**BIAM510 Week 2 iLab Correlational Summarizations**

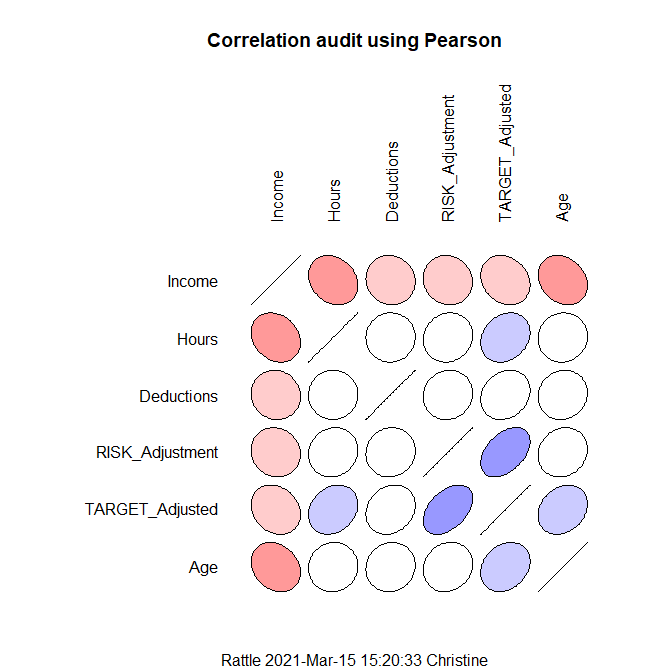
Name: \_\_\_Christine Baxter\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_March 14, 2021\_\_

**Step 8: Correlational Summarizations (Video 2.3)**

# Correlation Plot

Copy/paste correlation plot below:

Correlation summary using the 'Pearson' covariance.



Correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

Income Hours Deductions RISK\_Adjustment TARGET\_Adjusted Age

Income 1.00000000 -0.21269065 -0.05734147 -0.08339021 -0.1981176 -0.22686777

Hours -0.21269065 1.00000000 0.01365124 0.09060735 0.2106816 0.04236487

Deductions -0.05734147 0.01365124 1.00000000 0.06559720 0.1835169 0.08399899

RISK\_Adjustment -0.08339021 0.09060735 0.06559720 1.00000000 0.4415193 0.12274079

TARGET\_Adjusted -0.19811763 0.21068160 0.18351693 0.44151926 1.0000000 0.23400690

Age -0.22686777 0.04236487 0.08399899 0.12274079 0.2340069 1.00000000

Rattle timestamp: 2021-03-15 15:20:33 Christine

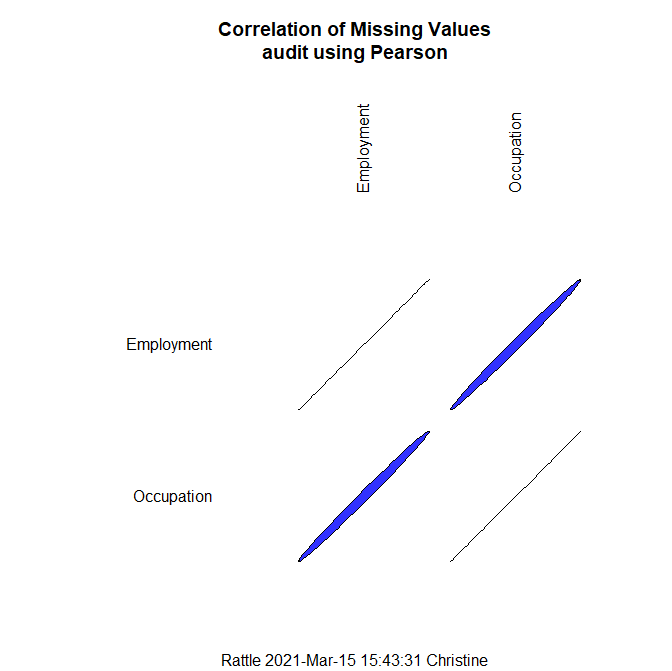
======================================================================

Your notes:

* Strength of correlation between any two variables
* Straight line = 1:1 correlation (i.e., income is correlated with income; hours to hours, deductions to deductions, etc.)
* Round circle with white in middle = almost zero correlation (hours and deductions)
* Symmetric matrix, both sides of diagonal are mirror of each other so you can confine yourself to one side
* Direction of the ellipse tells you something about the direction of the correlation
  + Pointing to left = negative correlation (i.e., income and hours)
  + Pointing to right = positive correlation (i.e., target\_adjusted and hours)
* Shading also depicts strength of correlation, darker the color…the higher correlation that exists
* Quick way of visualizing and scanning for important sets of relationships that might be used/useful
* Look for high degree of correlation between independent variables that may cause concern and may want to handle in later analyses

# Missing Value Correlations

Copy/paste missing value correlation plot below:



Your notes:

* Takes all variables together and Looks at missing variables and displays graphically
* Results tell us there is a high degree of correlation between Occupation and Employment
* After looking at all other possible combinations (including categorical and numeric) variables, Rattle found this is only significant correlation and high degree of correlation
* Thin, dark, blue line signifies strong correlation between two variables

Answer **any one** (your choice) of the following 2 questions:

Q18: What is the direction of the relationship of the variables Age and Income based on the correlation diagram?  
HINT: Examine the direction the oval is tilted.

The direction (oval pointing to left) and darker color of the oval at intersection of Age and Income signifies a strong, negative correlation.

Q19: What might explain the high correlation between the variables Employment and Occupation?  
HINT: Consider how these variables are defined.

The employment and occupation variables themselves are very similar in nature and relatable. For example, the employment category of Volunteer could also encompass Support, Home, and Service from the occupation category list. In addition, Self-Employed (SelfEmp) could encompass any number of categories from the occupation category list. And, PSFederal could encompass Clerical, Executive, Professional, Service, Support, or Transport occupation categories. Those are just a few examples to clarify the ambiguity/similarity of these two variables and their categories. Consolidating and reducing the number of category levels would greatly improve further summarization/analyses.

+----------+------------------------------------------------------------------------------------+

|Variable |Levels |

+----------+------------------------------------------------------------------------------------+

|Employment|Consultant, Private, PSFederal, PSLocal, PSState, SelfEmp, Unemployed, Volunteer |

+----------+------------------------------------------------------------------------------------+

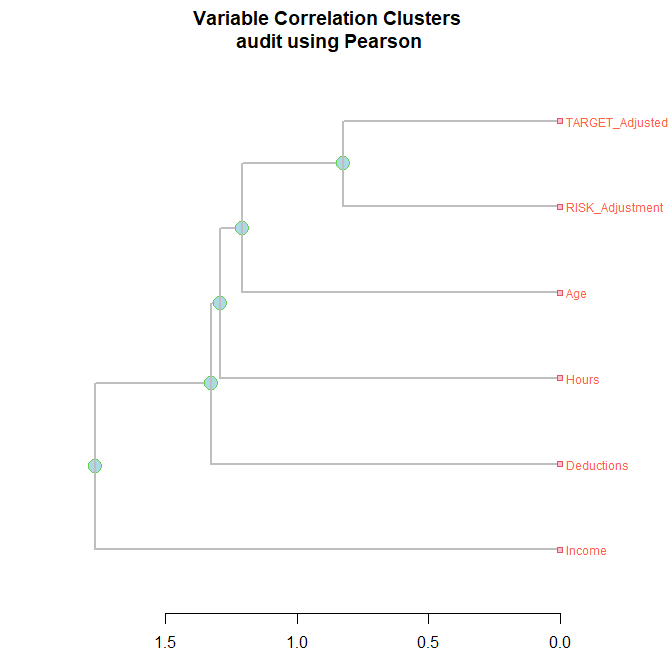
|Occupation|Cleaner, Clerical, Executive, Farming, Home, Machinist, Military, Professional, |

Protective, Repair, Sales, Service, Support, Transport |

+----------+------------------------------------------------------------------------------------+

# Hierarchical Correlation

Copy/paste hierarchical correlation plot (dendrogram) below:



Missing values correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

Employment Occupation

Employment 1.0000000 0.9947753

Occupation 0.9947753 1.0000000

Count of missing values:

Occupation Employment

101 100

Percent missing values:

Occupation Employment

5.05 5.00

Rattle timestamp: 2021-03-15 15:43:31 Christine

======================================================================

Your notes:

* Creates Dendrogram….like a decision tree laid on its side
* Provides overview of correlation between variables
* Max length of X axis = 3.0
* Length of line indicates amount of difference
* Shorter line means higher correlation
* Blue circles join related variables/groups
* Longer line, less likely there is correlation
* Starts at the top with shortest line and most correlated and works way down
* As you go up the Dendrogram, lines get shorter and correlation becomes higher

Answer the following question:

Q20: Using the information in the dendrogram, which variable Age or Income is likely to influence the variable Target Adjusted the most?  
HINT: Consider the lengths of the lines connecting the variables.

The Age variable is most likely to influence the variable TARGET\_Adjusted as it appears in the top part of the Dendrogram with a shorter line than the Income variable. The Age variable’s line length and location within the graph depict a higher correlation than the TARGET\_Adjusted variable.

# Grand Summary of Notes:

Review and briefly summarize key ideas from your notes for all three documents in this iLab, using a green font.

## 2.1 Numerical Summarizations

### Summary/Summary Results

* Possible issues with the Marital Status categorical variable, need further definition or categorization (consolidating and reducing number of levels)
  + Absent = No selection?
  + Unmarried = Divorced and/or Widowed?
  + Married = Absent and/or Married-spouse-absent
* Possible issues with the Occupation categorical variable, need further definition or categorization (consolidating and reducing number of levels)
  + Service include Cleaner, Machinist, Professional, Protective, Repair?
  + Support include Cleaner, Clerical, Farming?
  + Sales include Executive or Professional?
  + Professional/Executive include Sales, Service, Clerical, Transport?
* Review case with 1 missing observation in Occupation category
* Review observation data for Education, see if further summarization would reduce the number of (Other) category
* Deductions related to marital status and/or income?

Chart, treemap chart

Description automatically generated

### Summary/Describe Results

* 28-year-old’s fall within 25% of the mean (38.62)
* Majority fall between the ages of 28 and 48

### Summary/Basic Results

* Low variance = values stay near the average
* High variance = values move away from the average
* Standard Deviation = no squares so unit matches unit of observation

### Summary/Kurtosis Results

* Perfectly normal distributed data have a kurtosis value of 0
* Normally distributed data establishes the baseline for kurtosis
* Positive kurtosis value = heavier tails and a sharper peak
* Negative kurtosis value = lighter tails and a flatter peak
* Larger values mean sharper peaks and flatter tails
* Positive values indicate an acute peak around the mean
* Negative values indicate a smaller peak around the mean

### Summary/Skewness Results

* Symmetrical or non-skewed distributions = skewness value approaches zero
* Positive or right skewed = skewness value > 0 (tail points to the right)
* Negative or left skewed = skewness value < 0 (tail points to the left)

## 2.2 Visual Summarizations

### Age/Box Plot

* TFC target adjusted = comparing the distribution of entire data set to the distribution for people who did not get their taxes adjusted because of the audit and people who did get their taxes adjusted because of the audit
* Category Zero (0) = the people that did not to have to have any sort of additional payments
* Category One (1) = those that had to have additional payments

### Age/Histogram

* TFC target adjusted = comparing the distribution of entire data set to the distribution for people who did not get their taxes adjusted because of the audit and people who did get their taxes adjusted because of the audit
* Category Zero (0) = the people that did not to have to have any sort of additional payments
* Category One (1) = those that had to have additional payments.

### Age/Cumulative Distribution

No notes added to this section.

### Age/Benford

No notes added to this section.

### Income/Box Plot

* No overlap of 95% confidence interval around the median, showing there is a statistically significant difference between the values of these two groups
* Category 1 has many outliers
* Size of boxes indicate people that did not receive additional penalties from the audit had larger incomes than those that had additional payments due to penalties

### Income/Histogram

* TFC target adjusted = comparing the distribution of entire data set to the distribution for people who did not get their taxes adjusted because of the audit and people who did get their taxes adjusted because of the audit
* Category Zero (0) = the people that did not to have to have any sort of additional payments
* Category One (1) = those that had to have additional payments
* Combined group (All): the frequency peaks around 275 with an income range of $30,000 to $40,000
* Significant drop in frequency when income is > $40,000
* Significant increase in frequency when income is > $20,000

### Income/Cumulative Distribution

* Approximately 30% increase in income from $50,000 to $125,000
* 95% of people with income > $200,000 pay penalties

### Income/Benford

Income for Category 0 and Category 1 follows Benford’s Law; however, income levels for combined group (All) does not seem to completely follow Benford’s Law.

### Employment/Bar Plot

* Large frequency number for Private employment within Category 0 versus Category 1
* Category Zero’s frequency number for Private employment is approximately 77% more than Category One’s

### Employment/Dot Plot

* All employment categories other than Private have a frequency number <200
* Majority of Private employment category did not have to pay penalties

### Employment/Mosaic Plot

Higher number of Self Employed (SelfEmp) workers paid penalties versus the other employment categories.

### Marital Status/Bar Plot

No notes added to this section.

### Marital Status/Dot Plot

No notes added to this section.

### Marital Status/Mosaic Plot

* Married people seem more likely to pay penalties
* All Marital Status categories except for ‘Married’ are less likely to pay penalties

### Scatter Plot Matrix

* Age and Marital Status
  + Positive correlation between variables
  + Potential correlation/relation between independent variables
  + .49 = .52 = .25 = 25% of variance
* Age and Income
  + Continuous variables
  + Negative correlation
  + -.212 = .04 = 4% of variance
  + Two numeric variables are not correlated, no violation of assumption
* Age and Gender
  + Negative correlation
  + -.42 = .1764 = 17.6% of variance

### Employment x Age Pairwise Comparison

* Target = Employment Classification
* Outliers exist for…
  + Federal sector (1 outlier = 76 years old)
  + Private section (5 outliers = 74- to 83-years-old)
* Larger sized boxes indicate more people within that sector for the specified age range
* Adults between the ages of 53 and 57 are more likely to be self employed
* Adults between the ages of 50 and 53 are more likely to be consultants
* Federal sector…
  + Employs adults between 29 and 47 years of age
  + <25% within the 17 to 27
* Local and State sectors employ adults between the ages of 18 and 73 with half of the values falling in the 32 to 50 range
* Private sector seems more likely to hire younger workers

### Employment x Income Pairwise Comparison

* Target = Employment Classification
* Larger sized box indicates more people with specified income by Employment
* Outliers
  + Potential issue with number of outliers for Consultants, Private sector, and Self-Employed workers
  + Federal sector has two outliers
* Income for Private Sector workers within the interquartile range from $35,271.88 to $122,081.92 with a median value of $64,633.82.

### Gender x Income Pairwise Comparison

* The 95% confidence interval around the median does not overlap which means there is a statistically significant difference between the values of these two groups.
* Outliers exist for…
  + Males = large number of outliers
  + Females = 8 outliers ranging from $330,573 to $481,255
* Income range for…
  + Males = $29,960.07 to $80,133.13
  + Females = $69,453.85 to $176,175.66
* Income ranges for Females are distinctly higher than male’s income ranges.

## 2.3 Correlational Summarizations

### Correlation Plot

* Strength of correlation between any two variables
* Straight line = 1:1 correlation (i.e., income is correlated with income; hours to hours, deductions to deductions, etc.)
* Round circle with white in middle = almost zero correlation (hours and deductions)
* Symmetric matrix, both sides of diagonal are mirror of each other so you can confine yourself to one side
* Direction of the ellipse tells you something about the direction of the correlation
  + Pointing to left = negative correlation (i.e., income and hours)
  + Pointing to right = positive correlation (i.e., target\_adjusted and hours)
* Shading also depicts strength of correlation, darker the color…the higher correlation that exists
* Quick way of visualizing and scanning for important sets of relationships that might be used/useful
* Look for high degree of correlation between independent variables that may cause concern and may want to handle in later analyses

### Missing Value Correlations

* Takes all variables together and Looks at missing variables and displays graphically
* Results tell us there is a high degree of correlation between Occupation and Employment
* After looking at all other possible combinations (including categorical and numeric) variables, Rattle found this is only significant correlation and high degree of correlation
* Thin, dark, blue line signifies strong correlation between two variables

### Hierarchical Correlation

* Creates Dendrogram….like a decision tree laid on its side
* Provides overview of correlation between variables
* Max length of X axis = 3.0
* Length of line indicates amount of difference
* Shorter line means higher correlation
* Blue circles join related variables/groups
* Longer line, less likely there is correlation
* Starts at the top with shortest line and most correlated and works way down
* As you go up the Dendrogram, lines get shorter and correlation becomes higher