Predicting Individual Income Using US Census Data

MATH 2319 Machine Learning Applied Project Phase I

Yong Kai Wong (s9999999) & Vural Aksakalli (s0000000)

1 January 1900

Contents

1 Introduction

The objective of this project was to build classifiers to predict whether an individual earns more than USD 50,000 or less in a year from the 1994 US Census Data. The data sets were sourced from the UCI Machine Learning Repository. This project has two phases. Phase I focuses on data preprocessing and exploration, as covered in this report. We shall present model building in Phase II. The rest of this report is organised as follow. Section 2 describes the data sets and their attributes. Section 3 covers data pre-processing. In Section 4, we explore each attribute and their inter-relationships. The last section ends with a summary.

2 Data Set

The UCI Machine Learning Repository provides five data sets, but only `adult.data`, `adult.test`, and `adult.names` were useful in this project. `adult.data` and `adult.test` are the training and test data sets respectively. `adult.names` contains the details of attributes or variables. The training data set has 32,561 training observations. Meanwhile, the test data set has 16,281 test observations. Both data sets consist of 14 descriptives features and one target feature. In this project, we combined both training and test data into one. In Phase II, we would build the classifiers from the combined the data set and evaluate their performance using cross-validation.

2.1 Target Feature

The response feature is income which is given as:

$$\text{income} = \begin{cases} > 50K & \text{if the income exceeds USD 50,000} \\ \leq 50K & \text{otherwise} \end{cases} \tag{1}$$

The target feature has two classes and hence it is a binary classification problem. To reiterate, The goal is to predict **whether a person makes over $50K a year**.

2.2 Descriptive Features

The variable description is produced here from `adult.names` file:

- 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.

Most of the descriptive features are self-explanatory, except fnlwgt which stands for "Final Weight" defined by the US Census. The weight is an "estimate of the number of units in the target population that the responding unit represents". This feature aims to allocate similar weights to people with similar demographic characteristics. For more details, see US Census.

3 Data Pre-processing[1]

3.1 Preliminaries (Optional)

In this project, we used the following `R` packages.

```
library(knitr)
library(mlr)
library(tidyverse)
library(GGally)
library(cowplot)
```

We read and merged training and test datasets by treating the string values as characters. We would later convert the string columns to factor (categorical) after the data processing. For naming consistency with the data dictionary, we purposely skipped the headers and manually renamed the columns.

```
train <- read.csv('Datasets/adult.data.txt', stringsAsFactors = FALSE, header = FALSE)
test  <- read.csv('Datasets/adult.test.txt', stringsAsFactors = FALSE, skip = 1, header = FALSE)
adult <- rbind(train, test)


names( adult ) <- c('age', 'workclass', 'fnlwght', 'education', 'education_num',
                    'marital_status', 'occupation', 'relationship', 'race', 'sex',
                    'capital_gain', 'capital_loss', 'hours_per_week', 'native_country',
                    'income')
```

3.2 Data Cleaning and Transformation

With `str` and `summarizeColumns` (see Table 1), we noticed the following anomalies:

- All character columns contained excessive white space.
- The target feature, `income` had a cardinality of 4, which was supposed to be 2 since `income` must be binary.
- The `education_num` ranged from 1 to 16 which coincided with the cardinality of `education`. They might represent the same information.
- The max value of `capital_gain` was 99999, potentially a value to represent missing value.
- The max value of `hours_per_week` was 99. It could be a valid or missing value
- On surface, each feature had no missing value, especially the character features.

[1] In report submission, the code chunks are optional.

```
str(adult)
```

```
## 'data.frame':    48842 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
##  $ fnlwght       : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: chr  " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
##  $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Husband" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ sex           : chr  " Male" " Male" " Male" " Male" ...
##  $ capital_gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native_country: chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ income        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

```
summarizeColumns(adult) %>% knitr::kable( caption =  'Feature Summary before Data Preprocessing')
```

Table 1: Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
age	integer	0	38.64359	1.371051e+01	37.0	14.8260	17	90	0
workclass	character	0	NA	3.058024e-01	NA	NA	10	33906	9
fnlwght	integer	0	189664.13460	1.056040e+05	178144.5	89394.1083	12285	1490400	0
education	character	0	NA	6.768355e-01	NA	NA	83	15784	16
education_num	integer	0	10.07809	2.570973e+00	10.0	1.4826	1	16	0
marital_status	character	0	NA	5.418083e-01	NA	NA	37	22379	7
occupation	character	0	NA	8.736333e-01	NA	NA	15	6172	15
relationship	character	0	NA	5.963310e-01	NA	NA	1506	19716	6
race	character	0	NA	1.449572e-01	NA	NA	406	41762	5
sex	character	0	NA	3.315180e-01	NA	NA	16192	32650	2
capital_gain	integer	0	1079.06763	7.452019e+03	0.0	0.0000	0	99999	0
capital_loss	integer	0	87.50231	4.030046e+02	0.0	0.0000	0	4356	0
hours_per_week	integer	0	40.42238	1.239144e+01	40.0	4.4478	1	99	0
native_country	character	0	NA	1.025757e-01	NA	NA	1	43832	42
income	character	0	NA	4.938782e-01	NA	NA	3846	24720	4

Firstly, we removed the excessive white spaces for all character features.

```
adult[, sapply( adult, is.character )] <- sapply( adult[, sapply( adult, is.character )], trimws)
```

Second, we found that some `income` values were encoded with "`.`" as revealed in Table 2 and we removed them so it became binary[2]. Note that a higher proportion of people earning less than $50,000 in 1994. Therefore, we would require additional parameter-tuning in building models to cater such unbalanced class.

```
table(adult$income) %>% kable(caption = 'Number of Income Classes before Data Preprocessing')
```

[2]In fact, the "`.`" symbol is used to indicate target features from the test data.

Table 2: Number of Income Classes before Data Preprocessing

Var1	Freq
<=50K	24720
<=50K.	12435
>50K	7841
>50K.	3846

```
adult$income <- sub('K.', "K", adult$income)
```

We defined a new feature named `capital` which is `capital_gain` minus `capital_loss` because it is impossible that an individual had to pay taxes on capital gain from their financial investment and claim tax deductions from capital loss at the same time.

```
adult$capital <- adult$capital_gain - adult$capital_loss
adult$capital_gain <- NULL
adult$capital_loss <- NULL
```

We were ambivalent in removing 99999 values from `capital` because it could be a valid value. That is, an individual could see his or her investment increased in value by $99999. We decided to keep it as we could either bin it into categorical feature when deploying machine learning models which are robust to outliers in Phase II. Table 3 reports the summary statistics of `capital` after removing individuals with `capital` values of 99999. The maximum value drops substantially but the median is zero and far below the mean, suggesting presence of skewness. Consistent with the data exploration in later section, we shall show that scaling did not help solve the heavy skewness isssue. Table 4 reveals that individuals with `capital` values of 99999 were high income earners in 1994. Such finding is plausible and it explains why we were reluctant to remove these observations.

```
adult %>% select( capital ) %>% filter ( capital < 99999 ) %>%
  summary() %>% kable(caption = 'Summary Statistics of Capital after Removing 99999')
```

Table 3: Summary Statistics of Capital after Removing 99999

capital
Min. :-4356.0
1st Qu.: 0.0
Median : 0.0
Mean : 494.5
3rd Qu.: 0.0
Max. :41310.0

```
adult %>% filter ( capital == 99999 ) %>%
  select( income, capital) %>% table() %>%
  kable(caption = 'Income Classes for Individuals with 99999 Capital Value')
```

Table 4: Income Classes for Individuals with 99999 Capital Value

	99999
>50K	244

We computed the level table for each character feature.[3] The tables revealed:

- The missing values were encoded as `?` in `workclass`, `occuptation`, `native_country`
- Only 15 individuals worked in Armed Forces.
- `workclass` and `occupation` likely carried different information about the data. We reduced the cardinality of `workclass` by merging missing values, `Never-worked` and `Without=pay` as `Other` and grouping the government-related jobs as one.
- For `marital_status`, we aggregated the levels with the prefix of `Married` as one category.
- Almost all individuals were born in the United States. Therefore, we redefined `native_country` into two classes: `US` and `Non-US`.[4]
- Almost 50 % of individuals were married.
- There were more males than females.
- Most of individuals attended high schools, followed by some colleges. However, it might depend on age groups as the elder generations had fewer opportunities to enroll in higher educations. We grouped the post-graduate levels such as `Doctorate` and `Masters` into one level.

```
sapply( adult[ sapply(adult, is.character)], table)
```

```
## $workclass
##
##                 ?      Federal-gov        Local-gov     Never-worked
##              2799             1432             3136               10
##           Private     Self-emp-inc Self-emp-not-inc        State-gov
##             33906             1695             3862             1981
##      Without-pay
##                21
##
## $education
##
##         10th         11th         12th      1st-4th      5th-6th
##         1389         1812          657          247          509
##      7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
##          955          756         1601         2061         8025
##    Doctorate      HS-grad      Masters    Preschool  Prof-school
##          594        15784         2657           83          834
## Some-college
##        10878
##
## $marital_status
##
##              Divorced     Married-AF-spouse    Married-civ-spouse
##                  6633                   37                 22379
## Married-spouse-absent         Never-married              Separated
##                   628                 16117                  1530
##               Widowed
##                  1518
##
## $occupation
##
##                 ?      Adm-clerical      Armed-Forces      Craft-repair
##              2809              5611                15              6112
##   Exec-managerial   Farming-fishing Handlers-cleaners Machine-op-inspct
##              6086              1490              2072              3022
```

[3] Vural/Yong Kai: it would be more organised if each table is presented in "tabular" forms instead of a list.

[4] Another approach would be classifcation by continents, such as Africa, America, Asia, and Europe

```
##     Other-service   Priv-house-serv    Prof-specialty   Protective-serv
##             4923               242              6172               983
##             Sales      Tech-support  Transport-moving
##              5504              1446              2355
##
## $relationship
##
##        Husband  Not-in-family Other-relative      Own-child      Unmarried
##          19716          12583           1506           7581           5125
##           Wife
##           2331
##
## $race
##
## Amer-Indian-Eskimo Asian-Pac-Islander              Black
##                470               1519               4685
##              Other              White
##                406              41762
##
## $sex
##
## Female   Male
##  16192  32650
##
## $native_country
##
##                          ?                   Cambodia
##                        857                         28
##                     Canada                      China
##                        182                        122
##                   Columbia                       Cuba
##                         85                        138
##         Dominican-Republic                    Ecuador
##                        103                         45
##                El-Salvador                    England
##                        155                        127
##                     France                    Germany
##                         38                        206
##                     Greece                  Guatemala
##                         49                         88
##                      Haiti         Holand-Netherlands
##                         75                          1
##                   Honduras                       Hong
##                         20                         30
##                    Hungary                      India
##                         19                        151
##                       Iran                    Ireland
##                         59                         37
##                      Italy                    Jamaica
##                        105                        106
##                      Japan                       Laos
##                         92                         23
##                     Mexico                  Nicaragua
##                        951                         49
```

```
## Outlying-US(Guam-USVI-etc)                         Peru
##                         23                           46
##                Philippines                       Poland
##                        295                           87
##                   Portugal                  Puerto-Rico
##                        67                          184
##                   Scotland                        South
##                        21                          115
##                     Taiwan                     Thailand
##                        65                           30
##           Trinadad&Tobago                United-States
##                        27                        43832
##                    Vietnam                   Yugoslavia
##                        86                           23
##
## $income
##
## <=50K  >50K
## 37155 11687
```

Note that we kept the original features intact by defining new features from them. Each newly defined feature ended with `1` to differentiate them from their respective original features. Such approach would allow us to evaluate the model performance with various granularity of the data.

```
adult <- adult %>%
  mutate( workclass1 = ifelse( workclass %in% c('?', 'Never-worked', 'Without-pay'), "Other",
                               ifelse( grepl('gov', workclass), 'gov',
                                       ifelse( grepl('Self-emp', workclass), 'Self-emp', workclass )) ),

          education1 = ifelse( grepl('th', education) | education == 'Preschool', 'Before-HS',
                               ifelse( education %in% c('Doctorate', 'Masters', 'Prof-school'), 'Postgra
                                       ifelse( grepl( 'Assoc', education ), 'Assoc', education ) ) ),

          native_country1 = ifelse( native_country != 'United-States', 'Not US', 'United-States'),
          marital_status1 = ifelse( grepl( 'Married', marital_status ), 'Married', marital_status),
          occupation1 = ifelse( occupation == '?', 'Other', occupation)
          )
```

Lastly, we converted all character features into factor and removed `fnlwght`.

```
adult[, sapply( adult, is.character )] <- lapply( adult[, sapply( adult, is.character )], factor)
adult$fnlwght <- NULL
```

Table 5 presents the summary statistics after data-preprocessing.

```
summarizeColumns(adult) %>% kable( caption = 'Feature Summary before Data Preprocessing' )
```

Table 5: Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
age	integer	0	38.64359	13.7105099	37	14.8260	17	90	0
workclass	factor	0	NA	0.3058024	NA	NA	10	33906	9
education	factor	0	NA	0.6768355	NA	NA	83	15784	16
education_num	integer	0	10.07809	2.5709728	10	1.4826	1	16	0
marital_status	factor	0	NA	0.5418083	NA	NA	37	22379	7
occupation	factor	0	NA	0.8736333	NA	NA	15	6172	15

name	type	na	mean	disp	median	mad	min	max	nlevs
relationship	factor	0	NA	0.5963310	NA	NA	1506	19716	6
race	factor	0	NA	0.1449572	NA	NA	406	41762	5
sex	factor	0	NA	0.3315180	NA	NA	16192	32650	2
hours_per_week	integer	0	40.42238	12.3914440	40	4.4478	1	99	0
native_country	factor	0	NA	0.1025757	NA	NA	1	43832	42
income	factor	0	NA	0.2392818	NA	NA	11687	37155	2
capital	integer	0	991.56531	7475.5499061	0	0.0000	-4356	99999	0
workclass1	factor	0	NA	0.3058024	NA	NA	2830	33906	4
education1	factor	0	NA	0.6768355	NA	NA	3662	15784	6
native_country1	factor	0	NA	0.1025757	NA	NA	5010	43832	2
marital_status1	factor	0	NA	0.5281929	NA	NA	1518	23044	5
occupation1	factor	0	NA	0.8736333	NA	NA	15	6172	15

4 Data Exploration[5]

We explored the data for each feature individually and split them by the classes of target features. Then we proceeded to multivariate visualisation.

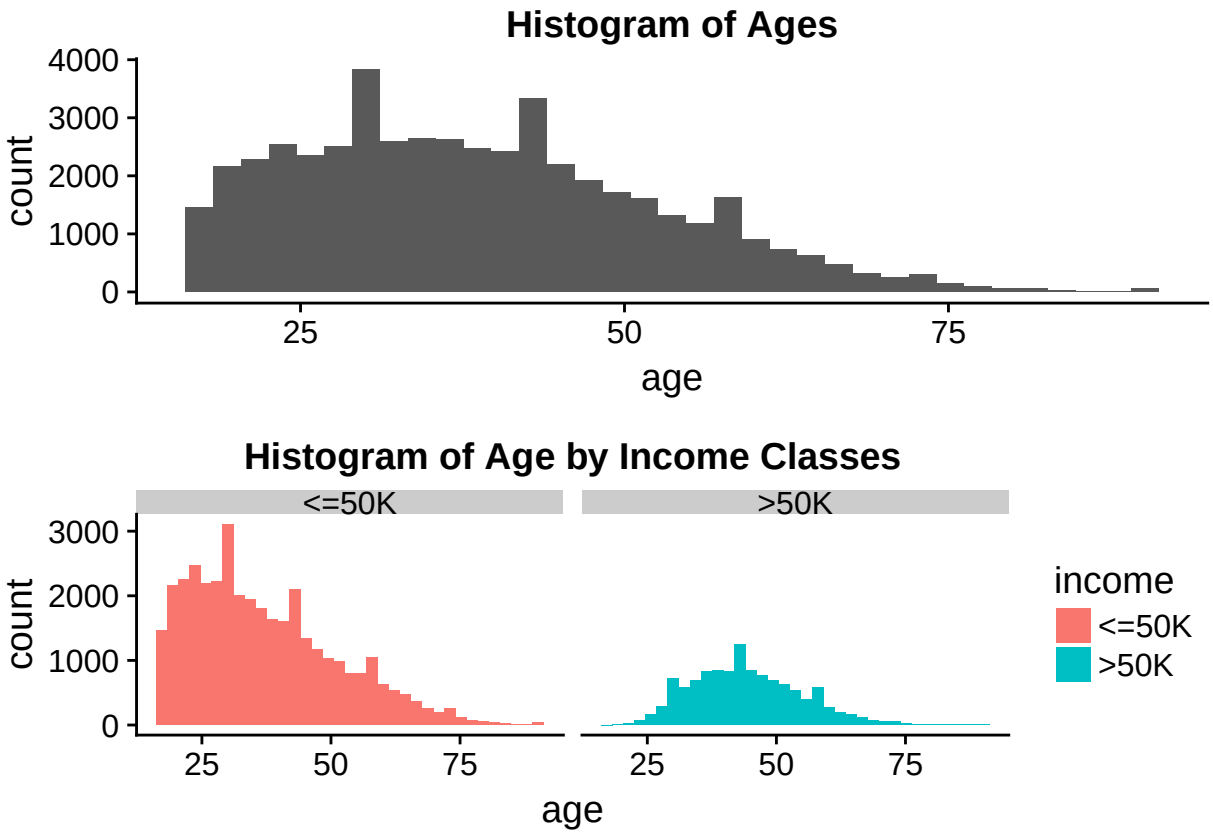
4.1 Univariate Visualisation

4.1.1 Numerical Features

4.1.1.1 Age

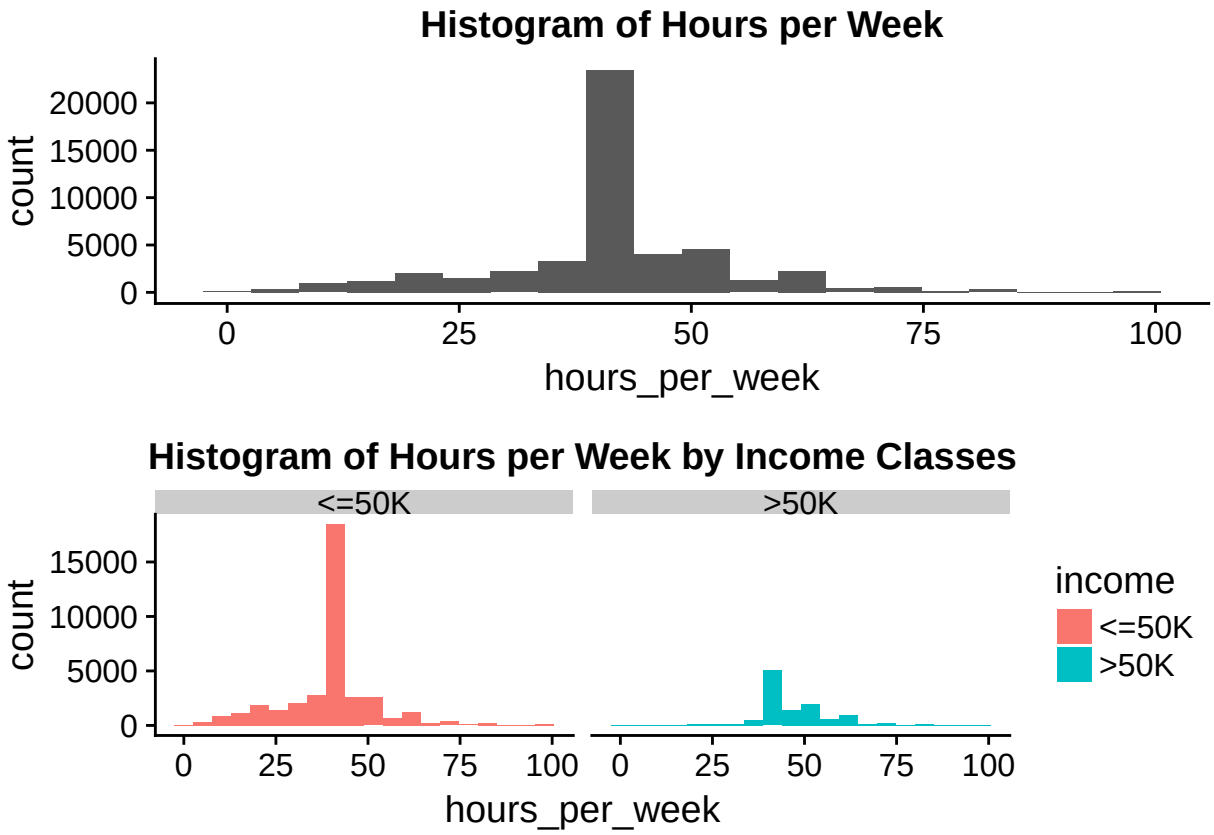
Most of individuals aged between 25 and 50 with younger generation tended to earn less than $50,0000. The higher-income individuals' ages appeared to following a normal distribution whereas the lower income group had a skewed distribution of age. Therefore, age would be a predictive feature.

[5] Yong Kai: Visualisation codes are hidden in the pdf version. See RMD version for codes.



4.1.1.2 Hours per week

On aggregated level, individuals worked for approximately 40 hours per week, as indicated by the sharp kurtosis in the the histogram of `hours_per_week`. When we segregated by income classes, the kurtosis did not change much although the lower income group had fatter tails in the distribution of hours per week. The change in the shape could be explained by the workclass or occupations.

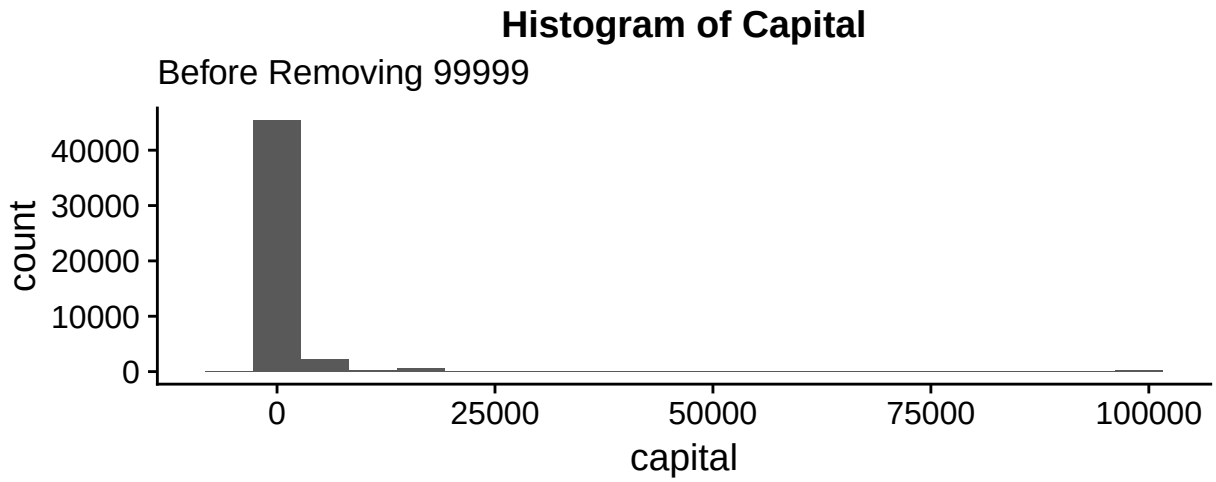


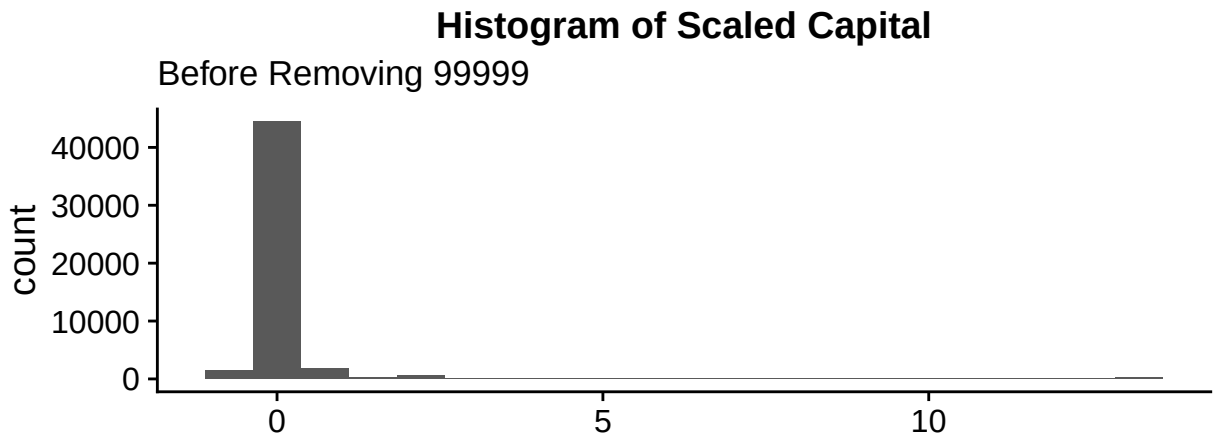
4.1.1.3 Education Num

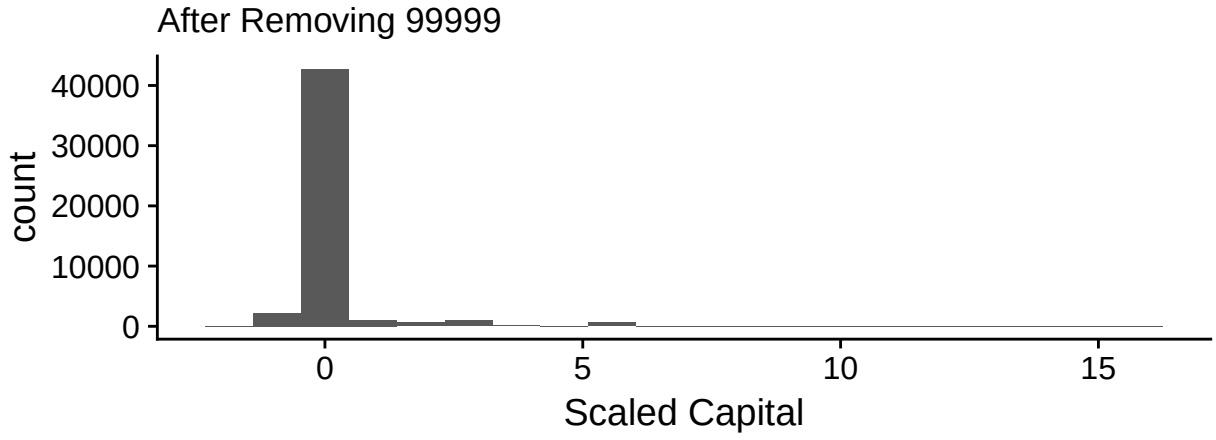
We did not display any visualisation for `education_num` as we shall show that it carried the same information of `education` in the later section.

4.1.1.4 Capital

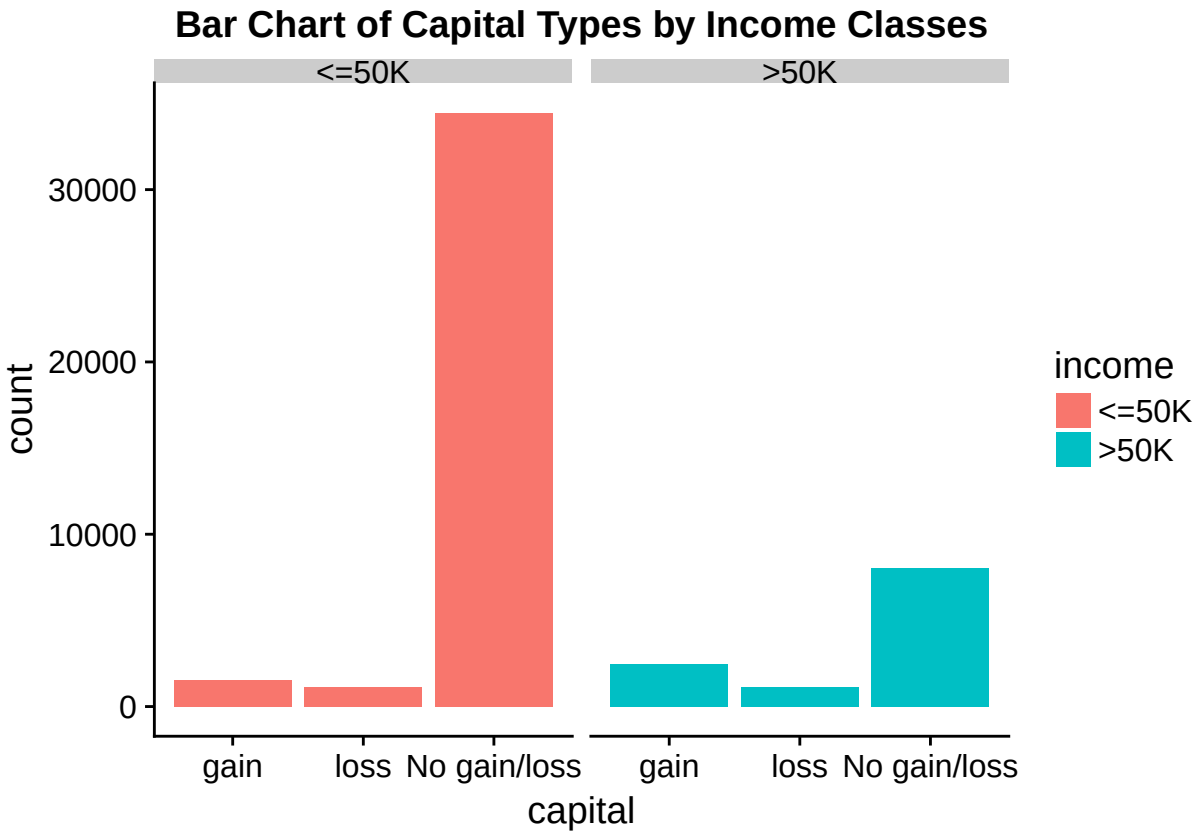
The histogram of `capital` was heavily skewed. The problem persisted even after scaling this variable or removing individuals with 99999 values.


Histogram of Capital
Before Removing 99999
count
40000
30000
20000
10000
0
0
25000
50000
75000
100000
capital


Histogram of Scaled Capital
Before Removing 99999
count
40000
30000
20000
10000
0
0
5
10


After Removing 99999
count
40000
30000
20000
10000
0
0
5
10
15
Scaled Capital

Due to skewness, we explored capital by binning it into three groups: “No gain/loss”, “loss”, and “gain” for both income classes. Although the higher income earners had more capital gains on a tiny margin, most of taxpayers had no gain or loss. Univariately, we suspected this feature would have little predictive power.

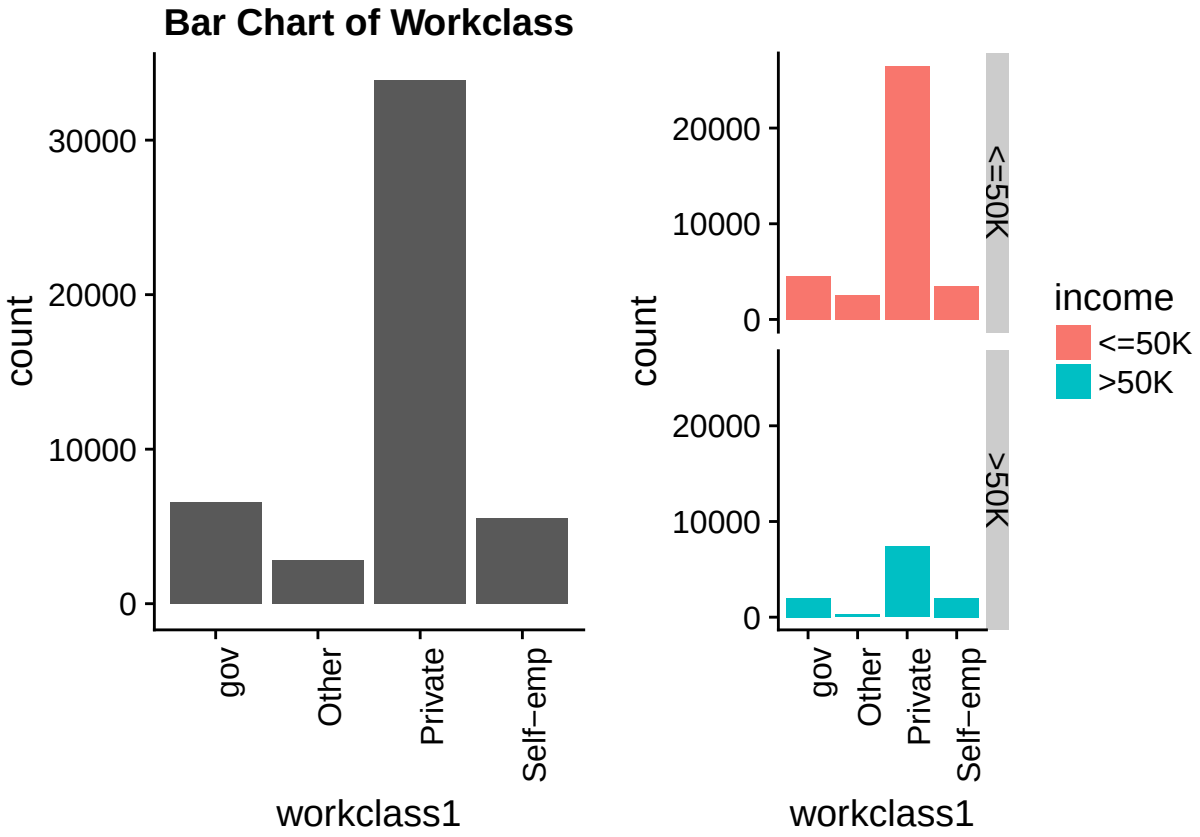


4.1.2 Categorical Features

For categorical features, we shall present the newly defined categorical features.

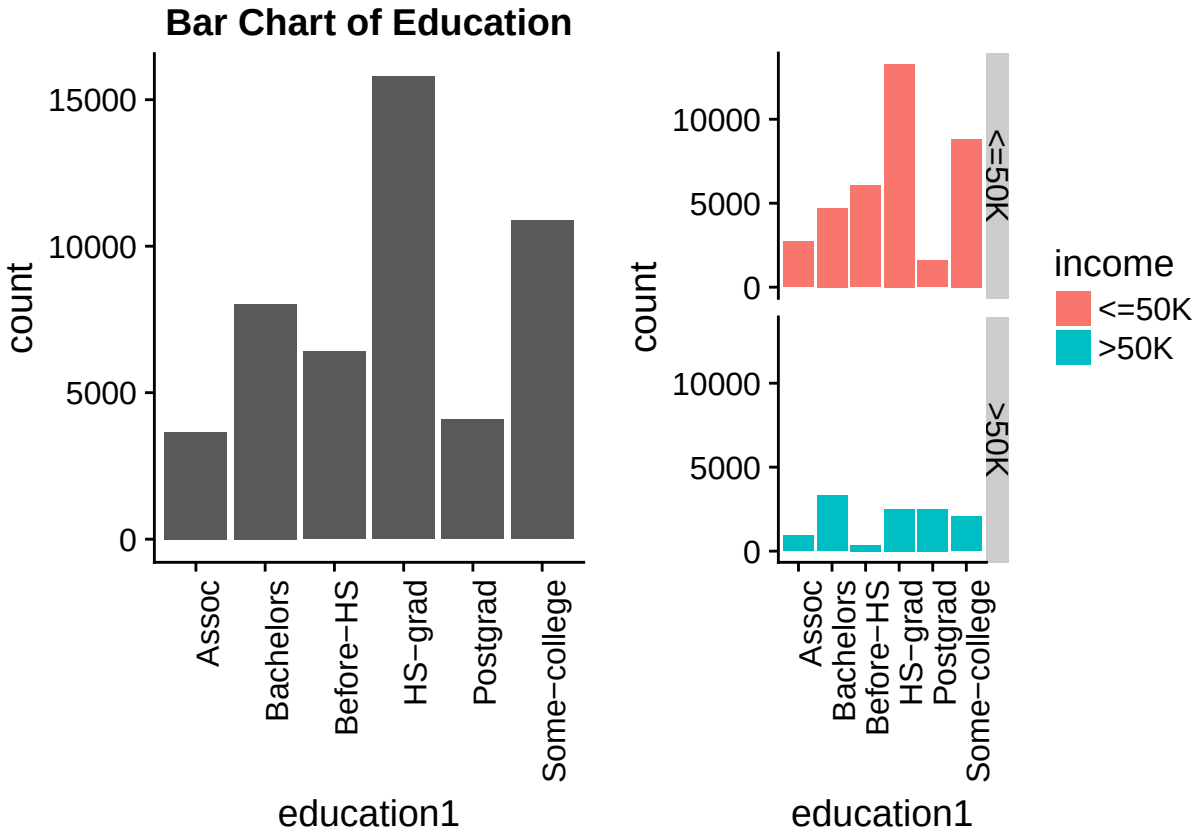
4.1.2.1 Work Class

Individuals worked mostly in private sector. There was no clear distinction between two income classes.



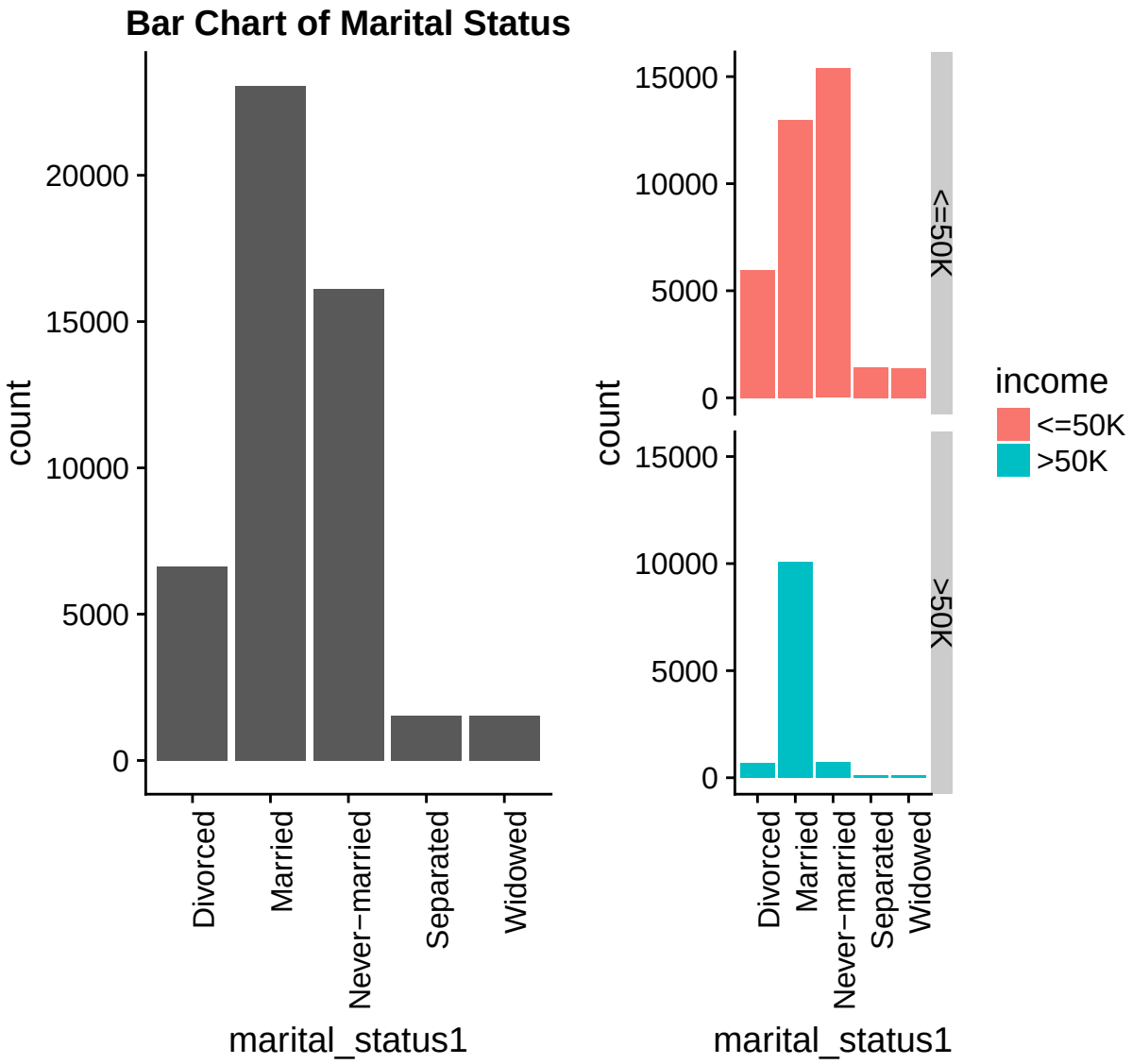
4.1.2.2 Education

Segregating education by income classess showed that higher income earners were at least bachelor degree holders whereas the lower income groups were mostly graduated from high schools.



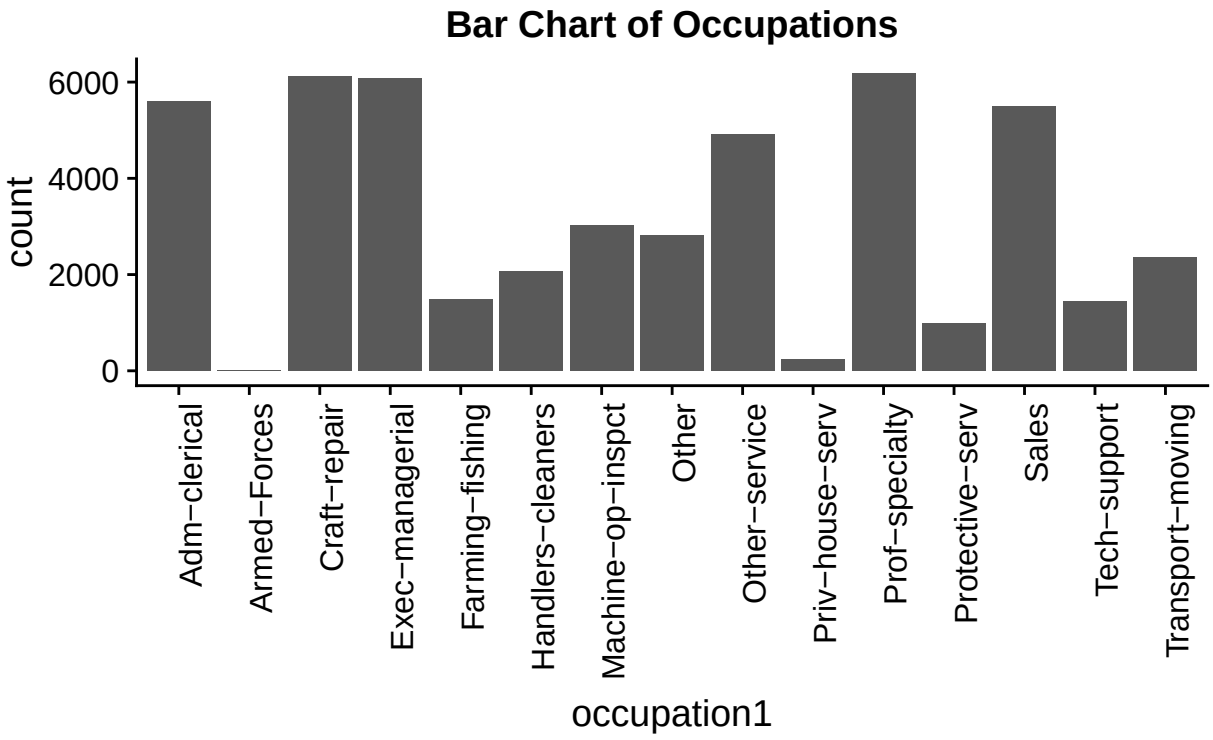
4.1.2.3 Marital Status

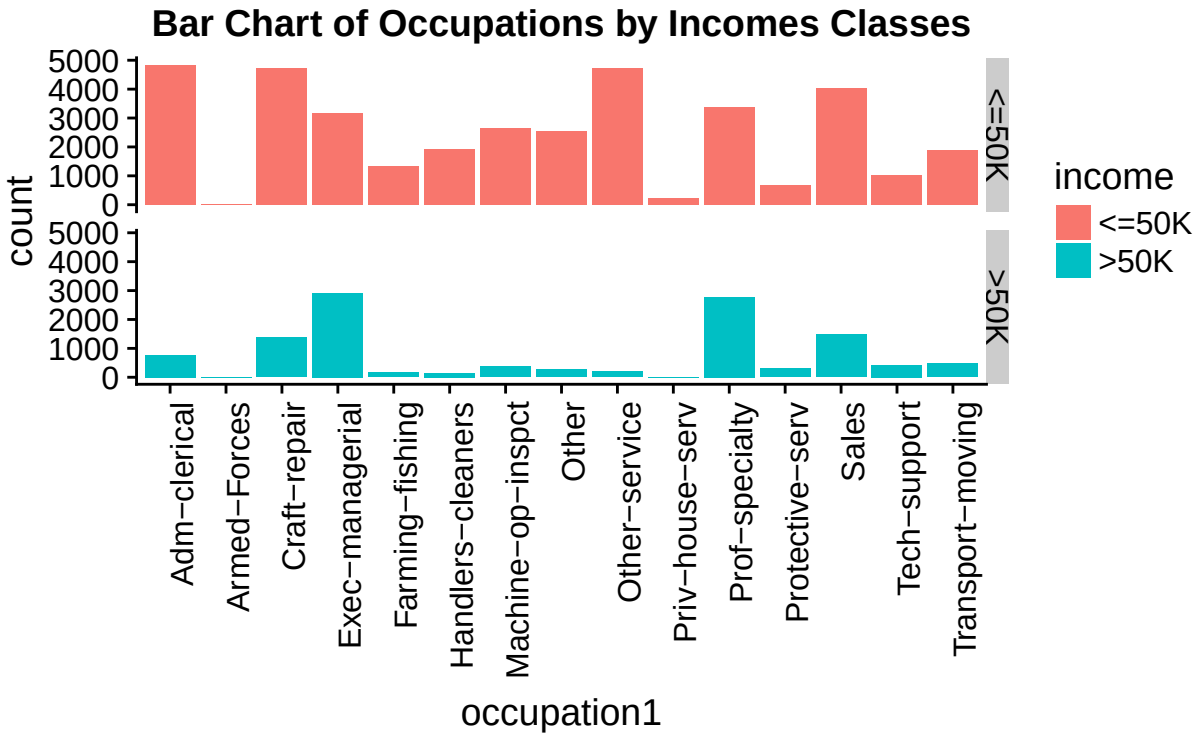
Most taxpayers were married; however, number of never married individuals was signficantly lower in the lower income class. If, *on average* income level was proportional to years of working experience, it could imply never married individuals might be younger and had started their career.



4.1.2.4 Occupation

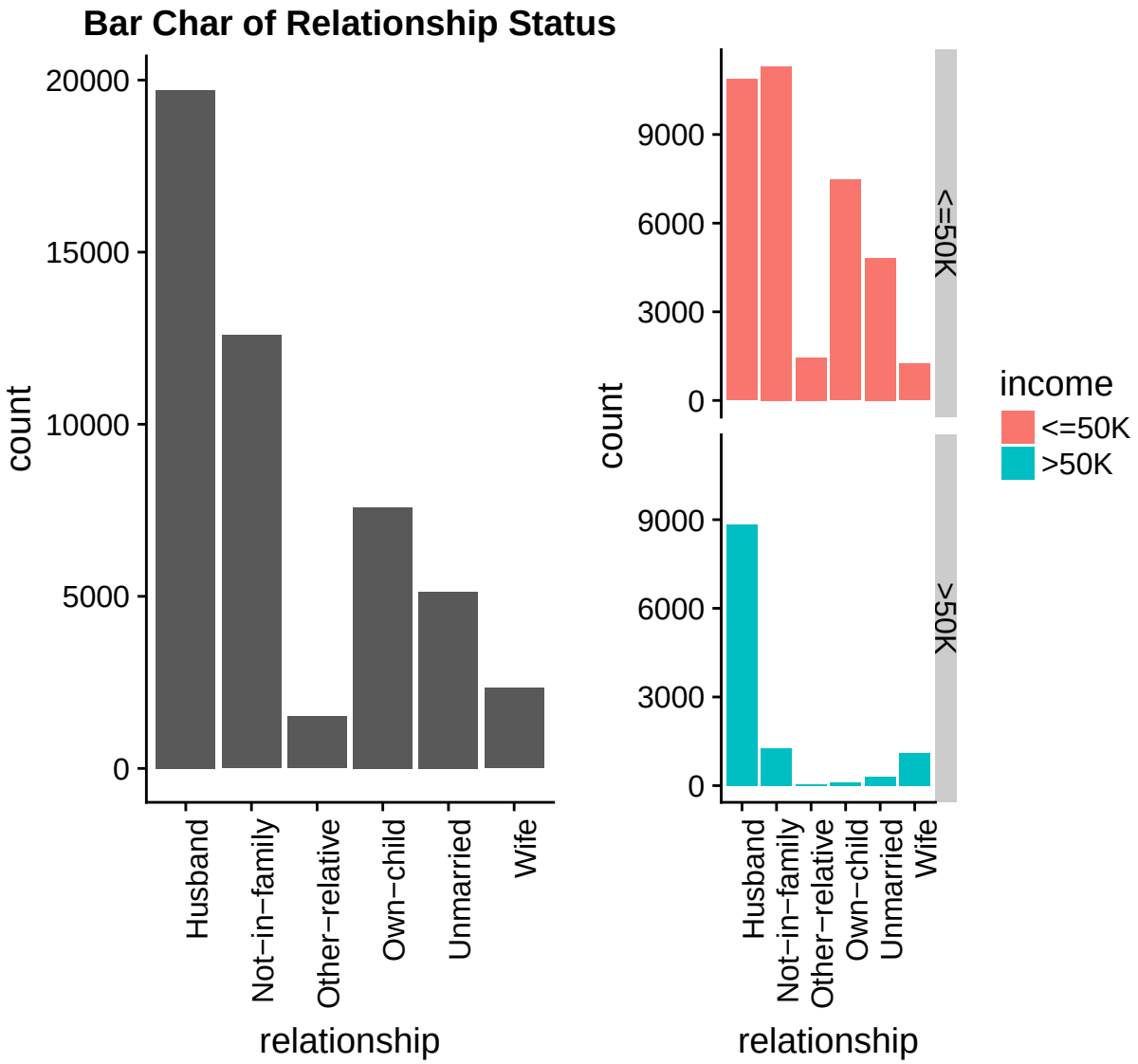
Occupation might be a predictive feature as higher income earners tended to be white-collar workers such as executive/managerial positions and professionals.





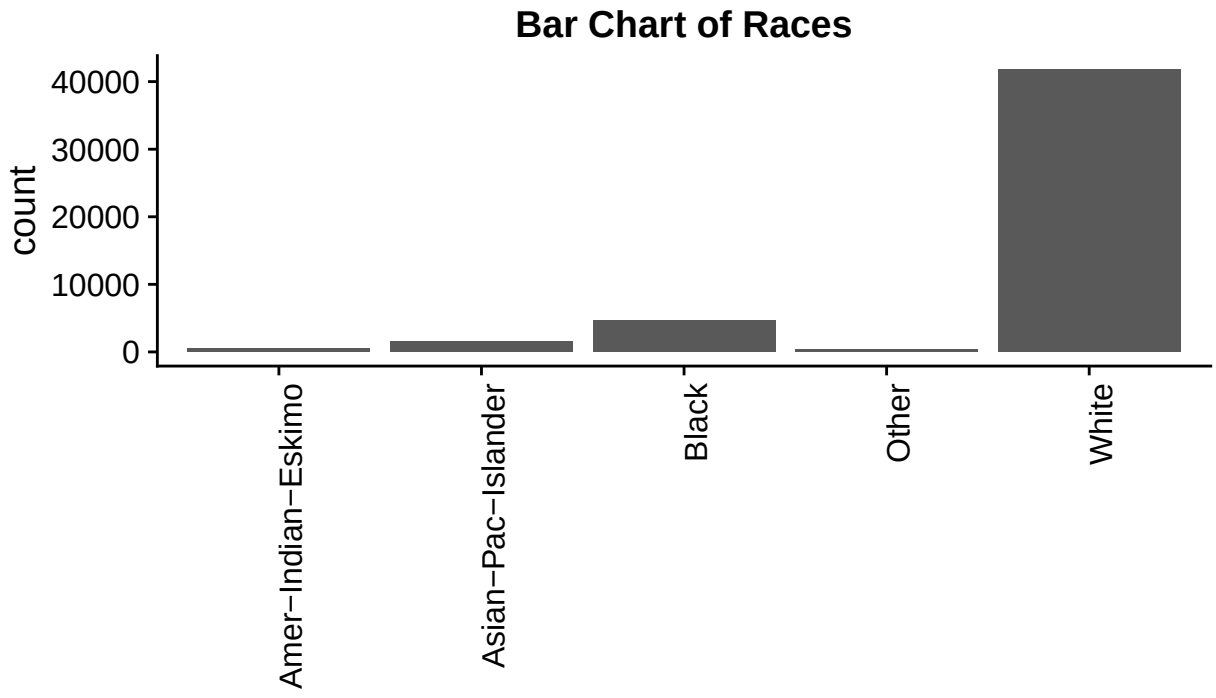
4.1.2.5 Relationship

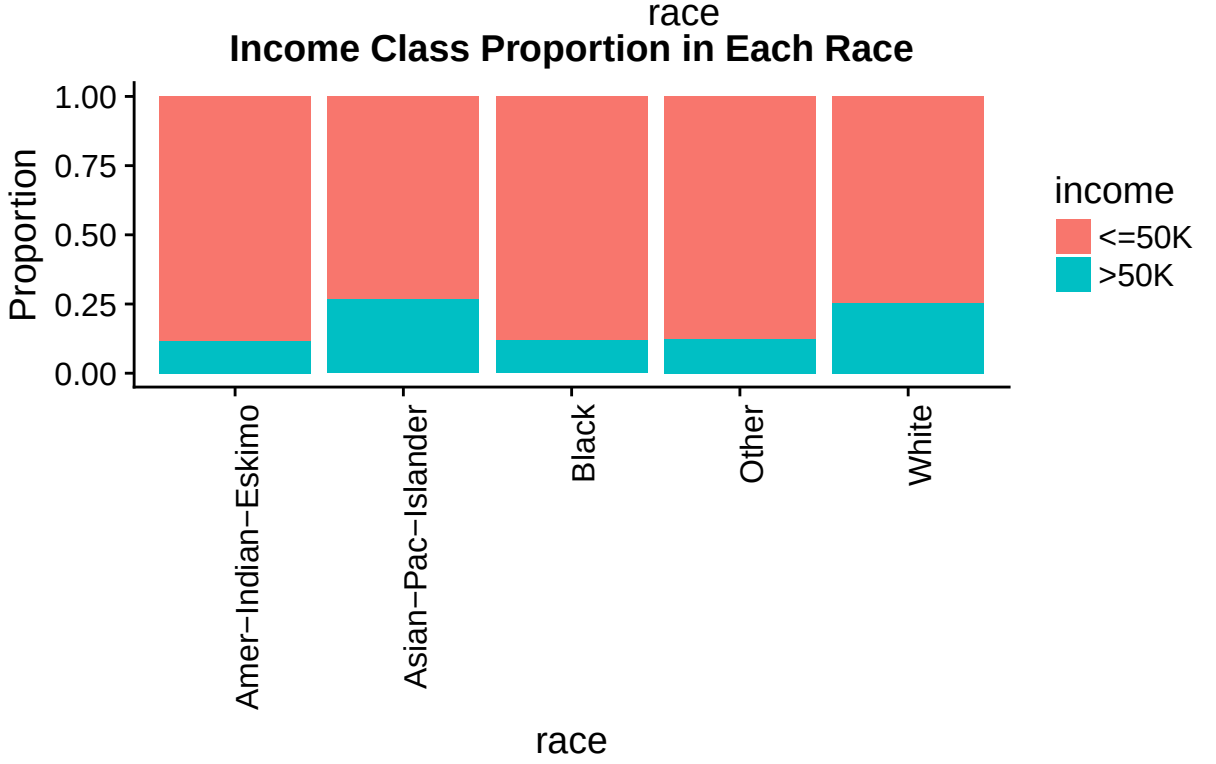
Most of individuals were husbands, regardless income classes. Consistent with marital status, unmarried and not-in-family individuals tended to earn less than $50,000. Unfortunately, a significant number of individuals who had children were from the lower income class.



4.1.2.6 Race

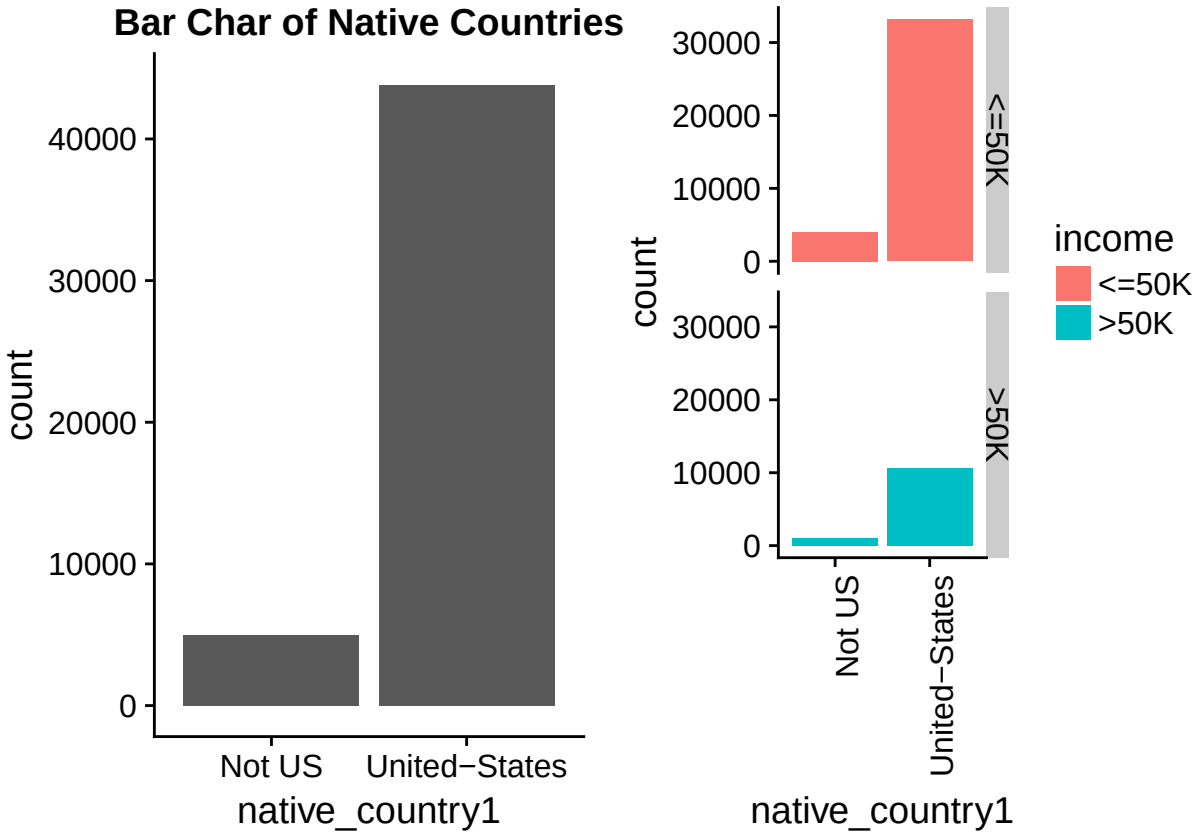
Consistent with the US demographic, the white were dominant ethnic followed by the black Americans. White and asian individuals had relatively higher proportion of high-income groups than other races.





4.1.2.7 Native Country

Most of individuals were born in the United States. There was no clear difference in the income classes.

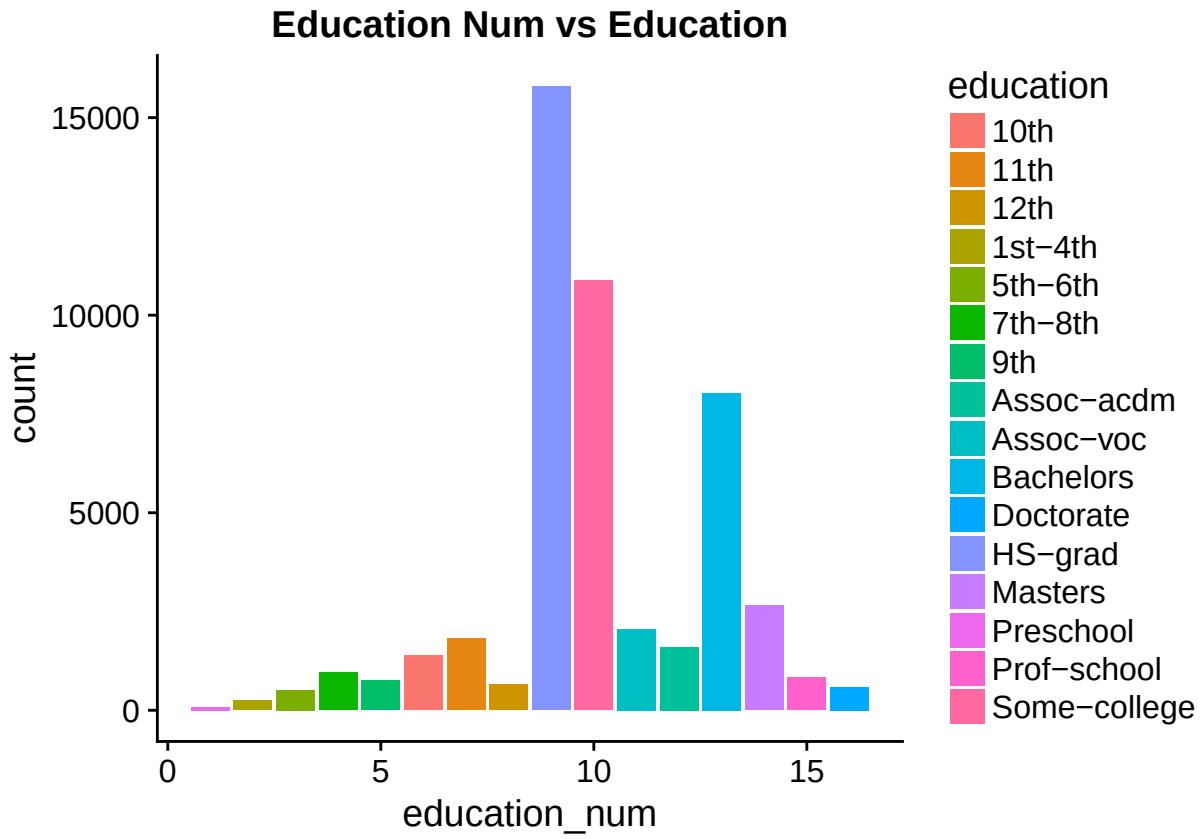


4.2 Multivariate Visualisation

4.2.1 Education Num vs Education

The following visual confirms that `education_num` was a code label for `education` and hence the former was redundant[6].

[6] Another approach is to compute the contingency table between these features and verify only non-diagonal entries are zeroes.



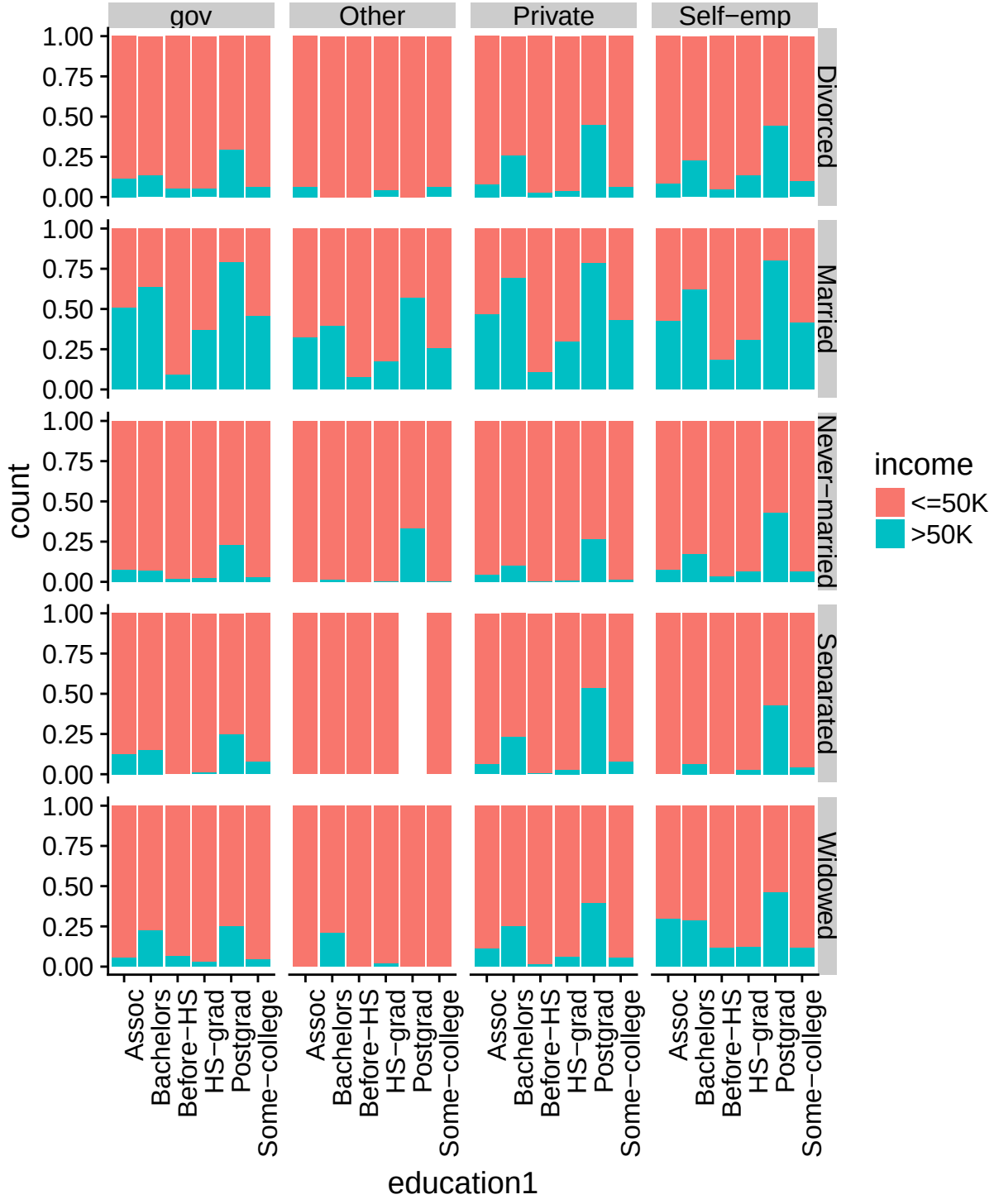
Therefore, we removed it.

```
adult$education_num <- NULL
```

4.2.2 Education, Marital Status, and Workclass

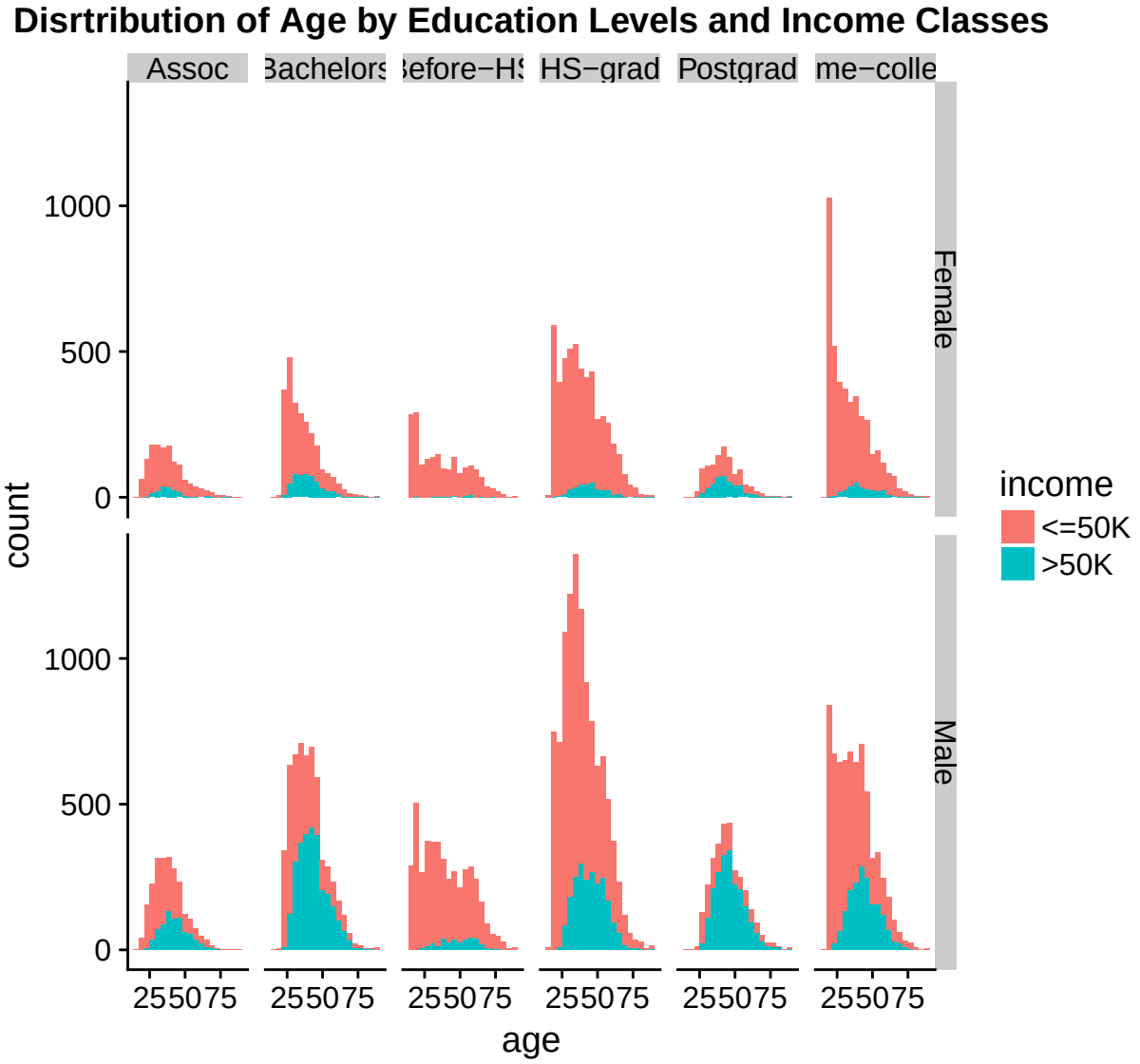
The following visual depicts most of indivuals were married, never married, or divorced working in private sectors. In particular, married individuals were high income earners in all workclasses. The education levels between income classes varied across different marital statuses. Overall, post-graduate degree holders stood out as the highest proportion in most of the combinations between marital statuses and workclasses. Note that for married individuals, the proportion by education levels were similar in each work class. That is, the higher income earners were mostly post-graduate holders, followed by bachelors, associates, some colleges, high school graduates, and lastly those did not complete high schools.

Proportional Bar Chart: Education, Marital Status, and Workclass



4.2.3 Education, Income Classes, and Age

The *stacked* histograms reveals that, regardless gender, the low income class had a positive skewed age distribution in all education levels, except post-graduate degree holders. The post graduate holders had a symmetric age distribution. Also, relatively higher proportion of female post-graduate holder earned less than $50,000 compared to their male peers.



5 Summary

For numerical features, we combined `capital_loss` and `capital_gain` as `capital`; however, we did not remove observations with a value of 99999 which could be a valid value too. For categorical features, we defined some new features which binned their corresponding original features into lower cardinalities. The binning process also imputed missing values encoded as `?`. Except `education_num` and `fnlwght`, we did not remove the original features since we would like to experiment the model building by changing the granularity of the data. From the data exploration, we found that education levels, workclasses, gender, ages, and marital statuses were potentially useful features in predicting the income classes.