nsw_data

August 2, 2022

Data Analysis Case Study

Performed by: Hyeokjin Kwon

[1]:
```
ls
```

```
Data Analysis - Data Sheets.xlsx
Screen Shot 2022-08-01 at 10.34.36 am.png
nsw_data.ipynb
~$Data Analysis - Data Sheets.xlsx
```

Find the xlsx file location and use the file for analysis

[2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
```

[3]:
```
xlsx = pd.ExcelFile('Data Analysis - Data Sheets.xlsx')
```

[4]:
```
xlsx.sheet_names
```

[4]:
```
['Title Page', 'PT & FT Data Table', 'PT & FT Data PivotTable format']
```

There are three sheets in one xlsx file

[5]:
```
df1 = pd.read_excel(xlsx, 'Title Page')
df2 = pd.read_excel(xlsx, 'PT & FT Data Table')
df3 = pd.read_excel(xlsx, 'PT & FT Data PivotTable format')
```

[6]:
```
df1
```

[6]:
```
   InsideSherpa Virtual Internship - Data Analyst Module - Data sheets  \
0                                                NaN
1                                          Glossary:
2                             Sector or Public Sector
3                                            Cluster
4                                          Headcount
5                                                 pp
6                                                NaN
```

```
7                                                   NaN
8                                                   NaN
9                                                Tips:
10  Break each part of the request down to its dat…
11  The two attached sheets contain the same data:…
12  Don't merge cells, it makes an excel file non-…

                                          Unnamed: 1
0                                                NaN
1                                                NaN
2   The term for the collective Agencies/people wh…
3   A group of agencies that share a common functi…
4                            The number of employees
5                                   Percentage Point
6                                                NaN
7                                                NaN
8                                                NaN
9                                                NaN
10                                               NaN
11                                               NaN
12                                               NaN
```

df1 is the first page, which is the description (I will not use this data for analysis)

[7]:
```
df2
```

[7]:
```
   Unnamed: 0          Unnamed: 1       2014     2014.1     2014.2     2014.3  \
0         NaN                 NaN  Full-Time  Full-Time  Part-Time  Part-Time
1     Cluster              Agency       Male     Female       Male     Female
2   Education  Education Agency 1        107        180          8         48
3   Education  Education Agency 2       2797       2463       1691        764
4   Education  Education Agency 3          6         32       1163      18410
..        …                   …          …          …          …
92   Treasury   Treasury Agency 2        272        578          5          5
93   Treasury   Treasury Agency 3        249        258          6         41
94        NaN               Total     123614     156793      13995      87983
95        NaN                 NaN        NaN     280407        NaN     101978
96        NaN                 NaN        NaN        NaN        NaN     382385

         2015     2015.1     2015.2     2015.3  …     2016.2     2016.3  \
0   Full-Time  Full-Time  Part-Time  Part-Time  …  Part-Time  Part-Time
1        Male     Female       Male     Female  …       Male     Female
2         105        176          6         38  …          7         38
3        2115       1767       1670        620  …       1724        665
4          14         40       1250      18852  …       1377      19727
..         …          …          …          … …        …          …
92        295        400         14        182  …         10        169
93        255        289          6         44  …          6         43
```

```
94     118504     152038      14302      89943  …      14678      88264
95        NaN     270542        NaN     104245  …        NaN     102942
96        NaN        NaN        NaN     374787  …        NaN     375407

         2017     2017.1     2017.2     2017.3       2018     2018.1  \
0   Full-Time  Full-Time  Part-Time  Part-Time  Full-Time  Full-Time
1        Male     Female       Male     Female       Male     Female
2         109        246          6         36        123        247
3        2154       2225       1712        746       2294       2666
4          24         33       2211      19415          6         13
..         …          …          …          …          …          …
92         19         15          5          6         18         21
93        270        284          6         42        278        274
94     114962     155408      18706      90721     111377     155833
95        NaN     270370        NaN     109427        NaN     267210
96        NaN        NaN        NaN     379797        NaN        NaN

       2018.2     2018.3
0   Part-Time  Part-Time
1        Male     Female
2           7         33
3        1687        764
4        2501      19110
..         …          …
92          6          6
93          6         49
94      22034      90216
95        NaN     112250
96        NaN     379460

[97 rows x 22 columns]
```

[8]:

```
df3
```

[8]:

```
                           Cluster                              Agency  Year  \
0                        Education                  Education Agency 1  2014
1                        Education                  Education Agency 2  2014
2                        Education                  Education Agency 3  2014
3                        Education                  Education Agency 4  2014
4     Family & Community Services  Family & Community Services Agency 1  2014
…                                …                                   …     …
1835                     Transport                  Transport Agency 5  2018
1836                     Transport                  Transport Agency 6  2018
1837                      Treasury                   Treasury Agency 1  2018
1838                      Treasury                   Treasury Agency 2  2018
1839                      Treasury                   Treasury Agency 3  2018
```

```
         PT/FT  Gender  Headcount
0     Full-Time  Female        180
1     Full-Time  Female       2463
2     Full-Time  Female         32
3     Full-Time  Female      39251
4     Full-Time  Female       9817
…           …      …          …
1835  Part-Time    Male       1354
1836  Part-Time    Male        579
1837  Part-Time    Male          6
1838  Part-Time    Male          6
1839  Part-Time    Male          6

[1840 rows x 6 columns]
```

df2, df3 are basically same file with different format, I will focus on df3 for analysis

[10]:
```
part_time=df3['PT/FT']=='Part-Time'

part=df3[part_time]
full=df3[~part_time]
```

Divide dataset into two categories: part-time, full-time

[11]:
```
male_flag=df3['Gender']=='Male'
male=df3[male_flag]
female=df3[~male_flag]
```

Divide dataset into two categories: male, female

[12]:
```
part
```

[12]:
```
                            Cluster                               Agency  Year  \
184                       Education                   Education Agency 1  2014
185                       Education                   Education Agency 2  2014
186                       Education                   Education Agency 3  2014
187                       Education                   Education Agency 4  2014
188   Family & Community Services  Family & Community Services Agency 1  2014
…                               …                                   …    …
1835                      Transport                   Transport Agency 5  2018
1836                      Transport                   Transport Agency 6  2018
1837                       Treasury                    Treasury Agency 1  2018
1838                       Treasury                    Treasury Agency 2  2018
1839                       Treasury                    Treasury Agency 3  2018

          PT/FT  Gender  Headcount
184   Part-Time  Female         48
185   Part-Time  Female        764
186   Part-Time  Female      18410
```

```
187   Part-Time  Female      16327
188   Part-Time  Female       5794
…          …     …         …
1835  Part-Time    Male       1354
1836  Part-Time    Male        579
1837  Part-Time    Male          6
1838  Part-Time    Male          6
1839  Part-Time    Male          6

[920 rows x 6 columns]
```

[13]:

```
full
```

[13]:

```
                          Cluster                                 Agency  Year  \
0                       Education                     Education Agency 1  2014
1                       Education                     Education Agency 2  2014
2                       Education                     Education Agency 3  2014
3                       Education                     Education Agency 4  2014
4     Family & Community Services  Family & Community Services Agency 1  2014
…                             …                                    …     …
1651                    Transport                     Transport Agency 5  2018
1652                    Transport                     Transport Agency 6  2018
1653                     Treasury                      Treasury Agency 1  2018
1654                     Treasury                      Treasury Agency 2  2018
1655                     Treasury                      Treasury Agency 3  2018

          PT/FT  Gender  Headcount
0     Full-Time  Female        180
1     Full-Time  Female       2463
2     Full-Time  Female         32
3     Full-Time  Female      39251
4     Full-Time  Female       9817
…          …     …         …
1651  Full-Time    Male       7845
1652  Full-Time    Male       1945
1653  Full-Time    Male        288
1654  Full-Time    Male         18
1655  Full-Time    Male        278

[920 rows x 6 columns]
```

[14]:

```
male_part=part[male_flag]
male_full=full[male_flag]
female_part=part[~male_flag]
female_full=full[~male_flag]
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
```

```
  """Entry point for launching an IPython kernel.
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  This is separate from the ipykernel package so we can avoid doing imports
until
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  after removing the cwd from sys.path.
```

Divide dataset into four categories: male part-time, male full-time, female part-time, female full-time

[15]:
```
male_part
```

[15]:
```
                          Cluster                               Agency  Year  \
276                     Education                   Education Agency 1  2014
277                     Education                   Education Agency 2  2014
278                     Education                   Education Agency 3  2014
279                     Education                   Education Agency 4  2014
280   Family & Community Services  Family & Community Services Agency 1  2014
…                             …                                    …     …
1835                    Transport                   Transport Agency 5  2018
1836                    Transport                   Transport Agency 6  2018
1837                     Treasury                    Treasury Agency 1  2018
1838                     Treasury                    Treasury Agency 2  2018
1839                     Treasury                    Treasury Agency 3  2018

          PT/FT Gender  Headcount
276   Part-Time   Male          8
277   Part-Time   Male       1691
278   Part-Time   Male       1163
279   Part-Time   Male       2021
280   Part-Time   Male       1034
…           …     …          …
1835  Part-Time   Male       1354
1836  Part-Time   Male        579
1837  Part-Time   Male          6
1838  Part-Time   Male          6
1839  Part-Time   Male          6

[460 rows x 6 columns]
```

[16]:
```
female_full
```

[16]:

```
                           Cluster                             Agency  Year  \
0                        Education                 Education Agency 1  2014
1                        Education                 Education Agency 2  2014
2                        Education                 Education Agency 3  2014
3                        Education                 Education Agency 4  2014
4     Family & Community Services  Family & Community Services Agency 1  2014
…                                …                                  …     …
1559                     Transport                 Transport Agency 5  2018
1560                     Transport                 Transport Agency 6  2018
1561                      Treasury                  Treasury Agency 1  2018
1562                      Treasury                  Treasury Agency 2  2018
1563                      Treasury                  Treasury Agency 3  2018

          PT/FT  Gender  Headcount
0     Full-Time  Female        180
1     Full-Time  Female       2463
2     Full-Time  Female         32
3     Full-Time  Female      39251
4     Full-Time  Female       9817
…           …      …         …
1559  Full-Time  Female       1922
1560  Full-Time  Female       1983
1561  Full-Time  Female        492
1562  Full-Time  Female         21
1563  Full-Time  Female        274

[460 rows x 6 columns]
```

[17]:

```python
male_part_trend=np.array([male_part[male_part['Year'] == 2014]['Headcount'].
 ↪sum(),male_part[male_part['Year'] == 2015]['Headcount'].
 ↪sum(),male_part[male_part['Year'] == 2016]['Headcount'].
 ↪sum(),male_part[male_part['Year'] == 2017]['Headcount'].sum(),
male_part[male_part['Year'] == 2018]['Headcount'].sum()])
female_part_trend=np.array([female_part[female_part['Year'] ==␣
 ↪2014]['Headcount'].sum(),female_part[female_part['Year'] ==␣
 ↪2015]['Headcount'].sum(),female_part[female_part['Year'] ==␣
 ↪2016]['Headcount'].sum(),female_part[female_part['Year'] ==␣
 ↪2017]['Headcount'].sum(),
female_part[female_part['Year'] == 2018]['Headcount'].sum()])
male_full_trend=np.array([male_full[male_full['Year'] == 2014]['Headcount'].
 ↪sum(),male_full[male_full['Year'] == 2015]['Headcount'].
 ↪sum(),male_full[male_full['Year'] == 2016]['Headcount'].
 ↪sum(),male_full[male_full['Year'] == 2017]['Headcount'].sum(),
male_full[male_full['Year'] == 2018]['Headcount'].sum()])
```

```
female_full_trend=np.array([female_full[female_full['Year'] ==␣
↪2014]['Headcount'].sum(),female_full[female_full['Year'] ==␣
↪2015]['Headcount'].sum(),female_full[female_full['Year'] ==␣
↪2016]['Headcount'].sum(),female_full[female_full['Year'] ==␣
↪2017]['Headcount'].sum(),
female_full[female_full['Year'] == 2018]['Headcount'].sum()])
```

Sum up the employee number by each year(2014 to 2018)

[18]:
```
male_part_trend
```

[18]:
```
array([13995, 14302, 14678, 18706, 22034])
```

[19]:
```
female_part_trend
```

[19]:
```
array([87983, 89943, 88264, 90721, 90216])
```

[20]:
```
male_full_trend
```

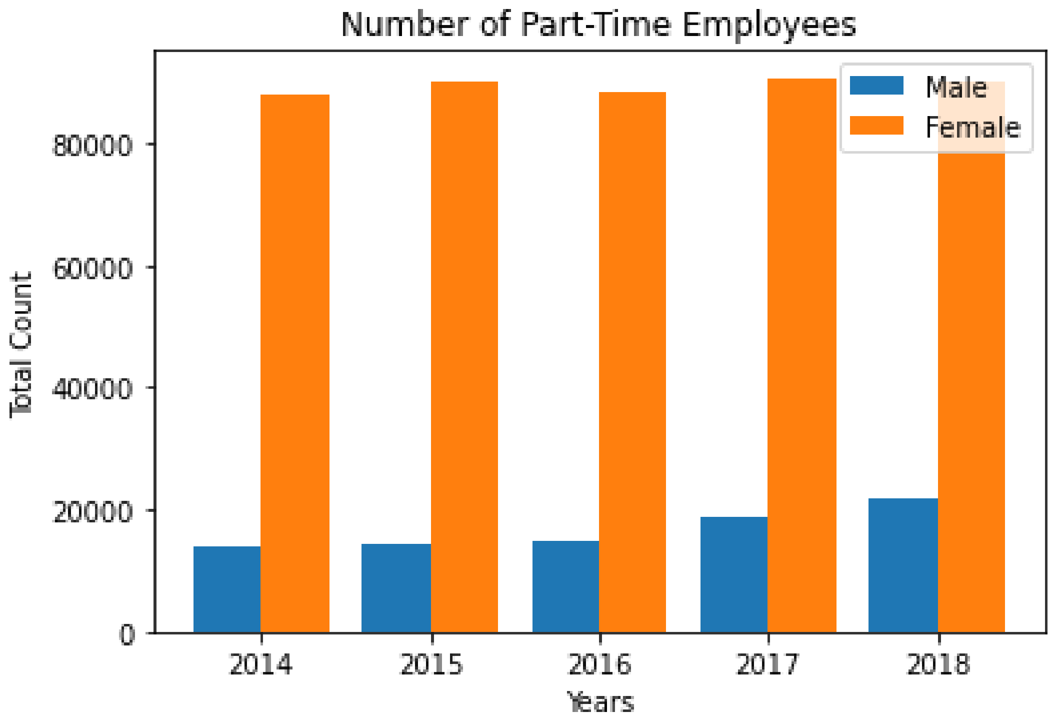
[20]:
```
array([123614, 118504, 117976, 114962, 111377])
```

[21]:
```
female_full_trend
```

[21]:
```
array([156793, 152038, 154489, 155408, 155833])
```

Part-Time Trend from 2014 to 2018 (Male vs Female)

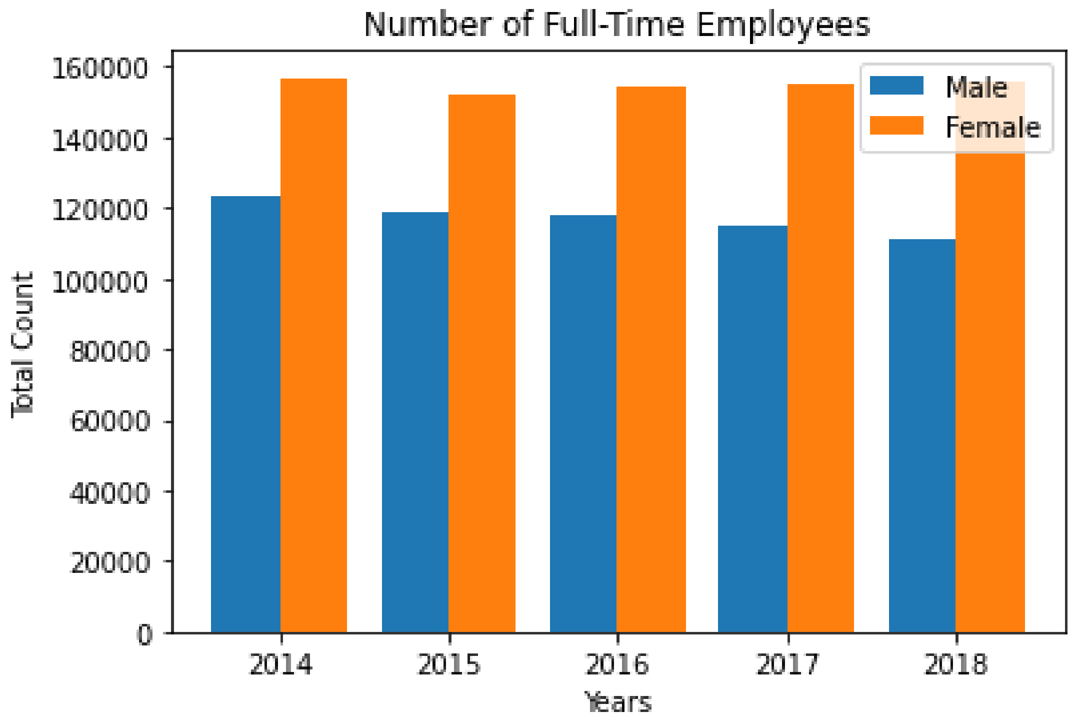
[23]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Number of Part-Time Employees")
plt.legend()
plt.show()
```



The number of Part-Time employees of female is extremely larger than males

Full-Time Trend from 2014 to 2018 (Male vs Female)

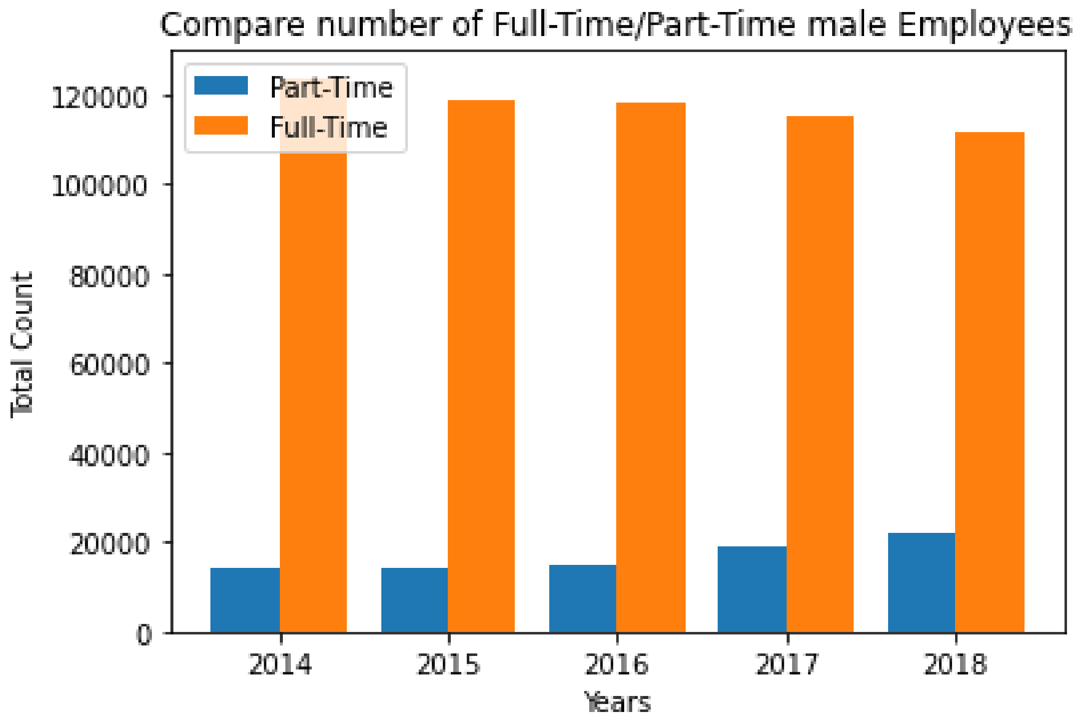
[24]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_full_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_full_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Number of Full-Time Employees")
plt.legend()
plt.show()
```



The number of Full-Time employees female is larger than males. However, less gap between the group compared to the number of part-time employees. In conclusion, females are more employed in both types of employment(Full-Time, Part-Time) than the males.

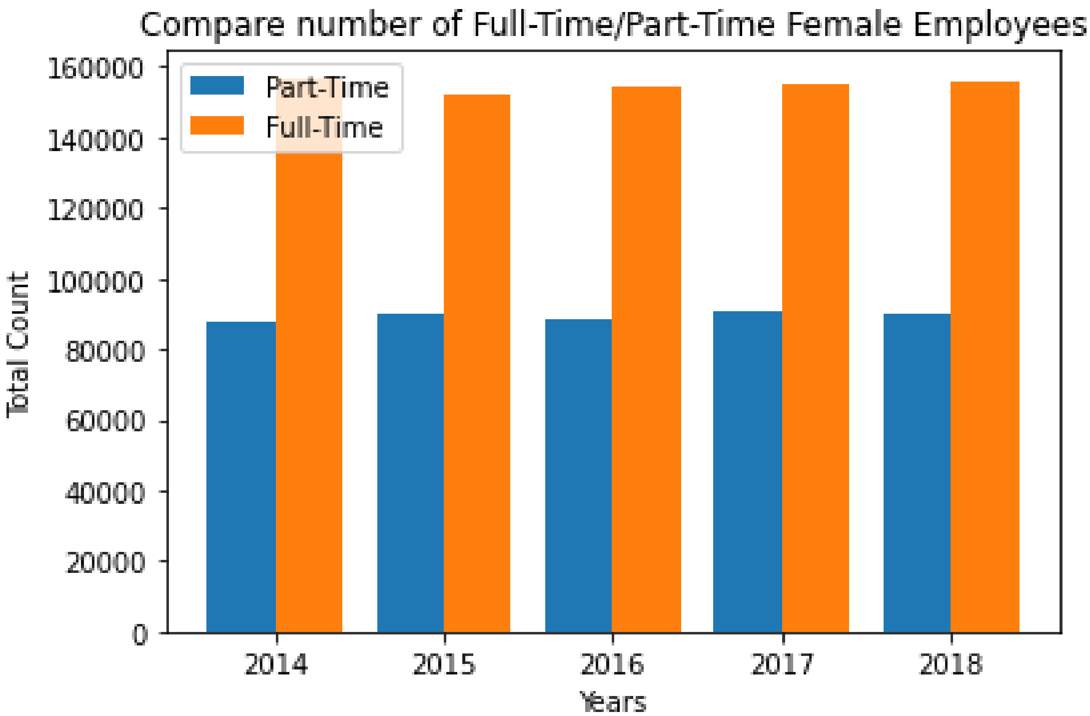
Employment Trend from 2014 to 2018 (Male)

[26]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_trend, 0.4, label = 'Part-Time')
plt.bar(X_axis + 0.2, male_full_trend, 0.4, label = 'Full-Time')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Compare number of Full-Time/Part-Time male Employees")
plt.legend()
plt.show()
```



In the male group, Full-Time workers are significantly more than Part-Time workers.

[25]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, female_part_trend, 0.4, label = 'Part-Time')
plt.bar(X_axis + 0.2, female_full_trend, 0.4, label = 'Full-Time')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Compare number of Full-Time/Part-Time Female Employees")
plt.legend()
plt.show()
```



In the female group, Full-Time workers are more than Part-Time workers. However, the gap between groups is smaller than the male group.

Comparing by each cluster

[27]:
```
df3.Cluster.unique()
```

[27]:
```
array(['Education', 'Family & Community Services',
       'Finance, Services & Innovation', 'Health', 'Industry', 'Justice',
       'Planning & Environment', 'Premier & Cabinet', 'Transport',
       'Treasury'], dtype=object)
```

There are ten clusters

[28]:
```
education_flag=df3['Cluster']=='Education'
family_flag=df3['Cluster']=='Family & Community Services'
finance_flag=df3['Cluster']=='Finance, Services & Innovation'
health_flag=df3['Cluster']=='Health'
industry_flag=df3['Cluster']=='Industry'
justice_flag=df3['Cluster']=='Justice'
planning_flag=df3['Cluster']=='Planning & Environment'
premier_flag=df3['Cluster']=='Premier & Cabinet'
transport_flag=df3['Cluster']=='Transport'
treasury_flag=df3['Cluster']=='Treasury'
```

Divide dataset into ten clusters for each gender (male, female)

[29]:
```
male_part_education=male_part[education_flag]
male_part_family=male_part[family_flag]
male_part_finance=male_part[finance_flag]
male_part_health=male_part[health_flag]
male_part_industry=male_part[industry_flag]
male_part_justice=male_part[justice_flag]
male_part_planning=male_part[planning_flag]
male_part_premier=male_part[premier_flag]
male_part_transport=male_part[transport_flag]
male_part_treasury=male_part[treasury_flag]
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  """Entry point for launching an IPython kernel.
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  This is separate from the ipykernel package so we can avoid doing imports
until
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  after removing the cwd from sys.path.
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:7: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  import sys
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:8: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:9: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  if __name__ == '__main__':
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:10: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  # Remove the CWD from sys.path while we load stuff.
```

[30]:
```
female_part_education=female_part[education_flag]
female_part_family=female_part[family_flag]
female_part_finance=female_part[finance_flag]
female_part_health=female_part[health_flag]
female_part_industry=female_part[industry_flag]
```

```
female_part_justice=female_part[justice_flag]
female_part_planning=female_part[planning_flag]
female_part_premier=female_part[premier_flag]
female_part_transport=female_part[transport_flag]
female_part_treasury=female_part[treasury_flag]
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  """Entry point for launching an IPython kernel.
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  This is separate from the ipykernel package so we can avoid doing imports
until
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  after removing the cwd from sys.path.
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  """
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:6: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:7: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  import sys
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:8: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:9: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  if __name__ == '__main__':
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:10: UserWarning:
Boolean Series key will be reindexed to match DataFrame index.
  # Remove the CWD from sys.path while we load stuff.
```

Trend from 2014 to 2018 for every dataset

[31]:
```
#education
male_part_edu_trend=np.array([male_part_education[male_part_education['Year']␣
 ↪== 2014]['Headcount'].sum(),male_part_education[male_part_education['Year']␣
 ↪== 2015]['Headcount'].sum(),male_part_education[male_part_education['Year']␣
 ↪== 2016]['Headcount'].sum(),male_part_education[male_part_education['Year']␣
 ↪== 2017]['Headcount'].sum(),
male_part_education[male_part_education['Year'] == 2018]['Headcount'].sum()])
```

```
female_part_edu_trend=np.
 ↪array([female_part_education[female_part_education['Year'] ==␣
 ↪2014]['Headcount'].sum(),female_part_education[female_part_education['Year']␣
 ↪== 2015]['Headcount'].
 ↪sum(),female_part_education[female_part_education['Year'] ==␣
 ↪2016]['Headcount'].sum(),female_part_education[female_part_education['Year']␣
 ↪== 2017]['Headcount'].sum(),
female_part_education[female_part_education['Year'] == 2018]['Headcount'].
 ↪sum()])

#family
male_part_fam_trend=np.array([male_part_family[male_part_family['Year'] ==␣
 ↪2014]['Headcount'].sum(),male_part_family[male_part_family['Year'] ==␣
 ↪2015]['Headcount'].sum(),male_part_family[male_part_family['Year'] ==␣
 ↪2016]['Headcount'].sum(),male_part_family[male_part_family['Year'] ==␣
 ↪2017]['Headcount'].sum(),
male_part_family[male_part_family['Year'] == 2018]['Headcount'].sum()])
female_part_fam_trend=np.array([female_part_family[female_part_family['Year']␣
 ↪== 2014]['Headcount'].sum(),female_part_family[female_part_family['Year'] ==␣
 ↪2015]['Headcount'].sum(),female_part_family[female_part_family['Year'] ==␣
 ↪2016]['Headcount'].sum(),female_part_family[female_part_family['Year'] ==␣
 ↪2017]['Headcount'].sum(),
female_part_family[female_part_family['Year'] == 2018]['Headcount'].sum()])

#finance
male_part_finance_trend=np.array([male_part_finance[male_part_finance['Year']␣
 ↪== 2014]['Headcount'].sum(),male_part_finance[male_part_finance['Year'] ==␣
 ↪2015]['Headcount'].sum(),male_part_finance[male_part_finance['Year'] ==␣
 ↪2016]['Headcount'].sum(),male_part_finance[male_part_finance['Year'] ==␣
 ↪2017]['Headcount'].sum(),
male_part_finance[male_part_finance['Year'] == 2018]['Headcount'].sum()])
female_part_finance_trend=np.
 ↪array([female_part_finance[female_part_finance['Year'] == 2014]['Headcount'].
 ↪sum(),female_part_finance[female_part_finance['Year'] == 2015]['Headcount'].
 ↪sum(),female_part_finance[female_part_finance['Year'] == 2016]['Headcount'].
 ↪sum(),female_part_finance[female_part_finance['Year'] == 2017]['Headcount'].
 ↪sum(),
female_part_finance[female_part_finance['Year'] == 2018]['Headcount'].sum()])

#health
male_part_health_trend=np.array([male_part_health[male_part_health['Year'] ==␣
 ↪2014]['Headcount'].sum(),male_part_health[male_part_health['Year'] ==␣
 ↪2015]['Headcount'].sum(),male_part_health[male_part_health['Year'] ==␣
 ↪2016]['Headcount'].sum(),male_part_health[male_part_health['Year'] ==␣
 ↪2017]['Headcount'].sum(),
male_part_health[male_part_health['Year'] == 2018]['Headcount'].sum()])
```

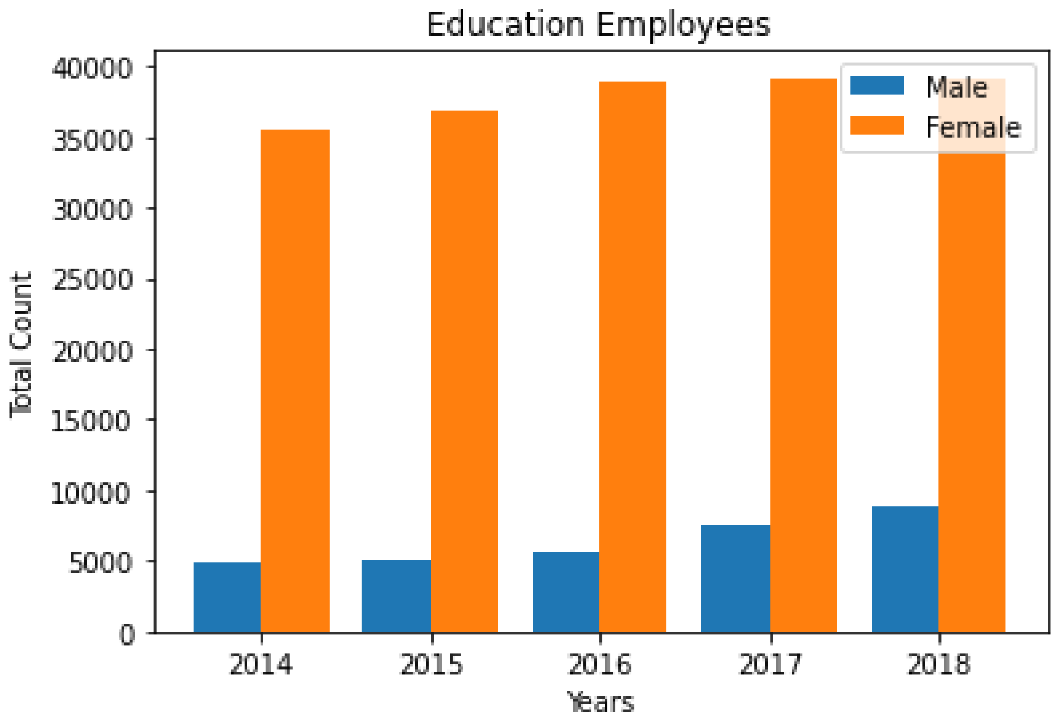
```
female_part_health_trend=np.
↪array([female_part_health[female_part_health['Year'] == 2014]['Headcount'].
↪sum(),female_part_health[female_part_health['Year'] == 2015]['Headcount'].
↪sum(),female_part_health[female_part_health['Year'] == 2016]['Headcount'].
↪sum(),female_part_health[female_part_health['Year'] == 2017]['Headcount'].
↪sum(),
female_part_health[female_part_health['Year'] == 2018]['Headcount'].sum()])

#industry
male_part_industry_trend=np.
↪array([male_part_industry[male_part_industry['Year'] == 2014]['Headcount'].
↪sum(),male_part_industry[male_part_industry['Year'] == 2015]['Headcount'].
↪sum(),male_part_industry[male_part_industry['Year'] == 2016]['Headcount'].
↪sum(),male_part_industry[male_part_industry['Year'] == 2017]['Headcount'].
↪sum(),
male_part_industry[male_part_industry['Year'] == 2018]['Headcount'].sum()])
female_part_industry_trend=np.
↪array([female_part_industry[female_part_industry['Year'] ==␣
↪2014]['Headcount'].sum(),female_part_industry[female_part_industry['Year']␣
↪== 2015]['Headcount'].
↪sum(),female_part_industry[female_part_industry['Year'] ==␣
↪2016]['Headcount'].sum(),female_part_industry[female_part_industry['Year']␣
↪== 2017]['Headcount'].sum(),
female_part_industry[female_part_industry['Year'] == 2018]['Headcount'].sum()])

#justice
male_part_justice_trend=np.array([male_part_justice[male_part_justice['Year']␣
↪== 2014]['Headcount'].sum(),male_part_justice[male_part_justice['Year'] ==␣
↪2015]['Headcount'].sum(),male_part_justice[male_part_justice['Year'] ==␣
↪2016]['Headcount'].sum(),male_part_justice[male_part_justice['Year'] ==␣
↪2017]['Headcount'].sum(),
male_part_justice[male_part_justice['Year'] == 2018]['Headcount'].sum()])
female_part_justice_trend=np.
↪array([female_part_justice[female_part_justice['Year'] == 2014]['Headcount'].
↪sum(),female_part_justice[female_part_justice['Year'] == 2015]['Headcount'].
↪sum(),female_part_justice[female_part_justice['Year'] == 2016]['Headcount'].
↪sum(),female_part_justice[female_part_justice['Year'] == 2017]['Headcount'].
↪sum(),
female_part_justice[female_part_justice['Year'] == 2018]['Headcount'].sum()])

#planning
male_part_planning_trend=np.
↪array([male_part_planning[male_part_planning['Year'] == 2014]['Headcount'].
↪sum(),male_part_planning[male_part_planning['Year'] == 2015]['Headcount'].
↪sum(),male_part_planning[male_part_planning['Year'] == 2016]['Headcount'].
↪sum(),male_part_planning[male_part_planning['Year'] == 2017]['Headcount'].
↪sum(),
```

```
male_part_planning[male_part_planning['Year'] == 2018]['Headcount'].sum()])
female_part_planning_trend=np.
↪array([female_part_planning[female_part_planning['Year'] ==␣
↪2014]['Headcount'].sum(),female_part_planning[female_part_planning['Year']␣
↪== 2015]['Headcount'].
↪sum(),female_part_planning[female_part_planning['Year'] ==␣
↪2016]['Headcount'].sum(),female_part_planning[female_part_planning['Year']␣
↪== 2017]['Headcount'].sum(),
female_part_planning[female_part_planning['Year'] == 2018]['Headcount'].sum()])

#premier
male_part_premier_trend=np.array([male_part_premier[male_part_premier['Year']␣
↪== 2014]['Headcount'].sum(),male_part_premier[male_part_premier['Year'] ==␣
↪2015]['Headcount'].sum(),male_part_premier[male_part_premier['Year'] ==␣
↪2016]['Headcount'].sum(),male_part_premier[male_part_premier['Year'] ==␣
↪2017]['Headcount'].sum(),
male_part_premier[male_part_premier['Year'] == 2018]['Headcount'].sum()])
female_part_premier_trend=np.
↪array([female_part_premier[female_part_premier['Year'] == 2014]['Headcount'].
↪sum(),female_part_premier[female_part_premier['Year'] == 2015]['Headcount'].
↪sum(),female_part_premier[female_part_premier['Year'] == 2016]['Headcount'].
↪sum(),female_part_premier[female_part_premier['Year'] == 2017]['Headcount'].
↪sum(),
female_part_premier[female_part_premier['Year'] == 2018]['Headcount'].sum()])

#transport
male_part_transport_trend=np.
↪array([male_part_transport[male_part_transport['Year'] == 2014]['Headcount'].
↪sum(),male_part_transport[male_part_transport['Year'] == 2015]['Headcount'].
↪sum(),male_part_transport[male_part_transport['Year'] == 2016]['Headcount'].
↪sum(),male_part_transport[male_part_transport['Year'] == 2017]['Headcount'].
↪sum(),
male_part_transport[male_part_transport['Year'] == 2018]['Headcount'].sum()])
female_part_transport_trend=np.
↪array([female_part_transport[female_part_transport['Year'] ==␣
↪2014]['Headcount'].sum(),female_part_transport[female_part_transport['Year']␣
↪== 2015]['Headcount'].
↪sum(),female_part_transport[female_part_transport['Year'] ==␣
↪2016]['Headcount'].sum(),female_part_transport[female_part_transport['Year']␣
↪== 2017]['Headcount'].sum(),
female_part_transport[female_part_transport['Year'] == 2018]['Headcount'].
↪sum()])

#treasury
```

```
male_part_treasury_trend=np.
↪array([male_part_treasury[male_part_treasury['Year'] == 2014]['Headcount'].
↪sum(),male_part_treasury[male_part_treasury['Year'] == 2015]['Headcount'].
↪sum(),male_part_treasury[male_part_treasury['Year'] == 2016]['Headcount'].
↪sum(),male_part_treasury[male_part_treasury['Year'] == 2017]['Headcount'].
↪sum(),
male_part_treasury[male_part_treasury['Year'] == 2018]['Headcount'].sum()])
female_part_treasury_trend=np.
↪array([female_part_treasury[female_part_treasury['Year'] ==␣
↪2014]['Headcount'].sum(),female_part_treasury[female_part_treasury['Year']␣
↪== 2015]['Headcount'].
↪sum(),female_part_treasury[female_part_treasury['Year'] ==␣
↪2016]['Headcount'].sum(),female_part_treasury[female_part_treasury['Year']␣
↪== 2017]['Headcount'].sum(),
female_part_treasury[female_part_treasury['Year'] == 2018]['Headcount'].sum()])
```

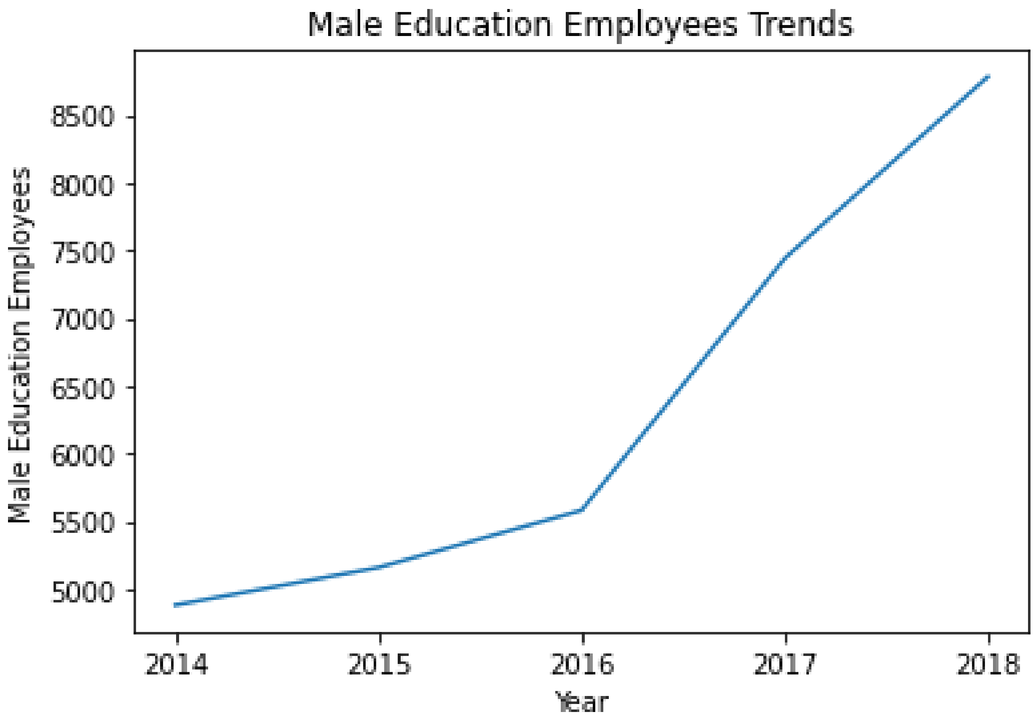
Education

[36]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_edu_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_edu_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Education Employees")
plt.legend()
plt.show()
```



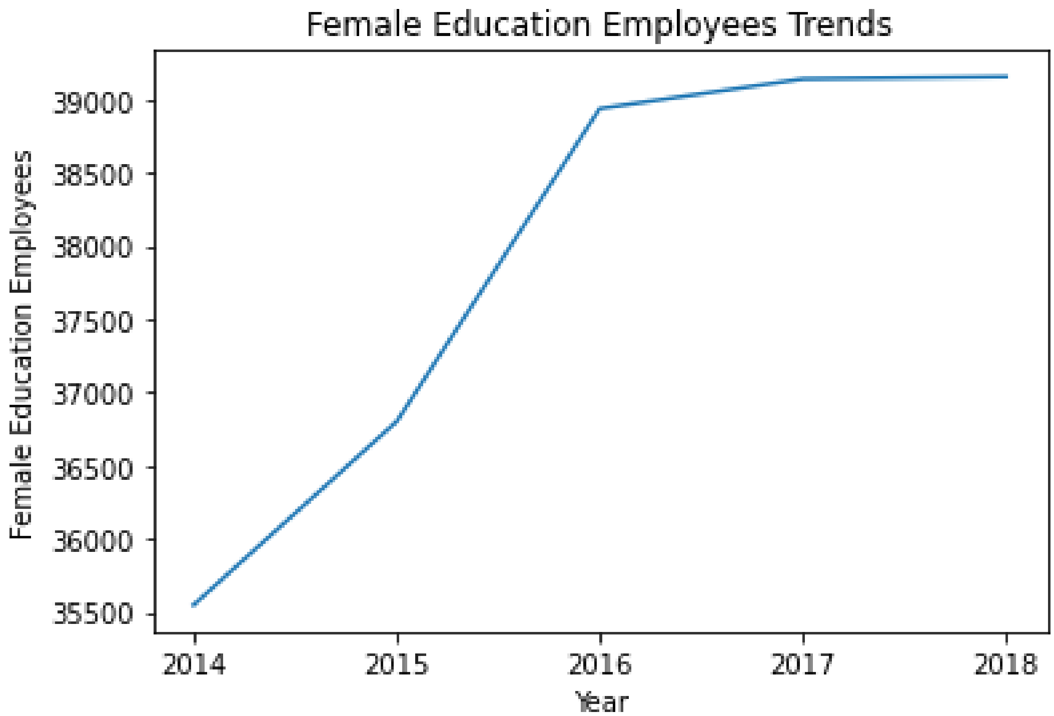
Female group has significantly bigger number of employees in education sector than males.

[38]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_edu_trend)
plt.title('Male Education Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Education Employees')
plt.show()
```



The male group shows drastic upward trends from 2016.

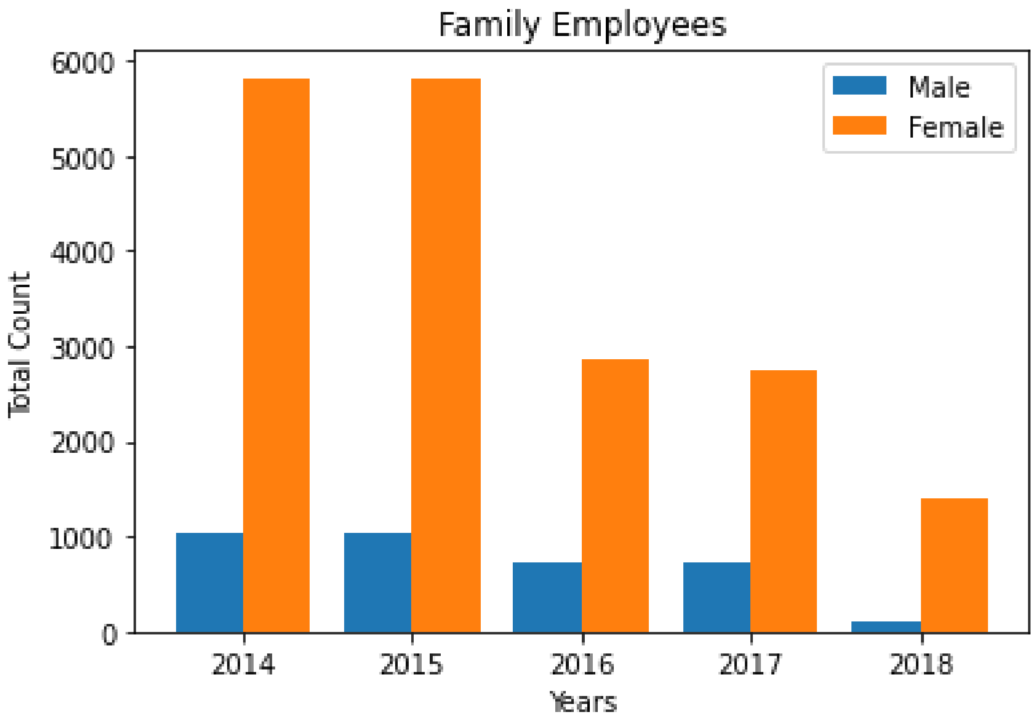
[39]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_edu_trend)
plt.title('Female Education Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Education Employees')
plt.show()
```



The female group shows upward trends too, but its growth slope is reduced from 2016. In conclusion, female employees growth will either stop or decrease around 2025. On the other hand, male employees growth will be keep increasing for few years.

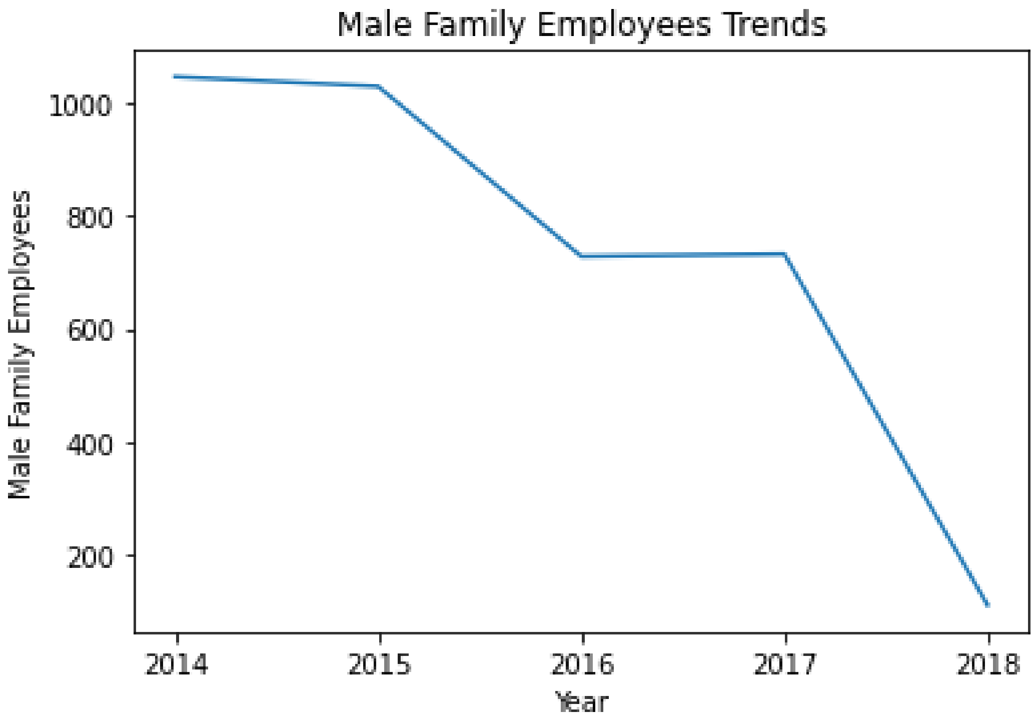
Family

[43]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_fam_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_fam_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Family Employees")
plt.legend()
plt.show()
```



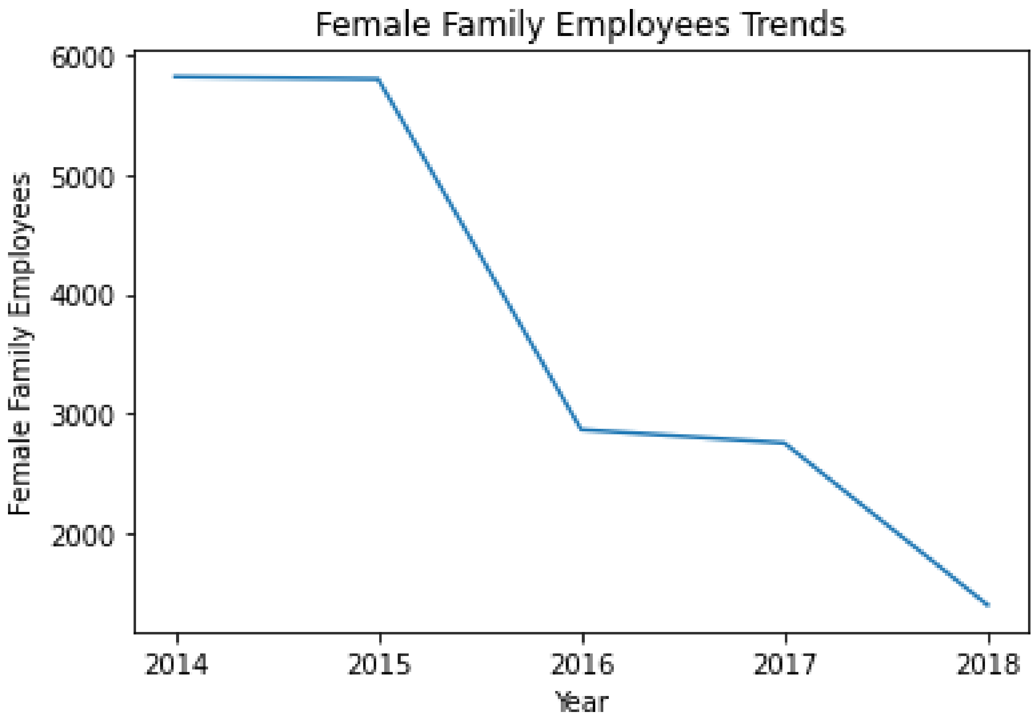
The female group is larger than males, and they both have downward trends.

[44]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_fam_trend)
plt.title('Male Family Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Family Employees')
plt.show()
```



The number of male employees in family sector drastically reduced after 2017.

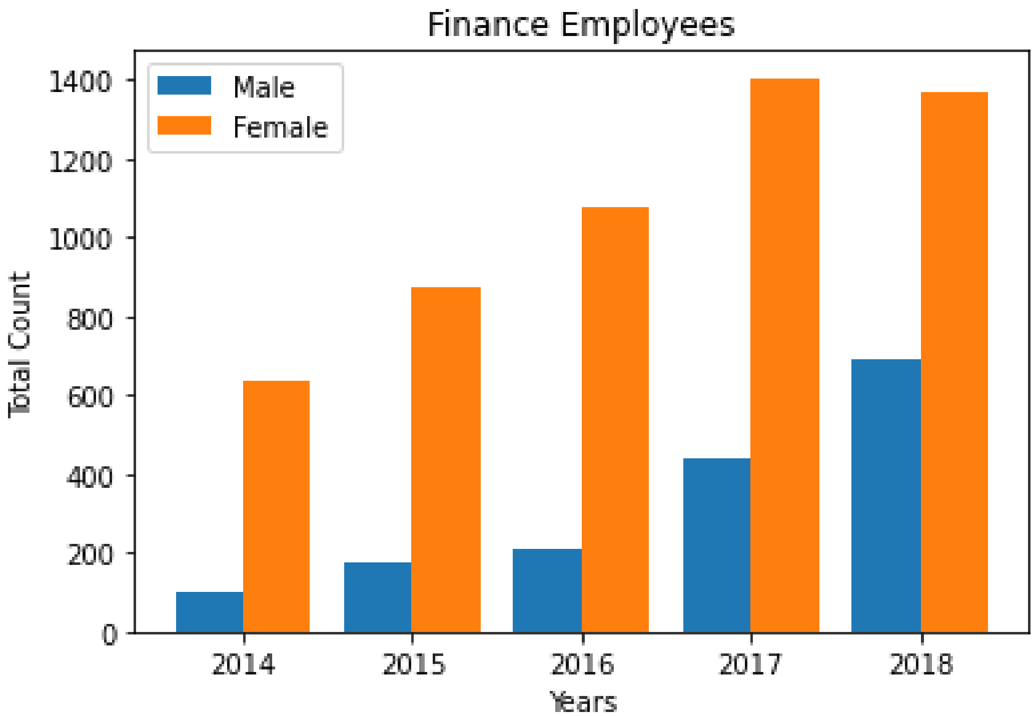
[45]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_fam_trend)
plt.title('Female Family Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Family Employees')
plt.show()
```



The number of female employees in family sector drastically reduced from 2015 to 2016.

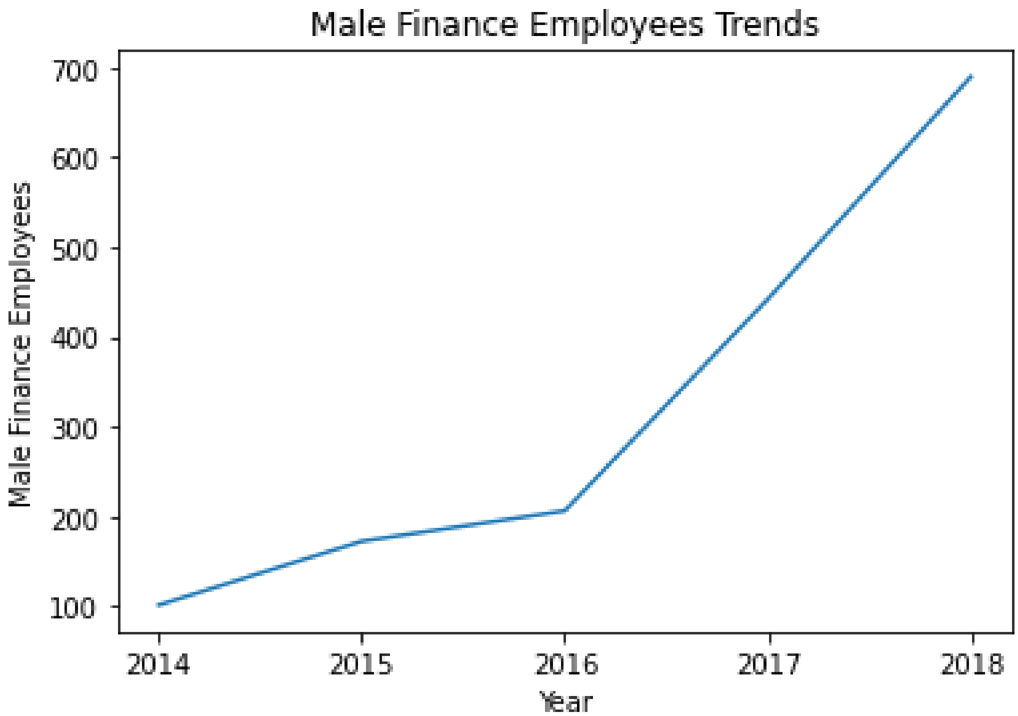
Finance

[46]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_finance_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_finance_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Finance Employees")
plt.legend()
plt.show()
```



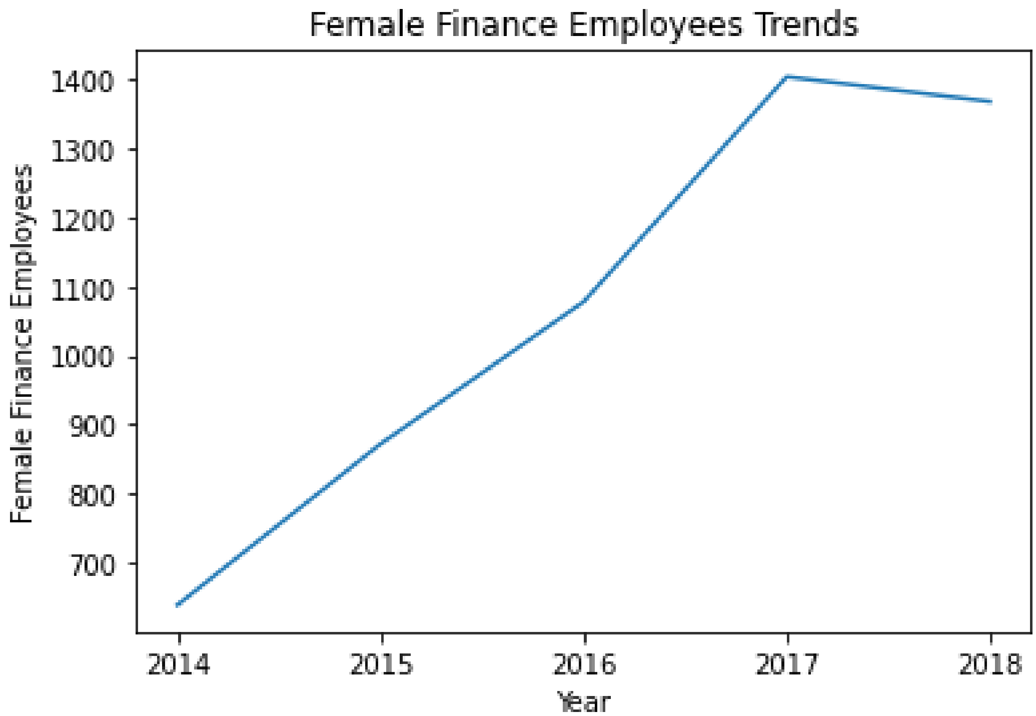
The number of female employees in the finance sector is more than double compare to males.

[47]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_finance_trend)
plt.title('Male Finance Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Finance Employees')
plt.show()
```



The number of male employees in finance sector drastically increased after 2016. It shows strong upward trends.

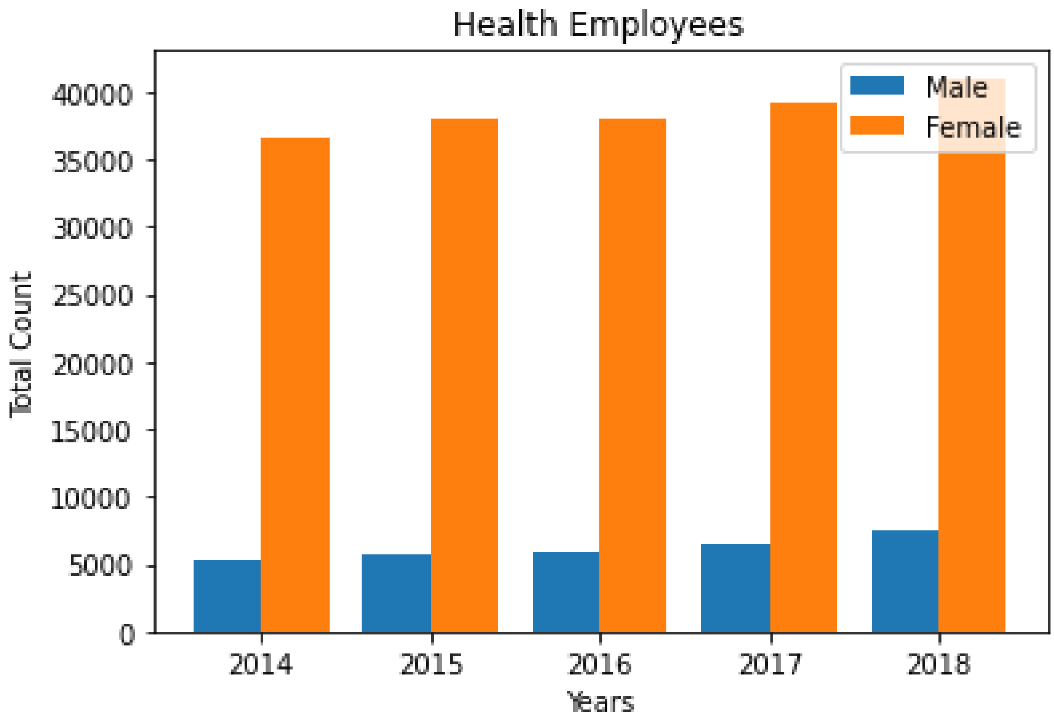
[48]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_finance_trend)
plt.title('Female Finance Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Finance Employees')
plt.show()
```



The number of female employees in the finance sector slightly shows downward trends after 2017. This downward trend might be continued till 2025.

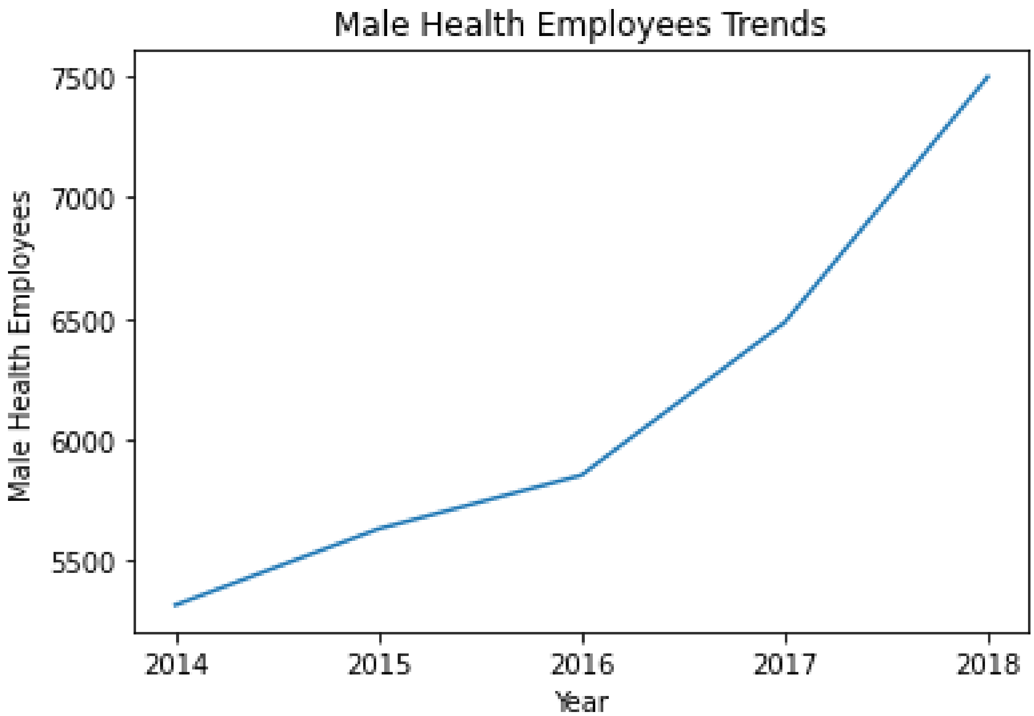
Health

[49]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_health_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_health_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Health Employees")
plt.legend()
plt.show()
```

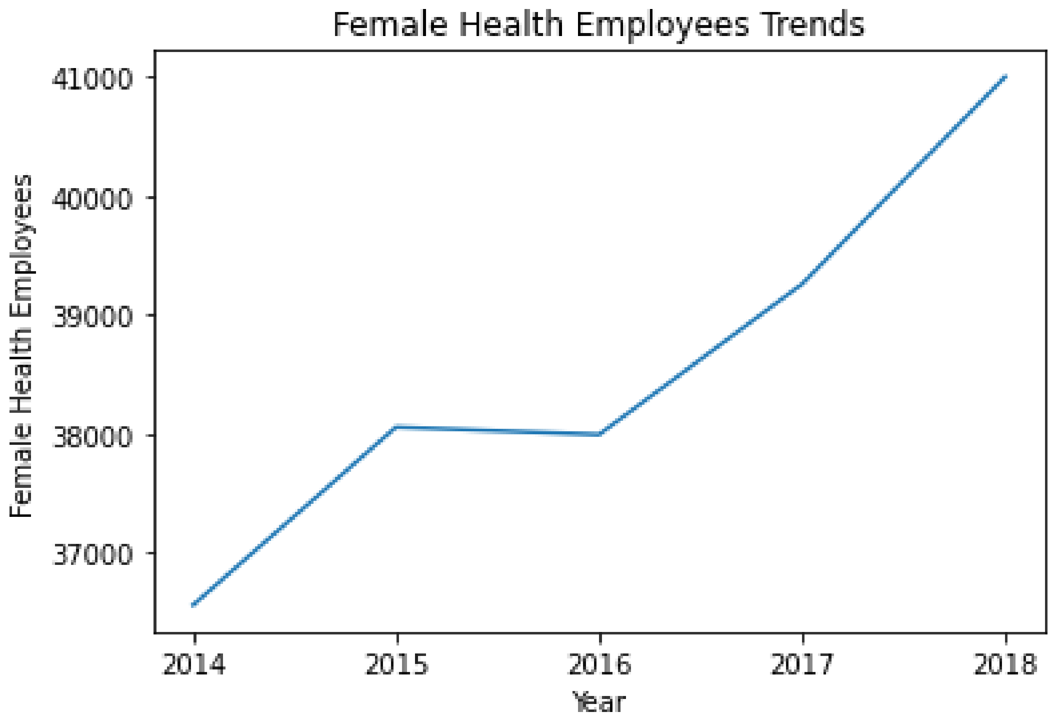


The number of female employees in the health sector indicates strong large gap compare to males.

[50]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_health_trend)
plt.title('Male Health Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Health Employees')
plt.show()
```



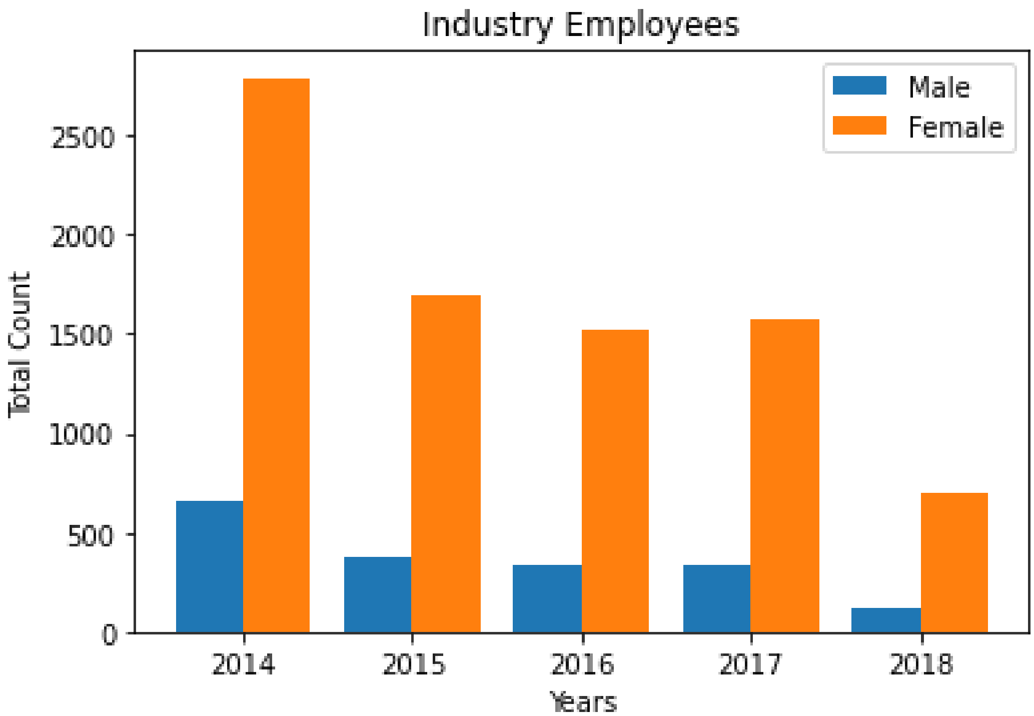
[51]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_health_trend)
plt.title('Female Health Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Health Employees')
plt.show()
```



Both graph represent strong upward trend, so I can guess health sector employees will be increased till 2025.

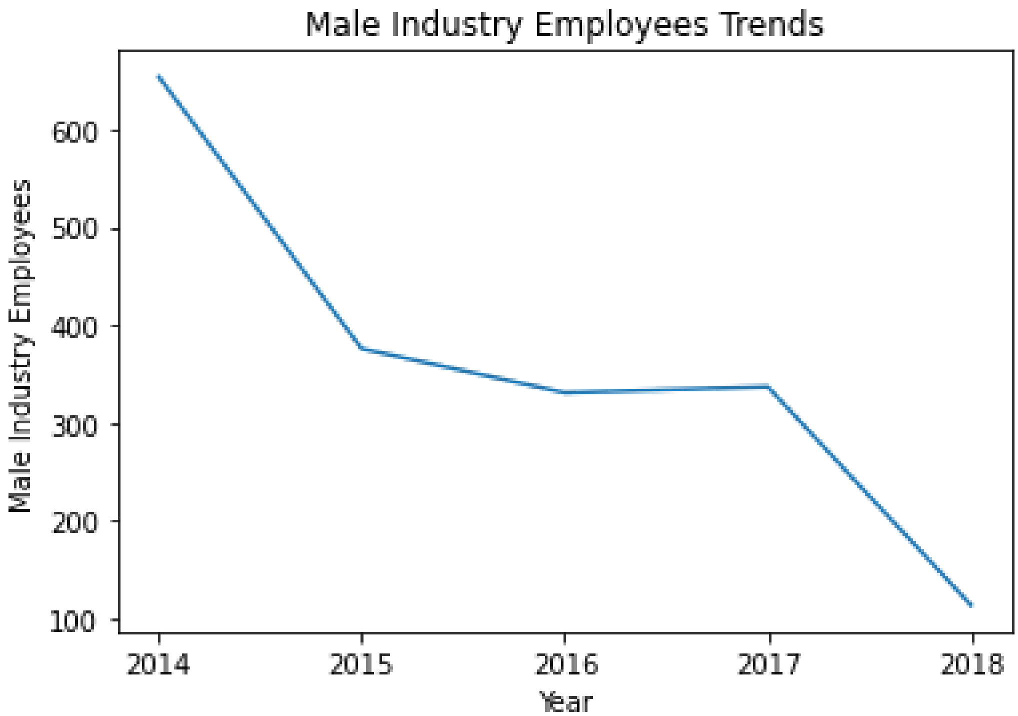
Industry

[52]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_industry_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_industry_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Industry Employees")
plt.legend()
plt.show()
```



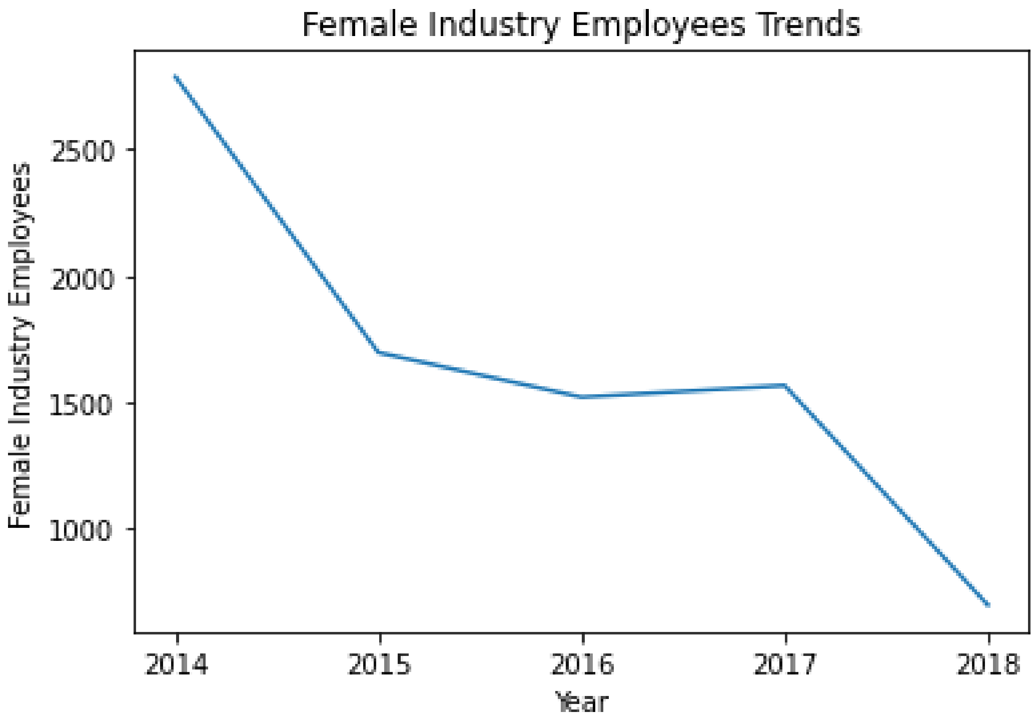
Both graph represent downward trend, so I can guess industry sector employees will be decreased till 2025.

[53]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_industry_trend)
plt.title('Male Industry Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Industry Employees')
plt.show()
```



Very few males are employeed in 2018(less than 150)

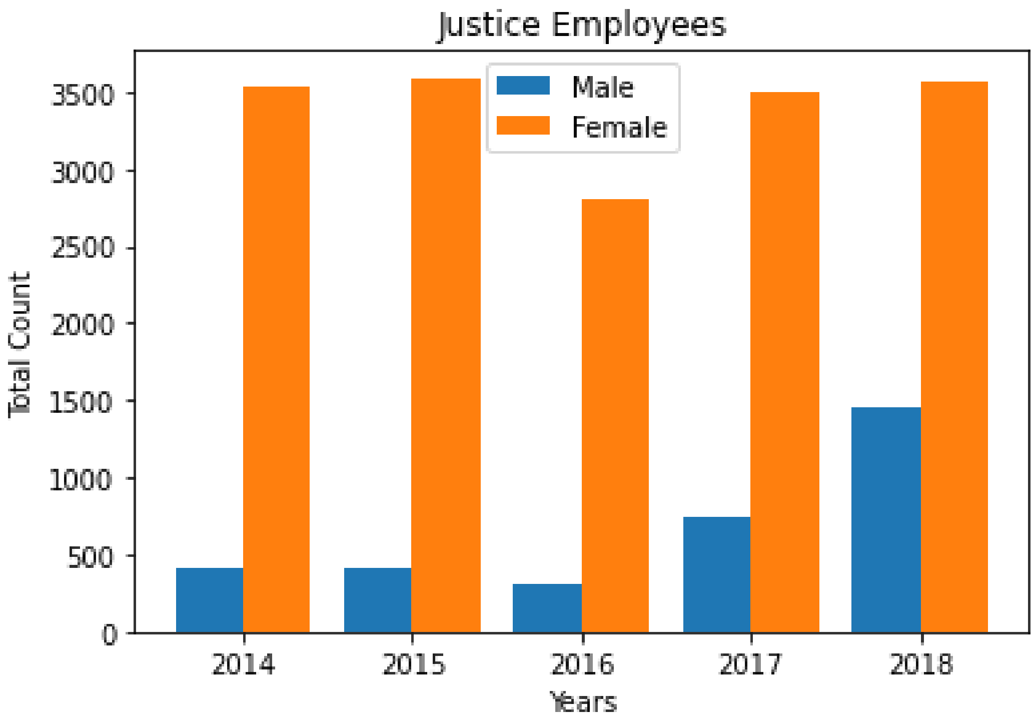
[54]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_industry_trend)
plt.title('Female Industry Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Industry Employees')
plt.show()
```



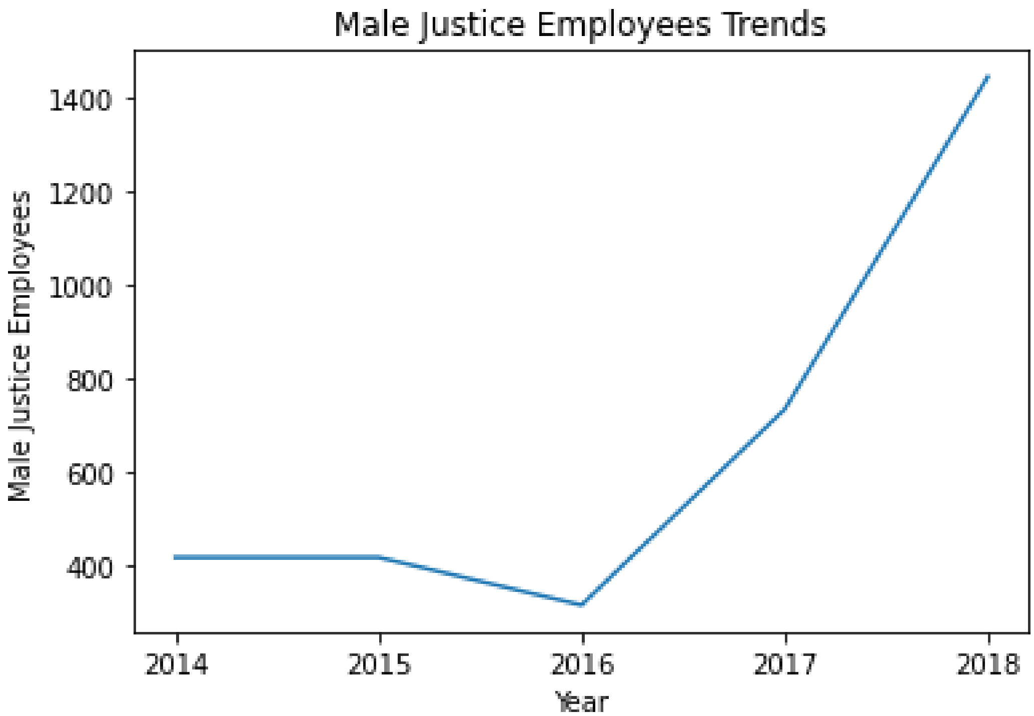
The graph represent downward trend.

Justice

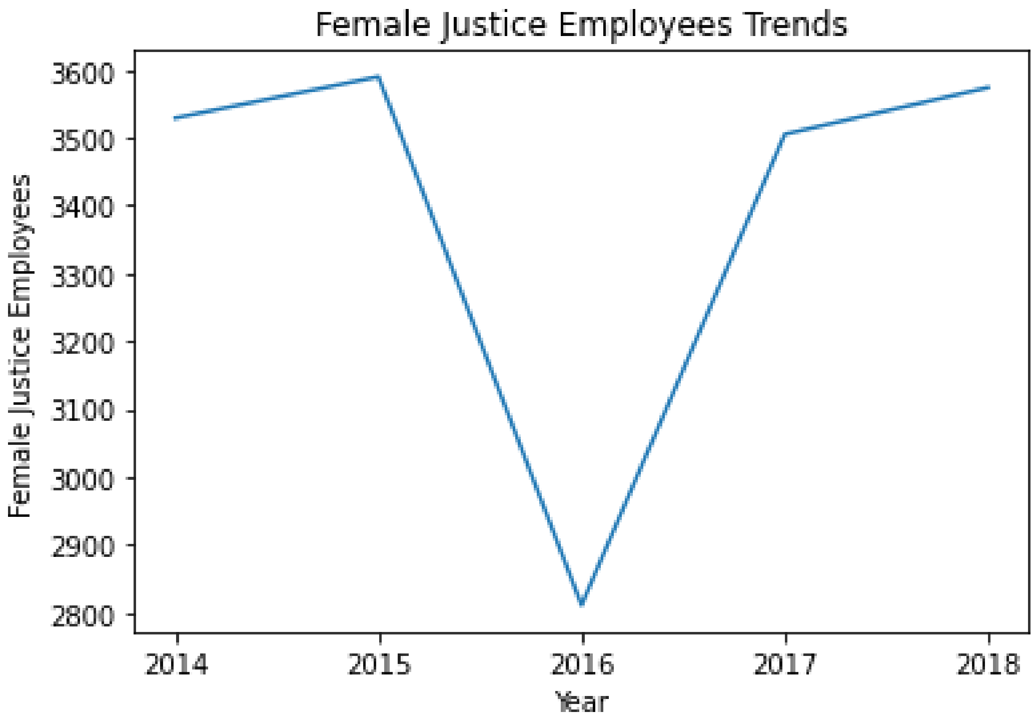
[55]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_justice_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_justice_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Justice Employees")
plt.legend()
plt.show()
```



[56]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_justice_trend)
plt.title('Male Justice Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Justice Employees')
plt.show()
```



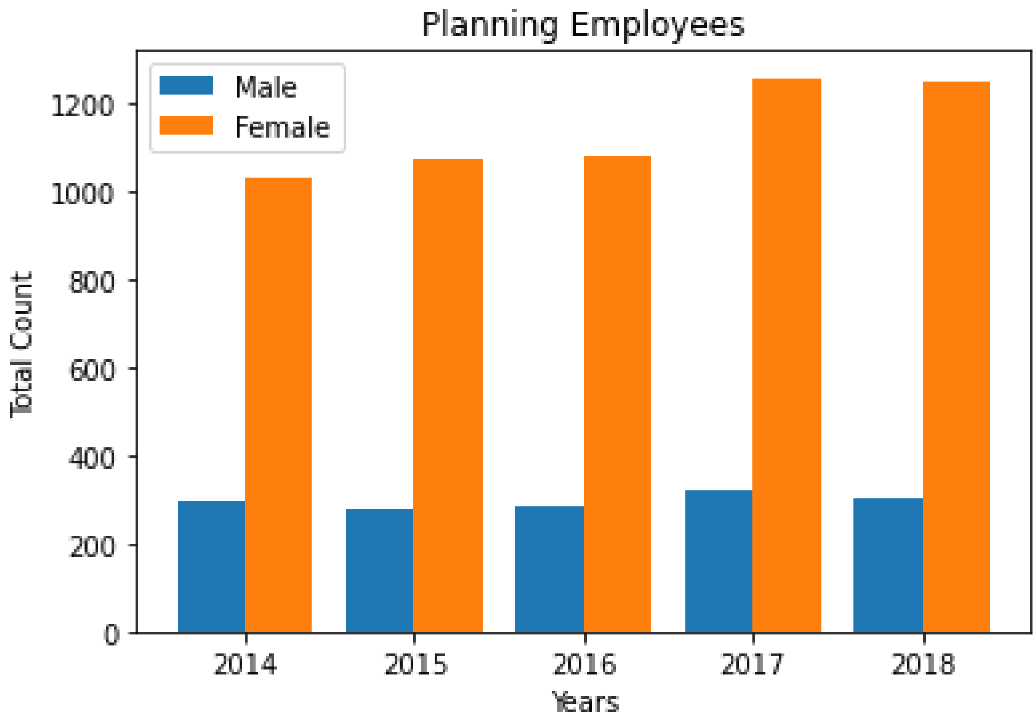
[57]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_justice_trend)
plt.title('Female Justice Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Justice Employees')
plt.show()
```



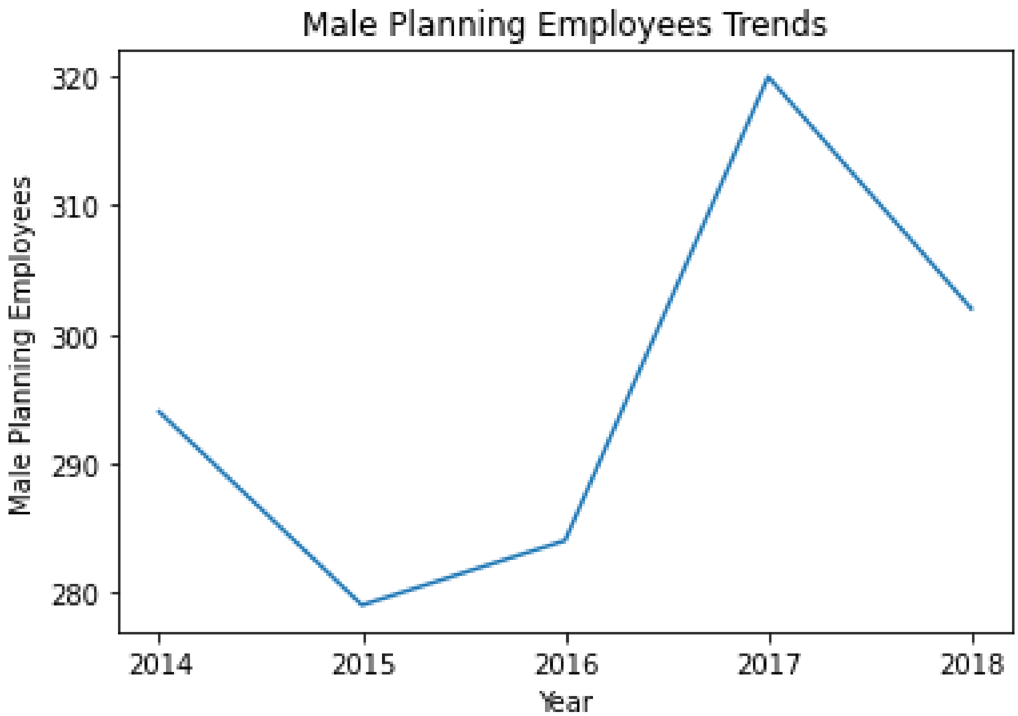
Both graphs represent a similar pattern. They had a slight fall in 2016 but recovered and showing a upward trend at the moment.

Planning

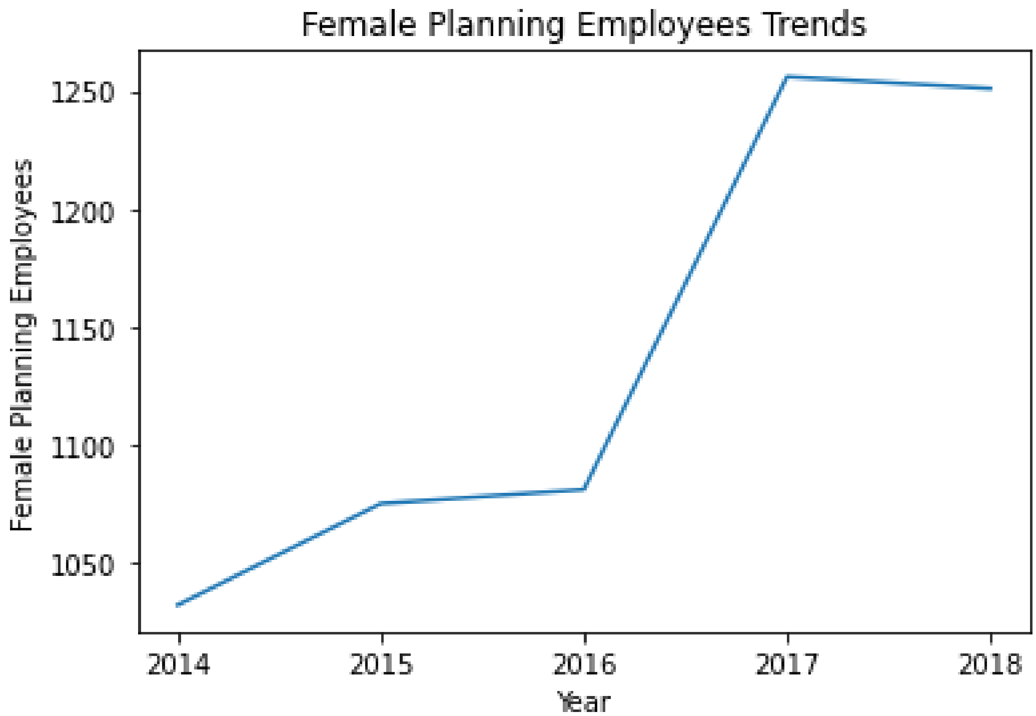
[58]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_planning_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_planning_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Planning Employees")
plt.legend()
plt.show()
```



[59]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_planning_trend)
plt.title('Male Planning Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Planning Employees')
plt.show()
```



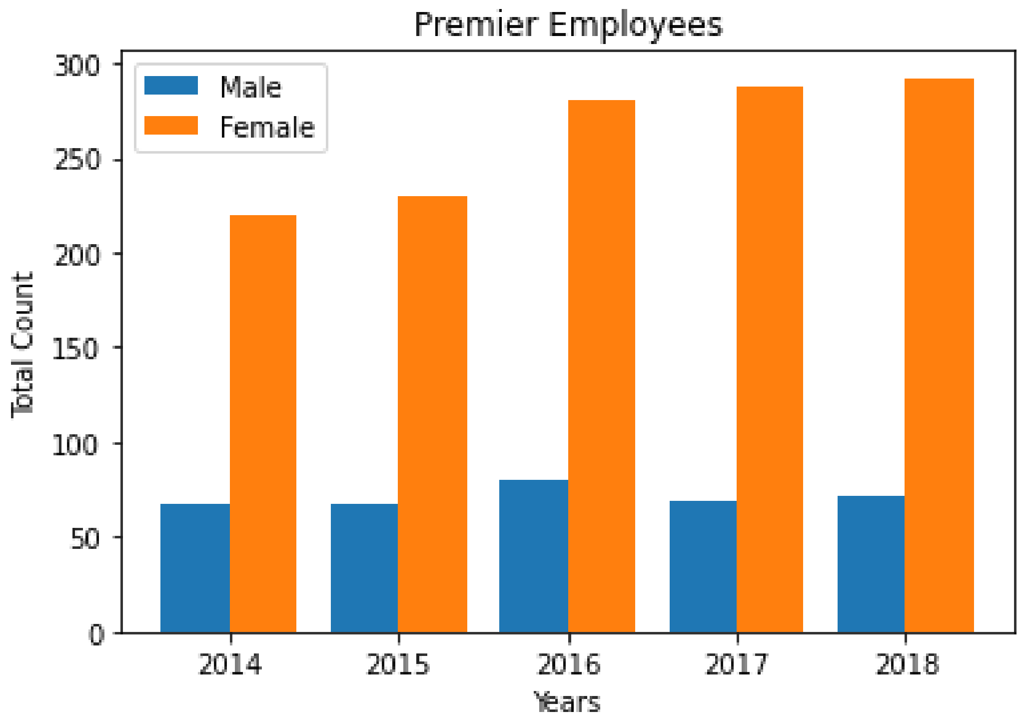
[60]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_planning_trend)
plt.title('Female Planning Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Planning Employees')
plt.show()
```



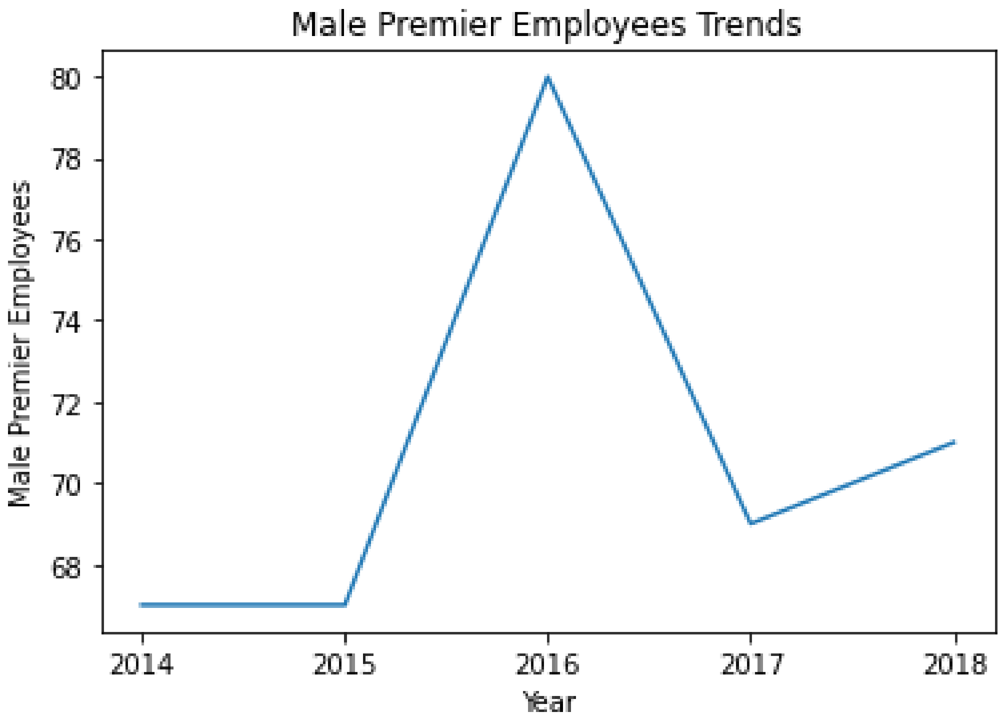
Both graphs represent a similar pattern. They had an increase from 2016 to 2017 but fall down right after, so it is really hard to predict 2025.

Premier

[61]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_premier_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_premier_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Premier Employees")
plt.legend()
plt.show()
```

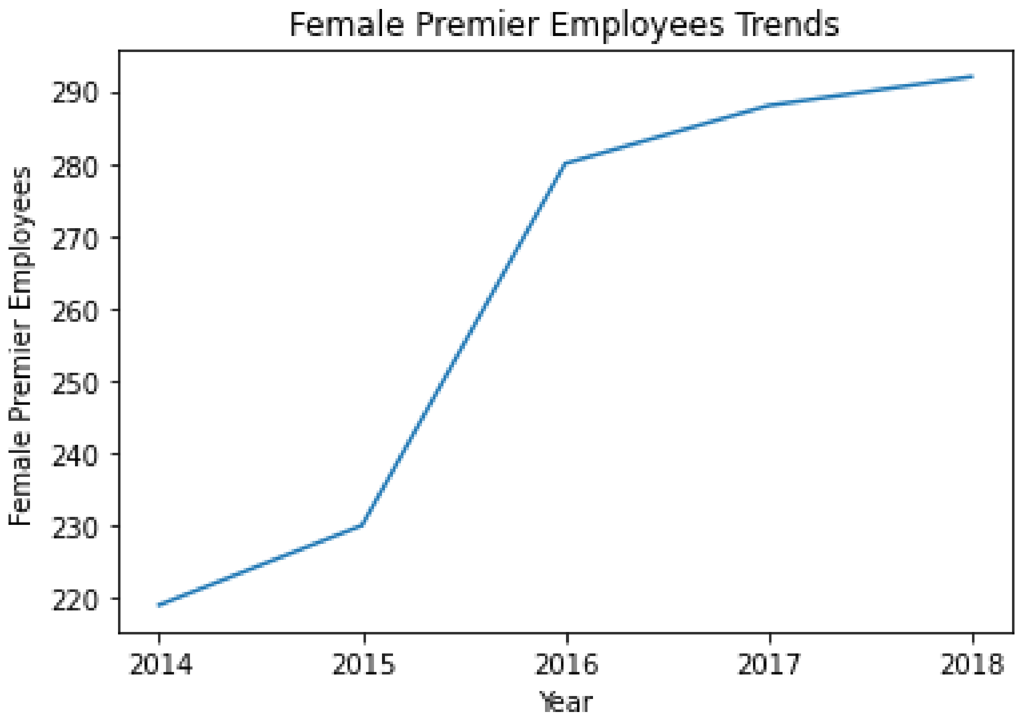


[62]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_premier_trend)
plt.title('Male Premier Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Premier Employees')
plt.show()
```



The graph is really unstable, so it is hard to guess the 2025 trend.

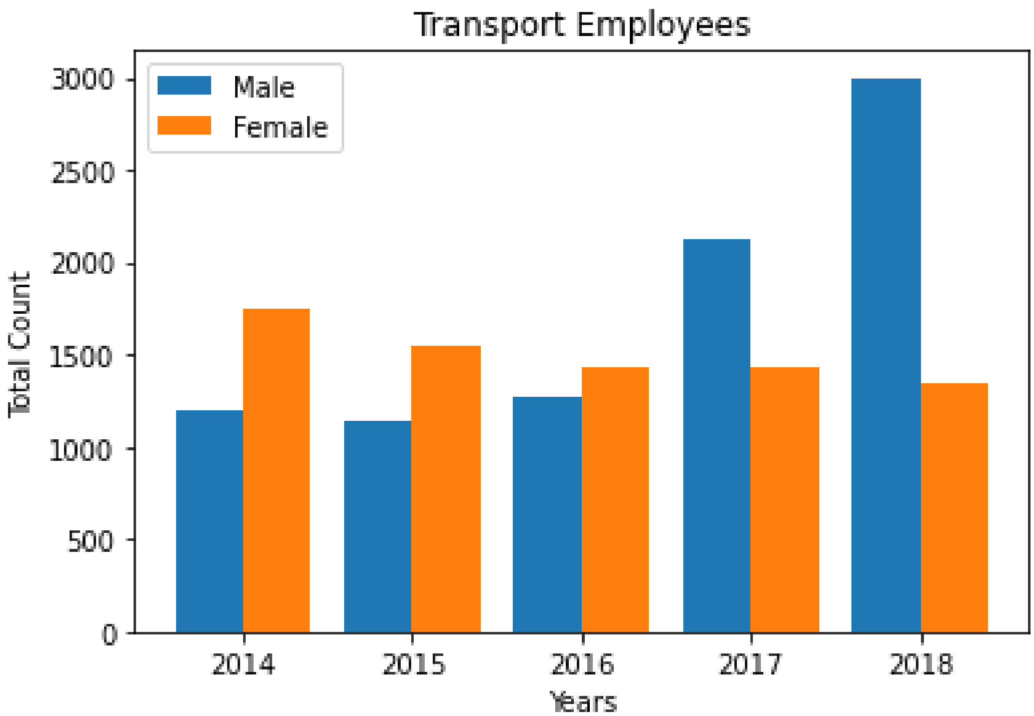
[63]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_premier_trend)
plt.title('Female Premier Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Premier Employees')
plt.show()
```



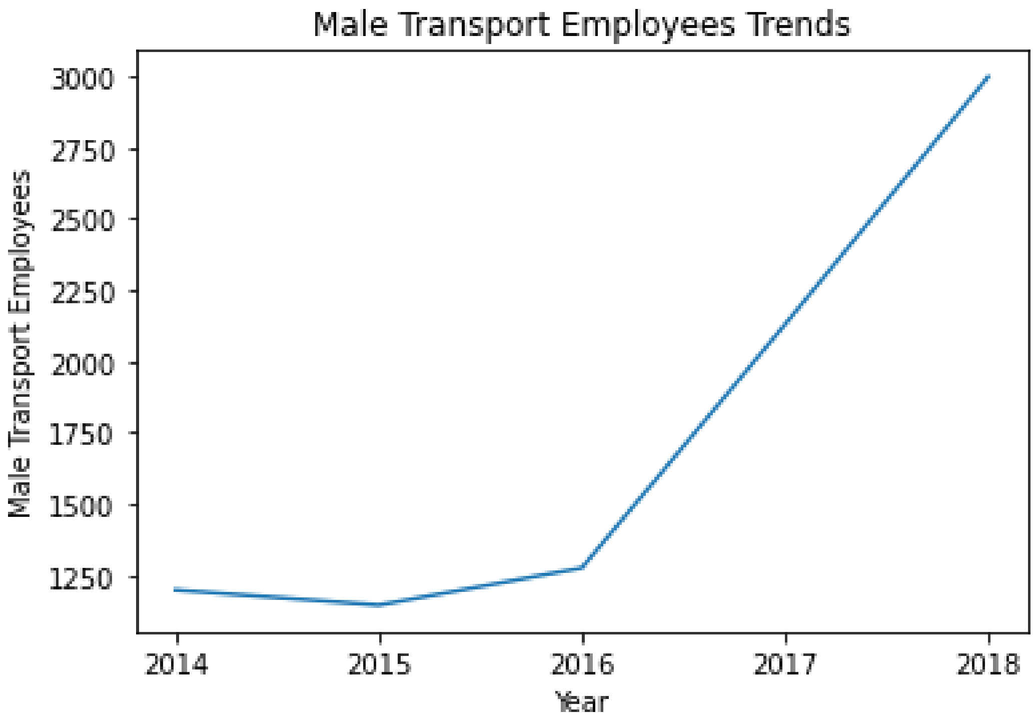
The female employees keep increasing since 2014, but the slope is getting lower. Therefore, I assume that the growth would be steady in 2025.

Transport

[64]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_transport_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_transport_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Transport Employees")
plt.legend()
plt.show()
```

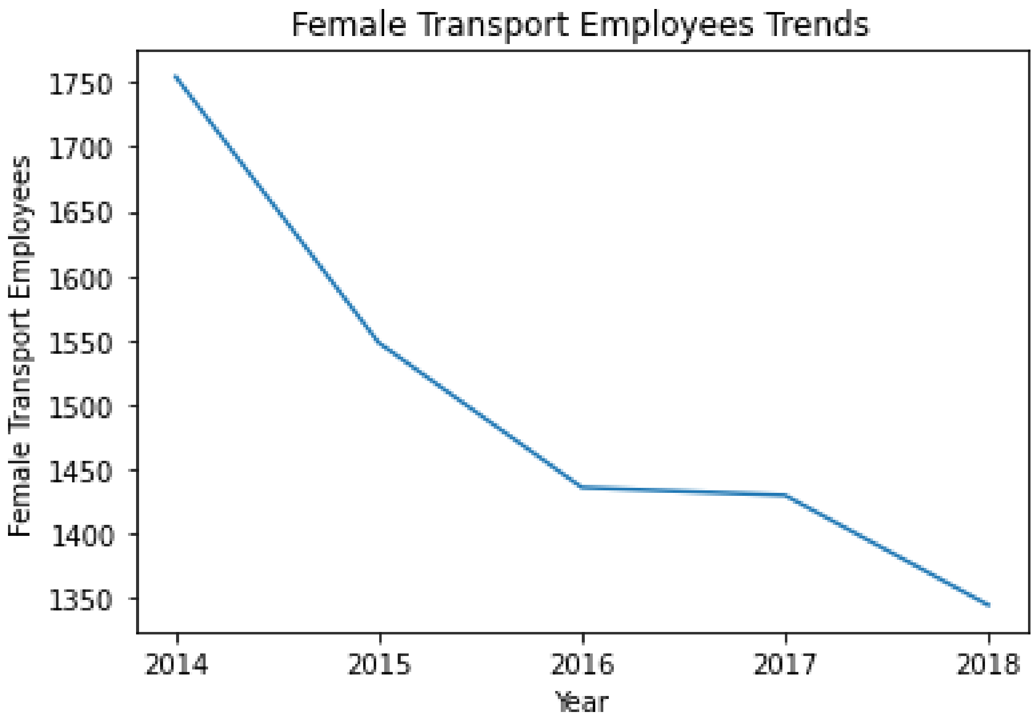


[65]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_transport_trend)
plt.title('Male Transport Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Transport Employees')
plt.show()
```



[66]:

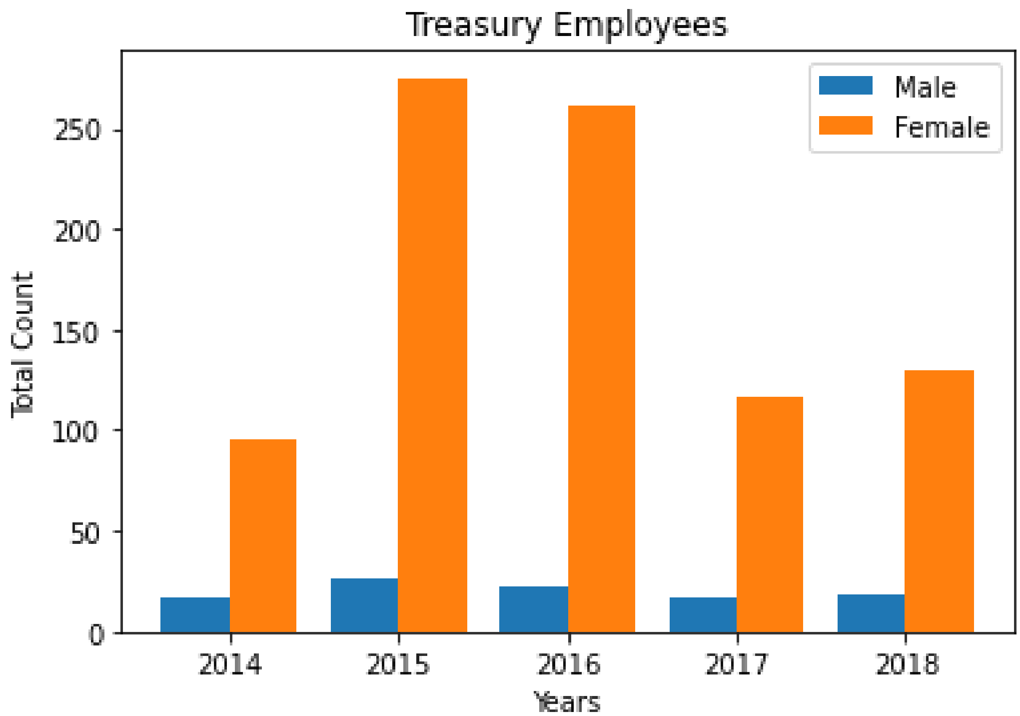
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_transport_trend)
plt.title('Female Transport Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Transport Employees')
plt.show()
```



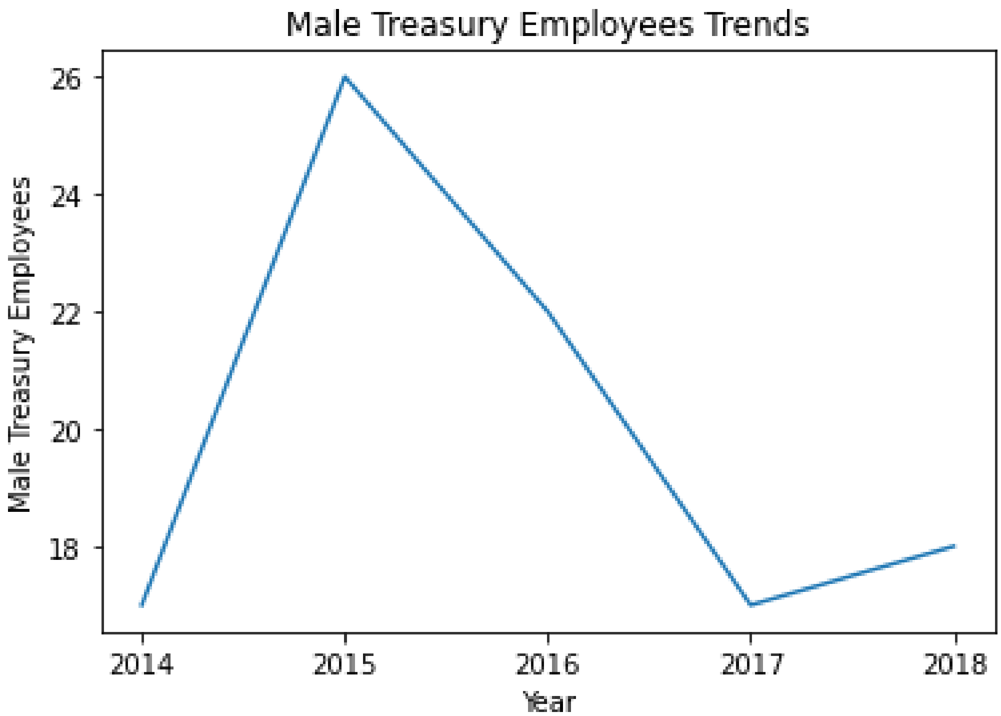
Each graph indicates the opposite trend. While the number of female employees is decreasing, the number of male employees is increasing. In 2017, for the first time, the number of male employees surpassed females.I can assume there will be more gaps in 2025.

Treasury

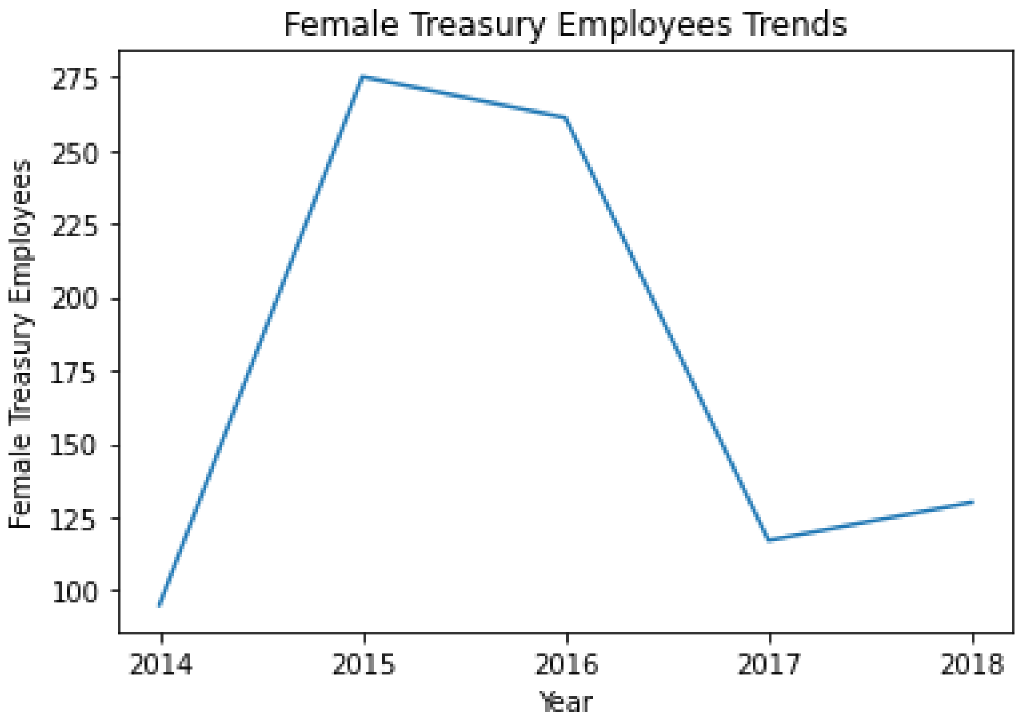
[67]:
```
X = ['2014','2015','2016','2017','2018']

X_axis = np.arange(len(X))

plt.bar(X_axis - 0.2, male_part_treasury_trend, 0.4, label = 'Male')
plt.bar(X_axis + 0.2, female_part_treasury_trend, 0.4, label = 'Female')

plt.xticks(X_axis, X)
plt.xlabel("Years")
plt.ylabel("Total Count")
plt.title("Treasury Employees")
plt.legend()
plt.show()
```



[68]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, male_part_treasury_trend)
plt.title('Male Treasury Employees Trends')
plt.xlabel('Year')
plt.ylabel('Male Treasury Employees')
plt.show()
```



[69]:
```
Year =['2014','2015','2016','2017','2018']

plt.plot(Year, female_part_treasury_trend)
plt.title('Female Treasury Employees Trends')
plt.xlabel('Year')
plt.ylabel('Female Treasury Employees')
plt.show()
```



Both graphs represent a similar pattern. They had an increase from 2014 to 2015 but fall down right after and recover from 2017 to 2018, so it is really hard to predict 2025.