

DEI Jobs Step 3 - Merge Small Sample

November 22, 2025

1 Intro

This code create a smaller version data to make it more managable

```
[1]: import nltk
import numpy as np
import pandas as pd
import pickle
import pprint
```

2 Data Preset

2.1 Shrink the dataset by selecting a relevant period.

```
[2]: jp1 = pd.read_pickle("C:/Users/dli/Downloads/job_descriptions_chosen_4-0.pkl")
jp2 = pd.read_pickle("C:/Users/dli/Downloads/job_descriptions_chosen_9-5.pkl")
jp3 = pd.read_pickle("C:/Users/dli/Downloads/job_descriptions_chosen_f-a.pkl")
jp1 = jp1.append(jp2)
jp1 = jp1.append(jp3)
len(jp1)
jp1.to_pickle("C:/Users/dli/Downloads/job_descriptions_chosen_full.pkl")
```

```
[3]: job_des_chosen = pd.read_pickle("Processed_Data/job_descriptions_chosen_full.
↪pkl")
```

```
[4]: job_record_chosen = pd.read_pickle("C:/Users/dli/Downloads/
↪total_job_posts_chosen.pkl")
```

```
[5]: job_record_chosen2 = job_record_chosen.merge(jp1, on = ['hash'], how = 'left')
job_record_chosen2 = job_record_chosen2[job_record_chosen2["description"].
↪notna()]
job_record_chosen2.to_pickle("Processed_Data/job_chosen_merged_full.pkl")
```

```
[6]: # Make the sample smaller
jp1 = pd.read_pickle("Processed_Data/job_chosen_merged_full.pkl")
jp_temp = jp1
jp_temp = jp_temp[jp_temp["created_date"]>="2019-06-30"]
jp_temp = jp_temp[jp_temp["created_date"]<="2022-06-30"]
```

```
jp_temp.to_pickle("Processed_Data/job_chosen_merged_small.pkl")  
len(jp_temp)
```

[6]: 20248699

```
[2]: jp_temp = pd.read_pickle("Processed_Data/job_chosen_merged_small.pkl")
```

```
[3]: jp_temp2 = jp_temp
```

```
[4]: jp_temp2.to_pickle("Processed_Data/jp_dimensions_medium_simple_new.pkl")
```

```
[5]: jp_temp2.to_csv("Processed_Data/jp_dimensions_medium_simple_new.csv", index =  
↪False)
```

```
[ ]:
```