DATA 511 Intro to Data Science
Course Project 1: Data Prep and EDA
Prof Larose

Name: Chad Zeller

Use the *proj1* data set for all items in this course project.

The project instructions are shown in bold. This is to distinguish the instructions from your work. Your work should be in not-bold.

- Work neatly. Aim for a professional-looking presentation. Report writing is part of the DATA 511 course description, so I will be grading your level of professionalism, as well as your English expression.
- Make sure all graphs and tables fit neatly on the page.
- Neither add nor delete pages.
- There is no need to post your R code in this document. That is what the R script is for.
- At the beginning of the project, set.seed to 12345.
- Whenever I ask you to *manually* construct a table, make sure you don't just copy and paste the R output. Non-coders sometimes have trouble reading R output, even when it is clear to you and me. Instead, use the table function from MS Word, and always include the variable names and categories.
- Pay attention to when I ask for column proportions. Not doing so will guarantee you will miss the most dramatic results in the data set and will cost you big points.

Apart from this document, which you will save as a pdf and submit, you must submit your R script, containing the code you used to solve the problems. The R script should be neat and easily understandable by people who are not you. It should be well-annotated, describing what you are doing so that anyone could understand it.

This Project may have typos, errors, etc, that I have missed. Please report these to me asap. For this and other reasons, this Project is subject to change at any time (though of course I will be reasonable.)

Good luck!

Prof Larose

1. **Insert your Executive Summary here. (A strategy for this is given at the end.)**

This project examines a set of 14,797 customers from our database, to determine which customer characteristics are associated with high-income, defined as having income greater than $50,000. Henceforth, those customers with an income greater than $50,000 will be referred to as "*high-income"* and those with an income of $50,000 or less will be referred to as "*lower income*." The following are among the key findings observed.

- The data set of customers is composed of roughly two-thirds males 9,885 (67%) and one-third females 4,912 (33%).
- Approximately 76% of customers are lower income, compared to 24% in the high-income category.
- Males are nearly three times as likely as females to fall into the high-income category. The proportion of high-income males is ~30% while the proportion of high-income females is only ~11%.
- Capital gains data was missing for 69 customers, *all* of which are in the high-income category. For these customers, the capital gains were imputed (estimated). Both the original and imputed mean and standard deviations are very close together, therefore no systematic differences were introduced because of the imputed data.
- In comparing the relative incomes by marital status, the data shows that ~44% married customers are high-income customers, whereas only ~6% of non-married customers are lower income.
- The income level of customers has an impact on whether the customer pays capital gains or losses. Of the customers who have capital gains or losses, roughly 81% are high-income earners whereas only 19% are in the lower income category. Among customers without capital gains or losses, 57% are in the high-income category compared to 43% in the lower income category.
- The relative level of educational attainment of customers appears to have a positive correlation with their level of income. Of customers with 13 years or less of education, 80% are in the lower income category whereas only 20% are categorized as high-income. In the next tier (14 years) of education, 44% of customers are lower income whereas 56% are high-income. At the highest level of education (14 years of more), 76% of customers are high-income earners while only 24% are lower income.

In summation, the data seems to indicate that some of the characteristics associated with high-income customers includes being a male, being married, having reported capital gains or losses, and higher levels of educational attainment.

Please note that this project is only exploratory in nature. No data modeling has been performed and the next step should be to perform data modeling to predict which customers will be high-income.

2. **Missing Data. Look at a histogram of *capital.gain*. (Don't insert here.) The extreme data values in the right tail all have the exact same value: 99999. It is unlikely that all these individuals have the exact same amount of capital gains. Thus it is likely that the 99999 entry is code for *missing*.**

   a. **Set these 99999 values to missing using something like the following code:**
   ```
   proj1$capital.gain[proj1$capital.gain == 99999] <- NA
   ```
   Then find the mean and standard deviation of *capital.gain*, after Step (a). Something like the following code might be helpful.
   ```
   cgm <- mean(proj1$capital.gain, na.rm = TRUE)
   cgsd <- sd(proj1$capital.gain, na.rm = TRUE)
   ```
   Fill in the following equations. $\mu_{orig} = 603.09$ $\sigma_{orig} = 2610.70$

   b. **Set your seed to 12345 (**set.seed(12345)**). Use *knnImpute* to impute the missing values for *capital.gain*. Make sure the output data set is not the same as the input data set. In other words, the second step uses code something like this:**
   ```
   proj1.imp <- predict(imputation_model, proj1)
   ```

   The *knnImpute* method standardizes all the variables. But we haven't done EDA yet, so this is inconvenient at this early stage. So, de-standardize the imputed *capital.gain*, so that it is on the original scale, with no missing values. Name this variable *cg.imp*, and make sure it belongs to the same data set that was input to the imputation algorithm (*proj1*). Something like the following code might be helpful.
   ```
   proj1$cg.imp <- round(proj1.imp$capital.gain * cgsd + cgm, 5)
   ```

   Complete the following table, which compares some statistics of the imputed capital gains against the original capital gains from Step (a).

Capital Gains		
	Mean	**Standard Deviation**
Original	603.09	2610.70
Imputed	605.94	2606.44

   c. **Comment that the mean and standard deviations are rather close, and that therefore the imputation method worked, and has not introduced systematic differences from the original distribution of *capital gains.* One sentence.**

   The original and imputed mean and standard deviations are both very close together, therefore the imputation method worked, with no systematic differences between the original capital gains distribution.

3. **Construct a flag variable, *cg.miss*, that takes value 1 when *capital.gain* (pre-imputation) is missing, and 0 otherwise. Hint: Use an *ifelse* command. *Manually* (not just R output), Using a table from MS Word, manually (not just R output), construct a contingency table, with *income* as the rows, and *cg.miss* as the columns, showing the counts. Discuss the interesting result you have found.**

	Capital Gains		
		Not Missing (Complete)	**Missing**
Income	<=50K	11,243	0
	>50K	3,485	69

The contingency table shows that capital gains data is missing for 69 customers in total, *all* of whom are categorized as having an income of greater than $50,000 per year. The capital gains values are categorized as not missing (complete) for *all* customers making $50,000 or less per year.

4. **Add a new field, *ID*. To show it is working, fill in the following values for Record #2001.**

Record #2001			
Education	**Sex**	**Income**	**Marital Status**
10	Male	<=50K	Never-married

5. **Consider *marital-status*. Rename *marital-status* as *marital-status-old.***

```
(names(proj1)[names(proj1)=="marital.status"] <- "marital.status.old")
```

Make a new variable, *marital.status,* where the three married categories are combined into the new category *Married*, and the other statuses are combined into the new category *Other*. Construct a contingency table with *income* for the rows and *marital.status* for the columns, *with column proportions* (see notes), rounded to two decimal places. Don't just copy and paste from R. Instead, manually construct a table using MS Word.

Compare the proportions of high income for the two categories and discuss.

	Marital Status		
		Married	Not Married
Income	<=50K	56.20%	93.65%
	>50K	43.80%	6.35%

The above table looks at the income categories of married and non-married (divorced, never-married, separated, widowed) customers. Of the married customers, 43.8% have an income greater than $50,000 per year, whereas only 6.35% of non-married customers have an income greater than $50,000 per year.

6. **Derive a new flag variable, called *capgl*. This flag variable should equal 1 whenever a customer has *either* any (imputed) capital gains *or* any capital losses. It should equal 0 otherwise. Using MS Word, construct a contingency table, with *income* as the rows, and *capgl* as the columns, showing the column proportions.**
   Clearly describe the effect of having any capital gains or losses on *Income*.

	Capital Gains or Losses		
		Does *Not* Have (Imputed) Capital Gains or Capital Losses	*Has* (Imputed) Capital Gains or Capital Losses
Income	<=50K	80.91%	42.68%
	>50K	19.09%	57.32%

The above table looks at the proportion of customers who have no (imputed) capital gains and no capital losses and those who have (imputed) capital gains or capital losses, broken down by income category. Among customers *without* capital gains or losses, just over 80% have an income of $50,000 or less while just under 20% have an income greater than $50,000. Among those customers *with* capital gains or losses, roughly 57% have an income greater than $50,000 per year while those making $50,000 or less account for about 43% of customers with capital gains or losses. Overall, the data seems to indicate that those customers in the higher income bracket are more likely to have imputed capital gains or capital losses while those with a lower income are less likely to have imputed capital gains or capital losses.

7. Outliers. Your professor is not a fan of deleting outliers at the EDA stage, because it often results in changing the character of the data set. Let's demonstrate this!

a. What is the proportion of records with high income, over all records in the data set?

24.02% of records fall into the high-income category of the data set.

b. Consider *capital.loss*. The upper cutoff point for identifying outliers is the mean ($\bar{x}$) plus three times the standard deviation (s), or $\bar{x} + 3s$. Write it here by completing the equation begun in Equation Editor.

$\bar{x} + 3s = 1307.547$

c. Select only the records with values of *capital.loss* greater than this cut-off value. How many are there?

There are *679* records above the cut-off value.

d. What is the proportion of records with high income, among these outlier records?

The proportion of high-income earners is *51.99%* among the outlier records.

e. Describe the change to the character of the data set that will result if we delete these outlier records. State your conclusion regarding deleting outliers at the EDA stage.
If we were to simply delete the outlier records, the integrity of the data set may be compromised. The outliers may provide patterns and insight about the data set that would be lost by omitting them. The unusually large values being removed may also cause the data to become biased towards the mean. In the case of removing the large capital losses, the subset of high-income earners would probably be particularly distorted considering they account for a majority of capital loss outliers yet are slightly below one-fourth of the overall number of cases in the data set. A quick calculation shows that removing the capital loss outliers would result in removing ~3% of the total lower income cases but would remove ~10% of the high-income cases. The latter number seems especially significant and should be avoided. All in all, it seems like a good idea to leave in the outliers during the EDA stage.

8. We would like to bin *education* based on predictive value.

a. Use the method and options shown in the notes to generate the decision tree for predicting *income* based on *education*. No need to copy it here.

Use the *cut* function, along with the split thresholds from your tree in part (a), to construct a new variable *educ.bin*. Manually construct a contingency table here of the counts, with *income* for the rows and *educ.bin* for the columns.

		Education		
		(0, 13]	(13, 14]	(14, 16]
Income	<=50K	10,784	344	115
	>50K	2,753	444	357

b. Redo the contingency table from (a), this time with the column proportions, rounded to two decimal places.

		Education		
		(0, 13]	(13, 14]	(14, 16]
Income	<=50K	79.66%	43.65%	24.36%
	>50K	20.34%	56.35%	75.64%

c. Discuss your results from (a) and (b).

Of those customers where the individual has less than 13 years or less of education, 10,784 (~80%) have an income of $50,000 or less, while only 2,753 (~20%) have an income greater than $50,000. Of those with 14 years of education, 344 (~44%), have an income of $50,000 or less, while 444 (~56%) have an income of greater than $50,000. Among those cases with greater than 14 years of education, 115 make $50,000 (~24%), while 357 (~76%) make more than $50,000. There appears to be a correlation between higher income and a higher level in educational attainment.

Exploratory Data Analysis

9. **Using *ggplot2,* construct the following stacked bar graphs of *educ.bin* with overlay of *income*. (The lingo here means that *educ.bin* will be on the horizontal axis, and *income* will represent the colors.) Insert them so that they both fit in the table I provided below.**
   a. **Non-normalized stacked bar graph.**
   b. **Normalized stacked bar graph.**
   c. **In one sentence, describe the distribution of *educ.bin* regardless of *income*. Don't be vague!**

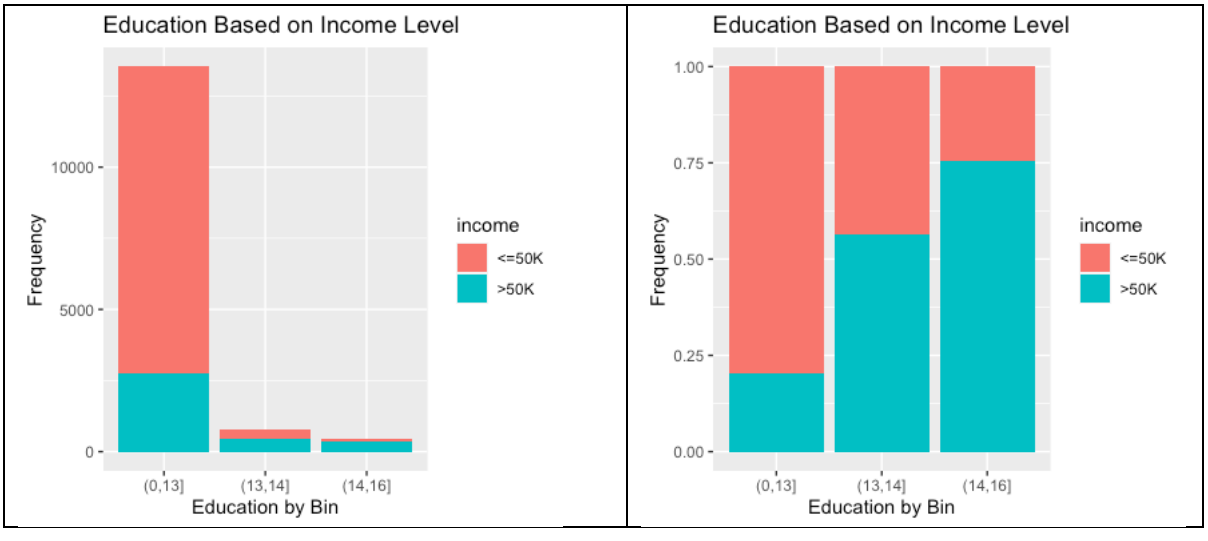
      The distribution shows that a large majority of customers at both levels of income, those making $50,000 or less and those making more than $50,000, fall into the bin of 13 years or less of education while a small proportion fall into the two bins encompassing more than 13 years of education.

   d. **In one sentence, describe the distribution of *educ.bin*, with respect to *income*. Don't be vague!**

      The distribution of educ.bin shows that as the levels of education attained rises with each bin, the proportion of customers with an income of greater than $50,000 rises.

   e. **Briefly state the benefit of using the non-normalized version and of using the normalized version.**

      A non-normalized stacked bar graph gives you a good graphical representation of the total (raw) *count* of customers falling into each bin whereas a normalized bar graph is more useful in illustrating the relative *proportions* of cases that fall into each bin.



10. Provide the following. For the contingency tables, make sure sex is the columns and income is the rows. For (c) – (f), don't be vague!

a. Manually construct a table of counts, with row and column totals.

Income	Sex			
		Female	Male	Total
	<=50K	4,348	6,895	11,243
	>50K	564	2,990	3,554
	Total	4,912	9,885	14,797

b. Manually construct a table of column proportions, rounded to two decimal places, with totals provided for the columns only (not the rows).

Income	Sex		
		Female	Male
	<=50K	88.52%	69.75%
	>50K	11.48%	30.25%
	Total	100.0%	100.00%

c. In one sentence, describe your results from (a).

Table (a) shows that there are slightly more than twice as many males (9,885) as females (4,912) out of a total of 14,797 cases.

d. In one sentence, describe the effect of *sex* on *income* from (b).

Table (b) shows that the proportion of males in the high-income bracket is 30.25% while the proportion of females in the high-income category is only 11.48%, which seems to indicate that males are substantially more likely to make more than $50,000 in income.

e. In one sentence, how is the table in (a) preferable to the table in (b)?

Table (a) is preferrable to table (b) if we are interested in seeing the *raw counts* of the cases broken down by sex and income level.

f. In one sentence, how is the table in (b) preferable to the table in (a)?

Table (b) is preferable to table (a) if we are interested in seeing the *relative proportions* of the cases broken down by sex and income level.

Craft your Executive Summary as follows.

Your boss makes more money than you. He or she has little time for the arcane details of all the data prep and other work you did to produce your report. Your boss is only interested in RESULTS.

A good executive summary should consist of the following.

1. **A quick summary of the *Objective* of the analysis. For this project, this would be something like the following: "This project examines a set of [this many] customers from our database, to determine which customer characteristics are associated with high income, defined as having income greater than $50,000." Feel free to copy and paste this sentence in your executive summary. Also include the original proportion of high-income customers.**

2. **Bullet points with explanations of your most salient results. I think you can make good bullet points with your results from the following problem numbers:**
    a. **Problem 3**
    b. **Problem 5**
    c. **Problem 6**
    d. **Problem 8c**
    e. **Problem 9a and 9b**
    f. **Problem 10c and 10d**

3. **What NOT to include in your Executive Summary is anything about data prep, unless it affects managerial policy. I think it is safe to assume nothing in this project affects managerial policy. Problem 3 is EDA drawn from data prep, so deserves a mention.**

4. **Brief mentioning of next steps. This helps to delimit the scope of the project. For this project, you may say something like, "Note that this project is exploratory only. No actual data modeling has been performed. Rather, our next step should be to perform data modeling to predict which customers will be high-income."**

5. **And, whatever you do, do not exceed one page!** 😉

==========

Well done!

<u>Deliverables:</u>

1. **Save your completed Word document as a pdf file, named *Doe_Jane_Project1* (if your name is Jane Doe, with last name first!). Because of virus issues, no Word documents will be accepted.**
2. **Your well-annotated R script, named *Doe_Jane_Project1_RScript.***

Do NOT zip these two files together. Rather, make two separate submissions using the Project Submissions Tool.