# Cleaning and Analyzing Employee Exit Surveys

## Introduction

In this project, I take on the role of a data analyst. I work with exit surveys from employees of the Department of Education, Training and Employment (DETE) and the Technical and Further Education (TAFE) institute in Queensland, Australia.

## Goals

My goals are to clean and analyze the data to answer the following questions to stakeholders:

- Are employees who only worked for the institutes for a short period of time resigning due to some kind of dissatisfaction? What about employees who have been there longer?
- Are younger employees resigning due to some kind of dissatisfaction? What about older employees?

## Data

Below is a preview of a couple columns I'll work with from the dete_survey.csv:

- `ID` : An id used to identify the participant of the survey
- `SeparationType` : The reason why the person's employment ended
- `Cease Date` : The year or month the person's employment ended
- `DETE Start Date` : The year the person began employment with the DETE

Below is a preview of a couple columns I'll work with from the tafe_survey.csv:

- `Record ID` : An id used to identify the participant of the survey
- `Reason for ceasing employment` : The reason why the person's employment ended
- `LengthofServiceOverall. Overall Length of Service at Institute (in years)` : The length of the person's employment (in years)

I'll start by reading the datasets into pandas and exploring them

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

dete_survey = pd.read_csv('dete_survey.csv')
tafe_survey = pd.read_csv('tafe_survey.csv')

print("DETE Information")
dete_survey.info()
dete_survey.tail()
```

```
DETE Information
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 822 entries, 0 to 821
Data columns (total 56 columns):
ID                                     822 non-null int64
SeparationType                         822 non-null object
Cease Date                             822 non-null object
DETE Start Date                        822 non-null object
Role Start Date                        822 non-null object
Position                               817 non-null object
Classification                         455 non-null object
Region                                 822 non-null object
Business Unit                          126 non-null object
Employment Status                      817 non-null object
Career move to public sector           822 non-null bool
Career move to private sector          822 non-null bool
Interpersonal conflicts                822 non-null bool
Job dissatisfaction                    822 non-null bool
Dissatisfaction with the department    822 non-null bool
Physical work environment              822 non-null bool
Lack of recognition                    822 non-null bool
Lack of job security                   822 non-null bool
Work location                          822 non-null bool
Employment conditions                  822 non-null bool
Maternity/family                       822 non-null bool
Relocation                             822 non-null bool
Study/Travel                           822 non-null bool
Ill Health                             822 non-null bool
Traumatic incident                     822 non-null bool
Work life balance                      822 non-null bool
Workload                               822 non-null bool
None of the above                      822 non-null bool
Professional Development               808 non-null object
Opportunities for promotion            735 non-null object
Staff morale                           816 non-null object
Workplace issue                        788 non-null object
Physical environment                   817 non-null object
Worklife balance                       815 non-null object
Stress and pressure support            810 non-null object
Performance of supervisor              813 non-null object
Peer support                           812 non-null object
Initiative                             813 non-null object
Skills                                 811 non-null object
Coach                                  767 non-null object
Career Aspirations                     746 non-null object
Feedback                               792 non-null object
Further PD                             768 non-null object
Communication                          814 non-null object
My say                                 812 non-null object
Information                            816 non-null object
Kept informed                          813 non-null object
Wellness programs                      766 non-null object
Health & Safety                        793 non-null object
Gender                                 798 non-null object
Age                                    811 non-null object
Aboriginal                             16 non-null object
Torres Strait                          3 non-null object
```

```
South Sea                                  7 non-null object
Disability                                 23 non-null object
NESB                                       32 non-null object
dtypes: bool(18), int64(1), object(37)
memory usage: 258.6+ KB
```

Out[1]:

| | ID | SeparationType | Cease Date | DETE Start Date | Role Start Date | Position | Classification | Region | Busi |
|---|---|---|---|---|---|---|---|---|---|
| **817** | 819 | Age Retirement | 02/2014 | 1977 | 1999 | Teacher | Primary | Central Queensland | |
| **818** | 820 | Age Retirement | 01/2014 | 1980 | 1980 | Teacher | Secondary | North Coast | |
| **819** | 821 | Resignation-Move overseas/interstate | 01/2014 | 2009 | 2009 | Public Servant | A01-A04 | Central Office | Educ Queens |
| **820** | 822 | Ill Health Retirement | 12/2013 | 2001 | 2009 | Teacher | Secondary | Darling Downs South West | |
| **821** | 823 | Resignation-Move overseas/interstate | 12/2013 | Not Stated | Not Stated | Teacher Aide | NaN | Metropolitan | |

5 rows × 56 columns

In [2]:
```
print("TAFE Information")
tafe_survey.info()
tafe_survey.head()
```

```
TAFE Information
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 702 entries, 0 to 701
Data columns (total 72 columns):
Record ID
702 non-null float64
Institute
702 non-null object
WorkArea
702 non-null object
CESSATION YEAR
695 non-null float64
Reason for ceasing employment
701 non-null object
Contributing Factors. Career Move - Public Sector
437 non-null object
Contributing Factors. Career Move - Private Sector
437 non-null object
Contributing Factors. Career Move - Self-employment
437 non-null object
Contributing Factors. Ill Health
437 non-null object
Contributing Factors. Maternity/Family
437 non-null object
Contributing Factors. Dissatisfaction
437 non-null object
Contributing Factors. Job Dissatisfaction
437 non-null object
Contributing Factors. Interpersonal Conflict
437 non-null object
Contributing Factors. Study
437 non-null object
Contributing Factors. Travel
437 non-null object
Contributing Factors. Other
437 non-null object
Contributing Factors. NONE
437 non-null object
Main Factor. Which of these was the main factor for leaving?
113 non-null object
InstituteViews. Topic:1. I feel the senior leadership had a clear vision and
direction
608 non-null object
InstituteViews. Topic:2. I was given access to skills training to help me do
my job better
613 non-null object
InstituteViews. Topic:3. I was given adequate opportunities for personal deve
lopment
610 non-null object
InstituteViews. Topic:4. I was given adequate opportunities for promotion wit
hin %Institute]Q25LBL%
608 non-null object
InstituteViews. Topic:5. I felt the salary for the job was right for the resp
onsibilities I had
615 non-null object
InstituteViews. Topic:6. The organisation recognised when staff did good work
607 non-null object
```

InstituteViews. Topic:7. Management was generally supportive of me
614 non-null object
InstituteViews. Topic:8. Management was generally supportive of my team
608 non-null object
InstituteViews. Topic:9. I was kept informed of the changes in the organisati
on which would affect me
610 non-null object
InstituteViews. Topic:10. Staff morale was positive within the Institute
602 non-null object
InstituteViews. Topic:11. If I had a workplace issue it was dealt with quickl
y
601 non-null object
InstituteViews. Topic:12. If I had a workplace issue it was dealt with effici
ently
597 non-null object
InstituteViews. Topic:13. If I had a workplace issue it was dealt with discre
etly
601 non-null object
WorkUnitViews. Topic:14. I was satisfied with the quality of the management a
nd supervision within my work unit
609 non-null object
WorkUnitViews. Topic:15. I worked well with my colleagues
605 non-null object
WorkUnitViews. Topic:16. My job was challenging and interesting
607 non-null object
WorkUnitViews. Topic:17. I was encouraged to use my initiative in the course
of my work
610 non-null object
WorkUnitViews. Topic:18. I had sufficient contact with other people in my job
613 non-null object
WorkUnitViews. Topic:19. I was given adequate support and co-operation by my
peers to enable me to do my job
609 non-null object
WorkUnitViews. Topic:20. I was able to use the full range of my skills in my
job
609 non-null object
WorkUnitViews. Topic:21. I was able to use the full range of my abilities in
my job. ; Category:Level of Agreement; Question:YOUR VIEWS ABOUT YOUR WORK UN
IT]    608 non-null object
WorkUnitViews. Topic:22. I was able to use the full range of my knowledge in
my job
608 non-null object
WorkUnitViews. Topic:23. My job provided sufficient variety
611 non-null object
WorkUnitViews. Topic:24. I was able to cope with the level of stress and pres
sure in my job
610 non-null object
WorkUnitViews. Topic:25. My job allowed me to balance the demands of work and
family to my satisfaction
611 non-null object
WorkUnitViews. Topic:26. My supervisor gave me adequate personal recognition
and feedback on my performance
606 non-null object
WorkUnitViews. Topic:27. My working environment was satisfactory e.g. suffici
ent space, good lighting, suitable seating and working area
610 non-null object
WorkUnitViews. Topic:28. I was given the opportunity to mentor and coach othe

rs in order for me to pass on my skills and knowledge prior to my cessation d
ate    609 non-null object
WorkUnitViews. Topic:29. There was adequate communication between staff in my
unit
603 non-null object
WorkUnitViews. Topic:30. Staff morale was positive within my work unit
606 non-null object
Induction. Did you undertake Workplace Induction?
619 non-null object
InductionInfo. Topic:Did you undertake a Corporate Induction?
432 non-null object
InductionInfo. Topic:Did you undertake a Institute Induction?
483 non-null object
InductionInfo. Topic: Did you undertake Team Induction?
440 non-null object
InductionInfo. Face to Face Topic:Did you undertake a Corporate Induction; Ca
tegory:How it was conducted?
555 non-null object
InductionInfo. On-line Topic:Did you undertake a Corporate Induction; Categor
y:How it was conducted?
555 non-null object
InductionInfo. Induction Manual Topic:Did you undertake a Corporate Inductio
n?
555 non-null object
InductionInfo. Face to Face Topic:Did you undertake a Institute Induction?
530 non-null object
InductionInfo. On-line Topic:Did you undertake a Institute Induction?
555 non-null object
InductionInfo. Induction Manual Topic:Did you undertake a Institute Inductio
n?
553 non-null object
InductionInfo. Face to Face Topic: Did you undertake Team Induction; Categor
y?
555 non-null object
InductionInfo. On-line Topic: Did you undertake Team Induction?process you un
dertook and how it was conducted.]
555 non-null object
InductionInfo. Induction Manual Topic: Did you undertake Team Induction?
555 non-null object
Workplace. Topic:Did you and your Manager develop a Performance and Professio
nal Development Plan (PPDP)?
608 non-null object
Workplace. Topic:Does your workplace promote a work culture free from all for
ms of unlawful discrimination?
594 non-null object
Workplace. Topic:Does your workplace promote and practice the principles of e
mployment equity?
587 non-null object
Workplace. Topic:Does your workplace value the diversity of its employees?
586 non-null object
Workplace. Topic:Would you recommend the Institute as an employer to others?
581 non-null object
Gender. What is your Gender?
596 non-null object
CurrentAge. Current Age
596 non-null object
Employment Type. Employment Type

```
596 non-null object
Classification. Classification
596 non-null object
LengthofServiceOverall. Overall Length of Service at Institute (in years)
596 non-null object
LengthofServiceCurrent. Length of Service at current workplace (in years)
596 non-null object
dtypes: float64(2), object(70)
memory usage: 395.0+ KB
```

Out[2]:

| | Record ID | Institute | WorkArea | CESSATION YEAR | Reason for ceasing employment | Contributing Factors. Career Move - Public Sector | Contributing Factors. Career Move - Private Sector | C<br>M<br>e |
|---|---|---|---|---|---|---|---|---|
| **0** | 6.341330e+17 | Southern Queensland Institute of TAFE | Non-Delivery (corporate) | 2010.0 | Contract Expired | 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

Some observations based on the outputs above:

- Both datasets have different shapes
  - DETE has 56 columns and 821 rows
  - TAFE has 72 columns and 702 rows
- This will need to be resolved prior to combining any results
- There are several missing/NaN entries in both datasets, some missing values are not represented as NaN
- Both datsets record answers in different ways and formats, there are some duplicate columns but with different names
- There are several entries that indicate dissatisfaction as a reason for resigning

To address the issue of missing values not being labeled as `NaN`, I'll reread the .csv files into pandas. I'll read `Not Stated` in as `NaN`. Also, I'll drop the unnecessary columns from both data frames.

In [3]:
```
dete_survey = pd.read_csv('dete_survey.csv', na_values='Not Stated')
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)
```

# Data Cleaning

Each dataframe contains many of the same columns, but the column names are different. Below are some of the columns I'd like to use for our final analysis:

| dete_survey | tafe_survey | Definition |
|---|---|---|
| ID | Record ID | An ID used to identify the participant of the survey |
| Separation Type | Reason for ceasing employment | The reason the participant's employment ended |
| Cease Date | CESSATION YEAR | The year or month the participant's employment ended |
| Age | CurrentAge. Current Age | The age of the participant |

Because I eventually want to combine them, I'll have to standardize the column names. I'll do this for the `dete_survey_updated` data frame by removing whitespace from the column names, and replacing spaces and backslashes with with underscores.

In [4]:
```
dete_survey_updated.columns = dete_survey_updated.columns.str.replace(" ","_")
.str.replace(" ","").str.replace("/", "_").str.lower()
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')
```

For the `tafe_survey_updated` data frame, I'll pass in a dictionary to rename some of the column names. The others will be handled later.

In [5]:
```
new_names = {'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_service'}
tafe_survey_updated = tafe_survey_updated.rename(columns=new_names)
tafe_survey_updated.head()
```

Out[5]:

| | id | Institute | WorkArea | cease_date | separationtype | Contributing Factors. Career Move - Public Sector | Contributing Factors. Career Move - Private Sector |
|---|---|---|---|---|---|---|---|
| **0** | 6.341330e+17 | Southern Queensland Institute of TAFE | Non-Delivery (corporate) | 2010.0 | Contract Expired | 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 × 23 columns

Since I'm only interested in employees that resigned due to dissatisfaction I'll look in the `separationtype` column for entries that contain `Resignation`. However, the `dete_survey_updated` dataframe contains multiple separation types with the string 'Resignation':

- Resignation-Other reasons
- Resignation-Other employer
- Resignation-Move overseas/interstate

So I'll have to account for each of these variations so I don't unintentionally drop data.

In [6]:
```
dete_survey_updated['separationtype'].value_counts()
dete_survey_updated['separationtype'] = dete_survey_updated['separationtype'].
str.split('-').str[0]
dete_survey_updated['separationtype'].value_counts()
```

Out[6]:
```
Resignation                          311
Age Retirement                       285
Voluntary Early Retirement (VER)      67
Ill Health Retirement                 61
Other                                 49
Contract Expired                      34
Termination                           15
Name: separationtype, dtype: int64
```

In [7]:
```
dete_resignations = dete_survey_updated[dete_survey_updated['separationtype']
== 'Resignation'].copy()
tafe_resignations = tafe_survey_updated[tafe_survey_updated['separationtype']
== 'Resignation'].copy()
```

In the two cells above, I used `.value_counts()` to review the unique values in the `separationtype` columns of both data frames. Then I assigned the corresponding resignation types to their own variable using the `.copy()` method. I used this method to avoid a 'SettingWithCopy' warning.

## Checking the Data for Errors

Before I start cleaning and manipulating the rest of the data, I'll verify that the data doesn't contain any major inconsistencies. I'll focus on verifying the years in `cease_date` and `dete_start_date` make sense. The `cease_date` should be after the `dete_start_date`. Given that most people in this field start working in their 20s, it's also unlikely that the `dete_start_date` was before the year 1940.

In [8]:
```
dete_resignations['cease_date'].value_counts()
```

Out[8]:
```
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
07/2012      1
07/2006      1
2010         1
Name: cease_date, dtype: int64
```

It looks like there are different formats in this column that prevent me from getting a clear picture of the data. Since I'm only interested in the year, I'll extract the year and make sure I convert it to a float.

In [9]:
```
dete_resignations['cease_date'] = dete_resignations['cease_date'].str.split(
'/').str[-1]
dete_resignations['cease_date'] = dete_resignations['cease_date'].astype("floa
t")

dete_resignations['cease_date'].value_counts()
```

Out[9]:
```
2013.0    146
2012.0    129
2014.0     22
2010.0      2
2006.0      1
Name: cease_date, dtype: int64
```

In [10]:
```
dete_resignations['dete_start_date'].value_counts()
```

Out[10]:
```
2011.0    24
2008.0    22
2007.0    21
2012.0    21
2010.0    17
2005.0    15
2004.0    14
2009.0    13
2006.0    13
2013.0    10
2000.0     9
1999.0     8
1996.0     6
2002.0     6
1992.0     6
1998.0     6
2003.0     6
1994.0     6
1993.0     5
1990.0     5
1980.0     5
1997.0     5
1991.0     4
1989.0     4
1988.0     4
1995.0     4
2001.0     3
1985.0     3
1986.0     3
1983.0     2
1976.0     2
1974.0     2
1971.0     1
1972.0     1
1984.0     1
1982.0     1
1987.0     1
1975.0     1
1973.0     1
1977.0     1
1963.0     1
Name: dete_start_date, dtype: int64
```

Interestingly, there is a wide range of start dates from 1963 to 2013, but most resignations occured in 2012/2013.

In [11]:
```
tafe_resignations['cease_date'].value_counts()
```

Out[11]:
```
2011.0    116
2012.0     94
2010.0     68
2013.0     55
2009.0      2
Name: cease_date, dtype: int64
```

The `tafe_resignations` data frame is in the correct format so I don't need to manipulate it. There doesn't seem to be any major issues with the years.

In order to answer the question: Are employees who have only worked for the institutes for a short period of time resigning due to some kind of dissatisfaction? What about employees who have been at the job longer?

I'll need to calculate how long employees spent in their workplace. The `tafe_resignations` data frame already has this information in the `institute_service` column. I'll need to create a similar column in `dete_resignations` so I'll be able to combine the two data frames.

## Calculating Length of Employment

I will create an `institute_service` column in `dete_resignations` and will subtract the `dete_start_date` from the `cease_date` .

In [12]:
```
dete_resignations['institute_service'] = dete_resignations['cease_date'] - dete_resignations['dete_start_date']
dete_resignations.head()
```

Out[12]:

| | id | separationtype | cease_date | dete_start_date | role_start_date | position | classification | |
|---|---|---|---|---|---|---|---|---|
| **3** | 4 | Resignation | 2012.0 | 2005.0 | 2006.0 | Teacher | Primary | Que |
| **5** | 6 | Resignation | 2012.0 | 1994.0 | 1997.0 | Guidance Officer | NaN | |
| **8** | 9 | Resignation | 2012.0 | 2009.0 | 2009.0 | Teacher | Secondary | Que |
| **9** | 10 | Resignation | 2012.0 | 1997.0 | 2008.0 | Teacher Aide | NaN | |
| **11** | 12 | Resignation | 2012.0 | 2009.0 | 2009.0 | Teacher | Secondary | F Que |

5 rows × 36 columns

# Identifying Dissatisfied Employees

Now, I'll identify any employees who resigned because they were dissatisfied.

Below are the columns I'll use to categorize employees as "dissatisfied" from each dataframe.

- tafe_survey_updated:
  - Contributing Factors. Dissatisfaction
  - Contributing Factors. Job Dissatisfaction
- dafe_survey_updated:
  - job_dissatisfaction
  - dissatisfaction_with_the_department
  - physical_work_environment
  - lack_of_recognition
  - lack_of_job_security
  - work_location
  - employment_conditions
  - work_life_balance
  - workload

If the employee indicated any of the factors above caused them to resign, I'll mark them as dissatisfied in a new column.

First, I'll examine the `tafe_resignation` data set.

```
In [13]: tafe_resignations['Contributing Factors. Dissatisfaction'].value_counts()
```

```
Out[13]: -                                         277
         Contributing Factors. Dissatisfaction      55
         Name: Contributing Factors. Dissatisfaction, dtype: int64
```

```
In [14]: tafe_resignations['Contributing Factors. Job Dissatisfaction'].value_counts()
```

```
Out[14]: -                       270
         Job Dissatisfaction      62
         Name: Contributing Factors. Job Dissatisfaction, dtype: int64
```

Here, I create a function that will update the values in the 'Contributing Factors. Dissatisfaction' and 'Contributing Factors. Job Dissatisfaction' in the tafe_resignations dataframe so that each contains only True, False, or NaN values. Then, I will use the `any()` methods to create a dissatisfied column in both data frames.

In [15]:
```
def update_vals(x):
    if x == '-':
        return False
    elif pd.isnull(x):
        return np.nan
    else:
        return True
tafe_resignations['dissatisfied'] = tafe_resignations[['Contributing Factors.
 Dissatisfaction', 'Contributing Factors. Job Dissatisfaction']].applymap(upda
te_vals).any(1, skipna=False)
tafe_resignations_up = tafe_resignations.copy()


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

Out[15]:
```
False    241
True      91
NaN        8
Name: dissatisfied, dtype: int64
```

In [16]:
```
dete_resignations['dissatisfied'] = dete_resignations[['job_dissatisfaction',
       'dissatisfaction_with_the_department', 'physical_work_environment',
       'lack_of_recognition', 'lack_of_job_security', 'work_location',
       'employment_conditions', 'work_life_balance',
       'workload']].any(1, skipna=False)
dete_resignations_up = dete_resignations.copy()
dete_resignations_up['dissatisfied'].value_counts(dropna=False)
```

Out[16]:
```
False    162
True     149
Name: dissatisfied, dtype: int64
```

## Combining Data Sets

Now the data ready to be combined. My end goal is to aggregate the data according to the `institute_service` column. First, I'll add an `institute` column to each data frame with the name of the organization that gave the surveys. This will allow me to easily distinguish between the two. Then I will combine the data frames using the `pd.concat()` method.

In [17]:
```
dete_resignations_up['institute'] = "DETE"
tafe_resignations_up['institute'] = "TAFE"
combined = pd.concat([dete_resignations_up, tafe_resignations_up], ignore_inde
x=True, axis=0)
```

Now that the data frames are combined into a single data frame, there are still some columns that I need to drop. I'll drop any columns that have less than 500 non null values.

In [18]:
```
combined.notnull().sum().sort_values()
```

Out[18]:
```
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
employment_conditions                           311
work_location                                   311
lack_of_job_security                            311
job_dissatisfaction                             311
dissatisfaction_with_the_department             311
workload                                        311
lack_of_recognition                             311
interpersonal_conflicts                         311
maternity_family                                311
none_of_the_above                               311
physical_work_environment                       311
relocation                                      311
study_travel                                    311
traumatic_incident                              311
work_life_balance                               311
career_move_to_private_sector                   311
ill_health                                      311
Contributing Factors. Career Move - Private Sector    332
Contributing Factors. Other                     332
Contributing Factors. Career Move - Public Sector     332
Contributing Factors. Career Move - Self-employment   332
Contributing Factors. Travel                    332
Contributing Factors. Study                     332
Contributing Factors. Dissatisfaction           332
Contributing Factors. Ill Health                332
Contributing Factors. NONE                      332
Contributing Factors. Maternity/Family          332
Contributing Factors. Job Dissatisfaction       332
Contributing Factors. Interpersonal Conflict    332
WorkArea                                        340
Institute                                       340
institute_service                               563
gender                                          592
age                                             596
employment_status                               597
position                                        598
cease_date                                      635
dissatisfied                                    643
id                                              651
separationtype                                  651
institute                                       651
dtype: int64
```

In [19]:
```
combined_updated = combined.dropna(thresh=500, axis=1).copy()
```

## Categorizing the Service Column

The `institute_service` column has values in a few different formats:

In [20]:
```
combined_updated['institute_service'].value_counts(dropna=False).head(10)
```

Out[20]:
```
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
0.0                 20
Name: institute_service, dtype: int64
```

To analyze the data, I'll convert the numbers into categories. My analysis will be based on this article (https://www.businesswire.com/news/home/20171108006002/en/Age-Number-Engage-Employees-Career-Stage), which makes the argument that understanding employee's needs according to career stage instead of age is more effective. We'll use the slightly modified definitions below:

- New: Less than 3 years at a company
- Experienced: 3-6 years at a company
- Established: 7-10 years at a company
- Veteran: 11 or more years at a company

I'll use string methods to extract the years of service and change the string to a float.

In [21]:
```
combined_updated['institute_service_up'] = combined_updated['institute_servic
e'].astype('str').str.extract(r'(\d+)')
combined_updated['institute_service_up'] = combined_updated['institute_service
_up'].astype('float')

# Check the years extracted are correct
combined_updated['institute_service_up'].value_counts()
```

```
/dataquest/system/env/python3/lib/python3.4/site-packages/ipykernel/__main__.
py:1: FutureWarning:

currently extract(expand=None) means expand=False (return Index/Series/DataFr
ame) but in a future version of pandas this will be changed to expand=True (r
eturn DataFrame)
```

Out[21]:

```
1.0     159
3.0      83
5.0      56
7.0      34
11.0     30
0.0      20
20.0     17
6.0      17
4.0      16
9.0      14
2.0      14
13.0      8
8.0       8
15.0      7
17.0      6
10.0      6
12.0      6
14.0      6
22.0      6
16.0      5
18.0      5
24.0      4
23.0      4
39.0      3
19.0      3
21.0      3
32.0      3
28.0      2
36.0      2
25.0      2
30.0      2
26.0      2
29.0      1
38.0      1
42.0      1
27.0      1
41.0      1
35.0      1
49.0      1
34.0      1
33.0      1
31.0      1
Name: institute_service_up, dtype: int64
```

Now that I have all the years as floats, I'll create a function that maps each value to one of the career stages.

In [22]:
```
# Convert years of service to categories
def transform_service(val):
    if val >= 11:
        return "Veteran"
    elif 7 <= val < 11:
        return "Established"
    elif 3 <= val < 7:
        return "Experienced"
    elif pd.isnull(val):
        return np.nan
    else:
        return "New"
combined_updated['service_cat'] = combined_updated['institute_service_up'].app
ly(transform_service)

# Quick check of the update
combined_updated['service_cat'].value_counts()
```

Out[22]:
```
New            193
Experienced    172
Veteran        136
Established     62
Name: service_cat, dtype: int64
```

## Initial Analysis

Since the `dissatisfied` column consists of Boolean values, which the `pivot_table()` method treats as integers, I can aggregate the `dissatisfied` column and calculate the number of people or the percentage of dissatisfied within each group.

In [23]:
```
combined_updated['dissatisfied'].value_counts(dropna=False)
```

Out[23]:
```
False    403
True     240
NaN        8
Name: dissatisfied, dtype: int64
```

From looking at the column, I have 8 missing values that need to be dealt with. I'll replace the missing values with the value the occurs most frequently in the column, which is False.

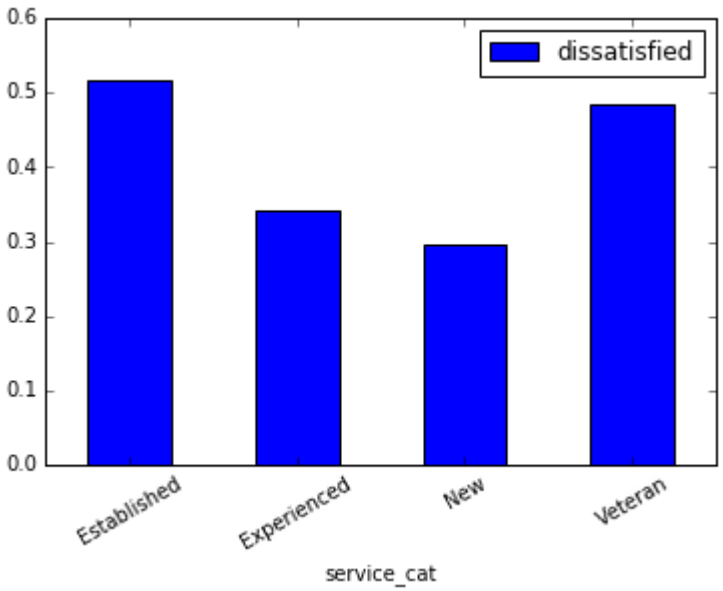
In [24]:
```
combined_updated['dissatisfied'] = combined_updated['dissatisfied'].fillna(Fal
se)
```

Now I'll calculate the percentage of employees who resigned due to dissatisfaction in each category, and plot the results on a bar chart.

In [25]:
```
dis_pct = combined_updated.pivot_table(index='service_cat', values='dissatisfied')

# Plot the results
dis_pct.plot(kind='bar', rot=30)
```

Out[25]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f56f0b2ee48>
```




From the initial analysis results, I can state that employees with 7 or more years of service are more likely to resign due to some kind of dissatisfaction with the job than employees with less than 7 years of service.

## Cleaning the Age Column

To answer one of the original questions: Are younger employees resigning due to some kind of dissatisfaction? What about older employees? I'll have to clean the `age` column in the data frame. However, I also want to see how much data is still missing. Using the `isnull().sum()` I can see that 55 entries have age missing along with 59 missing gender entries and 88 missing institute service.

In [26]:
```
combined_updated.isnull().sum()
```

Out[26]:
```
age                    55
cease_date             16
dissatisfied            0
employment_status      54
gender                 59
id                      0
institute               0
institute_service      88
position               53
separationtype          0
institute_service_up   88
service_cat            88
dtype: int64
```

Before cleaning, I'd also like to know how the age column is structured.

In [27]:
```
combined_updated['age'].value_counts(dropna=False, ascending=False)
```

Out[27]:
```
51-55            71
NaN              55
41-45            48
41  45           45
46-50            42
36-40            41
46  50           39
26-30            35
21  25           33
36  40           32
26  30           32
31  35           32
56 or older      29
31-35            29
21-25            29
56-60            26
61 or older      23
20 or younger    10
Name: age, dtype: int64
```

The data is organized by age being in the form of an age range. This is helpful because I can categorize the age column similar to the `institute_service` column. However, there are some repeat age ranges as separate categories due to the absence of a hyphen. Also, there is a '56 or older' and a '56-60' range that will need to be dealt with. I also see that the missing data category is the second largest. I'll take a closer look at these columns.

In [28]:
```
age_missing = combined_updated[combined_updated['age'].isnull()]
print(age_missing)
```

| | age | cease_date | dissatisfied | employment_status | gender | id |
|---|---|---|---|---|---|---|
| 68 | NaN | 2012.0 | False | Permanent Part-time | Female | 2.150000e+02 |
| 93 | NaN | 2012.0 | False | Permanent Full-time | Female | 2.860000e+02 |
| 141 | NaN | 2012.0 | False | NaN | NaN | 4.060000e+02 |
| 301 | NaN | 2013.0 | False | Permanent Part-time | NaN | 8.040000e+02 |
| 310 | NaN | 2013.0 | False | NaN | NaN | 8.230000e+02 |
| 311 | NaN | 2010.0 | False | NaN | NaN | 6.341399e+17 |
| 322 | NaN | 2010.0 | False | NaN | NaN | 6.341770e+17 |
| 324 | NaN | 2010.0 | False | NaN | NaN | 6.341779e+17 |
| 325 | NaN | 2010.0 | False | NaN | NaN | 6.341820e+17 |
| 326 | NaN | 2010.0 | True | NaN | NaN | 6.341821e+17 |
| 327 | NaN | 2010.0 | False | NaN | NaN | 6.341831e+17 |
| 331 | NaN | 2010.0 | True | NaN | NaN | 6.341934e+17 |
| 335 | NaN | 2010.0 | False | NaN | NaN | 6.342062e+17 |
| 336 | NaN | 2010.0 | False | NaN | NaN | 6.342080e+17 |
| 337 | NaN | 2010.0 | False | NaN | NaN | 6.342081e+17 |
| 345 | NaN | 2010.0 | False | NaN | NaN | 6.342141e+17 |
| 347 | NaN | 2010.0 | False | NaN | NaN | 6.342148e+17 |
| 348 | NaN | 2010.0 | True | NaN | NaN | 6.342174e+17 |
| 367 | NaN | 2010.0 | False | NaN | NaN | 6.342574e+17 |
| 370 | NaN | 2010.0 | False | NaN | NaN | 6.342661e+17 |
| 373 | NaN | 2011.0 | False | NaN | NaN | 6.342679e+17 |
| 375 | NaN | 2011.0 | True | NaN | NaN | 6.342686e+17 |
| 378 | NaN | 2010.0 | True | NaN | NaN | 6.342745e+17 |
| 379 | NaN | 2010.0 | True | NaN | NaN | 6.342746e+17 |
| 385 | NaN | NaN | True | NaN | NaN | 6.342978e+17 |
| 397 | NaN | 2011.0 | False | NaN | NaN | 6.343264e+17 |
| 402 | NaN | NaN | True | NaN | NaN | 6.343283e+17 |
| 405 | NaN | 2011.0 | False | NaN | NaN | 6.343333e+17 |
| 419 | NaN | 2011.0 | True | NaN | NaN | 6.343811e+17 |
| 440 | NaN | 2010.0 | True | NaN | NaN | 6.344568e+17 |
| 453 | NaN | 2010.0 | True | NaN | NaN | 6.344993e+17 |
| 461 | NaN | 2011.0 | False | NaN | NaN | 6.345234e+17 |
| 466 | NaN | 2011.0 | False | NaN | NaN | 6.345510e+17 |
| 472 | NaN | 2011.0 | False | NaN | NaN | 6.345581e+17 |
| 474 | NaN | 2011.0 | False | NaN | NaN | 6.345632e+17 |
| 476 | NaN | 2011.0 | False | NaN | NaN | 6.345647e+17 |
| 495 | NaN | 2011.0 | True | NaN | NaN | 6.345925e+17 |
| 513 | NaN | 2012.0 | True | NaN | NaN | 6.346668e+17 |
| 519 | NaN | 2012.0 | False | NaN | NaN | 6.346832e+17 |
| 523 | NaN | 2012.0 | False | NaN | NaN | 6.346963e+17 |
| 543 | NaN | NaN | False | NaN | NaN | 6.347827e+17 |
| 554 | NaN | 2012.0 | False | NaN | NaN | 6.348110e+17 |
| 556 | NaN | 2012.0 | False | NaN | NaN | 6.348112e+17 |
| 558 | NaN | 2012.0 | False | NaN | NaN | 6.348129e+17 |
| 562 | NaN | 2012.0 | False | NaN | NaN | 6.348187e+17 |
| 581 | NaN | 2012.0 | False | NaN | NaN | 6.348785e+17 |
| 596 | NaN | 2013.0 | False | NaN | NaN | 6.349156e+17 |
| 599 | NaN | 2013.0 | True | NaN | NaN | 6.349375e+17 |
| 602 | NaN | 2013.0 | False | NaN | NaN | 6.349384e+17 |
| 624 | NaN | 2013.0 | False | NaN | NaN | 6.350055e+17 |
| 625 | NaN | 2013.0 | False | NaN | NaN | 6.350055e+17 |
| 627 | NaN | 2013.0 | False | NaN | NaN | 6.350124e+17 |
| 642 | NaN | 2013.0 | False | NaN | NaN | 6.350496e+17 |
| 645 | NaN | 2013.0 | False | NaN | NaN | 6.350652e+17 |
| 648 | NaN | 2013.0 | False | NaN | NaN | 6.350677e+17 |

| | institute | institute_service | position | separationtype |
|---|---|---|---|---|
| 68 | DETE | 13 | School Administrative Staff | Resignation |
| 93 | DETE | 0 | Cleaner | Resignation |
| 141 | DETE | NaN | Teacher | Resignation |
| 301 | DETE | NaN | Teacher Aide | Resignation |
| 310 | DETE | NaN | Teacher Aide | Resignation |
| 311 | TAFE | NaN | NaN | Resignation |
| 322 | TAFE | NaN | NaN | Resignation |
| 324 | TAFE | NaN | NaN | Resignation |
| 325 | TAFE | NaN | NaN | Resignation |
| 326 | TAFE | NaN | NaN | Resignation |
| 327 | TAFE | NaN | NaN | Resignation |
| 331 | TAFE | NaN | NaN | Resignation |
| 335 | TAFE | NaN | NaN | Resignation |
| 336 | TAFE | NaN | NaN | Resignation |
| 337 | TAFE | NaN | NaN | Resignation |
| 345 | TAFE | NaN | NaN | Resignation |
| 347 | TAFE | NaN | NaN | Resignation |
| 348 | TAFE | NaN | NaN | Resignation |
| 367 | TAFE | NaN | NaN | Resignation |
| 370 | TAFE | NaN | NaN | Resignation |
| 373 | TAFE | NaN | NaN | Resignation |
| 375 | TAFE | NaN | NaN | Resignation |
| 378 | TAFE | NaN | NaN | Resignation |
| 379 | TAFE | NaN | NaN | Resignation |
| 385 | TAFE | NaN | NaN | Resignation |
| 397 | TAFE | NaN | NaN | Resignation |
| 402 | TAFE | NaN | NaN | Resignation |
| 405 | TAFE | NaN | NaN | Resignation |
| 419 | TAFE | NaN | NaN | Resignation |
| 440 | TAFE | NaN | NaN | Resignation |
| 453 | TAFE | NaN | NaN | Resignation |
| 461 | TAFE | NaN | NaN | Resignation |
| 466 | TAFE | NaN | NaN | Resignation |
| 472 | TAFE | NaN | NaN | Resignation |
| 474 | TAFE | NaN | NaN | Resignation |
| 476 | TAFE | NaN | NaN | Resignation |
| 495 | TAFE | NaN | NaN | Resignation |
| 513 | TAFE | NaN | NaN | Resignation |
| 519 | TAFE | NaN | NaN | Resignation |
| 523 | TAFE | NaN | NaN | Resignation |
| 543 | TAFE | NaN | NaN | Resignation |
| 554 | TAFE | NaN | NaN | Resignation |
| 556 | TAFE | NaN | NaN | Resignation |
| 558 | TAFE | NaN | NaN | Resignation |
| 562 | TAFE | NaN | NaN | Resignation |
| 581 | TAFE | NaN | NaN | Resignation |
| 596 | TAFE | NaN | NaN | Resignation |
| 599 | TAFE | NaN | NaN | Resignation |
| 602 | TAFE | NaN | NaN | Resignation |
| 624 | TAFE | NaN | NaN | Resignation |
| 625 | TAFE | NaN | NaN | Resignation |
| 627 | TAFE | NaN | NaN | Resignation |
| 642 | TAFE | NaN | NaN | Resignation |
| 645 | TAFE | NaN | NaN | Resignation |

| | | | | | |
|---|---|---|---|---|---|
| 648 | TAFE | NaN | | NaN | Resignation |

| | institute_service_up | service_cat |
|---|---|---|
| 68 | 13.0 | Veteran |
| 93 | 0.0 | New |
| 141 | NaN | NaN |
| 301 | NaN | NaN |
| 310 | NaN | NaN |
| 311 | NaN | NaN |
| 322 | NaN | NaN |
| 324 | NaN | NaN |
| 325 | NaN | NaN |
| 326 | NaN | NaN |
| 327 | NaN | NaN |
| 331 | NaN | NaN |
| 335 | NaN | NaN |
| 336 | NaN | NaN |
| 337 | NaN | NaN |
| 345 | NaN | NaN |
| 347 | NaN | NaN |
| 348 | NaN | NaN |
| 367 | NaN | NaN |
| 370 | NaN | NaN |
| 373 | NaN | NaN |
| 375 | NaN | NaN |
| 378 | NaN | NaN |
| 379 | NaN | NaN |
| 385 | NaN | NaN |
| 397 | NaN | NaN |
| 402 | NaN | NaN |
| 405 | NaN | NaN |
| 419 | NaN | NaN |
| 440 | NaN | NaN |
| 453 | NaN | NaN |
| 461 | NaN | NaN |
| 466 | NaN | NaN |
| 472 | NaN | NaN |
| 474 | NaN | NaN |
| 476 | NaN | NaN |
| 495 | NaN | NaN |
| 513 | NaN | NaN |
| 519 | NaN | NaN |
| 523 | NaN | NaN |
| 543 | NaN | NaN |
| 554 | NaN | NaN |
| 556 | NaN | NaN |
| 558 | NaN | NaN |
| 562 | NaN | NaN |
| 581 | NaN | NaN |
| 596 | NaN | NaN |
| 599 | NaN | NaN |
| 602 | NaN | NaN |
| 624 | NaN | NaN |
| 625 | NaN | NaN |
| 627 | NaN | NaN |
| 642 | NaN | NaN |

```
645                  NaN          NaN
648                  NaN          NaN
```

For rows with missing age data, it looks like most were not dissatisfied with the job, but most importantly, these rows are also missing several other pieces of data such as gender and employment status that would be unwise to impute. In this case, I'm going to drop these rows because they don't aid in the analysis I am trying to perform. Also, these rows only make up 8% of the overall data set.

In [29]:
```
combined_updated.dropna(subset=['age'], inplace=True)
```

Now I need to clean up the age categories. I'll format the strings to replace characters.

In [30]:
```
combined_updated['age'] = combined_updated['age'].str.replace("  ", " ").str.r
eplace(" or older", "+").str.replace(" ","-")
```

In [31]:
```
combined_updated['age'].unique()
```

Out[31]:
```
array(['36-40', '41-45', '31-35', '46-50', '61+', '56-60', '51-55',
       '21-25', '26-30', '20-or-younger', '56+'], dtype=object)
```

Age ranges are broken down into 5 year intervals, as a result there are 11 categories. There is also three similar categories: 56-60, 56+, and 61+. Below I can see that these categories are relatively small compared to others, so I'll combine them.

In [32]:
```
combined_updated['age'].value_counts(dropna=False, ascending=False)
```

Out[32]:
```
41-45            93
46-50            81
36-40            73
51-55            71
26-30            67
21-25            62
31-35            61
56+              29
56-60            26
61+              23
20-or-younger    10
Name: age, dtype: int64
```

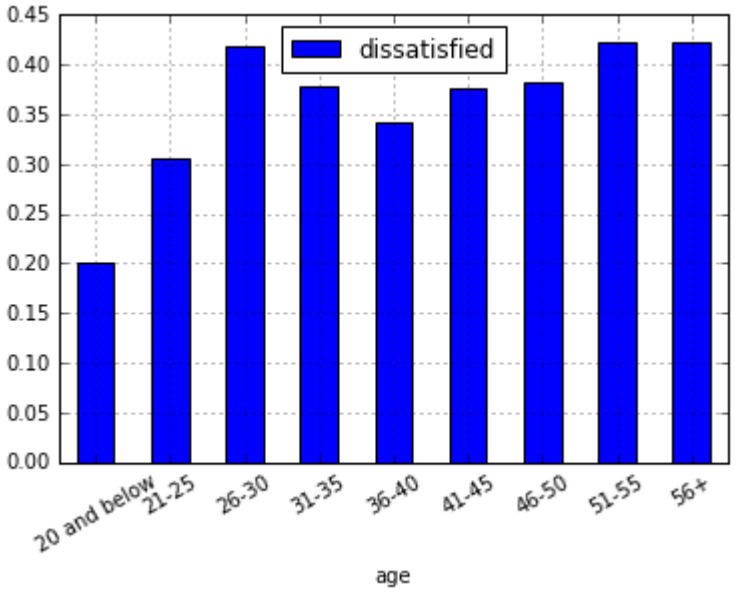
In [33]:
```
combined_updated['age'].replace({'56-60':'56+', '61+':'56+','20-or-younger':'2
0 and below'}, inplace=True)
```

Now that the Age column has been cleaned, I can see whether younger employees are more dissatisfied than older employees. I'll calculate the percent of dissatisfied employees within each age category and plot it as a bar chart.

In [34]:
```
age_dis_pct = combined_updated.pivot_table(index='age', values='dissatisfied')

age_dis_pct.plot(kind='bar', rot=30, grid=True)
```

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x7f56ee9cdd30>

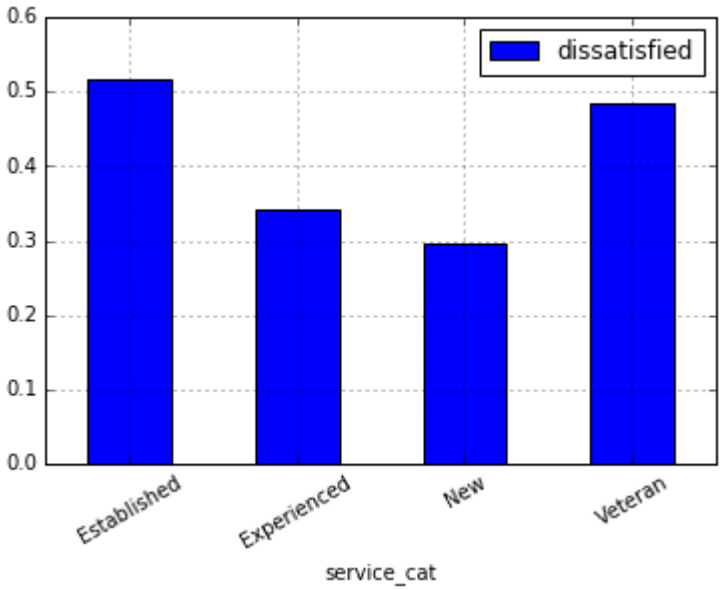



From the plot above, I made the following observations:

- Employees in their late 20s are just as dissatisfied as employees 51 and older.
- About 42% of employees in these age ranges cited dissatisfaction as the reason for resigning.
- However, age ranges between these two reported sligtly lower rates of resigning due to dissatisfaction.
- Only 20-30 % of employees aged 20-25 resigned due to dissatisfaction.

Before I move on, I want to check to see if the `service_cat` column changed at all since dropping missing age columns.

In [35]:
```
#replotting the service category data
dis_pct.plot(kind='bar', rot=30, grid=True)
plt.show()
```




The proportions are about the same as they were before dropping the missing age data. About half of established and Veteran employees reported dissatisfaction as the reason for resigning.

## Conclusions

The goal of this project was to answer the following questions:

- Are employees who only worked for the institutes for a short period of time resigning due to some kind of dissatisfaction? What about employees who have been there longer?
- Are younger employees resigning due to some kind of dissatisfaction? What about older employees?

It seems that employees that have worked for the institutes the longest are resigning due to dissatisfaction in greater numbers than newer employees. However, half of all resignations for established(7-11 years of service) and veteran (11+ years service) employees are due to dissatisfaction.

Young and older employees are resigning due to dissatisfaction with the highest groups being 26-30 year olds and employees 51 and older.

In [ ]:
```
```