

Demographic_analysis

Andrew

2024-05-18

```
#LOAD PACKAGES AND THE DATASET
```

```
#ggplot2
```

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'  
## (as 'lib' is unspecified)
```

```
library(ggplot2)
```

```
#tidyverse
```

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'  
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
## v dplyr      1.1.4      v readr      2.1.5
```

```
## v forcats    1.0.0      v stringr    1.5.1
```

```
## v lubridate  1.9.3      v tibble     3.2.1
```

```
## v purrr      1.0.2      v tidyr      1.3.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# dataset
```

```
demographic_data <- read.csv("demographic_data.csv")
```

```
# INDIVIDUALS WITH INCOME GREATER THAN 70K, EDUCATED AND OWN A HOUSE
```

```
# Logical statement
```

```
subset <- demographic_data[demographic_data$Income > 70000 &  
                           demographic_data$Education == 1 &  
                           demographic_data$House_Ownership == 1, ]
```

```
# print subset
```

```
print(subset)
```

```
##      Age      Sex Income Education Religion House_Ownership
## 3    43    male  80900          1         0             1
## 17   47  female  81955          1         0             1
## 37   38    male  98570          1         1             1
## 38   28    male  89840          1         0             1
## 39   65    male  89215          1         0             1
```

```
# INDIVIDUALS ABOVE 60 YEARS THAT DONT OWN A HOUSE
```

```
subset <-demographic_data[demographic_data$Age > 60 & demographic_data$House_Ownership==0,]
print(subset)
```

```
##      Age      Sex Income Education Religion House_Ownership
## 13   69  female  41475          1         1             0
## 20   62    male  46785          1         0             0
## 28   80  female  63215          0         1             0
## 43   68  female  37635          1         0             0
## 47   76    male  79320          1         0             0
```

```
# WHAT IS THE RELATIONSHIP BETWEEN INCOME AND HOME OWNERSHIP
```

```
# Load library
```

```
library(dplyr)
```

```
# Calculate the mean income for each house ownership status
```

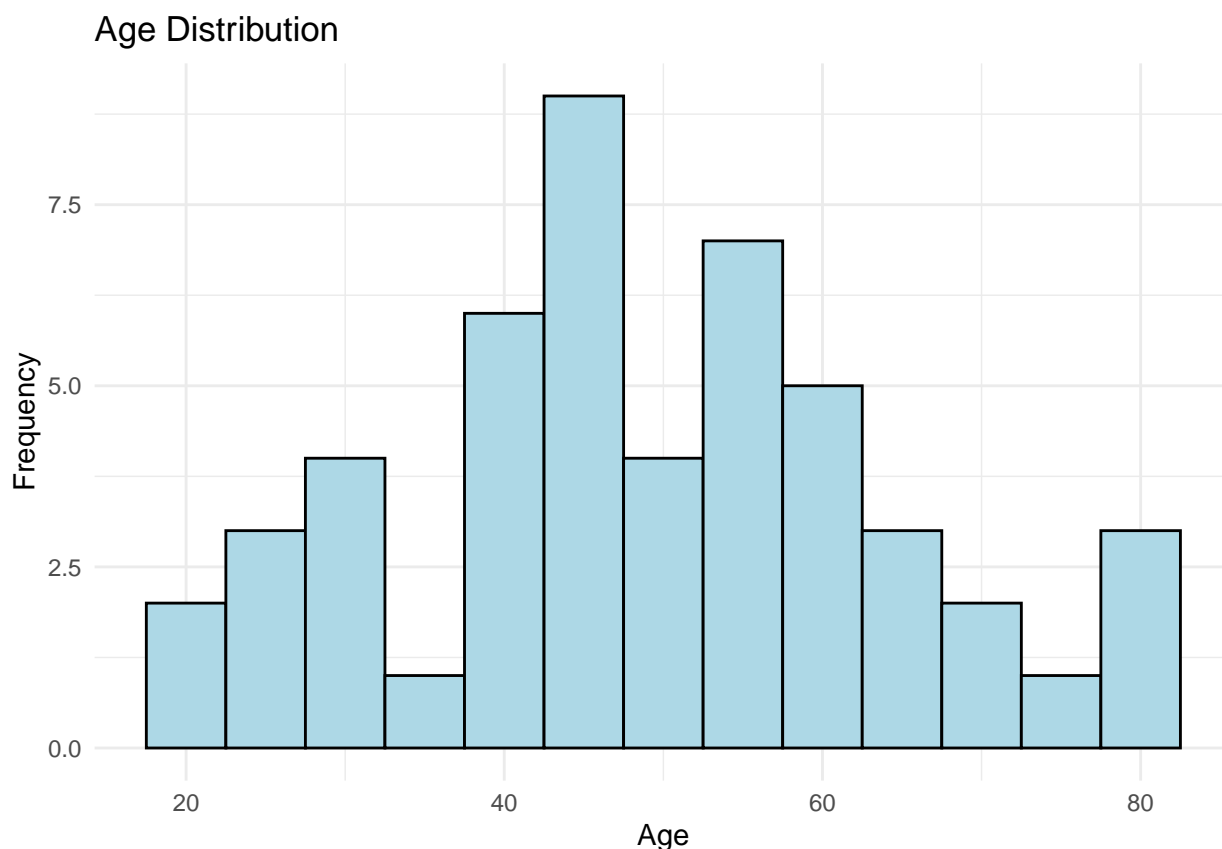
```
income_house_ownership <- demographic_data %>%
  group_by(House_Ownership) %>%
  summarize(mean_income = mean(Income))
```

```
# Print the results
```

```
# WHAT IS THE FREQUENCY OF AGE DISTRIBUTION
```

```
# Create a histogram of ages
```

```
library(ggplot2)
ggplot(demographic_data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(title = "Age Distribution",
       x = "Age",
       y = "Frequency") +
  theme_minimal()
```



```
if (mean(demographic_data$Income) > 50000) {
  print("Income exceeds $50,000 for most individuals.")
} else if (mean(demographic_data$Income) == 50000) {
  print("Income is exactly $50,000 for most individuals.")
} else {
  print("Income is less than $50,000 for most individuals.")
}
```

```
## [1] "Income exceeds $50,000 for most individuals."
```

```
# IS THERE A SIGNIFICANT DIFFERENCE IN INCOME BETWEEN MALES AND FEMALES
```

```
# Calculate the mean income for each sex
```

```
income_by_sex <- demographic_data %>%
  group_by(Sex) %>%
  summarise(mean_income=mean(Income))
```

```
# Print results
```

```
print(income_by_sex)
```

```
## # A tibble: 2 x 2
##   Sex      mean_income
##   <chr>      <dbl>
## 1 female    45542.
## 2 male     53501.
```

```
# Perform a t-test to compare incomes of males and females
```

```
t_test_result <- t.test(Income ~ Sex, data = demographic_data)
```

```
print(t_test_result)
```

```
##
```

```
## Welch Two Sample t-test
```

```
##
```

```
## data: Income by Sex
```

```
## t = -1.0432, df = 39.962, p-value = 0.3031
```

```
## alternative hypothesis: true difference in means between group female and group male is not equal to
```

```
## 95 percent confidence interval:
```

```
## -23378.292 7460.459
```

```
## sample estimates:
```

```
## mean in group female mean in group male
```

```
## 45542.25 53501.17
```

```
ggplot(data=demographic_data, mapping=aes(x=Income, y=Sex))+  
  geom_boxplot(fill="lightblue")+  
  labs(title="Income by Gender",  
        x="Income",  
        y="Sex")
```

