

LinkUp Step 4 Process Job Posting

November 18, 2025

1 Intro

This code further clean and processes US job records

```
[1]: import re
import os
import pandas as pd
import numpy as np
from tqdm import tqdm
import tarfile
import time
import datetime
import sqlite3
import us

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

[3]: # Need to run twice due to the size of the data
import pickle5 as pickle
with open("C:/Users/dli/Downloads/20220607US_Job_Record_batch1.pkl", "rb") as fh:
    fh:
        total_job_post1 = pickle.load(fh)
# with open("C:/Users/dli/Downloads/20220607US_Job_Record_batch2.pkl", "rb") as fh:
#     fh:
#         total_job_post1 = pickle.load(fh)

[4]: total_job_post2 = total_job_post1.copy()
```

1.1 Update Company ID

```
[5]: LU_idupdatetable = pd.read_csv("Processed_Data/20220612companyid_updatetable.
    ↪csv")

[6]: total_job_post3 = total_job_post2.merge(LU_idupdatetable, how = "left", left_on_
    ↪= "company_id", right_on = "company_id" )
```

```
[7]: # def updateid(x):
#     import math
#     if math.isnan(x["company_id_updated"]):
#         return(x["company_id"])
#     else:
#         return(x["company_id_updated"])

[8]: total_job_post3["company_id"] = total_job_post3.apply(lambda x:updateid(x),
↪axis = 1)
```

2 State Clean

```
[9]: total_job_post3 = total_job_post3.replace({"state": us.states.mapping('fips',
↪'abbr')}})
total_job_post3["state_full"] = total_job_post3["state"].apply(lambda x:us.
↪states.lookup(str(x)))

[10]: total_job_post3 = total_job_post3[total_job_post3["state_full"].notna()]
```

3 Clean duplicates

```
[11]: total2 = total_job_post3.copy()

[12]: total2['onet_occupation_code2'] = total2.
↪groupby('hash')['onet_occupation_code'].transform('first')
total2['created_date2'] = total2.groupby('hash')['created'].transform('first')
total2['delete_date2'] = total2.groupby('hash')['delete_date'].
↪transform('first')

[13]: total2 = total2.drop_duplicates(subset=['hash', 'onet_occupation_code2',
↪'created_date2', 'delete_date2'])

[14]: total2['onet_occupation_code'] = total2['onet_occupation_code2']
total2['created_date'] = total2['created_date2']
total2['delete_date'] = total2['delete_date2']

[15]: total2 = total2.drop(columns=['onet_occupation_code2', 'created_date2',
↪"delete_date2"])
```

4 Onet code update

Many Onet code used in LinkUp are outdated and not used any more

Update using new code from www.onetcodeconnector.org

```
[16]: # import packages
from fake_useragent import UserAgent
import requests
from bs4 import BeautifulSoup
```

4.1 Need to run for only once

```
[ ]: # onetcodes = total2.drop_duplicates(subset=['onet_occupation_code'])
# onetcodes["onet_occupation_code_NEW"] = ""
# onetcodes
```

```
[ ]: # len(onetcodes)
```

```
[ ]: # ua = UserAgent()
# headers = {'User-Agent': ua.random}
# finished = 700
# for i, x in tqdm(onetcodes.iterrows()):
#     if i % 50 == 0:
#         headers = {'User-Agent': ua.random}
#     if i > finished:
#         finished = i - 2
#     else:
#         continue

#     oldcode = onetcodes["onet_occupation_code"].at[i]
#     url = "https://www.onetcodeconnector.org/find/result?s=" + str(oldcode)
#     response = requests.get(url, headers=headers)

#     # Parse text obtained
#     soup = BeautifulSoup(response.text, 'html.parser')
#     # Find all hyperlinks present on webpage
#     text = soup.find_all('div')
#     text = str(text[0])

#     results = re.search("This occupational code is no longer in use\. In the
↪future, please use\n<a href=+.+\>(.)</a> instead\.", text)
#     if results:
#         results = results.group(1)
#         temp = re.search("(.)\ \((.+)\)", results)
#         results_code = temp.group(1)
#         onetcodes["onet_occupation_code_NEW"].at[i] = results_code
#     else:
#         onetcodes["onet_occupation_code_NEW"].at[i] = oldcode
```

```
[ ]: # onetcodes = onetcodes[["onet_occupation_code", "onet_occupation_code_NEW"]]
# onetcodes.to_csv("onet_code_updatetable_batch2.csv", index = False)
```

```
[17]: onetcodes2 = pd.read_csv("onet_code_updatetable_batch2.csv")

[18]: total2 = total2.merge(onetcodes2, how = "left", left_on =
    ↪ "onet_occupation_code", right_on = "onet_occupation_code")

[19]: total2["onet_occupation_code"] = total2["onet_occupation_code_NEW"]
```

5 Job Zone

```
[20]: filename = "Onet/Job Zones.xlsx"
    job_zone = pd.read_excel(filename, engine='openpyxl')

[21]: job_zone2 = job_zone[["O*NET-SOC Code", "Job Zone"]]

[22]: total2 = total2.merge(job_zone2, how = "left", left_on =
    ↪ "onet_occupation_code_NEW", right_on = "O*NET-SOC Code")

[23]: total2 = total2.drop(columns=['O*NET-SOC Code'])
```

6 Year-month generate

```
[24]: total2['year_month'] = pd.DatetimeIndex(total2['created']).strftime('%Y-%m')
```

7 Merge with links

```
[25]: total2['created'] = total2['created'].apply(lambda x:pd.to_datetime(str(x),
    ↪ format='%Y-%m-%d')).dt.date

[26]: total2 = total2.drop(['state_full'], axis=1)

[27]: LU_lnkhist3 = pd.read_csv("Processed_Data/LU_lnkhist3.csv")

[28]: LU_lnkhist3['start_date'] = LU_lnkhist3['start_date'].apply(lambda x:pd.
    ↪ to_datetime(str(x), format='%Y-%m-%d')).dt.date
    LU_lnkhist3['end_date'] = LU_lnkhist3['end_date'].apply(lambda x:pd.
    ↪ to_datetime(str(x), format='%Y-%m-%d')).dt.date

[29]: conn2 = sqlite3.connect(':memory:')

    LU_lnkhist3.to_sql('LU_lnkhist3', conn2, index=False)
    total2.to_sql('total2', conn2, index=False)

    qry = '''
        SELECT
            a.*, b.factset_entity_id, b.cusip, b.cusip_6
```

```

FROM total2 a
LEFT JOIN LU_lnhist3 b
ON a.company_id = b.company_id
and b.start_date <= a.created
and b.end_date >= a.created

'''
total3 = pd.read_sql_query(qry, conn2)

```

```

[30]: total3 = total3.drop_duplicates(subset=['hash', "factset_entity_id", "cusip"],
↳keep='first')

```

8 Merge with COMPUSTAT

```

[31]: # comp = pd.read_csv('Raw_Data/compuSTAT.csv');
comp = pd.read_sas("WRDS/funda_identifier.sas7bdat", encoding="utf-8")
comp = comp[["gvkey", "datadate", "cusip"]]
comp["cusip_6"] = comp["cusip"].apply(lambda x:str(x)[0:6])
comp['datadate'] = comp['datadate'].apply(lambda x:pd.to_datetime(str(x),
↳format='%Y-%m-%d')).dt.date

```

```

[32]: comp2 = comp.copy()
comp2['year'] = pd.DatetimeIndex(comp2['datadate']).year
comp2 = comp2[comp2["year"]>=2007]
comp2 = comp2[comp2["year"]<=2021]

```

```

[33]: comp['datadate_lag1'] = comp["datadate"].apply(lambda x:x - datetime.
↳timedelta(days = 365))
comp['datadate_lead1'] = comp["datadate"].apply(lambda x:x + datetime.
↳timedelta(days = 365))
comp.head()

```

```

[34]: conn2 = sqlite3.connect(':memory:')

total3.to_sql('total3', conn2, index=False)
comp.to_sql('comp', conn2, index=False)

qry = '''
SELECT
    a.*, b.gvkey
FROM total3 a
LEFT JOIN comp b
ON a.cusip_6 = b.cusip_6
and b.datadate_lag1 <= a.created
and a.created <= b.datadate_lead1

```

```
'''
total4 = pd.read_sql_query(qry, conn2)
```

```
[35]: total4 = total4.drop_duplicates(subset=['hash', "factset_entity_id", "gvkey"],
↳ keep='first')
total4 = total4.sort_values(by=['hash', "gvkey", "factset_entity_id"])
total4 = total4.drop_duplicates(subset=['hash'], keep='first')
```

```
[36]: def key_generation(x):
      if x["gvkey"] is not None:
          return(x["gvkey"])
      if x["factset_entity_id"] is not None:
          return("fskey_" + str(x["factset_entity_id"]))
      return ("cikey_" + str(int(x["company_id"])))
```

```
[37]: total4["the_key"] = total4.apply(key_generation, axis = 1)
```

```
[38]: total4.to_pickle("Processed_Data/20220618US_Job_Record_Final_batch1.pkl")
      # total4.to_pickle("Processed_Data/20220618US_Job_Record_Final_batch2.pkl")
```

9 Create filling time

```
[3]: total4_batch1 = pd.read_pickle("Processed_Data/
↳ 20220618US_Job_Record_Final_batch1.pkl")
total4_batch1['year'] = pd.DatetimeIndex(total4_batch1['created']).year
```

```
[4]: total4_batch1
```

```
[4]:
```

	hash \		title
9967765	000000082858d2d0b2c0d9a130cdf016		
9967766	0000003cf0ed2539cc4950a95f420261		
9967768	000000491be59966a6408d201350c9ce		
9967769	000000781fbcd4459bac5aa390464e47		
9967770	00000090d1133192dd83e2ed9dc33b1c		
...	...		
114303948	7fffff60ed07d167bbf16e2ced9d2dc2		
114303949	7fffff95e45d9fba64285161cd4f82fc		
114303950	7fffffc87badca288e755d9d7f801a16		
114303951	7fffffd6a96358ec2531a7bc7fad7303		
114303953	7fffffde7edd80bfc7eefbe0fa75c0c8		
9967765		Customer Service Representative	
9967766		Manager, Personal and Business Banking, Powder Springs, GA	
9967768		LPN General Surgery Full Time	

9967769	RETAIL CONSULTANT
9967770	Outside Solar Sales Consultant
...	...
114303948	Automotive Detailer
114303949	Sommelier - Olive's Bar (PT)
114303950	Transportation Attendant
114303951	Operations Analyst-21McLean, VA
114303953	Customer Care Advocates- Hybrid Roles \$20/hour. Launch your career!

	company_id	company_name	city	state	zip	\
9967765	11198.0	Aaron's, Inc	Anderson	SC	29626	
9967766	9025.0	Royal Bank of Canada	Irvine	CA	92606	
9967768	1960.0	Trinity Health Corporation	boise	ID	83701	
9967769	1731.0	Sprint	High Point	NC	27261	
9967770	14901.0	Solarcity Corporation	Sacramento	CA	94204	

...	...	...	...	...	...	...
114303948	36720.0	Penske Automotive Group	chantilly	VA	22022	
114303949	22797.0	MGM Resorts	Las Vegas	NV	89134	
114303950	26344.0	Emory Healthcare	atlanta	GA	30301	
114303951	420.0	SAIC	McLean	VA	22107	
114303953	20887.0	MetLife, Inc.	omaha	NE	68102	

	country	created	last_checked	last_updated	\
9967765	USA	2016-04-09	2016-04-18 04:03:00+00:00	None	
9967766	USA	2008-08-02	2008-08-09 05:16:00+00:00	None	
9967768	USA	2022-03-27	2022-05-23 16:29:00+00:00	None	
9967769	USA	2008-01-16	2008-01-16 20:50:00+00:00	None	
9967770	USA	2016-02-05	2016-02-16 23:10:00+00:00	None	
...	...	...	...	...	...
114303948	USA	2019-10-17	2020-01-07 12:01:00+00:00	None	
114303949	USA	2014-05-30	2014-07-11 07:52:00+00:00	None	
114303950	USA	2019-06-05	2019-09-13 10:43:00+00:00	None	
114303951	USA	2008-09-24	2008-09-24 16:26:00+00:00	None	
114303953	USA	2021-10-28	2022-01-07 10:29:00+00:00	None	

	delete_date	unmapped_location	onet_occupation_code	\
9967765	2016-04-20 09:50:00+00:00	0	43-4051.00	
9967766	2008-08-11 05:56:00+00:00	0	13-2072.00	
9967768	2022-05-26 03:21:00+00:00	0	29-2061.00	
9967769	2008-01-18 08:50:00+00:00	0	41-2031.00	
9967770	2016-02-19 07:20:00+00:00	0	41-4011.07	
...	...	...	...	...
114303948	2020-01-09 13:57:00+00:00	0	53-7061.00	
114303949	2014-07-16 00:20:00+00:00	0	35-3031.00	
114303950	2019-09-15 17:06:00+00:00	0	31-1131.00	
114303951	2008-09-27 03:41:00+00:00	0	15-2031.00	
114303953	2022-01-12 20:16:00+00:00	0	43-4051.00	

```

url \
9967765 https://jobs.aarons.com/job/anderson/customer-
service-representative/1618/1942568
9967766 https://careers.peopleclick.com/careerscp/client_rbc/external/gateway
.do?functionName=viewFromLi...
9967768 https://jobs.trinity-
health.org/job/TRHEUS00347492SAINTALPHONSUSENUS
9967769 https://careersearch.sprint.com/1033/ASP/TG/cim_jobdetail.asp?jobId=5
73278&type=search&JobReqLan...
9967770 https://solarcity.wd1.myworkdayjobs.com/en-US/01//job/Sacramento-
CA/Outside-Solar-Sales-Consulta...
...
...
114303948 https://recruiting.adp.com/srccar/public/nghome.guid?c=1151451&d=Exte
rnalCareerSite&prc=RMP0D3&r...
114303949
https://www.hrapply.com/mgmresorts/AppJobView.jsp?link=179871
114303950 https://non-clinical-
emory.icims.com/jobs/35655/transportation-attendant/job
114303951 https://cp-its-rmprd.saic.com/main/careerportal/Job_Profile.cfm?/2IG2
6H2IHSEK02Z9JGK9YSJX3GOMTBZ...
114303953 https://jobs.metlife.com/job/Omaha-Customer-Care-Advocates-Hybrid-
Roles-%2420hour_-Launch-your-c...

```

	company_id_updated	created_date \
9967765	NaN	2016-04-09 06:14:00+00:00
9967766	NaN	2008-08-02 14:44:00+00:00
9967768	NaN	2022-03-27 04:26:00+00:00
9967769	NaN	2008-01-16 20:50:00+00:00
9967770	NaN	2016-02-05 23:24:00+00:00
...	...	...
114303948	NaN	2019-10-17 20:35:00+00:00
114303949	NaN	2014-05-30 12:03:00+00:00
114303950	NaN	2019-06-05 04:08:00+00:00
114303951	NaN	2008-09-24 16:26:00+00:00
114303953	NaN	2021-10-28 15:48:00+00:00

	onet_occupation_code_NEW	Job Zone	year_month	factset_entity_id \
9967765	43-4051.00	2.0	2016-04	000BFT-E
9967766	13-2072.00	4.0	2008-08	0018FV-E
9967768	29-2061.00	3.0	2022-03	05LC20-E
9967769	41-2031.00	2.0	2008-01	0D7N1X-E
9967770	41-4011.07	3.0	2016-02	06KGBG-E
...	...	...	...	...
114303948	53-7061.00	2.0	2019-10	None
114303949	35-3031.00	2.0	2014-05	None

114303950	31-1131.00	3.0	2019-06	None
114303951	15-2031.00	5.0	2008-09	00FD97-E
114303953	43-4051.00	2.0	2021-10	None

	cusip	cusip_6	gvkey	the_key	year
9967765	002535102	002535	None	fskey_000BFT-E	2016
9967766	780087102	780087	015633	015633	2008
9967768	None	None	None	fskey_05LC20-E	2022
9967769	J7596P109	J7596P	None	fskey_0D7N1X-E	2008
9967770	83416T100	83416T	015314	015314	2016
...	...	...	...	...	...
114303948	None	None	None	cikey_36720	2019
114303949	None	None	None	cikey_22797	2014
114303950	None	None	None	cikey_26344	2019
114303951	525327102	525327	165123	165123	2008
114303953	None	None	None	cikey_20887	2021

[81671837 rows x 26 columns]

[4]: hash \

9967765	000000082858d2d0b2c0d9a130cdf016
9967766	0000003cf0ed2539cc4950a95f420261
9967768	000000491be59966a6408d201350c9ce
9967769	000000781fbcd4459bac5aa390464e47
9967770	00000090d1133192dd83e2ed9dc33b1c
...	...
114303948	7ffffff60ed07d167bbf16e2ced9d2dc2
114303949	7ffffff95e45d9fba64285161cd4f82fc
114303950	7ffffffc87badca288e755d9d7f801a16
114303951	7ffffffd6a96358ec2531a7bc7fad7303
114303953	7ffffffde7edd80bfc7eefbe0fa75c0c8

	title
9967765	Customer Service Representative
9967766	Manager, Personal and Business Banking, Powder Springs, GA
9967768	LPN General Surgery Full Time
9967769	RETAIL CONSULTANT
9967770	Outside Solar Sales Consultant
...	...
114303948	Automotive Detailer
114303949	Sommelier - Olive's Bar (PT)
114303950	Transportation Attendant
114303951	Operations Analyst-21McLean, VA
114303953	Customer Care Advocates- Hybrid Roles \$20/hour. Launch your career!

company_id	company_name	city	state	zip	\
------------	--------------	------	-------	-----	---

9967765	11198.0	Aaron's, Inc	Anderson	SC	29626
9967766	9025.0	Royal Bank of Canada	Irvine	CA	92606
9967768	1960.0	Trinity Health Corporation	boise	ID	83701
9967769	1731.0	Sprint	High Point	NC	27261
9967770	14901.0	Solarcity Corporation	Sacramento	CA	94204
...	...	...	...	...	...
114303948	36720.0	Penske Automotive Group	chantilly	VA	22022
114303949	22797.0	MGM Resorts	Las Vegas	NV	89134
114303950	26344.0	Emory Healthcare	atlanta	GA	30301
114303951	420.0	SAIC	McLean	VA	22107
114303953	20887.0	MetLife, Inc.	omaha	NE	68102

	country	created	last_checked	last_updated	\
9967765	USA	2016-04-09	2016-04-18 04:03:00+00:00		None
9967766	USA	2008-08-02	2008-08-09 05:16:00+00:00		None
9967768	USA	2022-03-27	2022-05-23 16:29:00+00:00		None
9967769	USA	2008-01-16	2008-01-16 20:50:00+00:00		None
9967770	USA	2016-02-05	2016-02-16 23:10:00+00:00		None
...	...	...	...	...	...
114303948	USA	2019-10-17	2020-01-07 12:01:00+00:00		None
114303949	USA	2014-05-30	2014-07-11 07:52:00+00:00		None
114303950	USA	2019-06-05	2019-09-13 10:43:00+00:00		None
114303951	USA	2008-09-24	2008-09-24 16:26:00+00:00		None
114303953	USA	2021-10-28	2022-01-07 10:29:00+00:00		None

	delete_date	unmapped_location	onet_occupation_code	\
9967765	2016-04-20 09:50:00+00:00	0	43-4051.00	
9967766	2008-08-11 05:56:00+00:00	0	13-2072.00	
9967768	2022-05-26 03:21:00+00:00	0	29-2061.00	
9967769	2008-01-18 08:50:00+00:00	0	41-2031.00	
9967770	2016-02-19 07:20:00+00:00	0	41-4011.07	
...	...	...	...	...
114303948	2020-01-09 13:57:00+00:00	0	53-7061.00	
114303949	2014-07-16 00:20:00+00:00	0	35-3031.00	
114303950	2019-09-15 17:06:00+00:00	0	31-1131.00	
114303951	2008-09-27 03:41:00+00:00	0	15-2031.00	
114303953	2022-01-12 20:16:00+00:00	0	43-4051.00	

	url	\
9967765	https://jobs.aarons.com/job/anderson/customer-service-representative/1618/1942568	
9967766	https://careers.peopleclick.com/careerscp/client_rbc/external/gateway.do?functionName=viewFromLi...	
9967768	https://jobs.trinity-health.org/job/TRHEUS00347492SAINTALPHONSUSENUS	
9967769	https://careersearch.sprint.com/1033/ASP/TG/cim_jobdetail.asp?jobId=573278&type=search&JobReqLan...	

9967770 <https://solarcity.wd1.myworkdayjobs.com/en-US/01//job/Sacramento-CA/Outside-Solar-Sales-Consulta...>
 ...
 ...
 114303948 <https://recruiting.adp.com/srccar/public/nghome.guid?c=1151451&d=ExternalCareerSite&prc=RMP0D3&r...>
 114303949
<https://www.hrapply.com/mgmresorts/AppJobView.jsp?link=179871>
 114303950 <https://non-clinical-emory.icims.com/jobs/35655/transportation-attendant/job>
 114303951 https://cp-its-rmprd.saic.com/main/careerportal/Job_Profile.cfm?/2IG26H2IHSEK02Z9JGK9YSJX3G0MTBZ...
 114303953 https://jobs.metlife.com/job/Omaha-Customer-Care-Advocates-Hybrid-Roles-%2420hour_-Launch-your-c...

	company_id_updated	created_date \
9967765	NaN	2016-04-09 06:14:00+00:00
9967766	NaN	2008-08-02 14:44:00+00:00
9967768	NaN	2022-03-27 04:26:00+00:00
9967769	NaN	2008-01-16 20:50:00+00:00
9967770	NaN	2016-02-05 23:24:00+00:00
...	...	...
114303948	NaN	2019-10-17 20:35:00+00:00
114303949	NaN	2014-05-30 12:03:00+00:00
114303950	NaN	2019-06-05 04:08:00+00:00
114303951	NaN	2008-09-24 16:26:00+00:00
114303953	NaN	2021-10-28 15:48:00+00:00

	onet_occupation_code_NEW	Job Zone	year_month	factset_entity_id \
9967765	43-4051.00	2.0	2016-04	000BFT-E
9967766	13-2072.00	4.0	2008-08	0018FV-E
9967768	29-2061.00	3.0	2022-03	05LC20-E
9967769	41-2031.00	2.0	2008-01	0D7N1X-E
9967770	41-4011.07	3.0	2016-02	06KGBG-E
...	...	...	...	...
114303948	53-7061.00	2.0	2019-10	None
114303949	35-3031.00	2.0	2014-05	None
114303950	31-1131.00	3.0	2019-06	None
114303951	15-2031.00	5.0	2008-09	00FD97-E
114303953	43-4051.00	2.0	2021-10	None

	cusip	cusip_6	gvkey	the_key	year
9967765	002535102	002535	None	fskey_000BFT-E	2016
9967766	780087102	780087	015633	015633	2008
9967768	None	None	None	fskey_05LC20-E	2022
9967769	J7596P109	J7596P	None	fskey_0D7N1X-E	2008
9967770	83416T100	83416T	015314	015314	2016

...	...	...	...	...	...
114303948	None	None	None	cikey_36720	2019
114303949	None	None	None	cikey_22797	2014
114303950	None	None	None	cikey_26344	2019
114303951	525327102	525327	165123	165123	2008
114303953	None	None	None	cikey_20887	2021

[81671837 rows x 26 columns]

[4]: hash \

9967765	000000082858d2d0b2c0d9a130cdf016
9967766	0000003cf0ed2539cc4950a95f420261
9967768	000000491be59966a6408d201350c9ce
9967769	000000781fbcd4459bac5aa390464e47
9967770	00000090d1133192dd83e2ed9dc33b1c

...	...
114303948	7ffffff60ed07d167bbf16e2ced9d2dc2
114303949	7ffffff95e45d9fba64285161cd4f82fc
114303950	7ffffffc87badca288e755d9d7f801a16
114303951	7ffffffd6a96358ec2531a7bc7fad7303
114303953	7ffffffde7edd80bfc7eefbe0fa75c0c8

		title
\		
9967765		Customer Service Representative
9967766	Manager, Personal and Business Banking, Powder Springs, GA	
9967768	LPN General Surgery Full Time	
9967769	RETAIL CONSULTANT	
9967770	Outside Solar Sales Consultant	
...		...
114303948		Automotive Detailer
114303949		Sommelier - Olive's Bar (PT)
114303950		Transportation Attendant
114303951		Operations Analyst-21McLean, VA
114303953	Customer Care Advocates- Hybrid Roles \$20/hour. Launch your career!	

	company_id	company_name	city	state	zip	\
9967765	11198.0	Aaron's, Inc	Anderson	SC	29626	
9967766	9025.0	Royal Bank of Canada	Irvine	CA	92606	
9967768	1960.0	Trinity Health Corporation	boise	ID	83701	
9967769	1731.0	Sprint	High Point	NC	27261	
9967770	14901.0	Solarcity Corporation	Sacramento	CA	94204	
...	...	...	...	...	...	...
114303948	36720.0	Penske Automotive Group	chantilly	VA	22022	
114303949	22797.0	MGM Resorts	Las Vegas	NV	89134	
114303950	26344.0	Emory Healthcare	atlanta	GA	30301	
114303951	420.0	SAIC	McLean	VA	22107	

114303953 20887.0 MetLife, Inc. omaha NE 68102

	country	created	last_checked	last_updated	\
9967765	USA	2016-04-09	2016-04-18 04:03:00+00:00	None	
9967766	USA	2008-08-02	2008-08-09 05:16:00+00:00	None	
9967768	USA	2022-03-27	2022-05-23 16:29:00+00:00	None	
9967769	USA	2008-01-16	2008-01-16 20:50:00+00:00	None	
9967770	USA	2016-02-05	2016-02-16 23:10:00+00:00	None	
...	...	...	...	...	
114303948	USA	2019-10-17	2020-01-07 12:01:00+00:00	None	
114303949	USA	2014-05-30	2014-07-11 07:52:00+00:00	None	
114303950	USA	2019-06-05	2019-09-13 10:43:00+00:00	None	
114303951	USA	2008-09-24	2008-09-24 16:26:00+00:00	None	
114303953	USA	2021-10-28	2022-01-07 10:29:00+00:00	None	

	delete_date	unmapped_location	onet_occupation_code	\
9967765	2016-04-20 09:50:00+00:00	0	43-4051.00	
9967766	2008-08-11 05:56:00+00:00	0	13-2072.00	
9967768	2022-05-26 03:21:00+00:00	0	29-2061.00	
9967769	2008-01-18 08:50:00+00:00	0	41-2031.00	
9967770	2016-02-19 07:20:00+00:00	0	41-4011.07	
...	...	...	...	
114303948	2020-01-09 13:57:00+00:00	0	53-7061.00	
114303949	2014-07-16 00:20:00+00:00	0	35-3031.00	
114303950	2019-09-15 17:06:00+00:00	0	31-1131.00	
114303951	2008-09-27 03:41:00+00:00	0	15-2031.00	
114303953	2022-01-12 20:16:00+00:00	0	43-4051.00	

	url	\
9967765	https://jobs.aarons.com/job/anderson/customer-service-representative/1618/1942568	
9967766	https://careers.peopleclick.com/careerscp/client_rbc/external/gateway.do?functionName=viewFromLi...	
9967768	https://jobs.trinity-health.org/job/TRHEUS00347492SAINTALPHONSUSENUS	
9967769	https://careersearch.sprint.com/1033/ASP/TG/cim_jobdetail.asp?jobId=573278&type=search&JobReqLan...	
9967770	https://solarcity.wd1.myworkdayjobs.com/en-US/01//job/Sacramento-CA/Outside-Solar-Sales-Consulta...	
...		
...		
114303948	https://recruiting.adp.com/srccar/public/nghome.guid?c=1151451&d=ExternalCareerSite&prc=RMP0D3&r...	
114303949	https://www.hrapply.com/mgmresorts/AppJobView.jsp?link=179871	
114303950	https://non-clinical-emory.icims.com/jobs/35655/transportation-attendant/job	

114303951 https://cp-its-rmprd.saic.com/main/careerportal/Job_Profile.cfm?/2IG26H2IHSEK02Z9JGK9YSJX3GOMTBZ...
 114303953 https://jobs.metlife.com/job/Omaha-Customer-Care-Advocates-Hybrid-Roles-%2420hour_-Launch-your-c...

	company_id_updated	created_date	\
9967765	NaN	2016-04-09 06:14:00+00:00	
9967766	NaN	2008-08-02 14:44:00+00:00	
9967768	NaN	2022-03-27 04:26:00+00:00	
9967769	NaN	2008-01-16 20:50:00+00:00	
9967770	NaN	2016-02-05 23:24:00+00:00	
...	...	...	
114303948	NaN	2019-10-17 20:35:00+00:00	
114303949	NaN	2014-05-30 12:03:00+00:00	
114303950	NaN	2019-06-05 04:08:00+00:00	
114303951	NaN	2008-09-24 16:26:00+00:00	
114303953	NaN	2021-10-28 15:48:00+00:00	

	onet_occupation_code_NEW	Job Zone	year_month	factset_entity_id	\
9967765	43-4051.00	2.0	2016-04	000BFT-E	
9967766	13-2072.00	4.0	2008-08	0018FV-E	
9967768	29-2061.00	3.0	2022-03	05LC20-E	
9967769	41-2031.00	2.0	2008-01	0D7N1X-E	
9967770	41-4011.07	3.0	2016-02	06KGBG-E	
...	...	...	...	...	
114303948	53-7061.00	2.0	2019-10	None	
114303949	35-3031.00	2.0	2014-05	None	
114303950	31-1131.00	3.0	2019-06	None	
114303951	15-2031.00	5.0	2008-09	00FD97-E	
114303953	43-4051.00	2.0	2021-10	None	

	cusip	cusip_6	gvkey	the_key	year
9967765	002535102	002535	None	fskey_000BFT-E	2016
9967766	780087102	780087	015633	015633	2008
9967768	None	None	None	fskey_05LC20-E	2022
9967769	J7596P109	J7596P	None	fskey_0D7N1X-E	2008
9967770	83416T100	83416T	015314	015314	2016
...	...	...	...	...	...
114303948	None	None	None	cikey_36720	2019
114303949	None	None	None	cikey_22797	2014
114303950	None	None	None	cikey_26344	2019
114303951	525327102	525327	165123	165123	2008
114303953	None	None	None	cikey_20887	2021

[81671837 rows x 26 columns]

[4]:

	hash	\		title
9967765	000000082858d2d0b2c0d9a130cdf016			
9967766	0000003cf0ed2539cc4950a95f420261			
9967768	000000491be59966a6408d201350c9ce			
9967769	000000781fbcd4459bac5aa390464e47			
9967770	00000090d1133192dd83e2ed9dc33b1c			
...	...			
114303948	7ffffff60ed07d167bbf16e2ced9d2dc2			
114303949	7ffffff95e45d9fba64285161cd4f82fc			
114303950	7ffffffc87badca288e755d9d7f801a16			
114303951	7ffffffd6a96358ec2531a7bc7fad7303			
114303953	7ffffffde7edd80bfc7eefbe0fa75c0c8			
...	...			
9967765				Customer Service Representative
9967766				Manager, Personal and Business Banking, Powder Springs, GA
9967768				LPN General Surgery Full Time
9967769				RETAIL CONSULTANT
9967770				Outside Solar Sales Consultant
...	...			...
114303948				Automotive Detailer
114303949				Sommelier - Olive's Bar (PT)
114303950				Transportation Attendant
114303951				Operations Analyst-21McLean, VA
114303953				Customer Care Advocates- Hybrid Roles \$20/hour. Launch your career!
	company_id	company_name	city state	zip \
9967765	11198.0	Aaron's, Inc	Anderson SC	29626
9967766	9025.0	Royal Bank of Canada	Irvine CA	92606
9967768	1960.0	Trinity Health Corporation	boise ID	83701
9967769	1731.0	Sprint	High Point NC	27261
9967770	14901.0	Solarcity Corporation	Sacramento CA	94204
...	...	...	...	...
114303948	36720.0	Penske Automotive Group	chantilly VA	22022
114303949	22797.0	MGM Resorts	Las Vegas NV	89134
114303950	26344.0	Emory Healthcare	atlanta GA	30301
114303951	420.0	SAIC	McLean VA	22107
114303953	20887.0	MetLife, Inc.	omaha NE	68102
	country	created	last_checked	last_updated \
9967765	USA	2016-04-09	2016-04-18 04:03:00+00:00	None
9967766	USA	2008-08-02	2008-08-09 05:16:00+00:00	None
9967768	USA	2022-03-27	2022-05-23 16:29:00+00:00	None
9967769	USA	2008-01-16	2008-01-16 20:50:00+00:00	None
9967770	USA	2016-02-05	2016-02-16 23:10:00+00:00	None
...	...	...	...	...

114303948	USA	2019-10-17	2020-01-07 12:01:00+00:00	None
114303949	USA	2014-05-30	2014-07-11 07:52:00+00:00	None
114303950	USA	2019-06-05	2019-09-13 10:43:00+00:00	None
114303951	USA	2008-09-24	2008-09-24 16:26:00+00:00	None
114303953	USA	2021-10-28	2022-01-07 10:29:00+00:00	None

		delete_date	unmapped_location	onet_occupation_code	\
9967765	2016-04-20	09:50:00+00:00	0	43-4051.00	
9967766	2008-08-11	05:56:00+00:00	0	13-2072.00	
9967768	2022-05-26	03:21:00+00:00	0	29-2061.00	
9967769	2008-01-18	08:50:00+00:00	0	41-2031.00	
9967770	2016-02-19	07:20:00+00:00	0	41-4011.07	
...		...	...	...	
114303948	2020-01-09	13:57:00+00:00	0	53-7061.00	
114303949	2014-07-16	00:20:00+00:00	0	35-3031.00	
114303950	2019-09-15	17:06:00+00:00	0	31-1131.00	
114303951	2008-09-27	03:41:00+00:00	0	15-2031.00	
114303953	2022-01-12	20:16:00+00:00	0	43-4051.00	

	url	\
9967765	https://jobs.aarons.com/job/anderson/customer-service-representative/1618/1942568	
9967766	https://careers.peopleclick.com/careerscp/client_rbc/external/gateway.do?functionName=viewFromLi...	
9967768	https://jobs.trinity-health.org/job/TRHEUS00347492SAINTALPHONSUSENUS	
9967769	https://careersearch.sprint.com/1033/ASP/TG/cim_jobdetail.asp?jobId=573278&type=search&JobReqLan...	
9967770	https://solarcity.wd1.myworkdayjobs.com/en-US/01//job/Sacramento-CA/Outside-Solar-Sales-Consulta...	
...		
...		
114303948	https://recruiting.adp.com/srccar/public/nghome.guid?c=1151451&d=ExternalCareerSite&prc=RMP0D3&r...	
114303949	https://www.hrapply.com/mgmresorts/AppJobView.jsp?link=179871	
114303950	https://non-clinical-emory.icims.com/jobs/35655/transportation-attendant/job	
114303951	https://cp-its-rmprd.saic.com/main/careerportal/Job_Profile.cfm?/2IG26H2IHSEK02Z9JGK9YSJX3GOMTBZ...	
114303953	https://jobs.metlife.com/job/Omaha-Customer-Care-Advocates-Hybrid-Roles-%2420hour_-Launch-your-c...	

	company_id_updated	created_date	\
9967765	NaN	2016-04-09 06:14:00+00:00	
9967766	NaN	2008-08-02 14:44:00+00:00	
9967768	NaN	2022-03-27 04:26:00+00:00	

9967769	NaN	2008-01-16 20:50:00+00:00
9967770	NaN	2016-02-05 23:24:00+00:00
...	...	...
114303948	NaN	2019-10-17 20:35:00+00:00
114303949	NaN	2014-05-30 12:03:00+00:00
114303950	NaN	2019-06-05 04:08:00+00:00
114303951	NaN	2008-09-24 16:26:00+00:00
114303953	NaN	2021-10-28 15:48:00+00:00

	onet_occupation_code_NEW	Job Zone	year_month	factset_entity_id \
9967765	43-4051.00	2.0	2016-04	000BFT-E
9967766	13-2072.00	4.0	2008-08	0018FV-E
9967768	29-2061.00	3.0	2022-03	05LC20-E
9967769	41-2031.00	2.0	2008-01	0D7N1X-E
9967770	41-4011.07	3.0	2016-02	06KGBG-E
...	...	...	...	...
114303948	53-7061.00	2.0	2019-10	None
114303949	35-3031.00	2.0	2014-05	None
114303950	31-1131.00	3.0	2019-06	None
114303951	15-2031.00	5.0	2008-09	00FD97-E
114303953	43-4051.00	2.0	2021-10	None

	cusip	cusip_6	gvkey	the_key	year
9967765	002535102	002535	None	fskey_000BFT-E	2016
9967766	780087102	780087	015633	015633	2008
9967768	None	None	None	fskey_05LC20-E	2022
9967769	J7596P109	J7596P	None	fskey_0D7N1X-E	2008
9967770	83416T100	83416T	015314	015314	2016
...	...	...	...	...	...
114303948	None	None	None	cikey_36720	2019
114303949	None	None	None	cikey_22797	2014
114303950	None	None	None	cikey_26344	2019
114303951	525327102	525327	165123	165123	2008
114303953	None	None	None	cikey_20887	2021

[81671837 rows x 26 columns]

```
[5]: total4_batch1_simple = total4_batch1[["hash", "state", "created",
↪ "delete_date", "onet_occupation_code", "the_key", "year"]]
```

```
[6]: total4_batch1_simple
```

```
[6]:
```

	hash	state	created \
9967765	000000082858d2d0b2c0d9a130cdf016	SC	2016-04-09
9967766	0000003cf0ed2539cc4950a95f420261	CA	2008-08-02
9967768	000000491be59966a6408d201350c9ce	ID	2022-03-27
9967769	000000781fbcd4459bac5aa390464e47	NC	2008-01-16

9967770	00000090d1133192dd83e2ed9dc33b1c	CA	2016-02-05
...	...	...	...
114303948	7ffffff60ed07d167bbf16e2ced9d2dc2	VA	2019-10-17
114303949	7ffffff95e45d9fba64285161cd4f82fc	NV	2014-05-30
114303950	7ffffffc87badca288e755d9d7f801a16	GA	2019-06-05
114303951	7ffffffd6a96358ec2531a7bc7fad7303	VA	2008-09-24
114303953	7ffffffde7edd80bfc7eefbe0fa75c0c8	NE	2021-10-28

	delete_date	onet_occupation_code	the_key \
9967765	2016-04-20 09:50:00+00:00	43-4051.00	fskey_000BFT-E
9967766	2008-08-11 05:56:00+00:00	13-2072.00	015633
9967768	2022-05-26 03:21:00+00:00	29-2061.00	fskey_05LC20-E
9967769	2008-01-18 08:50:00+00:00	41-2031.00	fskey_0D7N1X-E
9967770	2016-02-19 07:20:00+00:00	41-4011.07	015314
...	...	...	...
114303948	2020-01-09 13:57:00+00:00	53-7061.00	cikey_36720
114303949	2014-07-16 00:20:00+00:00	35-3031.00	cikey_22797
114303950	2019-09-15 17:06:00+00:00	31-1131.00	cikey_26344
114303951	2008-09-27 03:41:00+00:00	15-2031.00	165123
114303953	2022-01-12 20:16:00+00:00	43-4051.00	cikey_20887

	year
9967765	2016
9967766	2008
9967768	2022
9967769	2008
9967770	2016
...	...
114303948	2019
114303949	2014
114303950	2019
114303951	2008
114303953	2021

[81671837 rows x 7 columns]

[6]:

	hash	state	created \
9967765	000000082858d2d0b2c0d9a130cdf016	SC	2016-04-09
9967766	0000003cf0ed2539cc4950a95f420261	CA	2008-08-02
9967768	000000491be59966a6408d201350c9ce	ID	2022-03-27
9967769	000000781fbcd4459bac5aa390464e47	NC	2008-01-16
9967770	00000090d1133192dd83e2ed9dc33b1c	CA	2016-02-05
...	...	...	...
114303948	7ffffff60ed07d167bbf16e2ced9d2dc2	VA	2019-10-17
114303949	7ffffff95e45d9fba64285161cd4f82fc	NV	2014-05-30
114303950	7ffffffc87badca288e755d9d7f801a16	GA	2019-06-05
114303951	7ffffffd6a96358ec2531a7bc7fad7303	VA	2008-09-24

114303953 7fffffde7edd80bfc7eefbe0fa75c0c8 NE 2021-10-28

		delete_date	onet_occupation_code	the_key	\
9967765	2016-04-20	09:50:00+00:00	43-4051.00	fskey_000BFT-E	
9967766	2008-08-11	05:56:00+00:00	13-2072.00	015633	
9967768	2022-05-26	03:21:00+00:00	29-2061.00	fskey_05LC20-E	
9967769	2008-01-18	08:50:00+00:00	41-2031.00	fskey_0D7N1X-E	
9967770	2016-02-19	07:20:00+00:00	41-4011.07	015314	
...		...	...	...	
114303948	2020-01-09	13:57:00+00:00	53-7061.00	cikey_36720	
114303949	2014-07-16	00:20:00+00:00	35-3031.00	cikey_22797	
114303950	2019-09-15	17:06:00+00:00	31-1131.00	cikey_26344	
114303951	2008-09-27	03:41:00+00:00	15-2031.00	165123	
114303953	2022-01-12	20:16:00+00:00	43-4051.00	cikey_20887	

	year
9967765	2016
9967766	2008
9967768	2022
9967769	2008
9967770	2016

	...
114303948	2019
114303949	2014
114303950	2019
114303951	2008
114303953	2021

[81671837 rows x 7 columns]

[6]:

	hash	state	created	\
9967765	000000082858d2d0b2c0d9a130cdf016	SC	2016-04-09	
9967766	0000003cf0ed2539cc4950a95f420261	CA	2008-08-02	
9967768	000000491be59966a6408d201350c9ce	ID	2022-03-27	
9967769	000000781fbc4459bac5aa390464e47	NC	2008-01-16	
9967770	00000090d1133192dd83e2ed9dc33b1c	CA	2016-02-05	
...	...	...	...	
114303948	7fffff60ed07d167bbf16e2ced9d2dc2	VA	2019-10-17	
114303949	7fffff95e45d9fba64285161cd4f82fc	NV	2014-05-30	
114303950	7fffffc87badca288e755d9d7f801a16	GA	2019-06-05	
114303951	7fffffd6a96358ec2531a7bc7fad7303	VA	2008-09-24	
114303953	7fffffde7edd80bfc7eefbe0fa75c0c8	NE	2021-10-28	

		delete_date	onet_occupation_code	the_key	\
9967765	2016-04-20	09:50:00+00:00	43-4051.00	fskey_000BFT-E	
9967766	2008-08-11	05:56:00+00:00	13-2072.00	015633	
9967768	2022-05-26	03:21:00+00:00	29-2061.00	fskey_05LC20-E	

9967769	2008-01-18 08:50:00+00:00	41-2031.00	fskey_OD7N1X-E
9967770	2016-02-19 07:20:00+00:00	41-4011.07	015314
...	...	...	...
114303948	2020-01-09 13:57:00+00:00	53-7061.00	cikey_36720
114303949	2014-07-16 00:20:00+00:00	35-3031.00	cikey_22797
114303950	2019-09-15 17:06:00+00:00	31-1131.00	cikey_26344
114303951	2008-09-27 03:41:00+00:00	15-2031.00	165123
114303953	2022-01-12 20:16:00+00:00	43-4051.00	cikey_20887

	year
9967765	2016
9967766	2008
9967768	2022
9967769	2008
9967770	2016

...	...
114303948	2019
114303949	2014
114303950	2019
114303951	2008
114303953	2021

[81671837 rows x 7 columns]

[6]:

	hash	state	created	\
9967765	000000082858d2d0b2c0d9a130cdf016	SC	2016-04-09	
9967766	0000003cf0ed2539cc4950a95f420261	CA	2008-08-02	
9967768	000000491be59966a6408d201350c9ce	ID	2022-03-27	
9967769	000000781fbcd4459bac5aa390464e47	NC	2008-01-16	
9967770	00000090d1133192dd83e2ed9dc33b1c	CA	2016-02-05	
...	...	...	...	...
114303948	7fffff60ed07d167bbf16e2ced9d2dc2	VA	2019-10-17	
114303949	7fffff95e45d9fba64285161cd4f82fc	NV	2014-05-30	
114303950	7fffffc87badca288e755d9d7f801a16	GA	2019-06-05	
114303951	7fffffd6a96358ec2531a7bc7fad7303	VA	2008-09-24	
114303953	7fffffde7edd80bfc7eefbe0fa75c0c8	NE	2021-10-28	

	delete_date	onet_occupation_code	the_key	\
9967765	2016-04-20 09:50:00+00:00	43-4051.00	fskey_000BFT-E	
9967766	2008-08-11 05:56:00+00:00	13-2072.00	015633	
9967768	2022-05-26 03:21:00+00:00	29-2061.00	fskey_05LC20-E	
9967769	2008-01-18 08:50:00+00:00	41-2031.00	fskey_OD7N1X-E	
9967770	2016-02-19 07:20:00+00:00	41-4011.07	015314	
...	...	...	...	...
114303948	2020-01-09 13:57:00+00:00	53-7061.00	cikey_36720	
114303949	2014-07-16 00:20:00+00:00	35-3031.00	cikey_22797	
114303950	2019-09-15 17:06:00+00:00	31-1131.00	cikey_26344	

114303951	2008-09-27 03:41:00+00:00	15-2031.00	165123
114303953	2022-01-12 20:16:00+00:00	43-4051.00	cikey_20887

	year
9967765	2016
9967766	2008
9967768	2022
9967769	2008
9967770	2016
...	...
114303948	2019
114303949	2014
114303950	2019
114303951	2008
114303953	2021

[81671837 rows x 7 columns]

```
[7]: total4_batch2 = pd.read_pickle("Processed_Data/
    ↪20220618US_Job_Record_Final_batch2.pkl")
total4_batch2['year'] = pd.DatetimeIndex(total4_batch2['created']).year
```

```
[8]: total4_batch2_simple = total4_batch2[["hash", "state", "created",
    ↪"delete_date", "onet_occupation_code", "the_key", "year"]]
```

```
[9]: total4_batch1_simple = total4_batch1_simple.append(total4_batch2_simple)
```

```
[10]: total4_batch1_simple.to_csv("Processed_Data/20220626US_Job_Record_FillingTime.
    ↪csv", index = False)
```

```
[11]: total4_batch1_simple.to_pickle("Processed_Data/
    ↪20220626US_Job_Record_FillingTime.pkl")
```

10 Process Job Record Filling Time data

```
[61]: df1 = pd.read_pickle("Processed_Data/20220626US_Job_Record_FillingTime.pkl")
```

```
[67]: df1
```

```
[67]:
```

	hash	state	created	\
9967765	000000082858d2d0b2c0d9a130cdf016	SC	2016-04-09	
9967766	0000003cf0ed2539cc4950a95f420261	CA	2008-08-02	
9967768	000000491be59966a6408d201350c9ce	ID	2022-03-27	
9967769	000000781fbcd4459bac5aa390464e47	NC	2008-01-16	
9967770	00000090d1133192dd83e2ed9dc33b1c	CA	2016-02-05	
...	...	...	...	...
114362810	fffffcdbd8ce7b19363bf7303589cf98	WA	2016-12-16	

114362812	ffffffe08c5113049ed1db60b93d18ab	CA	2021-06-11
114362813	ffffffeb9c8de1e01ac35c82f791caaf	TX	2015-02-15
114362814	ffffff9e71ca2624e47b2735516024b	NJ	2022-04-10
114362815	ffffffd4a3e00cba52d6f06af79b986	PA	2021-07-28

	delete_date	onet_occupation_code	the_key	\
9967765	2016-04-20 09:50:00+00:00	43-4051.00	fskey_000BFT-E	
9967766	2008-08-11 05:56:00+00:00	13-2072.00	015633	
9967768	2022-05-26 03:21:00+00:00	29-2061.00	fskey_05LC20-E	
9967769	2008-01-18 08:50:00+00:00	41-2031.00	fskey_0D7N1X-E	
9967770	2016-02-19 07:20:00+00:00	41-4011.07	015314	
...	...	...	...	
114362810	2017-04-20 09:50:00+00:00	23-1011.00	064768	
114362812	2021-09-22 00:11:00+00:00	13-2051.00	fskey_05NY4J-E	
114362813	2015-02-28 02:50:00+00:00	15-1253.00	cikey_21617	
114362814	2022-04-22 07:42:00+00:00	49-9071.00	fskey_061R9D-E	
114362815	None	41-1011.00	cikey_24006	

	year
9967765	2016
9967766	2008
9967768	2022
9967769	2008
9967770	2016
...	...
114362810	2016
114362812	2021
114362813	2015
114362814	2022
114362815	2021

[163171800 rows x 7 columns]

```
[89]: df1_temp = df1
```

```
[90]: df1_temp.loc[df1_temp["delete_date"].isna(), "delete_date"] = "2099-12-31 09:20:
      ↪00+00:00"
df1_temp
```

	hash	state	created	\
9967765	000000082858d2d0b2c0d9a130cdf016	SC	2016-04-09	
9967766	0000003cf0ed2539cc4950a95f420261	CA	2008-08-02	
9967768	000000491be59966a6408d201350c9ce	ID	2022-03-27	
9967769	000000781fbcd4459bac5aa390464e47	NC	2008-01-16	
9967770	00000090d1133192dd83e2ed9dc33b1c	CA	2016-02-05	
...	...	...	...	
114362810	ffffffcbd8ce7b19363bf7303589cf98	WA	2016-12-16	

114362812	fffffffe08c5113049ed1db60b93d18ab	CA	2021-06-11
114362813	ffffffeb9c8de1e01ac35c82f791caaf	TX	2015-02-15
114362814	ffffff9e71ca2624e47b2735516024b	NJ	2022-04-10
114362815	fffffffd4a3e00cba52d6f06af79b986	PA	2021-07-28

	delete_date	onet_occupation_code	the_key	\
9967765	2016-04-20 09:50:00+00:00	43-4051.00	fskey_000BFT-E	
9967766	2008-08-11 05:56:00+00:00	13-2072.00	015633	
9967768	2022-05-26 03:21:00+00:00	29-2061.00	fskey_05LC20-E	
9967769	2008-01-18 08:50:00+00:00	41-2031.00	fskey_0D7N1X-E	
9967770	2016-02-19 07:20:00+00:00	41-4011.07	015314	
...	...	...	...	
114362810	2017-04-20 09:50:00+00:00	23-1011.00	064768	
114362812	2021-09-22 00:11:00+00:00	13-2051.00	fskey_05NY4J-E	
114362813	2015-02-28 02:50:00+00:00	15-1253.00	cikey_21617	
114362814	2022-04-22 07:42:00+00:00	49-9071.00	fskey_061R9D-E	
114362815	2099-12-31 09:20:00+00:00	41-1011.00	cikey_24006	

	year
9967765	2016
9967766	2008
9967768	2022
9967769	2008
9967770	2016
...	...
114362810	2016
114362812	2021
114362813	2015
114362814	2022
114362815	2021

[163171800 rows x 7 columns]

```
[91]: df1_temp["delete_date2"] = df1_temp["delete_date"].apply(lambda x:x[0:10])
df1_temp['delete_date2'] = pd.to_datetime(df1_temp['delete_date2'])
df1_temp['created_date2'] = pd.to_datetime(df1_temp['created'])
df1_temp['filling_days'] = df1_temp['delete_date2'] - df1_temp['created_date2']
df1_temp['filling_days'] = df1_temp['filling_days'].dt.days
df1_temp
```

	hash	state	created	\
9967765	000000082858d2d0b2c0d9a130cdf016	SC	2016-04-09	
9967766	0000003cf0ed2539cc4950a95f420261	CA	2008-08-02	
9967768	000000491be59966a6408d201350c9ce	ID	2022-03-27	
9967769	000000781fbc4459bac5aa390464e47	NC	2008-01-16	
9967770	00000090d1133192dd83e2ed9dc33b1c	CA	2016-02-05	
...	...	...	...	

114362810	ffffffcbd8ce7b19363bf7303589cf98	WA	2016-12-16
114362812	ffffffe08c5113049ed1db60b93d18ab	CA	2021-06-11
114362813	ffffffeb9c8de1e01ac35c82f791caaf	TX	2015-02-15
114362814	ffffff9e71ca2624e47b2735516024b	NJ	2022-04-10
114362815	ffffffd4a3e00cba52d6f06af79b986	PA	2021-07-28

		delete_date	onet_occupation_code	the_key \
9967765	2016-04-20	09:50:00+00:00	43-4051.00	fskey_000BFT-E
9967766	2008-08-11	05:56:00+00:00	13-2072.00	015633
9967768	2022-05-26	03:21:00+00:00	29-2061.00	fskey_05LC20-E
9967769	2008-01-18	08:50:00+00:00	41-2031.00	fskey_0D7N1X-E
9967770	2016-02-19	07:20:00+00:00	41-4011.07	015314
...	...	...	...	...
114362810	2017-04-20	09:50:00+00:00	23-1011.00	064768
114362812	2021-09-22	00:11:00+00:00	13-2051.00	fskey_05NY4J-E
114362813	2015-02-28	02:50:00+00:00	15-1253.00	cikey_21617
114362814	2022-04-22	07:42:00+00:00	49-9071.00	fskey_061R9D-E
114362815	2099-12-31	09:20:00+00:00	41-1011.00	cikey_24006

	year	delete_date2	created_date2	filling_days
9967765	2016	2016-04-20	2016-04-09	11
9967766	2008	2008-08-11	2008-08-02	9
9967768	2022	2022-05-26	2022-03-27	60
9967769	2008	2008-01-18	2008-01-16	2
9967770	2016	2016-02-19	2016-02-05	14
...	...	...	...	...
114362810	2016	2017-04-20	2016-12-16	125
114362812	2021	2021-09-22	2021-06-11	103
114362813	2015	2015-02-28	2015-02-15	13
114362814	2022	2022-04-22	2022-04-10	12
114362815	2021	2099-12-31	2021-07-28	28645

[163171800 rows x 10 columns]

```
[92]: df1_temp_clean = df1_temp.drop(['created', 'delete_date',
    ↪ "onet_occupation_code", "state", "year"], axis=1)
df1_temp_clean["filled"] = 1
df1_temp_clean.loc[df1_temp_clean["filling_days"]>20000, "filled"] = 0
df1_temp_clean.loc[df1_temp_clean["filling_days"]>20000, "filling_days"] = None
df1_temp_clean
```

```
[92]:
```

	hash	the_key	delete_date2 \
9967765	000000082858d2d0b2c0d9a130cdf016	fskey_000BFT-E	2016-04-20
9967766	0000003cf0ed2539cc4950a95f420261	015633	2008-08-11
9967768	000000491be59966a6408d201350c9ce	fskey_05LC20-E	2022-05-26
9967769	000000781fbcd4459bac5aa390464e47	fskey_0D7N1X-E	2008-01-18
9967770	00000090d1133192dd83e2ed9dc33b1c	015314	2016-02-19

...	...	...	...
114362810	ffffffcbd8ce7b19363bf7303589cf98	064768	2017-04-20
114362812	ffffffe08c5113049ed1db60b93d18ab	fskey_05NY4J-E	2021-09-22
114362813	ffffffeb9c8de1e01ac35c82f791caaf	cikey_21617	2015-02-28
114362814	ffffff9e71ca2624e47b2735516024b	fskey_061R9D-E	2022-04-22
114362815	fffffffd4a3e00cba52d6f06af79b986	cikey_24006	2099-12-31

	created_date2	filling_days	filled
9967765	2016-04-09	11.0	1
9967766	2008-08-02	9.0	1
9967768	2022-03-27	60.0	1
9967769	2008-01-16	2.0	1
9967770	2016-02-05	14.0	1
...	...	...	...
114362810	2016-12-16	125.0	1
114362812	2021-06-11	103.0	1
114362813	2015-02-15	13.0	1
114362814	2022-04-10	12.0	1
114362815	2021-07-28	NaN	0

[163171800 rows x 6 columns]

```
[93]: df1_temp_clean_p = df1_temp_clean[df1_temp_clean["the_key"].apply(lambda x: re.
      ↪search("_", x) == None) == True]
df1_temp_clean_p
```

```
[93]:
```

	hash	the_key	delete_date2	\
9967766	0000003cf0ed2539cc4950a95f420261	015633	2008-08-11	
9967770	00000090d1133192dd83e2ed9dc33b1c	015314	2016-02-19	
9967774	000000be62247aa211a827915bbd1ad1	028877	2020-09-16	
9967776	000000c11b16e4556fb058b041e5f630	065417	2019-11-10	
9967782	000001125c8bde7c446643f50211a129	006829	2018-03-17	
...	...	...	...	
114362796	ffffffeddd0f27dd90366a32e55f77ee3	014225	2014-02-03	
114362799	ffffff5329a837b8e89fb7dc5491afaa	005643	2015-08-23	
114362801	ffffff5a888986e1bdc932a935b066e4	176249	2021-08-25	
114362804	ffffff8728f125f04f0560bcfbf28c9f	009899	2014-04-30	
114362810	ffffffcbd8ce7b19363bf7303589cf98	064768	2017-04-20	

	created_date2	filling_days	filled
9967766	2008-08-02	9.0	1
9967770	2016-02-05	14.0	1
9967774	2020-07-30	48.0	1
9967776	2019-11-03	7.0	1
9967782	2017-12-18	89.0	1
...	...	...	...
114362796	2014-01-17	17.0	1

114362799	2015-07-20	34.0	1
114362801	2021-07-10	46.0	1
114362804	2014-02-12	77.0	1
114362810	2016-12-16	125.0	1

[64314018 rows x 6 columns]

```
[94]: df1_temp_clean.to_pickle("Processed_Data/  
      ↪20220909US_Job_Record_FillingTime_Clean.pkl")  
df1_temp_clean_p.to_pickle("Processed_Data/  
      ↪20220909US_Job_Record_FillingTime_Public_Clean.pkl")
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```