# DSFinal

Dhairya, Varun, Siddhartha, Subham

DataSet origin: https://archive.ics.uci.edu/dataset/2/adult

```
# Read the dataset from the URL
data <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data")
```

```
#cleaning Data
data <- data %>% rename(age = X39,workclass = State.gov,final_weight = X77516,education = Bachelors,edu

#trim data values
data <- data %>%
  mutate(across(everything(), trimws))

#change ? values to NA
data <- data %>%mutate(
  workclass = ifelse(workclass == "?",NA,workclass),
  occupation = ifelse(occupation == "?",NA,occupation),
  native_country = ifelse(native_country == "?",NA,native_country))

#reorder education column values
data$education <- factor(data$education, levels = c("Preschool", "1st-4th", "5th-6th", "7th-8th", "9th"

#changing datatype
data$occupation <- as.factor(data$occupation)
data$workclass <- as.factor(data$workclass)
data$maritial_status <- as.factor(data$maritial_status)
data$relationship <- as.factor(data$relationship)
data$sex <- as.factor(data$sex)
data$race <- as.factor(data$race)
data$native_country <- as.factor(data$native_country)
data$income <- as.factor(data$income)


data$age <- as.numeric(data$age)
data$education_num <- as.numeric(data$education_num)
data$capital_gain <- as.numeric(data$capital_gain)
data$capital_loss <- as.numeric(data$capital_loss)
data$hours_per_week <- as.numeric(data$hours_per_week)
```

The paramters are:

age: the age of an individual
workclass: a general term to represent the employment status of an individual
final_weight: final weight. This is the number of people the census believes the entry represents..
education: the highest level of education achieved by an individual.
education_num: the highest level of education achieved in numerical form.
marital_status: marital status of an individual.

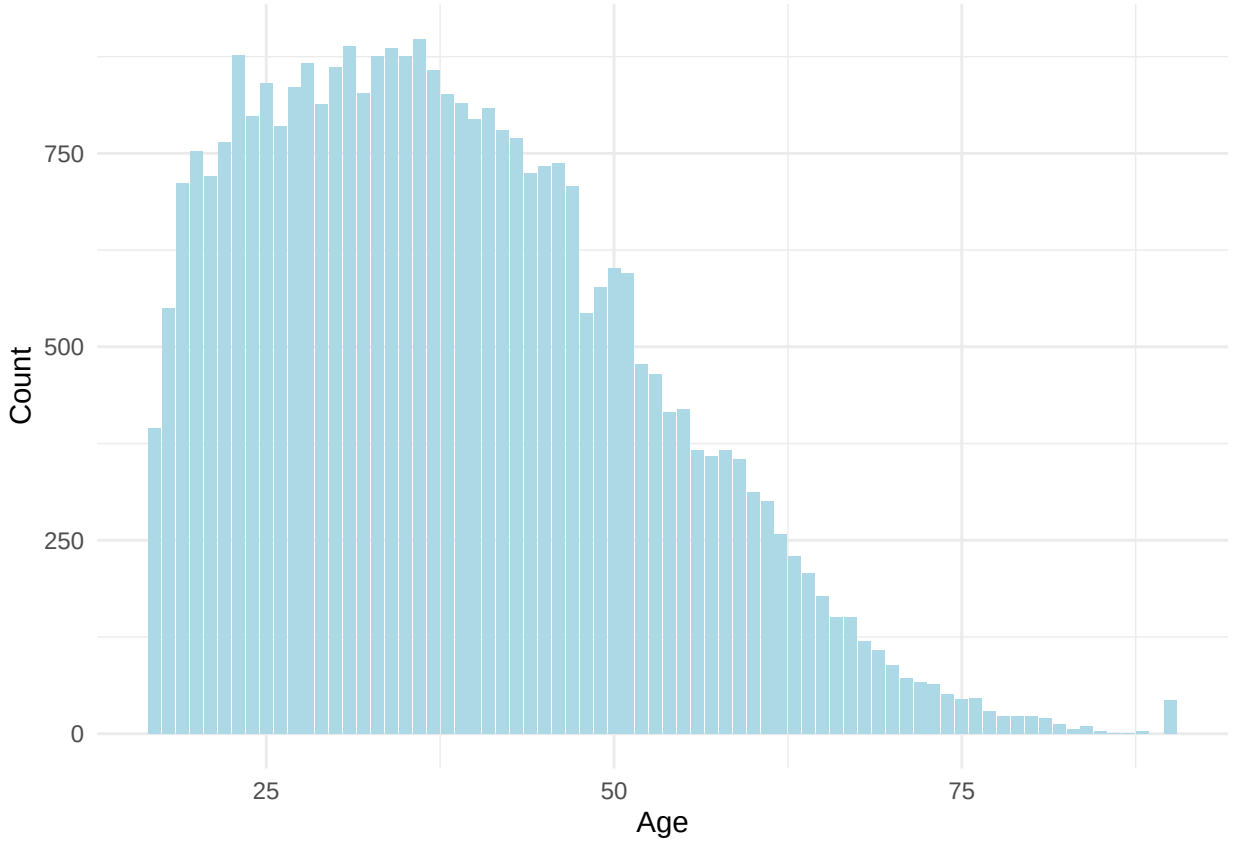
occupation: the general type of occupation of an individual
relationship: represents what this individual is relative to others.
race: Descriptions of an individual's race
sex: the sex of the individual
capital_gain: capital gains for an individual
capital_loss: capital loss for an individual
hours_per_week: the hours an individual has reported to work per week
native_country: country of origin for an individual
income: if the income of that person is less than or greater than 50,000

```
summary(data)
```

```
##       age                     workclass     final_weight
##  Min.   :17.00   Private         :22696   Length:32560
##  1st Qu.:28.00   Self-emp-not-inc: 2541   Class :character
##  Median :37.00   Local-gov       : 2093   Mode  :character
##  Mean   :38.58   State-gov       : 1297
##  3rd Qu.:48.00   Self-emp-inc    : 1116
##  Max.   :90.00   (Other)         :  981
##                  NA's            : 1836
##         education     education_num                maritial_status
##  HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
##  Bachelors   : 5354   Median :10.00   Married-civ-spouse   :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10682
##  11th        : 1175   Max.   :16.00   Separated            : 1025
##  (Other)     : 5134                   Widowed              :  993
##           occupation            relationship                    race
##  Prof-specialty : 4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   : 4099   Not-in-family : 8304   Asian-Pac-Islander: 1039
##  Exec-managerial: 4066   Other-relative:  981   Black             : 3124
##  Adm-clerical   : 3769   Own-child     : 5068   Other             :  271
##  Sales          : 3650   Unmarried     : 3446   White             :27815
##  (Other)        :10993   Wife          : 1568
##  NA's           : 1843
##      sex          capital_gain    capital_loss     hours_per_week
##  Female:10771   Min.   :    0   Min.   :   0.00   Min.   : 1.00
##  Male  :21789   1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.00
##                 Median :    0   Median :   0.00   Median :40.00
##                 Mean   : 1078   Mean   :  87.31   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.00   Max.   :99.00
##
##        native_country    income
##  United-States:29169   <=50K:24719
##  Mexico       :  643   >50K : 7841
##  Philippines  :  198
##  Germany      :  137
##  Canada       :  121
##  (Other)      : 1709
##  NA's         :  583
```

This shows that there are no missing values in the dataset, besides occupation,workclass and native_cou

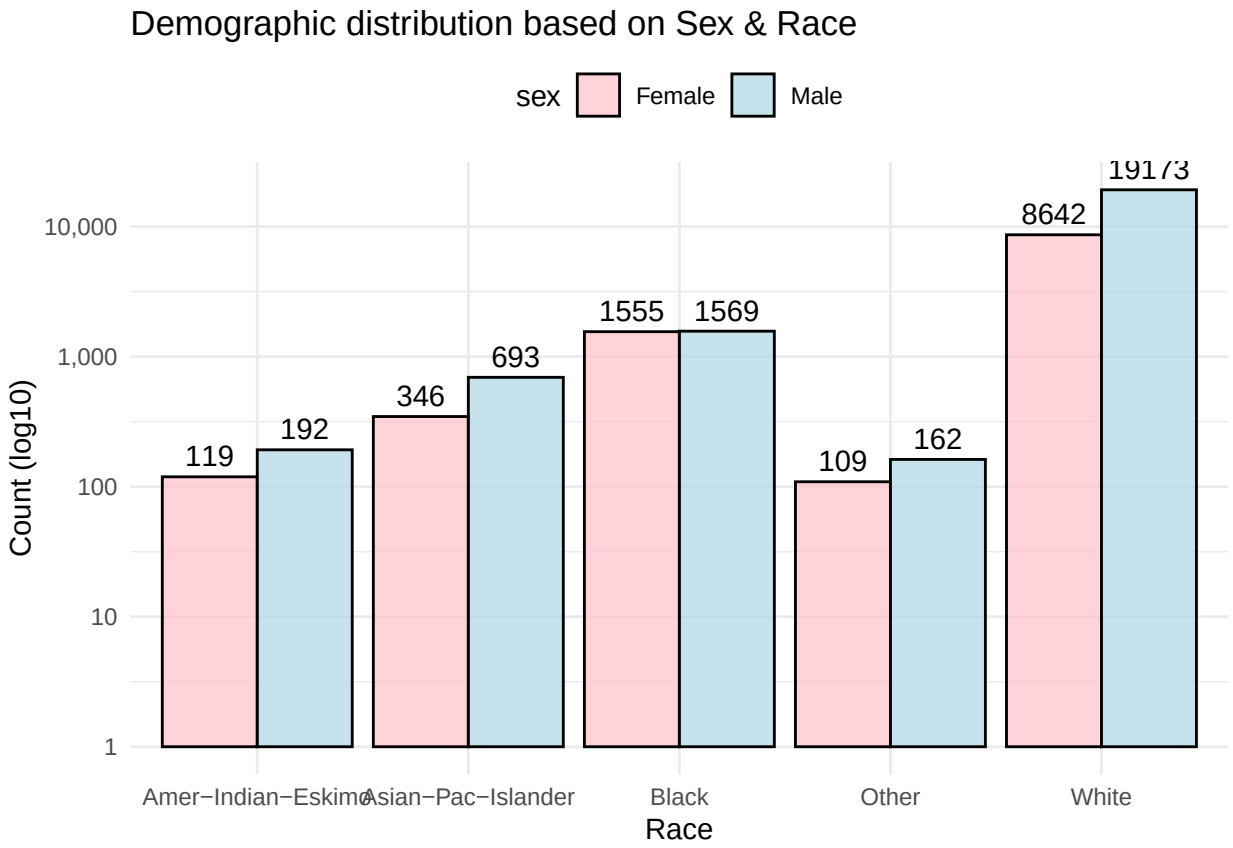
```r
#density plot for age
ggplot(data, aes(x = age)) +
  geom_bar(fill = "lightblue") +
  labs(x = "Age", y = "Count") +
  theme_minimal()
```




This bar plot is left skewed which shows that majority of the working population in ranging from 20 to 5

```r
#demographics of the working population
custom_fill_colors <- c("Female" = "pink", "Male" = "lightblue")

ggplot(data, aes(x = race, fill = sex)) +
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  geom_text(aes(label = ..count..), stat = "count", position =
              position_dodge(width = 0.9), vjust = -0.5) +
  labs(title = "Demographic distribution based on Sex & Race",
       x = "Race", y = "Count (log10)") +
  scale_y_continuous(trans = "log10", labels = scales::comma_format()) +
  scale_fill_manual(values = custom_fill_colors) +
  theme_minimal() +
  theme(legend.position = "top")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




This bar chart shows the demographic destribution of the population in the US. Most of the people are W

```
data <- data %>%
  group_by(native_country) %>%
  mutate(count_native_country = n())
data$log_count_native_country <- log(data$count_native_country +1) * 100  # Adding 1 to handle cases wh

# Create a treemap
treemap(data,
        index = c("native_country"),
        vSize = "log_count_native_country",
        title = "Treemap of Native Country (Log Scale)",
        palette = "Dark2",
        border.col = "black",
        border.lwds = 1,
        fontsize.labels = 12,
        fontcolor.labels = "black",
        align.labels = c("left", "top"),
        overlap.labels = 0,
        inflate.labels = TRUE        )
```

Treemap of Native Country (Log Scale)

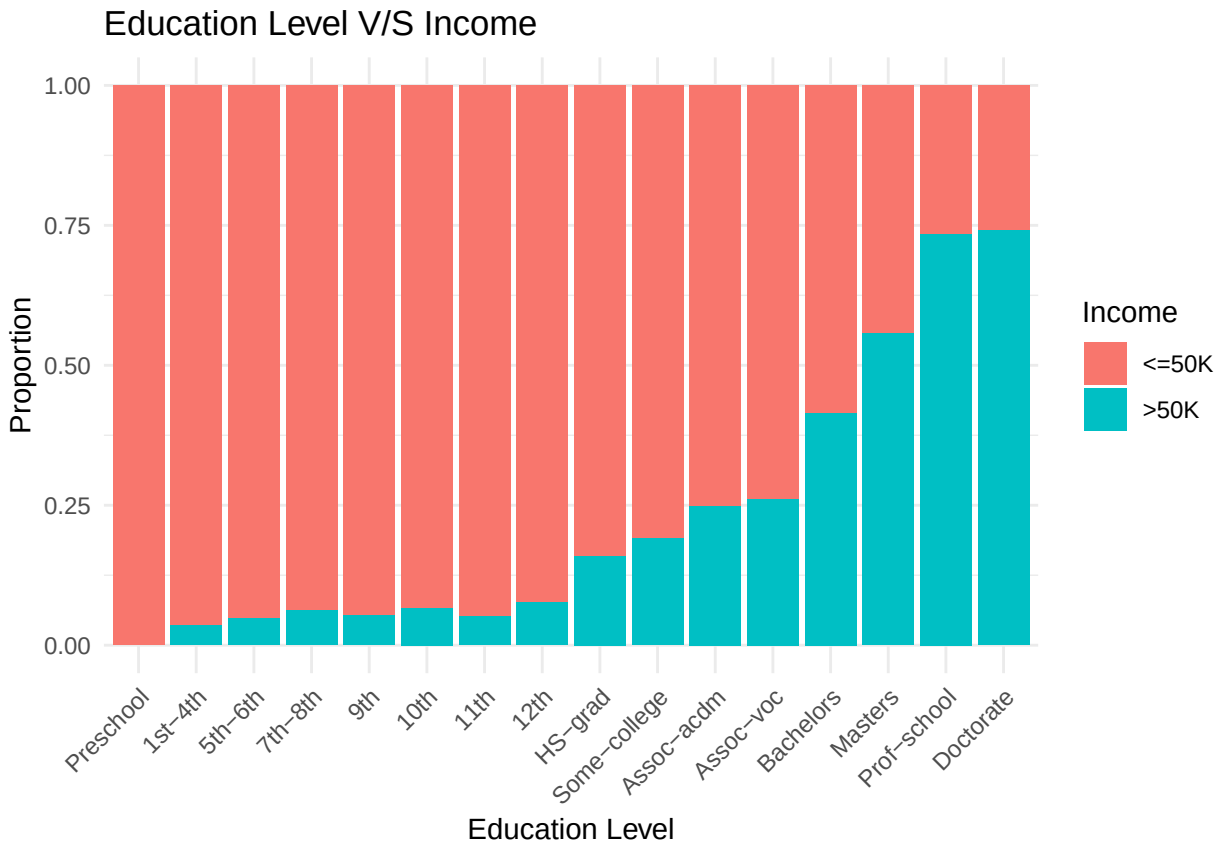

```
#number of observations where income <=50K
print(sum(data$income == "<=50K"))
```

```
## [1] 24719
```

```
#number of observations where income >50K
print(sum(data$income == ">50K"))
```

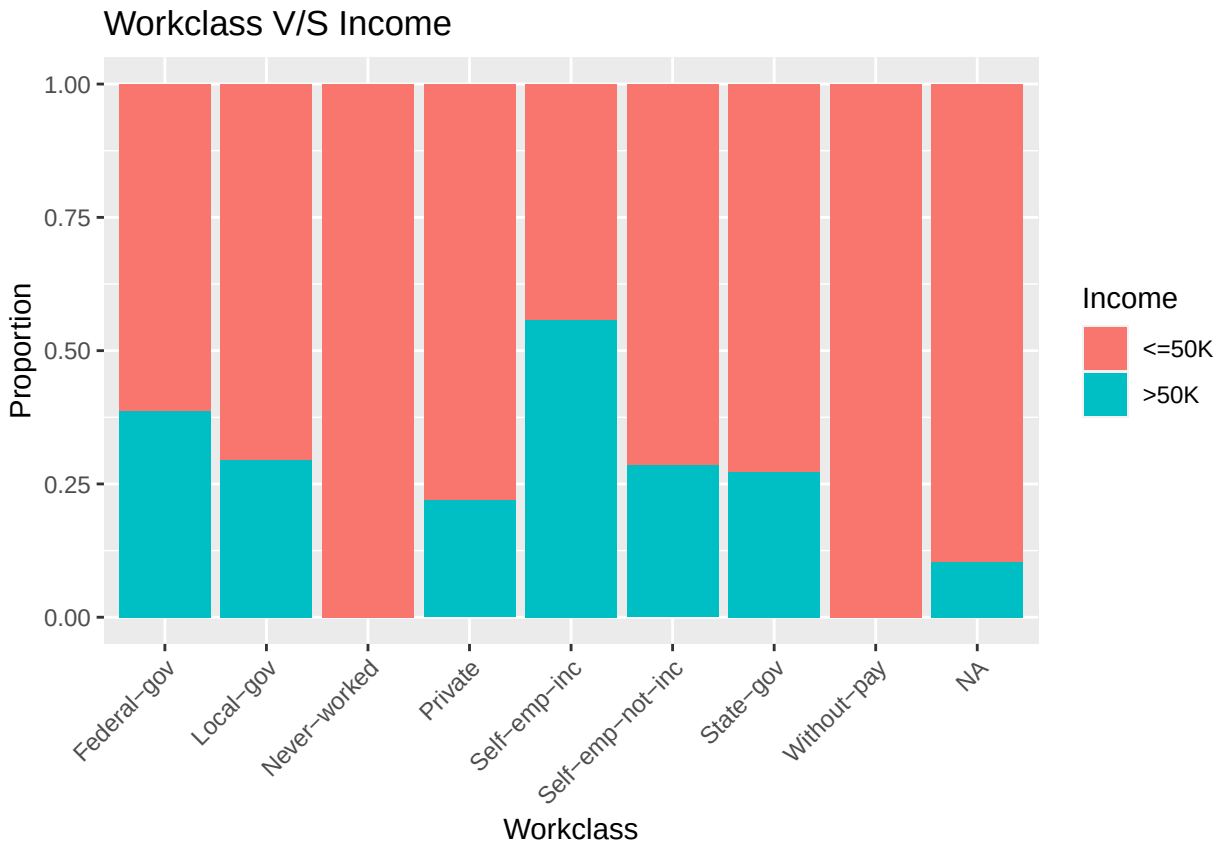
```
## [1] 7841
```

```
ggplot(data, aes(x = education, fill = income)) +
  geom_bar(position = "fill") +
  labs(title = "Education Level V/S Income",x = "Education Level", y = "Proportion", fill = "Income") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle =45,hjust=1))
```



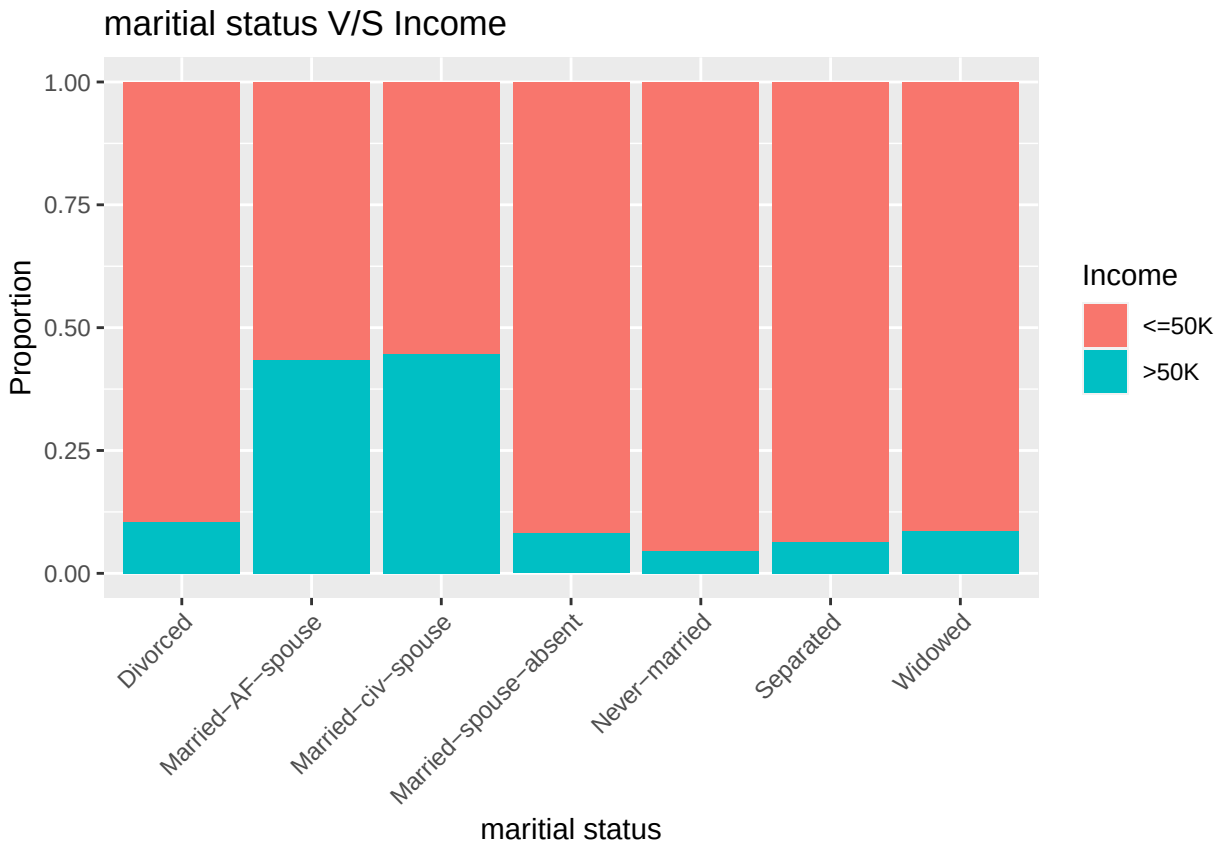

The above bar plot decribes the relation between Education Level and Income. The income values are of tw

```
ggplot(data, aes(x=workclass,fill=income))+
  geom_bar(position = "fill")+
  labs(title = "Workclass V/S Income",x = "Workclass", y = "Proportion", fill = "Income")+
  theme(axis.text.x = element_text(angle=45,hjust=1))
```

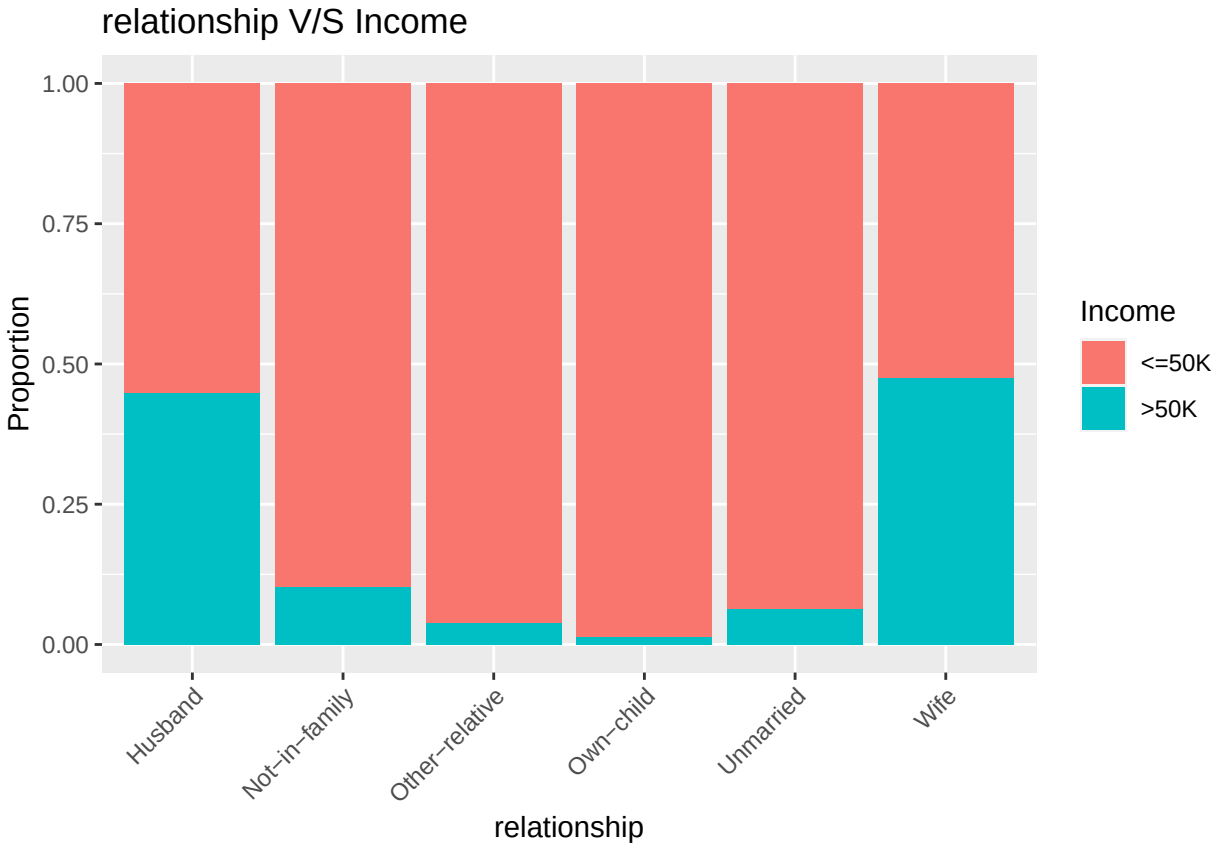



From this graph we can see that majority of the people work in private sector. And majority of the peopl

```
ggplot(data, aes(x=maritial_status,fill=income))+
  geom_bar(position = "fill")+
  labs(title = "maritial status V/S Income",x = "maritial status", y = "Proportion", fill = "Income")+
  theme(axis.text.x = element_text(angle=45,hjust=1))
```



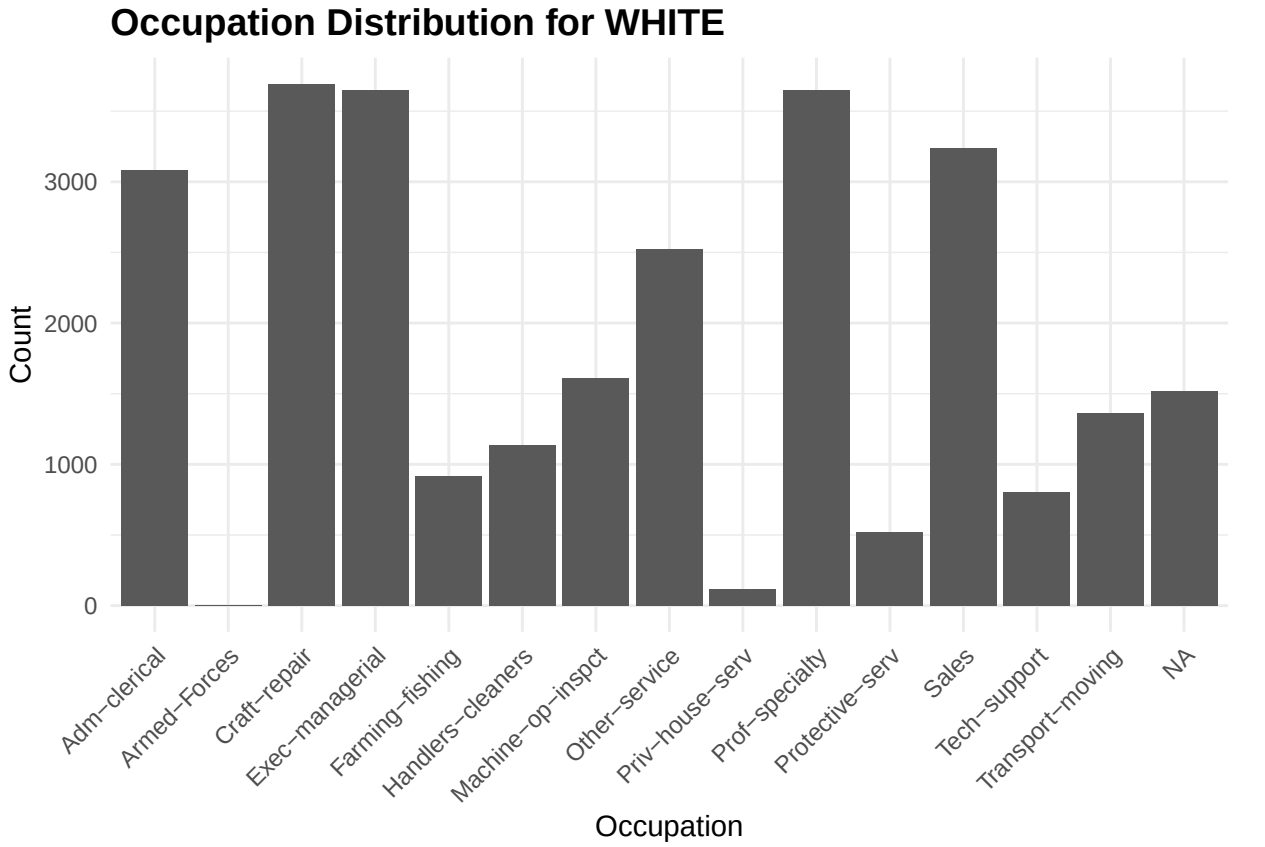

```
ggplot(data, aes(x=relationship,fill=income))+
  geom_bar(position = "fill")+
  labs(title = "relationship V/S Income",x = "relationship", y = "Proportion", fill = "Income")+
  theme(axis.text.x = element_text(angle=45,hjust=1))
```

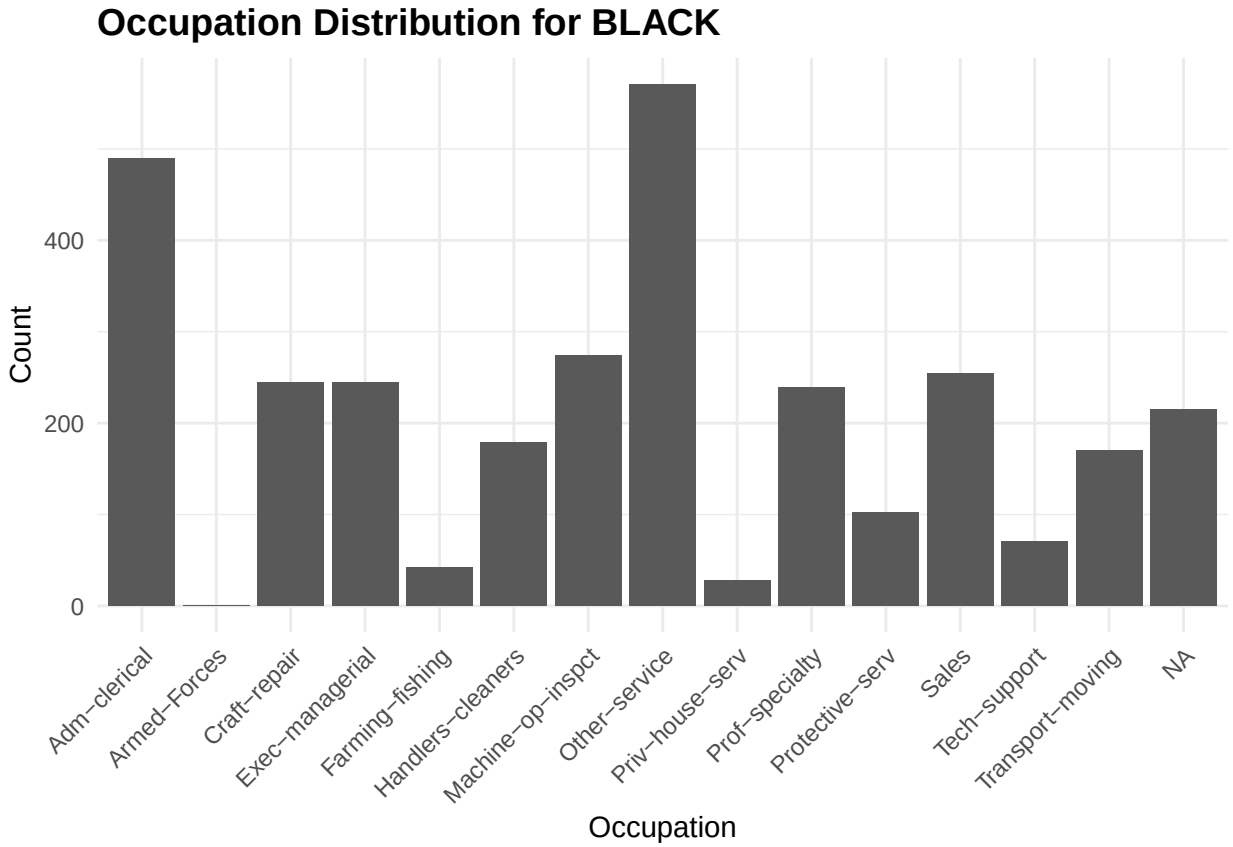


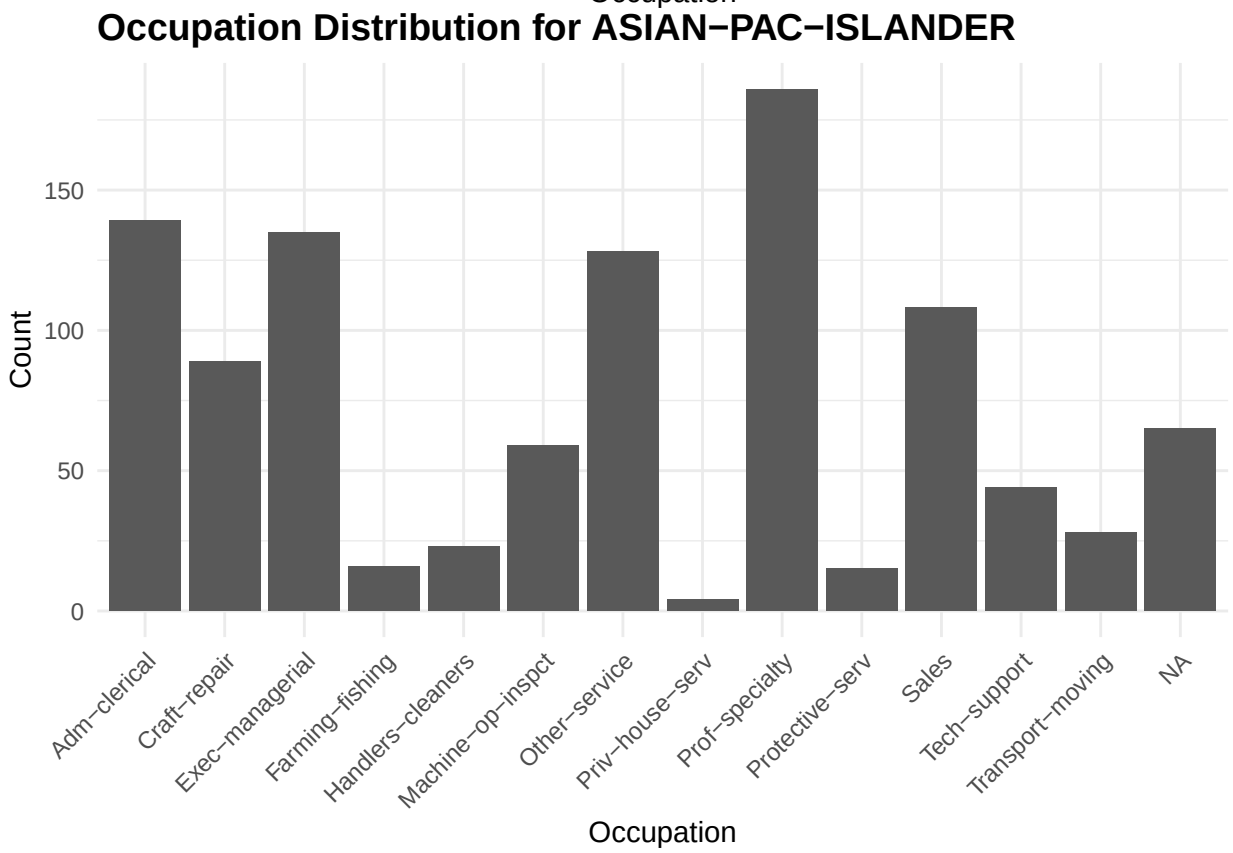

From these two graphs we can observe that families that are together have a better income proportion tha
The data is observered to be consistant with the relationships column. Husbands and wives have a better

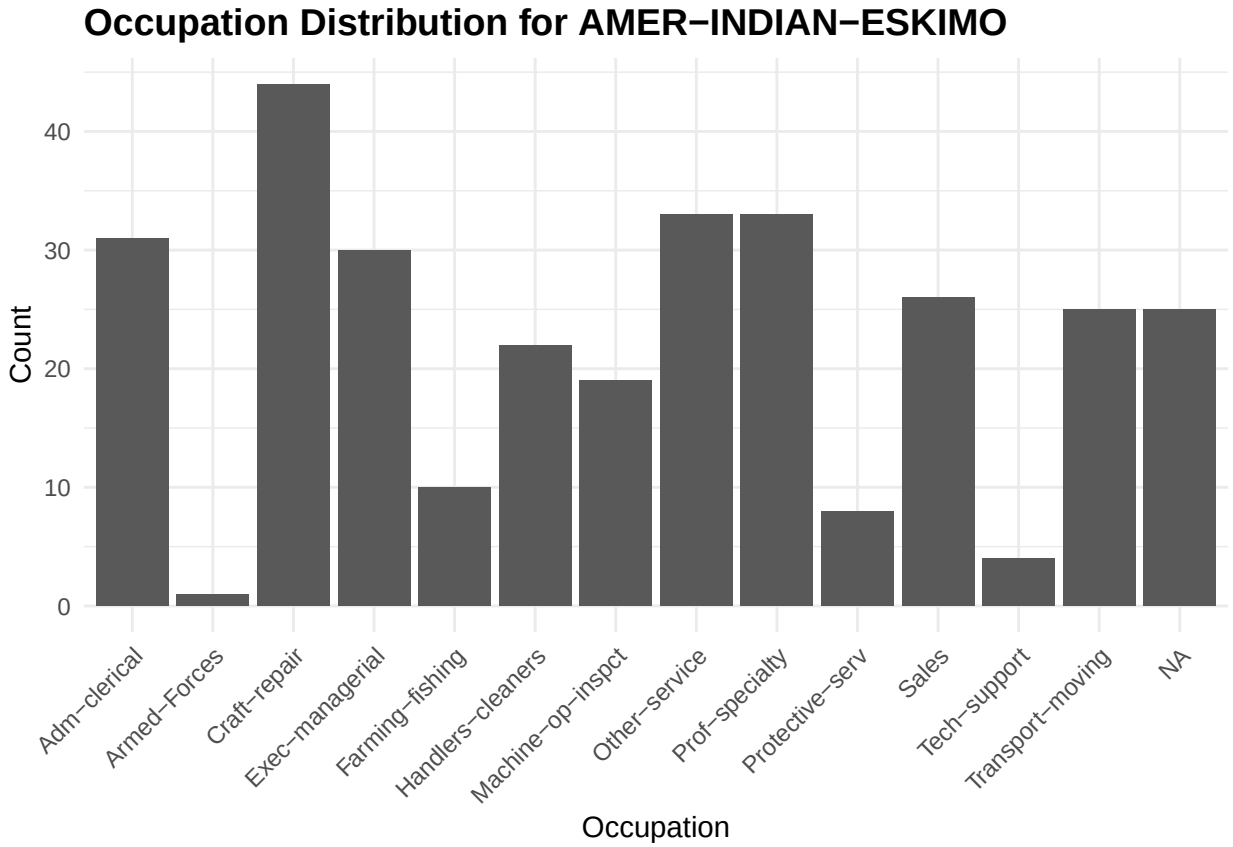
```r
# Create facetted plot with clearer labels
library(ggplot2)

# Assuming 'data' is your dataset
# Loop through each unique race
for(race in unique(data$race)) {
  # Filter the data for one race at a time
  race_data <- data[data$race == race, ]

  # Plotting
  p <- ggplot(race_data, aes(x = occupation)) +
    geom_bar() +
    labs(
      title = paste("Occupation Distribution for", toupper(race)),
      x = "Occupation",
      y = "Count"
    ) +
    theme_minimal() +
    theme(
      axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels for better readability
      plot.title = element_text(size = 14, face = "bold")  # Bold and larger title
    )

  # Print the plot
```

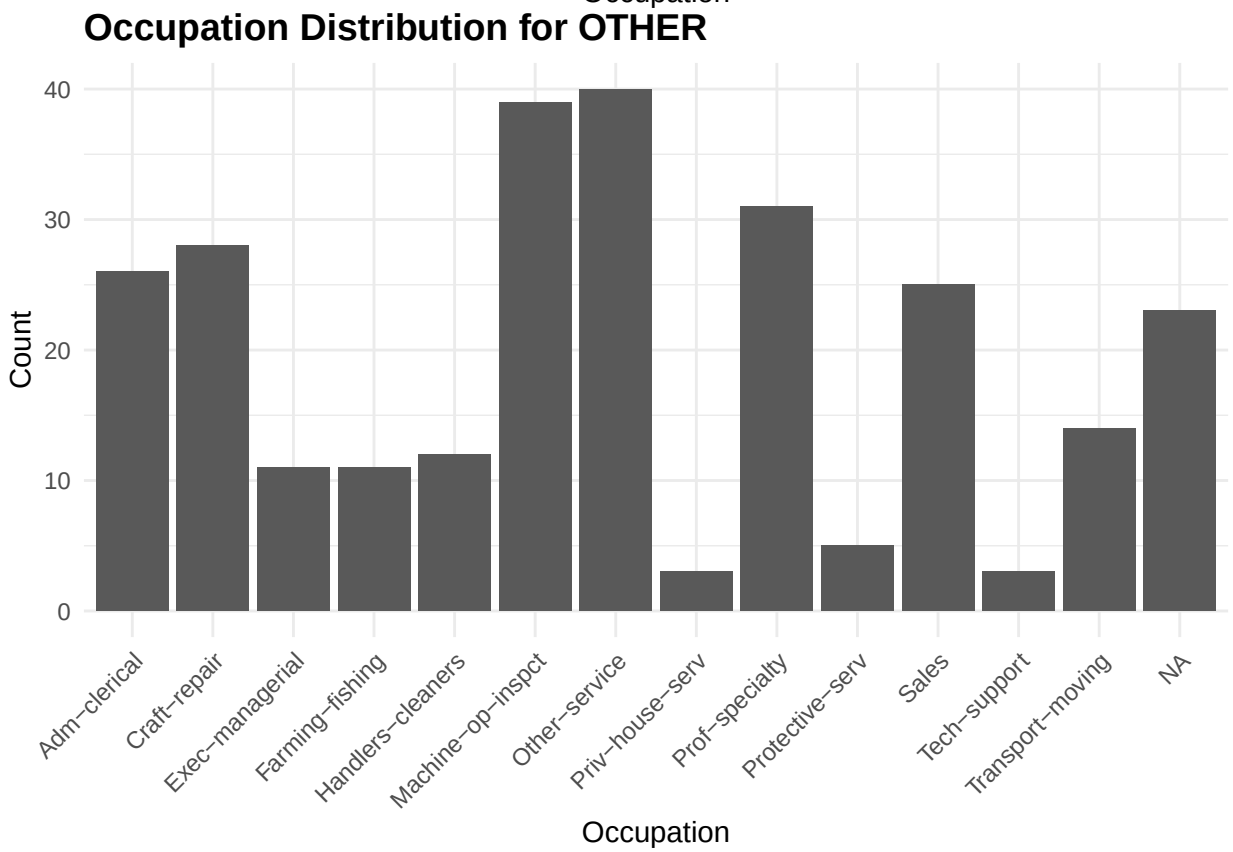
```
  print(p)
}
```


Occupation Distribution for WHITE
Count
3000
2000
1000
0
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
NA
Occupation


Occupation Distribution for BLACK
Count
400
200
0
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
NA
Occupation


Occupation Distribution for ASIAN−PAC−ISLANDER
Count
150
100
50
0
Adm−clerical
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
NA
Occupation


Occupation Distribution for AMER−INDIAN−ESKIMO
Count
0
10
20
30
40
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
NA
Occupation


Occupation Distribution for OTHER
Count
0
10
20
30
40
Adm−clerical
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
NA
Occupation