Web Scraping Job Postings

Background

This project forms part of the DSI+ course conducted by General Assembly, Melbourne from February 2018. The brief of the project is as follows.

You're working as a data scientist for a contracting firm that's rapidly expanding. Now that they have their most valuable employee (you!), they need to leverage data to win more contracts. The firm offers technology and scientific solutions and wants to be competitive in the hiring market.

Business Problem

Business principal has two main objectives:

Determine the industry factors that are most important in predicting the salary amounts for these data.

Determine the factors that distinguish job categories and titles from each other. For example, can required skills accurately predict job title?

To limit the scope, focus on data-related job postings, e.g. data scientist, data analyst, research scientist, business intelligence, and any others you might think of.

Web Site

Australian job posting company Seek.com operates job posting websites in 18 countries, including Australia. The Australian website was selected for web scraping job details for the following search criteria

Data Scientist

Data Analyst

Data Engineer

Business Intelligence

The website has the following structure for its content

Each posted job has a unique Job ID

A summary page that lists 20 jobs per page with summary details for the job including Job ID. The typical URL for summary page is like *https://www.seek.com.au/data-engineer-jobs/in-All-Australia?page=2 (https://www.seek.com.au/data-engineer-jobs/in-All-Australia?page=2)*

A details page that can be navigated from the summary page using a link embedded in the summary. The typical URL for detail page is like *https://www.seek.com.au/job/36226295 (https://www.seek.com.au/job/36226295)* where the number is the Job ID

Data Dictionary

The following data items were scraped from the websites

Job ID	Unique ID for posted job

Jobclass	Job Classification assigned as per search term. Values are: DS, DA, DE, BI
Salary	Salary / Contract Rate. This is a free-text field. Data was cleaned to get values
Jobdate	Date job was posted
Title	Job Title
Location	City in Australia. Eg: Sydney, Melbourne
Suburb	Suburb class within city. Eg: CBD, North Melbourne
Category	Job category assigned by job poster. Eg: IT&T, Marketing
Subcategory	Job sub-category assigned by job poster. Eg: Customer Analysis
Worktype	Type of work. Eg: Fulltime, Parttime, Contract, Casual
Company	Organisation posting the job. Could be a placement agency or an employer
Jobtext	Free text description of the job

Web Scraping Steps

The following steps were performed for scraping data

```
Iterate - for each <jobclass>:
    Iterate - for each <summary page>:
        Get list of <JobID>
        Iterate - for each <JobID>:
            Get data from summary page
            Navigate to detail page
            Get data from detail page
```

In [1]:

```
import requests
from bs4 import BeautifulSoup as bs
import pandas as pd
import numpy as np
import csv
import seaborn as sn
```

Functions to extract data using HTML elements and attributes

In [2]:

```
# Get all text for the summary page and extract jobids on the page
def get_jobids(summary):
    article = summary.find_all(name='article')
    for x in article:
        joblist.append(x['data-job-id'])
    return
```

In [3]:

```
# Check if next page block is there on the page. If not found then it is the last page for
def check_next_page(summary):
    try:
        x = summary.find('a', {'data-automation':'page-next'}).text
    except:
        x='NA'
    return x
```

In [4]:

```
# Get job salary from the summary page
def get_salary(jobdet):
    try:
        return jobdet.find('span', {'data-automation':'jobSalary'}).text
    except:
        return 'NA'
```

In [5]:

```
# Get job posted date from the summary page
def get_job_dd(jobdet):
    try:
        x=jobdet.find('dd', {'data-automation':'job-detail-date'}).text
    except:
        x='NA'
    return x
```

In [6]:

```
# Get job title from the summary page
def get_job_title(jobdet):
    try:
        x=jobdet.find('a', {'data-automation':'jobTitle'}).text
    except:
        x='NA'
    return x
```

In [7]:

```
# Get job location from the summary page
def get_job_location(jobdet):
    try:
        x=jobdet.find('a', {'data-automation':'jobLocation'}).text
    except:
        x='NA'
    return x
```

In [8]:

```
# Get job suburb from the summary page
def get_job_suburb(jobdet):
    try:
        x=jobdet.find('a', {'data-automation':'jobArea'}).text
    except:
        x='NA'
    return x
```

In [9]:

```
# Get job classification from the details page
def get_job_category(jobdet):
    try:
        x=jobdet.find('a', {'data-automation':'jobClassification'}).text
    except:
        x='NA'
    return x
```

In [10]:

```
# Get job sub classification from the details page
def get_job_subcategory(jobdet):
    try:
        x=jobdet.find('a', {'data-automation':'jobSubClassification'}).text
    except:
        x='NA'
    return x
```

In [11]:

```
def get_job_worktype(jobdet):
    try:
        x=jobdet.find('dd', {'data-automation':'job-detail-work-type'}).text
    except:
        x='NA'
    return x
```

In [12]:

```
# Get job posting company from the details page
def get_job_company(jobdet):
    try:
        x=jobdet.find('a', {'data-automation':'jobCompany'}).text
    except:
        x='NA'
    return x
```

In [13]:

```
# Get job free text description from the details page
def get_job_text(jobdet):
    try:
#         x=jobdet.find('div', {'data-automation':'jobDescription'}).text
        x=jobdet.find('div', {'class':'templatetext'}).text
    except:
        x='NA'
    return x
```

Get details from detail page for the job

In [14]:

```
# Go to job details page for job id and get details
def get_dets_from_detail(id):
    det=''
    url = 'https://www.seek.com.au/job/'+ id
#    print(url)
    r = requests.get(url)
    sc = r.status_code
    if sc != 200:
        print('Error',sc,'opening page for jobid', id)
        detail = ''
    else:
        detail = bs(r.text,'lxml')
    jobdate.append(get_job_dd(detail))
    worktype.append(get_job_worktype(detail))
    jobtext.append(get_job_text(detail))
    return
```

Get details from Summary Page

In [15]:

```
# Get details for each job from the current SUMMARY page
def get_details(pagetext,jobc):
    for id in joblist:
        #    print('Processing for', i)
        detail = pagetext.find(name='article',attrs={'data-job-id':id})
        jobid.append(id)
        jobclass.append(jobc)
        salary.append(get_salary(detail))
        title.append(get_job_title(detail))
        location.append(get_job_location(detail))
        suburb.append(get_job_suburb(detail))
        category.append(get_job_category(detail))
        subcategory.append(get_job_subcategory(detail))
        company.append(get_job_company(detail))
        # Get details from job details page
        get_dets_from_detail(id)
    return
```

Webscraping MAIN

In [16]:

```
###################################################
# Initialise Panda Series
###################################################
jobid=[]
jobclass=[]
salary=[]
jobdate=[]
title=[]
location=[]
suburb=[]
category=[]
subcategory=[]
worktype=[]
company=[]
jobtext=[]
###################################################
# Define Job Class and URLs for seek summary pages
###################################################
URL = [
        ['DS',"https://www.seek.com.au/data-scientist-jobs/in-All-Australia"],
        ['DA',"https://www.seek.com.au/data-analyst-jobs/in-All-Australia"],
        ['DE',"https://www.seek.com.au/data-engineer-jobs/in-All-Australia"]
        ['BI',"https://www.seek.com.au/business-intelligence-jobs/in-All-Australia"]
      ]
###################################################
# Web Scraping Start
###################################################
#
for jobc,loc_url in URL:                                            # For each JobClass and
    print("Location Processing", loc_url,'for jobclass',jobc)
#
    for i in range(1,100):                              # For each page of Location URL
        page_url = loc_url + '?page=' + str(i)          # Construct the URL for a page
        r = requests.get(page_url)                      # Request access to website
        sc = r.status_code
        if sc != 200:                                   # Status Code 200 is success
            print('Error',sc,'opening page',str(i))
            continue
        else:
            pagetext = bs(r.text,'lxml')                # Get text for FULL summary page
            joblist=[]                                  # Initialise JobID list for the pag
            get_jobids(pagetext)                        # Construct JobID list for the page
            get_details(pagetext,jobc)                  # This function iterates through Jo
            sc = check_next_page(pagetext)              # Last page does not return an erro
            print('  Page Processing',page_url,'>>',len(joblist),'jobs',sc)
            if sc == 'NA':
                break
###################################################
# Make DataFrame from lists
###################################################
jobsdf = pd.DataFrame(
    {'jobid':jobid, 'jobclass':jobclass,'salary':salary, 'jobdate':jobdate, 'title':title,
     'suburb':suburb, 'category':category, 'subcategory':subcategory, 'worktype':worktype,
     'jobtext':jobtext})
# Remove carriage returns from dataframe
jobsdf = jobsdf.replace(r'\n',' ', regex=True)
print('---------- Processing Completed ------------')
```

```
Location Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austr
```

```
alia (https://www.seek.com.au/data-scientist-jobs/in-All-Australia) for jobc
lass DS
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=1 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=1) >> 22 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=2 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=2) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=3 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=3) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=4 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=4) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=5 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=5) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=6 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=6) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=7 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=7) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=8 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=8) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=9 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?page
=9) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=10 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?pag
e=10) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=11 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?pag
e=11) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=12 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?pag
e=12) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-scientist-jobs/in-All-Austral
ia?page=13 (https://www.seek.com.au/data-scientist-jobs/in-All-Australia?pag
e=13) >> 16 jobs NA
Location Processing https://www.seek.com.au/data-analyst-jobs/in-All-Austral
ia (https://www.seek.com.au/data-analyst-jobs/in-All-Australia) for jobclass
 DA
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=1 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1)
 >> 22 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=2 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2)
 >> 22 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=3 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=3)
 >> 21 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=4 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=4)
 >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=5 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=5)
 >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=6 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=6)
```

```
 >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=7 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=7)
 >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=8 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=8)
 >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=9 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=9)
 >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=10 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
0) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=11 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
1) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=12 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
2) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=13 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
3) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=14 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
4) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=15 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
5) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=16 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
6) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=17 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
7) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=18 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
8) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=19 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=1
9) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=20 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
0) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=21 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
1) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=22 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
2) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=23 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
3) >> 20 jobs Next
Error 404 opening page for jobid 36078999
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=24 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
4) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=25 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
5) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=26 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
6) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=27 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
7) >> 20 jobs Next
Error 404 opening page for jobid 36062881
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=28 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
8) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=29 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=2
9) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=30 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=3
0) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=31 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=3
1) >> 20 jobs Next
  Page Processing https://www.seek.com.au/data-analyst-jobs/in-All-Australi
a?page=32 (https://www.seek.com.au/data-analyst-jobs/in-All-Australia?page=3
2) >> 14 jobs NA
Location Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia (https://www.seek.com.au/business-intelligence-jobs/in-All-Austr
alia) for jobclass BI
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=1 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=1) >> 22 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=2 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=2) >> 22 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=3 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=3) >> 22 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=4 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=4) >> 22 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=5 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=5) >> 22 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=6 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=6) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=7 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=7) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=8 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=8) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=9 (https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=9) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=10 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=10) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=11 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=11) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=12 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=12) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=13 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=13) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=14 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=14) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=15 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=15) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=16 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=16) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=17 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=17) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=18 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=18) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=19 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=19) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=20 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=20) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=21 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=21) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=22 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=22) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=23 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=23) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=24 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=24) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=25 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=25) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=26 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=26) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=27 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=27) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=28 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=28) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=29 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=29) >> 20 jobs Next
Error 404 opening page for jobid 36024053
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-All-
Australia?page=30 (https://www.seek.com.au/business-intelligence-jobs/in-All
-Australia?page=30) >> 20 jobs Next
```

```
Error 404 opening page for jobid 36015551
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=31 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=31) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=32 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=32) >> 20 jobs Next
```

```
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=33 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=33) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=34 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=34) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=35 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=35) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=36 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=36) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=37 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=37) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=38 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=38) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=39 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=39) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=40 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=40) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=41 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=41) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=42 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=42) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=43 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=43) >> 20 jobs Next
  Page Processing https://www.seek.com.au/business-intelligence-jobs/in-Al
l-Australia?page=44 (https://www.seek.com.au/business-intelligence-jobs/in
-All-Australia?page=44) >> 1 jobs NA
---------- Processing Completed ------------
```

EDA and Data Munging

Cleanup Salary text to get numbers.

In [17]:

```
# Step 1 - Create DataFrame with Salary column only, without affecting Main DataFrame Salar
#jobsdf.head()
x=jobsdf.salary
df=pd.DataFrame({'salary':x,'oldsal':x})
df.head()
```

Out[17]:

	oldsal	salary
0	NA	NA
1	Base + Super + Profit Share	Base + Super + Profit Share
2	NA	NA
3	NA	NA
4	NA	NA

In [18]:

```
# Step 2 - Replace characters not required at all
df['salary'] = df['salary'].str.replace(',','')
df['salary'] = df['salary'].str.replace('$','')
df['salary'] = df['salary'].str.replace('+','')
df['salary'] = df['salary'].str.replace('x','')
df['salary'] = df['salary'].str.replace('\n','')
df['salary'] = df['salary'].str.replace("Phone",'')
df['salary'] = df['salary'].str.replace("phone",'')
#df['salary'].replace(np.nan,'NA',inplace=True)
```

In [19]:

```
#Step 5: Convert text to lower case
df['salary'] = df['salary'].map(lambda x: x.lower())
```

In [20]:

```
# Step 3 - Identify Salary Periods
df['period'] = 'year'
df.loc[df['salary'].str.contains('pa'),'period'] = 'year'
df.loc[df['salary'].str.contains('p.a'),'period'] = 'year'
df.loc[df['salary'].str.contains('P.A'),'period'] = 'year'
df.loc[df['salary'].str.contains('per annum'),'period'] = 'year'
df.loc[df['salary'].str.contains('per year'),'period'] = 'year'
df.loc[df['salary'].str.contains('pd'),'period'] = 'day'
df.loc[df['salary'].str.contains('p.d'),'period'] = 'day'
df.loc[df['salary'].str.contains('PD'),'period'] = 'day'
df.loc[df['salary'].str.contains('day'),'period'] = 'day'
df.loc[df['salary'].str.contains('per hour'),'period'] = 'hour'
df.loc[df['salary'].str.contains('ph'),'period'] = 'hour'
df.loc[df['salary'].str.contains('p.h.'),'period'] = 'hour'
df.loc[df['salary'].str.contains('per day'),'period'] = 'day'
```

In [21]:

```
df.period.value_counts()
```

Out[21]:

```
year    1655
day       85
hour      28
Name: period, dtype: int64
```

In [22]:

```
# Step 4 - Remove period strings from Salary column
df['salary'] = df['salary'].str.replace('pa','')
df['salary'] = df['salary'].str.replace('p.a.','')
df['salary'] = df['salary'].str.replace('P.A.','')
df['salary'] = df['salary'].str.replace('per annum','')
df['salary'] = df['salary'].str.replace('per year','')
df['salary'] = df['salary'].str.replace('pd','')
df['salary'] = df['salary'].str.replace('p.d','')
df['salary'] = df['salary'].str.replace('per day','')
df['salary'] = df['salary'].str.replace('per hour','')
df['salary'] = df['salary'].str.replace('PD','')
```

In [23]:

```
# Step 6 - Replace k with 000s
#df['salary'].replace(r'[k]', '000', regex=True, inplace=True)
```

In [24]:

```
# Step 7 - Stripping most characters and saving numbers to Extras column
# alc = 'abcdefghijklmnopqrstuvwxyz'
# df['salary'] = df['salary'].map(lambda x: x.lstrip(alc+' ').rstrip(alc+' '+'&'))
```

In [25]:

```
# Step 8 - replace blank cells with NaN
df['salary'].replace('',0, inplace=True)
```

In [26]:

```
# Step 8 - Split salaries range in to two index values
import re
ctr=0
x=df['salary'].map(lambda x : re.findall('[0-9\,.]+', str(x)))
lower=[]
upper=[]
for i in x:
    ctr += 1
    #print(i,'len',len(i),'ctr',ctr)
    if (len(i)==1):
        if i[0].isdigit():
            lower.append(int(i[0]))
        else:
            lower.append(np.nan)
        upper.append(np.nan)
    elif (len(i) >= 2):
        if i[0].isdigit():
            lower.append(int(i[0]))
        else:
            lower.append(np.nan)
        if i[1].isdigit():
            upper.append(int(i[1]))
        else:
            upper.append(np.nan)
    else:
        lower.append(np.nan)
        upper.append(np.nan)
```

In [27]:

```
# Step 9 - Append columns to DataFrame
df['lowersal'] = lower
df['uppersal'] = upper
```

In [28]:

```
# Calculate final salary
period = df['period']
sal=[]
for i in range(0,len(lower)):
    #print(lower[i], upper[i], period[i])
    if (pd.isnull(upper[i]) or upper[i]==0.0):
        amt = lower[i]
        #print('first loop', amt)
    else:
        amt = (lower[i] + upper[i]) / 2
        #print('second', amt)
    if (period[i] == 'year' and amt < 1000):
        amt = amt * 1000
    if period[i] == 'day':
        if amt >= 5000:
            sal.append(amt)
        else:
            sal.append(amt * 250)
        #print('day', sal)
    elif period[i] == 'hour':
        sal.append(amt * 2000)
        #print('hour', sal)
    else:
        sal.append(amt)
        #print('final', sal)
```

In [29]:

```
df['finalsal'] = sal
```

In [30]:

```
jobsdf.drop(columns='salary',inplace=True)
```

In [31]:

```
jobsdf[['period','uppersal','lowersal','finalsal']] = df[['period','uppersal','lowersal','f
```

In [32]:

```
jobsdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1768 entries, 0 to 1767
Data columns (total 15 columns):
category       1768 non-null object
company        1768 non-null object
jobclass       1768 non-null object
jobdate        1768 non-null object
jobid          1768 non-null object
jobtext        1768 non-null object
location       1768 non-null object
subcategory    1768 non-null object
suburb         1768 non-null object
title          1768 non-null object
worktype       1768 non-null object
period         1768 non-null object
uppersal       289 non-null float64
lowersal       388 non-null float64
finalsal       388 non-null float64
dtypes: float64(3), object(12)
memory usage: 207.3+ KB
```

In [33]:

```
jobsdf.to_csv('./jobsdf.csv')
```

In [ ]:

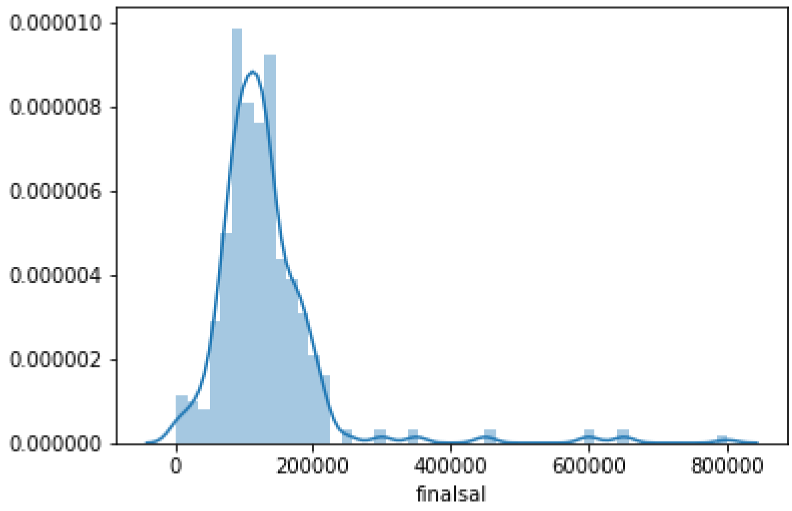
```

```

In [34]:

```
x=jobsdf['finalsal'][jobsdf['finalsal']>0]
sn.distplot(x,bins=50)
```

Out[34]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f7d396e7f0>
```


finalsal