

LinkUp Step 3 DEI - Get private firm JP

November 22, 2025

1 Intro

The code get job postings from private firms in our sample.

```
[1]: import pandas as pd
```

```
[4]: # Due to the large file size, each file needs to be read and processed
      ↪separately.

      # df_total = pd.read_pickle(r"D:\user_data\Dan\Job_
      ↪Posts\Processed_Data\20220618US_Job_Record_Final_batch1.pkl")
      df_total = pd.read_pickle(r"D:\user_data\Dan\Job_
      ↪Posts\Processed_Data\20220618US_Job_Record_Final_batch2.pkl")
```

```
[5]: df_total.head()
```

```
[5]:
```

	hash \
0	8000003e624f858208a2017becb57bcd
1	8000005153ff7de94c7db903c403bbf4
3	800000cb4090a45f053fc3a1b05b934b
4	800000df74dee5f8358f8b787370b975
6	800000e381fc76b63dff3bb41db7683a

	title	company_id \
0	Retail Coverage Merchandiser - Walmart	19555.0
1	Software Engineer	9271.0
3	Account Director	671.0
4	Biomedical Flight Controller (Houston, TX)	22381.0
6	Accounts Receivable Specialist	1946.0

	company_name	city	state	zip \
0	Acosta, Inc.	lamar	CO	81052
1	Raytheon Technologies Corp	schriever air force base	CO	None
3	iCrossing, Inc.	New York City	NY	10008
4	Wyle Laboratories, Inc.	Houston	TX	77020
6	Providence Health & Services	Portland	OR	97201

country	created	last_checked	...	company_id_updated \
---------	---------	--------------	-----	----------------------

0	USA	2020-11-15	2020-11-23	11:07:00+00:00	...	NaN
1	USA	2021-09-10	2021-09-21	13:25:00+00:00	...	NaN
3	USA	2009-04-14	2009-04-14	19:52:00+00:00	...	NaN
4	USA	2016-07-06	2016-07-11	00:25:00+00:00	...	NaN
6	USA	2012-02-25	2012-04-06	23:01:00+00:00	...	NaN

	created_date	onet_occupation_code_NEW	Job	Zone	year_month	\
0	2020-11-15 06:30:00+00:00	27-1026.00	3.0	2020-11		
1	2021-09-10 08:19:00+00:00	15-1253.00	4.0	2021-09		
3	2009-04-14 19:52:00+00:00	11-2011.00	4.0	2009-04		
4	2016-07-06 22:35:00+00:00	53-2021.00	3.0	2016-07		
6	2012-02-25 00:01:00+00:00	43-3031.00	3.0	2012-02		

	factset_entity_id	cusip	cusip_6	gvkey	the_key
0	0675SV-E	None	None	None	fskey_0675SV-E
1	000Y55-E	75513E101	75513E	010983	010983
3	064Q8J-E	None	None	None	fskey_064Q8J-E
4	05MMQS-E	48242W106	48242W	174310	174310
6	079CJD-E	None	None	None	fskey_079CJD-E

[5 rows x 25 columns]

```
[7]: firm_list = pd.read_excel(r"D:\user_data\Dan\JP_
↳ Informativeness\Raw_Data\Matched List.xlsx")
```

```
[8]: df_total["company_id"] = df_total["company_id"].apply(lambda x: int(x))
firm_list["company_id"] = firm_list["company_id"].apply(lambda x: int(x))
```

```
[10]: df_total2 = df_total.merge(firm_list, on = ["company_id"], how = "inner")
```

```
[11]: df_total2 = df_total2[df_total2["gvkey"].isna()]
```

```
[12]: df_total2.head()
```

	hash	title	\
401440	800004155f9dbdf6ba4bc6628892ff25	Driver - Local	
401441	803a4344dd835175640d2e0cf0adf77f	Driver - Local	
401442	8045fe21817b05092bfc7436facb7193	Driver - Linehaul	
401443	80ad8f1a486a6c3bf93c2a4a8dcc4334	Software Developer II - Java	
401444	811461aceace6236a8beb0f1ed580acf	Driver - Local	

	company_id	company_name	city	state	zip	country	\
401440	63074	Lynden Incorporated	sunnyside	WA	98944	USA	
401441	63074	Lynden Incorporated	mcminnville	OR	97128	USA	
401442	63074	Lynden Incorporated	bonanza	OR	97623	USA	
401443	63074	Lynden Incorporated	seattle	WA	98113	USA	
401444	63074	Lynden Incorporated	chehalis	WA	98532	USA	

	created		last_checked	...	Job	Zone	year_month	\
401440	2021-06-05	2021-06-15	05:48:00+00:00	...		2.0	2021-06	
401441	2020-11-18	2020-11-20	16:07:00+00:00	...		2.0	2020-11	
401442	2020-08-20	2020-09-07	09:06:00+00:00	...		2.0	2020-08	
401443	2021-11-05	2022-05-15	18:14:00+00:00	...		3.0	2021-11	
401444	2022-02-28	2022-03-10	17:02:00+00:00	...		2.0	2022-02	

	factset_entity_id	cusip	cusip_6	gvkey	the_key	\
401440	05L3DQ-E	551277106	551277	None	fskey_05L3DQ-E	
401441	05L3DQ-E	551277106	551277	None	fskey_05L3DQ-E	
401442	05L3DQ-E	551277106	551277	None	fskey_05L3DQ-E	
401443	05L3DQ-E	551277106	551277	None	fskey_05L3DQ-E	
401444	05L3DQ-E	551277106	551277	None	fskey_05L3DQ-E	

	Identifier	Name	Revenue
401440	05L3DQ-E	Lynden, Inc.	981.760911
401441	05L3DQ-E	Lynden, Inc.	981.760911
401442	05L3DQ-E	Lynden, Inc.	981.760911
401443	05L3DQ-E	Lynden, Inc.	981.760911
401444	05L3DQ-E	Lynden, Inc.	981.760911

[5 rows x 28 columns]

```
[13]: # df_total2.to_pickle("20230226 private_JP_chosen_1.pkl")
df_total2.to_pickle("20230226 private_JP_chosen_2.pkl")
```

2 Merge

```
[2]: df_total1 = pd.read_pickle("20230226 private_JP_chosen_1.pkl")
df_total2 = pd.read_pickle("20230226 private_JP_chosen_2.pkl")
```

```
[3]: df_total1 = df_total1.append(df_total2)
```

```
[5]: df_total1.to_pickle("20230226 private_JP_chosen_full.pkl")
```

```
[ ]:
```