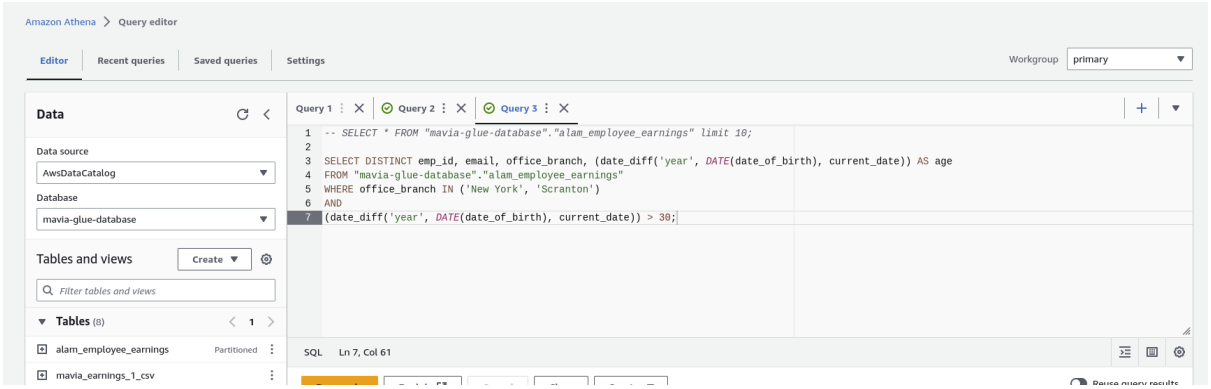
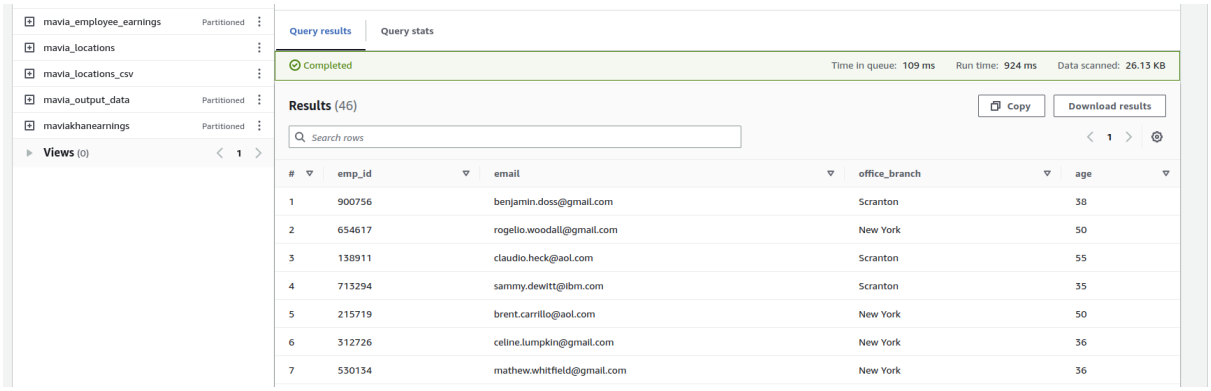
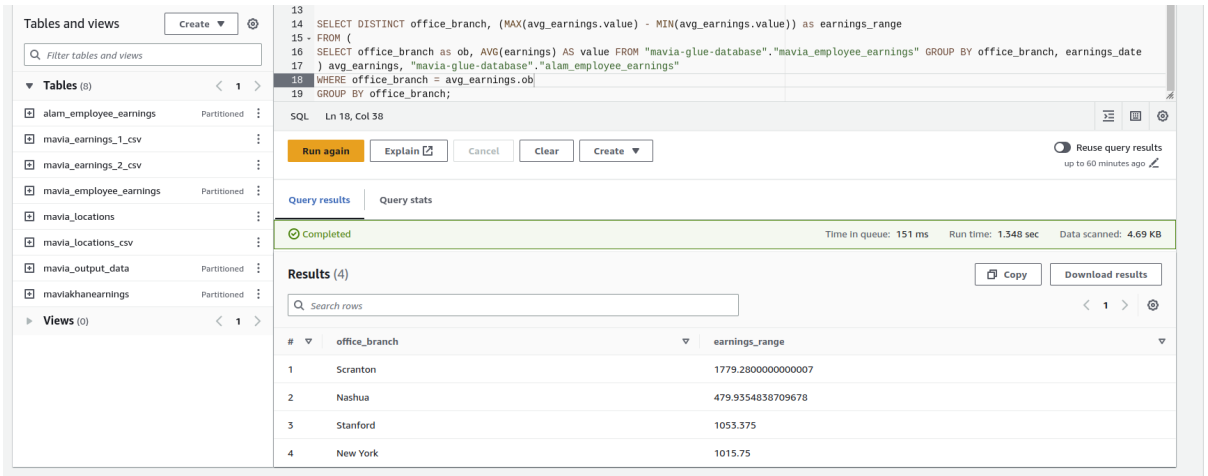
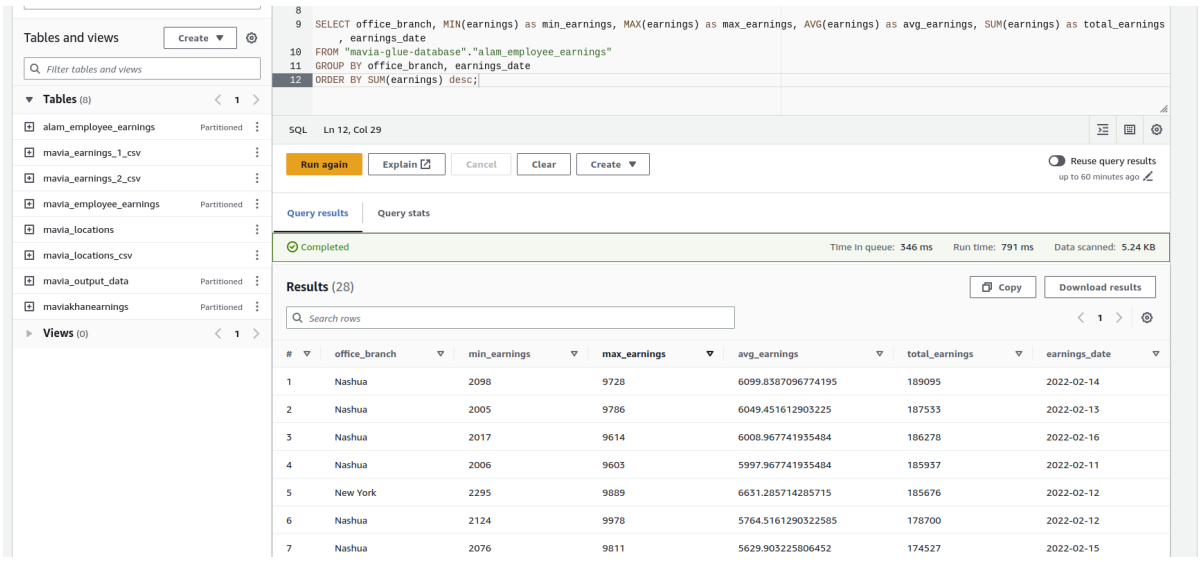
NAME: MAVIA ALAM KHAN (2303.KHI.DEG.017)
PAIRING WITH : Mohammad Hussam (2303.KHI.DEG.020)
ASSIGNMENT NO : 5.4

Here we run the task 3 queries

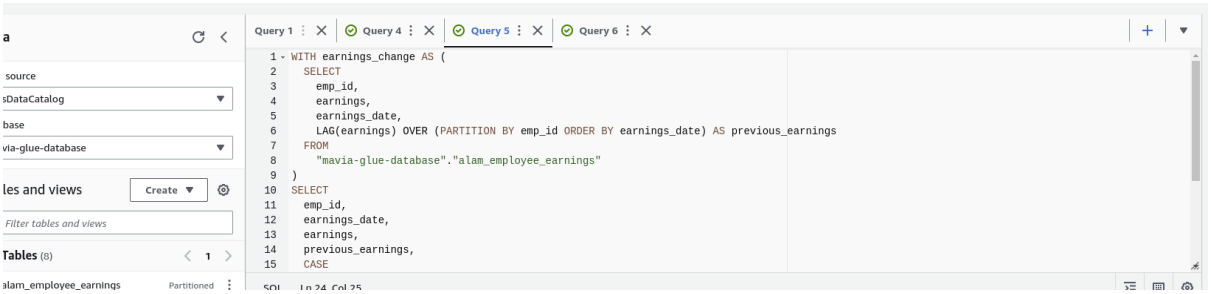


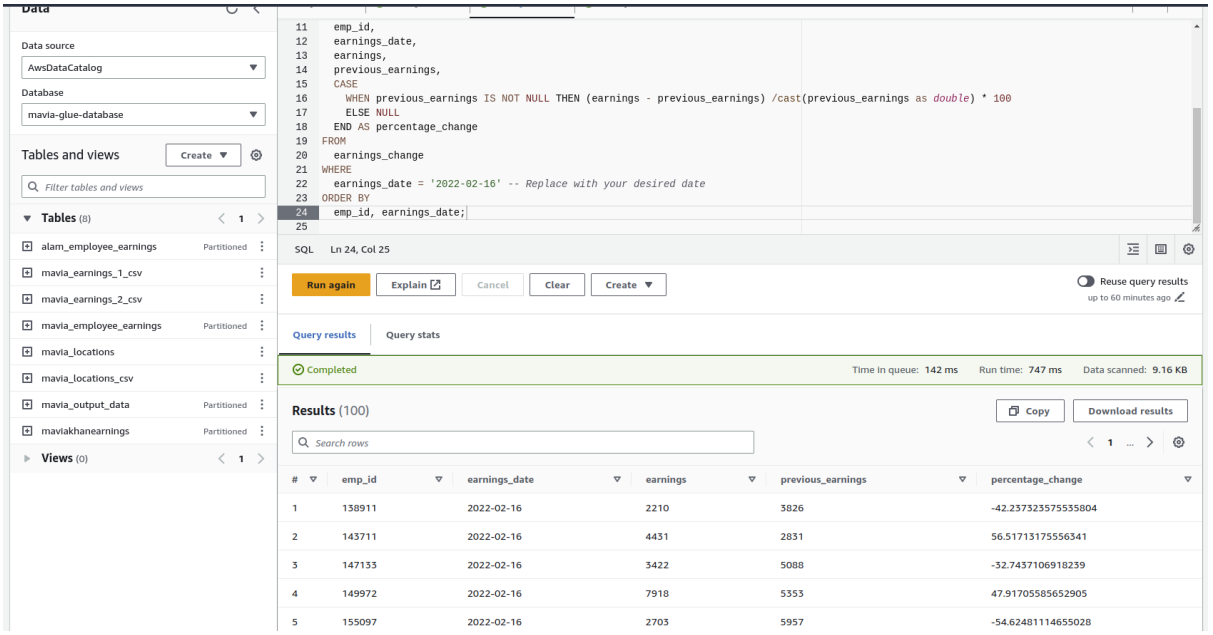




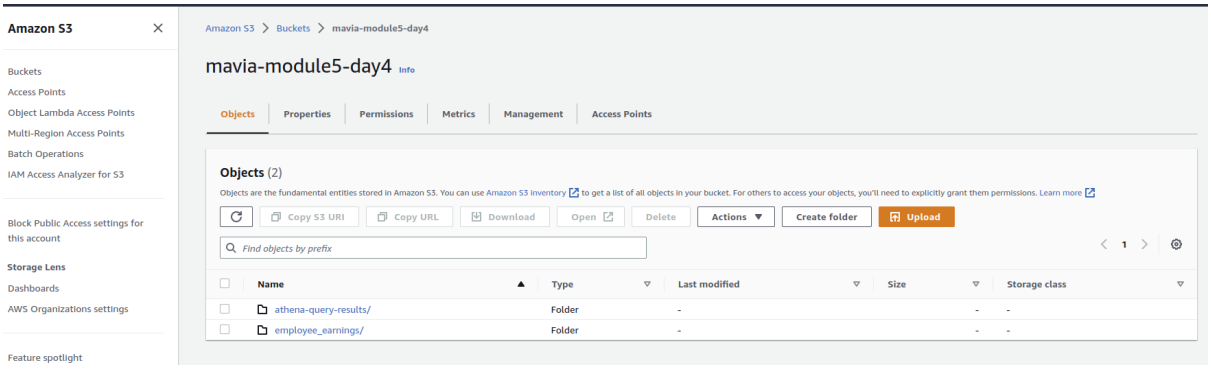


We calculates the % change in earnings for every employee from a given day compared to the previous day.

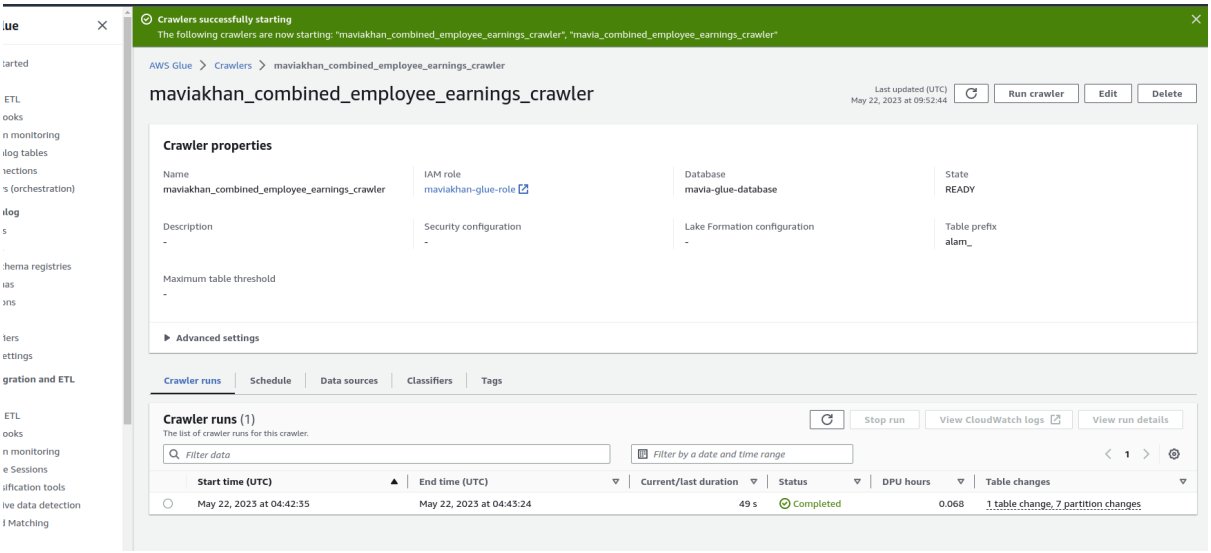




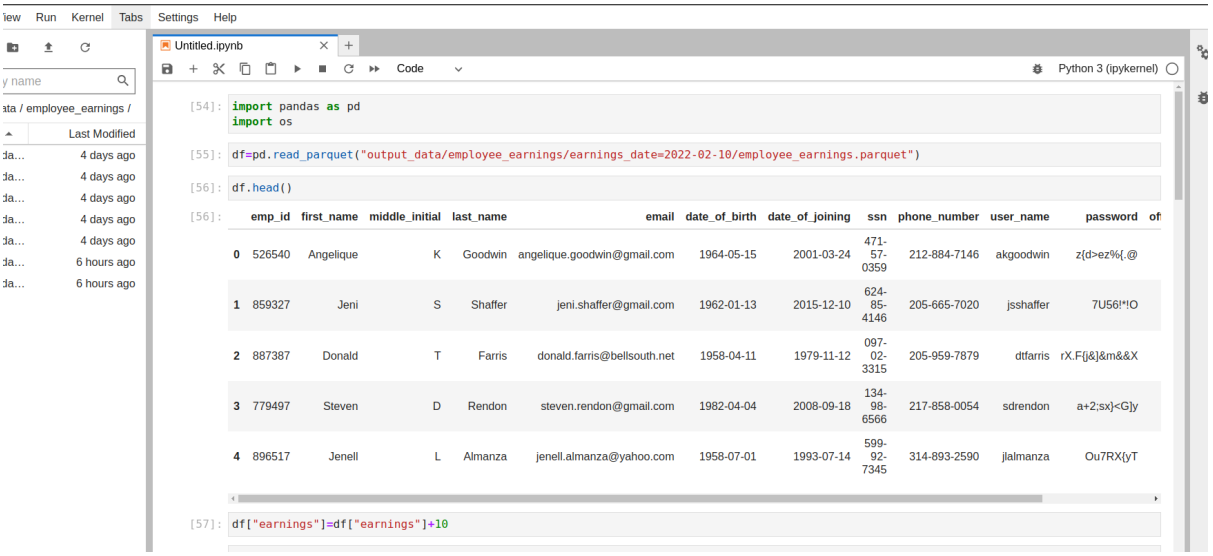
Here we create s3 bucket and create 2 folder one is employee earning and second is athena-query-result. In employee earning we stored a stored a dataset



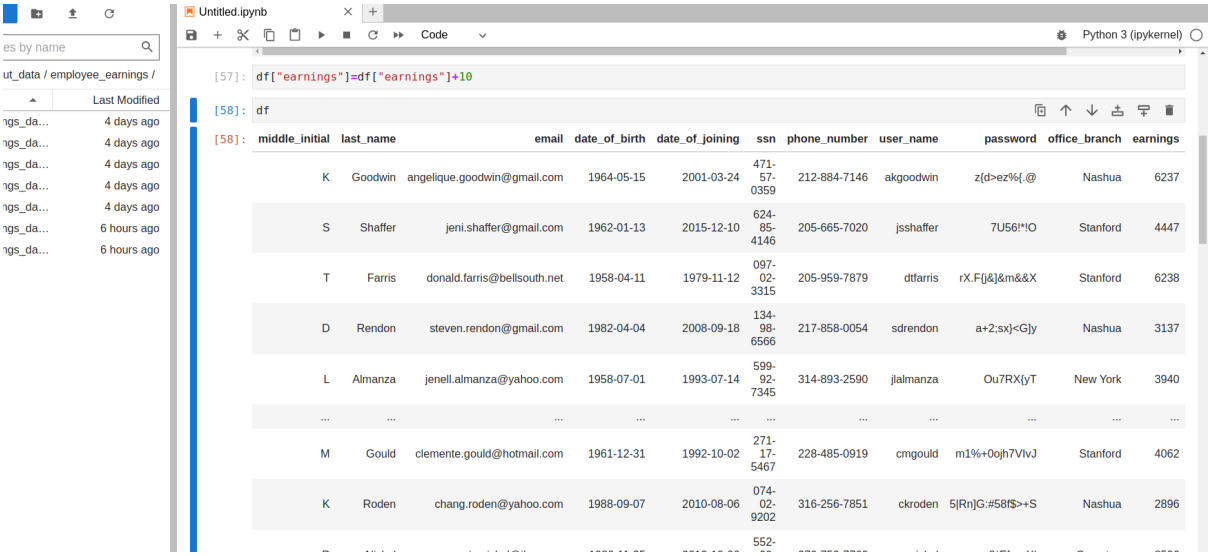
We create a crawler



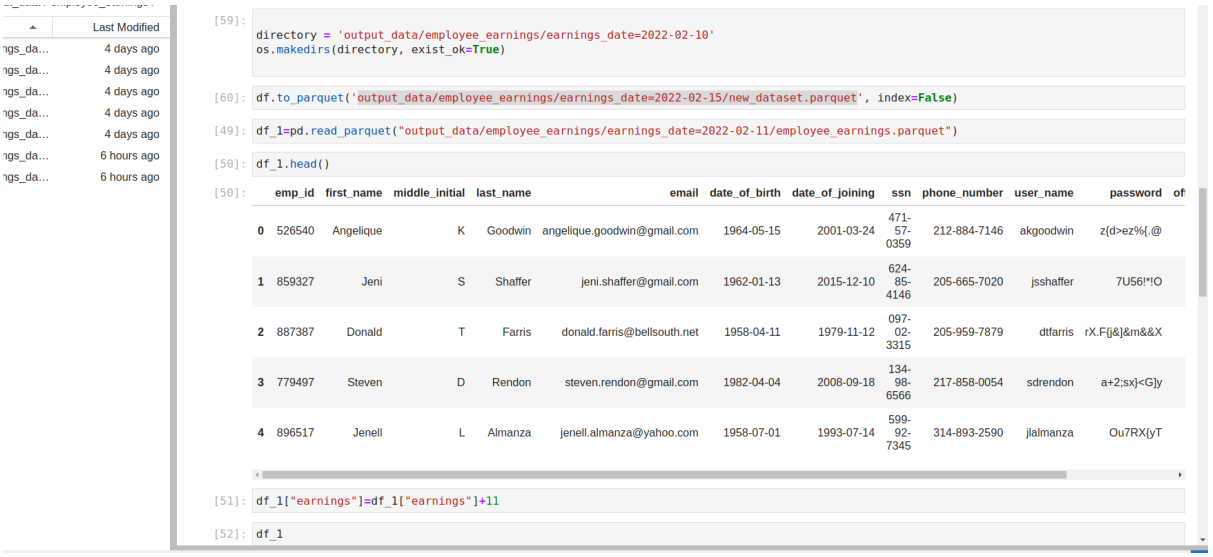
Using pandas we read a dataset



Here we change the value of earning column



We create a directory and stored a new parquet file



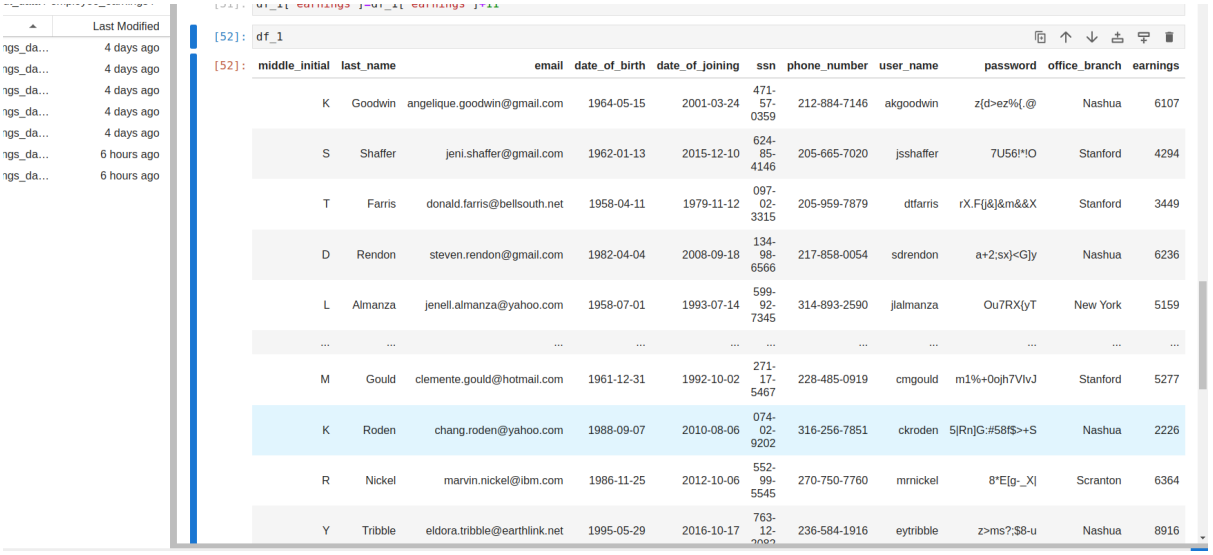
Last Modified
4 days ago
4 days ago
4 days ago
4 days ago
4 days ago
6 hours ago
6 hours ago

```
[59]:
directory = 'output_data/employee_earnings/earnings_date=2022-02-10'
os.makedirs(directory, exist_ok=True)

[60]: df.to_parquet('output_data/employee_earnings/earnings_date=2022-02-15/new_dataset.parquet', index=False)

[49]: df_1=pd.read_parquet("output_data/employee_earnings/earnings_date=2022-02-11/employee_earnings.parquet")

[50]: df_1.head()
```

[50]:

	emp_id	first_name	middle_initial	last_name	email	date_of_birth	date_of_joining	ssn	phone_number	user_name	password	of
0	526540	Angelique	K	Goodwin	angelique.goodwin@gmail.com	1964-05-15	2001-03-24	471-57-0359	212-884-7146	akgoodwin	z{d>ez%{.@	
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jsshaffer	7U56!*!O	
2	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879	dtfarris	rX.F{j&]&m&&X	
3	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04	2008-09-18	134-98-6566	217-858-0054	sdrendon	a+2;sx}<G]y	
4	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01	1993-07-14	599-92-7345	314-893-2590	jlalmanza	Ou7RX{yT	

```
[51]: df_1["earnings"]=df_1["earnings"]+11

[52]: df_1
```



Last Modified
4 days ago
4 days ago
4 days ago
4 days ago
4 days ago
6 hours ago
6 hours ago

```
[52]: df_1
```

[52]:

middle_initial	last_name	email	date_of_birth	date_of_joining	ssn	phone_number	user_name	password	office_branch	earnings
K	Goodwin	angelique.goodwin@gmail.com	1964-05-15	2001-03-24	471-57-0359	212-884-7146	akgoodwin	z{d>ez%{.@	Nashua	6107
S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jsshaffer	7U56!*!O	Stanford	4294
T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879	dtfarris	rX.F{j&]&m&&X	Stanford	3449
D	Rendon	steven.rendon@gmail.com	1982-04-04	2008-09-18	134-98-6566	217-858-0054	sdrendon	a+2;sx}<G]y	Nashua	6236
L	Almanza	jenell.almanza@yahoo.com	1958-07-01	1993-07-14	599-92-7345	314-893-2590	jlalmanza	Ou7RX{yT	New York	5159
...	...	...	...	...	...	...	...	...	...	...
M	Gould	clemente.gould@hotmail.com	1961-12-31	1992-10-02	271-17-5467	228-485-0919	cmgould	m1%+0ojh7VIvJ	Stanford	5277
K	Roden	chang.roden@yahoo.com	1988-09-07	2010-08-06	074-02-9202	316-256-7851	ckroden	5|Rn]G:#58f$>+S	Nashua	2226
R	Nickel	marvin.nickel@ibm.com	1986-11-25	2012-10-06	552-99-5545	270-750-7760	mrnickel	8*E[g-_X|	Scranton	6364
Y	Tribble	eldora.tribble@earthlink.net	1995-05-29	2016-10-17	763-12-2082	236-584-1916	eytribble	z>ms?;$8-u	Nashua	8916

```
[53]: df_1.to_parquet('output_data/employee_earnings/earnings_date=2022-02-16/new1_dataset.parquet', index=False)

[ ]:
```

We run the task 4 queries but all the queries are failed
May be appropriate filters, reducing the data volume, or using partitioning techniques, depending on the query engine .