MSDS 660 Week 1 Assignment

Jeremy Beard

2022-07-06

Introduction In this week's assignment, we will be continuing to explore the WVS dataset and trying to find correlations between variables. I wanted to continue to explore what is correlated with higher income levels and as such, I found variables which indicated priorities in life, and how much weight one places on work vs. leisure. I chose the following two (2) hypotheses to explore in the analysis: 1. A person who prioritizes thrift/saving money more will have a higher income 2. A person who prioritizes work more has a higher income Through my analysis, I found that prioritizing determination above all else is correlated with a higher income level, NOT thrift. I also found that significantly prioritizing work over leisure was correlated with a lower income level. This type of analysis is important because it teaches us the basics of exploratory data analysis and finding associations between variables in a dataset. At a more functional level, it can show us what can be shown to be correlated with a higher income level! This could be profitable in the future.

```
# 20220706
# MSDS660
# Week 1 - Homework Assignment
# Jeremy Beard

# To start, I will list what variables I will focus on for the assignment

# V10: happiness
# V25: 1st priority for children to learn
# V72: work vs. leisure scale
# V227: income
# V228: tv

# V10 Taking all things together, would you say you are:
# 1 Very happy
# 2 Quite happy
# 3 Not very happy
# 4 Not at all happy
# 9 Don't know

# V25 Here is a shorter list of things that children can be encouraged to learn. If you had
# to choose, which one of these do you consider to be the most important thing for a child
# to learn at home?
# 1=Thrift, saving money and things
# 2=Obedience
# 3=Determination, perseverence
# 4=Religious faith
# 9=Don't know

# V72. Which point on this scale most clearly describes how much weight you place on
# work (including housework and school work), as compared with leisure or recreation?
# 1. It's leisure that makes life worth living, not work
```

```
# 2.
# 3.
# 4.
# 5. Work is what makes life worth living, not leisure
# 9. DK

# V227 Here is a scale of incomes. We would like to know in what group your household
# is, counting all wages, salaries, pensions and other incomes that come in. Just give
# the letter of the group your household falls into, before taxes and other deductions.
# 1 2 3 4 5 6 7 8 9 10
# C D E F G H I J K L
# No answer = 98
# [CODE INCOME CATEGORIES BY DECILES FOR YOUR SOCIETY,
# 1=LOWEST DECILE, 10=HIGHEST DECILE]


##### HYPOTHESIS
# 1. A person who prioritizes thrift/saving money more will have a higher income
# 2. A person who prioritizes work more has a higher income

##### Setup, Reading data

setwd("C:\\Users\\jerem\\OneDrive\\Documents\\School\\_REGIS\\2022-05_Summer\\MSDS660\\Week1") #this wi

# read in our .rdata file
load("WorldValues_Data.rdata")

# load necessary packages
library(data.table)
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

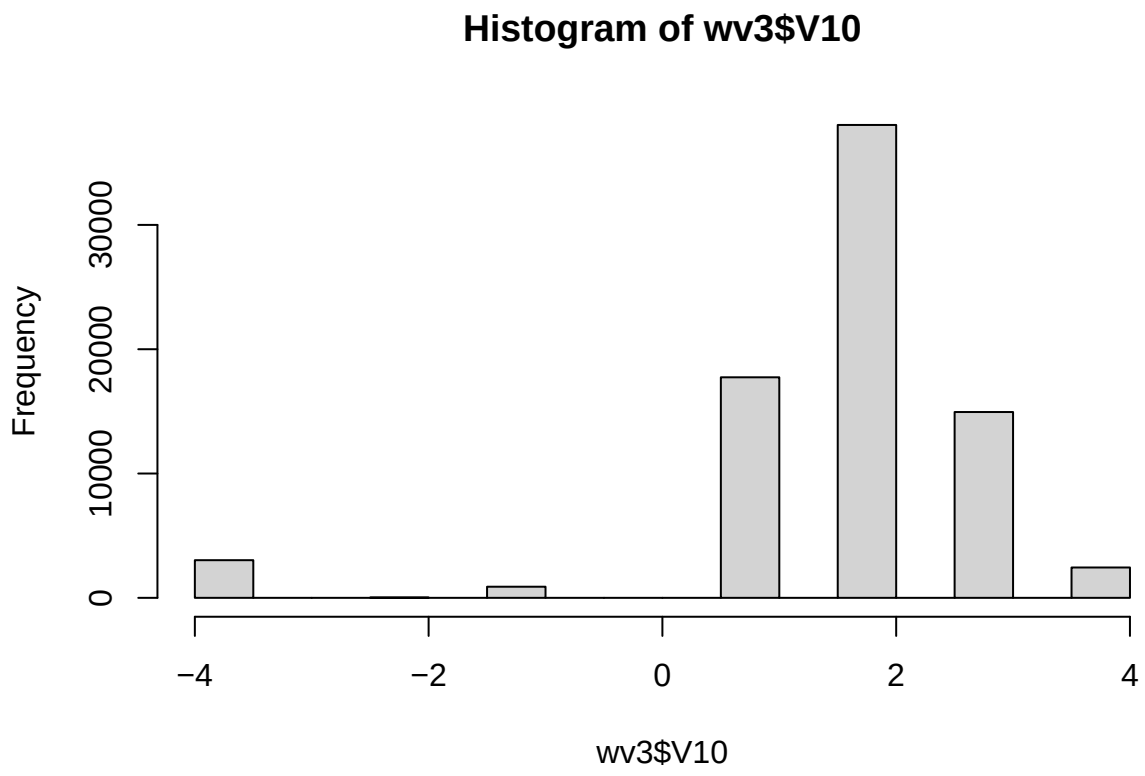
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

# this converts to a data.table, which is easier
# to work with than other formats in R
wv3 = data.table(`WV3_Data_R_v_2015-04-18`)
rm("WV3_Data_R_v_2015-04-18")
```

Methods For this week's assignment, I will not be using specific statistical tests but rather displaying correlations between variables in a dataset. This will involve loading the data into R, cleaning the data, and

then plotting the correlations. For this specific assignment, cleaning the data involved merely eliminating the negative values as those values were not valid responses in the WVS survey. In addition to cleaning the data, I provided text labels for the variables so they can be easily recognizable. I used statistical summaries to confirm that the data cleaning had indeed worked the way I intended it to.

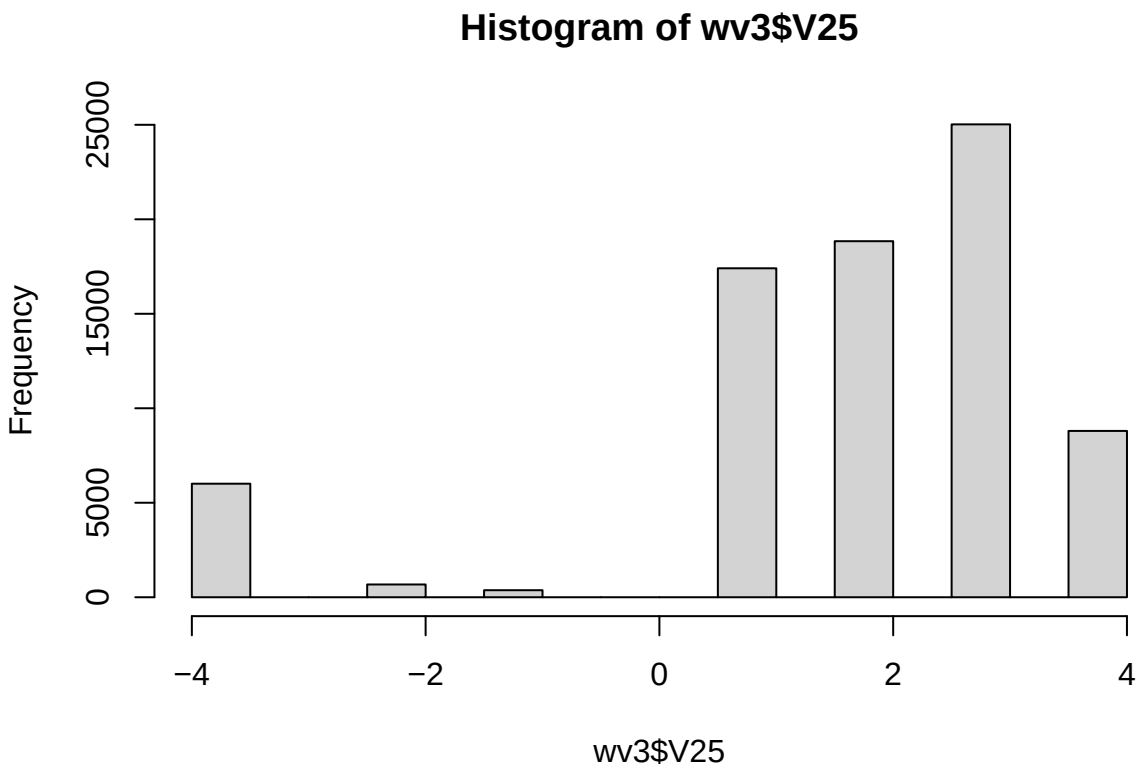
```
##### EDA
# V10 (happiness)
hist(wv3$V10)
```




```
summary(wv3$V10)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.000   1.000   2.000   1.755   2.000   4.000
```

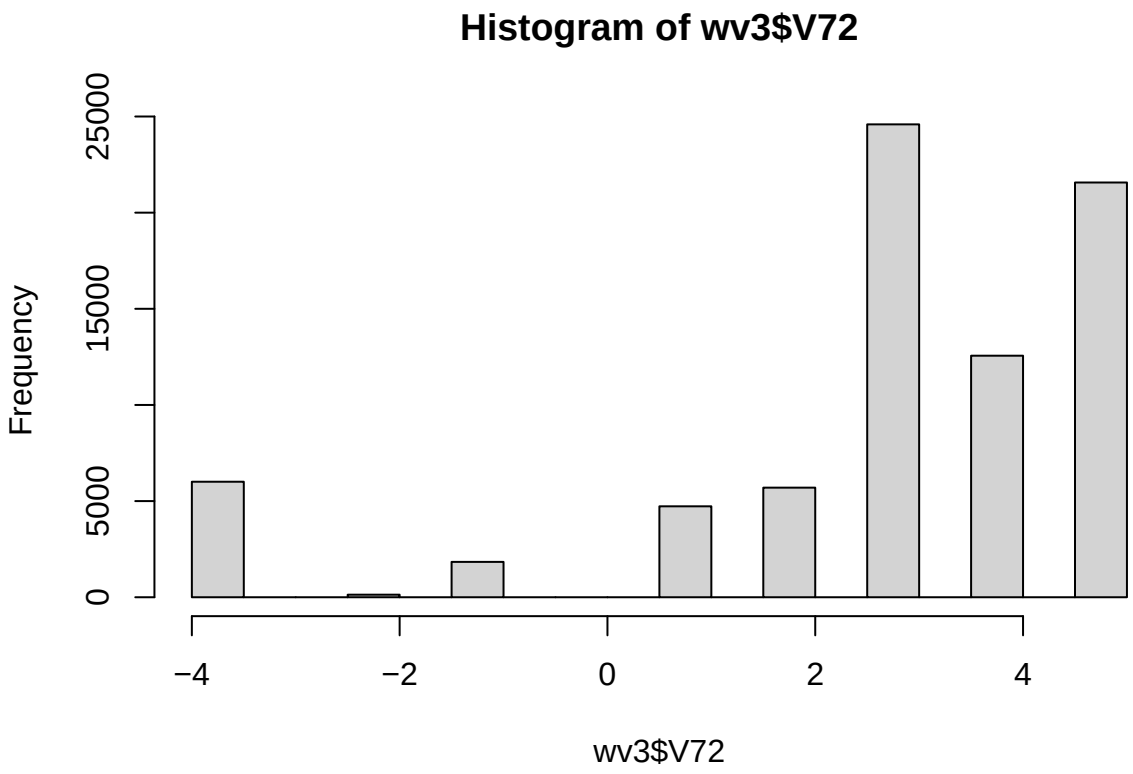
```
# NOTE: there are negative values which must be cleaned

# V25 (priority)
hist(wv3$V25)
```




```
summary(wv3$V25)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -4.00    1.00    2.00    1.81    3.00    4.00
```

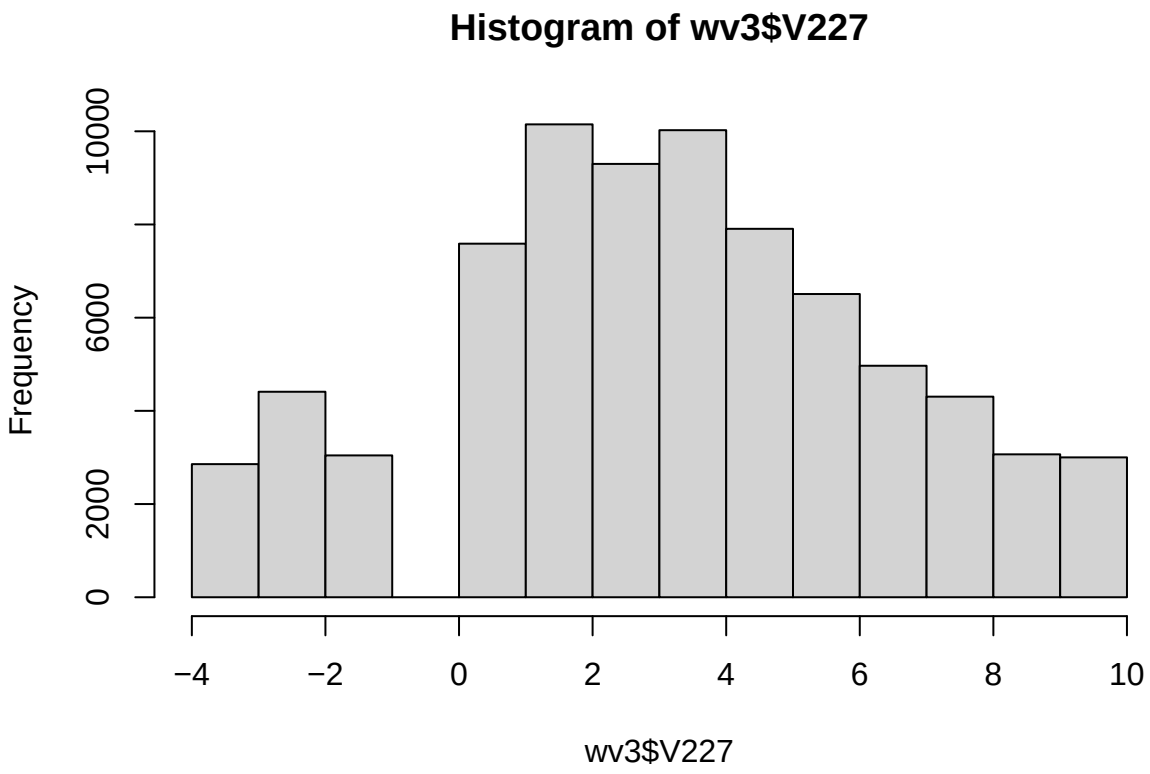
```
# NOTE: there are negative values which must be cleaned

# V72 (work priority)
hist(wv3$V72)
```




```
summary(wv3$V72)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.000   3.000   3.000   2.876   5.000   5.000
```

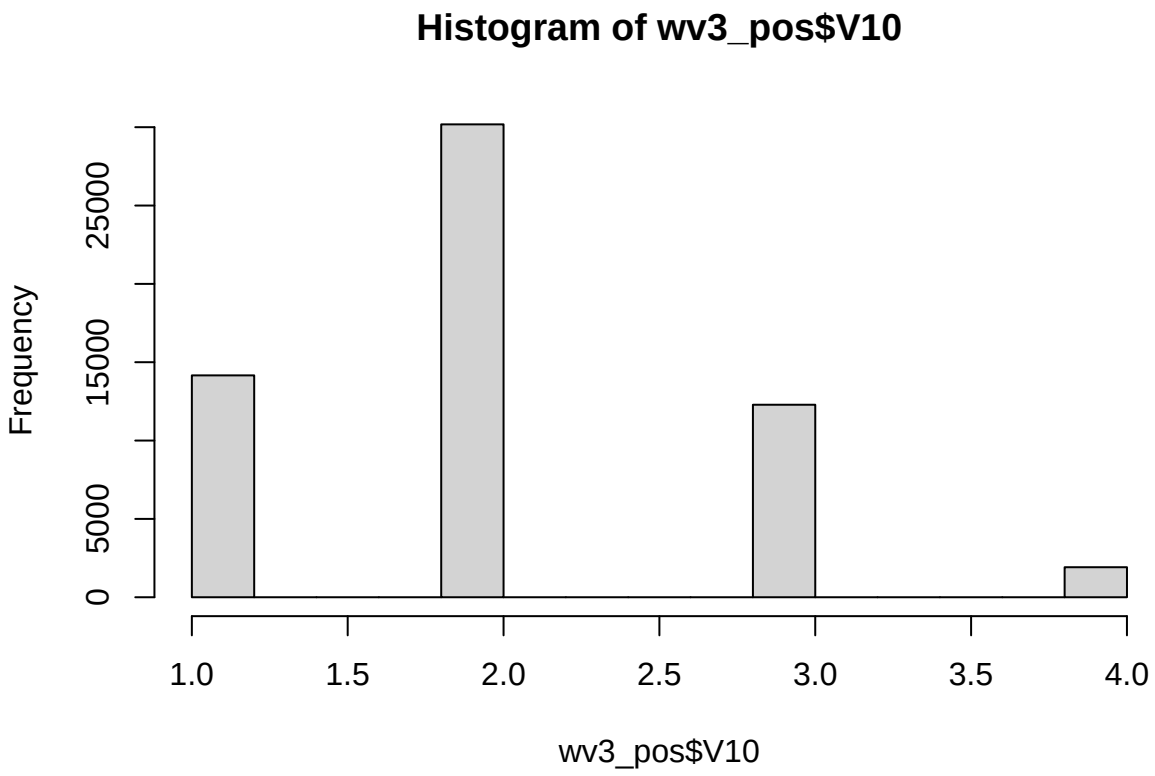
```
# NOTE: there are negative values which must be cleaned

# V227 (income)
hist(wv3$V227)
```




```
summary(wv3$V227)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.000   2.000   4.000   3.605   6.000  10.000
```

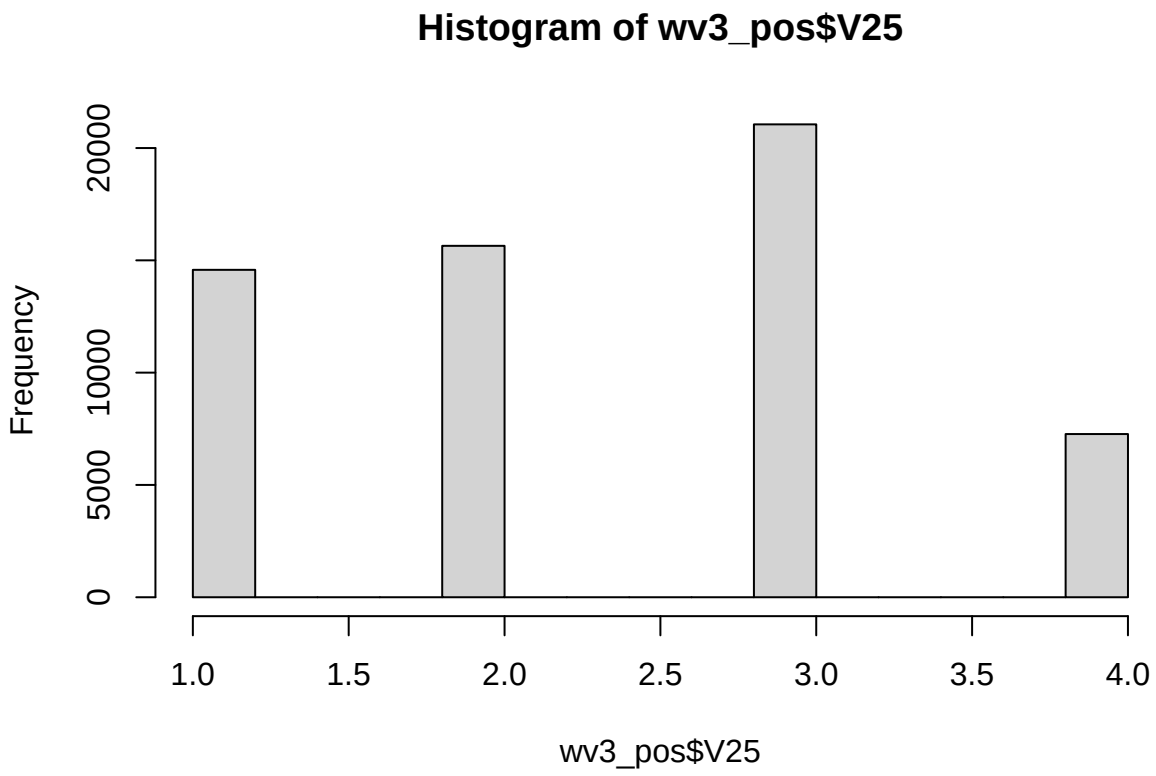
```
# NOTE: there are negative values which must be cleaned

##### PROBLEMS OR CONCERNS WITH THE DATA
# From the summary, we can see that the data has some negative values
# These negative values do not correspond to any responses from the survey

##### DATA CLEANING
# Let's remove the negative values
wv3_pos <- wv3[V10 > 0 & V25 > 0 & V72 > 0 & V227 > 0]

# Let's see how the distributions changed
# V10 (happiness)
hist(wv3_pos$V10)
```




```
summary(wv3_pos$V10)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.033   2.000   4.000
```

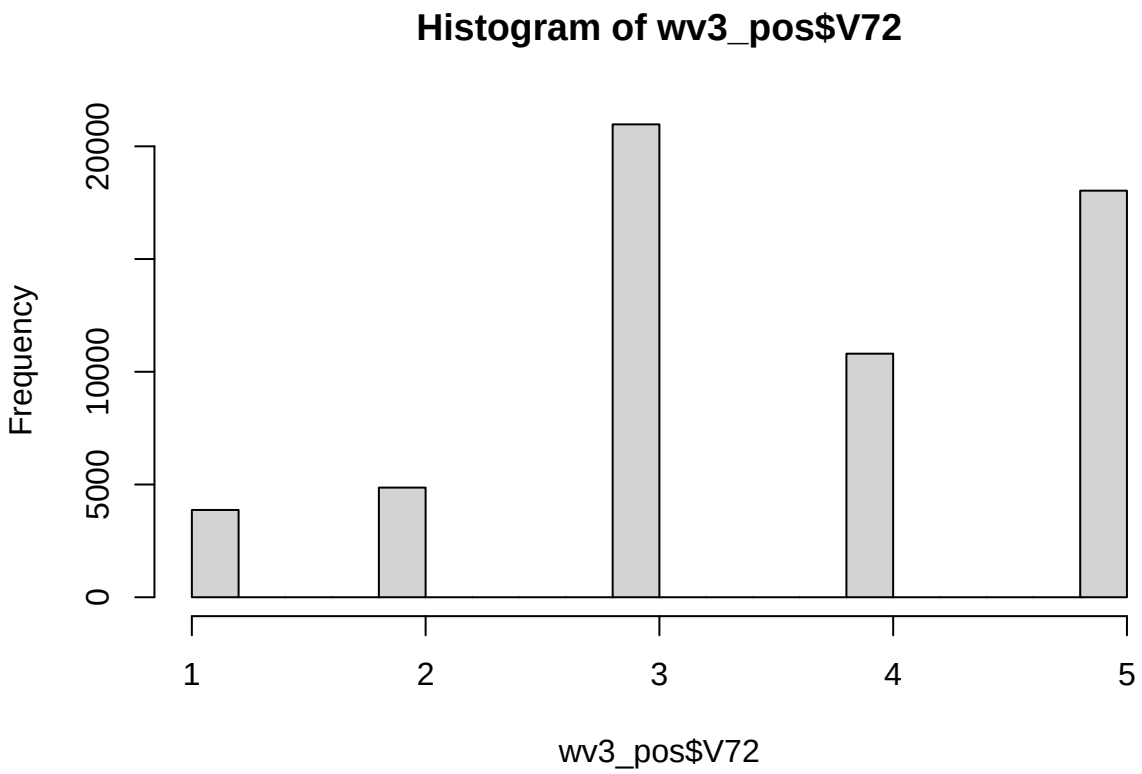
```
# V25 (priority)
hist(wv3_pos$V25)
```




```
summary(wv3_pos$V25)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.359   3.000   4.000
```

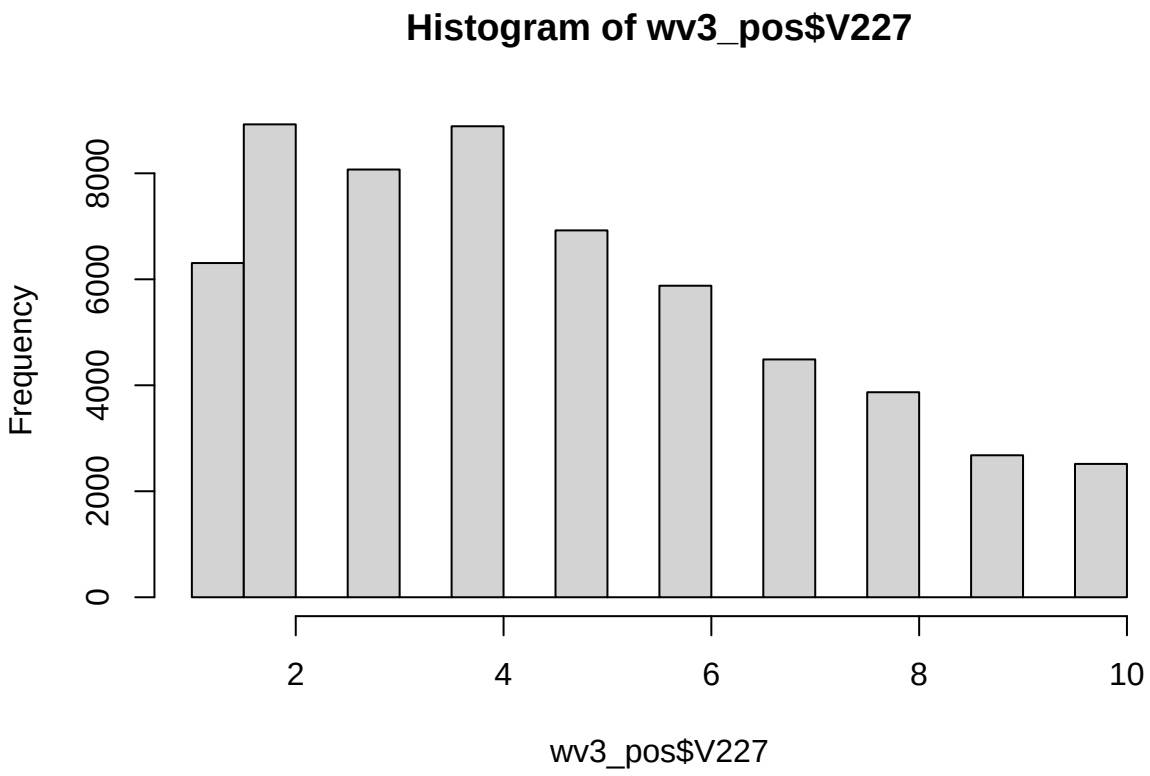
```
# V72 (work priority)
hist(wv3_pos$V72)
```




```
summary(wv3_pos$V72)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.585   5.000   5.000
```

```
# V227 (income)
hist(wv3_pos$V227)
```



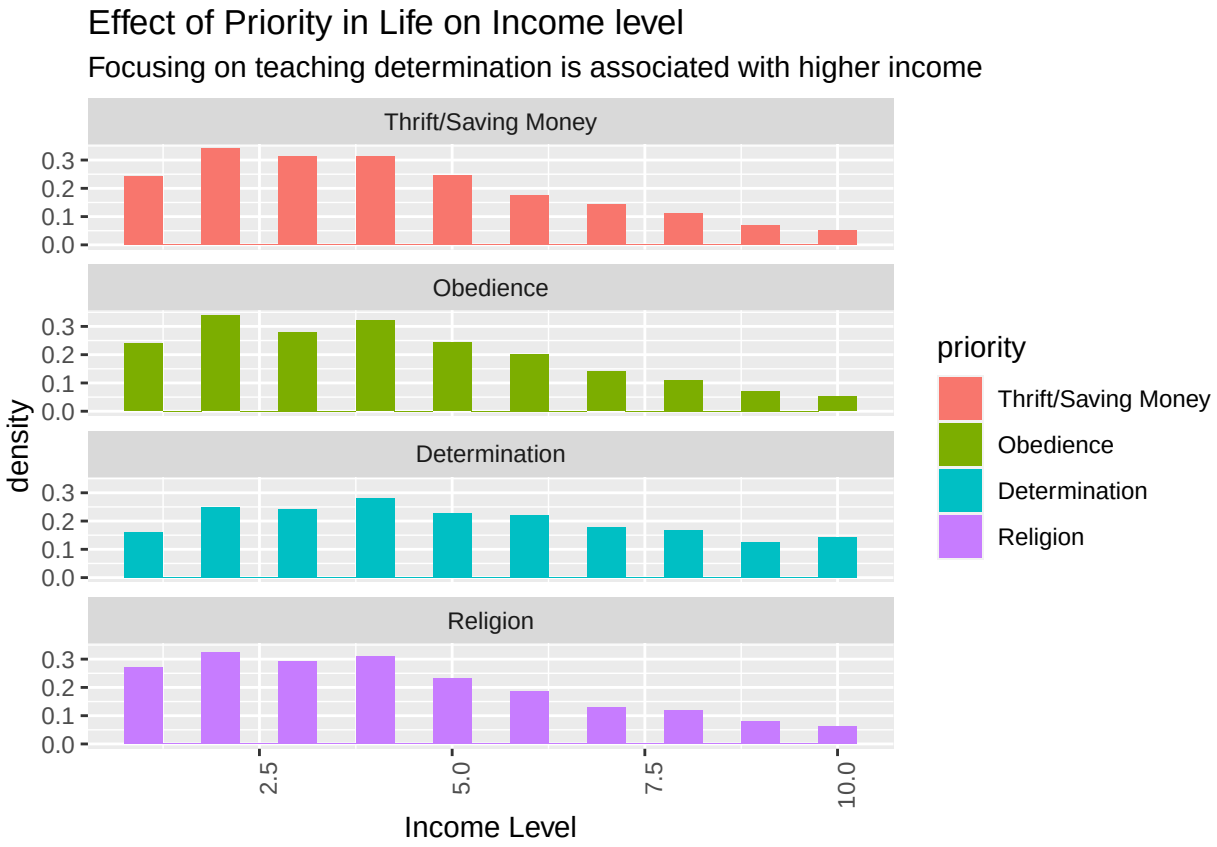

```
summary(wv3_pos$V227)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   4.534   6.000  10.000
```

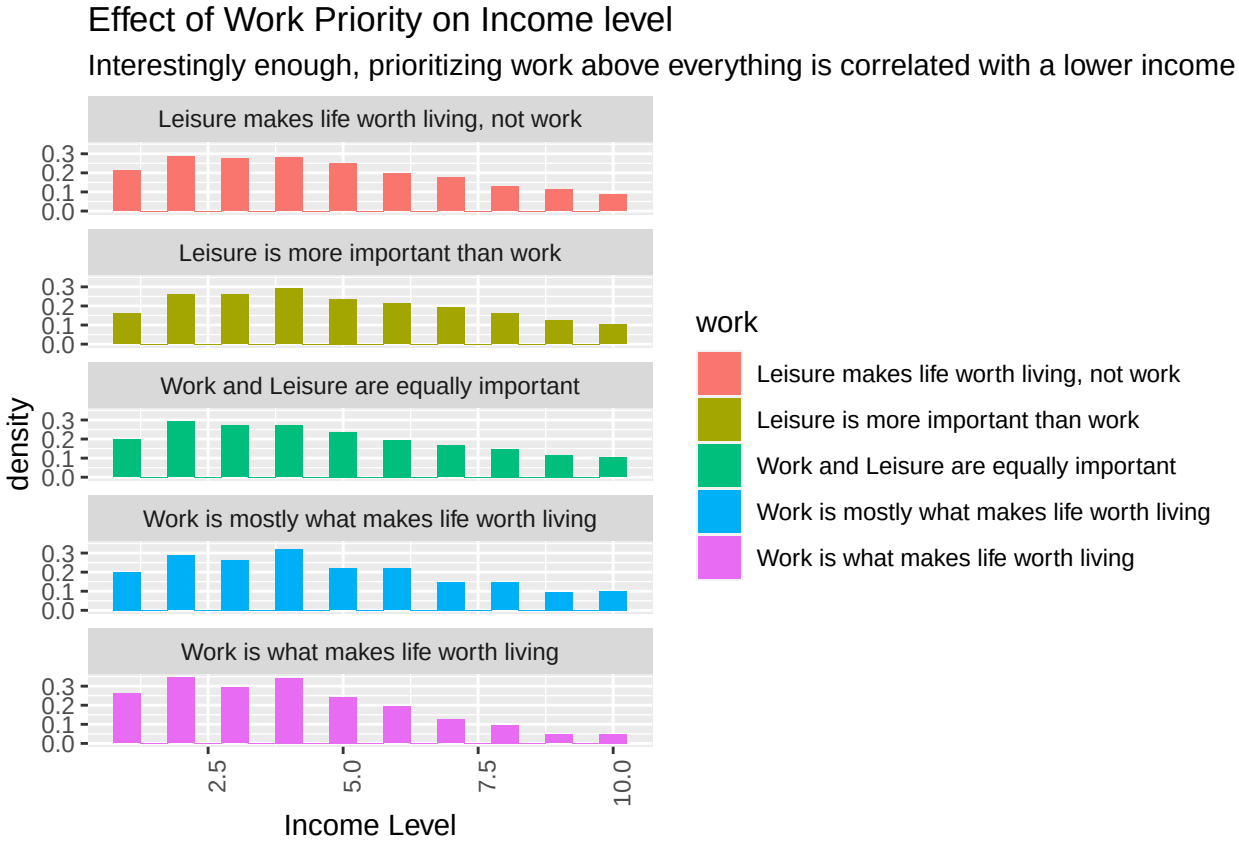
```
# Looks much better now, no more erroneous data

# Let's use the as.factor function to provide labels to the variables
wv3_pos$happy <- as.factor(wv3_pos$V10)
wv3_pos$priority <- as.factor(wv3_pos$V25)
wv3_pos$work <- as.factor(wv3_pos$V72)
wv3_pos$income <- as.factor(wv3_pos$V227)

# Let's give labels to the amount of TV people watch
wv3_pos$priority <- plyr::revalue(wv3_pos$priority,
                                  c("1" = "Thrift/Saving Money",
                                    "2" = "Obedience",
                                    "3" = "Determination",
                                    "4" = "Religion"))

wv3_pos$work <- plyr::revalue(wv3_pos$work,
                              c("1" = "Leisure makes life worth living, not work",
                                "2" = "Leisure is more important than work",
                                "3" = "Work and Leisure are equally important",
                                "4" = "Work is mostly what makes life worth living",
                                "5" = "Work is what makes life worth living"))
```

Results Below I show how I plotted the correlation between life priority and income level, and also work priority and income level. I use ggplot to show multiple plots on the same chart. I also title the axes and give a title to the entire chart, in addition to a subtitle.

```
# Now let's plot the correlations between life priorities, priority of work, and income
ggplot(data=wv3_pos, aes(x=V227, fill=priority)) +
  geom_histogram(aes(y=..density..), binwidth=0.5) +
  facet_wrap(~priority, nrow=4) +
  #scale_x_continuous(breaks = 1:4, labels=c("1" = "Very happy", "2" = "Quite happy",
  #                                          "3" = "Not very happy", "4" = "Not at all happy")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = 'Income Level',
       title = 'Effect of Priority in Life on Income level' ,
       subtitle = 'Focusing on teaching determination is associated with higher income')
```




```
ggplot(data=wv3_pos, aes(x=V227, fill=work)) +
  geom_histogram(aes(y=..density..), binwidth=0.5) +
  facet_wrap(~work, nrow=5) +
  #scale_x_continuous(breaks = 1:4, labels=c("1" = "Very happy", "2" = "Quite happy",
  #                                          "3" = "Not very happy", "4" = "Not at all happy")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = 'Income Level',
       title = 'Effect of Work Priority on Income level' ,
       subtitle = 'Interestingly enough, prioritizing work above everything is correlated with a lower
```



Conclusion

In this analysis of the WVS data, I chose to analyze life priorities, and work priority, and how they both are correlated with income level. I found that both my hypotheses were incorrect! I hypothesized that wanting to teaching children thrift/saving money would be correlated most with a higher income, but this was actually false. Teaching children determination was most correlated with a higher income. This does make sense, determination would have been my second choice for being highly correlated with a higher income level.

Additionally, people who stated "work is what makes life worth living" were found to have lower income levels than any other category. This may be due to the fact that they choose a job which they love but which may have a lower income level, over another job which pays well but they hate.

Thanks! Jeremy B