MATH 9102 - Probability and Statistical Inference Assignment
Final Assignment

Antonio Silva (D23129331@mytudublin.ie)

2024-05-26

Contents

Method

Participants

The dataset[1] for this study originates from the 1994 U.S. Census, which involved a diverse range of participants to provide a comprehensive snapshot of the U.S. population. This extensive coverage makes the findings relevant to the adult population of the United States during that period.

The data captures key demographic details critical for examining income disparities, including age, gender, race/ethnicity, education level, marital status, and employment status. Specifically, it categorizes gender into male or female; race/ethnicity includes White, Black, Asian, Hispanic, among others; and education levels vary from none to advanced degrees.

The census participants were not specifically recruited like in targeted research studies but were included through a national initiative aimed at documenting the demographic attributes of the U.S. population. Participation was solicited from households nationwide, covering both urban and rural areas to ensure a representative sample.

Although the census gathered data from millions, this study analyzes a subset of 32,561 complete records, chosen for their comprehensive demographic details needed to study income variations. No power analysis was conducted as the census was designed to encompass as broad a population segment as possible, not to meet specific sample size calculations for statistical power.

The data was gathered specifically to analyze the earned income of the population, and the collection was conducted according to the following criteria:

AAGE $>$ 16 Keep only working-age adults;

AGI $>$ 100 Adjusted Gross Income;

AFNLWGT $>$ 1 Final Weight meaning the number of people that is believed that this entry represents;

HRSWK $>$ 0 Participants with working hours reported.

The dataset comes in comma-separated values (CSV) format has *14 features* and has 0.9% missing values. The target variable is the *income* that can be either *more* or *less or equal* than $50000.

Due to the census's scope, there was no allocation of participants into experimental groups or need for randomization typical of controlled experiments. The dataset naturally includes a broad demographic diversity reflective of the national census effort.

Procedure

Hypothesis

As an initial analysis, we aim to explore how social factors such as gender and race affect income levels. We want to understand if a specific gender is more likely to earn more.

We also intend to investigate how other factors interact with gender. One interesting analysis could be to examine how ethnic groups might experience advantages or disadvantages combined with gender in terms of income disparity.

Another intriguing hypothesis would consider the income progression influenced by combinations of age and gender.

Exploration

Features Description

First, we need to conduct a thorough exploration of the dataset to understand the underlying structure and quality of the data. This will involve assessing various features for their relevance and potential impact on the study's outcomes, particularly how they relate to income levels. We will also present a summary in Table X.

```r
census <- read.csv("../adult.csv", na.strings =  c("?"))
head(census)
```

```
##   age workclass fnlwgt    education education.num marital.status
## 1  90      <NA>  77053      HS-grad             9        Widowed
## 2  82   Private 132870      HS-grad             9        Widowed
## 3  66      <NA> 186061 Some-college            10        Widowed
## 4  54   Private 140359      7th-8th             4       Divorced
## 5  41   Private 264663 Some-college            10      Separated
## 6  34   Private 216864      HS-grad             9       Divorced
##          occupation  relationship  race    sex capital.gain capital.loss
## 1              <NA> Not-in-family White Female            0         4356
## 2   Exec-managerial Not-in-family White Female            0         4356
## 3              <NA>     Unmarried Black Female            0         4356
## 4 Machine-op-inspct     Unmarried White Female            0         3900
## 5    Prof-specialty     Own-child White Female            0         3900
## 6     Other-service     Unmarried White Female            0         3770
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             18  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40  United-States  <=50K
## 6             45  United-States  <=50K
```

```r
missing_values <- census %>% summarise(across(everything(), ~sum(is.na(.))))
missing_values_ft <- missing_values %>%
  pivot_longer(cols = everything(), names_to = "Feature", values_to = "MissingValues")

ggplot(missing_values_ft, aes(x = Feature, y = MissingValues, fill = Feature)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Missing Values Per Feature", x = "Feature", y = "Missing Values") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none") +
  theme(plot.title = element_text(hjust = 0.5))
```

We get the following resume for the categories:

Scoring (Optional)

Appendix

Statistics Description

So do describe the variables we will

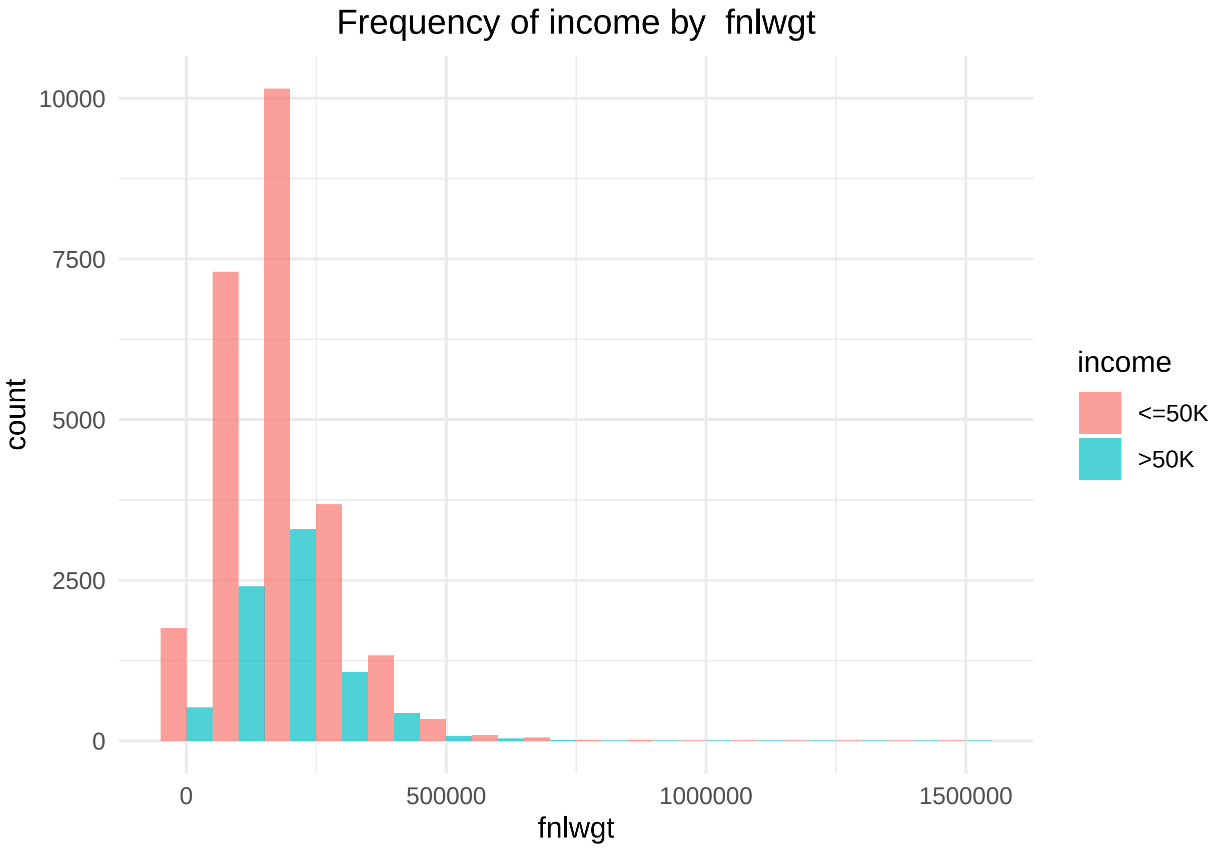
```
feat_desc_num <- function(variable_name, binwidth = 5) {
  summary_stats <- census %>%
    group_by(income) %>%
    summarise(count = n(),
              mean = mean(!!sym(variable_name), na.rm = TRUE),
              median = median(age, na.rm = TRUE),
              min = min(!!sym(variable_name)),
              max = max(!!sym(variable_name), na.rm = TRUE),
              kurtosis = kurtosis(!!sym(variable_name), na.rm = TRUE),
              skewness = skewness(!!sym(variable_name), na.rm = TRUE),
              .groups = "drop") %>%
    ungroup()

  g <- ggplot(census, aes(x = !!sym(variable_name), fill = income)) +
      geom_histogram(binwidth = binwidth, position = "dodge", alpha = 0.7) +
      theme_minimal() +
      labs(title = paste("Frequency of income by ", variable_name)) +
      theme(plot.title = element_text(hjust = 0.5))

  grid.draw(g)
  kable(summary_stats, "latex")
}

feat_desc_cat <- function(variable_name, render_table = TRUE) {
  summary_stats <- census %>%
    group_by(!!sym(variable_name), income) %>%
    summarise(count = n(), .groups = "drop") %>%
    mutate(proportion = count / sum(count)) %>%
    ungroup()  %>%
    mutate(percent = scales::percent(proportion, 1))

  g <- ggplot(summary_stats, aes(x = !!sym(variable_name), y = proportion, fill = income)) +
    geom_bar(stat = "identity", position = position_dodge()) +
    theme_minimal() +
    labs(title = paste("Proportion of income by ", variable_name)) +
    theme(plot.title = element_text(hjust = 0.5)) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

  grid.draw(g)
  if (render_table) {
    kable(summary_stats, "latex")
  }
}
```

fnlwgt

The majority of individuals have a fnlwgt (final weight) value below 200,000, with a significant drop-off beyond this range. Most individuals with higher fnlwgt values earn <=50K.

The histogram is heavily right-skewed, indicating that most individuals have lower fnlwgt values.

```r
feat_desc_num('fnlwgt', 100000)
```



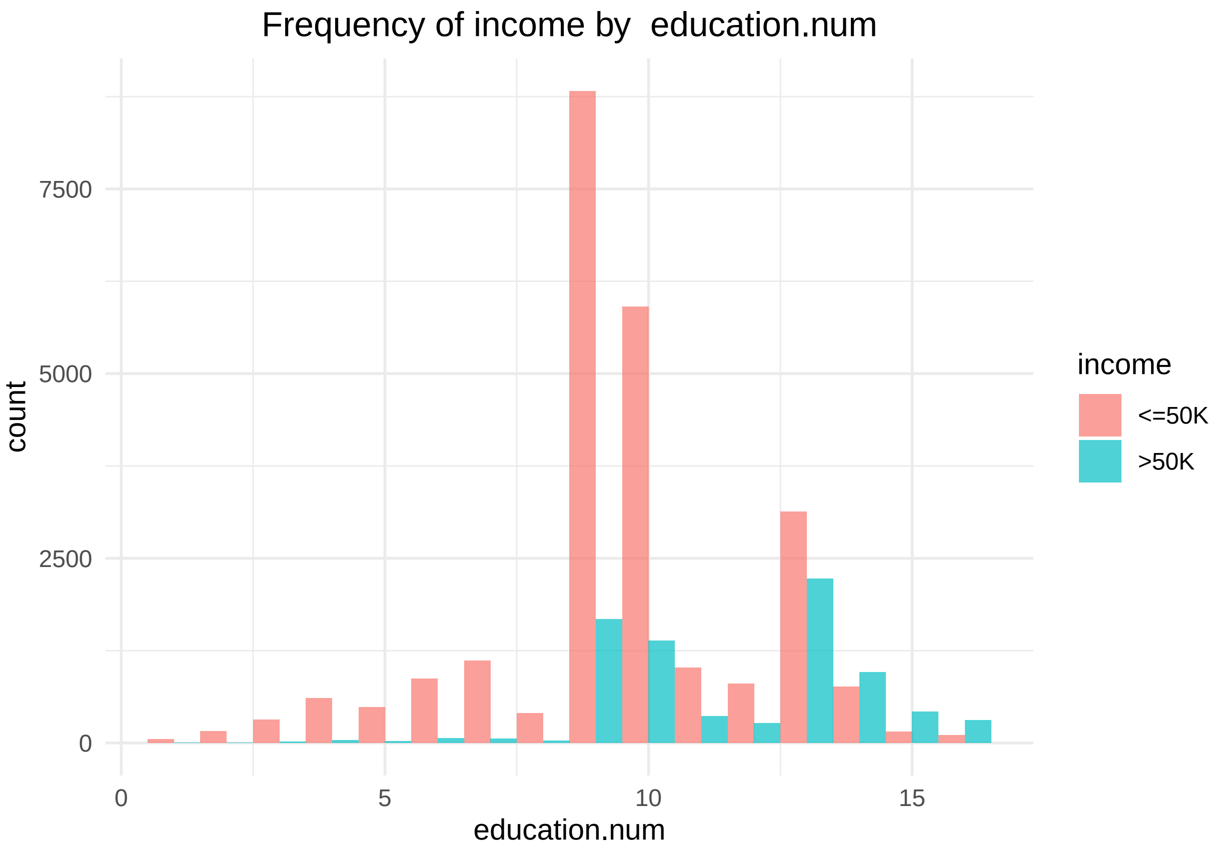
income	count	mean	median	min	max	kurtosis	skewness
<=50K	24720	190340.9	34	12285	1484705	9.495538	1.461008
>50K	7841	188005.0	44	14878	1226583	8.144175	1.391488

education.num

education.num, represents the number of years of education. Observations:

- Majority of individuals in the dataset havevaround 9-10 years of education;
- Individuals with fewer years of education (1-9 years) tend to earn <=50K;
- Individuals with more than 10 years of education, there is a noticeable increase in the proportion of those earning >50K.

```r
feat_desc_num('education.num', 1)
```

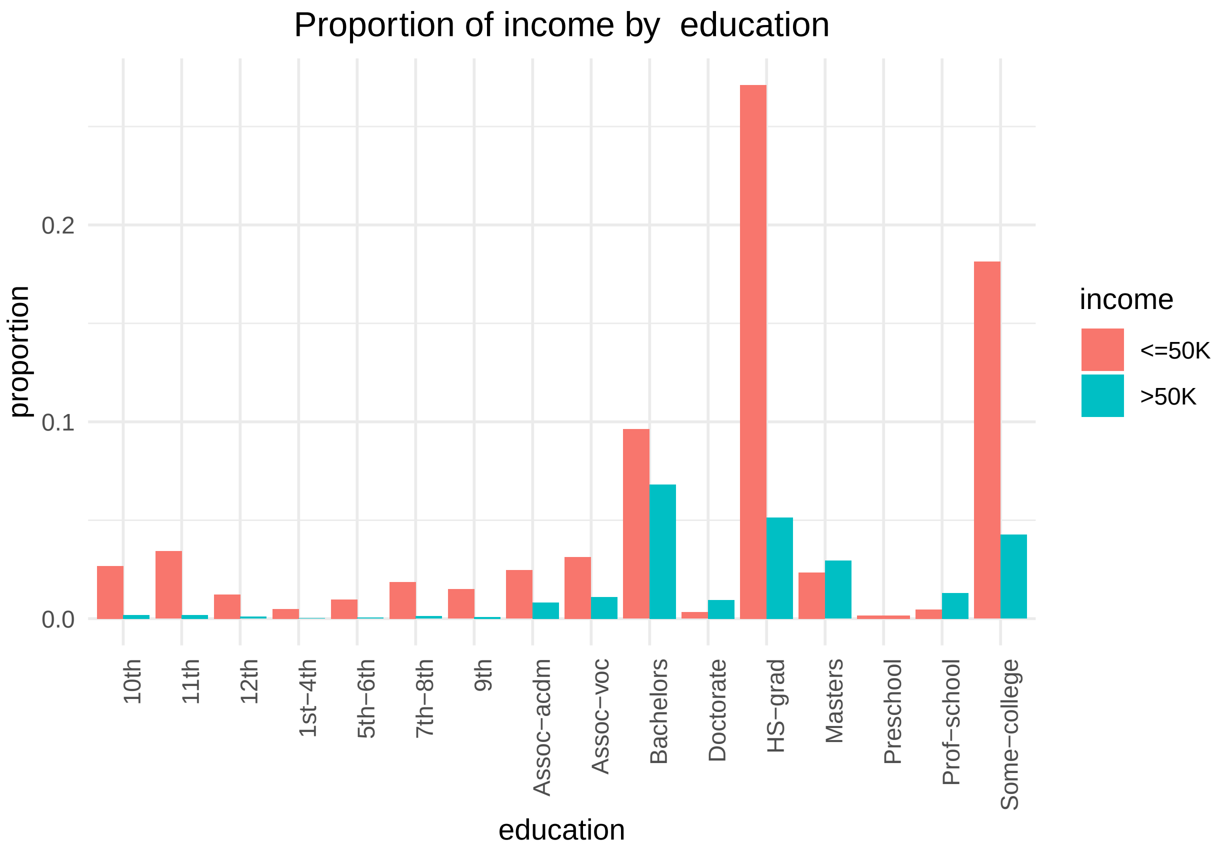


income	count	mean	median	min	max	kurtosis	skewness
<=50K	24720	9.595065	34	1	16	3.995261	-0.4349842
>50K	7841	11.611657	44	2	16	2.819701	-0.3175003

education

There is a positive correlation between higher education levels and the likelihood of earning >50K.

```
feat_desc_cat('education')
```


Proportion of income by education
proportion
0.0
0.1
0.2
education
10th
11th
12th
1st−4th
5th−6th
7th−8th
9th
Assoc−acdm
Assoc−voc
Bachelors
Doctorate
HS−grad
Masters
Preschool
Prof−school
Some−college
income
<=50K
>50K

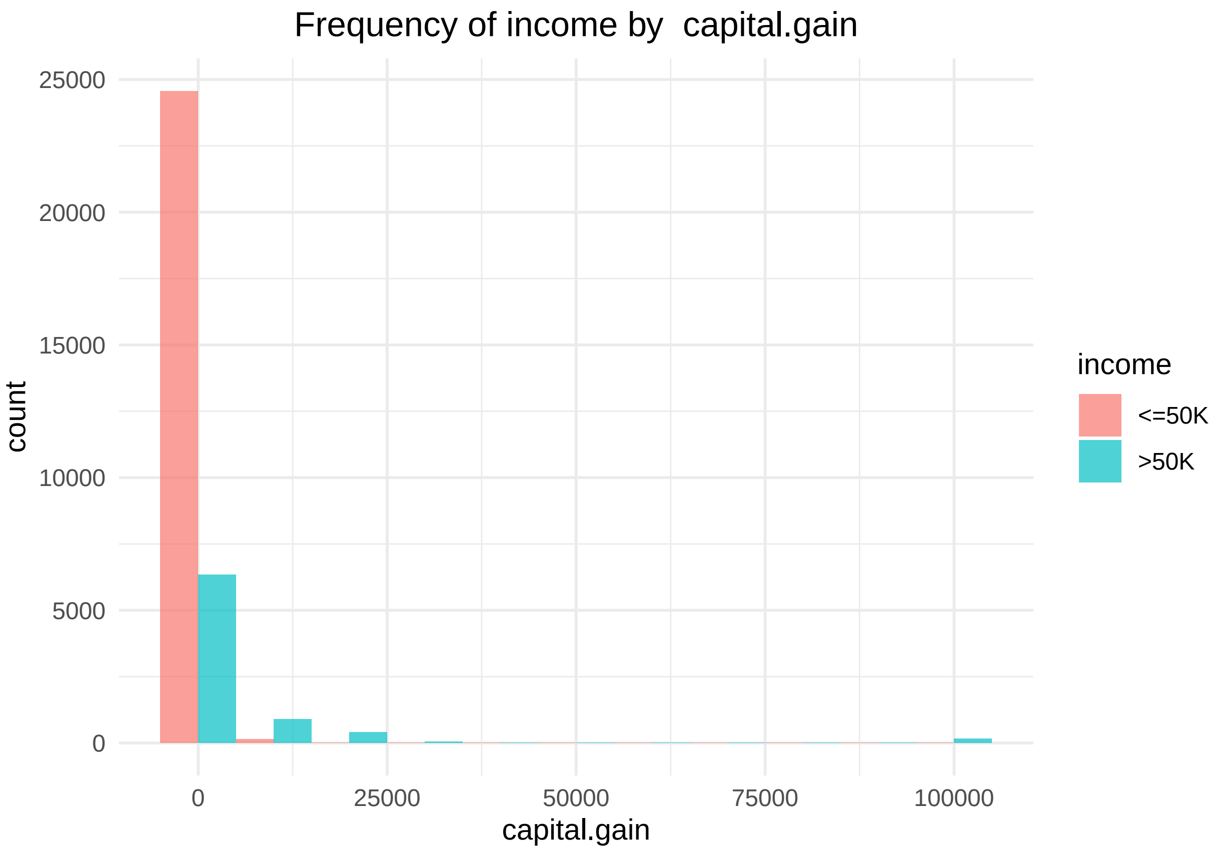
education	income	count	proportion	percent
10th	<=50K	871	0.0267498	3%
10th	>50K	62	0.0019041	0%
11th	<=50K	1115	0.0342434	3%
11th	>50K	60	0.0018427	0%
12th	<=50K	400	0.0122846	1%
12th	>50K	33	0.0010135	0%
1st-4th	<=50K	162	0.0049753	0%
1st-4th	>50K	6	0.0001843	0%
5th-6th	<=50K	317	0.0097356	1%
5th-6th	>50K	16	0.0004914	0%
7th-8th	<=50K	606	0.0186112	2%
7th-8th	>50K	40	0.0012285	0%
9th	<=50K	487	0.0149565	1%
9th	>50K	27	0.0008292	0%
Assoc-acdm	<=50K	802	0.0246307	2%
Assoc-acdm	>50K	265	0.0081386	1%
Assoc-voc	<=50K	1021	0.0313565	3%
Assoc-voc	>50K	361	0.0110869	1%
Bachelors	<=50K	3134	0.0962501	10%
Bachelors	>50K	2221	0.0682104	7%
Doctorate	<=50K	107	0.0032861	0%
Doctorate	>50K	306	0.0093977	1%
HS-grad	<=50K	8826	0.2710605	27%
HS-grad	>50K	1675	0.0514419	5%
Masters	<=50K	764	0.0234637	2%
Masters	>50K	959	0.0294524	3%
Preschool	<=50K	51	0.0015663	0%
Prof-school	<=50K	153	0.0046989	0%
Prof-school	>50K	423	0.0129910	1%
Some-college	<=50K	5904	0.1813212	18%
Some-college	>50K	1387	0.0425970	4%

capital.gain

The majority of individuals have a capital gain of 0. This is evident from the tall bar at the 0 mark for both income categories.

Among individuals with non-zero capital gains, those earning >50K are more prevalent. This indicates that individuals with higher capital gains are more likely to earn >50K.

```
feat_desc_num('capital.gain', 10000)
```

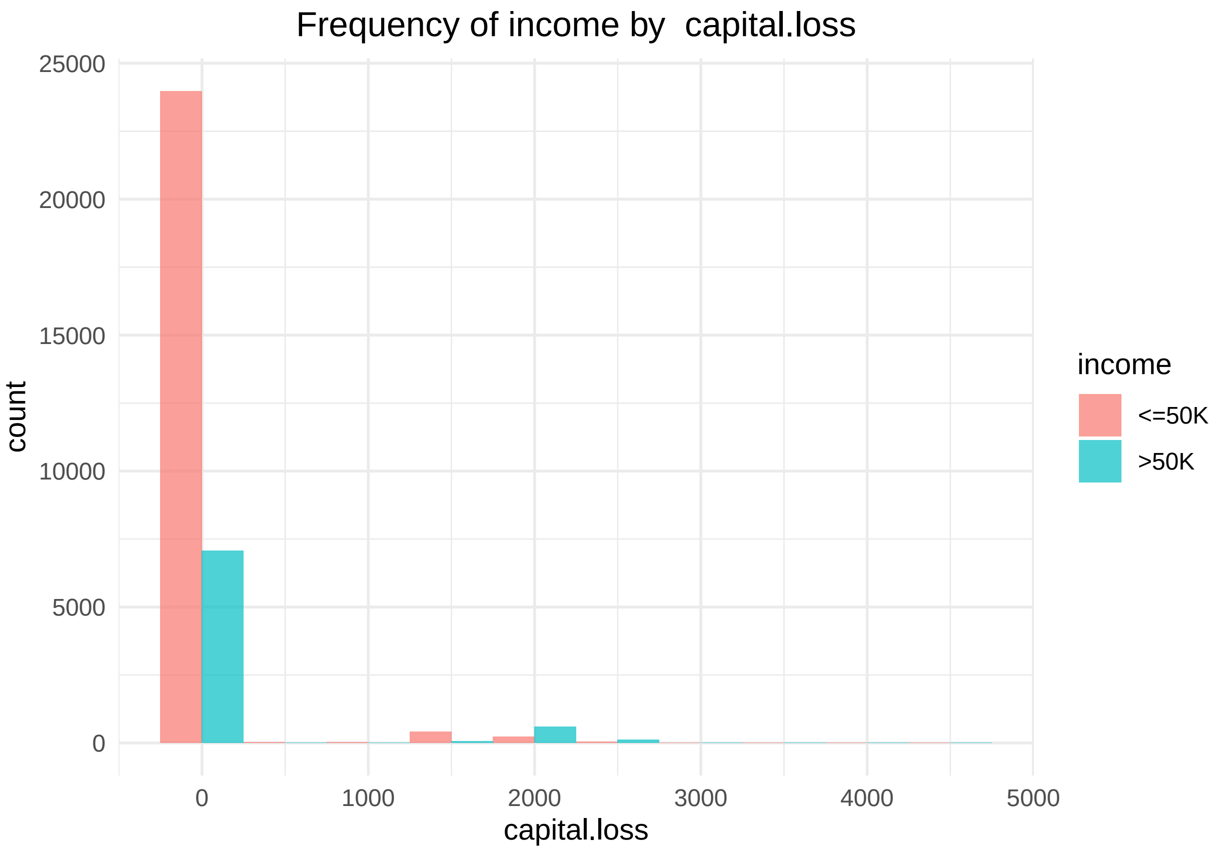


income	count	mean	median	min	max	kurtosis	skewness
<=50K	24720	148.7525	34	0	41310	608.1919	18.339804
>50K	7841	4006.1425	44	0	99999	38.2824	5.840607

capital.loss

Capital losses are generally low and infrequent. They are more common among higher earners when they do occur. This may reflect the financial activities and investment behaviors of higher-income individuals who might have more complex financial portfolios that include both gains and losses.

```
feat_desc_num('capital.loss', 500)
```

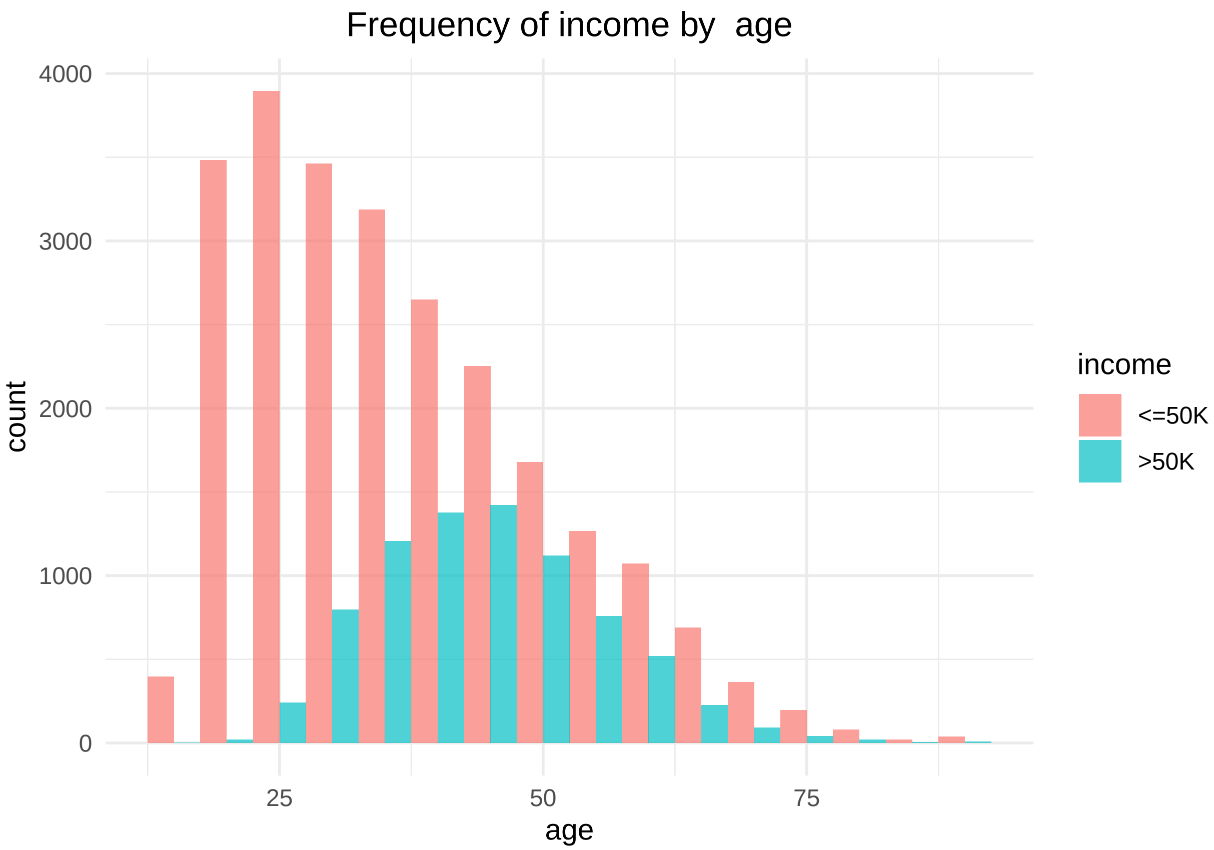


income	count	mean	median	min	max	kurtosis	skewness
<=50K	24720	53.14292	34	0	4356	41.214506	6.059831
>50K	7841	195.00153	44	0	3683	9.103034	2.797187

age

As a person get older your income tends to grow. This has to do with the career progression. In older age groups this number the earning decline, possibly because of retirement.

```
feat_desc_num('age')
```

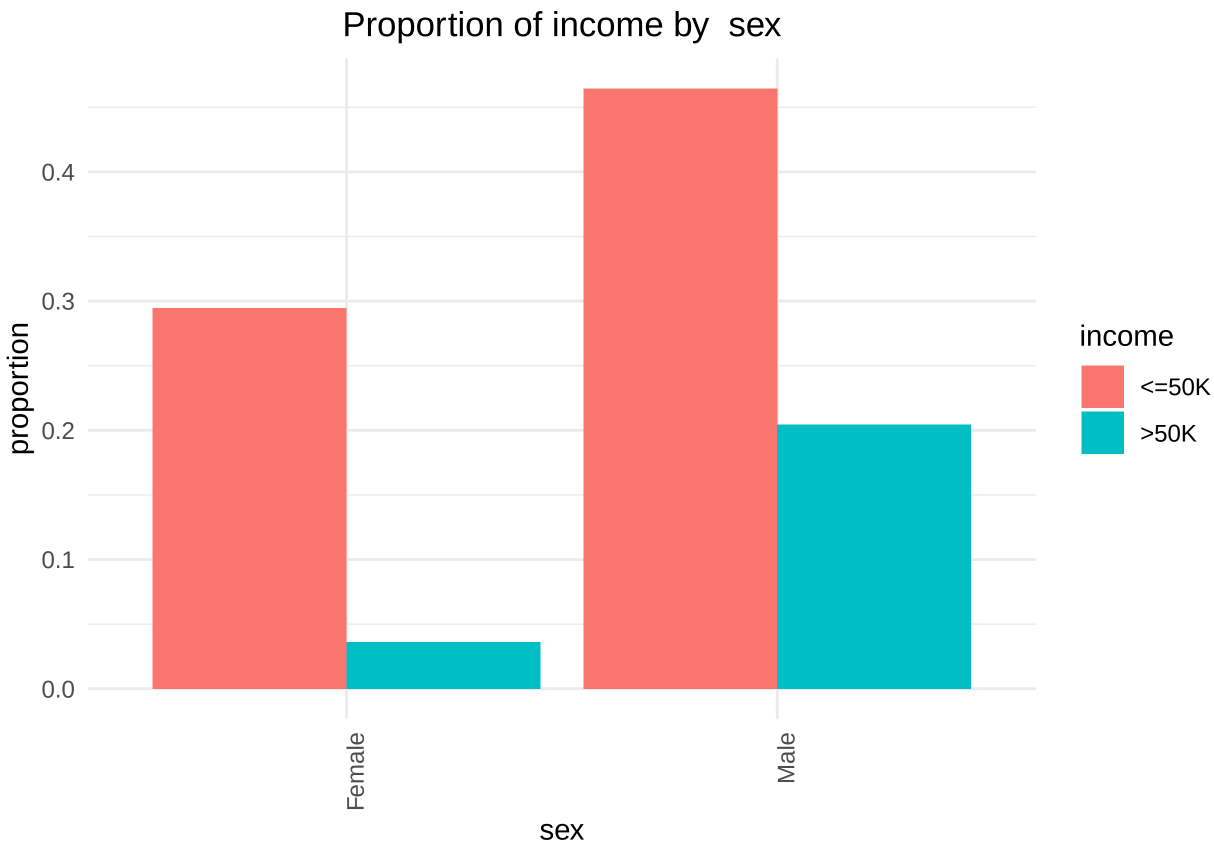


income	count	mean	median	min	max	kurtosis	skewness
<=50K	24720	36.78374	34	17	90	3.032914	0.7631435
>50K	7841	44.24984	44	19	90	3.154975	0.4773580

sex

A higher proportion of males earn >50K compared to females. The proportion of females earning <=50K is significantly higher than the proportion of females earning >50K.

```
feat_desc_cat('sex')
```

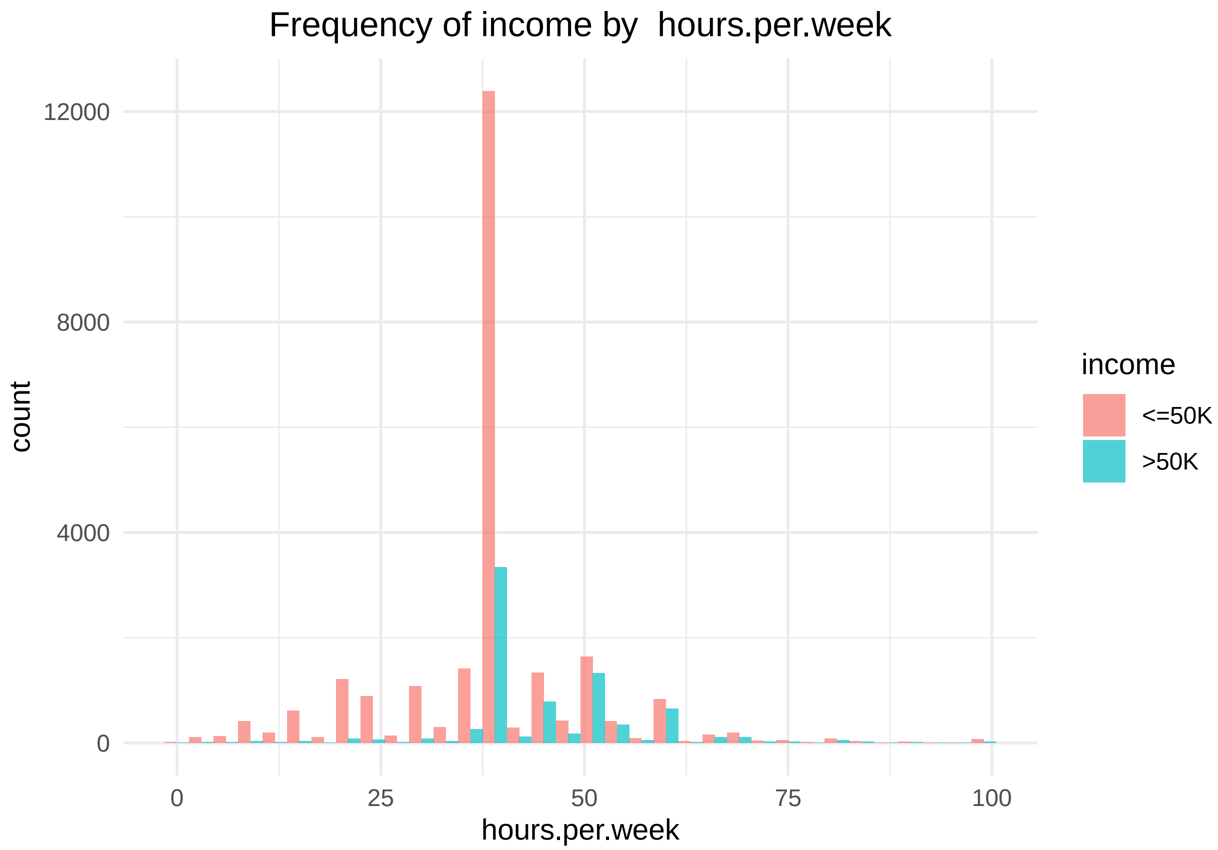


sex	income	count	proportion	percent
Female	<=50K	9592	0.2945855	29%
Female	>50K	1179	0.0362090	4%
Male	<=50K	15128	0.4646049	46%
Male	>50K	6662	0.2046006	20%

hours.per.week

There is a positive correlation between the number of hours worked per week and the likelihood of earning >50K. This trend is particularly evident among those working more than the standard 40 hours per week.

```
feat_desc_num('hours.per.week', 3)
```

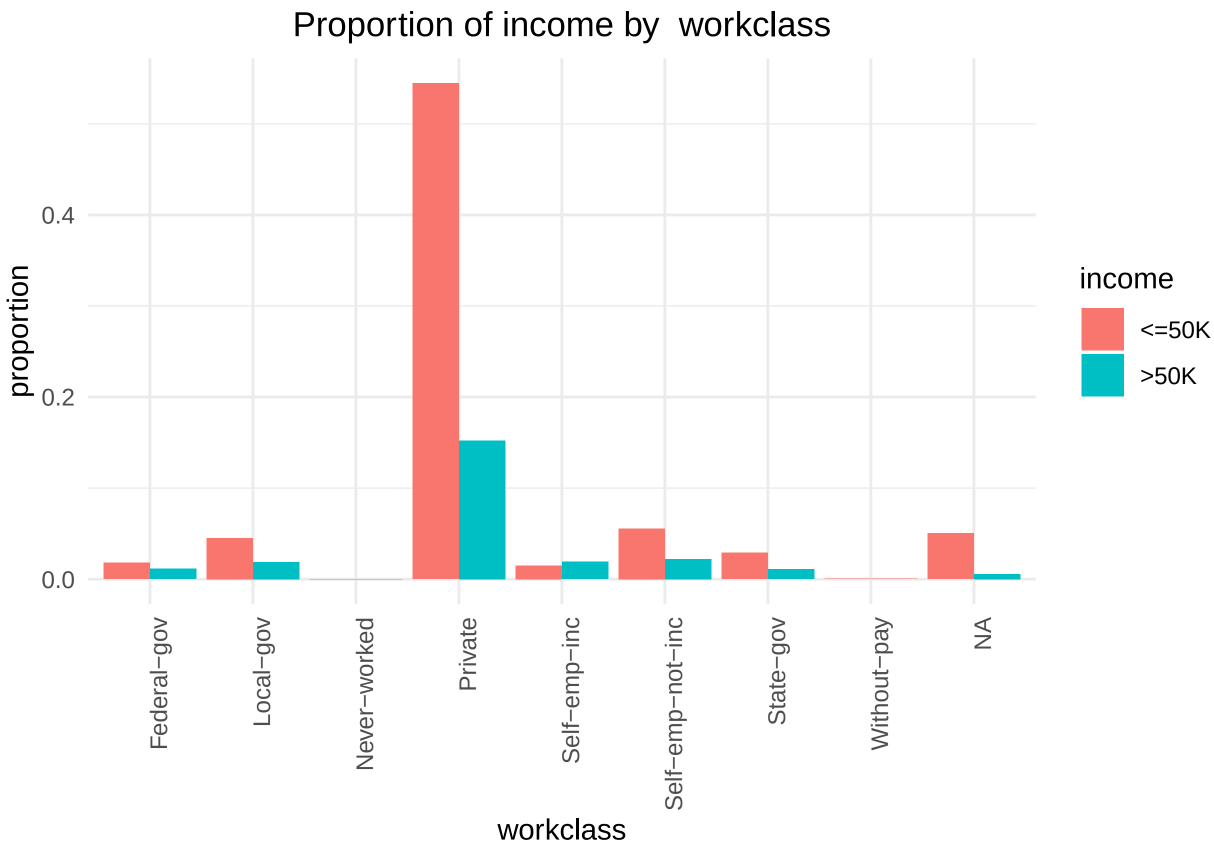


income	count	mean	median	min	max	kurtosis	skewness
<=50K	24720	38.84021	34	1	99	5.970909	0.2136417
>50K	7841	45.47303	44	1	99	6.913578	0.6483987

workclass

Private sector and Self-employment achieves higher income.

```
feat_desc_cat('workclass')
```

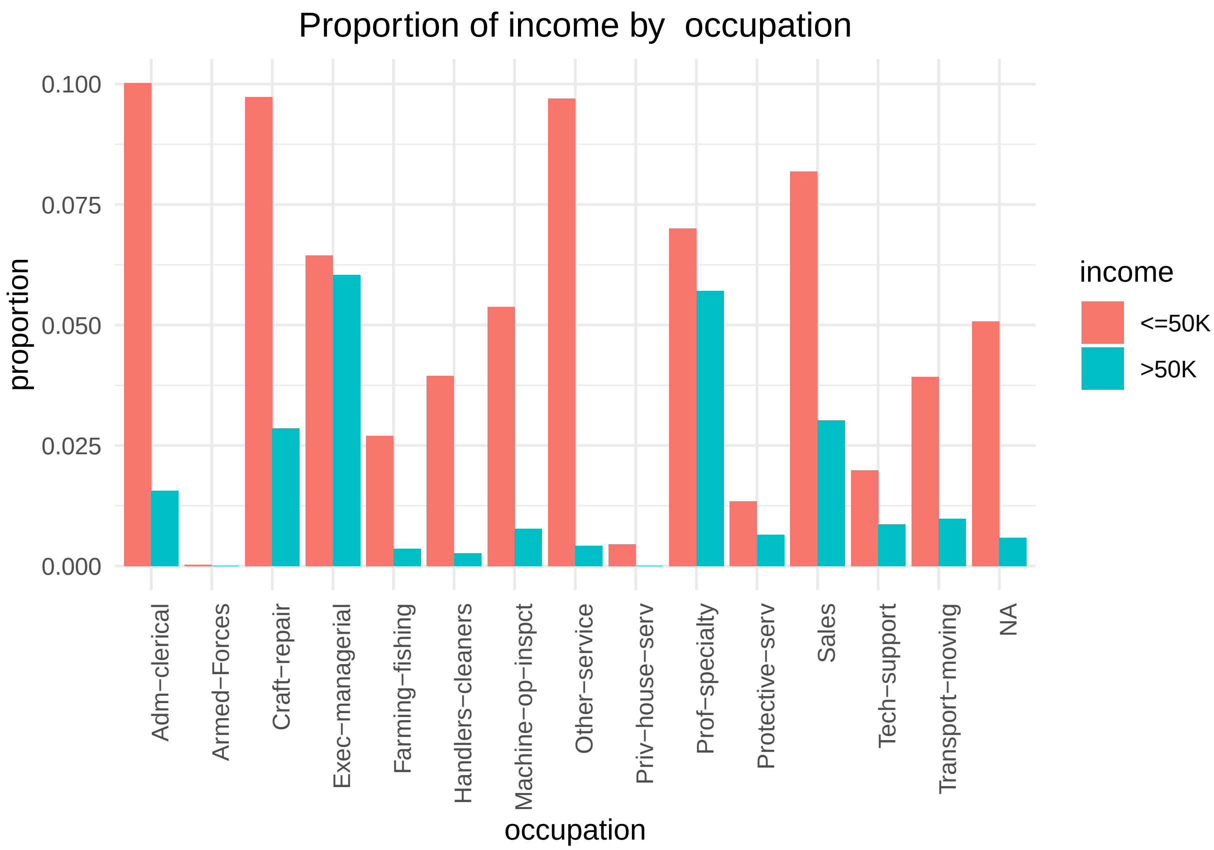


workclass	income	count	proportion	percent
Federal-gov	<=50K	589	0.0180891	2%
Federal-gov	>50K	371	0.0113940	1%
Local-gov	<=50K	1476	0.0453303	5%
Local-gov	>50K	617	0.0189490	2%
Never-worked	<=50K	7	0.0002150	0%
Private	<=50K	17733	0.5446086	54%
Private	>50K	4963	0.1524216	15%
Self-emp-inc	<=50K	494	0.0151715	2%
Self-emp-inc	>50K	622	0.0191026	2%
Self-emp-not-inc	<=50K	1817	0.0558030	6%
Self-emp-not-inc	>50K	724	0.0222352	2%
State-gov	<=50K	945	0.0290225	3%
State-gov	>50K	353	0.0108412	1%
Without-pay	<=50K	14	0.0004300	0%
NA	<=50K	1645	0.0505206	5%
NA	>50K	191	0.0058659	1%

occupation

Occupations such as executive/managerial, professional specialties, and protective services show higher earning potential, while roles in administrative, clerical, farming, and private household services are associated with lower incomes.

```r
feat_desc_cat('occupation')
```

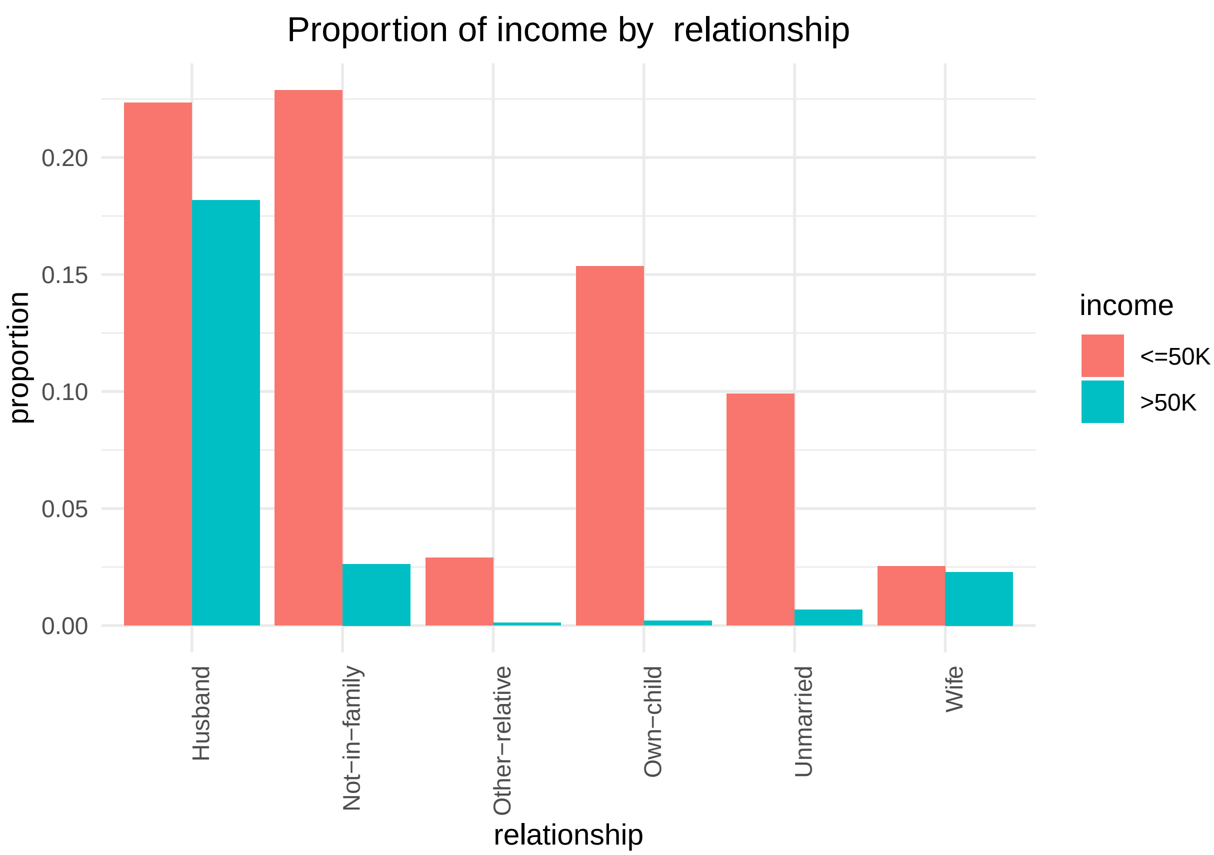


occupation	income	count	proportion	percent
Adm-clerical	<=50K	3263	0.1002119	10%
Adm-clerical	>50K	507	0.0155708	2%
Armed-Forces	<=50K	8	0.0002457	0%
Armed-Forces	>50K	1	0.0000307	0%
Craft-repair	<=50K	3170	0.0973557	10%
Craft-repair	>50K	929	0.0285311	3%
Exec-managerial	<=50K	2098	0.0644329	6%
Exec-managerial	>50K	1968	0.0604404	6%
Farming-fishing	<=50K	879	0.0269955	3%
Farming-fishing	>50K	115	0.0035318	0%
Handlers-cleaners	<=50K	1284	0.0394337	4%
Handlers-cleaners	>50K	86	0.0026412	0%
Machine-op-inspct	<=50K	1752	0.0538067	5%
Machine-op-inspct	>50K	250	0.0076779	1%
Other-service	<=50K	3158	0.0969872	10%
Other-service	>50K	137	0.0042075	0%
Priv-house-serv	<=50K	148	0.0045453	0%
Priv-house-serv	>50K	1	0.0000307	0%
Prof-specialty	<=50K	2281	0.0700531	7%
Prof-specialty	>50K	1859	0.0570928	6%
Protective-serv	<=50K	438	0.0134517	1%
Protective-serv	>50K	211	0.0064801	1%
Sales	<=50K	2667	0.0819078	8%
Sales	>50K	983	0.0301895	3%
Tech-support	<=50K	645	0.0198090	2%
Tech-support	>50K	283	0.0086914	1%
Transport-moving	<=50K	1277	0.0392187	4%
Transport-moving	>50K	320	0.0098277	1%
NA	<=50K	1652	0.0507355	5%
NA	>50K	191	0.0058659	1%

relationship

Husbands tend to have higher incomes compared to wives, and individuals in other categories such as not-in-family, other relatives, own children, and unmarried are more likely to have lower incomes.

```
feat_desc_cat('relationship')
```

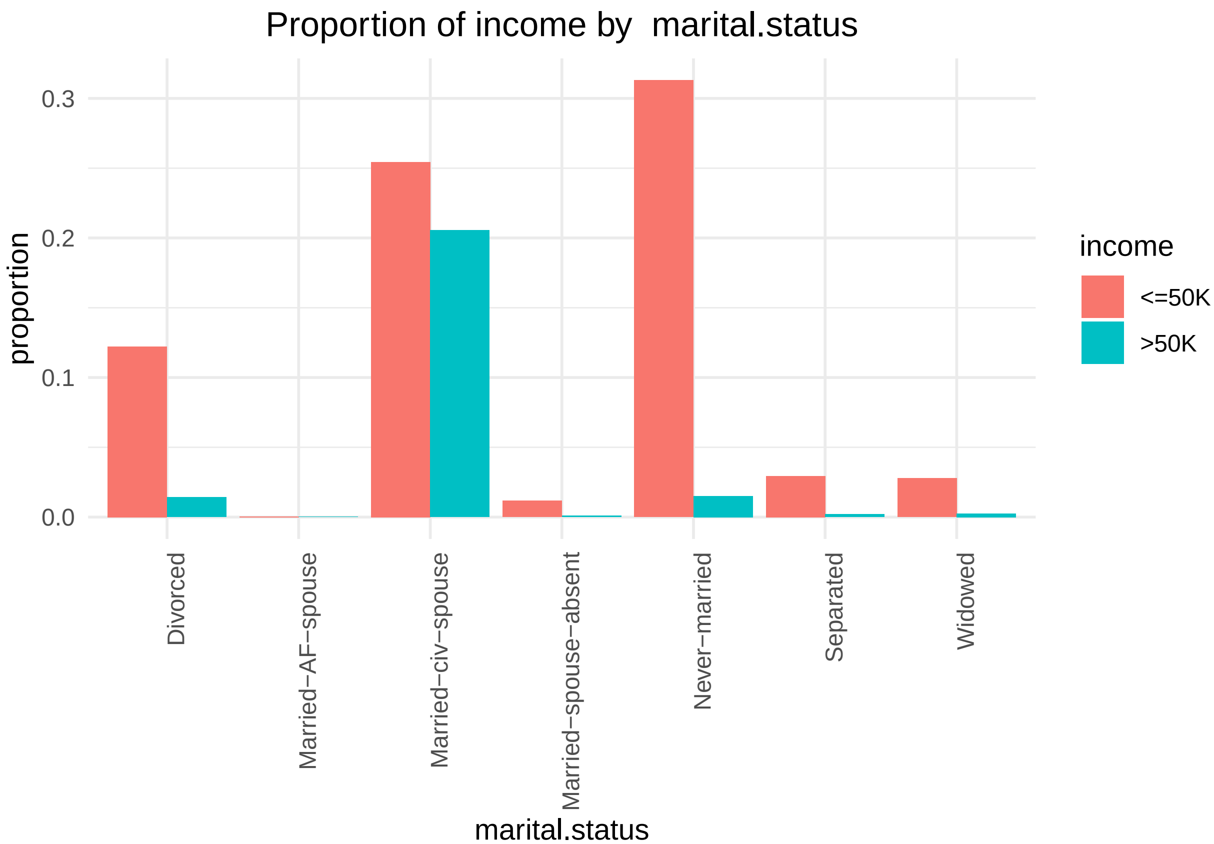


relationship	income	count	proportion	percent
Husband	<=50K	7275	0.2234268	22%
Husband	>50K	5918	0.1817512	18%
Not-in-family	<=50K	7449	0.2287706	23%
Not-in-family	>50K	856	0.0262891	3%
Other-relative	<=50K	944	0.0289917	3%
Other-relative	>50K	37	0.0011363	0%
Own-child	<=50K	5001	0.1535886	15%
Own-child	>50K	67	0.0020577	0%
Unmarried	<=50K	3228	0.0991370	10%
Unmarried	>50K	218	0.0066951	1%
Wife	<=50K	823	0.0252756	3%
Wife	>50K	745	0.0228801	2%

marital.status

Married individuals, particularly those with civilian spouses, have higher earning potential, while those who are divorced, separated, never-married, or widowed are more likely to have lower incomes.

```
feat_desc_cat('marital.status')
```

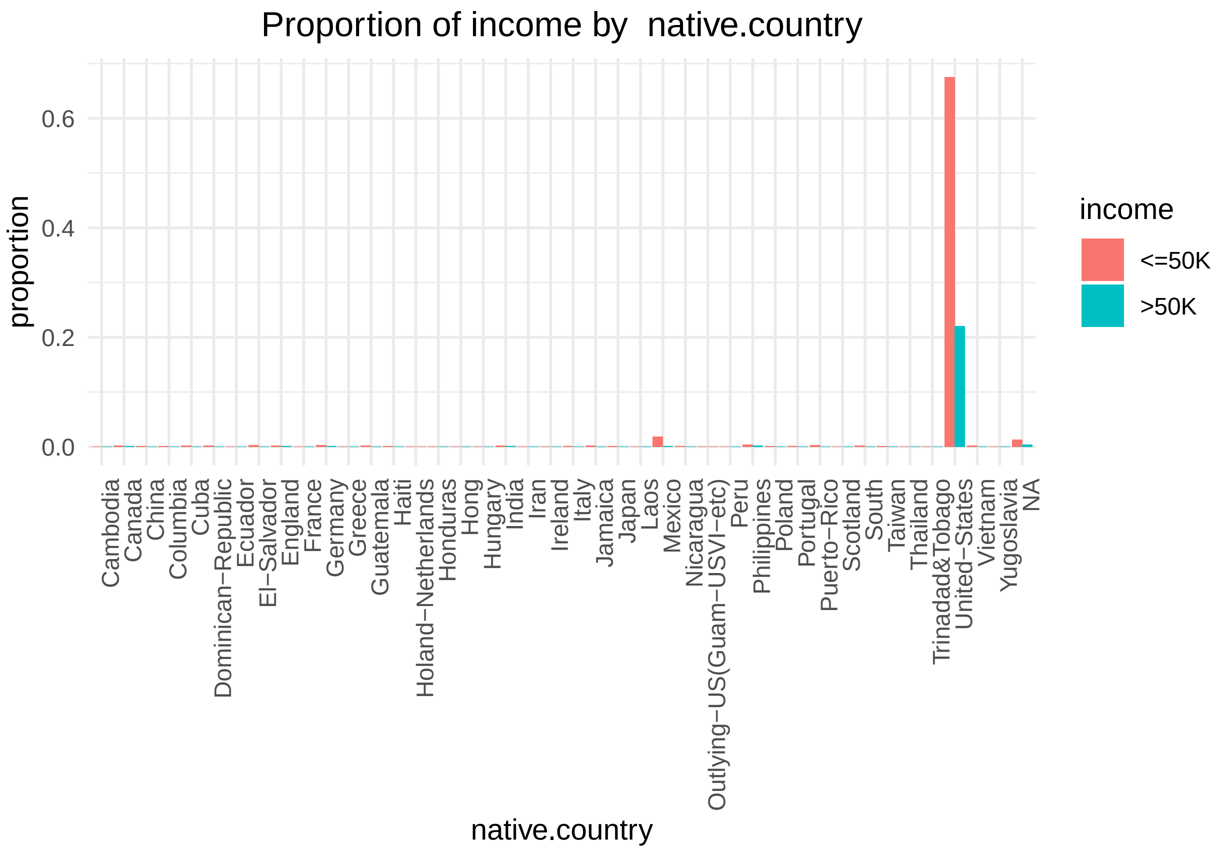


marital.status	income	count	proportion	percent
Divorced	<=50K	3980	0.1222321	12%
Divorced	>50K	463	0.0142195	1%
Married-AF-spouse	<=50K	13	0.0003993	0%
Married-AF-spouse	>50K	10	0.0003071	0%
Married-civ-spouse	<=50K	8284	0.2544148	25%
Married-civ-spouse	>50K	6692	0.2055219	21%
Married-spouse-absent	<=50K	384	0.0117932	1%
Married-spouse-absent	>50K	34	0.0010442	0%
Never-married	<=50K	10192	0.3130125	31%
Never-married	>50K	491	0.0150794	2%
Separated	<=50K	959	0.0294524	3%
Separated	>50K	66	0.0020270	0%
Widowed	<=50K	908	0.0278861	3%
Widowed	>50K	85	0.0026105	0%

native.country

Individuals from the United States having a higher likelihood of earning >50K compared to those from other countries.

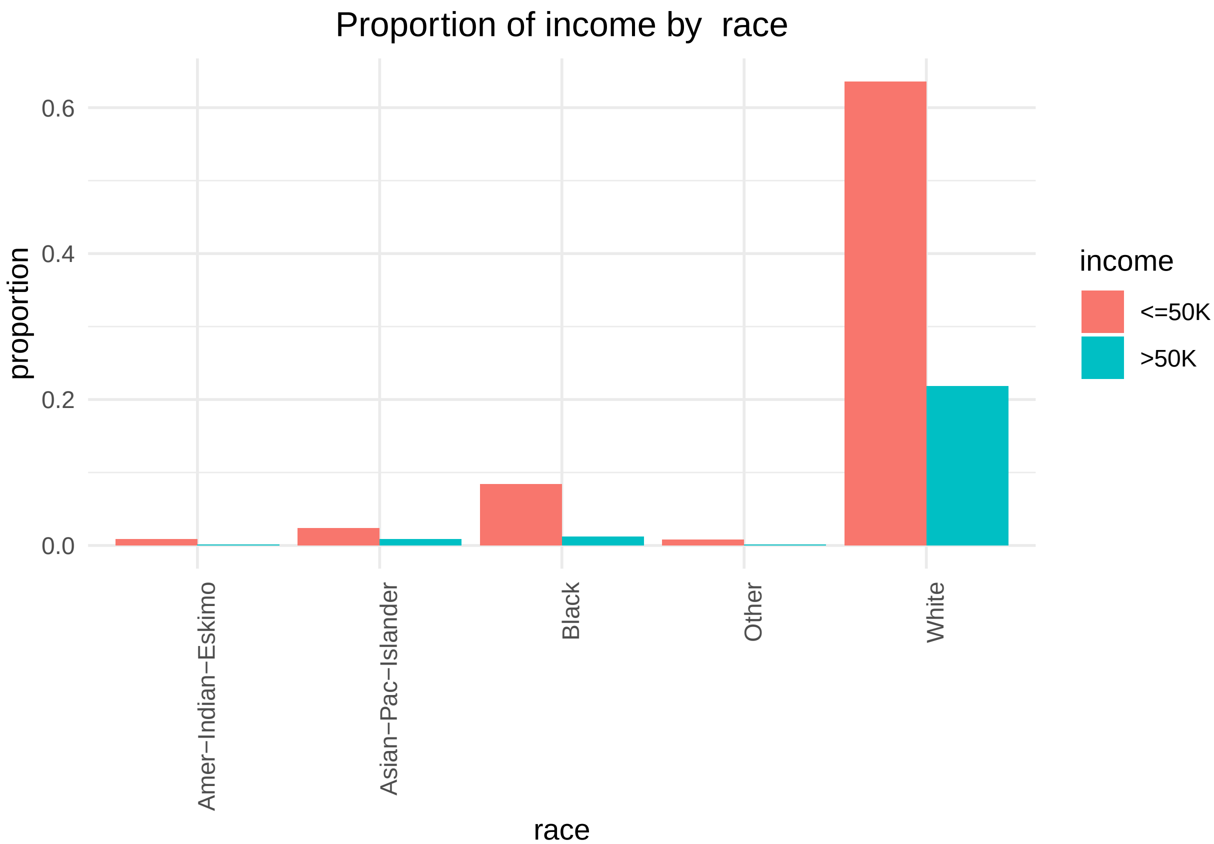
```
feat_desc_cat('native.country', FALSE)
```



race

The majority of individuals identified as White earn <=50K, but there is also a significant proportion earning >50K. This group has the highest overall number of individuals in both income categories.

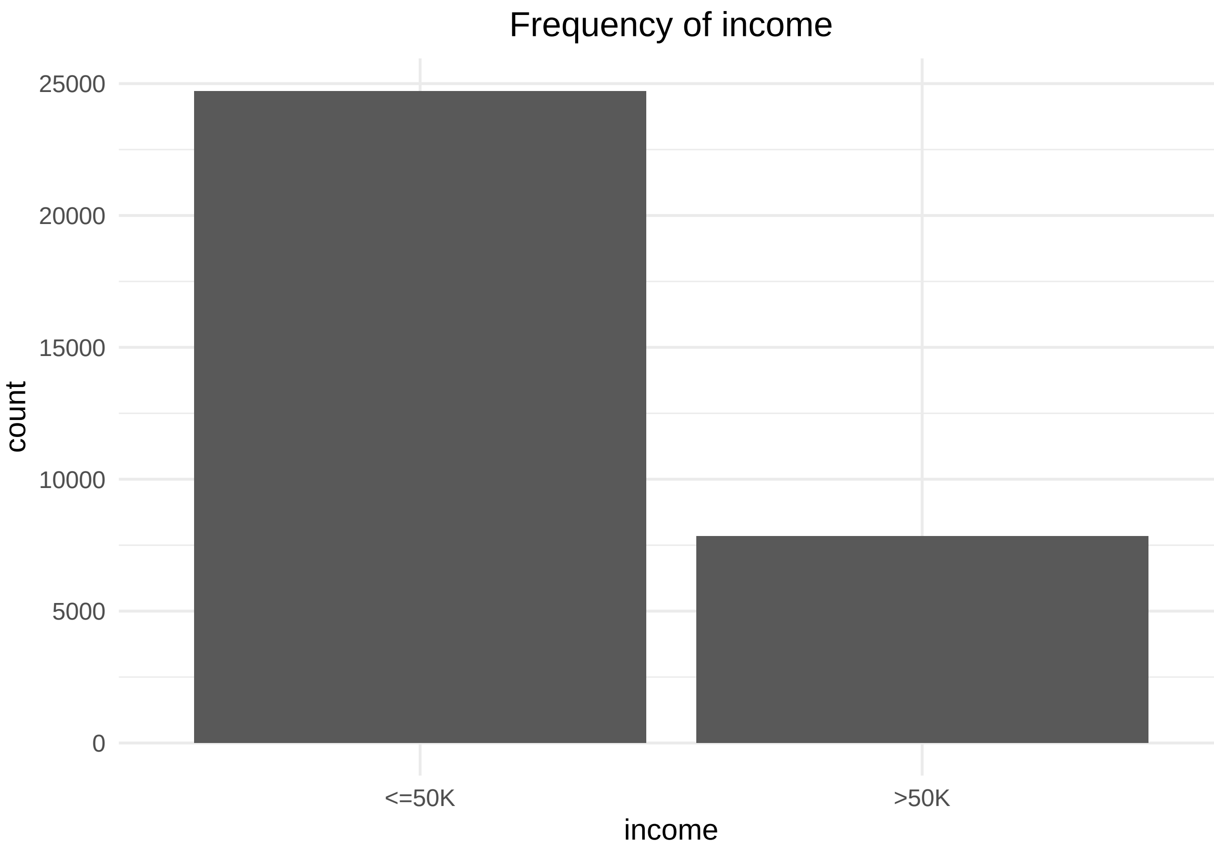
```
feat_desc_cat('race')
```



race	income	count	proportion	percent
Amer-Indian-Eskimo	<=50K	275	0.0084457	1%
Amer-Indian-Eskimo	>50K	36	0.0011056	0%
Asian-Pac-Islander	<=50K	763	0.0234329	2%
Asian-Pac-Islander	>50K	276	0.0084764	1%
Black	<=50K	2737	0.0840576	8%
Black	>50K	387	0.0118854	1%
Other	<=50K	246	0.0075551	1%
Other	>50K	25	0.0007678	0%
White	<=50K	20699	0.6356991	64%
White	>50K	7117	0.2185744	22%

target variable

```
stats <- table(census$income)

g <- ggplot(census, aes(x = income)) +
     geom_bar() +
     theme_minimal() +
     labs(title = "Frequency of income") +
     theme(plot.title = element_text(hjust = 0.5))

grid.draw(g)
```




```r
kable(stats, "latex")
```

Var1	Freq
<=50K	24720
>50K	7841

References

[1] UCI UC Irvine Machine Learning Repository. *Census Income*. 1996. URL: https://archive.ics.uci.edu/dataset/20/census+income (visited on 05/11/2024).