HRS Data Exploration

Maria Navarro

```
import pandas as pd
import matplotlib.pyplot as plt

pensions = pd.read_csv('/Users/marianavarro/Desktop/pdi_pensions/csv_files/h22j2_p.csv')
```

Notable Variables

SJ2Z503 - Job Type

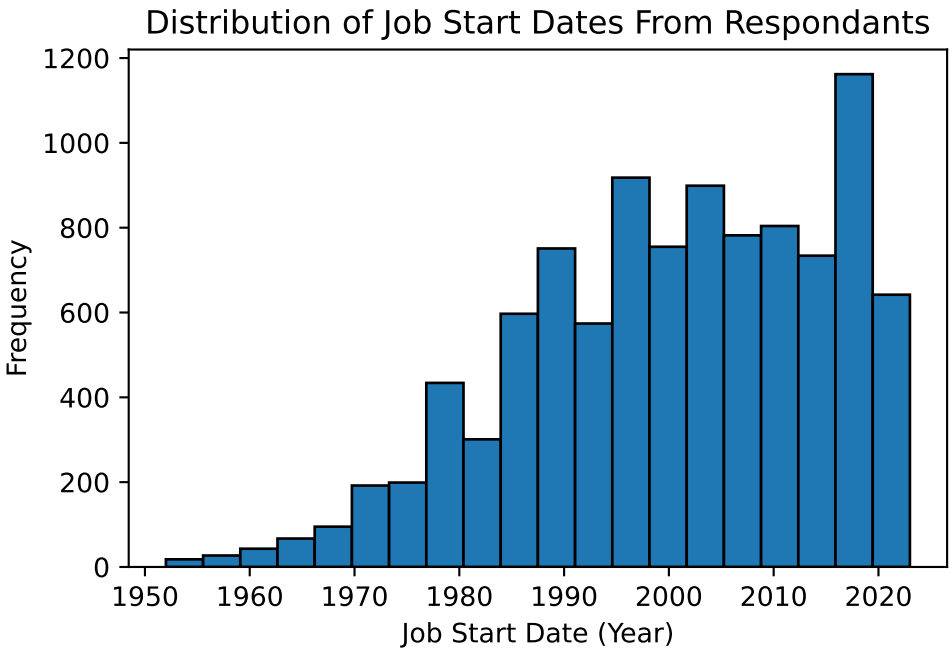
```
1.  DESIGNATED PREVIOUS WAVE JOB      3316
2.  OLD JOB                           3489
Blank.  Newly added past pension      3866
```

SJ2W410 - Year Job - Begin

```
This column will need to be cleaned as there are some job start years as low as -
8 and as high as 9999.
```

```
pensions['SJ2W410'].plot.hist(bins = 20, range = (1952, 2023), edgecolor = 'black')
plt.xlabel('Job Start Date (Year)')
plt.title('Distribution of Job Start Dates From Respondants')
print(pensions['SJ2W410'].max())
print(pensions['SJ2W410'].min())
```

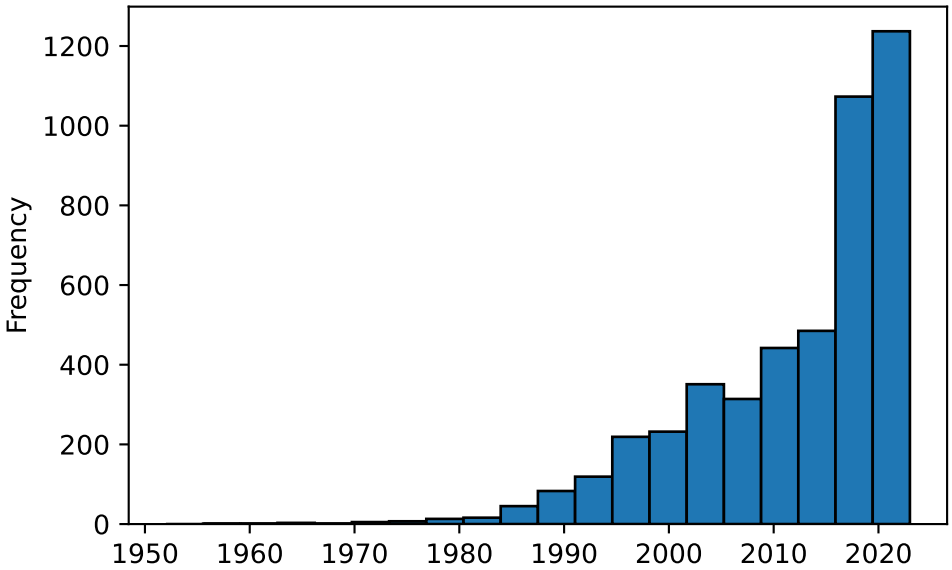
```
9999.0
-8.0
```



SJ2W411 - Year Job - End

This column will also need to be cleaned for same issue as Year Job - Begin.

```
pensions['SJ2W411'].plot.hist(bins = 20, range = (1952, 2023), edgecolor = 'black')
```



SJ2W405 - Plan Provider

```
                              count
-8. Web non-response           48
1.  PRIVATE EMPLOYER           5095
2.  GOVERNMENT EMPLOYER        2533
3.  R'S OWN BUSINESS           151
4.  UNION                      495
97. OTHER                      430
98. DK (Don't Know);           89
    NA (Not Ascertained)
99. RF (Refused)               37
Blank.  INAP (Inapplicable); 1793
        Partial Interview
```

```
target_variables = pensions[['HHID', 'PN', 'SJ2W410', 'SJ2W411']]
target_variables = target_variables.rename(columns = {'SJ2W410': 'Year Job - Begin', 'SJ2W41
target_variables['HHID'] = target_variables['HHID'].astype(str)
target_variables['PN'] = target_variables['PN'].astype(str)
target_variables['HHID + PN'] = target_variables['HHID'] + target_variables['PN']
target_variables
```

	HHID	PN	Year Job - Begin	Year Job - End	HHID + PN
0	10038	40	1989.0	1996.0	1003840
1	10038	40	1996.0	2014.0	1003840
2	10059	30	2004.0	NaN	1005930
3	10059	30	2004.0	NaN	1005930
4	10059	30	NaN	2004.0	1005930
...	...	...	...	...	...
10666	923489	10	2012.0	NaN	92348910
10667	923489	20	2017.0	NaN	92348920
10668	923497	20	1999.0	NaN	92349720
10669	923497	20	1999.0	NaN	92349720
10670	952836	10	2011.0	NaN	95283610

```
pensions['PN'].value_counts()
```

```
PN
10    6301
20    3603
```

```
11     349
40     158
21     105
30      86
12      22
31      16
41      16
22      10
32       4
33       1
Name: count, dtype: int64
```