**BIAM510 Week 2 iLab Numerical Summarizations**

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

**Step 3: Looking at the Numbers (Video 2.1.2.1)**

# Summary/Summary Results

Copy/paste Summary output from Rattle below:

Chart, treemap chart

Description automatically generated

Below we summarise the dataset.

Note that the data contains 101 observations with missing values.

Enable the 'Show Missing' check box for details.

Data frame:crs$dataset[, c(crs$input, crs$risk, crs$target)]

2000 observations and 11 variables Maximum # NAs:101

Levels Storage NAs

Age integer 0

Employment 8 integer 100

Education 16 integer 0

Marital 6 integer 0

Occupation 14 integer 101

Income double 0

Gender 2 integer 0

Deductions double 0

Hours integer 0

RISK\_Adjustment integer 0

TARGET\_Adjusted integer 0

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

|Variable |Levels |

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

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

| |Volunteer |

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

|Education |Associate,Bachelor,College,Doctorate,HSgrad,Master,Preschool |

| |Professional,Vocational,Yr10,Yr11,Yr12,Yr1t4,Yr5t6,Yr7t8,Yr9 |

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

|Marital |Absent,Divorced,Married,Married-spouse-absent,Unmarried,Widowed|

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

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

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

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

|Gender |Female,Male |

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

For the simple distribution tables below the 1st and 3rd Qu.

refer to the first and third quartiles, indicating that 25%

of the observations have values of that variable which are

less than or greater than (respectively) the value listed.

Age Employment Education Marital

Min. :17.00 Private :1411 HSgrad :660 Absent :669

1st Qu.:28.00 Consultant: 148 College :442 Divorced :266

Median :37.00 PSLocal : 119 Bachelor :345 Married :917

Mean :38.62 SelfEmp : 79 Master :102 Married-spouse-absent: 22

3rd Qu.:48.00 PSState : 72 Vocational: 86 Unmarried : 67

Max. :90.00 (Other) : 71 Yr11 : 74 Widowed : 59

NA's : 100 (Other) :291

Occupation Income Gender Deductions

Executive :289 Min. : 609.7 Female: 632 Min. : 0.00

Professional:247 1st Qu.: 34433.1 Male :1368 1st Qu.: 0.00

Clerical :232 Median : 59768.9 Median : 0.00

Repair :225 Mean : 84688.5 Mean : 67.57

Service :210 3rd Qu.:113842.9 3rd Qu.: 0.00

(Other) :696 Max. :481259.5 Max. :2904.00

NA's :101

Hours RISK\_Adjustment TARGET\_Adjusted

Min. : 1.00 Min. : -1453 Min. :0.0000

1st Qu.:38.00 1st Qu.: 0 1st Qu.:0.0000

Median :40.00 Median : 0 Median :0.0000

Mean :40.07 Mean : 2021 Mean :0.2315

3rd Qu.:45.00 3rd Qu.: 0 3rd Qu.:0.0000

Max. :99.00 Max. :112243 Max. :1.0000

Rattle timestamp: 2021-03-14 19:01:55 Christine

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

Missing Value Summary

Age Education Marital Income Gender Deductions Hours TARGET\_Adjusted

1899 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1

100 1 1 1 1 1 1 1 1

0 0 0 0 0 0 0 0

Employment Occupation

1899 1 1 0

1 1 0 1

100 0 0 2

100 101 201

Rattle timestamp: 2021-03-14 19:01:56 Christine

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

Answer **all** of the following 5 questions below:

Q1: What potential problems do you see in the Marital Status variable?   
HINT: Consider the definition of Absent.

Of the six Marital Status variables, the ‘Absent’ variable could have multiple definitions/meaning. More specifically, does ‘Absent’ mean the person is married and them/their spouse is absent or does it mean the person did not complete this section and there is an absence of selection. For the latter, the summarization data would be adversely affected since ‘Absent’ category is not considered ‘missing’ when in fact the person did not select an answer for this question. In addition, one person could have answered ‘Unmarried’ when another person would have chosen ‘Widowed’ and yet another person chose ‘Divorced’ when in fact all three are technically unmarried. A similar situation could arise with ‘Married’, one married person could simply answer ‘Married’ while another one answered ‘Married-spouse-absent’ and yet another chose ‘Absent’. Again, this data confusion would affect the confidence level for any marital status summarization/analyses that were based on this data.

If viewing this dataset under the view of ‘Absent’ means no selection, it would not be adequate for any marital status analyses since 33% of the 2000 values are missing. If an analysis was interested in unmarried/single participants and the analyst only used the ‘Unmarried’ statistic, a low percentage (3%) of unmarried individuals would be reported; however, if the analyst combines the ‘Unmarried’, ‘Divorced’, and ‘Widowed’ categories they would report a 20% level of unmarried individuals. This is definitely an area that could fall under the ethical/unethical reporting aspects of data analyses.

Q2: What do you consider to be the most important characteristic of the Employment distribution?  
HINT: Consider whether certain categories are much larger or smaller than others.

The top Employment category encompasses the Private sector which includes 1,411 (71%) of the audited records while the Self Employed and PSState sectors are on the lower end of the audited records. There is a 90% drop in audited records for the Consultant category.

Q3: What do you consider to be the most important characteristic of the Income distribution?  
HINT: Consider the shape of the distribution and its basic parameters such as mean, median, mode, and range.

The mean/average income level ($84,688.50) seems to be the most important characteristic of all audited records with the median ($59,768.90) and 3rd quartile ($113,842.90) being next important. Since the mean and median are not approximately equal, there is a good chance we are dealing with a non-symmetrical distribution. In addition, the 3rd quartile number indicates that 25% of the observations are greater than $113,842.90.

Q4: What do you consider to be the most important characteristic of the Deductions distribution?  
HINT: Consider whether certain values occur much more frequently than others.

The mean of $67.57 seems to be the most important characteristic of all audited records and it may be best to use the mean for a more typical measure of center.

Q5: How would you summarize the relative quality of the data so far? (Especially in regards to missing observations.)

In regard to missing observations, the relative quality of the data is categorized as good since 95% (1,899) of the cases are complete. Out of the 5% (101) of cases with missing data, there are 100 cases missing both the Employment and Occupation categories values and only 1 case missing Occupation data. When considering the categorical variables, the data quality definitely decreases due to the ambiguity of the Marital and Occupation data observations/responses.

Your notes:

* 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?

**Step 4: Looking at the Numbers Using a Deeper Analysis (Video 2.1.2.2)**

# Summary/Describe Results

Copy/paste Describe output from Rattle below:

Below is a description of the dataset.

crs$dataset[, c(crs$input, crs$risk, crs$target)]

11 Variables 2000 Observations

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Age

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95

2000 0 67 0.999 38.62 15.4 20 22 28 37 48 58 63

lowest : 17 18 19 20 21, highest: 79 81 82 83 90

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Employment

n missing distinct

1900 100 8

lowest : Consultant Private PSFederal PSLocal PSState , highest: PSLocal PSState SelfEmp Unemployed Volunteer

Value Consultant Private PSFederal PSLocal PSState SelfEmp Unemployed Volunteer

Frequency 148 1411 69 119 72 79 1 1

Proportion 0.078 0.743 0.036 0.063 0.038 0.042 0.001 0.001

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Education

n missing distinct

2000 0 16

lowest : Associate Bachelor College Doctorate HSgrad , highest: Yr12 Yr1t4 Yr5t6 Yr7t8 Yr9

Value Associate Bachelor College Doctorate HSgrad Master Preschool Professional Vocational Yr10 Yr11 Yr12 Yr1t4 Yr5t6

Frequency 69 345 442 27 660 102 6 24 86 58 74 17 6 23

Proportion 0.034 0.172 0.221 0.014 0.330 0.051 0.003 0.012 0.043 0.029 0.037 0.009 0.003 0.012

Value Yr7t8 Yr9

Frequency 34 27

Proportion 0.017 0.014

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Marital

n missing distinct

2000 0 6

lowest : Absent Divorced Married Married-spouse-absent Unmarried

highest: Divorced Married Married-spouse-absent Unmarried Widowed

Value Absent Divorced Married Married-spouse-absent Unmarried Widowed

Frequency 669 266 917 22 67 59

Proportion 0.334 0.133 0.458 0.011 0.034 0.029

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Occupation

n missing distinct

1899 101 14

lowest : Cleaner Clerical Executive Farming Home , highest: Repair Sales Service Support Transport

Value Cleaner Clerical Executive Farming Home Machinist Military Professional Protective Repair Sales Service Support Transport

Frequency 91 232 289 58 5 139 1 247 40 225 206 210 49 107

Proportion 0.048 0.122 0.152 0.031 0.003 0.073 0.001 0.130 0.021 0.118 0.108 0.111 0.026 0.056

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Income

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95

2000 0 2000 1 84688 72090 15708 22446 34433 59769 113843 189507 233348

lowest : 609.72 1428.27 1598.95 2164.59 2378.63, highest: 385067.54 393393.42 411783.25 421075.30 481259.50

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Gender

n missing distinct

2000 0 2

Value Female Male

Frequency 632 1368

Proportion 0.316 0.684

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Deductions

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95

2000 0 41 0.119 67.57 130.6 0 0 0 0 0 0 0

lowest : 0.0000 416.6667 649.3333 893.3333 920.0000, highest: 2392.0000 2415.0000 2513.3333 2824.0000 2904.0000

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Hours

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95

2000 0 68 0.908 40.07 12.2 18 24 38 40 45 55 60

lowest : 1 2 3 4 5, highest: 80 84 85 98 99

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

RISK\_Adjustment

n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95

2000 0 310 0.487 2021 3658 0 0 0 0 0 7224 10520

lowest : -1453 -961 -615 -402 -199, highest: 28235 77901 88402 99999 112243

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

TARGET\_Adjusted

n missing distinct Info Sum Mean Gmd

2000 0 2 0.534 463 0.2315 0.356

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Rattle timestamp: 2021-03-14 22:47:34 Christine

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

Missing Value Summary

Age Education Marital Income Gender Deductions Hours TARGET\_Adjusted

1899 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1

100 1 1 1 1 1 1 1 1

0 0 0 0 0 0 0 0

Employment Occupation

1899 1 1 0

1 1 0 1

100 0 0 2

100 101 201

Rattle timestamp: 2021-03-14 22:47:35 Christine

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

Answer **any one** (your choice) of the following 3 questions:

Q6: The Age distribution shows a value of 28 underneath the number .25. In your own words describe what this represents.

Out of the 2000 cases, most people fall between the ages of 28 and 48 with 28 year olds representing the 25th percentile or stated another way, 28 year olds fall within 25% of the mean/average age (38.62).

Q7: The Income variable has a mean of $84,688 and its 50% has a value of $59,769. Why do you think the mean is so much larger?  
HINT: Consider how extreme (very large and very small) values affect the mean and median.

Enter your answer in a red font.

Q8: For which one of the categorical variables (if any) would this not be an appropriate order when presenting the results of this project? Why?  
HINT: This question concerns the order in which the individual values within a categorical variable are listed in the Rattle output. For example, for the categorical variable Employment, the values are listed as Consultant, Private, PSFederal, PSLocal, and so on. Look at each categorical variable and consider whether the order in which the values are listed is logical, or would a different ordering make more sense.

Enter your answer in a red font.

Your notes:

Enter your notes on these results in a green font.

# Summary/Basics Results

Copy/paste Basics output from Rattle below:

Basic statistics for each numeric variable of the dataset.

$Age

X...X.i

nobs 2000.000000

NAs 0.000000

Minimum 17.000000

Maximum 90.000000

1. Quartile 28.000000

3. Quartile 48.000000

Mean 38.622000

Median 37.000000

Sum 77244.000000

SE Mean 0.303764

LCL Mean 38.026272

UCL Mean 39.217728

Variance 184.545389

Stdev 13.584748

Skewness 0.499070

Kurtosis -0.396665

$Income

X...X.i

nobs 2000.000000

NAs 0.000000

Minimum 609.720000

Maximum 481259.500000

1. Quartile 34433.117500

3. Quartile 113842.897500

Mean 84688.460045

Median 59768.950000

Sum 169376920.090000

SE Mean 1556.787298

LCL Mean 81635.364424

UCL Mean 87741.555666

Variance 4847173383.239407

Stdev 69621.644503

Skewness 1.488821

Kurtosis 2.208765

$Deductions

X...X.i

nobs 2000.000000

NAs 0.000000

Minimum 0.000000

Maximum 2904.000000

1. Quartile 0.000000

3. Quartile 0.000000

Mean 67.568333

Median 0.000000

Sum 135136.666667

SE Mean 7.618389

LCL Mean 52.627519

UCL Mean 82.509147

Variance 116079.695845

Stdev 340.704705

Skewness 5.249432

Kurtosis 27.541989

$Hours

X...X.i

nobs 2000.000000

NAs 0.000000

Minimum 1.000000

Maximum 99.000000

1. Quartile 38.000000

3. Quartile 45.000000

Mean 40.074500

Median 40.000000

Sum 80149.000000

SE Mean 0.271765

LCL Mean 39.541527

UCL Mean 40.607473

Variance 147.712806

Stdev 12.153716

Skewness 0.132331

Kurtosis 2.914707

$RISK\_Adjustment

X...X.i

nobs 2000.000000

NAs 0.000000

Minimum -1453.000000

Maximum 112243.000000

1. Quartile 0.000000

3. Quartile 0.000000

Mean 2020.962000

Median 0.000000

Sum 4041924.000000

SE Mean 186.529935

LCL Mean 1655.148553

UCL Mean 2386.775447

Variance 69586833.376244

Stdev 8341.872294

Skewness 9.591535

Kurtosis 107.475665

$TARGET\_Adjusted

X...X.i

nobs 2000.000000

NAs 0.000000

Minimum 0.000000

Maximum 1.000000

1. Quartile 0.000000

3. Quartile 0.000000

Mean 0.231500

Median 0.000000

Sum 463.000000

SE Mean 0.009434

LCL Mean 0.212999

UCL Mean 0.250001

Variance 0.177997

Stdev 0.421897

Skewness 1.272187

Kurtosis -0.381730

Rattle timestamp: 2021-03-15 00:38:48 Christine

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

Missing Value Summary

Age Education Marital Income Gender Deductions Hours TARGET\_Adjusted Employment Occupation

1899 1 1 1 1 1 1 1 1 1 1 0

1 1 1 1 1 1 1 1 1 1 0 1

100 1 1 1 1 1 1 1 1 0 0 2

0 0 0 0 0 0 0 0 100 101 201

Rattle timestamp: 2021-03-15 00:38:48 Christine

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

Answer **any one** (your choice) of the following 3 questions:

Q9 If the unit of measure of Age is years then what is the unit of measure for Variance? Standard Deviation?  
HINT: Consider the formulas for calculating variance and standard deviation.

The formula for variance uses squares so the variance has different units than the data for which it was calculated. When interpreting standard deviation, there are no squares so unit of standard deviation matches unit of observation.

Unit of measure for Age (observation) = years

Sample Mean = 38.622 years

Unit of measure for Variance = S2 = = = = 184.5453892 years

Unit of measure for Standard Deviation = S = = = = 13.58 years

Q10 Describe in your own words what being 95% Confident means in this circumstance.  
HINT: Consider what does it mean that, for Age, the lower confident limit of the mean is 38.026 and the upper confidence limit (UCL) of the mean is 39.218.

Enter your answer in a red font.

Q11: Compare the sizes of the confidence intervals for Age and Income. How can you compare the differences?   
HINT: Consider the Coefficient of Variation.

Enter your answer in a red font.

Your notes:

* Low variance = values stay near the average
* High variance = values move away from the average

# Summary/Kurtosis Results

Copy/paste Kurtosis output from Rattle below:

Kurtosis for each numeric variable of the dataset.

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.

Age Income Deductions Hours RISK\_Adjustment TARGET\_Adjusted

-0.3966648 2.2087650 27.5419888 2.9147066 107.4756652 -0.3817296

Rattle timestamp: 2021-03-15 00:39:55 Christine

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

Missing Value Summary

Age Education Marital Income Gender Deductions Hours TARGET\_Adjusted Employment Occupation

1899 1 1 1 1 1 1 1 1 1 1 0

1 1 1 1 1 1 1 1 1 1 0 1

100 1 1 1 1 1 1 1 1 0 0 2

0 0 0 0 0 0 0 0 100 101 201

Rattle timestamp: 2021-03-15 00:39:55 Christine

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

Answer the following question:

Q12 Compare the kurtosis values for Age, Income and Deductions. Which one is the most peaked distribution?

At 27.5419888, Deductions has a sample kurtosis that significantly deviates from 0, indicating the data are not normally distributed. It also has a sharper peak with flatter tails and is the most peaked distribution due its larger Kurtosis value. Age has a smaller peak around the mean and relatively flat distribution with lighter tails due to the lower, negative Kurtosis value (-0.3966648). Income’s positive 2.2087650 Kurtosis value indicates that the distribution has an acute peak around the mean with heavier tails.

Your notes:

* 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

Copy/paste Kurtosis output from Rattle below:

Skewness for each numeric variable of the dataset.

Positive means the right tail is longer.

Age Income Deductions Hours RISK\_Adjustment TARGET\_Adjusted

0.4990696 1.4888209 5.2494323 0.1323312 9.5915351 1.2721873

Rattle timestamp: 2021-03-15 00:40:20 Christine

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

Missing Value Summary

Age Education Marital Income Gender Deductions Hours TARGET\_Adjusted Employment Occupation

1899 1 1 1 1 1 1 1 1 1 1 0

1 1 1 1 1 1 1 1 1 1 0 1

100 1 1 1 1 1 1 1 1 0 0 2

0 0 0 0 0 0 0 0 100 101 201

Rattle timestamp: 2021-03-15 00:40:21 Christine

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

Answer the following question:

Q13 Compare the skewness values for Age, Income and Deductions. Which one is the most skewed? In what direction?

With a Skewness result of 5.2494323, the Deductions category is the most skewed to the right. No negative (or left skewness) exists among these three variables since all skewness results are greater than zero. According to the skewness numbers, the Age category would be closest to a symmetrical/non-skewed distribution since its value is approaching zero.

Your notes:

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)