Gender Pay Gap in Germany

Abstract

In this project, I try to determine the level and reason of gender inequality in Germany 1994 with respect to the annual salary. This inequality manifest itself in a pay gap between male and female workers.

I will use the so called “Census Dataset”, a dataset to predict whether the income exceeds the level of $50 000 per year depending on various census data. The dataset was extracted by Barry Becker from the 1994 Census database and is available in the Machine Learning Depository under https://archive.ics.uci.edu/ml/datasets/Adult (last access: 13.07.2019).

From this dataset, I derive that in case of males the weekly workload, the age, and being a professional specialist can result in the belonging to the class of people earning more that $50K. As for women, many of them work part-time which can explain that they mostly earn less than $50K. Although the data does not allow for a precise analysis, an existence of a pay gap can be derived from the number of people belonging to the class of those making > $50K.

Motivation

Despite the fact that the biggest German political parties Christian Democrats (CDU) and Social Democrats (SPD) are currently lead by women, Dr. Angela Merkel and Andrea Nahles, respectively, Germany is still far away from achieving an equality between genders, both, in politics and in private life. Thus, most of the high-rank politicians are males[1] and women still face a pay gap of 21% as for 2017 and 2018[2] despite the government's effort to help women in child care to allow them to better concentrate on their jobs[3].

I assume that the problem of gender inequality with respect to the difference in salary between men and women was at least the same (if not even worse) in the 90's. In this project, I analyse the level and reasons of pay gap using the dataset collected in 1994.

1. Myfanwy Craigie: Despite appearance, women don't rule in Germany. POLITICO.EU, 24.04.2018, URL: *https://www.politico.eu/article/despite-appearances-women-dont-rule-in-germany (last access: 13.07.2019)*
2. German gender pay gap unchanged at 21 percent. Deutsche Welle, 14.03.2019, URL: https://p.dw.com/p/3F0YL (last access: 13.07.2019)
3. Addressing the gender pay gap: Government and social partner actions – Germany. European Foundation for the Improvement of Living and Working, 24.04.2010, URL: https://www.eurofound.europa.eu/publications/report/2010/addressing-the-gender-pay-gap-government-and-social-partner-actions-germany (last access. 13.07.2019)

Dataset I

The dataset assigns the level of the yearly salary (> $50K or <= $50K) to 14 different attributes containing the

- Age of a person
- Their sex
- Native country
- Race
- Belonging to a work class
- Education
- Martial status
- Relationship
- Occupation
- Working hours per week (Weekly workload)
- Capital gain
- Capital loss
- Fnlwgt: The number of people the census takers believe that observation represents

Dataset II

The parameters “Capital gain” and “Capital loss” denote the gain and loss from private investment activities (like, for instance, buying and selling stokes) and, thus, cannot be connected to the income earned at the working place. Therefore, I will exclude these parameters from consideration.

The further parameter to be excluded is “Fnlwgt”. This is because this parameter denotes the data collection methods and has no importance for my studies.

Data Preparation and Cleaning I

In the Machine Learning Depository, the dataset is saved as a table with the name adult.data.

To work with this dataset using Python, it has to be downloaded and converted to a .csv-file („comma separated file“). The conversion can be easily done by appending the ending .csv to the name of the dataset: adult.data.csv or by appending the ending .txt to filename (adult.data.txt), loading this file to Excel and saving it as a .csv file.

There are several missing values in the dataset. Those are denoted by question marks „?“. These values are replaced by NaN and then dropped.

Since I am interested in a pay gap as indicator of gender inequality specifically in Germany, the data for any countries except for Germany is filtered out.

Data Preparation and Cleaning II

In introduce two classes for year salary:

- > $50K : class 1 and <= $50K : class 0

Any categorical values like education or occupation are replaced by numerical values to train a decision tree for prediction of the payment class (1 or 0) depending on the sex, age, education, and occupation of a person.

Research Questions

- Is there a pay gap between men and women in Germany in 1994?
- What is the averaged weekly workload in hours? Are there are some gender-based specifics?
- What is the weekly workload distribution between women and men?
- Does the increase of weekly workload (or even doing extra hours) lead to a better payment, i.e. belonging to class 1?
- Does the age of people have an impact on the salary?
- Does the education level of people have an impact on the salary?
- Does the occupation have an impact on the salary?
- What is the payment class of a person depending on their sex, age, education, and occupation?

Methods

- Descriptive and exploratory data analysis will be applied to derive knowledge about the gender pay gap in Germany in 1994

- A decision tree will be used to predict a payment class of a person depending on their sex, age, education, and occupation.

Findings I: The pay gap

Question: is there a pay gap between men and women in Germany in 1994?

Results:

⇒ Only 13% of women belong to the better-paid class of people earning > $50K

⇒ But 50% of men belong to this group

So, it seems that there ***is a pay gap*** between women and men in Germany in 1994. However, we cannot determine how big is this pay gap because the data does not provide the information on the precise salary of participants of the census collection.

To make sure that we really have an example of a gender pay gap in Germany in 1994, we have to look at other parameters of the survey, namely on the weekly workload of women and men.

Findings II: Weekly workload

Question: what is the averaged weekly workload in hours? Are there are some gender-based specifics?

Results:

⇒ Women work 36.91 hours per week on average

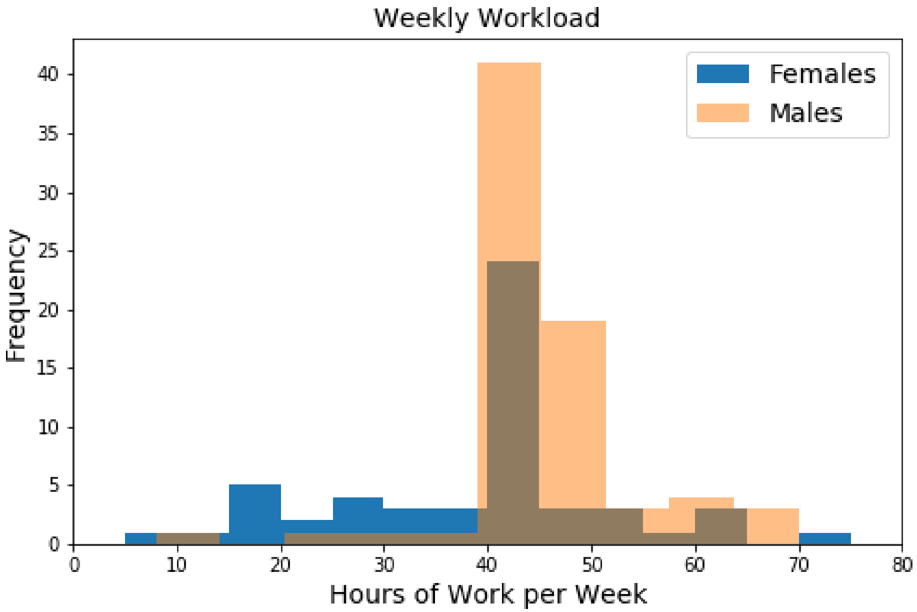
⇒ Men work 45.3 hours per week on average

⇒ ***Assuming a 40-hour week, many women seem to work part-time, whereas many men work extra hours to achieved the averaged values from above. Assumingly, many women work part-time due to child care obligations. However, the dataset does not allow to prove this assumption.***

Findings II: Weekly workload

Question: what is the weekly workload distribution between women and men?

Results:



Findings II: Weekly workload

Results:

⇒ As the figure from the previous slide shows, ***the mode of both distributions*** (men and women) lies slightly above of 40 hours per weeks meaning that working a few additional extra hours occurs quite often for both genders

⇒ The distributions show that there are indeed ***many women working part-time***, whereas ***men often do extra hours***. There are only 3 males from the dataset working part-time

⇒ ***A bigger workload might explain the difference between less- and better-paid men and women. This result needs further studies.***

Findings III: Salary vs. Workload

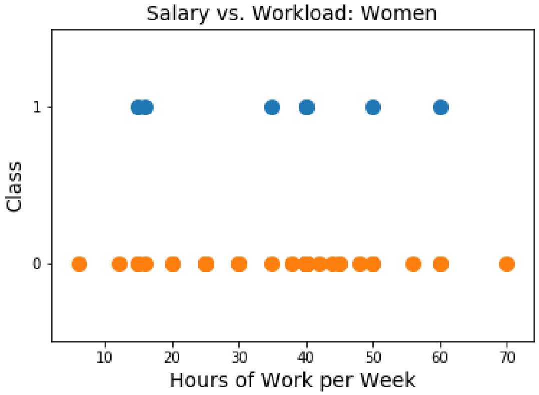
Question: does the increase of weekly workload (or even doing extra hours) lead to a better payment, i.e. belonging to class 1?

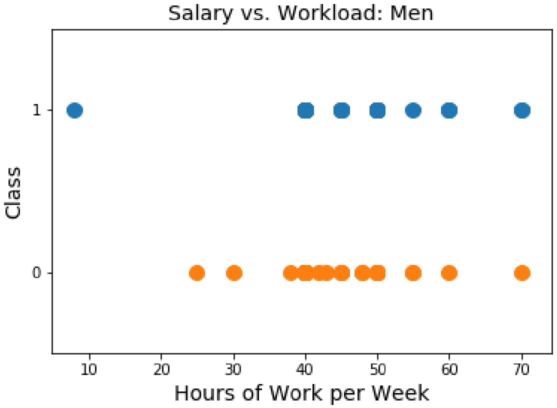
Results:

⇒ There is ***no dependence*** between the workload and payment class in case of ***women***

⇒ In case of ***men***, we can observe ***the tendency*** that an increased workload can indeed result in the belonging to the class of better-paid people (class 1)

⇒ However, in both cases, increased workload ***does not guarantee*** better payment per year.





Findings IV: Salary vs. Age

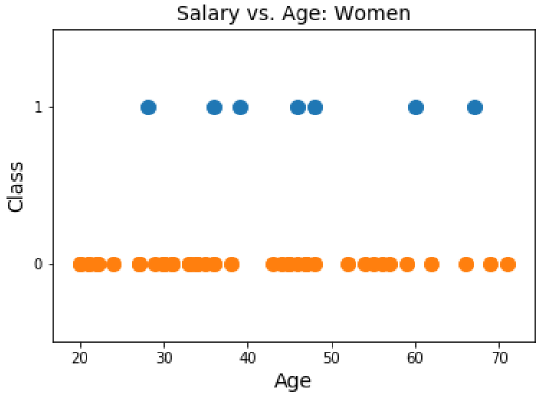
Question: does the age of people have an impact on the salary?

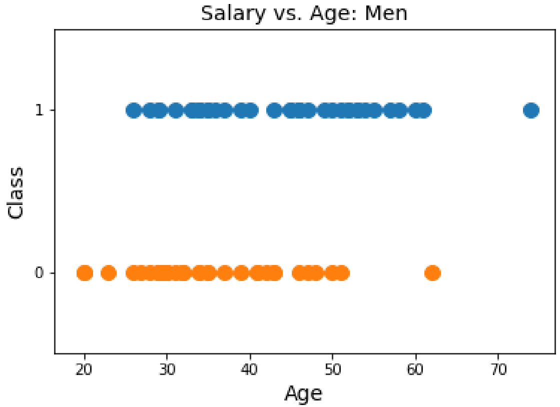
Results:

⇒ There is ***no dependence*** between the age and payment class in case of ***women***

⇒ In case of ***men***, we can observe ***the tendency*** that the probability to belong to class 1 increases with age (especially for men over 40). This can be explained with increased and, thus, a better paid professional experience

⇒ In both cases, there are ***no people*** belonging to class 1 under the ***age of 25***.



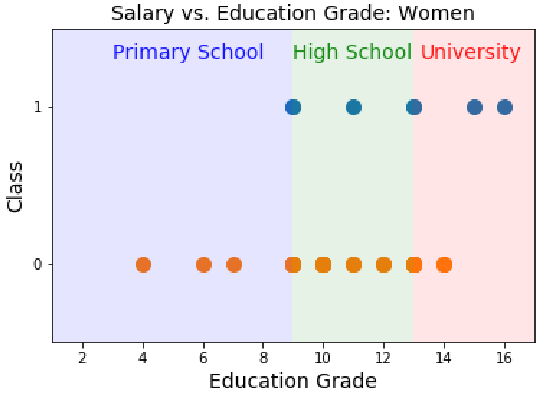


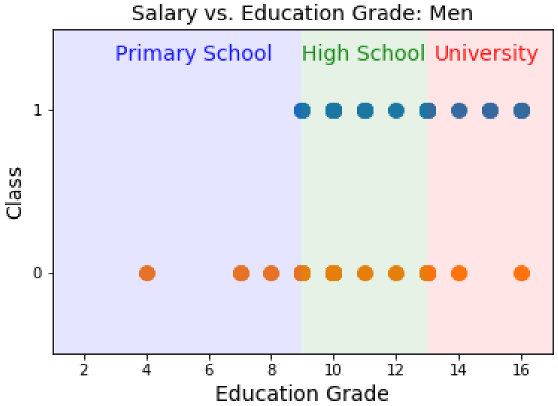
Findings V: Salary vs. Education

Question: does the education level of people have an impact on the salary?
Results:

⇒ In both, ***females and males***, one ***should have*** at least a ***high-school*** certificate to belong to class 1

⇒ However, having a high-school or even a university certificated ***does not guarantee*** the belonging to class 1





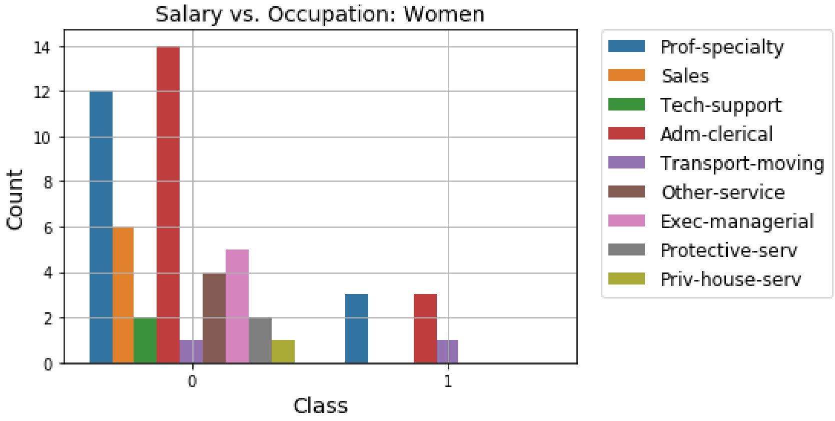
Findings V: Salary vs. Occupation

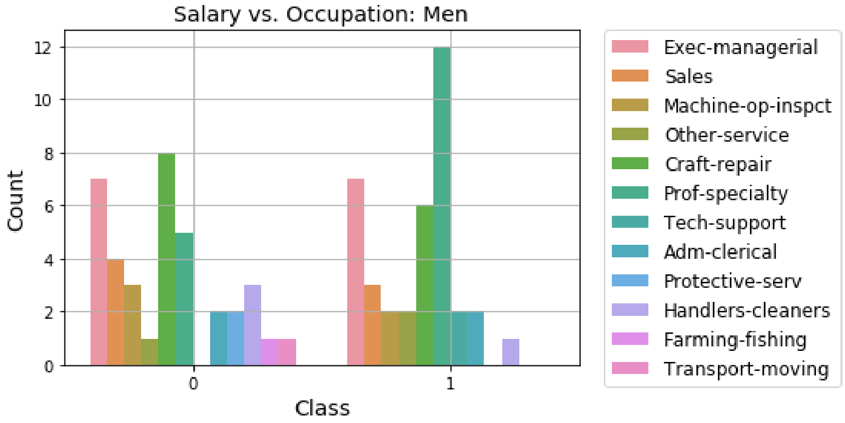
Question: does the occupation have an impact on the salary?

Results:

⇒ There are ***no clear results*** on the dependence between the job and salary ***for women***

⇒ As for **men**, being a ***professional specialist*** (Prof-specialty) increases the chances to belong to class 1. Any other occupations are almost equally distributed between class 0 and 1.





Findings V: Prediction of the payment class

Question: what is the payment class of a person depending on their sex, age, education, and occupation?

Results:

⇒ Using a decision tree, I predict that both, a 35-old male and 35-old female having a university degree and working in Sales would belong to class 1

⇒ However, since the dataset used for training of the decision tree is quite small having only 128 samples for both, males and females, the prediction results are not trustworthy.

Limitations

⇒ To better determine the pay gap and its reasons, the value of the annual salary should be continuous representing the salary of every asked person and not categorical having only two classes (> $50K and <= $50K)

⇒ The dataset specifically for Germany was to small (128 samples) for a detailed analysis and deploying machine learning techniques. Especially the data for women do not provide sufficient amount of information to make solid conclusions. A much bigger set is needed for detailed studies of the pay gap.

Conclusions

- There seem to be a gender pay gap: 50% of men make > $50K per year, whereas only 13% of women belong to this group.
- Many women seem to work part-time, whereas many men work extra hours.
- For women, there is no obvious dependence between the weekly workload, whereas for men we can see that working extra hours might result in a salary > $50K. However, extra hours do not guarantee a better salary.
- There is no obvious dependence between the age of women and their belonging to the class of better-paid people (> $50K). For men, we observe the tendency to belong to the class of people making > $50K for the for the age over 40. This might have to do with the development of their career.
- To belong to the class of better paid people (> $50K) both groups, men and women, should at least have a high school certificate. However, neither a high school nor a university certificate guarantees a better-paid job.
- There is no obvious dependence between the occupation of a woman and her belonging to a certain salary class. As for men, being a professional specialist may pay off: there are many male professional specialists in the class making > $50K.

Thanks a lot for your attention!

In [1]:
```
# Import Python moduls

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from numpy import array
from numpy import argmax
from sklearn.preprocessing import LabelEncoder
from sklearn import tree
```

Data information

Attribute Information:

"Listing of attributes: >50K, <=50K.

age: continuous.

- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands."

Data Source: https://archive.ics.uci.edu/ml/datasets/Adult (https://archive.ics.uci.edu/ml/datasets/Adult)

In [2]:
```
# Read the data
names = ['age', 'workclass', 'fnlwgt', 'education', 'education-num', 'martial-st
atus',
          'occupation', 'relationship', 'race', 'sex',
         'capital-gain', 'capital-loss', 'hours-per-week', 'native-country', 'sal
ary_year']
data = pd.read_csv('./Data/adult.csv', names = names, sep = ' ')

print(data.shape)
```

```
(32561, 15)
```

In [3]:

```
# Choose Germany

data_ger = data[data['native-country'] == 'Germany']
print(data_ger.shape)
data_ger.head(5)
```

```
(137, 15)
```

Out[3]:

	age	workclass	fnlwgt	education	education-num	martial-status	occupation	relationship	race	
122	30	Private	77143	Bachelors	13	Never-married	Exec-managerial	Own-child	Black	Ma
280	22	Private	34918	Bachelors	13	Never-married	Prof-specialty	Not-in-family	White	Fe
767	22	Private	151790	Some-college	10	Married-civ-spouse	Sales	Wife	White	Fe
1009	26	Private	109186	Some-college	10	Married-civ-spouse	Sales	Husband	White	Ma
1142	59	Federal-gov	212448	HS-grad	9	Widowed	Sales	Unmarried	White	Fe

In [4]:

```
# Pop columns that are not useful ('fnlwgt', 'race', 'capital-gain', 'capital-lo
ss')
# or meanwhile redundant ('native-country')

data_ger.pop('fnlwgt')
data_ger.pop('race')
data_ger.pop('capital-gain')
data_ger.pop('capital-loss')
data_ger.pop('native-country')
print(data_ger.shape)
data_ger.head(5)
```

```
(137, 10)
```

Out[4]:

	age	workclass	education	education-num	martial-status	occupation	relationship	sex	hours-per-week	sa
122	30	Private	Bachelors	13	Never-married	Exec-managerial	Own-child	Male	40	<
280	22	Private	Bachelors	13	Never-married	Prof-specialty	Not-in-family	Female	15	<
767	22	Private	Some-college	10	Married-civ-spouse	Sales	Wife	Female	30	<
1009	26	Private	Some-college	10	Married-civ-spouse	Sales	Husband	Male	50	<
1142	59	Federal-gov	HS-grad	9	Widowed	Sales	Unmarried	Female	40	<

In [5]:

```
data_ger.is_copy = False
```

```
C:\Users\zajnulim\AppData\Local\Continuum\anaconda3\lib\site-packages\pandas\c
ore\generic.py:4384: FutureWarning: Attribute 'is_copy' is deprecated and will
be removed in a future version.
  object.__getattribute__(self, name)
C:\Users\zajnulim\AppData\Local\Continuum\anaconda3\lib\site-packages\pandas\c
ore\generic.py:4385: FutureWarning: Attribute 'is_copy' is deprecated and will
be removed in a future version.
  return object.__setattr__(self, name, value)
```

In [6]:

```
# Class assignment: Class 1: > 50K, class 0: <= 50K

data_ger.salary_year.replace(['>50K', '<=50K'], [1, 0], inplace = True)
```

In [7]:

```
# Handle missing values: some fiels are filled with string signes '?'.
# They are to be replaced by NaN values and then droped.

data_ger1 = data_ger.copy()
data_ger1 = data_ger1.replace('?', np.NaN)
data_ger2 = data_ger1.dropna(subset=['workclass', 'occupation'])
data_ger2.shape
```

Out[7]:

```
(128, 10)
```

Exploratory Data Analysis

1) What is the percentage of women who earn more than 50K per year in Germany in 1994. What is the according percentage of men?

In [9]:

```
# Split the data in a dataset containing only males and a dataset containing onl
y females

male = data_ger2[data_ger2.sex == 'Male']
female = data_ger2[data_ger2.sex == 'Female']
```

In [10]:

```
# Shows how many counts belong to class 1 and how many belong to class 0 in male
and female
ff = female.salary_year.value_counts()
mm = male.salary_year.value_counts()
print('Class distribution for females :')
print(ff)
print('-------------------------------')
print('Class distribution for males :')
print(mm)
```

```
Class distribution for females :
0    47
1     7
Name: salary_year, dtype: int64
-------------------------------
Class distribution for males :
1    37
0    37
Name: salary_year, dtype: int64
```

In [11]:

```
# Calculate the percentage of females and males earning more than 50K per year

ff_percent = round((ff[1] / (ff[0] + ff[1])) , 2) * 100

mm_percent = round((mm[1] / (mm[0] + mm[1])) , 2) * 100

print('The percentage of women who earn more than 50K per year is: ' + str(ff_pe
rcent) + '%')
print('The percentage of men who earn more than 50K per year is: ' + str(mm_perc
ent) + '%')
```

```
The percentage of women who earn more than 50K per year is: 13.0%
The percentage of men who earn more than 50K per year is: 50.0%
```

Result: the percentage of women who are well paid, i.e. earn more than 50K per year, is only 13%, whereas the percentage of well-paid men is 50%. This might lead to the conclusion that there is a huge gender inequality in Germany in 1994, the year the data were collected. However, further studies are needed to find out if women in Germany are really disadvantaged in terms of payment.

How many hours do males and females work in Germany?

In [12]:

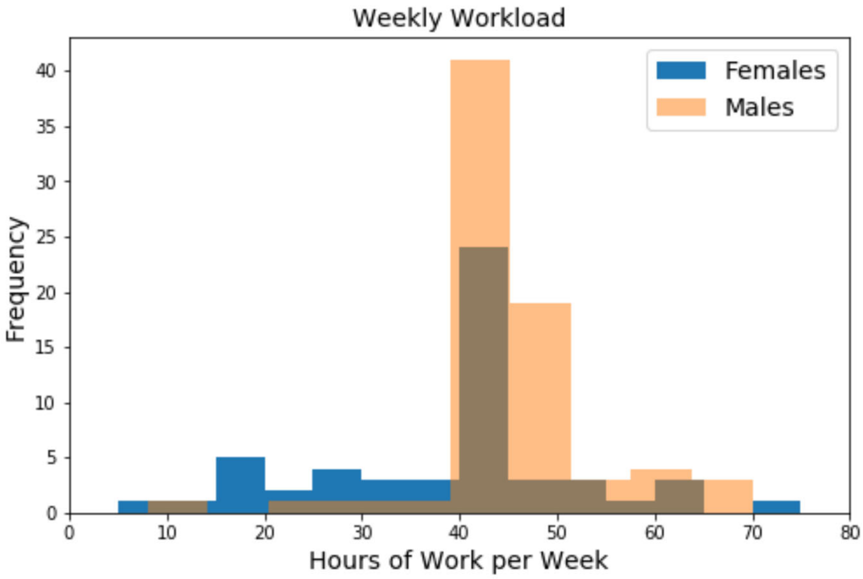
```
# Hours per week on avarage

fem_work = female['hours-per-week'].mean()
man_work = male['hours-per-week'].mean()

print('In Germany, women work ' + str(round(fem_work, 2)) + ' hours per week on
average.')
print('In Germany, men work ' + str(round(man_work, 2)) + ' hours per week on av
erage.')
```

```
In Germany, women work 36.91 hours per week on average.
In Germany, men work 45.3 hours per week on average.
```

In [13]:

```
# Considering the workload of 40 hours per week, find out what is the percentage
#of men / women who work extra hours and part time

# Extra hours
extra_hours_men = male[male['hours-per-week'] > 40]
extra_hours_men_percent = round((extra_hours_men.shape[0] / (mm[0] + mm[1])) * 1
00)

extra_hours_female = female[female['hours-per-week'] > 40]
extra_hours_female_percent = round((extra_hours_female.shape[0] / (ff[0] + ff
[1])) * 100)

# Part time work
part_time_men = male[male['hours-per-week'] < 40]
part_time_men_percent = round((part_time_men.shape[0] / (mm[0] + mm[1])) * 100)

part_time_female = female[female['hours-per-week'] < 40]
part_time_female_percent = round((part_time_female.shape[0] / (ff[0] + ff[1])) *
100)

print('The percentage of men who worked extra hours is: ' + str(extra_hours_men_
percent) + '%')
print('The percentage of women who worked extra hours is: ' + str(extra_hours_fe
male_percent) + '%')
print('-----------------------------------------------------')
print('The percentage of men who worked part time is: ' + str(part_time_men_perc
ent) + '%')
print('The percentage of women who worked part time is: ' + str(part_time_female
_percent) + '%')

# Plot distribution of workload per week for men and women

plt.figure(figsize=(8,5))
plt.hist(female['hours-per-week'], bins= np.arange(0, 80, 5))
plt.hist(male['hours-per-week'], alpha=0.5)
plt.title('Weekly Workload', fontsize = 14)
plt.xlabel('Hours of Work per Week', fontsize = 14)
plt.xlim([0, 80])
plt.ylabel('Frequency', fontsize = 14)
plt.legend(['Females', 'Males'], fontsize = 14)
plt.savefig('workload_distribution.png')
plt.show()
```

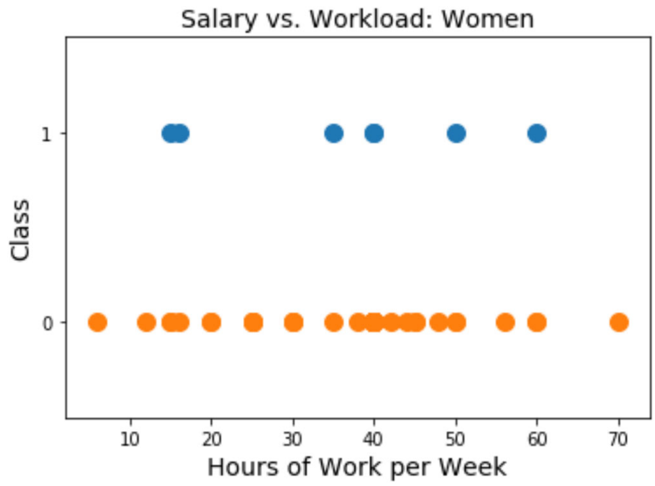
```
The percentage of men who worked extra hours is: 54.0%
The percentage of women who worked extra hours is: 24.0%
---------------------------------------------------
The percentage of men who worked part time is: 5.0%
The percentage of women who worked part time is: 35.0%
```

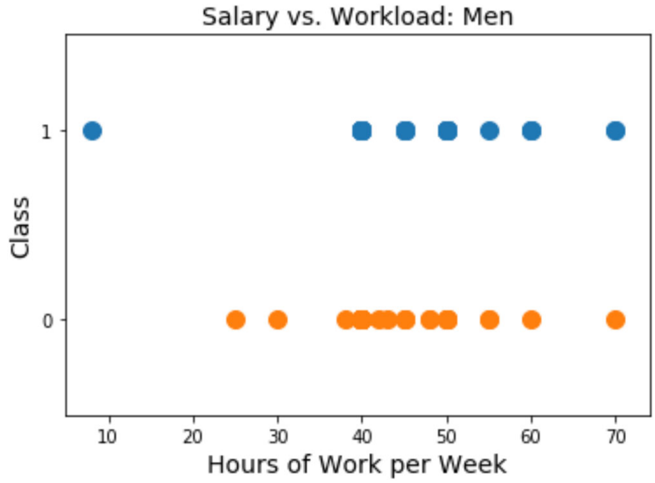


Result: more than the half (54%) of men work extra hours, whereas only 5% work part-time. In case of women, only 24% work extra hours, whereas 35% work part-time. The result is that men work 45.3 hours per week on average, whereas women work almost 36 hours per week. The fact that men in general work more hours per week could explain why then earn more money than women. Let's now consider how the belonging to the better-paid group (class 1) depends on the weekly workload in hours in case of men and women.

In [14]:

```
# Relation between the number of hours per week and the salary

salary_female = female[['hours-per-week','salary_year']]
salary_female_cl1 = salary_female[salary_female['salary_year'] == 1]
salary_female_cl0 = salary_female[salary_female['salary_year'] == 0]

#plt.figure(figsize=(8,5))
plt.scatter(salary_female_cl1['hours-per-week'],
            salary_female_cl1['salary_year'], s = 100)
plt.scatter(salary_female_cl0['hours-per-week'],
            salary_female_cl0['salary_year'], s = 100)
plt.ylim([-0.5, 1.5])
frame1 = plt.gca()
frame1.axes.get_yaxis().set_ticks([0, 1])
plt.xlabel('Hours of Work per Week', fontsize = 14)
plt.ylabel('Class', fontsize = 14)
plt.title('Salary vs. Workload: Women', fontsize = 14)
plt.savefig('salary_workload_female.png')
plt.show()

# -------------------------------------------------------------
salary_male = male[['hours-per-week','salary_year']]
salary_male_cl1 = salary_male[salary_male['salary_year'] == 1]
salary_male_cl0 = salary_male[salary_male['salary_year'] == 0]

#plt.figure(figsize=(8,5))
plt.scatter(salary_male_cl1['hours-per-week'],
            salary_male_cl1['salary_year'], s = 100)
plt.scatter(salary_male_cl0['hours-per-week'],
            salary_male_cl0['salary_year'], s = 100)
plt.ylim([-0.5, 1.5])
frame1 = plt.gca()
frame1.axes.get_yaxis().set_ticks([0, 1])
plt.xlabel('Hours of Work per Week', fontsize = 14)
plt.ylabel('Class', fontsize = 14)
plt.title('Salary vs. Workload: Men', fontsize = 14)
plt.savefig('salary_workload_male.png')
plt.show()
```





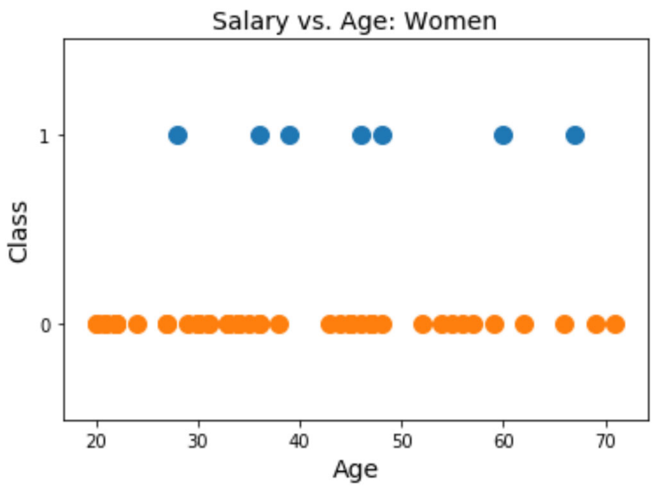
Result: there is no relationship between the weekly workload in hours and the belonging to a class of the better-paid in case of women. Thus, there are two women working only 15 hours per week and earning more than 50K dollars, but there is also a woman who works 60 hours per week to belong to the same class. Unfortunately, the same applies also to women from class 0: you can work 10, 40, or even 70 hours per week and still not be able to make more than 50K dollars per year. As for men, we see the tendency that the belonging to the group of better-paid (class 1) correlates with the weekly workload: the more you work, the better you can be paid. However, also in case of men we have some instances when working more hours per week (50, 60, 70) is not paying off, the men remain in class 0.

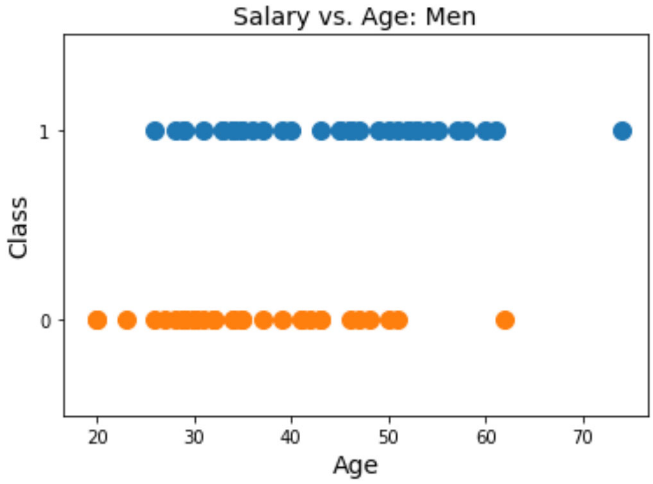
3) What is the reason for gender inequality in terms of payment / salary in Germany in 1994?

3.1 Age

In [15]:

```
# Relation between the age of a person and their salary

salary_female_age = female[['age','salary_year']]
salary_female_age_cl1 = salary_female_age[salary_female_age['salary_year'] == 1]
salary_female_age_cl0 = salary_female_age[salary_female_age['salary_year'] == 0]

plt.scatter(salary_female_age_cl1['age'],
            salary_female_age_cl1['salary_year'], s = 100)
plt.scatter(salary_female_age_cl0['age'],
            salary_female_age_cl0['salary_year'], s = 100)
plt.ylim([-0.5, 1.5])
frame1 = plt.gca()
frame1.axes.get_yaxis().set_ticks([0, 1])
plt.xlabel('Age', fontsize = 14)
plt.ylabel('Class', fontsize = 14)
plt.title('Salary vs. Age: Women', fontsize = 14)
plt.savefig('salary_age_female.png')
plt.show()


salary_male_age = male[['age','salary_year']]
salary_male_age_cl1 = salary_male_age[salary_male_age['salary_year'] == 1]
salary_male_age_cl0 = salary_male_age[salary_male_age['salary_year'] == 0]

plt.scatter(salary_male_age_cl1['age'],
            salary_male_age_cl1['salary_year'], s = 100)
plt.scatter(salary_male_age_cl0['age'],
            salary_male_age_cl0['salary_year'], s = 100)
plt.ylim([-0.5, 1.5])
frame1 = plt.gca()
frame1.axes.get_yaxis().set_ticks([0, 1])
plt.xlabel('Age', fontsize = 14)
plt.ylabel('Class', fontsize = 14)
plt.title('Salary vs. Age: Men', fontsize = 14)
plt.savefig('salary_age_male.png')
plt.show()
```



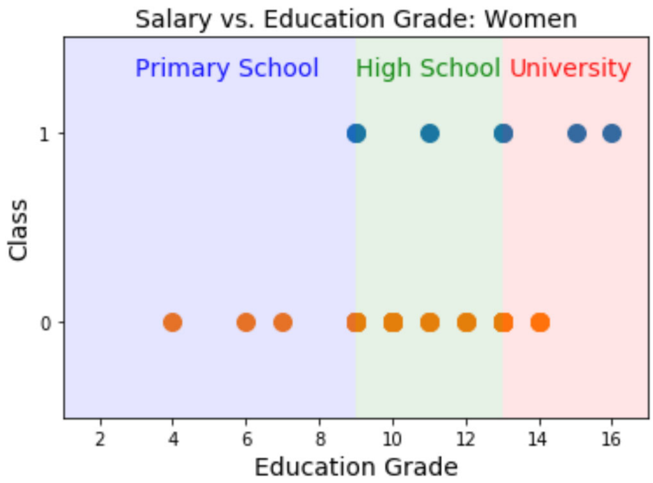


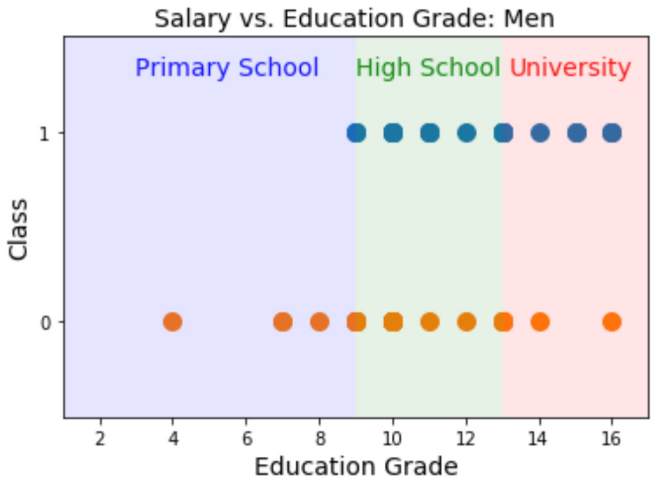
Result: in both cases, there are no people belonging to class 1 under the age of 25. In case of males, there is a tendency that their payment level increases with age. In case, no dependence between the age and the level of payment is recognisable.

3.2 Education

In [17]:

```
# Relation between education and the salary

salary_female_edu = female[['education-num','salary_year']]
salary_female_edu_cl1 = salary_female_edu[salary_female_edu['salary_year'] == 1]
salary_female_edu_cl0 = salary_female_edu[salary_female_edu['salary_year'] == 0]

fig1, ax1 = plt.subplots()
for i in range(9):
    plt.axvspan(i, i+1, facecolor='b', alpha=0.1)
for i in range(9,13):
    plt.axvspan(i, i+1, facecolor='g', alpha=0.1)
for i in range(13,17):
    plt.axvspan(i, i+1, facecolor='r', alpha=0.1)
plt.scatter(salary_female_edu_cl1['education-num'],
            salary_female_edu_cl1['salary_year'], s = 100)
plt.scatter(salary_female_edu_cl0['education-num'],
            salary_female_edu_cl0['salary_year'], s = 100)
plt.ylim([-0.5, 1.5])
plt.xlim([1, 17])
frame1 = plt.gca()
frame1.axes.get_yaxis().set_ticks([0, 1])
plt.xlabel('Education Grade', fontsize = 14)
plt.ylabel('Class', fontsize = 14)
plt.title('Salary vs. Education Grade: Women', fontsize = 14)
ax1.text(3, 1.3, 'Primary School', fontsize=14, color = 'b', alpha = 0.9)
ax1.text(9, 1.3, 'High School', fontsize=14, color = 'g', alpha = 0.9)
ax1.text(13.2, 1.3, 'University', fontsize=14, color = 'r', alpha = 0.9)
plt.savefig('salary_education_female.png')
plt.show()


salary_male_edu = male[['education-num','salary_year']]
salary_male_edu_cl1 = salary_male_edu[salary_male_edu['salary_year'] == 1]
salary_male_edu_cl0 = salary_male_edu[salary_male_edu['salary_year'] == 0]

fig2, ax2 = plt.subplots()
for i in range(9):
    plt.axvspan(i, i+1, facecolor='b', alpha=0.1)
for i in range(9,13):
    plt.axvspan(i, i+1, facecolor='g', alpha=0.1)
for i in range(13,17):
    plt.axvspan(i, i+1, facecolor='r', alpha=0.1)
plt.scatter(salary_male_edu_cl1['education-num'],
            salary_male_edu_cl1['salary_year'], s = 100)
plt.scatter(salary_male_edu_cl0['education-num'],
            salary_male_edu_cl0['salary_year'], s = 100)
plt.ylim([-0.5, 1.5])
plt.xlim([1, 17])
frame1 = plt.gca()
frame1.axes.get_yaxis().set_ticks([0, 1])
plt.xlabel('Education Grade', fontsize = 14)
plt.ylabel('Class', fontsize = 14)
plt.title('Salary vs. Education Grade: Men', fontsize = 14)
ax2.text(3, 1.3, 'Primary School', fontsize=14, color = 'b', alpha = 0.9)
ax2.text(9, 1.3, 'High School', fontsize=14, color = 'g', alpha = 0.9)
ax2.text(13.2, 1.3, 'University', fontsize=14, color = 'r', alpha = 0.9)
plt.savefig('salary_education_male.png')
plt.show()
```





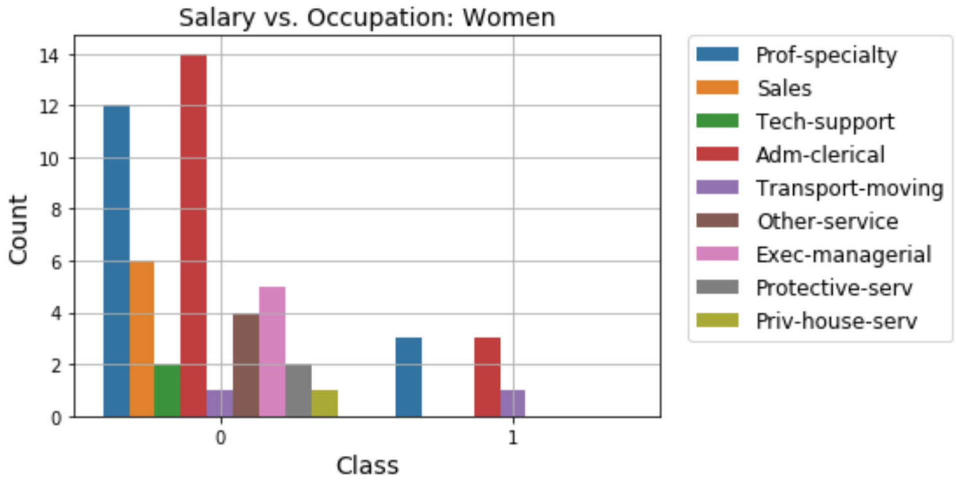
Result: For both, males and females, to have at least at high school certificate is presupposing to get more than 50K Dollars per year. However, neither a high school nor a university certificate guarantees that you will be indeed in the group of better-paid people making >50K.

3.3 Occupation

Let's take a look whether the occupation might have an impact on the belonging to class 1 or class 0.

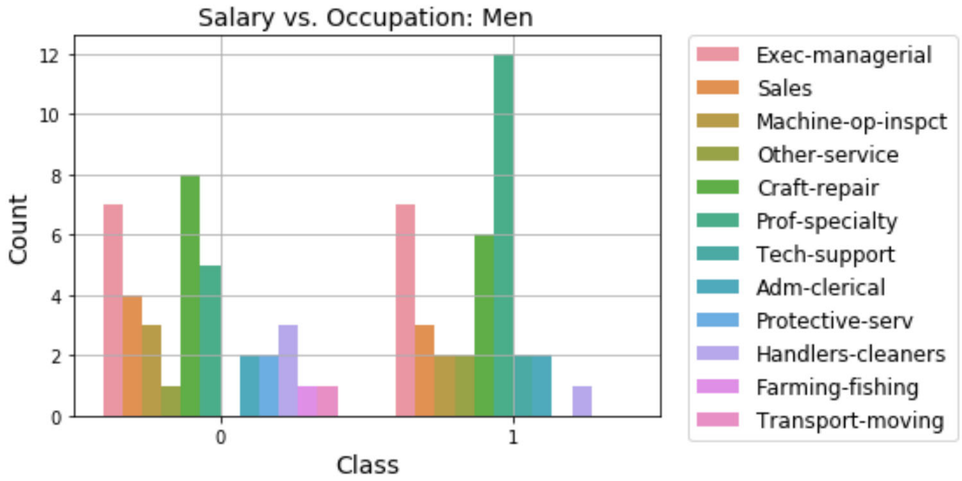
In [22]:

```
# Class depending on the occupation of women
#plt.figure(figsize=(8,5))
ax = sns.countplot(x = "salary_year", hue = "occupation", data=female)
legend = plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0., fontsize
= 12)
plt.grid()
plt.xlabel('Class', fontsize = 14)
plt.ylabel('Count', fontsize = 14)
plt.title('Salary vs. Occupation: Women', fontsize = 14)
plt.savefig('salary_occupation_female.png', bbox_inches='tight')
plt.show()
```



In [24]:

```
# Class depending on the occupation of men

ax = sns.countplot(x = "salary_year", hue = "occupation", data = male)
legend = plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0., fontsize
= 12)
#plt.setp(legend.get_texts(), color='k')
plt.grid()
plt.xlabel('Class', fontsize = 14)
plt.ylabel('Count', fontsize = 14)
plt.title('Salary vs. Occupation: Men', fontsize = 14)
plt.savefig('salary_occupation_male.png', bbox_inches='tight')
plt.show()
```



Result: no clear dependence between the class and the occupation in case of women because there is no enough data available. As for men, we see that being a professional specialist (Prof-specialty) can indeed help you to belong to the group of better-paid people (class 1)

4) Prediction of payment class depending on sex, age, education, occupation

In [303]:

```
# Dataset preparation
xx = data_ger2[['sex', 'age', 'education-num', 'occupation']]
Y = data_ger2['salary_year']
```

In [304]:

```
# Replace cagegorical value 'sex' by 'Male' : 1, 'Female' : 0

values_s = array(xx['sex'])
label_encoder_s = LabelEncoder()
integer_encoded_s = label_encoder_s.fit_transform(values_s)

X1 = xx.copy()
X1.sex.replace(values_s, integer_encoded_s, inplace = True)
```

In [305]:

```
# Replace cagegorical value 'occupation' by
#'Exec-managerial' : 1, 'Prof-specialty' : 2, 'Sales' : 3, 'Machine-op-inspct' : 4,
#       'Other-service' : 5, 'Tech-support' : 6, 'Craft-repair' : 7, 'Adm-clerical' : 8,
#       'Transport-moving' : 9, 'Protective-serv' : 10, 'Handlers-cleaners' : 11,
#       'Farming-fishing' : 12, 'Priv-house-serv' : 13

values_o = X1['occupation'].unique()

X = X1.copy()
X.occupation.replace(values_o, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13], inplace = True)
X.head(5)
```

Out[305]:

	sex	age	education-num	occupation
122	1	30	13	1
280	0	22	13	2
767	0	22	10	3
1009	1	26	10	3
1142	0	59	9	3

In [306]:

```
# DecisionTreeClassifier
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, Y)
```

In [307]:

```
# Predict the class for Male, age 35, education grade = 14, occupation = sales
# ['sex', 'age', 'education-num', 'occupation'] = [1, 35, 14, 3])
prediction_male = clf.predict([[1, 35, 14, 3]])

prediction_female = clf.predict([[0, 35, 14, 3]])

print('A 35-old male working in sales is predicted to be in payment class: ' + str(prediction_male))
print('A 35-old female working in sales is predicted to be in payment class: ' + str(prediction_male))
```

```
A 35-old male working in sales is predicted to be in payment class: [1]
A 35-old female working in sales is predicted to be in payment class: [1]
```

Result: Both, a 35-old male and a 35-old female having a university degree and working in sales are predicted to earn more than 50K per year.