Exit Survey of Australian TAFE and DETE employees

This project takes in data of exit surveys of Australian employees who worked in Australian institutions that concern employment and education. The data are taken from two different sources, namely Department of Education, Training and Employment (DETE) and the Technical and Further Education (TAFE) institute. This project emphasizes on how to deal with invalid and missing data when analyzing a dataset.

We would like to answer the following question:

> What are the profiles of those who resigned from these institutes?

Filter dataset to include only relevant data

First we import the data and do some exploration

In [42]:
```
import warnings
warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np

#read in the data
dete_survey=pd.read_csv("dete_survey.csv")
tafe_survey=pd.read_csv("tafe_survey.csv")
```

```
#explore dete_survey
dete_survey.head()
```

	ID	SeparationType	Cease Date	DETE Start Date	Role Start Date	Position	Classification	Region	Business Unit	Employment Status	...	Kept informed	W pr
0	1	Ill Health Retirement	08/2012	1984	2004	Public Servant	A01-A04	Central Office	Corporate Strategy and Peformance	Permanent Full-time	...	N	N
1	2	Voluntary Early Retirement (VER)	08/2012	Not Stated	Not Stated	Public Servant	AO5-AO7	Central Office	Corporate Strategy and Peformance	Permanent Full-time	...	N	N
2	3	Voluntary Early Retirement (VER)	05/2012	2011	2011	Schools Officer	NaN	Central Office	Education Queensland	Permanent Full-time	...	N	N
3	4	Resignation-Other reasons	05/2012	2005	2006	Teacher	Primary	Central Queensland	NaN	Permanent Full-time	...	A	N
4	5	Age Retirement	05/2012	1970	1989	Head of Curriculum/Head of Special Education	NaN	South East	NaN	Permanent Full-time	...	N	A

5 rows × 56 columns

In [43]:
```
#explore tafe_survey
tafe_survey.head()
```

	Record ID	Institute	WorkArea	CESSATION YEAR	Reason for ceasing employment	Contributing Factors. Career Move - Public Sector	Contributing Factors. Career Move - Private Sector	Contributing Factors. Career Move - Self-employment	Contributing Factors. Ill Health	Co Matern
0	6.341330e+17	Southern Queensland Institute of TAFE	Non-Delivery (corporate)	2010.0	Contract Expired	NaN	NaN	NaN	NaN	NaN
1	6.341337e+17	Mount Isa Institute of TAFE	Non-Delivery (corporate)	2010.0	Retirement	-	-	-	-	-
2	6.341388e+17	Mount Isa Institute of TAFE	Delivery (teaching)	2010.0	Retirement	-	-	-	-	-
3	6.341399e+17	Mount Isa Institute of TAFE	Non-Delivery (corporate)	2010.0	Resignation	-	-	-	-	-
4	6.341466e+17	Southern Queensland Institute of TAFE	Delivery (teaching)	2010.0	Resignation	-	Career Move - Private Sector	-	-	-

5 rows × 72 columns

From looking at the above some notes:

1. The dete_survey dataframe contains 'Not Stated' values that indicate values are missing, but they aren't represented as NaN.
2. Both the dete_survey and tafe_survey contain many columns that we don't need to complete our analysis.
3. Each dataframe contains many of the same columns, but the column names are different.

4. The survey included everyone who left the company for various reasons, not exclusively resignations

---

To address point 1 we can read the data again but this time setting na_values="Not Stated"

In [44]:
```
#Re-read the data with more consistent NaN values
dete_survey=pd.read_csv("dete_survey.csv",na_values="Not Stated")
tafe_survey=pd.read_csv("tafe_survey.csv",na_values="Not Stated")
```

To address point 2, columns 28 to 48 of dete_survey and 17 to 65 of tafe survey were decided to be irrelevant in answering the question; lets drop them

In [45]:
```
#Drop unnecessary columns
dete_survey_updated=dete_survey.drop(dete_survey.columns[28:49],axis=1)
tafe_survey_updated=tafe_survey.drop(tafe_survey.columns[17:66],axis=1)
```

To address point 3, standardize the column names on both surveys that are relevant to answering our question and convey the same information. Since the DETE columns have more comprehensible column names, we adapt TAFE's column names to follow DETE's.

In [46]:
```
#First we standardize DETE's column names for easy referencing later. We standardize to
 snake case
dete_survey_updated.columns=dete_survey_updated.columns.str.lower()
dete_survey_updated.columns=dete_survey_updated.columns.str.strip()
dete_survey_updated.columns=dete_survey_updated.columns.str.replace(" ","_")
print(dete_survey_updated.columns)
```

```
Index(['id', 'separationtype', 'cease_date', 'dete_start_date',
       'role_start_date', 'position', 'classification', 'region',
       'business_unit', 'employment_status', 'career_move_to_public_sector',
```

```
       'career_move_to_private_sector', 'interpersonal_conflicts',
       'job_dissatisfaction', 'dissatisfaction_with_the_department',
       'physical_work_environment', 'lack_of_recognition',
       'lack_of_job_security', 'work_location', 'employment_conditions',
       'maternity/family', 'relocation', 'study/travel', 'ill_health',
       'traumatic_incident', 'work_life_balance', 'workload',
       'none_of_the_above', 'gender', 'age', 'aboriginal', 'torres_strait',
       'south_sea', 'disability', 'nesb'],
      dtype='object')
```

Now standardize column names of TAFE that convey the same information as that of DETE's

In [47]:
```
#Dictionary containing column names in TAFE that are translated to corresponding columns
 in DETE
old_to_new_column_name={'Record ID': 'id',
'CESSATION YEAR': 'cease_date',
'Reason for ceasing employment': 'separationtype',
'Gender. What is your Gender?': 'gender',
'CurrentAge. Current Age': 'age',
'Employment Type. Employment Type': 'employment_status',
'Classification. Classification': 'position',
'LengthofServiceOverall. Overall Length of Service at Institute (in years)': 'institute_
service',
'LengthofServiceCurrent. Length of Service at current workplace (in years)': 'role_servi
ce'
}

#Rename TAFE's columns
tafe_survey_updated=tafe_survey_updated.rename(old_to_new_column_name,axis=1)
print(tafe_survey_updated.columns)
```

```
Index(['id', 'Institute', 'WorkArea', 'cease_date', 'separationtype',
       'Contributing Factors. Career Move - Public Sector ',
```

```
 'Contributing Factors. Career Move - Private Sector ',
 'Contributing Factors. Career Move - Self-employment',
 'Contributing Factors. Ill Health',
 'Contributing Factors. Maternity/Family',
 'Contributing Factors. Dissatisfaction',
 'Contributing Factors. Job Dissatisfaction',
 'Contributing Factors. Interpersonal Conflict',
 'Contributing Factors. Study', 'Contributing Factors. Travel',
 'Contributing Factors. Other', 'Contributing Factors. NONE', 'gender',
 'age', 'employment_status', 'position', 'institute_service',
 'role_service'],
dtype='object')
```

Thus now we ensure that in both dataframes the columns that convey the same information have the same name. This is important for concatenating both dataframes later

---

To address point 4, we first need to filter out the dataset to only include employees who leave the company by resignation. To do this, we check out the types of separation that employers have with the company (e.g. resigned, retired, etc.). This information is under the separationtype column of each dataset.

In [48]:
```
#Access separationtype column statistics
tafe_survey_updated["separationtype"].value_counts()
```

```
Resignation                 340
Contract Expired            127
Retrenchment/ Redundancy    104
Retirement                   82
Transfer                     25
Termination                  23
Name: separationtype, dtype: int64
```

In [49]:
```
dete_survey_updated["separationtype"].value_counts()
```

```
Age Retirement                          285
Resignation-Other reasons               150
```

```
Resignation-Other employer              91
Resignation-Move overseas/interstate    70
Voluntary Early Retirement (VER)        67
Ill Health Retirement                   61
Other                                   49
Contract Expired                        34
Termination                             15
Name: separationtype, dtype: int64
```

From the above, there are 4 types of resignations; for our objective we will only focus on these. We filter each dataset so that only rows with "resignation" are kept

In [50]:
```
#Create regex pattern for filtering
pattern=r"Resignation"
#Filter both datasets to only include employees who resigned
dete_resignations=dete_survey_updated[dete_survey_updated["separationtype"].str.contains
(pattern)]
tafe_resignations=tafe_survey_updated[tafe_survey_updated["separationtype"].str.contains
(pattern,na=False)]
```

We would also like to standardize the length of employment of these resigned employees. The datasets contain data about the start and end year of employment. The TAFE dataset already has the employment duration under institute_service column, however the DETE doesn't. Let's create this column for DETE.

First, some of the data in cease_date and dete_start_date columns of the DETE dataset are not clean. Let's check them

In [51]:
```
#Check resignation year in DETE
dete_resignations["cease_date"].value_counts()
```

```
2012       126
2013        74
01/2014     22
```

```
12/2013     17
06/2013     14
09/2013     11
07/2013      9
11/2013      9
10/2013      6
08/2013      4
05/2012      2
05/2013      2
09/2010      1
07/2006      1
07/2012      1
2010         1
Name: cease_date, dtype: int64
```

In [52]:

```
#Check start year in DETE
dete_resignations["dete_start_date"].value_counts(sort=True).sort_index(ascending=False)
```

```
2013.0    10
2012.0    21
2011.0    24
2010.0    17
2009.0    13
2008.0    22
2007.0    21
2006.0    13
2005.0    15
2004.0    14
2003.0     6
2002.0     6
2001.0     3
2000.0     9
1999.0     8
1998.0     6
1997.0     5
1996.0     6
1995.0     4
1994.0     6
```

```
1993.0     5
1992.0     6
1991.0     4
1990.0     5
1989.0     4
1988.0     4
1987.0     1
1986.0     3
1985.0     3
1984.0     1
1983.0     2
1982.0     1
1980.0     5
1977.0     1
1976.0     2
1975.0     1
1974.0     2
1973.0     1
1972.0     1
1971.0     1
1963.0     1
Name: dete_start_date, dtype: int64
```

The data in cease_date are all not in years. Let's fix this

In [53]:
```
#Fix cease_date column of DETE to only include year of resignation
pattern=r"([1-2][0-9]{3})"
dete_resignations["cease_date"]=dete_resignations["cease_date"].copy().str.extract(patte
rn).astype(float)
```

In [54]:
```
#Calculate employment length of resigned employees in DETE
dete_resignations["institute_service"]=dete_resignations["cease_date"].copy()-dete_resig
nations["dete_start_date"].copy()
```

With this data, we can find how many years do each employee work in the DETE institute

Now that the datasets only include relevant data, we can start profiling these resigned employees.

Classify resignations based on dissapointment

First we would like to profile resigned employees that resigned because they are dissapointed.

The datasets have columns of possible reasons of resignation, with True/False answer for each reason. In each data set, we have chosen the columns that indicate dissapointment, and the choice are as shown:

TAFE

1. Contributing Factors. Dissatisfaction
2. Contributing Factors. Job Dissatisfaction

DETE

1. job_dissatisfaction
2. dissatisfaction_with_the_department
3. physical_work_environment
4. lack_of_recognition
5. lack_of_job_security
6. work_location
7. employment_conditions
8. work_life_balance
9. workload

Let's create a new column called "dissatisfied", which will have True whenever any of the factors above are True for each resigned employee.

In [55]:
```
#We do the above first for TAFE
#In the two columns indicating disappointment specified above, "-" indicates False and c
```

```
olumn name indicates True

#Initialize tafe dissatisfaction columns
tafe_dissatisfaction_columns=["Contributing Factors. Dissatisfaction","Contributing Fact
ors. Job Dissatisfaction"]

#Check if there are null/invalid entries in the columns
for i in tafe_dissatisfaction_columns:
    print(tafe_resignations[tafe_resignations[i].isnull()])
#The print statement above shows there are 8 NaN entries in each column, and each pair r
eside in the same row.
#This is fortunate, since this means the invalid entries are grouped together. Thus when
 we create the dissatisfied column
#later, valid data in that column comes from valid data and invalid data from invalid da
ta

#Create a function that maps - to False (i.e. not disappointed) and column name to True
 (i.e. disappointed)
def update_vals(value):
    if value=="-":
        return False
    elif pd.isnull(value):
        return np.nan
    return True

#Create the dissatisfied column for TAFE; if the employee answers True in any of the two
 columns, the dissatisfied value for
#that employee will be True. The NaN employees will remain NaN.

X=tafe_resignations[tafe_dissatisfaction_columns].applymap(update_vals).copy()
```

```
#used copy() in the above line so that if I change X (as I do below with any(), the colu
mns in tafe_resignations remain intact)
tafe_resignations["dissatisfied"]=X.any(axis=1,skipna=False)
tafe_resignations_up=tafe_resignations.copy()

tafe_resignations_up['dissatisfied'].value_counts(dropna=False).head()
```

```
               id                          Institute  \
16   6.341770e+17   Brisbane North Institute of TAFE
18   6.341779e+17   Brisbane North Institute of TAFE
51   6.342141e+17  Southbank Institute of Technology
258  6.345510e+17   Tropical North Institute of TAFE
276  6.345581e+17               SkillsTech Australia
437  6.346963e+17   Tropical North Institute of TAFE
513  6.347827e+17  Southbank Institute of Technology
670  6.350124e+17   Tropical North Institute of TAFE

                    WorkArea  cease_date separationtype  \
16   Non-Delivery (corporate)      2010.0    Resignation
18        Delivery (teaching)      2010.0    Resignation
51   Non-Delivery (corporate)      2010.0    Resignation
258  Non-Delivery (corporate)      2011.0    Resignation
276       Delivery (teaching)      2011.0    Resignation
437  Non-Delivery (corporate)      2012.0    Resignation
513  Non-Delivery (corporate)         NaN    Resignation
670       Delivery (teaching)      2013.0    Resignation

    Contributing Factors. Career Move - Public Sector   \
16                                                 NaN
18                                                 NaN
51                                                 NaN
258                                                NaN
276                                                NaN
437                                                NaN
513                                                NaN
670                                                NaN
```

```
    Contributing Factors. Career Move - Private Sector   \
16                                                  NaN
18                                                  NaN
51                                                  NaN
258                                                 NaN
276                                                 NaN
437                                                 NaN
513                                                 NaN
670                                                 NaN

    Contributing Factors. Career Move - Self-employment  \
16                                                  NaN
18                                                  NaN
51                                                  NaN
258                                                 NaN
276                                                 NaN
437                                                 NaN
513                                                 NaN
670                                                 NaN

    Contributing Factors. Ill Health Contributing Factors. Maternity/Family  \
16                              NaN                                     NaN
18                              NaN                                     NaN
51                              NaN                                     NaN
258                             NaN                                     NaN
276                             NaN                                     NaN
437                             NaN                                     NaN
513                             NaN                                     NaN
670                             NaN                                     NaN

     ... Contributing Factors. Study Contributing Factors. Travel  \
16   ...                         NaN                          NaN
18   ...                         NaN                          NaN
51   ...                         NaN                          NaN
258  ...                         NaN                          NaN
276  ...                         NaN                          NaN
437  ...                         NaN                          NaN
513  ...                         NaN                          NaN
670  ...                         NaN                          NaN
```

```
    Contributing Factors. Other Contributing Factors. NONE gender  age  \
16                          NaN                         NaN    NaN  NaN
18                          NaN                         NaN    NaN  NaN
51                          NaN                         NaN    NaN  NaN
258                         NaN                         NaN    NaN  NaN
276                         NaN                         NaN    NaN  NaN
437                         NaN                         NaN    NaN  NaN
513                         NaN                         NaN    NaN  NaN
670                         NaN                         NaN    NaN  NaN

    employment_status position institute_service role_service
16                NaN      NaN               NaN          NaN
18                NaN      NaN               NaN          NaN
51                NaN      NaN               NaN          NaN
258               NaN      NaN               NaN          NaN
276               NaN      NaN               NaN          NaN
437               NaN      NaN               NaN          NaN
513               NaN      NaN               NaN          NaN
670               NaN      NaN               NaN          NaN

[8 rows x 23 columns]
               id                          Institute  \
16   6.341770e+17   Brisbane North Institute of TAFE
18   6.341779e+17   Brisbane North Institute of TAFE
51   6.342141e+17  Southbank Institute of Technology
258  6.345510e+17   Tropical North Institute of TAFE
276  6.345581e+17               SkillsTech Australia
437  6.346963e+17   Tropical North Institute of TAFE
513  6.347827e+17  Southbank Institute of Technology
670  6.350124e+17   Tropical North Institute of TAFE

                     WorkArea  cease_date separationtype  \
16   Non-Delivery (corporate)      2010.0    Resignation
18        Delivery (teaching)      2010.0    Resignation
51   Non-Delivery (corporate)      2010.0    Resignation
258  Non-Delivery (corporate)      2011.0    Resignation
276       Delivery (teaching)      2011.0    Resignation
437  Non-Delivery (corporate)      2012.0    Resignation
```

```
513  Non-Delivery (corporate)         NaN    Resignation
670       Delivery (teaching)      2013.0    Resignation

    Contributing Factors. Career Move - Public Sector   \
16                                                 NaN
18                                                 NaN
51                                                 NaN
258                                                NaN
276                                                NaN
437                                                NaN
513                                                NaN
670                                                NaN

    Contributing Factors. Career Move - Private Sector   \
16                                                  NaN
18                                                  NaN
51                                                  NaN
258                                                 NaN
276                                                 NaN
437                                                 NaN
513                                                 NaN
670                                                 NaN

    Contributing Factors. Career Move - Self-employment  \
16                                                  NaN
18                                                  NaN
51                                                  NaN
258                                                 NaN
276                                                 NaN
437                                                 NaN
513                                                 NaN
670                                                 NaN

    Contributing Factors. Ill Health Contributing Factors. Maternity/Family  \
16                              NaN                                     NaN
18                              NaN                                     NaN
51                              NaN                                     NaN
258                             NaN                                     NaN
276                             NaN                                     NaN
```

```
437                            NaN                                  NaN
513                            NaN                                  NaN
670                            NaN                                  NaN

     ... Contributing Factors. Study Contributing Factors. Travel  \
16   ...                          NaN                           NaN
18   ...                          NaN                           NaN
51   ...                          NaN                           NaN
258  ...                          NaN                           NaN
276  ...                          NaN                           NaN
437  ...                          NaN                           NaN
513  ...                          NaN                           NaN
670  ...                          NaN                           NaN

    Contributing Factors. Other Contributing Factors. NONE gender  age  \
16                         NaN                         NaN    NaN  NaN
18                         NaN                         NaN    NaN  NaN
51                         NaN                         NaN    NaN  NaN
258                        NaN                         NaN    NaN  NaN
276                        NaN                         NaN    NaN  NaN
437                        NaN                         NaN    NaN  NaN
513                        NaN                         NaN    NaN  NaN
670                        NaN                         NaN    NaN  NaN

    employment_status position institute_service role_service
16                NaN      NaN               NaN          NaN
18                NaN      NaN               NaN          NaN
51                NaN      NaN               NaN          NaN
258               NaN      NaN               NaN          NaN
276               NaN      NaN               NaN          NaN
437               NaN      NaN               NaN          NaN
513               NaN      NaN               NaN          NaN
670               NaN      NaN               NaN          NaN

[8 rows x 23 columns]


False    241
True      91
```

```
NaN        8
Name: dissatisfied, dtype: int64
```

In [56]:

```
#Now do the same operation for DETE

dete_dissatisfaction_columns=["job_dissatisfaction",
"dissatisfaction_with_the_department",
"physical_work_environment",
"lack_of_recognition",
"lack_of_job_security",
"work_location",
"employment_conditions",
"work_life_balance",
"workload"]

#Check if there are any NaN values in each column
#for i in dete_dissatisfaction_columns:
#    print(dete_resignations[i].isnull().sum())
#The above returned 0 for all result, and so we can be sure that all the data inside is
 valid.

Y=dete_resignations[dete_dissatisfaction_columns].copy()
dete_resignations["dissatisfied"]=Y.any(axis=1,skipna=False)
dete_resignations_up=dete_resignations.copy()

dete_resignations_up['dissatisfied'].value_counts(dropna=False)
```

```
False    162
True     149
Name: dissatisfied, dtype: int64
```

Now we combine our datasets.

In [57]:
```
# We concatenate the rows.
# All columns will be available and NaN will be added to cells that are in one dataset b
ut not in other

# First we create a column in each dataset to help signify which dataset the rows belong
 to.
dete_resignations_up["institute"]="DETE"
tafe_resignations_up["institute"]="TAFE"
# Now concatenate
combined=pd.concat([tafe_resignations_up,dete_resignations_up],axis=0)

# Next we do an extra filtering task
combined.notnull().sum().sort_values()
```

```
torres_strait                          0
south_sea                              3
aboriginal                             7
disability                             8
nesb                                   9
business_unit                         32
classification                       161
region                               265
role_start_date                      271
dete_start_date                      283
role_service                         290
career_move_to_public_sector         311
interpersonal_conflicts              311
job_dissatisfaction                  311
dissatisfaction_with_the_department  311
physical_work_environment            311
lack_of_recognition                  311
lack_of_job_security                 311
career_move_to_private_sector        311
```

```
work_location                                    311
maternity/family                                 311
relocation                                       311
study/travel                                     311
ill_health                                       311
traumatic_incident                               311
work_life_balance                                311
workload                                         311
none_of_the_above                                311
employment_conditions                            311
Contributing Factors. Ill Health                 332
Contributing Factors. Maternity/Family           332
Contributing Factors. Career Move - Public Sector 332
Contributing Factors. Dissatisfaction            332
Contributing Factors. Job Dissatisfaction        332
Contributing Factors. Interpersonal Conflict     332
Contributing Factors. Study                      332
Contributing Factors. Travel                     332
Contributing Factors. Other                      332
Contributing Factors. NONE                       332
Contributing Factors. Career Move - Self-employment 332
Contributing Factors. Career Move - Private Sector 332
WorkArea                                         340
Institute                                        340
institute_service                                563
gender                                           592
age                                              596
employment_status                                597
position                                         598
cease_date                                       635
dissatisfied                                     643
separationtype                                   651
institute                                        651
id                                               651
dtype: int64
```

In [58]:
```
#We saw that many columns have a lot of NaNs; we take them out and create a cleaner comb
ined dataset.
```

```
combined_updated=combined.dropna(thresh=500,axis=1).copy()
```

After combining, we can plot how many people resigned with disappointment based on their years of service in the institutes. To more easily categorize each dissatisfied person, we can categorize them based on their career stage, defined as:

1. New: less than 3 years working in the company
2. Experienced: 3-6 years
3. Established: 7-10 years
4. Veteran: >=11 years

In [59]:
```
#We check the institute_service column for this purpose, and later create a new column s
ervice_cat to categorize each person
combined_updated["institute_service"].astype("str")
combined_updated["institute_service"].value_counts(dropna=False)
```

```
NaN                   88
Less than 1 year      73
1-2                   64
3-4                   63
5-6                   33
11-20                 26
5.0                   23
1.0                   22
7-10                  21
3.0                   20
0.0                   20
6.0                   17
4.0                   16
2.0                   14
9.0                   14
7.0                   13
More than 20 years    10
8.0                    8
13.0                   8
```

```
15.0                  7
20.0                  7
12.0                  6
10.0                  6
14.0                  6
22.0                  6
17.0                  6
18.0                  5
16.0                  5
11.0                  4
23.0                  4
24.0                  4
19.0                  3
39.0                  3
32.0                  3
21.0                  3
25.0                  2
26.0                  2
30.0                  2
36.0                  2
28.0                  2
41.0                  1
27.0                  1
29.0                  1
31.0                  1
33.0                  1
34.0                  1
35.0                  1
38.0                  1
42.0                  1
49.0                  1
Name: institute_service, dtype: int64
```

In [60]:

```
#Since the institute_service column names are irregular, we need to extract the numbers
 corresponding to the period of employment

#Extract only the years inside institute_service. For those with ranges, we take the low
```

```
er bound
pattern=r"(\d+)"

combined_updated["institute_service"].apply(type).nunique()

#The above indicates insitute_service contains both str and float, so we need to make al
l str so that we can use the
#the extract function; use astype(str) to the Series as shown below
result_pattern=combined_updated["institute_service"].astype("str").str.extract(pattern)
#only extract first occurence in each
result_pattern=result_pattern.astype(float)

#Create category
def categorize(value):
    if value<3:
        return "New"
    elif 3 <= value <7:
        return "Experienced"
    elif 7<=value<11:
        return "Established"
    elif value>=11:
        return "Veteran"
    elif pd.isnull(value):
        return np.nan

#Create the new column specifying the employee's place in the category
combined_updated["service_cat"]=result_pattern.apply(categorize,axis=1,raw=True)
```

Thus now we have two columns; one signifying whether the employee resigned with dissatisfaction or not, and the other categorizing how long have they worked in the company. We can finally plot, for each category (i.e.

the x axis), how many of them resigned with dissapointment (in percentage) (i.e. y axis). However, each column still has NaN values, seen through the code below:

In [61]:
```
combined_updated["dissatisfied"].value_counts(dropna=False)
```

```
False    403
True     240
NaN        8
Name: dissatisfied, dtype: int64
```

In [62]:
```
combined_updated["service_cat"].value_counts(dropna=False)
```

```
New            193
Experienced    172
Veteran        136
NaN             88
Established     62
Name: service_cat, dtype: int64
```

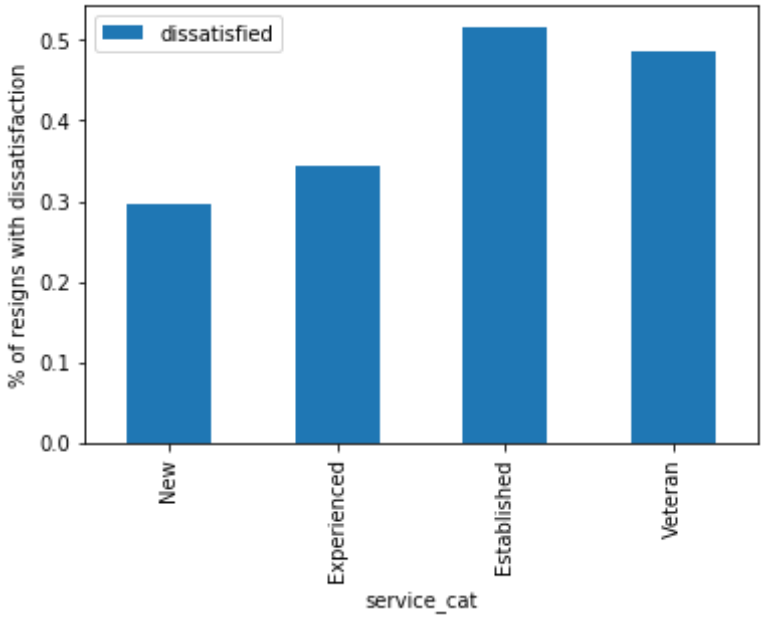
We need to decide on what to do with these NaN values, since they are quite significant. For instance, service_cat has 88 NaN values, which is about 13 percent of the whole data. We decide to drop these from our analysis. For the dissatisfied column, we decide to classify NaN's as False (i.e. not disappointed). This is done below.

In [63]:
```
# Replace missing values in combined_updated with its most frequent value (i.e. False)
combined_updated['dissatisfied'] = combined_updated['dissatisfied'].fillna(False)

# Pivot table so that index is category and values is dissatisfaction. This operation au
tomatically drop the NaN indices.
plot_dissatisfaction=combined_updated.pivot_table(index="service_cat",values="dissatisfi
ed")
```

```
#Finally plot the data
%matplotlib inline
import matplotlib.pyplot as plt
fig=plt.figure()
ax1=fig.add_subplot(1,1,1)
x_label_ordered=["New","Experienced","Established","Veteran"]
plot_diss=plot_dissatisfaction.reindex(x_label_ordered).plot(kind="bar",ax=ax1)
ax1.set_ylabel("% of resigns with dissatisfaction")
```

```
Text(0, 0.5, '% of resigns with dissatisfaction')
```



We can conclude that as an employer serves in the company longer, those who resign are likely to do so in dissatisfaction.

Remember this is preliminary result; we have assumed that 8 people who answered NaN in the dissatisfied column are dissatisfied and we classified them as dissatisfied. We have also excluded the 88 people who was not categorizable due to their start or/and end year of employment being invalid.