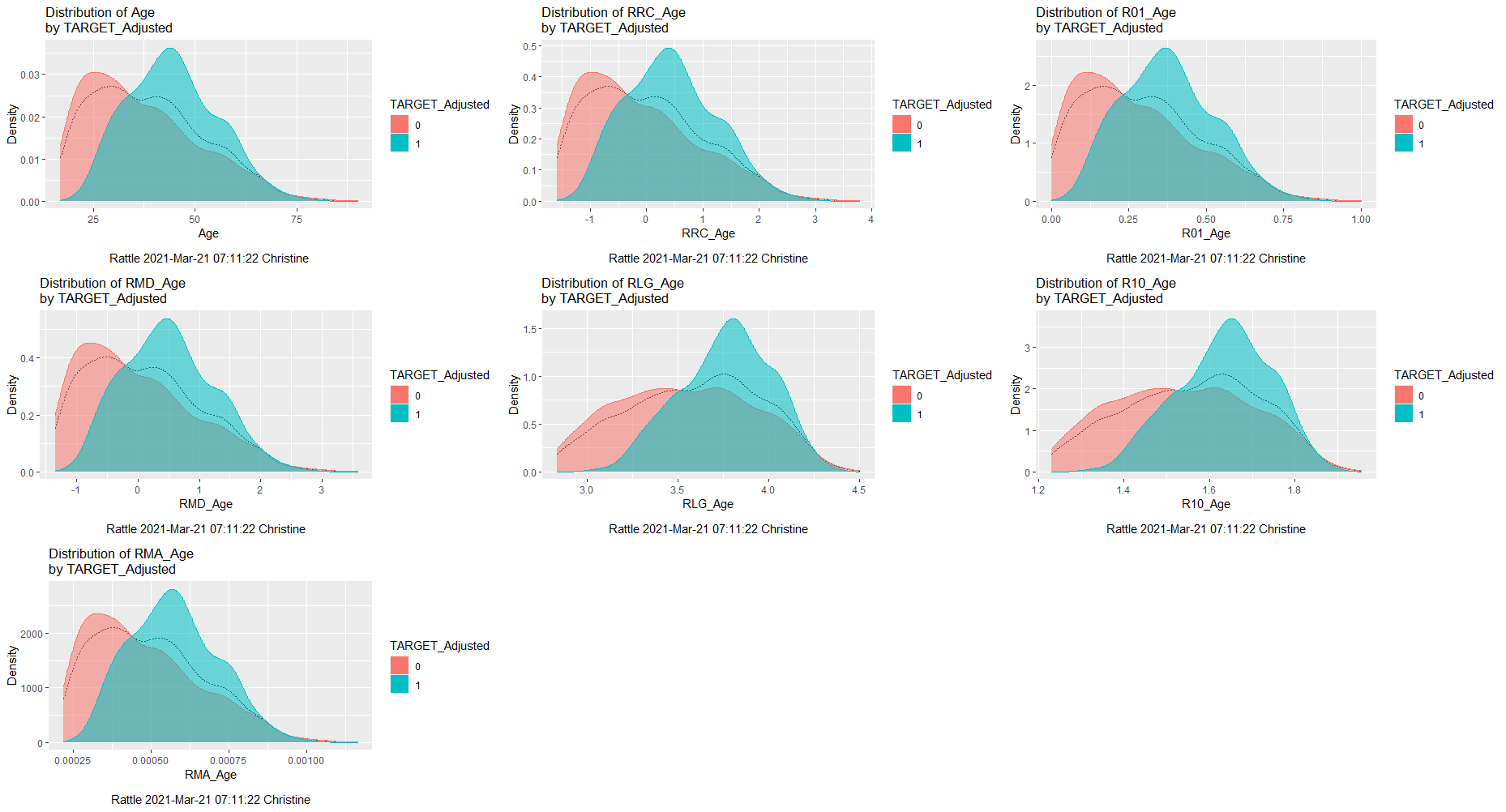
# Rescale - Age

**Q1:  Is Age a symmetric distribution?  Why or why not?**



When viewing the distribution shape for the two different groups, there are two almost equal peaks (mean=38; median=37) where the most frequently occurring values are close to each other, producing a bi-modal distribution. However, AGE is not a symmetrical distribution as the shape to the left and to the right of a vertical line drawn at mean/medium/mode are not mirror images of each other. Category 1 group (penalized with fees) is closer to being symmetric but still has a slight positive skew with tail towards the right. Category 0 group (no fees) has a more prominent positive skew with tail towards the right. All Rescale Normalize Transforms minus the Natural Log and Log 10 produce very similar shapes and larger ranges.

Rescaling using the Natural Log and Log 10 distributions produce a negative skew with tail towards the left as well as smaller ranges for Category 1 group. The Natural Log and Log 10 distributions also produce a flatter, positive skew with tail towards the right.

**Q2:  If we decide to use a Rescale Recenter (RRC) transformation on Income do we need to use one on Age? Why or why not?**

The INCOME variable ranges from 609.72 to 481259.50; whereas the AGE variable ranges from 17 to 90, making the scale of INCOME quite a bit larger than AGE. When using distance measures, that size of a variable (INCOME) may overwhelm/outweigh any influence by AGE. So, transforming both INCOME and AGE will bring the data to comparable scales, equalizing the range.

**Q3:  Which transformations change the scale of a numeric variable?  Which might change the shape?**

To change the scale of a numeric variable, use:

* Recenter
* Scale [0-1]
* -Median/MAD
* Matrix

To change the shape of a numeric variable, use:

* Natural Log
* Log 10

Your notes:

* Rescale = transform numeric variables to standard scales
* AGE and INCOME variables are a good example for rescaling as the variables are not approximately the same size or roughly same order of magnitude
  + AGE varies from about 17 to about 80 or 90
  + INCOME varies up to around $500,000.00
  + Scale of income is a lot larger than age
  + When using distance measures, that size of a variable may overwhelm any influence by age
* Rescale Function has 2 major categories
  + Normalize
    - Recenter = rescale mean to zero and standard deviation is 1
    - Scale [0-1] = rescale to the range zero to one
    - -Median/MAD (median absolute deviation) = rescale median=0 & MAD=1  
      (Similar to recenter, median is used instead of mean)
    - Natural Log = rescale using a natural log transform
    - Log 10 = rescale using a log 10 transform
    - Matrix = divide each cell of selected numeric variables by matrix total
  + Order (focusing on Normalize function for this lesson)
    - Rank = used when more interested in position of value versus value itself, scoring new observation may prove difficult since distribution has already been calculated
    - Interval = rescale to integer 0-99 (using default of 100 groups), use to break numeric variable into categorical variable

# Impute

**Q4:  Missing Value Table**

| **Variable** | **Type** | **Number Missing** | **Comments** |
| --- | --- | --- | --- |
| Age | Numeric | 0 | There is no missing data so there will be no significate issues. |
| Employment | Categoric | 100 | Out of 2000 cases, approximately 5% of the Employment category data are missing, giving us 100 missing observations for this variable. As noted in *Figure Q4-1*, there are 8 different levels and/or categories. It is very probable that the available options (see *Figure Q4-2* below) confused the participants since, for example, a person could be a Consultant in the Private sector and consider themselves a Volunteer, so they left the field blank. In addition, Week 2’s Correlation Plot (*Figure Q4-3*) displays that the Employment and Occupation variables are highly correlated. As a result, the missing data can be classified as Missing at random (MAR). Typically, in these instances, the mechanism producing the missing data can be ignored and once the missing data are replaced or deleted the relationships of interest can be modeled directly. Lastly, it seems very likely that this missing data will not be an issue as income/age seem to have a better relationship. |
| Education | Categoric | 0 | There is no missing data so there will be no significate issues. |
| Marital | Categoric | 0 | There is no missing data so there will be no significate issues. |
| Occupation | Categoric | 101 | This is almost identical to the Employment variable comment. Main differences being 5.1% Occupation category data are missing, giving us 101 missing observations and, as noted in *Figure Q4-1*, there are 14 different levels and/or categories. Again, the missing data can be classified as MAR. As a result, it seems very likely that this missing data will not be an issue. |
| Income | Numeric | 0 | There is no missing data so there will be no significate issues. |
| Gender | Categoric | 0 | There is no missing data so there will be no significate issues. |

Figure Q4-

**A picture containing table

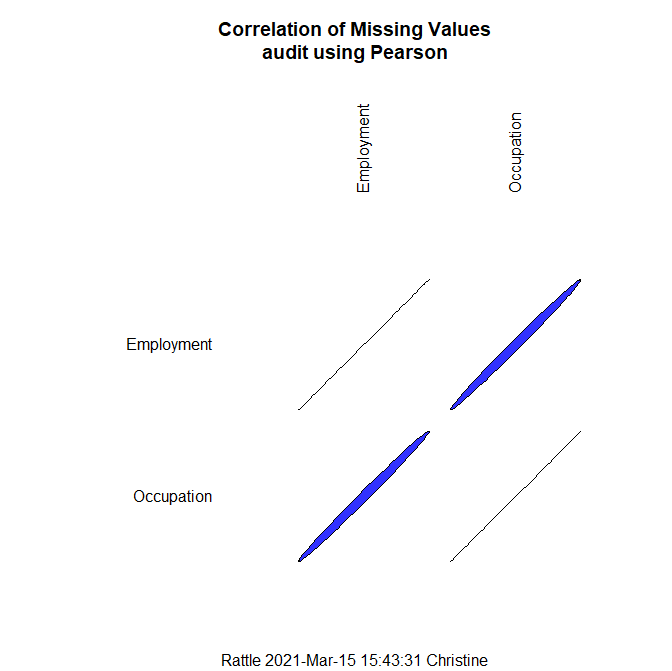
Description automatically generated**

Figure Q4-

**Table

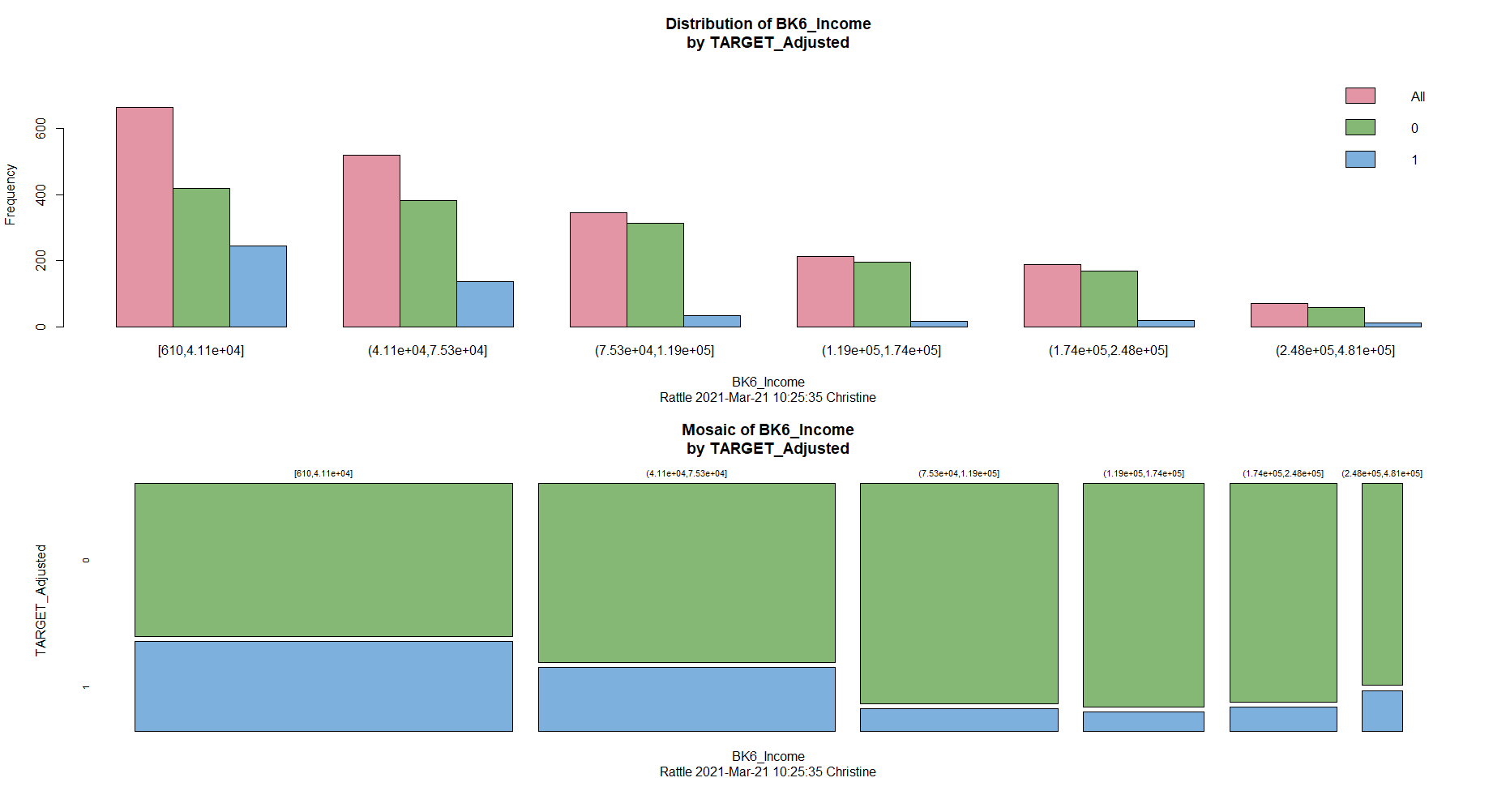
Description automatically generated**

Figure Q4-



# Recode/ Binning/ Kmeans/ Income

**Q5: Using the Binning technique of Kmeans generate a variable that uses Income as the source and has 6 bins. What can you say about the possible relationship of Income and Incurring an Audit Adjustment? Comment on the advisability using this new variable as a strategy to sample the population.**



Using the Binning technique of Kmeans, a new variable, named BK6\_Income, was created that uses Income as source and has 6 bins. When reviewing the Mosaic of BK6\_Income graph, audit adjustments are most likely to occur for people who make between $610 to $41,100 on an annual basis. Individuals who make between $75,300 and $248,000 per year are the least likely to have audit adjustments. However, there is a slight spike in audit adjustments for individuals within the $248,000 and $481,000 per year range.

When choosing strategy for sampling the population, it is recommended to concentrate sampling resources in the first two groups ($610 to $41,100 and $41,100 to $75,300), oversampling them in a stratified sample. It would be best to also include the third group ($75,300 to $119,000) and under sample this category. However, do be careful when following this strategy and confirm that age has does or does not have a significant correlation with annual income.

# Rescale - Income

Diagram

Description automatically generated

Your notes:

* Income
  + Large range [609.72 to 481259.50]
  + Skewed to right
  + Unique=2000; Mean=84688.46; Median=59768.95
* RRC\_Income
  + Mean set to 0, standard deviation=1
  + Collapsed to lower range [-1.21 to 5.70]
  + Range closer to AGE but still not there
  + AGE is now a larger range than INCOME
  + Still skewed to right, little changed
  + Unique=2000; Mean=-0.00; Median=-0.36
* R01\_Income
  + Range is now 0 to 1
  + No change in distribution
  + Still skewed to right
  + Unique=2000; Mean=0.17; Median=0.12
* RMD\_Income
  + Range = -1.25 to 8.92
  + Median set to 0, MAD=1
  + No change in distribution
  + Still skewed to right
  + Unique=2000; Mean=0.53; Median=0.00
* RLG\_Income
  + Rescale from 7 to 13 [6.41 to 13.08], quite a bit of change here
  + Distribution closer to being symmetric (mean and median are very close to each other)
  + Closer to fulfilling assumption that distribution is normal
  + Could possibly use in Regression without violating normal distribution
  + Unique=2000; Mean=11.02; Median=11.00
* R10\_Income
  + Rescale from 3 to 5 [2.79 to 5.68]
  + Distribution closer to being symmetric (mean and median are same value)
  + Closer to fulfilling assumption that distribution is normal
  + Could possibly use in Regression without violating normal distribution
  + Unique=2000; Mean=4.78; Median=4.78
* RMA\_Income
  + Range is now 0 to 0.003 [0.00 to 0.00]
  + No change in distribution
  + Still skewed to right
  + Unique=2000; Mean=0.00; Median=0.00]

# Impute

Your notes:

* Missing Values
  + Pretty much in almost every dataset
  + Can have significant impact on analysis results
* Impute
  + R missing value = NA
  + Implies replacing missing value with something else (where the value is NA)
  + Questions to ask:
    - Significant number of missing values in any given variable/combination of variables?
    - Missing values impact analytical method used?
      * R automatically eliminates any case where input variable is missing
      * Decision trees are less of a problem
    - Variable type = categorical or numeric
  + Rattle Methods
    - Zero/Missing: numeric=0; categories=’missing’
    - Mean: replace missing values with population mean
    - Mode: replace missing values with population mode
    - Constant: replace missing values with specified value
  + Approaches to missing values
    - Remove: use only complete cases, will reduce number of observations which can be issue with small dataset as well as introduction bias
    - Impute: use Rattle’s method
    - Model: doing a sub-analysis within your analyses, not imputation because those values will vary, need to investigate and/research
* Resource: Kabacoff, Robert I. *R in Action: Data Analysis and Graphics with R.* Shelter Island, NY: Manning Publications, 2011. Chapter 15

# Recode/ Binning/ Kmeans/Age

Chart, bar chart

Description automatically generated

Your notes:

* Recode
  + Remap variables through…
    - Binning (uses numeric variables)
    - Changing variable type (use categoric variables)
    - Can change categorical to numeric
  + Binning
    - Partition numeric variable into several bins (buckets)
    - Reduces information provided by individual values of the variable, reduction being meaningful is judgement call
    - Use to…
      * Produce visualizations of data (i.e., mosaic plots)
      * Set as a stratifying variable in various plots
      * Simplify models
    - Quantiles
      * Each bin has approximately same number of observations
      * Observations are weighted if weight variable is present on data tab
    - Kmeans
      * Clustering used to bin variable
      * Common clustering algorithm covered in another course
        + Clustering = method of discovering groups,
        + Partition numeric variables in a way that maximizes the differences in the groups and more meaningful
        + Kmeans uses a geometric interpretation of data as points in space
        + Chapter 13 discusses further
      * Kmeans to bin Age
        + Bins not in numerical order but by count (binning process
        + 17-25 age group very few audit adjustments
        + 42,51 and 51,62 age groups where most adjustments are made
        + EXAMPLES:

May want to concentrate sampling resources in 42,51 and 51,62…oversample them in stratified sample

Will want to include 17,25 group and under sample this category

Careful when doing this as there may be another thing at play….worry about how work out from economic standpoint

If goal is maximize number of people found, this works just fine

If goal is to maximize number of adjustments found…this does not account for that so need satisfy self that age has no significant correlation with dollar size adjustments

* + - Equal Width = Min to max range of continuous numeric variable will be split into equal width bins
    - Number = set number of bins to construct for any of above methods
  + Type changing
    - Not on Rattle GUI but underneath Binning and related options
    - Indicator Variable
      * Turn category into collection of numeric (0, 1) variables
      * Called dummy variables
      * Useful in regression analysis for handling categoric variables
      * EXAMPLE:
        + Category = 3 levels
        + 3 new (0, 1) variables are generated
        + If observation in 1st category, then it is assigned 1
        + Remaining is assigned 0
        + 3 level category create 2 new variables
        + Rattle generates all three, assign Ignore to first one
    - Join Categories
      * Combine multiple categoric variables into one
      * EXAMPLE: Gender and Ethnicity (missing values may be issue)
      * Bin numeric variable and combine with categoric
    - As Categoric = convert numeric to categoric by turning each distinct value of numeric variable into level for categoric variable
    - As Numeric
      * Convert categoric to a numeric
      * Replacing each level with the numeric index of the level
      * CAUTION: careful doing this and using in other methods, implicit assumption if you take that and assume interval type variable….meaning level 5 is 5 times bigger than 1….majority of the time this is not the case