Lab 2

Ray Will

2023-10-22

```
#Load the libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(descr)
library(forcats)
library(ggplot2)
#view the data
?gss_cat
gss_cat
```

```
## # A tibble: 21,483 x 9
##     year marital         age race  rincome        partyid    relig denom tvhours
##    <int> <fct>         <int> <fct> <fct>          <fct>      <fct> <fct>   <int>
##  1  2000 Never married    26 White $8000 to 9999  Ind,near ~ Prot~ Sout~      12
##  2  2000 Divorced         48 White $8000 to 9999  Not str r~ Prot~ Bapt~      NA
##  3  2000 Widowed          67 White Not applicable Independe~ Prot~ No d~       2
##  4  2000 Never married    39 White Not applicable Ind,near ~ Orth~ Not ~       4
##  5  2000 Divorced         25 White Not applicable Not str d~ None  Not ~       1
##  6  2000 Married          25 White $20000 - 24999 Strong de~ Prot~ Sout~      NA
##  7  2000 Never married    36 White $25000 or more Not str r~ Chri~ Not ~       3
##  8  2000 Divorced         44 White $7000 to 7999  Ind,near ~ Prot~ Luth~      NA
##  9  2000 Married          44 White $25000 or more Not str d~ Prot~ Other       0
## 10  2000 Married          47 White $25000 or more Strong re~ Prot~ Sout~       3
## # i 21,473 more rows
```

research question: how marital status relate to the income

Part 2

Data Cleaning and Manipulation

```
# Select the two columns
gss_subset <- gss_cat %>% select(marital, rincome)
```

```r
# Remove rows with missing values
gss_subset <- na.omit(gss_subset)
gss_subset
```

```
## # A tibble: 21,483 x 2
##    marital       rincome
##    <fct>         <fct>
##  1 Never married $8000 to 9999
##  2 Divorced      $8000 to 9999
##  3 Widowed       Not applicable
##  4 Never married Not applicable
##  5 Divorced      Not applicable
##  6 Married       $20000 - 24999
##  7 Never married $25000 or more
##  8 Divorced      $7000 to 7999
##  9 Married       $25000 or more
## 10 Married       $25000 or more
## # i 21,473 more rows
```

convert data to numerical

```r
 gss_subset$rincome <- as.numeric(gss_subset$rincome)
```

#explore the data

```r
summary(gss_subset)
```

```
##           marital         rincome
##  No answer    :   17   Min.   : 1.00
##  Never married: 5416   1st Qu.: 4.00
##  Separated    :  743   Median : 6.00
##  Divorced     : 3383   Mean   : 8.93
##  Widowed      : 1807   3rd Qu.:16.00
##  Married      :10117   Max.   :16.00
```

summary statistics

```r
summary_stats <- gss_subset %>%
  group_by(marital) %>%
  summarize(mean_income = mean(rincome), median_income = median(rincome))
```

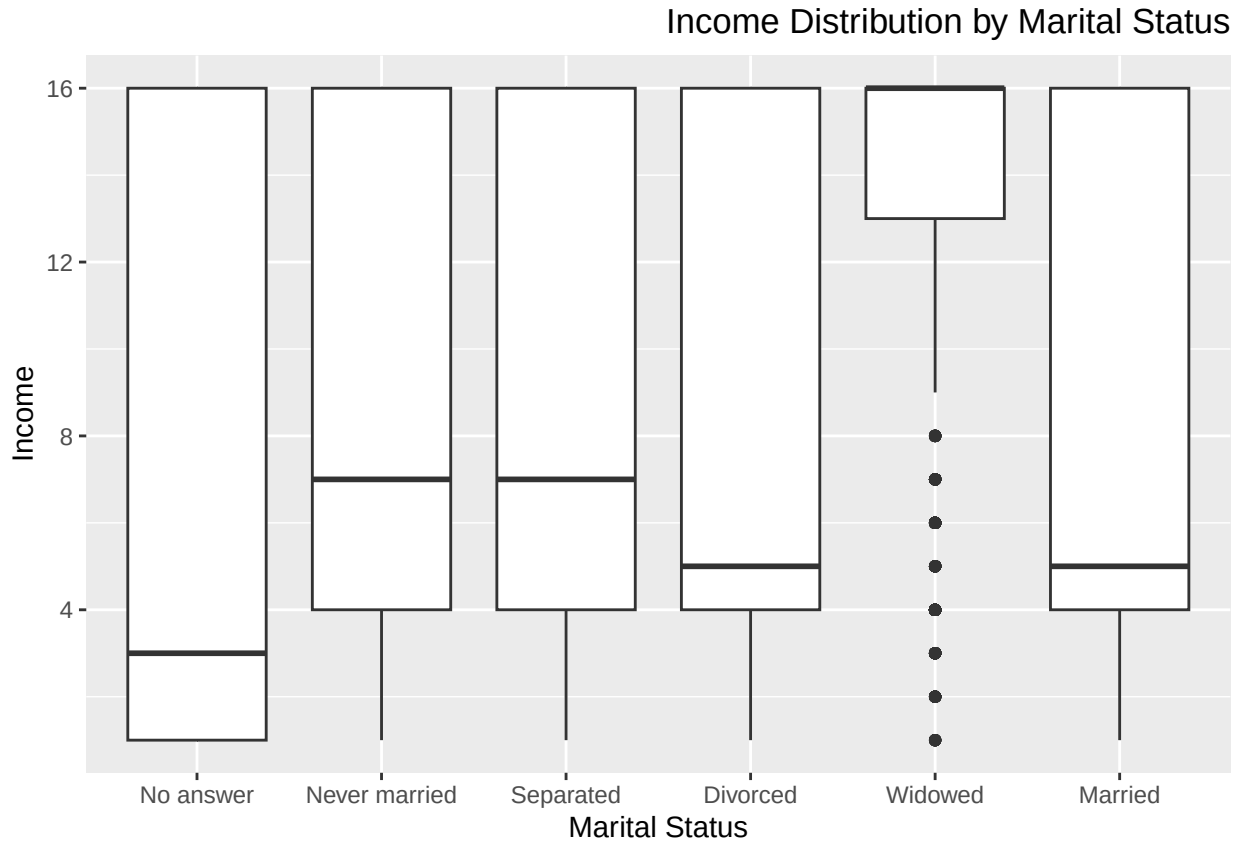
show summary statistics

```r
summary_stats
```

```
## # A tibble: 6 x 3
##   marital       mean_income median_income
##   <fct>               <dbl>         <dbl>
## 1 No answer            6.47             3
## 2 Never married        8.91             7
## 3 Separated            9.25             7
## 4 Divorced             8.39             5
## 5 Widowed             13.3             16
## 6 Married              8.32             5
```

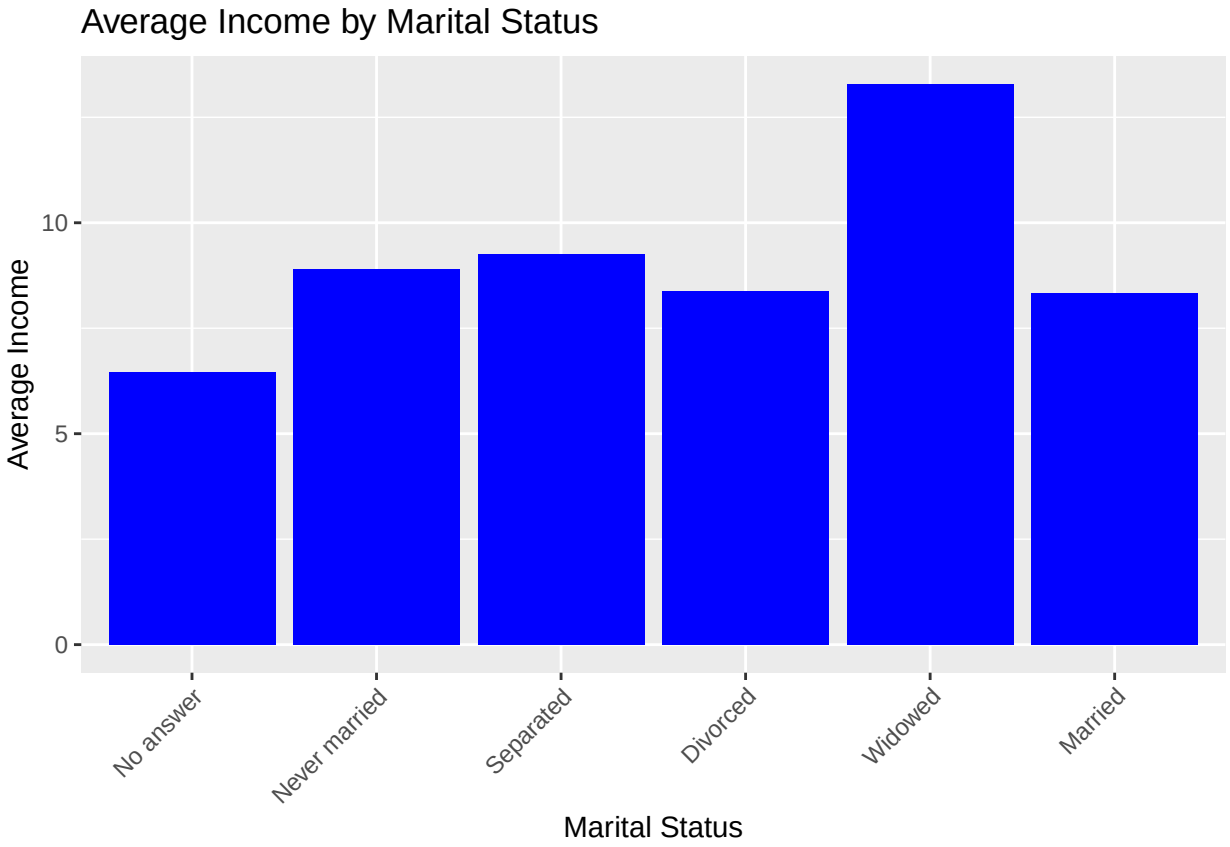
Create a boxplot with an adjusted title

```
ggplot(gss_subset, aes(x = marital, y = rincome)) +
  geom_boxplot() +
  labs(
    title = "Income Distribution by Marital Status",
    x = "Marital Status",
    y = "Income"
  ) +
  theme(plot.title = element_text(hjust =1))
```



Create a bar plot

```
ggplot(gss_subset, aes(x = marital, y = rincome)) +
  geom_bar(stat = "summary", fun = "mean", fill = "blue") +
  labs(
    title = "Average Income by Marital Status",
    x = "Marital Status",
    y = "Average Income"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Average Income by Marital Status
Average Income
10
5
0
No answer
Never married
Separated
Divorced
Widowed
Married
Marital Status

Part 1 Reports

Report 1.1

What data set have you decided to use? gss_cat

Report 1.2

Which two variables from your data set will be analyzed? marital vs rincome

Report 1.3* What is your research question? **how marital status relate to the level of income**

Report 1.4 What is your data analysis plan? Please be descriptive. -Load the "gss_cat" dataset ,Check for missing values and outliers in the "rincome" variable. -Calculate descriptive statistics for the "rincome" variable, such as mean, median, and standard deviation. -Create visualizations, such as histograms, box plots, or bar plots, to understand the distribution of income by marital status. -Interpret the results in the context of your research question, explaining the implications of the findings. -Document the findings in a structured report or presentation. -Use visual summaries and tables to highlight key results and insights. **What are some potential limitations for your analysis?** -Causation vs. correlation The analysi show a correlation between marital status and income but cannot prove causation

Part 2 Reports

Report 1.6 Does your data contain missing values? If so, how have you dealt with these values? **Yes** #any(is.na(gss_cat))

Report 1.7 Please include all code used to clean and manipulate the variables. **Code Used**

any(is.na(gss_cat)) #Data Cleaning and Manipulation # Select the two columns gss_subset <- gss_cat %>% select(marital, rincome) # Remove rows with missing values gss_subset <- na.omit(gss_subset) gss_subset #convert data to numerical gss_subset$rincome <- -as.numeric(gss_subset$rincome) #explore the data

Report 1.8 ##What relationship, if any, exists between the two variables? -"Widowed" have a higher average income compared to the other marital status categories. -This observation is based on the "Mean" income and the visualizations. **Report 1.9** === How do these findings relate to your research question and theory? -The findings are directly related to my research question, as they provide initial insights into how different marital status categories are associated with income levels. The "Widowed" have a higher average income while the no answer group has the lowest average income

Report 1.10

What limitations exist as a result of the data analysis? -The presence of missing values in the dataset limited the scope of may have introduced bias if not handled appropriately. -The analysis show a correlation between marital status and income, but it does not prove causation.