**CS277 Data Mining**

Anbang Xu(35086995)

**Part1:**

1. How many unique values there are for each variable?

[('age', '73'), ('capital-gain', '119'), ('capital-loss', '92'), ('education', '16'), ('education-num', '16'), ('fnlwgt', '21648'), ('hours-per-week', '94'), ('marital-status', '7'), ('native-country', '42'), ('occupation', '15'), ('race', '5'), ('relationship', '6'), ('sex', '2'), ('workclass', '9')]

1. A dict that indicates which variables are categorical and which are numeric.

[('age', ‘numeric’), ('capital-gain', ‘numeric’), ('capital-loss', ‘numeric’), ('education', 'categorical'), ('education-num', ‘numeric’), ('fnlwgt', ‘numeric’), ('hours-per-week', ‘numeric’), ('marital-status', 'categorical'), ('native-country', 'categorical'), ('occupation', 'categorical'), ('race', 'categorical'), ('relationship', 'categorical'), ('sex', 'categorical'), ('workclass', 'categorical')]

**Part2:**

1. Variable Definitions

#age: the age of the individual as reported by the individual at the time of the 1990 census, in integer units of years.

#workclass: a section of society dependent on physical labor, especially when compensated with an hourly wage

#fnlwgt: final weight(no clear definition)

#education: the highest level of education

#education-num: the amount of people receiving education

#marital-status: indicates whether the person is married

#occupation: a job or profession

#relationship: the way in which two or more people or things are connected

#race: a classification system used to categorize humans into large and distinct populations or groups

#sex: specialized into male and female varieties

#capital-gain: a profit that results from a disposition of a capital asset, such as stock, bond or real estate

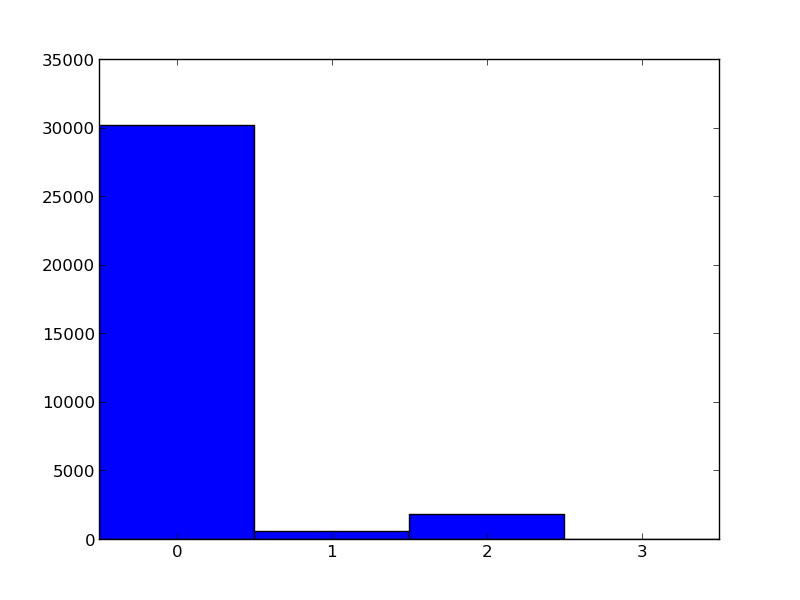
#capital-loss: the difference between a lower selling price and a higher purchase price, resulting in a financial loss for the seller

#hours-per-week: the number of hours to work per week

1. Missing Data
   1. Calculate and list what percentage of rows have missing values for each variable.

Three missing data in ‘’workclass’’, “occupation”, “native-country” and the percentages of missing data are 5.6%, 5.7% and 1.8% for the three variables respectively.

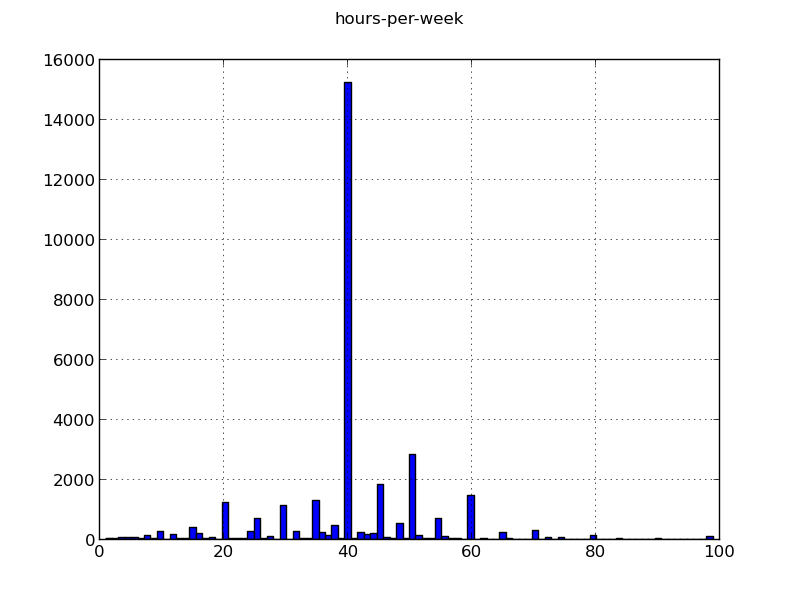
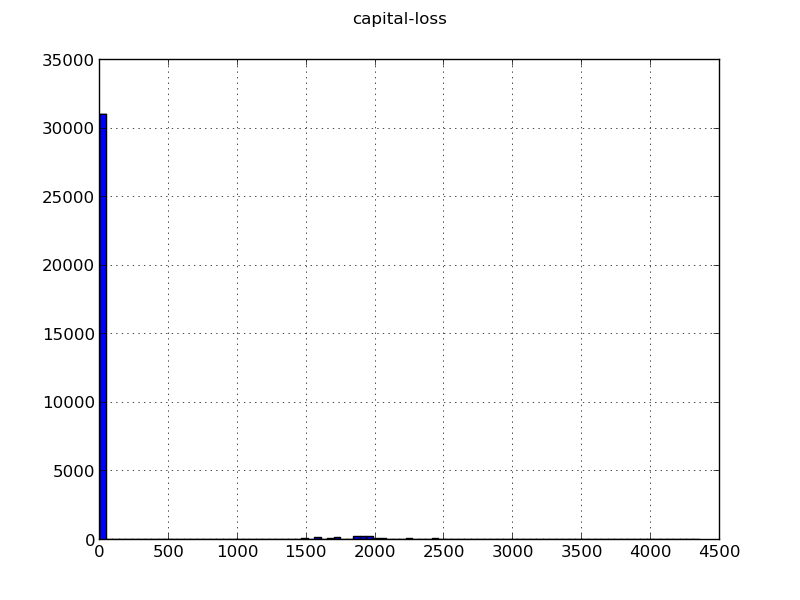
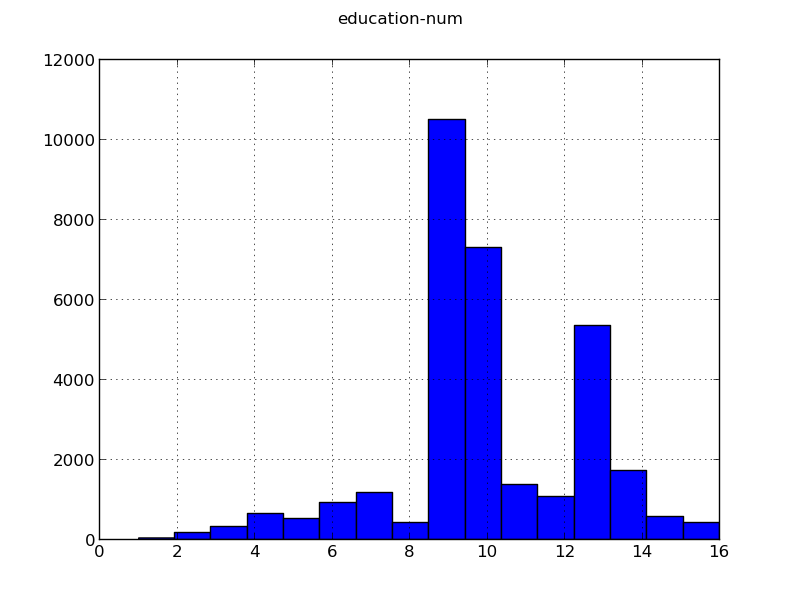
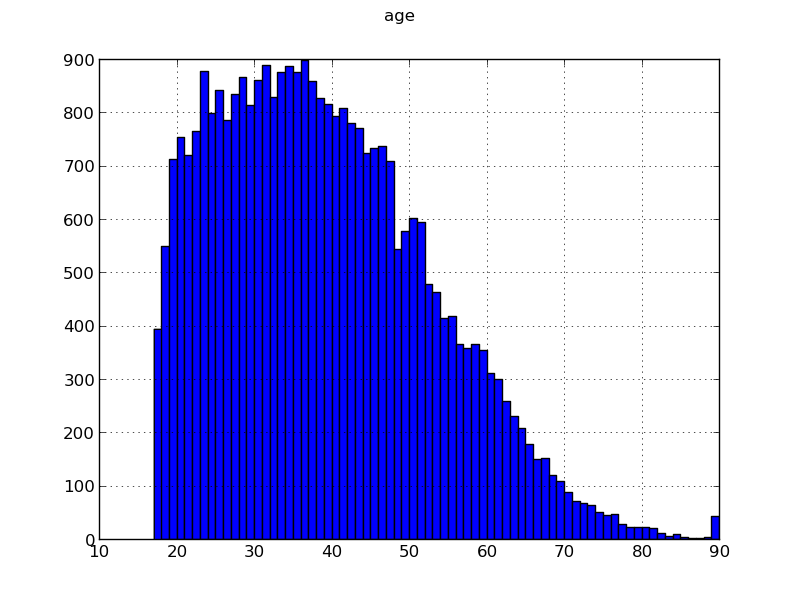
* 1. Generate a histogram indicating how many rows have 0, 1, 2, 3, .... missing values.



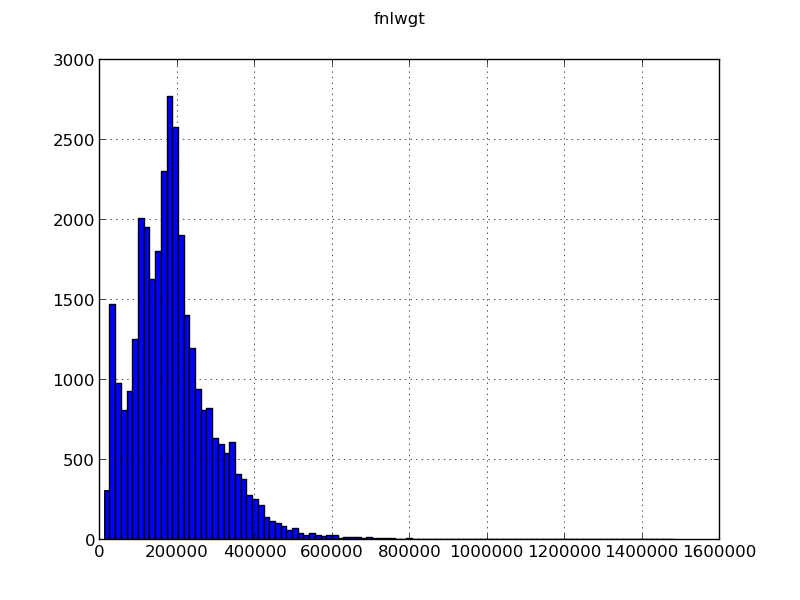
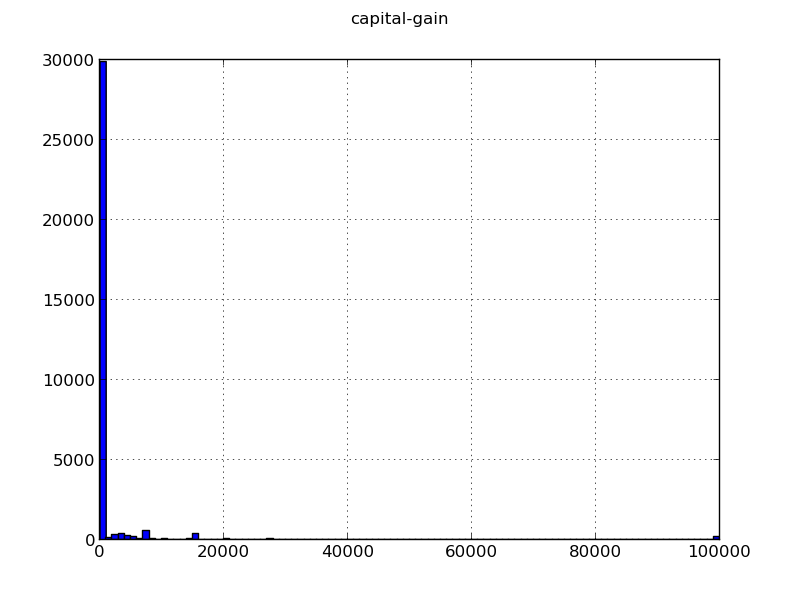
1. Numeric Variables
   1. List the number of unique values for each variable

[('age', 73), ('capital-gain', 119), ('capital-loss', 92), ('education-num', 16), ('fnlwgt', 21648), ('hours-per-week', 94)]

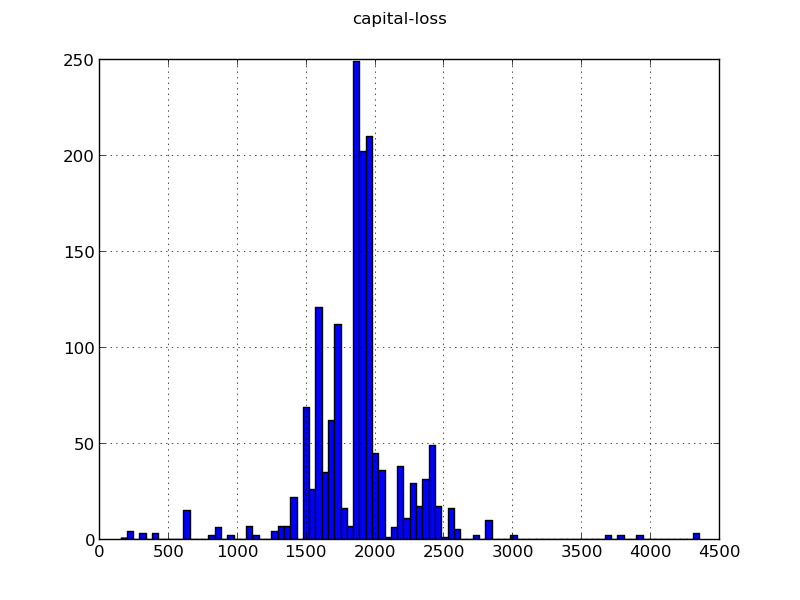
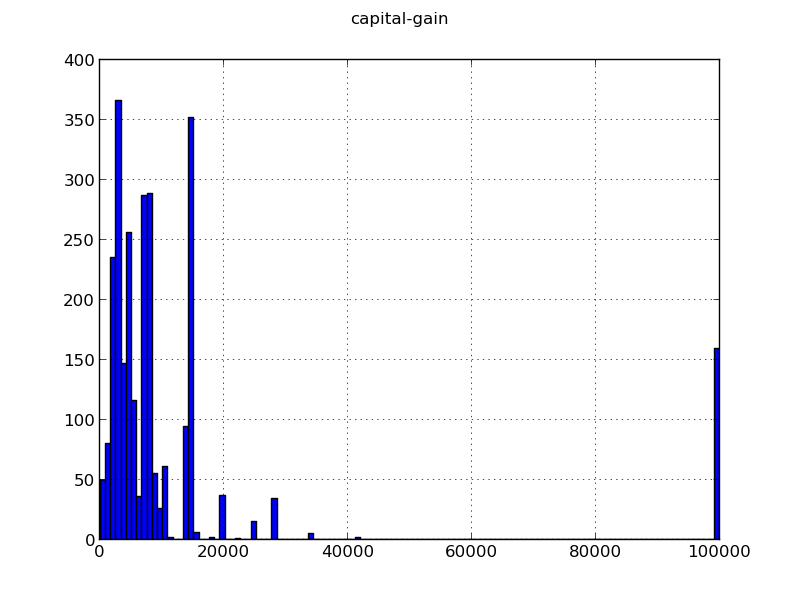
* 1. Generate a histogram for variables with less than 100 values



* 1. Generate a histogram using 100 bins for variables with 100 or more values



* 1. Just plot histograms of the non-zero values

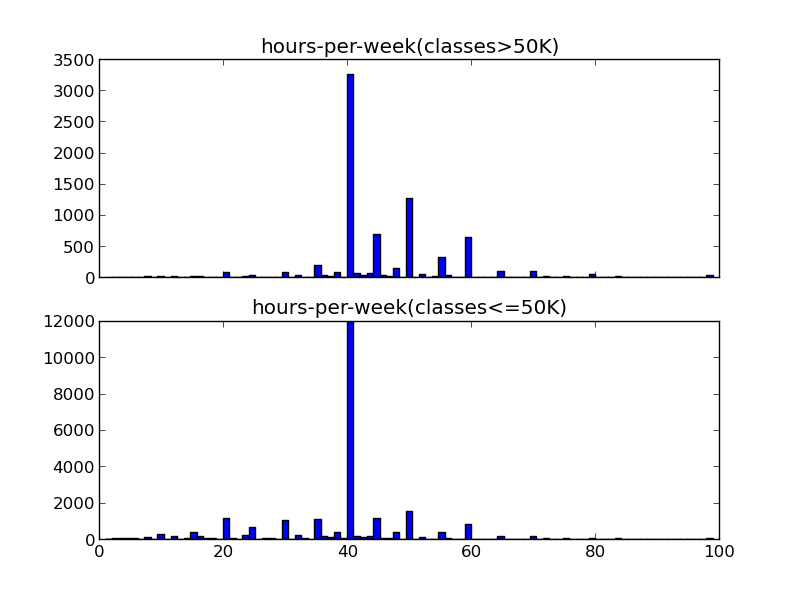
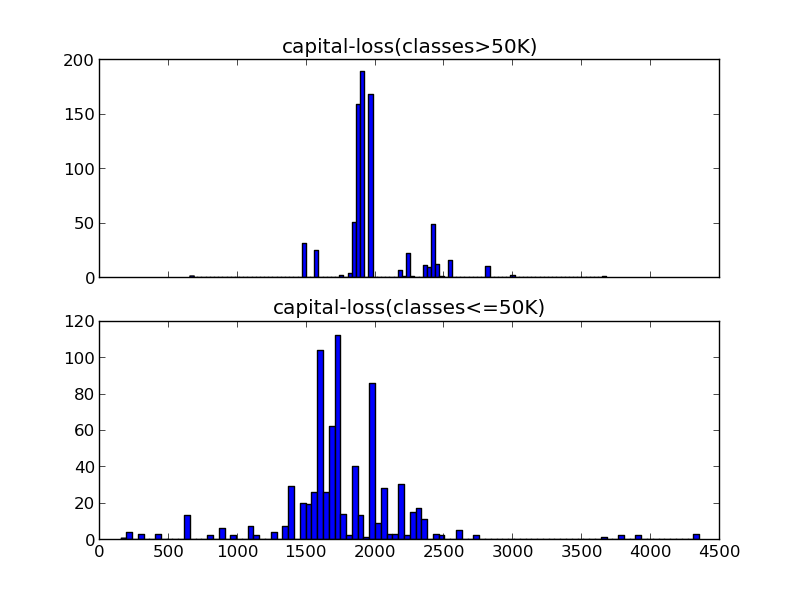
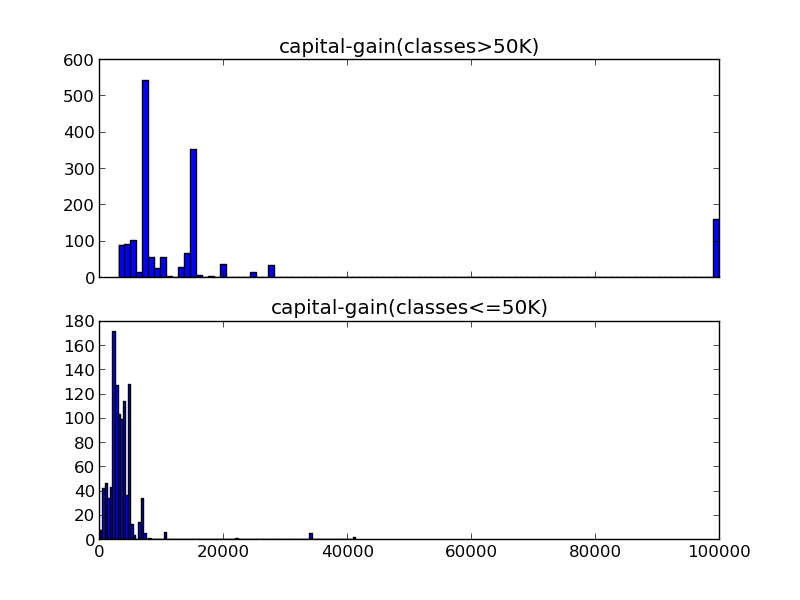
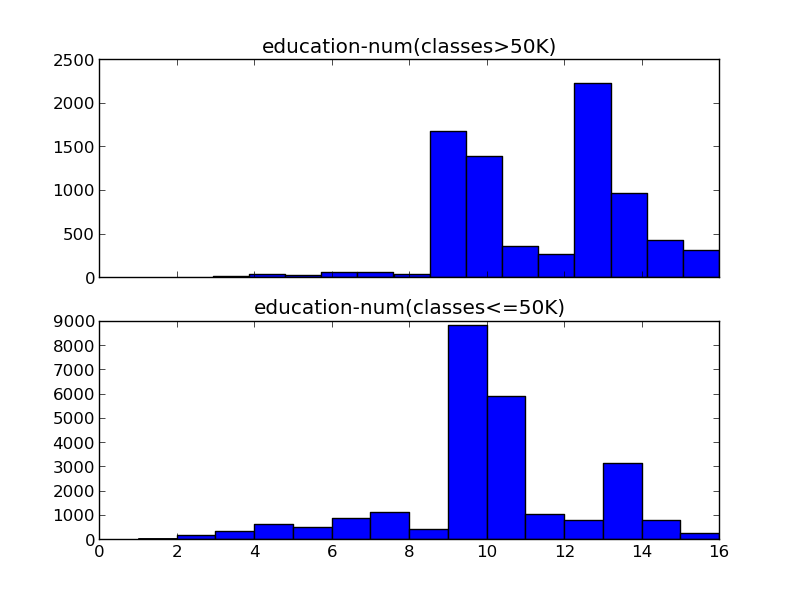
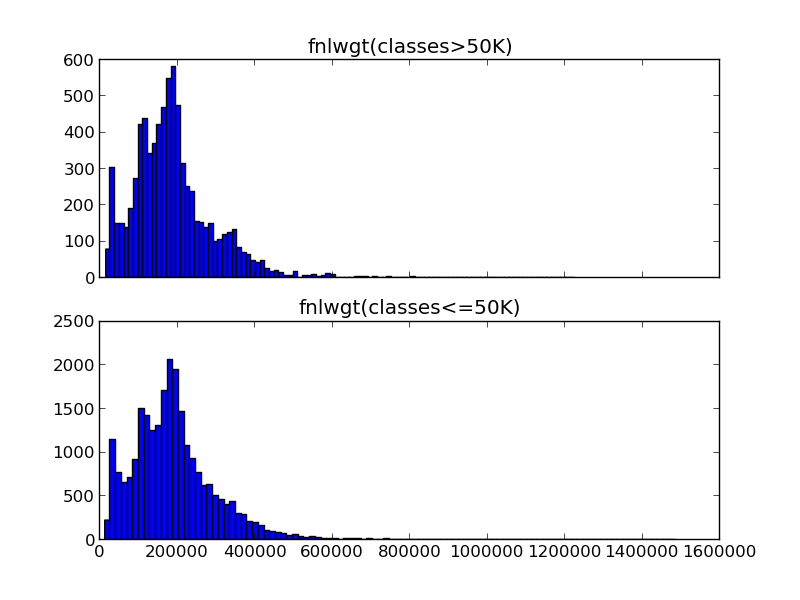
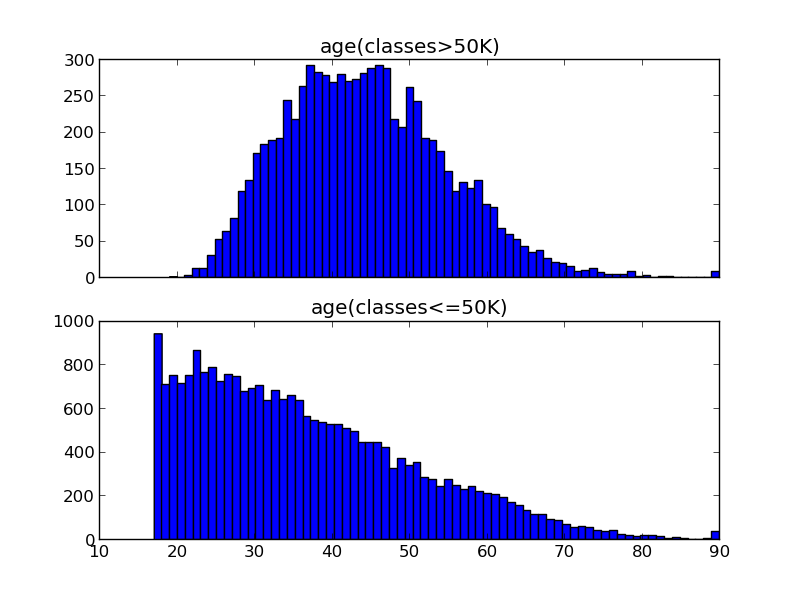


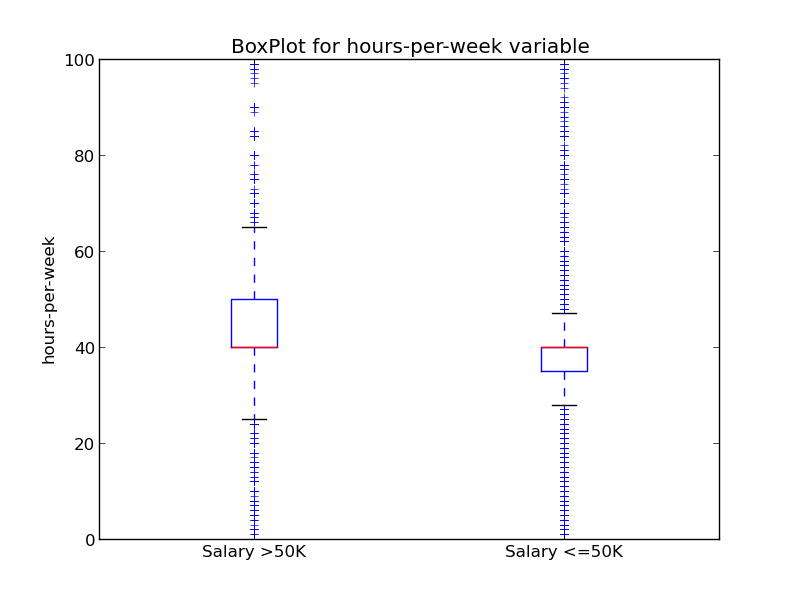
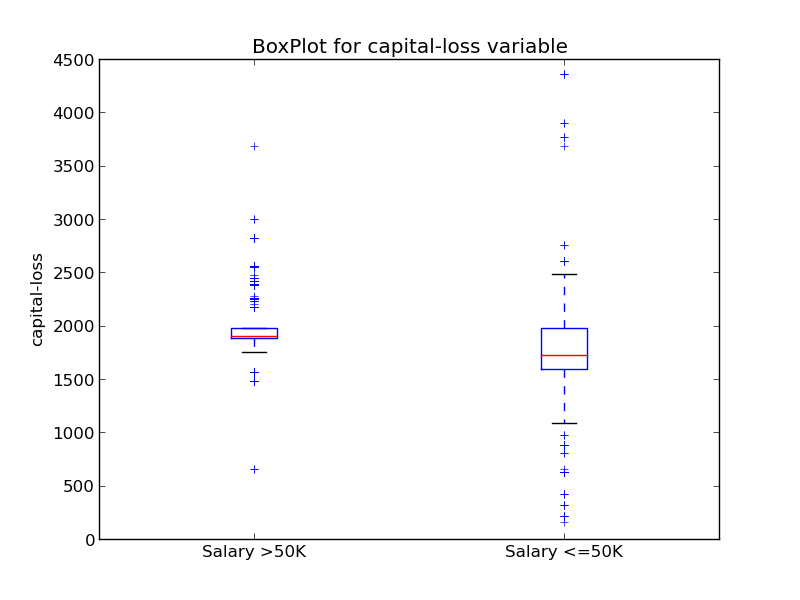
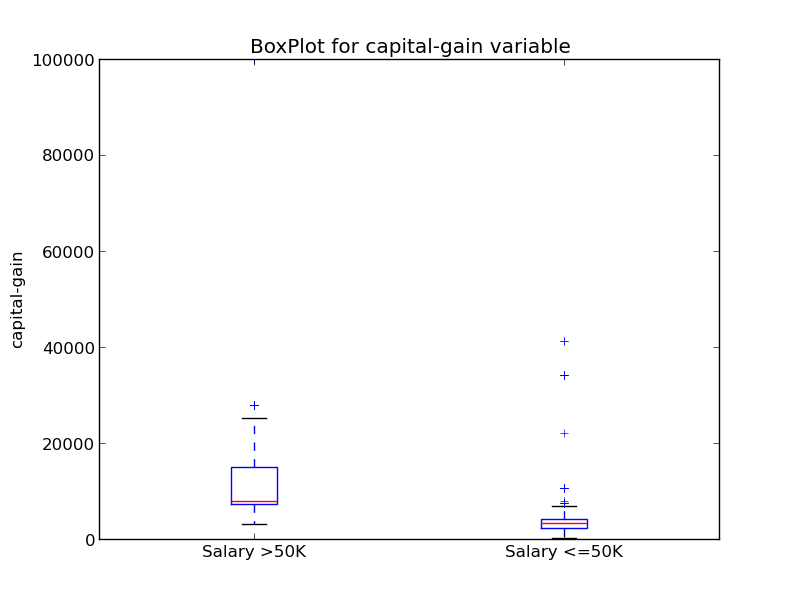
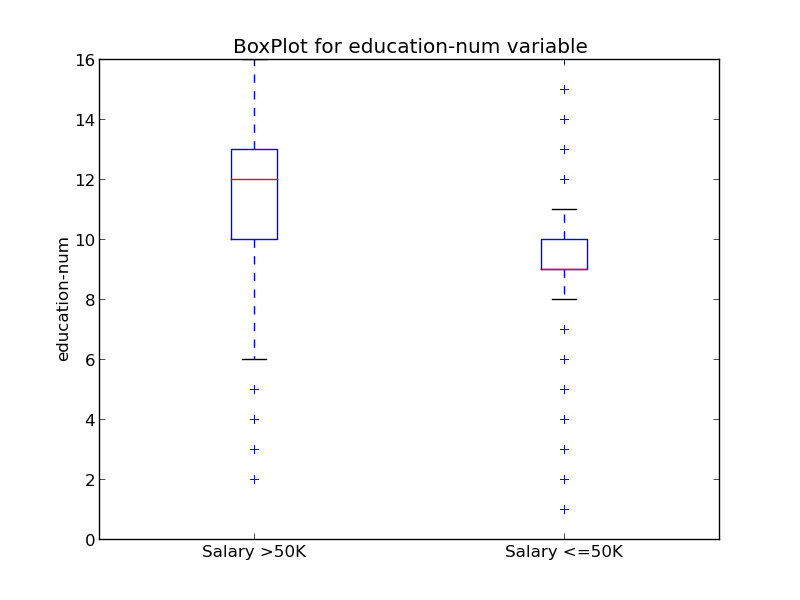
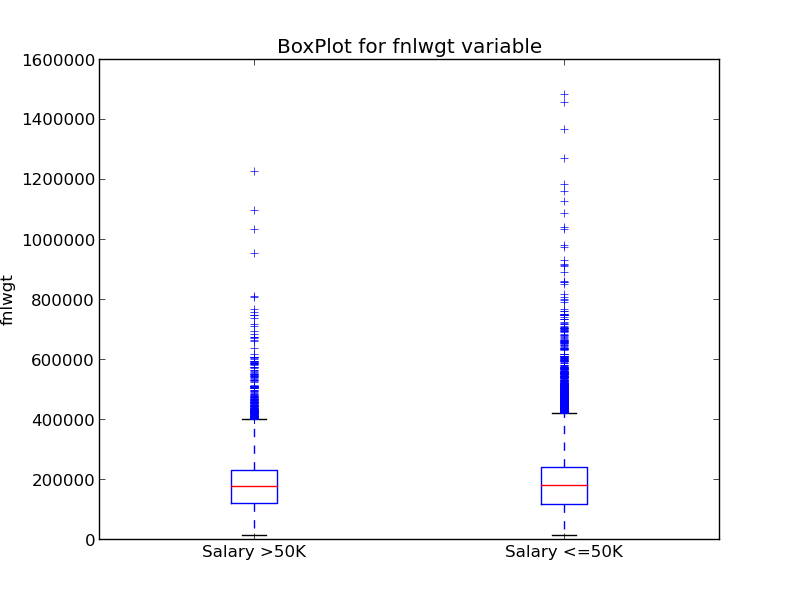
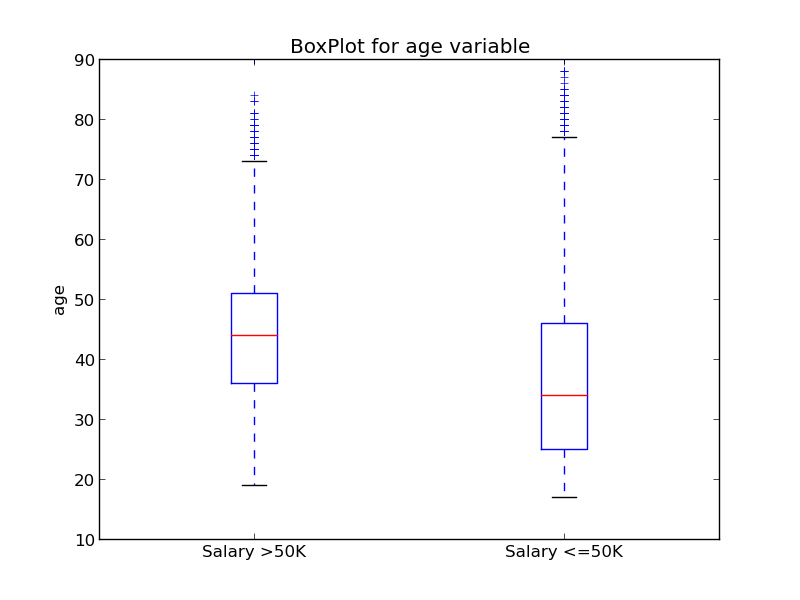
capital-gain 0 fraction is: 0.916682021989

capital-loss 0 fraction is: 0.95331982065

* 1. Plot 2 histograms(>50K and <=50K)
  2. Plot 2 boxplots(>50K and <=50K) for each variable
  3. Comment for numeric variables:

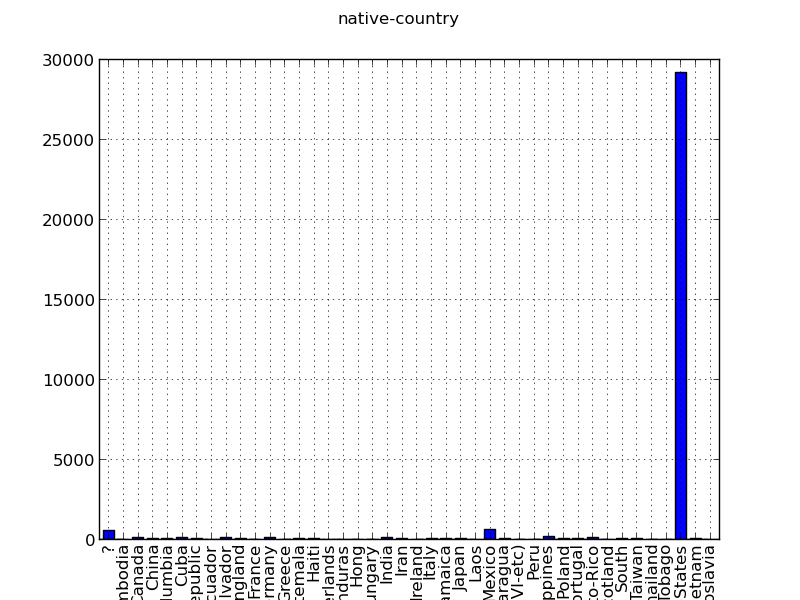
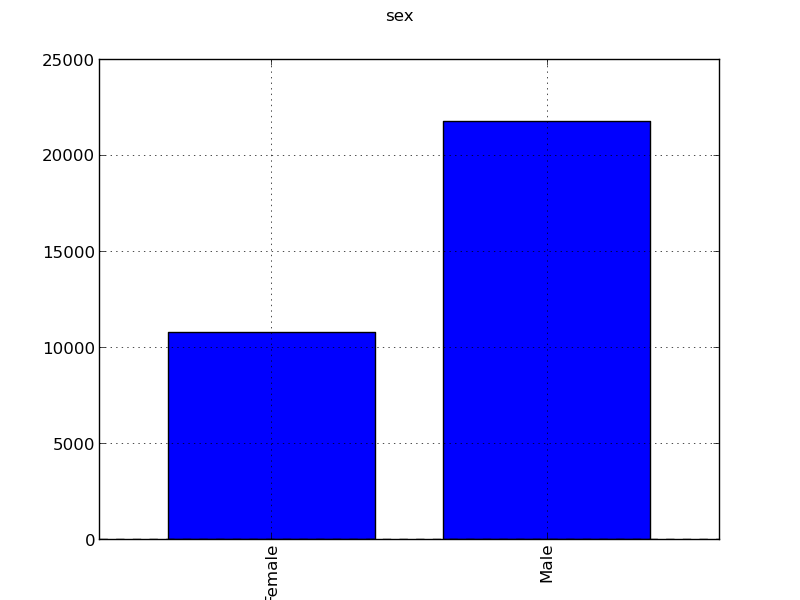
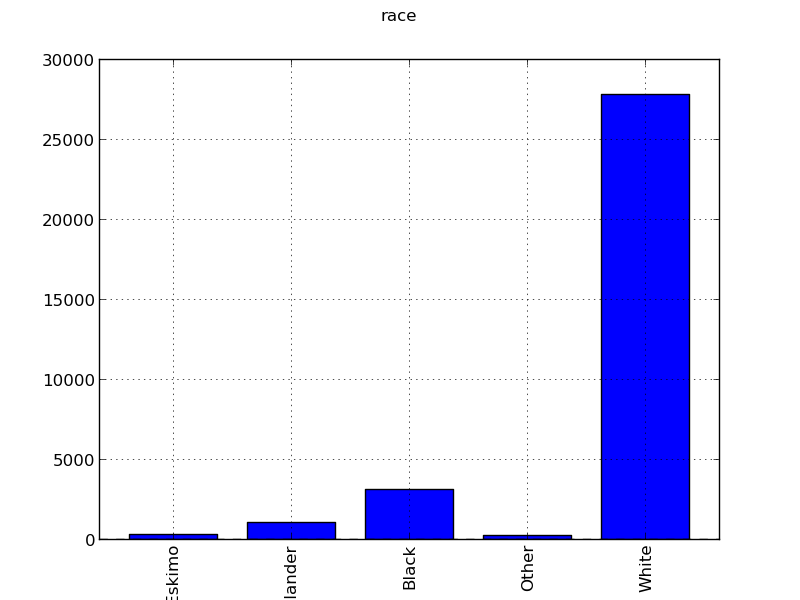
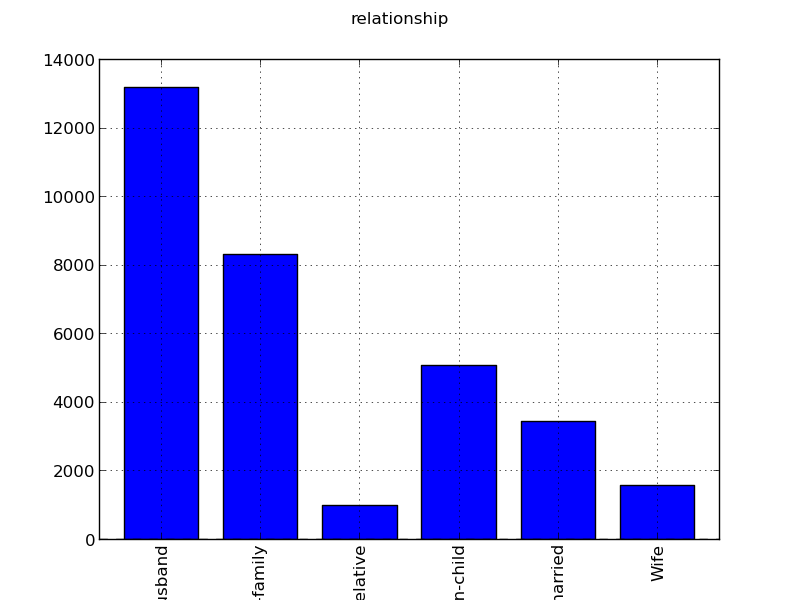
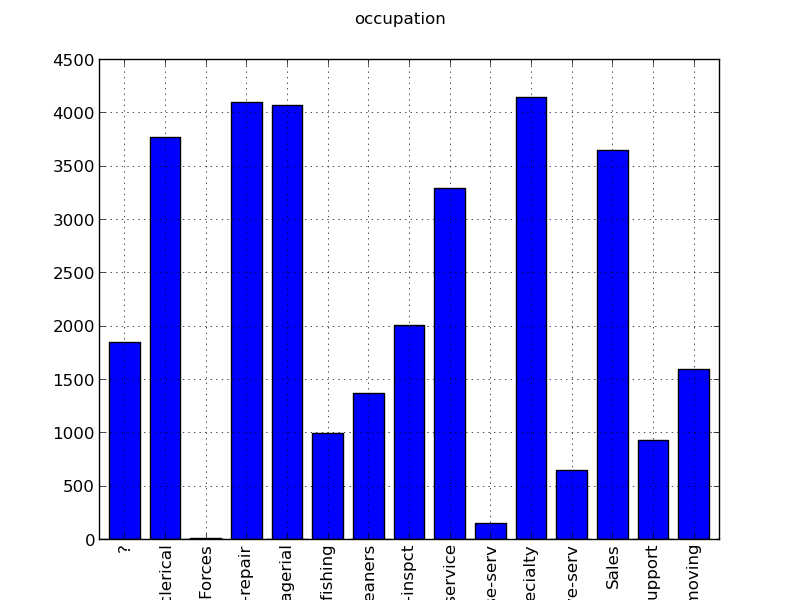
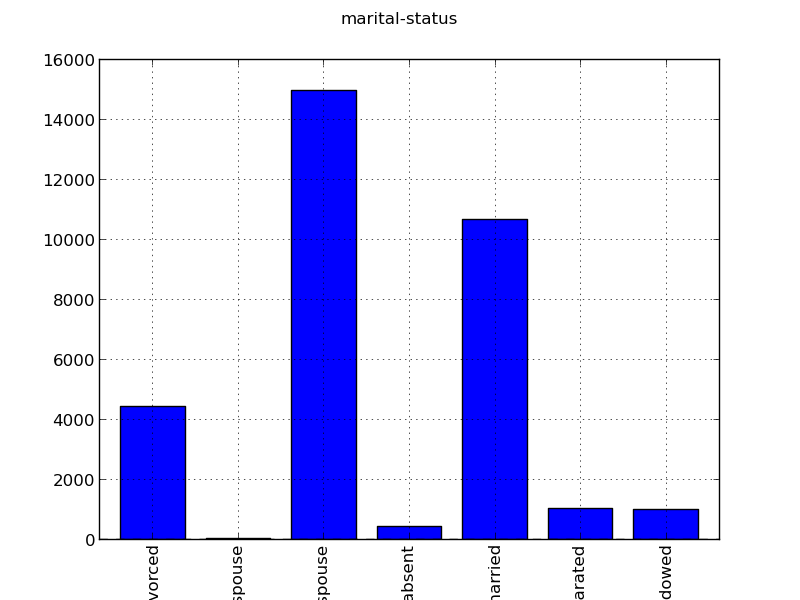
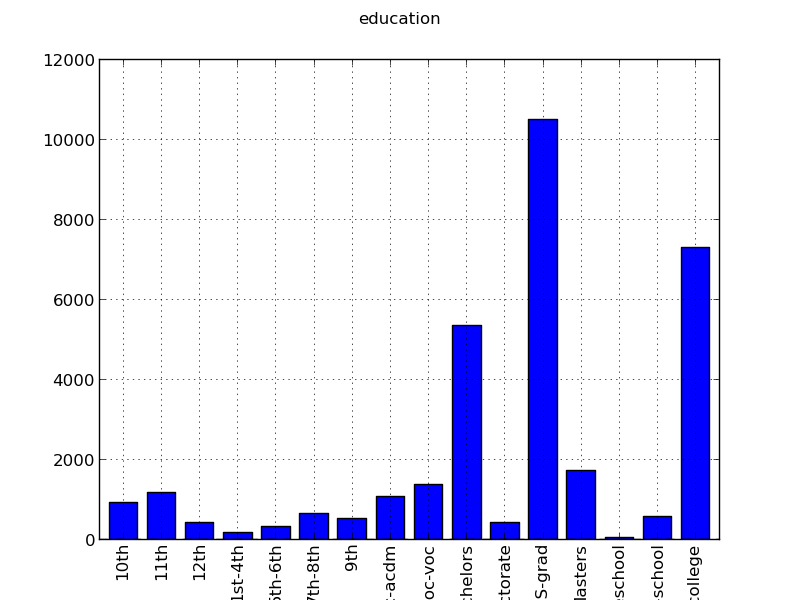
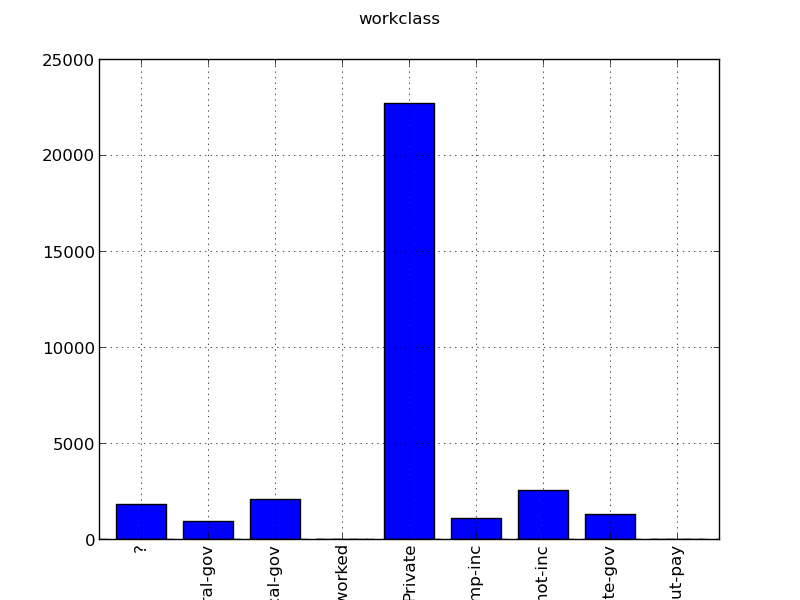
It is obvious that “age”, “education\_num” and “capital gain” are significantly associated with “salary”. The histograms and boxplots show that people who is older and/or has a higher education background tend to have higher salary. And people with higher salary have more capital gain. From the figures, “gnlwgt” seems to not to be too much related with “salary” because the two “fnlwgt” are almost the same. However, the “hour-per-work” surprised me. It shows that people with higher salary tend to work much less hours per week than people with lower salary.



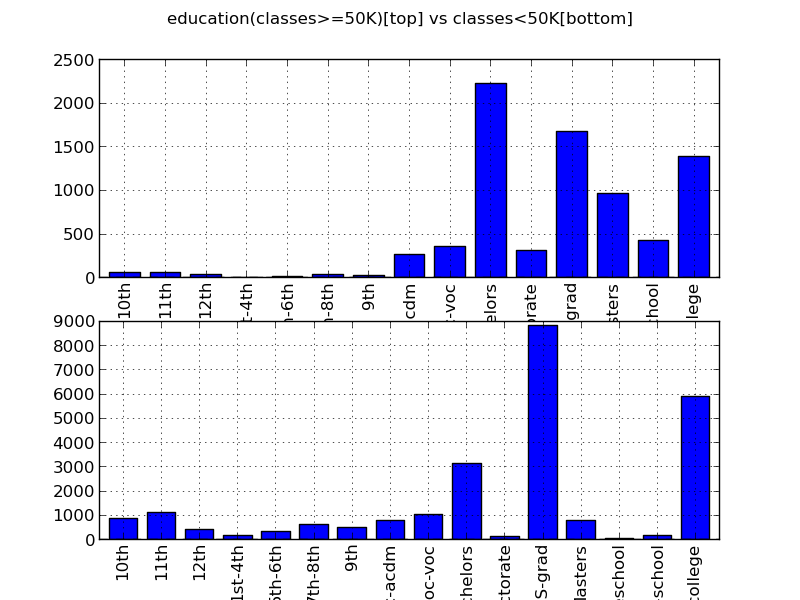
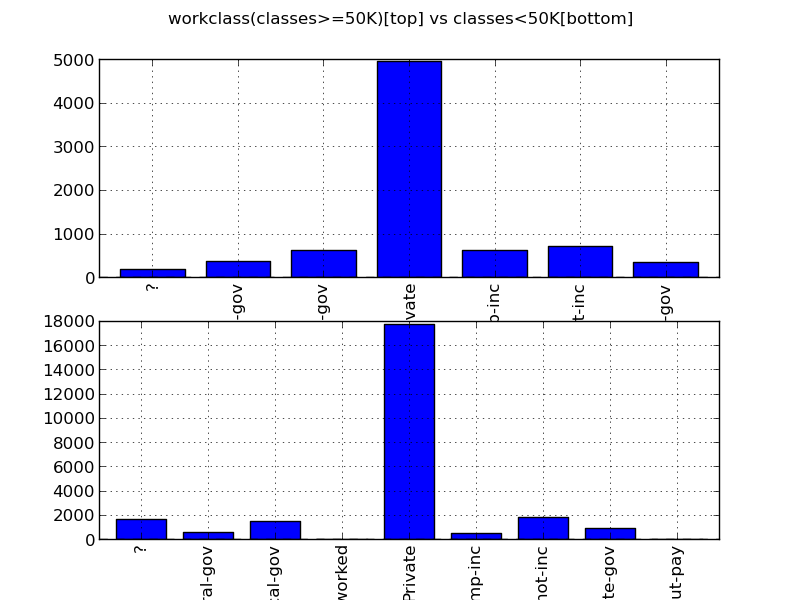


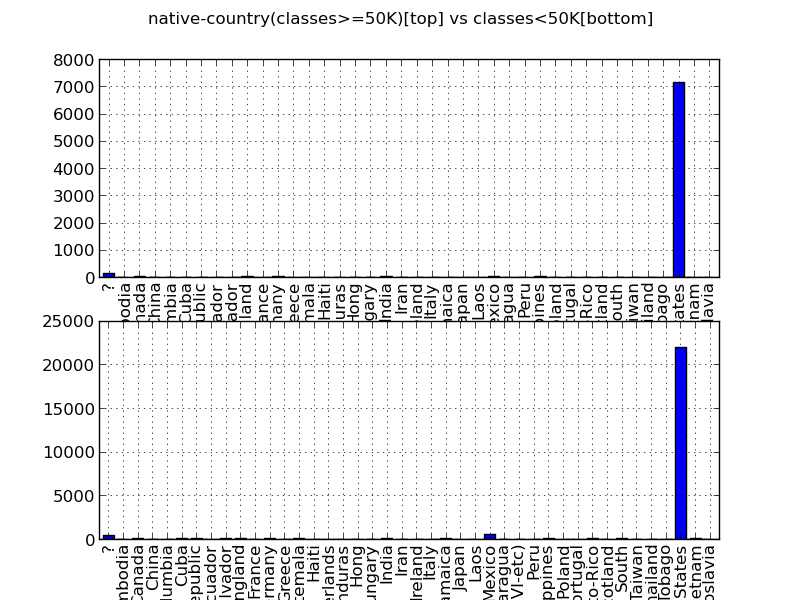
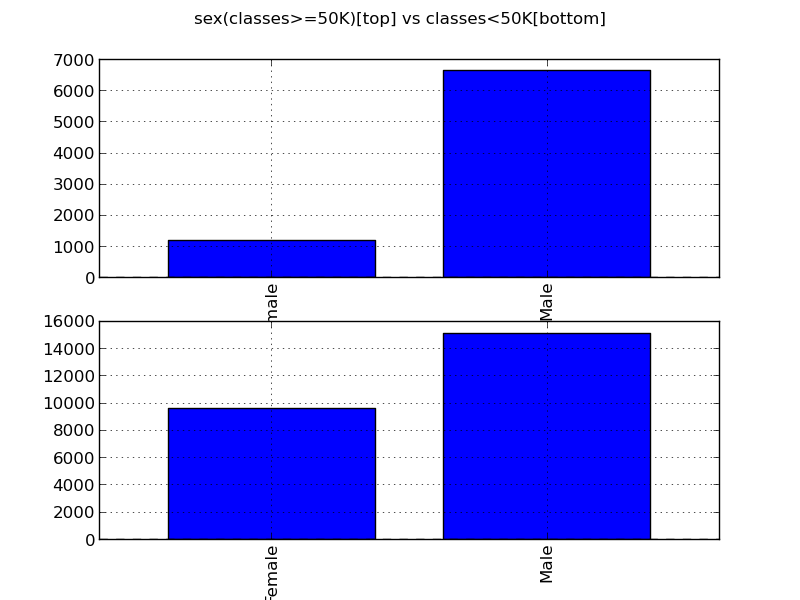
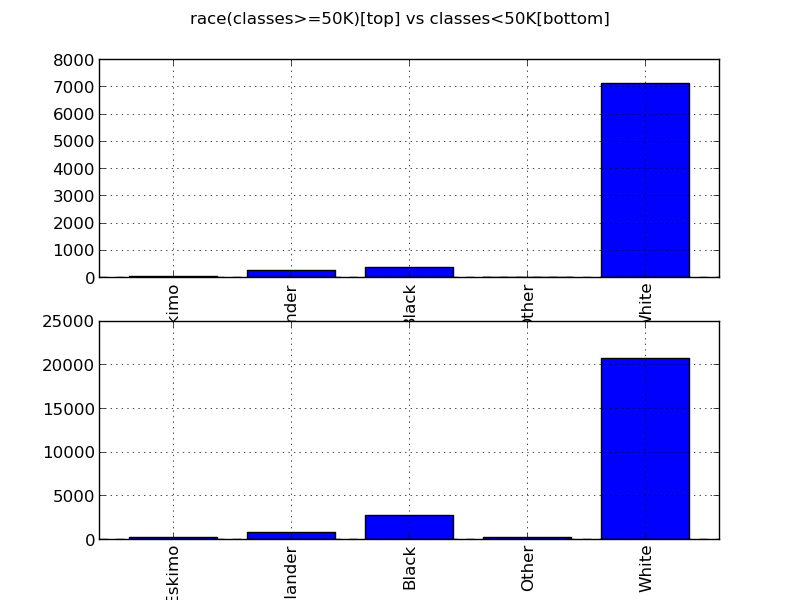
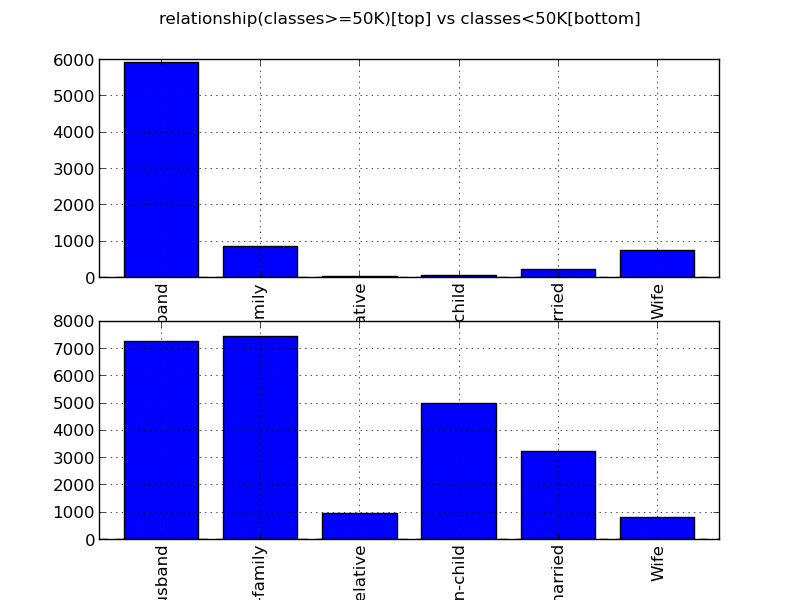
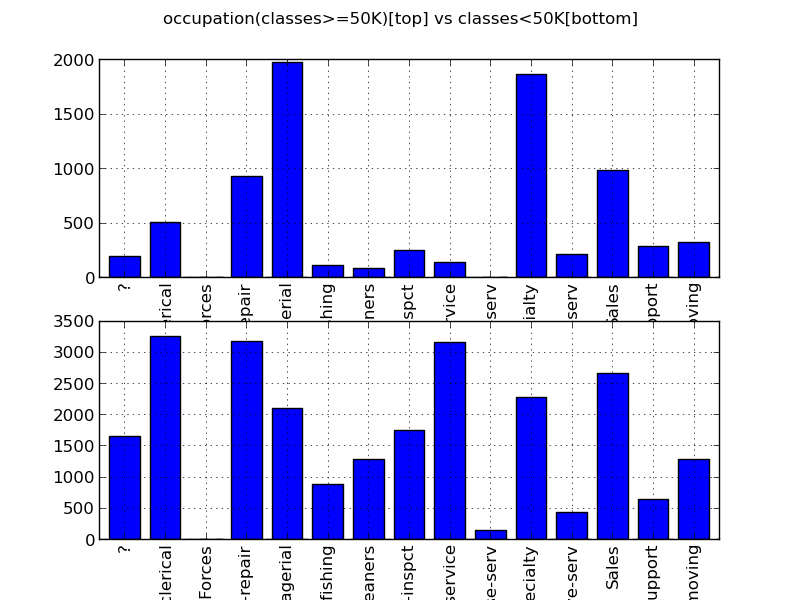
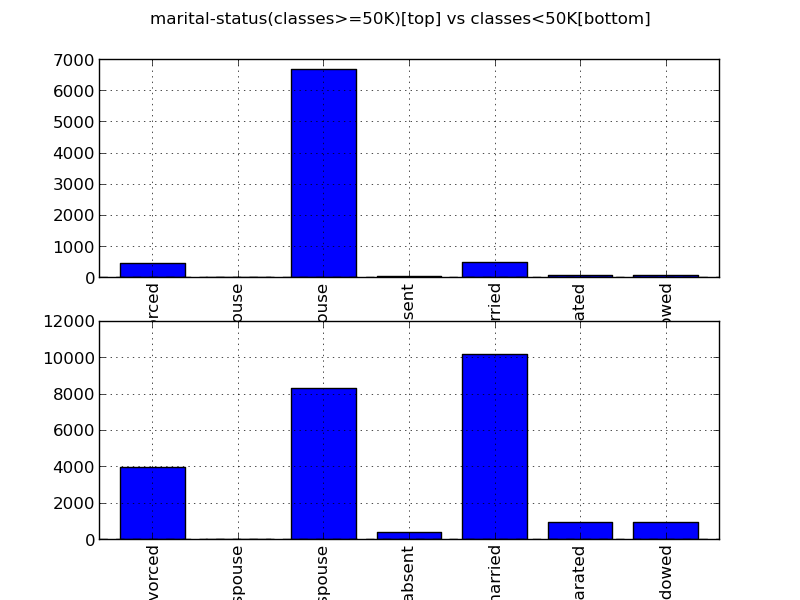
**4. Categorical Variables**

4.1 Generate a bar-plot for each categorical variable



4.2. Generate 2 barplots for the class >50k and the class <=50k





4.3. Compute the expected information gain

(1)Old Entropy = 0.796403803152

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Var | Workclass | Education | Marital\_status | Occupation | Relationship | Race | Sex | Native\_ctry |
| Entrop | .77 | .702 | .640 | .703 | .631 | .788 | .76 | .788 |
| IG | .22 | .094 | .157 | .093 | .165 | .008 | .04 | .009 |

(2) Every value of a variable as a binary and its IG

|  |  |  |
| --- | --- | --- |
| Work-Class | Missing | 0.0052 |
| Private | 0.0043 |
| Self-emp-not-inc | 0.0006 |
| Self-emp-inc | 0.0118 |
| Federal-gov | 0.0023 |
| Local-gov | 0.0008 |
| State-gov | 0.0002 |
| Without-pay | NaN |
| Never-worked | NaN |
| Education | Bachelors | 0.0214 |
| Some-college | 0.003 |
| 11th | 0.0071 |
| HS-grad | 0.0131 |
| Prof-school | 0.0141 |
| Assoc-acdm | 0 |
| Assoc-voc | 0.0001 |
| 9th | 0.003 |
| 7th-8th | 0.0033 |
| 12th | 0.0018 |
| Masters | 0.0186 |
| 1st-4th | 0.0012 |
| 10th | 0.0045 |
| Doctorate | 0.0103 |
| 5th-6th | 0.0021 |
| Preschool | NaN |
| Marital-Status | Married-civ-spouse | 0.1521 |
| Divorced | 0.0136 |
| Never-married | 0.0894 |
| Separated | 0.0051 |
| Widowed | 0.0037 |
| Married-spouse-absent | 0.0016 |
| Married-AF-spouse | 0.0001 |
| Occupation | Missing | 0.0053 |
| Tech-support | 0.0005 |
| Craft-repair | 0.0001 |
| Other-service | 0.0237 |
| Sales | 0.0004 |
| Exec-managerial | 0.0293 |
| Prof-specialty | 0.0222 |
| Handlers-cleaners | 0.0071 |
| Machine-op-inspct | 0.004 |
| Adm-clerical | 0.0065 |
| Farming-fishing | 0.0023 |
| Transport-moving | 0.0003 |
| Priv-house-serv | 0.0016 |
| Protective-serv | 0.0005 |
| Armed-Forces | 0 |
| Relationship | Wife | 0.0095 |
| Own-child | 0.0551 |
| Husband | 0.1167 |
| Not-in-family | 0.0291 |
| Other-relative | 0.007 |
| Unmarried | 0.0187 |
| Race | White | 0.0057 |
| Asian-Pac-Islander | 0.0001 |
| Amer-Indian-Eskimo | 0.0007 |
| Other | 0.0009 |
| Black | 0.0065 |
| Sex | Female | 0.0372 |
| Male | 0.0372 |
| Native-Country | Missing | 0 |
| United-States | 0.0009 |
| Cambodia | 0 |
| England | 0.0001 |
| Puerto-Rico | 0.0003 |
| Canada | 0.0001 |
| Germany | 0.0001 |
| Outlying-US(Guam-USVI-etc) | NaN |
| India | 0.0003 |
| Japan | 0.0001 |
| Greece | 0 |
| South | 0 |
| China | 0 |
| Cuba | 0 |
| Iran | 0.0001 |
| Honduras | 0.0001 |
| Philippines | 0.0001 |
| Italy | 0.0001 |
| Poland | 0 |
| Jamaica | 0.0002 |
| Vietnam | 0.0003 |
| Mexico | 0.0038 |
| Portugal | 0.0001 |
| Ireland | 0 |
| France | 0.0001 |
| Dominican-Republic | 0.0006 |
| Laos | 0 |
| Ecuador | 0 |
| Taiwan | 0.0001 |
| Haiti | 0.0001 |
| Columbia | 0.0004 |
| Hungary | 0 |
| Guatemala | 0.0004 |
| Nicaragua | 0.0002 |
| Scotland | 0 |
| Thailand | 0 |
| Yugoslavia | 0 |
| El-Salvador | 0.0004 |
| Trinadad&Tobago | 0.0001 |
| Peru | 0.0002 |
| Hong | 0 |
| Holand-Netherlands | NaN |

4.4. Comments for categorical variables

Firstly, people with higher education background tend to have higher salary, which is consistent with the result from numerical variable – “education\_num”. The barplots of “marital\_status” show that “spouse” almost is the only group of people who can have high salary. The ”never-married” has a significant decreased number from <50K to >=50K. For “Sex”, it also affects “salary” significantly since the proportion of male in the high salary group is much higher than in the low salary group. For “Occupation”, only two kinds of people make high salary – “managerial” and prof-pecialty. For “native-country”, it seems to be not related to the salary because the distribution doesn’t have too much difference in barplot.

**5. Pairwise dependency on Age**

5.1. Pick one of the categorical variables and check pairwise dependency on Age

I pick ‘’sex”, the partial conditional probability:

P(sex = Female | age = (16.927, 24.3]) = 0.450269

P(sex = Female | age = (24.3, 31.6]) = 0.345161

P(sex = Female | age = (31.6, 38.9]) = 0.291832

P(sex = Female | age = (38.9, 46.2]) = 0.301963

P(sex = Female | age = (46.2, 53.5]) = 0.288631

There are some differences among the empirical distributions of “sex” for different “age”, but not very significant. So they have some dependences, but not very strong.

5.2. Pick any 2 of the numeric variables and check dependence.

I pick “age” and “hours-per-week”, the partial conditional probability

P(age = (16.927, 24.3] | hours-per-week = (0.902, 25.5]) = 0.454476

P(age = (16.927, 24.3] | hours-per-week = (25.5, 50]) = 0.142583

P(age = (16.927, 24.3] | hours-per-week = (50, 74.5]) = 0.060578

It is significant that the empirical distributions of “age” are different for different “hours-per-week”. So they are dependent on each other. And the Correlation Coefficient of these two variables is 0.06875571. So, that means that these two variables are dependent on each other. The degree is 0.06875571 and it is positive correlative.

And here are some of two variables correlation coefficients values

|  |  |  |
| --- | --- | --- |
| Variable1 | Variable2 | Correlation Coefficients |
| Age | Hours-per-week | 0.06875571 |
| Education-Num | Fnlwgt | 0.03652718 |
| Capital-gain | Capital-loss | -0.03161506 |