

DEI Jobs Step 1 - Sample Selection

November 22, 2025

1 Intro

This code processes job postings sample from processed LinkUp files

```
[1]: import re
import os
import pandas as pd
import numpy as np
from tqdm import tqdm
```

```
[2]: pd.set_option('display.max_colwidth', 100)
pd.set_option('display.max_rows', 2000)
pd.set_option('display.max_columns', 2000)
```

2 Read Job Posts

```
[3]: total_job_posts = pd.read_csv("Processed_Data/20220618US_Job_Record_Final_batch1.
    ↪ csv")
total_job_posts_2 = pd.read_csv("Processed_Data/
    ↪ 20220618US_Job_Record_Final_batch2.csv")
total_job_posts = total_job_posts.append(total_job_posts_2)
```

```
[4]: total_job_posts = total_job_posts[["hash", "created", "the_key", "Job_Zone",
    ↪ "title", "company_name", "company_id", "state", "city"]]
```

```
[5]: # Find a sample of public firms covered in years after 2019
df_temp = total_job_posts[["the_key", "year"]].copy()
df_temp["d_public"] = df_temp["the_key"].apply(lambda x: re.search("_", x) is_
    ↪ None)
df_temp = df_temp.drop_duplicates(subset=["the_key", "year"])
df_temp = df_temp[df_temp["d_public"]]

df_temp["year"] = df_temp["year"].astype(int)
df_temp2 = df_temp[df_temp["year"] >= 2019]

year_counts = df_temp2.groupby("the_key").size().reset_index(name="n_years")
gvkey_chosen = year_counts[year_counts["n_years"] == 4]
```

```
[6]: gvkey_chosen["the_key"] = gvkey_chosen["the_key"].astype(str)
gvkey_chosen["the_key"] = gvkey_chosen["the_key"].str.zfill(6)
gvkey_chosen

[7]: total_job_posts = total_job_posts.merge(gvkey_chosen, on = "the_key", how = "left")

[8]: total_job_posts_chosen = total_job_posts[total_job_posts["0"].notna()]

[9]: total_job_posts_chosen
```

[9]:	hash	created	the_key	Job Zone	\
1	0000003cf0ed2539cc4950a95f420261	2008-08-02	015633	4.0	
8	000000be62247aa211a827915bbd1ad1	2020-07-30	028877	2.0	
9	000000c11b16e4556fb058b041e5f630	2019-11-03	065417	2.0	
14	000001125c8bde7c446643f50211a129	2017-12-18	006829	2.0	
18	000001a5f7022973b3d8b7b2c3e24082	2008-12-01	165123	4.0	
...	...	...	...	...	
163171786	fffffeddd0f27dd90366a32e55f77ee3	2014-01-17	014225	2.0	
163171788	ffffff5329a837b8e89fb7dc5491afaa	2015-07-20	005643	2.0	
163171789	ffffff5a888986e1bdc932a935b066e4	2021-07-10	176249	2.0	
163171791	ffffff8728f125f04f0560bcfbf28c9f	2014-02-12	009899	2.0	
163171795	ffffffcbd8ce7b19363bf7303589cf98	2016-12-16	064768	5.0	
	title \				
1	Manager, Personal and Business Banking, Powder Springs, GA				
8	2nd Shift- Warehouse Picker Order Selector				
9	Shift Lead				
14	Receiver/Stocker				
18	Sr. Environmental Engineer				
...					
...					
163171786	Sales Trainee PT				
163171788	Aloha Greeter - Casual				
163171789	Breakfast Coordinator				
163171791	Part Time Retail Sales Consultant Jacksonville NC (Jacksonville Western Blvd)				
163171795	Corporate Counsel - AWS				

	company_name	company_id	state	city	0
1	Royal Bank of Canada	9025.0	CA	Irvine	4.0
8	Keurig Dr Pepper	19172.0	CA	san leandro	4.0
9	Taco Bell	15692.0	CA	manteca	4.0
14	Lowe's Companies, Inc.	2839.0	DE	dover	4.0
18	SAIC	420.0	PA	Harrisburg	4.0
...	...	...	...	...	...
163171786	Fastenal	26.0	MS	Byram	4.0
163171788	Hilton Worldwide	14160.0	HI	Wailea	4.0
163171789	Carrols Restaurant Group, Inc.	39266.0	MI	plymouth	4.0
163171791	AT&T	2787.0	PI	NaN	4.0
163171795	Amazon.com, Inc.	469.0	WA	Seattle	4.0

[55502632 rows x 10 columns]

```
[10]: total_job_posts_chosen["created_date"] = total_job_posts_chosen["created"].
      ↪apply(lambda x: x[0:10])
total_job_posts_chosen['year_month'] = pd.
      ↪DatetimeIndex(total_job_posts_chosen['created_date']).strftime('%Y-%m')
total_job_posts_chosen['year'] = pd.
      ↪DatetimeIndex(total_job_posts_chosen['created_date']).year
total_job_posts_chosen
```

C:\Program Files (x86)\Microsoft Visual Studio\Shared\Anaconda3_64\lib\site-packages\ipykernel_launcher.py:1: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

"""Entry point for launching an IPython kernel.

C:\Program Files (x86)\Microsoft Visual Studio\Shared\Anaconda3_64\lib\site-packages\ipykernel_launcher.py:2: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

C:\Program Files (x86)\Microsoft Visual Studio\Shared\Anaconda3_64\lib\site-packages\ipykernel_launcher.py:3: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

This is separate from the ipykernel package so we can avoid doing imports until

```
[10]:
```

	hash	created	the_key	Job Zone	\
1	0000003cf0ed2539cc4950a95f420261	2008-08-02	015633	4.0	
8	000000be62247aa211a827915bbd1ad1	2020-07-30	028877	2.0	
9	000000c11b16e4556fb058b041e5f630	2019-11-03	065417	2.0	
14	000001125c8bde7c446643f50211a129	2017-12-18	006829	2.0	
18	000001a5f7022973b3d8b7b2c3e24082	2008-12-01	165123	4.0	
...	...	...	...	...	
163171786	ffffffeddd0f27dd90366a32e55f77ee3	2014-01-17	014225	2.0	
163171788	ffffff5329a837b8e89fb7dc5491afaa	2015-07-20	005643	2.0	
163171789	ffffff5a888986e1bdc932a935b066e4	2021-07-10	176249	2.0	
163171791	ffffff8728f125f04f0560bcfbf28c9f	2014-02-12	009899	2.0	
163171795	ffffffcbd8ce7b19363bf7303589cf98	2016-12-16	064768	5.0	

	title	\
1	Manager, Personal and Business Banking, Powder Springs, GA	
8	2nd Shift- Warehouse Picker Order Selector	
9	Shift Lead	
14	Receiver/Stocker	
18	Sr. Environmental Engineer	
...		
...		
163171786	Sales Trainee PT	
163171788	Aloha Greeter - Casual	
163171789	Breakfast Coordinator	
163171791	Part Time Retail Sales Consultant Jacksonville NC (Jacksonville Western Blvd)	
163171795	Corporate Counsel - AWS	

	company_name	company_id	state	city	0	\
1	Royal Bank of Canada	9025.0	CA	Irvine	4.0	
8	Keurig Dr Pepper	19172.0	CA	san leandro	4.0	
9	Taco Bell	15692.0	CA	manteca	4.0	
14	Lowe's Companies, Inc.	2839.0	DE	dover	4.0	
18	SAIC	420.0	PA	Harrisburg	4.0	
...	...	...	...	...	...	

163171786	Fastenal	26.0	MS	Byram	4.0
163171788	Hilton Worldwide	14160.0	HI	Wailea	4.0
163171789	Carrols Restaurant Group, Inc.	39266.0	MI	plymouth	4.0
163171791	AT&T	2787.0	PI	NaN	4.0
163171795	Amazon.com, Inc.	469.0	WA	Seattle	4.0

	created_date	year_month	year
1	2008-08-02	2008-08	2008
8	2020-07-30	2020-07	2020
9	2019-11-03	2019-11	2019
14	2017-12-18	2017-12	2017
18	2008-12-01	2008-12	2008
...	...	...	...
163171786	2014-01-17	2014-01	2014
163171788	2015-07-20	2015-07	2015
163171789	2021-07-10	2021-07	2021
163171791	2014-02-12	2014-02	2014
163171795	2016-12-16	2016-12	2016

[55502632 rows x 13 columns]

```
[11]: total_job_posts_chosen2 =
      ↪total_job_posts_chosen[total_job_posts_chosen["year"]>=2018]
```

```
[12]: total_job_posts_chosen2
```

```
[12]:
```

	hash	created	the_key	Job Zone	\
8	000000be62247aa211a827915bbd1ad1	2020-07-30	028877	2.0	
9	000000c11b16e4556fb058b041e5f630	2019-11-03	065417	2.0	
19	000001af53b758d6e190755ef3daad9d	2020-12-17	005680	2.0	
32	0000030edf766859e8ab8af37aaf5e4e	2019-12-30	011264	2.0	
33	00000311459cb84b752f46dc64aff4c6	2018-09-05	003231	2.0	
...	...	...	...	...	
163171746	fffffa6433e6ad29ece999e0c787eae0	2021-01-13	004040	2.0	
163171756	fffffb99f36038465d5e57bb5ba6b7e	2019-11-22	064166	3.0	
163171761	fffffc2287578c11ef90900af4f0b74b	2022-05-26	004016	2.0	
163171767	fffffc896f044e867919947196636d08	2021-09-22	011264	2.0	
163171789	fffff5a888986e1bdc932a935b066e4	2021-07-10	176249	2.0	

	title	\
8	2nd Shift- Warehouse Picker Order Selector	
9	Shift Lead	
19	Store Support	
32	Customer Service Associate	
33	Part Time Teller at Ann Arbor Lilley, Plymouth	
...	...	
163171746	Material Handler 3rd Shift	

```

163171756      Phlebotomist I - P/T- Hartford, CT req36398
163171761      Store Manager Candidate-recovery, Oh
163171767      Customer Service Associate
163171789      Breakfast Coordinator

```

```

                                company_name  company_id state      city \
8                                Keurig Dr Pepper    19172.0   CA      san leandro
9                                Taco Bell          15692.0   CA      manteca
19                               Home Depot           378.0    FL      bradenton
32                               Walgreens           8382.0    NV      reno
33                               Comerica           23101.0   MI      plymouth
...                               ...              ...      ...
163171746      RR Donnelley & Sons          20118.0   WI      plover
163171756      Quest Diagnostics Incorporated    2136.0    CT      hartford
163171761      Dollar General Corporation        503.0    OH      fort recovery
163171767      Walgreens                       8382.0    LA      boutte
163171789      Carrols Restaurant Group, Inc.   39266.0   MI      plymouth

```

```

0 created_date year_month  year
8      4.0    2020-07-30    2020-07    2020
9      4.0    2019-11-03    2019-11    2019
19     4.0    2020-12-17    2020-12    2020
32     4.0    2019-12-30    2019-12    2019
33     4.0    2018-09-05    2018-09    2018
...     ...     ...      ...      ...
163171746  4.0    2021-01-13    2021-01    2021
163171756  4.0    2019-11-22    2019-11    2019
163171761  4.0    2022-05-26    2022-05    2022
163171767  4.0    2021-09-22    2021-09    2021
163171789  4.0    2021-07-10    2021-07    2021

```

[29927927 rows x 13 columns]

```
[13]: total_job_posts_chosen2.to_pickle("Processed_Data/total_job_posts_chosen.pkl")
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```