

# STAT 531 Assignment 1

Chloe Moore

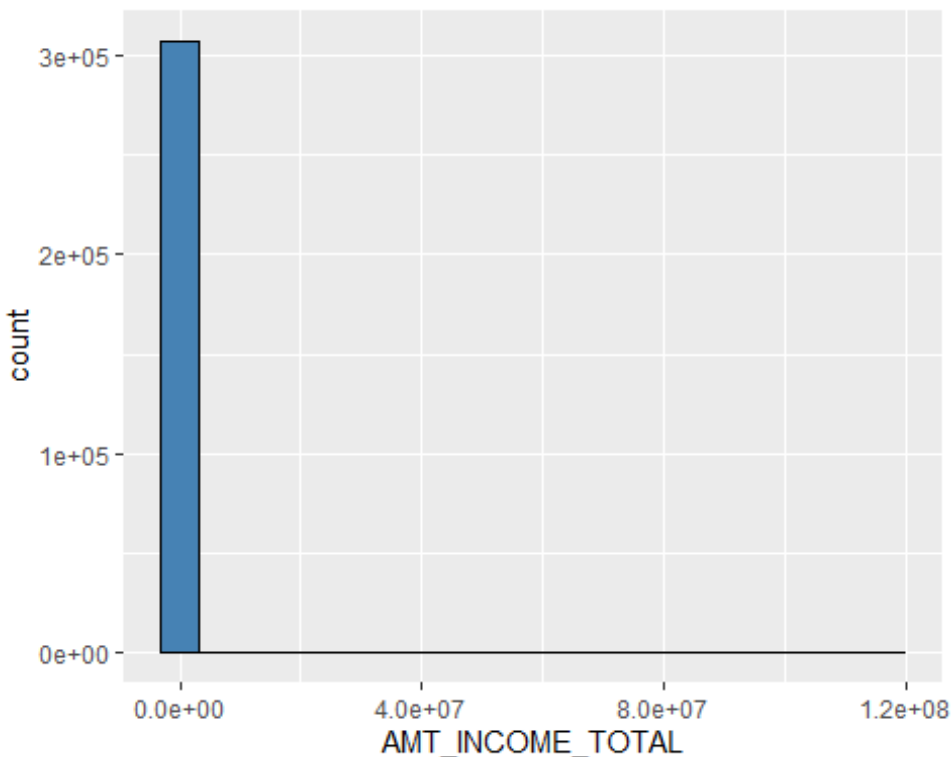
2024-09-27

```
library(ggplot2)
data = read.csv("Loandefault.csv", header = TRUE)
attach(data)
```

## Question 1:

*AMT\_INCOME\_TOTAL*

```
ggplot(data, aes(x=AMT_INCOME_TOTAL)) +
  geom_histogram(bins=20, fill="steelblue", color="black")
```



```
summary(AMT_INCOME_TOTAL)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	25650	112500	147150	168798	202500	117000000

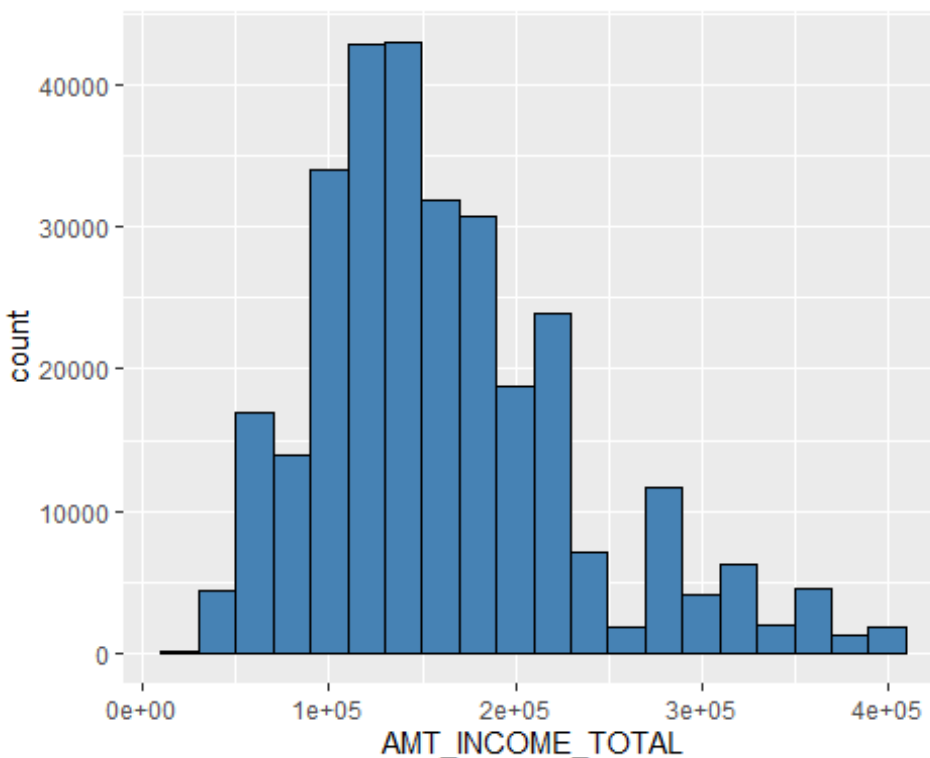
## Response (part 1):

This distribution is extremely skewed right. The range between the maximum and minimum values is also huge. The right skew indicates there are not many outlying high

values, as otherwise their bars would show up on the histogram. Since there are so few of these high outliers, I think we can safely eliminate them. To do this, while still preserving as much data as possible, I removed the top 2.5% of the values. In statistical studies, it is common to use an alpha level of 0.05, where the highest and lowest 2.5% of the distribution are considered significant. I am simply following this precedent, except I have elected to keep the lower 2.5% because the right skew indicates the data on this side of the graph that may not be outliers (at least not as extreme).

```
upper_bound = quantile(AMT_INCOME_TOTAL, 0.975)
data_capped = subset(data, AMT_INCOME_TOTAL <= upper_bound)

ggplot(data_capped, aes(x=AMT_INCOME_TOTAL)) +
  geom_histogram(bins=20, fill="steelblue", color="black")
```



```
summary(data_capped$AMT_INCOME_TOTAL)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  25650  112500  144000  159861  202500  405000
```

*Response (part 2):*

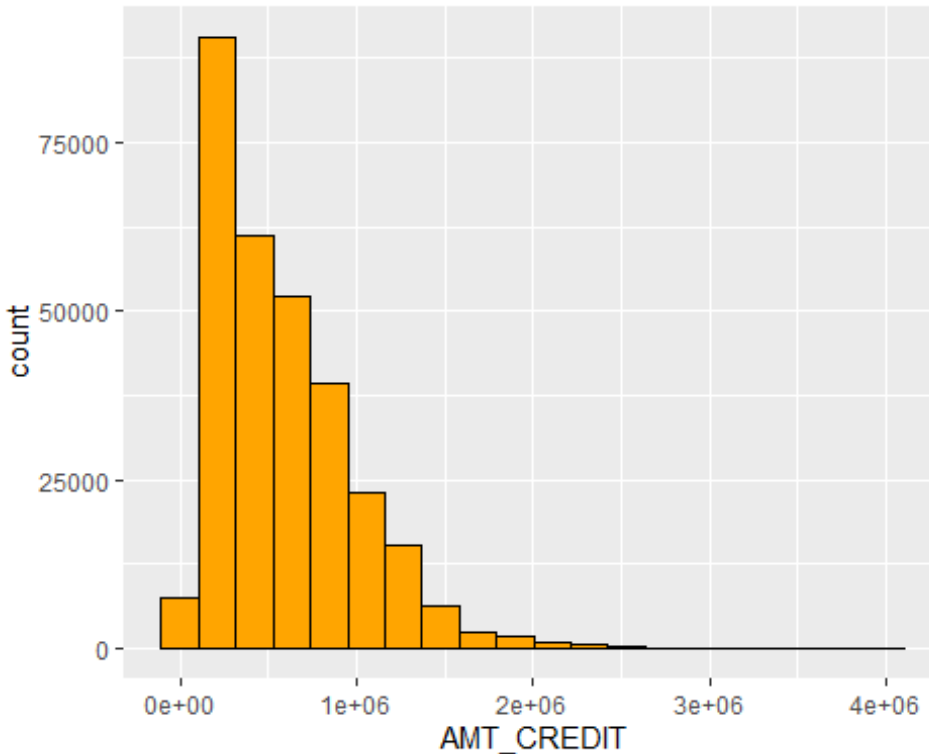
The new graph is still skewed right, but now we can actually see where the majority of our values lie. The distribution has one mode and the values cluster around 150,000. We can see the affect the outliers were having on the mean, as it is much lower now than when the outliers were included.

We're going to keep using the capped data set for the rest of the assignment.

```
detach(data)
attach(data_capped)
```

*AMT\_CREDIT*

```
ggplot(data_capped, aes(x=AMT_CREDIT)) +
  geom_histogram(bins=20, fill="orange", color="black")
```



```
summary(AMT_CREDIT)
```

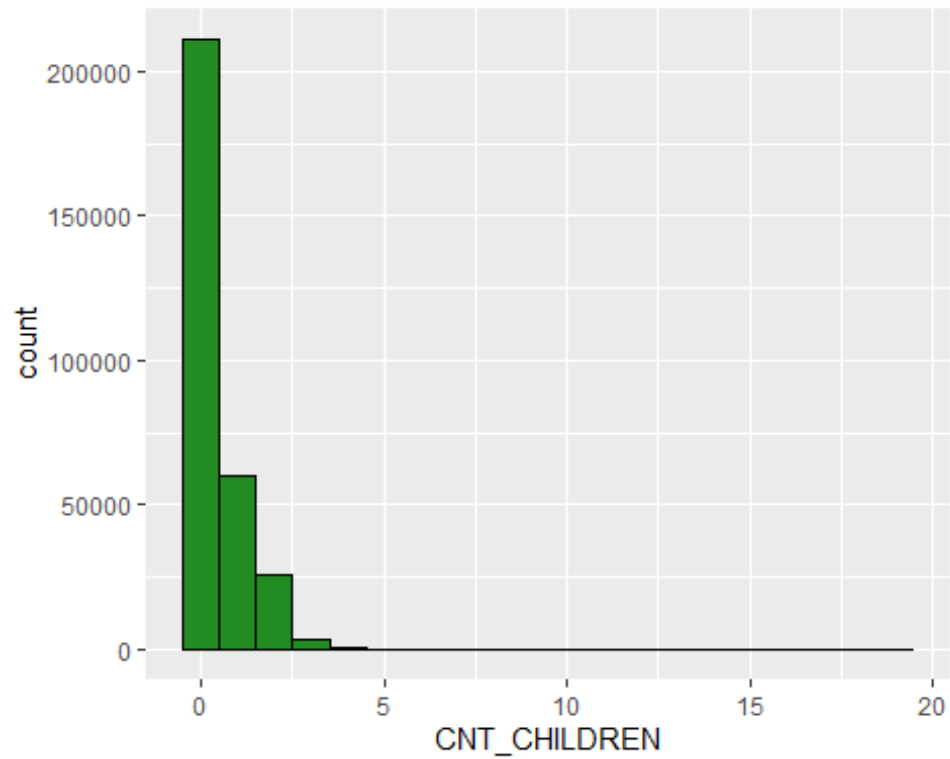
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  45000   270000   508496   589225   808650  4050000
```

*Response:*

Even on our new capped data set, this distribution is skewed right as well. All of its data are clustered on the left side of the chart, so the mode is one of our lowest values. I considered cutting off the top 2.5% of this variable too, but ultimately I decided not to. While the distribution is heavily skewed, I think the data are spread out enough that we can work with it.

*CNT\_CHILDREN*

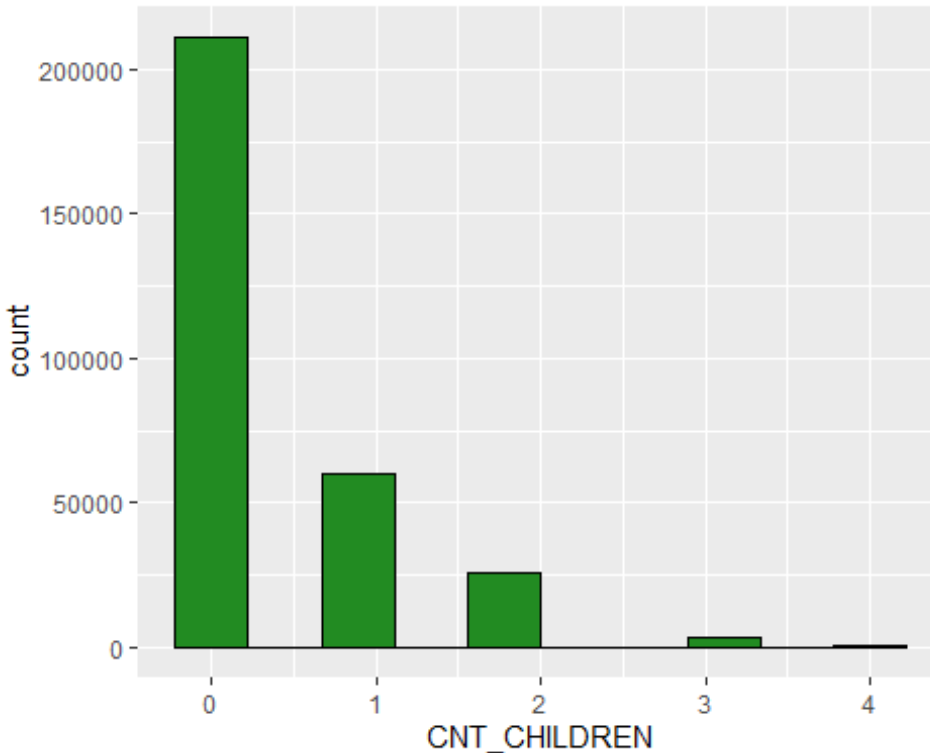
```
ggplot(data_capped, aes(x=CNT_CHILDREN)) +
  geom_histogram(bins=20, fill="forestgreen", color="black")
```



```
summary(CNT_CHILDREN)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4154  1.0000 19.0000
```

```
upper_bound2 = quantile(CNT_CHILDREN, 0.999)
data_capped2 = subset(data_capped, CNT_CHILDREN <= upper_bound2)
ggplot(data_capped2, aes(x=CNT_CHILDREN)) +
  geom_histogram(bins=10, fill="forestgreen", color="black")
```



```
summary(data_capped2$CNT_CHILDREN)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4131  1.0000  4.0000
```

*Response:*

Here we have a similar case to our AMT\_INCOME\_TOTAL, although not quite as dramatic. I still wanted to remove outliers so our histogram was more than a couple bars and a very long tail. Since the range of unique values for this variable is comparatively tiny (only 0-19), I had to be careful not to set the cap too low. I chose to remove only the top 0.1%.

Even though we kept most of the data, we dramatically reduced the range of unique values for this variable. Like before, the outliers we removed were so few in number that dropping them shouldn't hurt our analysis. This data resembles the shape of our AMT\_CREDIT distribution. It is heavily skewed right and the mode is, in this case, our smallest value.

**We're going to keep using this NEW capped data set for the rest of the assignment.**

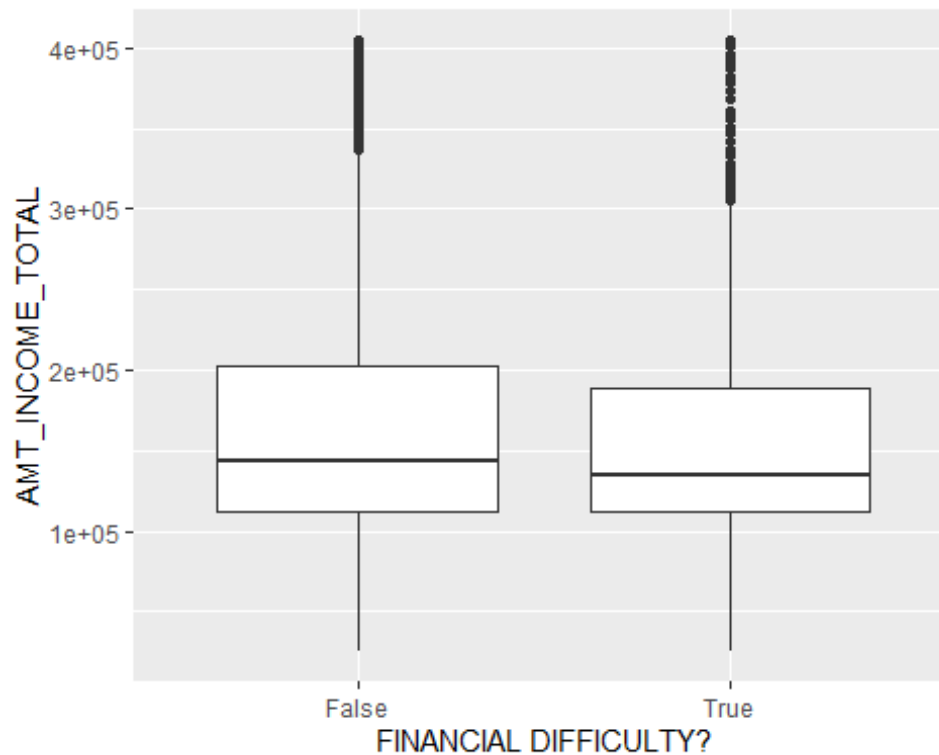
```
detach(data_capped)
attach(data_capped2)
```

## Question 2

*AMT\_INCOME\_TOTAL*

```
ggplot(data_capped2, aes(y=AMT_INCOME_TOTAL, x=factor(TARGET))) +
  geom_boxplot() +
```

```
labs(x="FINANCIAL DIFFICULTY?") +  
scale_x_discrete(label=c("False", "True"))
```



```
summary(AMT_INCOME_TOTAL[TARGET==1]) # True column
```

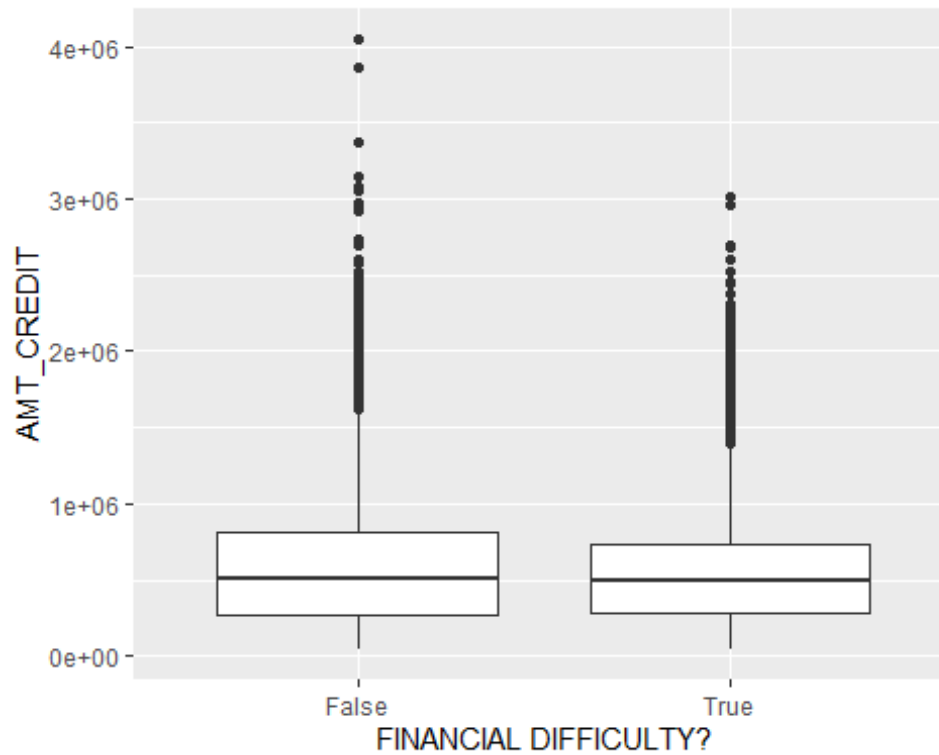
```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  25650 112500  135000  155108 189000  405000
```

```
summary(AMT_INCOME_TOTAL[TARGET==0]) # False column
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  25650 112500  144000  160282 202500  405000
```

AMT\_CREDIT

```
ggplot(data_capped2, aes(y=AMT_CREDIT, x=factor(TARGET))) +  
  geom_boxplot() +  
  labs(x="FINANCIAL DIFFICULTY?") +  
  scale_x_discrete(label=c("False", "True"))
```



```
summary(AMT_CREDIT[TARGET==1]) # True
```

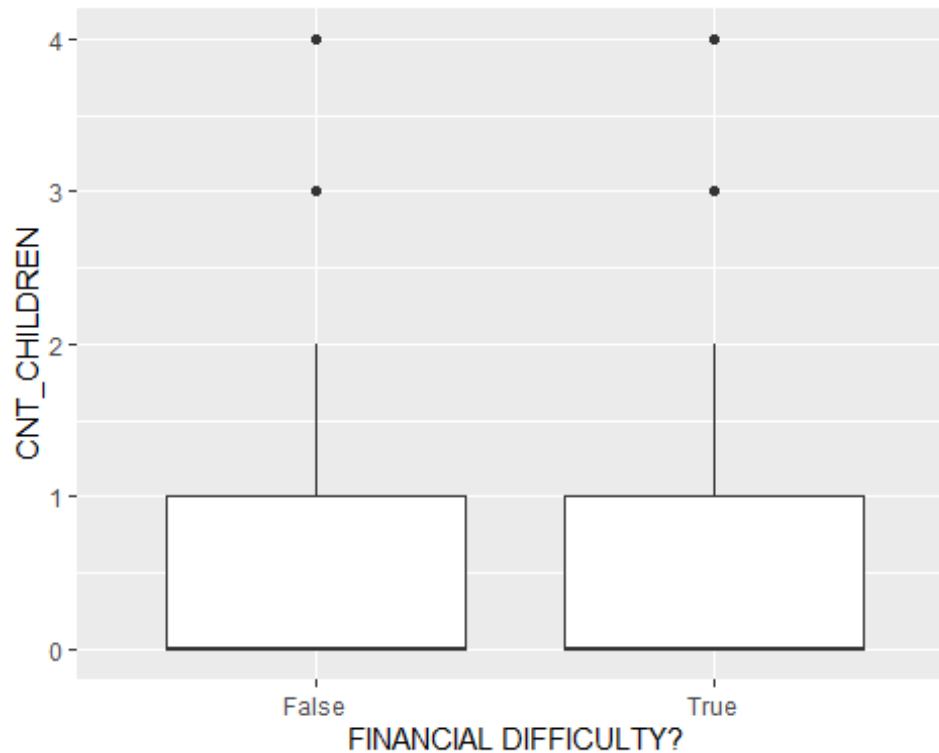
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  45000  284400  495000  552503  728460 3020760
```

```
summary(AMT_CREDIT[TARGET==0]) # False
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  45000  270000  508496  592490  808650 4050000
```

CNT\_CHILDREN

```
ggplot(data_capped2, aes(y=CNT_CHILDREN, x=factor(TARGET))) +
  geom_boxplot() +
  labs(x="FINANCIAL DIFFICULTY?") +
  scale_x_discrete(label=c("False", "True"))
```



```
summary(CNT_CHILDREN[TARGET==1]) # True

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000  0.0000  0.0000  0.4593  1.0000  4.0000

summary(CNT_CHILDREN[TARGET==0]) # False

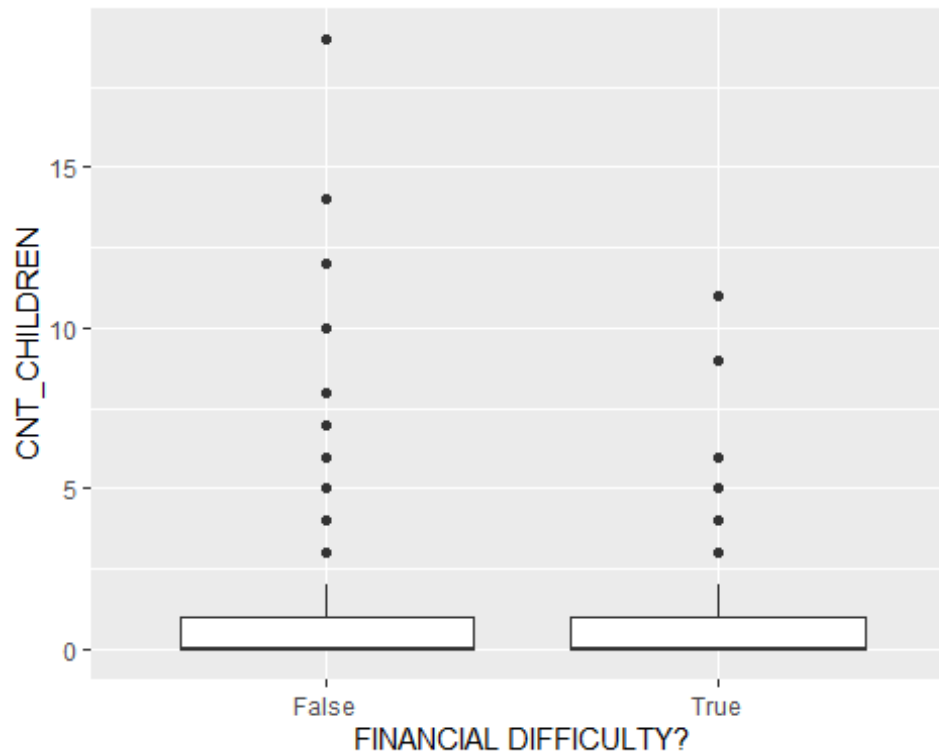
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.000  0.000  0.000  0.409  1.000  4.000
```

*Response (part 1):*

For all three of our variables, the boxplots appear very similar. Their quantile values and even their outliers are relatively close. Although it's true that the data for customers with a history of financial difficulty (True column) have a mean, median, and quantile values that are slightly lower than the other customers, the difference is slight and I'm not prepared to make any conclusions based on it. We'd need to do more analysis determine that. For now, it looks like total income, credit and number of children do not have a significant impact on whether or not customers have a history of financial difficulty.

```
ggplot(data_capped, aes(y=CNT_CHILDREN, x=factor(TARGET))) +
  geom_boxplot() +
  labs(x="FINANCIAL DIFFICULTY?") +
  scale_x_discrete(label=c("False", "True"))
```





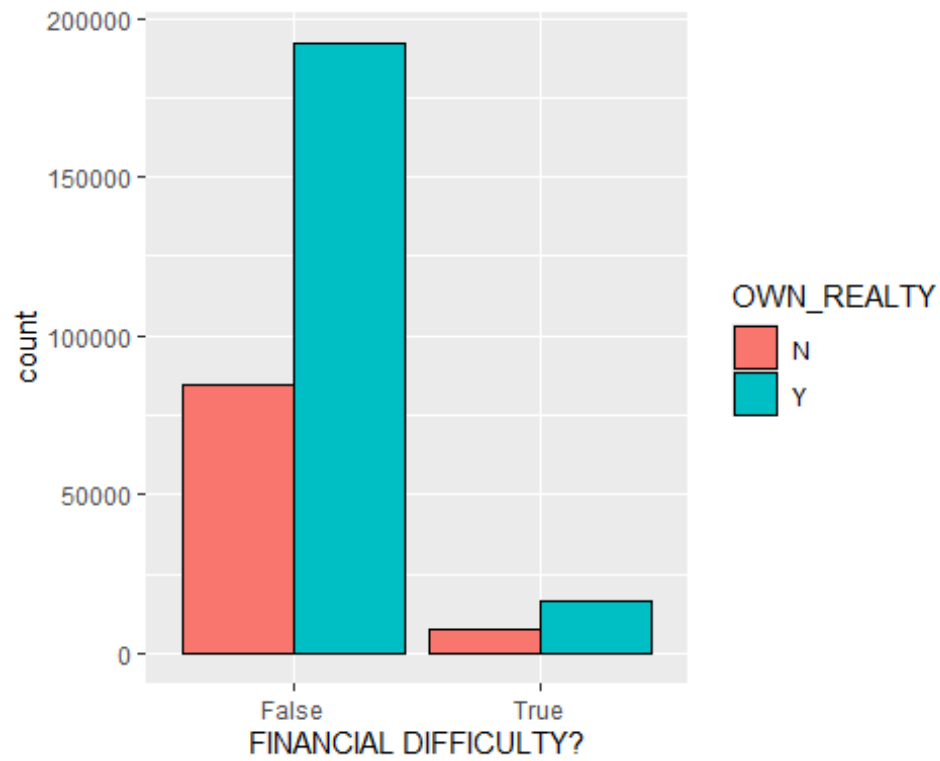
#### Response(part 2):

Since the boxplots for CNT\_CHILDREN were so similar, I also made a box plot for that variable using its full range (0-19). It possible that the impact of having children does not appear until the count is much greater than average. I was surprised to find there wasn't a large difference with this data either. If anything, there were more data points for families with greater numbers of children in the "False" column than there were in the True" one! This tells me that number of children mostly likely doesn't have an impact of whether or not the client had past financial difficulties.

#### Question 3

##### OWN\_REALTY

```
ggplot(data_capped2, aes(x=factor(TARGET), fill=OWN_REALTY)) +
  geom_bar(color="black", position="dodge") +
  labs(x="FINANCIAL DIFFICULTY?") +
  scale_x_discrete(label=c("False", "True"))
```

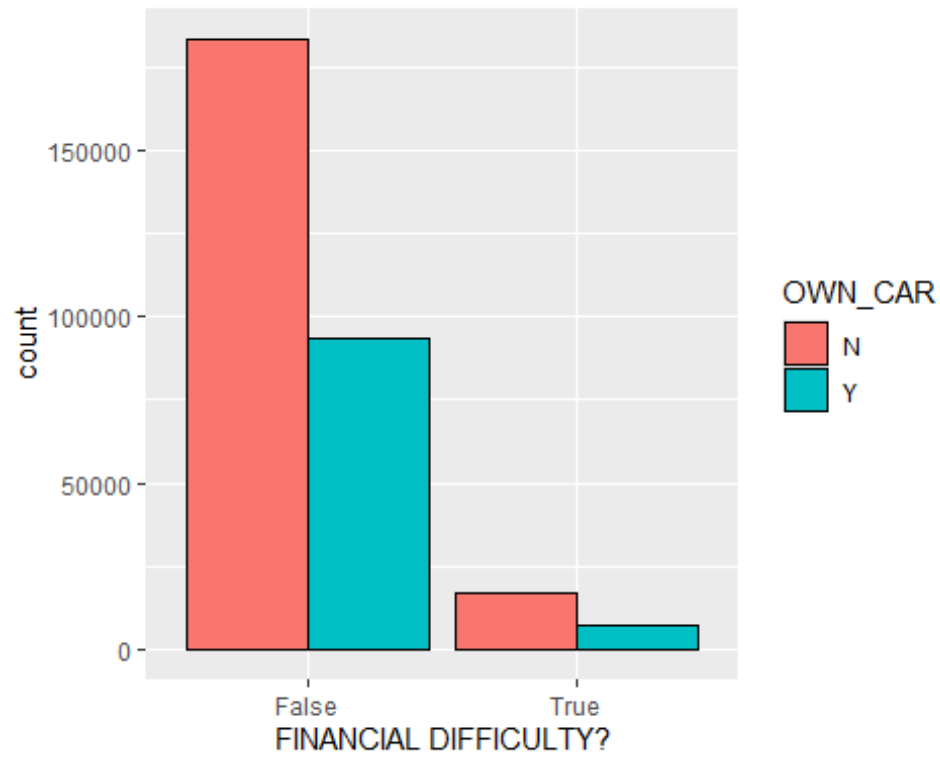


```
ggplot(data_capped2, aes(x=factor(TARGET), fill=OWN_REALTY)) +  
  geom_bar(color="black", position="fill") +  
  labs(x="FINANCIAL DIFFICULTY?") +  
  scale_x_discrete(label=c("False", "True"))
```

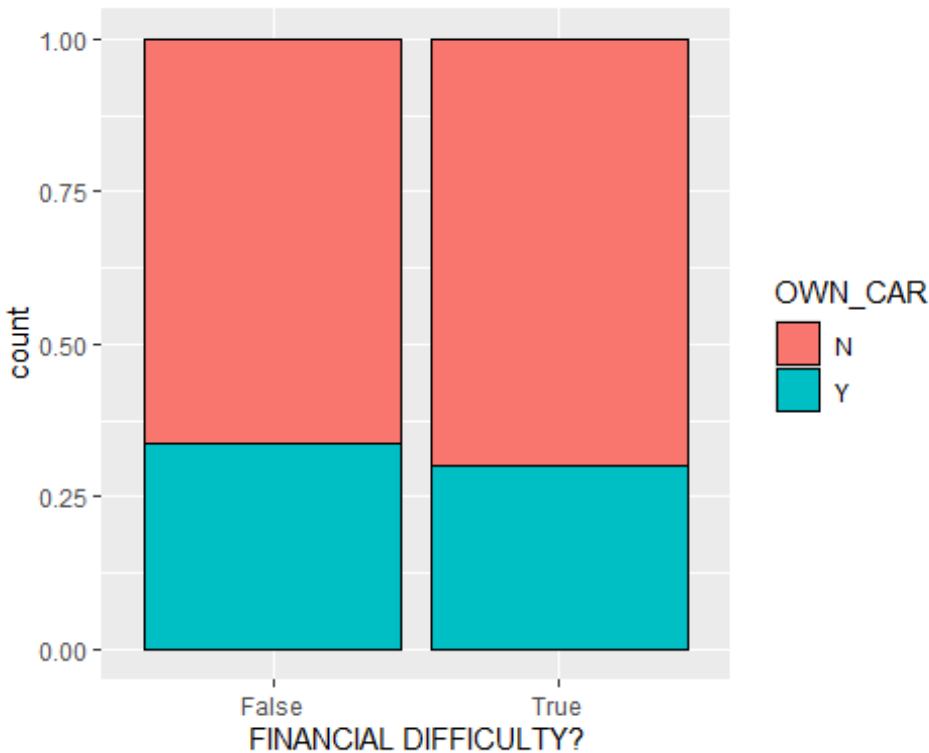


OWN\_CAR

```
ggplot(data_capped2, aes(x=factor(TARGET), fill=OWN_CAR)) +  
  geom_bar(color="black", position="dodge") +  
  labs(x="FINANCIAL DIFFICULTY?") +  
  scale_x_discrete(label=c("False", "True"))
```



```
ggplot(data_capped2, aes(x=factor(TARGET), fill=OWN_CAR)) +  
  geom_bar(color="black", position="fill") +  
  labs(x="FINANCIAL DIFFICULTY?") +  
  scale_x_discrete(label=c("False", "True"))
```

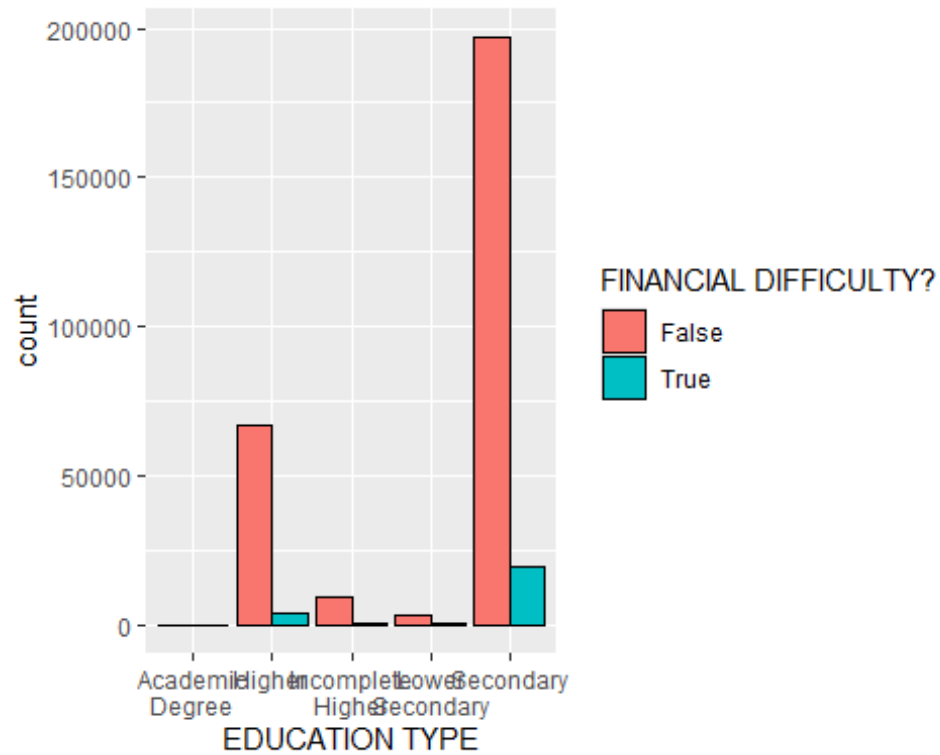


#### Response (part 1):

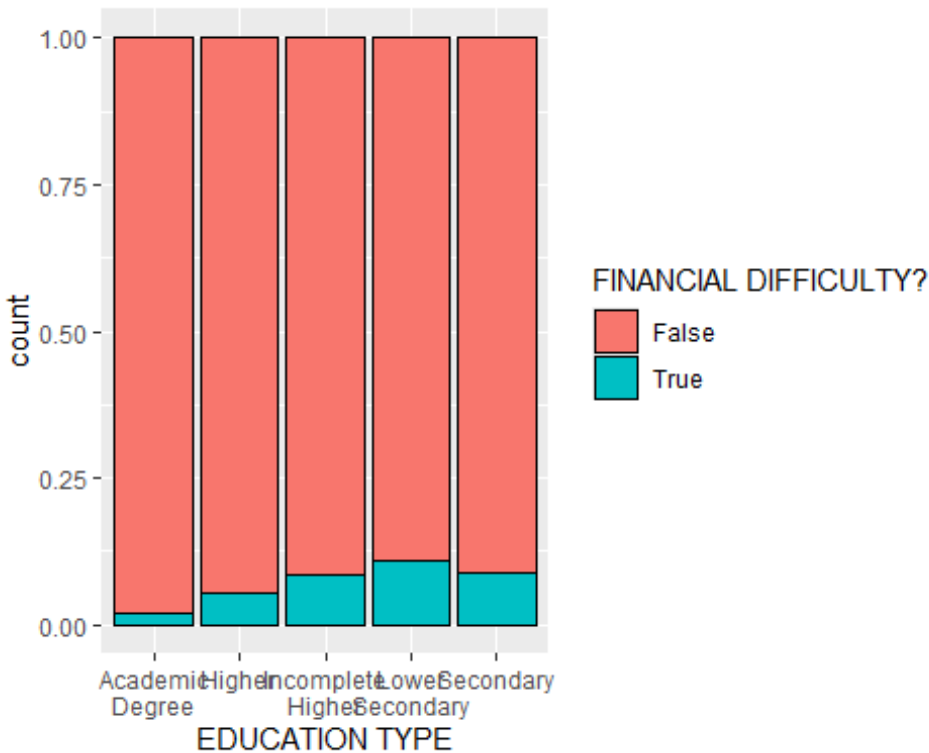
For the two variables, OWN\_REALTY and OWN\_CAR, I do not see any significant differences. Since there were a lot more data in the “False” column, I needed to normalize it by putting them into a stacked bar graph. Once normalized, it was quite apparent that owning realty or cars does not have any significant influence over whether or not someone has had past financial struggles.

#### NAME\_EDUCATION\_TYPE

```
ggplot(data_capped2, aes(x=NAME_EDUCATION_TYPE, fill=factor(TARGET))) +
  geom_bar(color="black", position="dodge") +
  labs(fill="FINANCIAL DIFFICULTY?", x="EDUCATION TYPE") +
  scale_fill_discrete(label=c("False", "True")) +
  scale_x_discrete(label=c("Academic\nDegree", "Higher",
    "Incomplete\nHigher", "Lower\nSecondary", "Secondary"))
```



```
ggplot(data_capped2, aes(fill=factor(TARGET), x=NAME_EDUCATION_TYPE)) +
  geom_bar(color="black", position="fill") +
  labs(fill="FINANCIAL DIFFICULTY?", x="EDUCATION TYPE") +
  scale_fill_discrete(label=c("False", "True")) +
  scale_x_discrete(label=c("Academic\nDegree", "Higher",
    "Incomplete\nHigher", "Lower\nSecondary", "Secondary"))
```



#### Response (part 2)

Here, we finally see something happening! As we can see in the stacked bar chart, education type does influence whether or not someone has a history of financial difficulty. From the normalized chart, we can see that as the level of formal education decreases, the occurrences of past financial difficulty rises. While it's not a dramatic jump, this does support this conclusion. (Note, "Lower secondary" is lower than "Secondary", so really it should be the last column.)

#### Question 4

##### My question:

Provide three research questions that relate the variables: total income, credit score, number of children, owning realty, owning a car, and education level. Your questions should not be related to past financial history and you don't have to use every variable.

##### ChatGPT response:

Here are three focused research questions that incorporate some of the specified variables:

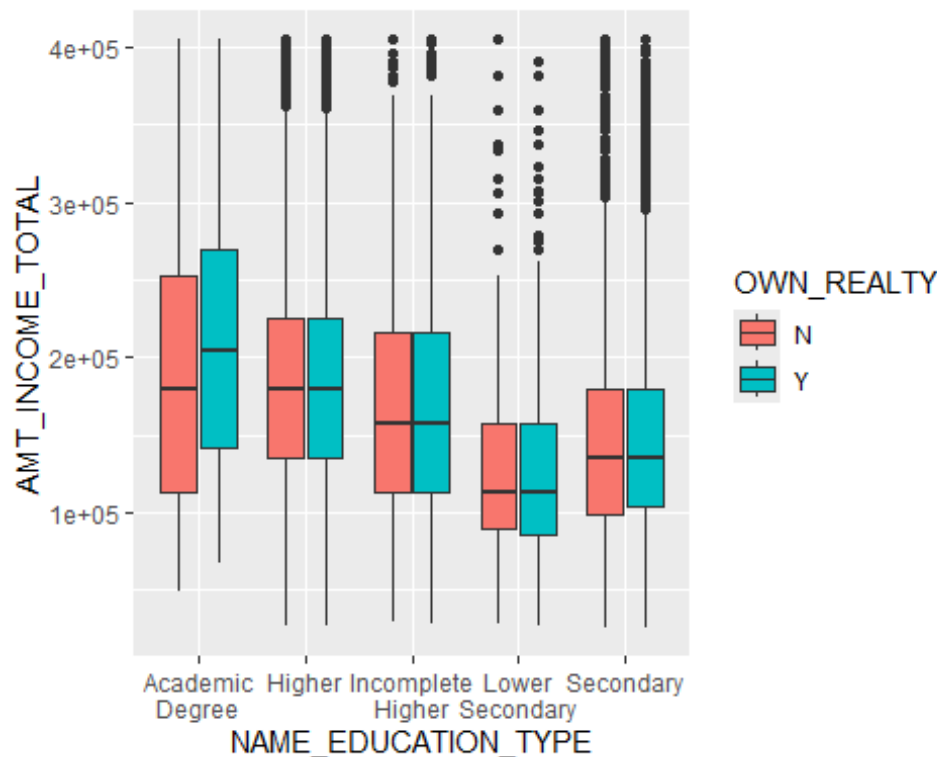
1. How does the level of education correlate with total income among individuals who own realty compared to those who do not?
2. What is the relationship between credit score and the likelihood of owning a car among families with varying numbers of children?

3. In households with children, how does total income influence the decision to invest in realty, and what role does education level play in this decision-making process?

### Question 5

1. How does the level of education correlate with total income among individuals who own realty compared to those who do not?

```
ggplot(data_capped2, aes(x=NAME_EDUCATION_TYPE, y=AMT_INCOME_TOTAL,  
fill=OWN_REALTY)) +  
  geom_boxplot() +  
  scale_x_discrete(label=c("Academic\nDegree", "Higher",  
"Incomplete\nHigher", "Lower\nSecondary", "Secondary"))
```



*Response:*

Based off this visualization, level of education and total income are positively correlated. The boxes represent income level, and they rise on the chart as education rises (recall, LowerSecondary should be the rightmost column). There are also no outliers at the top of the chart for the “Academic degree” column, which makes it unique compared to the other categories. This trend is the same among those who own realty and those who do not. The only interaction between academic level and realty ownership is at the “Academic degree” level, where among these academics, those who own realty also tend to have higher total incomes (or vice versa).