

RP Data-Merge

November 22, 2025

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

[2]: job_record_sum_output_dir = "D:/User.Data/Dan.Li/RP LinkUp_Sum/"

[3]: # List to store DataFrames
dfs = []

# Iterate over all files in the directory
for filename in os.listdir(directory):
    if filename.endswith("_sum.pkl"):
        # Load the pkl file and append it to the list
        file_path = os.path.join(directory, filename)
        print(file_path)
        df = pd.read_pickle(file_path)
        dfs.append(df)

# Merge all DataFrames in the list
merged_df = pd.concat(dfs, ignore_index=True)

# Save the merged DataFrame to a new file
output_file = os.path.join(directory, "merged_sum.pkl")
merged_df.to_pickle(output_file)

print(f"Merged file saved to {output_file}")
```

```
D:/User.Data/Dan.Li/RP LinkUp_Sum/2007-08_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2007-09_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2007-10_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2007-11_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2007-12_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2008-01_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2008-02_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2008-03_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2008-04_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2008-05_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2008-06_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2008-07_sum.pkl
```

[illegible]

[illegible]

[illegible]

```

D:/User.Data/Dan.Li/RP LinkUp_Sum/2020-08_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2020-09_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2020-10_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2020-11_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2020-12_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-01_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-02_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-03_sum.pkl

D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-04_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-05_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-06_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-07_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-08_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-09_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-10_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-11_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2021-12_sum.pkl
D:/User.Data/Dan.Li/RP LinkUp_Sum/2022-01_sum.pkl
Merged file saved to D:/User.Data/Dan.Li/RP LinkUp_Sum/merged_sum.pkl

```

1 Merge job record

```

[1]: import pandas as pd

[2]: merged_df = pd.read_pickle("D:/User.Data/Dan.Li/RP LinkUp_Sum/merged_sum.pkl")

[3]: import pickle5 as pickle
    with open("Processed_Data/20220607US_Job_Record_batch1.pkl", "rb") as fh:
        data = pickle.load(fh)

[4]: import pickle5 as pickle
    with open("Processed_Data/20220607US_Job_Record_batch2.pkl", "rb") as fh:
        data_batch2 = pickle.load(fh)

[6]: # PROVIDER_DOCUMENT_CHAIN_ID is hash in LinkUp
    data2 = data.merge(merged_df, left_on = ["hash"], right_on =
        ↳ ["PROVIDER_DOCUMENT_CHAIN_ID"])

[7]: data_batch2_2 = data_batch2.merge(merged_df, left_on = ["hash"], right_on =
        ↳ ["PROVIDER_DOCUMENT_CHAIN_ID"])

[8]: data_batch2_2

[8]:
                                hash \
0          8000003e624f858208a2017becb57bcd
1          8000005153ff7de94c7db903c403bbf4

```

```

2      800000df74dee5f8358f8b787370b975
3      800000f38c8e16e2df2f6f70bd6235b3
4      800000f38c8e16e2df2f6f70bd6235b3
...
74815979 ff6fdaea6b3c31643623b56310b00ead
74815980 ff6fdaf5139deeda56527abb4207f81e
74815981 ff6fdaf5139deeda56527abb4207f81e
74815982 ff6fdb04ef3d737a2b737811160a7a5d
74815983 ff6fdb4e02ab6d4c8d9ac952ac09dec1

```

	title	company_id \
0	Retail Coverage Merchandiser - Walmart	19555
1	Software Engineer	9271
2	Biomedical Flight Controller (Houston, TX)	22381
3	Registered Polysomnographic Technologist - (PS...	24764
4	Registered Polysomnographic Technologist - (PS...	24764
...	...	...
74815979	Managing Director, Consumer/Retail Investment ...	19010
74815980	Management Trainee - FT	24954
74815981	Management Trainee - FT	24954
74815982	Day Stocker	36335
74815983	Sort Observation for Package Handler Positions	17354

	company_name	city	state \
0	Acosta, Inc.	lamar	CO
1	Raytheon Technologies Corp	schriever air force base	CO
2	Wyle Laboratories, Inc.	Houston	TX
3	AMN Healthcare Services, Inc.	little rock	AR
4	AMN Healthcare Services, Inc.	little rock	AR
...	...	...	...
74815979	Wedbush Morgan Securities	new york	NY
74815980	Southeastern Freight Lines	Knoxville	TN
74815981	Southeastern Freight Lines	Knoxville	TN
74815982	Food City	wise	VA
74815983	FedEx	Cheyenne	WY

	zip	country	created	last_checked \
0	81052	USA	2020-11-15 06:30:00+00:00	2020-11-23 11:07:00+00:00
1	NaN	USA	2021-09-10 08:19:00+00:00	2021-09-21 13:25:00+00:00
2	77020	USA	2016-07-06 22:35:00+00:00	2016-07-11 00:25:00+00:00
3	72215	USA	2018-02-25 14:57:00+00:00	2018-03-08 11:27:00+00:00
4	72215	USA	2018-02-25 14:57:00+00:00	2018-03-08 11:27:00+00:00
...	...	...	...	...
74815979	10007	USA	2020-11-14 07:07:00+00:00	2020-12-16 15:22:00+00:00
74815980	37902	USA	2015-05-05 10:32:00+00:00	2017-01-20 13:27:00+00:00
74815981	37902	USA	2015-05-05 10:32:00+00:00	2017-01-20 13:27:00+00:00
74815982	24293	USA	2021-11-21 08:01:00+00:00	2021-11-23 08:06:00+00:00

74815983 82003 USA 2016-07-01 21:31:00+00:00 2016-07-04 21:55:00+00:00

	...	Health & Wellness Benefits	Other Employment Perks	\
0	...	NaN	NaN	
1	...	2.0	1.0	
2	...	NaN	NaN	
3	...	2.0	NaN	
4	...	2.0	NaN	
...	...	...	...	
74815979	...	NaN	NaN	
74815980	...	NaN	NaN	
74815981	...	NaN	NaN	
74815982	...	2.0	NaN	
74815983	...	NaN	NaN	

	Personality Traits	Retirement Benefits	Soft Skills	\
0	1.0	1.0	9.0	
1	1.0	1.0	4.0	
2	NaN	NaN	10.0	
3	NaN	1.0	NaN	
4	NaN	1.0	NaN	
...	...	...	...	
74815979	NaN	NaN	4.0	
74815980	NaN	NaN	5.0	
74815981	NaN	NaN	5.0	
74815982	2.0	1.0	2.0	
74815983	2.0	NaN	6.0	

	Sustainability Knowledge	Technical Skills	Technology Skills	\
0	NaN	2.0	1.0	
1	NaN	1.0	13.0	
2	NaN	NaN	6.0	
3	NaN	2.0	NaN	
4	NaN	2.0	NaN	
...	...	...	...	
74815979	NaN	4.0	2.0	
74815980	NaN	1.0	NaN	
74815981	NaN	1.0	NaN	
74815982	NaN	2.0	NaN	
74815983	NaN	NaN	NaN	

	Time Off Benefits	Work Values
0	1.0	1.0
1	1.0	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN

...	...	...
74815979	NaN	NaN
74815980	NaN	2.0
74815981	NaN	2.0
74815982	NaN	NaN
74815983	2.0	2.0

[74815984 rows x 38 columns]

```
[9]: data3 = data2.append(data_batch2_2)
```

```
[10]: data3.to_pickle("df_record_withRP.pkl")
```

```
[ ]:
```

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

```
[2]: data3 = pd.read_pickle("df_record_withRP.pkl")
```

```
[3]: import pickle5 as pickle
with open("Processed_Data/20220811US_Job_Records_DateOnetKeyMSA.pkl", "rb") as f:
    fh:
        df_otherinfo = pickle.load(fh)
```

```
[4]: df_otherinfo.head()
```

```
[4]:
```

	hash	onet_occupation_code_NEW	created	\
0	000000082858d2d0b2c0d9a130cdf016	43-4051.00	2016-04-09	
1	0000003cf0ed2539cc4950a95f420261	13-2072.00	2008-08-02	
2	000000491be59966a6408d201350c9ce	29-2061.00	2022-03-27	
3	000000781fbcd4459bac5aa390464e47	41-2031.00	2008-01-16	
4	00000090d1133192dd83e2ed9dc33b1c	41-4011.07	2016-02-05	

	year_month	gvkey	the_key	Job	Zone	MSA_code	\
0	2016-04	None	fskey_000BFT-E		2.0	24860	
1	2008-08	015633	015633		4.0	31080	
2	2022-03	None	fskey_05LC20-E		3.0	14260	
3	2008-01	None	fskey_0D7N1X-E		2.0	24660	
4	2016-02	015314	015314		3.0	40900	

	MSA_name	MSA_type
0	Greenville-Anderson, SC	metropolitan
1	Los Angeles-Long Beach-Anaheim, CA	metropolitan
2	Boise City, ID	metropolitan
3	Greensboro-High Point, NC	metropolitan
4	Sacramento-Roseville-Folsom, CA	metropolitan

```
[5]: data3 = data3.merge(df_otherinfo, left_on = ["hash"], right_on = ["hash"])

[6]: data3.to_pickle("df_record_withRP2.pkl")

[7]: # Generate dummy variable
      data3['d_public'] = data3['the_key'].apply(lambda x: 0 if "_" in x else 1)

[8]: data3_public = data3[data3["d_public"] == 1]

[9]: data3_public.to_pickle("df_record_withRP2_public.pkl")

[10]: data3_public.to_csv("df_record_withRP2_public.csv")
```