CS6220: Data Mining Homework 01

Due: Feb. 17th
Name: Mingzhe Xu

Problem 1:

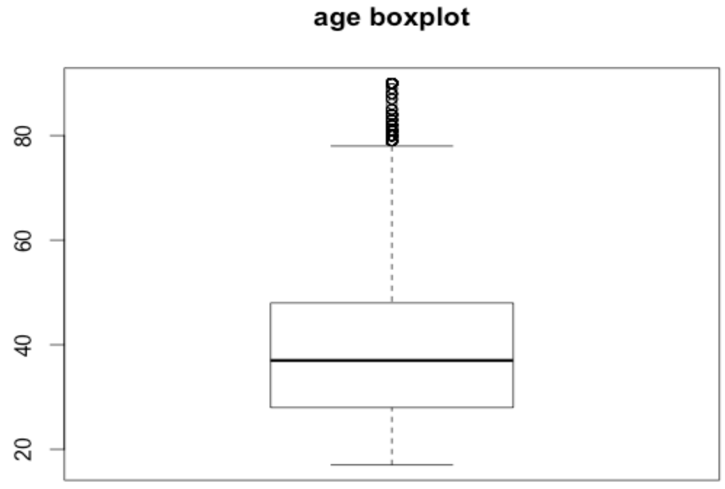
i) Compute mean, min, 1st Quartile, median, 3rd Quartile, Max, mode
(See R code)

ii) Table for summary statistics for continuous variables

Attribute	Min	Q1	Median	Q3	Max	Mean	Mode	Missing
Age	17	28	37	48	90	38.65	36	470
Fnlwgt	12285	117549	178145	237630	1490400	189664	203488	0
Edu.num	1	9	10	12	16	10	9	254
Cap.gain	0	0	0	0	99999	1079	0	0
Cap.loss	0	0	0	0	4356	87.5	0	0
Hours	1	40	40	45	99	40	40	737

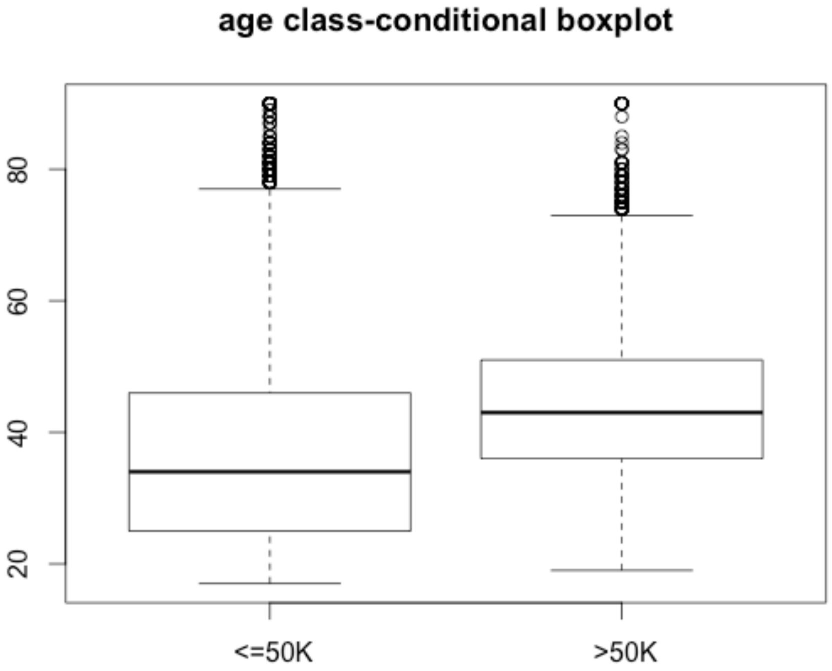
(b) Visualizing Data

ii) Box plot and class-conditional box plot

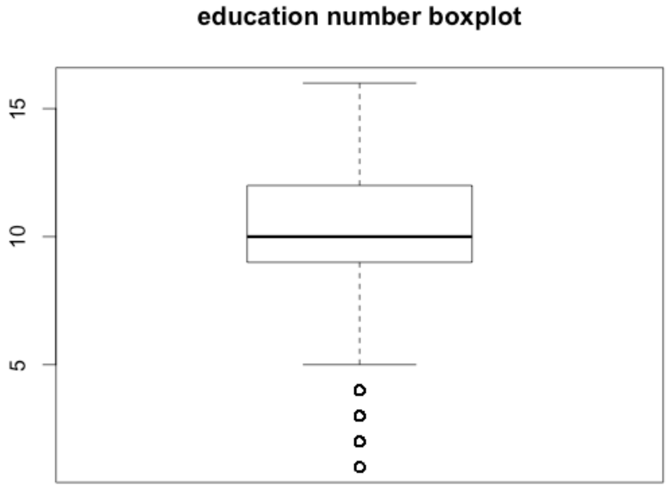


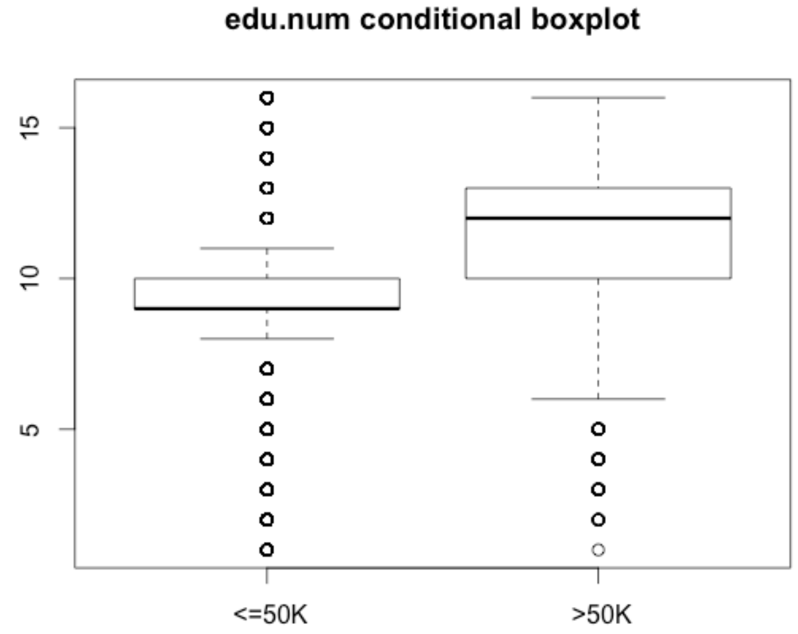
- Based on the box plot of **age**, we can derive several arguments:

# - the median is 37 so half of the people are below 37, half above.
# - half of the people are from 28 to 48
# - the conditional boxplot:
# -- the median age for income <= 50k is less than 37 and median age for income > 50k is more than 40

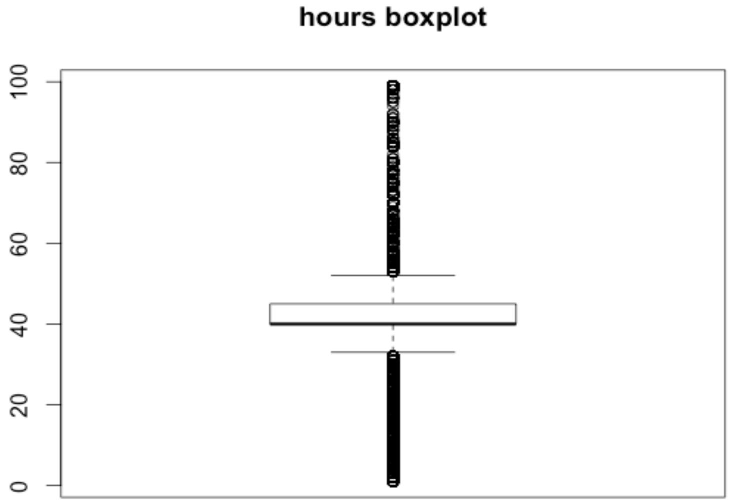


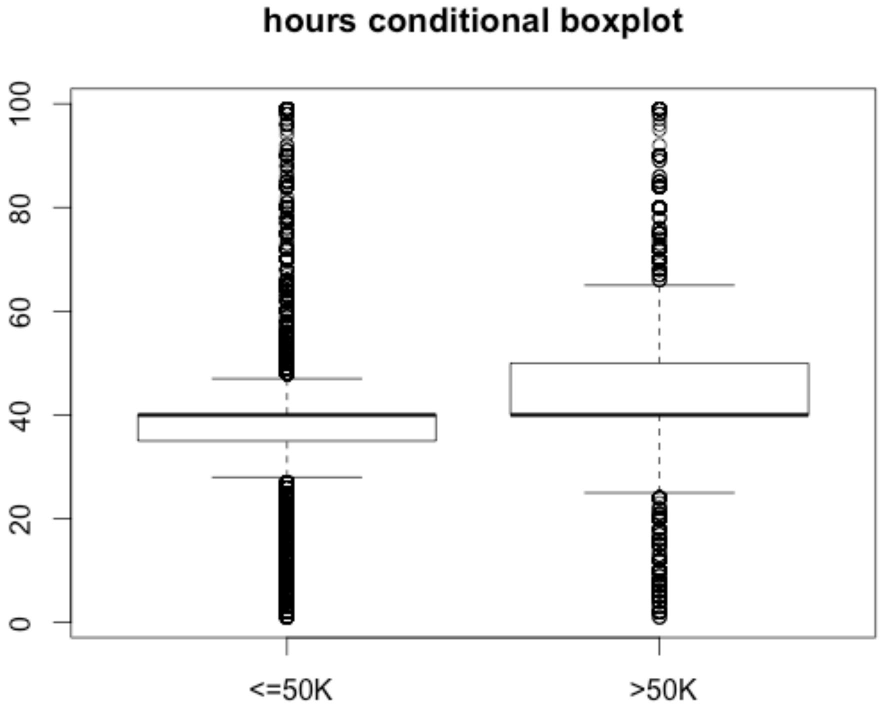
- Based on the box plot of **education number**, we can derive several arguments:
# - the median is 10 and half of the sample is within 9 - 12
# - 25% of the sample is 9 and 25% are 11 or 12
# - median of education number for those income <= 50k is 9, while for income more than 50k, the median education number is obviously more than 10





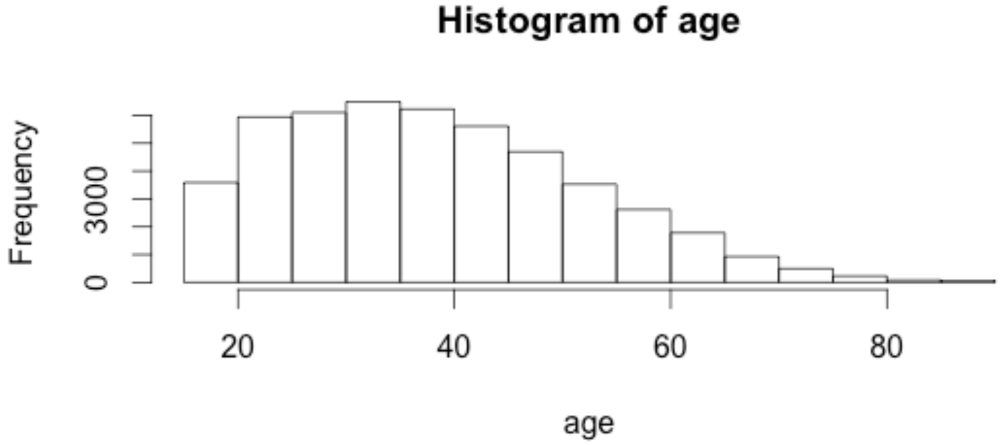
- Based on the box plot **of hours per week**, we can derive several arguments:
# - the median is 40 and around 25% of the sample is 40
# - 25% of the sample is between 40 and 45
# - the conditional boxplot shows that the median hours from people who has income less or equal to 50k is 40, the same as those who earn more than 50k
# - 75% people whose income is less or equal to 50k work less than or equal to 40 hour
# - 75% people whose income is more than 50k work more than or equal to 40 hours

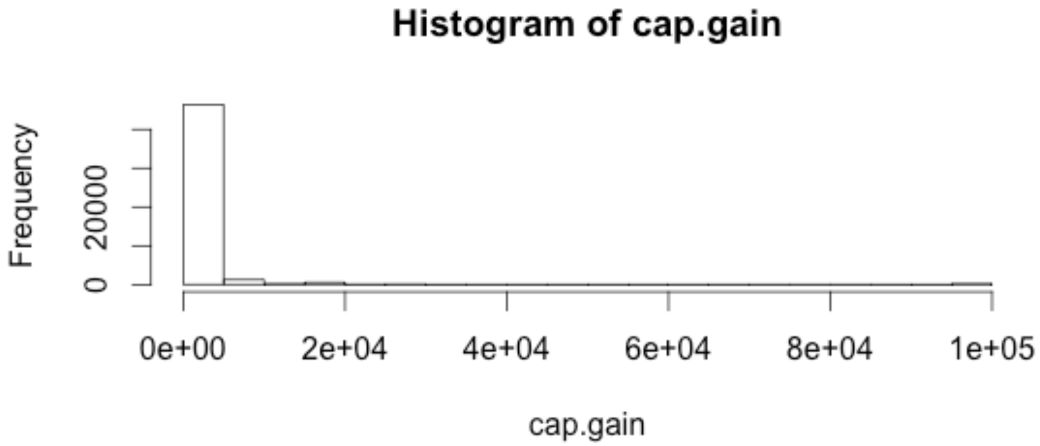


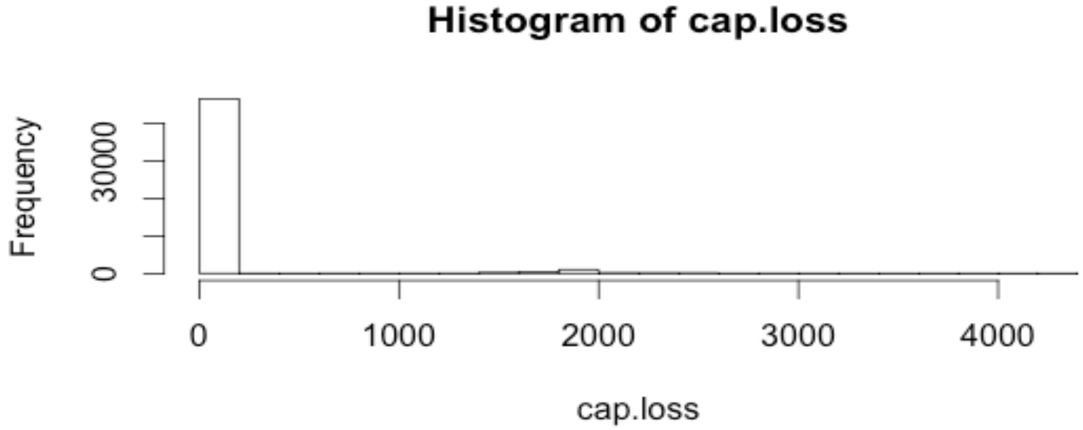


For capital.gain or capital.loss or capital gain, the value is either 0 or a large number and mostly is 0, so we couldn't tell much from the boxplot with median q1 q3 being 0.

iii) Histograms for age, cap.loss, cap.gain



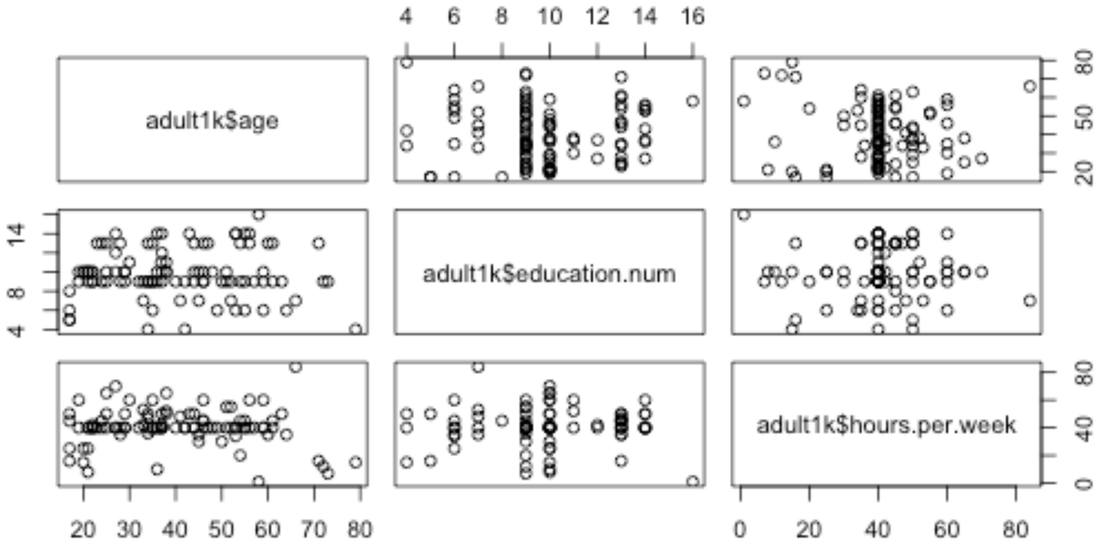




# Interpret:
# The histogram tells more than boxplots.
# For age, it is a little bit skewed to the left.
# And there are more people range from age 20 to 45 than other ranges
# cap.gain and cap.loss have similar characteristics - both are heavily skewed with most values being 0.

iv) barplots and class conditional bar plots for categorical variables
The one that most interest me is the bar plots of sex based on income. Obviously, among all male, the percentage of them with income >50k is higher than the percentage of female with income >50k. Another one is the relationship, it seems that the percentage of married people (husband and wife) with income >50k is higher than unmarried or other.
(The plots are in the Appendix)

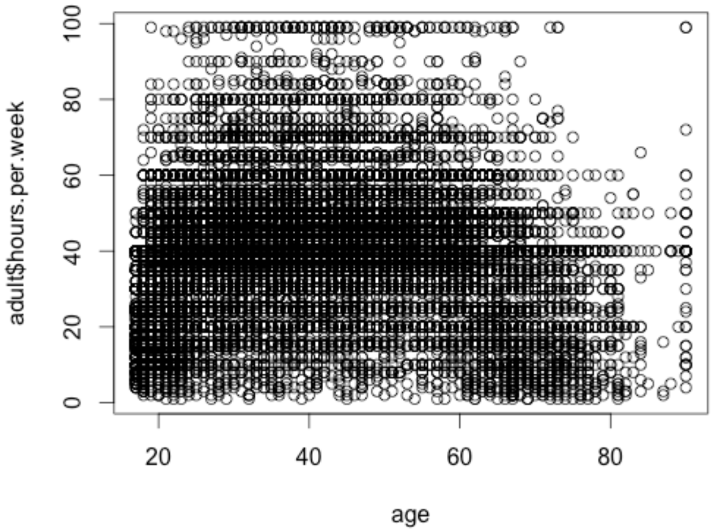
v) scatter plot matrix:



vi) Closer look at the scatter plot for age and hours per week.

Interpret:

- The most densely scattered points are among the range of age 20 – 60 and around the 40 hours per week line.
- Another area that has strong density is lower age and lower hours per week. Age around 20 and the hours scattered ranging from 5 to 25 hours per week.
- For elderly, people age 60+, they work less hours. Much less people work 40 hours with age above 60. Hours around 20 or bellow is more popular for age 60+.



-

vii) Computer chi-squared test for nominal variable workclass, education, marital.status, occupation, relationship, race, sex and native_country.

Test for workclass:

```
data:  table(adult$income, adult$workclass)
X-squared = 1610.752, df = 8, p-value < 2.2e-16
```

Test for education:

```
data:  table(adult$income, adult$education)
X-squared = 6537.973, df = 15, p-value < 2.2e-16
```

Test for marital status:

```
data:  table(adult$income, adult$marital.status)
X-squared = 9816.015, df = 6, p-value < 2.2e-16
```

```
data:  table(adult$income, adult$occupation)
X-squared = 5983.164, df = 14, p-value < 2.2e-16
```

```
data:  table(adult$income, adult$relationship)
X-squared = 10088.72, df = 5, p-value < 2.2e-16
```

```
data:  table(adult$income, adult$race)
X-squared = 487.0263, df = 4, p-value < 2.2e-16
```

```
data:  table(adult$income, adult$sex)
X-squared = 2248.848, df = 1, p-value < 2.2e-16
data:  table(adult$income, adult$native.country)
X-squared = 452.229, df = 41, p-value < 2.2e-16
```

Interpretation:
The p-values for all of them are all lower than < 2.2e-16, indicating significant correlation between the given feature and income.

Problem 2: Item Similarity

(a) two instances: x1 and x2

i) Euclidean Distance

```
# sqrt((-1-5)^2 + (6-2)^2 + 3^2 + (-1)^2)
# The result is 7.874008
```

ii) Manhattan Distance

```
# abs(-1-5) + abs(6-2)+ abs(3) + abs(-1)
# The result is 14
```

iii) Minkowski Distance

# h = 0: when h = 0, the distance is (-1-5)^0 + (6-2)^0 + 3^0 + (-1)^0 = 4
# when h = infinity, the result will be the max value of among the 4 dimensions, which is 4

(b) Prove that the Euclidean distance is always less than or equal to Manhattan distance

In order to prove it, I need to let the formula to the power of 2 on both side to get rid of the sqrt in Euclidean Distance formula

- Euclidean side: (x1-x2)^2 + (y1-y2)^2 + ...+(d1 - d2)^2
- Manhattan side:(abs(x1-x2) + abs(y1-y2) + ... + abs(d1-d2))^2

Then open the power of 2 for Manhattan. The result will be large than or equal to the value on Euclidean side.

(c) (See R code)

	Standardized				Non-standardized			
	Manhattan		Euclidean		Manhattan		Euclidean	
9	189	0	189	0	189	0	189	0
31	32475	0.07325	32475	0.07325	3007	1	3007	1
45	17	0	17	0	17	0	17	0

Problem 3 finding similar items

(a)

Element	S1	S2	S3	S4	H1	H2	H3
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

If we use these hash functions to get the signature:
Signature Matrix:

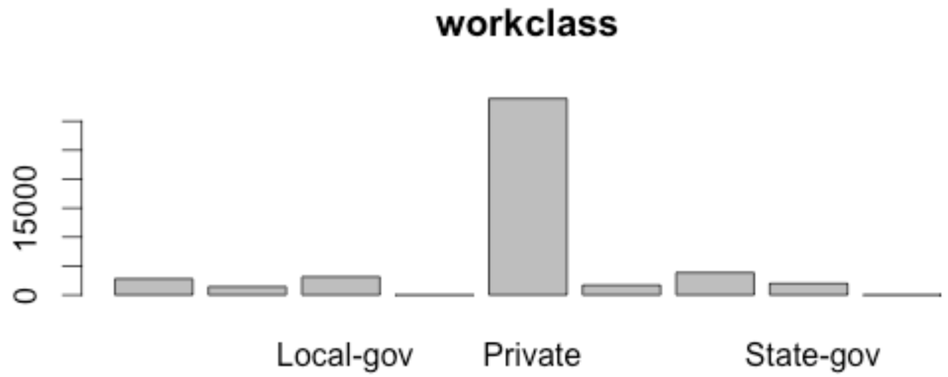
H1	5	1	1	1
H2	2	2	2	2
H3	0	1	4	0

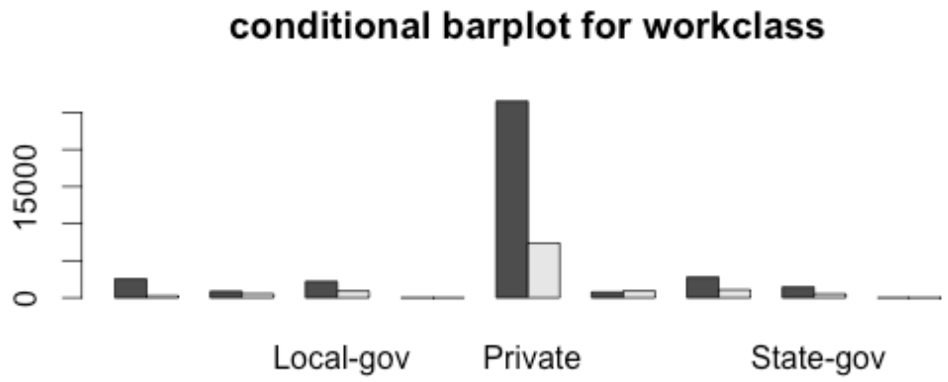
(b) Only h3 is a true permutation

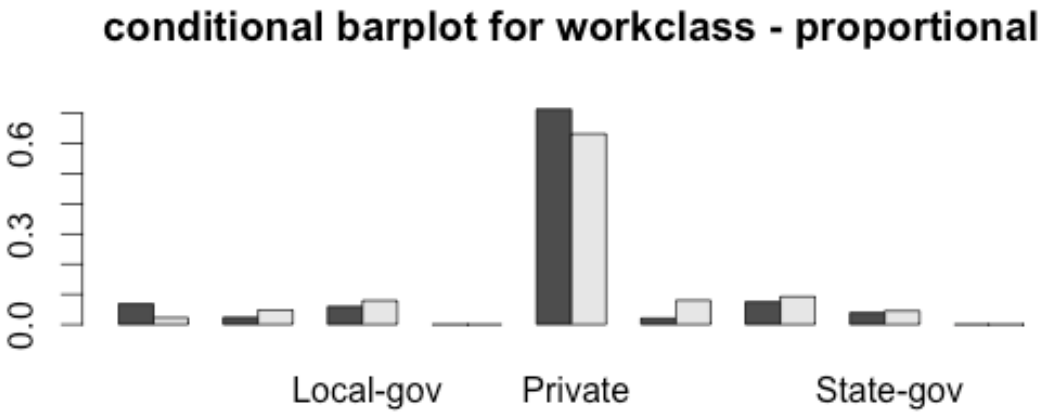
(c)

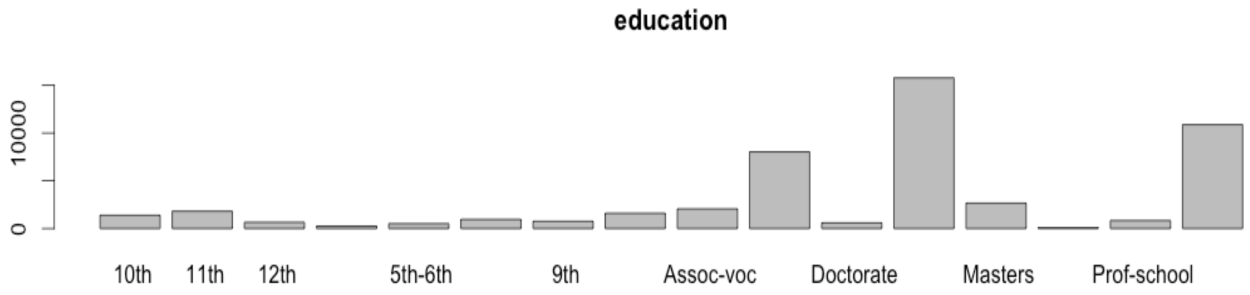
	S1-S2	S1-S3	S1-S4	S2-S3	S2-S4	S3-S4
Col/col	0	0	0.25	0	0.25	0.25
Sig/Sig	0.33	0.33	0.67	0.67	0.67	0.67

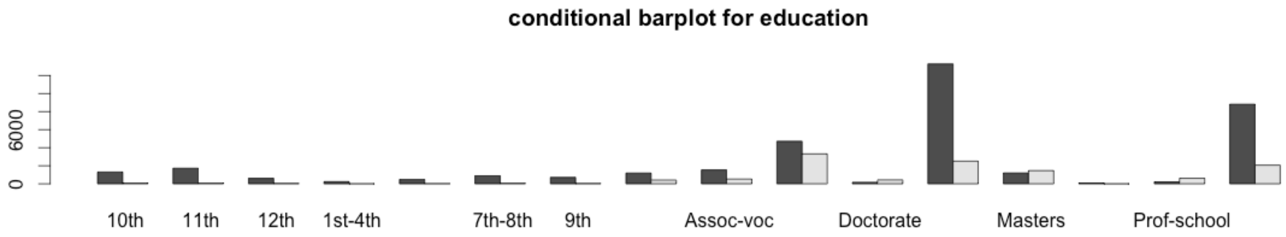
Appendix

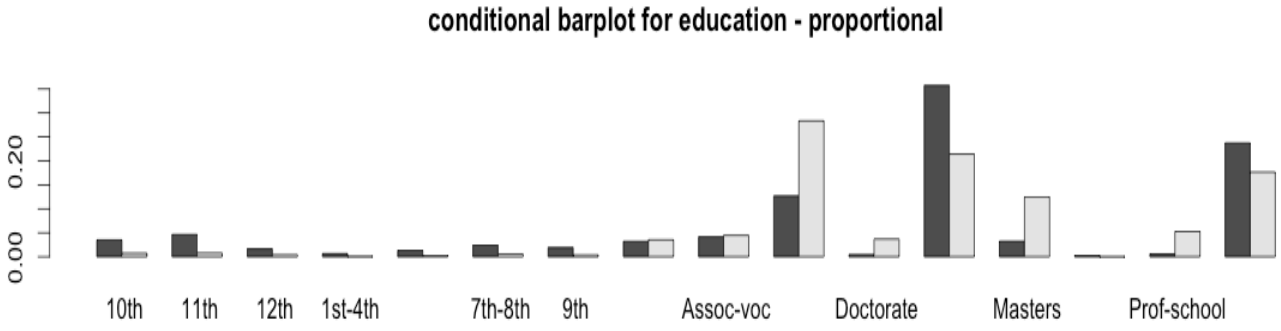

workclass
0
15000
Local-gov
Private
State-gov


conditional barplot for workclass
0
15000
Local-gov
Private
State-gov

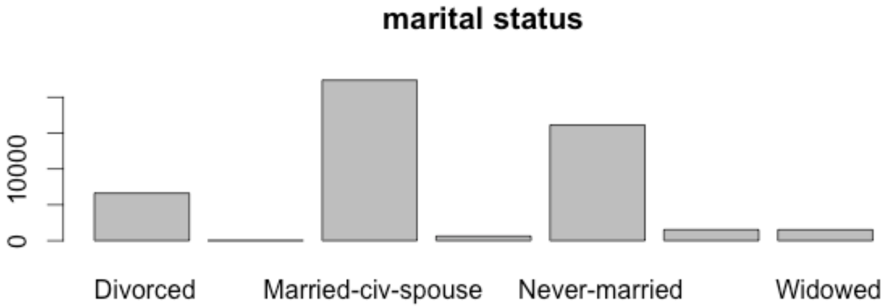

conditional barplot for workclass - proportional
0.0
0.3
0.6
Local-gov
Private
State-gov


education
0
10000
10th
11th
12th
5th-6th
9th
Assoc-voc
Doctorate
Masters
Prof-school

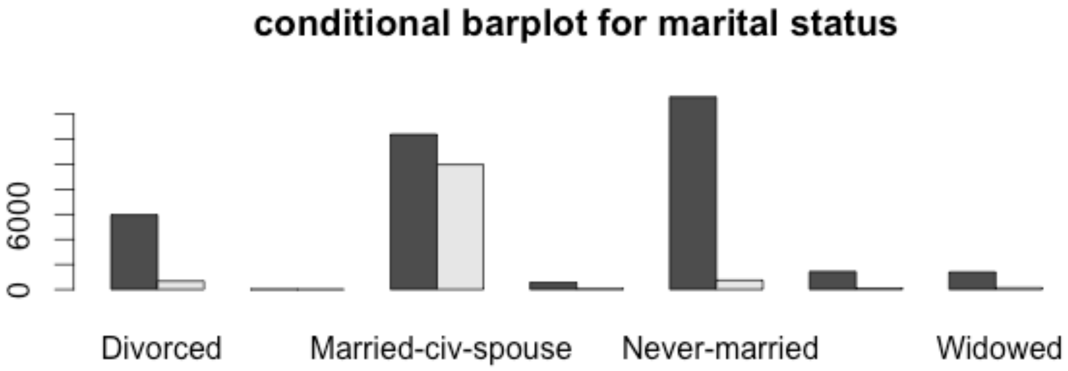

conditional barplot for education
0
6000
10th
11th
12th
1st-4th
7th-8th
9th
Assoc-voc
Doctorate
Masters
Prof-school



conditional barplot for education - proportional

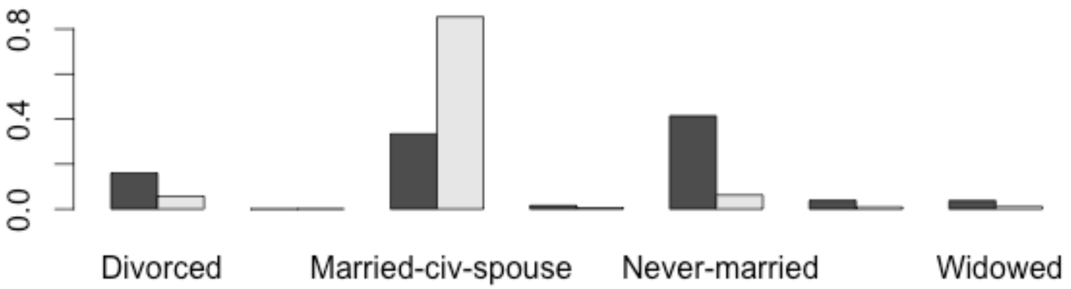


marital status

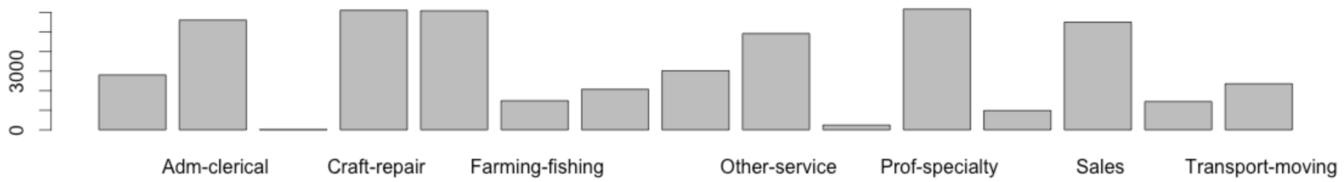


conditional barplot for marital status

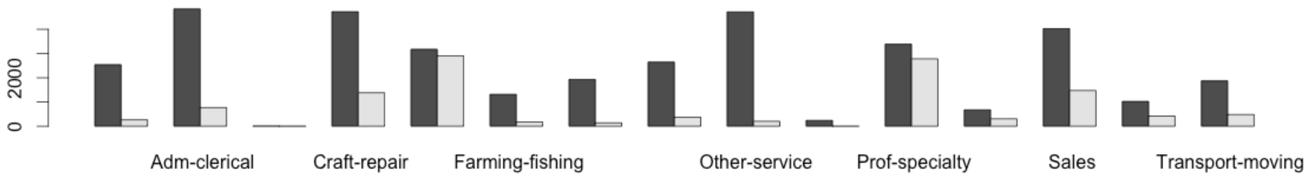
conditional barplot for marital status - proportional



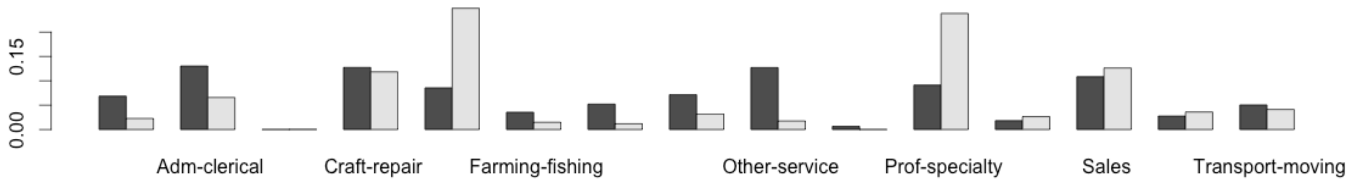
occupation

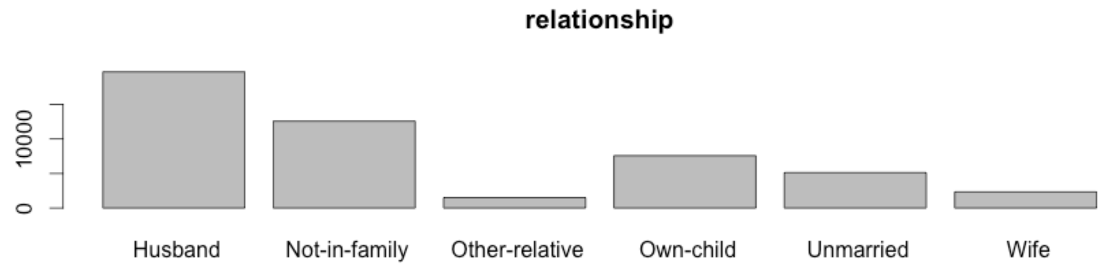


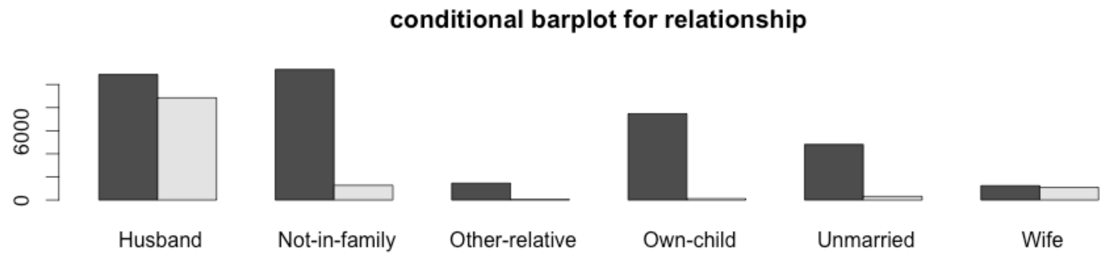
conditional barplot for occupation

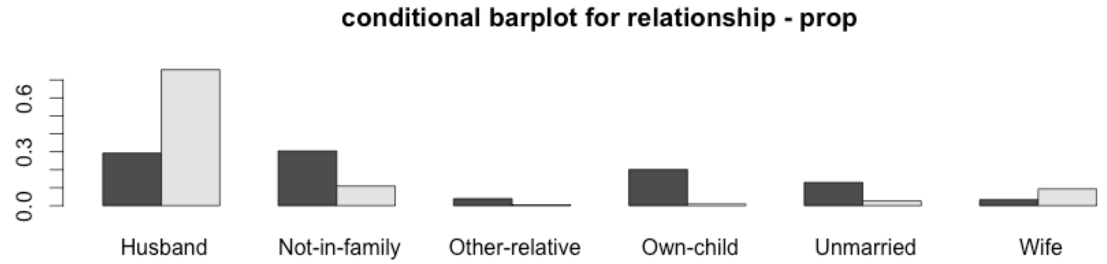


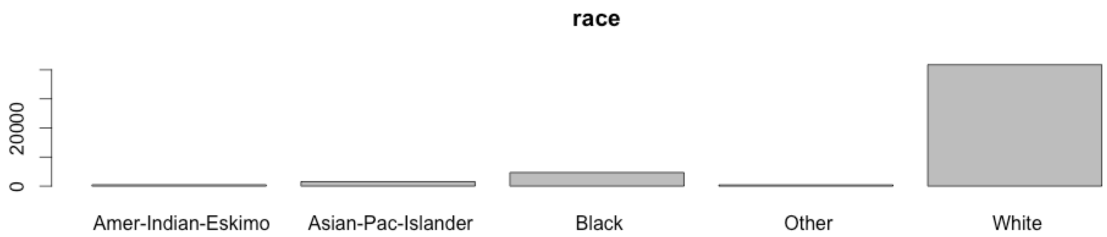
conditional barplot for occupation - proportional



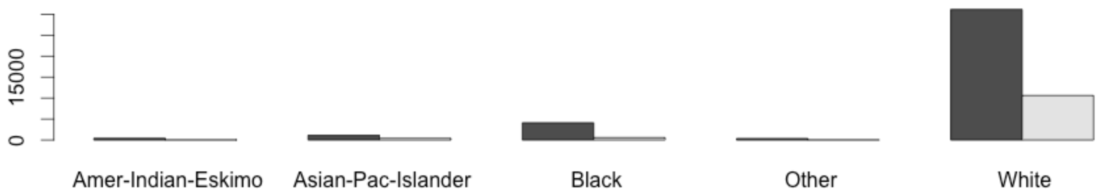

relationship
0
10000
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife


conditional barplot for relationship
0
6000
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife

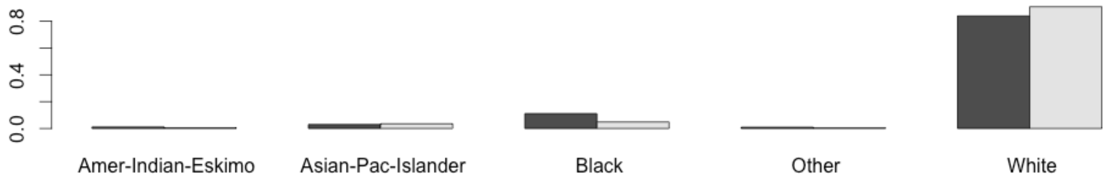

conditional barplot for relationship - prop
0.0
0.3
0.6
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife


race
0
20000
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

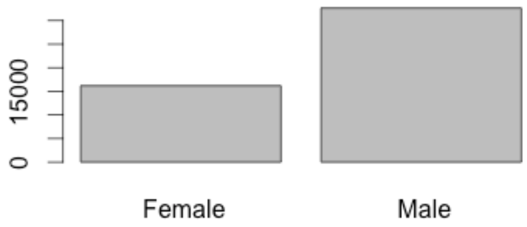
conditional barplot for race



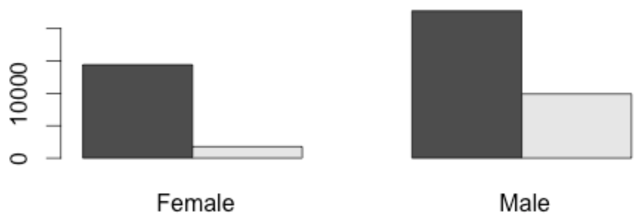
conditional barplot for race - prop

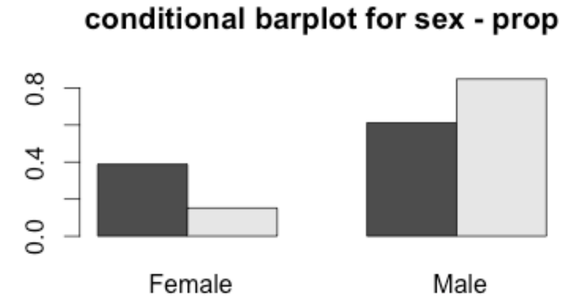


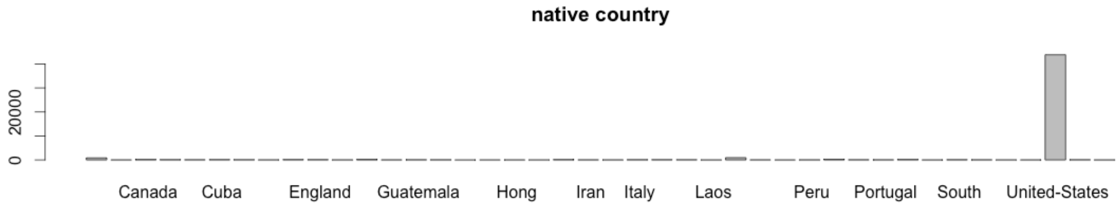
sex

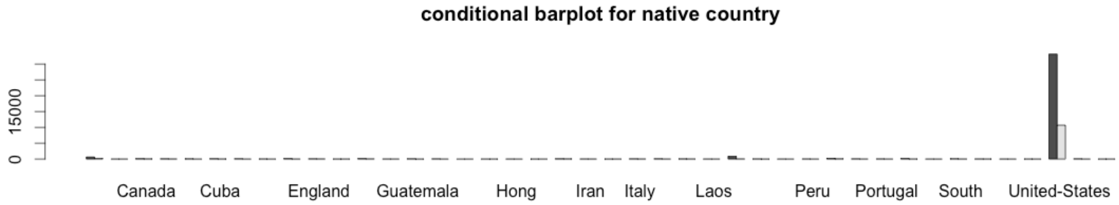


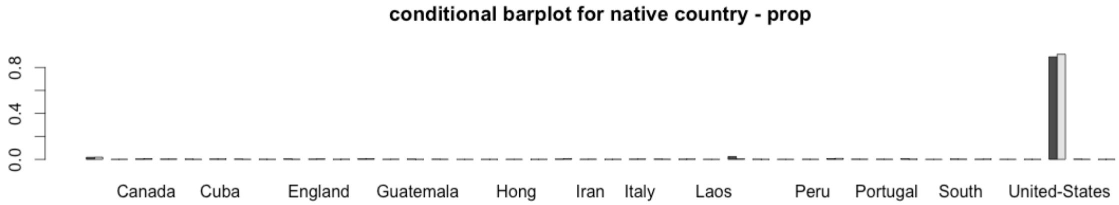
conditional barplot for sex




conditional barplot for sex - prop
0.0
0.4
0.8
Female
Male


native country
0
20000
Canada
Cuba
England
Guatemala
Hong
Iran
Italy
Laos
Peru
Portugal
South
United-States


conditional barplot for native country
0
15000
Canada
Cuba
England
Guatemala
Hong
Iran
Italy
Laos
Peru
Portugal
South
United-States


conditional barplot for native country - prop
0.0
0.4
0.8
Canada
Cuba
England
Guatemala
Hong
Iran
Italy
Laos
Peru
Portugal
South
United-States