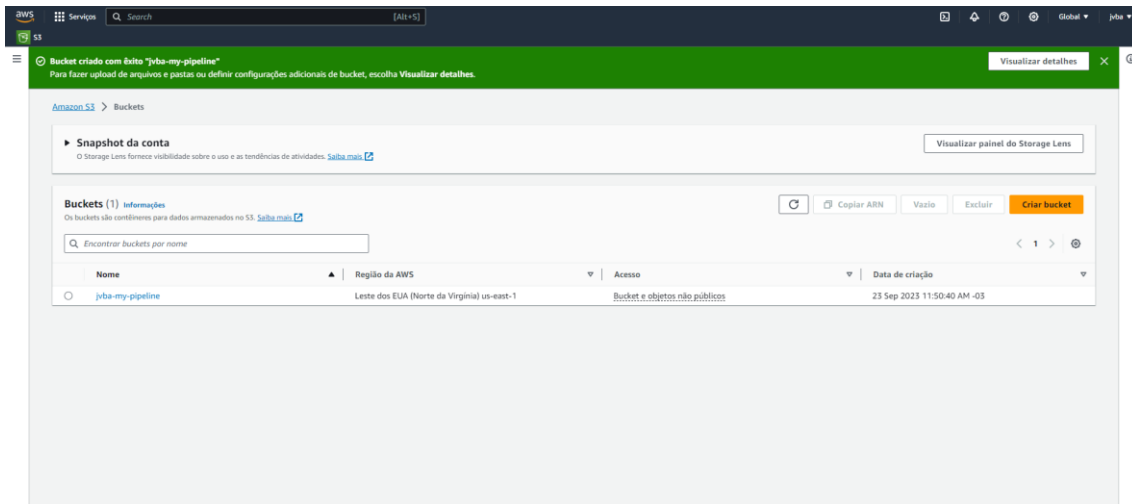
O Banco de Dados que será utilizado será o “London Housing”. Este conjunto de dados contém informações sobre casas listadas à venda na área metropolitana de Londres.

Link do Dataset: <https://www.kaggle.com/datasets/bertiemackie/london-house-listings>

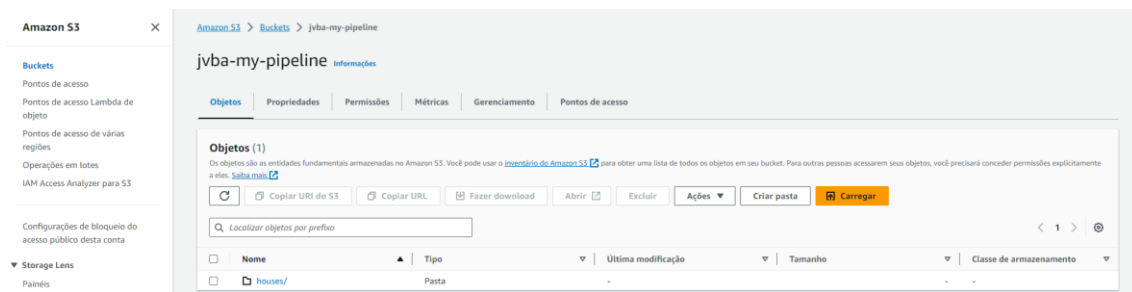
O Data Warehouse que será utilizado nesse trabalho será a AWS.

2. Coleta

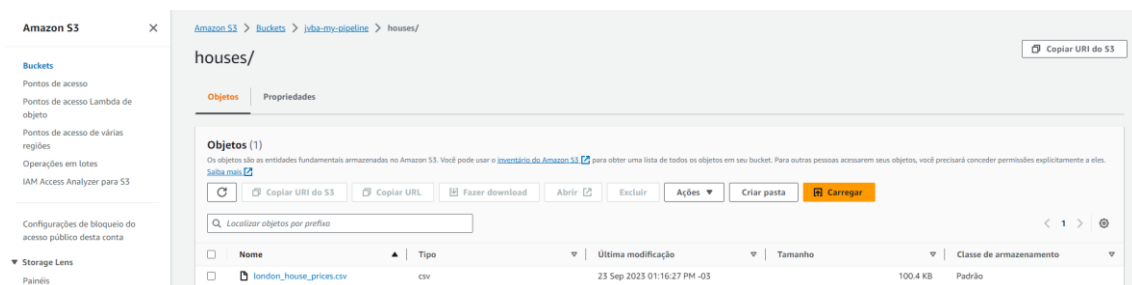
Passo 1: Criar o Bucket no S3



Passo 2: Criar pasta



Passo 3: Dar upload do arquivo



3. Modelagem

Nome da tabela: data_houses

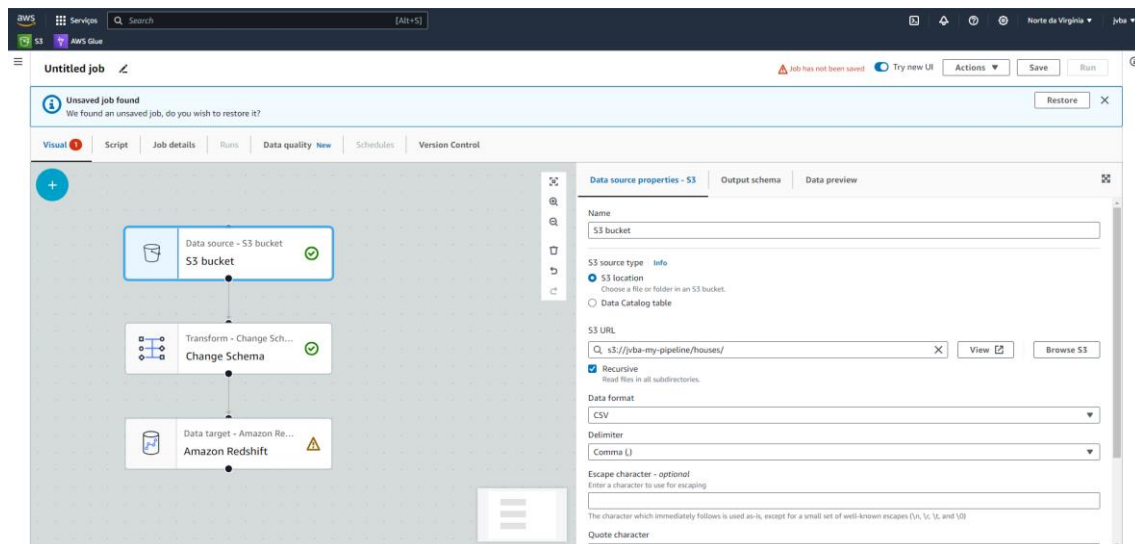
- id (INT): Identificador único para cada casa.
- street (VARCHAR): Nome da rua onde a casa está situada.
- bedrooms (STRING): Número de quartos na casa.
- bathrooms (STRING): Número de banheiros na casa.
- tenure (VARCHAR): Tipo de posse da propriedade (por exemplo, freehold, leasehold, etc.).

- garden (BOOLEAN): Indica se a casa possui um jardim (verdadeiro/falso).
- size (STRING): Tamanho da casa.
- price (VARCHAR): Preço da casa.
- nearest_station_name (VARCHAR): Nome da estação de trem mais próxima.
- nearest_station_miles (FLOAT): Distância até a estação de trem mais próxima em milhas.
- postcode_outer (VARCHAR): Código postal da área externa.

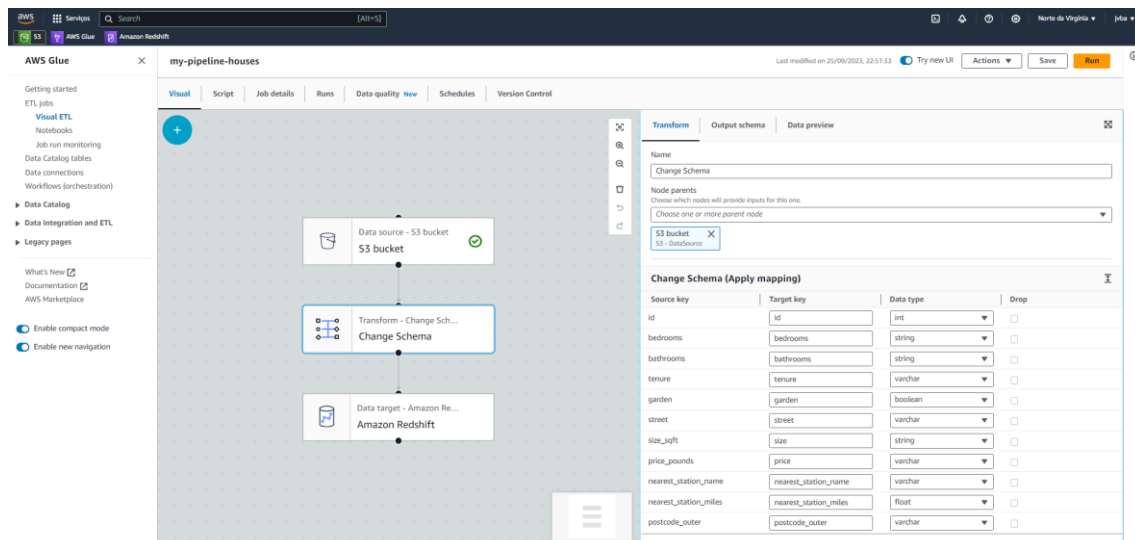
4. Carga

Agora é hora de criar um estúdio para o ETL usando o AWS Glue Studio

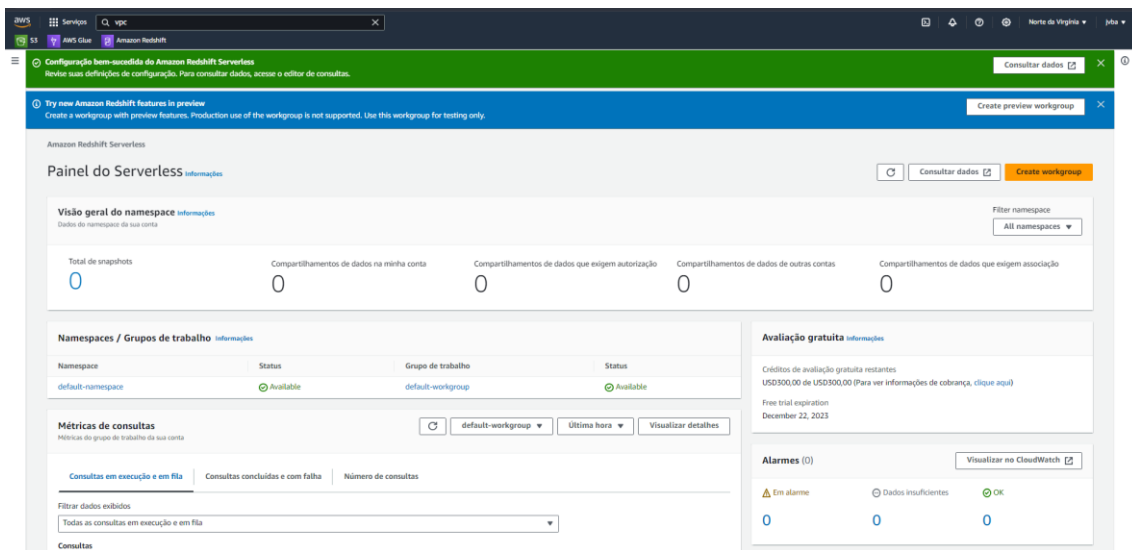
Na etapa 1, utilizei o S3 bucket para fazer a extração dos dataset



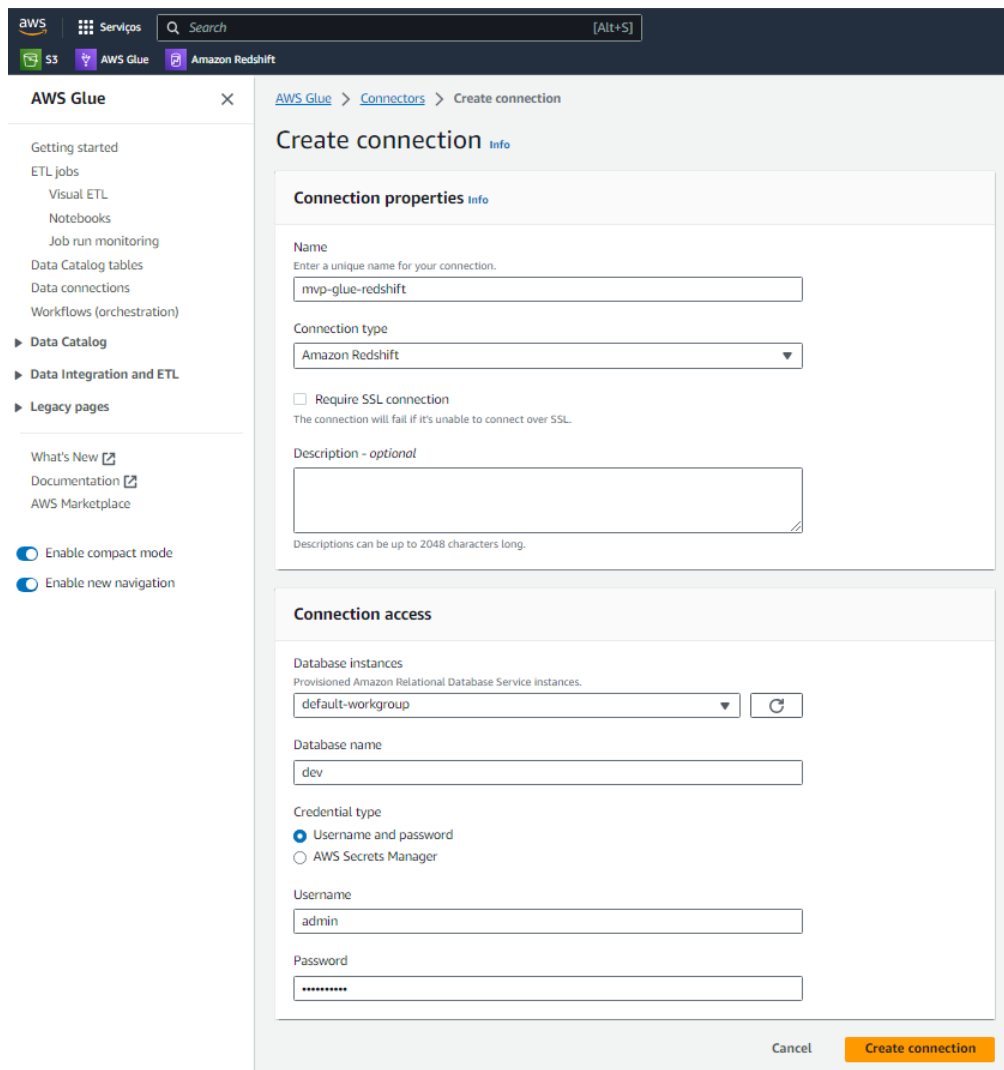
Já na etapa 2, realizei a Transformação dos dados

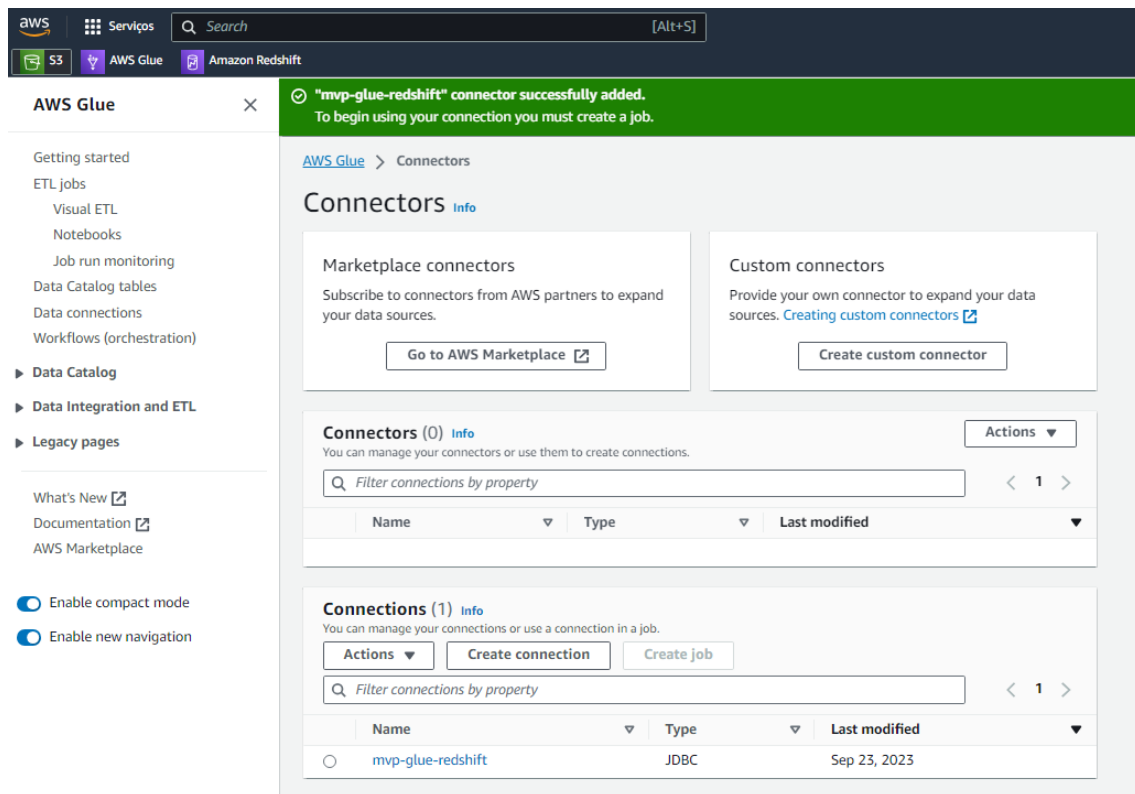


Na etapa 3, é necessário primeiro criar o Amazon Redshift Serveless

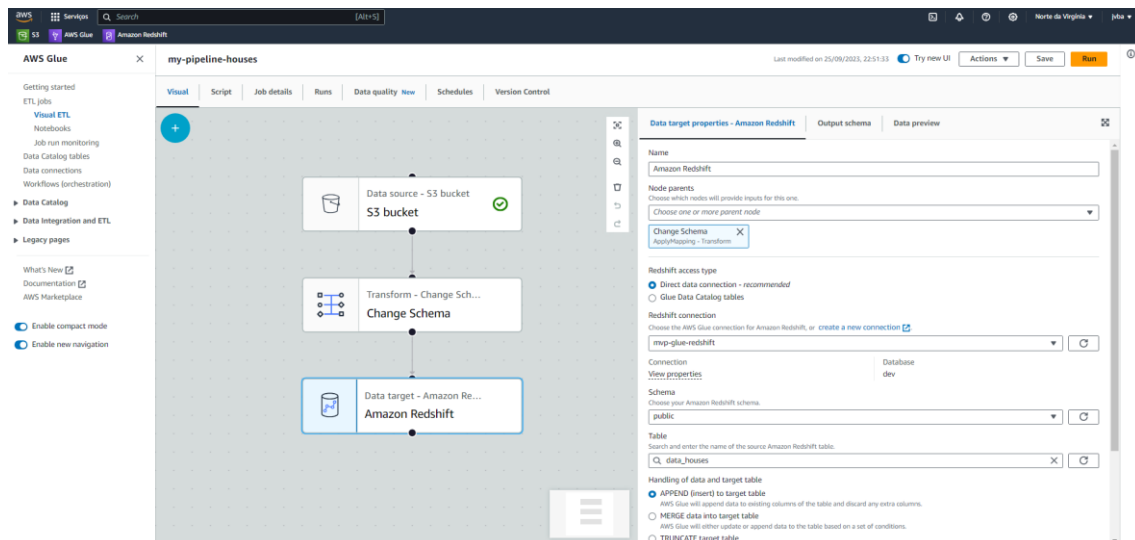


A próxima etapa 4 é a conexão do AWS Glue com o Redshift

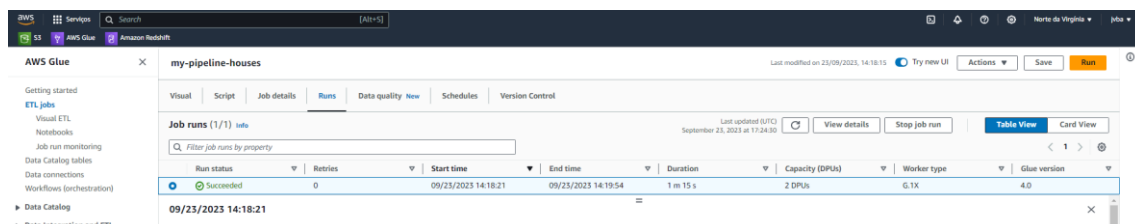




Na última etapa, registrei a execução dos Jobs



Após todas essas etapas, foi dado o Run do job e com status de “Sucesso”



Como Pode ser visto, conseguimos carregar o dataset no Redshift

The screenshot shows the Amazon Redshift Query Editor v2 interface. The left sidebar displays the database structure with a tree view showing 'sample_data_dev' and 'public' schemas. The main area shows a SQL query: `SELECT * FROM data_houses`. The results are displayed in a table with columns: id, postcode_out, price, nearest_station_miles, size, bedrooms, tenure, street, and nearest_station. The table contains 10 rows of data.

id	postcode_out	price	nearest_station_miles	size	bedrooms	tenure	street	nearest_station
132180206		10500000	0.2		4.0	freehold	Ladbroke Grove, London	Holland Park St
134996630	SW19	8950000	0.3		5.0	freehold	Murray Road, Wembley	Wembley Sta
134169233	SW19	11950000	0.7		5.0	freehold	Southside Common, Wils...	Wembley Sta
132180206		10500000	0.2		4.0	freehold	Ladbroke Grove, London	Holland Park St
134996630	SW19	8950000	0.3		5.0	freehold	Murray Road, Wembley	Wembley Sta
134169233	SW19	11950000	0.7		5.0	freehold	Southside Common, Wils...	Wembley Sta
124365635	W1J	10950000	0.3	4402.0	6.0	freehold	Chesterfield Hill, London	Green Park Sta
132654419	W8	18500000	0.3		4.0	freehold	Cottlemore Gardens, Len...	High Street Kbr
125741842	W18	8950000	0		3.0	freehold	The Regents Crescent, P...	Regent's Park S
126980469	SW1X	11750000	0.3	4563.0	6.0	freehold	Wilton Street, Belgravia	Hyde Park Cor
136488719	NW7	20000000	1.1		5.0	freehold	Highwood Lodge Farm E...	Mill Hill Broadw
134329288	SW1X	43700000	0.1		7.0	leasehold	Belgrave Gate, Grosvenor	Hyde Park Cor
132180206		10500000	0.2		4.0	freehold	Ladbroke Grove, London	Holland Park St
134996630	SW19	8950000	0.3		5.0	freehold	Murray Road, Wembley	Wembley Sta

5. Análise

- Qual o total de casas a venda?

The screenshot shows the Amazon Redshift Query Editor v2 interface. The SQL query is: `SELECT COUNT(*) as total_casas_a_venda FROM data_houses;`. The results are displayed in a table with one row: `total_casas_a_venda` with the value `934`.

total_casas_a_venda
934

Depois dessa pesquisa, concluímos que de acordo com esse dataset, que existem 934 casas a vendas em Londres.

- Qual a rua com mais casas a venda?

Run Limit 100 Explain Isolated session

```

1 SELECT
2   street,
3   COUNT(*) as casas_a_venda
4 FROM data_houses
5 GROUP BY street
6 ORDER BY casas_a_venda DESC;

```

Result 1 (100)

street	casas_a_venda
Winnington Road, N2	8
Moxon Street, London, W...	8
Moxon Street, W1U	7
St. Mary Abbots Place, Lo...	5
Grosvenor Square, Mayfa...	5
Cumberland Terrace, Reg...	5
Canary Wharf,	5
Chesham Street, SW1X	5
Lowndes Square, London...	5
Old Church Street, Londo...	5
Damac Tower, Nine Elms,...	5
Cambridge Gate, Regent'	4
Damac Tower, Nine Elms,...	4
Southbank Place, Belved...	4

Nessa query, listamos em ordem decrescente, o total de casas pelas ruas de Londres. A “Winnington Road” lidera junto com a “Moxon Street” com 8 casas a venda cada um.

- Qual o preço médio das casas das 5 ruas mais caras? E as mais baratas?

Top 5 mais caras:

Run Limit 100 Explain Isolated session

```

1 SELECT
2   street,
3   AVG(price) as preco_medio
4 FROM data_houses
5 GROUP BY street
6 ORDER BY preco_medio DESC
7 LIMIT 5;

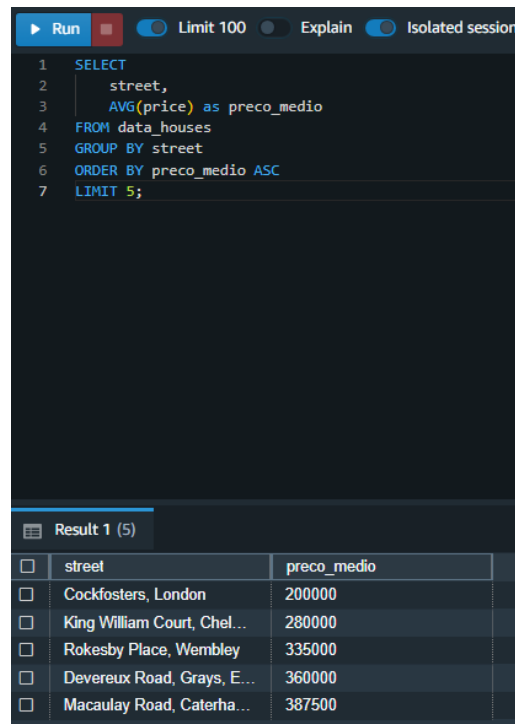
```

Result 1 (5)

street	preco_medio
Denham Place, Denham, ...	75000000
Denham Place, Denham, ...	75000000
Lygon Place, Belgravia, S...	45000000
Pitt Street, London, W8	44000000
Pitt Street, Kensington, W8	44000000

Analizamos com essa pesquisa que a “Denham Place” possui a média de casa mais cara de Londres com £75.000.000.

Top 5 mais baratas:

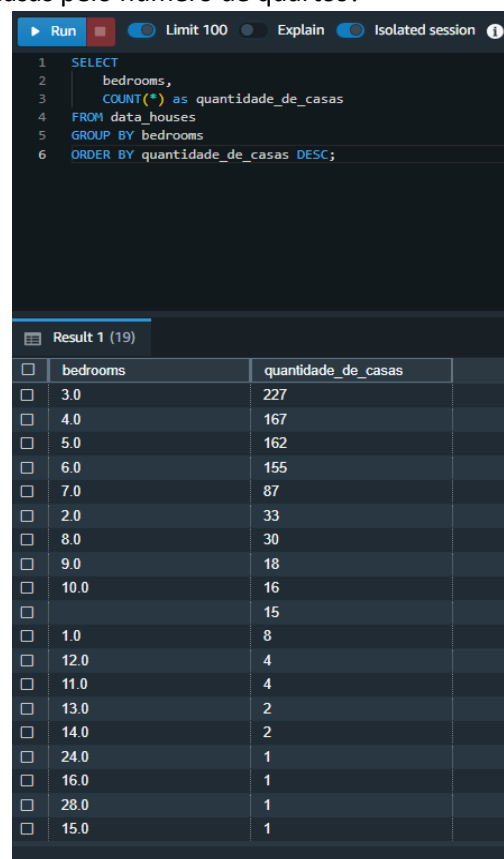


The screenshot shows a SQL query in a dark-themed editor. The query is: `SELECT street, AVG(price) as preco_medio FROM data_houses GROUP BY street ORDER BY preco_medio ASC LIMIT 5;`. Below the query, the results are displayed in a table with two columns: 'street' and 'preco_medio'. The results are ordered from lowest to highest average price.

street	preco_medio
Cockfosters, London	200000
King William Court, Chel...	280000
Rokesby Place, Wembley	335000
Devereux Road, Grays, E...	360000
Macaulay Road, Caterha...	387500

Já a mais barata é a “Cockfoster Street” com o preço médio de £200.000.

- Qual a contagem de casas pelo número de quartos?

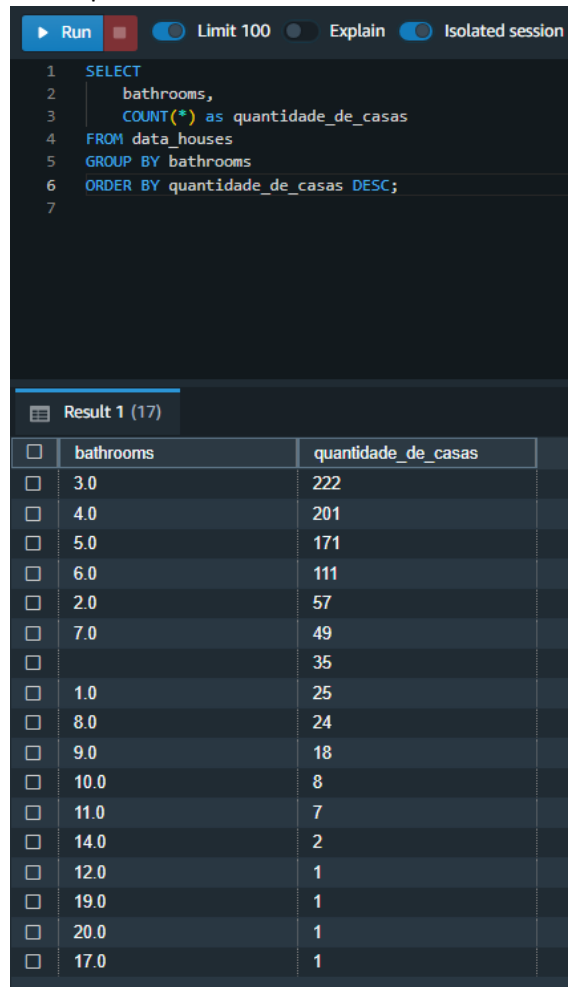


The screenshot shows a SQL query in a dark-themed editor. The query is: `SELECT bedrooms, COUNT(*) as quantidade_de_casas FROM data_houses GROUP BY bedrooms ORDER BY quantidade_de_casas DESC;`. Below the query, the results are displayed in a table with two columns: 'bedrooms' and 'quantidade_de_casas'. The results are ordered from highest to lowest count.

bedrooms	quantidade_de_casas
3.0	227
4.0	167
5.0	162
6.0	155
7.0	87
2.0	33
8.0	30
9.0	18
10.0	16
1.0	15
1.0	8
12.0	4
11.0	4
13.0	2
14.0	2
24.0	1
16.0	1
28.0	1
15.0	1

As casas com 3 quartos são as que lideram o ranking com 227 casas a venda, em seguida de 4 quartos com 167 casas e depois 5 quartos com 162 residências. Podemos analisar que o tipo de casa que as pessoas mais almejam são a de 1 e 2 quartos, já que são as que tem menos ofertas. Podemos decifrar que existem varias casas com 1 ou 2 quartos, porém a maioria já se encontram ocupadas, e não é de muito interesse dos britânicos possuírem uma casa com 3 ou mais quartos.

- Qual a contagem de casas pelo número de banheiros?



The screenshot shows a SQL query execution interface. At the top, there are buttons for 'Run', 'Limit 100', 'Explain', and 'Isolated session'. Below these, the SQL query is displayed in a dark-themed editor:

```
1 SELECT
2     bathrooms,
3     COUNT(*) as quantidade_de_casas
4 FROM data_houses
5 GROUP BY bathrooms
6 ORDER BY quantidade_de_casas DESC;
7
```

Below the query editor, the results are shown in a table titled 'Result 1 (17)'. The table has two columns: 'bathrooms' and 'quantidade_de_casas'. The data is sorted in descending order of the number of houses.

bathrooms	quantidade_de_casas
3.0	222
4.0	201
5.0	171
6.0	111
2.0	57
7.0	49
	35
1.0	25
8.0	24
9.0	18
10.0	8
11.0	7
14.0	2
12.0	1
19.0	1
20.0	1
17.0	1

A análise sobre os banheiros segue a mesma lógica que a dos quartos. Podemos imaginar que o padrão das casas de Londres são 1 banheiro para cada quarto. Seguindo essa lógica, podemos fazer a mesma análise passada, onde as pessoas não possuem muita preferencia por casas com 3 quartos ou mais, consequentemente as casas com mais de 3 banheiros são as que mais estão a venda. Até se olharmos os números, são muito semelhantes

- Qual a contagem do tipo de posse da propriedade?

▶ Run	Limit 100	Explain	Isolated session
-------	-----------	---------	------------------

```

1 SELECT
2   tenure,
3   COUNT(*) as quantidade
4 FROM data_houses
5 GROUP BY tenure;

```

Result 1 (5)		
<input type="checkbox"/>	tenure	quantidade
<input type="checkbox"/>	freehold	438
<input type="checkbox"/>	leasehold	372
<input type="checkbox"/>	share of freehold	58
<input type="checkbox"/>	ask agent	54
<input type="checkbox"/>		12

Freehold (438): O termo "freehold" indica que o proprietário tem posse total da propriedade, incluindo o terreno sobre o qual ela está construída. Isso é geralmente considerado uma forma mais desejável de posse, pois dá ao proprietário mais controle e liberdade sobre a propriedade.

Leasehold (372): No caso de "leasehold", o proprietário possui a propriedade apenas por um período específico, conforme especificado em um contrato de arrendamento. Após o término do contrato, a propriedade retorna ao locador.

Share of Freehold (58): Esta é uma situação em que os proprietários de apartamentos em um edifício também compartilham a propriedade do terreno e partes comuns do edifício.

Ask Agent (54): "Ask agent" pode ser uma categoria especializada ou uma instrução para entrar em contato com um agente imobiliário para obter informações adicionais.

- Quantas casas possuem jardins? E quantas não possuem?

```
▶ Run Limit 100 Explain Isolated session
1 SELECT
2   garden,
3   COUNT(*) as quantidade_de_casas
4 FROM data_houses
5 GROUP BY garden
6 ORDER BY garden;
```

Result 1 (3)

garden	quantidade_de_casas
false	392
true	402
NULL	140

Com a query vemos que a maioria das casas de Londres possuem algum jardim em sua estrutura, mas fica bem parelho com as que não possuem.

- Qual a maior casa a venda?

```
▶ Run Limit 100
1 SELECT
2   *
3 FROM data_houses
4 ORDER BY size DESC
5 LIMIT 1;
```

id	postcode_out	price	nearest_station_miles	size	bathrooms	tenure	street	nearest_station_name	garden	bedrooms
84082359	N6	19500000	0.8	9858.0	3.0	ask agent	Guildens, Courtenay Ave...	East Finchley Station	true	6.0

Aqui está a maior casa do dataset. Mesmo ela sendo a maior, não possui o maior preço, custa £19.500.000. Interessante também que ela possui apenas 3 banheiros e 6 quartos, então deve ser uma casa bem espaçosa e com um jardim enorme. Bom também é que se precisar utilizar o metrô para se deslocar, precisará caminhar menos de uma milha.

Autoavaliação

Foi uma saga bem legal e desafiadora até o final dessa pós-graduação, foi um desafio exaustivo do início até o fim, mas agradeço por me ensinarem muito durante esses meses.

O início do MVP pareceu muito trabalhoso, vendo primeiro as lives de dúvidas sobre o trabalho, me subiu um desespero, mas quando coloquei na prática, ficou mais fácil. Comecei utilizando o Google Cloud, mas no momento final de subir a carga sempre me dava problema, então decidir rever a aula de AWS e seguir com esse DW. Vendo as aulas vi como era uma ferramenta bem mais completa, mas com suas dificuldades, um exemplo foi o VPC e o teste de conexão, bem complicado, mas felizmente consegui tirar de letra.

Já focado no Redshift e meus objetivos, sinto que consegui extrair tudo e responder o que eu estava em mente com possíveis perguntas para o dataset escolhido, então sinto que foi uma ótima conclusão de trabalho e uma ótima conclusão de graduação.