Data_Cleaning

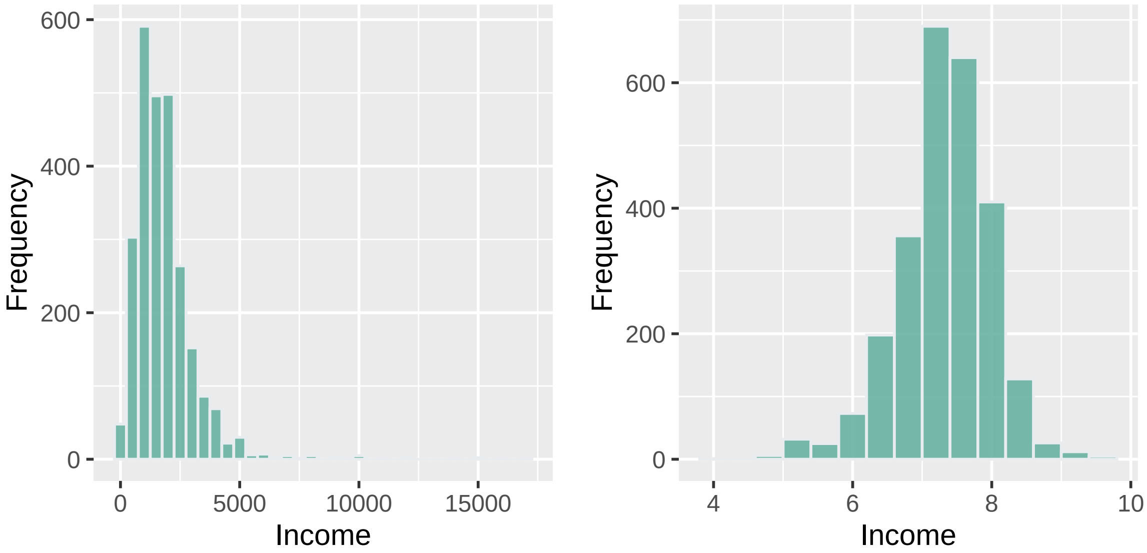
```r
head(da20520.0001[1:14])
```

```
##   CASEID   V1                 V2             V4                 V5      V7
## 1      1  257          (1) Miami  (07) School 7 (08) Eighth grade (1) Yes
## 2      2 2347          (1) Miami (13) School 13  (09) Ninth grade (1) Yes
## 3      3  860          (1) Miami (12) School 12  (09) Ninth grade  (2) No
## 4      4 5178 (3) Ft. Lauderdale (20) School 20  (09) Ninth grade (1) Yes
## 5      5 1984          (1) Miami (12) School 12  (09) Ninth grade (1) Yes
## 6      6 1067          (1) Miami  (07) School 7 (08) Eighth grade  (2) No
##               V8             V9 V10     V11     V13  V14            V15 V16
## 1          <NA> (044) Pakistan  85    <NA> (1) Yes <NA> (044) Pakistan  85
## 2          <NA>  (102) Ecuador  NA    <NA> (1) Yes <NA>  (102) Ecuador  NA
## 3 (1) Same city     (078) Cuba  69 (1) Yes (1) Yes <NA>     (078) Cuba  69
## 4          <NA>    (082) Haiti  78 (1) Yes (1) Yes <NA>    (082) Haiti  80
## 5          <NA> (101) Colombia  78  (2) No (1) Yes <NA> (101) Colombia  83
## 6 (1) Same city     (078) Cuba  68 (1) Yes (1) Yes <NA>     (078) Cuba  68
```

Our original dataset, "da20520.0001," is vast, consisting of 5262 observations with 665 variables. The chart above shows the first 14 columns of our dataset. Since we wanted to use the LASSO method to select the model, we had to do lots of data cleaning. Here are some steps that we used for the data cleaning.

(1) Our goal here is to select the best model with the most predictive power. Our response variable is income, so we concluded that we don't want rows with NA in `V421`, which represents the income of the individuals. We deleted all of such rows.

(2) For our prediction, we checked if our response variable, `V421`, is approximately normal. As you see on the left histogram, our original `V421` is heavily skewed, so we had performed a log transformation to make it approximately normal (See the right histogram).



(3) This data frame consists of results of all responses from a total of 4 different surveys. In other words, this data frame recorded the response of individuals who took the four different surveys at different stages of their lifetime. Since using all four surveys will create dependency in terms of time series, we decided to use only the responses collected when individuals were attending high school.

One of the big issues we faced for using LASSO for our variable selection was that we had many NA values across the data frame, and LASSO couldn't handle the NAs. To tackle this issue, we used many assumptions to either replace or delete the NAs.

(1) We assumed that the column with more than 50% of NAs is not informative, so those columns were removed.

(2) Second, for the categorical variables, we assumed that the columns with more than 25 categories are causing a problem with the predictive power of LASSO, so we removed all of such variables. For example, variables `V32`, `V37`, `V233`, `V238`, `V263`, `V62`, `V64`, `V264` have more than 100 categories (occupations), and they have only 2~3 data points for many of them. We assumed that this lack of data points in each category is causing problems in the predictive power of the LASSO. We wanted to ensure that each category had at least 100 data points for each column, so we divided our number of observations (around 2500) by 100, equal to 25. We know that this calculation is not great, but we used this because we needed a rule of thumb for the numbers of categories in each column.

(3) Since the LASSO couldn't handle the NAs, we used the imputation to replace the NA values. For the numerical variable, we used the mean of each variable to replace the NA. Here we did not want to simply replace the NA because it would raise a problem of non-response bias, so we included the dummy column with 1 and 0 so that we can still keep track of all NA values for each variable. For categorical variables, we just simply made NA as a level.