Supplementary Material

Zhihan Xian

2021-12-11

# 1 Exploratory Analysis Details

We are only interested in data from US. We first filter out all American data (Country = USA/United States, Currency = USD)

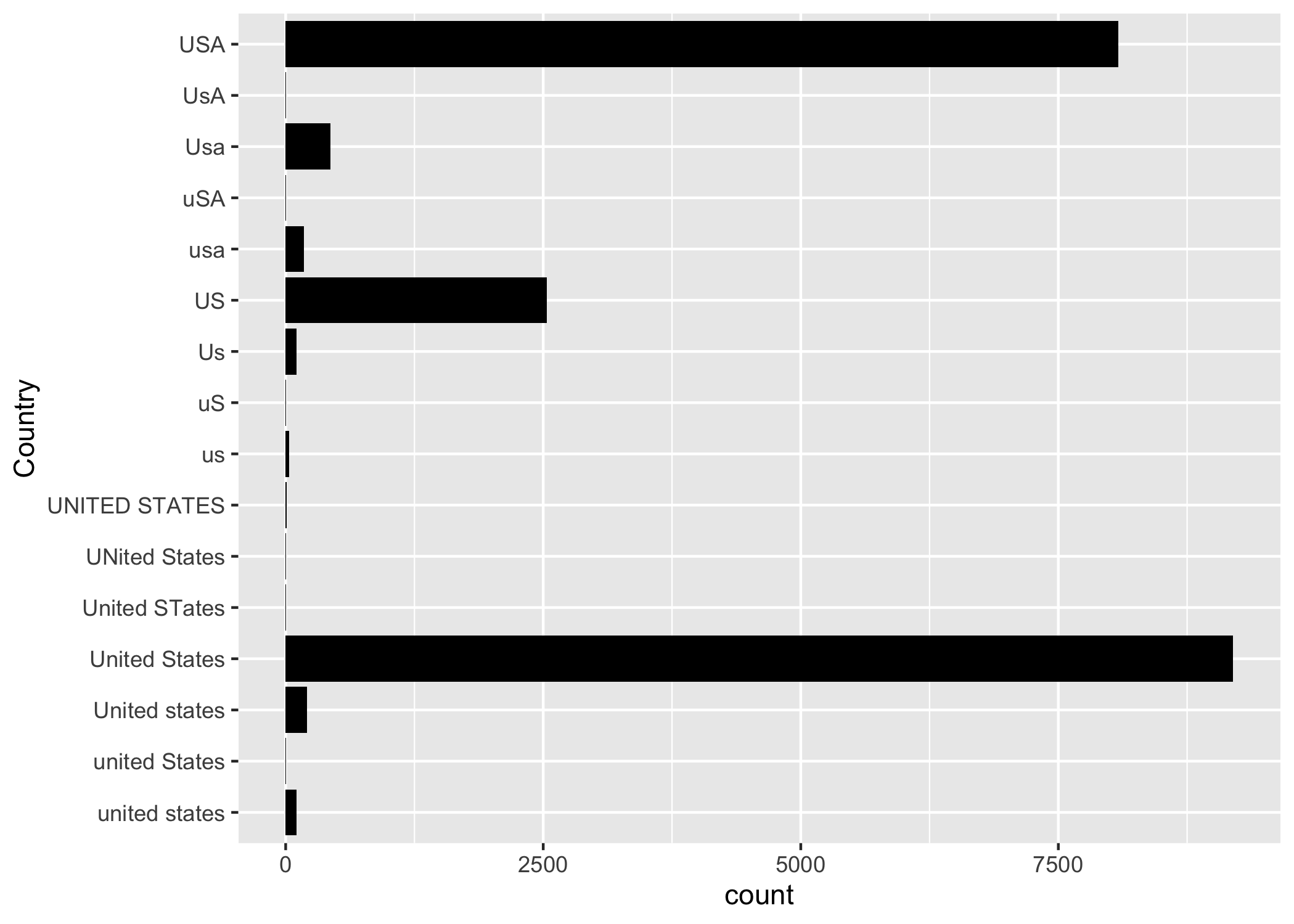
Figure 1.1 shows the remaining country entries are USA only after clean up, regardless of upper or lower cases. 

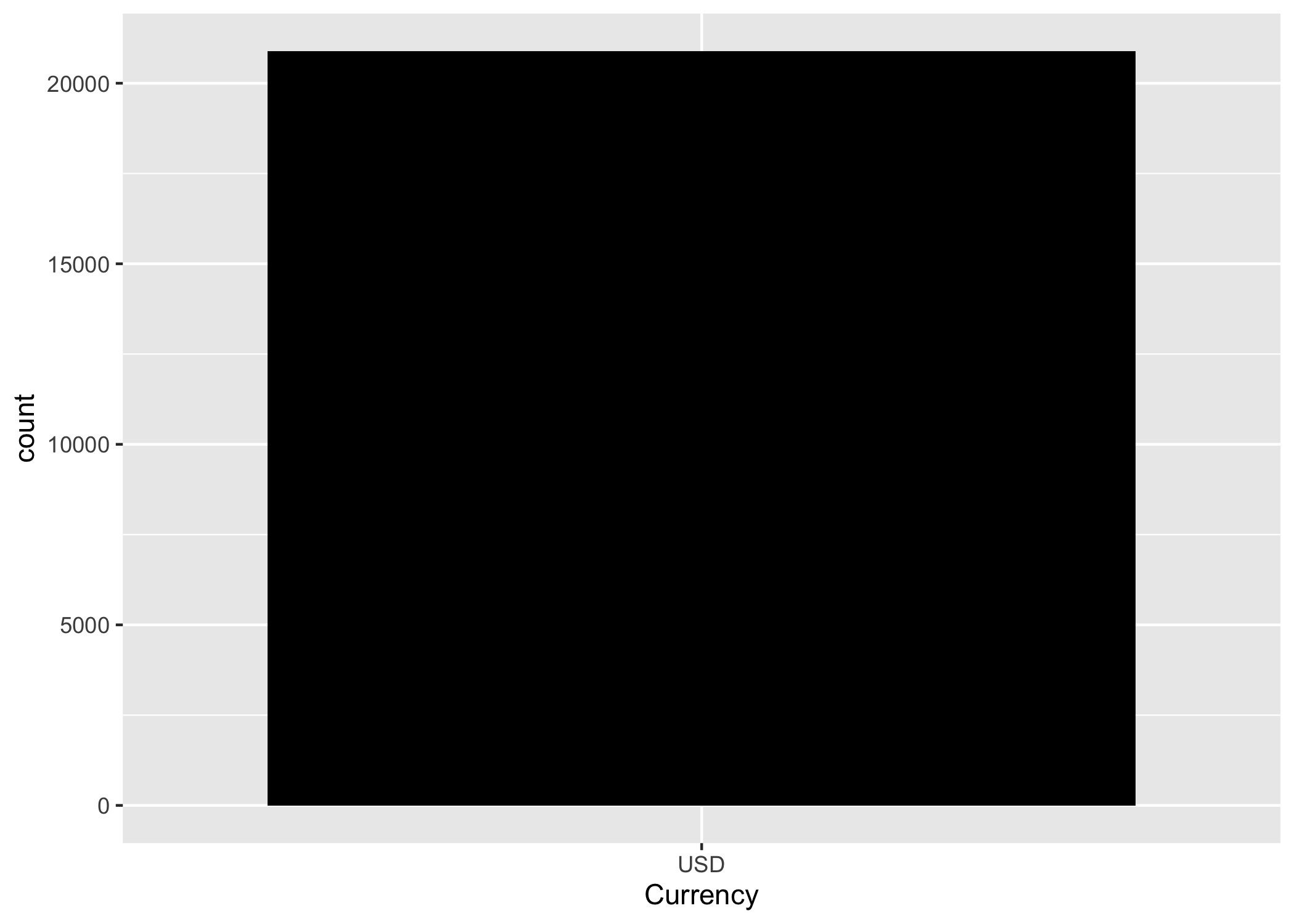
Figure 1.2 shows the remaining currency entries are USD only after clean up. 

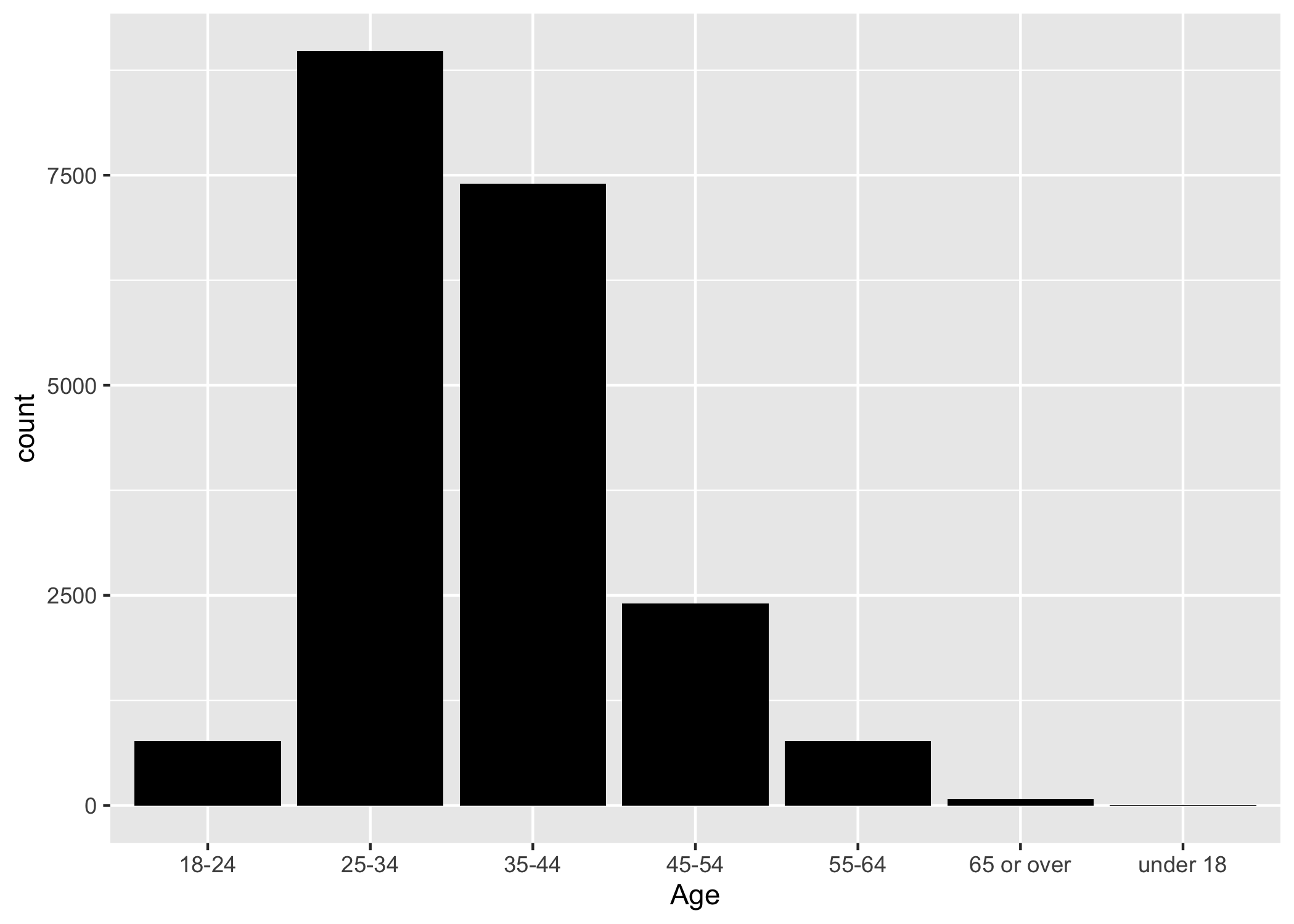
Figure 1.3 shows the original age data before clean up.  As shown above, the age entries are divided into 7 groups, we have to remove 2 groups with almost no datapoints since they do not contain enough data for model training.

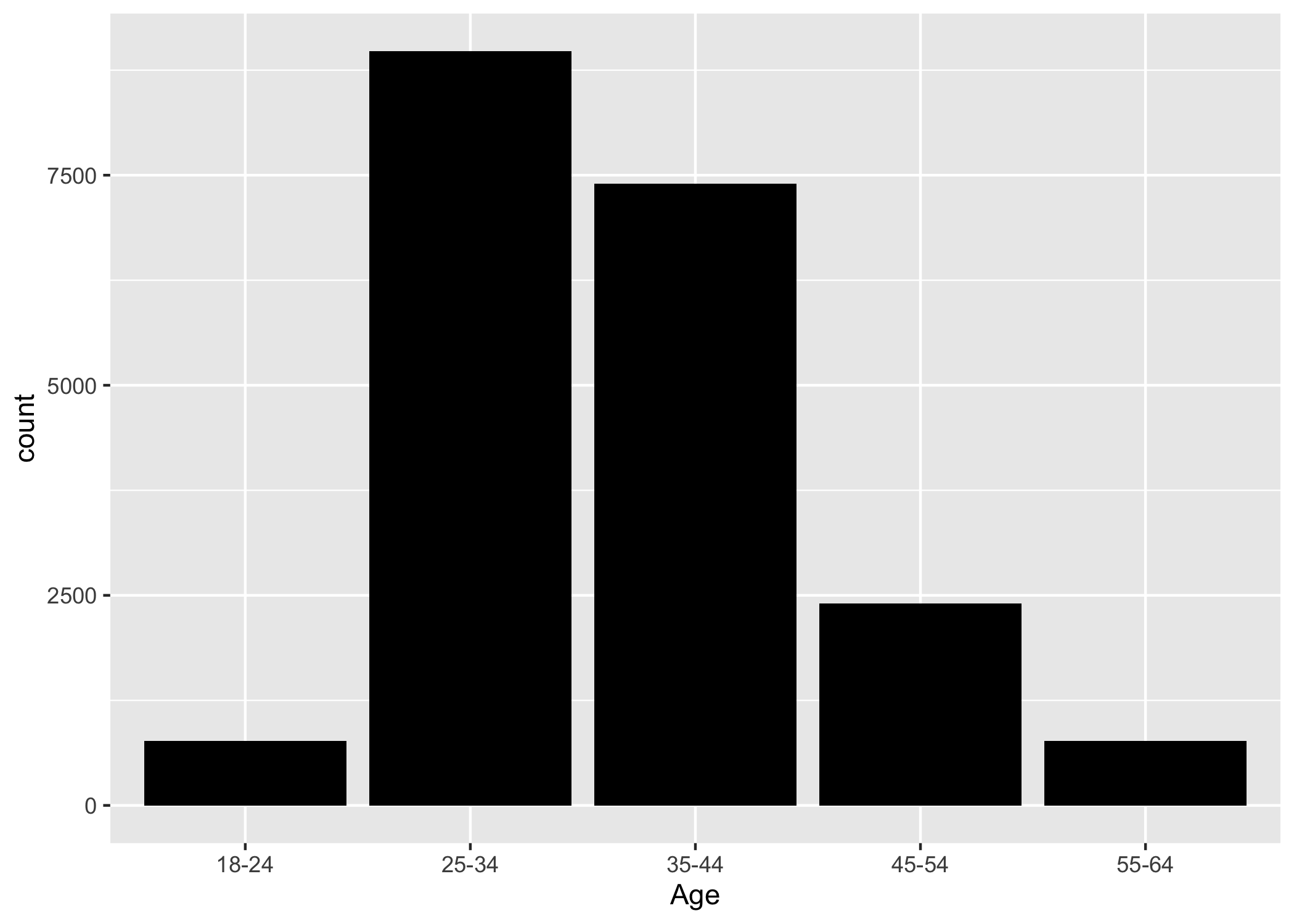
Figure 1.4 shows the age data after clean up. 

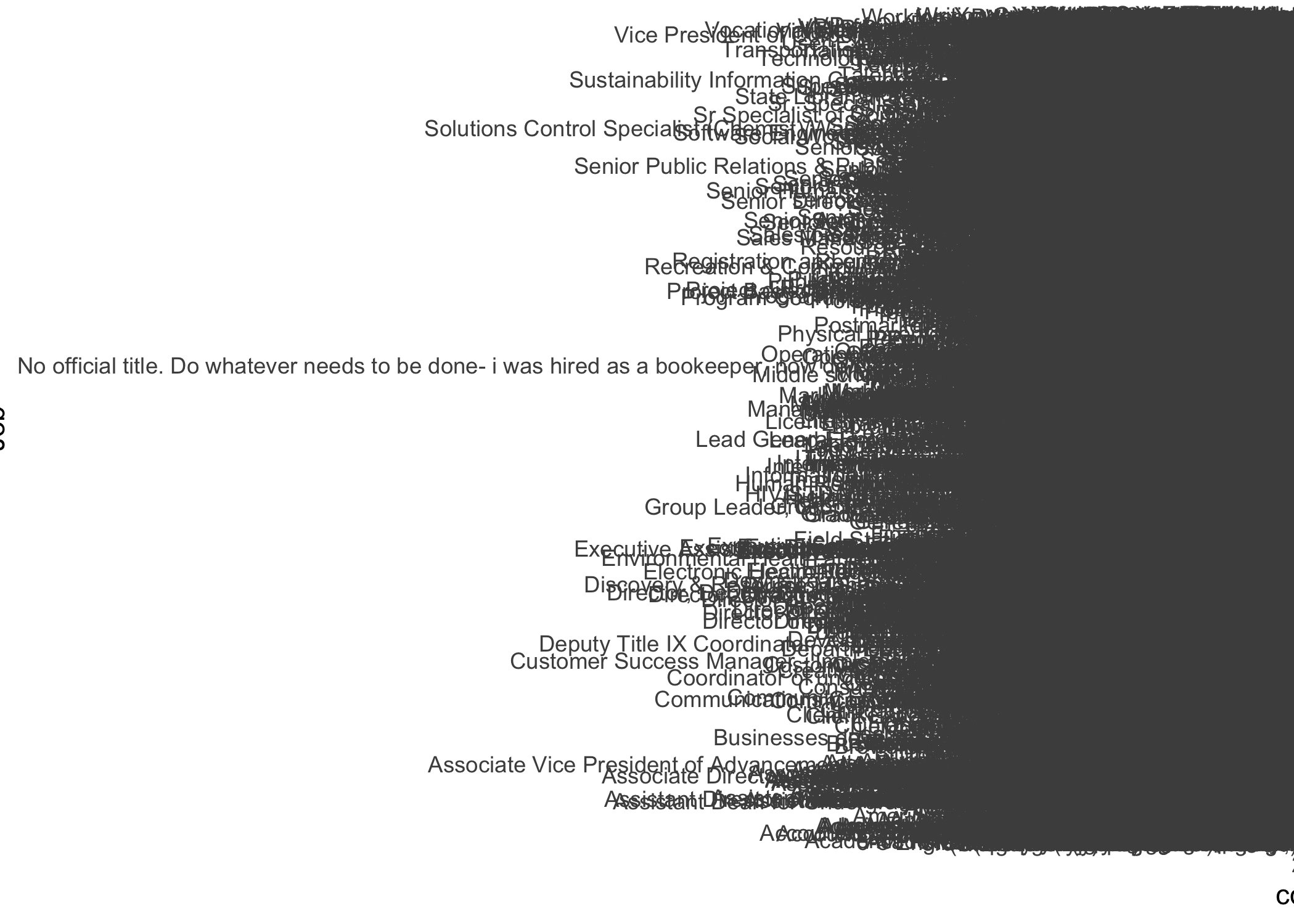
Figure 1.5 shows the original job title data.  As shown above, the job title entries are too messy to clean up. We just drop this column for simplicity.

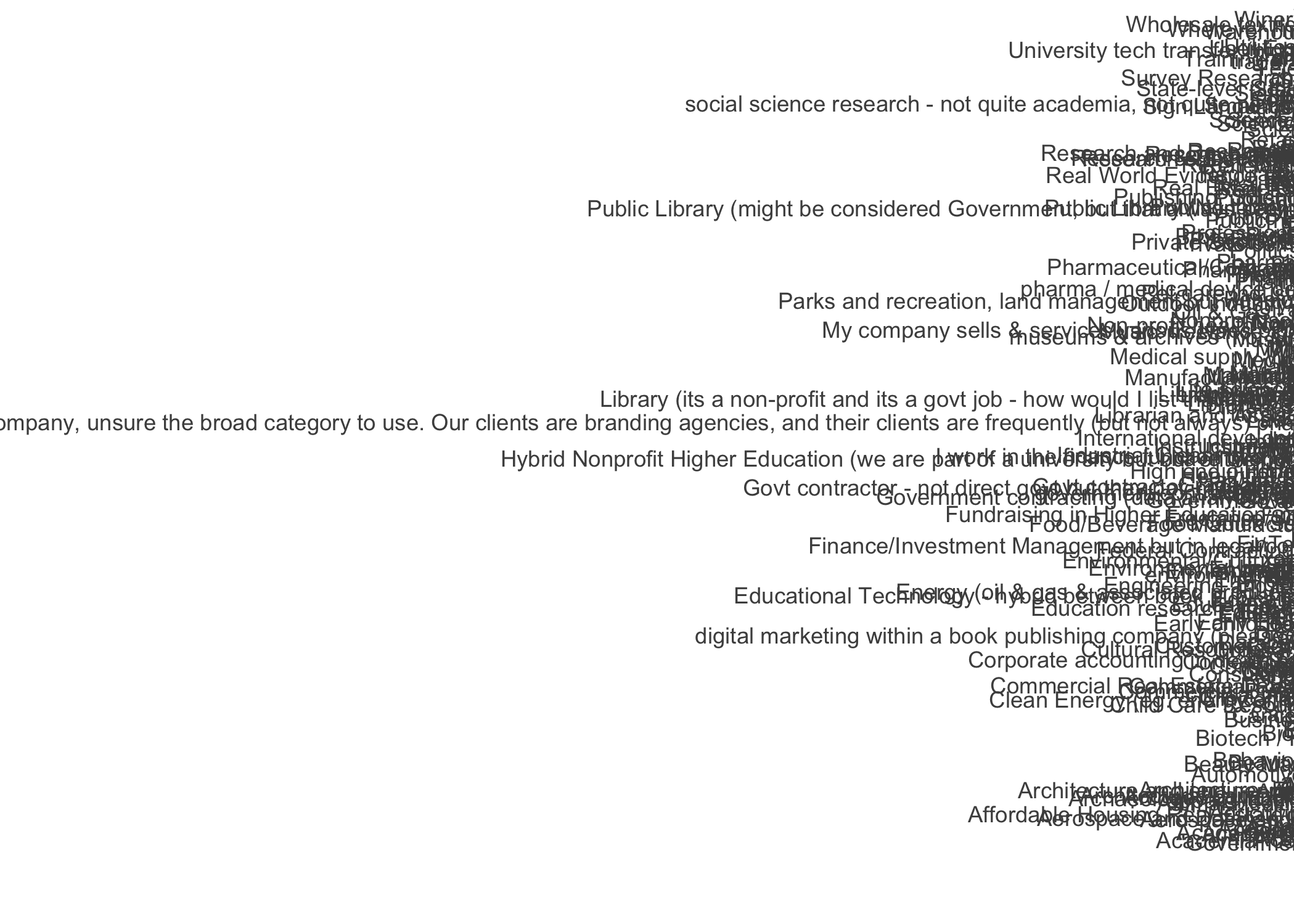
Figure 1.6 shows the original industry data before clean up.  As shown above, the industry entries are messy. However, different from the job title data, most entries are among the top 20 occurrence. Therefore, we extract top 20 industries. Moreover, all top 20 industries have enough data for model training. Therefore, we choose to drop the rest.

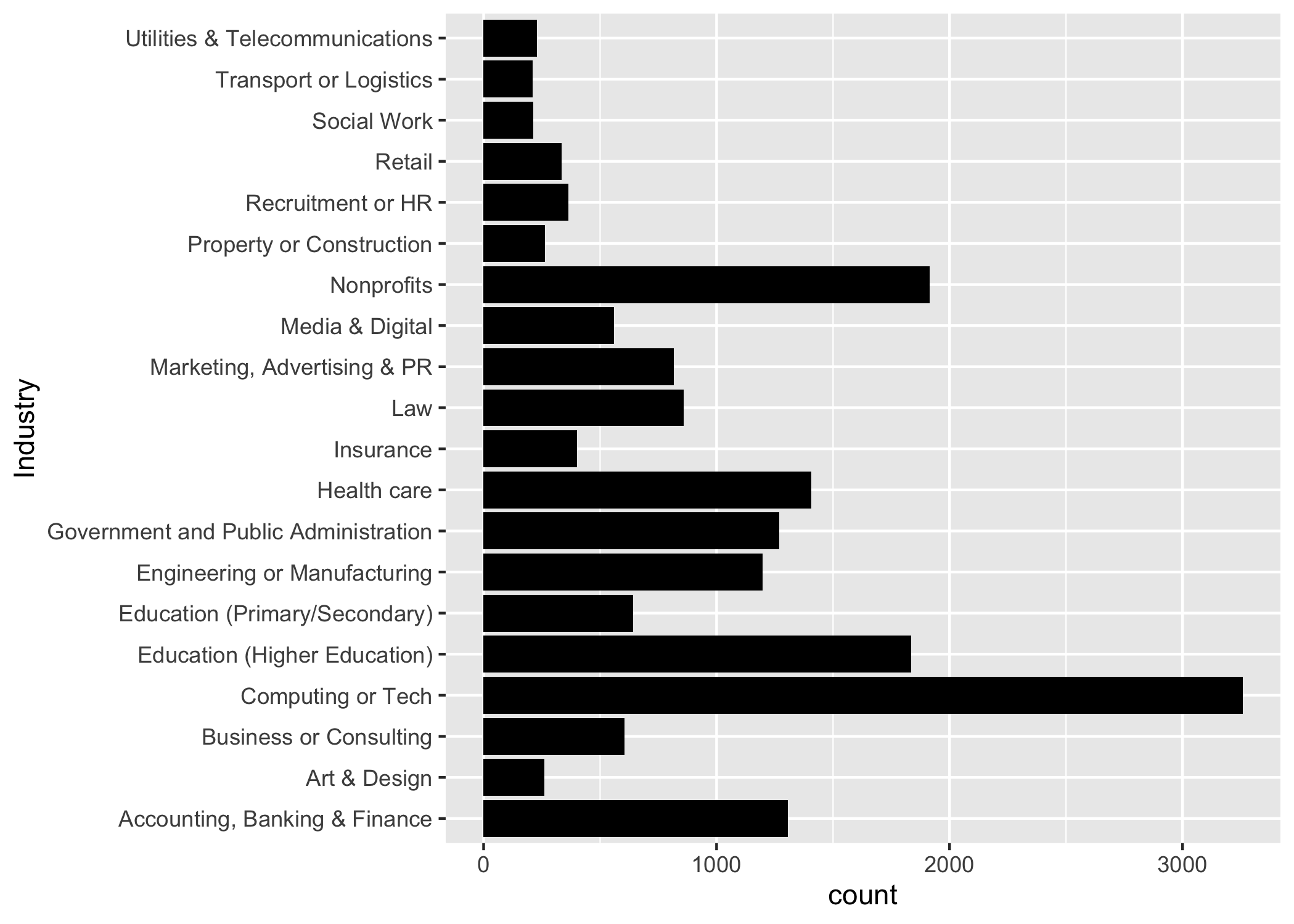
Figure 1.7 shows the industry data after clean up. 

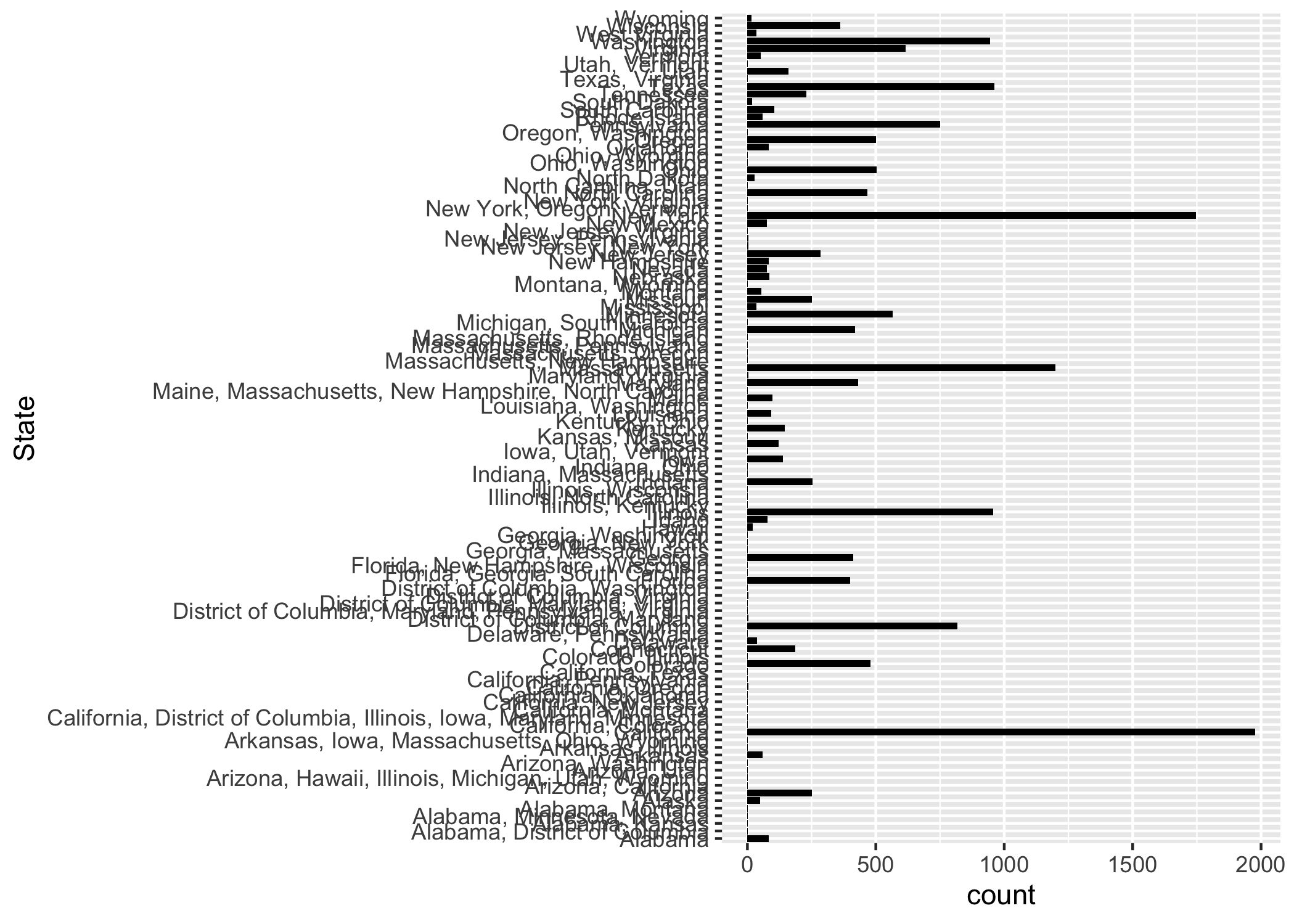
Figure 1.8 shows the original state data before clean up.  As shown above, the state entries are messy too. We have to filter out the 50 states.

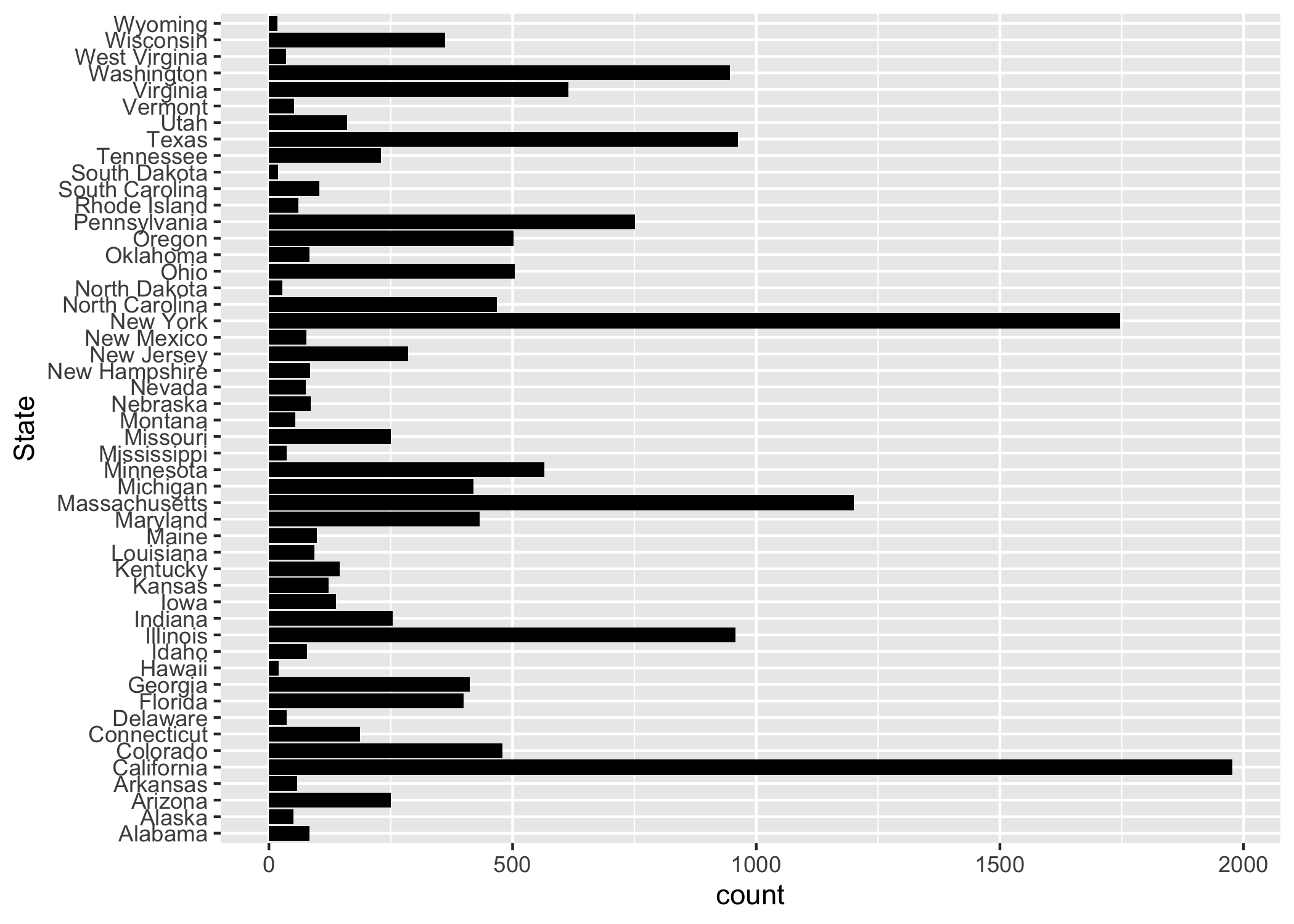
Figure 1.9 shows the state data after clean up. 

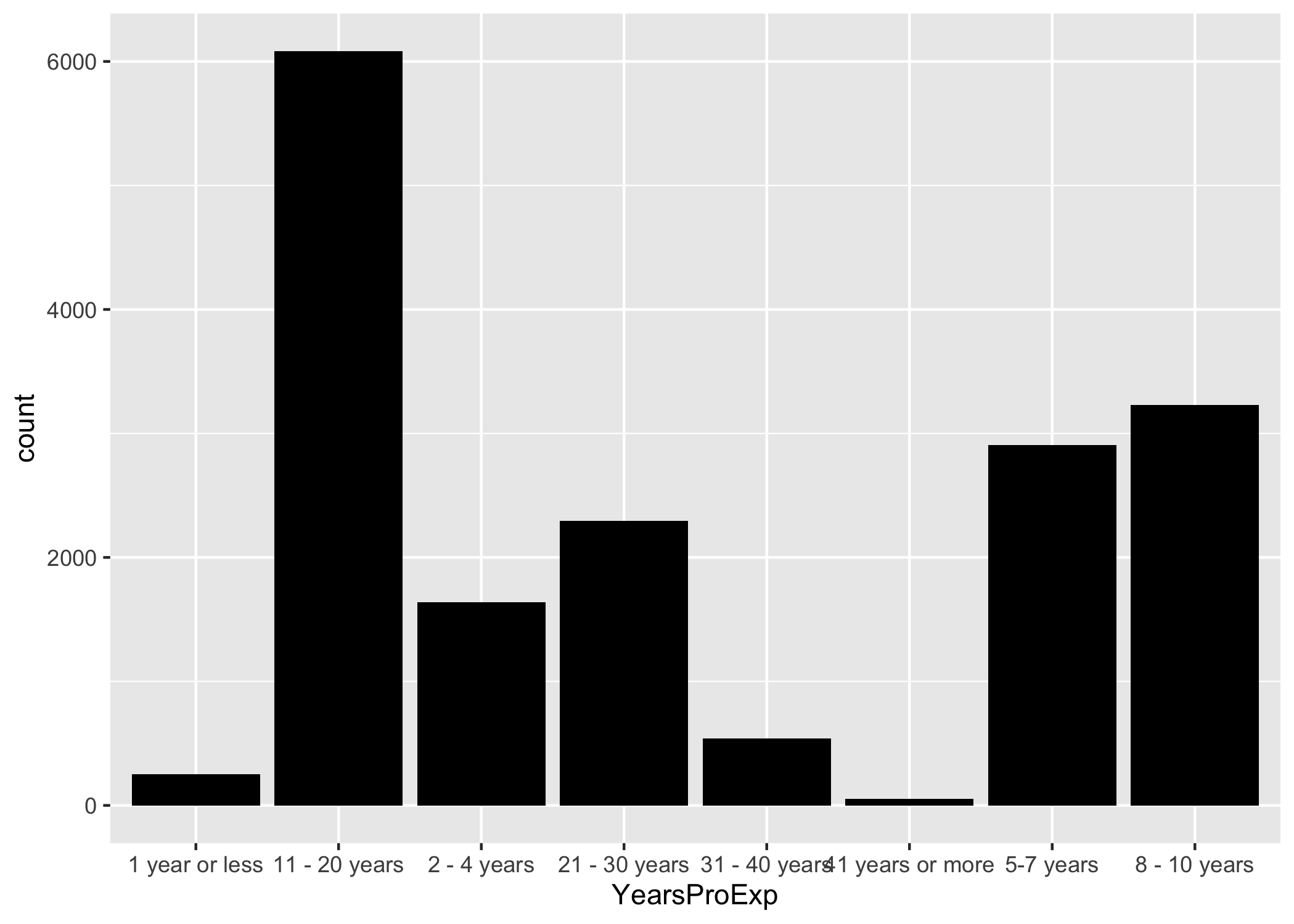
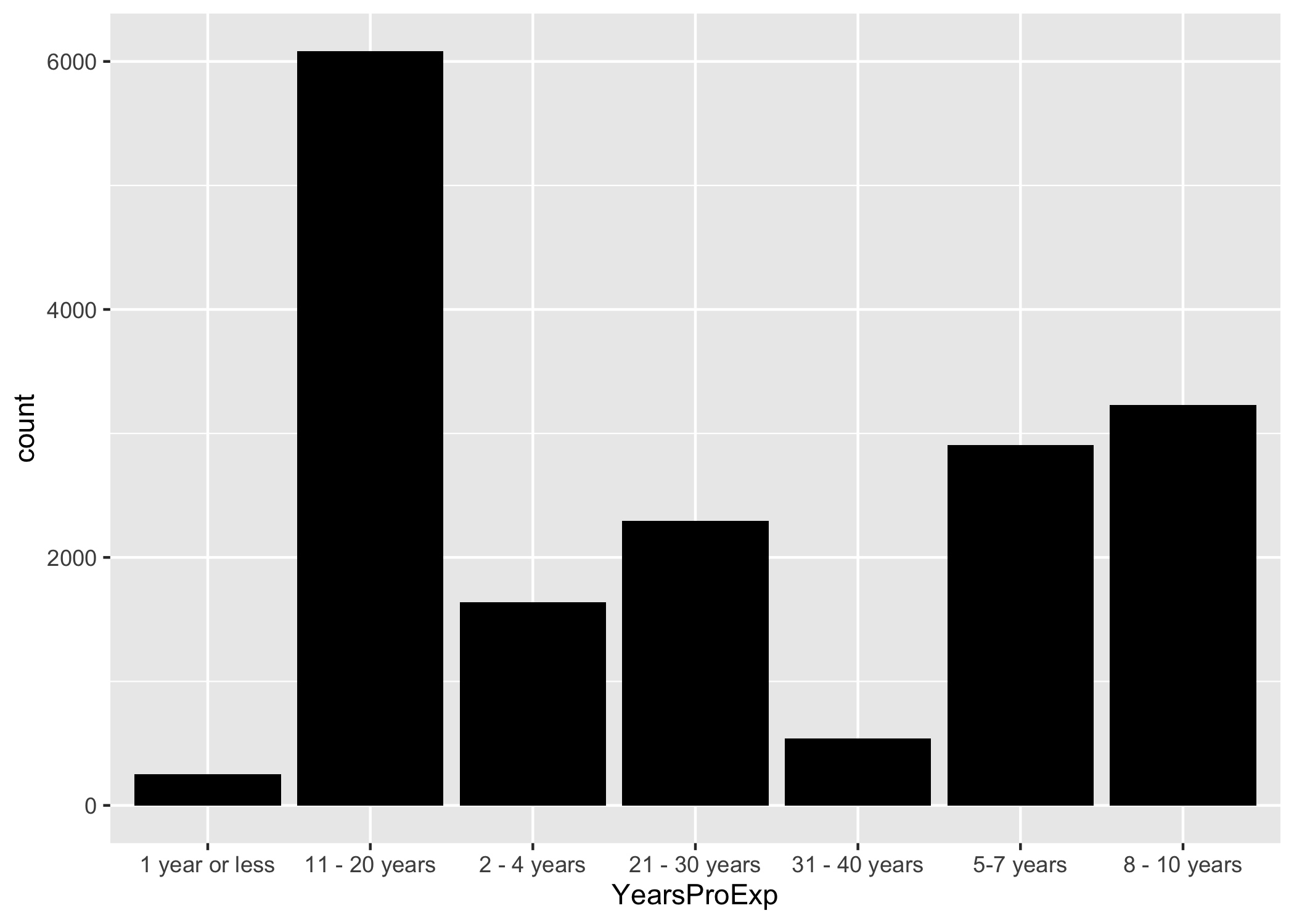
Figure 1.10 shows the original years of professional experience data before clean up.  As shown above, the entries “41 years or above” have few datapoints and do not contain enough data for model training. Thus, we have to drop them.

Figure 1.11 shows the years of professional experience data after clean up. 

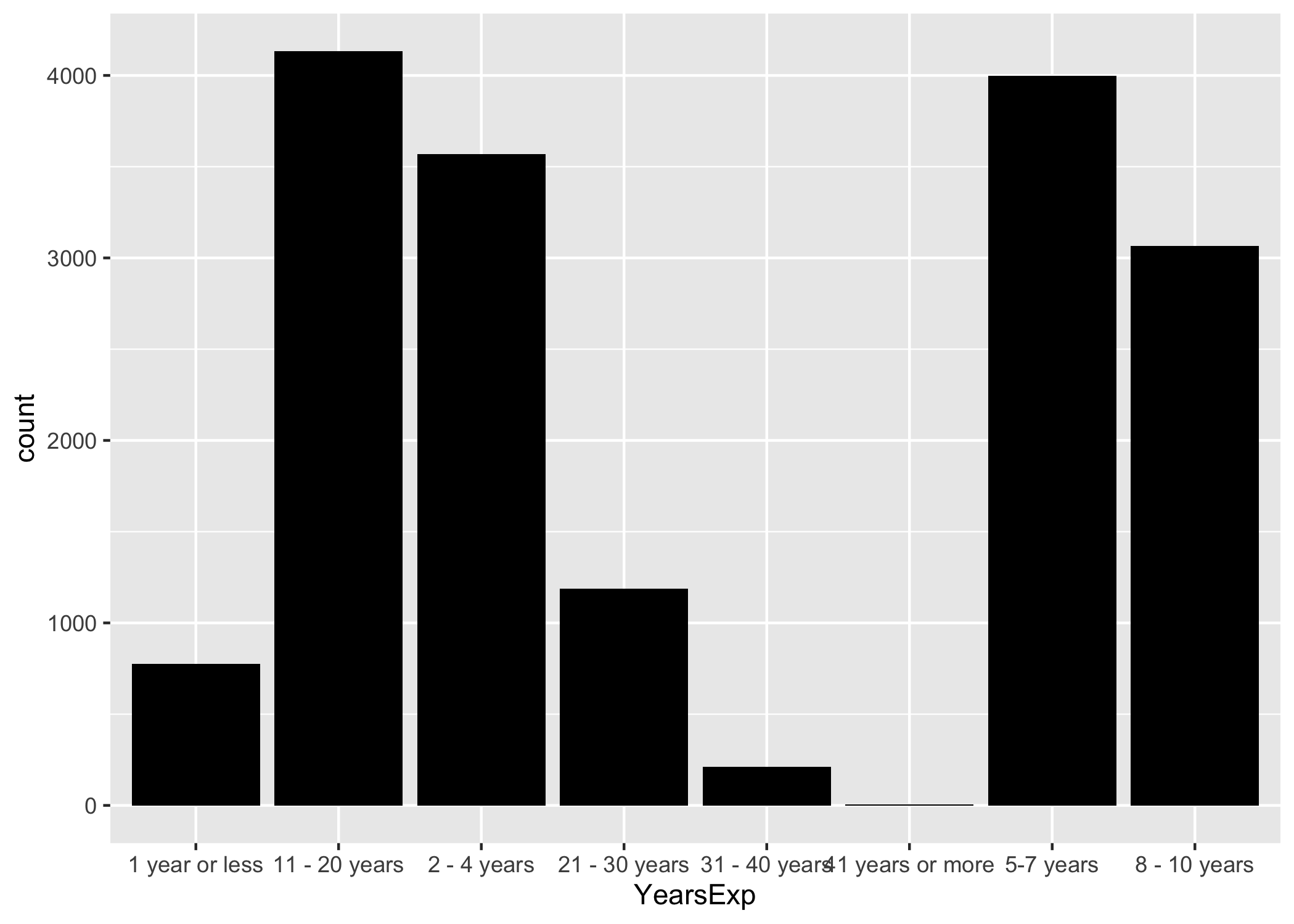
Similarly, we do the same thing for “years of total experience”. Figure 1.12 shows the original years of total experience data before clean up.  As shown above, the entries “41 years or above” have few datapoints and do not contain enough data for model training. Thus, we have to drop them.

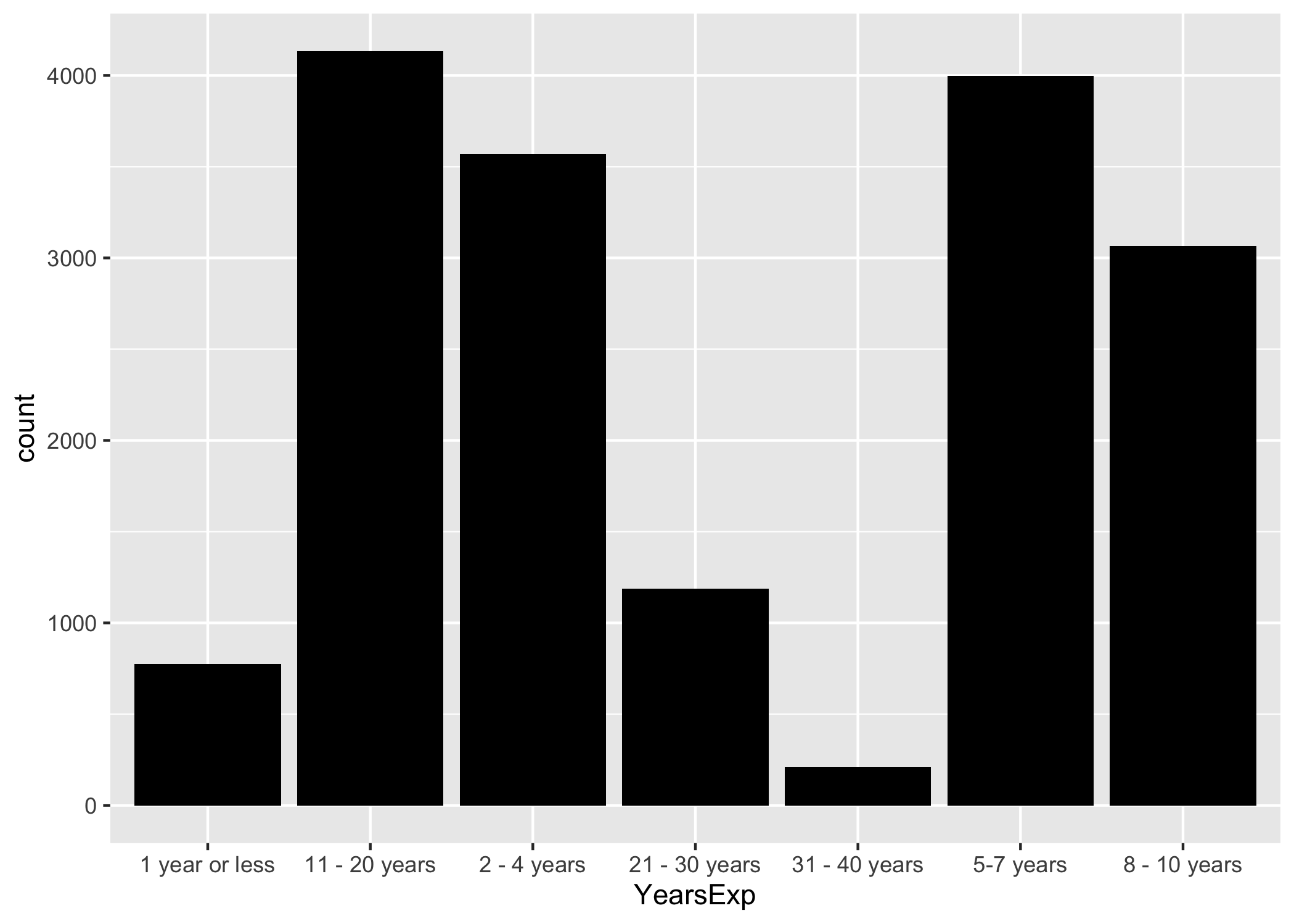
Figure 1.13 shows the years of total experience data after clean up. 

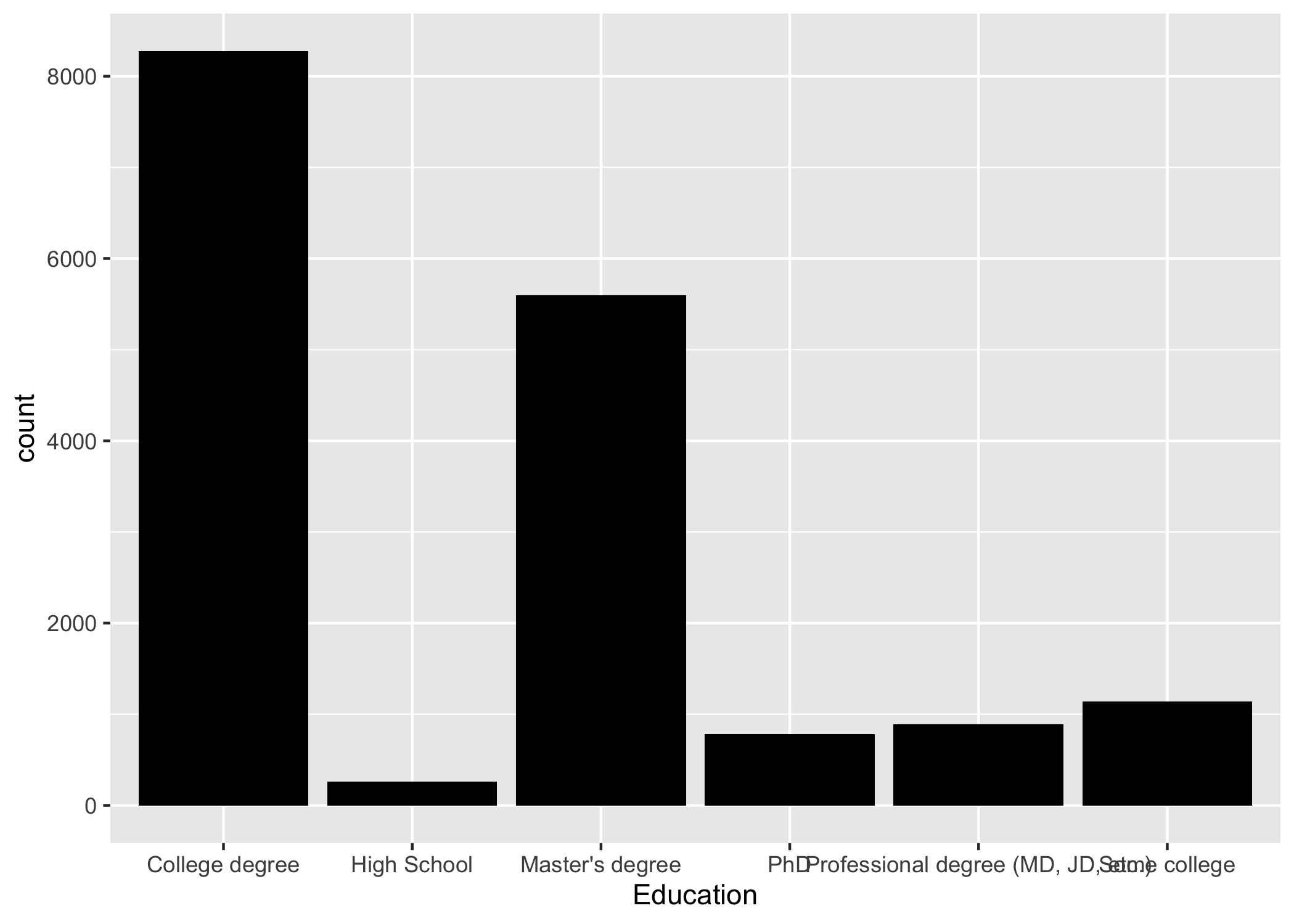
Figure 1.14 shows the original education data, which are decent and do not need clean-up. 

Figure 1.15 shows the original race data, which show little variation.  As shown above, the entries are mostly “white” and contains little variation. Therefore, this variable should be dropped.

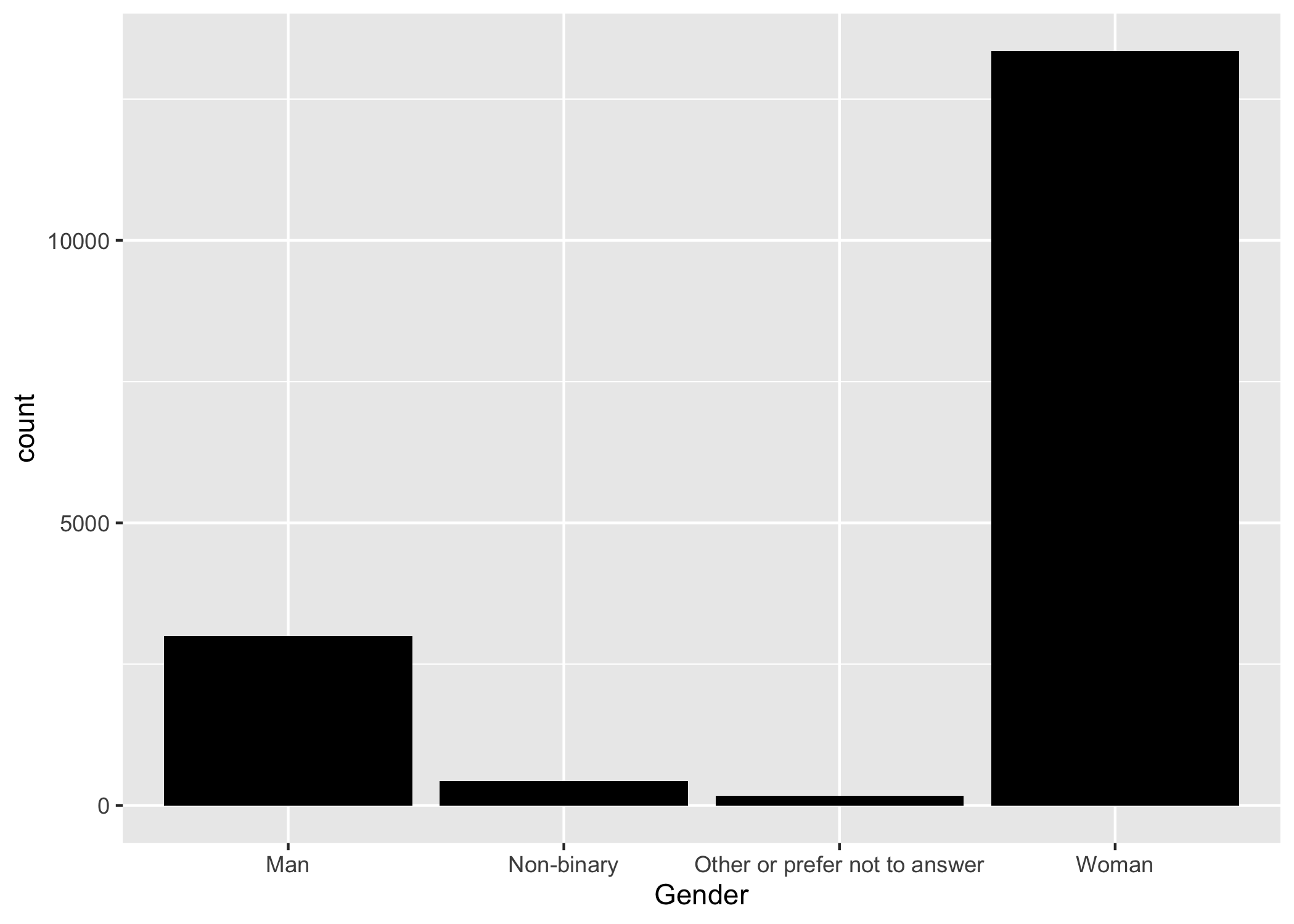
Figure 1.16 shows the original gender data before clean up.  As shown above, the entries “Other or prefer not to answer” have few datapoints and do not contain enough data for model training. Thus, we have to drop them.

Figure 1.17 shows the gender data after clean up. 