COURSE PROJECT FINAL REPORT

Vasanta Prayaga

Goals and a business objective: The most important goal of this project was to use the United States Census Bureau data provided by XYZ company, to develop an application to find the factors that influence and determine an individual's income. The application needs to predict the salary of an individual based on their profile created with the attributes influencing the salary the most.

For this project, the US Census Bureau dataset provided for analysis had fourteen other attributes or factors that may have influence on the salary an individual might earn. Eight of the fourteen attributes are categorical data. Also, some of the data is continuous and some discrete.

Before starting analysis, this data was loaded and cleaned and made ready for the analysis. Both univariate and multivariate analysis was done on all the factors/attributes in the given dataset, to see their influence in determining the salary of an individual. Visualizations were developed to give a clear perception of how the data was related and how much influence it has on determining the salary of a profile with different factors.

A few of the factors that had the strongest influence on salary were then used to develop marketing profiles to predict the income of individual based on the different factors. The marketing team wants to group these factors to tailor their marketing efforts to cater to individual customers.

The next goal is to select the model to train and test the data for the selected factors and predict the salary based on them. Accuracy of the model needs to be good before handing over to the client. Logistic regression algorithm was chosen to train and evaluate the model while keeping the focus on $50,000 as the key number for salary.

The business objective of this project is to provide many companies like the UVW company's marketing team with this salary prediction application. The UVW college marketing team wants to use this application to bolster enrollments by showing salary as a motivating criterion and thereby achieve their enrollment targets.

Assumptions: There were a few technical and business assumptions that were made in this project. Assumptions were made about conversion of original data, influence of weights, demographic characteristics as follows :

1. There are 32561 rows, mix of continuous and discrete data and 14 columns in the given dataset with 6 duplicate or conflicting instances.
2. This data needed to be scaled as needed, to do the analysis.
3. People with similar demographic characteristics had similar weight.
4. Salary in the dataset is assumed to be categorical as it is in the form of either greater than $50,000 or less than or equal to $50,000.
5. For the analysis, the categorical salary was imputed with binary weights of 0 and 1 for less than or equal to 50,000 and greater than $50,000 respectively.
6. After analysis it was assumed that age, education, marital status, occupation, and sex were some of the factors that had strong influence on determining the salary earned by an individual.
7. The train test split was assumed to be 1/3 of the data.
8. The information given in the adult.names file was accurate and can be used to get more information about the attributes and dataset.
9. The UVW marketing team would be able to reach its target of enrollment using the developed application.
10. It was assumed that age was continuous.
11. The workclass consisted of Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked categories.

12. Fnlwgt was continuous.
13. The education categories were Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
14. The education-num is a number assigned to each education category and is continuous.
15. The marital-status categories consisted of Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
16. The different occupations in the data set were 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.
17. The relationship, yet another categorical attribute has categories, Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
18. The race attribute consisted of White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
19. The sex attribute was limited to Female and Male.
20. The capital-gain is continuous.
21. The capital-loss is also continuous.
22. Hours-per-week is also continuous.
23. The native-country attribute consisted of the following countries: 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, Holland-Netherlands.
24. Education and education-num are redundant and one can be ignored.
25. Fnlwgt has no significant impact on salary and can be ignored.

User stories: The user stories that I prioritized for this project are the following :

1. The staff member in UVW marketing team would like to know what percentage of individuals earned above $50,000 and how many individuals earned below $50,000.
2. UVW Marketing wants to know if age was a detrimental factor in predicting salary of an individual.
3. Marketing team wants to know how education influenced the salary earning capability.
4. As the head of the UVW marketing team, I want to know a good combination of factors to earn a salary greater than $50,000.
5. Marketing team would like to know how much salary is influenced by the sex of the individual.
6. Marketing wants to know if the marital status of an individual effects the salary and if it would be a good factor to consider for the model.

Visualizations: To provide answers to the user's stories I prioritized, I had to analyze the data and come up with various visualizations to make the interpretation of the analysis easier.

1. *User Story #1:* The staff member in UVW marketing team would like to know what percentage of individuals earned above $50,000 and how many individuals earned below $50,000.
   - Refer to **US1.png** in Appendix for the visualization
   - To show the percentage of individuals earning above and below $50,000, the given data set was divided into two categories. This visualization shows the percentage in each category along with the labels to make it easily understandable.
   - Since, it was categorical data, a pie chart was used to show the percentages of each category. A darker color red was chosen to represent the bigger portion of the salary category, which is less than or

equal to $50K in this case. A relatively lighter color green was chosen to represent the category of salary greater than $50K. The title and labels give the needed extra information to understand which category each percentage belongs to.

- This data set has 75.1% earning less than 50K and 24.9% earning more than 50K. From this we can conclude that the data set is unbalanced.

2. *User Story #2:* UVW Marketing wants to know if age was a detrimental factor in predicting salary of an individual.
   - Refer to **US2.png** in the Appendix for the visualization.
   - This visualization gives us an insight on whether salary is varying with age and if it is a detrimental factor in predicting the salary of an individual.
   - For this visualization, a bar plot was chosen to show the relationship between salary and the age of people. Highest salaries earned mostly were in the age ranges 40-50 years, peaking at 50. Some outliers can also be seen as salary greater than 50K at 79 and 83 years
   - From the graph, we can conclude that age is an important factor to be considered when predicting the salary as it varies tremendously with age.

3. *User Story #3:* Marketing team wants to know how education influenced the salary earning capability.
   - Refer to **US3.png** in the Appendix for the visualization.
   - This visualization helps us understand how each category of education influences the salary earned by the individual in that category.
   - The choice of plot was again a bar plot to show counts of individuals in that category. In the legend 0 represents people earning less than 50K and 1 represents people earning more than 50K.
   - We can clearly see that education is influencing the salary earned by an individual. We see those individuals with a Bachelor's degree constituted the majority earning more than 50K in the dataset. They were followed by individuals who were High School graduates and some college degree students. UVW college Marketing team can use this information to bolster enrollment into the bachelor's program offered by their college.

4. *User Story #4:* As the head of the UVW marketing team, I want to know a good combination of factors to earn a salary greater than $50,000.
   - Refer to **US4.png** in the Appendix for the visualization
   - This visualization shows us the correlation between different attributes/factors given in the dataset.
   - This visualization is a correlation matrix generated using Pearson's correlation coefficient. It measures the linear association between two variables. It has a value between -1 and 1. -1 indicates negative correlation. 0 indicates no correlation and 1 indicates a positive correlation. This matrix gives us a summary of overall correlations in the dataset. The darker green the cell is more its correlation to salary.
   - Along with that we can also get insights of how the individual attributes are correlated to each other based on the numbers. Any number between 0 and 1 shows a positive correlation and closer to 1 indicates most correlation.
   - Looking at the visualization, we can see that age and hours worked per week have a good correlation of 0.1. Education also has a good correlation with hours of work per week of 0.15.

- From this, we can conclude that age, education, and hours of work per week are a good combination of factors to predict the salary of an individual and see if it is greater than 50K or not.

5. *User Story #5:* Marketing team would like to know how much salary is influenced by the sex of the individual.
   - Refer to **US5.png** in the Appendix for the visualization.
   - This visualization shows us the influence of sex of an individual on the salary earned. In this dataset only Male and Female sexes were considered.
   - The choice of plot for this visualization was a stacked bar chart. Here again the 0 in the legend represents salary less than 50K and 1 represents salary more than 50K. The bar chart again is a good visualization to use when counts of two categorical variables are taken into account.
   - In the plot we can see that the number of males earning more than 50K is far too greater than the number of females earning more than 50K.
   - This can be useful to predict salary for an individual based on sex. The conclusion that we can draw from this is that if you are a female, the chances of earning more that 50K are relatively less than if you are a male.

6. *User Story #6:* Marketing wants to know if the marital status of an individual effects the salary and if it would be a good factor to consider for the model.
   - Refer to **US6.png** in the Appendix for the visualization.
   - The visualization shows how different categories in the marital status which include never-married, married civ spouse, divorced, married spouse absent, separated , married af spouse and widowed influence the salary of the individual.
   - For the visualization a bar plot was used to count the number of individuals falling into each category with salary greater than 50k shown in orange and salary less than 50K shown in green.
   - 0 in the legend represents salary greater than 50K and 1 represents salary less than 50K.
   - From the visualization we can see that married civ spouse category has the highest chance of earning more than 50K followed by never married and divorced. This can be used as a factor in the model for predicting salary .

Questions: A lot of questions aroused during the project progression. The first one was when trying to clean the dataset and make it ready for analysis. Some of the unknown values in the dataset were converted to "?". What was the best way to impute them? Mean and median were some of the choices. But, as some of the data was categorical , those choices did not provide a meaningful dataset. Dropping those rows didn't affect the dataset and thus was chosen as the best solution.

Next, the choice of analysis and visualizations to get the answers for the user stories I prioritized was another question that needed to be answered. Salary being a categorical in the given data set, the best solution was to replace the greater than $50,000 with a '1' and less than $50,000 with a '0' helped in the analysis making it binary.

Finally, it was what to choose for training the dataset? The choice of going for logistic regression to train the dataset was made because the dataset was dichotomous or binary. The logistic regression machine learning algorithm, was used to predict the salary of an individual, based on his/her profile.

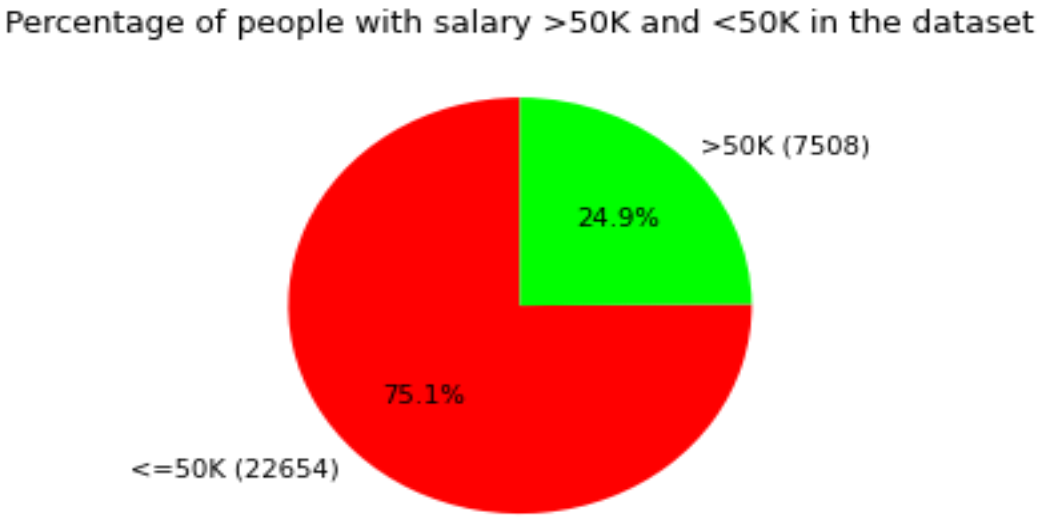
Not doing: In this phase of the project implementation, more importance was given to the analysis and visualization of the data. Only logistic regression was used to train and test the data for creating a salary prediction model.

More research and analysis need to be done to train the data set using different machine learning algorithms to compare their efficiency and accuracy with respect to predictions and pick the best one and use it train and test the data to create a better prediction model. Not done creating a working application with good UI to predict the salary. Coming up with a solution for that has been moved to the next phase.

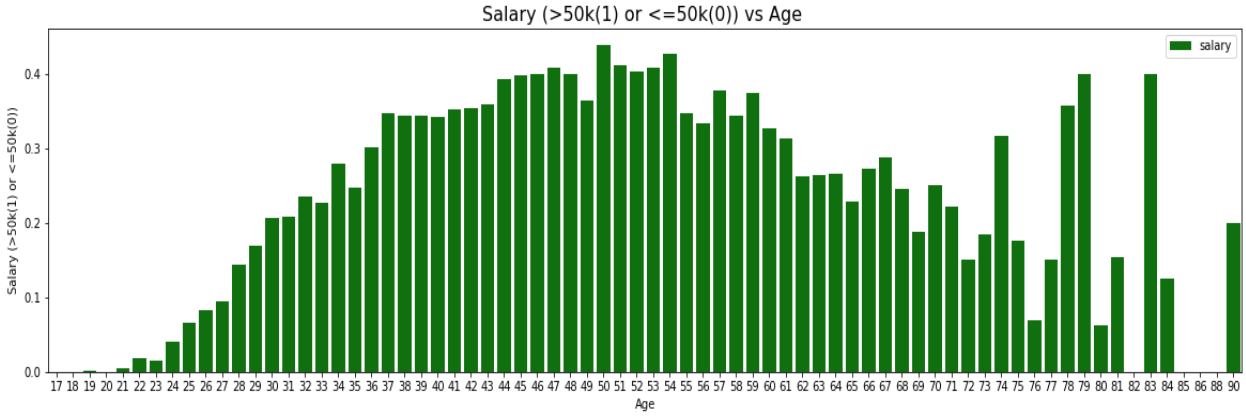
Appendix: For the implementation of this project, Python, Pandas, Numpy were used for analysis and calculations. Seaborn, Matplotlib were used for the visualizations.

Below are the visualizations referenced in the user stories:

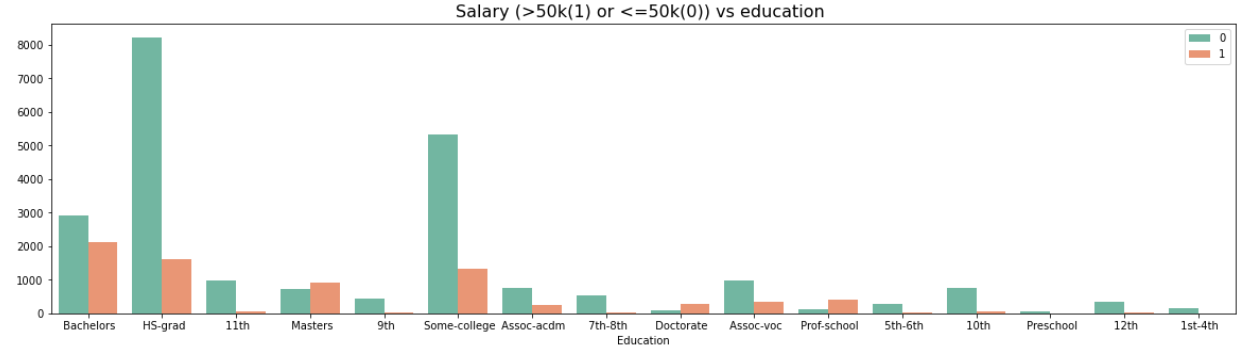
US1.png- The staff member in UVW marketing team would like to know what percentage of individuals earned above $50,000 and how many individuals earned below $50,000.



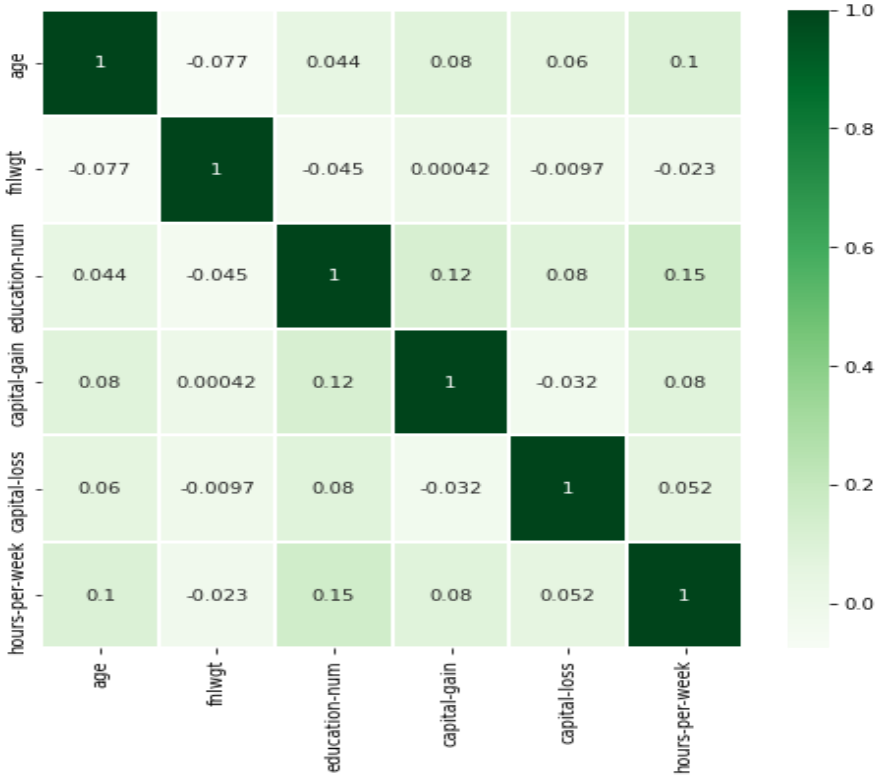
US2.png- UVW Marketing wants to know if age was a detrimental factor in predicting salary of an individual.



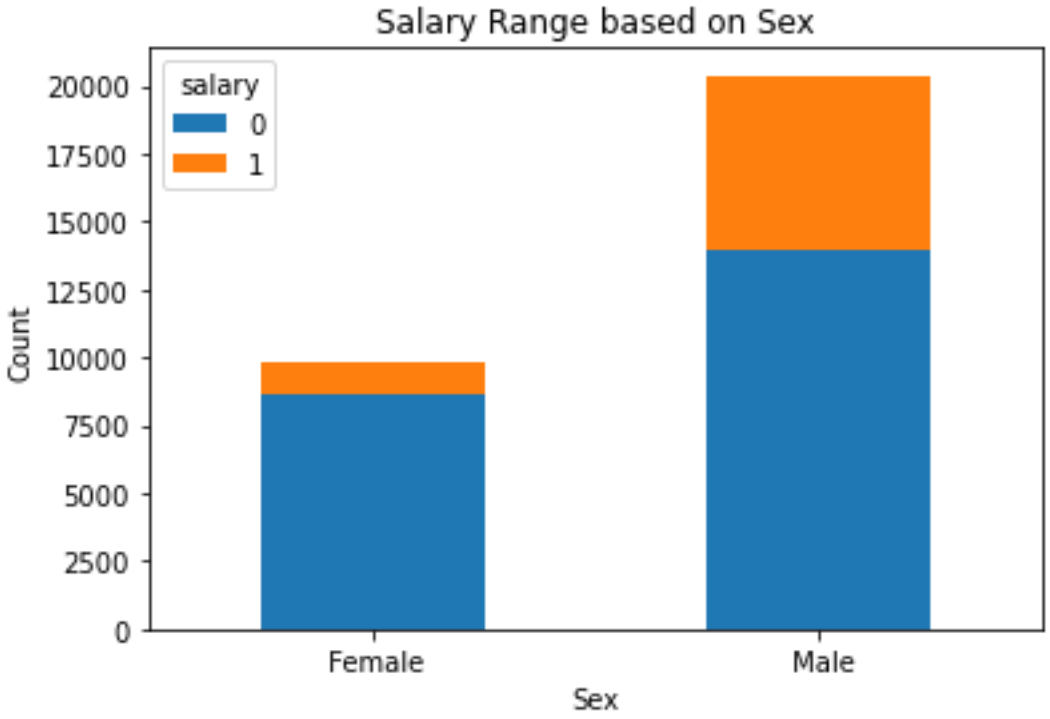
US3.png- Marketing team wants to know how education influenced the salary earning capability.



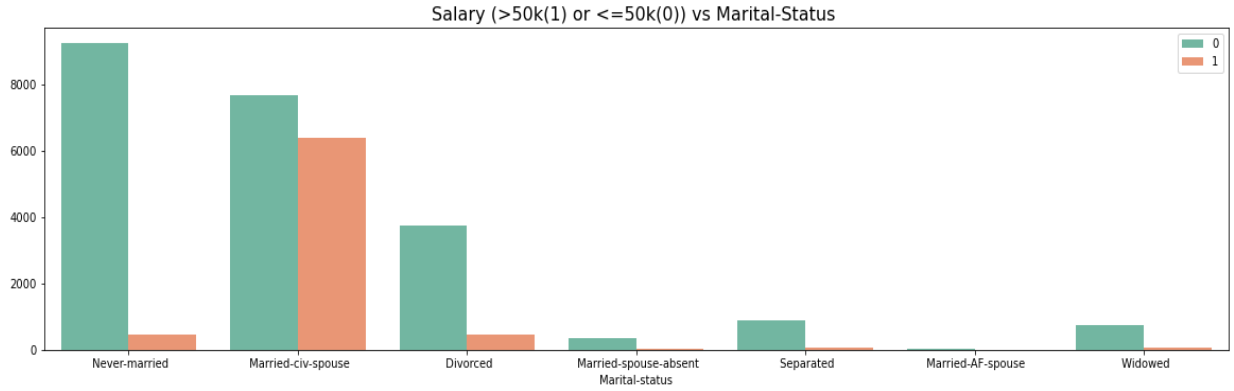
US4.png -As the head of the UVW marketing team, I want to know a good combination of factors to earn a salary greater than $50,000.



US5.png - Marketing team would like to know how much salary is influenced by the sex of the individual.



US6.png - Marketing wants to know if the marital status of an individual effects the salary and if it would be a good factor to consider for the model.



The pdf link to the code and the visualizations:

vizprojectmain.pdf

The code for the project is below:

Goals and a business objective

To develop an application to predict the income of an individual, based on the factors used in developing marketing profiles on people.Data provided by the United States Census Bureau was used. The focus was kept on $50,000 as the key number for salary.

Assumptions

Assumptions were made about conversion of original data, final weights, similar demographic characteristics should have similar weights. It was assumed that the salary was binary and categorical. Salary greater than 50,000wasreplacedwith′1′andsalarylessthan50,000wasreplacedwith′1′andsalarylessthan**50,000 was replaced with '0' for analysis.**

Import libraries, Load data and clean data

In [140]:

```
# Import dependencies
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn import metrics
import os
import sys
%matplotlib inline
```

In [2]:

```
# Load data
adult_df=pd.read_csv('adult.data',header=None)
adult_df.columns=['age','workclass','fnlwgt','education','education-
num','marital-status','occupation', 'relationship','race', 'sex', 'capital-
gain', 'capital-loss','hours-per-week','native-country','salary']
adult_df.head()
```

Out[2]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

In [3]:

```
# Clean the dataset by dropping rows with ?
clean_df = adult_df.replace(' ?', np.NaN).dropna()
clean_df.shape
```

Out[3]:

```
(30162, 15)
```

In [4]:

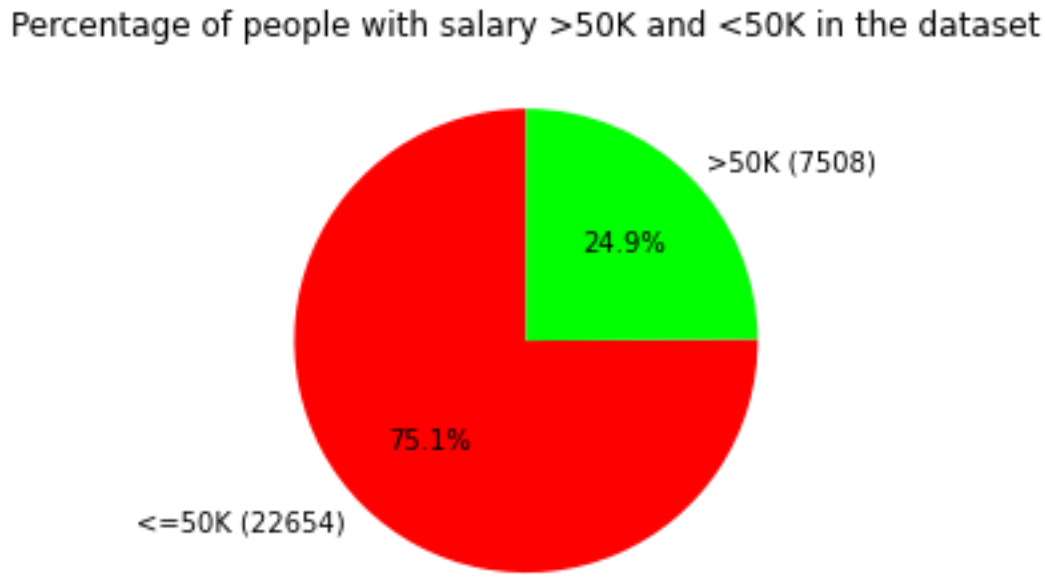
```
# Find the number of people with salary >50K and <50K in the dataset
salary_count = clean_df.groupby('salary').size()
salary_count
```

Out[4]:

```
salary
 <=50K    22654
 >50K      7508
dtype: int64
```

In [5]:

```
# Pie chart of the number of people with salary >50K and <50K
plt.pie(salary_count, labels=['<=50K (22654)', '>50K (7508)'],
autopct='%1.1f%%', startangle=90, colors=['#ff0000', '#00ff00'])
plt.title('Percentage of people with salary >50K and <50K in the dataset')
plt.show()
```



Analyze the data to see which attributes or factors contribute to a higher individual salary, and help create various marketing profiles using this analysis.

In [6]:

```
# Assume that the salary is binary and categorical variable and convert it to
numeric by replacing >50K with 1 and <=50K with 0
log_df = clean_df.copy()
log_df.loc[clean_df.salary.str.strip() == "<=50K", 'salary'] = 0
log_df.loc[clean_df.salary.str.strip() == ">50K", 'salary'] = 1
log_df.tail()
```

Out[6]:

	age	work class	fnl wgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
32556	27	Private	257302	Assoc-acdm	12	Married-civ-	Tech-support	Wife	White	Female	0	0	38	United-States	0

	age	work class	fnl wgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
						spouse									
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	1
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	0
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	0
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	1

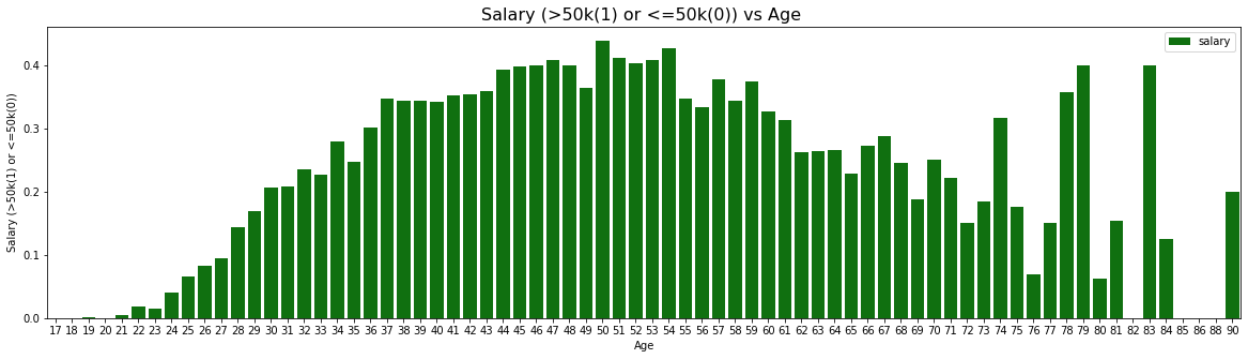
Salary vs Age analysis

In [109]:

```
# Plot showing relation between salary and age
fig = plt.figure(figsize=(20,5))
ax_age = sns.barplot(x='age', y='salary', data=log_df, color='green',
label='salary', ci=None)
ax_age.set_title("Salary (>50k(1) or <=50k(0)) vs Age ", loc='center',
fontsize=16)
ax_age.set_xlabel("Age")
ax_age.set_ylabel("Salary (>50k(1) or <=50k(0))")
ax_age.legend(loc="upper right")
```

Out[109]:

```
<matplotlib.legend.Legend at 0x2925c0926a0>
```



Conclusion: There is a significant correlation or influence of age on salary earned.

Salary vs fnlwgt

In [112]:

```
# Relation between salary and fnlwgt below and above 50K
below50_df = log_df[log_df.salary == 0]
above50_df = log_df[log_df.salary == 1]
sumweightbelow = below50_df['fnlwgt'].sum()
sumweightabove = above50_df['fnlwgt'].sum()
print("Percentage weight below 50k: ", (sumweightbelow/(sumweightbelow +
sumweightabove))*100)
print("Percentage weight above 50k: ", (sumweightabove/(sumweightbelow +
sumweightabove))*100)
```

```
Percentage weight below 50k:  75.32335207448135
Percentage weight above 50k:  24.676647925518644
```

Conclusion: As it is showing the distribution of the dataset, we can ignore this as an influencing factor.

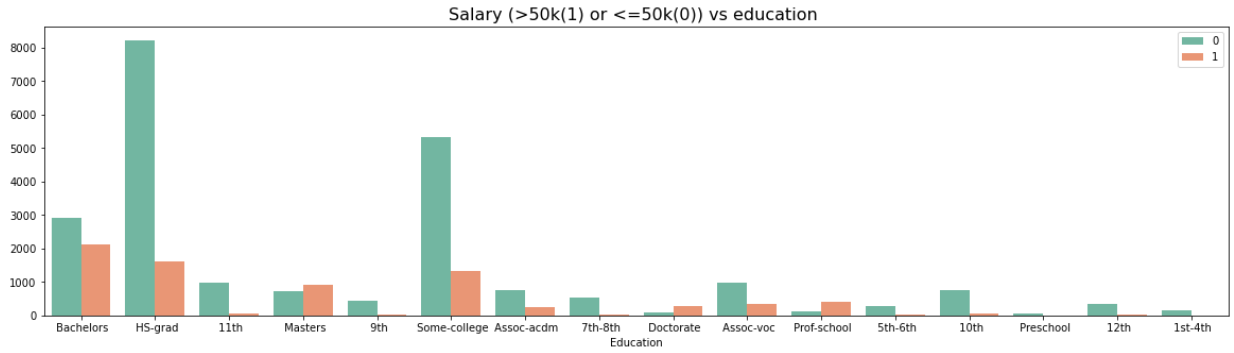
Salary vs Education

In [9]:

```
# Plot showing relation between salary and education
fig = plt.figure(figsize=(20,5))
ax_edu = sns.countplot(data=log_df, x='education', hue='salary',
palette='Set2')
ax_edu.set_title("Salary (>50k(1) or <=50k(0)) vs education ", loc='center',
fontsize=16)
ax_edu.set_xlabel("Education")
ax_edu.set_ylabel(" ")
ax_edu.legend(loc="upper right")
```

Out[9]:

```
<matplotlib.legend.Legend at 0x29242a6f100>
```



Conclusion: We see that education does influence the earning power and hence can be considered a factor.

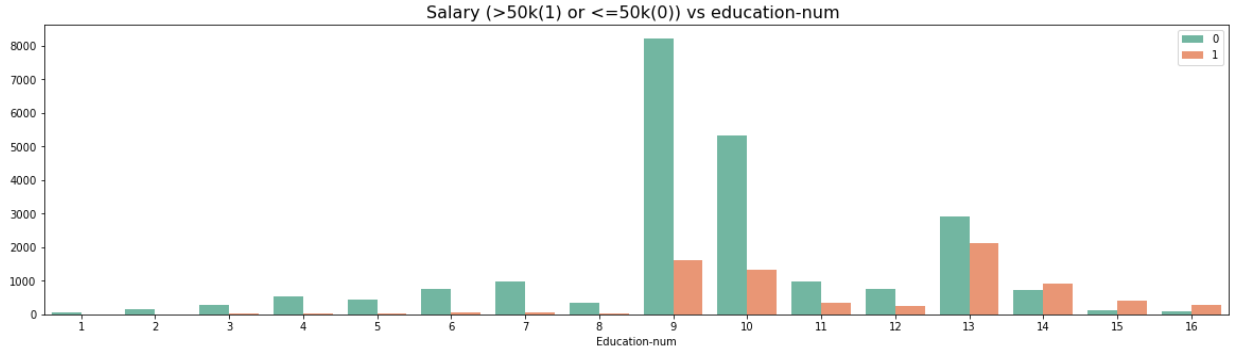
Salary vs Education-num

In [10]:

```
# Plot showing correlation between salary and education-num
fig = plt.figure(figsize=(20,5))
ax_edu_num = sns.countplot(data=log_df, x='education-num', hue='salary',
palette='Set2')
ax_edu_num.set_title("Salary (>50k(1) or <=50k(0)) vs education-num ",
loc='center', fontsize=16)
ax_edu_num.set_xlabel("Education-num")
ax_edu_num.set_ylabel(" ")
ax_edu_num.legend(loc="upper right")
```

Out[10]:

```
<matplotlib.legend.Legend at 0x29243c2b700>
```



Conclusion: This is same as education and can be considered for the model.

Salary vs Marital status

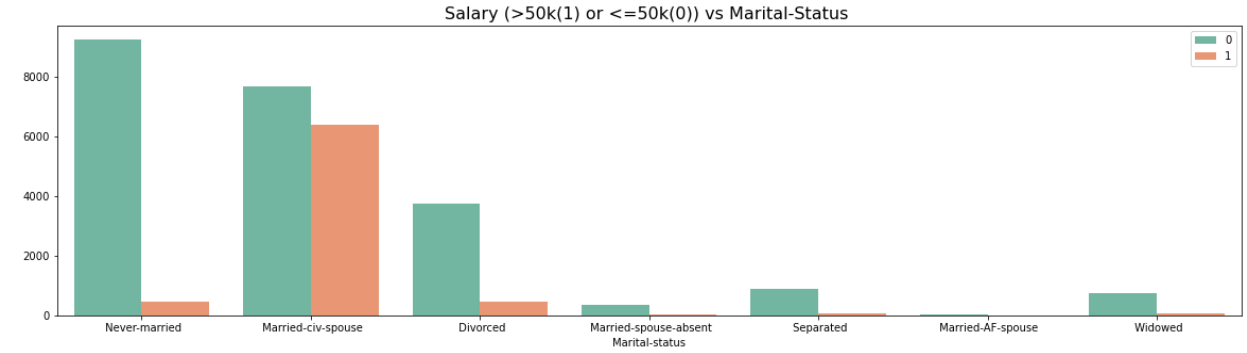
In [11]:

```
# Plot showing relation between salary and matrial-status matplotlib
fig = plt.figure(figsize=(20,5))
ax_mat = sns.countplot(data=log_df, x='marital-status', hue='salary',
palette='Set2')
ax_mat.set_title("Salary (>50k(1) or <=50k(0)) vs Marital-Status ",
loc='center', fontsize=16)
ax_mat.set_xlabel("Marital-status")
```

```
ax_mat.set_ylabel(" ")
ax_mat.legend(loc="upper right")
```

Out[11]:

```
<matplotlib.legend.Legend at 0x29242a6f160>
```



Conclusion: Again we see a strong influence on salary.

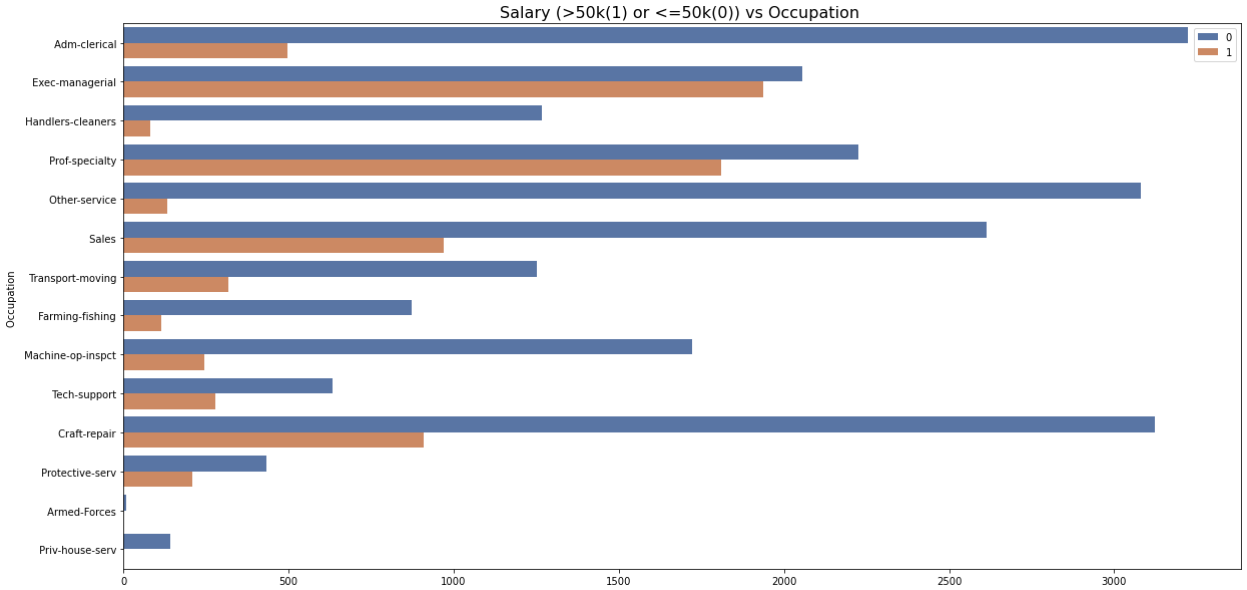
Salary vs Occupation

In [12]:

```
# Plot showing relation between salary and occupation
fig = plt.figure(figsize=(20,10))
ax_occ = sns.countplot(data=log_df, y='occupation', hue='salary',
palette='deep',)
ax_occ.set_title("Salary (>50k(1) or <=50k(0)) vs Occupation ", loc='center',
fontsize=16)
ax_occ.set_xlabel("")
ax_occ.set_ylabel("Occupation ")
ax_occ.legend(loc="upper right")
```

Out[12]:

```
<matplotlib.legend.Legend at 0x2924417b1c0>
```



Conclusion: We can conclude that occupations strongly influence salary earned.

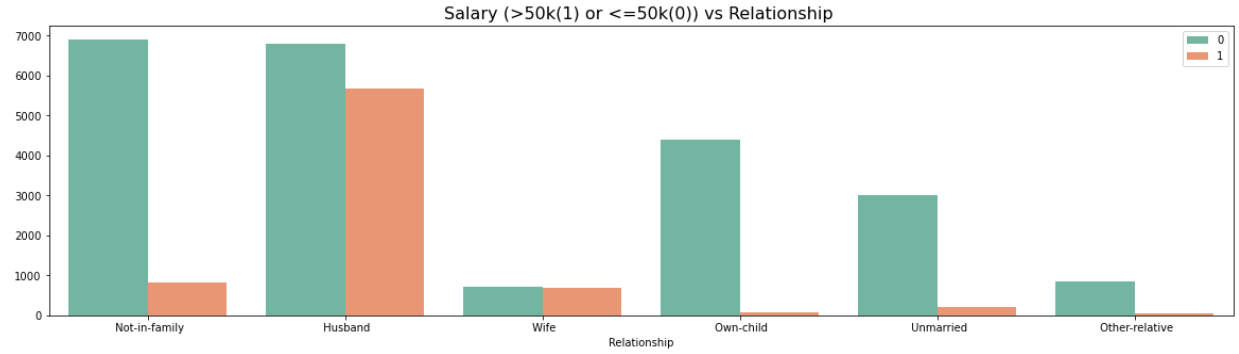
Salary vs Relationship

In [13]:

```
# Plot showing relation between salary and relationship
fig = plt.figure(figsize=(20,5))
ax_rel = sns.countplot(data=log_df, x='relationship', hue='salary',
palette='Set2')
ax_rel.set_title("Salary (>50k(1) or <=50k(0)) vs Relationship ",
loc='center', fontsize=16)
ax_rel.set_xlabel("Relationship")
ax_rel.set_ylabel(" ")
ax_rel.legend(loc="upper right")
```

Out[13]:

```
<matplotlib.legend.Legend at 0x2924411dbe0>
```



Conclusion: We can clearly see that relationships have impact on salaries.

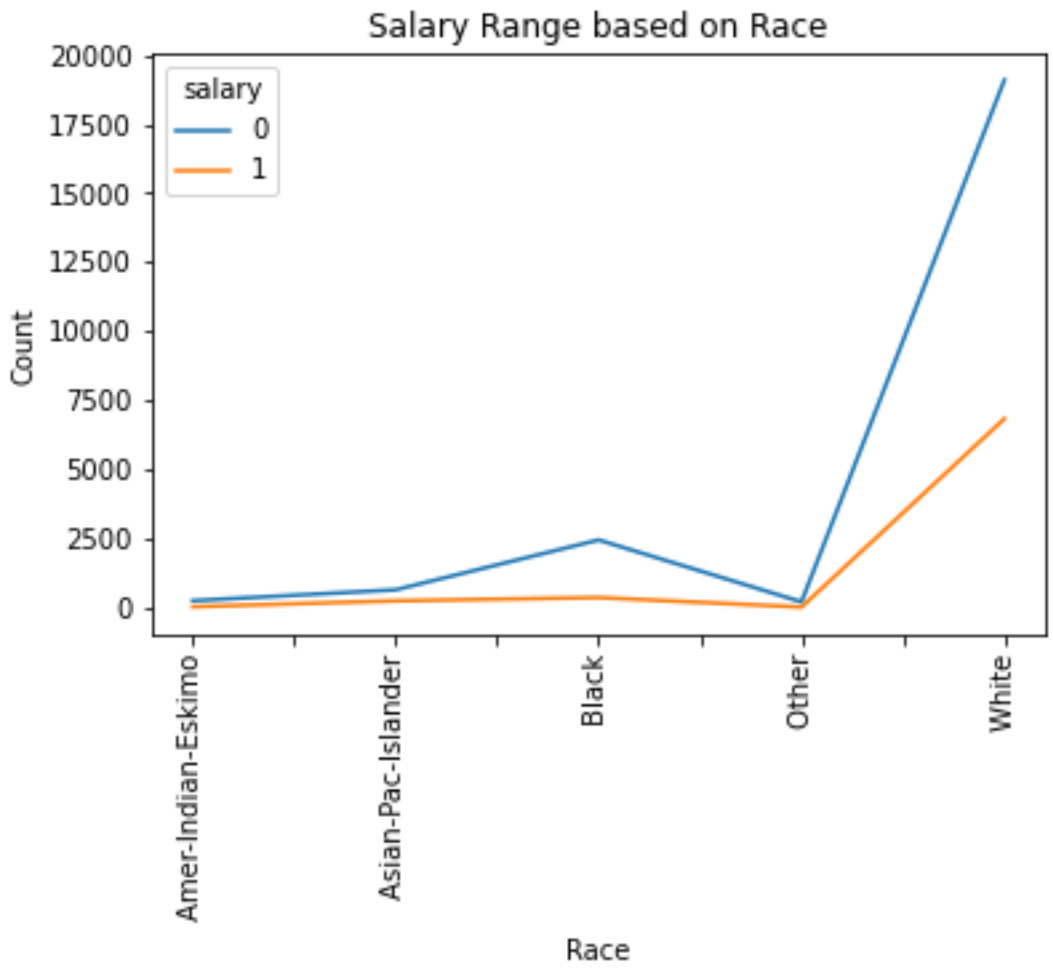
Salary vs Race

In [26]:

```
# Plot showing relation between salary and race
#Race Analysis: This appears to have greater impact on the salary
pd.crosstab(log_df['race'], log_df['salary']).plot(kind='line', rot=90)
plt.xlabel('Race')
plt.ylabel('Count')
plt.title('Salary Range based on Race')
```

Out[26]:

```
Text(0.5, 1.0, 'Salary Range based on Race')
```



Conclusion: We can see that it has an impact only if you are white. We can ignore this.

Salary vs Sex

In [15]:

```
# Plot showing correlation between salary and sex
#Compute the percentage of male and female earning below 50k.
male_df = log_df[log_df['sex'] == ' Male']
male_lessthan50k_df = log_df[(log_df['sex'] == ' Male') & (log_df['salary']
== 0)]
female_df = log_df[log_df['sex'] == ' Female']
female_lessthan50k_df = log_df[(log_df['sex'] == ' Female') &
(log_df['salary'] == 0)]
print("Male & Female Percentage with salary <=50k: ",
round((male_lessthan50k_df.shape[0]/male_df.shape[0])*100,2), ", ",
round((female_lessthan50k_df.shape[0]/female_df.shape[0])*100,2))
male_df = log_df[log_df['sex'] == ' Male']
male_morethan50k_df = log_df[(log_df['sex'] == ' Male') & (log_df['salary']
== 1)]
female_df = log_df[log_df['sex'] == ' Female']
```

```
female_morethan50k_df = log_df[(log_df['sex'] == ' Female') &
(log_df['salary'] == 1)]
print("Male & Female Percentage with salary >50k: ",
round((male_morethan50k_df.shape[0]/male_df.shape[0])*100,2), ", ",
round((female_morethan50k_df.shape[0]/female_df.shape[0])*100,2))
```

```
Male & Female Percentage with salary <=50k:  68.62 ,  88.63
Male & Female Percentage with salary >50k:  31.38 ,  11.37
```

In [23]:

```
#Sex Analysis: It can be seen that sex has an important role in determining
the salary, especially in males.
pd.crosstab(log_df['sex'], log_df['salary']).plot(kind='bar', rot=0,
stacked=True)
plt.xlabel('Sex')
plt.ylabel('Count')
plt.title('Salary Range based on Sex')
```

Out[23]:

```
Text(0.5, 1.0, 'Salary Range based on Sex')
```

Conclusion: We can see that sex definitely defines your earning power.

Salary vs Native country

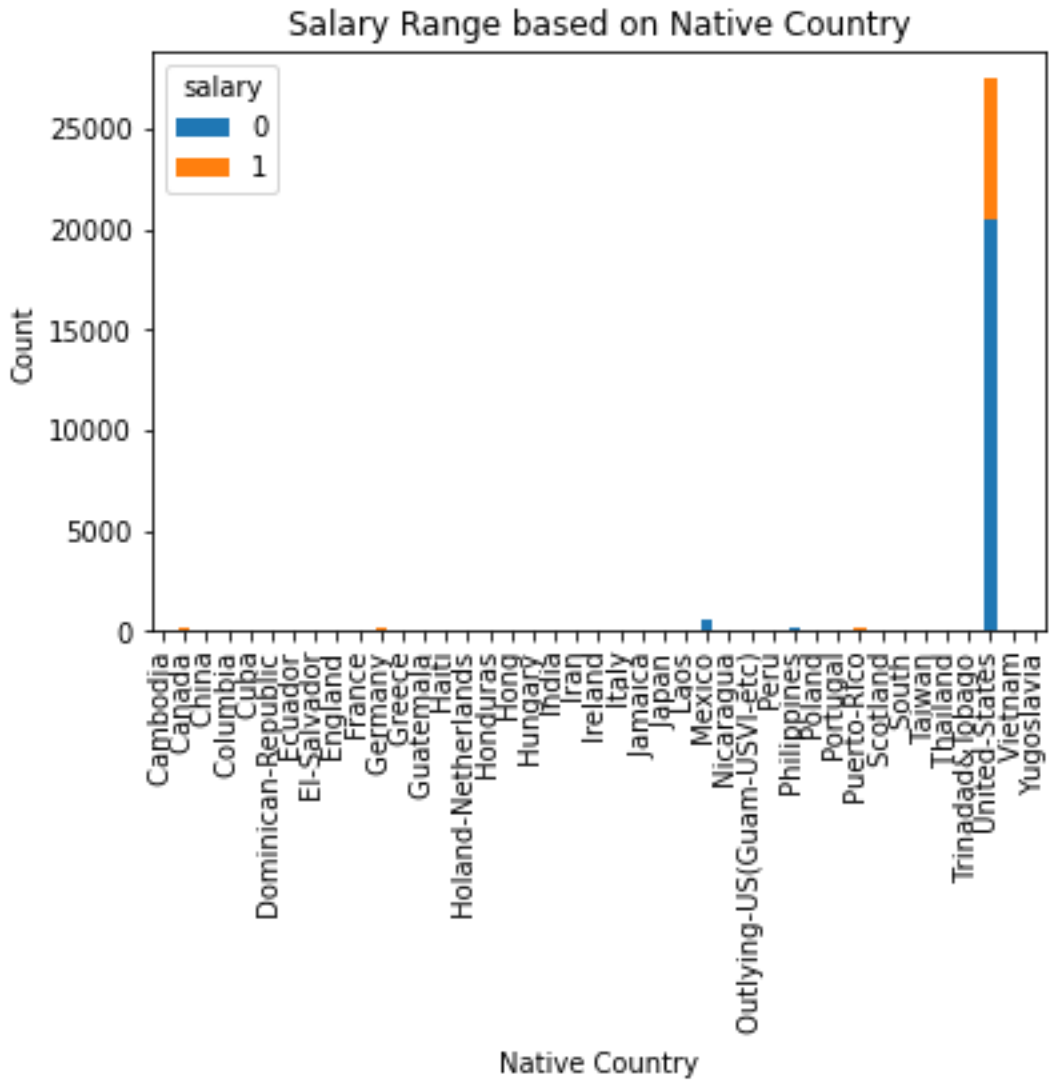
In [24]:

```
# Plot showing rrelation between salary and native country
#Native Country Analysis: This doesn't appear to have a clear info of the
impact on the salary.
pd.crosstab(log_df['native-country'], log_df['salary']).plot(kind='bar',
stacked=True)
```

```
plt.xlabel('Native Country')
plt.ylabel('Count')
plt.title('Salary Range based on Native Country')
```

Out[24]:

```
Text(0.5, 1.0, 'Salary Range based on Native Country')
```



Conclusion: The influence is limited to United states and can be ignored.

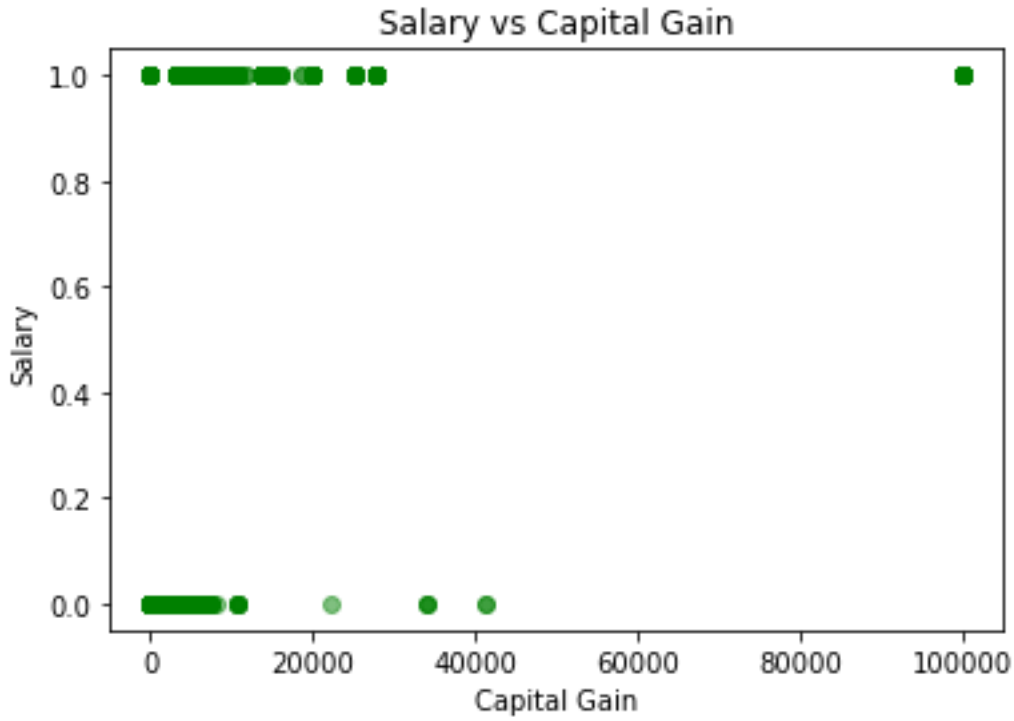
Salary vs Capital gain

In [115]:

```
# Plot showing correlation between salary and capital-gain
#Capital Gain Analysis: This doesn't appear to have a clear info of the
impact on the salary.
plt.scatter(log_df['capital-gain'], log_df['salary'], alpha=0.5, c='green')
plt.xlabel('Capital Gain')
plt.ylabel('Salary')
plt.title('Salary vs Capital Gain')
```

Out[115]:

```
Text(0.5, 1.0, 'Salary vs Capital Gain')
```



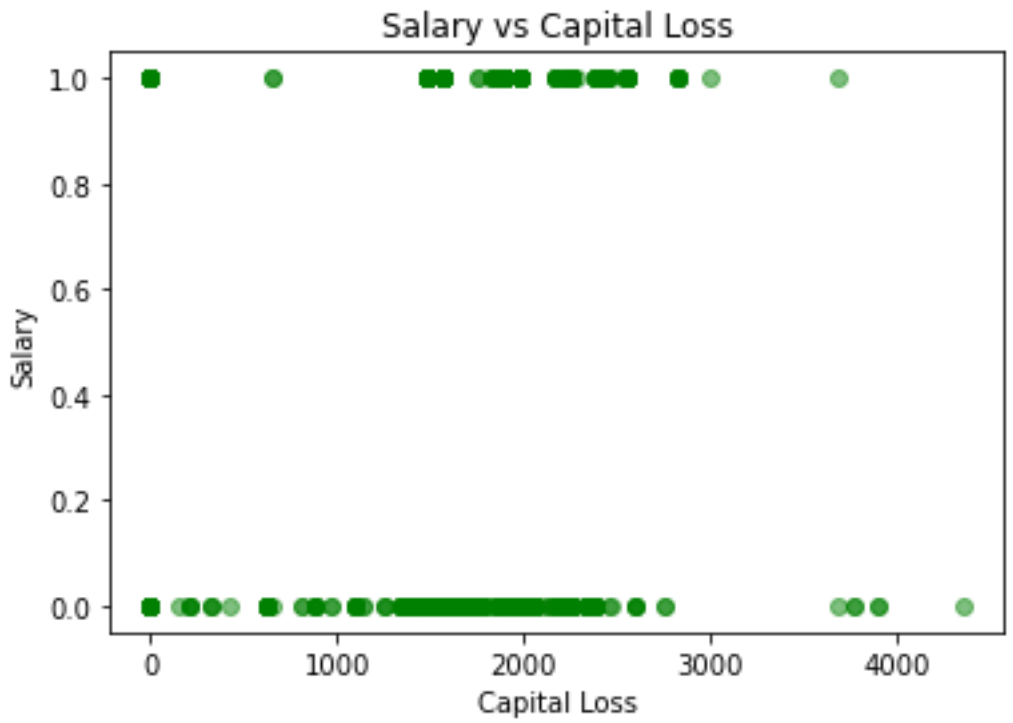
Salary vs Capital loss

In [33]:

```
# Plot showing correlation between salary and capital-loss
plt.scatter(log_df['capital-loss'], log_df['salary'], alpha=0.5, c='green')
plt.xlabel('Capital Loss')
plt.ylabel('Salary')
plt.title('Salary vs Capital Loss')
```

Out[33]:

```
Text(0.5, 1.0, 'Salary vs Capital Loss')
```



Conclusion:Both Capital gain and capital loss have a weak relationship with salary and can be ignored.

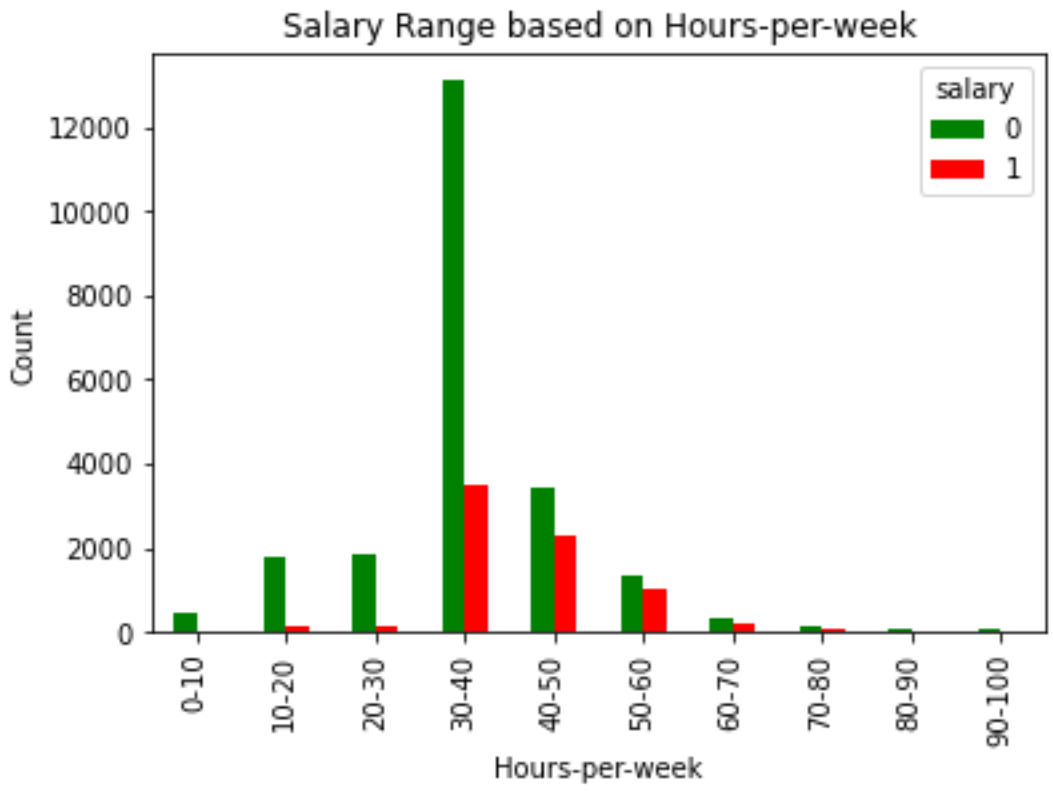
Salary vs Hours per week

In [77]:

```
# pie Plot showing correlation between salary and hours-per-week
#Hours-per-week Analysis: This doesn't appear to have a clear info of the
impact on the salary.
log_df['hours-per-week-grouped'] = pd.cut(log_df['hours-per-week'],
bins=[0,10,20,30,40,50,60,70,80,90,100], labels=['0-10','10-20','20-30','30-
40','40-50','50-60','60-70','70-80','80-90','90-100'])
pd.crosstab(log_df['hours-per-week-grouped'],
log_df['salary']).plot(kind='bar',color= ['green','red'])
plt.xlabel('Hours-per-week')
plt.ylabel('Count')
plt.title('Salary Range based on Hours-per-week')
```

Out[77]:

```
Text(0.5, 1.0, 'Salary Range based on Hours-per-week')
```



Conclusion: This can be considered a factor influencing salary.

Looking at the analysis so far, we can pick age, education, marital status, occupation, sex and hours per week as factors which influence the salary for our model.

Correlation between all the other factors.

In [17]:

```
# relation between salary and age, fnlwgt, education-num, capital-gain,
capital-loss, hours-per-week
fig = plt.figure(figsize=(8,8))
sns.heatmap(log_df.corr(), annot=True, cmap= 'Greens', linewidths=0.2)
```

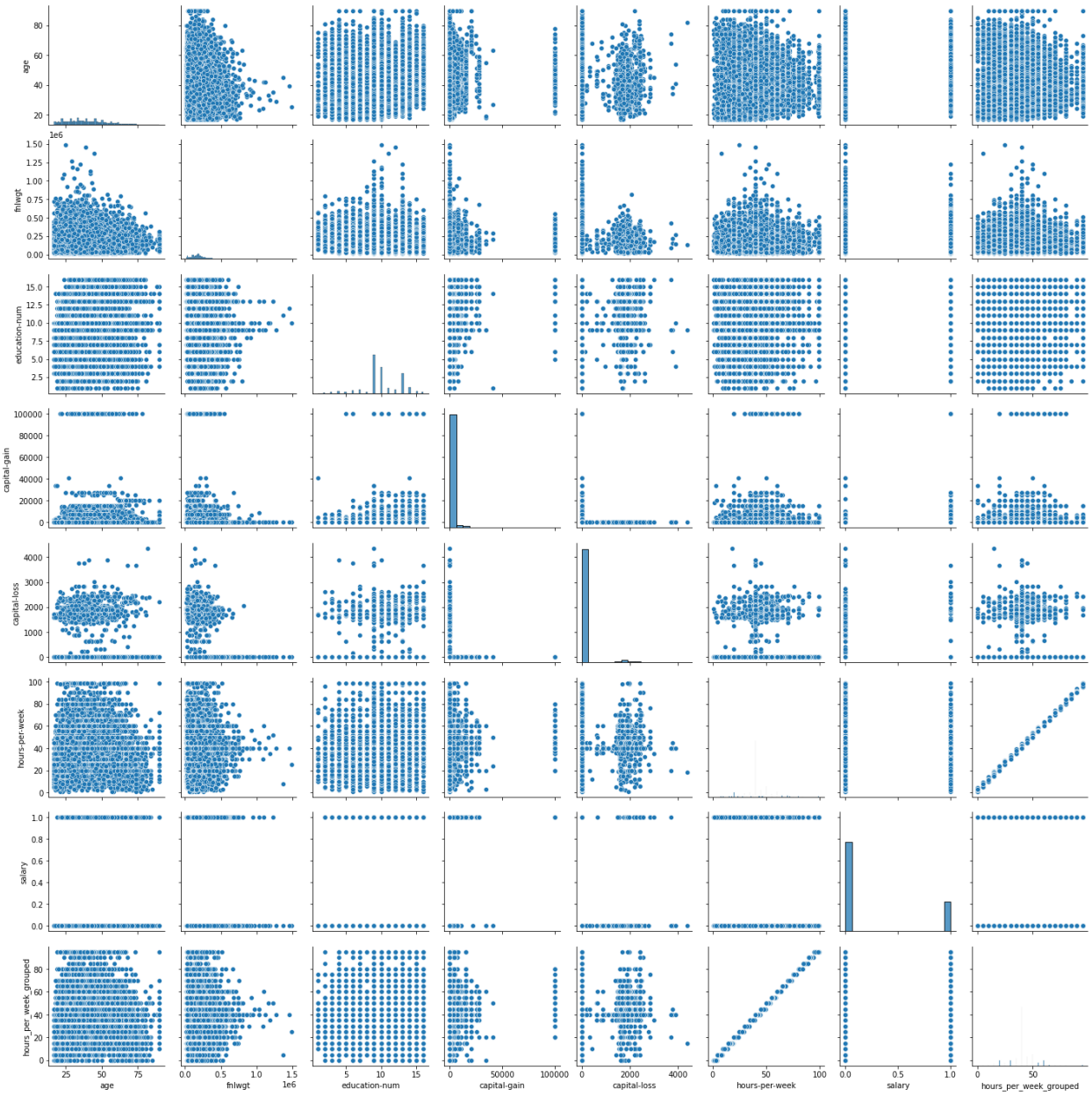
Out[17]:

```
<AxesSubplot:>
```

In [99]:

```
fig = plt.figure(figsize=(20,5))
ax_age = sns.pairplot(log_df)
```

```
<Figure size 1440x360 with 0 Axes>
```



Conclusion: From the above heat map and pairplot, we can conclude that there is a further strong correlation between age, education, hours per week and are good factors for the model.

Now that the influencing factors are determined to build marketing profiles, a logistic regression model can be built on the training dataset, and the salary category (<=50k and >50k) can be predicted. Model accuracy can also be determined utilizing the confusion matrix.

Based on the above analysis, we can consider the following attributes/factors of the dataset are used for building the model.

In [118]:

```
cols = ['age', 'education', 'marital-status', 'occupation', 'relationship',
'sex']
```

In [144]:

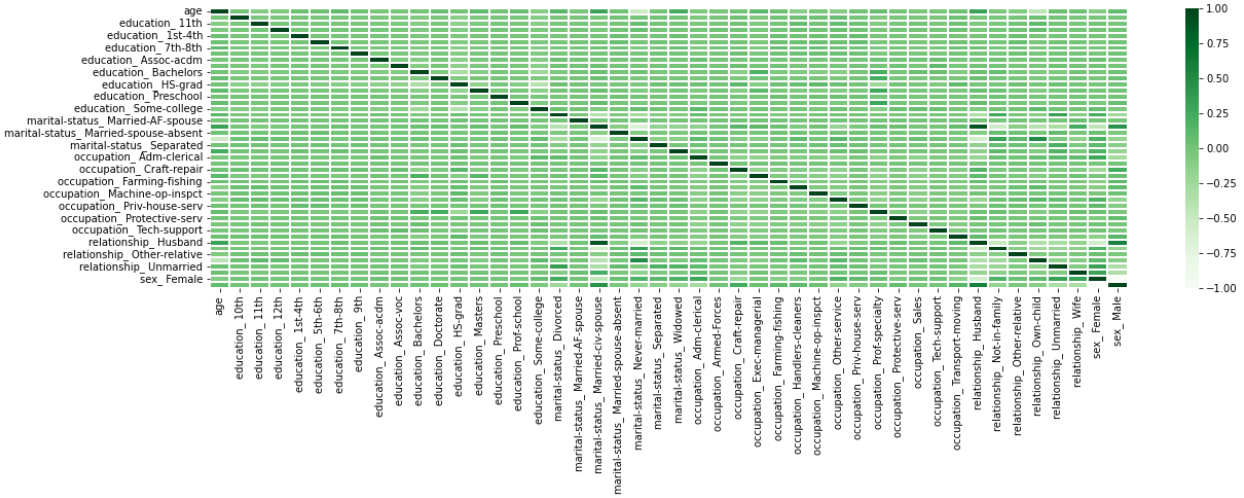
```
X = log_df[cols]
X_dummies = pd.get_dummies(X)
y = log_df['salary']

# Split the dataset into train and test datasets, considering 70:30 ratio.
X_train, X_test, y_train, y_test = train_test_split(X_dummies, y,
test_size=0.3, random_state=0)

# Scale the train and test datasets
min_max_scaler = MinMaxScaler()
X_train_minmax = min_max_scaler.fit_transform(X_train)
X_test_minmax = min_max_scaler.fit_transform(X_test)

# Create a Logistic Regression model
logreg = LogisticRegression(max_iter=200)
y1 = y_train.astype(int)
logreg.fit(X_train_minmax, y1)

# Predicting the Test set results
y_pred=logreg.predict(X_test_minmax)
y_pred
```

Out[144]:

```
array([0, 0, 0, ..., 0, 0, 0])
```

In [131]:

```
X_dummies = pd.get_dummies(X)
# heatmap
fig = plt.figure(figsize=(20,5))
ax_age = sns.heatmap(X_dummies.corr(), cmap= 'Greens', linewidths=0.2)
```



Salary predicted based on the selected factors

In [135]:

```
pred_df = X_test.copy()
pred_df['salary'] = pd.Series(y_test, index=pred_df.index)
pred_df['predicted salary'] = pd.Series(y_pred, index=pred_df.index)
```

```
pred_df['predicted salary'] = pred_df['predicted salary'].map({0: '<=50k', 1:
'>50k'})
pred_df[['salary','predicted salary']]
```

Out[135]:

	salary	predicted salary
2135	0	<=50k
15639	0	<=50k
29059	0	<=50k
27523	0	<=50k
9280	0	<=50k
...	...	...
16826	0	<=50k
25246	0	<=50k
18980	1	<=50k
953	0	<=50k
30925	0	<=50k

9049 rows × 2 columns

Checking for the accuracy of the model

In [145]:

```
y1_test = y_test.astype(int)
cnf_matrix = metrics.confusion_matrix(y1_test, y_pred)
cnf_matrix
```

Out[145]:

```
array([[6194,  570],
       [1050, 1235]], dtype=int64)
```

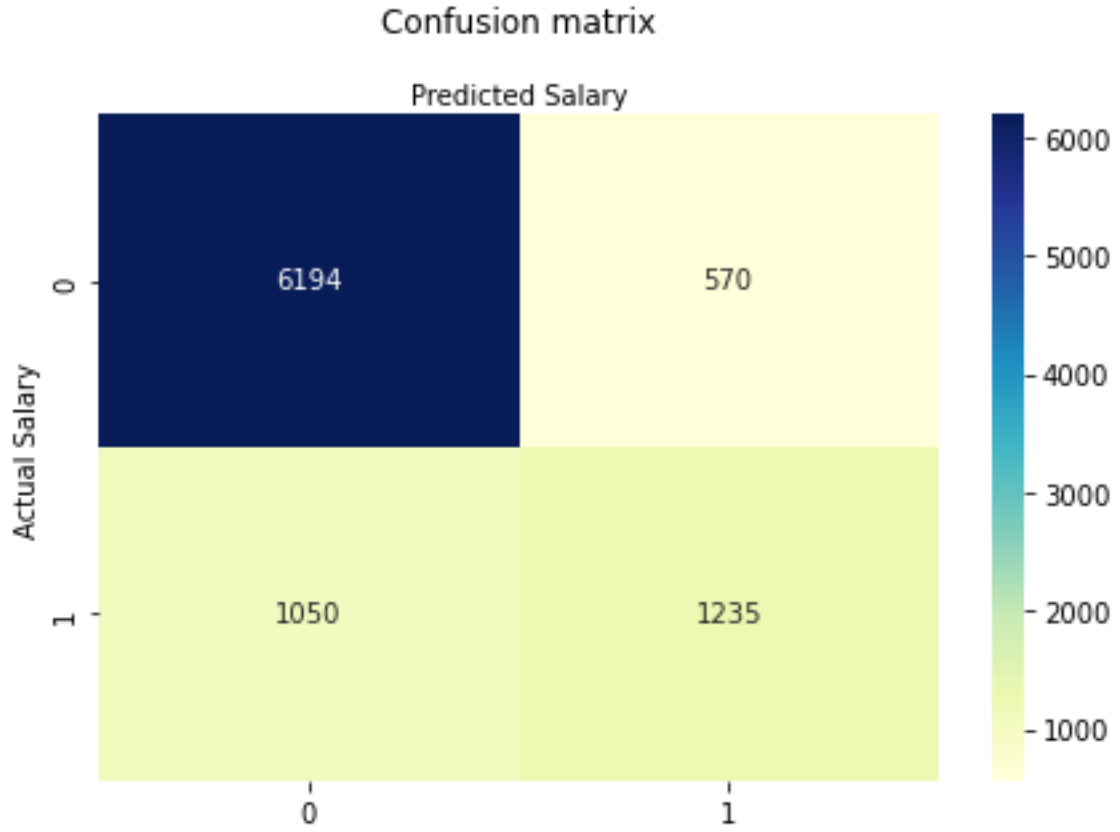
In [146]:

```
class_names=[0,1] # name  of classes
```

```
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual Salary')
plt.xlabel('Predicted Salary')
```

Out[146]:

```
Text(0.5, 257.44, 'Predicted Salary')
```



In [150]:

```
print("Accuracy:", round(metrics.accuracy_score(y1_test, y_pred)*100, 2))
print("Precision:",round(metrics.precision_score(y1_test, y_pred)*100, 2))
```

```
Accuracy: 82.1
Precision: 68.42
```

Conclusion: We can see that based on the factors selected the model predicts the salary with an 82.10% accuracy and 68.42% precision.