midterm

2023-10-25

Packages

```
library(tidyverse)
```

Data

```
incomeData <- read.csv("data/adult.income.csv")
```

```
glimpse(incomeData)
```

```
## Rows: 32,561
## Columns: 15
## $ age            <int> 39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30, 23, 32,~
## $ workClass      <chr> " State-gov", " Self-emp-not-inc", " Private", " Privat~
## $ fnlwgt         <int> 77516, 83311, 215646, 234721, 338409, 284582, 160187, 2~
## $ education      <chr> " Bachelors", " Bachelors", " HS-grad", " 11th", " Bach~
## $ education.num  <int> 13, 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13, 12, 11,~
## $ marital.status <chr> " Never-married", " Married-civ-spouse", " Divorced", "~
## $ occupation     <chr> " Adm-clerical", " Exec-managerial", " Handlers-cleaner~
## $ relationship   <chr> " Not-in-family", " Husband", " Not-in-family", " Husba~
## $ race           <chr> " White", " White", " White", " Black", " Black", " Whi~
## $ sex            <chr> " Male", " Male", " Male", " Male", " Female", " Female~
## $ capital.gain   <int> 2174, 0, 0, 0, 0, 0, 0, 0, 14084, 5178, 0, 0, 0, 0, 0, ~
## $ capital.loss   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ hours.per.week <int> 40, 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40, 30, 50,~
## $ native.country <chr> " United-States", " United-States", " United-States", "~
## $ income         <chr> " <=50K", " <=50K", " <=50K", " <=50K", " <=50K", " <=5~
```

Task 1.1

```
# Remove specified columns
incomeData <- incomeData %>%
  select(-fnlwgt, -capital.gain, -capital.loss)
```

Task 1.2

```
# Fill missing values in 'native-country' column with 'others'
incomeData <- incomeData %>%
  mutate(native.country = ifelse(is.na(native.country), "others", native.country))
```

Task 2.1

```
# Calculate mean and median working hours by education level
education_working_hours <- incomeData %>%
  group_by(education) %>%
  summarize(working_mean = mean(hours.per.week), working_median = median(hours.per.week))

# Create a data frame with three columns: "education", "working_mean", and "working_median"
education_working_hours_table <- education_working_hours %>%
  select(education, working_mean, working_median)

# Print the resulting table
print(education_working_hours_table)
```

```
## # A tibble: 16 x 3
##    education       working_mean working_median
##    <chr>                  <dbl>          <dbl>
##  1 " 10th"                 37.1             40
##  2 " 11th"                 33.9             40
##  3 " 12th"                 35.8             40
##  4 " 1st-4th"              38.3             40
##  5 " 5th-6th"              38.9             40
##  6 " 7th-8th"              39.4             40
##  7 " 9th"                  38.0             40
##  8 " Assoc-acdm"           40.5             40
##  9 " Assoc-voc"            41.6             40
## 10 " Bachelors"            42.6             40
## 11 " Doctorate"            47.0             45
## 12 " HS-grad"              40.6             40
## 13 " Masters"              43.8             40
## 14 " Preschool"            36.6             40
## 15 " Prof-school"          47.4             48
## 16 " Some-college"         38.9             40
```

Is there any relation between education level and number of weekly working hours?

Observations:

Higher Education, Higher Hours (Mean): Generally, individuals with higher education levels tend to work more hours per week on average. For instance, individuals with education levels like 'Doctorate', 'Prof-school', and 'Masters' have higher mean working hours compared to those with lower education levels. This suggests that individuals with advanced degrees tend to work longer hours on average.

Consistent Median Hours: Across various education levels, the median working hours are consistently around 40 hours per week. This indicates that, regardless of education level, a significant portion of individuals work standard full-time hours.

Exceptions: There are a few exceptions. For instance, individuals with 'HS-grad' (high school graduates) have a mean working hours of 40.58, which is similar to the median, indicating a consistent workweek. On the other hand, individuals with 'Some-college' education have a lower mean working hours (38.85) compared to the median, indicating potential variability in working hours within this group.

In summary, while the mean working hours vary across education levels, the median working hours remain consistent around 40 hours per week. Higher education levels tend to correlate with higher mean working hours, suggesting a relationship between advanced education and longer workweeks on average. However, it's important to note that individual circumstances, job types, and industries can influence these patterns.

Task 2.2

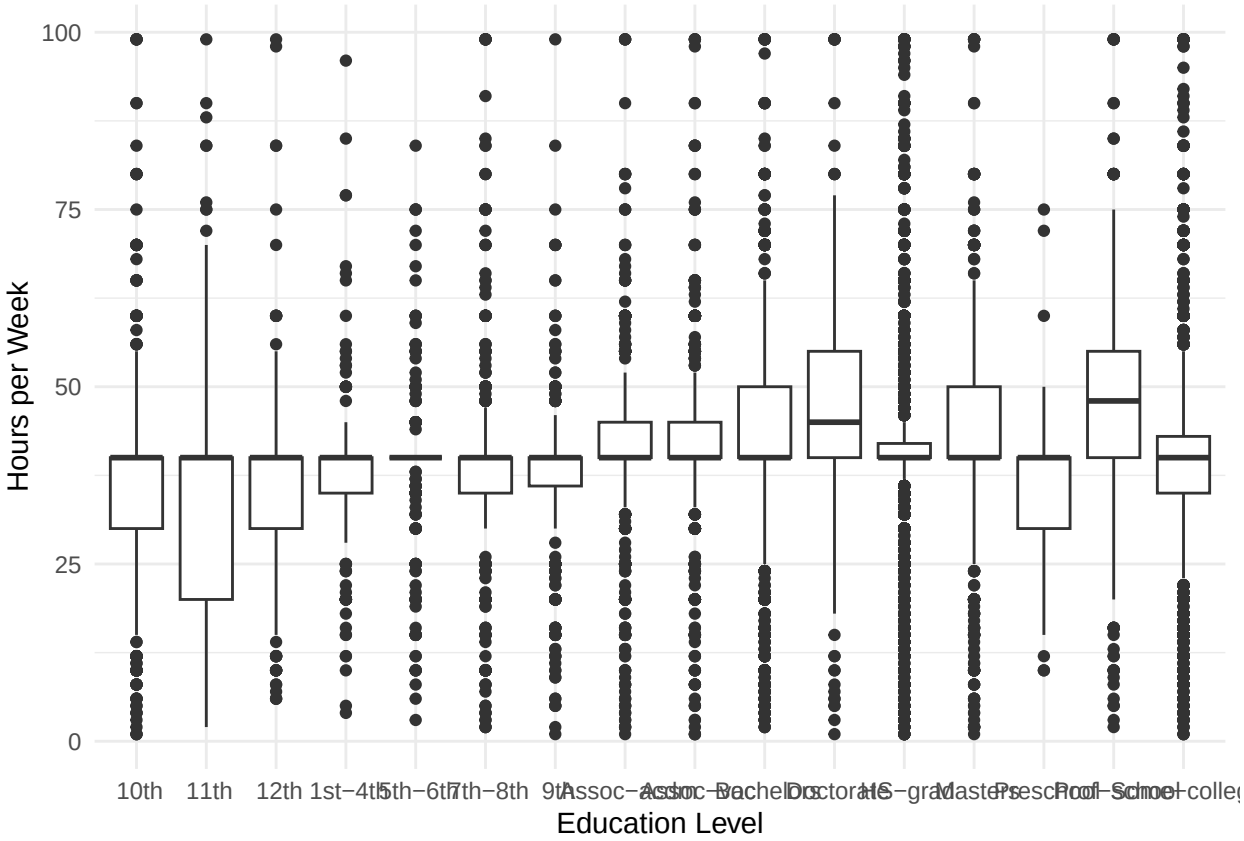
```
# Task 2.2: Total Number of Individuals by Native Country
native_country_counts <- incomeData %>%
  group_by(native.country) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

# Print the first 5 countries with the highest number of individuals
print(head(native_country_counts, 5))
```

```
## # A tibble: 5 x 2
##   native.country   count
##   <chr>            <int>
## 1 " United-States" 29170
## 2 " Mexico"          643
## 3 ""                 583
## 4 " Philippines"     198
## 5 " Germany"         137
```

Task 3.1

```
# Box plot for 'hours-per-week' by education level
ggplot(incomeData, aes(x = education, y = hours.per.week)) +
  geom_boxplot() +
  labs(x = "Education Level", y = "Hours per Week") +
  theme_minimal()
```

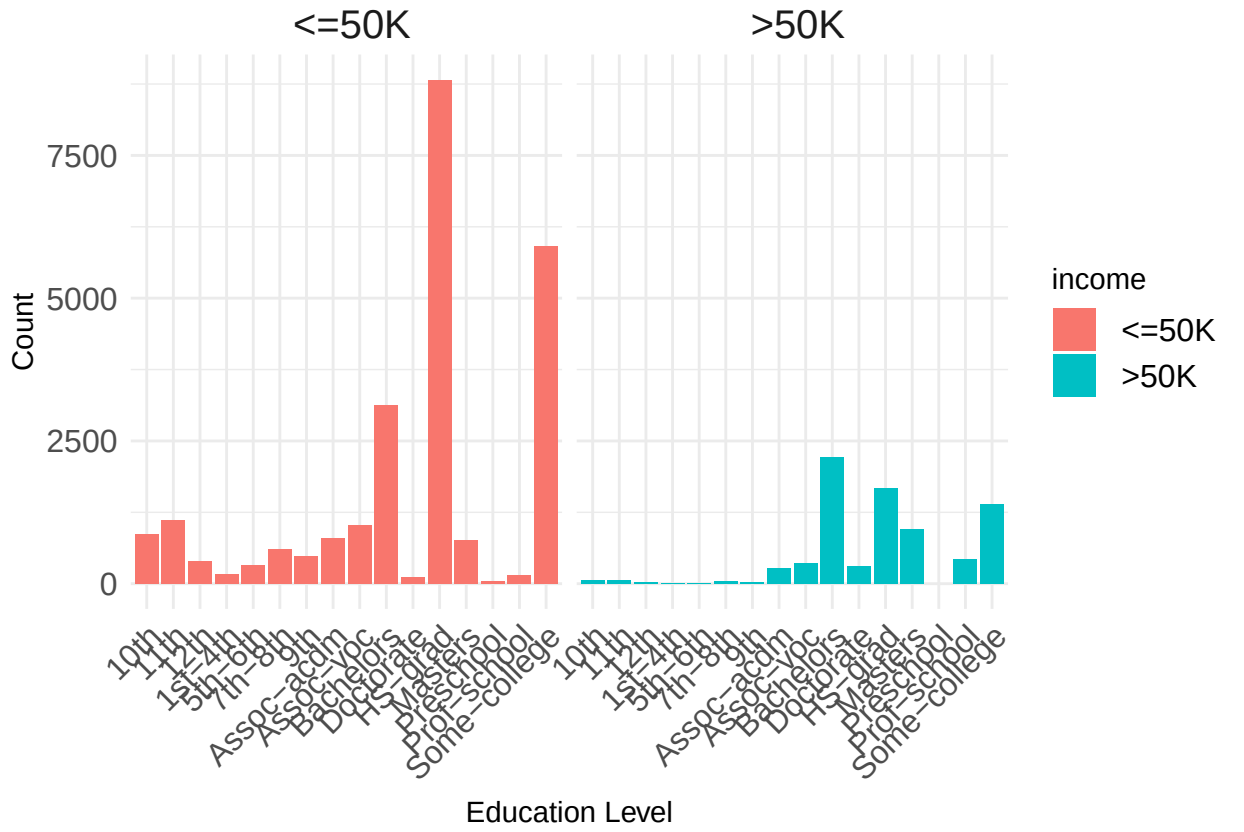


What insights can you draw from this plot?

Most education levels have the same range of working hours however Assoc-acdm, Assoc-voc, Bachelors, Doctorate, Masters, and Prof-school have higher working hours, with Doctorate and Prof-school going over 50 hours. Also, the highest median weekly working hours are observed for individuals with a degree in Prof-School. HS-grad shows the highest number of outliers in the graph, indicating significant variability in weekly working hours among high school graduates. The lower bound of the graph corresponds to individuals with an education level of 11th grade, indicating the minimum weekly working hours observed within this category.

Task 3.2

```
# Faceted bar chart for count of individuals by education level and income
ggplot(incomeData, aes(x = education, fill = income)) +
  geom_bar() +
  facet_wrap(~income) +
  labs(x = "Education Level", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels for better readabili
        plot.title = element_text(size = 20),  # Increase plot title size
        strip.text = element_text(size = 15),  # Increase facet labels size
        axis.text = element_text(size = 12),  # Increase axis text size
        legend.text = element_text(size = 12))  # Increase legend text size
```

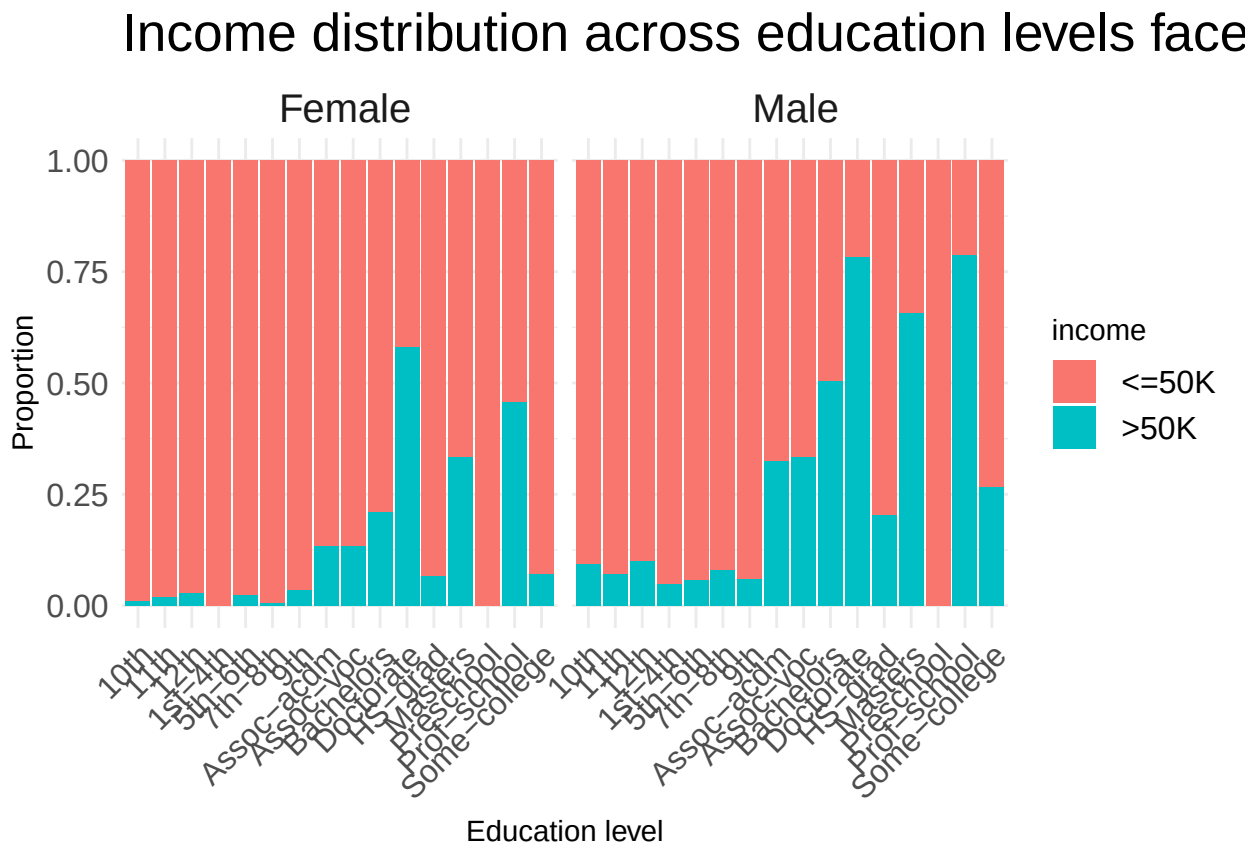


What insights can you draw from these facetted plots?

Most of the education levels tend to fall under the catergory earning less thank 50k, except for Masters, prof-school, and doctorate where there were almost equally the same under both categories. HS-grad has the highest count, exceeding 7500 individuals, followed by some-college in terms of frequency.

Task 3.3

```
ggplot(incomeData, aes(x = education, fill = income)) +
  geom_bar(position = "fill") +
  facet_wrap(~ sex, nrow = 1) +
  xlab("Education level") +
  ylab("Proportion") +
  ggtitle("Income distribution across education levels faceted by gender") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels for better readabili
        plot.title = element_text(size = 20),  # Increase plot title size
        strip.text = element_text(size = 15),  # Increase facet labels size
        axis.text = element_text(size = 12),  # Increase axis text size
        legend.text = element_text(size = 12))  # Increase legend text size
```



What does the chart reveal about income distribution across education levels?

More males tend to fall under the category that earn more than 50k compared to women even when they have the same eqduaction level. As the level of education increases, the income tends to rise, with a higher proportion exceeding the $50,000 threshold.

Task 4.1

```
# Create 'work_hours_per_year' column
incomeData <- incomeData %>%
  mutate(work_hours_per_year = hours.per.week * 52)
```

Task 5.1

```
# Read country income data
countryIncome <- read.csv("data/country_income.csv")
```

Task 5.2

```r
# Create 'income_category' column
countryIncome <- countryIncome %>%
  mutate(income_category = ifelse(avg_income <= 50000, '<=50K', '>50K'))
```

```r
glimpse(incomeData)
```

```
## Rows: 32,561
## Columns: 13
## $ age                 <int> 39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30, 23~
## $ workClass           <chr> " State-gov", " Self-emp-not-inc", " Private", " P~
## $ education           <chr> " Bachelors", " Bachelors", " HS-grad", " 11th", "~
## $ education.num       <int> 13, 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13, 12~
## $ marital.status      <chr> " Never-married", " Married-civ-spouse", " Divorce~
## $ occupation          <chr> " Adm-clerical", " Exec-managerial", " Handlers-cl~
## $ relationship        <chr> " Not-in-family", " Husband", " Not-in-family", " ~
## $ race                <chr> " White", " White", " White", " Black", " Black", ~
## $ sex                 <chr> " Male", " Male", " Male", " Male", " Female", " F~
## $ hours.per.week      <int> 40, 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40, 30~
## $ native.country      <chr> " United-States", " United-States", " United-State~
## $ income              <chr> " <=50K", " <=50K", " <=50K", " <=50K", " <=50K", ~
## $ work_hours_per_year <dbl> 2080, 676, 2080, 2080, 2080, 2080, 832, 2340, 2600~
```

```r
glimpse(countryIncome)
```

```
## Rows: 206
## Columns: 3
## $ Country         <chr> "Afghanistan", "Albania", "Algeria", "Andorra", "Angol~
## $ avg_income      <int> 1746, 12300, 13639, 48641, 5555, 22201, 17611, 9277, 4~
## $ income_category <chr> "<=50K", "<=50K", "<=50K", "<=50K", "<=50K", "<=50K", ~
```

Task 5.3

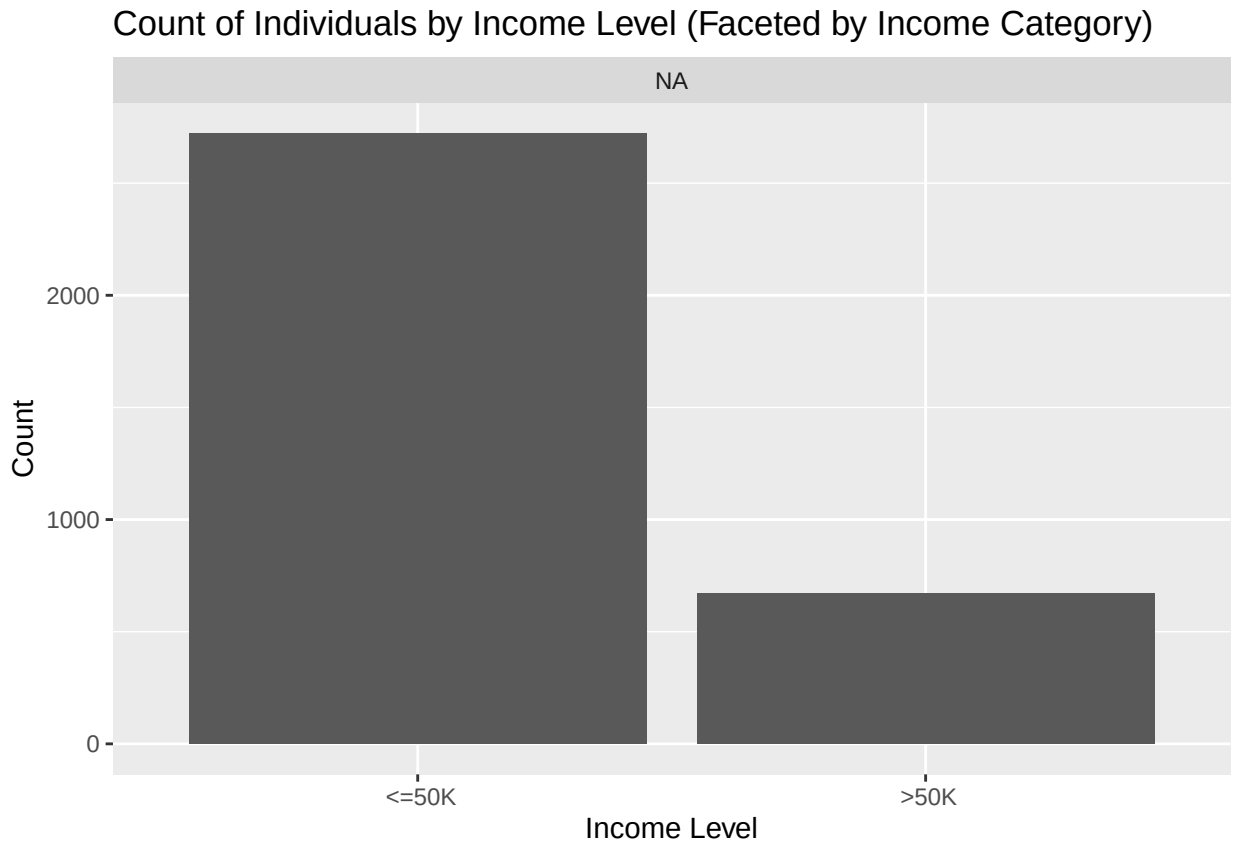
```r
# Merge 'incomeData' and 'country_income' data frames using 'native_country'
incomeData <- left_join(incomeData, countryIncome, by = c("native.country" = "Country"))
```

Task 5.4

```r
# Remove observations where 'native_country' is 'United-States' or NA
incomeData <- incomeData %>%
  filter(!is.na(native.country) & native.country != " United-States")
```

Task 5.5

```r
# Faceted Bar Chart for Count of Individuals by Income Level and Income Category
ggplot(incomeData, aes(x = income)) +
  geom_bar() +
  facet_wrap(~ income_category, nrow = 1) +
  xlab("Income Level") +
  ylab("Count") +
  ggtitle("Count of Individuals by Income Level (Faceted by Income Category)")
```



What insights can you draw from these facetted plots?

Only a very few percatnages of people earn more than 50k