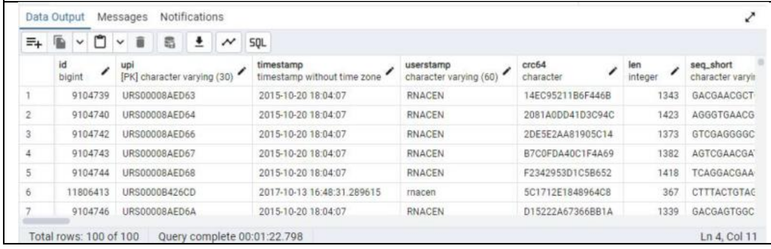
Uzair Khan

BWT - Data Engineering

Task 02:

1. **Write a query to get data having length of Rna structures more than 12 with them being added after 2008.**

```
SELECT * FROM rna
WHERE len > 12
AND timestamp > ‘2008-12-31’
LIMIT 100;
```

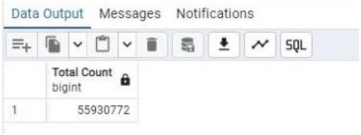


	id	upi	timestamp	userstamp	crc64	len	seq_short
	bigint	[PK] character varying (30)	timestamp without time zone	character varying (60)	character	integer	character varyi
1	9104739	URS00008AED63	2015-10-20 18:04:07	RNACEN	14EC95211B6F446B	1343	GACGAACGCT
2	9104740	URS00008AED64	2015-10-20 18:04:07	RNACEN	2081A0DD41D3C94C	1423	AGGGTGAACG
3	9104742	URS00008AED66	2015-10-20 18:04:07	RNACEN	2DE5E2AA81905C14	1373	GTCGAGGGGC
4	9104743	URS00008AED67	2015-10-20 18:04:07	RNACEN	B7C0FDA40C1F4A69	1382	AGTCGAACGA
5	9104744	URS00008AED68	2015-10-20 18:04:07	RNACEN	F2342953D1C5B652	1418	TCAGGACGAA
6	11806413	URS0000B426CD	2017-10-13 16:48:31.289615	rnacen	5C1712E1848964C8	367	CTTTACTGTAC
7	9104746	URS00008AED6A	2015-10-20 18:04:07	RNACEN	D15222A67366BB1A	1339	GACGAGTGGC

Total rows: 100 of 100 Query complete 00:01:22.798 Ln 4, Col 11

2. **How many pre computed RNA are present that are still active and got their last release update before 2022:**

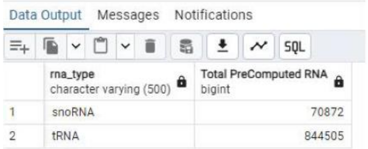
```
SELECT COUNT(*)
FROM rnc_rna_precomputed
WHERE is_active = 1
AND updated_date < '2022-01-01';
```



	Total Count
	bigint
1	55930772

3. **How many total pre computed RNA records for snoRNA and tRNA were recorded in 2011, 2016, 2014, and 2020:**

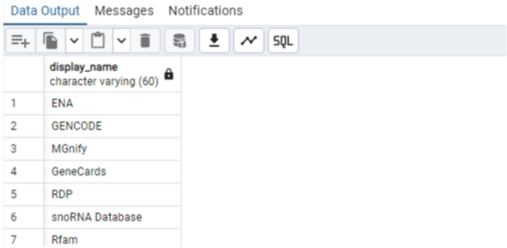
```
SELECT EXTRACT (YEAR FROM update_date) AS year, rna_type, COUNT (*) FROM
rnc_rna_precomputed WHERE rna_type IN('snoRNA','tRNA') AND EXTRACT (YEAR FROM
update_date) IN (2011, 2014, 2016, 2020) GROUP BY year, rna_type
ORDER BY year;
```



	rna_type	Total PreComputed RNA
	character varying (500)	bigint
1	snoRNA	70872
2	tRNA	844505

4. **Can you give me the names of all databases built for RNA with minimum length other than 100, 200, 300, 400, and 15**

```
SELECT display_name
FROM rnc_database
WHERE min_length NOT IN (100, 200, 300, 400, 15);
```

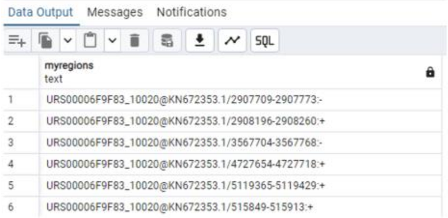


	display_name character varying (60)
1	ENA
2	GENCODE
3	MGnify
4	GeneCards
5	RDP
6	snoRNA Database
7	Rfam

5. **Can you get complete 500 records of sequences for active regions and name your column as myregions in which you are getting the region name column vakue. Then tell me what different chromosomes with exon_count we have for regions including center, east and north using the name you set for your column**

```
SELECT region_name AS myregions FROM rnc_sequence_regions WHERE region_name IS NOT NULL LIMIT 500;

SELECT DISTINCT chromosome, exon_count FROM rnc_sequence_regions WHERE region_name IN ('center', 'east', 'north') LIMIT 1000
```



	myregions text
1	URS00006F9F83_10020@KN672353.1/2907709-2907773:-
2	URS00006F9F83_10020@KN672353.1/2908196-2908260:+
3	URS00006F9F83_10020@KN672353.1/3567704-3567768:-
4	URS00006F9F83_10020@KN672353.1/4727654-4727718:+
5	URS00006F9F83_10020@KN672353.1/5119365-5119429:+
6	URS00006F9F83_10020@KN672353.1/515849-515913:+